Introduction to Big Data for Social Science

Daniela Hochfellner
Clayton Hunter

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
- Providing Rich Context to Research & Data Discovery
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

NYU Center for Urban Science & Progress

Daniela Hochfellner

Senior Research Scientist and Research Assistant Professor at CUSP

CUSP Data Facility lead in data privacy and data provider support

Clayton Hunter

Deputy Director of Training & Outreach for the Coleridge Initiative

NYU's Center for Urban Science & Progress: http://cusp.nyu.edu/
Coleridge Initiative (https://coleridgeinitiative.org/) is a collaboration between CUSP, UMD's JPSM, and Harris at UChicago

Coleridge Initiative's Applied Data Analytics Program - Directors

Rayid Ghani

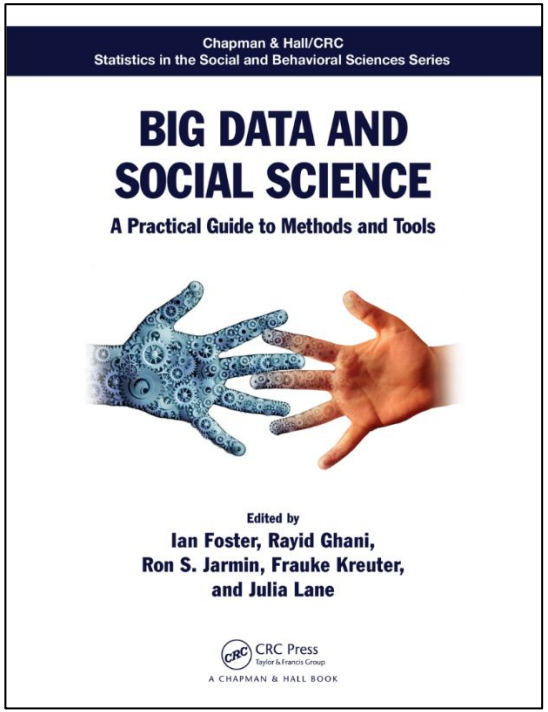
- Director, Center for Data Science and Public Policy
  Senior Fellow, Harris School of Public Policy
  Senior Fellow, Computational Institute, The University of Chicago

Frauke Kreuter

- Professor, Joint Program in Survey Methodology, University of Maryland
  Professor, Methods and Statistics, University of Mannheim
  Head, Statistical Methods Group, German Institute for Employment Research, Nuremberg

Julia Lane

- Professor, Robert F. Wagner Graduate School of Public Service, NYU
  Professor, Center for Urban Science and Progress, NYU
  Provostial Fellow, Innovation Analytics



Structure & contents of Big Data for Social Science

Capture and Curation	Working with Web Data and APIs, Record Linkage, Databases, Programming with Big Data
Modeling and Analysis	Machine Learning, Text Analysis, Networks
Inference and Ethics	Information Visualization, Errors and Inference, Privacy and Confidentiality

http://www.bigdatasocialscience.com/contents

Structure & contents of Big Data for Social Science (Today)

Capture and Curation	Working with Web Data and APIs
Modeling and Analysis	Machine Learning, Text Analysis
Inference and Ethics	Privacy and Confidentiality

http://www.bigdatasocialscience.com/contents

Learning Objectives

- You will be able to understand fundamental concepts of applying new analytical techniques to social science
- You will learn how to perform your own data analytics projects
- You will learn the underlying methodology of data analytics toolkits
- This course will also teach you limitation of these methods

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
- Providing Rich Context to Research & Data Discovery
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

Big Data are ...

- Huge in volume
- High in velocity
- Diverse in variety
- Exhaustive in scope (populations)
- Fine grained in resolution
- Relational in nature
- Flexible (can expand in size rapidly)

From Various Sources

- Directed
  - Data are generated by traditional form of surveillance (transaction data, sensors, mobile phones, clickstream data, scanning of machine readable objects.
- Volunteered
  - Social media, crowdsourcing projects (open maps)
- Automated
- Structured, semi-Structured, and unstructured

Importance of Big Data for Social Science

- New analytical paradigm
  - It’s more than just data -> the entire approach
- New business model
  - Cost
  - Coverage
  - Timeliness
  - Competition
- Agencies need new approaches
- Evidence-based Policymaking Commission Act of 2016

However, Be Aware of Problems

- Large data sets from Internet sources are often unreliable
- Understanding biases requires understanding the properties and limits of a data set
- We need to know where data is coming from
- Big data risk: seeing patterns where none actually exist, simply because enormous quantities of data
- Ownership?
- Processing complexity

Big Data Analytics Lifecycle

- Preparation:
  - Is there enough good data to be potentially useful?
- Planning:
  - There are many ML models to choose from. Which one(s) to use?
- Building:
  - Is the model robust? Is it appropriate?
- Operationalizing:
  - Scale the model(s) for deployment

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
- Providing Rich Context to Research & Data Discovery
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

Data collection from the web

Web-scraping: automated collection of information directly from web pages, that may or may not be intended for that use.

Application Programming Interface (API): a system made to help us use some software in an automated way

Web-scraping vs APIs

Web Scraping / Crawling	APIs
Work Intensive	Easily Automated
Subject to Change	Consistent / Predictable
Referred to as ‘Soup’	Intentionally Structured
Generally Inefficient	Efficient

There is a *lot* of information on the internet

> “The Internet is a busy place. **Every second**, approximately **6,000 tweets** are tweeted; more than **40,000 Google queries** are searched; and more than **2 million emails** are sent...”

- *How Big Is the Internet, Really?* By Stephanie Pappas, March 2016

https://www.livescience.com/54094-how-big-is-the-internet.html

Web-scraping & web-crawling

Web Scraping

Automated collection of information directly from web pages, that may or may not be intended for that use. This uses a few core tools for (1) making requests via HTTP and (2) organizing data with HTML Parsing. But can also require a wider set of skills for browser automation or advanced searches (regex).

Web Crawling

Scraping, except across many different webpages

Motivating question: what data do you need?

Does ex-offender access to social services reduce the likelihood of recidivism?

Does welfare benefit recipient’s access to social services matter?

Data required for either question:

- Social service locations
- Individuals’ locations
- Travel time between the two

Web-scraping workflow

Find the site

Ask maintainers of site if underlying data available

View structure of site

Write & test scraper

Warn site you're going to scrape from it

Scrape information

Practical example of web-scraping

API - Application Programming Interface

Using APIs entails interacting with a system made to help us use some software in an automated way.

We can use APIs to (1) obtain data & (2) use tools.

An agreement that if you ask for information in a specified way the system will return information in a specific structure.

Practical example of API using OpenTripPlanner

Web-scraping vs APIs

Web Scraping / Crawling	APIs
Work Intensive	Easily Automated
Subject to Change	Consistent / Predictable
Referred to as ‘Soup’	Intentionally Structured
Generally Inefficient	Efficient

BREAK UNTIL 10:45

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
- Providing Rich Context to Research & Data Discovery
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

What is Machine Learning?

“A computer program is said to learn from experience E

with respect to some task T

and some performance measure P,

if its performance on T, as measured by P, improves with experience E.”

Advantages of Machine Learning

- Goal: Adaptive, Scalable systems that are cost effective to build and maintain
- Rules-based systems are rigid and expensive
- Lots of data is available to “train” the system

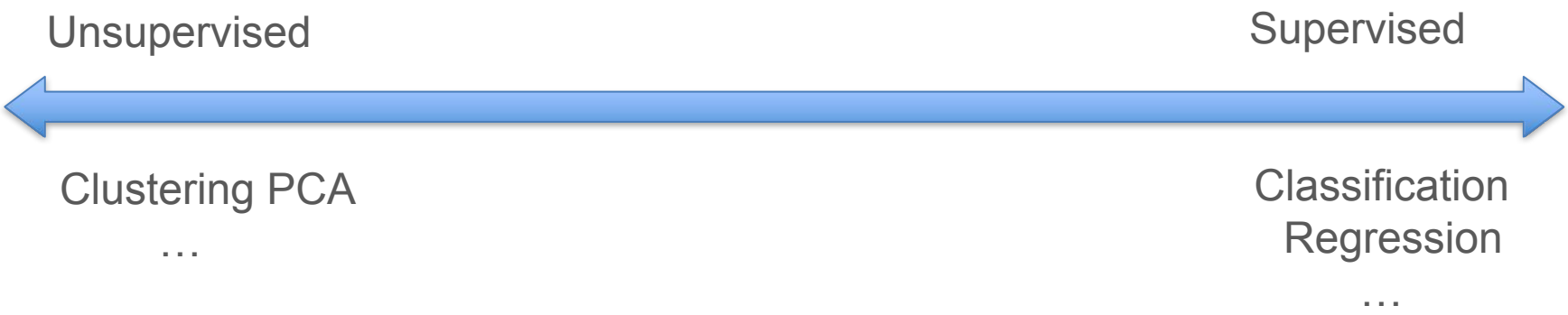
What can we use it for in Public Policy

- Description (Understand the past)
- Detection (Anomalies, Events, Patterns)
- Prediction (Predict the Future)
- Behavior Change (Causal Inference)

Workflow of Machine Learning Projects

- Formulate your policy problem
- Map the problem to a Machine Learning problem
- Get (and integrate) the data you need to work on the problem
- Explore, process and clean the data
- Create “Features” you want to include in the model
- Select which methods to test
- Evaluate methods
- Deploy, Maintain, Update

Different Types of Models



Unsupervised Learning

- No outcome or dependent variable is present
- Goal is exploration, understanding historical data, finding patterns and/or groups in the data

Examples

- Clustering (cluster analysis)
- Principal Components Analysis
- Association Rules (beer and diapers)

Clustering Algorithms

- Partitioning a group of data points into a small number of clusters
- Allows you to find and analyze the groups that have formed organically
- A good clustering method will produce clusters with
  - High intra-cluster similarity
  - Low inter-cluster similarity

- K-Means is the simplest and the most common algorithm
  - Goal: find groups in the data, with the number of groups represented by the variable K
  - Assign each data point to one of K groups based on the features that are provided

