Natural Language Processing
Session 1

Nick Kadochnikov

University of Chicago – MS Applied Data Science

About the Instructor (Nick!!!)

Nick Kadochnikov; kadochnikov@uchicago.edu
Associate Clinical Professor at UChicago PSD
20+ years of hands-on experience in data analytics to solve a variety of business problems

- ○ Social media and digital marketing analytics
- ○ Population health, medical prognosis, healthcare interoperability
- ○ Customer targeting and segmentation
- ○ Recommender systems, propensity to buy, cross-sell and up-sell modeling
- ○ Product development analytics
- ○ Fraud prevention

- Multiple data mining packages:
  - Python, Spark, SAS, SPSS, ILOG CPLEX, Netezza, Hadoop, Hive, Pig
- Worked with large volumes of structured and unstructured data (billions of records), including: transactional, financial, firmographic / demographic, organizational and macroeconomic data
- Worked in multiple countries across the globe
- Education:
  - MS in Global Marketing Management, Virginia Commonwealth University
  - MS and BS in Economics, St. Petersburg State University

Personal:

- ○ Love everything high speed: auto racing, go-karts, skiing and rollerblading
- ○ Passion for Renaissance and Baroque architecture

WORK EXPERIENCE

Recent work data includes company industry & sector classification, document classification, and information extraction

TA EXPERIENCE

NLP, Conversation AI Capstone, Reinforcement Learning

EDUCATION

M.S. in Analytics | University of Chicago | 2022

CFA charter | CFA Institute | 2019

B.S. in Physics | University of Chicago | 2014

IGNAS GRABAUSKAS

Data Scientist @ STBLaw

About Me

Avid learner and a data science enthusiast with close to 5 years of experience working at the intersection of data and technology across various domains including Supply Chain, Investment Banking and Legal Consulting. Passionate about NLP and truly appreciate the importance of regular expressions.

SWATHI GANESAN

AI Engineer

Education

2015 - 2019

SASTRA UNIVERSITY, India

B.Tech in Computer Sci. & Engg.

2022 - 2023

University of Chicago

M.S. in Applied Data Science

Experience

2019 - 2022

Decision Scientist - Mu Sigma, India

Jun ‘23 - Sep ‘23

Data Science Intern - William Blair, Chicago

Currently

AI Engineer - Harbor Global, Chicago

Personal

Love exploring new cultures and experiences through food, travel and books.
I enjoy cooking and grocery shopping.
Trader Joe’s is my happy place.
<Book recommendation: Becoming Trader Joe>

PROFESSIONAL BACKGROUND

Management Consultant @PwC

Data Analyst @One Mount Group

AI Engineer @Harbor Global

EDUCATION

BSc in Information System @Boston University

MSc in Applied Data Science @University of Chicago

PERSONAL

My Spotify Wrapped 2023

LINH LE

linhcle@uchicago.edu | Chicago, IL

Class rules

- I am not a “professor”
- My role is not to “teach”, but to share expertise and facilitate learning
- Please turn your Zoom video-on
- Please join the class on-time

Syllabus highlights

- Use of Generative AI
  - In this course, students are allowed to use AI tools (such as ChatGPT) on all assignments.
- Academic Integrity
  - Sharing course content is strictly prohibited. Under no circumstances are students permitted to:
  - Download lecture videos;
  - Share sample code on open-source websites;
  - Plagiarize individual assignments;
  - Discuss course content, whether verbally, or electronically, with other cohorts of MS in Applied Data Science students.

Class materials

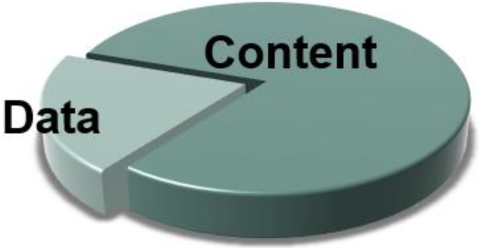
- Book
  - Speech and Language Processing, 2nd Edition by Daniel Jurafsky and James Martin
    - https://www.pearson.com/us/higher-education/program/Jurafsky-Speech-and-Language-Processing-2nd-Edition/PGM181706.html
  - You can also use free online PDF chapters from the draft version of the 3rd edition:
    - https://web.stanford.edu/~jurafsky/slp3/
- Software:
  - Python with Jupyter Notebooks
  - Word processing and presentation software
- Assignments
  - Hands-on experience with software and techniques
  - Your TA will be able to assists you each week as needed

