Математики в IT. Data Scientist

Зражевський Олексій Григорович

PhD, Sr. Data Scientist in ML department, Quantfury

Популярні спеціальності на українському IT ринку для математиків

- Data Engineer
- Data Scientist
- Data Analyst
- Machine Learning Engineer/ Researcher

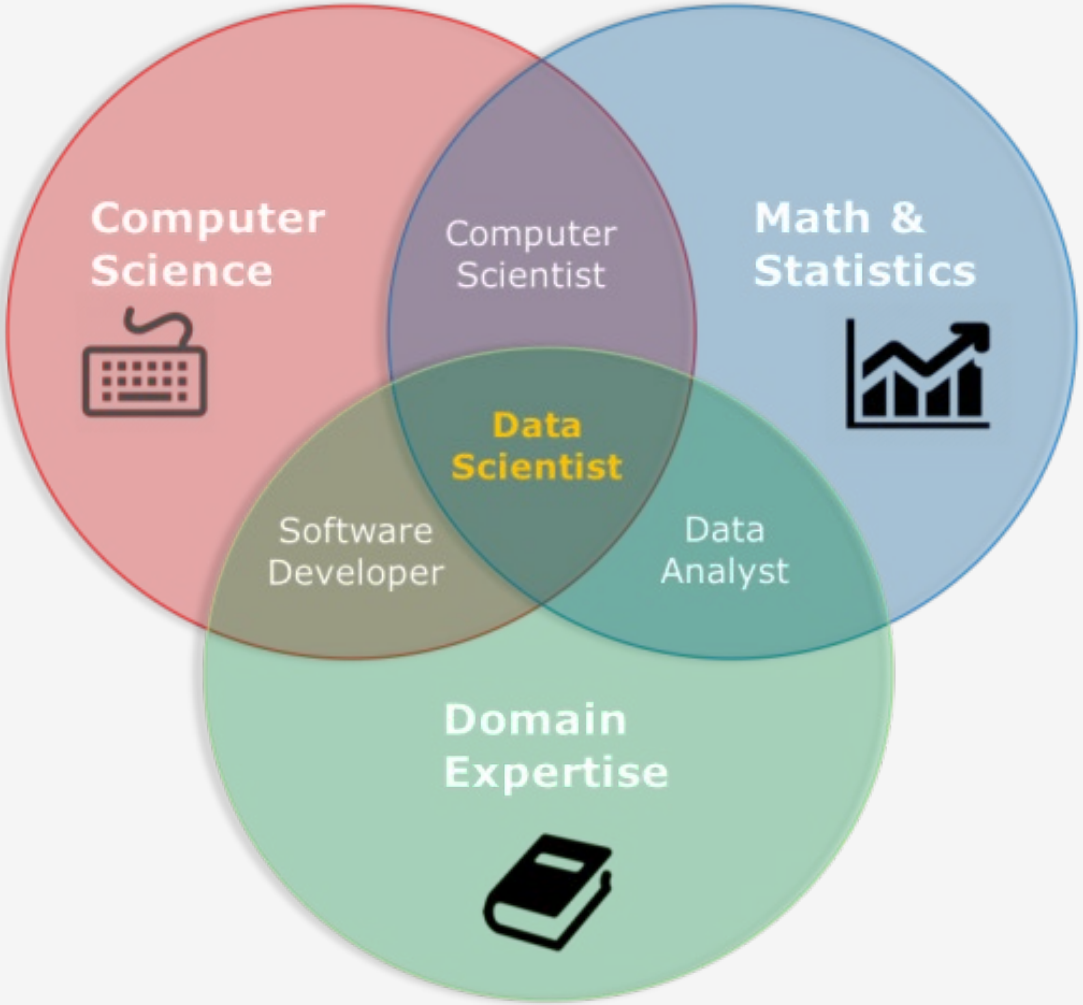
Загальні вимоги, функціональні обов’язки, відмінності спеціалізацій

- Data Engineer
  - організація збору даних;
  - організація зберігання даних: SQL та NoSQL-бази, Data Lakes,…;
  - трансформація даних.

Загальні вимоги, функціональні обов’язки, відмінності спеціалізацій

- Data Scientist
  - Exploratory Data Analysis (EDA): аналіз даних з метою виявлення трендів і залежностей;
  - створення моделей прогнозу важливих метрик;
  - автоматизація системи прогнозів: створення Model Pipelines;
  - Model Maintenance: моніторинг і підтримка усіх Model Pipelines;
  - створення вимог до інженерних команд (Data Engineer) щодо збору даних.

Як утворилась спеціалізація?



