



# Data Science and Business Intelligence

BU.330.780

Session 6

Instructor: Changmi Jung, Ph.D.

# Announcement



» Next week: Assignment 4 is due (under Canvas > Week 7)

» Quiz #2

- Onsite test: closed-book, closed-notes
- Covers week 4 through week 6: from CT to what we learn today (No Text-mining)
- Format: up to 25 multiple-choice questions (including True/False)
- Logistics: first 40-45 minutes of the class. After submission, you may leave the classroom and come back to join the class within the given amount of time
- Sample questions: Week 7 > “Quiz 2 Study Guide and Sample Questions.pdf”



# Probability Estimation to Classification: Classification Threshold and Expected Value Framework



# Cutoff (Threshold) for classification

- » Most ML algorithms classify individuals via a 2-step process
- » For each record,
  1. Compute the probability of belonging to class “1”
  2. Compare to the cutoff value (threshold) and classify accordingly
- » The default cutoff value is 0.50
  - If predicted probability  $\geq 0.5$ , classified as “1” – positive
  - If predicted probability  $< 0.5$ , classified as “0” – negative
- » Can use different cutoff values (arbitrary values, traditionally used values, or what seems to be the best from the ROC curve)

# Cutoff Table



| Actual Class | P of being "1" | Actual Class | P of being "1" |
|--------------|----------------|--------------|----------------|
| 1            | 0.996          | 1            | 0.506          |
| 1            | 0.988          | 0            | 0.471          |
| 1            | 0.984          | 0            | 0.337          |
| 1            | 0.980          | 1            | 0.218          |
| 1            | 0.948          | 0            | 0.199          |
| 1            | 0.889          | 0            | 0.149          |
| 1            | 0.848          | 0            | 0.048          |
| 0            | 0.762          | 0            | 0.038          |
| 1            | 0.707          | 0            | 0.025          |
| 1            | 0.681          | 0            | 0.022          |
| 1            | 0.656          | 0            | 0.016          |
| 0            | 0.622          | 0            | 0.004          |

24 Records in total

- Respond to a special offer: "1"
- Otherwise: "0"

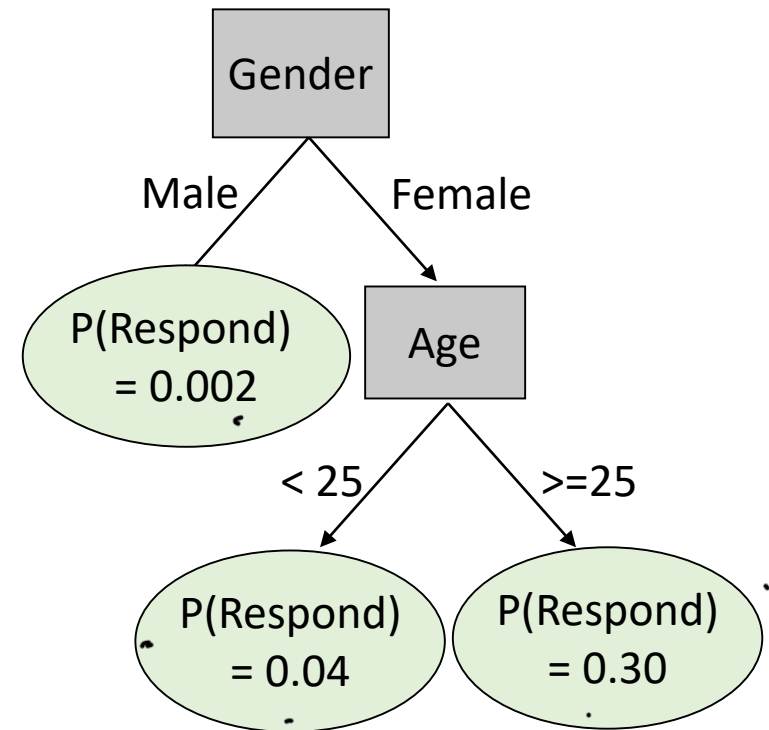
Let's say that this model predicts which customer will likely respond to a special offer. We will send a promotional catalog to the identified customers.

- If the cutoff is 0.50: thirteen records are classified as "1"
- If the cutoff is 0.80: seven records are classified as "1"

# Targeted Marketing Example



- » Consider a retailer that sells their product only through catalogs
- » Assign each customer to either a class “likely responder” or “unlikely responder” for an offer for a product that is only available via this offer.



None of the leaves has more than 50% chance to respond!





# Expected value calculation

» Expected value calculation includes enumeration of the possible outcomes of a situation: the outcomes are mutually exclusive and collectively exhaustive.

