SUSTAINABLE SOFTWARE:
ISSUES OF *BIAS, PROXIES AND GROUND TRUTHING* IN MACHINE LEARNING

Mireille Hildebrandt, FBA
@VUB Law
@Radboud CS

COHUBICOL
ERC Advanced Grant 2019-2024

About News Subscribe Blog Team Typology

It would be nice if all of the data which sociologists require could be enumerated because then we could run them through IBM machines and draw charts as the economists do. However, not everything that can be counted counts, and not everything that counts can be counted

– William Cameron, *Informal Sociology* (1963)

- **Code-driven systems:**
  - systems that do not learn based on training data (for instance legal expert systems, rules as code), including dedicated programming languages (though they are not systems)
- **Data-driven systems:**
  - systems that learn based on training data (whether supervised, unsupervised or reinforcement learning), including training datasets (though they are not systems)

- Obviously, many systems are ***hybrid*** in various ways

COHUBICOL research Qs

1. What are these systems claimed to achieve in terms of functionality?
2. How could this be substantiated (or not)?
3. *What upstream design decisions impact law and legal effect, and how?*

- Relevance: AI Act, GDPR, AI Liability Directive

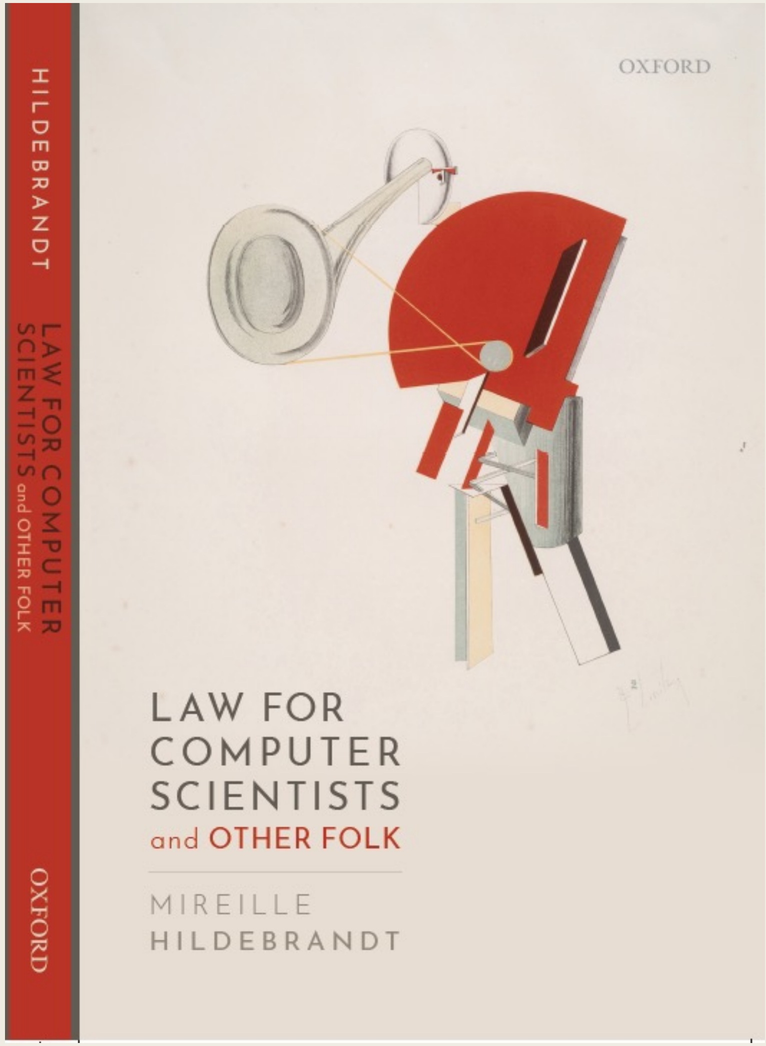
COHUBICOL research Qs

Upstream design decisions have normative effects, depending on the use case

- What **upstream design decisions** impact fundamental rights, and how?
  - In case of ***downstream deployment***
  - Depending on ***reasonably foreseeable*** use cases

