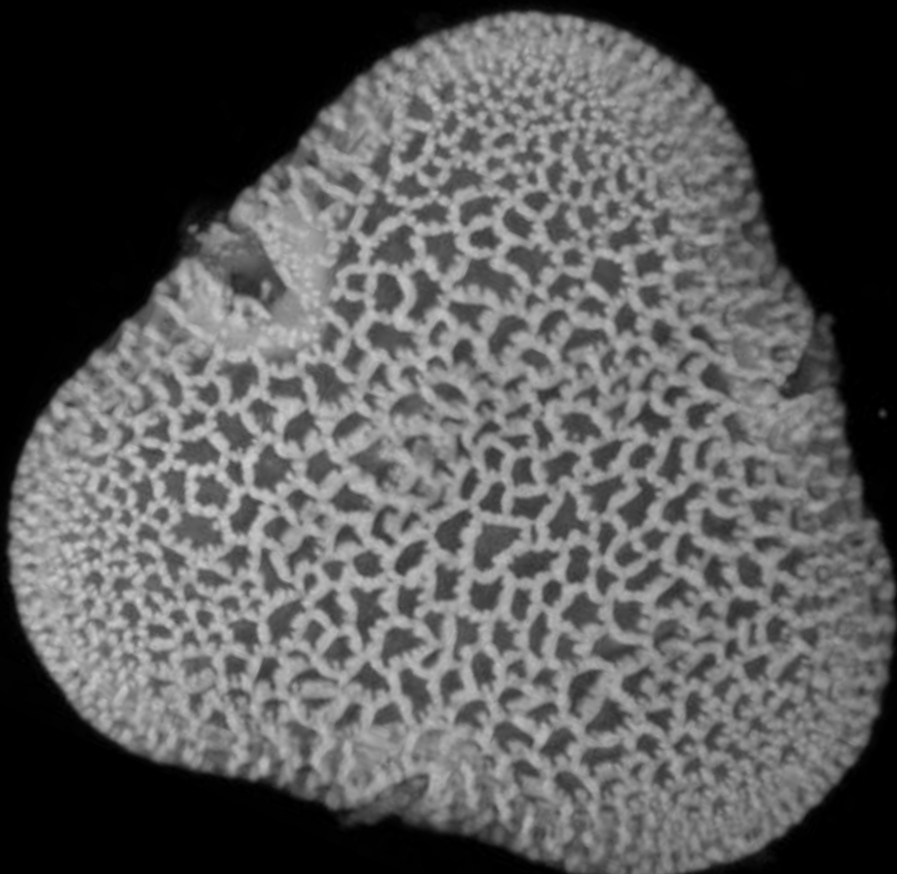
Digital palynology: imaging, automation, and intelligent databases



Surangi W. Punyasena

University of Illinois, Urbana-Champaign
Plant Biology, Geology, Geography, Informatics

punyasena@life.illinois.edu

Digitization and databasing effort is significant.
Databases require active curation.

Existing image databases (exclusively?)
employ text-based search
based on manual, expert tags

Are there alternative models for approaching visual specimen data?

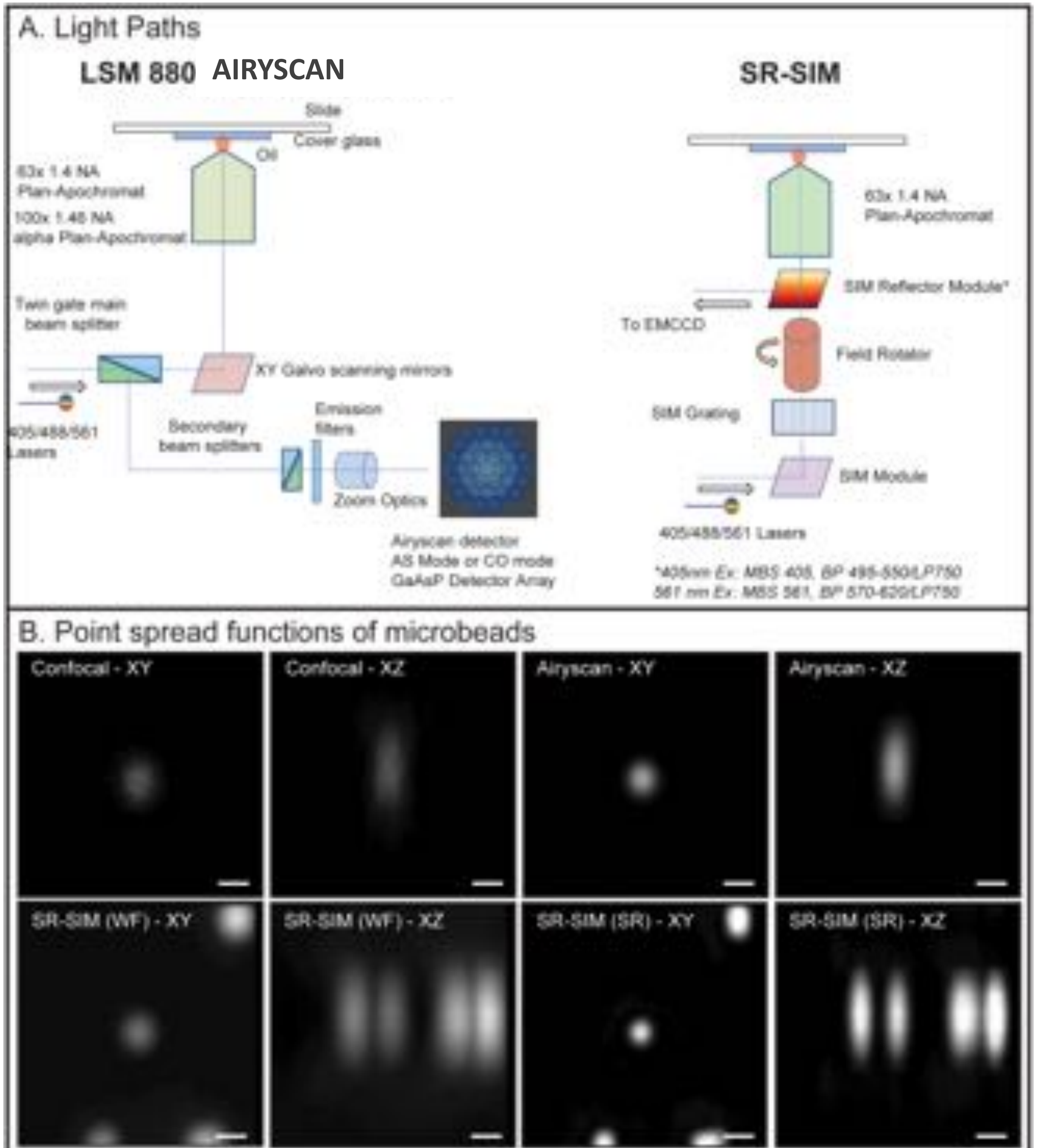
HANDLING VISUAL (MICROSCOPIC) DATA

- Increase throughput of microscopic imaging
  - Adapting existing slide scanning microscopes to pollen specimen slides
  - Purpose-built software to work with microscope APIs/image documentation
- Improve resolution of imaging
  - Archival images must capture the same visual information evident on a (diffraction-limited) microscope or better
- Automate “knowledge capture”
  - Archives must efficiently record and efficiently transfer expert knowledge
  - Utilize computer vision/machine learning

HANDLING VISUAL (MICROSCOPIC) DATA

- Increase throughput of microscopic imaging
  - Adapting existing slide scanning microscopes to pollen specimen slides
  - Purpose-built software to work with microscope APIs/image documentation
- Improve resolution of imaging
  - Archival images must capture the same visual information evident on a (diffraction-limited) microscope or better
- Automate "knowledge capture"
  - Archives must efficiently record and efficiently transfer expert knowledge
  - Utilize computer vision/machine learning

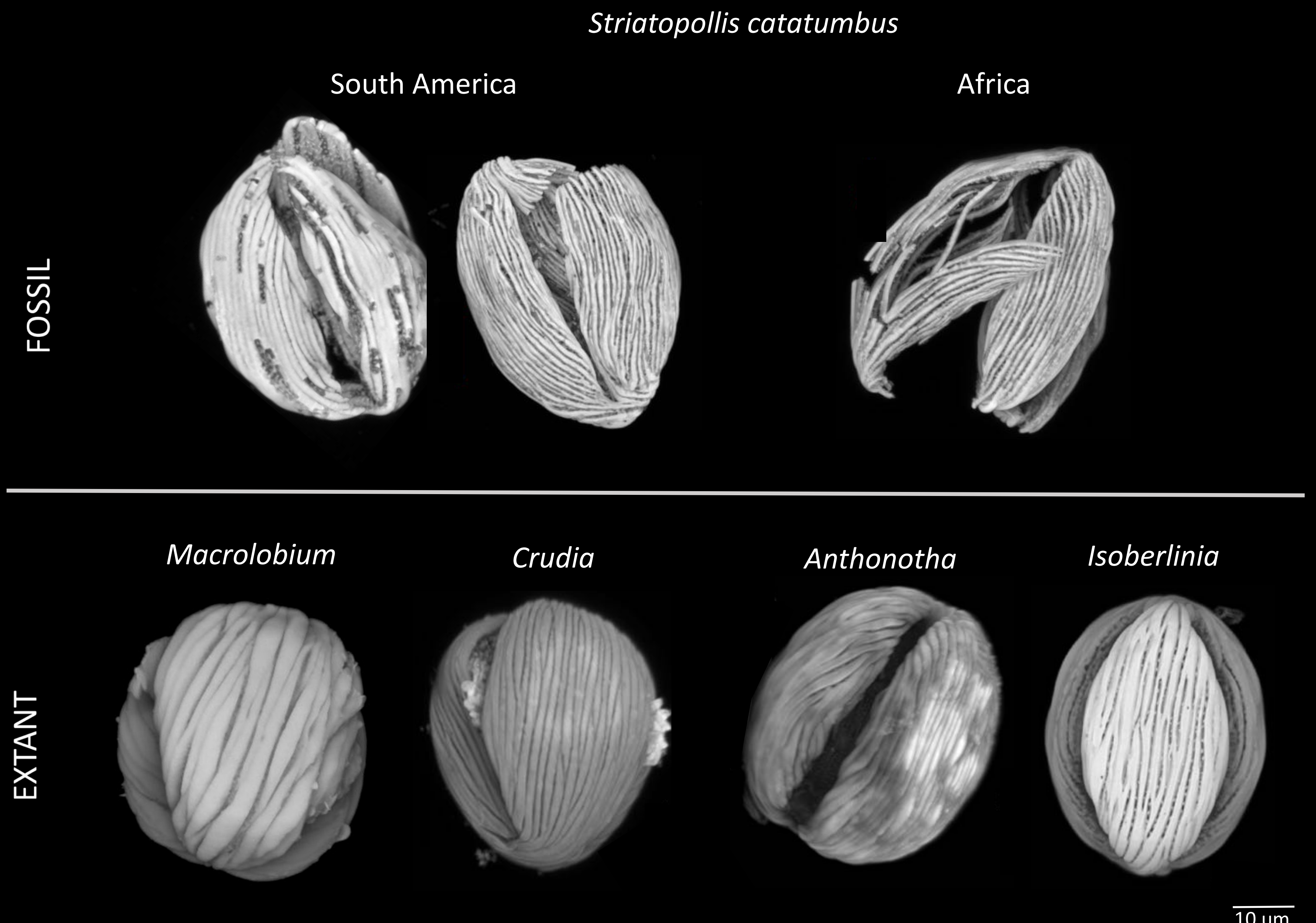
Resolution



Scale bars
represent 0.2 µm

Sivaguru, Urban, Fried, Wesseln, Mander, and Punyasena, *MRT* (2016, in press)

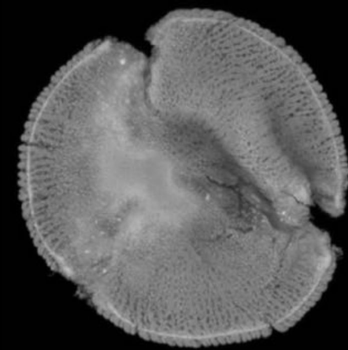
EQUATORIAL VIEW



POLAR VIEW

Imagen
Unavailable

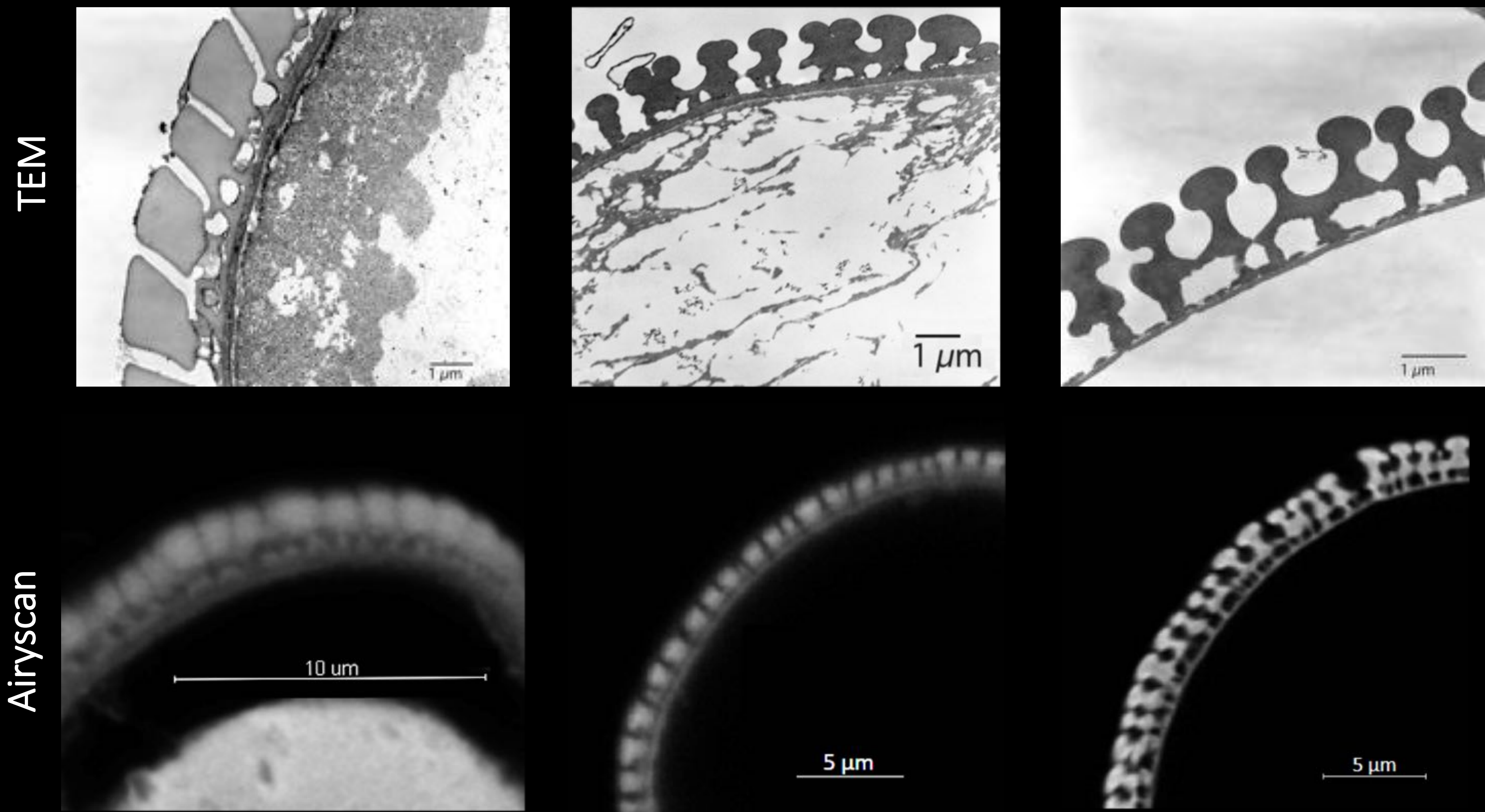
S. catatumbus
(South America)



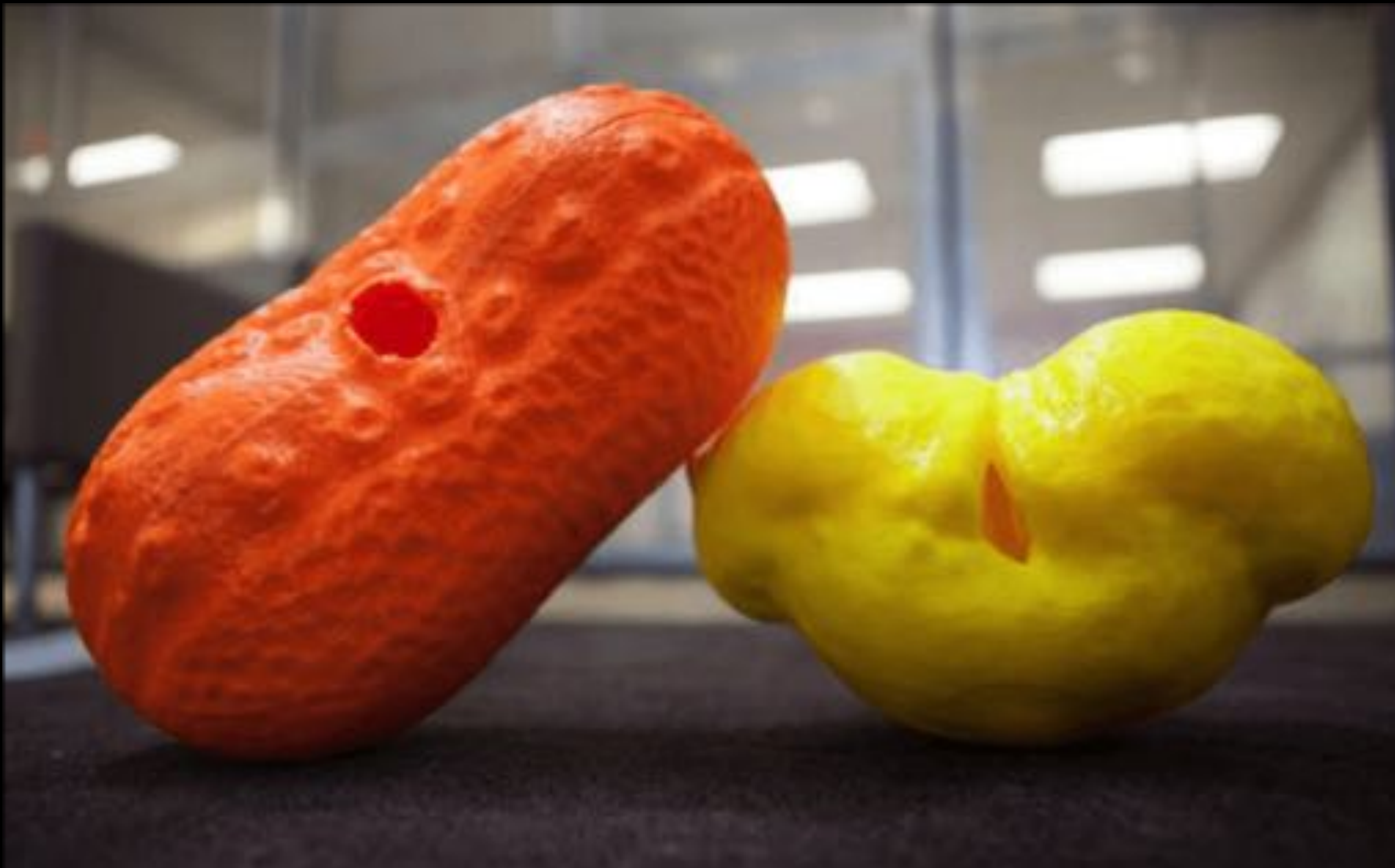
S. catatumbus
(Africa)

