Research and Documentation Centre

On interpretation of predictive AI models

Mortaza S. Bargh & Sunil Choenni

18 Nov. 2021

Wetenschappelijk onderzoeks- en kennisinstituut
voor het ministerie van Justitie en Veiligheid

Outline

Introduction

- Data-driven / AI applications

Reasoning on AI models

- Logic types
- Naive way
- Statistical truth

On two solution directions

- Responsible AI
- Dealing with AI uncertainty

Conclusion

- Takeaways

Introduction

Data-driven / AI applications

Using various data sources, like

- Big data (v's: volume, variety, velocity, …)
- Registration data (related to daily operations of orgs)

Applying various data processing techniques

- From statistics and AI (machine learning)
- For data mining, classification, clustering, ...

To learn a relevant model of a phenomena in the real world

Why AI models?

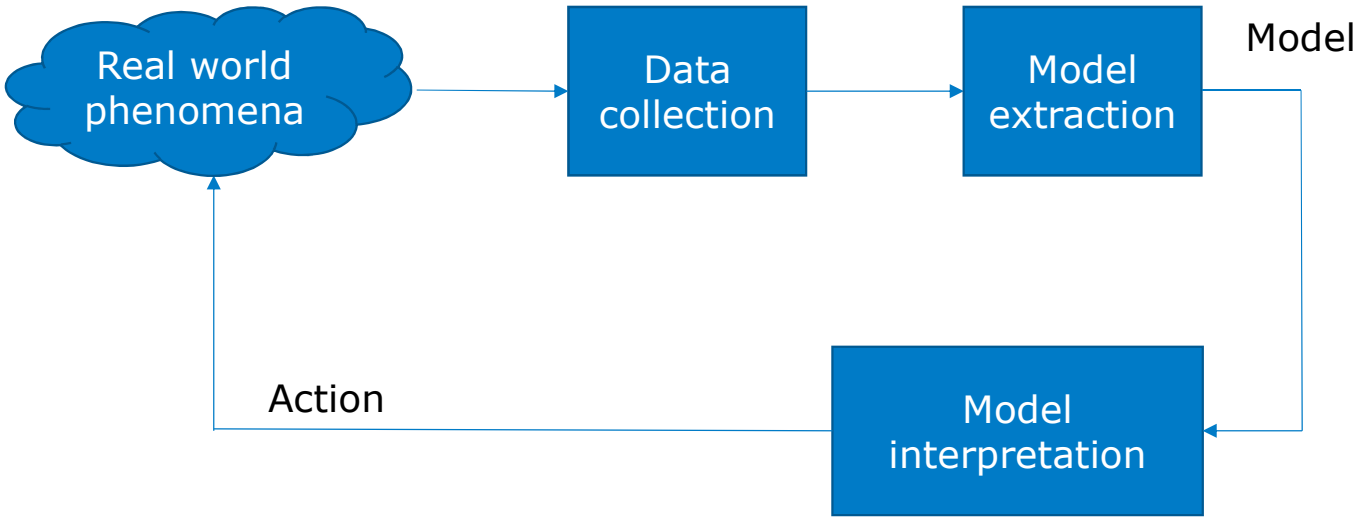
Two main reasons

- To predict a real-world phenomena
- To understand the real-world phenomena

Being both

- Contrary: models that predict well, do not necessarily offer much insight
- Reinforcing: insight can help to improve the predictive accuracy of models

Use of AI models



Reasoning on AI models

Reasoning schemes (Pierce,1903)

Deduction

1) All men are mortal
2) Socrates is a man
3) **So: Socrates is mortal**

Induction

1) Socrates died
2) Kant died
3) Plato died
4) …
5) **So: All men are mortal**

Abduction

1) All men are mortal
2) Socrates died
3) **So: Socrates is a man**

Statistical truth

Extracted model

Young men driving leased cars have 80% chance of being involved in car accidents

Interpreted model (in the case of Bob)

- Bob is a young man driving a leased car
- **Thus:** Bob has 80% chance to be involved in a car accident

