Course Overview

Seyed Abbas Hosseini
Sharif University of Technology

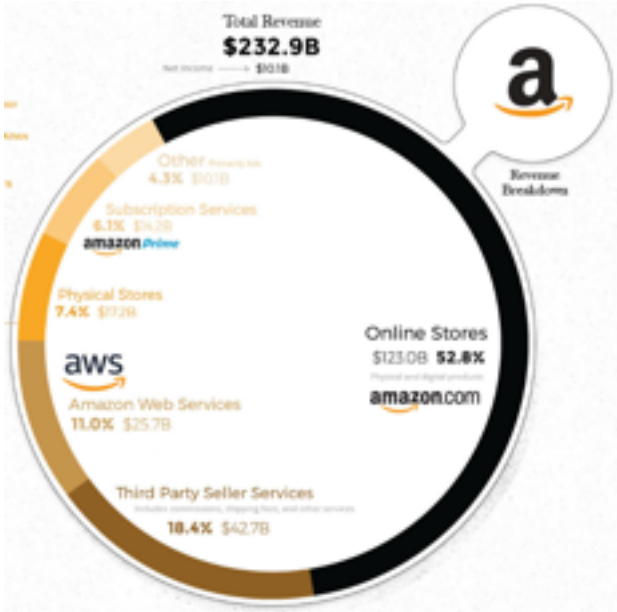
Outline

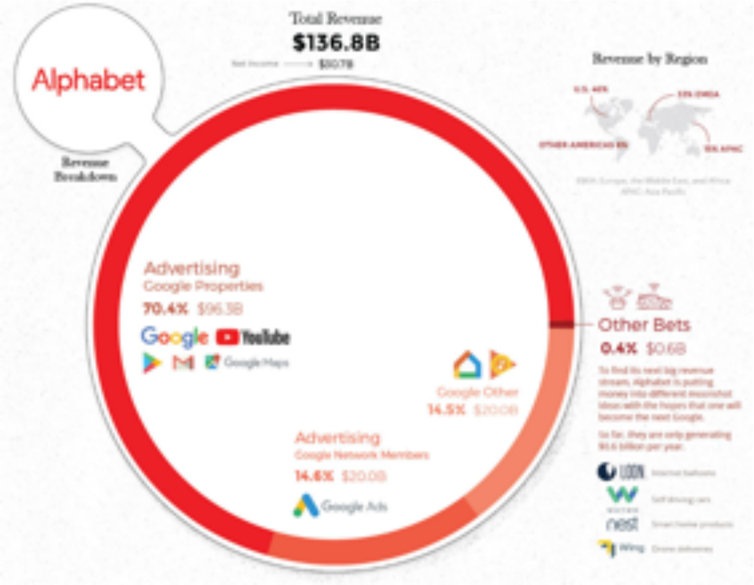
- ❑ Why am I excited about Machine Learning?
- ❑ What is Machine Learning?
- ❑ What is Data Science?
- ❑ What you will learn in this class?
- ❑ Course Logistics

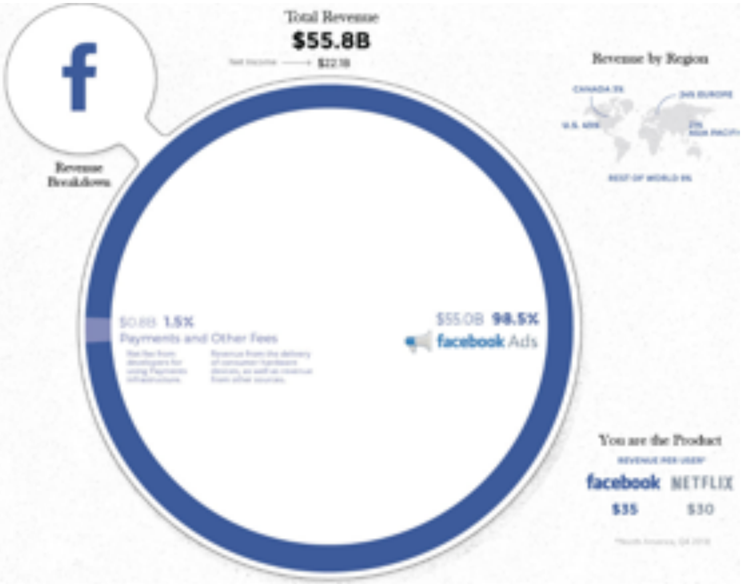
Why am I excited about

Machine Learning?

