MICROSCOPIA QUANTITATIVA: DADOS MASSIVOS, INTELIGÊNCIA ARTIFICIAL E NOVAS FERRAMENTAS

Dani Ushizima, Ph.D.

Staff Scientist - CRD, Lawrence Berkeley National Lab.

Data Scientist - BIDS, University of California, Berkeley

BERKELEY LAB

Bringing Science Solutions to the World

WHAT IS BERKELEY LAB?

One of 17 Department of Energy National Laboratories, Berkeley Lab — managed by the University of California — conducts non-classified, basic and applied scientific research in the public interest.

BERKELEY LAB

MISSION STATEMENT

Bringing Science Solutions to the World

Berkeley Lab fosters the groundbreaking fundamental science that brings transformational solutions to the world's most urgent energy and environmental challenges and a greater understanding of the universe.

BERKELEY LAB

SCIENTISTS, ENGINEERS, SUPPORT STAFF, & STUDENTS

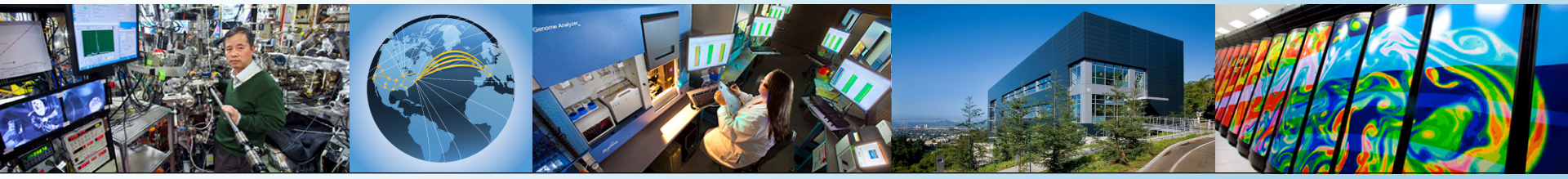
- 962 Scientists
- 740 Engineers
- 473 Postdoctoral Associates
- 644 Research Support
- 296 Graduate Students
- 141 Undergraduate Students
- 758 Operations/Administrative Support

4014

BERKELEY LAB

NATIONAL USER FACILITIES

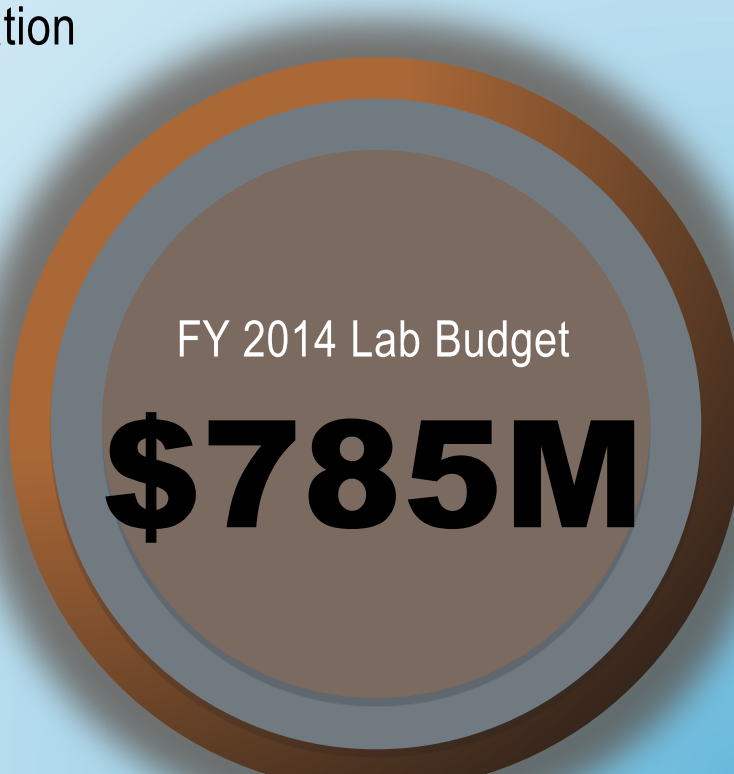
ALS ESnet JGI Molecular Foundry NERSC



9,911 external scientific facility users per year

ECONOMIC IMPACT

- $441 million in employee salary and benefits have directly impacted the regional economy
- $326 million in procurements were awarded to a diverse group of large and small businesses for a broad range of products and service across the nation
- $141 million in purchasing made in California
- $4.4 million in licensing and royalty income
- 114 patents issued (70 US, 44 foreign)
- 739 active income-bearing license agreements



EXCELLENCE

- 13 Nobel Prizes (1 group prize)
- 70 National Academy of Sciences members
- 14 National Medal of Science recipients
- 18 National Academy of Engineering members
- 16 Elements of the Periodic Table discovered at the Lab

