

# Social Media & Text Analysis

lecture 1 - Introduction

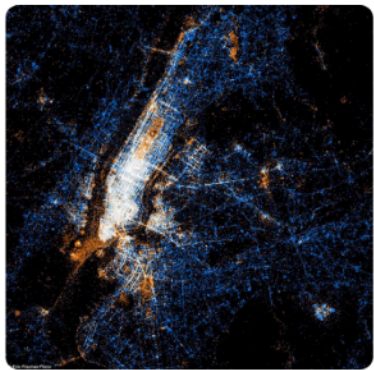
**CSE 5539 Ohio State University**

**Instructor: Alan Ritter**

**Website: [socialmedia-class.org](http://socialmedia-class.org)**

# Course Website

## <http://socialmedia-class.org/>



*A visualization showing the location of Twitter messages (blue) and Flickr photos (orange) in New York City by Eric Fischer*

Social media provides a massive amount of valuable information and shows us how language is actually used by lots of people. This course will give an overview of prominent research findings on language use in social media. The course will also cover several machine learning algorithms and the core natural language processing techniques for obtaining and processing Twitter data.

### Instructor

[Alan Ritter](#)

### Current

**Autumn 2019, CSE 5539-0010** The Ohio State University

dual-listed undergraduate and graduate course

**Time/Place:** Fri 11:30am-1:35pm | Jennings Hall 140

**Office Hours:** Fri 4:00pm-5:00pm | Dreese Lab 595

### Prerequisites

In order to succeed in this course, you should know basic probability and statistics, such as the chain rule of probability and Bayes' rule; some basic calculus and linear algebra will also help, such as knowing what is gradient. On the programming side, all projects will be in Python. You should understand basic computer science concepts (like recursion), basic data structures (trees, graphs), and basic algorithms (search, sorting, etc).

### Course Readings

Each lecture has an accompanying set of [academic papers](#)

### Resources

**Piazza** (discussion, announcements and restricted resources)

**Carmen** (homework submission and grades)

# This is a **special topic class**

- hobby (not a mandatory course)
- but is lecture-based and project-based
- advanced and research-oriented
- but strong undergraduate students (sophomore, junior, senior) are encouraged to take this course

Who am I?



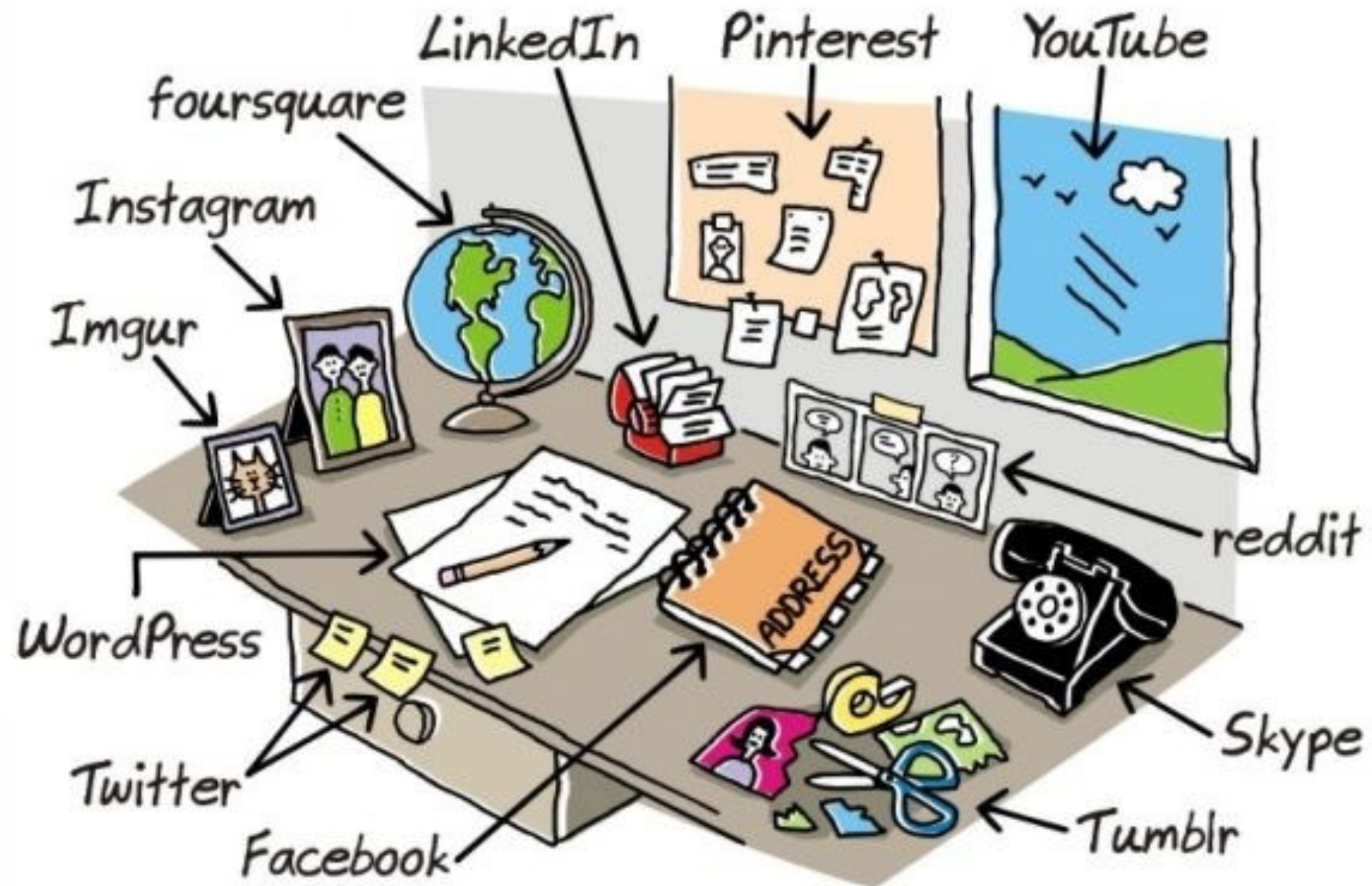


# Alan Ritter

- Assistant Professor in CSE at the Ohio State University
- Research Areas:
  - Natural Language Processing
  - Information Extraction
  - Dialogue
  - Social Media Analysis
  - Machine Learning

Why Social Media?

# Vintage Social Media



<http://wronghands1.wordpress.com>

© John Atkinson, Wrong Hands





**skip**  
@han\_horan

so my plane just crashed...  
[pic.twitter.com/X51BLwa5PS](https://pic.twitter.com/X51BLwa5PS)

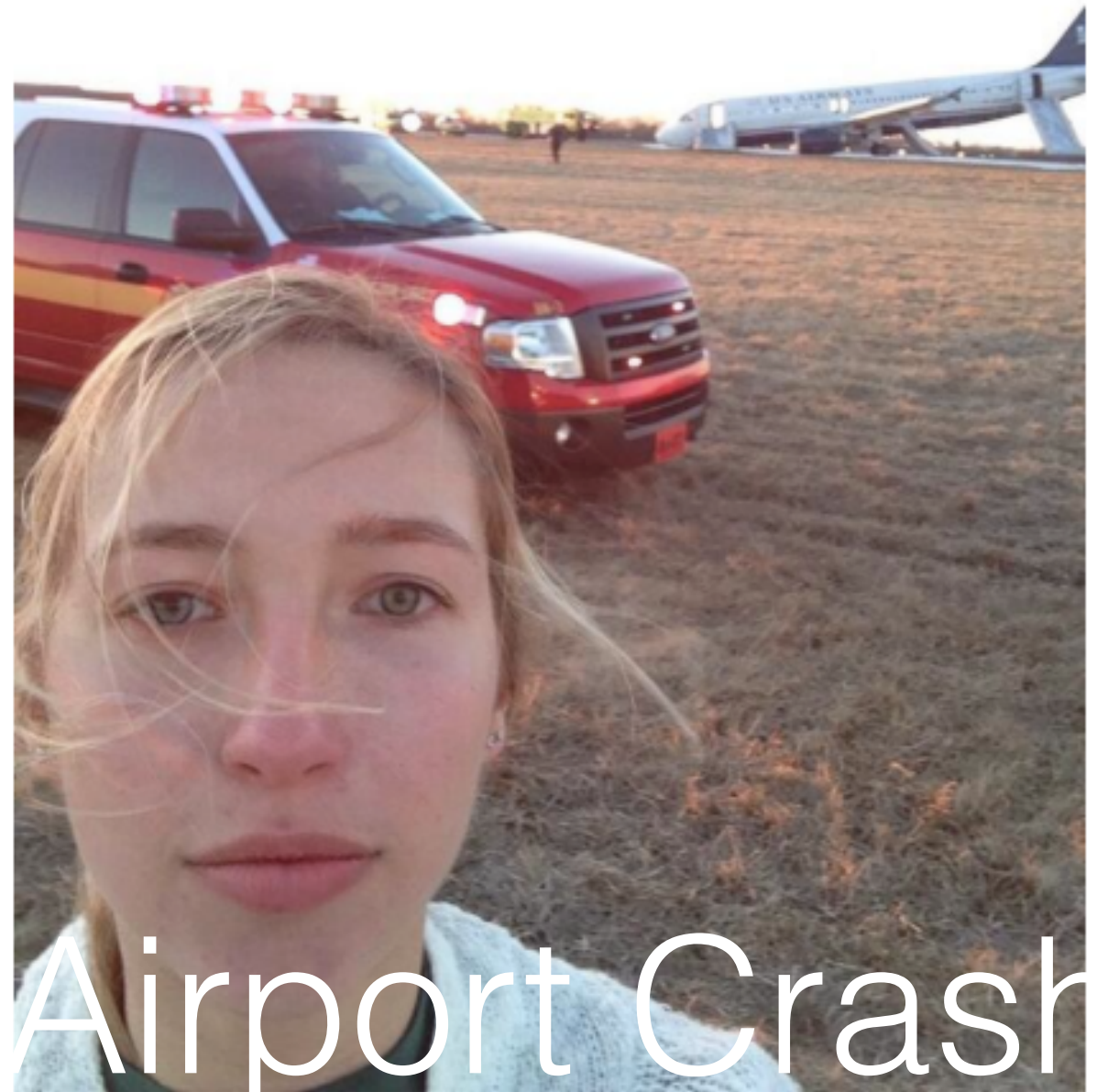
↩ Reply ↻ Retweet ★ Favorite ⋮ More



**skip**  
@han\_horan

so yup [pic.twitter.com/2WuLUWzpND](https://pic.twitter.com/2WuLUWzpND)

