Overview and Data Science

Brian Wright

brianwright@virginia.edu

Course Administration

- Everybody Reads Even Computers: Text Mining

Final Projects

- Work individually and use one of the areas below to answer a broad questions related to a given dataset. I'll provide several datasets for you to potential use, but you are also welcome to chose your own. You can also use any dataset from the class if you choose.

- Topics we will/have covered that can be a focus of the final project:
  - Data Visualization
  - Fairness/Bias
  - Text Mining
  - KNN
  - Tree based methods
    - Ensemble – Random Forrest – time permitting

Final Projects

- Generate a publishable Rmarkdown document with the following sections:

1. **Question and background** information on the data and why you are asking this question(s). References to previous research/evidence generally would be nice to include.

2. **Exploratory Data Analysis** – Initial summary statistics and graphs with an emphasis on variables you believe to be important for your analysis.

3. **Methods** – Techniques you are using to address your question and the results of those methods.

4. **Conclusions** – What can you say about the results of the methods section as it relates to your question.

5. **Future work** – What additional analysis is needed or what limited your analysis on this project.

“Text Mining”

Broader field: What is Exploratory Text Analytics? (ETA)

Much of the following content is from the Exploratory Text Analytics Class as part of UVA's MSDS taught by Rafael Alvarado

Text Mining

ETA refers to text analytics applied to long-form texts with the purpose of surfacing their latent cognitive, cultural, and social content

Texts: Novels, essays, newspaper articles, letters, blogs, journal articles, etc.

Content: concepts, categories, themes, emotions, events ...

Text Mining

Its called "exploratory" because it methods are primarily **unsupervised** and designed to support human-in-the-loop interpretation

“interpretation support”

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

1.3 Unsupervised learning

We now consider **unsupervised learning**, where we are just given output data, without any inputs. The goal is to discover "interesting structure" in the data; this is sometimes called **knowledge discovery**. Unlike supervised learning, we are not told what the desired output is for each input. Instead, we will formalize our task as one of **density estimation**, that is, we want to build models of the form $p(\mathbf{x}_i|\boldsymbol{\theta})$. There are two differences from the supervised case. First, we have written $p(\mathbf{x}_i|\boldsymbol{\theta})$ instead of $p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$; that is, supervised learning is conditional density estimation, whereas unsupervised learning is unconditional density estimation. Second, $\mathbf{x}_i$ is a vector of features, so we need to create multivariate probability models. By contrast, in supervised learning, y_i is usually just a single variable that we are trying to predict. This means that for most supervised learning problems, we can use univariate probability models (with input-dependent parameters), which significantly simplifies the problem. (We will discuss multi-output classification in Chapter 19, where we will see that it also involves multivariate probability models.)

From Kevin Murphy, 2012, *Machine Learning: A Probabilistic Perspective,* p. 9.

Unsupervised learning is about knowledge discovery

Text Mining

Some Unsupervised Methods:

Clustering — K-means, hierarchical, etc.

Topic Modeling — PCA, LSI/A, NMF, LDA, etc.

Word Embedding — SGNS (word2vec), etc.

Sentiment Analysis -- dictionary-based methods, etc.

Text Mining: Applications of ETA

Extracting features from unstructured data to support **machine learning**

E.g. principal components are document features

Information retrieval tasks such as document summarization, grouping, classification, and knowledge discovery

Provide data to support language modeling, including grammar, syntax, and pragmatics for **NLP** and **computational linguistics**

Extraction and representation of cultural and social **patterns** from text —

See **cultural analytics** and **culturomics**

Text Mining: Culturomics

Coined by Harvard researchers Jean-Baptiste **Michel** and Erez Lieberman **Aiden**, who helped create Google's NGram Viewer

Michel and Aiden (2010):

