Social Media & Text Analysis

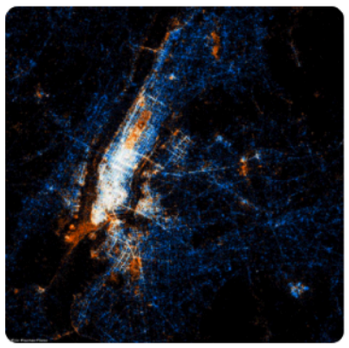
lecture 1 - Introduction

CSE 5539-0010 Ohio State University
Instructor: @alan_ritter
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 2017, CSE 5539-0010 The Ohio State University
Bolz Hall Room 318 | Tuesday 2:20PM – 4:10PM
dual-listed undergraduate and graduate course
[Office Hour] Dreese 495 | Tuesday 4:15PM – 5:15PM

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

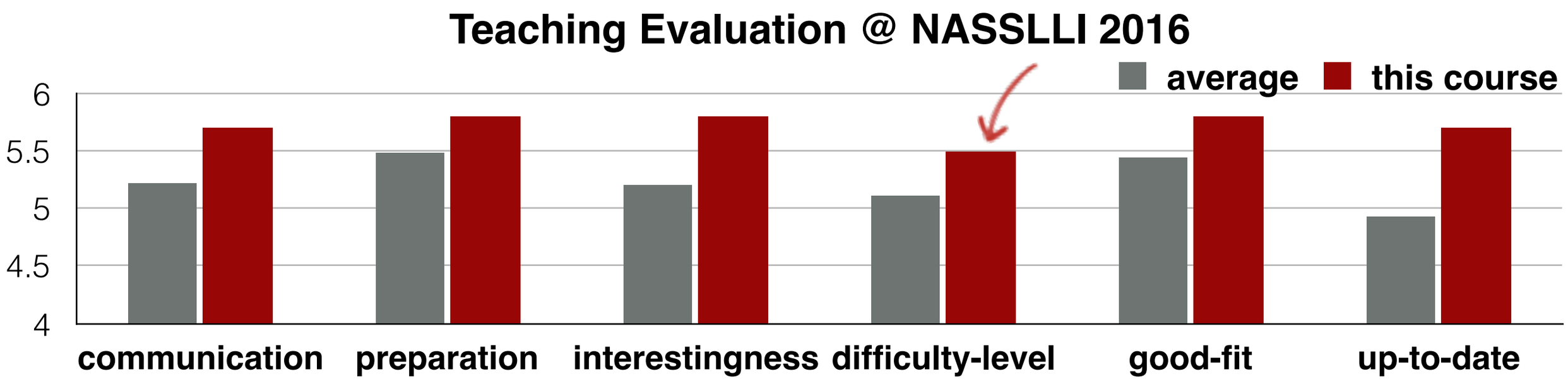
Various academic papers

Discussion Board

Piazza (TBA)

