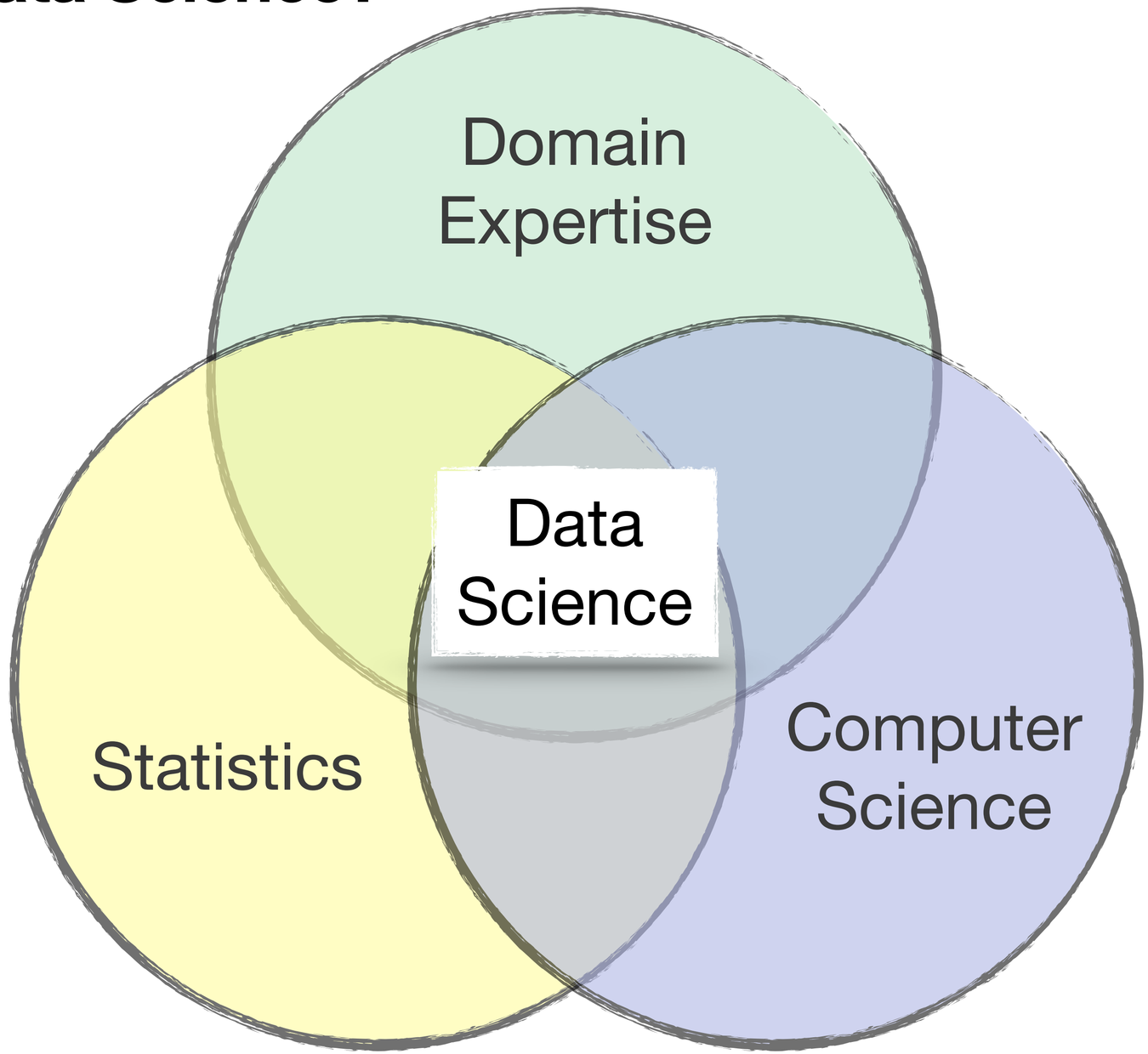
Mid Journey V4 prompt: Data Science, creative, inspiring, digital --ar 3:2 --upbeta --q 2 --v 4

Lecture 1-1: Introduction to Data Science

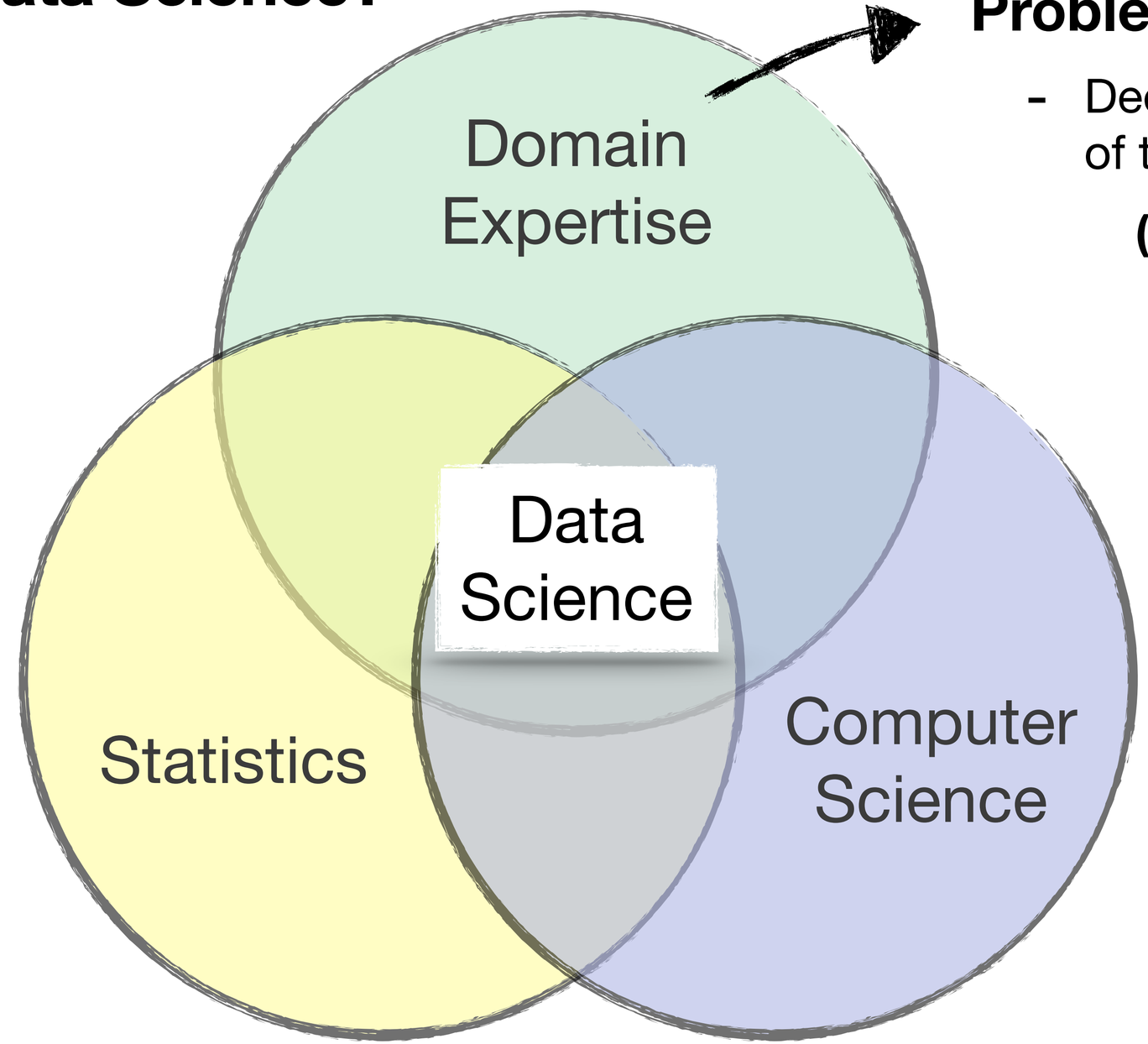
Dr. James Chen, Animal Data Scientist, School of Animal Sciences

