EMORY UNIVERSITY

GOIZUETA BUSINESS SCHOOL

Introduction to Business Analytics

Vilma Todri
Associate Professor
Goizueta Business School
Emory University
vtodri@emory.edu

Class Syllabus

Introduction to Business Analytics (MSBA)

Week	Date	Topic	Assignments
1	August 23	Course introduction and overview Basic terminology and data objects Predictive modeling framework - Supervised vs. unsupervised methods - Classification vs. numeric prediction	**HW0** Due (**Optional**) (August 27 11:59pm EST)
2	August 28	Fundamentals of classification - Building and evaluating classification models - Technique: Decision Trees - Technique: k Nearest Neighbors (k-NN) Introduction to RapidMiner & scikit-learn - Repeatable analytics tasks and workflows	
2	August 30		
3/4	September 6 September 11	Sept 4 – Labor Day (No Class) Classification, class probability estimation, and ranking - Technique: Logistic Regression - Generalization and the issue of overfitting - Regularization	**HW1** Due (Sept 8 11:59pm EST)

Class Syllabus

4	September 13	In-depth view at classifier performance and evaluation - N-fold cross-validation approach - Advanced evaluation metrics - Visualization of predictive performance (ROC Curves, etc.)	**HW2** Due (Sept 15 11:59pm EST) **Final Project Proposal** (1 page) (Due Sept 17 11.59pm EST)
5	September 18	Predictive modeling applications using other tools Additional predictive modeling applications - Case studies	
5	September 20	**MIDTERM EXAM**	**HW3** Due (Due Sept 22 11:59pm EST)
6	September 25 September 27	Fundamentals of numeric prediction - Technique: Linear Regression, Lasso Regression, Ridge Regression - Technique: k-NN and combining functions - Technique: Regression Trees	

Class Syllabus

7	October 2	Unsupervised predictive analytics - Technique: Clustering Final Exam Review	**HW4** Due (Due Sept 29 11:59pm EST; if time permits)
7	October 4 October 6	October 4 (No Class) October 6: Class Project Presentations	**Final Project Deliverables** (Due Oct 5 11:59pm EST)
8	October 11	**FINAL EXAM**	

Note: The schedule is tentative and the list of topics may be adjusted over the course of the term. Even though we will cover all of the topics, the pace of learning/discovery will dictate the actual schedule. The changes (if any) will be indicated on the course Canvas website.

Course Overview: Instructor

- Professor: Prof. Vilma Todri, PhD
  - Graduated from **New York University** (Stern School of Business)
    - PhD in Information Systems
  - **Industry Experience**:
    - Google, Data Scientist and Strategist
    - Toyota, Business Analyst
    - Co-founder to a tech start-up that introduced a new business model in the market and earned angel investors' funding
  - **Research Expertise**:
    - Topics: Digital advertising, Social Media, IoT, Effects of technology
    - Methods: Data Analytics, Machine Learning and Experimental Designs
- Contact Information:
  - **Emails**: [Please address your e-mails to the **TAs** of the course cc-ing me. Use "ISOM 672" in subject line.]
  - Office hours: Please Check Canvas for Detailed Schedule

Course Overview: Teaching Assistant

Please check the Canvas website for the detailed schedule of office hours.

- Office Hours every day – please see Canvas schedule

Teaching Assistants:

- Chen Tian (chen.tian@emory.edu)
- Jeffrey de Groot (jeffrey.de.groot@emory.edu)
- Cassie Srb (cassie.srb@emory.edu)
- Ragip Gurlek (rgurlek@emory.edu)

Course Overview: Student Evaluation

- **Participation and Class Contribution** (10%)
  - Both in-class and online
- **Homeworks** (15%)
  - ~ 3-4 team assignments
- **Group Project** (15%)
  - Student presentations
- **Midterm Exam** (25%)
  - In-class, closed-book and closed-notes
- **Final Exam** (35%)
  - In-class, closed-book/closed-notes

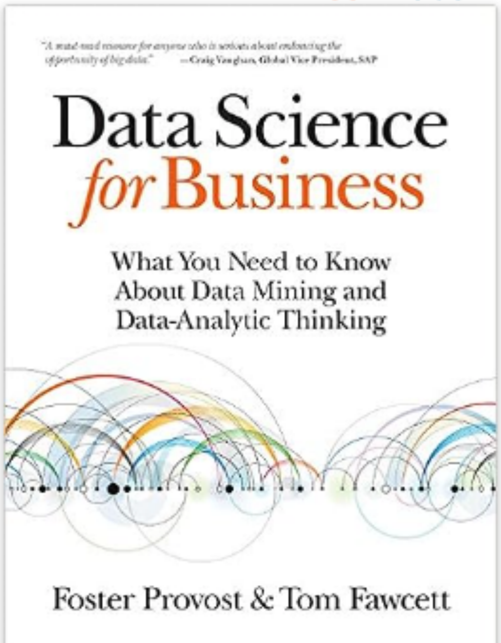
Course Overview: Textbooks

- **Textbooks**
  - **Required book**: Foster Provost, Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013. ISBN-10: 1449361323, ISBN-13: 978-1449361327.
  - **Optional but recommended**: Jiawei Han, Micheline Kamber, Jian Pei. *Data Mining: Concepts and Techniques* (3rd edition). Morgan Kaufmann, 2011. ISBN-10: 0123814790, ISBN-13: 978-0123814791.

