DETECTANDO AS FACES DE IMAGENS CIENTÍFICAS

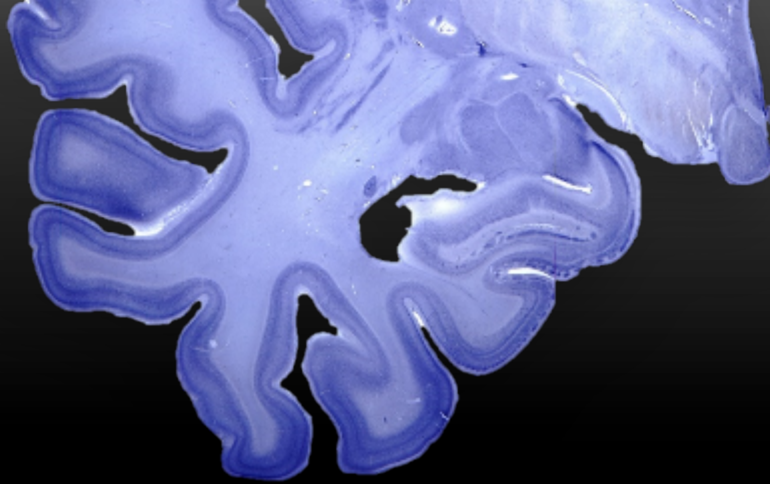
COM REDES CONVOLUCIONAIS

Dani Ushizima, Ph.D.

Staff Scientist - CRD, Lawrence Berkeley National Lab.

Data Scientist - BIDS, University of California, Berkeley

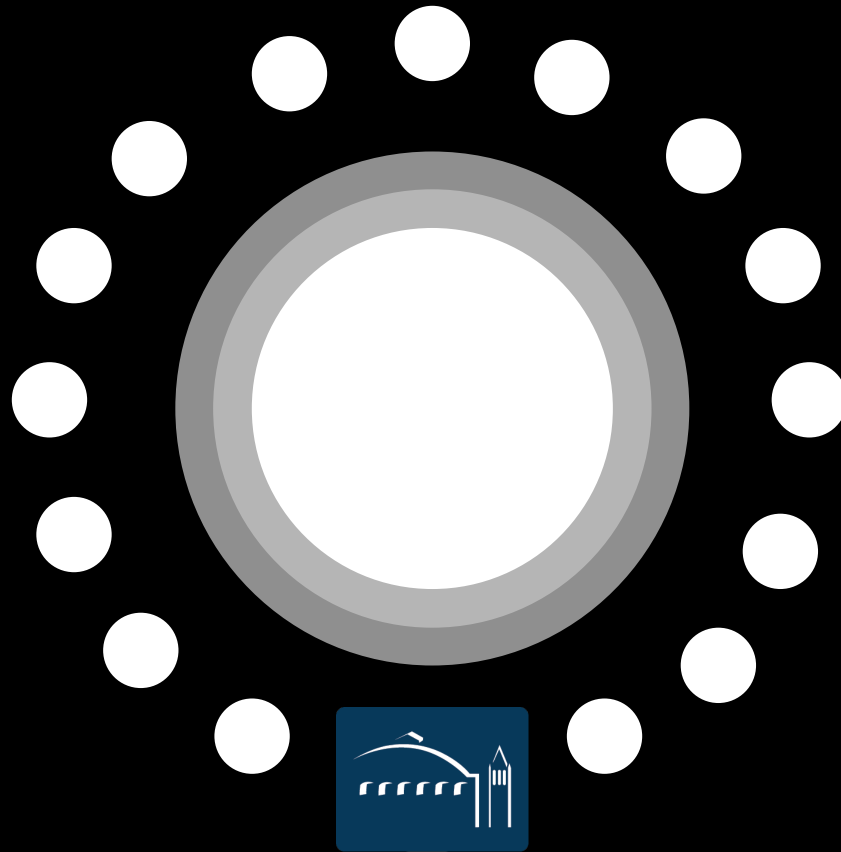
OUTLINE



1. Lawrence Berkeley National Lab
2. Segredos e super poderes
3. Projetos em visão computacional
4. Força tarefa “Made in Piauí”
5. Missão: busca de imagem por conteúdo
6. Oportunidades de estágio no exterior

BERKELEY LAB
Bringing Science Solutions to the World

O QUE É O 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

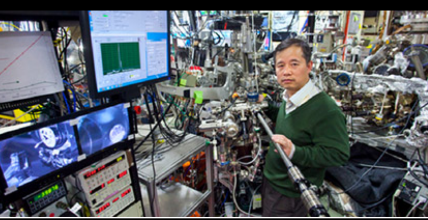
- 962 Scientists
- 740 Engineers
- 473 Postdoctoral Associates
- 644 Research Support
- 296 Graduate Students
- 141 Undergraduate Students
- 758 Operations/Administrative Support

4014

BERKELEY LAB

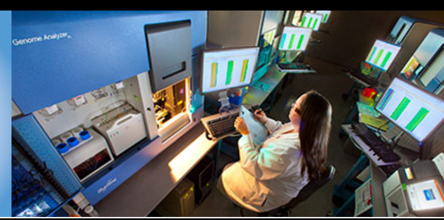
NATIONAL USER FACILITIES

ALS

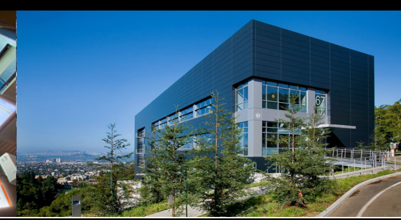


ESnet

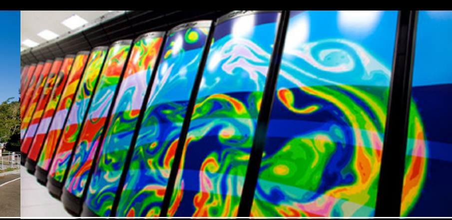
JGI



Molecular Foundry



NERSC



9,911 external scientific facility users per year

EXCELLENCE

13

70

- 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

14

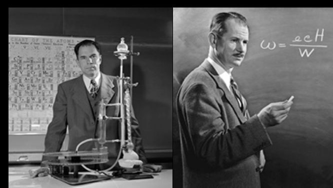
18

16

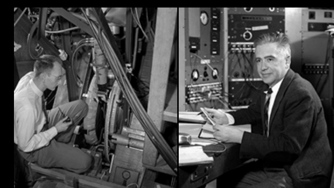
BERKELEY LAB

NOBEL LAUREATES

E.O. Lawrence
1939: Physics

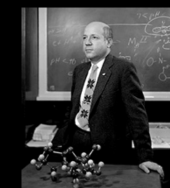


Glenn T. Seaborg and Edwin M. McMillan
1951: Chemistry



Owen Chamberlain and Emilio Segrè
1959: Physics

Donald Glaser
1960: Physics



Melvin Calvin
1961: Chemistry

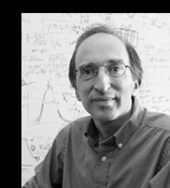
Luis W. Alvarez
1968: Physics

Yuan T. Lee
1986: Chemistry

Steven Chu
1997: Physics

George F. Smoot III
2006: Physics

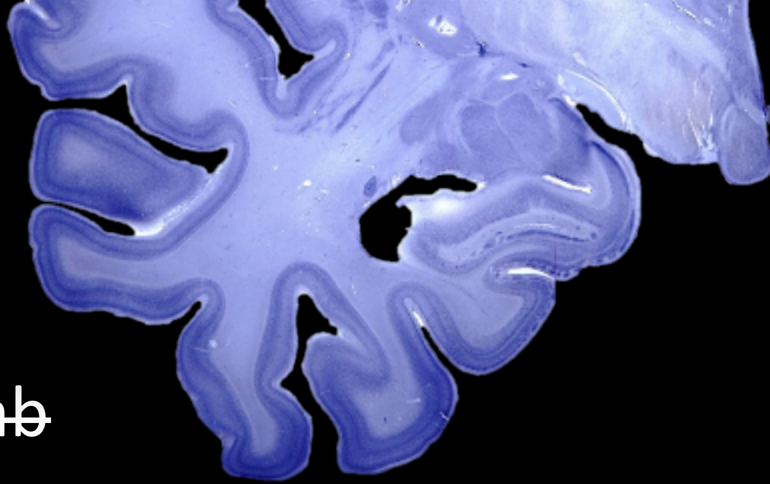
Intergovernmental Panel on Climate Change
2007: Peace



Saul Perlmutter
2011: Physics

BERKELEY LAB

Sumário

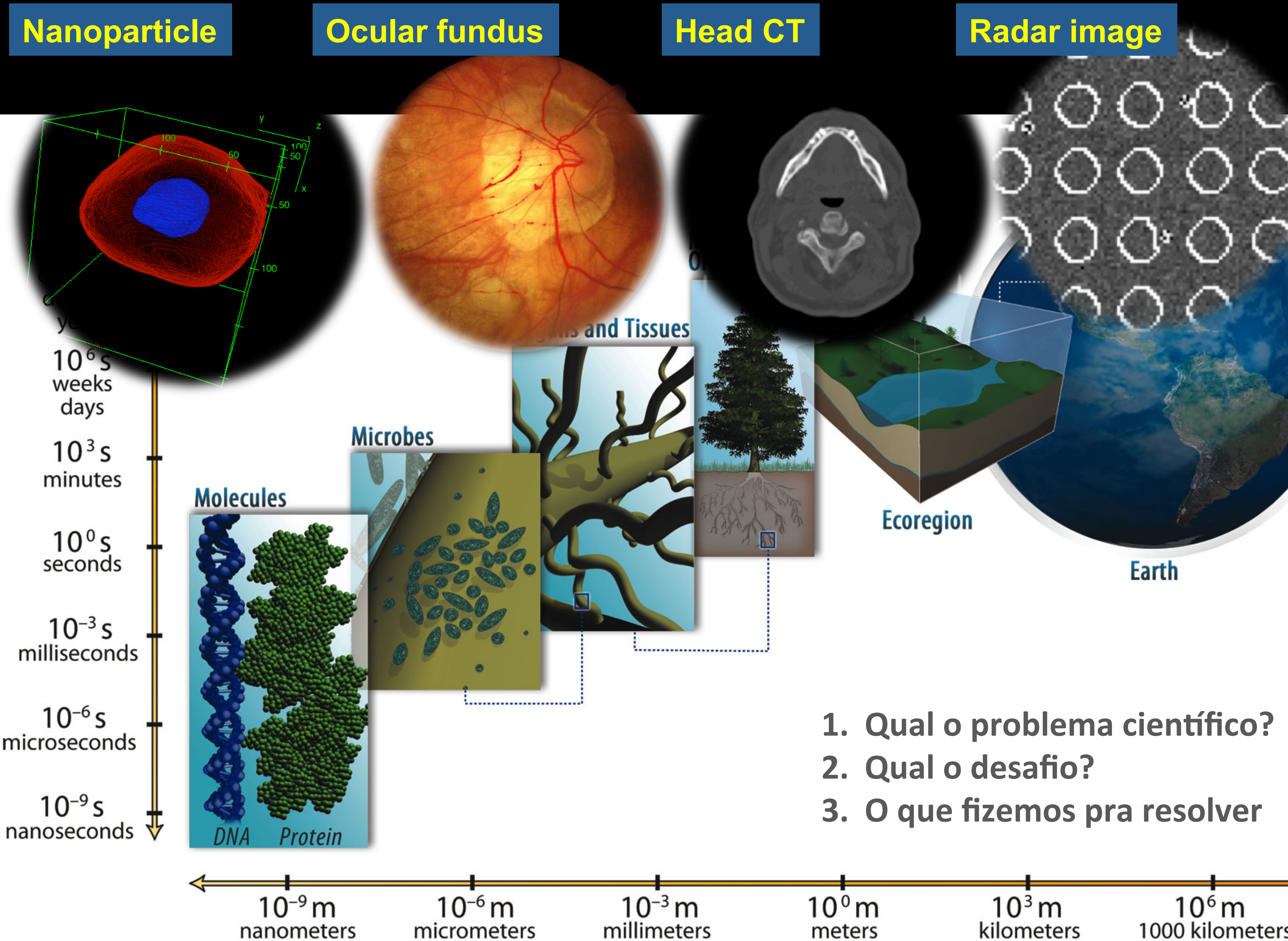


1. ~~Lawrence Berkeley National Lab~~
2. Segredos e super poderes
3. Projetos em visão computacional
4. Força tarefa “Made in Piauí”
5. Missão: busca de imagem por conteúdo
6. Oportunidades de estágio no exterior

Minha busca por energia, poderes extraordinários e engenhocas

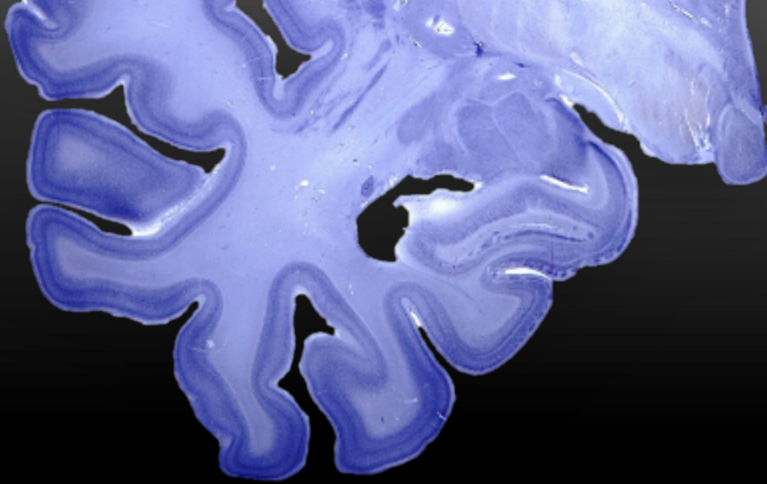
POR QUE TRABALHAR NO BERKELEY LAB, DOE?

