

MACHINE LEARNING FOR SCIENTIFIC WORKFLOWS

MANAGING THE DATA SCIENCE PROCESS

BALÁZS KÉGL

Université Paris-Saclay / CNRS

WHO AM I?

Balázs Kégl

- Senior researcher **CNRS**
 - machine learning (20+ years)
interfacing with particle physics (10+ years)
- Head of the **Paris-Saclay Center for Data Science**
 - interfacing with biology, economy, climatology, chemistry, etc. (4 years)
 - industrial ML projects (4 years)

OUTLINE

- Machine learning for **scientific workflows**
 - challenges
 - **use cases**: data collection, inference, simulation
 - **examples**: pollinating insects, autism, variable stars, Mars craters, drug spectra, the Higgs boson, El Nino
- Managing the data science process: the **RAMP framework**
 - **roles** and **tasks** in the **data science process**
 - **building** the workflow: **who does what**
 - what is a **predictive workflow**, what are the **parametrizable components**
 - how to make **data scientists efficient**

WHY IS THIS RELEVANT FOR YOU?

- Typical **ML** research project
 - take an **existing ML problem** (e.g., image classification)
 - scan **literature**
 - install/develop **experimental environment**
 - **explore ideas** to find new solutions
 - **optimize**, show they work **better than existing solutions**, ideally on established and accepted **benchmarks**
 - publish

WHY IS THIS RELEVANT FOR YOU?

- Typical **applied** research project
 - take an **existing domain-scientific or industrial problem** (e.g., galaxy deblending)
 - scan **literature**
 - install/develop **experimental environment**
 - collect **data** and **establish benchmark**
 - apply **existing ML solution**
 - optionally fine tune, explore a **small number** of alternatives
 - show that the **ML solution is better than the classical “manual” solution on your own benchmark**
 - publish

Both take **years**, typically

**How to make applied projects
faster?**

**How to explore a
large number of ML solutions
in a short time?**

LONGER TERM VISION

- Put more focus on **problem formulation**
 - **detect** important applied problems
 - agree on **metrics** and **benchmarks** and **organize data collection**
 - organizationally **separate setting up benchmarks and optimizing solutions**
 - establish a **fair and possibly “third party” framework** (see ImageNet)
 - hammer the message that **formulating scientific problems into predictive workflows is valuable research**

WHY IS THIS RELEVANT FOR YOU?

- You may use it to **accelerate your own research**
- You may remember this when you **join the data science industry**

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

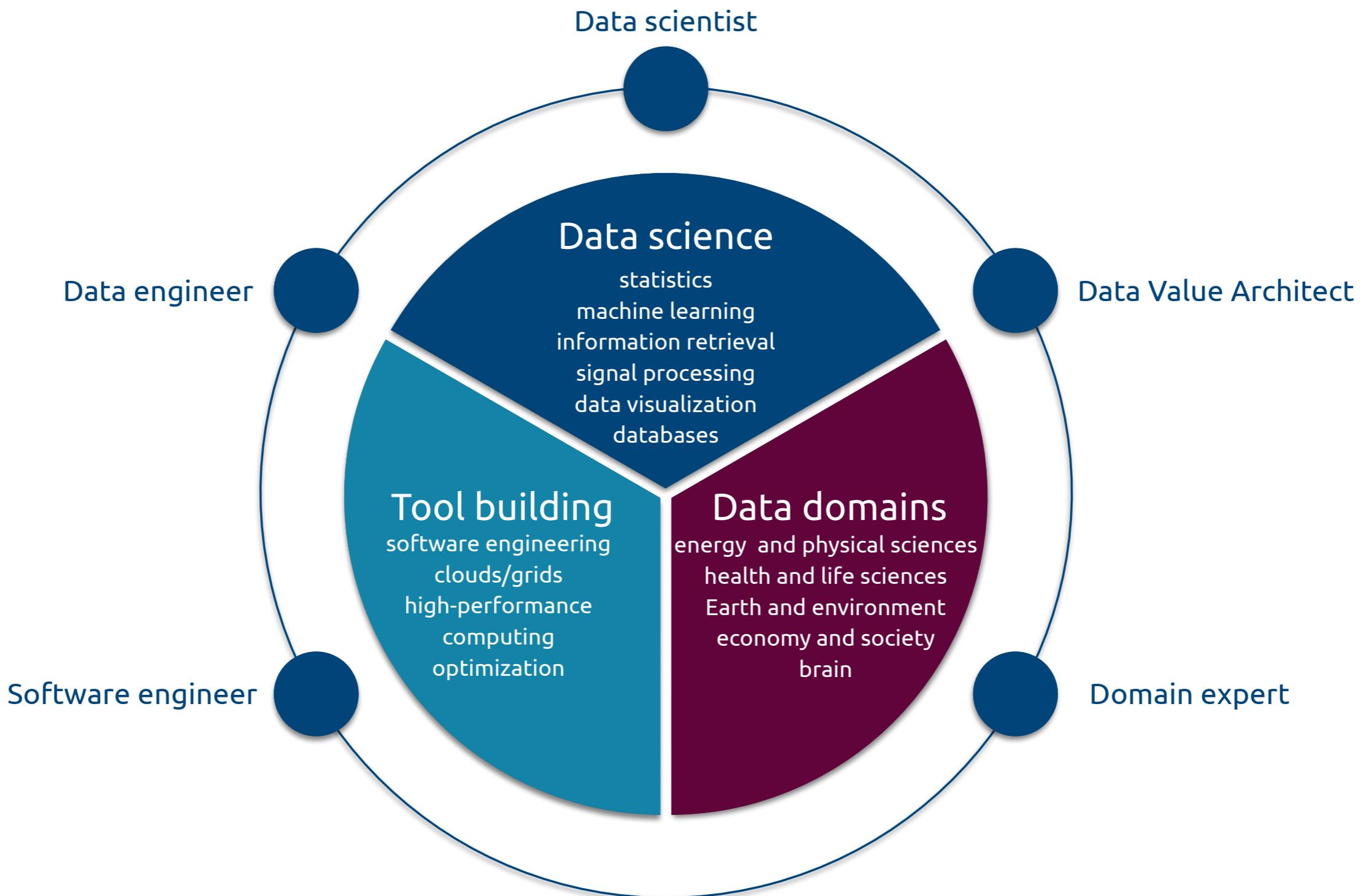
LTCI/Telecom
CMA/Polytechnique
CVN/Centrale
LSS/Supélec
CMLA/Cachan
LIMSI
DTIM/ONERA

Statistics

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

THE DATA SCIENCE ECOSYSTEM

<https://medium.com/@balazskegl>



MANAGEMENT AND ORGANIZATIONAL

- 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

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**

ML USE CASES IN SCIENCES

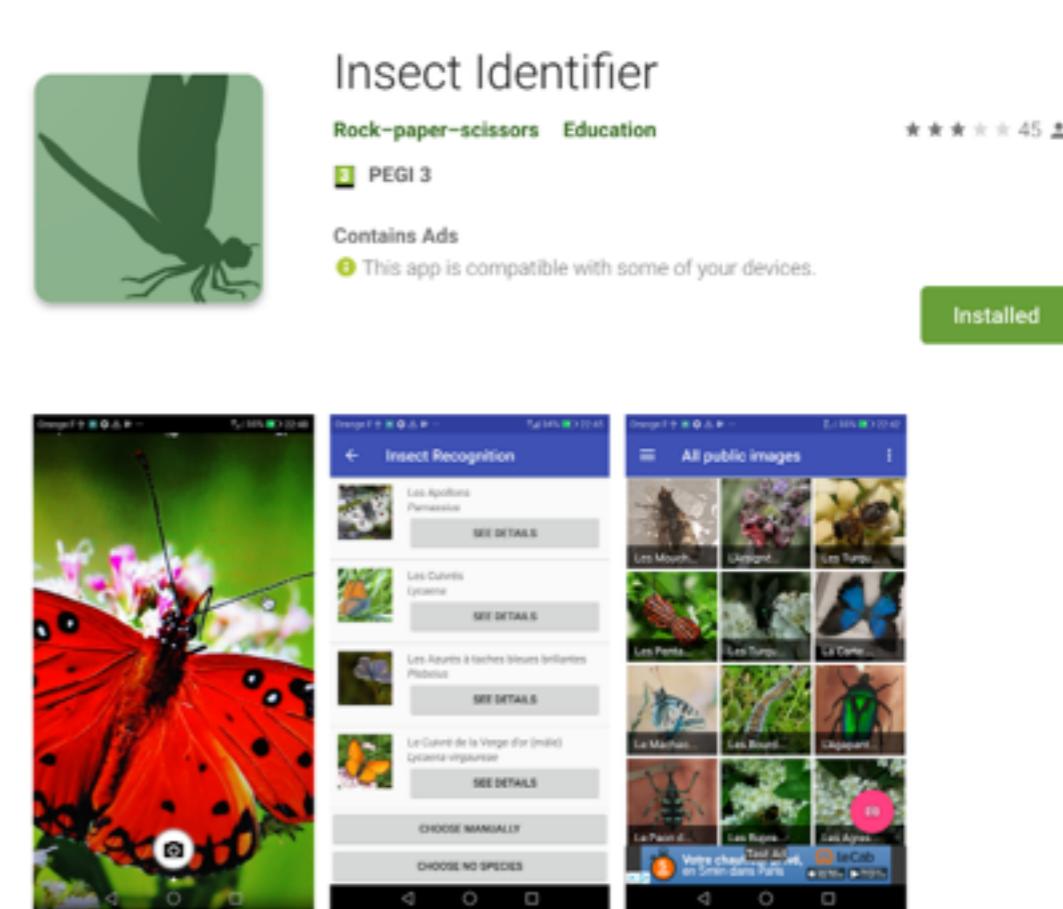
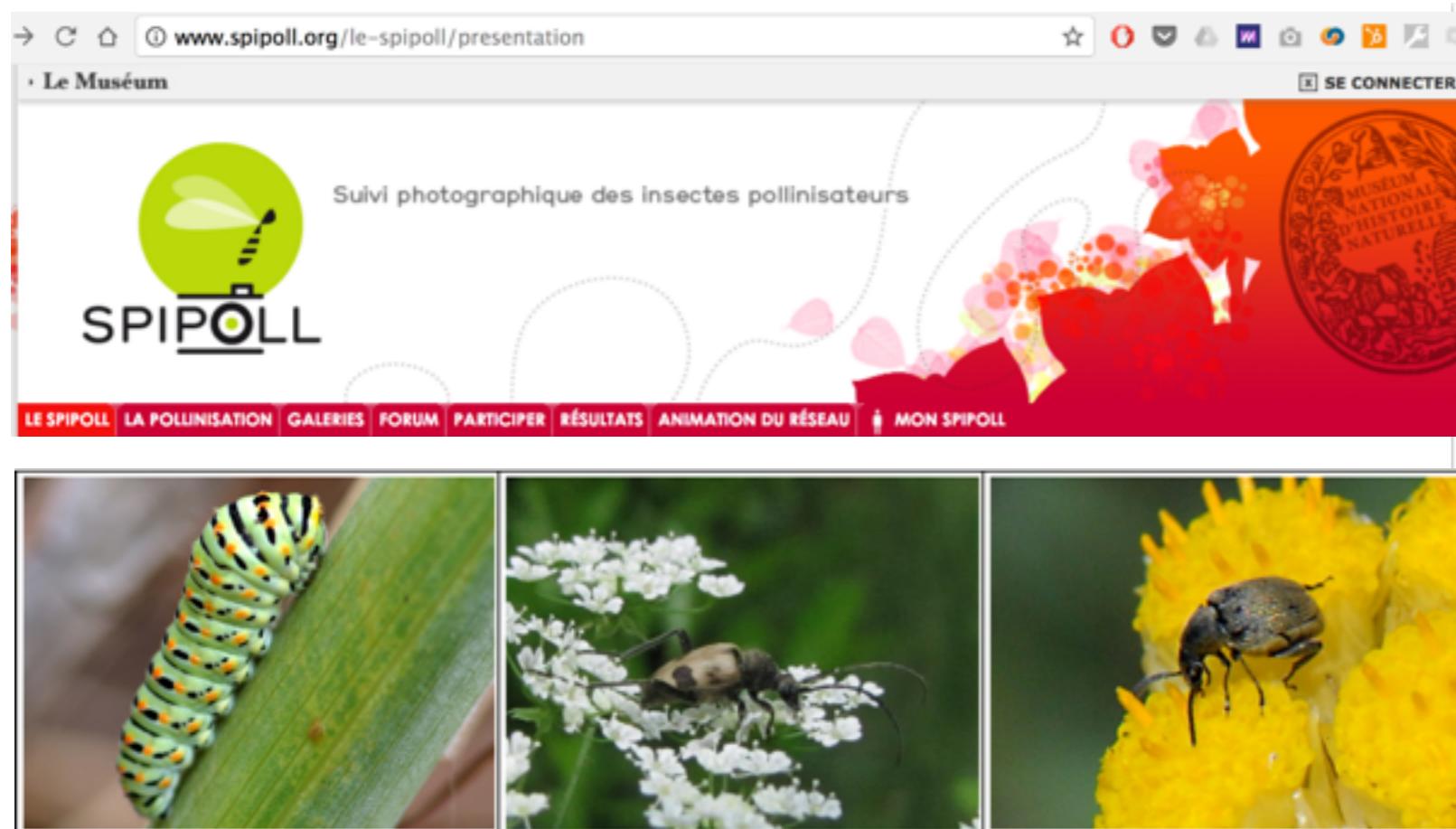
<https://www.ramp.studio/problems>

- **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

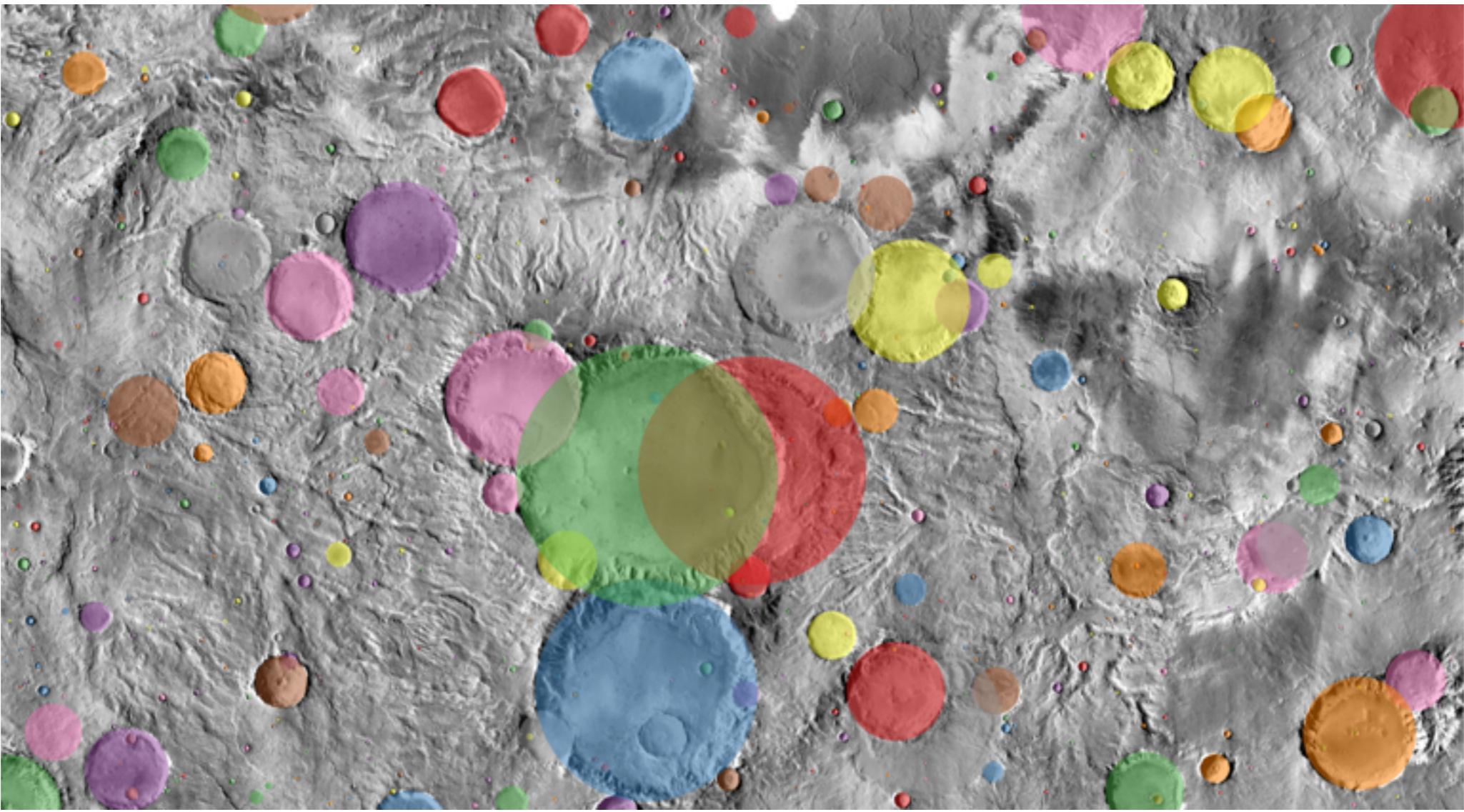
CLASSIFYING POLLENATING INSECT PHOTOS

- collaboration with **ecologists at the Paris Museum of Natural History**
- **400 classes, 150K photos, long tail**
- great benchmark for **transfer** and **few-shot** learning
- developed models in **production**, powering an **android app**



