Text Mining

Spring 2018

Week 1

Where do we find text?

- Fiction
- News
- Scientific books, articles
- Every-day communication (email, twitter messages, SMS messages)
- Reviews (Amazon product reviews)
- Etc…
- Text is everywhere

Course Goals

- Provide an introduction to both Natural Language Processing (NLP) and Data Mining ➔ Text Mining
  - Simple: counting word frequencies to compare different writing styles.
  - Difficult: "understanding" complete human utterances, at least to the extent of being able to give useful responses to them.

About myself

- Elena Filatova, PhD in CS from Columbia University
- efilatova@citytech.cuny.edu
- Current affiliation: CUNY CityTech (NYC College of Technology)
- Research interests:
  - Information extraction
  - Summarization
  - Sarcasm detection
  - Crowdsourcing

About you

- Name
- Major: Linguistics, Computer Science, Electrical Engineering, other?
- Coursework and other background in each of NLP, Data Mining
- Prior research and current research Interests
- Future plans: academia or industry

Course information

- Blackboard
  - Syllabus
  - Weekly reading assignments
  - Programming assignments and submissions
  - Project
  - Lecture notes

  (4 programming assignments plus a term project)
- Technology
  - Python
  - NLTK
  - Azure Notebooks

Term Projects

- Question Answering
- Specialized Search
- Reviews/Recommendations Analysis
- Fraud Detection
- Sentiment Analysis

History

- What was the main NLP task in the dawn of Computer Science? When?
- What is the first example of an NLP task that comes to your mind now?

- Desk Set movie (8^{th} out of 9 Kathrine Hepburn and Spencer Tracy movies and their last comedy)

https://www.youtube.com/watch?v=ZK3zmPUxblk (4:20)

https://www.youtube.com/watch?v=nBT1oHGSeFc (2:45)

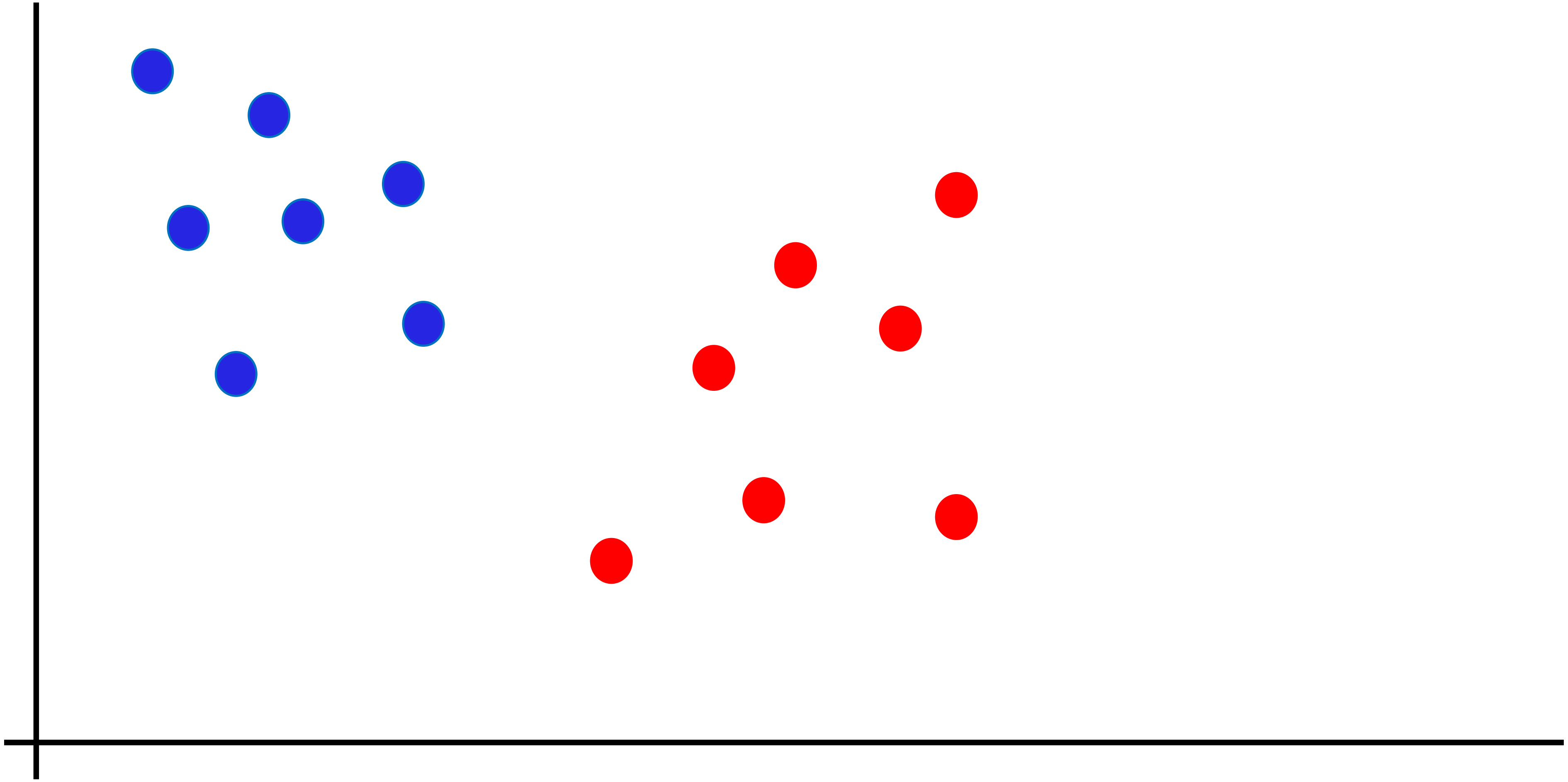
IBM Watson

Stanford Reading Comprehension Task

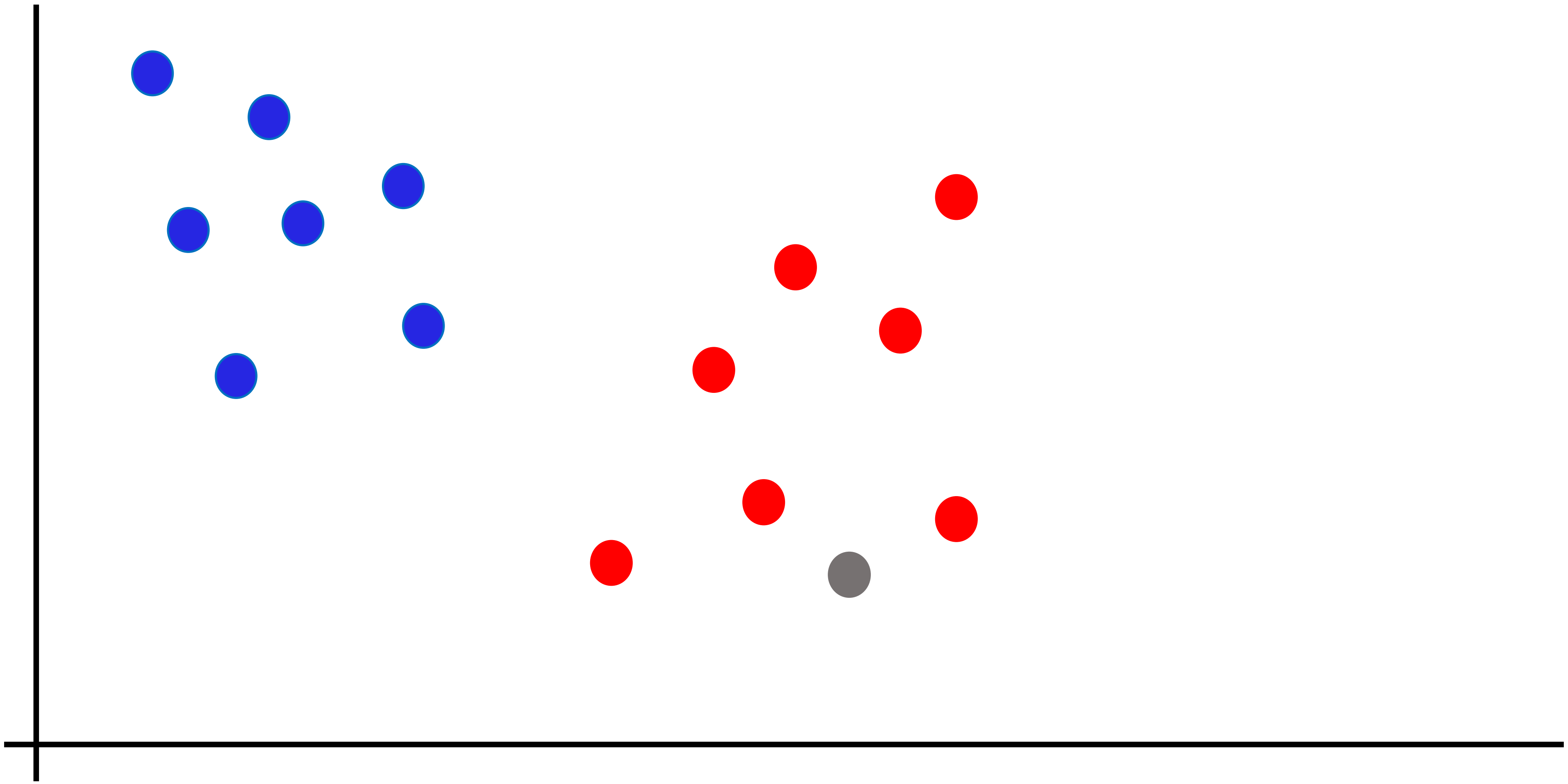
Major Data Mining Tasks

- Regression
  - Predict a numeric value given “other information”
- Classification
  - Predict a categorical value given “other information”
- Clustering
  - Identify groups of similar entities.
- Learning Feature Representations
  - What’s the best way to describe this data?
- Evaluation