» General form of expected value computation

$$EV = p(o_1) \cdot v(o_1) + p(o_2) \cdot v(o_2) \dots$$

where  $o_i$  is a possible decision outcome:

ex.  $o_R$  (Responding),  $o_{NR}$  (Not responding)

$p(o_i)$  is its probability

$v(o_i)$  as its value



Let's relate this to the previous targeted marketing example →



# Using expected value to frame classifier use (1/2): Targeted Marketing Example

- » Possible outcomes – Respond (R) or Not Respond (NR)
- » Expected benefit (value) of targeting
$$p_R(x) \cdot v_R + [1 - p_R(x)] \cdot v_{NR}$$
where  $p_R$  is an estimated probability of response  
 $v_R$  is the value we get from a response  
 $v_{NR}$  is the value we get from no response
- » Price of product: \$200, costs of product: \$100 ✓
- » Cost of targeting a customer: \$2 (marketing materials + shipping)
$$v_R = \$200 - \$100 - \$2 = \$98$$
$$v_{NR} = -\$2$$
- » Expected value of targeting =  $p_R(x) \cdot \$98 + [1 - p_R(x)] \cdot (-2)$
- » Expected benefit of not targeting = 0 in this case.






# Using expected value to frame classifier use (2/2): Targeted Marketing Example

» For which customer do we make a profit?

Expected value of targeting > Expected value of not targeting


$$p_R(x) \cdot v_R + [1 - p_R(x)] \cdot v_{NR} > 0$$

$$p_R(x) \cdot \$98 + [1 - p_R(x)] \cdot (-2) > 0$$

$$p_R > 0.02$$

We should target the consumer as long as the estimated probability of responding is greater than 2%!



# Text Mining

*Text Mining will not be on the Quiz 2*

# Text data



» Until now we have been dealing with structured data

- Binary (yes/no)
- Numerical
- Multi-category

| 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        |

*We are clearly defined! You know what 'age' means, right?*

» Now we turn to unstructured data – text!



# What exactly is text mining?

- » Case 1: Extract meaning from a single document – interpreting it like a human reads language
  - “Natural language processing” (not predictive modeling, not our focus)
  
- » Case 2: Classify (label) thousands of documents ← **Our Focus**
  - Extension of predictive modeling
  - No attempt to extract overall meaning from a single document
  - Focus is extracting useful features to be used to classify numerous documents



# Text Mining Example (I)

- » Insurance fraud – notes in claim forms can be mined and transformed into predictor variables for a predictive model
- » The model is trained on prior claims in two classes – found to be fraudulent, and not found to be fraudulent
- » The model is then applied to new claims



## CLAIM FORM AND INSTRUCTIONS

If you have any questions regarding benefits available, or how to file your claim, or if you would like to appeal any determination, please contact our Customer Care Center at 1-800-348-4489, 8:00 A.M. to 8:00 P.M. Eastern Standard Time

The furnishing of this form, or its acceptance by the Company as proof, must not be construed as an admission of any liability on the part of the Company, nor a waiver of any of the conditions of the insurance contract.

### INSTRUCTIONS FOR FILING YOUR GROUP ACCIDENT CLAIM

# Text Mining Example (II)



» Clinics could use patient online appointment request forms to route their requests to appropriate personnel:

- Physician assistant
- Nurse Practitioner
- Doctor

The screenshot shows the 'one MEDICAL GROUP' website. The navigation bar includes links for 'HOW WE'RE DIFFERENT', 'PRIMARY CARE TEAM', 'LOCATIONS', 'INSURANCE', 'MEMBERSHIP', 'HELP', and 'BLOG'. The main heading is 'BOOK A NEW APPOINTMENT'. A location dropdown menu is set to 'Washington, D.C.'. A yellow banner states: 'Can't find an appointment that works for you? Feel free to give us a call at 202-706-7634 and we'll do our best to help.' Below this, the form is divided into four numbered steps: 1. 'I would like to see' (with buttons for 'My Primary Care Team', 'Any Available Provider', and a green 'Specific Provider' link), 2. 'I want to be seen for', 3. 'I want to be seen on', and 4. 'I want to cover'.

