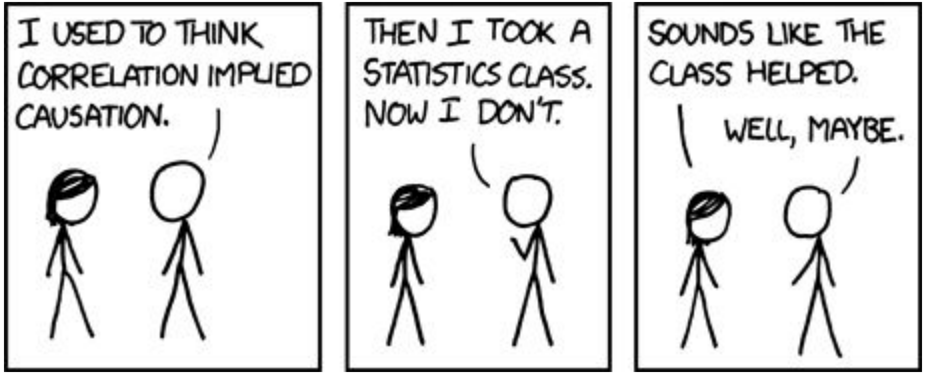
Causal Inference in ML Track

Track goal

Create causal inference guide for data scientists



Track organizer

i.o.goloshchapova@gmail.com

: @irina_g

: @irina_goloshchapova

- Head of Risks, Macro and Research at X5 Retail Group
- DS Lead at Manchester University
- PhD in Economics, doing data science since 2009:
    - 2009-2017: statistical analysis and modeling at CMASF
    - 2017-present: DS and Big Data projects in risks, macroeconomics and business processes optimization

Track Content

#1 - Causal Inference Introduction

Anton Lebedevich

Data science engineer with a background in backend performance optimization

#2 - Mastering Causal

Ivan Komarov

Chief Data Scientist @ CFT

#3 - Would you like a Cup of LATE?

Philipp Kartaev

Doctor of economics. Head of the Department of mathematical methods in economics, Lomonosov Moscow state University

