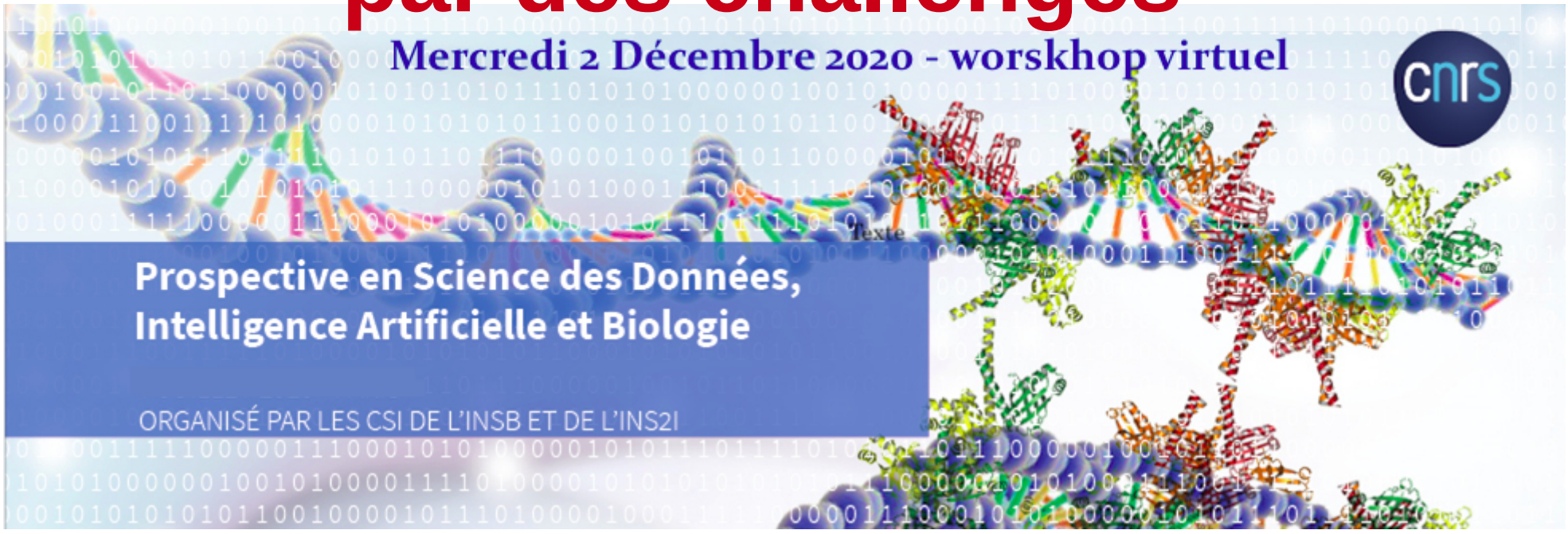
Faire progresser la science des données par des challenges

Bertrand Thirion, bertrand.thirion@inria.fr

Faire progresser la science par des challenges



Bertrand Thirion, bertrand.thirion@inria.fr

AI-Data science at Paris-Saclay

université PARIS-SACLAY

Paris-Saclay Center for Data Science

- Teaching
- Challenge organization
- Open data, OSS

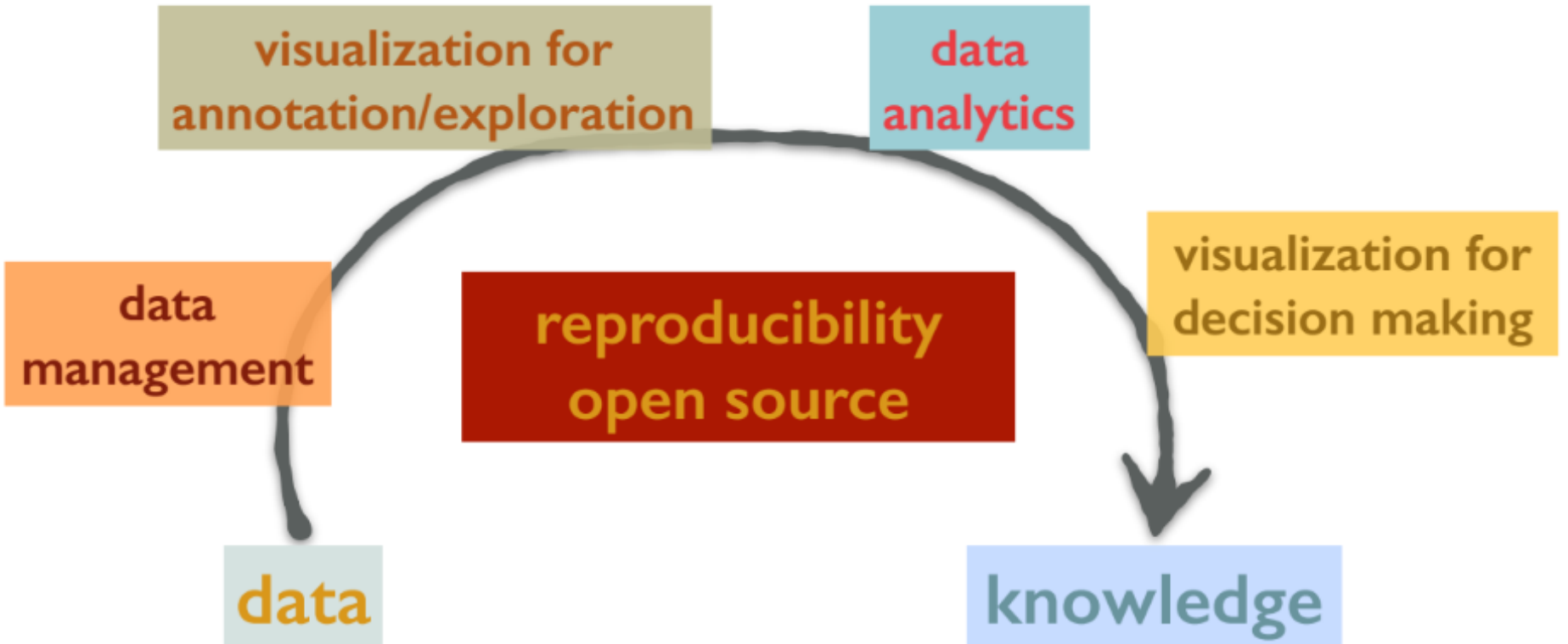
INSTITUT DATAIA
Science des données, Intelligence & Société