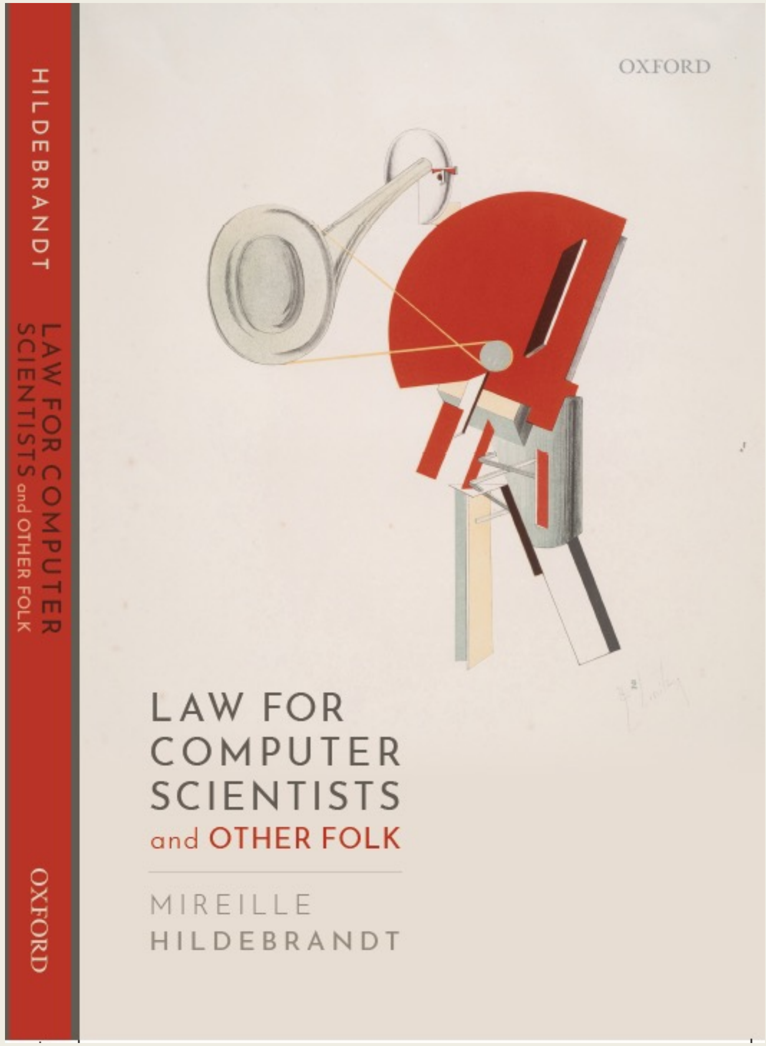
IPA talks

- Mireille: data-driven (machine learning)
- Paulus: code-driven (software engineering)



2020

- Law is ***not a bag of rules***
- Singling out a specific rule may misfire:
  - unwritten law probably applies
  - fundamental rights may be relevant
  - the complex context of the entire legal framework counts when interpreting the rule