#4 - Causal Inference with Panel Data

Dmitry Arkhangelsky

Associate professor (untenured) at CEMFI

#5 - Solving discrete optimization problems using continuous optimization

Artyom Gadetsky

Researcher in Bayesian methods group, PhD student at NRU HSE

#6 - Causal inference for a steel mill

Boris Voskresenskii

Chief Digital Officer at Severstal

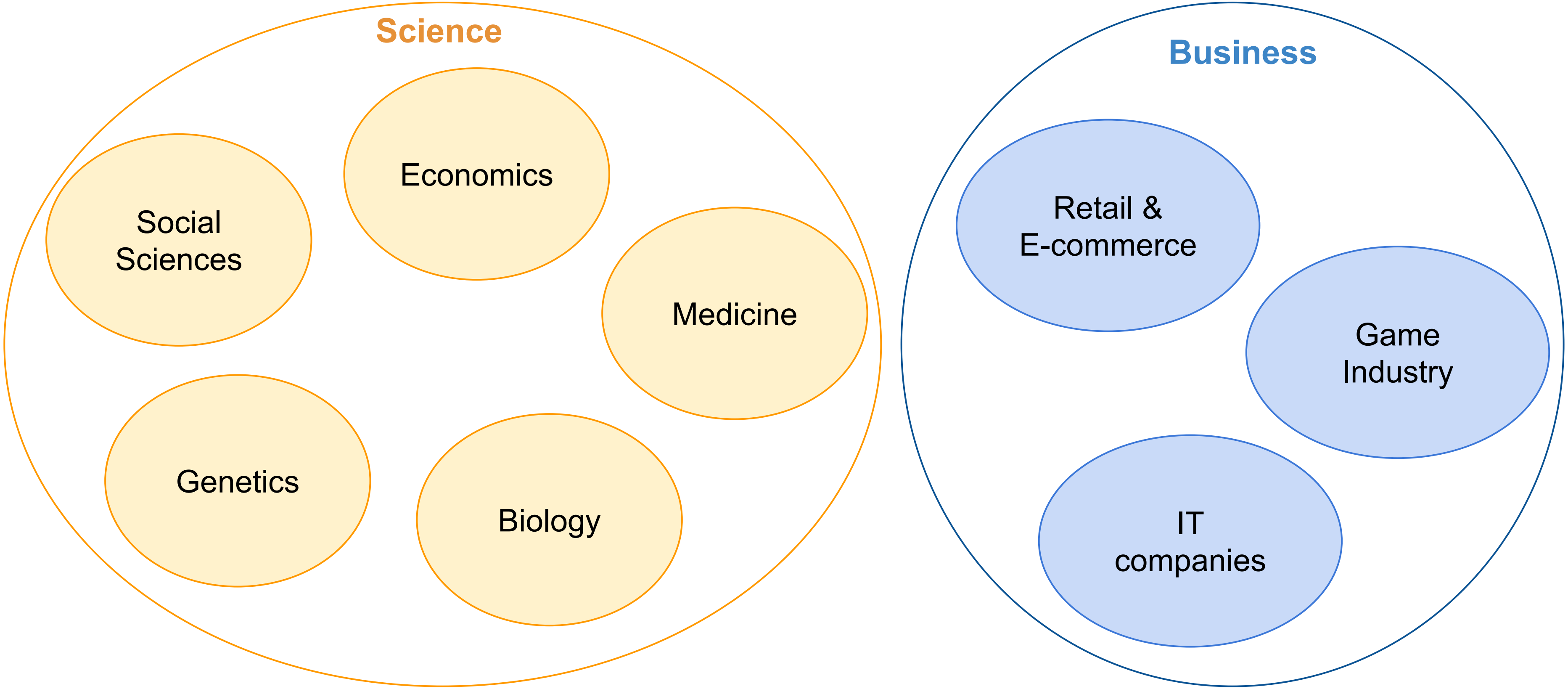
#7 - How to target the Uplift model for profit

