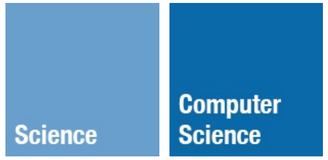
The Road to Data Science



CS 100: Introduction to the Profession

Matthew Bauer & Michael Saelee

ILLINOIS INSTITUTE OF TECHNOLOGY
College of Science

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

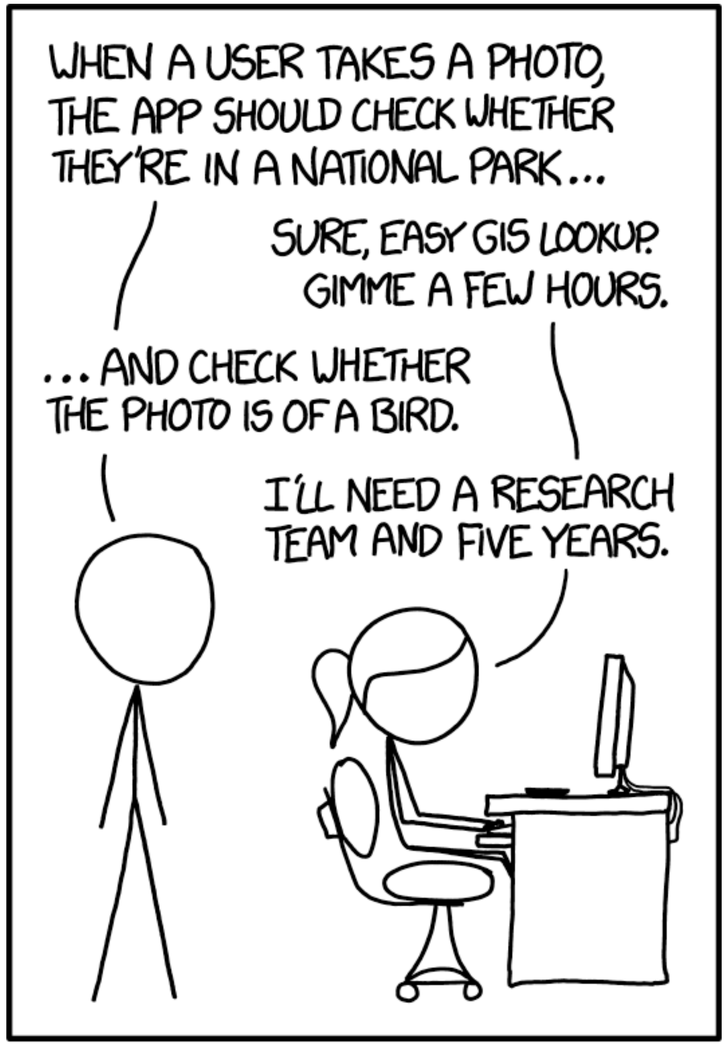
What is Data Science, and how is it related to Computer Science?

§ Artificial Intelligence

When programmable computers were first conceived, people wondered whether such machines might become intelligent, over a hundred years before one was built (Lovelace, 1842). Today, artificial intelligence (AI) is a thriving field with many practical applications and active research topics. We look to intelligent software to automate routine labor, understand speech or images, make diagnoses in medicine and support basic scientific research.

In the early days of artificial intelligence, the field rapidly tackled and solved problems that are intellectually difficult for human beings but relatively straightforward for computers — problems that can be described by a list of formal, mathematical rules. The true challenge to artificial intelligence proved to be solving the tasks that are easy for people to perform but hard for people to describe formally — problems that we solve intuitively, that feel automatic, like recognizing spoken words or faces in images.

- Goodfellow, Bengio, Courville. *Deep Learning*.



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

“The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”

- John McCarthy, 1955

“The art of creating machines that perform functions that require intelligence when performed by people”

- Ray Kurzweil, 1990

“A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes”

- Robert Schalkoff, 1990

“The study of how to make computers do things at which, at the moment, people are better.”

- Elaine Rich and Kevin Knight, 1991

“AI is whatever hasn't been done yet.”

- Douglas Hofstadter, paraphrasing Tesler's Theorem

“Every time we figure out a piece of it, it stops being magical; we say, ‘Oh, that’s just a computation.’”

