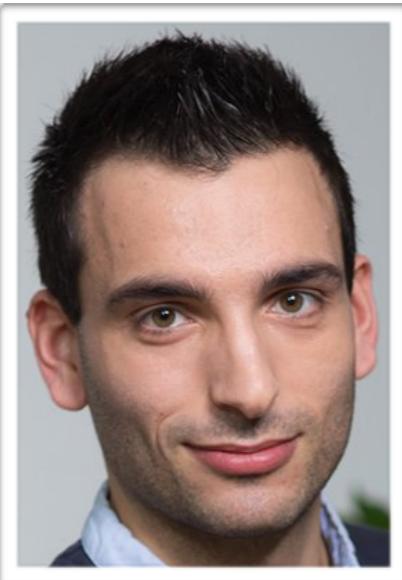


# Introduction to LIME

A further step towards ML model interpretability

# Who are we?



**Florent Pajot**

@FlorentPajot



**Samia Drappeau**

@samiadrappeau

Data Scientists at

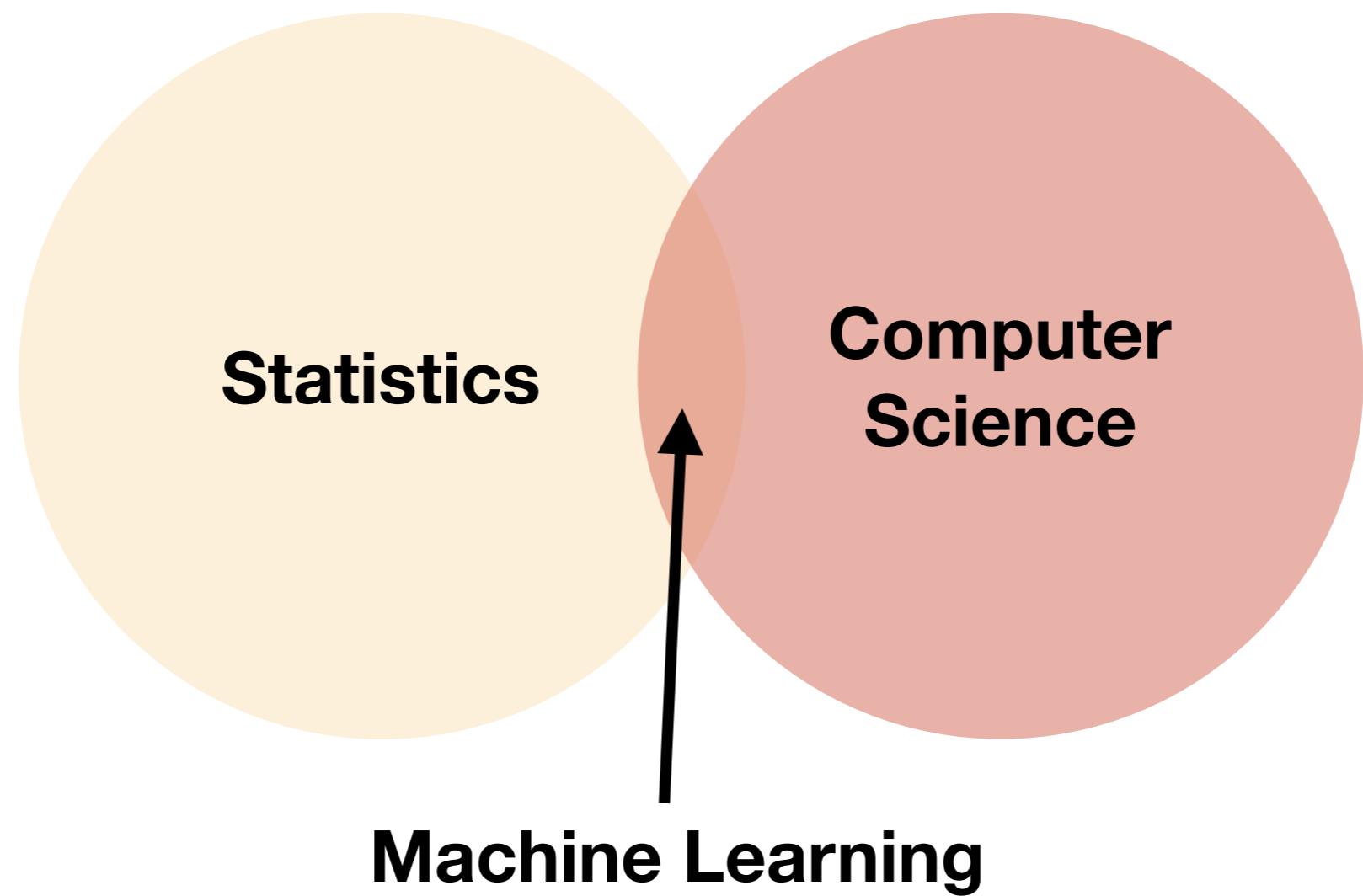


Elles bougent

S'engager  
pour plus  
de mixité  
professionnelle



# Machine Learning in two slides



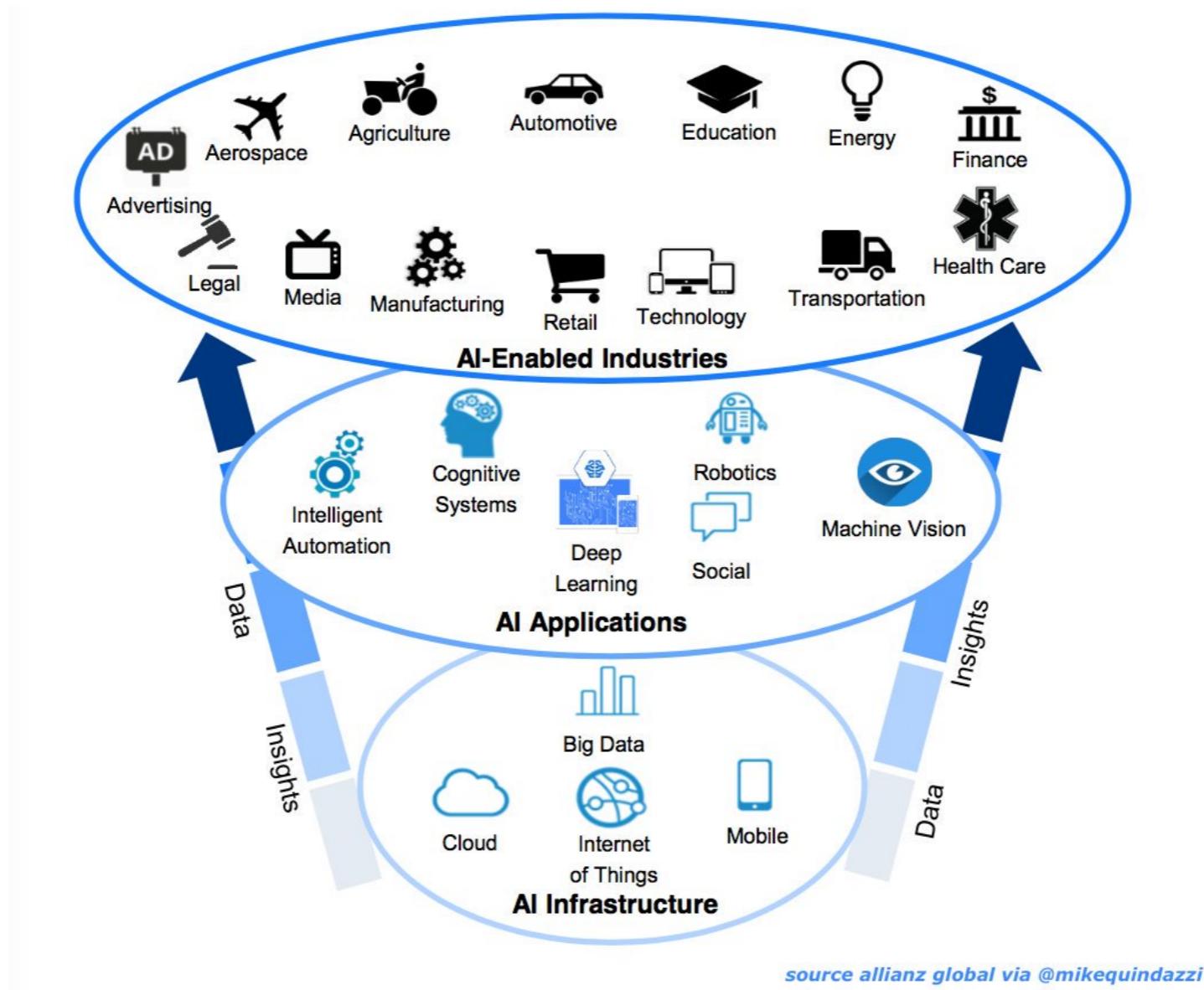
# Machine Learning in two slides

## Automation Spectrum

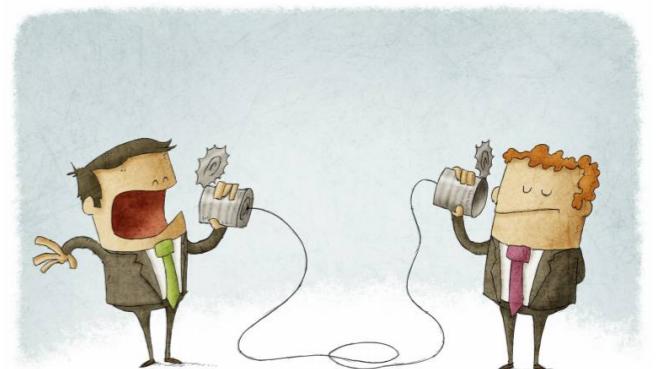


- Handcrafted code → Semi-supervised learning → Supervised learning → Reinforcement learning
- Machine Learning**
- if... elif. “*Computers [...] ability to learn without [...] explicit programming.*” networks  
~ Arthur Samuel (1959)
- › DON'T TOUCH code
  - › Magic constants
  - › p values
  - › Bayesian stats
  - › MCMC sampling
  - › SVM
  - › Random forests
  - › Recurrent net
  - › Convolutional net

# Why do we need interpretability?



# Why do we need interpretability?



# General Data Protection Regulation

- EU Data Protection Law // Regulation 2016/679 – Directive 95/46/EC
- Takes effect in May 2018
- Concerned for the ML community – *Article 22: Automated individual decision-making, including profiling.*