# Why Text is Difficult

## » Text is “unstructured”

- Linguistic structure is intended for human communication and not computers

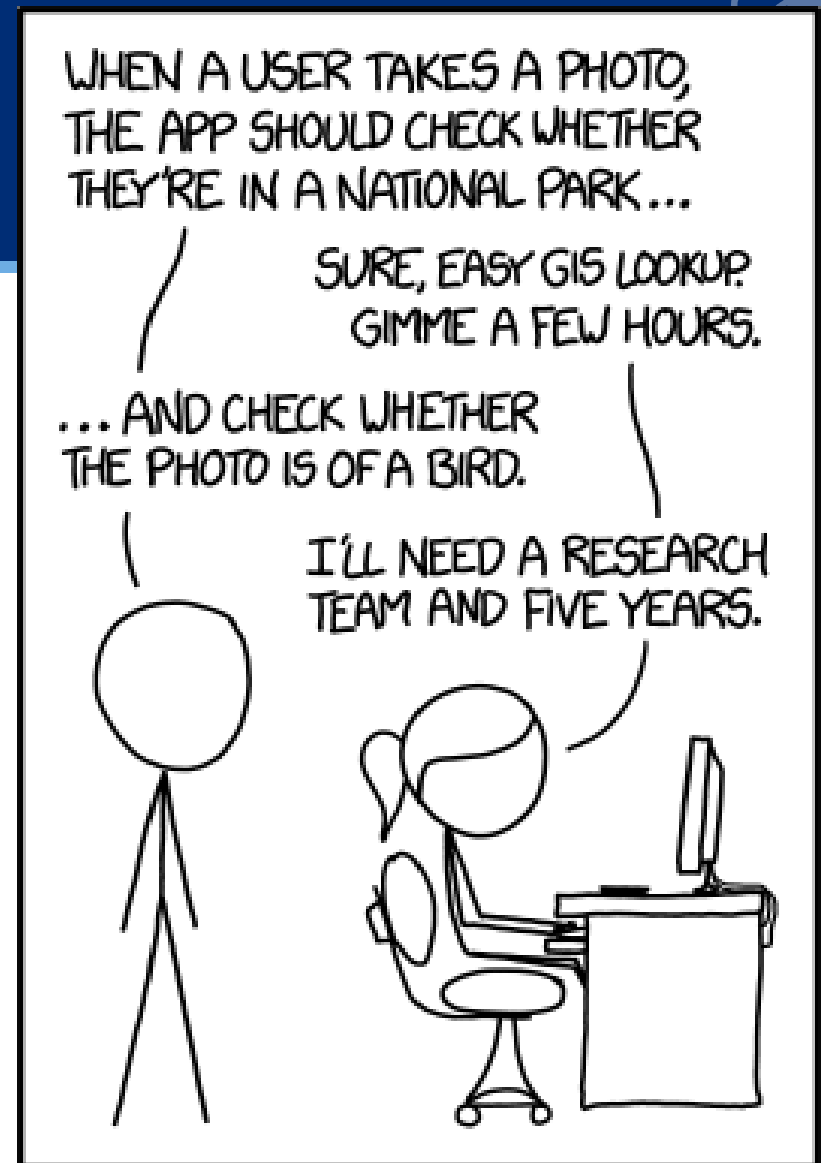
## » Word order matters sometimes

## » Context matters

## » Text can be dirty

- People write ungrammatically, misspell words, abbreviate unpredictably, and punctuate randomly

For example, you need to update your system to understand wdyam, ngl, lol, etc.

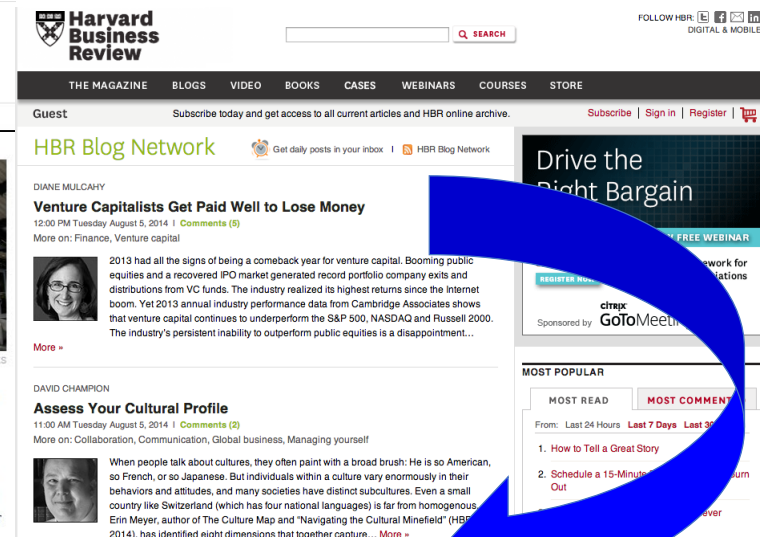


IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

# Text Representation



- » **Goal:** Take a set of documents – each of which is a relatively free-form sequence of words– and turn it into our familiar “feature-vector (table)” form



Amazon Un...  
The results were a se...  
landmark victory at a large...  
23m ago • By KAREN WEISE and NOAM SCHEIBER

| Attribute 1 | Attribute 2 | Attribute 3 | Attribute 4 | Attribute 5 | Attribute 6 |
|-------------|-------------|-------------|-------------|-------------|-------------|
|             |             |             |             |             |             |
|             |             |             |             |             |             |
|             |             |             |             |             |             |





# Get Data from Twitter using Twitter API

## » What is API?

- Application Programming Interface
- A standardized protocols for sending request messages over the internet.
- <https://apps.twitter.com/>
- Useful reference: <https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html>