Inferences about culture made from **trends in n-gram usage**

An n-gram is a sequence of n words

Based on **Google Books**

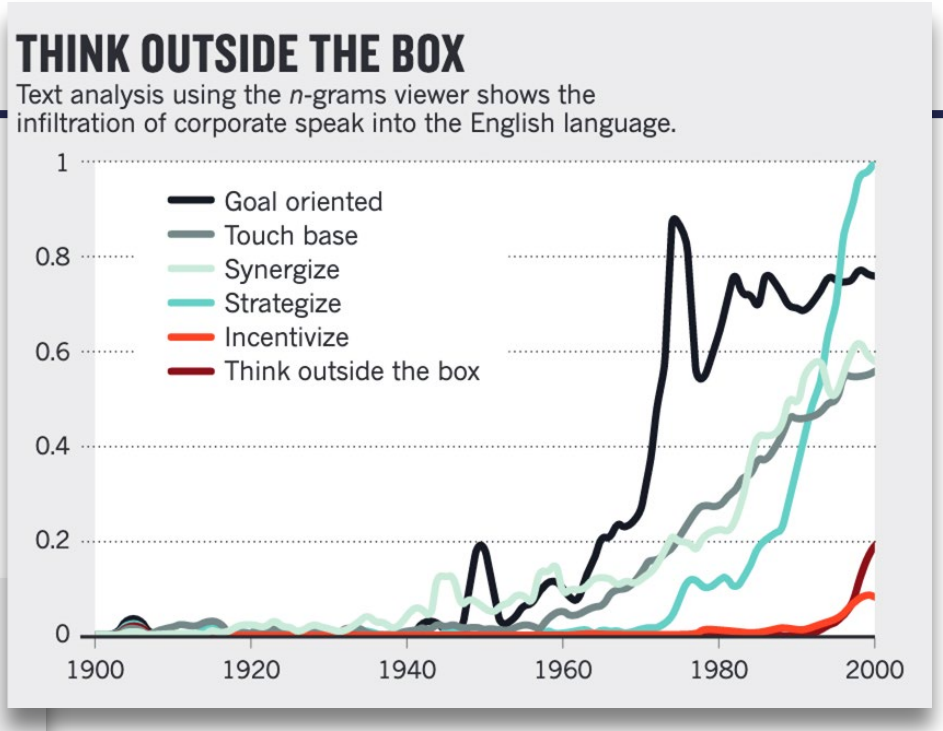
Transformed the field of text analytics

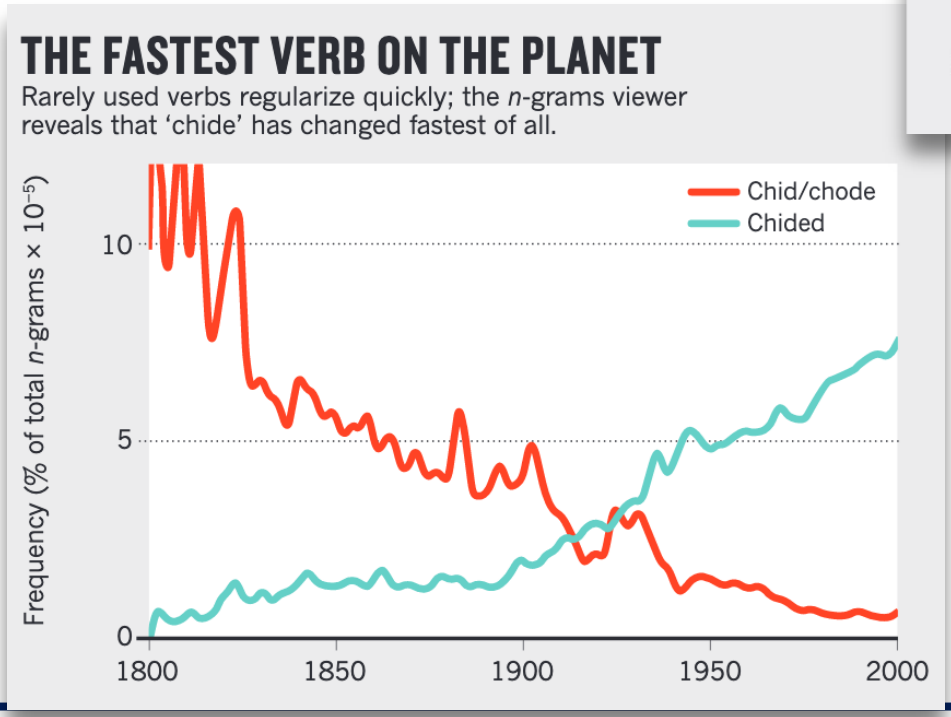
Based on application of **genomic sequencing** techniques to text (See recent book, *Uncharted)*

