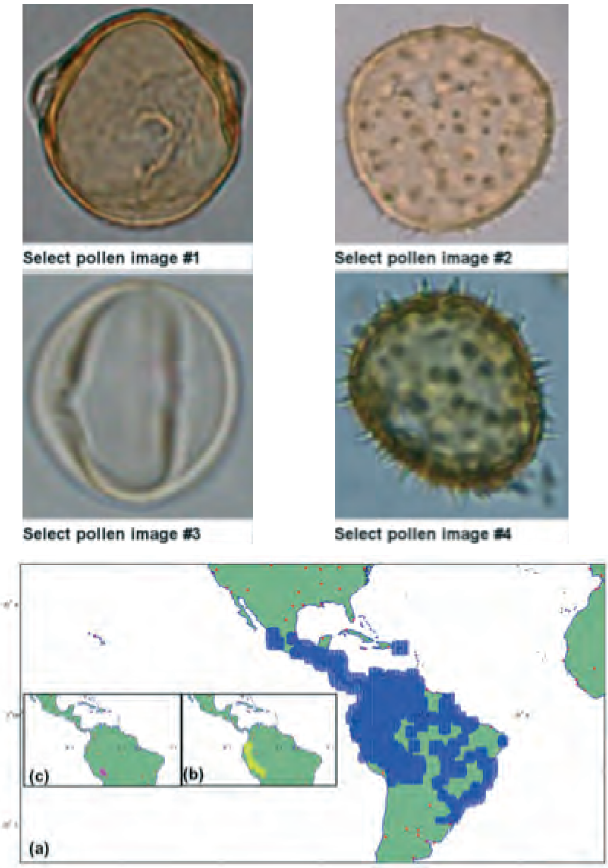
AASP- THE PALYNOLOGICAL SOCIETY



Example of pollen and geolocation.
© 2013 The MITRE Corporation. All Rights Reserved.
Approved for Public Release; Distribution Unlimited.
Case # 13- 2331

NEWSLETTER

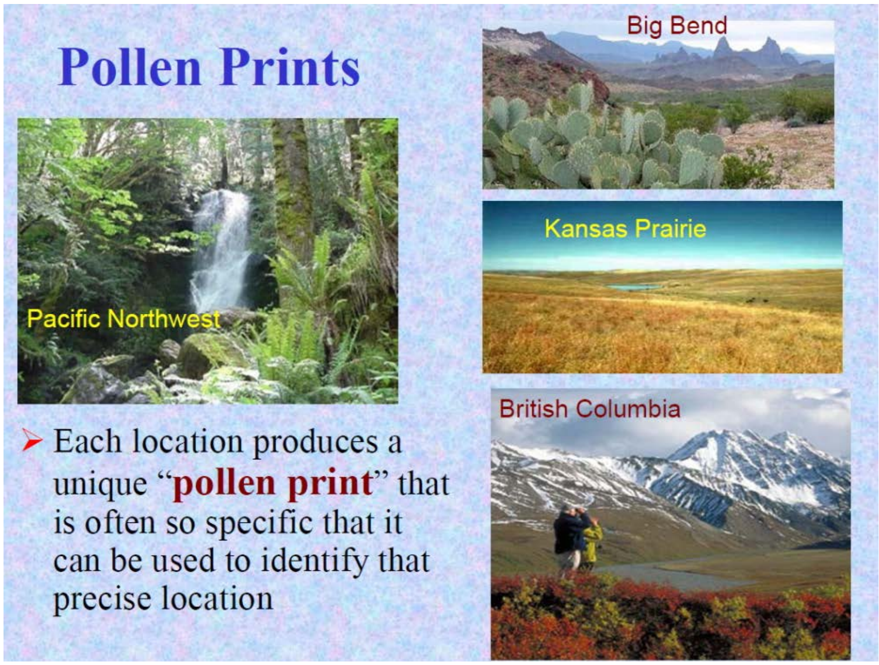
July 2013

Special Issue:
Palynology and Geolocation

ISSN 0732-6041

Published Quarterly by AASP — The Palynological Society

Forensic Geolocation Challenge: Is Pollen Analysis the Answer?



Pollen Prints

- Each location produces a unique "**pollen print**" that is often so specific that it can be used to identify that precise location

Figure credit: Vaughn Bryant, Texas A&M University

A Report of the 09/12/2012
Pollen Coalition Workshop

Grace M. Hwang
The MITRE Corporation
McLean, VA

David Masters
US Department of Homeland Security
Washington, DC

MITRE Innovation Program

FORENSIC GEOLOCATION
POLLEN COALITION MEMBERS

Grace M. Hwang (Organizer), Principal Scientist, *The MITRE Corporation*
David Masters (Co-Organizer), Scientific Advisor, *US Department of Homeland Security (DHS)/Science & Technology (S&T)*

Vaughn M. Bryant (Keynote Speaker) Professor and Director of the Palynology Laboratory, Department of Anthropology, *Texas A&M University and Trustee of the American Association of Stratigraphic Palynologists (AASP) – The Palynological Society*
Mark Bush, Professor, Department of Biological Sciences, *Florida Institute of Technology (FIT)*
Carol Christou, Principal Sensor Systems Engineer, *The MITRE Corporation*
Eric Grimm, Curator and Chair of Botany, *Illinois State Museum*
Garry Jacyna, Fellow, *The MITRE Corporation*
Susan McCarthy, Acting Associate Director for Public Services, Head, Strategic Programs Branch, National Agricultural Library, *US Department of Agriculture (USDA)*
David Korejwo, Geologist, *Department of Justice (DoJ)/Federal Bureau of Investigation (FBI)*
Nicholas M. Orlans, Senior Principal Engineer and Division Chief Engineer for Biometrics, *The MITRE Corporation*
Surangi Punyasena, Assistant Professor of Plant Biology, *University of Illinois*
Kim Riley, Senior Signal Processing Engineer, *The MITRE Corporation*
Peter Siska, Professor and Chair of Regional Studies, *United States Military Academy*
Libby Stern, Research Chemist - Stable Isotope Specialist, *DoJ/FBI*
James Sweet, Director Southwest Regional Science Center, *DHS/Customs & Border Protection (CBP)*
David Tcheng, Researcher, *University of Illinois*
Sophie Warny, Assistant Professor of Palynology/Department of Geology and Geophysics Curator/Museum of Natural Science, *Louisiana State University (LSU) and AASP Board member*
Debra Willard, Program Coordinator, Climate and Land Use Change Research & Development, *US Geological Survey (USGS)*

MITRE Editorial Staff
Janice Ballo, Lead Information Management
Margaret MacDonald, Senior Writer-Editor
Elaine Mullen, Senior Multi-Discipline Systems Engineer

MITRE Reviewers
Tonia Korves, Senior Multi-Discipline Systems Engineer
Regina Ryan, Group Leader – Geospatial Computing
Paul Silvey, Chief Engineer – Data Mining
Jeff Woodard, Principal Signal Processing Engineer

PREFACE AND ACKNOWLEDGEMENTS

In August 2012, the Department of Homeland Security (DHS) Science and Technology (S&T) Directorate tasked The MITRE Corporation to conduct an *advanced study* to determine the feasibility of applying palynology – the analysis of pollen grains – to forensic geolocation. This effort was organized by the Systems Engineering Development Institute (SEDI), one of the Federally Funded Research and Development Centers (FFRDC) managed by MITRE that focuses on DHS systems engineering challenges. The advanced study addressed the operational and technical requirements for determining the geographic origin of physical goods based on the relative abundance of indicator pollen types. Geographic identification of goods based on factors that do not rely on electronic surveillance or tracking will critically support numerous applications related to national and economic security. Potential applications of this technology include improving compliance with US import controls and tariffs and supporting criminal and civil forensic investigations. Specifically, the top priority of the DHS for this advanced study has been the development of statistically rigorous methods for determining the accuracy and precision of pollen-based geolocation of products originating in Mexico, Central America, and South America.

To conduct the advanced study, MITRE surveyed the palynology and forensic communities and discovered that several disparate pollen databases covering the Neotropics are under development by the research community. These databases include Illinois State Museum's Neotoma Paleoecology Database; Louisiana State University's (LSU) Pollen Database, which is maintained by the Center for Excellence in Palynology (CENEX); and the Neotropical Pollen Database, which is maintained by the Florida Institute of Technology. As separate projects, these databases have different purposes (*e.g.*, paleoecology, paleontology, paleoclimatology, or palynology) and unfortunately have not created the metadata that would be required by applications in forensic geolocation. Collectively, the survey of the pollen research community reflected a consensus that the development of a comprehensive and versatile database of pollen grain images would be valuable to multiple stakeholders; such a database would be oriented around visual search and would include scalable and reconfigurable schema.

Already, the US Department of Agriculture's National Agricultural Library has plans to digitize a national archival collection of pollen flora that focuses on the Southeastern United States and Northern Mexico. Moreover, the US Geological Survey (USGS) cooperates with law enforcement agencies by providing georeferenced information about identified pollen assemblages that help determine place-of-origin within the United States [1]. Recently, DHS Customs and Border Protection built a forensics laboratory to conduct high-throughput palynology at the Southwest borders; DHS has solicited information on specific equipment and processes that are currently used for pollen-grain identification. Given the diversity of these ongoing efforts, MITRE decided to bring together researchers and stakeholders to discuss current state-of-the-art methods for applying palynology to forensic geolocation. DHS S&T and MITRE hosted the first Pollen Coalition Workshop on September 12, 2012, bringing together world-renowned palynologists, forensics lab directors and examiners, climatologists, and other scientists from both government and academic communities.

This consensus report captures the ideas and viewpoints presented at the Pollen Coalition Workshop. The report begins with an Introduction, followed by three chapters that reflect each of the central themes of the workshop: 1) Forensic Geolocation Based on Pollen Analysis, 2) Unified Massive Global Pollen Database, and 3) Automation Algorithms for Pollen Identification. These thematic chapters are followed by a chapter of recommendations developed by consensus during the workshop. Finally, the report concludes by presenting the workshop agenda and speaker biographies.

This workshop and the associated consensus document were financially supported in part by The MITRE Corporation's Innovation Program, and also by internal funding from various participants' organizations.

This report could not have been prepared without the dedication of the workshop participants. Although several participants reviewed this report in draft form, and many reviewers provided constructive comments and suggestions, they were not asked to endorse the content of this report prior to public release. The institutions in which authors are employed do not necessarily endorse opinions expressed in this report.

Dr. Hwang is indebted to Professors Vaughn Bryant and Mark Bush for their thorough critique of this report. Dr. Hwang is also grateful to all the Workshop presenters, moderators, and MITRE reviewers – especially Carol Christou and Garry Jacyna – who provided insightful comments, corrections, and original content, many of which are reflected in this report.

EXECUTIVE SUMMARY

The Department of Homeland Security (DHS) Customs & Border Protection (CBP) identified an operational need for improved geolocation analytics based on pollen from Mexico, Central America, and South America; these regions will now be collectively referred to as the Neotropics. DHS Science and Technology (S&T) asked The MITRE Corporation to quantify the precision and accuracy of pollen-based geolocation. Accordingly, MITRE hosted a workshop on 9/12/2012 with participants from government and academia to identify the best available data sources, pollen identification techniques, and geolocation approaches. This report summarizes the outcome of the Workshop.

Chapter 1 describes the challenges associated with improving precision and accuracy of pollen grain identification for the purpose of understanding the reliability of using information on pollen grain distribution as the sole geographic indicator for goods originating in the Neotropics. The Pollen Coalition believes that the minimum number of plant families required to determine country-of-origin from the Neotropics depends on whether key insect-dispersed pollen types belonging to a few plant families can be identified. The framework for prioritizing the top 1,000–10,000 species within these families should consider which pollination strategies and geographic locations have the greatest forensics value. DHS should focus on certain families such as the *Anacardiaceae, Araliaceae, Arecaceae, Asteraceae, Euphorbiaceae, Fabaceae, Malvaceae*, *Marcgraviaceae*, and *Rubiaceae*. Accurate identification of key pollen types from a few families would contribute significantly to the success of geolocation efforts.

The biological resolution at which pollen should be classified is a function of cost and time. This calls for a tiered strategy in which genus-level classification is conducted first, followed by species-level classification, if such biological resolution is necessary.

Of equal importance is the geographic coverage of pollen data. Palynologists rarely take spatial analysis into consideration when designing their surface sampling plans; this omission could lead to geographic gaps. Clearly a standard for determining collection points, as well as for methodology of collection, is warranted.

Chapter 2 describes the types of databases and slide repositories that are already available, and associated gaps that would have to be filled to enable pollen-based geolocation in the Neotropics. As the presentations summarized in this chapter reveal, several organizations are already developing digitized sets of global pollen data. The Coalition recognizes an opportunity to align resources and develop a unified standard that would allow continued sharing of resources via a "unified massive global pollen database." A target schema that includes statistically sampled, recognized, and counted taxon instances at genus and species levels, as well as spatial-temporal information, is highly desirable. Additional data layers for biomass, temperature (max, min, and seasonal ranges), soil type, slope, surface roughness, and low cloud cover could be overlain to refine the distributional model for each pollen type. Current databases provide many of these parameters, although the scale and degree of interpolation of the data vary somewhat. In addition, this next-generation database must be flexible enough to account for taxonomic changes due to reclassification by the plant botanical community. The Coalition should reach consensus on a schema for the next-generation unified massive global pollen database as soon as possible. Furthermore, the community needs a set of auditable standardized approaches for sample processing and identification.

Chapter 3 summarizes the state of the art in automatic pollen identification techniques, with particular emphasis on semi-automatic pollen identification algorithms for use in conjunction with light microscopy (LM). Sample preparation methods are likely to affect machine-learning algorithms that rely on the wavelength of light, especially preparation methods that change the spectrum of the grain (*e.g.*, fresh versus acetylized pollen grains). Most of the classification algorithms discussed operated on manually segmented data, so a sizeable gap remains between true automation and human-assisted automation. By the conclusion of the Workshop it became apparent that the Coalition's goals should focus on developing algorithms that would assist a human-in-the-loop rather than pursuing true autonomous automation for pollen identification. For example, a reliable query tool that would rapidly conduct searches in a next-generation "unified massive global pollen database" could be based on one of three classes of algorithms: graph-theoretic, computer vision, and group-theoretic methods. No *a priori* method exists to determine which of these three classes of algorithms would be the most accurate and computationally efficient for classifying pollen grains from the Neotropics. Instead, the community would have to devise a benchmark to test each class of algorithms using a common set of representative test images.

Chapter 4 contains consensus recommendations in furtherance of DHS' geolocation goals, categorized by near-term (1 yr), mid-term (2–5 yrs), and long-term (5–10 yrs) for five thrust areas, as follows:

Thrust 1 – Design a schema and populate the unified massive global pollen database comprises near-, mid-, and long-term tasks. In the near term, the Coalition should generate a list of no more than 1,000 indicator taxa that are especially important for forensic geolocation in the Neotropics. The Coalition should then obtain or borrow good reference slides for those pollen types, photograph them, and enter them into a new pollen database. Key members of the Coalition should have the authority to add further metadata to each image (*e.g.*, pollination mechanism(s), whether pollen is associated with arid or mesic regions, whether the plant grows best in sunlight or closed canopy, elevation, geographic coordinates, and scientific references to support each entry when available). In the mid-term, the Coalition should consider visiting key herbaria and requesting permission to gather flowers and make pollen reference slides from indicator genera that are missing from any of the existing pollen reference collections. In the long term, to determine whether the magnification of microscope objectives used in forensic lab instrumentation should match that of high-quality, low-throughput archival labs (*e.g.,* 1000X), the Coalition may wish to address discrepancies in resolution (*e.g.,* 40X, 100X, 1000X) and dimensionality (2D versus 3D) used in pollen identification algorithms

Thrust 2 – Develop predictive algorithms and geolocation analytics has two distinct near-term tasks and one mid-term task. For the first near-term task, the Coalition should focus on developing weighted habitat isolation and range (WHIR) models needed to assist in determining the best indicator taxa in relation to Thrust 1. The initial step would be to create WHIR models for Central and South American locations commensurate with areas of greatest interest to the government. For the second near-term task, the Coalition should further develop continuous joint probability estimators based on the relative abundance of indicator taxa at the genus or species level, using pollen data from existing collections (*e.g.*, Florida Institute of Technology and others), and taking into account elevation, latitude, and longitude. For the mid-term task, the Coalition should evaluate air filters to determine whether any meaningful geolocational pollen grains could be extracted and identified.

Thrust 3 – Set standards for pollen collection/purification/mounting/acquisition is a near-term task that is likely to be a continuous activity. The Coalition should devise a set of standardized protocols governing collection strategy, extraction methods, microscopy mounting media, and magnification/acquisition modes specific to forensic geolocation. It is critical that other trace materials be preserved in the sample preprocessing steps. These trace materials include soils and anthropogenic matter (*e.g.*, glass, chemicals and fly ash). Therefore, nondestructive extraction methods are preferred over destructive methods. Although this thrust is written for microscopic methods, the Coalition should be flexible in considering spectroscopic and biological techniques for pollen identification in the future, should these other methods become financially tractable and technically feasible.

Thrust 4 – Initiate collection and digitization for low-throughput, high-quality archives (*e.g.*, scanning electron microscopes [SEMs]) is a long-term task. The Coalition recognizes that collecting up to 10,000 specimens of targeted pollen types will require extensive travel, as individual herbaria only hold a small proportion of the total flora. It is also possible that the flowers of some taxa have never been collected or that the herbarium sheets are of fruiting specimens, spent flowers or sterile plants. Therefore, the Coalition proposes that DHS establish a distributed Center for Global Pollen Informatics over the next decade to build a large reference database based on to-be-developed standards (see Thrust 3). Regional herbaria, *e.g.*, Lima and Quito, have many specimens, and identification at the genus level is probably accurate, although species identifications are unlikely to be correct. Combined with the knowledge that most Neotropic pollen grains are small (< 13 microns) and featureless, the Coalition therefore recommends the use of SEM over LM in the long run for building a high-quality reference database.

Thrust 5 – Initiate digitization for a high-throughput forensics laboratory has both near-term and long-term components. The Coalition recognizes that database integration and curation of existing resources can be advanced in the near term even if new standardized workflows specific to data collection for geolocation are not yet funded. As demonstrated by the various presentations at the Workshop, different organizations have different database schemas to serve their particular research objectives and requirements. Since a thorough effort to fully populate the "unified massive global pollen database" may not occur in the near term due to limited resources and risk-reward investment decisions, the Coalition suggests that DHS make better use of resources that already exist by incorporating new and improved data and processes in evolutionary ways. In the long term, DHS should encourage experiments in algorithm improvement (*e.g.*, automatic segmentation, grain classification), especially if benchmarks that establish best-of-breed algorithms were to be completed under Thrust 1.

Table of Contents

List of Figures

List of Tables

Introduction

The field of palynology has many applications, including paleoecology [2, 3], paleoclimatology [4], ecosystem restoration [5], medical research on asthma [6, 7] and pollen allergies [8-10], forensics [11-16], biostratigraphy [17], archeology [18] and oil exploration [19]. The workshop covered by this report focused primarily on forensic geolocation: the ability to ascertain where a product (or person) of interest has been geographically within a spatio-temporal context [20]. The underlying hypothesis is that certain assemblages (or distributions) of indicator pollen taxa could be used to infer the place of origin of an artifact with a given statistical confidence, and that this confidence level would far exceed chance. Completeness, precision, and accuracy are all important aspects that must be considered when computing geolocation probabilities. In this context, *precision* refers to the size of the area inferred from the pollen assemblage, while *accuracy* refers to the degree of confidence of the inference. *Completeness* indicates how thoroughly pollen has been cataloged in a particular region of the world geographically and taxonomically. The consensus opinion is that it is difficult to quantify the degree of uncertainty associated with geolocation probabilities based upon pollen assemblages given the scale of the problem and cost constraints associated with pollen genus and species identification. Forensic pollen identification is very challenging; and one underlying, universal technical enabler would be the development of a massive global pollen database with adequate metadata, including georeferencing information.

In the spring of 2012, MITRE learned through the Department of Homeland Security Science & Technology (DHS S&T) that the DHS Customs and Border Protection (CBP) laboratories were interested in enhancing geolocation analysis capabilities, and that DHS S&T had assumed the responsibility for developing a strategy. In the months that followed, MITRE and DHS S&T confirmed that so far, no commercial off-the-shelf (COTS) hardware/software products can perform automated pollen analysis without a human-in-the-loop, and that algorithms needed to correlate pollen distribution with geolocation are also not commercially available. Although scientists have tried to develop automated pollen identification (ID) systems since 1968, no validated system that would work for any region of the world, especially Mexico and Central and South America, is close to being commercial availability. DHS S&T, however, cannot wait for a COTS solution, given the importance and timeliness of forensic geolocation, but instead will fund low-risk exploratory research to arrive at a near-term, if partial, solution.

In August 2012, DHS S&T asked MITRE to conduct an advanced study to quantify the best achievable precision and accuracy of pollen-based geolocation using publicly available datasets. MITRE soon realized that several academic institutions had already developed databases that would fulfill some aspects of DHS's needs. Therefore, to attain the best value for DHS S&T, MITRE recommended that a workshop be organized involving disparate academic communities, along with several members of the forensics community and civilian agencies that have overlapping pollen identification interests. The first Pollen Coalition Workshop convened on 9/12/2012 and contributed to the generation of the present report.

Workshop Invitation

In collaboration with DHS S&T MITRE circulated an invitation, excerpts of which are included in this section.

Pollen Coalition Workshop

Organizers:

Grace M. Hwang gmhwang@mitre.org
The MITRE Corporation
David Masters david.masters@HQ.dhs.gov
Department of Homeland Security (DHS)
Science and Technology (S&T) Directorate

www.paldat.org www.neotomadb.org

The MITRE Corporation – a not-for-profit organization that exclusively operates Federally Funded Research and Development Centers (FFRDC) – invites you to participate at the first Pollen Coalition Workshop scheduled for 12 September 2012 from 0815 to 1800 in McLean, VA, USA.

Purpose of meeting: In collaboration with DHS S&T, The MITRE Corporation has identified an operational and technical need to evaluate the potential for determining the place of origin of pollen grains to improve compliance with US import controls and tariffs and aid forensic investigations of criminal activities. Recognizing that law enforcement and other agencies have developed methods to identify pollen grains and their places of origin, DHS S&T would like to learn about equipment and processes already in use to make these identifications. In parallel, DHS S&T has tasked The MITRE Corporation to conduct an *advanced study* on the accuracy and precision of geolocation determinations for pollen grains from the Neotropics (*i.e.*, Mexico, Central and South America). Associated with the advanced study is the present effort to survey the palynology community and develop a series of agenda topics for future Pollen Coalition workshops. MITRE has learned through surveying the palynology and sponsor communities that the development of a visually queryable structured dataset of pollen grains with an infinitely scalable and reconfigurable schema would be valuable to multiple stakeholders. We understand that there are many issues that can potentially be addressed through a detailed knowledge of foreign pollen grains entering the United States. Although our focus is on geolocation for treaty verification purposes in Mexico and Central and South American countries, our goal is to construct a database that could be used to facilitate investigations in multiple domains and locations. Figure 1 describes the pollen identification and geolocation process from the viewpoint of DHS CBP.

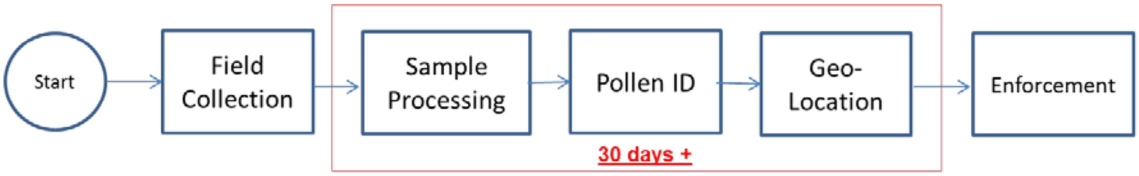


Figure 1. Steps Involved in Pollen Collection, Analysis, and Exploitation

Problem statement: Pollen sample processing, recognition, and geolocation can take a long time and currently requires participation of highly trained scientists.

Operational requirements:

1. Establish methods to prevent evidence contamination while preserving the integrity of pollen grain samples (and other trace materials) during collection and processing.
2. Develop a visual database of reference pollen-feature sets and images for geographical regions of interest.
3. Capture, archive, and visualize images of pollen samples and associated morphologies.
4. Enable lab technicians to process, identify, and geolocate most samples within a week in laboratories at or near the place where the samples were collected; some samples may still have to be shipped offsite for evaluation by pollen experts.
5. Compile a geospatial database of existing botanical datasets for regions of interest.
6. Produce match statistics (by frequency of occurrence and taxa) for accurate geolocation at regional levels within a country.

Should a successful pollen-based forensics capability be realized, DHS CBP anticipates savings from reduced sample transportation and warehouse costs, and lower laboratory service fees; reduced time required to determine samples' points-of-origin and routes of transit to the United States, and better control of the process for tracing them; the ability to derive new evidence for criminal investigations that would otherwise not be available; and overall enhancement of CBP's ability to enforce trade compliance and tariff collections (*i.e.*, geolocation accuracy and volume).

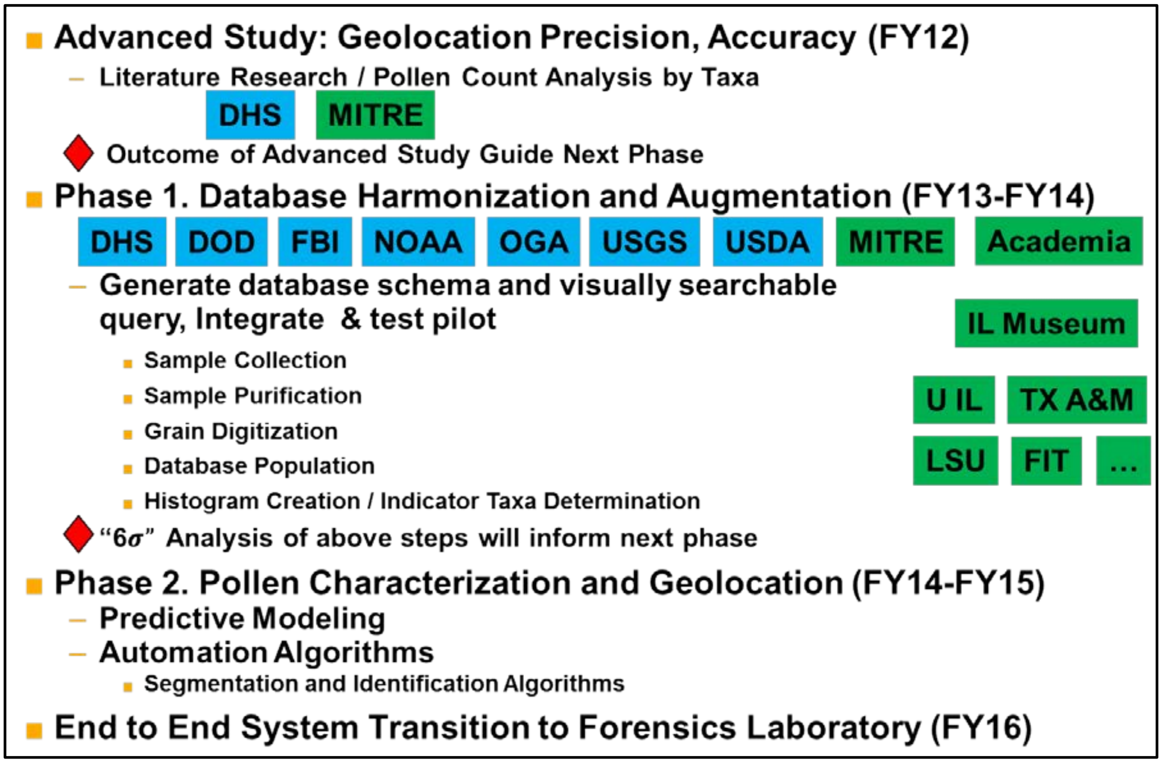
The workshop invitation summarized a set of example questions that provided context for invited presenters:

- Geolocation: What is the minimum number of species needed to conduct country-of-origin determination for Mexican and Central and South American flora?
- Geolocation: How accurate must the pollen classifier be to provide country- or regional-level geolocation information, assuming that updated pollen data are available on a weekly (or monthly) time-scale? How is the freshness of pollen grains determined?
- Database: What arguments support choosing a Hadoop map-reduced solution over an open SQL solution?
- Automated Pollen Recognition Algorithms: What typical sample sizes (*e.g*., number of images/orientations per taxon) are required to train algorithms to segment and classify pollen grains accurately? What are the typical distributions of pollen grain orientation on slides? What is the relationship between classifier accuracy and the size of training sets?
- Is there a universal standard for pollen grain collection, processing, and identification?

Finally, the invitation asked all participants to contribute their expertise and experience to the formation of a visually searchable, structured pollen database. Methods that would enable the automation of steps involved in populating the pollen grain database were of particular interest.

Workshop Objectives

As stated in the original workshop invitation, DHS expected to hold a series of meetings. The first meeting would permit participants to propose and formalize the objectives of the Pollen Coalition, and establish collaboration among stakeholders. DHS anticipated two phases for accomplishing the objectives of the Coalition: Phase 1: Database Schema Development (24 months); and Phase 2: Geospatial Database Integration, bringing together data from existing datasets, and determining methods to improve related efficiencies (24 months). Increased efficiency may be realized through automation at multiple levels, from sample collection, purification, segmentation, and grain digitization to database population and histogram creation (see Fig. 2). Successful implementation of the first two phases should result in the transition of a visually searchable database to the DHS CBP forensics laboratory.



- Advanced Study: Geolocation Precision, Accuracy (FY12)
  - Literature Research / Pollen Count Analysis by Taxa

    DHS MITRE

  ◆ Outcome of Advanced Study Guide Next Phase
- Phase 1. Database Harmonization and Augmentation (FY13-FY14)

  DHS DOD FBI NOAA OGA USGS USDA MITRE Academia

  - Generate database schema and visually searchable query, Integrate & test pilot
    - Sample Collection
    - Sample Purification
    - Grain Digitization
    - Database Population
    - Histogram Creation / Indicator Taxa Determination

  IL Museum

  U IL TX A&M

  LSU FIT …

  ◆ "6σ" Analysis of above steps will inform next phase
- Phase 2. Pollen Characterization and Geolocation (FY14-FY15)
  - Predictive Modeling
  - Automation Algorithms
    - Segmentation and Identification Algorithms
- End to End System Transition to Forensics Laboratory (FY16)

Figure 2. DHS S&T's Goals for the Pollen Coalition

Workshop Organization

The workshop began with a 40-minute keynote speech by Professor Vaughn Bryant of Texas A&M University. The keynote speech was followed by 20-minute presentations given by academic researchers and government stakeholders from law enforcement and civilian agencies. Afternoon sessions centered around three themes addressed sequentially: 1) Gaps in Forensic Geolocation, 2) Challenges Associated with Developing a Unified Massive Global Pollen Identification Database for Geolocation, and 3) Automatic Recognition and Classification of Pollen Grains: from Taxa to Species.

