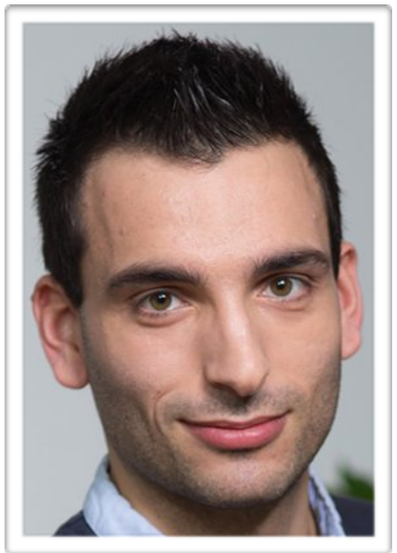
Introduction to LIME

A further step towards ML model interpretability

September 28, 2017 Florent Pajot — Samia Drappeau

DevFest Toulouse 2017

Who are we?



Florent Pajot

@FlorentPajot

Samia Drappeau

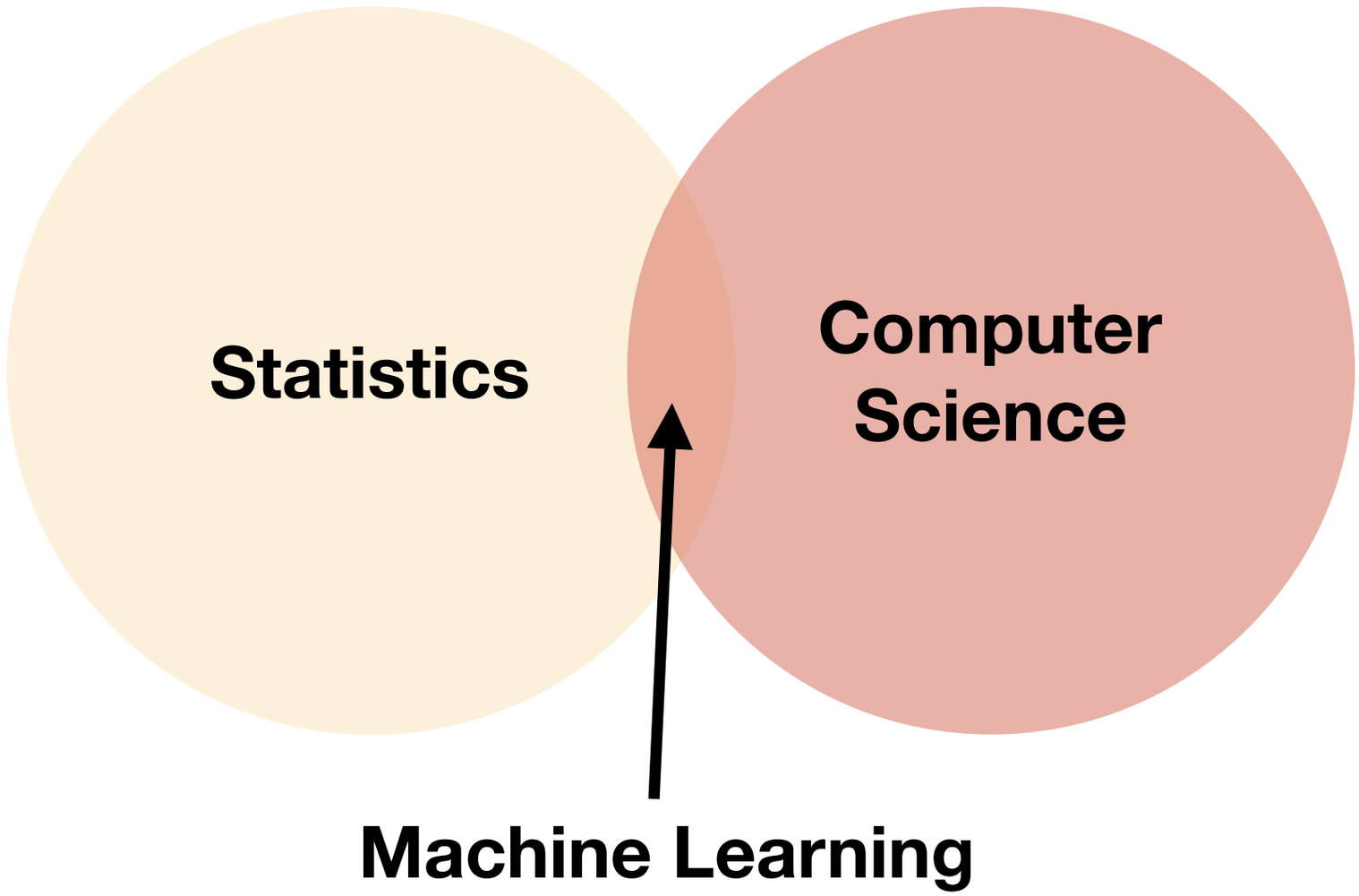
@samiadrappeau

Data Scientists at CONTINENTAL SINCE 1871

Elles bougent — S'engager pour plus de mixité professionnelle