History of the Course

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



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

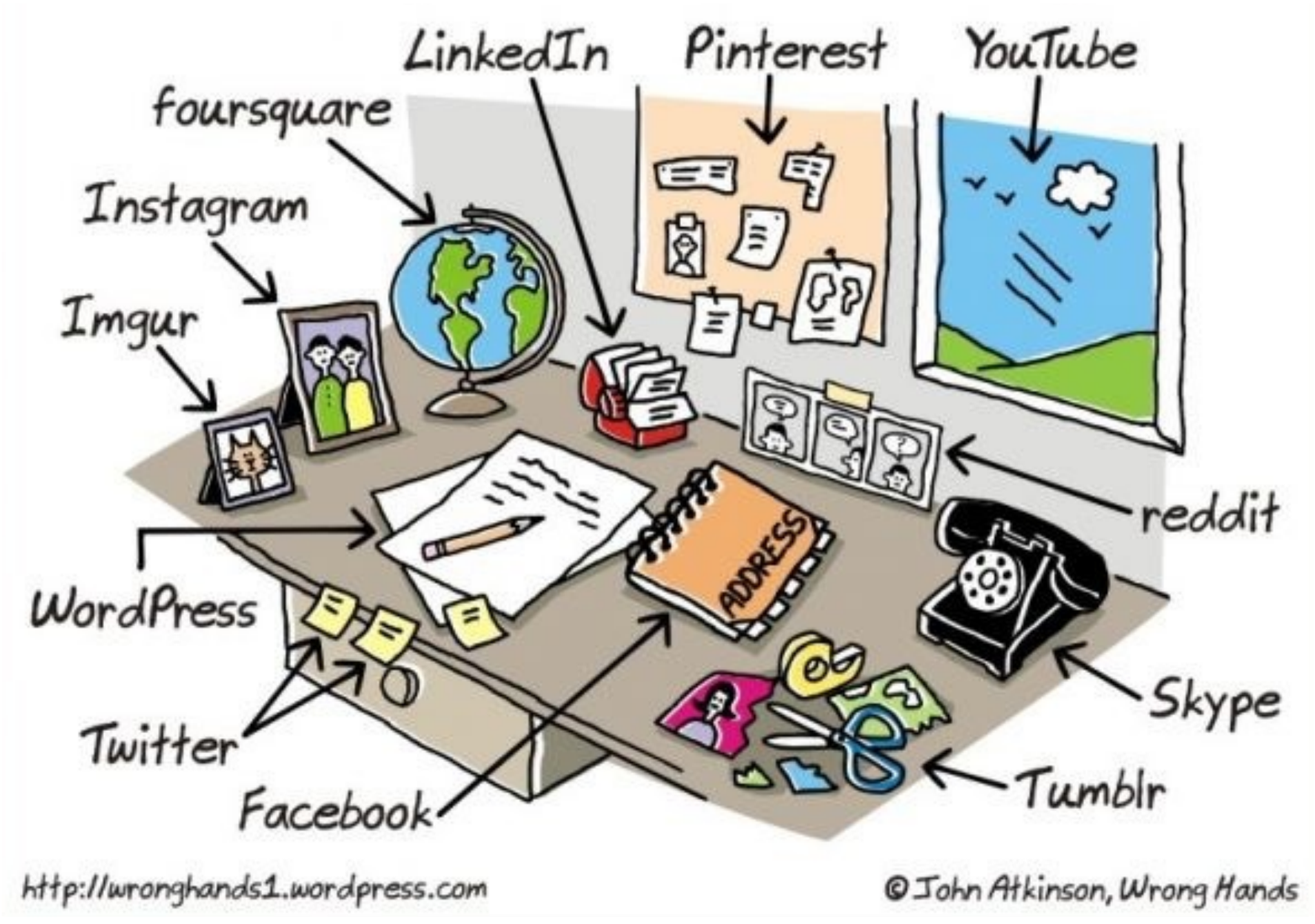
- Assistant Professor in CSE at the Ohio State University
- Postdoctoral researcher at Carnegie Mellon University Machine Learning Department
- PhD from University of Washington in Computer Science
- Research Areas:
  - Natural Language Processing
  - Machine Learning
  - Information Extraction
  - Social Media Analysis

TA: TBA

(supported by my research fund)

Why Social Media?