Machine Learning is the shovel to mine gold

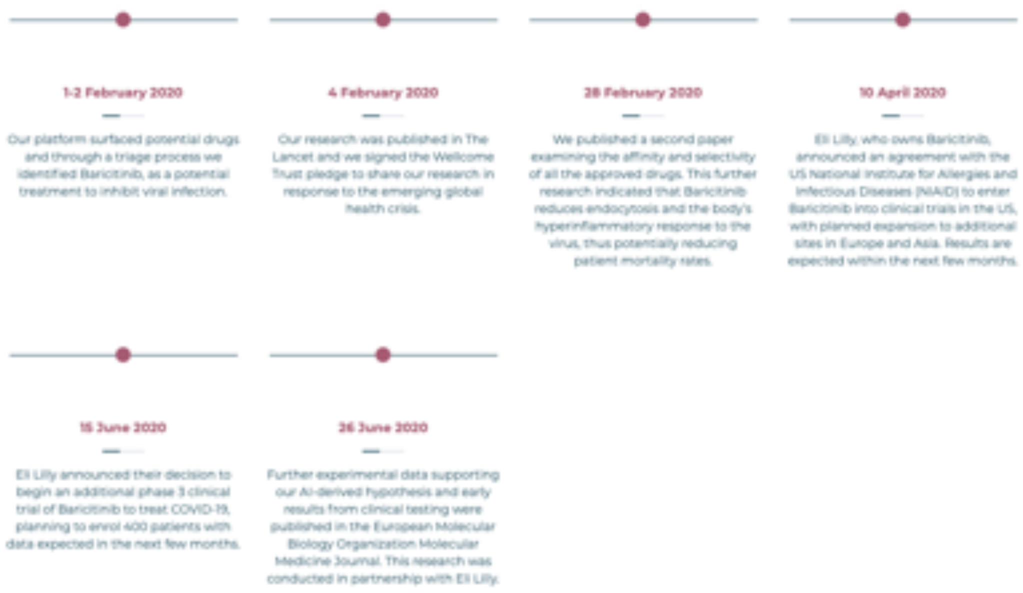






ML changes the method to tackle challenges

BenevolentAI identified a potential *coronavirus* treatment using their Knowledge Graph 4 months earlier than the owner company of the drug



1-2 February 2020

Our platform surfaced potential drugs and through a triage process we identified Baricitinib, as a potential treatment to inhibit viral infection.

4 February 2020

Our research was published in The Lancet and we signed the Wellcome Trust pledge to share our research in response to the emerging global health crisis.

28 February 2020

We published a second paper examining the affinity and selectivity of all the approved drugs. This further research indicated that Baricitinib reduces endocytosis and the body's hyperinflammatory response to the virus, thus potentially reducing patient mortality rates.

