Seldon: Ethics in AI

Explainability and Interpretability in Machine Learning

The Microsoft Reactor London June 2019

SELDON

Agenda

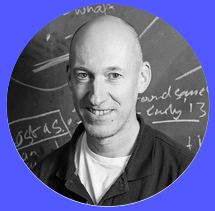
1. Seldon
2. Our journey
3. The promise & the pitfalls
4. Explanation
5. Anchors
6. Counterfactuals
7. Trust scores

We help people shape the future by bringing machine learning deployment to life.

Meet the core team.

SELDON

Alex Housley
CEO & Founder



Clive Cox PhD
CTO

Gurminder Sunner
VP Engineering

Ryan Dawson PhD
Cloud Engineer

Giovanni Vacanti PhD
Machine Learning Engineer

Inga Veidmane
Operations Manager

Andrew Turner
Sales Manager

Janis Klaise PhD
Data Scientist

Arnaud Van Looveren
Data Scientist

@seldon_io

hello@seldon.io

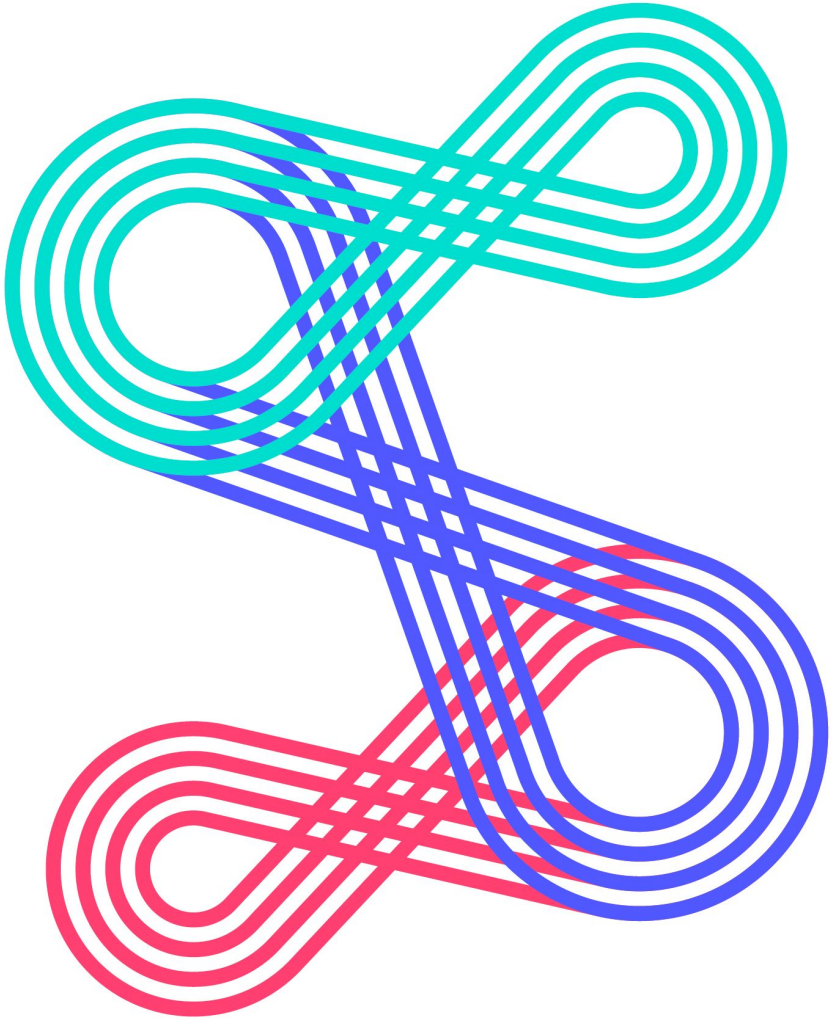
bit.ly/SeldonCoreSlack


seldon.io/careers

Janis Klaise

Architect of Alibi by Seldon
(Any question answered!)