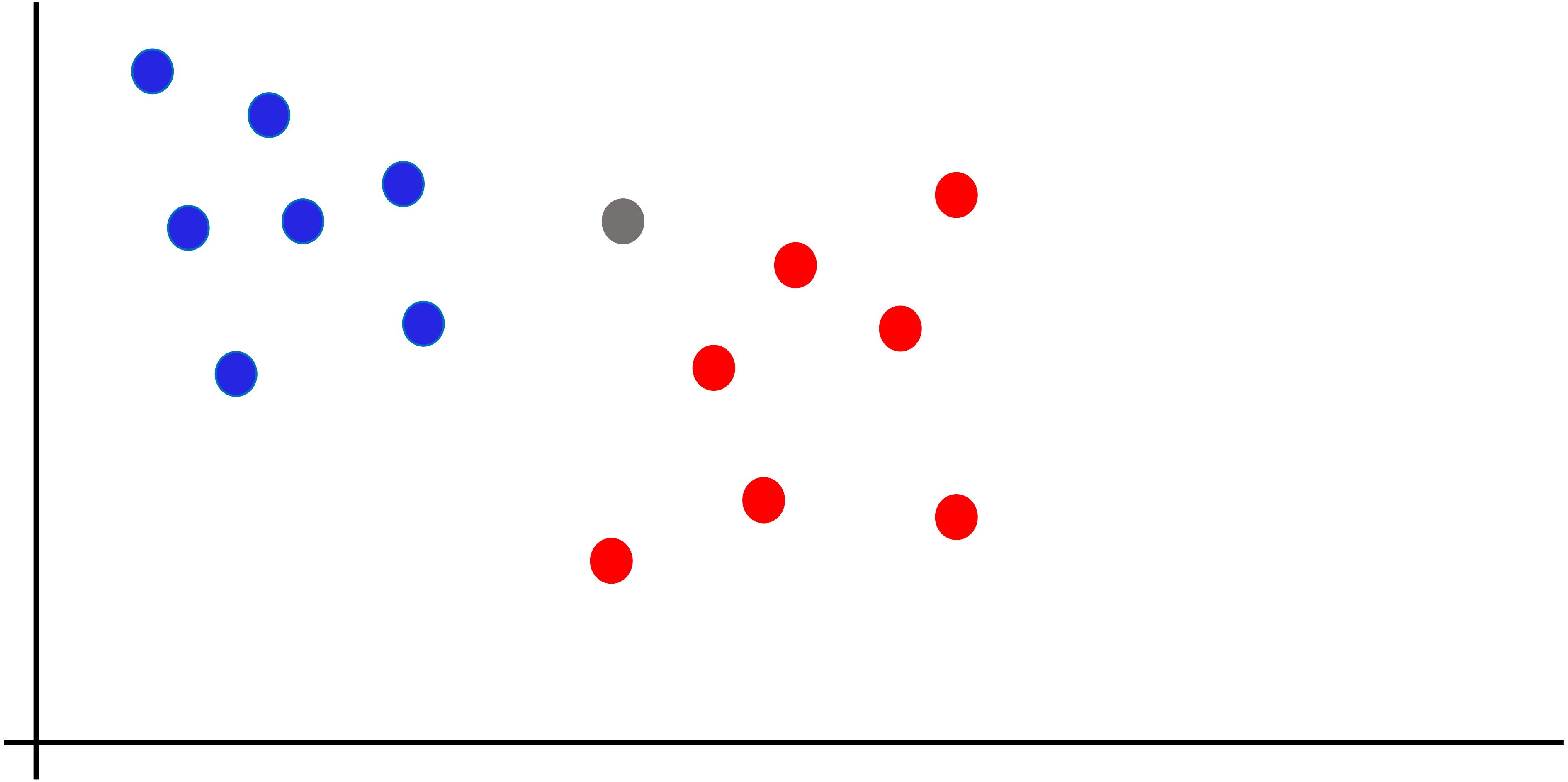
Classification



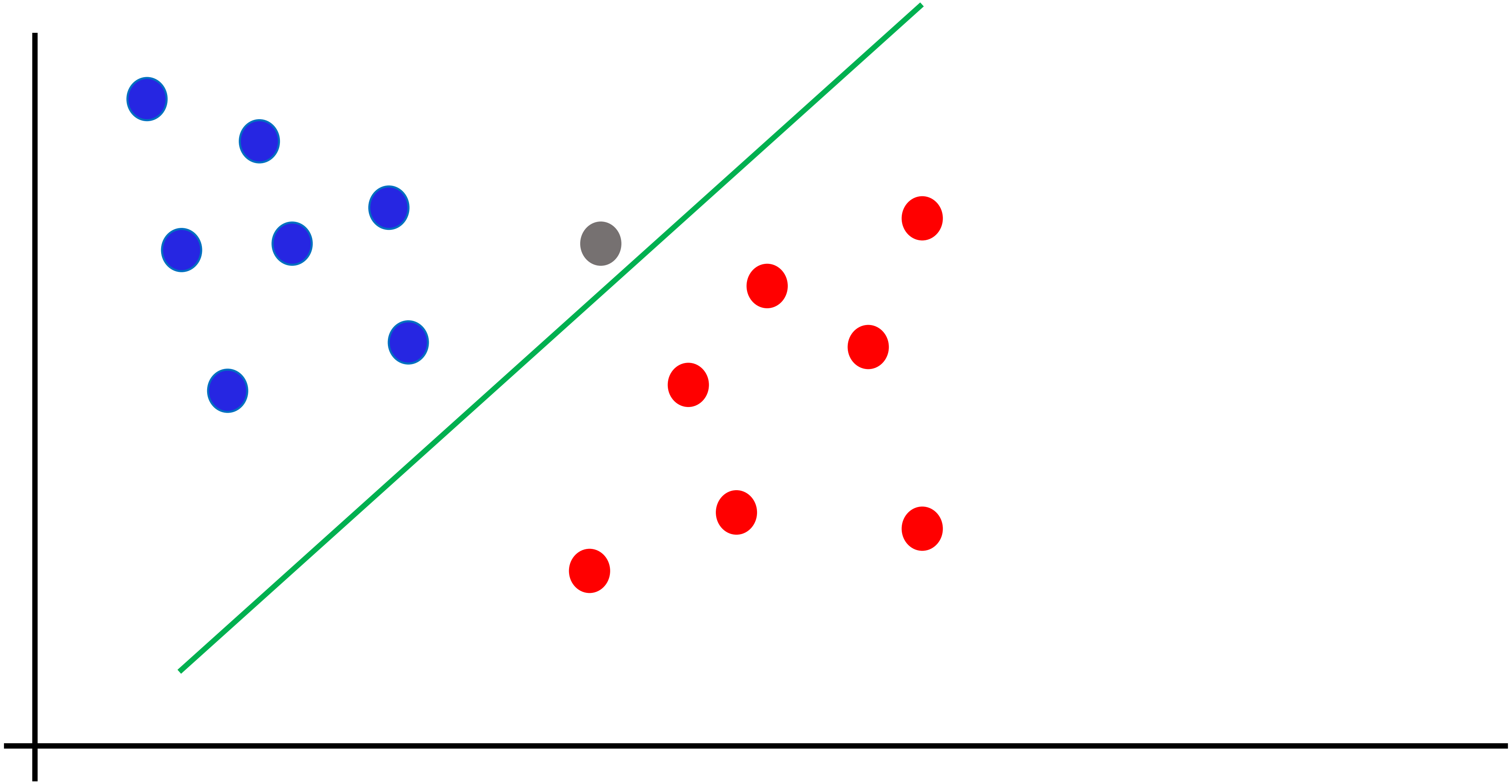
Classification



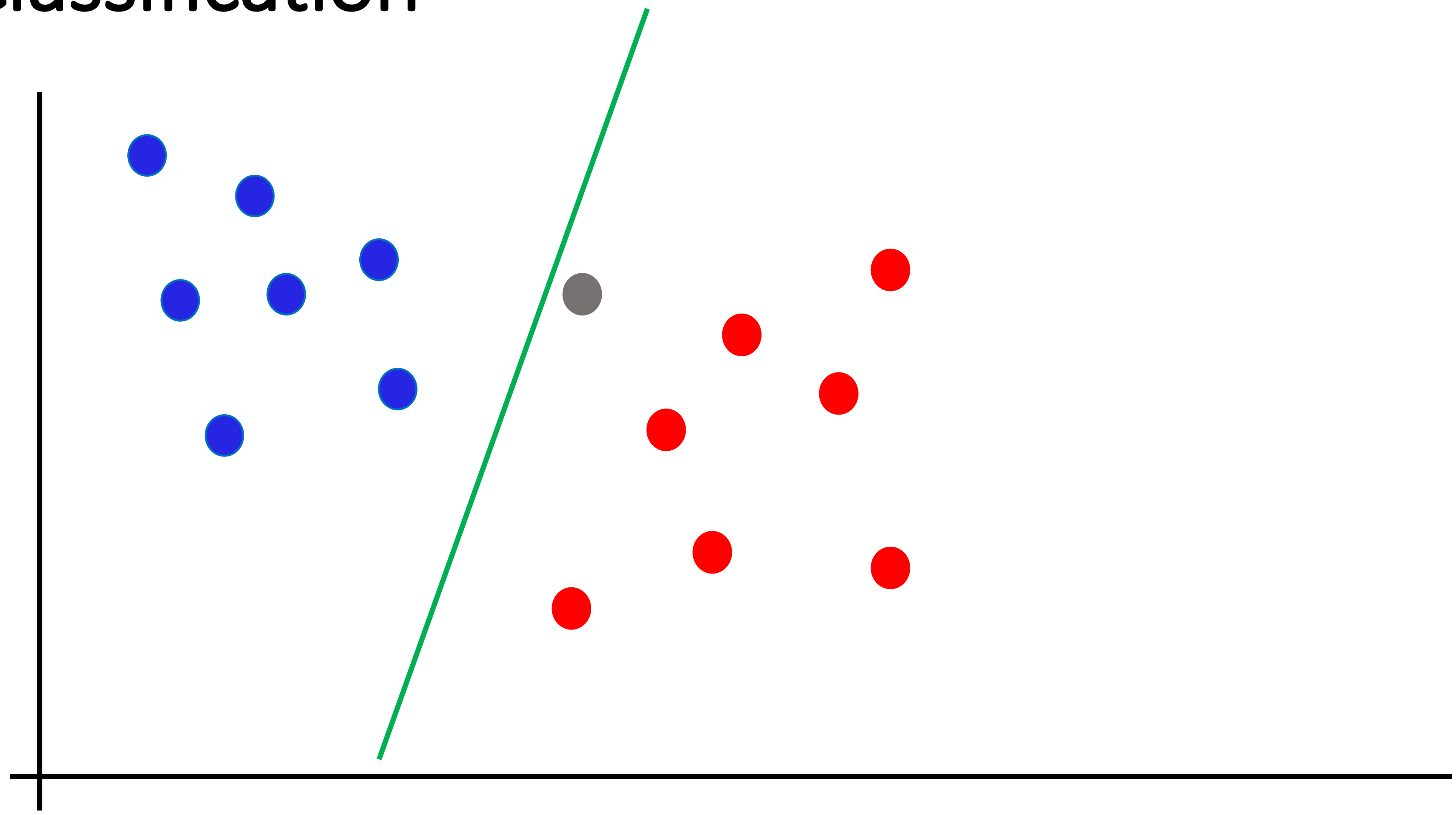
Classification



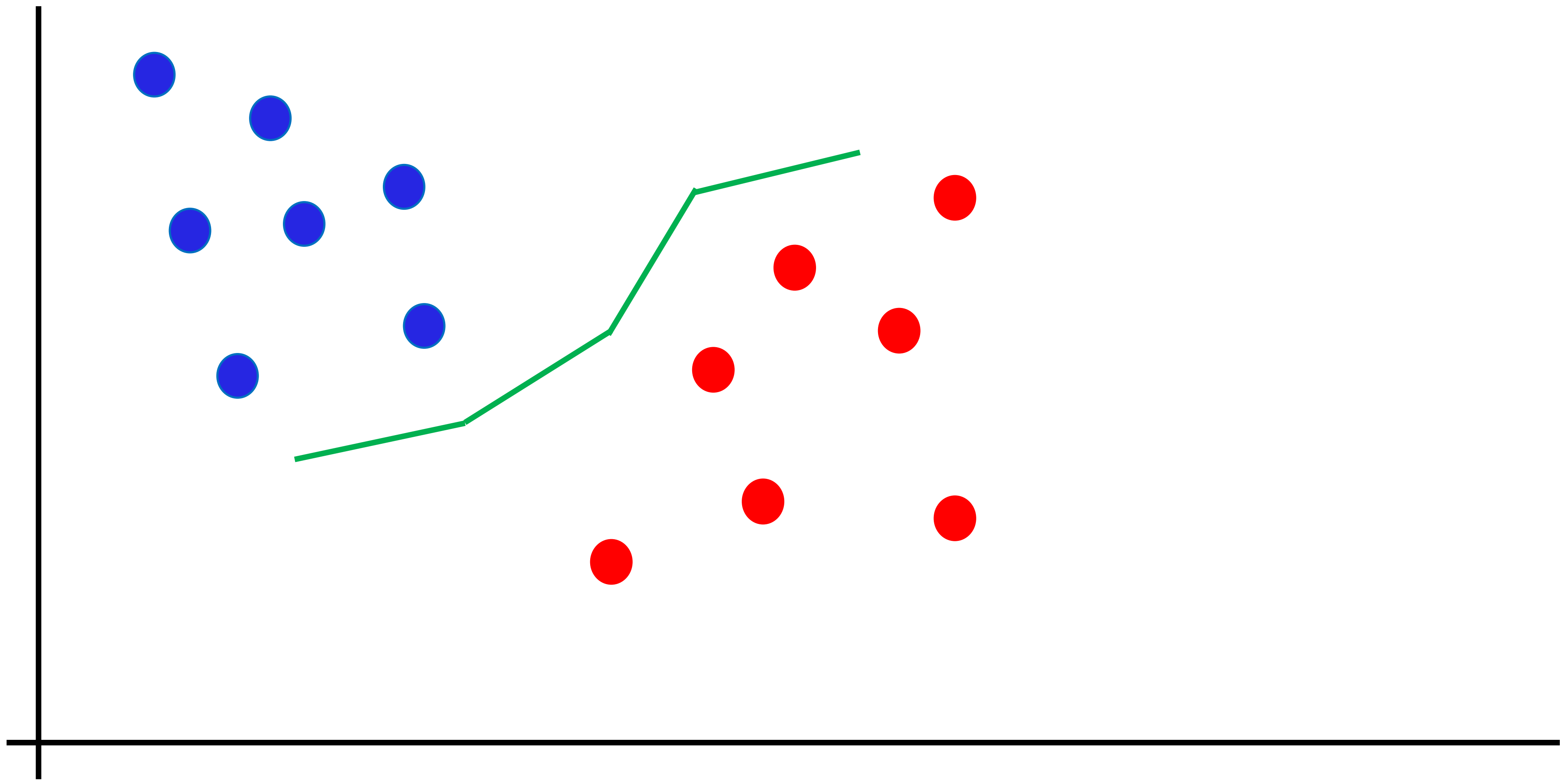
Classification



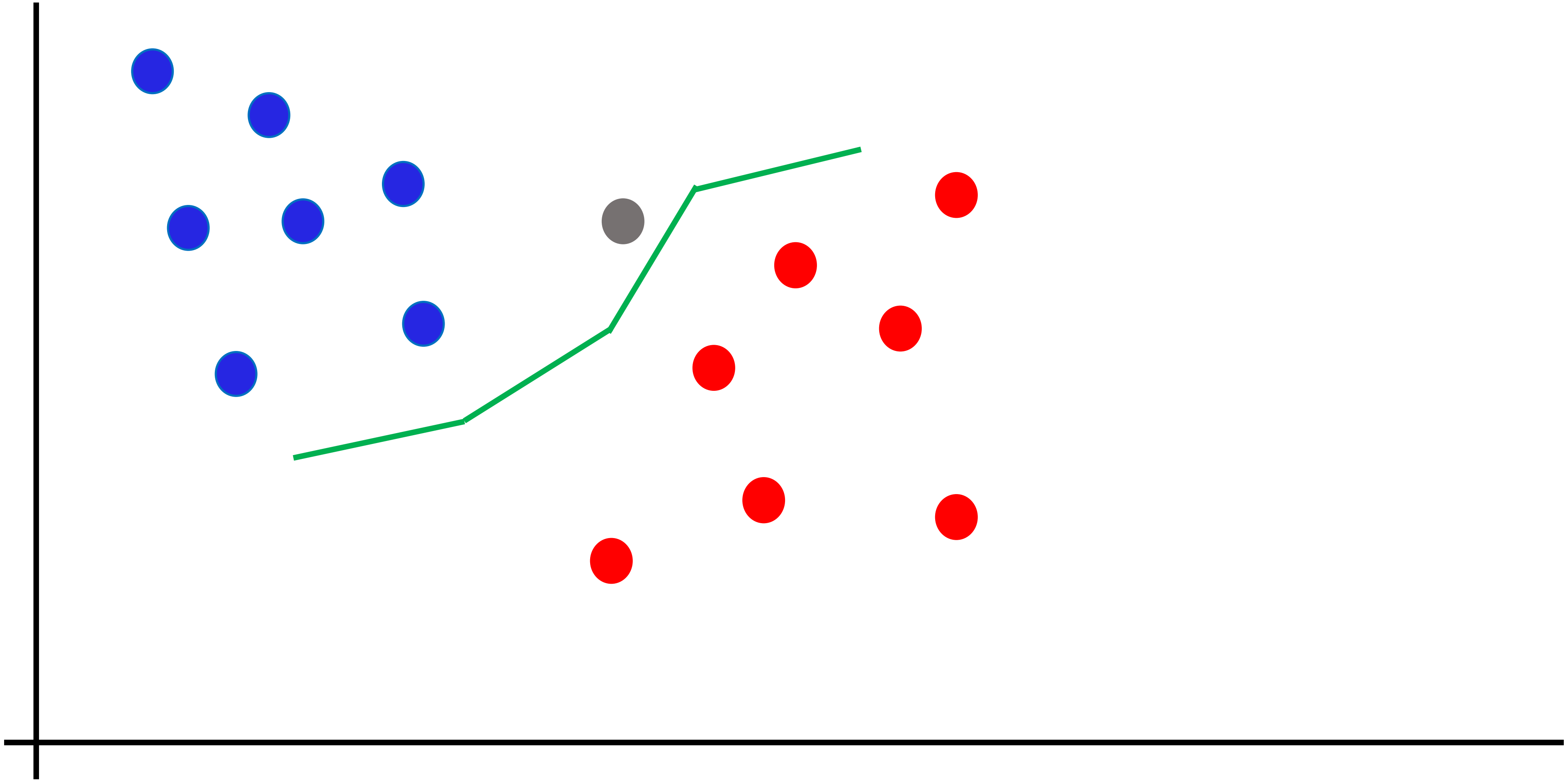
Classification



Classification

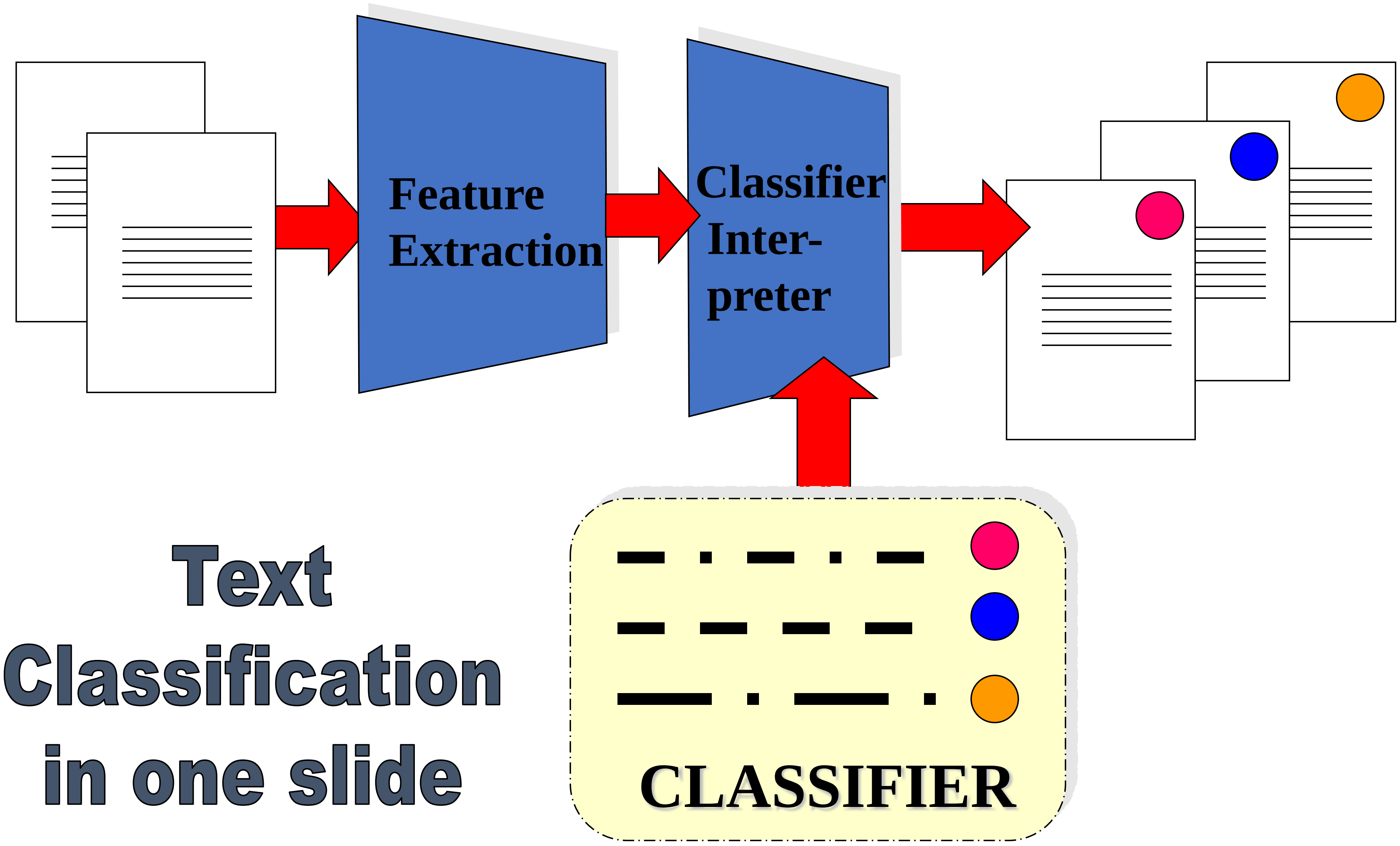


Classification



NLP, Text Mining and Classification

- Document classification:
  - Spam / not spam
  - By topic
- POS tagging
- Syntactic parsing



Text Classification in one slide

Lexical Ambiguity

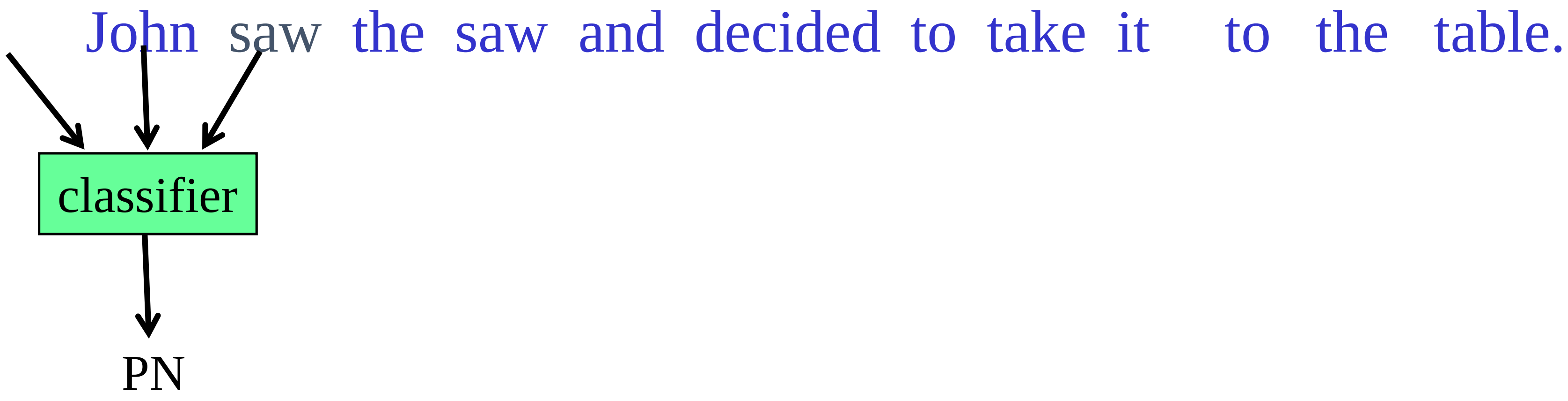
- Most words in natural languages have multiple possible meanings.
  - “pen” (noun)
    - The dog is in the pen.
    - The ink is in the pen.
  - “take” (verb)
    - Take one pill every morning.
    - Take the first right past the stoplight.

- Syntax helps distinguish meanings for different parts of speech of an ambiguous word.
  - “conduct” (noun or verb)
    - John’s conduct in class is unacceptable.
    - John must will conduct the orchestra on Thursday.

- Word Sense Disambiguation (WSD)

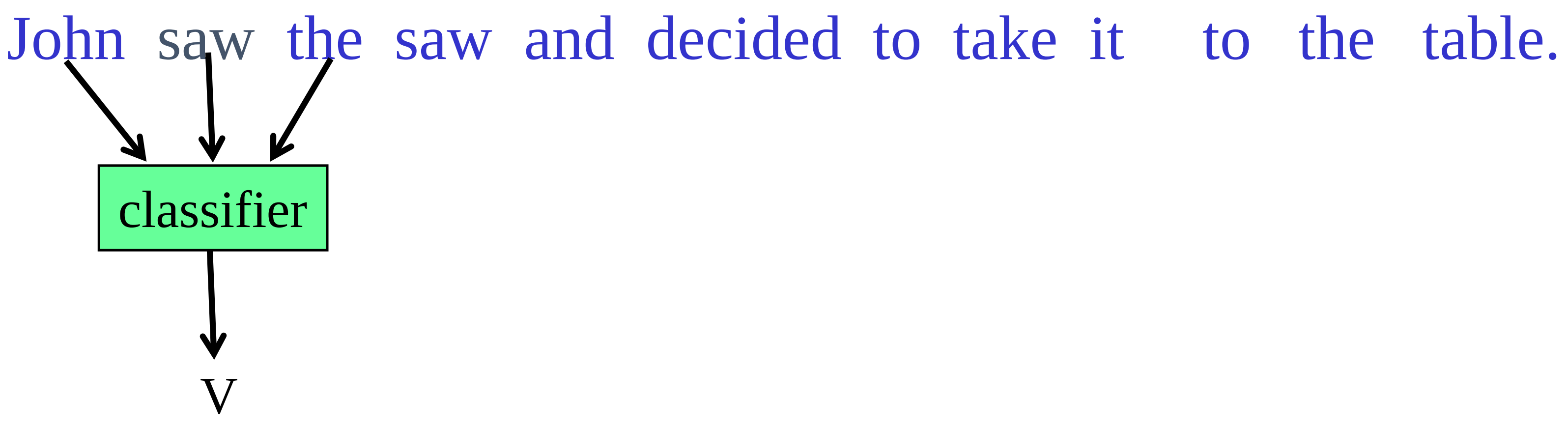
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



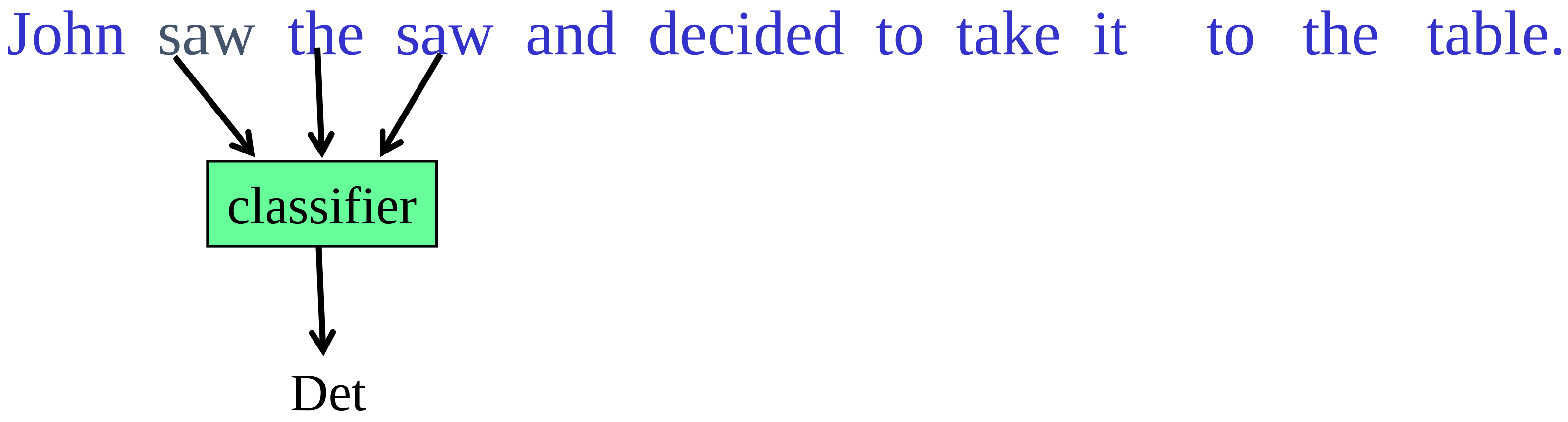
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