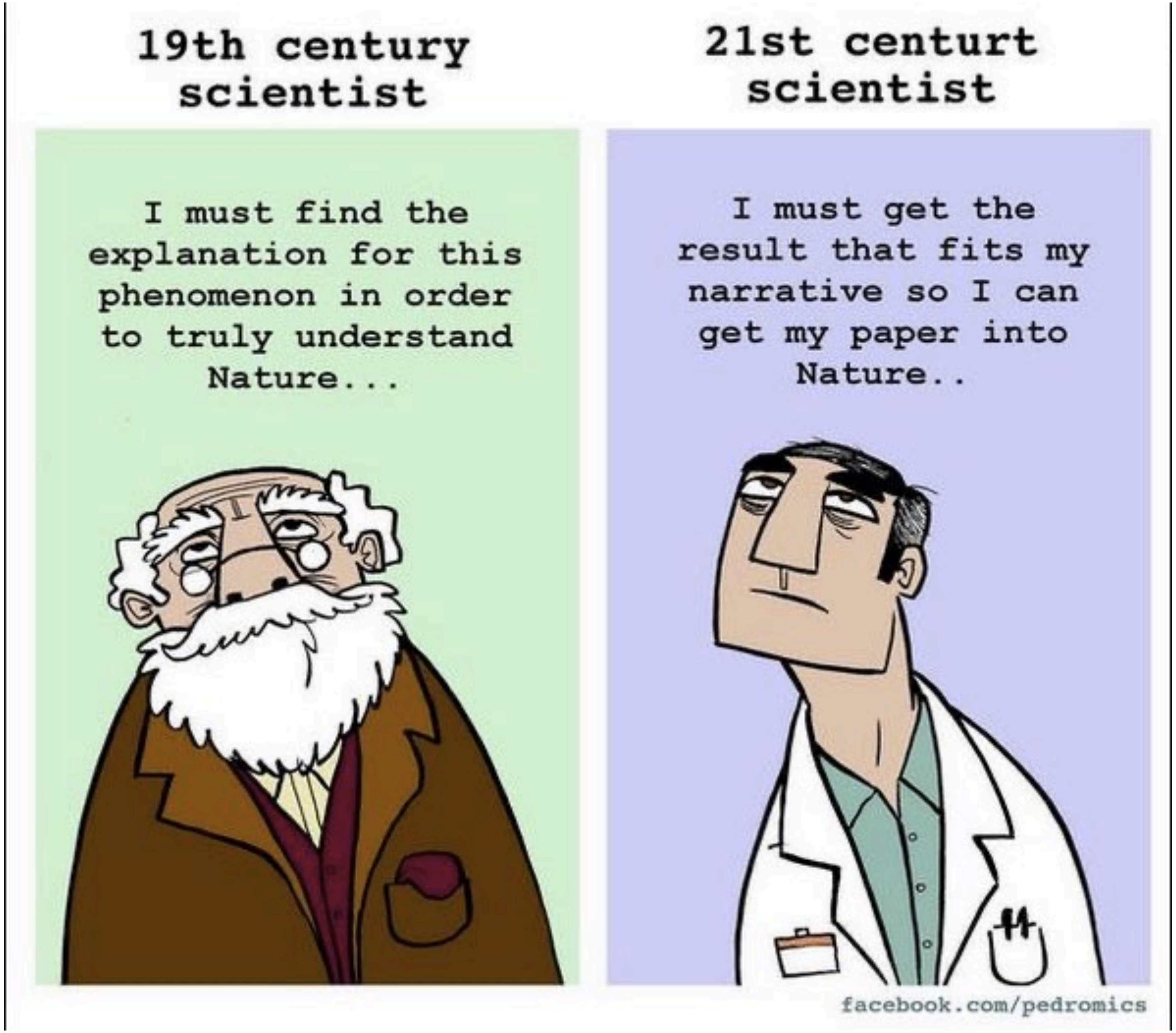


2020

- Law is ***not a bag of data***
- A statistical approach to legal norms misses the point:
  - unwritten law is normative
  - fundamental rights may be relevant
  - the complex context of the entire legal framework counts when interpreting the data

IPA talks

- Sustainable *in the real world*?
  - “correct, reliable, secure and fair”
  - 4R AI: robust, resilient, reliable, responsible
  - downstream impact of *upstream design decisions* is key
  - connection with
    - *reasonably foreseeable* unreliability, unlawfulness, unfairness