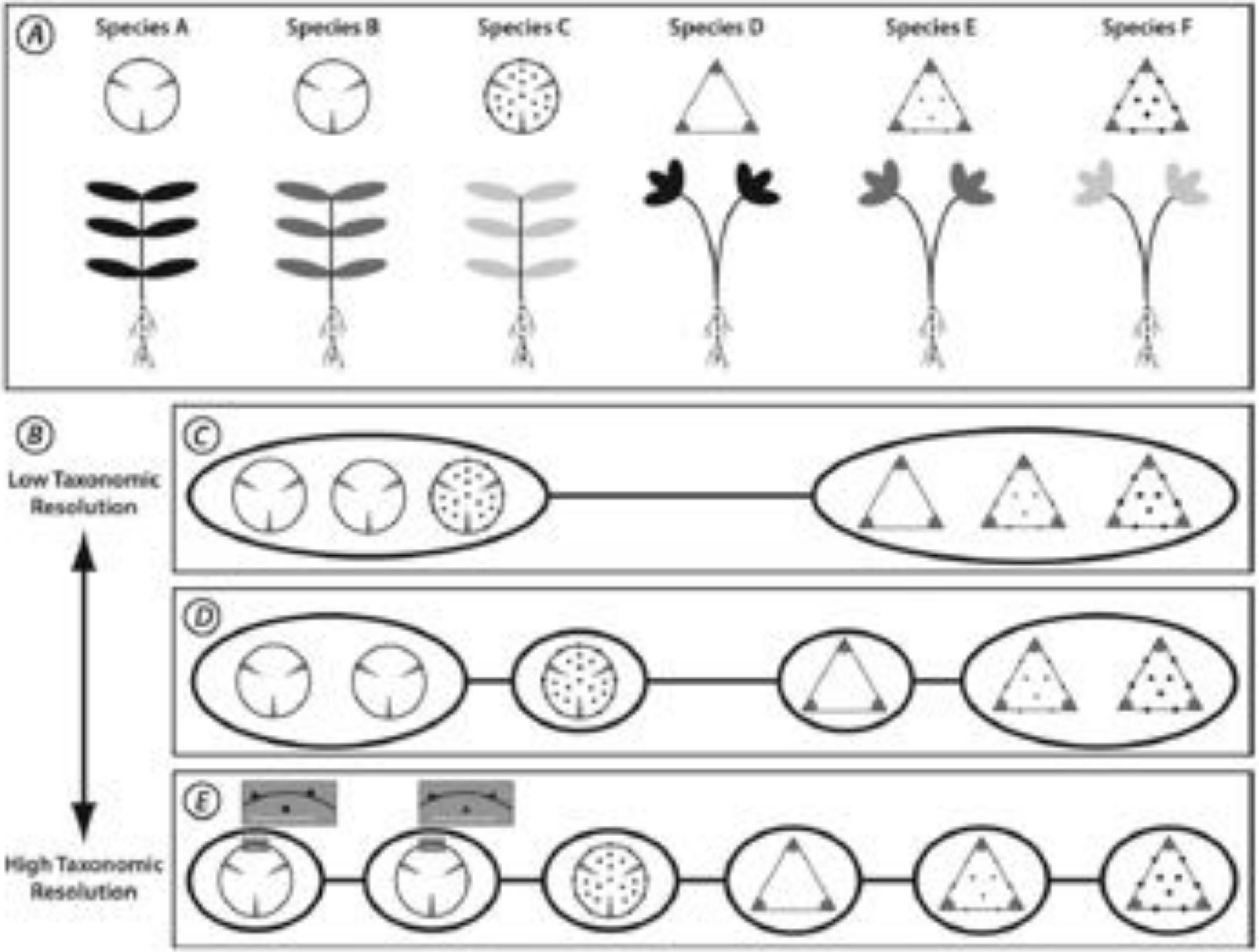
Macrolobium *Crudia* *Anthonotha* *Isoberlinia*

AIRYSCAN SUPERRESOLUTION MICROSCOPY



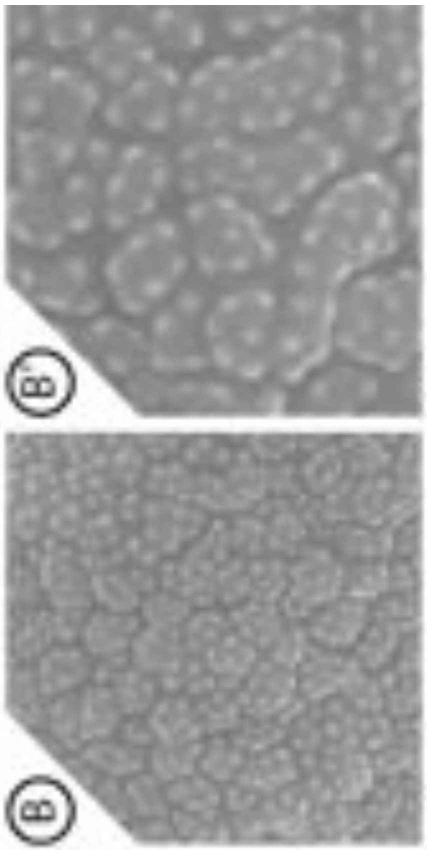
Macrolobium angustifolium *Crudia amazonica* *Anthonotha macrophylla*

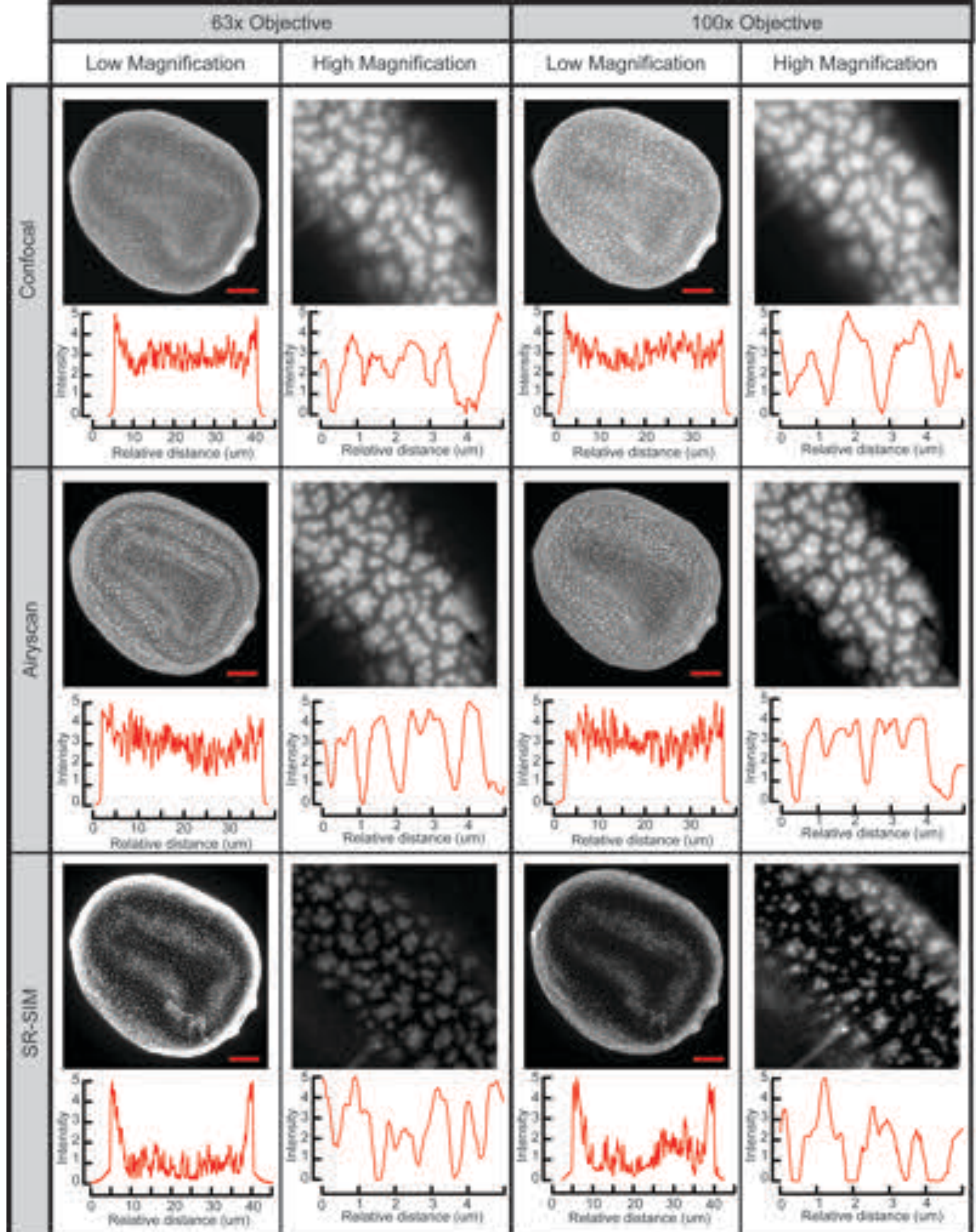




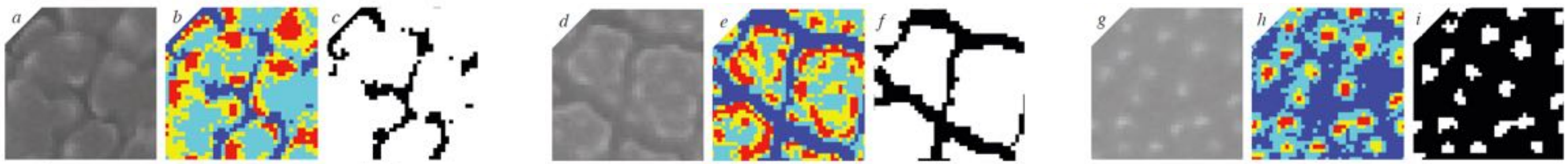
Mander and Punyasena, *IJPS* (2016)

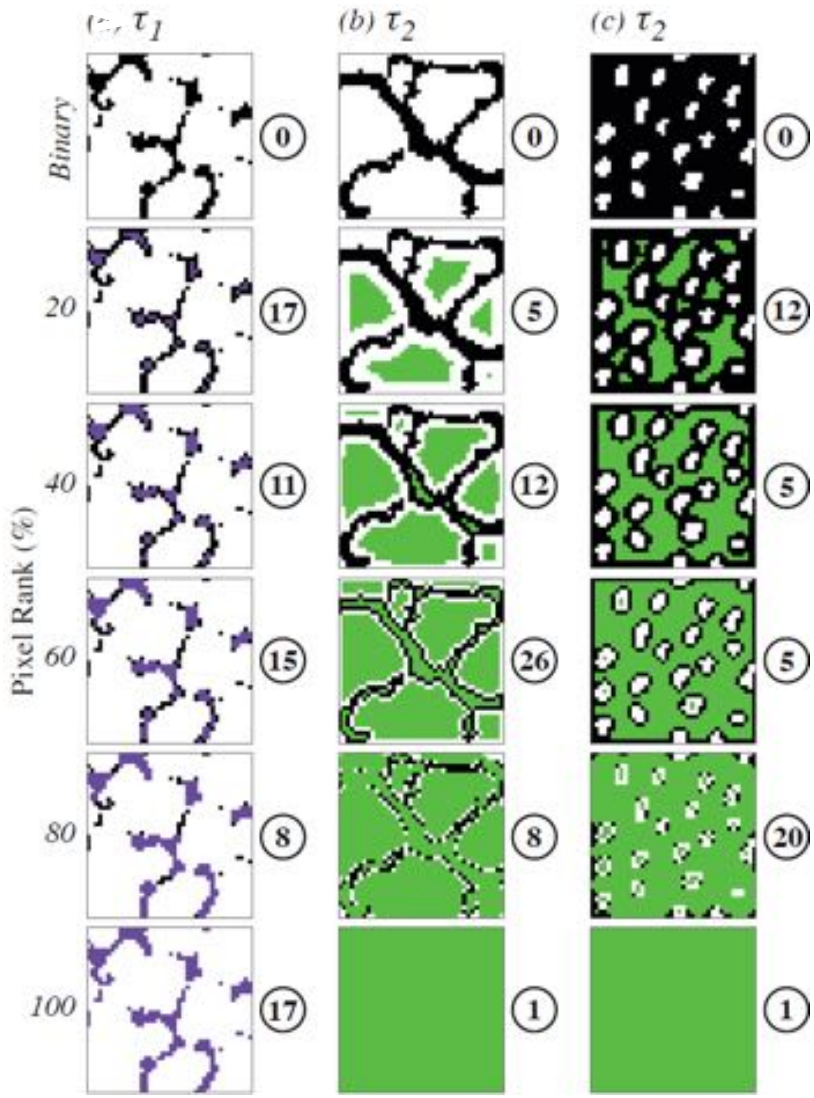
SEM





Sivaguru, Urban, Fried, Wesseln, Mander, and Punyasena, *MRT* (2016, in press)





Mander, Li, Mio, Fowlkes and Punyasena, *Proceedings of the Royal Society B* (2013)

