

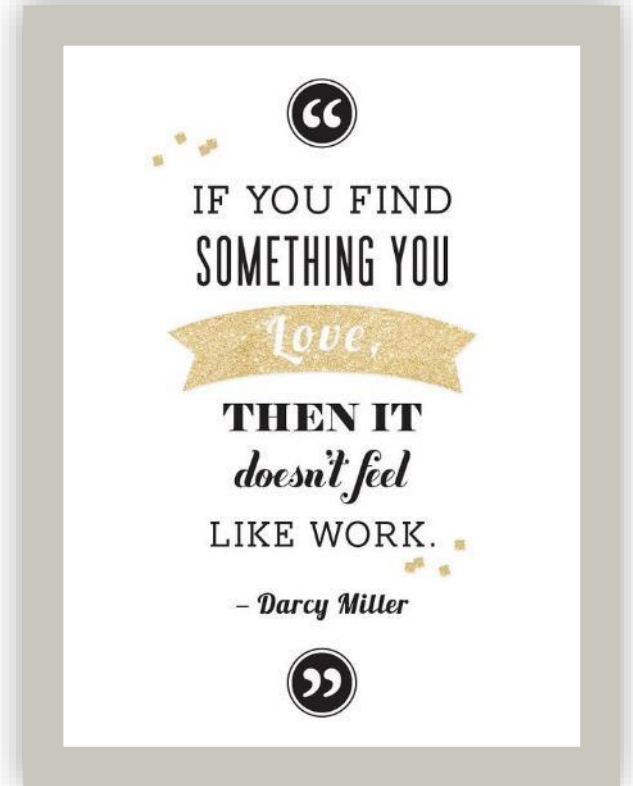
01 – Introduction

CSAI 325: Introduction to Data Science
Week 1



Course Principles

- Fast paced
- Understanding > Memorizing
- Practicing > Understanding
- Enjoy!
- You have what it takes:
 - experience with Python,
 - background in probability and statistics,
 - and linear algebra



Team


- Dr Mohamed Gamal
- Eng. Asmaa El-Hadidy
- Eng. Yomna Ramadan
- Eng. Nourhan Khalaf

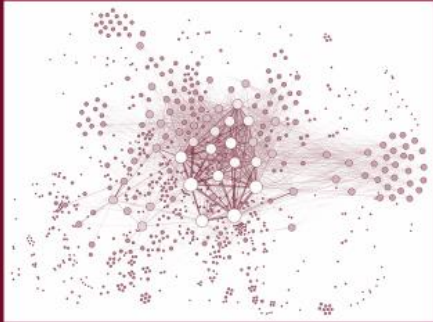
Lectures

- Concepts + Hands-On
- The second lecture will be in the Lab
- We try to practice what we discussed

Week 1
Introduction to Data
Science

CSAI 325: Introduction to Data Science





```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

"""
This script performs a Hyperparameter Search for a Neural Network
using Grid Search and Cross-Validation. It uses the sklearn library
for model selection and evaluation.
"""

# Imports
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.neural_network import MLPClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import load_digits

# Load data
X, y = load_digits(return_X_y=True)

# Standardize the data
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Define the model
model = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(100, 100), max_iter=1000)

# Define the parameter grid
param_grid = {
    'hidden_layer_sizes': [(100, 100), (100, 50), (50, 100)],
    'activation': ['tanh', 'relu'],
    'solver': ['lbfgs', 'adam'],
    'alpha': [1e-5, 1e-4, 1e-3],
    'max_iter': [100, 200, 500, 1000]
}

# Perform Grid Search
grid_search = GridSearchCV(model, param_grid, cv=5, scoring='accuracy')
grid_search.fit(X, y)

# Print the best parameters and score
print("Best parameters found: %s" % grid_search.best_params_)
print("Best cross-validated accuracy: %s" % grid_search.best_score_)

# Train the final model
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X)
accuracy = accuracy_score(y, y_pred)
print("Final accuracy: %s" % accuracy)
```

Lab Sessions

- Once a week, supervised by TA
- Cover:
 - Quick review of the concepts covered in the lectures.
 - Demo code
 - Problem solving
- Code will be made available after each session



Exams and Assignments

- 1 Mid-term exam
- Final exam
- Best 2 of 3 Quizzes
- Weekly Assignments
- Two Projects
 - Midterm Mini Project: Individual
 - Final Project: Teams of 3
- Hackathons
 - Two hackathons one before the midterm and one before the final

Assignments, Integrity

- **Assignments:**
 - Integral part of learning – NOT optional!
 - Gauge how well you are keeping up
 - Provide individual assistance
 - Make adjustments to the course if necessary
- **Academic Integrity:**
 - Your future, your investment!
 - You are here to learn, not just to earn grades
 - I expect mature and responsible behaviour
 - **Don't make a habit of cheating your way out of challenges... it will hurt you your whole life!**
 - If you find an assignment difficult, contact me or the TA and we will help you
 - Investigations, penalties, etc.



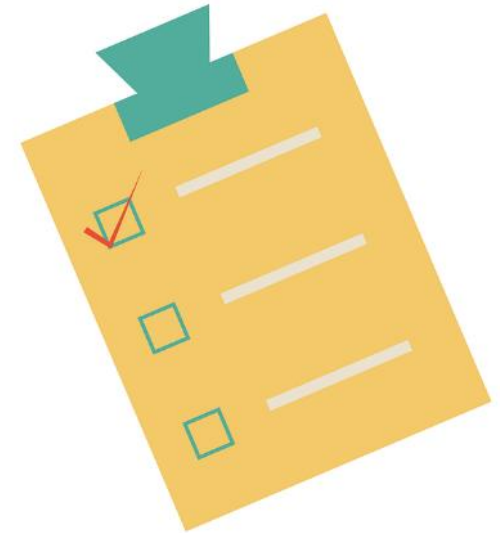
Grading Summary

NEW GIZA UNIVERSITY

Item	Weight
Lab Exercises	10%
Quizzes	10%
Assignments	10%
Mid-term	15%
Final exam	30%
Projects	15%
Hackathons	10%

Learning Objectives

- Understand the phases of a DS project
- Choose the right DS approach for a task
- Communicate using data visualisations
- Import and explore datasets using statistics and visualisations
- Format, clean, and prepare data for use in predictive tools
- Engineer features that will improve model performance
- Understand, select, apply, and tune Black-Box Machine Learning models
- Write code in Python and the standard DS libraries, while following best practices for code organization and reusability
- Understand the specificities of DS in corporate environments.



Prepare

Prepare students for advanced courses in **data management, machine learning, and statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**.

Empower

Empower students to apply computational and inferential thinking to address **real-world problems**.