10 April 2020

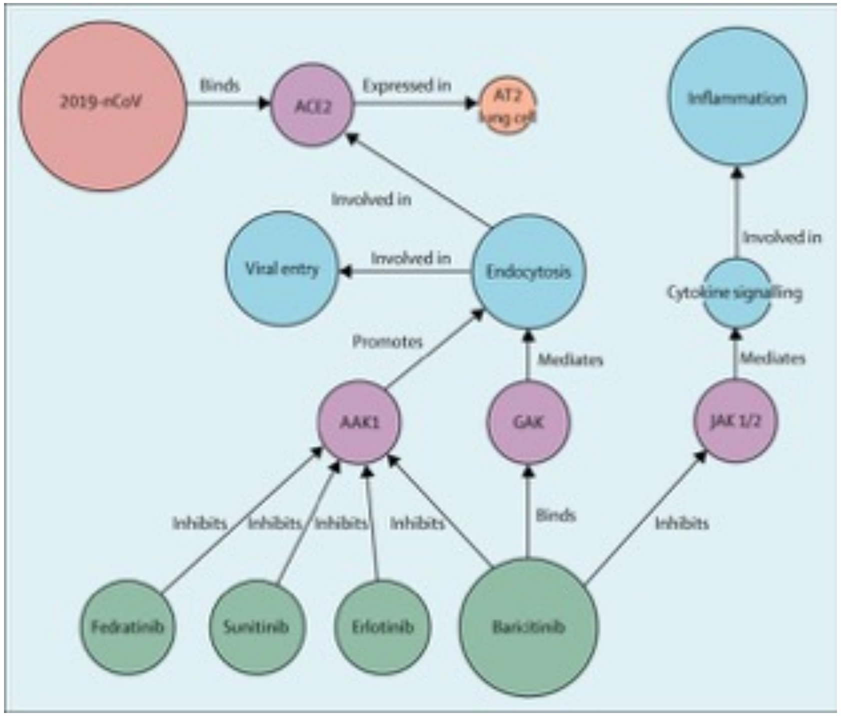
Eli Lilly, who owns Baricitinib, announced an agreement with the US National Institute for Allergies and Infectious Diseases (NIAID) to enter Baricitinib into clinical trials in the US, with planned expansion to additional sites in Europe and Asia. Results are expected within the next few months.

15 June 2020

Eli Lilly announced their decision to begin an additional phase 3 clinical trial of Baricitinib to treat COVID-19, planning to enrol 400 patients with data expected in the next few months.

26 June 2020

Further experimental data supporting our AI-derived hypothesis and early results from clinical testing were published in the European Molecular Biology Organization Molecular Medicine Journal. This research was conducted in partnership with Eli Lilly.



What is

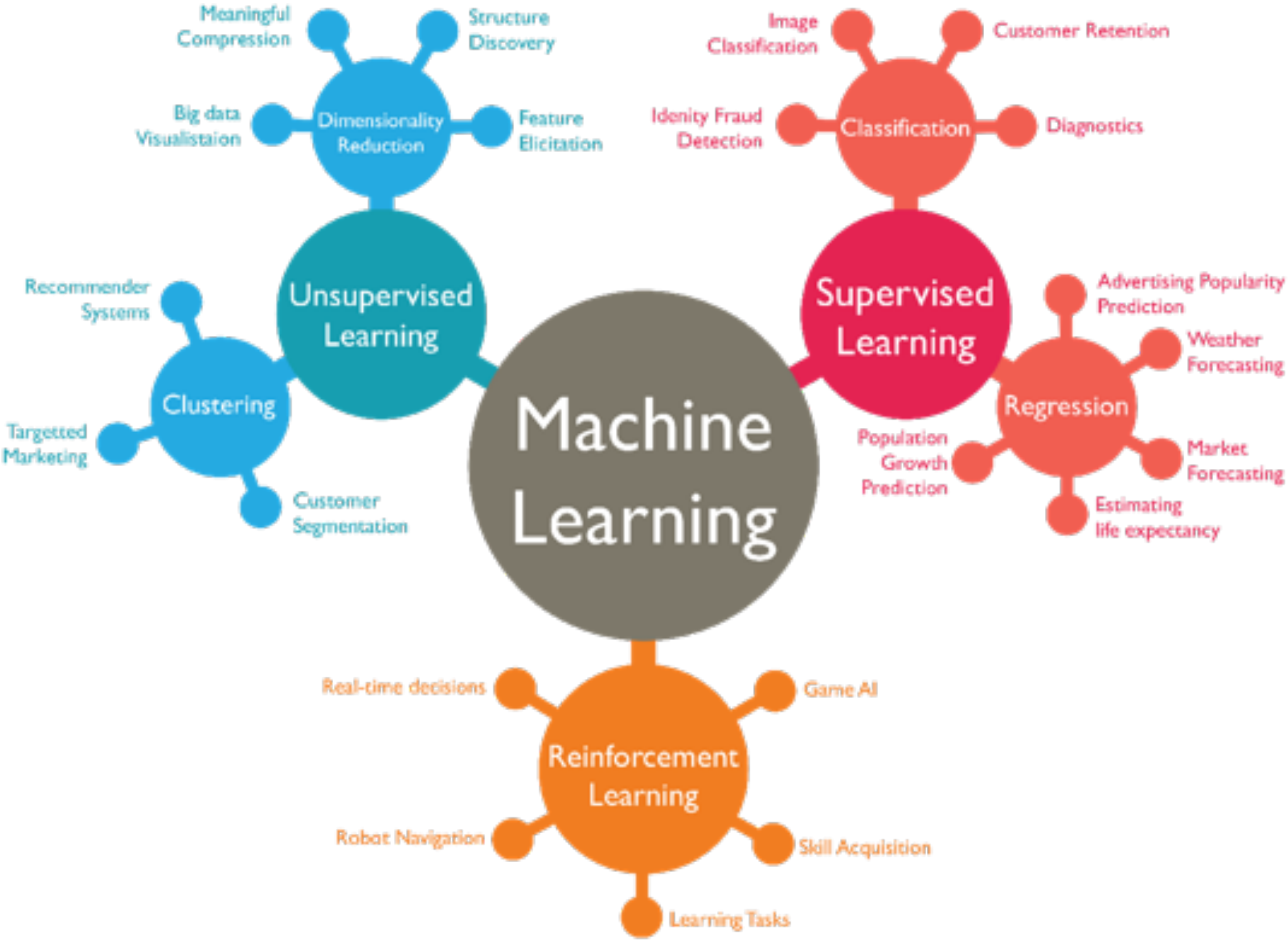
Machine Learning?