Challenges as a crowdsourcing endeavour

- crowdsourcing combines
  - bottom-up creative intelligence of a community that volunteers solutions
  - top-down management of an organization that poses the problem.
- **1714**, **British Board of Longitude Prize**: determine the longitude of a ship at sea.
  - Won by unknown clockmaker John Harrison for his invention of the marine chronometer.

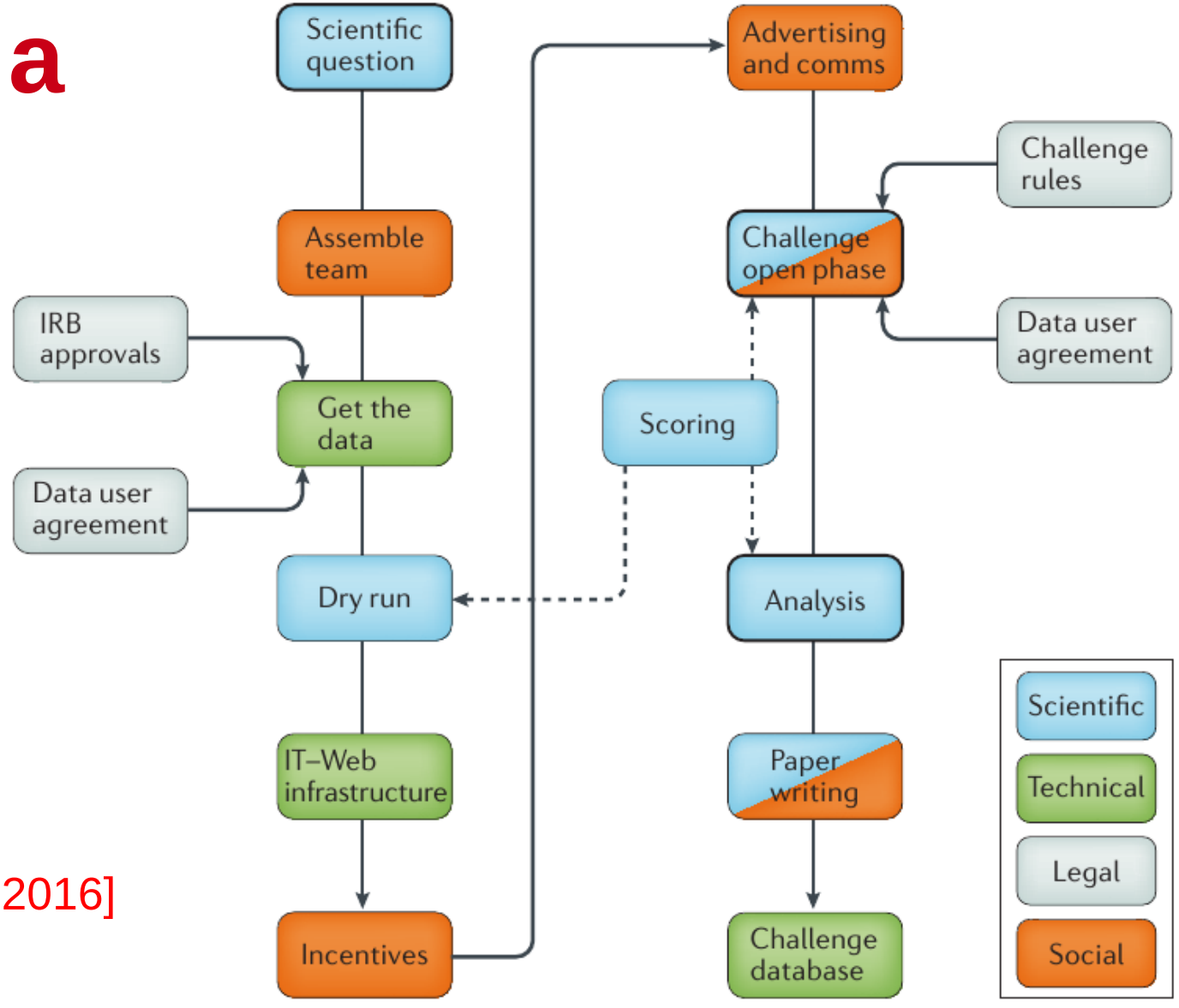
[Daez-Rodrigues et al. Nature 2016]

It’s all about inference



Organizing a challenge

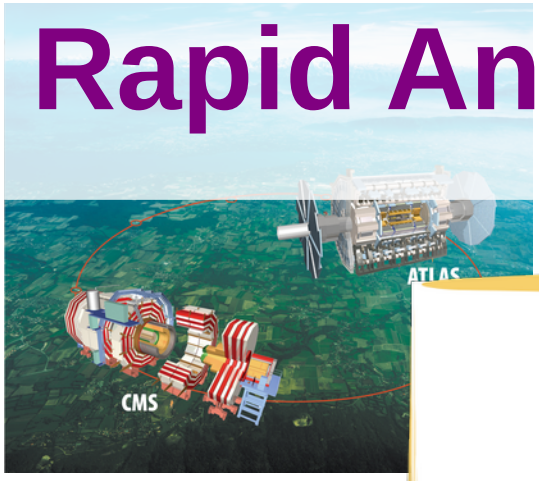
4 kinds of steps:
scientific,
technical,
social, legal