Typical AI-based reasoning

Deduction

1) All men are mortal
2) Socrates is a man
3) **So: Socrates is mortal**

Induction

1) Socrates died
2) Kant died
3) Plato died
4) …
5) **So: All men are mortal**

Abduction

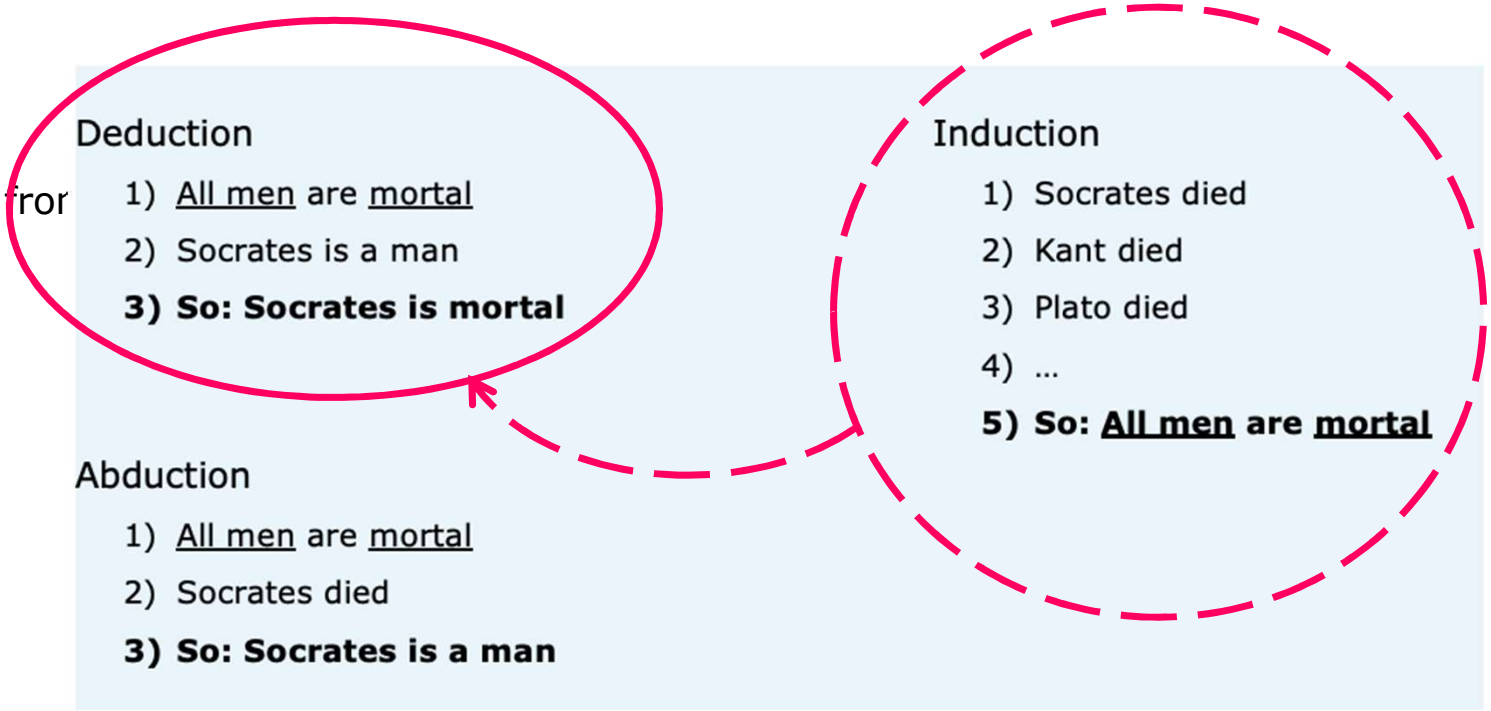
1) All men are mortal
2) Socrates died
3) **So: Socrates is a man**