↩ Reply ↻ Retweet ★ Favorite ⋮ More



2014 Philly Airport Crash



# 2014 Ukrainian Revolution



# AP Account Hack



# Impact

- Politics
- Business
- Socialization
- Journalism
- Cyber Bullying
- Rumors / Fake News
- Productivity
- Privacy
- Emotions
- ...
- and our language (!)



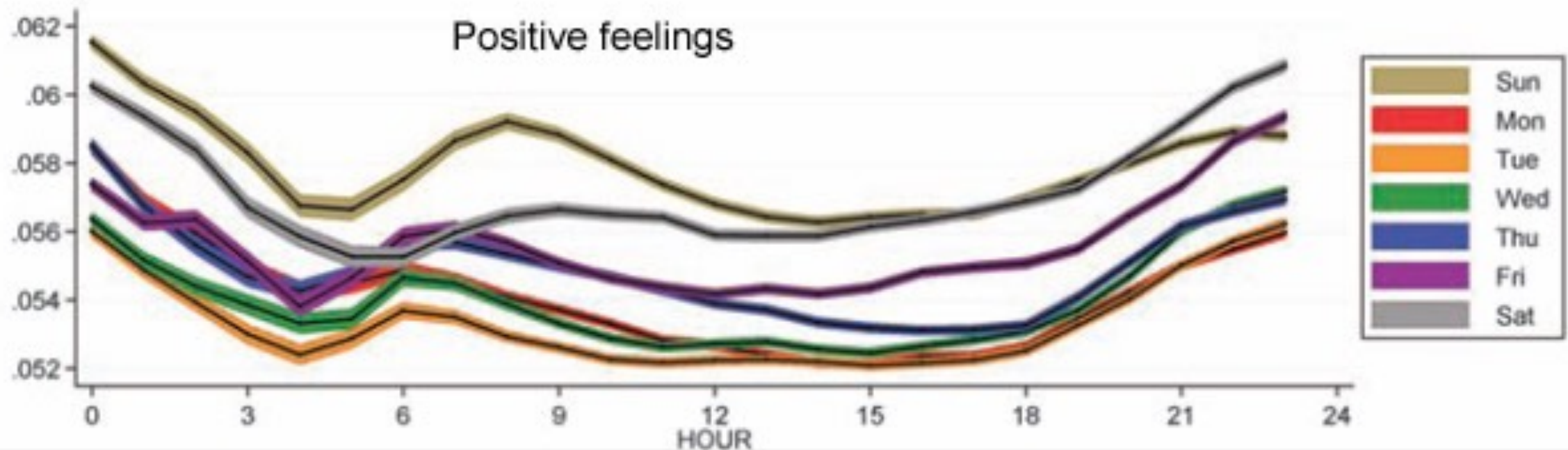


# Research Value

- ▶ In contrast to survey/self-report
- ▶ A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- ▶ Easy to access and aggregate **a lot** of data
- ▶ thus **a lot** of information



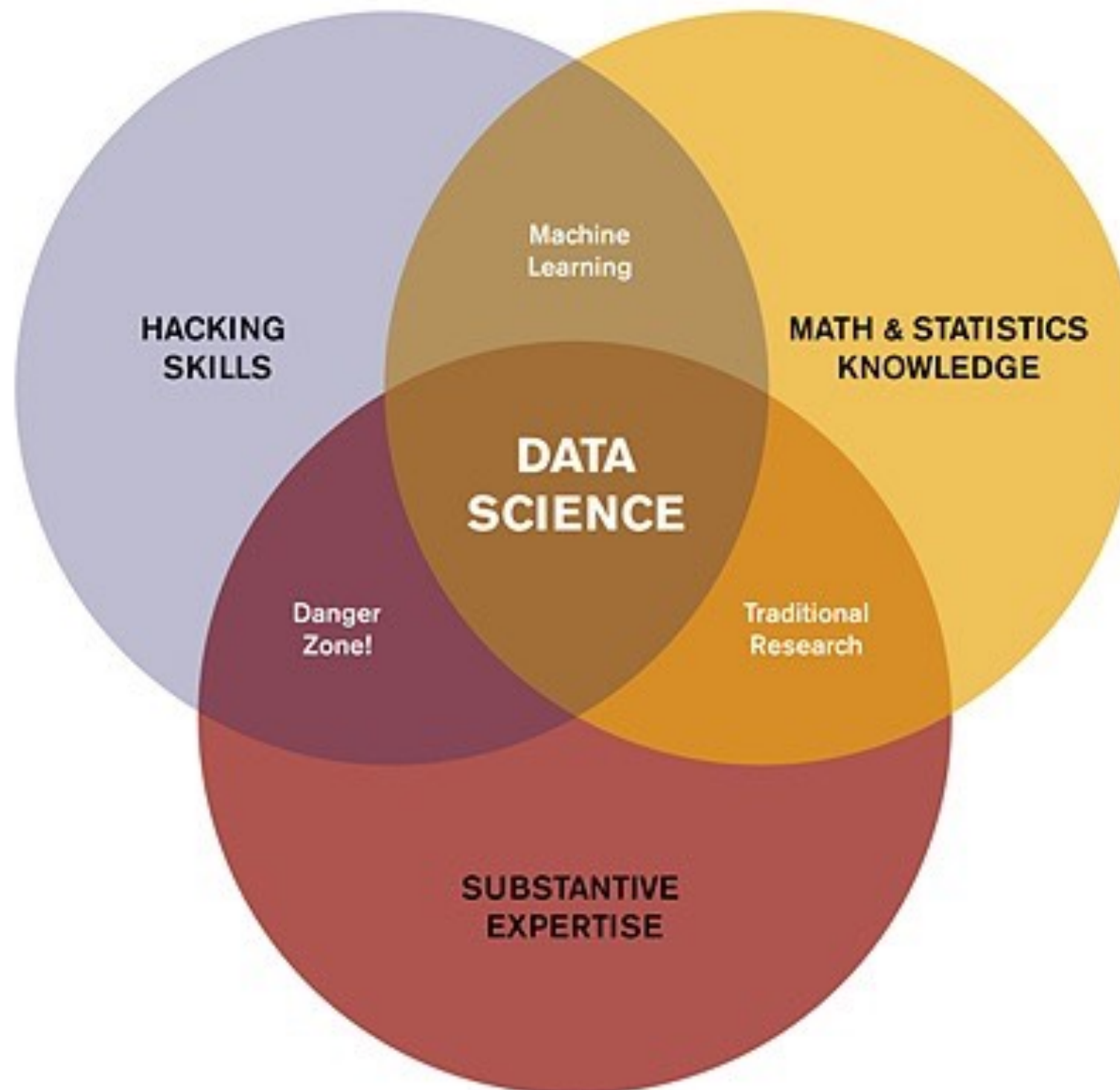
# Mood



“We found that individuals awaken in a good mood that deteriorates as the day progresses—which is consistent with the effects of sleep and circadian rhythm”

“People are happier on weekends, but the morning peak in positive affect is delayed by 2 hours, which suggests that people awaken later on weekends.”