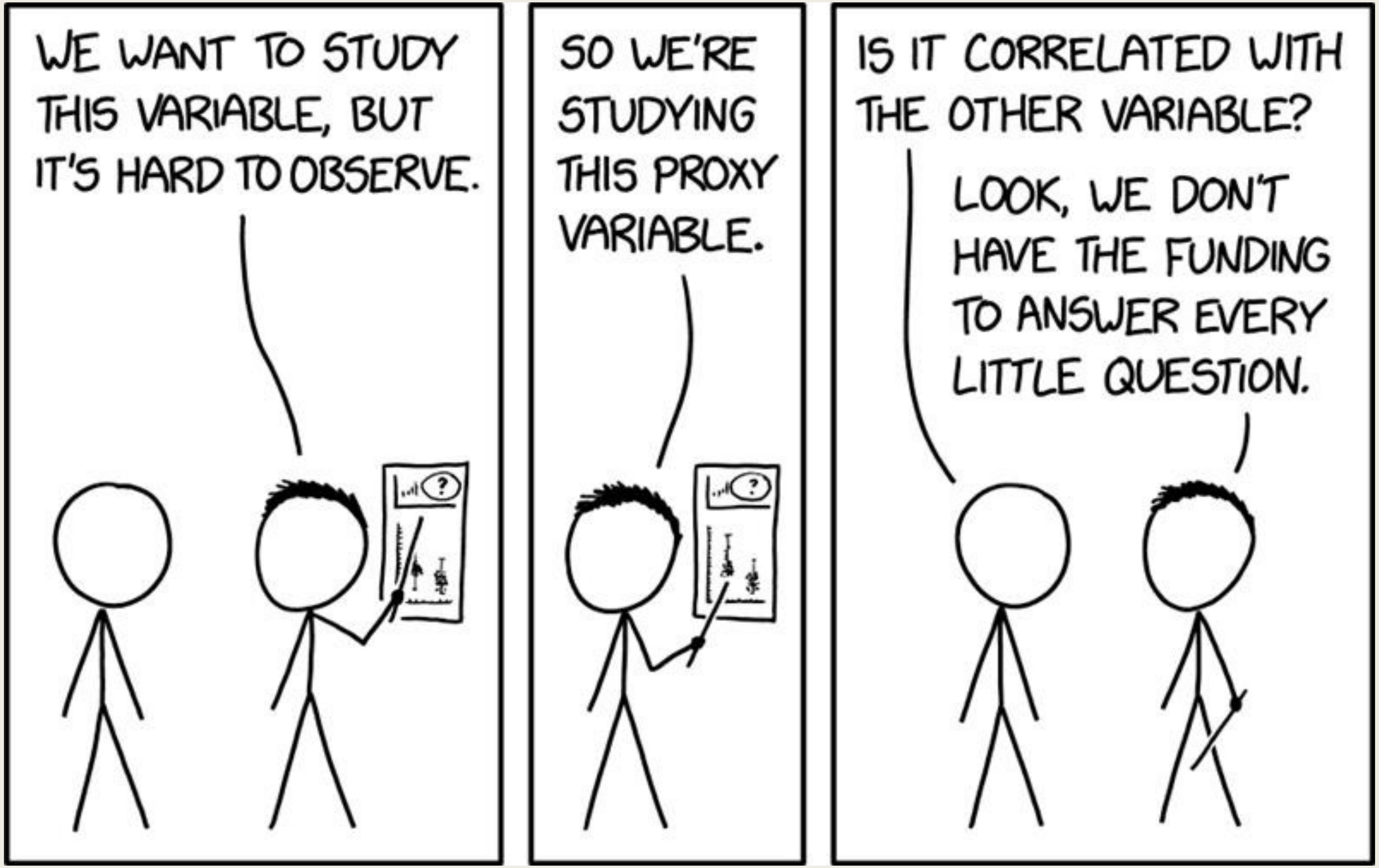

19th century scientist
I must find the explanation for this phenomenon in order to truly understand Nature...
21st centurt scientist
I must get the result that fits my narrative so I can get my paper into Nature..
facebook.com/pedromics

What's next?

- 3 types of proxies
- 3 types of ground-truthing
- *performance metrics and the ML pipeline*
- 3 types of bias

