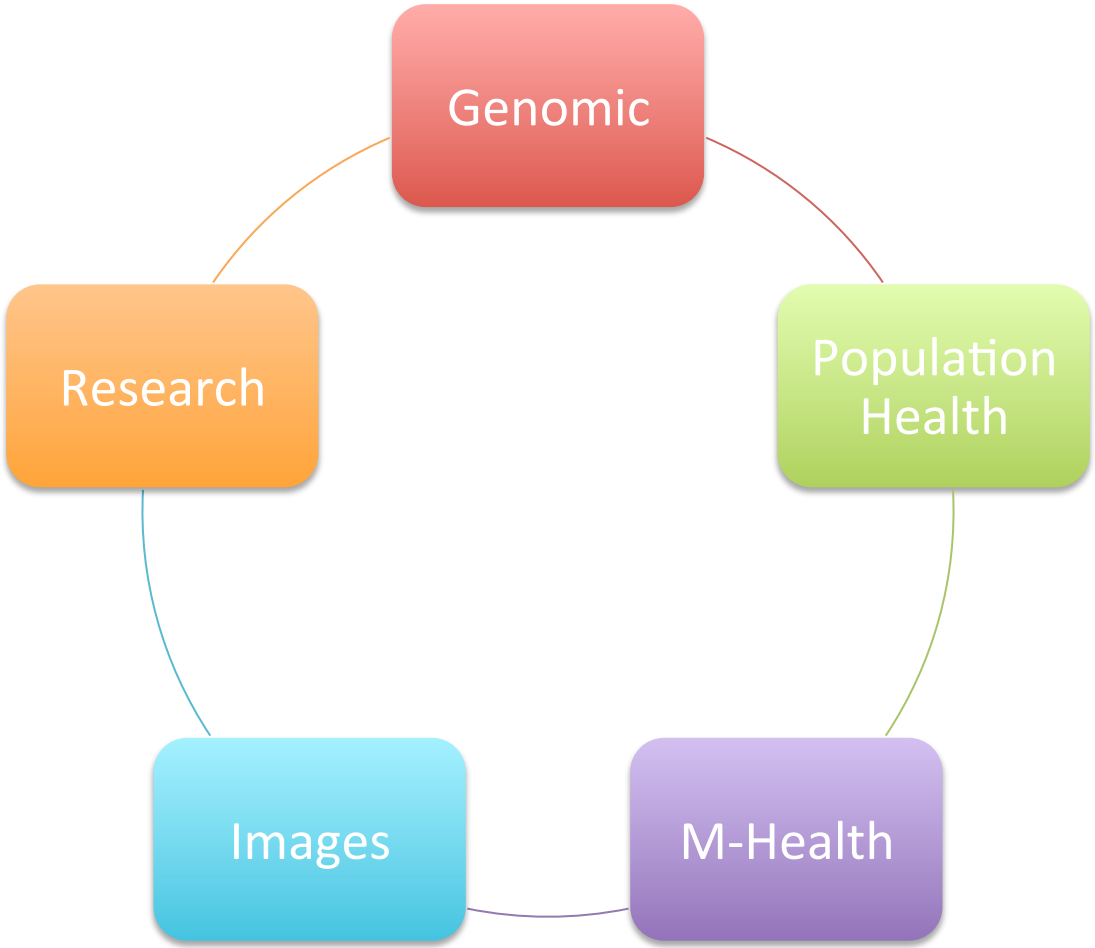
Big Data in medical image processing

Konstantin Bychenkov, CEO

Aligned Research Group LLC

bychenkov@alignedresearch.com

Big data in medicine



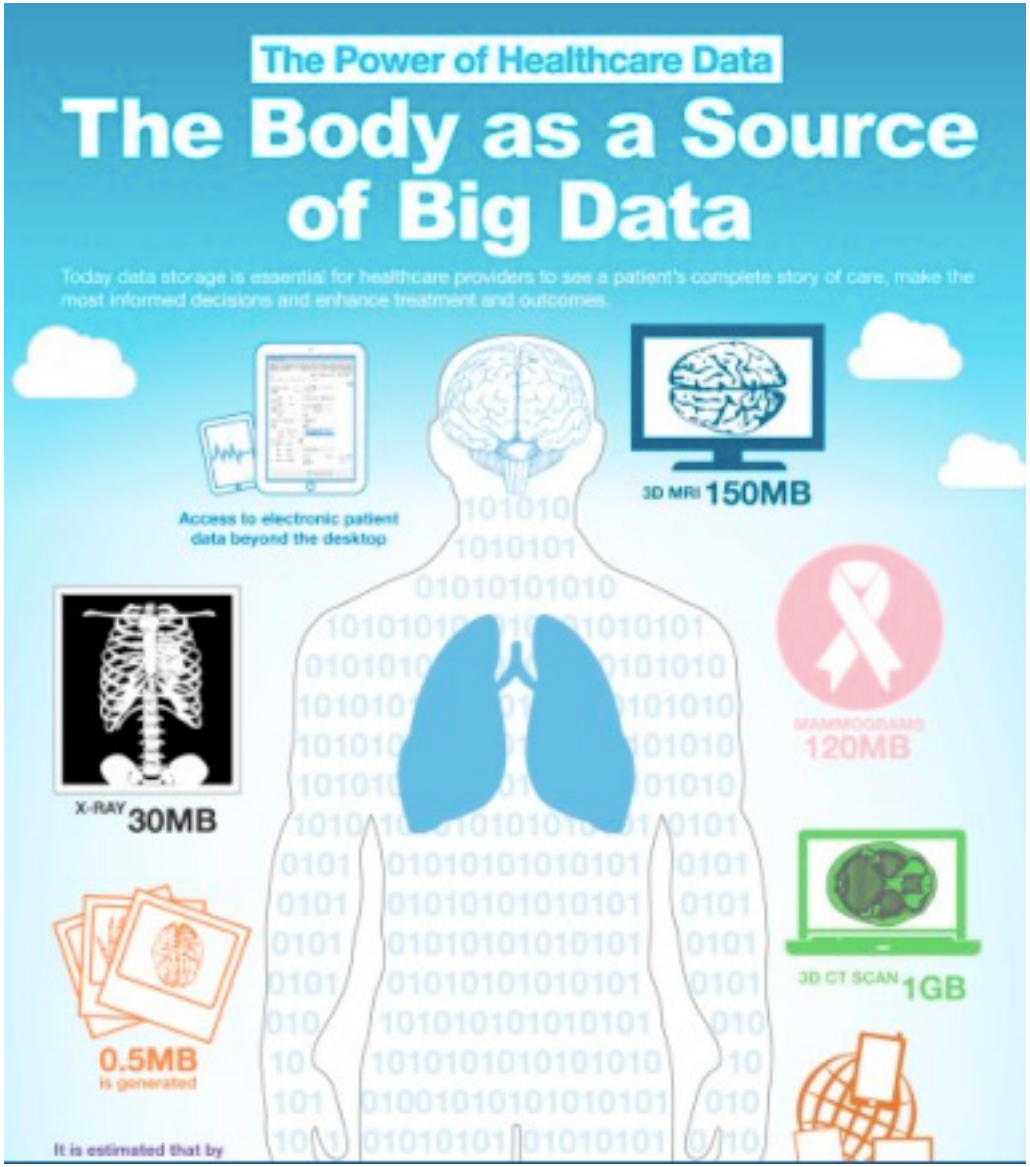
ISB AWARDED $6.5 MILLION NIH CONTRACT TO DEVELOP 'CANCER GENOMICS CLOUD' WITH GOOGLE AND SRA INTERNATIONAL

