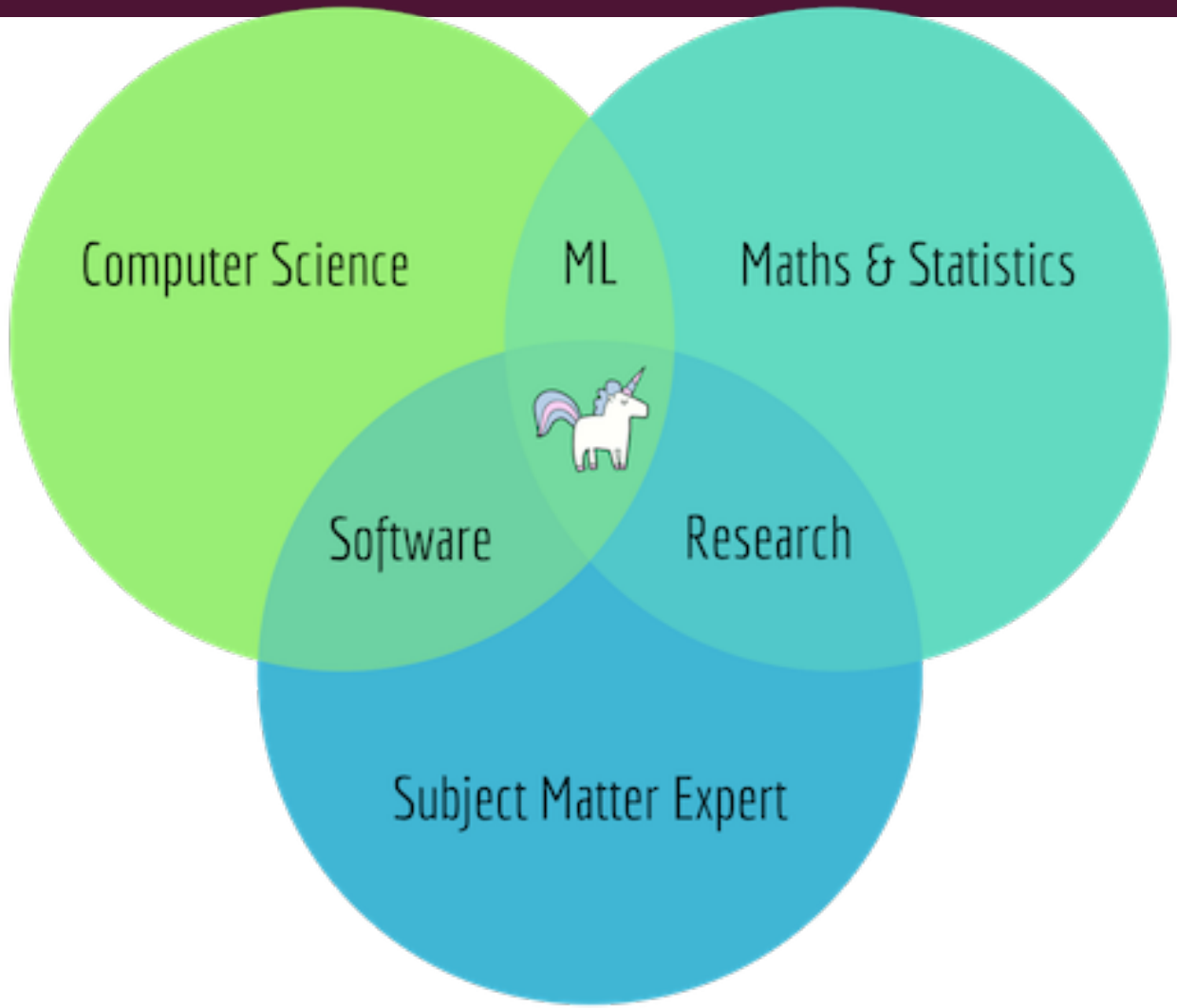
# COMPUTATIONAL BEHAVIOR MODELING

MULTI-DISCIPLINARY RESEARCH

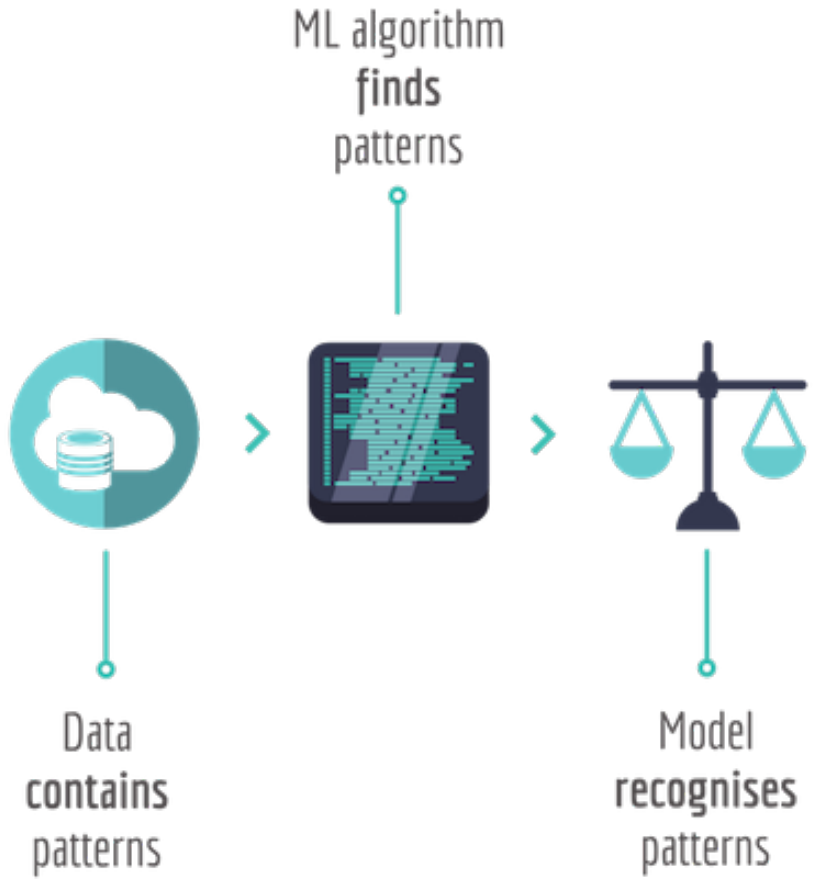
# DATA SCIENTIST



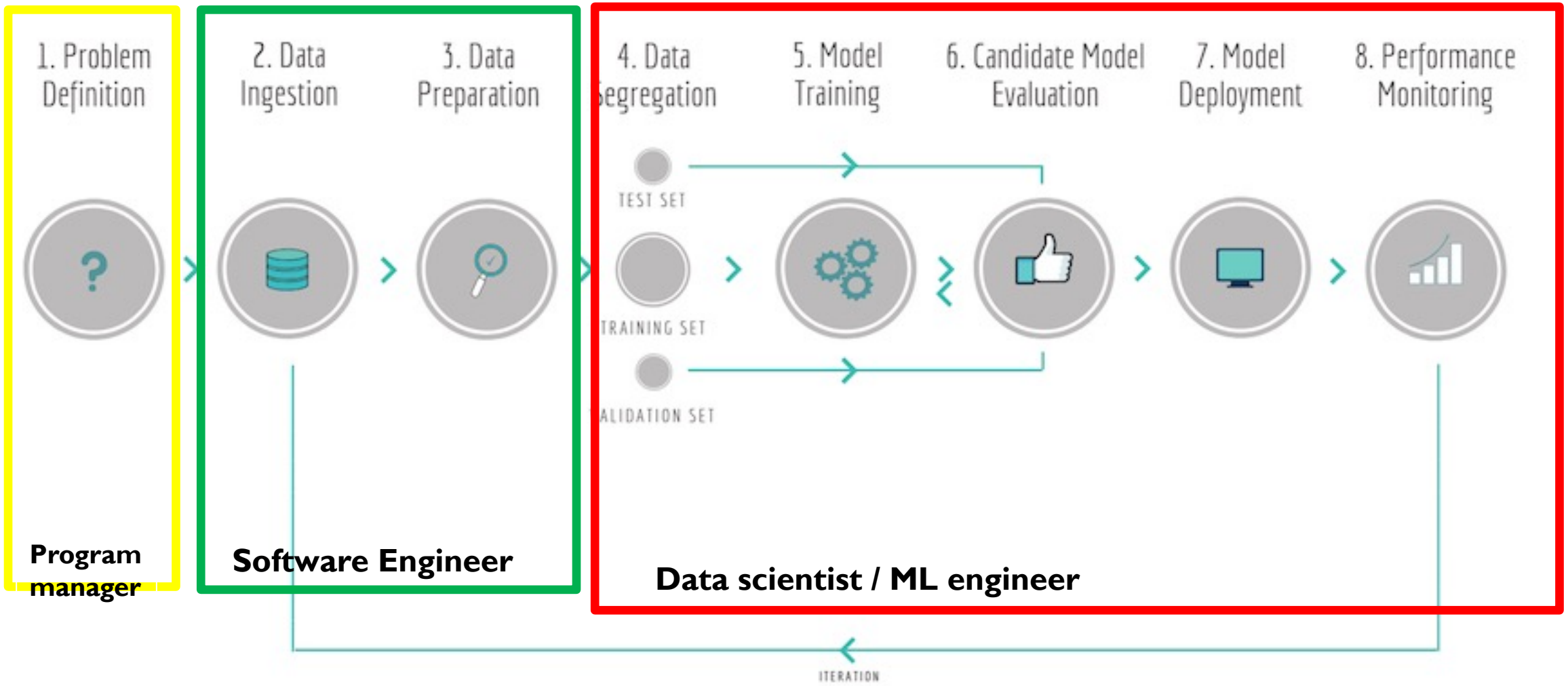