TDS

Machine Learning in two slides



Machine Learning in two slides

Automation Spectrum

Handcraft… Rules — Statistics — Machine Learning — Deep learning

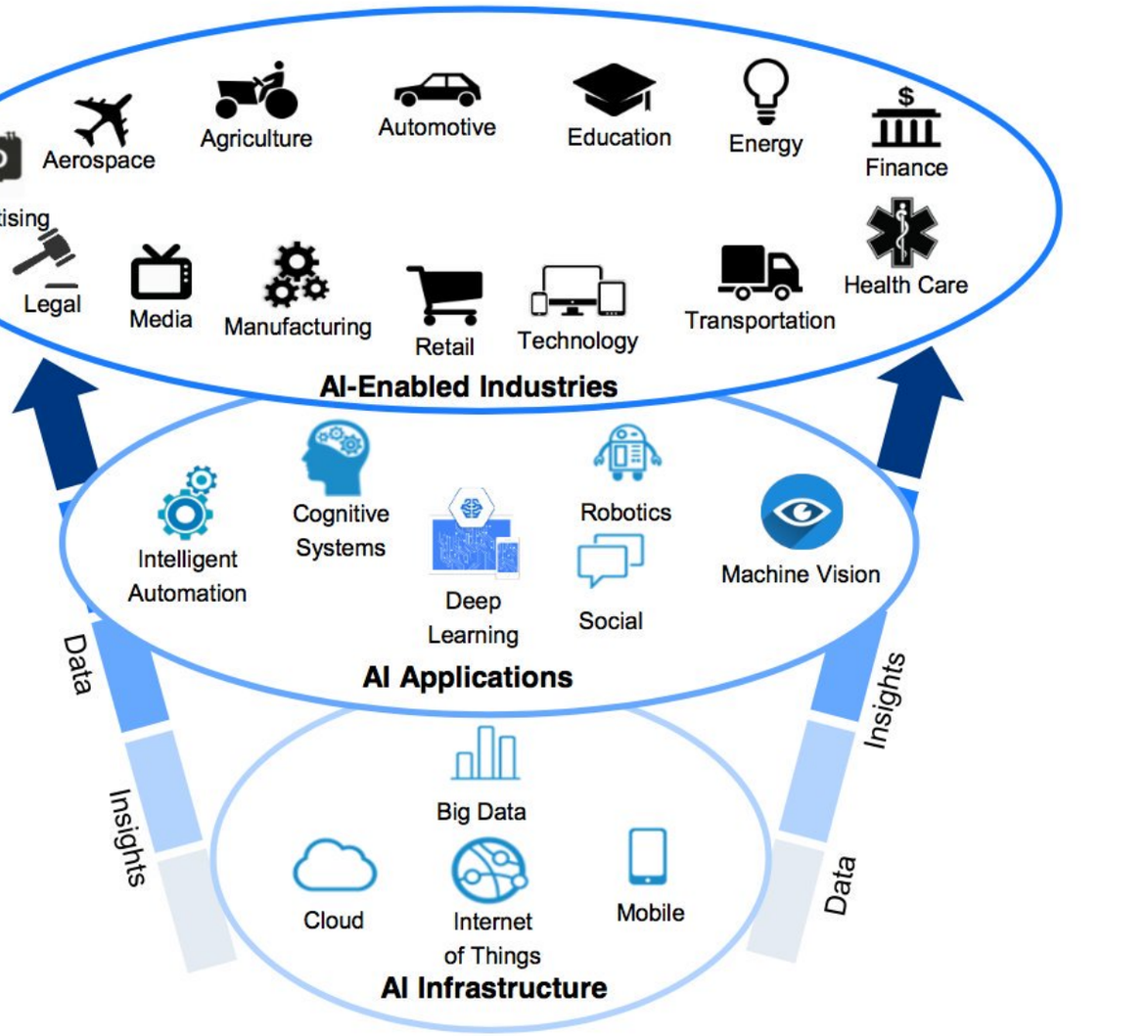
Machine Learning

“Computers […] ability to learn without […] explicit programming.”

~ Arthur Samuel (1959)

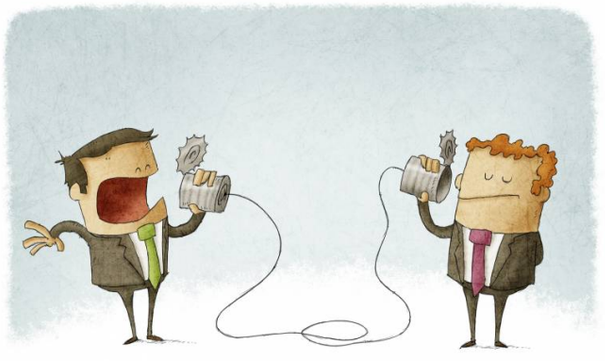
- if… elif… networks
- DON’T TOUCH code