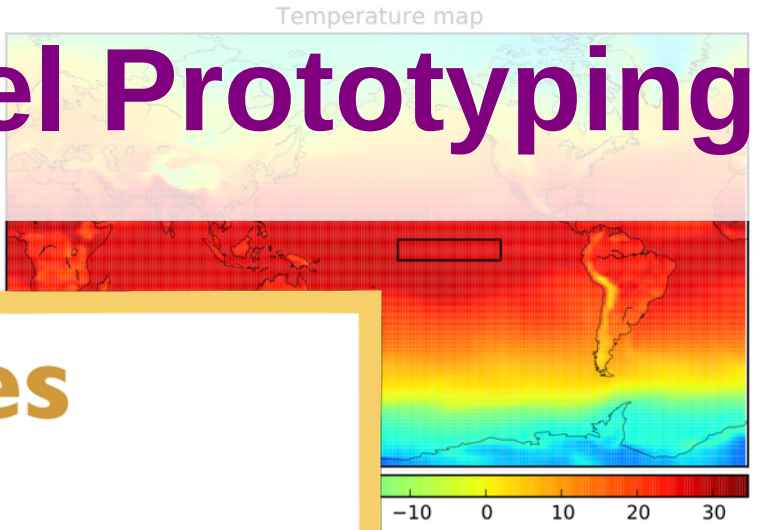


[Daez-Rodrigues et al. Nature 2016]