# MACHINE LEARNING

- **Purpose**
- Find patterns in data
- Use the learned patterns to predict for future
- Use the learned patterns to make decisions

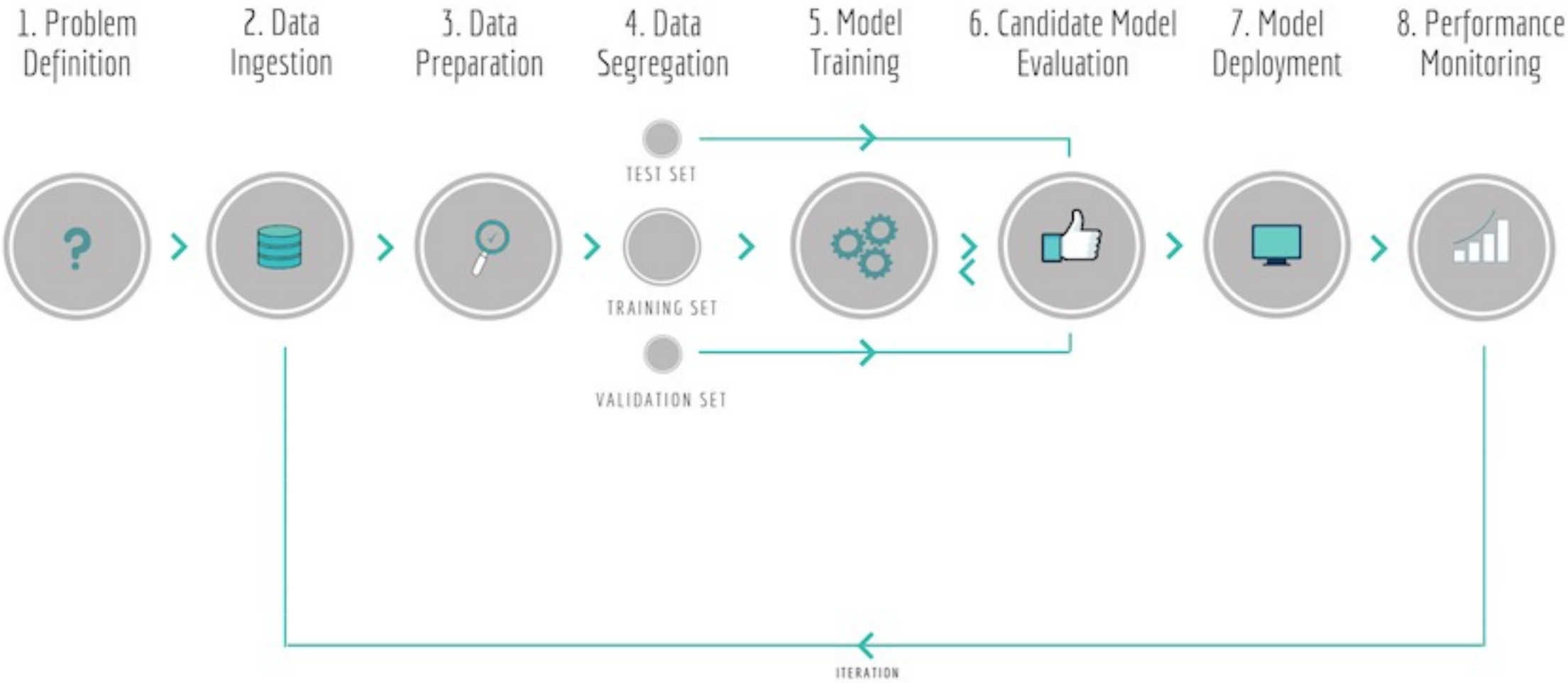
- **Data**
- Data that contains patterns
- ML algorithm finds the patterns and generates a model
- Given new data, the model recognizes these patterns.