PRESS RELEASE Posted on October 10, 2014

Institute for Systems Biology (ISB) has received a $6.5 million, up to two-year, federally funded contract from the National Cancer Institute (NCI), National Institutes of Health (NIH). ISB is one of three organizations awarded a contract by NCI to develop a cloud--based platform that will serve as a large-scale data repository and provide the computational infrastructure necessary to carry out cancer genomics research at unprecedented scales. ISB's Shmulevich group will serve as the lead on the project with two partners: Google, Inc., and SRA International, Inc. (a Fairfax, Va.-based provider of IT solutions and professional services to government organizations).

https://cloud.google.com/genomics/v1beta2/reference/

- By 2015, the average hospital will hold 665 TB of patient data
  - 80% of which will be unstructured image data like CT scans and X-rays
- Medical imaging archives increasing by 20-40%

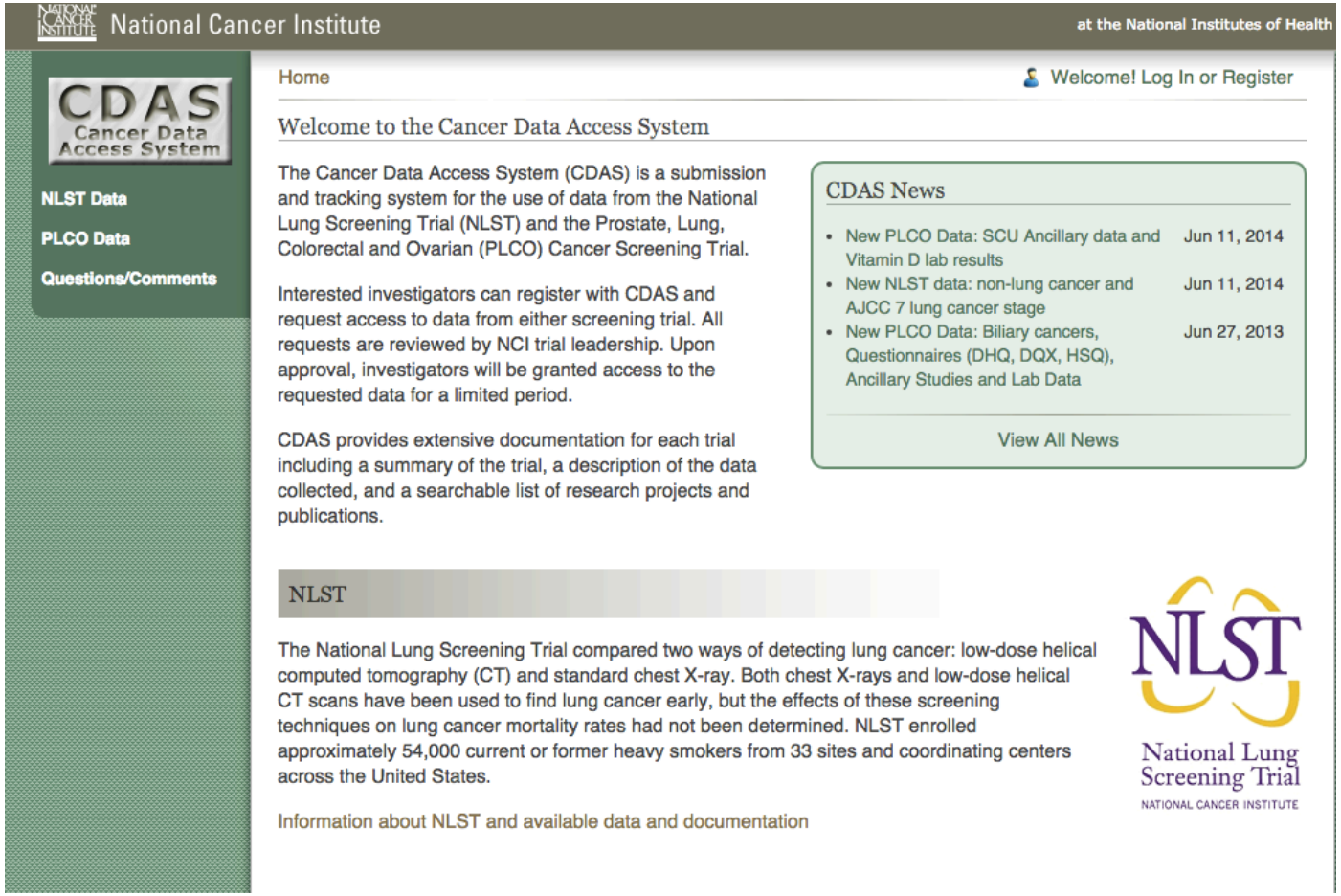


New possibilities

- Personalized medicine became a reality
- It is possible to set up a secure, anonymous and large-scale collection images
- High precision of the new age of diagnostic devices
- The lack of objectification is a considerable risk