# General Data Protection Regulation

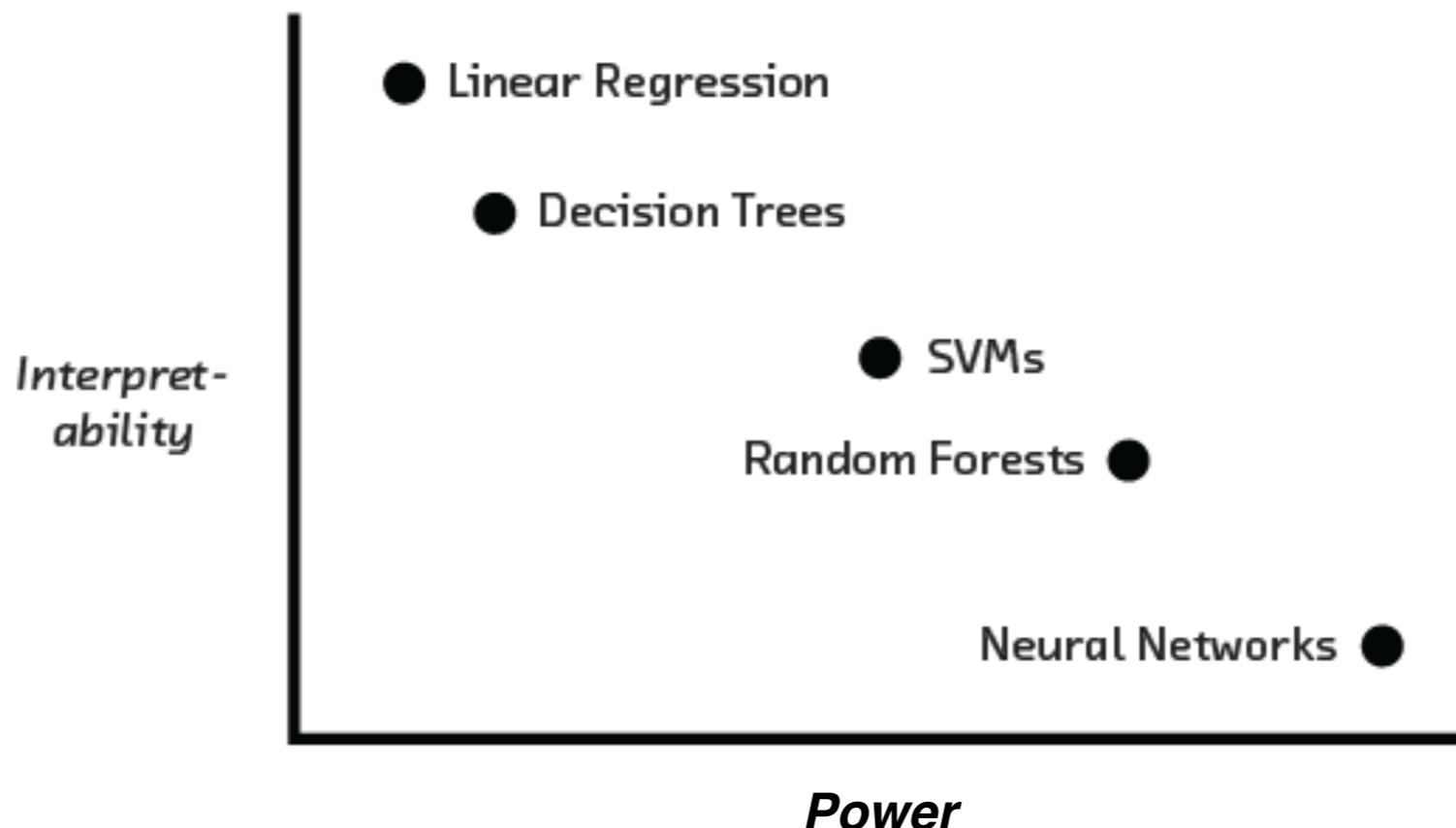
## Article 22

- Automated decision are contestable
  - ▶ what data was used?
  - ▶ need to explain a decision-making to EU citizen
  - ▶ non-discrimination
  - ▶ Up to 4% of the world gross revenues as fees

Art. 22 GDPR  
**Automated individual decision-making, including profiling**

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
  - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
  - (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(2)(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

# Interpretability vs Power



# The two approaches

1. Use simple models that are interpretable
2. Use complex models and try to explain/interpret them

# The two approaches

1. Use simple models that are interpretable
2. Use complex models and try to explain/interpret them

# Local Interpretable Model-agnostic Explanations

**"Why Should I Trust You?": Explaining the Predictions of Any Classifier (Feb 2016)**

**Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin**

As of September 26, 2017

**marcotcr / lime** First release March 2016

Code Issues 16 Pull requests 0 Projects 0 Wiki Insights ▾

263 commits 5 branches 10 releases 18 contributors BSD-2-Clause

Branch: master ▾ New pull request Create new file Upload files Find file Clone or download ▾