Vintage Social Media



Broader Point of View

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

skip
@han_horan

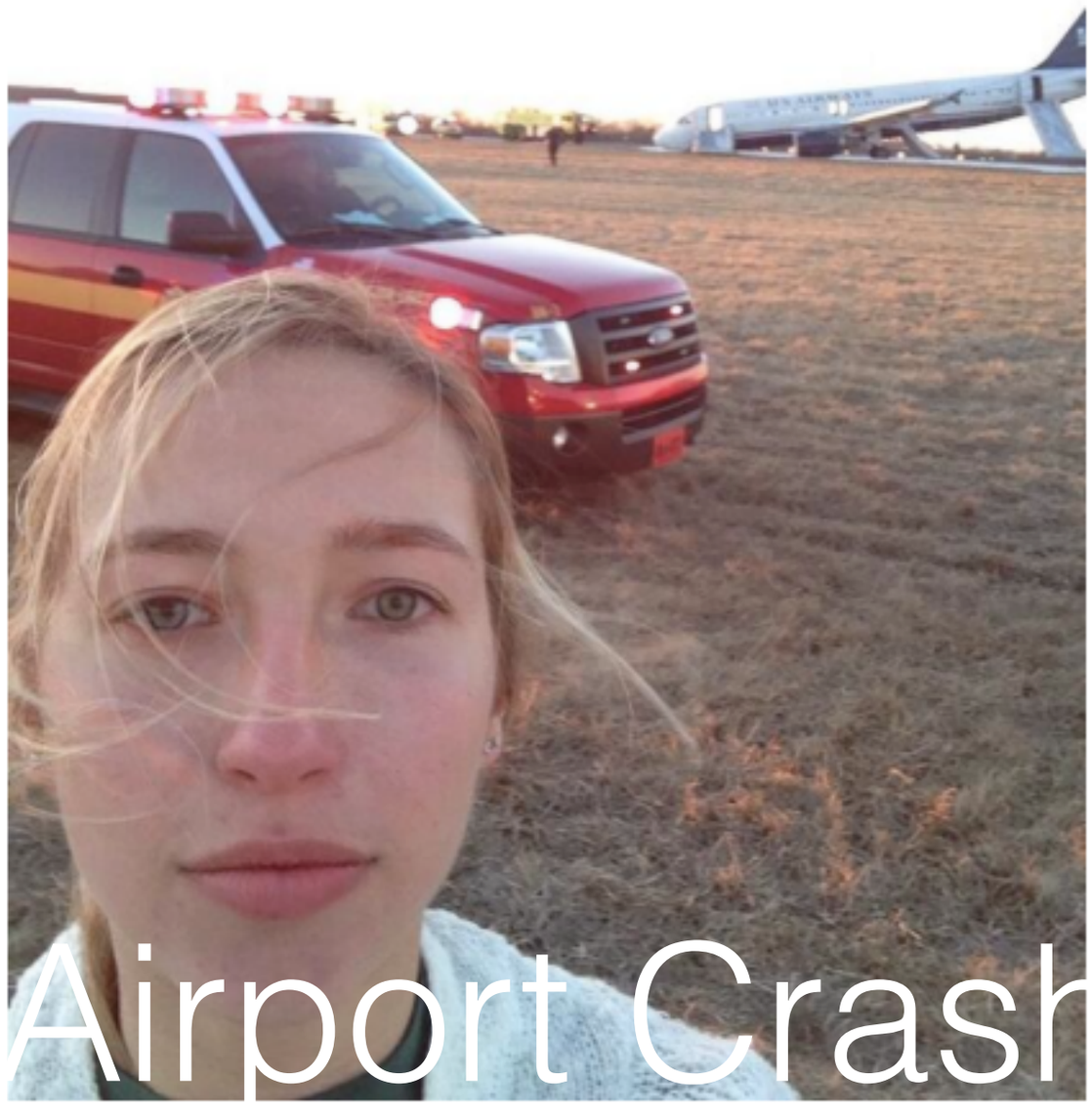
so my plane just crashed...
pic.twitter.com/X51BLwa5PS

