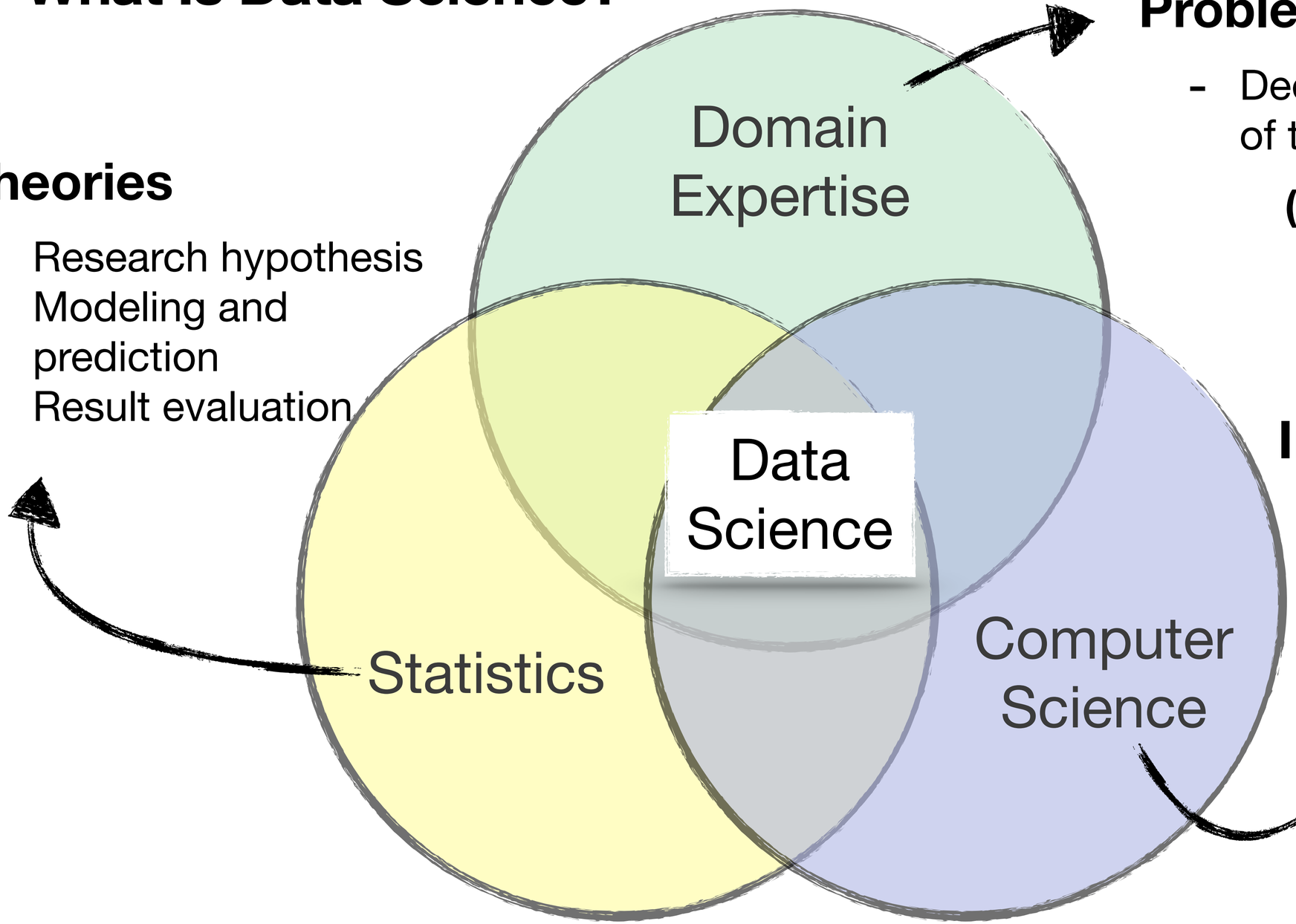
Mid Journey V4 prompt: Data Science, creative, inspiring, digital --ar 3:2 --upbeta --q 2 --v 4

Lecture 1-1: Introduction to Data Science

Dr. James Chen, Animal Data Scientist, School of Animal Sciences

2023 APSC-5984 SS: Agriculture Data Science

What Is Data Science?