What is Machine Learning

Developing systems that are able to automatically ***Learn*** and ***Improve*** from ***Experience***

- **Modeling**
  - Proposing a (probabilistic) model for data
- **Learning Model Parameters**
  - Using Estimation theory to find an objective function
  - Use (large scale) optimization to find optimal parameters
  - Evaluation and error analysis
- **Generalization & Prediction**
  - Using Learned model to make informed guesses or predict the future
- **Decision Under Uncertainty**

Machine Learning Paradigms



What is

Data Science?

The recurring question across industry and academia.

Data Science Definition

The application of **data centric**, **computational**, and **inferential thinking** to

understand the world & *solve problems*

Science **Engineering**

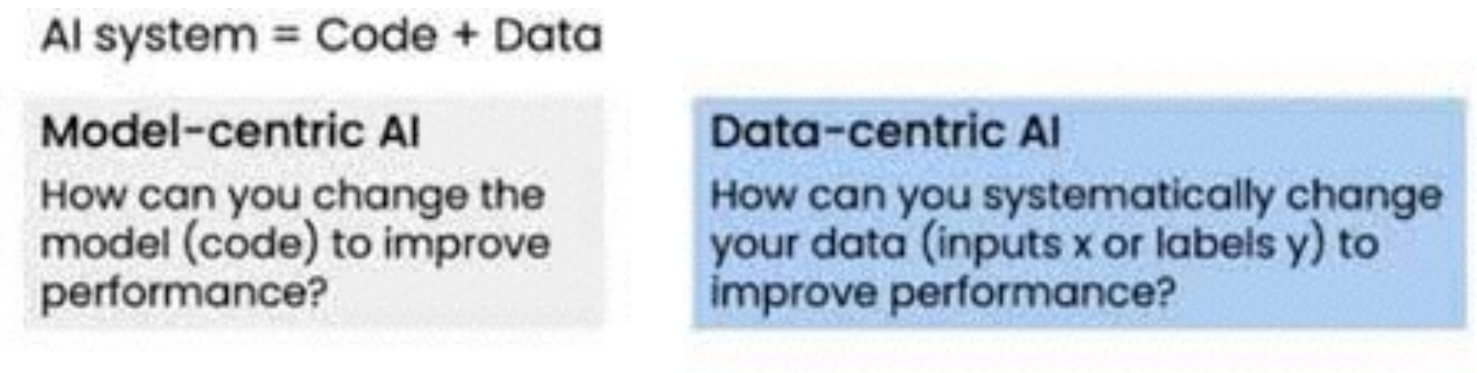
From Joey Gonzalez.