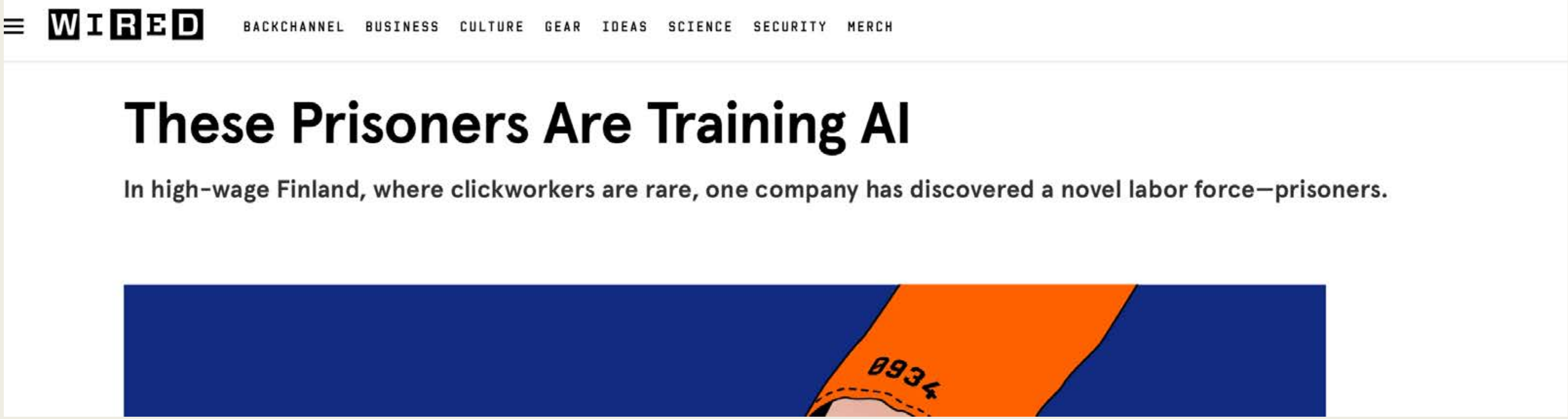
3 types of proxies

- A proxy is something that 'stands in' for something else
- In ML we need ***machine-readable*** proxies that stand for:
  - Relevant features deemed to define or influence real world phenomena
  - Real world representation in the form of data
  - Real world goals (or targets, cf the approximation of a target function)


WE WANT TO STUDY THIS VARIABLE, BUT IT'S HARD TO OBSERVE.
SO WE'RE STUDYING THIS PROXY VARIABLE.
IS IT CORRELATED WITH THE OTHER VARIABLE?
LOOK, WE DON'T HAVE THE FUNDING TO ANSWER EVERY LITTLE QUESTION.

3 types of proxies

- **Labels** in supervised ML
  - E.g. male/female, positive/negative, violation/non-violation
- **Training data** in unsupervised ML
  - E.g. Legal text corpora (case law) to enable case outcome prediction
- **Prompts** in reinforcement learning with human feedback
  - E.g. telling the system what output to avoid or prioritise