Workshop Observations

During the workshop, it became clear that forensic pollen analysis for geolocation in the United States has largely been the work of Professor Vaughn Bryant. Professor Bryant identified Pat Wilshire and Dallas Mildenhall as leaders in forensic palynology in the United Kingdom and in New Zealand, respectively. Some experts also work in Australia, Canada, Europe, and a few Asian countries where forensic palynology is more commonly practiced than in the United States, primarily because the United States has been unable to create academic programs that teach forensic palynology at the doctoral level and to offer jobs that would encourage future palynologists to specialize in forensics.

Professor Bryant credited Lennart von Post for developing the science of pollen analysis in 1916 [21]. The earliest reported application of forensic pollen analysis took place in 1959 in Sweden, when the scene of a murder was determined to be different from the location at which the body was found [22], and in Austria, where a murderer was identified by pollen from the crime scene that was trapped on his boots [22-24]. The earliest known forensic application of palynology in the United States occurred in 1975, when Professor Bryant was asked to assist the USDA in confirming the origin of domestic honey as part of the required compliance with the US-subsidized Farm Loan Program.

To determine geolocation, Professor Bryant explained that pollen samples can be taken from a myriad of objects including boats, airplanes, skeletons and buried remains, carpets and floor surfaces, improvised explosive devices and letter bombs, clothing of all sorts, books and documents, computers and cell phones, all parts of vehicles including air filters, and the contents of transported products [11, 12, 20]. Following Professor Bryant's overview, the FBI's Trace Evidence team presented a case in which pollen taken from the interior and tires of a car, and from dirt deposits found on a shovel, assisted investigators in narrowing the search area for a roadside grave [1]. This particular example illustrates the limitations of pollen analysis for locating a structure of such relatively small scale as a grave: the FBI could narrow the search to an area spanning Northwest Pennsylvania through Central Ohio, but could not resolve the area to a scale that would readily help to determine the location of the grave. In general, pollen is considered one of several elements of trace evidence that, in combination with geologic and anthropogenic data, can contribute to geo-attribution by collectively improving the confidence of the inference [20, 25, 26].

Furthermore, it became clear during the workshop that research goals will affect the physical collection methods used by pollen experts who are asked to identify and report on pollen grains. For example, in academia, formal standards for sample collection are created *ad hoc*, and site selection is based on the motivation of the particular study (*e.g.*, air traps above buildings would be appropriate for studies focused on allergies; sediment traps would be appropriate for paleoecology studies; moss polsters would be appropriate for palynology studies). In law enforcement scenarios, reference samples are collected relative to the evidence in question. But whether pollen grains collected and catalogued at or near a particular crime scene arrived prior to or immediately after the crime is difficult to determine. Without ample pollen coverage, the operational efficacy of geolocation is likely to be compromised.

Unlike the United States, where the USGS has published atlases of a large collection of US pollen grain images taken with a high-resolution (1000X) light microscope (LM) [27-29], the pollen coverage for the Neotropics is marginal at best. Moreover, the USGS emphasized the importance of tailoring physical processing techniques to the depositional environment and goals of the project in order to maximize the pollen concentration, minimize potential pollen losses and reduce counting time. In the experience of USGS palynologists, pollen concentration in samples could vary from nearly zero to five million grains per gram.

It remained unclear throughout the workshop how many taxa would be needed to conduct forensic geolocation of goods and evidence from the Neotropics, and whether pollen ID at the species level would be superior to ID at the genus level. While pollen ID at the species level is ideal and represents the most accurate means of differentiating pollen grains, the cost of establishing a species-specific database is prohibitive.

During the workshop, Mr. David Masters emphasized the importance to the US government of big problems such as geolocation. Although prior endeavors to employ pollen analysis for geolocation may not have succeeded, he believes it is important to hear all stakeholders' thoughts and perceptions on the potential of pollen forensics research to support government requirements. While security is of paramount importance to many US Government agencies, other stakeholders, such as climate change researchers, can benefit from research in this area. A pollen database is critical to this research, and the relevant government agencies are trying to figure out how to structure one and which methods to employ. Building a coalition is very important, and this workshop represents a first step in that direction.

Dr. James Sweet, Laboratory Director of the CBP Southwest Regional Science Center located in Houston, Texas, introduced DHS CBP's seven regional labs and associated staff who attended the Workshop. Dr. Sweet asserted that customs and tariffs are the second largest revenue generator [30] for the United States. For example, in the honey market, false claims of origin cost the United States a great deal of revenue. Although most of the analyses conducted by forensics technicians are empirical, DHS would like to leverage the statistical techniques developed in other countries, in particular the United Kingdom (*e.g.*, Scotland Yard), for forensic pollen geolocation. In examining how forensic palynology could facilitate customs and crime scene investigations, Dr. Sweet commented on the need to compare samples to a reference. Moreover, he suggested that the best solution in terms of time and efficiency would be to develop a massive global pollen database. The ultimate goal would be to build a database that others could utilize for their particular applications. While a pollen database may not be relevant to every case, it could still help immensely in solving crimes by enabling law enforcement officers to make good decisions about where to focus their efforts.

Several workshop participants reinforced Dr. Sweet's statement that a massive global database of pollen grains is needed. The database should feature a structured vocabulary and over 100 images per pollen record, including many different orientations per record. A subset of workshop participants cited a lack of high-quality pollen maps and georeferencing information in many regions of the world. The consensus view was that the benefits of building a global pollen database that encompasses both morphological descriptors and geographic referencing include: 1) the potential for expanding the use of forensic pollen data in criminal cases in the United States, 2) facilitating worldwide geolocation searches, and 3) helping to link illegal drug activity to precise locations. In fact, Professor Bryant asserted that "A good database can help identify terrorists and prevent terrorism, and aid with other criminal activities (pharmaceutical, food, illegal drugs, etc.). Lots can be done, but a database is integral to getting it done."

One of the themes explored in the workshop was how automation algorithms could aid in identifying pollen grains. Several seasoned palynologists asserted vehemently that no automation algorithm is likely to replace trained experts. However, a human can only hold several hundred to a thousand images in visual memory; therefore, an automated system would be of tremendous value to palynologists and other forensic scientists in preventing fatigue-related errors while improving reproducibility of identification. In addition, an automation algorithm would greatly reduce the time needed to query a massive global database for images to aid in pollen ID.

Several questions remained unanswered at the conclusion of the workshop (see Box I-1). These questions may be addressed in future workshops.

BOX I-1 – FUTURE RESEARCH QUESTIONS

1) Assuming that one has a complete reference set of pollen grains in the regions of interest, what is the relationship between geolocation precision and pollen ID accuracy?
2) What are the top 10% or 20% indicator taxa that should be included in a reference library with which to accomplish operational forensic geolocation? What are acceptable false positive and false negative rates?
3) Palynologists who study airborne pollen grains tend to use 40X objectives, while others image with 100X objectives. Many Amazonian pollen grains are under 13 microns in diameter [31]. Image quality undoubtedly depends on the magnification of the collection optics. The coalition is concerned that the absence of universal standards for collection/extraction/purification/imaging/counting may result in detrimental impact on geolocation? Because the Coalition wishes to collaborate with various communities without imposing cumbersome requirements, the Coalition members would like to hear comments on prior efforts in standards development.
    a. Given the diversity of plant species, would a new unified standard have to be tailored to each region of interest?
    b. Can regional libraries of morphotypes be developed, *i.e.*, pollen grains that may be distinctive, but as yet unidentified, that would contribute to geolocational identification?
    c. Are there instances for which Scanning Electron Microscopy (SEM) (or other techniques) would be recommended over optical microscopy?
    d. Does the magnification of the field unit need to be on par with the magnification of the microscope used to build a reference database?
4) Pollen grains stay entrapped in clothing and may even survive washes. Without the ability to determine when a grain was first trapped in clothing, how reliable is the analysis? Are there methods to time-stamp pollen grains, or to determine freshness?
5) As USDA scientists digitize archives of pollen collections, they are considering which business model would be sustainable. Are there potential partners who might be willing to support this effort? Can costs be defrayed by charging customers for high-resolution images? How will the USDA handle the volume of customer questions?

Report Structure

Chapter 1 summarizes the presentations related to geolocation and the challenges specifically discussed during the Geolocation roundtable. Chapter 2 summarizes the workshop presentations that centered on the themes of image digitization and database infrastructure, and describes the foundational research needed for implementing a unified massive global pollen database. Chapter 3 summarizes an in-depth discussion on the role of automation algorithms for pollen identification. The report concludes with consensus recommendations, summarized in Chapter 4.

Chapter 1. Forensic Geolocation Based on Pollen Analysis

This chapter summarizes the work of Dr. Vaughn Bryant, a leading US forensic palynologist, regarding the gaps in geolocation discussed during the workshop – in particular, during the first roundtable. The chapter concludes with a strategy for geolocation based on pollen grains from the Neotropics (*i.e*., Mexico and Central and South America).

Forensic Palynology: why it works. Vaughn M. Bryant, *Texas A&M University*

Professor Vaughn Bryant is currently the leading expert on forensic palynology in the United States. He directs the Palynology Laboratory at Texas A&M University, one of the very few facilities in the United States that routinely conducts forensic palynology analyses for government agencies and law enforcement personnel.

In Professor Bryant's keynote address on the importance and uses of pollen in forensic applications, he discussed the current status of attempts to develop an automated pollen counting device. He pointed out that the value of using pollen and spores (collectively called palynomorphs) as forensic tools relies on four important aspects.

First, most flowering plants produce small amounts of sticky pollen that relies on an insect or other animal to carry the pollen to a location where fertilization can occur. Since these pollen types are not dispersed into the wind, they represent excellent clues about the location of their origin-. This makes the pollen from this group of plants ideal geolocation markers in samples. Many other types of flowering and spore-producing plants disperse vast quantities of these palynomorphs that are carried by air currents and eventually fall to the ground in a thin coating called the "pollen rain." Whereas a German study found that pollen rain on average falls within 700 meters of the source [32], a US study reported inferred dispersal distances as great as 140 km [33]. In some regions, the amount of pollen and spores dispersed is so great that exposed land and water surfaces turn yellow from the pollen rain. Although not a precise measurement of the surrounding vegetation, and thus by inference of the climate of the area, the pollen rain in each region of the world nevertheless offers a snapshot of that area and becomes a "pollen print" that can be used to assist in the geolocation and identification of the region [24]. "Pollen rain," however, is also suspect in that it is readily carried by air currents, which could transport the pollen to areas in which the particular plant species that produced it may not be present. This contamination, in turn, could lead to a false assumption that certain taxa were present in a location of interest to forensic scientists.

Second, pollen and spores are microscopic in size, invisible to the naked eye, and can become trapped on almost any type of surface. This means that at any geographic location, pollen grains or spores from plants in that region can become evidence that could enable linking (or geolocating) a suspect or object with their original location [34].

Third, there are nearly one-half million different plant species that produce either pollen grains or spores[1]. Fortunately, each of these species produces pollen or spores that are unique and can be identified by association with the parent species; however, the pollen and spores of closely related species or even related genera in one family may appear so similar that precise identification can only be achieved through detailed studies using the resolution capabilities of an SEM and/or transmission electron microscope (TEM) [23].

Fourth, most pollen and spores are highly resistant to destruction or decay. This means that pollen and spore evidence from a region or crime scene can remain intact for years, hundreds of years, or even thousands and millions of years. If items of interest or items used as evidence are handled correctly and stored safely, years or decades later the trapped pollen and spores can still be recovered and used to assist investigators in determining the original geolocation of those items, depending upon the rate of land-degradation and associated rate of change in pollen rain composition. Fortunately, species generally produce a distinct pollen rain or "pollen print" [35-39].

Professor Bryant's introduction to pollen studies came during his graduate studies as he worked on questions related to paleo-environments and archaeology. His expanded interest leading to forensic palynology began in the mid-1970s, when he was asked to assist the USDA in using pollen to identify the origin of honey samples. This work would ensure that US loan subsidies were given only for domestically produced rather than imported honey fraudulently declared as coming from domestic sources [11]. During those studies the USDA generated pollen data for most regions of the United States; the research resulted in identifying about 6% of examined samples as having originated in foreign countries and thus illegally represented as domestic honey.

Over the next decade Professor Bryant pioneered methods for melissopalynology in the United States [11]. As world honey prices rose, US honey producers no longer relied on the loan subsidy program, and thus the need for further honey analyses by the USDA subsided during the 1980s. Professor Bryant continued to examine honey samples for a wide range of beekeepers and private importers and exporters. Most recently he also was asked to examine a sample of honey produced from hives located on the grounds of the White House [40, 41]. Beginning with his early work for the USDA in the 1970s and to date with his continued melissopalynology studies, he has examined more than 2,000 samples of North American honey and a wide range of additional samples from other regions of the world, mostly to determine the geolocation of honey production and to determine the primary and secondary nectar sources in the honey [42].

Shortly after the destruction of the New York World Trade Center and Twin Towers on September 11, 2001, the US government asked Professor Bryant to lend his forensic expertise to helping catch and prevent future terrorist activities both in the United States and abroad. Since that initial contact, he has applied his forensic palynology expertise to studies of materials associated with national security, terrorist activities, the import of illegal drugs, the geolocation of illegal contraband materials, other import/export issues, and, to a limited extent, domestic criminal cases ([11]; personal communication).

[1] (http://www.currentresults.com/Environment-Facts/Plants-Animals/estimate-of-worlds-total-number-of-species.php, accessed 1/3/2013).

One recent example is a case of possible illegal smuggling of archeological artifacts into the United States from a foreign country. Pollen analysis suggested the artifact probably did not come from purported sites in Texas as the possessor claimed [11]. Another example is the evaluation of the source of an illegal ivory shipment, where Bryant and his colleagues used geolocation techniques to narrow the origin of the ivory to regions within a few African countries, which then enabled a more focused investigation. Pollen ID and quantification suggested geolocation to particular ecoregions in Africa. The pollen study, combined with known plant distributions, and then integrated with mineral/geological analysis, proved useful in determining the potential geolocation of the ivory [20]. A third example is the cold case of a teenager murdered in 1979. When the case was reopened in 2006, pollen analysis conducted by Bryant suggested that the young woman had lived in Southern California and most probably had traveled more than 2000 miles to New York before she was killed. That new information broadened the search to a teenager reported missing in 1979 on the West Coast of the United States[2].

Through work on a variety of cases, Professor Bryant has amassed extensive information about pollen distributions in North America [11, 43, 44], and developed and refined forensic palynology methods which he has described in multiple review papers [13, 23, 24, 34]. In particular, Professor Bryant explained the distinguishing physical features of pollen used for identification, the roles of pollen species' presence and abundance, pollen dispersion characteristics such as wind vs. insect dispersal, the importance of pollen size, the potential effects of pollen degradation, and how all of these factors affect forensic interpretations [24, 34]. He also elaborated on types of pollen preparation methods. His current research centers on various examination and identification techniques applied to forensic pollen interpretations. Some of these analytical methods include comparisons of identification using LM, SEM, pollen DNA analysis, stable isotope analysis, and Raman spectrometry [23, 45]. In addition, he has detailed the importance of proper pollen sample collection and the collection of control samples, and the variety of materials from which pollen can be obtained [13, 24, 34].

Professor Bryant has also been an advocate for the greater use of forensic palynology in the United States where it is very rarely used in criminal investigations, although forensic palynology receives much broader application and acceptance in other countries such as New Zealand and the United Kingdom. However, US interest in using forensic palynology has increased since the Sept 2001 terrorist attacks in New York City. Professor Bryant pointed out that forensic palynology has been successfully used in a variety of cases in other countries, how it could have been used in various well-known American criminal cases such as the OJ Simpson trial, and the current limitations and potential opportunities for the broader application of forensic palynology in the United States [11, 13].

Geolocation Using Pollen: present capabilities and potential improvements. Mark Bush, *Florida Institute of Technology*

Dr. Bush asserted that, if one thought broadly, geolocation could represent an elevation band, a habitat setting, or a location on an ecological gradient. Identification of pollen can be difficult because pollen rain does not respect boundaries. Although distinguishing Chinese pollen from Mexican pollen is relatively easy, distinguishing Mexican pollen from Guatemalan

[2] http://seattletimes.com/html/nationworld/2003288037_pollen04.html, accessed 1/3,2013.

pollen is more difficult. In southern Mexico and Guatemala, pollen data have been quantified for approximately 80 modern lakes with known climates [46]. The pollen data reflect that every species of plant has an ecological 'sweet spot' where it is most abundant. Therefore, different pollen assemblages characterize a location. Modern plant distributions provide a great deal of geolocational data because they vary predictably with elevation and other gradients. How does this translate into forensic pollen analysis?

To conduct pollen analyses, Dr. Bush prefers to use detrended correspondence analysis (DCA), an ordination technique based on modified reciprocal averaging that preserves ecological distance between samples [47-49]. A pollen dataset (*e.g*., sample site vs. pollen type abundance) typically contains many zero values given the sparse nature of pollen distribution and differences in pollination methods. Even though some statisticians see DCA as "old school," the technique performs well with the sparse (many zero values) datasets inherent in pollen analysis, whereas the abundance of zero values impedes the use of other common data analysis tools such as principal component analysis (PCA) [47, 50, 51]. Another technique, non-metric multi-dimensional scaling (NMDS, [52]) can be valuable, especially when comparing samples that have highly similar species composition [49, 53]. This technique has been successfully used with Neotropical pollen data [54, 55] and offers a viable alternative to DCA, although DCA allows more detailed estimates of ecological distance between samples than NMDS [56].

Dr. Bush presented a study that he conducted in southeastern Peru in which he collected 44 samples (Fig. 3 Left Panel). Pollen types were defined based on climate in the study area, and grouped statistically by elevation across two different years (Fig. 3 Right Panel [57]). Statistical analysis placed 6 unknown samples within ± 1^{o}C or within approximately 200 meters of the elevation relative to where the sample was actually collected.

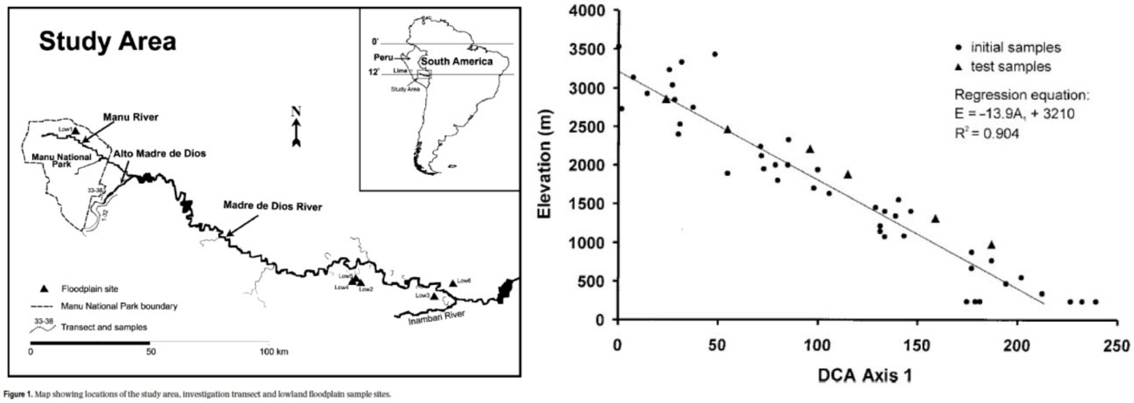


Figure 3. Southeastern Peru Study.
Left panel presents a map of the study area. The right panel is regression of the elevation and detrended correspondence analysis (DCA) axis 1 values based on the percentages of pollen data and mapping of the test sites [57].

In a study of pollen grains from Guatemala and Mexico, again using climate to define pollen characteristics, Dr. Bush applied precipitation and temperature as predictors. He found that pollen could be used to predict the actual temperature of a modern setting (± 1^{o}C), whereas precipitation was less certain (c. ± 10–20%). The latter finding is not surprising and is consistent with other means of modeling precipitation, *i.e.*, the most sensitive global circulation models run on supercomputers have about the same or worse predictive power [58].

In 2001, Dr. Bush conducted a comparative study of areas in the Amazon in which he extracted habitat-specific data from forests that were only 10 km apart [59]. The results also confirmed that the forests of Brazil and Ecuador were ecologically distinct. On the basis of his research over the years, Dr. Bush concluded that pollen is a powerful tool to use in assigning habitat attributes. Optimum spatial identification is possible when a sample can be attributed to a range of known settings, *i.e.*, the ability to choose between contrasting habitats that occur close together. One way forward would involve improving the geographic coverage of modern pollen samples and performing empirical testing of modeled pollen distributions. In landscapes that are changing because of the introduction of new species, or where deforestation is rapid, emphasis would have to be placed on repeated sampling to capture current trends. Again, this level of effort could be prioritized for locations of greatest interest.

Geospatial Analysis of Pollen for Developing Anti-Terrorist Tracking System. Peter Siska, *United States Military Academy*

Dr. Siska presented work on remote sensing of vegetation, a technique that can resolve plant communities at the family-level [60]. For most regions of the world, the pollen rain provides a fairly reliable record of the plants that produce and disperse airborne pollen within a radius of about 30 km from the sampled location. To some extent, the local pollen rain can also reflect limited information about the insect-pollinated plants in a region. For some regions of North America, existing studies of the pollen rain and the regional vegetation associated with those data demonstrate a reliable relationship between these two vegetation aspects. For other regions of North America, pollen rain studies exist, but have not yet been linked or correlated with the regional vegetation [43, 61].

One of Dr. Siska's objectives is to develop a method for using geographic information systems (GIS) to correlate existing data on pollen rain with remote sensing, based on classified vegetation patterns, especially in the forested biomes of North America. However, one problem with this type of correlation results from the assumed averaging of areas in between specific point locations where pollen rain is actually reconstructed through sampling. This problem becomes critical in areas with few point sample locations from which to retrieve the pollen data for the whole region. A better system might be to use spatial interpolation methods in between areas of point sampling where GIS can be used to predict the expected pollen rain in regions where remote sensing data is available, but no pollen rain data currently exist. Once completed, these correlations can be used to produce actual and projected pollen rain distributions for many regions of temperate North America. However, Dr. Siska recognizes that in the heavily forested regions of tropical Central and South America, where the majority vegetation is insect pollinated, these types of reconstructions cannot provide the types of pollen data that will make this a useful technique.

Understanding the relationships between pollen rain data and the vegetation biomes they represent will enable researchers and practitioners to use existing fossil pollen records to map past environmental changes in forested regions of North America and to predict future global changes of the biosphere. A secondary benefit of this research will be actual and projected pollen rain maps for North America. Those maps will provide law enforcement agencies with data on the expected pollen rain, which can act as geographic markers. Such mappings, however, will require significant parameterization of uncertainty, given the non-discrete boundaries to which the pollen rain would map as well as the very high rate of contamination of the canopy from wind and precipitation. Researchers must also consider the gradual, yet significant, changes in plant taxa as one moves through a vegetative canopy. Even if the canopy presents a unique signature in a remotely sensed image, the population of vegetation at the ground level will be heterogeneous. Pollen species will likewise have non-uniform distributions throughout a canopy.

During his talk, Dr. Siska showed how he applied a model with 36 bioclimatic variables and found five major factors that described the vegetation of the Columbia River Basin located across Washington, Idaho, and Oregon (Fig. 4). Vegetation and pollen distribution type 1 is correlated with geomorphology – defined as the shape of landscape -- which explained 46% of variance (var) in the data. Vegetation and pollen distribution of type 2 is correlated with the length of periods of sunlight and seasonality (10.2% var). Vegetation and pollen type 3 is correlated with moisture content (9% var). Vegetation and pollen type 4 is correlated with changes caused by increasing/decreasing elevation (6.2% var). Finally, vegetation and pollen type 5 is correlated with stream networks and changes caused by streams running through the basin (5.5%). Collectively, these five factors explained 77% of the variation in the data. Dr. Siska concluded by presenting a vision of a dynamic spatial pollen monitoring system, leveraging information from the Whitmore Spatial Database ([62],[3]), and integrating satellite feeds with data from airplanes and local collection sites.

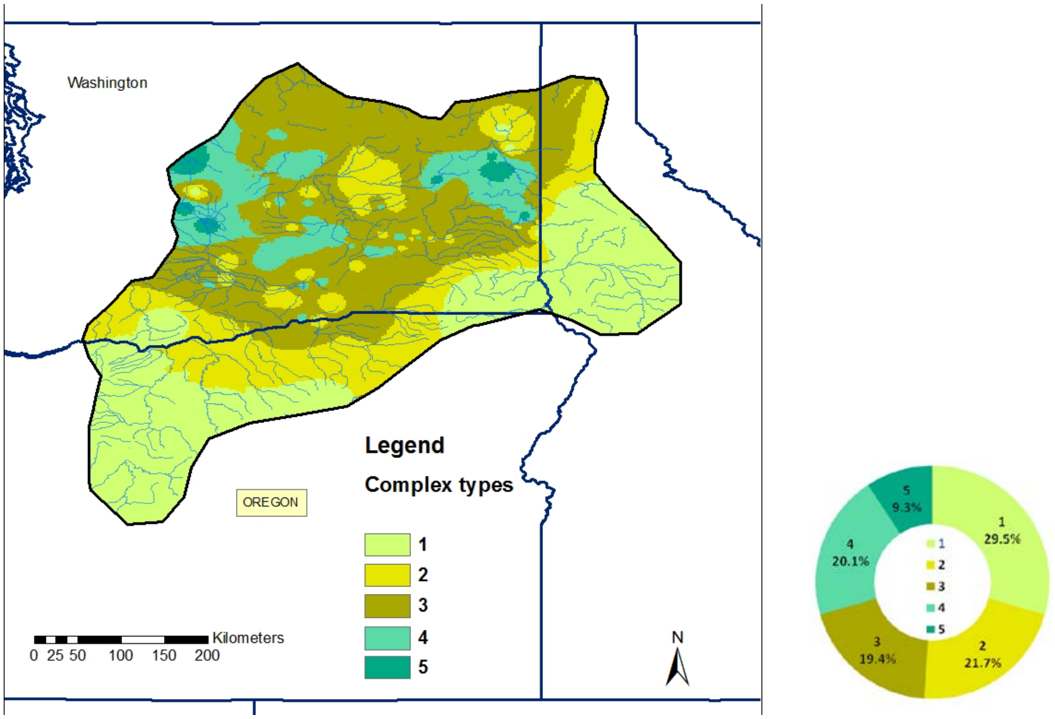


Figure 4. Results of Complex Regional Analysis of the pollen and vegetation of the Columbia Basin.
Figure credit: P. Siska and E. Mullen.

[3] http://www.geography.wisc.edu/faculty/williams/lab/Downloads.html; accessed 11/3/2012

Roundtable #1: Update on Forensic Geolocation Research.

Panelists: Vaughn Bryant – *Texas A&M University*, Mark Bush – *Florida Institute of Technology*, Peter Siska – *United States Military Academy.*
Moderator: Grace Hwang, *The MITRE Corporation*

The first roundtable focused on best achievable geolocation if researchers had perfect knowledge of pollen distribution. Dr. Hwang began the roundtable by introducing the panelists and providing an overview of pollen research initiatives at The MITRE Corporation. The MITRE team [63] is currently working on pollen identification using pollen grain images supplied by Dr. Surangi Punyasena and Dr. Bush. Geolocation accuracy is inextricably linked to the quality of pollen grain classification. Small uncertainties in pollen grain classification could possibly translate into large uncertainties in pollen grain geolocation. This error propagation of misidentified pollen specimens could impact subsequent models or analyses that directly link to the misclassified specimens at the very initial stages of a model or spatial data analysis. Researchers must explore this link and perform a careful sensitivity analysis to determine the required upper bounds on the classification error probabilities deemed acceptable for geolocation.

Dr. Hwang and David Masters have been trying to answer the questions: If we had perfect data, how useful would it be for geolocation purposes? What are the critical elements that contribute to precision, accuracy and completeness? Sometimes the presence of a single type of pollen in a sample is enough to offer precise geolocation; situations where exotic plants have been imported to regions outside their normal ecological range and used as ornamentals represent a case in point. If, for example, their pollen is found in a sample it could tie the sample to the exact location where those exotic ornamental plants were being grown [16]. In other cases, it is essential to know the pollen production levels in plants, the rate at which their pollen grains sink to the surface after being dispersed, and whether or not their pollen grains are dispersed by wind or insects. The presence in a sample of a few larch (Larix) pollen grains, which are traditionally highly underrepresented in terms of dispersion, would mean the geolocation of that object was in or very near larch trees, but the presence in a sample of a few pine pollen grains, which are dispersed by the millions, would rarely offer such a precise clue to location. Therefore, both the spectrum of pollen types and the percentages of each taxon in a sample provide the essential data needed to estimate geolocation of samples.

The exact relationship between pollen identification and geolocation is not entirely clear; current precision depends a great deal on the analyst’s knowledge of a particular region of interest as well as on the sample size, resolution of pollen types, and the degree of uncertainty associated with the pollen spectrum. For example, remotely sensed imagery is analyzed for specific spectrum characteristics indicative of specific vegetative signatures. In remote sensing, spectrum characteristics of the vegetative surface are considered unique; however, the same vegetative signature may coincide with different spectrum characteristics, such that the vegetative signature, and consequently the pollen markers, may provide false positives at a given location.

Finally, how accurate must the pollen ID step be to accomplish precise and accurate geolocation of pollen grain distribution in the field? Using the Neotropics as a starting point, and based on a review of the literature, records in the Neotoma database, and inputs from Dr. Bush, MITRE identified over 1409 pollen samples (see Fig. 5). An initial challenge for the Pollen

Coalition will lie in systematically organizing the data to compute geolocational precision and accuracy.