Session 1 Agenda

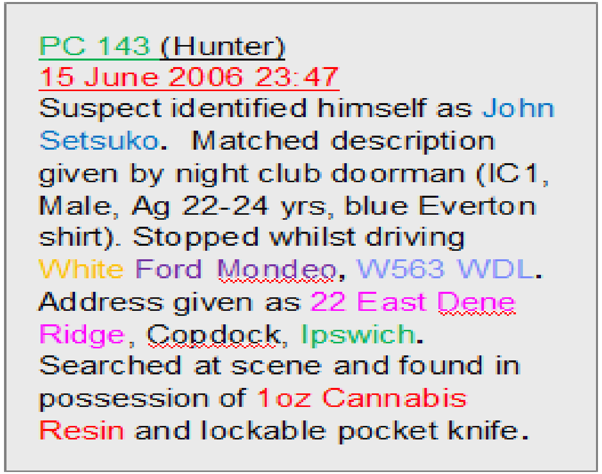
- NLP applications
    - Text Mining, Survey mining, Social Media Analytics
- Basic text processing
- Regular expressions

Introduction to Text Analytics

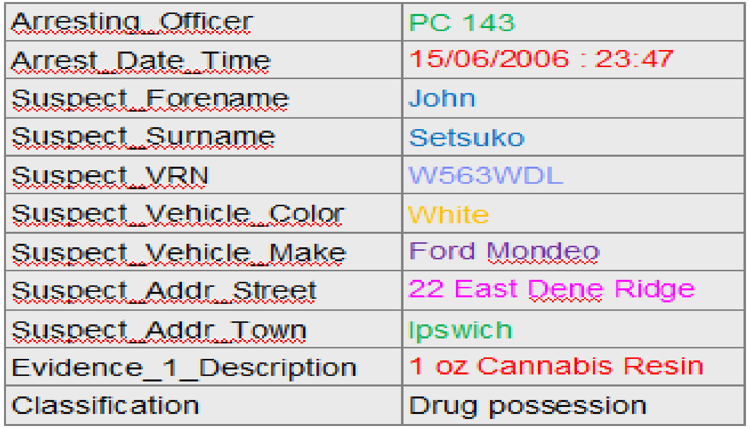
What is Text Analytics?



- Over 80% of information being stored is unstructured
- Text analytics unlocks the power of that information for a variety of functions and applications



PC 143 (Hunter)
15 June 2006 23:47
Suspect identified himself as John Setsuko. Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving White Ford Mondeo, W563 WDL. Address given as 22 East Dene Ridge, Copdock, Ipswich. Searched at scene and found in possession of 1oz Cannabis Resin and lockable pocket knife.

