Social Media & Text Analysis

lecture 1 - Introduction

Follow @cocoweixu

CSE 5539-0010 Ohio State University

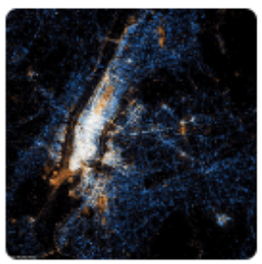
Instructor: Wei Xu

Website: socialmedia-class.org

Course Website

http://socialmedia-class.org/

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 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

Wei Xu is an assistant professor in the Department of Computer Science and Engineering at the Ohio State University. Her research interests lie at the intersection of machine learning, natural language processing, and social media. She holds a PhD in Computer Science from New York University. Prior to joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the ACL/COLING Workshop on Noisy User-generated Text, serving as a workshop co-chair for ACL 2017, an area chair for EMNLP 2016 and the publicity chair for NAACL 2016.

Time/Place new

Fall 2016, CSE 5539-0010 The Ohio State University
Cockins Hall Room 218 | Wednesday 2:20PM – 4:10PM
dual-listed undergraduate and graduate course

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. 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

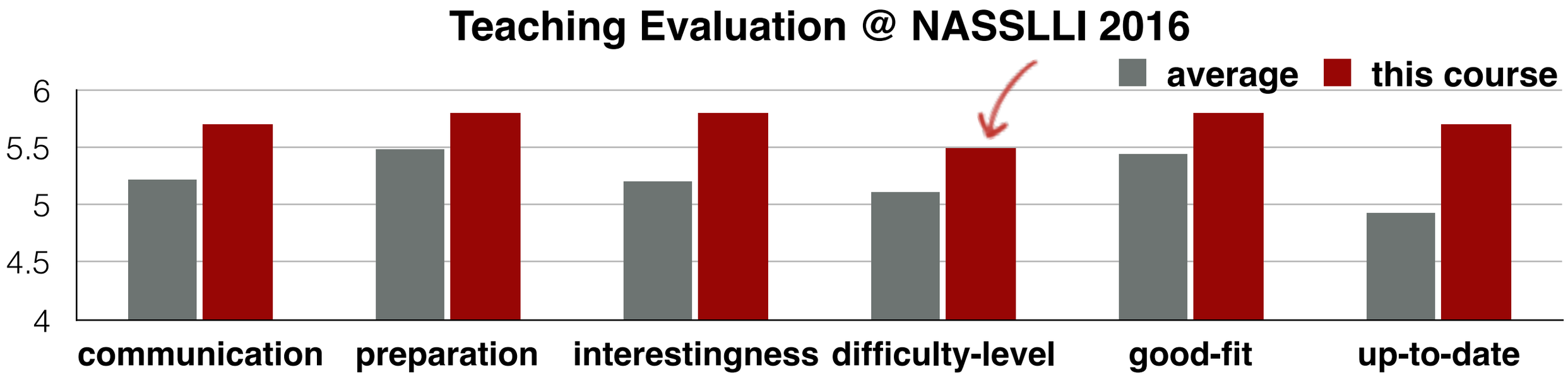
Various academic papers

Previous Offerings

Summer 2016, The North American Summer School on Logic, Language, and Information (NASSLLI)
Teaching evaluation was 5.72 out of 6 at NASSLLI; average across all instructors was 5.23.
Summer 2015, University of Pennsylvania (where this course was first designed and taught)

History of the Course

- Summer 2015, University of Pennsylvania
- Summer 2016, North American Summer School on Logic, Language, and Information (NASSLLI)
- Now, Ohio State University — continue developing



This is a special topic class

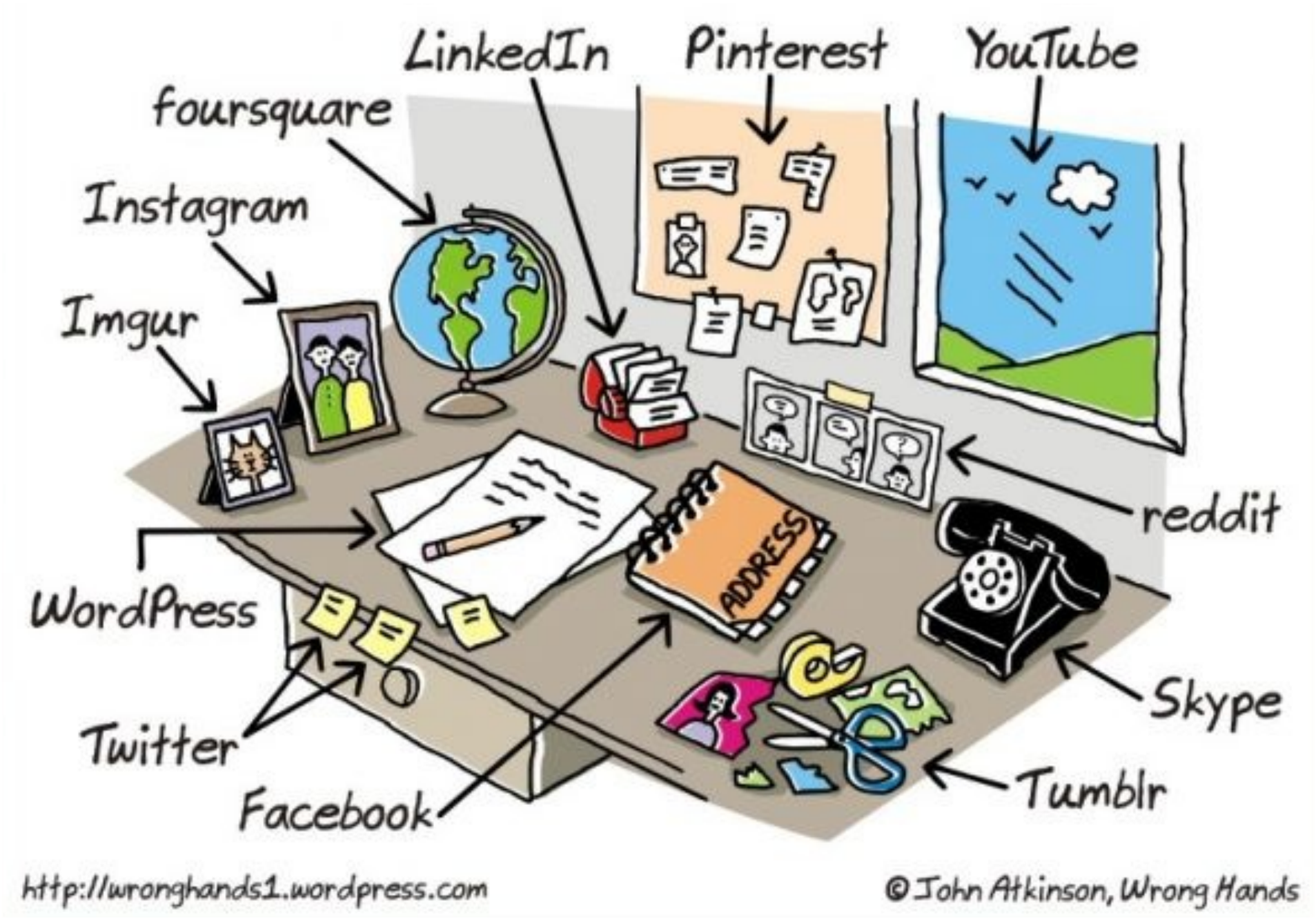
- hobby (not a mandatory course)
- but is lecture-based and project-based
- advanced and research-oriented
- but strong undergraduate students are also encouraged to take this course

Who am I?