13

70

14

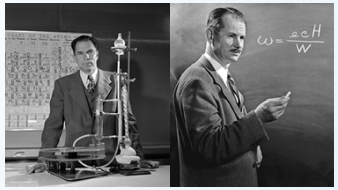
18

16

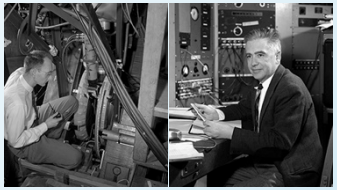
BERKELEY LAB

NOBEL LAUREATES

E.O. Lawrence
1939: Physics

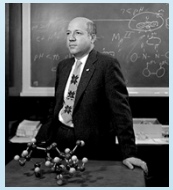


Glenn T. Seaborg and Edwin M. McMillan
1951: Chemistry



Owen Chamberlain and Emilio Segrè
1959: Physics

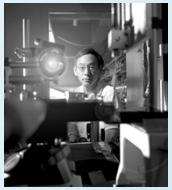
Donald Glaser
1960: Physics



Melvin Calvin
1961: Chemistry

Luis W. Alvarez
1968: Physics

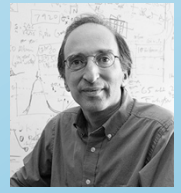
Yuan T. Lee
1986: Chemistry



Steven Chu
1997: Physics

George F. Smoot III
2006: Physics

Intergovernmental Panel on Climate Change
2007: Peace

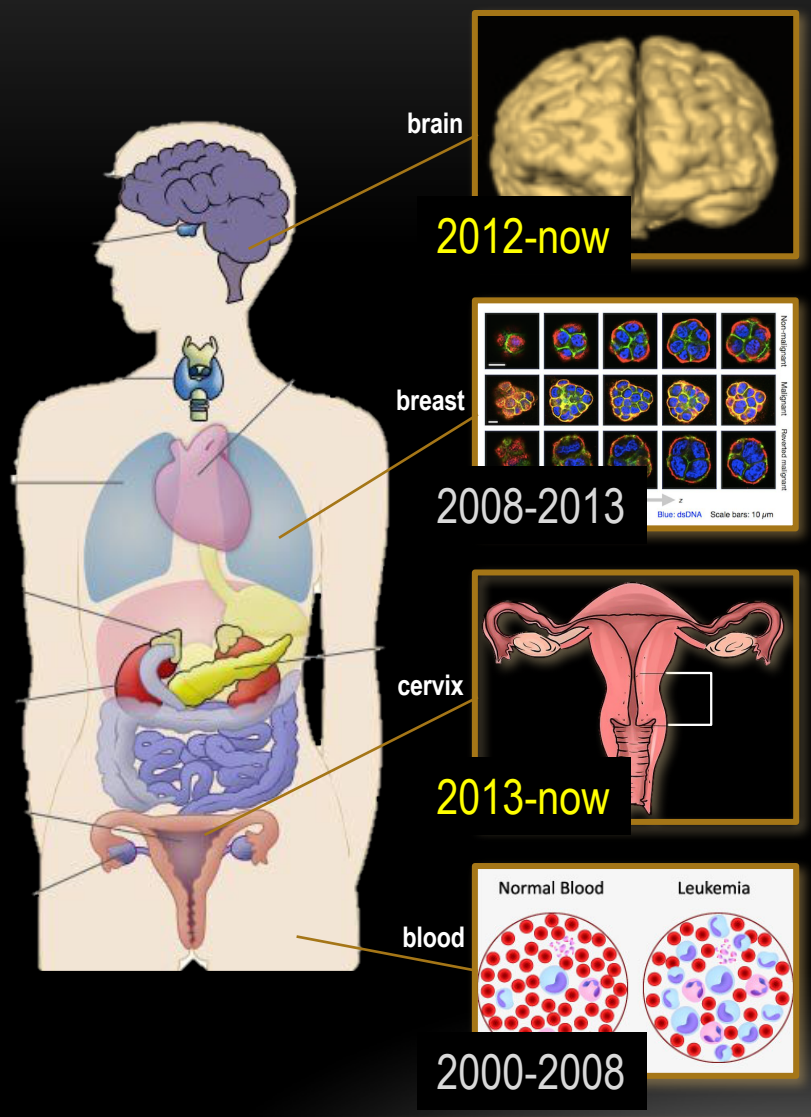


Saul Perlmutter
2011: Physics

BERKELEY LAB

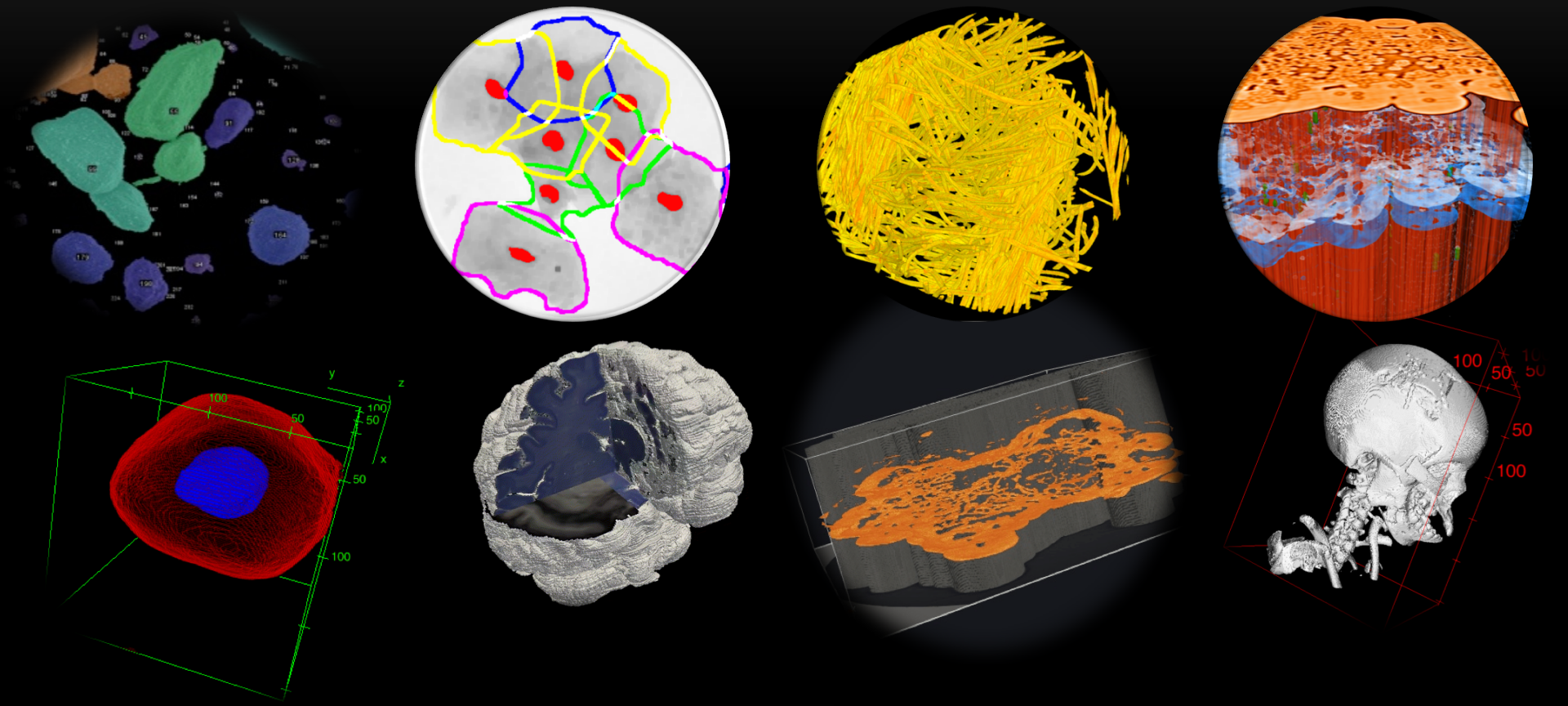
HELPING THE GOVERNMENT WITH DATA

COMPUTATIONAL HEALTH SCIENCE



- Problem:
  - Quantification from images;
- Approach:
  - Bio: detect and track biomarkers associated to the progression of diseases;
  - Math: schemes for image representation, segmentation, characterization, classification, ML;
- Expected impact:
  - Software to measure and reproduce experiments;
  - Develop new treatments that target individuals more precisely.

QUANTITATIVE IMAGE ANALYSIS ACROSS DOMAINS



COMO DINAMIZAR O PROCESSO DE MICROSCOPIA QUANTITATIVE?

Tarefas e ferramentas

ALGORITHMS

1. Image format and metadata
2. Image enhancement and filtering
3. Image statistics
4. Image segmentation: super-pixel to object
5. Object description
6. Object classification
7. Object visualization

SOFTWARE TOOLS

A. ImageJ, Fiji
B. Python, ipython, jupyter notebooks
C. R, Rstudio
D. Paraview, Tomviz

CONNECTING THE DOTS

1. Image format and metadata
2. Image enhancement and filtering
3. Image statistics
4. Image segmentation
5. Object description
6. Object classification
7. Object visualization

A. Fiji

B. Python

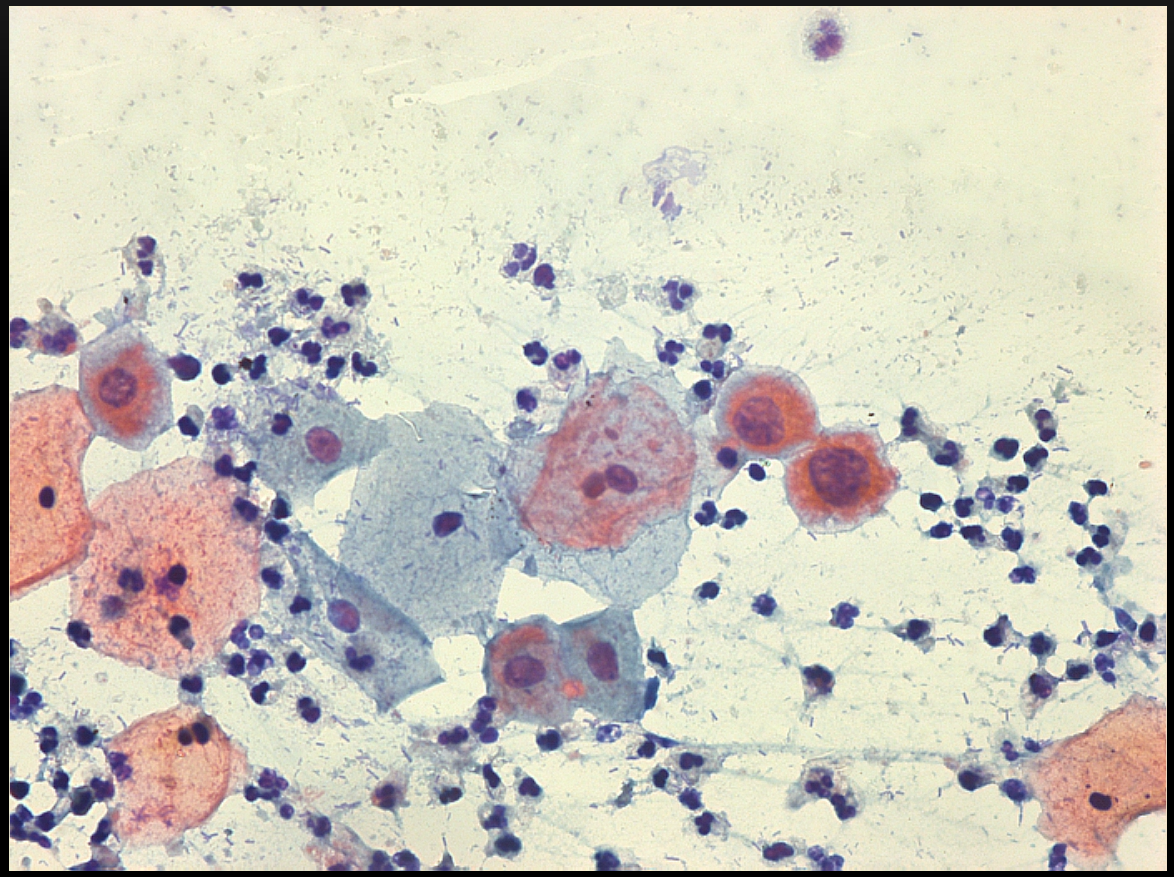
C. R

D. Tomviz

BEING A DATA SCIENTIST

1. Look at the big picture
2. Get the data (“easy” and “difficult”)
3. Discover: visualize data to gain insights
4. Prepare data for Machine Learning algorithms
5. Select a model and train it
6. Fine-tune your model
7. Present your solution
8. Launch, monitor, and maintain your system

DATA: CERVICAL CELLS



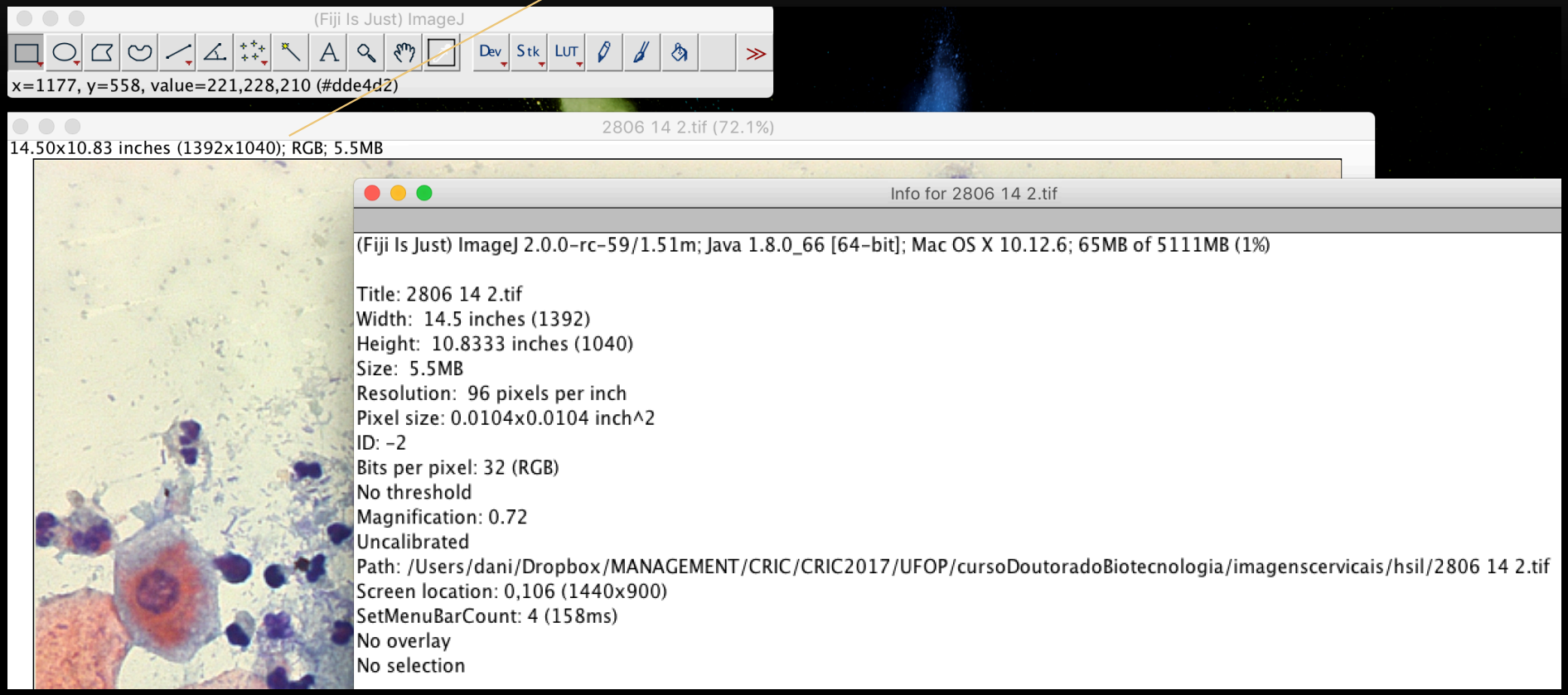
Courtesy: AMBAR lab, UFOP, Brazil – prof. Dr. Claudia Medeiros
https://www.facebook.com/programa.ambar/

