outrageously
AMBITIOUS

Fundamentals of Data Science

EGMGMT 585-01, 02

Fall 2024

Monday 3:05-5:50 PM, Teer 203

Duke
PRATT SCHOOL of ENGINEERING

Daniel Egger

BA in Philosophy, JD in Law, **Yale University**

Graduate study in philosophy, **Ludwig Maximilian's University**, Munich, Germany

30 Years **Software Industry** experience

Invented and commercialized patented search technology licensed to **Google, Yahoo, Microsoft**, etc.

Founded Three Venture-backed Companies and assisted many others as Board Member and investor - most recently **Little Otter** ($20 million "A" financing 2021).

Ran an early-stage **Venture Capital Fund** for 7 Years

Duke University **Entrepreneur in Residence** (2003-2009)

Since 2009, teach Data Science, Mathematical Finance, and AI and Data Visualization at **Pratt School of Engineering**, Duke University (MEM, MIDS, AIPI Programs)

Over **1.2 Million On-line Students** in 80+ Countries on Coursera

Duke
PRATT SCHOOL of ENGINEERING

No Coding Required in this Course
Problem Sets can be done 100% Using MS Excel

Why Excel?

Excel remains ***the most widely used data science tool in industry*** especially by senior executives. Fluency in Excel is an essential business skill for managers.

It is easy to see ***how algorithms work*** in Excel – and the focus of this course is on understanding the inner workings of the most important data science methods ***and algorithms***.

What is Data?

Anything that can be measured:
money, calories, heart rates, click-throughs, you name it!

Anything that can be stored in digital form, including:
Written and spoken language
Images and video
Software (Code)

So is Data Science the Science of Everything? Yes, in a way.

Data science is about **extracting value from data mathematically**

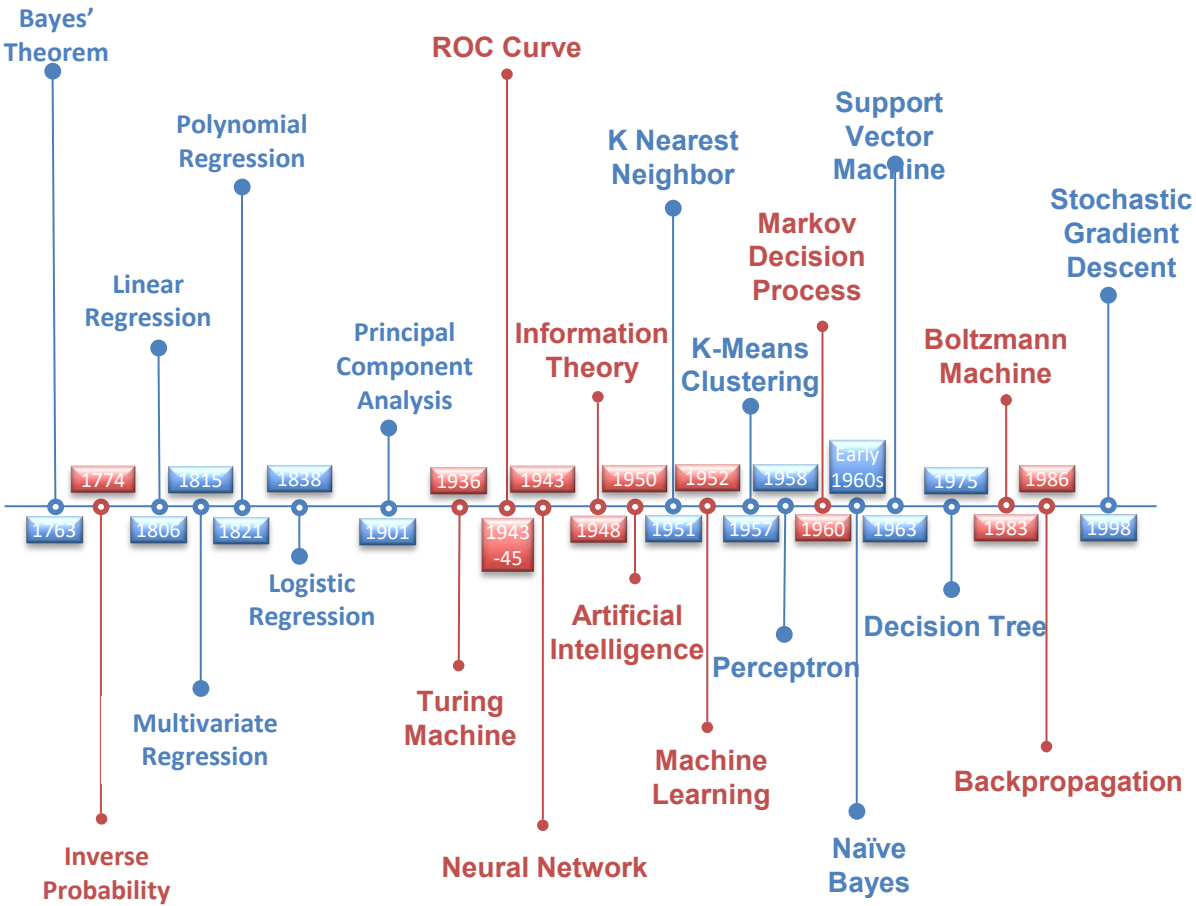
Creating **inferences**,
predictive models, and **insights**
to guide important decisions and **influence behavior**

We cover the step-by-step engineering insights that led from simple statistics to Machine Learning and Generative AI/ LLMS

Duke
PRATT SCHOOL of ENGINEERING

A long history of ideas

MACHINE LEARNING Concepts and Algorithms



Thanks to Shiyan Chen for creating this visualization

Course Goals

Goal is to **prepare you to collaborate effectively** with the data science and AI teams in high-tech companies

Become fluent with the most important data science **vocabulary** and **concepts**

Understand how and why the key **algorithms** and **mathematical methods** work and get **hands-on experience** with real-world data-driven problem-solving

Duke
PRATT SCHOOL of ENGINEERING

Why is Data Science Central to Business?