Wei Xu

- Assistant Professor in CSE at the Ohio State University
- Postdoctoral researcher at University of Pennsylvania
- PhD from New York University in Computer Science
- Research Interests:
  - Natural Language Processing
  - Social Media
  - Machine Learning

Vintage Social Media



Broader Point of View

Source: http://www.conversationprism.com/

Why Social Media?

skip
@han_horan

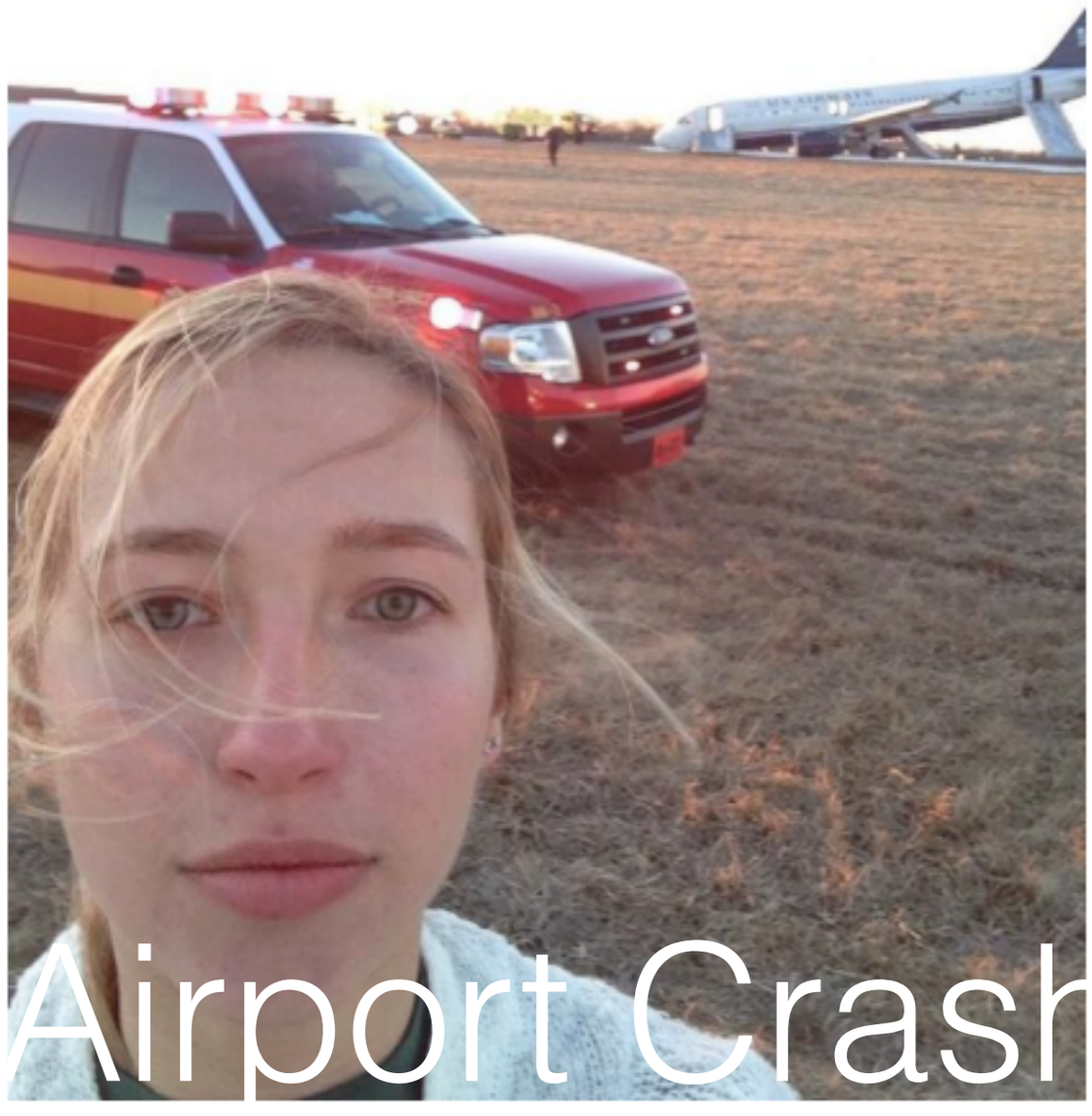
so my plane just crashed...
pic.twitter.com/X51BLwa5PS

Reply Retweet Favorite ••• More

skip
@han_horan

so yup pic.twitter.com/2WuLUWzpND

Reply Retweet Favorite ••• More



2014 Philly Airport Crash

Impact

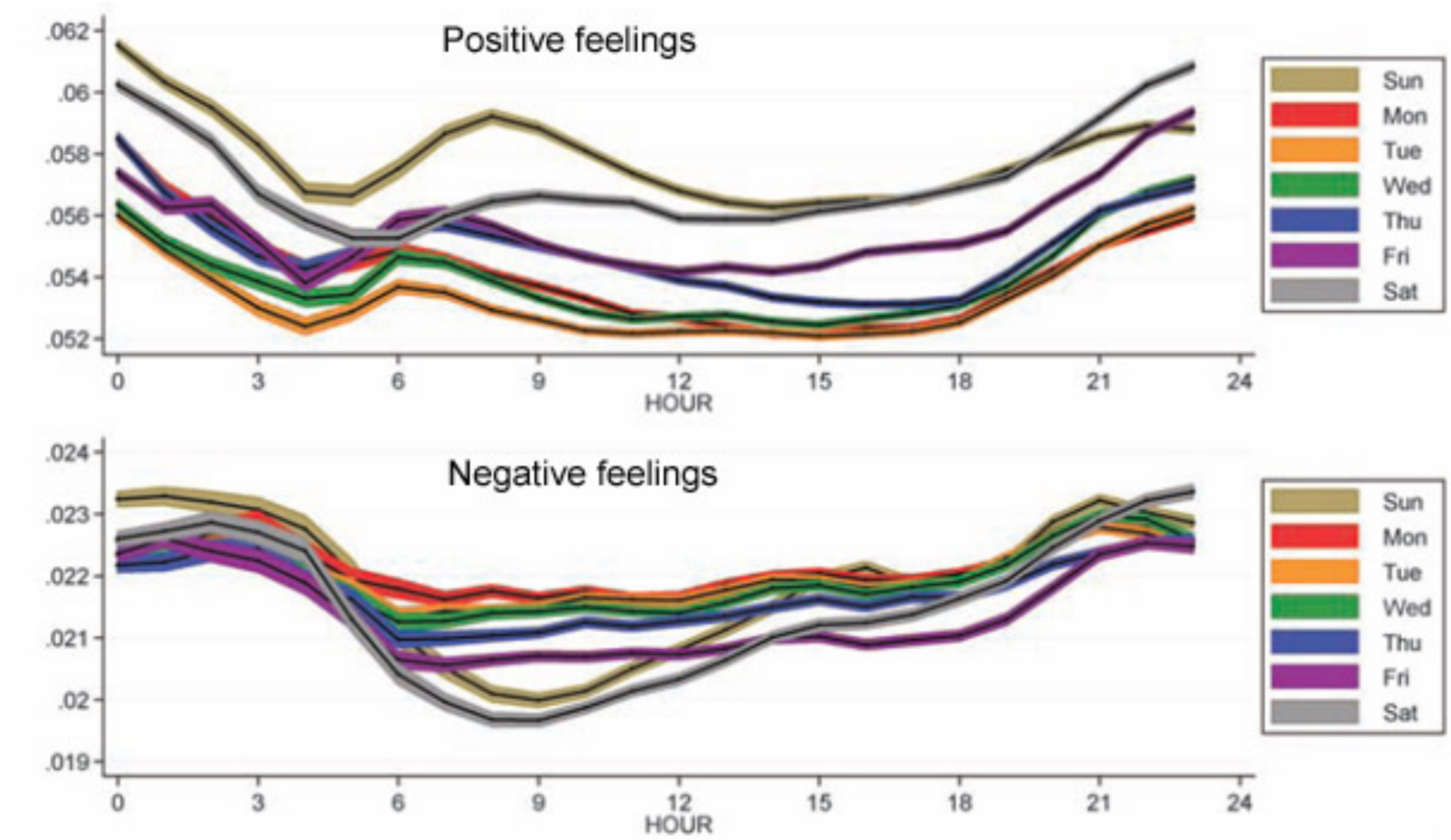
- Politics
- Business
- Socialization
- Journalism
- Cyber Bullying
- Productivity
- Privacy
- Emotions
- …
- and our language (!)

