Machine learning for sciences

Balázs Kégl
Université Paris-Saclay / CNRS

OUTLINE

- The bumpy 60-year history of ML
- ML in sciences
  - use cases and challenges
  - examples
  - the RAMP tool
- Automatic hypothesis generation

Machine learning is an engineering toolkit for induction

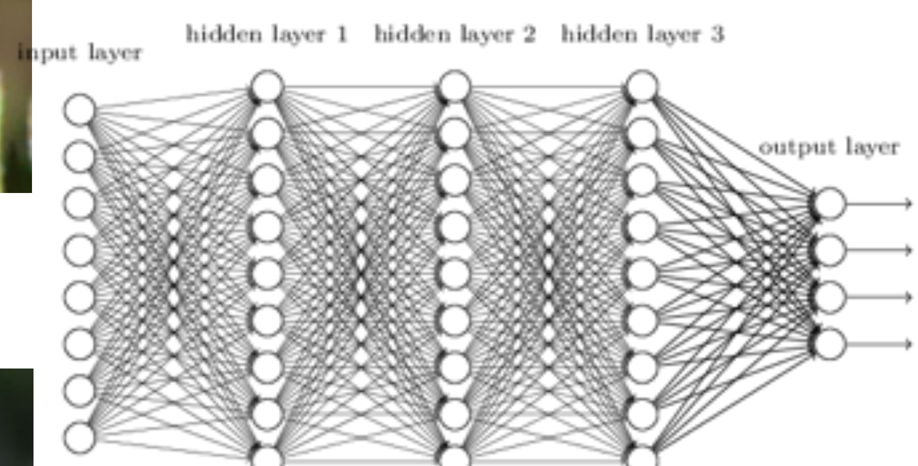
Learn a function $y = f(x)$ from a large number of (x, y) pairs

DATA-DRIVEN INFERENCE

- **Classification** problem $y = f(x)$

x

f → y *'Stomorhina'*



x

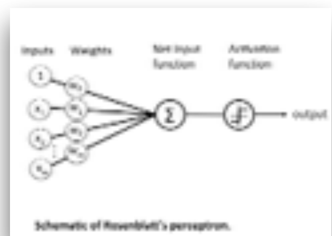
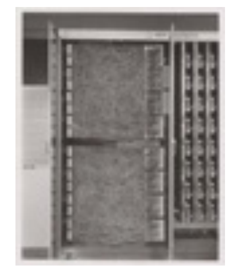
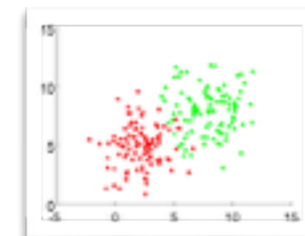
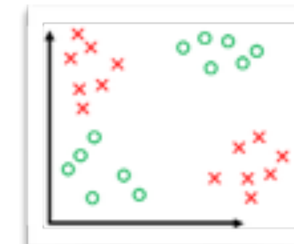
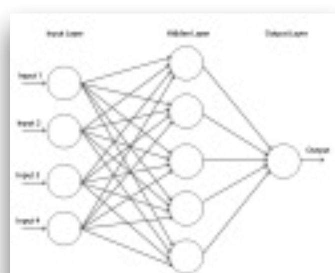
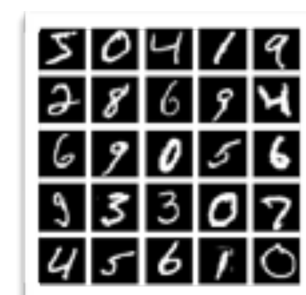
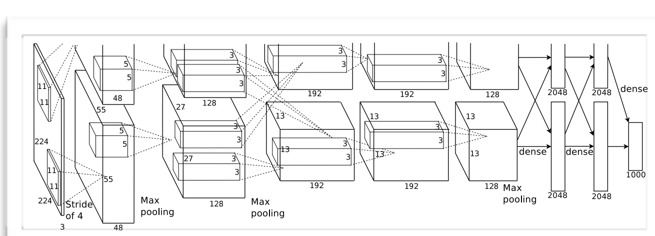
f → y *'Scaeva'*

Data-driven inference

- **Classification** problem $y = f(x)$
- **No model to fit**, but a **large set** of (x, y) pairs
  - The source is typically **observation** + human **labeling**
  - Or **computer simulation**
- And a **loss function** $L(y, y_{pred})$

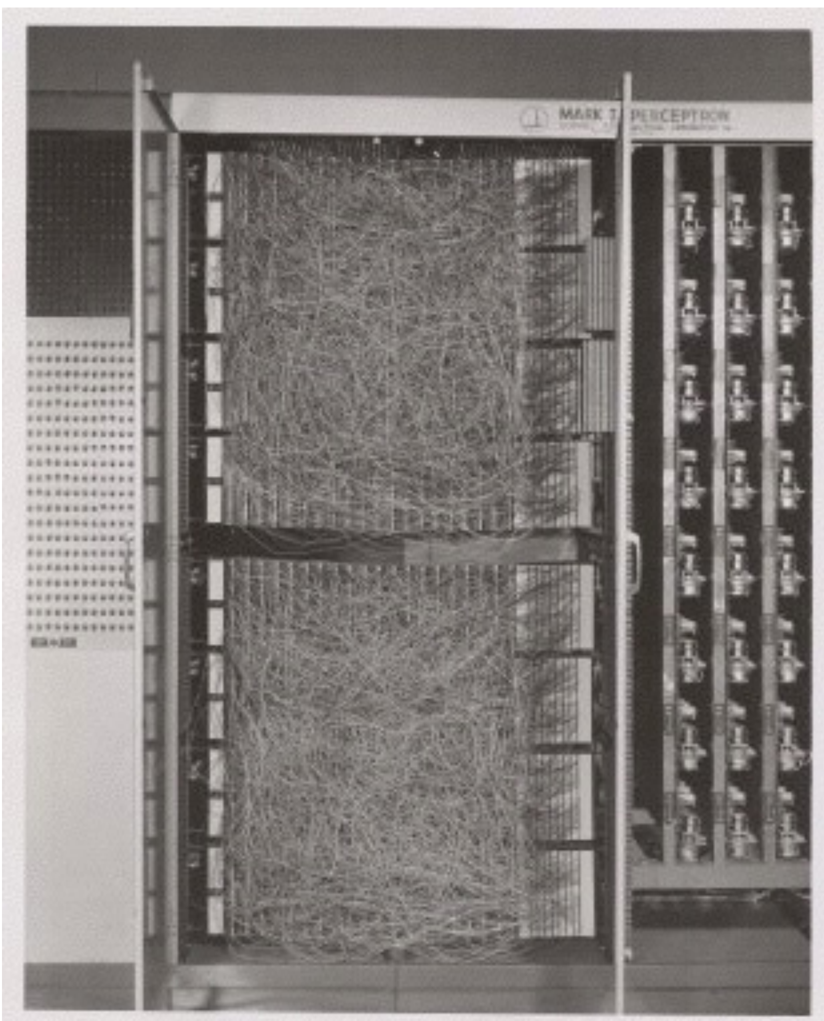
The bumpy 60-year history that led to the current state of the art

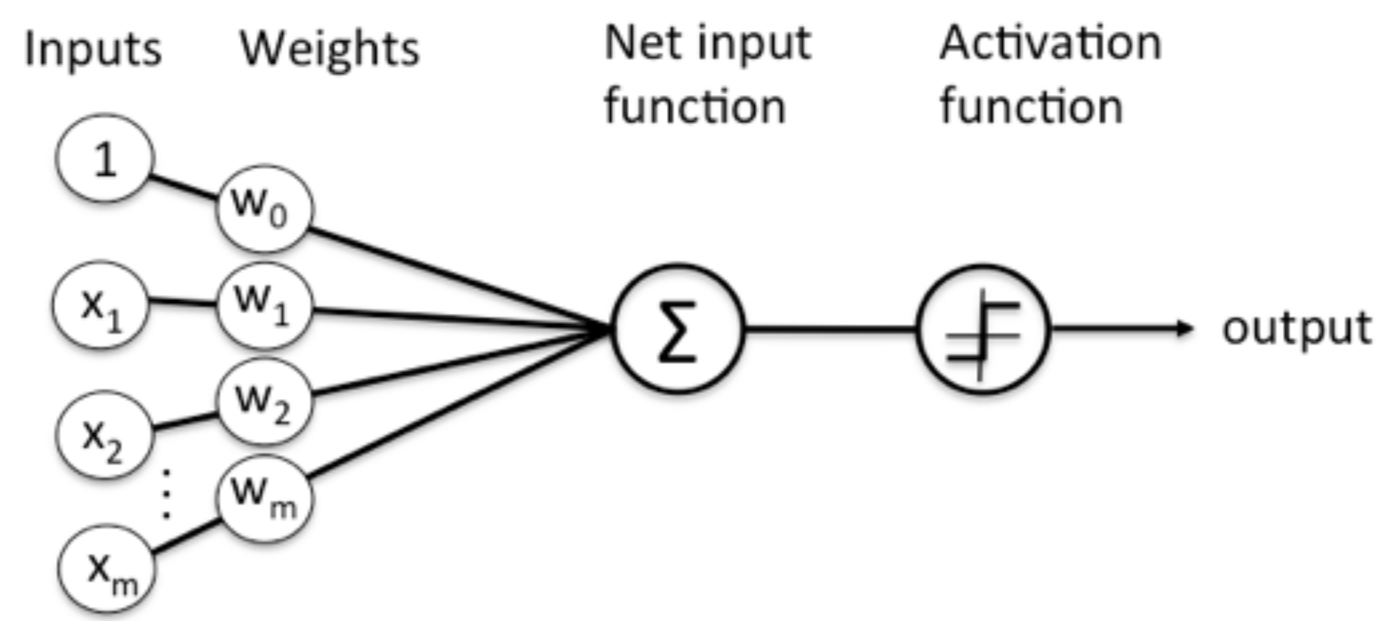
Deep learning = three intertwining story

	techniques / tricks	hardware	data
1957-69 dawn	perceptron	early mainframes	toy linear, small images, XOR
			 
1986-95 golden age	early NNs	workstations	MNIST
			
2006- deep learning	deep NNs	GPU, TPU, Intel Xeon Phi	Imagenet
		  	

THE PERCEPTRON (ROSENBLATT 1957)

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$





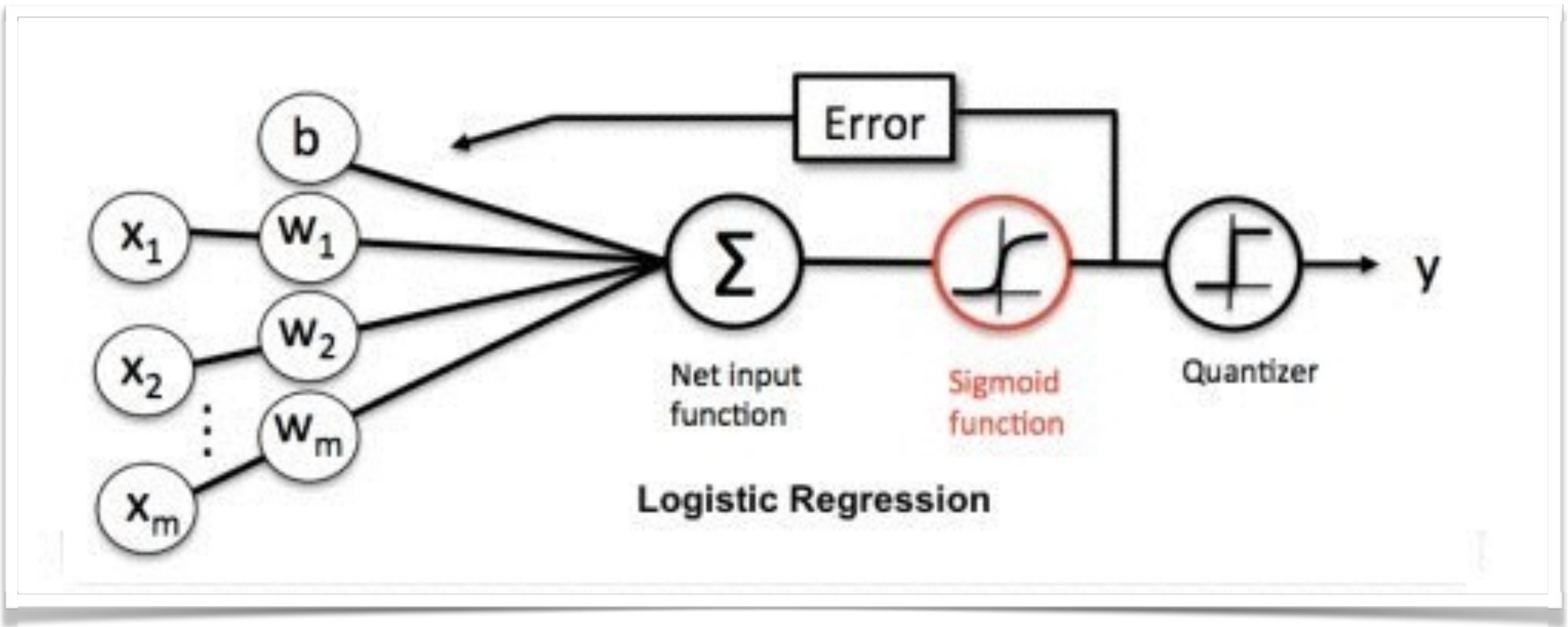
Schematic of Rosenblatt's perceptron.

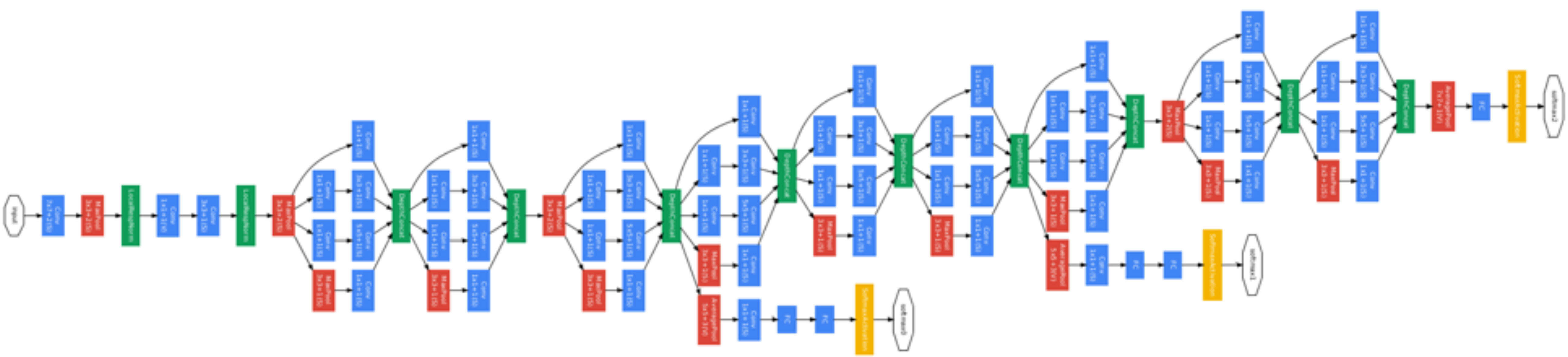
Weights were encoded in potentiometers, and weight updates during learning were performed by electric motors.