Vahe Brsoyan

DS Team Lead at Big Data X5

Causal Inference Methods

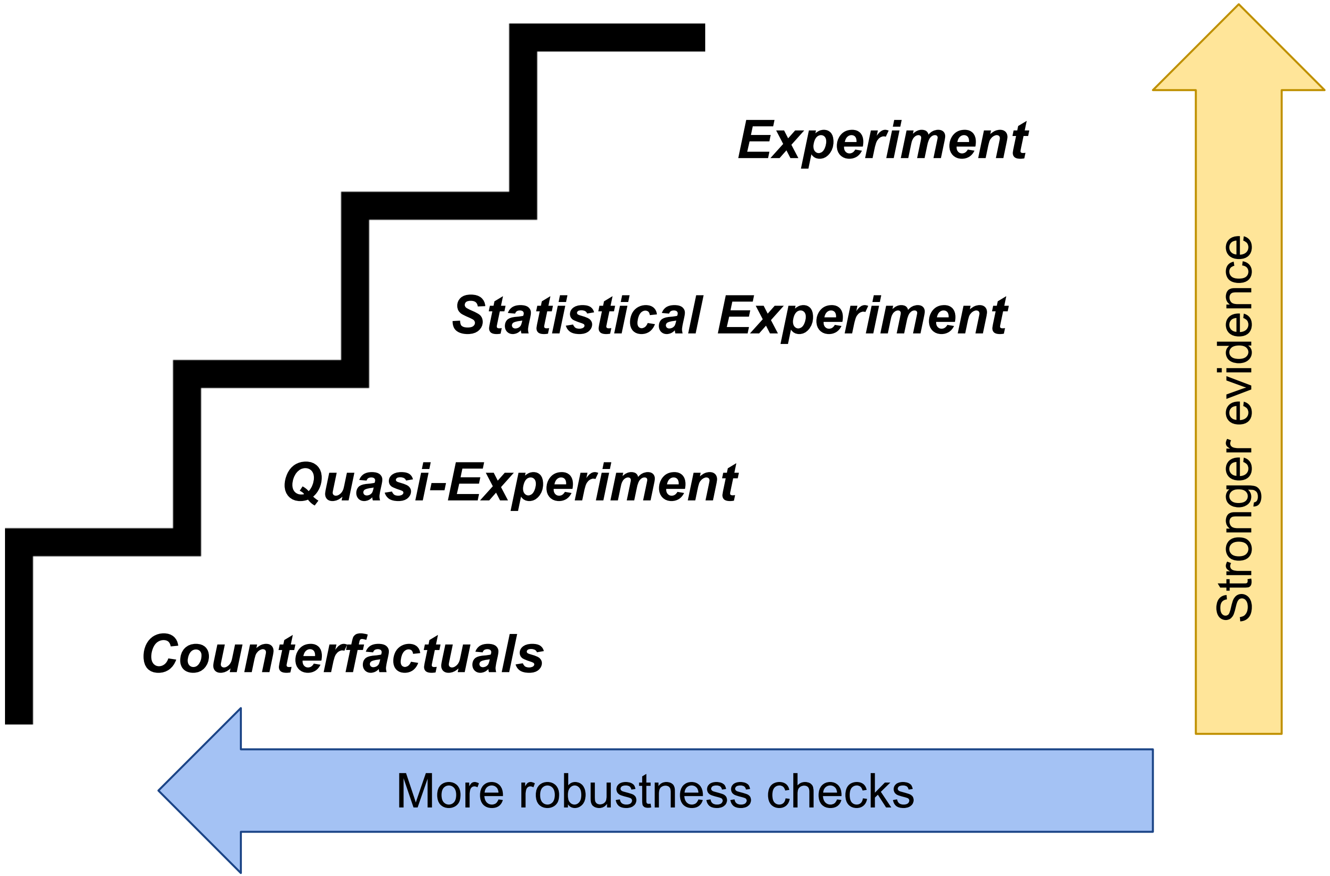
Need for causal inference



Definition

	Intervention	***Counterfactual***
	What happens to Y if I do X?	***Let's imagine Y if we did/didn't do X***
X causes Y iff	changing X leads to a change in Y	the world of Y with X is different from the one without X
A causal effect is	the magnitude by which Y is changed by a unit change in X	the magnitude by which Y with X is different from Y without X

Evidence Ladder



Source: [A.Rebecq (2020)]

Evidence Ladder

Expectation

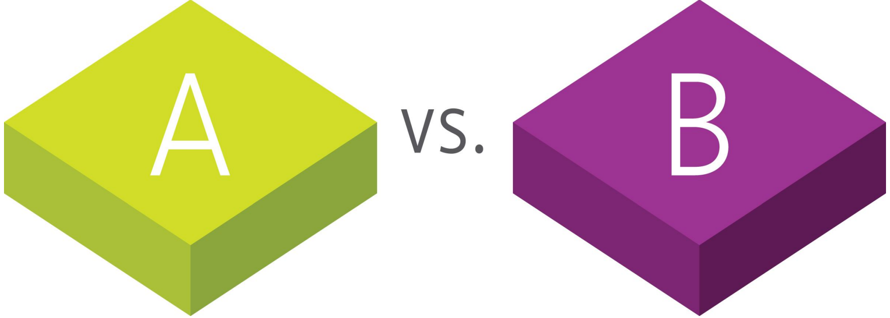
Reality

Experiments

The reference method

- ➔ Almost identical conditions for test and control groups - all is strictly identical but the X you're experimenting with
- ➔ Allows to conclude that X causes Y
- ➔ Typical in physics, chemistry
- ➔ Not applicable in social sciences and business cases

Statistical Experiment



A/B Testing or Randomised Controlled Trials

- ➔ Test and control groups are not identical but divided at random
- ➔ Comparison with some statistical significance and power
- ➔ Necessity for data science