Problem identification

- Deep understanding of the data

(KEY factor!)

Theories

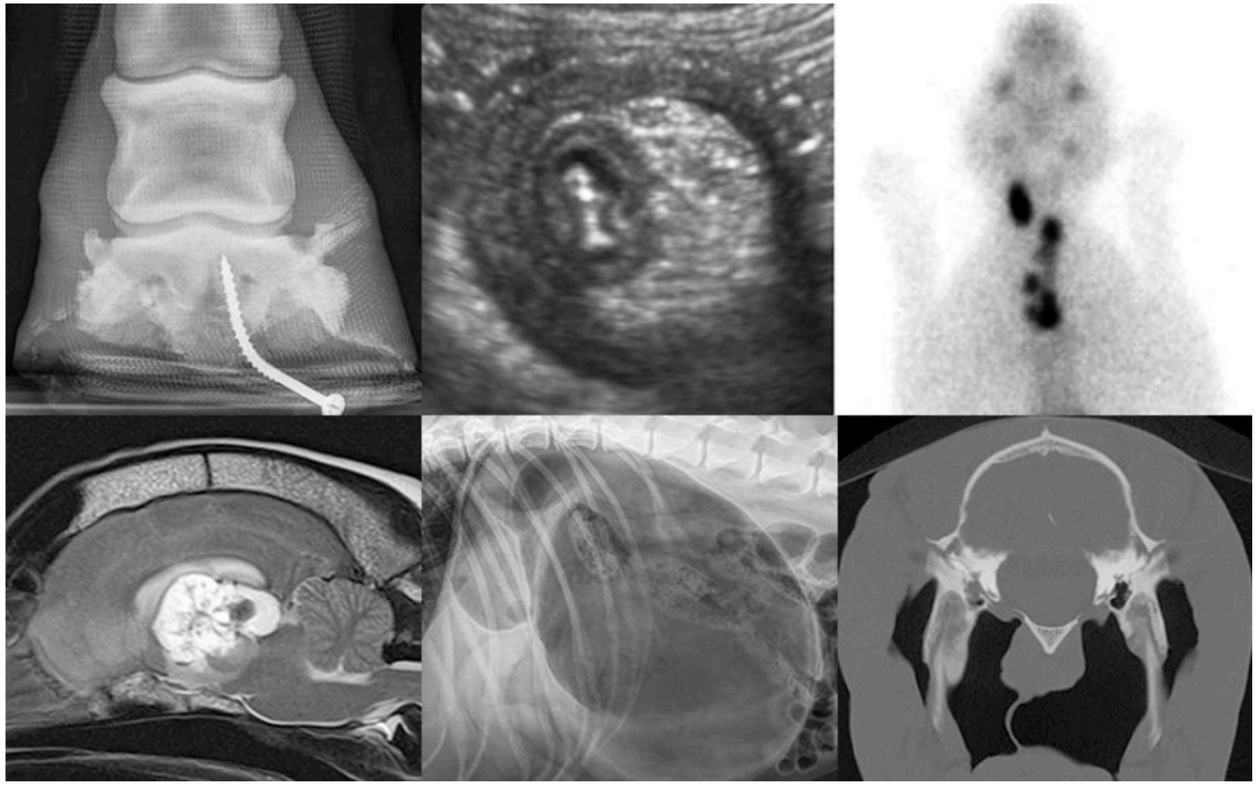
- Research hypothesis
- Modeling and prediction
- Result evaluation

Implementation

- Algorithm
- Functionality
- Optimization (performance, memory efficiency, etc)

Domain Expertise

Computed tomography (CT)



https://www.vet.cornell.edu/hospitals/services/imaging-0

Plant disease identification

Ahmed, N., Asif, H.M., & Saleem, G. (2021). Leaf Image-based Plant Disease Identification using Color and Texture Features. *ArXiv, abs/2102.04515*.