DETECTING MARS CRATERS

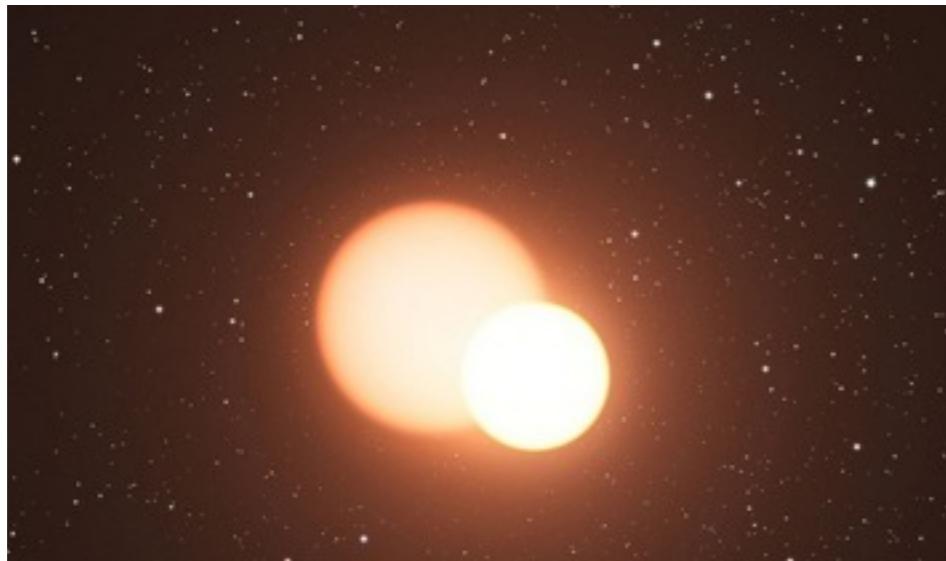
- collaboration with **planetary geologists at Paris-Saclay**
- new **metrics** and **workflow**
- great benchmark for **detection in satellite imagery**



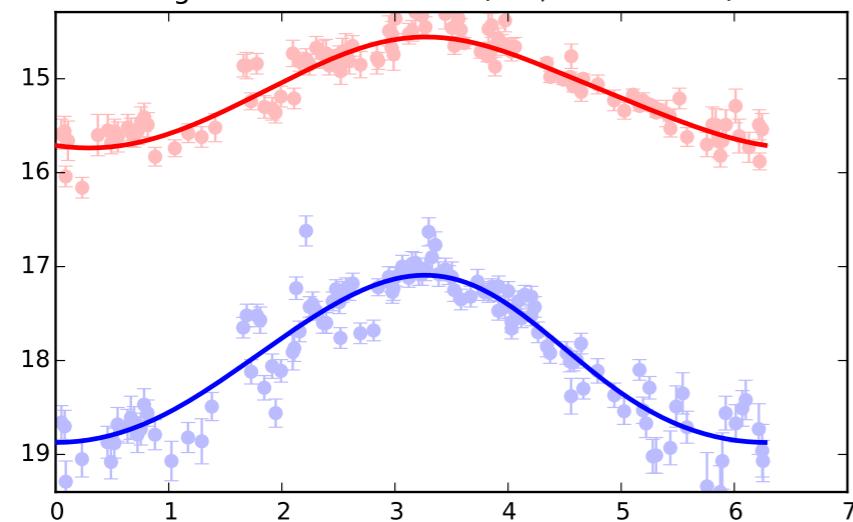
Inference

CLASSIFYING VARIABLE STARS

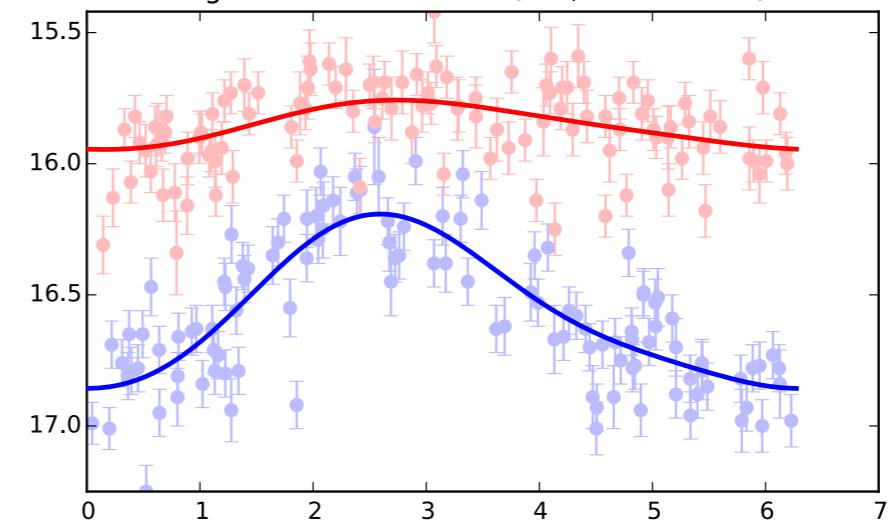
- collaboration with **astrophysicists at Paris-Saclay**
- variable-length **functional data**



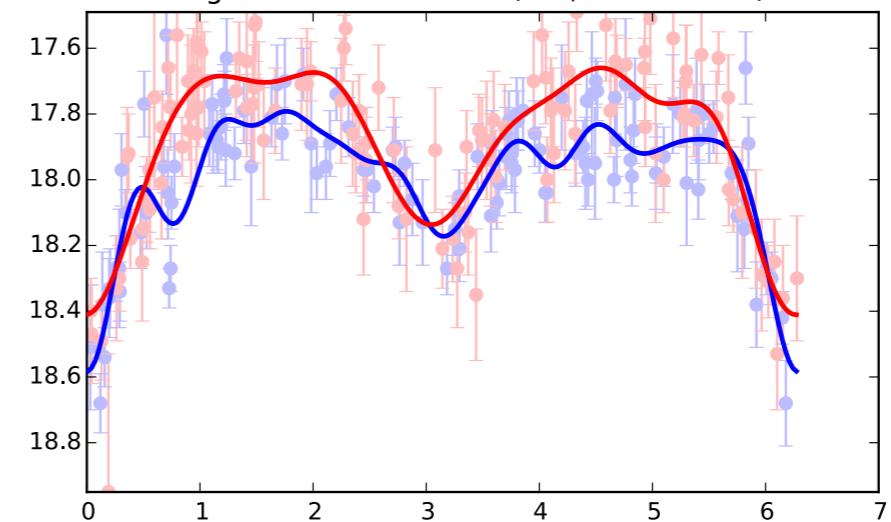
patch = 327, star = 1726, $\alpha = 5^\circ 25'27''$, $\delta = -69^\circ 23'43''$
type = mira, period = 214.28 day
Length scale blue = $2.48 / 2\pi$, red = $2.09 / 2\pi$



patch = 717, star = 2162, $\alpha = 4^\circ 55'31''$, $\delta = -68^\circ 53'0''$
type = cepheid, period = 2.77 day
Length scale blue = $2.14 / 2\pi$, red = $2.96 / 2\pi$



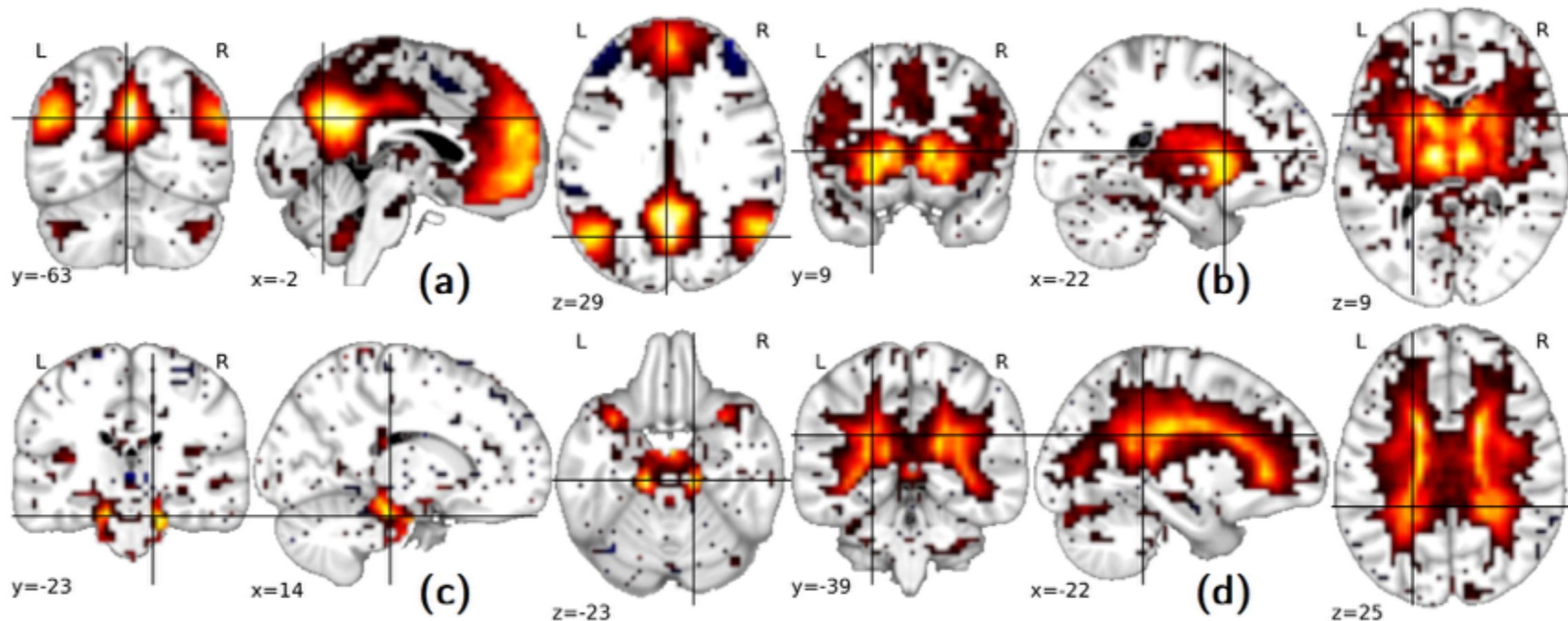
patch = 747, star = 2945, $\alpha = 4^\circ 52'33''$, $\delta = -69^\circ 13'17''$
type = binary, period = 1.18 day
Length scale blue = $0.29 / 2\pi$, red = $0.49 / 2\pi$



PREDICT AUTISM FROM BRAIN SCANS

https://paris-saclay-cds.github.io/autism_challenge

- collaboration with **neurologists of Institut Pasteur**
- **3000** subjects: a **major major data collection effort**
- heavy **preprocessing** and **quality control**
- **ongoing** challenge till **July 1** with **9.5K€** money prizes



CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

- collaboration with the **pharmacy department** of Georges Pompidou Hospital
- **major data collection effort**
- functional data
- probably the first ever research paper where the **ML workflow optimization was entirely crowdsourced**

Classifying and quantifying monoclonal antibody preparations for cancer therapy using machine learning

Laetitia Le ^{ab}, Camille Marini ^{ce}, Alexandre Gramfort ^{cfg},
David Nguyen ^a, Mehdi Cherti ^{ch}, Sana Tfaili ^b, Ali
Tfayli ^b, Arlette Baillet-Guffroy ^b, Eric Caudron ^{ab}, Balázs
Kégl ^{ch}

^a European Georges Pompidou Hospital (AP-HP), Pharmacy department, Paris, France

^b Lip(Sys) Chimi Analytique Pharmaceutique, Univ. Paris-Sud, Universit Paris Saclay, F92290 Chatenay-Malabry, France (EA4041 Groupe de Chimie Analytique de Paris Sud)

^c Center of Data Science, Université Paris-Saclay

^d Université Paris-Sud

^e CMAP, Ecole Polytechnique, Palaiseau, France

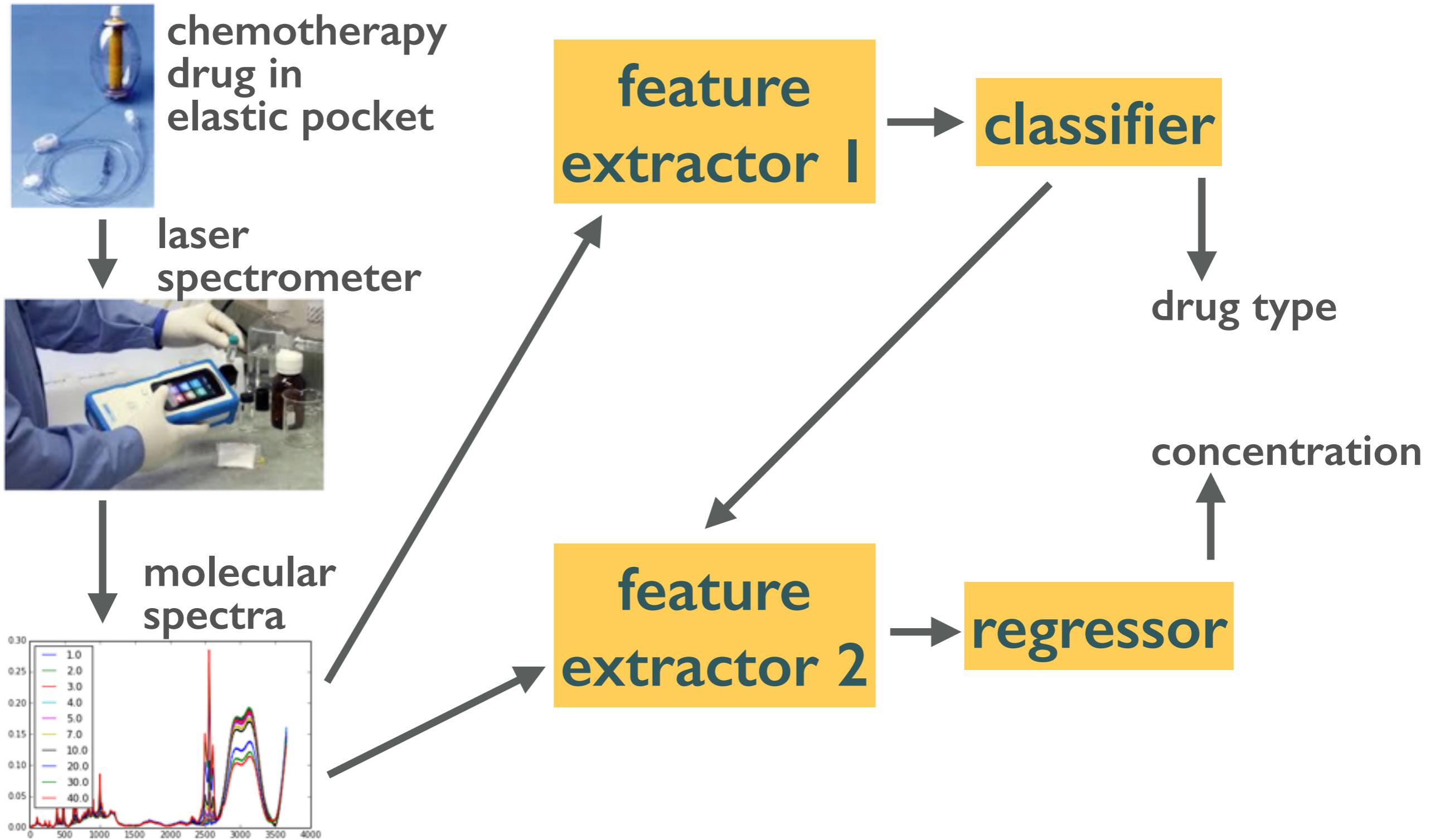
^f INRIA, Parietal team, Saclay, France

^g LTCI, Télécom ParisTech

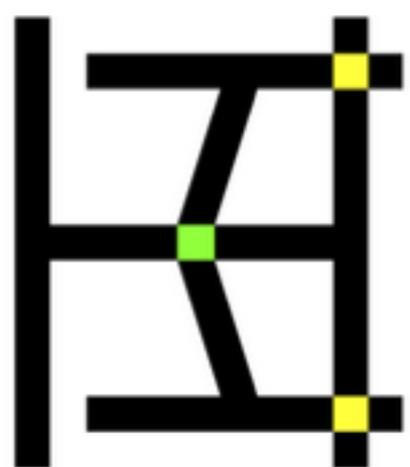
^h LAL, CNRS, France

26 March 2017

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA



Learning to discover: the Higgs boson machine learning challenge

Higgs 
challenge

Claire Adam-Bourdarios^a, Glen Cowan^b, Cécile Germain^c,
Isabelle Guyon^d, Balázs Kégl^{a,c}, David Rousseau^a