# Data Science



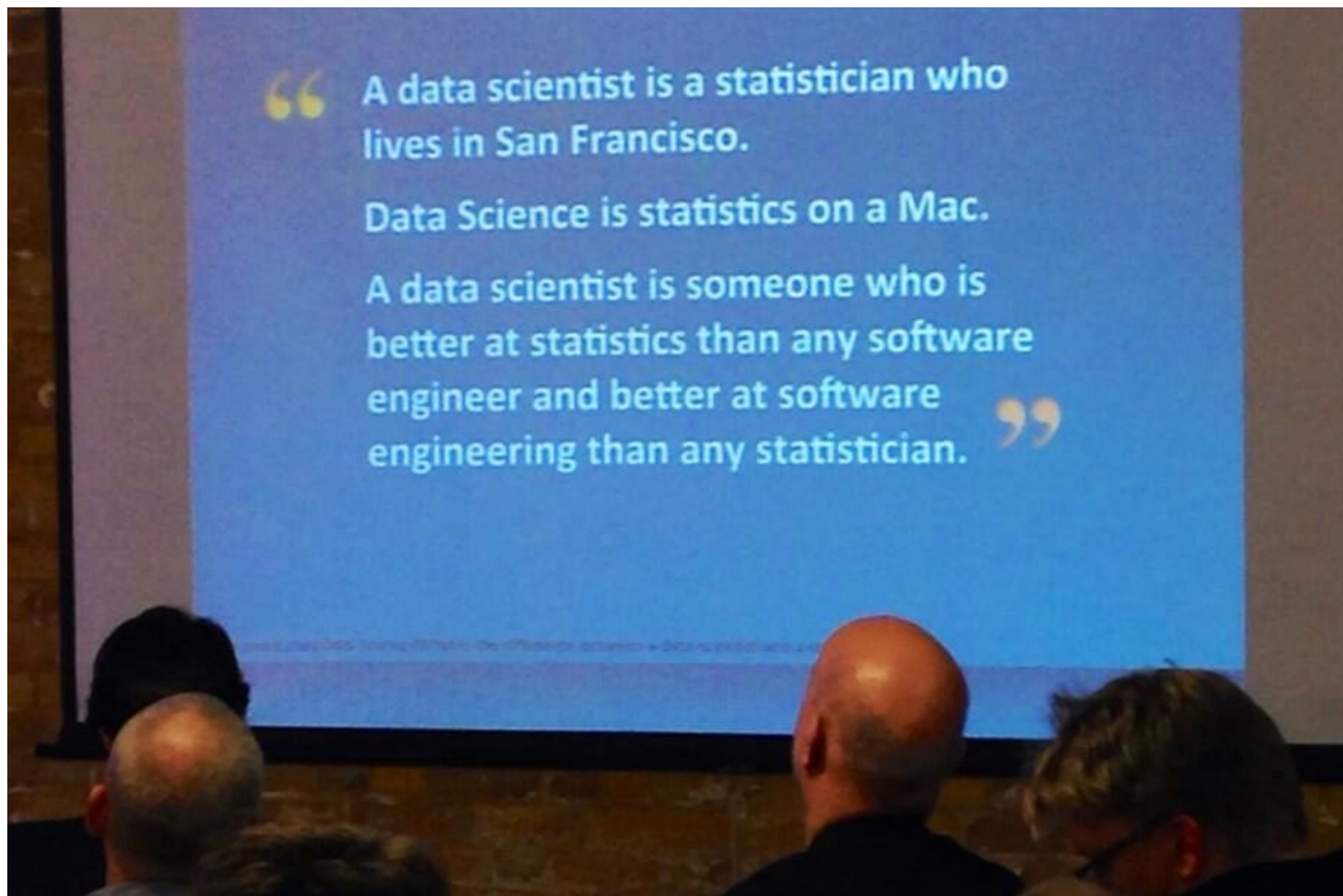
# Data Science

- ▶ is the **practice** of:
  - asking question (formulating hypothesis)
  - finding and collecting the data needed (often big data)
  - performing statistical and/or predictive analytics (often machine learning)
  - discovering important information and/or insights



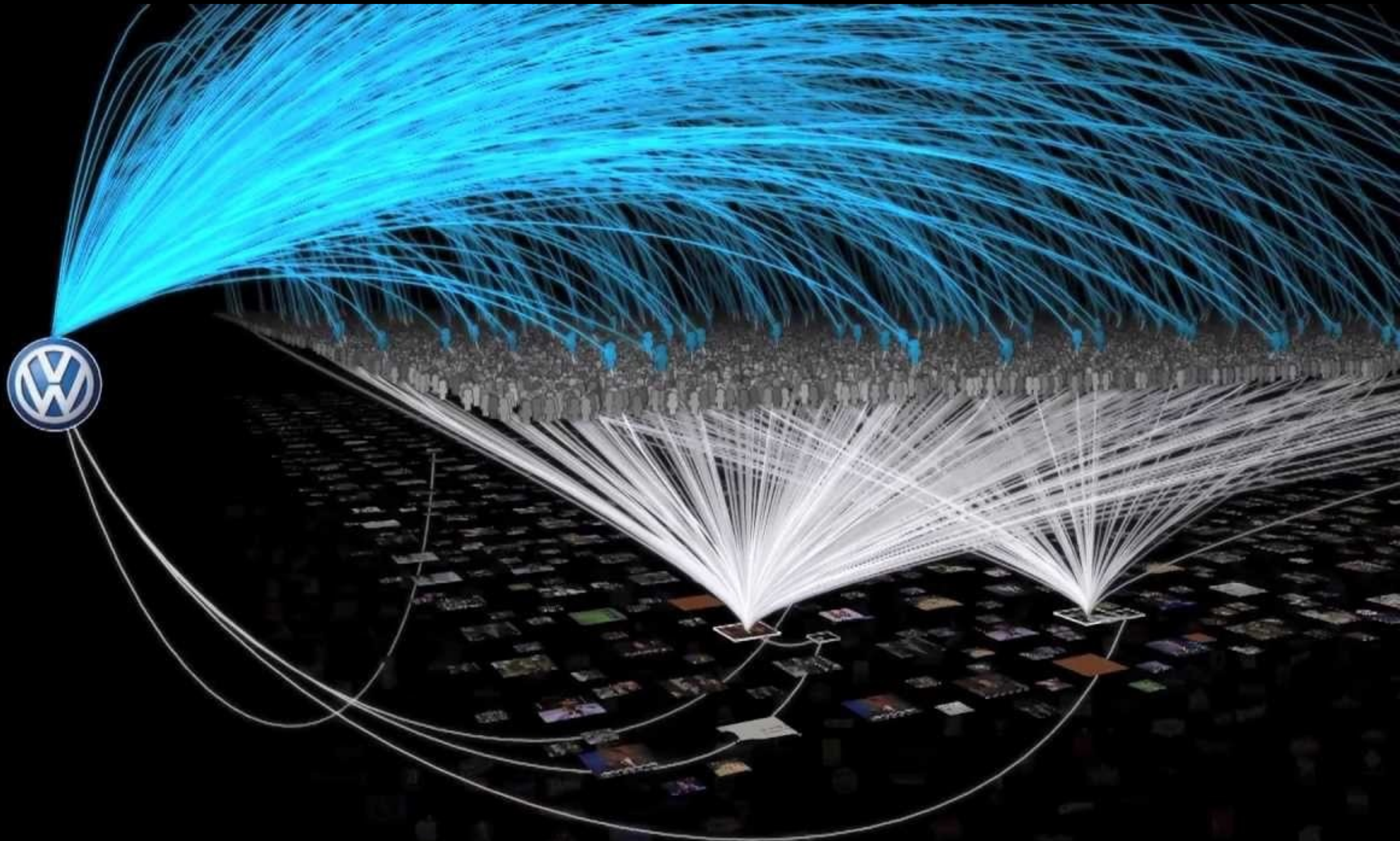
# Data Science

- the infamous definition:





# Marketing



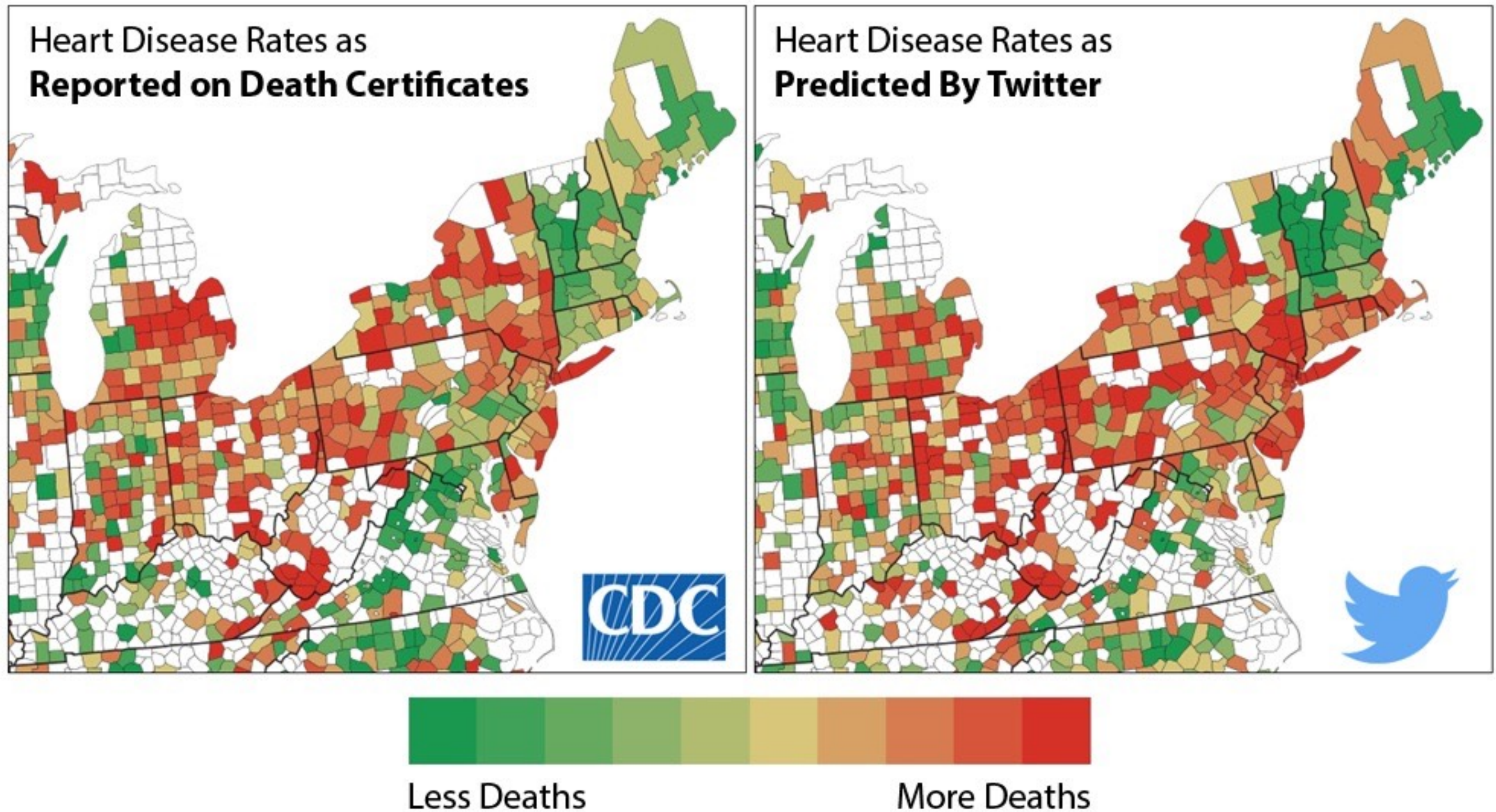
# User Profiling

    
Delighted I kept my Xmas vouchers - Happy  
Friday to me 😊 #shopping





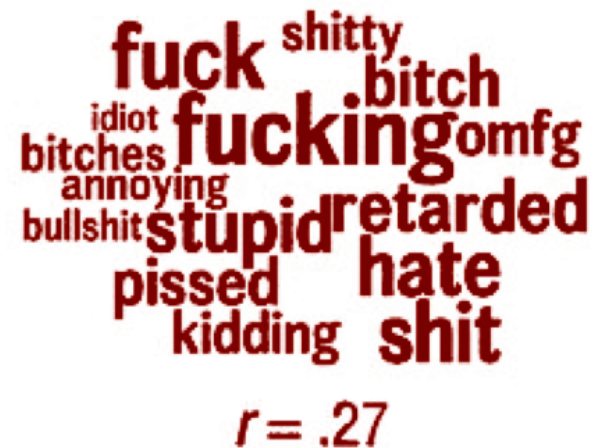
# Health



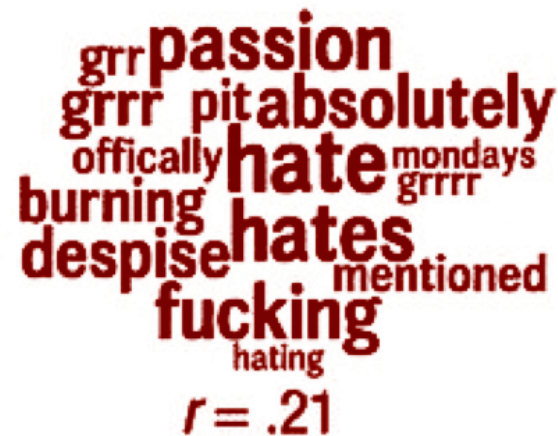


# Health

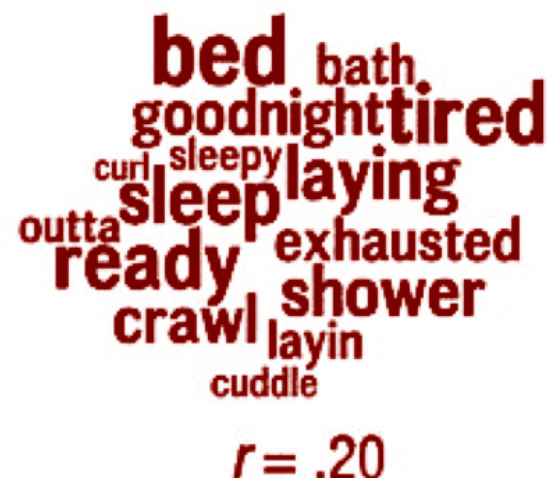
Hostility,  
Aggression



Hate,  
Interpersonal  
Tension



Boredom,  
Fatigue



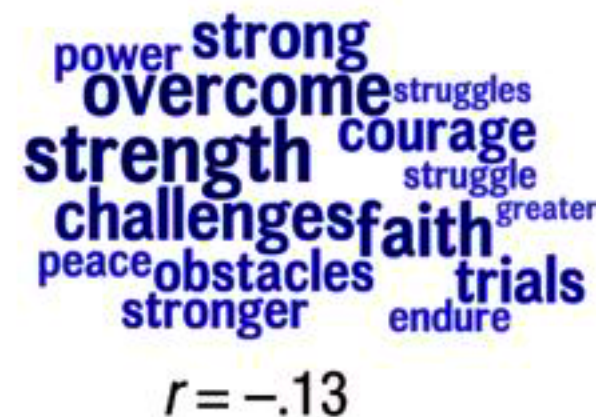
Skilled  
Occupations



Positive  
Experiences



Optimism





# Problems on Social Media

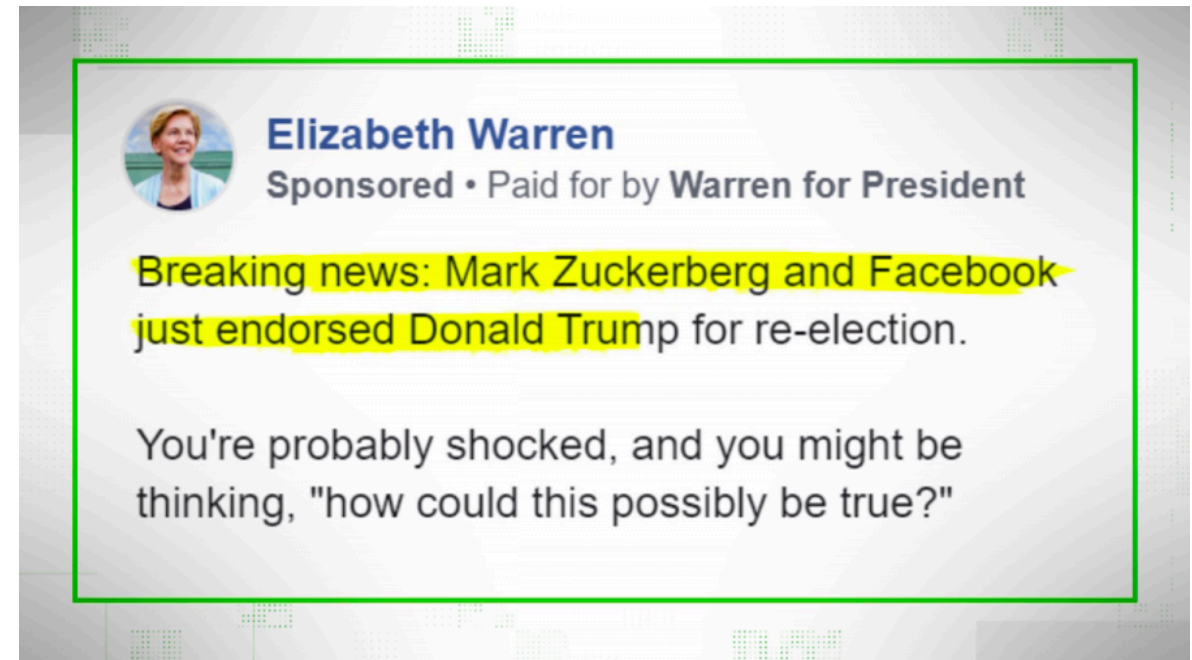
Hate Speech



*The New York Times*

***Facebook Admits It Was Used  
to Incite Violence in Myanmar***

False and Misleading Content



*The New York Times*

***Warren Dares Facebook With  
Intentionally False Political Ad***

**Maybe Natural Language Processing can Help?**

What is Natural  
Language Processing?

# Sentiment Analysis



*This nets vs bulls game is **great***

*This Nets vs Bulls game is **nuts***

***Wowzers** to this nets bulls game*

*this Nets vs Bulls game is **too live***

*This Nets and Bulls game is a **good** game*

*This netsbulls game is **too good***

*This NetsBulls series is **intense***

# Named Entity Recognition

sportsteam sportsteam geo-loc  
India vs Australia 2014-15 , 4th Test in Sydney

company product  
Samsung to launch Galaxy S6 in March

tvshow tvshow  
New Suits and Brooklyn Nine-Nine tomorrow ... Happy days

# Machine Translation

The screenshot shows the Google Translate web interface. At the top is the Google logo. Below it, the word "Translate" is displayed in red. To the right of "Translate" is a link that says "Turn off instant translation" and a star icon. Below this, there are two language selection bars. The left bar has buttons for "English", "Spanish", "French", and "Detect language". The right bar has buttons for "English", "Spanish", and "German", followed by a dropdown arrow. A blue "Translate" button is positioned to the right of the right language bar. Below the language bars, there are two text input areas. The left area contains the text "To the airport, please." and has a close button (X) in the top right corner. Below this text are icons for voice input, speaker output, and a keyboard icon. The right area contains the translated text "Bis zum Flughafen, bitte." and has icons for a star, copy, speaker output, and share, along with a pencil icon for editing.

Google

Translate

Turn off instant translation

English Spanish French Detect language

English Spanish German

Translate

To the airport, please.

