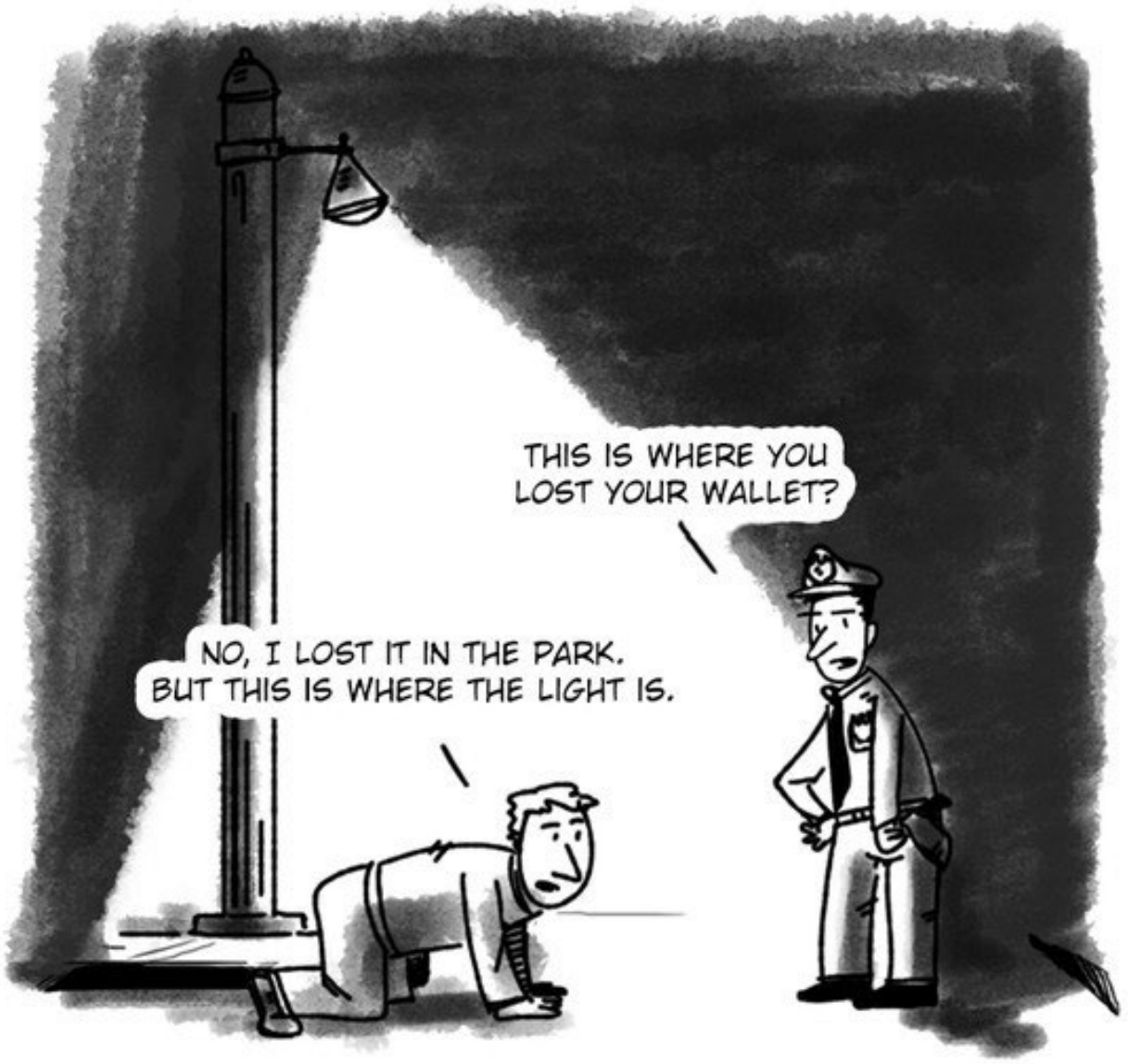
Введение

Структура курса

- Лекции – основы методов
- Семинары – реализация в коде, обсуждение
- Практикумы – самостоятельная реализация в коде

- Темы – химические данные, машинное обучение, генеративные модели; молекулярные данные, кристаллические данные; python, фреймворки, оптимизация кода.

Motivation


THIS IS WHERE YOU LOST YOUR WALLET?
NO, I LOST IT IN THE PARK. BUT THIS IS WHERE THE LIGHT IS.