Defendendo dados do governo



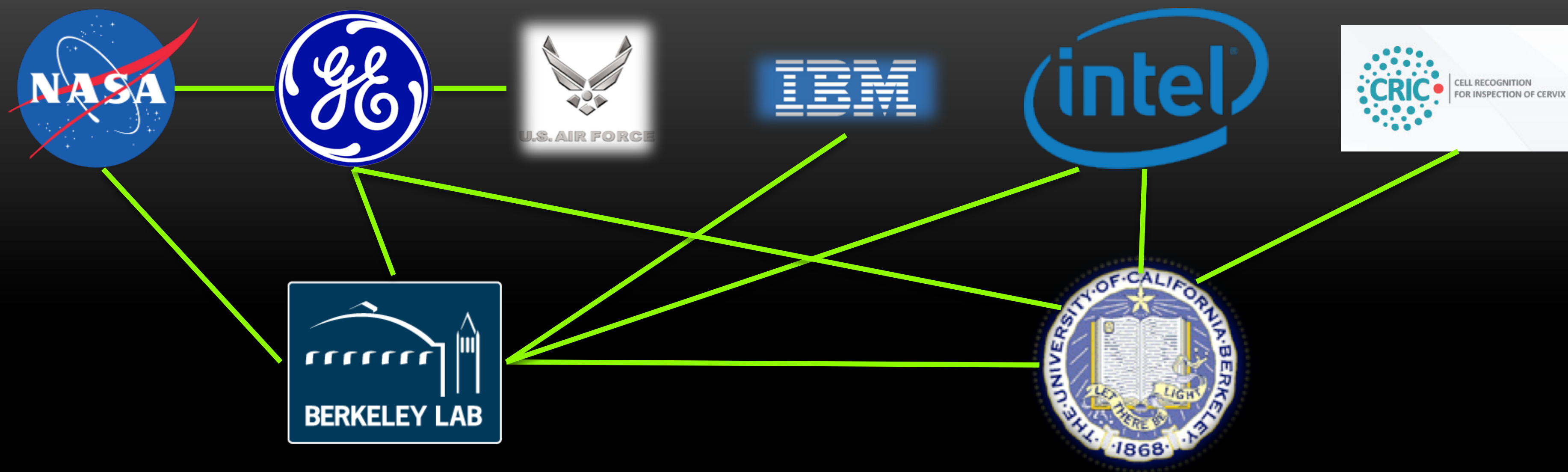
1. Qual o problema científico?
2. Qual o desafio?
3. O que fizemos pra resolver

SUMÁRIO



1. ~~Lawrence Berkeley National Lab~~
2. ~~Segredos e super poderes~~
3. Projetos em visão computacional
4. Força tarefa “Made in Piauí”
5. Missão: busca de imagem por conteúdo
6. Oportunidades de estágio no exterior

SBC'17 ERIPI - DANI USHIZIMA



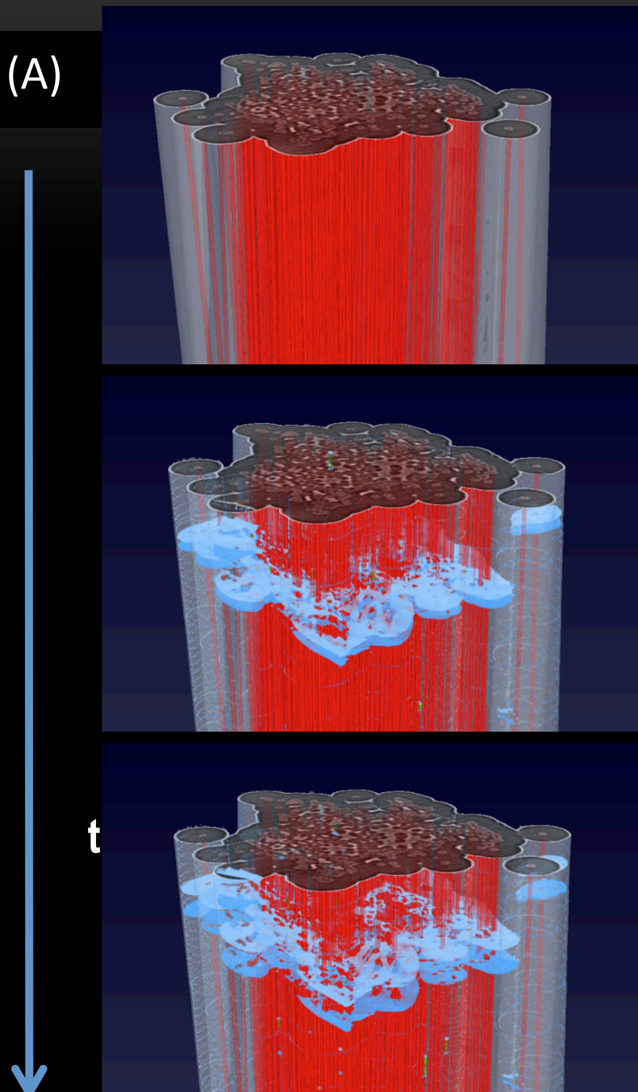
Collaborations among national labs, academia and industry

USE-CASES

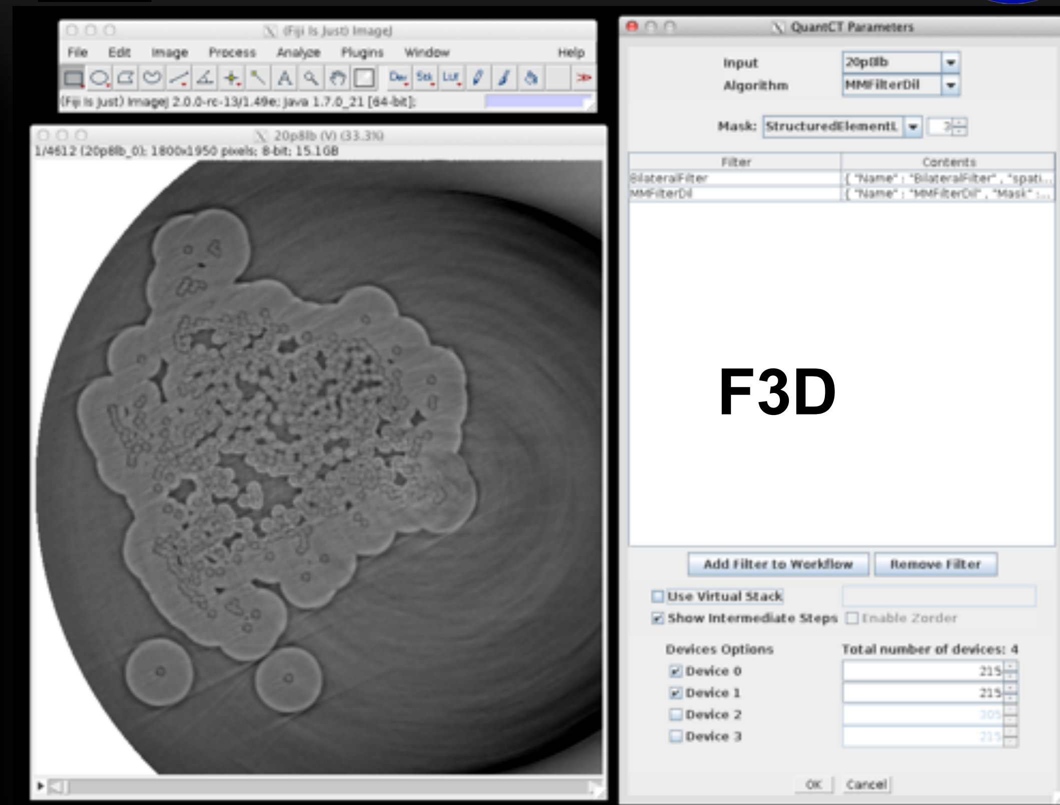
Indústria aeronáutica

1. Quantificação de danos micro-structurais

(A)



(B)



(A) Rendering of samples under deformation after image processing and analysis to detect the matrix (gray) and fibers (red) of a single tow at different applied loads, leading to microfractures (blue); (B) Deployed tool: F3D.

. Ushizima, Perciano, Krishnan, Parkinson. "Fast Detection of Material Deformation through Structural Dissimilarity", IEEE Int. Conf on Big Data, 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.

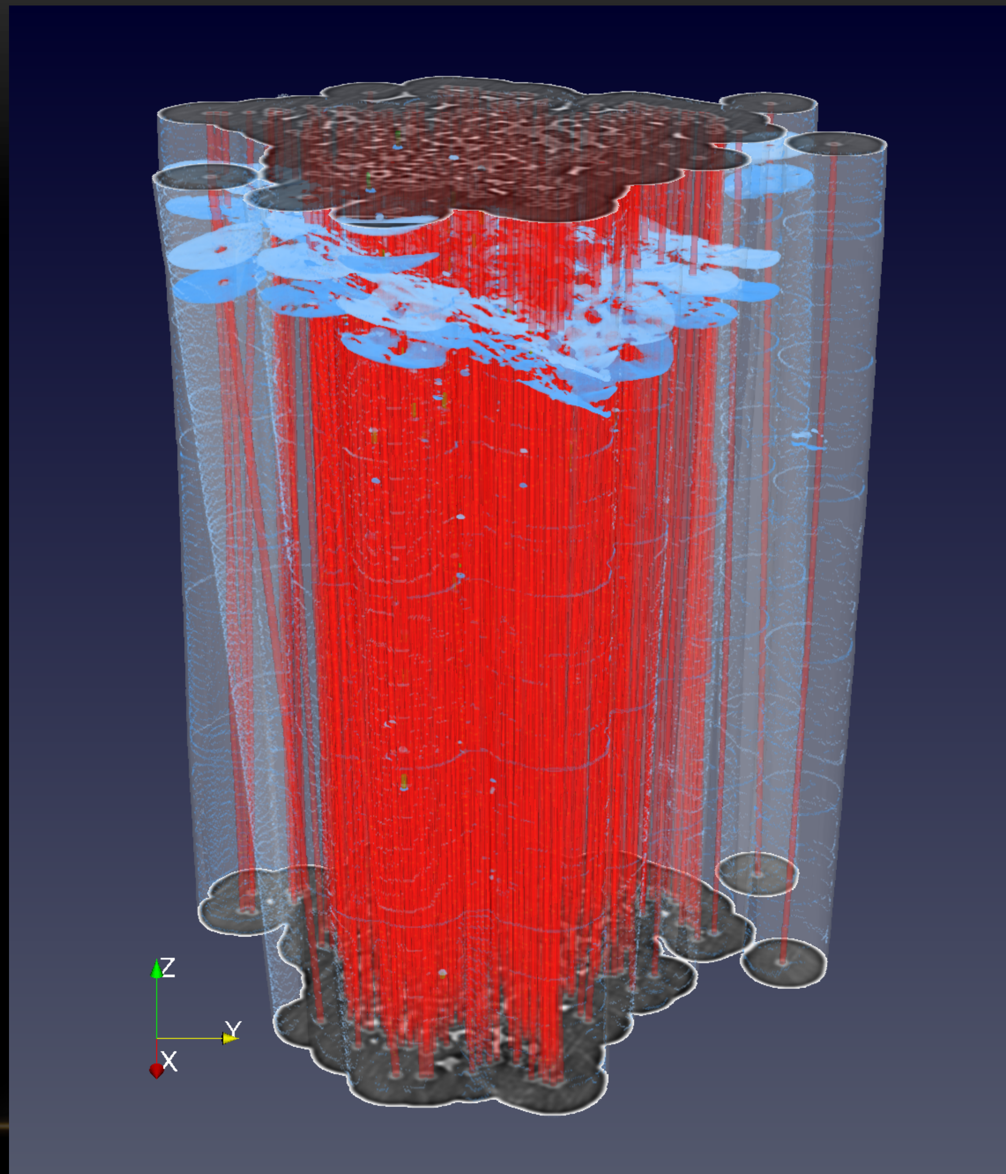
Advanced technique: team work

CAMERA Applied Math

ALS

DAV

ECRP-2015



Novos microeletrônicos

2. ESPECIFICAÇÃO DE FILMES E CARACTERIZAÇÃO

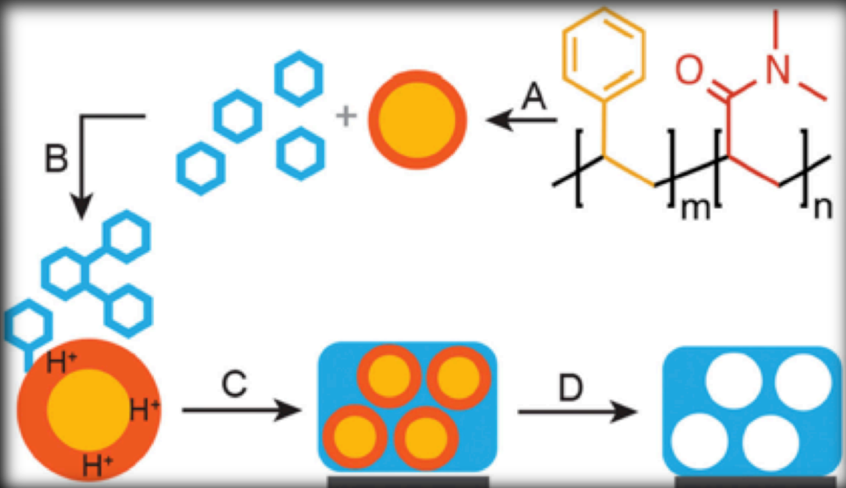
Porogens

Block copolymer

Right properties?