Popular interpretation of statistical models

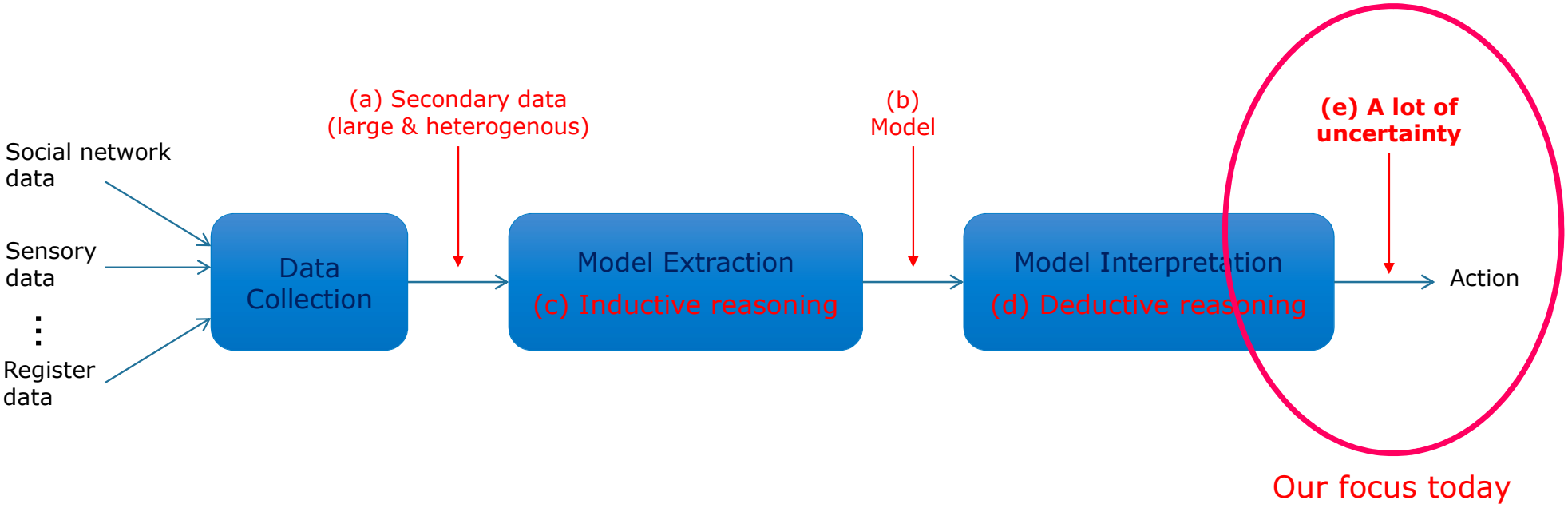
Being scientific

- Data driven
- Scientific (stemming fror

Actually: Being naive



Typical AI-based reasoning



Reasoning on a statistical truth

Extracted model

Young men driving leased cars have 80% chance of being involved in car accidents

Interpreted model (in the case of Bob)

- Bob is a young man driving a leased car
- **Thus:** Bob has 80% chance to be involved in a car accident

Reasoning on a statistical truth

Extracted model

Young men driving leased cars have 80% chance of being involved in car accidents

Interpreted model (in the case of Bob)

- Bob is a young man driving a leased car
- **Thus:** Bob has 80% chance to be involved in a car accident

(a) Frequentist approach: **Is p=80% relative frequency?**

- Many drivers/clones
- Many drives by a driver

Reasoning on a statistical truth

Extracted model

Young men driving leased cars have 80% chance of being involved in car accidents

Interpreted model (in the case of Bob)