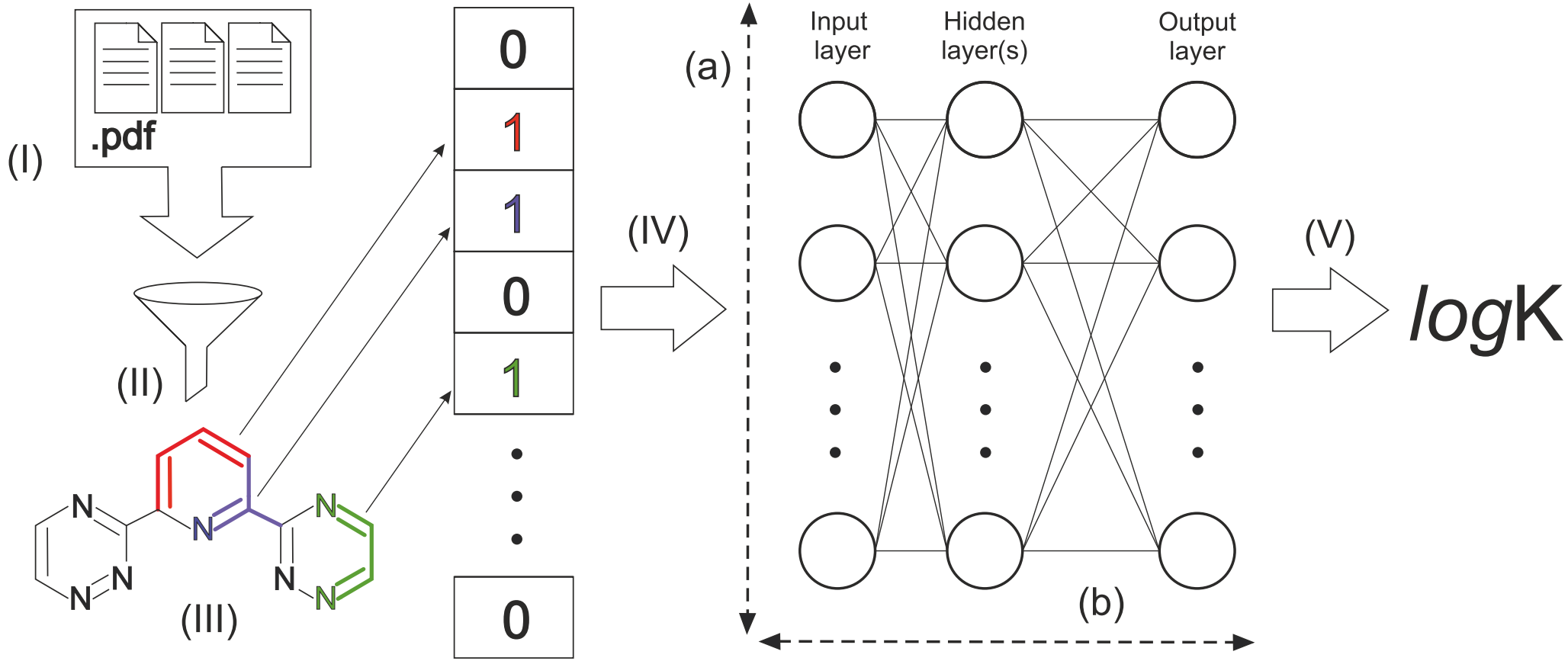
Artificial intelligence

- “Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.”
- “Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans. Leading AI textbooks define the field as the study of "intelligent agents": any system that perceives its environment and takes actions that maximize its chance of achieving its goals.”
- “Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data.”

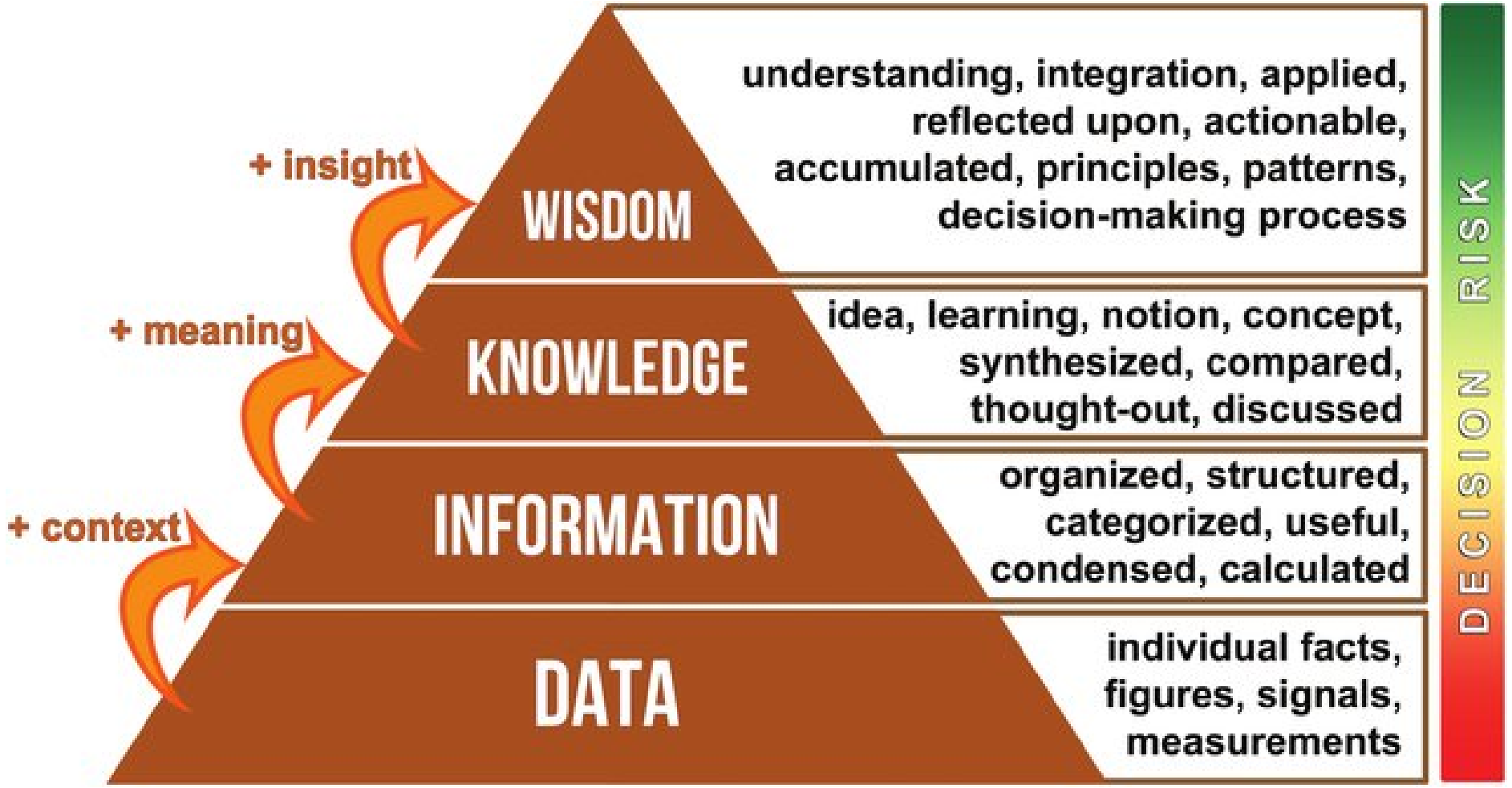
General principles

- Данные: сбор, очистка, разметка
- Задачи: классификация, регрессия, кластеризация, генерация, оптимизация, анализ
- Методы: «классическое» машинное обучение, нейронные сети, эволюционные алгоритмы, роевой интеллект
- Human-readable vs. machine-readable
- Функции потерь и метрики качества: классификационные, регрессионные, подобия
- Тестовые наборы данных: как избежать переобучения?
- Область применимости модели