1. IMAGE FORMAT AND METADATA

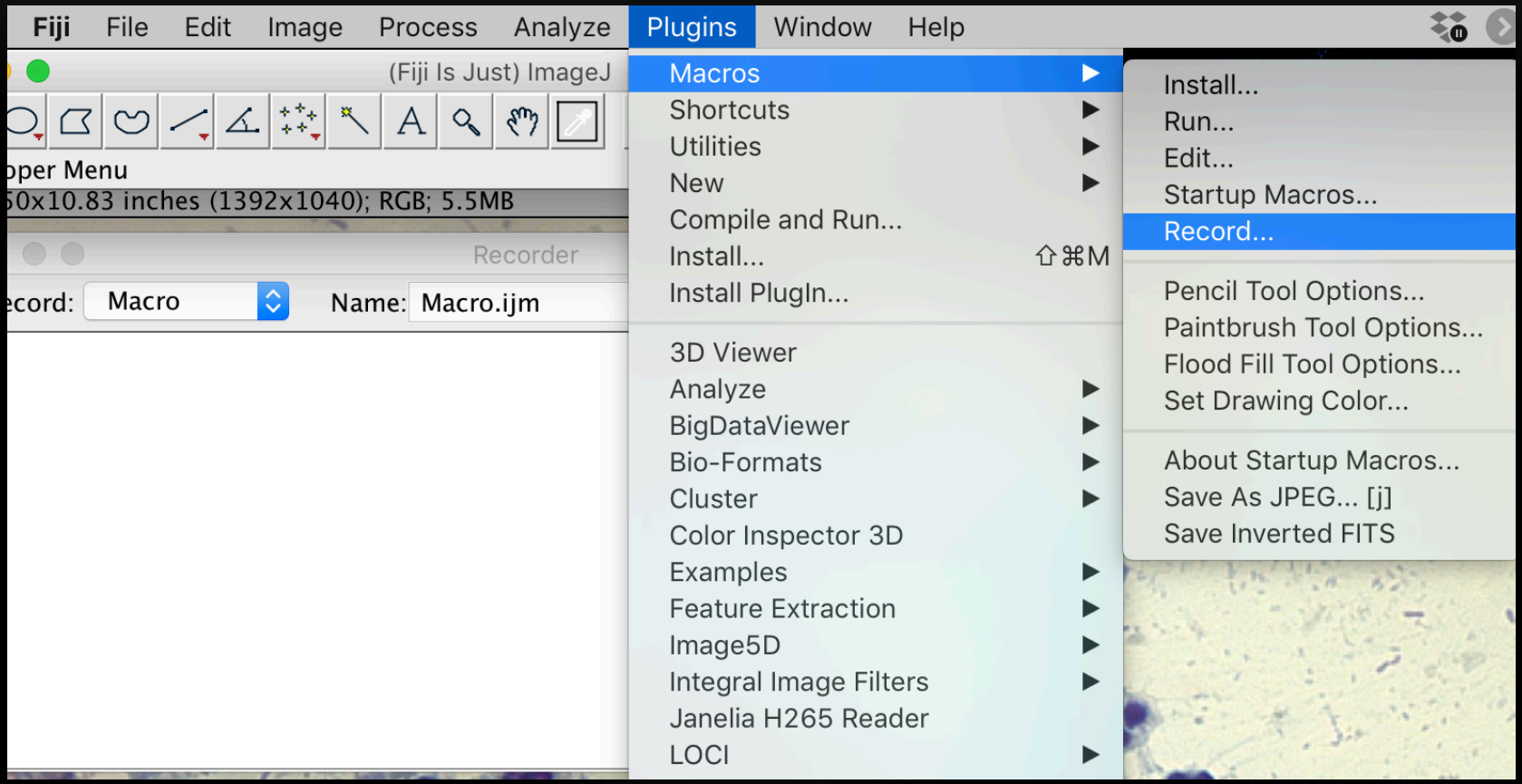
a) Other info & metadata
b) Macro
c) Image size
d) Best sample for prototyping
e) Color channels and stacks
f) Montage

METADATA: A GIFT AND A CURSE

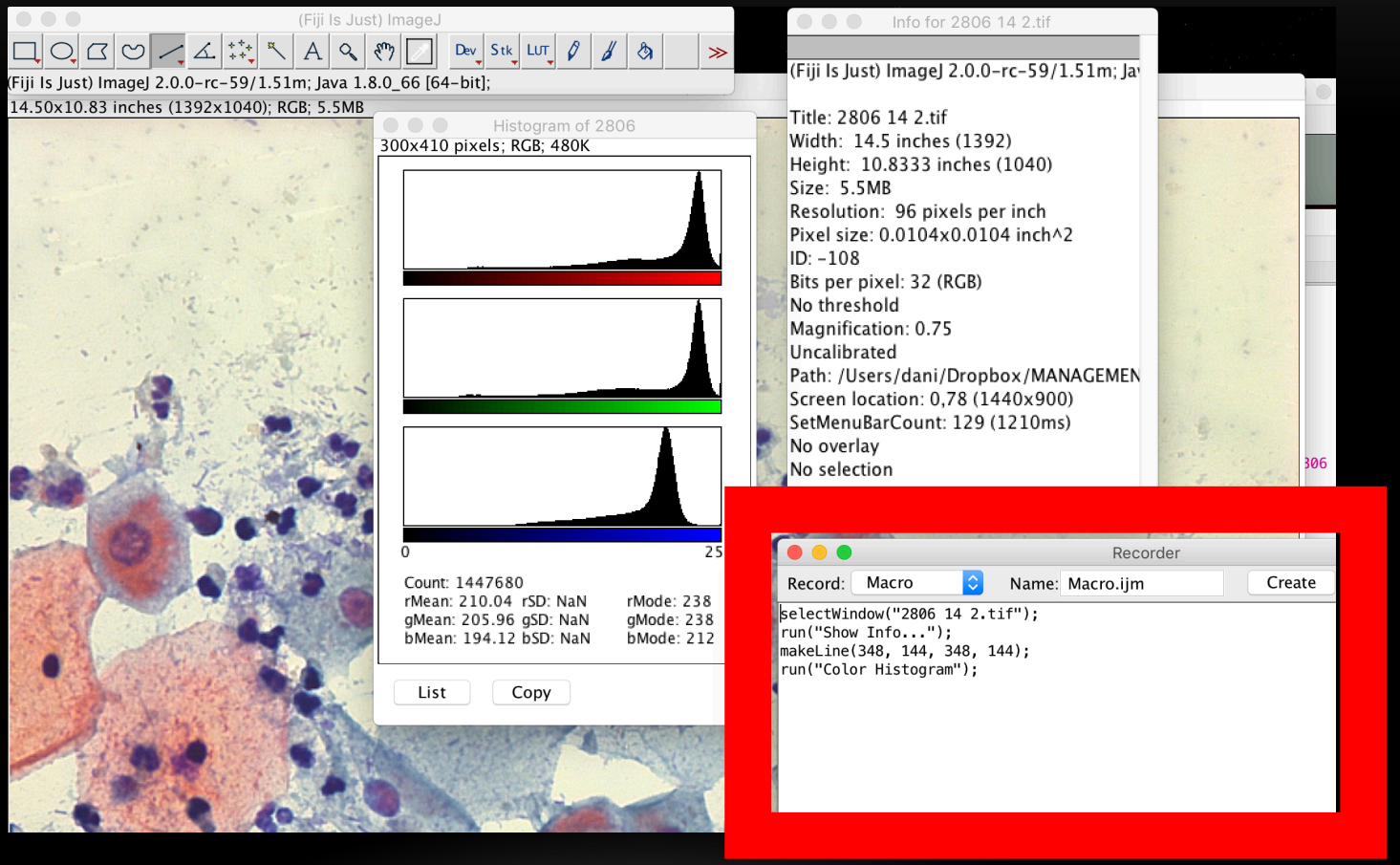
Resolution, size, channels, file size



RECORD YOUR STEPS IN A MACRO



PROVENANCE: CLICKING AROUND

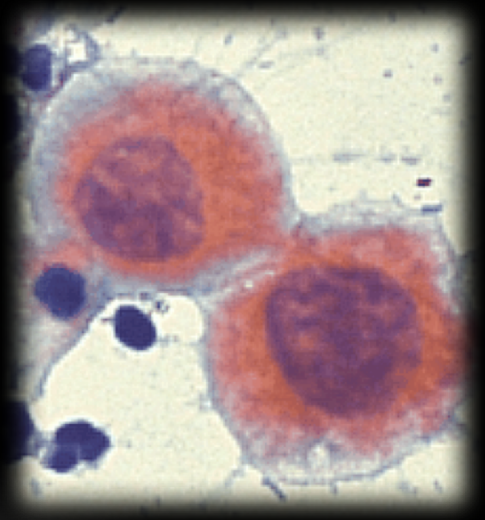


BOREDOM CHECK

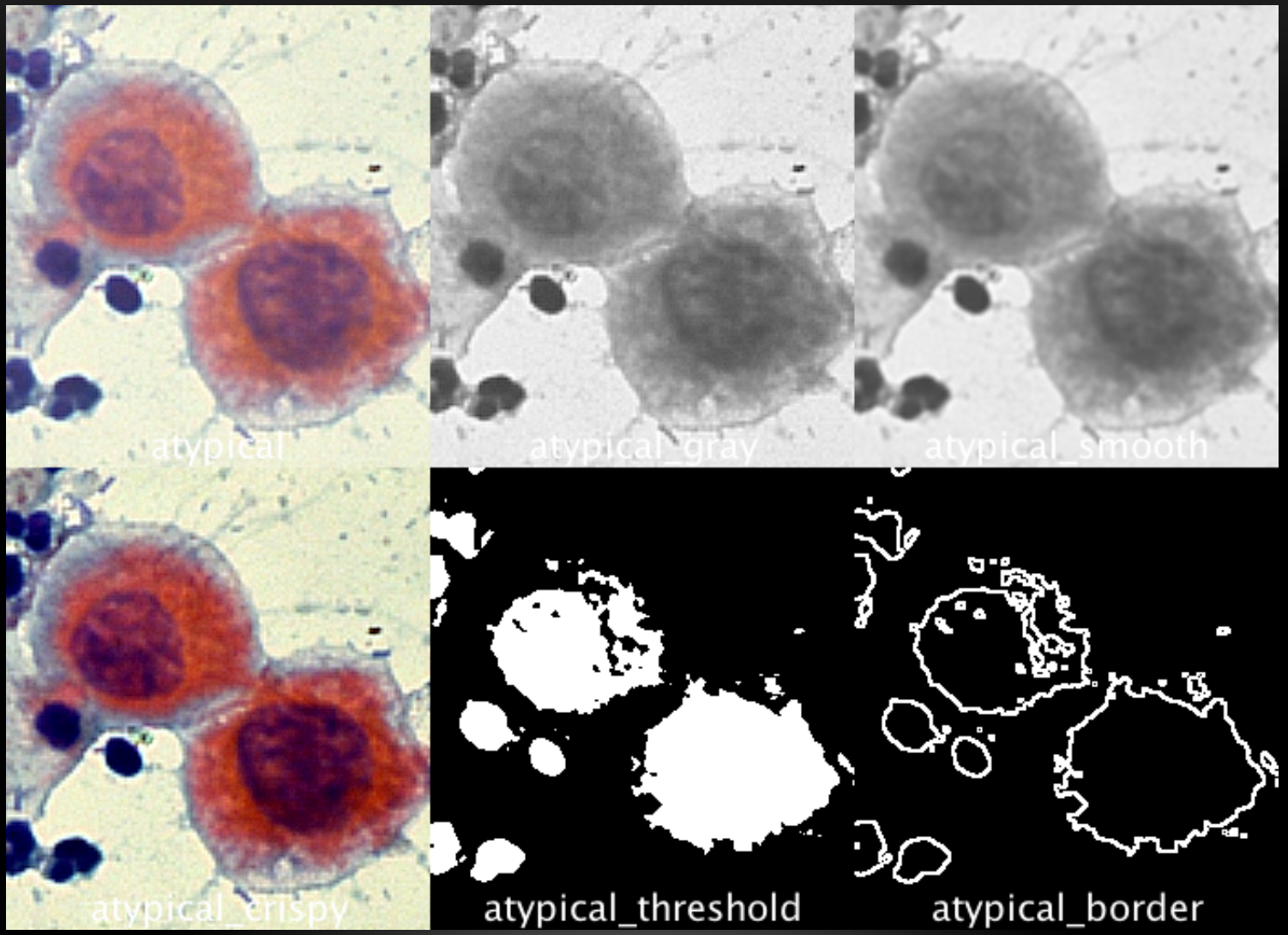
- How many of you knew these points?