What We Do in Data Science?

Drawing Useful **Conclusions** from **Data** using Computation

- **Exploration**
  - Collecting, integrating and cleaning data
  - identifying patterns in data using visualizations
- **Prediction**
  - Model data and train a model using ***Machine Learning***
  - Making informed guesses using learned model
- **Analyze and Make Decision**
  - Analyze the results
  - Making decision under uncertainty

Data Centric AI



AI system = Code + Data

Model-centric AI

How can you change the model (code) to improve performance?

Data-centric AI

How can you systematically change your data (inputs x or labels y) to improve performance?

Model-centric

- Collect as much data as we can
- Optimize the model so it can deal with the noise in the data

Approach:

- Data is fixed after standard preprocessing
- Model is improved iteratively

Data-centric

- Data consistency is key
- Higher investment in data quality tools rather collecting more data
- Allows more models to do well

Approach:

- Hold the code/algorithms fixed
- Iterated the data quality

Data Centric AI

We have to answer the following questions to have a data-centric approach

- **Is the data complete?**
- **Is the data relevant for the use case**
- **If labels are available, are they consistent?**
- **Is the impact of bias impacting the performance?**
- **Do I have enough data?**

Data quality has to be *monitored* and *improved* at every step of the AI development in a *continuous manner* which makes **MLOps** a much-needed ally to achieve a proper and successful *data-centric* paradigm.

What are we looking for in data science?

Insight

Good data analysis is not:

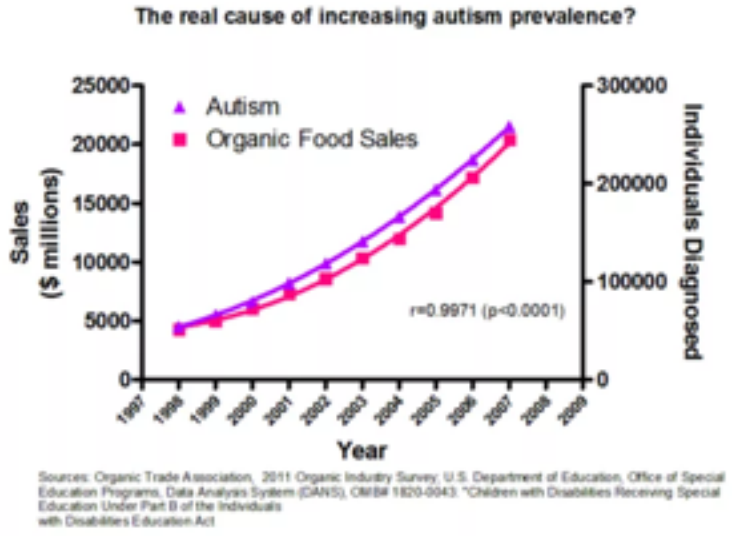
- Simple application of a statistics recipe.
- Simple application of statistical software.

There are many **tools** out there for data science, but they are merely tools.

- **They don’t do any of the important thinking!**

“The purpose of computing is insight, not numbers.” - R. Hamming. *Numerical Methods for Scientists and Engineers (1962).*

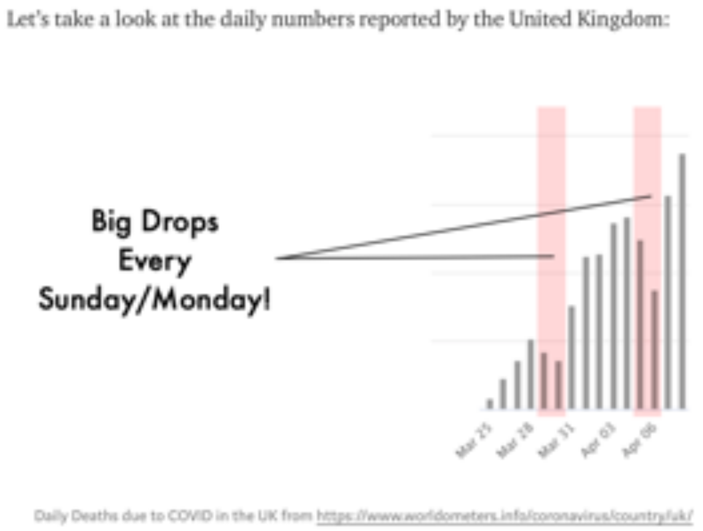
Question what you see!



