Data Science: What It Is and How It Can Help Your Company

with Eliezer Kanal and Brian Lindauer

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM-0003841

Which of the below is closest to your job title?

- ❑ Engineer, Architect, Analyst
- ❑ Manager, Director, VP
- ❑ CEO, COO, CTO
- ❑ Project Manager
- ❑ Consultant
- ❑ Student
- ❑ Other

What industry do you work in?

- ❑ Aerospace
- ❑ Academia
- ❑ Banking
- ❑ Computer Security
- ❑ Consulting
- ❑ Insurance
- ❑ Retail
- ❑ Software Engineering
- ❑ Other *(please specify in comments)*

Eliezer Kanal

- PhD in biomedical engineering from University of Pittsburgh
- Technical Manager at CERT (Software Engineering Institute, CMU)
- Previous experience:
  - Quantitative analyst at PNC Bank
  - Manager of Metrics and Analytics at Highmark, Inc.
- Expertise in signal processing, operations research, machine learning

Brian Lindauer

- Graduate certificate in Mining Massive Data Sets from Stanford University, B.S. in Computer Science from Columbia University
- Research Scientist in CERT (Software Engineering Institute, CMU)
- Previous experience:
  - Principal Software Engineer at Raytheon
  - Software Engineer at CounterStorm, built network anomaly detection systems
- Expertise in machine learning, software engineering, and cybersecurity

“Data Science” – note differences!

Eliezer

- Biomedical Engineering
- Predicting when businesses default on loans
- Optimizing call center staffing

Brian

- Computer Science
- Network security defense
- Predicting cybersecurity attacks

Similarities:

- Shared skill set
- Similar tools
- (Somewhat) similar training
- *Problem solving!*

Data Science skills

- **Statistical/Machine Learning techniques**
- **Data manipulation**
- **Information visualization**
- **Intermediate programming**
- **Quick learners**
- **Able to think of** <adjective> **questions**
  - …answerable...
  - ...intelligent...
  - ...relevant...
  - ...value-added...

Any particular topic you want to hear about?

- ❑ Insurance Operations Optimization
- ❑ Retail Traffic Maximization
- ❑ Credit Card Fraud
- ❑ Call Center Management
- ❑ Insider Threat Management
- ❑ Network Security
- ❑ Airline Safety

Case Study: Call Center Optimization

Project goal:

Determine optimal staffing pattern so as to meet a variety of contractual performance guarantees while minimizing staffing costs.

Data Science approach:

Identify historical call time distributions, create software to automatically recommend staffing needed to meet guarantees.

Skills used:

Quick learning, distribution fitting, algorithm development, data manipulation, programming

Case Study: Network Anomaly Detection

Project goal:

Detect abnormal, possibly malicious activity on a computer network and provide information for an analyst to determine a proper course of action.

Data Science approach:

Model normal network traffic and alert on unusual behavior matching known malicious patterns

Skills used:

Machine learning, systems programming, data manipulation and analysis, high-speed data processing, cybersecurity domain knowledge

What's in a name?

Business Analyst

Decision Support Analyst

Technical Analyst

Business value:

- Gather, condense, and present information
- Minimal-to-moderate technical expertise

Data Scientist

Business value:

- Extract actionable insights from existing data
- Think of as-of-yet unasked questions and answer them
- Machine learning and statistical expertise
- Assist with solution implementation

Questions on what Data Science means?

Questions on distinctions between Data Scientists and other positions?

Case Study: Insurance Claims Optimization

Project goal:

Determine trends in failure for claims to auto-adjudicate, identify actionable fixes

Data Science approach:

Use big data tools capable of handling >1B claims. Cluster similar claims to identify common attributes.

Skills used:

Big data retrieval and manipulation, clustering, data visualization

Case Study: CMU/Boeing Partnership

Project goal:

Determine improvements to be made to aircraft maintenance and safety through automated techniques.