- Some occasional online readings

Required Textbook

Look inside



Listen

See this image

Follow the Authors

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking 1st Edition

by Foster Provost (Author), Tom Fawcett (Author)

4.5 1,254 ratings

4.1 on Goodreads 2,322 ratings

See all formats and editions

Written by renowned data science experts Foster Provost and Tom Fawcett, Data Science for Business introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining techniques in use today.

Based on an MBA course Provost has taught at New York University over the past ten years, *Data Science for Business* provides examples of real-world business problems to illustrate these principles. You'll not only learn how to improve communication between business

Read more

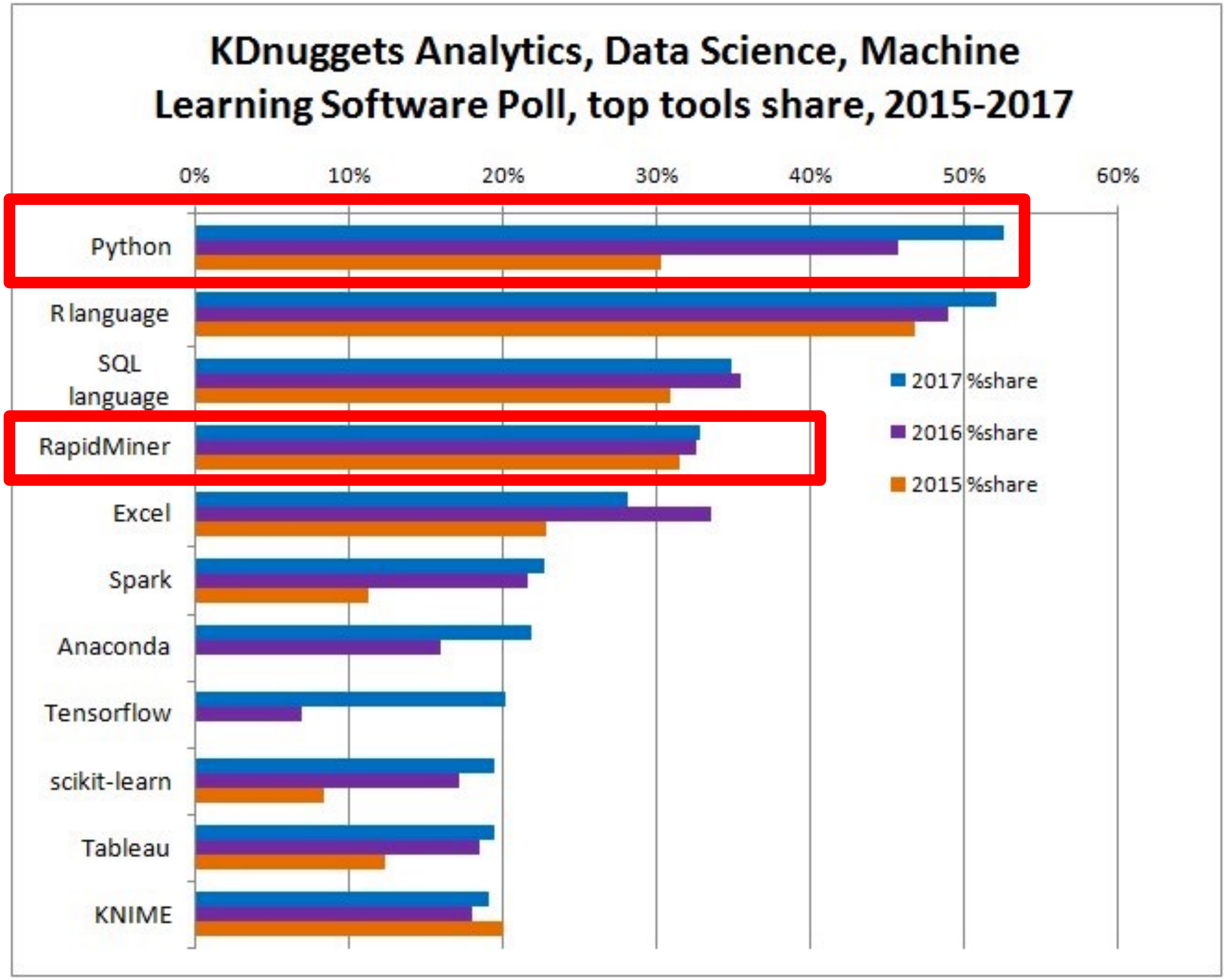
Report incorrect product information.

