INTRO TO MACHINE LEARNING

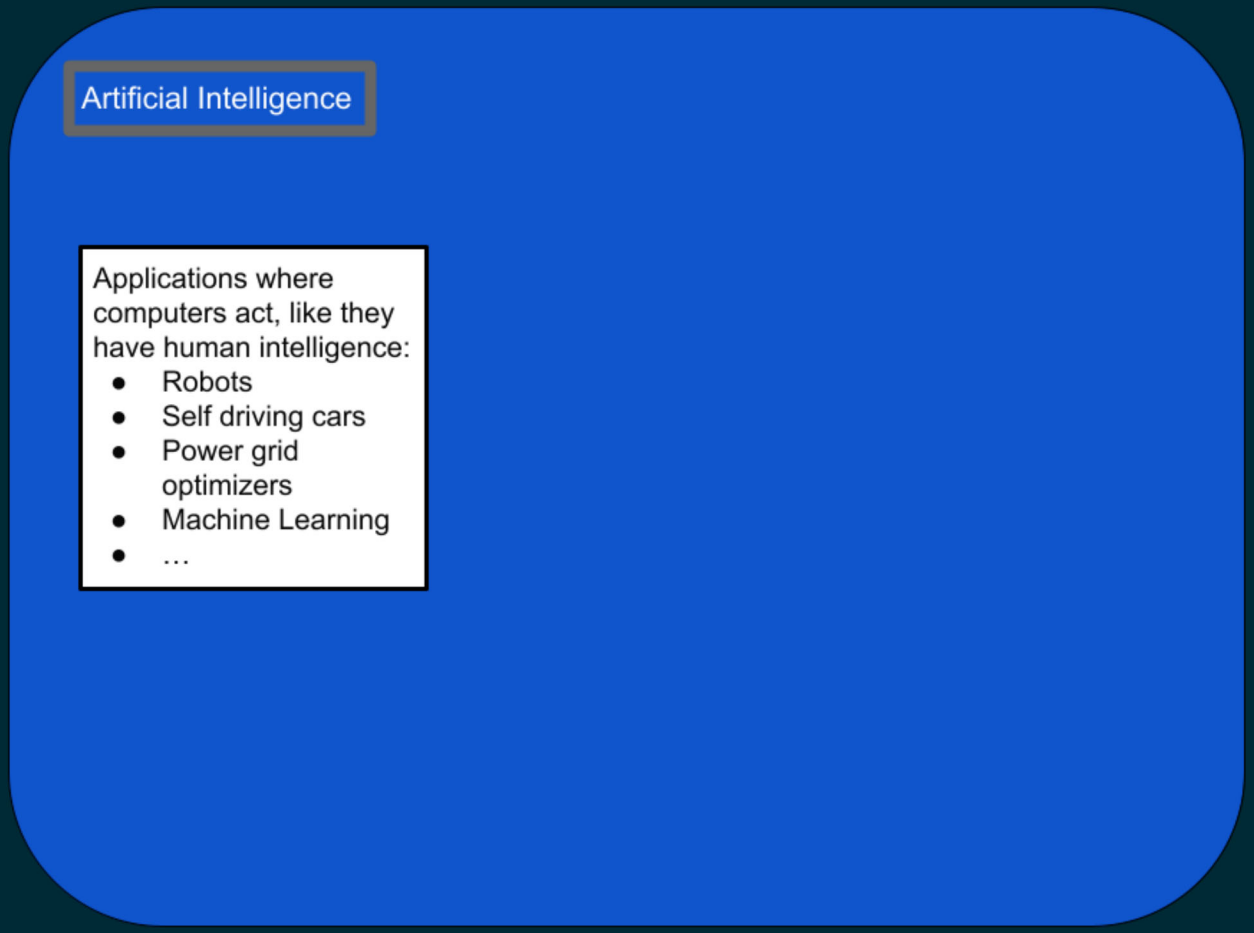
Types, Tasks, Terminology

INTRODUCTION TO MACHINE LEARNING

WHAT IS WHAT?