marcotcr removed legacy stuff Latest commit 73f0313 5 days ago

# Local Interpretable Model-agnostic Explanations

"Why should I trust you?"  
Explaining the predictions of any classifier



marcotcr removed legacy stuff

File	Description	Time Ago
benchmark	removing author tag	a year ago
doc	documentation fixes	2 months ago
lime	removed legacy stuff	3 days ago
.gitignore	CI: run unit tests & expand Travis configuration	6 months ago
.travis.yml	Merge branch 'v1-release' into master	5 months ago
CONTRIBUTING.md	Update CONTRIBUTING.md	a year ago
LICENSE	Initial commit	2 years ago
MANIFEST.in	Add LICENSE to MANIFEST.in	10 months ago
README.md	Readme	3 months ago
setup.cfg	flake arg in setup.cfg	5 months ago
setup.py	fix import in discretize	3 days ago

lime

This project is about explaining what machine learning classifiers (or models) are doing. At the moment, we support explaining individual predictions for text classifiers or classifiers that act on tables (numpy arrays of numerical or categorical data), with a package called lime (short for local interpretable model-agnostic explanations). Lime is based on the work presented in [this paper](#). Here is a link to the promo video:

[Promo Video](#)

Introduction to Local Interpretable Model-Agnostic Explanations (LIME)

A technique to explain the predictions of any machine learning classifier.

