

# Faire progresser la science des données par des challenges

Mercredi 2 Décembre 2020 - workshop virtuel



Prospective en Science des Données,  
Intelligence Artificielle et Biologie

ORGANISÉ PAR LES CSI DE L'INSB ET DE L'INS2I

Bertrand Thirion, [bertrand.thirion@inria.fr](mailto:bertrand.thirion@inria.fr)

# Faire progresser la science par des challenges

Mercredi 2 Décembre 2020 - workshop virtuel



Prospective en Science des Données,  
Intelligence Artificielle et Biologie

ORGANISÉ PAR LES CSI DE L'INSB ET DE L'INS2I

Bertrand Thirion, [bertrand.thirion@inria.fr](mailto: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

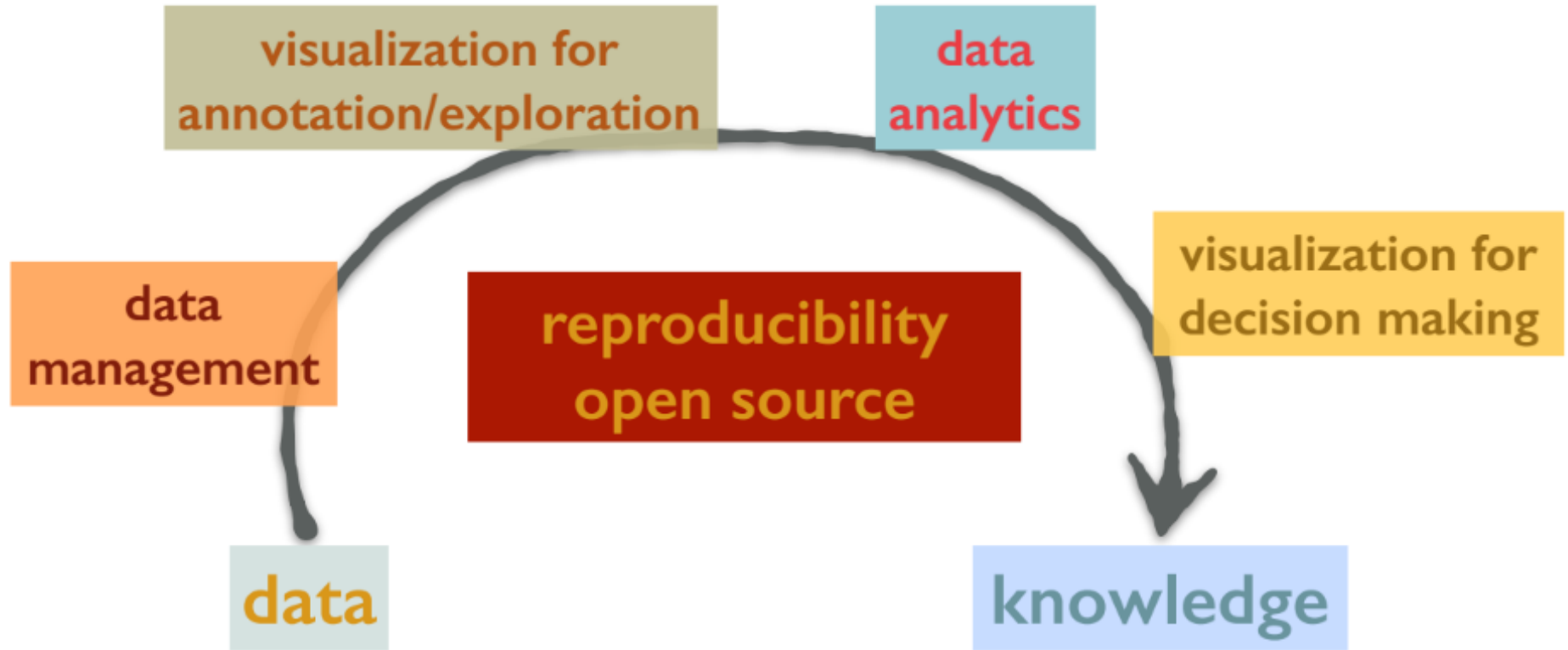


# Challenges as a crowdsourcing endeavour

- 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]

