Lecture 1

QM 701: Advanced Data Analytics and Applications

Fuqua School of Business

Fall 2024

Yehua Wei

About Me

- My background:
  - Bachelor of Mathematics @ University of Waterloo
  - Ph.D. in Operations Research @ MIT
  - Worked at Fuqua, Decision Sciences, from 2013 - 16 and 2019 - present
- Teaching and Research Interest: Decision and Data Analytics

Lecture outline

- Course Objectives
- Logistics
- Intro to Sentiment Analysis

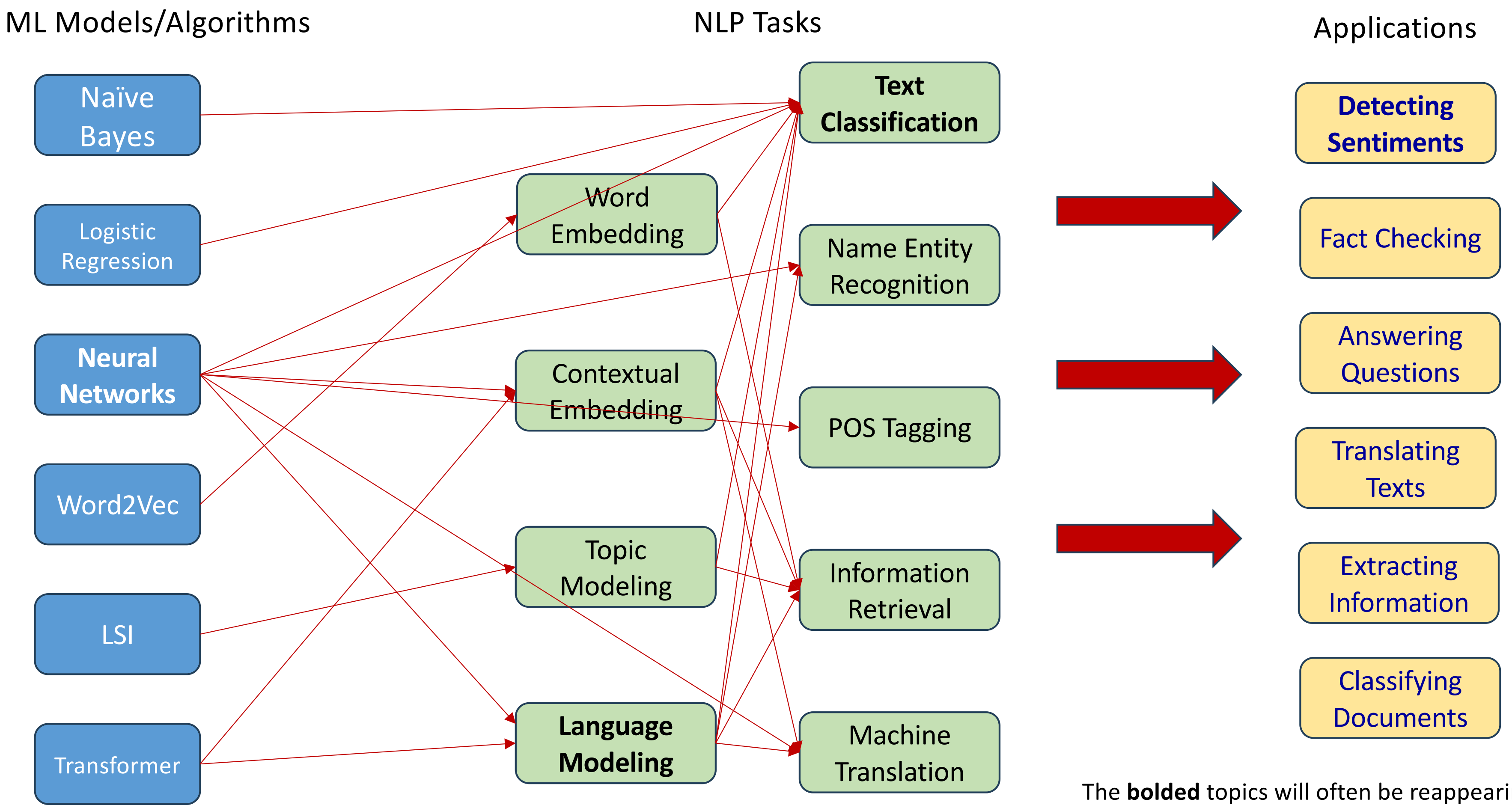
Course Objectives

Course Objectives

Developing ability to apply NLP and machine learning techniques quickly to create business value. To do this, we will

1. Explore practical applications of NLP
   - While we will engage with real code and datasets, the focus will be on application rather than in-depth programming knowledge
2. Develop proficiency in a variety of advanced machine learning tools by looking at the reasoning and intuition behind the models (without delve into all the specifics)

