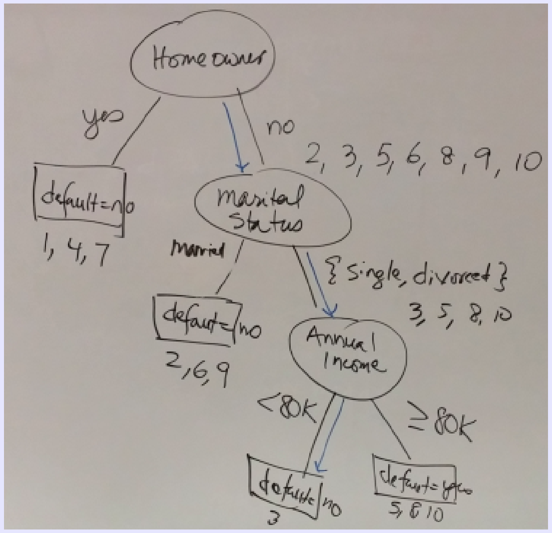
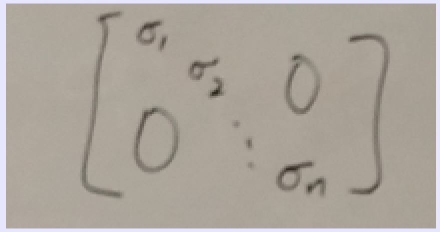
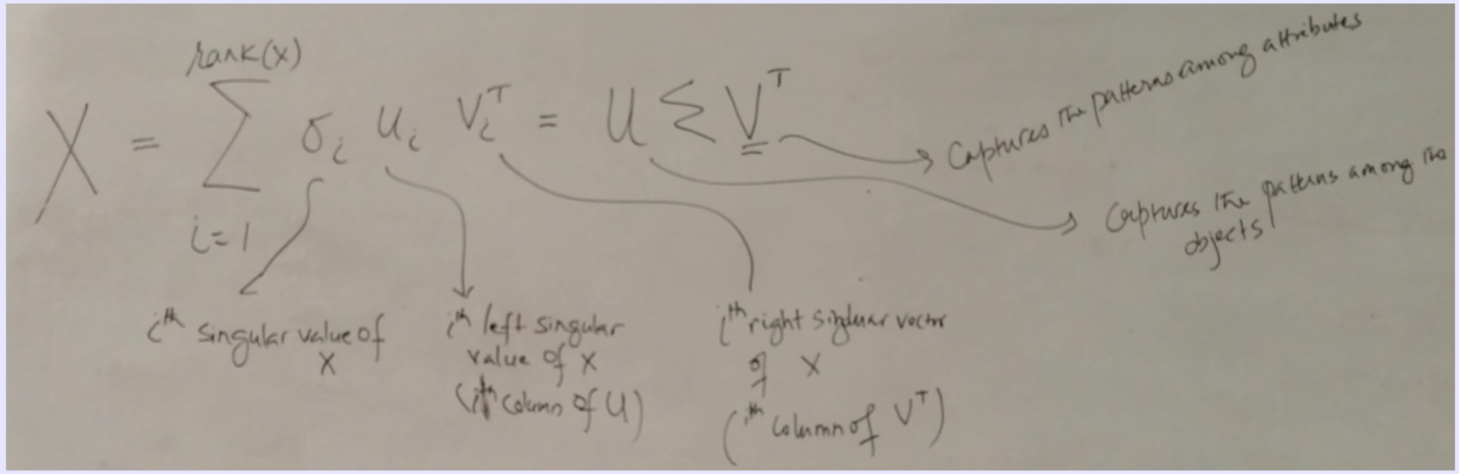
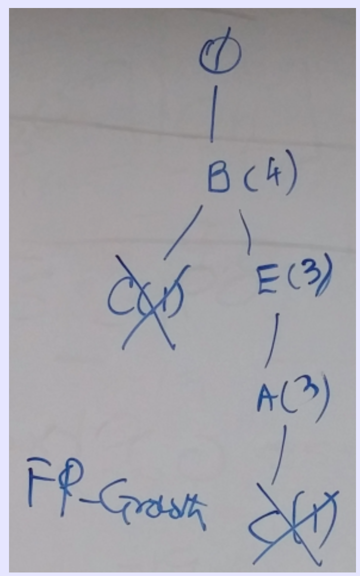


