Imperial College
London

MSc Machine Learning & Data Science

Applicable Mathematics

Ethics of ML/DS
Part I

Week 5: Explainability

Dr. Chris Anagnostopoulos, Hon. Assoc. Professor

Imperial College
London

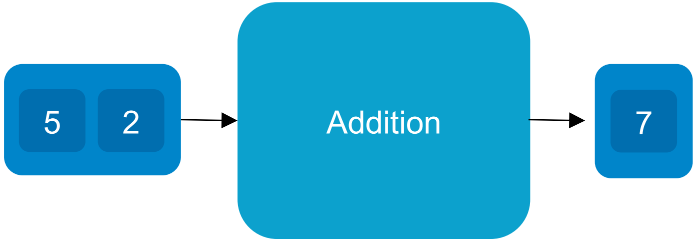
Explainability

- **The right to an explanation**
- Classical Interpretability and Partial Dependence Plots
- An overview of XAI techniques
- Are all explanations causal?

Code that writes itself

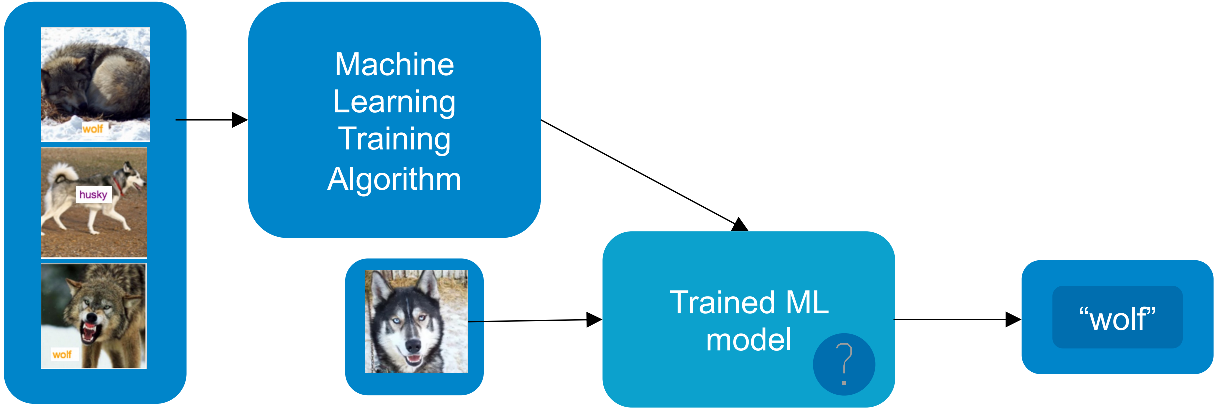
Algorithm

- A list of prespecified steps that complete a task (recipe).



Machine Learning model

- A computer program that programs parts of itself.



Do we have a right to an explanation?

- Morally, perhaps, but it is contested whether the GDPR makes it legally binding
- It would have to involve a negative decision with significant repercussions
- Technically challenging and unclear whether understandable
- Explanations can “leak” data or details of algorithm, or invite “gaming the system”

Purpose of explanations is:

(1) to inform and help the subject understand why a particular decision was reached

(2) to provide grounds to contest adverse decisions

(3) to understand what could be changed to receive a desired result

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31 (2017): 841.

What could an explanation look like?

Counterfactual explanations give the smallest possible change(s) that would overturn the decision.