ISBN-10	ISBN-13	Edition	Publisher
1449361323	978-1449361327	1st	O'Reilly Media

Course Overview: Software

- Throughout the course:
  - **RapidMiner Studio**
    - *ACTION REQUIRED*: Please **install** the academic version of RapidMiner ASAP (link provided on Canvas)
  - **Python**
    - Recommended distribution: Python Anaconda
  - **Jupyter**
    - A web-based interactive computational environment for creating Jupyter notebooks documents.
- Few sessions of the course:
  - Examples of some other tools as well, if time permits
- Important to have your laptop in class for hands-on exercises

Analytics Software Popularity



Source: www.kdnuggets.com

Course Overview: Group Projects

- You will mine actual data for a problem of interest
  - You will mine the data given a predictive task and describe the results

- Deliverables:
  - Project update
  - Final write-up

- Evaluation criteria:
  - Business understanding
  - Data understanding
  - Data preparation
  - Modeling
  - Evaluation
  - Deployment

Course Overview: Communication Platforms

- Course Website
  - Canvas platform

- Emails
  - **“ISOM 672”** as part of the **subject line** in your email messages will ensure prompter response
  - For questions related to the course materials please **address** them to the **TAs cc’ing me**

Feedback and Course Personalization

- General Feedback:
  - Email: **vtodri@emory.edu**

- Anonymous Feedback:
  - https://goo.gl/forms/nRts106ysvvkLmEz2

A few words about yourself…

1. Name or preferred nickname?
2. Educational background and/or Industry Experience?
3. How did you get interested in Business Analytics?
4. What are your aspirations after the MSBA program?
   a) Are there any particular <u>industries</u> you would like to apply your business analytics skills?

Flipped Classroom Approach

- **Personalized Pace:** Students learn at their own pace.
- **Flexibility:** Students choose when and where they want to learn.
- **Active In-Class Learning:** Classroom time is used for interactive activities, promoting deeper understanding.
- **Immediate Feedback:** Opportunities to clarify doubts and receive feedback during in-person sessions.
- **Increased Engagement:** More meaningful discussions and peer interactions.

Team-based Learning

- Parts B of homework assignments and the class project require collaborative work (team submissions)
  - **Collaborative Learning:** Sharing knowledge & techniques enhances overall understanding.
  - **Peer Debugging:** Teammates can help spot and fix errors quickly.
  - **Diverse Skill Sets:** Combining individual strengths leads to more robust solutions.
  - **Improved Communication:** Explaining code & logic to teammates hones communication skills.
  - **Real-world Simulation:** Most coding and data science projects in the industry involve teamwork.
- Declare your team of 5 students on Canvas!

THE MAGAZINE
October 2012

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Back in the 1990s, computer engineer and Wall Street "quant" were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for these experts has raced well ahead of supply. Indeed, Greylock Partners, the VC firm that backed Facebook and LinkedIn, is so worried about the shortage of data scientists that it has a recruiting team dedicated to channeling them to the businesses in its portfolio.

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions. They find the story buried in the data and communicate it. And they don't just deliver reports: They get at the questions at the heart of problems and devise creative approaches to them. One data scientist who was studying a fraud problem, for example, realized it was analogous to a type of DNA sequencing problem. Bringing those disparate worlds together, he crafted a solution that dramatically reduced fraud losses.