An (Incomplete) Overview for NLP Topics for the Course



The **bolded** topics will often be reappearing throughout this course

Why Natural Language Processing?

- Text data is everywhere; vast amount of text data are being collected every day
- However, the structure of the text are linguistics for humans, not computers
  - Not in numerical format and unstructured
  - Difficult to standardize
    - Many words can have multiple meanings in different context
  - In addition, text data such as tweets may contain
    - Spelling errors
    - Hyperlinks/tags/emojis
    - etc.

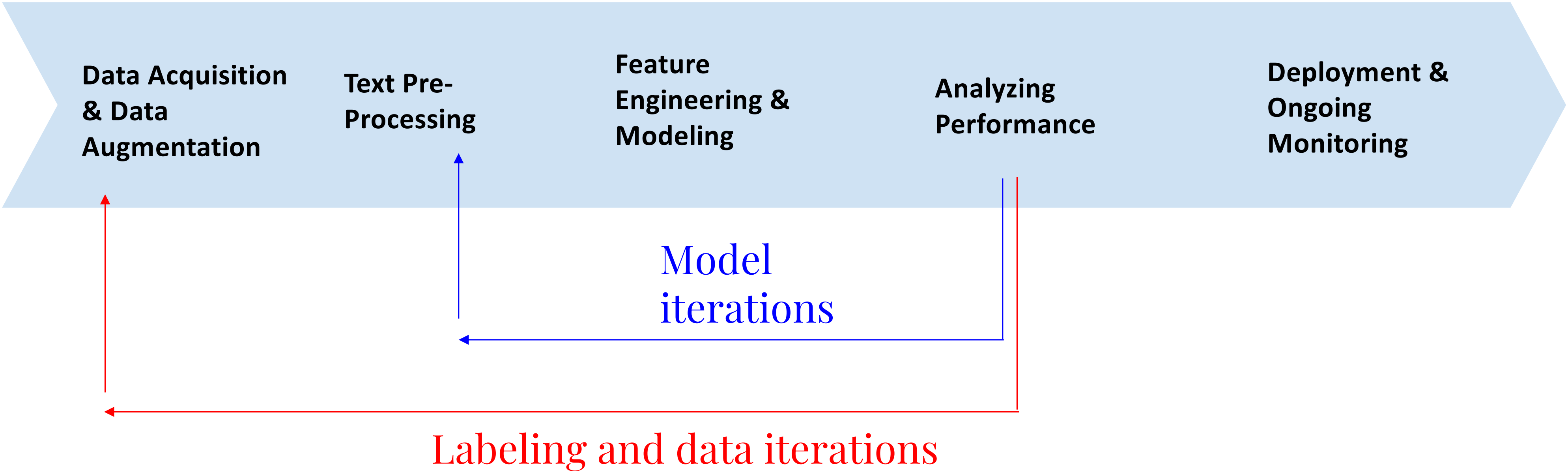
A Typical NLP Application Pipeline

Data:
Corpus for training
Labeling
OCR for digital text
Augmentation

Pre-Processing:
Stopword removal
Lowercasing
Lemmatization
Tokenization
Domain adjustments
Parsing

