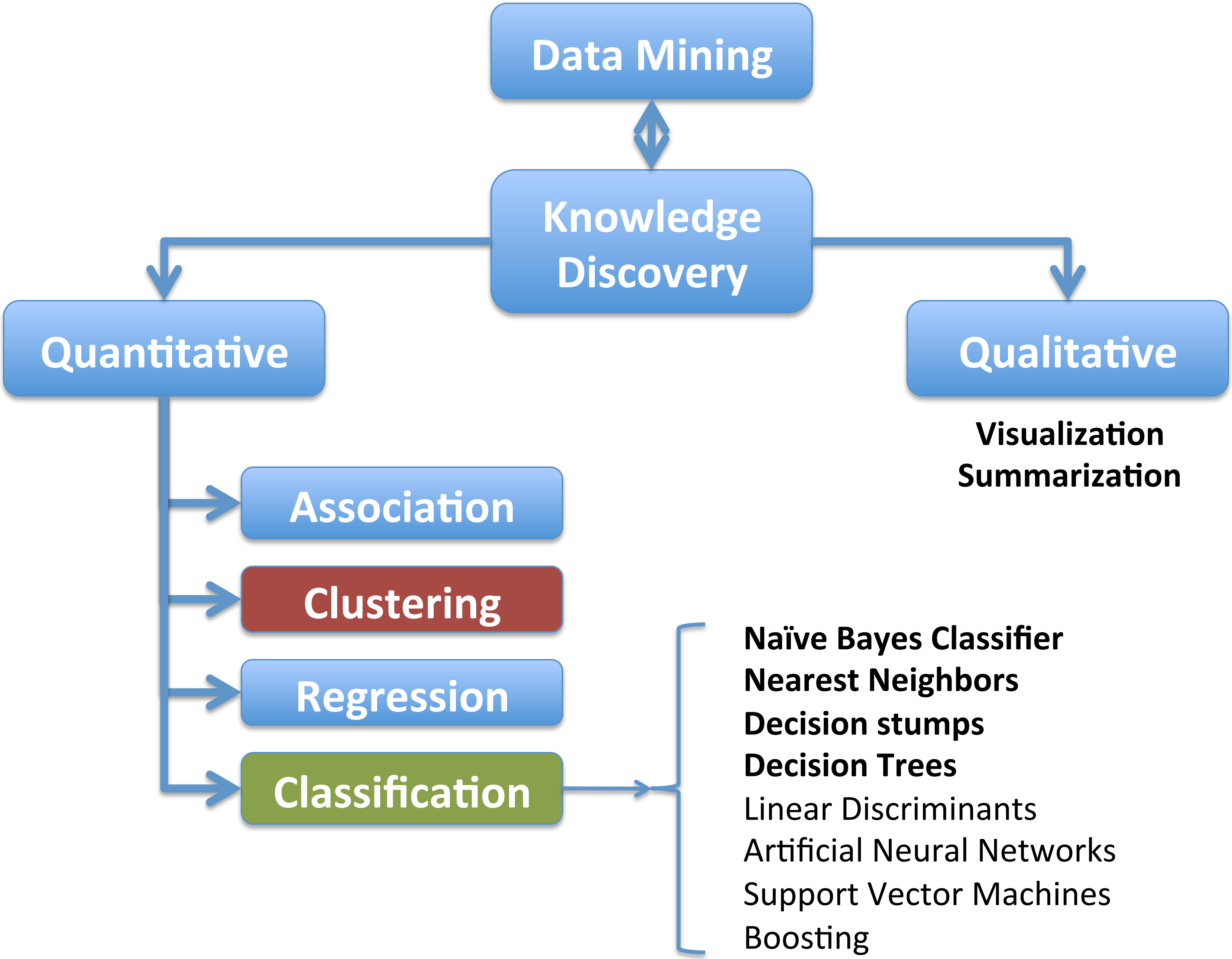
dads 2009|2010 spring

Theme: ***Data Mining for Architecture and Urban Planning***

Lecture V
Classification

Ceyhun Burak Akgül, PhD in EE

Ahu Sökmenoğlu, M. Arch.