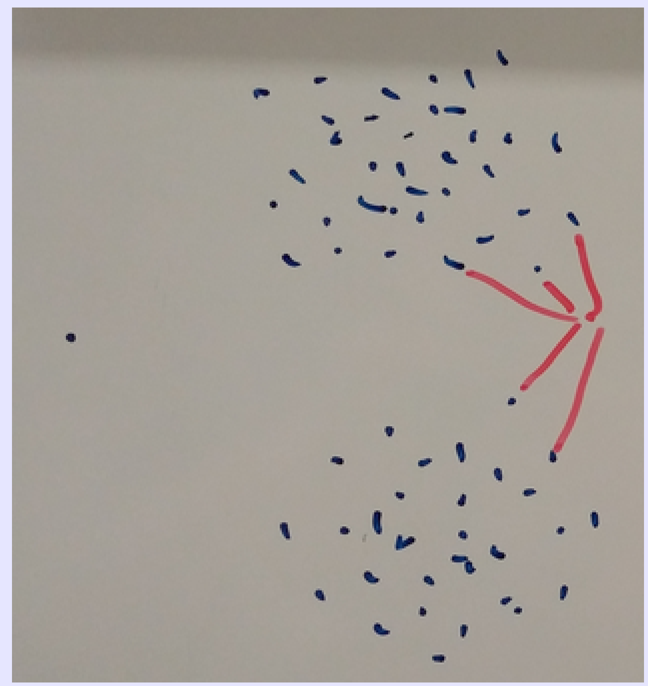




CS 422: Data Mining
Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

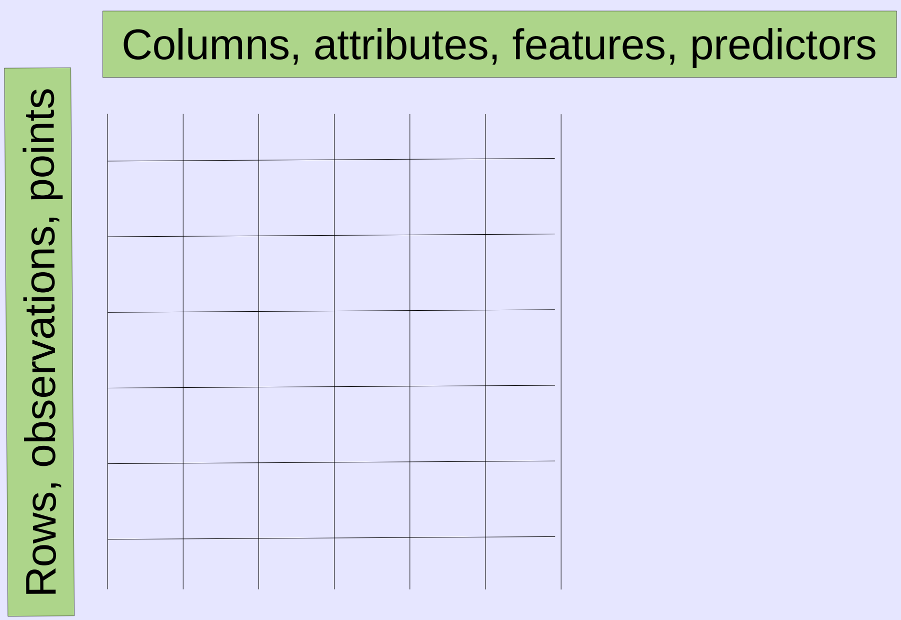
Lecture 4: **Components of Learning, Decision Trees**





Components of learning

- Recall, most data mining / machine learning algorithms operate on matrices.
- The canonical picture to keep in mind is this:



Components of learning

- Example of a *matrix* data layout.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Components of learning

- Example of a *document* data layout.

	token1	token2	token3	token4	token5	token6	token7	token8	token9	token10
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

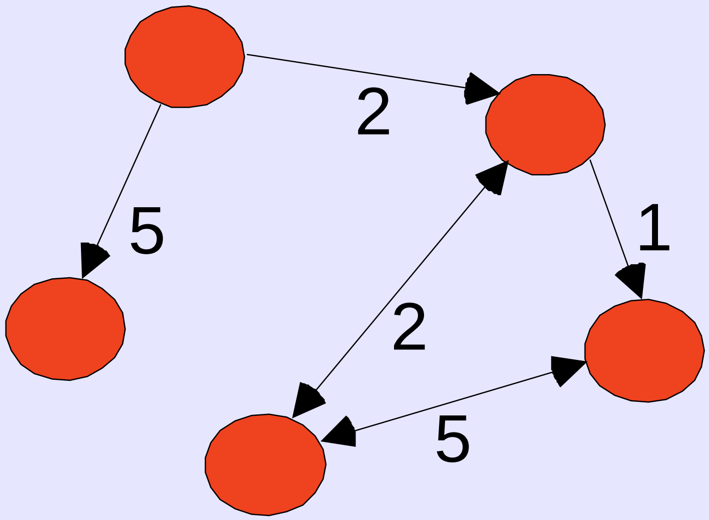
Components of learning

- Example of a *transaction* data layout.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, M**

Components of learning

- Example of a *graph* data layout.

