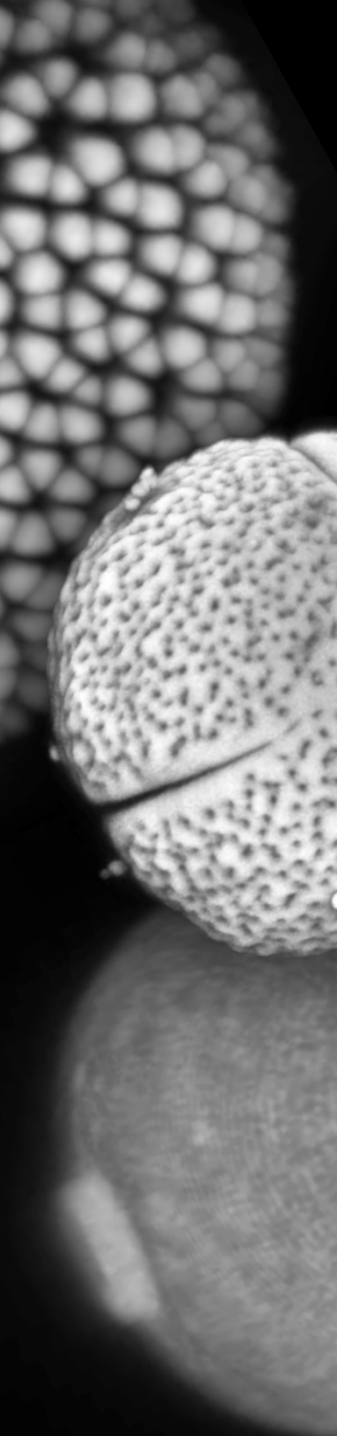


An automated image analysis platform for palynological specimens

---

Surangi W. Punyasena
Department of Plant Biology
University of Illinois

Kenton McHenry
National Center for Supercomputing Applications
University of Illinois

Fossil pollen =

a microscopic census of past vegetation, preserved in geologic sediments

POLLEN APPLICATIONS

Biostratigraphy

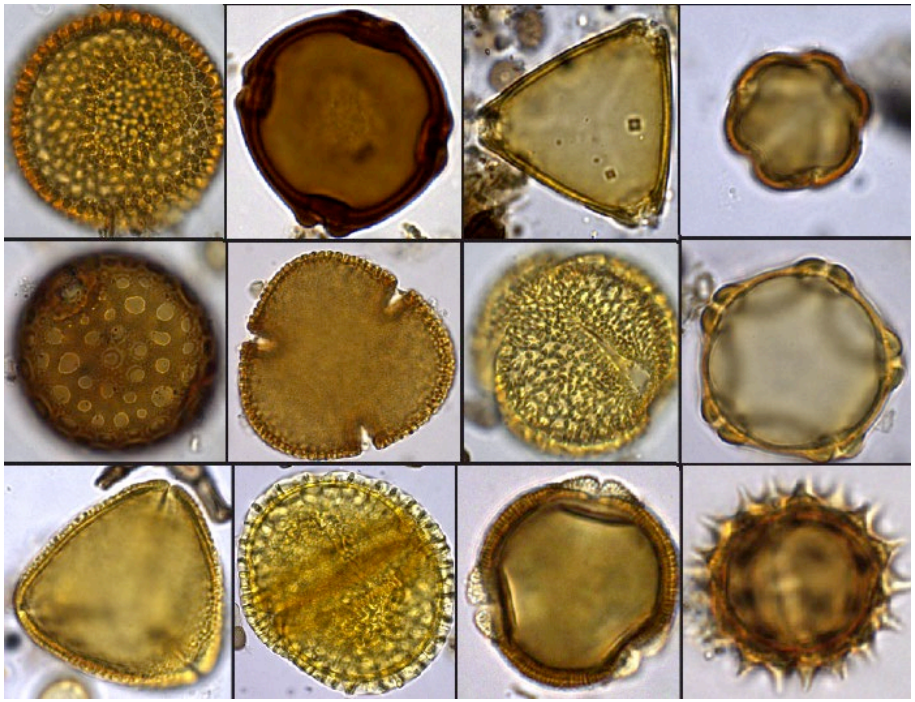
Paleoclimate

Paleoecology

Plant evolution

Forensics

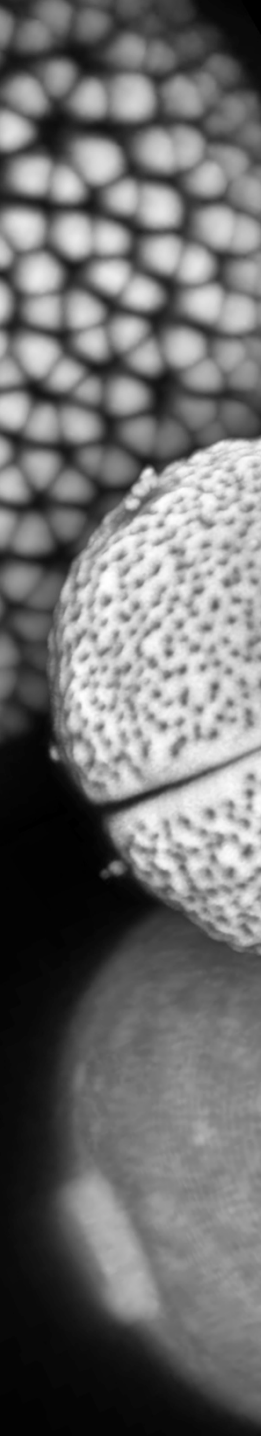
POLLEN AS “BIG DATA”



470 million years of plant history

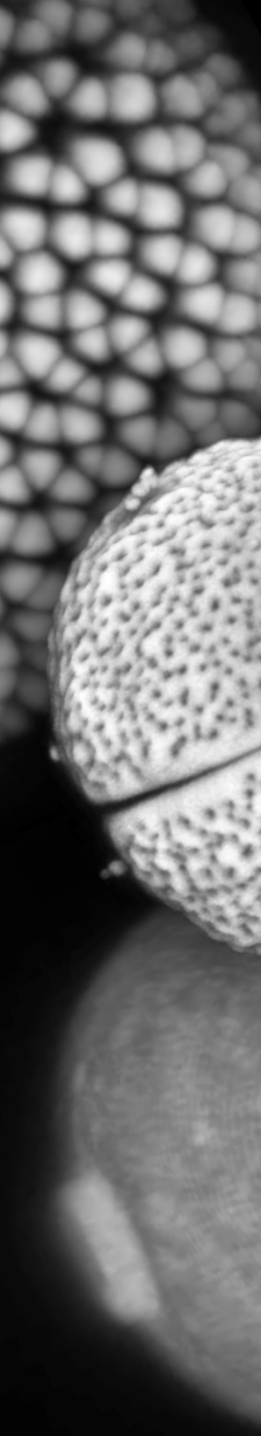
Billions of potential specimens

Continuous deposition across a range of environments



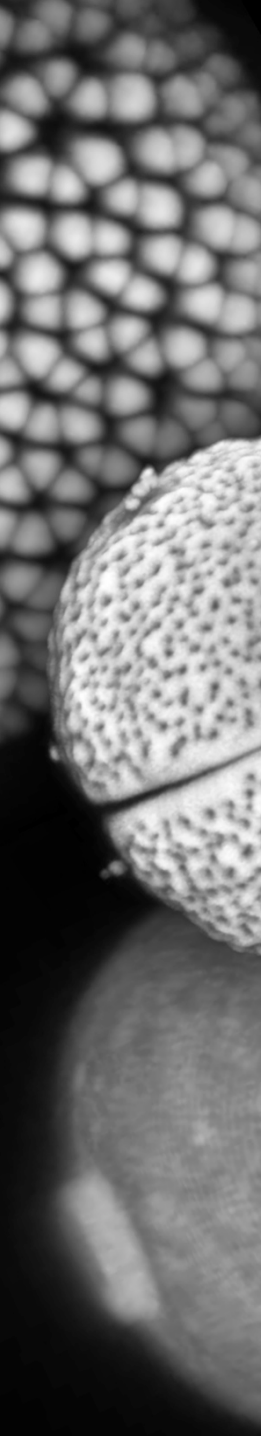
WHAT HOLDS US BACK?