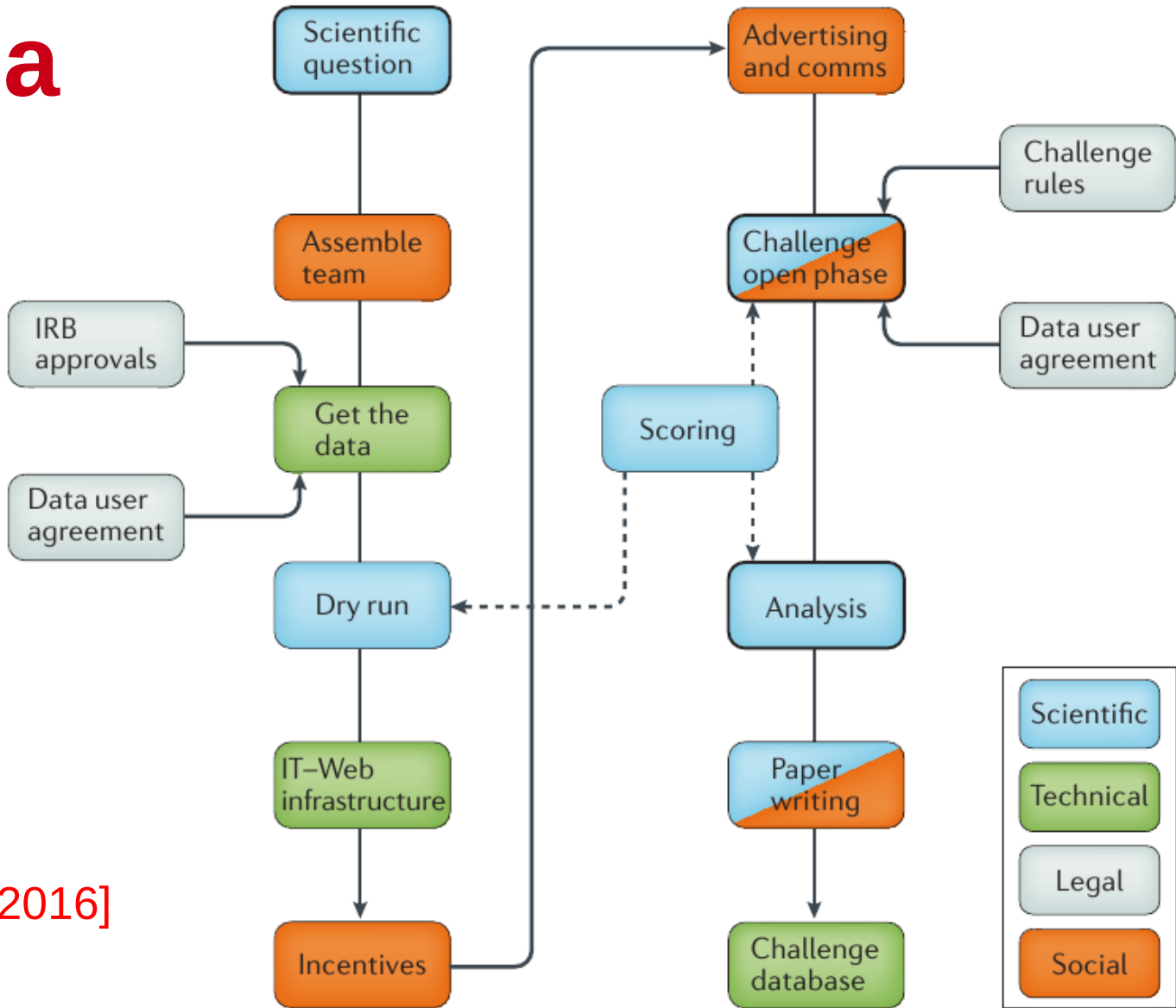
# It's all about inference





# Organizing a challenge

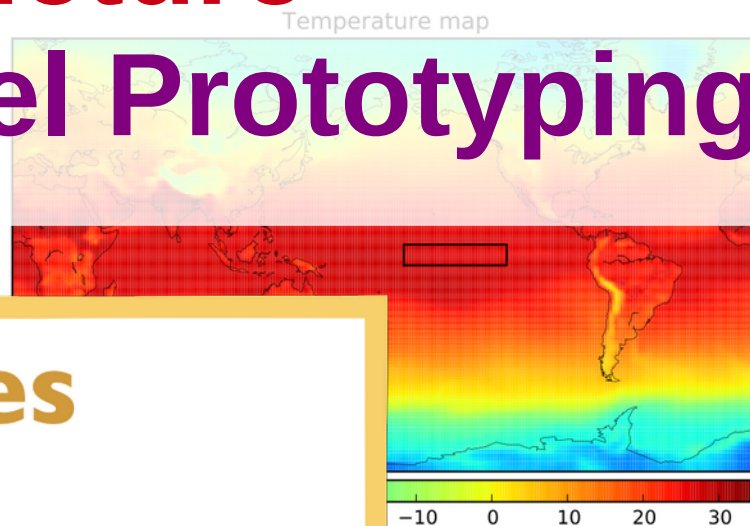
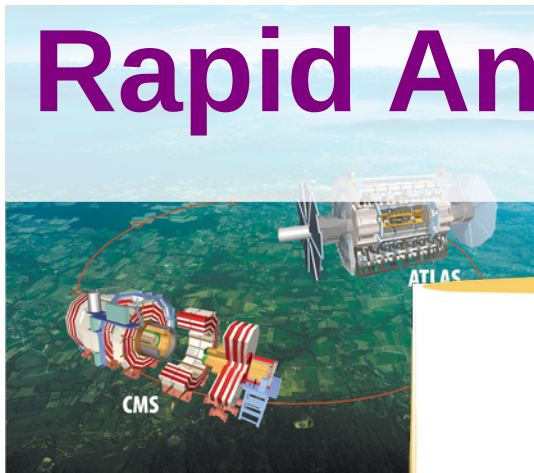
4 kinds of steps:  
scientific,  
technical,  
social, legal



[Daez-Rodrigues et al. Nature 2016]

# RAMP infrastructure

## Rapid Analytics & Model Prototyping



**16 challenges**  
**31 events**  
**996 users**  
**5685 predictive models**



# RAMP.STUDIO

github.com/ramp-kits

Hi Balázs!