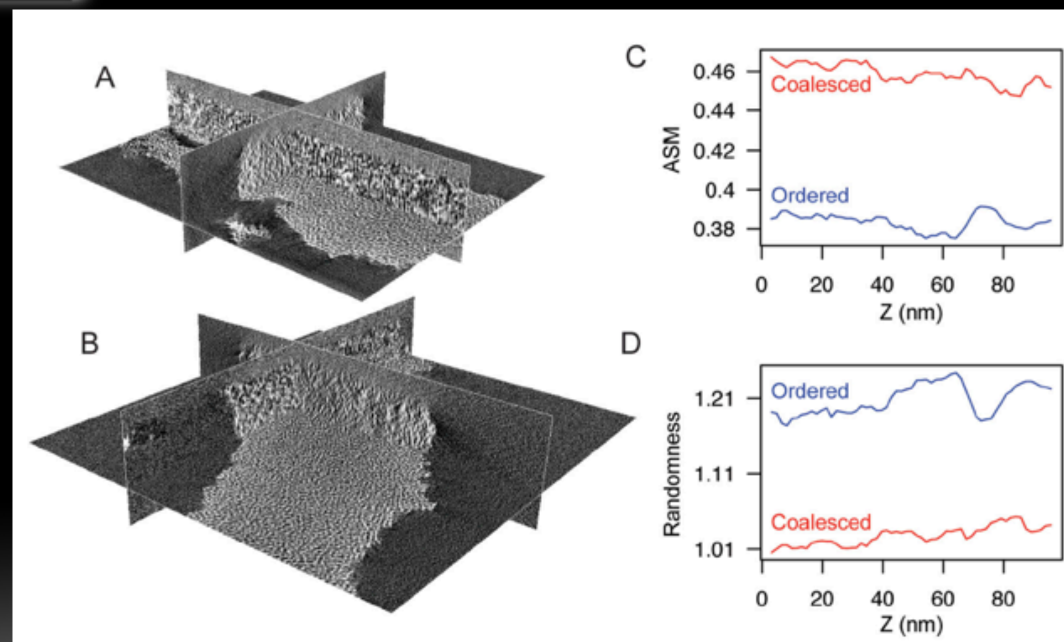
Perfect PMO

Controle de qualidade durante montagem de filmes



- Porosimetry for finding the lowest ever dielectric constants for PMO matrix material, used in microelectronics;
- Delivered analysis method, second-order statistics of

Metrology of material architecture and porosimetry



. Wills, Michalak, Ercius, Rosenberg, Perciano, Ushizima, Runser and Helms. “Block Copolymer Packing Limits and Interfacial Recongurability in the Assembly of Periodic Mesoporous Organosilicas”. Advanced Functional Materials, 4120-4128, 2015.

ASSEMBLY AND THERMAL PROCESSING OF PMO

Porogens

introduced to

BCP

assemble

Advanced PMO

films

BCP packing geometries?

analysis

Instrumentation

1 of data modalities

STEM tomography

texture analysis

ock Copolymer Packing Limits and Interfacial
configurability in the Assembly of Periodic Mesoporous
ganosilicas

rew W. Wills, David J. Michalak, Peter Ercius, Ethan R. Rosenberg, Talita Perciano,
iela Ushizima, Rory Runser, and Brett A. Helms*

2.C. WHEN COMPLEX STRUCTURES ARE HARD TO OBSERVE

$$f(i,j|d,\theta) = \sharp\{(i,j) \in I_g \mid j = \rho(i|d,\theta)\}$$

$$\begin{array}{rcc}
\rho(i|d,\theta = 0^o) \rightarrow & \bar{x} = x + d; & \bar{y} = y; \\
\rho(i|d,\theta = 45^o) \rightarrow & \bar{x} = x + d; & \bar{y} = y + d; \\
\rho(i|d,\theta = 90^o) \rightarrow & \bar{x} = x; & \bar{y} = y + d; \\
\rho(i|d,\theta = 135^o) \rightarrow & \bar{x} = x - d; & \bar{y} = y + d;
\end{array}$$

- Pore packing relationships through a method called texture analysis;
- Measuring local variations in image intensity which are often too fine to be distinguished as separate objects by the observer;

- Textural features can be statistically evaluated using the joint probability distribution of every two pixels in an image, given a direction (Θ) and distance (here d=1);
- Gray-level co-occurrence matrix (GLCM) (Equations) consider not only the distribution of intensities, but also the spatial organization of the pixels.

UNDERSTANDING GLCM CONSTRUCTION

1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

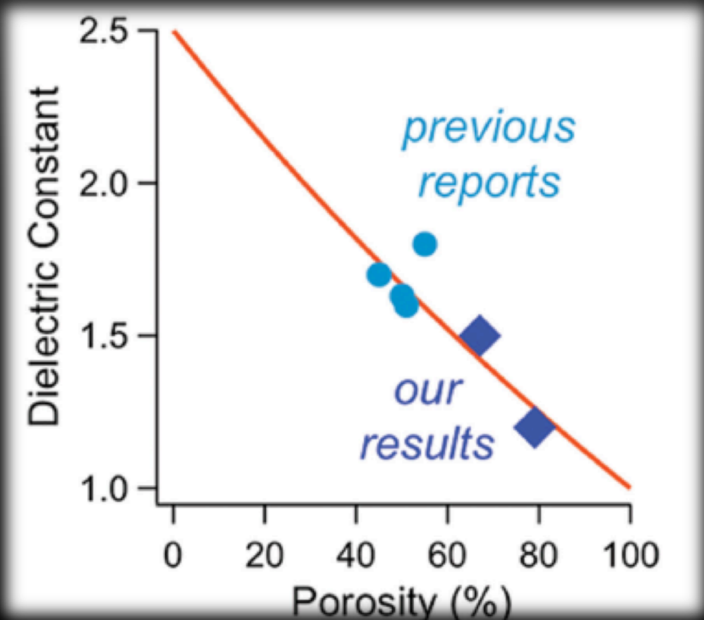
GLCM

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

EXTRACTING FEATURES FROM GLCM

$$
\begin{aligned}
f_1 = \text{Energy} &= \sum_{i,j} \hat{f}(i,j)^2 \quad , \\
f_2 = \text{Entropy} &= -\sum_{i,j} \hat{f}(i,j) \log \hat{f}(i,j) \quad , \\
f_3 = \text{Correlation} &= \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)\hat{f}(i,j)}{\sigma_i \sigma_j} \quad , \\
f_4 = \text{Inverse Difference Moment} &= \sum_{i,j} \frac{1}{1+(i-j)^2} \hat{f}(i,j) \quad , \\
f_5 = \text{Inertia} &= \sum_{i,j} (i-j)^2 \hat{f}(i,j) \quad , \\
f_6 = \text{Cluster Shade} &= \sum_{i,j} ((i-\mu_i)+(j-\mu_j))^3 \hat{f}(i,j) \quad , \\
f_7 = \text{Cluster Prominence} &= \sum_{i,j} ((i-\mu_i)+(j-\mu_j))^4 \hat{f}(i,j) \quad , \\
\text{where } \mu_i &= \sum_i i \sum_j \hat{f}(i,j) \quad , \\
\mu_j &= \sum_j j \sum_i \hat{f}(i,j) \quad , \\
\sigma_i &= \sum_i (i-\mu_i)^2 \sum_j \hat{f}(i,j) \quad , \\
\sigma_j &= \sum_j (j-\mu_j)^2 \sum_i \hat{f}(i,j) \quad , \\
f_8 = \text{Haralick's Correlation} &= \frac{\sum_{i,j} (ij)\hat{f}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}
\end{aligned}
$$

E DAÍ?

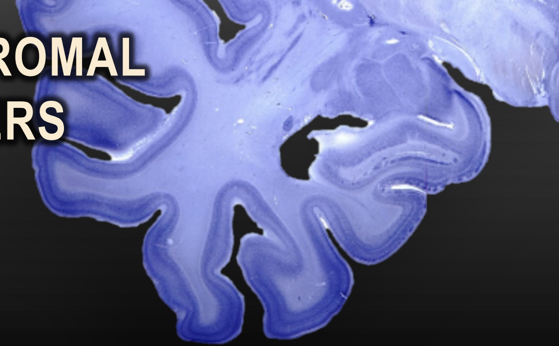


- Teoria e experimentos ligando propriedades dielectricas de organosilica de mesoporo a porosidade total

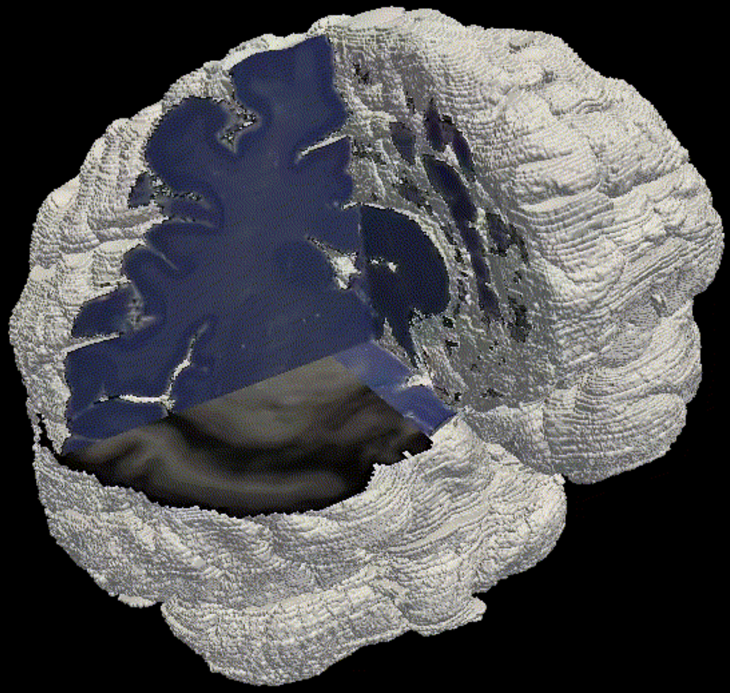
- Prior to our work, PMO porosity had been controlled up to 55% without significant loss in order, resulting in dielectric constants of 1.6–1.8 (aqua dots);
- Theoretical predictions of dielectric constant for a given porosity were made using the Maxwell-Garnett effective-medium approximation (red line);
- Our PMO films: higher porosity yielded dielectric constants of 1.5 (ordered, at packing limit) and 1.2 (disordered, above packing limit) (blue diamonds).

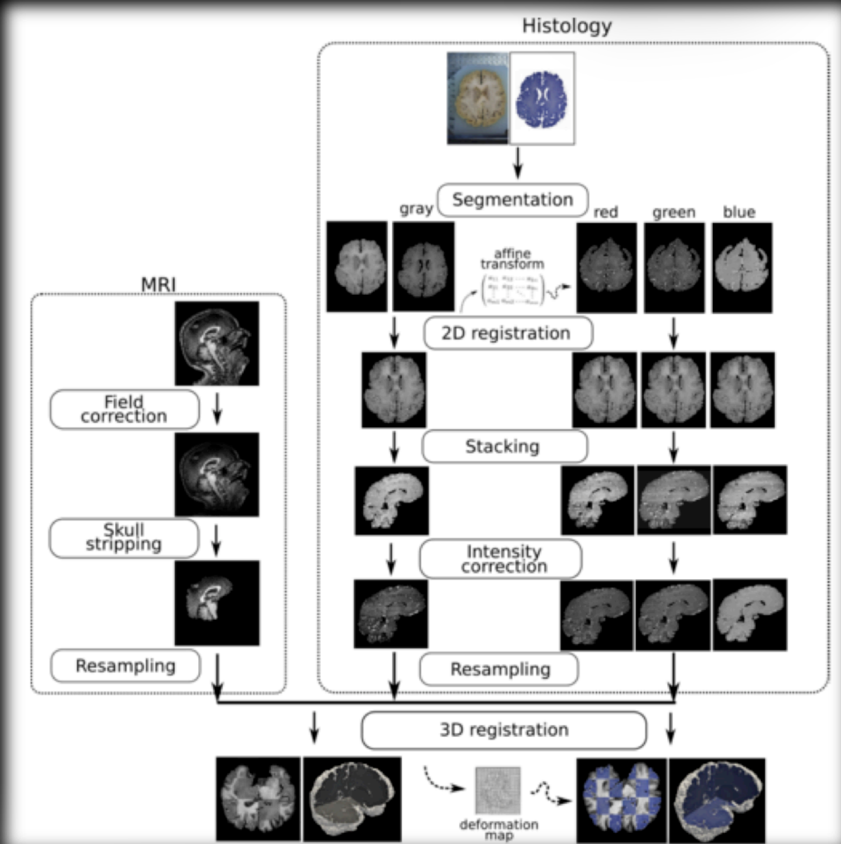
Procurando demência

PATHWAY TOWARD DIAGNOSING AND MONITORING PRODROMAL ALZHEIMER'S DISEASE USING NOVEL IMAGING BIOMARKERS



Lea T. Grinberg, M.D, Ph.D (UCSF), Maryana Alegro, Ph.D (UCSF/LBNL) and Dani Ushizima, Ph.D (LBNL/BIDS)





Alegro, Amaro, Heisen, Loring, Alho, Zollei, Ushizima, Grinberg, “Multimodal whole brain registrations: MRI and high resolution histology”, IEEE CVPR 2016.