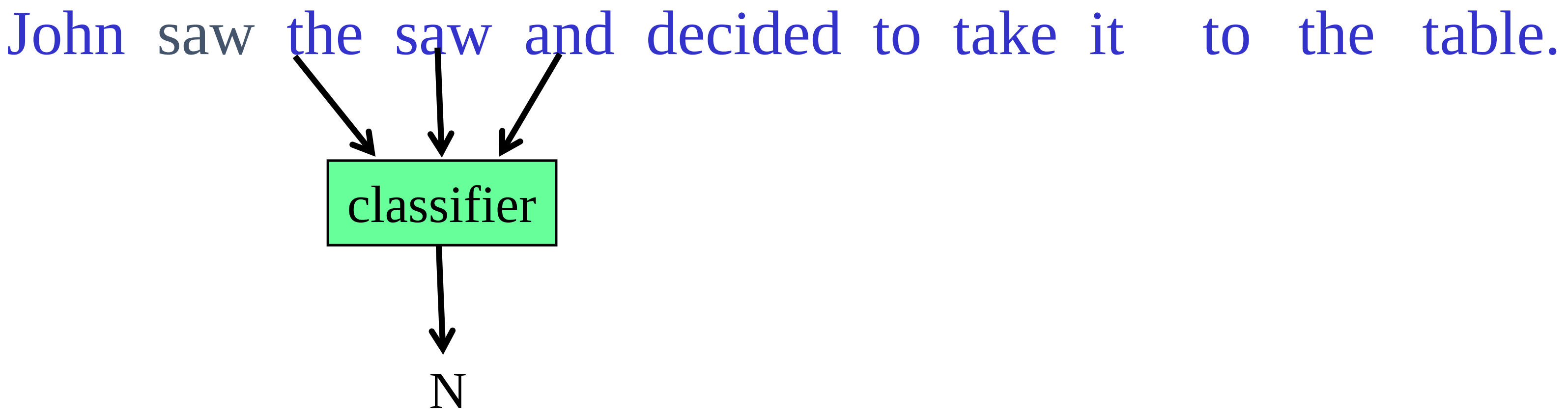
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



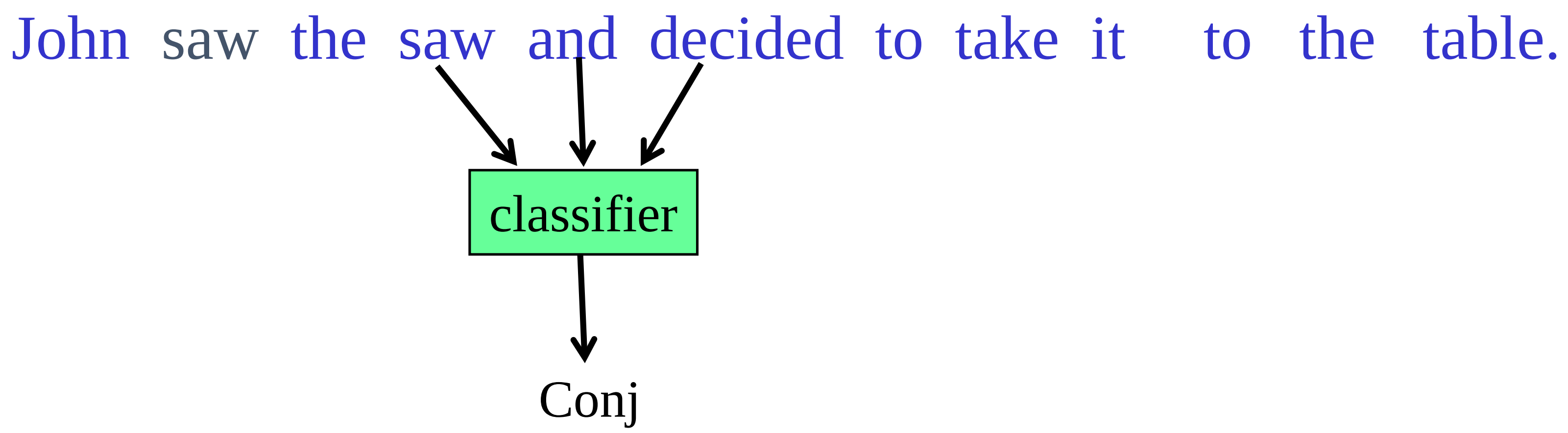
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



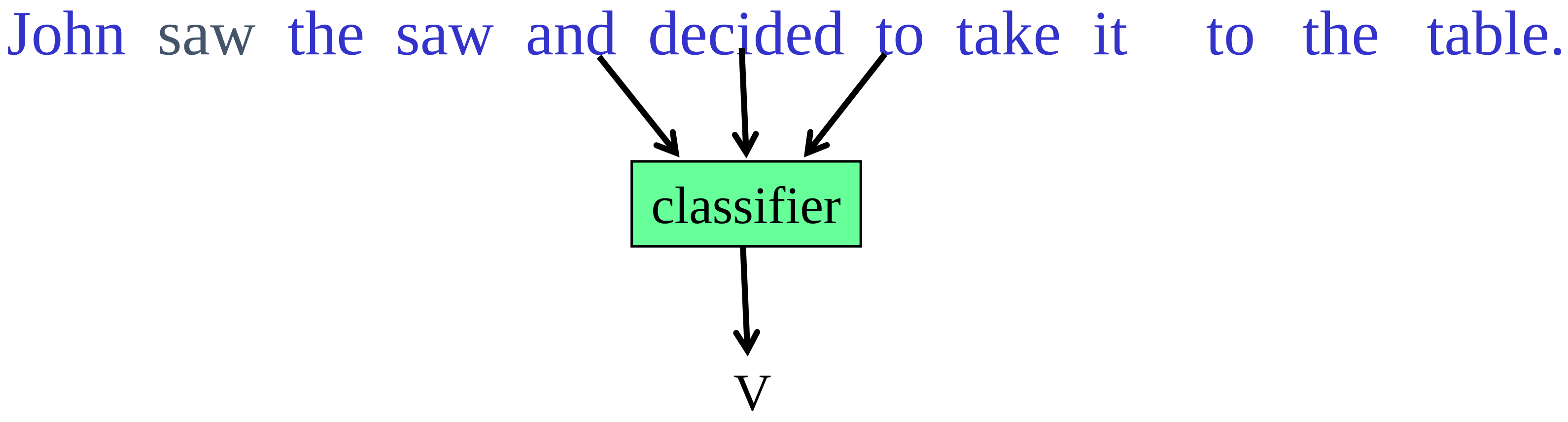
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



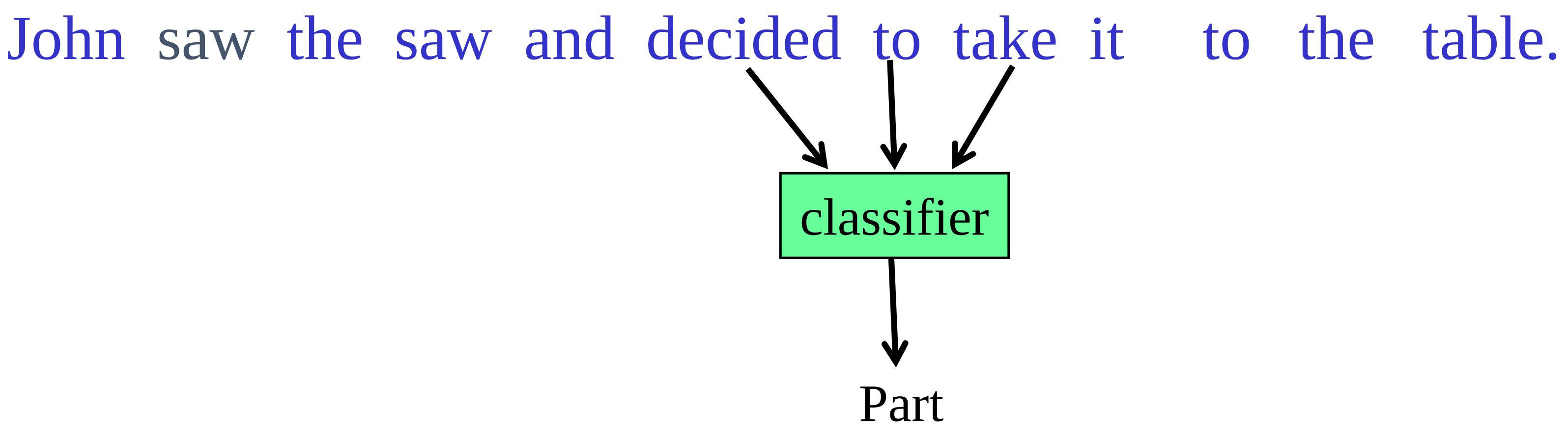
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



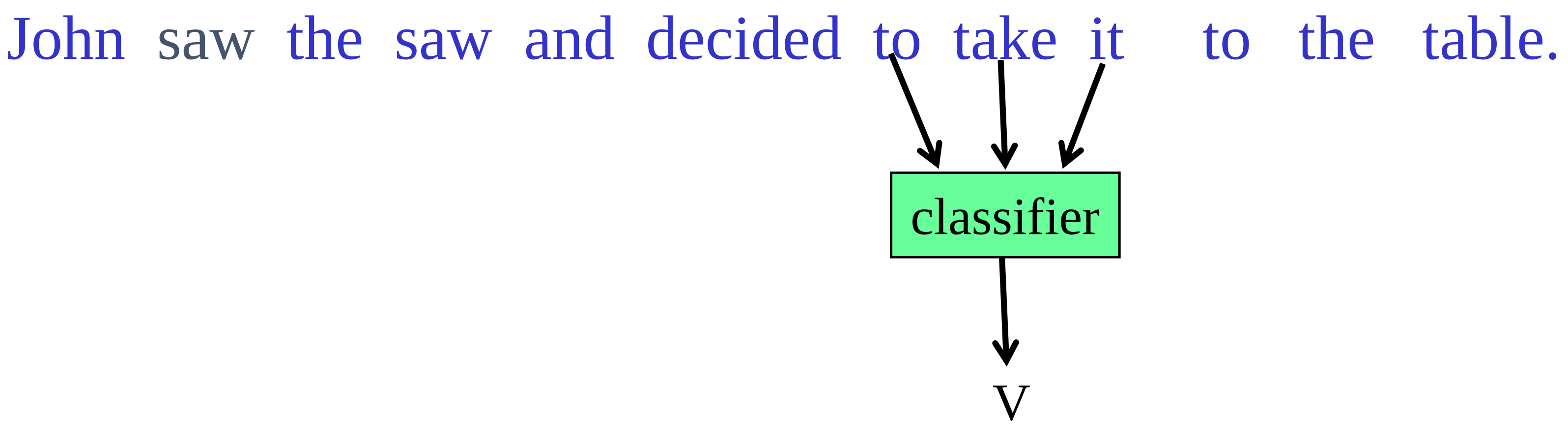
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



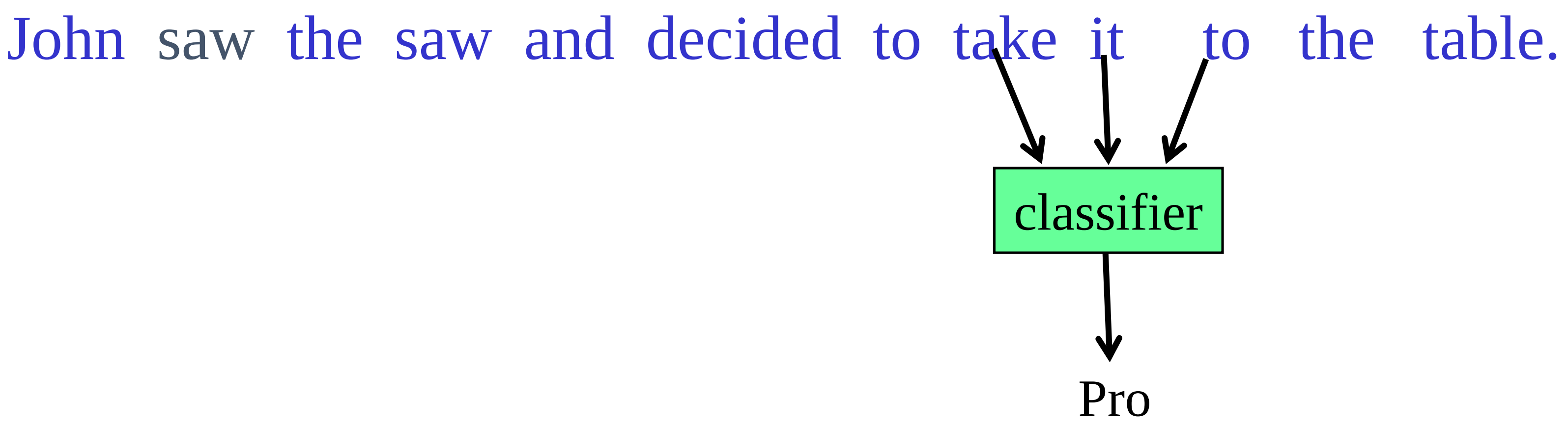
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



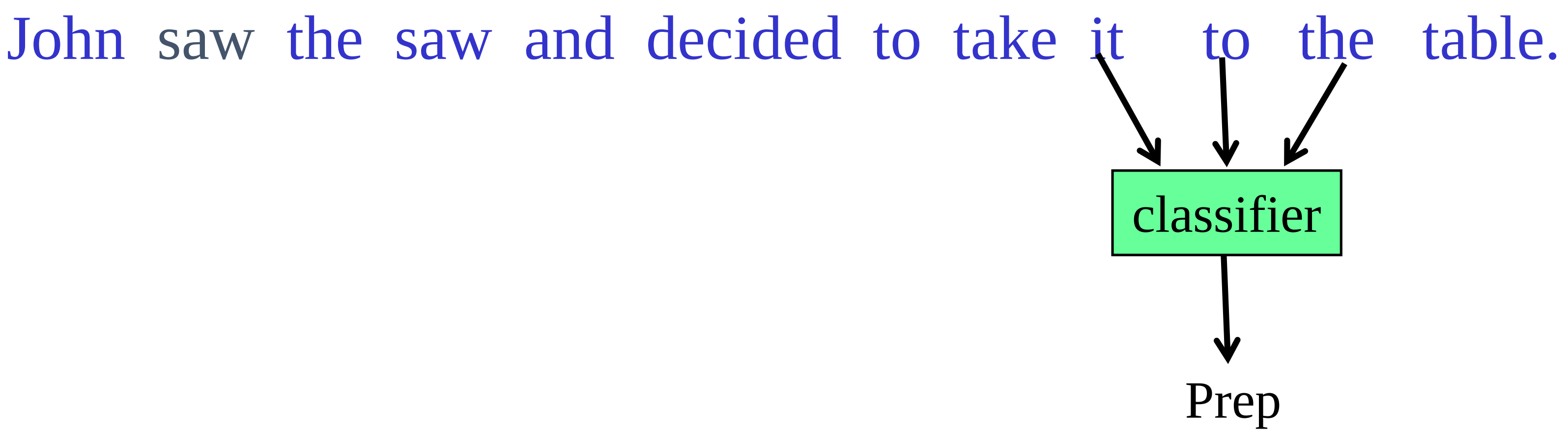
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



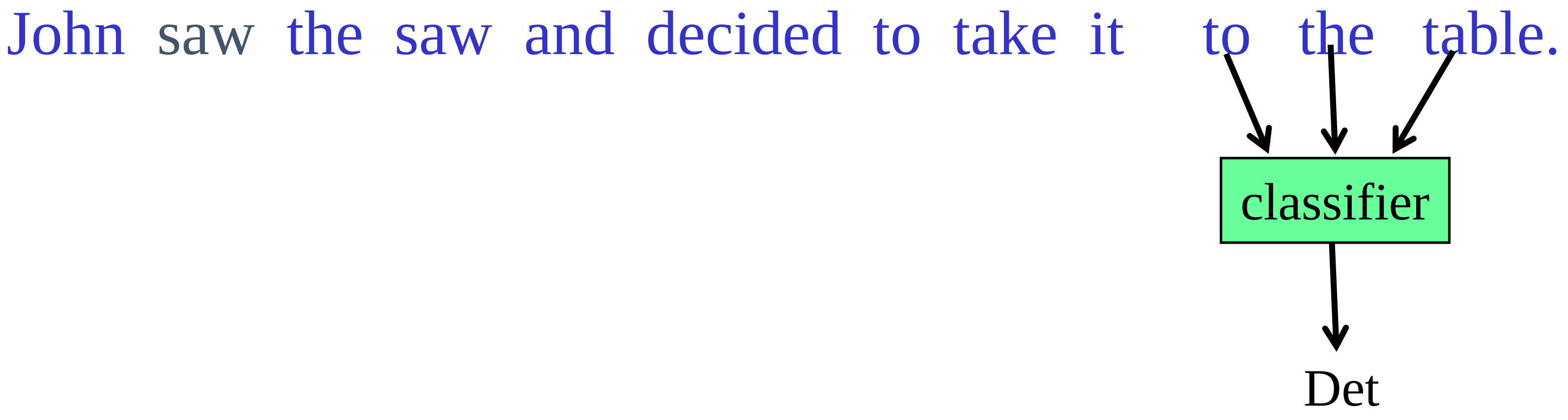
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



Sequence Labeling as Classification

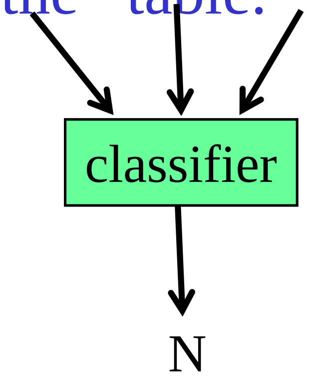
- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

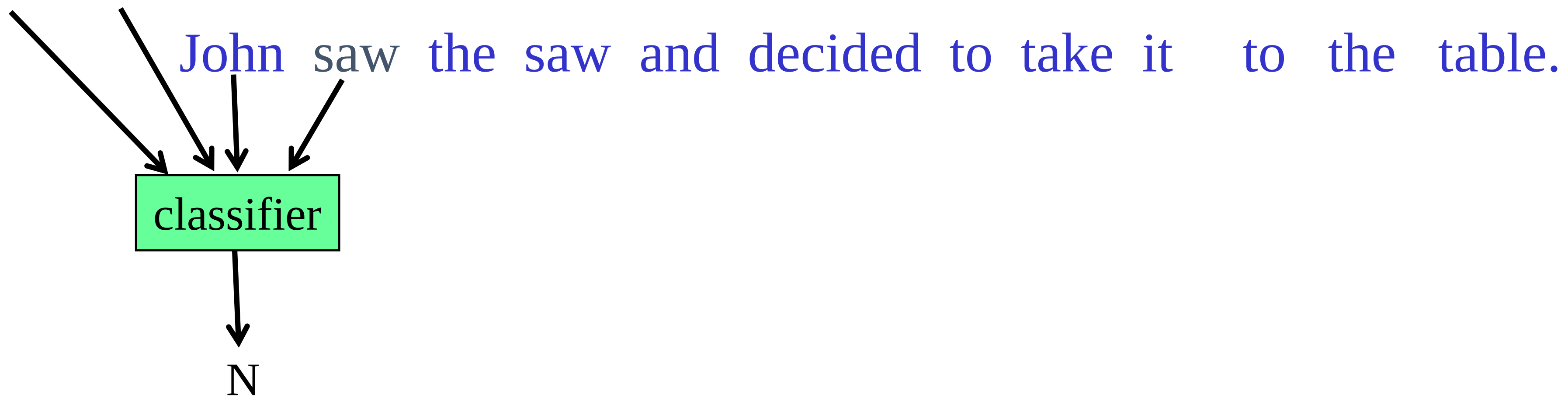
John saw the saw and decided to take it to the table.