TOP MAGAZINE ARTICLES

24 HOURS | 7 DAYS | 30 DAYS

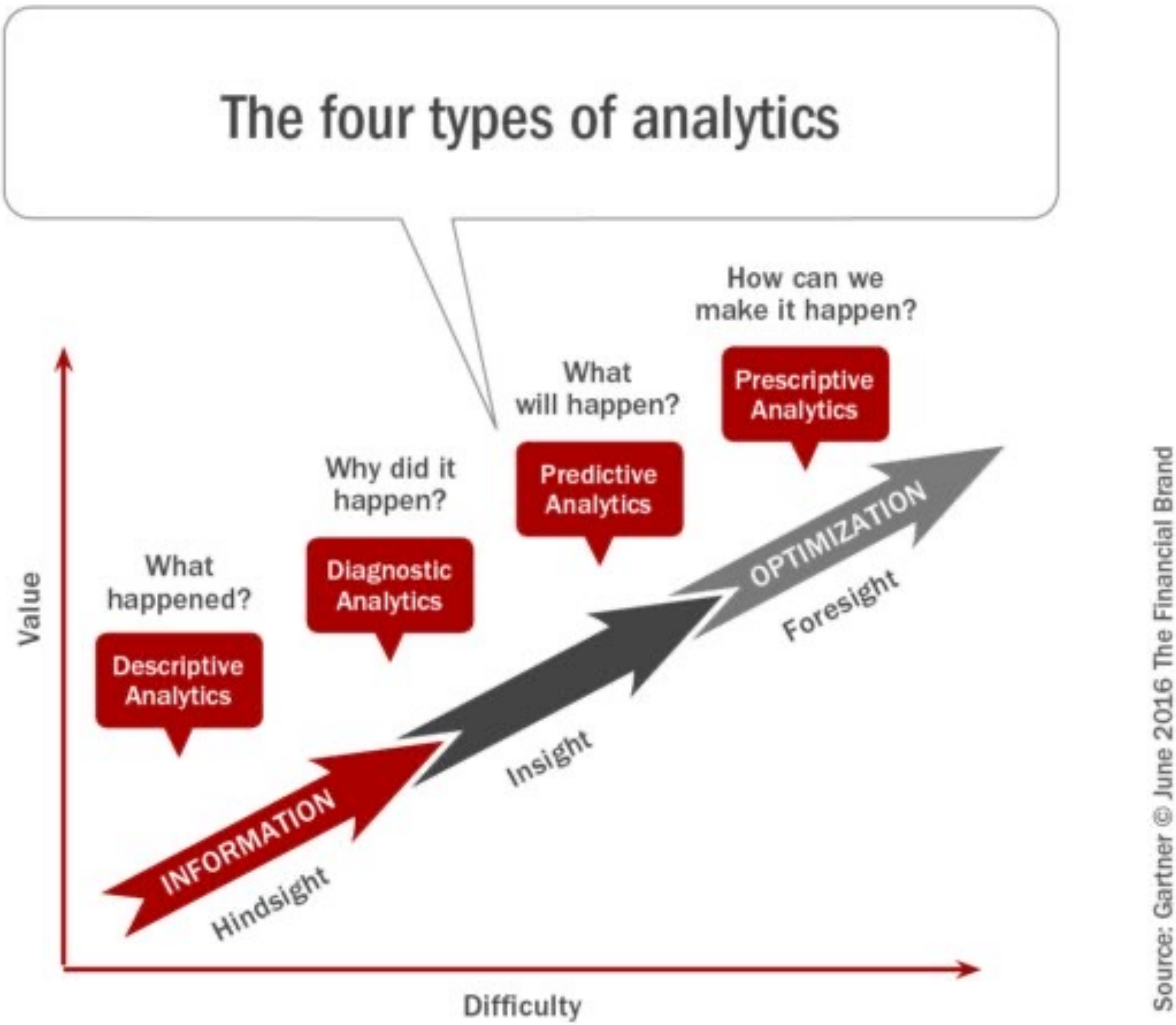
1. Lean Knowledge Work
2. How Netflix Reinvented HR
3. The Five Competitive Forces That Shape Strategy
4. The Big Lie of Strategic Planning
5. Smart Rules: Six Ways to Get People to Solve Problems Without You
6. Find the Coaching in Criticism
7. Salman Khan

All Most Popular »

HBR.ORG ON FACEBOOK

Great Leaders Don't Need Experience

Predictive Analytics: Difficult but Valuable



Data Opportunities

- **Volume** of data
- **Variety** of data
- Powerful **computers**
- **Better algorithms**
  - Traditional (statistical) techniques might not be viable
- These factors have given rise to the widespread applications of **data science principles** and **data-mining techniques**.

Creating Value with Data

- Making information **transparent** and **usable** at much higher frequency
  - Real-time decision-making
- Introducing **sophisticated** analytics
  - Improved decision-making
- Narrowing **segmentation** of customers
  - Much more precisely tailored products or service
- Improving the development of the **next generation** of products and services
- ..

Moneyball: The Competitive Advantage of Data

Roles in Predictive Analytics

- "Data Scientist" (Geek?)
  - can do the actual **modeling**
  - applied statistician X computer scientist
- Collaborator in a data-centric project
  - can **translate** from business to the execution
- Managing a data-mining project
  - understanding the **potential**
  - ability to evaluate a **proposal** and **execution**
  - ability to interface with a broad variety of people
- Strategist, Investor, …
  - envision opportunities, come up with novel ideas, design data science projects/companies conceptually
  - evaluate the promise of new ideas

Learning Goals

- Approach business **problems data-analytically**
- Interact competently on the topic of predictive analytics for business intelligence (fundamental **principles and techniques**)
- **Hands-on experience** data science