SUS, células e

câncer cervical:

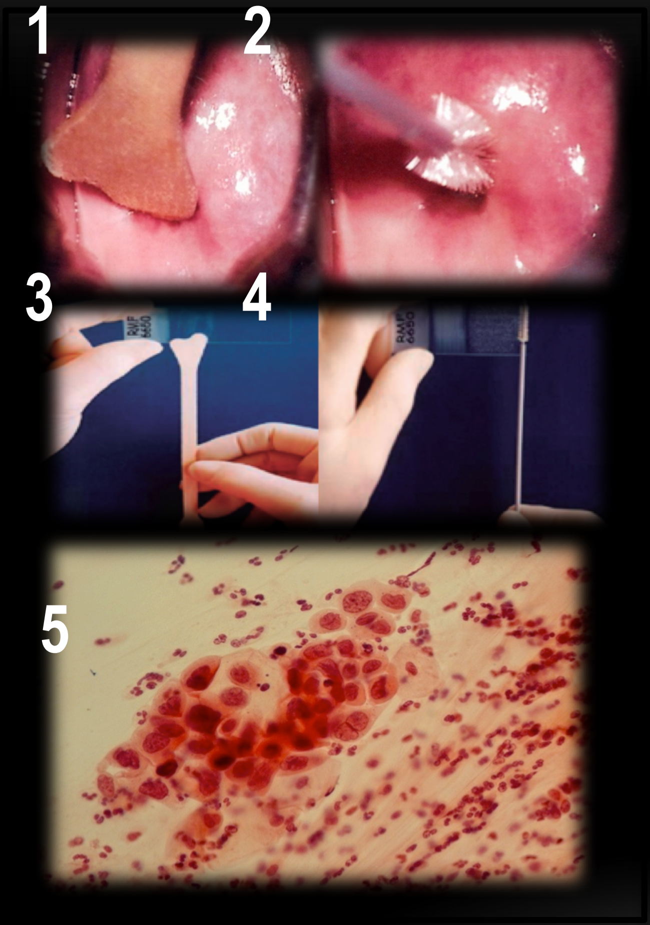
18,503 brasileiras/ano

8,414 mortes/ano*

*ICO Information Centre on HPV and Cancer - 2017

CERVICAL CANCER

COLLABORATION BETWEEN UCB, LBNL, UFOP, UFC, UFPI, SUS

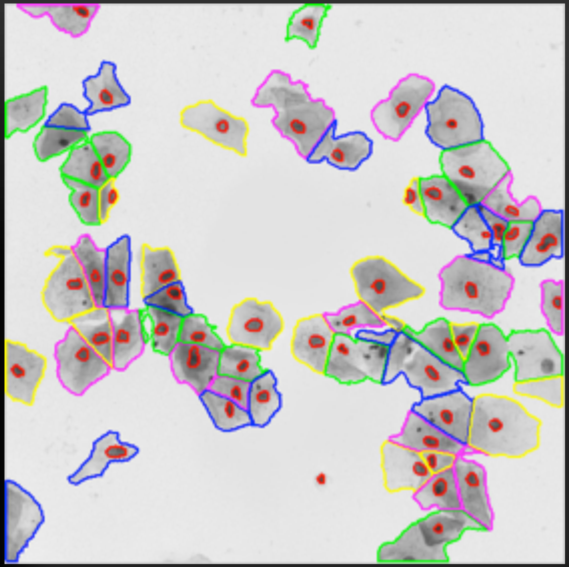


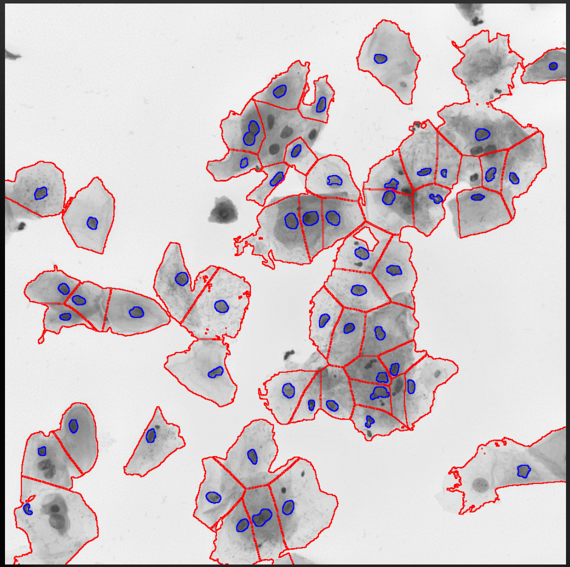
- Data collection with spatula and brush (1-2);
- Pap smear prep (3-4);
- Microscopy (5);
- Next:
  - Digital image
  - Cell identification
  - Cell classification
  - Pre-screening

BIDS BERKELEY INSTITUTE FOR DATA SCIENCE

Science Without Borders

Cell identification





- ***Problem:*** segment cervical cells from Pap smears images, and identify main subcellular components;
- ***Approach:*** invented method for cell identification based on super-pixel definition, combined to Voronoi diagrams;
- ***Results and Impact:***
  - Algorithms that can detect subcellular structures in 12 seconds;
  - Highest precision in comparison with other available algorithms;
  - **1st place in code competition: IEEE Int. Symp. on Biomedical Imaging 2014.**

Cell recognition for Inspection of Cervix – CRIC

SUMÁRIO

1. ~~Lawrence Berkeley National Lab~~
2. ~~Segredos e super poderes~~
3. ~~Projetos em visão computacional~~
4. Força tarefa “Made in Piauí”
5. SUS, células e professores da UFPI-Picos
6. Missão: busca de imagem por conteúdo
7. Oportunidades de estágio no exterior

SBC'17 ERIPI - DANI USHIZIMA

SISTEMA DE RECOMENDAÇÃO PARA IMAGEM CIENTÍFICA

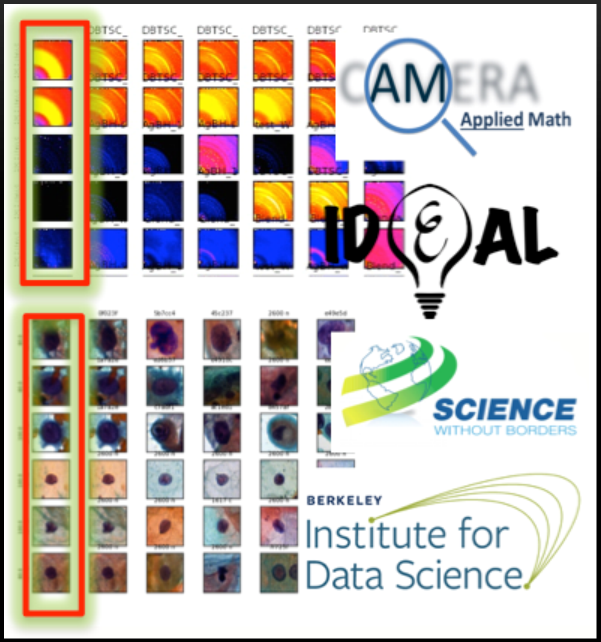
F. Araujo (UFC,UFPI), R. Silva (UFC,UFPI), D. Ushizima (LBNL, BIDS)

Problema

- Carência de ferramentas para recuperacao de imagens usando imagens, com código aberto;

Impacto

- **Recuperacao em tempo real usando representacao compacta (assinaturas)**;
- **Auxilio na** investigacao de padroes abstratos que herdem dados historicos, adquiridos com algo custo;



Research Details

- Deployed **CNN**-based tools for pattern recognition using optimized libraries, such as TensorFlow, cuDNN, cuFFT;
- **New visual search engine**: pyCBIR prototype for scientific image retrieval based on pictorial similarity;
- **Upcoming developments:** promote guided data explorations and new discoveries among scientists through enhanced online collaborative environments.

[1] Araujo, Romuere and Ushizima, "Searchable datasets in Python: images across domains, experiments, algorithms and learning – pyCBIR", *PyData San Francisco* 2016.

[2] Ushizima, Bale, Bethel, Ercius, Helms, Krishnam, Grinberg, Haranczyk, Macdowell, Odziomek, Parkinson, Perciano, Ritchie, and Yang. IDEAL: Images across Domains, Experiments, Algorithms and Learning, *Journal of Minerals, Metals and Materials*, 2016.

SEARCHING VISUALLY: NOW AND THE FUTURE

- Searching visually will become virtually interface-free;
- Search within context: science domain.

