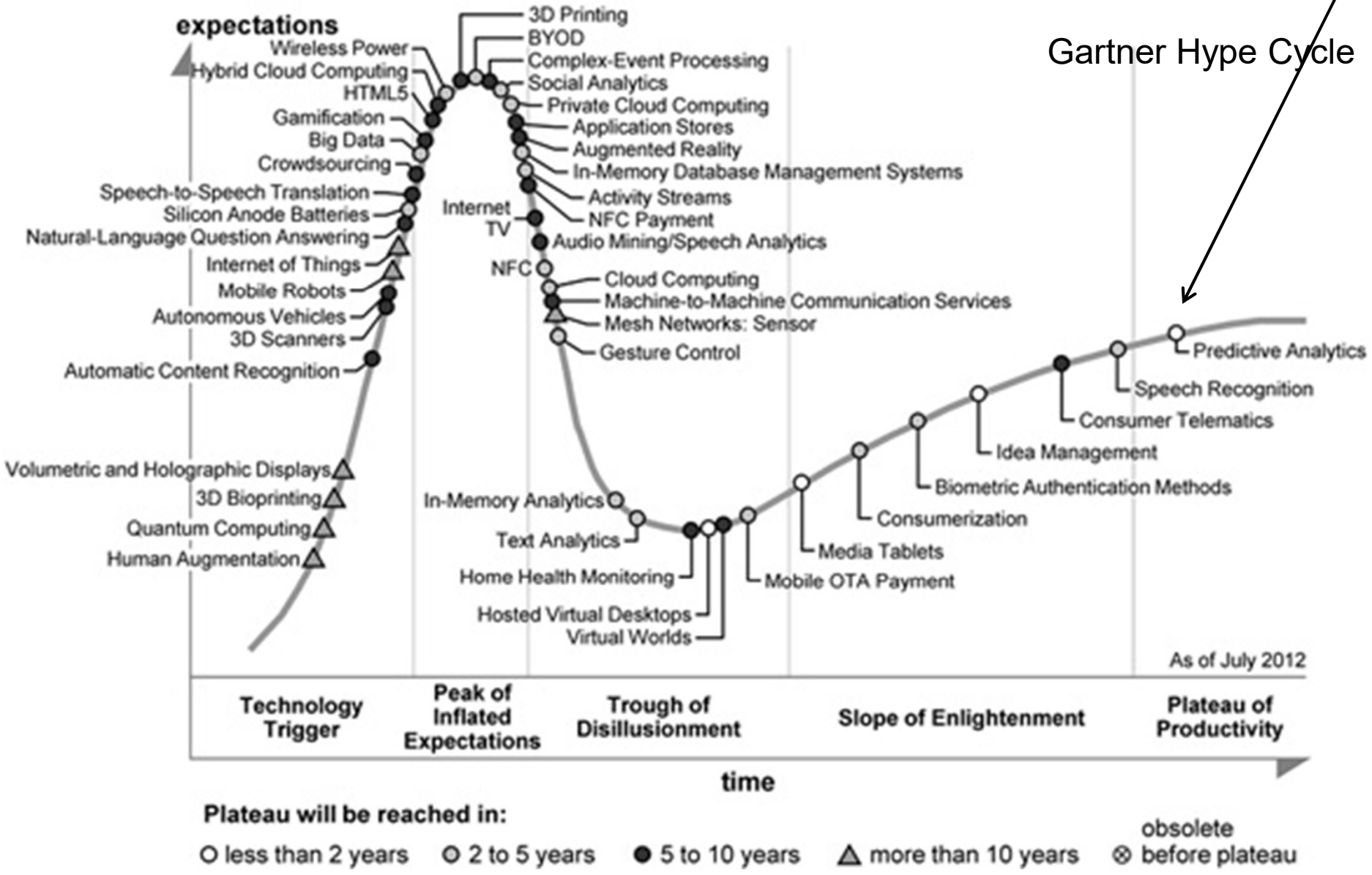
Statistical Learning and Analytics

Predictive Modeling I

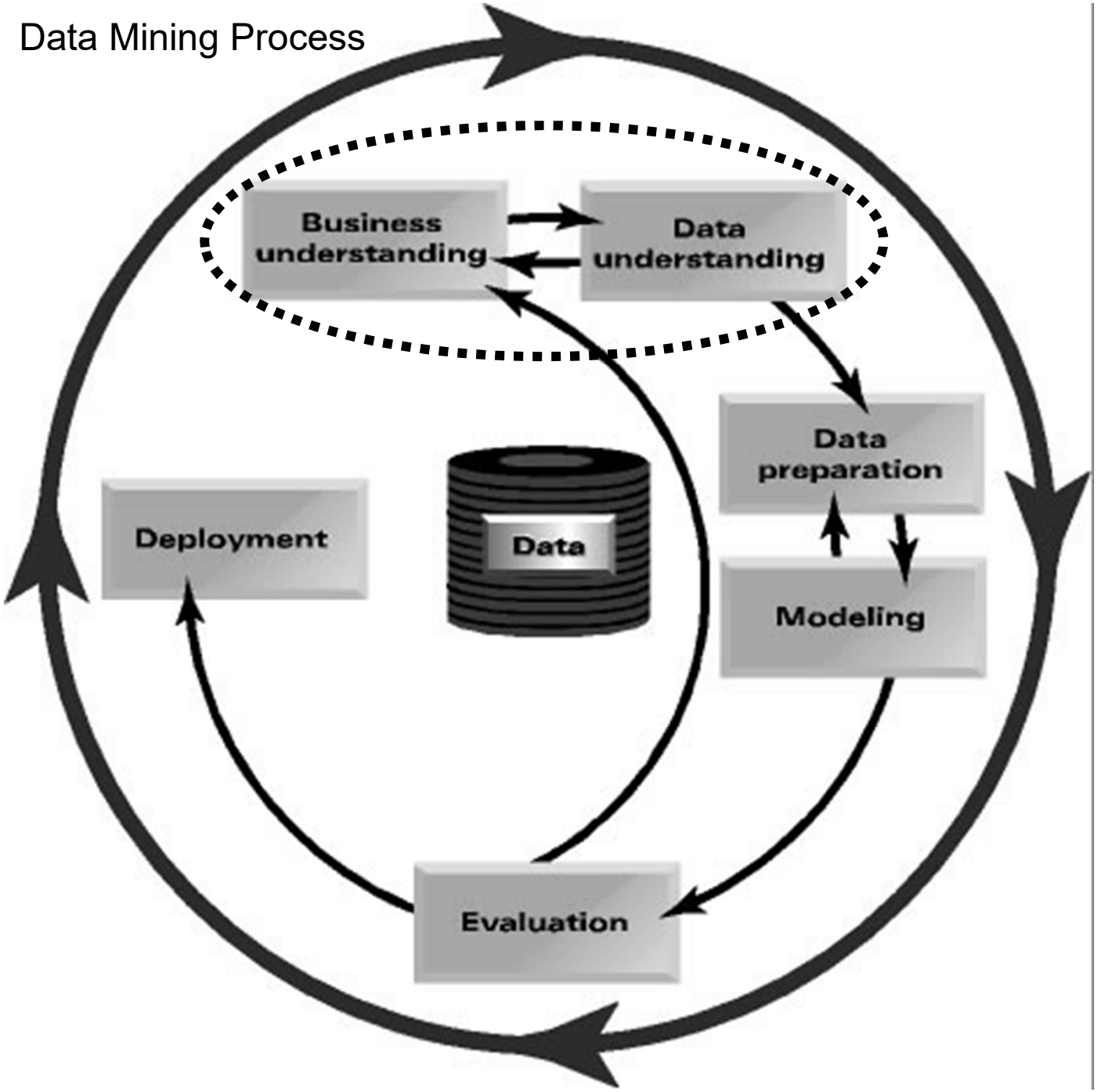
Source: Provost and Fawcett (2013).
Thanks to Maytal Saar-Tsechansky, and Claudia Perlich

Toward Predictive Analytics



Topic: Predictive Modeling 101

Data Mining Process



Supervised Data Mining/ Predictive Modeling

Key (part 1): is there a specific, quantifiable target that we are interested in or trying to predict?

Examples:

- What will the IBM stock price be tomorrow? (e.g., $200)
  What would you do if you could predict this?

- Will this prospect default her loan? (e.g. yes/no)
  What would you do if you could predict this?

- Do my customers naturally fall into different groups?
  [Unsupervised: no objective target stated.]
  What would you do if you could predict this?

Supervised Data Mining/ Predictive Modeling

*Key (part 2): do we have **data** on this target?*

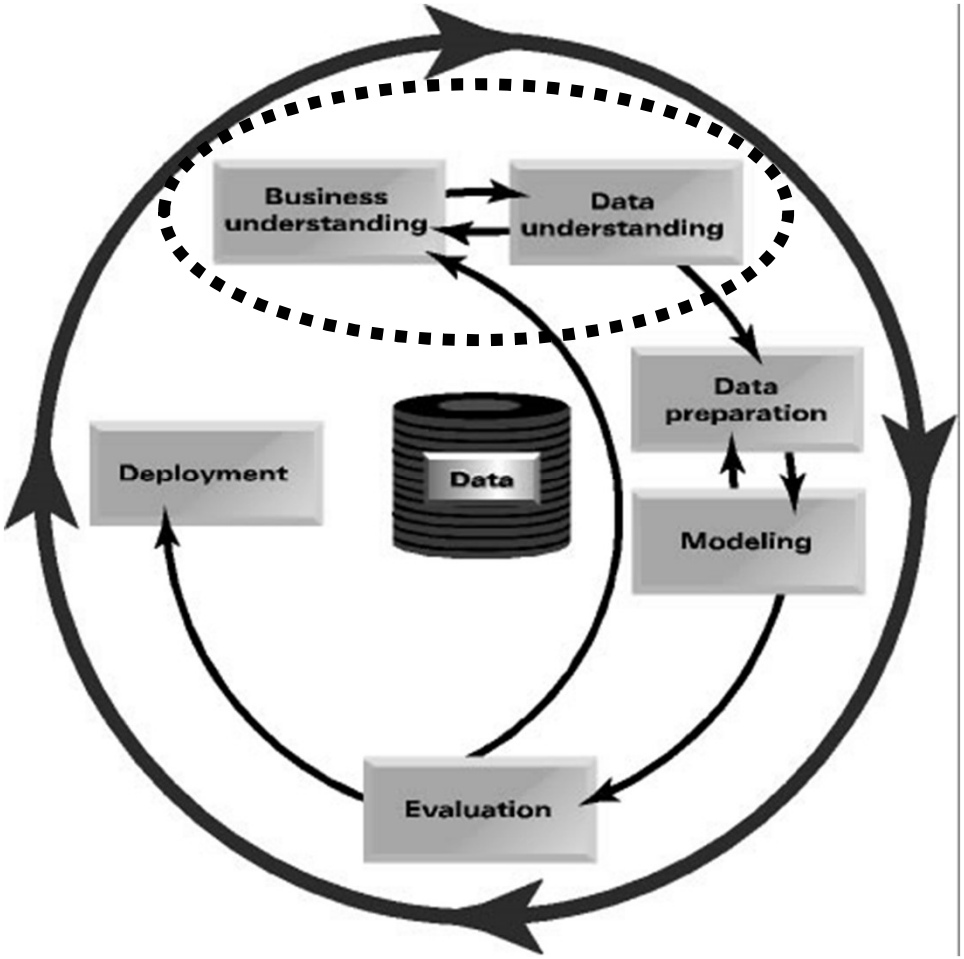
supervised data mining requires both parts 1 & 2

We don't need the exact data (e.g., whether the current customers will leave) BUT we need data for the same or a related phenomonon (e.g., from customers from last quarter)

→ *we will use these data to build a model to predict the phenomenon of interest*

- Think: Is the phenomenon of "who left the company" last quarter the same as the phenomenon of "who will leave the company" next quarter?
- Think: Who might buy this completely new product I have never sold before?

What would target be for ourTelCo churn management problem?



“Supervised Segmentation” Example: Market Life Insurance

- We have a particular life insurance product we would like to sell
- We have a nice offer, but we incur a cost to target it

- How should we proceed?