Countries	No. of Samples (published)	No. of Samples (unpublished)
Argentina	269	
Bolivia	20	
Brazil	60	
Brazil, Ecuador, Venezuela, Panama	21	
Chile	68	
Chile, Argentina	102	
Columbia	214	
Costa Rica	29	
Dominican Republic	47	
Ecuador, Brazil	79	
Guatemala	62	
Mexico	36	45
Mexico, Guatemala	81	
Peru	133	45
Peru, Bolivia	40	60

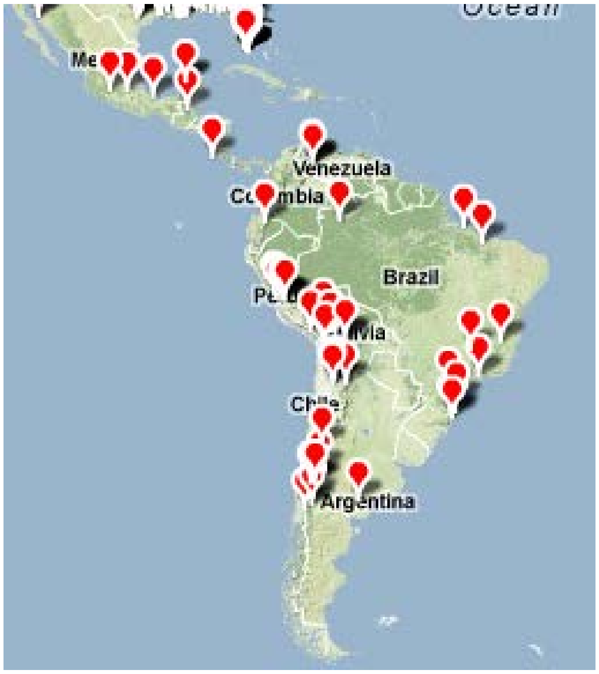


Figure credit: www.neotoma.org

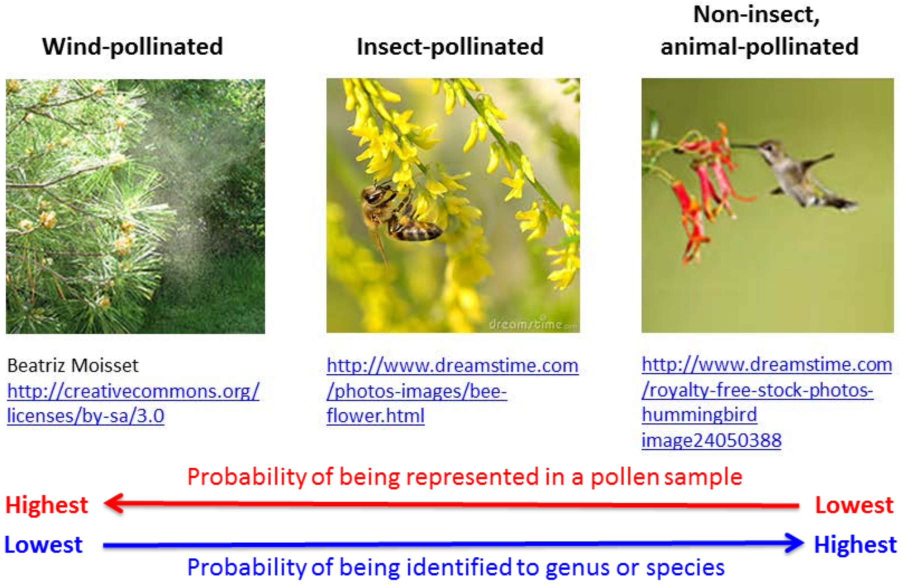
Well over 1259 published and 150 unpublished pollen samples have already been identified from the Neotropics.

Figure 5. The Neotoma database currently houses several surface pollen samples, more specifically a set of diverse pollen grains collected from a given piece of material, from the Neotropics.

Challenges

Question 1. In the context of developing a unified database for geolocation of pollen grains from the Neotropics, it is not feasible to collect from all possible plant species (80,000–90,000) in the Neotropics. MITRE's preliminary research indicates that many pollen grains (*e.g.*, grass, orchids, composites, and rushes) are not informative or indicative of a particular region. What is the minimum number of plant families required to conduct country-of-origin determination based on Neotropical flora? How would one identify and prioritize the top 10% (or 20%) of species that should be collected?

Answer: The minimum number of plant families required to determine country-of-origin depends on the end-user's needs. The good news in South America is that relatively few plants produce windblown pollen grains and many of them tend to be small, round, and rather featureless. Professor Bryant believes that insect-borne pollen types will be key to high-precision geolocation in the Neotropics, provided those pollen types are dislodged onto items of forensic importance. He also suggested that the DHS should focus on certain families such as the *Anacardiaceae, Fabaceae, Arecacea,* and *Moraceae*. If key pollen types in a few families can be identified, DHS can achieve success with geolocation efforts. Dr. Bush concurred and further explained that while only 2% of tree types in the Neotropics produce windblown pollen, pollen grains from anemophilous (wind-pollinated) plants generally make up 20–50% of most pollen counts (Fig. 6). Entomophilous (insect-pollinated) plants produce pollen grains that are diverse and highly ornamented, which offer better opportunities for identification; however, those plant types produce only moderate amounts of pollen. Zoophilous species (a category that includes plants pollinated by birds and bats, as well as insects) produce the most distinctive pollen types, but they are also the rarest in the pollen rains that fall into lake sediments and pollen traps. Importantly, though, in contrast to studies of temperate regions of the world, no study has investigated the proportions of pollen types acquired through direct contact when a person walks through different types of tropical vegetation.


Wind-pollinated
Insect-pollinated
Non-insect, animal-pollinated
dreamstime.com
Probability of being represented in a pollen sample
Highest
Lowest
Lowest
Highest
Probability of being identified to genus or species

Beatriz Moisset http://creativecommons.org/licenses/by-sa/3.0

http://www.dreamstime.com/photos-images/bee-flower.html

http://www.dreamstime.com/royalty-free-stock-photos-hummingbird image24050388

Figure 6. Pollen Strategies and Pollen Representation. Opposing relationship between the probability of being represented in a pollen sample and the probability of being identified to genus or species level.
Figure credit: M. Bush and E. Mullen.

Question 2. Will identification of pollen grains at the species level offer higher precision and accuracy relative to family or genus levels?

Answer: The majority of participants were in agreement that cost and time are the most important factors preventing the identification of pollen grain down to the species level [45]. Ideally researchers would use scanning or transmission electron microscopes SEMs/TEMs to obtain species-level information, but SEMs/TEMs are low throughput and significantly more costly than conventional LMs. In the Neotropics, many taxa may be classified at the family levels. However, genera in the tropics may be very localized [64-67]. Dr. Bush suggested a tiered strategy in which genus-level classification is conducted first, followed by classification at the species level if such biological resolution is necessary.

Dr. Punyasena, who specializes in pollen from Panama, expressed interest in translating classification into something measureable. She noted that pollen analysis is largely described solely in words, but capturing morphology digitally would represent a big step in the identification problem. In fact, Dr. Punyasena pioneered an ultra-high resolution technique that, while operating at the optical limit of diffraction, successfully distinguished pollen from several plant families at the species level [68, 69]. Recently, Dr. Punyasena generated the first demonstration of species-level classification of grasses based on ultra-high-resolution, confocal, LM [68].

Dr. Grimm disagreed with the assertion that grass genera or species are unimportant for geolocation and stated that the major issue is where one devotes resources to build the needed type of geolocation-reference pollen database. Dr. Bush commented that *Asteraceae* was a better choice than grass for determining geolocation, a suggestion with which Dr. Punyasena agreed. She explained that the identification of grass pollen grains represents a challenge in identifying texture features, which may not provide the level or reliability of identification needed for primary use in forensic geolocation.

Question 3. What types of models are needed to quantify the gain in precision and accuracy by identifying pollen grains at the species level?

Answer: Spatial analysis is difficult. Palynology should take into account the collection of samples that would benefit spatial analysis. Dr. Siska asserted that determining the number of species needed for precise geolocation is different from conducting spatial analysis. Researchers need to identify geographic points, not just the species, in order to conduct good spatial analysis. The network of collection points must also be determined with respect to the grid pattern one seeks to develop for a given collection scenario. Spatial analyses of data samples collected in clustered patterns differ from those analyses performed on data samples collected in a rigorously uniform sampling pattern. Often palynologists do not take spatial analysis into consideration when designing their surface sampling plan. Clearly a standard is needed for determining collection points in addition to methodology of collection.

Outlook

Currently about 25–40% of Neotropical pollen types encountered in samples collected from the field are unidentified or 'under-identified' (*i.e.*, marked 'unknown'), are included in morphological dump groups, or are identified only at the family level [70]. For some taxa, this level of identification may be as good as possible, *i.e.*, *Poaceae* (grasses) cannot be reliably subdivided using traditional LM. Other families, however, offer considerable scope for improved recognition, *e.g.*, *Anacardiaceae, Araliaceae, Arecaceae, Asteraceae, Euphorbiaceae, Fabaceae, Malvaceae*, *Marcgraviaceae*, and *Rubiaceae,* to name just a few. Prioritization of pollen types for forensic geolocation is essential and must be driven by end-user needs and expert palynological input. Such prioritization is critical to avoid wasting enormous amounts of time, effort, and money. The difference in utility between 10,000 thoughtfully prioritized taxa and a random subset of the total flora would be considerable, and consequences could potentially be devastating to the forensics mission if the wrong taxa were compiled.

The most comprehensive pollen atlas for a particular location in the Neotropics is the *Pollen and Spores of Barro Colorado Island* [71], comprising 1210 taxa. An online website is currently available[4]. Although this atlas offers a valuable resource for palynologists, it is an incomplete list and not sufficiently targeted for forensic geolocation. Specifically, it does not provide enough prioritization to allow one to distinguish the 10,000 species of pollen found in the study area from pollen collected in other Neotropical regions. The Neotropical Pollen Search Tool (NPST), which houses about 1300 pollen types, is the broadest and most user-friendly database available [72]. Statistical analyses of published records of pollen recovered from Neotropical sediment show that a hierarchical ranking of sexual mechanism, floral structure, and pollination strategies predict which species are most likely to be collected [73-75]. As such, the framework used to prioritize the top 10,000 species should consider pollination strategies (Fig. 6) and geographic locations with the most forensics value. To ascertain the latter, analyses of past manifest records may be a good predictor of future incursion locations. An additional promising method for predicting pollen distribution at the genus and species levels could be achieved by coupling plant distribution models (*e.g.*, [76]) with pollination strategies backed by suitable field collects.

[4] https://www.sup.org/ancillary.cgi?isbn=0804709505;gvp=1 , accessed 10/31/2012.

Chapter 2. Unified Massive Global Pollen Database

Several organizations are clearly engaged in developing digitized sets of pollen data for the world. The Pollen Coalition recognizes an opportunity to align resources and develop a standard that would allow continued sharing of resources for the generations ahead. Palynology is poised to take full advantage of informational and technological advances; the time is ripe for development of new tools that would increase the usefulness of palynology to investigative agencies.

End-user needs will drive the stored dataset and associated metadata. For example, the DHS forensics laboratory may prefer a high-throughput capability, preferably based on light microscopes commonly used by technicians. In contrast, national archives may prefer low-throughput and higher resolution imaging techniques to enable a broader set of applications. This chapter summarizes the talks of presenters who discussed their database plans, as well as presentations specific to the creation of a massive global pollen database.

DHS CBP Operational Requirements.

David Masters, *DHS S&T*;
James Sweet, *DHS/Customs & Border Protection (CBP)*

David Masters explained that DHS CBP laboratories would like to perform geolocation analysis, so DHS S&T has become responsible for developing a strategy to fund this research effort. Automated pollen analysis is not a COTS capability. Some scientists have attempted to develop such a system, but for the most part, the capability is still relatively unknown. DHS, however, cannot wait for the capability "to be at Wal-mart," but instead will fund the research necessary to arrive at a near-term solution that improves on the *status quo*. A major driver for DHS as a sponsor agency is to distinguish "the definite from the not-so-definite." Forensic pollen identification is very challenging; Mr. Masters sees the database as a critical element in achieving effective identification. The database will not be perfect; but how good is the science? DHS S&T has launched an Advanced Study to answer this question, and will canvas the scientific community on state-of-the-art methods regarding pollen analysis. Using this Advanced Study as a starting point, DHS S&T can request more work. Mr. Masters would like to have a fundamental basis for determining where funds can best be spent. Should efforts focus on building a better database, or on pollen identification? He does not want to predispose research in either direction, but rather would like to evaluate all steps in the process.

In addition to research, DHS S&T funds hardware development. But while it might be wonderful to develop a fully automated system, that may not be feasible at this time. DHS prefers to make all elements of the system available in each laboratory in order to make the effort effective and useful. There is always more to be done, but usage remains the best means of identifying unknowns and the next steps to take.

Dr. Sweet asserted that customs and tariffs are the second largest revenue generator for the United States [30]. For example, in the honey market, false claims of origin have cost a great deal of money. Most of the analyses conducted at DHS laboratories are empirical, but DHS would like to leverage the work of other countries in forensic pollen geolocation. How could forensic palynology facilitate customs and crime scene investigations? In comparing samples, when one contains an unknown, a database of other samples is essential for identification. Dr. Sweet suggested that the best solution, in terms of time and efficiency, would be to develop a database. The ultimate goal would be to build a database that others can use for their particular purposes or applications. While a pollen database may not lead to definitive results in all cases, it could still help immensely in solving crimes and aid law enforcement in making good decisions about where to focus their efforts.

Conceptual Plan for the Digital Reformatting of the Stanley D. and Gretchen D. Jones Pollen Collection.

Susan McCarthy, *US Department of Agriculture*

Dr. McCarthy presented a different perspective on pollen analysis. Her organization is tasked with preservation of the Stanley D. and Gretchen D. Jones Pollen Collection[5]. Libraries have undergone changes in the past decades, and while the way in which the USDA National Agricultural Library (NAL) accomplishes its mission has changed, the mission itself has not. The library has a new organizational structure and a new Director who is bringing fresh energy to the USDA. The library is embarking on a mass digitization of its extensive collection (over 10 million items) and plans to implement automated indexing in the near future in AGRICOLA.

Although the USDA houses several collections -- David and Susan Jarzen, 1300 tropical taxa; Meredith Hoag Lieux, 450 Southeastern US taxa; and the Areawide Pest Management Research Unit, 7500 taxa worldwide -- its first priority is the Jones Collection comprising ~2500 species from specimens collected by Stanley D. and Gretchen D. Jones (Fig. 7). The USDA's near-term objectives include digitizing prints and negatives and posting lower resolution images on the Web. In the mid-term, USDA wishes to establish a preservation/curation plan and engage key stakeholders. In the long term, the Department plans to develop discovery tools to speed identification and link the image collection to the scientific literature. The project is in its infancy, and Dr. McCarthy is looking forward to input from participants in this Workshop as well as other experts.

The immediate goal is to connect the specimen records with the literature. The physical formats in the collection include glass slides, light micrographs, scanning electron micrographs, and vouchered pollen specimens. Although the geographic focus of the collection is the State of Texas, it includes other US locations, as well as samples from Canada, Belize, and Mexico. The Jones and Meredith Hoag Lieux collections are sources of micrographs in the *Pollen of the Southeastern United States*[6]*.*

The NAL plans to scan negatives of pollen images at a high resolution to accommodate future research and aid in the development of software applications. Scanning the prints is likely to take place in the Jones Laboratory at the NAL in Beltsville, MD, because this is an active research collection, to which Dr. Gretchen Jones needs to have continuous access. A goal for this approach is to minimize handling of the collection, especially since some samples are as old as 25 years.

[5] http://pollen.usda.gov/Reference_Collection/Jones_Collection.htm, accessed 01/03/2013.
[6] http://www.scirpus.ca/cap/reviews/review13.htm, accessed 01/03/2013.

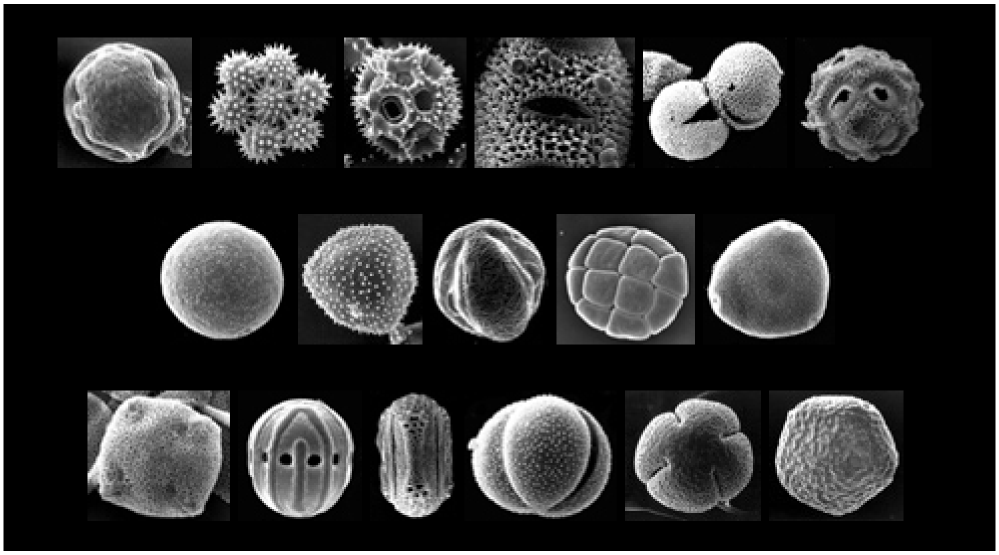


Figure 7. The Stanley D. and Gretchen D. Jones Collection houses glass slides, light micrographs, scanning electron micrographs, and vouchered pollen materials.
Figure credit: Gretchen D. Jones.

Dr. McCarthy acknowledges that NAL must address gaps, needs, and unknowns. Curating the collection over the long term creates challenges in taxonomy development, refreshing digital formats, and linking to literature. Plant taxonomy is being revolutionized by molecular tools (*e.g.*, DNA technologies) being used to sequence plant genomes. This new information sometimes leads to reclassification at the family, genus, or species level. Consequently, taxonomy is in a constant state of flux in several subfields of biology. Solutions may include linking into expert-maintained taxonomic systems such as ITIS[7] or GRIN[8]. Developing a partnership with the American Association of Stratigraphic Palynologists is also seen as advantageous, since the NAL is not likely to hire palynology experts. In considering the unknowns, USDA is contemplating which business models would be sustainable. Are there potential partners who might be willing to support this effort? Can the USDA defray costs by charging customers for high-resolution images? How will the USDA handle the anticipated volume of customer questions?

[7] http://www.itis.gov/, accessed 01/03/2013.
[8] http://www.ars-grin.gov/, accessed 01/03/2013.

USGS Palynological Research and Applications – An Overview.

Debra Willard, *US Geological Survey*

The USGS serves the nation by providing reliable scientific information to describe and understand the Earth; minimize loss of life and property from natural disasters; manage water, biological, energy, and mineral resources; and enhance and protect our quality of life. The USGS employs seven palynologists distributed across the country. Palynological research at USGS includes analysis of pollen, spores, and dinoflagellate cysts to inform and support paleoclimate research, ecosystem restoration, biostratigraphy and provenance studies, and forensic studies.

One aim of palynological research in the USGS is to document patterns of natural climate and vegetation variability to provide a baseline to examine impacts of anthropogenic activities. Comparison of fossil and modern pollen assemblages provides details of past vegetation response to climate change and rising sea levels, as well as the ability to reconstruct hydrologic variability through time. Palynomorph biostratigraphy uses pollen, spores, dinocysts, and other organic-walled microfossils preserved in sedimentary rocks to provide information on the geologic age of the unit. Pollen and spores from land plants can be used to correlate geologic units up to ~470 million years old. Palynomorphs also are used to identify sources of sediment in provenance studies for geological and forensic applications.

There are several requirements for interpreting palynological results: 1) the ability to identify pollen from different taxa based on morphological uniqueness of pollen and spores, 2) calibration datasets that relate pollen assemblages to source vegetation and, when possible, environmental parameters, 3) an understanding of depositional environment of the sample, and 4) the biology and distribution of source plants. Laboratory techniques should be tailored to individual sample types (for example, peats, lake muds, marine sediments, sedimentary rocks, fabrics) to ensure that elements of the palynoflora are preserved in slides for final analysis. USGS palynologists have compiled pollen reference slide collections that represent 1570 species from across North America, 192 from rainforests in Guatemala and Mexico, and others from Alaska and the western United States. USGS has also published a pollen atlas for Everglades plants, including illustrations and descriptions of 121 species [27]. Two additional atlases are in preparation for mid-Atlantic wetlands and the Chesapeake Bay Watershed, comprising 142 and 150 species, respectively.

Previously published regional calibration datasets have been based primarily on pollen from lake and pond surfaces and some moss polsters. Pollen typically is most abundant near source plants, although the pollen of some wind-pollinated species is dispersed over hundreds of kilometers. Pollen of insect-pollinated plants may be entirely absent from sediments. USGS has compiled local calibration datasets for some wetland sites, revealing significant changes in pollen assemblages over less than 100 meters. Such local to regional variability in assemblages must be considered when using palynomorphs as forensic tools; some pollen assemblages are better suited to identifying the source environment than specific locations (Fig. 8).

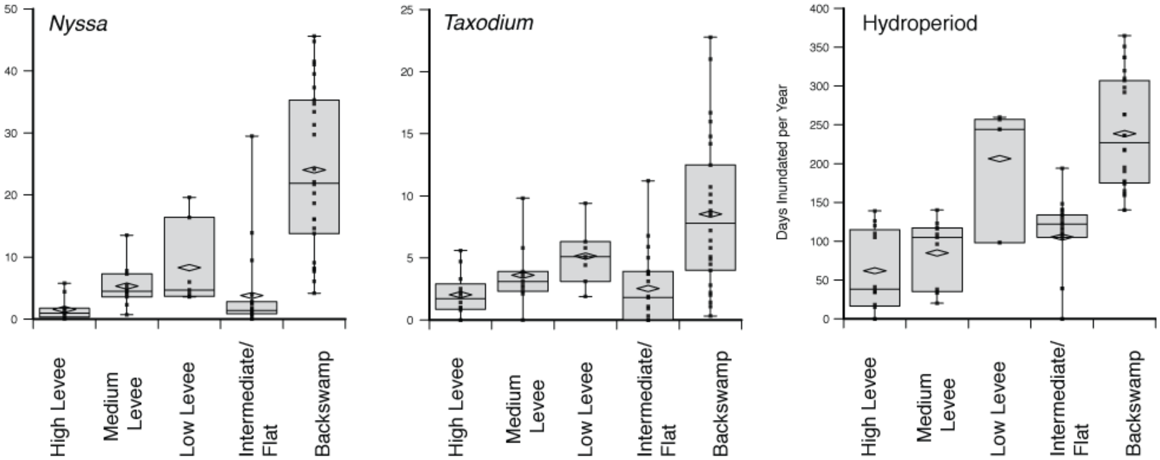


Figure 8. Tukey box and whisker plots of percent abundance of *Nyssa* and *Taxodium* pollen and annual hydroperiod in surface samples from five landforms in the lower Roanoke River basin. Points indicate individual sample abundance, and the diamond indicates the mean. The horizontal line within the box indicates median abundance for the landform type, and the box indicates values for the lower and upper quartile. Figure is from [77].

The depositional environment influences composition of palynomorph assemblages. Lake/estuary samples preserve regional vegetation signatures (and typically are the cleanest preparations). Marine records provide regional to extra-regional vegetation signatures, depending on the distance from shore; they include pollen, spores, dinoflagellate cysts, and organic remains of other marine organisms. Wetlands integrate both regional and local vegetational signatures; they also may include remains of fungi and other non-pollen palynomorphs. Soil samples provide local to regional signatures, depending on land cover type; they may be highly degraded and include non-pollen palynomorphs. Examples of strew slides of palynomorph assemblages from different depositional environments are shown in Fig. 9.

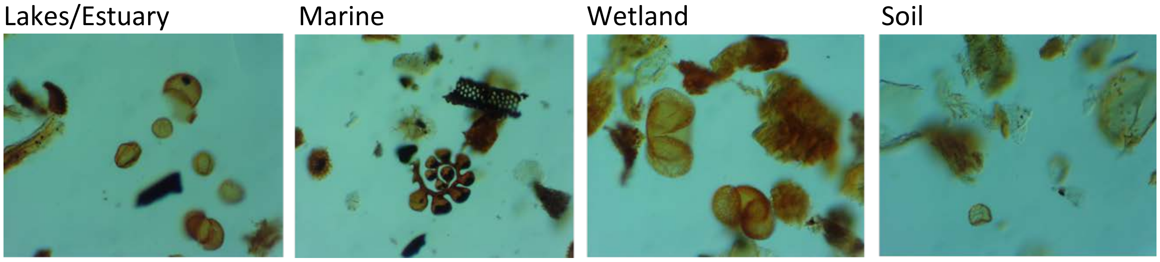


Figure 9. Example of palynomorph assemblages as a function of depositional environment: Lakes/Estuary, Marine, Wetland, Soil.
Figure credit: D. Willard.

There also are biological influences on pollen assemblage composition. The mechanism of pollination (wind, insect, or self-pollination) can make a difference in pollen taxa distribution in sediments. Pollen transported by the wind is usually abundant in sediments, whereas pollen deposited by insects or self-pollinating plants is rare. Timing of flowering also influences the pollen composition.

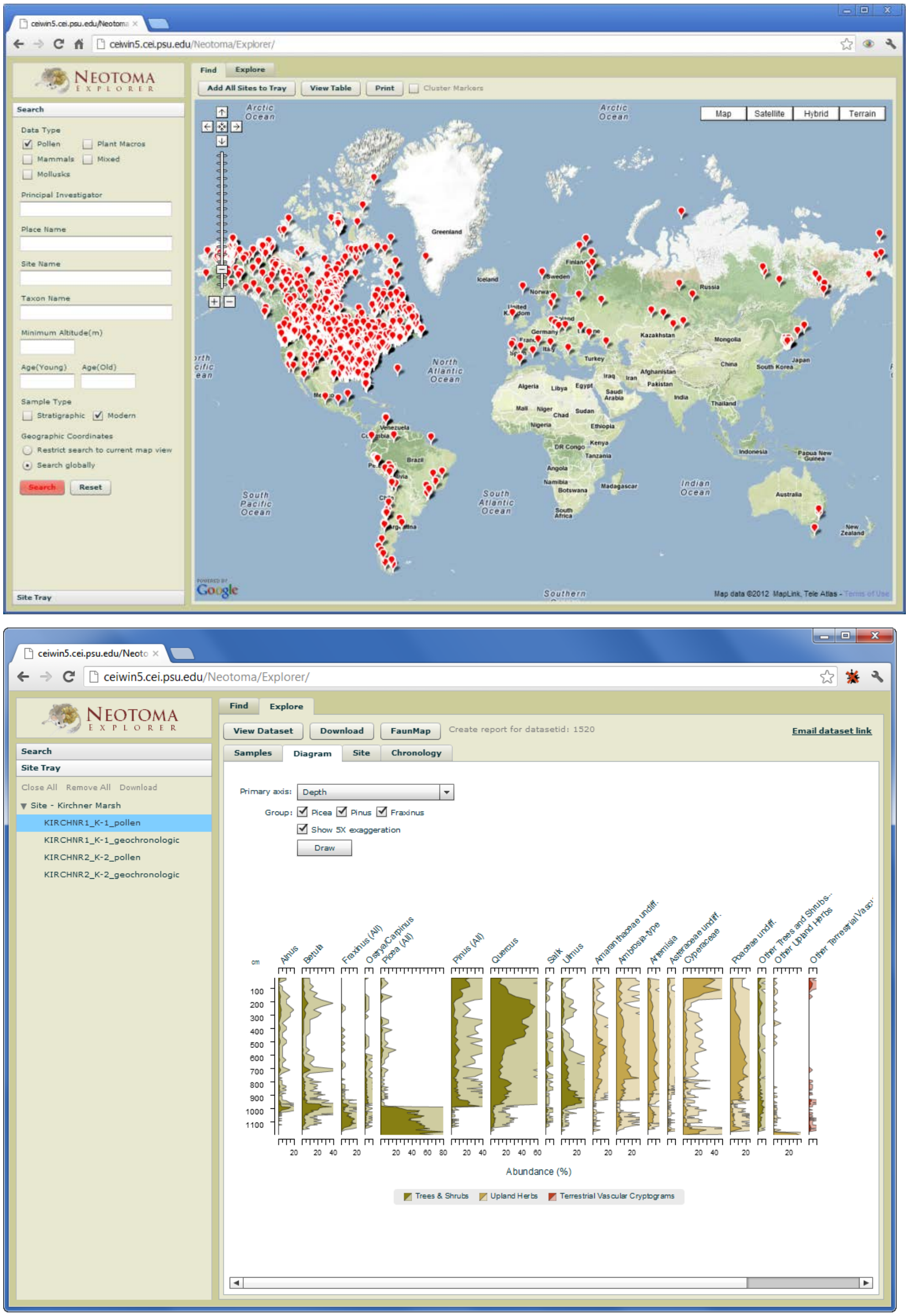
Dr. Willard emphasized the need to tailor lab preparation techniques to sample type. The goal of laboratory processing is to concentrate palynomorphs on the microscope slide by removing non-palynomorph material, thus producing a clean preparation. Pollen concentration in different types of samples may vary from nearly zero to five million grains per gram, and processing protocols and processing time vary accordingly. Choosing the appropriate sample preparation lowers the chance of damaging the pollen grains and shortens the time needed to analyze pollen assemblages.

Dr. Willard concluded that years of palynological research in academic and government institutions have provided a good overview of regional patterns of pollen distribution across much of North America. Some more localized studies also have been completed. Fairly comprehensive pollen collections exist and are scattered throughout government, academic, and industrial labs. Processing techniques must be tailored to the depositional environment of samples to maximize pollen concentration and minimize counting time. Use of multiple proxies (palynomorphs, minerals, etc.) can improve the accuracy of provenance estimates for forensic studies.

Roundtable #2: Database Schema/Ontology/Standards for Pollen ID.
Panelists: Mark Bush – *Florida Institute of Technology*,
Nicholas Orlans – *The MITRE Corporation.*
Moderator: Eric Grimm, *Illinois State Museum*

The Neotoma Paleoecology Database.
Eric Grimm, *Illinois State Museum*

The Neotoma Paleoecology Database (Fig. 10) represents the culmination of years of work with approximately 10 principal investigators and 15 international collaborators. Neotoma is an open access, multi-proxy community database for studies of the Pliocene Epoch-Quaternary period (5.3 million years ago to present). Neotoma was developed to facilitate studies of paleoenvironments, ecosystem development, and response to climate change. It enables joint analysis of multi-proxy datasets to address paleoenvironmental questions, and provides historical context for understanding biodiversity and the data for climate model validation. Neotoma also facilitates analyses of biostratigraphy and provides a long-term, low-cost archive for a wide variety of paleobiological data.



Search

Data Type
☑ Pollen □ Plant Macros
□ Mammals □ Mixed
□ Mollusks

Principal Investigator

Place Name

Site Name

Taxon Name

Minimum Altitude(m)

Age(Young) Age(Old)

Sample Type
□ Stratigraphic ☑ Modern

Geographic Coordinates
○ Restrict search to current map view
● Search globally

Search Reset

View Dataset | Download | FaunMap | Create report for datasetid: 1520 | Email dataset link

Samples | Diagram | Site | Chronology

Site Tray
Close All Remove All Download
Site - Kirchner Marsh
KIRCHNR1_K-1_pollen
KIRCHNR1_K-1_geochronologic
KIRCHNR2_K-2_pollen
KIRCHNR2_K-2_geochronologic

Primary axis: Depth
Group: ☑ Picea ☑ Pinus ☑ Fraxinus
☑ Show 5X exaggeration
Draw

Figure 10. Screen capture of the Neotoma Database. Top panel shows all location for which Neotoma has pollen sample data. Bottom panel illustrates percent abundance by family for a core sample.
Figure credit: www.neotoma.org

Neotoma was designed through partnerships between domain scientists and information technology specialists. With Neotoma, the science drives the information technology (IT). The database can accommodate most any type of fossil and modern data. Constituent database cooperatives may develop individualized websites as front ends to the database if they wish to do so. Neotoma provides the capability for 'data stewards,' who perform specific tasks with respect to the overall data management policies of this database, to remotely input and update the data in completely open access fashion.