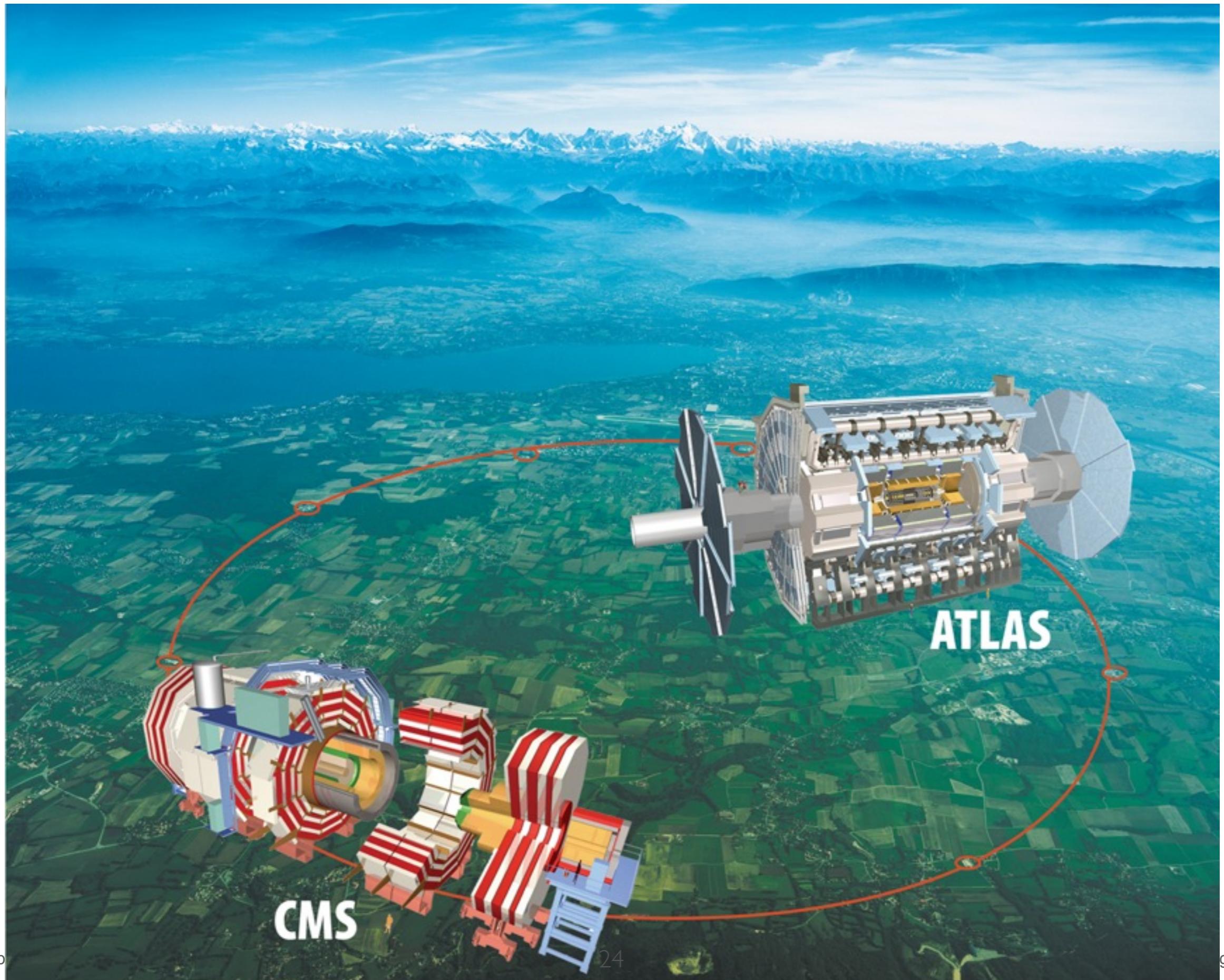
^a LAL, IN2P3/CNRS & University Paris-Sud, France

^b Physics Department, Royal Holloway, University of London, UK

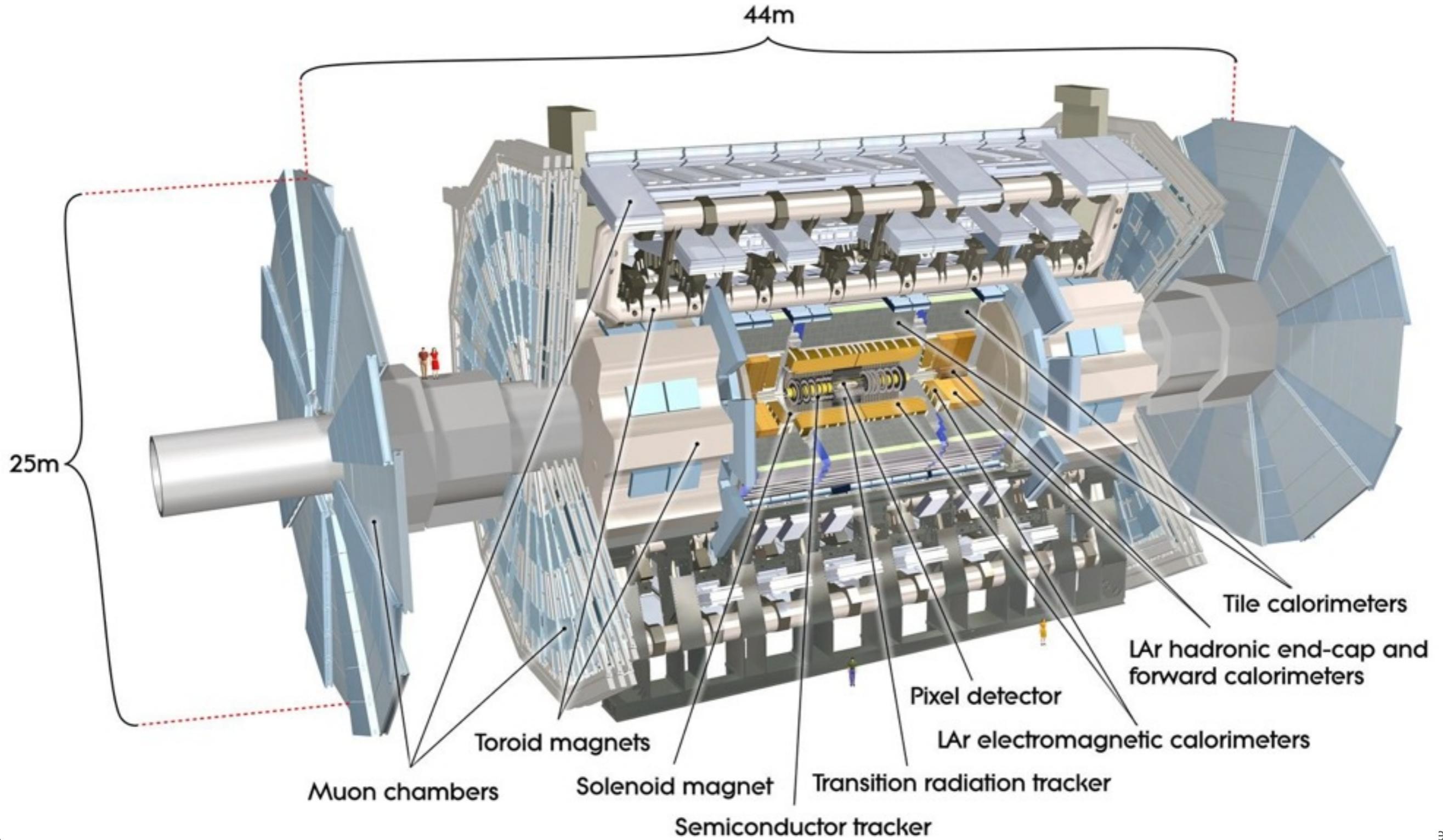
^c TAO team, INRIA & LRI, CNRS & University Paris-Sud, France

^d ChaLearn

THE LHC IN GENEVA

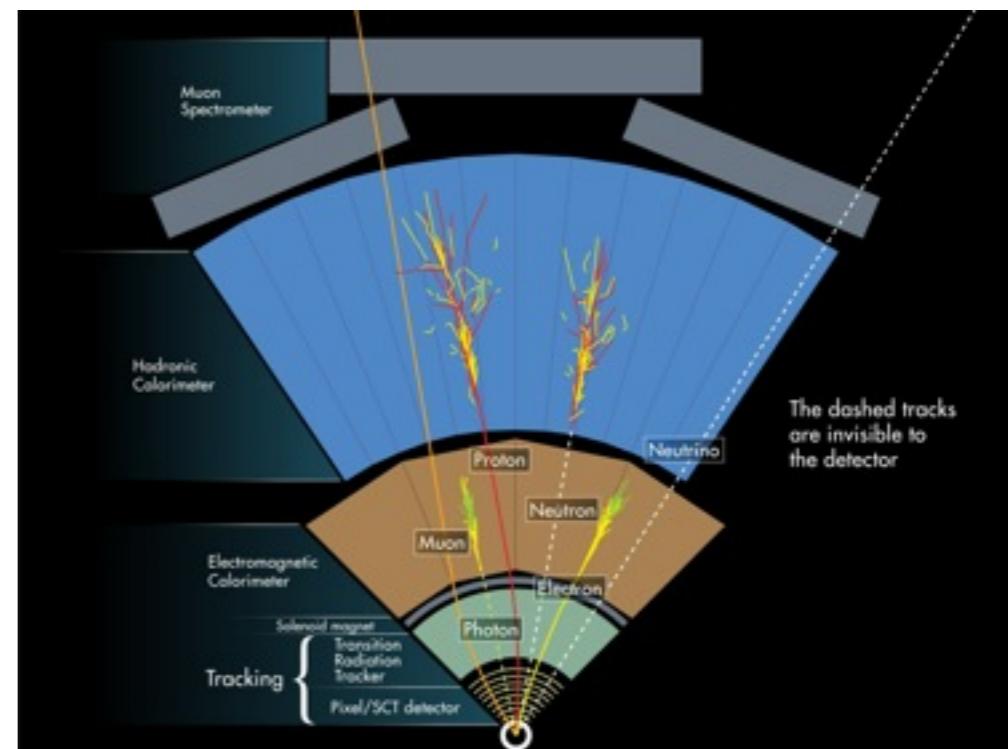
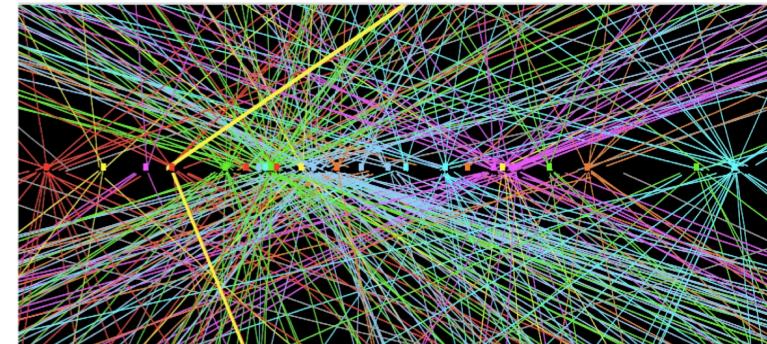