Modeling:
Language Model
Classifier approach
Search tools

Performance:
Precision
F1-Score
Perplexity
Business Results



Data Acquisition & Data Augmentation

Text Pre-Processing

Feature Engineering & Modeling

Analyzing Performance

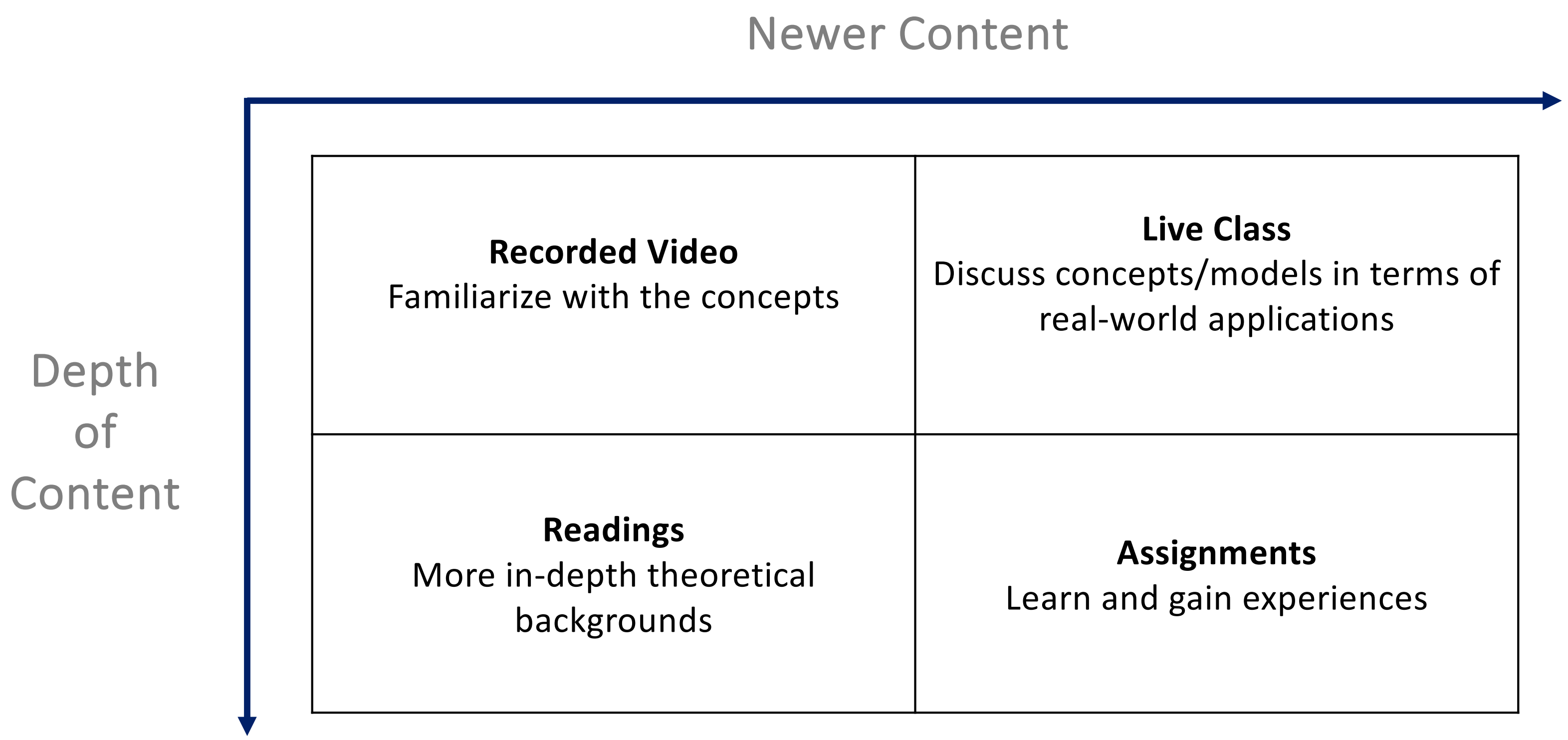
Deployment & Ongoing Monitoring

Model iterations

Labeling and data iterations

Course Logistics

Class Structure



Newer Content

Depth of Content

Recorded Video Familiarize with the concepts	**Live Class** Discuss concepts/models in terms of real-world applications
Readings More in-depth theoretical backgrounds	**Assignments** Learn and gain experiences