Predictive Analytics \ Data Mining \ Machine Learning

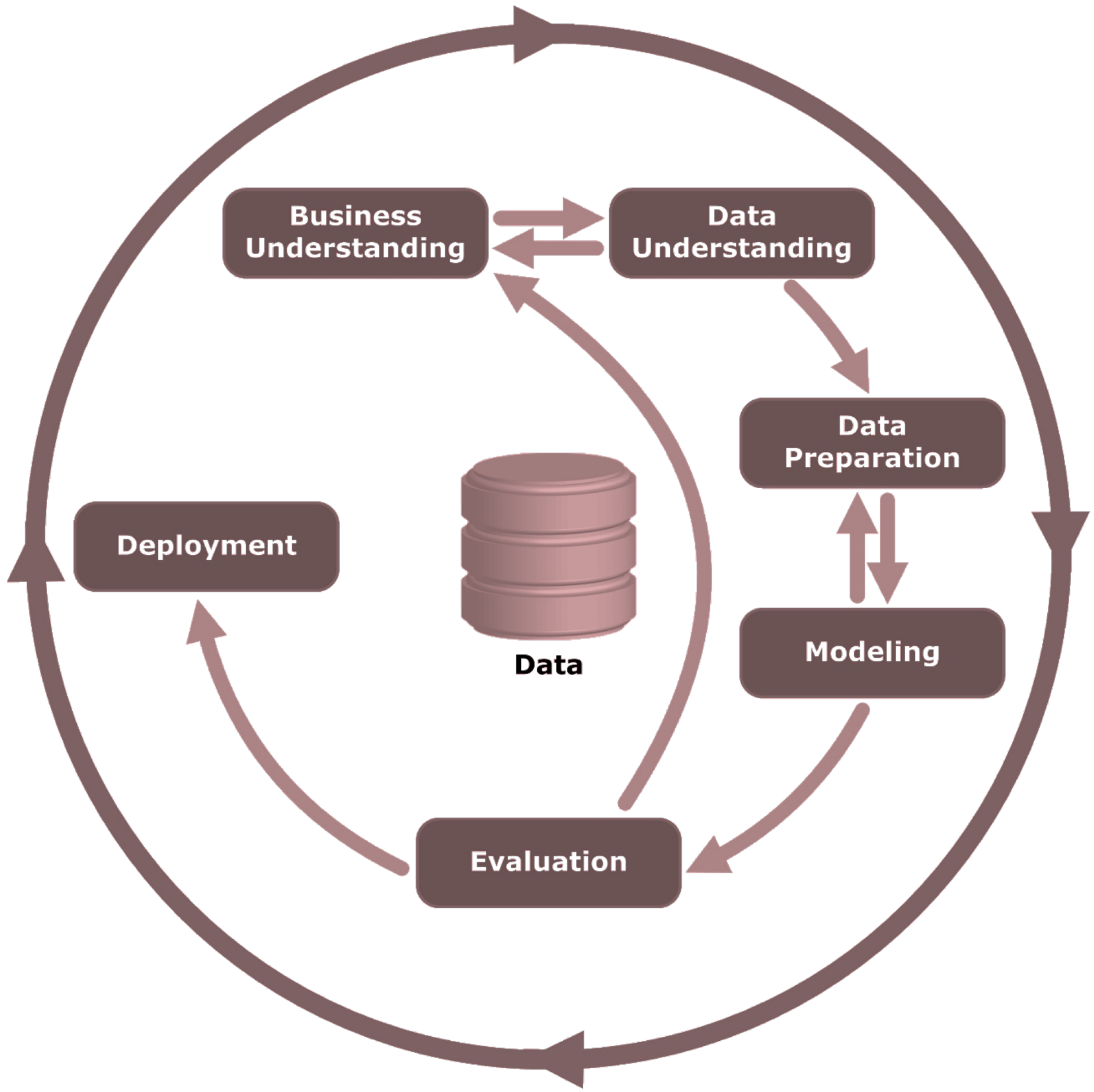
- A set of **principles**, **concepts**, and **techniques** that **structure** thinking and **analysis** of **data**
- Extracts useful **information** and **knowledge** from large volumes of data by following a process with reasonably **well-defined steps**
- Changes the way you **think about data** and its role in business

Business data mining is a process..

science + craft + creativity + common sense

a *process*

Data Mining Process



Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Python Exercise

- Dataset: https://bit.ly/44g60sx

- Using Python perform the following tasks:
  1. Read the data and store it
  2. Print the first 5 observations of the data set
  3. Print the number of rows AND the number of columns in the data set
  4. What is the most ordered item in the dataset and how many times was this item ordered?
  5. What is the total revenue the firm earned?
  6. What is the average amount per order?
  7. How many distinct items were ordered in the data set?

Data Mining versus…

- Data Warehousing / Storage
  - Data warehouses coalesce **data** from **across** an **enterprise**, often from multiple transaction-processing systems
  - Hadoop, Voldemore, Cassandra, etc.

- Querying / Reporting (SQL, Excel, QBE, other GUI-based querying)
  - Very flexible interface to ask **factual questions** about data
  - **No modeling** or sophisticated pattern finding
  - Most of the cool visualizations

- OLAP – On-line Analytical Processing
  - OLAP provides easy-to-use GUI to explore large data collections
  - **Exploration** is <u>**manual**</u>; no modeling
  - Dimensions of analysis preprogrammed into OLAP system

Data Mining versus…

- Traditional **statistical analysis**
  - Mainly based on hypothesis testing or estimation/quantification of uncertainty
  - Should be used to follow-up on data mining's hypothesis generation
- **Automated statistical modeling** (e.g., advanced regression)
  - This is data mining; one type -- usually based on linear models
  - Massive databases allow non-linear alternatives

Answering Business Questions with Such Techniques

- Q: Who are the most profitable customers?
  - **Database querying**
- Q: Is there really a difference between profitable customers and the average customer?
  - **Statistical hypothesis testing**
- Q: But who really are these customers? Can I characterize them?
  - **OLAP** (manual search), **Data mining** (automated pattern finding)
- Q: Will some particular new customer be profitable? How much revenue should I expect this customer to generate?
  - **Data mining** (predictive modeling)

What is a model?