Reply Retweet Favorite ••• More

skip
@han_horan

so yup pic.twitter.com/2WuLUWzpND

Reply Retweet Favorite ••• More

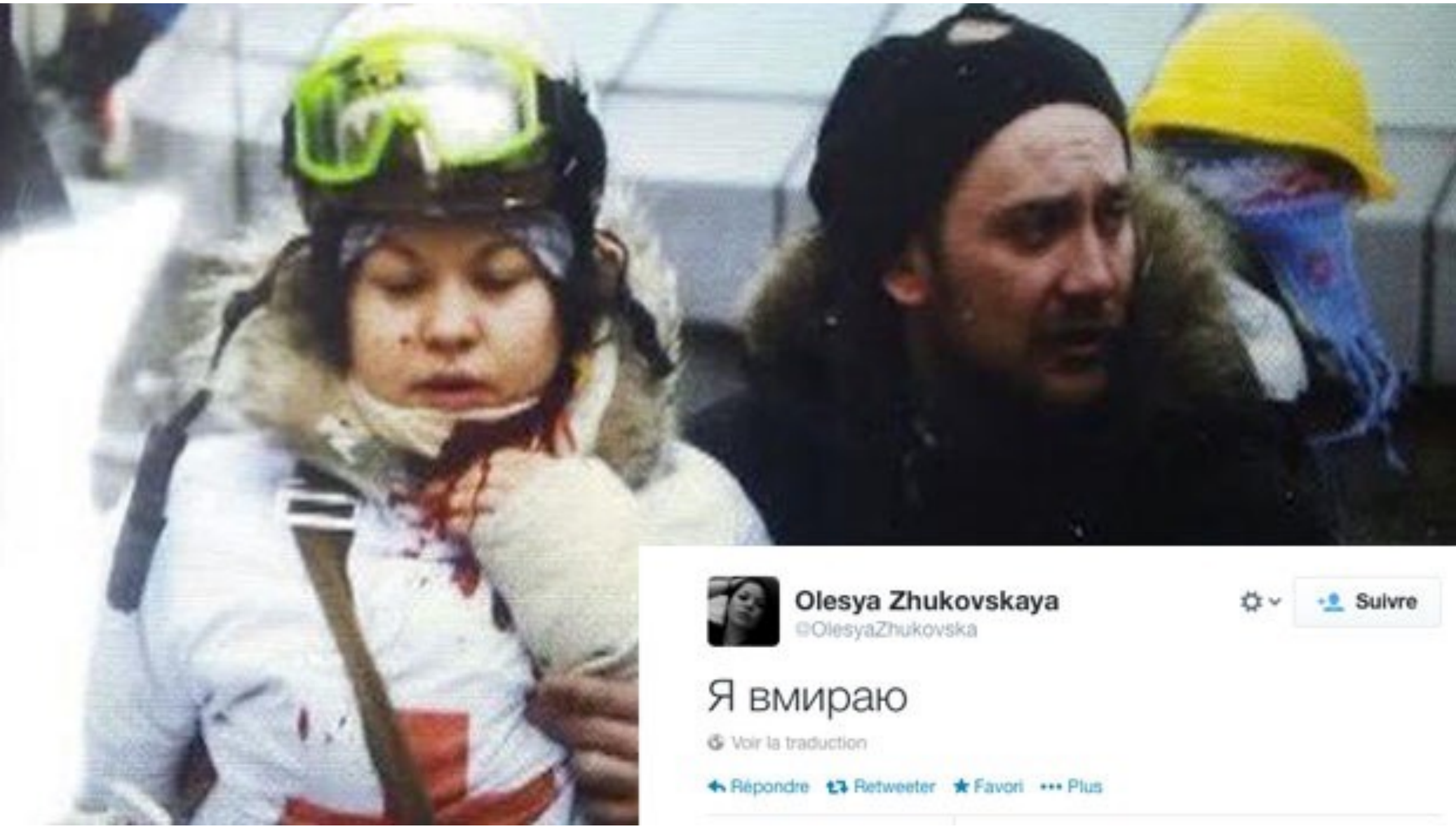


2014 Philly Airport Crash

Impact

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

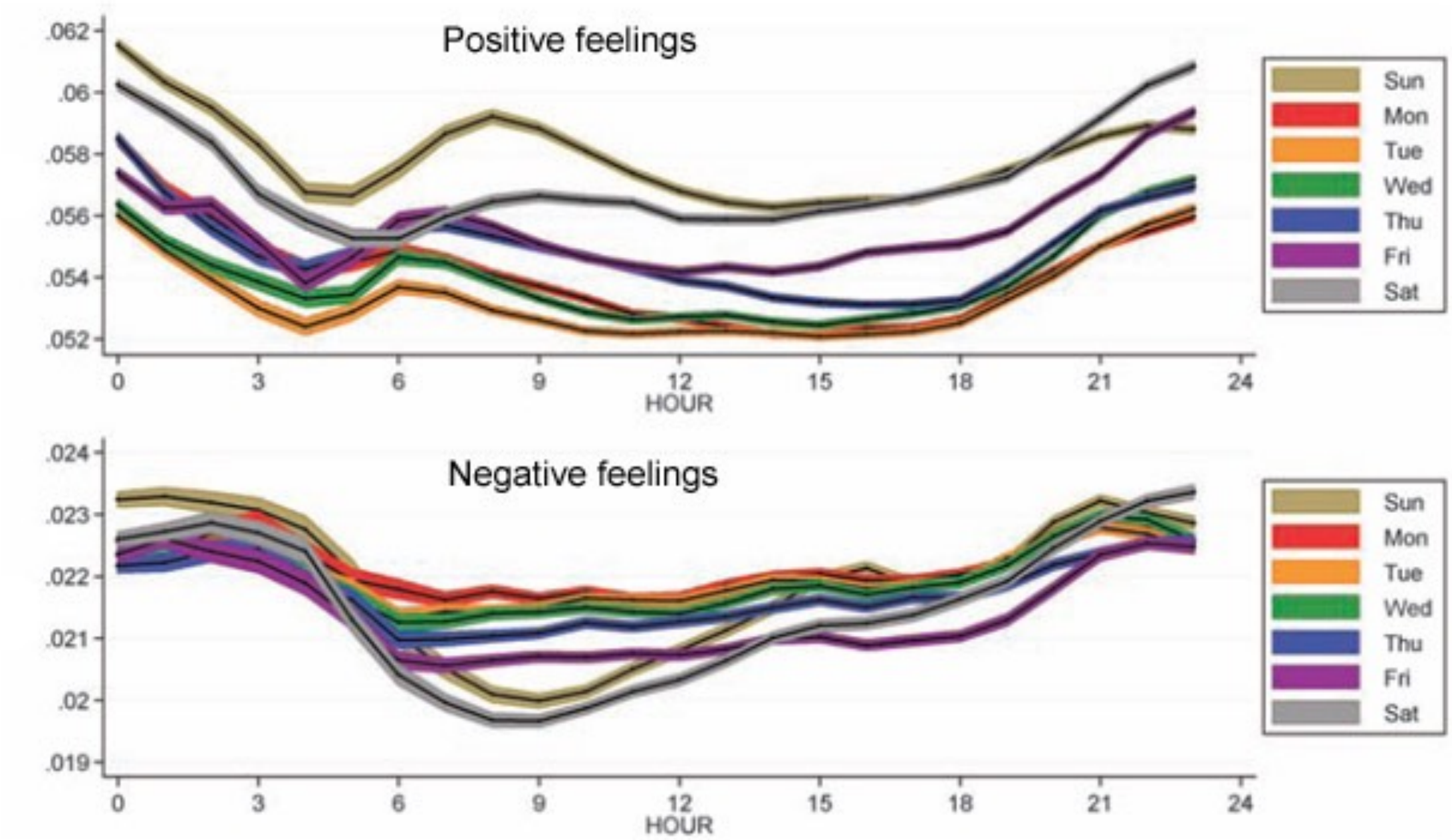