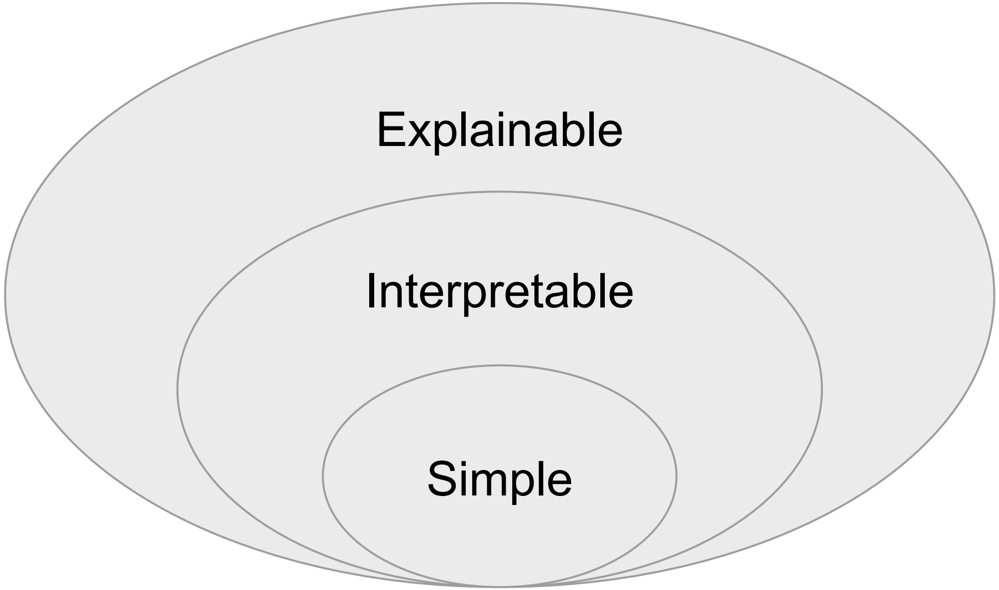
For relevance, we constrain explanations to mutable properties.

Nothing about the internal state of the algorithm is shared.

“You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.”

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31 (2017): 841.

Explainability, interpretability, simplicity



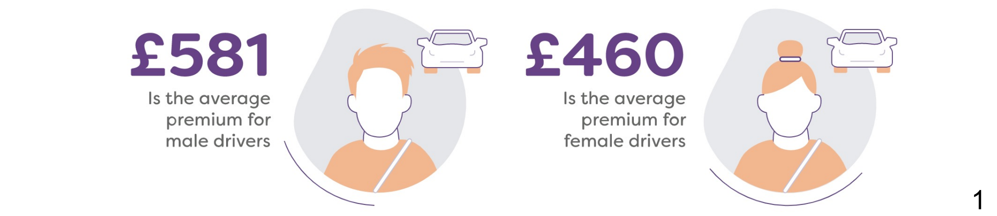
- Any model + counterfactual explanations
- Deep Neural Nets + XAI techniques
- Ensembles, Boosted/Bagged, Random Forest, XGBoost, + feature importance
- Hierarchical Bayesian Models / Mixed Models, Decision Trees
- Generalised Linear Models (e.g., linear regression, logistic regression), Bayesian Networks

How does XAI interplay with other principles?

- Explanations offer the chance for developers and users to scrutinize the model
- This can improve both its fairness and safety and security profile, and promotes transparency

Interplay with Fairness

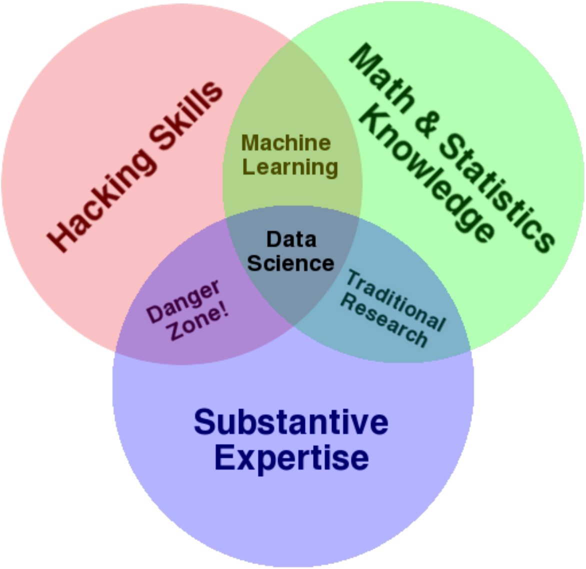
- Motivation behind request for explanation is often a sense of fairness
- Sensitive attributes can be correlated or proxied by other variables that are used by algorithm
  - *Insurers cannot by law discriminate against women*
  - *"Men pose higher per-km risk to others than women for all modes except buses"*[2]


£581
Is the average premium for male drivers
£460
Is the average premium for female drivers