Knowledge Capture

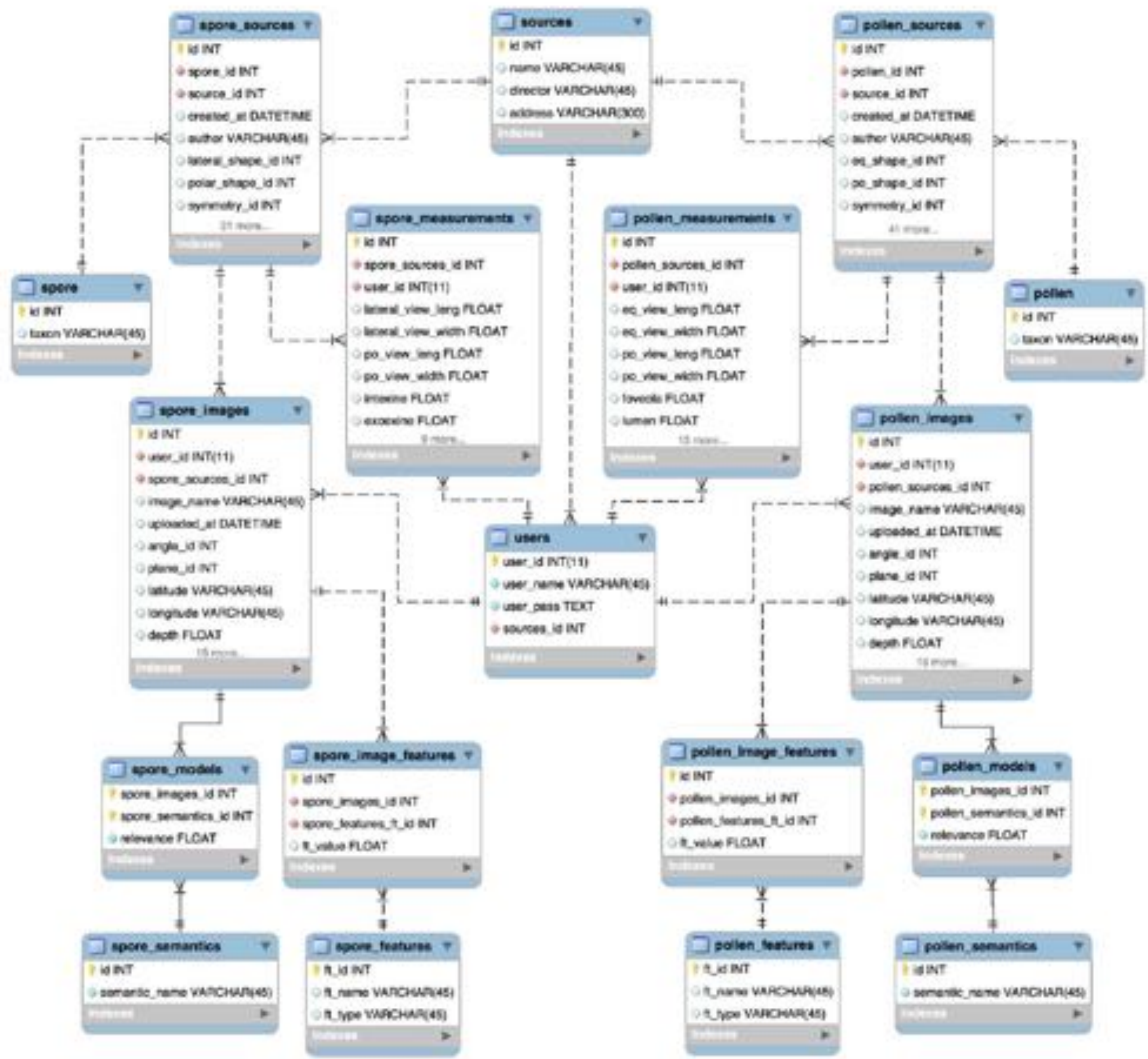
Smart Databases

IMAGE-BASED SEARCH AND SEMANTIC MODELING

http://www.bioshapes.org/mioceneDB

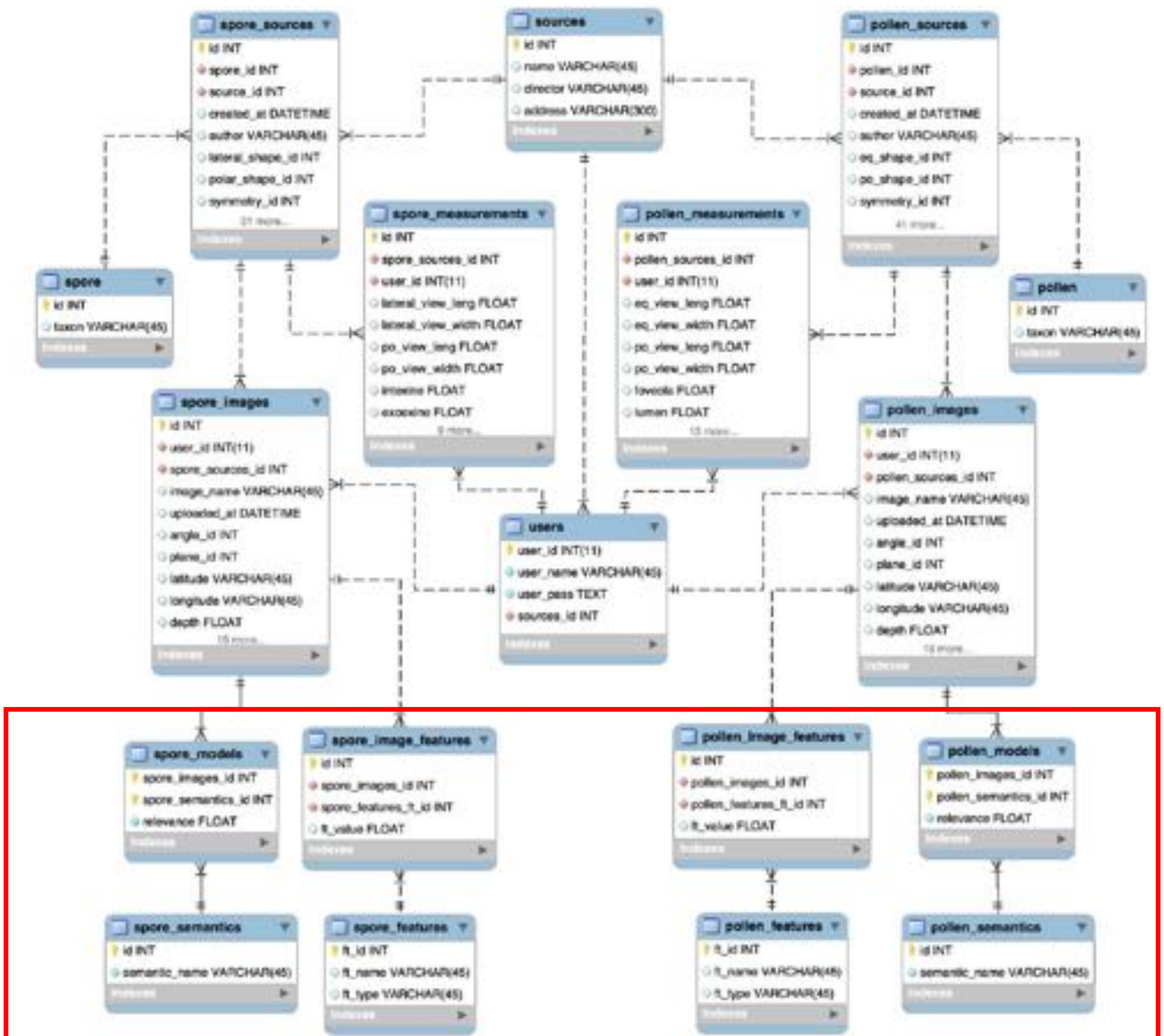
Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

SUSTAINABLE RELATIONAL DB DESIGN



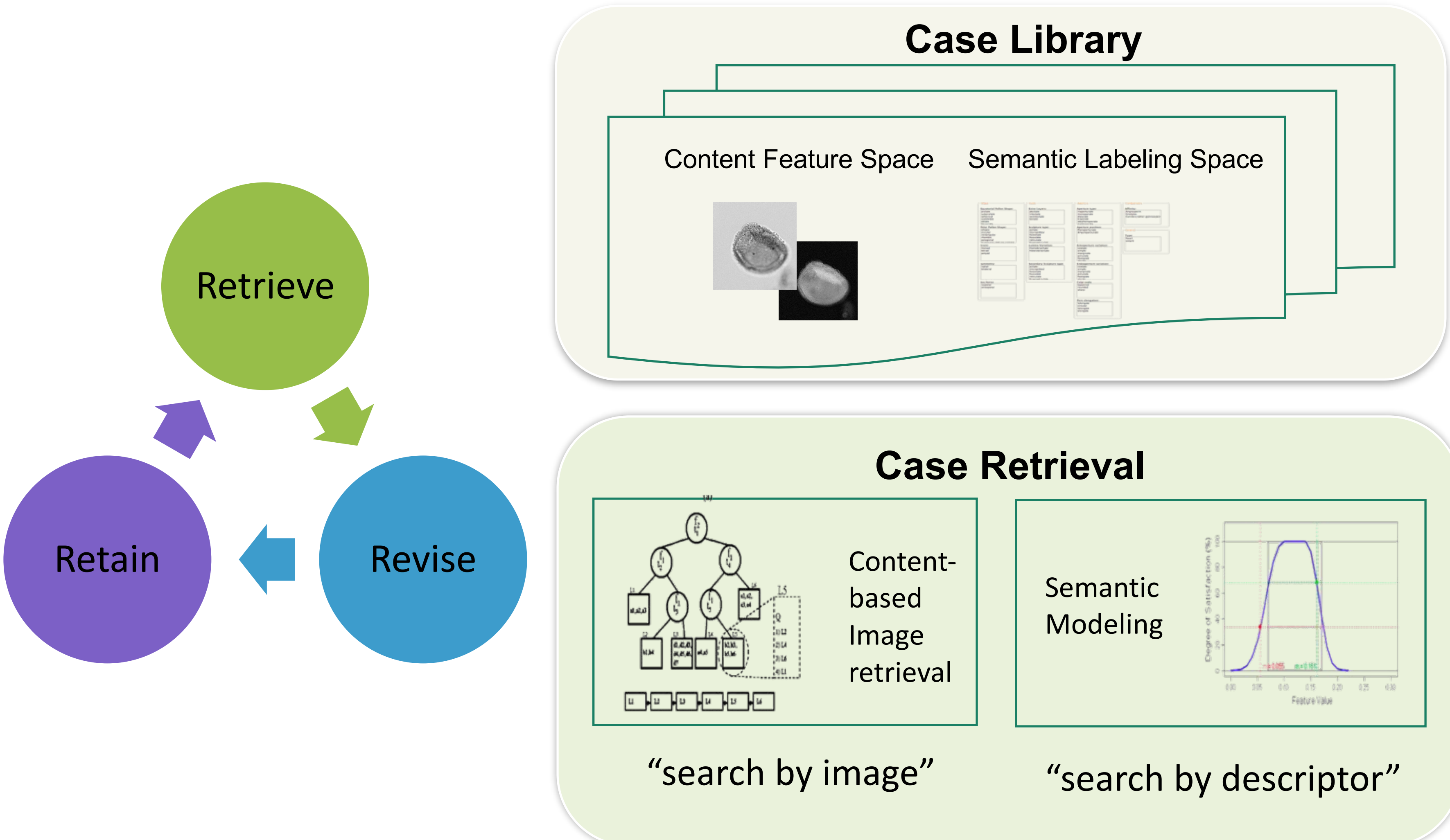
Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

SUSTAINABLE RELATIONAL DB DESIGN



Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

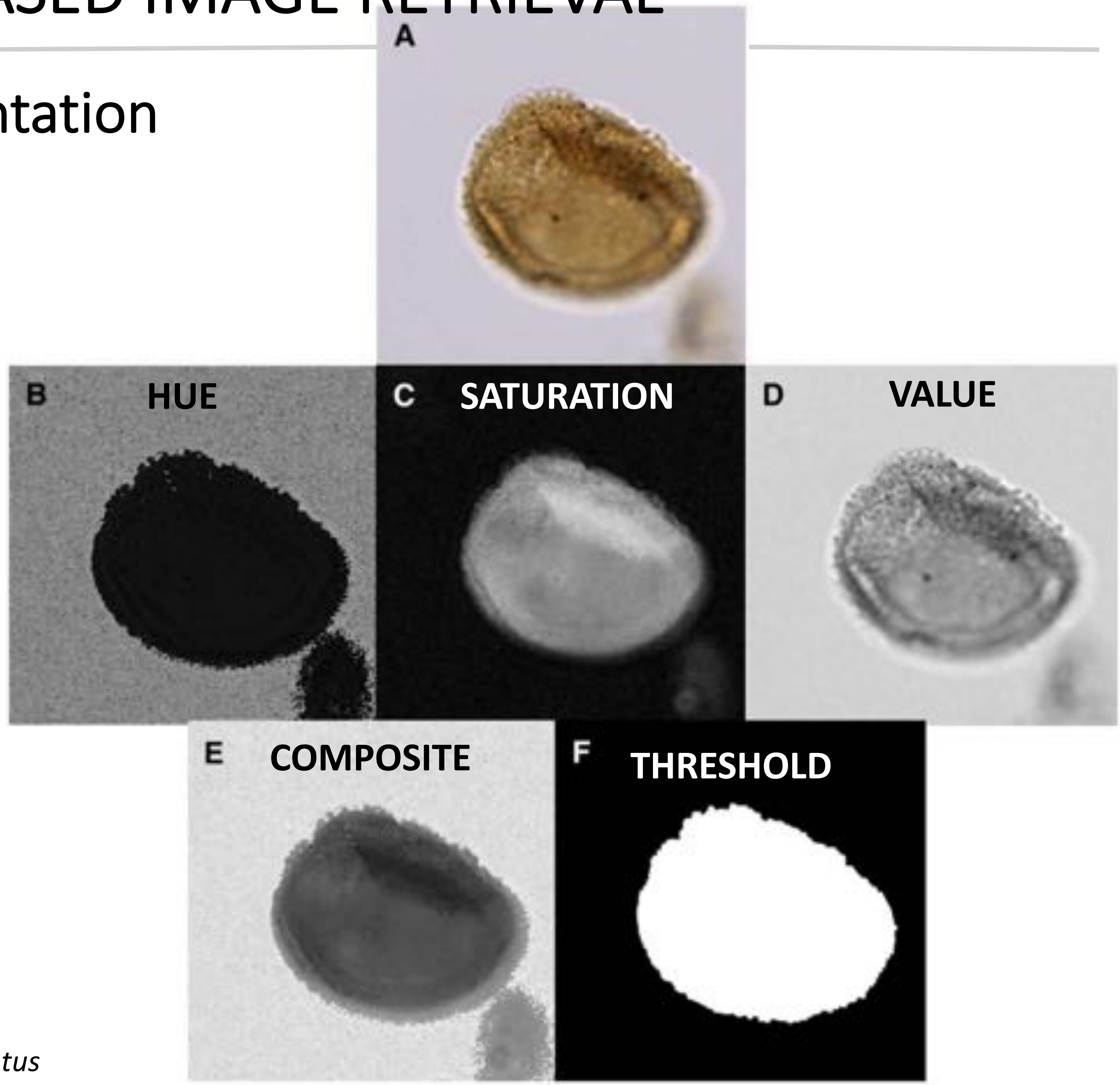
CASE-BASED REASONING (CBR)



Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

CONTENT-BASED IMAGE RETRIEVAL

Image Segmentation

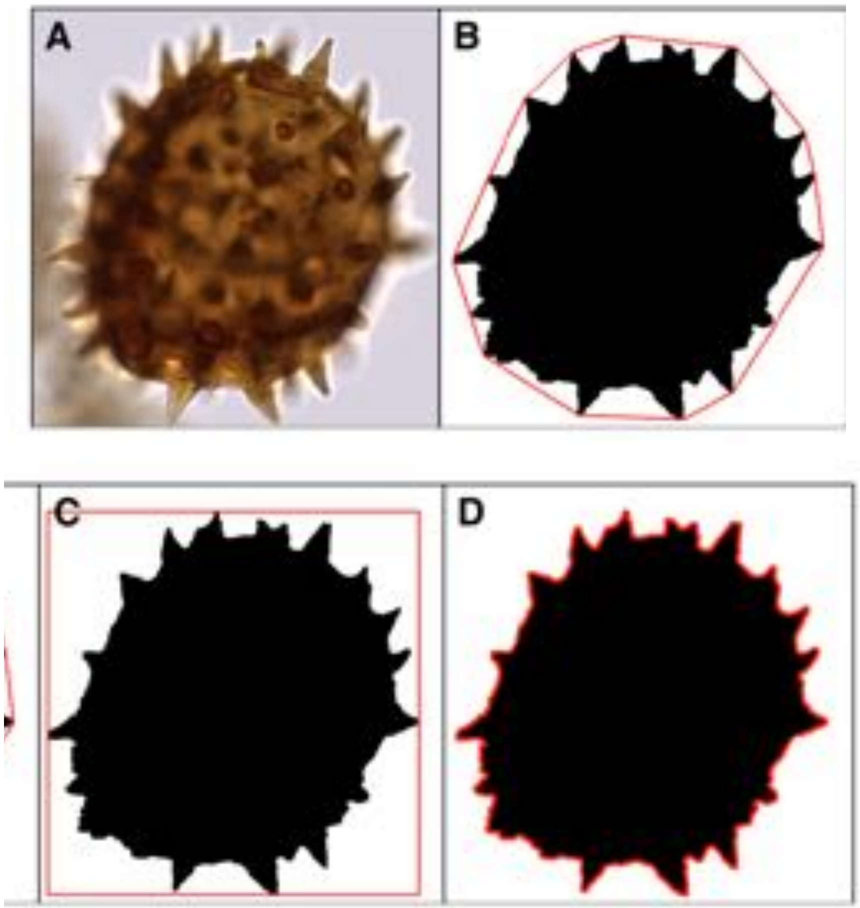


Clavainaperturites microclavatus

Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

CONTENT-BASED IMAGE RETRIEVAL

Image Characterization



Grimsdalea magnaclavata

COLOR

- Otsu threshold
- Mean pixel value
- Standard deviation of pixel value
- 1D histogram (16 bins)

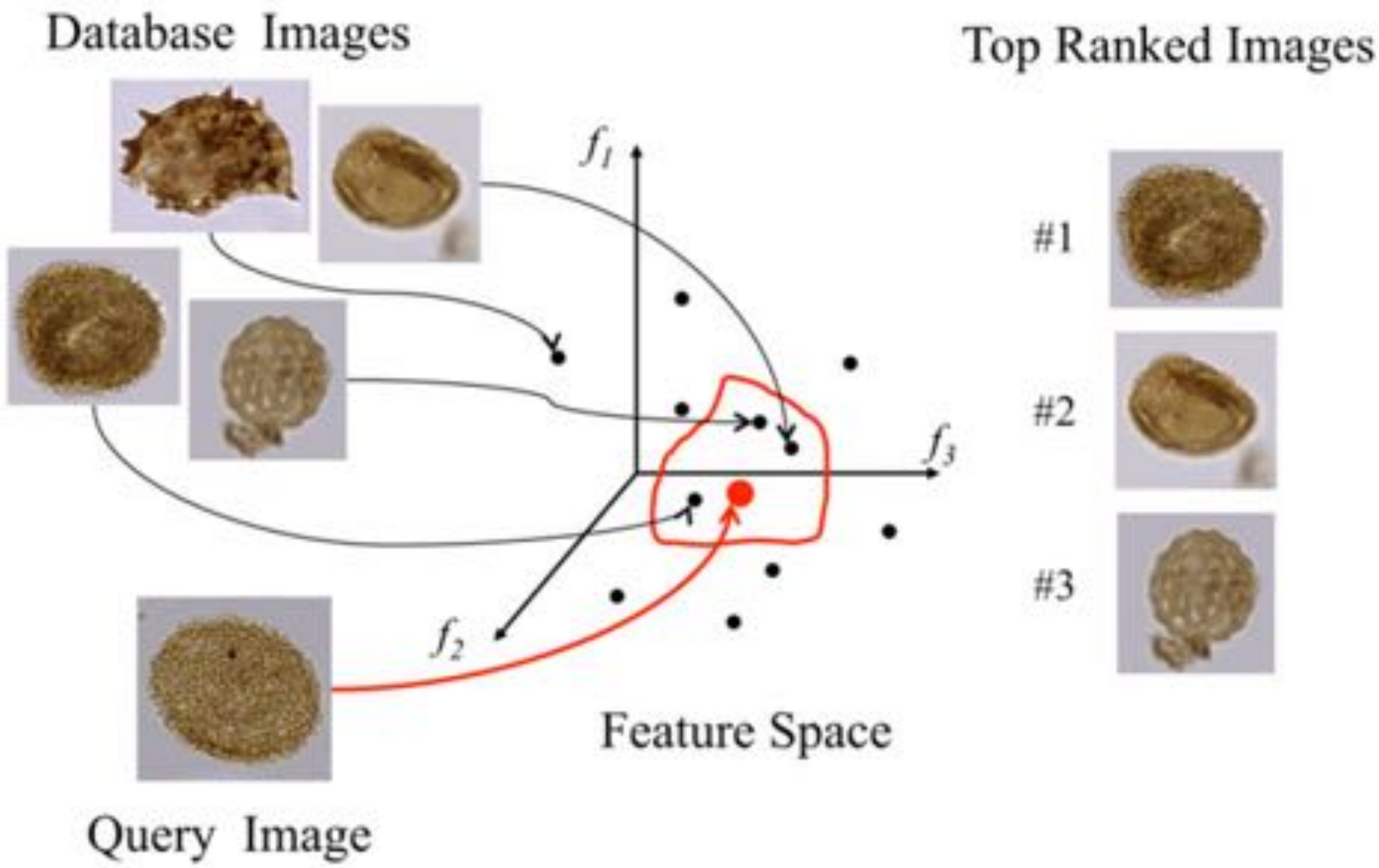
SHAPE

- Grain size
- Hu shape descriptor (7 values)
- Aspect ratio
- Compactness
- Convexity
- Form factor
- Roundness
- Solidity
- Perimeter
- Texture

Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

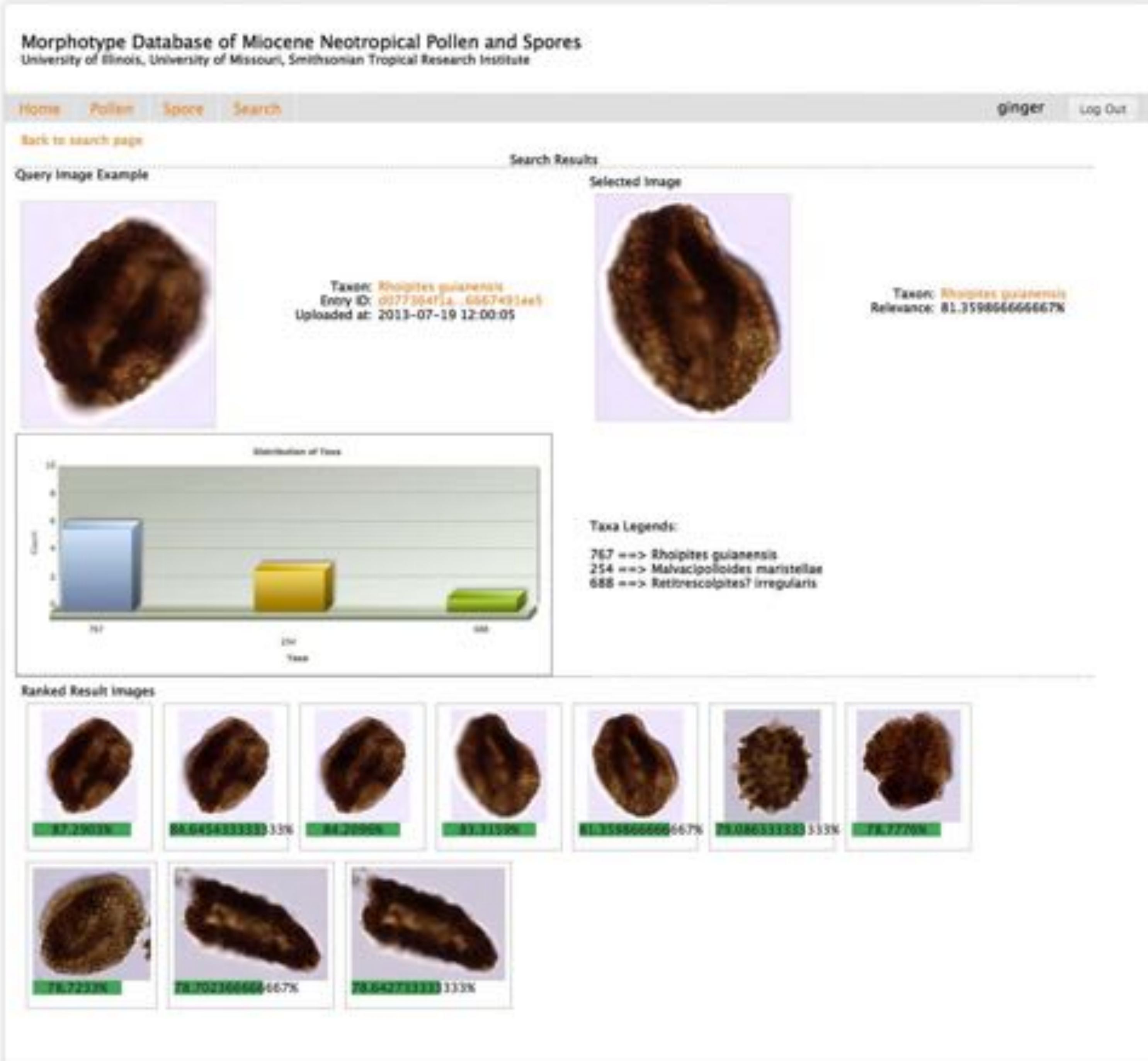
CONTENT-BASED IMAGE RETRIEVAL

Image Search



69 features, 4 channels (hue, saturation, value, grey-scale) = 276 dimensions

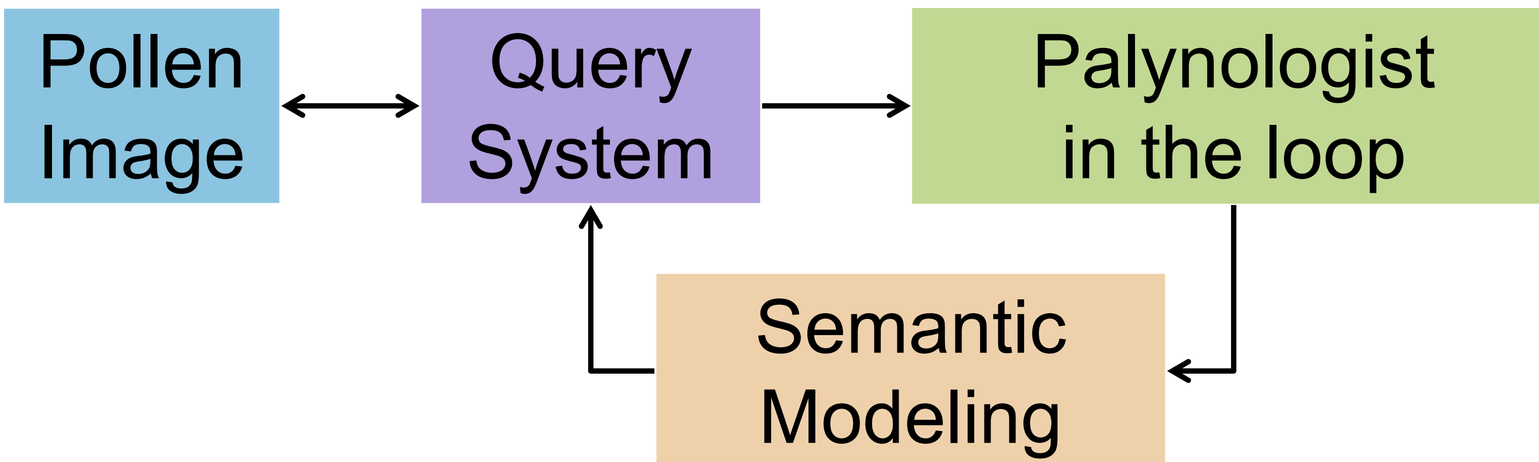
Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)



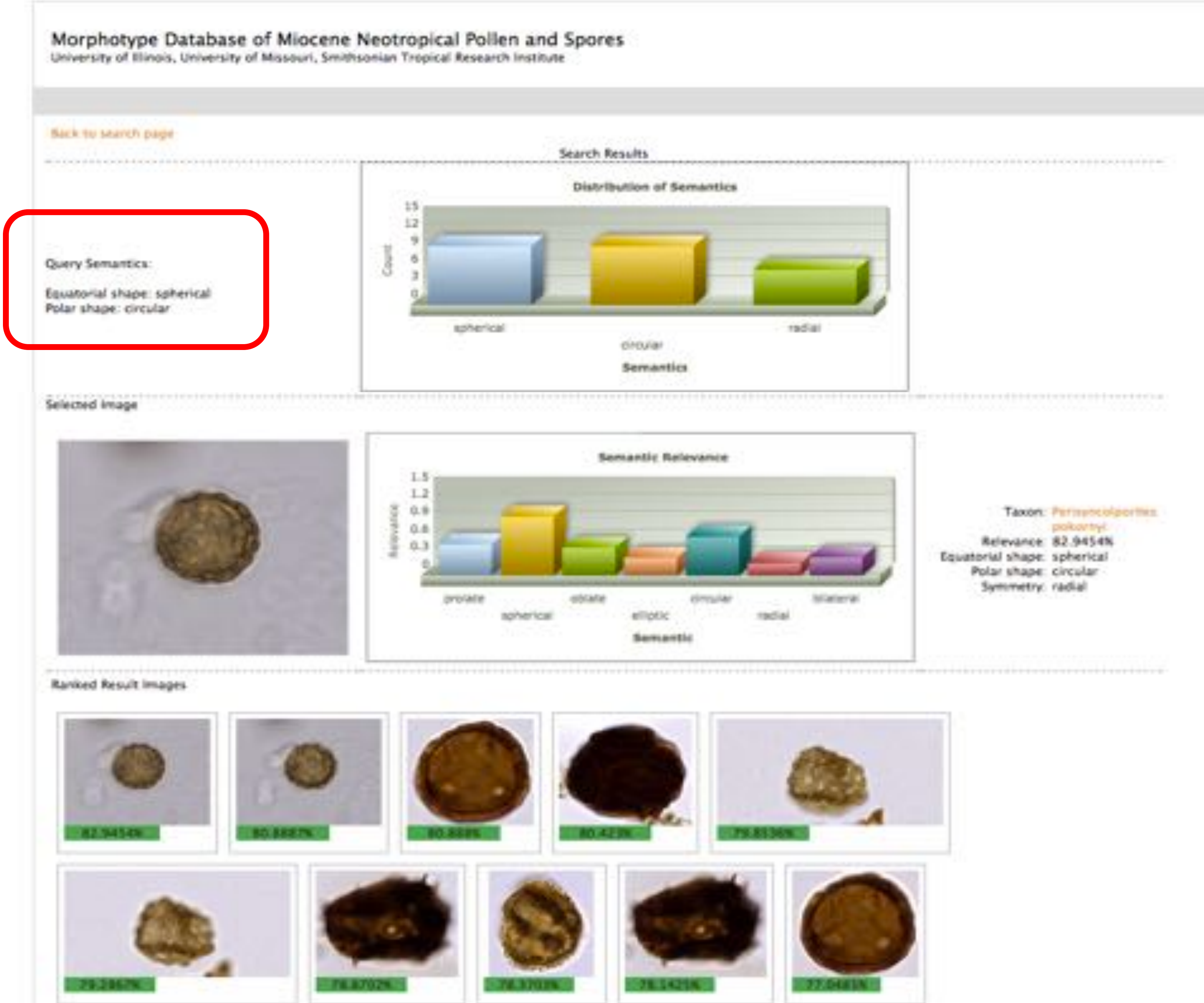
Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

SEMANTIC MODELING

automated feature recognition



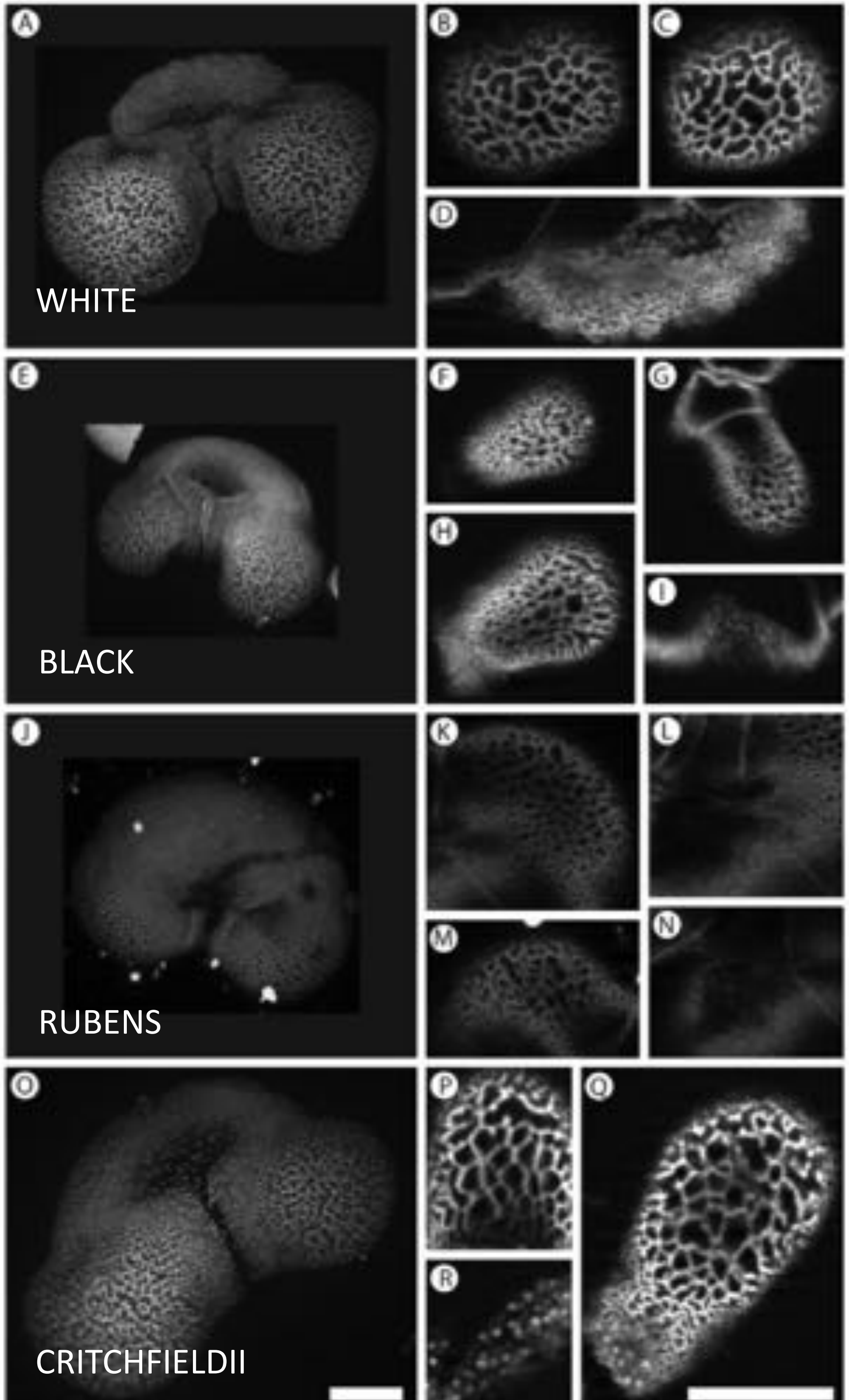
low level features ⟷ high level semantics

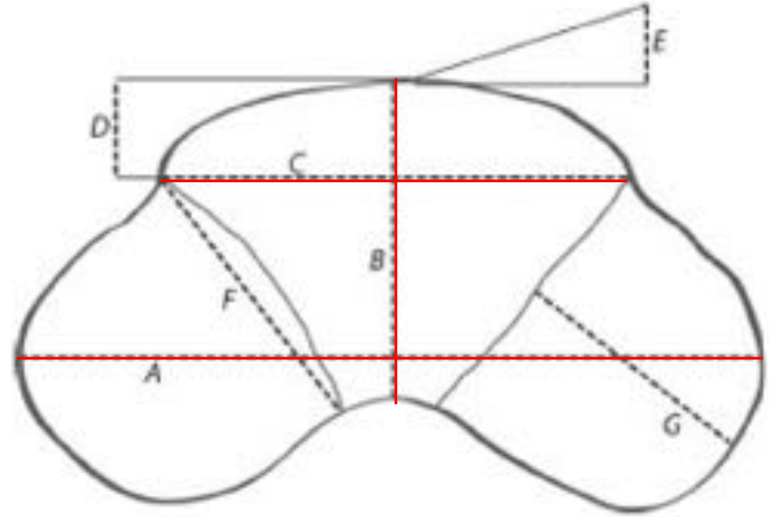


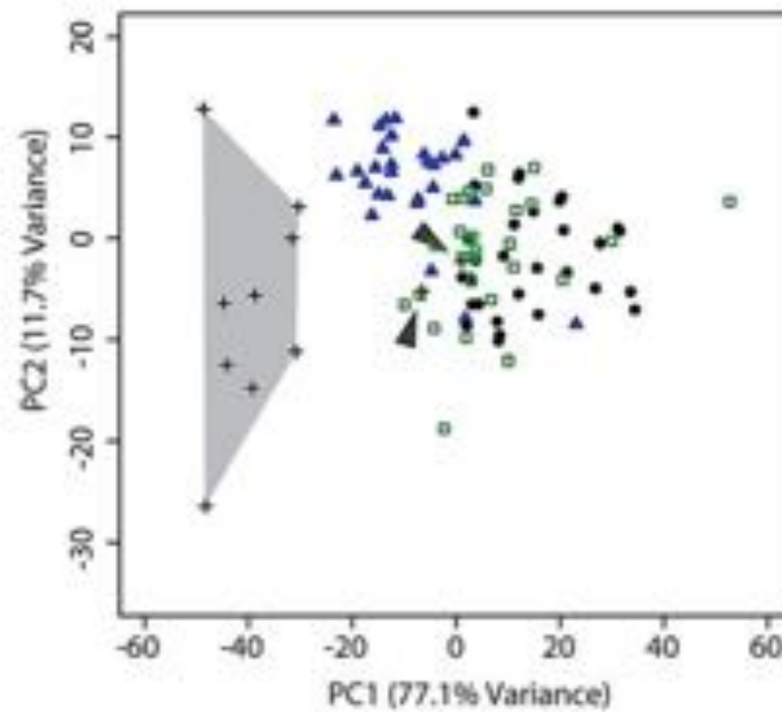
Han, Cao, Barb, Punyasena, Jaramillo, and Shyu, *Applications in Plant Sciences* (2014)