2. IMAGE ENHANCEMENT AND FILTERING

a) Crop and resize
b) Color, grayscale and LUT
c) Enhancement for what? Smooth vs. crispy
d) Border detection
e) Create an animated gif



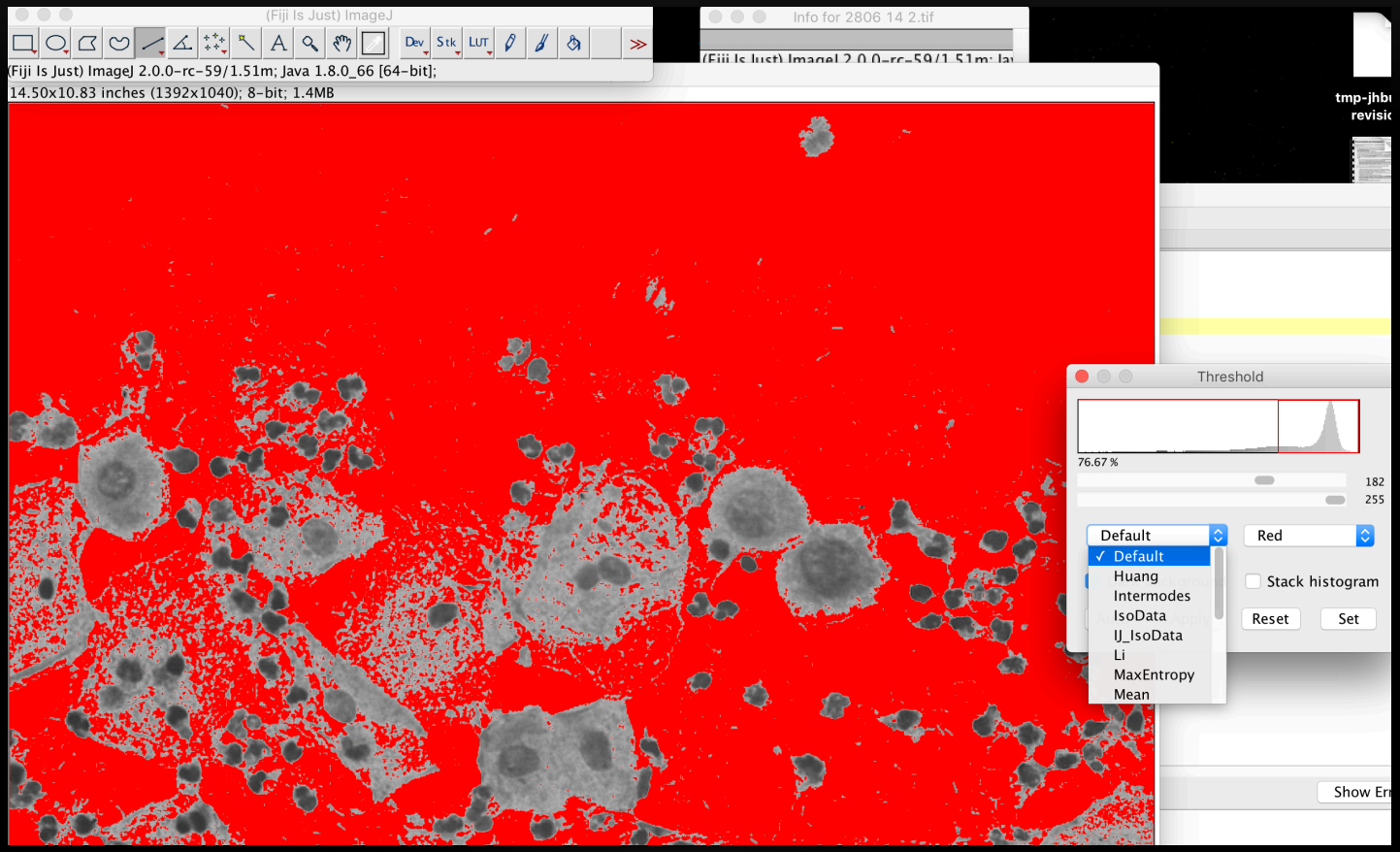
STEP BY STEP: REPRODUCIBLE RESEARCH



3. IMAGE STATISTICS

a) Profile
b) Image histogram
c) Enhance with histogram
d) Segment with thresholding
e) Algorithms to automate thresholding
f) Overlay your results

FIND THRESHOLD FOR GRAYSCALE CHANNEL



(Fiji Is Just) ImageJ

(Fiji Is Just) ImageJ 2.0.0-rc-43/1.51n / Java 1.8.0_31 (64-bit)

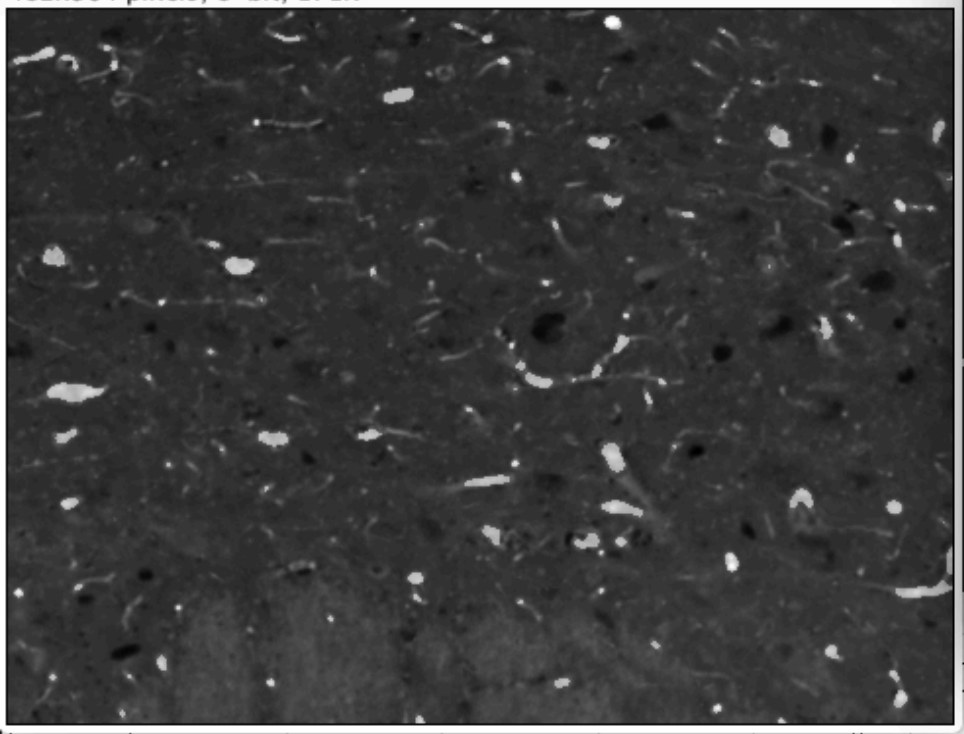
Recorder

Record: Macro Name: Macro.ijm

Smoothed with median

Composite (RGB)

482x364 pixels; 8-bit; 171K



Auto Threshold

Auto Threshold v1.16.1

Method: Try all

☐ Ignore black ☐ Ignore white

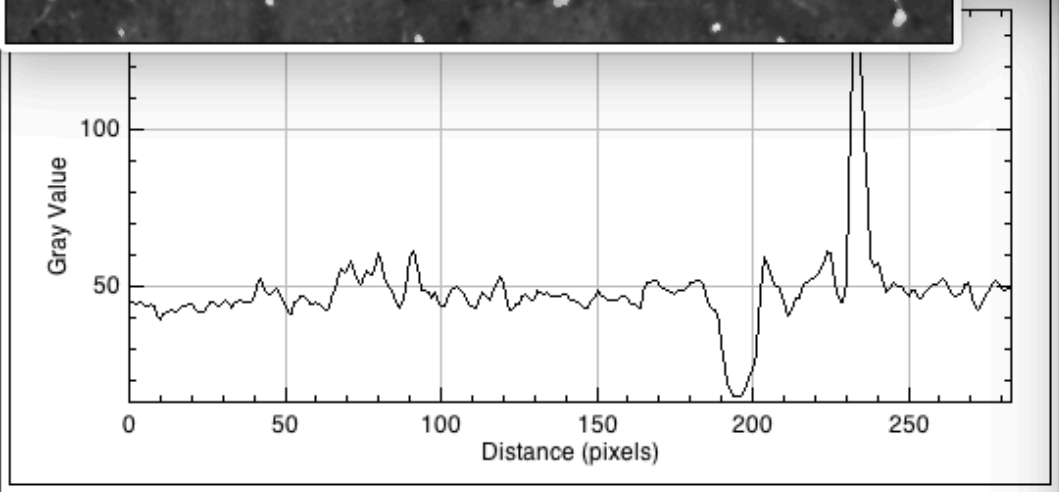
☑ White objects on black background

☐ SetThreshold instead of Threshold (single images)

☐ Show threshold values in log window

The thresholded result of 8 & 16 bit images is shown in white [255] in 8 bits.
For 16 bit images, results of 'Try all' and single slices of a stack are shown in white [65535] in 16 bits.
Unsuccessfully thresholded images are left unchanged.