## DATA CHALLENGE WITH CODE SUBMISSION

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

1 from sklearn.base import BaseEstimator
2 from sklearn.ensemble import RandomForestClassifier
3
4
5 class Classifier(BaseEstimator):
6     def __init__(self):
7         pass
8
9     def fit(self, X, y):
10        self.clf = RandomForestClassifier(
11            n_estimators=2, max_leaf_nodes=3, random_state=61)
12        self.clf.fit(X, y)
13
14    def predict(self, X):
15        return self.clf.predict(X)
16
17    def predict_proba(self, X):
18        return self.clf.predict_proba(X)
```

Upload your files!

File list

- classifier.py

Upload file

Choose File No file chosen

Upload

Leaderboard

Combined score: 0.899

Show 10 entries

team	submission	contributivity	historical contributivity	auc	accuracy	nll	train time
diego.souza	tuning_xgboost3	9	5	0.896	0.820	0.385	3074
ndeye-fatou.diop	kit_from_all	5	1	0.896	0.819	0.382	1167
diego.souza	tuning_xgboost2	4	2	0.896	0.819	0.385	4900
ndeye-fatou.diop	kit_from_all_clearer	3	0	0.896	0.819	0.384	1175
etienne.boursier	combine_features	2	7	0.896	0.820	0.383	2712
clement.vignac	boursier_improved_1	1	0	0.896	0.819	0.385	2499





# 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](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:**

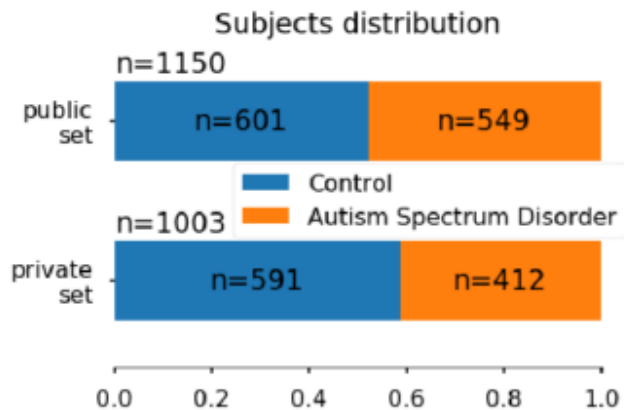
- 3000€ for the best prediction of autism status

### **Web-based:**

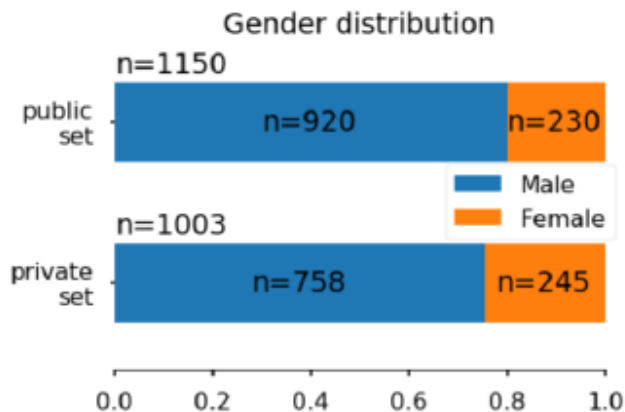
- Participants submit code
- Competition open during 3 months

# Blind assessment of biomarkers

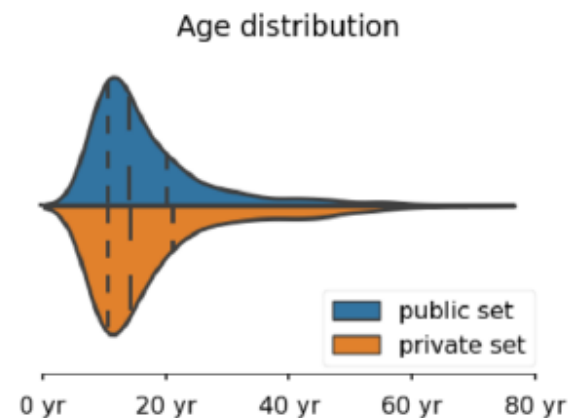
Patient vs Control distribution



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

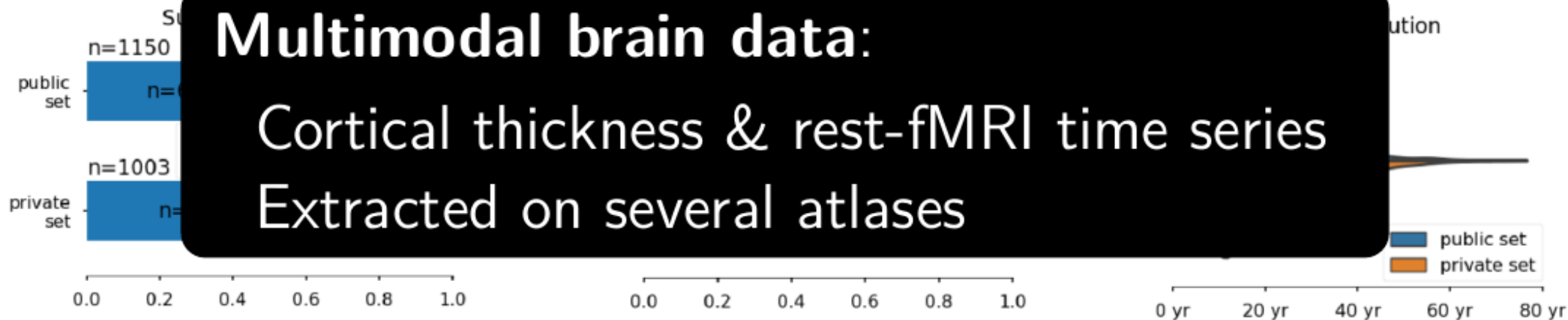
Patient vs Control distribution

