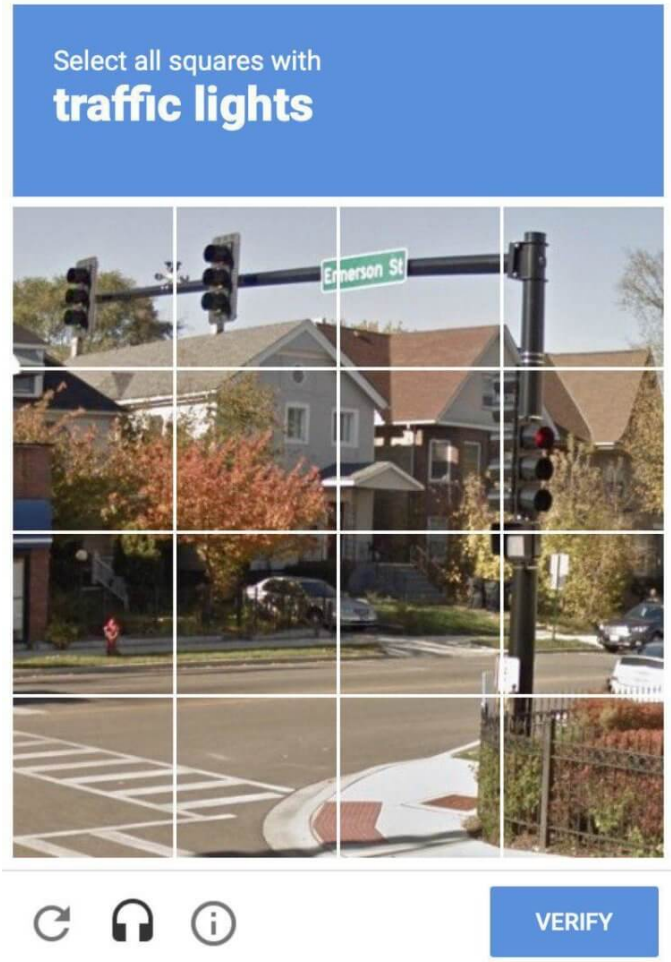
MUNI LAW Institute of Law and Technology

Data-driven technologies

... and the problem of responsibility

Rule-based and data-driven

- rule-based vs data-driven systems
- AVs combine these approaches
- this talk is about data-driven technologies

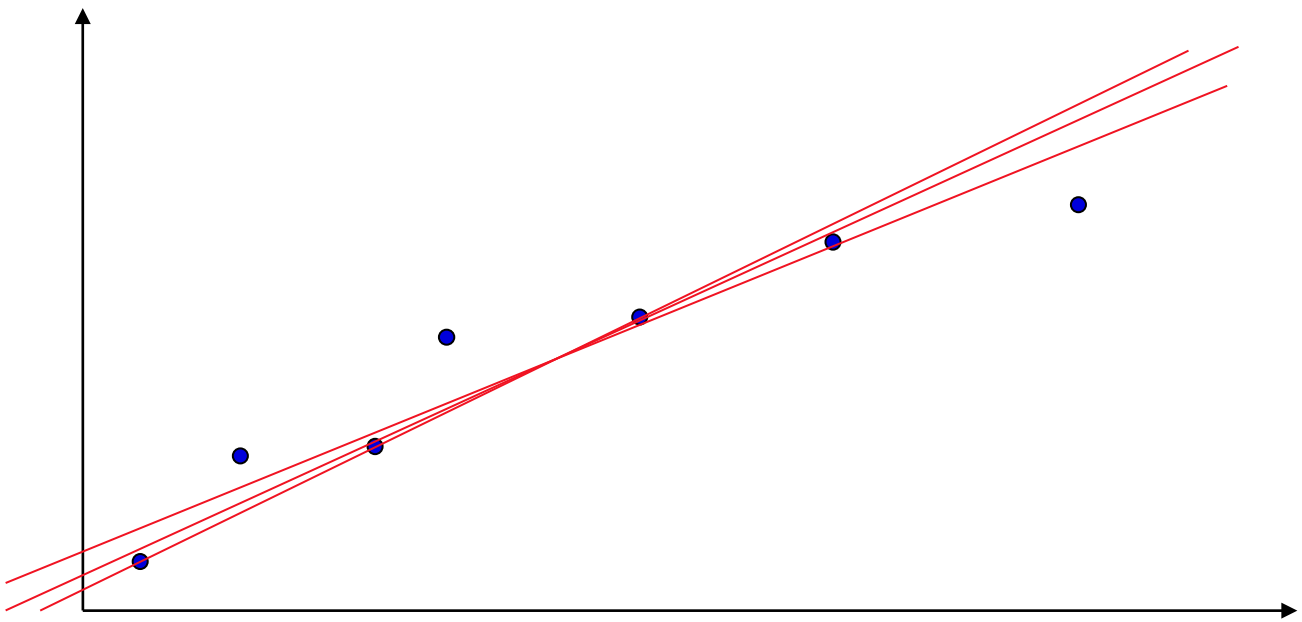


MUNI LAW

Statistical algorithmic models

- Data-driven models = statistical AI (a dominant stream of AI today)
- Linear regression (a highly interpretable model)

Weather conditions	Actual speed	Braking distance
1.032	50 kph	5.1 m
2.502	50 kph	6.2 m
2.750	50 kph	7.0 m
3.400	50 kph	8.7 m
3.625	50 kph	9.1 m
4.857	50 kph	10.7 m
5.000	50 kph	11.1 m