Course Website

- All course materials will be posted on Canvas, including:
  - The class slides and notebooks used in lectures (posted before each lecture)
  - Pre-class notebooks (posted before the beginning of each class)
  - Assignment solutions
  - Additional course materials and the syllabus containing course schedule, grading, TA info, etc
- There will also be important announcements posted on Canvas, please check the course website regularly

Gradings

- Homework assignments (50%)
  - You discusss with others for the assignments, but always submit your own solution
  - Due on Canvas at 11:59 pm on Friday, the night before class. Late submissions are accepted with penalties, at a rate of 1%*(Hours late, rounded up to whole hours).
- Final exam (35%)
- Class participation (15%)
  - Attend live sessions **on time**
  - Participate with good/answers that contributes to the learning
    - You do not need to speak at every class to receive full participation credit

Additional resources

- Teaching assistant: Ruifeng Ding, Dhaval Potdar
  - Office hours: 5:30 – 7:00 pm, 8:30 – 10:00 pm on Monday, Tuesday, Wednesday and Thursday during the week an assignment is due
  - You can also reach out to them via email
- My office hours: 12 to 1 pm on every Thursday
  - Also feel free to email me with questions or schedule Zoom appointments

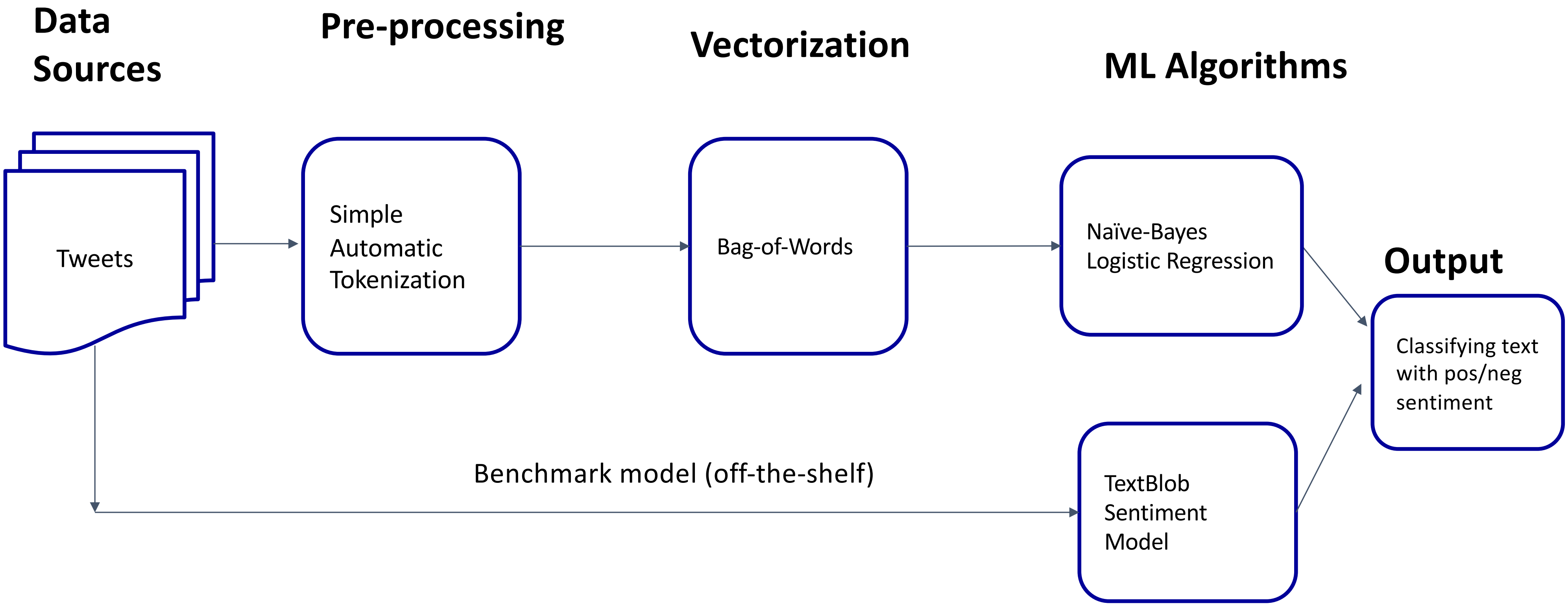
Text Classification

Determining if a tweet has positive or negative sentiment

Review: Some Important NLP Terminologies

1. **Token**: The smallest unit of processing, usually a word, but it can also be a username or an emoji/emoticon.
2. **Document**: A single data point or piece of text in NLP, often corresponding to a tweet, an article, a text file, a message, etc.
3. **Corpus (plural: Corpora)**: A collection of documents. For instance, all tweets in the NLTK dataset constitutes a corpus.
4. **Vocabulary**: The set of all unique tokens in a corpus.
5. **Lexicon**: a dictionary that contains information about words and their properties.