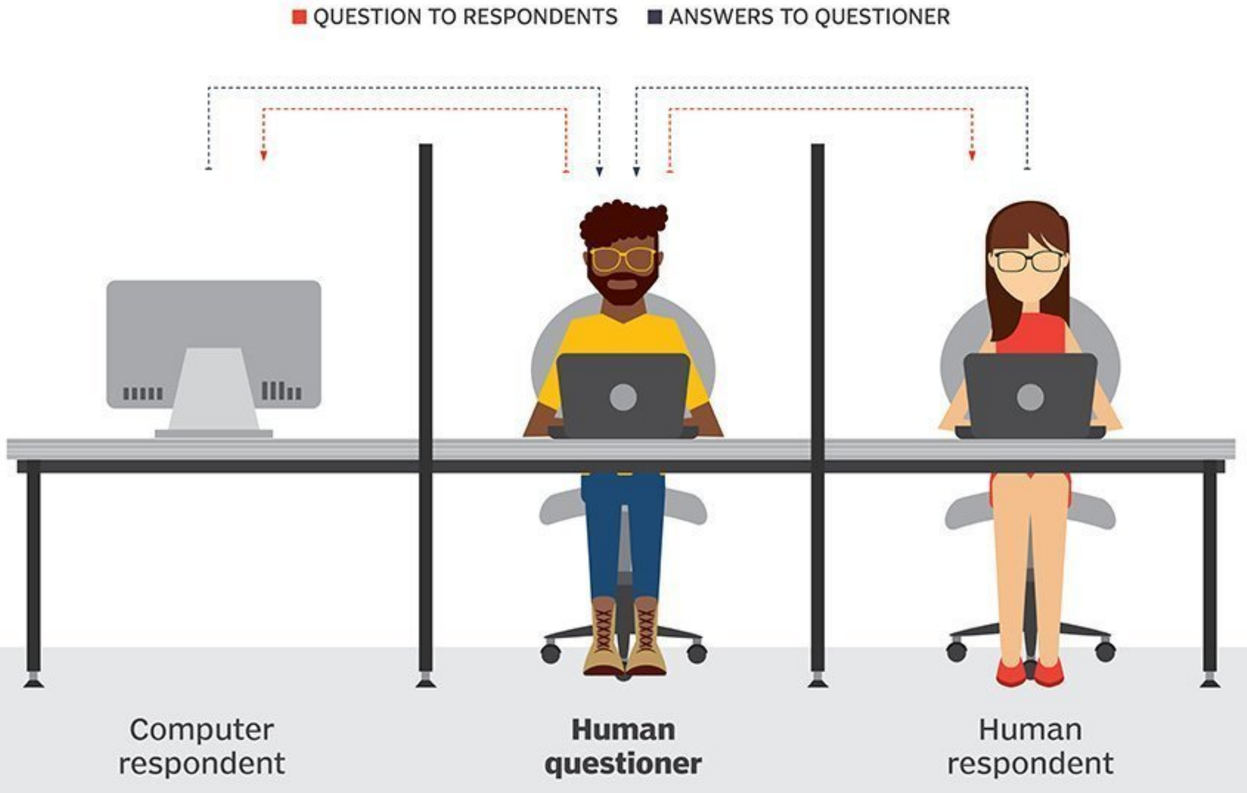
- Rodney Brooks

“Strong” AI

- Program computers to “think” like humans
  - Thinking vs. Intelligence vs. Consciousness
- Grail: a general purpose artificial intelligence that can learn to solve arbitrary tasks, as a human would
- How to identify a thinking program?

Turing test

During the Turing test, the human questioner asks a series of questions to both respondents. After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.



1950s-60s: Great Expectations

- Computer scientists widely expected general purpose strong AIs to be developed within a decade
  - Era of “Good old-fashioned AI” (GOFAI) — mostly symbolic programming aided by search & logic
- FAIL!
- Led to “AI Winter” of 70s & 80s

“Weak” AI

- Narrower definition of intelligence
- 1980s-Present
  - Less focus on GOFAI, more on application-specific AI

AI problem characteristics

- Fully observable vs. Partially observable
- Single/Multi/Adversarial agent
- Deterministic vs. Stochastic
- Static vs. Dynamic
- Discrete vs. Continuous

CS vs. AI

- Algorithms vs. Heuristics
- Exact vs. Approximate
- Optimal vs. Close
- All AI is CS; not all CS is AI

AI Tools

- Symbolic programming
- Search
- Formal logic
- Probability theory
- Statistical learning → Machine Learning

§ Machine Learning

Types of Learning

- *Supervised*: human expert indicates desired output for given input to system
- *Reinforcement*: human indicates when system is doing good/ bad in an environment
- *Unsupervised*: no human feedback

Common ML Problems

- *Categorization*: assign discrete labels to different inputs (e.g., spam/ham)
- *Regression*: assign continuous "scores" to inputs (e.g., strength of a chess move)
- *Clustering*: grouping like inputs together (e.g., biological classification)