RAMP infrastructure

Rapid Analytics & Model Prototyping





16 challenges

31 events

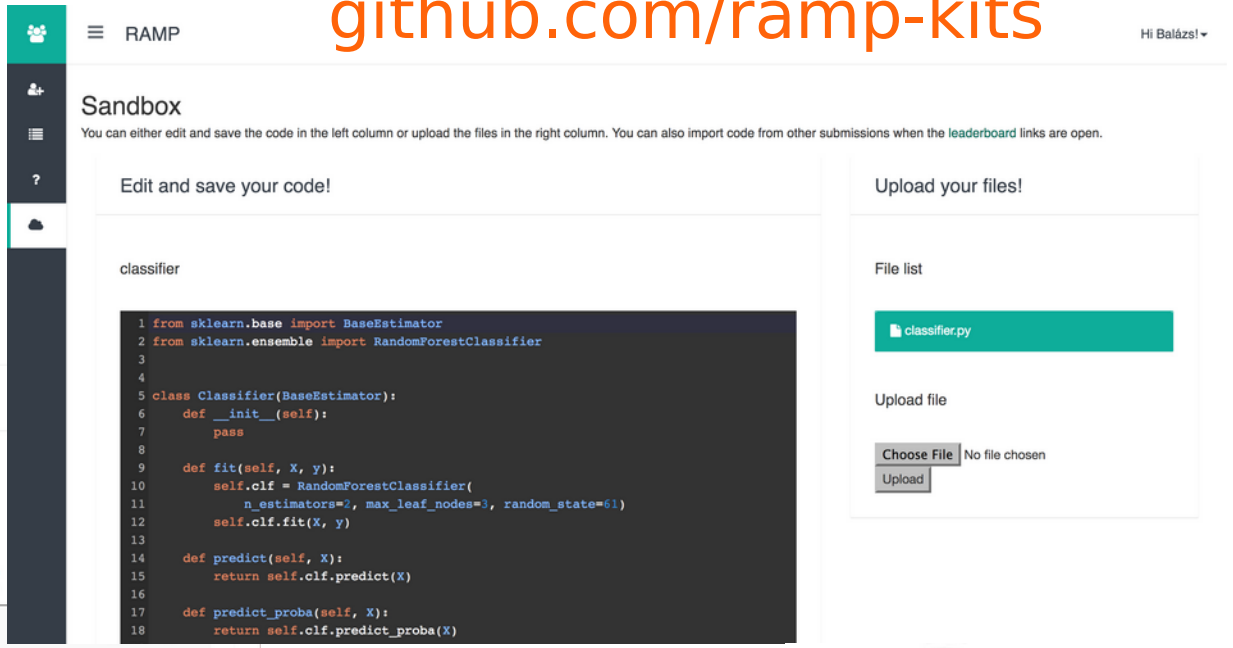
996 users

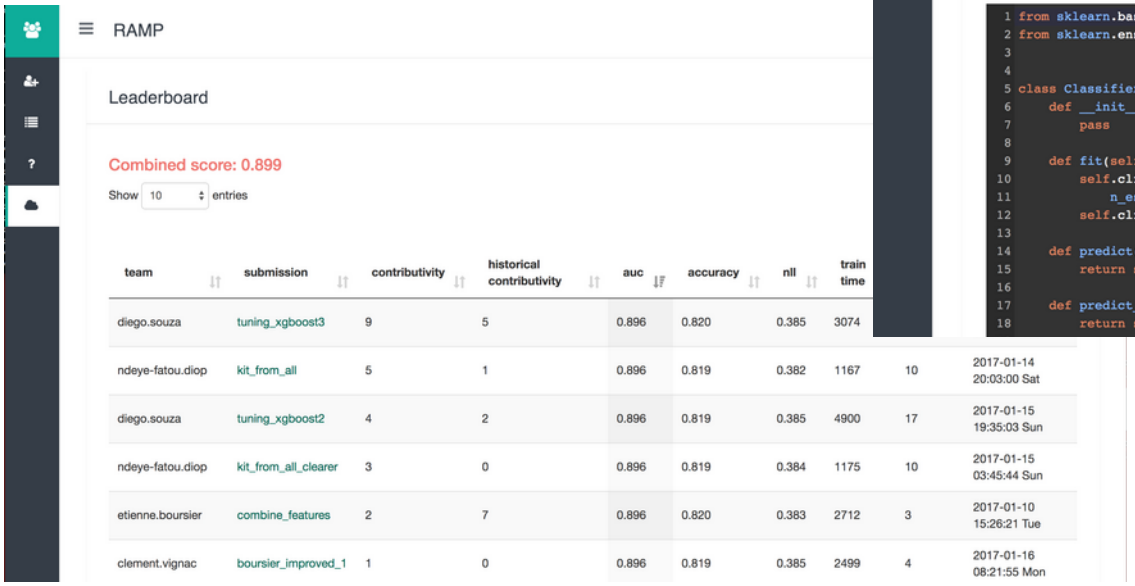
5685 predictive models

RAMP.STUDIO

DATA CHALLENGE WITH **CODE SUBMISSION**

github.com/ramp-kits





OVHcloud

amazon web services™

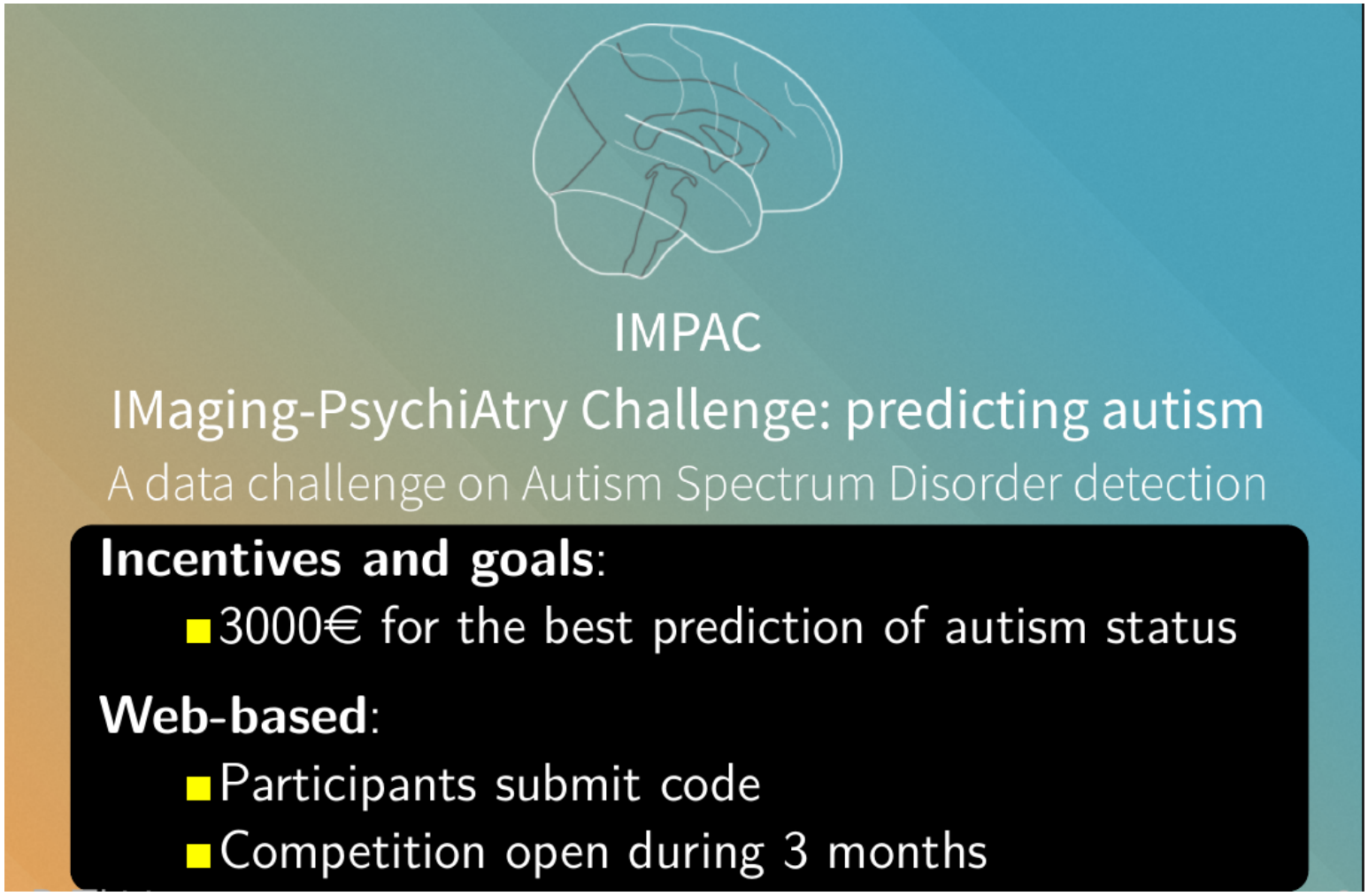
An autism challenge

Collaboration with:
Roberto Torro, Pasteur
Balazs Kegl, Center for Data Science
Gael Varoquaux

Guillaume Lemaître (Inria/CDS2) did the hard work

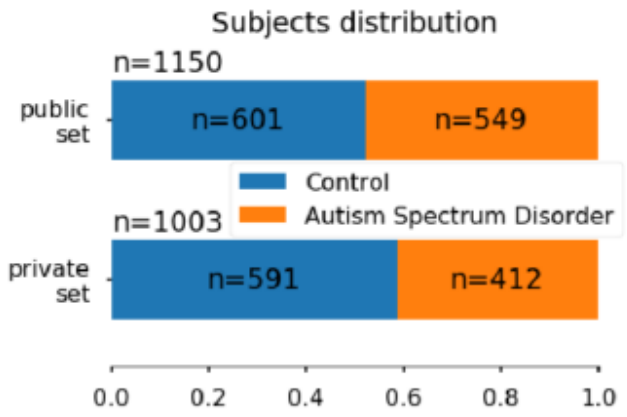
http://paris-saclay-cds.github.io/autism_challenge