Course Contents

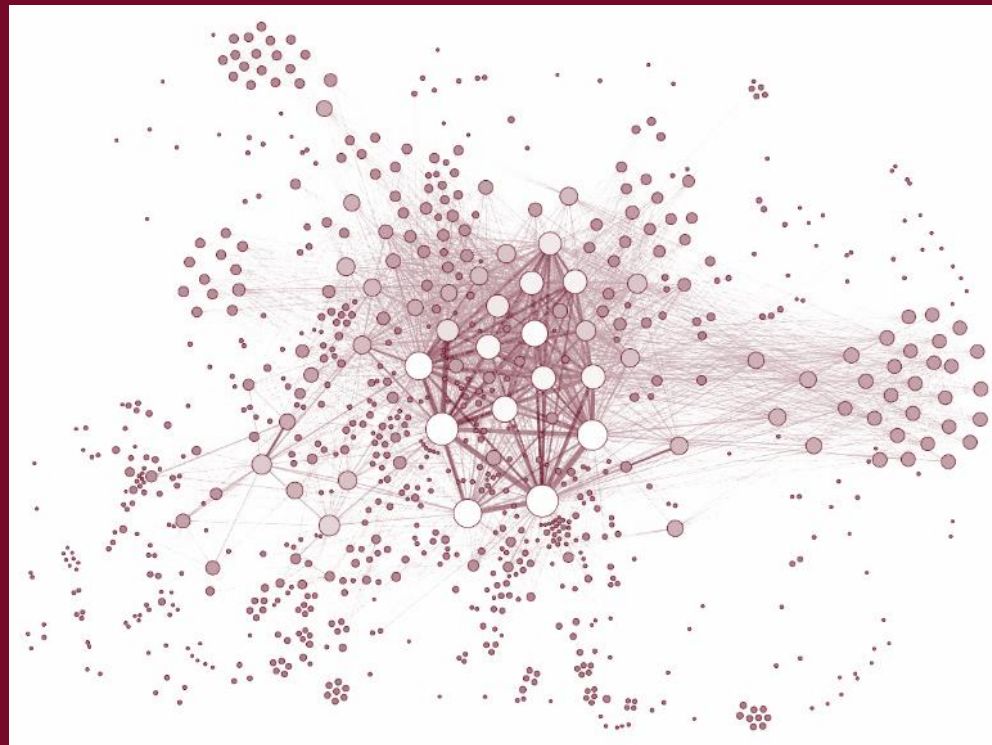
Week	Topic	Main References
1	Introduction to Data Science	M. Herman et al., <i>The Field Guide to Data Science</i>
2	Communicating with Data (Visualisation)	H. Wickham, <i>A Layered Grammar of Graphics</i>
3	Tools of the Trade: Numpy	W. McKinney, <i>Python for Data Analysis</i> , notebook
4	Tools of the Trade: Pandas	W. McKinney, <i>Python for Data Analysis</i> , notebook
5	Tools of the Trade: Matplotlib / Seaborn	W. McKinney, <i>Python for Data Analysis</i> , notebook
6	Acquiring and Inspecting Data	Notebook
7	Exploratory Data Analysis	Notebook
8	Data Cleaning	Notebook
9	Feature Selection and Engineering	A. Müller & S. Guido, <i>Introduction to Machine Learning with Python</i> , notebook
10	Introduction to Machine Learning (1)	A. Müller & S. Guido, <i>Introduction to Machine Learning with Python</i> ; M.P. Deisenroth et al., <i>Mathematics for Machine Learning</i> ; T. Hastie et al., <i>The Elements of Statistical Learning</i>
11	Introduction to Machine Learning (2)	A. Müller & S. Guido, <i>Introduction to Machine Learning with Python</i> ; AM.P. Deisenroth et al., <i>Mathematics for Machine Learning</i> ; T. Hastie et al., <i>The Elements of Statistical Learning</i>
12	Linear Models, k-Nearest Neighbours	A. Müller & S. Guido, <i>Introduction to Machine Learning with Python</i> ; AM.P. Deisenroth et al., <i>Mathematics for Machine Learning</i> ; T. Hastie et al., <i>The Elements of Statistical Learning</i>
13	Naïve Bayes, Tree-based Models, Unsupervised Learning	A. Müller & S. Guido, <i>Introduction to Machine Learning with Python</i> ; AM.P. Deisenroth et al., <i>Mathematics for Machine Learning</i> ; T. Hastie et al., <i>The Elements of Statistical Learning</i>
14	Practical Data Science in a Corporate Environment	
15	Wrap-Up & Project Completion	

- Quizzes, course content, notebooks, assignments, submissions, etc.
 - Moodle
- Set up a Conda environment with required libraries:
 - Lab 1
- You can run your Jupyter Notebooks on Google Collab for free:
 - Lab 1

Week 1

Introduction to Data Science

CSAI 325: Introduction to Data Science



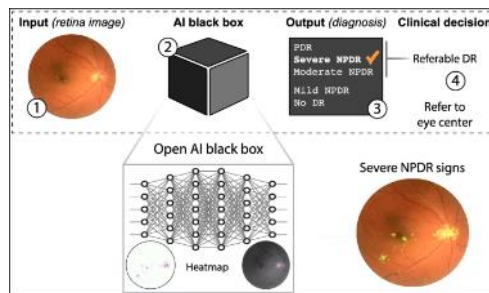
Why Learn Data Science?

How Is Data Science Used?

- Recommendation systems



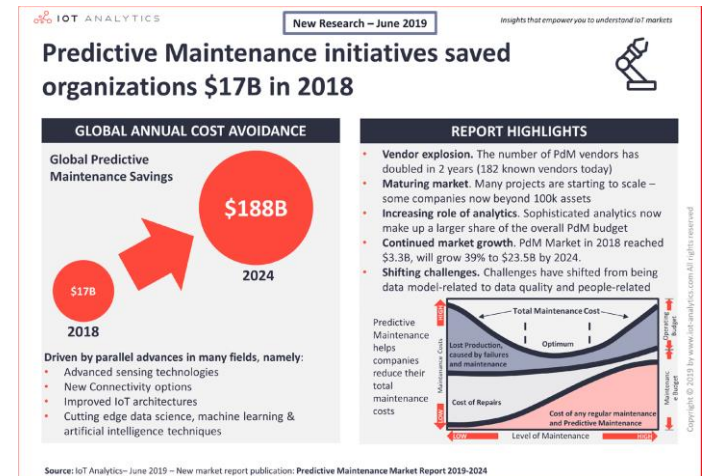
- Health diagnosis assistance



Lim et al. 2020, *Different Fundus Imaging Modalities And Technical Factors In AI Screening For Diabetic Retinopathy: A Review*

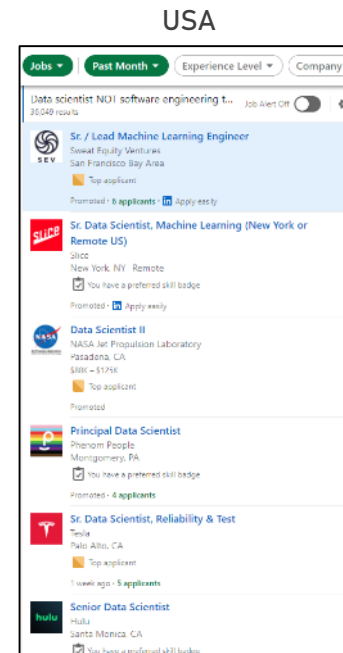
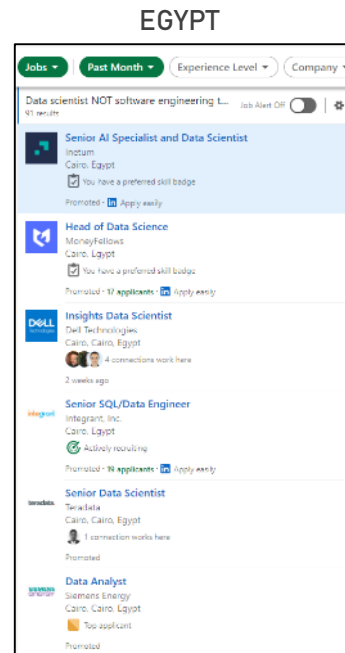
How Is Data Science Used?

- Preventive maintenance of industrial equipment
- Commodity price prediction
- Targeted advertising
- Supply chain optimisation
- Performance improvement in Sports, etc.



Where Is Data Science in Demand?

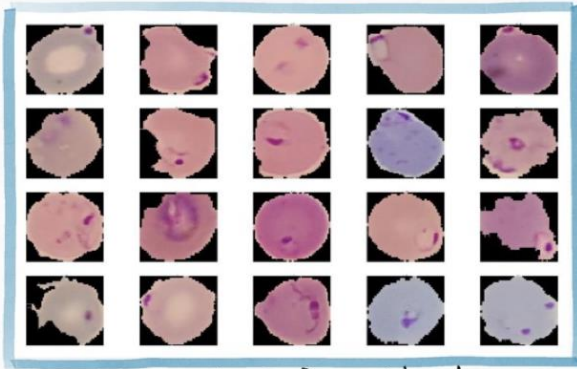
- In almost every industry, everywhere in the world
- Job ads on LinkedIn:



Where Is Data Science in Demand?

NEW GIZA UNIVERSITY

Disease Diagnosis



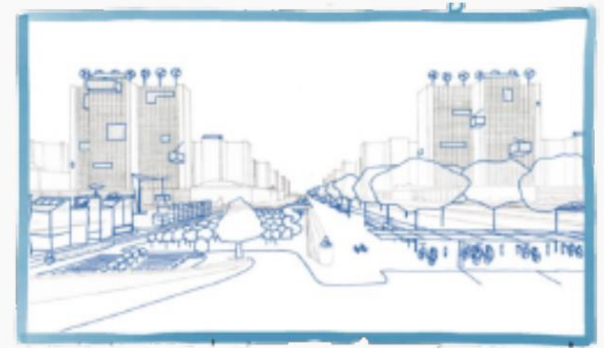
Detecting malaria from blood smears

Drug Discovery



Quickly discovering new drugs for COVID

Urban Planning



Predicting and planning for resource need
Agriculture



Precision agriculture

Why Data Science?

The world is complicated! Decisions are hard.

- It's a lot of fun!
- You will be at the cutting edge of research and product
- You will make lots of money doing something you will enjoy.
- It's not that hard to start and do!

After this course, you should be able to take data and produce useful insights on the world's most challenging and ambiguous problems.



Why Data Science?

Jobs!


50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? [Find out how.](#)

United States | 2017 | 12k Shares | [f](#) [t](#) [in](#) [✉](#)

1 Data Scientist



4.8 / 5
Job Score

\$110,000
Median Base Salary

4.4 / 5
Job Satisfaction

4,184
Job Openings

[View Jobs](#)

2 DevOps Engineer

What Is Data Science?