THE PERCEPTRON (ROSENBLATT 1957)

Based on Rosenblatt's statements, The New York Times reported the perceptron to be "the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

Back propagation (Rumelhart et al. 1986)

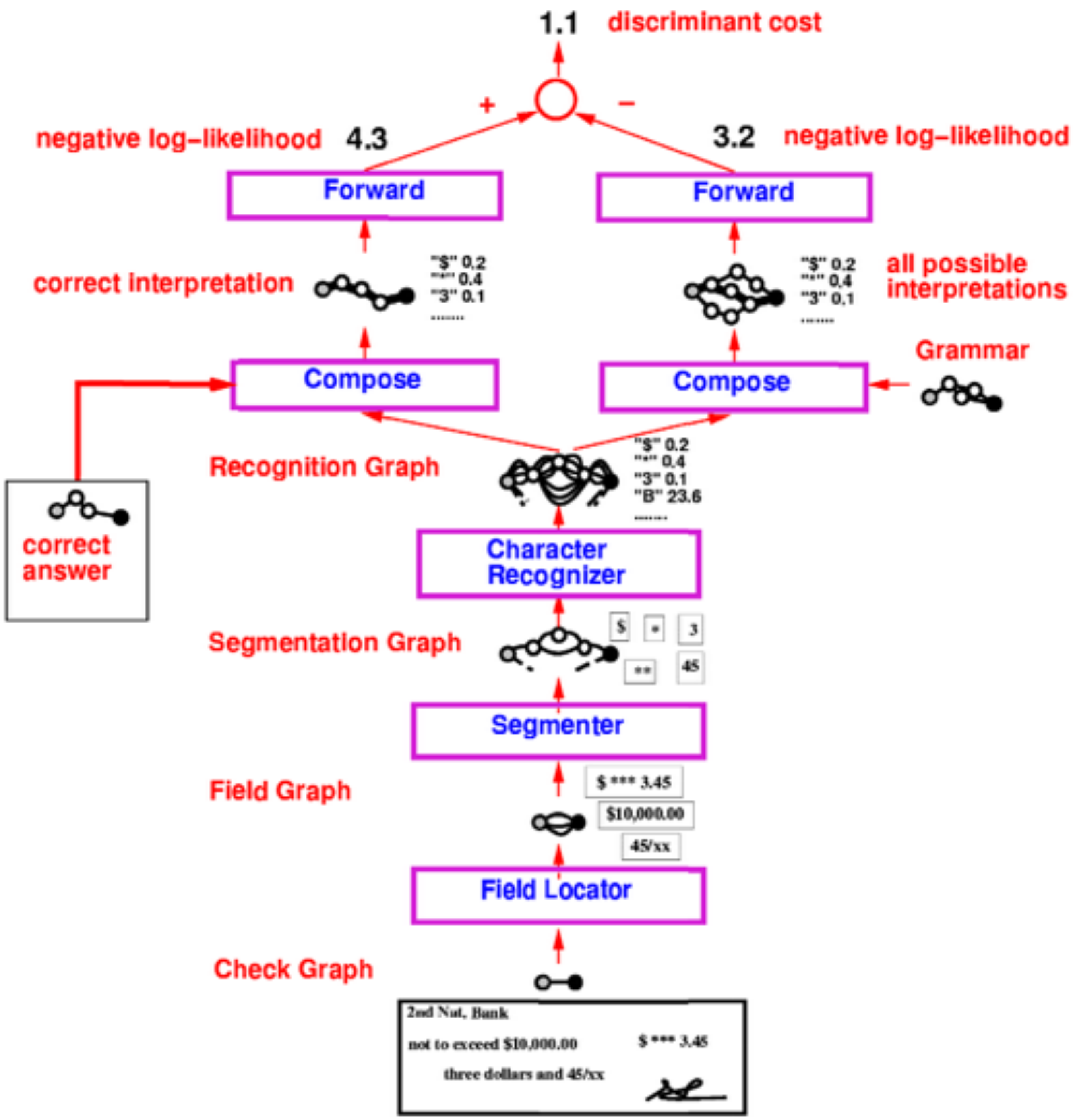




The golden age (1986-95)

- Convolutional nets
- The first algorithmic tricks: initialization, weight decay, early stopping
- Some limited understanding of the theory
- First commercial success: AT&T check reader (Bottou, LeCun, Burges, Nohl, Bengio, Haffner, 1996)

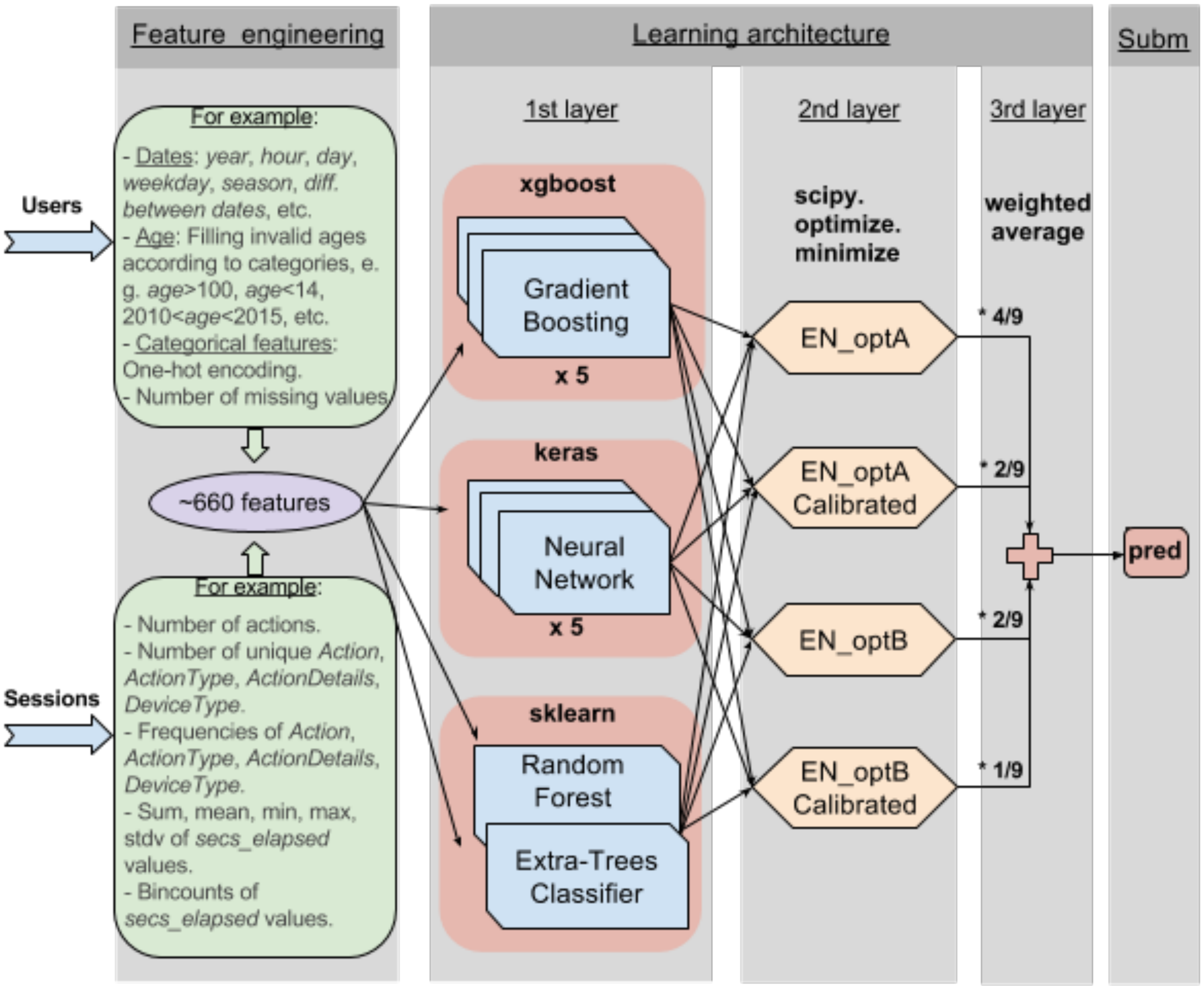
THE AT&T CHECK READER



- Reading checks is **more than character recognition**
- If all steps are differentiable, the whole system can be **trained end-to-end** by backdrop

Yann Lecun:
“Deep learning is dead,
long live
differentiable programming”

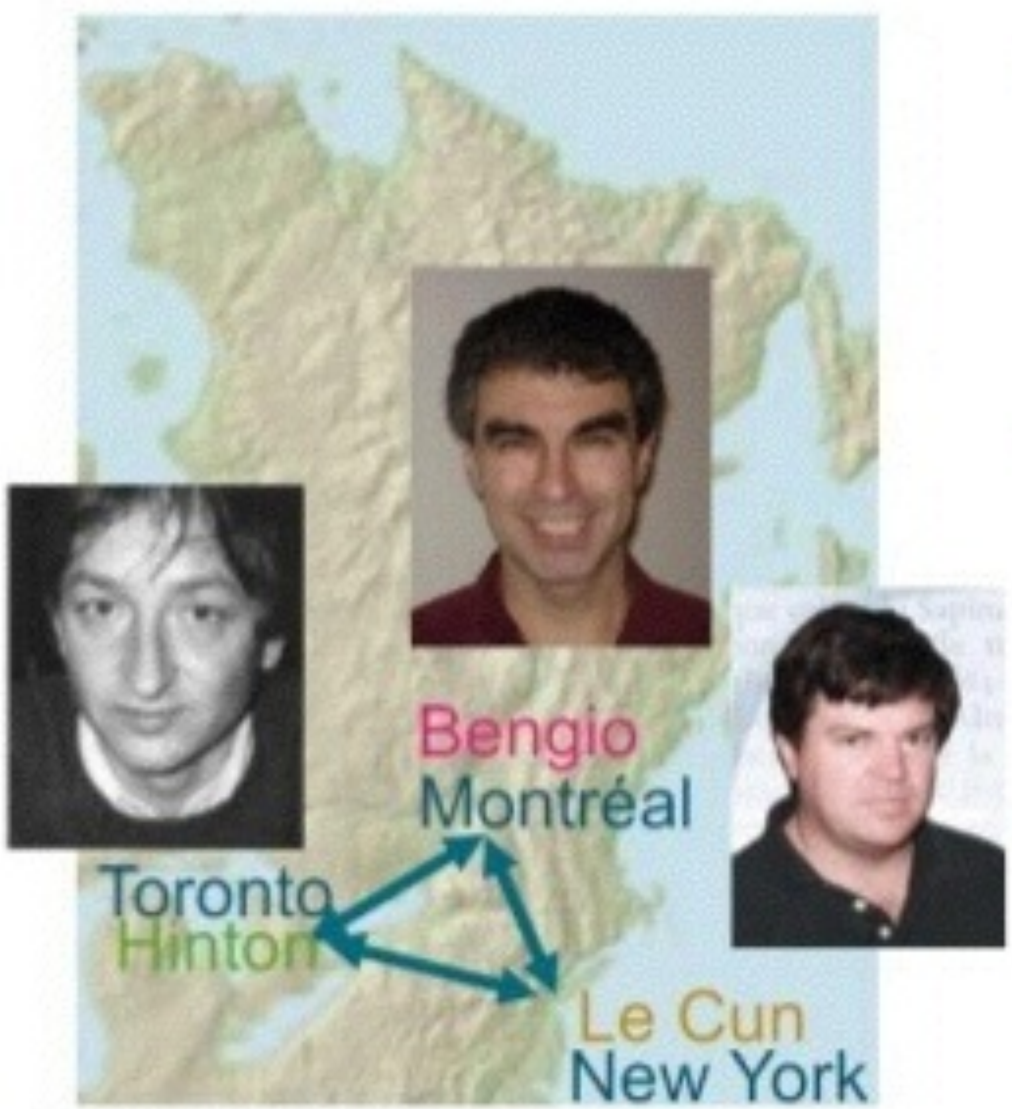
FIRST TAKE-HOME MESSAGE



Before you jump on the deep learning bandwagon: scikit-learn forests + xgboost gets >90% performance on >90% of the industrial problems, cautious estimate

2006: A NEW WAVE BEGINS

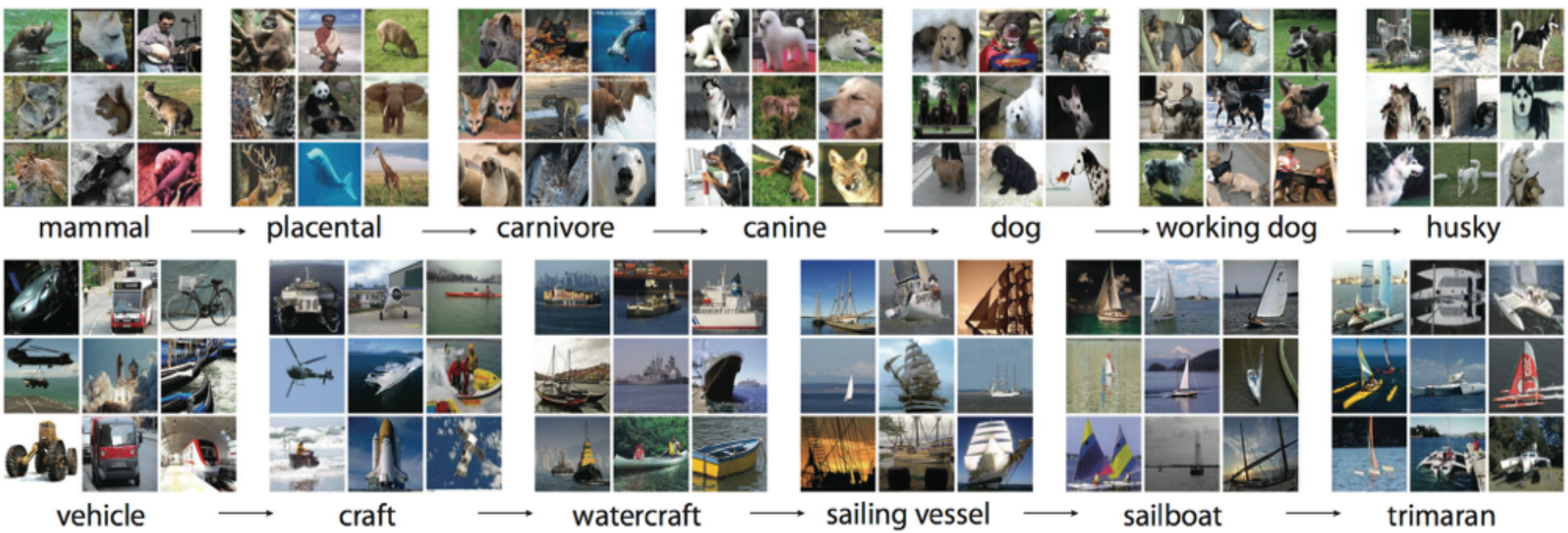
- NNs are **back on the research agenda**



- Hinton, Osindero & Teh « A Fast Learning Algorithm for Deep Belief Nets », *Neural Computation*, 2006
- Bengio, Lamblin, Popovici, Larochelle « Greedy Layer-Wise Training of Deep Networks », *NIPS'2006*
- Ranzato, Poultney, Chopra, LeCun « Efficient Learning of Sparse Representations with an Energy-Based Model », *NIPS'2006*

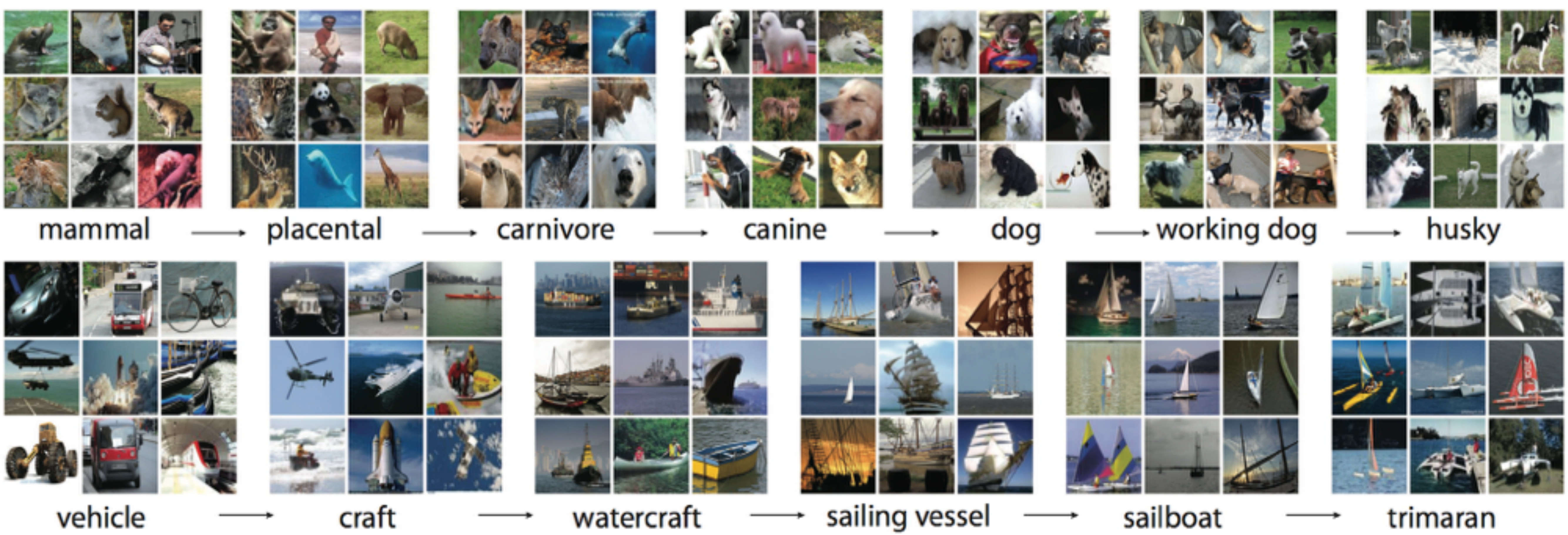
2009: IMAGENET

"We believe that a large-scale ontology of images is a critical resource for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms." (Fei Fei Li et al CVPR09)