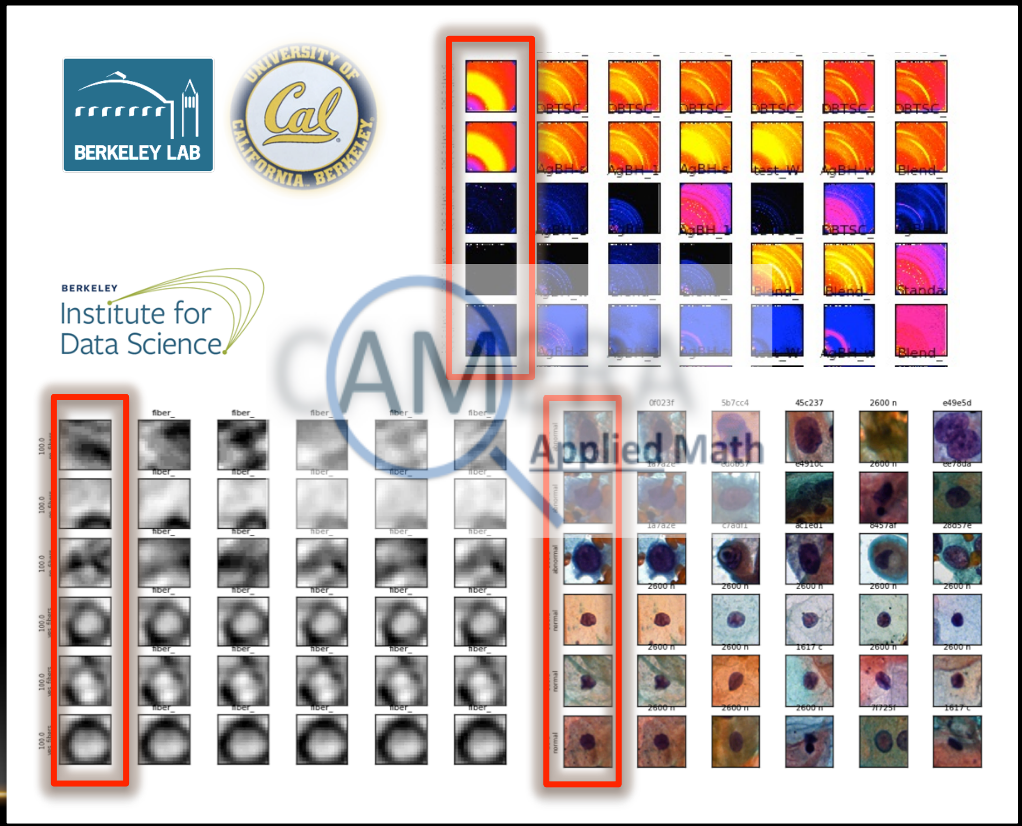
Made in Berkeley

pyCBIR

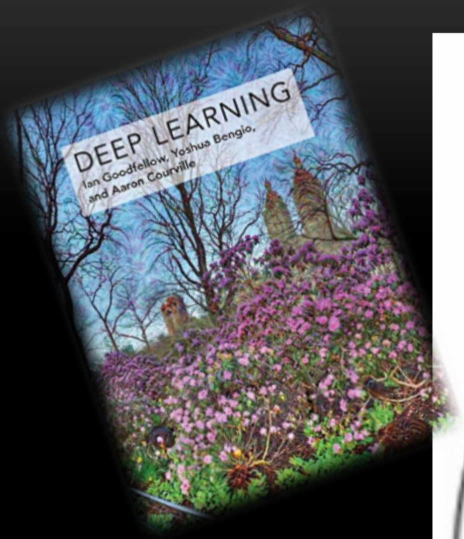
image retrieval library in python

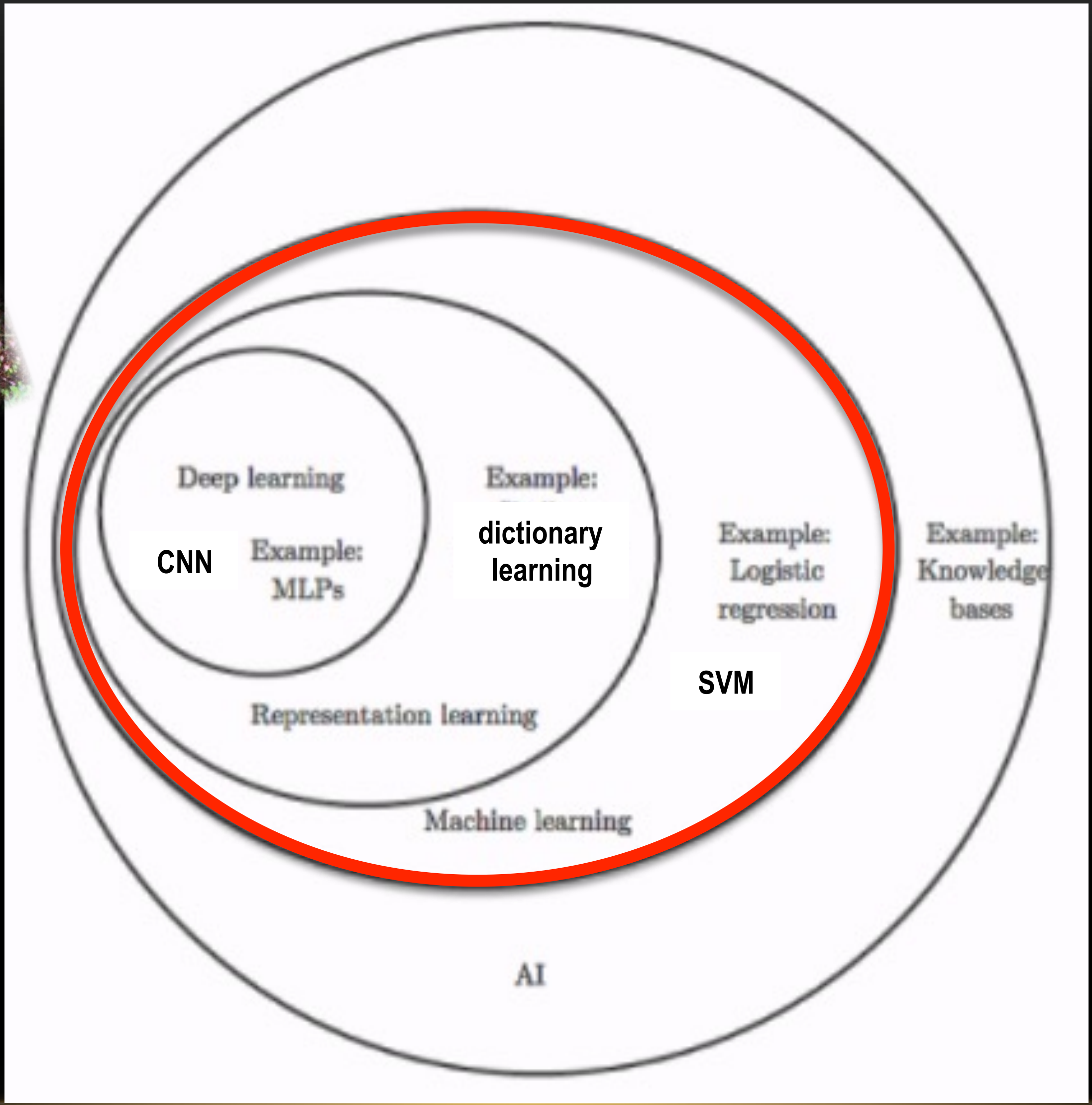
scikit learn

TensorFlow



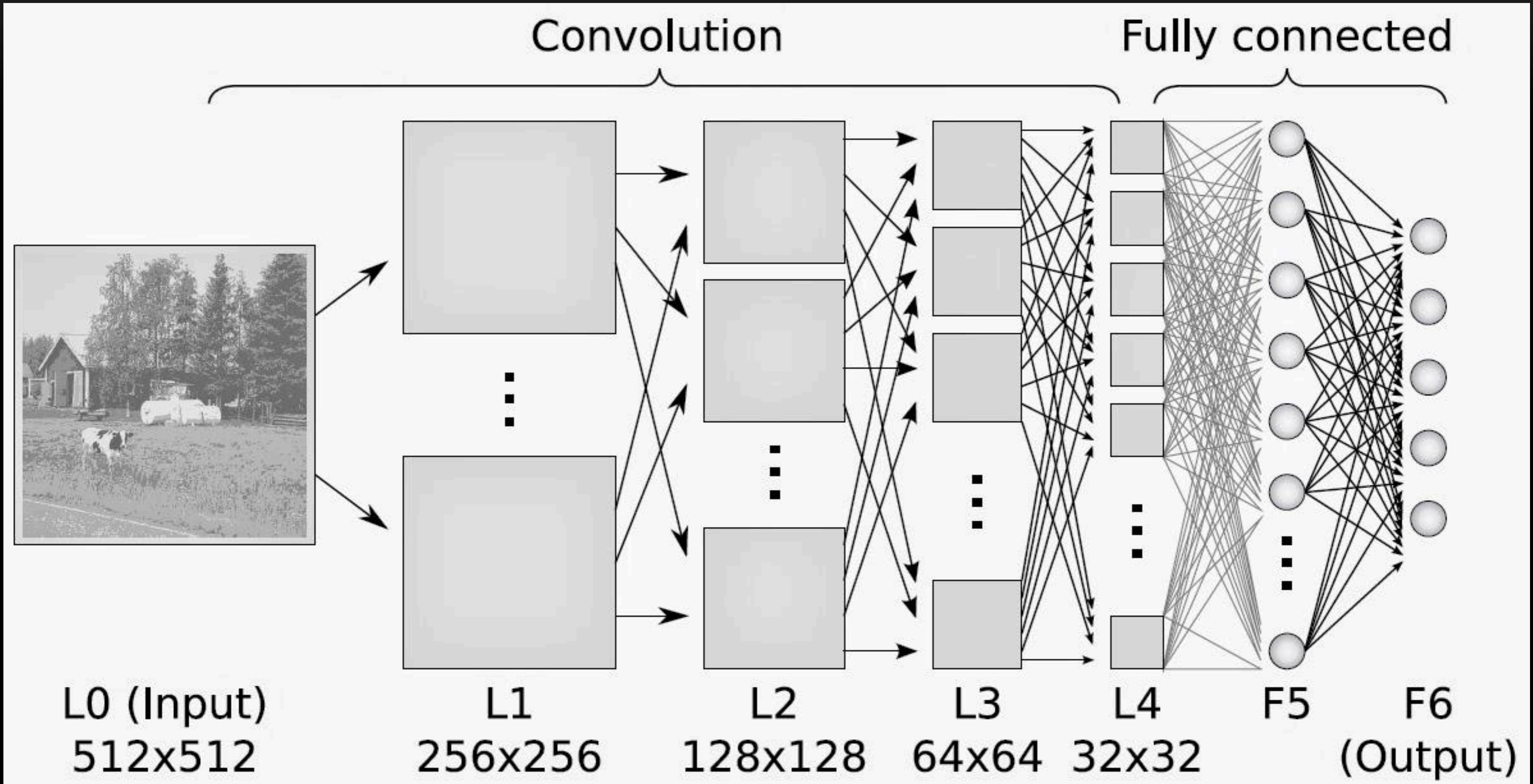
ML = DEEP LEARNING?





Source: Goodfellow, Bengio and Courville

CONVOLUTIONAL NEURAL NETWORK (CONVNET)



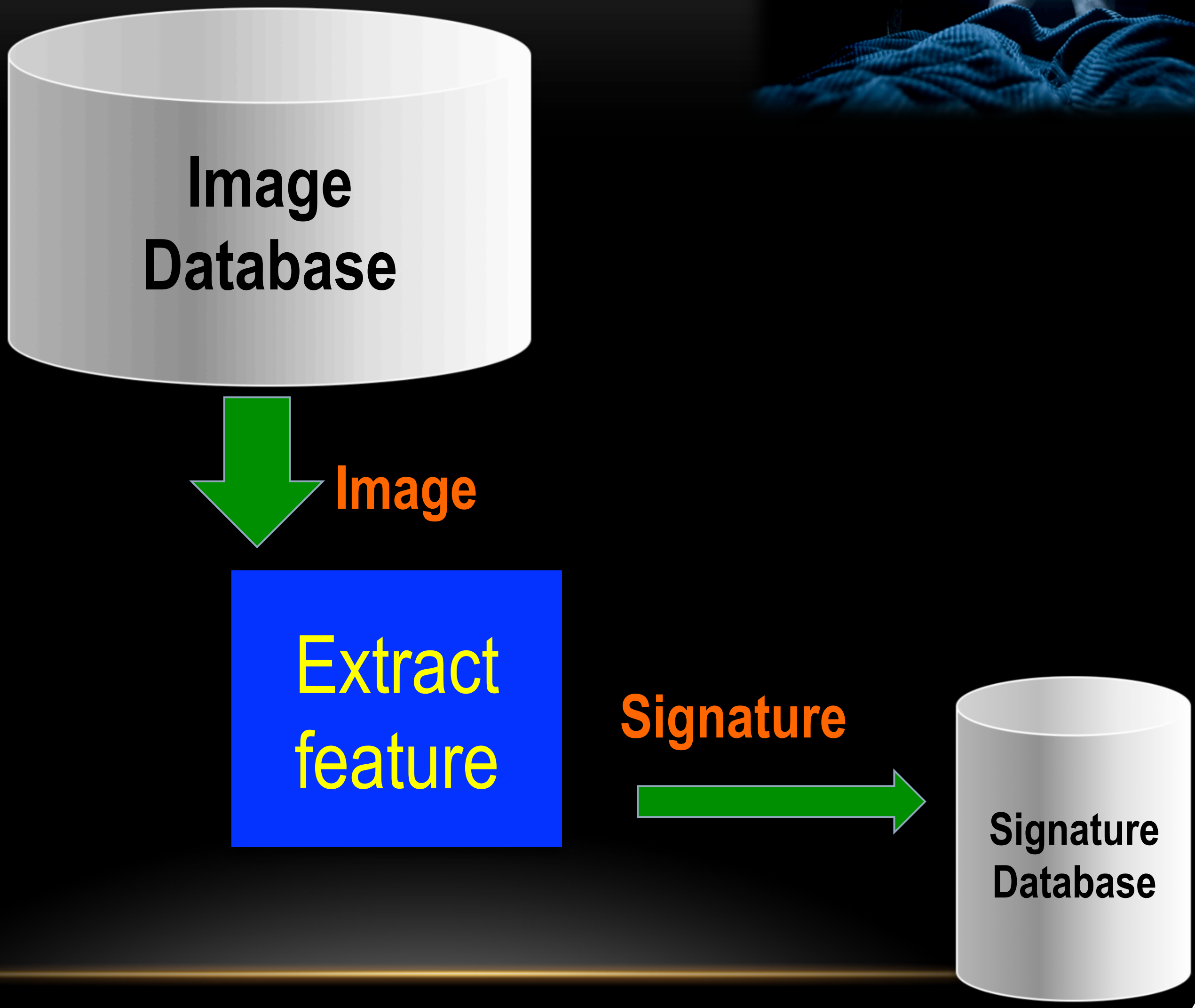
Source: http://www.ais.uni-bonn.de/deep_learning/

WHY PYCBIR?