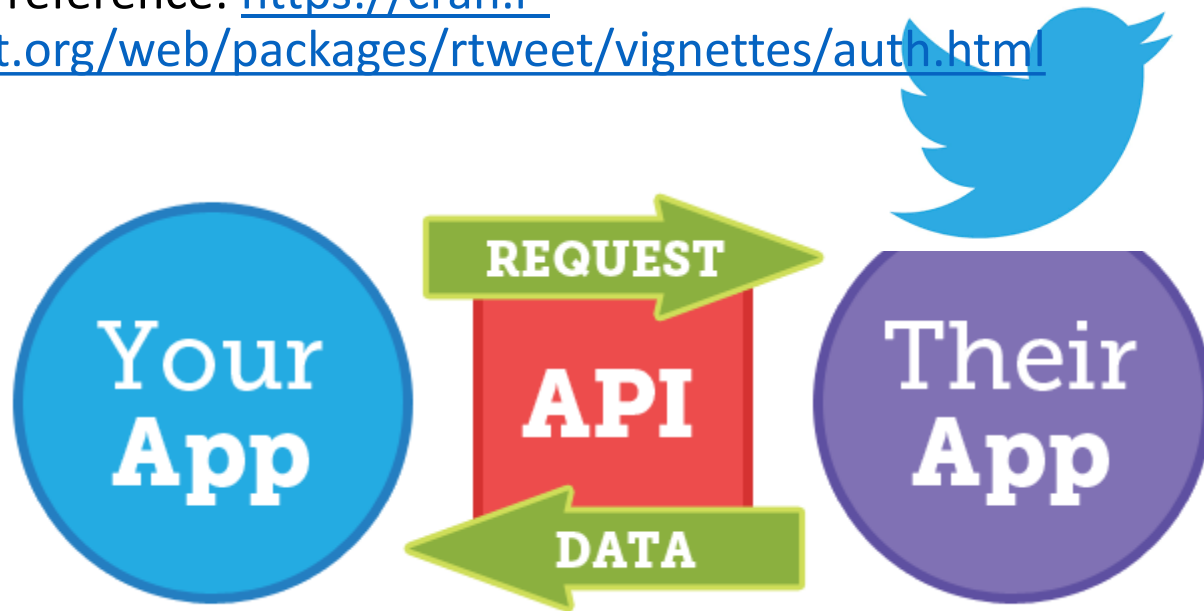


Image from <http://moz.com/blog/apis-for-datadriven-marketers>. This site provides a list of examples of web API.

# Text Mining Example (III) + Text Representation



"The first part of this movie is far better than the second. The acting is poor and it gets out-of-control by the end, but it's till fun to watch."

## » Document

- One piece of text, no matter how large or small.
- It is composed of individual **tokens** or terms (e.g. a word).

## » Corpus: a collection of documents.

## » Each document is one instance

- *but we don't know in advance what the features will be*





# Feature Extraction (1/2) – Bag of words

- » Feature extraction involves converting the text string into some sort of quantifiable measures.
- » Bag of words
  - treats every document as a collection of individual words
  - ignores grammar, word order, sentence structure, and punctuation
  - Treat every word in a document as a potentially important keyword of the document
- Each document is represented by the frequency of tokens (the number of times the word appears in the document)

|           |                                  |
|-----------|----------------------------------|
| <b>d1</b> | jazz music has a swing rhythm    |
| <b>d2</b> | swing is hard to explain         |
| <b>d3</b> | swing rhythm is a natural rhythm |

Simple documentation



|           | a | explain | hard | has | is | jazz | music | natural | rhythm | swing | to |
|-----------|---|---------|------|-----|----|------|-------|---------|--------|-------|----|
| <b>d1</b> | 1 | 0       | 0    | 1   | 0  | 1    | 1     | 0       | 1      | 1     | 0  |
| <b>d2</b> | 0 | 1       | 1    | 0   | 1  | 0    | 0     | 0       | 0      | 1     | 1  |
| <b>d3</b> | 1 | 0       | 0    | 0   | 1  | 0    | 0     | 1       | 2      | 1     | 0  |

Document Term  
Matrix (DTM)



# Feature Extraction (2/2): Lots of things to process ...

*Microsoft Corp and Skype Global today announced that they have entered into a definitive agreement under which Microsoft will acquire Skype, the leading Internet communication company, for \$8.5 billion in cash from the investor group led by Silver Lake. The agreement has been approved by the boards of directors of both Microsoft and Skype.*



| Term    | Count | Term      | Count | Term      | Count | Term     | Count |
|---------|-------|-----------|-------|-----------|-------|----------|-------|
| skype   | 3     | microsoft | 3     | agreement | 2     | global   | 1     |
| approv  | 1     | announc   | 1     | acquir    | 1     | lead     | 1     |
| definit | 1     | lake      | 1     | communic  | 1     | internet | 1     |
| board   | 1     | led       | 1     | director  | 1     | corp     | 1     |
| compani | 1     | investor  | 1     | silver    | 1     | billion  | 1     |