2023 APSC-5984 SS: Agriculture Data Science

What Is Data Science?



What Is Data Science?



Problem identification

- Deep understanding of the data

(KEY factor!)

What Is Data Science?

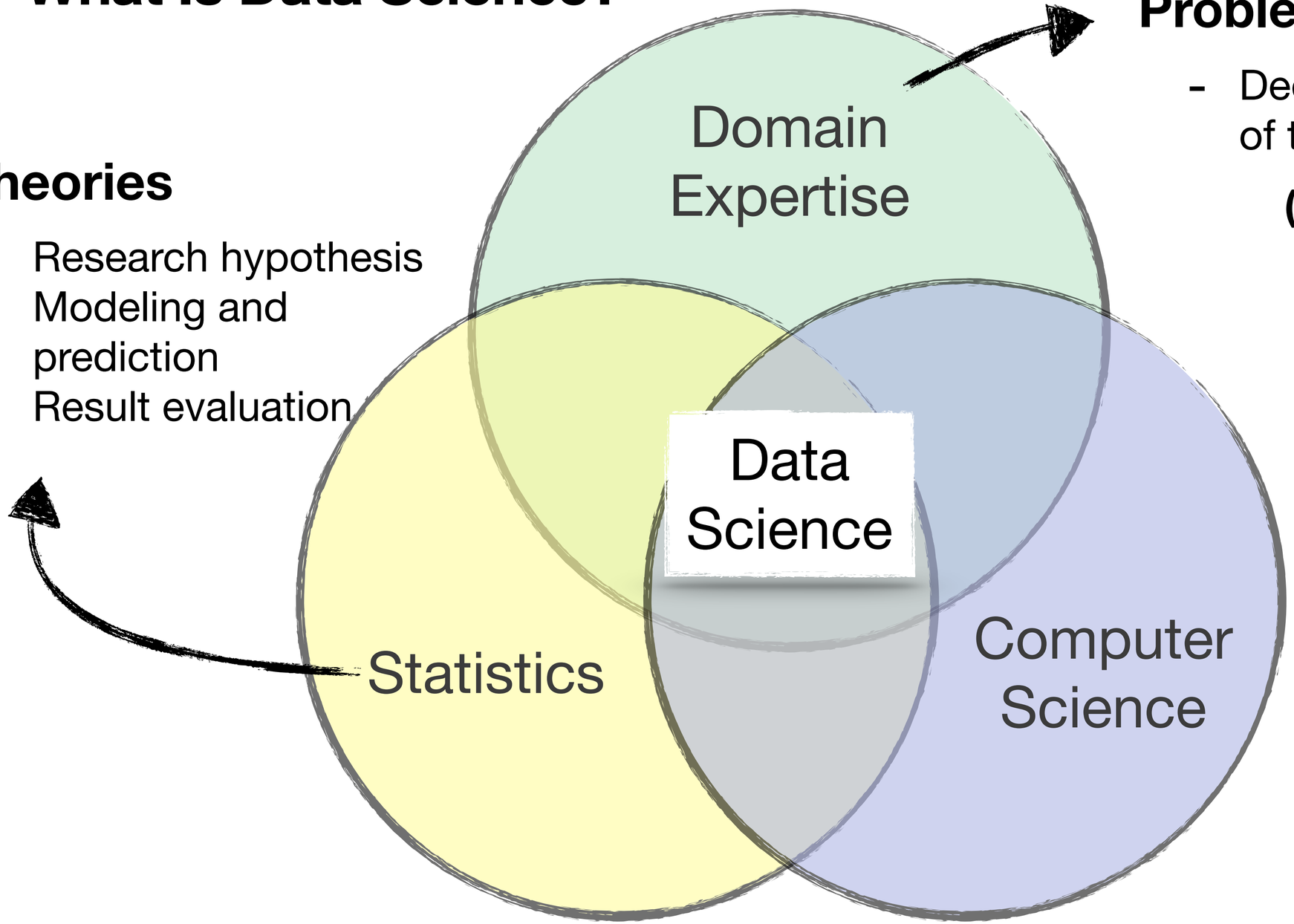
Problem identification

- Deep understanding of the data

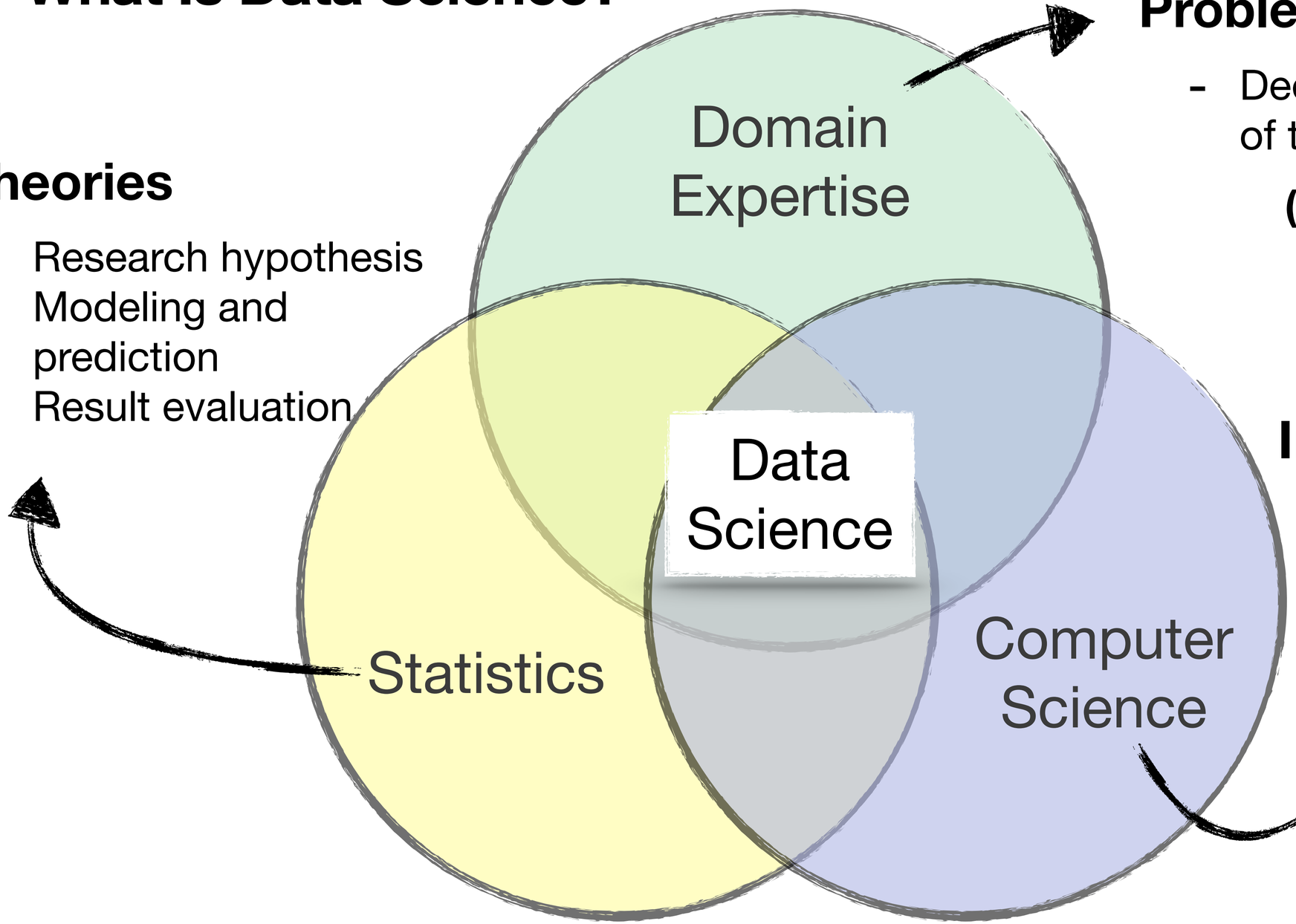
(KEY factor!)

Theories

- Research hypothesis
- Modeling and prediction
- Result evaluation



What Is Data Science?



Problem identification

- Deep understanding of the data

(KEY factor!)

Theories

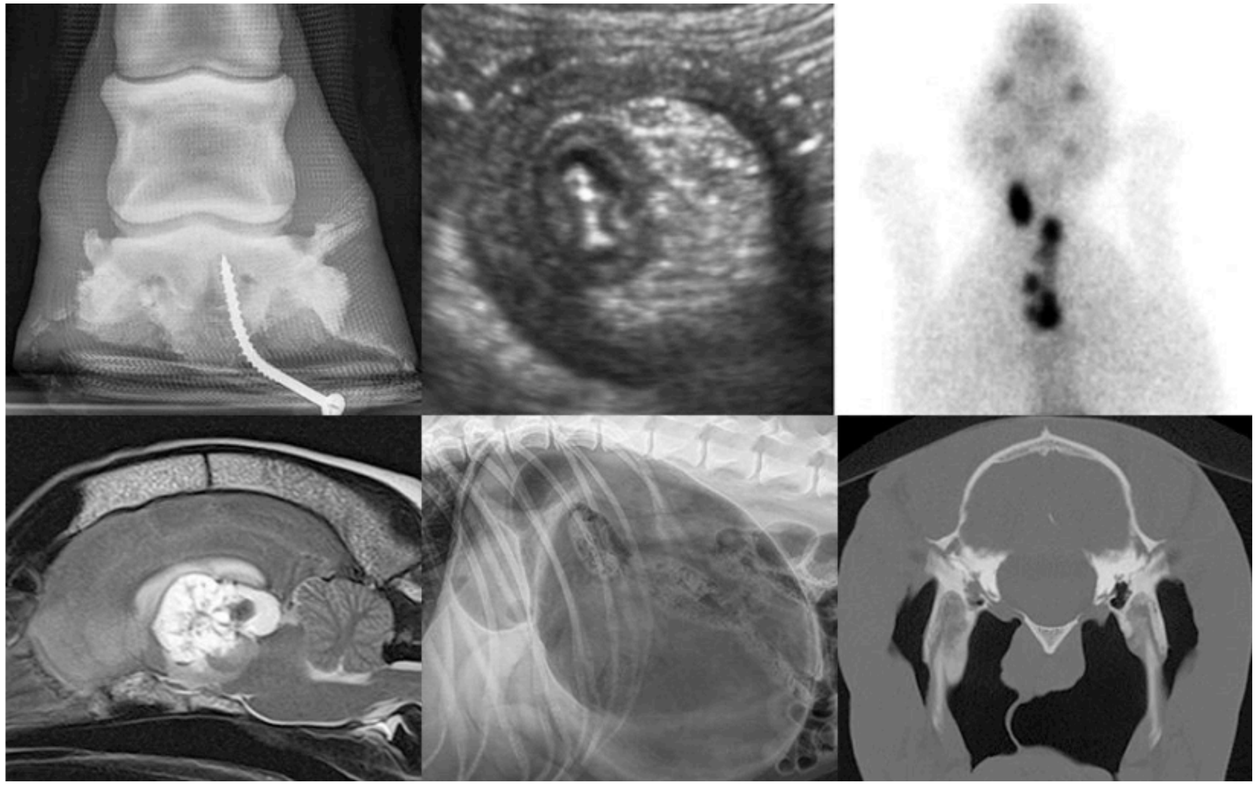
- Research hypothesis
- Modeling and prediction
- Result evaluation