The Neotoma project began with a 2-year grant from NSF that was renewed for 5 years[9]. At present, the Neotoma project is in its third year of the five. The database contains surface and core samples from North America and Europe, and a smaller dataset from Latin America. The entire database can be downloaded as an Access or SQL Server file. Alternatively, users can specify the type of data they wish to download, search the site by taxon or site, and download results. Metadata and chronology are included. Applications on the Neotoma website include Neotoma Explorer, which provides a Google Maps interface to the database, and TaxaMapper, which allows users to view geographic taxon distributions for specified time intervals.

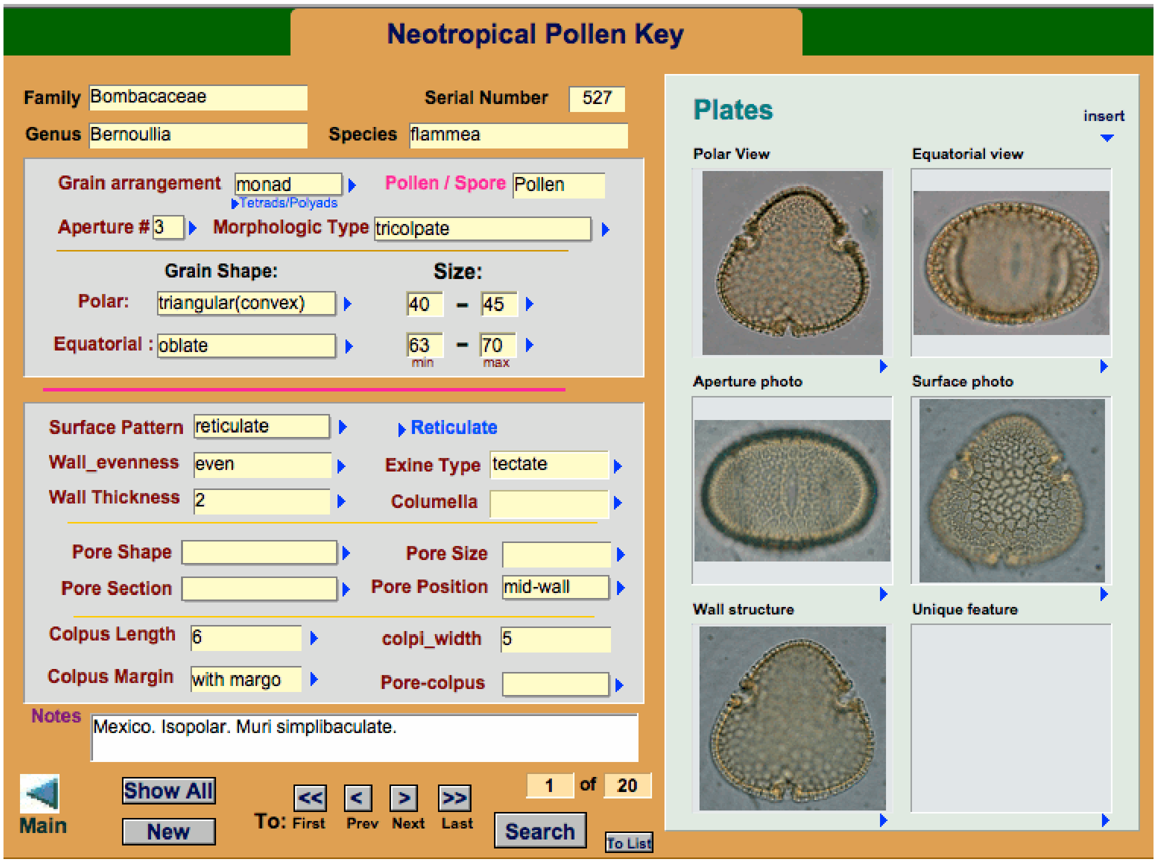
Neotropical Pollen Database and Key: the next steps.
Mark Bush, *Florida Institute of Technology*

Dr. Bush created his database in 2003 under National Science Foundation (NSF) funding. At present, it contains approximately 1300 species and serves over 750 registered users (downloadable Freeware version[10]; see Fig. 11). The search tool was developed over a period of two years and made as user friendly as possible, requiring little prior knowledge to conduct a search. Search capabilities allow highlighting as many plant families as one chooses. As mentioned in Dr. McCarthy's talk, plant taxonomy is being revolutionized by molecular tools that sometimes lead to reclassification at the family, genus, or species level. In Dr. Bush's database, older families map to newer ones. Dr. Bush did not create the database for geolocation purposes, but designed it to aid users in pollen classification. In fact, it has a very forgiving key system: users can fill in minimum fields and still obtain reasonable results with an initial best first guess.

Dr. Bush explained that training a palynologist take years. One of his goals is to develop machine-learning algorithms that would aid a palynologist in the identification process by reducing the amount of time it takes to query a database. Dr. Bush does not think it is possible to develop a fully-automated system (refer to Chapter 3 for additional details on Dr. Bush's automation algorithm research).

[9] http://www.nsf.gov/awardsearch/showAward?AWD_ID=0622289&HistoricalAwards=false. Accessed 01/03/2013.

[10] http://research.fit.edu/paleolab/downloads/pc_download.php, accessed 01/03/2013.



Neotropical Pollen Key

Family	Bombacaceae	Serial Number	527
Genus	Bernoullia	Species	flammea
Grain arrangement	monad	Pollen / Spore	Pollen
Aperture #	3	Morphologic Type	tricolpate

	Grain Shape:	Size: min	Size: max
Polar:	triangular(convex)	40	45
Equatorial :	oblate	63	70

Surface Pattern	reticulate	Reticulate	
Wall_evenness	even	Exine Type	tectate
Wall Thickness	2	Columella	
Pore Shape		Pore Size	
Pore Section		Pore Position	mid-wall
Colpus Length	6	colpi_width	5
Colpus Margin	with margo	Pore-colpus	

Notes Mexico. Isopolar. Muri simplibaculate.

1 of 20

Figure 11. Screen Capture from the Neotropical Pollen Database

Dr. Bush concluded his presentation by asking "What are the next steps?" He proposed performing a gap analysis to determine which species a database should include to enable forensic geolocation in the Neotropics. He would like to expand his database to house up to 20,000 pollen types, and to include species metadata from other databases such as TROPICOS[11] and metadata that describe the pollen collection and extraction methods. Dr. Bush plans to upgrade his database so that it will be web-based and integrated with mobile platforms. He asserted that improved image quality (including SEM images), along with machine-learning algorithms to achieve partial automation of pollen recognition and a substantial online database, are all essential elements in reducing the long learning curve currently associated with becoming an expert in pollen identification.

Furthermore, Dr. Bush warned the audience that existing databases and slides will not be sufficient for future research, in part due to the deterioration of pollen specimens on slides. The palynology community must develop a protocol to specify the way in which a reference collection should be acquired, processed, and stored. In general, pollen is best photographed in a low-viscosity mounting medium, the medium is commonly termed mountant, with a refractive index close to 1.47. Traditional mountants include 1) glycerol jelly, which is known to swell grains and therefore obscures important surface patterns – an effect most pronounced in grains > 10 years in age [78] – and 2) silicone oil, which can develop bubbles around the grain. These

[11] http://www.tropicos.org/, accessed 01/03/2013.

mountant-induced artifacts may affect the accuracy of ID for either a human analyst or a computer algorithm. Consequently, these effects should not be ignored or underestimated.

On the day after the Workshop, Dr. Bush returned to his laboratory in Florida and acquired images from the same pollen grain of *Fabaceae* in glycerol and silicone. Whereas the left panel of Fig. 12 shows an old specimen captured in glycerol, the right panel shows the same pollen grain in freshly prepared new silicone oil. The optical conditions employed to capture the images in Fig. 12 were identical. Notice the difference in the surface texture. The image captured in glycerol appears to have a swollen wall.

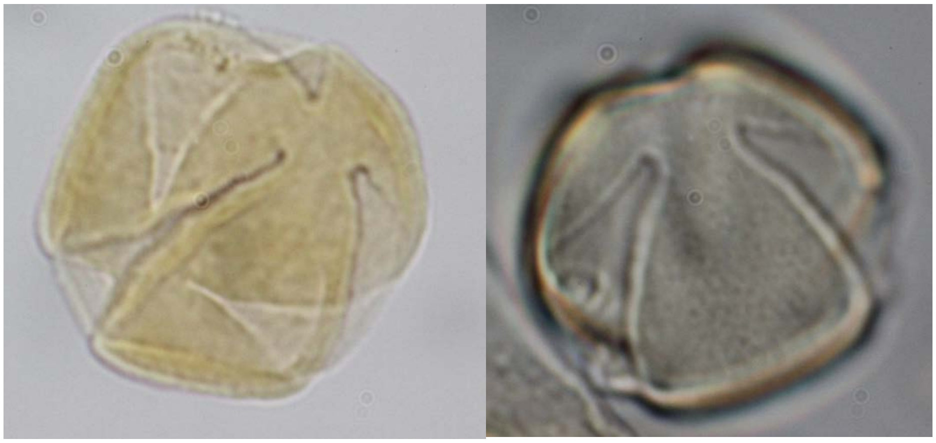


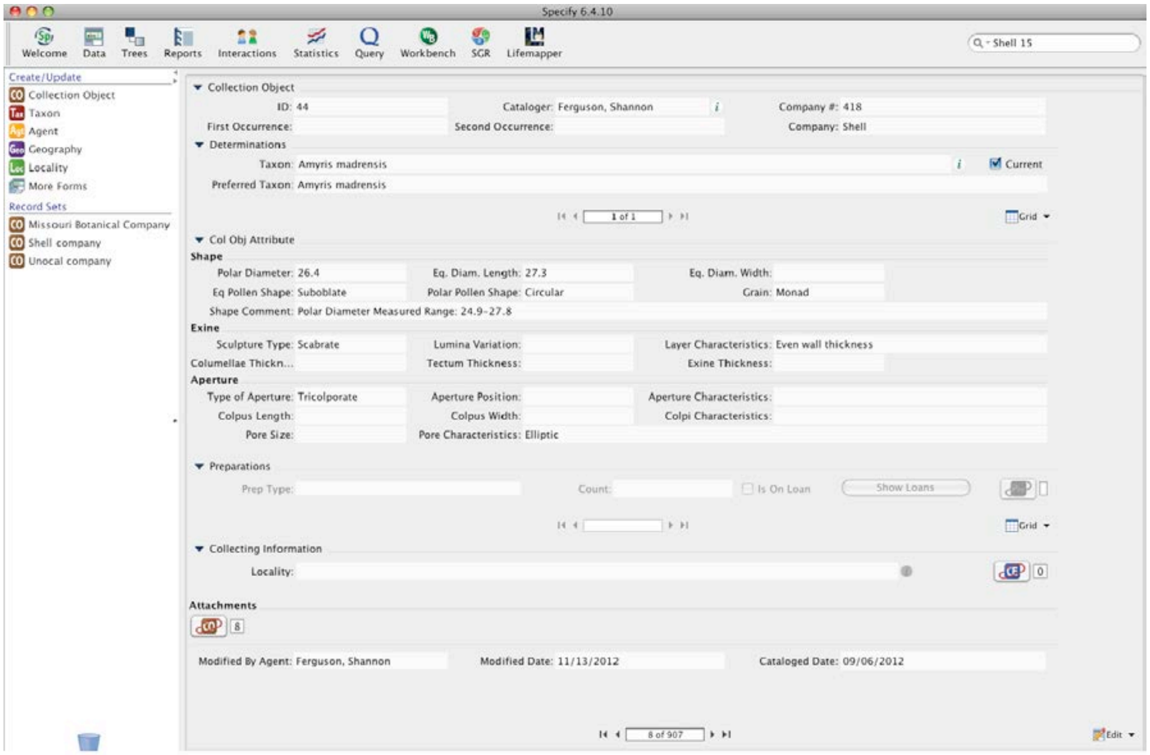
Figure 12. *Fabaceae* pollen grains in two different mounting media: glycerol (left) and silicone oil (right).
Photo credit: Mark Bush

The CENEX Pollen Database — History, Perspective and Goals.

Sophie Warny, *Louisiana State University*

The Center for Excellence in Palynology (CENEX) database comprises more than 10,000 pollen species, largely donated by the petroleum industry (*e.g.*, Shell, Exxon, and Unocal). Dr. Warny concentrates on training graduate students, and her large research group (four PhD students and three Master's students) allows her to quickly expand the CENEX collection with samples from Neogene and Quaternary projects in Antarctica, Mexico, Canada, Morocco, and Papua New Guinea, as well as a project in Texas. Her research has historically focused on paleoclimate reconstruction and biostratigraphy, but her lab is starting to develop expertise in forensic palynology. One reason she was interested in coming to this workshop was to learn about other palynological databases. She has spent the last four years planning what data processing/curation software to use, and has decided on SPECIFY, an open source software used by many museums worldwide.

Dr. Warny gave a brief overview of the CENEX database, which includes a sophisticated workbench with built-in error control (Fig. 13). The workbench allows researchers to link a record to images, export data directly into Google Earth, and even specify the uncertainty radius of a geolocation. Moreover, the workbench interacts with LifeMapper[12], which enables users to map to species in other museum collections and share data entered by others. The database supports robust querying, browsing, and reporting. At present, SPECIFY provides secure online data entry capabilities for sharing data with scientists worldwide (Fig. 14). Eventually, SPECIFY will allow users to enter data from smartphones. A SPECIFY application for the iPad is also being developed. In the future, SPECIFY will have the capability to be integrated into PostGRES SQL servers, enabling seamless integration with geospatial capabilities.



ID: 44
Cataloger: Ferguson, Shannon
Company #: 418
First Occurrence:
Second Occurrence:
Company: Shell
Taxon: Amyris madrensis
☑ Current
Preferred Taxon: Amyris madrensis
Polar Diameter: 26.4
Eq. Diam. Length: 27.3
Eq. Diam. Width:
Eq Pollen Shape: Suboblate
Polar Pollen Shape: Circular
Grain: Monad
Shape Comment: Polar Diameter Measured Range: 24.9-27.8
Sculpture Type: Scabrate
Lumina Variation:
Layer Characteristics: Even wall thickness
Columellae Thickn...
Tectum Thickness:
Exine Thickness:
Type of Aperture: Tricolporate
Aperture Position:
Aperture Characteristics:
Colpus Length:
Colpus Width:
Colpi Characteristics:
Pore Size:
Pore Characteristics: Elliptic
Prep Type:
Count:
□ Is On Loan
Locality:
Modified By Agent: Ferguson, Shannon
Modified Date: 11/13/2012
Cataloged Date: 09/06/2012

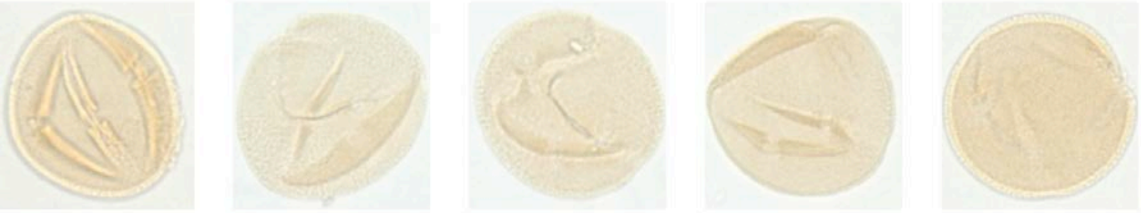
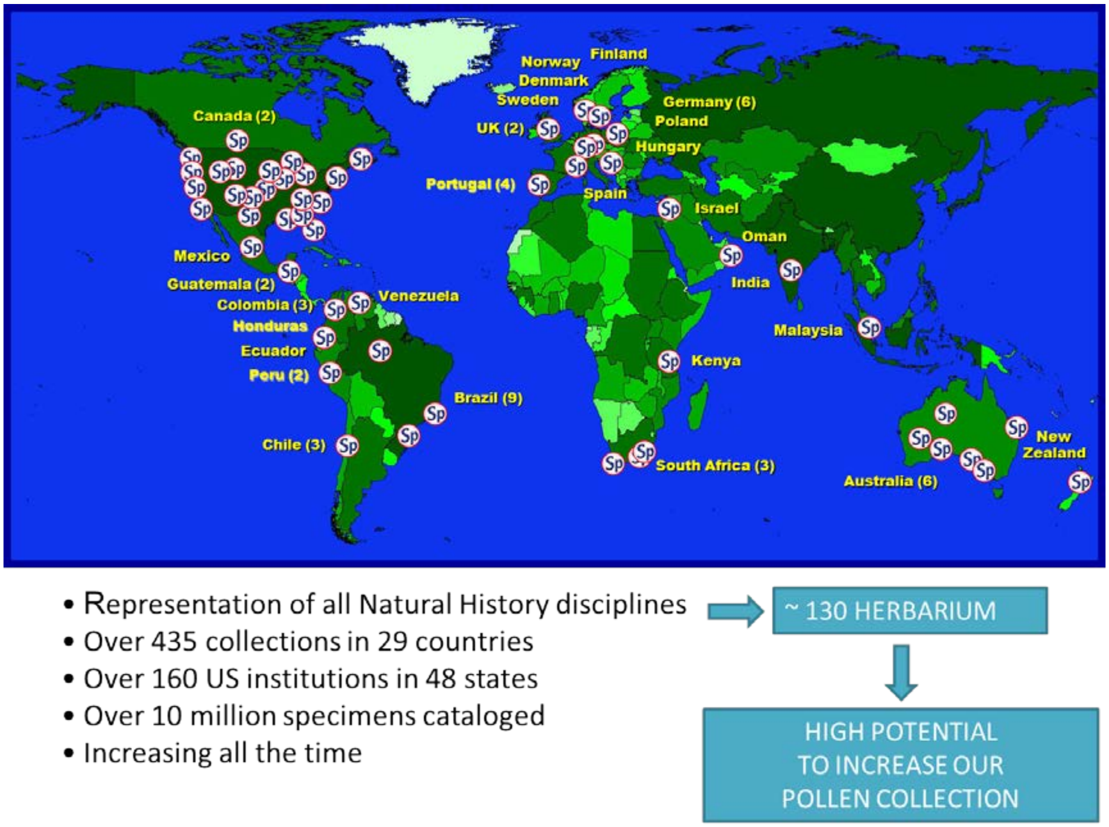


Figure 13. Screen capture of the CENEX SPECIFY workbench showing an example record. Figure credit: Sophie Warny.

[12] http://www.lifemapper.org/, accessed 01/03/2013.



- Representation of all Natural History disciplines
- Over 435 collections in 29 countries
- Over 160 US institutions in 48 states
- Over 10 million specimens cataloged
- Increasing all the time

~ 130 HERBARIUM

HIGH POTENTIAL TO INCREASE OUR POLLEN COLLECTION

Figure 14. Geographic distribution of museum collections being curated worldwide via the Specify database.
Figure credit: Andy Bentley, University of Kansas

Challenges

This session of the workshop presented three impressive databases. Dr. Grimm's Neotoma is a fantastic paleoecology database that has great referencing capability, but it contains no imagery. Dr. Bush's Neotropical Pollen database houses a rich set of images along with morphological descriptors, but lacks georeferencing information or metadata. Dr. Warny's plans for SPECIFY sounds promising. However, none of the databases presented was designed with forensic geolocation in mind. Thus, for each database to be useful for geolocation, several new data-fields would need to be added, including sample locations (latitude, longitude, elevation), medium used to mount the pollen on the slide, collection methods, and processing methods — the last of which may affect the size of the pollen grain during post-processing. For example, hydrofluoric acid (HF), used in pollen sample processing, is known to decrease the diameter of pollen by approximately 10%. In retrospect, the use of HF is troublesome owing to the forensics value that geologic and anthropogenic trace materials may possess when used in combination with pollen analysis [79]. Future standards development on sample collection should consider the extraction and preservation of a variety of trace materials, one of which is pollen.

Question 1. Would it be possible to link the three databases or is it cheaper to start all over?

Answer: Testing the feasibility of interchanges from CENEX to Dr. Bush's Neotropical database would be an interesting and potentially costly experiment. Incorrect labeling of specimens would present a large problem. The consensus was that, in academia, one cannot ordinarily publish papers based on database integration experiments. Therefore, technicians/engineers would likely have to run interchange experiments to determine if interchanges would be an efficient way to implement a massive global database of pollen for forensic geolocation.

Question 2. There has been a revolution in taxonomy in the last 20 years. How will ever-changing taxonomy be addressed in the unified database?

Answer: The International Plant Index and TROPICOS are updated regularly, but neither resource is complete or perfect for the Neotropics. In the forensics community, it is typical to create a look-up table that enables many-to-one relationships to account for the use of synonyms and taxonomic updates, both of which are imposed by the larger scientific community.

Question 3. Another issue with using pollen for forensic geolocation is the assumption that vouchered pollen reference collections and authenticated comparison reference soil samples exist for all regions of interest. How would the absence of vouchered / authenticated samples affect forensic geolocation?

Answer: Vouchered /authenticated samples imply that samples have been processed by expert collectors with local information, and that the identification of the soil sample is accurate and/or that plant specimens have been confirmed by experts from herbaria. Once authenticated, if a genus or species is revised, the herbarium record would not ordinarily be updated. Herbarium identification of vouchered specimens is a critical part of palynology, but digitizing the field would be a major undertaking.

The problem of using unvouchered pollen specimens is the potential for error. Granted, accidents might be rare, but accidental misidentifications would lead others who use the specimens to “believe” that the pollen types actually represent the taxon information that accompanies the reference specimen. Likewise, using pollen spectra from surface soil samples as a basis for making comparisons between actual locations and samples being examined will be useful only if the soil sample is a true reflection of the purported location that accompanies the sample. Surface soil samples collected incorrectly could introduce contaminants or could contain information that would skew the resulting pollen spectrum so that the results would not be a true comparison with the sampled location.

With regard to locally authenticated samples, remote sensing and predictive modeling may fulfill an important technical gap by enabling access to previously denied or inaccessible areas. Additionally, the modeling can fill gaps in the geographic coverage until samples can be obtained. The reality is that palynologists may not be able to obtain empirical genus level information for all regions of interest.

Outlook

The real problem DHS wants to solve involves more applied research, and may require databases that do not currently exist, but that could perhaps be prototyped using existing collections. The Coalition views the target solution as a “visually searchable” image database (backed by automation algorithms for rapid query results), that can take a scan of a pollen grain and output a genus (or species) name with some confidence. A second type of database would then be used to join the match or possible matches with location statistics. This two-step process might be combined into a single lookup on a single database; if so, the scientific names of the family or species might not even be essential (except to explain results and have them vetted by human domain experts). This approach might lead to the formation of a “Big Data” database over time by creating automated collection processes that simply gather samples from known (precise) locations and date/times, isolate pollen grain images, and store them.

Another potential data pipeline would incorporate data about plant species and climate distributions. As pollen tracks actual occurrence of the plant, using existing geolocational data from conventional herbaria (international and local) could yield a large amount of basic information. Conventional herbaria should contain orders of magnitude more specimens that could be gathered relative to collections of modern pollen samples. The <pollen sample slide image, date-time, long-lat, elevation> raw data might feed <taxa, family, genus, species, confidence, count, date-time, long-lat regional descriptors> data from which temporal-spatial models and visualizations could be built. Overlaid onto this schema could be ecological and precipitation information that would contribute to likelihood estimates. Additional data layers for biomass, temperature (max, min and seasonal ranges), soil type, slope, surface roughness, and cloud cover could be overlaid to refine the distributional model for each pollen type. All of these parameters are available from databases, although the scale and degree of interpolation of the data are somewhat variable.

A target schema that includes statistically sampled, recognized, and counted taxon instances at genus and species levels, and spatial-temporal information is highly desirable. The palynology community needs a ‘next generation’ database that integrates graphic and metadata components. The construction of that database must be coordinated with end-users so that images are collected and stored at the resolution required for optimal identification, and are also available on machines accessible to the ‘front-line’ forensics personnel. Using very similar image quality in the investigative unit as is used to compile the database is important to ensure that images can be matched accurately.

As the pollen samples are collected from herbaria, the metadata for all specimens held by a particular herbarium will also be acquired. In some cases, the data will already be a dataset, *i.e.*, at large herbaria, and the data can simply be downloaded. Where metadata are not available digitally, the collection sheets that accompany every herbarium specimen can be imaged, and metadata captured. In this way, for every sample collected, many more herbarium records should be acquired that will assist in defining the sample’s distribution. A caveat here is that latitude and longitude co-ordinates in herbaria are often approximate and should not be used to quantify elevation in steep regions [80]. The collection notes usually contain an estimate of collection elevation, which is probably more accurate.

The Coalition is aware of organizations such as the Global Biodiversity Infrastructure Facility[13] and realizes that species information is often available. However, a more careful analysis of the accuracy of species-level classification in the Neotropics revealed a study in which 45-60% of specimens within one family in Peru were either unidentified or misidentified [81]. Another significant problem with using existing plant data is that very few specimens exist for most species, which makes inferring a distribution difficult. In fact, a recent workshop on species distribution for Neotropical plants reported that 72% of Brazilian plant species have fewer than 10 recorded collections, and about a third have only one [82]. Another complication is that only recently collected specimens might be useful to infer current species distributions, because some species distributions are likely to have changed in the time since the herbaria specimens were collected. Moreover, morphological identification of plant species often requires flowers in addition to leaves, and therefore collection may not be possible in non-flowering times. The rarity of identification based on DNA from leaves exacerbates this situation.

Despite these caveats, the Coalition nonetheless proposes creation of a "next-generation" unified massive global database for forensic pollen analysis. The success of the database will depend on upfront discussion of the ontology and accompanying semantics. The palynology community should settle on a consensus schema for the next-generation database as soon as possible – certainly before a database is actually populated. Finally, the community needs a set of auditable, standardized approaches for sample processing/sample identification using a standard taxonomy for naming conventions of identified taxa.

[13] http://data.gbif.org, accessed 01/03/2013.

Chapter 3. Automation Algorithms for Pollen Identification

This chapter first presents a state-of-the-art survey covering a wide range of technologies used for pollen grain identification at the family and genus levels. This survey is followed by reviews of presentations at the Workshop which showed significant progress in image processing, especially in the areas of machine-learning and computer vision algorithms. Collectively, these algorithms have enabled the accurate classification of morphologically similar pollen grains at both the genus and species levels. In addition, the chapter summarizes perspectives on automation algorithms expressed during the Workshop, followed by presentations specific to the roundtable dedicated to the development of algorithms for automatic pollen identification.

There are several facets in the pollen identification process that can be automated. For example, the mechanical and chemical processing used to extract pollen from different sample types (*e.g.*, top-soil, sediment, air) can be automated in a scalable fashion for routine forensics laboratory use [83]. Recently, a sizable (1x1 m^2 base x <3 m height), end-to-end integrated automatic pollen collector, identifier, and counter that operated on grains collected for a week was demonstrated at an air monitoring station in Germany [10]. This system employed particle impaction and electrostatic deposition for collection and single-particle fluorescence imaging ($\lambda_{\text{excitation}}$=380nm, $\lambda_{\text{emission}}$ =455 nm) for identification and counting. However, it is a very big step from air sampling to soil/trace materials sampling: air samples will be almost pure pollen with just a hint of mineral and organic matter, whereas soil samples will be almost pure mineral and organic matter with just a hint of pollen. In this chapter, focus will be given to the algorithmic aspects of automating pollen ID.

Palynologists regularly spend countless hours manually classifying pollen grains using transmission light microscopes (LM). To analyze each sample, they must manually separate and count well over 100 pollen grains per sample in order to compute a histogram based on *a priori* indicator taxa. In addition to aiding the palynologist in forensics case work, automation has the potential to reduce the time required to build a database of reference pollen images necessary for calculating geolocation precision and accuracy for existing and future areas of concern. Whereas palynologists typically devote their expertise to one geographic area, automation algorithms can recognize pollen grains from numerous disparate geographic areas, thereby providing a capability that enables an increase in geographic coverage relative to the knowledge possessed by one palynologist.

Background on Automated Pollen Recognition

The concept of automating the pollen recognition process was first instantiated in 1968 by J.R. Flenley [84]. Thereafter, palynologists working with computer scientists continued to develop methods that could automate morphological identification and quantification of pollen grains [85]. The first of these automation endeavors involved template-matching holography, which resulted in 93% accuracy when the hologram of one pollen grain was compared with another hologram from the same grain [86]. This demonstrated the limited utility of holography in pollen ID.

Owing to advances in analysis of textures in digital signal processing, John R. Flenley's group analyzed texture features (*e.g.*, contrast and angular second moment) and, by employing a leave-one-out classifier scheme, achieved 94% accuracy from 192 exine images acquired from SEM [87]. Initially, Flenley chose SEM over LM on the basis of SEM's higher surface resolution and depth of focus. Furthermore, Flenley reasoned that the higher cost of SEM could be offset by the benefits of automation [87].

During the early 1990s, classification of pollen grains at the family or genus level was likely considered adequate, in part since this period predated the era of genomics in which species-level identification became commonplace. By 1996, Stillman and Flenley suggested that finer determinations – even to the species level – would be possible if one could measure sufficient parameters of size, exine structure, and sculpture. They further reasoned that without automation, these finer structures would be too laborious to record on a regular basis in spite of the added benefits in terms of ecological information [19]. However, the high cost of SEM and the laborious nature of operating SEM prevented it from becoming a routine, high-throughput, laboratory tool. As such, Flenley and team [88] eventually applied their algorithms to images acquired from LM [89] in addition to SEM [90].

For LM, Li *et al.* used a pre-classification algorithm that first grouped images based on shape, followed by a neural network classifier that operated on texture features, resulting in 100% accuracy from 13 pollen families. The average size of the training set was 54% with a maximum of 74%. Similarly, for SEM, Treloar *et al.* achieved 95% accuracy for the best set of input variables for 12 taxa.

Presently, the gold standard for pollen ID is a panel of well-trained human palynologists. Orthogonal identification techniques, such as quantitative PCR, immunoassays, and various forms of optical spectroscopy, described in [85], are seldom implemented to complement gold standard verification. Suffice it to say, the idea of distinguishing morphologically similar grains at the species level has probably been considered an impossibility since the 1990s. Most research groups that conducted image processing in furtherance of pollen ID (see Table 1) tested their pre-processing algorithms and classifiers using fresh, pristine pollen grains, easily differentiated by eye, and from known pollen types. Recent progress in 3D microscopy has enabled the capture of stacks of images acquired from confocal microscope systems that rely on fluorescence intensities [91]. Moreover, in 2012 machine-learning algorithms based on super-resolution (250 nm, λ=405 nm) confocal LM for the first time demonstrated the ability to automatically distinguish morphologically similar pollen grains from modern and fossil grains, reaching precision accuracies upwards of 93% [68]. Nonetheless, one must keep in mind that most automatic recognition techniques cannot recognize damaged pollen grains consistently – yet, on occasion, precisely these deformed/damaged grains are most critical in real-world forensics palynology.

Table 1. Summary of automatic pollen recognition approaches, identification rates, and accuracy.