2014 Ukrainian Revolution

Research Value

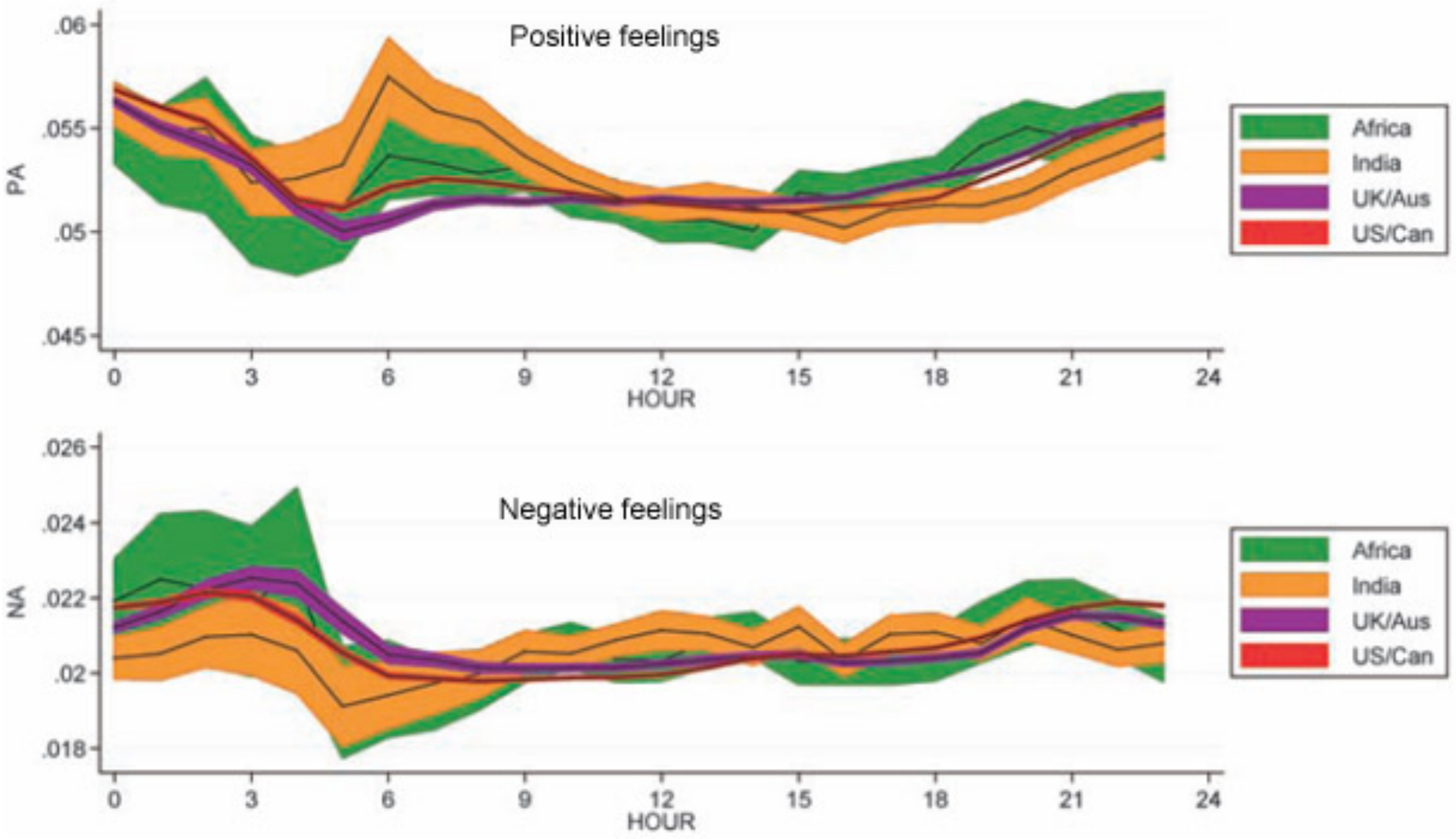
- In contrast to survey/self-report
- A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- Easy to access and aggregate **a lot** of data
- thus **a lot** of information

Mood



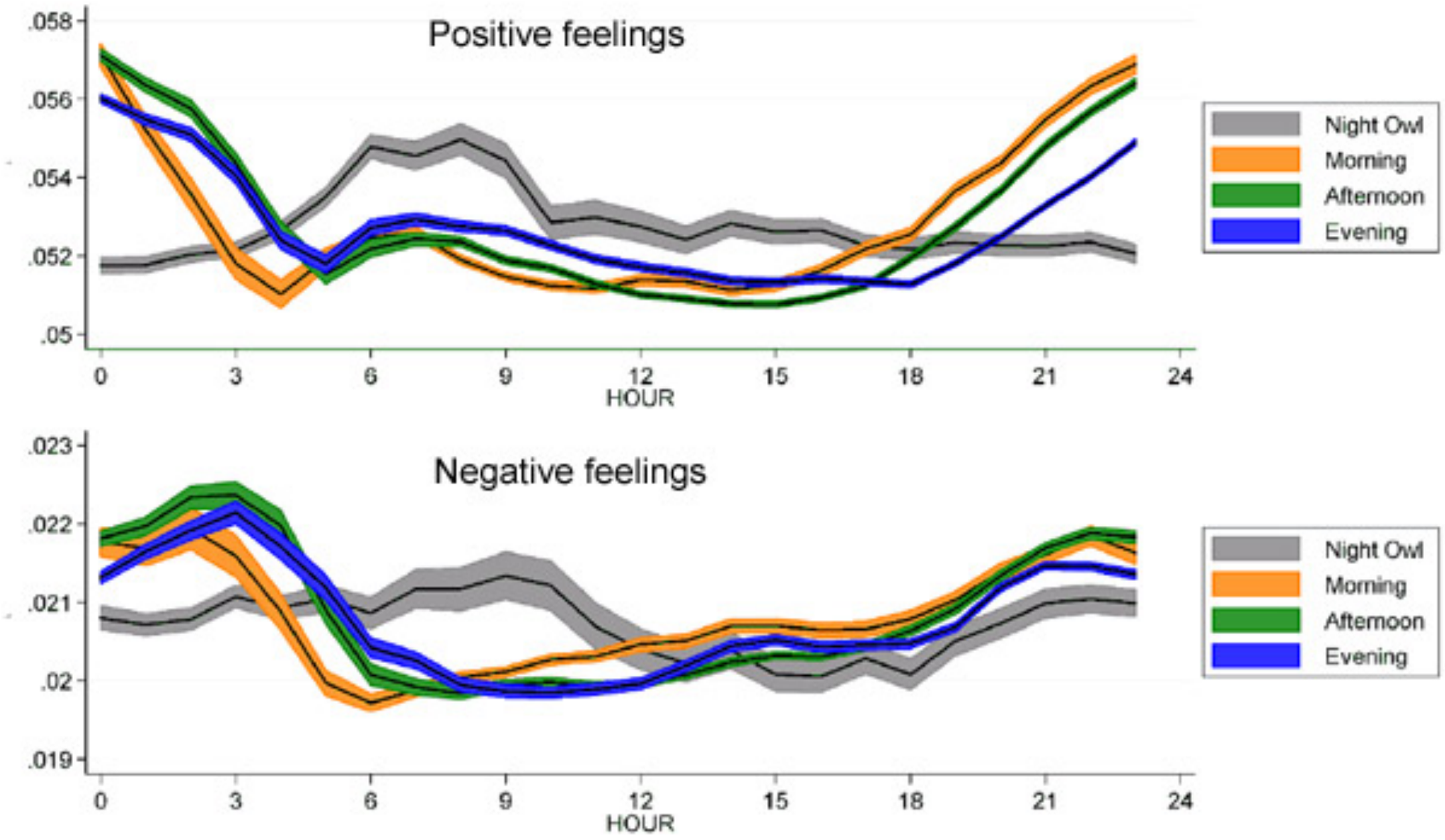
Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