Statistics and Machine Learning (Generic Categories)

Statistical inferences

Supervised learning

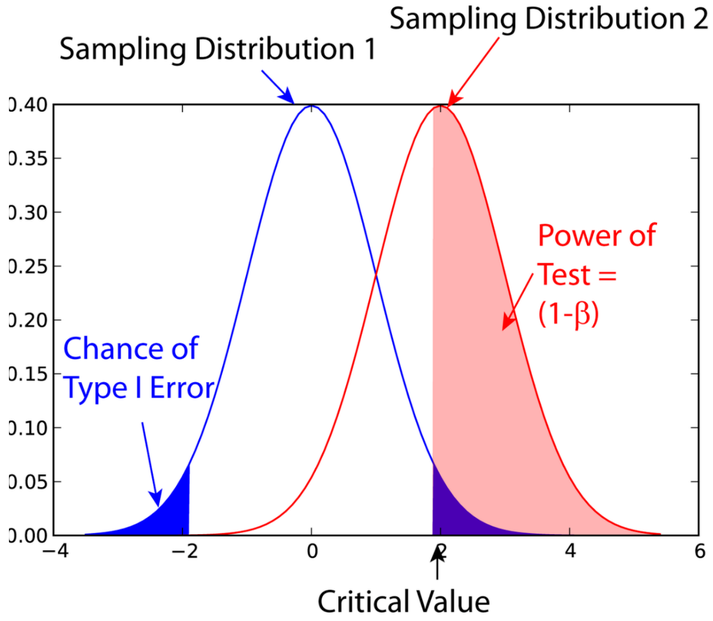
Unsupervised learning

Reinforcement learning

Hypothesis testing
(t-test, Chi-squared test)

ANOVA
(compare **within-group** and **between-group** variances)

Power analysis
(Determine the sample size)



And a lot more variants …

Statistics and Machine Learning

- Statistical inferences
- Supervised learning
- Unsupervised learning
- Reinforcement learning

X	**y**
Features (Predictors, independent variables)	Labels (Response, dependent variable)
It can be a matrix, an image, or a sequence of signals …	It can be a number, a prediction probability, or a category…

Parametric models → Linear model, deep learning model

Non-parametric models → Decision tree, random forest, boosting trees

More applications:
Image classification, event detection, prediction

Statistics and Machine Learning

Statistical inferences

Supervised learning

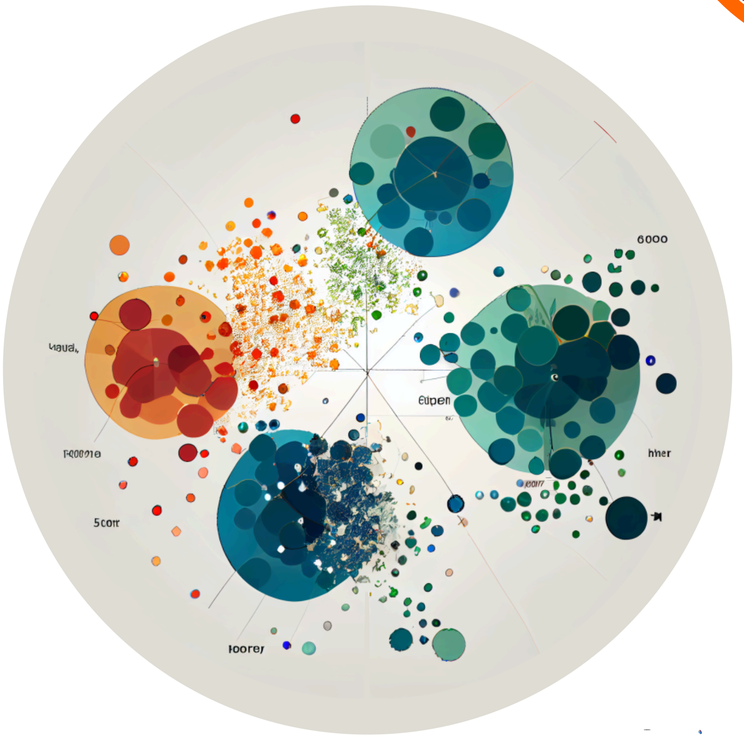
Unsupervised learning

Reinforcement learning

X

Features
(Predictors, independent variables)