WIRED
BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY MERCH

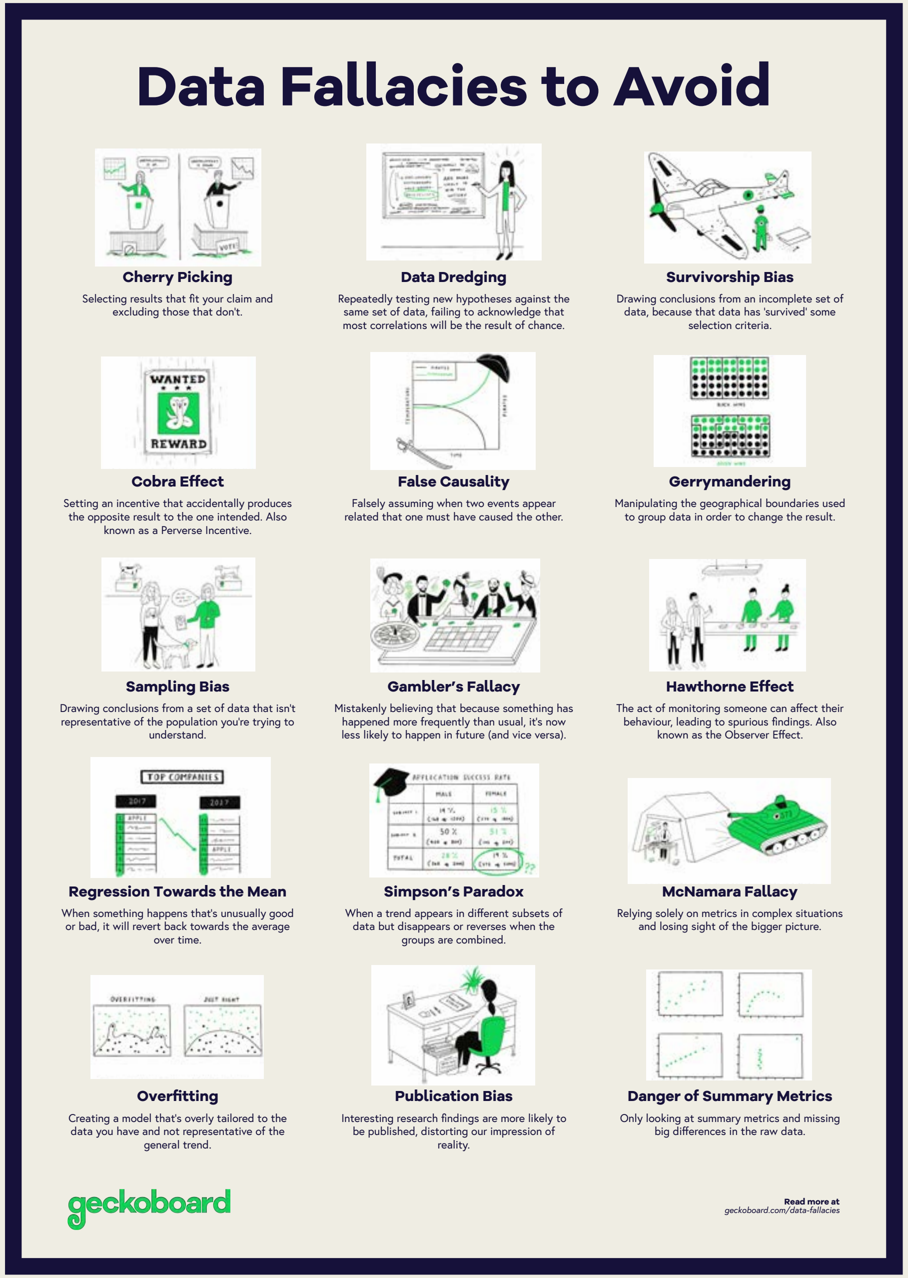
These Prisoners Are Training AI

In high-wage Finland, where clickworkers are rare, one company has discovered a novel labor force—prisoners.

- **Labels** in supervised ML
  - Who determines the label (defining the feature)?
  - Who does the labelling (attributing a label to the training set)?
  - What is the relationship between the labels and the real-world trigger?
    - The proxy relationship, e.g. in sentiment analysis
    - This concerns the framing problem

3 types of proxies

- Training data in unsupervised ML
  - Low hanging fruit
    (easy but irrelevant or incomplete data)
  - Benchmark datasets
    (may be a local minimum)
  - Assuming the distribution of the training data is that of future data



3 types of proxies

- Prompts in reinforcement learning with human feedback
  - RLHF
  - alignment
  - (with whose values?)
  - adversarial manipulation

3 types of ground truthing

- *Ground truth is a proxy*
  - for a slice or real-world (representation)
  - for an intended real-world (goal to be achieved)

OR

- *Ground truth as the incomputable real-world slice or goal*
  - To be approximated with data/variables/output models

3 types of ground truthing

Upstream design decisions 'define' the ground truth (as a proxy)

- Labelling: solutions to ***intra- and inter-rating disagreement***
  - Cp radiologists and judges
- Training data: trade-offs when ***selecting and curating training data***
  - *Cp medical treatment data and legal text corpora*
- Test data: choices made when ***deciding on test-data***
  - Cp post-treatment health data and legal text corpora