2014 Ukrainian Revolution



Research Value

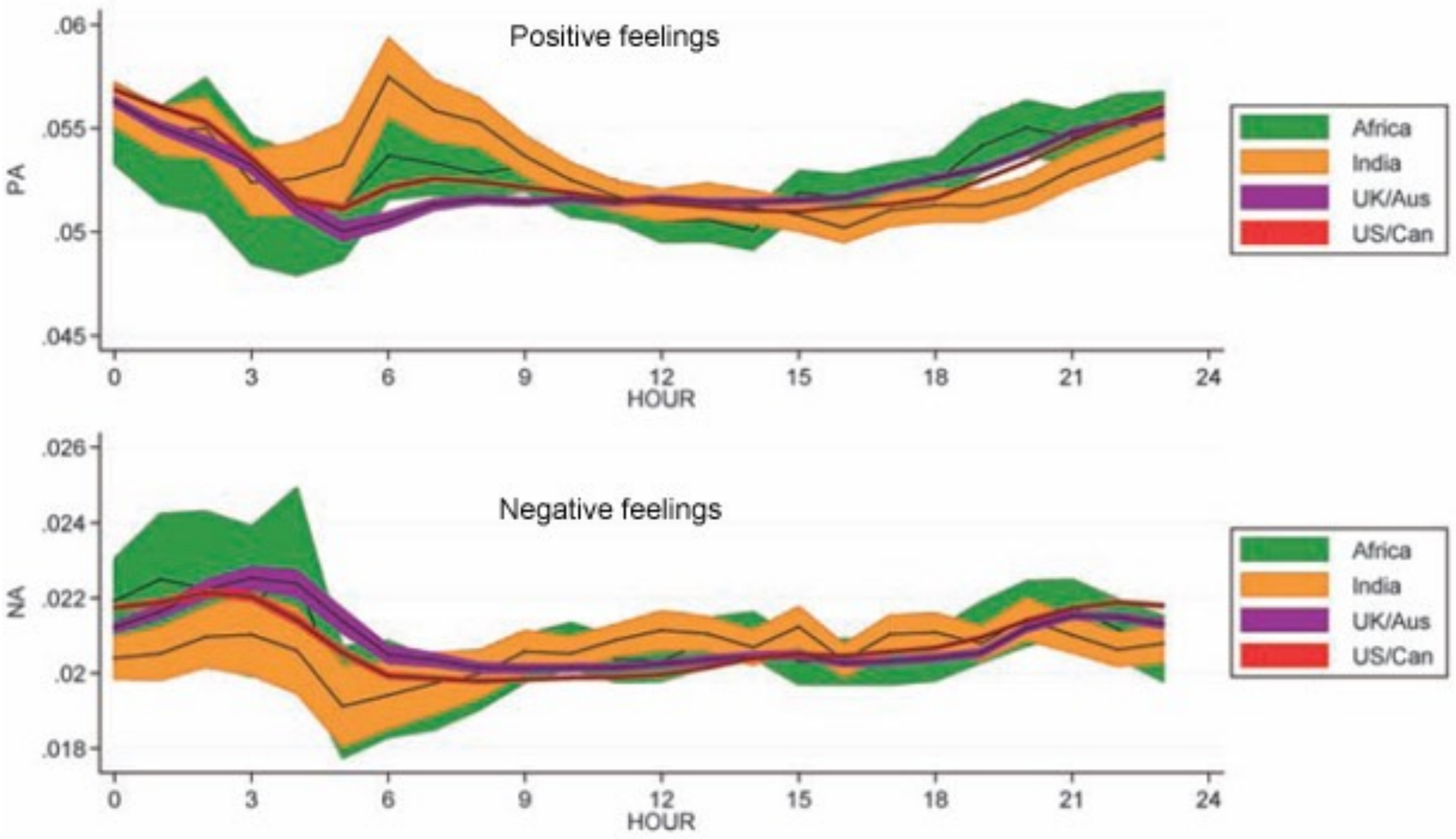
- In contrast to survey/self-report
- A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- Easy to access and aggregate **a lot** of data
- thus **a lot** of information