K-Means Clustering

We have n data points $x_i, i = 1 \dots n$

which have to be partitioned in k clusters

Goal: Assign a data point to each cluster

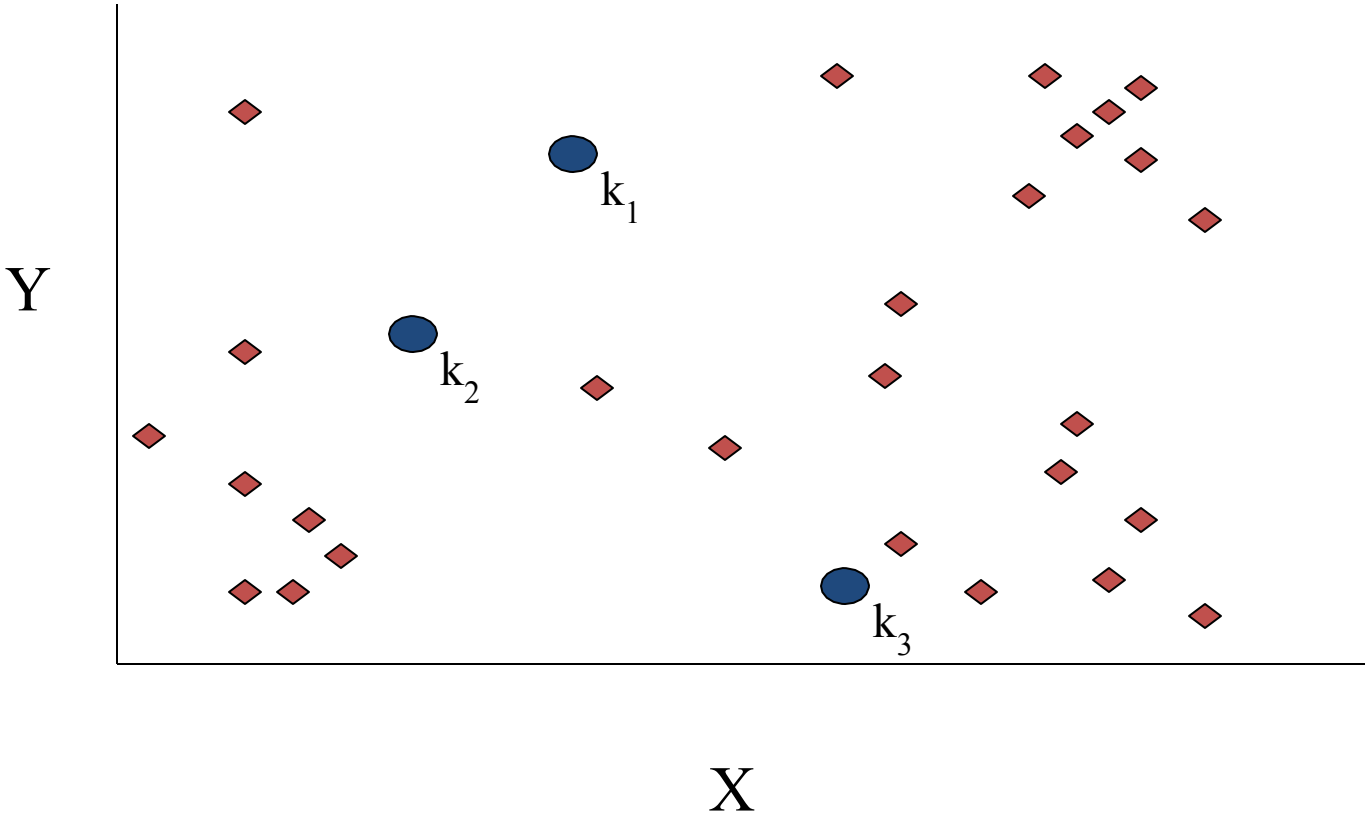
Aims to find the positions $\mu_i, i = 1 \dots k$ of the clusters that minimize the distance from the data points to the cluster

$$\arg\min_{\mathbf{c}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in c_i} \|\mathbf{x} - \mu_i\|_2^2$$

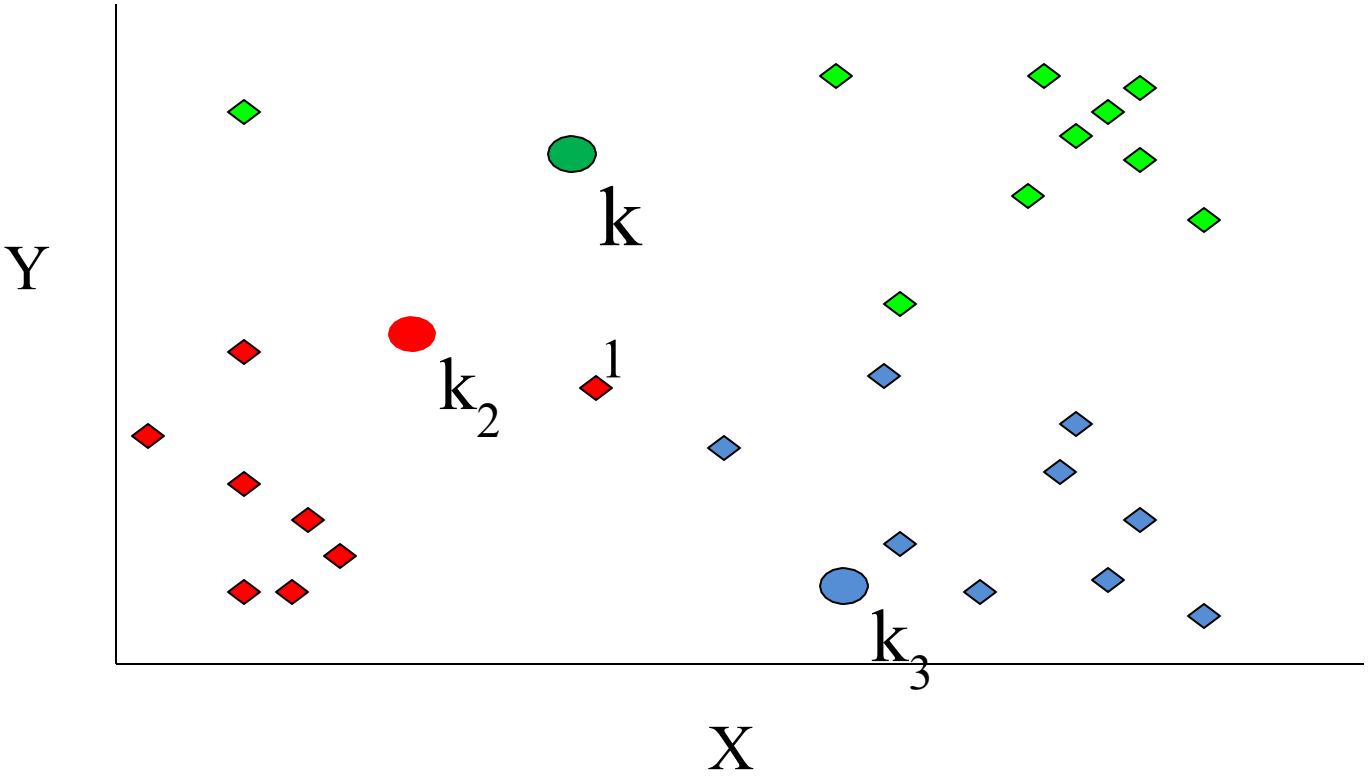
This means...

1. Randomly choose k data points (seeds) to be the initial centroids, cluster centers
2. Assign each data point to the closest centroid
3. Re-compute the centroids using the current cluster memberships.
4. If a convergence criterion is not met, go to 2).

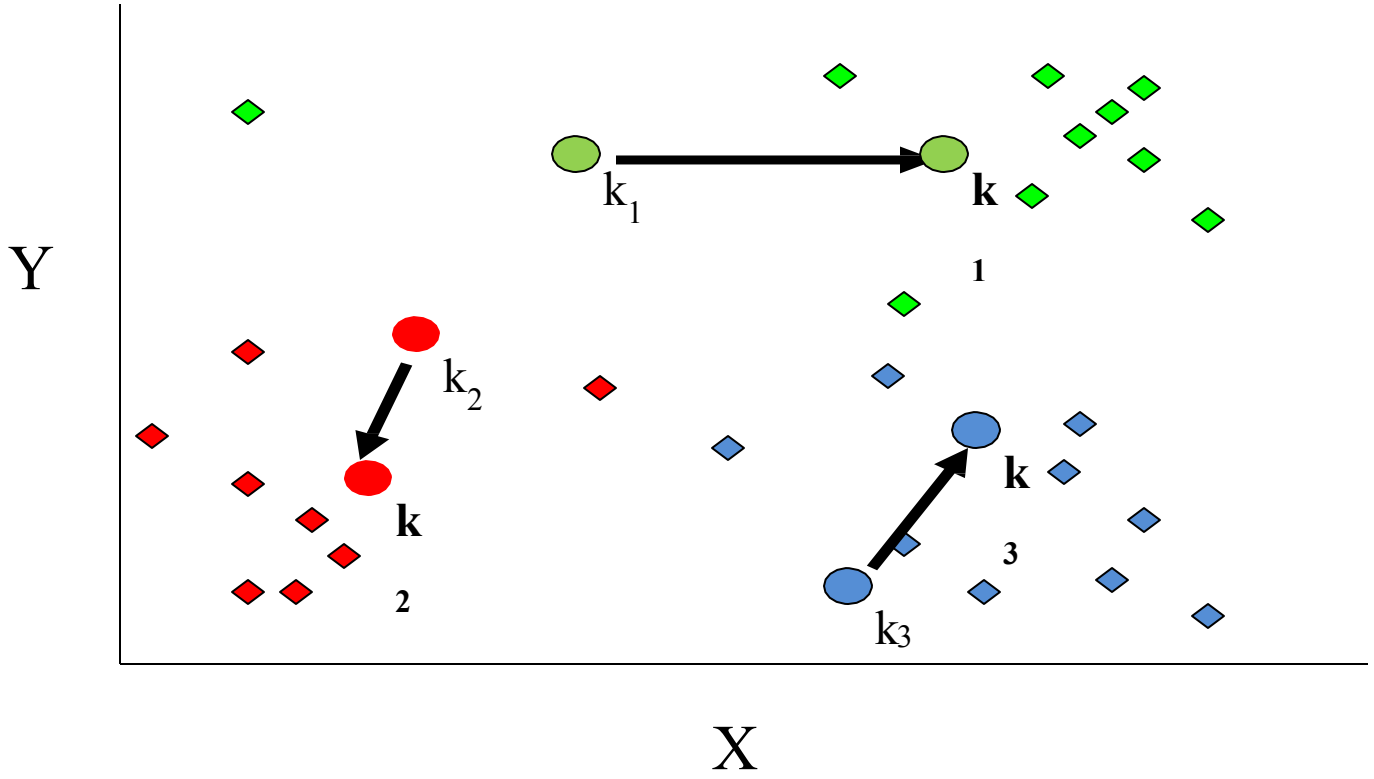
Randomly choose cluster centers



Assign each point to the closest center



Recompute centers (mean of cluster)



Evaluation of Clusters

- Objective Evaluation
- Task-Specific Evaluation
- Same data can be clustered in different ways in different number of clusters
- Can be used to interactively explore data

Supervised Learning

Supervised learning is where you have

- input variables (X), also called features
- and an output variable (y)
- and you use an algorithm to learn the mapping function from the input to the output

$$y = f(X)$$

The task is to find an f that minimizes the error in recovering y

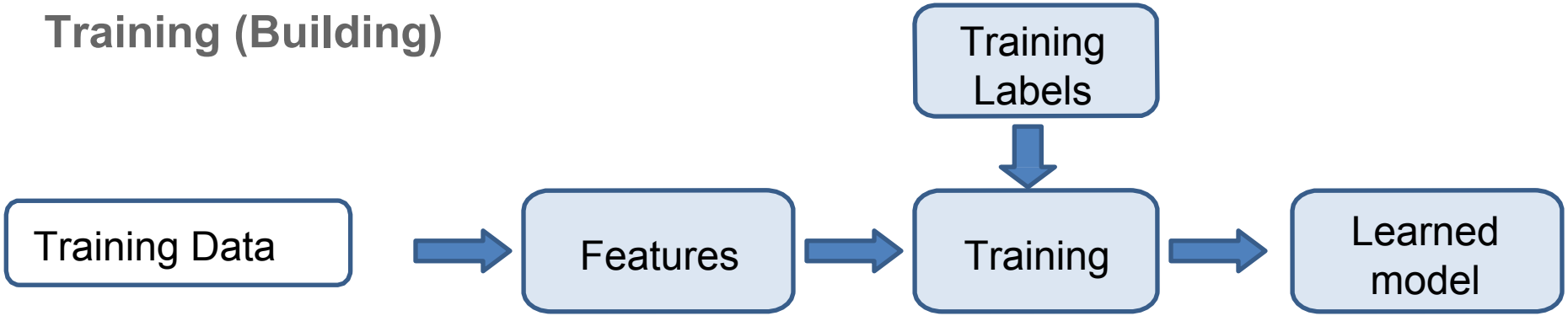
Steps to Solve a Classification Problem

1. Define and Create label (outcome variable)
2. Define and Create Features (predictors)
3. Create Training and Validation Sets
4. Train model(s) on Training Set
5. Validate model(s) on Validation Set