It can be a matrix, an image,
or a sequence of signals …



Midjourney V4 prompt: clustering scatter plot --upbeta --q 2 --v 4

Clustering → K-means, hierarchical clustering

Dimensionality reduction → PCA, factor analysis, t-SNE

More applications:
Image classification, event detection, prediction

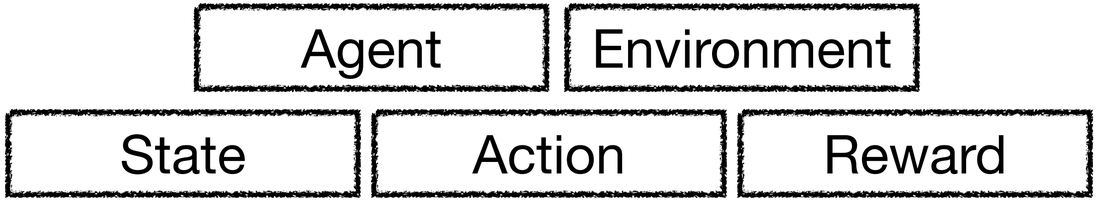
Statistics and Machine Learning

Statistical inferences

Supervised learning

Unsupervised learning

Reinforcement learning



Q-Learning

$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{current value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q(s_t, a_t)}_{\text{current value}} \right)}^{\text{temporal difference}}$$

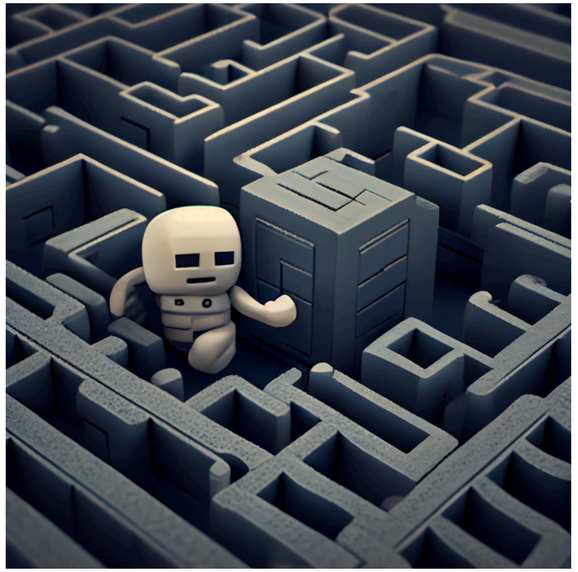
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