Approach [identification rate in grains/min]	**Reported Accuracy (number of taxa / families / genus / species)**	**References**
2D Light Microscopy & Neural net [10]	100% (13 families)	[88, 89] Massey University, New Zealand
Scanning electron microscopy & Neural net [10]	95% (12 taxa)	[90]
3D Fluorescence Microscopy (λ=450-490 nm) & 26 SVM [1]	92% (26 taxa) 97.4% Allergy database	[92] German Weather Scvs
2D Optical Microscopy and Gaussian kernel [3000]	83% (27 species)	[93] New York University and CalTech
2D Contrast Microscopy & Neural net [10]	90% (6 species)	[94, 95] Massey University, New Zealand
2D Microscopy / Gabor Transforms & Neural Net [not provided]	97% (5 taxa)	[96] Massey University, New Zealand
FTIR micro-spectroscopy & machine-learning [40]	84% (11 taxa)	[97] Verona Italy
NIR (λ=785 nm) Vibrational Microscopy	Qualitative study of tree pollen (15 species)	[98] Humboldt-University, Germany

Non-traditional Methods for Pollen Identification

In addition to traditional optical microscopy, researchers in Greece and Croatia reported non-traditional methods based on vibrational microscopy /spectroscopy [98-100], researchers in Italy used techniques based on Fourier transform infrared (FTIR) micro-spectroscopy [97, 101], and researchers in Japan used techniques based on coherent anti-stokes Raman Scattering (CARS) microspectroscopy [102]. A research group from Korea evaluated sedimentation and gravitational field-flow fractionation to size pollen grains using three field-flow fractionation techniques [103]. Although pollen types cannot be identified solely on the basis of size, size can be a valuable guide when analyzed in combination with surface texture and morphology. In practice, very small and very large grains immediately prompt trained palynologists to think of just a few families. Some important tropical families (*e.g.*, *Cunoniaceae*, *Melastomataceae*, and *Clethraceae*), comprise mostly small grains, whereas other families may contain only a few species with atypically small grains compared with the rest of the family. Hence, when analysts see a small grain they make an immediate first association with the 'small-grain' families, which would actually be an incorrect path. In this circumstance a large database that is interrogated by an algorithm may serve palynologists especially well by preventing them from leaping to an incorrect conclusion.

Research has also investigated stable isotope ratio analysis of pollen grains for differentiating plant functional groups [104, 105]. Reports have appeared on DNA analysis for pollen identification, including the use of terminal restriction fragment length polymorphism (tRFLP) [106], polymerase chain reaction (PCR) to identify species within a family [107], and real-time PCR for quantitative assessment of pollen counts from multiple species [108, 109] . Next-generation sequencing techniques and metagenomic analyses may provide another method for pollen ID based on DNA sequence. Although workshop participants did not discuss DNA in great detail as a potentially powerful orthogonal approach for automatically achieving species-level pollen ID, they anticipate that DNA sequencing will become more broadly reported in the future as the ecology community starts to conduct metagenomics research.

The March for an Automated System.
Vaughn M. Bryant, *Texas A&M University*

Professor Bryant's keynote address emphasized that LM was a higher-throughput method for pollen ID, relative to SEM/TEM. Confocal or fluorescence microscopy, though comparable in speed with LM, requires highly trained personnel and is more expensive than standard LM. Professor Bryant also agreed with the idea of using machine-learning, particularly with artificial neural networks. However, he is concerned with overloading a system created by using gigantic datasets. Instead, he believes feature selection is critical. Key capabilities are the ability to image a slide, automatically select the pollen present on the slide (a process referred to as segmentation), disambiguate the pollen from other debris, and, finally, identify the grain relative to a known databank of images.

Professor Bryant pointed out that Gary Allen of Massey University, a student of Dr. Flenley's, developed the AutoStage system (Fig. 15) that can achieve 88–100% accuracy consistently on pristine samples composed of 3–40 pollen types. The AutoStage, which has since been rebranded to ClassiFynder™ is not able to reliably and accurately identify fossil pollen grains that are broken, degraded, partly obscured, or clumped without user intervention.[14] ClassiFynder appears to be a reasonable system for identifying and counting low-diversity assemblages such as those found in the Arctic. Because of the limitations in the recognition of damaged, broken, or slightly degraded grains – all of which would have to be quantified in order to obtain an accurate pollen record of past or present events – workshop participants did not recommend the use of ClassiFynder for high-diversity settings such as pollen grains from the Neotropics.

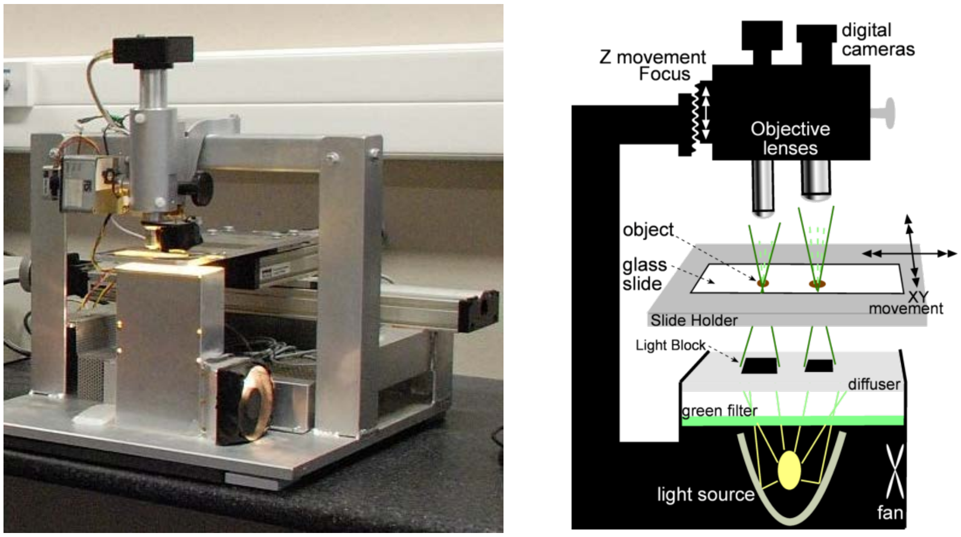


Figure 15. ClassiFynder™ Photo and Optical Schematic.
Left: Photo shows that the ClassiFynder™ scope designed by Gary Allen has two linear step motors. These motors permit precise x/y movements (resolution = 0.6 μm). Right: Schematic illustrates the optical illumination of the AutoStage system, which employs a monochromatic dark field to identify pollen grains on a slide at 4X magnification (objective on the left). Subsequently, classification of pollen grain is achieved at 20X magnification (objective on the right). Figures from [95].

Since the workshop, several members of the Pollen Coalition contacted Massey University to obtain clarification regarding 1) accuracy of classification for degraded or damaged grains, 2) unknowns, 3) flexibility in classification algorithm, and 4) throughput improvement. Collectively as of December of 2012, the participants learned that:

1) The ClassiFynder operates best on acetylized undamaged or fresh pollen grains from air, soil, or honey, it can also classify degraded or broken pollen grains provided that more than 50% of the grain remains on the slide. The system can also identify pollen clumps and debris, and the operator can check and manually correct for misclassification. This is usually a quick procedure (*e.g.*, dragging and dropping images from folder to folder using a mouse), resulting in a greatly improved level of accuracy. The system uniquely

[14] http://www.classifynder.com/, accessed 01/03/2013.

provides automated digital image capture and catalogs all target objects on the slide with focus across the full depth of field. This allows an examiner to return to the virtual slide without any remounting to further verify accuracy of pollen ID.

2) The neural network classifier embedded in ClassiFynder is a 'forced classification' system that operates on 43 shape, histogram, moment, and texture parameters. Consequently, the system does not have the capacity to classify a true unknown. True unknown grains would be dispersed among the known and correctly classified taxa. The proportional counts produced by ClassiFynder are very close to human counts and the variation in these proportional counts is less than the variation between multiple human counts of the same sample.
3) Even though the current release of the ClassiFynder system software provides only the native neural net classifier, users can apply other algorithms using Intel's Open CV library of statistical machine-learning tools[15]. Users can also export raw parameter data to statistical analysis programs such as 'R' for further analysis.
4) The typical rate of classification is up to 12 objects per minute, which equates to approximately one hour for a slide comprising 600 pollen grains. Massey University continues to work on improvements and anticipates that Classifynder's throughput could be increased by 100-fold by using programmable gate arrays. The researchers regularly look for collaborators to test and develop their system in order to cope with the wide variety of real-world demands, and would be happy to learn about the US government's interest.

Given this added knowledge, it would seem congruent with the goals of the Pollen Coalition for the US government to invite Massey University to the next Pollen Coalition workshop.

A Neotropical Pollen Database and Key: the next steps.
Mark Bush, *Florida Institute of Technology*

In addition to developing his pioneering morphology database of the Neotropics presented in Chapter 2, Dr. Bush also applied machine-learning [110, 111] to pollen grain identification. Unlike Dr. Punyasena's approach, Dr. Bush's machine-learning algorithm focused primarily on extracting information from SEM images. Although SEM is a low-throughput technique, the resolution is significantly (~40X) better than optical diffraction limit of light ($\lambda/2$, for NA=1.4 typical of modern microscopes). Furthermore, the depth-of-field in SEM depends upon the magnification and size of the final lens. For a 100 μm lens, the depth-of-field of SEM is either 20X or 67X larger than conventional LM operating at 100X or 1000X magnification, respectively[16]. Considering that many Neotropical pollen grains are under 13 microns in diameter [31] and decorated with even finer features that would otherwise be missed by conventional LM, the SEM imaging technique is clearly mandatory for conducting geolocation at the species-level. Dr. Bush's machine-learning algorithm, based on graph theory (Ribeiro and Bush unpublished data), considers spatial structure, texture, and shape to detect features that a human could not see.

[15] http://en.wikipedia.org/wiki/OpenCV, accessed 01/03/2013.

[16] http://www.emal.engin.umich.edu/courses/semlectures/focus.html, accessed 01/03/2013.

As illustrated in Fig. 16, spatial structure and appearance can be combined using probabilistic graphical models [112, 113], while the shape of salient (representative) regions can be encoded using descriptors such as shape-context [114] or other shape representations such as moments of deformation [115]. The degree of confidence of each visual cue associated with a given pollen type can be estimated/learned from a training dataset of images. The degree of descriptiveness of visual cues varies for different pollen types. Thus, the use of multiple cues and their corresponding confidence levels should help improve identification accuracies.

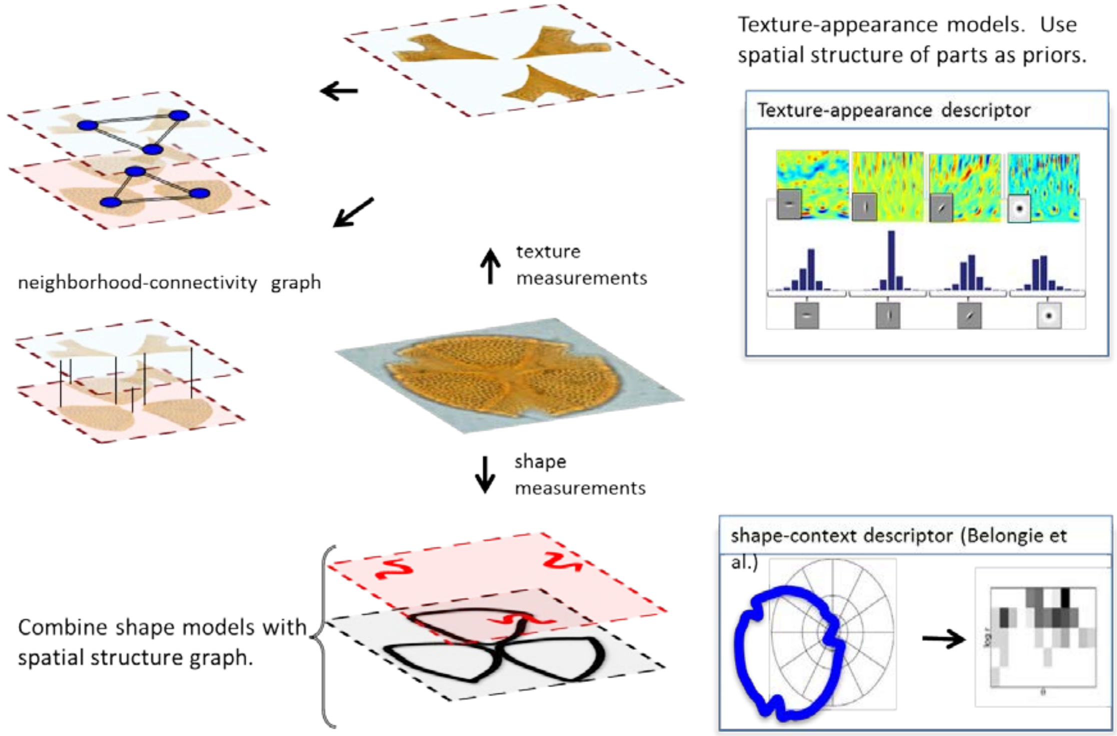

Texture-appearance models. Use spatial structure of parts as priors.
Texture-appearance descriptor
neighborhood-connectivity graph
texture measurements
shape measurements
Combine shape models with spatial structure graph.
shape-context descriptor (Belongie et al.)
log r
θ

Figure 16. Machine-Learning: an appearance model for pollen grains using a combination of multiple visual cues such as visual texture, shape, and spatial neighborhood structure.
Figure credit: Ribeiro and Bush.

Two genera, *Acaena* and *Polylepis* in the family *Rosaceae*, have proven inseparable using conventional LM. Even using SEM, prior researchers had failed to separate *Acaena* from *Polylepis*. To determine the potential to separate these species, Drs. Ribeiro and Bush used a machine-learning approach that identified ’corners’ in the surface texture of the pollen grains (Fig 17).

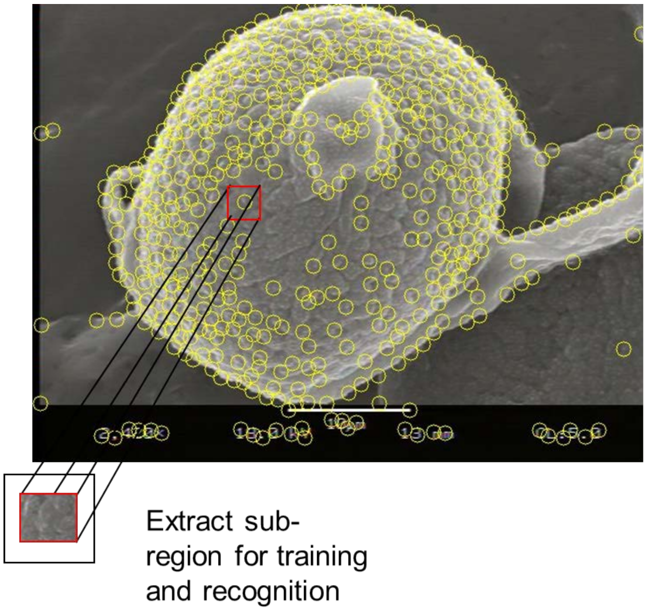
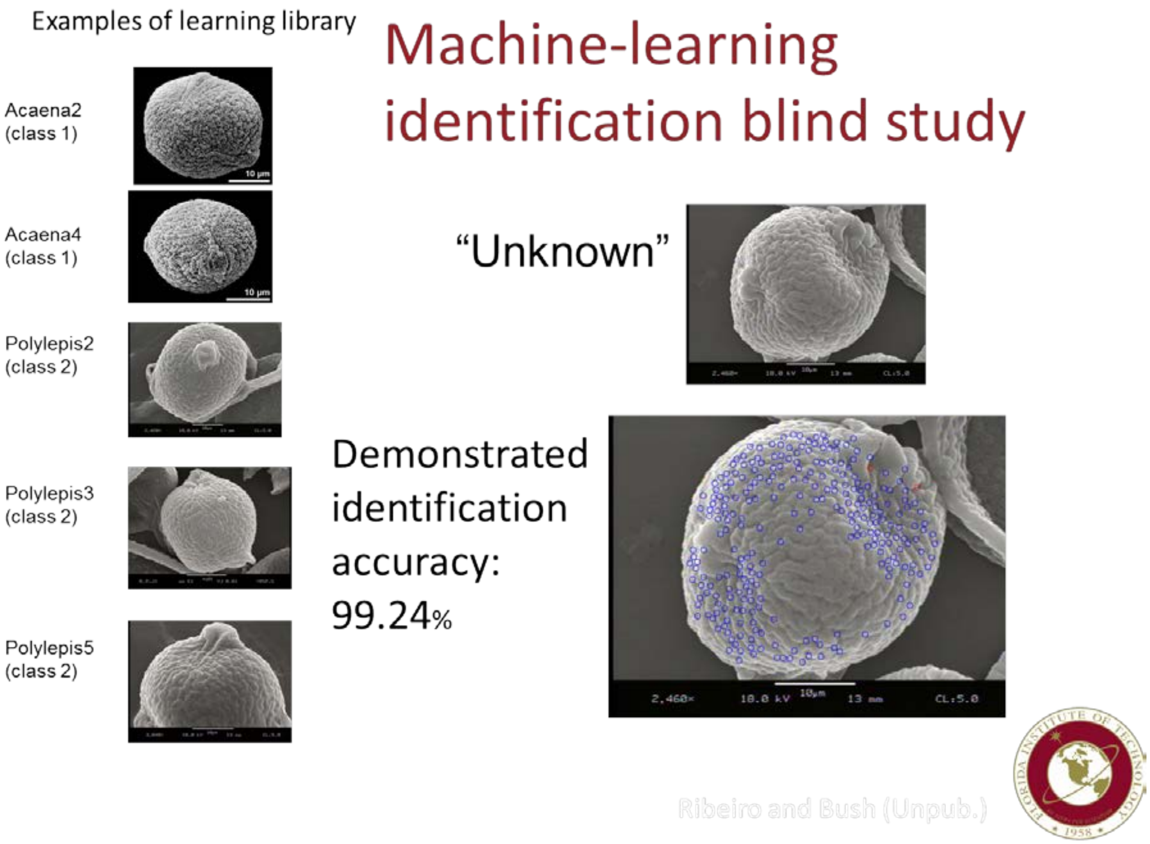


Figure 17. A SEM image of a *Polylepis* pollen grain showing ‘corners’ identified by Ribeiro and Bush’s algorithm.
Figure credit: Ribeiro and Bush.

Twenty individual grains of five species in each genus were photographed to create a library of types that was used to teach the computer the pattern of ‘corners’ found in the two genera. Drs. Bush and Ribeiro then conducted a blind study in which over 200 individual points on the surface of each grain were scored as ‘Polylepis’ or ‘Acaena’. This algorithm achieved a 99.24% classification accuracy regardless of grain distortion prior to imaging for the particular example presented in Fig. 18. Overall accuracies ranged from 86% to >99%. Where incorrect values were found they conformed to a pattern in terms of location in depressions on the surface. It is therefore hypothesized that a second algorithm could be used to screen out incorrect readings (Ribeiro and Bush unpublished). Dr. Bush concluded that machine-learning, partial automation of pollen recognition and a substantial on-line database are essential to reducing the long learning curve for becoming expert in pollen identification.



Examples of learning library

Machine-learning identification blind study

Acaena2 (class 1)

Acaena4 (class 1)

Polylepis2 (class 2)

Polylepis3 (class 2)

Polylepis5 (class 2)

"Unknown"

Demonstrated identification accuracy: 99.24%

Ribeiro and Bush (Unpub.)

10 µm
10 µm
2,460× 18.0 kV 10µm 13 mm CL:5.0
2,460× 18.0 kV 10µm 13 mm CL:5.0

Figure 18. Machine-Learning Identification Blind Study of Acaena and Polylepsis Pollen.
In a test of the ability of machine-learning to go beyond the human eye in pollen analysis, a blind study was conducted to identify *Acaena* and *Polylepis* pollen. Libraries of images (left-hand side) were used to train the computer (20 images of each of five species in each genus). An 'unknown' type was then presented (right-hand image) and the identification was based on >200 point identifications (lower right) with correct attributions showing in blue and incorrect in red. An accuracy of 99.24% was obtained in this trial.
Figure credit: Ribeiro and Bush.

Roundtable #3: Algorithm Deep-dive – Automatic Recognition and Classification of Pollen Grains.

Panelists: Kim Riley – *The MITRE Corporation*,
David Tcheng –*University of Illinois.*
Moderator: Surangi Punyasena, *University of Illinois*

Progress on Automated Pollen Identification.

Surangi Punyasena, *University of Illinois*

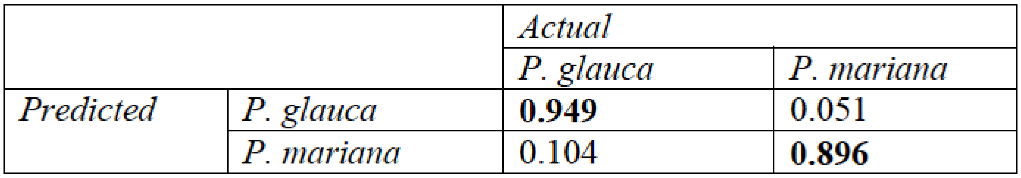
Dr. Punyasena asserted that pollen analysis could potentially address a broad range of research questions, but several limitations prevent it from being a universal solution to ecological or forensic research. First, palynology is a very specialized skill that has undergone no drastic changes in methodology or training over the last century. Second, the skill of the analyst often determines data quality, yet no formal mechanism exists for evaluating accuracy and consistency of expert identifications. Third, the precision of taxonomic classification varies; some pollen types are classified at the family level, others at the genus level. However, classification at the species level is seldom practiced.

Dr. Punyasena would like to see palynology evolve into a higher-throughput, “big-data” science. Her lab’s recent research has focused on three areas:

1) Improving the consistency and accuracy of pollen identification. She believes analyses need to be probabilistic, and identifications need to include measures of uncertainty.
2) Increasing speed and pollen sample sizes analyzed. Larger sample sizes are needed to improve pollen interpretations.
3) Producing accurate, repeatable pollen species identifications. She is a strong proponent of reintroducing species-level classification into pollen analysis because some families and genera are very widespread and would therefore not be useful for geolocation, especially in the Neotropics.

Dr. Punyasena and her colleague, David Tcheng (University of Illinois, Illinois Informatics Institute) recently published successful discriminations of two species of spruce (black and white spruce) in both modern and damaged fossil collections [68]. They recorded the human expert’s confidence in her fossil identifications. Using this criterion, Dr. Punyasena and Mr. Tcheng could subsample their training examples based on confidence thresholds, and were therefore able to test the correlation of machine and human confidence levels for each identification. In the most rigorous learning experiments with fossil samples, they used different training and testing examples from different slides. This minimized the influence of slide-level artifacts on the results. When training examples were limited to grains identified with ≥95% expert confidence, Tcheng’s algorithm achieved an average of 93% accuracy (n=263 grains; see Table 2 for details). However, when the goal was to maximize the correlation of the machine-counted ratios of white to black spruce with human expert ratios, and the system was allowed to “choose” its training and testing data sets, the machine chose to count grains with confidence values ≥70%. As a consequence, accuracy decreased to 77.5% but the total grain counts correspondingly increased to ~800 and the results captured the trend exhibited by human experts for each sample evaluated (Fig 19). These results demonstrate that, with a sufficient number of images and proper selection of discriminating features, automated systems can recognize species and reconstruct expert proportions. These results therefore support the development of automated approaches to forensic geolocation.

Table 2. Confusion matrix based on first-layer leave one out.
"Machine-learning outputs for slide-level analysis of the fossil two-class problem, the discrimination of Nelson Lake *Picea mariana* and *P. glauca*, using grains identified with high expert confidence (≥95%) (n = 264 grains). Slide-level analysis means that training and testing examples were taken from different slides. So all spruce grains on a given slide were used for training or testing, but never both. Final classification accuracy was 93.8% averaged per grain."
Caption and table from supplemental in [68].



		Actual	
		P. glauca	*P. mariana*
Predicted	*P. glauca*	**0.949**	0.051
	P. mariana	0.104	**0.896**

Dr. Punyasena believes that high-throughput imaging and data annotation are the keys to bringing "Big Data" engineering solutions to the field of palynology. Researchers need to record metadata (*e.g.*, expert's confidence level, collection/extraction/processing methods) as an integral part of creating a digital archive in order to obtain reproducibility. Equally important, she cautioned about over-tailoring a machine-learning algorithm to the training dataset, because that could lead to overly optimistic results – a situation referred to as "overfit." She presented work by David Tcheng that demonstrated that highly overfit models can achieve high ID accuracies, but when the learning system encounters a novel sample the identification accuracy is much lower. This difference is a measure of optimism. Optimism becomes a significant factor as the number of variables, classes, and degree of bias optimization increases, as is the case for highly diverse tropical samples.

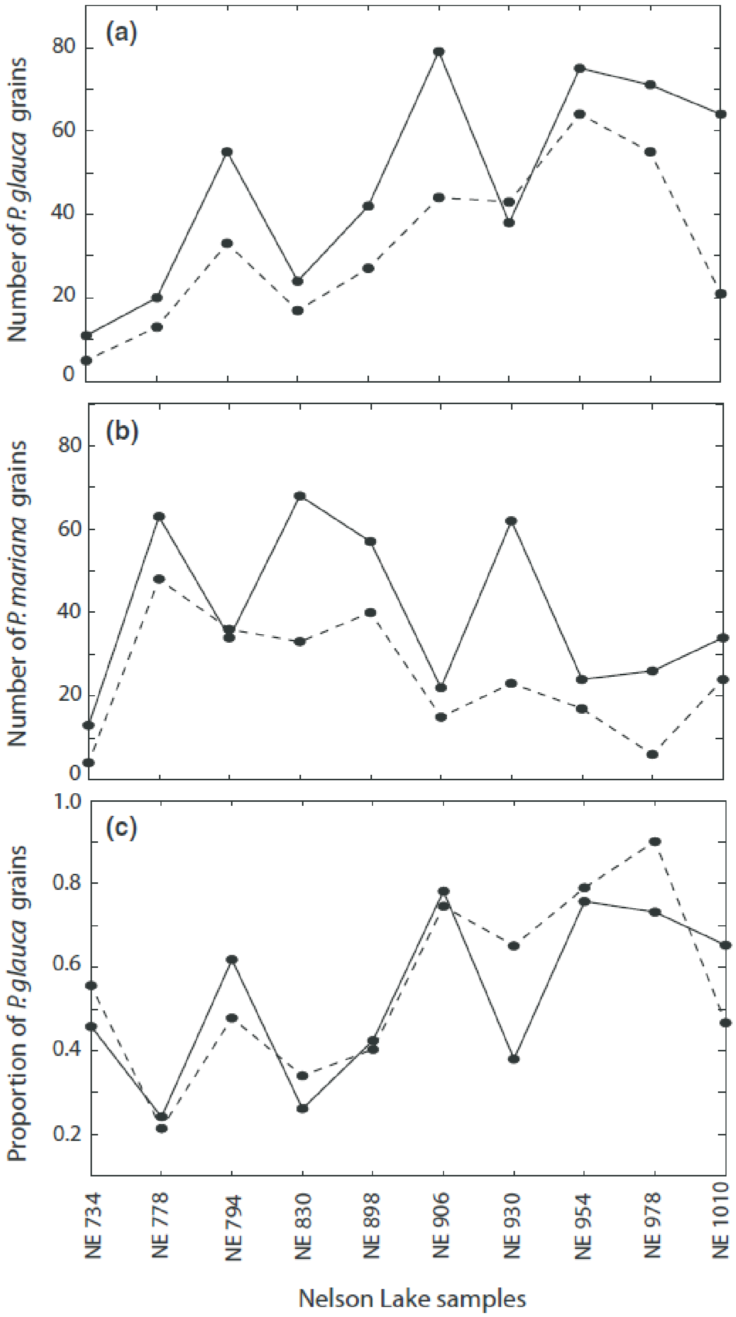


Figure 19. Comparison of machine and human counts for the fossil Lake Nelson samples.
(a, b) Number of grains counted by the classification system (solid line) and number of grains identified with > 80% confidence by the human expert (dashed line) for *Picea glauca* and *Picea mariana*, respectively. (c) Proportion of spruce grains that were identified as *P. glauca*. The machine was able to match human ratios of white to black spruce, based on identifications made with > 80% confidence, by counting as many available grains as possible.
Figure and caption from [68].

The latest version of Tcheng's pollen identification system (ARLO 2.0[17]) explicitly and iteratively quantifies this degree of overfitting by measuring internal model accuracy (as in [68]) and the difference in accuracy measured by a second tier of analysis. As shown in Fig. 20, note how the highest accuracies are achieved with a large amount of overfitting. The explicit measurement of overfitting/optimism prevents the development of learning systems that only produce high identification accuracies with the original training data. In fact, most researchers who develop machine-learning algorithms do not account for overfit, making it difficult to compare and contrast studies. Optimism must be minimized when developing a generalizable machine-learning algorithm.

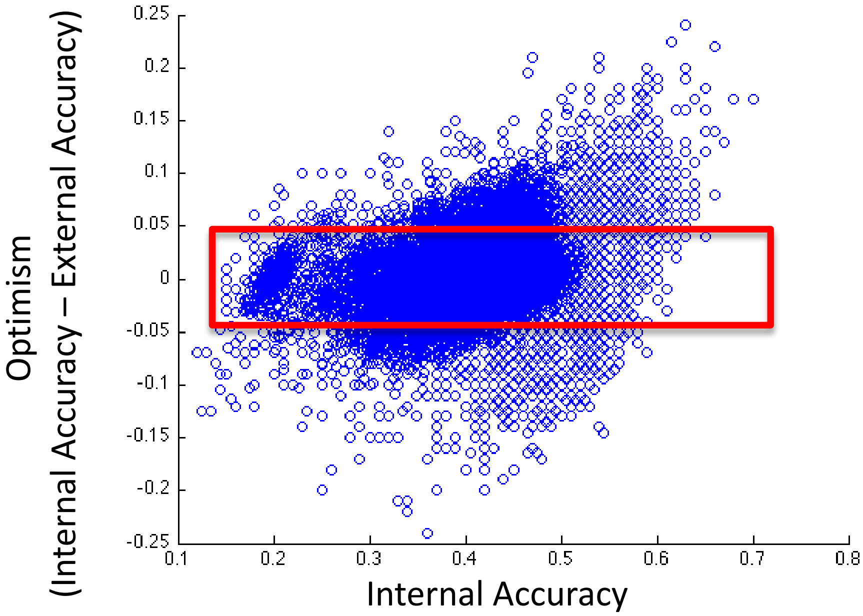


Figure 20. Measurement of optimism relative to internal accuracy for 17,813 bias optimization experiments.
Figure credit: Surangi Punyasena

Additionally, Dr. Punyasena is working with a colleague, J. Enrique Moreno, who has collected over 1000 pollen samples since 1994 from more than two locations from Panama. This includes two sets that are tower samples in which pollen traps are installed every 5m from ground to above the canopy (45m), and 20 traps within the monitored forest plot on Barro Colorado Island (BCI). Thus far, over 76 samples have been analyzed using a high-throughput scanning microscope that takes high-resolution images over the full XYZ volume of a pollen sample. Average scans take ~3 hours per slide, and 210 slides can be placed in a queue at once. The scans generate upwards of 5–6 terabytes per day. These samples form the basis of the automated pollen analyses and optimization experiments described above. Traditional pollen analysis of 110 tower BCI samples suggest that the samples capture seasonality and interannual differences in pollen signatures [116]; and that these data may be of value to the overall forensic geolocation mission.