Text Mining

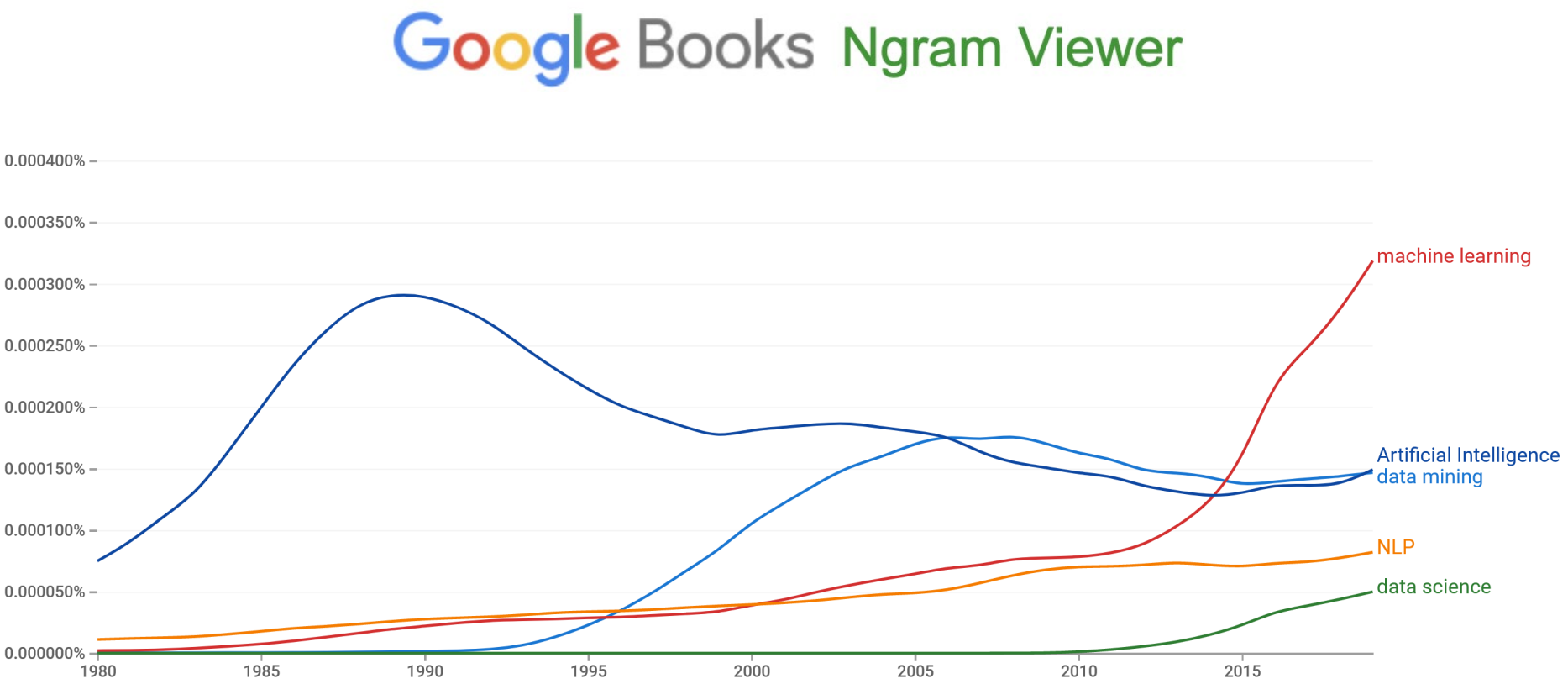
Two examples of inferences drawn from n-gram trends

(Hand 2011)





Text Mining



https://books.google.com/ngrams

Text Mining

ETA builds on the domain knowledge of textual theory and criticism from history, literary studies, anthropology, sociolinguistics, religious studies, etc.