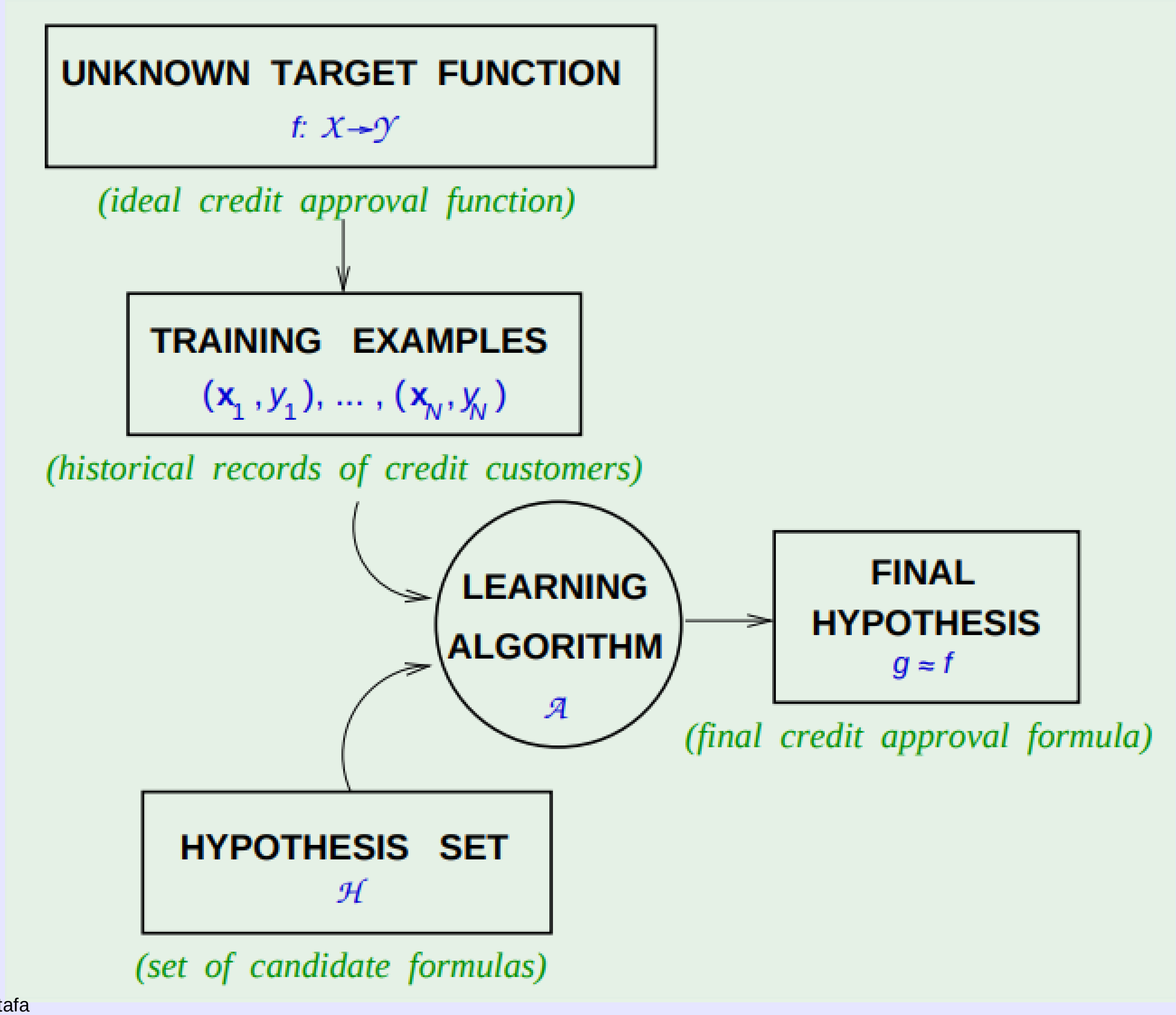


- As it turns out, graphs can be represented as matrices.