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation



WHAT HOLDS US BACK?

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation



WHAT HOLDS US BACK?

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation

DATA QUANTITY

How do we transform pollen analysis into a higher-throughput “big data” discipline?

DATA REPRODUCIBILITY

How do we improve the consistency and accuracy of pollen identifications?

TAXONOMIC RESOLUTION

How do we discriminate among morphologically similar taxa?

DATA QUANTITY

How do we transform pollen analysis into a higher-throughput “big data” discipline?

DATA REPRODUCIBILITY

How do we improve the consistency and accuracy of pollen identifications?

TAXONOMIC RESOLUTION

How do we discriminate among morphologically similar taxa?

DATA QUANTITY

How do we transform pollen analysis into a higher-throughput “big data” discipline?

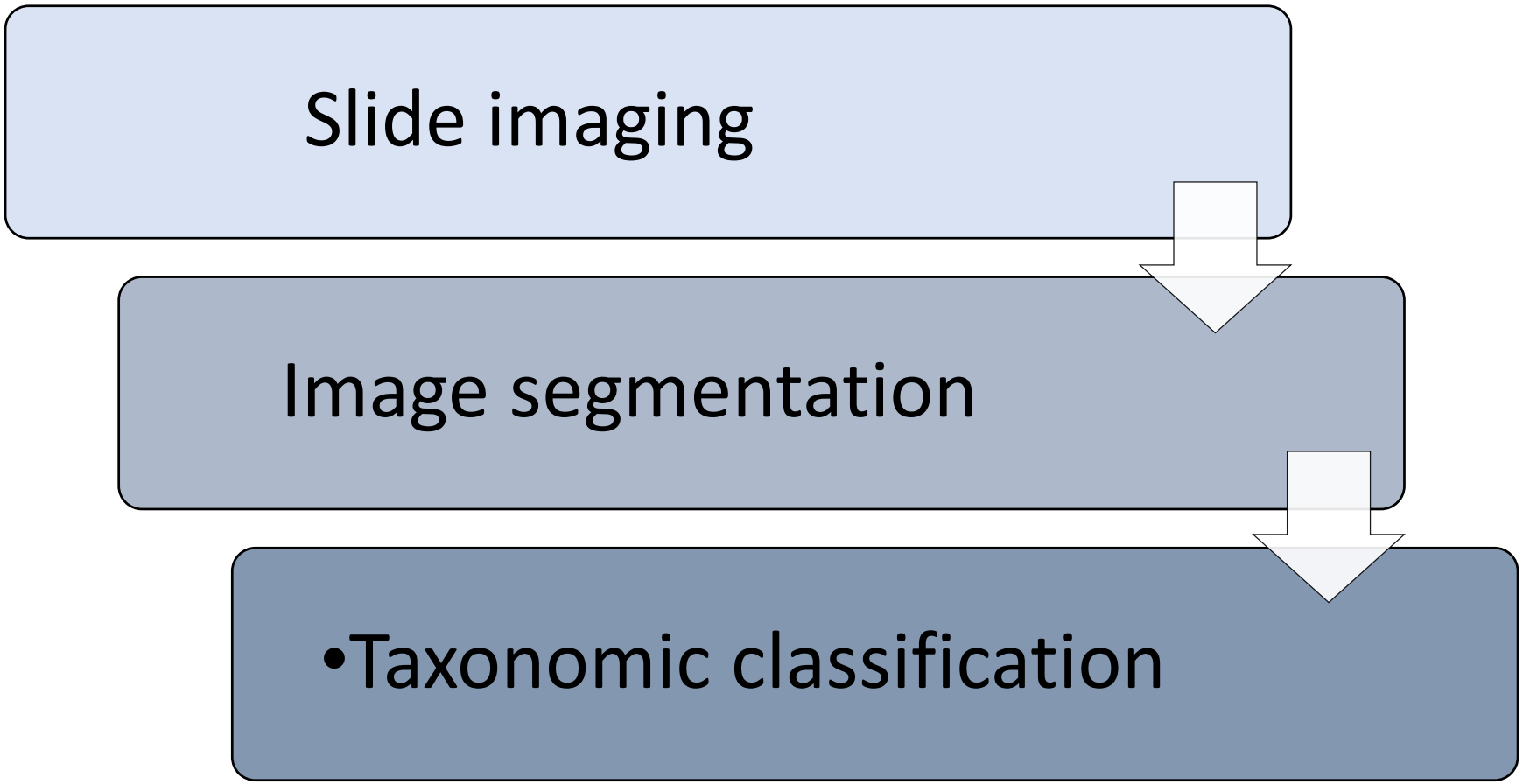
DATA REPRODUCIBILITY

How do we improve the consistency and accuracy of pollen identifications?

TAXONOMIC RESOLUTION

How do we discriminate among morphologically similar taxa?

AUTOMATED POLLEN CLASSIFICATION WORKFLOW

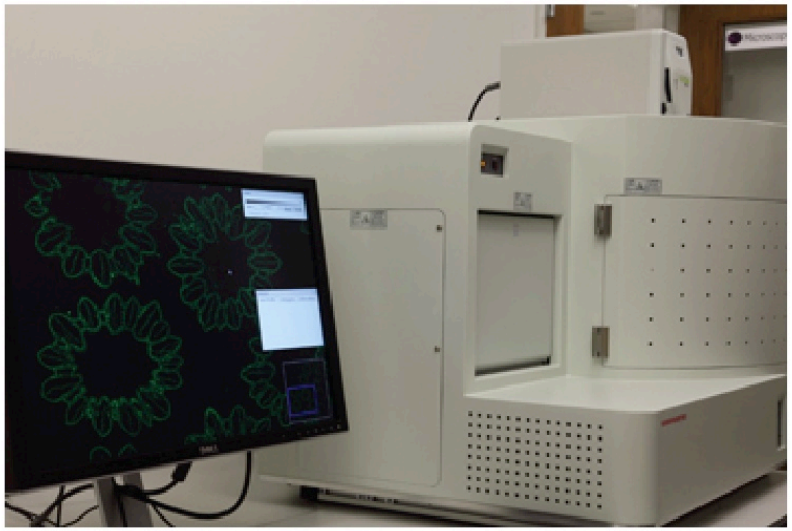


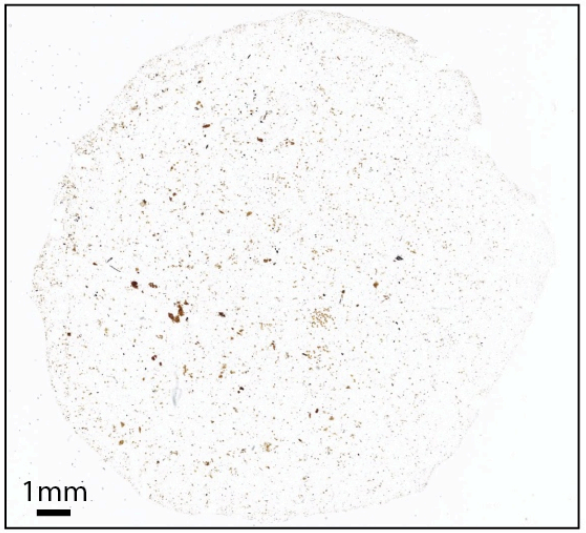
Charless Fowlkes
(UC Irvine)