2009: ImageNet

- 80K hierarchical categories
- 80M images of size >100x100
- labeled by 50K Amazon Turks



GPUs (2004 -)

System and method for accelerating and optimizing the processing of machine learning techniques using a graphics processing unit

US 7219085 B2

Publication number	US7219085 B2
Publication type	Grant
Application number	US 10/837,382
Publication date	May 15, 2007
Filing date	Apr 30, 2004
Priority date ?	Dec 9, 2003
Fee status ?	Paid
Also published as	CN1627251A, 6 More »
Inventors	Ian Andrew Buck, Patrice Y. Simard, David W. Steinkraus
Original Assignee	Microsoft Corporation
Export Citation	BiBTeX, EndNote, RefMan

Patent Citations (1), Non-Patent Citations (11), Referenced by (35), Classifications (13), Legal Events (5)

External Links: USPTO, USPTO Assignment, Espacenet

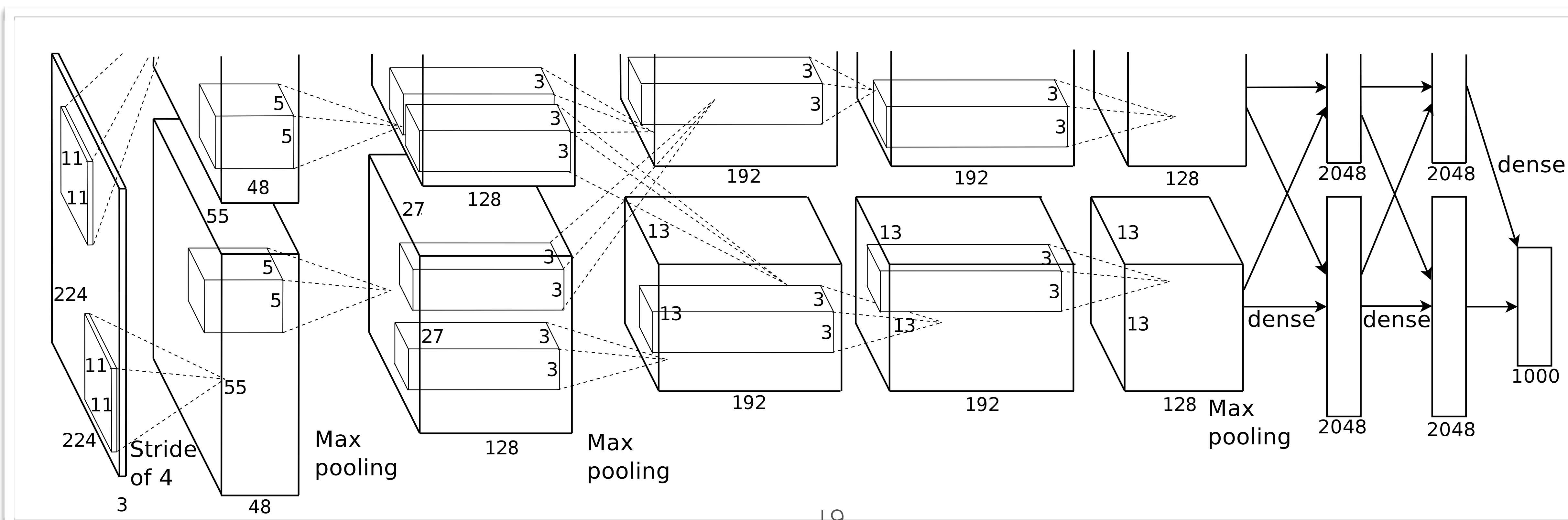
ABSTRACT

A system and method for processing machine learning techniques (such as neural networks) and other non-graphics applications using a graphics processing unit (GPU) to accelerate and optimize the processing. The system and method transfers an architecture that can be used for a wide variety of machine learning techniques from the CPU to the GPU. The transfer of processing to the GPU is accomplished using several novel techniques that overcome the limitations and work well within the framework of the GPU architecture. With these limitations overcome, machine learning techniques are particularly well suited for processing on the GPU because the GPU is typically much more powerful than the typical CPU. Moreover, similar to graphics processing, processing of machine learning techniques involves problems with solving non-trivial solutions and large amounts of data.

IMAGES (11)

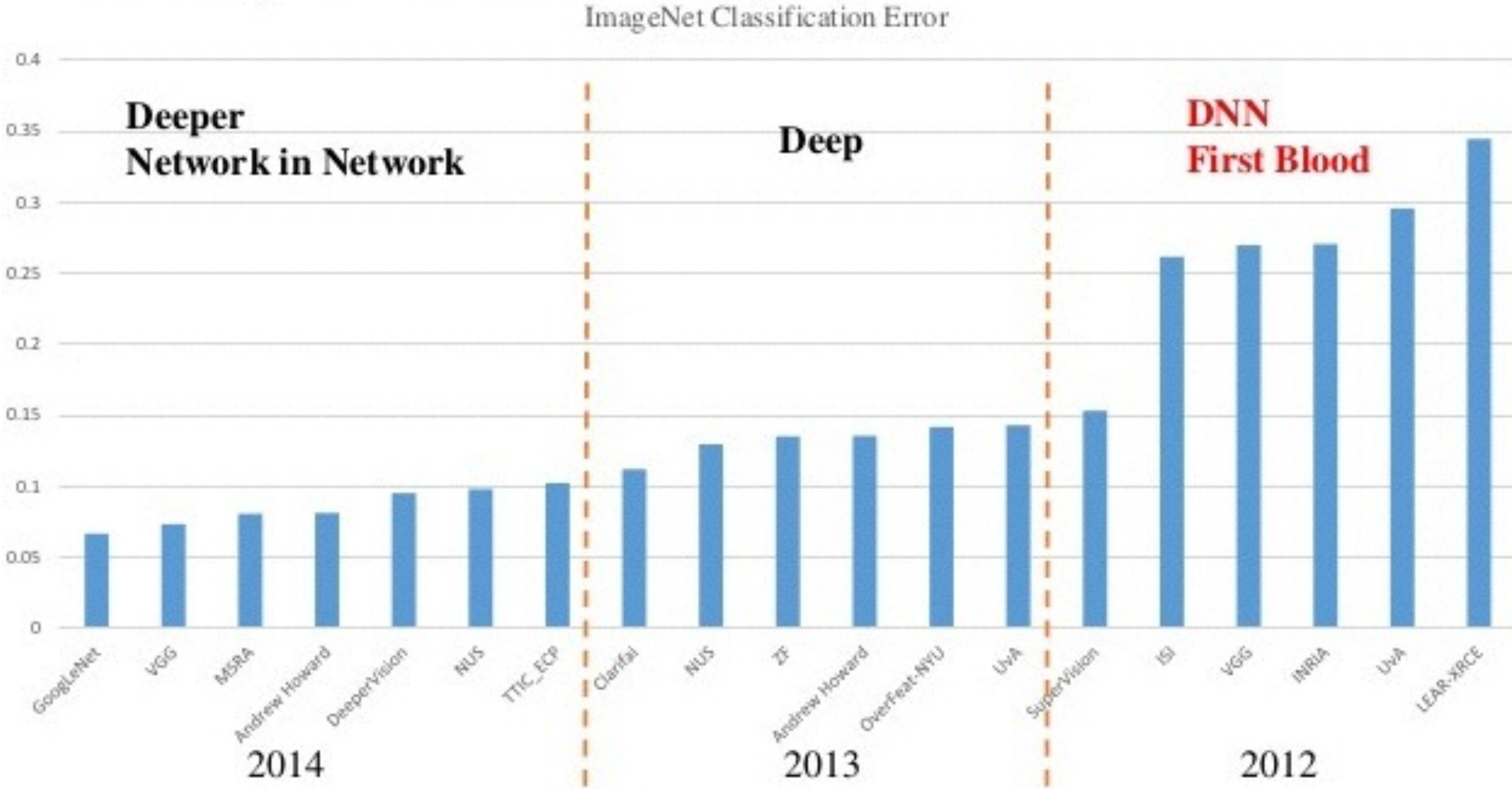
Techniques & tricks

- dropout, ReLU, max-pooling, data augmentation, batch normalization, automatic differentiation, end-to-end training, lots of layers
- Krizhevsky, Sutskever, Hinton (2012): **1.2M images**, **60M parameters**, **6 days training** on two GPUs



IMAGENET COMPETITIONS

- **1000** categories and **1.2** million training images



Li Fei-Fei: ImageNet Large Scale Visual Recognition Challenge, 2014 http://image-net.org/

SECOND TAKE-HOME MESSAGE

To make deep learning shine,
you need huge labeled data sets and time to train

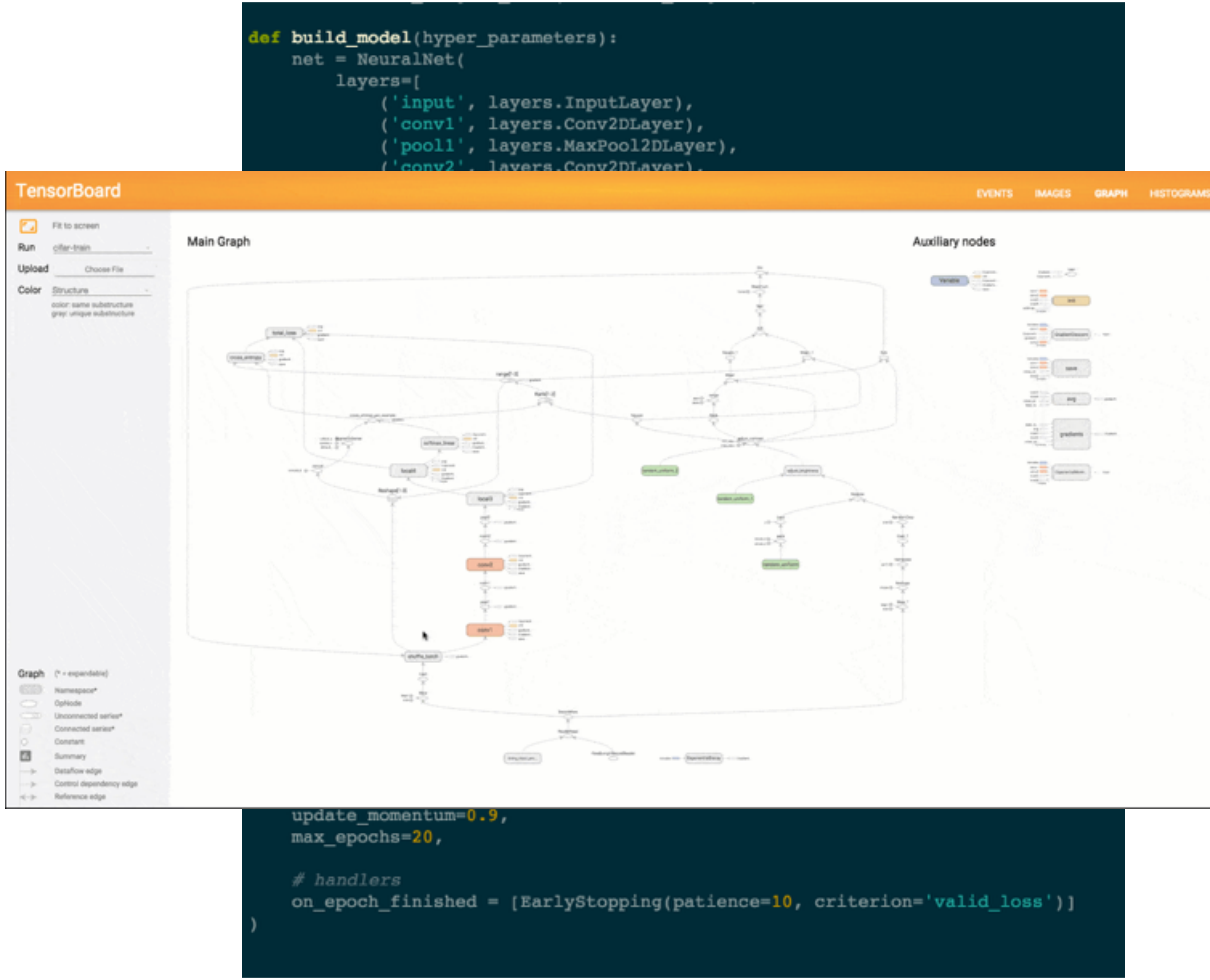
SECOND TAKE-HOME MESSAGE

To make deep learning shine,
you need huge labeled data sets and time to train

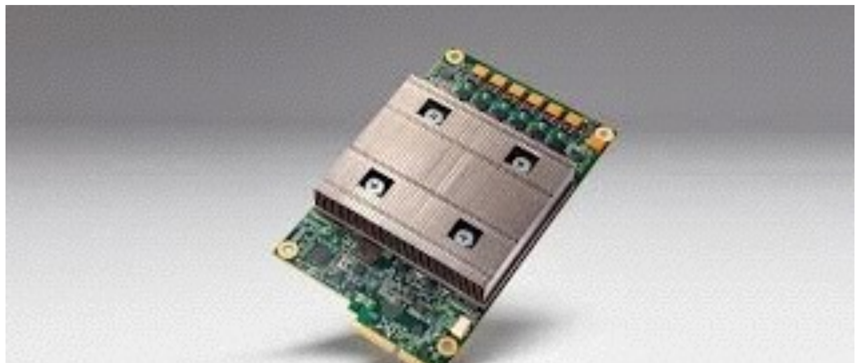
- Imagenet (80M >100x100 color images, 80K classes)
- FaceBook (300M photos/day)
- Google (300h of video/minute)

TODAY: EASY-TO-USE LIBRARIES

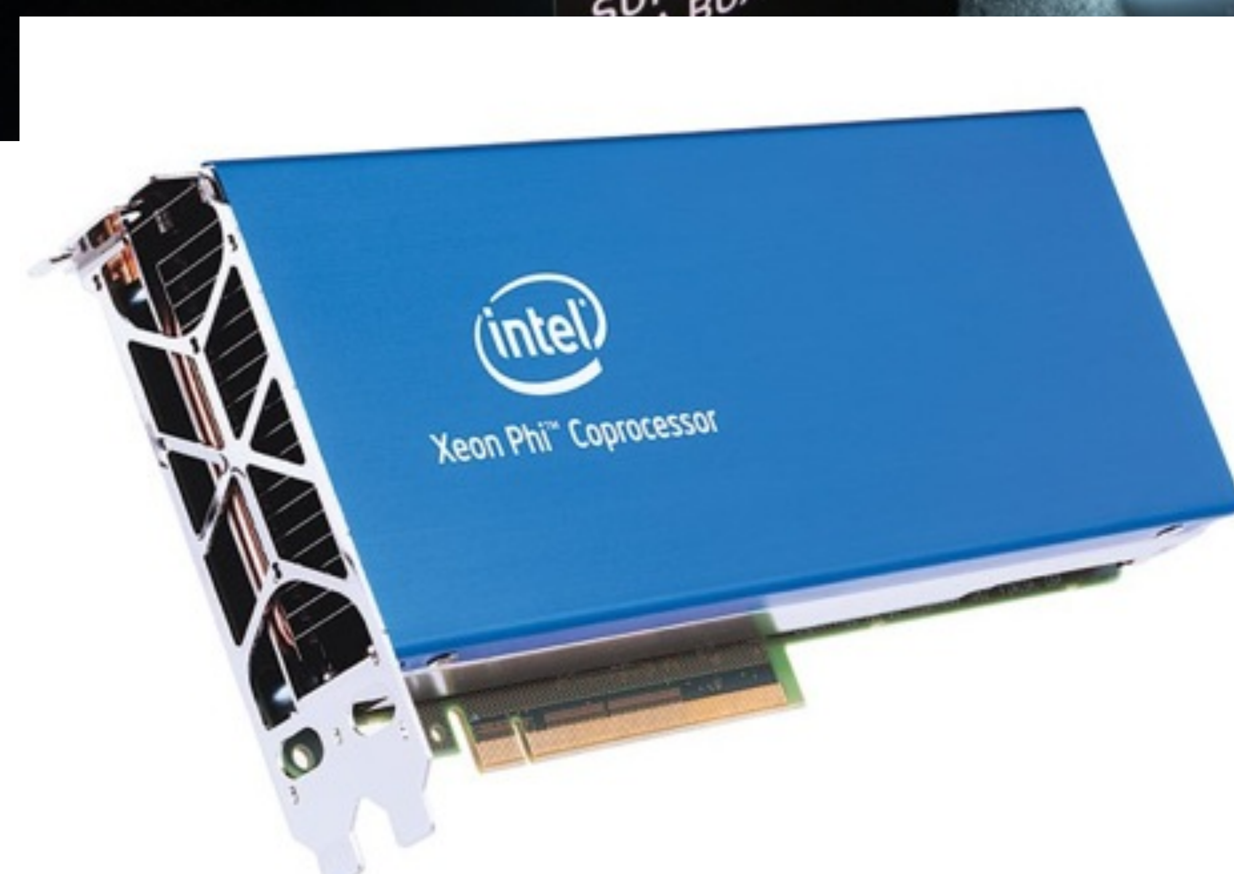
- Theano
- TensorFlow
- Keras
- Caffe
- Torch