Data Science approach:

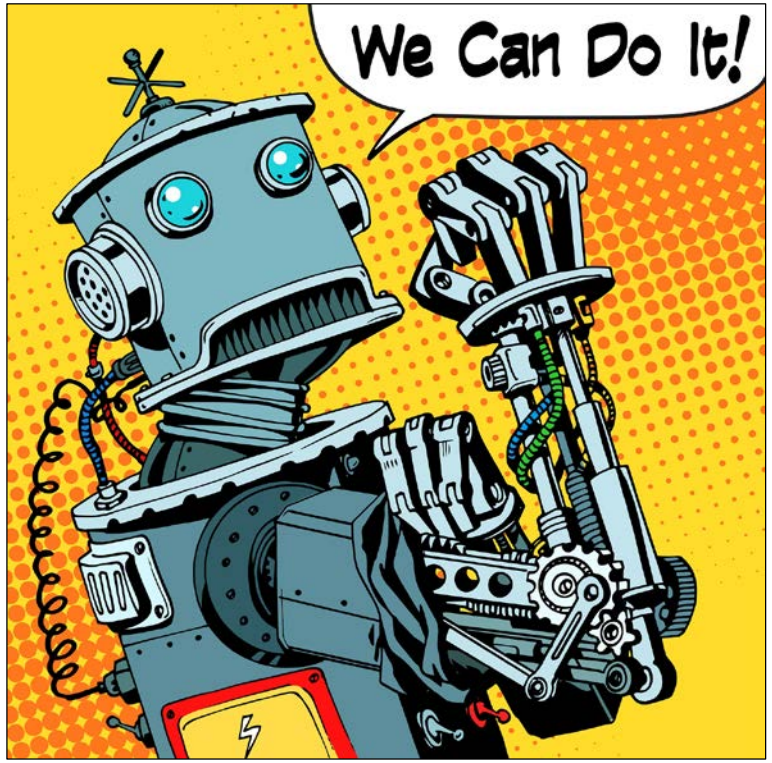
Collect and study data from countless sensors deployed throughout the aircraft to determine numerous aspects related to safety.

Skills used:

Big data collection & manipulation, quick learning, asking new questions

https://www.cs.cmu.edu/news/boeing-establishes-analytics-lab-aerospace-data-carnegie-mellon

What is Machine Learning?



Tom Mitchell, former CMU Machine Learning department chair:

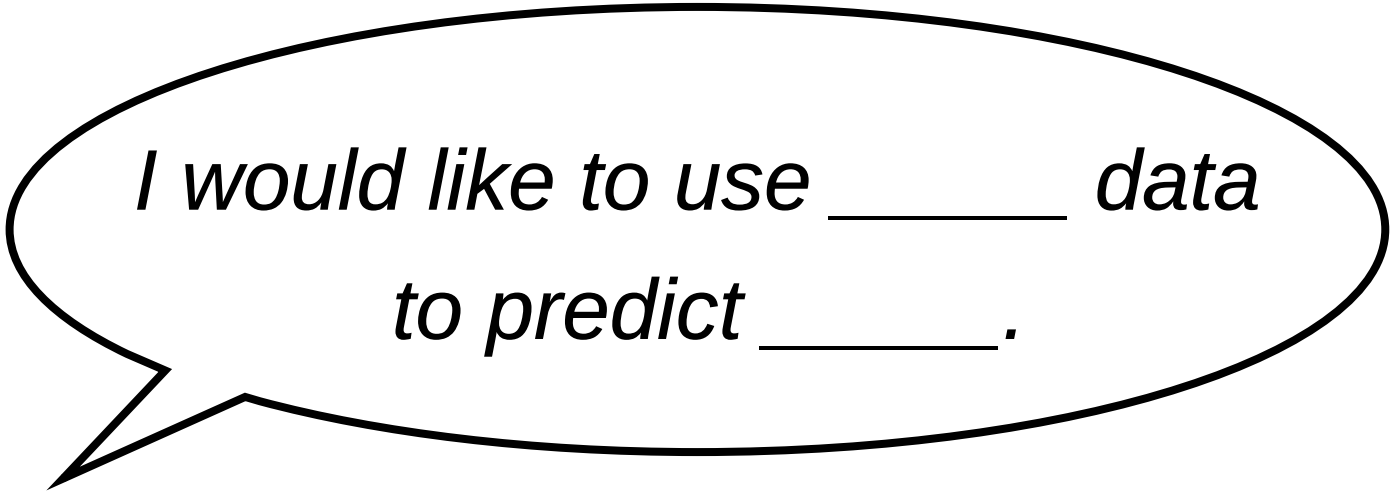
> *The field of Machine Learning asks the question, "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