https://en.wikipedia.org/wiki/Q-learning

Chess AI

Midjourney V4 prompt: Reinforcement learning --upbeta --q 2 --v 4

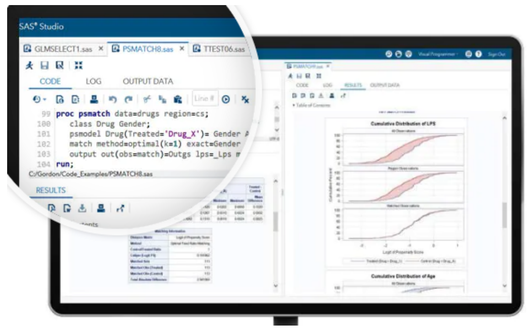
Maze-solving algorithm

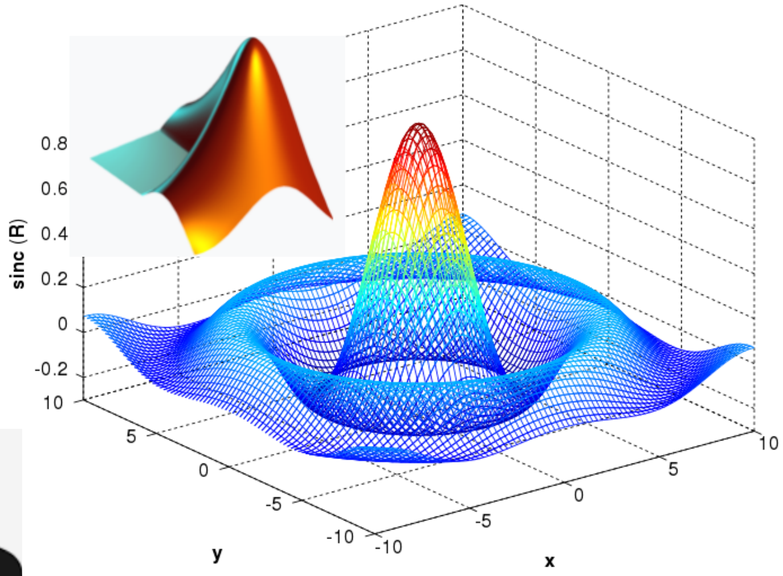


Midjourney V4 prompt: robot trying to escape a maze --upbeta --q 2 --v 4

Computer Engineering and Science

How do we implement the idea? Luckily, we have …





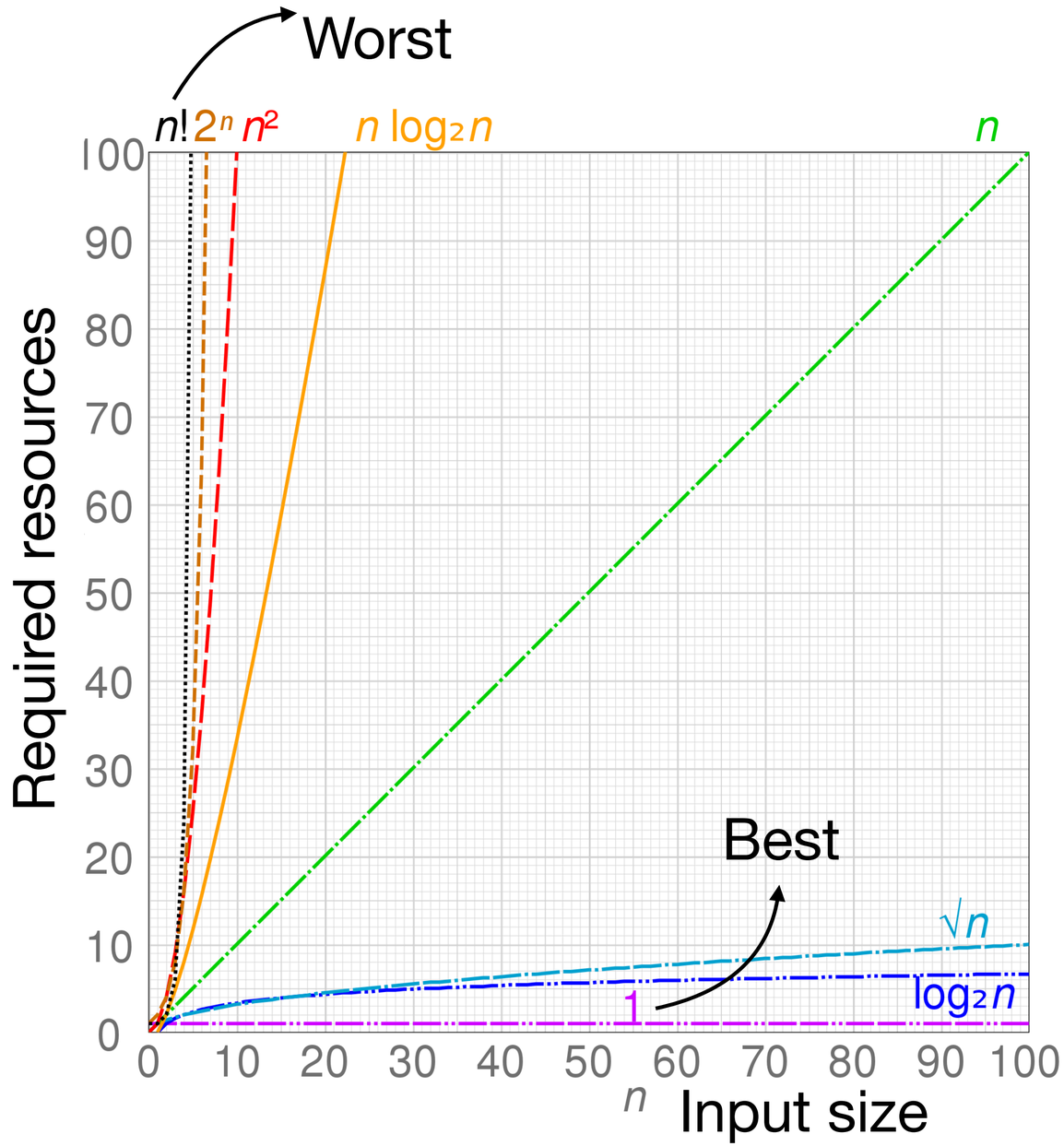
https://en.wikipedia.org/wiki/MATLAB

We will discuss how to choose a right tool (for you) in the next lecture

Computer Engineering and Science

Computational Complexity

- It's an upper bound of the required resource (time/memory)
- For example, when the complexity is $O(n^2)$. If we triple our sample size, the number of required time/memory will increase by a factor of 9 (3^2)
- Computing a **correlation matrix** is an example of time complexity $O(n^2)$. The number of operations increases with the **square** of the number of data points.

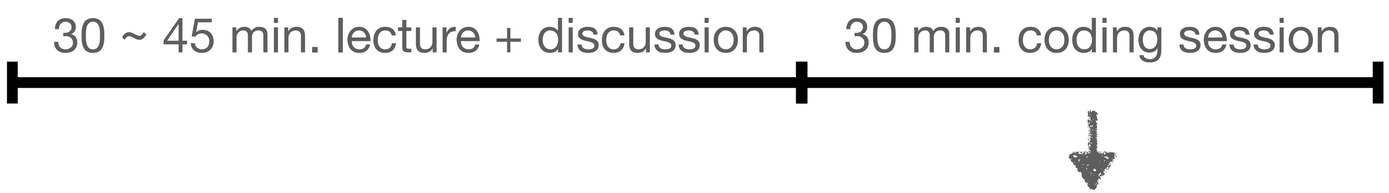


https://en.wikipedia.org/wiki/Time_complexity

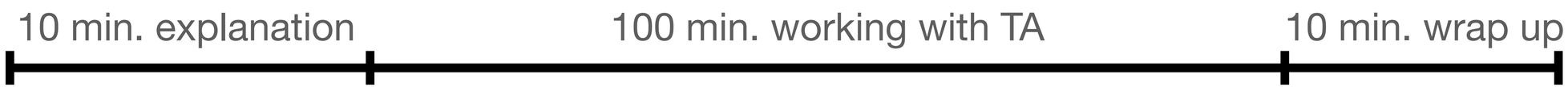
Course Overview

Course Format

- Lecture (75 minutes)



- Lab (120 minutes)



- You will be assigned a 2- to 3-page lab assignment (coding questions or questions that need you to elaborate)
- Attendance is **not required**, but you have to make sure to submit the assignment by the end of the day (23:59)

