



# The AutoML challenge on



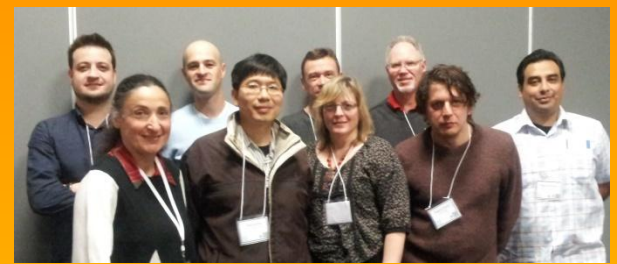
*Isabelle Guyon, Kristin Bennett, Gavin Cawley, Hugo Jair Escalante, Sergio Escalera, Tin Kam Ho, Nuria Macia, Bisakha Ray, Mehreen Saeed, Alexander Statnikov, Evelyne Viegas*

*[automl@ChaLearn.org](mailto:automl@ChaLearn.org)*





# Thanks



## *Hackathon team:*

Marc Boullé  
Lukasz Romaszco  
Sébastien Treger  
Emilia Vaajoensuu  
Philippe Vanderersch

## *Software development:*

Eric Carmichael  
Ivan Judson  
Christophe Poulain  
Percy Liang  
Arthur Pesah  
Xavier Baro Solé  
Lukasz Romaszco  
Michael Zyskowski

## *Advisors and beta testers:*

Kristin Bennett  
Marc Boullé

Cecile Germain  
Cecile Capponi  
Richard Caruana  
Gavin Cawley  
Gideon Dror  
Sergio Escalera  
Tin Kam Ho  
Balasz Kégl  
Hugo Larochelle  
Víctor Ponce López  
Nuria Macia  
Simon Mercer  
Florin Popescu  
Michèle Sebag  
Danny Silver,