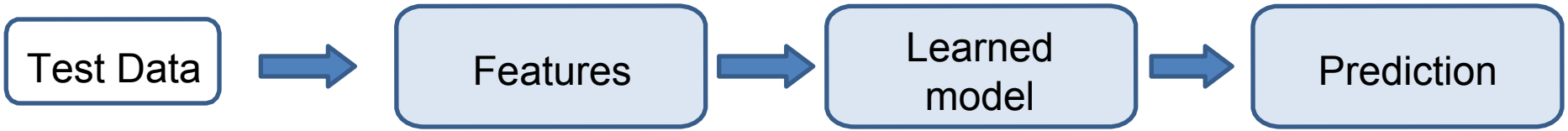
Build model that is useful for prediction, don’t build model that fits the data best
Predictions, focus is not causality

Modeling and Validation

Training (Building)



Testing (Validating)



Various Classifiers to Choose From

- **Nearest Neighbor**
- Decision Trees
- Regression
- **Support Vector Machines**
- Bayes Classifier
- Neural Networks
- Ensembles
  - Bagging
  - Boosting
  - Random Forests

K Nearest Neighbor (KNN)

- Simple to understand, works well in practice
- Very versatile, plenty of applications (for example economic forecasting)
- Non parametric: no assumption on the underlying data distribution
- Lazy algorithm: does not use training data for generalization
- Forming a majority vote between the K most similar instances to a given “unseen” observation

Assumptions

- The data is in a feature space, more exactly in a metric space
- There is a notion of distance (for example euclidean)
- Each of the training data consists of a set of vectors and class label associated with each vector
- Number "k" . This number decides how many neighbors (where neighbors is defined based on the distance metric) influence the classification

K Nearest Neighbor (KNN)

Given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

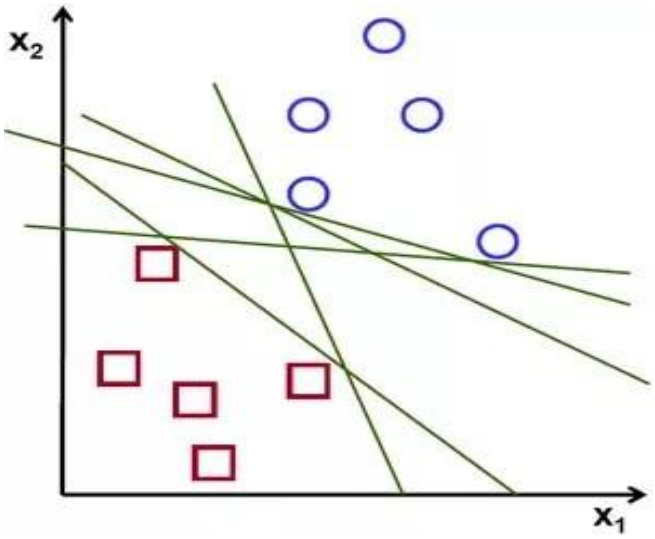
- It runs through the whole dataset computing d between x and each training observation. We'll call the K points in the training data that are closest to x the set A.
- It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label
- Finally, our input x gets assigned to the class with the largest probability

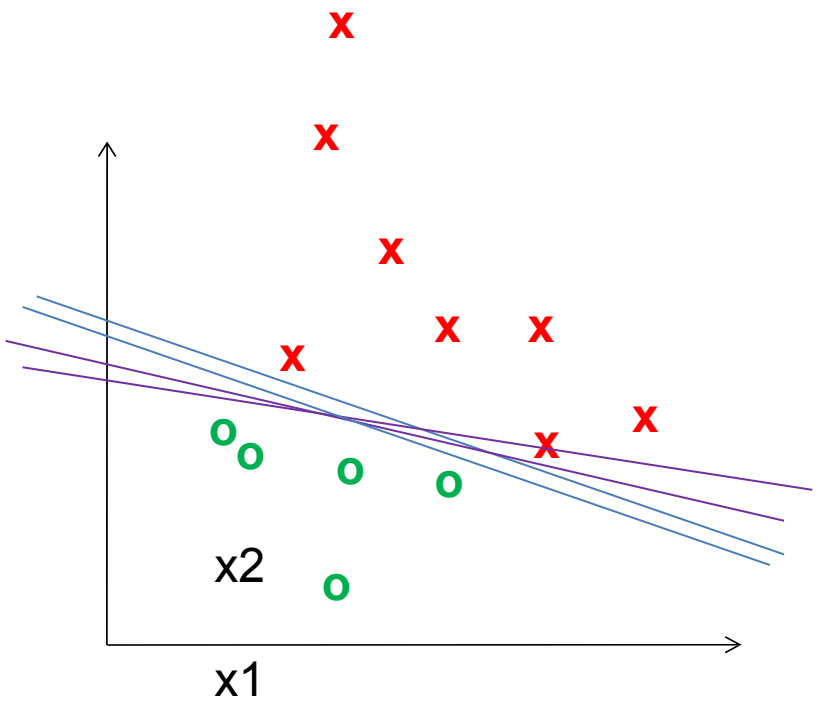
Example

Model	Durability	Strength	Class	Distance	Rank
Type 1	7	7	Bad	$\text{Sqrt}((7-3)^2+(7-7)^2)=4$	3
Type 2	7	4	Bad	5	4
Type 3	3	4	Good	3	1
Type 4	1	4	Good	3.6	2

Data to be classified: durability 3, strength 7, class ?

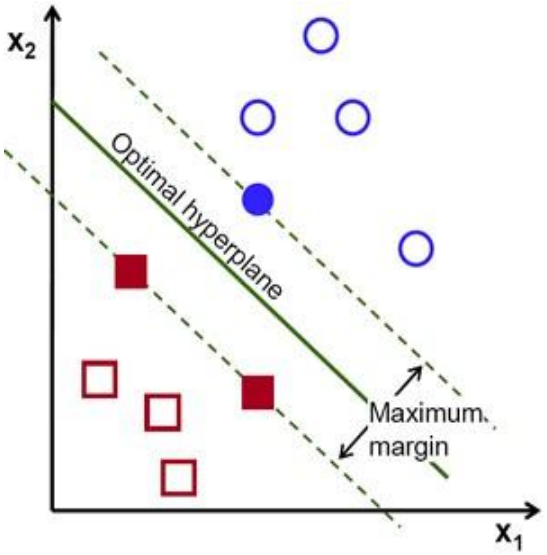
Support Vector Machines

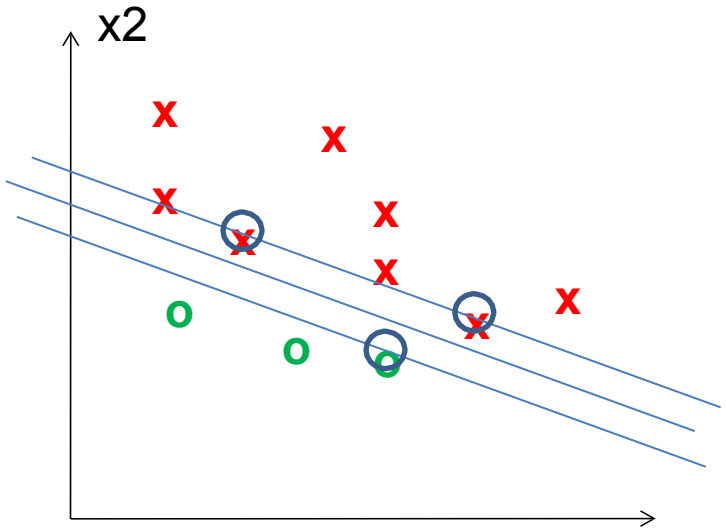




Find a *linear function* to separate the classes: f(**x**) = sgn(**w** · **x** + b)

Support Vector Machines II





- Too many possible boundaries
- SVMs attempt to maximize the “margin”
- Optimization problem

Classifiers...

- Better to have smart features and simple classifiers than simple features and smart classifiers
- Need more training data with increasingly powerful/complex classifiers

Evaluation

- In-sample
- Out of sample
- Multiple Out-of-sample (Hold-out) Splits
- Cross Validation
  - Leave one out (LOO)
  - K fold
- Temporal Holdouts

Evaluation Metric

- Predictions are often scores between 0 and 1
- We need to first turn them into 0 or 1 by selecting a threshold

Predicted Class

Actual Class	Yes	No
Yes	**True Positives (TP)**	**False Negatives (FN)**
No	**False Positives (FP)**	**True Negatives (TN)**

Evaluation Metrics II

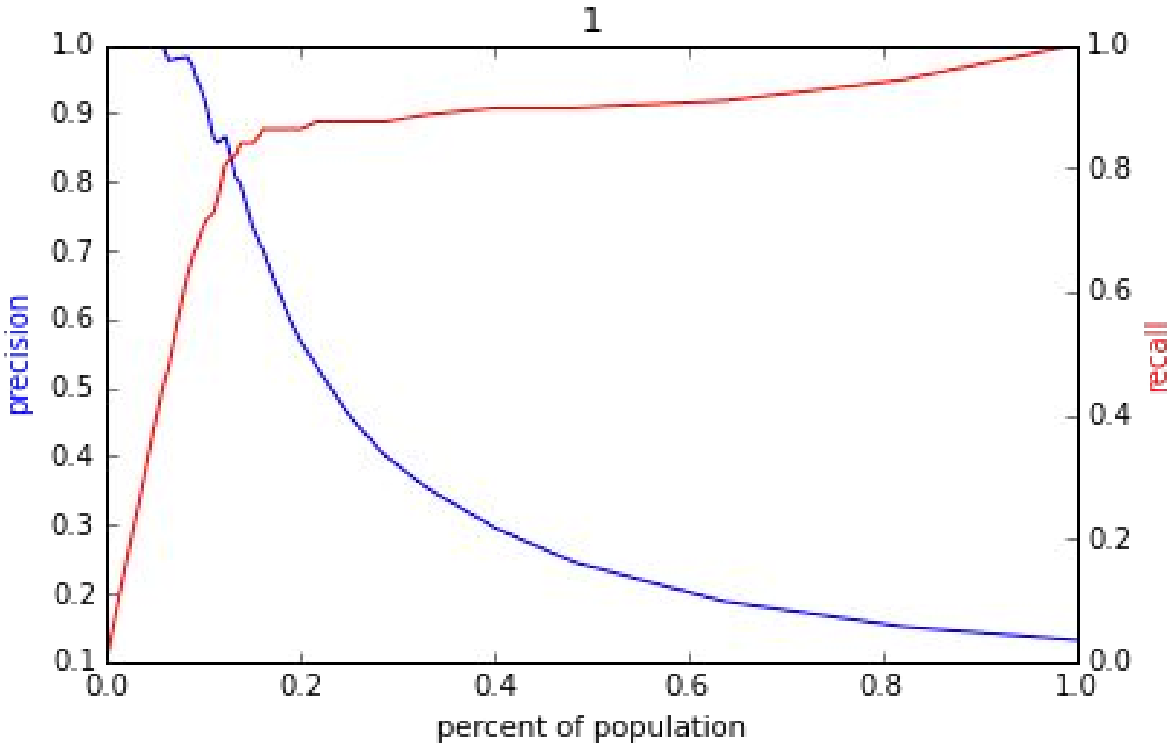
Accuracy = **(TP + TN) / (TP + TN + FP + FN)**