- Artificial Intelligence (AI),
- Machine Learning,
- Deep Learning
- Big Data

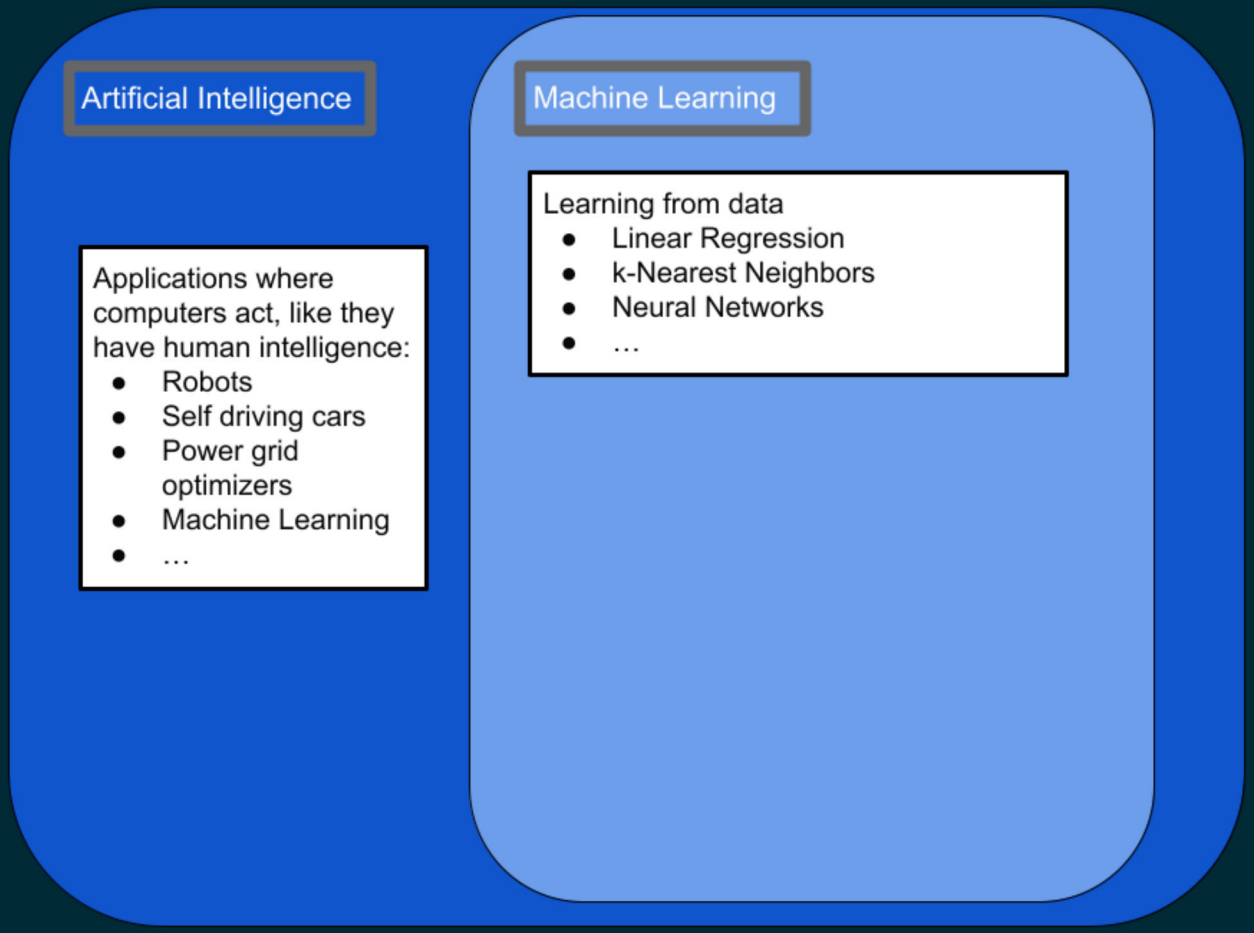
WHAT IS WHAT?