Cancel OK

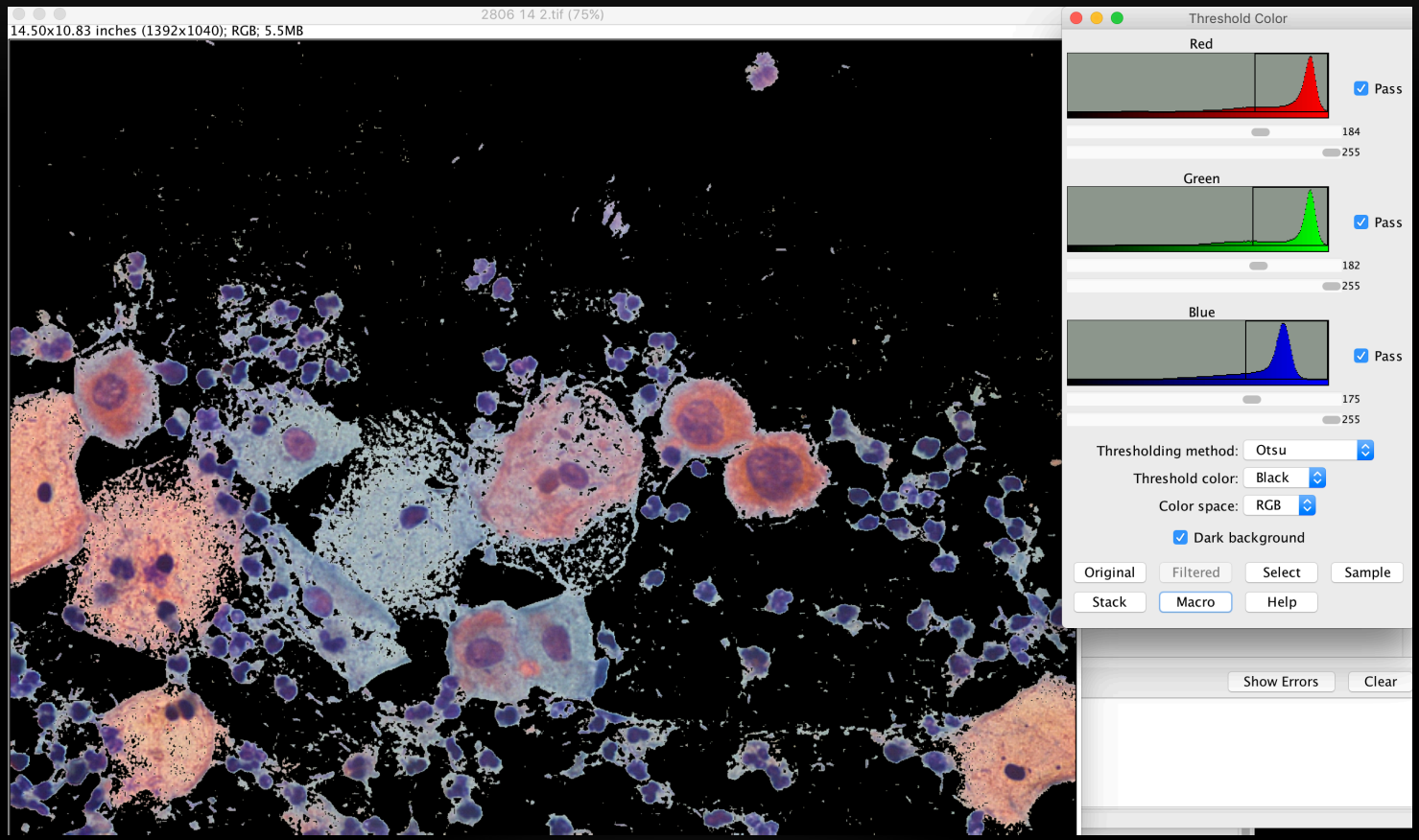


List Save... More » Live

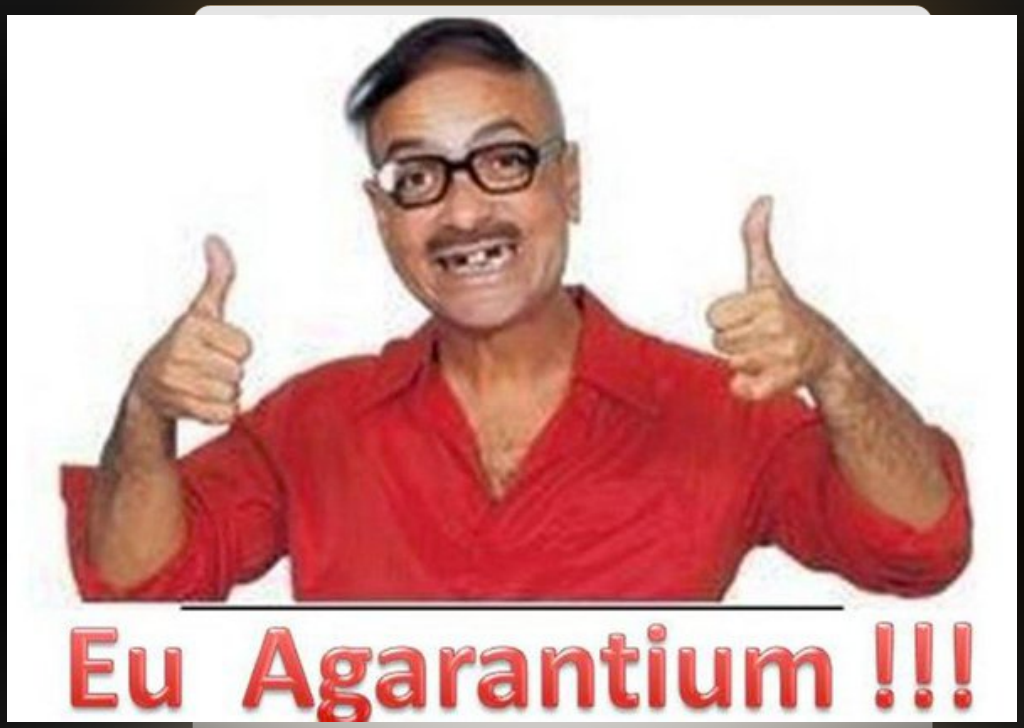
Run Kill

Started Macro.ijm.ijm at Sat May 20 14:34:54 PDT
Started Macro.ijm.ijm at Sat May 20 14:53:21 PDT
Started Macro.ijm.ijm at Sat May 20 15:02:59 PDT

FIND THRESHOLD FOR COLOR CHANNELS



Seus problemas acabaram!



WHAT IF THE CLICK GENERATES **NO** CODE?

You do it yourself:

```
open(“c://mypath//myfolder//image.png”);
run("Duplicate...", "title=myNewImage") ;
```

You might need to clean up the code as in the threshold example:

```
run("8-bit");
setAutoThreshold("Default");
//run("Threshold...");
setAutoThreshold("Default dark");
//setThreshold(182, 255);setOption("BlackBackground", true);run("Convert to Mask");
```

COMMON PITFALLS –LIMITATIONS AND ISSUES

- Slower than you want: Fiji macros are interpreted
- Even slower when flashing pictures: use setBatchMode(true)
- Race conditions: programs that make use of **multiple threads** are subject to **race conditions**
- It's open source and dynamically ingesting community code, therefore…
- "*ImageJ is not responding*"