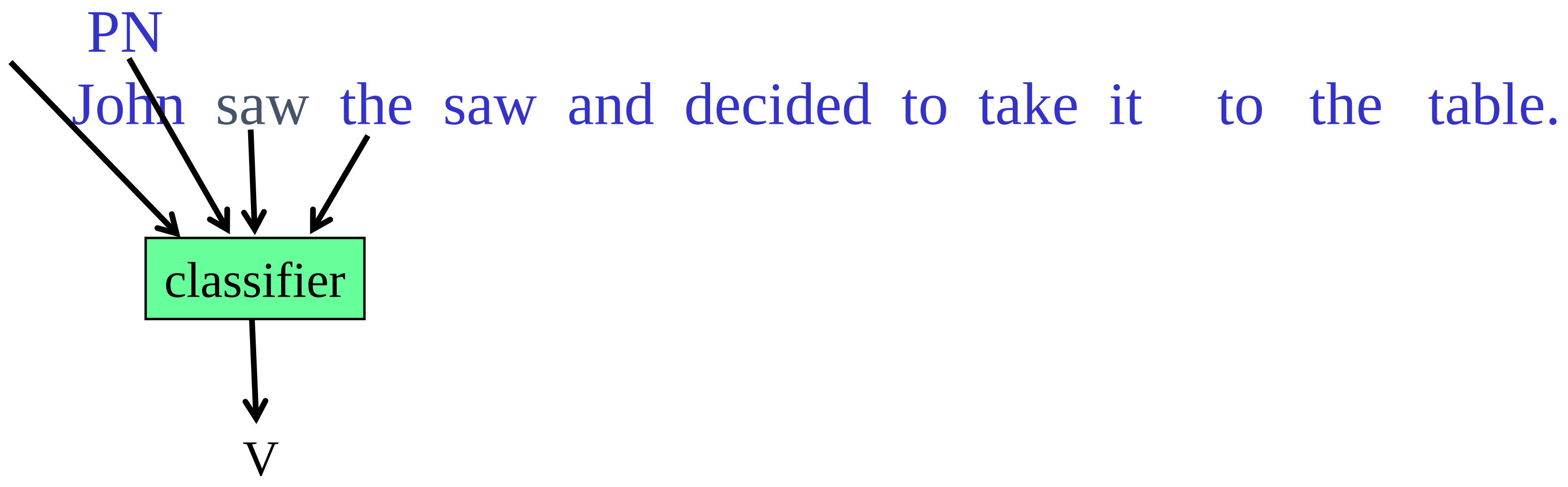
Sequence Labeling as Classification Using Outputs as Inputs

- Better input features are usually the categories of the surrounding tokens, but these are not available yet.
- Can use category of either the preceding or succeeding tokens by going forward or back and using previous output.

