# Hypotheses and Models in Data Intensive Domains @hamc2023, Masters 1/2

**Dmitry Kovalev**

Lomonosov Moscow State University
Federal Research Center “Computer Science and Control”
of the Russian Academy of Sciences

*https://github.com/dmkovalev/hamc2023*

# Course outcomes

- **Objective** - overview hypothesis-driven approach and the skills needed to do empirical research in data-intensive domains
- **Aim** - provide students with techniques and receipts for applying statistical/probabilistic framework to form and assess hypotheses and models
- The course will also emphasize recent developments in hypothesis management and will present some open questions and areas of ongoing research

# Course overview and positioning

МГУ ВМК

Информатика и Управление

- Topics covered include overview of different approaches to hypotheses and models:
  - Formulation and representation
  - Statistical tests, probabilistic inference
  - Ranging/ranking using metrics-based approaches
  - Hypotheses/models and ML/DL with data-intensive examples on Spark
  - Causality and graph models

- This course is part of a sequence of courses on Big Data track and is taught for 1st and 2nd year masters students.

# How students time is spent

- 2 hours per week - lectures
- 4 hour per week – homework

*Course website:*

https://github.com/dmkovalev/hamc2023

# Tools to be used

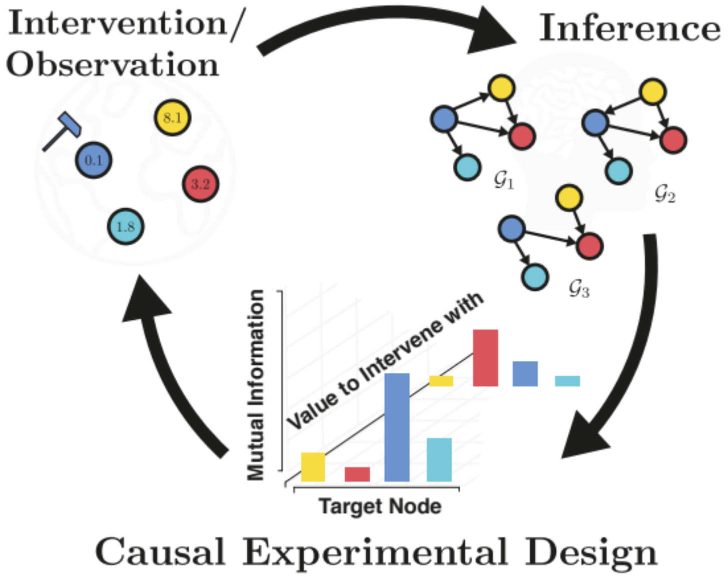
- Python with following libs:
  - Statsmodels
  - Scipy
  - Sympy
  - PySindy
  - PySpark
- Jupyter notebooks/docker container will be provided with instructions

# Assessment

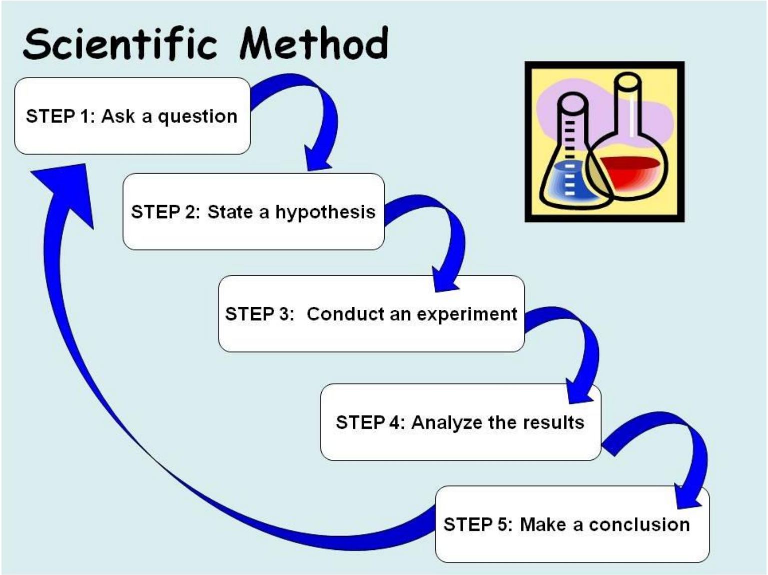
- 50% - Final Oral Exam
- 50% - Homeworks
- grade
  - 5: 80 - 100%;
  - 4: 60 - 79%;
  - 3: 40 - 59%;
  - <3: 0 - 39%.
- Instructors
  - Dmitry Kovalev

# Importance of scientific method in data-intensive sciences

- Industry vs Academic research
- Emergence of industry labs
- Open access to papers
- AI/ML/Statistics tools



*From P. Tigas et al. (2022)*
*Interventions, Where and How? Experimental Design for Causal Models at Scale*



Scientific Method

STEP 1: Ask a question

STEP 2: State a hypothesis

STEP 3: Conduct an experiment

STEP 4: Analyze the results

STEP 5: Make a conclusion

# Importance of scientific method in data-intensive sciences

**M. Jordan:** *... current focus on doing AI research via the gathering of data, the deployment of "deep learning" infrastructure, and the demonstration of systems that mimic certain narrowly-defined human skills—with little in the way of emerging explanatory principles— tends to deflect attention from major open problems in classical AI. These problems include the need to bring meaning and reasoning into systems that perform natural language processing, the need to infer and represent causality, the need to develop computationally-tractable representations of uncertainty and the need to develop systems that formulate and pursue long-term goals. These are classical goals in human-imitative AI, but in the current hubbub over the "AI revolution," it is easy to forget that they are not yet solved.*

# Importance of scientific method in data-intensive sciences

**M. Brodie:** *We have to have error bars around all our predictions ... Otherwise it's gambling, and too many failed predictions can lead to big disappointment with Big Data - a Big Data Winter.*
*Yet there is a potential Big Data Winter ahead if people blindly apply Big Data and more specifically machine learning.*

# Importance of scientific method in data-intensive sciences

***Causality*** *has a long history, providing it with many principled approaches to identify a causal effect (or even distill cause from effect). However, these approaches are often restricted to very specific situations, requiring very specific assumptions. This contrasts heavily with recent advances in machine learning. Real-world problems aren’t granted the luxury of making strict assumptions, yet still require causal thinking to solve. Armed with the rigor of causality, and the can-do-attitude of machine learning, we believe the time is ripe to start working towards solving real-world problems.*

# Importance of scientific method in data-intensive sciences

Дмитрий Ветров: ChatGPT глазами ученого. Когда будет создан искусственный интеллект?

- https://www.youtube.com/watch?v=IMP1zZ9K4Wc

ChatGPT is 'not particularly innovative,' and 'nothing revolutionary', says Meta's chief AI scientist

https://www.zdnet.com/article/chatgpt-is-not-particularly-innovative-and-nothing-revolutionary-says-metas-chief-ai-scientist/