Knowledge Capture

Automated Analysis

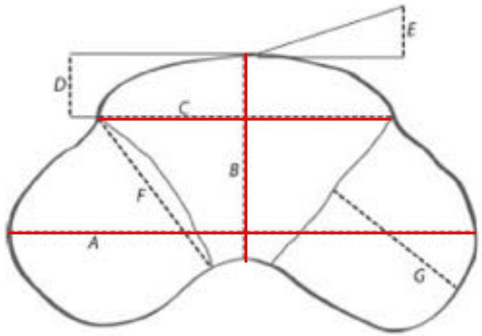






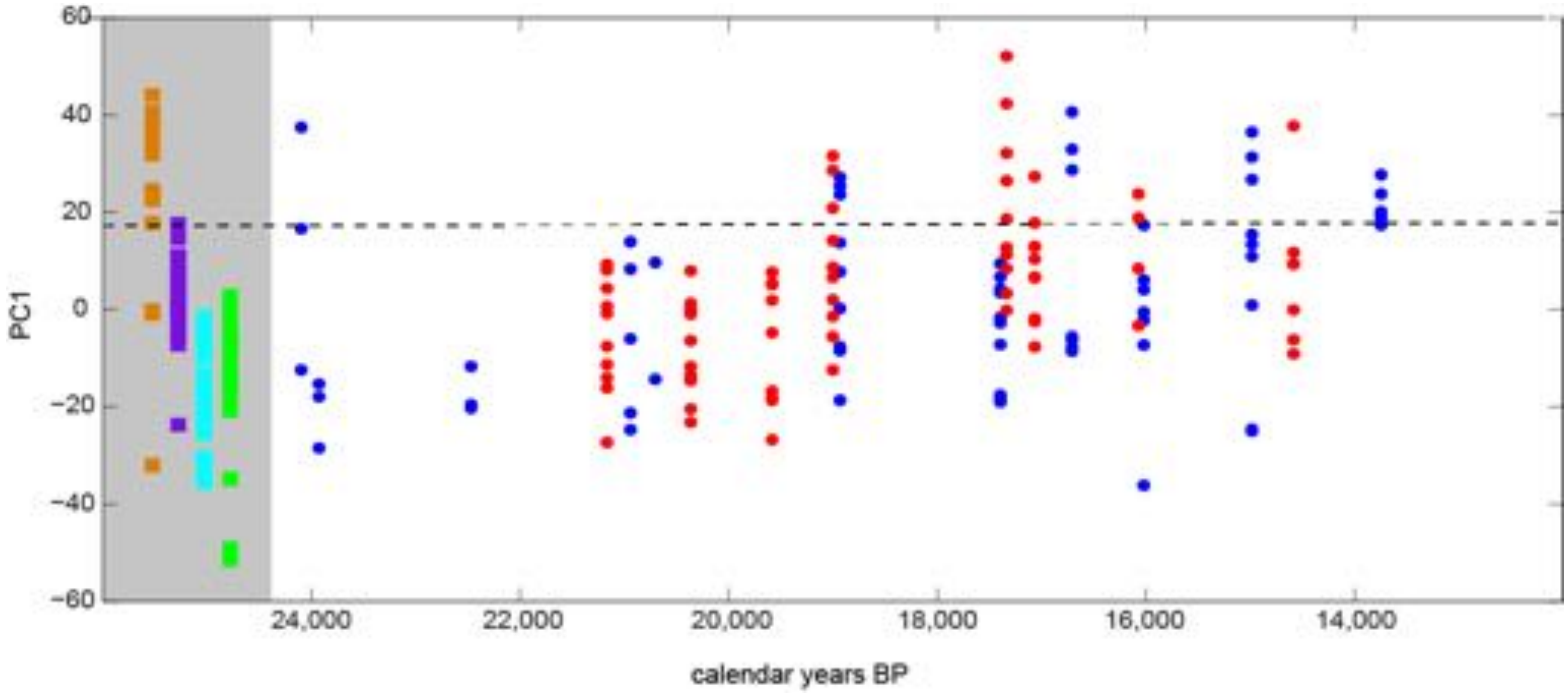
Mander, Rodriguez, Mueller, Jackson and Punyasena, *JQS (*2014)

RECONSTRUCTING PALEO-POPULATION DYNAMICS

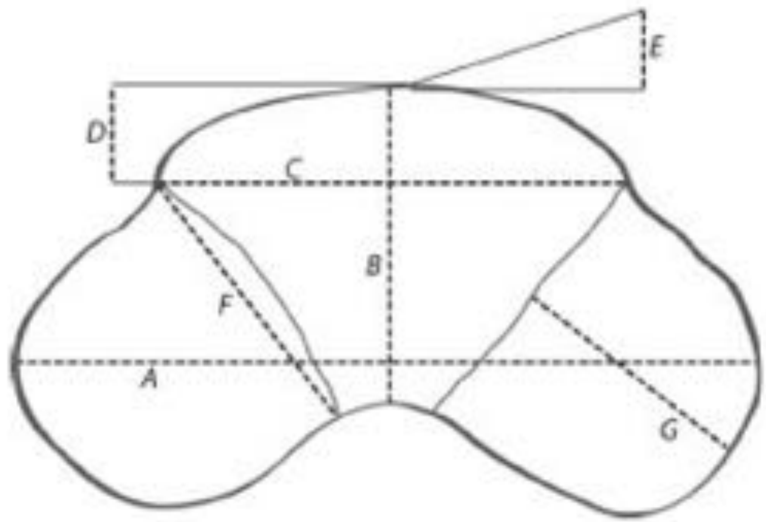


Anderson Pond, TN

Cupola Pond, MO

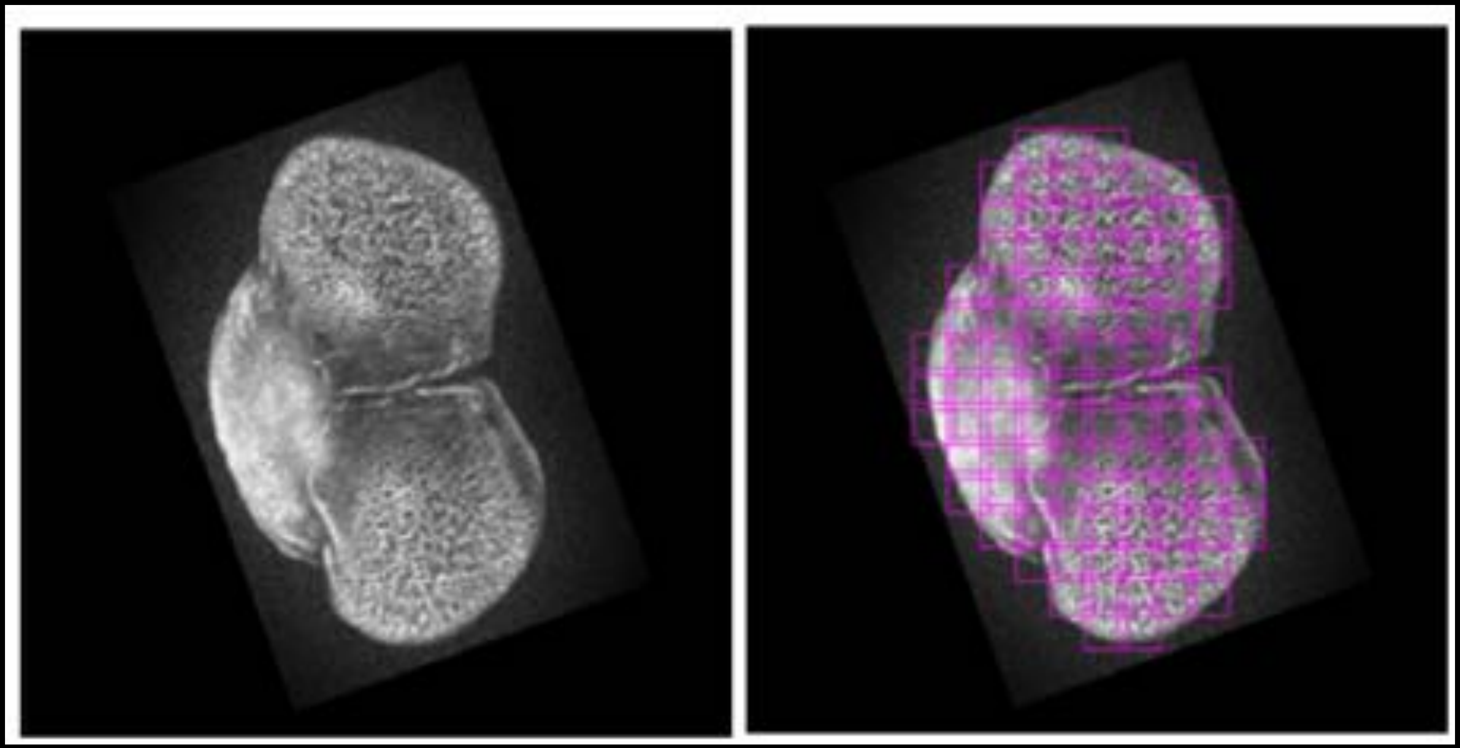


Punyasena, Rodriguez, Mander, and Jackson, unpublished



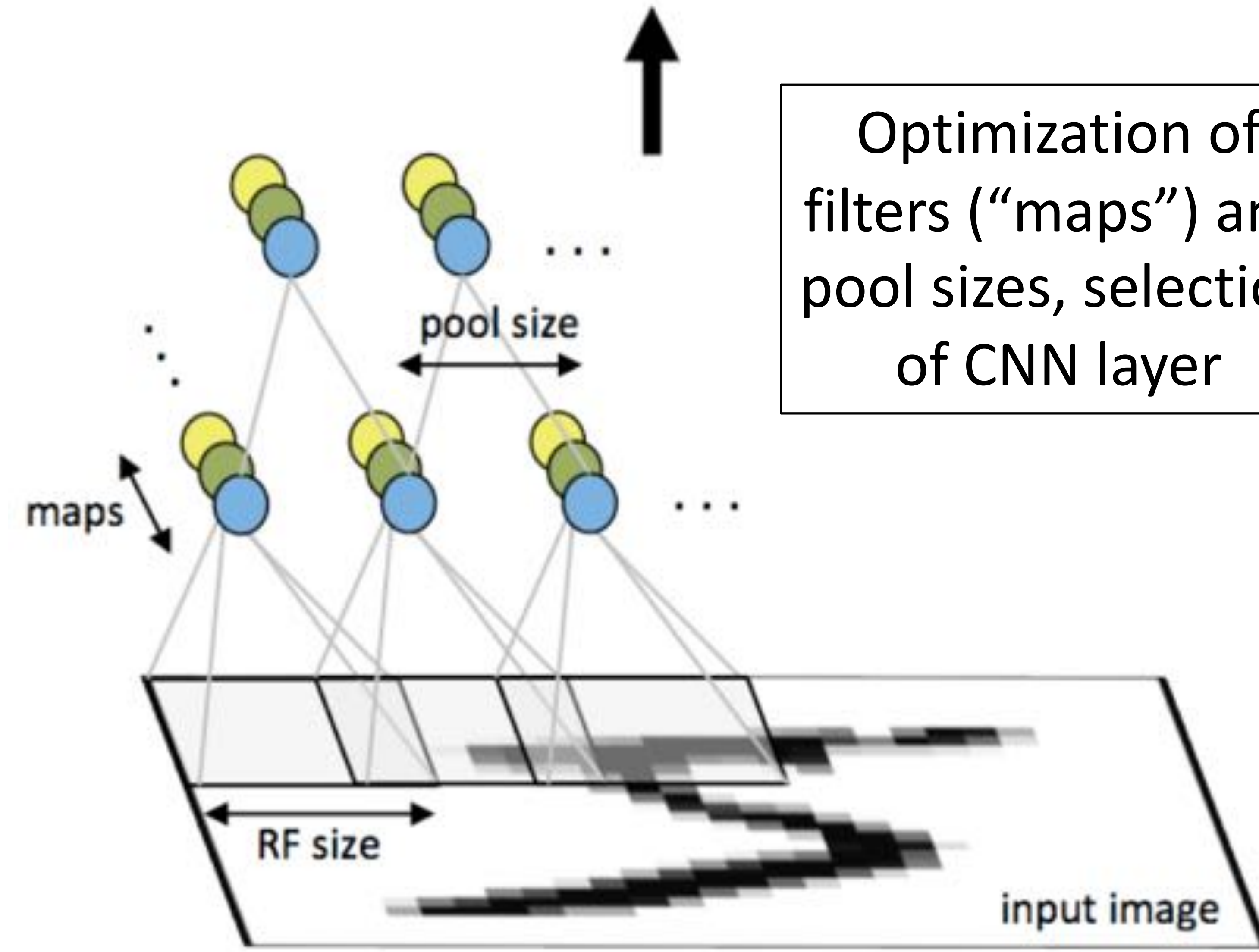
Morphometric approaches only allow us to classify grains in a specific orientation.

Collaborators at UC Irvine, Computer Science (Shu Kong and Charless Fowlkes) experimented with “deep learning” convolution neural nets as an alternative.



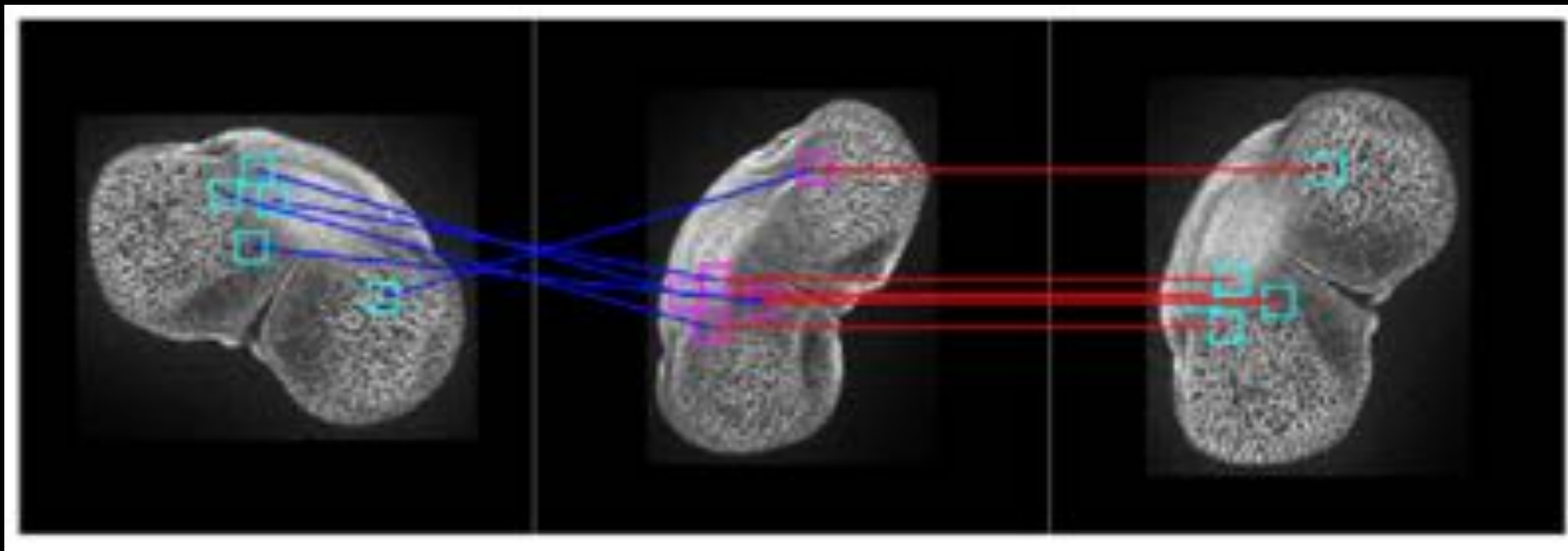
Created a set of overlapping patches (52 x 52 pixels)

Patch features were learned using a CNN (VGGVeryDeep-19) model that created an output of hidden layers (512-dimensional vectors) that were used as our classifiers.



Optimization of filters ("maps") and pool sizes, selection of CNN layer

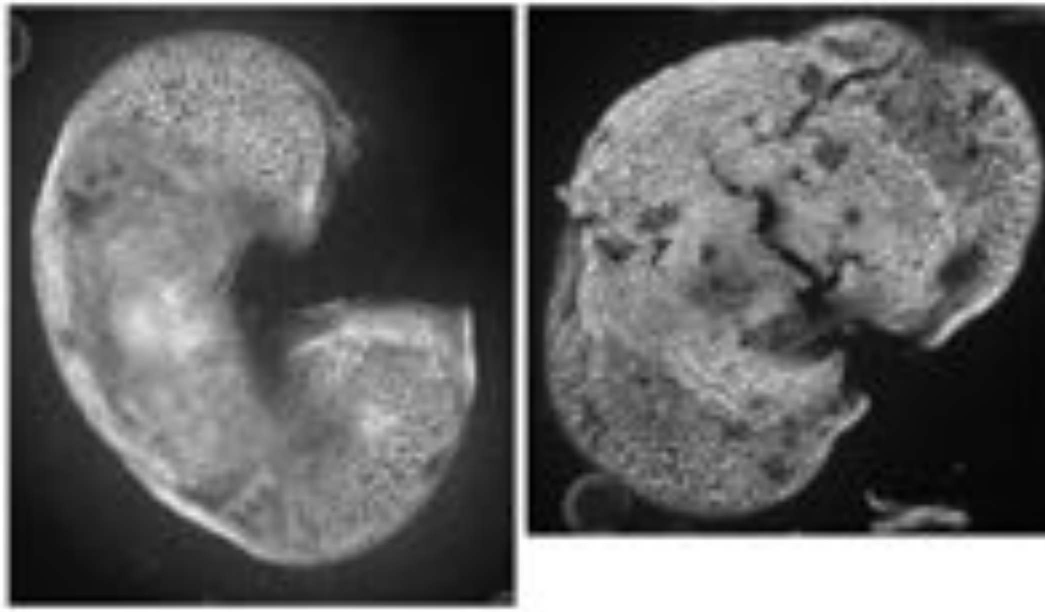
http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/



Used *k*-medoids clustering of downsized images to identify canonical shapes. Only two representations were necessary: equatorial and polar views.

Images were rotated to match canonical shapes.