Data Mining

Knowledge Discovery

Quantitative

Qualitative

Visualization
Summarization

Association

Clustering

Regression

Classification

Naïve Bayes Classifier
Nearest Neighbors
Decision stumps
Decision Trees
Linear Discriminants
Artificial Neural Networks
Support Vector Machines
Boosting

Working Example – 1/2

Concept A “Taşkışla” Student

Attributes

Age
Gender
Year
Department
#Projects
FirstChoice
HappyWithChoice
ColleageInFamily
AltProfession

LivesWith
LivesWhere
TimeSpentTo…

BeenAbroad
#Languages
PlaysInstrument
Dancing
BeenInFestival
PracticeSports
ReadNewspapers
ReadComics
EnjoyLiterature

Uses3DModelingSW
UsesGraphicsSoftware
WorkedAtOffice
DesignDraftChoice
OfficeOrSite
FaveCourses

FollowsProfPeriodicals
FollowsProfActivities
VoluntaryProfActivities
SpendsTimeAtTaskisla
ActiveStudentClubber

Working Example – 2/2

Data Table: N Subjects (i.e. Student Instances)

	Age	Gender	Year	Dept	#Projects	Happy
Subject ID	**Att1**	**Att2**	**Att3**	**Att4**	**Att5**	**Att6**
1	21	F	3	Arch	6	Yes
2						
3						
4						
5						
6						
7						
8						
9						
…						
N						

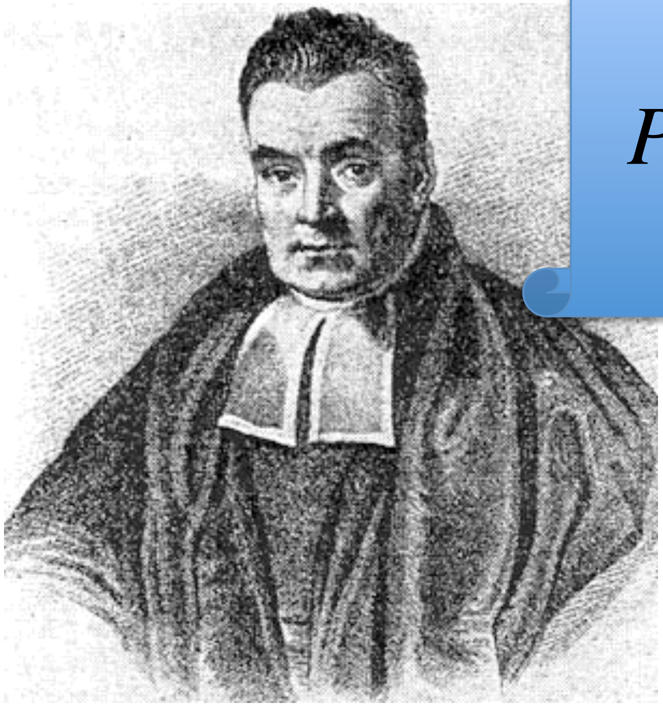
Predictor attributes

vs

Target attributes

Naïve Bayes Classifier – 1/3

*Reverend Thomas Bayes once said…**



$$P(C \mid F) = \frac{P(F \mid C)P(C)}{P(F)}$$

What' s all this about?
- see whiteboard

* Essay Towards Solving a Problem in the Doctrine of Chances (1764)

Naïve Bayes Classifier – 2/3

The classification problem is, having observed a set of attributes *F* about an entity *X*, to assign *X* to one of the possible classes *C = C1, C2, …, Ck*

The best possible classification rule is

$$\begin{gathered} Assign\ X\ to\ C^* \\ such\ that\ P(C^* \mid F) > P(C \mid F) \\ for\ all\ C \neq C^* \end{gathered}$$