The real cause of increasing autism prevalence?

Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

Are autism rates and organic food sales inherently related? Seems unlikely.



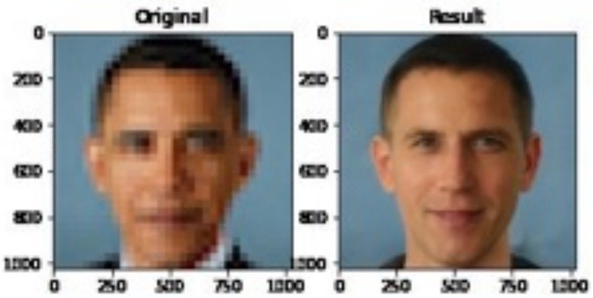
Let's take a look at the daily numbers reported by the United Kingdom:

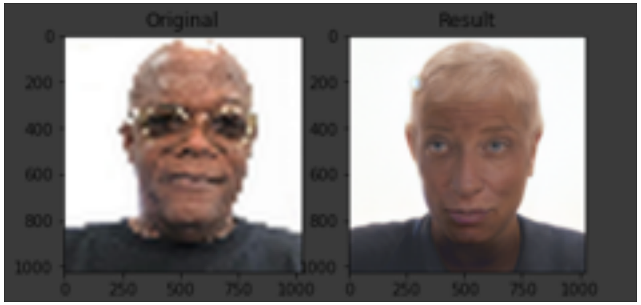
Daily Deaths due to COVID in the UK from https://www.worldometers.info/coronavirus/country/uk/

The problem is that this weekly cycle is fake. It's an artifact of how the data is collected and reported.

Unconscious bias is real – be mindful of it

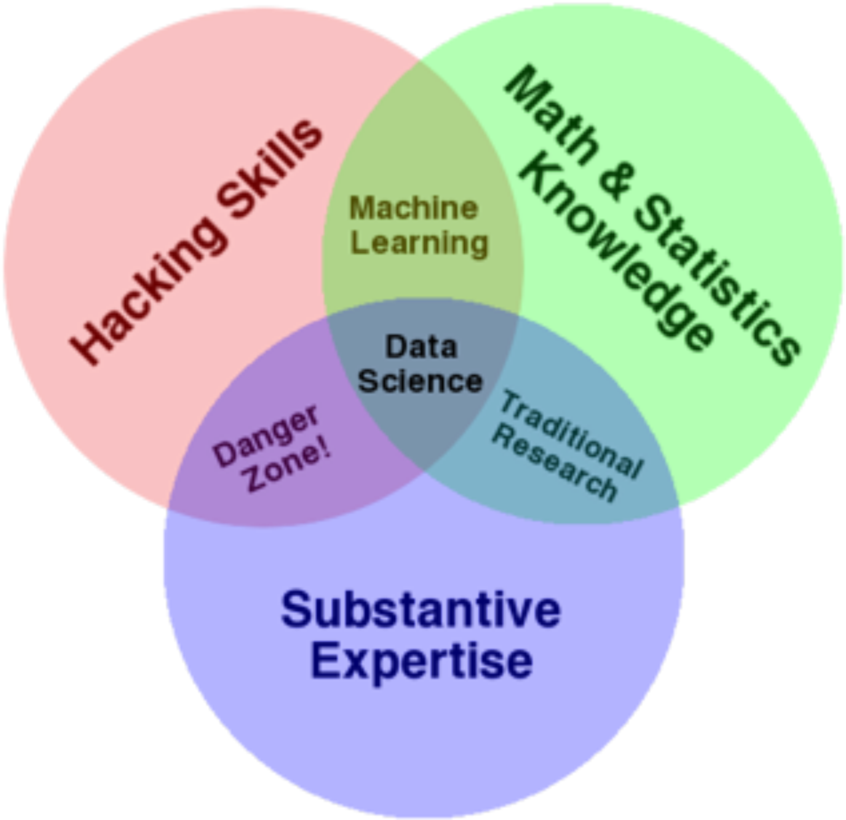
A "depixelizer" was built that takes pixelated images and generates images that are perceptually realistic and downscale correctly.