What do we have

https://biometry.nci.nih.gov/cdas/



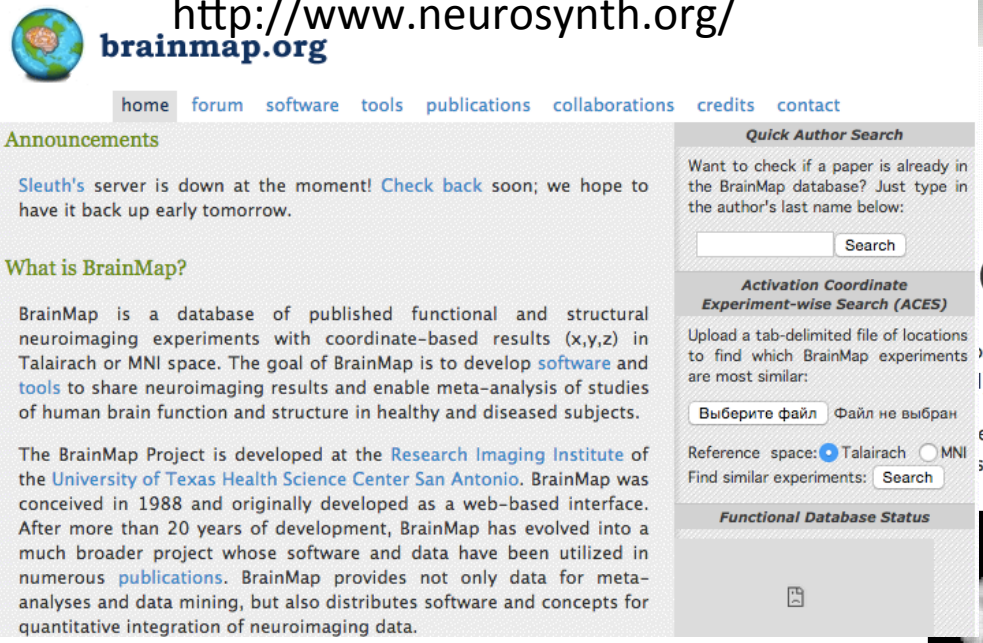
What do we have

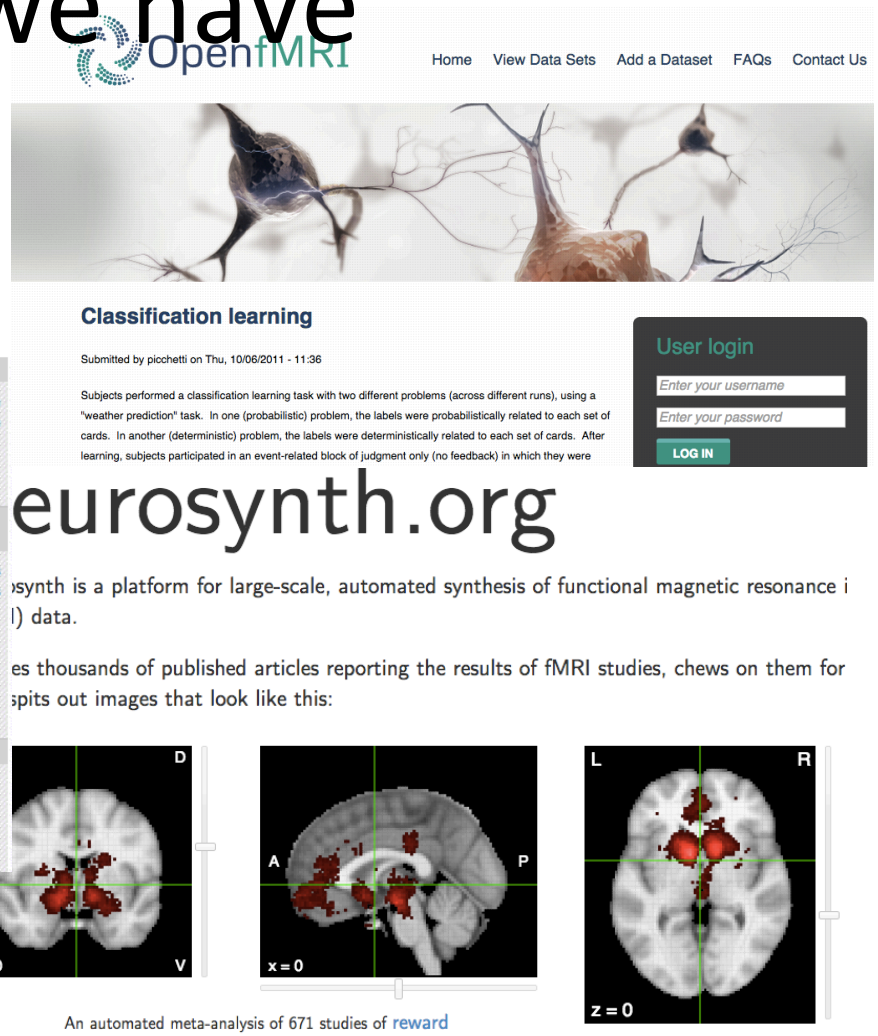
Section 2: Spiral CT Screening

Variable	Label	Description	Format Text
attempts	Number of screening attempts on last screening visit this year	Number of screening attempts on the last screening visit for this study year.	Numeric .M="Missing"
ct_recon_filter1-4	CT reconstruction algorithm / filter	What CT reconstruction algorithm / filter was used for the screen? These variables come from thedata collection forms. They may disagree with data extracted from the CT images' DICOM headers. Header data may be obtained from the NLST CT image collectionat TCIA or from ACRIN.	.M="Missing or less than 4 algorithms/filters" 1="GE Bone" 2="GE Standard" 3="GE, other" 4="Phillips D" 5="Phillips C" 6="Phillips, other" 7="Siemens B50F" 8="Siemens B30" 9="Siemens, other" 10="Toshiba FC10" 11="Toshiba FC51" 12="Toshiba, other"
ctdxqual	Overall diagnostic quality		.M="Missing"

What do we have

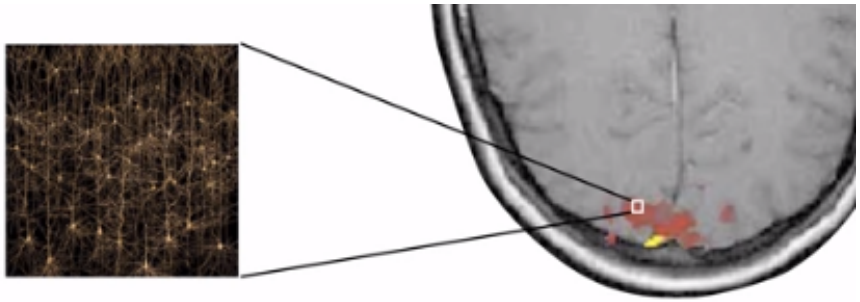
https://openfmri.org/dataset/ds000002
http://www.brainmap.org/
http://www.neurosynth.org/





Neurosynth:
413429 activations reported in 11406 studies
Interactive, downloadable meta-analyses of 3107 terms
Functional connectivity and coactivation maps for over 150,000 brain locations

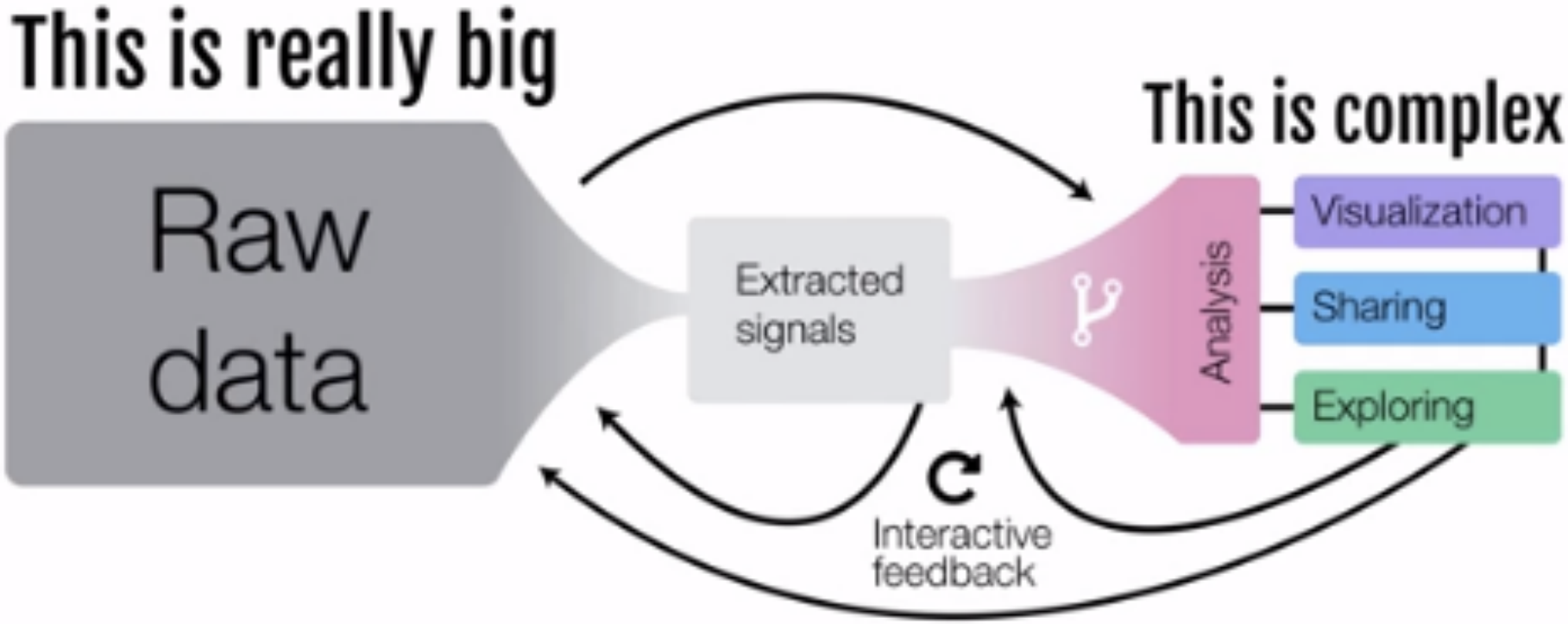
What can we do with fMRI



~50,000 neurons per cubic millimeter
-> need higher resolution!