“Supervised Segmentation” Example: Market Life Insurance

- Buy a large mailing list with demographic information

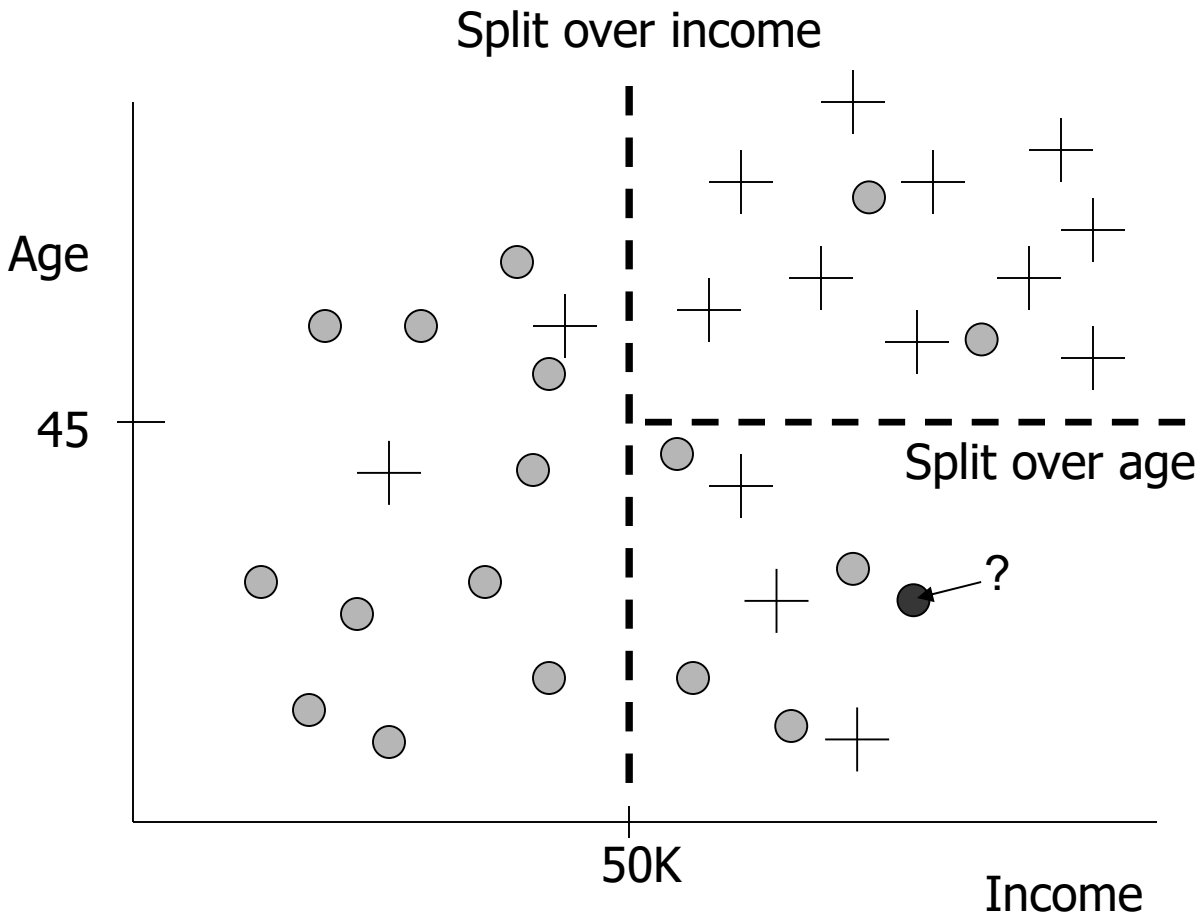
Age	**Income**
35	75K
68	83K
43	61K
71	56K
…	…

Example: Market Life Insurance

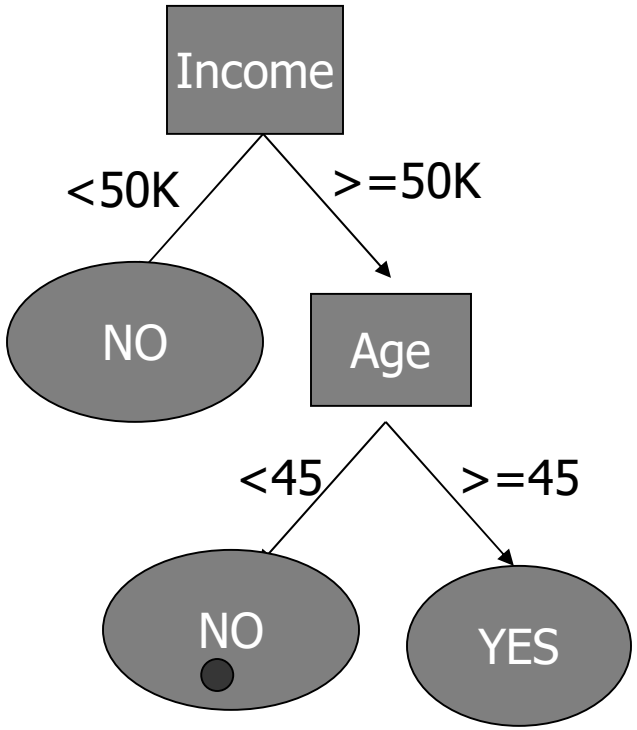
- Send a letter to some prospects in a mailing list
- Wait for a response …

Age	Income	Response?
35	75K	no
68	83K	yes
43	61K	no
71	56K	yes
…	…	…

A supervised segmentation for targeting our Life Insurance product



Classification tree

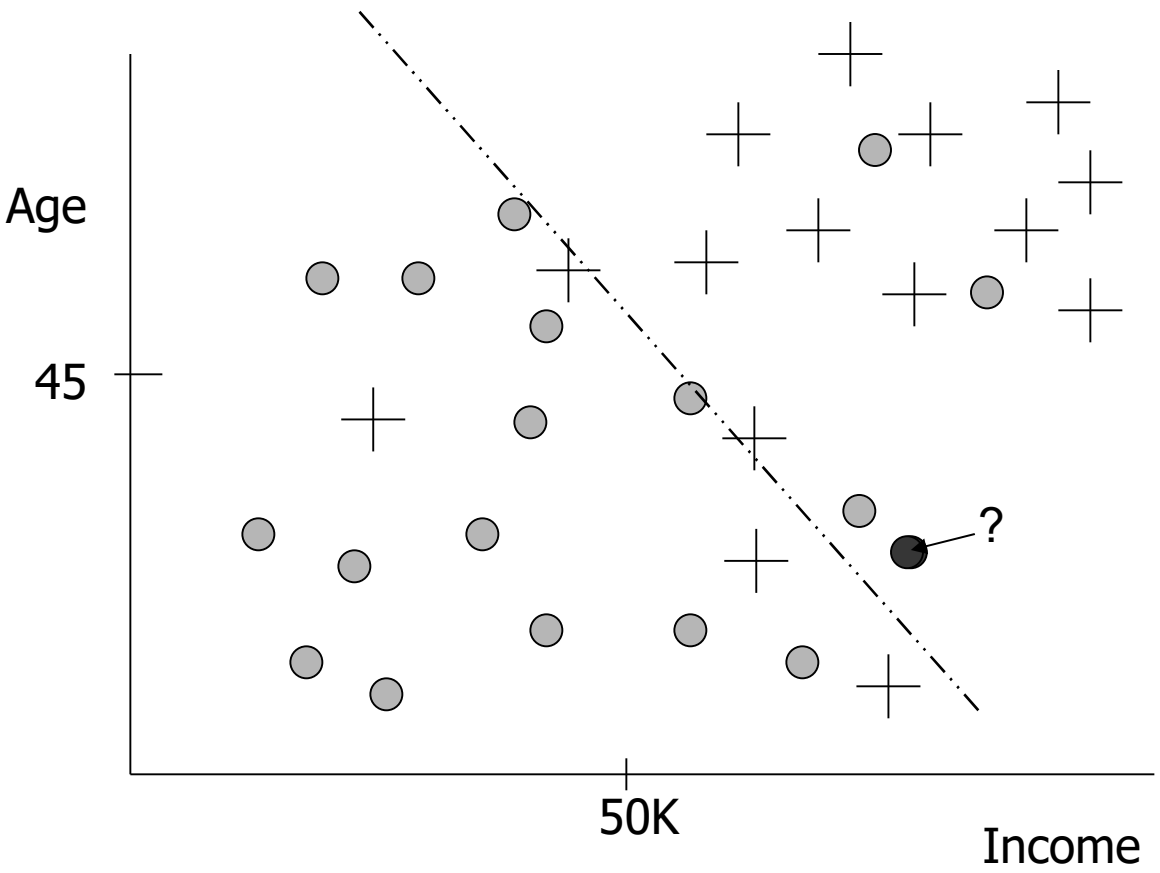


Did not buy life insurance

Bought life insurance

Interested in LI? NO

A different sort of supervised segmentation for our Life Insurance product



Logistic Regression

$$p(+|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$\beta_0 = 123$
$\beta_1 = -1.3$

p(LI|x) = 0.48

Credit Card Application – 16 cases
No Credit Card Application – 14 cases

Types of Data Mining Tasks

Many business problems have as an important component one of these data mining tasks:

- Affinity grouping (a.k.a. "associations", "market-basket analysis")
  - What items are commonly purchased together?
- Similarity Matching
  - What other companies are like our best small business customers?
- Description/Profiling
  - What does "normal behavior" look like?
    (for example, as baseline to detect fraud)
- Clustering
  - Do my customers form natural groups?

Unsupervised

- Predictive Modeling (including causal modeling & link prediction)
  - Will customer X churn next month/default on her loan?
  - How much would prospect X spend?
  - Who might be good "friends" on our social networking site?

Supervised

Supervised Data Mining/ Predictive Modeling

Key (part 3): the result of supervised data mining is a MODEL that given data predicts some quantity

- if (income <50K)

  then no Life Insurance

  else Life Insurance

  [Result of Supervised: you can apply this rule to any customer and it gives you prediction]

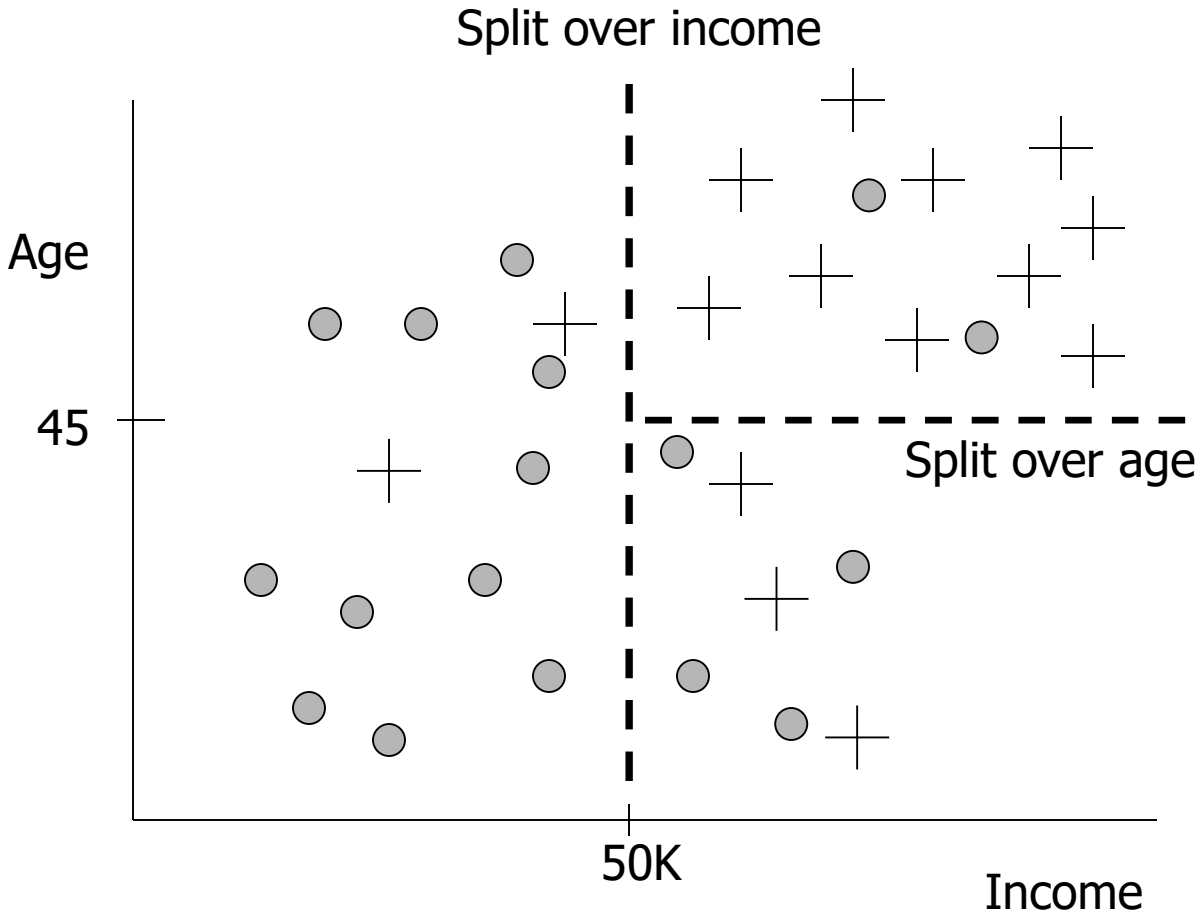
What might a data mining model look like?