@JKlaise



Lee Baker

Seldon Commercial Leader
(Dumbest man in the room @ Seldon)

@BakerLJ

Our journey

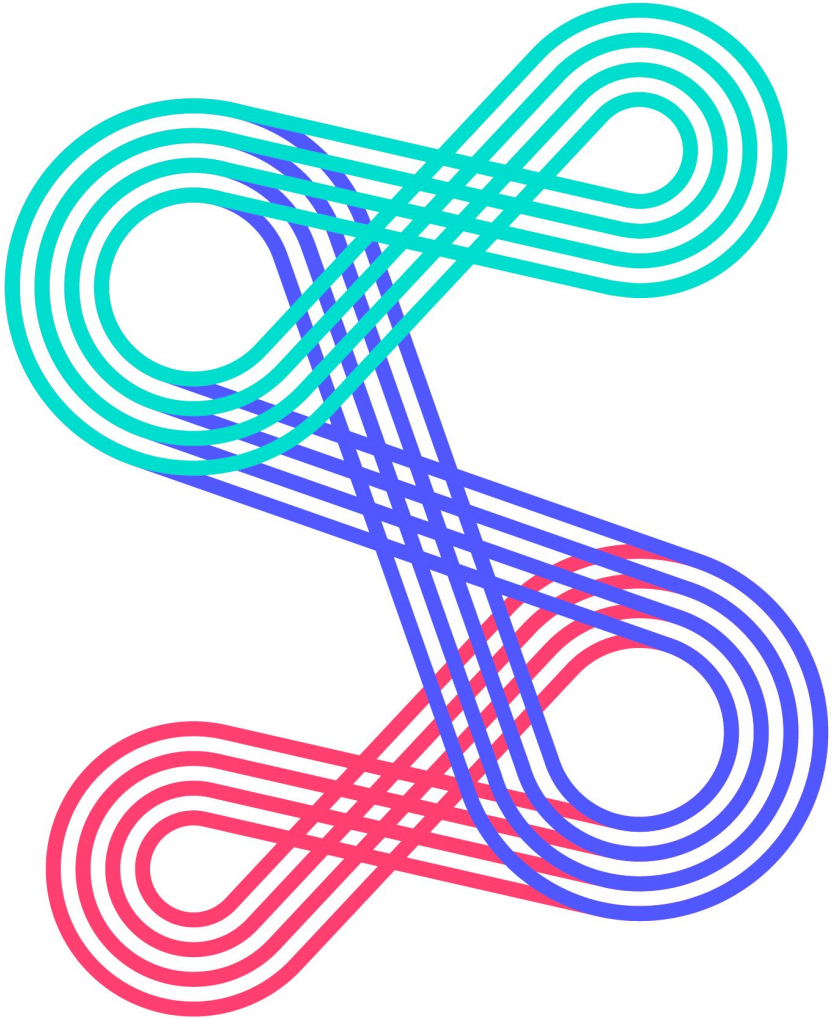
SELDON + BARCLAYS

Open Source Machine Learning

Barclays Accelerator London

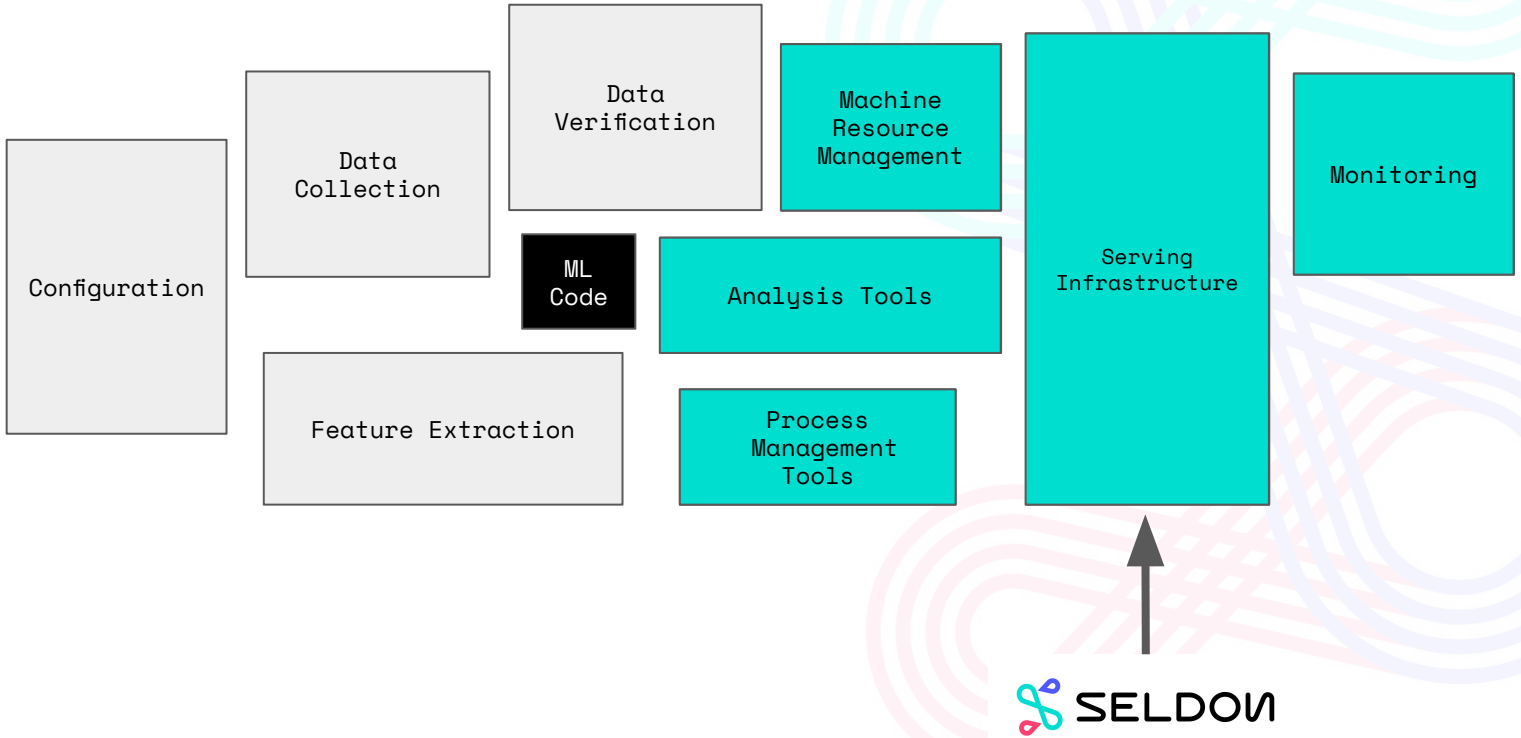
2014	2015	2016	2017
Data scientists in short supply Organisations want control & on-premise solutions	Recommendation & Prediction Platform agnostic = no lock-in Deploy on-premise or in the cloud	Barclays POCs Mentoring Network growth	FinTech Products Machine Intelligence infrastructure Global ML Models

Joining up data science, devops and business users.



Hidden Technical Debt in Machine Learning Systems

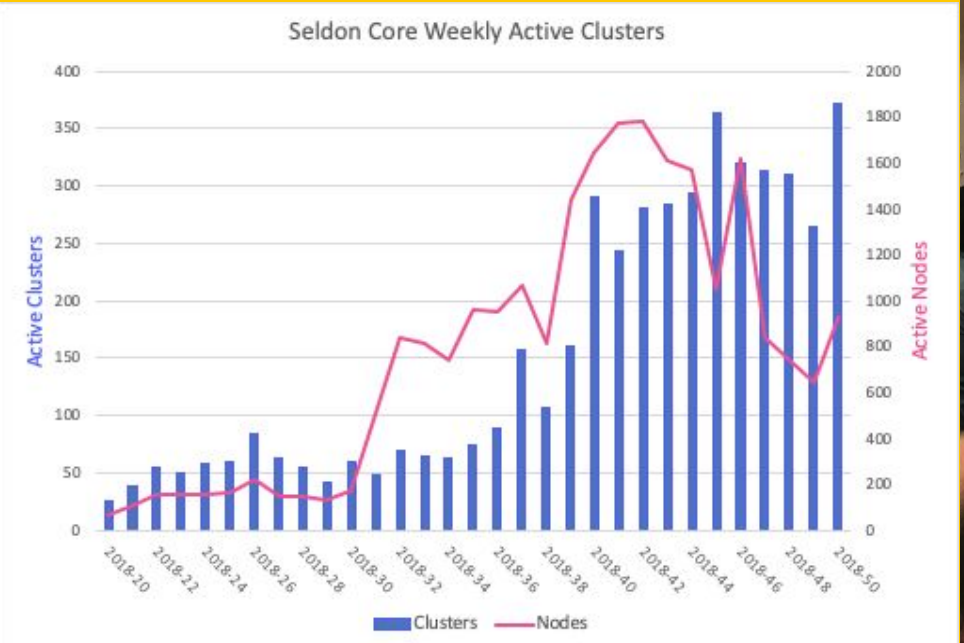
Size of the boxes corresponds to lines of code required