Mood



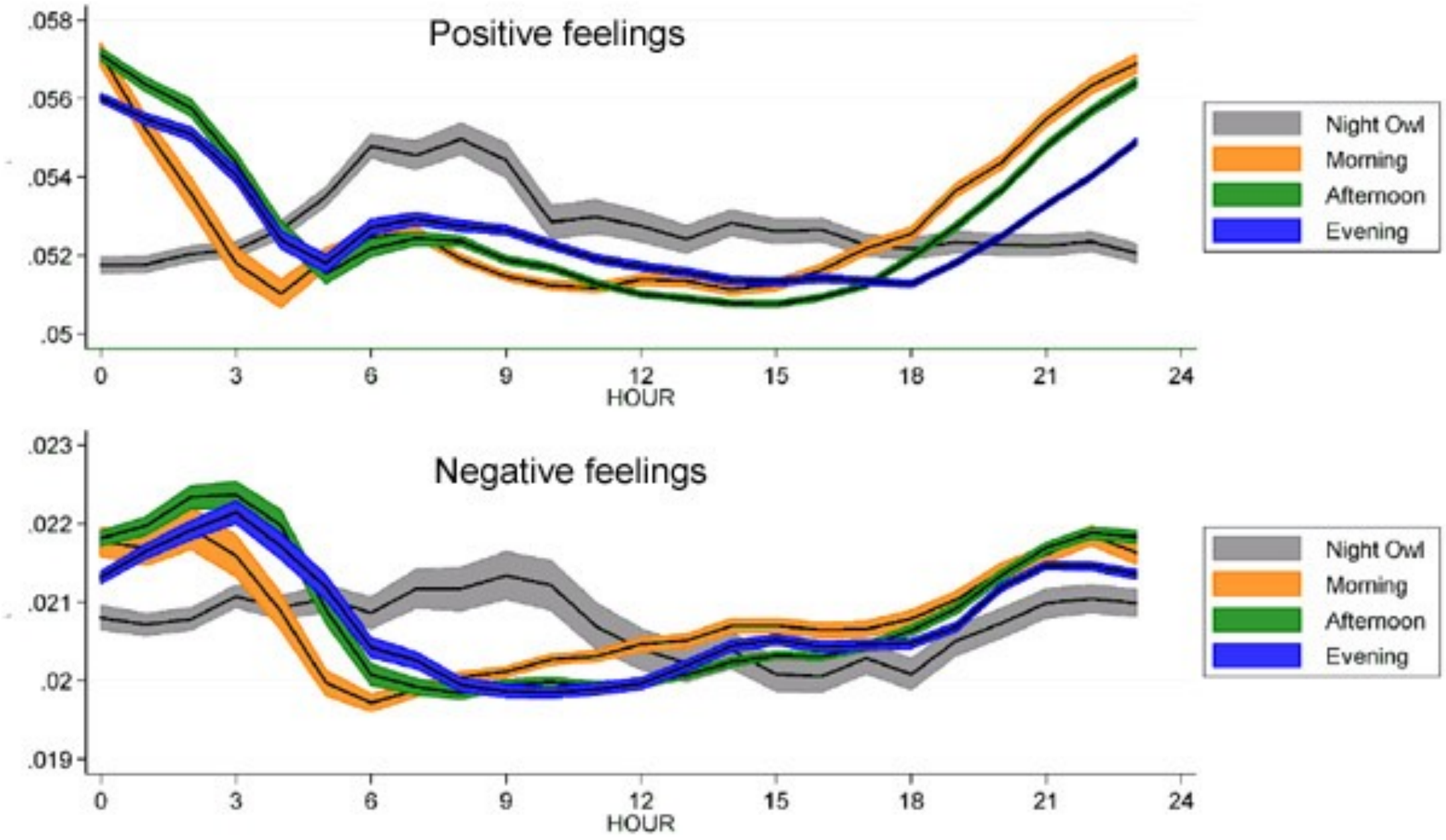
Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