Bis zum Flughafen, bitte.

# Information Extraction

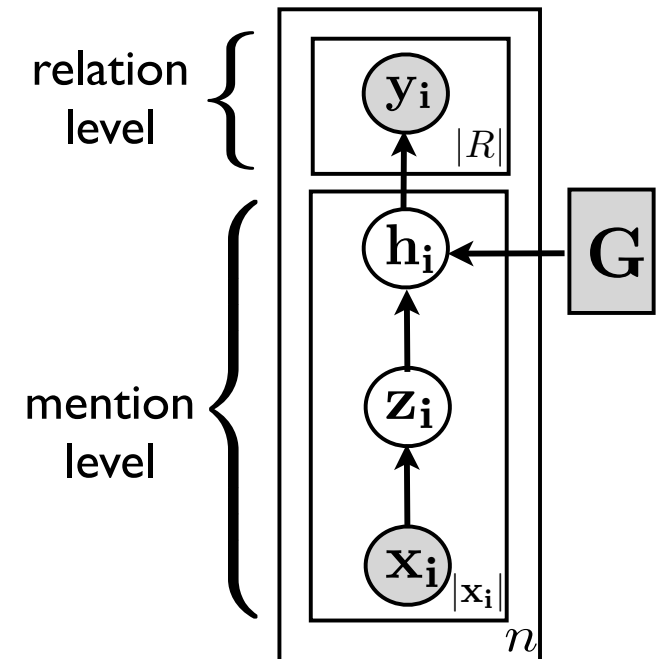
*... the forced resignation  
of the CEO of Boeing,  
Harry Stonecipher, for ...*

**Harry Stonecipher**



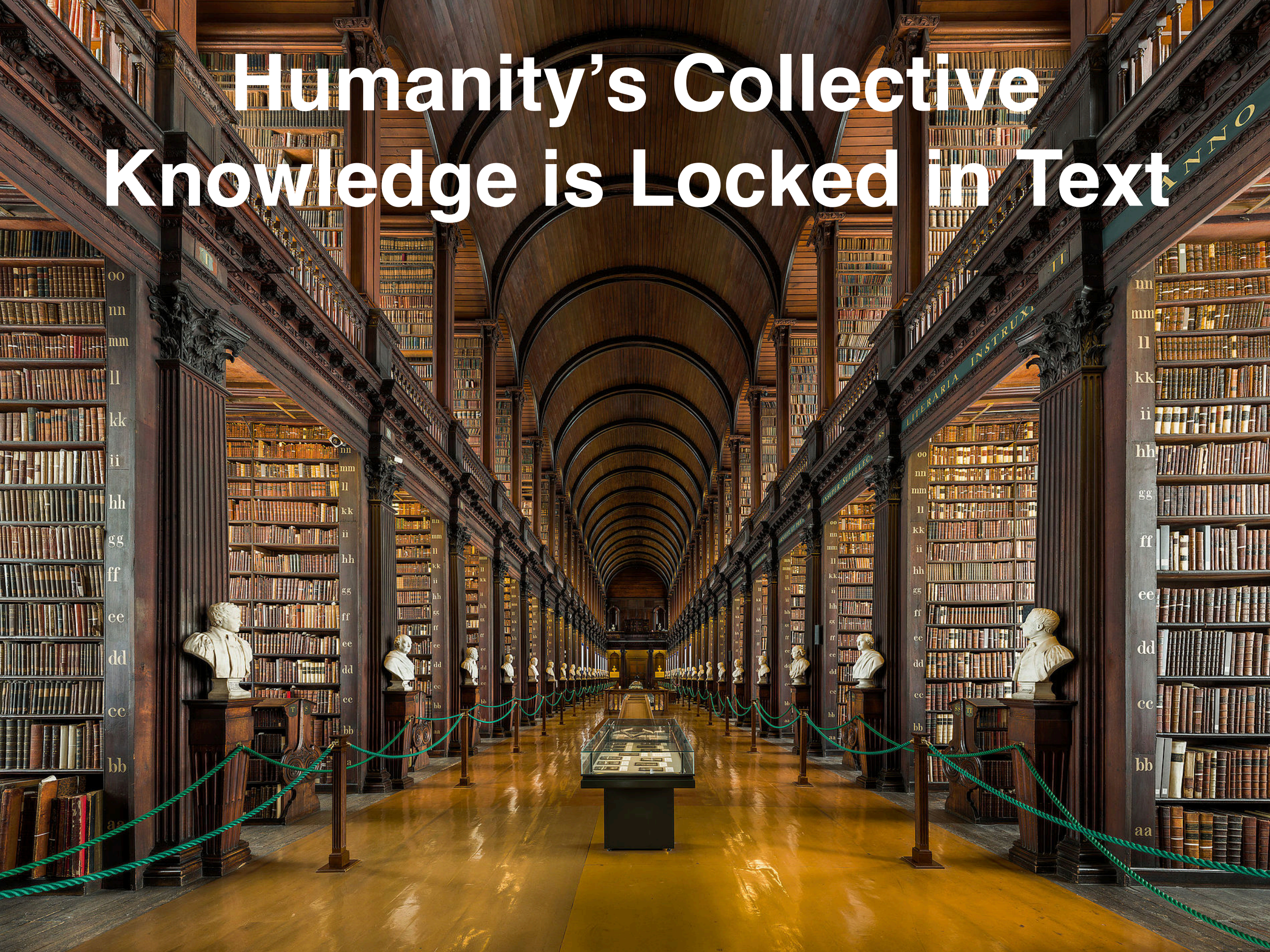
**CEO, Boeing**

**In office**  
2003-2005





# Humanity's Collective Knowledge is Locked in Text

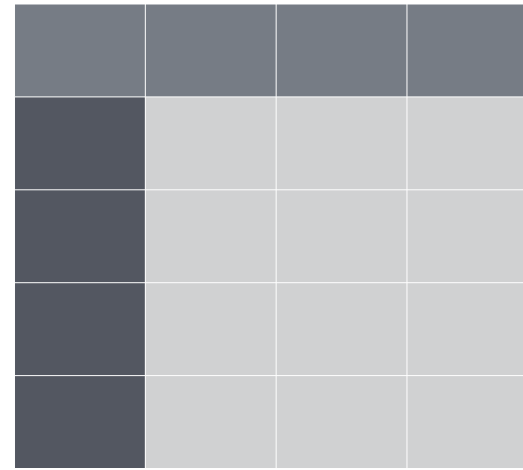




# Information Extraction



Text



Structured Data



# Information Extraction

*“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for \$250”*

# Information Extraction

*“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for \$250”*

# Information Extraction

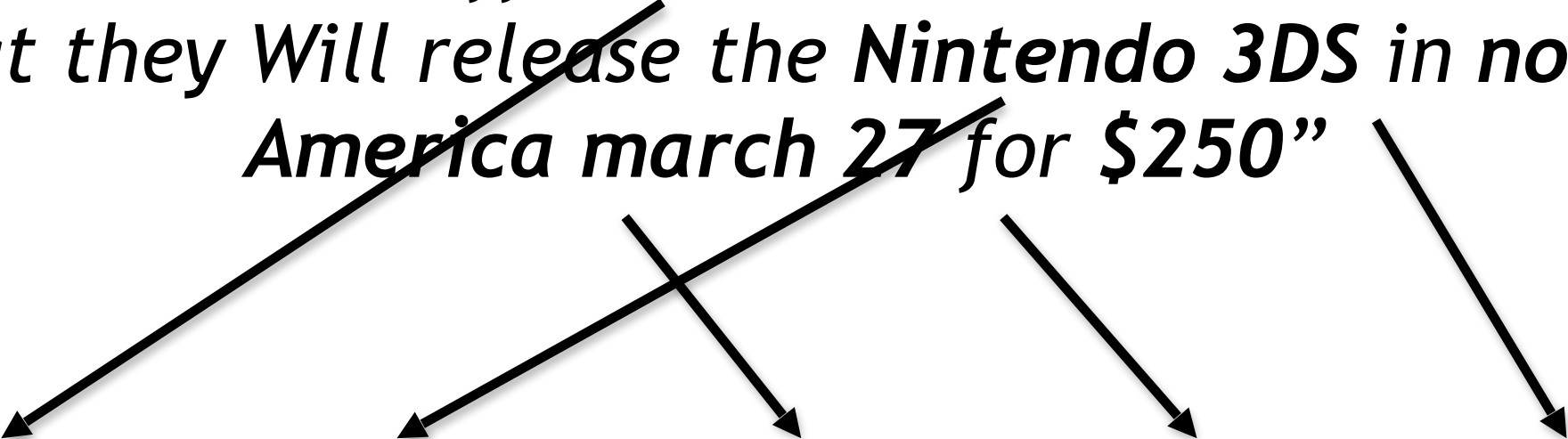
*“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for \$250”*

COMPANY	PRODUCT	DATE	PRICE	REGION

PRODUCT RELEASE

# Information Extraction

*“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for \$250”*



COMPANY	PRODUCT	DATE	PRICE	REGION
Nintendo	3DS	March 27	\$250	North America

**PRODUCT RELEASE**

# Information Extraction

## *Samsung Galaxy S5 Coming to All Major U.S.*

- State of the art is maybe 80%, for single easy fields: 90%+
- Redundancy helps a lot!
- Much of human knowledge is waiting to be harvested from the Web!

COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
Nintendo	3DS	March 27	\$250	North America

PRODUCT RELEASE

# Paraphrase

*cup*

**word**

*mug*

*the king's speech*

**phrase**

*His Majesty's address*

*... the forced resignation of  
the CEO of Boeing, Harry  
Stonecipher, for ...*

**sentence**

*... after Boeing Co. Chief  
Executive Harry Stonecipher  
was ousted from ...*

**Wuwei Lan, Wei Xu.** “Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering” COLING (2018)

**Wuwei Lan, Siyu Qiu, Hua He, Wei Xu.** “A Continuously Growing Dataset of Sentential Paraphrases” EMNLP (2017)

**Wei Xu, Alan Ritter,** Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)

**Wei Xu, Alan Ritter,** Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” In COLING (2012)

# Question Answering

Who is the CEO stepping down from Boeing?

**match**

```
graph TD; Q[Who is the CEO stepping down from Boeing?]; A1[... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...]; A2[... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...]; Q -- match --> A1; Q -- match --> A2;
```

*... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...*

*... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...*

# Watson leverages multiple algorithms to perform deeper analysis






## [Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

## [Supporting Evidence]

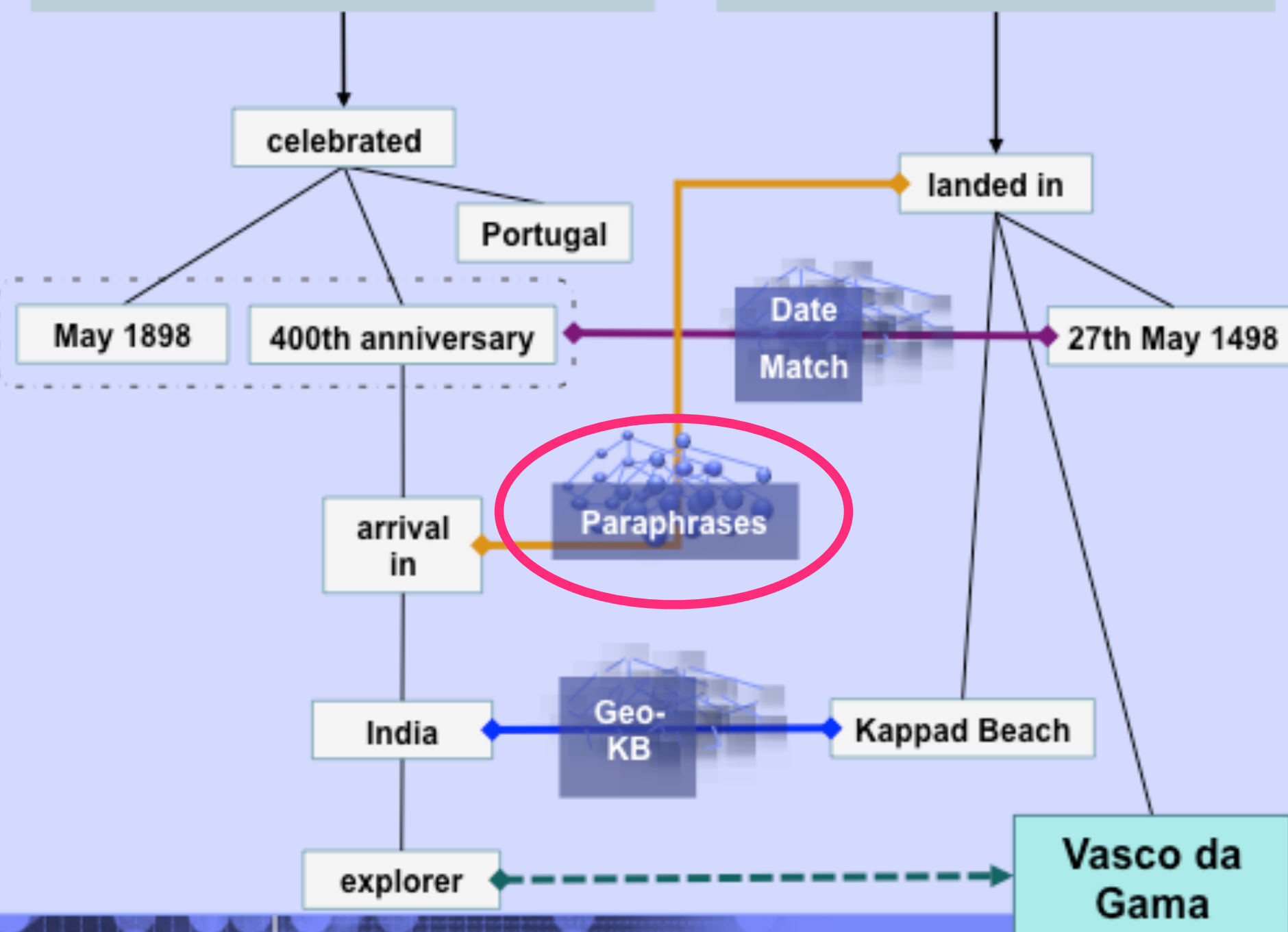
On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

### Legend

-  Temporal Reasoning
-  Statistical Paraphrasing
-  GeoSpatial Reasoning
-  Reference Text
-  Answer

*Stronger evidence can be much harder to find and score...*

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms





# Natural Language Generation

*who wants to get a beer?*



*want to get a beer?*

*who else wants to get a beer?*

*who wants to go get a beer?*

*who wants to buy a beer?*

*who else wants to get a beer?*

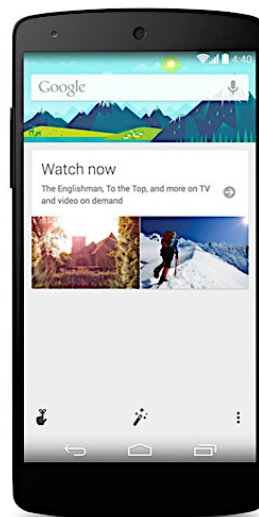
*trying to get a beer?*

*... (21 different ways)*

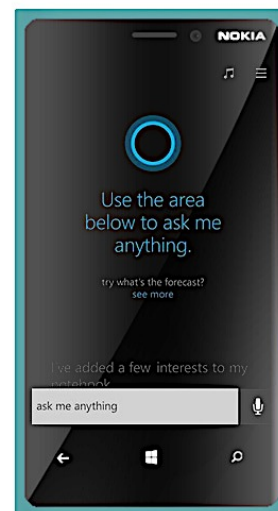
Apple Siri



Google Now



Windows Cortana



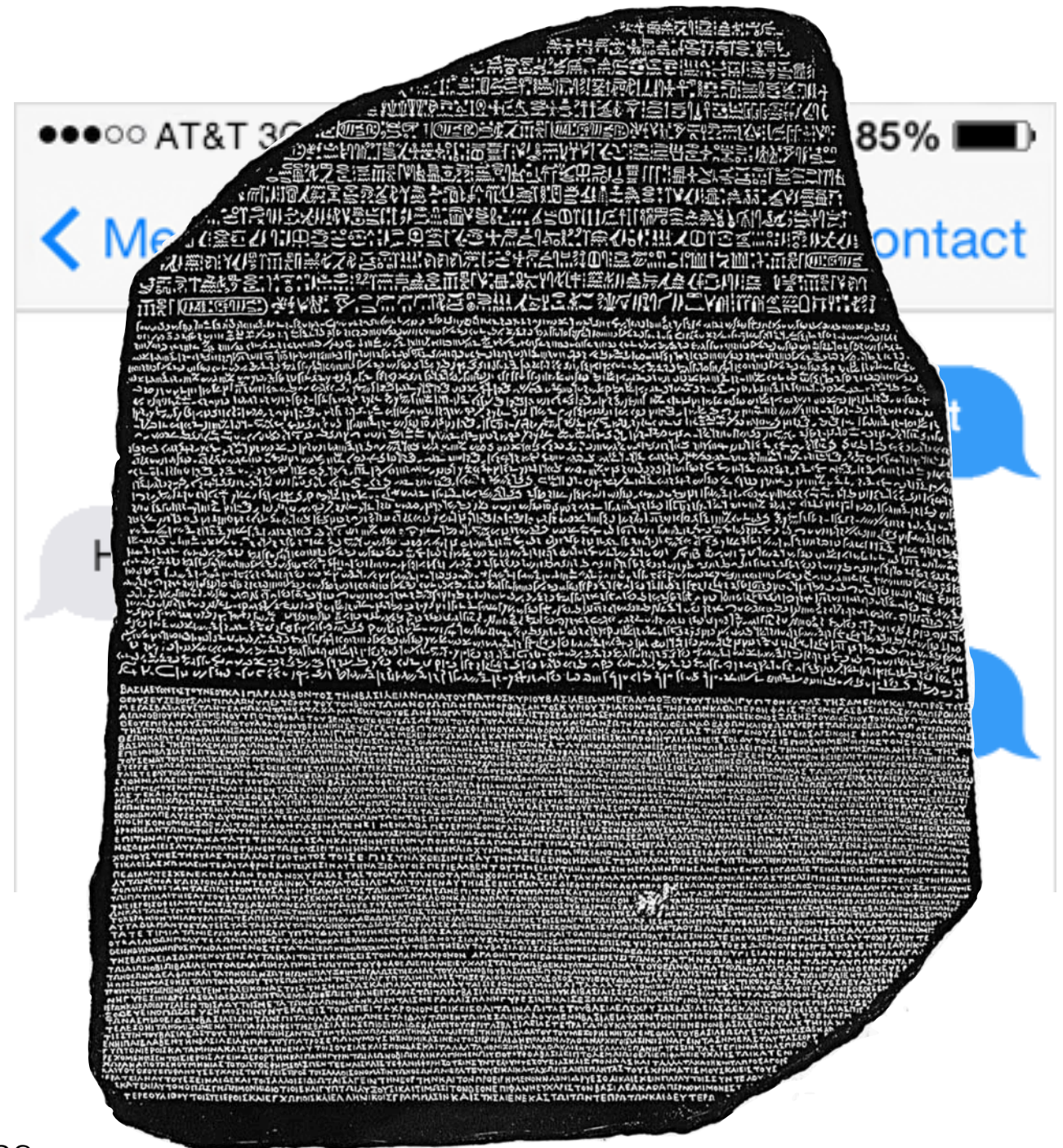
**Wei Xu**, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. “Optimizing Statistical Machine Translation for Simplification” in TACL (2016)