[17] Code located at www.sourceforge.net/projects/arlo

These new automation techniques are one area of the emerging field of "bioimage informatics." Dr. Punyasena asserts that developing new algorithms to describe complex continuous characters such as texture and shape, and maximizing the information content of images captured by new microscopy methods, will lead to the next breakthroughs in automated pollen analysis. Her most recent work has taken place in collaboration with Dr. Washington Mio, an applied mathematician at Florida State University. Together, their labs have developed algorithms that describe texture on the surface of grass pollen smaller than the diffraction limit of light. The identification of grasses at the species level is often considered impossible, but preliminary classification experiments for 12 species using 240 grains from reference samples (*i.e.*, pollen from vouchered modern plant samples, where the species identity is known) were 75% accurate.

Dr. Punyasena concluded her presentation with the following key points. Quantitative approaches hold the potential to revolutionize classification consistency and accuracy and true measures of identification confidences. Classification at the species level is difficult, but may not be as intractable as previously thought, given her recent results with spruce and grasses (unpublished work). To enable a species-level database, which she contends would be especially useful for forensic geolocation, researchers need improved microscopy techniques that provide more morphological characteristics required for accurate bioimage analysis. Finally, associated development of feature algorithms and machine-learning models is needed to create robust, generalizable models.

Computer Vision Applied to Pollen Classification.
Kim Riley, *The MITRE Corporation*

Ms. Riley asserted that, although observations of highly trained palynologists are indispensable, their knowledge is often localized to one region of the world. One solution would be to leverage state-of-the-art computer vision techniques to achieve automation. This would provide faster turnaround for sponsors' geolocation needs while decreasing requirements for manpower, which translates into cost reduction. In addition, automation has the potential to aid in building a trusted database as well as in authenticating the identity of pollen grains in existing databases. MITRE's Signal Processing and Communications Analysis Department previously developed computer vision algorithms that classified butterfly images with an accuracy equivalent to that achieved by an entomologist [117]; and Ms. Riley has applied MITRE's computer vision toolbox to classify pollen at the genus level. Her analysis would not have been possible without the dataset that Dr. Punyasena kindly made available to MITRE.

The Illinois dataset, comprising 641 pollen grains with approximately 50 slices per grain (min=30, max=80), presented some initial challenges. These challenges include the orientation of pollen grains, which differed by over 20 degrees (Fig. 21, top panel), coupled with the morphological similarity across genera (Fig. 21, bottom panel). Despite these challenges, Ms. Riley verified via MITRE's computer vision toolbox that higher accuracy was obtained consistently when an entire image-stack was used relative to choosing any one slice (or subset of slices) from the image-stack.

Subsequently, Ms. Riley analyzed all image slices based on the techniques presented in Table 3 with and without a median filter (kernel size K= 6). The main advantage of a median filter is that it removes noise while preserving edges. The median filter was applied to the images prior to feature extraction using a built-in MATLAB function named medfilt2. This 2D median filter works by using a sliding window of size K-by-2, replacing the central point of the window with the median value of the neighbors. As evidenced by Table 3, in most cases, applying a median filter to the dataset improved the accuracy of recognition. In her preliminary work, the global feature Local Binary Patterns (LBP; [118]) with a grain-fold nearest neighbor classifier outperformed all other methods used to characterize the same data set (see row 2), achieving an error rate of 2.03%.

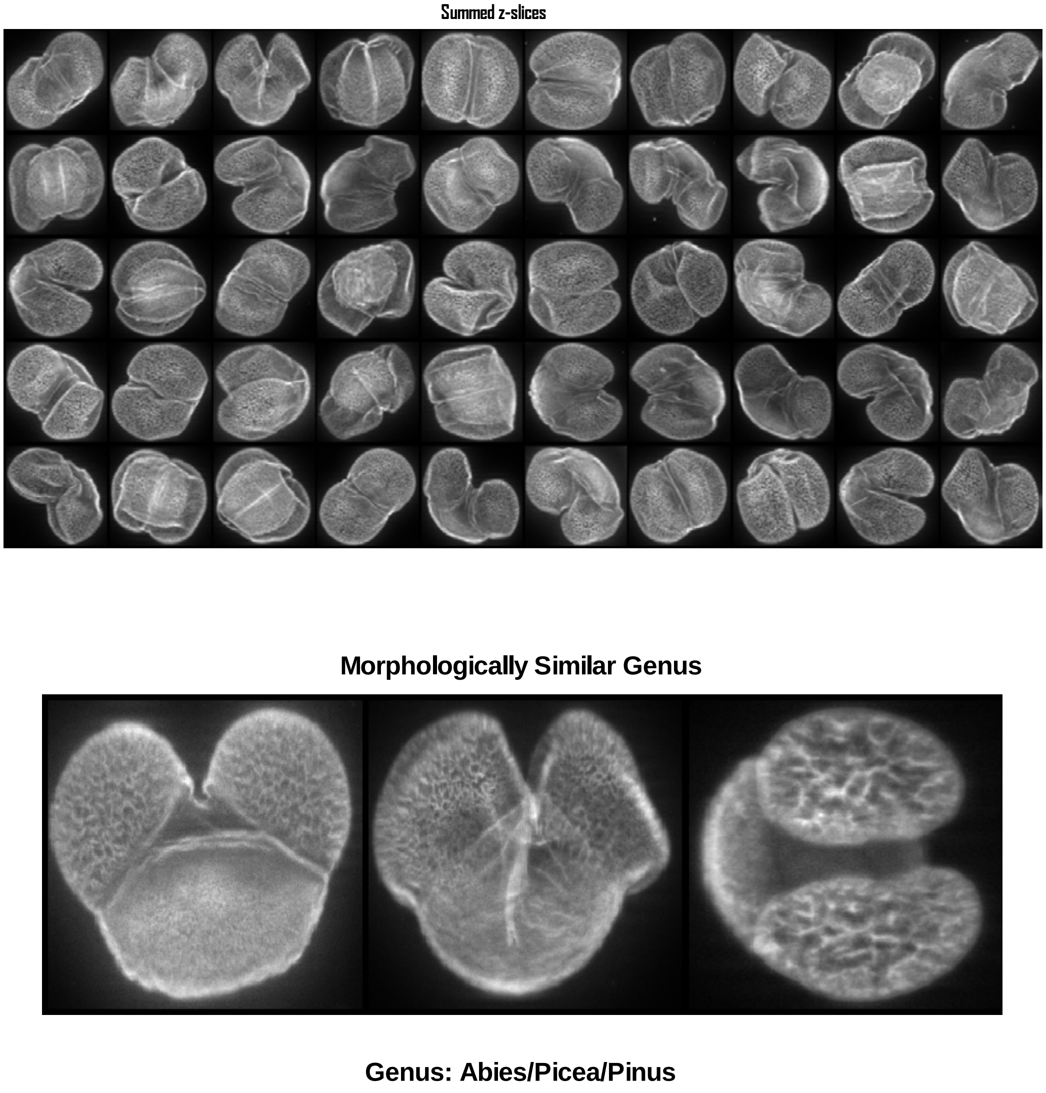
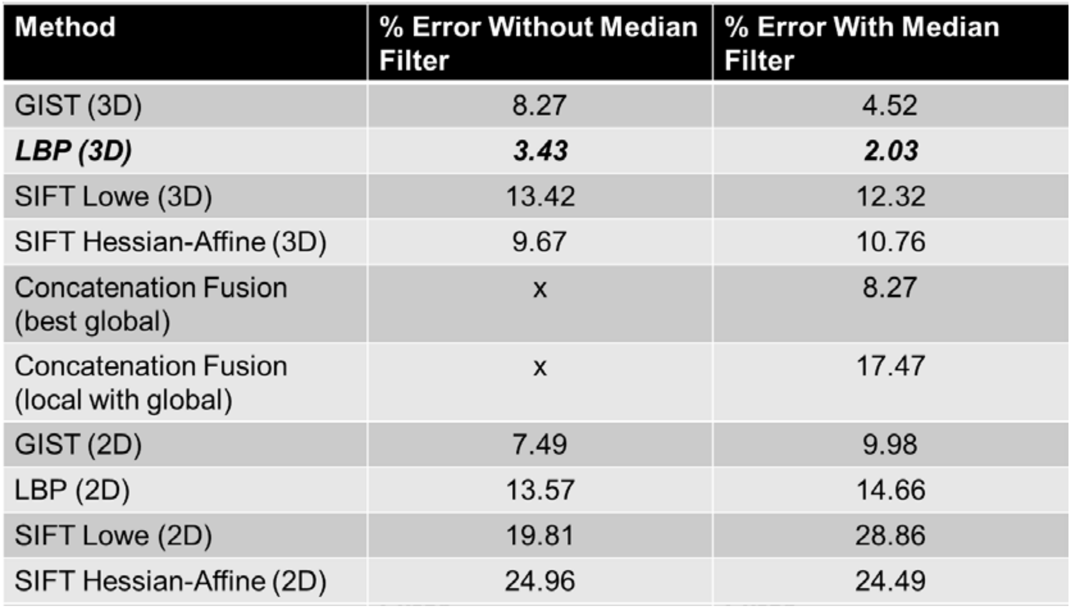


Figure 21. Morphologically Similar Pollen Grains: *Abies/Picea/Pinus.*
Top panel presents pollen grains and their different orientations based on the summation of all images in a grain-stack. Bottom panel presents morphological similarities across three genera. Figure credit: Surangi Punyasena, University of Illinois.

Table 3. Comparison of algorithm performance based on percent error and median filter usage. LBP: Local Binary Patterns; SIFT: Scale-invariant Feature Transform.



Method	% Error Without Median Filter	% Error With Median Filter
GIST (3D)	8.27	4.52
LBP (3D)	***3.43***	***2.03***
SIFT Lowe (3D)	13.42	12.32
SIFT Hessian-Affine (3D)	9.67	10.76
Concatenation Fusion (best global)	x	8.27
Concatenation Fusion (local with global)	x	17.47
GIST (2D)	7.49	9.98
LBP (2D)	13.57	14.66
SIFT Lowe (2D)	19.81	28.86
SIFT Hessian-Affine (2D)	24.96	24.49

To understand the performance trade-off of the LBP method, Ms. Riley constructed a Receiver Operator Characteristic (ROC) curve (Fig. 22), which resulted favorably in an AUC (Area Under the Curve) of 0.961. She also performed an error analysis as a function of the number of pollen grain-image stacks employed by the LBP method. It is evident from Fig. 23 that accuracy declined sharply when fewer than 100 grain-image stacks were used.

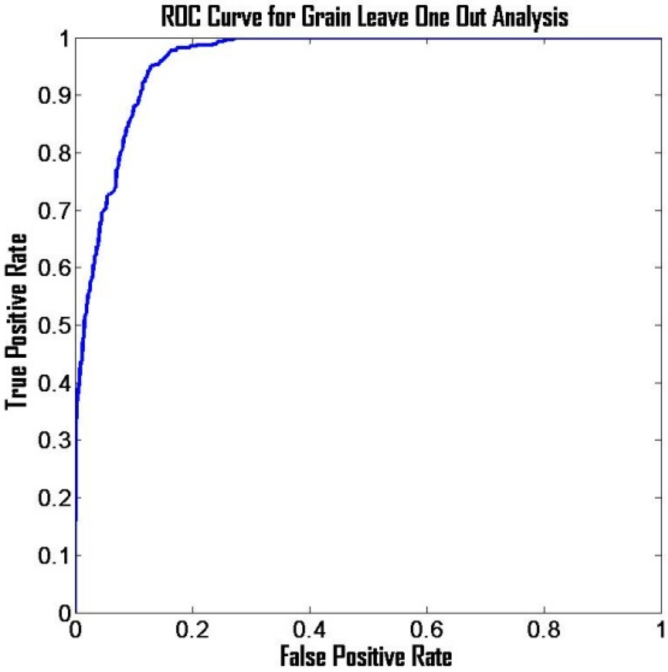


Figure 22. Receiver Operator Characteristic (ROC) curves for Grain-Level Leave-One-Out based on the Local Binary Patten 3D Method. Figure credit: Kim Riley.

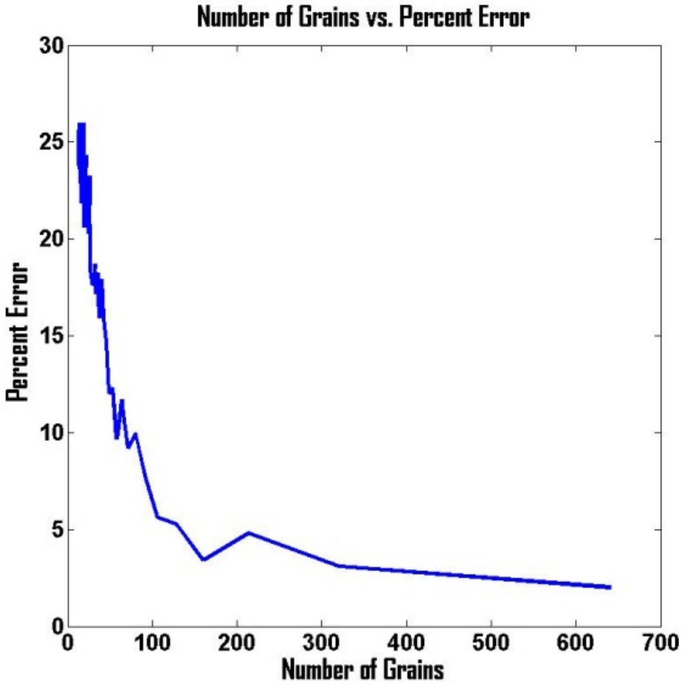


Figure 23. Percent error as the number of grain-image stack in the training dataset. Number of grains on the x-axis refers to the total number of images associated with the grain-image stack. Figure credit: Kim Riley.

Ms. Riley concluded that research must address several other concerns, including the automatic classification of damaged or occluded grains, understanding the effects on classifier accuracy imposed by variation in grain rotation, and determining whether differences in magnification between the test data and the reference data, as well as any other changes in how the data is collected, would be detrimental to any automatic classifier. In addition, understanding the scalability of this problem is another major concern. The most diverse test in this study only accounted for 5 species. Future studies should provide a best estimate model of how accuracy changes as the number of taxa increases.

It is worth noting that MITRE used the same machine-learning protocol, as described in [68], to enable direct comparison of results. Experiments run by and the University of Illinois indicated the potential for an "overfit" situation that Dr. Punyasena described earlier, due to possible slide level correlations among grain. After the workshop, MITRE ran comparable slide level and grain level leave-one-out machine-learning protocols and failed to observe any differences in error rates. Therefore, MITRE's results do not seem to reflect this sort of overfit effect, and are therefore not overly optimistic. Furthermore, Ms. Riley's classification was based on manually-segmented data, as were results published by the University of Illinois. Automatic segmentation of pollen grains is an equally important challenge in the classification problem. Finally, having a sufficiently large and representative training dataset is essential to the success of this classifier; the effects can be seen in Fig. 23.

ARLO: An Advanced Human Machine-Learning System for Automated Pollen Classification.
David Tcheng, *University of Illinois at Urbana/Champaign*

Mr. Tcheng presented his work on the application of a robust machine-learning system – Automated Recognition and Layered Optimization (ARLO) – which operates by first creating a diverse space of features and learning algorithms, then searching these combined spaces (features and algorithms) with bias optimization (Fig. 24), and finally performing layered cross-validation [119, 120]. Collectively, these steps result in avoidance of overfitting. Mr. Tcheng then introduced two types of performance metrics for ARLO: 1) Grain level, which includes classification accuracy, menu distance accuracy, and classification speed, and 2) Slide level, which includes correlation, ratio, or absolute difference between human and machine counts. He cautioned that cross-validation at the grain level is not recommended because of bias introduced by the physical slide when the two grains reside on the same slide. Furthermore, he asserted that validation at the slide level would avoid the overfit regime known to yield accurate classification.

Mr. Tcheng is a strong advocate for central processing of slides and sharing of digital imagery. The idea behind the virtual microscope is to make it faster and alleviate eye strain on the palynologist. Mr. Tcheng envisions that the future of palynology will be digital. He explained that one can easily find and train experts in pollen spotting, whereas it is much harder to train experts in pollen classification. ARLO can be taught to spot and classify pollen automatically by employing a random forest of random trees, randomly selected features, randomly selecting examples for picking feature split values, and splitting features until purity is achieved.

Mr. Tcheng also described some recent successes in scaling up ARLO's capabilities (*e.g.*, from recognizing one color band to three, from classifying 2 to 5 pollen species to 119, and to counting larger numbers of grains (13,000) and slides [76]). He described big numbers in one of the case studies presented: 30,940,000 image examples (5.47% positive), large binary trees (5.2 M leaf nodes), and forest (64 trees). The process required 91GB memory for learning, 8.5 hours for forest building, and 5GB for model storage. Mr. Tcheng noted that in the future he will investigate using more image features and learning algorithms in ARLO, accelerate GPUs (Graphic Processing Units), and explore methods of integrating the virtual microscope on the web so that experts can access the imagery collected through a user interface. Fig. 24 shows an example work-flow, and Fig. 25 shows an example image collected by the virtual microscope.

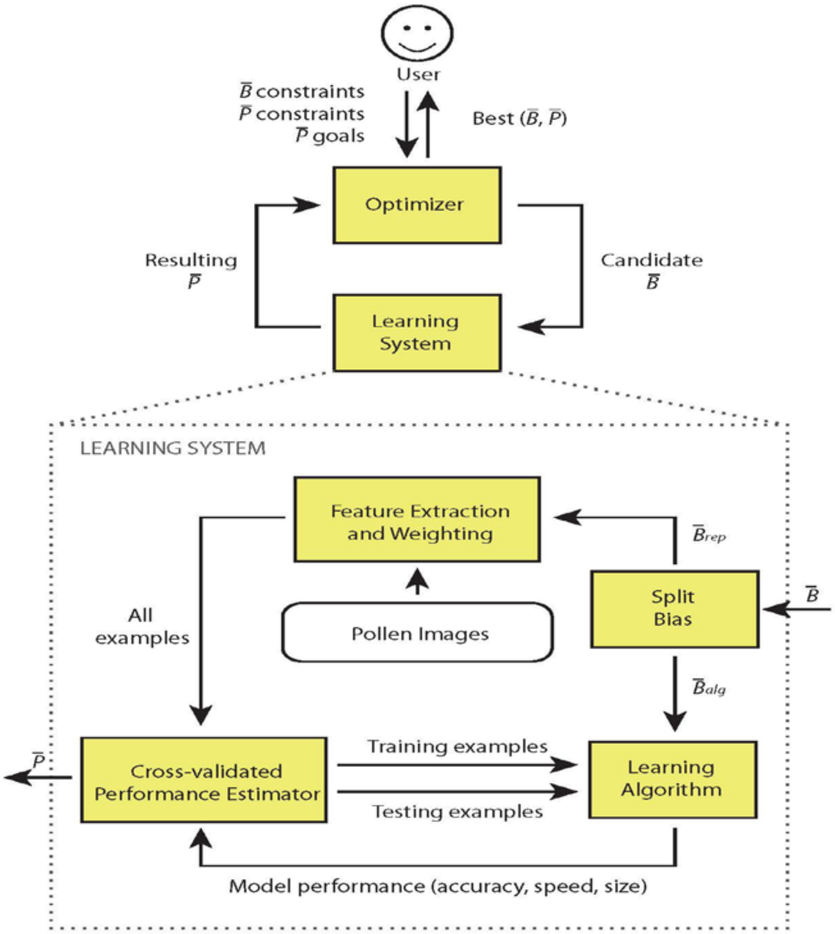


Figure 24. Schematic of Layered Classification System.
Figure is from [68].

As shown in Fig 24, there are two components to the layered classification system: an optimizer and the learning system. The optimizer searches for bias values ($\overline{\beta}$) that result in the highest performance value ($\overline{P}$); the best value discovered within an experimental run is reported as the optimized result. The learning system tries bias values ($\overline{\beta}$) from the optimizer and outputs a resulting performance values ($\overline{P}$). Within the learning system, there are two areas where bias is optimized: feature representation ($\overline{\beta}_{rep}$) and learning algorithm ($\overline{\beta}_{alg}$). Feature representation refers to how the image features are extracted from the pollen images, as described in the Materials and Methods section of [68], and how they are weighted by the learning system. Pollen examples are then divided into training and testing sets, which are used to evaluate both the learning algorithm and the feature representation with a cross-validation performance estimator. The performance estimator provides the resulting performance value, $\overline{P}$, to the optimizer.

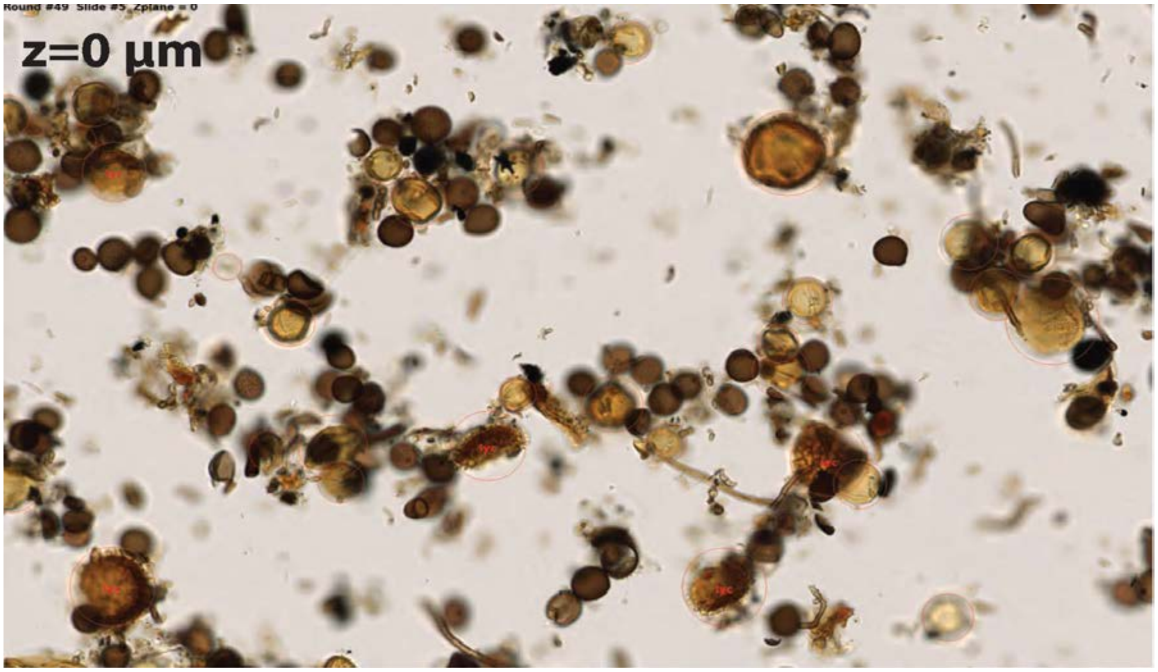


Figure 25. Mixture of Tropical Pollen Grains and Debris from Traps.
Figure credit: D. Tcheng.

During virtual microscope operation, random areas are selected for counting within a field-of-view of 2560x1600 (square pixels). The center of screen (1024x1024) is automatically designated the tagging area. The best Z-plane is selected automatically, followed by automatic pollen grain identification (see red circles in Fig. 25). Thereafter, expert classification and confirmation can be included and all confidence levels can be recorded.

Challenges

During the question-and-answer period, it became clear that sample preparation methods are likely to affect machine-learning algorithms that rely on the wavelength of light, especially preparation methods that change the spectrum of the grain (*e.g.*, fresh versus acetylized pollen grains). It was unclear how color and grain-size change(s) induced by preparation would affect pollen classification algorithms that did not rely on color or size. In addition, most of the work presented was based on undamaged, non-occluded, non-hybrid grains. One way to account for imperfect grains might be to add damaged or occluded training sets. However, adding more dimensions to the training set would exponentially scale up the memory and storage requirements, and could potentially increase the training time for all algorithms that are needed to build a library of features. Furthermore, most presentations on classification were based on manually segmented data, so it appears that a sizeable gap remains between true automation and human-assisted automation. Finally, the amount of time required to produce a super-resolution image (UI approach) is on the order of 2–3 hours at 1-micron increments, while image acquisition using LM is significantly shorter (5–20 minutes).

During the workshop, discussions on storage and memory requirements for approaches that require an entire stack of images compared to traditional single-images were inconclusive or absent. Another underexplored challenge is how algorithms would handle previously unidentified pollen grains (the unknowns).

Outlook

If the Coalition's goal centers on developing algorithms that would assist a human, the requirements would be quite different than if the goal is to develop algorithms that would replace the palynologist's role in classification. The latter would potentially require an entire stack of images with well over 100 image-stacks per pollen type (at many rotational angles) for both pristine and occluded pollen grains of interest. Typical humans have both depth perception and the ability to mentally rotate shapes. Therefore, automation algorithms designed to work collaboratively with humans (*e.g.*, in a forensics lab) would necessarily differ from those designed for an autonomous device (*e.g.*, inside an unmanned pollen monitoring station). The Coalition should continue to fine-tune the requirements for automation by holding in-depth discussions with end-users and stakeholders. The scope and complexity of the technical solution will be driven by user and sponsor needs.

In general, there are three distinct classes of algorithms that are well established for semi-automatic pollen ID using light/confocal microcopy and fluorescence: graph-theoretic, computer vision, and group-theoretic methods. The graph-theoretic methods use graphs to provide a compact representation of the data and to uncover similarities (isometries) in the graphs that can be combined to form a base subset of features for classification [121, 122]. In the computer vision methods, pollen grain images are transformed into *n*-dimensional vectors (features vectors) which are derived based on their shapes. Images are then classified using a nearest neighbor approach to evaluating pairwise distances. Local (*i.e.*, SIFT [123], ASIFT [124]) and global (*i.e.*, LBP [125], GIST [126]) feature vectors can be used while observing a range of pairwise distance types (cosine, Euclidean, cityblock, etc.) to determine the closest match. It is important to note that the success of a nearest neighbor classifier depends on a representative training dataset. In the group-theoretic methods [92, 93, 127], invariant features are formed using Gabor and/or orthogonal wavelets, spatial derivatives, centralized moments, or equivalence classes (average invariant features). For each method, a number of different classification algorithms have been proposed. These include naïve Bayes classifiers, Chi-squared classifiers, symmetric Kullback-Leiber classifiers [128], support vector machines, neural networks, k-nearest neighbor classifiers, and Gaussian mixture classifiers [129-131].

In addition, the development of a robust and scalable machine-assisted pollen ID system depends on a number of critical preprocessing steps that include image segmentation and 3D image fusion. Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as superpixels). The goal of segmentation is to simplify and/or change the representation of an image into something more meaningful and easier to analyze. Image segmentation would be used to locate individual pollen grains in an image and thus simplify the pollen ID problem.

Data generated by the University of Illinois contain pollen grain images at different focal depths. At issue is whether to optimally combine 2D images at different focal depths to preserve features representative of the original 3D image or to form features directly from the 3D image based on an extension of the SIFT algorithm. The latter approach would be more computationally intensive. Another issue is that government forensics labs may not be able to afford a $600K ultra-high resolution fluorescent microscope. Even if a researcher could extract features from the original 3D image stack, how would these features be compared with the low-resolution 2D LM images commonly used today?

Finally, decimation is a process used in signal processing where a signal is down-sampled or subsampled in order to transform it into a lower dimensional space. The ability to decimate high-resolution data to form features that could be compared with features generated from low-resolution images presents an important operational challenge. If multi-resolution feature comparisons can be accomplished, they would allow flexible sharing of datasets.

Chapter 4. Recommendations

The recommendations in this chapter are based on topics that were raised at the workshop and in several individual follow-up discussions with workshop participants. Overall, five thrust areas are warranted (Box 4-1), each of which has been assigned near-term (1 yr), mid-term (2–5 yrs), and long-term (5–10 yrs) milestones. The formation of a distributed Global Center for Forensic Pollen Informatics would be a cost-effective method for initiating the foundation work envisioned. The Pollen Coalition presents an example strategy for applying knowledge of the Neotropics as a starting point.

BOX 4-1

Thrust 1 – Design a schema and populate the "unified massive global pollen" database
Thrust 2 – Develop predictive algorithms and geolocation analytics
Thrust 3 – Set standards for pollen collection/purification/mounting/acquisition
Thrust 4 – Initiate collection and digitization for high-quality archives
Thrust 5 – Initiate digitization for a high-throughput forensics laboratory

Thrust 1 – Design Schema and Populate the "Unified Massive Global Pollen" Database

Near term: Develop a usable deliverable for the Neotropics within 12 months, the Pollen Coalition recommends creating a next-generation database and populating it with ~1,000 pollen taxa in a format that others could use to key out potential unknown pollen grains from the Neotropics. The Coalition should establish a list of taxa that are the most important for forensic geolocation. The next step would be to obtain or borrow good reference slides for those types, which could then be photographed and entered into the new pollen database. Once this list has been completed, the Coalition will need to search for individuals or existing pollen reference collections that have good examples of these specific taxa. It would be best to designate one location where these reference materials (slides) could be sent, gathered, and then photographed. Finally the images and metadata should be entered into a database. Key members of the Coalition should have the ability to add metadata to each image (*e.g.*, pollination mechanism(s), whether pollen is associated with arid or mesic regions, whether the plant grows best in sunlight or closed canopy, elevation, geographic coordinates, and scientific references to support each entry, if available).

The Coalition should structure the database in a searchable manner that would provide the easiest access for users searching for unknown pollen grains. For example, when Drs. Gretchen Jones and Vaughn Bryant produced a pollen atlas of Southeastern US flora [44], they arranged the pollen taxa by ornamentation (*e.g.*, from psilate grains through basic ornamentation types to the most complex forms). This system disregarded whether the pollen grains had one, two, or many apertures, and whether the pollen grain was big or small. In other words, all a user needed to know about an unknown pollen type was that it lacked or had spines as an ornamentation feature. With that information the user could consult the section in the atlas where all pollen grains with spines were displayed. This system would immediately enable the Pollen Coalition and the US Government to begin working on obtaining the key pollen types on a temporary loan basis from other collections.

Mid-term: As the database grows, a sophisticated query tool will be needed to aid the user in rapidly retrieving the best set of records. In order to properly evaluate the efficacy of using automated pollen ID approaches described in Chapter 3 as query tools, machine-learning and computer vision must be compared with other methods of automating pollen identification, including graph-theoretic and group-theoretic techniques. This analysis should also include a careful performance comparison of the various classification algorithms currently in use. Several well-known measures-of-effectiveness (*e.g.*, timeliness, efficiency, accuracy, cost) should be used to determine which algorithms are most suitable based on each sponsor's mission-specific needs; these missions must be defined in collaboration with Government sponsors. Some sponsors value speed while others value quality and high-resolution; therefore, a suite of algorithms is likely to be needed.