- There are different sorts of data mining. Here are just two examples:

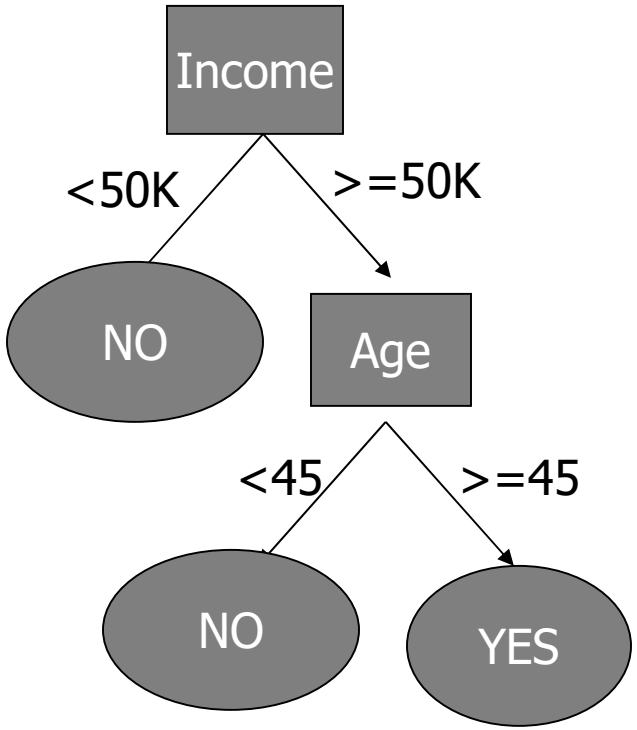
- Tree/Rule: (a supervised segmentation)
- **If (income > $50K) & (age > 45) then LI = YES**
- **If** … **then** …

- Numeric function:
- **P(LI) = f(x_1,x_2,…, x_k)**

What is the model?



Classification tree



Did not buy life insurance

Bought life insurance

Within supervised learning: classification vs. regression?

The difference is the type of target variable:

- classification → categorical target (in historical data)
- regression → numeric target

Supervised Data Mining/ Predictive Modeling

Recall: Key (part 1): is there a specific, quantifiable target that we are interested in or trying to predict?

Which one is classification, which is regression?

- What will the IBM stock price be tomorrow? (e.g., $200)

  What would you do if you could predict this?

- Will this prospect default her loan? (e.g. yes/no)

  What would you do if you could predict this?

- Will the person sign up for life insurance (e.g. yes/no)

  What would you do if you could predict this?

Example: Life Insurance Marketing

Classification vs. Regression?

Think:

- What is the target variable?
- What values can it take in your data?

Age	**Income**	**Response?**
35	75K	no
68	83K	yes
43	61K	no
71	56K	yes
…	…	…

Supervised Data Mining/ Predictive Modeling

*<u>Key</u> (part 4): a data-driven model can either be used to predict or to understand**

**Explanatory modeling can be quite complex. We will return to it*
It turns out that you need to understand the fundamentals of predictive modeling first.

Caveat of classification?

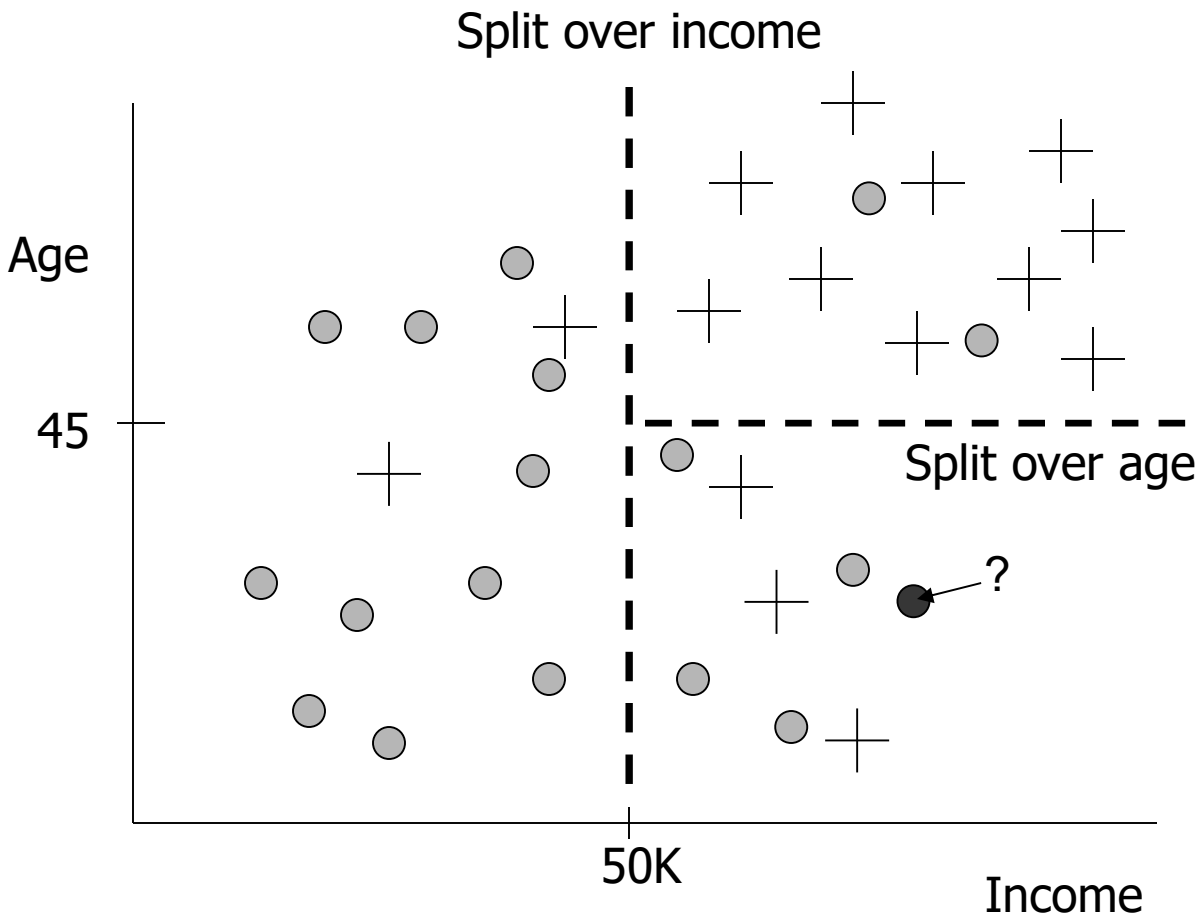
Type of target variable:

- classification → categorical target

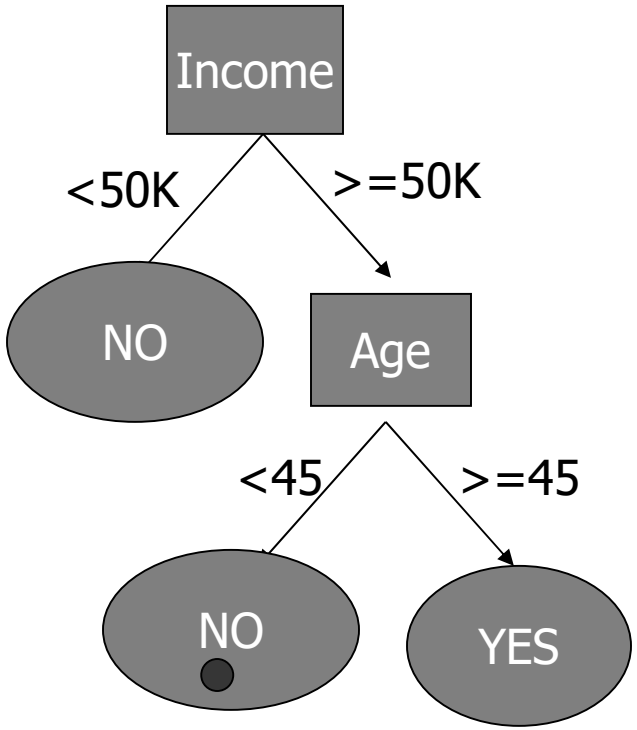
Many classification models can predict continuous values (probabilities, or “ranks”/“scores”)

In that case classification can also be referred to as probability estimation or ranking

What are we predicting?



Classification tree



Did not buy life insurance

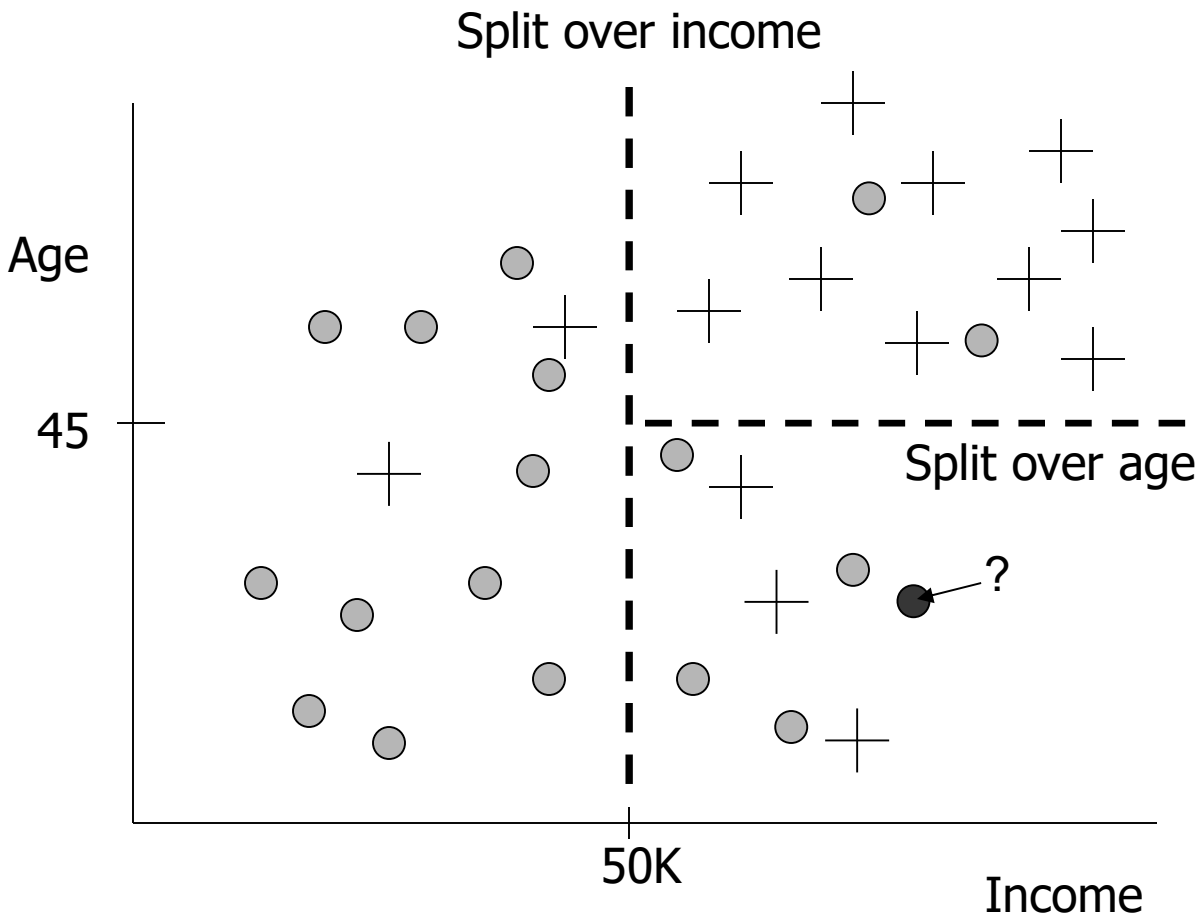
Bought life insurance

Interested in LI? NO

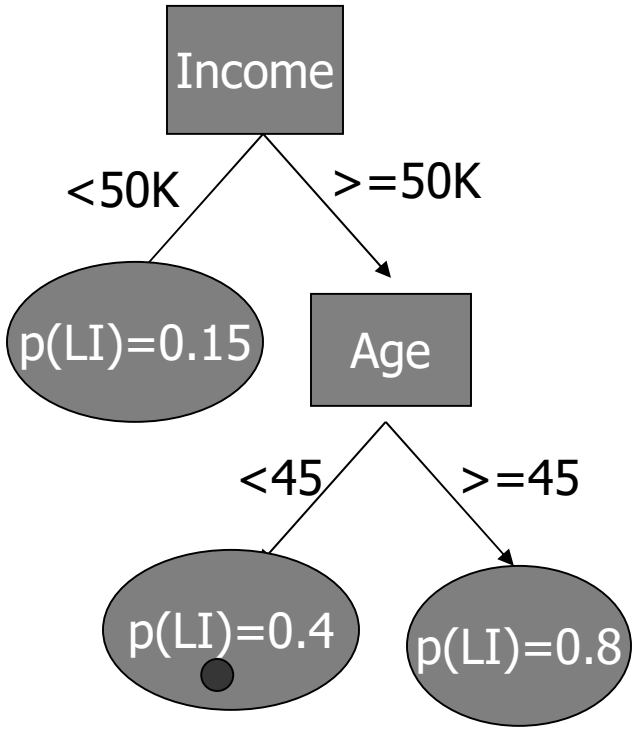
When might a probability be more useful than a yes/no?