Precision (or PPV) = **TP / (TP + FP)**

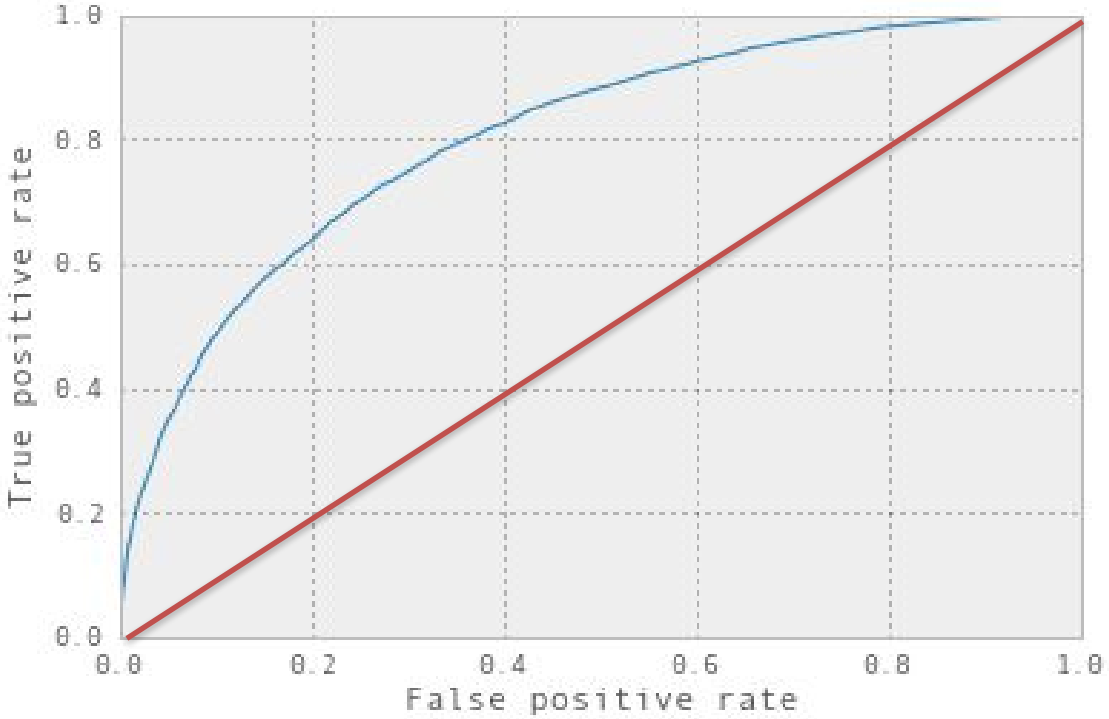
Recall (or Sensitivity) = **TP / (TP + FN)**

Specificity = TNR

Threshold


1
precision
recall
percent of population
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve



Factors To Consider in ML

- Complexity
- Overfitting
- Robustness
- Interpretability
- Training Time
- Test Time

Understanding Potential Biases: Selection

- Is your sample randomly drawn? What about APIs?
- Who is represented?
  - Facebook, Twitter,
  - Open data/transactional data
- Your model is biased towards to people who are in your sample

- Assumption: data in the design set are randomly drawn from the same distribution as the points to be classified in the future. True?
  - *Population drift*, you are dealing with something that no longer exists

Examples: Selection Bias

- Netflix movie recommendations
  - Biased toward people who have the time and interest in rating a bunch of movies online
- Predicting customer behavior
  - Customers will change their behavior with price changes etc. -> distribution is changing over time (even if N=All, N ≠ N')
- Classification rule aimed at differential medical diagnosis:
  - What if patients are drawn from only one hospital? -> demographic, social, and other factors influence who seeks treatment, might vary across hospitals

Understanding Potential Biases: Training Data

Your learning algorithm is designed to pick up statistical patterns in training data, thus the quality and representativeness of records might vary in ways that correlate with class membership

- Social Bias
  - If the training data reflect existing social biases against a minority, the algorithm is likely to incorporate these biases
  - If not everybody contributes in the same way to data collection
- Sample Size
  - Classifier generally improves with the number of data points used to train it, thus less data leads to worse predictions

Understanding Potential Biases: Target variables

Analysts must determine how to solve the problem at hand by translating it into a question about the value of some “target variable”. The proper specification is not always obvious

The “ideal” consumer:

- The person who is most likely to click on an ad
- The person who is most likely to make a purchase
- The person who will establish a long-term relationship with the company
- The person who will generate the most profit for the company

Understanding Potential Biases: Features

- We need good proxies
  - Don’t pretend to measure something you can’t. The way to avoid this mistake is by avoiding being vague in describing what the model does with the input.
- Good Proxies
  - We can’t measure someone’s interest in a website, but we can measure how many pages they went to and how long they spent on each page
- Bad Proxies
  - Measuring teachers performance via students’ grades
- Concept drift
  - Assumption: there are no errors and classes are well defined
  - However: classes might change over time

Consequences?

If biases are not taken into account ...

- It might lead to decisions/procedures which limit the future contact an agency will have with specific groups,
  - Skewing further the sample upon which subsequent analyses will be performed
- It might lead to members of those populations not having access to specific services or to less advantageous decisions for them
- Your model might be less accurate for certain populations

What We Can Do

- We need to understand where the data comes from
- We need to know who is represented and if this changes over time
- Then we can assess possibles biases and errors we are making
- Then we can decide if we can actually answer the questions we want to answer

Practical example of machine learning: record linkage

LUNCH BREAK (UNTIL 1:15)

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

Outline

1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

Outline

1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

Information Retrieval

Automatic Extraction of Dataset Names from Publications

We have developed a system for mining free text publications to automatically extract references to dataset. To bootstrap our algorithms, we have leveraged ICPSR to develop a training set of linked datasets and publications.

Our algorithms apply NLP techniques to identify datasets within papers. We have found the following two areas to focus on that produce accurate dataset names used in the papers.

1. Papers often reference tables or figures from published datasets. In these cases, the name of the dataset is usually included in the caption that precedes or follows the table or figure. (Examples shown in Fig. 1)

Age difference with partner	Age difference between the respondents and their partner	Integer	NA	NA	− 11	30	2.026	4.63
Partner is of Hispanic origin	Respondent's partner was of Hispanic origin.	Yes or No	NA	NA	0	1	0.898	0.30

Source: Fragile Families and Child Wellbeing (waves 1–4).

† Refers to $p < 0.10$.

••• Refers to $p < 0.001$.

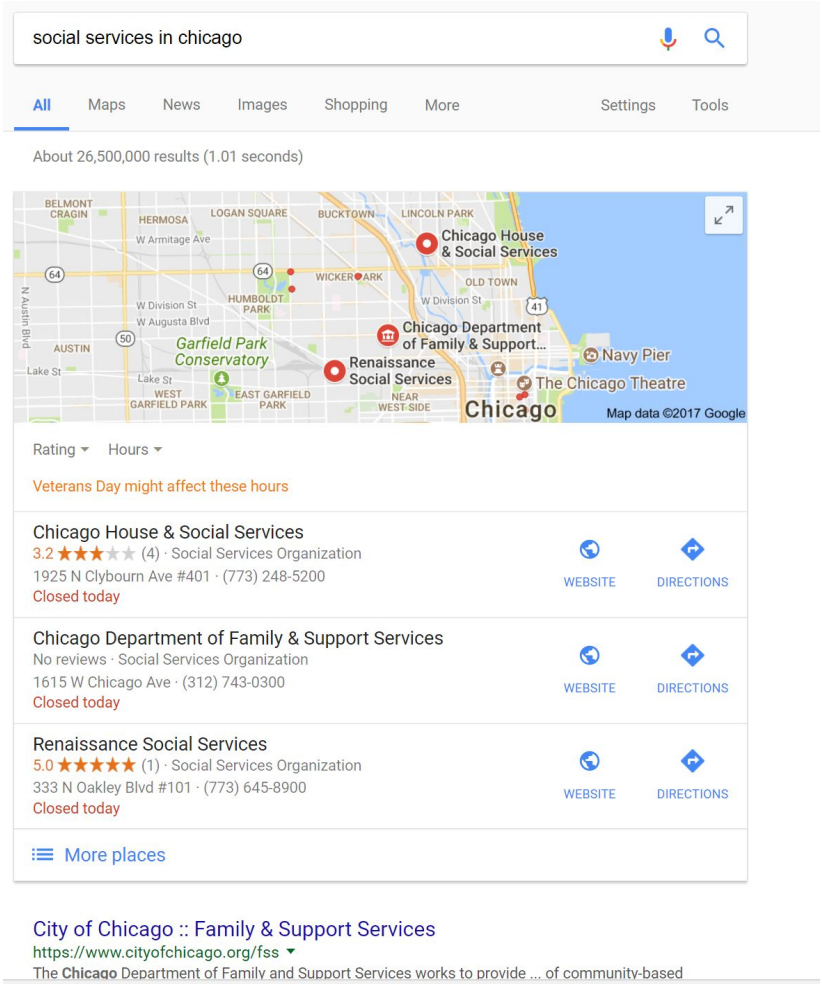
Results

As an example of our process the table below shows the datasets that are identified for three ICPSR papers. The first two show correct matches. The third shows three suggested datasets of which only the first is correct.

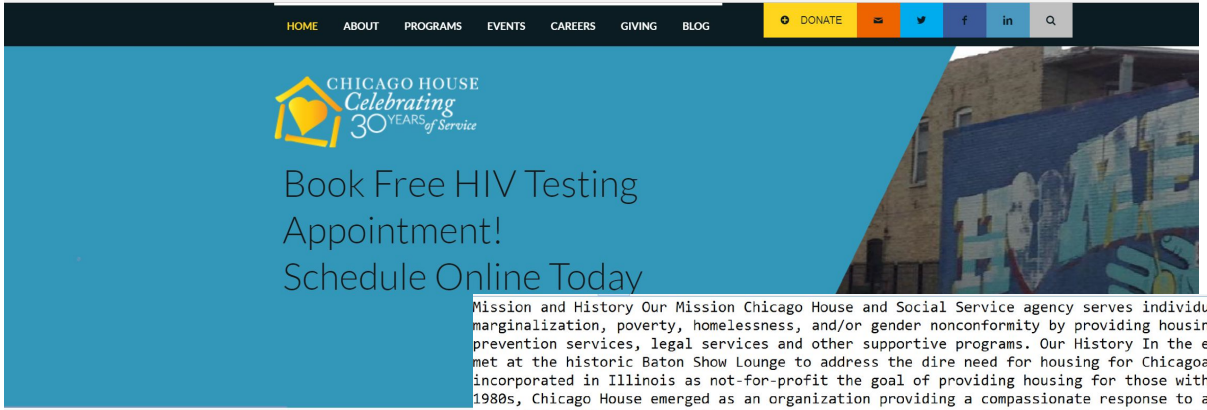
Paper	Datasets identified
Fragile families in the American welfare state https://doi.org/10.1016/j.childyouth.2015.05.018	{'Fragile Families and Child Wellbeing'}
Children in no-parent households: The continuity of arrangements and the composition of households https://doi.org/10.1016/j.childyouth.2007.02.001	{'Survey of Income and Program Participation'}
Effects of infant health on family food insecurity: Evidence from two U.S. birth cohort studies https://doi.org/10.1016/j.socscimed.2014.10.041	{'Fragile Families and Child Wellbeing Study', 'Supplemental Nutrition Assistance Program.', 'Special Supplemental Nutrition Program for Women,', 'Supplemental Nutrition Program for Women,', 'Supplemental Nutrition Supplement Program'} (Note: these false positives are not dataset names, but because they have the keyword "program" in them they didn't get filtered out)

Of our 50 papers, we correctly obtained the dataset in 82% of cases. We completely missed the dataset in 8% and incorrectly extracted a dataset in 10% of cases. These results, while preliminary, highlight the ability to establish these linkages with even simple heuristics.