What do you notice? **Why** might this be happening?

Data Science Venn Diagram



by Drew Conway in 2010 (link)

What you will learn in this class?

Course Goals

Familiarize — Familiarize students with fundamental concepts and popular algorithms in Machine Learning

Empower — Empower Students to apply computational and inferential thinking to tackle real world problems

Enable — Enable Students to start career as data scientist by providing experience working with real world data, tools and technologies.

Topics covered in this course

- Pandas and NumPy
- Exploratory Data Analysis
- Visualization
- Dimensionality reduction for visualization
- Model design and loss formulation
  - Gradient Descent
  - Regularization, Bias-Variance Tradeoff, Cross-Validation
- Linear Regression
- Classification
  - Logistic Regression
  - Decision Trees
- Ensemble Learning
- Deep Neural Networks
  - Multilayer Neural Networks
  - Backpropagation
  - Training DNN challenges
  - Convolutional Neural Networks
- ML for Production (MLOps)
  - ML Lifecycle in Production
  - Data Lifecycle in Production
  - ML Modeling Pipelines
  - Deploying ML in Production

Course Logistics

Instructor

Seyed Abbas Hosseini

- I got a Ph.D. In Machine Learning from SUT.
- Currently I'm an assistant professor in SUT and working as a data scientist in industry.
- My contact info is available at https://mlclass.ir/staff/.
- Office Hours: Contact me to set an appointment.

Head TA

Seyed Mohammad javad

Feizabadi Sani

- Mohammad Javad is the Head TA of the course
- Contact info is available at https://mlclass.ir/staff/.
- With any logistic concerns email Mohammadjavad

References

- Chris Bishop, ***Pattern Recognition and Machine Learning***, 2nd edition, 2006
    - Main reference for the ML parts.
- A. Zhang, Z. Lipton, M. Li, A. Smola, **Dive into Deep Learning**
- S. Lau, J. Gonzalez, D. Nolan, ***Principles and Techniques of Data Science***.
    - In first portion of the class, we will cover some parts of this book

Remote Instruction

This is the third time ***entirely remote*** offering of Machine Learning and it is the third time offered ***specially for B.Sc. Students.***

- There will also be a lot of ***experimentation***! We want your ***feedback*** on what works and what doesn't.
  - We will have weekly surveys.
  - These are released on **Tuesday**, and are due that **Friday**.
    - These deadlines are flexible, but we really would like for you to fill them out!
  - Weekly surveys may also contain logistical questions.
- The following information is all on the **syllabus** on the website.
- The **calendar** page contains the scheduling for all live events.

Online Platforms

- **Course website** (**https://mlclass.ir**)
    - Where all lectures, assignments, and discussions are posted.
- **Piazza** (https://piazza.com/sharif/fall2021/ce7172/)
    - A place to ask and answer questions about assignments and concepts.
    - Where all announcements are posted (exam logistics, new assignment released, etc).
- **Quizify** (**mlclass.ir/quizify**)
    - A website developed by TAs to take quizzes online.
    - The username and password for each student will be posted via email

Homework, Quizzes and Projects

Be informed that this is a **graduate level course** although offered for B.Sc. students. We expect you devote at least **2 days per week** to this course.

- There will be 5 HW series (every other weeks)
  - Each containing some theoretical and programming problems.
  - Homework will be released on course website
  - Use Piazza to ask any question regarding HW problems
  - The late submission policy is announced on course website
- There will be two random quizzes (totally 5%)
- There will be two mini-exams to wrap up course materials
- There will be a project instead of two last HWs to make you appropriate work with ML tools in real world scenarios
  - The details will be announced on Piazza

Grading

- 25% Homeworks
  - each 5%
- 30% Mini exams
- 5% Quizzes
- 15% Project
- 25% Final Exam

Any Questions?!