Gender distribution

Age distribution

**Multimodal brain data:**

Cortical thickness & rest-fMRI time series  
Extracted on several atlases



 Multimodal imaging data

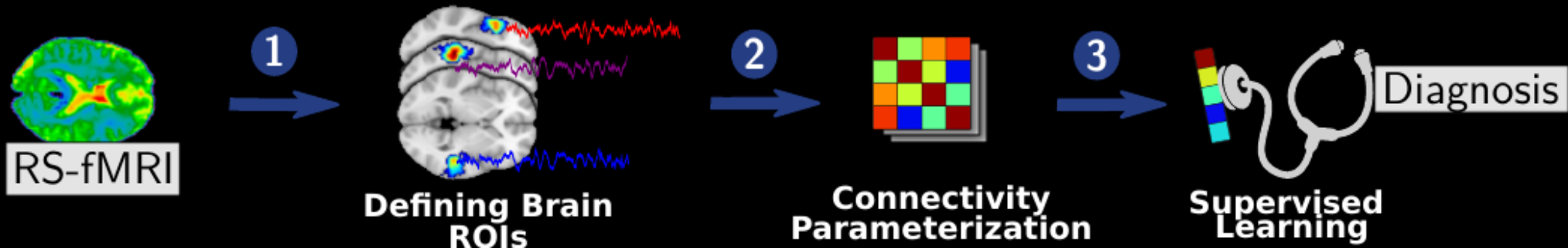
Structural MRI

Functional MRI

# MRI biomarker extraction

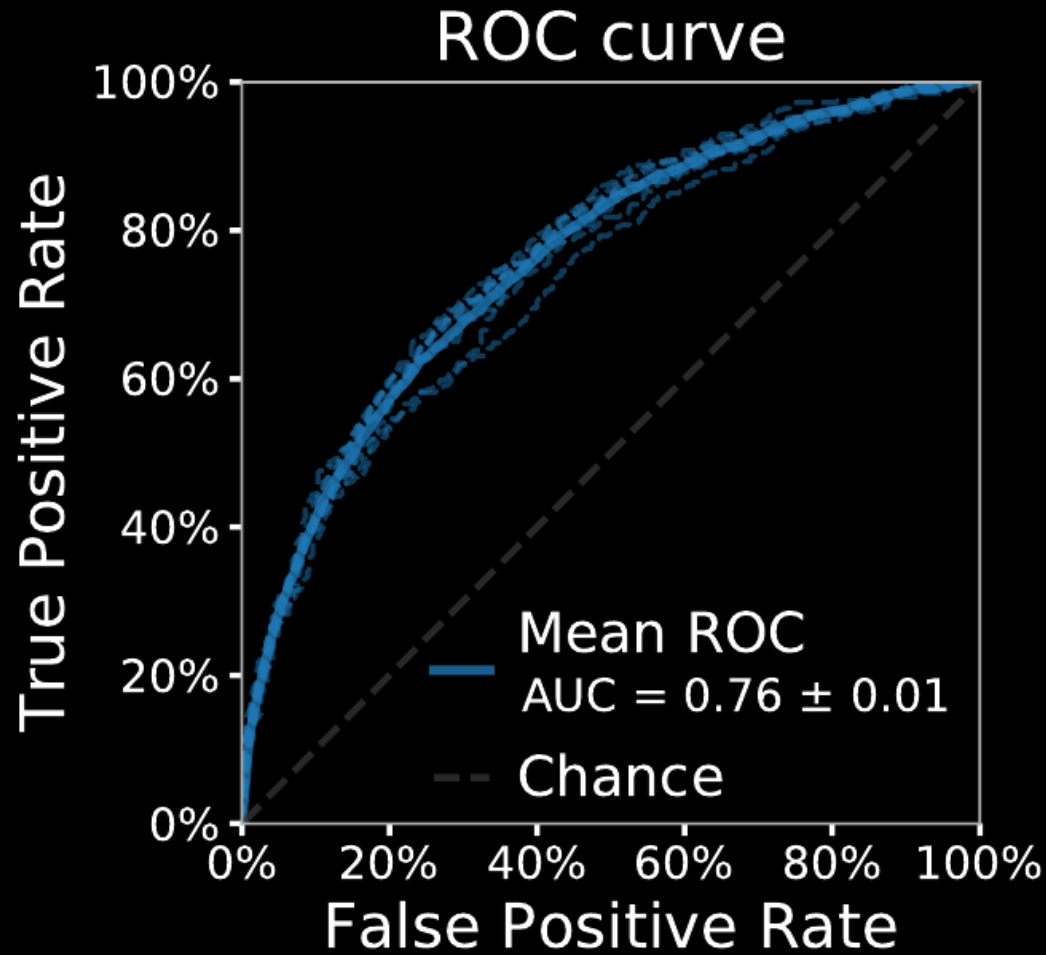
## Rest-fMRI biomarkers extraction

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

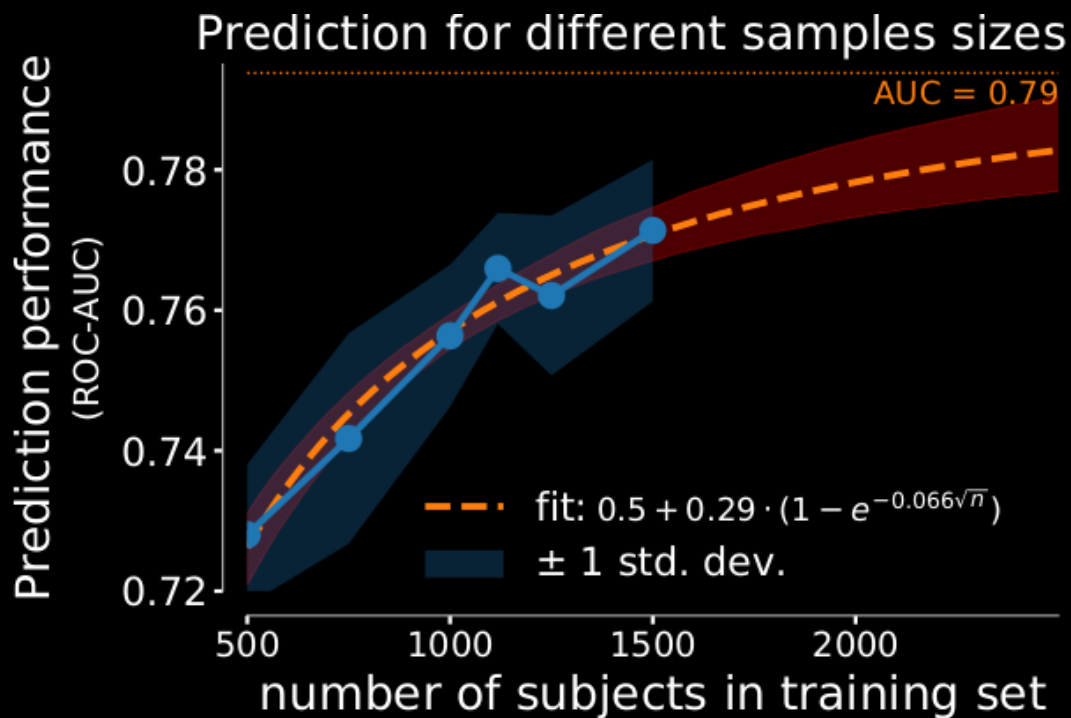




# Final prediction score



# 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

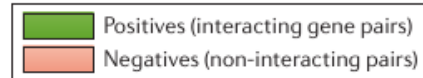
[Daez-Rodrigues et al. Nature 2016]



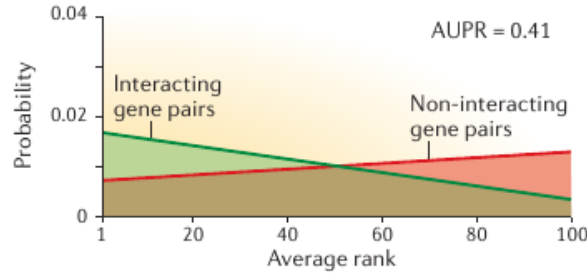
# The “wisdom of crowds”

**a Hypothetical Challenge**

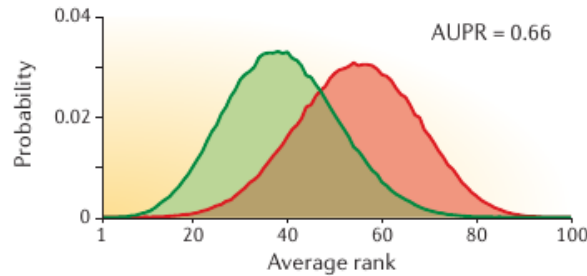
Rank	Team 1	Team 2	Team 3
1	A	C	
2	B		A
3			
4			
5	C		
6			
7		A	
8			
9			
10		X	Y
11		B	
12			B
13			