ML as Computational Statistics

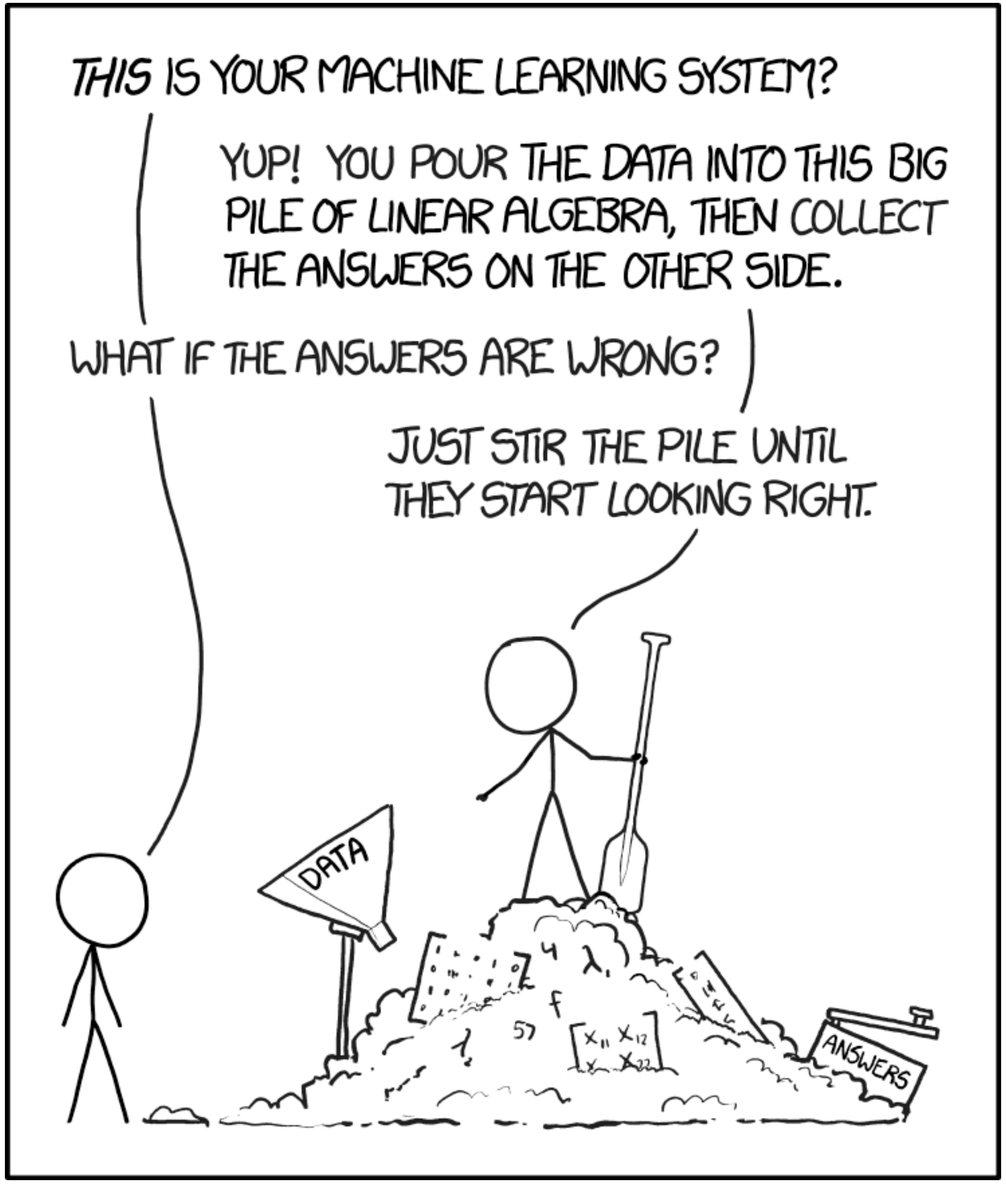
- Train models on existing data to identify and “learn” trends to predict aspects of future data

Ali Rahimi, a researcher in artificial intelligence (AI) at Google … charged that machine learning algorithms, in which computers learn through trial and error, have become a form of "alchemy." Researchers, he said, do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another.

"There's an anguish in the field," Rahimi says. "Many of us feel like we're operating on an alien technology."

The issue is distinct from AI's reproducibility problem, in which researchers can't replicate each other's results because of inconsistent experimental and publication practices. It also differs from the "black box" or "interpretability" problem in machine learning: the difficulty of explaining how a particular AI has come to its conclusions. As Rahimi puts it, "I'm trying to draw a distinction between a machine learning system that's a black box and an entire field that's become a black box."

- Matthew Hutson. *Science*, May 3, 2018.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Models

- Decision trees
- Bayesian networks
- Support Vector Machines
- Neural networks

Enter the Information Age!

- Massive quantities of data (PB+ scale)
- Lots of domain specific expertise necessary to acquire, manage, parse this data
- Applying machine learning and other techniques from CS + domain expertise to extract knowledge from data

§ Data Science