D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, Dan Dennison Google, NIPS 2015 Conference

Organic traction and a global community.

- Over **190,000 installs**
- **MoM growth: clusters 43.3%, nodes 85.97%.**
- WoW growth in actives over L12W: clusters 14.2%, nodes 6.6%. **Week 50 - record active clusters 372**!



TensorFlow™ LONDON

Organisers of the UK's largest TensorFlow meetup
*- now **over 2,100 members**!*

bit.ly/SeldonCoreSlack

Enterprise machine learning deployment

Kubeflow

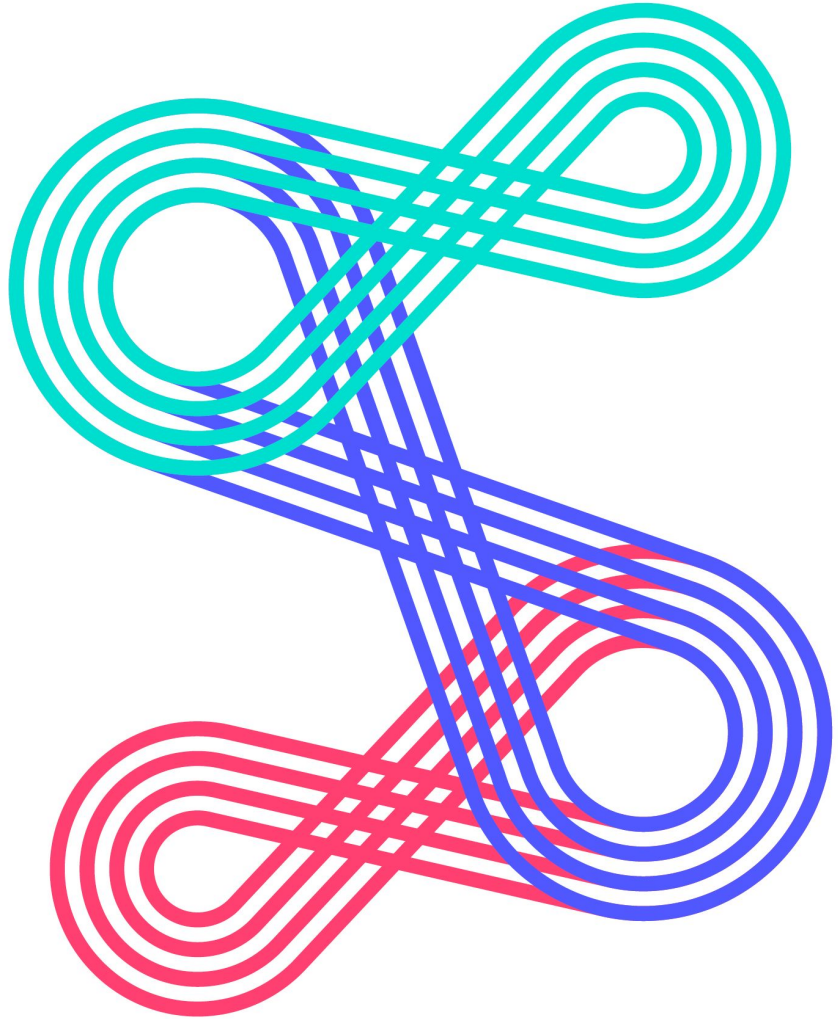
Data Ingestion → Data Analysis → Data Transform -ation → Data Validation → Data Splitting →