Mood



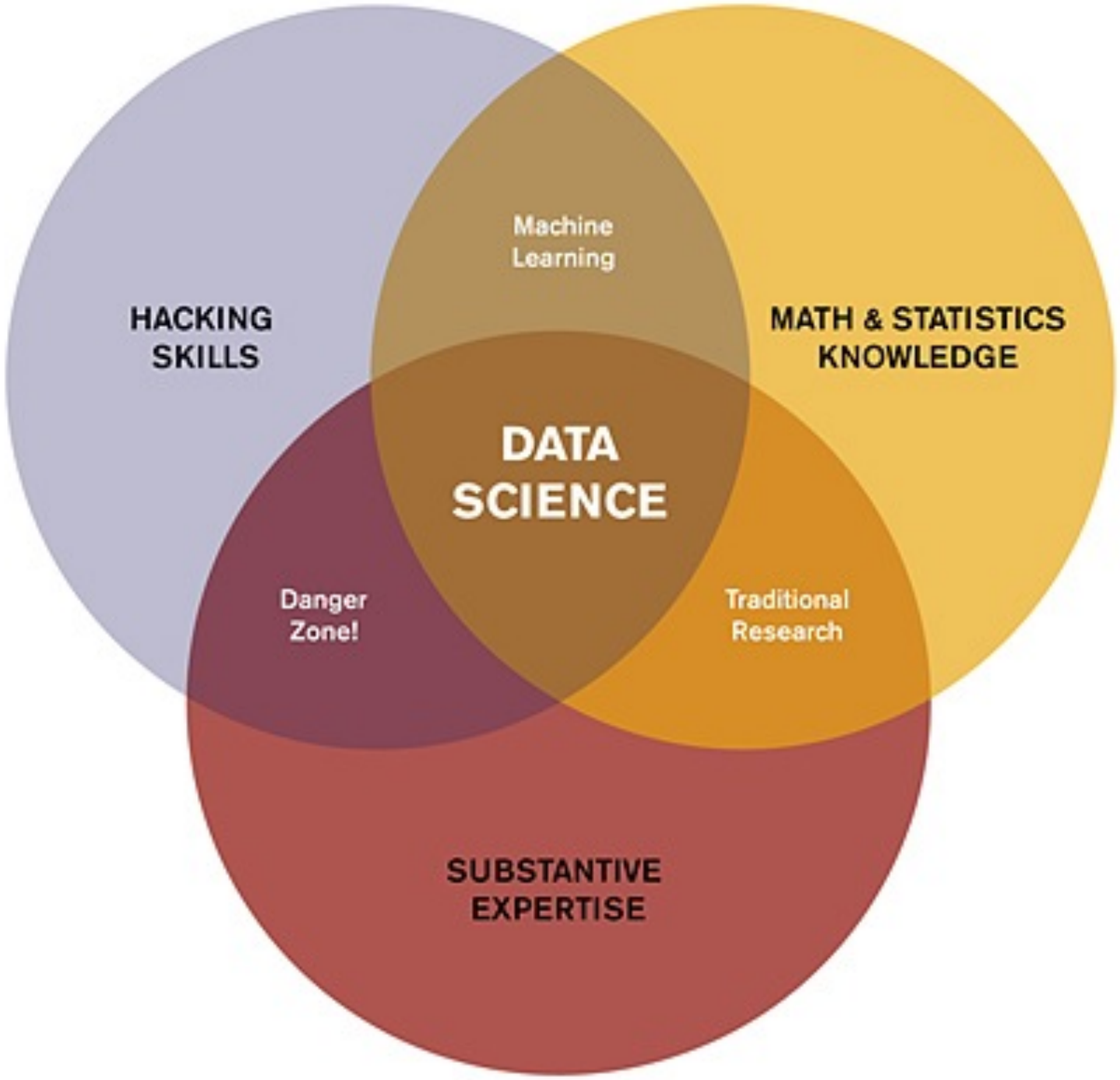
Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