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

**based on some assumptions*

- *Examples: map, engineering prototype, etc.*
- *Data Mining Example:*

  "formula" for predicting probability of customer attrition at contract expiration → "classification model" or "class-probability estimation model"

Terminology

- Model
  - A simplified representation of reality created to serve a purpose
- Predictive Model
  - A formula for estimating the unknown value of interest: **the target**
    - The formula can be mathematical, logical statement (e.g. rule), etc.
- Prediction
  - Estimate an unknown value (i.e. the target)
- Instance / example
  - Represents a fact or a data point
  - Described by a set of **attributes** (fields, columns, variables, or features)
- Model induction
  - The creation of **models** from data
- Training data
  - The input data for the induction algorithm

Terminology

Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

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 (-1)
- Food for thought: Names, Ratings, SIC

Dimensionality of 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 <u>dataset</u>** is the sum of the dimensions of the features
  - the sum of the number of numeric features and the number of values of categorical features

Supervised vs Unsupervised Methods

- "Do our customers naturally fall into different groups?"
  - **No specific purpose or target** specified
  - No guarantee that the results are meaningful / useful

UNSUPERVISED

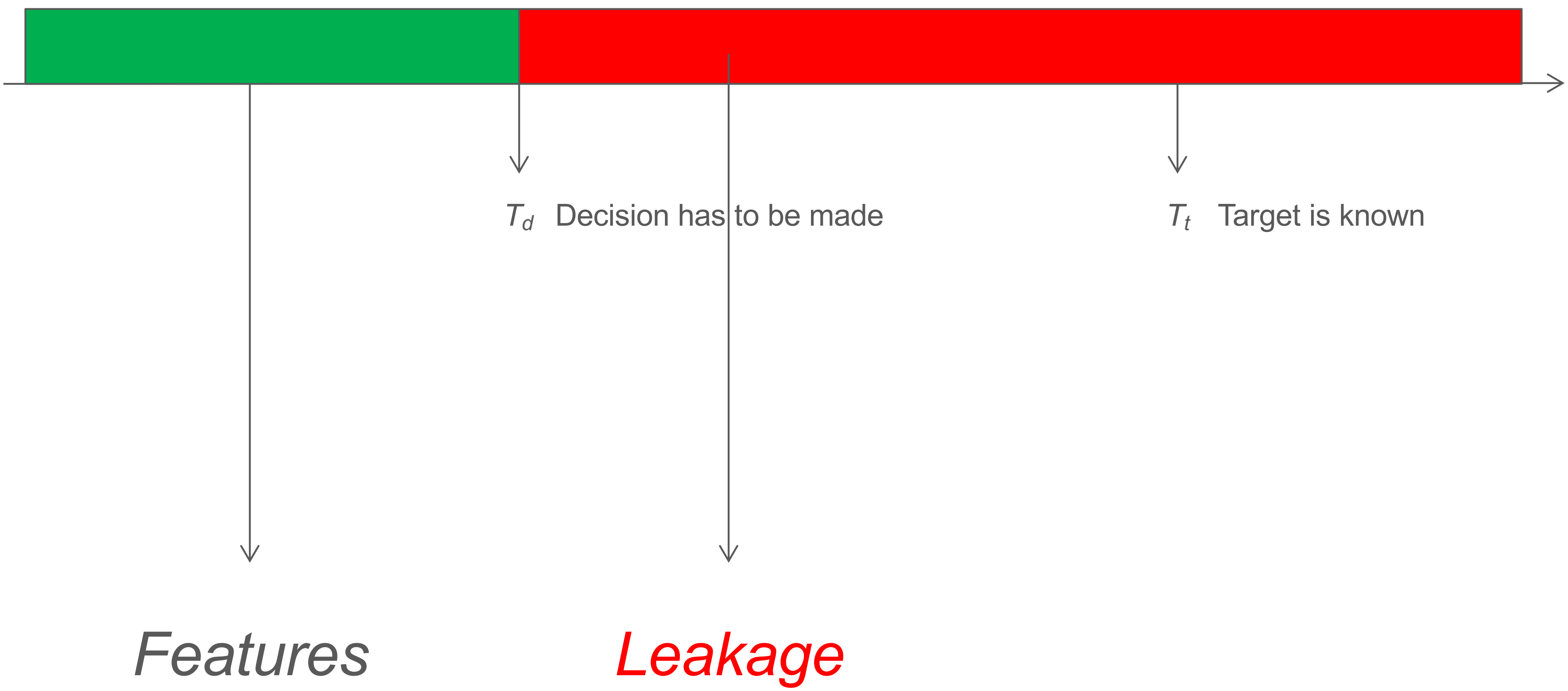
- "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after contracts expire?"
  - **A specific purpose**
  - Much more useful results (usually)
  - Different techniques
  - **Requires data on the target**
    - The individual's label

SUPERVISED

Supervised Data Mining & Predictive Modeling

- Is there a specific, quantifiable **target** that we are interested in or trying to predict?
  - think about the decision
- Do we have **data** on this target?
  - Do we have enough data on this target?
    - Need a min ~500 of each type of classification
- Do we have relevant data prior to decision?
  - think timing of decision and action
- The result of supervised data mining is a model that predicts some quantity
- A model can either be used to predict or to understand

Digression on features: It is all about the timing in use!