Naïve Bayes Classifier – 3/3

One of the problems is how to obtain $P(C \mid F)$ when *F* consists of **multiple attributes**, i.e.,
$F = (F1, F2, \ldots, Ft)$

Naïve Bayes is a **naïve but working approach**

$$P(C \mid F) \propto P(F \mid C)P(C)$$
$$= P(F1, F2, \ldots, Ft \mid C)P(C)$$
$$= P(F1 \mid C) \times P(F2 \mid C) \times \ldots \times P(Ft \mid C)P(C)$$

-by statistical independence-

Naïve Bayes in Action

On a blank paper, write

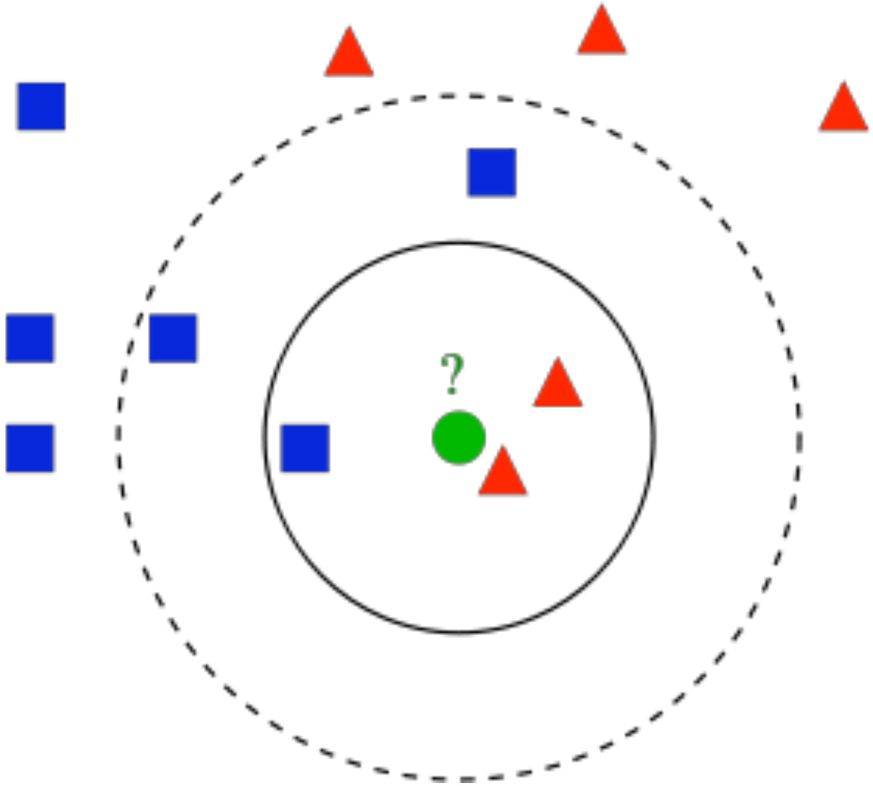
- Your preferred movie:

  English Patient or **The Godfather**?
- Your preferred color: **red** or **blue** or **else**?
- Do you like Obama? **YES** or **NO**.
- Your gender