Check the issue of ***data leakage*** between health and law

ML performance metrics

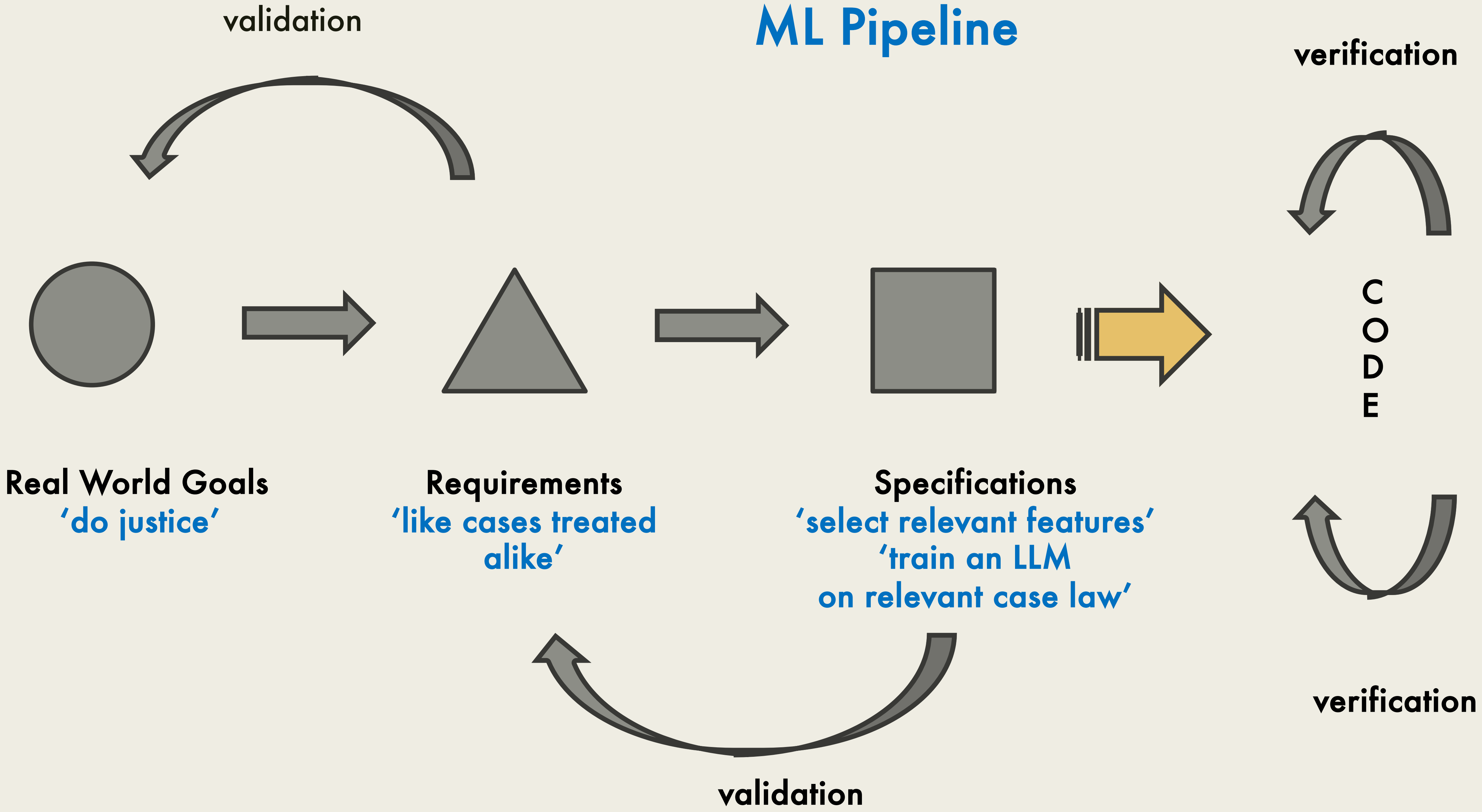
ML output-testing:

- **Accuracy**
- **Precision**
- **Recall**

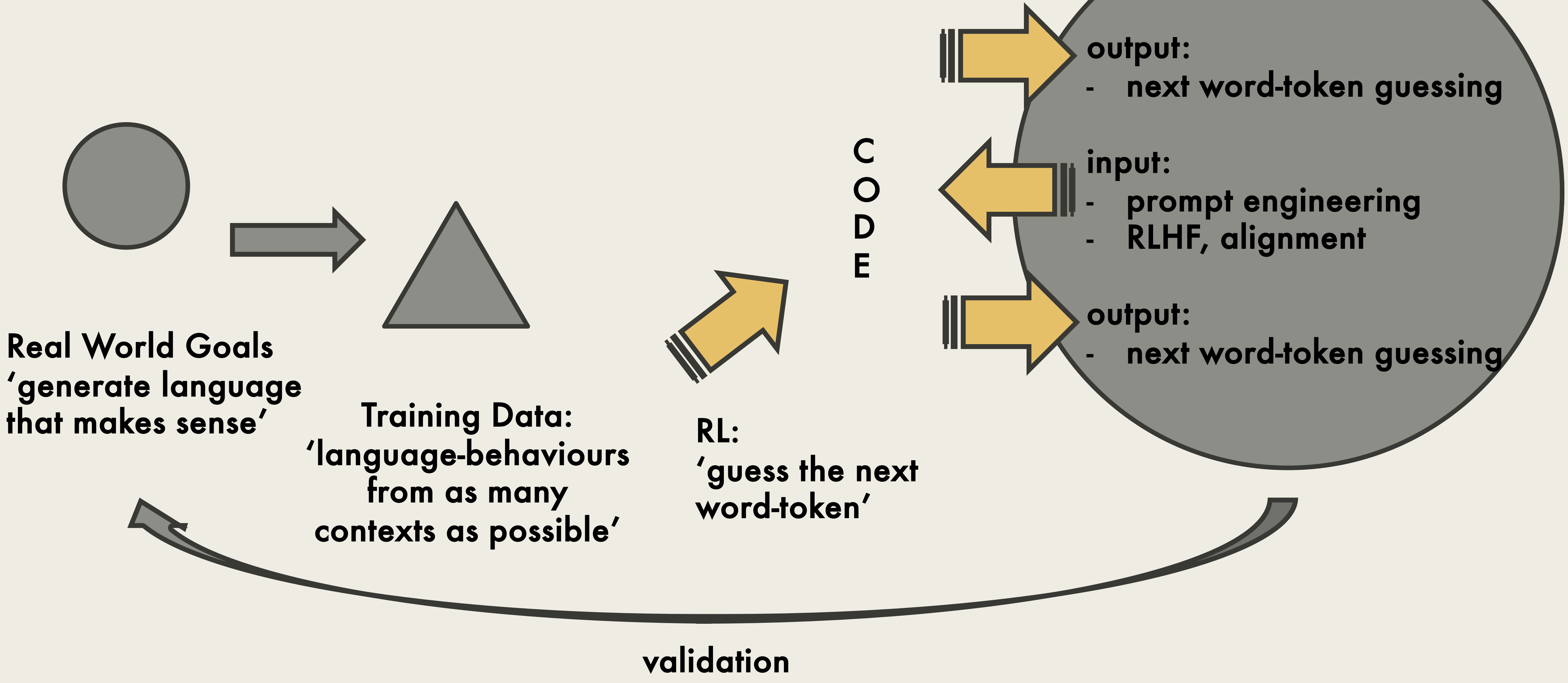
These *performance metrics depend on* assumptions inherent in:
- the training data
- the learners, hypothesis space
- the feedback provided

- Performance metrics offer an *internal* test
- Just like verification and some types of validation
- What we very much need is *external validation*, testing against real-world goals
- *Real-world goals* is not the same as real-world data (which is a proxy)

ML Pipeline


validation
verification
C O D E
Real World Goals
'do justice'
Requirements
'like cases treated alike'
Specifications
'select relevant features'
'train an LLM
on relevant case law'
verification
validation

LLM Pipeline



3 types of bias

- **Productive bias,** that is ***key to machine learning***
- **Ethical bias**, that may reinforce existing or introduce new *unfairness*
- **Unlawful bias,** that implies *discrimination* based on a prohibited ground

3 types of bias

- **Productive bias,** that is *key to machine learning*
  - Inductive bias, which depends on the training data, the labelling, prompts
  - Inductive bias is inevitable, productive and generative (Mitchell):
  - "a learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances'
  - *Simple comme bonjour*, but what about DL and LLMs?
  - Supervised and reinforcement learners: target concept (the goal)
  - Unsupervised learner: no target concept, loss function and optimisation method

3 types of bias

- **Ethical bias,** that is *inevitable in machine learning*
  - ML upstream design decisions (training data, target function, hypothesis space, loss function and optimisation method, goals and prompts)
  - are neither good nor bad, but never neutral
  - they will have normative impact insofar as they e.g.
    - produce different output models used for ADM
    - change the 'choice architecture' for deployers and end-users
  - this may also result in moral impact, e.g. unfairness
  - however, who defines what counts as unfair?

3 types of bias

- Unlawful bias, that depends on the violation of e.g. the right to non-discrimination
  - ML upstream design decisions (training data, target function, hypothesis space, loss function and optimisation method, goals and prompts)
  - may result in decisions or a 'choice architecture' that:
    - discriminates on a prohibited ground, such as
      - gender, sexual orientation, ethnic background, religion
      - note that economic deprivation is not a prohibited ground:
        - a higher premium for low-income folk does not involve a prohibited ground, unless
        - E.g. low-income coincides with a specific ethnic background

So what?

- Upstream design decisions matter – a lot:
  - for sustainable output models:
    - robust, resilient, reliable and responsible
    - fairness and human dignity
    - avoiding violation of fundamental rights