THE ATLAS DETECTOR



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

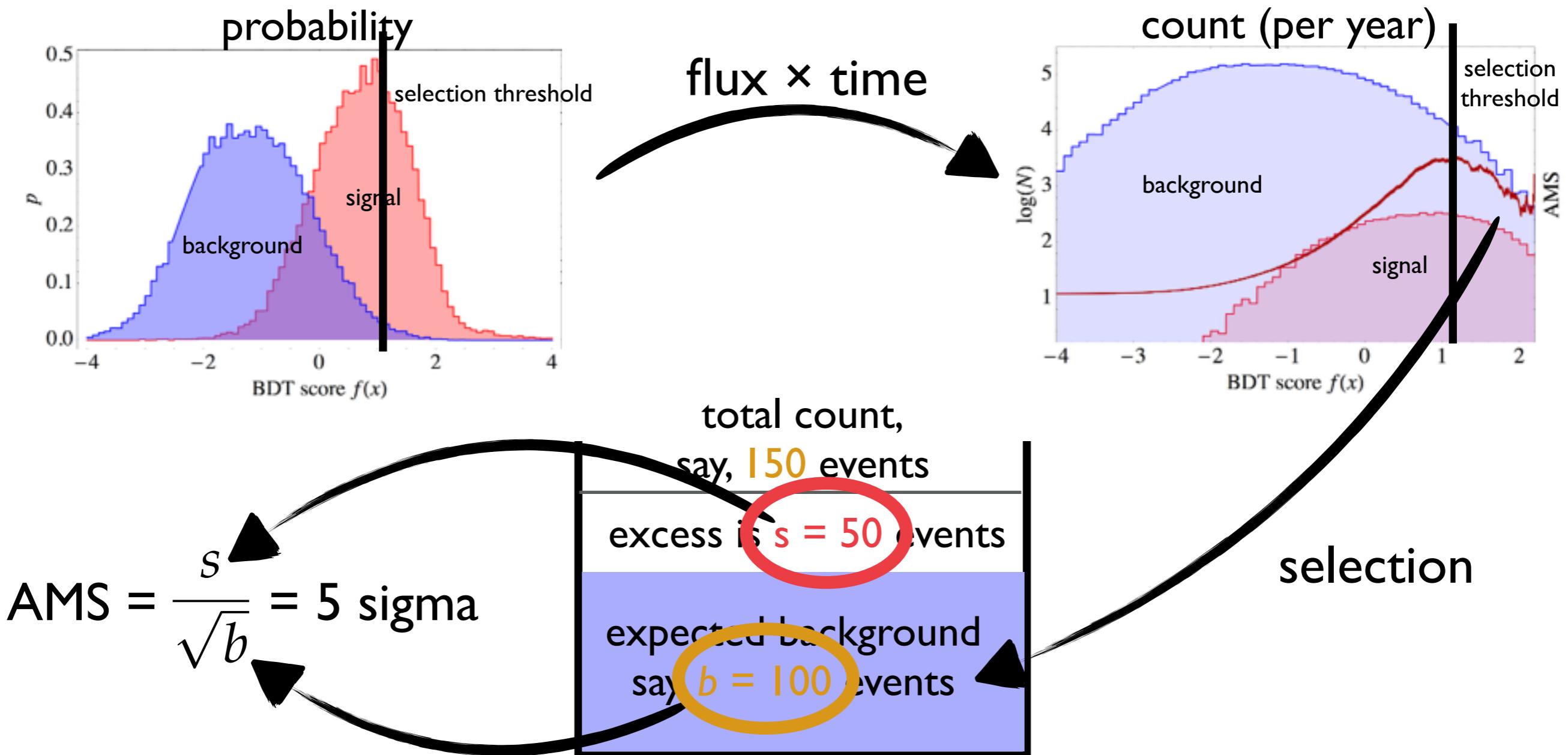


FEATURE ENGINEERING

- 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**



Generation and model reduction

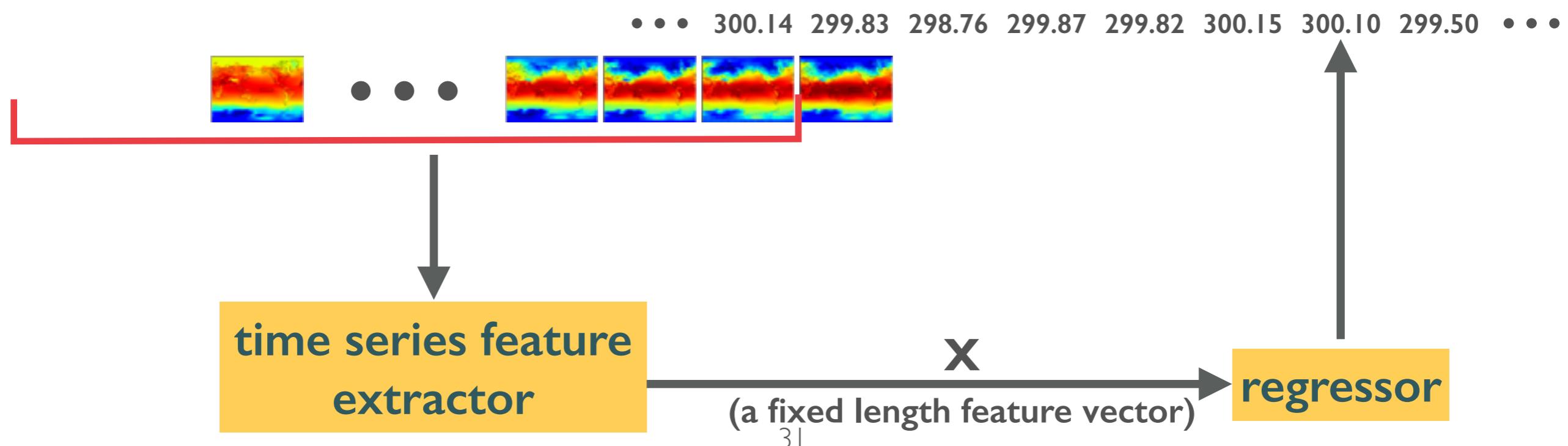
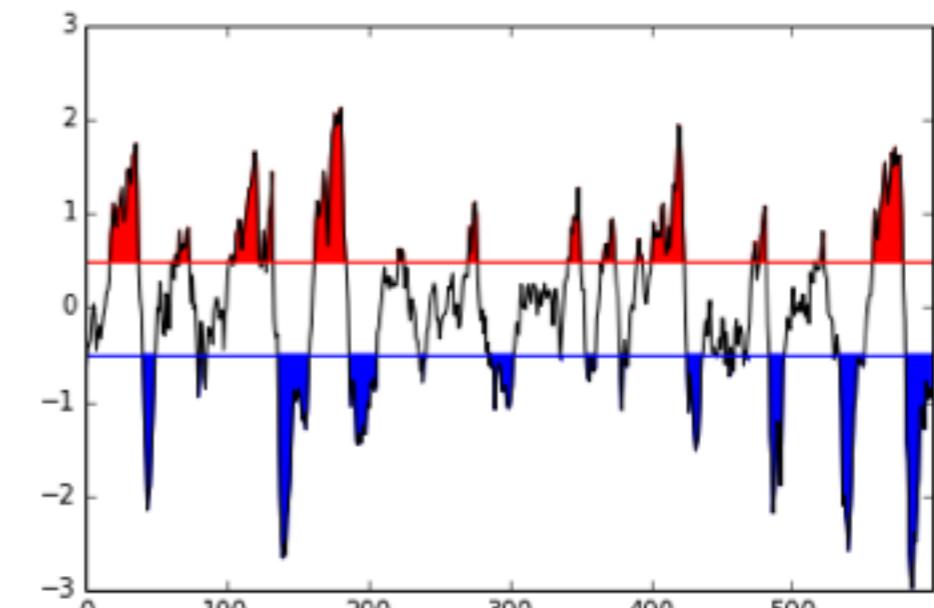
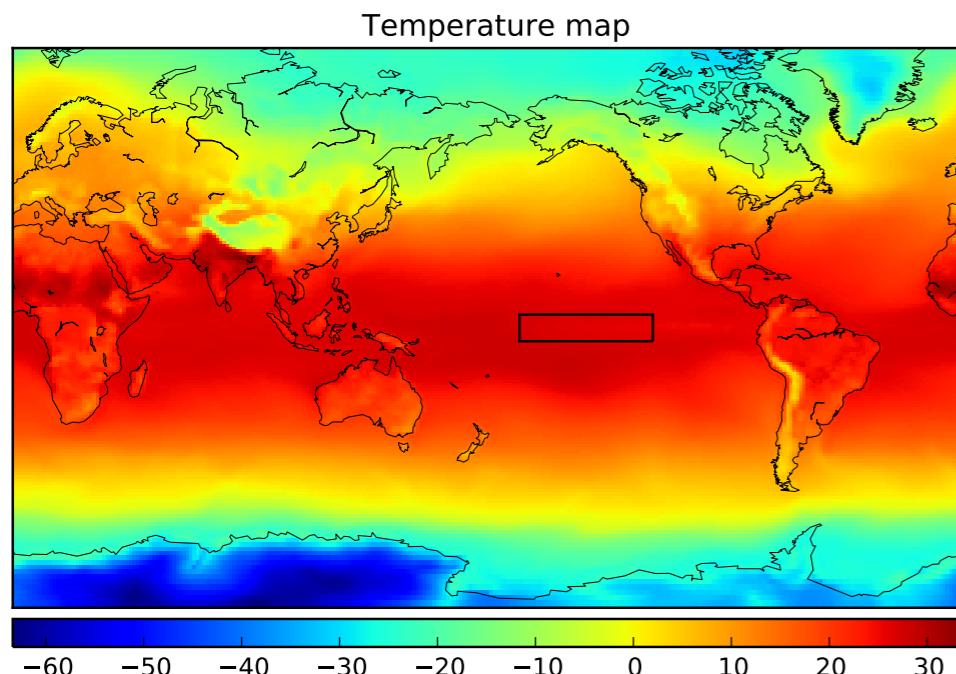
GENERATION AND MODEL REDUCTION

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

FORECASTING EL NIÑO: SPATIOTEMPORAL TIME SERIES

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



We have built and optimized
~20 scientific predictive workflows
for three years

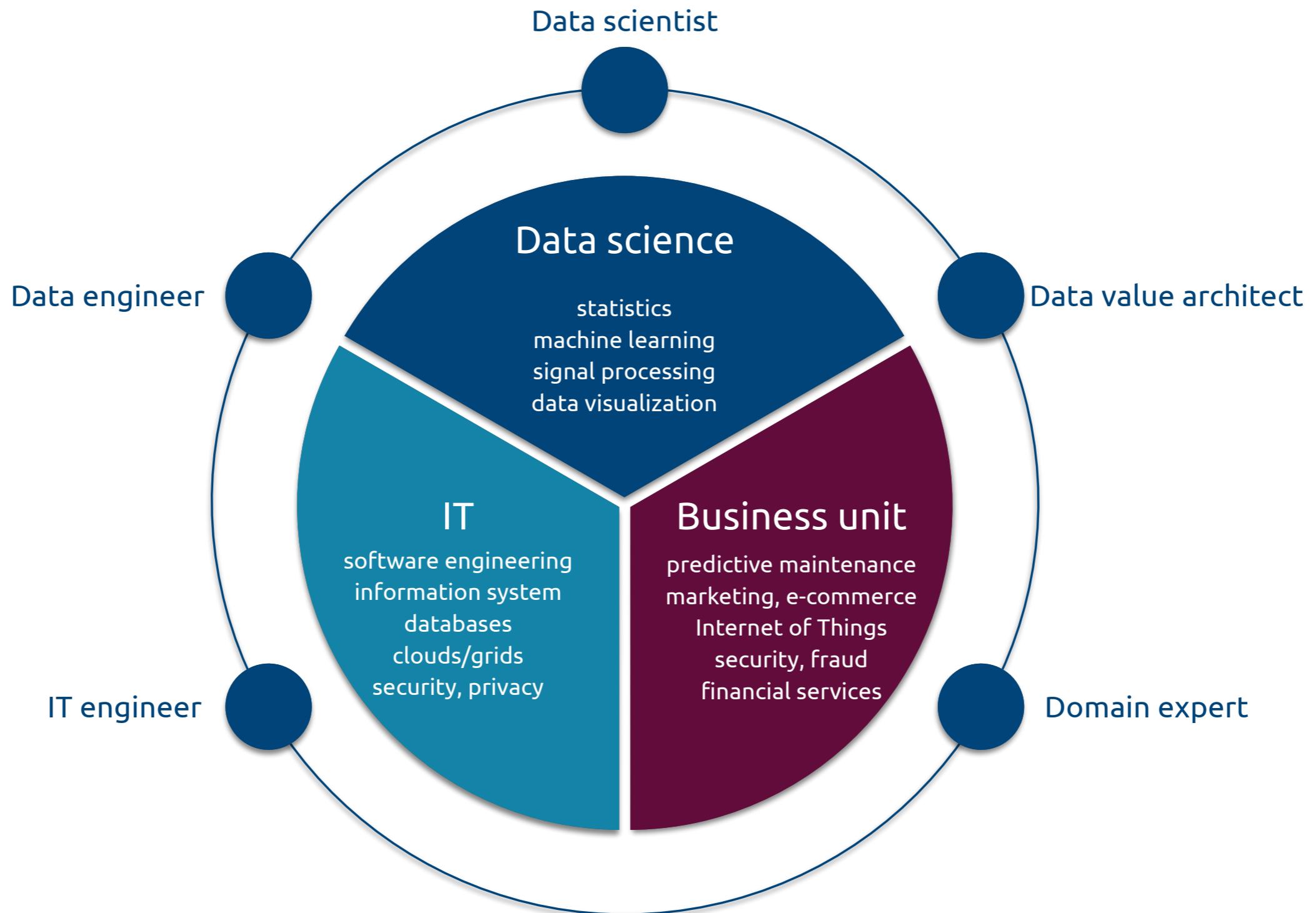
What have we learned?

BUILDING SCIENTIFIC WORKFLOWS

WHAT HAVE WE LEARNED?

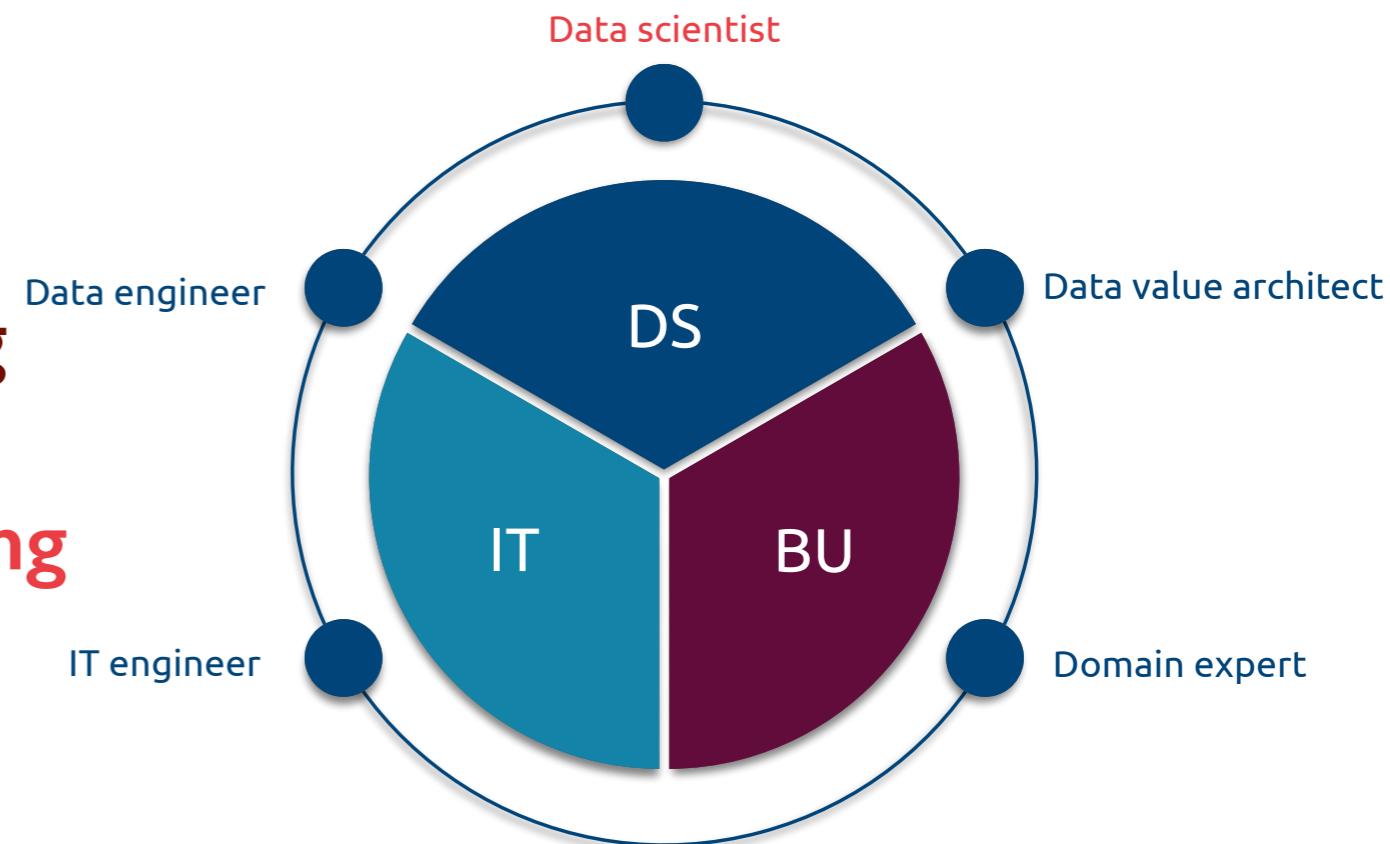
Roles and tasks in the data science process