IMPAC

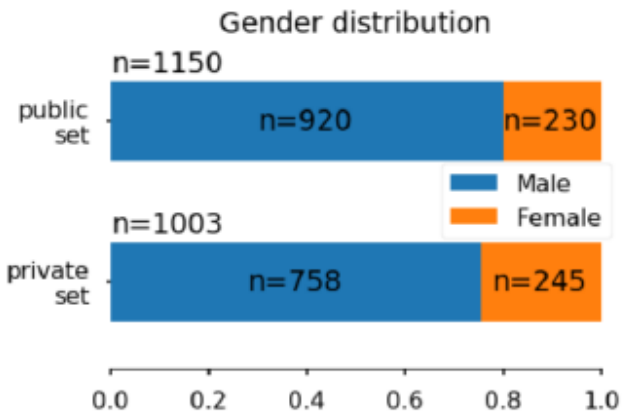

IMPAC
IMaging-PsychiAtry Challenge: predicting autism
A data challenge on Autism Spectrum Disorder detection
Incentives and goals:
3000€ for the best prediction of autism status
Web-based:
Participants submit code
Competition open during 3 months

Blind assessment of biomarkers

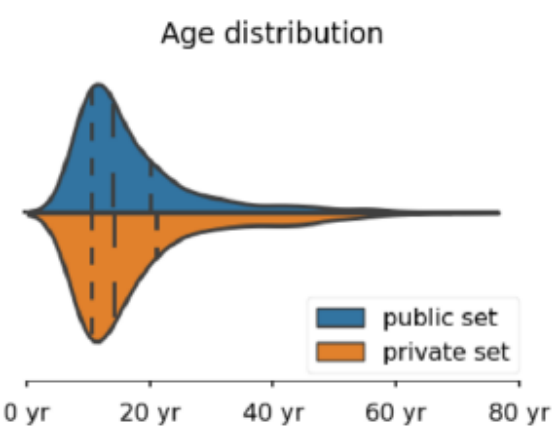
Patient vs Control distribution



Gender distribution



Age distribution

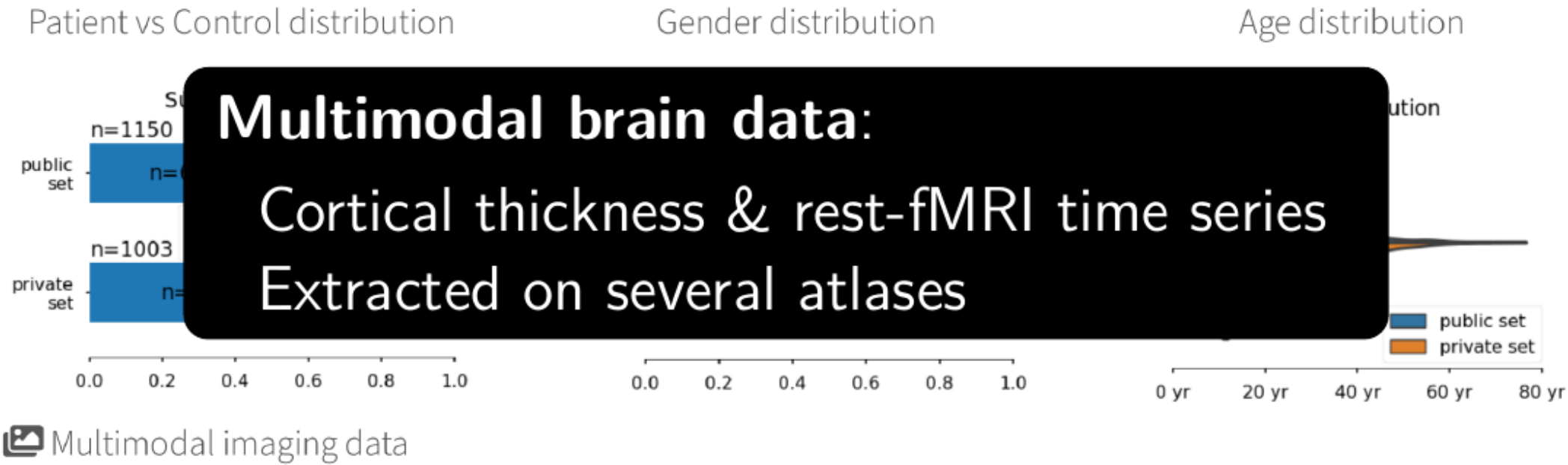


Hidden test set:

Participants never see the private set

Private-set prediction scores are published at the end

Blind assessment of biomarkers



Multimodal imaging data

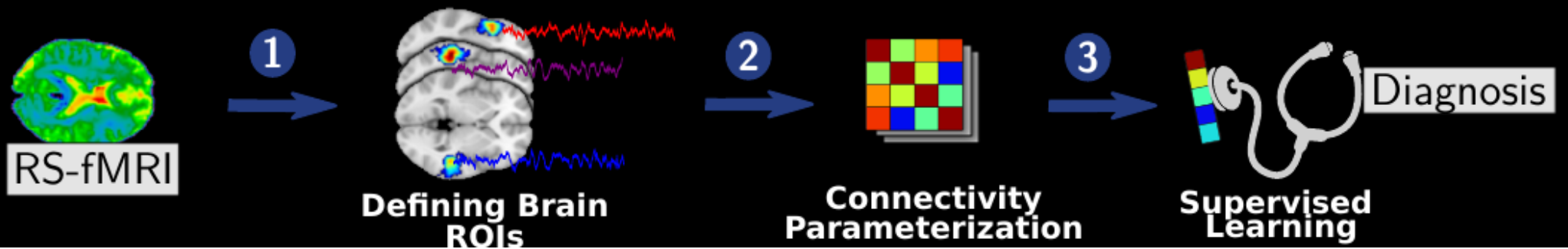
Structural MRI

Functional MRI

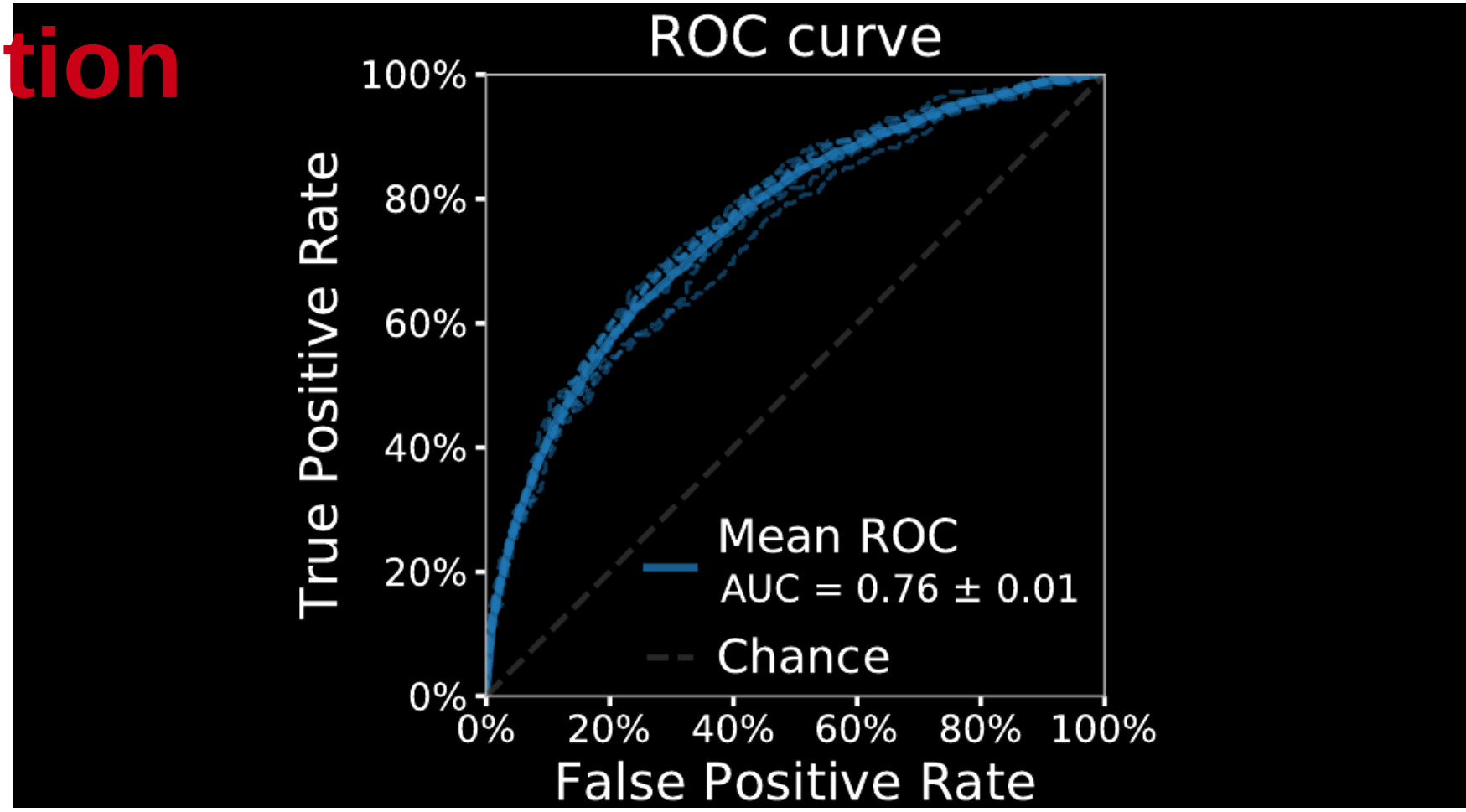
MRI biomarker extraction

Rest-fMRI biomarkers extraction

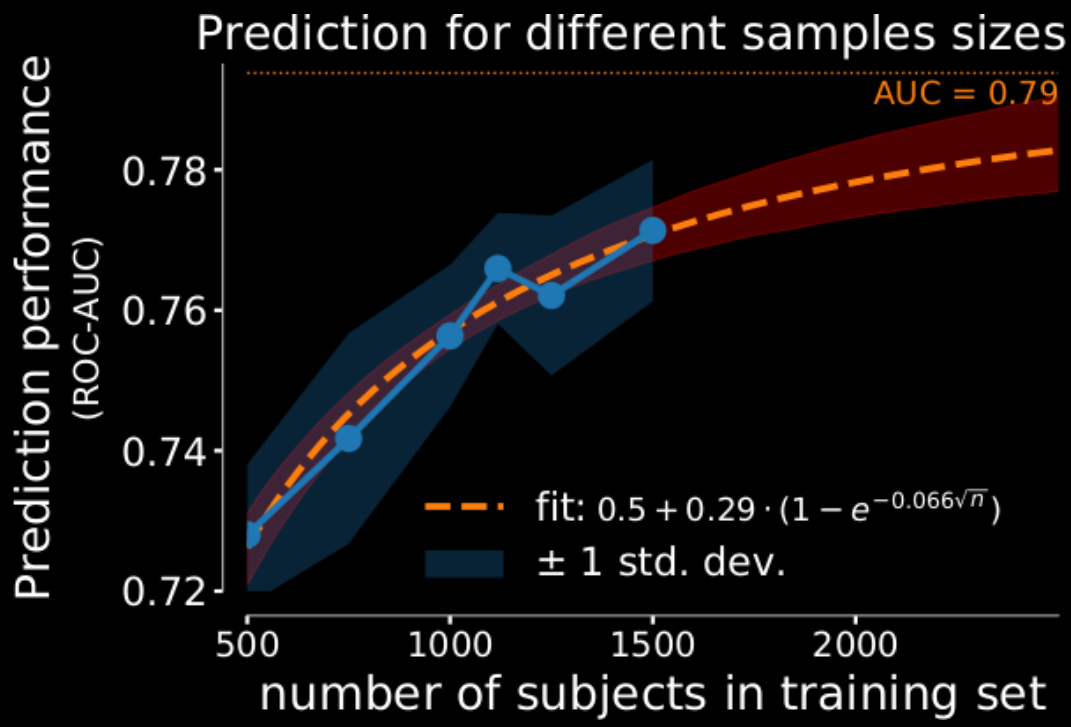
- Functional regions (extracted by dictionary learning)
- Tangent space to compare connectomes
- Linear model for supervised learning