Potential problems

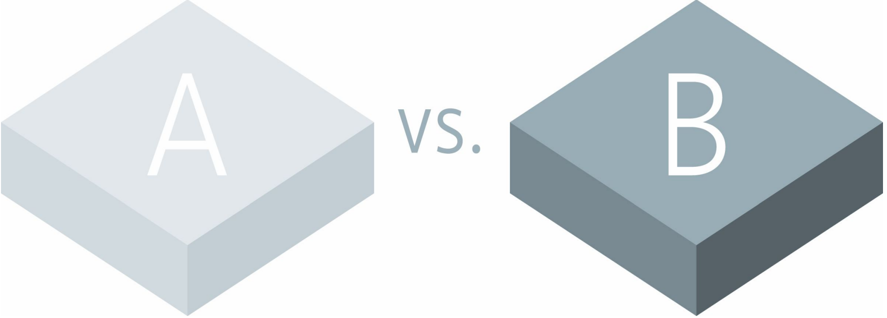
- ➔ Generalization ability
- ➔ Groups similarity
- ➔ External factors / confounders
- ➔ Spillover and network effects

Robustness checks

- ➔ Domain knowledge
- ➔ Causal graphical models - Backdoor criterion
- ➔ ...

Source: [Kohavi et al. (2009)]

Quasi-Experiment



Natural Experiments

- ➔ Test and control groups are not identical but similar - divided by natural criteria
- ➔ Methods to find the most similar objects
- ➔ Comparison with some statistical significance and power
- ➔ Necessity for data science

Methods

- ➔ Difference-in-Difference
- ➔ Matching / Propensity score
- ➔ Regression Discontinuity Design
- ➔ Instrumental Variables
- ➔ Doubly robust

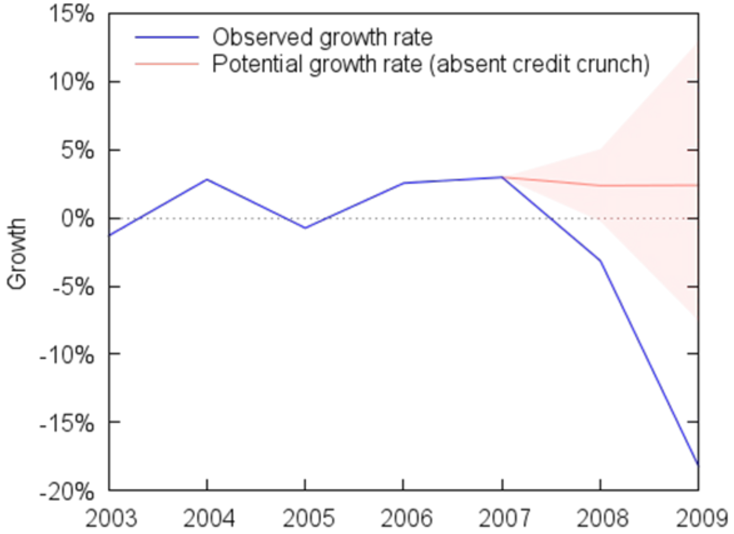
Potential problems

- ➔ All from A/B Testing
- ➔ Similarity through time
- ➔ Factors to calculate similarity
- ➔ Experiment design
- ➔ Randomisation in case of latent factors

Robustness checks

- ➔ Individual for each method

Counterfactuals



Experiments on observational data

➔ Without test and control groups
➔ Comparison with a modelled *counterfactual* control group
➔ Necessity for data science

Potential problems

➔ All from higher rungs of the ladder
➔ Quality of prediction
➔ Reveal underlying factors for the model

Robustness checks

➔ Individual for each method

Methods

➔ Synthetic diff-in-diff
➔ Athey & Imbens (2017): policy evaluation
➔ Structural time series models
➔ ...

The new frontier

- ➔ Discovery of causal relationships from data (Scholkopf et al, 2017)
- ➔ Heterogeneous treatment effects (Athey and Wager, 2015)
- ➔ Machine learning, representations, and causal inference
- ➔ Reinforcement learning and causal inference
- ➔ Automated causal inference

Source: Scharma and Kiciman (2018)

Best practices

➔ **Always follow the four steps: Model, Identify, Estimate, Refute**
Refute is the most important step

➔ **Aim for simplicity**
If your analysis is too complicated it is most likely wrong

➔ **Try at least two methods with different assumptions**
Higher confidence in estimate if both methods agree

Source: Scharma and Kiciman (2018)

Causal Inference in ML Track