Implementation

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

I. 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*.

II. Statistics and Machine Learning

Statistical inferences

Supervised learning

Unsupervised learning

Reinforcement learning

II. Statistics and Machine Learning

Statistical inferences

Supervised learning

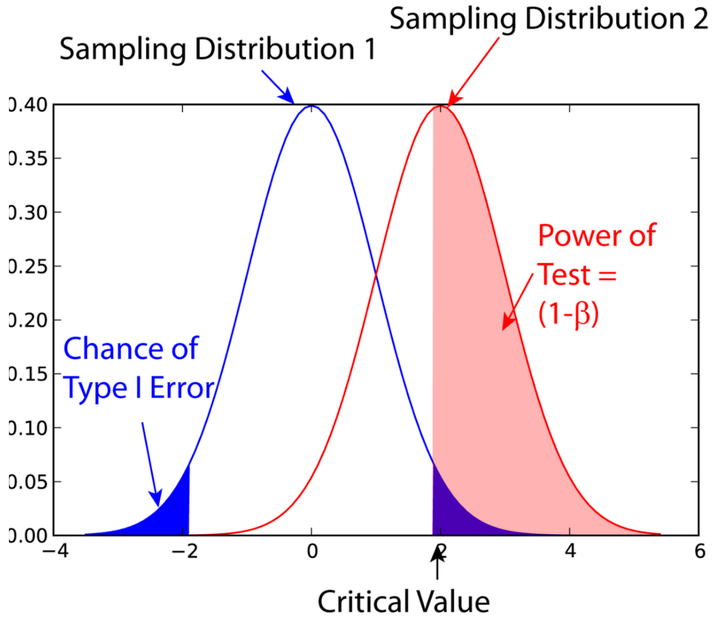
Unsupervised learning

Reinforcement learning

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

ANOVA
(To test whether the means among groups are equal)

Power analysis
(Determine the sample size)



And a lot more variants …

II. 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

II. Statistics and Machine Learning

Statistical inferences

Supervised learning

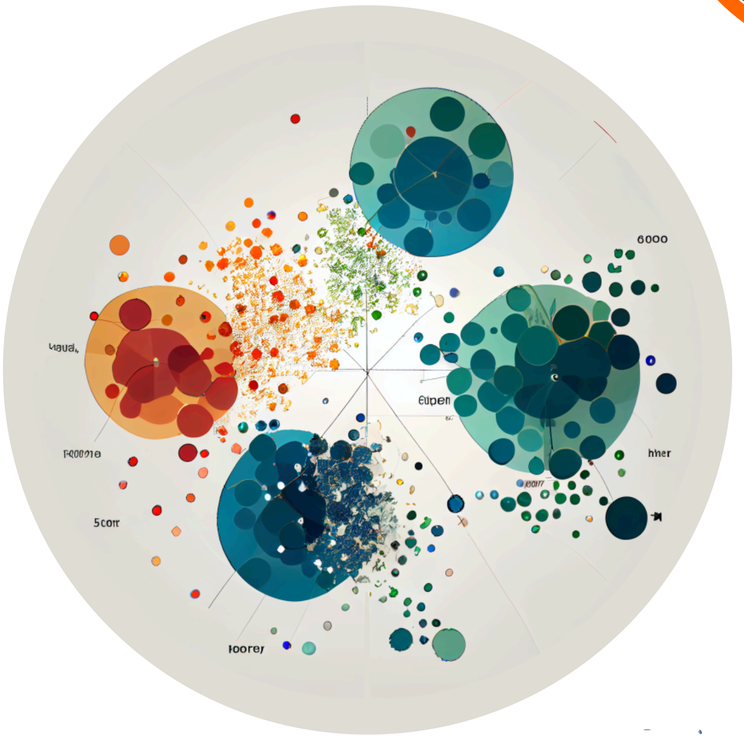
Unsupervised learning

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

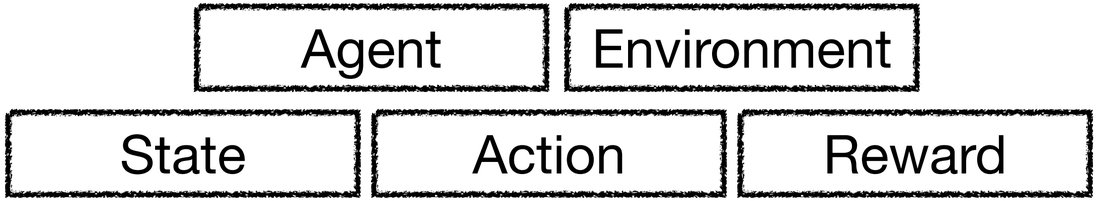
II. Statistics and Machine Learning

Statistical inferences

Supervised learning

Unsupervised learning

Reinforcement learning



Q-Learning

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

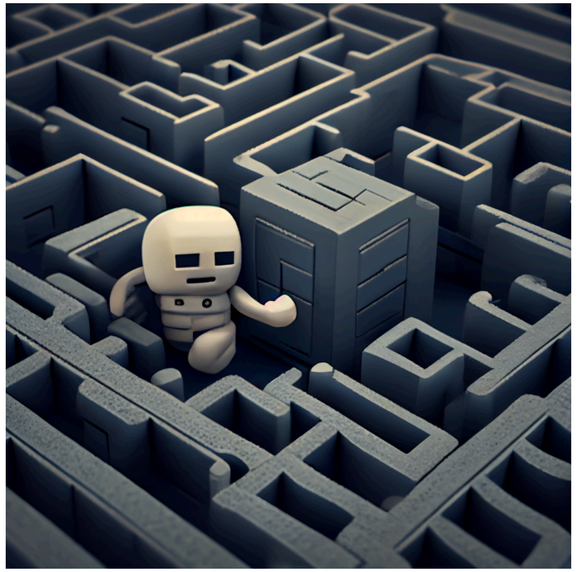
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