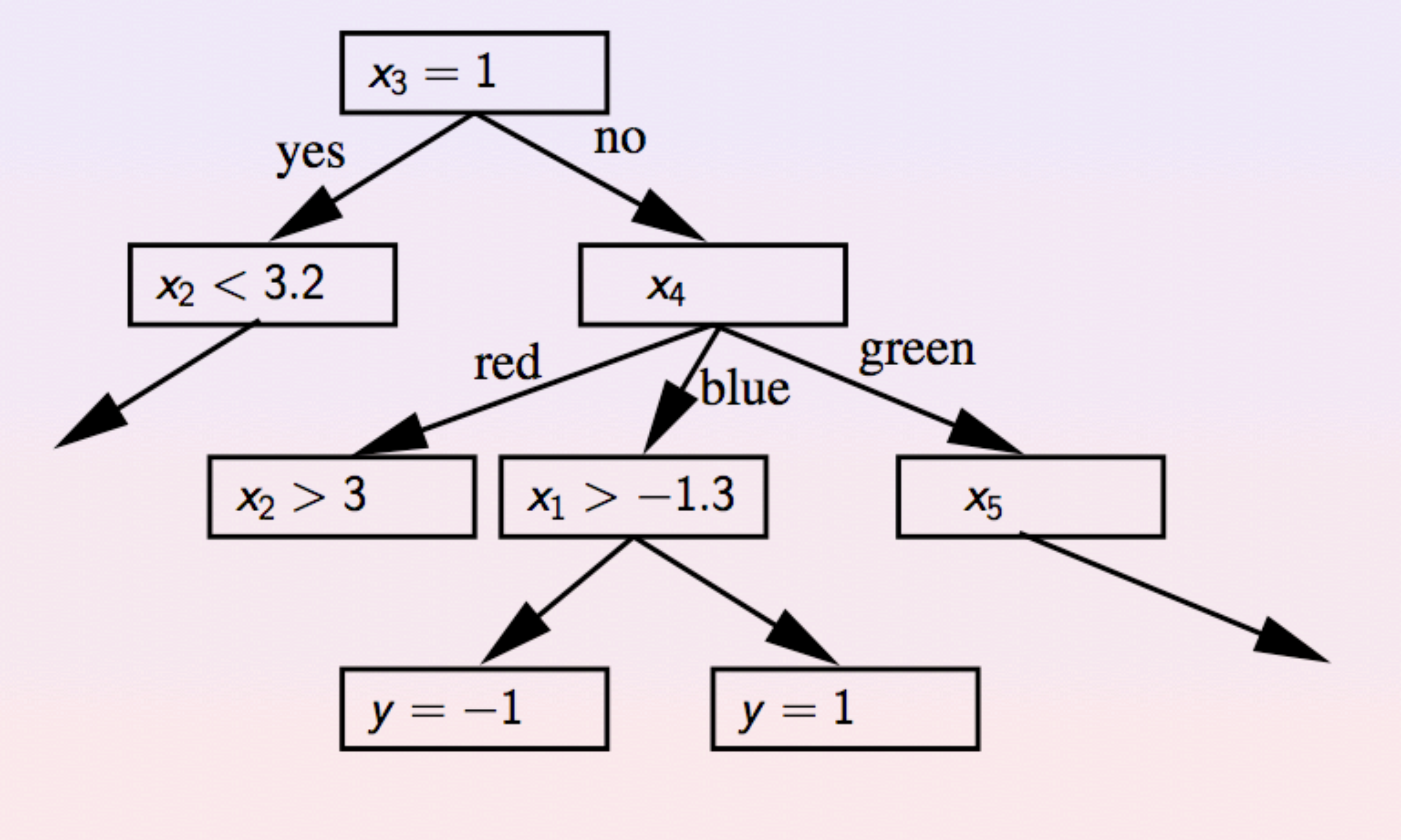
Then we'll have some fun…

Nearest Neighbors

A picture is worth one thousand words…



Decision Trees – 1/2



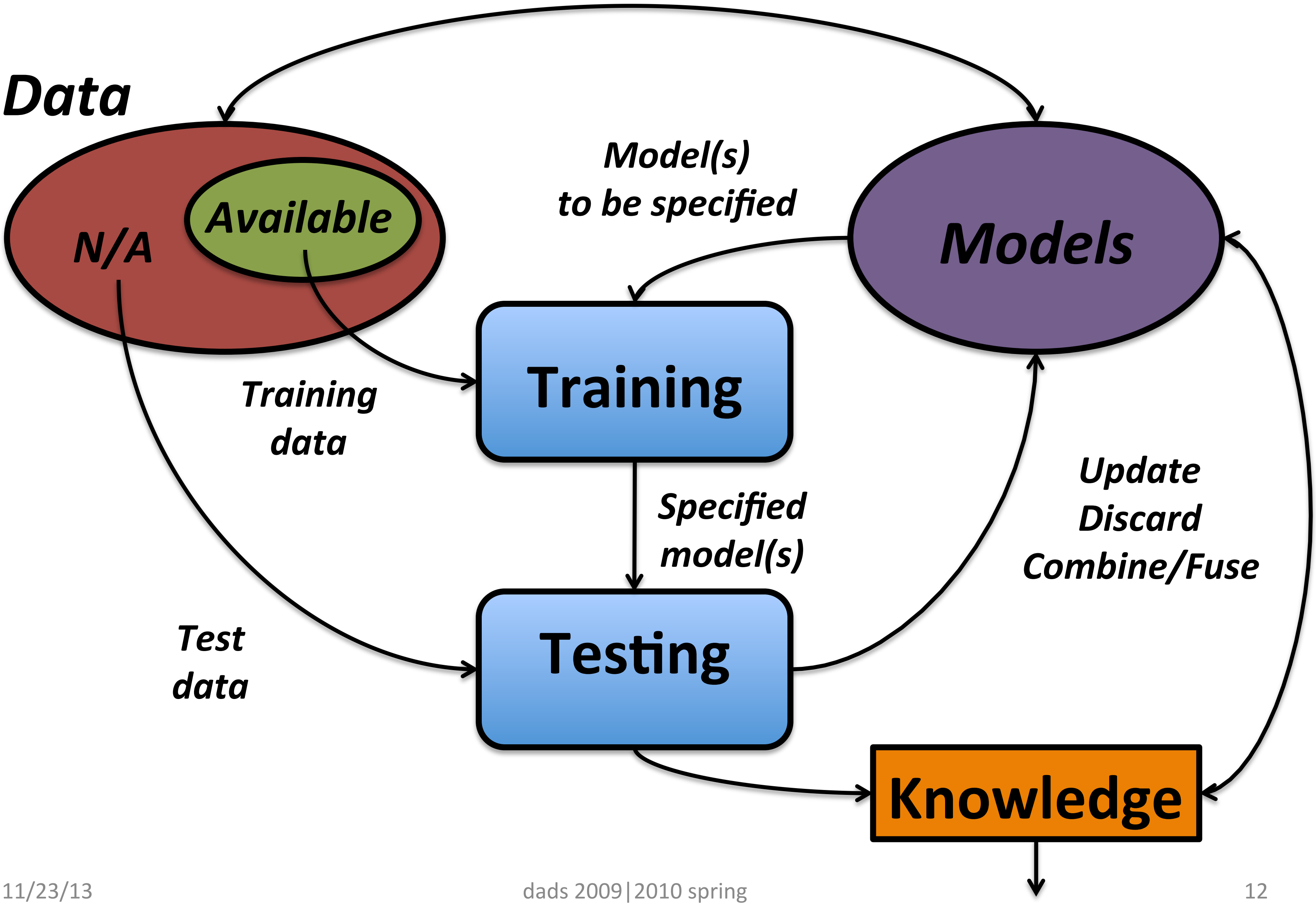
Decision Trees – 2/2

How to grow a decision tree? – A Generic Option

- Sort the attributes by the amount of information they individually contain on the target variable
- Start with the most informative attribute
- Find a splitting point on the current attribute's range of values so as to obtain the least misclassification possible
- Exhaust all the features recursively

Informativeness? Splitting point? Misclassification? Recursively?

Recall…


Data
N/A
Available
Model(s)
to be specified
Models
Training
data
Training
Specified
model(s)
Update
Discard
Combine/Fuse
Test
data
Testing
Knowledge

Assignments for next week

Consider your part in **The Survey**

(A) Identify the attributes

(B) Determine how to code these attributes (their range of values)

(C) Specify one by one the relationships you aim at discovering, and for each case

- **(C.1.)** Determine **predictor** attributes
- **(C.2.)** Determine **target** attributes