Types of AI

https://econ.lange-analytics.com/aibook/

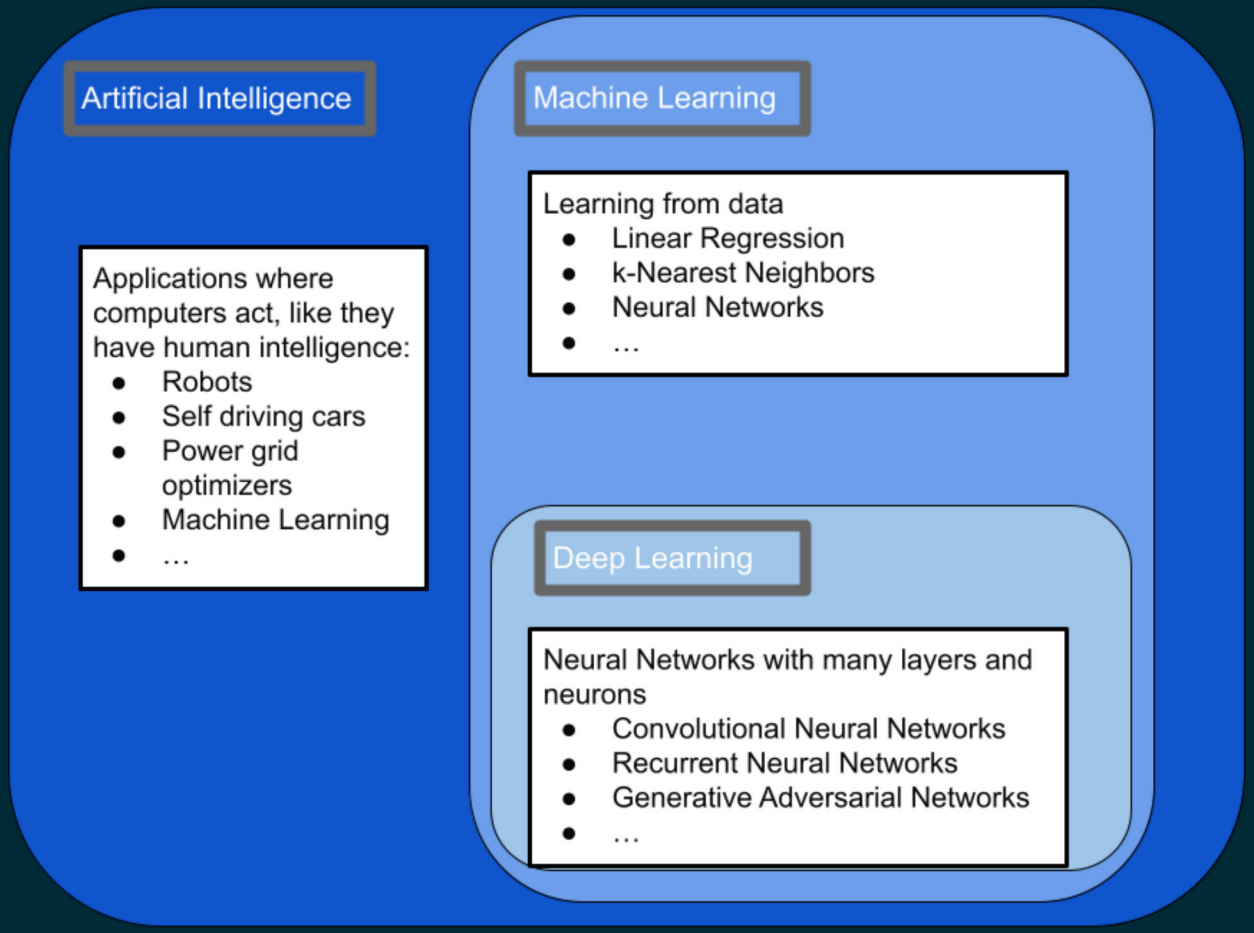
WHAT IS WHAT?



Types of AI

https://econ.lange-analytics.com/aibook/

WHAT IS WHAT?



Types of AI

https://econ.lange-analytics.com/aibook/

WHAT ABOUT BIG DATA

WHAT ABOUT BIG DATA

- Big Data is not a category of learning. It is a category of data!!!
- Two common definitions
  - Laymen: Many records (thousands?, millions?, billions?)
  - Experts: So many records that they do not fit in the memory of one computer.
    - At least billions of records.
    - Requires distributed computing.

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
- Cluster

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
  - Outcome variable is continuous
  - We try to predict a numerical value
- Classification
- Cluster

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
- Cluster

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
  - Outcome variable is categorial
  - Most of the times 2 categories such as:
    - Yes/No
    - Red Wine/White Wine
    - True/False
    - often represented as dummies: 1/0
  - Sometimes more than two catogories (ordered or unordered):
    - good, fair, bad (ordered)
    - red, blue, green (unordered)
    - strongly agree, agree, disagree, strongly disagree (ordered)
- Cluster

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
- Cluster

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
- Cluster
  - Sorting observations into a number of groups based on feature variables.
  - Groups are as homogenous inside as possible.
  - Groups are as diverse between groups (when comparing groups)

THREE APPLICATIONS OF MACHINE LEARNING

- Regression
- Classification
- Cluster

TERMINOLGY

First 3 Observations (records) of the Housing Dataset (to predict house prices)

▶ Code

```
   Price Sqft Bedrooms Waterfront
1 221900 1180        3         no
2 538000 2570        3         no
3 180000  770        2         no
```

Tidy data:

- Observations (synonym: records) are in the rows.
- Variables (synonym: features) are in the columns.
- Variable names (column names) are in the first row.
- Data are in individual cells (and they form vectors; column names can be interpreted as vector names).

TERMINOLGY

Main | Synonyms

First 3 Observations (records) of the Housing Dataset (predict house prices)

▶ Code

```
   Price Sqft Bedrooms Waterfront
1 221900 1180        3         no
2 538000 2570        3         no
3 180000  770        2         no
```

- **Outcome Variable**: The variables that is the outcome of the prediction ($Price$)
- **Predictor Variables**: The variables that **predict** an outcome ($Sqft, Bedrooms, Waterfront$)
- **Example linear regression:** $Price = \beta_1 \cdot Sqft + \beta_2 \cdot Bedrooms + \beta_3 \cdot Waterfront + \beta_4$

WHY USING R FOR MACHINE LEARNING?

https://econ.lange-analytics.com/aibook/