1

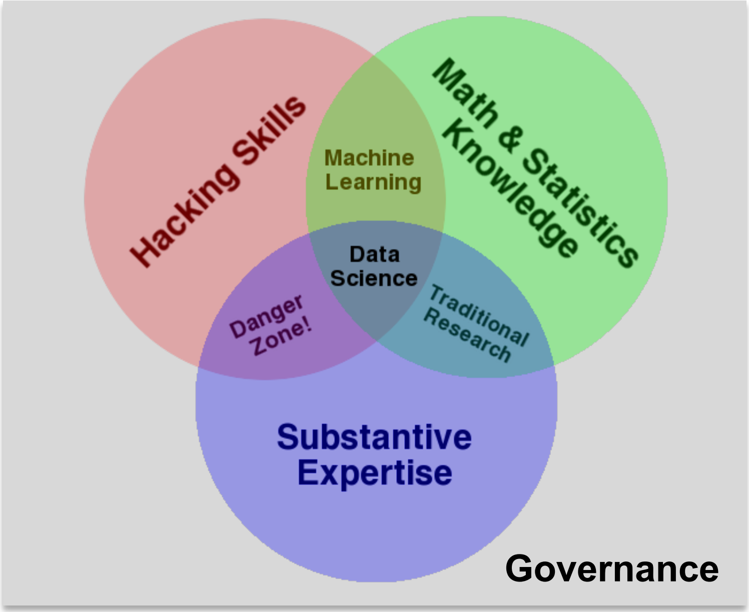
[1] https://www.moneysupermarket.com/car-insurance/why-do-women-pay-less/
[2] https://injuryprevention.bmj.com/content/27/1/71.abstract

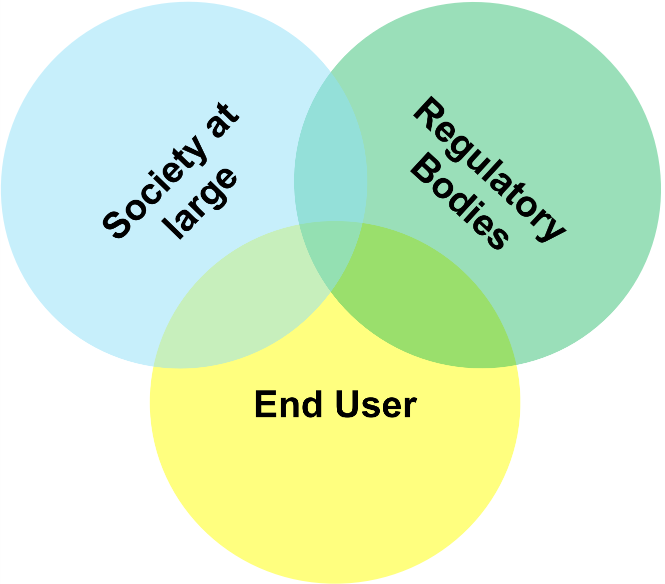
Explanations for whom?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Explanations for whom?