Once Upon a Time

Long time ago (thousands of years) science was only empirical and people counted stars



Then Came Theory

Few hundred years ago: theoretical approaches, try to derive equations to describe general phenomena.

$$F = G \frac{m_1 m_2}{d^2}$$

$$i\hbar \frac{\partial}{\partial t} \Psi = \hat{H} \Psi$$

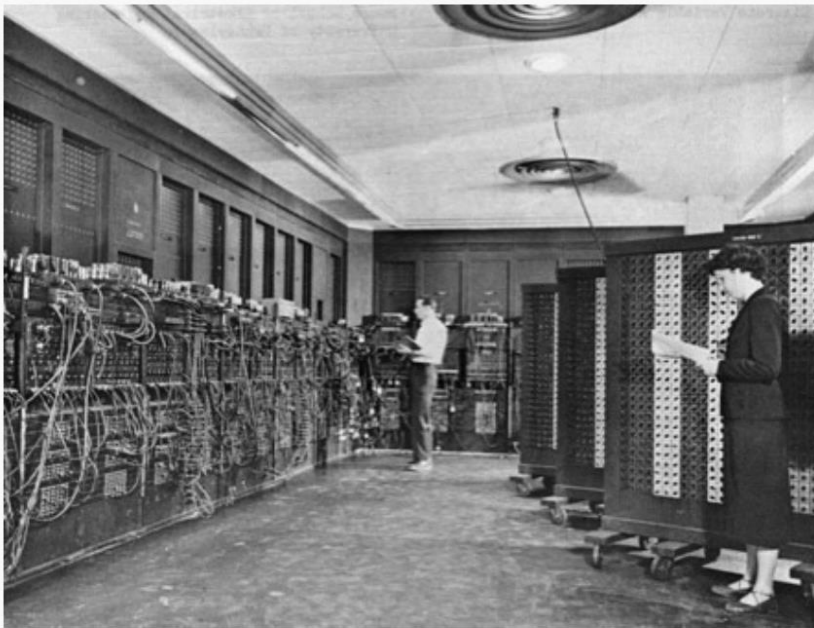
$$\begin{aligned} \nabla \cdot E &= 0 & \nabla \times E &= -\frac{1}{c} \frac{\partial H}{\partial t} \\ \nabla \cdot H &= 0 & \nabla \times H &= \frac{1}{c} \frac{\partial E}{\partial t} \end{aligned}$$

$$E = mc^2$$

$$\rho \left(\frac{\partial v}{\partial t} + v \cdot \nabla v \right) = -\nabla p + \nabla \cdot T + f$$

After that: Computers

About a hundred years ago: computational approaches appeared

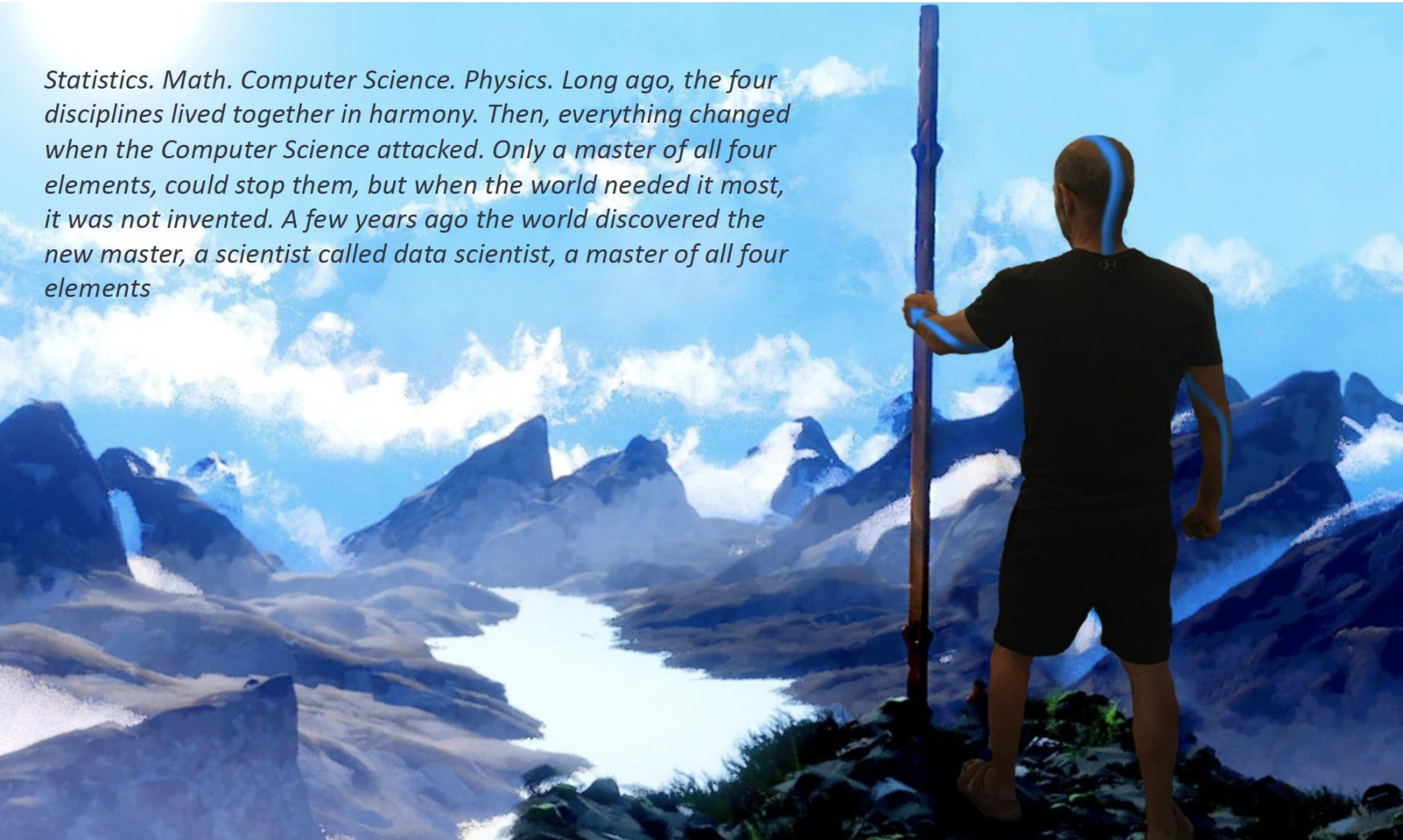


“The purpose of computing is insight, not numbers.”

R. Hamming. *Numerical Methods for Scientists and Engineers* (1962).

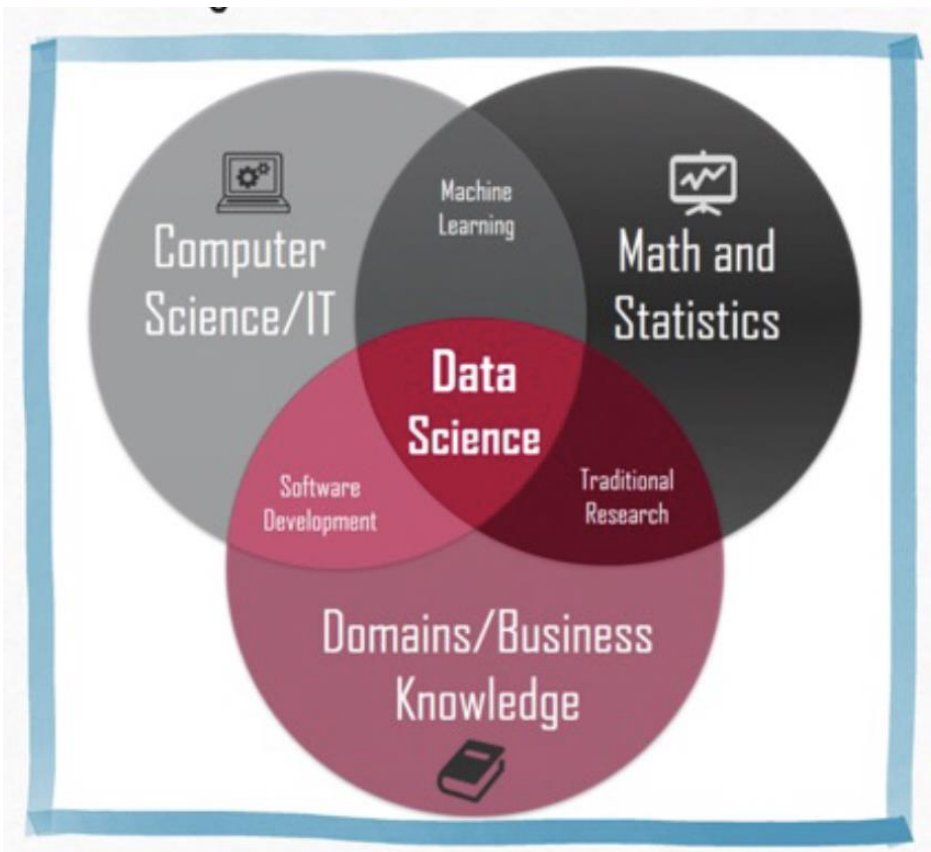
Behold: The Master

Statistics. Math. Computer Science. Physics. Long ago, the four disciplines lived together in harmony. Then, everything changed when the Computer Science attacked. Only a master of all four elements, could stop them, but when the world needed it most, it was not invented. A few years ago the world discovered the new master, a scientist called data scientist, a master of all four elements



Data Science

In both data science and machine learning we extract pattern and insights from data.