Mid-term: The Coalition should consider requesting permission from key herbaria to gather flowers and make pollen reference slides from key genera that are missing from the existing pollen reference collections. These genera can be deduced from predictive modeling based on botanical databases, or through the use of Bayesian predictors which should allow the Coalition to incorporate important expert knowledge (such as geographic coordinates, season, and co-occurring common pollen types) into the models.

Long-term: The Coalition may also wish to address discrepancies in resolution and dimensionality for pollen ID. Extracting features from SEM images and mapping them to a feature space would permit comparison with light microscopy or 2D or 3D fluorescence microscopy.

Thrust 2 – Predictive Algorithms and Geolocation Analytics

a) Weighted habitat isolation and range (WHIR) models

Near-term: Owing to the questionable quality of species-level plant classification (see Chapter 2 – Outlook) and the sparsity of existing pollen collections for many regions of interest, the Coalition will probably need WHIR models to assist in determining the best indicator taxa. In the Neotropics, geolocational data based on optical images is accurate in predicting temperature and precipitation to within *c.* ± 10% and ± 1^{o}C, respectively [46, 57]. However, local comparators are needed to achieve this accuracy. So far, moss polsters, pollen traps, soils, and sediments have all shown that local habitats can be recreated from pollen profiles [39, 46, 132-135]. Many more samples are needed to provide reliable regional profiles that could be used to infer geolocation. Targeted sampling of key habitat types and locations will be needed. In addition, sampling methodologies must match evidence streams; *i.e.*, if automobile air filters are to be used to obtain forensic evidence, then a regional library of pollen in air filter samples must be developed.

It will be impossible to survey all sites for their pollen distribution at all times; therefore, interpolation and modeling will be critical to achieving the Coalition's goal for geolocation. An approach that builds on environmental variables derived from multispectral remote sensing (*e.g.*, seasonality, biomass, seasonal greenness, ground level cloud inundation, temperature maxima, minima, and means, precipitation, elevation, slope, aspect, longitude, latitude, known vegetation distributions derived from databases such as GBIF and TROPICOS, and the actual modern pollen spectra of sampled locations) will allow the Coalition to generate WHIR models for any given species. Bayesian statistics have been used previously to create similar models based on modern vegetation data [2]. Such tools, coupled with GIS, constitute a powerful way to model landscapes. For example, Crystal McMichael, Michael Palace, and Mark Bush are currently using 30 remotely sensed variables to predict the spatial occurrence of rare soil types in the Amazon Basin (McMichael, Palace and Bush; new data). Half of the known occurrences of soil types are used to train the model, and then the other half is used to test it and assess its precision and accuracy. Modifying these approaches would make it possible to produce WHIRs for pollen distributions in Central and South America as an initial step. The first versions of these WHIRs would be crude, and would benefit from refinements such as improved weighting of parameters, ground-truthing, and collecting from additional sites. In an iterative approach, the models could be tested and improved as more modern pollen samples are added.

b) Joint Probability Estimators Based on Indicator Taxa

Near-term: To complement WHIR models, geolocation of pollen samples can also be appropriately expressed in terms of probabilities of species occurrence in a particular region (*e.g.*, country, area defined by latitude and longitude, elevation, etc.), or combinations of environmental parameters. Shortly after the 9/12/2012 workshop, Dr. Carol Christou of MITRE developed a probabilistic Bayesian-based approach using a set of species occurrence measures for each collection site. In early October 2012, Dr. Mark Bush of the Florida Institute of Technology provided MITRE with a table of 120 pollen species abundances from 364 collection sites located in the Neotropics. The table included the associated latitude, longitude, elevation, and site description. From the 120 pollen types, Dr. Bush chose 11 taxa as the best geographic indicators based on his knowledge of flora in the Neotropics.

To estimate geolocation probabilities, discrete probability distributions for both single and multiple species may be computed directly from the data tables by first normalizing across species (rows) for each site (column), and then plotting these abundances versus latitude and longitude (Fig. 26). The abundance values may also be color-coded by elevation (not shown). However, due to data sparseness, the resulting probabilities could not give a robust representation of quantitative occurrence or a reliable estimate of geolocation precision. For this reason, MITRE derived continuous conditional probability distributions for each species, as well as joint conditional distributions for pairs of species (Fig. 27). To extend discrete values to continuous ones, MITRE applied smoothing kernels in such a way that probability densities were normalized to unity across latitude and longitude, and variances associated with local peaks reflected the data as closely as possible. Although several kernels may be implemented (*e.g.*, uniform, triangular, Epanechnikov, quartic, triweight, tricube, Gaussian, and cosine), the researchers selected Gaussian kernel, based on Dr. Bush's assurance that this would be the choice of most ecologists. Subsequently, MITRE derived methods for calculating partial probabilities under local peaks, as illustrated in Fig. 27. Note, however, that some taxa do not show any correlations.

In general, the choice of species combinations for joint probability computation should be based on correlation magnitude. Species that are most highly correlated with respect to site will give the most useful information for geolocation. Toward this end, MITRE computed a species correlation matrix and chose species pairs based on the magnitudes of correlation coefficients across each row. Site similarities and clustering were derived through statistical techniques used extensively by the ecology community – again based on Dr. Bush's guidance. Examples are site clusterings based on dissimilarity measures, *e.g.*, chord metric and Ward clustering. The resulting "heatmaps" (site correlation matrix ordered by cluster, not shown) and silhouette plots (vertical plots of number of sites per cluster, not shown) compared favorably to those provided by Dr. Bush and his team. Furthermore, joint probability estimators based on indicator taxa for a larger dataset would represent a promising way forward in calculating reliable geolocation probabilities. The ability to reproducibly derive reliable geolocation probability underlies the precision and accuracy computations that are paramount in forensics applications.

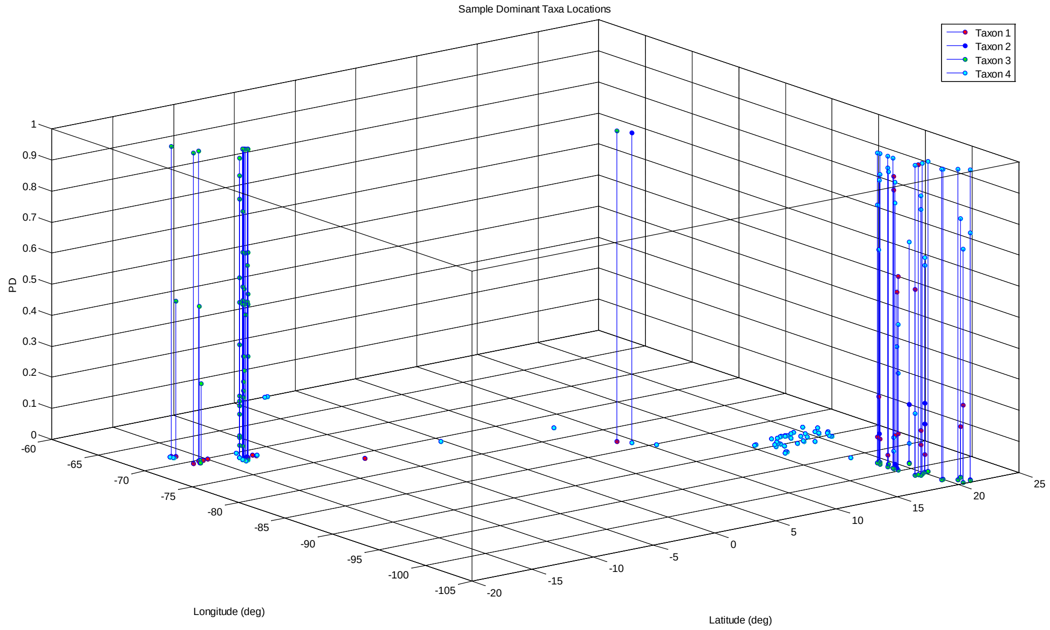


Figure 26. Discrete Probability Distributions for Taxa *Abies, Carya, Clethra,* and *Cupressus* and *Juniperus*.
Plot credit: Carol Christou.

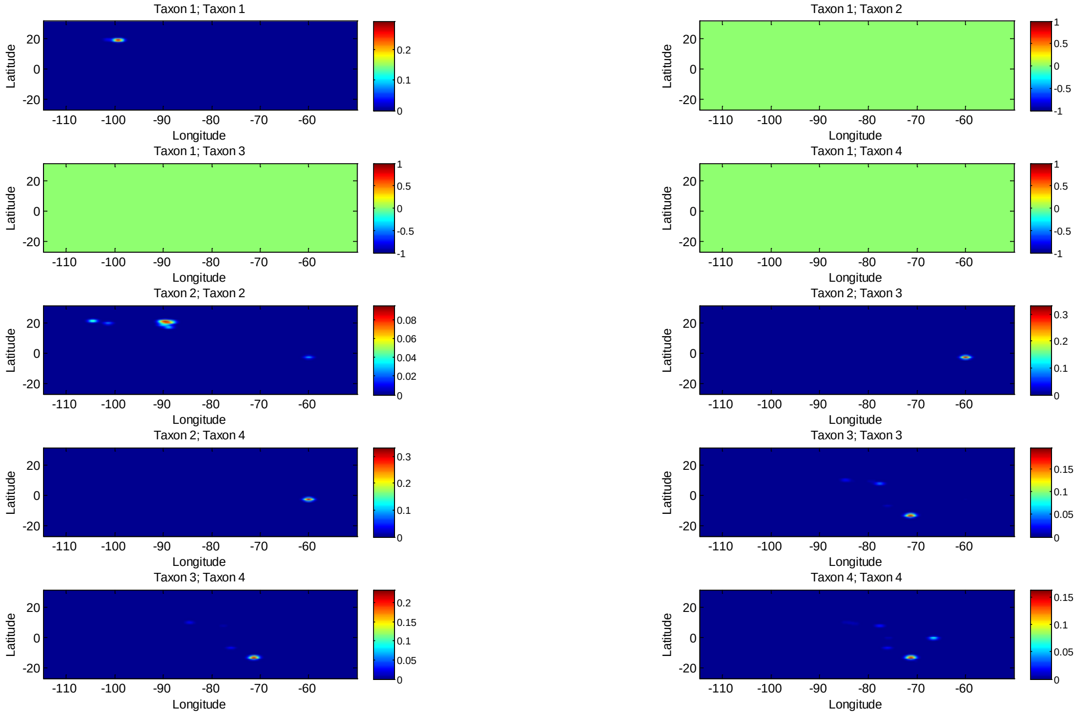


Figure 27. Continuous Joint Probability Distributions for Taxa *Abies, Carya, Clethra,* and *Cupressus* and *Juniperus*.
Figure credit: Carol Christou and Garry Jacyna.

c) Empirical estimators based on pollen grains from well-mixed aircraft air

Mid-term: Another layer of accuracy could be obtained from datasets of well-mixed pollen grains (for identifying the likely false positives). MITRE possesses over 100 freezer-preserved aircraft High Efficiency Particulate Air (HEPA) filter swatches, each of which represents up to 6 months of air travel. Analyses of these filters for pollen grains at the genus or species levels may provide an indication of likely confounders based on their high frequency of occurrence or abundance in aircraft from airlines that operated largely in North America. At a minimum, researchers should conduct an exploratory effort to investigate the forensics value of pollen grains trapped in aircraft filters.

Thrust 3 – Standards Setting for Pollen Collection/Purification/ Mounting/Acquisition

Near-term: The Coalition should devise and adopt a set of standardized protocols for collection strategy, extraction methods, microscopy mounting media, magnification, and acquisition modes for samples that will be used for forensic geolocation. It is critical that other trace materials be preserved in the sample preprocessing steps. These trace materials include soils and anthropogenic matter (*e.g.*, glass, chemicals, fly ash). Therefore, nondestructive extraction methods are preferred over destructive methods.

For example, the proposed metadata fields should include:

1) Identity of collectors (will allow researchers to verify information if needed).
2) Imaging techniques (model of microscopes used, magnification, light source, data acquisition mode).
3) Identity of mountant applied to slide.
4) Originating geolocation (latitude, longitude).
5) Date of collection.
6) Sample type (*e.g.*, soil, sediment, air).
7) Date of field collection.
8) Date of slide preparation (*e.g.*, application of mountant)
9) Date of grain digitization.
10) Identity of all solvents used in slide preparation.
11) Value of expert's confidence level that the pollen grain was labeled correctly, and name of expert.
12) Condition of pollen grains (damaged, pure, occluded, hybrid, genetic drift).

Although Thrust #3 is written consistently for microscopic methods, the Coalition should be flexible in considering spectroscopy and biological techniques for pollen identification in the long-term, should the opportunity become financially tractable.

Thrust 4 – Initiate Collection and Digitization for Low-throughput, High-Quality Archives

Long-term: It is important to recognize that collecting up to 10,000 specimens of targeted pollen types will require extensive travel, as individual herbaria only hold a small proportion of the total flora of the Neotropics. It is also possible that the flowers of some taxa have never been collected or that the herbarium sheets are of fruiting specimens, spent flowers, or sterile plants. Therefore, the Coalition proposes that a distributed Center for Global Pollen Informatics be established over the next decade to build a large reference database based on to-be-developed standards (see Thrust 3). The main herbaria that the Center for Global Pollen Informatics should visit include the Missouri Botanical Gardens, New York Botanical Gardens, University of Chicago, University of Michigan, University of California Berkeley, Fairchild Botanical Gardens, Herbaria of the Royal Botanical Gardens at Kew, Leyden in Amsterdam, and the National Autonomous University of Mexico (Universidad Nacional Autónoma de México; UNAM). Regional herbaria, *e.g.*, those at Lima and Quito, have many specimens, and identification to genus is probably accurate, although species identifications are unlikely to be correct. Nevertheless, the geolocational metadata associated with each specimen is likely to be robust (probably more locally complete than that of the first-tier herbaria). Where there is a low probability that pollen can be identified at the species level, researchers may have to balance the geolocational data with precision in taxonomy.

Palynologists funded by DHS S&T could obtain high-quality images under 1000X magnification using oil immersion and differential inference phase contrast with bright-field illumination under LM (other illuminations may reveal more, but are harder to replicate consistently). Black-and-white imaging would provide the optimal image quality for automatic pollen identification. Representative color photographs could be taken of each pollen type, as these are easier for humans to interpret, and be made available as an on-line resource. (However, color is not useful in pollen image analysis, as it is a product of processing conditions.) SEM images could be collected of each pollen type, with emphasis on groups that are difficult to distinguish under LM (because of small size or lack of distinguishing features). Image-for-image SEM imaging is slower than LM imaging; but the most time-consuming tasks associated with initial pollen preparation and pollen description should already be completed as part of the Coalition's LM imaging process (see Thrust 1). Predicting the rate of SEM imaging is difficult, as sample quality, size of grain, tendency of a grain to collapse during preparation, and number of magnifications needed (typically 3000X, 7500X and 15,000X) are taxon-specific. An average of 6 taxa per day is a realistic imaging rate to anticipate for one instrument (assuming 220 working days per year = 1320 taxa per year).

Thrust 5 – Initiate Digitization for High-throughput Forensic Laboratory

Near-term: The Coalition recognizes that database integration and curation of existing resources can be pursued now, even if new standardized workflows specific to geolocation for data collection are not yet funded. As evidenced by the various presentations at the Workshop, different organizations have different database schemas to serve their particular research objectives and requirements. Clearly, not all presenters shared a geolocation-forensics requirement. Since a thorough effort to fully populate the envisioned "unified massive global pollen database" will probably not occur due to limited resources and risk-reward investment decisions, the Coalition suggests that DHS should make better use of existing resources and incorporate new and improved data processing methods in evolutionary ways.

Long-term: DHS should encourage algorithm improvement experiments (*e.g.*, automatic segmentation, grain classification), especially after benchmarks have been completed that establish best-of-breed algorithm classes (see Thrust 1, mid-term recommendation).

Workshop Agenda

The MITRE Corporation
MITRE-3 Building, 7594 Colshire Drive, McLean, VA 22102
South Lobby Entrance

0815	Registration and Continental Breakfast	
0845	Welcome Remarks	Grace Hwang, *MITRE* and David Masters, *DHS S&T*
0900	Keynote Address: Forensic Palynology & the March for an Automated System	Vaughn M. Bryant, *Texas A&M, The American Association of Stratigraphic Palynologists*
0950	Progress on Automated Pollen Identification	Surangi Punyasena, *University of Illinois*
1030	Break	
1045	DHS CBP Operational Requirements	David Masters, *DHS S&T*; James Sweet, *DHS/Customs & Border Protection (CBP)*
1115	Geographic Attribution of Trace Evidence	Libby Stern and David Korejwo, DoJ/FBI
1145	Conceptual Plan for the Digital Reformatting of the Stanley D. and Gretchen D. Jones Pollen Collection	Susan McCarthy, United States Department of Agriculture
1215	Networking Lunch	(Boxed lunch will be provided)
1245	USGS Palynological Research and Applications – An Overview	Debra Willard, United States Geological Survey
1315	Roundtable #1: Update on Forensic Geolocation Research	Presenter/Moderator: Grace Hwang, *MITRE* Panelists: Vaughn Bryant, *Texas A&M , The American Association of Stratigraphic Palynologists*; Mark Bush, *Florida Institute of Technology*; Peter Siska, *United States Military Academy*
1320	Geolocation using pollen: present capabilities and potential improvements	Mark Bush, *Florida Institute of Technology*
1350	Geospatial Analysis of Pollen for Developing Anti-Terrorist Tracking System	Peter Siska, *United States Military Academy*
1430	Break	
1445	Roundtable #2: Database Schema / Ontology / Standards for Pollen ID	Presenter/Moderator: Eric Grimm, *Illinois State Museum* Panelists: Mark Bush, *Florida Institute of Technology*; Nick Orlans, *MITRE*
1450	The Neotoma Paleoecology Database	Eric Grimm, *Illinois State Museum*
1520	A Neotropical Pollen Database and Key: the next steps	Mark Bush, *Florida Institute of Technology*

1540	*Pollen Database: Center for Excellence in Palynology (CENEX)*--History, Perspective, and Goals	Sophie Warny, *Louisiana State University*
1600	Break	
1615	Roundtable #3: Algorithm Deep-Dive	Presenter/Moderator: Surangi Punyasena, *University of Illinois* Panelists: Kim Riley, *MITRE*; David Tcheng, *University of Illinois*
1620	ARLO: An Advanced Human and Machine-Learning System for Automated Pollen Classification	David Tcheng, *University of Illinois*
1640	Computer Vision Applied to Pollen Classification	Kim Riley, *MITRE*
1730	Next Steps	Moderator: David Masters, *DHS S&T*
1800	Adjourn	
1900	Optional Networking Dinner	No-host dinner at Maggiano's (driving directions will be provided)

Speaker Biographical Information

Mark Bush, Ph.D., *Florida Institute of Technology*

Mark B. Bush received his Ph.D. in 1986 from the University of Hull, U.K., and is currently a Professor of Biology at the Florida Institute of Technology. His research centers on using modern and fossil pollen to detect human influences on landscape and ecological responses to climate change. He has been working with modern and subfossil pollen for over 30 years, with over 25 years of that experience centered on Neotropical palynology. He pioneered modern pollen rain trapping in Central and South America, and the development of pollen-climate transfer functions for the Neotropics. He has conducted fieldwork collecting modern and fossil pollen samples in almost every major habitat type found between highland Mexico and the subtropics of Brazil. He has developed a modern reference collection that includes about 3300 Neotropical pollen types, 1100 of which have been made public via a downloadable pollen database. He has published more than 100 peer-reviewed papers and two books. He has served as an educational consultant to the World Bank and a climate change expert for Conservation International. He is an editor for the *Journal of Biogeography*.

Vaughn Bryant, Ph.D., *Texas A&M University*

Vaughn M. Bryant received his Ph.D. in Botany in 1969 from The University of Texas at Austin and is currently a Professor and the Director of the Palynology Laboratory at Texas A&M University. During the past 40 years he has pioneered the fields of forensic palynology and melissopalynology in the United States. Beginning in 1975 he examined honey samples purchased by the USDA to determine the country of origin. He then presented forensic pollen workshops over many years during the early 1980s at the McCrone Research Institute. He also held a one-week forensic pollen workshop during the late 1980s for US Customs, and since the early 2000s he has assisted various forensic labs and federal agencies by conducting forensic pollen studies of materials associated with national security, terrorist activities, and the import of illegal drugs. During the past decade he has also worked closely with importers, exporters, and private agencies concerned with the illegal sale and shipment of foreign honey into the United States. On a limited basis he has also used pollen data to assist law enforcement personnel in various criminal cases.

Eric Grimm, Ph.D., *Illinois State Museum*

Eric C. Grimm earned his B.S. (1973) in Biology at South Dakota State University and his M.S. (1975) and Ph.D. (1981) in Ecology at the University of Minnesota. He received an NSF Postdoctoral Fellowship focused on quantitative paleoecology at Cambridge University, UK (1981-82). He pursued further NSF-funded postdoctoral work at the University of Minnesota (1983–87), until moving to the Illinois State Museum, where he is now Curator and Chair of the Botany Section and Director of the Landscape History Program. In recent years, his research has centered on the upper Midwest and the northern Great Plains, especially on high-resolution decadal- to centennial-scale oscillations in vegetation and climate, which are important for understanding the near-term response to greenhouse warming. Dr. Grimm has coordinated the North American Pollen Database since its inception in the early 1990s. The database includes fossil-pollen datasets for the Quaternary period and modern pollen surface samples used for calibration. He is now co-Principal Investigator of the Neotoma Paleoecology Database (www.neotomadb.org), an open access multi-proxy community database for the Pliocene and Quaternary.

Grace Hwang, Ph.D., *The MITRE Corporation*

Grace M. Hwang is a Principal Scientist at The MITRE Corporation. She holds a Ph.D. and M.S. in Biophysics and Structural Biology from Brandeis University and an S.M. in Civil and Environmental Engineering from the Massachusetts Institute of Technology, where she developed a microlaser-based plasma-emission sensor for rapid detection of heavy metals. Dr. Hwang joined MITRE in 2005 and was awarded internal research funds to develop photonic sensing technologies for detecting infectious pathogens onboard aircraft. She has worked with the Department of Health and Human Services/Centers for Disease Control on disease spread modeling and serves as a panel member for two National Academies' studies: one on the role of air travel in the transmission of infectious and insect-borne diseases, and the other on evaluating the risk of disease at airports and on aircraft. In parallel, she develops high-risk-and-reward biotechnology research programs for advanced-research government agencies. She enjoys applying her background in engineering, physics, and biology and her experience with advanced sensor technologies to help develop solutions to our most challenging public health and forensics problems. In 2010, one of Dr. Hwang's papers received the prestigious 2010 IEEE Sensors Council Best Paper Award.

David Korejwo, *FBI*

David A. Korejwo is a Geologist/Forensic Examiner in the Trace Evidence Unit of the FBI Laboratory located in Quantico, Virginia. His duties include the forensic examination of soil, rocks, minerals, glass, building materials, and gemstones. He received a BA degree in Environmental Sciences from the University of Virginia and an MS degree in Geosciences from the University of Massachusetts. Prior to joining the FBI, he worked at the US Geological Survey, Reston, Virginia, in the Pollen Laboratory of the Eastern Earth Surface Processes Team, where his duties included collecting, recovering, and identifying fossil and modern pollen for a variety of environmental studies.

David Masters, *Borders and Maritime Division, DHS S&T*

David Masters joined DHS S&T in 2007 as Deputy Director of Research. He currently assists the Director of the Borders and Maritime Division on program development and manages threat detection and tunnel detection projects for the directorate. As the deputy to the Director of the Homeland Security Advanced Research Projects Agency (HSARPA), he planned innovative efforts and oversaw project direction and goals in collaboration with DHS S&T program managers and division leadership. Mr. Masters is a graduate of the School of Information and Computer Science, University of California, Irvine. He has completed many courses and workshops in management, contracting, and acquisition, as well as technical courses in computer science, radar, electro-optics, process control, systems engineering, and materials science. From 1999 to 2007, he served as Deputy Technical Director/Chief Scientist at the Office of Naval Research (ONR), where he managed, analyzed, and evaluated the programs under ONR's Discovery and Invention (D&I) 6.1 (Basic Research) and early 6.2 (Applied Research or Exploratory Development) portfolios. For basic research, he developed goals, strategic objectives, and oversaw a program to develop multidisciplinary ways to predict, detect, neutralize, and mitigate improvised explosive devices. He also led technology workshops to stimulate small and large business participation in Navy and DoD S&T programs. He has earned numerous Spot Awards and Certificates of Commendation for programmatic and technical excellence.

Susan McCarthy, Ph.D., *USDA, National Agricultural Library*

Susan McCarthy is currently serving as the Acting Associate Director for Public Services and Head of the Strategic Programs Branch at the USDA, National Agricultural Library (NAL). NAL is one of four national libraries established in 1862 with the founding of the Department of Agriculture. Susan received a B.S. in Botany from the University of Michigan in 1975 and her Ph.D. in Biology in 1981 from the University of Illinois, with a specialization in plant physiology. Susan had more than 10 years of research experience in chloroplast biogenesis prior to joining NAL in 1990. She began her career at NAL by serving for nine years as the Coordinator for the Plant Genome Data and Information Center. She headed a number of special project teams, and was the program manager for the National Invasive Species Information Center between 2000 and 2010. Today, she is the Head of the Strategic Programs Branch, with seven subject-focused information centers. In the near future, Susan will lead the newly created Knowledge Services Division, which focuses on scientific data management, access, specialized services, and dataset preservation.

Nicholas Orlans, *The MITRE Corporation*

Nicholas M. Orlans is a Senior Principal Engineer at The MITRE Corporation and a Division (Cyber Security) Chief Scientist for Biometrics. He has a Bachelor of Science in mathematics from the University of North Carolina, and received his Master's degree from North Carolina State

University in Computer Science and Architecture. Mr. Orlans works primarily in biometrics and identity management technologies. In the area of biometrics, he supports location-based collection, and performance and conformance topics for accuracy and computation. His current work centers on face and iris recognition, aging, and quality studies. Mr. Orlans was a principal author and editor for the SABER reports and co-authored the book *Biometrics: Identity Assurance in the Information Age.*

Surangi Punyasena, Ph.D., *University of Illinois*

Surangi W. Punyasena is a palynologist, paleoecologist, and evolutionary biologist, whose research focuses on the response of tropical forests to the long-term history of environmental change. Her expertise is in the modern and fossil pollen of the Neotropics, its morphological diversity, and quantitative analyses of the palynological record. Her lab at the University of Illinois, Urbana-Champaign, is currently engaged in two National Science Foundation-funded projects on quantifying the morphological diversity of pollen and developing high-throughput automated analyses of high-diversity Neotropical pollen samples. She completed her BA in Organismal Biology at Yale University (1998), and her SM and Ph.D. in Evolutionary Biology at the University of Chicago (2007). She is a former Fulbright Fellow (to Sri Lanka) and recipient of a Smithsonian Postdoctoral Fellowship for research based at the Center for Tropical Paleoecology and Archaeology at the Smithsonian Research Institute, Panama. She began her career at the University of Illinois in 2008 as an Assistant Professor. Her primary appointment is in Plant Biology, with affiliate positions in the Program in Ecology, Evolution, and Conservation Biology; the Department of Geology; the Department of Geography; and the Illinois Informatics Institute.

Libby Stern, Ph.D., *FBI*

Libby A. Stern is a research chemist in the FBI Laboratory. Her educational background is in geochemistry with a specialization in stable isotopes. Current research efforts include forensic application of isotope ratio measurements and geographic constraints on evidence to enhance investigative lead development. Prior to working for the FBI, Libby Stern was as postdoctoral researcher at UC Berkeley in soil science and on the faculty of the University of Texas at Austin in the Department of Geological Sciences conducting research in environmental geochemistry.

Kim Riley, *The MITRE Corporation*

Kim Riley is a signal processing engineer at The MITRE Corporation. She received her Bachelor's degree in Electrical Engineering with a concentration in Signal Processing from the University of Pittsburgh and received her Master's degree from George Mason University, also in Electrical Engineering with a concentration in Signal Processing. Kim has worked in MITRE's Signal Processing and Communications Analysis Department for approximately 7 years. She has spent the last 3 years concentrating her work on computer vision and pattern recognition.

Peter Siska, Ph.D., *US Military Academy, West Point, NY*

Peter P. Siska received his first Ph.D. in Geography in 1984 from The Komensky University in Bratislava, Slovakia, and a second one in 1995 from Texas A&M University. He is currently a Professor and the Chair of Regional Studies at the United States Military Academy, West Point, NY. During the past 25 years he has worked on a variety of projects involving mapping of ecosystems using field techniques combined with geographic information systems (GIS). For more than a decade he has been working on the development of new applications of statistical methods to map expected pollen distributional patterns over the landscape. One important application of this research would be to develop a method to predict pollen distributional patterns using data from ecosystems, GIS, and geostatistics.

James Sweet, Ph.D., *US Customs and Border Protection Southwest Regional Science Center, Houston, TX*

James D. Sweet holds Bachelor of Science and Doctor of Philosophy degrees in Chemistry from Texas Tech University. He is currently enrolled in the Naval War College. Dr. Sweet is the Laboratory Director of the US Customs and Border Protection (CBP) Southwest Regional Science Center located in Houston, Texas. He oversees direct scientific and forensic support to CBP, the Homeland Security Investigations (HSI) Directorate, and A&M for the majority of the Southern Border and Gulf of Mexico region: nine states and their points of entry (POEs) extending from Alabama to Arizona. The Laboratory supports five Field Offices, nine Border Patrol Sectors, and three Air and Marine Regions with over 100 POEs. In addition to trade support, the laboratory assists the POEs within its area of responsibility with their daily activities in the areas of countervailing, counter narcotics, counter weapons of mass destruction and crime deterrence. Dr. Sweet began his Customs career in 2007 as a US Customs Forensic Scientist in Houston, Texas. He was recruited to support the WMD team as a Special Projects Team Leader, was promoted to Director in 2008, and assumed responsibility for managing the Customs Chicago Field Laboratory. In 2011, Dr. Sweet became the second Director of the Southwest Regional Science Center for CBP. He has been responsible for expanding forensic and trade capabilities of this modern facility and has added two area laboratories: the El Paso Area Satellite Laboratory located in El Paso, TX, and the Joint Arizona Forensic Center in Tucson, AZ. Dr. Sweet is a Commander in the US Navy Reserve and has served four tours in support of combat operations: one in the Balkans and three in the Middle East.

David Tcheng, *University of Illinois, Urbana-Champaign*

David K. Tcheng works as a Research Scientist for the Illinois Informatics Institute (I3) at the University of Illinois at Urbana-Champaign (UIUC). David received a BS from Illinois State University and is currently pursuing a Ph.D. in Informatics at UIUC. David is a machine-learning (ML) specialist and has applied ML to many difficult real-world problems in domains ranging from art to science and involving media types including sound, image, and symbolic sequences. Prior to I3, David worked many years with NCSA and co-founded the Automated Learning Group. Backed by venture funding from I-Ventures, David took a one-year leave of absence from UIUC to start up a music analysis and recommendation company called One Llama Media, Inc.