**Wei Xu**, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” in TACL (2015)

**Wei Xu**, **Alan Ritter**, Ralph Grishman. “Gathering and Generating Paraphrases from Twitter with Application to Normalization” In BUCC (2013)

# Data-Driven Conversation

- **Twitter:** ~ 500 Million Public SMS-Style Conversations *per Month*
- **Goal:** Learn conversational agents directly from massive volumes of data.



# Noisy Channel Model

Input:

**Who wants to come over for dinner tomorrow?**

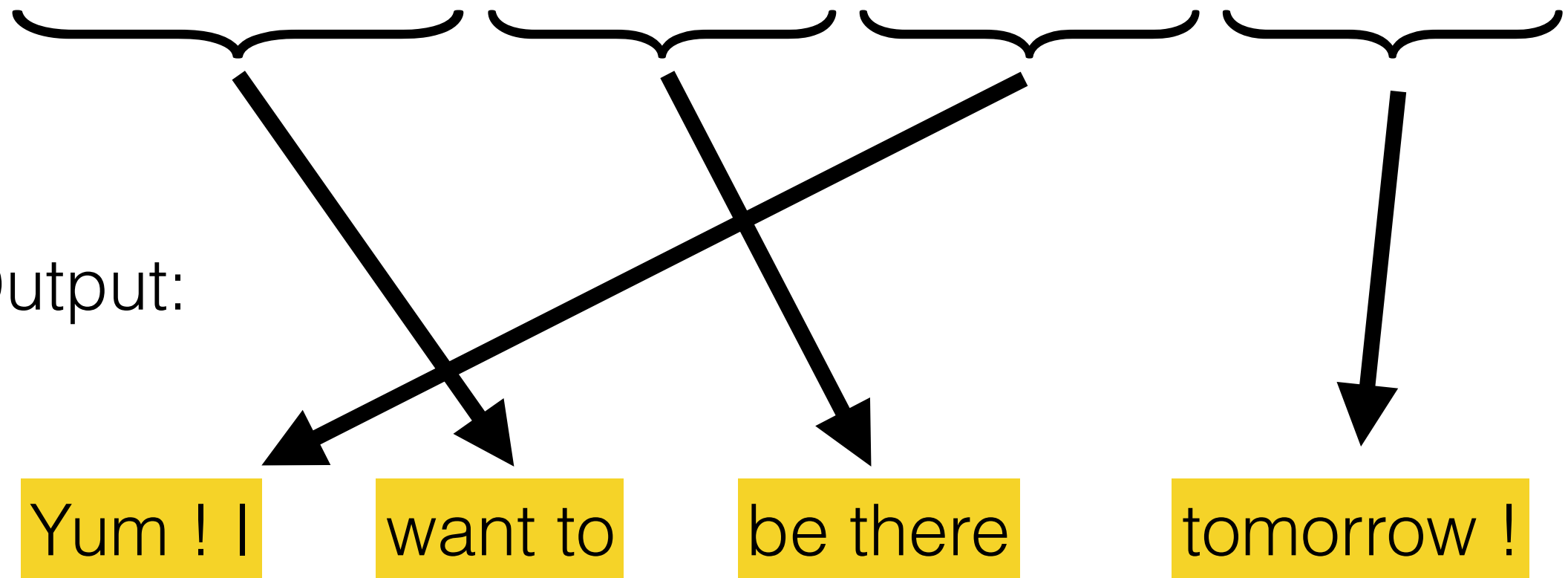
Output:

Yum ! I

want to

be there

tomorrow !





# Neural Conversation

[Sordoni et. al. 2015] [Xu et. al. 2016] [Wen et. al. 2016]  
[Li et. al. 2016] [Kannan et. al. 2016] [Serban et. al. 2016]

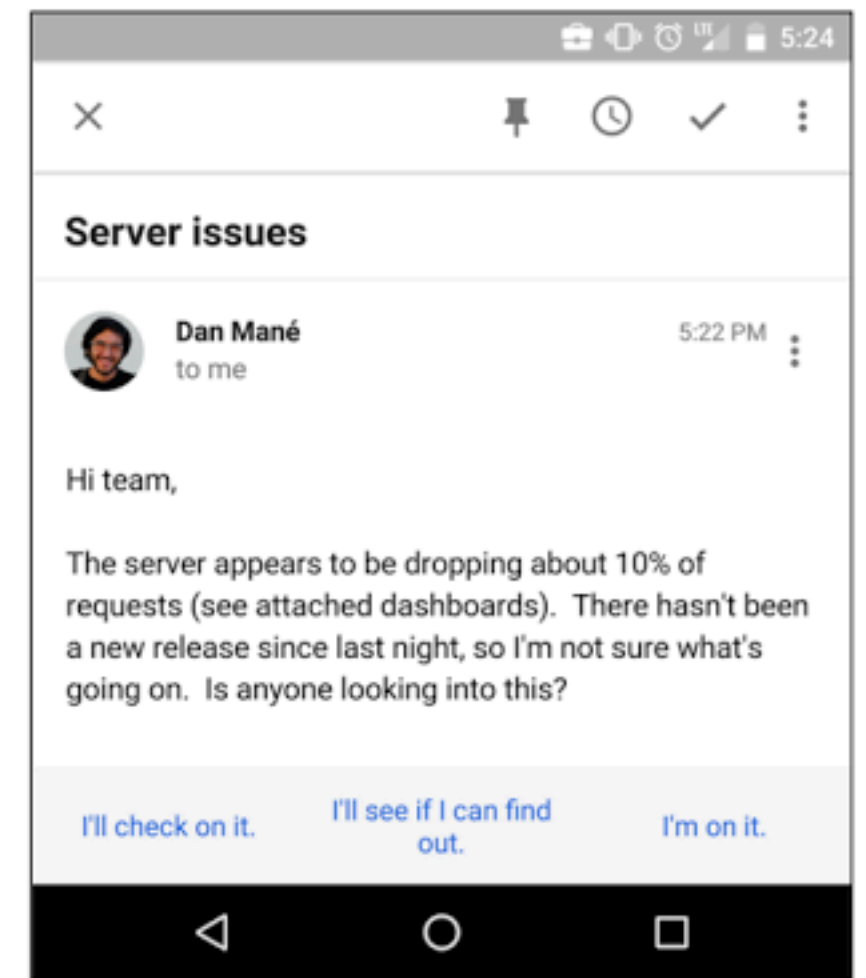


## Google Research Blog

Computer, respond to this email.

Tuesday, November 03, 2015

Posted by Greg Corrado\*, Senior Research Scientist



Another bizarre feature of our early prototype was its propensity to respond with “I love you” to seemingly anything. As adorable as this sounds, it wasn’t really what we were hoping for. Some analysis revealed that the system was doing exactly what we’d trained it to do, generate likely responses -- and it turns out that responses like “Thanks”, “Sounds good”, and “I love you” are super common -- so the system would lean on them as a safe bet if it was unsure. Normalizing the



# Language Technology

making good progress

mostly solved

## Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

## Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

## Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

## Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

## Coreference resolution

Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

## Parsing

I can see Alcatraz from the window!

## Machine translation (MT)

第13届上海国际电影节开幕...

The 13<sup>th</sup> Shanghai International Film Festival...

## Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

still really hard

## Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

## Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

## Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

## Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?



What will we cover in  
this class (and should  
you take it)?

# What do you expect to learn

- Twitter API for obtaining Twitter data
- cutting edge research on:
  - Natural Language Processing (NLP)
  - Machine Learning
- useful NLP tools, especially for Twitter text
- basic machine learning algorithms:
  - Naïve Bayes, Logistic Regression
  - Probabilistic Graphical Models
  - Some deep learning basics

# Guest Lectures

- At least one guest lecture from other NLP faculty members and/or industry, student researchers



# Grading

- two programming assignments (30% individual)
- A 3rd assignment/research project (**optional**, 20% bonus)
- in-class presentation (20% group of two)
- paper summaries (20% individual, about 10 papers)
- several take-home Quizzes (15% individual)
- participation in class discussions (15%)

# Grading

- two programming assignments (30% individual)
- in-class presentation (20% group of two)
- paper summaries (20% individual, about 10 papers)
- Grading on a 12-point scale — 10 for normal completion, 2 for going above and beyond. Final letter grade of the class will be graded on the curve.