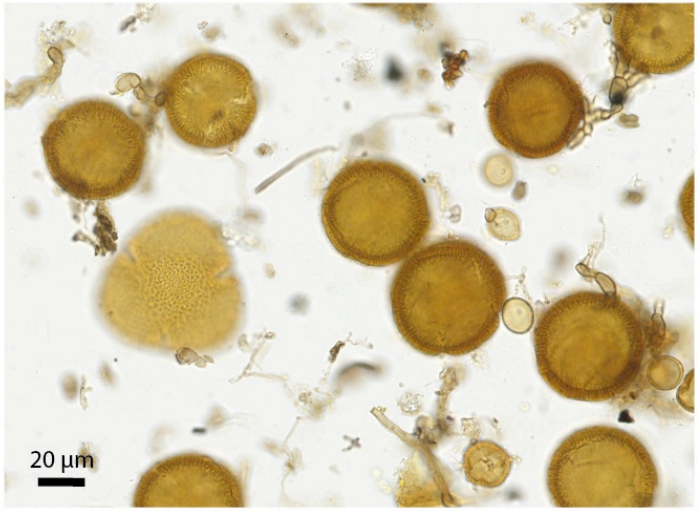
Shu Kong
(UC Irvine)

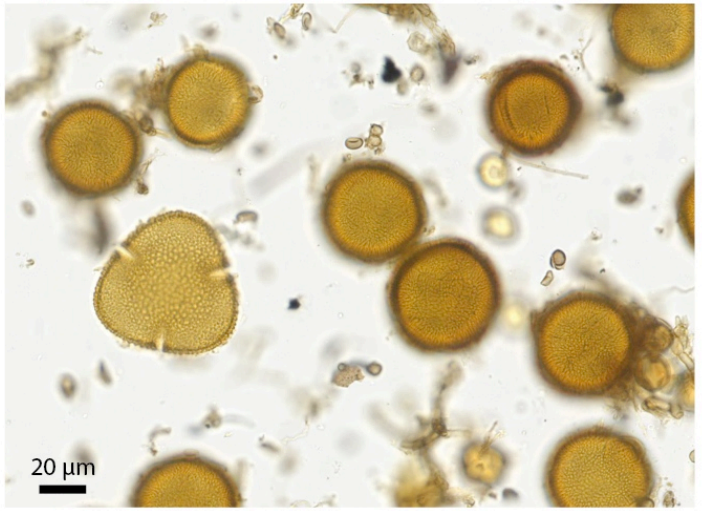
Derek Haselhorst
(U Illinois)

AUTOMATED IMAGING / VIRTUAL SLIDES









400x, 0.23 µm/pixel

One sample (41 @ 1 µm axial planes) = ~400 GB

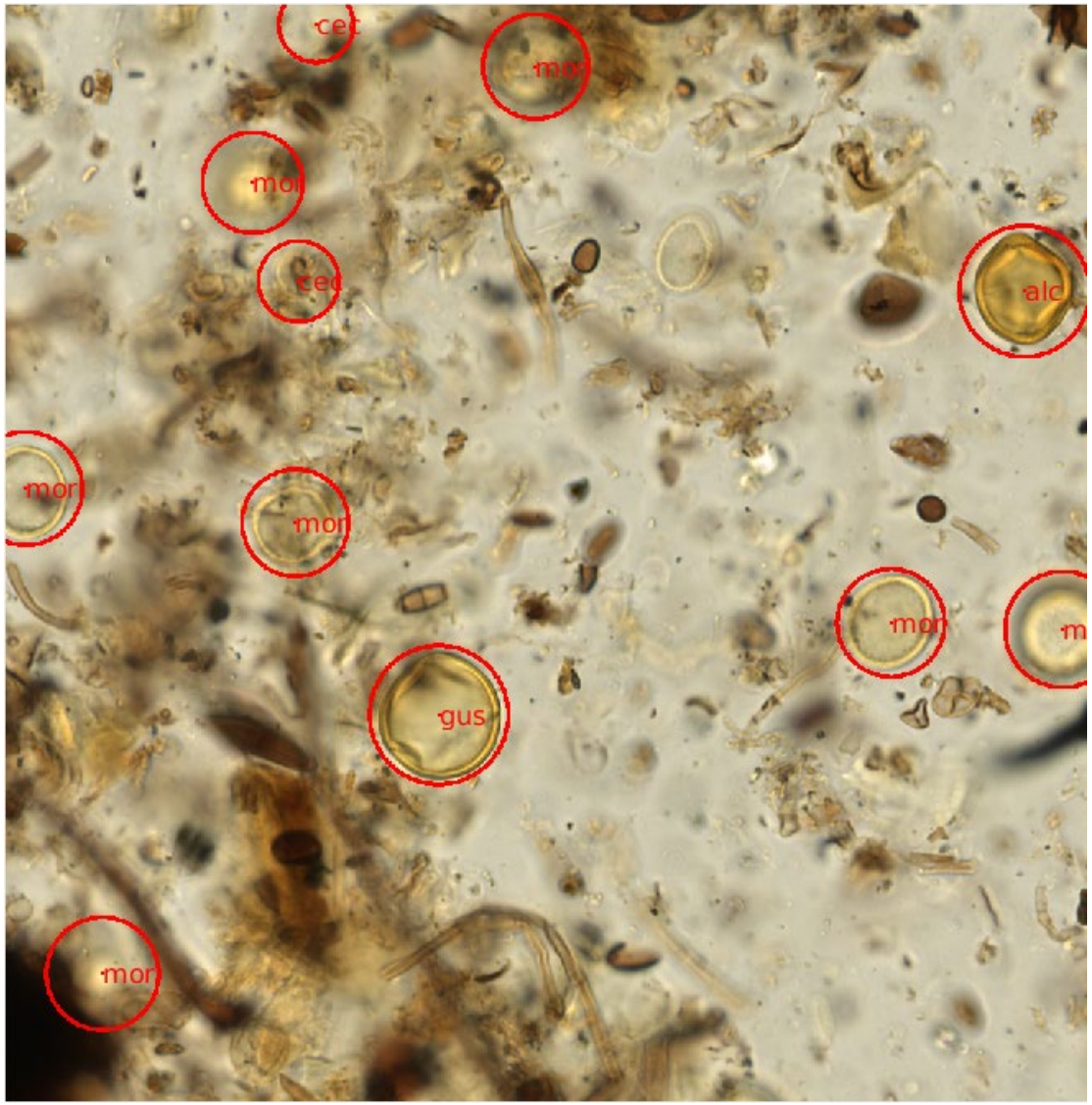
EXPERT ANNOTATIONS

PNG image stacks randomly subsampled

Subsample windows viewed with virtual microscope by expert

- Metadata tagged for each grain: taxon, ID confidence, coordinates, radius, slide info

EXPERT ANNOTATIONS



One of 41 planes of view

SEGMENTATION & CLASSIFICATION

Annotated images divided into a training set (70%) and a validation set (30%)