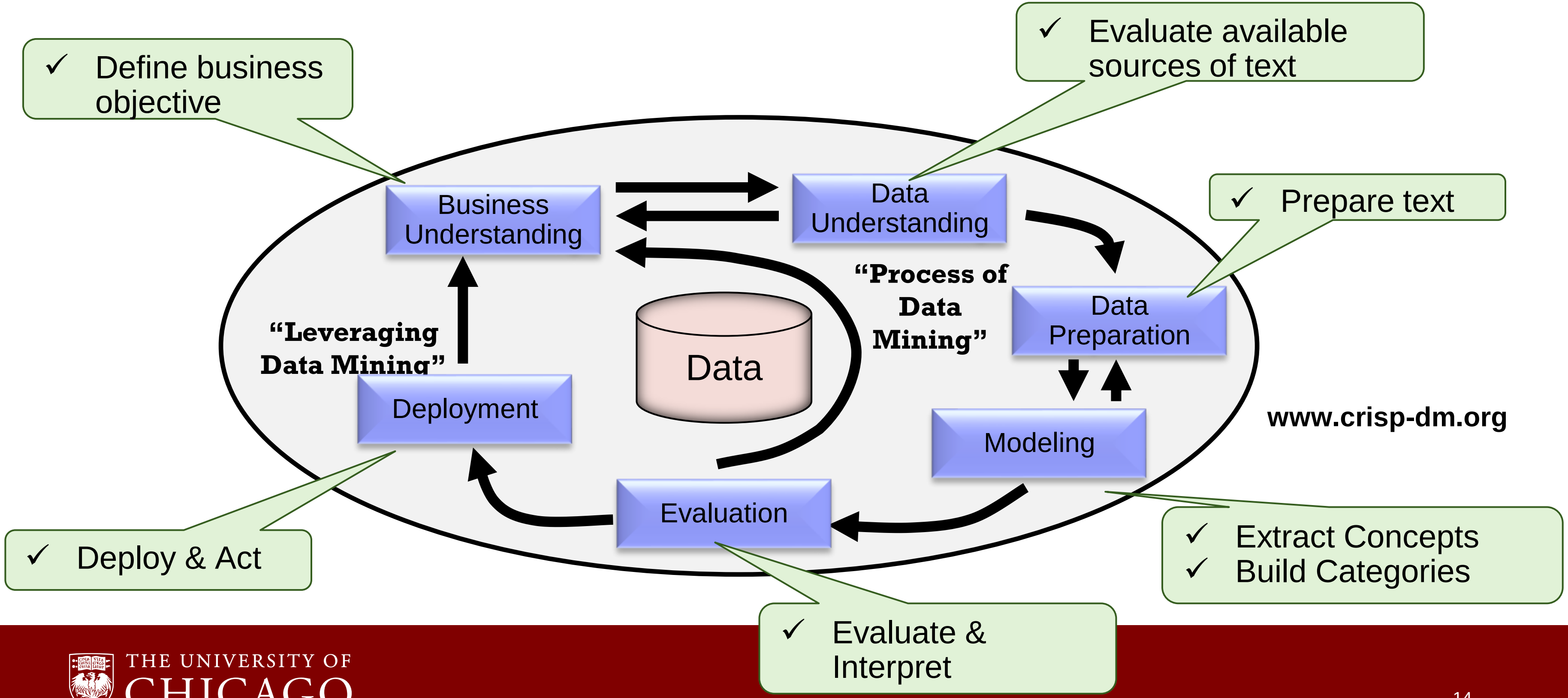


Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Color	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession

Identify email address in police report

Identify car make / model in police report

CRISP-DM Methodology applies to text mining



- ✓ Define business objective
- ✓ Evaluate available sources of text
- ✓ Prepare text
- ✓ Extract Concepts
- ✓ Build Categories
- ✓ Evaluate & Interpret
- ✓ Deploy & Act

Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

Data

“Process of Data Mining”

“Leveraging Data Mining”

www.crisp-dm.org

Data Mining vs. Text Mining

- In traditional data mining application you can either train the model on target variable (supervised modeling), or let the model find natural patterns in the data (i.e. unsupervised clustering)
- Same concepts apply to NLP problems!
    - You can build NLP classifiers
        - Target variable can be topic, sentiment, etc.
    - You can build NLP clusters
        - i.e. topic modeling, document / sentence clustering, etc.

Text Mining and Data Preparation

Data Mining

- Data cleaning
  - Selecting relevant data
  - Data quality – errors
  - Interpreting and handling missing data
- Data transformation
  - Get the data into the right form to ask the relevant questions…
- An iterative process

Text Mining

- Data cleaning
  - Synonyms, abbreviations, specialized vocabulary, common typos
- Data transformation
  - Simplify text:
    - Stemming, lemmatizations, etc.

Social Media Analytics

What is the difference between Social Media Analytics

and

traditional text mining / natural language processing (NLP)?