THE DATA SCIENCE ECOSYSTEM



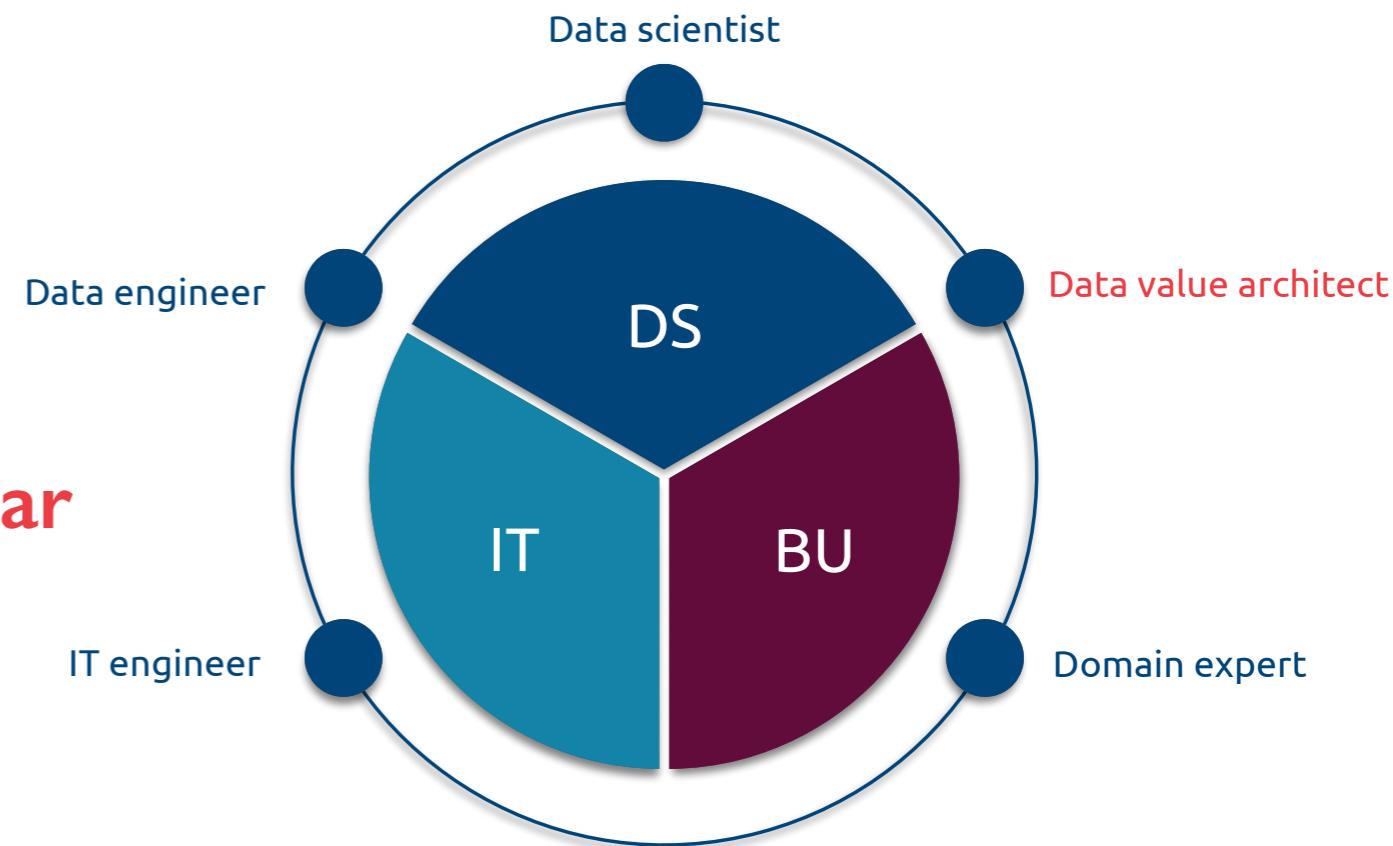
DATA SCIENTIST

- Technical expert in **machine learning, statistics, visualization, signal processing**
- Efficient in **cleaning** and **munging**
- Knows the latest **techniques** and **tools**
- Can handle different **data types** and **loss metrics**
- Can build adequate **prototype workflows**
- Knows how to **tune** (optimize) and **blend** models



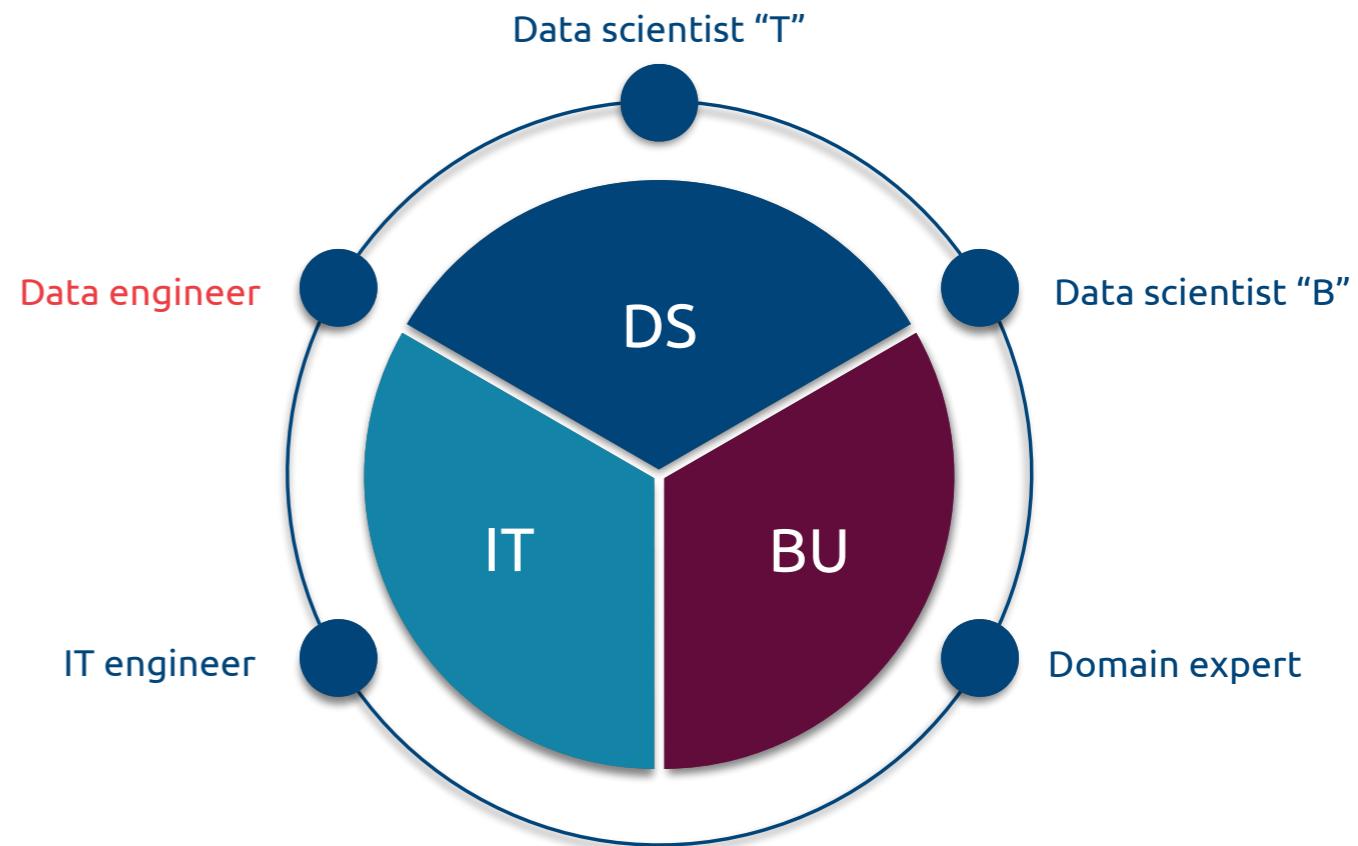
DATA VALUE ARCHITECT

- Has experience with a **wide variety of problems** and **technical solutions**
- Is possibly **expert in the particular domain**, or at least can **converse** with the domain expert
- Can translate **business/science goals into loss metrics**
- Can **formalize** adequate **prototype workflows**
- Can **estimate the costs** of building and running workflows
- Can define and dimension the **data collection** effort



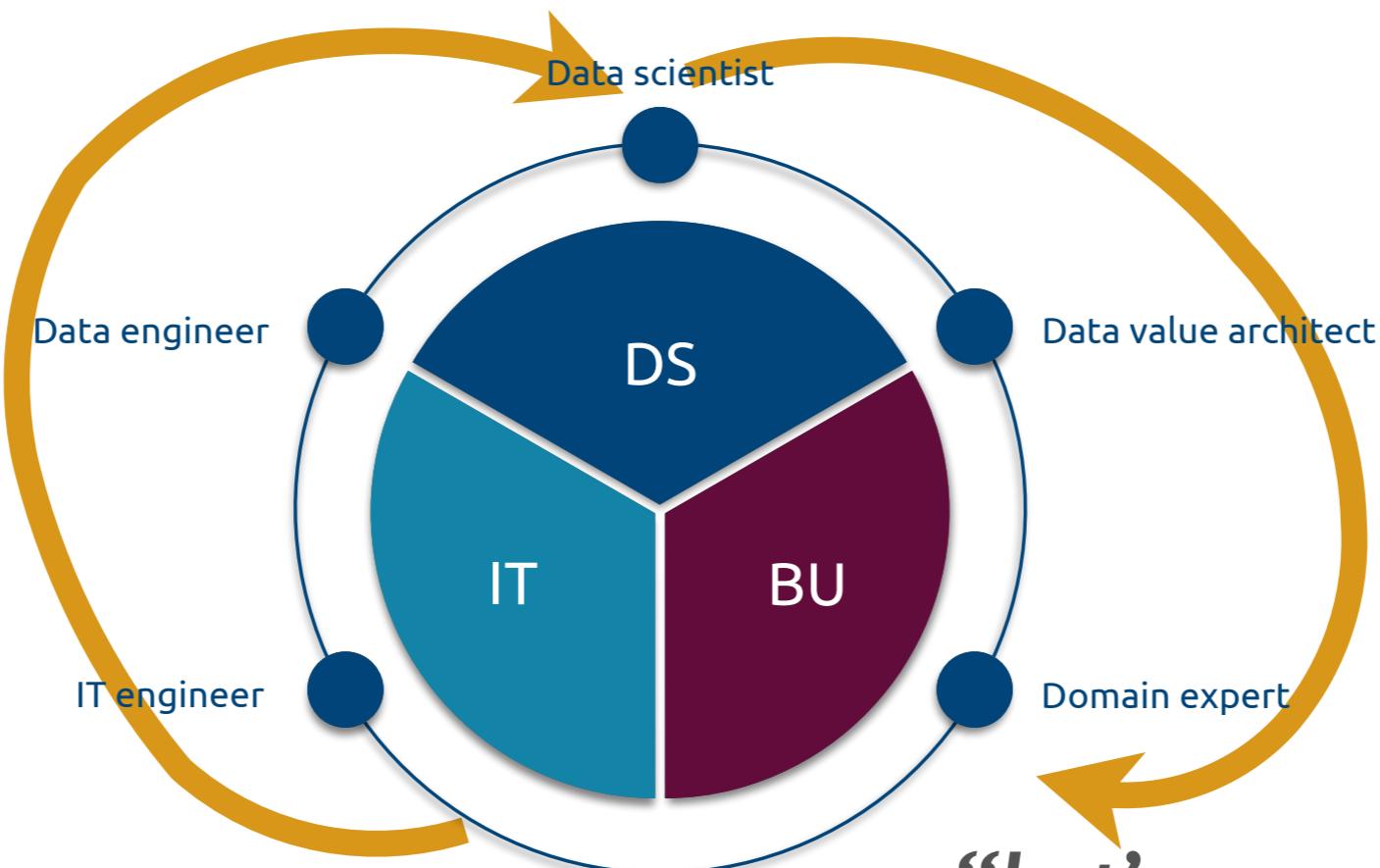
DATA ENGINEER

- Translates prototypes into **production workflows**, runs and maintains them
- Knows the latest **data engineering systems** and **architectures**
- Knows the **existing IT**
- Can **dimension the production workflows** and **estimate their costs**
- Knows the **basics of building** a data science workflow, and can feed the process by **extracting** and possibly **cleaning/munging** adequate data



BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT

“Let’s hire data scientists”

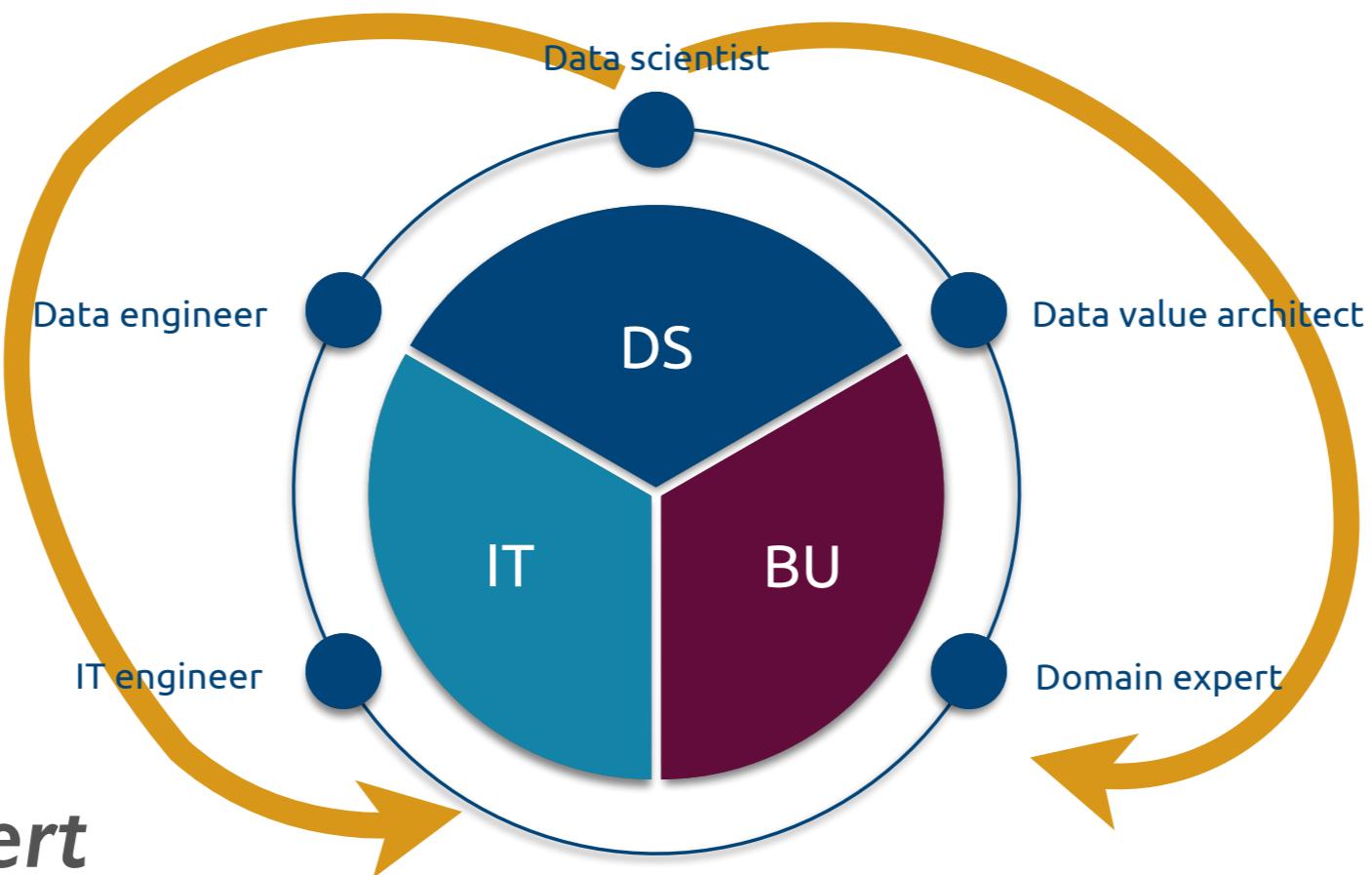


“Let’s install Hadoop”

“Let’s see what business problems we can solve with the existing data science team and the infrastructure we bought”

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

“Let’s hire data scientists.”