# MACHINE LEARNING PIPELINE



# MACHINE LEARNING PIPELINE



# TRAINING MODELS

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

# MODELS – SUPERVISED LEARNING

- A credit card company receives thousands of applications for new cards. Each application contains information about an applicant,
  - age
  - Marital status
  - annual salary
  - outstanding debts
  - credit rating
  - etc.
- Problem: to decide whether an application should be approved, or to classify applications into two categories, approved and not approved.

# MODELS – SUPERVISED LEARNING

| ID | Age | Has_Job | Own_House | Credit_Rating | Class |
|---|---|---|---|---|---|
| 1 | young | false | false | fair | No |
| 2 | young | false | false | good | No |
| 3 | young | true | false | good | Yes |
| 4 | young | true | true | fair | Yes |
| 5 | young | false | false | fair | No |
| 6 | middle | false | false | fair | No |
| 7 | middle | false | false | good | No |
| 8 | middle | true | true | good | Yes |
| 9 | middle | false | true | excellent | Yes |
| 10 | middle | false | true | excellent | Yes |
| 11 | old | false | true | excellent | Yes |
| 12 | old | false | true | good | Yes |
| 13 | old | true | false | good | Yes |
| 14 | old | true | false | excellent | Yes |
| 15 | old | false | false | fair | No |

labels

# MODELS – SUPERVISED LEARNING

- Like human learning from past experiences or historical data.
- A computer does not have "experiences".
- A computer system learns from data, which represent some "past experiences" of an application domain.
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.
- The task is commonly called: Supervised learning, classification, or inductive learning.

# MODELS – SUPERVISED LEARNING

- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following case/instance?

| Age | Has_Job | Own_house | Credit-Rating | Class |
|---|---|---|---|---|
| young | false | false | good | ? |

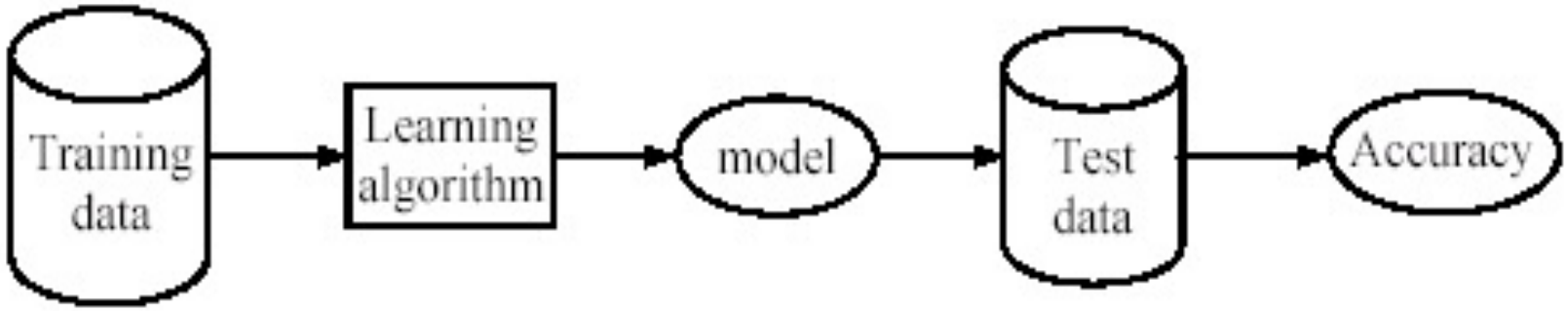
# MODELS – SUPERVISED LEARNING

- Supervised learning: classification is seen as supervised learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a "teacher" gives the classes (supervision).
  - Test data are classified into these classes too.
- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