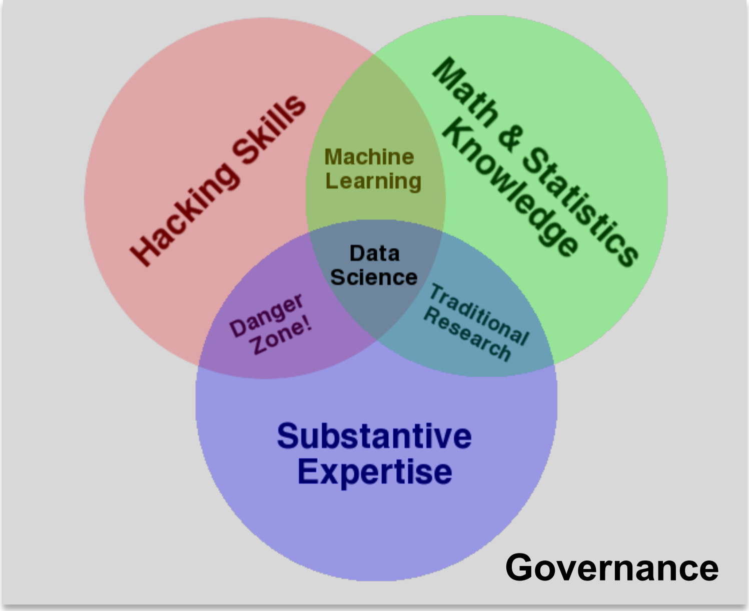


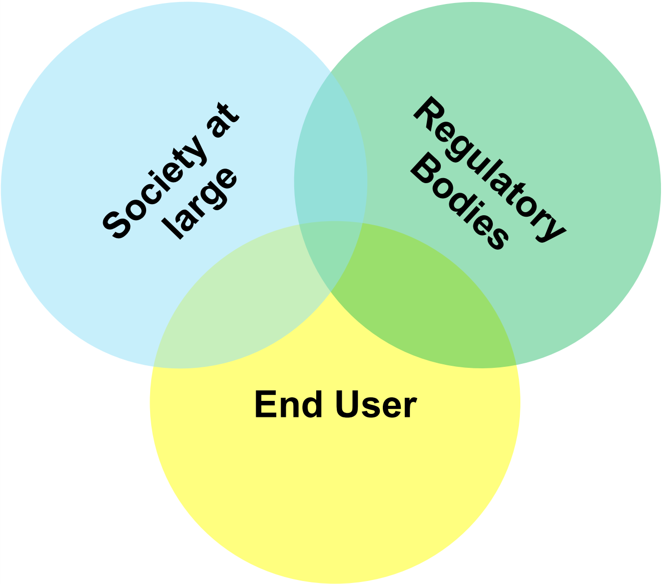


- Data scientists
- Domain experts
- Executives
- End user
- Regulators and society at large

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Explanations for whom?



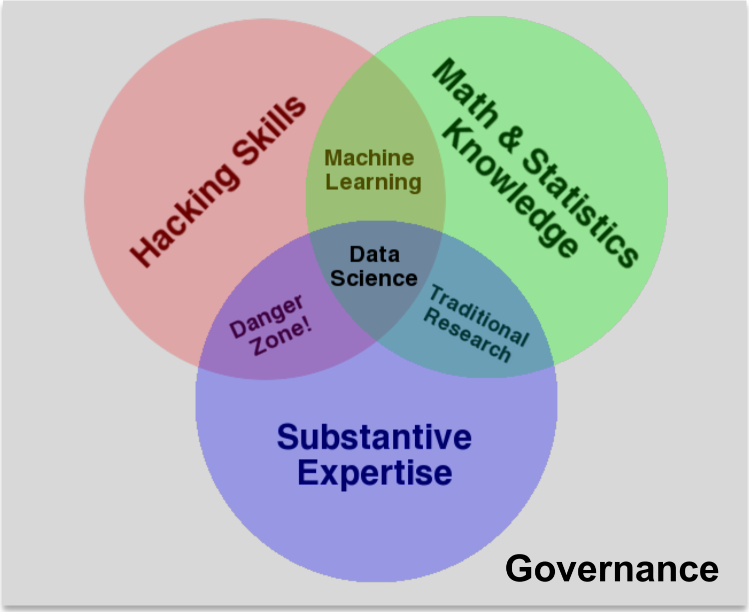


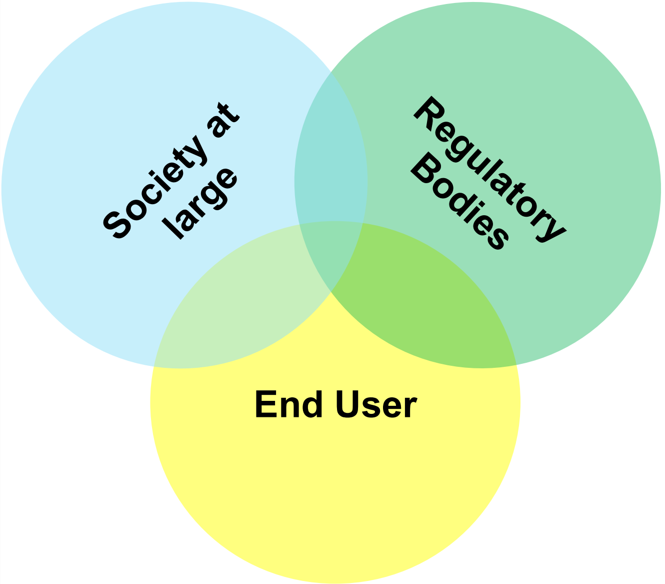
- Data scientists
- Domain experts
- Executives
- End user
- Regulators and society at large

Counterfactual explanations

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Explanations for whom?