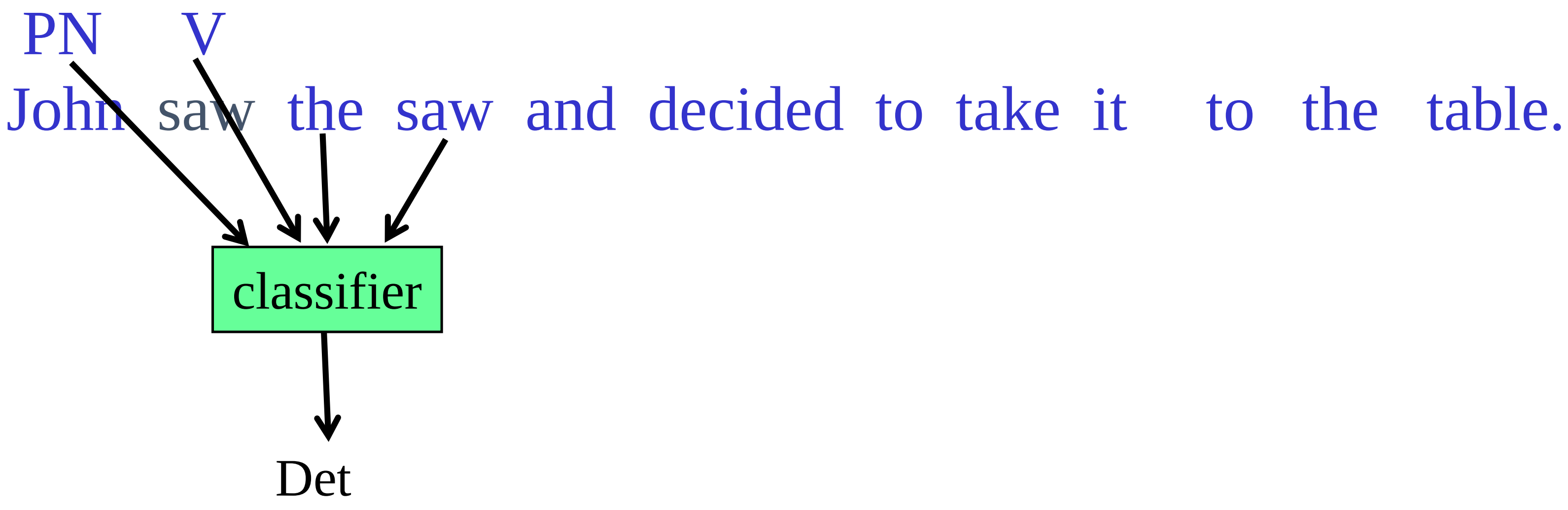
Forward Classification



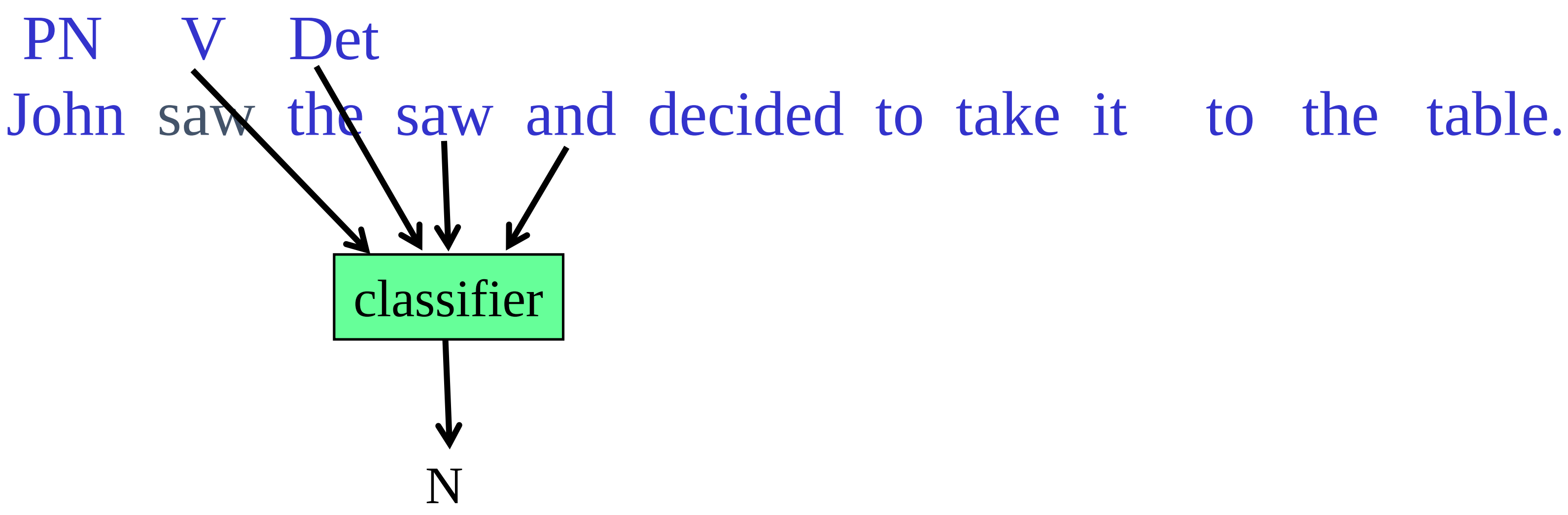
Forward Classification



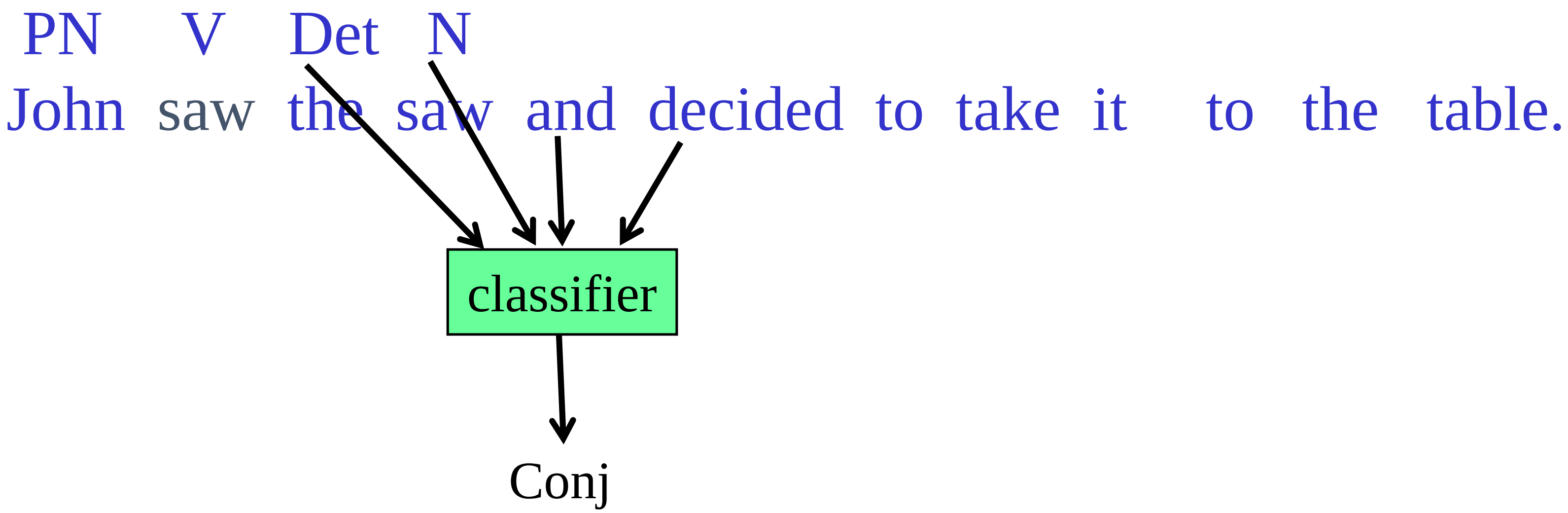
Forward Classification



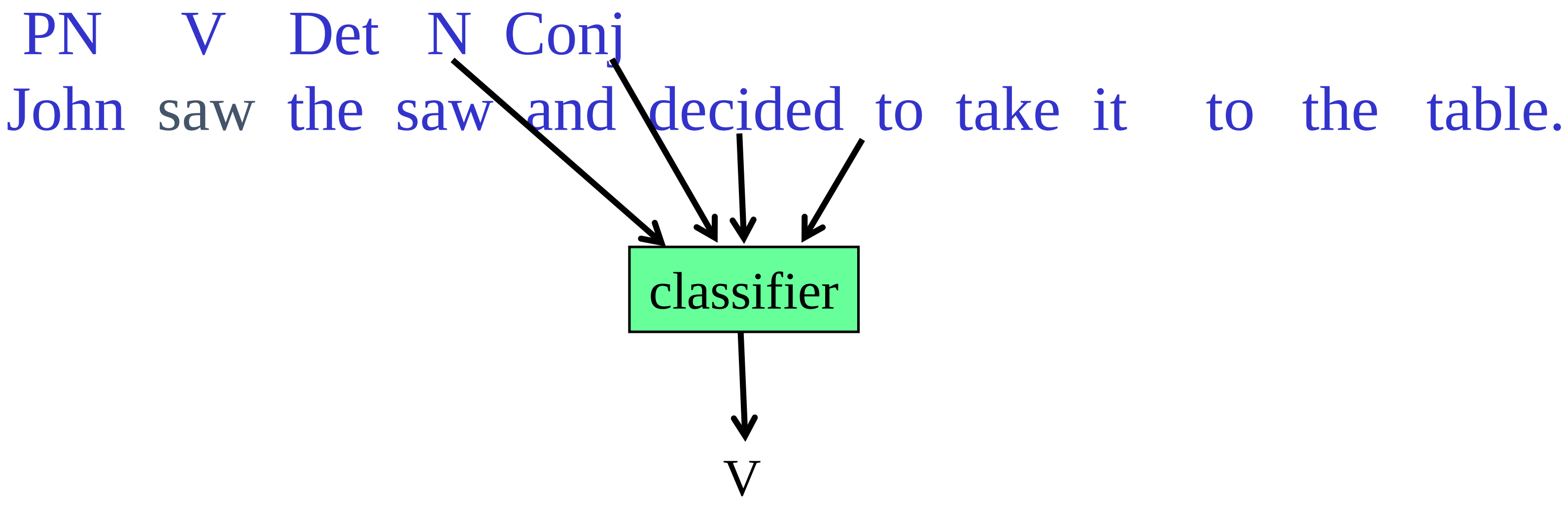
Forward Classification



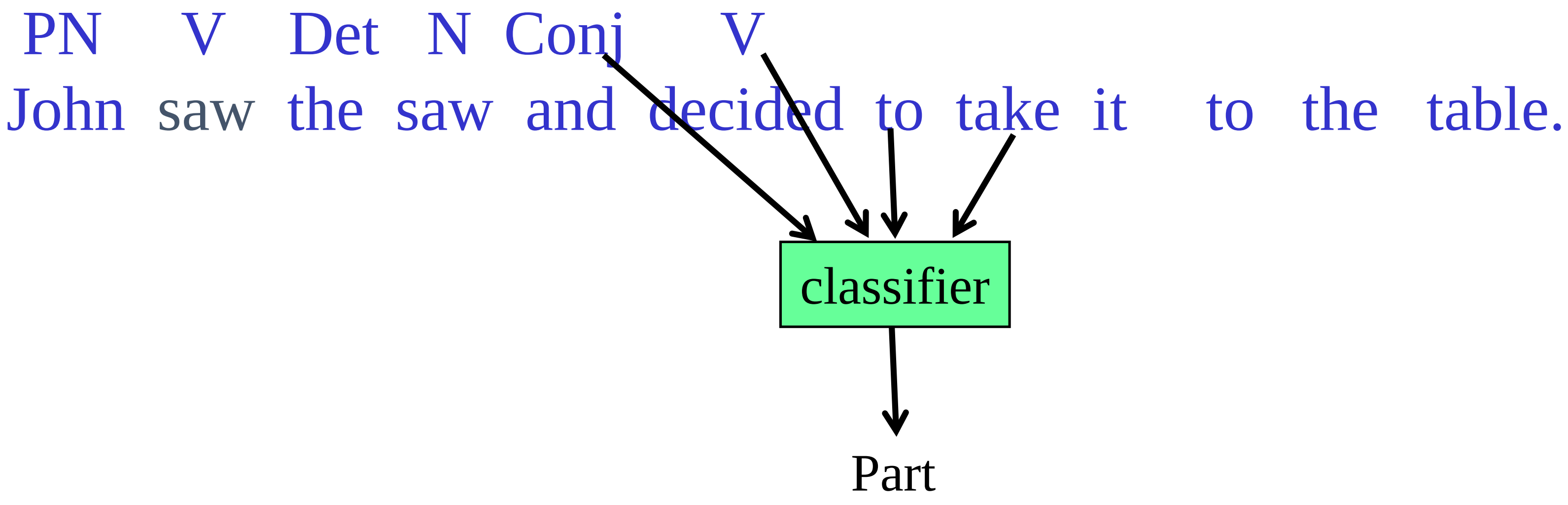
Forward Classification



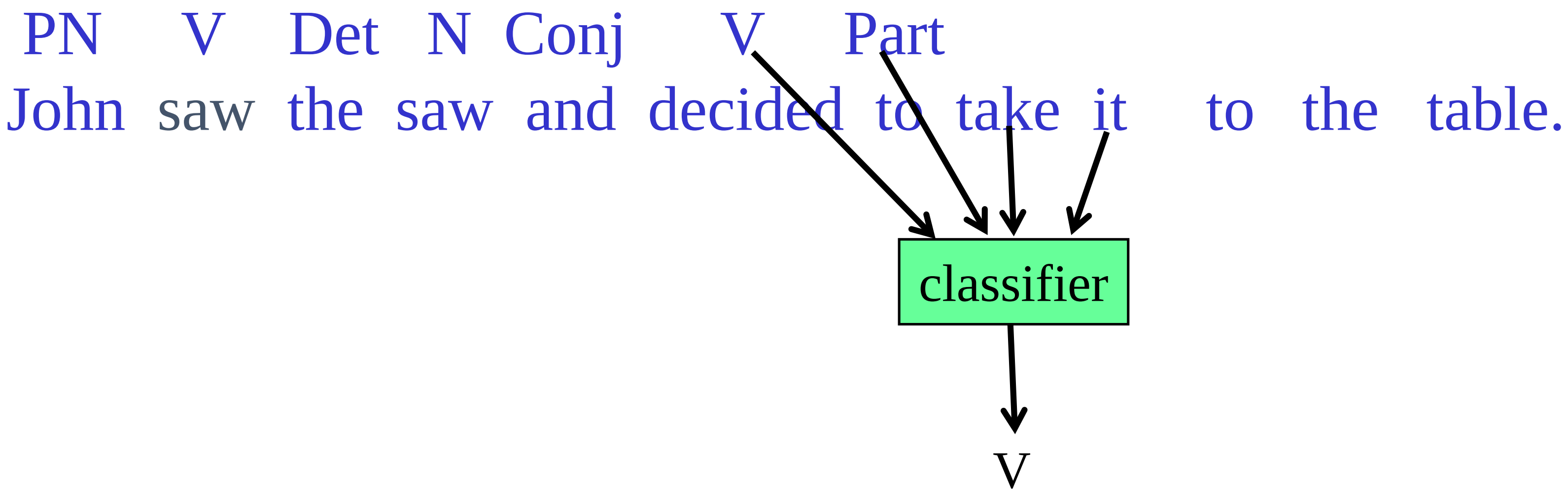
Forward Classification



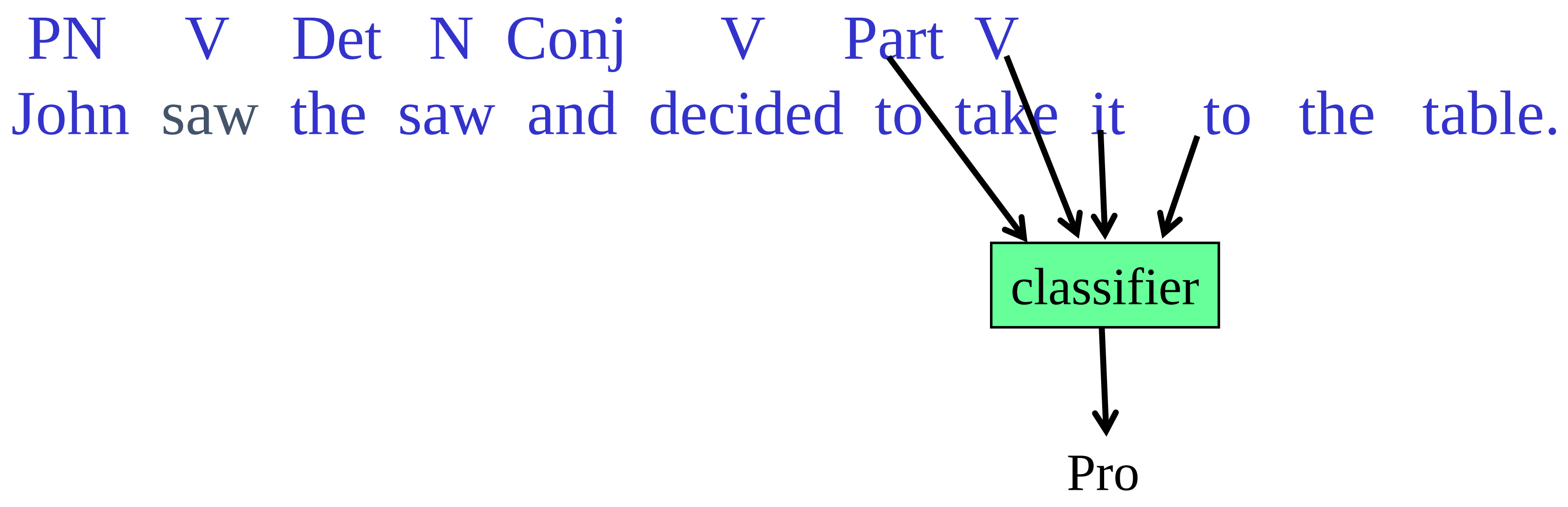
Forward Classification



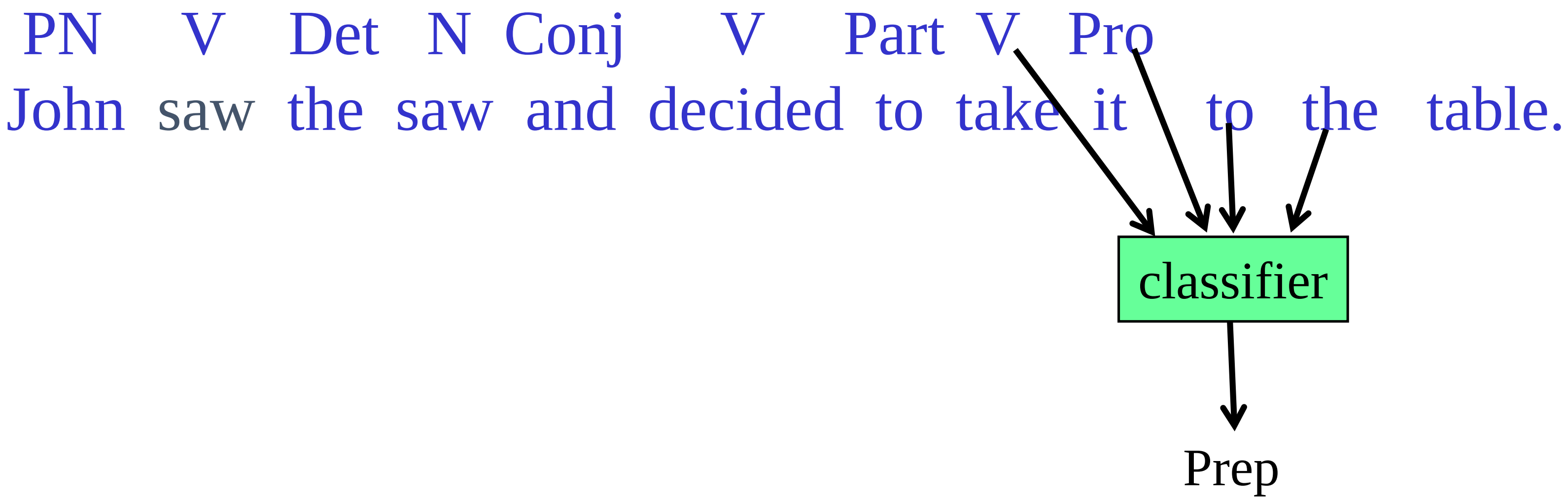
Forward Classification



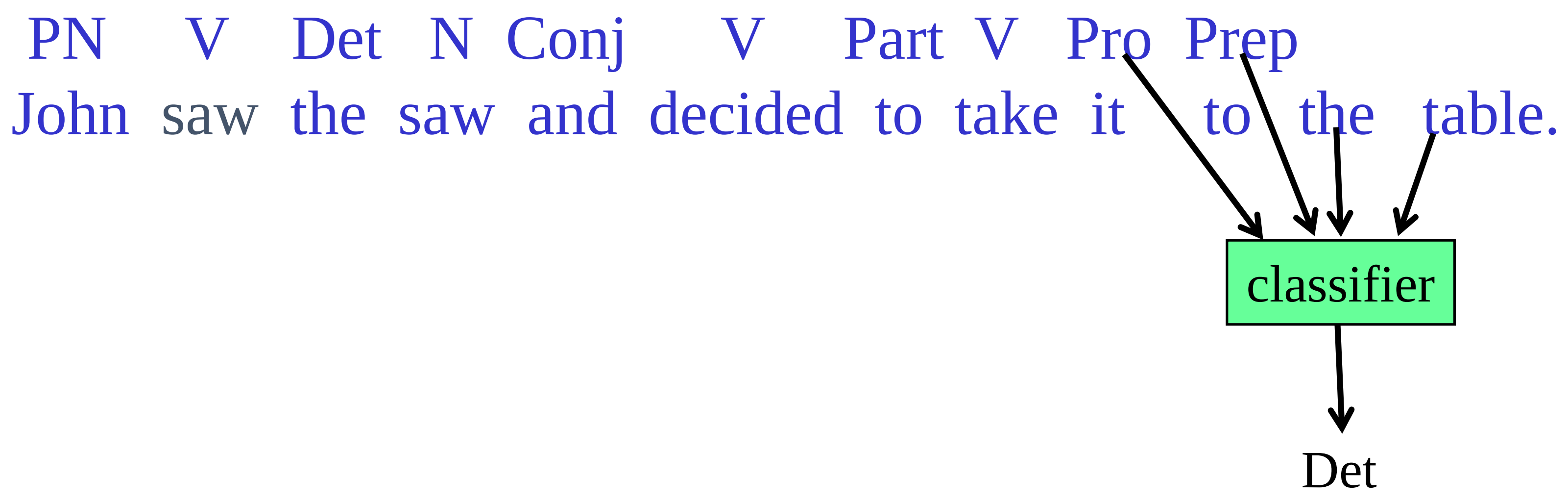
Forward Classification



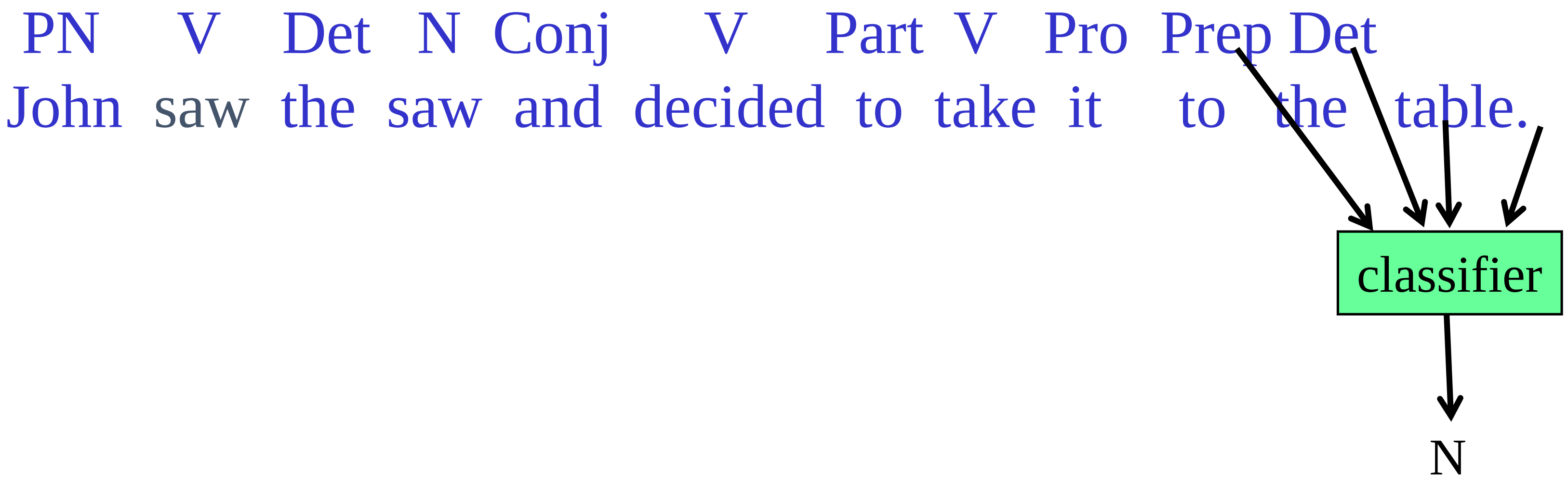
Forward Classification



Forward Classification

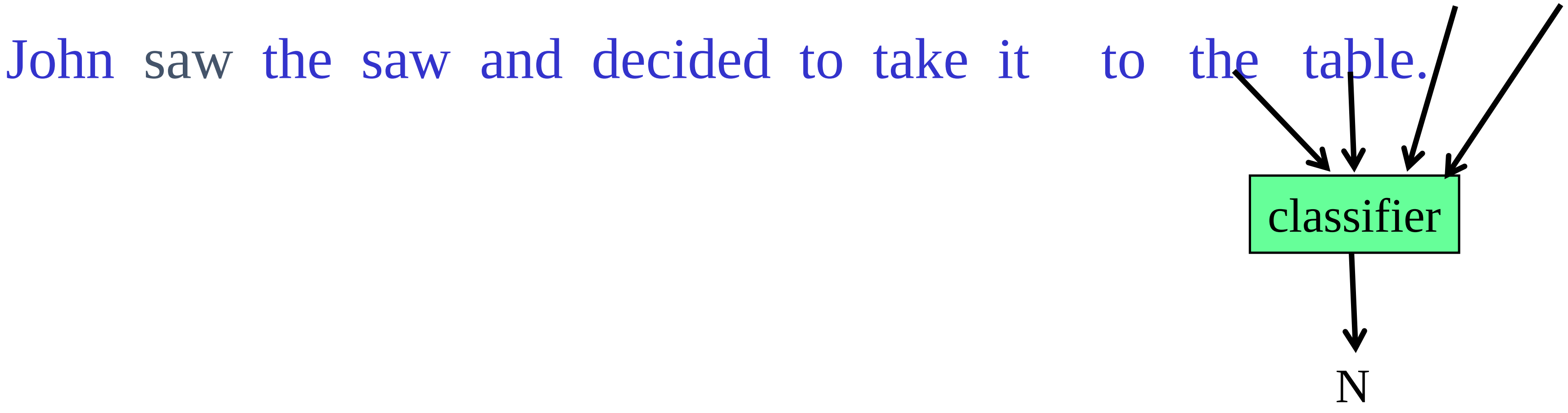


Forward Classification



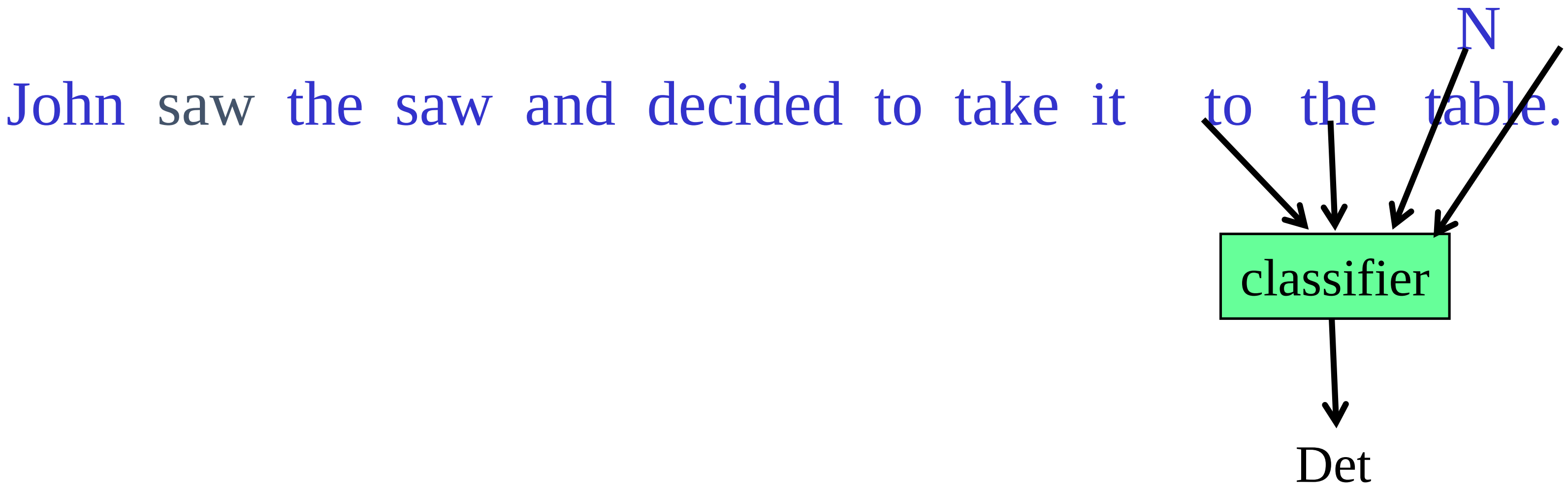
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



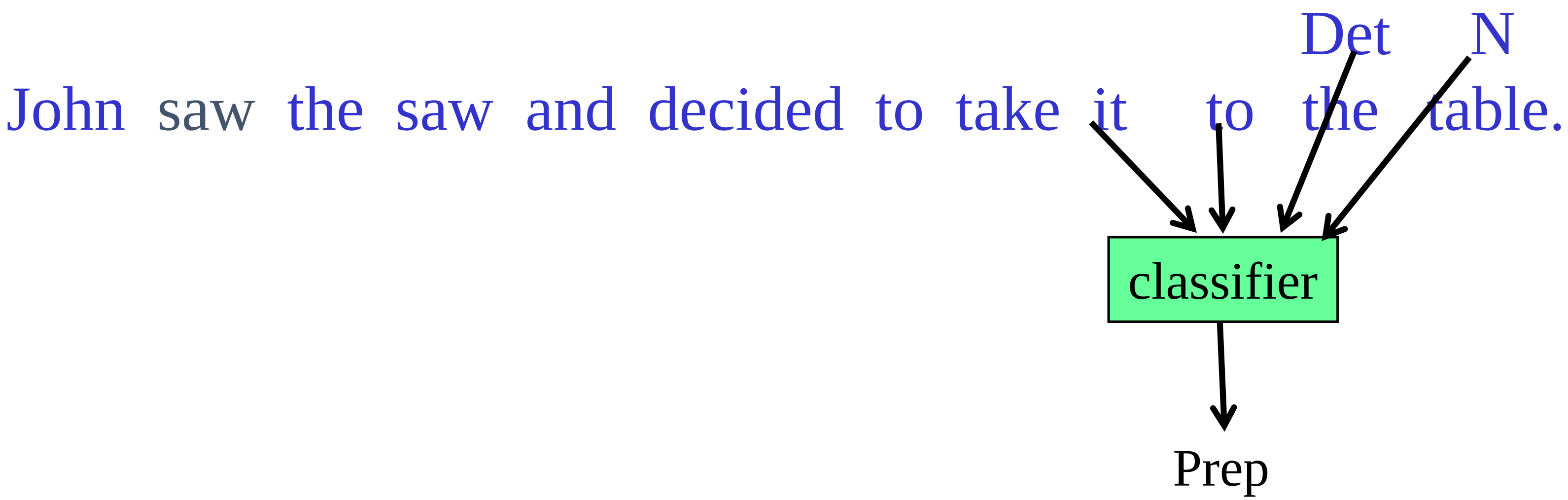
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



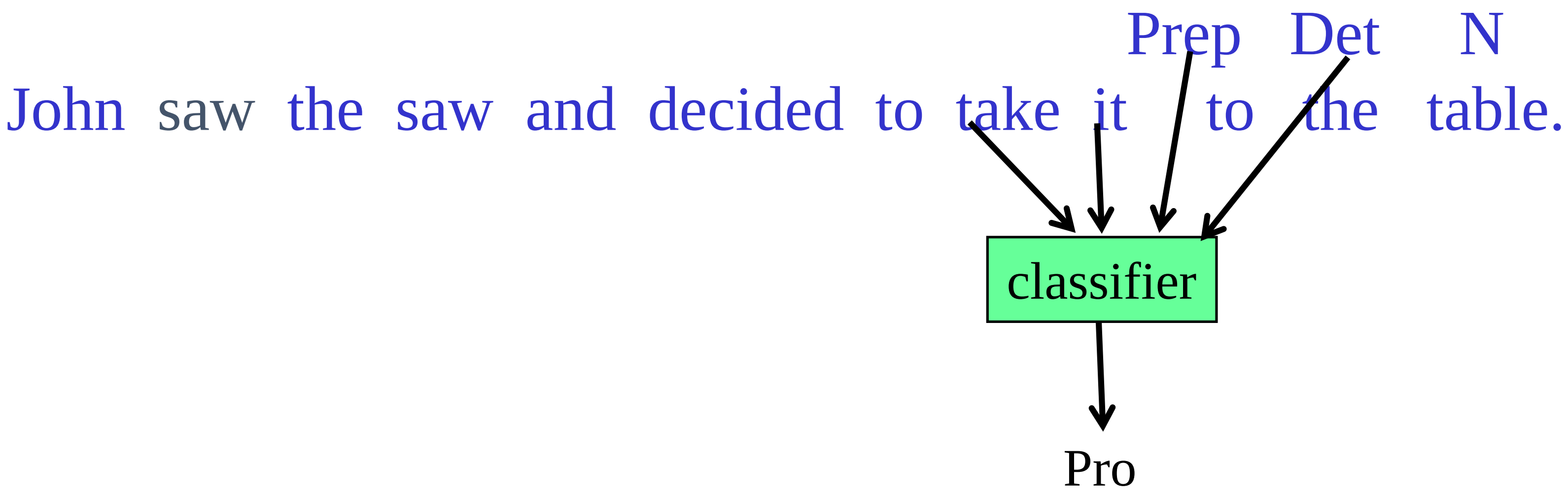
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



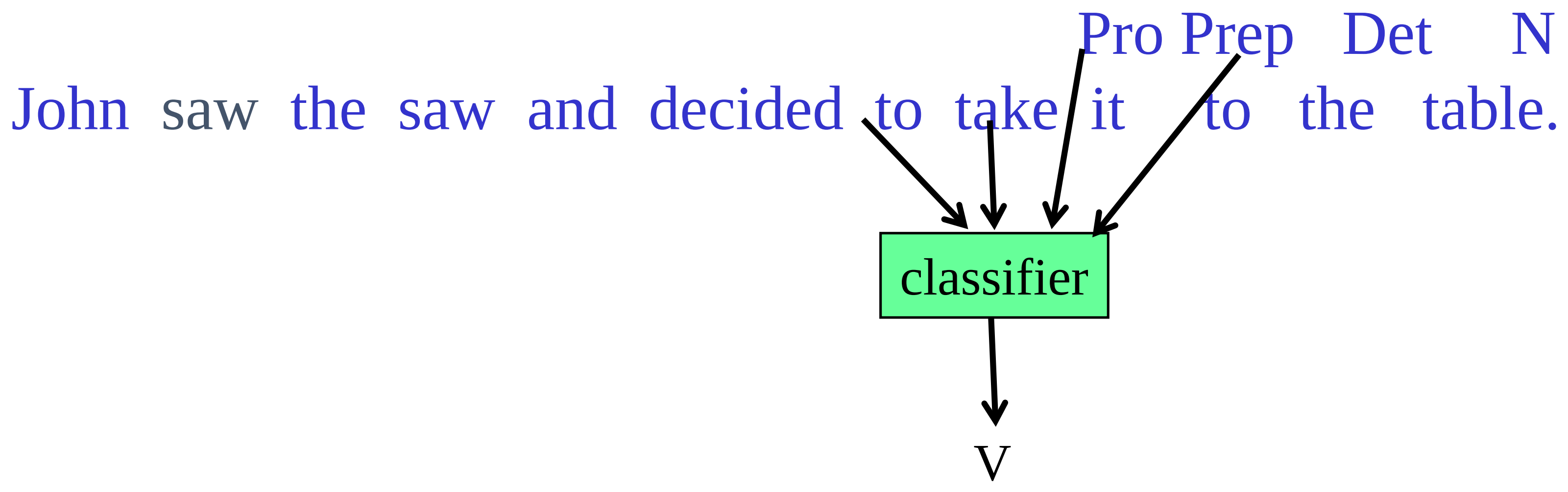
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