# MODELS – SUPERVISED LEARNING

- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



# MODELS – SUPERVISED LEARNING

- Data: credit card application data
- Task: Predict whether a credit card application should be approved or not.
- Performance measure: accuracy.

No learning: classify all future applications (test data) to the majority class (i.e., Yes):

Accuracy = 9/15 = 60%.

- We can do better than 60% with learning.

# MODELS – SUPERVISED LEARNING

Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples).

- In practice, this assumption is often violated to certain degree.
- Strong violations will clearly result in poor classification accuracy.
- To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data.

# MODELS – SUPERVISED LEARNING ALGORITHMS

- Decision tree induction/classification
- Random Forest (the average of the results from 100 random trees)
- Naïve Bayesian classification (0 or 1 cases – binary cases)
- Naïve Bayes for text classification
- Support vector machines (binary cases)
- K-nearest neighbor
- Ensemble methods: Bagging and Boosting

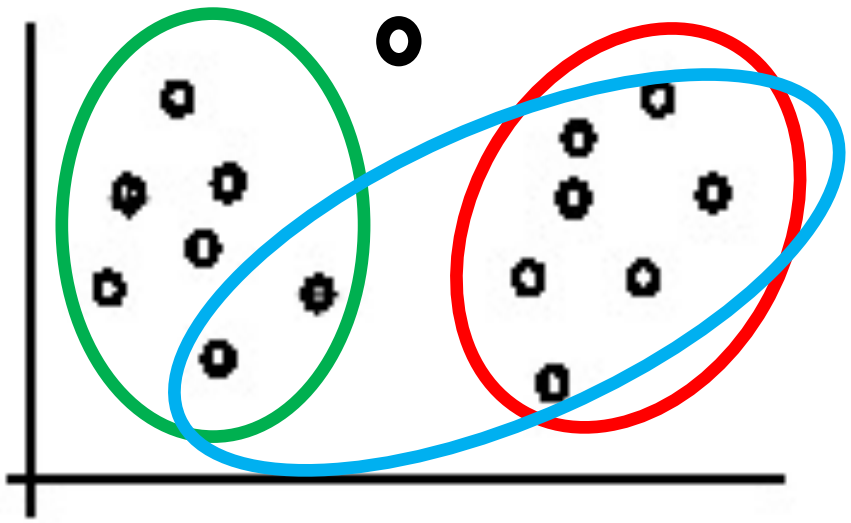
# MODELS

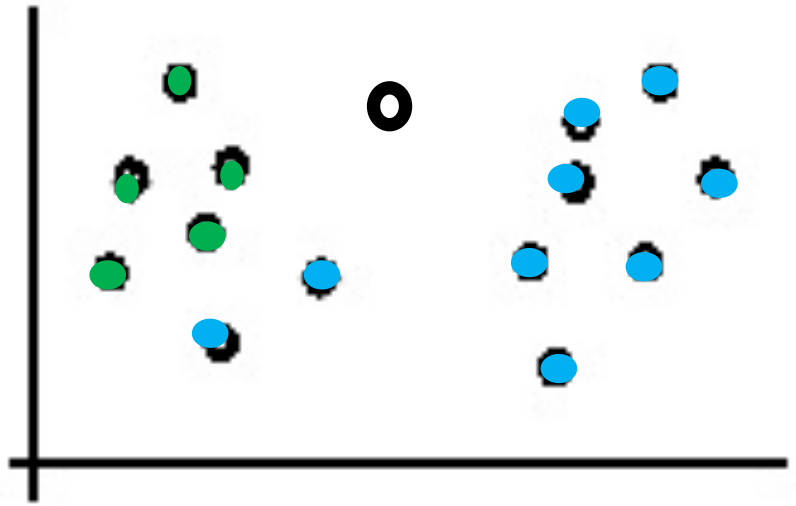
- Supervised learning: discover patterns in the data with a target (class).
  - to predict the target attribute in future data.
- Unsupervised learning: without target attribute.
  - learn intrinsic structures in data.