Mood



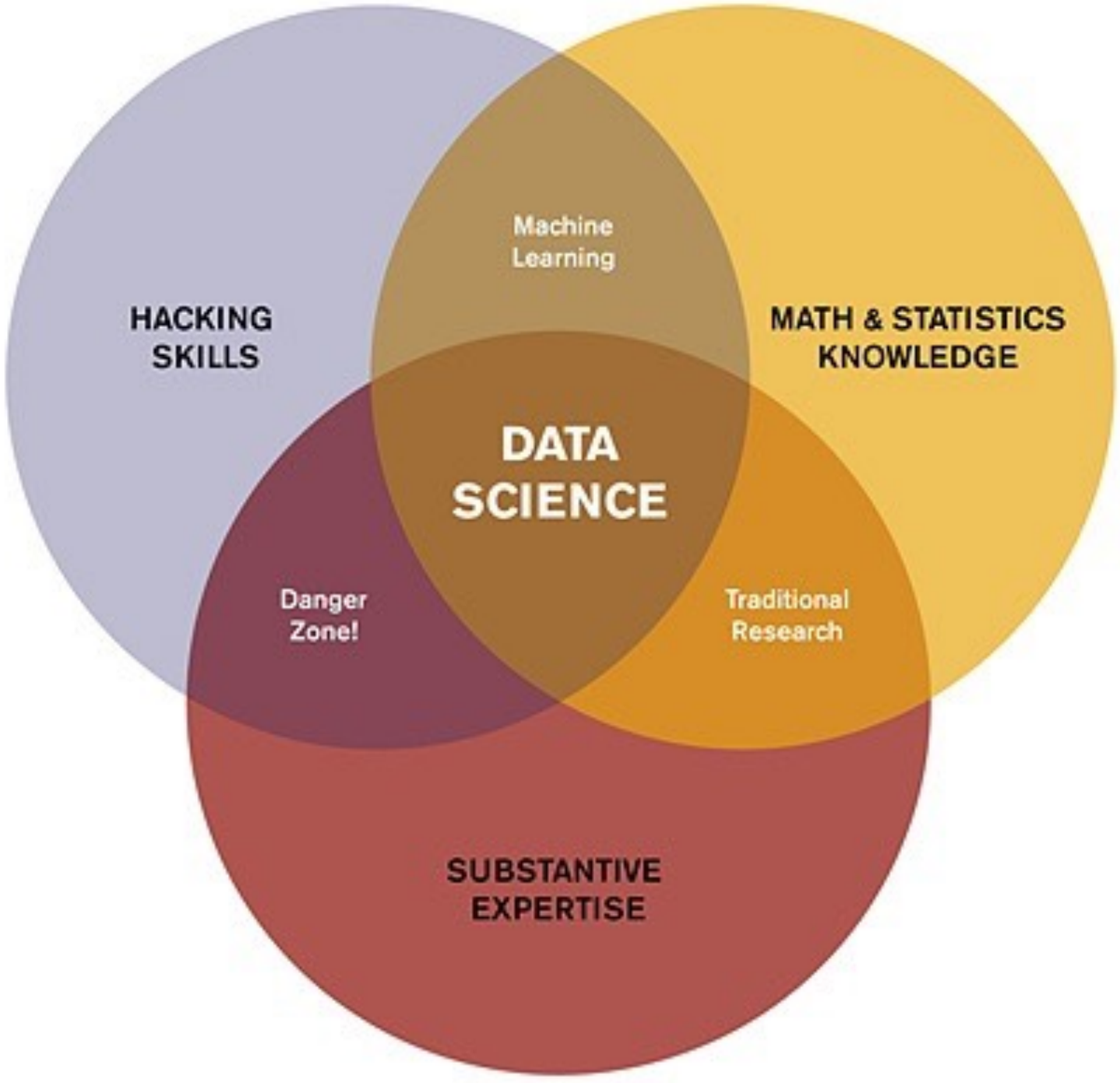
Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