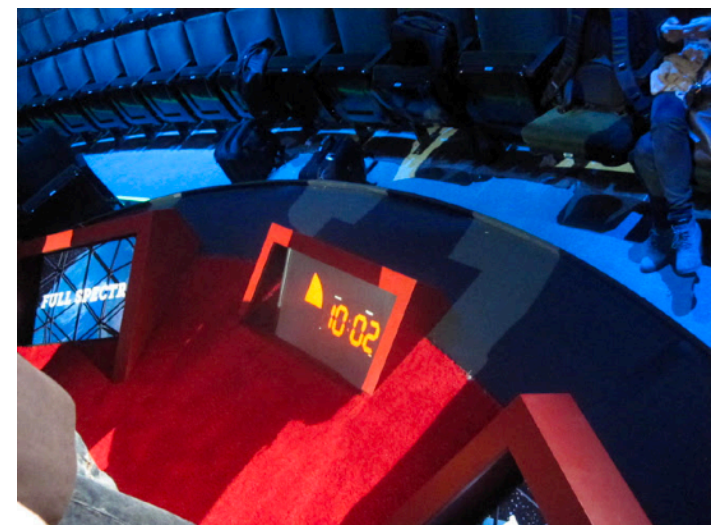
# Programming Assignments

- All in Python
- two programming assignments (30% — individual)
  1. Twitter's Language Mix (on the course website **now**)
  2. Logistic Regression Algorithm (use Numpy package)
- a third assignment (**optional** — group recommended)
  3. Deep Learning Basics and Word2Vec



# In-class Presentation

- a 12 minute presentation (20%)
  - A Social Media Platform
  - Or a research paper from NLP Researchers
  - Rehearse! We will use a timer as TED Talk



# In-class Presentation

Social Media & Text Analytics

Syllabus

Twitter API Tutorial

Homework ▾

High School Outreach



Social Media Map  
for 2016

## Survey a Social Media Platform, NLP Researcher or Dataset : In-class Presentation (20 points)

You will pair together (2 students) and give a 10-minute presentation (plus 2-minute Q&A) in class about a social media platform (an incomplete list [here](#)) or a paper from NLP researchers of your choice (an incomplete list of NLP groups [here](#)). You are also encouraged to find other NLP researchers that are not on this list through CS department homepages or top NLP conferences/journals (e.g. ACL, NAACL, TACL, EMNLP).

First, please [sign up](#) to pick a date you want to present, and pick a social media platform or a NLP researcher.

After your presentation in the class, **upload your slides** to [OSU's Carmen](#) system. Your slides will be also published on this course website.

### For NLP researchers, you may focus on

- Who: You are encouraged to consider NLP researchers who are current phd students and post-docs, as well as researchers in industrial labs. Summarize his/her career. How and why do they become successful?
- What: What research topics they are working on? What are they famous for? What does his/her first NLP paper look like? Present one of his/her important or recent work.

### For social media platforms, you may focus on:

- Market: When it was founded, purchased, and etc?
- Interface: How people use it, and why?
- Software Development: Any API available?
- Academic Research: Any interesting studies? Any useful datasets?
- and any other things you think are important

# In-class Presentation

5539 Presentations (2019AU)



# Quizzes

- several simple take-home quizzes
- hard-copy on paper
- will not be graded; but count for 10 points
- We have **Quiz #0 today** on class survey!

# Paper Summaries

- roughly one paper assigned for reading per week
- about 10 papers in total
- allowed to skip two papers throughout the semester
- write a short summary between 100-200 words:
  - discuss positive aspects and limitations
  - suggest potential improvement or extensions

# Paper Summaries

- Hal Daumé III's infamous NLP blog



**P16-1009: Rico Sennrich; Barry Haddow; Alexandra Birch**

*Improving Neural Machine Translation Models with Monolingual Data*

I like this paper because it has a nice solution to a problem I spent a year thinking about on-and-off and never came up with. The problem is: suppose that you're training a discriminative MT system (they're doing neural; that's essentially irrelevant). You usually have far more monolingual data than parallel data, which typically gets thrown away in neural systems because we have no idea how to incorporate it (other than as a feature, but that's blech). What they do here is, assuming you have translation systems in both directions, back translate your monolingual target-side data, and then use that faux-parallel-data to train your MT system on. Obvious question is: how much of the improvement in performance is due to language modeling versus due to some weird kind of reverse-self-training, but regardless the answer, this is a really cool (if somewhat computationally expensive) answer to a question that's been around for at least five years. Oh and it also works *really* well.



# Research Project

- **Optional**
- Build a machine translation system and **web demo** that can transfer contemporary English text into Shakespearean style!

# Stylistic Language Generation



Palpatine:

*If you will not be turned, you will be destroyed!*



*If you will not be turn'd, you will be undone!*

Luke:

*Father, please! Help me!*



*Father, I pray you! Help me!*





# Stylistic Language Generation

- Data and code:

<https://github.com/cocoxu/Shakespeare/>



# Stylistic Language Generation

- It has yet become a popular student research project:
  - Stanford students: <https://web.stanford.edu/class/cs224n/reports/2757511.pdf>
  - University of Maryland students: [http://xingniu.org/pub/styvar\\_emnlp17.pdf](http://xingniu.org/pub/styvar_emnlp17.pdf)
  - CMU students: <https://arxiv.org/abs/1707.01161>



# Language Styles



wonderfully delightfully beautifully fine well good nicely superbly



**she says**



**he says**

# What will you get out of this class?

- Understanding of an emerging field of CS
- Programming and machine learning skills useful in industry companies and academic research
- Getting a taste of research and being prepared

# Office Hour

- Have a question? Ask in/after class
- Or ask on Piazza discussion board
- Office Hour: TBA

# Piazza Discussion Broad

PIAZZA

CSE 5539 AU2017 (35985) ▾

Q & A

Resources

Statistics

Manage Class

Wei Xu

polls hw1 hw2 hw3 project exam logistics other

Unread Updated Unresolved Following

New Post Search or add a post...

PINNED

Private Search for Teammates! 8/21/17

FAVORITES

Instr How to Read a Technical... 9/19/17  
One of you asked a good question -- "how to read a paper?". In general, I

Instr Instructions for installing... 9/11/17  
This is a small write-up with steps for installing Jupyter and NumPy. Jupyter is

An instructor thinks this is a good note

WEEK 4/22 - 4/28

Ignore 4/25/18

WEEK 12/3 - 12/9

Word2Vec 12/4/17  
For the last homework, if two people are doing in a group, do both need to

WEEK 11/26 - 12/2

Instr no final exam 12/1/17  
Just a quick confirmation -- there will be no final examination for this class

Final cost value after gradient\_des... 11/29/17  
My cost value decreases over each

Note History:

This class has been made inactive. No posts will be allowed until an instructor reactivates the class.

note ★ stop following 18 views

How to Read a Technical Paper

One of you asked a good question -- "how to read a paper?".

In general, I think there is no single best way to read a paper -- it depends on. Many of you are writing very good and thoughtful reading notes in Carmen. We will discuss from time to time in the class, so hopefully you will learn from those discussions and from other people's thoughts.

That being said, Jason Esiner has written down some good advice on how to read a technical paper:  
<http://cs.jhu.edu/~jason/advice/how-to-read-a-paper.html>

As I mentioned in class earlier, you may find other useful advice on Quora, and just by Googling it.

logistics

edit · good note 0 Updated 1 year ago by Wei Xu

followup discussions for lingering questions and comments

Start a new followup discussion

Compose a new followup discussion

Average Response Ti... Special Mentions:

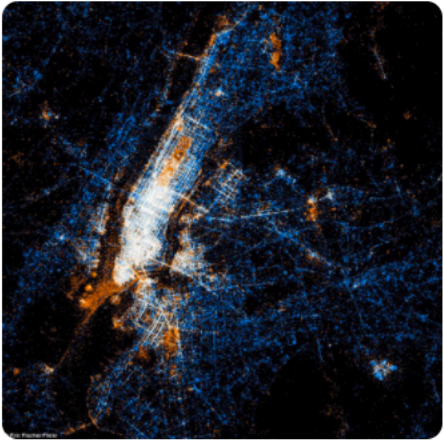
Online Now | This...



# By Next Class:

- Sign up for in-class presentation
- HW#1 Twitter's Language Mix

[Social Media & Text Analytics](#) [Syllabus](#) [Twitter API Tutorial](#) [Homework ▾](#)



*A visualization showing the location of Twitter messages (blue) and Flickr photos (orange) in New York City by Eric Fischer*

Social media provides a massive amount of data. This page provides an overview of prominent research findings and core natural language processing techniques.

**Instructor**  
[Wei Xu](#) is an assistant professor in the Department of Computer Science and Engineering at The Ohio State University. She is at the intersection of machine learning, natural language processing, and social media. She holds a PhD from the University of Pennsylvania. Before joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the ACL 2017 workshop on Social Media and NLP, serving as a workshop co-chair for [ACL 2017](#), an area chair for [EMNLP 2016](#) and the public chair for [ACL 2016](#).

**Time/Place** new  
**Fall 2017, CSE 5539-0010** The Ohio State University  
**Bolz Hall Room 318 | Tuesday 2:20PM – 4:10PM**

- 0. Become a Twitter User
- 1. Twitter's Language Mix
  - A. In-class Presentation
  - 2. Implement Logistic Regression
  - 3. Implement Word2vec (extracurricular)

[socialmedia-class.org](http://socialmedia-class.org)