# MODELS

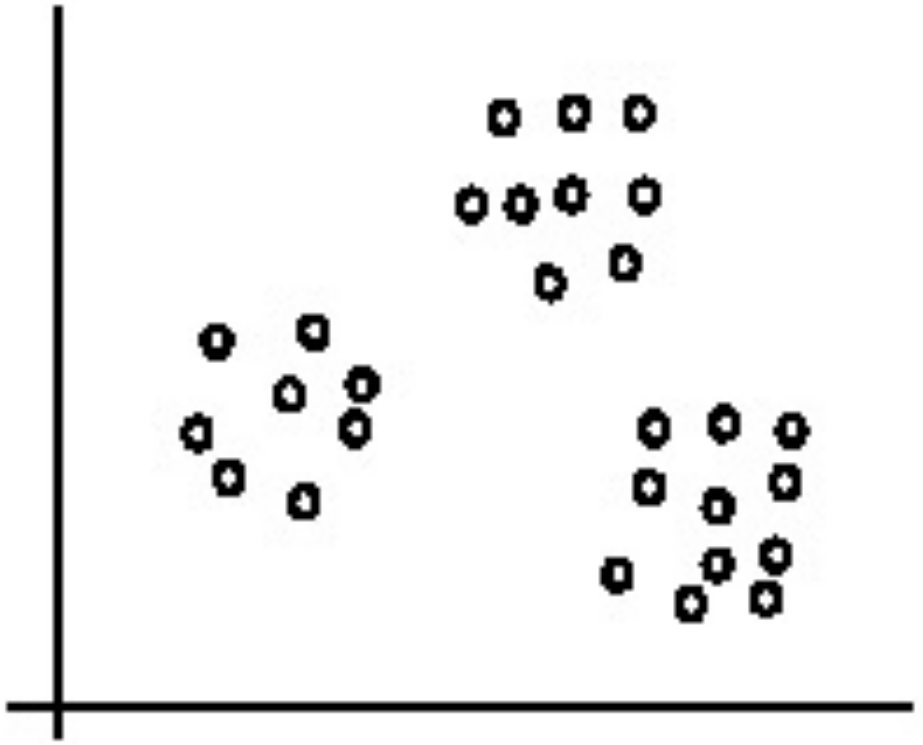
| ID | Age | Has_Job | Own_House | Credit_Rating |
|---|---|---|---|---|
| 1 | young | false | false | fair |
| 2 | young | false | false | good |
| 3 | young | true | false | good |
| 4 | young | true | true | fair |
| 5 | young | false | false | fair |
| 6 | middle | false | false | fair |
| 7 | middle | false | false | good |
| 8 | middle | true | true | good |
| 9 | middle | false | true | excellent |
| 10 | middle | false | true | excellent |
| 11 | old | false | true | excellent |
| 12 | old | false | true | good |
| 13 | old | true | false | good |
| 14 | old | true | false | excellent |
| 15 | old | false | false | fair |