General Extraction



www.chicagohouse.org



Mission and History Our Mission Chicago House and Social Service agency serves individuals and families who are disenfranchised by HIV/AIDS, LGBTQ marginalization, poverty, homelessness, and/or gender nonconformity by providing housing, employment services, medical linkage and retention services, HIV prevention services, legal services and other supportive programs. Our History In the early years of the HIV/AIDS epidemic in the US, nearly 100 activists met at the historic Baton Show Lounge to address the dire need for housing for Chicagoans living with AIDS. On September 9, 1985, Chicago House was incorporated in Illinois as not-for-profit the goal of providing housing for those with AIDS. A Chicago House float in the Pride Parade, 1988 During the 1980s, Chicago House emerged as an organization providing a compassionate response to a fearful disease. We built our infrastructure as we established several facilities to meet the growing and ever-evolving needs of our clients. Our first residence, opened in February 1986 in Uptown, accommodated 8 individuals in private bedrooms with shared cooking, dining, bathing, and living areas. We opened two more residences in 1987 and, responding to demand, a 24-hour care program and hospice in 1988 for clients in need of additional support. In 1992, Chicago House formed the Family Support Program, becoming the first provider of housing and related services to HIV-affected families. Before advances in HIV medication, these early years often meant that we were providing our clients a place to die with dignity after they had been abandoned by friends and family or forced out of their apartments. An early agency newsletter from 1986 reads, "Through May 15, 1987, Chicago House has provided a home for 28 individuals, all men. Of the 17 men no longer with Chicago House, 10 are deceased." In these early years, Chicago House was the last home many people had before they died. Early interior of a housing unit in Chicago's West Town The latter half of the 1990s, however, ushered in new hope as medical advances came to the fore with the advent of new drug therapies. As drug advances and approvals began to dominate the news in the AIDS community, increased hope was also somewhat tempered by the reality that the annual cost of drug therapy could easily exceed $20,000 per person. In 2000, the annual AIDS death toll in the United States dropped to 17,741 after it had peaked in 1995 at over 48,000. Many people with HIV/AIDS found themselves getting better and were confronted with the reality that they now faced rebuilding their lives while living with HIV rather than preparing to die. In 1997, we bought the building that would house our Independent Living Program, responding to then need that many of those living with HIV/AIDS had to live in a more independent supportive housing program. Then, in 1999, after experiencing empty beds in our hospice facility and increased demand in our other housing programs, Chicago House made the difficult decision to close down the hospice. Chicago House's hospice provided essential and compassionate care to over 400 people. In the four housing facilities that we currently operate, Chicago House now provides permanent supportive housing to more than 100 people each year. We also provide these residents, along with those living independently in the community, with vital case management services. In 2004, Chicago House became the lead agency on a 5-year project funded by the Centers for Disease Control (CDC). Running this program, Chicago House helped our HIV-positive clients reach out and protect their peers and partners from infection. Chicago House continues to run a CDC-funded Prevention With Positives intervention that, in 2013, reached 110 people. Also in 2004, U.S. Senator Dick Durbin visited Chicago House to learn more about our programs and the needs of our clients. The residents he met with expressed their desire to return to work, but explained the major obstacles that they faced. After listening to their concerns and the efforts of the agency, Senator Durbin helped facilitate a start-up grant to support the initial development and implementation of an HIV employment service system, which became our Employment Services Program. The program launched in the fall of 2005 with much support from leaders in the HIV and workforce development community and, since its inception, has served nearly 1,000 participants. Chicago House began its Scattered-Site Housing Program, which provides rental assistance and support services to clients, in 2006. Initially offering 18 apartment units connecting people to housing and employment services, the Scattered-Site Housing Program now manages about 150 units each year. 2013 was another momentous year for Chicago House as we officially launched the TransLife Center (TLC), responding to the great need for culturally-competent, expert social services specifically for transgender individuals. In the same facility that we once used as a hospice in the early years of the HIV epidemic, we now deliver housing, employment, linkage-to-medical care, case management, and legal services. With community partners, we are currently leading 5-year Special Project of National Significance funded by the Health Resources and Services Adminsitration to study the link between HIV and retention in care in transgender women of color.

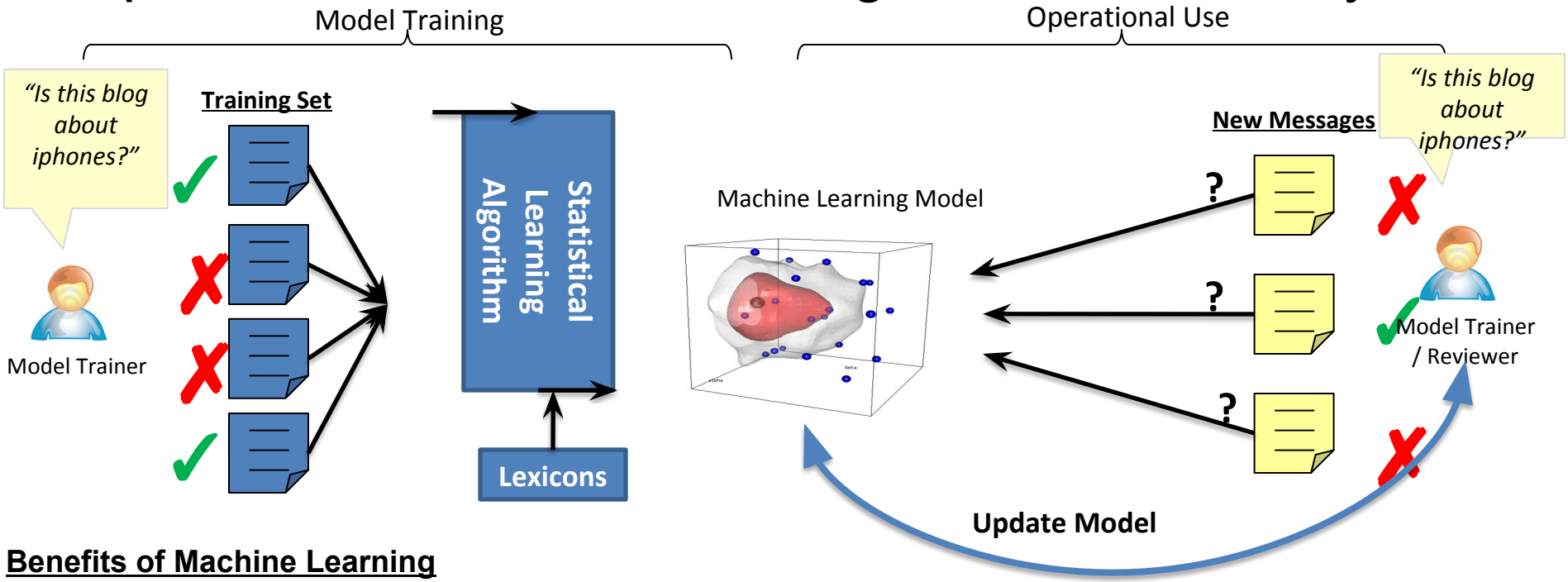
Summary: What can text mining do?

Capability	Description	Maturity
Classification	Classification of a document into one or more previously defined content categories. (e.g. Classify incoming emails as personal, business, news article, etc.)	High
Word Clustering / Synonyms	Finding groups of words that are similar to each other. Depending on the strictness of the definition of similarity, similar words can be synonyms (IBM, Deep Blue) or companies in similar industry (Oracle, Teradata)	Low
Topic Detection / Clustering	Finding emerging or existing topics in large amounts of text data. Clustering news articles can help detect emerging topics.	High
Opinion & Sentiment Analysis	Detection of sentiment and opinions in different levels of granularity (word, sentence, message).	Medium
Named Entity Extraction (People, Locations, Organizations)	Recognition, tagging, and extraction of named entities of type Person, Location, and Organization. Typically limited to proper nouns and not much customization possible.	High
General Extraction (Entities, Events, Facts, Relationships)	Recognition, tagging, and extraction of specified classes of words /phrases as an entity (client, competitor, company), event (acquisition), relationship (John King works for MSFT)	Low
Search	Ranked retrieval of a document or component based on the presence of one or more supplied search terms.	High
Visualization	Visualization of text data and /or visual mashups combining text with other forms of data (maps, networks, etc).	Medium
Summarization	Summarization of a document, intended to produce a readable abstract of the source document which captures salient points using fewer words.	Low

Different approaches

	Lexicon-based Rules	Linguistic Rules	Statistical Machine Learning (augmented with linguistics & lexicons)
Description	Rules based on lists of words	Rules using words and linguistic operators (parts of speech for example)	Statistical approaches that can be trained and learn over time. Can incorporate lexicons and linguistics as well
Ease of creation & maintenance	Low	Low	High
Accuracy	Low	Medium	High
Context Sensitiveness	Low	High	High
Interpretability	High (unless the rules get large)	Medium	Medium

Statistical Machine Learning based approaches compensate for the shortcomings of rule-based systems



Benefits of Machine Learning

- ✓ **Significantly cheaper approach to achieve a given level of accuracy compared to manual rule or lexicon creation**
- ✓ **No advanced linguistic or technical skills to train and maintain the system (business users or analysts are the maintainers)**

Does it have to be BIG data?

- Words are a universe of meanings and context:
  - ✓ one word is often enough to effectively communicate the message
- Documents bear different importance:
  - ✓ several directives and laws can cover an overlook of a subject area better than thousands of background notes
- The more documents the more confusing the message:
  - ✓ depends on context and application but generally cleaner datasets will convey cleaner messages

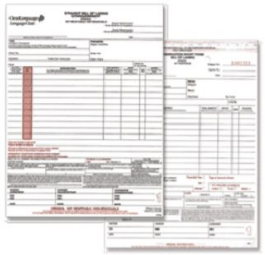
Outline

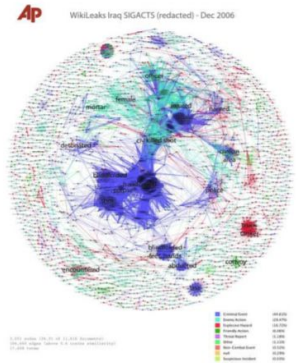
1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

What is text

- A bunch of words
- A bunch of linked words
- A bunch of ambiguous linked words
- A bunch of ambiguous linked words in multiple languages

Where does text data come from?





Why is text data different?

- Structured Data: Humans have already pre-determined the attributes and relations of interest (e.g. relational databases)
- Text data can have too many possible dimensions (millions)
- Text data often reflects human observations that are exceptions to regular business processes.
  - Complaints, Suggestions, (the ubiquitous “other” field)

What is a document

A purposefully organized text written in a given genre and style and serving a specific audience.

What is a corpus of documents

A corpus is a collection of documents that are typically similar in purpose, genre and style and share similar statistical properties.