MODERN REFERENCE



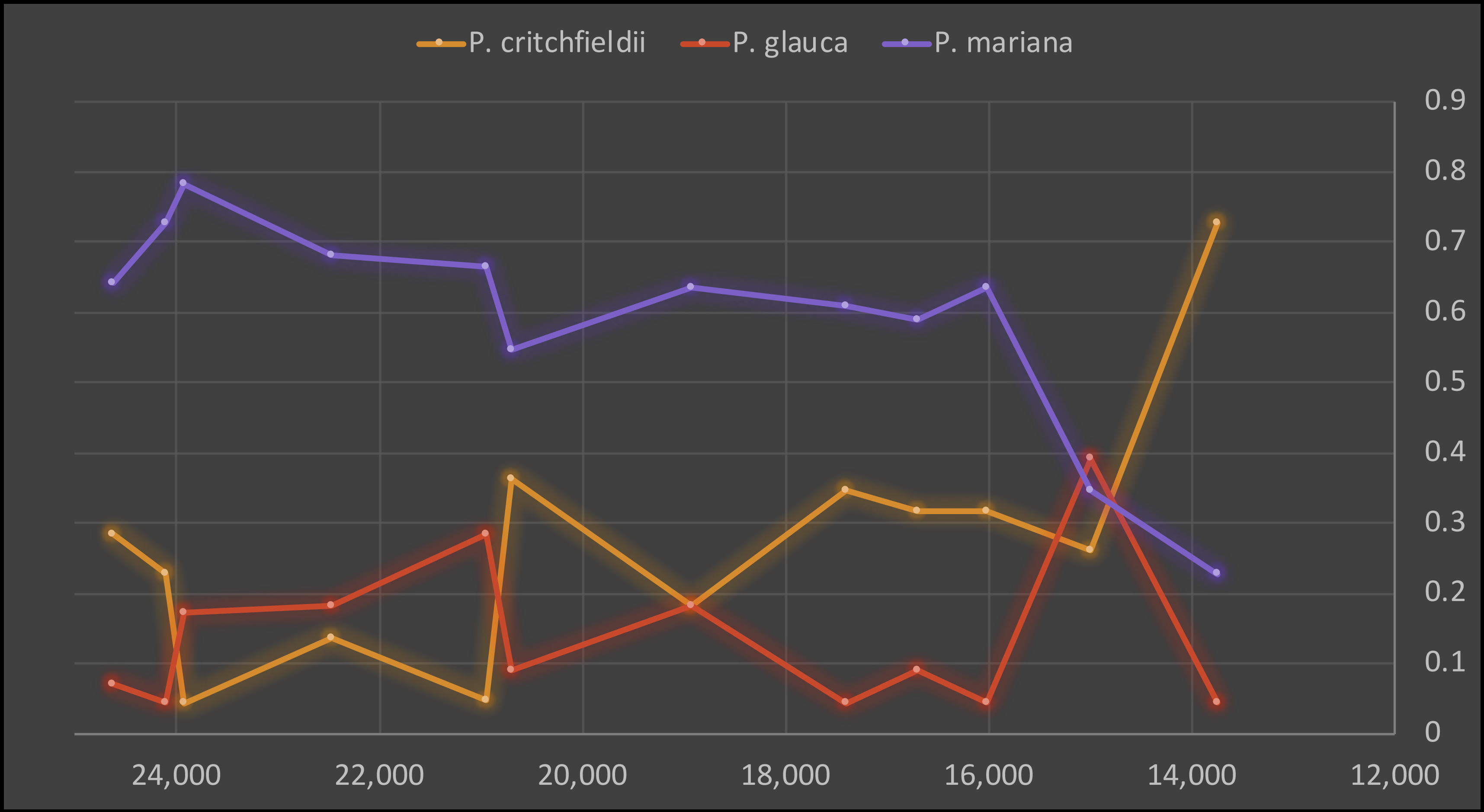
FOSSIL

System was trained on reference material and <u>used to classify fossil samples</u>.

70% accuracy (no published attempt can replicate)

A PRELIMINARY APPLICATION

Anderson Pond, TN



Years BP

HANDLING VISUAL (MICROSCOPIC) DATA

- Many of the technologies needed to do this work well (advanced microscopy, slide scanners, computer vision algorithms) already exist.
- Applicable to a range of microfossil (and macrofossil) datatypes.
- Potential to streamline data capture and analysis and improve our science (new hypotheses could be tested, more ambitious research performed)
- As a community, we should apply these tools to our work (and not fear that we are relinquishing our expertise)

Funders and Collaborators

Michael Urban, Ingrid Romero, Derek Haselhorst, Jacklyn Rodriguez, Luke Mander

Mayandi Sivaguru, Glenn Fried
Institute for Genomic Biology, UIUC

Charless Fowlkes, Shu Kong
U California Irvine, Computer Science

Chi-Ren Shyu, Jing Han
U Missouri, Computer Science

Washington Mio, Mao Li, Nigel Nye
Florida State, Mathematics

Carlos Jaramillo and lab
Smithsonian Tropical Research Institute

NSF-DBI – Advances in Biological Informatics

NSF-DBI/EF – Innovations in Biological Imaging & Visualization

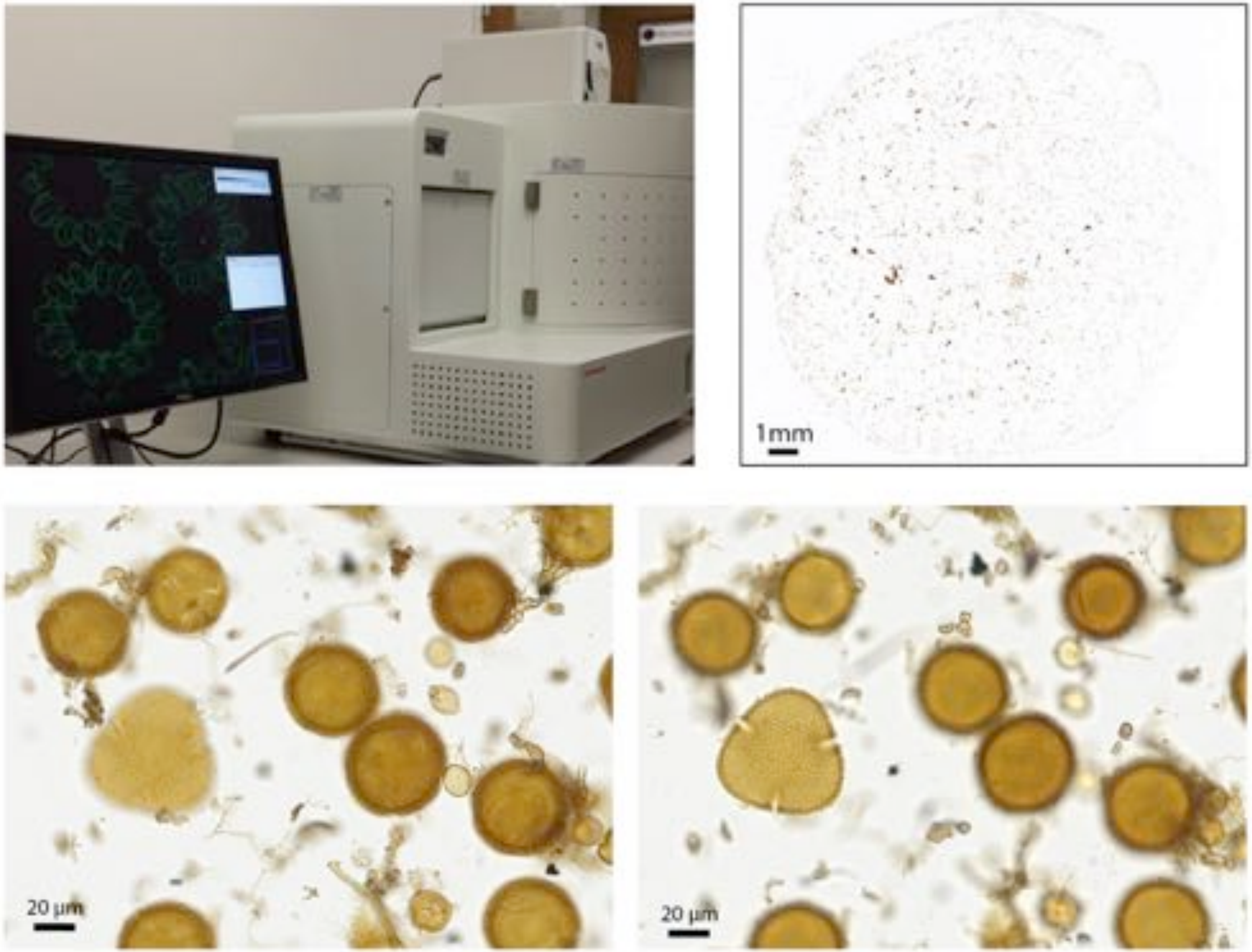
NSF-EF – Macrosystems Biology

National Center for Supercomputing Applications (NCSA) Institute for Advanced Computing Applications and Technologies (IACAT)

UIUC Campus Research Board

NSF XSEDE : Texas Advanced Computing Cluster (TACC); National Center for Supercomputing Applications (NCSA)

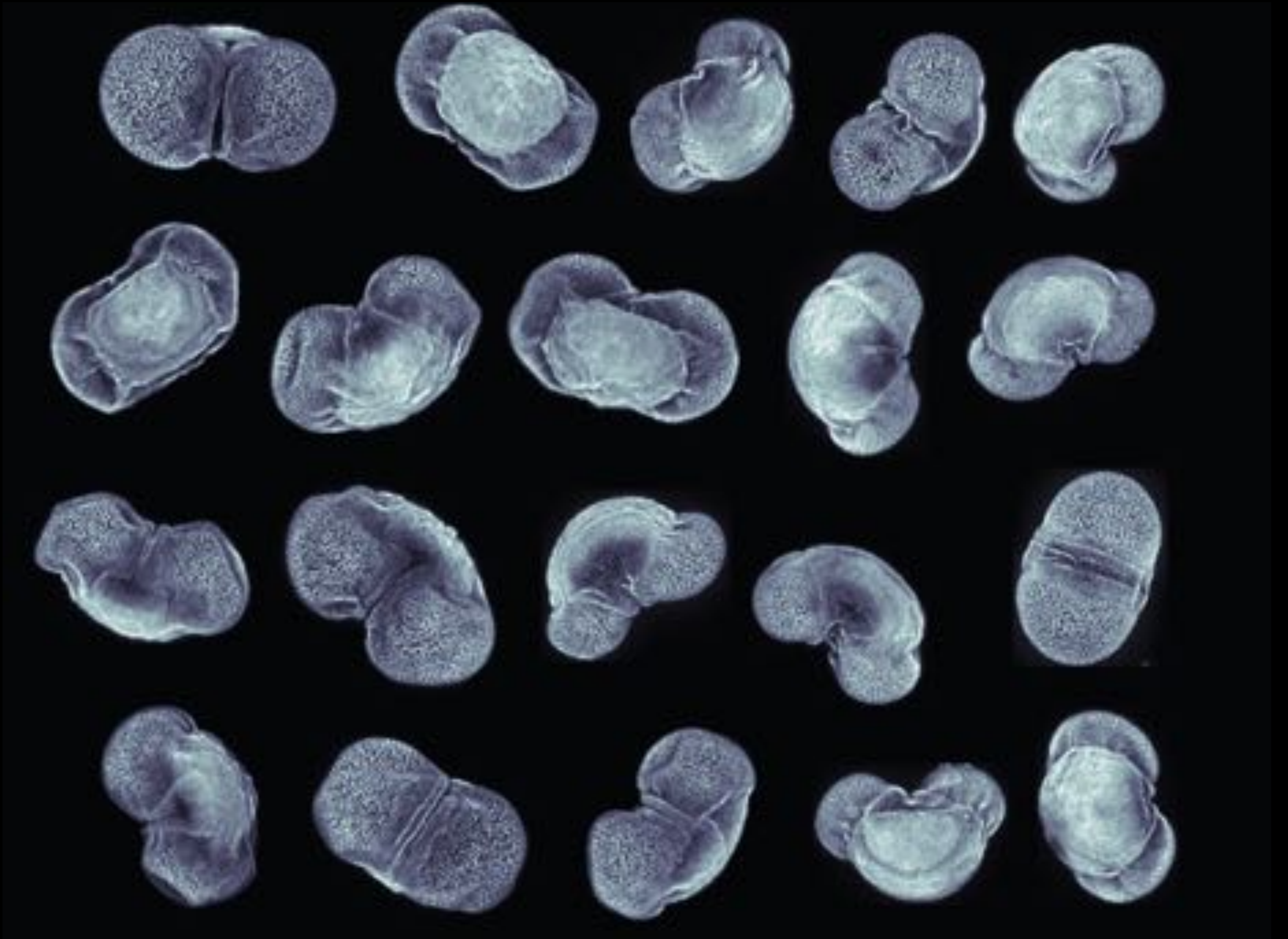
Throughput

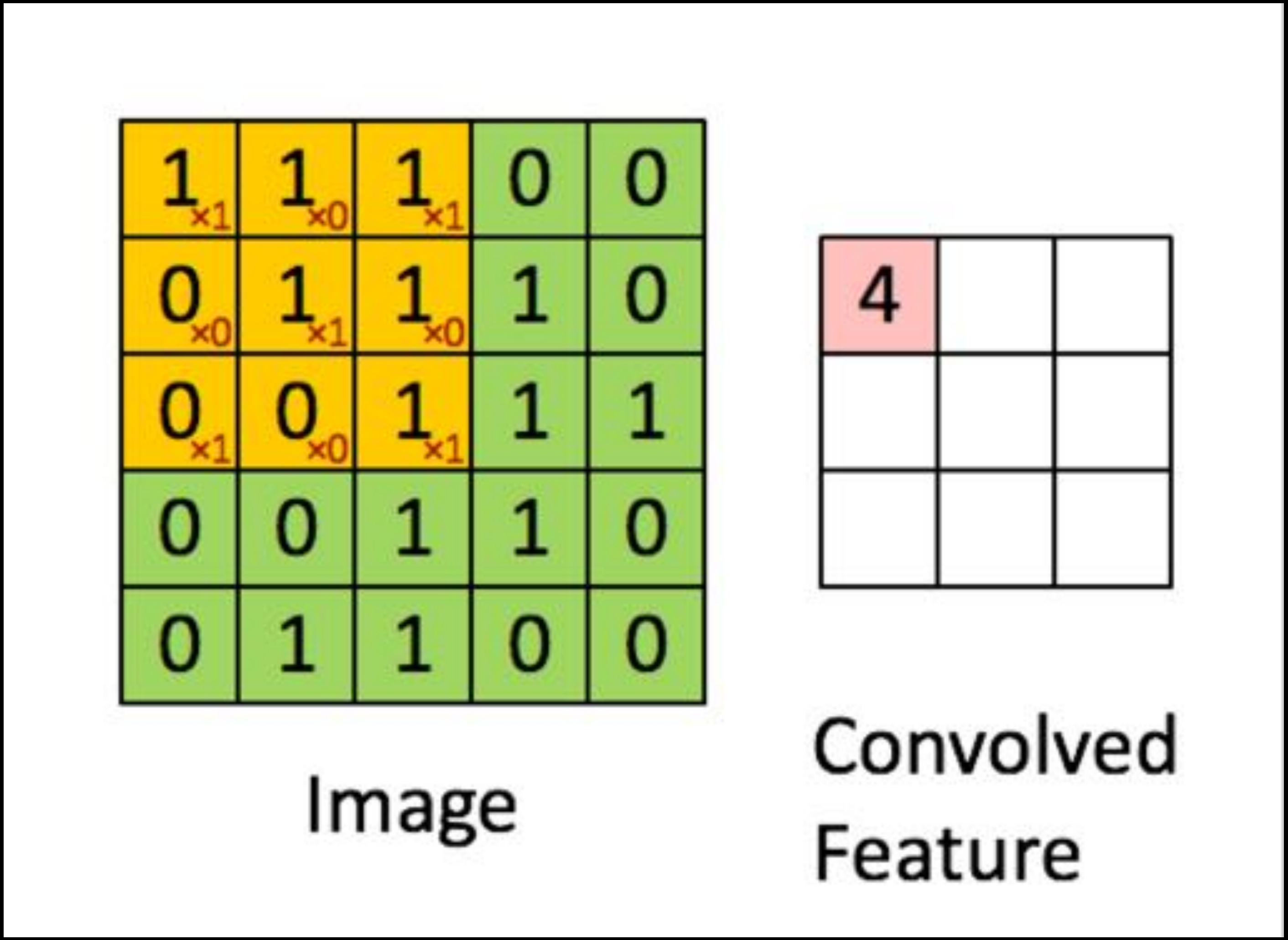


One sample (1 cm^2, 60 axial planes, 400x, 0.23 µm/pixel)
~500 GB, 2 hours

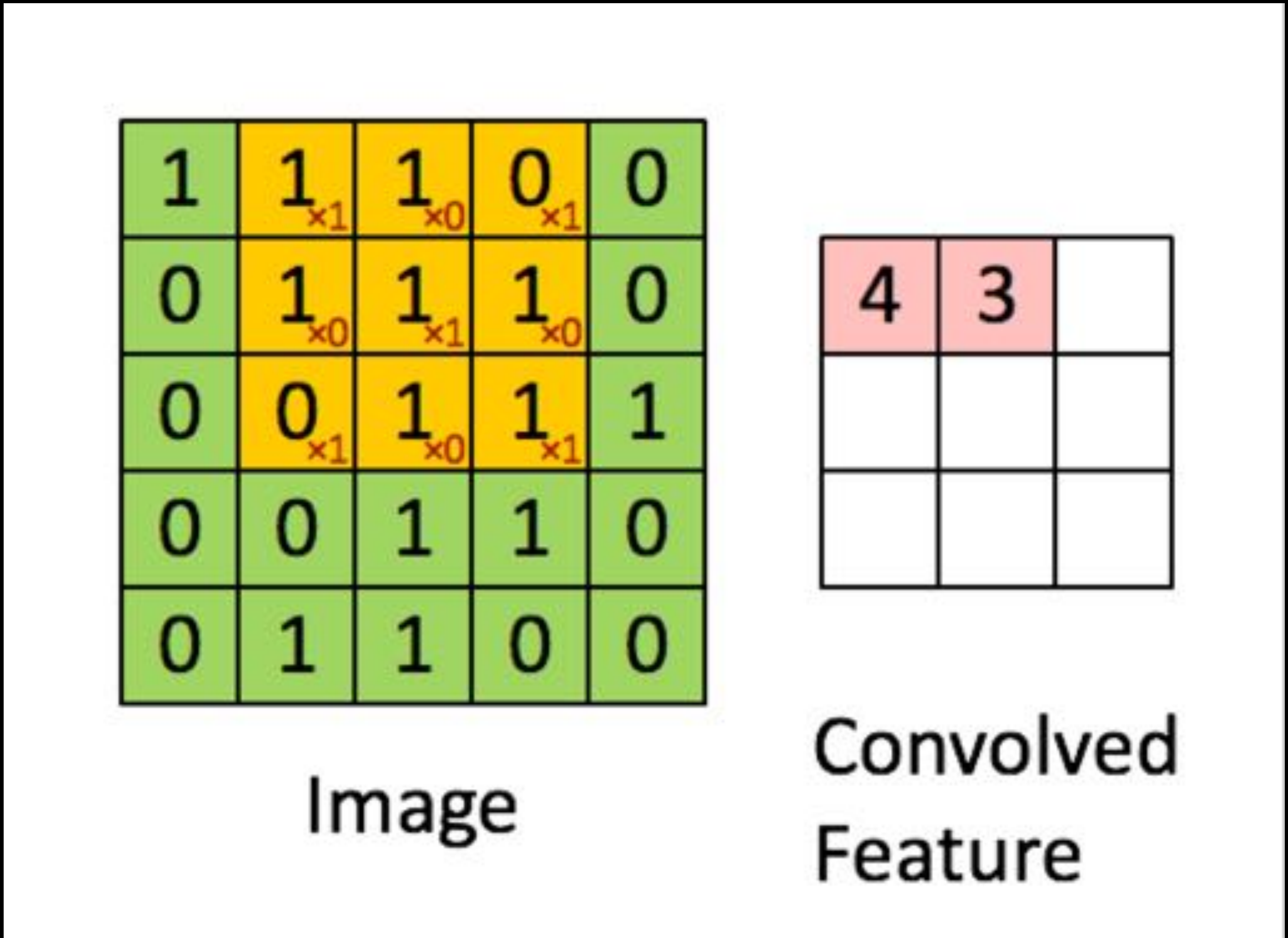
Used in: Tcheng, Nayak, Fowlkes and Punyasena, *PLOS One* (2016)