- » We want to remove extraneous information from the text and standardize it into a uniform format
- » Typical **text processing** include
  - Normalizing the case: e.g. Skype, SKYPE, skype → skype
  - Stemming words: e.g. suffixes – announces, announced, announcing
  - Removing stopwords: e.g. a, the, and, of, on
  - Often, numbers are discarded



# Tokenization and basic text analysis with Tidytext

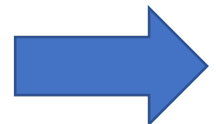
## » Tokenization with `unnest_tokens()`

- Punctuation are stripped
- Words are converted to lower-case

## » Recall: Why do we need these process?

- To reduce the number of text features in the set of documents!

R exercise





# Sentiment Analysis

- » Consider a document as a combination of its individual words (i.e. bag of words)
- » The sentiment of the whole text can be inferred as the sum of the sentiment of the individual words.
- » General-purpose, publicly-available lexicons
  - Constructed via crowdsourcing or by the labor of the authors.
- » While this lexicon-based method has limitations (e.g., loss of information from disregarding sentence structure), it is widely used for its simplicity and shows sufficiently good performance.



# Predictive modeling?

» Now we have a clean, structured dataset similar to what we have used in our numerical data mining:

- Class identifications (labels) for training
- Numerical predictors

*R exercise*





# Course Review & Project Evaluation

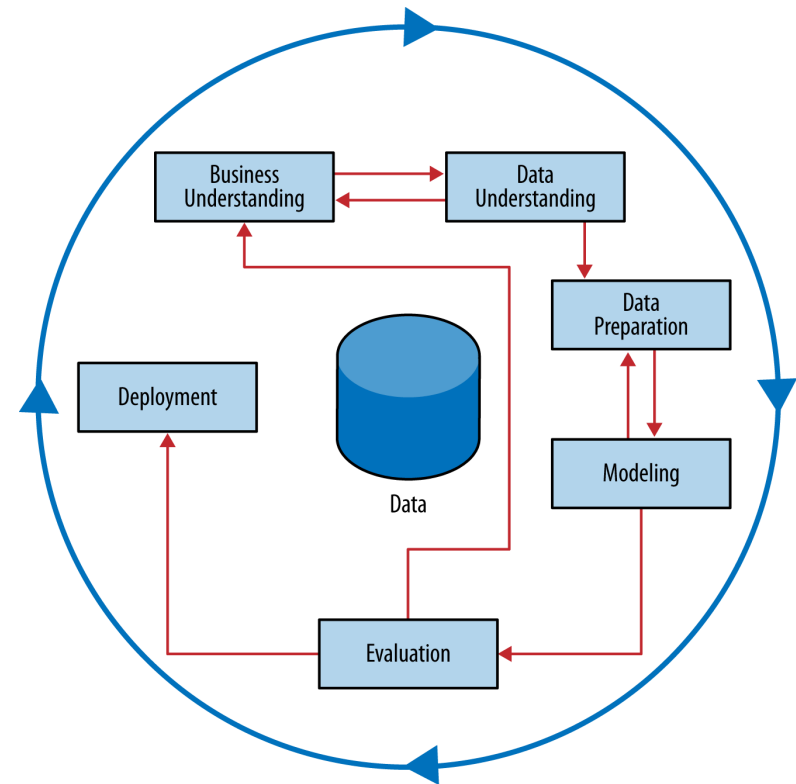


# What is data science?




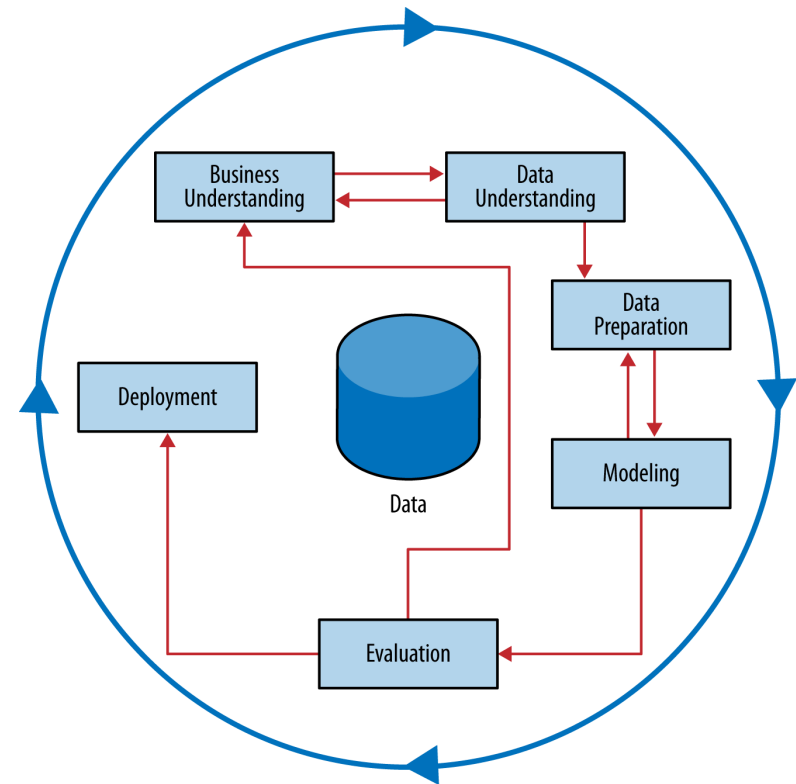
» Data science is a process with well-understood stages based on