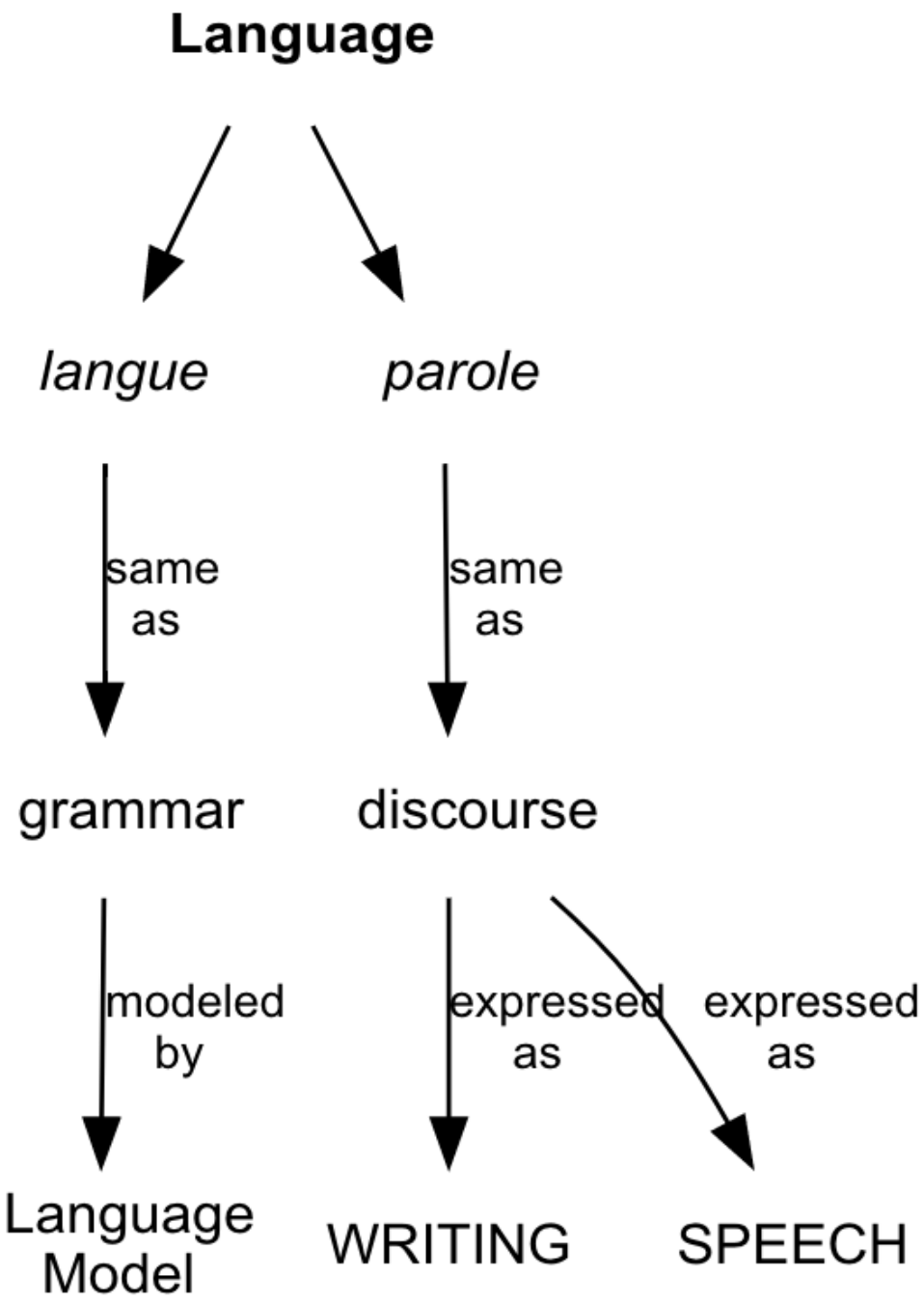
Text is regarded as a first-class object of study, not an incidental container of language data