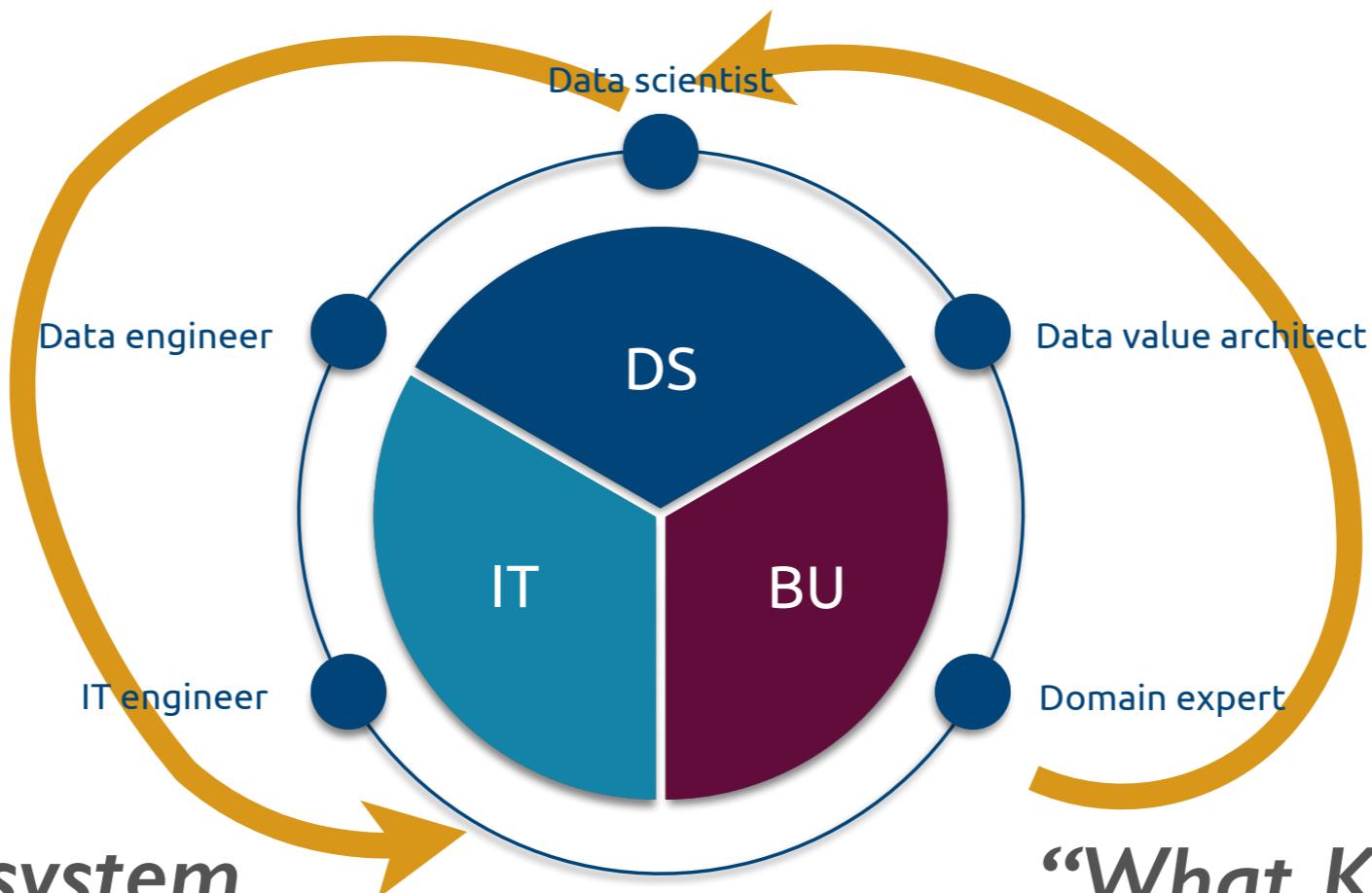


*“I’m an expert
of deep learning,
let’s buy a GPU cluster.”*

*“I’m an expert of deep
learning, let’s see what it
can do for your business.”*

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS

“Let’s hire data scientists for prototyping the business case.”



*“Let’s build a system
for putting the prototype
into production.”*

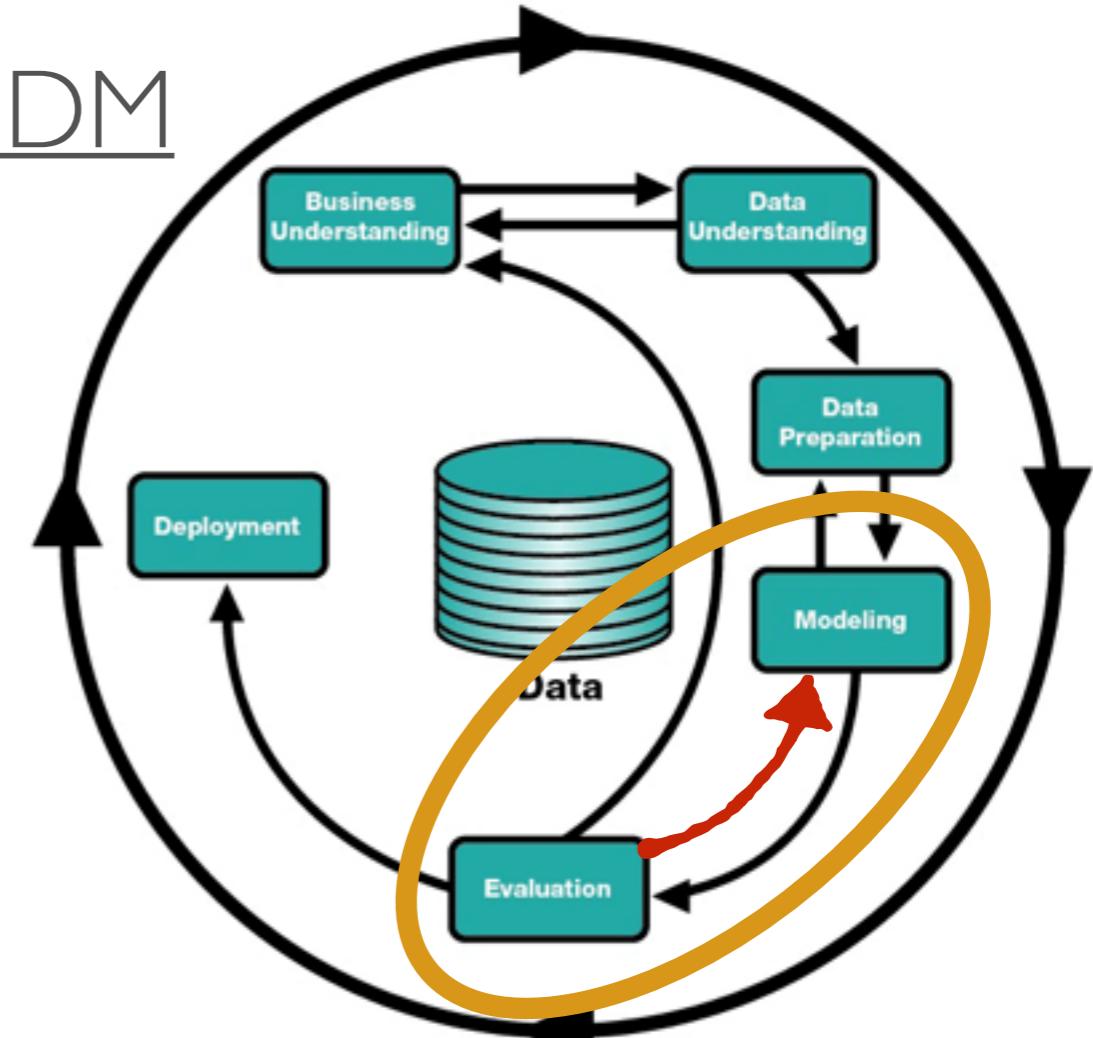
*“What KPI can we
improve with data?
What data should we
collect?”*

BUILDING PREDICTIVE SOLUTIONS

IBM CRISP-DM

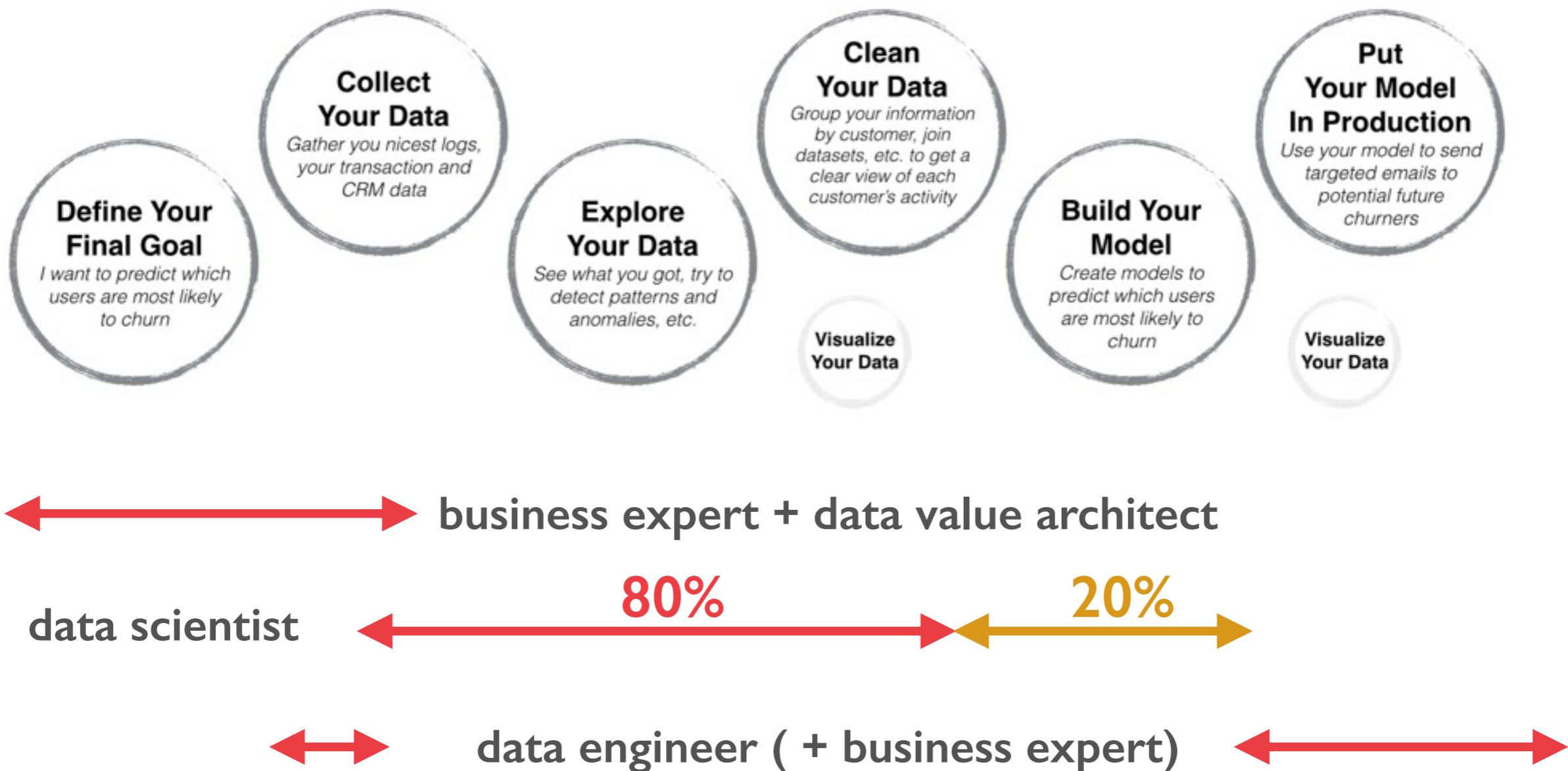
1996
2016

Dataiku



THE DATA ANALYTICS BUILDING PIPELINE

WHO DOES WHAT AND WHEN



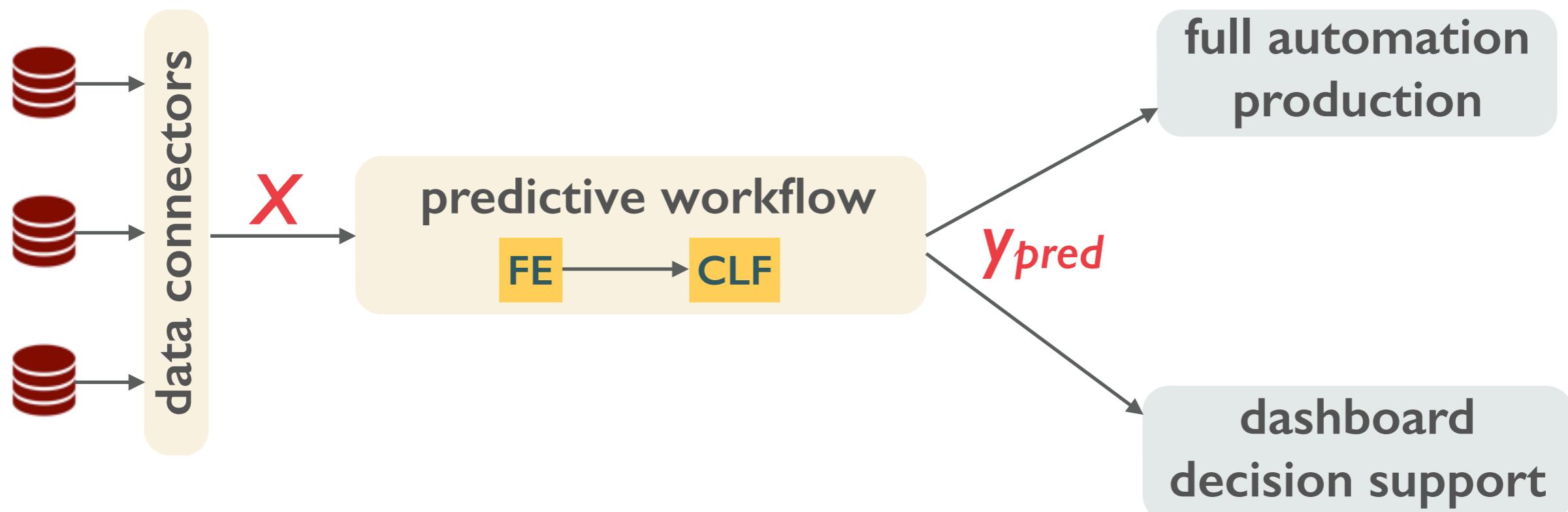
BUILDING SCIENTIFIC WORKFLOWS

WHAT HAVE WE LEARNED?

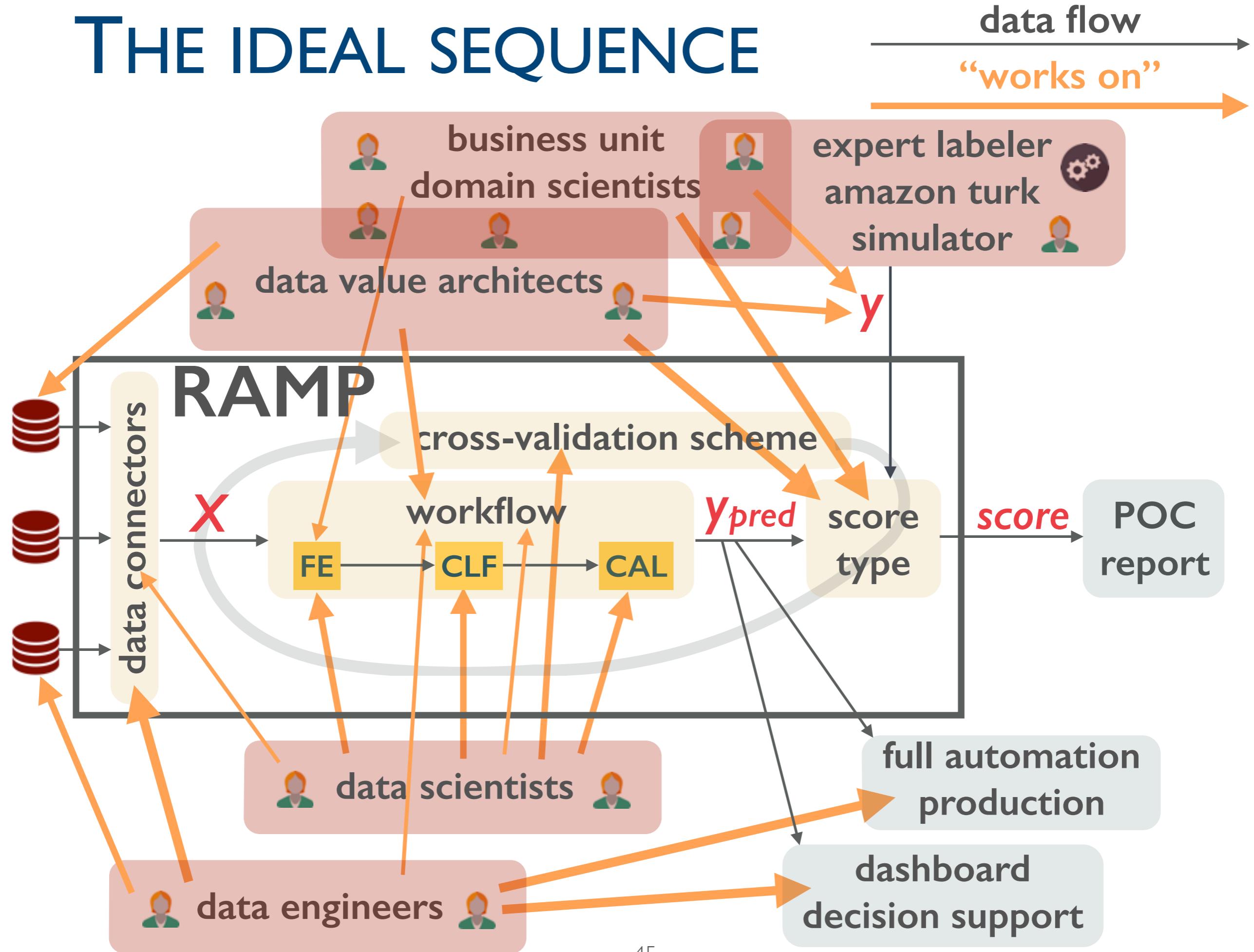
**Building the workflow:
what are the **tasks** and who does what**

THE PREDICTIVE WORKFLOW

data flow



THE IDEAL SEQUENCE



BUILDING THE PREDICTIVE WORKFLOW

- It is **trial and error**
 - little if any theory-based, model-based design
 - even research (development of new algorithms) is (mostly) trial and error
 - the data scientist's best friend is a **well-designed experimental studio** for facilitating fast iterations of
 - what **data** to use
 - what **features** to select or engineer
 - what **predictors** to use
 - how to **parametrize** the predictors

BUILDING SCIENTIFIC WORKFLOWS

WHAT HAVE WE LEARNED?