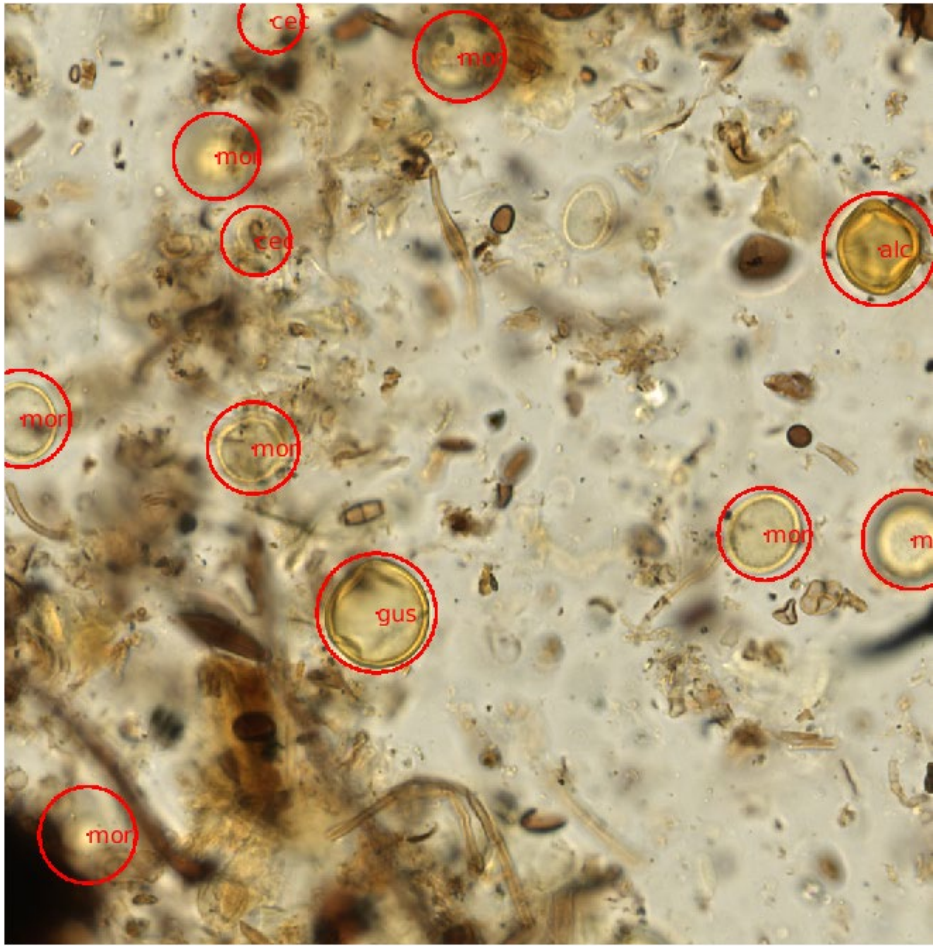
Data augmentation by randomly flipping and rotating images (~2,000 images per pollen type)

Fine-tune pre-trained convolutional neural net (CNN) (Alexnet) (Krizhevsky et al, 2012)

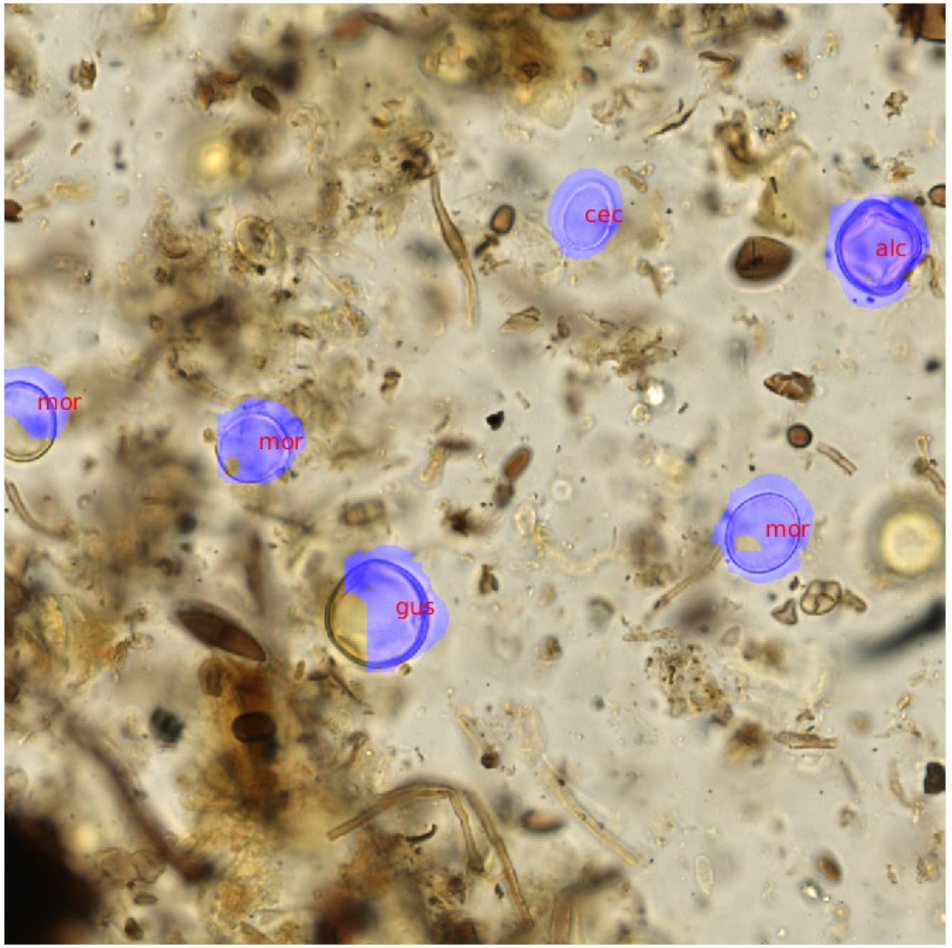
Utilize the radius information to train segmentation masks