“We”study text as text

Text is not necessarily language

Text as Text: *Langue* and *Parole*

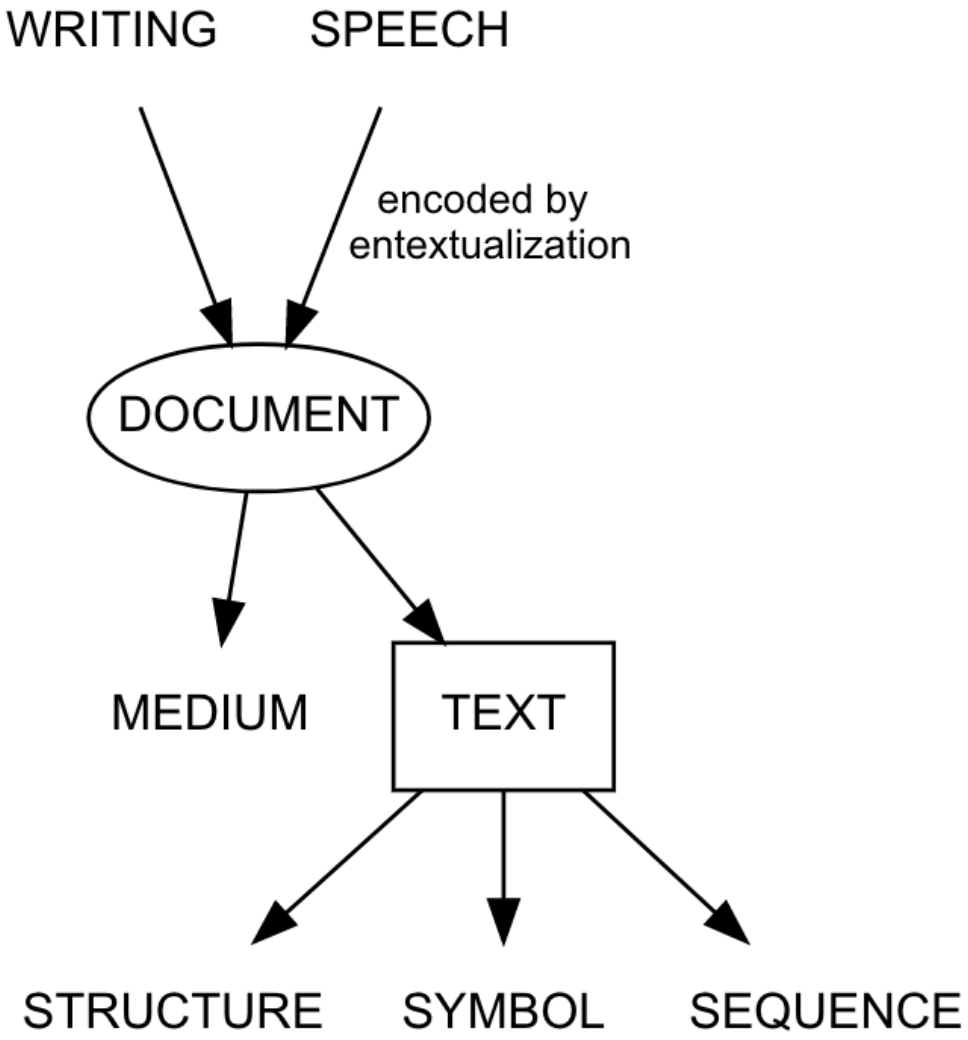
Language (*langue*)	Speech (*parole*)
Grammar	***Discourse***
Competence	Performance
Finite rules (grammar)	***Indefinite patterns (discourse)***
System	Usage
Collective	Individual
Unconscious	Conscious
Structure	Event
Latent	**Observed**



"Language" is divided into grammar and discourse

Discourse is expressed a speech and writing

Writing is "fixed discourse"



Writing is the direct **entextualization** of discourse in a document

Documents have a material form (**medium**) and an "immaterial" dimension -- the text as **structured sequence of symbols**

Text is not "unstructured"!

Text Mining: Some Substantive Properties of Text

Above all, texts contain **cultural information**

- They function as **social genes** that encode and express beliefs, opinions, ideas, symbolism, etc. -- think of Homer, the Bible, etc.
- As discourse, **distinctive of human beings**

Texts may also represent **events**

- Social media and newspapers are like **social sensors**

As more and more social life becomes **entextualized** through social media and other conduits (e.g. Internet of Things)

- Texts contain granular representations of **human behavior**
- **It is the principal means by which behavioral surplus is captured**

So, **culture** is a complex **system** of human **thought** and **behavior** that exhibits a **consistent pattern** in society

It is **expressed** and **communicated** by **symbolic forms**

A **primary vehicle** in our society for the expression and transmission of symbolic forms is the written word — **texts**

A premise of ETA is that **texts "contain" cultural patterns** and these may be discovered through **unsupervised methods**

ETA Related Fields (Antecedents)

Computational Linguistics (CL)
Use of computers to represent and study human language