Pipeline for Our First Twitter Sentiment Analysis



Data Sources

Pre-processing

Vectorization

ML Algorithms

Output

Tweets

Simple Automatic Tokenization

Bag-of-Words

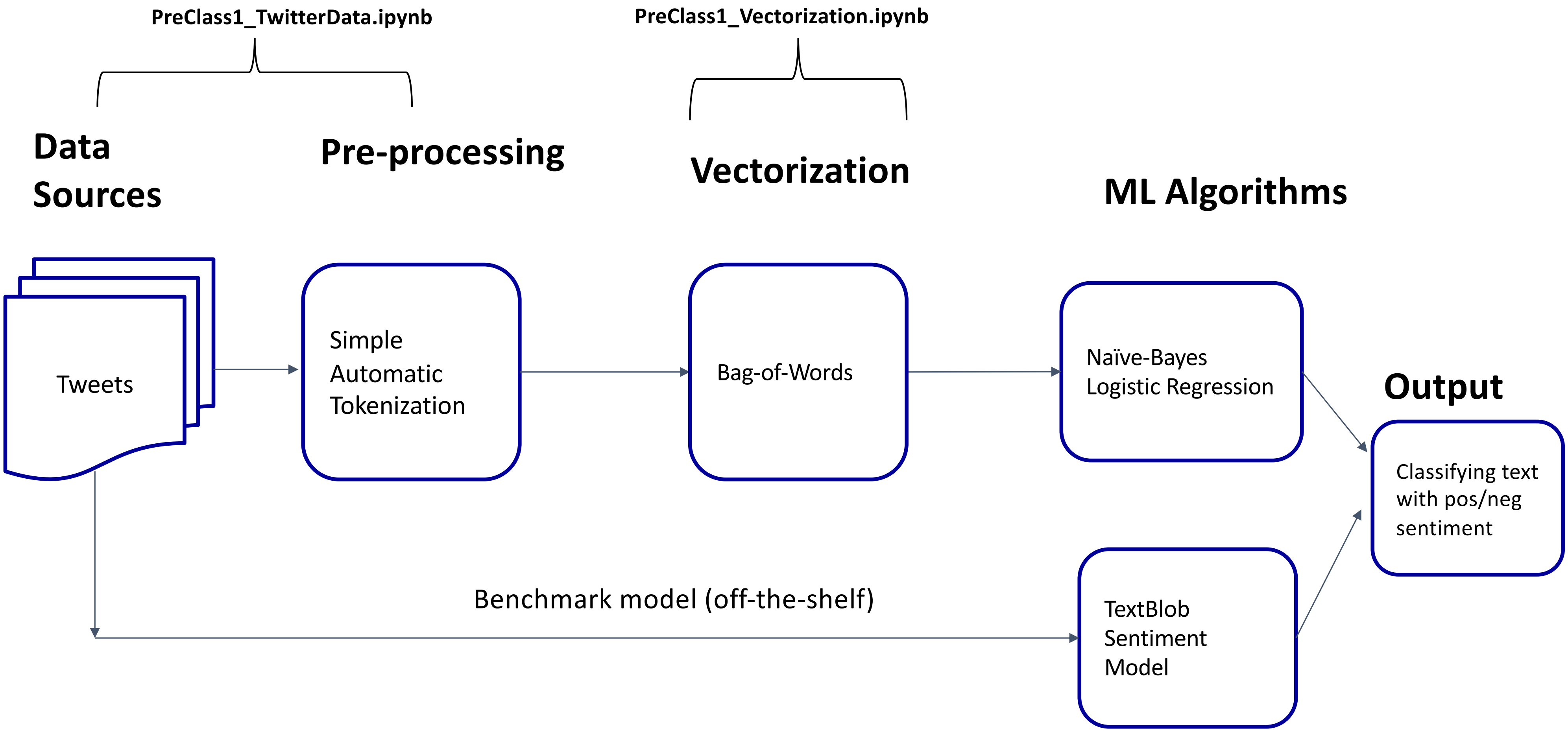
Naïve-Bayes
Logistic Regression

Classifying text with pos/neg sentiment

Benchmark model (off-the-shelf)