Text corpora: examples

- Brown University Standard Corpus of Present-Day American English (a.k.a. the Brown Corpus)
- Lancaster-Oslo-Bergen Corpus
- The Penn Treebank
- TREC (NIST)
- The Enron Corpus

- Social services websites…

Varieties of text

- **Language:** English, Spanish, Chinese…
- **Genre:** fiction, non-fiction…
- **Style:** concise, longer sentences, many adjectives…
- **Purpose:** policy document, scientific report, student essay…
- **Sentiment:** positive, negative, neutral
- **Truth** or **lie**
- **Target audience:** electorate, funders, general public…

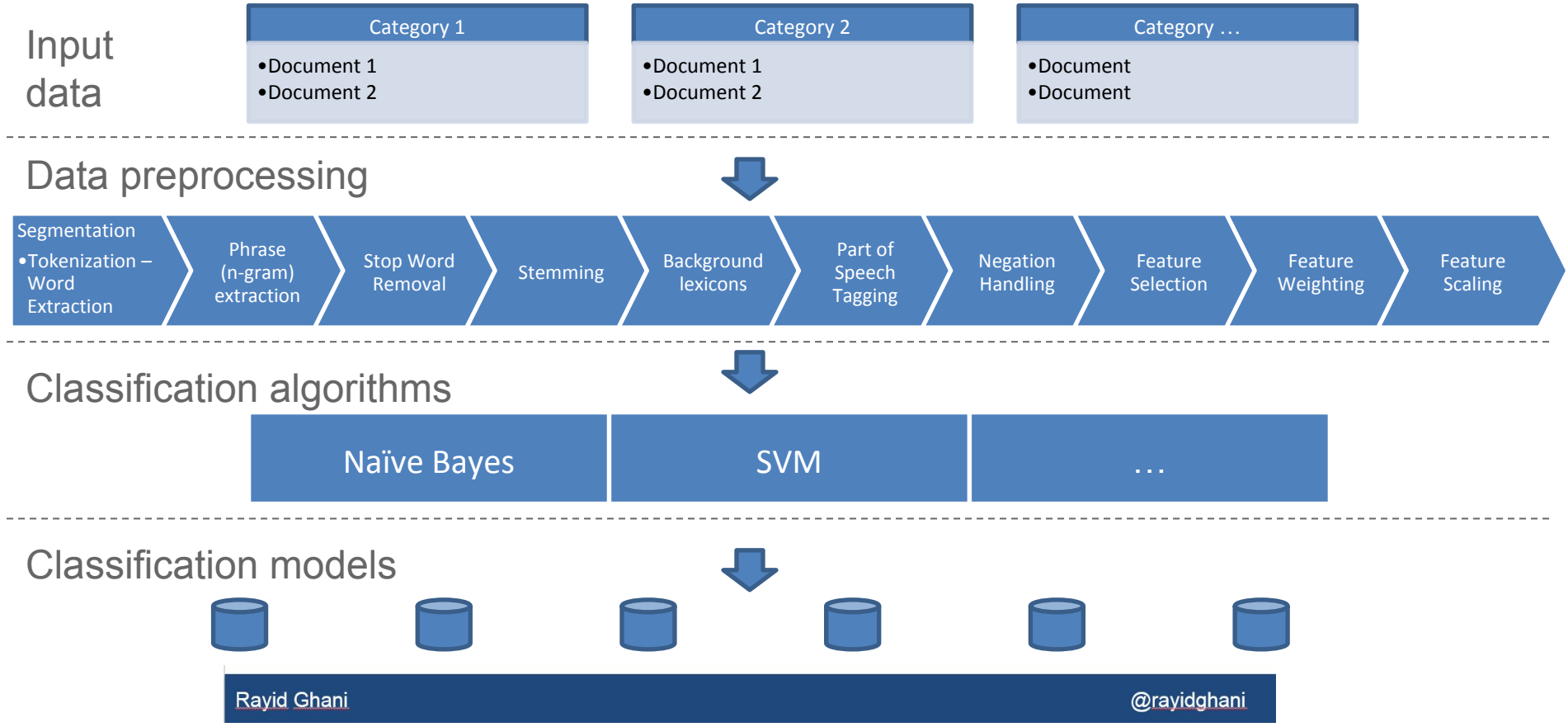
Varieties of text: methods

- **Language:** machine translation, PLTM
- **Genre and Style:** word sense disambiguation
- **Purpose:** document classification
- **Sentiment:** sentiment analysis
- **Truth** or **lie:** artificial intelligence
- **Target audience:** topic modeling

Outline

1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

Text Analytics: Classification Pipeline



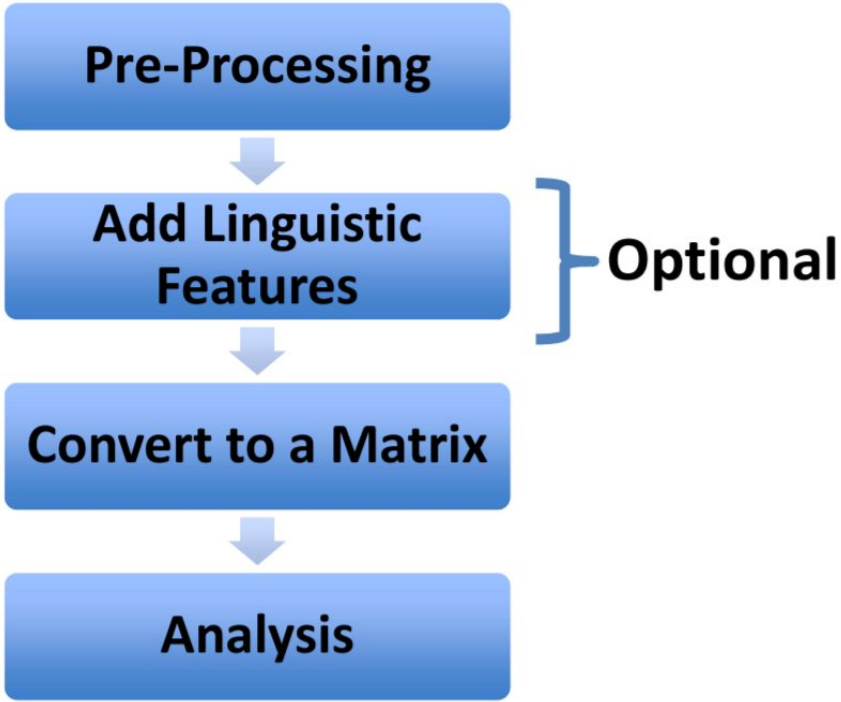
Statistical properties of text

- Length (of sentences, of documents, of corpora)
- Term frequency
- Document frequency
- Term co-occurrence

Tokenization

- Process of separating sentences and words from each other by syntax
- Different types of tokenizers: the simplest is by splitting at common punctuation
- **BUT:** a quote can be part of the same sentence; a dash can be part of same word or separate two parts of the same sentence; data quality can be low

NLP Pipeline

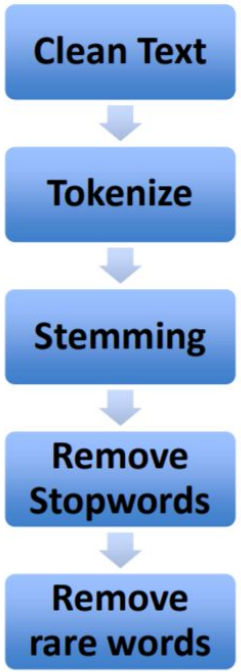


Raw data from a webpage

Our mission is to provide comprehensive ;nbsp social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.

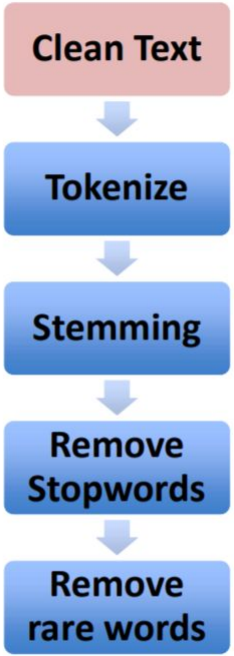
NLP Pipeline – Pre-Processing

Our mission is to provide comprehensive ;nbsp social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.



NLP Pipeline – Pre-Processing

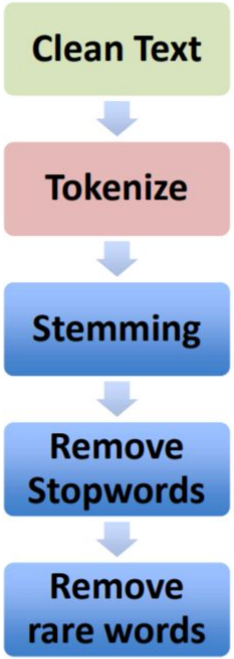
Our mission is to provide comprehensive social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.



NLP Pipeline – Pre-Processing

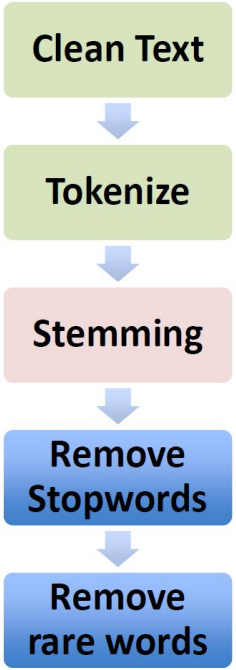
Our mission is to provide comprehensive social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.

Our	mission	is	to	provide	comprehensive	social	services	to	
refugees	to	help	them	overcome	the	societal	and	language	barriers
and	become	productive	members	of	the	US	society	.	



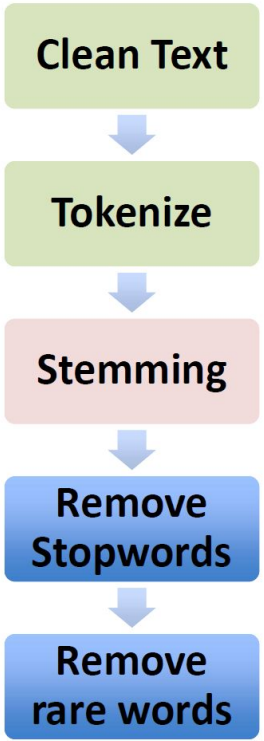
NLP Pipeline – Pre-Processing

Our mission is to provid comprehens social servic to refuge to help them overcom the societ and languag barrier and becom product member of the US societi .



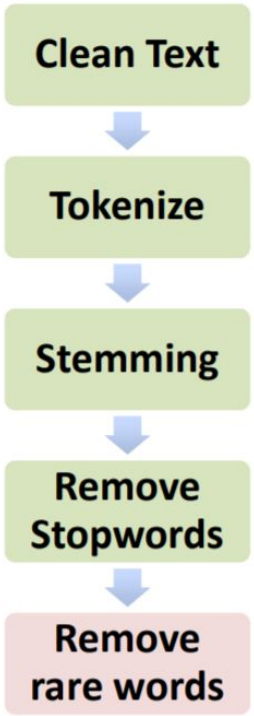
NLP Pipeline – Pre-Processing

Our mission is to provid comprehens social servic to refuge to help them overcom the societ and languag barrier and becom product member of the US societi .



NLP Pipeline – Pre-Processing

mission provid comprehens social servic refuge help overcom societ languag barrier becom product member US societi .