What is a predictive workflow?

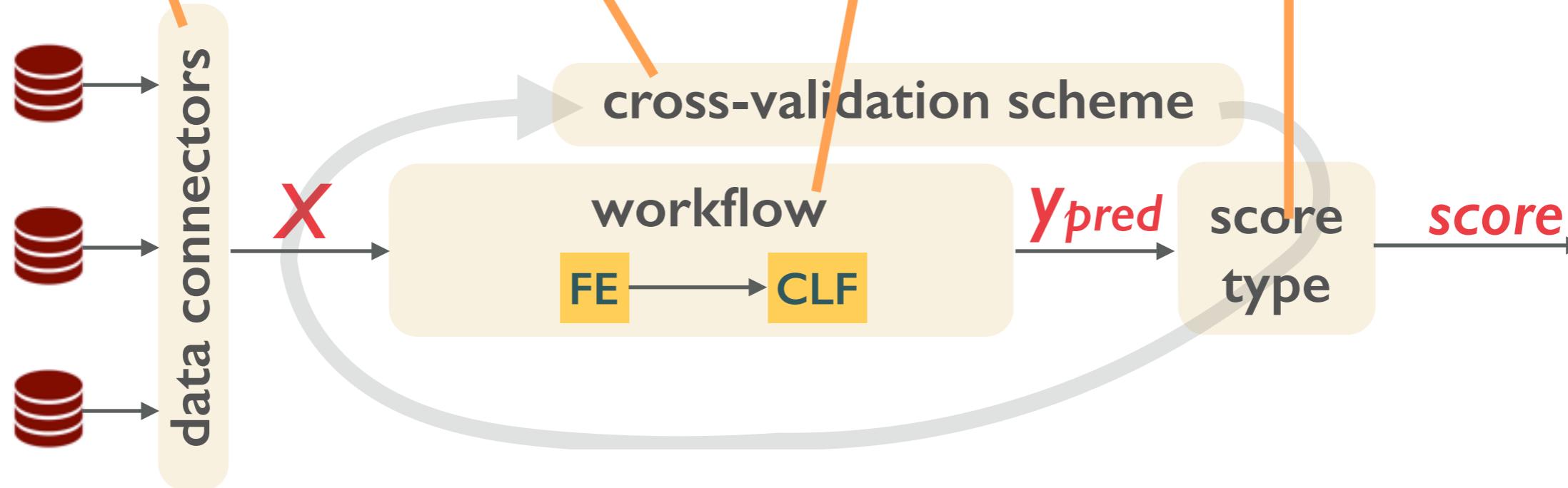
What are the parametrizable components?

**What can be put into a
unique training/scoring script?**

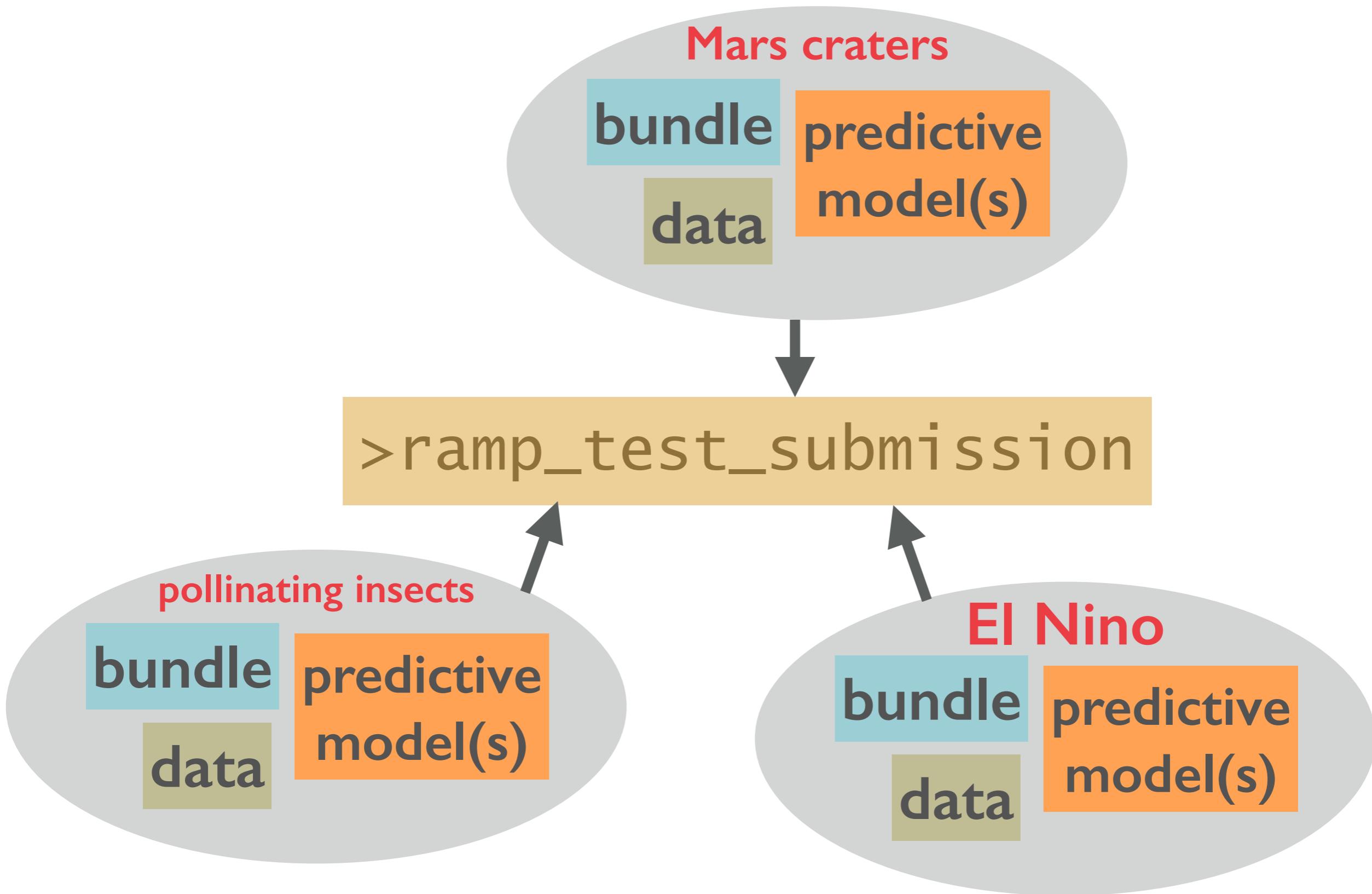
A SINGLE SCRIPT TO DEFINE THE BUNDLE

```
27  
28  
29 def get_cv(X, y):  
30     unique_replicates = np.unique(X['replicate'])  
31     r = np.arange(len(X))  
32     for replicate in unique_replicates:  
33         train_is = r[(X['replicate'] != replicate).values]  
34         test_is = r[(X['replicate'] == replicate).values]  
35         yield train_is, test_is  
36  
37  
38 def _read_data(path, f_name):  
39     data = pd.read_csv(os.path.join(path, 'data', f_name))  
40     y_array = data[_target_column_name]  
41     X_df = data.drop({_target_column_name}, axis=1)  
42     return X_df, y_array  
43  
44  
45 def get_train_data(path='.'):  
46     f_name = 'train.csv.gz'  
47     return _read_data(path, f_name)  
48  
49  
50 def get_test_data(path='.'):  
51     f_name = 'test.csv.gz'  
52     return _read_data(path, f_name)
```

```
1 import os  
2 import numpy as np  
3 import pandas as pd  
4 import rampwf as rw  
5  
6 problem_title = 'Cell population identification from single-cell mass cytometry data'  
7 _target_column_name = 'cell type'  
8 _prediction_label_names = [  
9     'B-cell Frac A-C (pro-B cells)', 'Basophils', 'CD4 T cells', 'CD8 T cells',  
10    'CLP', 'CMP', 'Classical Monocytes', 'Eosinophils', 'GMP', 'HSC',  
11    'IgD- IgMpos B cells', 'IgDpos IgMpos B cells', 'IgM- IgD- B-cells',  
12    'Intermediate Monocytes', 'MEP', 'MPP', 'Macrophages', 'NK cells',  
13    'NKT cells', 'Non-Classical Monocytes', 'Plasma Cells', 'gd T cells',  
14    'mDCs', 'pDCs']  
15  
16 # A type (class) which will be used to create wrapper objects for y_pred  
17 Predictions = rw.prediction_types.make_multiclass(  
18     label_names=_prediction_label_names)  
19 # An object implementing the workflow  
20 workflow = rw.workflows.FeatureExtractorClassifier()  
21  
22 score_types = [  
23     rw.score_types.BalancedAccuracy(name='bac', precision=3),  
24     rw.score_types.Accuracy(name='acc', precision=3),  
25     rw.score_types.NegativeLogLikelihood(name='nll', precision=3),  
26 ]
```



A UNIQUE SCRIPT TO RUN THE BUNDLES



A UNIQUE SCRIPT TO RUN THE BUNDLES

- 1 read training and test data
- 2 read submission
- 3 create train and valid folds on training data
- 4 for all train and valid folds:
 - 5 train submission on train
 - 6 score submission on train, valid, and test
- 7 summarize scores

```
silver6:autism kegl$ ramp_test_submission
Testing Autism Spectrum Disorder classification
Reading train and test files from ./data ...
Reading cv ...
Training ./submissions/startng_kit ...
CV fold 0
Couldn't re-order the score matrix..
  score  acc  auc
  test  0.696  0.765
  train  0.767  0.847
  valid  0.611  0.647
CV fold 1
Couldn't re-order the score matrix..
  score  acc  auc
  test  0.478  0.659
  train  0.766  0.842
  valid  0.628  0.662
CV fold 2
Couldn't re-order the score matrix..
  score  acc  auc
  test  0.609  0.720
  train  0.786  0.854
  valid  0.615  0.645
CV fold 3
Couldn't re-order the score matrix..
  score  acc  auc
  test  0.565  0.758
  train  0.769  0.849
  valid  0.619  0.645
-----
Mean CV scores
-----
Couldn't re-order the score matrix..
  score          acc          auc
  test  0.587 ± 0.0784  0.725 ± 0.042
  train  0.772 ± 0.0081  0.848 ± 0.0042
  valid  0.618 ± 0.0065  0.65 ± 0.0072
-----
Bagged scores
-----
Couldn't re-order the score matrix..
  score  auc
  test  0.735
  valid  0.647
```

RAMP-WORKFLOW & RAMP-KITS

- toolkit: <https://github.com/paris-saclay-cds/ramp-workflow>
 - for **designing workflows**
 - set of ready-made **metrics**, **workflows**, **CV schemes**, data readers
 - unique command-line **test script**
- examples: <https://github.com/ramp-kits>
 - a zoo of **problems**, **experiments**, **workflows**
 - (at least) one **initial solution**

BUILDING SCIENTIFIC WORKFLOWS

WHAT HAVE WE LEARNED?

How to make
(novice) data scientists efficient

HOW TO MAKE DATA SCIENTISTS EFFICIENT

- Principles
 - incite them to work on the problem
 - give them a working (but unoptimized) model to start with
 - make incremental contributions easy
 - gamify optimization
 - help them to collaborate and to learn from each other
 - “hide” heavy engineering and computational obstacles

THE JUPYTER NOTEBOOK

[
](http://www.datascience-paris-saclay.fr)

Paris Saclay Center for Data Science

RAMP on Pollinating insect classification

Mehdi Cherti (CNRS), Romain Julliard (MNHN), Gregoire Lois (MNHN), Balázs Kégl (CNRS)

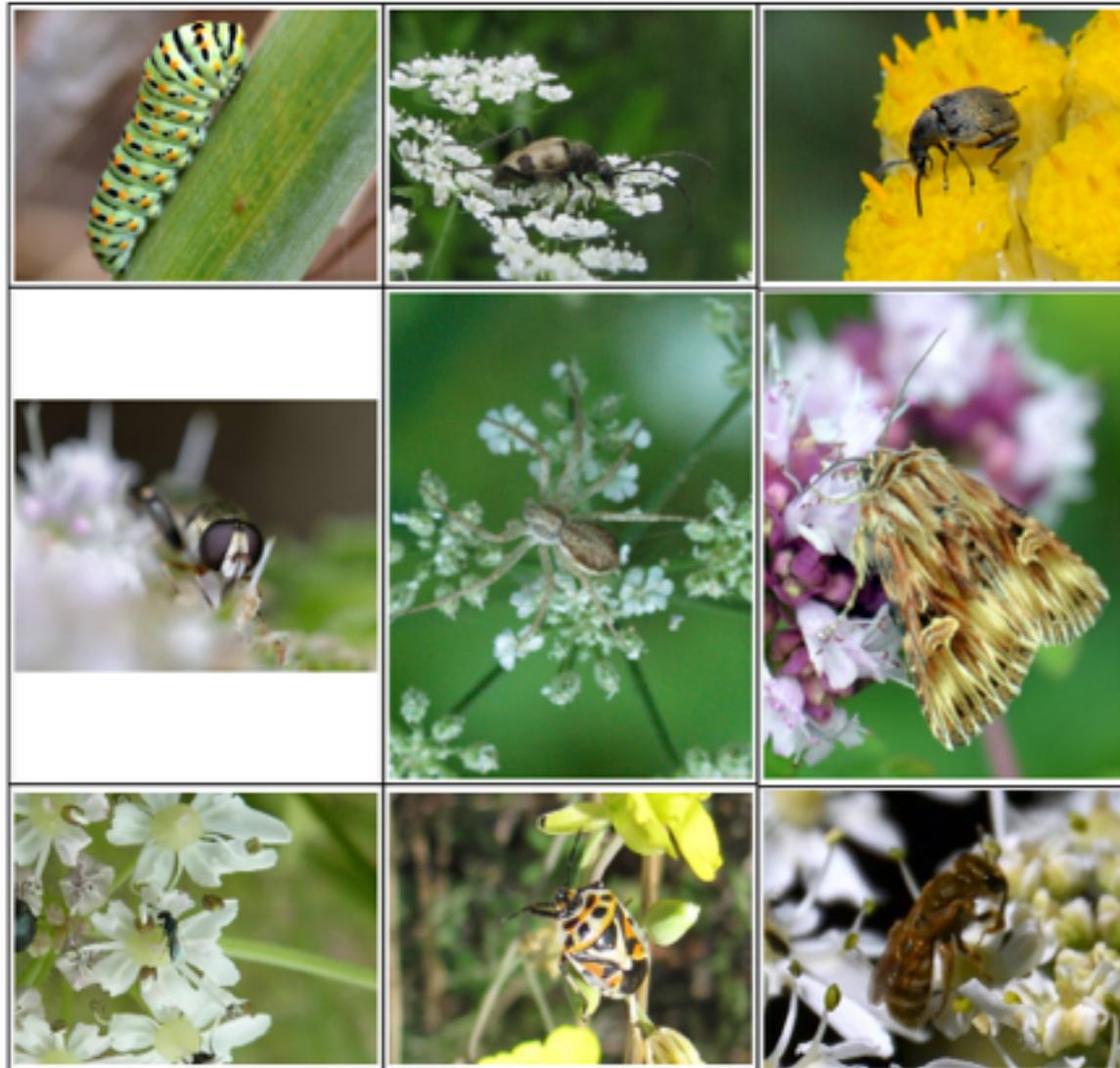
Introduction

Pollinating insects play a fundamental role in the stability of ecosystems. An insect is said to be pollinator when it transports pollen from one flower to another, helping them to accomplish fertilization. The vast majority of plants pollinates using insects, and at the same time, these insects depend on plants for their survival. However, because of human intensified agriculture, urbanisation and climate change, these species are threatened. 35% of human alimentation is based on plants pollinated by insects. Diversity of these insects is also important, the more diverse they are the best overall assistance is provided by these insects.