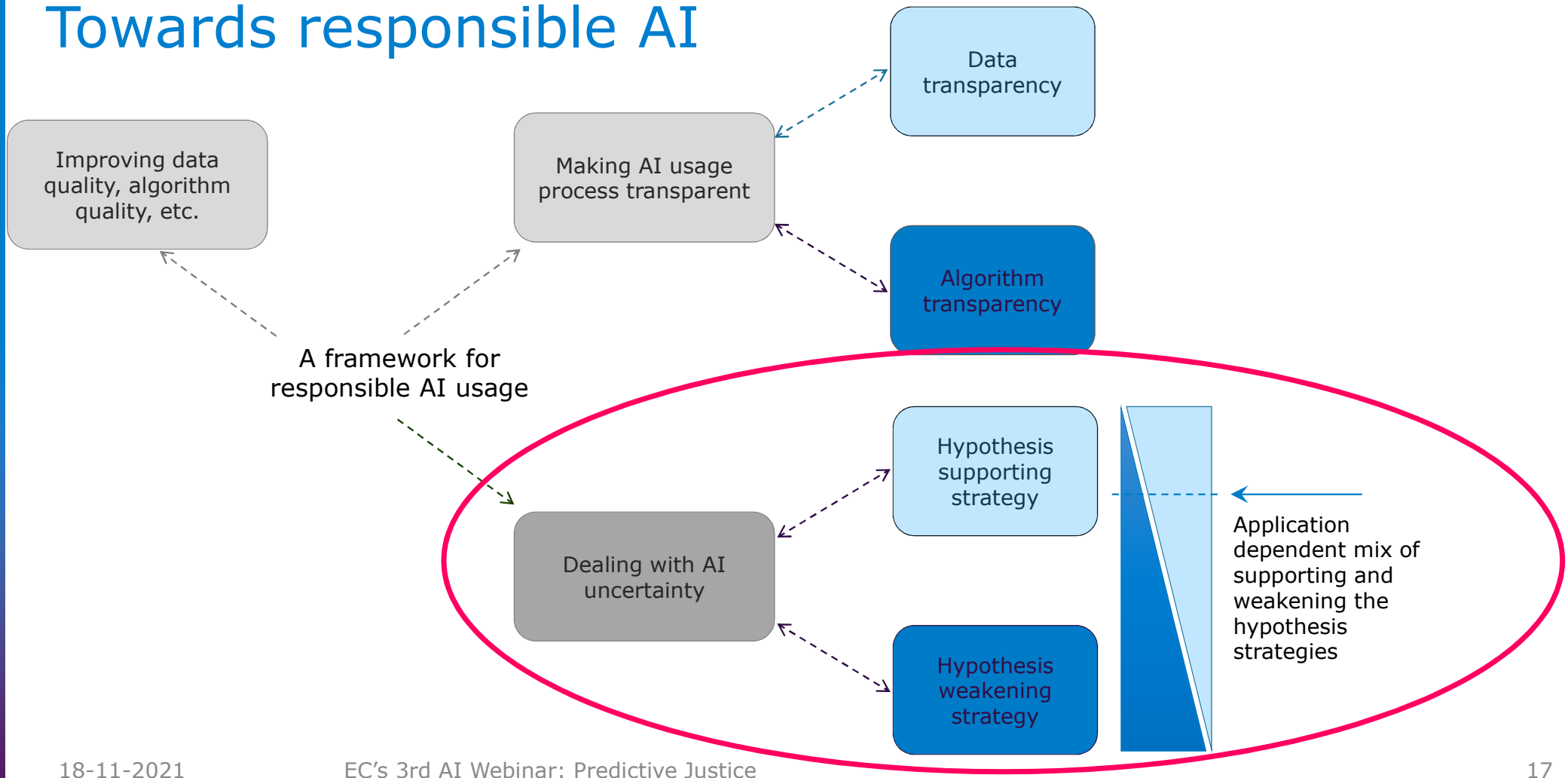
- Bob is a young man driving a leased car
- **Thus:** Bob has 80% chance to be involved in a car accident