Components of learning

- Formalism:
  - Input: $\mathcal{X}$, A matrix (n-dimension, n >= 1) of attributes
  - Output: $\overrightarrow{\mathcal{Y}}$, the response vector
  - Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$
  - Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$
  - Hope: $g \approx f$

Components of learning

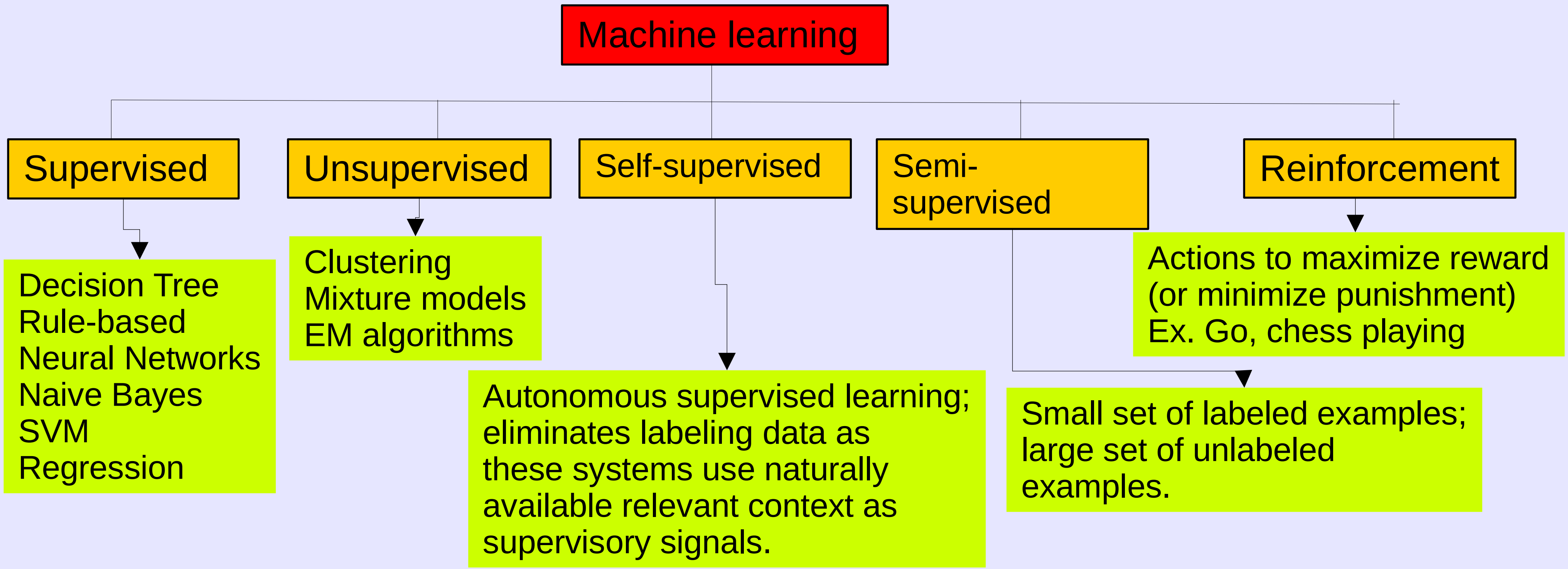


Slide source: Prof. Yaser S. Abu-Mostafa
Learning from Data, 2012.

Components of learning

- Terminology: learner, classifier, model … which is which?
  - ***Learner*** takes as input x_1, x_2, …, x_n, y_i and produces a ***classifier***.
  - A ***classifier*** takes as input x'_1, x'_2, …, x'_n, and produces y’.
  - **Model** is an artifact; a *learner* builds a model and a *classifier* uses that model to predict.