Sophie Warny, *Louisiana State University*

Sophie Warny is an Assistant Professor of Palynology in the Louisiana State University (LSU) Department of Geology and Geophysics and Curator at LSU's Museum of Natural Science. She has received one of the most prestigious awards bestowed by the National Science Foundation (NSF): its CAREER Award, meant to support junior faculty who exemplify the role of teacher, mentor, and scholar through outstanding research, scholarship, and educational outreach. Warny's research focuses on climate change and biostratigraphy in the historical past of Antarctica using pollen and spores. Already, Warny and her research team have used analysis of fossils of pollen and spores to discover that a previously unknown Antarctic warm period occurred approximately 15.7 million years ago. Her research has been published in journals such as *Nature, Geoscience*, *PNAS*, *Palynology*, *Palaeo3*, and *GSA*. Her current research group is composed of four Ph.D. students and three MS students. These students work on coupled isotopic and palynological analysis. Areas of investigations include Antarctica, Africa, Europe, Canada, South America, and the Gulf of Mexico region. She recently remodeled CENEX, the Center for Excellence in Palynology at LSU, where she currently serves as Interim Director. She is leading the effort to digitize the CENEX pollen collection.

Debra Willard, *US Geological Survey, Reston, VA*

Debra A. Willard has been a research geologist with the US Geological Survey in Reston, Virginia, since 1991. She received her M.S. and Ph.D. in Botany from the University of Illinois at Urbana-Champaign (1985, 1990) and a B.S. in Biology from The Pennsylvania State University (1982). Her palynological research focuses on paleoclimatology and paleoecology of wetland and other terrestrial ecosystems, specifically the response of plant communities to changes in climate and land use. Since 1994, Dr. Willard has researched the ecosystem history of the Everglades wetlands, focusing on the roles played by climate and land use in structuring critical habitats such as tree islands, the ridge and slough landscape, and marl prairies. She has directed and/or collaborated on projects examining response plant communities in the Chesapeake Bay, Tampa Bay, and Roanoke River watersheds to anthropogenic change and climate variability. She has authored or co-authored ~75 scientific articles in ~30 different peer-reviewed journals, including *Frontiers in Ecology and the Environment*, *Geology*, *Ecological Applications*, and *Global and Planetary Change*. She currently is the Program Coordinator for the USGS Climate and Land Use Change Research & Development Program.

References

1. Korejwo, D.A., J.B. Webb, D.A. Willard, and T.P. Sheehan, *Pollen Analysis: An Underutilized Discipline in the U.S. Forensic Science Community*, in *Trace Evidence Symposium*. 2007, FBI Laboratory: Quantico, Virginia.
2. Bush, M.B., M.R. Silman, and D.H. Urrego, 48,000 years of climate and forest change from a biodiversity hot spot. *Science*, 2004. **303**(5659): p. 827-829.
3. Bryant, V.M. and R.G. Holloway, *A late-Quaternary paleoenvironmental record for Texas: an overview of the pollen evidence*, in *Pollen records of late-Quaternary North American sediments*, V.M. Bryant and R.G. Holloway, Editors. 1985, American Association of Stratigraphic Palynologists Foundation: Dallas, TX. p. 39-70.
4. Behling, H., J.C. Berrio, and H. Hooghiemstra, Late Quaternary pollen records from the middle Caqueta river basin in central Colombian Amazon. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 1999. **145**(1-3): p. 193-213.
5. Forup, M.L. and J. Memmott, The restoration of plant-pollinator interactions in hay meadows. *Restoration Ecology*, 2005. **13**(2): p. 265-274.
6. Sears, M.R., G.P. Herbison, M.D. Holdaway, C.J. Hewitt, E.M. Flannery, and P.A. Silva, The relative risks of sensitivity to grass pollen, house dust mite and cat dander in the development of childhood asthma. *Clinical & Experimental Allergy*, 1989. **19**(4): p. 419-424.
7. Warner, J.A., S.A. Little, I. Pollock, J.L. Longbottom, and J.O. Warner, The influence of exposure to house dust mite, cat, pollen and fungal allergens in the home on primary sensitisation in asthma. *Pediatric Allergy and Immunology*, 1990. **1**(2): p. 79-86.
8. Gonzalez-Diaz, S.N., P.G. Rodriguez-Ortiz, A. Arias-Cruz, A. Macias-Weinmann, D. Cid-Guerrero, and G.A. Sedo-Mejia, Atmospheric pollen count in Monterrey, Mexico. *Allergy Asthma Proc*, 2010. **31**(4): p. 341-8.
9. Rocha-Estrada, A., M.A. Alvarado-Vazquez, T.E. Torres-Cepeda, and R. Foroughbakhch-Pournavab, Principales tipos polínicos presentes en el aire de la zona norte del área metropolitana de Monterrey, Nuevo León. *Ciencia Unal*, 2008. **11**(1): p. 69-76.
10. Scharring, S., A. Brandenburg, G. Breitfuss, H. Burkhardt, W. Dunkhorst, M. von Ehr, M. Fratz, D. Giel, U. Heimann, W. Koch, H. Lodding, W. Muller, O. Ronneberger, E. Schultz, G. Sulz, and Q. Wang, *Online monitoring of airborne allergenic particles (OMNIBUSS)*, in *Biophotonics: Visions for Better Health Care*, J. Popp and M. Strehle, Editors. 2006, Wiley-VCH Verlag GmbH & Co. : Weinheim, Germany. p. 31-87.
11. Bryant, V.M. and G.D. Jones, Forensic palynology: Current status of a rarely used technique in the United States of America. *Forensic Science International*, 2006. **163**(3): p. 183-197.
12. Bryant, V.M. and D.C. Mildenhall, *Forensic palynology: a new way to catch crooks*, in *New developments in palynomorph sampling, extraction, and analysis*, V.M. Bryant and J.W. Wrenn, Editors. 1998, American Association of Stratigraphic Palynologists Foundation: Dallas, Texas. p. 145-155.
13. Mildenhall, D.C., P.E.J. Wiltshire, and V.M. Bryant, Forensic palynology: Why do it and how it works. *Forensic Science International*, 2006. **163**(3): p. 163-172.
14. Wiltshire, P.E.J., Consideration of some taphonomic variables of relevance to forensic palynological investigation in the United Kingdom. *Forensic Science International*, 2006. **163**(3): p. 173-182.

15. Rawlins, B.G., S.J. Kemp, E.H. Hodgkinson, J.B. Riding, C.H. Vane, C. Poulton, and K. Freeborough, Potential and pitfalls in establishing the provenance of earth-related samples in forensic investigations. *Journal of Forensic Sciences*, 2006. **51**(4): p. 832-845.
16. Milne, L.A., *A Grain of Truth: How Pollen Brought a Murderer to Justice*. 2005, Sydney: Reed New Holland Press.
17. Nichols, D.J. and H.L. Ott, Biostratigraphy and evolution of the Momipites-Caryapollenites lineage in the early tertiary in the Wind River Basin, Wyoming. *Palynology*, 1978. **2**(1): p. 93-112.
18. Dimbleby, G.W., *The Palynology of Archaeological Sites*. 1985, New York: Academic Press.
19. Stillman, E.C. and J.R. Flenley, The needs and prospects for automation in palynology. *Quaternary Science Reviews*, 1996. **15**(1): p. 1-5.
20. Stoney, D.A., A.M. Bowen, V.M. Bryant, E.A. Caven, M.T. Cimino, and P.L. Stoney, Particle combination analysis for predictive source attribution: Tracing a shipment of contraband ivory. *Journal of American Society of Trace Evidence Examiners*, 2011. **2**(1): p. 13-72.
21. Manten, A.A., Lennart Von Post and the foundation of modern palynology. *Review of Palaeobotany and Palynology*, 1967. **1**(1–4): p. 11-22.
22. Erdtman, *Handbook of Palynology*. 1969, New York: Hafner Publishing Co. 486.
23. Bryant, V.M., *Analytical Techniques in Forensic Palynology*, in *Encyclopedia of Quaternary Science, 2nd Edition* 2013, Elsevier: Oxford, U.K.
24. Bryant, V., *Pollen and spore use in forensics*, in *Wiley Encyclopedia of Forensic Science (2nd Edition)*, A. Jamieson and A. Moenssens, Editors. 2013, John Wiley & Sons, Ltd.: Chichester, U.K.
25. Lark, R.M. and B.G. Rawlins, Can we predict the provenance of a soil sample for forensic purposes by reference to a spatial database? *European Journal of Soil Science*, 2008. **59**(5): p. 1000-1006.
26. Stoney, D.A., A.M. Bowen, P.L. Stoney, and S.B. Sparenga, (U) Review and analysis of geolocation and related forensic source determination efforts I: Overview and method-imposed limitations. *Journal of the Intelligence Community Research and Development*, 2006: p. 1-19.
27. Willard, D.A., C.E. Bernhardt, L. Weimer, S.R. Cooper, D. Gamez, and J. Jensen, Atlas of pollen and spores of the Florida everglades. *Palynology*, 2004. **28**(1): p. 175-227.
28. Bernhardt, C.E., D.A. Willard, and R. Peet, *Atlas of pollen spores of mid-Atlantic wetlands, USA*. In Prep.
29. Landacre, B., C.E. Bernhardt, and D.A. Willard, *Atlas of pollen and spores of the Chesapeake Bay watershed*. In Prep.
30. U.S. Customs and Border Protection. *Overview of revenue: A priority trade issue (PTI)*. January 22, 2013.
31. Burn, M.J. and F.E. Mayle, Palynological differentiation between genera of the Moraceae family and implications for Amazonian palaeoecology. *Review of Palaeobotany and Palynology*, 2008. **149**(3–4): p. 187-201.
32. Persson, A., Frequenzen von Kiefernpollen in Sudschweden 1953 und 1954. *Zeitung Forstgenetische Forstpflanzenzuchtung* 1954. **4**: p. 129.
33. Rogers, C.A. and E. Levetin, Evidence of long-distance transport of mountain cedar pollen into Tulsa, Oklahoma. *International Journal of Biometeorology*, 1998. **42**(2): p. 65-72.

34. Milne, L.A., V.M. Bryant, and Mildenhall. D.C., *Chapter 14, Forensic Palynology*. 2005, CRC Press LLC: Boca Raton. p. 318.

35. Islebe, G.A. and H. Hooghiemstra, Recent pollen spectra of highland Guatemala. *Journal of Biogeography*, 1995. **22**(6): p. 1091-1099.

36. Kennedy, L.M., S.P. Horn, and K.H. Orvis, Modern pollen spectra from the highlands of the Cordillera Central, Dominican Republic. *Review of Palaeobotany and Palynology*, 2005. **137**(1-2): p. 51-68.

37. Kuentz, A., A.G. De Mera, M.P. Ledru, and J.C. Thouret, Phytogeographical data and modern pollen rain of the puna belt in southern Peru (Nevado Coropuna, Western Cordillera). *Journal of Biogeography*, 2007. **34**(10): p. 1762-1776.

38. Reese, C.A. and K.B. Liu, A modern pollen rain study from the central Andes region of South America. *Journal of Biogeography*, 2005. **32**(4): p. 709-718.

39. Rodgers, J.C., III and S.P. Horn, Modern pollen spectra from Costa Rica. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 1996. **124**: p. 53-71.

40. McNeil, M.E.A., *Meet Vaughn Bryant, Honey Sleuth*, in *Bee Culture* 2012, A.I. Root & The Wright Brothers: Medina, Ohio.

41. Smith, P.A., *The Hive Minder*, in *WIRED.* 2012, Advance Publications Inc.: San Francisco, CA. p. 024.

42. Bryant, V.M. and G.D. Jones, The R-values of honey: Pollen coefficients. *Palynology*, 2001. **25**(1): p. 11-28.

43. Siska, P.P., V.M. Bryant, and J.G. Jones, The spatial analysis of modern pollen rain in Big Bend national park. *Palynology*, 2001. **25**(1): p. 199-216.

44. Jones, G.D., V.M. Bryant, Jr., M.H. Lieux, S.D. Jones, and P.D. Lingren, *Pollen of the Southeastern United States with Emphasis on Melissopalynology and Entomopalynology*. AASP Foundation Contribution Series 1995, Houston, Texas: American Association of Stratigraphic Polynologists Foundation. 76 + 104 plates.

45. Jones, G.D. and V.M. Bryant, A comparison of pollen counts: Light versus scanning electron microscopy. *Grana*, 2007. **46**(1): p. 20-33.

46. Correa-Metrio, A., M.B. Bush, L. Pérez, A. Schwalb, and K.R. Cabrera, Pollen distribution along climatic and biogeographic gradients in northern Central America. *The Holocene*, 2011. **21**(4): p. 681-692.

47. Gauch, J.H.G., *Multivariate Analysis in Community Ecology*. 1982, Cambridge: Cambridge University Press.

48. Hill, M.O., *Ecology and Systematics*, in *DECORANA - A FORTRAN program for detrended correspondence analysis and reciprocal averaging*. 1979, Cornell University New York.

49. McCune, B. and J.B. Grace, *Analysis of ecological communities*. 2002, Gleneden Beach, Oregon: MJM Software Design.

50. Birks, H.J.B., *Quantitative palaeoenvironmental reconstructions from Holocene biological data*, in *Global change in the Holocene*, A. MacKay, R. Battarbee, and J. Birks, Editors. 2003, Arnold: London. p. 107-123.

51. Birks, H.J.B. and A.D. Gordon, *Numerical Methods in Quaternary Pollen Analysis*. 1985, London: Academic Press. 317.

52. Anderson, A.J.B., Ordination Methods in Ecology. *Journal of Ecology*, 1971. **59**(3): p. 713-726.

53. McCune, B. and M.J. Mefford, *PC-ORD: Multivariate analysis of ecological data: Version 4 for Windows*. 1999, Gleneden Beach, Oregon: MJM Software Design.

54. Urrego, D.H., M.B. Bush, and M.R. Silman, A long history of cloud and forest migration from Lake Consuelo, Peru. *Quaternary Research*, 2010. **73**(2): p. 364-373.

55. Urrego, D.H., M.B. Bush, M.R. Silman, B.A. Niccum, P. De La Rosa, C.H. McMichael, S. Hagen, and M. Palace, Holocene fires, forest stability and human occupation in south-western Amazonia. *Journal of Biogeography*, 2013: **40**(3): p. 521-533.
56. Restrepo, A., P. Colinvaux, M. Bush, A. Correa-Metrio, J. Conroy, M.R. Gardener, P. Jaramillo, M. Steinitz-Kannan, and J. Overpeck, Impacts of climate variability and human colonization on the vegetation of the Galápagos Islands. *Ecology*, 2012. **93**(8): p. 1853-1866.
57. Weng, C., M.B. Bush, and M.R. Silman, An analysis of modern pollen rain on an elevational gradient in southern Peru. *Journal of Tropical Ecology*, 2004. **20**(1): p. 113-124.
58. Cochrane, M.A. and C.P. Barber, Climate change, human land use and future fires in the Amazon. *Global Change Biology*, 2009. **15**(3): p. 601-612.
59. Bush, M.B., E. Moreno, P.E. De Oliveira, E. Asanza, and P.A. Colinvaux, The influence of biogeographic and ecological heterogeneity on Amazonian pollen spectra. *Journal of Tropical Ecology*, 2001. **17**(5): p. 729-743.
60. Ferrier, S. and A. Guisan, Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 2006. **43**(3): p. 393-404.
61. Siska, P.P., I.-K. Hung, and V.M. Bryant Jr, *Correlation between pollen dispersion and forest spatial distribution patterns in the Southeastern United States*, in *Proceedings of the 5th Southern Forestry and Natural Resources GIS Conference*, S.P. Prisley, I.-K. Bettinger, and J. Kushla, Editors. 2006, Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA: Asheville, NC.
62. Whitmore, J., K. Gajewski, M. Sawada, J.W. Williams, B. Shuman, P.J. Bartlein, T. Minckley, A.E. Viau, T. Webb III, S. Shafer, P. Anderson, and L. Brubaker, Modern pollen data from North America and Greenland for multi-scale paleoenvironmental applications. *Quaternary Science Reviews*, 2005. **24**(16-17): p. 1828-1848.
63. Riley, K., J. Woodard, and G.M. Hwang, *Automated pollen classification*, In Prep: McLean, VA.
64. Clinebell, I., O.L. Phillips, A.H. Gentry, N. Stark, and H. Zuuring, Prediction of neotropical tree and liana species richness from soil and climatic data. *Biodiversity and Conservation*, 1995. **4**(1): p. 56-90.
65. Gentry, A.H., ed. *Contrasting phytogeographic patterns of upland and lowland Panamanian plants*. The botany and natural history of Panama: La botánica e historia natural de Panamá. Monogr. Syst. Bot. , ed. W.G. D'Arcy and M.D. Correa-A. 1985, Missouri Botanical Garden: St. Louis. 147-160.
66. Gentry, A.H., Changes in Plant Community Diversity and Floristic Composition on Environmental and Geographical Gradients. *Annals of the Missouri Botanical Garden*, 1988. **75**(1): p. 1-34.
67. Gentry, A.H., Tropical Forest Biodiversity: Distributional Patterns and Their Conservational Significance. *Oikos*, 1992. **63**(1): p. 19-28.
68. Punyasena, S.W., D.K. Tcheng, C. Wesseln, and P.G. Mueller, Classifying black and white spruce pollen using layered machine learning. *New Phytologist*, 2012. **196**(3): p. 937-944.
69. Sivaguru, M., L. Mander, G. Fried, and S.W. Punyasena, Capturing the surface texture and shape of pollen: a comparison of microscopy techniques. *PLoS One*, 2012. **7**(6).
70. Colinvaux, P.A., P.E. De Oliveira, and J.E. Moreno, *Amazon pollen manual and atlas*. 1999, New York: Harwood Academic Press. 332.

71. Roubik, D.W. and P.J.E. Moreno, *Pollen and Spores of Barro Colorado Island*. Monographs in Systematic Botany 36. 1991, St. Louis, MO Missouri Botanical Garden Press. 268.

72. Bush, M.B. and C. Weng, Introducing a new (freeware) tool for palynology. *Journal of Biogeography*, 2007. **34**(3): p. 377-380.

73. Bush, M.B., Neotropical plant reproductive strategies and fossil pollen representation. *American Naturalist*, 1995. **145**(4): p. 594-609.

74. Bush, M.B. and R. Rivera, Pollen dispersal and representation in a neotropical rain forest. *Global Ecology and Biogeography Letters*, 1998. **7**(5): p. 379-392.

75. Bush, M.B. and R. Rivera, Reproductive ecology and pollen representation among neotropical trees. *Global Ecology and Biogeography*, 2001. **10**(4): p. 359-368.

76. Pearson, R.G., *Species' Distribution Modeling for Conservation Educators and Practitioners*, 2007, American Museum of Natural History: New York.

77. Willard, D., C. Bernhardt, R. Brown, B. Landacre, and P. Townsend, Development and application of a pollen-based paleohydrologic reconstruction from the Lower Roanoke River Basin, North Carolina, USA. *The Holocene*, 2011. **21**(2): p. 305-317.

78. Faegri, K. and P. Deuse, Size variations in pollen grains with different treatments. *Pollen et Spores*, 1960. **2**: p. 293-298.

79. Brown, A.G., A. Smith, and O. Elmhurst, The combined use of pollen and soil analyses in a search and subsequent murder investigation. *Journal of Forensic Sciences*, 2002. **47**(3): p. 614-618.

80. Feeley, K.J. and M.R. Silman, Land-use and climate change effects on population size and extinction risk of Andean plants. *Global Change Biology*, 2010. **16**(12): p. 3215-3222.

81. Tobler, M., E. Honorio, J. Janovec, and C. Reynel, Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (*Moraceae and Myristicaceae*) in Peru. *Biodiversity and Conservation*, 2007. **16**(3): p. 659-677.

82. Kamino, L.H.Y., J.R. Stehmann, S. Amaral, P. De Marco, T.F. Rangel, M.F. de Siqueira, R. De Giovanni, and J. Hortal, Challenges and perspectives for species distribution modelling in the neotropics. *Biology Letters*, 2012. **8**(3): p. 324-326.

83. Klemic, J.F., *Pollen and Spores as Trace Evidence on IED Components: An Overview of Forensic Palynology with Recommendations for its Development as a Weapons Technical Intelligence Tool*, 2010, MITRE Corporation: McLean, VA.

84. Flenley, J.R., ed. *The problem of pollen recognition*. Problems in Picture Interpretation, ed. M.B. Clowes and J.P. Penny. 1968, C.S.I.R.O.: Canberra. 141-145.

85. Rittenour, W.R., R.G. Hamilton, D.H. Beezhold, and B.J. Green, Immunologic, spectrophotometric and nucleic acid based methods for the detection and quantification of airborne pollen. *Journal of Immunological Methods*, 2012. **383**(1–2): p. 47-53.

86. Mirkin, G.R. and L.L. Bagdasaryan, The feasibility of identifying paleontological objects with the aid of optical analyzing systems. *Paleontological Journal*, 1972. **6**: p. 103-8.

87. Langford, M., G.E. Taylor, and J.R. Flenley, Computerized identification of pollen grains by texture analysis. *Review of Palaeobotany and Palynology*, 1990. **64**(1–4): p. 197-203.

88. Li, P. and J.R. Flenley, Pollen texture identification using neural networks. *Grana*, 1999. **38**: p. 59-64.

89. Li, P., W.J. Treloar, J.R. Flenley, and L. Empson, Towards automation of palynology 2: The use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *Journal of Quaternary Science*, 2004. **19**(8): p. 755-762.

90. Treloar, W.J., G.E. Taylor, and J.R. Flenley, Towards automation of palynology 1: Analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *Journal of Quaternary Science*, 2004. **19**(8): p. 745-754.

91. Vitha, S., V.M. Bryant, A. Zwa, and A. Holzenburg, 3D confocal imaging of pollen. *Microscopy and Microanalysis*, 2009. **15**(SUPPL. 2): p. 622-623.

92. Ronneberger, O., E. Schultz, and H. Burkhardt, Automated pollen recognition using 3D volume images from fluorescence microscopy. *Aerobiologia*, 2002. **18**(2): p. 107-115.

93. Ranzato, M., P.E. Taylor, J.M. House, R.C. Flagan, Y. LeCun, and P. Perona, Automatic recognition of biological particles in microscopic images. *Pattern Recognition Letters*, 2007. **28**(1): p. 31-39.

94. Allen, G., *An Automated Pollen Recognition System*, in *Institute of Information Sciences and Technology* 2006, Massey University: Turitea, Palmerston North, New Zealand.

95. Holt, K., G. Allen, R. Hodgson, S. Marsland, and J. Flenley, Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology*, 2011. **167**(3-4): p. 175-183.

96. Zhang, Y., D.W. Fountain, R.M. Hodgson, J.R. Flenley, and S. Gunetileke, Towards automation of palynology 3: Pollen pattern recognition using Gabor transforms and digital moments. *Journal of Quaternary Science*, 2004. **19**(8): p. 763-768.

97. Dell'Anna, R., P. Lazzeri, M. Frisanco, F. Monti, F. Malvezzi Campeggi, E. Gottardini, and M. Bersani, Pollen discrimination and classification by Fourier transform infrared (FT-IR) microspectroscopy and machine learning. *Analytical and Bioanalytical Chemistry*, 2009. **394**(5): p. 1443-1452.

98. Schulte, F., J. Lingott, U. Panne, and J. Kneipp, Chemical characterization and classification of pollen. *Analytical Chemistry*, 2008. **80**(24): p. 9551-9556.

99. Pappas, C.S., P.A. Tarantilis, P.C. Harizanis, and M.G. Polissiou, New method for pollen identification by FT-IR spectroscopy. *Applied Spectroscopy*, 2003. **57**(1): p. 23-27.

100. Zimmermann, B., Characterization of pollen by vibrational spectroscopy. *Appl Spectrosc*, 2010. **64**(12): p. 1364-73.

101. Gottardini, E., S. Rossi, F. Cristofolini, and L. Benedetti, Use of Fourier transform infrared (FT-IR) spectroscopy as a tool for pollen identification. *Aerobiologia*, 2007. **23**(3): p. 211-219.

102. Kano, H. and H.O. Hamaguchi, Vibrational imaging of a single pollen grain by ultrabroadband multiplex coherent anti-stokes Raman scattering microspectroscopy. *Chemistry Letters*, 2006. **35**(10): p. 1124-1125.

103. Kang, D.Y., M.S. Son, C.H. Eum, W.S. Kim, and S. Lee, Size determination of pollens using gravitational and sedimentation field-flow fractionation. *Bulletin of the Korean Chemical Society*, 2007. **28**(4): p. 613-618.

104. Descolas-Gros, C. and C. Schölzel, Stable isotope ratios of carbon and nitrogen in pollen grains in order to characterize plant functional groups and photosynthetic pathway types. *New Phytologist*, 2007. **176**(2): p. 390-401.

105. Nelson, D.M., F.S. Hu, D.R. Scholes, N. Joshi, and A. Pearson, Using SPIRAL (Single Pollen Isotope Ratio AnaLysis) to estimate C3- and C4-grass abundance in the paleorecord. *Earth and Planetary Science Letters*, 2008. **269**(1–2): p. 11-16.

106. Eliet, J.R. and S.A. Harbison, *The development of a DNA analysis system for pollen*, International Congress Series, 2006. 1288: p. 825-827.

107. Zhou, L.J., K.Q. Pei, B. Zhou, and K.P. Ma, A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil. *American Journal of Botany*, 2007. **94**(3): p. 477-481.

108. Folloni, S., D.M. Kagkli, B. Rajcevic, N.C. Guimaraes, B. Van Droogenbroeck, F.H. Valicente, G. Van den Eede, and M. Van den Bulcke, Detection of airborne genetically modified maize pollen by real-time PCR. *Mol Ecol Resour*, 2012. **12**(5): p. 810-21.

109. Longhi, S., A. Cristofori, P. Gatto, F. Cristofolini, M.S. Grando, and E. Gottardini, Biomolecular identification of allergenic pollen: A new perspective for aerobiological monitoring? *Annals of Allergy, Asthma and Immunology*, 2009. **103**(6): p. 508-514.

110. Brieman, L., *Random Forests*, in *Machine Learning*, R.E. Shapiro, Editor. 2001, Kluwer Academic Publisher: Netherlands. p. 5-32.

111. Cortes, C. and V. Vapnik, Support-vector networks. *Machine Learning*, 1995. **20**(3): p. 273-297.

112. Filipovych, R. and E. Ribeiro, *Learning human motion models from unsegmented videos*, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: Anchorage, Alaska, June 2008.

113. Filipovych, R. and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR):* Anchorage, Alaska, June 2008.

114. Belongie, S., J. Malik, and J. Puzicha, Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002. **24**: p. 509-521.

115. Liu, W. and E. Ribeiro, Incremental variations of image moments for nonlinear image registration. *Signal, Image and Video Processing*, 2012: p. 1-10.

116. Haselhorst, D.S., J.E. Moreno, and S.W. Punyasena, Variability within the 10-Year Pollen Rain of a Seasonal Neotropical Forest and Its Implications for Paleoenvironmental and Phenological Research. *PLoS One*, 2013. **8**(1): p. 1-13.

117. Woodard, J.P., *Limited Release - Available upon request. Classification of Pieris and Colias butterflies: A feasibility study*, 2010, MITRE Corporation: McLean, VA.

118. Ojala, T., M. Pietikäinen, and T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002. **24**(7): p. 971-987.

119. Tcheng, D., B. Lambert, S.C.Y. Yu, and L. Rendell. Building robust learning systems by combining induction and optimization. in *IJCAI-89 Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*. 1989, Morgan Kaufman: San Mateo, California. p. 806-812.

120. Tcheng, D.K., B.L. Lambert, S.C.Y. Lu, and L.A. Rendell, *Aims - an Adaptive Interactive Modeling System for Supporting Engineering Decision-Making*, in *Machine Learning: Proceedings of the Eighth International Workshop*. 1991, Morgan Kaufman: San Mateo, California. p. 645-649.

121. Serratosa, F., R. Alquézar, and A. Sanfeliu, Function-described graphs for modelling objects represented by sets of attributed graphs. *Pattern Recognition*, 2003. **36**(3): p. 781-798.

122. Wiskott, L., J.-M. Fellous, N. Kruger, and C. von der Malsburg, *Face recognition by elastic bunch graph matching*, in *Intelligent biometric techniques in fingerprint and face recogntion*, L.C. Jain, et al., Editors. 1999, CRC Press: Boca Raton, FL.

123. Lowe, D.G., Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. **60**(2): p. 91-110.

124. Morel, J.-M. and G. Yu, ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *Siam Journal on Imaging Sciences*, 2009. **2**(2): p. 438-469.

125. Ahonen, T., J. Matas, C. He, and M. Pietikainen, Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features. *Lecture Notes in Computer Science*, 2009. **5575**: p. 61-70.

126. Torralba, A., A. Oliva, and W.T. Freeman, Object recognition by scene alignment. *Journal of Vision*, 2003. **3**(9): p. 196.

127. France, I., A.W.G. Duller, G.A.T. Duller, and H.F. Lamb, A new approach to automated pollen analysis. *Quaternary Science Reviews*, 2000. **19**(6): p. 537-546.

128. Saunders, C.P., L.J. Davis, A.C. Lamas, J.J. Miller, and D.T. Gantz, Construction and Evaluation of Classifiers for Forensic Document Analysis. *Annals of Applied Statistics*, 2011. **5**(1): p. 381-399.

129. Yiu, K.K., M.W. Mak, and C.K. Li, Gaussian Mixture Models and Probabilistic Decision-Based Neural Networks for Pattern Classification: A Comparative Study. *Neural Computing & Applications*, 1999. **8**(3): p. 235-245.

130. Yiu, K.K., M.W. Mak, and C.K. Li. Probabilistic decision-based neural networks for speech pattern classification. in *Proceedings of ICSP'98 Fourth International Conference on Signal Processing, 12-16 Oct. 1998*. Piscataway, NJ, USA: IEEE.

131. Wu, M. and Z. Zhang, *Handwritten digit classification using the MNIST data set.*, in *Course project CSE802: Pattern Classification & Analysis.* 2010, Michigan State University: Lansing, Michigan.

132. Grabandt, R.A.J., Pollen rain in relation to arboreal vegetation in the Columbian Cordillera Oriental. *Review of Palaeobotany and Palynology*, 1980. **29**: p. 65-147.

133. Bush, M.B., Deriving response matrices from Central American modern pollen rain. *Quaternary Research*, 2000. **54**(1): p. 132-143.

134. Collins, A. and M.B. Bush, An analysis of modern pollen representation and climatic conditions on the Galápagos Islands *The Holocene*, 2011. **21**(2): p. 237-250.

135. Urrego, D.H., M.R. Silman, A. Correa-Metrio, and M.B. Bush, Pollen-vegetation relationships along steep climatic gradients in western Amazonia. *Journal of Vegetation Science*, 2011. **22**(5): p. 795-806.