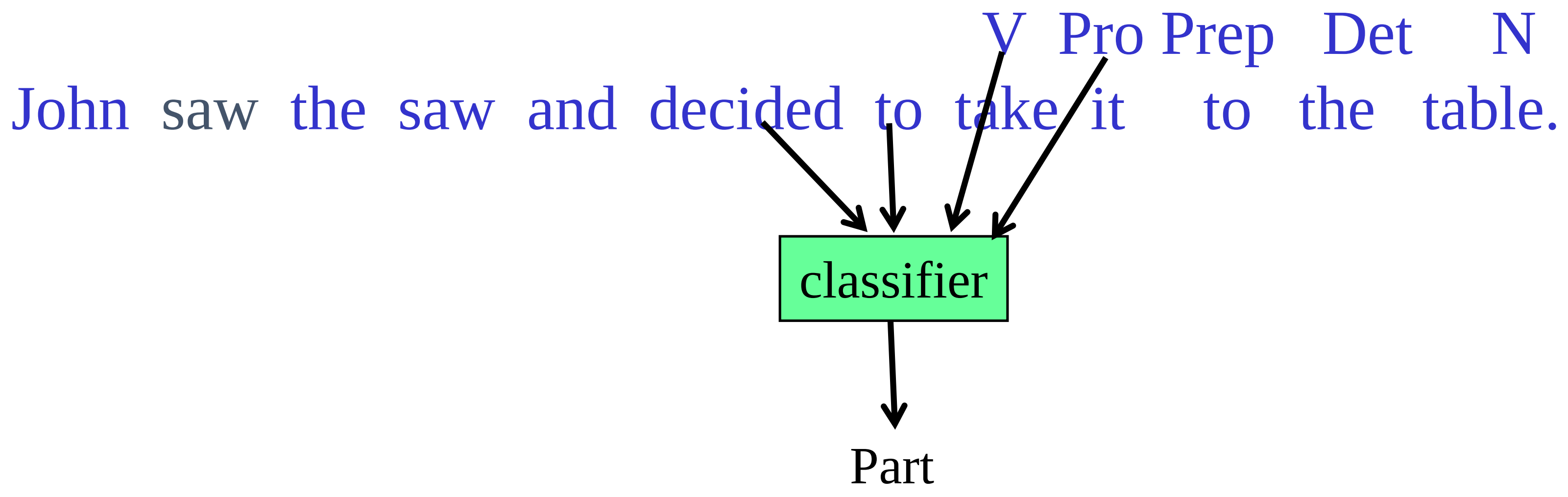
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



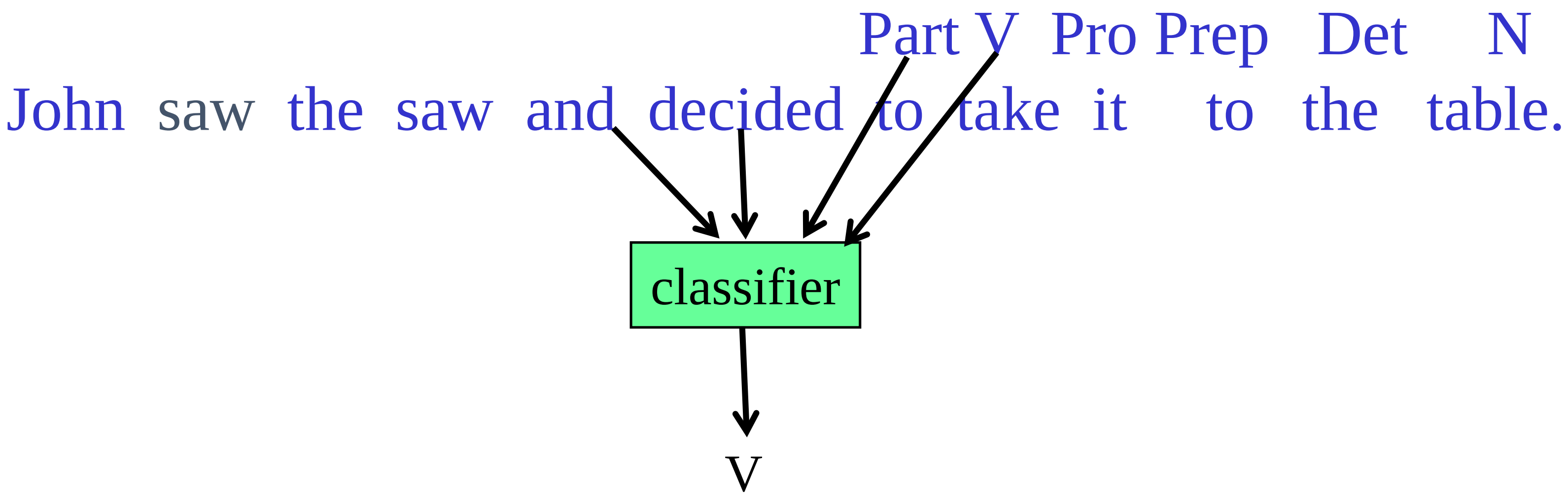
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



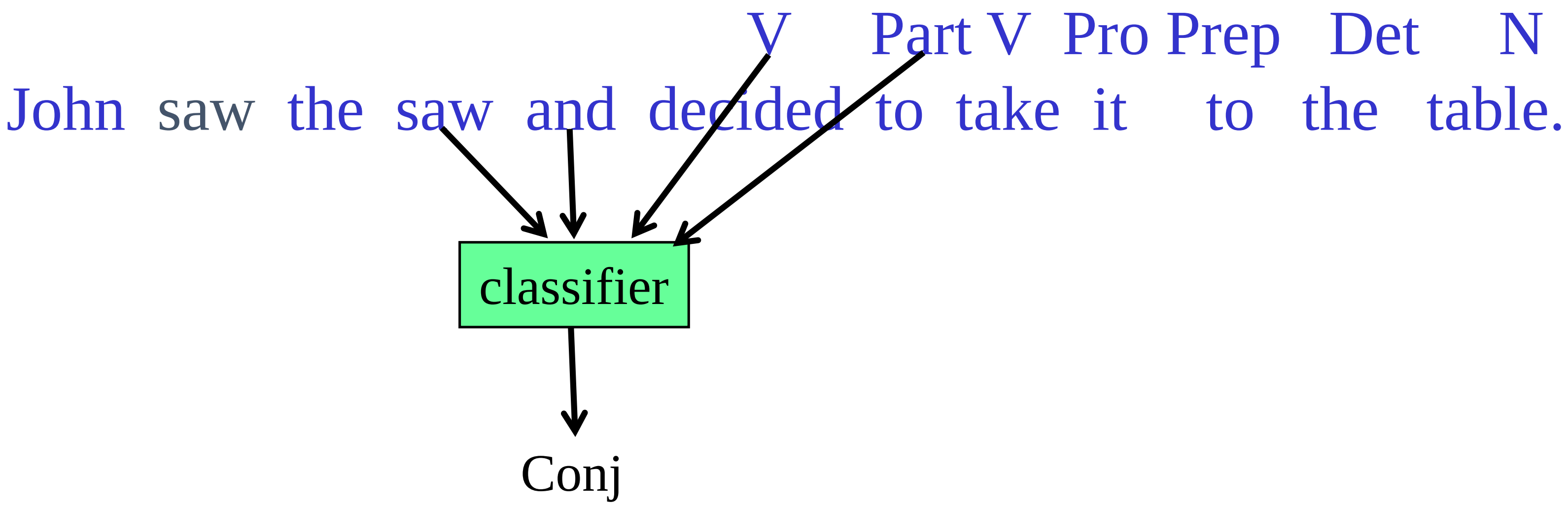
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



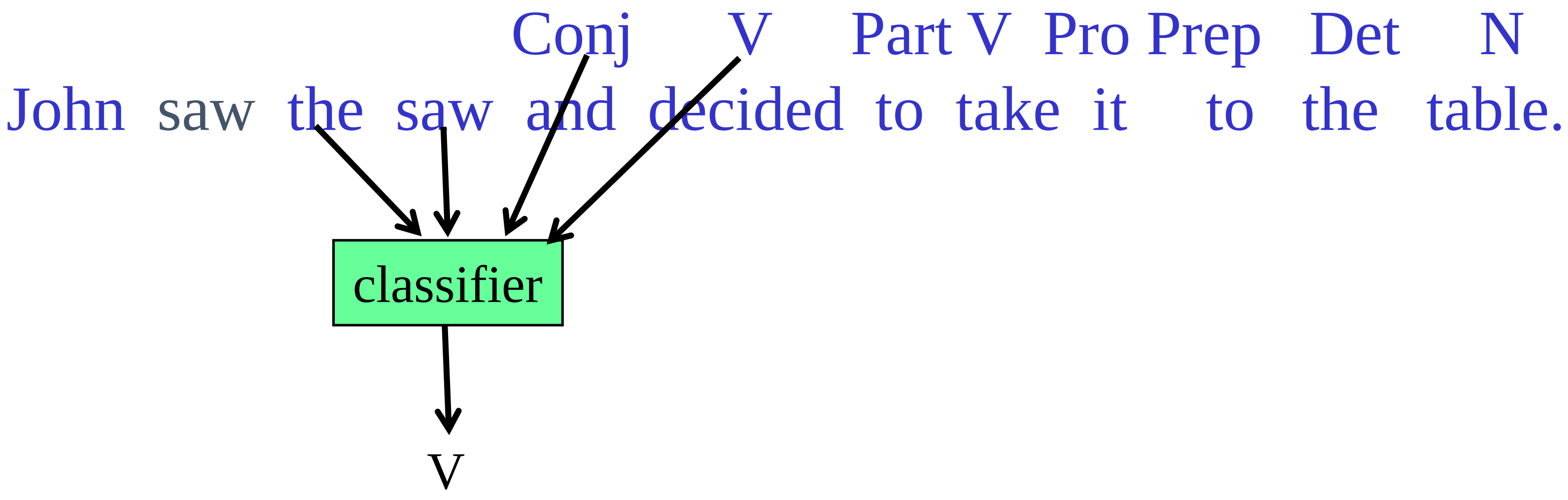
Backward Classification

- Disambiguating "to" in this case would be even easier backward.



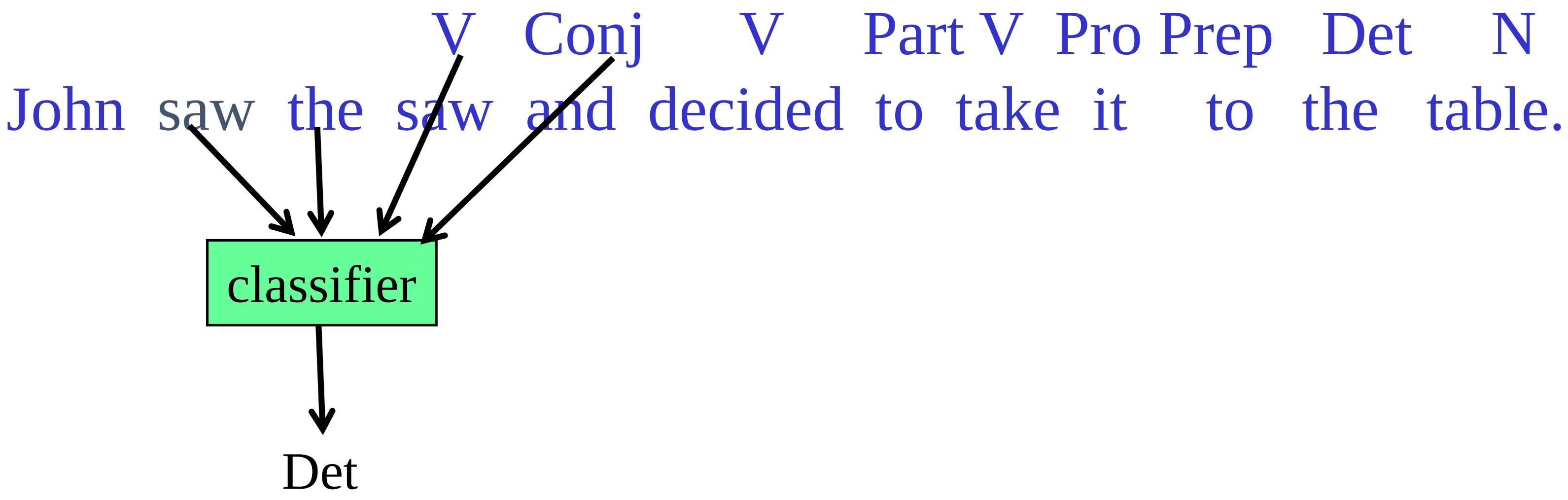
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



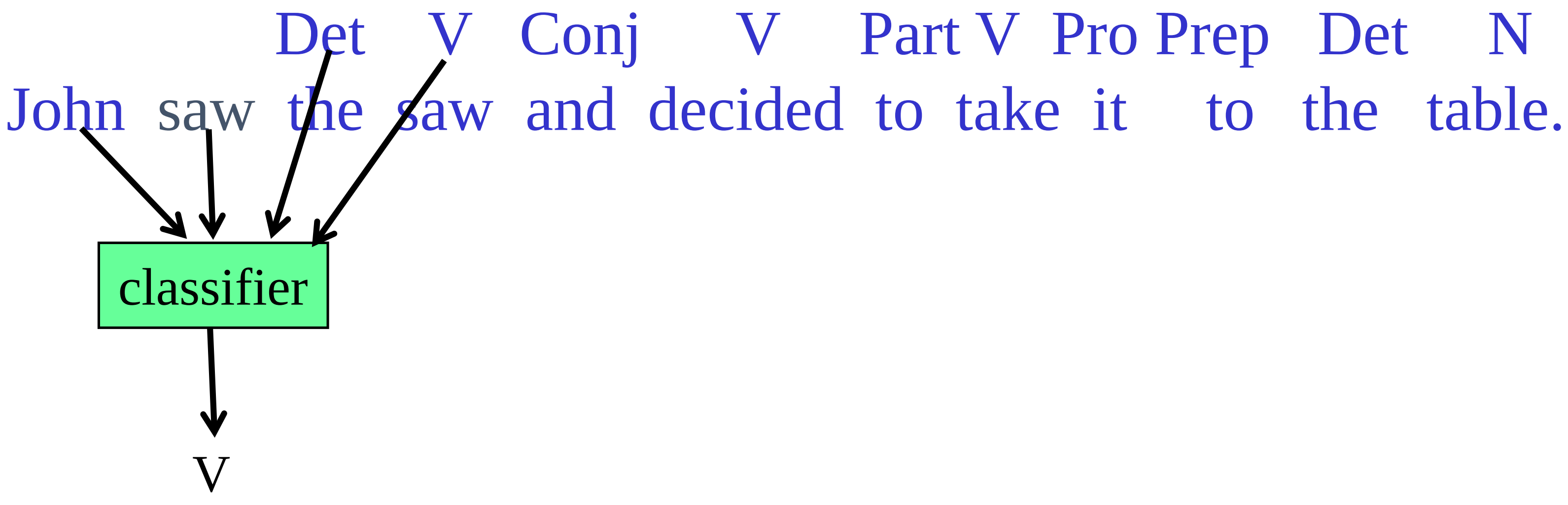
Backward Classification

- Disambiguating “to” in this case would be even easier backward.



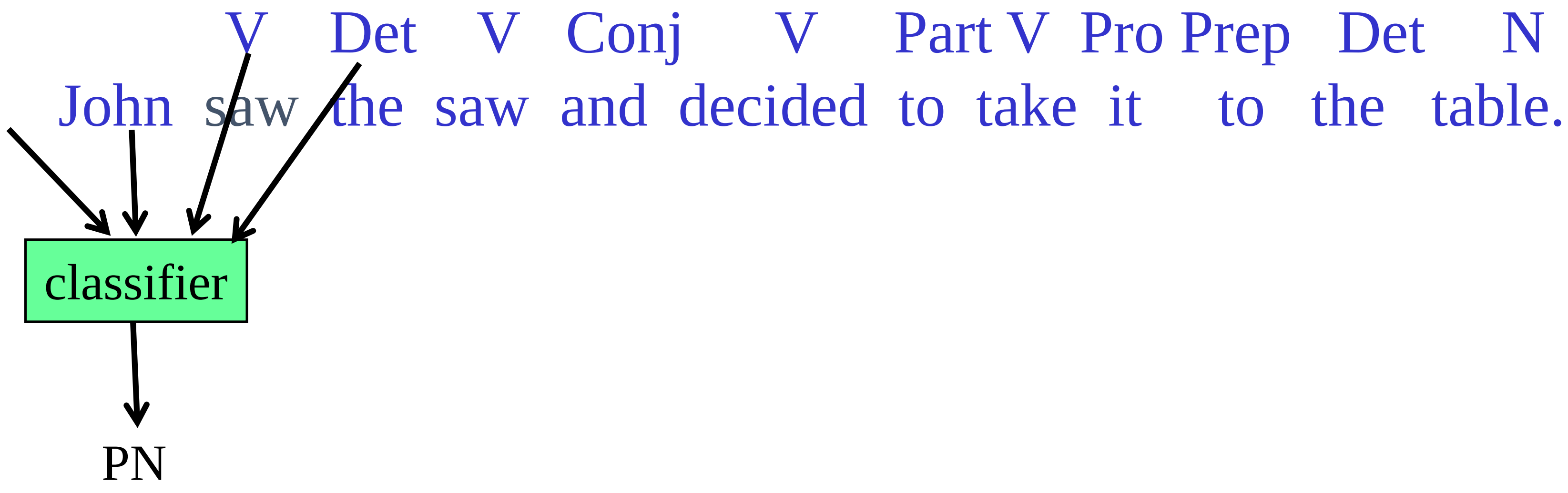
Backward Classification

- Disambiguating “to” in this case would be even easier backward.