- Life insurance targeting?
- Default prediction?
- ….. ?

What are we predicting?



Classification tree

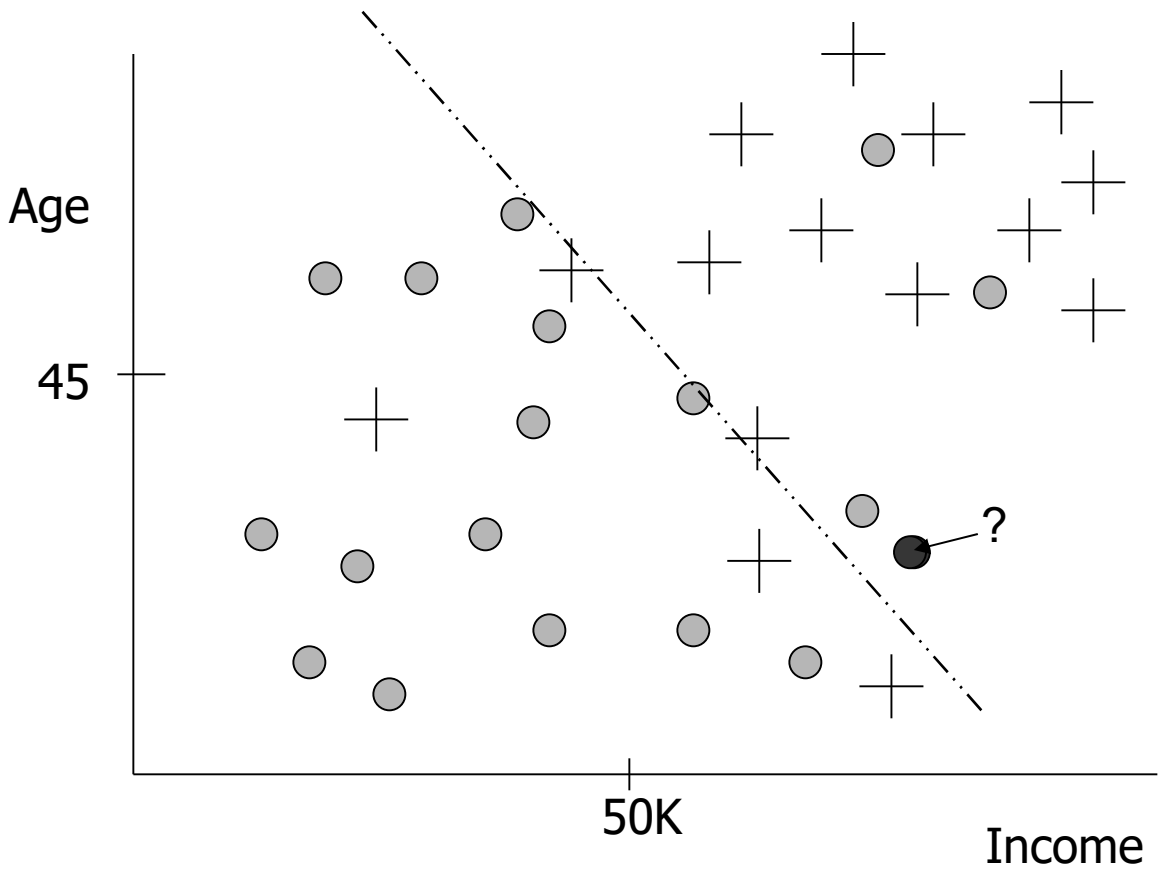


Did not buy life insurance

Bought life insurance

Interested in LI?=3/7

Classification, ranking, or probability estimation?



Credit Card Application – 16 cases

No Credit Card Application – 14 cases

Logistic Regression

$$p(+|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

$\beta_0 = 123$
$\beta_1 = -1.3$

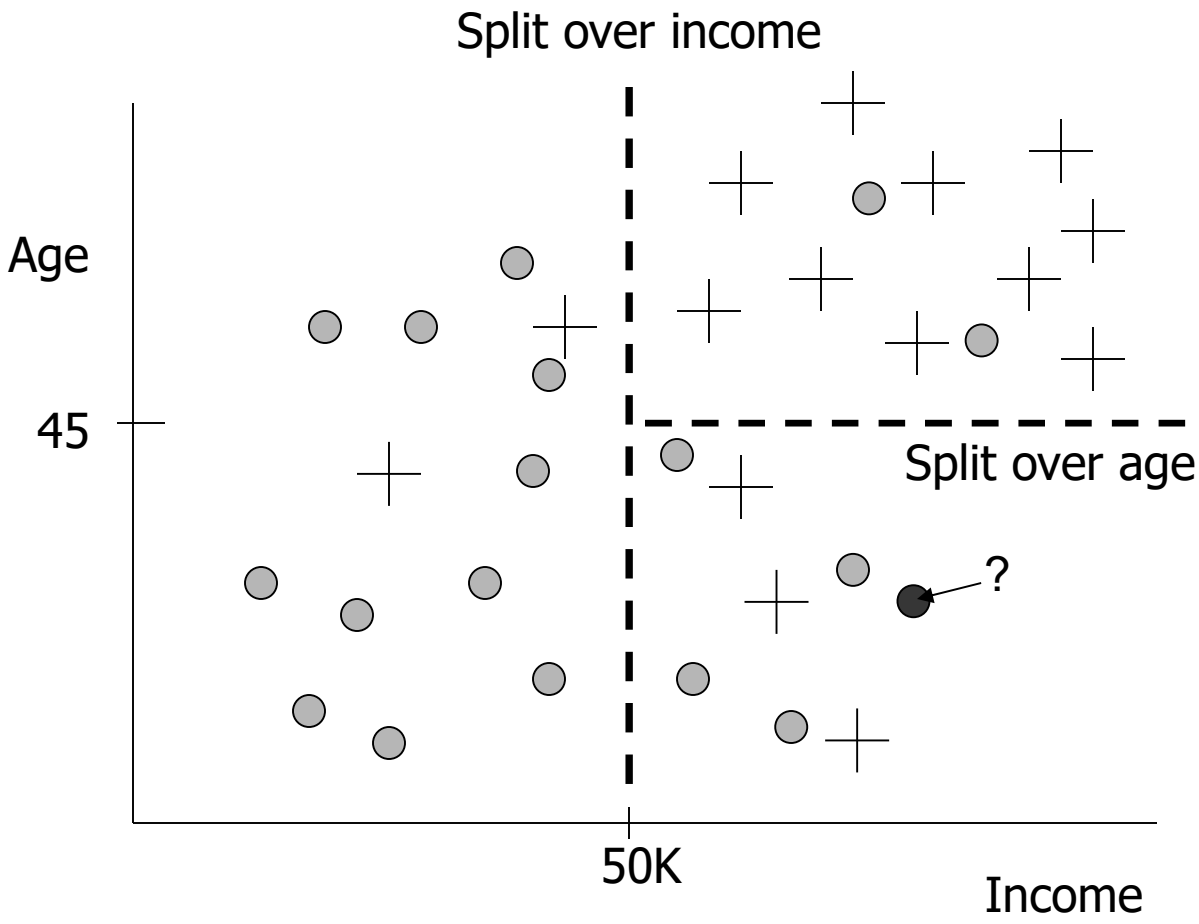
p(LI|x) = 0.48

Supervised Data Mining/ Predictive Modeling

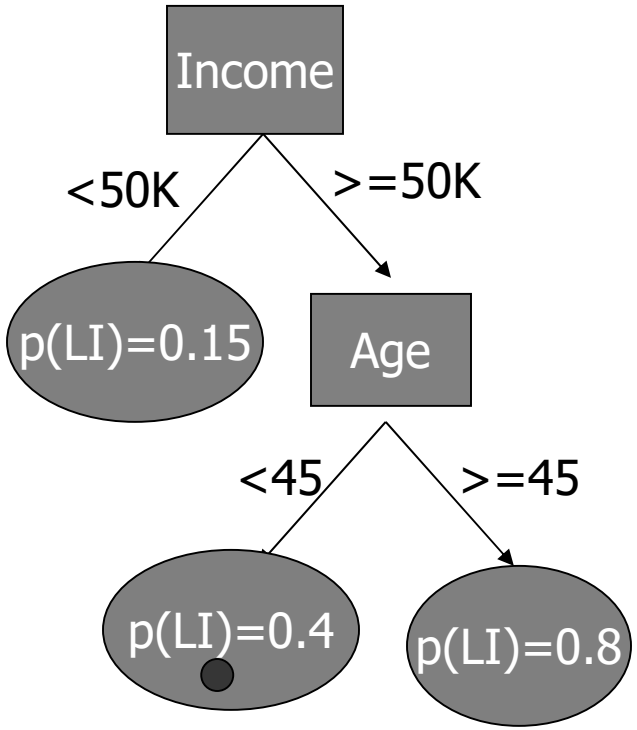
*<u>Key</u> (part 4): a data-driven model can either be used to predict or to understand**

**Explanatory modeling can be quite complex. We will return to it. It turns out that you need to understand the fundamentals of predictive modeling first.*

Which part is prediction?
Which part is understanding?