Mouse, somatosensory cortex ~1,000 neurons	**0.1 TB / experiment**
Larval zebrafish, whole-brain ~100,000 neurons	**1 TB**
* Entire mouse brain ~80,000,000 neurons	**>100 TB**

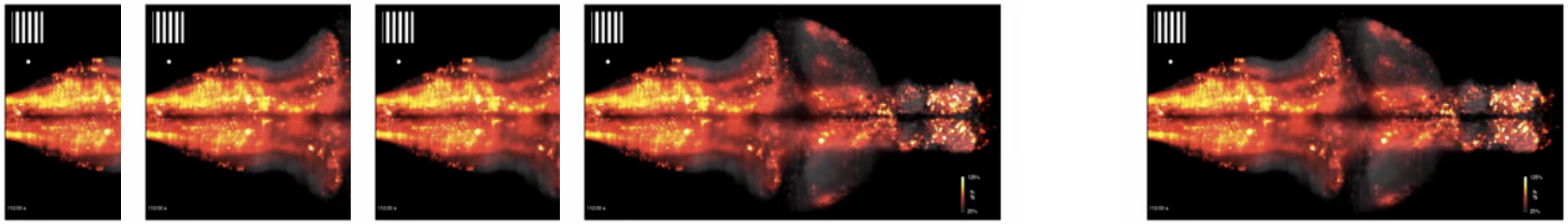
* hypothetical


This is really big
Raw data
Extracted signals
This is complex
Analysis
Visualization
Sharing
Exploring
Interactive feedback

Patterns in neurons activity

which brain areas are active at which times?
which brain areas are activated by different directions of the visual stimulus?

100 000 neurons of zebra-fish in 200 sec video
Map images into lines of pixels intensity
Reduce dimension and find out patterns in time



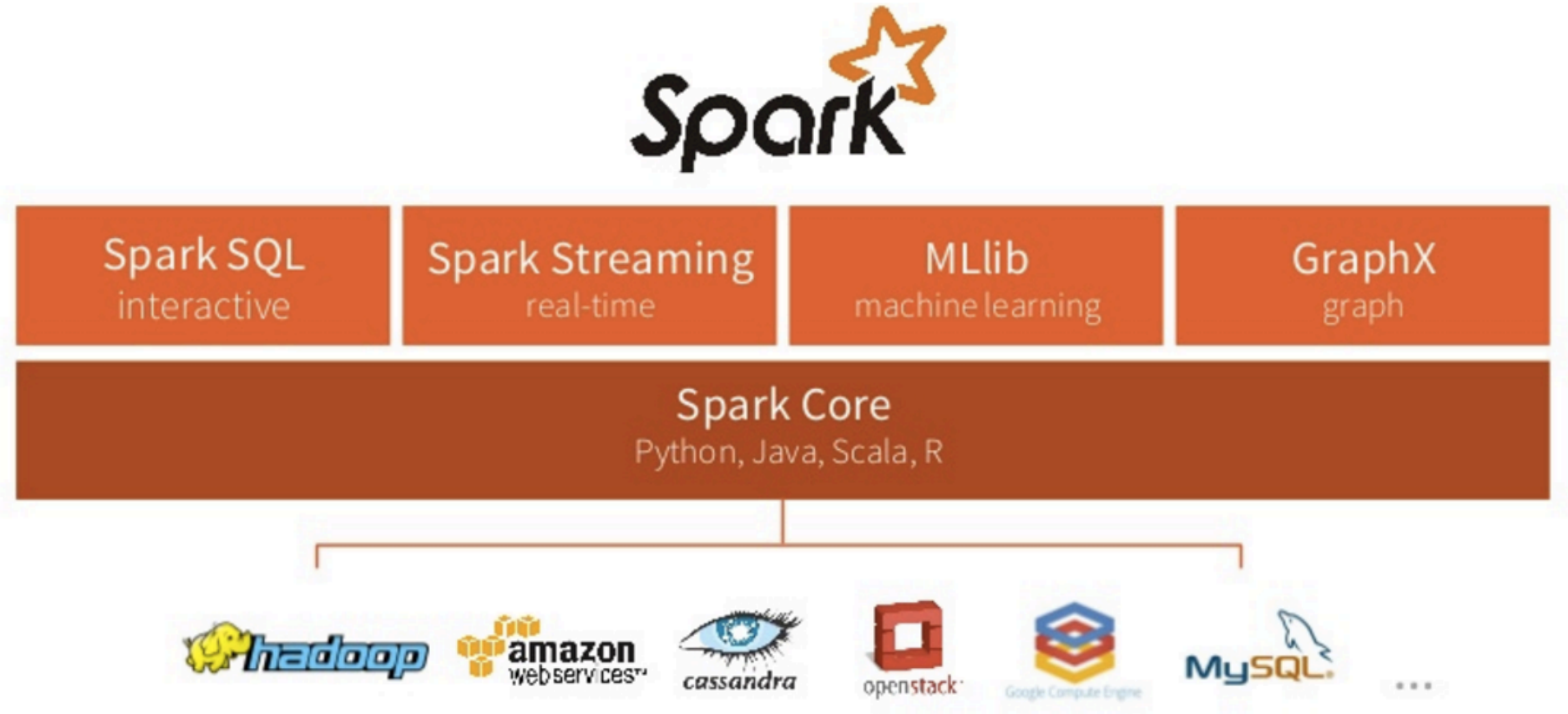
time

Method

PCA method to reduce dimension and eliminate patterns in data

Spark to process data in reasonable time and in different configuration

Simplifies Big Data Processing

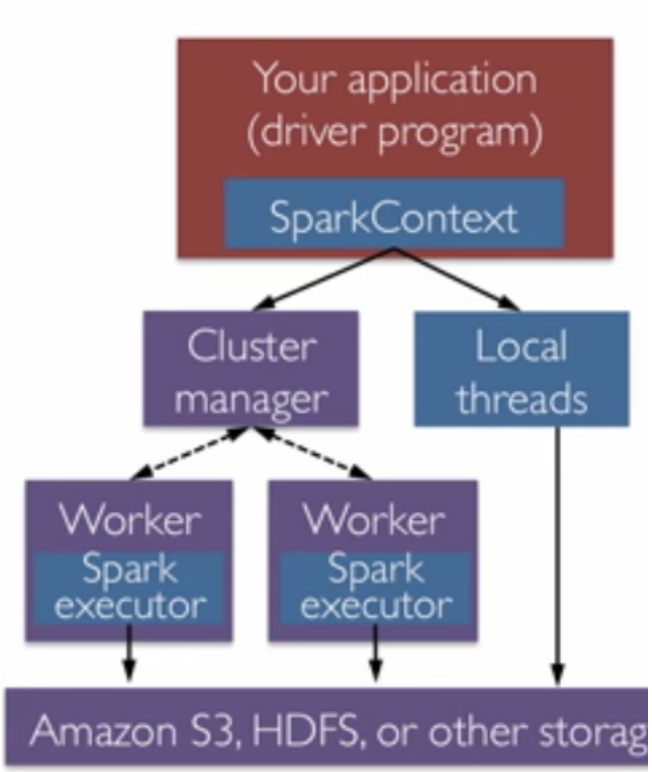


Fast Scalable General

Spark core

- Task Management
- Memory Management
- Sustainability
- Data Source Management
- API for data collections (RDD)