Components of learning

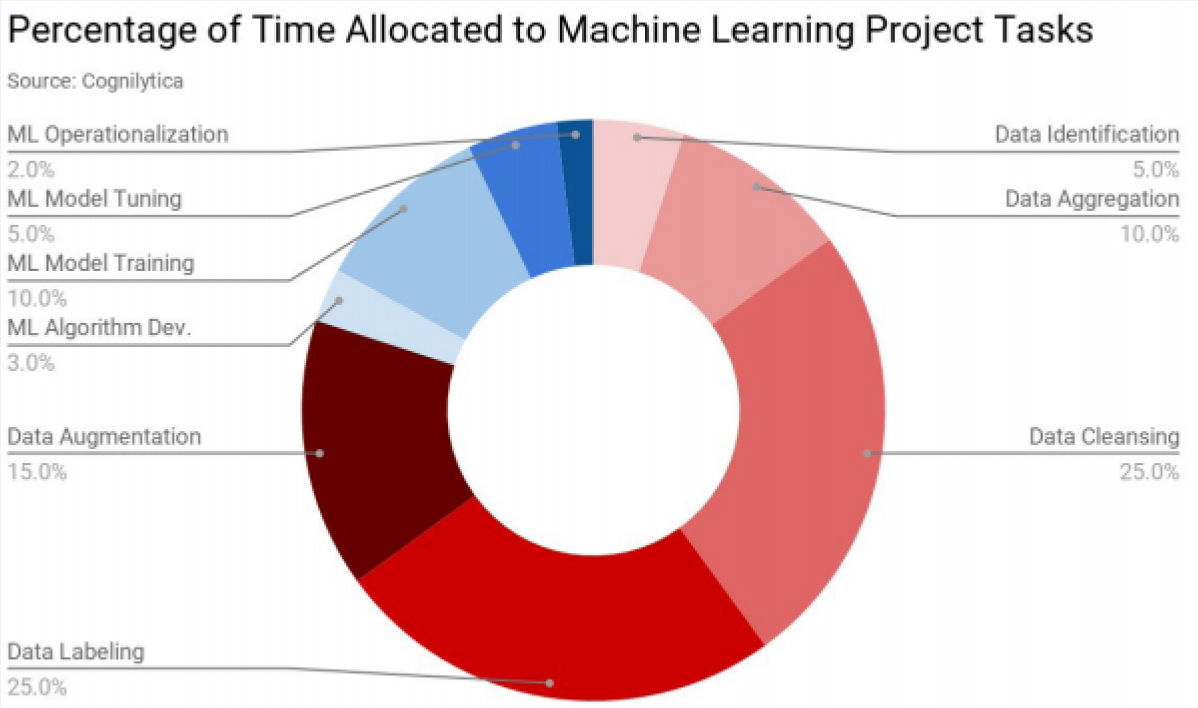


Components of learning

- *Generalizing to cases we have not seen before!* (Curse of dimensionality, see first lecture.)
- A data scientist’s time allocation.

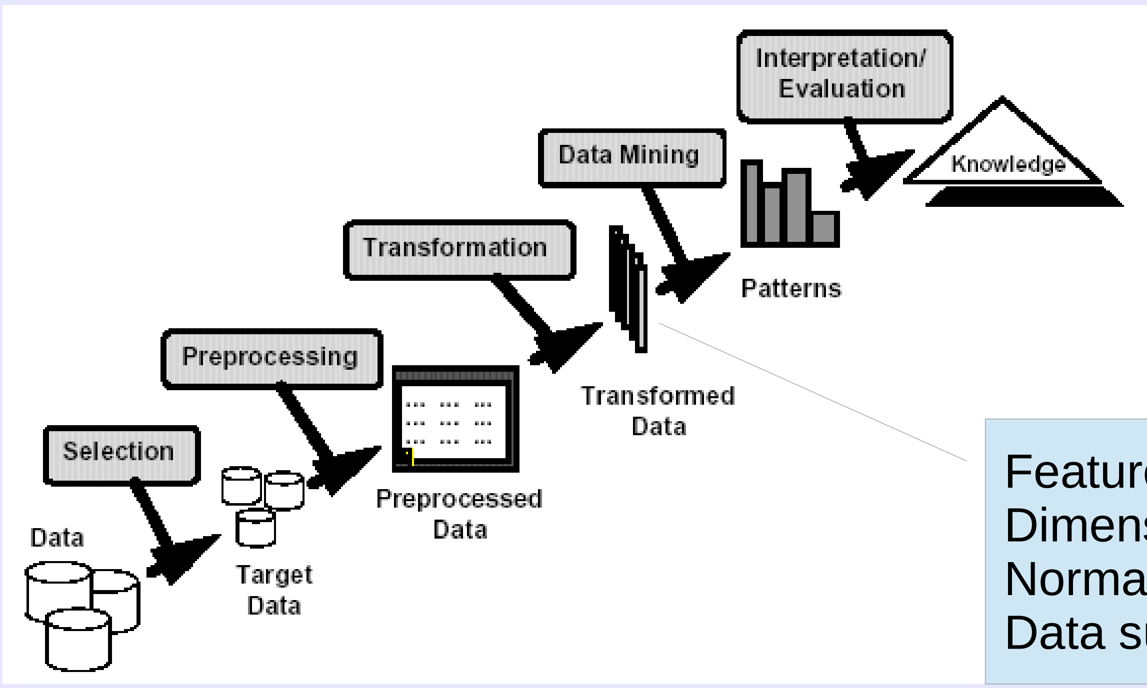
Components of learning

- *Generalizing to cases we have not seen before!* (Curse of dimensionality, see first lecture.)
- A data scientist's time allocation.



Components of learning

- The workflow.