Course Overview

Grading

Assignment	60%
Final project	15%
Lab final exam	25%

Grade	Aggregate Score Range
A	94 – 100
A-	87 – 93
B+	80 – 86
B	75 – 79
B-	70 - 74
C+	65 – 69
C	60 – 64
F	< 60

Course Overview

Weekly schedule

Weeks	Lecture 1	Lecture 2	Lab
Week 1	Why Data Science?	Introduction to Python Programming	Basic Python implementation I
Week 2	Python operators	If statements and for-loops	Basic Python implementation II
Week 3	Data frames I	Data frames II	Pandas library
Week 4	Data preprocessing I	Data preprocessing II	PCDART database
Week 5	Data quality control	Introduction to database I	Futures market database
Week 6	Introduction to database II	Introduction to database III	Database application
Week 7	Design of database I	Design of database II	Construct a database
Week 8	Spring break		
Week 9	Supervised learning: regression model I	Supervised learning: regression model II	Python Scikit-learn
Week 10	Feature selection	Model validation	Feature selection
Week 11	Principal Component Analysis	Unsupervised learning: K-means clustering	Python Scikit-learn
Week 12	Intro to computer vision	Computer vision: convolution	Python OpenCV
Week 13	Data visualization	Web app deployment	Python Plotly
Week 14	What is object-oriented programming	Encapsulation	OOP implementation
Week 15	Object and Class in Python	Inheritance and polymorphism	OOP implementation
Week 16	Project presentation	Project presentation	Final exam