The [SPIPOLL](#) (Suivi Photographique des Insectes POLLinisateurs) project proposes to quantitatively study pollinating insects in France. For this, they created a crowdsourcing platform where anyone can upload pictures of insects and identify their species through a series of questions. These data are then used by specialists for further analyses.

Data

In this RAMP, we propose a dataset of pictures of insects from different species gathered from the SPIPOLL project and labeled by specialists. The dataset contains a set of 72939 labeled pictures of insects coming from 403 different insect species. Each picture is a color image. The size of the images (number of pixels) vary.



The prediction task

RAMP on Mars craters detection

Alexandre Boucaud (CDS), Joris van den Bossche (CDS), Balázs Kégl (CDS), Frédéric Schmidt (GEOPS), Anthony Lagain (GEOPS)

1. [Introduction](#)
2. [Preprocessing](#)
3. [Workflow](#)
4. [Evaluation](#)
5. [Local testing/exploration](#)
6. [Submission](#)

Introduction

Impact craters in planetary science are used to date planetary surfaces, to characterize surface processes and to study the upper crust of terrestrial bodies in our Solar System (Melosh, 1989). Thanks to the Martian crater morphology, a wide amount of information could be deduced on the geological history of Mars, as for example the evolution of the surface erosion rate, the presence of liquid water in the past, the volcanic episodes or the volatiles layer in the subsurface (Carr & Head, 2010). These studies are widely facilitated by the availability of reference crater databases.

Surveying impact craters is therefore an important task which traditionally has been achieved by means of visual inspection of images. The enormous number of craters smaller than one kilometer in diameter, present on high resolution images, makes visual counting of such craters impractical. In order to overcome this problem, several algorithms have been developed to automatically detect impact structures on planetary images (Bandeira et al., 2007 ; Martins et al., 2009). Nevertheless, these methods allow to detect only 70-80 % of craters (Urbach & Stepinski, 2009).

The prediction task

This challenge proposes to design the best algorithm to detect crater position and size starting from the most complete Martian crater database containing 384 584 verified impact structures larger than one kilometer of diameter (Lagain et al. 2017). We propose to give to the users a subset of this large dataset in order to test and calibrate their algorithm.



THE STARTING KIT

feature_extractor.py

```
1 from sklearn.base import BaseEstimator  
2 from sklearn.base import TransformerMixin  
3  
4  
5 class FeatureExtractor(BaseEstimator, TransformerMixin):  
6     def fit(self, X_df, y):  
7         return self  
8  
9     def transform(self, X_df):  
10        # get only the anatomical information  
11        X = X_df[[  
12            col for col in X_df.columns  
13            if col.startswith('anatomy')]]  
14        return X.drop(columns='anatomy_select')  
15
```

classifier.py

```
1 from sklearn.base import BaseEstimator  
2 from sklearn.preprocessing import StandardScaler  
3 from sklearn.linear_model import LogisticRegression  
4 from sklearn.pipeline import make_pipeline  
5  
6  
7 class Classifier(BaseEstimator):  
8     def __init__(self):  
9         self.clf = make_pipeline(  
10             StandardScaler(), LogisticRegression(C=1.))  
11  
12     def fit(self, X, y):  
13         self.clf.fit(X, y)  
14         return self  
15  
16     def predict(self, X, y):  
17         return self.clf.predict(X)  
18  
19     def predict_proba(self, X):  
20         return self.clf.predict_proba(X)  
21
```

THE FRONTEND

The screenshot shows the RAMP Sandbox interface. On the left, there's a sidebar with icons for user profile, dashboard, and help. The main header says "Secure | https://www.ramp.studio/events/autism/sandbox". The top navigation bar includes links for "RAMP", "Leaderboard", "Events", "Help", and "Logout". A message "Hi Balázs!" is displayed.

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!

feature_extractor

```
1 from sklearn.base import BaseEstimator
2 from sklearn.base import TransformerMixin
3
4
5 class FeatureExtractor(BaseEstimator, TransformerMixin):
6     def fit(self, X_df, y):
7         return self
8
9     def transform(self, X_df):
10        # get only the anatomical information
11        X = X_df[[col for col in X_df.columns if col.startswith('anatomy')]]
12        return X.drop(columns='anatomy_select')
13
```

Upload your files!

File list

- feature_extractor.py**
- classifier.py**

Upload file

No file chosen

classifier

```
1 from sklearn.base import BaseEstimator
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.linear_model import LogisticRegression
4 from sklearn.pipeline import make_pipeline
```

THE LEADERBOARD

Secure | https://www.ramp.studio/events/autism/competition_leaderboard

RAMP Hi Balázs!

autism

Leaderboard

Combined score: 1.0

Show 10 entries Search:

rank	team	submission	auc	train time [s]	test time [s]	submitted at (UTC)
1	Warvito	sub_EP	0.997	151	29	2018-06-18 21:21:25 Mon
2	vfdev.5	lr_f23_fmri_nv=0.998	0.996	50	50	2018-06-16 17:18:28 Sat
3	jmr	test13	0.976	559	149	2018-06-15 10:40:37 Fri
4	pearrr	test9	0.948	786	153	2018-06-18 13:40:47 Mon
5	Gowtham_Murugesan	test_009	0.934	695	202	2018-06-13 17:05:50 Wed
6	nherran	test_22_part_4	0.908	350	131	2018-06-12 18:16:51 Tue
7	zhipeng	con_test1_1	0.806	542	170	2018-06-22 12:37:56 Fri
8	mmunoz	subm_3	0.776	152	71	2018-06-22 22:25:54 Fri
9	wwwmmmm	mengxb-3	0.772	773	50	2018-06-06 01:01:45 Wed
10	SRSteinkamp	2atl_mlp	0.772	72	26	2018-06-18 04:50:24 Mon

Showing 1 to 10 of 39 entries

Previous 1 2 3 4 Next

THE BACKEND ON AMAZON WEB SERVICES

The screenshot shows the AWS CloudWatch Metrics console interface. At the top, there are three buttons: "Launch Instance" (blue), "Connect" (grey), and "Actions" (grey). Below these is a search bar with a magnifying glass icon and the placeholder text "Filter by tags and attributes or search by keyword". The main area is a table listing eight Amazon EC2 instances. The columns are: a checkbox column, "Name", "Instance ID", "Instance Type", "Availability Zone", and "Instance Status". The "Name" column is sorted by name. The "Status" column uses colored circles to indicate instance status: orange for terminated and green for running.

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance Status
<input type="checkbox"/>	8475_submission_id0	i-0f9411820a116930e	m5.xlarge	us-west-2a	terminated
<input type="checkbox"/>	8474_test2	i-0a8afc1be7d824cb1	m5.xlarge	us-west-2a	terminated
<input type="checkbox"/>	8473_test	i-0322ebb27a7e4e5...	m5.xlarge	us-west-2a	terminated
<input type="checkbox"/>	8473_test	i-09f50b72b6a3a0155	m5.xlarge	us-west-2a	terminated
<input type="checkbox"/>	8472_TunedClassif+svc	i-04dca6a8fd8738b5e	t2.small	us-west-2c	terminated
<input type="checkbox"/>	8472_TunedClassif+svc	i-0f0529a6ab890f17c	t2.small	us-west-2b	terminated
<input type="checkbox"/>	8469_test2	i-049ba0f2c8075fdd6	m5.xlarge	us-west-2a	terminated
<input type="checkbox"/>	8467_TunedClassif+svm	i-08018a0ee2b90b2ec	t2.small	us-west-2b	running

Funded by Université Paris-Saclay and CNRS

Team



Balázs Kégl



Alex Gramfort



Akin Kazakçı



Mehdi Cherti



Yohann Sitruk



Guillaume Lemaître



Alexandre Boucaud



Joris Van den Bossche

Alumni



Djalel Benbouzid



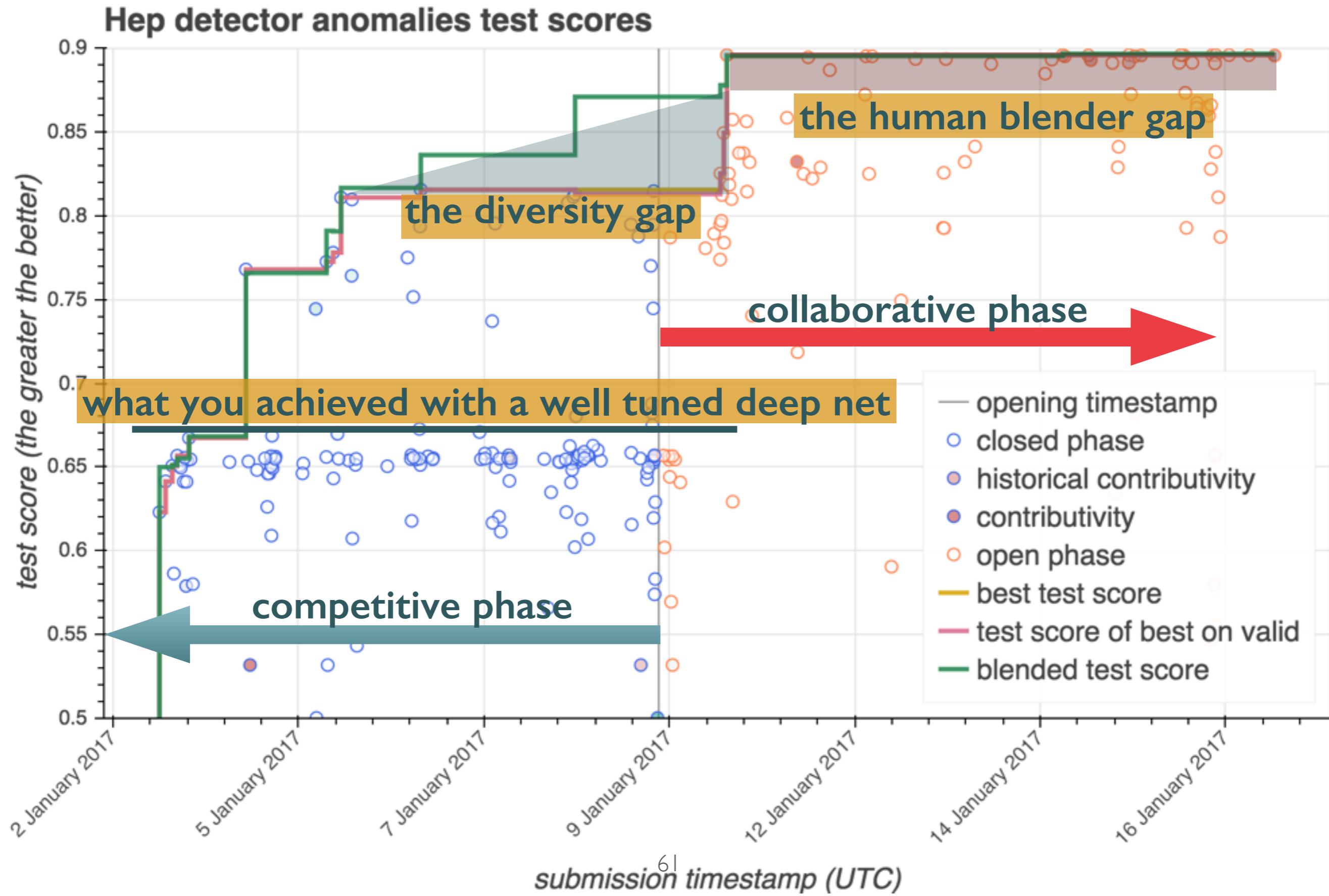
Camille Marini

Why code submission

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

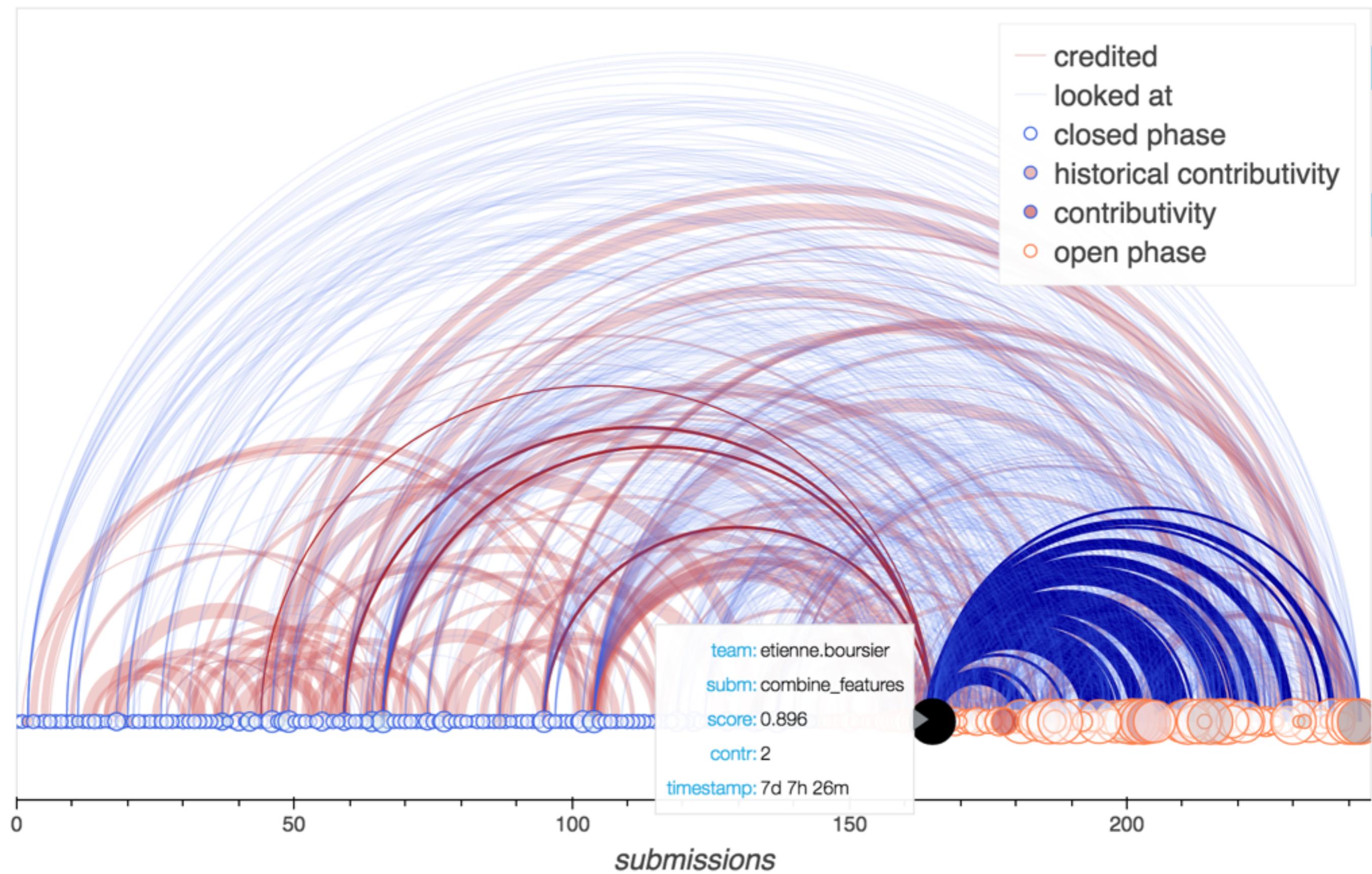
THE POWER OF THE (COLLABORATING) CROWD

OPTIMIZING GRADUATE STUDENT DESCENT



COMMUNICATION AND REUSE

Hep detector anomalies submissions



You can

1. Participate in upcoming RAMPs
2. Use RAMP in teaching or training
3. Use the toolkit for your own workflows
4. Submit it to us if you want to run a data challenge

LINKS

frontend:

www.ramp.studio

toolkit:

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

examples:

github.com/ramp-kits

slack:

ramp-studio.slack.com

READING MATERIAL

- medium.com/@balazskegl
 - The data science ecosystem (industrial edition)
 - Teaching the data science process
 - How to build a data science pipeline
- **RAMP paper**
 - <https://openreview.net/forum?id=Syg4NHz4eQ>