Mood



Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

Data Science



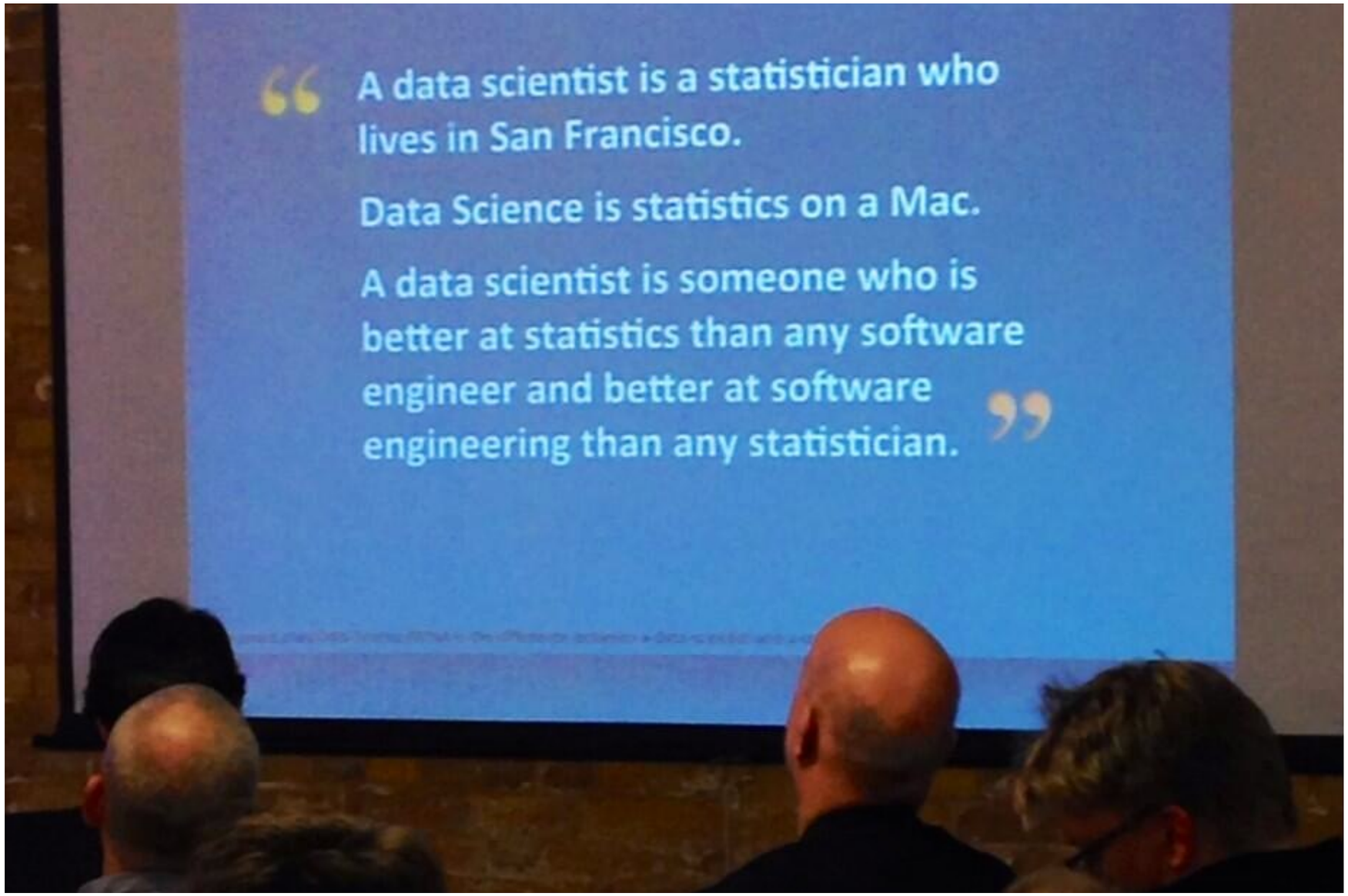
Source: Drew Conway

Data Science

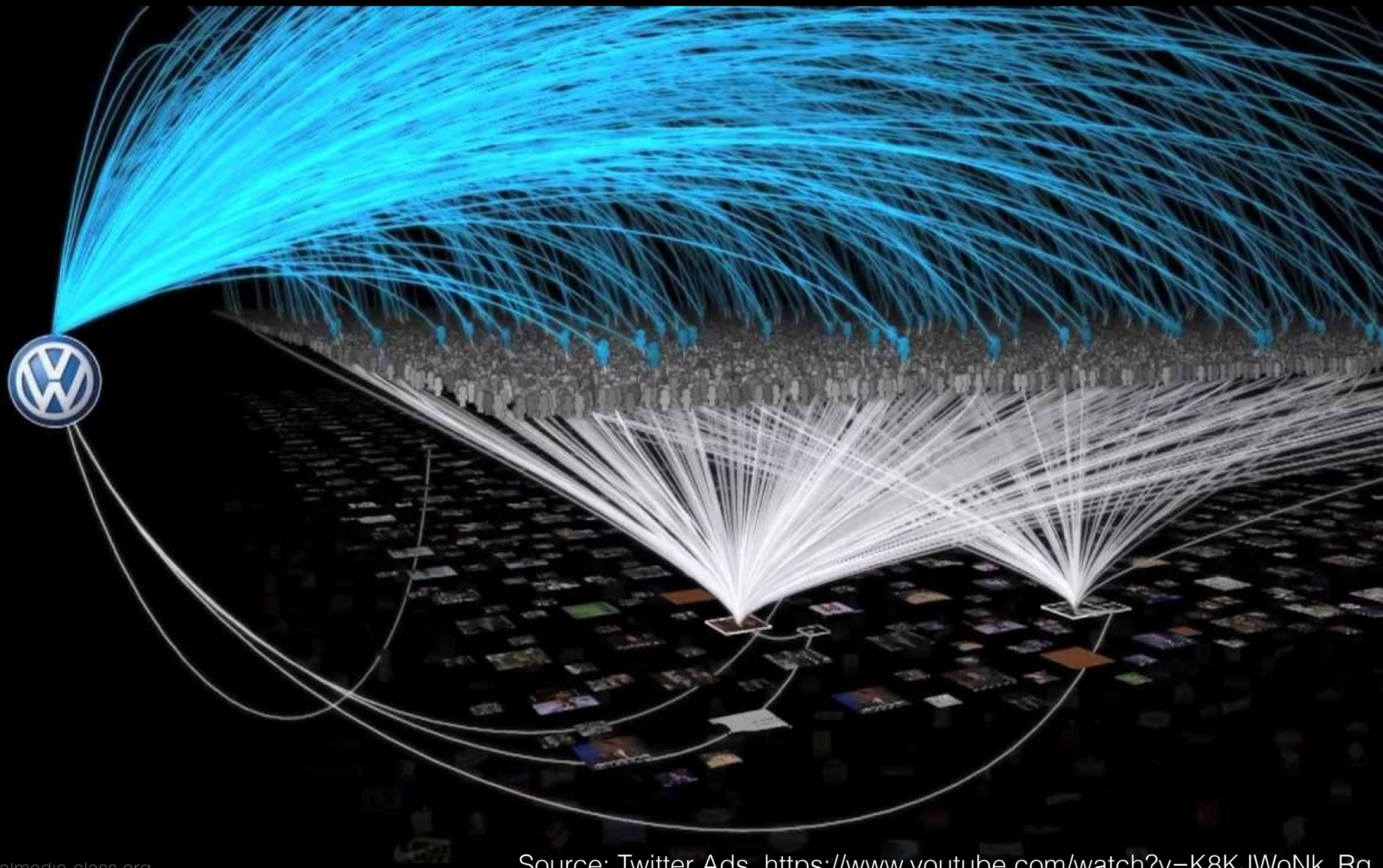
- is the **practice** of:
  - asking question (formulating hypothesis)
  - finding and collecting the data needed (often big data)
  - performing statistical and/or predictive analytics (often machine learning)
  - discovering important information and/or insights

Data Science

- the infamous definition:

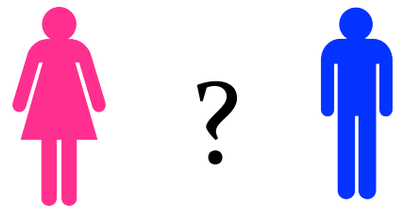


Marketing



Source: Twitter Ads https://www.youtube.com/watch?v=K8KJWoNk_Rg

User Profiling



Source: Volkova, Van Durme, Yarowsky, Bachrach
"Tutorial on Social Media Predictive Analytics" NAACL 2015

User Profiling

Follow

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

Follow

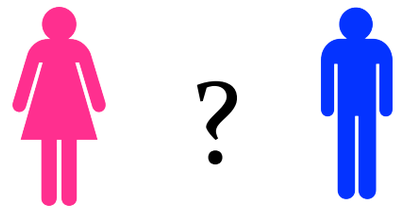
Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots

Source: Volkova, Van Durme, Yarowsky, Bachrach
"Tutorial on Social Media Predictive Analytics" NAACL 2015

User Profiling

Follow

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

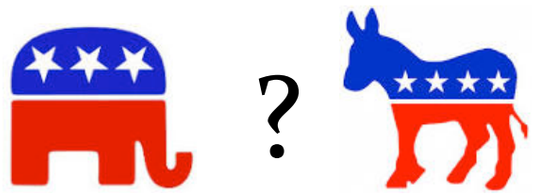


Follow

Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots

Follow

We've already tripled wind energy in America, but there's more we can do.

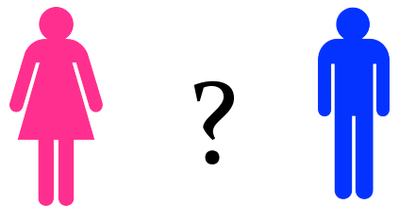


Source: Volkova, Van Durme, Yarowsky, Bachrach
"Tutorial on Social Media Predictive Analytics" NAACL 2015

User Profiling

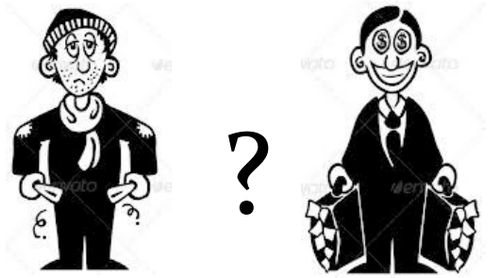
Follow

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



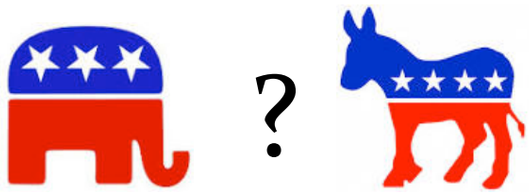
Follow

Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots



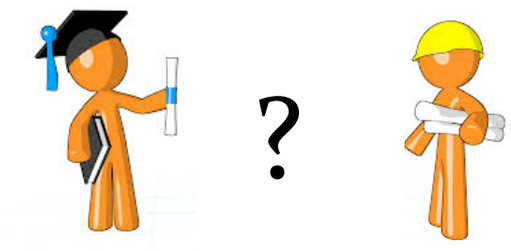
Follow

We've already tripled wind energy in America, but there's more we can do.