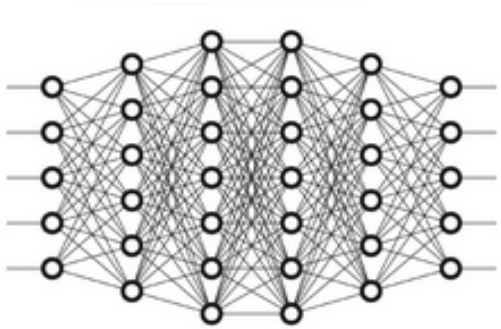
MUNI LAW

Liability and responsibility

- liability (that) v responsibility (why)
- In Slavic languages indistinguishable (e.g. in Czech law)
- BUT it is perfectly possible to have liability without responsibility and vice versa

Explainability of algorithms

- black box
- explainability as an attempt to address the responsibility issue
- linear models have high explainability



MUNI LAW

Explainability techniques

- Post-hoc explanations
- Definition: Interpretable description of the model behaviour
- Trade-offs between accuracy and interpretability
- Local and global explanations

Approaches for Post hoc Explainability

Local Explanations
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

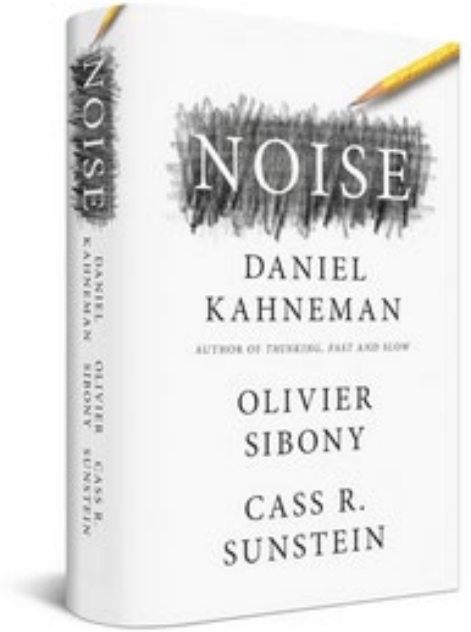
Global Explanations
- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

- More at https://explainml-tutorial.github.io/neurips20

MUNI LAW

Explainability of failures, but not faults

- Faults vs Failures - basic concepts in programming
- Explainability tracks “behaviour” → failures (not faults)
- At best, these techniques are relevant for identifying irrelevant features (those that should not have been considered) and discriminatory algorithms (those that should not have been developed)
- This is good (reveals noise), but not for responsibility



MUNI LAW

Benchmarking failure – AI metrics

- Liability is not about explanations, but about benchmarking (risk of failure, risk of damage).
- Not WHY a model fails, but how likely it is THAT it fails

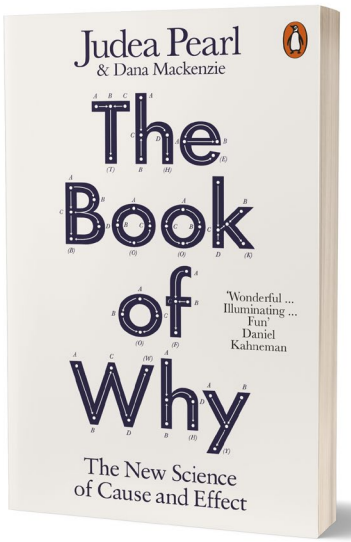
- Accuracy, precision, sensitivity
- AI metrics are thus relevant for failure + rules on causation for risk of damage (resulting from the failure or, in case of strict liability, from the application of the data-driven tool).

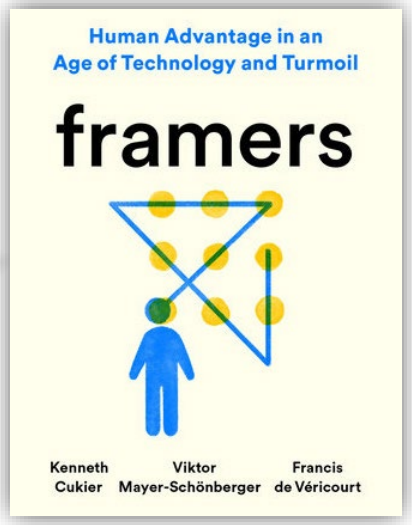
Responsibility does not always reduce the risk of failure, but can point out the fault

- Covid and restaurants: Who is liable in case of an infection?
- People have duties and responsibilities (which can differ from the observed behaviour of a “model” person)
- We are responsible and can give reasons because we are moral agents and can frame the issues as moral problems

Responsibility as a matter of framing

- For data-driven systems that feed the AVs' software, it is us who frame the problem and model the world (e.g. reCAPTCHA – both the framers and those who answer the problem).
- Modelling of the moral reality vs (statistical) modelling of the factual data





MUNI LAW

Responsibility as moral framing

- The responsibility rests with those who “frame” the algorithmic models by defining the tasks and with those who curate the data.
- Nietzsche and the morality of truth.
- We have moral sensibility but data-driven technologies are not sentient (Véliz, Moral zombies: Why algorithms are not moral agents 2021) → we cannot (in the strong sense) discuss the behaviour of an algorithmic model in terms of responsibility.

A global fault and the local failure

- With data-driven technologies, the responsibility issue is a global one, pertaining to the model that we see in the design of the algorithm; not a local one that we see in the concrete application of that algorithm.
- That is also what upsets us – there's no fault in the concrete circumstances because there's no local responsibility.

MUNI LAW

Conclusion: Model or reality?

- So how to bring more of some “responsibility” to the local level?
- Responsible innovation as an engaged discussion with the system, making the relevant features and desirable algorithms consistently explicit.
- Only then the statistical model will be getting closer to the reality. Well, the reality will be getting closer to the statistical model.
- Argument in a nutshell: Responsibility is about meaningful modelling of reality in global and local contexts, BUT data-driven technologies cannot provide that in relation to “moral” modelling.