Classification tree



Did not buy life insurance

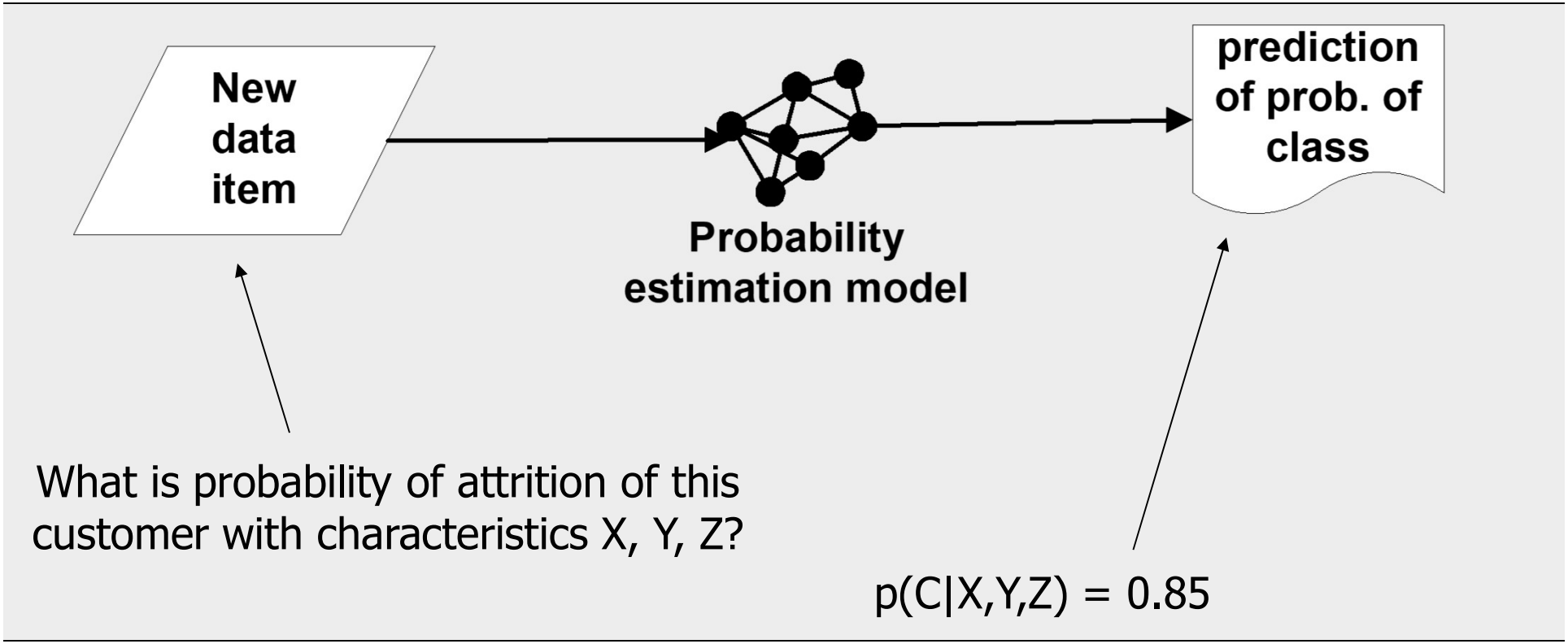
Bought life insurance

Interested in LI?=3/7

Heatmap of Westin Hotels geographic brand affinity

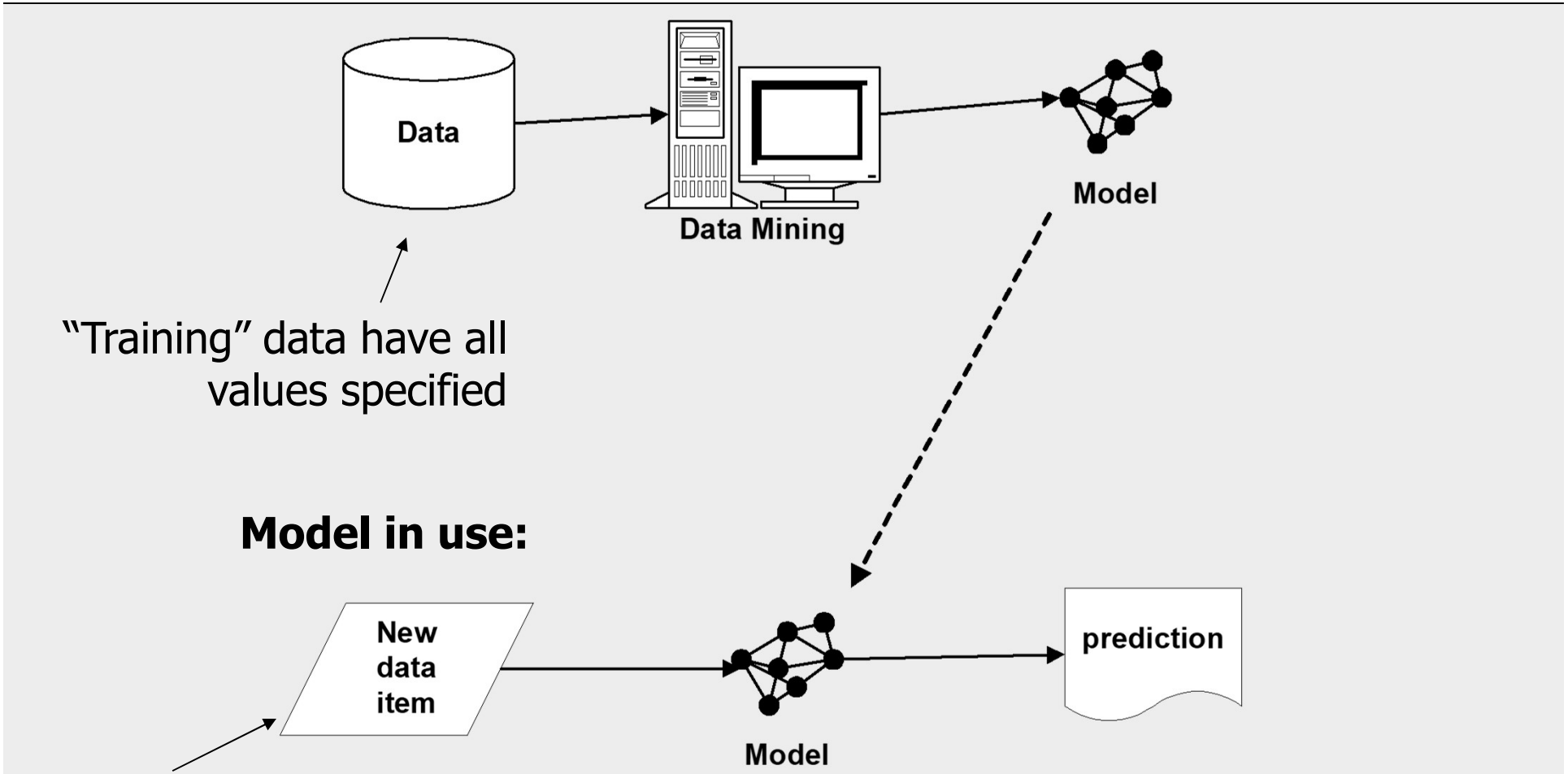
Data mining example:
Predictive model in use

- Example:

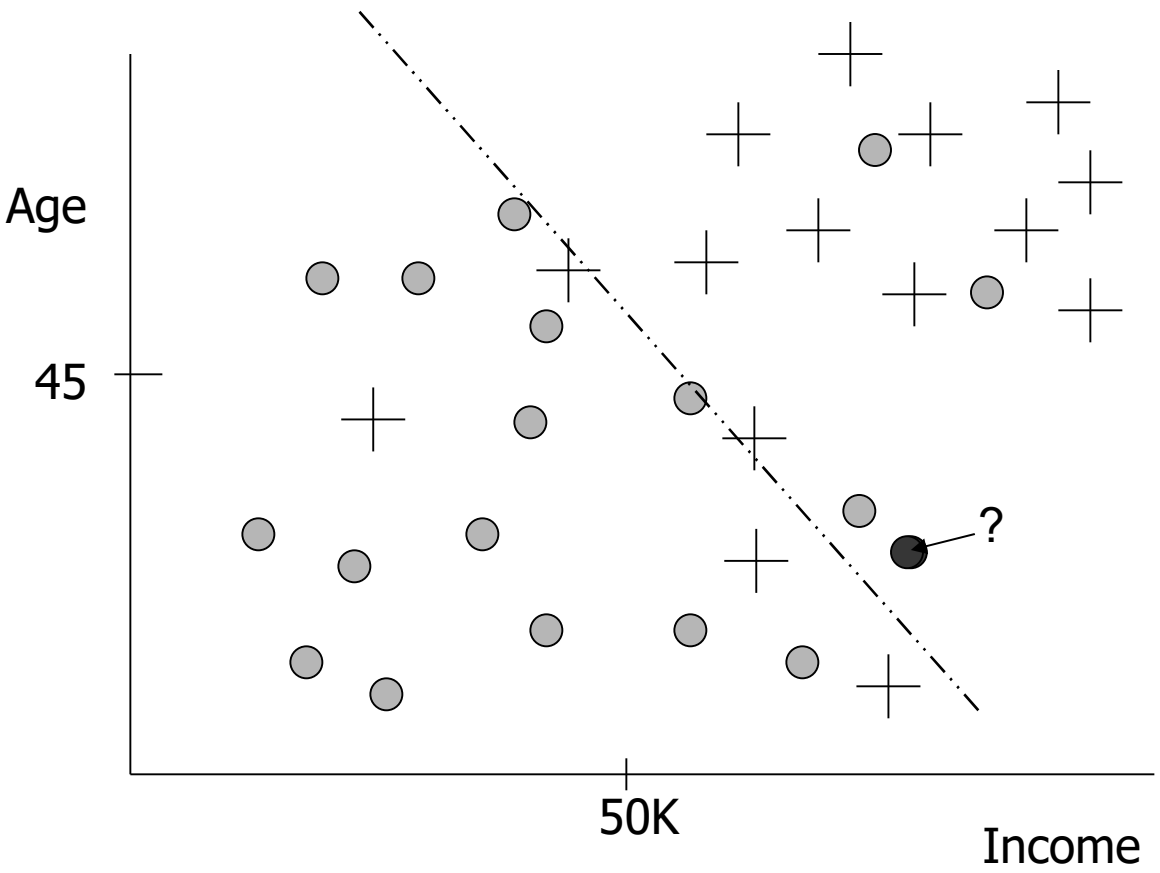


Data mining *versus* Use of the model

"Supervised" modeling:



Which part is Mining? Which is Use?



Logistic Regression

$$p(+|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$\beta_0 = 123$
$\beta_1 = -1.3$

p(LI|x) = 0.48

Credit Card Application – 16 cases
No Credit Card Application – 14 cases

Sidebar: Nuance with classification modeling

MINING:

Type of target variable:

- classification → categorical target

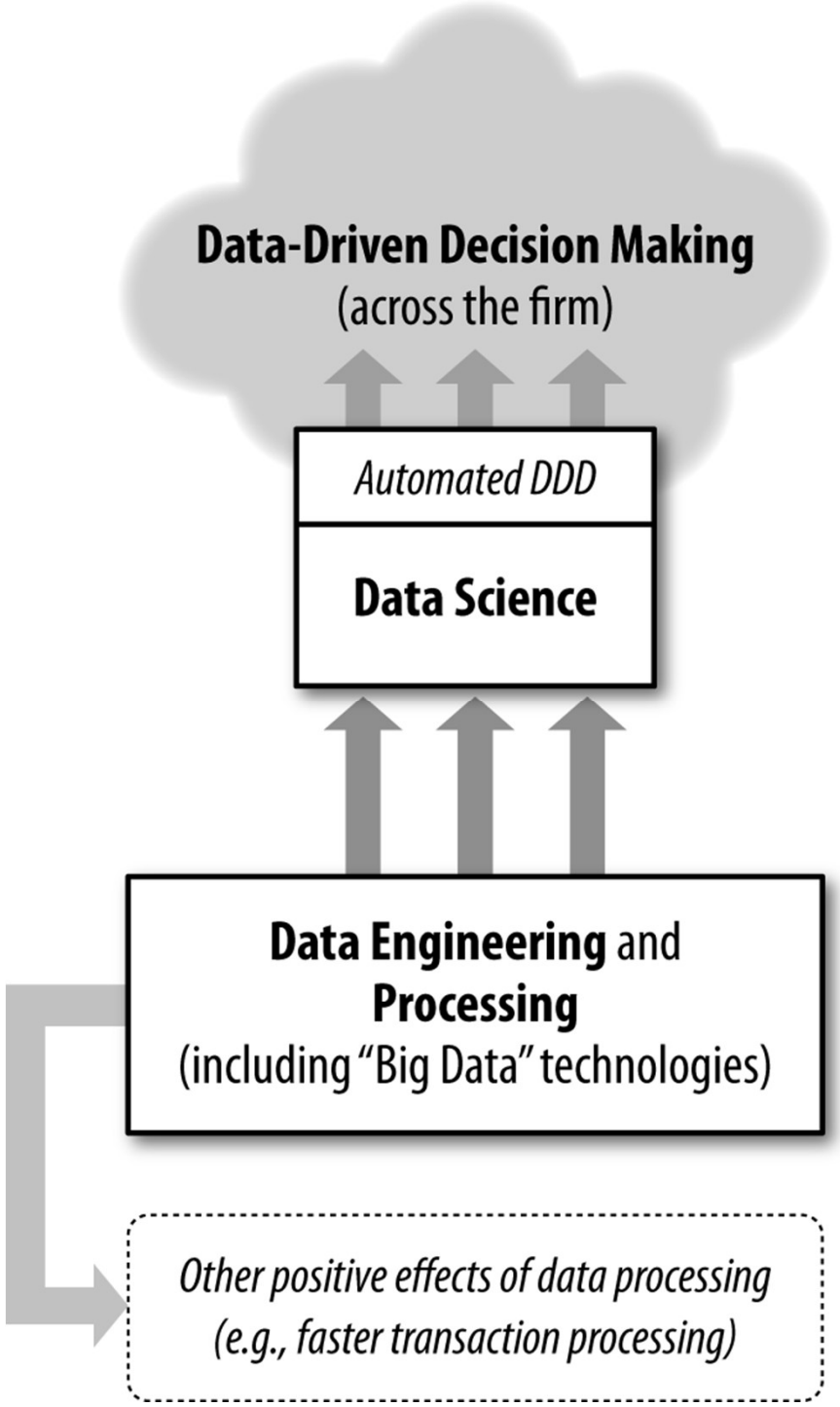
USE:

- Many classification models can predict probabilities and continuous values

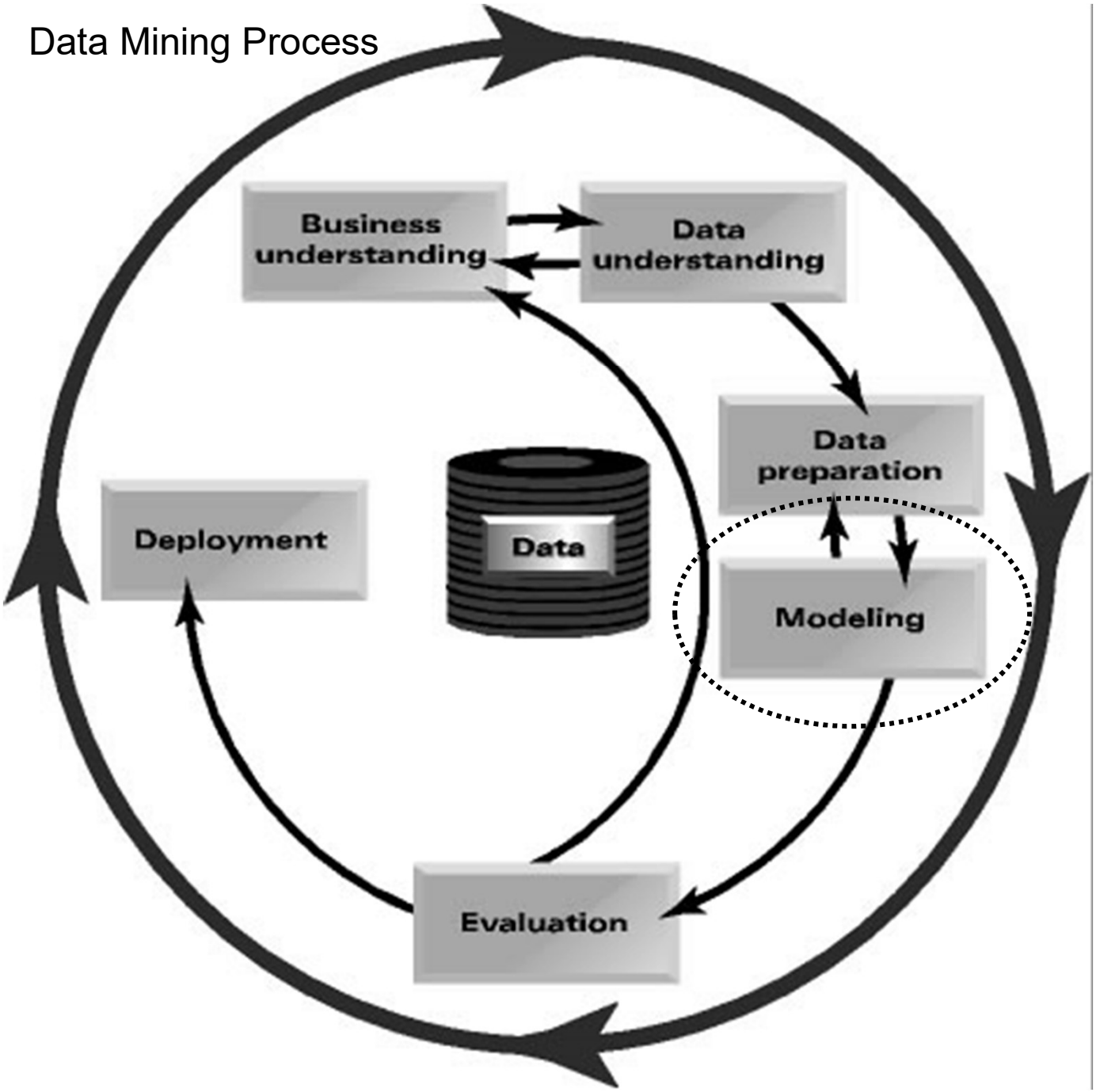
New Example: Credit Default

- Think of yourself as
  - working for lending club
  - investing in lending club
  - legal compliance

Where does predictive modeling come into play?



Data Mining Process



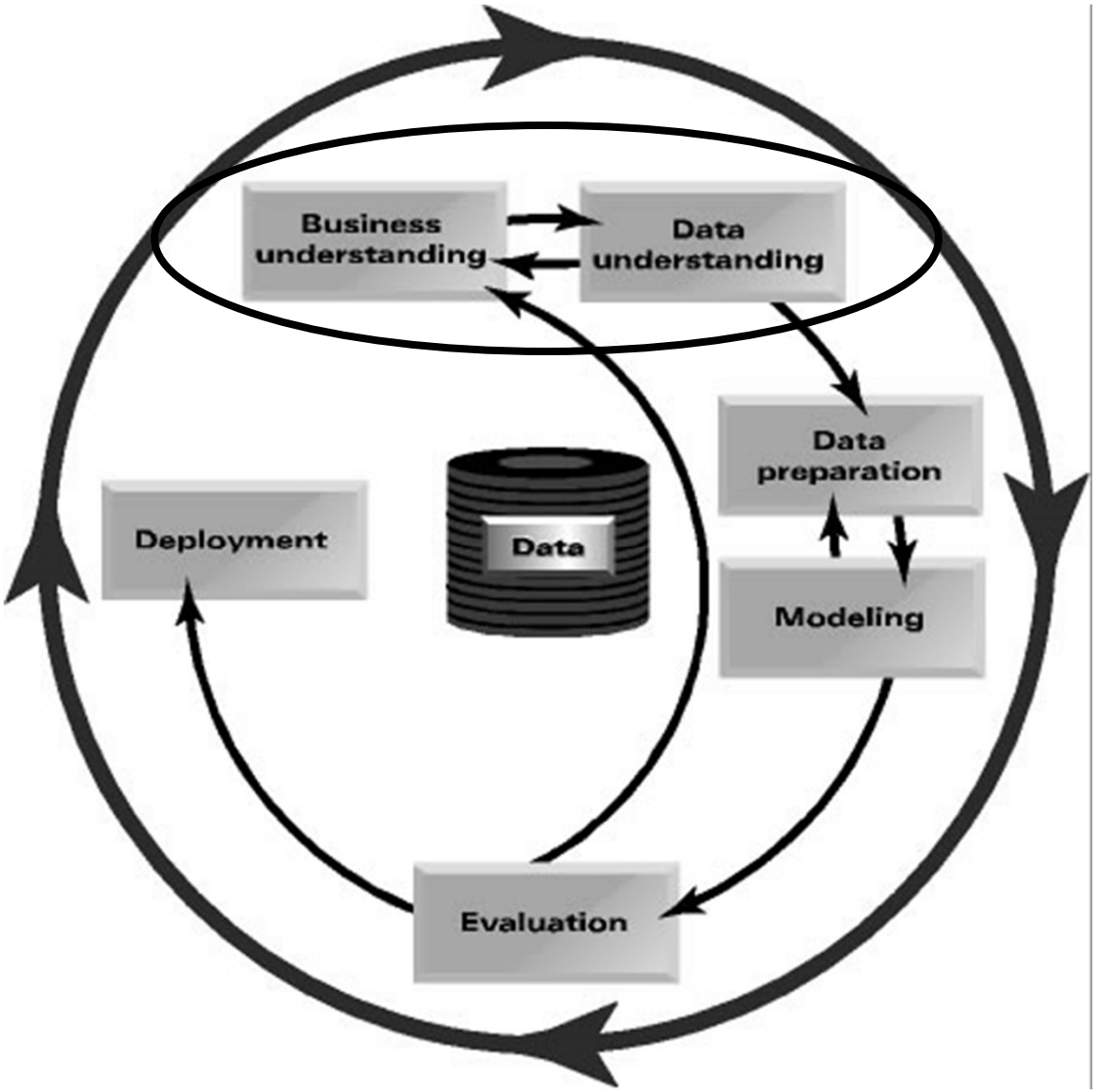
From analysis to analytics…

Why Model?

Progress from an intuitive, "seat of the pants" approach to data-driven decision-making to one based on science & process-based craft

- Frames data selection, acquisition, and investment
- Allows leverage of existing techniques & technology
- Improves consistency of analyses
- Helps to explore data interactively – understand impact of variables
- Helps with communication of results, "selling" ideas
- Can facilitate automated decision-making

Target Case



TARGET

Topic: Terminology

Supervised Data Mining: Terminology

Example, Instance
A fact; a data point

Attributes/Features

One example →

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

A data set/ sample *(as noun)* : A set of examples
"To sample": to choose certain examples

an example of this form sometimes is called a "**feature vector**"

Feature Types

- Numeric: anything that has some order
  - Numbers (that mean numbers)
  - Dates (that look like numbers …)
  - **Dimension** of 1
- Categorical: stuff that does not have an order
  - Binary
  - Text
  - **Dimension** = number of possible values (minus 1)
- Food for thought: Moody's Ratings,Industry codes

Dimensionality of the data?

Attributes/Features

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

"Dimensionality" of a dataset is the sum of the number of numeric features and the number of values of categorical features

Data Mining : Basic Terminology

Induction (a.k.a. ***learning***, ***inductive learning, model induction***)

A process by which a pattern/model is generalized from factual data

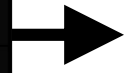
NAME	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

Data Mining: Terminology

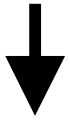
A learner, inducer, induction algorithm

A method or algorithm used to generalize a model or pattern from a set of examples

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no



Learner:
Induces a model from examples



Classification Model:

If **Balance** >= 50K and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

What is a model?

A *simplified** *representation* of reality created for a *specific purpose*.

**based on some assumptions*

- Examples: map, prototype, Black-Scholes model

- Data Mining Example:

  "formula" for predicting probability of customer attrition at contract expiration

  --> "classification model" or "class-probability estimation model"

Pattern/Model?

NAME	Balance	Age	Default
Mike	123,000	35	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	40	no
Mark	23,000	47	yes
Anna	100,000	49	no

Pattern 1:
If **Names starts with M**
Then **Default** = 'yes'
Else **Default** = 'no'

Pattern 2:
Age is inversely proportional to alphabetical order

Pattern 3:
Young people are more likely to default

Pattern 4:
If **Balance** >= 50K and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

Pattern 5:
If **Names ends with 'e'**
Then **Balance > 100000**
Else **Balance** <100000

Good vs bad patterns?

Supervised Data Mining

- *is there a specific, quantifiable target that we are interested in or trying to predict?*
  - *think about the decision in the business problem*
- *do we have enough data on this target?*
  - *need a min ~500 of each type for classification*
- *do we have relevant data prior to decision?*
  - *think timing of decision and action*

Supervised Data Mining: Terminology

Example, Instance

A fact; a data point

Attributes → ← Target

One example →

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

typically described by
a set of attributes
(fields, variables,
features) and a
target variable (label).

Equivalent statistical terminology :

- Attributes - independent variables
- Target - dependent variable

Dimensionality: sum of dimensionality of the attributes excluding target

Supervised Default Model?

Target

NAME	Balance	Age	Default
Mike	123,000	35	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	40	no
Mark	23,000	47	yes
Anna	100,000	49	no

Pattern 1:
If **Names starts with M**
Then **Default** = 'yes'
Else **Default** = 'no'

Pattern 2:
Age is inversely proportional to alphabetical order

Pattern 3:
Young people are more likely to default

Pattern 4:
If **Balance** >= 50K and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

Pattern 5:
If **Names ends with 'e'**
Then **Balance > 100000**
Else **Balance** <100000

Topic: Distinctions in Supervised Predictive Modeling

The many faces of classification: Classification/Probability Estimation/Ranking

Classification Problem

- Most general case: The target takes on discrete values that are NOT ordered
- Most common: binary classification where the target is either 0 or 1

3 Different Solutions to Classification

- **Classifier model**: Model predicts the same set of discrete value as the data had
- **Ranking (binary case):** Model predicts a score where a higher score indicates that the model think the example to be more likely to be in one class
- **Probability estimation:** Model predicts for each class a score between 0 and 1 that is meant to be the probability of being in that class. For mutually exclusive classes, the predicted probs should add up to 1.