Feature selection
Dimensionality reduction
Normalization
Data subsetting

Data Types

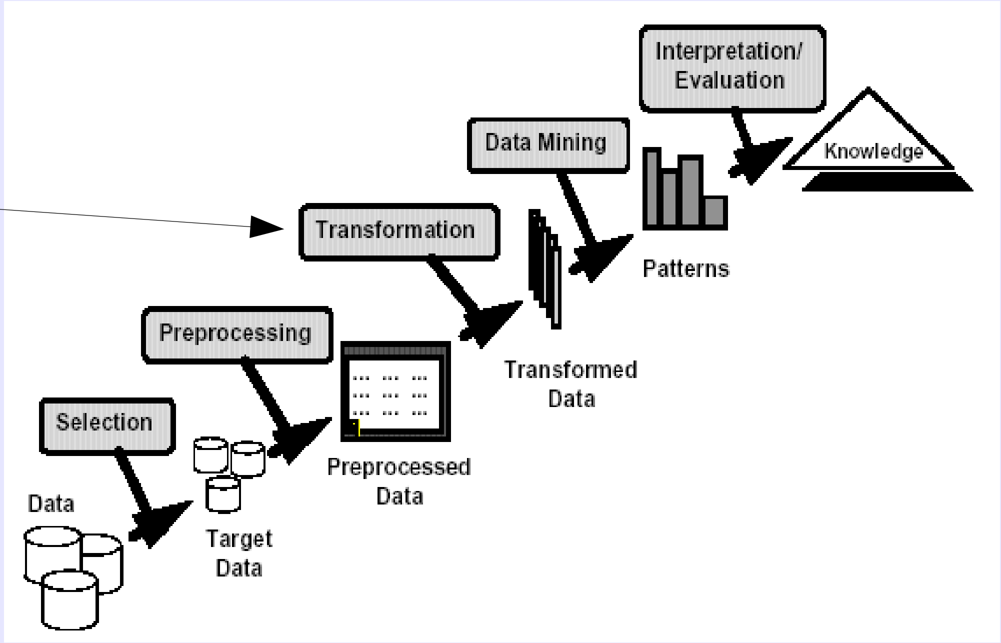
- R has the following data types to represent attributes:
  - Numeric
  - Integer
  - Factor
  - Character

Data Types

- R has the following data types to represent attributes:
  - Numeric: Can take “float” or “double” values.
  - Integer: Cannot take decimal or fraction values.
  - Factor: An enumeration data type that takes only certain values: {“blue”, “green”, “red”}; or {0, 1, 2}.
    - Values of a factor can be
      - *ordinal*, i.e., order of values matter. Example: {“small”, “medium”, “large”} is different than {“small”, “large”, “medium”}.
      - *nominal*, i.e., order of values does not matter. Example: {“blue”, “green”, “red”}.
    - Factors are also referred to as *categorical* variables.
  - Character: Single character or character strings.

Data Types

- Certain algorithms have an affinity for certain data types:
  - Certain classification requires that numeric (or continuous) data be represented as categorical (factor) attributes.
  - Association algorithms prefer a binary attribute (a factor of 0 and 1).
- One of the important step during the transformation phase is to ensure that algorithms get the attribute in the form they can operate on it. (More on this in later lecture.)