- Magic constants
- p values
- Bayesian stats
- MCMC sampling
- SVM
- Random forests
- Convolutional net
- Recurrent net

Why do we need interpretability?



source allianz global via @mikequindazzi

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;

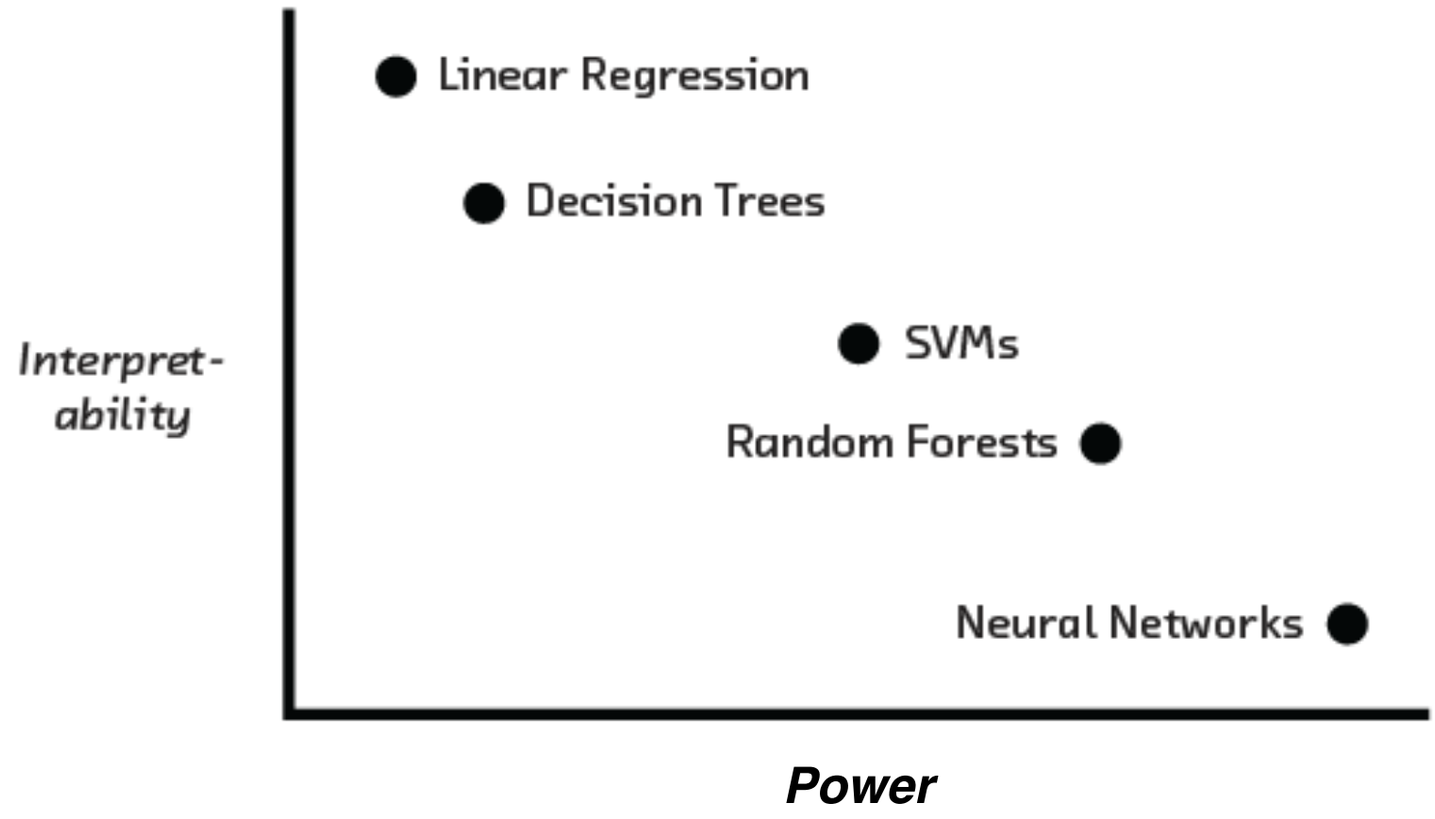
   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   (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