General principles: QSPR example

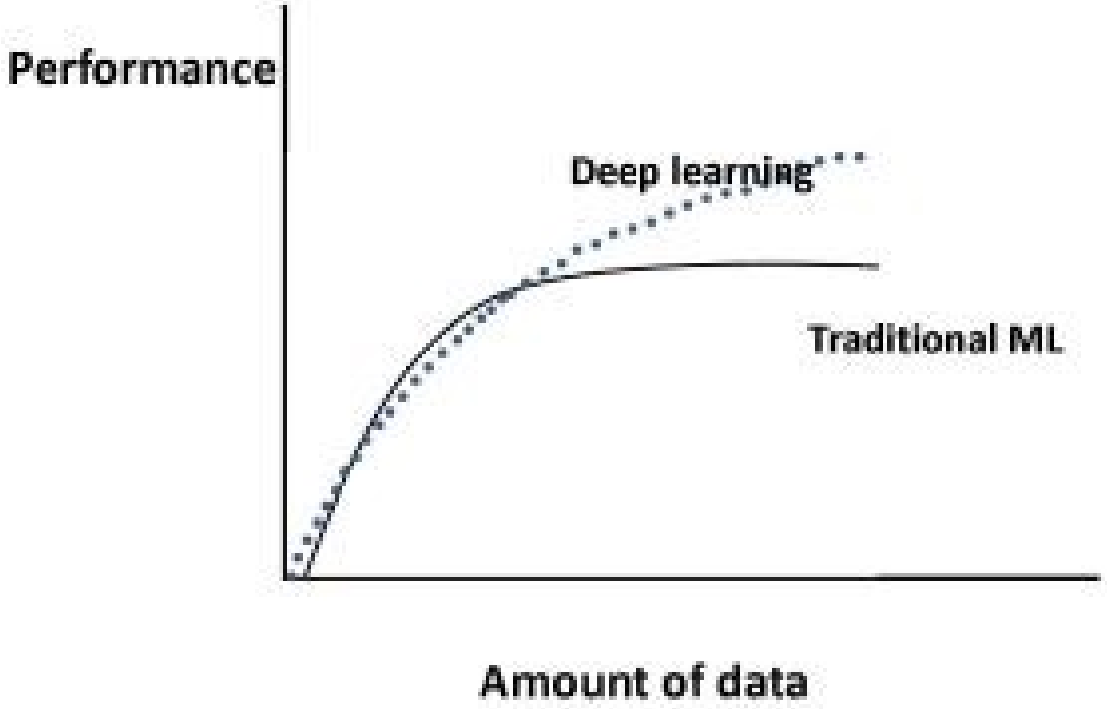


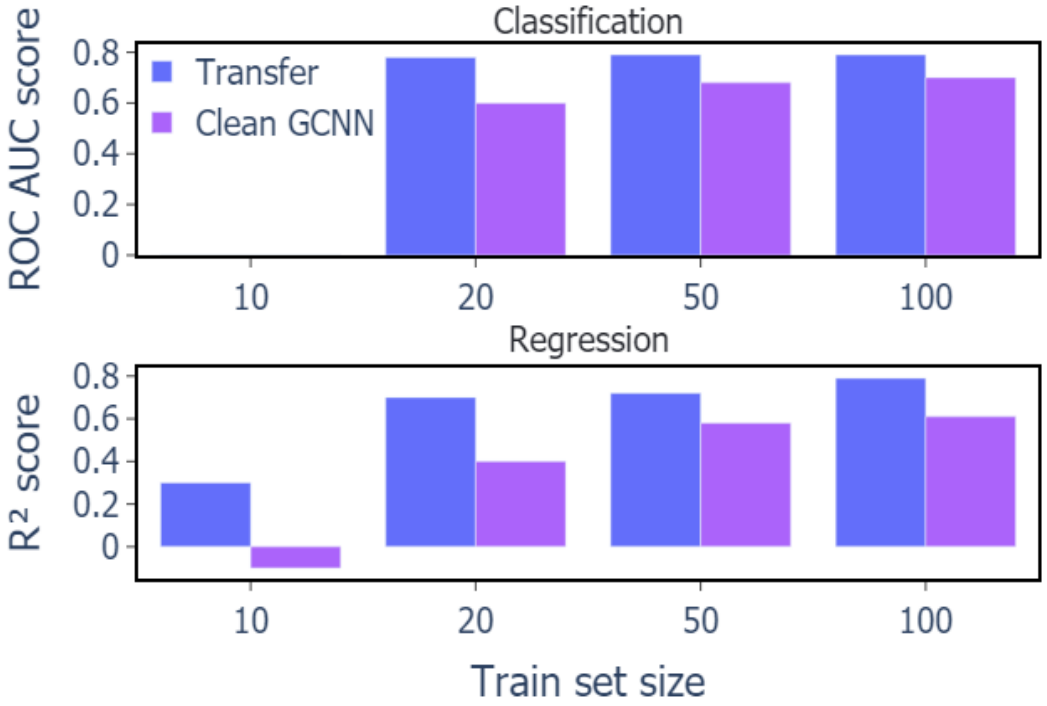
Big data



DOI: 10.1093/af/vfz005

Big/Small data?





towardsdatascience.com
DOI:10.1021/acs.jpclett.1c02477

Feature vectors

- Могут представлять собой данные (численные)
- Скорее всего, придется векторизовать данные других типов – текст, картинки, структурные формулы:
  - Записываем признаковое описание – много разных тестов
  - Записываем структуру – кодируем 2/3D разного размера в 1/2D постоянного размера
  - Предлагаем алгоритму сделать это за нас
- Проверяем масштабирование (scaling)
- Формируют feature space
- Определяют Applicability domain

Data: Summary

- Основные определения: data science, artificial intelligence, machine learning
- Переход количества в качество
- Нужно предварительно обрабатывать
- Могут быть размеченные и неразмеченные

Tasks

- Обучение с учителем:
  - Классификация
  - Регрессия
- Обучение без учителя
  - Кластеризация
- Обучение с подкреплением:
  - Генерация

Methods

Machine learning:

- Naïve Bayes
- Support vector machine
- Nearest neighbors
- Decision tree/Random forest/XGBoost
- Linear regression
- Deep learning

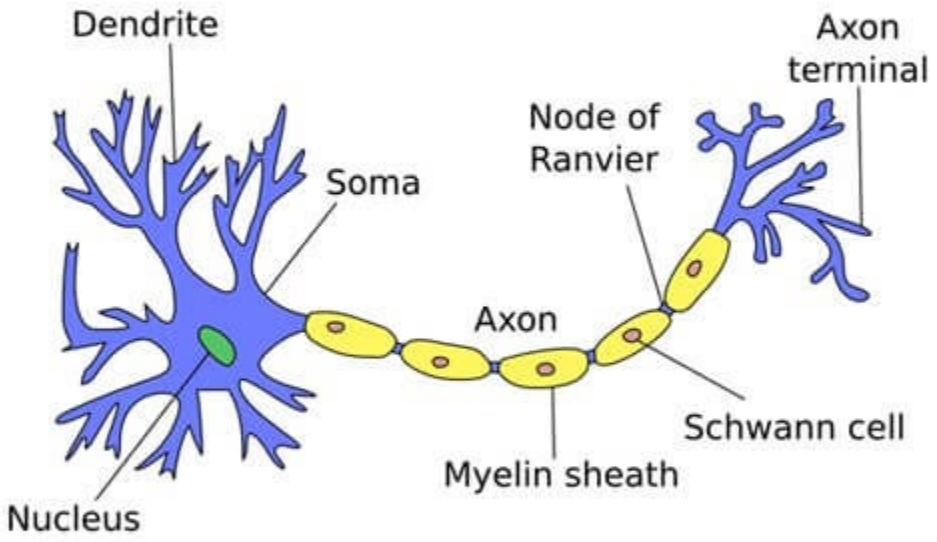
Generative:

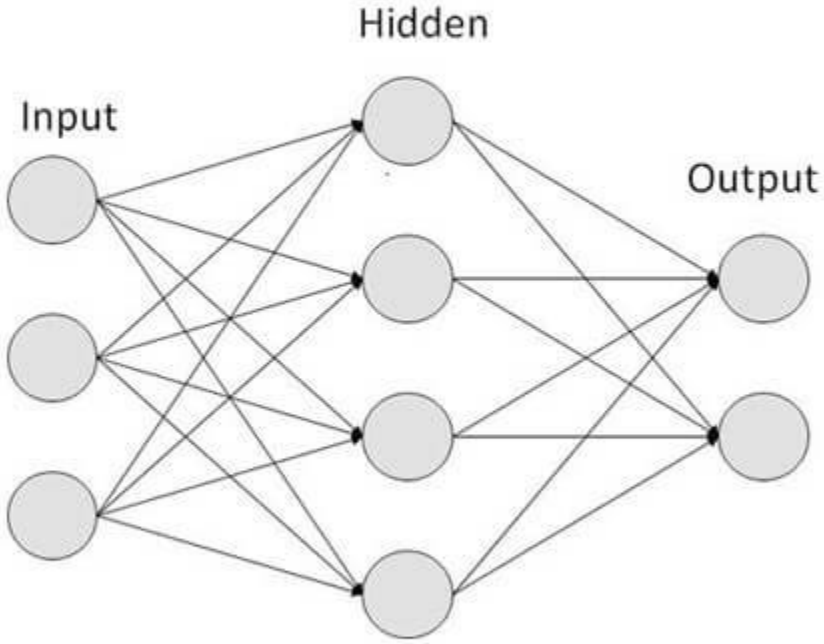
- GAN
- Evolutionary algorithms
- Swarm Intelligence

Other:

- Autoencoders
- PCA/t-SNE

Neural networks

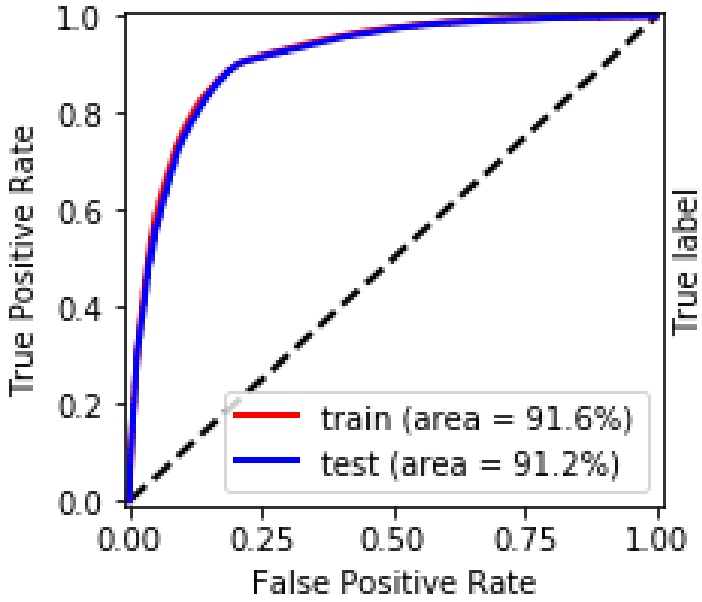


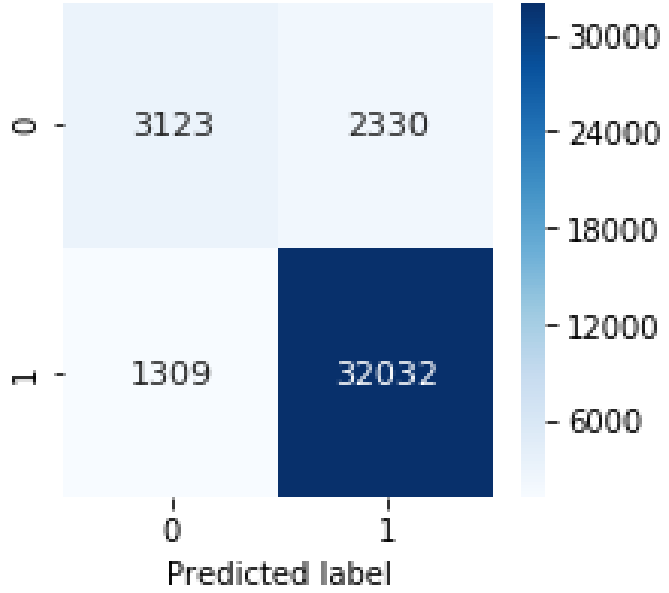


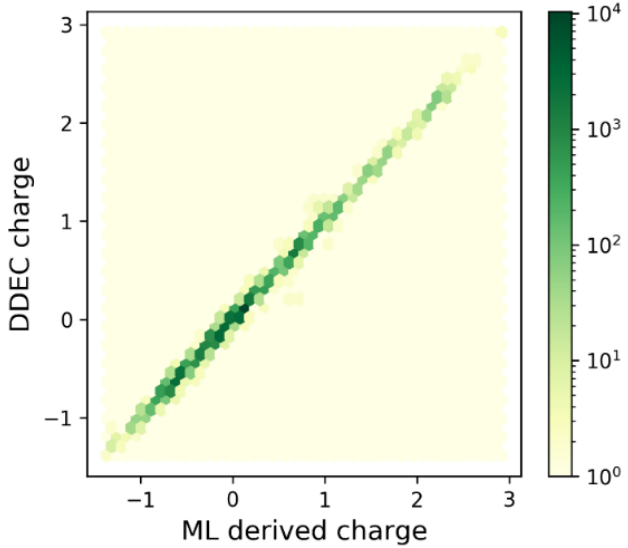
Metrics/Loss functions

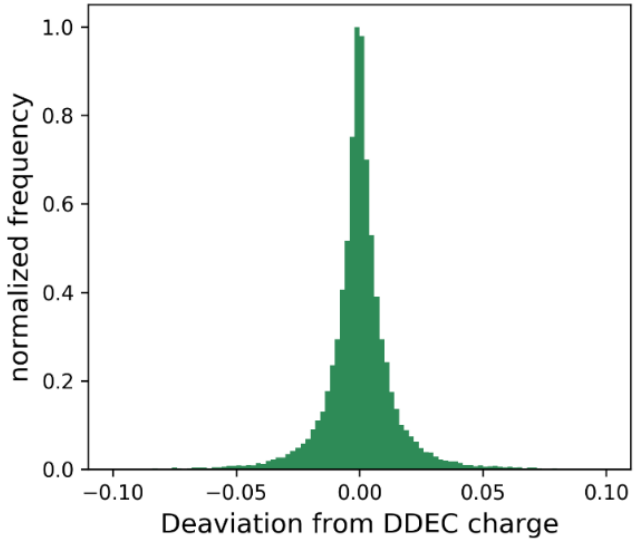
- Функция потерь – оптимизируется (минимизируется) в задаче:
  - Определена для типа задачи (классификация/регрессия/генерация)
  - Должна учитывать несбалансированность набора данных
  - Часто должна быть дифференцируема
  - Часто должна иметь физический/химический смысл
  - Может быть составной
  - Может быть связана с метрикой
- Метрика – оценивается после обучения:
  - Определена для типа задачи (классификация/регрессия/генерация)
  - Должна учитывать несбалансированность набора данных
  - Часто должны иметь физический/химический/логический смысл
  - Может быть составной
  - Лучше, когда их несколько

Metrics: examples





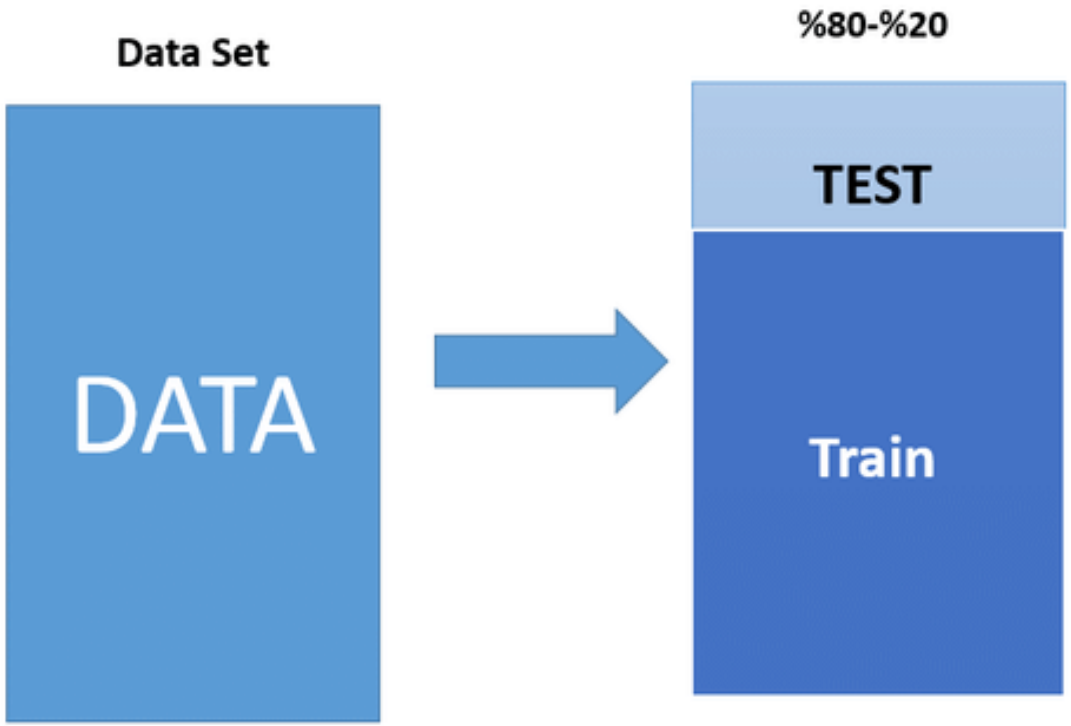




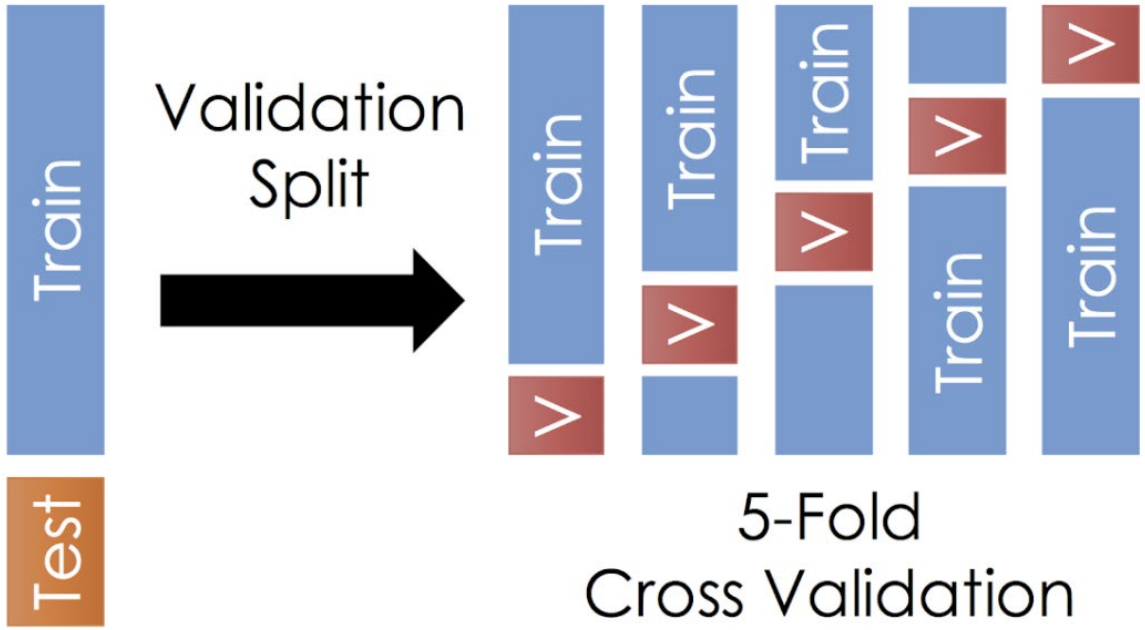
Quality: Summary

- Нужна функция, которую мы хотим оптимизировать
- Нужны метрики, которые мы хотим демонстрировать миру