Final prediction score



Use of open data: challenges



MRI biomarker extraction

An Autism challenge

- Trustworthy biomarkers (blind assesment)
- Rest fMRI is most useful
- More data is crucial
- Multi-site heterogeneity not a roadblock
- Overfit is easy

Some lessons from challenges

technical

- Simple is often better
- Prior knowledge
- Wisdom of crowds
- Multitask learning boosts performance.

social

- Incentivize participation
- avoid ‘winner-takes-all’ approach
- Detect unsportsmanlike behaviour
- Community building

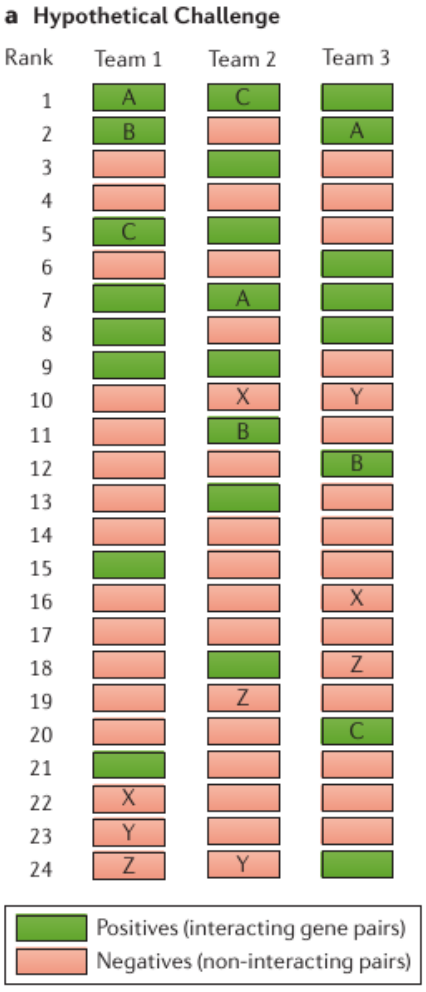
[Daez-Rodrigues et al. Nature 2016]

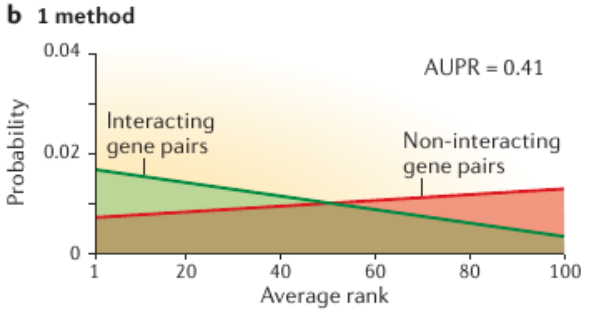
Some lessons from challenges

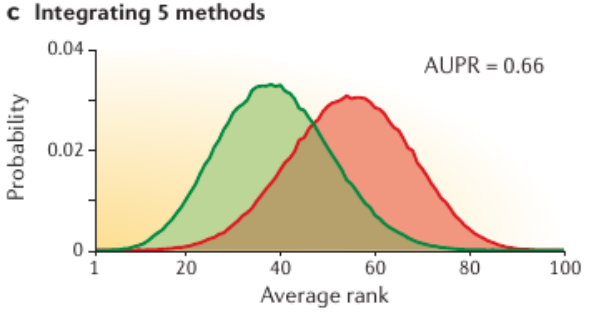
- Scoring strategies generally need to be made transparent
- Prevent data leakage and overfitting
- Do not provide any information about the test set
- multiple testing during scoring → less significance
- Dry runs to assess data quality

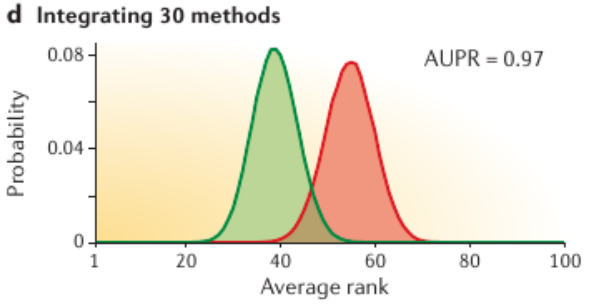
[Daez-Rodrigues et al. Nature 2016]