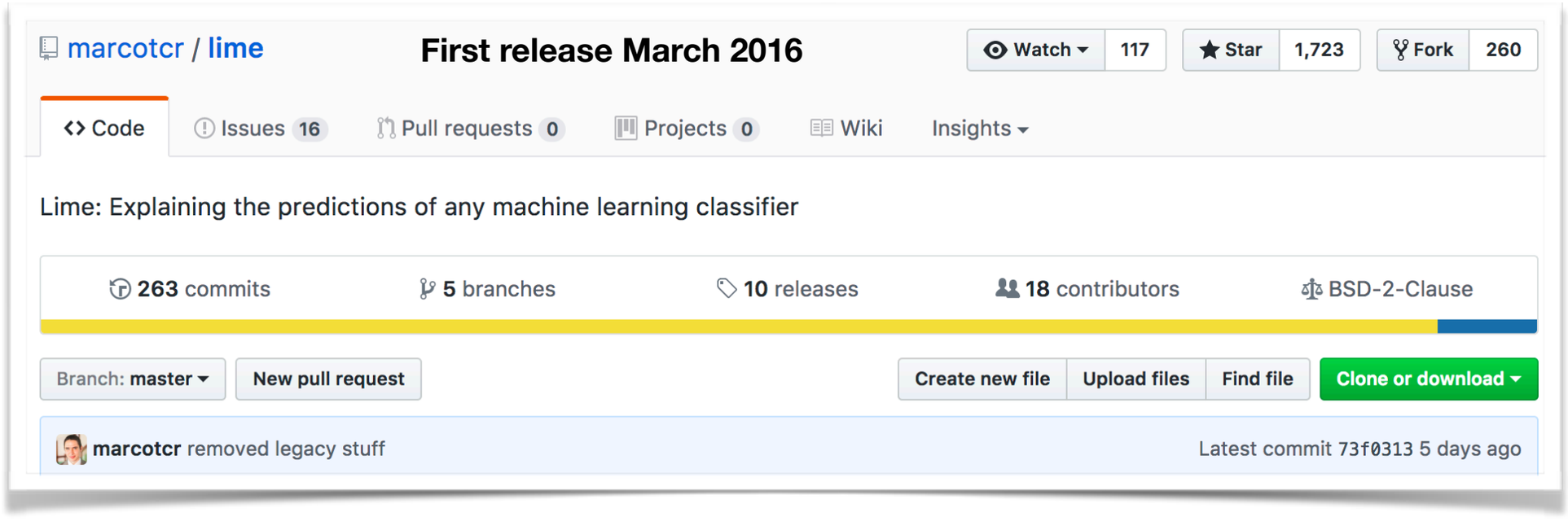
2. Use complex models and try to explain/interpret them