Arrives at Illinois May 2017; 630x magnification
Highest resolution available with light microscopy
(brightfield, fluorescence)

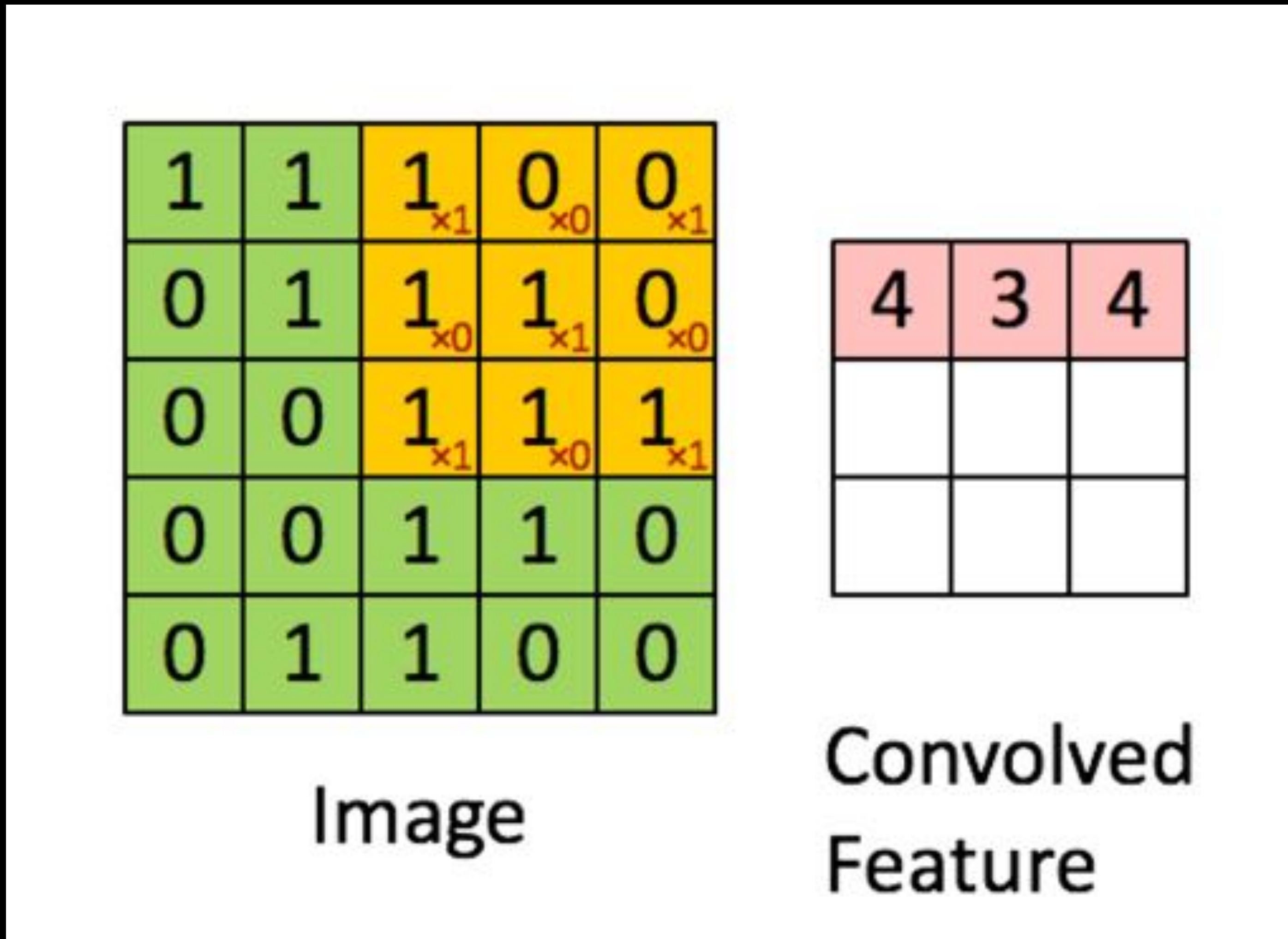




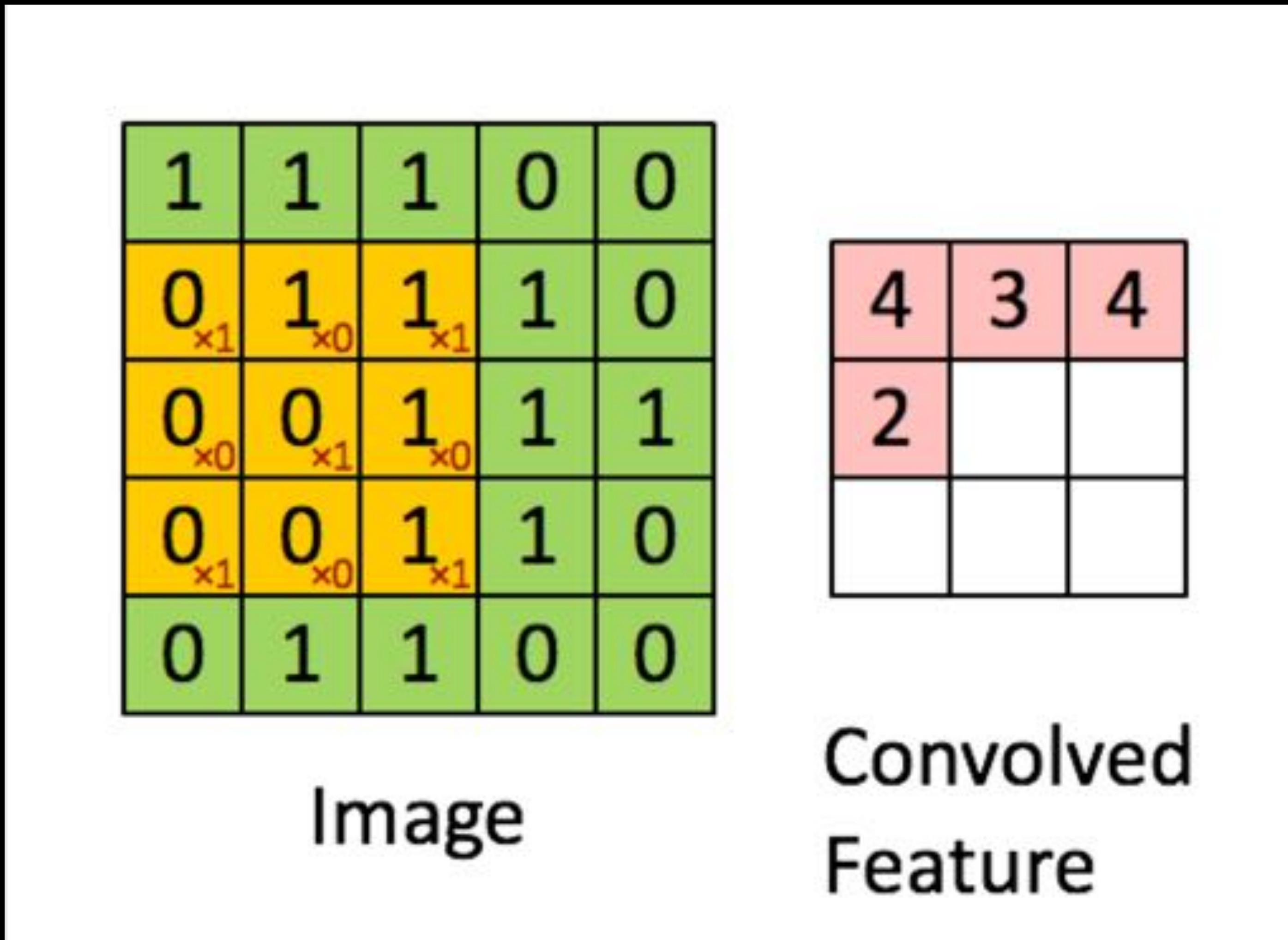
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



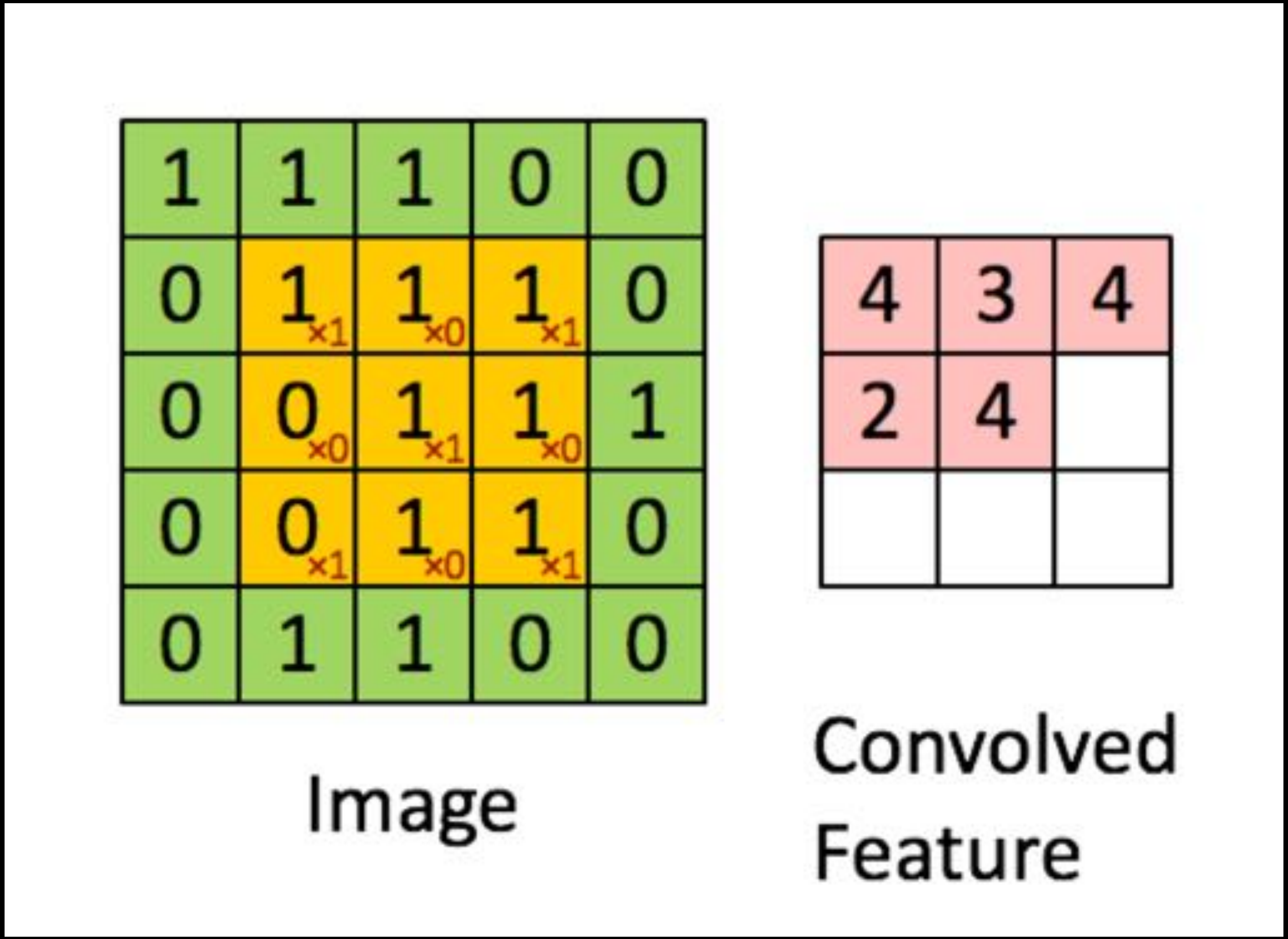
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



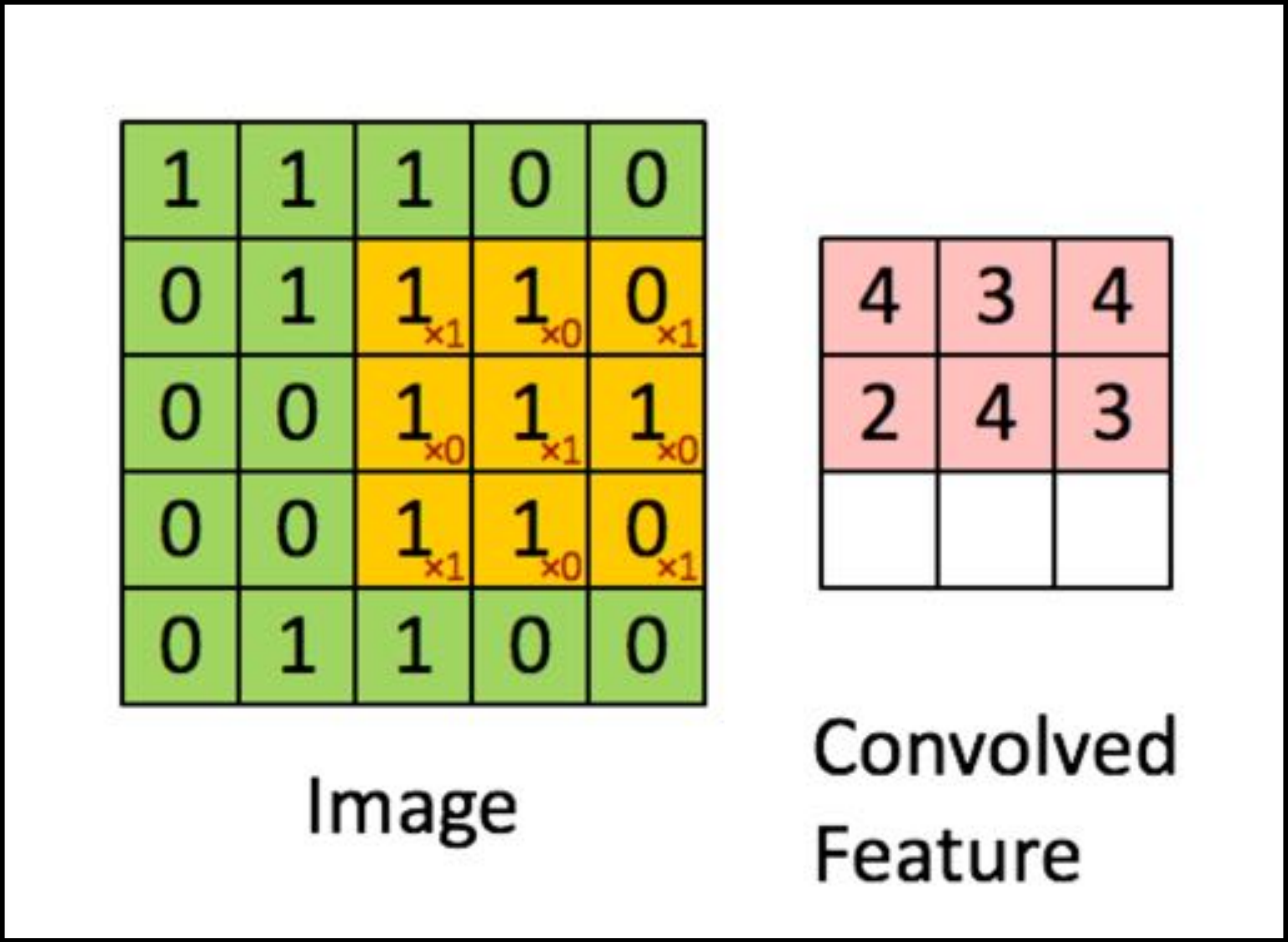
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