External testing



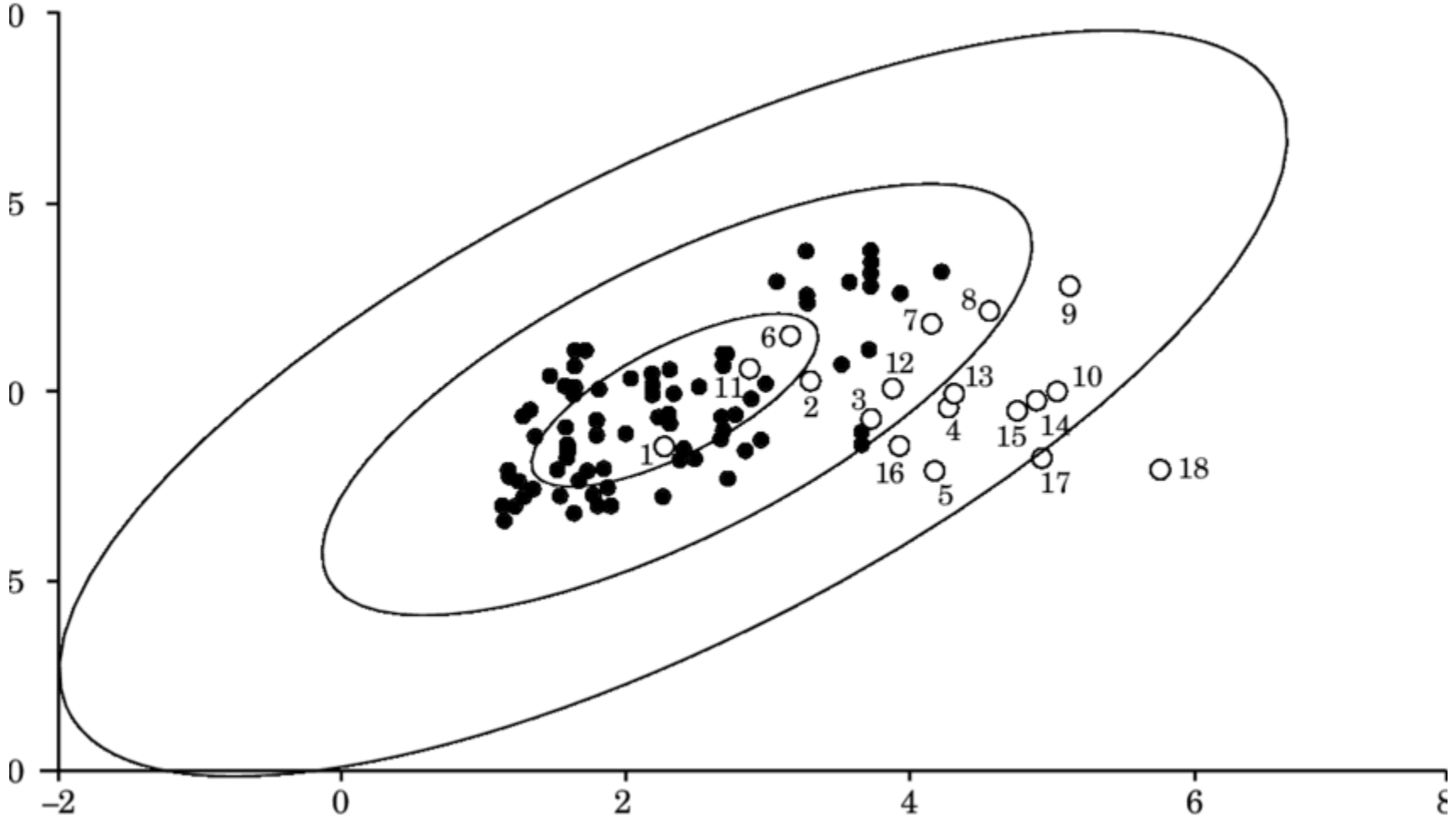
Internal/external testing: cross-fold validation