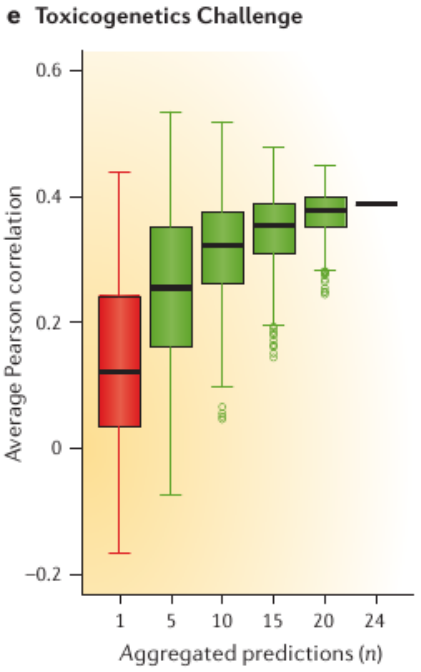
The “wisdom of crowds”











Cooperation leads to less variable, more accurate models

Conclusion

- The final goal is to do better science

Deep Learning for Hurricane Track Forecasting from Aligned Spatio-temporal Climate Datasets

Sophie Giffard-Roisin*
University of Colorado
Boulder, USA

Mo Yang*
Linear Accelerator Laboratory
Université Paris-Sud, CNRS

Guillaume Charpiat
Inria Saclay–Ile-de-France
LRI, Université Paris-Sud

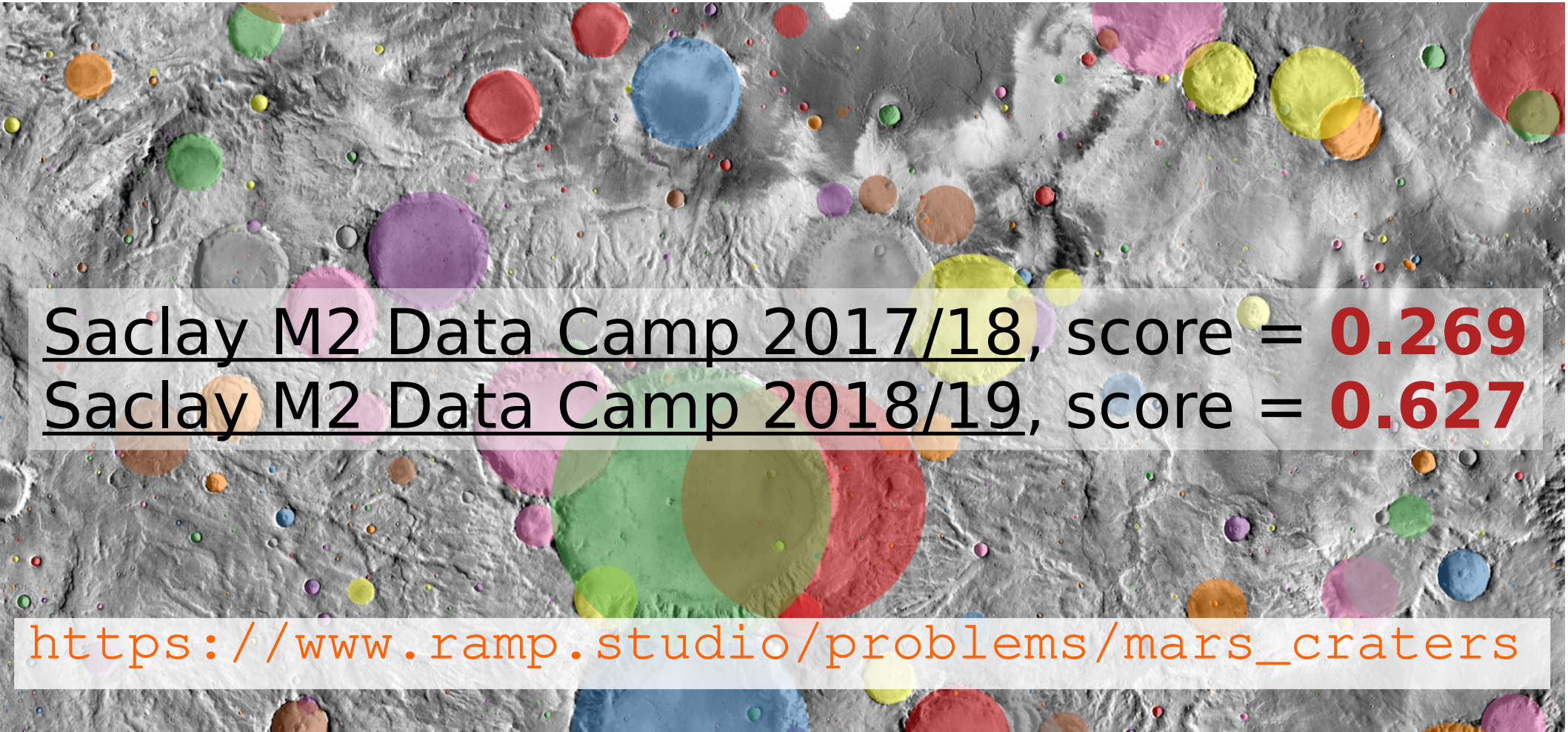
Balázs Kégl
Linear Accelerator Laboratory
Université Paris-Sud, CNRS

Claire Monteleoni
University of Colorado
Boulder, USA

Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach

Laetitia Minh Maï Le [a, b, 1], Balázs Kégl [c, g, 1], Alexandre Gramfort [c, e, f], Camille Marini [c, d], David Nguyen [a], Mehdi Cherti [c, g], Sana Tfaili [b], Ali Tfayli [b], Arlette Baillet-Guffroy [b], Patrice Prognon [a, b], Pierre Chaminade [b], Eric Caudron [a, b]

… with better scientists



Saclay M2 Data Camp 2017/18, score = **0.269**
Saclay M2 Data Camp 2018/19, score = **0.627**

https://www.ramp.studio/problems/mars_craters

Thanks

- Isabelle Guyon
- Paris Saclay CDS
  - Alexandre Gramfort
  - Sarah Cohen-Boulakia
  - David Rousseau
  - Balazs Kegl
  - Gael Varoquaux