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

Chess AI

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

Maze-solving algorithm



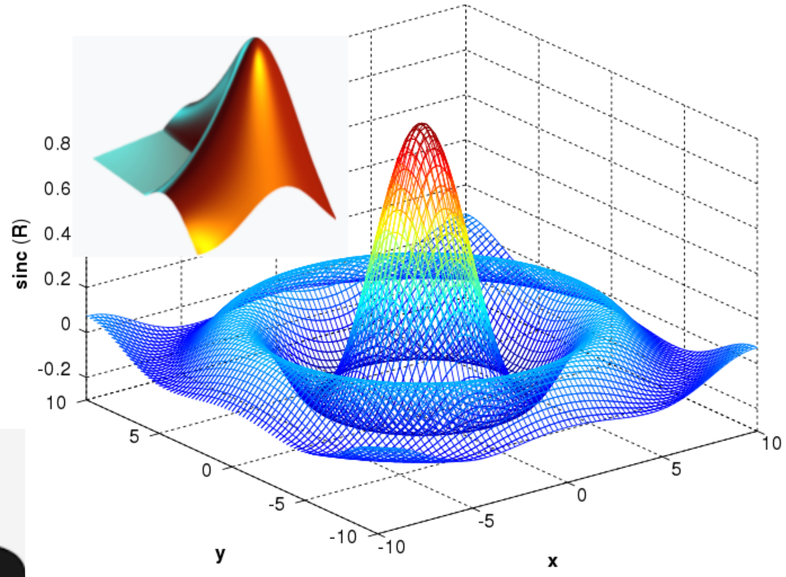
Midjourney V4 prompt: robot trying to escape a maze --upbeta --q 2 --v 4