Today: Hardware

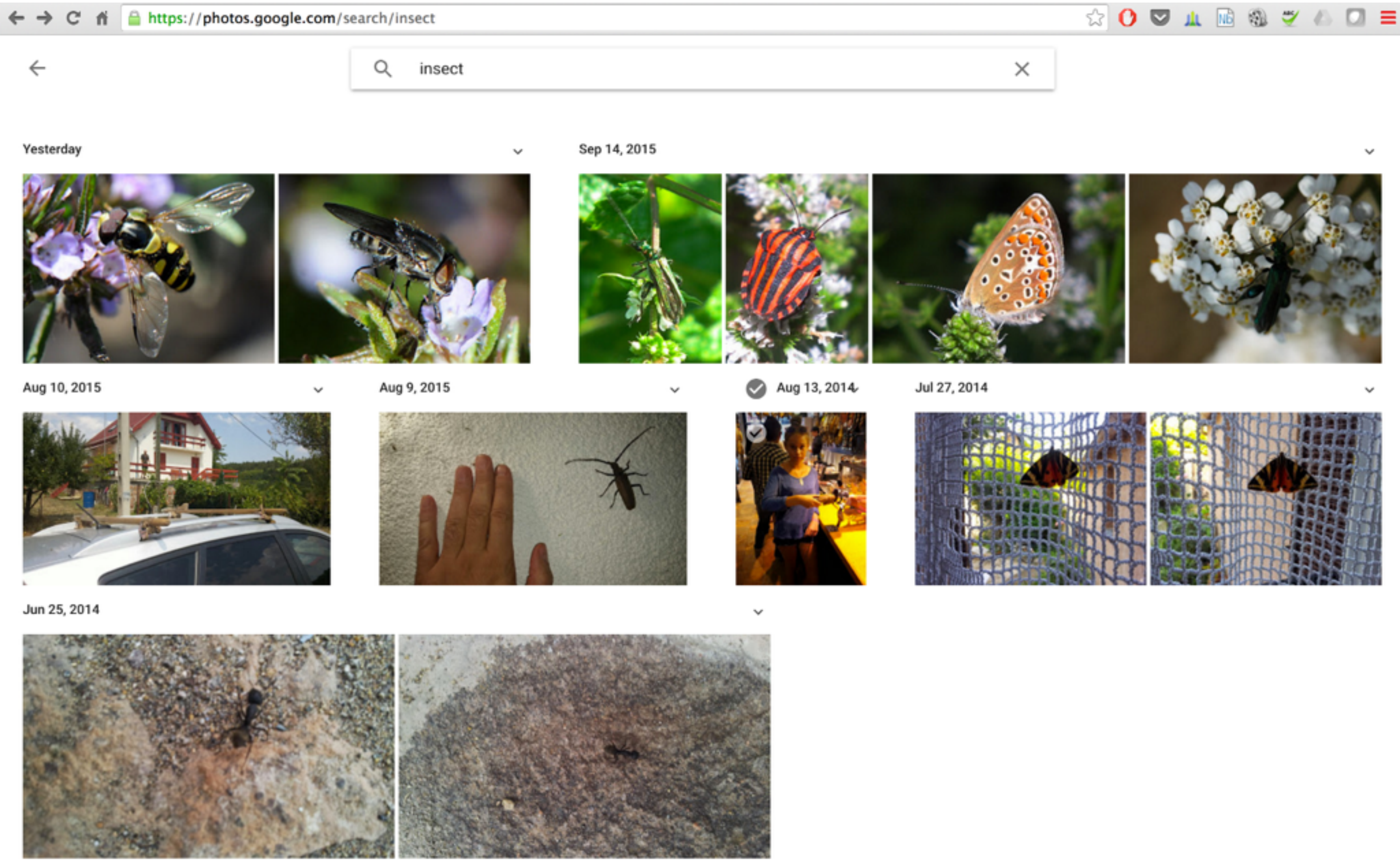


Google TPU



Commercial applications

GOOGLE IMAGE SEARCH



FACE RECOGNITION/DETECTION
A 6B$ MARKET IN 2020

Why Facial Recognition Will Change
Mark

by Ivan

221

Facial recognition tech is allowing stores to reward customers

'Who is doing what with th

By Kim Brunhuber, CBC News Poste

Facebook facial recognition tech helps parents control who is sharing pics of their children

Usher's new music video uses facial recognition technology to make sure you're paying attention to its message

Cadie Thompson
Oct. 23, 2015, 11:13 AM 884 1

FACEBOOK LINKEDIN TWITTER EMAIL PR

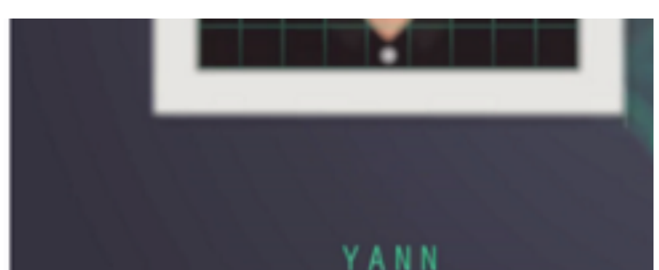
Usher has released a gripping interactive video for his new song "Chains" that requir utmost attention.

The video, which can be viewed on the streaming platform Tidal, uses facial recognit technology so that the video only plays when you are looking at the screen.

APPS

Finding Rover: The app using facial recognition software to find lost dogs

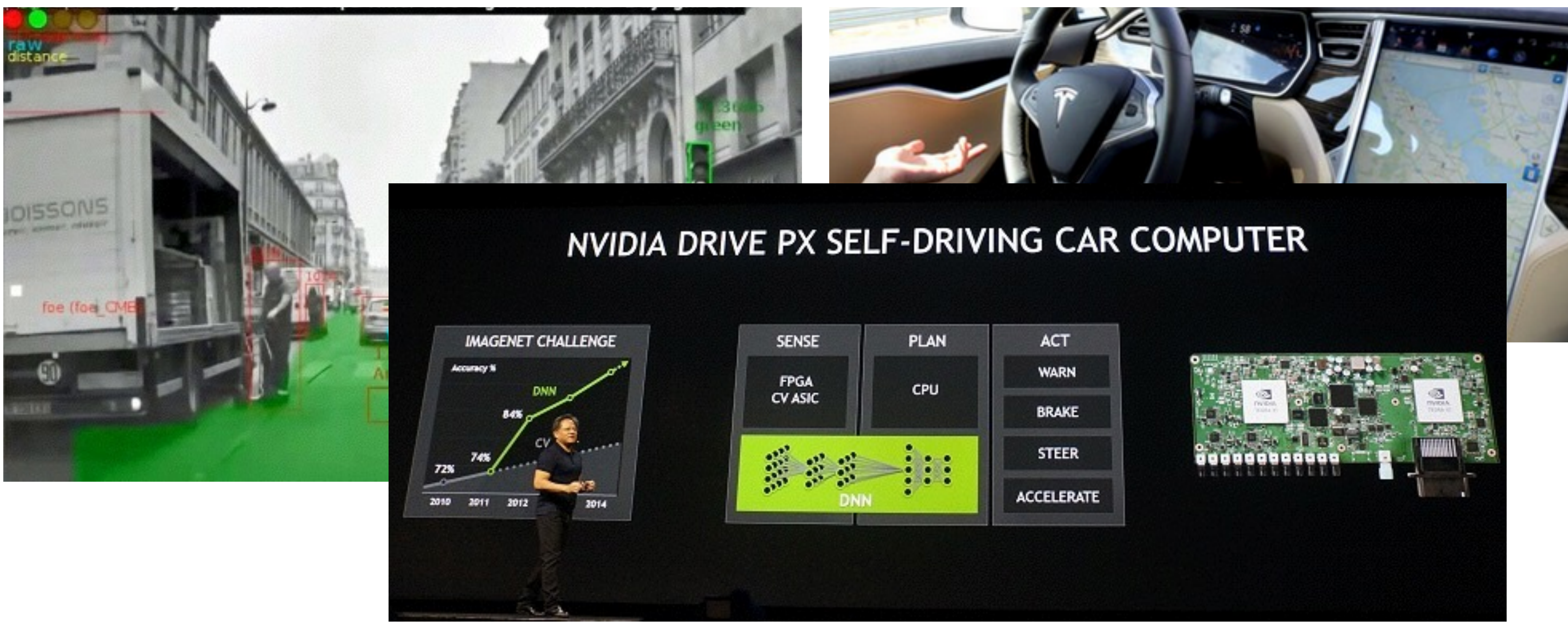
Most facial recognition systems are

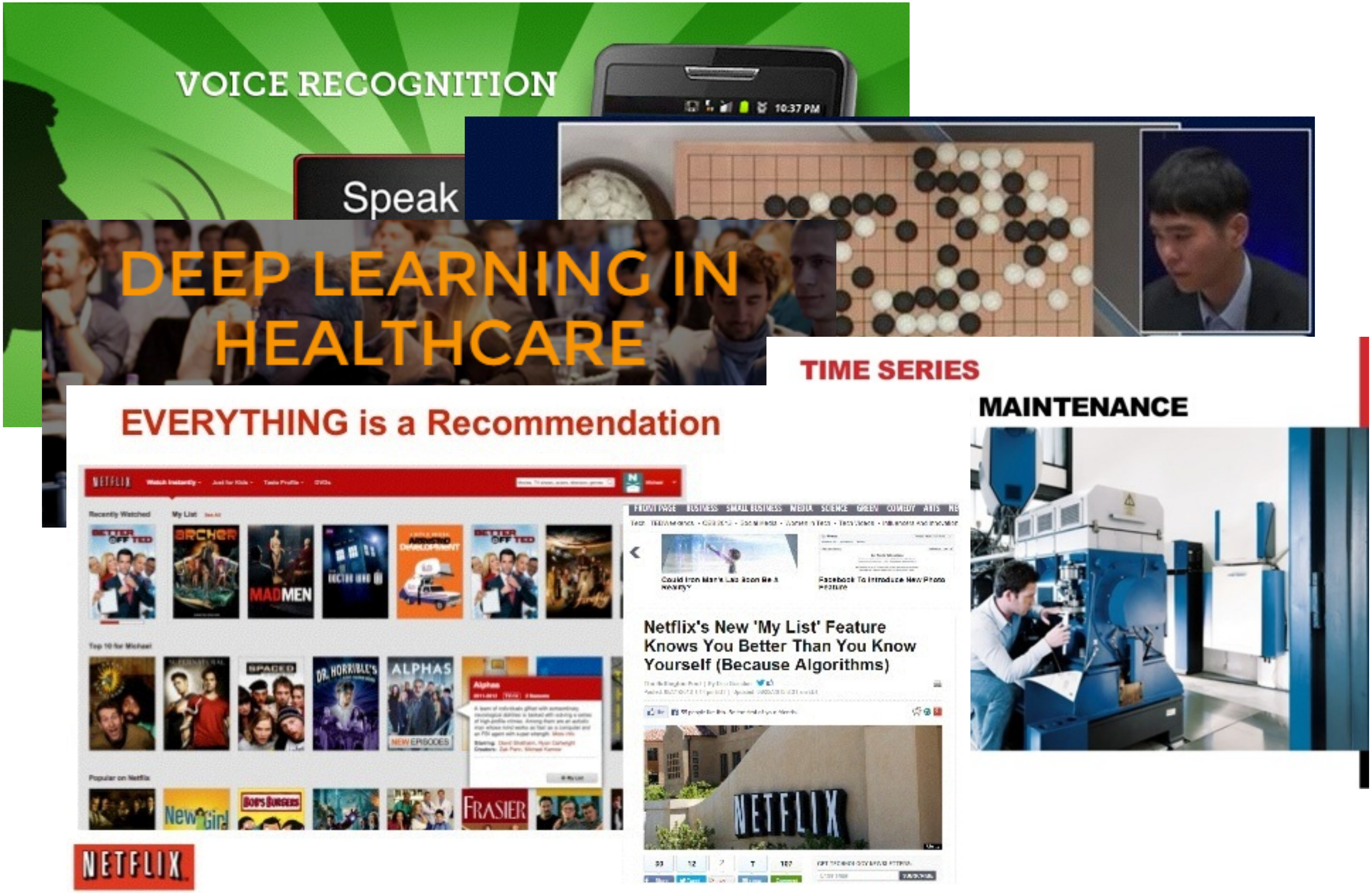


Facial recognition is the next biggest tre
pay strong attention to as it is taking fro
MasterCard announced selfie payments – a payment method using a selfie

SELF-DRVING CARS

Elon Musk Promises to Provide Self-Driving Cars within 2 Years





VOICE RECOGNITION
10:37 PM
Speak
DEEP LEARNING IN HEALTHCARE
TIME SERIES
MAINTENANCE
EVERYTHING is a Recommendation
NETFLIX
Watch Instantly
Just for Kids
Taste Profile
DVDs
Recently Watched
My List
Top 10 for Michael
Popular on Netflix
Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)
NETFLIX

ML USE CASES IN SCIENCES

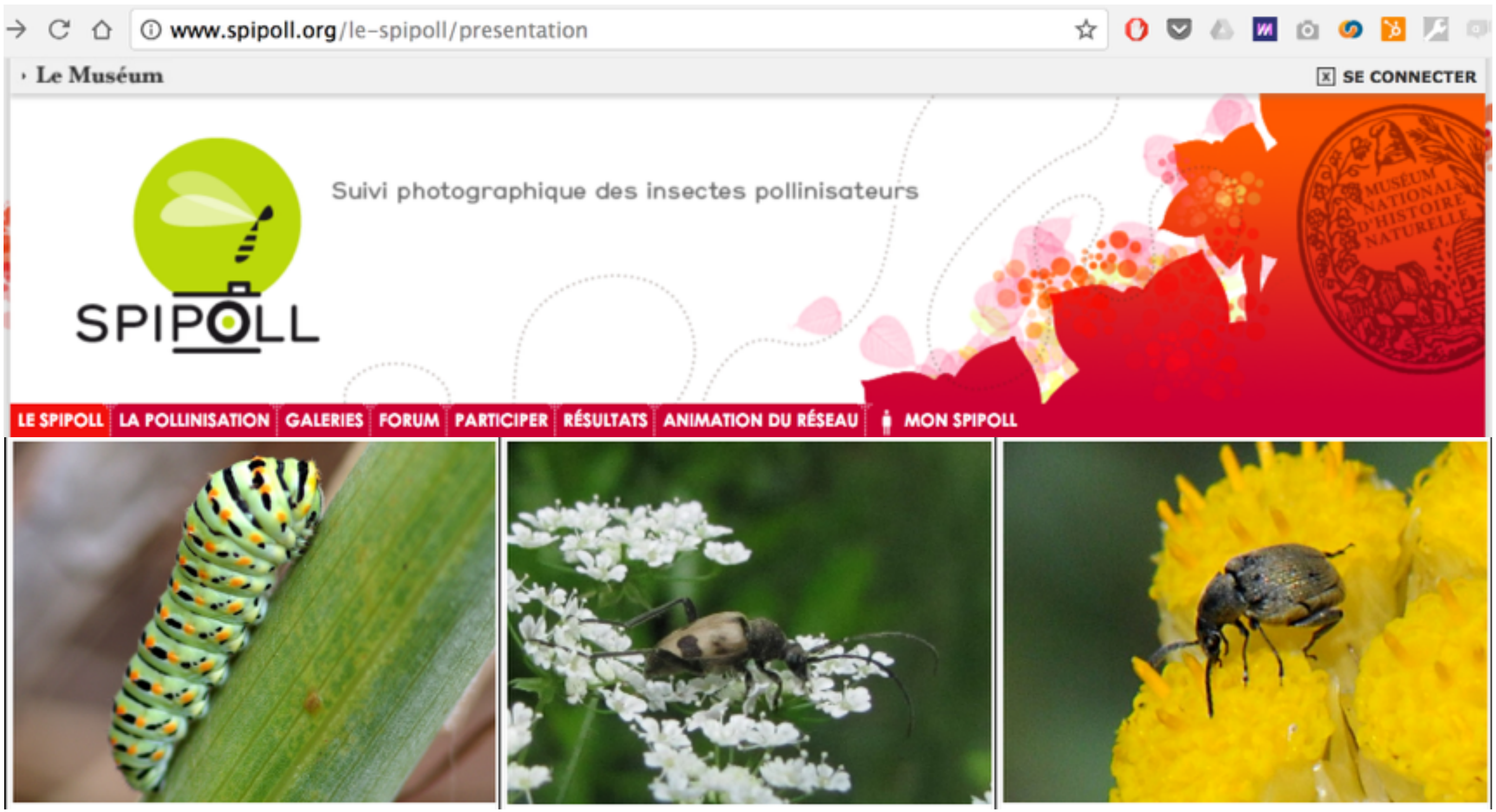
https://www.ramp.studio/problems

- **Data collection**: replace human or algorithmic collector or annotator
  - label insect photos, detect Mars craters, detect particle tracks
- **Inference**: to invert the generative model
  - "predict" a particle, detect an anomaly, **infer a parameter y from observation x**
- **Generation, model reduction**: to replace expensive simulations
  - "learn" a physics simulation or an agent based micro-economical model with a neural net
- **Hypothesis generation**: to "replace" theoreticians
  - **learn, represent structural knowledge** and **generate novelty in model space**, e.g., molecule generation in drug discovery