Decision tree

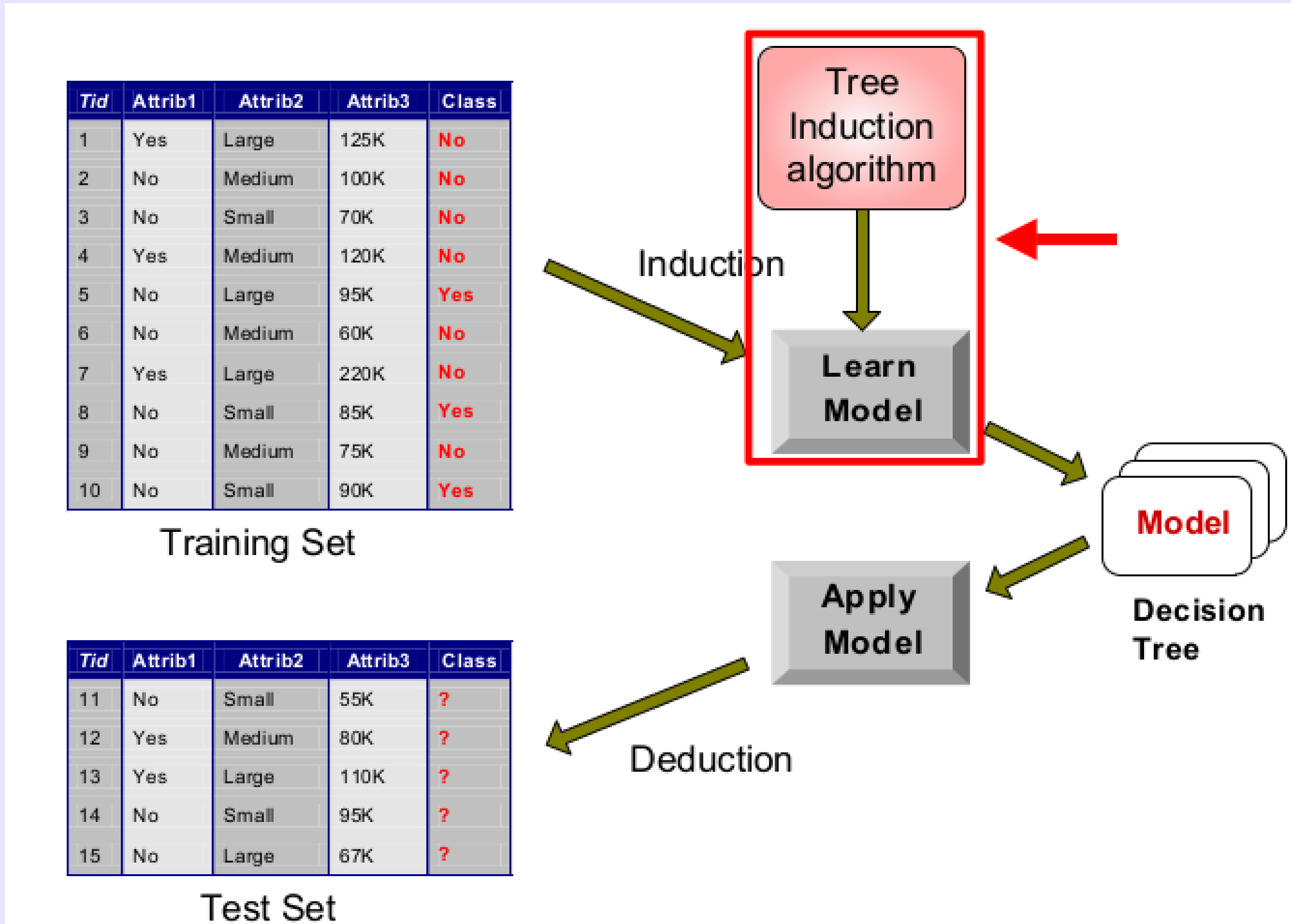
- Our first classification algorithm.
- **Classification**: The task of learning a target function, *g*, that maps each attribute set $\mathcal{X}$ to one of the predefined class labels, $\overrightarrow{\mathcal{Y}}$, or

$$f : x \to y \text{ where } x \in \mathbb{R}^n \text{, and } y \in \mathbb{R}$$

- Let's play a game.
  - Problem: A bank wants to determine who they should make loans to.
  - You are the loan officer.
  - What will **you** look for in potential loan applicants?

Decision tree

- A bird’s eye view.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

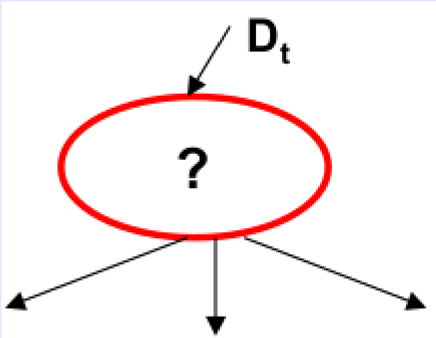
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Decision tree: Hunt's algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Decision tree: Hunt's algorithm

Default class =

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Married	100	No
No	Single	70	No
Yes	Married	120	No
No	Divorced	95	Yes
No	Married	60	No
Yes	Divorced	220	No
No	Single	85	Yes
No	Married	75	No
No	Single	90	Yes

Decision tree: Hunt's algorithm

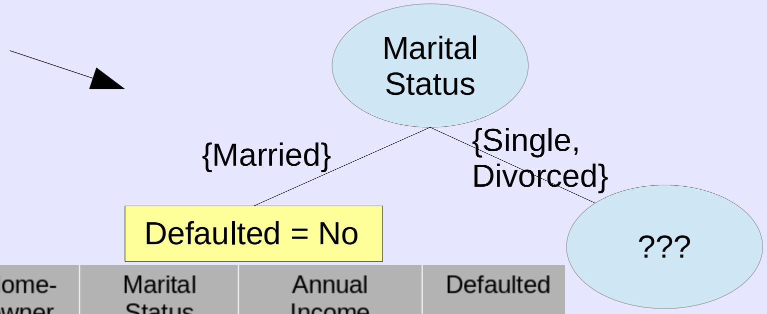
Default class = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Married	100	No
No	Single	70	No
Yes	Married	120	No
No	Divorced	95	Yes
No	Married	60	No
Yes	Divorced	220	No
No	Single	85	Yes
No	Married	75	No
No	Single	90	Yes