Cluster mode overview

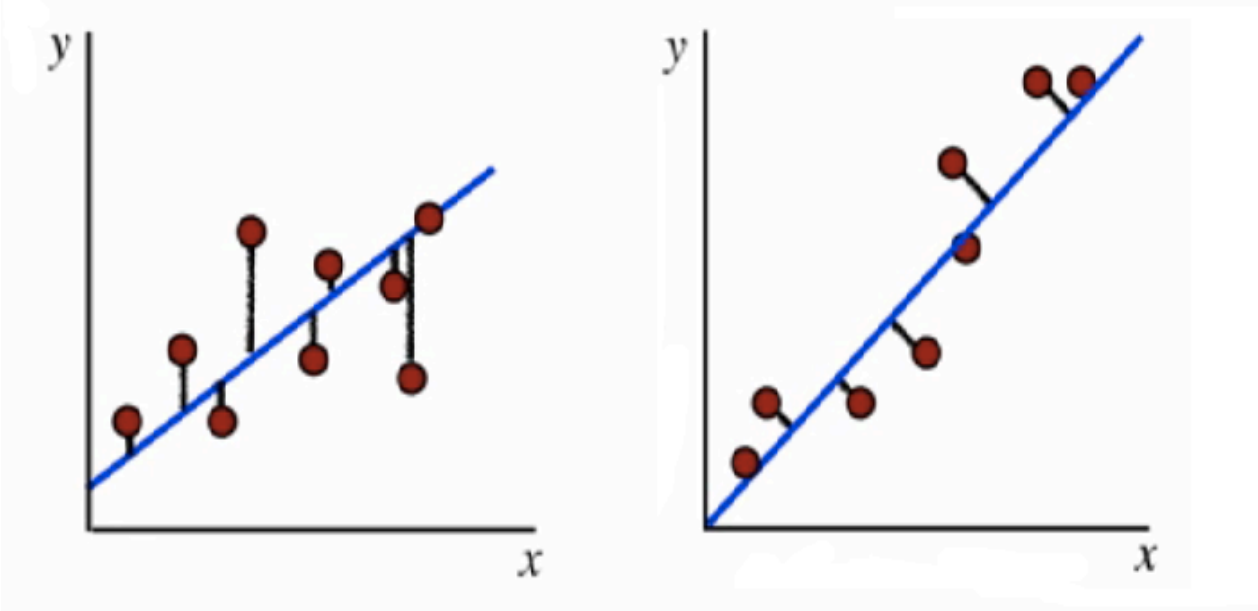


- A Spark program consists of two programs, a driver program and a workers program.
- The drivers program runs on the driver machine.
- The worker programs run on cluster nodes or in local threads.
- Then RDDs are distributed across the workers.

What is PCA

Linear Regression – predict y from x. Evaluate accuracy predictions by **vertical distances** between points and the line

PCA– reconstruct 2D data via 2D data with single degree of freedom. Evaluate reconstructions by **Euclidean distances**



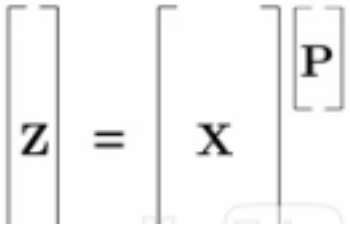
PCA solution finds direction of maximal variance
Can be done in parallel mode

PCA formulation

PCA: find lower-dimensional representation of raw data

- $\mathbf{X}$ is $n \times d$ (raw data)
- $\mathbf{Z} = \mathbf{XP}$ is $n \times k$ (reduced representation, PCA 'scores')
- $\mathbf{P}$ is $d \times k$ (columns are k principal components)

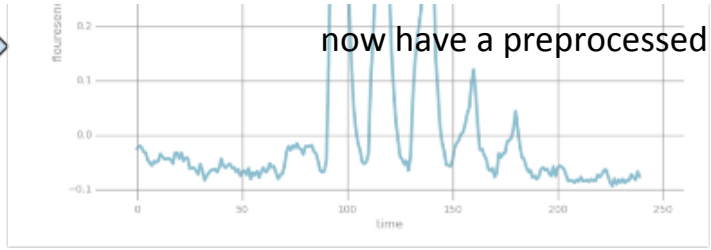
Linearity assumption ($\mathbf{Z} = \mathbf{XP}$) simplifies problem



PCA and Spark in scratch

dataset with n=46460 pixels and d=240 seconds of time series data for each pixel.

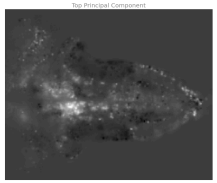
Load neuroscience data → ~~Normalized pixel intensity~~

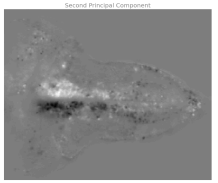


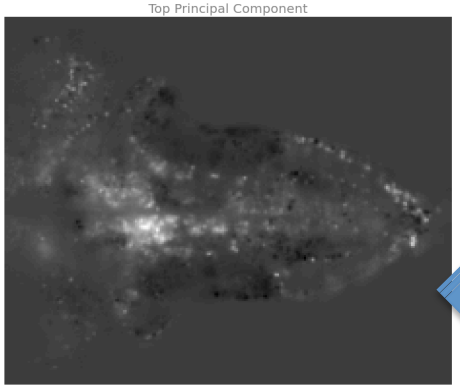
now have a preprocessed dataset

```
import os
baseDir = os.path.join('data')
inputPath = os.path.join('cs190', 'neuro.txt')

inputFile = os.path.join(baseDir, inputPath)

lines = sc.textFile(inputFile)
print lines.first()[0:100]

# Check that everything loaded properly
assert len(lines.first()) == 1397
assert lines.count() == 46460
```

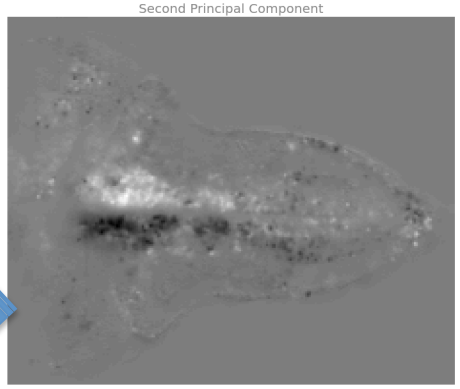
```
# Run pca using scaledData
componentsScaled, scaledScores,
eigenvaluesScaled =
pca(scaledData.values(),3)
```



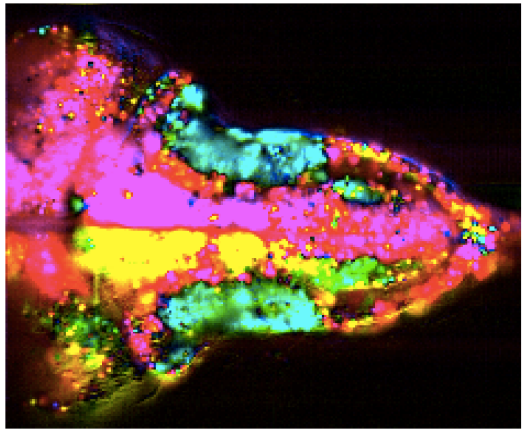




Top principal component



Second principal component



Top two components as one image

Dynamic causal modeling

Aim

The aim of dynamic causal modeling (DCM) is to infer the causal architecture of coupled or distributed dynamical systems.