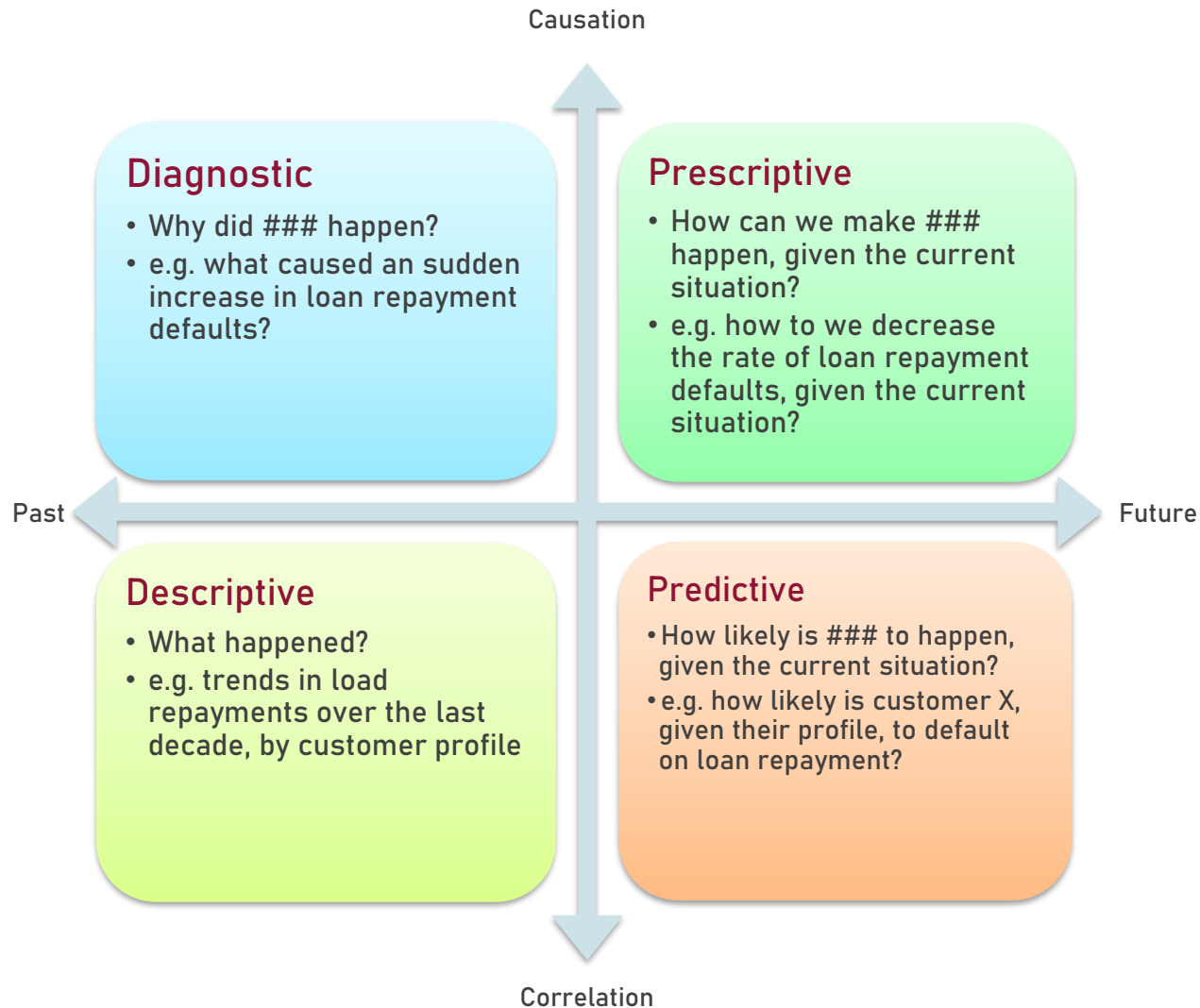


- Inter-disciplinary
- Data and task focused
- Resource aware
- Adaptable to changes in the environment and needs

One Definition of Data Science

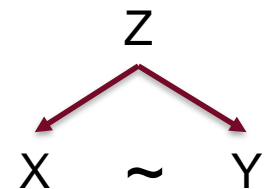
- From *The Field Guide to Data Science*:
 - The art of turning data into actions,
 - accomplished through the creation of data products, which...
 - provide actionable information...
 - without exposing decision makers to the underlying data or analytics.

Four Types of Data Science Projects

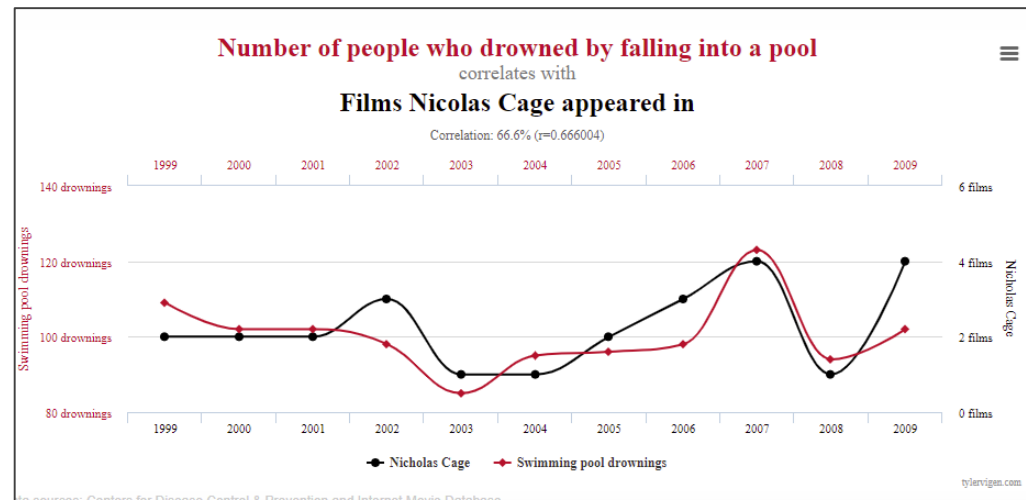


Correlation vs. Causation

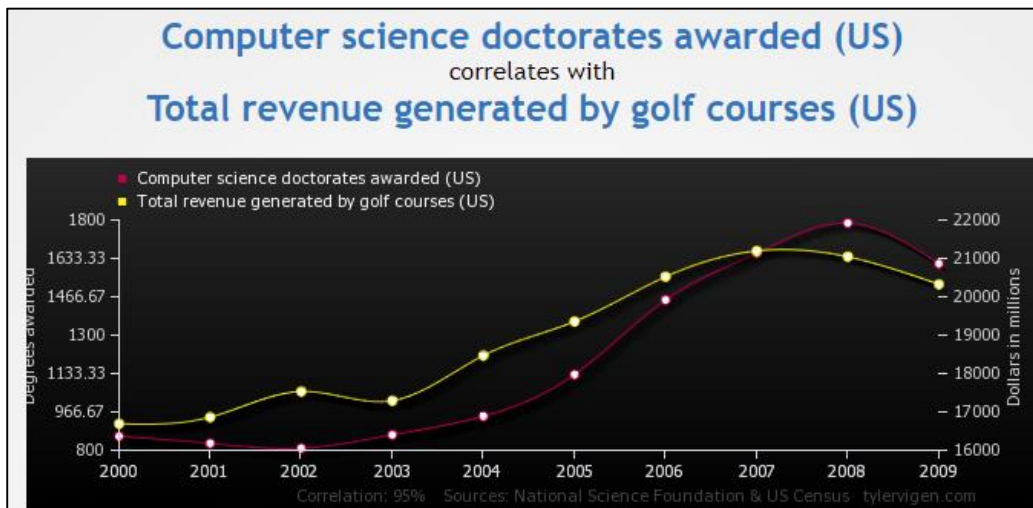
- Correlation \neq Causation!
- Correlation may imply causation,
- But causality between two events can be hard to establish rigorously. Given 2 correlated events X and Y, we need to show:
 - that X came before Y
 - that the relationship is not a coincidence
 - that nothing else can explain the relationship between X and Y (“confounding variable”)
- Requires comparisons controlling for external variables (A/B testing)