TextBlob Sentiment Model

Pipeline for Our First Twitter Sentiment Analysis



PreClass1_TwitterData.ipynb

PreClass1_Vectorization.ipynb

Data Sources

Pre-processing

Vectorization

ML Algorithms

Output

Tweets

Simple Automatic Tokenization

Bag-of-Words

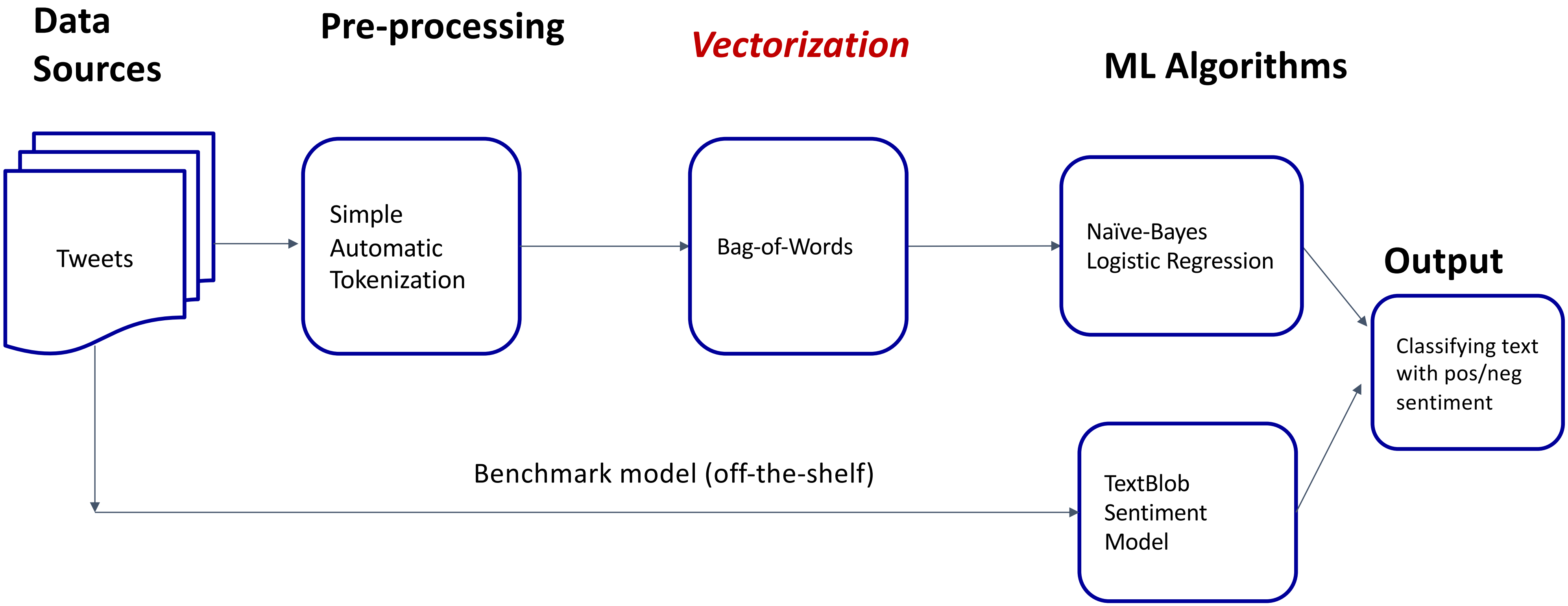
Naïve-Bayes
Logistic Regression

Classifying text with pos/neg sentiment

Benchmark model (off-the-shelf)

TextBlob Sentiment Model

Pipeline for Our First Twitter Sentiment Analysis



An Example of the Bag of Words (BOW) Vectorization Technique

Suppose that we want to identify whether a sentence contains positive sentiment. And we have the following training set (with 3 documents):

A) Bob likes to drink beer. (contains positive sentiment)
B) My kid likes to dance. (contains positive sentiment)
C) Babies drink milk. (contains no positive sentiment)