Trainer → Building a Model → Model Validation → Training At Scale → Logging →

Roll-out → Serving → Monitoring → Optimization

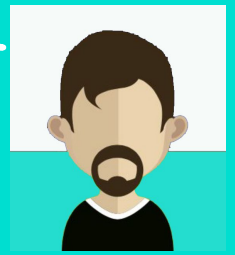
Seldon Core

Promise & Pitfalls



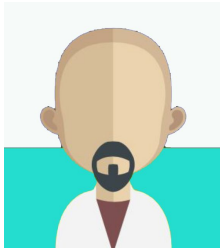
The ML pipeline is broken

DevOps do not speak the language of data science



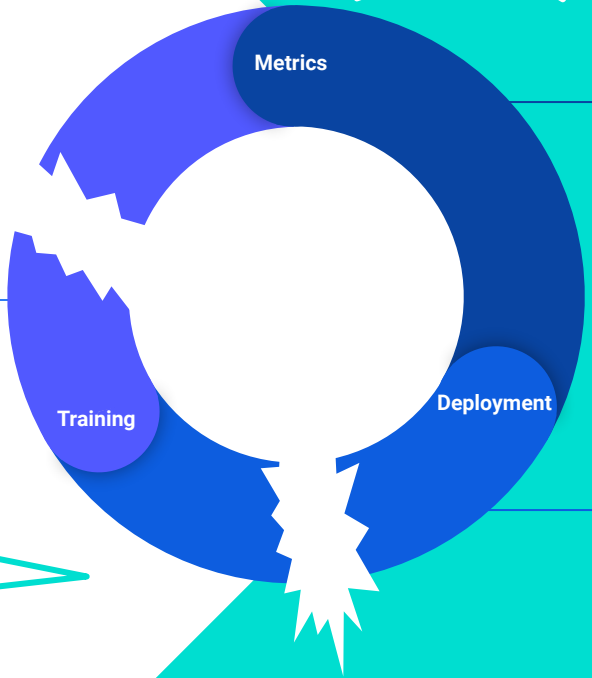
DevOps

Want to be able scale and repeat

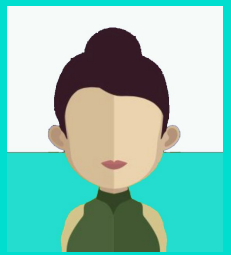


Data Scientists

Trains and builds ML models



Without feedback, models degrade and benchmarking is impossible



Data Engineers

In charge of development of Data pipelines

SELDON

Do you know what you're switching on?

FT Trading Room Knight Capital Group Inc + Add to myFT

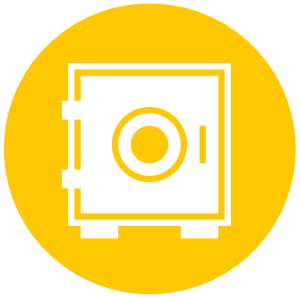
Knight Capital glitch loss hits $461m

"Knight did not have appropriate risk controls in place. ...No automated system to alert anyone to the discrepancy"

6 Machine Learning Trends in Financial Services

Regulatory requirements

Auditability & provenance



Security & data governance

Finance + tech niche skill sets

Integration with legacy ecosystem

Low latency & high data volumes

SELDON

APPG AI and Ethical Institute for AI

As part of our role as ML governance thought leaders we're on the board for **Ethics and AI Institute**.
http://ethical.institute/rfx.html

ETHICAL ML INSTITUTE

We're influencing ML compliance on the committee at **APPG AI**.

appg ARTIFICIAL INTELLIGENCE

BIG INNOVATION CENTRE

We're building out components to speed up compliance operations and reduce regulatory risk.

Seldon CEO Alex Housley giving evidence on AI alongside Oxford University, PwC, EDF, StatusToday to Co-Chairs, Stephen Metcalfe MP and Lord Clement-Jones

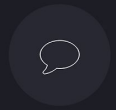
SELDON

The Institute for Ethical AI & Machine Learning

The 8 principles of responsible ML

1. Human Augmentation

I commit to assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes



2. Bias Evaluation

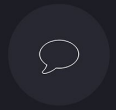
I commit to continuously develop processes that allow me to understand, document and monitor bias in development and production.