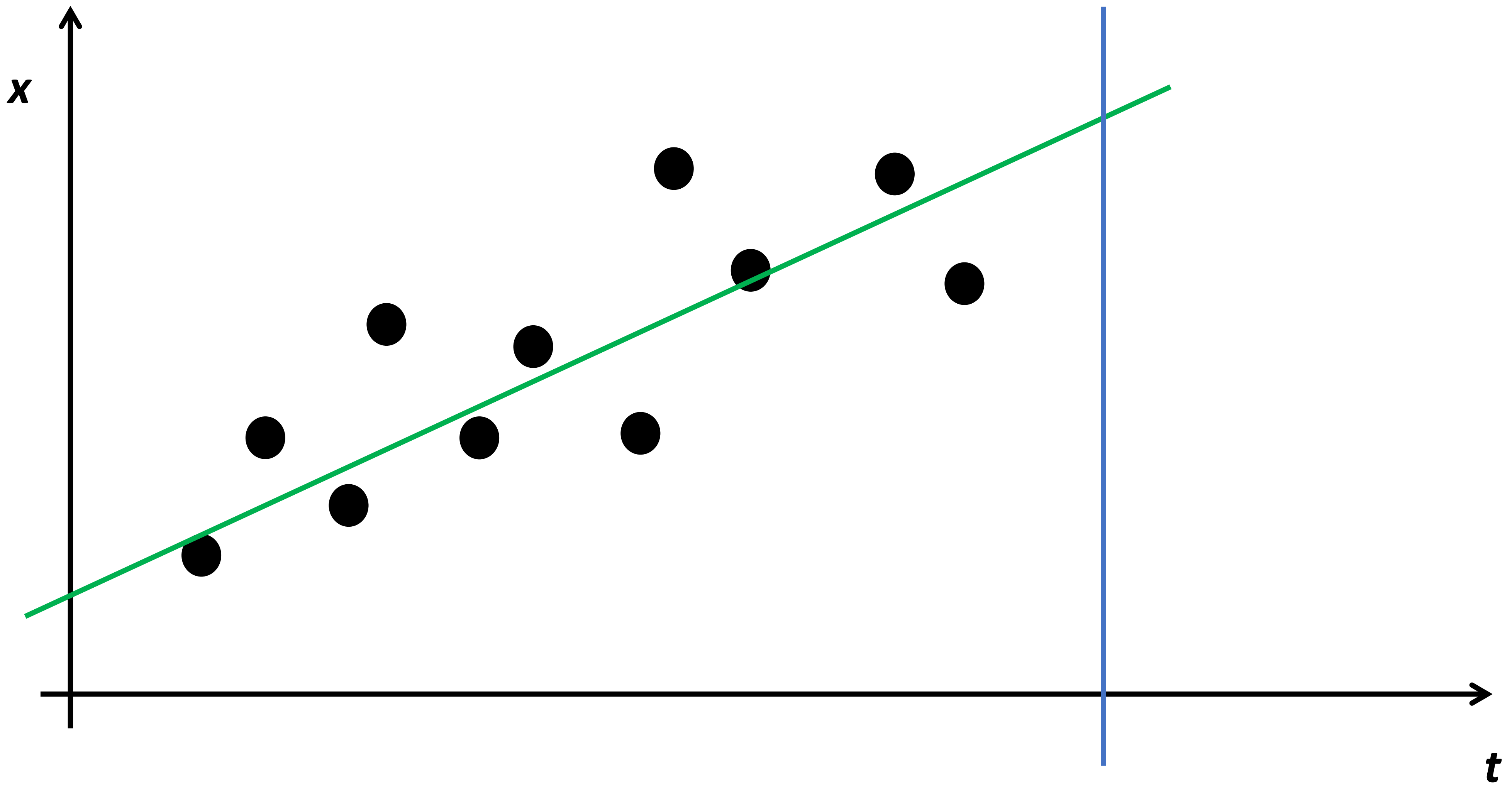


Backward Classification

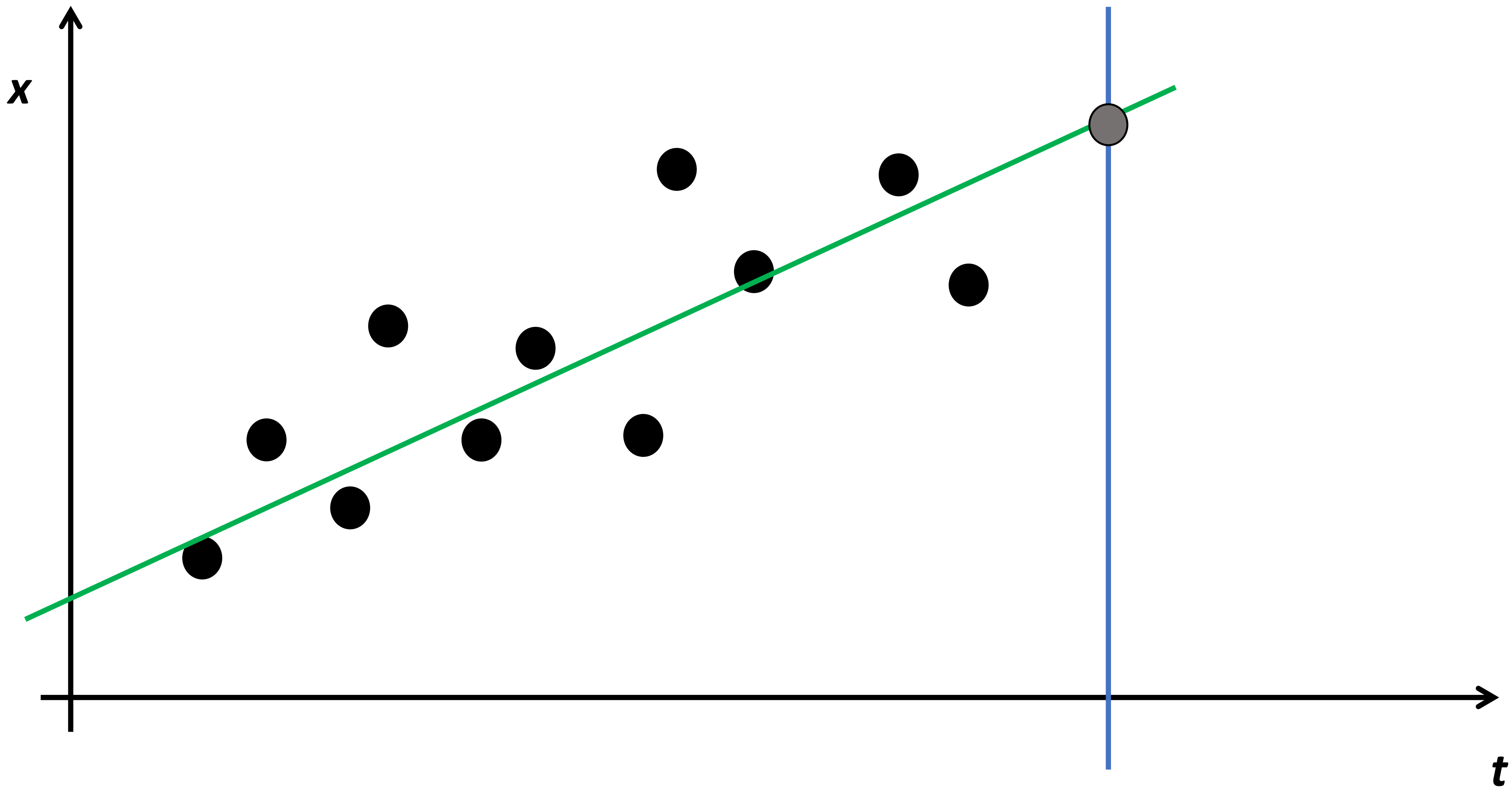
- Disambiguating “to” in this case would be even easier backward.



Regression



Regression



Linear regression

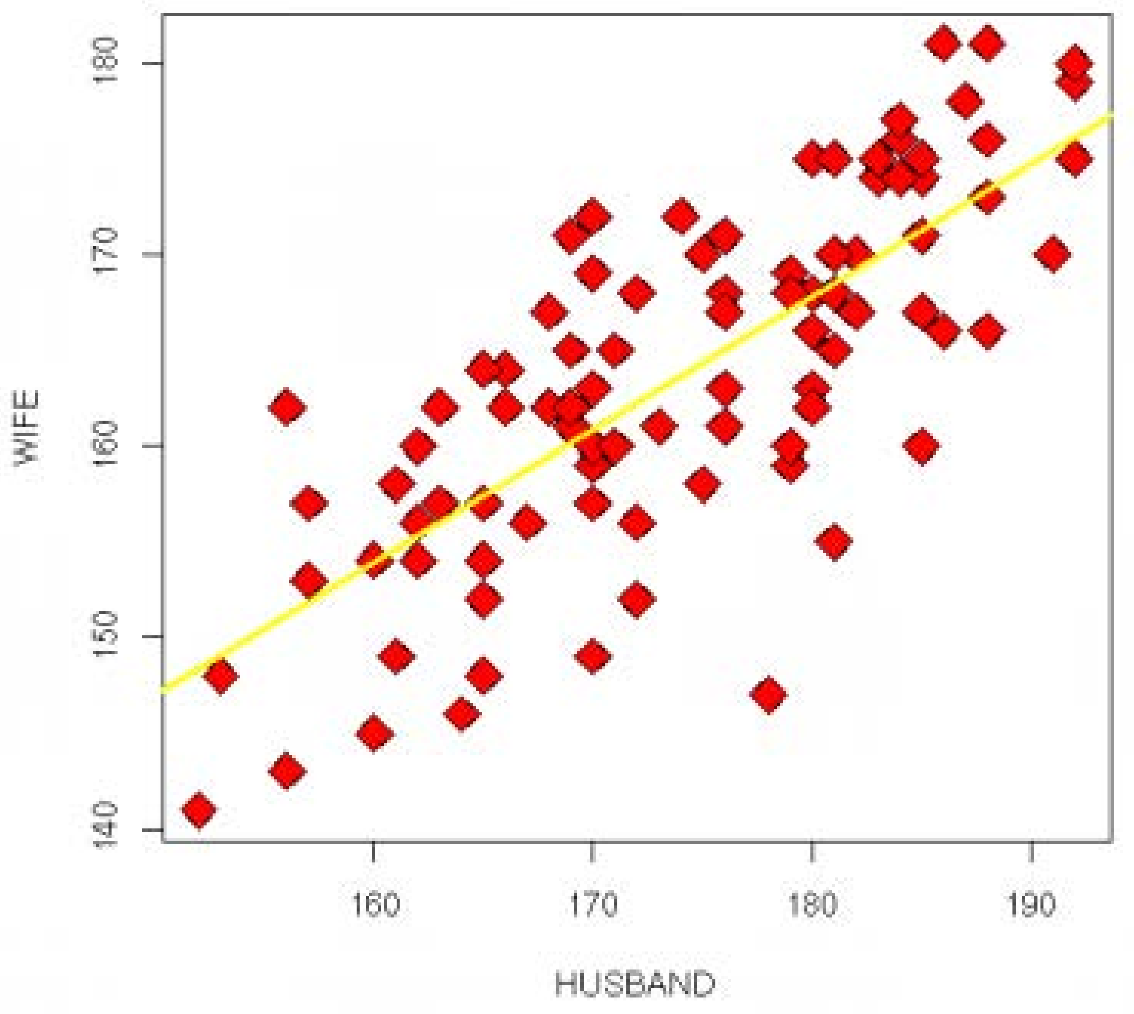
- What is a *regression* model?
  - A regression model is a model of the relationships between some covariates (predictors) and an outcome. Specifically, regression is a model of the average outcome given the covariates

- For height of couples data: a mathematical model, using only Husband's height:

  Wife = f (Husband) + ε

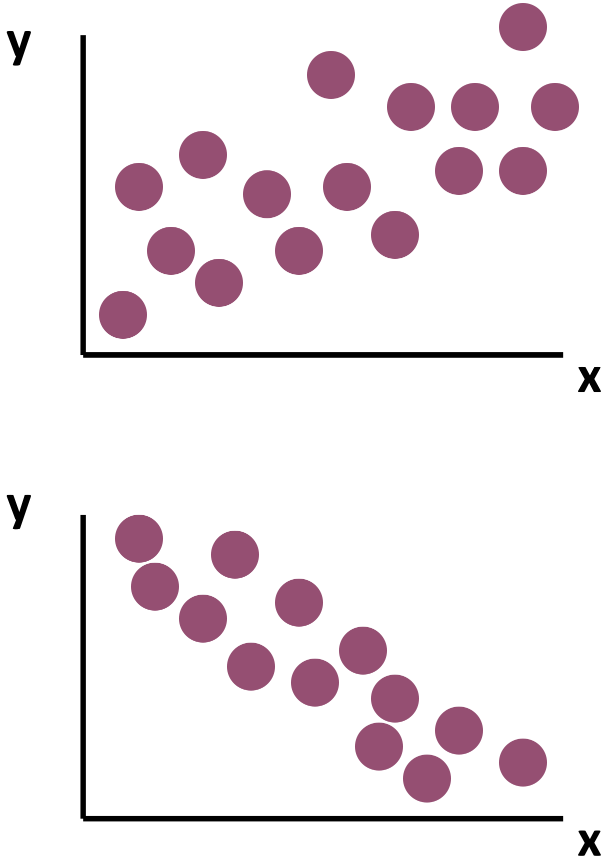
- where f gives the average height of the wife of a man of height Husband and ***ε*** is the random error.