NLP Pipeline – Pre-Processing

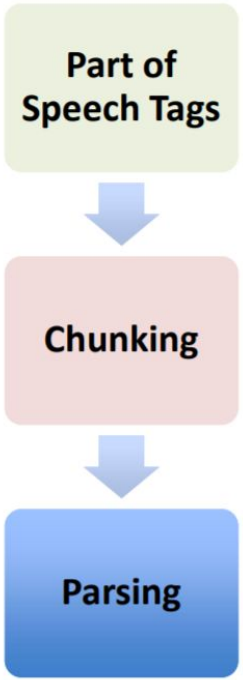
PRP$ Our **NN** mission **VBZ** is **TO** to **VB** provi de **JJ** comprehensive **JJ** social **NNS** services **T O** to **NNS** refugees **TO** to **VB** help **PRP** them **VB** overcome **DT** the **NN** societal **CC** and **NN** language **NNS** barriers **CC** and **VB** become **JJ** productive **NNS** members **IN** of **DT** the **NNP** US **NN** society **. .**

PRP Personal Pronoun
IN Preposition
NN Singular Noun
VBZ Verb, 3rd ps. sing. present

NLP Pipeline – Pre-Processing

NP Our mission **VP** is **VP** to provide **NP** comprehensive social services **PP** to **NP** refugees **VP** to help **NP** them **VP** overcome **NP** the societal and language barriers and **VP** become **NP** productive members **PP** of **NP** the US society .

NP	Noun Phrase
VP	Verb Phrase
PP	Prepositional Phrase



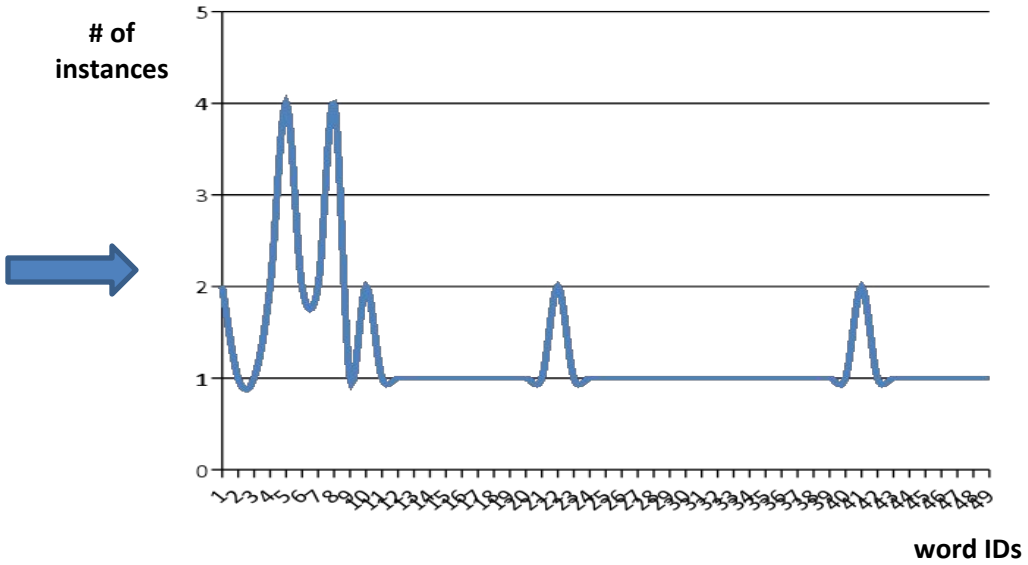
NLP Pipeline – Turning in to a Matrix

- What do you want the columns to be?
  - Words, phrases, POS tags, …
- What value do you put in a cell?
  - If a word appears in a document – binary variable
  - # of times a word in a document – word frequency
  - Words that uniquely characterize this document - tfidf

Text as a ‘bag-of-words’

Vector space representation of text – every word has its unique id (e.g., ‘comprehensive’=0, ‘social’=1, ‘language’=2, ‘members’=3, etc.) and the number of occurrences within the document:

Our mission is to provide comprehensive social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.



TFIDF – Term Frequency Inverse Document Frequency

- TF = Term Frequency (word count/ # words in the document)
- IDF = Inverse Document Frequency (how many documents does this word occur in?)

  log (# total documents / # documents this word appears in)

- TFIDF = TF X IDF

- highest when the word occurs many times within a small number of documents (thus lending high discriminating power to those documents)
- lower when the word occurs fewer times in a document, or occurs in many documents
- lowest when the word occurs in virtually all documents.

Processing text data: summary

- GIGO (Garbage In Garbage Out) – proper data processing is critical for the outcome
- Techniques differ dependent on the corpus properties and applications (what do we want to learn?)
- NLTK (Natural Language Toolkit in Python) and other software packages have good tools for processing text data

Outline

1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

Practical example of text analysis

Outline

1. Motivation
2. Definitions
3. Processing text data
4. Example approach
5. Evaluation
6. Approaches and Applications

Key takeaways

- Text is a useful resource to answer policy and research questions
- Text data is very diverse and requires substantial processing
- Data cleaning and processing are critical to success
- The choice of the right method depends on the question and application
- Evaluation is key to robust and valid results

Online Resources

- Interactive clustering tool Ontogen: ontogen.ijs.si
- Tutorial at http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf
- Online lectures at videolectures.net
- List of commercial software http://www.kdnuggets.com/software/text.html

BREAK UNTIL 3:00

Agenda

- Who we are
- Introduction to Big Data
- Web-scraping & APIs
  - BREAK
- Machine Learning
- Machine Learning applied to Record Linkage
  - LUNCH
- Text Analysis
- Providing Rich Context to Research & Data Discovery
  - BREAK
- Privacy & Confidentiality
  - ADJOURN

What is …

Privacy

Includes the famous “right to be left alone,” **and** the ability to share information selectively but not publicly (White House 2014)

Confidentiality

Means “preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information” (McCallister, Grance, and Scarfone 2010).

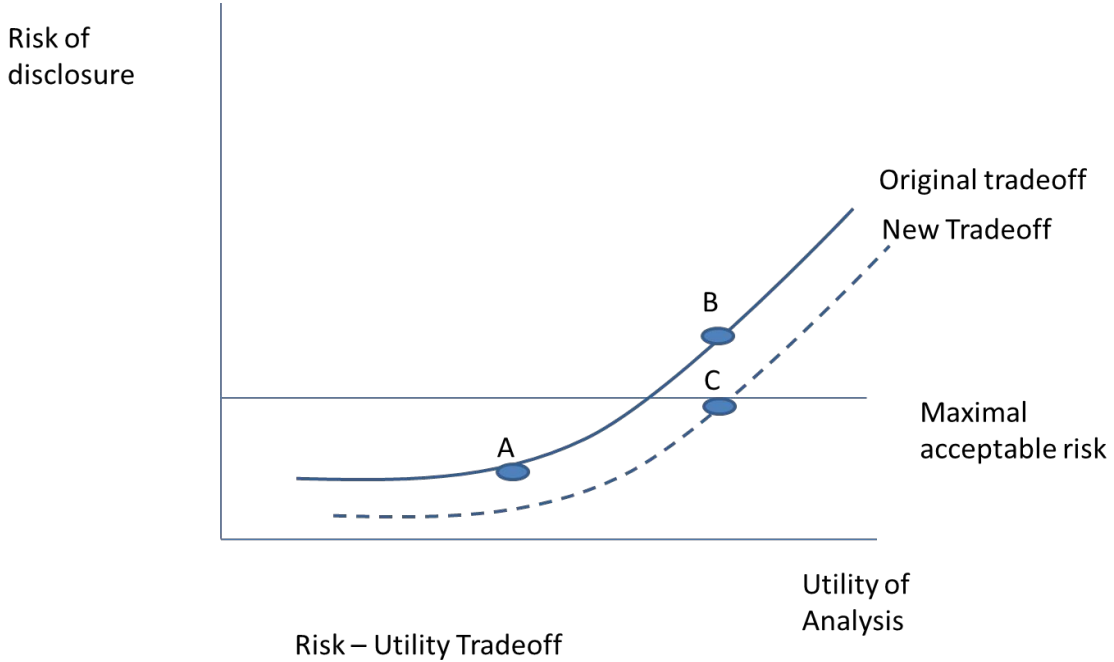
Why is confidentiality important

- Promise to respondents
- Ethical requirement
- Legal requirement
- Practical implications

Challenge

How to balance the ***risk*** of providing access with the associated utility?

Ballance & trade-off



Risk – Utility Tradeoff

Operational framework

Valid research purpose
- Statistical purpose
- Need “research benefit”

Trusted researchers

Limits on data use

Remote access to secure results
- **Disclosure control of results**

Safe projects

+ Safe people

+ Safe setting

+ Safe outputs

⇒ Safe use

What is disclosure?

Identity disclosure occurs when an individual can be identified from the released output, leading to information being provided about that identified subject.

Attribute disclosure occurs when confidential information is revealed and can be attributed to an individual. It is not necessary for a specific individual to be identified or for a specific value to be given for attribute disclosure to occur. For example, publishing a narrow range for the salary of persons exercising a particular profession in one region may constitute a disclosure.

Residual disclosure can occur when released information can be combined to obtain confidential data. While a table on its own might not disclose confidential information, disclosure can occur by combining information from several sources, including external ones.

Practicalities of disclosure control

The aim of disclosure control is to ensure that no unauthorised individual, technically competent with public data and private information could:

1) Identify any information not already public knowledge with a reasonable degree of confidence, and

2) Associate that information with the supplier of the information

Traditional Approaches (SDL)

Traditional approaches – tables

- cell suppression
- controlled tabular adjustment
- rounding
- cell perturbation

Traditional approaches – microdata

- top coding
- sampling
- rounding
- swapping
- added noise
- data shuffling

Specific examples

- Topcoding
  - Upper limit on values of a given variable, all cases above a certain part of the distribution are placed into one single category (wages)
  - Mean corrected top coding: choose the value for topcoded cells that the mean of the distribution is correct

- Noise addition
  - Multiplying or adding a stochastic or randomized number
  - Multiplicative noise: generating random numbers with mean=1

Specific examples (contd)

- Grouping, aggregating
  - Geographic population thresholds
  - Sensitive variables (nationality)
  - Rounding (age)

- Data Swapping
  - Introduce uncertainty, does not change the marginal distribution, but it distorts joint distributions of swapped and un-swapped variables

Disclosure control: primary disclosure

Threshold Rule: no cells with less than 3 units (individuals/enterprises)

Note: local unit analysis must show the enterprise count

This rule is applied even when there is no information associated with each cell

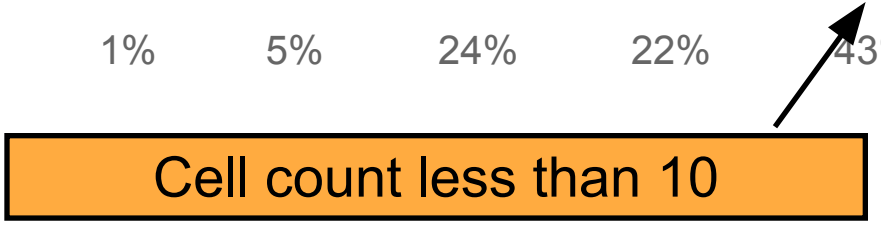
Example: manufacturing firms with over 1,000 employees by region

Region	Number of firms
North	152
South	8
East	12
West	6

Disclosure control: primary disclosure

% breakdown of hourly earnings by occupation

Pay bands: per hour	$5 to $6	$6 to $7	$7 to $8	$8 to $9	$9 to $10	>$10	Total numbers
Mechanics	15%	13%	32%	25%	10%	5%	1846
Nurses	13%	22%	57%	7%	1%	0%	949
Bankers	1%	5%	24%	22%	43%	5%	2059



Class disclosure: 0% values and 100% values might be problematic

Disclosure control: primary disclosure

Variable	Obs	Mean	Std. Dev	Min	Max
% PC users	3439	0.32	0.341	0	1
employees	3439	1413.7	5379.95		
Sales	3439	183323.7	694490.9		
Firm age	3439	6.5	2.08	1	15

Remove cells

?