Non-maximum suppression used to identify pollen grains

SEGMENTATION & CLASSIFICATION

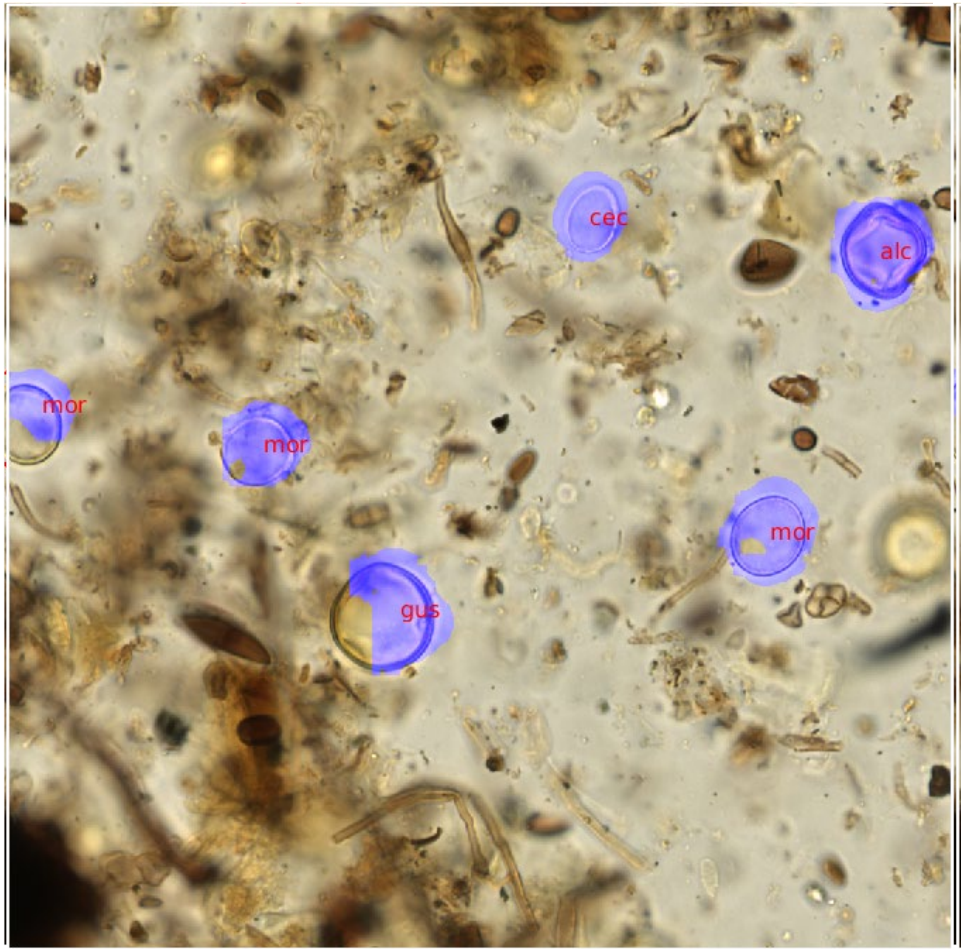


Human

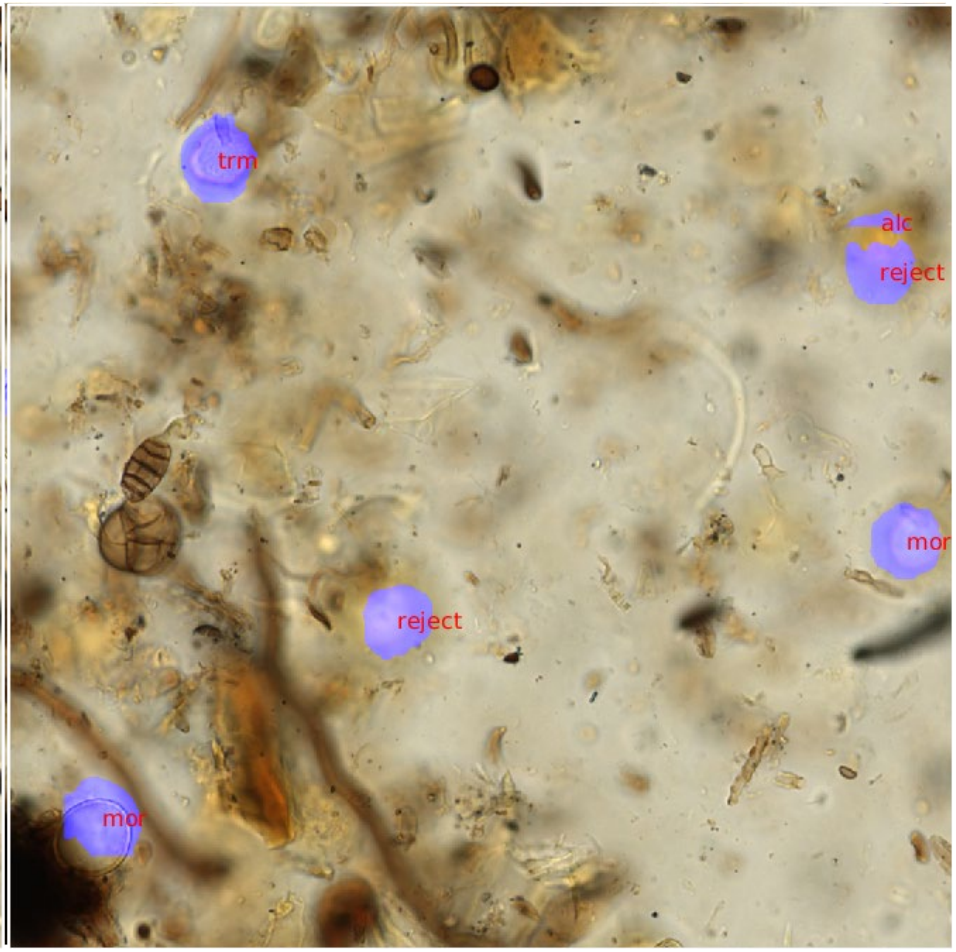


Machine

SEGMENTATION & CLASSIFICATION



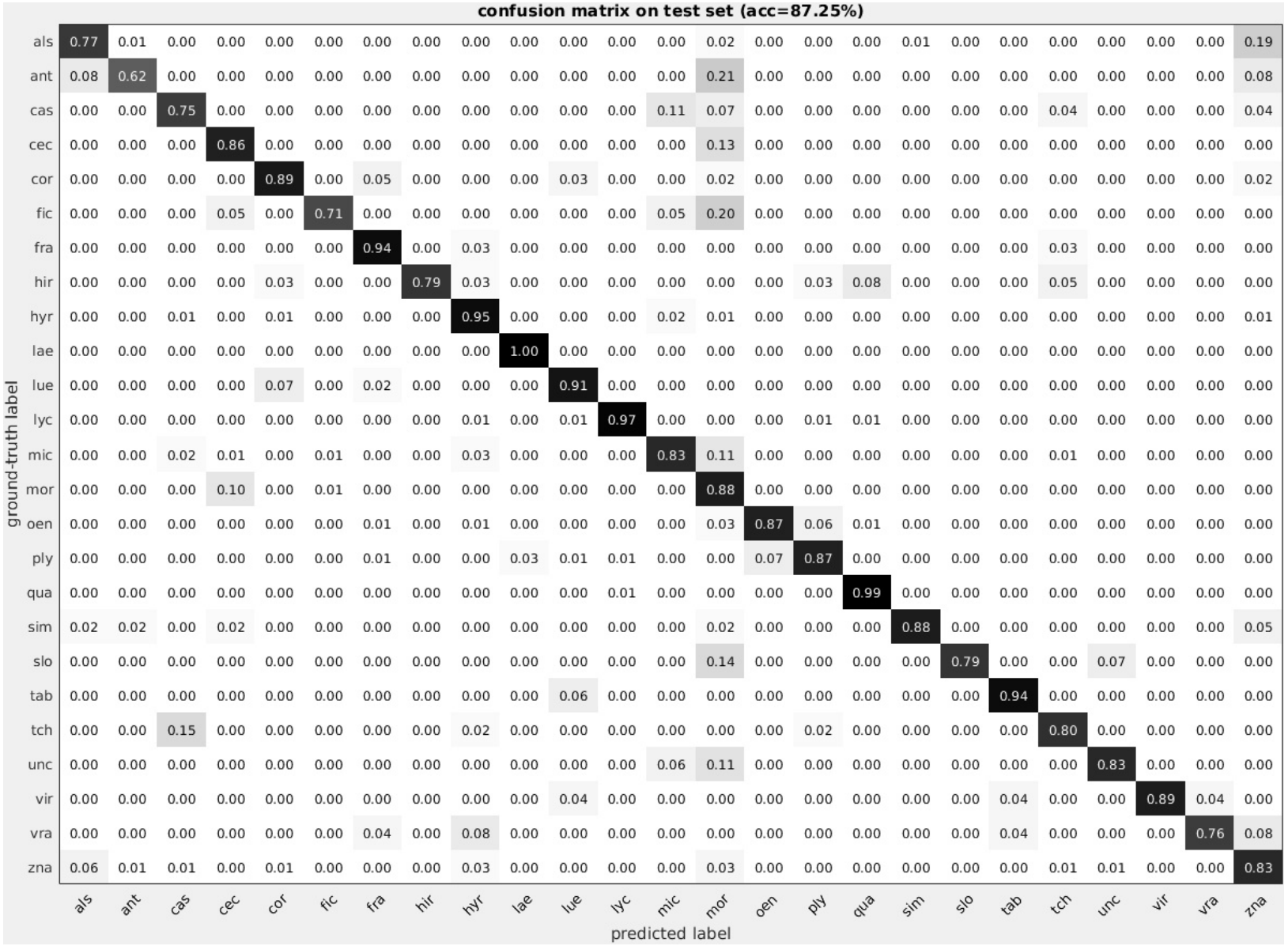
Machine (plane 0)



Machine (plane +20)