Data collection

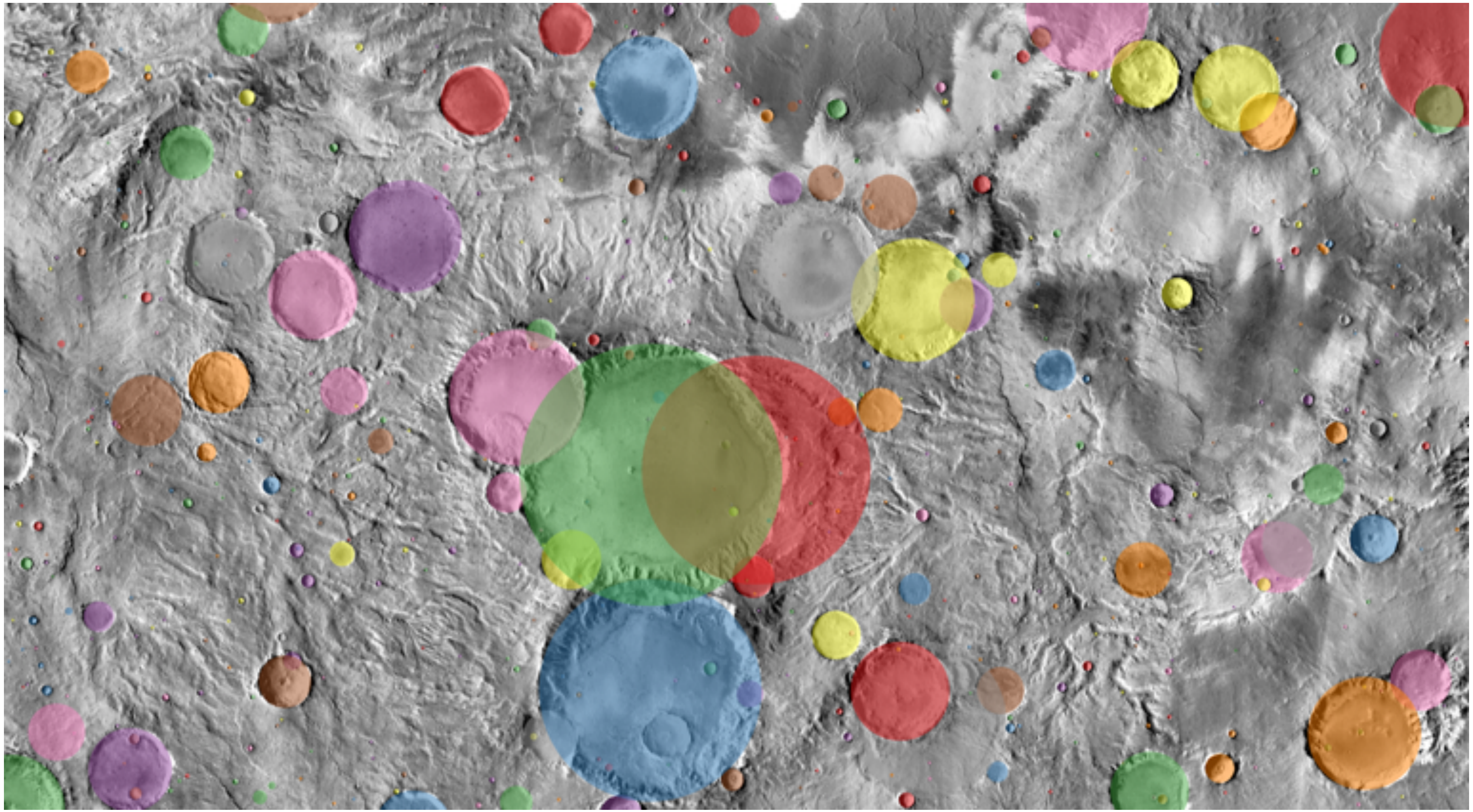
CLASSIFYING POLLENATING INSECT PHOTOS

- collaboration with ecologists at the Paris Museum of Natural History
- 400 classes, 150K photos, long tail



Detecting Mars craters

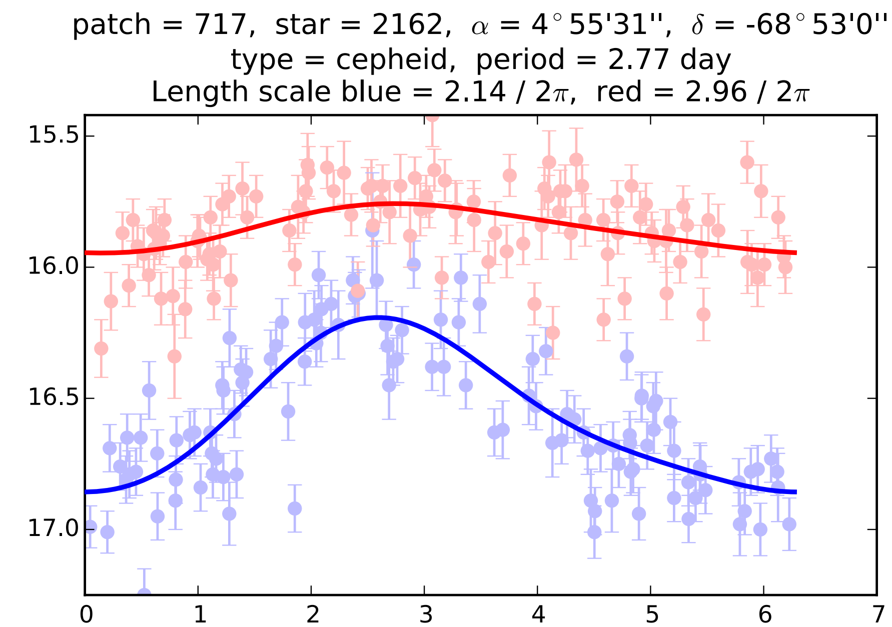
- collaboration with planetary geologists at Paris-Saclay
- complex metrics and detection workflow

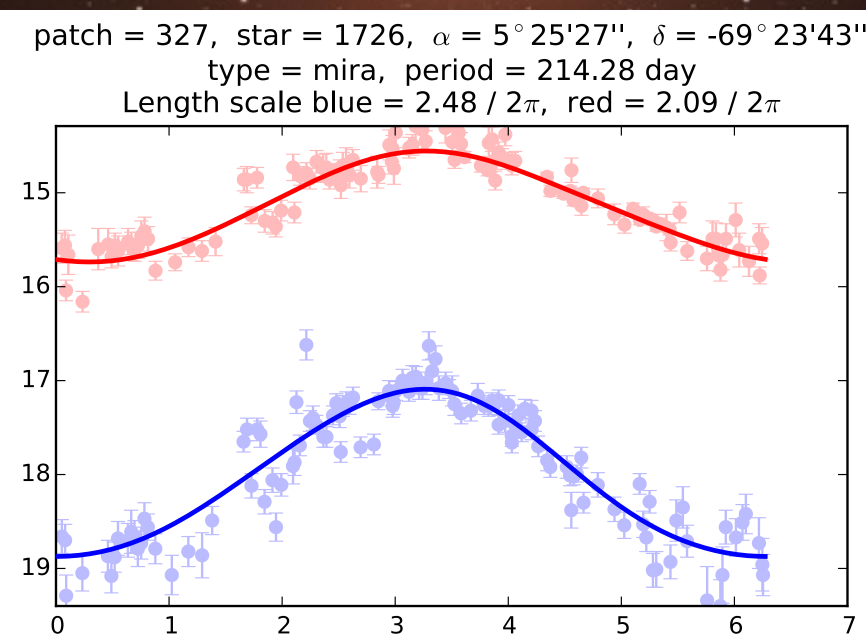


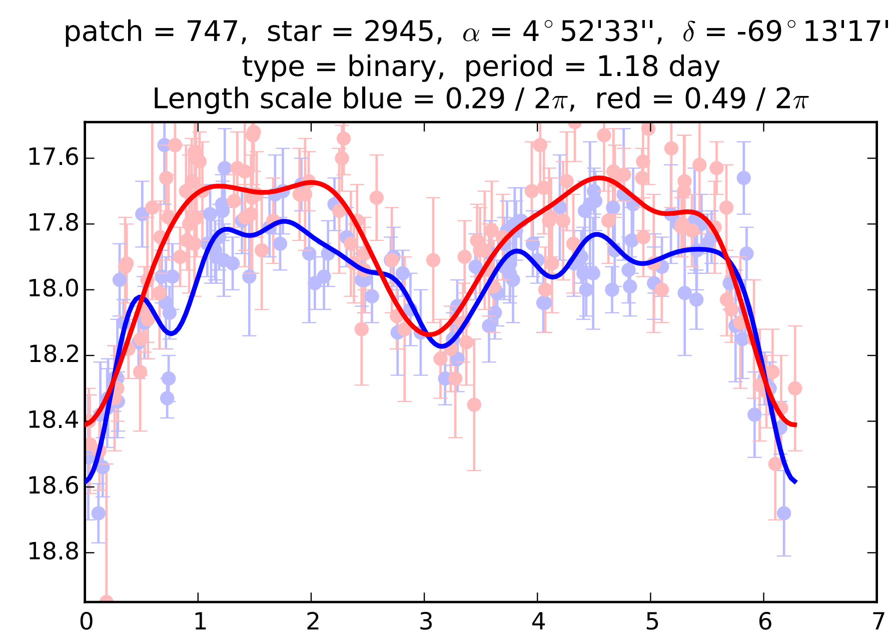
Inference

CLASSIFYING VARIABLE STARS

- collaboration with **astrophysicists at Paris-Saclay**
- variable-length **functional data**

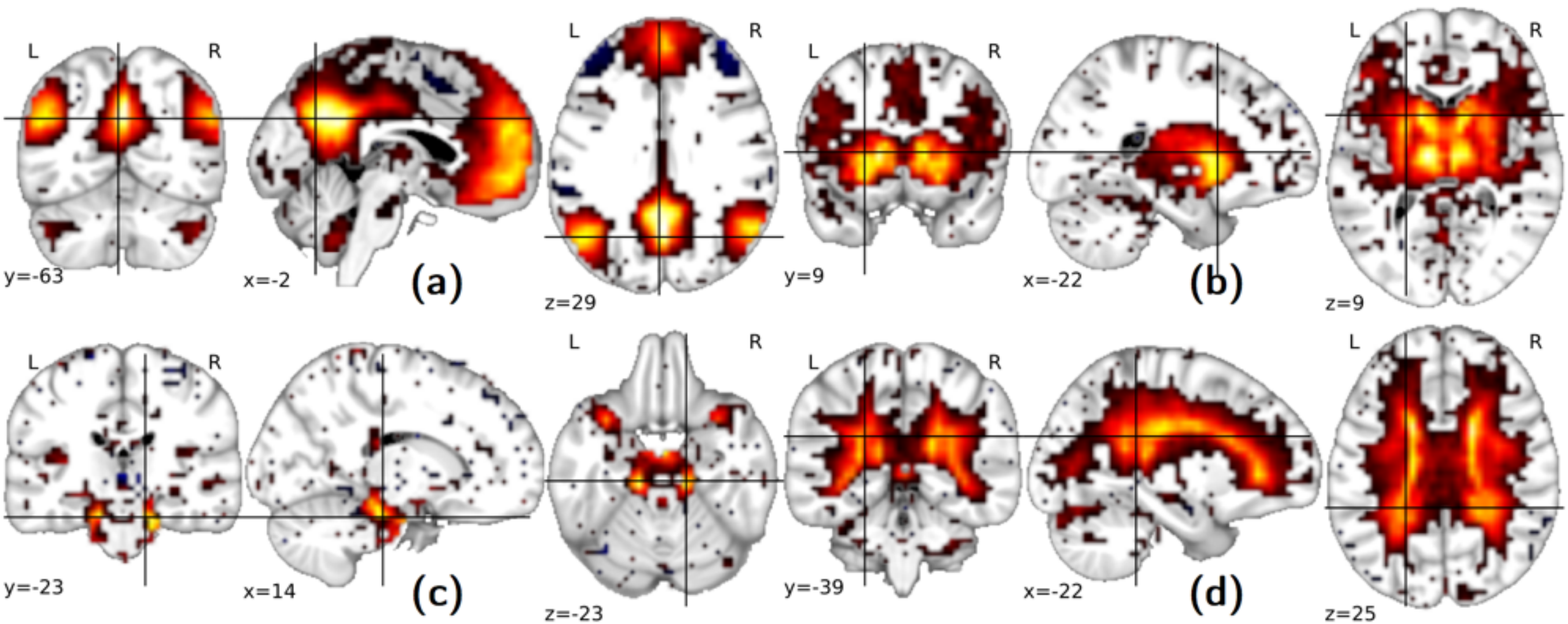






Predict autism from brain scans

- collaboration with **neurologists of Institut Pasteur**
- **3000** subjects: a **major major data collection effort**
- heavy **preprocessing** and **quality control**



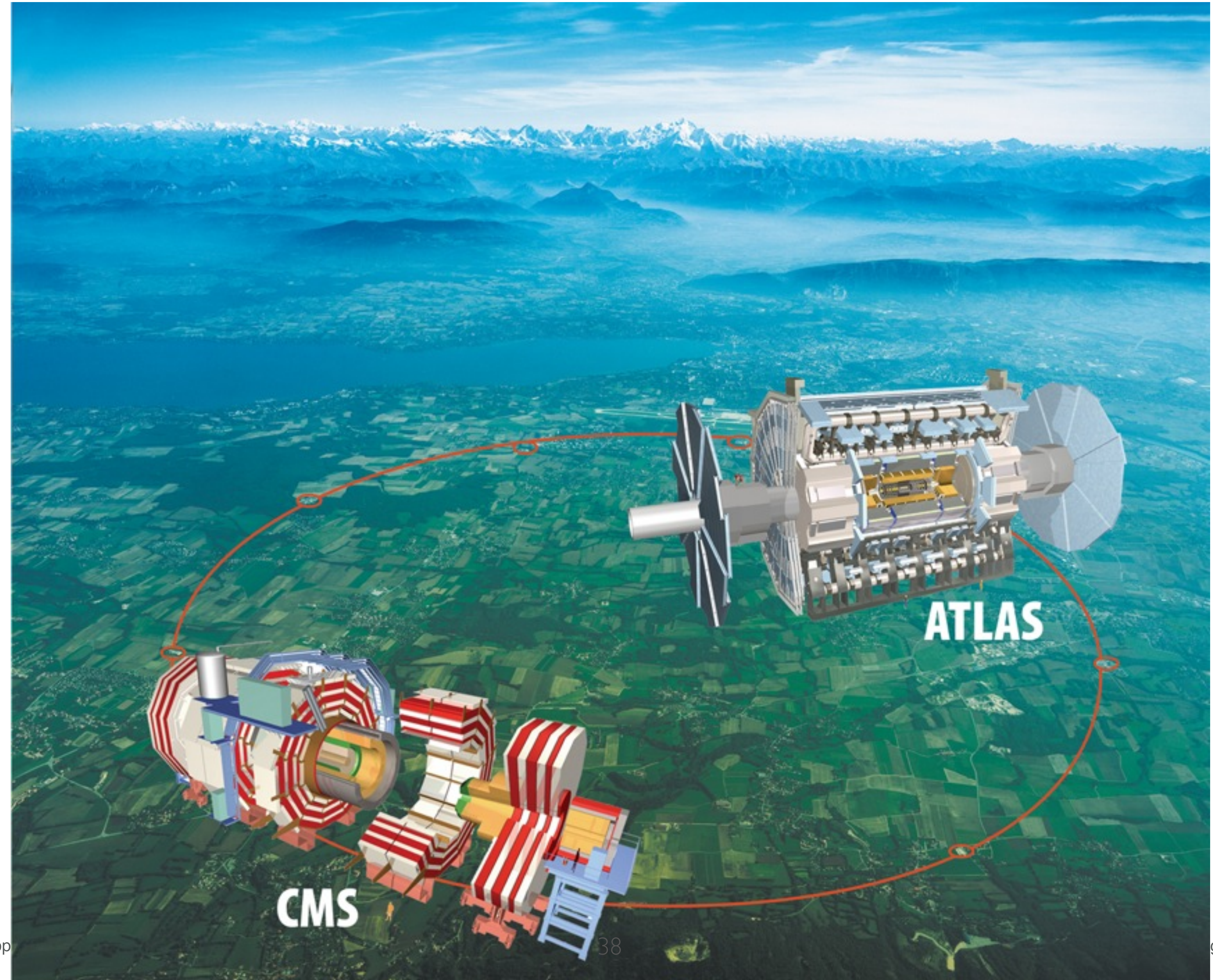
Learning to discover: the Higgs boson machine learning challenge