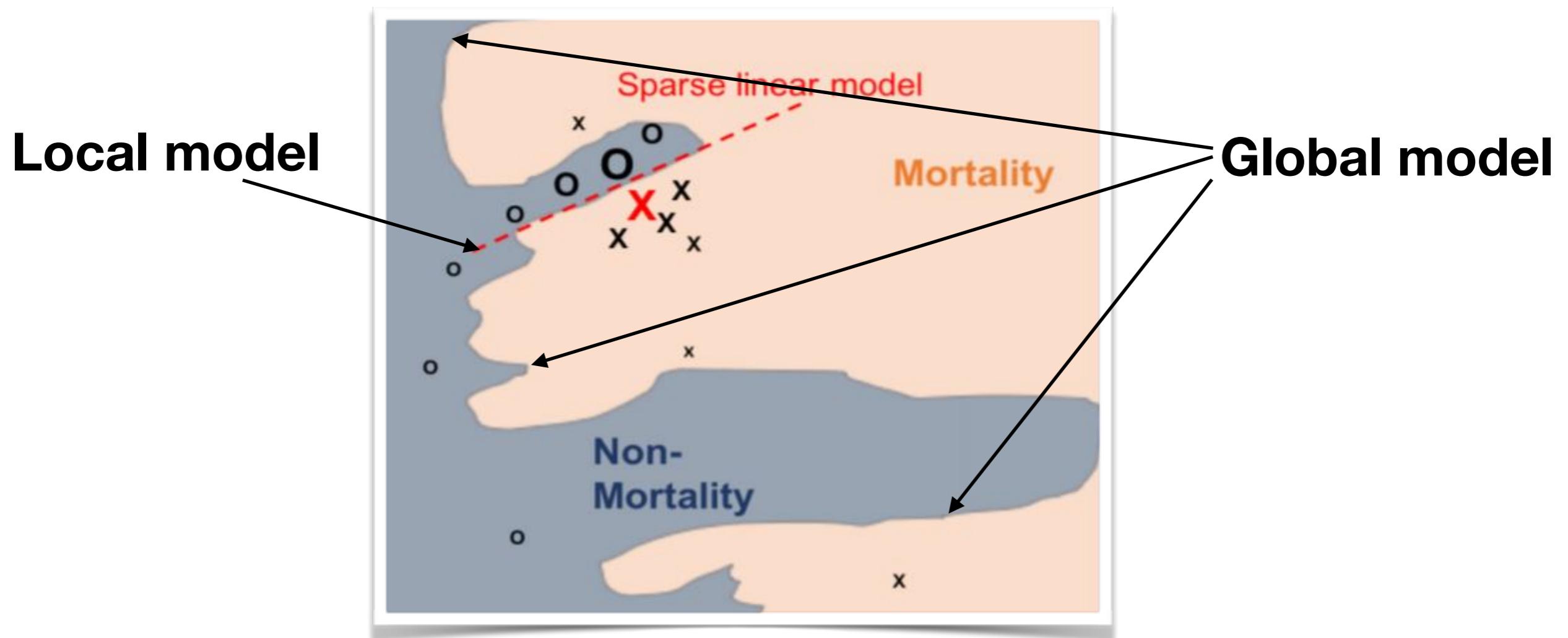
By Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. August 12, 2016

Check out the Data science and machine learning sessions at Strata Data in New York, September 25-28, 2017, for more on current trends and practical use cases in applied data science.

Machine learning is at the core of many recent advances in science and technology. With computers beating professionals in games like Go, many people have started asking if machines would also make for better drivers or even better doctors.

Happy predictions. (source: Jared Hersch on Flickr)

# LIME in one slide



# LIME in practice

- Go to the Jupiter notebook LIME tutorial
- [https://github.com/SamAstro/  
devfest toulouse 2017 LIME presentation/blob/master/  
LIME tutorial devfest-executed.ipynb](https://github.com/SamAstro/devfest_toulouse_2017_LIME_presentation/blob/master/LIME%20tutorial%20devfest-executed.ipynb)

# A field in expansion...

**Quora**  Search for questions, people, and topics

Quora uses cookies to improve your experience. Re

Working Model Machine Learning

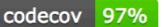
## What is the most interesting work in ML model interpretability in 2017?

3 Answers

 Surya Bhupatiraju, Google Brain Resident 2017  
Answered May 12 · Upvoted by Justin Rising, MSE in CS, PhD in Statistics

There's likely tons of work going on in this area, but one particularly cool recent paper that I enjoyed was: [\[1703.04730\] Understanding Black-box Predictions via Influence Functions ↗](#).

**ELI5**

[PyPI Version](#)   

ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions.

hi there, i am here looking for some help. my friend is a interic graphics software on pc. any suggestion on which software tc sophisticated software(the more features it has,the better)



## FairML: Auditing Black-Box Predictive Models

FairML is a python toolbox auditing the machine learning models for bias.



# Take away message

- Interpretability helps us understand the inner working of models
- However, one should never forget the three most important questions in ML:
  - Do I understand my data?
  - Do I understand the model and answers my ML algorithm is giving me?
  - Can I trust them?

# References

- <https://github.com/marcotcr/lime>
- <https://arxiv.org/pdf/1606.05386.pdf>
- <https://lime-ml.readthedocs.io/en/latest/>
- <https://medium.com/@thommash/local-interpretable-model-agnostic-explanations-lime-and-gdpr-9e3d66b64207>
- <https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning>
- <http://blog.fastforwardlabs.com/2017/09/01/LIME-for-couples.html>