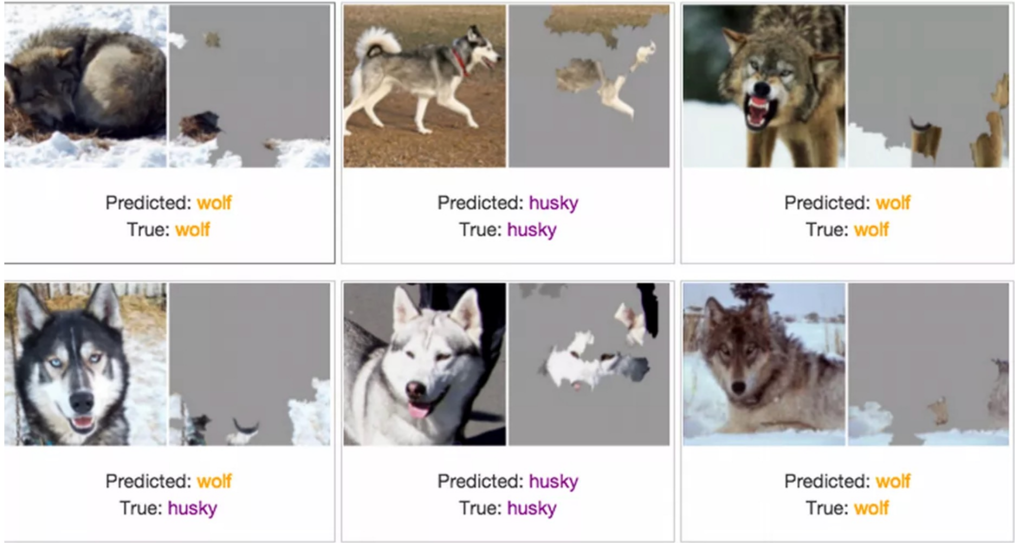




- Data scientists
  ↓
- Domain experts
- Executives
- End user
- Regulators and society at large

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

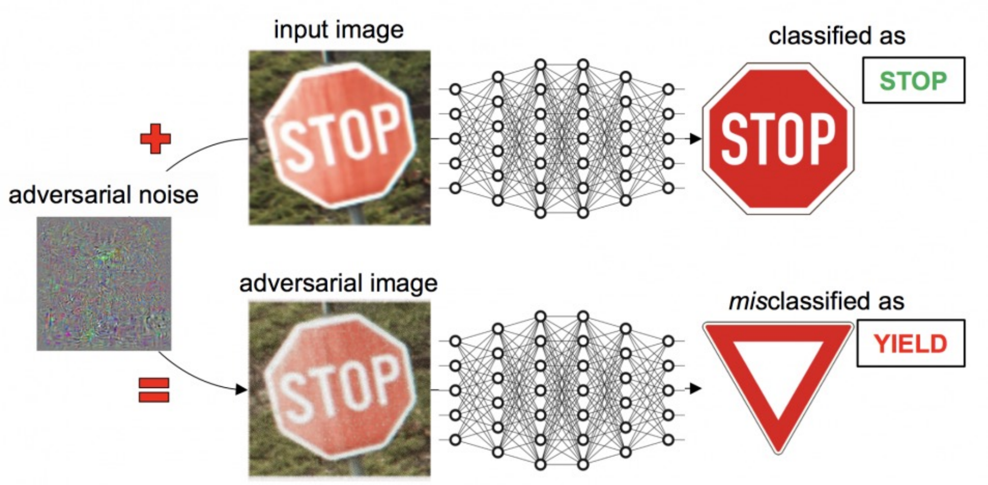
How can we trust an algorithm we don’t understand?



What makes a husky a husky? The snow!

Imperial College London

How can we trust an algorithm we don’t understand?



You can’t guarantee the safety of a system you do not understand

Summary

- Machine learning algorithms are unique because they are simultaneously autonomous, and self-programming (i.e., the exact logic by which the decision is made is determined by the training algo).
- There is an ethical and increasingly regulatory requirement to offer explanations to end users.
- Such explanations need not reveal internals of the algorithm, though they might have to
- Simplicity ensures interpretability, interpretability ensures explainability, but the research drive is focused on ways to achieve the latter without simple or inherently interpretable algorithms.
- Explanations have different requirements depending on the stakeholder it is directed to.
- In this course we focus mostly on explanations designed for data scientists and/or domain experts.
- Explanations can promote fairness, transparency, generalization ability, and safety and security.