- Application of information technology
- Analyst's creativity
- Business knowledge
- Common sense



# Business understanding

- » What exactly is the business problem to be solved?
  - » Do we really need data science?
  - » Is the problem a supervised or unsupervised problem?
  - » If supervised,
    - What is a target variable?
    - Is it defined precisely?





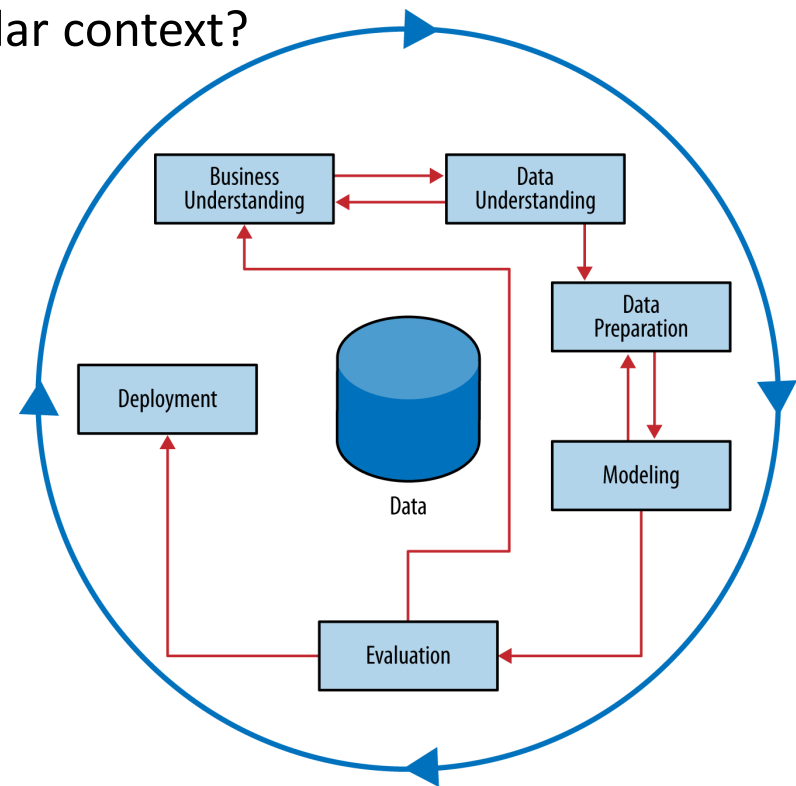
# Data understanding & preparation (1/2)

» Do we have sufficient data for solving the business problem?

- Do we have data on the target variable?
- Are the training data from the same/similar phenomenon?
- Are the training data from the same/similar context?

» Attributes understanding

- Think about the values the attributes can take (e.g. categorical, ordinal, numerical)
- Visualization is often very useful



- 
- The diagram illustrates the CRISP-DM (Cross-Industry Standard Process for Data Mining) process flow. It consists of six main stages arranged in a cycle, connected by a large blue arrow indicating the overall flow. The stages are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. A central blue cylinder labeled 'Data' represents the data source. Red arrows show the flow between stages: Business Understanding and Data Understanding are connected by a double-headed arrow. Data Understanding leads to Data Preparation, which leads to Modeling. Modeling leads to Evaluation, which leads to Deployment. Evaluation also leads back to Business Understanding, completing the cycle. Additionally, a red arrow points from the 'Data' cylinder to the Evaluation stage.