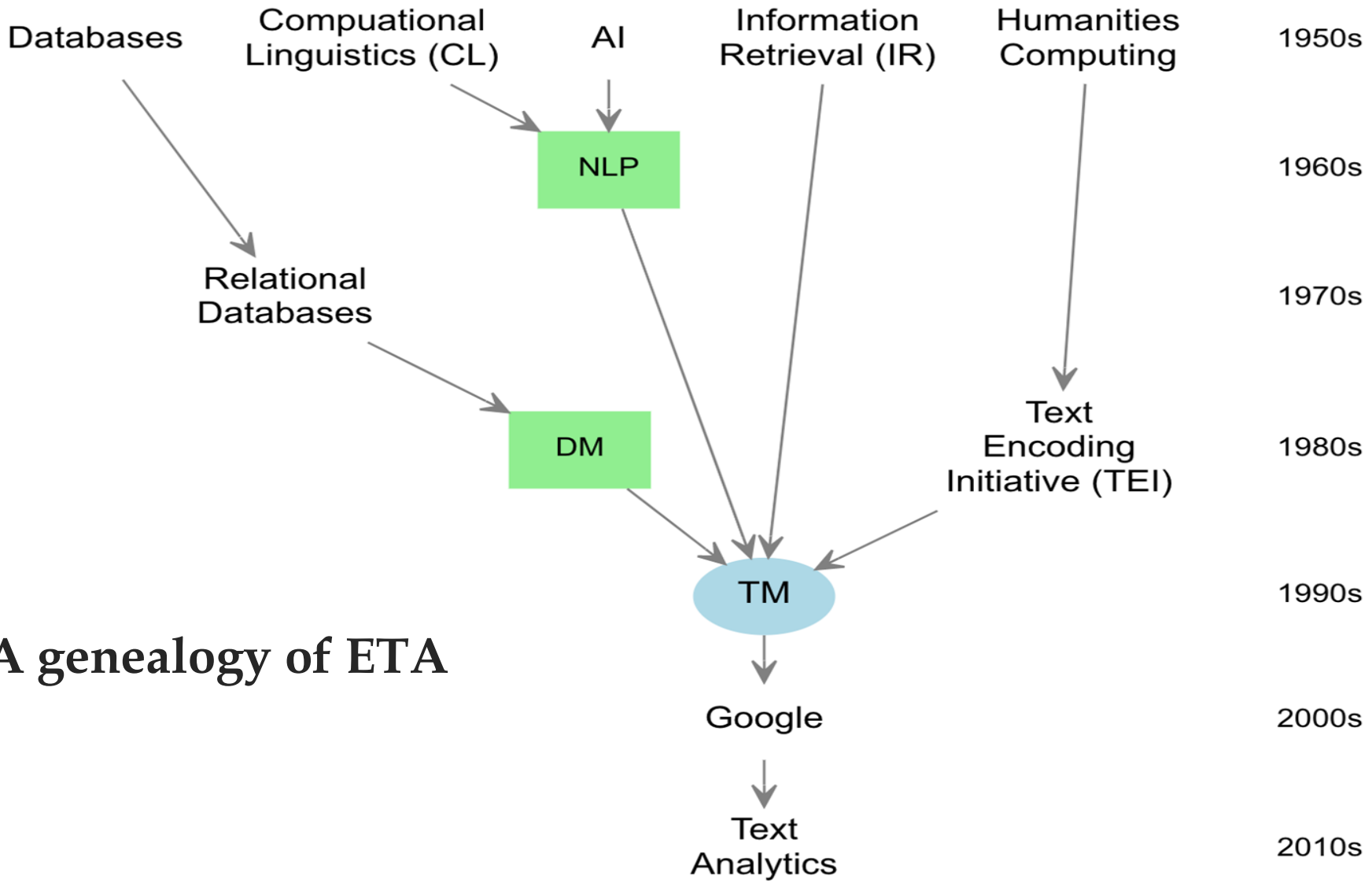
Information Retrieval (IR)
Document summarization, retrieval, indexing, classification based on contents and metadata

Natural Language Processing (NLP)
Get computers to understand and produce human language

Text Mining (TM)
Convert text-as-unstructured-data into features for data mining + ML

Digital Humanities (DH) / Humanities Computing
Create digital collections of primary textual sources and new forms of scholarship → 1949 Father Busa's *Index Thomisticus*

A genealogy of ETA



Text Mining

Note that **text mining** (TM)
and **natural language processing** (NLP)
are **not the same thing**