Decision tree: Hunt's algorithm

Default class = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Married	100	No
No	Single	70	No
Yes	Married	120	No
No	Divorced	95	Yes
No	Married	60	No
Yes	Divorced	220	No
No	Single	85	Yes
No	Married	75	No
No	Single	90	Yes



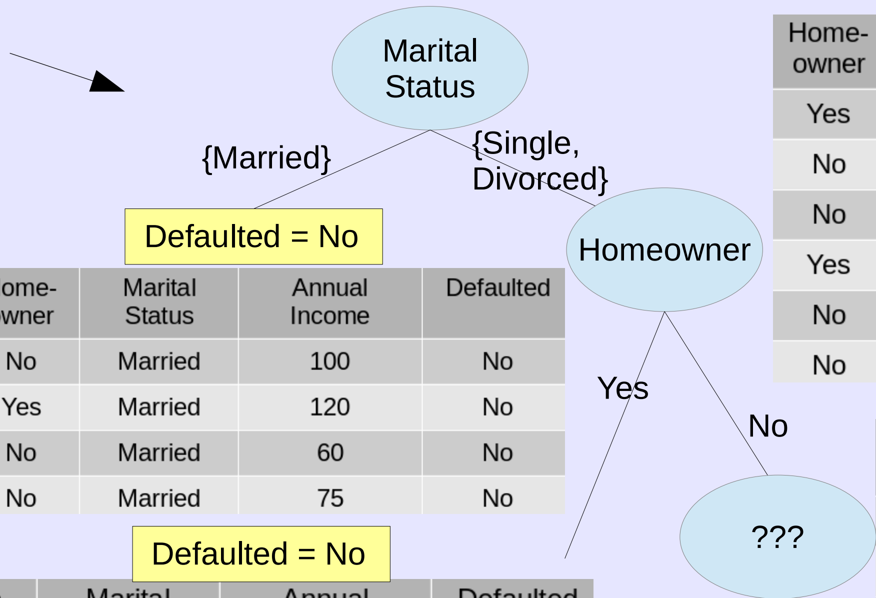
Home-owner	Marital Status	Annual Income	Defaulted
No	Married	100	No
Yes	Married	120	No
No	Married	60	No
No	Married	75	No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Single	70	No
No	Divorced	95	Yes
Yes	Divorced	220	No
No	Single	85	Yes
No	Single	90	Yes

Decision tree: Hunt's algorithm

Default class = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Married	100	No
No	Single	70	No
Yes	Married	120	No
No	Divorced	95	Yes
No	Married	60	No
Yes	Divorced	220	No
No	Single	85	Yes
No	Married	75	No
No	Single	90	Yes



Defaulted = No

Home-owner	Marital Status	Annual Income	Defaulted
No	Married	100	No
Yes	Married	120	No
No	Married	60	No
No	Married	75	No

Defaulted = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
Yes	Divorced	220	No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Single	70	No
No	Divorced	95	Yes
Yes	Divorced	220	No
No	Single	85	Yes
No	Single	90	Yes

Home-owner	Marital Status	Annual Income	Defaulted
No	Single	70	No
No	Divorced	95	Yes
No	Single	85	Yes
No	Single	90	Yes

Decision tree: Hunt’s algorithm

Default class = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Married	100	No
No	Single	70	No
Yes	Married	120	No
No	Divorced	95	Yes
No	Married	60	No
Yes	Divorced	220	No
No	Single	85	Yes
No	Married	75	No
No	Single	90	Yes

Marital Status

{Married}

Defaulted = No

Home-owner	Marital Status	Annual Income	Defaulted
No	Married	100	No
Yes	Married	120	No
No	Married	60	No
No	Married	75	No

{Single, Divorced}

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
No	Single	70	No
No	Divorced	95	Yes
Yes	Divorced	220	No
No	Single	85	Yes
No	Single	90	Yes

Homeowner

Yes

Defaulted = No

Home-owner	Marital Status	Annual Income	Defaulted
Yes	Single	125	No
Yes	Divorced	220	No

No

Home-owner	Marital Status	Annual Income	Defaulted
No	Single	70	No
No	Divorced	95	Yes
No	Single	85	Yes
No	Single	90	Yes

Annual Income < 78

Yes

Defaulted = No

Home-owner	Marital Status	Annual Income	Defaulted
No	Single	70	No

No

Defaulted = Yes

Home-owner	Marital Status	Annual Income	Defaulted
No	Divorced	95	Yes
No	Single	85	Yes
No	Single	90	Yes

Decision tree: Code

- See loan.r and loan.csv