Classification?
Probability Estimation?
Ranking

NAME	Balance	Age	Default
Mike	123,000	35	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

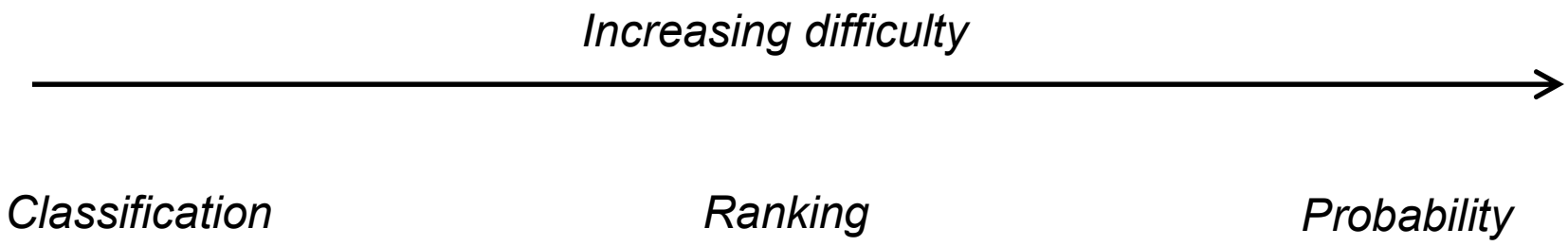
Pattern 1:
If **Names starts with M**
Then **Default** = 'yes'
Else **Default** = 'no'

Pattern 2:
Young people are more likely to default

Pattern 3:
If **Balance** >= 50K and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

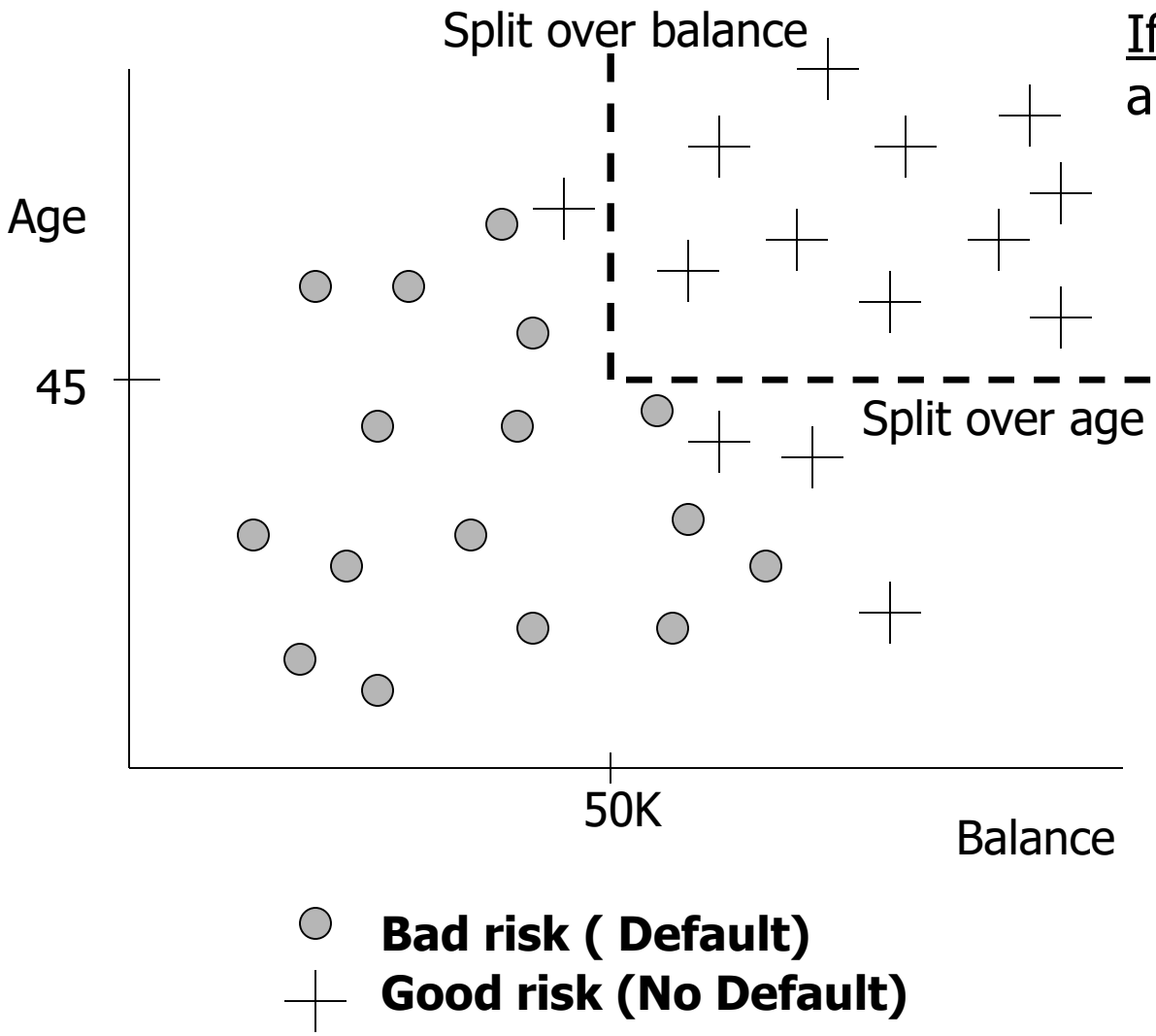
Pattern 4:
If **AGE** >= 45
Then p(**Default)** = 0.25
Else p(**Default)** = 1

When do we need classification, probability estimation, or ranking?



- Classification: (never)
- Ranking:
  - cost/benefit is unknown or difficult to calculate
  - cost/benefit is constant across instances
  - business context determines “how far down the list”
- Probability:
  - cost/benefit is known relatively precisely
  - cost/benefit may be not constant across instances
  - you can always rank/classify if you have probabilities!

Geometric interpretation of a model

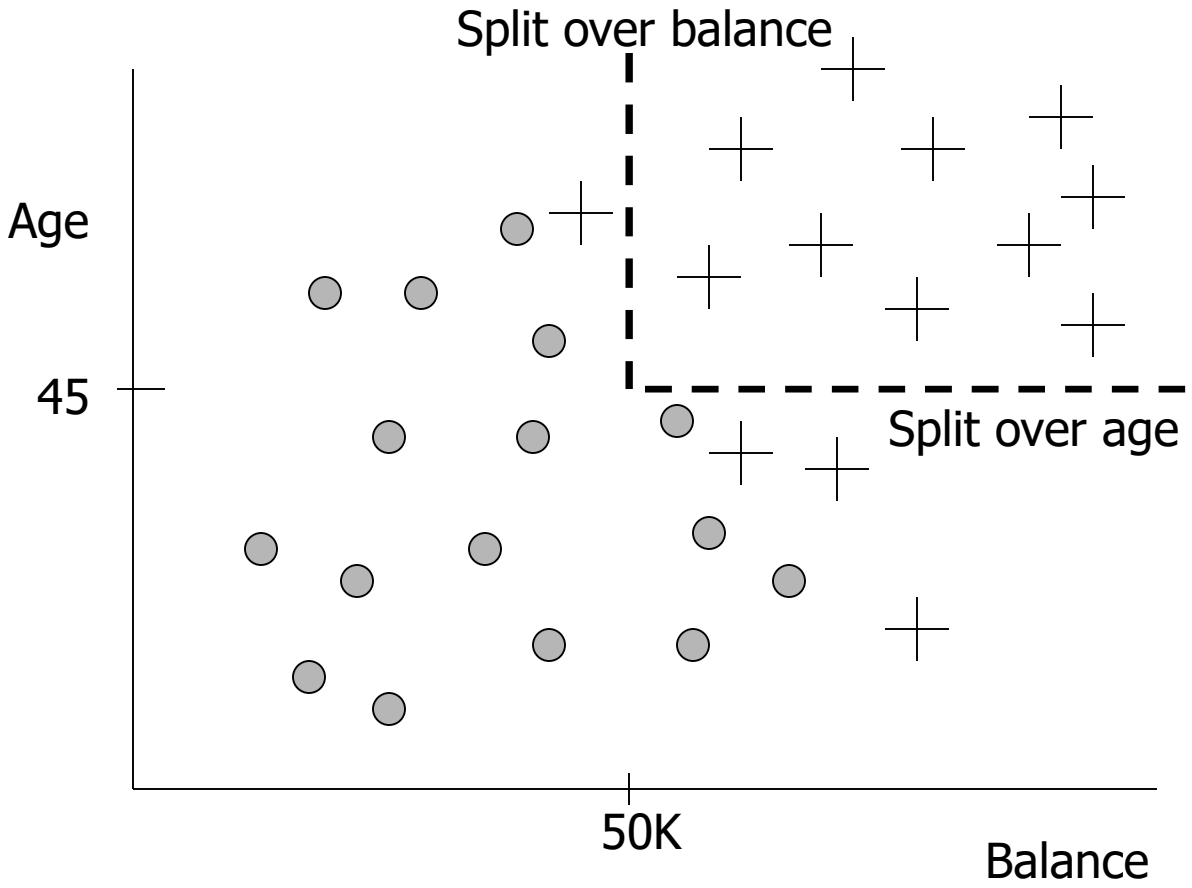


Pattern 2:

If **Balance** >= 50K
and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

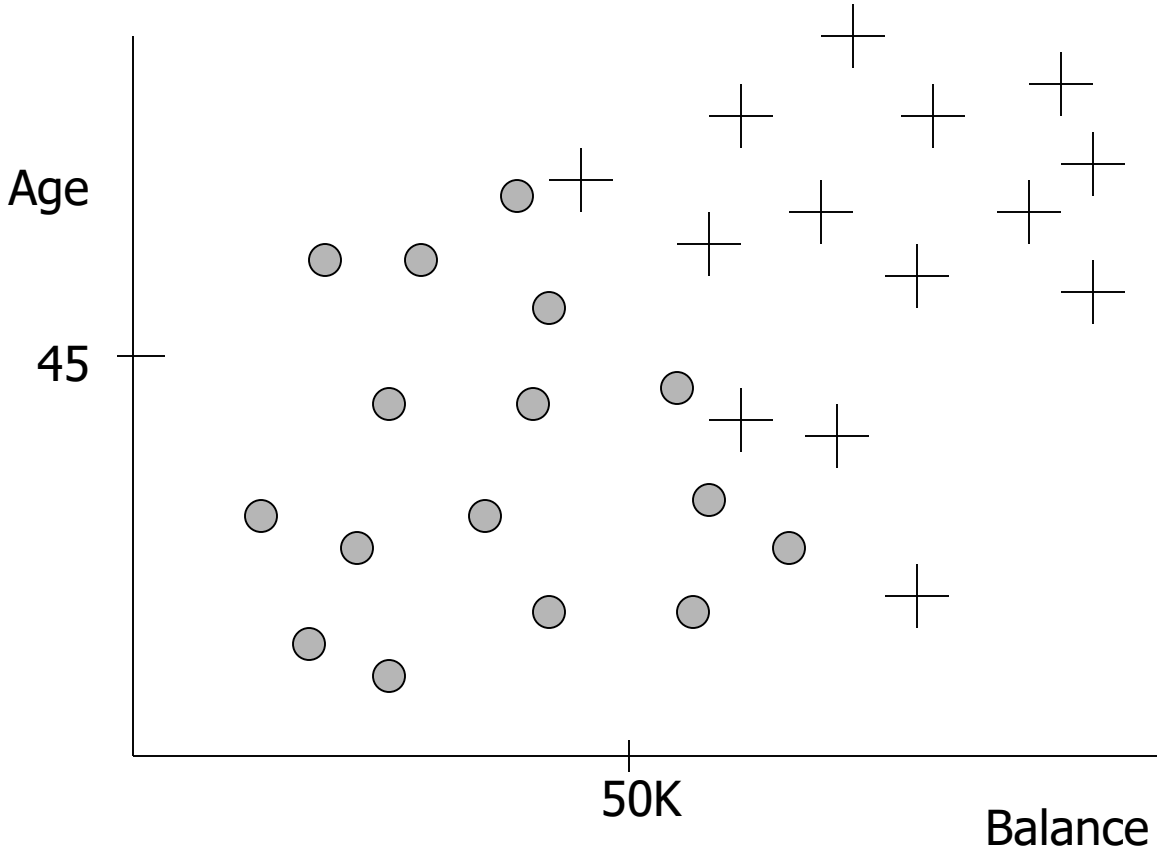
Geometric interpretation of a model

What alternatives are there to partitioning <u>this way</u>?