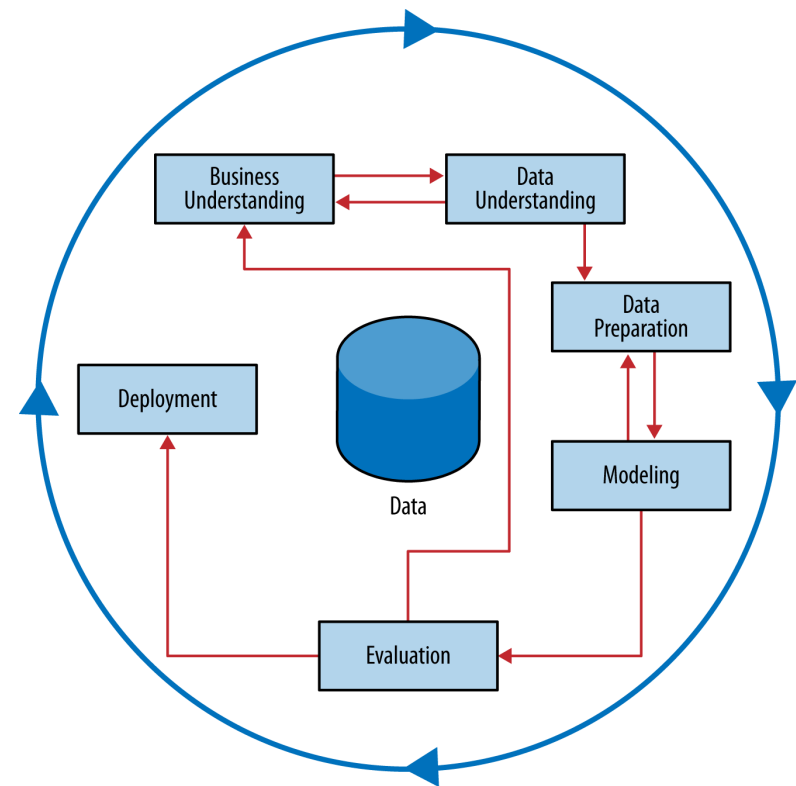
# Modeling



» Is the choice of model appropriate for the choice of target variable?

» Modeling techniques for classification problems

- Tree-based models
  - Classification trees
  - Random forest classifier
- Linear models (a.k.a linear classifier)
  - logit regression
  - SVM





# Recap: Important differences between Tree induction and Linear models

## »A classification tree

- uses decision boundaries that are **perpendicular** to the instance-space axes
- is a “**piecewise**” classifier that segments the instance space recursively → cut in arbitrarily small regions possible

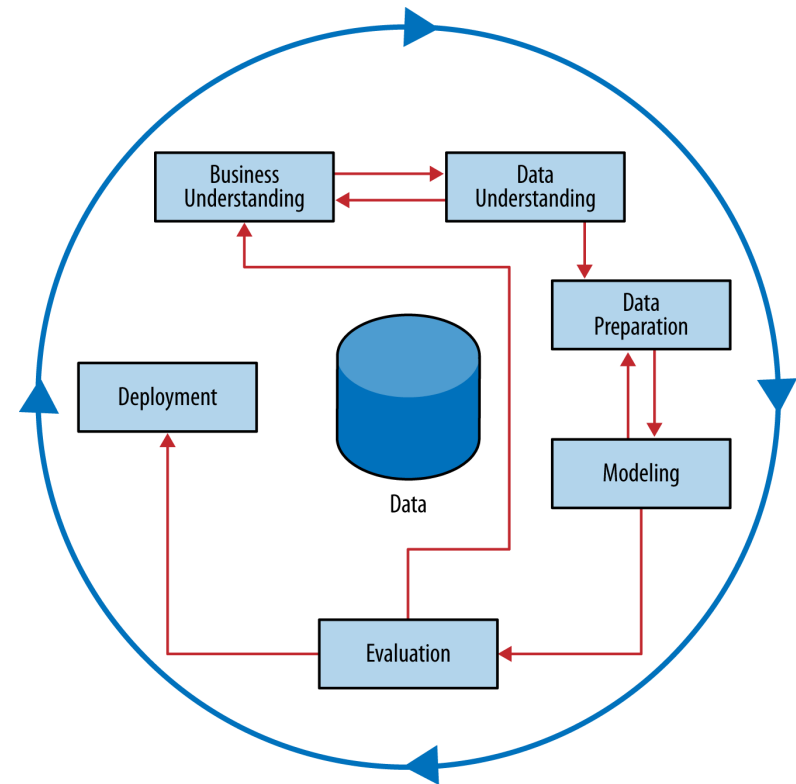
## »A linear classifier

- use decision boundaries of **any direction** or orientation
- places a **single decision** surface through the entire space



# Model Validation & Evaluation (1)

- » Generalization performance Fitting curve
- » Various models might be tried and compared.
- » Hyperparameter tuning
- » Model validation
  - Domain knowledge validation
  - Holdout validation
    - k-fold cross-validation
    - OOB errors (for Random Forest)