III. Computer Engineering and Computer Science

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

III. Computer Engineering and Computer Science

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

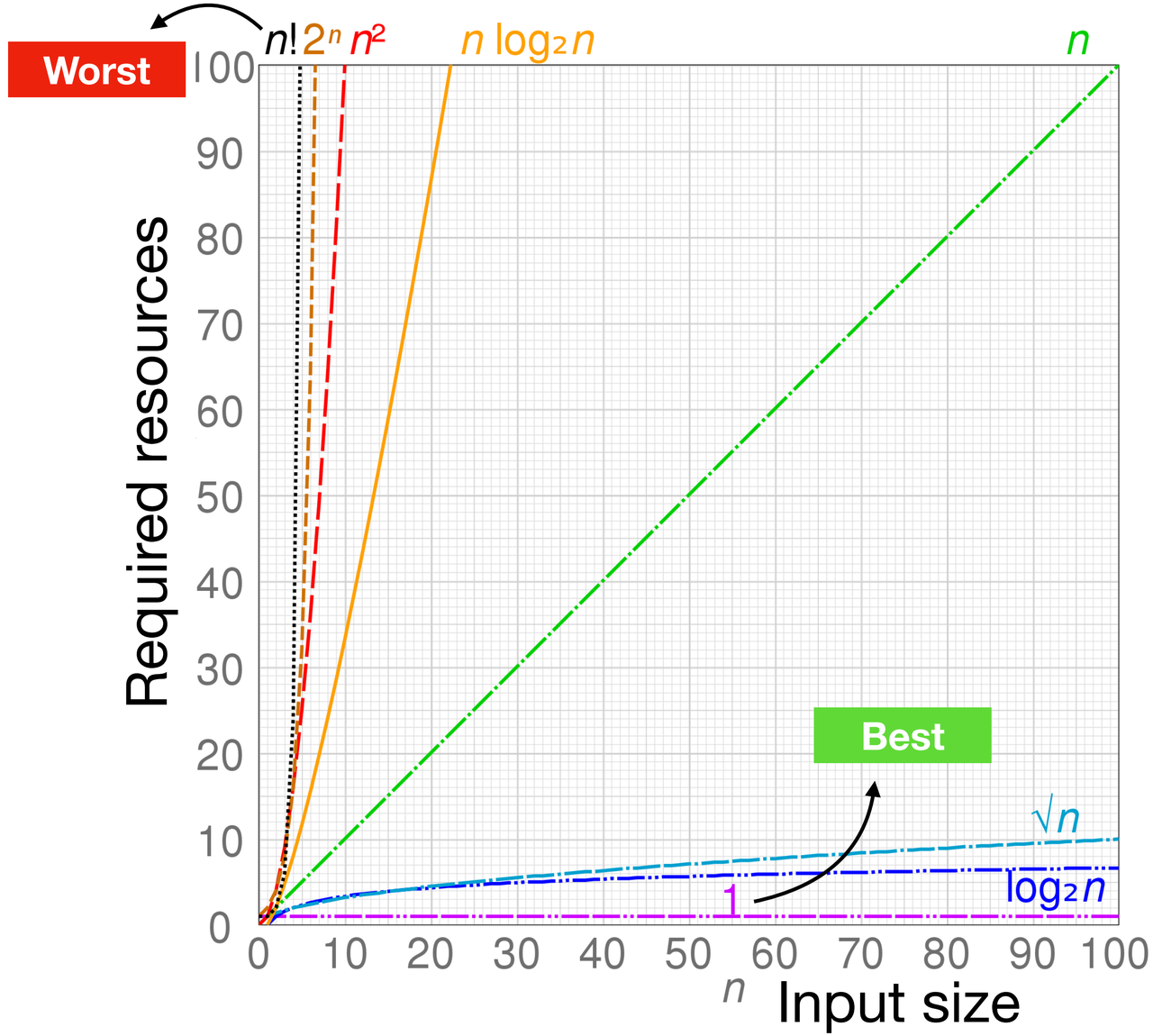


https://en.wikipedia.org/wiki/MATLAB

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