## *Codalab management:*

Evelyne Viegas  
Percy Liang  
Erick Watson



## *ChaLearn board*

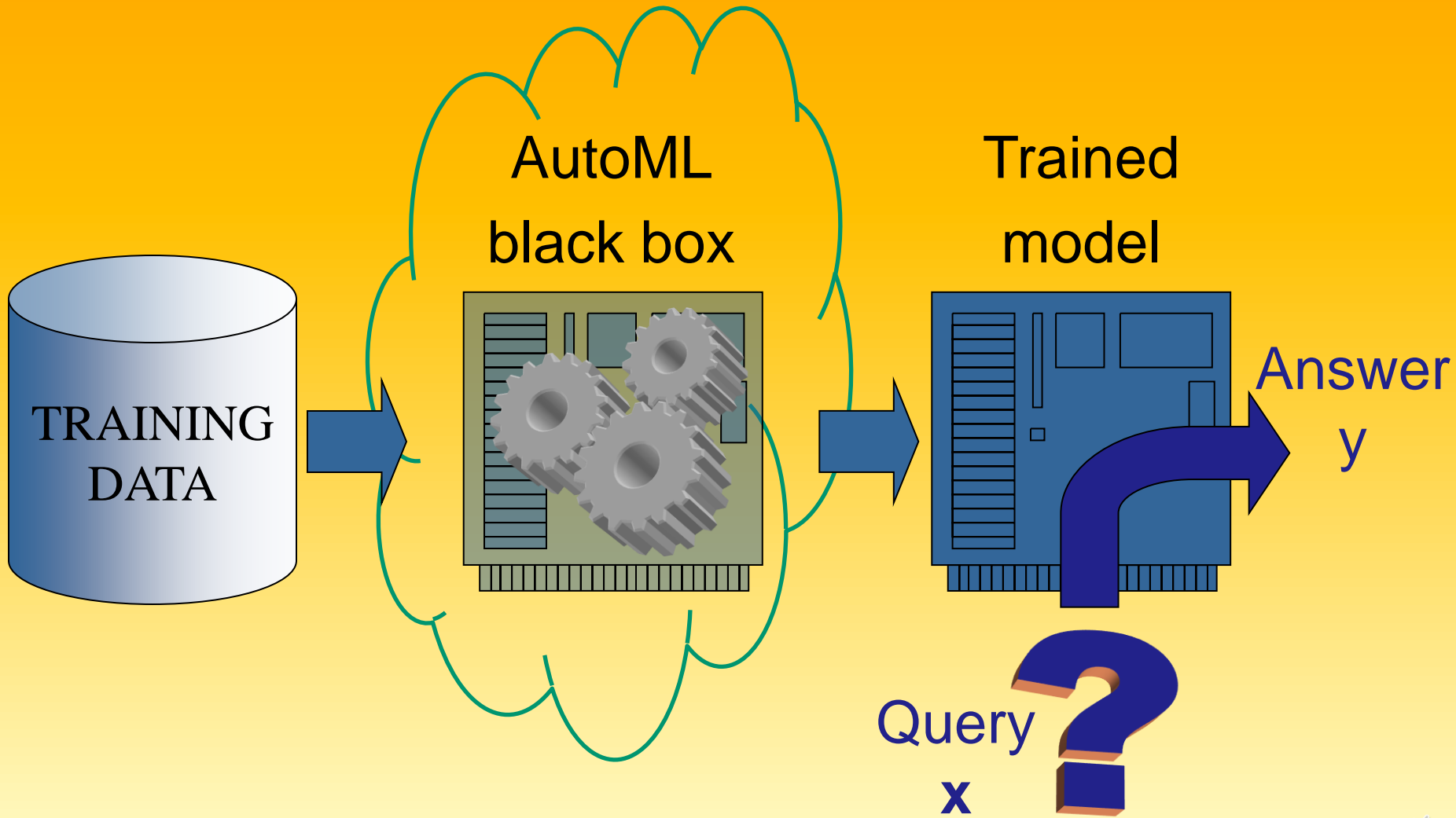
## *Data providers:*

Yindalon Aphinyanaphongs  
Olivier Chapelle  
Hugo Jair Escalante  
Sergio Escalera  
Zainab Iftikhar Malhi  
Vincent Lemaire  
Chih Jen Lin  
Meysam Madani  
Bisakha Ray  
Mehreen Saeed  
Alexander Statnikov  
Gustavo Stolovitzky  
H-J. Thiesen  
Ioannis Tsamardinos