(b) Subjective approach: **Is p = 80% a quantified judgement?**

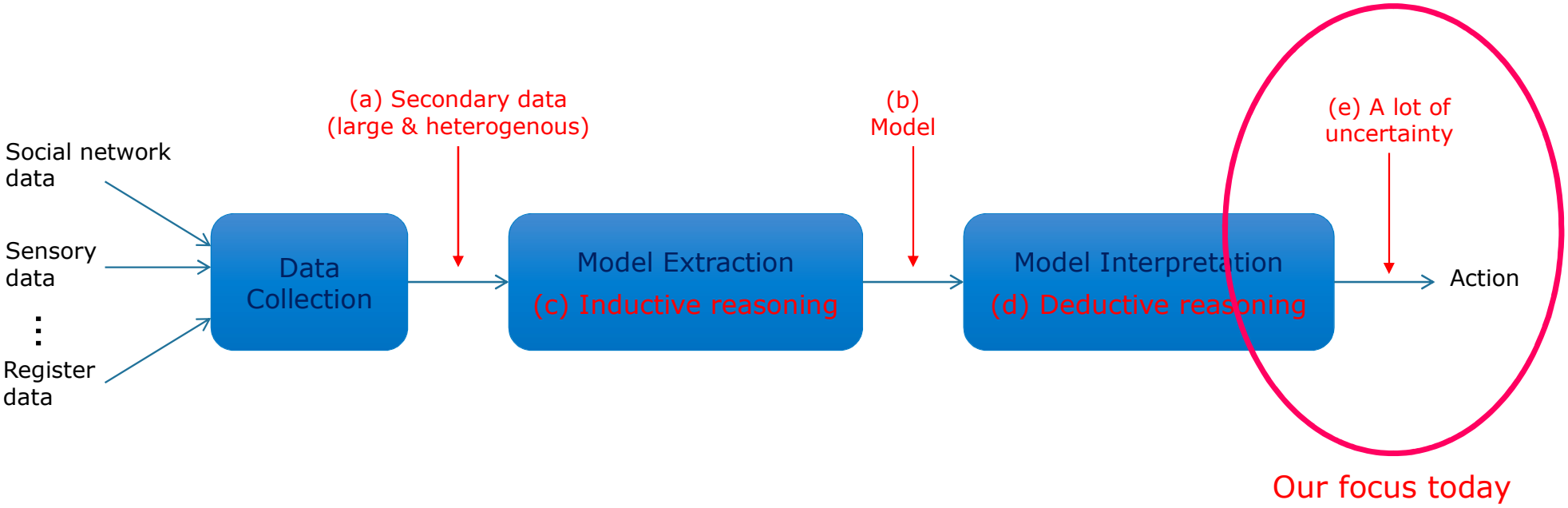
- Prior probability (may include "frequentist approach")
- Interpretation maybe different for the receiver entity (Bob, system user) and the probability generating entity (the AI algorithm, AI experts)

On two solution directions

Towards responsible AI



Typical AI-based reasoning



Strategy 1 to deal with the AI outcome

Consider the AI outcome as a central body of evidence and

- Extract a hypothesis like: Bob is a risky driver
- Search for evidences that **weaken** the hypothesis like: Bob is a cautious man
- If enough evidences are found to weaken the hypothesis; then the hypothesis is **rejected**