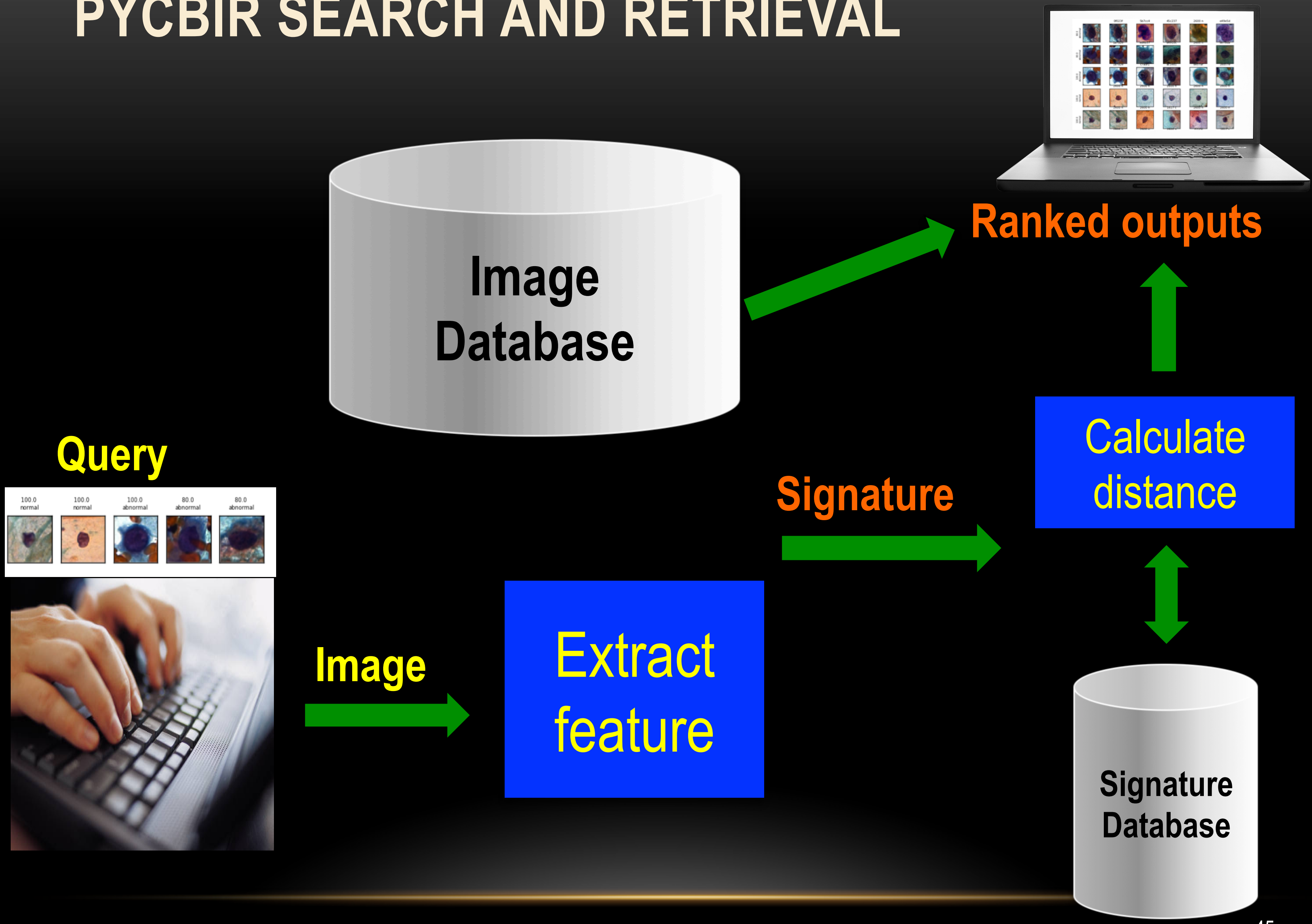
- New python tool for content-based image retrieval (CBIR);
- Query by example: capable of searching relevant items in large databases, given image samples;
- pyCBIR allows general purpose investigation across image domains;
- Our experiments: can we recover high-level abstraction from data using:

a. Color, texture, shape?
b. Learn signatures using CNN?
c. Similarity = distance?

PYCBIR TRAINING



PYCBIR SEARCH AND RETRIEVAL



FEATURE EXTRACTION METHODS

- Signature = index = feature vector = descriptors;

1. Gray Level Co-Occurrence Matrix;
2. Histogram of Oriented Gradient;
3. First Order Texture Features;
4. Local Binary Pattern;
5. Convolutional Neural Network.

FIRST ORDER TEXTURE FEATURES

- Mean

$$\mu = \sum_{i=0}^{G-1} i p(i)$$

- Variance

$$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 p(i)$$

- Skewness

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 p(i)$$

- Kurtosis

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 p(i) - 3$$

- Energy

$$H = -\sum_{i=0}^{G-1} p(i) \log_2 [p(i)]$$

- Entropy

$$E = \sum_{i=0}^{G-1} [p(i)]^2$$

GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM)

I

1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

GLCM

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

Features:

1. Contrast
2. Dissimilarity
3. Homogeneity
4. Energy
5. Correlation
6. ASM

CONVOLUTIONAL NEURAL NETWORK (CNN)

- We used the CNN in two different ways:
  1. Trained with the **same** database of the image retrieve;
     - 2 convolutional layers;
  2. Trained with the **imageNet** Database: CNN Inception*
     - 6 convolutional layers;

TensorFlow

theano

*Available in: http://arxiv.org/pdf/1602.07261v1.pdf

SIMILARITY

- IF Image = multidimensional vector,
  THEN similarity = distance!

1. Euclidean
2. Infinity
3. Cosine
4. Pearson
5. Chi-Square
6. Kullback-Liebler Divergence
7. Jeffrey Divergence
8. Kolmogorov-Smirnov Divergence
9. Cramer-von Mises Divergence
10. Cityblock Distance

DISTANCE METRICS

- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Infinity Distance

$$d(\mathbf{x}, \mathbf{y}) = max_{i=1}^{n} |x_i - y_i|$$

- Cosine Similarity

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||}$$

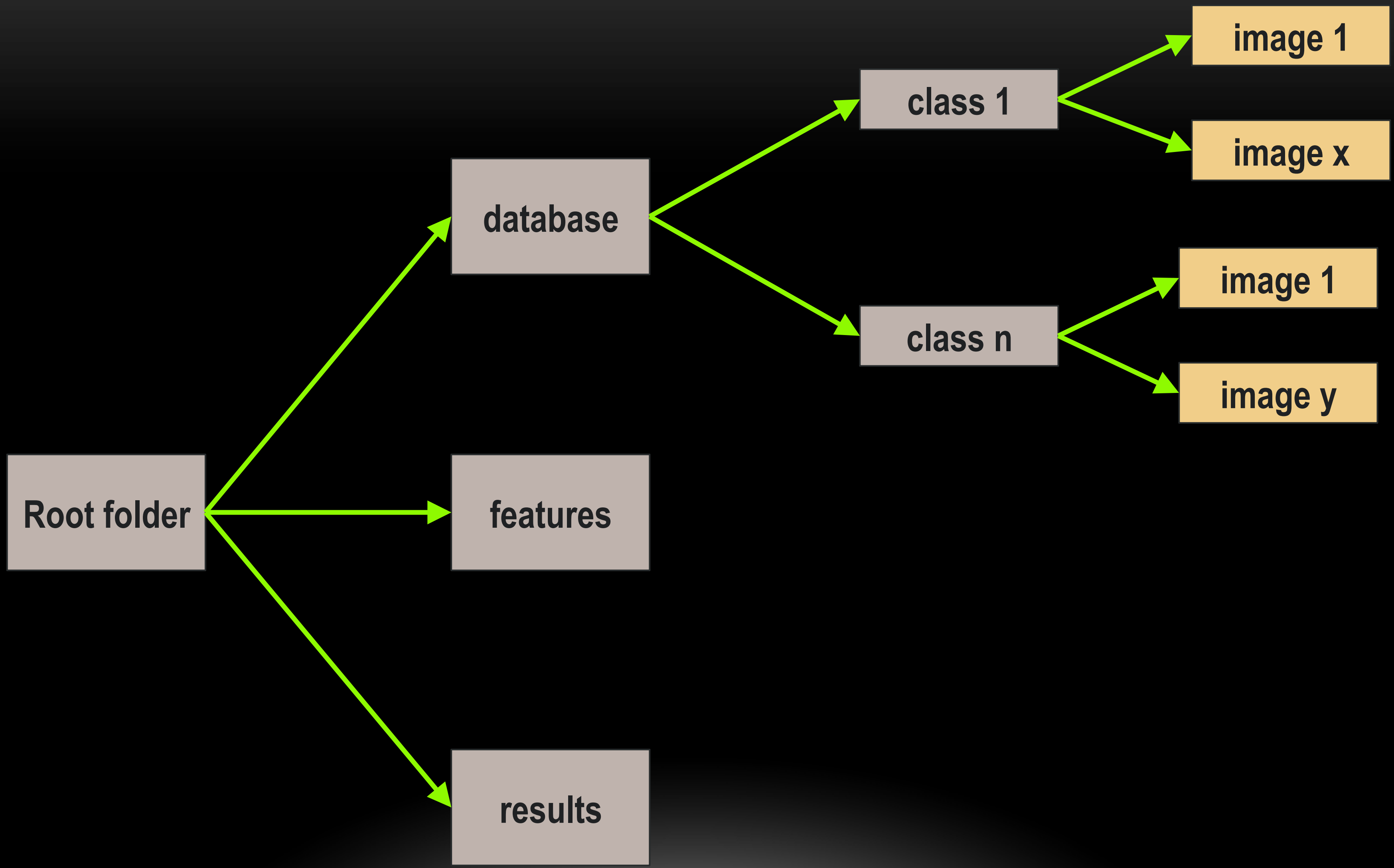
- Pearson Correlation Coefficient
- Chi-Square Dissimilarity

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

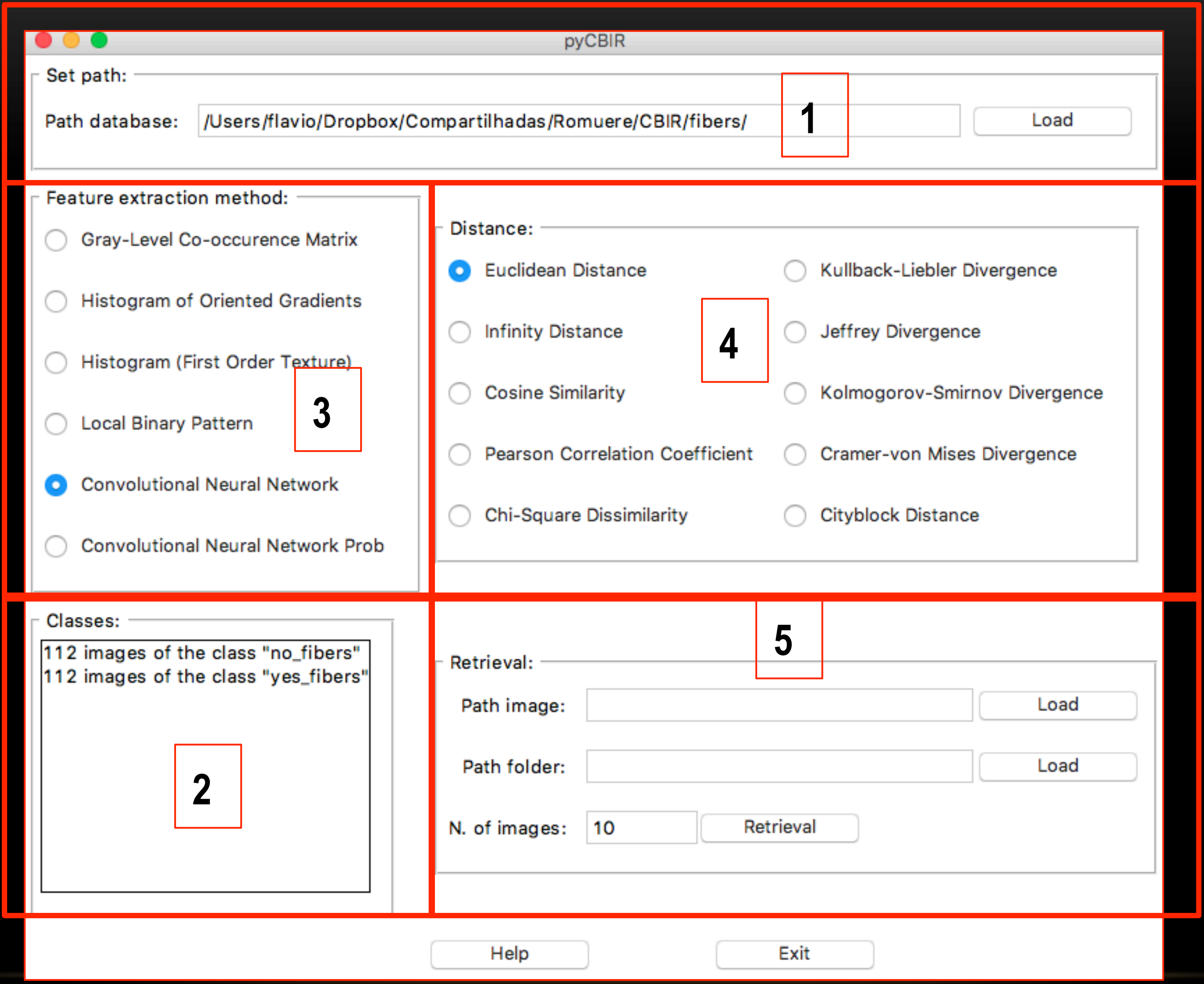
DISTANCE METRICS

- Kullback-Liebler Divergence $d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$
- Jeffrey Divergence $d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{\mu_i} + y_i \log \frac{y_i}{\mu_i}$
- Kolmogorov-Smirnov Divergence $d(\mathbf{x},\mathbf{y}) = max_{i=1}^{n} |X_i - Y_i|$
- Cramer-von Mises Divergence $d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} (X_i - Y_i)^2$
- Cityblock Distance $(L_1)\ d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$

HOW TO ORGANIZE THE DATABASE?