3. Explainability by design

I commit to develop tools and processes to continuously improve transparency and explainability of machine learning models where reasonable

4. Reproducible systems

I commit to continuously improve my machine learning infrastructure to enable for a reasonable level of reproducibility

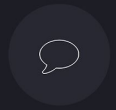


5. Displacement Strategy

I commit to identify and document relevant information so that business change processes can be developed to mitigate the impact towards workers being automated

6. Practical Accuracy

I commit to develop processes to ensure my accuracy and cost metric functions are aligned to the domain-specific applications



7. Trust beyond the user

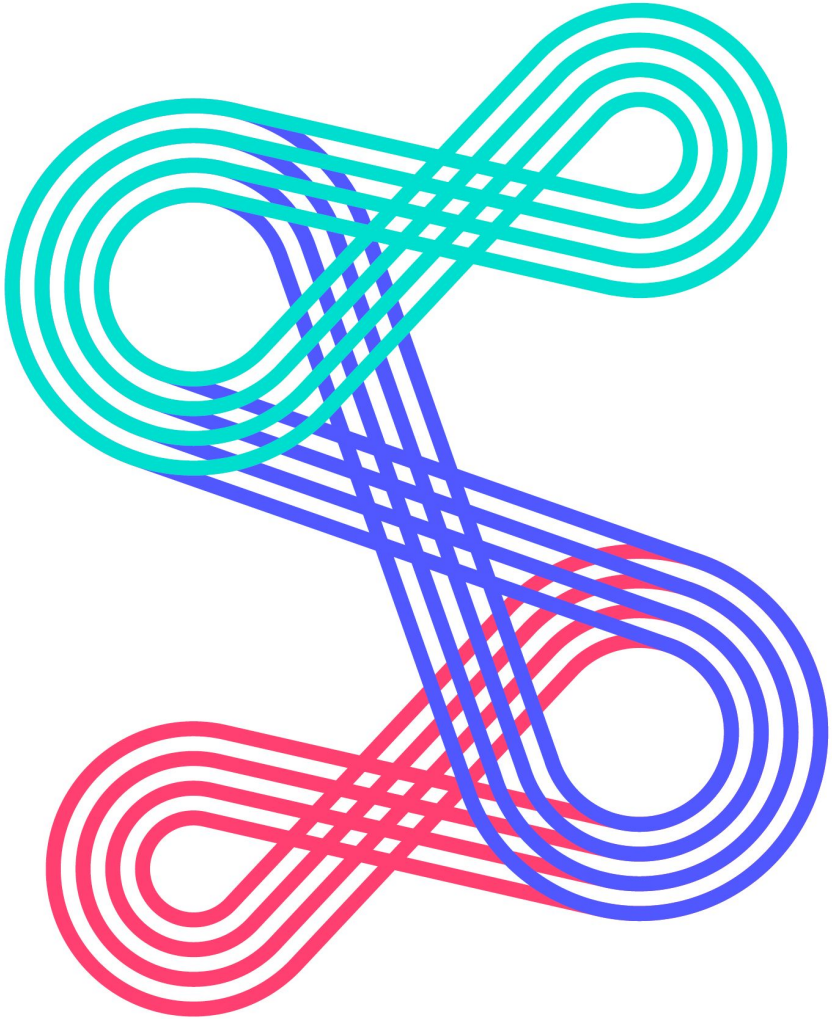
I commit to build and communicate processes that protect and handle data with stakeholders that may interact with the system directly and/or indirectly