Claire Adam-Bourdarios[a], Glen Cowan[b], Cécile Germain[c], Isabelle Guyon[d], Balázs Kégl[a,c], David Rousseau[a]

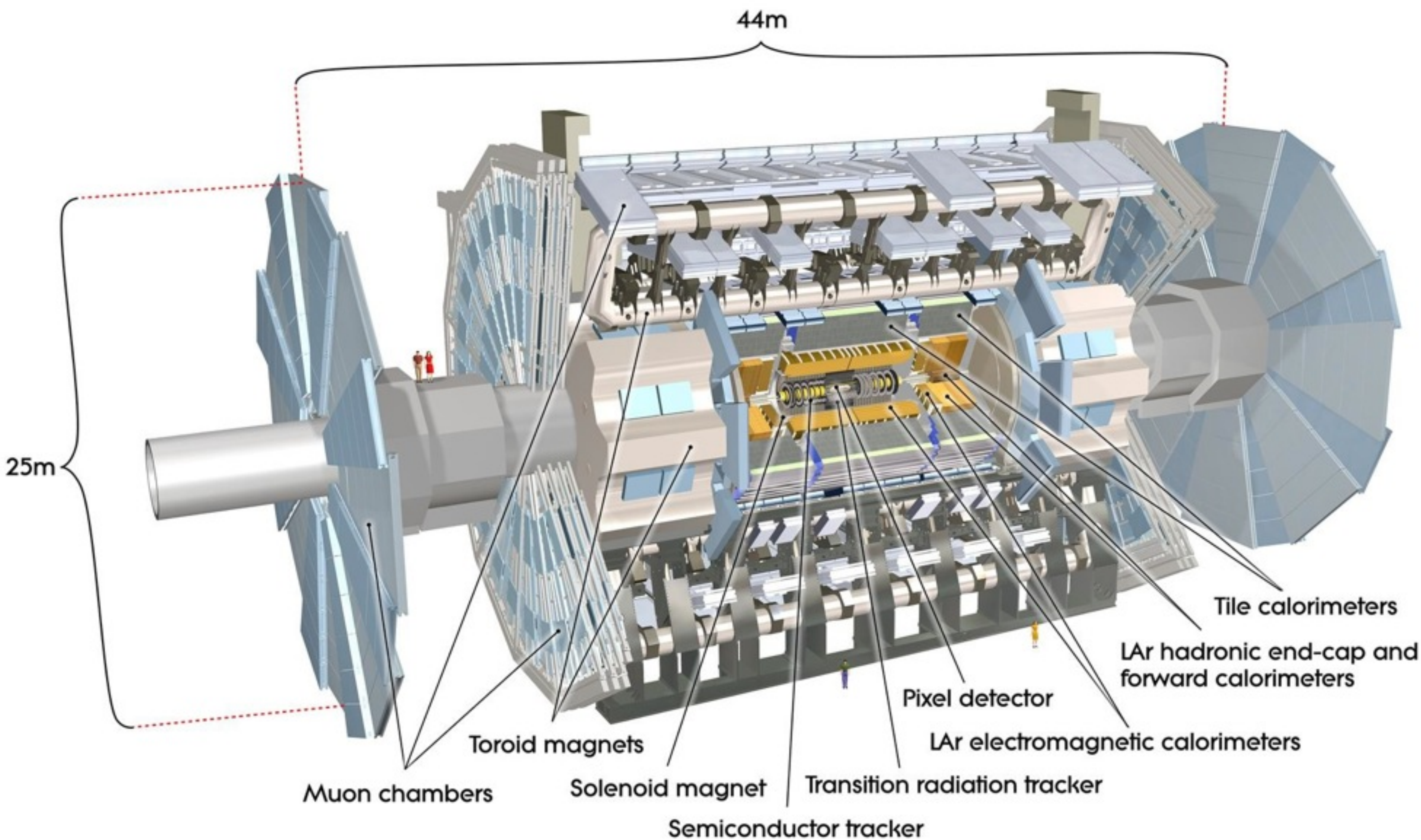
[a] LAL, IN2P3/CNRS & University Paris-Sud, France
[b] Physics Department, Royal Holloway, University of London, UK
[c] TAO team, INRIA & LRI, CNRS & University Paris-Sud, France
[d] ChaLearn

21 July 2014, version 1.8

THE LHC IN GENEVA

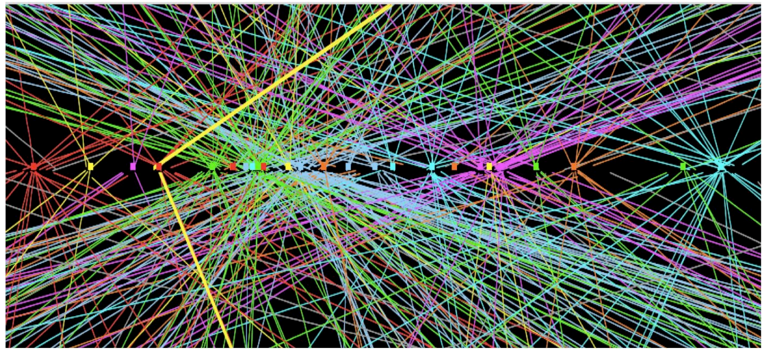


THE ATLAS DETECTOR

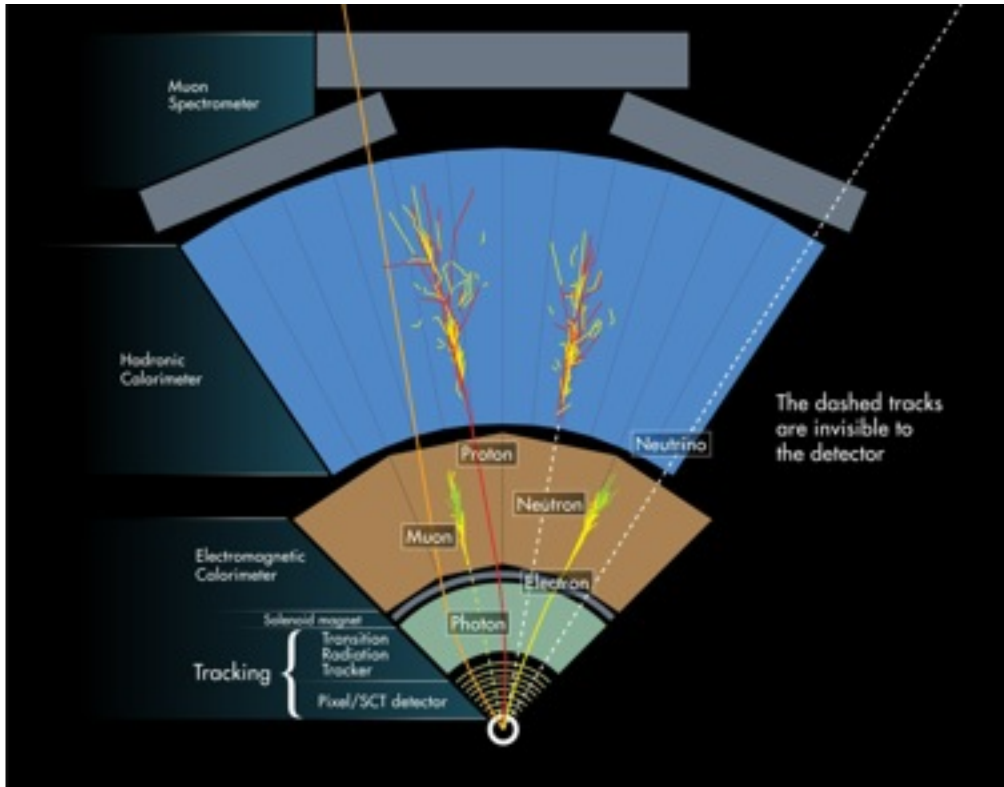


DATA COLLECTION

- **Hundreds of millions** of proton-proton collisions **per second**



- Filtered down to **400 events per second**
  - still **petabytes per year**
  - **real-time** (budgeted) classification: trigger
  - a research theme on its own

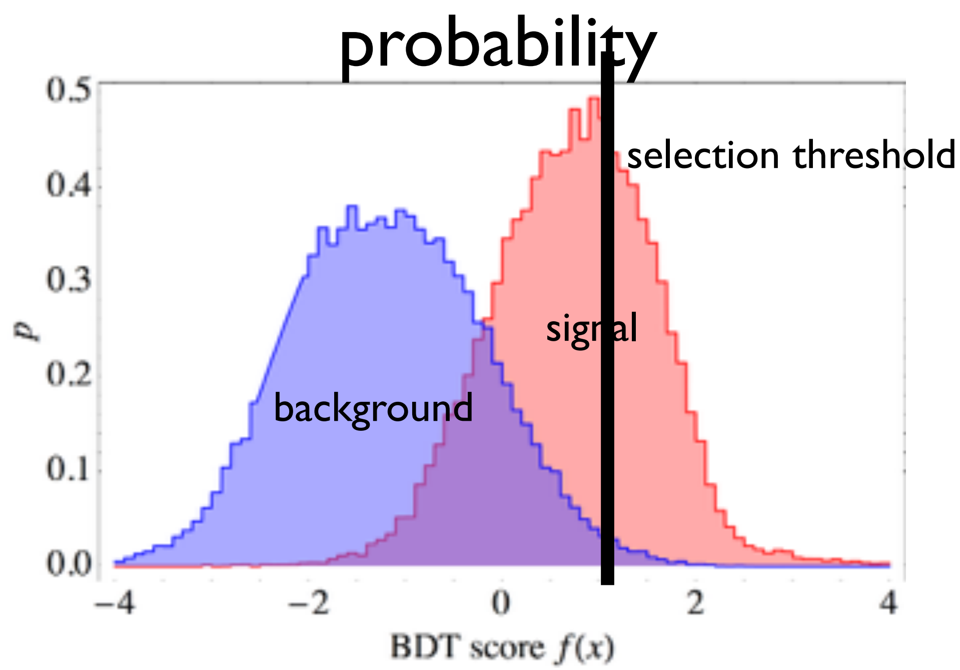


FEATURE ENGINEERING

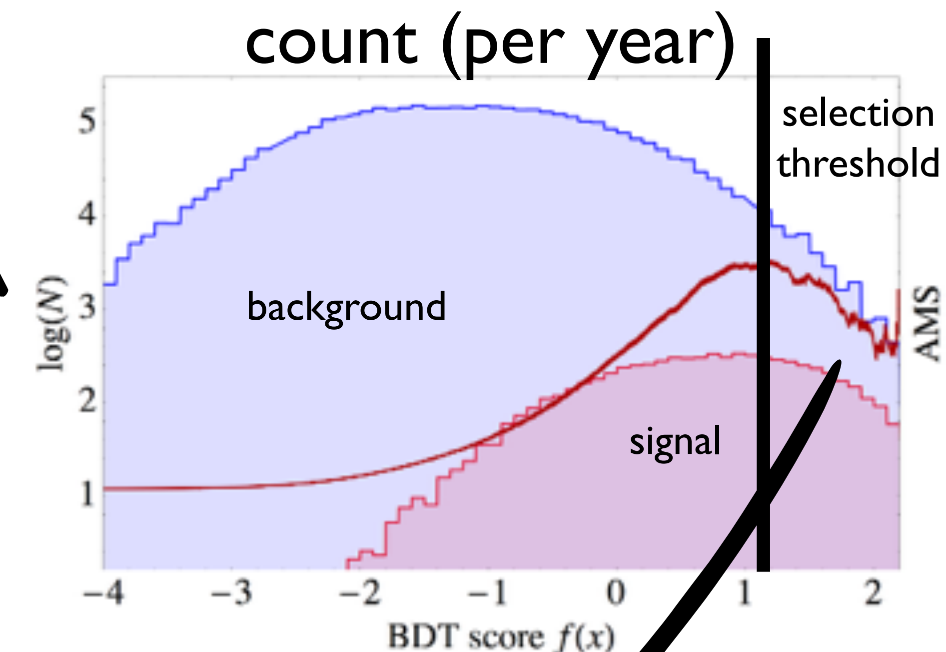
- Each collision is an **event**
  - **hundreds of particles**: decay products
  - **hundreds of thousands of sensors** (but sparse)
  - for each particle: **type**, **energy**, **direction** is measured
  - a fixed-length list of **~30-40 extracted features**: x
  - e.g., angles, energies, directions, reconstructed mass
  - based on **50 years** of accumulated **domain knowledge**

CLASSIFICATION FOR DISCOVERY

Goal: optimize the expected discovery significance



flux × time



total count,
say, 150 events

excess is s = 50 events

expected background
say b = 100 events

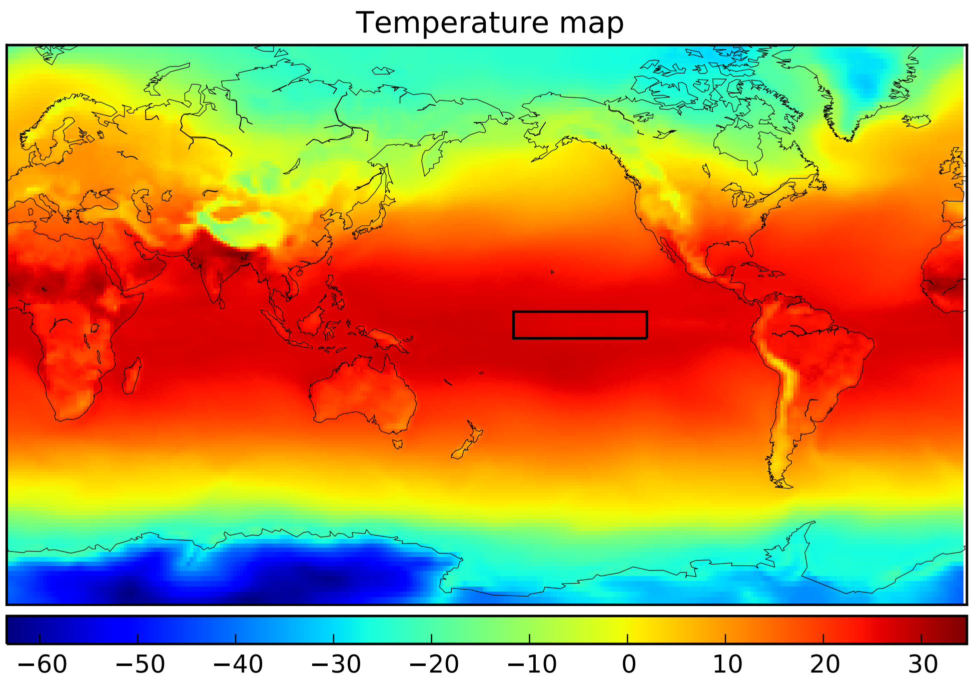
selection

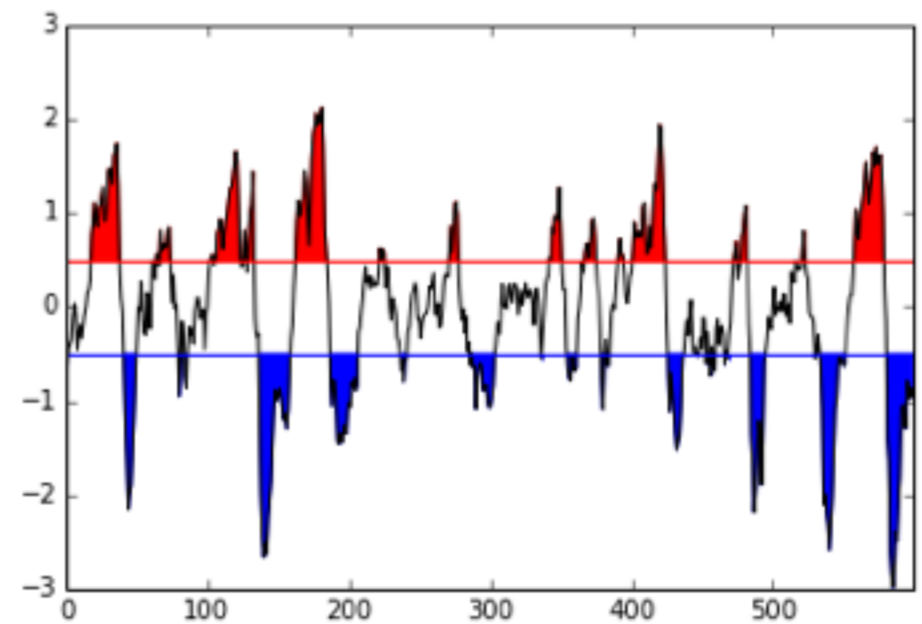
$$\text{AMS} = \frac{s}{\sqrt{b}} = 5 \text{ sigma}$$