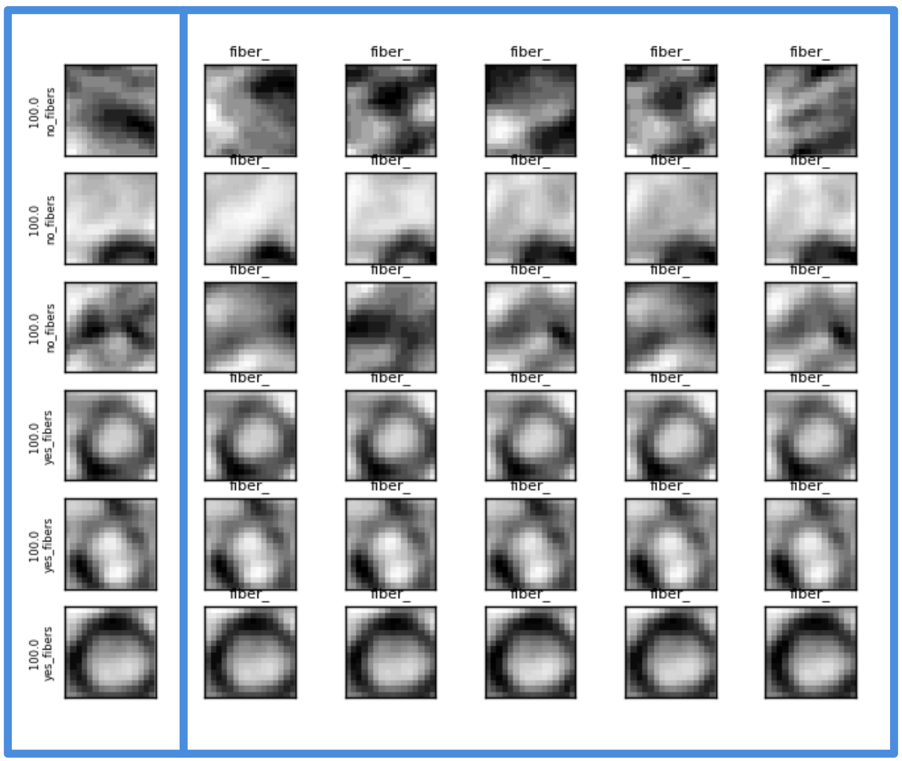
GRAPHICAL USER INTERFACE



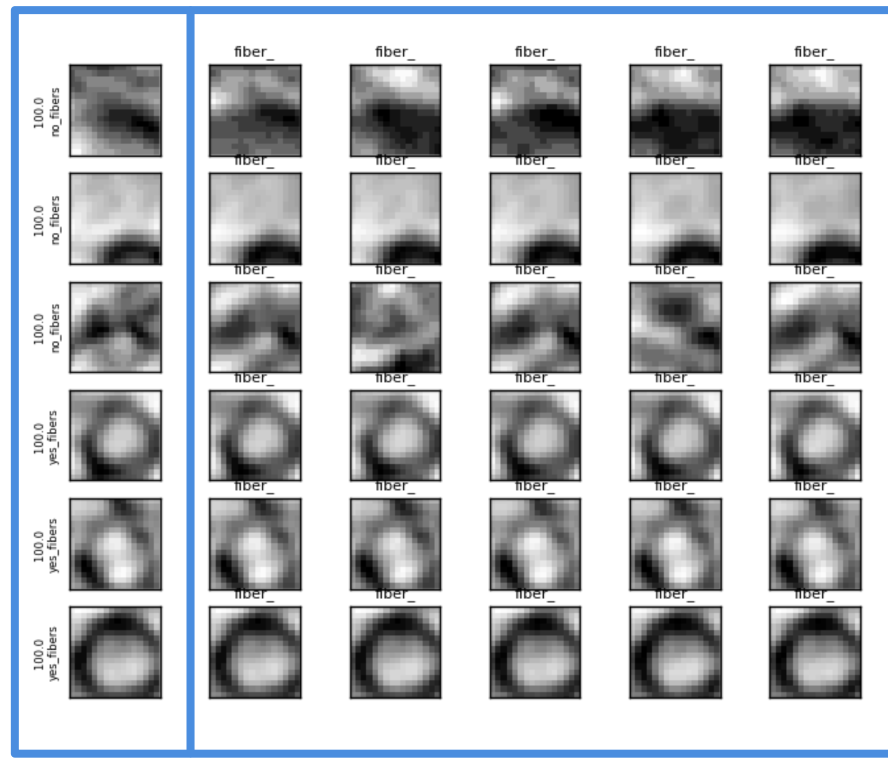
EXPERIMENTS - FIBERS DATASET

4,000 images of 16 X 16 for two balanced classes.

Query	Top 5 retrieved	Query	Top 5 retrieved



Result obtained using the CNN trained with the same dataset.



Result obtained using the inception network.

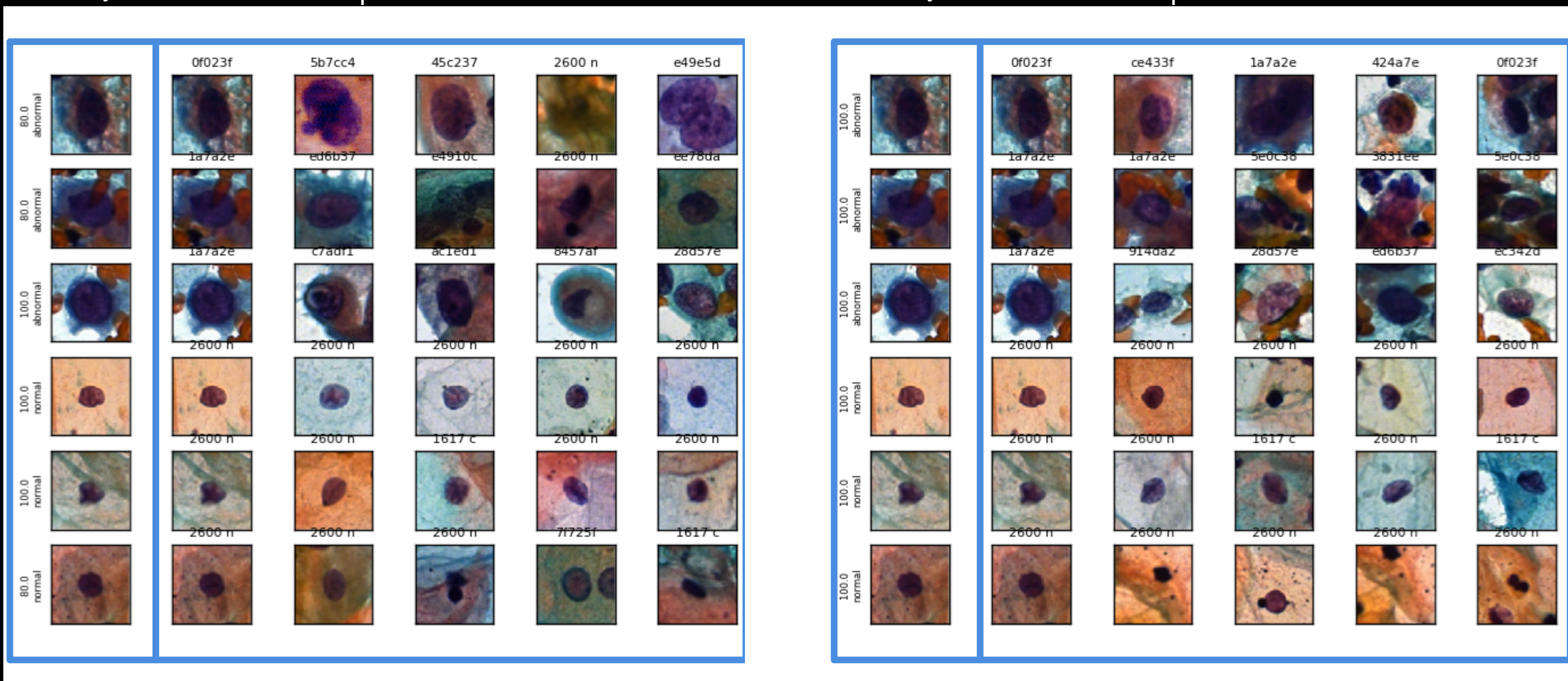
EXPERIMENTS - TIME

	Training	Extraction of features for the whole database	Top 10 retrieved for a query image
Approach 1 (same DB)	3.4 minutes	9 seconds	4 seconds
Approach 2 (inception)	-	29 minutes	15 seconds

EXPERIMENTS - CELLS DATASET

1,886 images of normal cells and 1,509 of abnormal - 100 X 100 pel;

Query | Top 5 retrieved | Query | Top 5 retrieved



Result obtained using the CNN trained with the same datase.

Result obtained using the inception network.

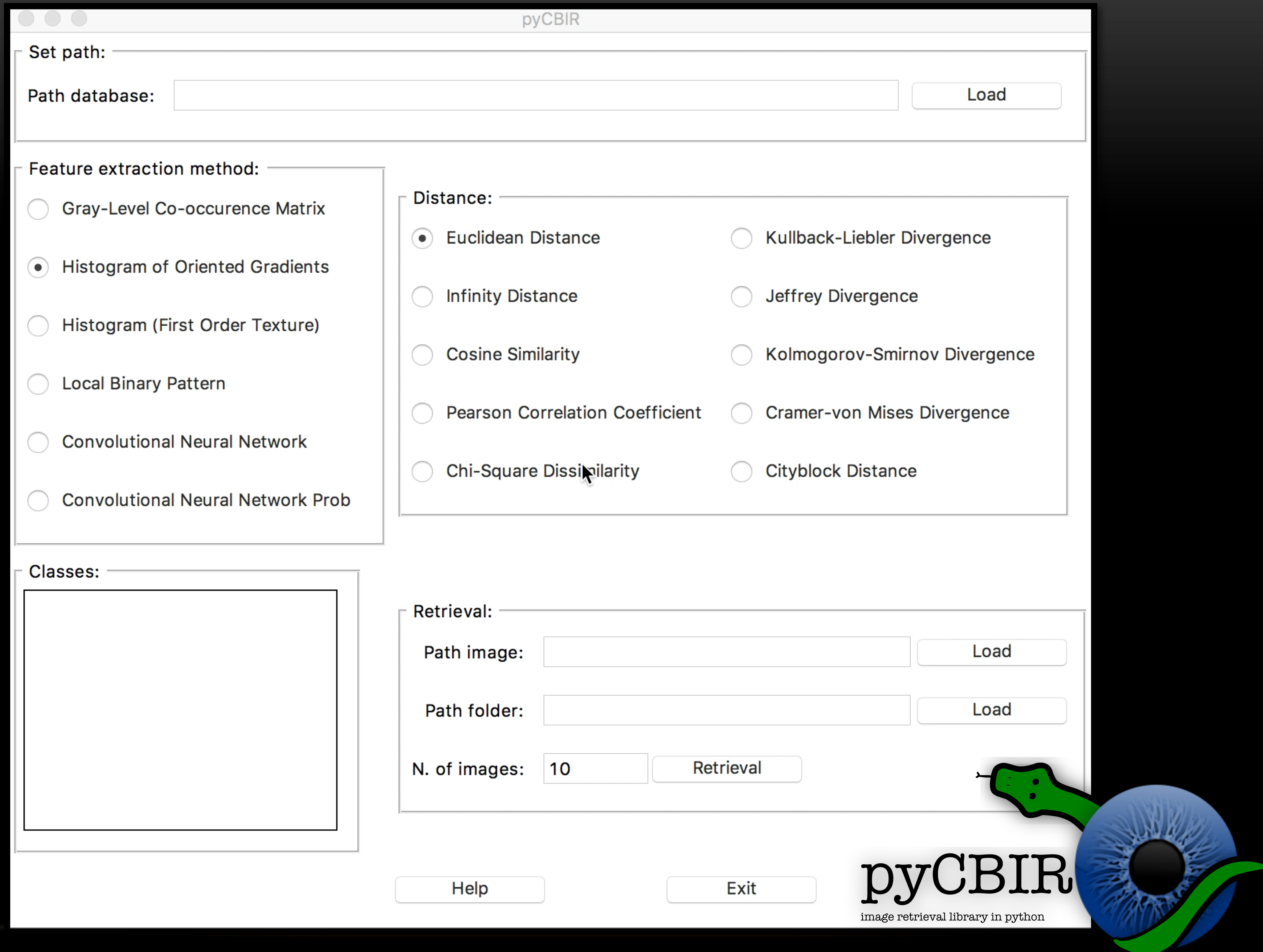
EXPERIMENTS - TIME

	Training	Extraction of features for the whole database	Top 10 retrieved for a query image
Approach 1	94 minutes	48 seconds	5 seconds
Approach 2	-	23 minutes	12 seconds