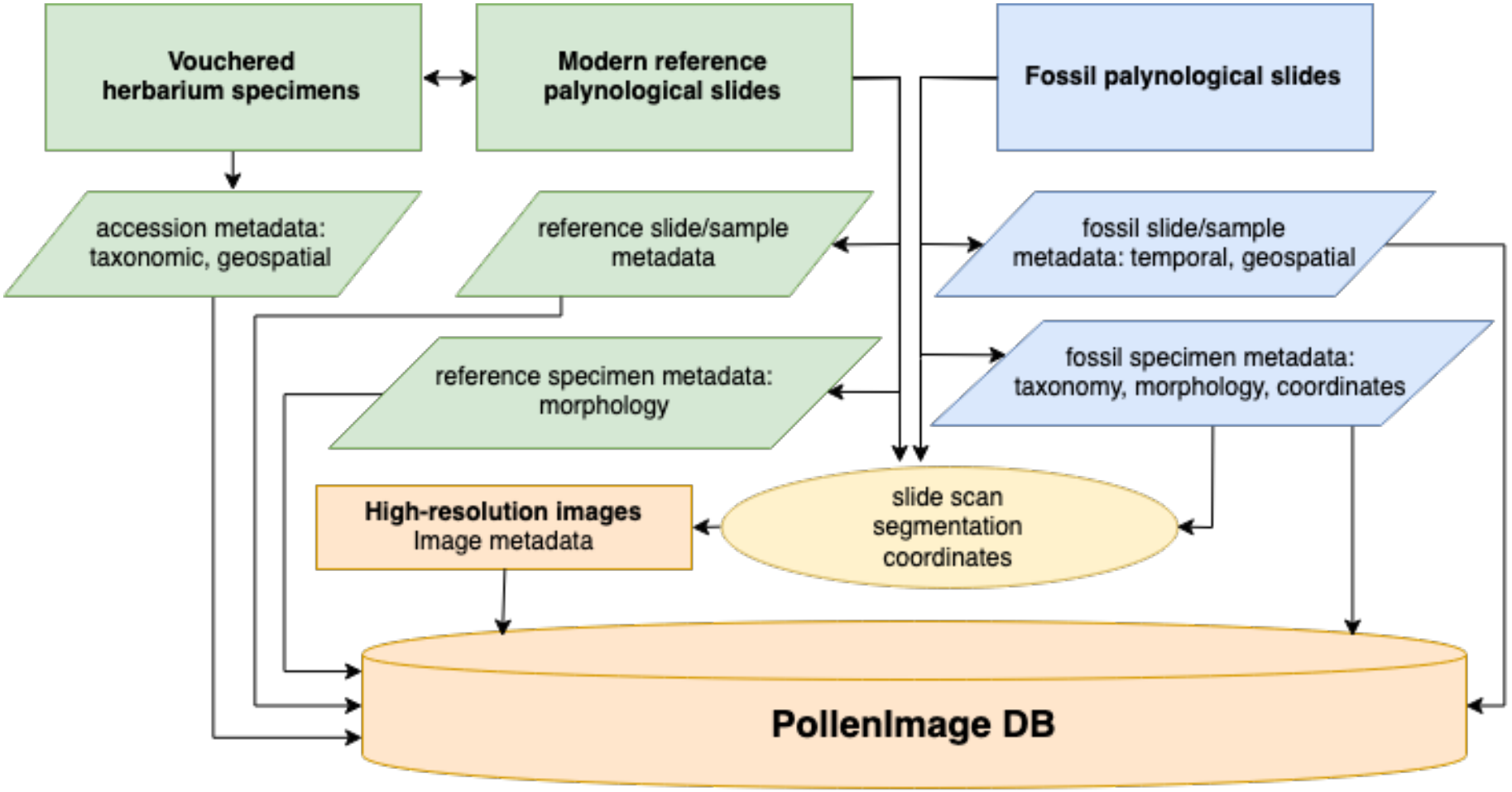
CLASSIFICATION RESULTS

87.25% accuracy for 25 most common/distinctive types



the next step

WORKFLOW FOR LARGE-SCALE IMAGING AND ANALYSIS



PARTNERSHIP WITH KEW AND SMITHSONIAN

Alex Antonelli
(Kew)

Carlos Jaramillo
(Smithsonian Tropical
Research Institute)

PARTNERSHIP WITH KEW AND SMITHSONIAN

Alex Antonelli
(Kew)

<u>Kew Royal Botanic Gardens</u>

- 40,000 slides/pollen residues
- ~7,000,000 herbarium specimens
  (unique species and cultivars)

PARTNERSHIP WITH KEW AND SMITHSONIAN

Smithsonian Tropical Research Institute

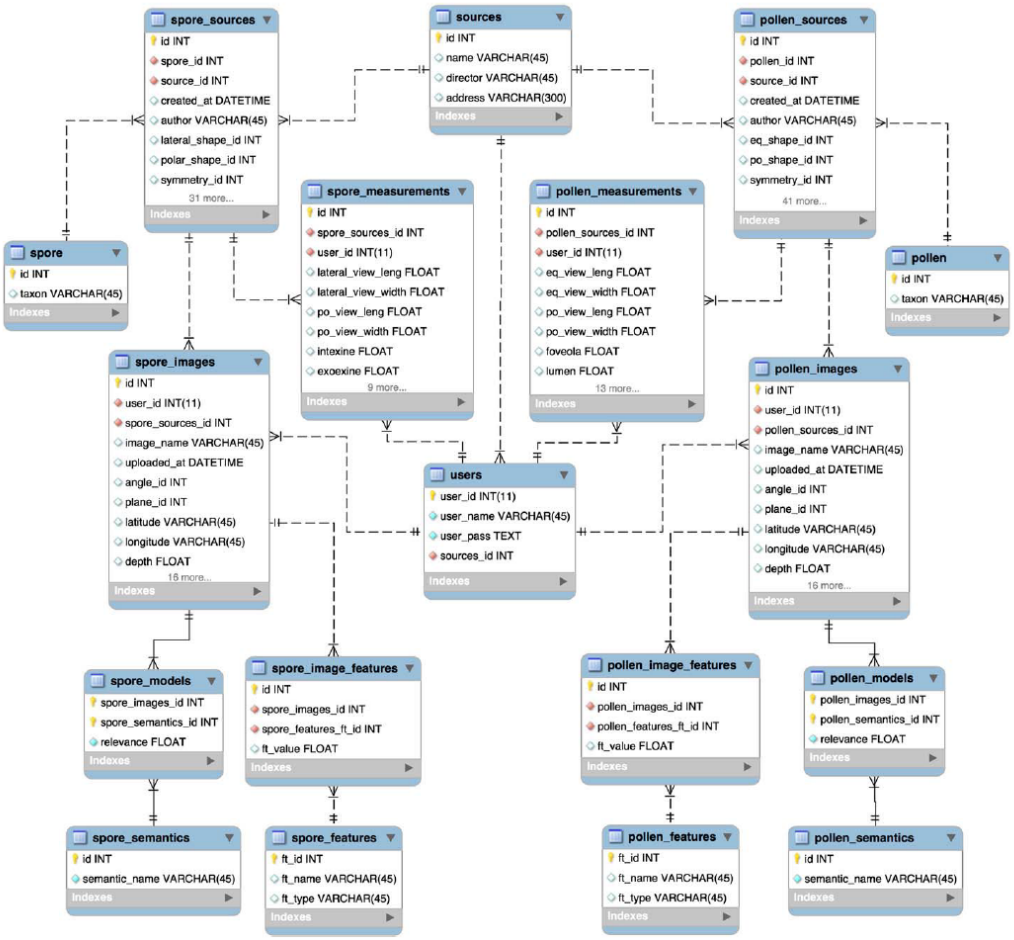
- 25,000 species, Alan Graham pollen reference collection (historic slides)
- Smithsonian Pollen Database

Carlos Jaramillo
(Smithsonian Tropical Research Institute)

PARTNERSHIP WITH KEW AND SMITHSONIAN

Smithsonian Pollen Database

Fossil specimens and morphological descriptions



Han et al., *Applications in Plant Sciences* (2014)