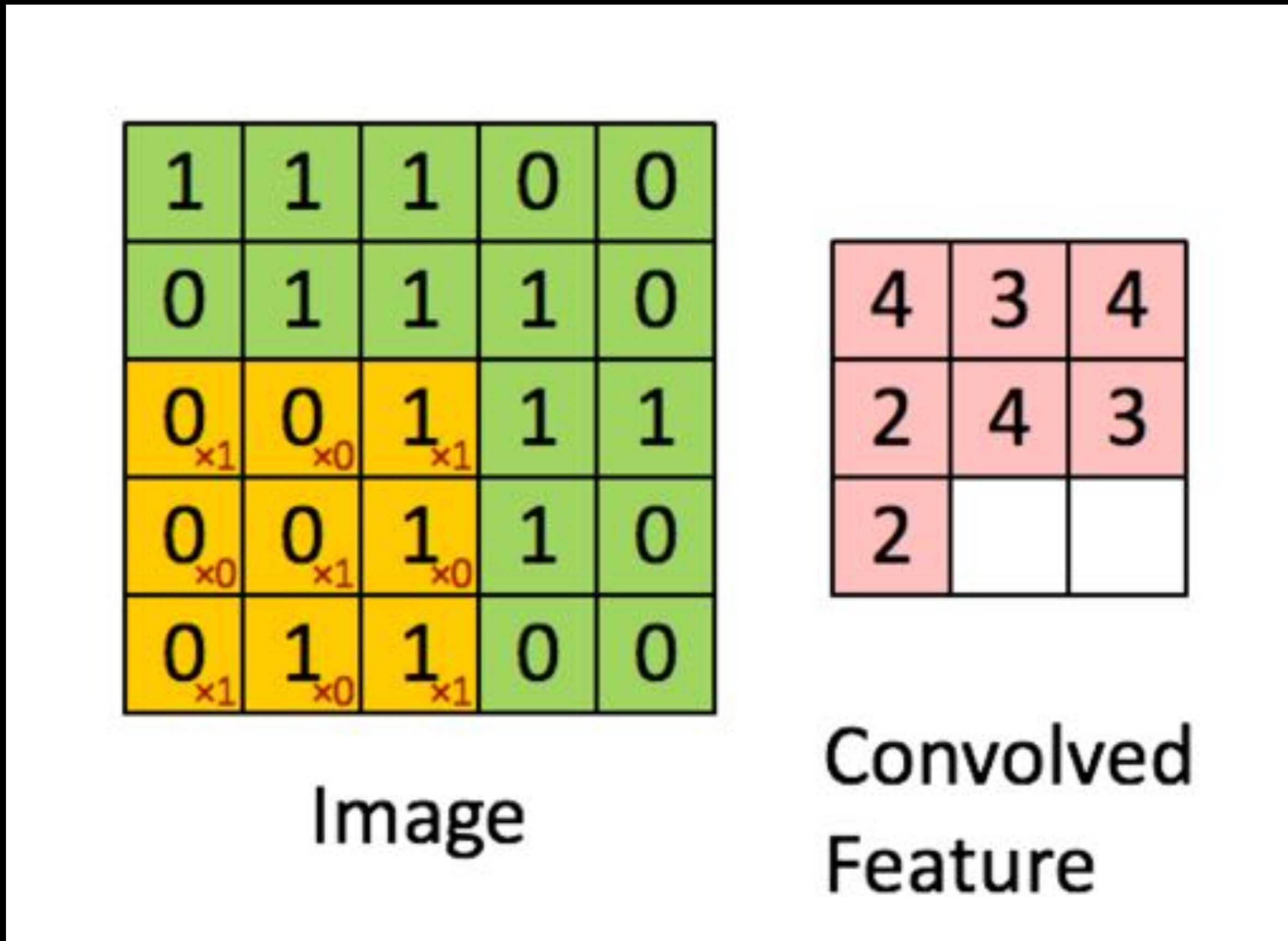
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



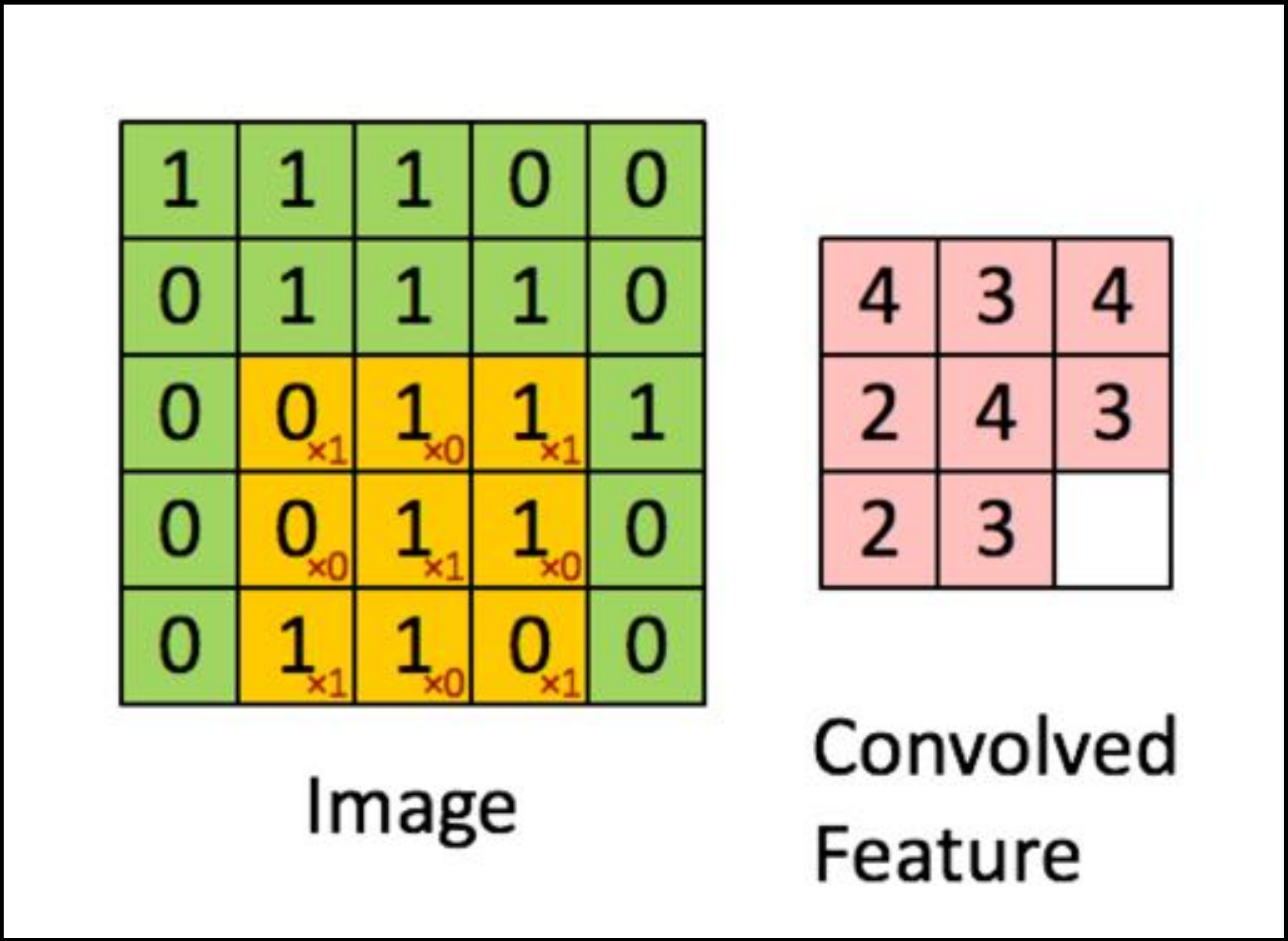
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



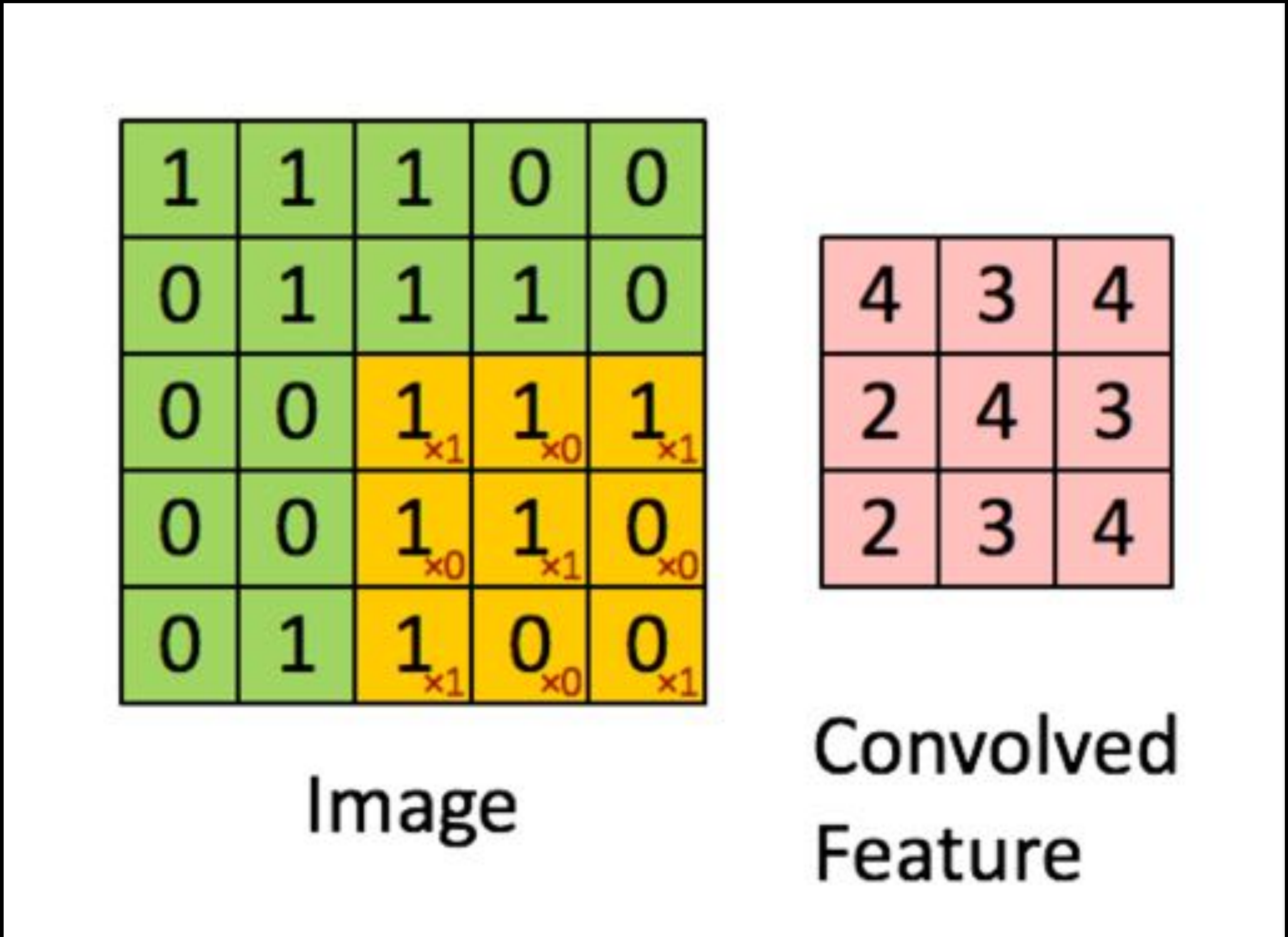
http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/



http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/

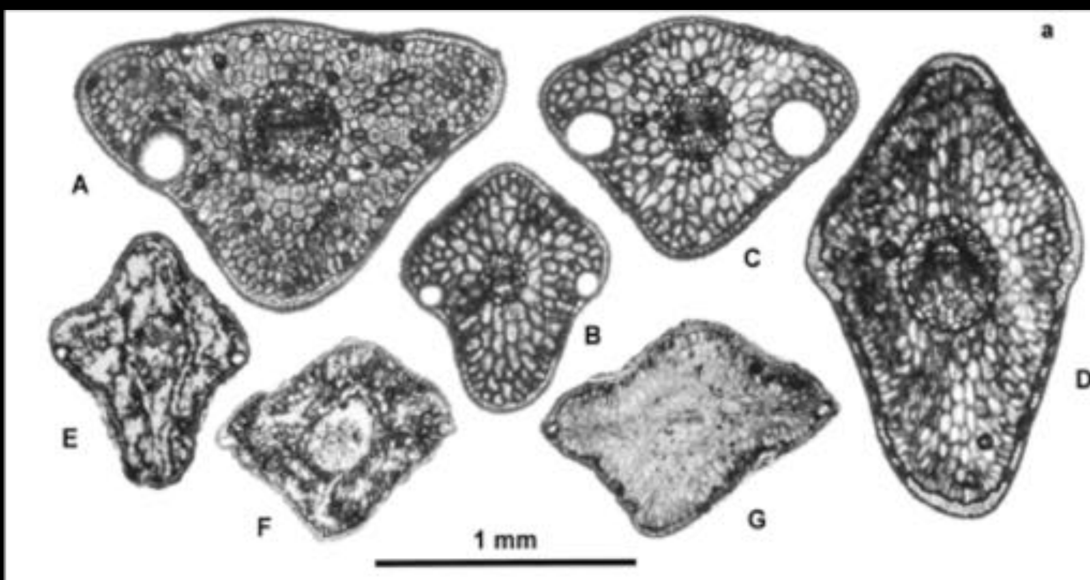
Late Quaternary extinction of a tree species in eastern North America

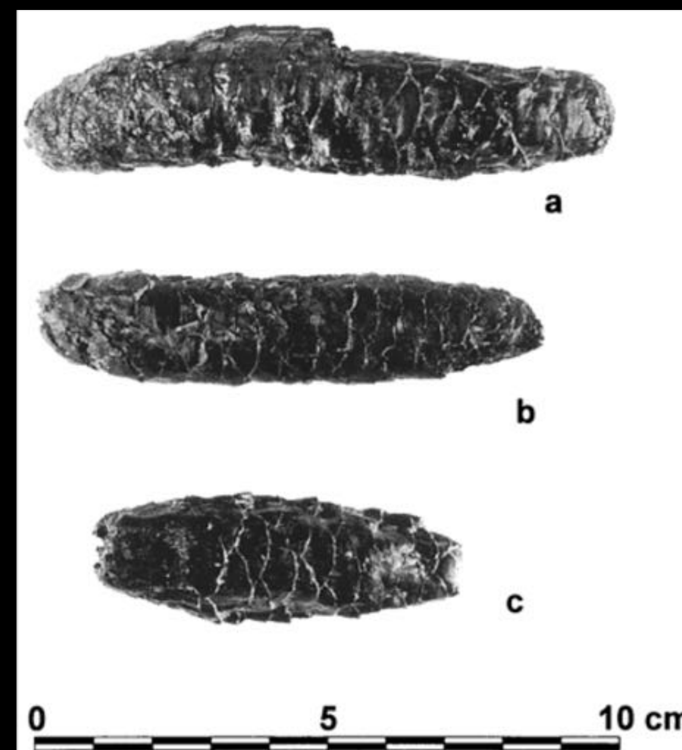
Stephen T. Jackson* and Chengyu Weng†

Department of Botany, University of Wyoming, Laramie, WY 82071

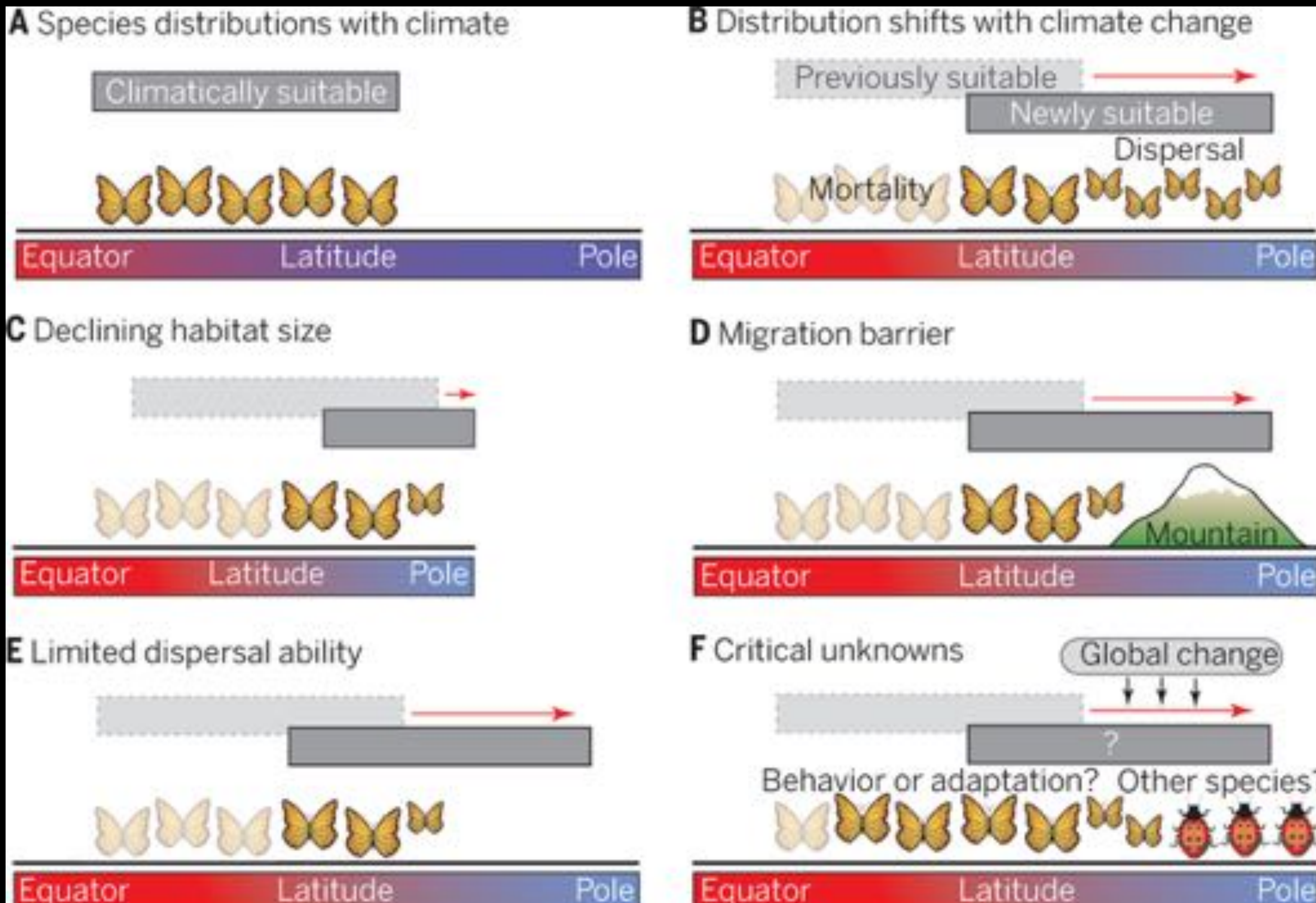
Edited by Margaret Bryan Davis, University of Minnesota, St. Paul, MN, and approved September 27, 1999 (received for review July 8, 1999)

Proceedings of the National Academy of Sciences 1999





HYPOTHESES FOR EXTINCTION



Lambers, *Science*, 2015

“Southeastern” *sensu lato* localities with >1% Late Pleistocene/Holocene spruce pollen

AIRYSCAN SUPERRESOLUTION MICROSCOPY

Brightfield	SEM	TEM	Airyscan
-Most common -Traditional method	Highest resolution of ornamentation or superficial (>5nm)	Highest resolution of internal ultrastructures (>0.5nm)	- High resolution (250-125nm) -Internal structure of the wall
Limited morphological (>250nm)	Only surface detail	High time consuming Loss of material	Slightly lower resolution than EM