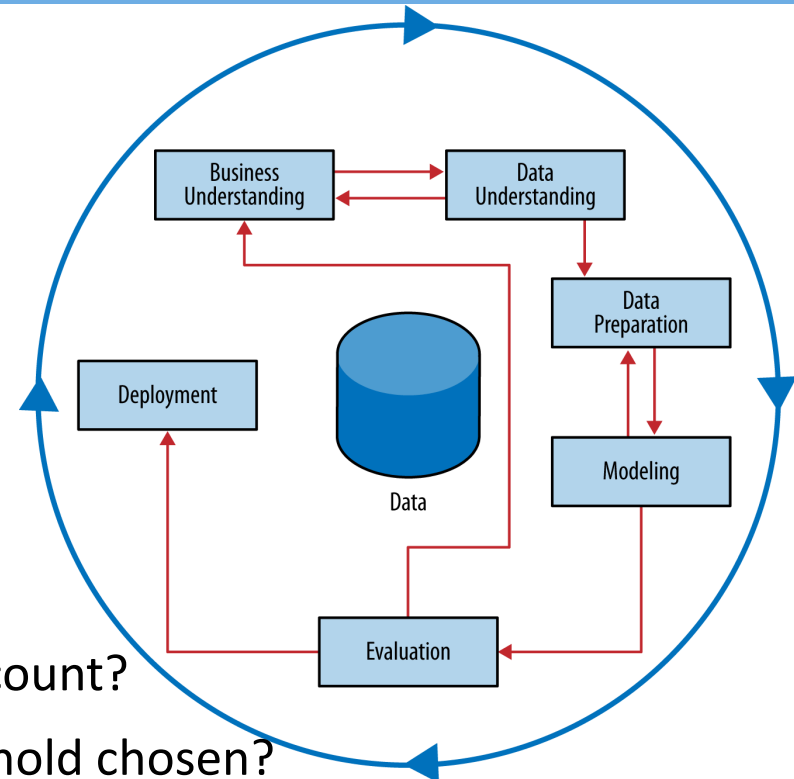
# Model Validation & Evaluation (2)

## » Performance evaluation metrics

- Accuracy
- Confusion matrix
- ROC analysis and AUC

## » Is the evaluation setup and metric appropriate for the business task?

- Are business costs and benefits taken into account?
- For classification, how is a classification threshold chosen?



## » Expected value frameworks for classifier use (i.e. selection of classification threshold)





# Sample Question

» Consider this model result. What is the *false positive rate* of this model when we set the threshold at 0.79. That is, a case is considered (i.e. predicted) as positive if the score is greater than 0.79 and negative otherwise.

Threshold

| True class | Score |
|------------|-------|
| P          | 0.99  |
| P          | 0.98  |
| N          | 0.96  |
| N          | 0.9   |
| P          | 0.88  |
| N          | 0.87  |
| P          | 0.85  |
| P          | 0.8   |
| N          | 0.7   |
| P          | 0.65  |

# Sample Question



- » Consider this model result. What is the false positive rate of this model when we set the threshold at 0.79. That is, a case is considered (i.e. predicted) as positive if the score is greater than 0.79 and negative otherwise.

| True class | Score | Pred. class |
|------------|-------|-------------|
| P          | 0.99  | P           |
| P          | 0.98  | P           |
| N          | 0.96  | P           |
| N          | 0.9   | P           |
| P          | 0.88  | P           |
| N          | 0.87  | P           |
| P          | 0.85  | P           |
| P          | 0.8   | P           |
| N          | 0.7   | N           |
| P          | 0.65  | N           |

|           |   | Actual |   |
|-----------|---|--------|---|
|           |   | +      | - |
| Predicted | Y | 5      | 3 |
|           | N | 1      | 1 |

$$\begin{aligned}\text{False Positive Rate (a.k.a. 1- specificity)} &= \frac{\text{False positive (b)}}{\text{True Negative (d) + False Positive (b)}} \\ &= 3/(3+1) = 75\%\end{aligned}$$

Can you plot this confusion matrix on ROC space?

# Sample Proposal Evaluation - Improve this plan



## **Targeted Whiz-bang Customer Migration --- prepared by Big Red Consulting, Inc.**

We will develop a predictive model using modern data-mining technology. As discussed in our last meeting, we assume a budget of \$5,000,000 for this phase of customer migration; adjusting the plan for other budgets is straightforward. Thus we can target 20,000 customers under this budget. Here is how we will select those customers:

We will use data to build a model of whether or not a customer will migrate given the incentive. The data set will comprise a set of attributes of customers, such as the number and type of prior customer service interactions, level of usage of the widget, location of the customer, estimated technical sophistication, tenure with the firm, and other loyalty indicators, such as number of other firm products and services in use. The target will be whether or not the customer will migrate to the new widget if he/she is given the incentive. Using these data, we will build a linear regression to estimate the target variable. The model will be evaluated based on its accuracy on these data; in particular, we want to ensure that the accuracy is substantially greater than if we targeted randomly.

To use the model: for each customer we will apply the regression model to estimate the target variable. If the estimate is greater than 0.5, we will predict that the customer will migrate; otherwise, we will say the customer will not migrate. We then will select at random 20,000 customers from those predicted to migrate, and these 20,000 will be the recommended targets.