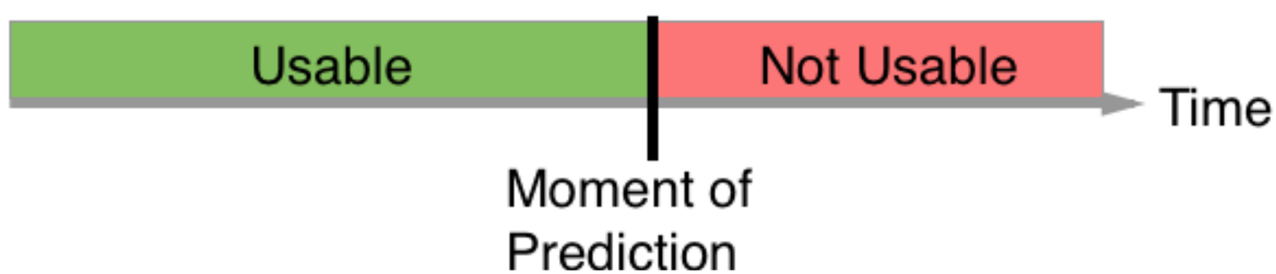


Data Leakage: Example

- Imagine you want to predict who will get sick with pneumonia. The top few rows of your raw data might look like this:

got_pneumonia	age	weight	male	took_antibiotic_medicine	...
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

- People take antibiotic medicines **<u>after</u>** getting pneumonia in order to recover. Using the feature “took_antibiotic_medicine” to predict “got_pneumonia” will lead to data leakage



Leakage

- If any other feature whose value **would not actually be available in practice** at the time you'd want to use the model to make a prediction, is a feature that can introduce leakage to your model
- Data Leakage allows a model or machine learning algorithm to make ***unrealistically good*** predictions.
- It's hard because we cannot evaluate the model on something we don't have.
- Some types of data leakage include:
  - Leaking of information from the future into the past.
  - Leaking test data / ground truth into the training data.

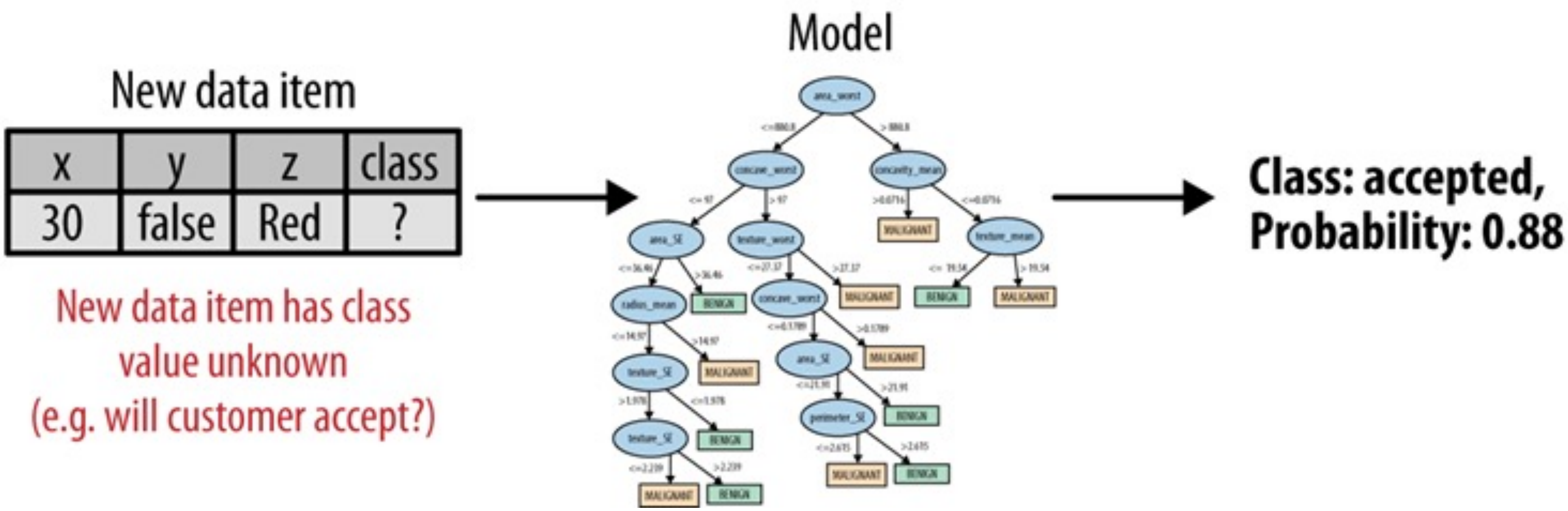
Subclasses of Supervised Data Mining

- Subclasses of supervised data mining are distinguished by the type of target.
- Classification
  - Categorical target
    - Often binary
  - Includes “class probability estimation”
  - e.g., How likely is this consumer to respond to our campaign?
- Regression
  - Numeric target
  - e.g., How much will this customer use the service?