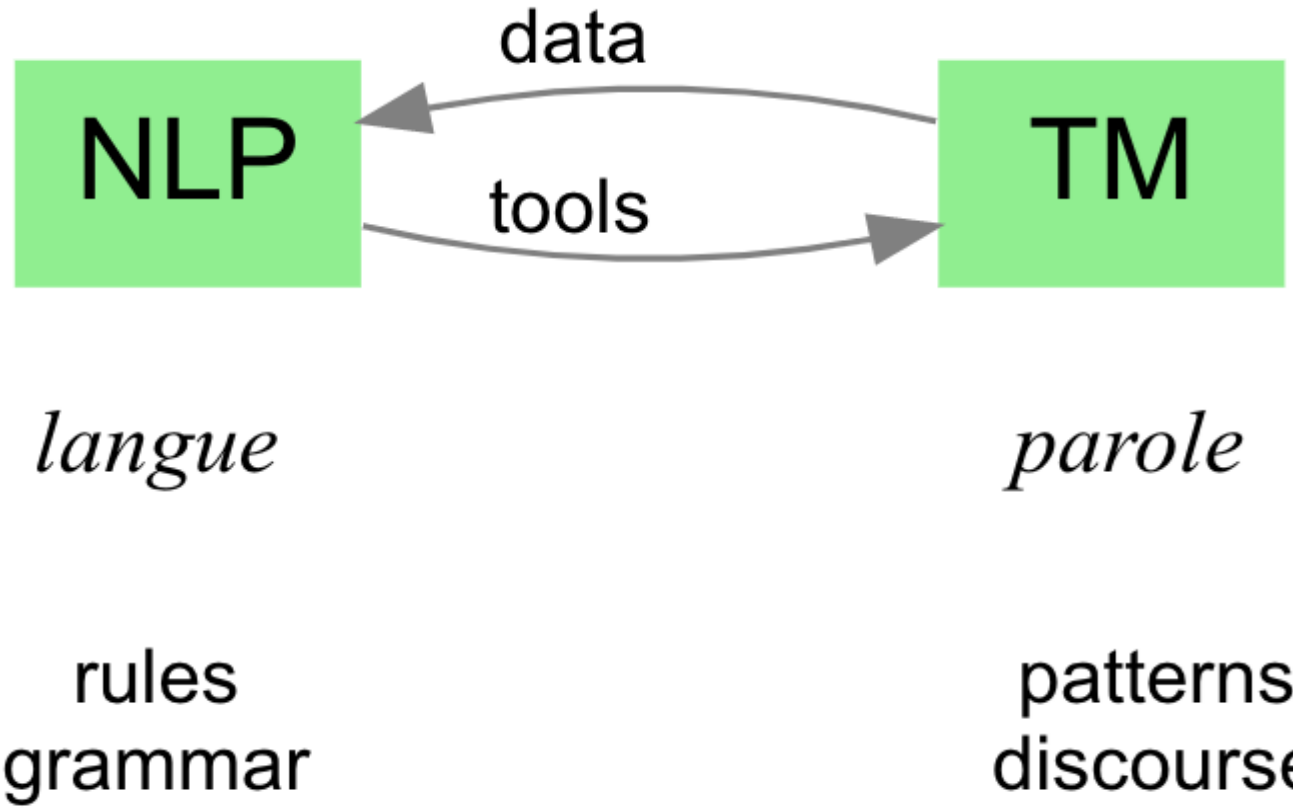
Although often used as synonyms, they have **different concerns, approaches, and methods**

They are, however, **closely related**

Areas of Focus

NLP	TM
Language models	*Text as structured data*
Tokenization	Document classification
Part of speech labeling	Content summarization
Named entity recognition	Network analysis
Dependency parsing	Knowledge discovery
Speech generation	Hypothesis discovery

Text Mining



The functional relationship between NLP and TM

Text Mining: Implementation in R

- Couple of packages in R that specialize in using text data:
  - tm – fairly popular (used often with the corpus package)
  - quanteda – developed by political scientist to analyze politically oriented text data
  - Tidytext – tidyverse of text analysis – this is what we will focus on this week.
  - textmineR – developed mostly for topic modelling

Text Mining: Implementation in R

- The first step with conducting text analysis is getting the data loaded into R so we can tokenize the dataset

- Text data comes in a wide variety of forms and can be difficult to wrangle into a data frame. We are going to use dataset that are in CSV but note this is often not the case.

- Tokenization means that we take a block of text and separate it into separate observations for each
  - word,
  - combination of 2, 3, or 4 words,
  - sentence,
  - or paragraph.