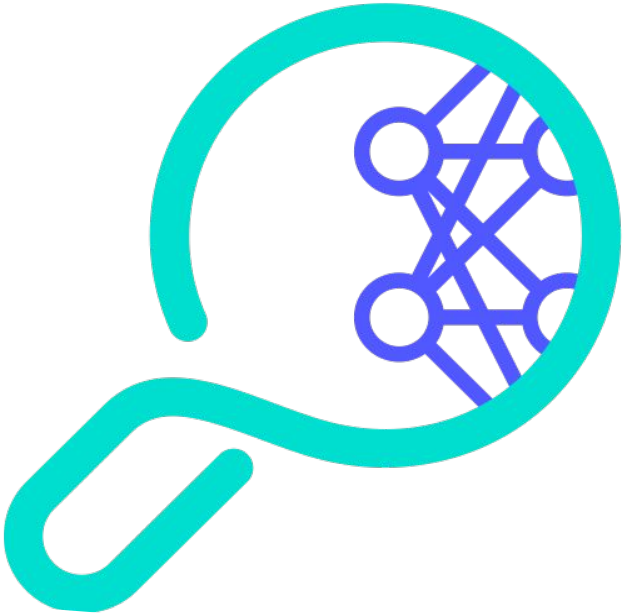
8. Data risk awareness

I commit to develop and improve reasonable processes and infrastructure to ensure data and model security are being taken into consideration during the development of machine learning systems

http://ethical.institute

Where next?





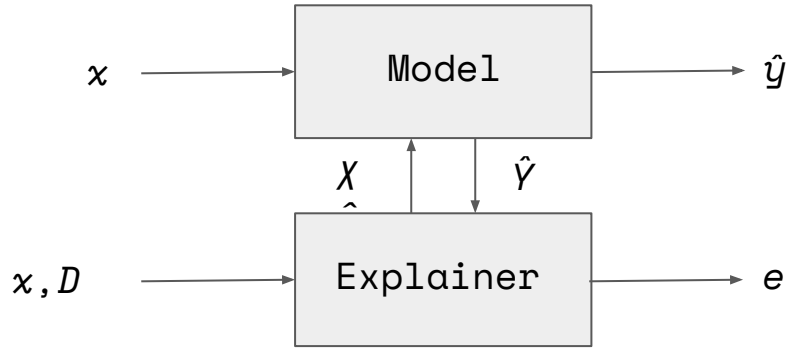
ALIBI

Machine Learning Explanations

SELDON

What is an explanation?

- Explain the model (global)
- Explain a single prediction (local)



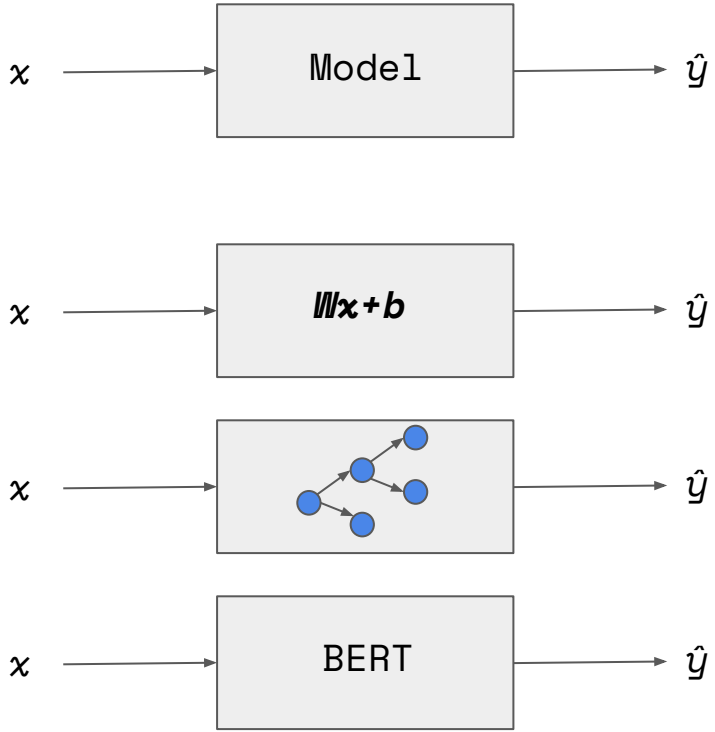
Black-box vs white-box models

Black-box:

- Predict fn.

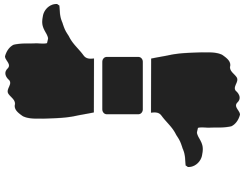
White-box:

- Predict fn.
- Loss fn.
- Architecture
- Training data (if applicable)



When to explain?

- Explanations are expensive
- Explanations are consumed by humans
- Low confidence predictions (Trust Scores)
- Anomalous instances (Outlier Detection)



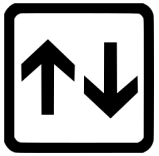
What makes a good explanation?