Итого: деление на три части – train, validation (dev), test

ISBN: 9781491912058

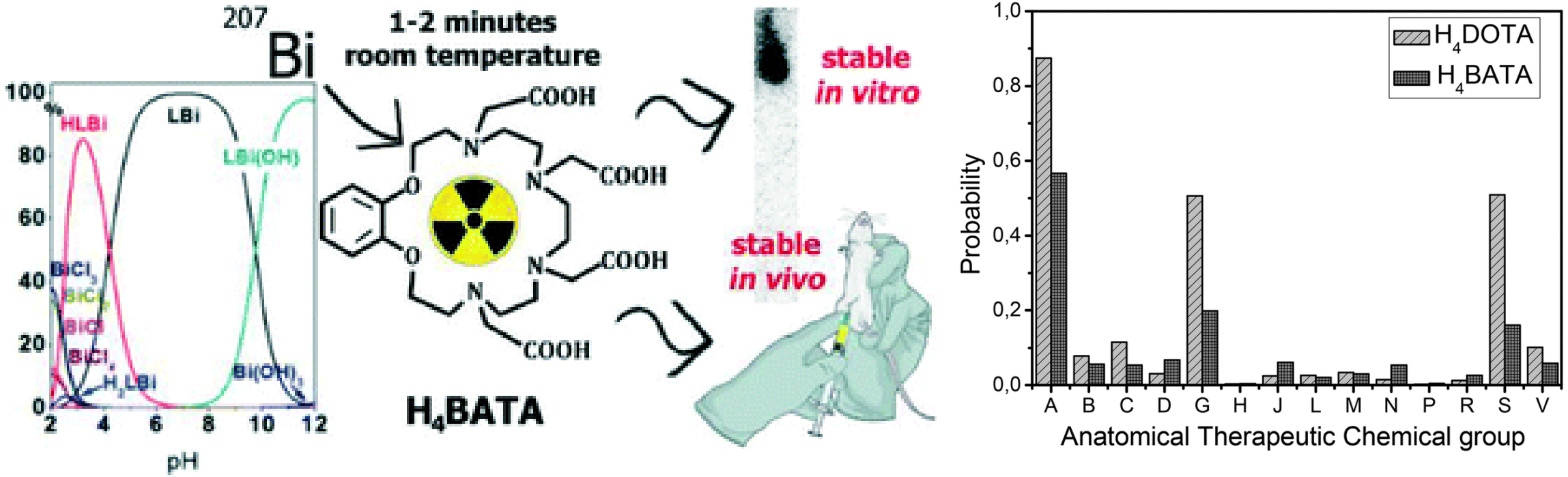
Applicability domain



DOI: 10.1177/026119290503300209

Examples

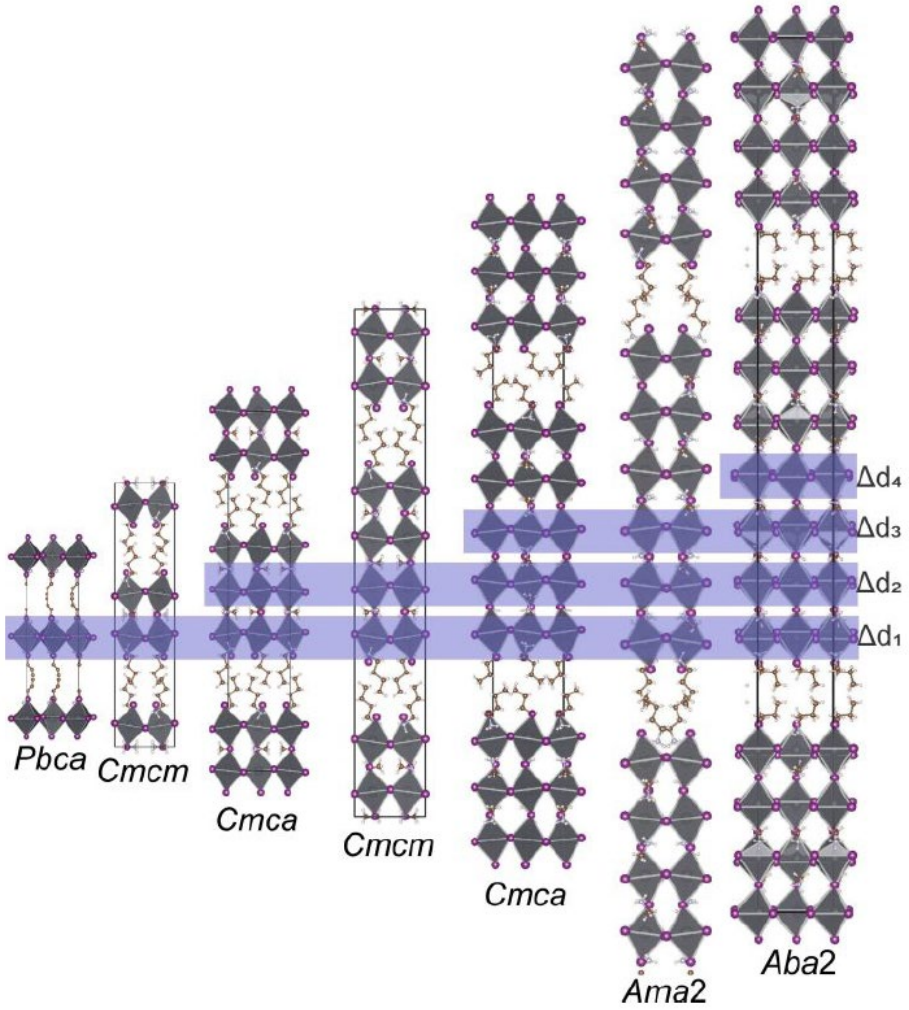
AI for the radiopharmaceuticals design

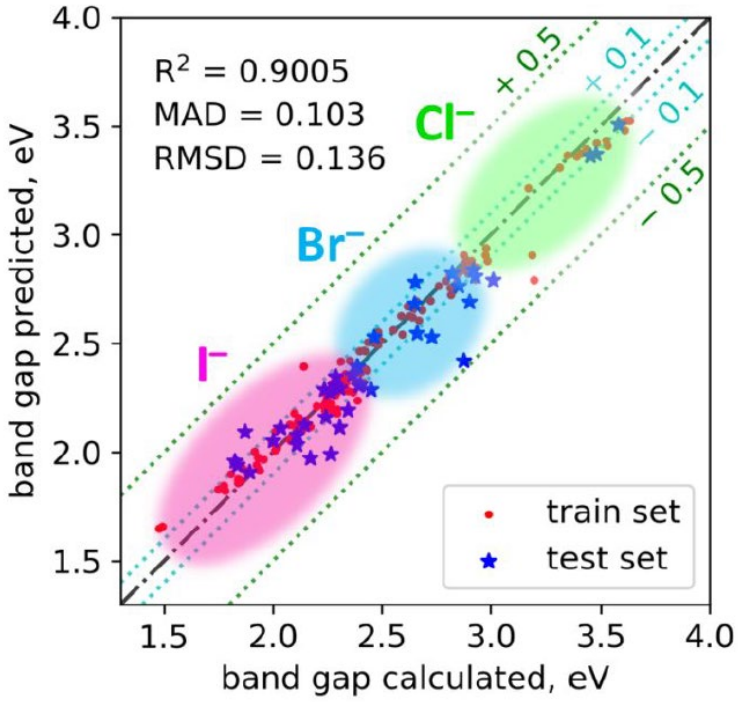


E. V. Matazova, B. V. Egorova, E. A. Konopkina, G. Y. Aleshin, A. D. Zubenko, A. A. Mitrofanov, K. V. Karpov, O. A. Fedorova, Y. V. Fedorov, and S. N. Kalmykov, “Benzoazacrown compound: a highly effective chelator for therapeutic bismuth radioisotopes,” *MedChemComm*, vol. 10, pp. 1641–1645, 2019

B. V. Egorova, E. V. Matazova, A. A. Mitrofanov, G. Y. Aleshin, A. L. Trigub, A. D. Zubenko, O. A. Fedorova, Y. V. Fedorov, and S. N. Kalmykov, “Novel pyridine-containing azacrown-ethers for the chelation of therapeutic bismuth radioisotopes: complexation study, radiolabeling, serum stability and biodistribution,” *Nuclear Medicine and Biology*, vol. 60, pp. 1–10, 2018