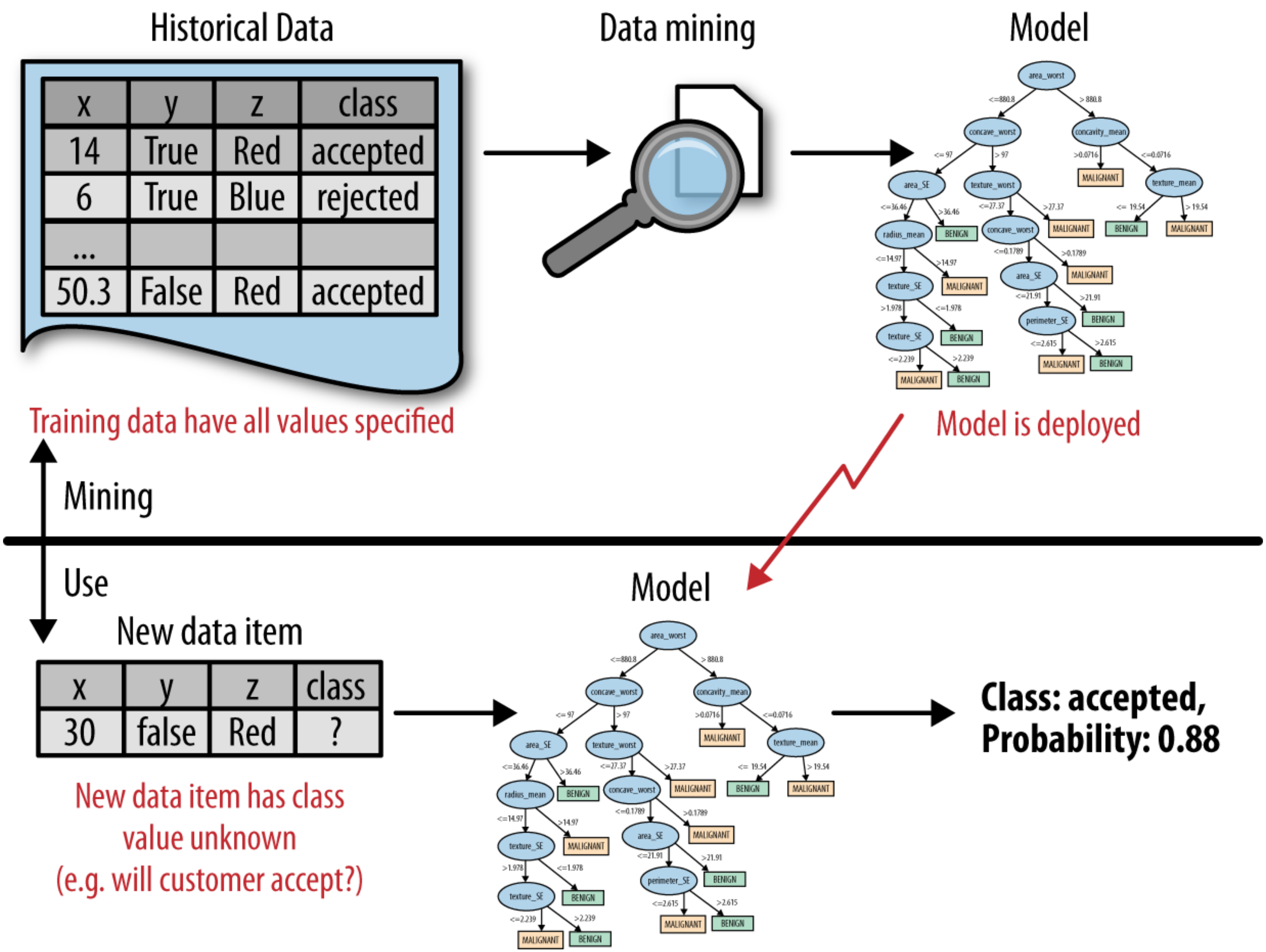
Subclasses of Supervised Data Mining

- "Will this customer purchase service $S1$ if given incentive $I1$?"
  - Classification problem
    - Binary target (the customer either purchases or does not)
- "Which service package ($S1$, $S2$, or none) will a customer likely purchase if given incentive $I1$?"
  - Classification problem
    - Three-valued target
- "How much will this customer use the service?"
  - Regression problem
    - Numeric target
    - Target variable: amount of usage per customer

Data Mining versus Use of the Model



Data Mining versus Use of the Model



Thank you!

Questions?