Local Interpretable Model-agnostic Explanations

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

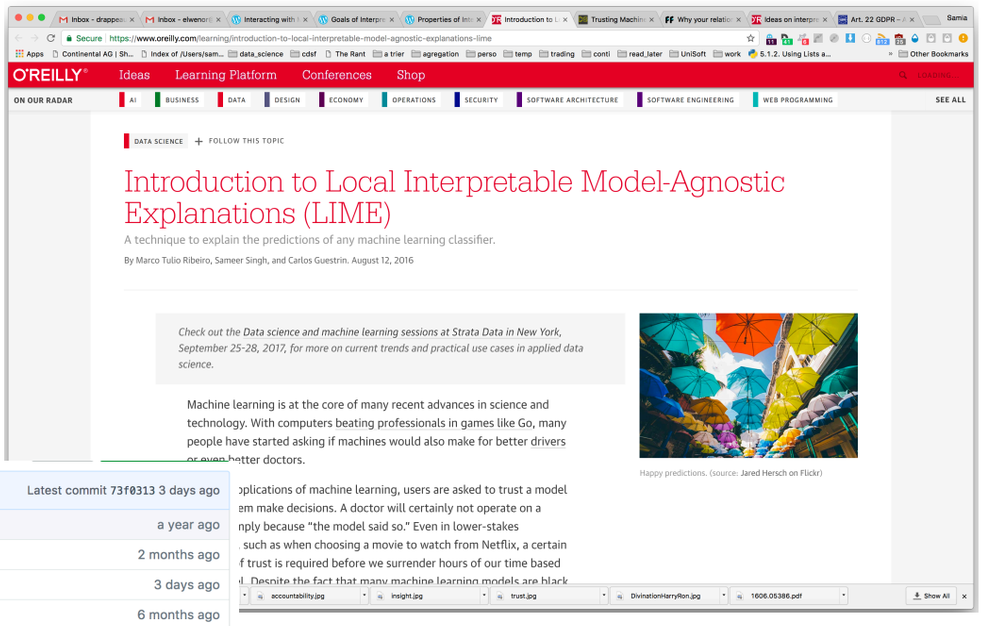
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