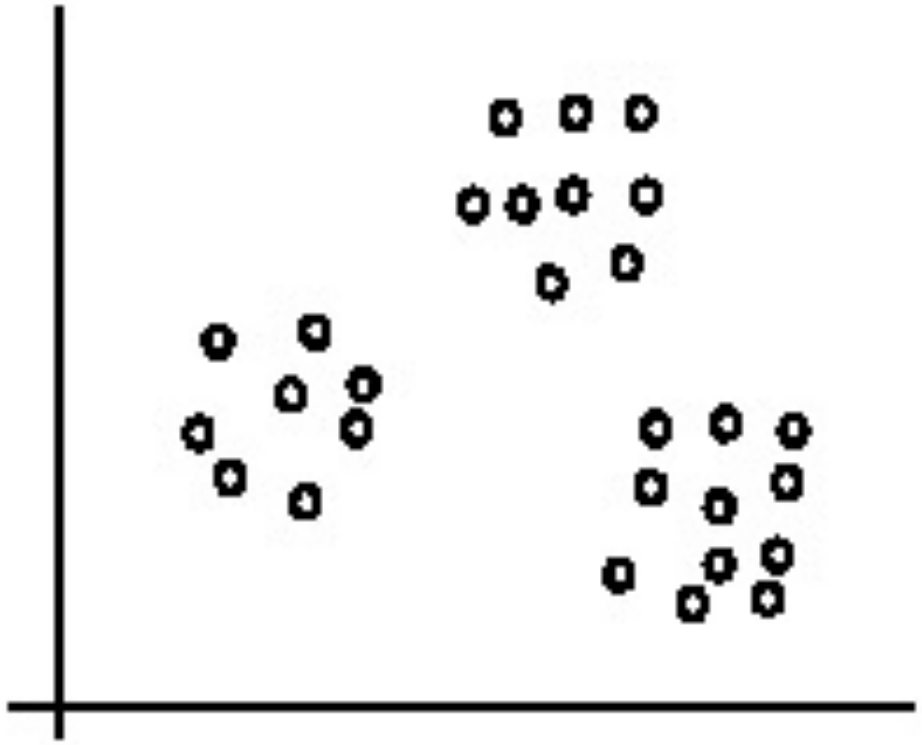
# MODELS – UNSUPERVISED LEARNING (CLUSTERING)





# MODELS – UNSUPERVISED LEARNING





# MODELS – UNSUPERVISED LEARNING MODELS

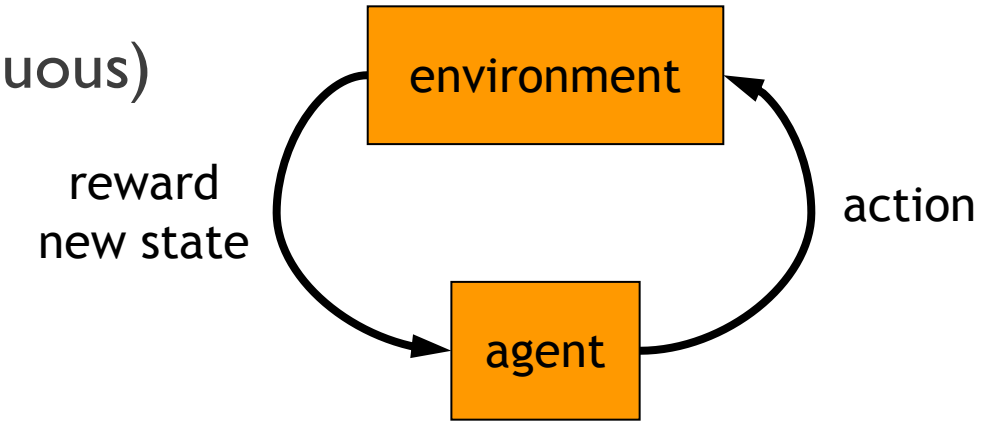
- **K-means algorithm**
- **Representation of clusters**
- **Hierarchical clustering**

# MODELS

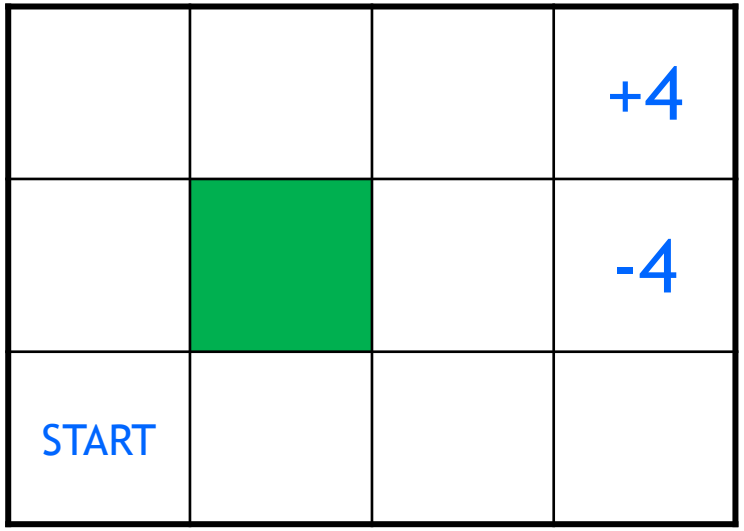
- Supervised learning
  - Classification (discrete), regression(continuous)
- Unsupervised learning
  - clustering
- Reinforcement learning
  - more general than supervised/unsupervised learning
  - learn from interaction w/ environment to achieve a goal