Height data

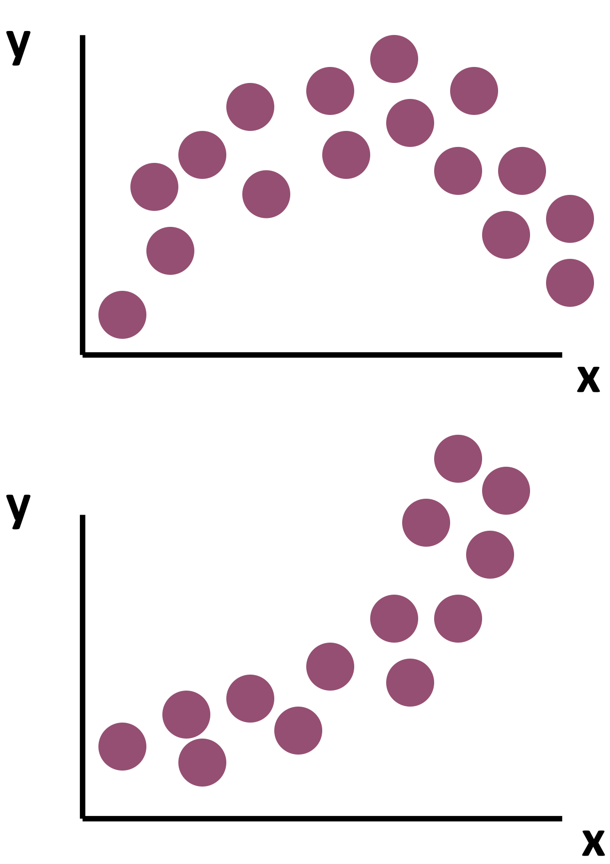


Scatter Plot Examples

Linear relationships

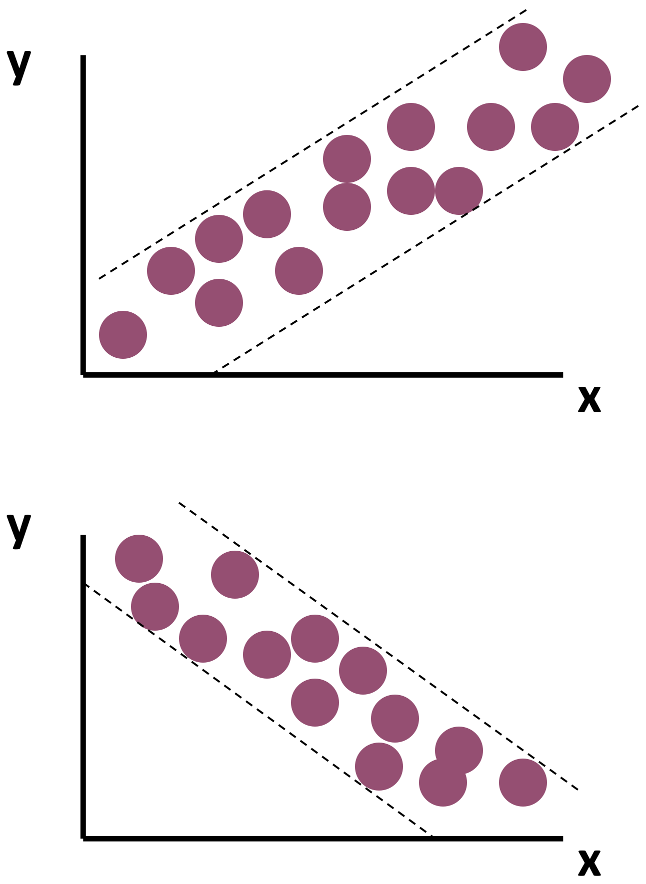


Curvilinear relationships

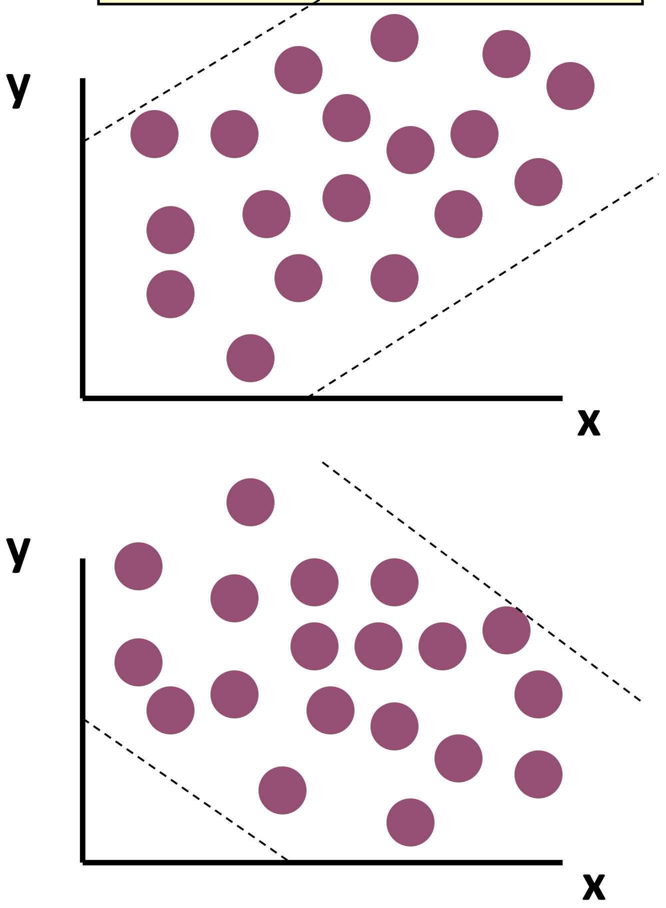


Scatter Plot Examples

Strong relationships

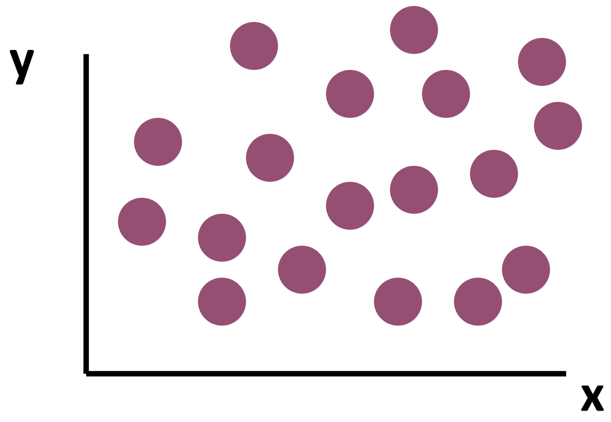


Weak relationships

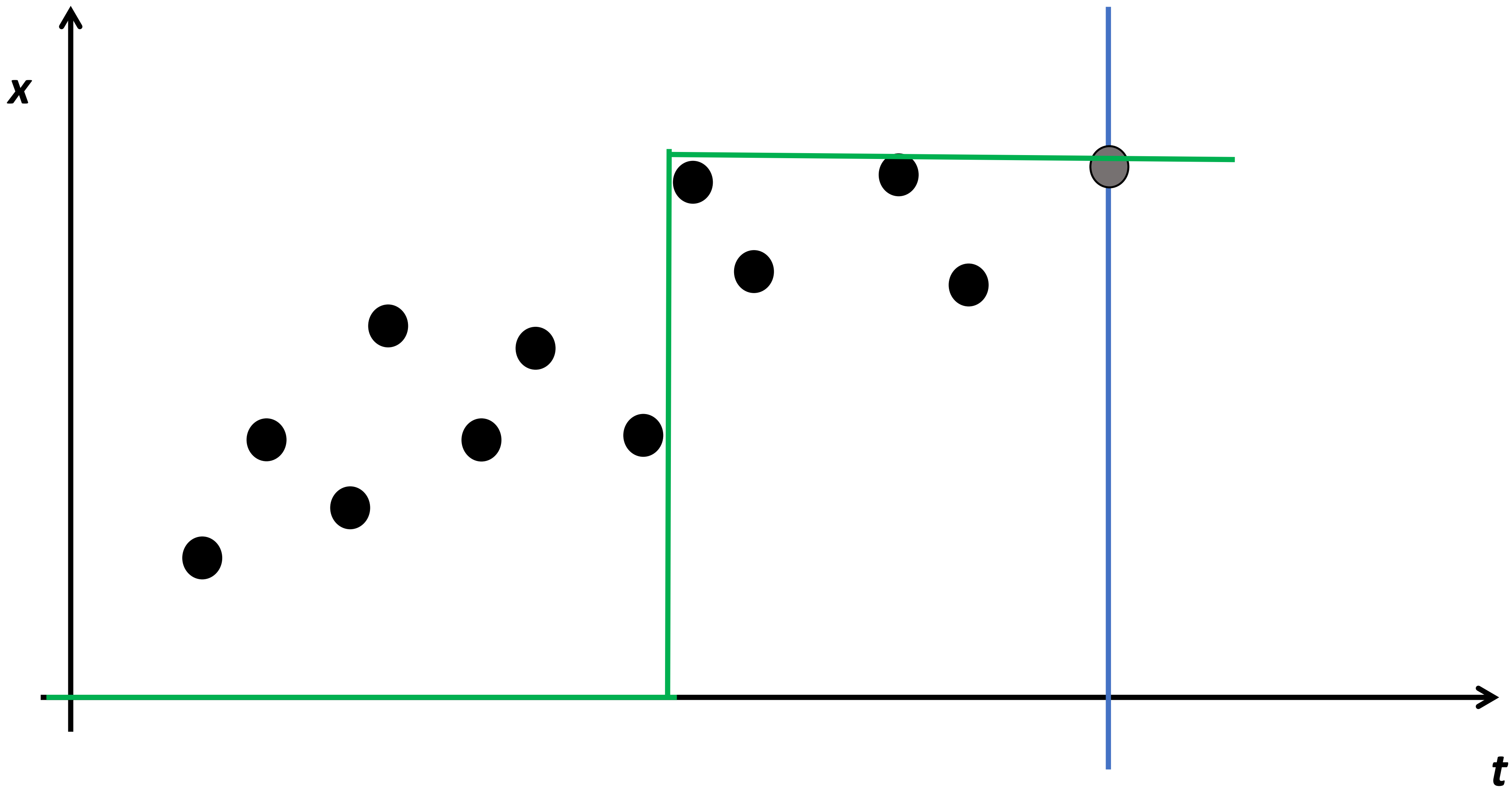


Scatter Plot Examples

No relationship



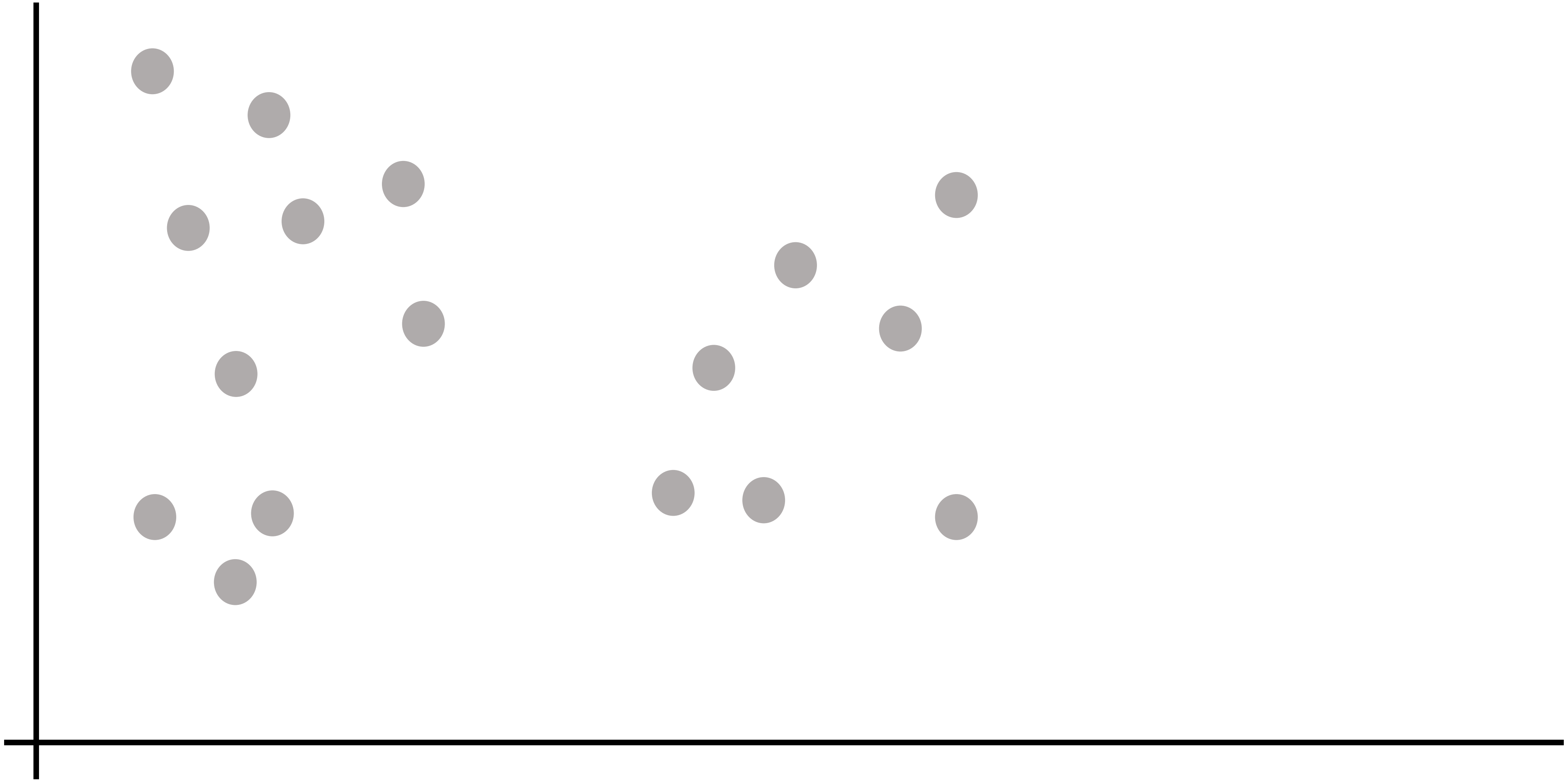
Logistic Regression



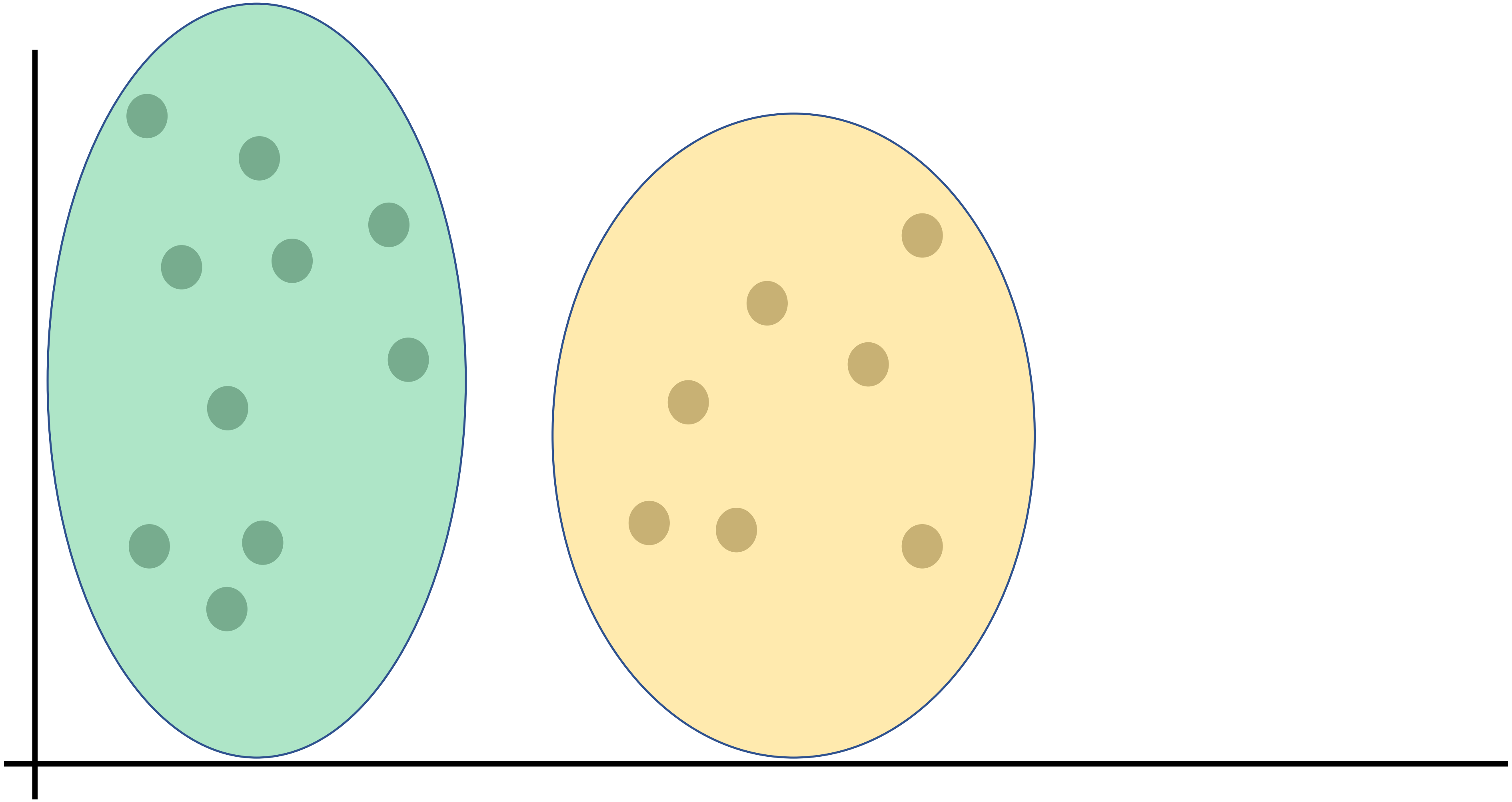
NLP, Text Mining and Regression

- Mainly logistic regression

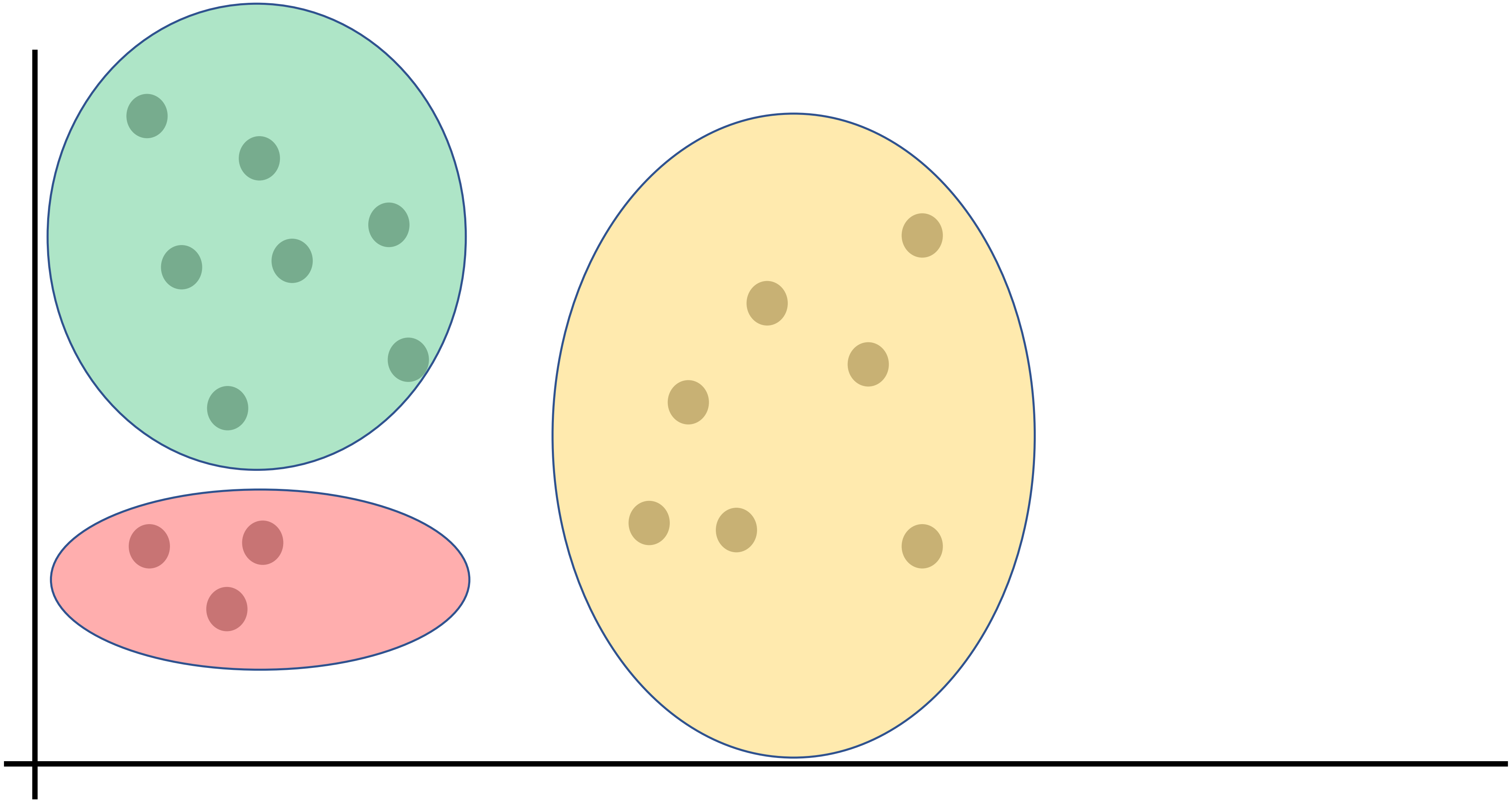
Clustering



Clustering



Clustering

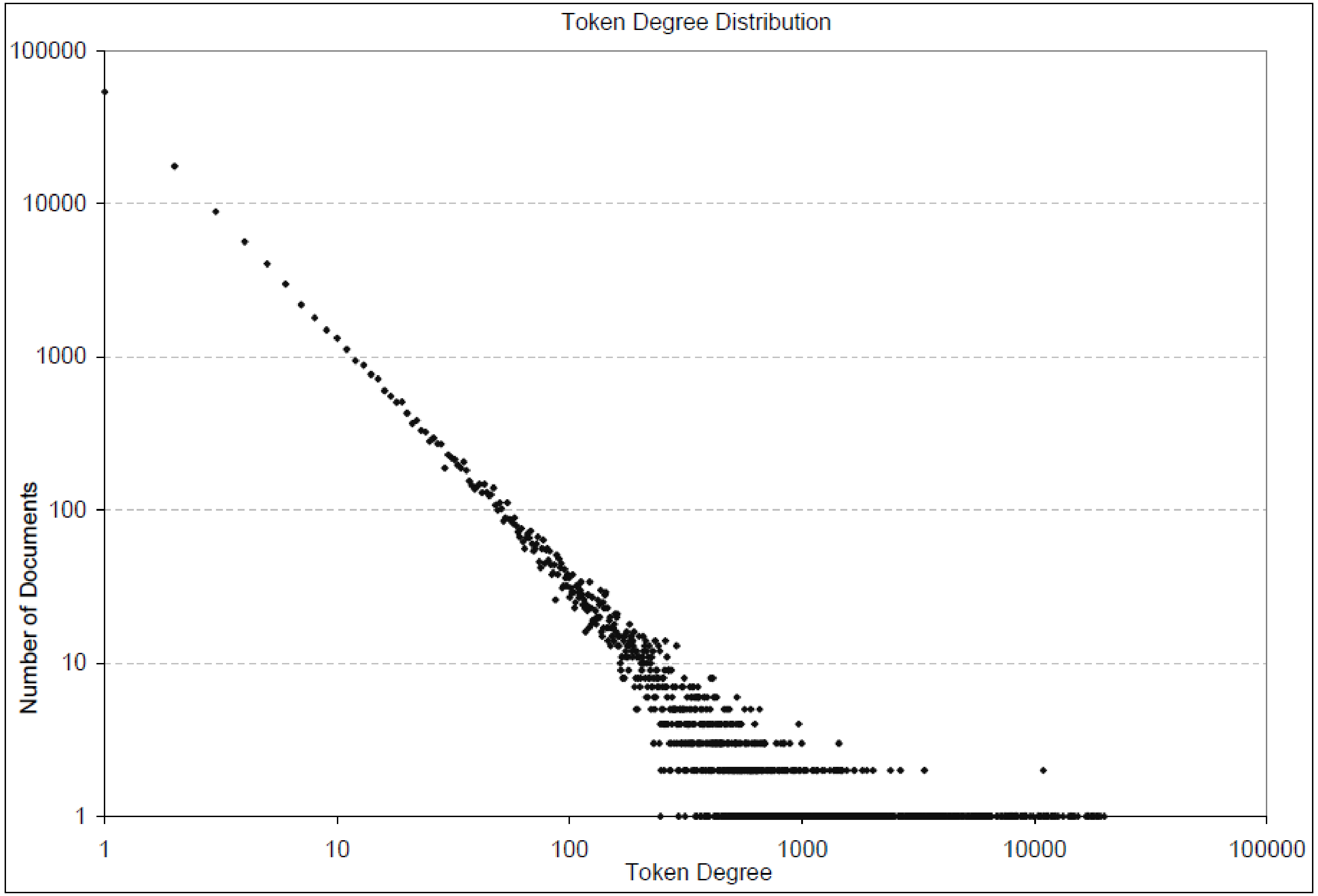


NLP, Text Mining and Clustering

- Grouping documents by topic

Words… A Lot of Them

Words ↔ Documents: Zipf's law: the frequency of any word is inversely proportional to its rank in the frequency table



Data Mining

- Statistical Estimation
- Feature Manipulation
- Similarity Measures

Math in Machine Learning

- Probability
- Statistics
- Calculus
- Vector Calculus
- Linear Algebra

Natural Language Processing

- Question Answering
- Machine Translation
- Sentiment Analysis
- Automatic Summarization
- Information Extraction
- Search
- (Spoken) Dialog Systems

Natural Language Processing =/= How a human process language

NLP Machinery

- Part-of-speech tagging
- Parsing
- Language modeling
- Named-entity recognition
- Coreference Resolution
- Word Sense disambiguation
- Word Representations

Feature Engineering

- The success of machine learning requires instances to be represented using an effective set of features that are correlated with the categories of interest.
- Feature engineering can be a laborious process that requires substantial human expertise and knowledge of the domain.
- In NLP it is common to extract many (even thousands of) potentially features and use a learning algorithm that works well with many relevant and irrelevant features.

Contextual Features

- Surrounding bag of words
- POS of neighboring words
- Local collocations
- Syntactic relations

Experimental evaluations indicate that all of these features are useful; and the best results comes from integrating all of these cues in the disambiguation process.

Data

- Structured data:
  - Wikipedia
  - Google N-grams
  - Yelp
  - Amazon