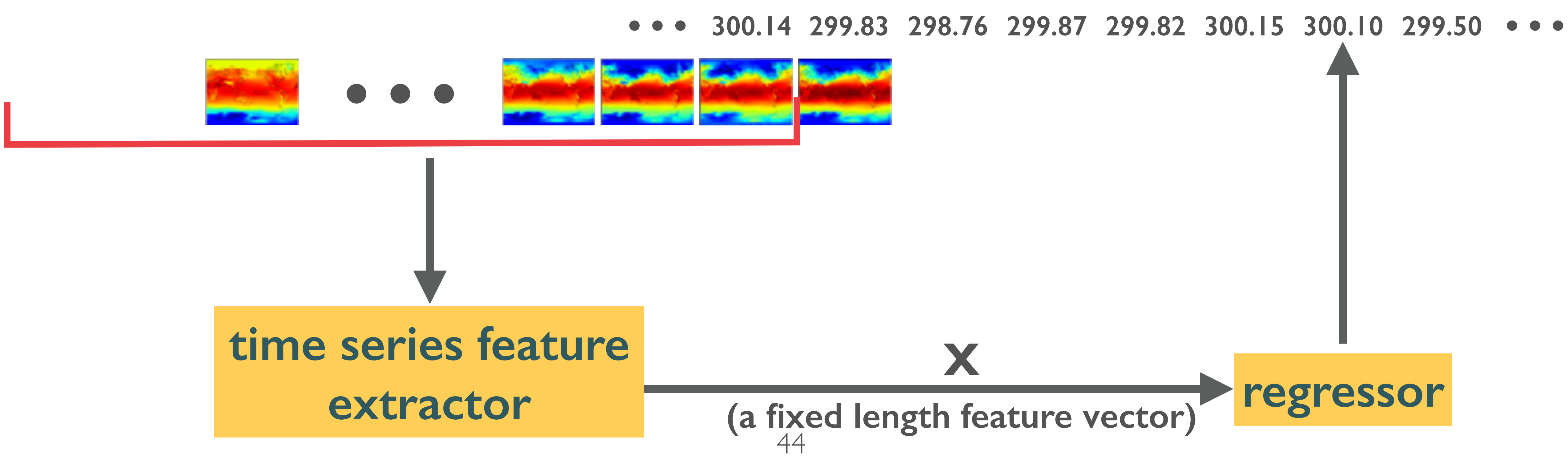
Generation and model reduction

FORECASTING EL NINO: SPATIOTEMPORAL TIME SERIES

- collaboration with the Climate Informatics workshop
- also on Arctic sea ice and California rainfall prediction







GENERATION AND MODEL REDUCTION

Why?

- Cost cutting 1: looking at the form of *f*, I can **place** my fixed number of **temperature sensors** **optimally**
- Cost cutting 2: *f* can replace costly simulation in a **detector optimization loop**
- Cost cutting 3: if I can generate realistic galaxy images, I can **replace costly manual labeling** of real photos

A multi-disciplinary initiative, building interfaces, matching people, helping them launching projects

345 affiliated researchers, 50 laboratories

Biology & bioinformatics
IBISC/UEvry
LRI/UPSud
Hepatinov
CESP/UPSud-UVSQ-Inserm
IGM-I2BC/UPSud
MIA/Agro
MIAj-MIG/INRA
LMAS/Centrale

Chemistry
EA4041/UPSud

Earth sciences
LATMOS/UVSQ
GEOPS/UPSud
IPSL/UVSQ
LSCE/UVSQ
LMD/Polytechnique

Economy
LM/ENSAE
RITM/UPSud
LFA/ENSAE

Neuroscience
UNICOG/Inserm
U1000/Inserm
NeuroSpin/CEA

Particle physics astrophysics & cosmology
LPP/Polytechnique
DMPH/ONERA
CosmoStat/CEA
IAS/UPSud
AIM/CEA
LAL/UPSud

Machine learning
LRI/UPSud
LTCI/Telecom
CMLA/Cachan
LS/ENSAE
LIX/Polytechnique
MIA/Agro
CMA/Polytechnique
LSS/Supélec
CVN/Centrale
LMAS/Centrale
DTIM/ONERA
IBISC/UEvry
LIST/CEA

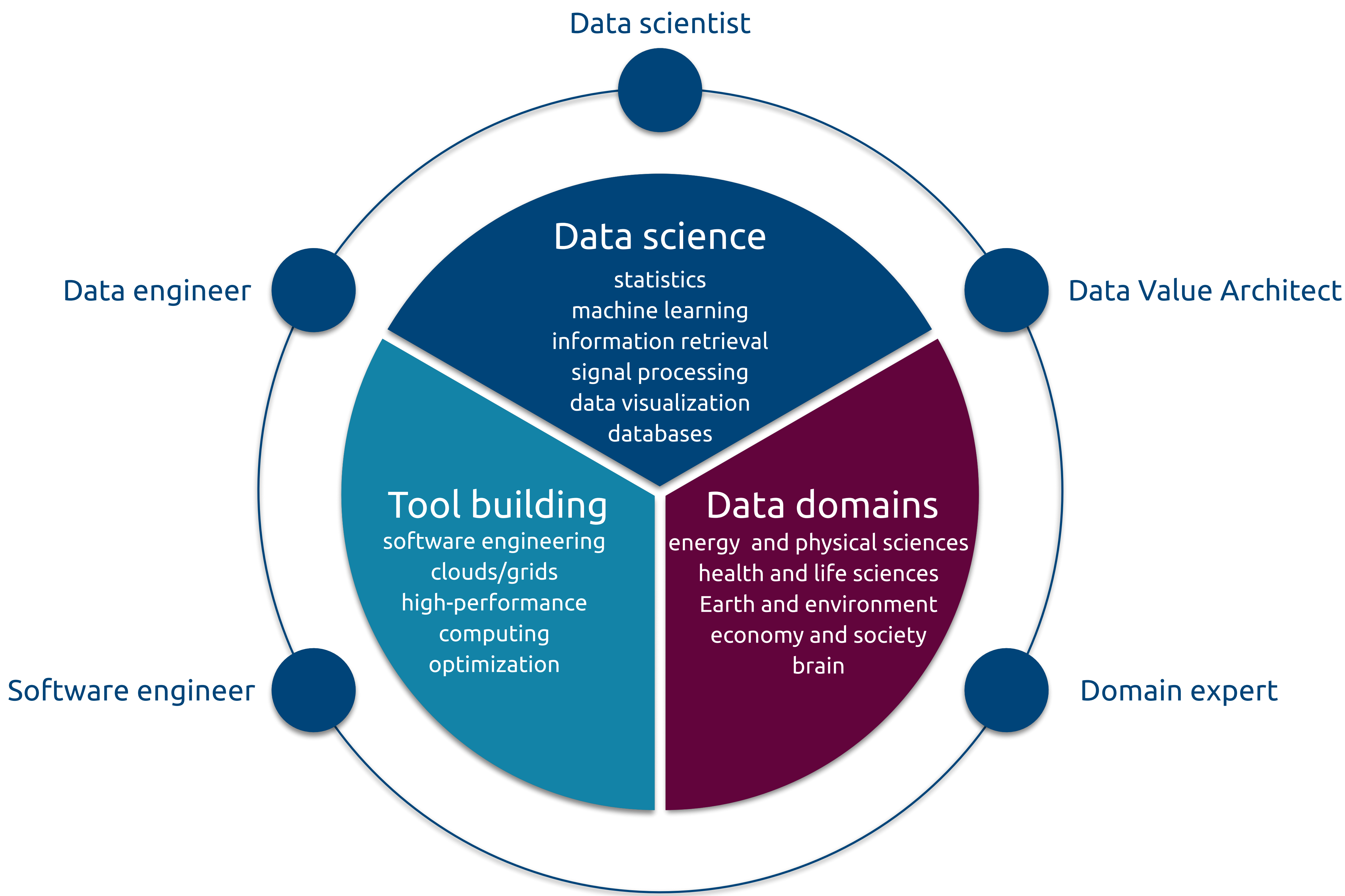
Visualization
INRIA
LIMSI

Signal processing
LTCI/Telecom
CMA/Polytechnique
CVN/Centrale
LSS/Supélec
CMLA/Cachan
LIMSI
DTIM/ONERA

Statistics
LMO/UPSud
LS/ENSAE
LSS/Supélec
CMA/Polytechnique
LMAS/Centrale
MIA/AgroParisTech

THE DATA SCIENCE ECOSYSTEM

https://medium.com/@balazskegl



TECHNICAL CHALLENGES

- Workflows and metrics
  - Designing the **workflow**, interaction with the **rest of the pipeline**, **metrics** is often **more important than “hyperopting” the predictor**
- Data generation
  - training is often done on **simulations**, so we need to **design data generation**
  - **systematic** uncertainties
  - the **iid oracle is a fairy tale**, happening only in machine learning textbooks
  - opportunity for **diversifying ML benchmarks**

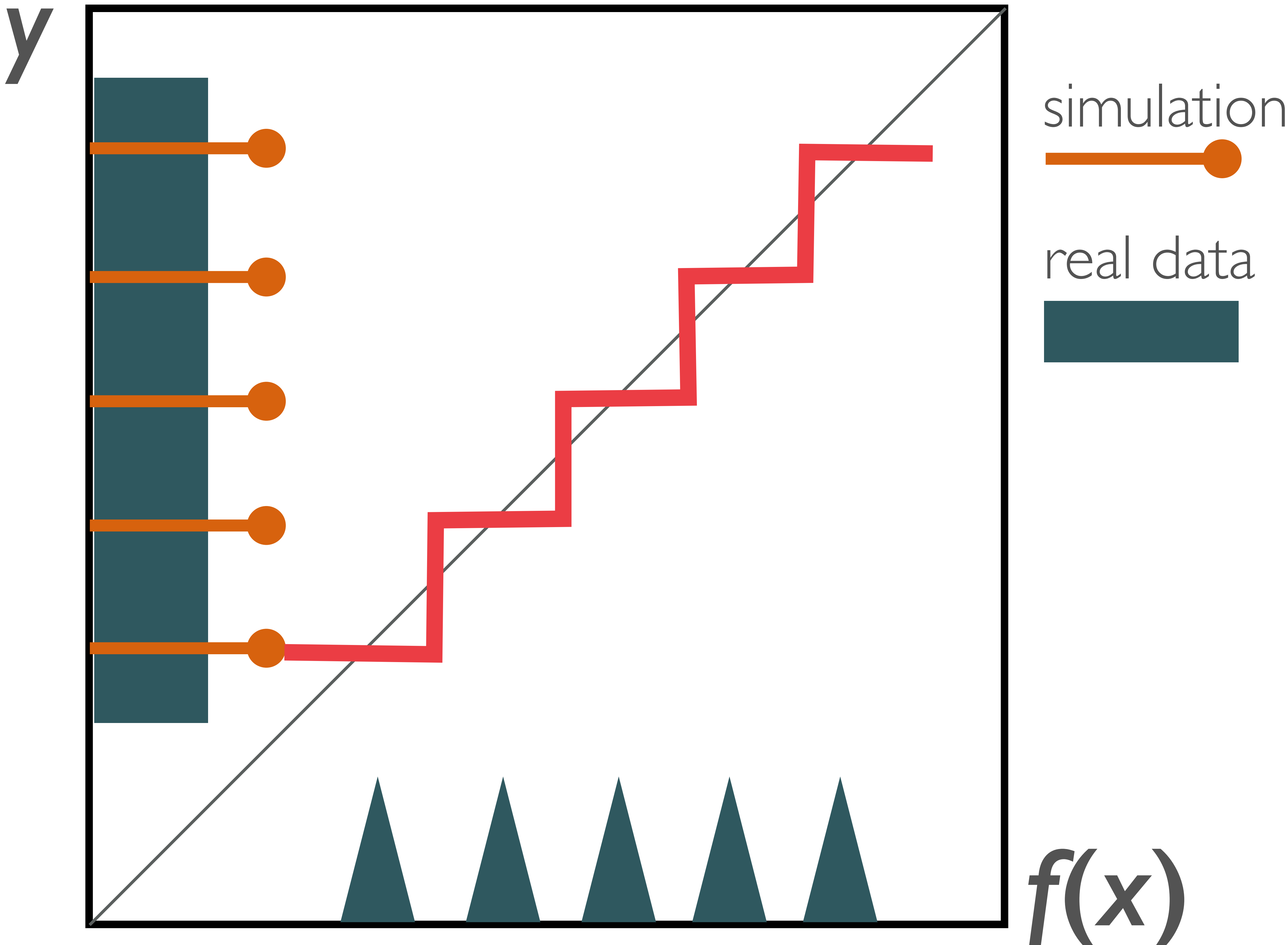
Systematics

- Your estimators f(x) should not only be efficient but also **insensitive to variables/parameters you don't know**
  - I know this problem because I worked with physicists
  - unsolved and even unknown in machine learning
  - google **"Fair ML"** to learn about the closest problem in ML
- Simulation-based training is **biased by design**
  - Because if we new all the distributions and parameters, we would not need to simulate