All Businesses track performance through **Business Metrics**

Data can be used to identify, through **experience** (historical data) or **experiments** (new, deliberately-generated data) **Business process changes** that can positively impact one or more **business metrics**

Data are identified, scrubbed, and analyzed, then **process change recommendations** are **identified** and **communicated to decisionmakers**

This process is **ongoing** an **iterative** because competition is relentless

New Categories of Products and Services

Many of the **most transformative** new products and services of the last 20 years could not have existed without **new forms of digital data**

They often benefit from **ongoing optimization** through tracking customer behavior

Creating a “virtuous cycle” of competitive advantage that can lead to complete market dominance for a single corporation

Course Topics

The **CRISP-DM** Data Science Project Methodology

Definitions of **Machine Learning**.

Supervised Learning Methods & Algorithms. **Binary Classification** and **Linear Regression, Logistic Regression**, **Multi-Class Classification**, **Avoiding Overfitting**

Unsupervised Learning Methods and Algorithms: Clustering (**K Means Algorithm**), use of **Scree Plots** for Feature Selection

Topics

Probability Distributions in Data Science: Uniform, Gaussian, Bernoulli, Binomial, Exponential and Poisson

Parameter Estimation, Histograms, **Kernel Density Estimation**. Use of **Bayes' Theorem** and **Maximum Likelihood Estimation**

Introduction to **Information Theory**: Why Machine Learning Algorithms Optimize on the **Log Loss** Cost Function

Experimental Design, **A/B testing**, **p-Values** and **Power**

Key Concepts in **Time Series Analysis**

Topics, Continued

Artificial Intelligence

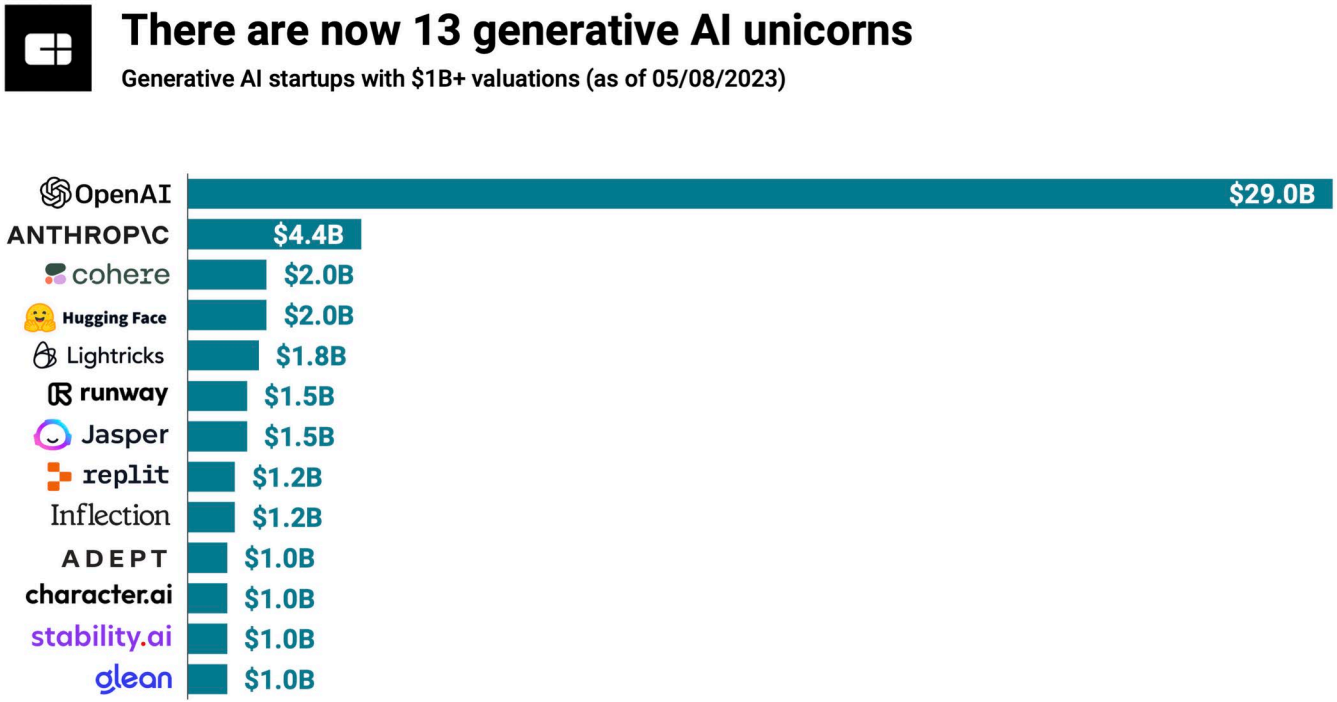
Generative Adversarial Networks

Large Language models

Exciting areas of current research and venture-backed development

GANs trained on image Data

Valuations as of May, 2023



CBINSIGHTS

Duke
PRATT SCHOOL of ENGINEERING

See You August 26!

Duke
PRATT SCHOOL of ENGINEERING