- Human interpretable
- Not over-simplified
- Trade-off between interpretability and fidelity

Interpretable questions, going beyond WHY?

- What is crucial for prediction to hold?

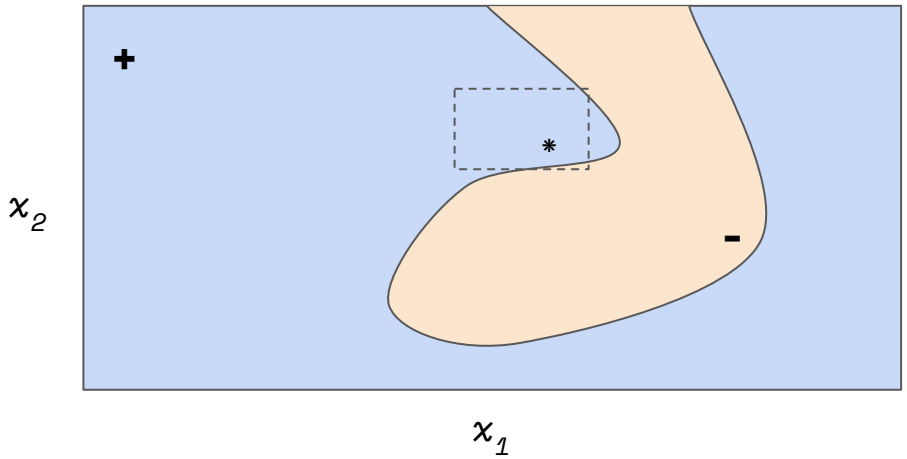
- What can you change to flip a prediction?



- How do the features relate to the prediction? (SHAP)
- What is the effect of the training data? (Influence fn.)

Anchors demo

- Output: if-then rules (interpretable)
- Guarantee explanation holds locally (faithful)

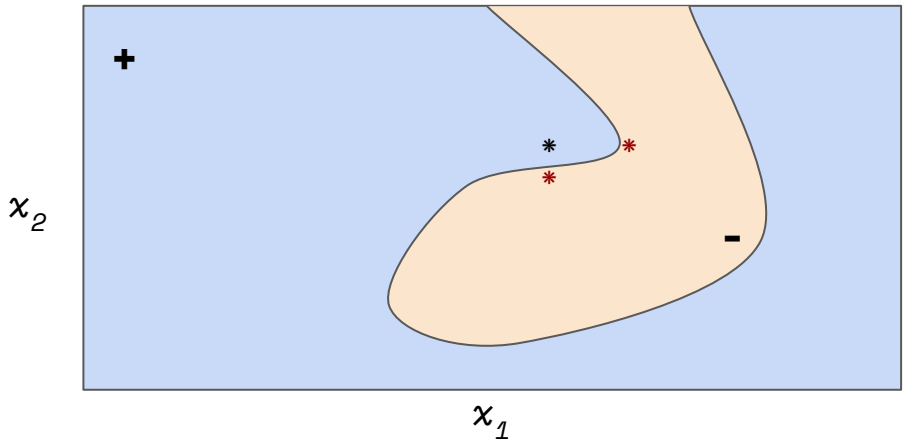


Anchor: if x_1 and x_2 in the box, will always predict + with 95% probability

Ribeiro et al., *Anchors: High-Precision Model-Agnostic Explanations* (2018)

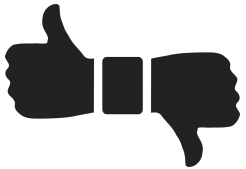
Counterfactuals demo

- Output: instance with a different prediction
- Some notion of minimal change (interpretable)



1. Van Looveren et al., *Prototypical Counterfactuals* (in preparation)
2. Wachter et al., *Counterfactual Explanations without Opening the Black Box* (2017)

Trust Scores demo



- Output: scalar score of model confidence
- $p(x \in C_1)=0.85$, but is it trustworthy?

Jiang et al., *To Trust or not to Trust a Classifier* (2018)

Thank You.

Please get in touch with any further questions

@seldon_io

hello@seldon.io

http://bit.ly/SeldonCoreSlack

SELDON