4. IMAGE SEGMENTATION

a) ~~Thresholding~~

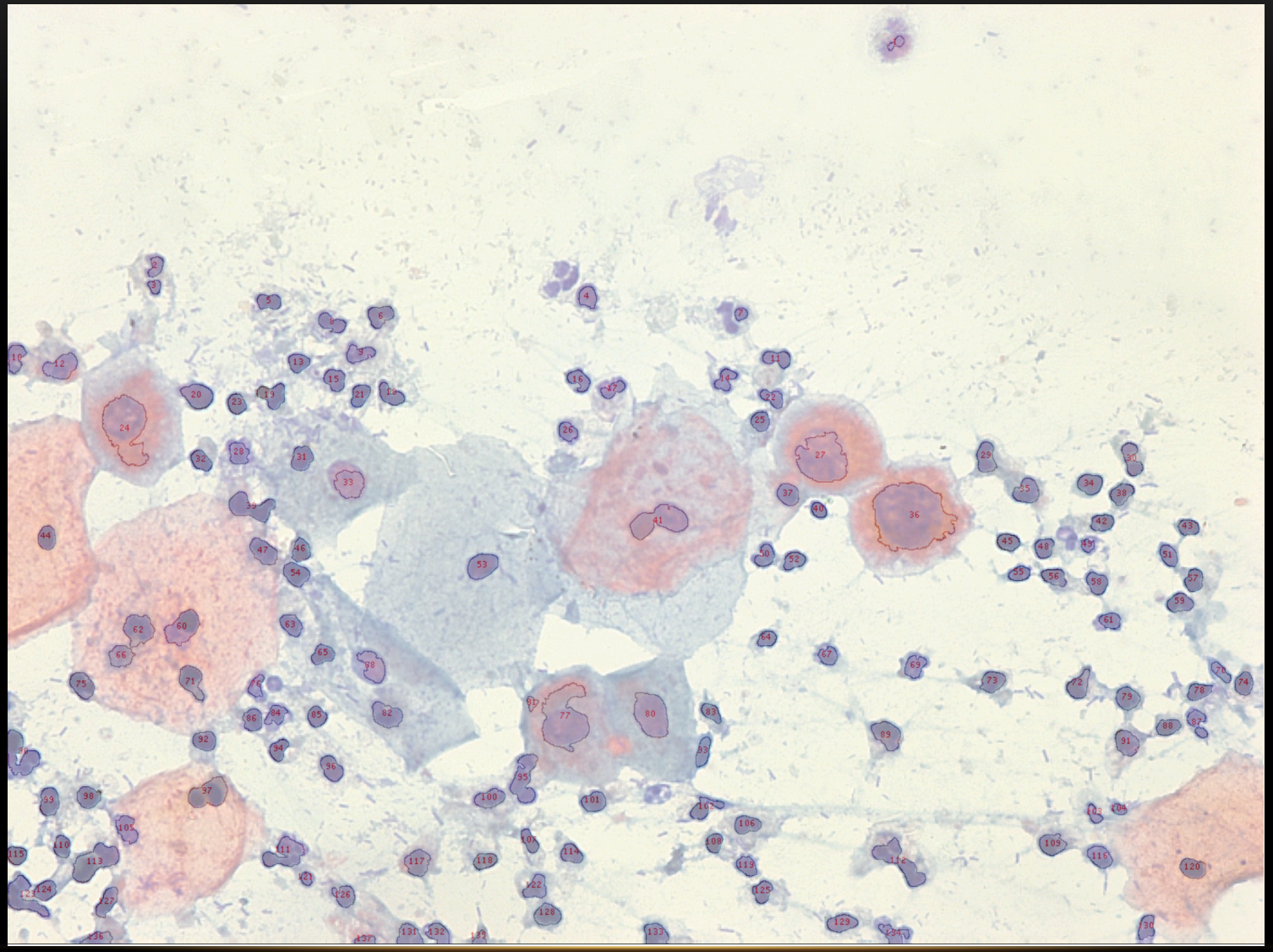
b) Statistical region merging

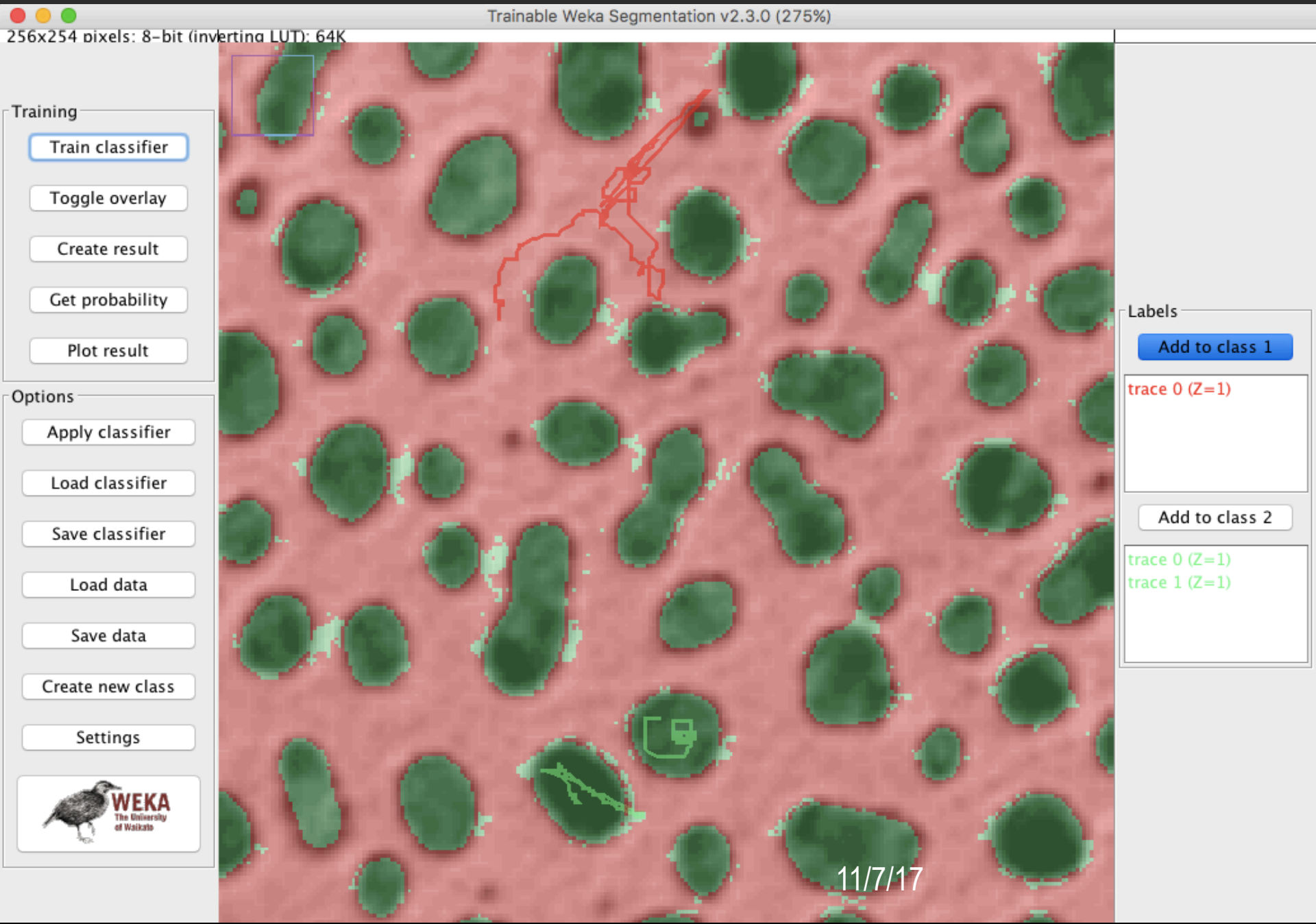
c) Super-pixel

d) Machine learning and images

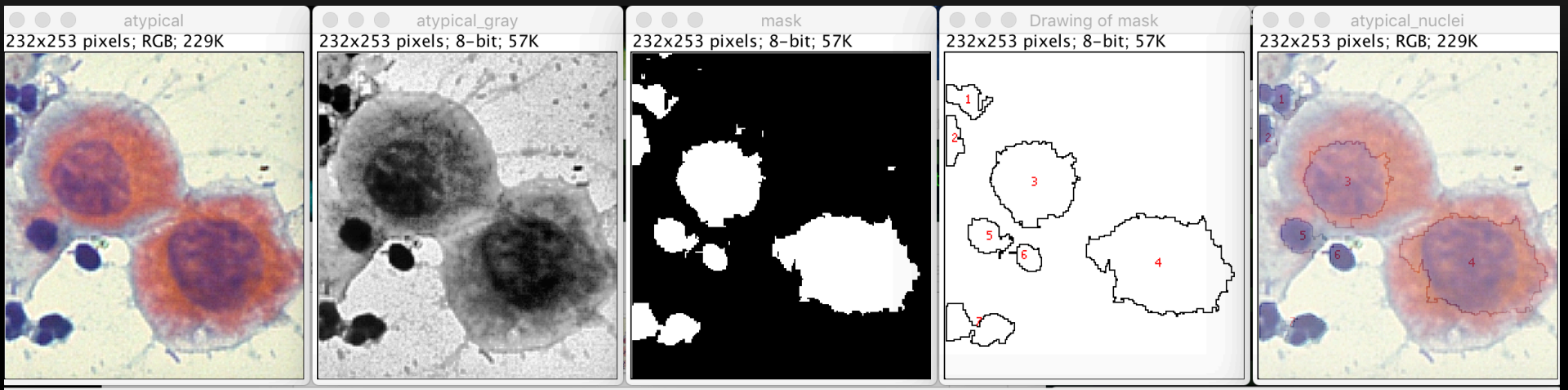
e) Weka

RESULT OF NUCLEI DETECTION




Trainable Weka Segmentation v2.3.0 (275%)
256x254 pixels; 8-bit (inverting LUT); 64K
Training
Train classifier
Toggle overlay
Create result
Get probability
Plot result
Options
Apply classifier
Load classifier
Save classifier
Load data
Save data
Create new class
Settings
WEKA
The University of Waikato
Labels
Add to class 1
trace 0 (Z=1)
Add to class 2
trace 0 (Z=1)
trace 1 (Z=1)
11/7/17

5. OBJECT DESCRIPTION

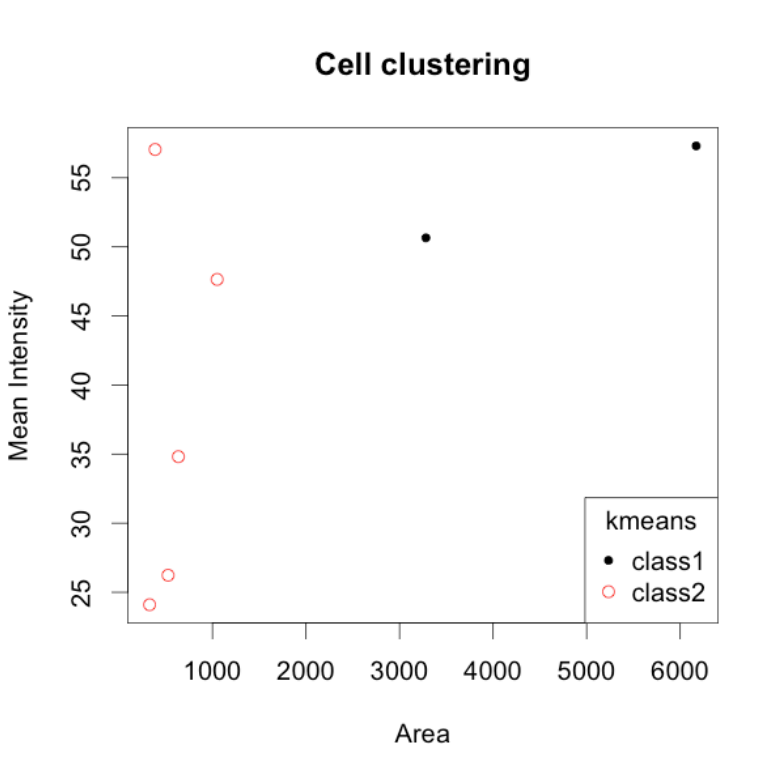


quantMic_results.ijm

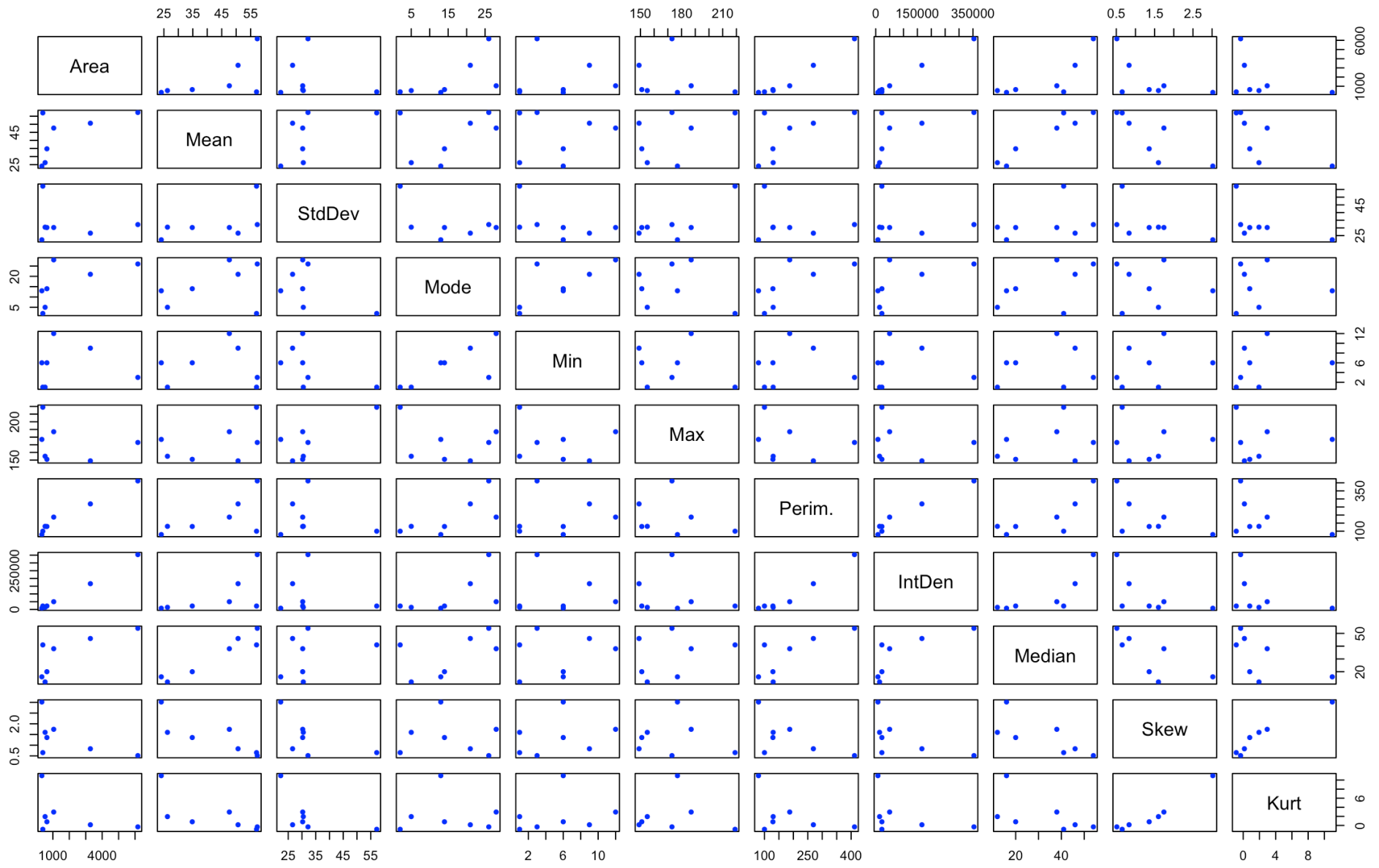
quantMic_results.ijm

```
45    }
46 */
47 setOption("BlackBackground", true);
48 setThreshold(0,115);
49 run("Convert to Mask");
50 rename("mask");
51 wait(500);
52 //Measurements
53 run("Set Measurements...", "area mean standard modal min perimeter integrated median skewness kurtosis stack redirect="+"atypical_gray"+" decimal=3");
54 run("Analyze Particles...", "size="+100+"-Infinity pixel show=Outlines display clear include");
55 //Overlay
56 selectWindow("atypical");
57 run("Duplicate...", "title=atypical_nuclei");
58 run("Add Image...", "image=[Drawing of mask] x=0 y=0 opacity=30 zero");
59 run("Tile");
60 saveAs("Results", path+"quantMic_results.csv");
61
```