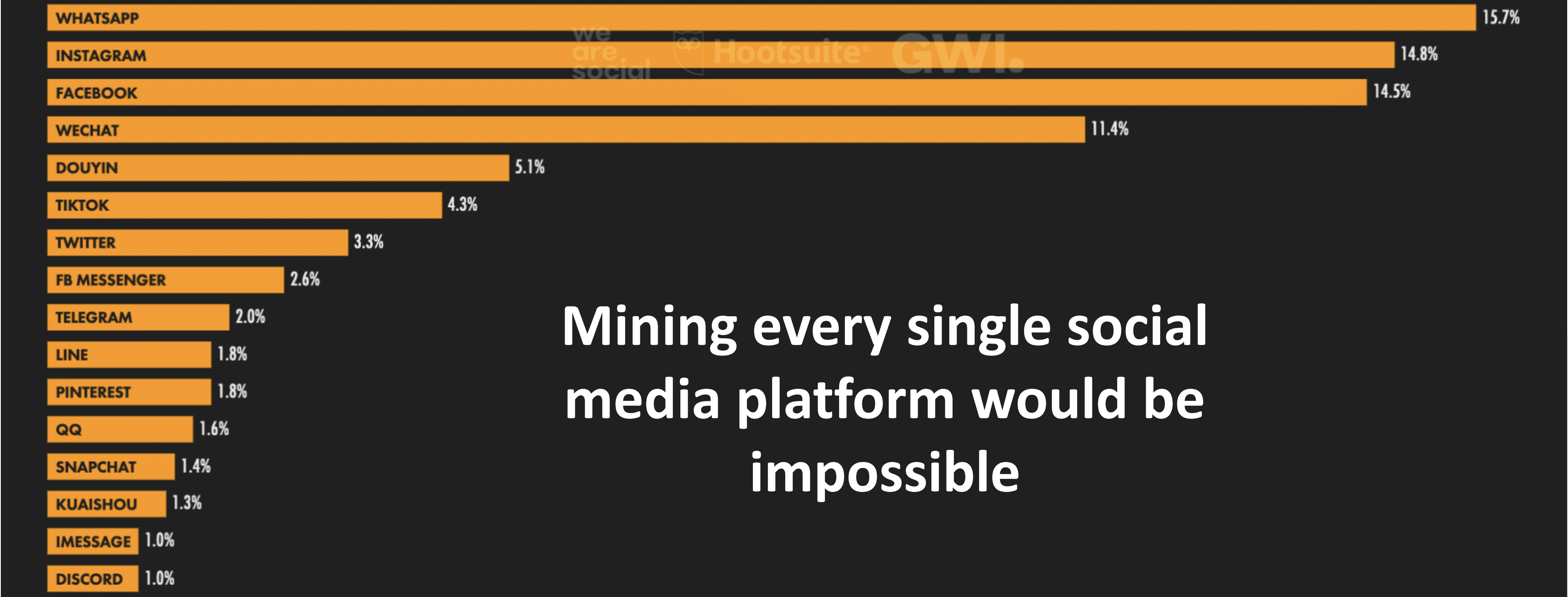
What about survey mining

JAN 2022

FAVOURITE SOCIAL MEDIA PLATFORMS

PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO SAY THAT EACH OPTION IS THEIR "FAVOURITE" SOCIAL MEDIA PLATFORM

GLOBAL OVERVIEW



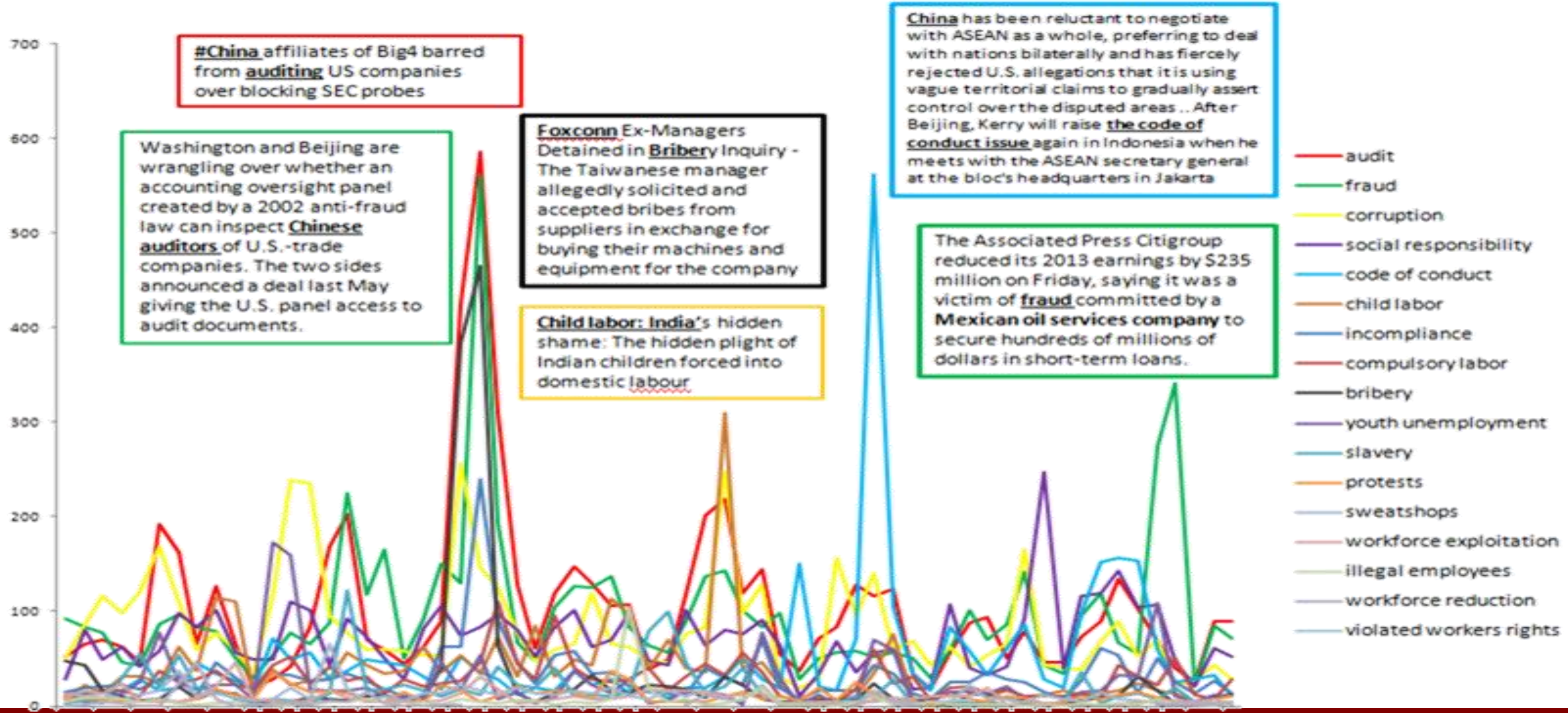
Mining every single social media platform would be impossible

SOURCE: GWI (Q3 2021). SEE GWI.COM FOR FULL DETAILS. **NOTES:** ONLY INCLUDES USERS AGED 16 TO 64. SURVEY RESPONDENTS COULD CHOOSE FROM OTHER OPTIONS NOT SHOWN ON THIS CHART, SO VALUES MAY NOT SUM TO 100%. YOUTUBE IS NOT AVAILABLE AS AN ANSWER FOR THIS QUESTION IN GWI'S SURVEY. WE REPORT GWI'S VALUES FOR TIKTOK IN CHINA SEPARATELY AS DOUYIN, AS PER BYTEDANCE'S CORPORATE REPORTING. **COMPARABILITY:** VERSIONS OF THIS CHART THAT FEATURED IN OUR PREVIOUS REPORTS DID NOT INCLUDE DATA FOR CHINA, SO VALUES ARE NOT COMPARABLE.

The solution comes in form of data aggregation services

	Twitter	Reviews/message boards	Blogs	Forums	Online news articles	Video comments
Data Aggregators						
Twitter (Gnip)	•	•	•	•	•	•
Topsy	•					
DataSift	•	•		•	•	•
Boardreader		•	•	•	•	•
Trendiction	•	•	•	•	•	•

Supply Chain Risks: Distributions of mentions by topics



#China affiliates of Big4 barred from auditing US companies over blocking SEC probes

Washington and Beijing are wrangling over whether an accounting oversight panel created by a 2002 anti-fraud law can inspect Chinese auditors of U.S.-trade companies. The two sides announced a deal last May giving the U.S. panel access to audit documents.