Machine Learning seeks to automate data analysis and inference.

What is Machine Learning?

Picture (optional)

If your problem can be stated as:



…you would likely benefit from machine learning.

Case Study – Credit Card Fraud

Project goal:

Automatically detect fraudulent credit card transactions as they occur.

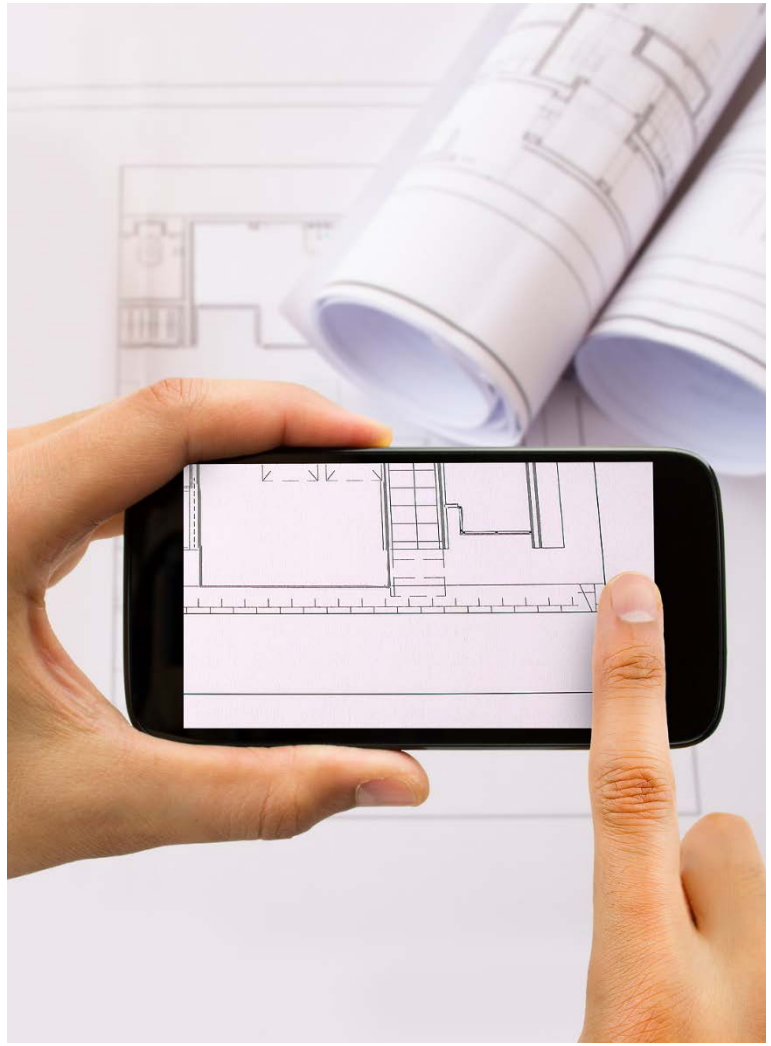
Data Science approach:

Create machine learning model based on historical transaction data to determine what combinations of attributes signify a fraudulent transaction.

Skills used:

Big data manipulation, machine learning and statistical techniques, value-added questions

Case Study – Insider Threat Detection



Project goal:

Identify employees either acting maliciously or at risk of becoming malicious actors.

Data Science approach:

Automatically mine email, workstation activity, badging records, etc. for unusual activity matching patterns of suspicious behavior, indicate where intervention is required.