6. OBJECT CLASSIFICATION




```
# Plot the measuruments extract from Fiji

filepath = "quantMic_results.csv";

outputdir = ”c:\metrics\”

d=read.csv(filepath,sep=",",head=T);

km=kmeans(dnew,iter.max=100,2,algorithm='L')

plot(d$Area,d$Mean, col=km$cluster,xlab =
"Area", ylab = "Mean Intensity", main = "Cell
clustering",pch=km$cluster+19)

legend('bottomright',
legend=c("class1","class2"), col=c(1,2),
pch=c(20,21), title="kmeans")
```


Area
Mean
StdDev
Mode
Min
Max
Perim.
IntDen
Median
Skew
Kurt

PYTHON

- Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte;
- Lançada por Guido van Rossum em 1991;
- Maneira mais rapida de comecar:
  1. Instalar Anaconda
  2. Entrar no Anaconda Navigator e instalar pacotes
  3. Abrir um Jupyter Notebook

PYTHON ESSENTIALS

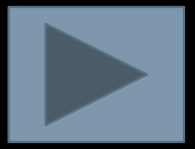
Image exploration using Python - essentials

1. Read image from web
2. Querying image: matrix, sub-matrices, ROI
3. Image transformations: filtering
4. Immunohistochemistry example from scikit-image
5. Segmentation and feature extraction
6. Save information as a xls file
7. Simulating 2D images - "cells"
8. Simulate particles
9. Check particle neighborhood: groups (clustering algorithms)
10. Pandas and Seaborn

1. Read image from web and scikit-image

In [1]:
```
%matplotlib inline
import numpy as np

from matplotlib import pyplot as plt
from skimage import data, io
```

In [2]:
```
fname = 'http://petfarmacia.ufop.br/sites/default/files/styles/media_gallery
imageFromWeb = data.imread(fname, as_grey=False, plugin=None, flatten=None)
plt.imshow(imageFromWeb)
```



TAKE HOME MESSAGE

New data regimes:

- High-throughput, high-resolution instruments;
- Multimodal experiments and massive datasets;
- Searching, ranking, benchmarking, reproducibility;

Hybrid computing architectures:

- Data and task parallel;
- GPU, Neuromorphic, TPU, FPGA.

DISCLAIMER

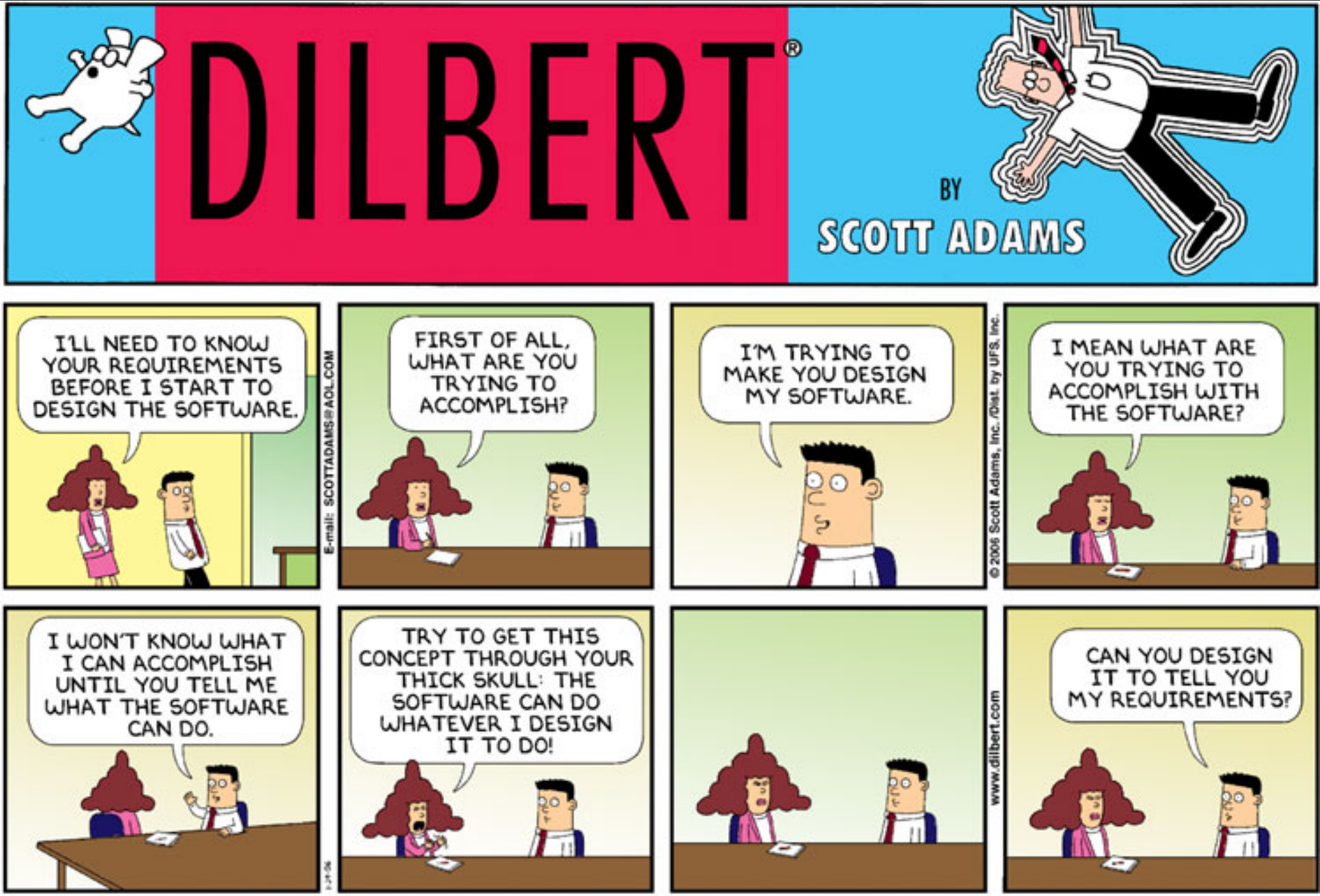
- This presentation was prepared with basis on work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

THANKS

DANI.LBNL@GMAIL.COM

SELECTED JOURNALS

- Alegro, Theofilas, Nguy, Castruita, Seeley, Ushizima, Grinberg, *Automating Cell Detection and Classification in Human Brain Fluorescent Microscopy Images Using Dictionary Learning and Sparse Coding*, Journal of Neuroscience Methods, 2017.
- Ushizima, Bale, Bethel, Ercius, Helms, Krishnam, Grinberg, Haranczyk, Macdowell, Odziomek, Parkinson, Perciano, Ritchie, Yang. *IDEAL: Images across Domains, Experiments, Algorithms and Learning*, Journal of Minerals, Metals and Materials, 2016, 68(11), 2963-2972.
- Odziomek, Ushizima, Oberbek, Kurzydlowski, Puzyn, *Scanning electron microscopy image representativeness: morphological data on nanoparticles*, Journal of Microscopy, 2016.
- Santos, Bianchi, Ushizima, Pavinatto, Bianchi, *Ammonia gas sensor based on the frequency-dependent impedance characteristics of ultrathin polyaniline films*, Sensors and Actuators A: Physical, 2016.
- Venkatakrishnan, Mohan, Beattie, Correa, Dart, Deslippe, Hexemer, Krishnan, MacDowell, Marchesini, Patton, Perciano, Sethian, Stromsness, Tierney, Ushizima, Parkinson, *Making Advanced Scientific Algorithms and Big Scientific Data Management More Accessible*, Electronic Imaging, 19, pp.1-7, 2016.
- Wills, Michalak, Ercius, Rosenberg, Perciano, Ushizima, Runser, Helms, *Block Copolymer Packing Limits and Interfacial Reconfigurability in the Assembly of Periodic Mesoporous Organosilicas*. Advanced Functional Materials, 2015.
- Donatelli, Haranczyk, Hexemer, Krishnan, Li, Lin, Maia, Marchesini, Parkinson, Perciano, Shapiro, Ushizima, Yang, Sethian. *CAMERA: The Center for Advanced Mathematics for Energy Research Applications*. Synchrotron Radiation News, 28:2, 2015.

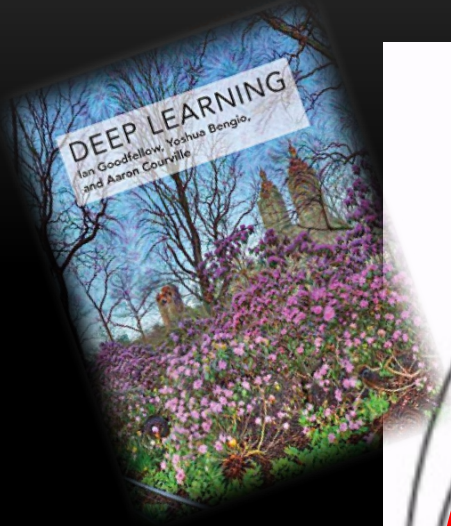


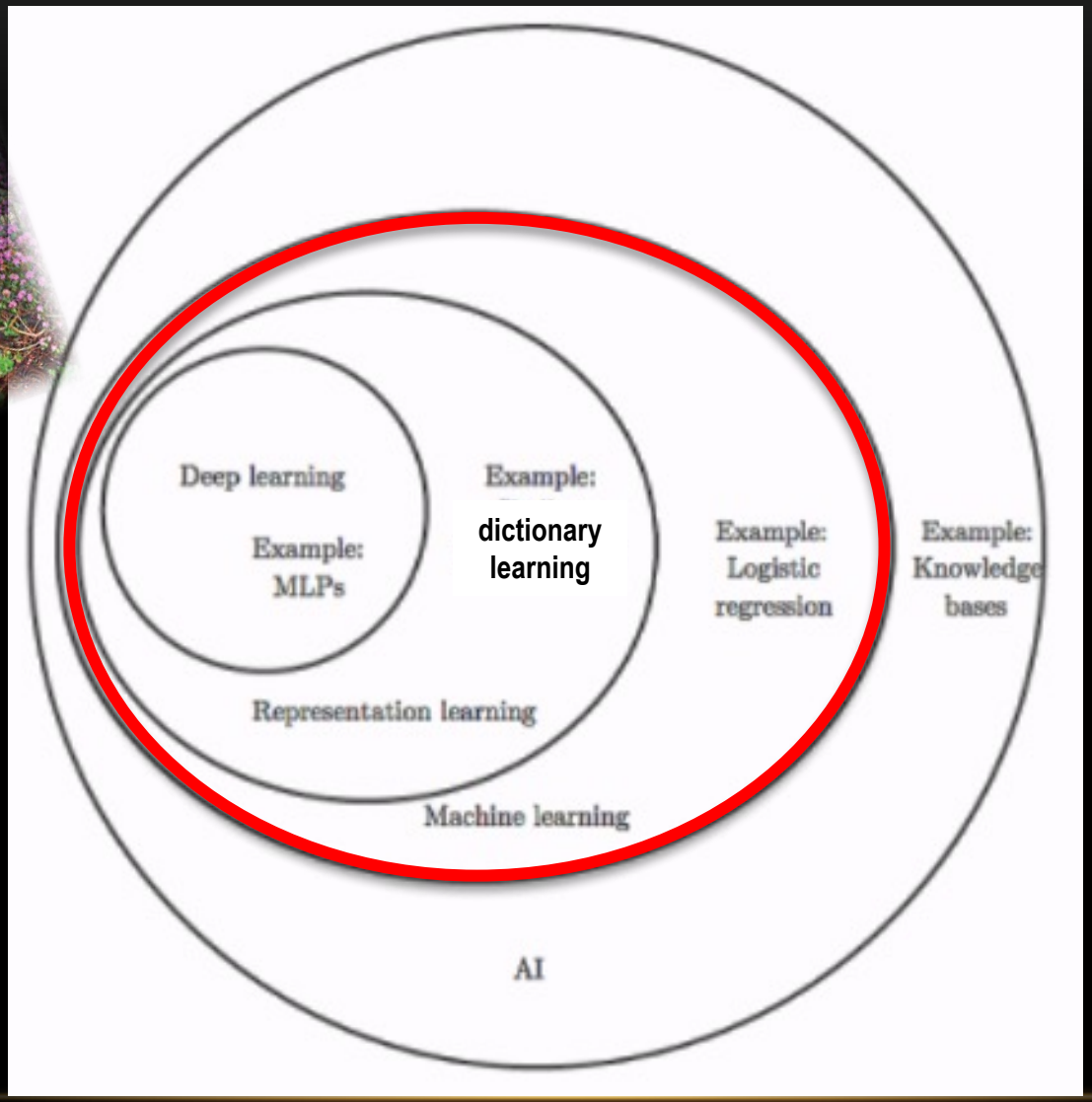
WHAT IS MACHINE LEARNING?