Correlation vs. Causation

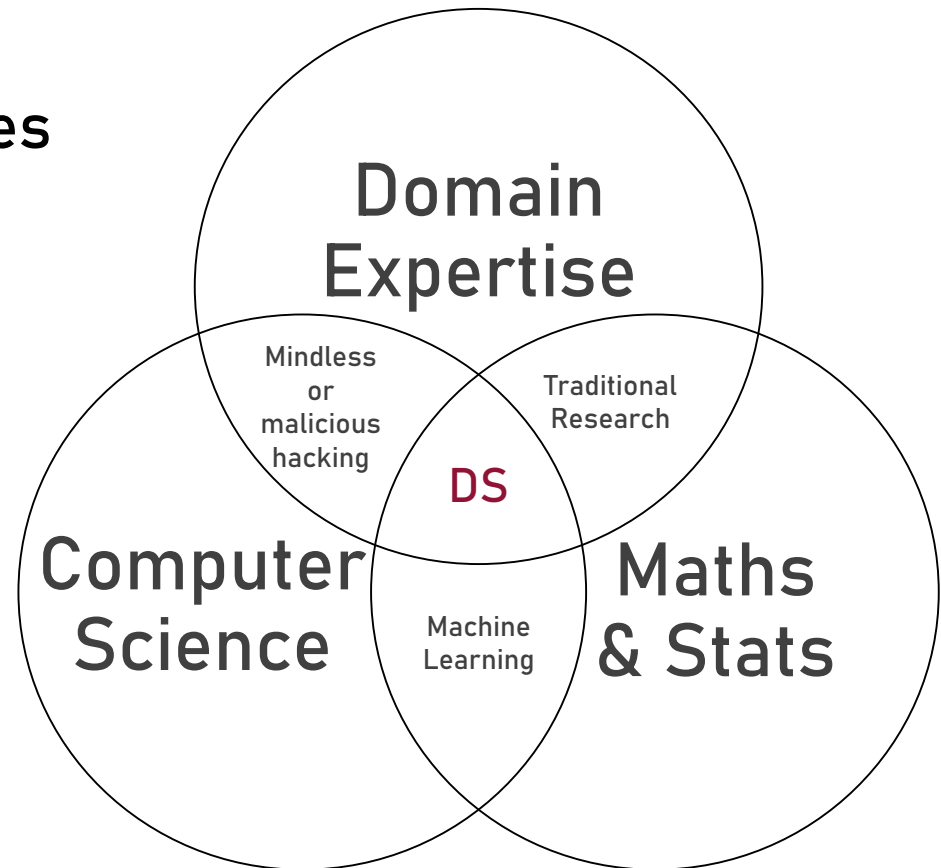


<https://www.tylervigen.com/spurious-correlations>



- **Domain expertise**
 - Ask the right questions
 - Form relevant hypotheses
 - Provide usable answers
- **Mathematics, statistics**
 - Guide analysis
 - Extract insights
 - Test hypotheses
- **Computer science**
 - Extract and manipulate data
 - Write reusable code
 - Productivity

The Data Science Venn Diagram



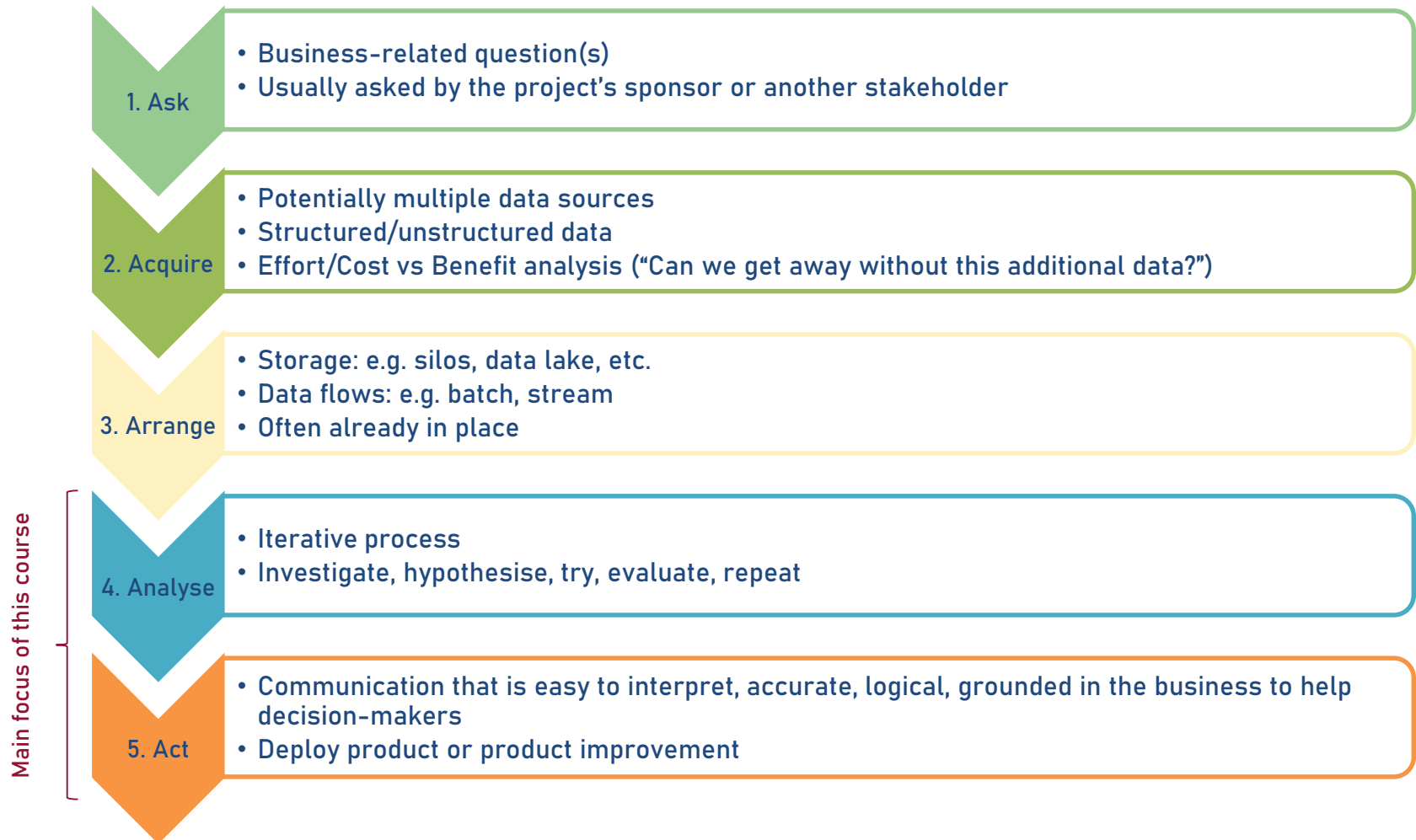
Adapted from <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Key Qualities of a Data Scientist

- **Curiosity**
 - “Why...?”, “What if...?”
- **Creativity**
 - e.g. creating new predictors, clever visualisations
- **Pragmatism**
 - Start simple
 - Complexify if there is a compelling reason to do so
- **Honesty**
 - Do not start with preconceived ideas of what you want to find
- **Organisation**
 - Structure your project to avoid getting lost or side-tracked
 - Document everything!
- **Team work**
 - Always a group initiative
- **Verbal and Written Communication**
 - To investigate, gain domain knowledge, present results, recommend actions...

Phases of a Data Science Project

Phases of a Data Science Initiative

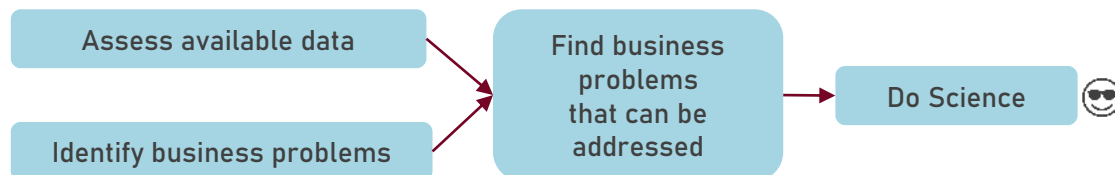


1. Ask

- Initial question usually comes from project sponsor
- In mature organisations, corresponds to a business problem
 - “Does the new version of our website increase customer spend?”
 - “Predict each customer’s probability of leaving us.”