Follow

Two giant planets may cruise unseen beyond Pluto - space - June 2014 - New Scientist: newscientist.com/article/dn2571



Source: Volkova, Van Durme, Yarowsky, Bachrach "Tutorial on Social Media Predictive Analytics" NAACL 2015

Language Styles

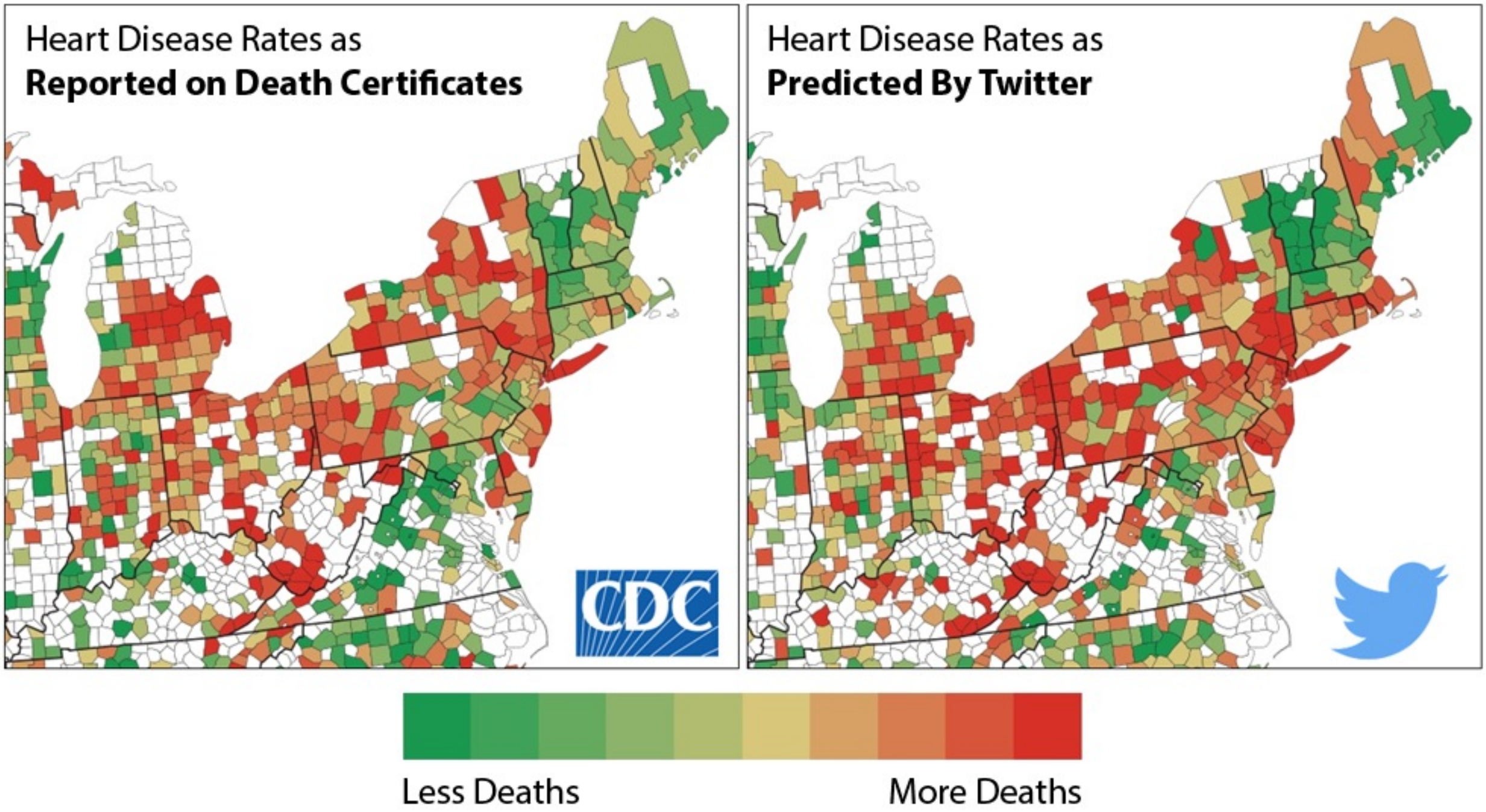
wonderfully delightfully beautifully fine well good nicely superbly

she says

he says

Source: Daniel Preoţiuc-Pietro, **Wei Xu** and Lyle Ungar
"Discovering User Attribute Stylistic Differences via Paraphrasing" AAAI 2016

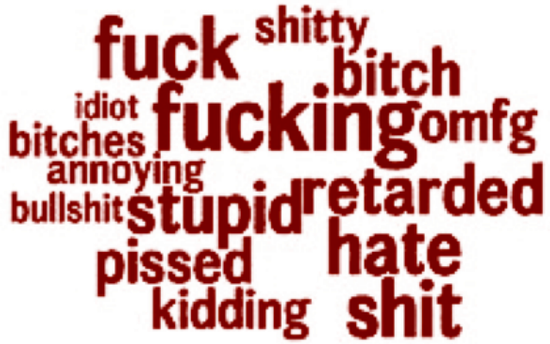
Health



Source: World Well-Being Project @ University of Penn

Health

Hostility, Aggression



$r = .27$

Skilled Occupations

$r = -.17$

Hate, Interpersonal Tension

$r = .21$

Positive Experiences

$r = -.15$

Boredom, Fatigue

$r = .20$

Optimism

$r = -.13$

Source: World Well-Being Project @ UPenn

What is Natural Language Processing?

Sentiment Analysis

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

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

***Wowsers** to this nets bulls game*

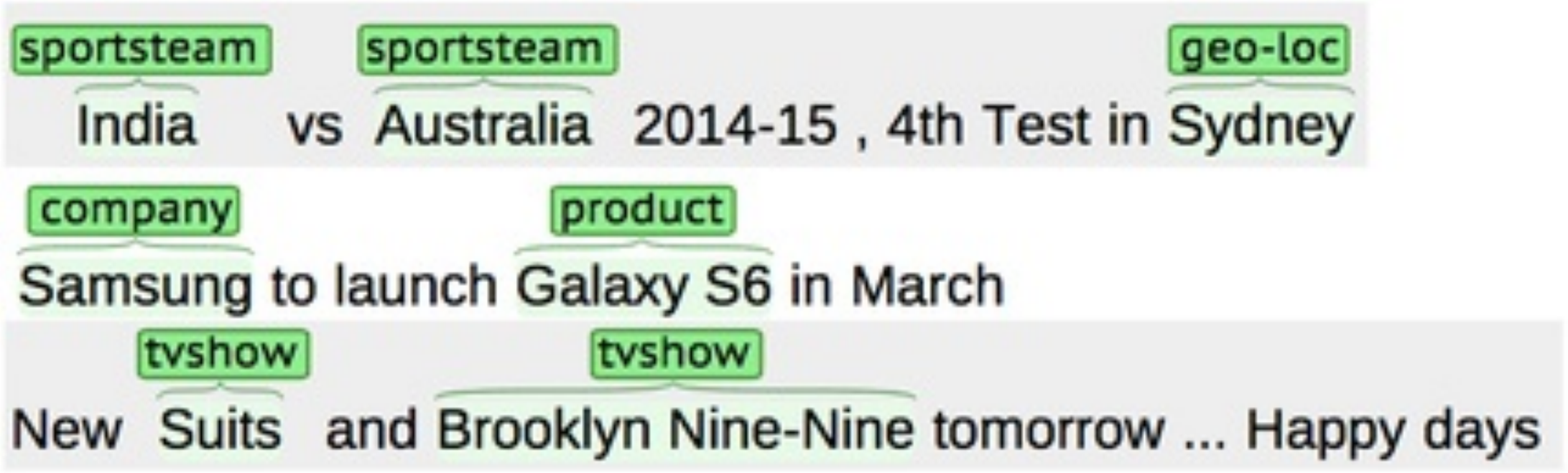
*this Nets vs Bulls game is **too live***