Foxconn Ex-Managers Detained in Bribery Inquiry - The Taiwanese manager allegedly solicited and accepted bribes from suppliers in exchange for buying their machines and equipment for the company

Child labor: India's hidden shame: The hidden plight of Indian children forced into domestic labour

China has been reluctant to negotiate with ASEAN as a whole, preferring to deal with nations bilaterally and has fiercely rejected U.S. allegations that it is using vague territorial claims to gradually assert control over the disputed areas.. After Beijing, Kerry will raise the code of conduct issue again in Indonesia when he meets with the ASEAN secretary general at the bloc's headquarters in Jakarta

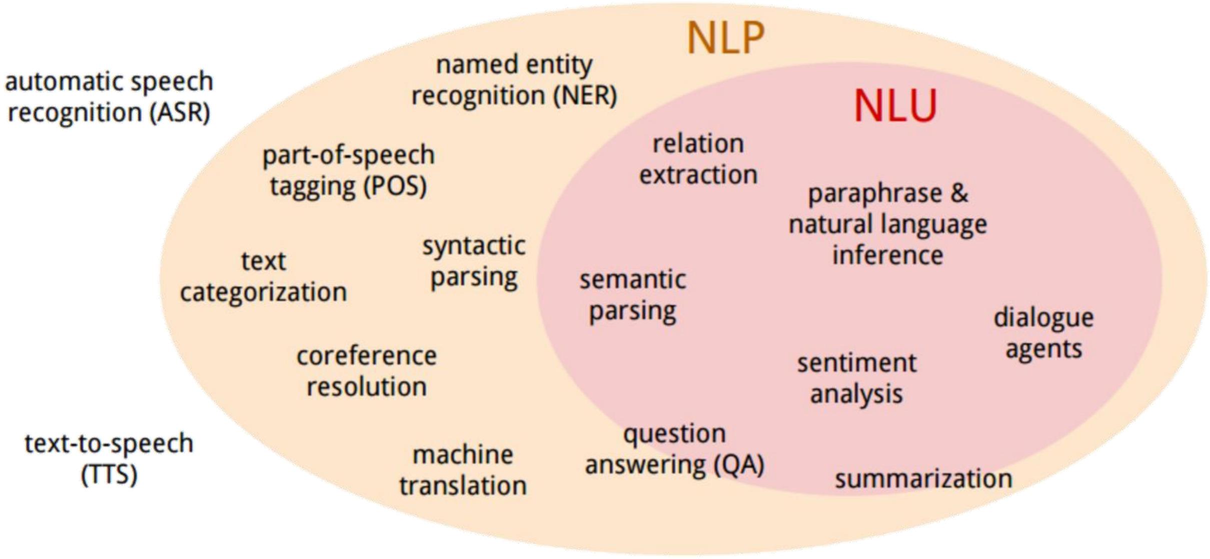
The Associated Press Citigroup reduced its 2013 earnings by $235 million on Friday, saying it was a victim of fraud committed by a **Mexican oil services company** to secure hundreds of millions of dollars in short-term loans.