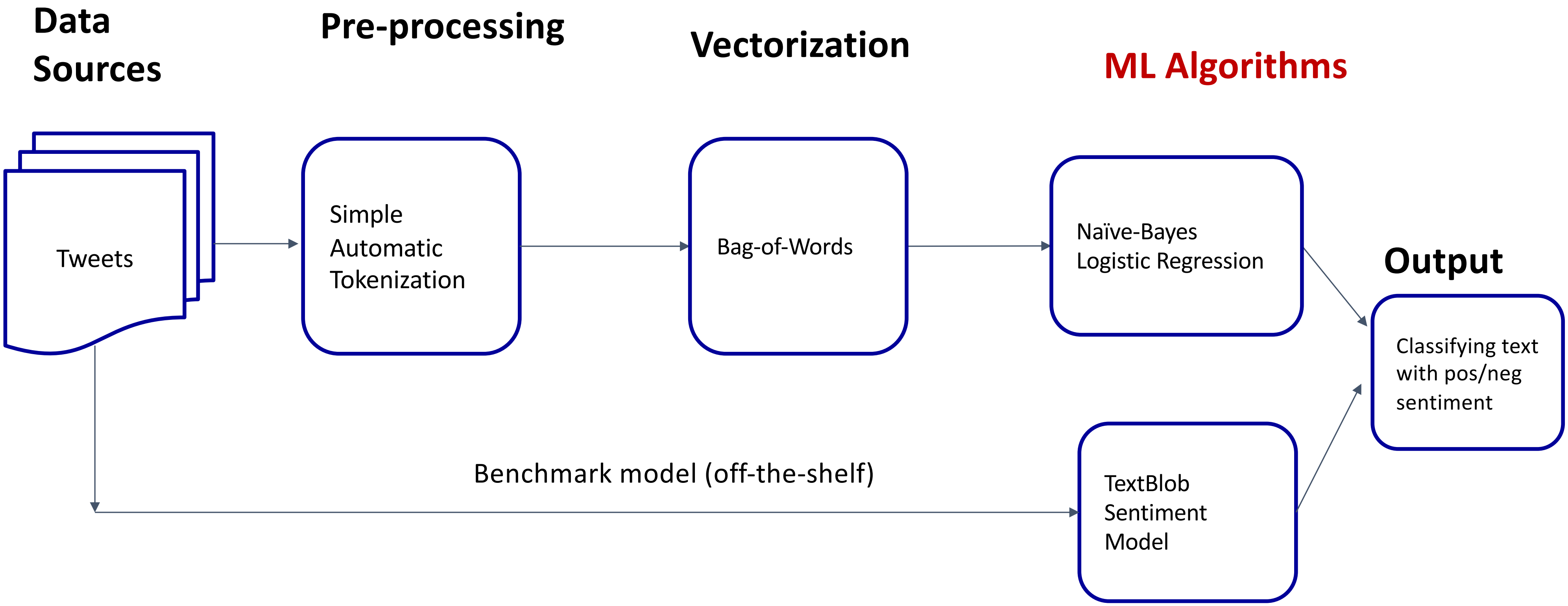
3x 10 matrix **Tokens (words)**

	Bob	likes	to	drink	beer	My	kid	dance	babies	milk	y
A	1	1	1	1	1	0	0	0	0	0	1
B	0	1	1	0	0	1	1	1	0	0	1
C	0	0	0	1	0	1	0	0	1	1	0

X y

Matrix is *r by c* where *r* = # documents or in this, sentences; *c* = size of dictionary (tokens)
We form a sparse matrix with counts of tokens by document (sentences)
We build models using the vectorized matrix (X) relating to the sentiment labels (y)

Pipeline for Our First Twitter Sentiment Analysis



Data Sources

Pre-processing

Vectorization

ML Algorithms

Tweets

Simple Automatic Tokenization

Bag-of-Words

Naïve-Bayes
Logistic Regression

Output

Classifying text with pos/neg sentiment

Benchmark model (off-the-shelf)

TextBlob Sentiment Model

Review: Training Machine Learning Models

Two important consideration when we apply a machine learning model:

1. **The underlying model:** a mathematical representation of how data is generated.
   - Consists of an equation or rules describing the input features to the output
   - Consists of a set of parameters that will be learned from the data
2. **Algorithms to train the model**: the algorithm for determining the best parameters that make the model fit the training data
   - Often, there are various training algorithms to choose from
   - Other important considerations include tuning model hyperparameters and techniques to avoid overfitting