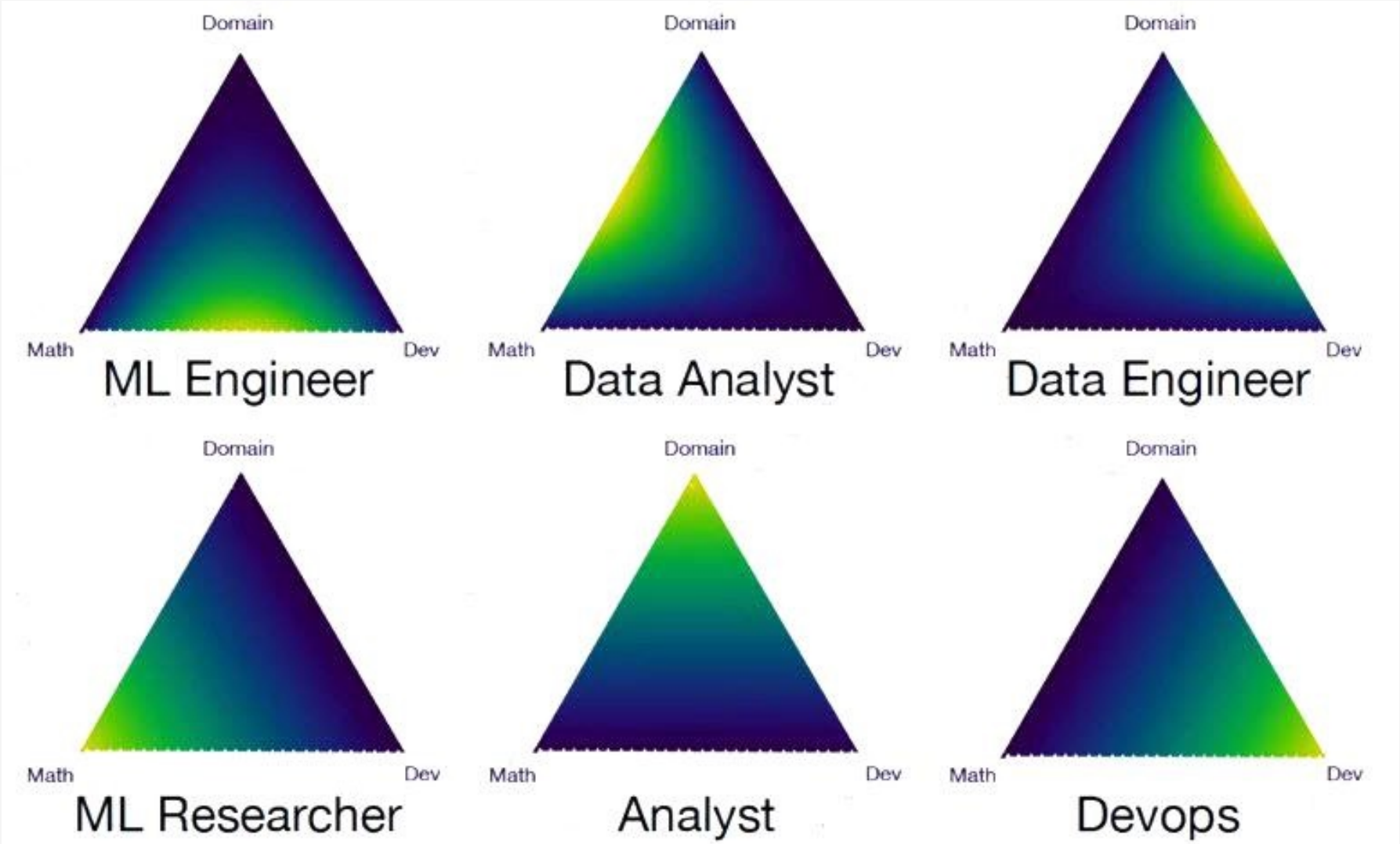
Спеціалізації всередині Data Science

- Data Analyst:
  - робота з даними: SQL, Big Data, тощо;
  - Exploratory Data Analysis: аналіз даних з метою виявлення трендів і залежностей;
  - розуміння домену походження даних та бізнес-контексту.

Спеціалізації всередині Data Science

- Machine Learning Engineer/Researcher:
  - знання моделей машинного навчання
  - розгортання моделей машинного навчання та їх інтеграція з іншими системами
  - розробка нових методів та алгоритмів машинного навчання.

Спеціалізації всередині Data Science



Нішеві спеціалізації ML Engineer за доменом

- specialist in Natural Language Processing (NLP)
- specialist in object detection
- web analytics
- game data analytics

Machine Learning

Машинне навчання (ML) — це підгалузь штучного інтелекту, яка застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» з даних.

Методи ML:

- навчання з учителем;
- навчання без учителя;
- навчання з підкріпленням.

Основні задачі ML

- класифікація (навчання з учителем);
- кластеризація (навчання без учителя);
- регресія (навчання з учителем);
- зниження розмірності даних та їх візуалізація (навчання без вчителя);
- відновлення щільності розподілу ймовірності набору даних;
- побудова рангових залежностей;
- виявлення аномалій.