- Machine learning is the capability of extracting patterns from raw data without the need to hard code a knowledge base for use with logical inference rules. – *Goodfellow, Bengio and Courville.*

- Machine learning is a type of artificial intelligence (AI) that allows computers to learn without being explicitly programmed. -*Whatisit*

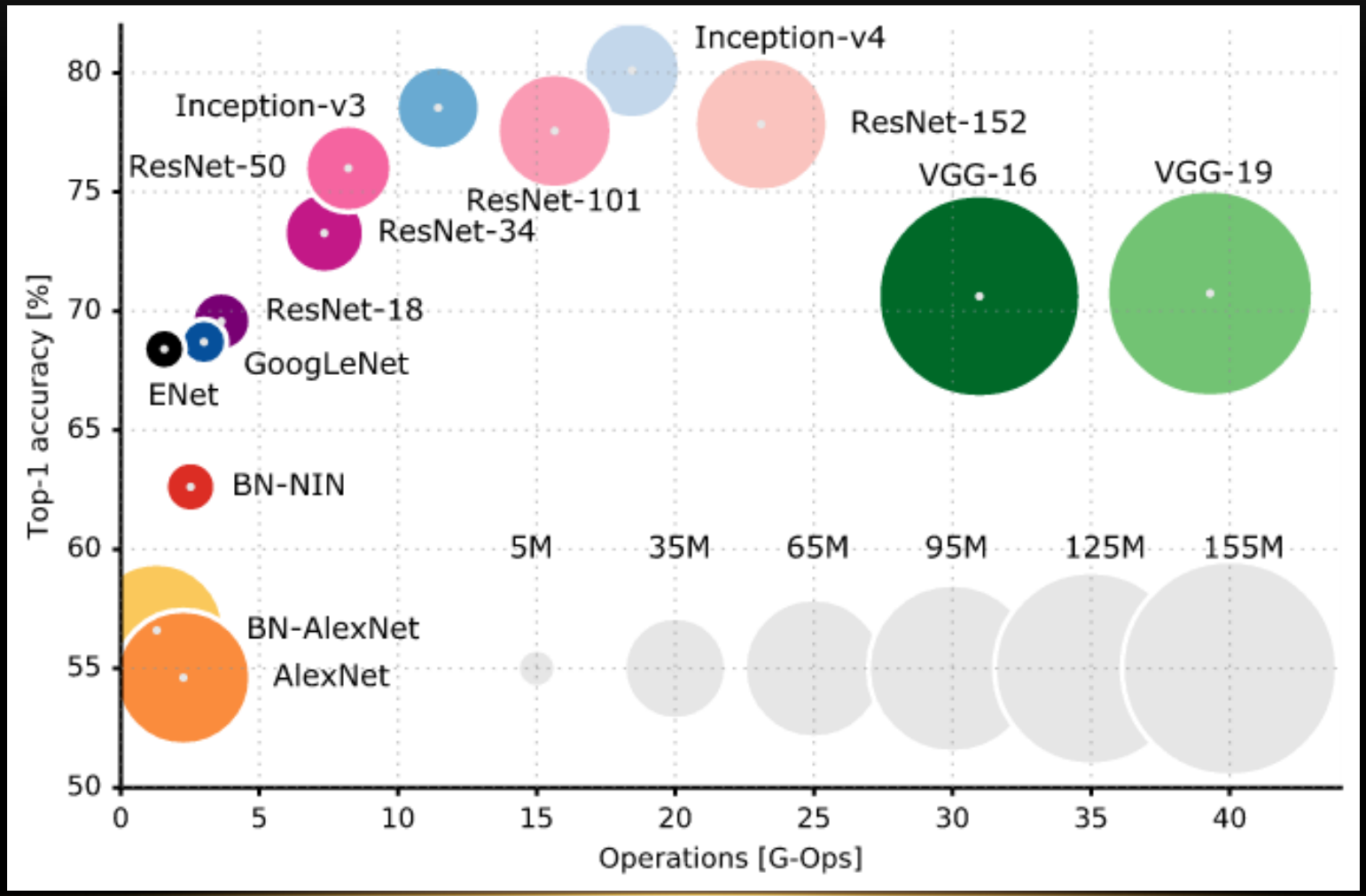
ML = DEEP LEARNING?

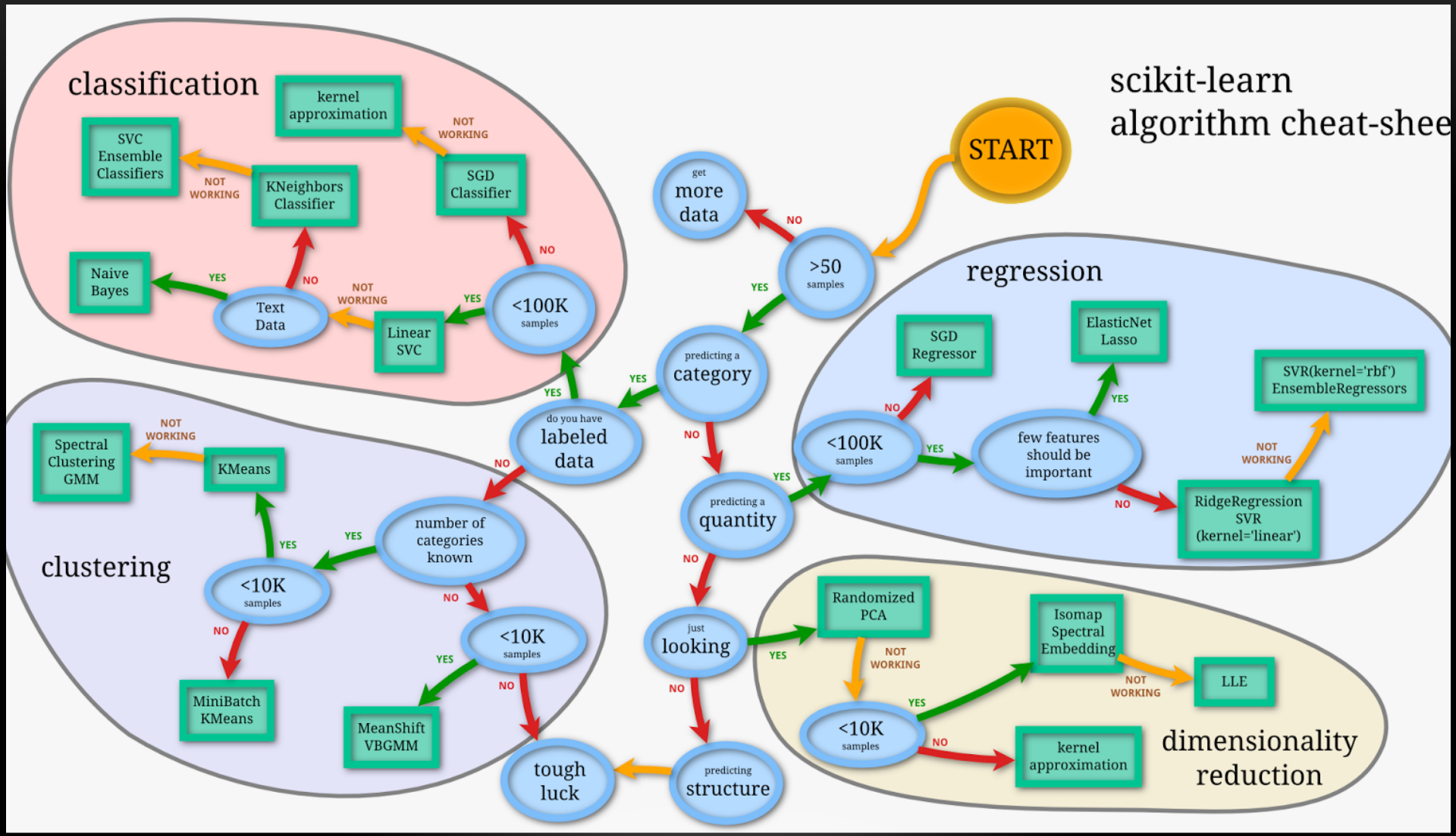




According to Goodfellow, Bengio and Courville

WHY THE BUZZ NOW?




scikit-learn
algorithm cheat-shee
START
get more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
just looking
YES
NO
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
YES
regression
SGD Regressor
NO
ElasticNet Lasso
YES
SVR(kernel='rbf') EnsembleRegressors
<100K samples
YES
few features should be important
NO
NOT WORKING
RidgeRegression SVR (kernel='linear')
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift VBGMM
Randomized PCA
YES
NOT WORKING
Isomap Spectral Embedding
NOT WORKING
LLE
<10K samples
YES
NO
kernel approximation
dimensionality reduction


How Do I Get Started?
Frustration!!!
I tried a textbook
I tried a MOOC
I tried blog posts
I tried Kaggle
Desire
Now: I'm a dev and I can ship solutions
Soon: I'm a dev that can do ML, I can create predictive models
One-off predictive models
Models embedded in software
Deep models for core business
Future: I'm a data scientist, I think up things to model...
Traditional answer...
Learn linear algebra
Learn probability
Learn calculus
Learn multivariate stats
Learn information theory
...now you're ready to learn ML algorithms
Tradition answer... is broken
Bottom-up approach
Slow
Focused on algorithms
Gap, never gets to delivering solutions
A better answer...
1. Select a systematic process for working through problems
2. Map process onto best tools
3. Targeted practice
Well understood datasets
Build a portfolio of work to show-off your work
A better answer... for devs!
Do not HAVE to write code
Do not need to be good at math
Do not need a higher degree
Do not need a super computer
Do not need large blocks of time
Do not need big data
Biggest mistakes
Not taking action! (so take action!)
Picking problems that are too big/hard (stay small)
Implementing algorithms from scratch (use tools!)
Not sticking to the process (be systematic!)
Not using resources (read, a lot!)

CR