(How our brain work)

Model

DCM formulated in terms of ordinary differential equations, that model dynamics of hidden states in the nodes of a probabilistic graphical model, where conditional dependencies are parameterised in terms of directed effective connectivity.

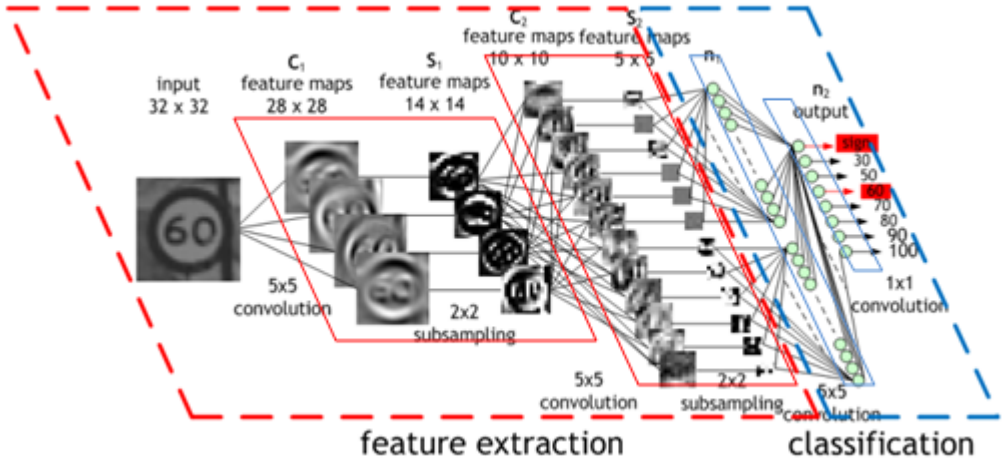
(graph of n interacting brain regions)

Choosing

Bayesian model comparison procedure that rests on comparing models of how data were generated.

CNN

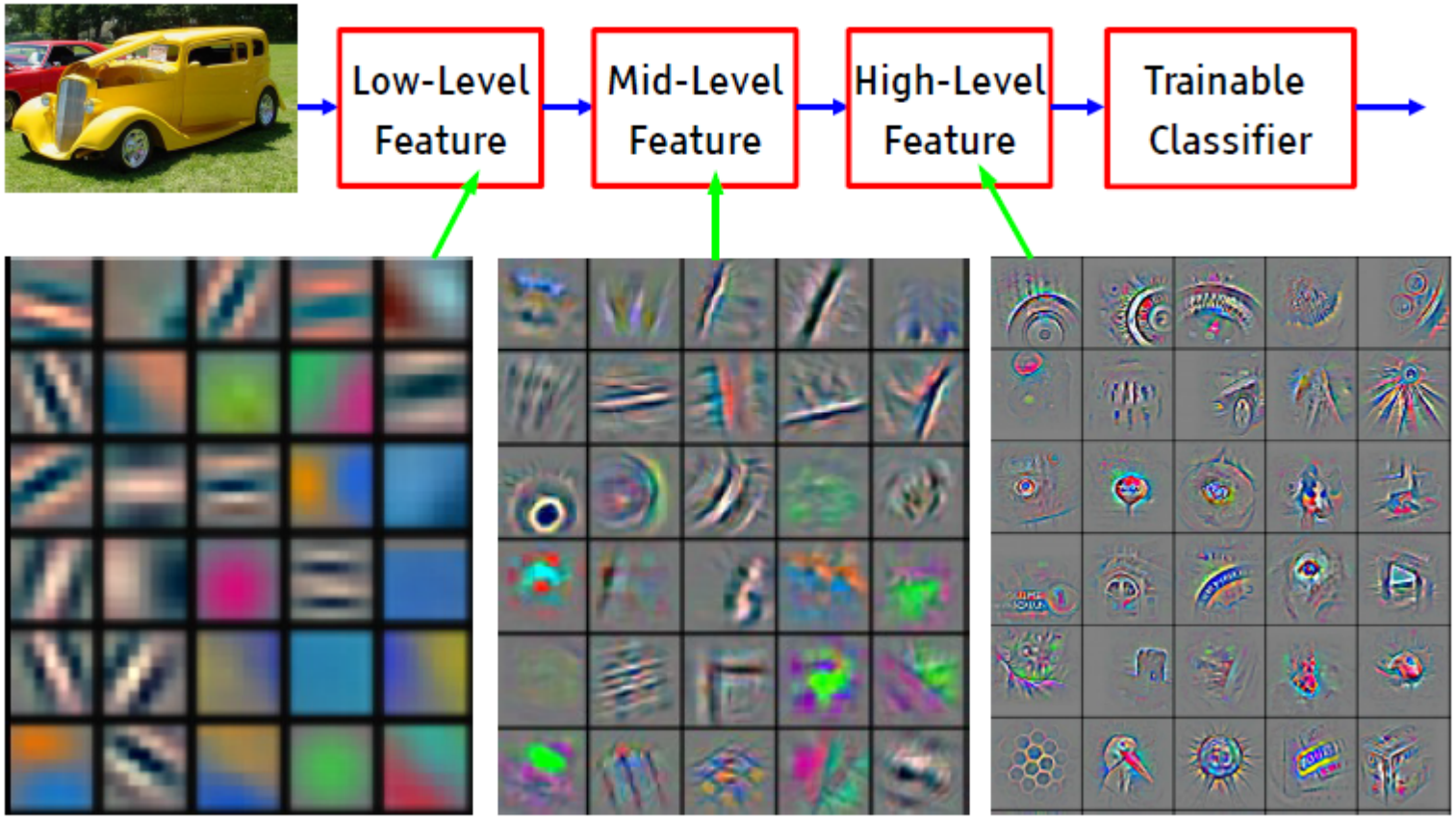
Previous supervised learning approach –
Recognition = Art of Feature Creation + Learning Technique
Convolutional Neural Network – feature based supervised learning without need of feature creation
CNN based recognition = Technique of feature extraction from data set + Learning Technique