AI for the materials

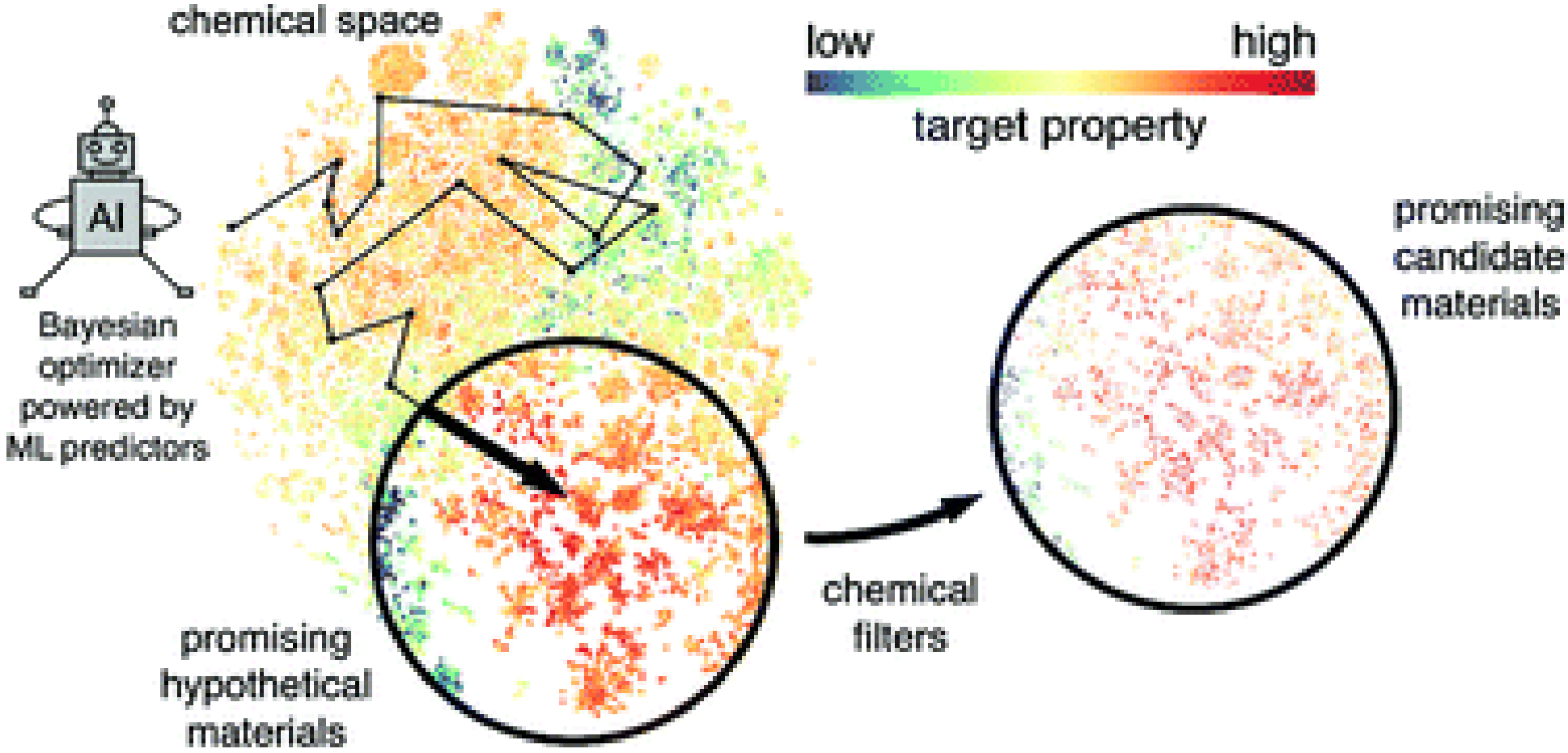




E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin, and A. B. Tarasov, “Database of 2d hybrid perovskite materials: open-access collection of crystal structures, band gaps and atomic partial charges predicted by machine learning,” *Chemistry of Materials*, , 2020

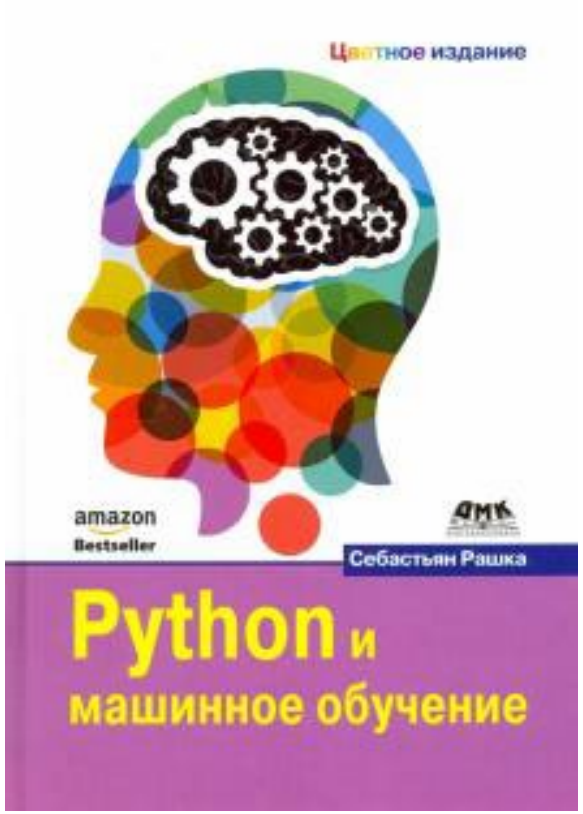
V. V. Korolev, A. Mitrofanov, E. I. Marchenko et al. Transferable and extensible machine learning derived atomic charges for modeling hybrid nanoporous materials // *Chemistry of Materials*. , American Chemical Society (United States), DOI: 10.1021/acs.chemmater.0c02468, 2020

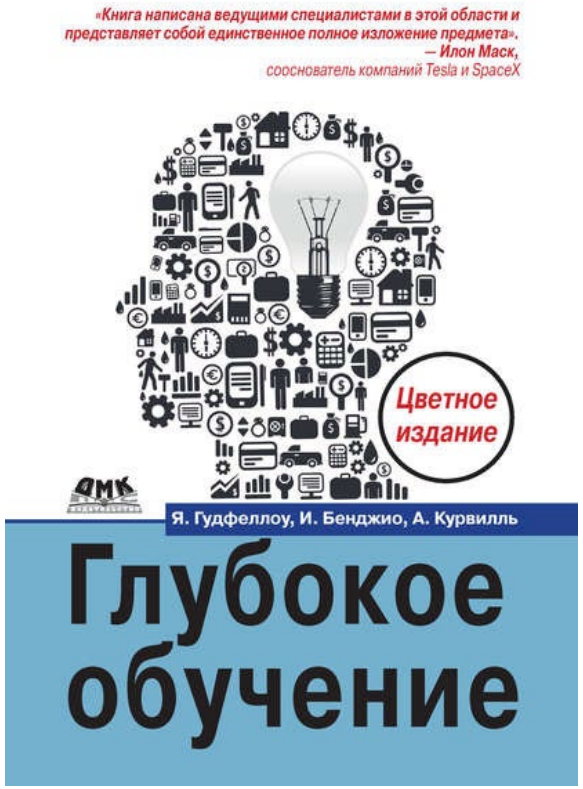
AI proposes materials



V. Korolev, A. Mitrofanov, A. Eliseev, and V. Tkachenko, “Machine-learning-assisted search for functional materials over extended chemical space,” *Materials Horizons*, 2020
https://nplus1.ru/news/2020/08/27/my-chemical-space

Try it yourself:





- https://scikit-learn.org/stable/tutorial/basic/tutorial.html
- https://www.tensorflow.org/tutorials/keras/classification?hl=ru