Skills used:

Big data, natural language processing, machine learning/statistics, domain expertise

Case Study – Retail

Project goal:

Increase store activity by providing customers with coupons they are most likely to find useful

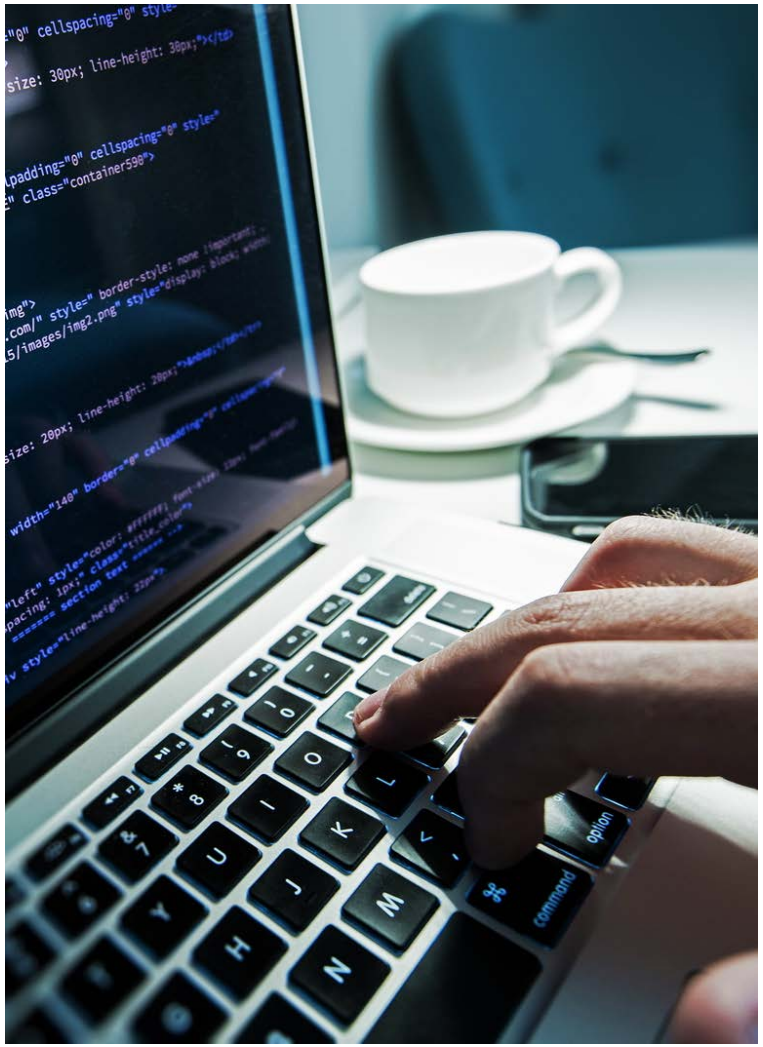
Data Science approach:

Build a recommender system using historical data, recommend future purchases. Identify which products are likely to be bought together and provide coupons for groups of products.

Skills used:

Big data manipulation, machine learning and statistical techniques

Case Study – Software Development



Project goal:

Decrease the number of bugs in released software.

Data Science approach:

1) Create a deep learning model to detect bugs
2) Build a grammar model and examine for irregular usage or known bad practices

Skills used:

Natural language processing, data extraction and manipulation, statistics and machine learning

Takeaway Points

- Data Scientists have very diverse backgrounds
- Critical Data Scientist skills:
  - Statistics and Machine Learning skills
  - Data manipulation
  - Knows how to add value to your organization
  - Quick learner
- Machine Learning: "*I want to use* _____ *data to predict* _____*.*"
- Data Science can add value to your organization

Contact Us

SEI can assist you with your data and analytics needs… reach out to us!

Eliezer Kanal – ekanal@cert.org

Brian Lindauer – lindauer@cert.org

Software Engineering Institute
Carnegie Mellon University
4500 Fifth Avenue
Pittsburgh, PA 15213