- No max/min unless shown to be uninformative

Disclosure control: secondary disclosure

Some simple examples:

Sales: US-based companies

Industry	Companies	Sales
101	11	12003
102	10	5434
103	15	45644

Sales: All companies

Industry	Companies	Sales
101	14	16013
102	11	5579
103	19	65744

1 company in industry 102, foreign owned, with Sales of 145

Disclosure control, summary:

- Difficult to tell in advance – context specific
- Avoid / be careful when:
- Tabulating raw data (threshold rule)
- Using “lumpy” variables such as investment
- Researching small geographical areas (dominance rule)
- Graphs are tables in another form -> always display frequencies
- Treat quantiles as tables -> always display frequencies
- Avoid min/max values

Big Data Approaches

1. Hashing
2. Differential Privacy
3. Synthetic Data

Changes in the way of how information is collected (record linkage) and analysed are eroding the effectiveness of SDL

Hashing example of de-identification

- De-identification procedure that replaces direct identifiers with artificial IDs that do not allow for re-identification
- Applying an algorithm (hash function) according to NIST, FIPS Secure Hash Standards (FIPS PUB 180-4)
- Recommended security level of 112-bit security (provided for example by SHA-256)
- Adding a salt (secret or public)
- Computationally infeasible 1) to find an original ID that corresponds to a given hashed ID, or 2) to find two different original IDs that produce the same hashed ID, and thus re-identification is not possible (if properly designed).
- If two people apply the same algorithm to the same data, they will produce identical outputs, known as the hash value.

Example

Original Data:
Name: Peter Miller
SSN: 234-56-295
DOB: 07/12/76

Hashed Data:
Name: 9660ea4d6a0953035372678dd36da57a3e6f5c165605cbf73b6cfd778b7724a7
SSN: 6462c7069be22fc103f2edbad09b5763e0f1866af12ad9b578b8f9e1a89f8d87
DOB: 6c2c2cca5715ec5cd08bb7889dd12837976ead3b04201d88bc34173df52f7b11

Requires Pre-processing

Make sure IDs are unique, especially if there are different datasets that are supposed to be linked afterwards

Name harmonization through name standardization tables.
Rob, Bob, Robert

Harmonize date formats
07/12/76 (mm/dd/yyyy)
6c2c2cca5715ec5cd08bb7889dd12837976ead3b04201d88bc34173df52f7b11

12/07/76 (dd/mm/yyyy)
8365ad555ede04fdda11d0e9150358c4202213fdd4146f70470626327f75dee7

```
import datetime
import hashlib

class HashCache( object ):

    SALT_APPEND_RIGHT = "right"
    SALT_APPEND_LEFT = "left"

    def __init__( self, *args, **kwargs ):

        self.string_to_hash_map = {}
        self.cache_hit_count = 0
        print( "Cache initialized at " + str( datetime.datetime.now() ) )

    #-- END __init__() method --#

    def get_hash( self, string_to_hash_IN, salt_IN = "", append_salt_to_IN = SALT_APPEND_RIGHT ):

        '''
        Accepts string to hash and optional salt value.  If salt, appends it to the right of the string.
            Then, checks to see if that string is already in the hash map.  If so, retrieves hash of it.
            If not, hashes it and caches the hash.  Returns the hash, else None if there was an error.
        '''

        # return reference
        hash_OUT = ""

        # declare variables
        working_string = ""
        temp_hash = ""

        # pull string into working string:
        working_string = string_to_hash_IN

        # empty?
        if ( ( working_string is not None ) and ( working_string != "" ) and ( working_string != "NaN" ) ):

            # Got a value.  hash it.

            # is there a salt?
            if ( ( salt_IN is not None ) and ( salt_IN != "" ) ):

                # yes - use it.  Append to right or left?
                if ( append_salt_to_IN == self.SALT_APPEND_RIGHT ):

                    # right.
                    working_string = working_string + salt_IN

                else:

                    # if not right, left.
                    working_string = salt_IN + working_string

                #-- END check to see if right or left append --#

            #-- END check to see if salt. --#

            # check to see if hash in cache
            if ( working_string in self.string_to_hash_map ):

                # cached - record cache hit.
                self.cache_hit_count += 1

            else:

                # not cached.  Hash and cache.

                # encode to UTF-8
                temp_hash = working_string.encode( "utf-8" )

                # hash
                temp_hash = hashlib.sha256( temp_hash ).hexdigest()

                # cache
                self.string_to_hash_map[ working_string ] = temp_hash

            #-- END check to see if cached. --#

            # retrieve hash from cache.
            hash_OUT = self.string_to_hash_map[ working_string ]

        else:

            # No string in.  Leave empty.
            hash_OUT = ""

        #-- END check to see if empty --#


        return hash_OUT

    #-- END function get_hash() --#

#-- END class HashCache --#

print( "Object HashCache defined at " + str( datetime.datetime.now() ) )
```

Differential Privacy

- Differential privacy is a rigorous mathematical definition of privacy
- An algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not.
- The guarantee of a differentially private algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset
- This guarantee holds for *any* individual and *any* dataset

Differential Privacy II

- Motivation: bound the risk, if an individual decides to become part of the database
- Differential privacy ensures that the addition or removal of a single row in the database has (almost) no impact on the results of interest
- Implies that individuals can participate without risk
- Privacy is ensured through randomization
- Primarily studied in the context of collection, analysis and release of data
- DP is not a single tool but a definition or standard to manage privacy risks

What is the DP guarantee?

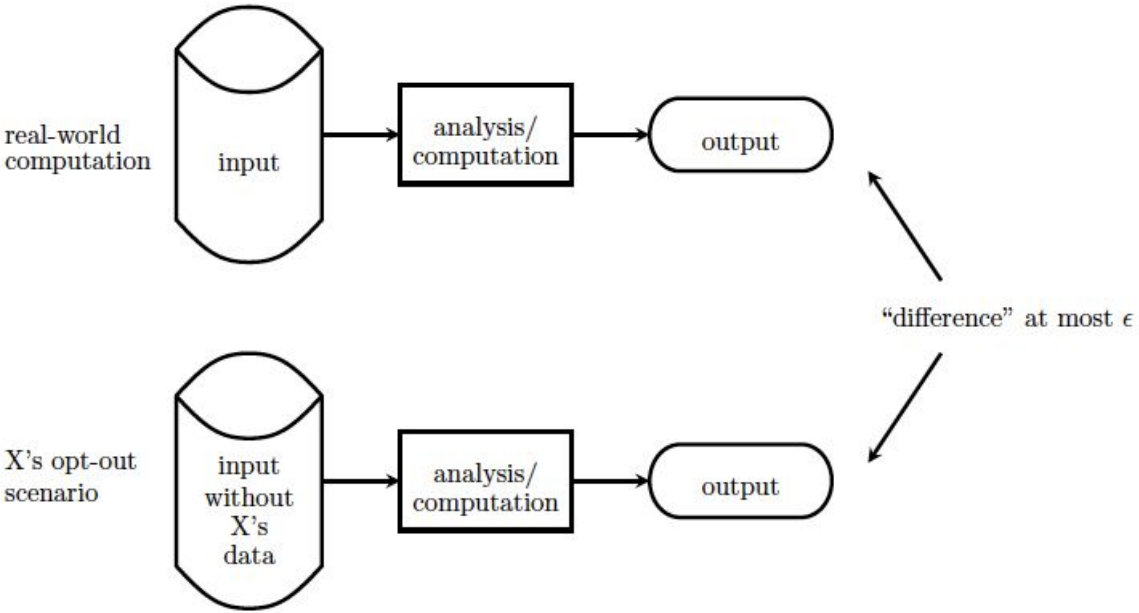
Researchers selected a sample of individuals to participate in a survey exploring the relationship between socioeconomic status and medical outcomes across a number of U.S. cities. Individual respondents were asked to complete a questionnaire covering topics such as where they live, their finances, and their medical history. One of the participants, John, is aware that individuals have been re-identified in previous releases of de-identified data and is concerned that personal information he provides about himself, such as his HIV status or annual income, could one day be revealed in de-identified data released from this study. If leaked, the personal information John provides in response to the questionnaire used in this tudy could lead to an increase in his life insurance premium or an adverse decision on a mortgage application he submits in the future.

Differential Privacy can be used to address John's concerns

Example

- If the analyses of this study is designed to be differentially private John is guaranteed that even though his information used the outcome will not disclose anything that is specific to him
- John's **opt out scenario**: The analysis is performed without including John's data. Thus his privacy is protected
- **Real world scenario**: The analysis is performed with all people's data
- DP: protect John in the real world scenario in a way that mimics the privacy protection of his opt out scenario
- Is achieved by adding randomness -> output is not exact but approximation

How Do We Add Randomness



Epsilon: privacy loss parameter

Captures deviation between opt-out and real world scenario

The effect of each individual's information on the output of the analysis

Smaller value is more privacy (0 = opt-out scenario)

Consider computing an estimate of the number of HIV-positive individuals in a sample, where the sample contains $n = 10,000$ individuals of whom $m = 38$ are HIV-positive. In a differentially private version of the computation, random noise Y is introduced into the count so as to hide the contribution of a single individual. That is, the result of the computation would be $m' = m + Y = 38 + Y$ instead of $m = 38$.

A researcher uses the estimate m', as defined in the previous example, to approximate the fraction p of HIV-positive people in the population. The computation would result in the estimate

$$p' = \frac{m'}{n} = \frac{38 + Y}{10,000}.$$

For instance, suppose the sampled noise is $Y = 4.2$. Then, the estimate would be

$$p' = \frac{38 + Y}{10,000} = \frac{38 + 4.2}{10,000} = \frac{42.2}{10,000} = 0.42\%,$$

whereas without added noise, the estimate would have been $p = 0.38\%$.

Types of analysis that can be done

- Count queries
- Histograms
- Cumulative Distribution Functions
- Linear Regressions
- Clustering
- Classification

Why Differential Privacy is so Attractive

- Only concept that offers formal privacy guarantees
- Independent of the data
- Doesn’t make any assumption about background knowledge
- No ex-post risk assessment required
- Privacy guarantees still hold even if other data sources are published later

Synthetic Data

Synthetic data is any production data applicable to a given situation that are not obtained by direct measurement

Imputed datasets are draws from the posterior predictive distribution of responses for a sample of individuals not included in the study, given the responses of the observed study sample

- Fit a model to the observed data. This is the imputation model
- Draw from the posterior distribution of the model parameters
- Generate new data from the predictive distribution of the data, given the imputation model structure and the value of the model parameters drawn in step 1

Example

Suppose we have a dataset with sex, age, gender:

- First we take a bootstrap sample of age to make the first column of the synthetic data age.syn
- Then we fit a logistic model to predict sex from age, using the real data and make the next column of the synthetic data by predicting sex from age.syn to get sex.syn
- Then we fit a model of marital status in terms of age and sex with the real data and make the next column of the synthetic data by predicting from age.syn and sex.syn to get maritalstatus.syn