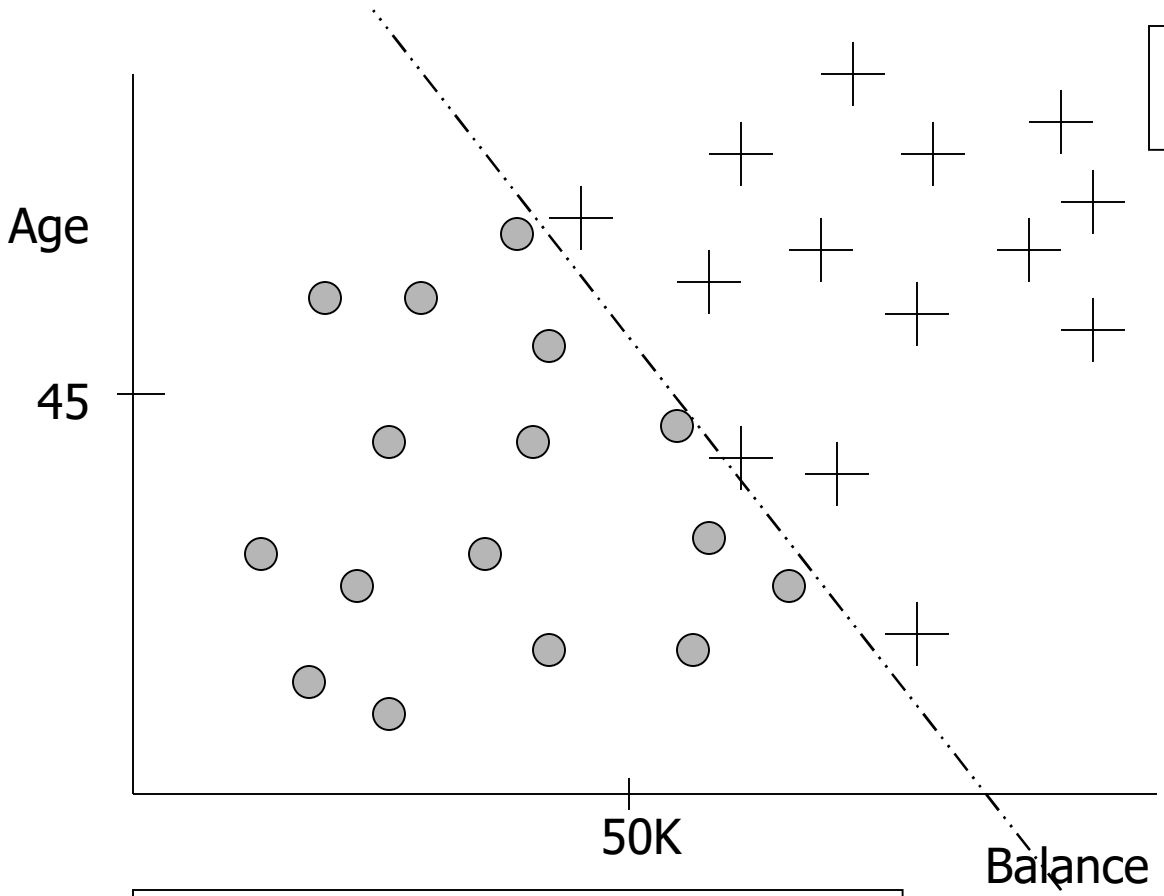
Geometric interpretation of a model

What alternatives are there to partitioning?



Geometric interpretation of a model

What alternatives are there to partitioning?



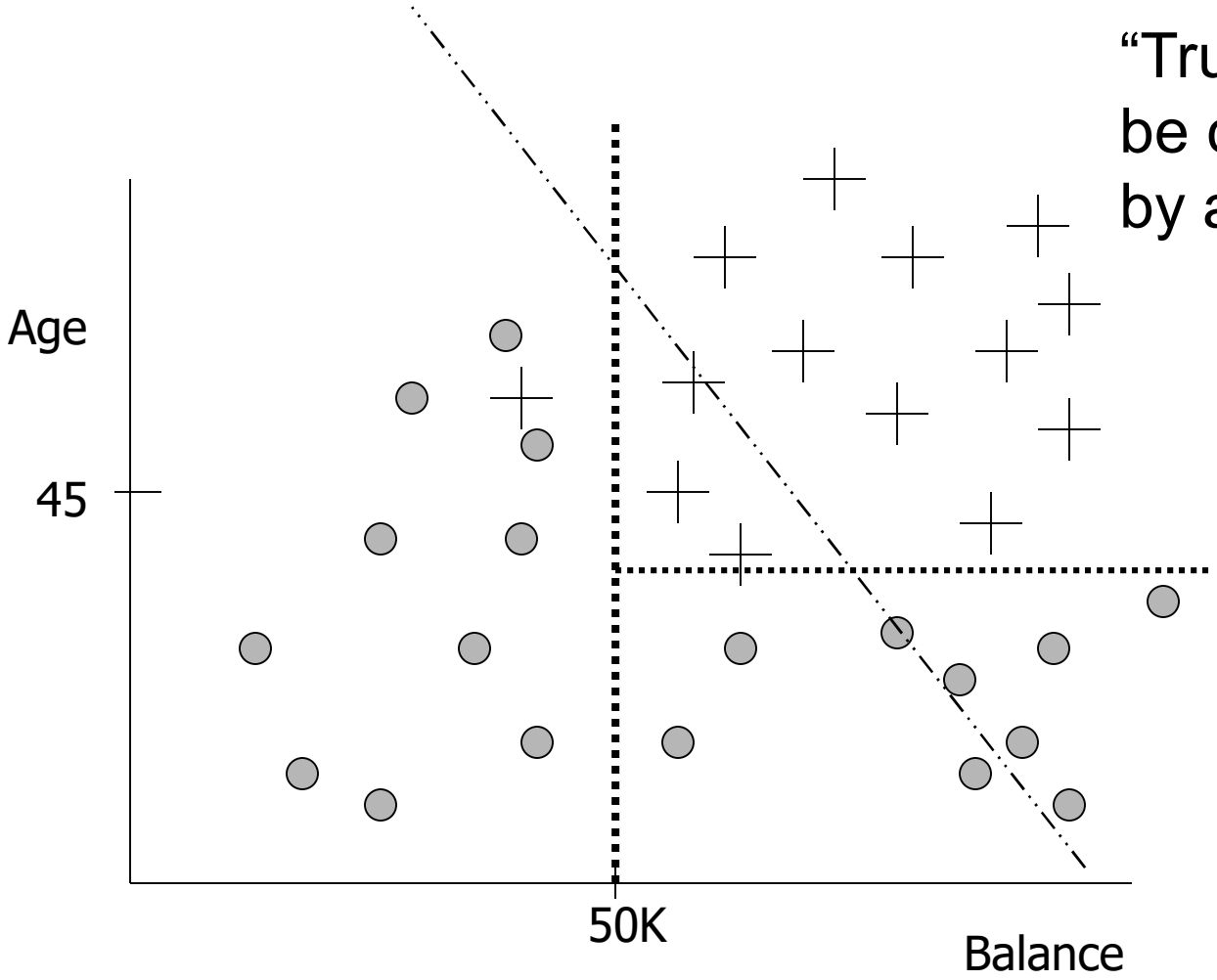
+ if age > -2*Balance + 160

This is called linear discriminant analysis; also basis of "support-vector machines"

logit(-2*Balance + 160 - age)

logistic regression

Geometric interpretation of a model



"True" boundary may not be closely approximated by a linear boundary

Data Mining: Terminology

Regression modeling (rather than classification modeling)

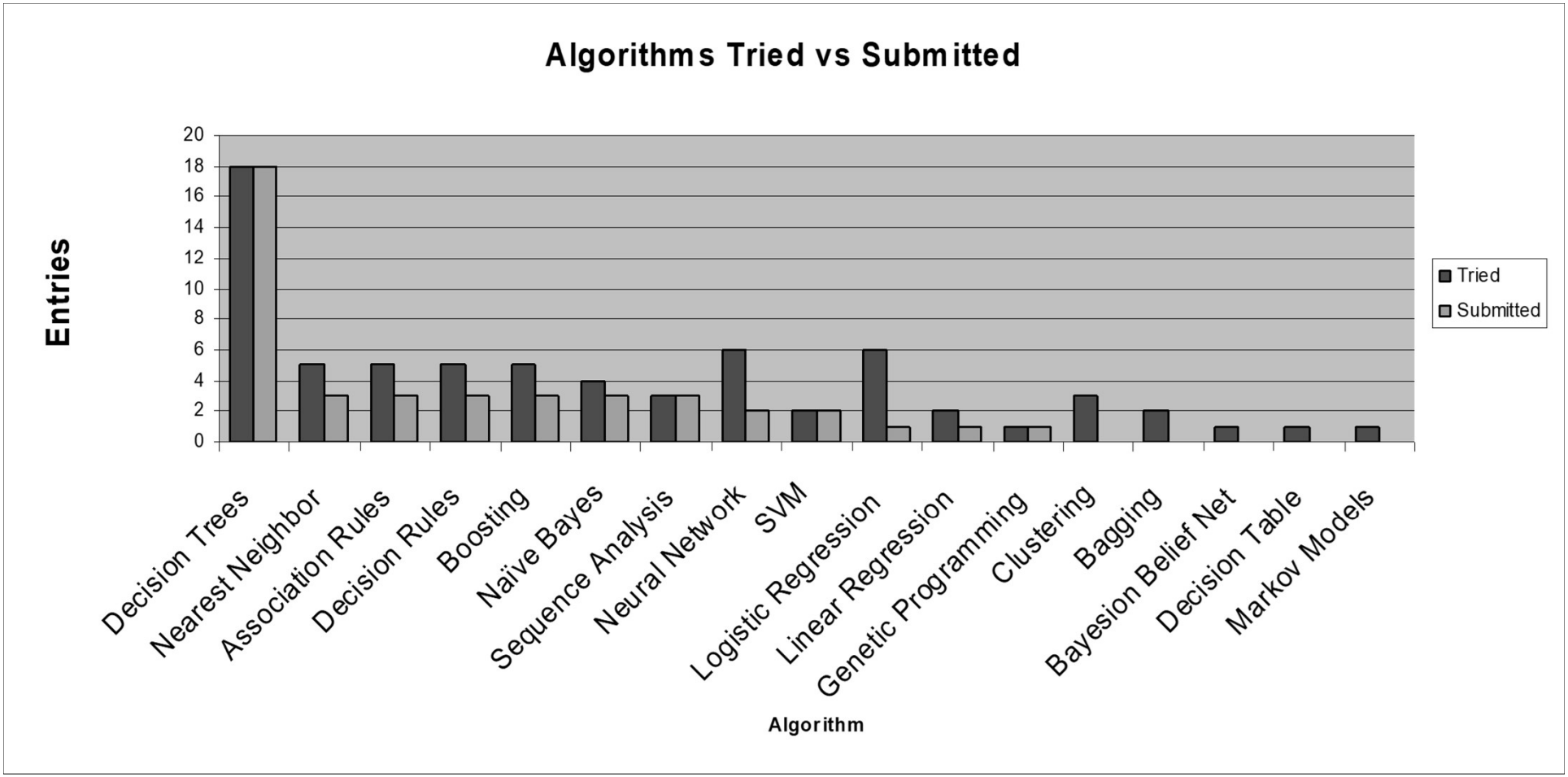
Target Variable

Name	Income	Age	Order $ Amount
Mike	123,000	30	183
Mary	51,100	40	131
Bill	68,000	55	178
Jim	74,000	46	166
Mark	23,000	47	117
Anne	100,000	49	198

Learner: Linear Regression

Model:

Amount= 0.001***Income**+2***Age**



Post-mortem analysis of a popular data mining competition

Thanks to Carla Brodley & Ron Kohavi