Text Mining: Implementation in R

- The first step with conducting text analysis is getting the data loaded into R so we can tokenize the dataset

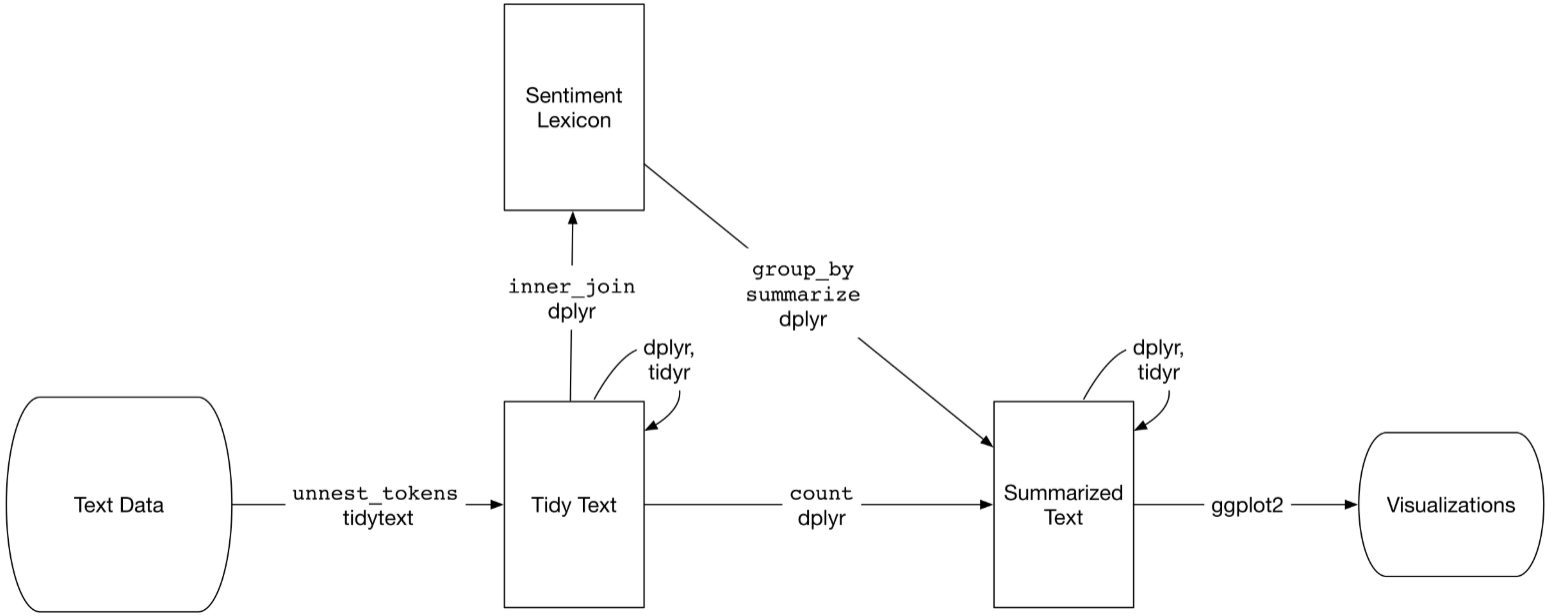
- Text data comes in a wide variety of forms and can be difficult to wrangle into a data frame. We are going to use dataset that are in CSV but note this is often not the case.

- Tokenization means that we take a block of text and separate it into separate observations for each
  - word,
  - combination of 2, 3, or 4 words,
  - sentence,
  - or paragraph.

Text Mining: Implementation in R

Switch over to R

Text Mining: Sentiment Analysis



Source: Text Mining with R

Text Mining: Sentiment Analysis

- Sentiment analysis for our purposes considers text to composed of individual words that can have positive or negative meaning.

- The tidytext package provides access to several lexicons that can work to classify words in our documents in a variety of ways according to sentiment. Examples:

  - AFINN provides a scale from -5 to 5 for included words
  - NRC – classifies words in categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.
  - Bing – Straight poss or neg

- Let's take a look….back to R

Source: Text Mining with R

Text Mining: Topic Modelling

- Next Step in the text journal is Topic Modelling
  - Topic models are "[probabilistic] latent variable models of documents that exploit the correlations among the words and latent semantic themes" (Blei and Lafferty, 2007).
  - The name "topics" signifies the hidden, to be estimated, variable relations (=distributions)that link words in a vocabulary and their occurrence in documents.
  - Essentially think of Topic Modelling as creating clusters of words that are associated with a set of similar documents in a corpus.
    - As an example if you were to gather newspaper articles from across the country from three sections: Politics, Sports and Entertainment. If we ran LDA it would like classify the individual stories into these three topics.

Source: Laten Dirichlet Allocation in R: Martin Ponweiser