# MODELS –REINFORCEMENT LEARNING

- Supervised learning
  - Classification (discrete), regression(continuous)
- Unsupervised learning
  - clustering
- Reinforcement learning
  - more general than supervised/unsupervised learning
  - learn from interaction w/ environment to achieve a goal

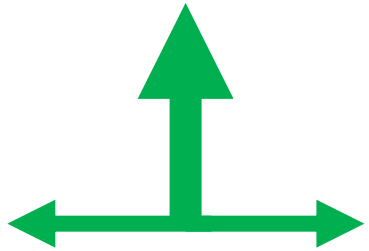


# MODELS –REINFORCEMENT LEARNING




actions: UP, DOWN, LEFT, RIGHT

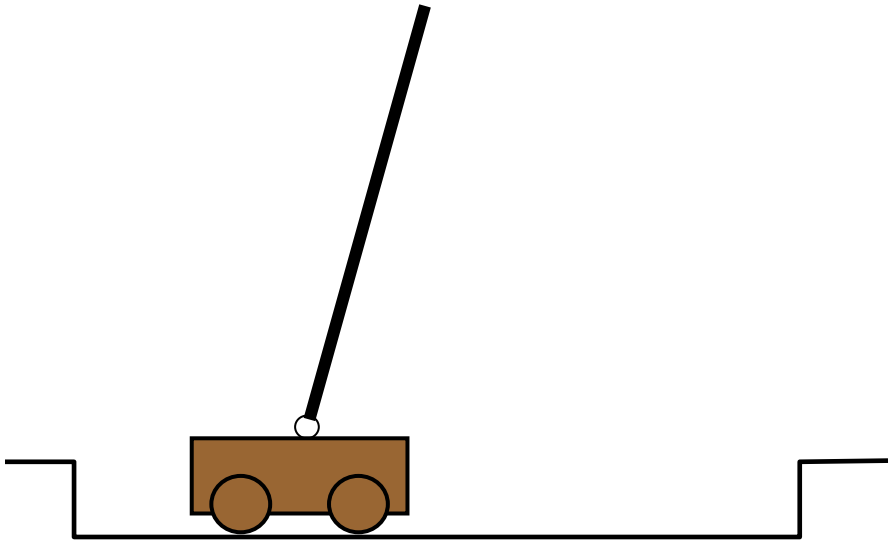
60% move UP
15% move Down
15% move LEFT
10% move RIGHT



- reward +1 at [4,3], -1 at [4,2]
- reward -0.01 for each step
- what's the strategy to achieve max reward?

# MODELS –REINFORCEMENT LEARNING

- pole-balancing: move car left/right
- no teacher who would say "good" or "bad"
  - is reward "10" good or bad?
  - rewards could be delayed
- more general, fewer constraints
- explore the environment and learn from experience



# MACHINE LEARNING ALGORITHMS FOR TOPICS

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

- car or bicycle driving patterns (traffic management)
  - Resource re-allocation.
- Human mobility patterns based on GPS data
  - Crowd flow
- Crime Analysis for Chicago
  - Algorithm: Explainable AI (bonus: design a website / app system)
- Weather prediction and its impact to human behavior
  - Maximize the energy efficiency to help the weather.
- Mining the spread patterns of COVID-19
  - GPS/ sensor;  combine spatio-temporal data into one model: IRL
- Robotic deep inverse reinforcement learning
  - Smart and connected community

car or bicycle driving patterns (traffic management)

Resource re-allocation.

Outline

Dataset from Data.world

We want to design features that could be used to train the ML models to better allocate resources (e.g., shared bikes).

Innovative parts: design features (date cleaning + data manipulation + data analysis ⇔ part of data science) + incorporate with the current ML algorithms.

- Features:
  - over activity level in each place (defined what is a place – cells with longtitude / latitude)
  - Duration for a bike that has been parking in a spot.
  - Number of the bikes in that spot.
  - Other features