NLP Tasks and Process

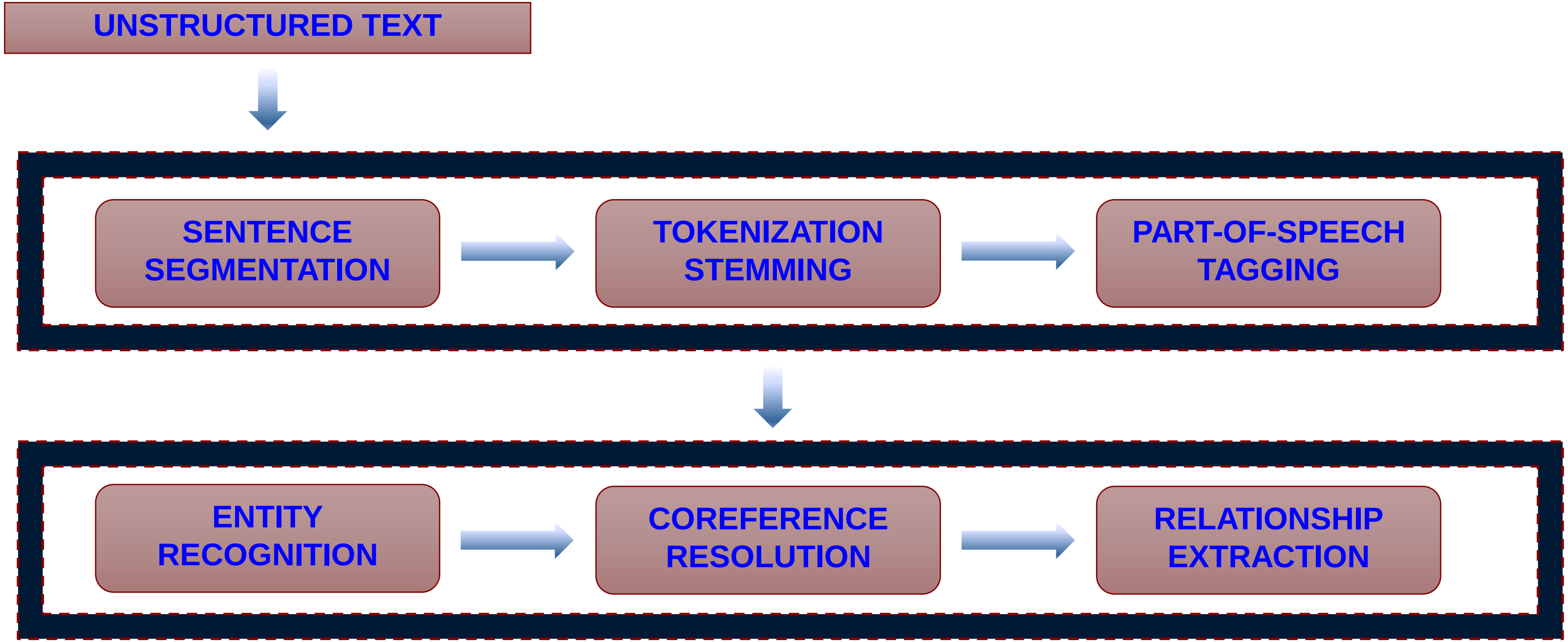
NLP: Big Picture

Natural Language Processing (NLP) is a field of computer science and artificial intelligence, concerned with making computers process natural (human) language

Computational Linguistics (CL) is the field of using computers to understand language



Basic Text Processing & Information Extraction (IE)



UNSTRUCTURED TEXT

SENTENCE SEGMENTATION

TOKENIZATION STEMMING

PART-OF-SPEECH TAGGING

ENTITY RECOGNITION

COREFERENCE RESOLUTION

RELATIONSHIP EXTRACTION

Identify top three most frequent beverages consumed in the book

Basic Text Processing

Regular Expressions

Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges `[A-Z]`

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Caret means negation only when first in []

Pattern	Matches	
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither ‘S’ nor ‘s’	`I have no exquisite reason”`
`[^e^]`	Neither e nor ^	`Look here`
`a^b`	The pattern a caret b	`Look up a^b now`

Regular Expressions: More Disjunction

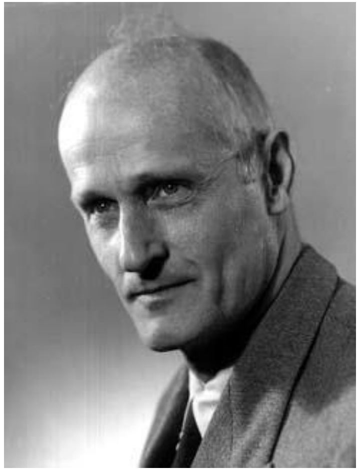
- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
groundhog|woodchuck	
yours|mine	yours mine
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	

Photo D. Fletcher

Regular Expressions: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n



Stephen C Kleene

Kleene *, Kleene +

Regular Expressions: Anchors ^ $

Pattern	Matches
`^[A-Z]`	Palo Alto
`^[^A-Za-z]`	1 "Hello"
`\.$`	The end.
`.$`	The end? The end!

Example

- Find me all instances of the word “the” in a text.

`the` **Misses capitalized examples**

`[tT]he` **Incorrectly returns `other` or `theology`**

`[^a-zA-Z][tT]he[^a-zA-Z]`

Regular Expressions in Python (re)

Errors

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Errors cont.

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
  - Increasing accuracy or precision (minimizing false positives)
  - Increasing coverage or recall (minimizing false negatives).

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

Thank You!