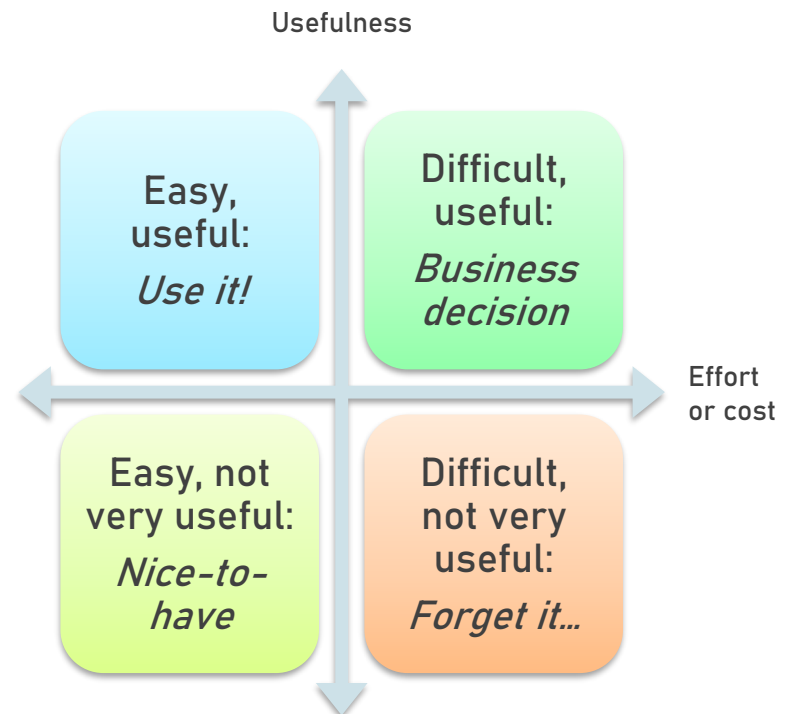
- In less mature organisations:
 - “We have all this data, what can we do with it?”



2. Acquire

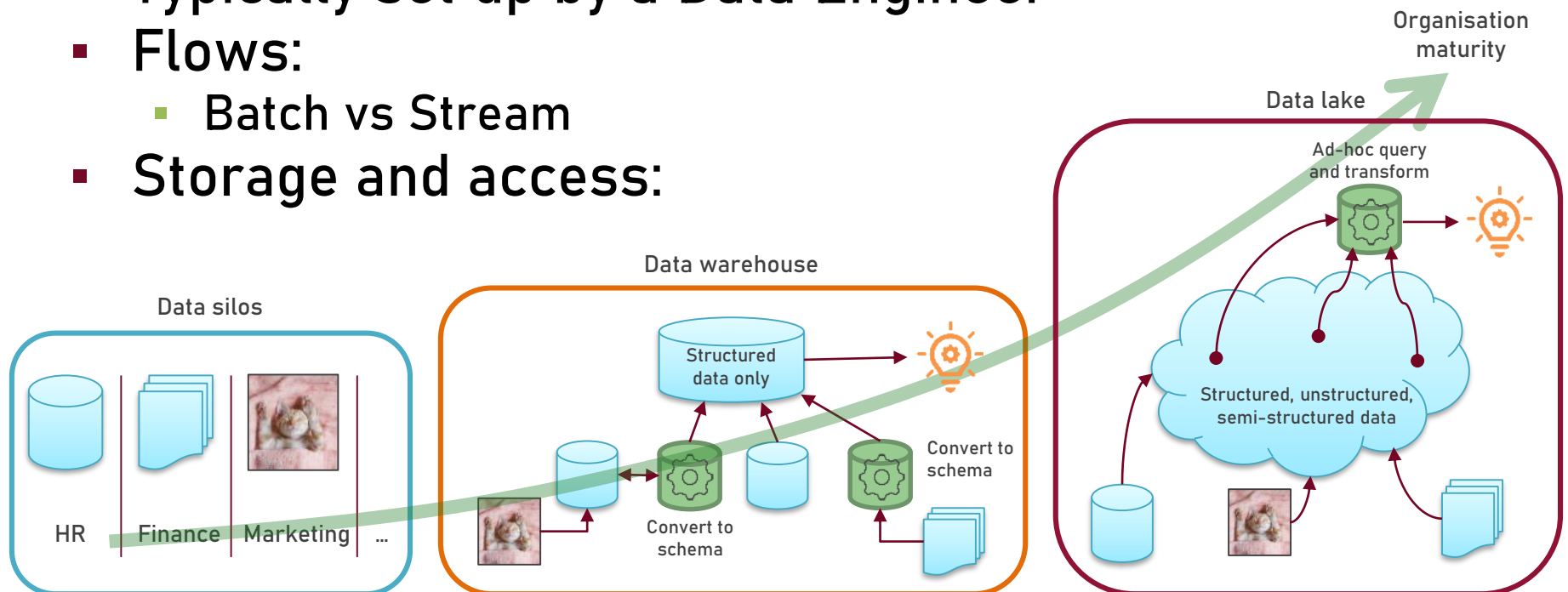
- What data do we already collect?
- What format is the data?
 - Structured, unstructured
- Can/Should we collect additional data?
 - e.g. from additional sensors, from more customers, from external sources, etc.
- How difficult/expensive is it to collect and use the data we want?
- Does it help us achieve significantly better outcomes?
 - Often only found out by trial-and-error during the prototyping phase

Cost vs. Benefit Analysis



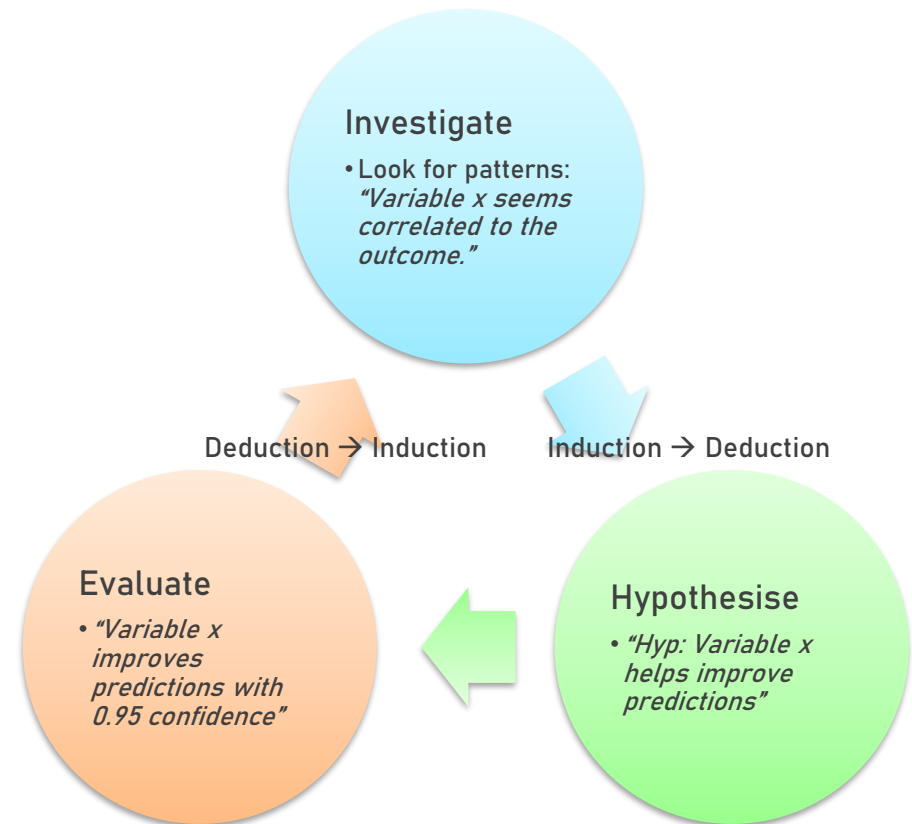
3. Arrange

- Prepare the infrastructure for data flow, storage, and access
- Often already in place and largely immutable
- Typically set up by a Data Engineer
- Flows:
 - Batch vs Stream
- Storage and access:



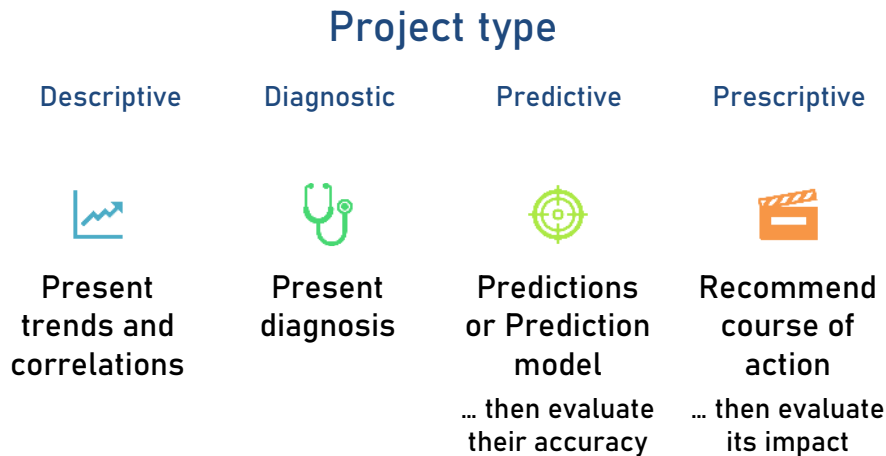
4. Analyse

- Deductive reasoning (logic-based):
 - Form hypotheses about relationships and underlying models
 - Run experiments to test hypotheses and models
- Inductive reasoning (pattern-based):
 - Explore data to discover candidate hypotheses to test
 - Discover possible relationships and insights from the data



5. Act

WHAT A DATA SCIENTIST MUST DELIVER



HOW TO DELIVER IT

- Little pre-requisite knowledge of statistics / analytics
- Easy to see and interpret
- Accurate
- Clear and compelling logic
- Grounded in the business

→ Allow stakeholders to make decisions

→ Deploy product / improvement



- What problem are we trying to solve?
- Does our approach make sense?
- Does our analysis address the original intent?
- Do our answers make sense?
 - If a result is too surprising, it is most likely false, even if it is convenient...
- Can we reach the same answers in a simpler way?
- Do our answers address the whole problem?
- What do we do with this knowledge?