# The dream

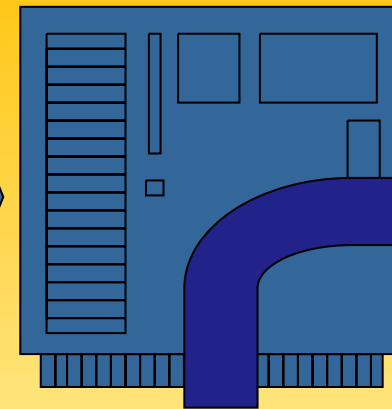




# *The REALITY*

Hyper-parameter  
tuning

Trained  
model

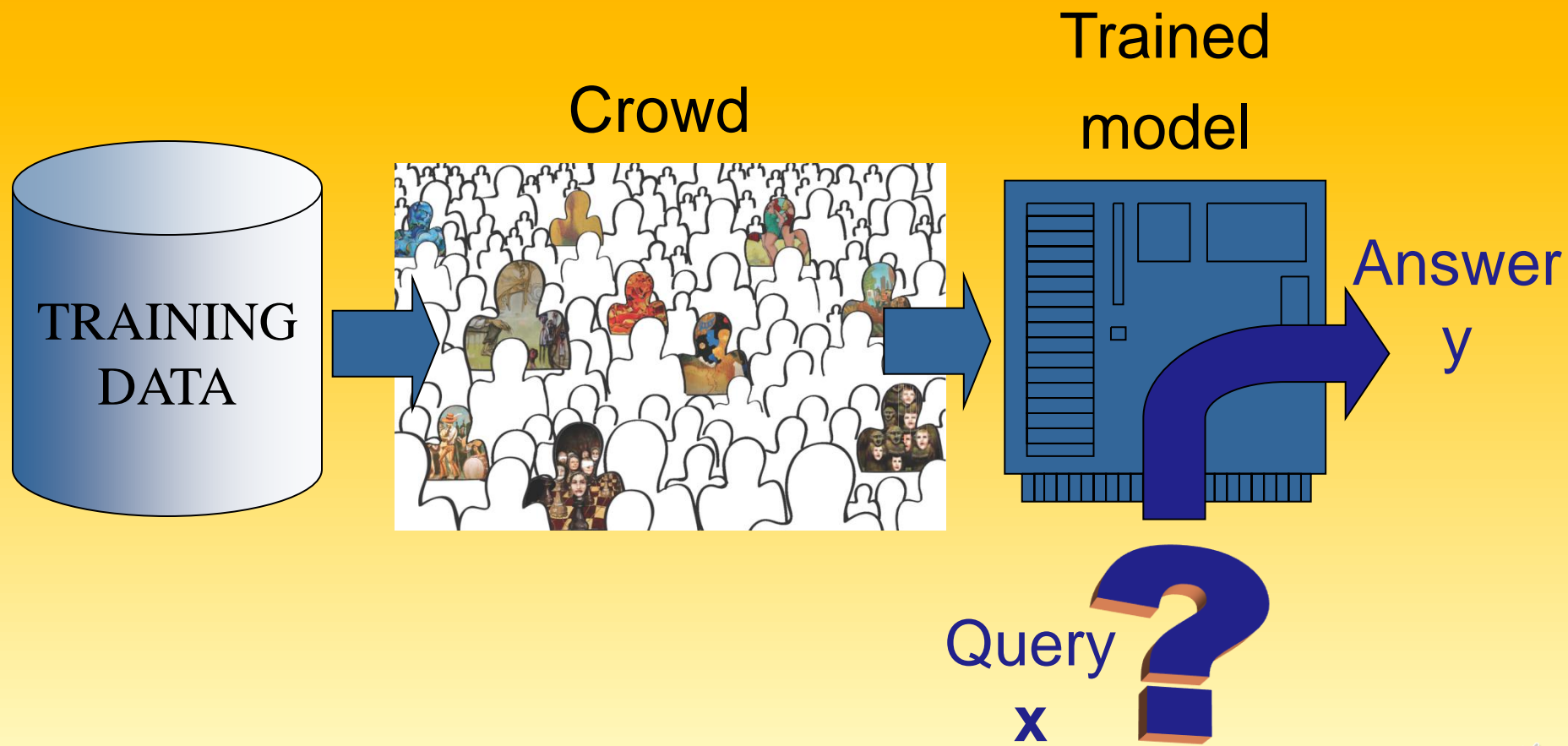


Answer  
 $y$

Query  
 $x$  ?