III. Computer Engineering and Computer Science

Computational Complexity

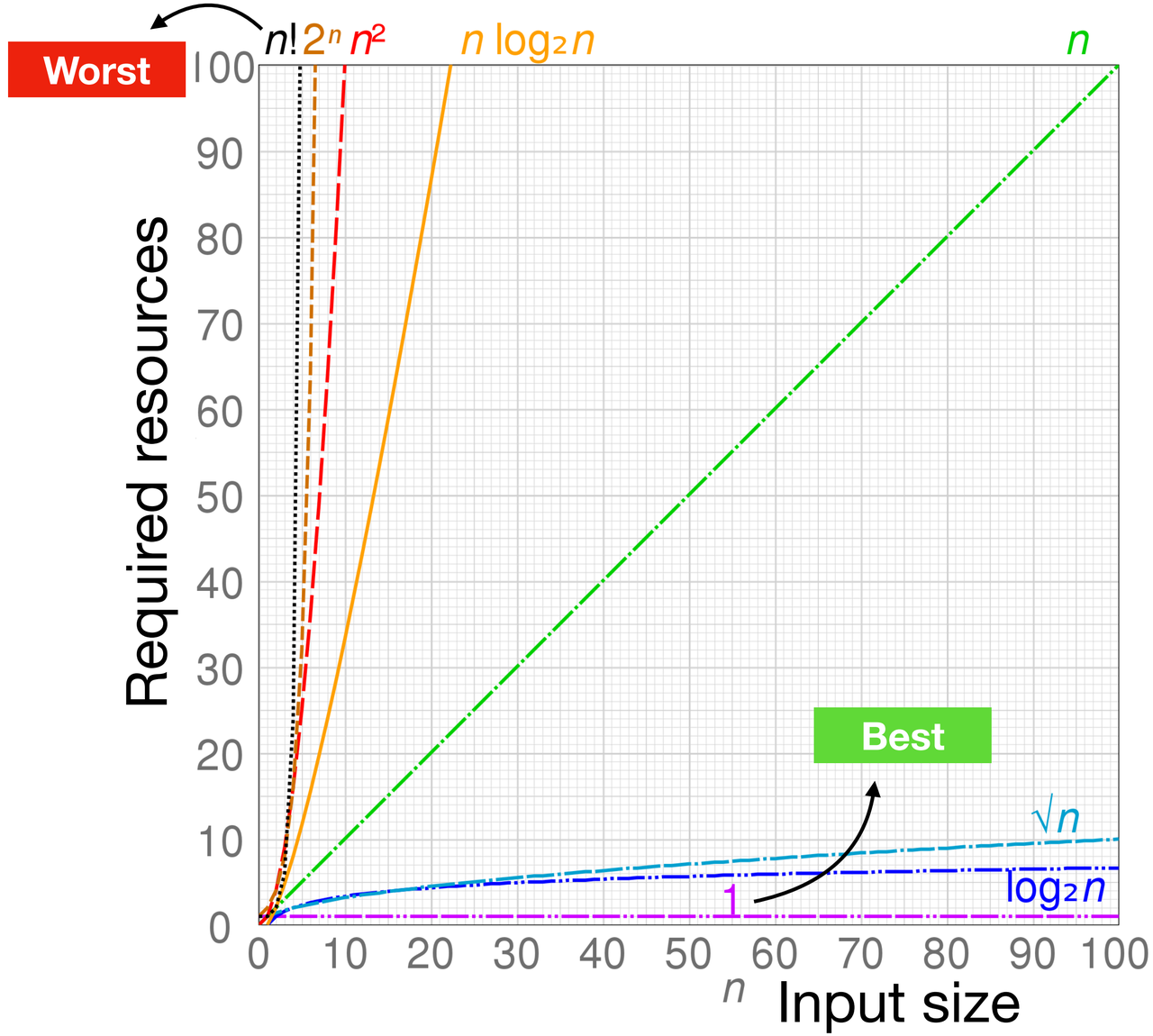


https://en.wikipedia.org/wiki/Time_complexity

III. Computer Engineering and Computer Science

Computational Complexity

- It's an upper bound of the required resource (time/memory)



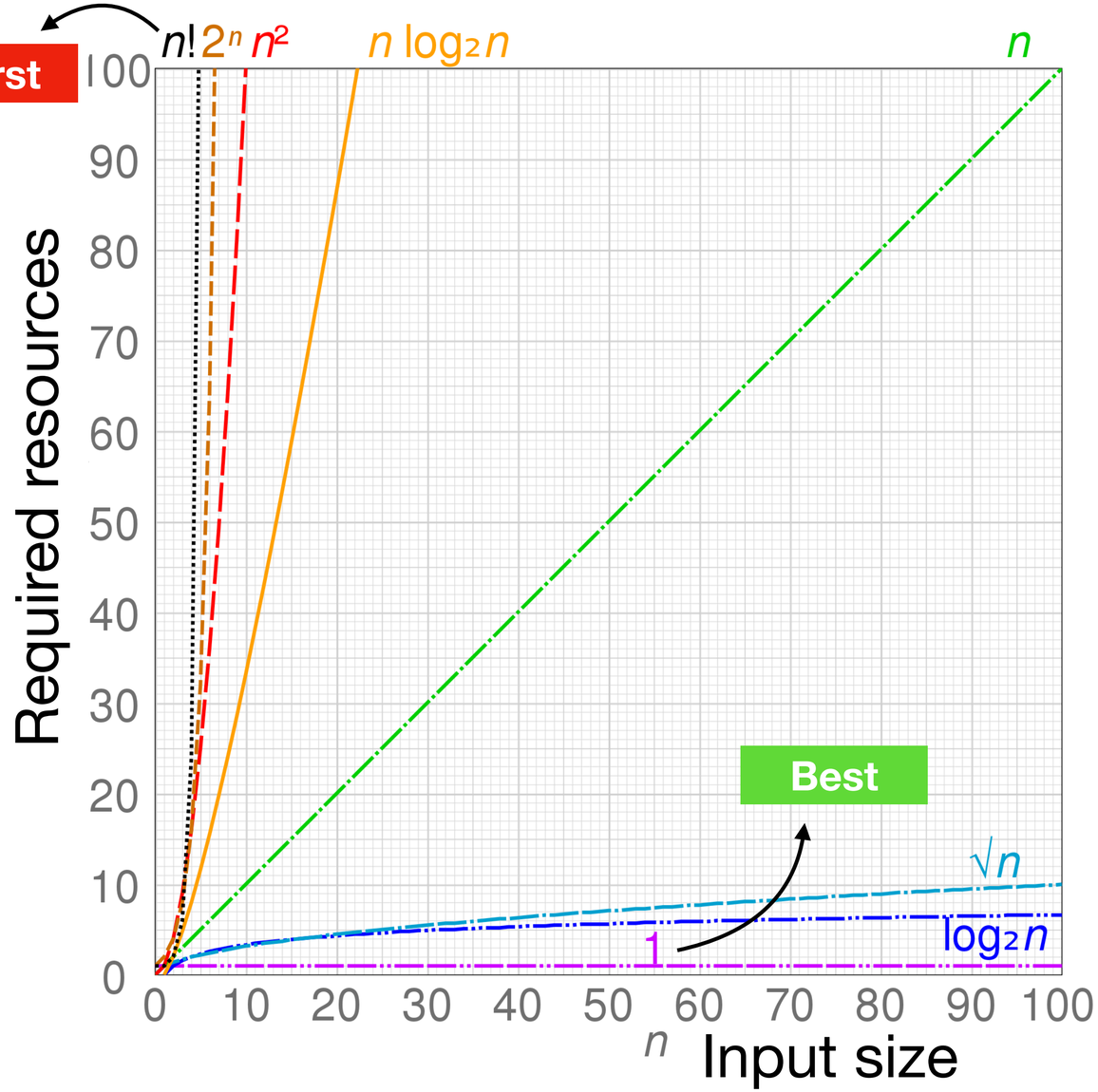
https://en.wikipedia.org/wiki/Time_complexity