PARTNERSHIP WITH NCSA

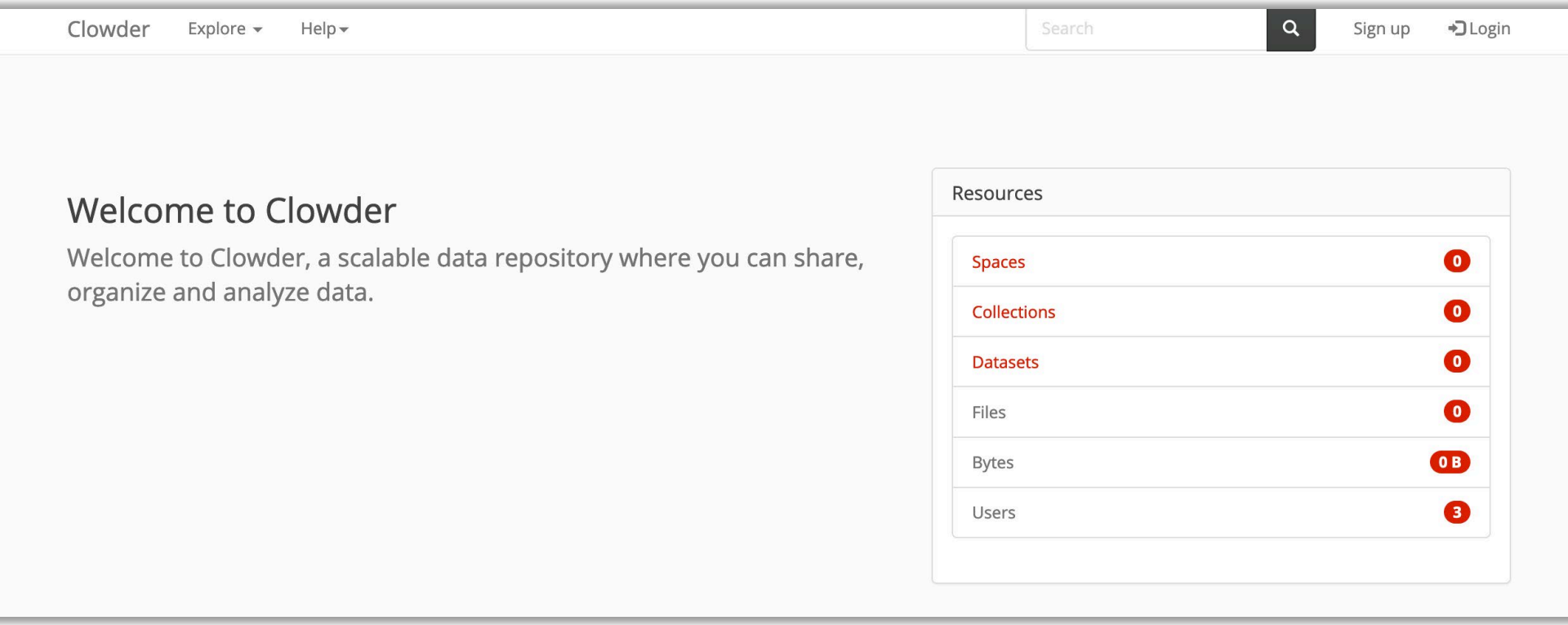
Kenton McHenry
(Principal Research Scientist)

Luigi Marini
(Senior Programmer)

Rob Kooper
(Senior Programmer)

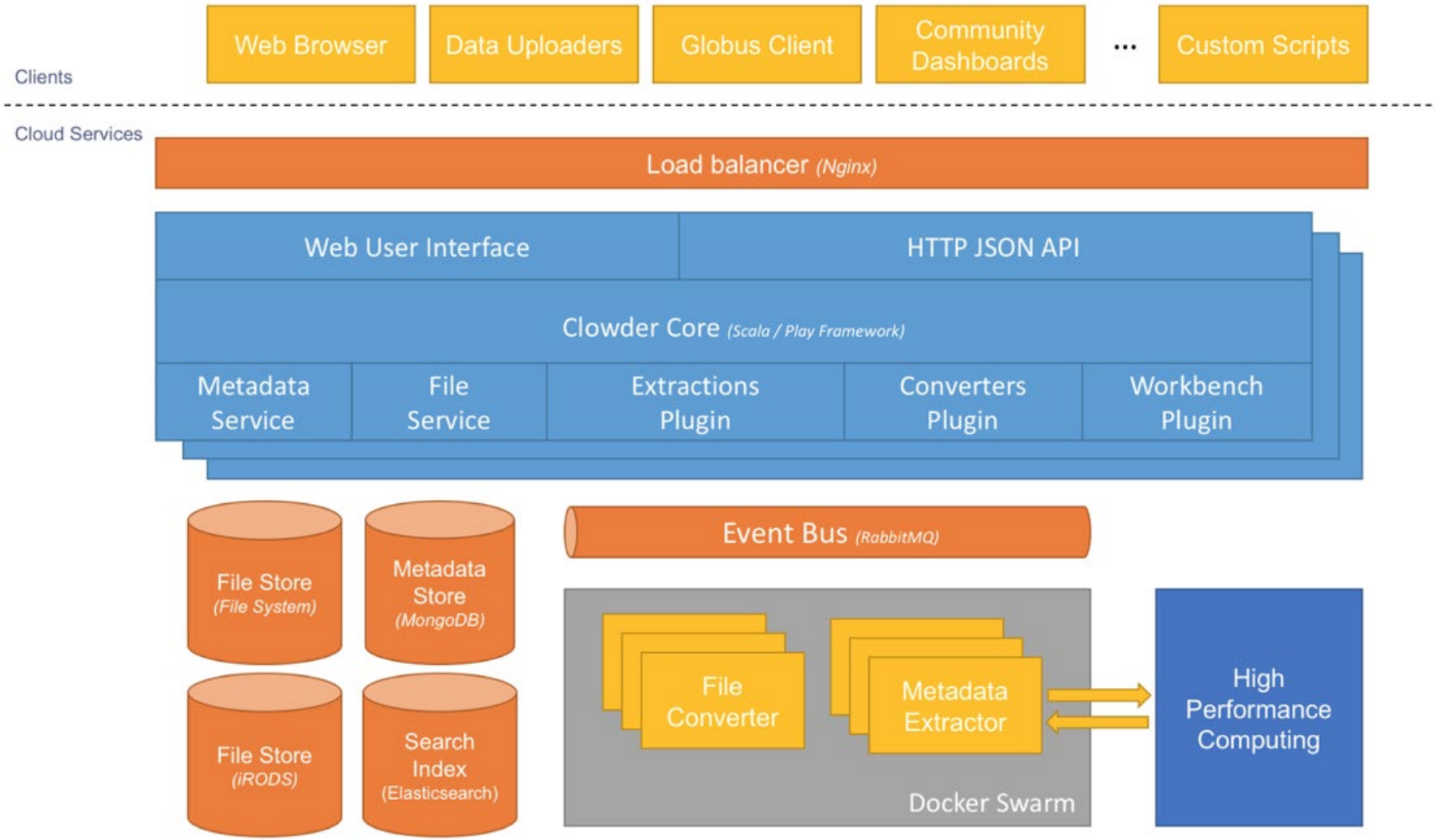
CLOWDER (NCSA)

customizable web interface for data ingestion and analysis



CLOWDER (NCSA)