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

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

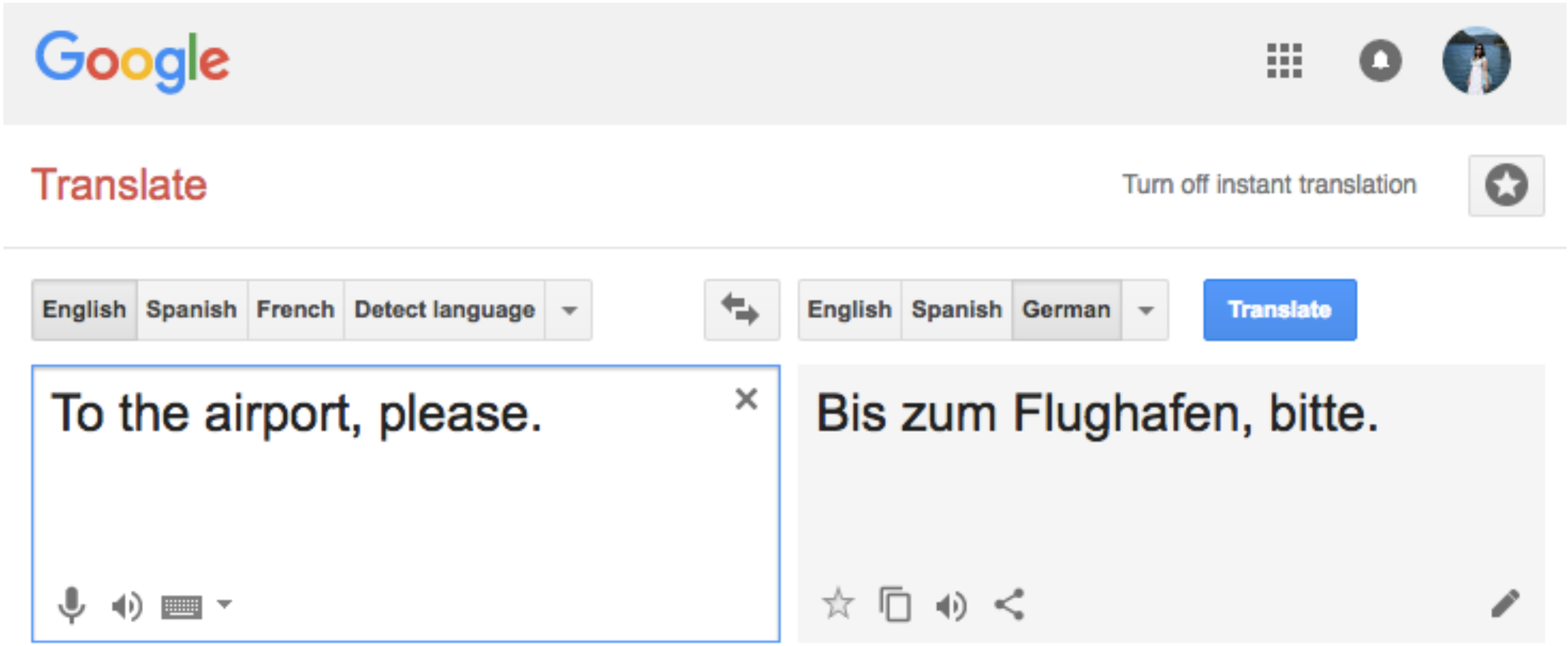
*This NetsBulls series is **intense***

Named Entity Recognition



Tim Baldwin, **Marie-Catherine de Marneffe**, Bo Han, Young-Bum Kim, **Ritter**, **Wei Xu**
Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition

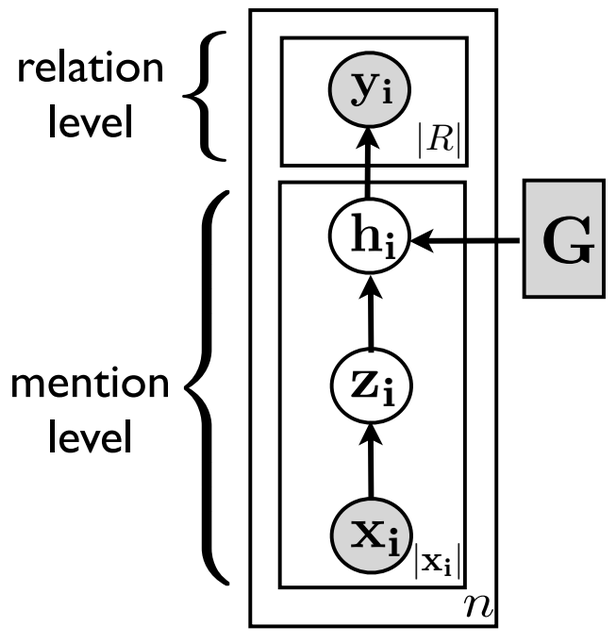
Machine Translation



Mingkun Gao, **Wei Xu**, Chris Callison-Burch. “Cost Optimization for Crowdsourcing Translation” In TACL (2014)

Information Extraction

… the forced <u>resignation</u> of the CEO of Boeing, Harry Stonecipher, for …



Maria Pershina, Bonan Min, **Wei Xu**, Ralph Grishman. "Infusion of Labeled Data into Distant Supervision for Relation Extraction" In ACL (2014)

Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman. "Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction" In ACL (2013)

Wei Xu, Alan Ritter, Ralph Grishman. "A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2013)

Wei Xu, Ralph Grishman, Le Zhao. "Passage Retrieval for Information Extraction using Distant Supervision" In IJCNLP (2011)

Question Answering

Who is the CEO stepping down from Boeing?

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

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

Question Answering

Who is the CEO stepping down from Boeing?

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

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

Question Answering

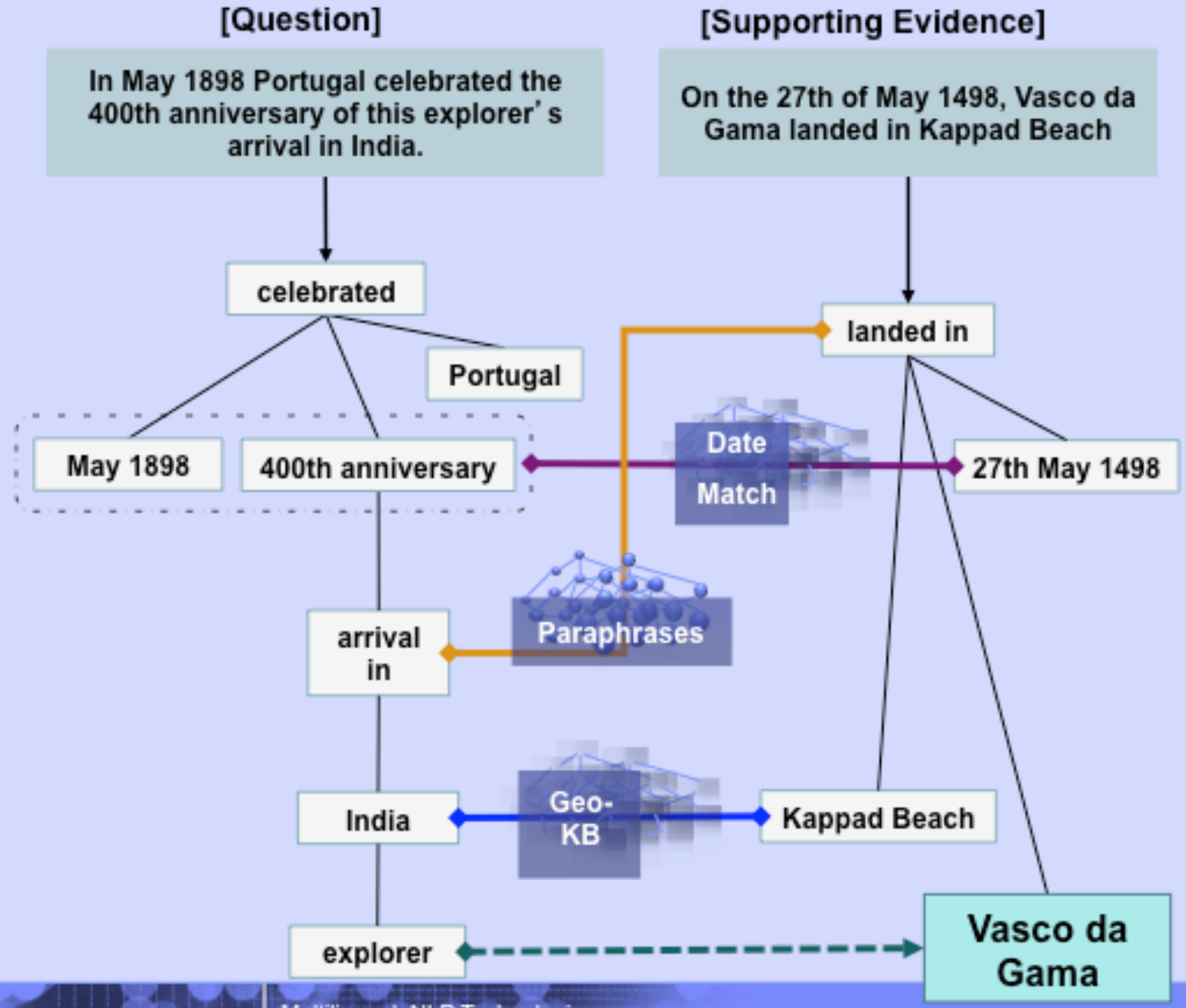
Who is the CEO stepping down from Boeing?