III. Computer Engineering and Computer 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)

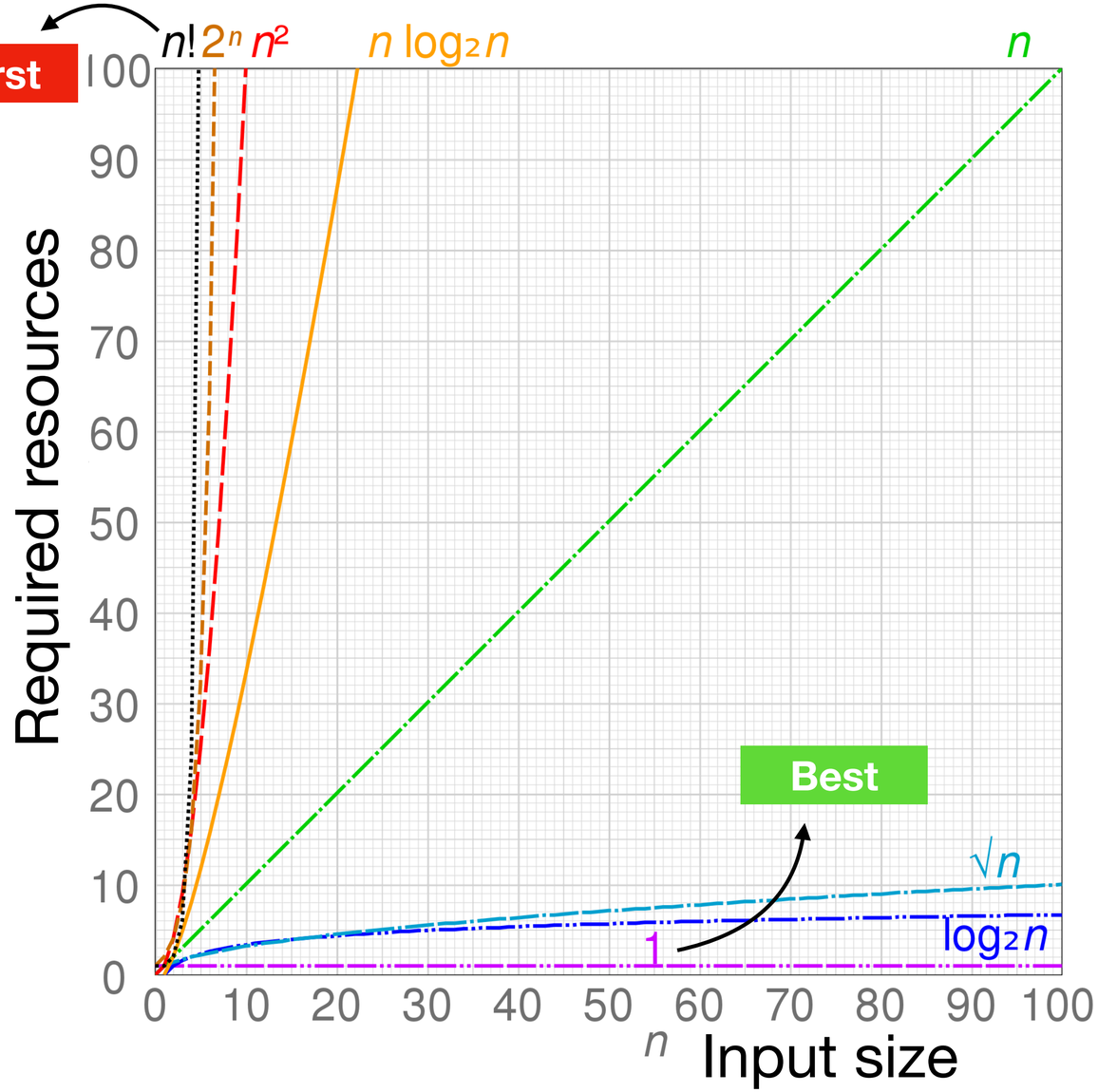


https://en.wikipedia.org/wiki/Time_complexity

III. Computer Engineering and Computer 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

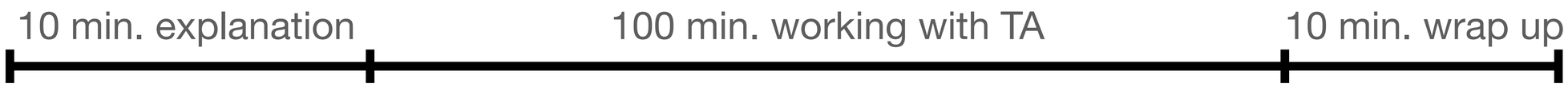
Course Overview

Course Format

- Lecture (75 minutes)