14			
15			
16			X
17			
18			Z
19		Z	
20			C
21			
22	X		
23	Y		
24	Z	Y	



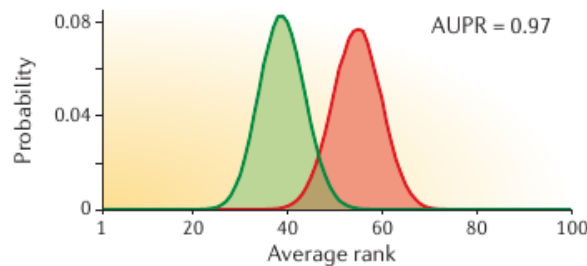
**b 1 method**



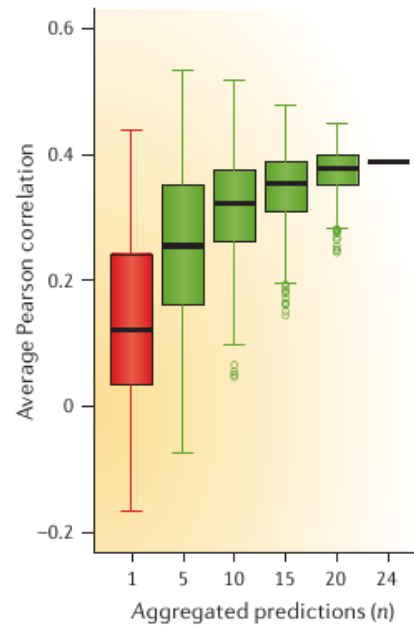
**c Integrating 5 methods**



**d Integrating 30 methods**



**e Toxicogenetics Challenge**



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 Mai Le <sup>a, b, 1</sup>, Balázs Kégl <sup>c, 8, 1</sup>, Alexandre Gramfort <sup>c, c, f</sup>, Camille Marini <sup>c, d</sup>, David Nguyen <sup>a</sup>, Mehdi Cherti <sup>c, 8</sup>, Sana Tfaili <sup>b</sup> ✉, Ali Tfayli <sup>b</sup>, Arlette Baillet-Guffroy <sup>b</sup>, Patrice Prognon <sup>a, b</sup>, Pierre Chaminade <sup>b</sup>, Eric Caudron <sup>a, b</sup>

# ... 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](https://www.ramp.studio/problems/mars_craters)

# Thanks

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