The material of the course will integrate the five key facets of an investigation using data:

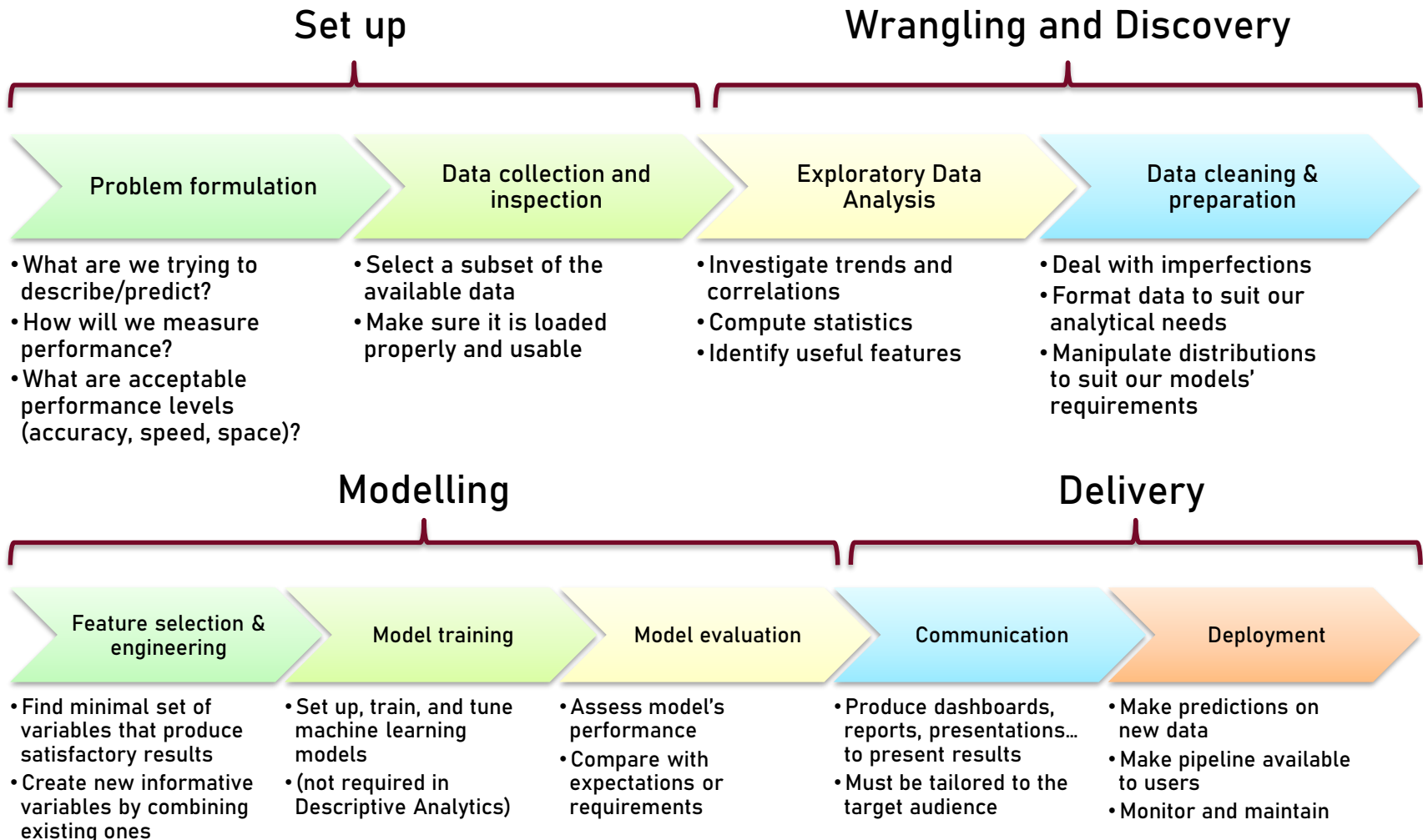
1. **Data collection:** data wrangling, cleaning, and sampling to get a suitable data set.
2. **Data management:** accessing data quickly and reliably.
3. **Exploratory data analysis;** generating hypotheses and building intuition.
4. **Prediction or statistical learning.**
5. **Communication:** summarizing results through visualization, stories, and interpretable summaries.

The Data Science Pipeline

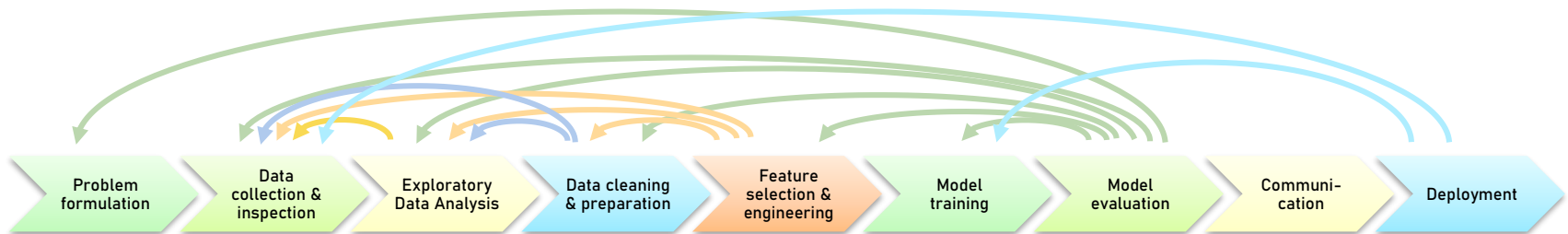
The Story So Far

- We have a business question to solve
 - e.g. Predict the probability of default of each loan applicant
- We are collecting data needed
 - e.g. Personal and financial data about applicants, data about the loans themselves, and macro-economic data
- We have data storage and flow systems in place
 - e.g. Data warehouse with historical loan and reimbursement data, with a SQL interface
- We are ready to build and deliver our data product
 - This section is about the Analyse phase, and the Data Scientist's role in the Act phase

Theoretical Pipeline



Actual Pipeline

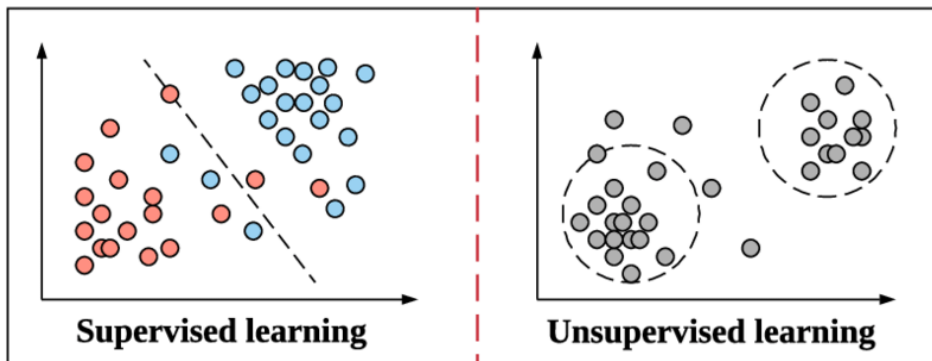


- Data Science is iterative and cyclical in nature
 - Each step depends on the previous **and** the next ones
 - Be organised!
 - Write flexible and modular code!