CNN

Feature extraction + learning classifier

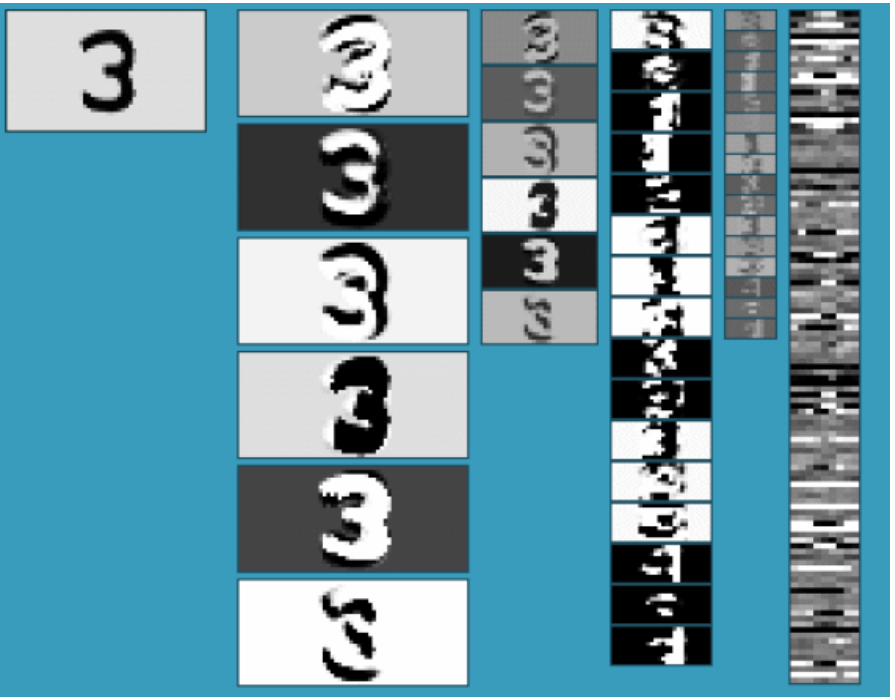
- It's deep if it has more than one stage of non-linear feature transformation



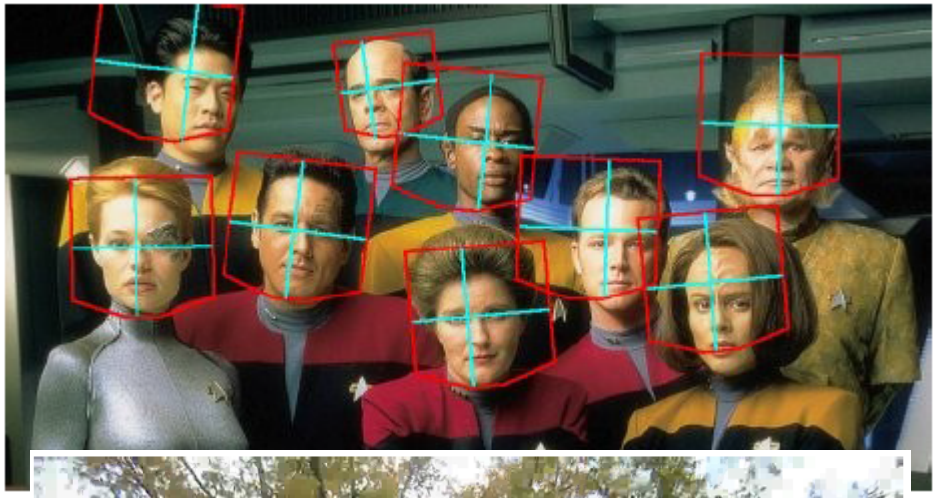
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Typical and semi-typical apps

Object and text recognition

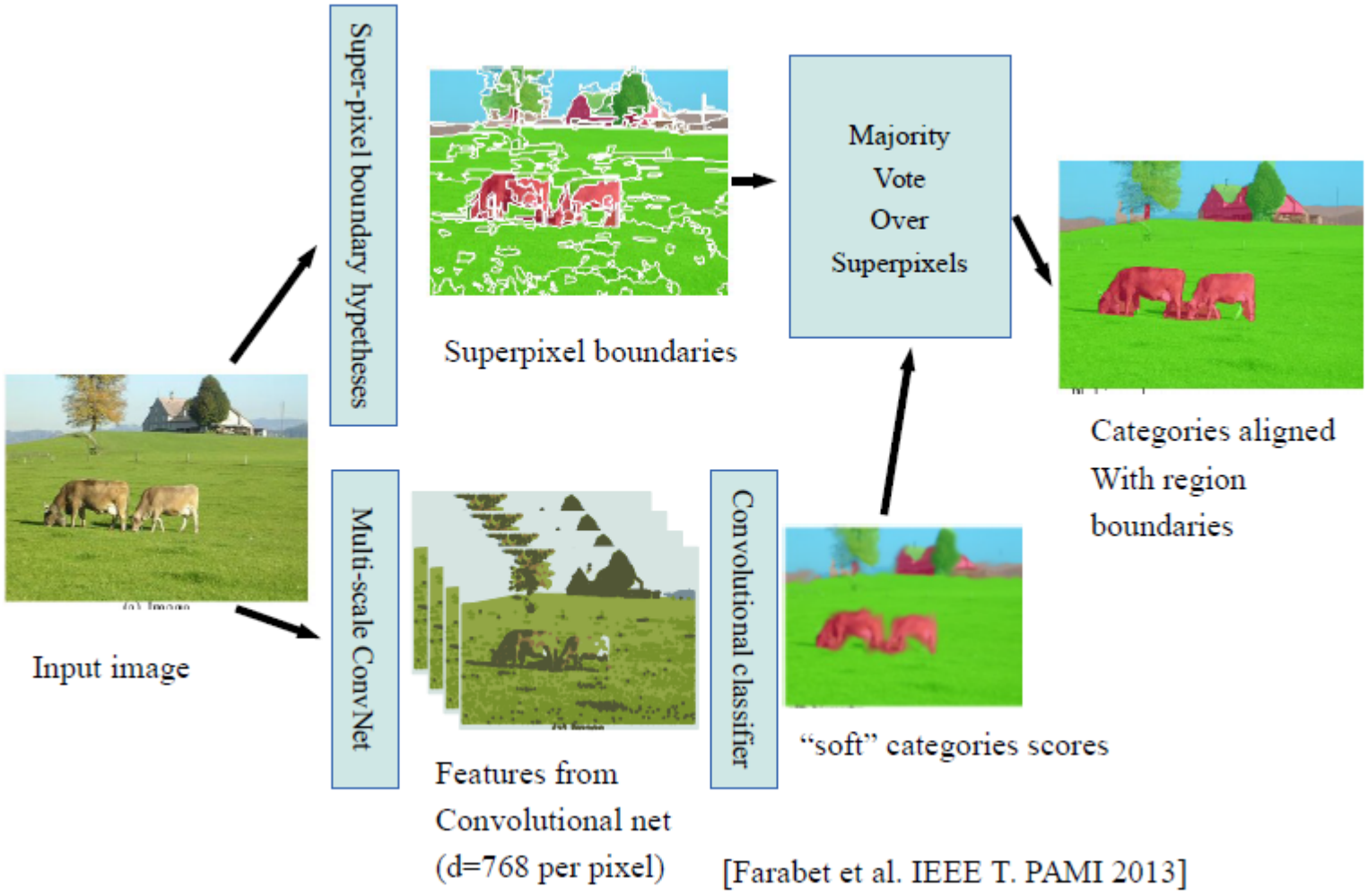


Detection



Combined nets for complex tasks

Scene segmentation



[Farabet et al. IEEE T. PAMI 2013]

Deep Nets = Big Data + GPUs

Huge training set is better then cute Net architecture!