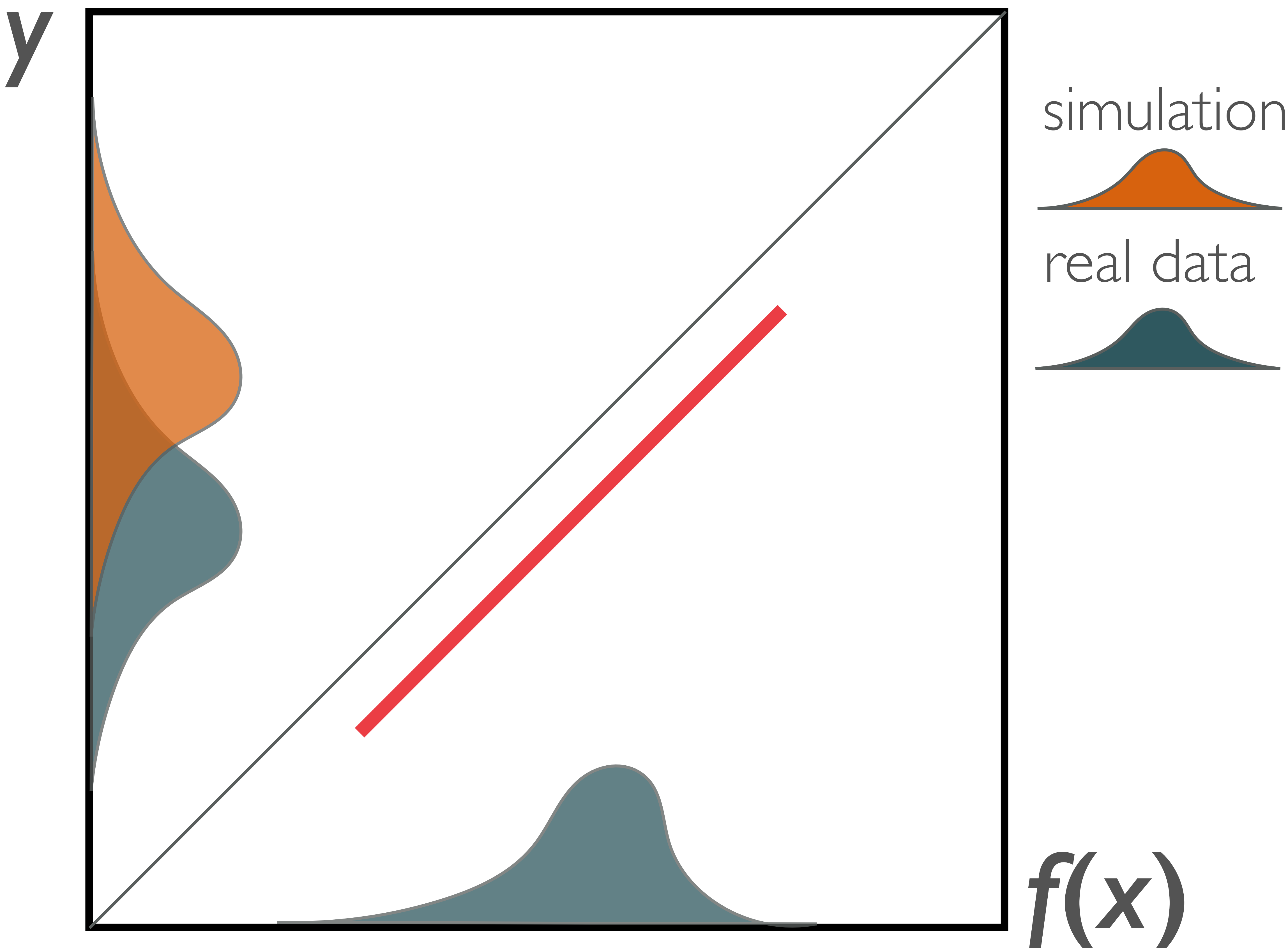
SYSTEMATICS

- ML is exacerbating the problem because it is so **efficient in optimizing the score**
  - unless the score contains systematics, which is hard because **systematics is usually not an event-wise metrics**
  - makes it similar to **adversarial generative models**, see the works of Kyle Cranmer and Gilles Louppe
- The classical approach: vary unknown parameters within their known range, train on one extreme, evaluate on the other
  - **exponential explosion** which makes computation-heavy deep learning even heavier
  - doesn't minimize systematics, but at least **measures it**
  - some are uncomfortable of **not "understanding" the black-box estimator**

SYSTEMATICS



SYSTEMATICS



Third take-home message

Systematics

MANAGEMENT AND ORGANIZATIONAL CHALLENGES

- Lack of manpower, misplaced incentives
  - hammers & nails
  - engineering: who deals with production?
- Lack of collaboration/innovation management tools
- Bottleneck is sometimes data collection/annotation
  - since domain scientists do not know ML, they do not collect the *right* data

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

We have "industrialized" workflow-building and optimization

By separating them

Then optimizing "graduate student descent"

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

RAMP

Hi Balázs!

Sandbox

You can either edit and save the code in the left column or upload the files in the right column. You can also import code from other submissions when the leaderboard links are open.

Edit and save your code!

classifier

```python
from sklearn.base import BaseEstimator
from sklearn.ensemble import RandomForestClassifier


class Classifier(BaseEstimator):
    def __init__(self):
        pass

    def fit(self, X, y):
        self.clf = RandomForestClassifier(
            n_estimators=2, max_leaf_nodes=3, random_state=61)
        self.clf.fit(X, y)

    def predict(self, X):
        return self.clf.predict(X)

    def predict_proba(self, X):
        return self.clf.predict_proba(X)
```

Upload your files!

File list

classifier.py

Upload file

Choose File No file chosen

Upload

RAMP

Leaderboard

Combined score: 0.899

Show 10 entries

team	submission	contributivity	historical contributivity	auc	accuracy	nll	train time	test time	submitted at (UTC)
diego.souza	tuning_xgboost3	9	5	0.896	0.820	0.385	3074	30	2017-01-17 11:34:53 Tue
ndeye-fatou.diop	kit_from_all	5	1	0.896	0.819	0.382	1167	10	2017-01-14 20:03:00 Sat
diego.souza	tuning_xgboost2	4	2	0.896	0.819	0.385	4900	17	2017-01-15 19:35:03 Sun
ndeye-fatou.diop	kit_from_all_clearer	3	0	0.896	0.819	0.384	1175	10	2017-01-15 03:45:44 Sun
etienne.boursier	combine_features	2	7	0.896	0.820	0.383	2712	3	2017-01-10 15:26:21 Tue
clement.vignac	boursier_improved_1	1	0	0.896	0.819	0.385	2499	4	2017-01-16 08:21:55 Mon

RAMP is a tool for

1. Collaborative prototyping
2. Teaching
3. Data science process management

Code submission

1. lets us deliver a working prototype
2. lets the participants collaborate

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

20+ challenges

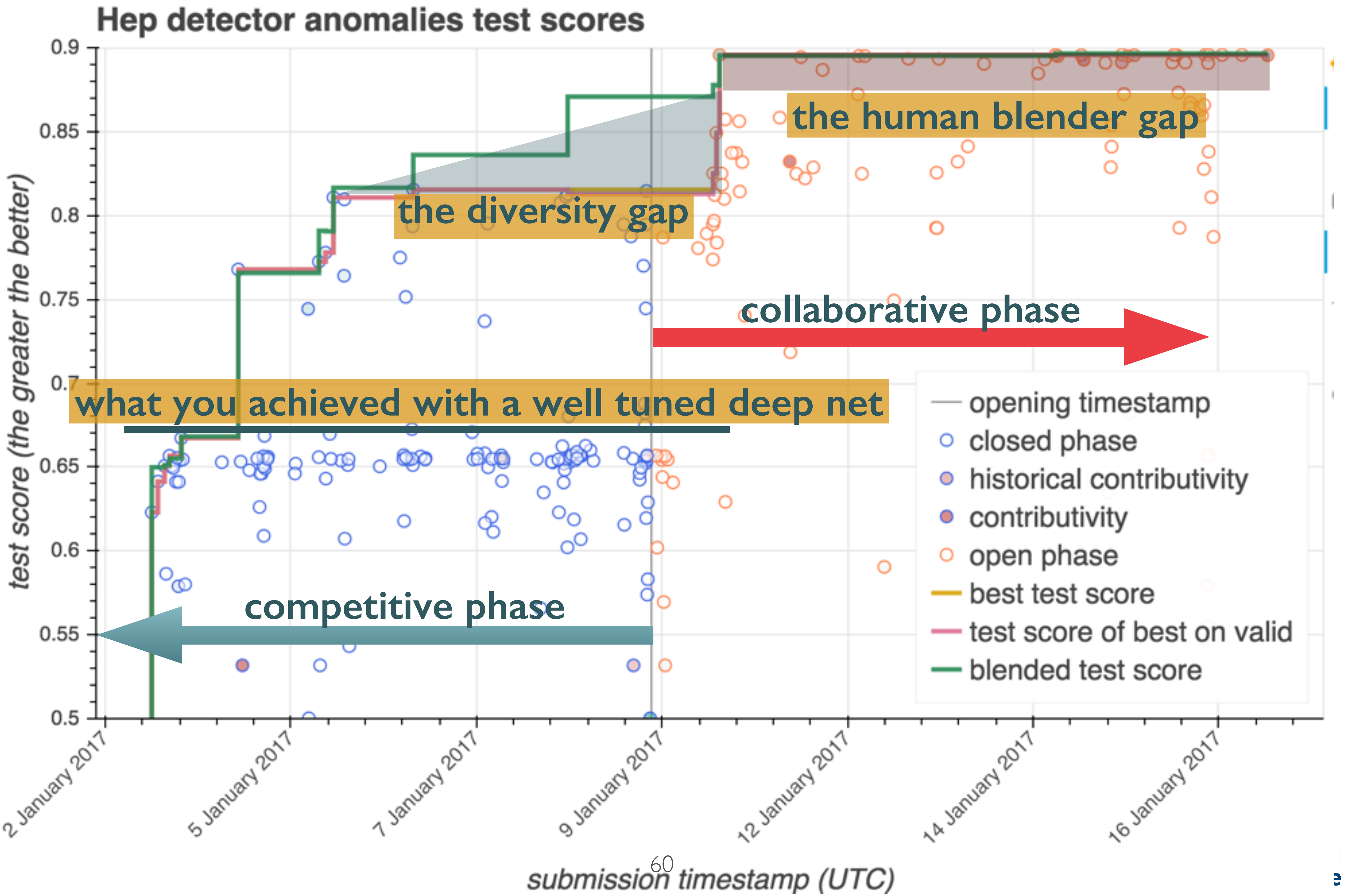
40+ events

1200+ users

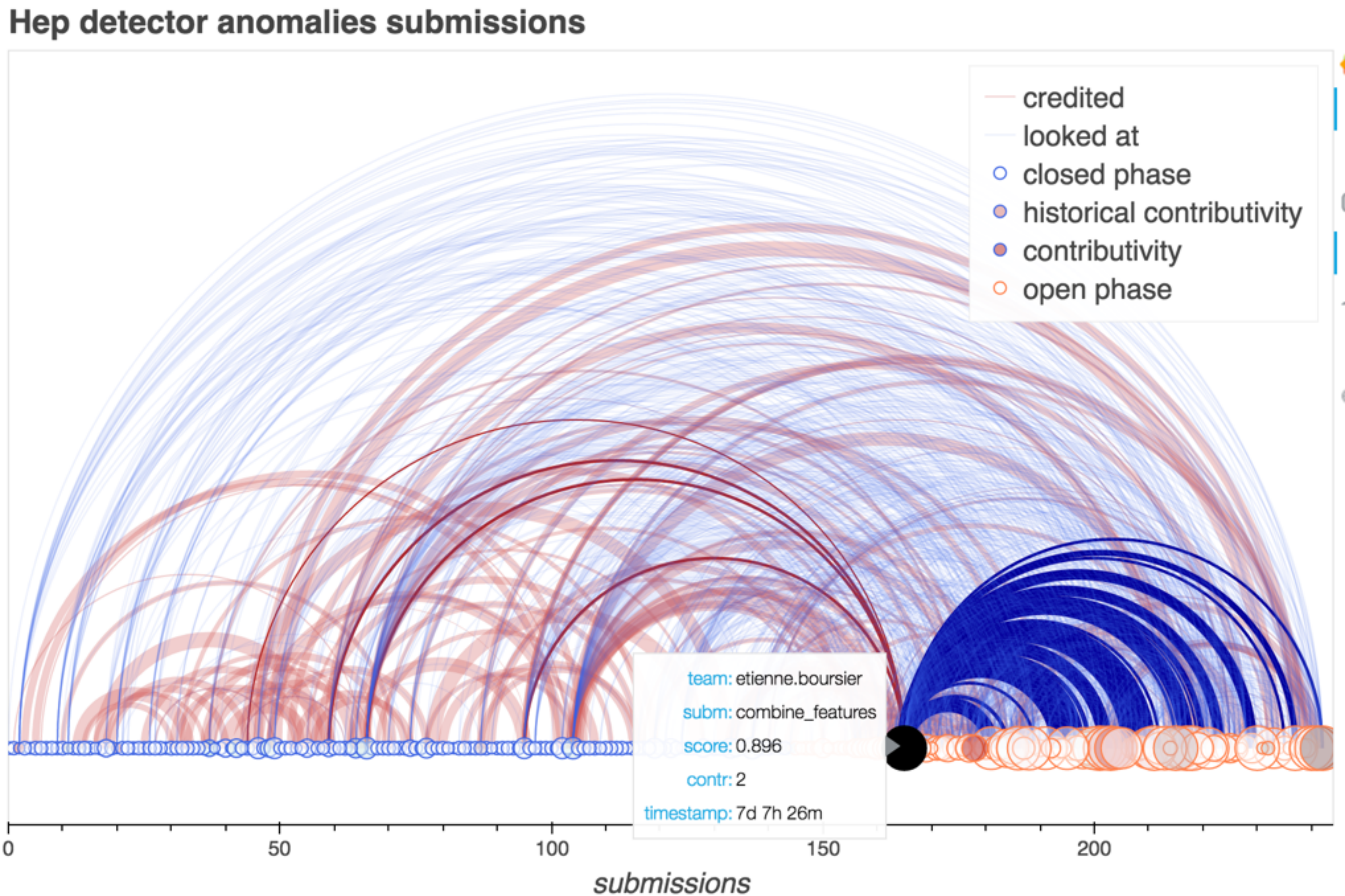
7000+ predictive models

https://www.ramp.studio/problems

The power of the (collaborating) crowd
optimizing graduate student descent



COMMUNICATION AND REUSE



frontend:

www.ramp.studio

toolkit:

github.com/paris-saclay-cds/ramp-workflow

examples:

github.com/ramp-kits

slack:

ramp-studio.slack.com

blogs:

medium.com/@balazskegl

mail:

balazs.kegl@gmail.com

Fourth take-home message

If you want ML experts to tackle your problem, make benchmarks, make it easy for them to contribute, without having to become a physicist

Machine learning in science

Inference
Generation/simulation and model reduction

- We can **automate almost everything**
  - simulation, inference, experimental design
  - this is not even controversial, just an extension of the current paradigm
- But not the **hypothesis generation**: what model to test?

Hypothesis generation is crucial and, at the same time, not covered by the scientific method

ROBOT SCIENTIST

Artificially-intelligent Robot Scientist 'Eve' could boost search for new drugs

Published

04 Feb 2015

Image

Eve, the Robot Scientist

Credit: University of Manchester

ROBOT SCIENTIST

"Robot scientists are a natural extension of the trend of increased involvement of automation in science. They can **automatically develop and test hypotheses** to explain observations, **run experiments** using laboratory robotics, **interpret the results** to amend their hypotheses, and then **repeat the cycle**, automating high-throughput **hypothesis-led research**."

http://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs

Hypothesis generation is crucial
and, at the same time,
not covered by the scientific method

This ignorance has already bitten us,
but with the appearance of the robot
scientist, it is unavoidable

THE SCIENTIFIC METHOD IN THE TRENCHES

- Come up with a **hypothesis**
- Design an **experiment to exclude it**
- Use a **statistical test** to show that the data is unlikely to be generated by a world in which the hypothesis does not hold (“background”)

The scientific method in the trenches

- Rutherford: "*If your experiment needs statistics, you ought to have done a better experiment*"
- **Without statistics, science would be over**
  - we went out of slam dunk infinite significance ("background free") hypotheses
  - phenomena are inherently noisy: **nobody has seen or will ever see a Higgs boson**

rsos.royalsocietypublishing.org/content/1/3/140216

THE ROYAL SOCIETY PUBLISHING

ROYAL SOCIETY OPEN SCIENCE

search

Advanced

Home | Content | Information for | About us | Sign up | Submit

Check for updates

An investigation of the false discovery rate and the misinterpretation of *p*-values

David Colquhoun

Published 19 November 2014. DOI: 10.1098/rsos.140216

Article | Figures & Data | Info & Metrics | Review History | PDF

Previous | Next

Abstract

If you use p=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time. If, as is often the case, experiments are underpowered, you will be wrong most of the time. This conclusion is demonstrated from several points of view. First, tree diagrams which show

November 2014

Table of Contents

About the Cover

Index by author

THE P-VALUE CONTROVERSY

But the main problem is a tautology:
if none of your hypotheses are true,
all your positives are false

But of course: if all your hypotheses are true,
you are not exploring

Guidelines

- Register all experiments and publish negatives
- Don't do underpowered experiments
- Put the significance bar high enough
- **Test only "plausible" hypotheses**

QUESTIONS

- What is a **plausible but non-trivial** hypothesis?
- How to **measure** plausibility?
- How to **generate** them (automatically)?
- How are **hypotheses related to prior/current knowledge**?

Generative models in ML

Interesting tools but it's a whole new ballgame and paradigmatically we are in the dark

Train on digits,
test on letters

Train on all music up to the Beatles, test on Sex Pistols

Train on all phones up to 2006,
test on the iPhone

Train on all scientific knowledge up to Einstein, test on relativity theory

CAN WE GENERATE NEW TYPES?

Existing objects of known types.

learning

generative model

generation

New objects. New types?

Some written stuff

https://medium.com/@balazskegl/the-epistemological-challenges-of-automating-a-b-testing-or-how-will-ai-do-science-b724f821781l#.q041gyvkt

https://arxiv.org/abs/1606.04345

http://openreview.net/forum?id=ByEPMj5el

FIFTH TAKE-HOME MESSAGE

Scientific knowledge representation and hypothesis generation is where real AI will go

Thank you!