Naïve Bayes and Logistic Regression Models

In our application, we use a ML model to classify whether a tweet has positive (y=1) or negative sentiment (y=0)

Naïve Bayes

Model:

$$\hat{P}(y = 1|\,\boldsymbol{x}) = \frac{\hat{P}(y = 1) \cdot \hat{P}(x_1\,|y = 1) \cdot \ldots \cdot \hat{P}(x_n\,|y = 1)}{Z}$$

Generative model, readily applies to >2 classes

Trained Parameters:

$\hat{P}(y = 1), \hat{P}(x_1\,|y = 1), \ldots, \hat{P}(x_n\,|y = 1)$ (trained through a simple count-based algorithm)

Logistic Regression

Model:

$$\mathrm{P}(y = 1|\,\boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{w}^T\boldsymbol{x}}} \ \textbf{OR} \ \frac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

Discriminative model, applies only to 2 classes (can be generalized to multi-class logistic regression models and multi-layered neural networks)

Trained Parameters:

vector $\mathbf{w}$ (trained using a gradient-based algorithm)

Homework 1

Analyze tweets using the NLTK Twitter Sample Corpus

- Q1: Concepts of Sentiment Analysis in Business. (20 points)
- Q2: Loading and Viewing Data (20 points)
- Q3: Naive Bayes for Sample Tweets (20 points)
- Q4: Pre-Built Sentiment Analyzers (20 points)
- Q5: Lexicon Matching (20 points)
- Bonus: Manipulating Scores of Sentiment Analyzers (10 points)

Hints: review the pre-class notebooks and class 1 notebook as you work through the questions