customizable web interface for data ingestion and analysis



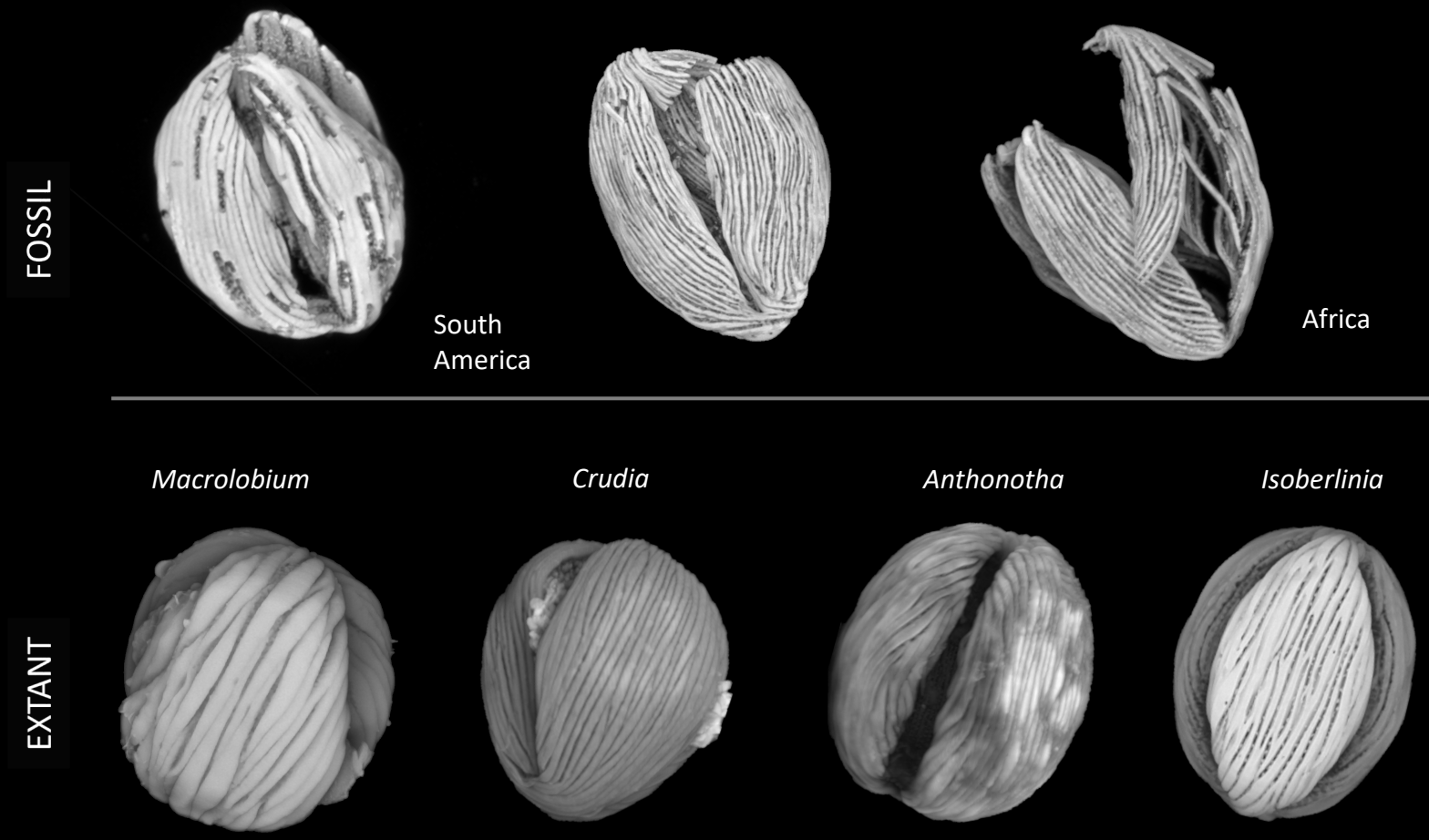
other workflows

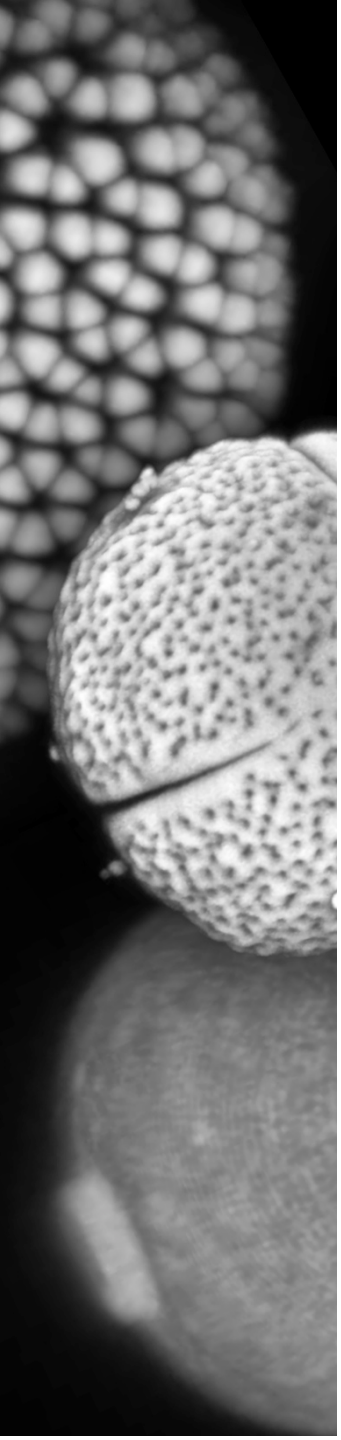
Ingrid Romero
(U Illinois)

Charless Fowlkes
(UC Irvine)

Shu Kong
(UC Irvine)

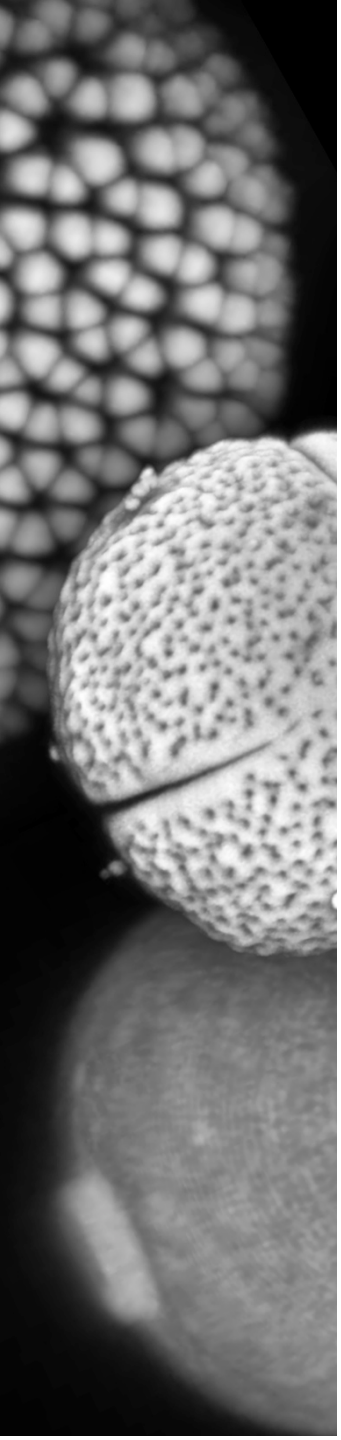
Putative Detarioideae legume – *Striatopollis catatumbus*





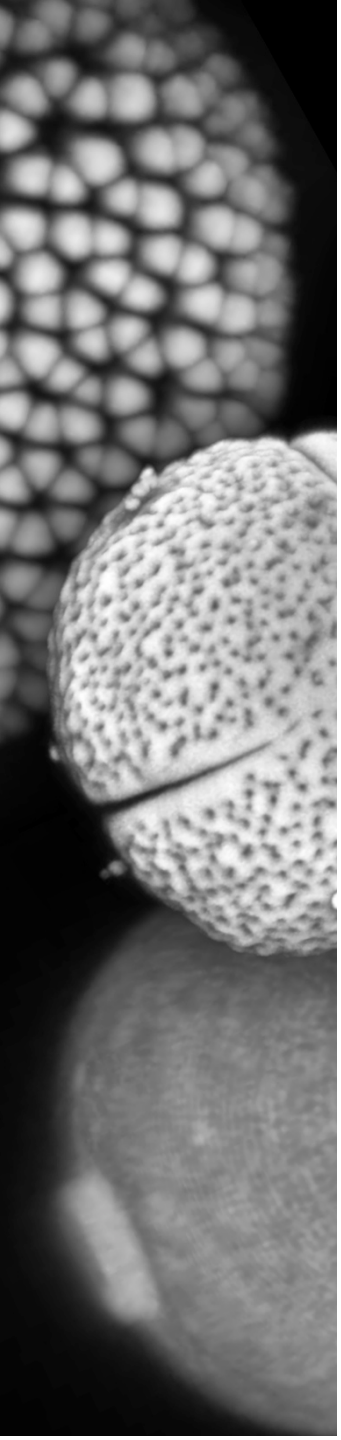
CONCLUSIONS

- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!



CONCLUSIONS

- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!



CONCLUSIONS

- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!

Much of the necessary computational infrastructure already exists

Ultimate bottleneck is in
the availability of
well-curated, high-resolution images

COLLABORATORS AND FUNDERS

David Tcheng
Illinois Informatics Institute, NCSA

Glenn Fried & Mayandi Sivaguru
Institute for Genomic Biology, UIUC

Enrique Moreno
Smithsonian Tropical Research Institute

NSF-DBI – Advances in Biological Informatics
NSF-DBI – Innovations in Biological Imaging & Visualization
NSF-EF – Macrosystems Biology
National Center for Supercomputing Applications (NCSA)
U Illinois Campus Research Board