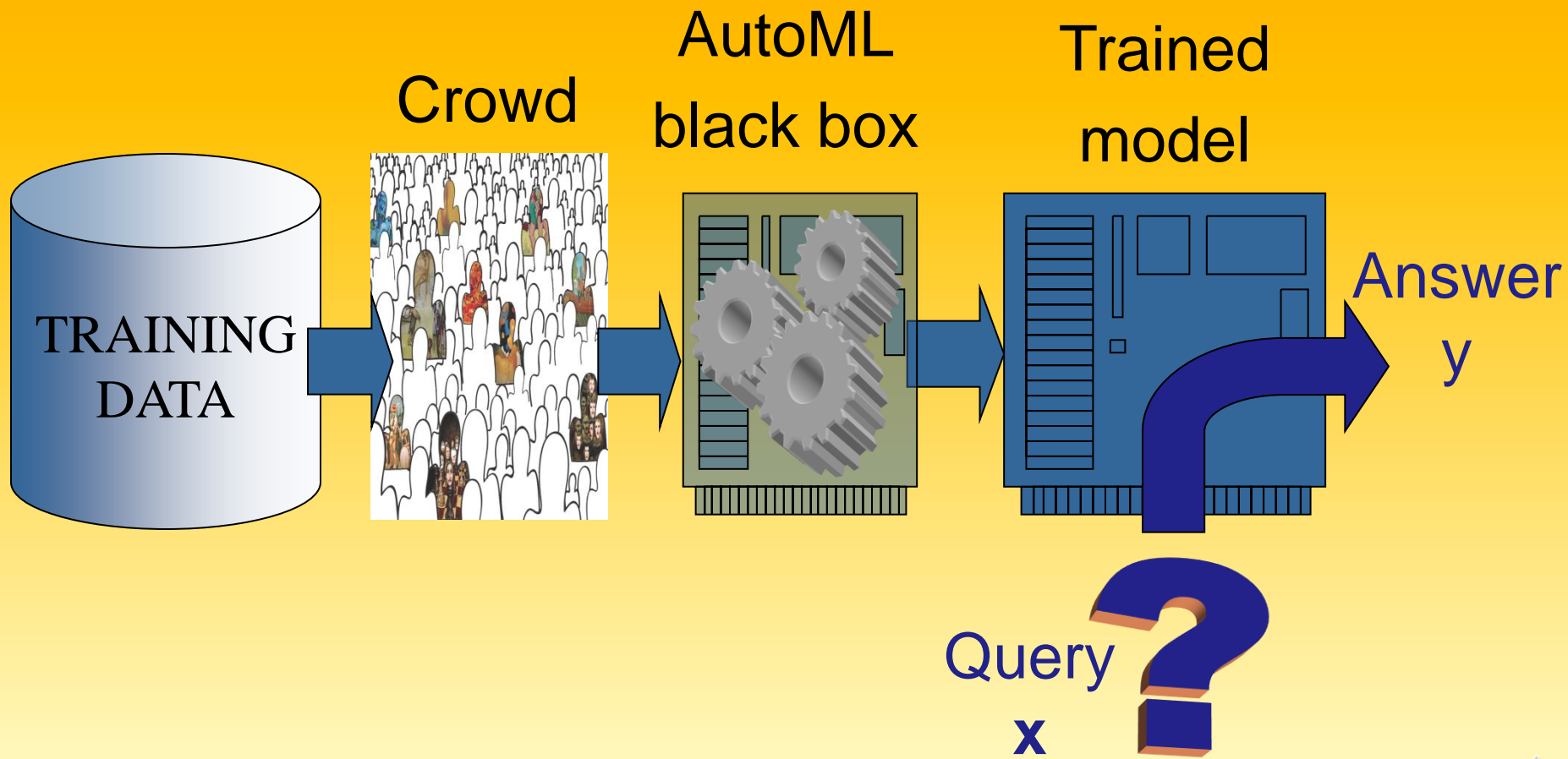


# ChaLearn ML challenges





# *AutoML challenge*





# Why would the crowd do that?



\$30000



Fame



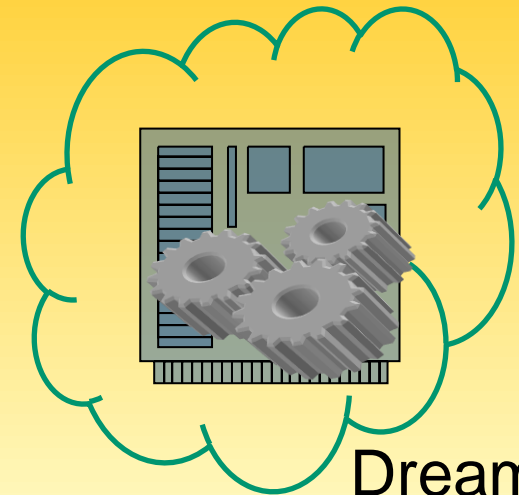
Learning



Fun



Workshop



Dream



# What tasks?

- **INPUT = I.I.D. data in feature representation, but:**
  - Sparse or full matrices.
  - Numerical/categorical/binary variables.
  - Missing values or not.
  - Noisy data or not.
  - Various proportions  $N_{train} / N_{feat}$ .
- **OUTPUT = one target, but:**
  - Binary (two-classes, balanced or not).
  - Categorical (multi-class: tens, hundreds of classes)
  - Multi-label.
  - Regression.
- **OBJECTIVE = miscellaneous loss functions.**
- **COMPUTATIONAL RESOURCES = fixed.**







# What data?

- **30 datasets spanning various domains:** pharmacology, medicine, marketing, ecology, text, image, video and speech processing.
- **6 rounds** (x 5 datasets).
- **Early phases:** partially “recycled data”.
- **Late phases:** novel data.
- **Medium hard tasks** (10-20% error) to separate algorithms best.
- **Large test sets** ( $\sim > 10,000$  examples) to ensure statistical significance.