- The ImageNet dataset [Fei-Fei et al. 2012]
  - 1.2 million training samples
  - 1000 categories

- Fast Graphical Processing Units (GPU)
  - Capable of 1 trillion operations/second


Matchstick


Sea lion


Flute


Strawberry


Backpack


Bathing cap


Racket

Huston, we've got a problem

Huge data is a solution for a lot of questions, but…

1. Why are CNN good architecture?
2. How many layers do we need?
3. How many free parameters do we need?
4. What can we do with local minima?
5. What about spatial relations?
6. We have neurons – layers – net itself – is it enough?

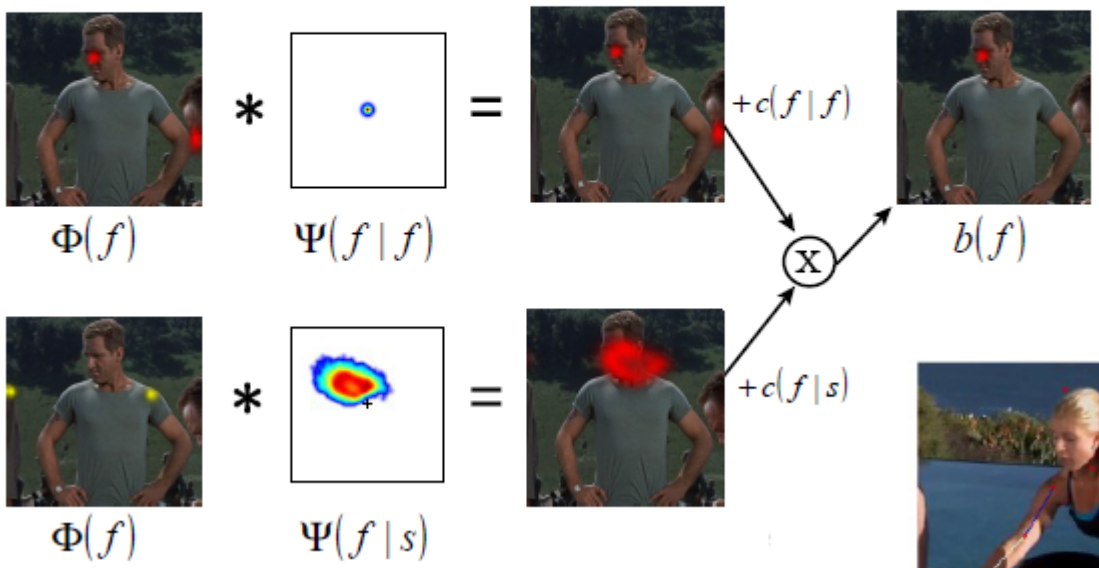
CNN + CRF = Spatial correlations

Learning Graphs model over the CNN

Start with a tree graphical model

... And approximate it

$$b(f) = \Phi(f) \prod_i (\Phi(x_i) * \Psi(f \mid x_i) + c(f \mid x_i))$$

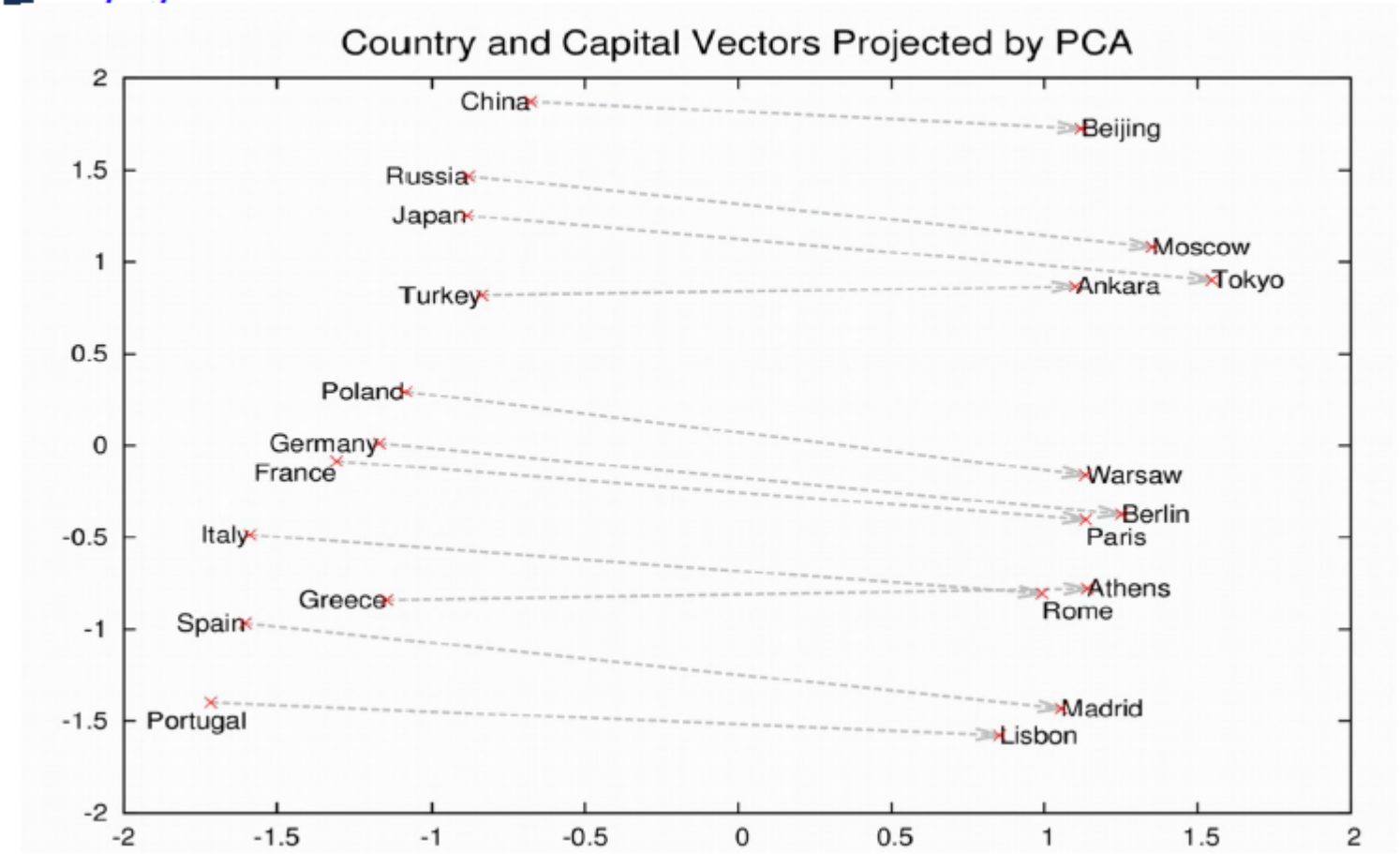


Tompson, Goroshin, Jain, LeCun, Bregler arXiv:1411.4280 (2014)

Implementing memory

Bag-of-words – simple associative NN-like memory for billions of words.

Beijing – China + France = Paris



Caffe

Key architecture idea – a config-file with network's configuration and training strategy

Pro

- You not need to make a program in order to start
- GPU in the box
- You are in good company (Berkley, MIT, Google)
- Growing fast

Contra

- You need to know how to make a program (if you want a bit more then a prototyping)
- Growing too fast (even without backward compatibility!)
- Better for contest than for production