Different Modelling Tasks

Main Types of Statistical Learning

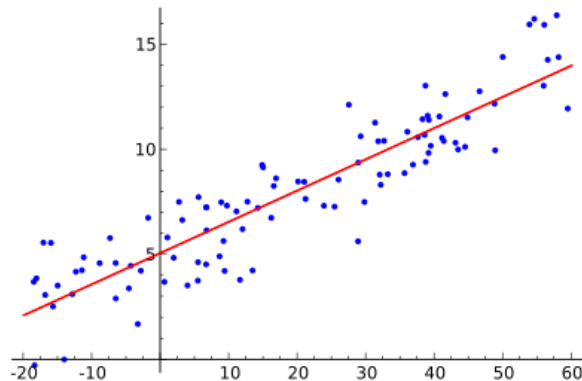
- The modelling phase includes EDA, model training (aka statistical learning) and model evaluation
- 2 main types of statistical learning:
 - Supervised: Provide the model with examples including the correct answer, and the model learns to produce the correct prediction
 - Unsupervised: Provide the model with data, the model learns about the structure of the data (e.g. clusters)



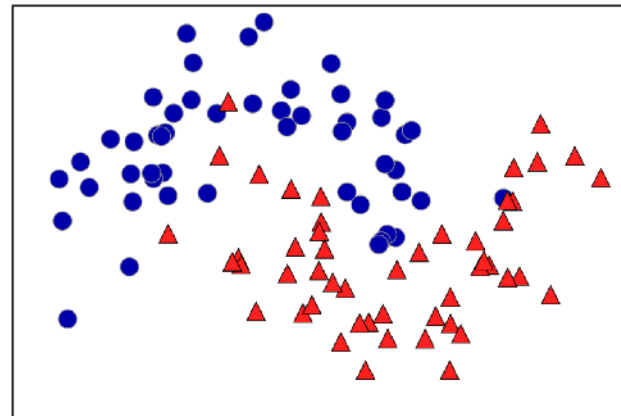
B. Qian et al. (2019). Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey.

Modelling: Main Types of Supervised Learning

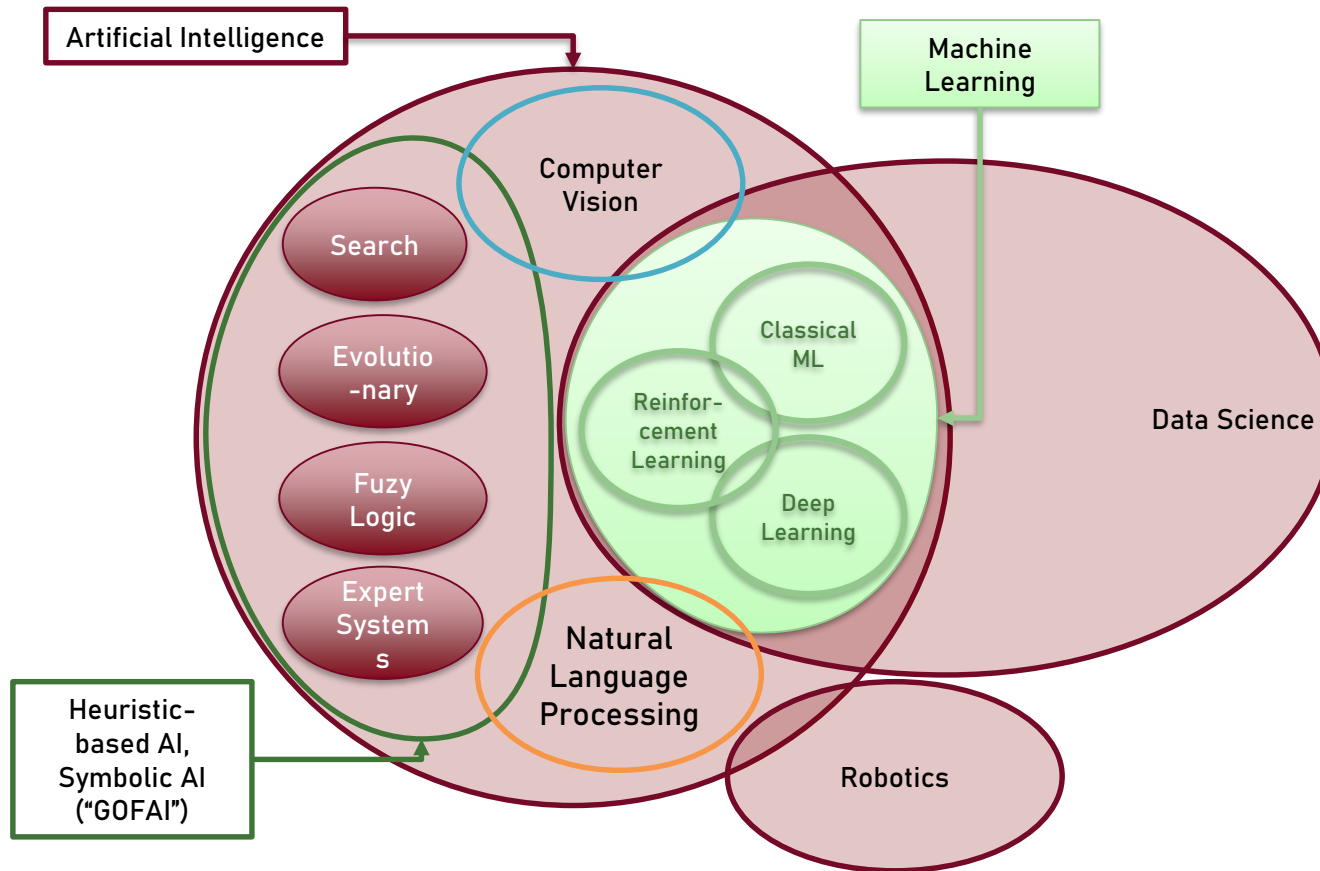
- Within supervised learning, 2 main tasks:
 - Regression:
 - Predict a value from variables, e.g. house price, satisfaction rating, etc.
 - Classification:
 - Predict a categorical label from variables, e.g. spam/not spam, cat/dog, benign/malignant tumour, etc.



Wikipedia



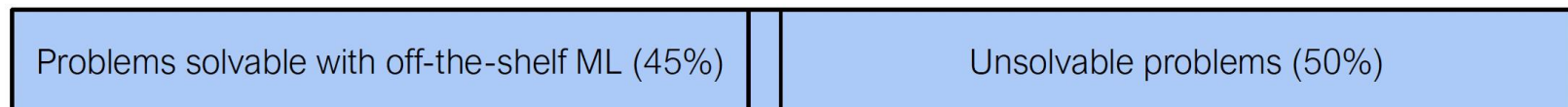
Modelling: AI, ML and DS



Note the numerous overlaps!

Data Science is Not Machine Learning

- Machine learning involves computation and statistics, but has not (traditionally) been very concerned about answering *scientific questions*
- Machine learning has a heavy focus on fancy algorithms...
- ... but sometimes the best way to solve a problem is just by visualizing the data, for instance



Problems requiring state-of-the-art ML (5%)

- “Analyzing data computationally, to understand some phenomenon in the real world, you say? ... that sounds an awful lot like statistics”
- Statistics (at least the academic type) has evolved a lot more along the mathematical/theoretical frontier
- Not many statistics courses have a lecture on e.g., web scraping, or a lot of data processing more generally
- Plus, statisticians use R, while data scientists use Python ... clearly these are completely different fields

Product Development vs. Deployment

Typical Steps in DS Research & Development

- Define the problem
- Assess available data
- Literature search for possible approaches
- Choose one approach, then:
 - Explore the data
 - Preprocess the data
 - Train a machine learning algorithm
 - Evaluate the algorithm
 - Repeat
- Final evaluation

Typical Steps in DS Deployment

- Design ETL (extract-transform-load) pipeline to feed data to the trained ML model
- Host trained model on a server
- Serve the model's predictions to user
- But also: (ML Ops)
 - Log all operations
 - Monitor the data to detect possible changes in schema or quality
 - Monitor the model's prediction quality
 - Record the user's outcome
 - Regularly retrain the ML model
 - Keep versions of all model and data
 - etc...

Conclusion

- Data Science is becoming ubiquitous
- Has elements of science and art, grounded in the business, and aiming at taking action
- Can be descriptive, diagnostic, predictive, or prescriptive
- At the crossroads of math/stats, CS, domain expertise
- Requires technical, personal, interpersonal skills

- A DS initiative has 5 phases:
 - Ask: A business question, usually provided to us
 - Acquire: Weigh Cost/Effort vs. Benefit
 - Arrange: Data engineering, often a given
 - Analyse: Cyclical, iterative process switching between deductive and inductive reasoning
 - Act: Produce clear recommendations to support decisions, or deploy product
- The DS pipeline:
 - Set up, wrangle & discover, model, deliver
 - Iterative and entangled
- Deploying a data product is much more than just DS.

- Do not confuse correlation with causation
- Be your harshest critic
 - Double-check any result that is too good or too bad to be true
 - Be honest with yourself and others – a user will eventually stumble upon any half-truths (even unintentional)!
- Be very organised and tidy, document every decision, store all iterations of your pipeline
- Avoid data leakage!

References

References

- M. Herman, S. Rivera, S. Mills, J. Sullivan, P. Guerra, A. Cosmas, *The Field Guide to Data Science*, Booz Allen Hamilton, 2013
- <https://www.tylervigen.com/spurious-correlations>, consulted June 14, 2021
- <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>, consulted June 13, 2021
- Harvard's Introduction to Data Science
- Berkley's Data 100

Thank You!

NEW GIZA UNIVERSITY