We will discuss one or two questions from the lab assignment

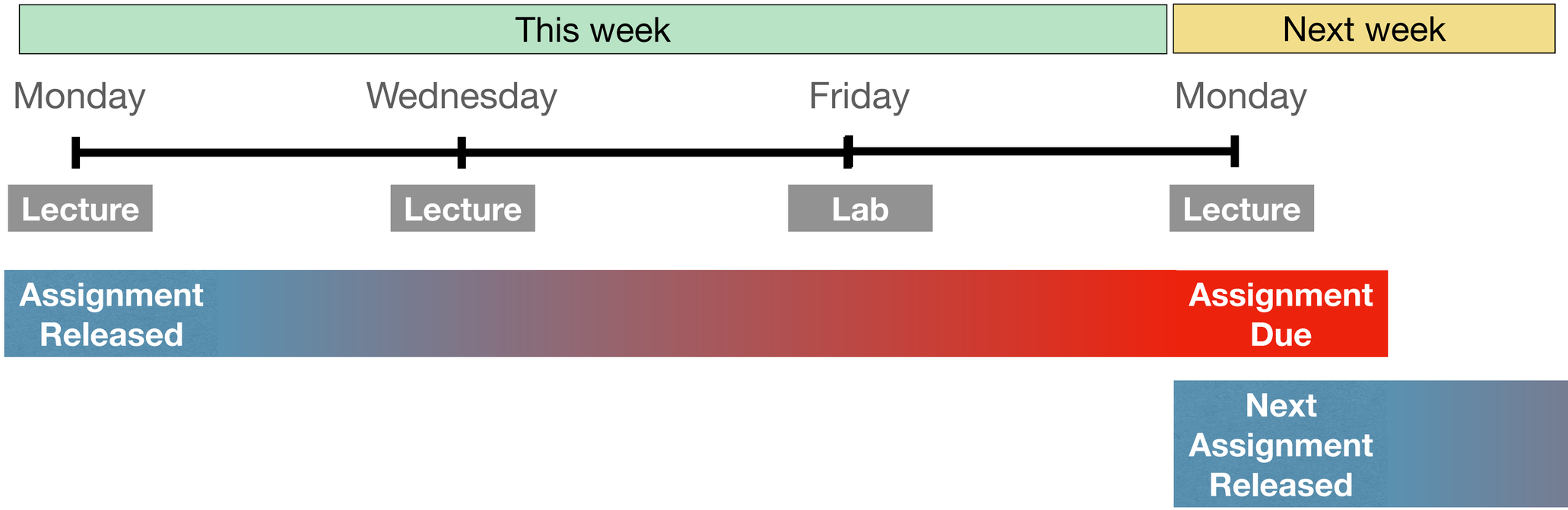
- Lab (120 minutes)



- 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 **next Monday** (23:59)

Course Overview

Weekly Routine



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 (Tentative)

Python basics and data wrangling

Week	Lecture 1	Lecture 2	Lab
1	[No class] Martin Luther King Jr. Day	What is Data Science?	Environment setup
2	Coding environment	Python basics	Python basics I
3	List and Dictionary	Loop	Python basics II
4	File system	String processing	File system
5	Data frames	Data frames	Pandas library
6	[No class] Presidents' Day	Intro to database	Public datasets
7	SQL basics	SQLite 3	SQLite 3
8	[No class] Spring break	[No class] Spring break	[No class] Spring break

Model theories and Product deployment

Week	Lecture 1	Lecture 2	Lab
9	Regression model	Regularization	Scikit-learn
10	Feature selection	Model validation	Feature selection
11	Principal Component Analysis	K-means clustering	PCA and K-means
12	Intro to computer vision	Convolution	OpenCV
13	Data visualization	Web app	Plotly
14	Object-oriented programming	Encapsulation	OOP implementation I
15	Object and Class	Inheritance and polymorphsm	OOP implementation II
16	Project presentation	Project presentation	Lab final exam

Course Overview

Final Project and Exam

Final project (15%)

Each student will give a short literature review (1-3 papers) on the applications of data science that are relevant to your research.

20-minute talk + 15-minute QA

Try to use the terminology and knowledge you learn from the course

Lab exam (25%)

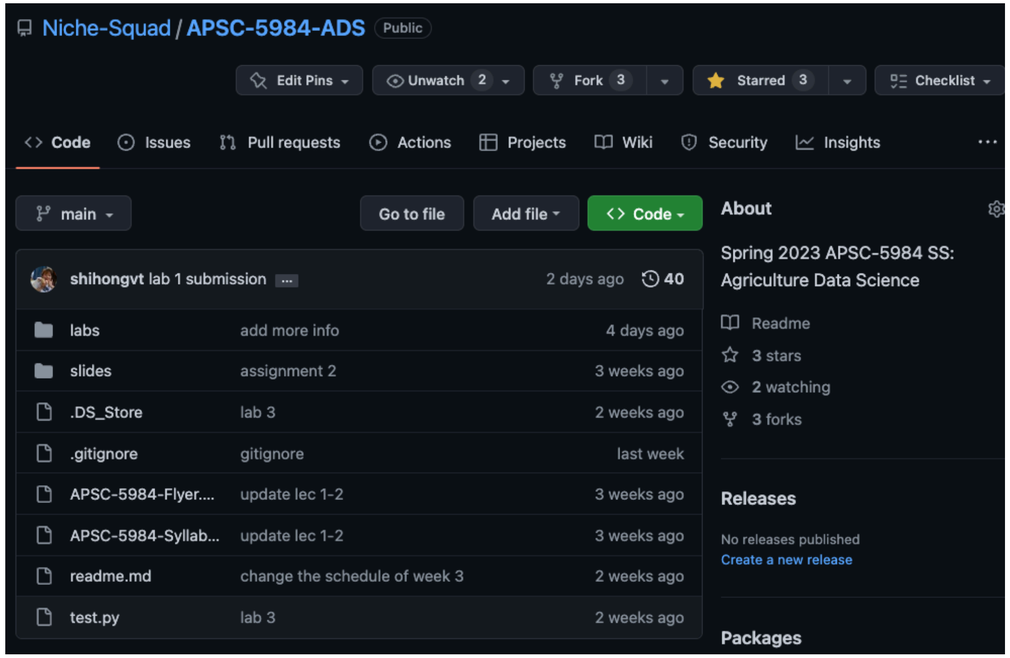
Will be similar questions with our assignments

You will be guided to make a simple Python library

It's Your Turn!

We will use Microsoft Teams for the communication

Create an GitHub account to access the course materials



https://github.com/Niche-Squad/APSC-5984-ADS