# Round 0: sample data



US census 1994 (Kohavi and Becker)



MNIST HWR.  
LeCun, Cortes, and Burges (1998)

Num	Name	Task	Metric	Time	Cnum	Cbal	Sparse	Missing	Catvar	Irrvar	Pte	Pva	Ptr	N	Ptr/N
1	adult	multilabel	f1_metric	300	5	1	0.16	0.011	1	0.5	9768	4884	34190	24	1424.58
2	cadata	regression	r2_metric	200	0	NaN	0	0	0	0.5	10640	5000	5000	16	312.5
3	digits	multiclass	bac_metric	300	10	1	0.42	0	0	0.5	35000	20000	15000	1568	9.57
4	florothea	binary	auc_metric	100	2	0.46	0.99	0	0	0.5	800	350	800	100000	0.01
5	newsgroups	multiclass	pac_metric	300	20	1	1	0	0	0	3755	1877	13142	61188	0.21



Pharmacology.  
DuPont Pharmaceuticals  
(2001)



House prices.  
Pace and Barry (1997)

20 NewsGroups.  
Lang and Mitchell  
(1997)





# *Next rounds*

1. **NOVICE:** Binary classification..
2. **INTERMEDIATE:** Multiclass classification.
3. **ADVANCED:** Multiclass and multilabel.
4. **EXPERT:** Classification and regression.
5. **MASTER:** All of the above.





# *What's exciting?*

- **Intellectually challenging:**
  - Completely autonomous learner
  - Beat the “no free lunch theorem”
  - Model selection
  - Meta learning
  - Two-level objectives
  - Any overall objective ( $R^2$ , ABS, BAC, AUC, F1, PAC)
  - Any time
- **Practically important:**
  - Improve cost effectiveness
  - Improve reliability
  - Reach out to more applications





# How?



- **AutoML:** Automatic code execution on Codalab platform.
- **Tweakathon:** Result or code submission.
- To earn prizes: code should be made open source.





# How?

**ROUND**

**PHASE**

**SUBMISSION / EVALUATION**

**n-1**

---

**n**

---

**n+1**





# How?

ROUND

n-1

*Blind test on fresh data*

AutoML [+]

Test set

n

n+1





# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]



Test set

n

n+1







# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]

Data  
release

n

n+1





# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]



Test set

Data release

n

Tweakathon

n+1





# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]

Test set

Data release

n

Tweakathon

Code

and/or

Results

Validation set

n+1





# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]



Test set

Data release

n

Tweakathon

Code

and/or

Results



Validation set



Final [+]



Test set

n+1





# How?

ROUND

PHASE

SUBMISSION / EVALUATION

n-1

Code

*Blind test on fresh data*

AutoML [+]



Test set

Data release

n

Tweakathon

Code

and/or

Results



Validation set



Test set

*Blind test on fresh data*

n+1

Final [+]

AutoML [+]



NEW Test set





# How?

- We provide sample submissions.
- We provide a starting kit in Python using the scikit-learn library.
- The platform accepts:
  - Python scripts.
  - Linux executables.
  - Java JRE executables.
- On-line resources: <http://automl.org/>,  
<http://www.e-lico.eu/>.



# *What should I work on?*

The datasets span a range of difficulties:

- **Different tasks:** regression, **binary classification**, multi-class classification, multi-label classification.
- **Class balance:** **Balanced** or unbalanced class proportions.
- **Sparsity:** **Full matrices** or sparse matrices.
- **Missing values:** Presence or **absence of missing values**.
- **Categorical variables:** Presence or **absence of categorical variables**.
- **Irrelevant variables:** Presence or absence of additional irrelevant variables (distractors).
- **Number  $P_{tr}$  of training examples:** Small or large number of training examples.
- **Number  $N$  of variables/features:** Small or large number of variables.
- **Aspect ratio  $P_{tr}/N$  of the training data matrix:**  $P_{tr} \gg N$ ,  $P_{tr} \sim N$  or  $P_{tr} \ll N$ .





# *Schedule*

## **2014**

- Start: December 2014.

----- NIPS 2014, Montreal (December) -----

## **2015**

- Phase 0 ends: Mid February. **[Done]**

- Phase 1 ends: Mid June. **[Done]**

----- ICML 2015, Lille (July) ← NOW

- Phase 2 ends: Mid September.

- Phase 3 ends: Mid November.

----- NIPS 2015, Montreal (December) -----

## **2016**

Phase 4/5 end: Mid January 2016.







# *Hackathon*

TODAY

**Saturday July 11**

**Lille Grand Palais (LGP) conference center**

**Room Rotterdam 1 and 2, 7 pm to 11 pm**

PROGRAMME

**7 pm: Reception and award ceremony.**

**7:45 pm: Welcome. Emilia Vaajoensuu and Lukasz Romaszko.**

**8 pm to 11 pm: Hacking by teams (with expert coaches).**

**<http://automl.chalearn.org/hackathon-icml>**

***(complete prerequisites!)***