Self-denying prophecy

- True hypothesis might become false
- Advantage: Reducing false positives

Strategy 2 to deal with the AI outcome

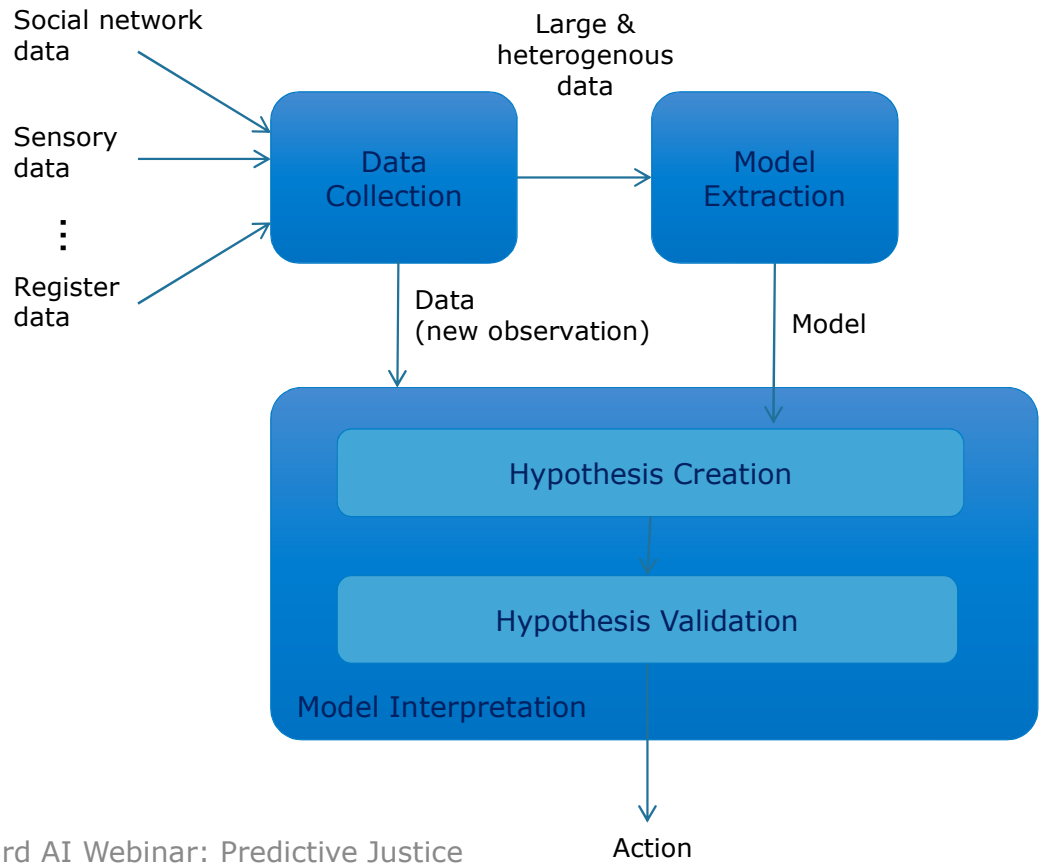
Consider the AI outcome as a central body of evidence and

- Extract a hypothesis like: Bob is a risky driver
- Search for evidences that **strengthen** the hypothesis like: Bob is a reckless man
- If enough evidences are found to strengthen the hypothesis; then the hypothesis is **accepted**

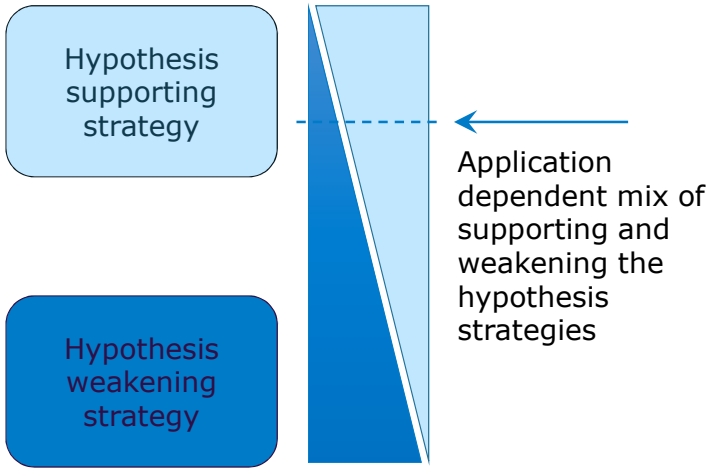
Self-fulfilling prophecy

- False hypothesis might become true
- Advantage: Reducing false negatives

Dealing with AI uncertainty



Which strategy to use?



Depends on the application and its

- **Costs/harms**
  - Impact of false positives and false negatives
  - The procedures to deal with false positives and false negatives
  - Searching for extra evidences
  - On affected individuals, affected groups, and the society
  - Before (ex-ante), during and after (ex-post, short term and long term) applying actions
- **Benefits:** Impact of true positives and true negatives

One threshold or two thresholds (when choosing for strategy 1 or strategy 2)

For in between cases: How to tailor the strategy to the application at hand?

Conclusion

Which strategy to use?

Statistical truth, and its uncertainty

- Stablished via induction
- Applied via deduction (naive way → responsible way)

Two promising strategies

- Self-denying prophecy
- Self-fulfilling prophecy

Still many challenges

- How to trade off costs/harms and benefits
- How to define threshold(s)
- How to devise strategies for in between cases