match

… 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

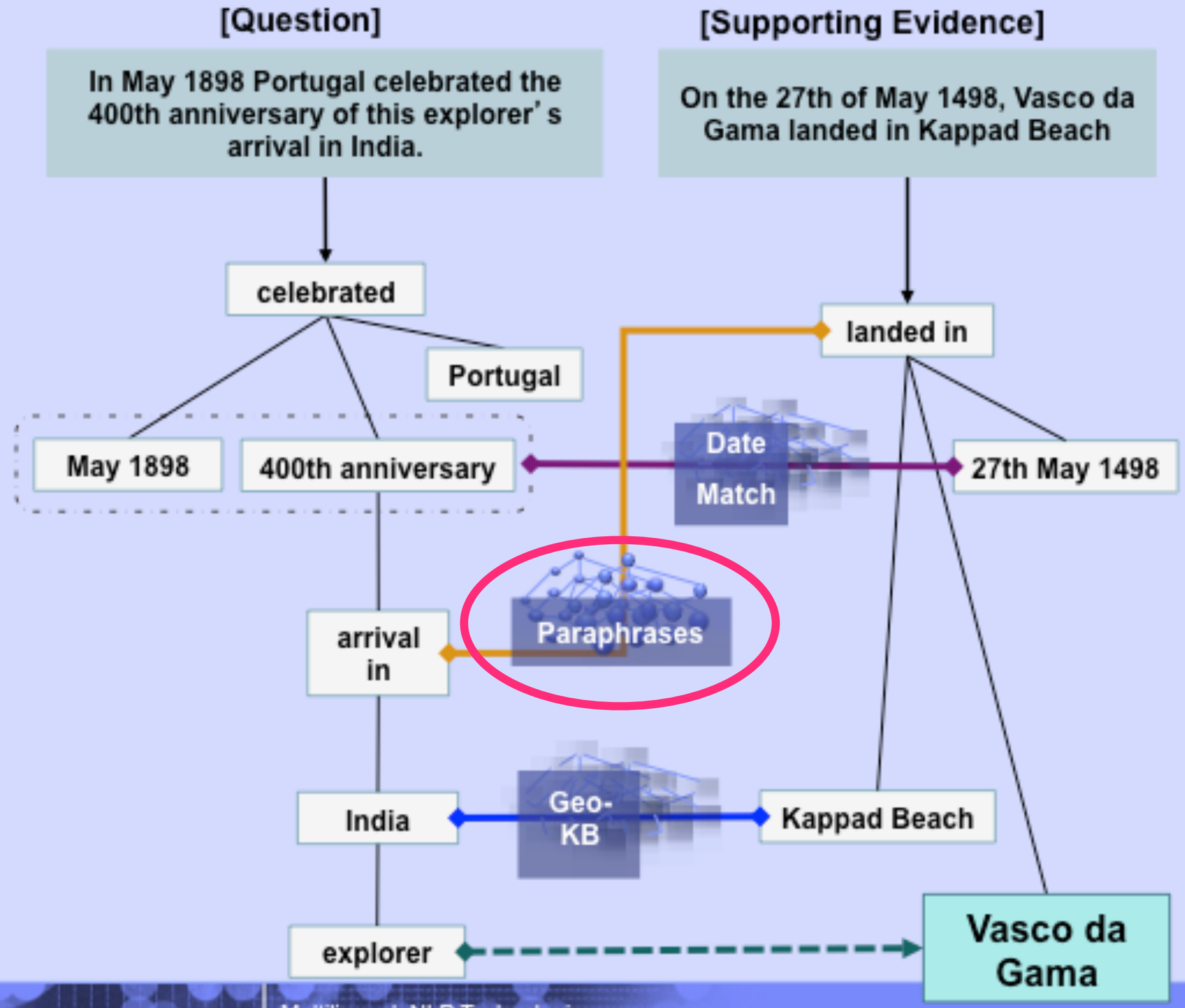


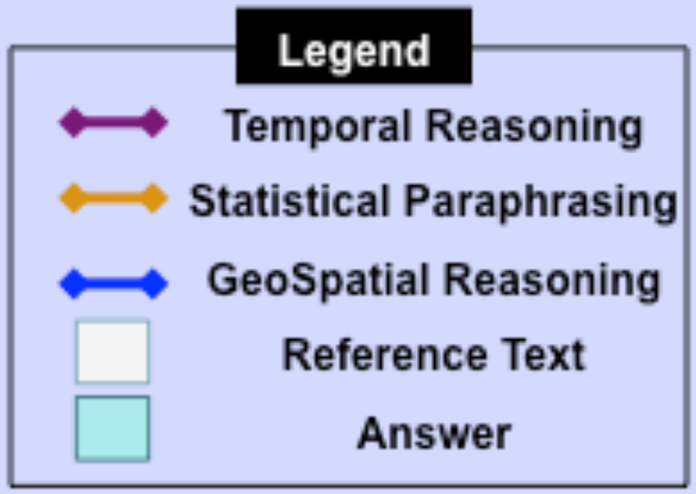
Stronger evidence can be much harder to find and score…

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

(courtesy: Salim Roukos)

Watson leverages multiple algorithms to perform deeper analysis





Stronger evidence can be much harder to find and score…

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

(courtesy: Salim Roukos)

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 …*

Wei Xu, Chris Callison-Burch, Bill Dolan. "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter" In SemEval (2015)
Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis. (2014)
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" In BUCC (2013)
Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. "Paraphrasing for Style" In COLING (2012)

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?

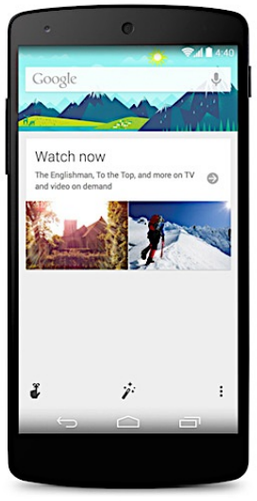
who else wants to get a beer?

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)

Dan Jurafsky

Language Technology

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

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

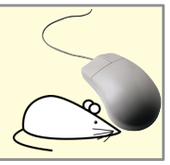
The waiter ignored us for 20 minutes.

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 13th 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

Guest Lectures

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

Grading

- two programing assignments (50 points)
- a 3rd assignment or a research project (30 points)
- in-class presentation (20 points)
- several simple Quiz/Survey (bonus 10 points)

Programming Assignments

- All in Python
- two programing assignments (50 points — individual)

  1. Twitter's Language Mix (on the course website **now**)

  2. Logistic Regression and Paraphrase Identification (tentative)

option 1

- a third assignment (25 points — group of 2/3)

  3. Basic Deep Learning and Word2Vec (tentative)

option 2

or a Research Project

- Group of 2/3
- open-end research problems
- aim for an academic publication

In-class Presentation

- a 10-15 minute presentation (20 points)
  - A Social Media Platform
  - or a NLP Research Group
  - or show off your research project (tentative)

Quiz/Survey

- several, very simple (much simpler than Quiz #1)
- hard-copy on paper
- will not be graded
- but give you 10 bonus points
- We have Quiz #1 today on **pre-requirements!**

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
- Summer internships?

TA and Office Hour

- Have a question? Ask at/after each class
- Office hour — TBA
- No TA and grader for this special topic class

Next Class:

- Hand in Quiz #1

- Bring Your Laptop

socialmedia-class.org