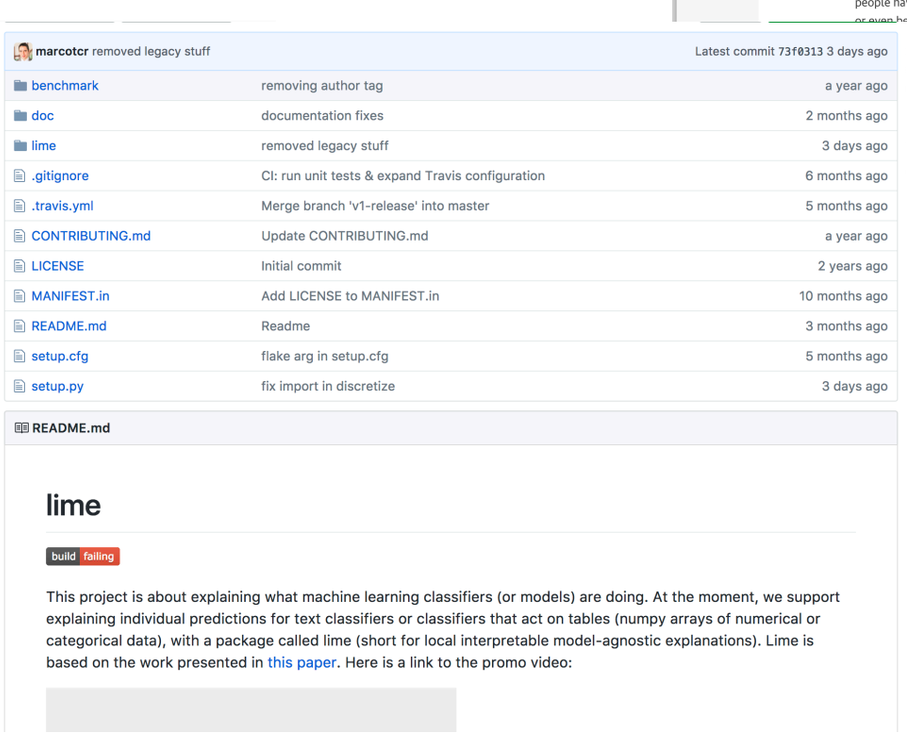
As of Septembre 26, 2017



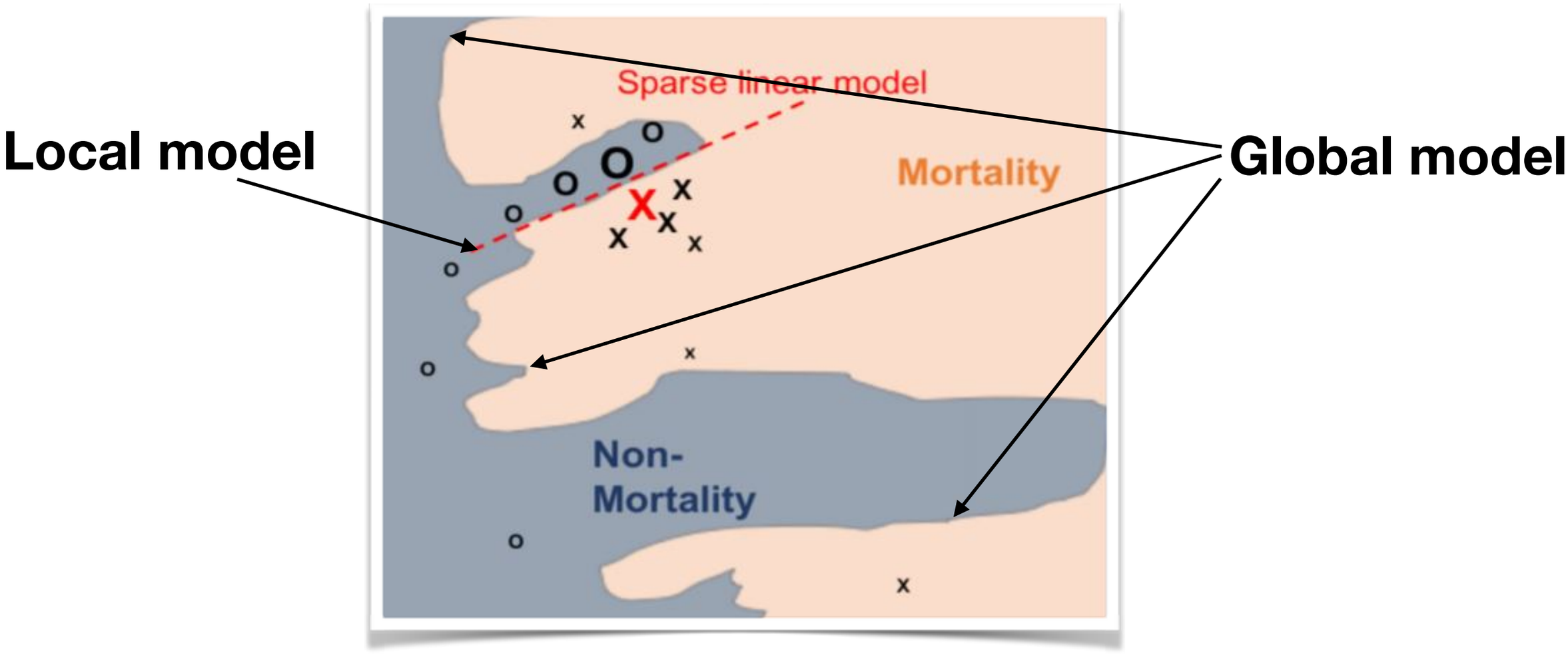
Local Interpretable Model-agnostic Explanations

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





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

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.

DATASCIENCE.COM
SKATER

ELI5

PyPI Version | build passing | codecov 97% | docs passing

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