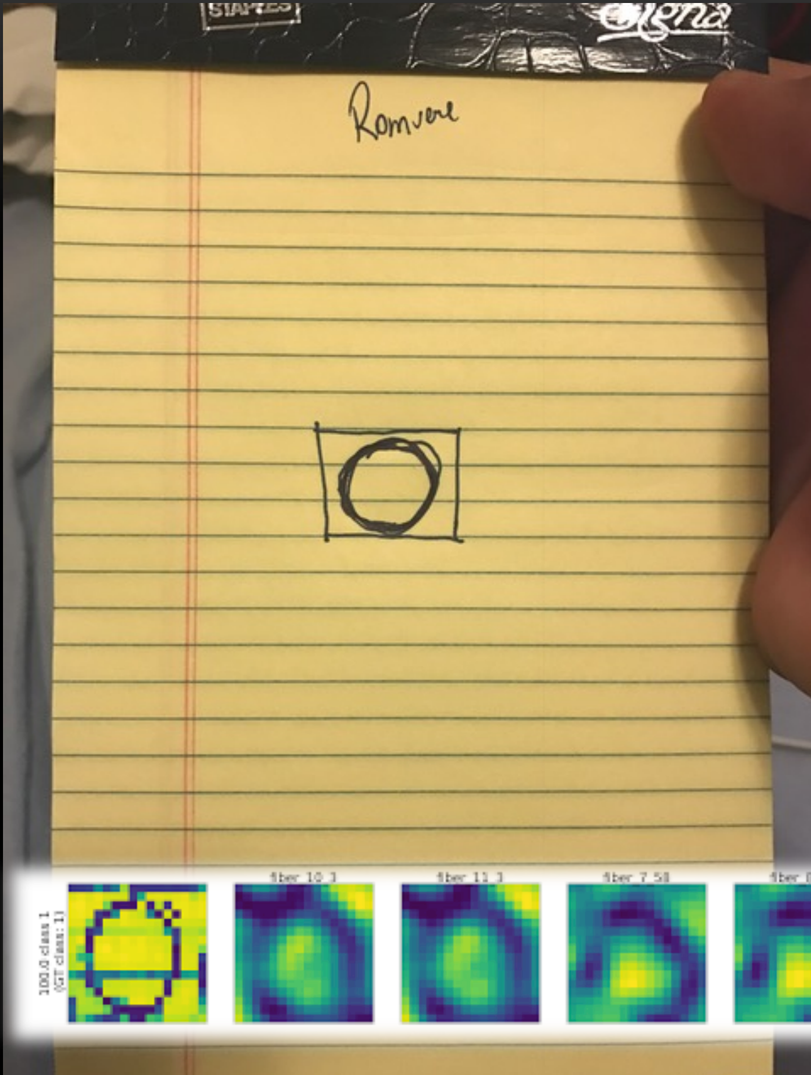
CONCLUSIONS

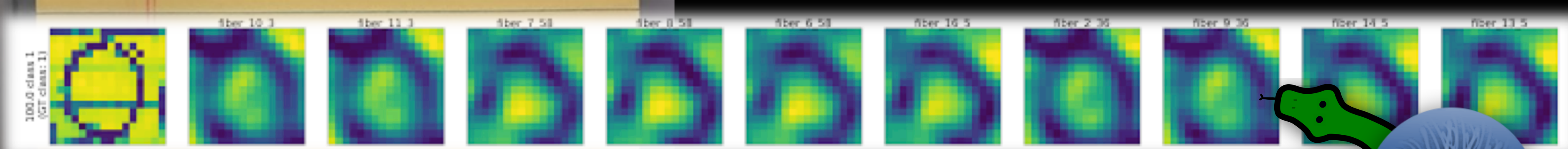
Approach 1 (same DB)	Approach 2 (inception)
Advantages	
Feature extraction faster after training	Doesn't need training
Training done only once	
Each image = 256 features	
Disadvantages	
For datasets with big images and a lot of classes the training is slow	Feature extraction is slow
	Each image = 2,048 features

SEARCHING VISUALLY NOW



ANY MORE TRICKS WITH PYCBIR VISUAL SEARCH?





OPORTUNIDADES

Converse com quem já esteve fora do país:

- Ex: Prof. Flavio e Prof. Romuere

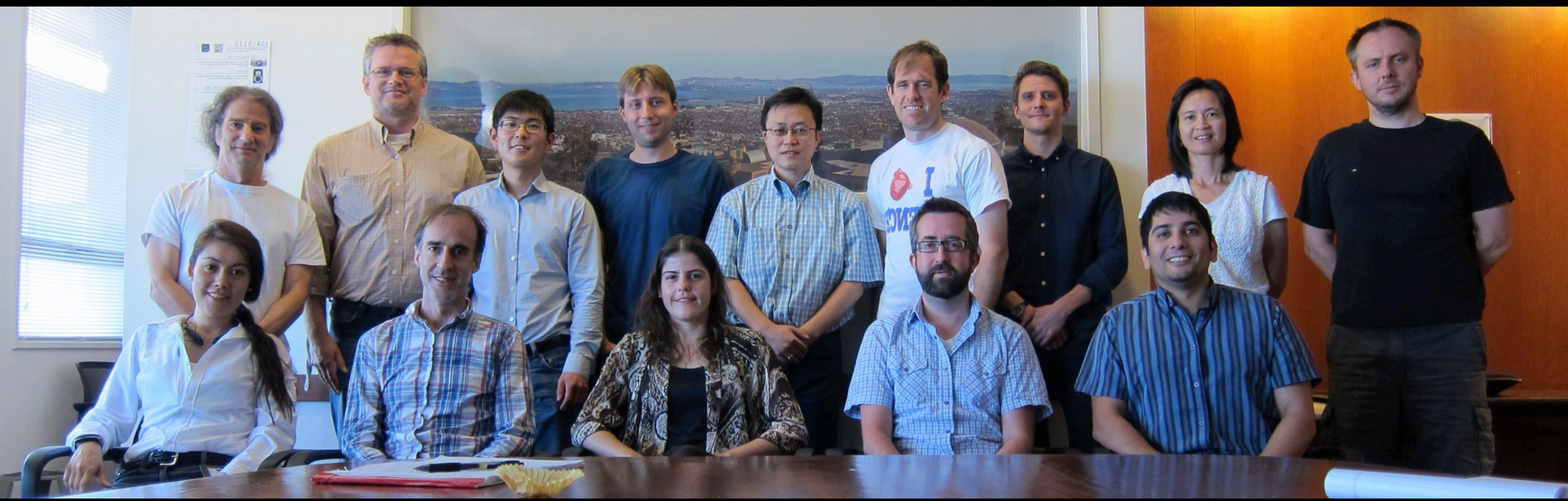
Governo brasileiro, estrangeiro e iniciativa privada:

- Fullbright
- Fundação Leeman
- Laboratórios nacionais

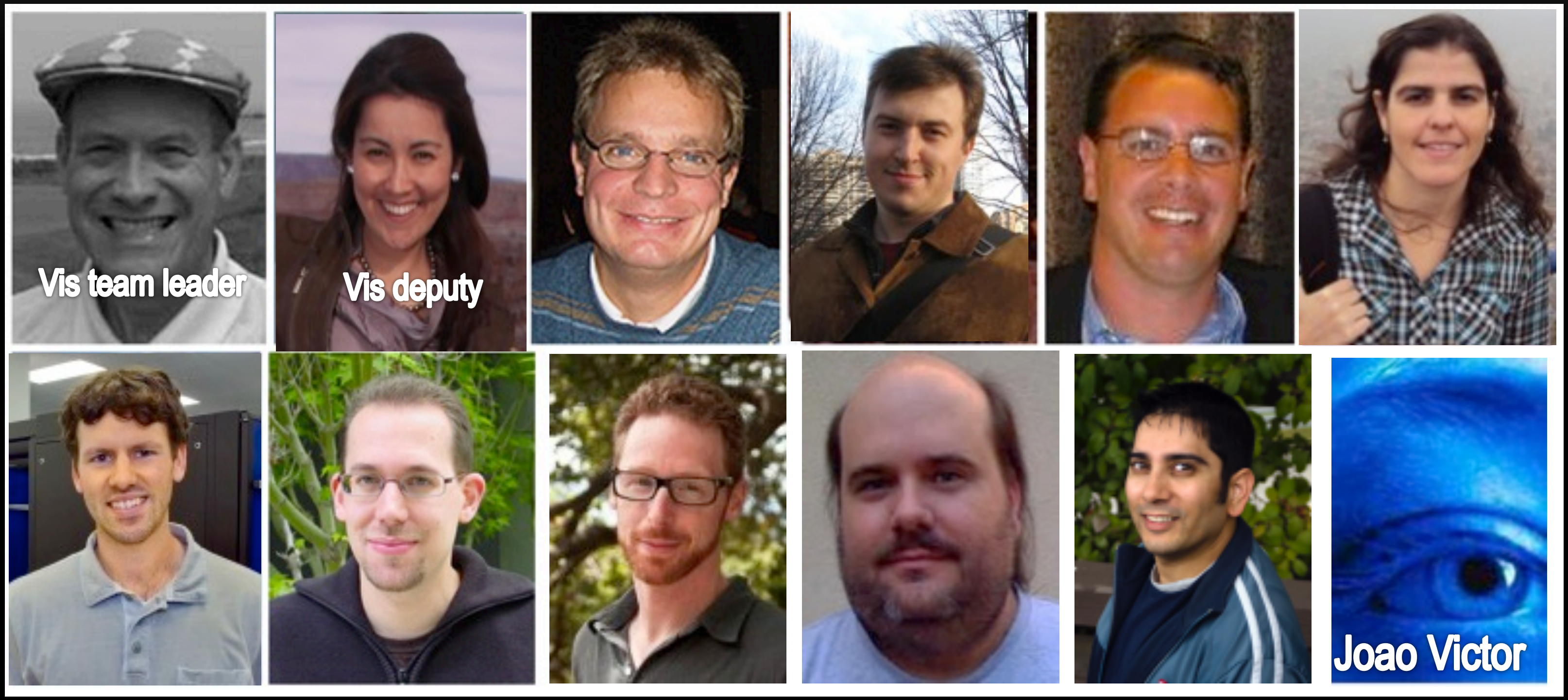
CRIC

CELL RECOGNITION FOR INSPECTION OF CERVIX

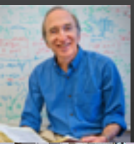
CAMERA Applied Math



DATA ANALYTICS & VISUALIZATION
DAV GROUP



Wes Bethel, Daniela Ushizima, Gunther Weber, Dmitriy Morozov, Hank Childs, Talita Perciano, Mark Howison, Oliver Ruebel, Burlen Loring, David Camp, Hari Krishnam



BERKELEY INSTITUTE FOR DATA SCIENCE

THANKS

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

Berkeley UNIVERSITY OF CALIFORNIA

SELECTED JOURNALS

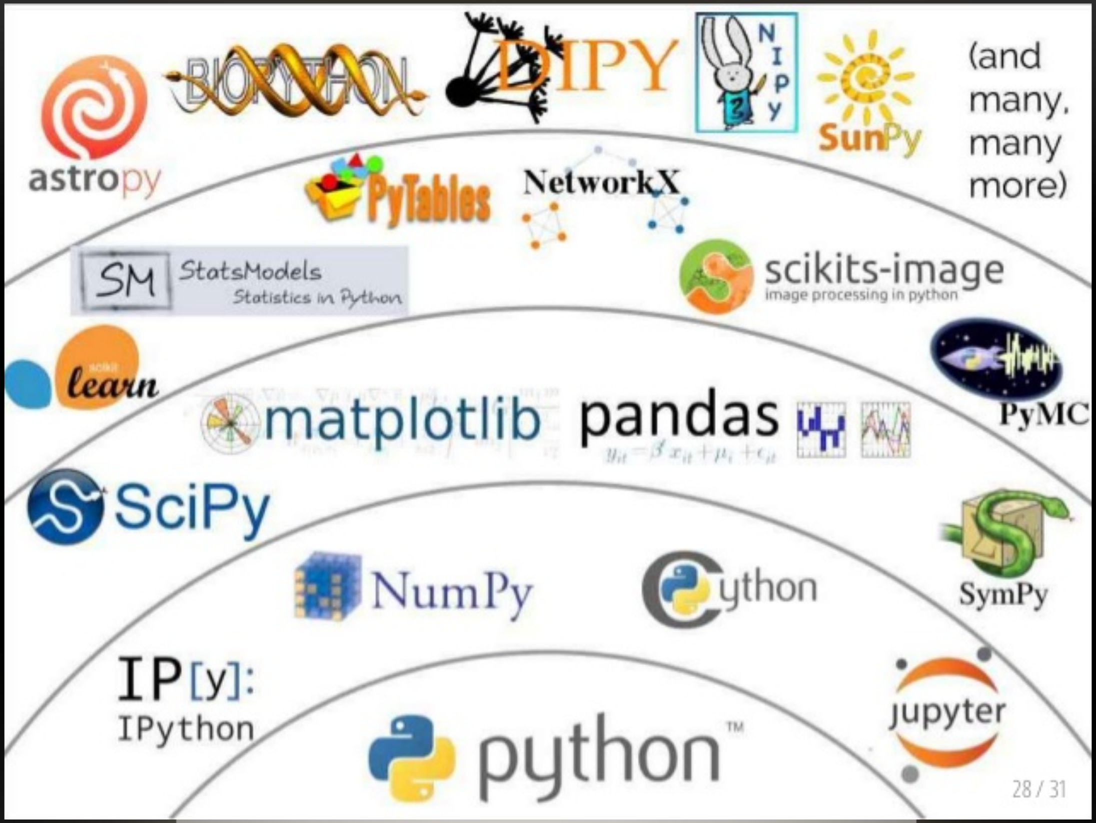
- Alegro, Theofilas, Nguy, Castruita, Seeley, Ushizima, Grinberg, *Automating Cell Detection and Classification in Human Brain Fluorescent Microscopy Images Using Dictionary Learning and Sparse Coding*, Journal of Neuroscience Methods, 2017.
- Ushizima, Bale, Bethel, Ercius, Helms, Krishnam, Grinberg, 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.

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.

Python

How to navigate and choose among the several scientific computing packages freely available for Python?



@jakevdp

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*