Mood



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

Data Science



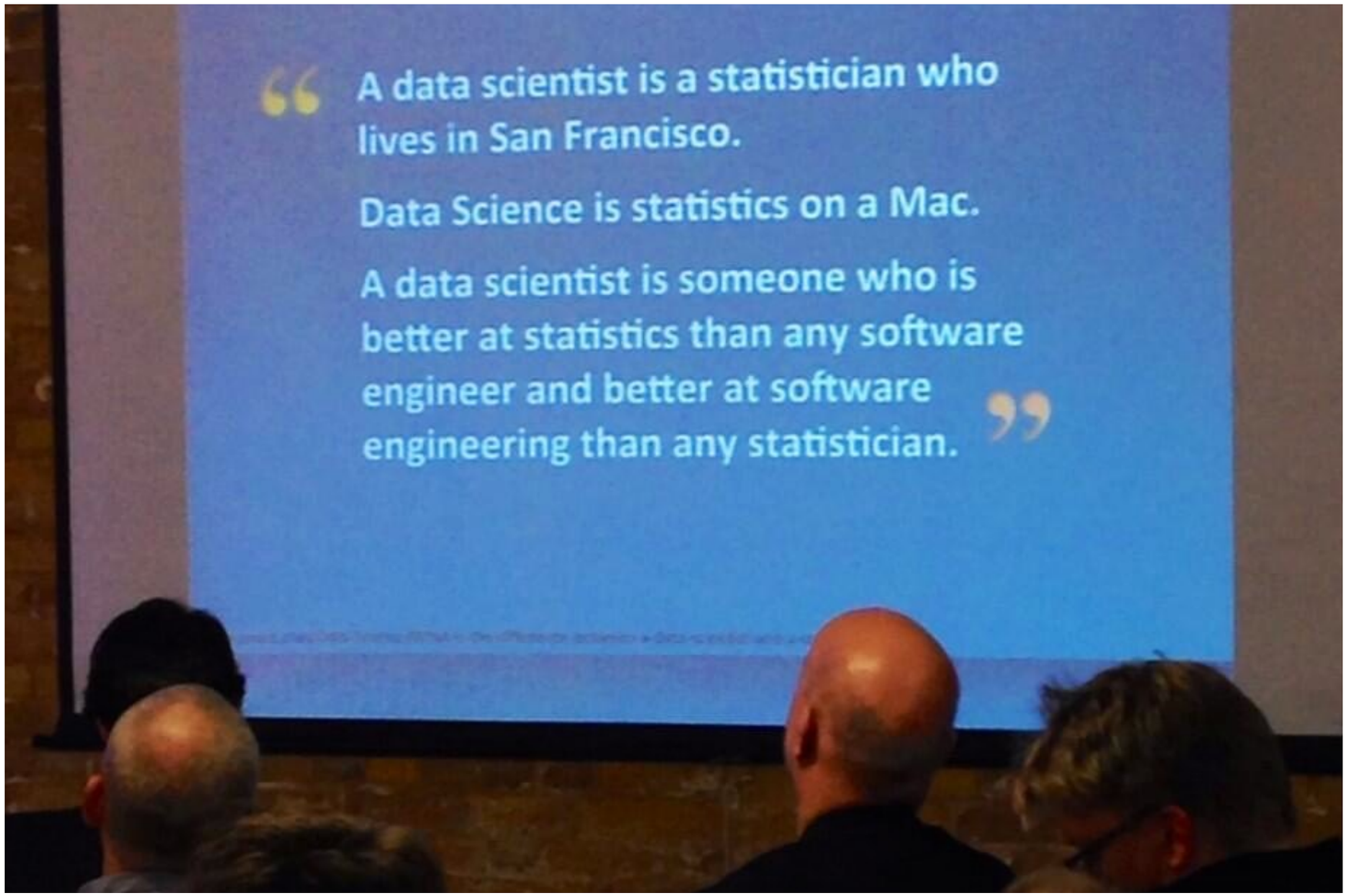
Source: Drew Conway

Data Science

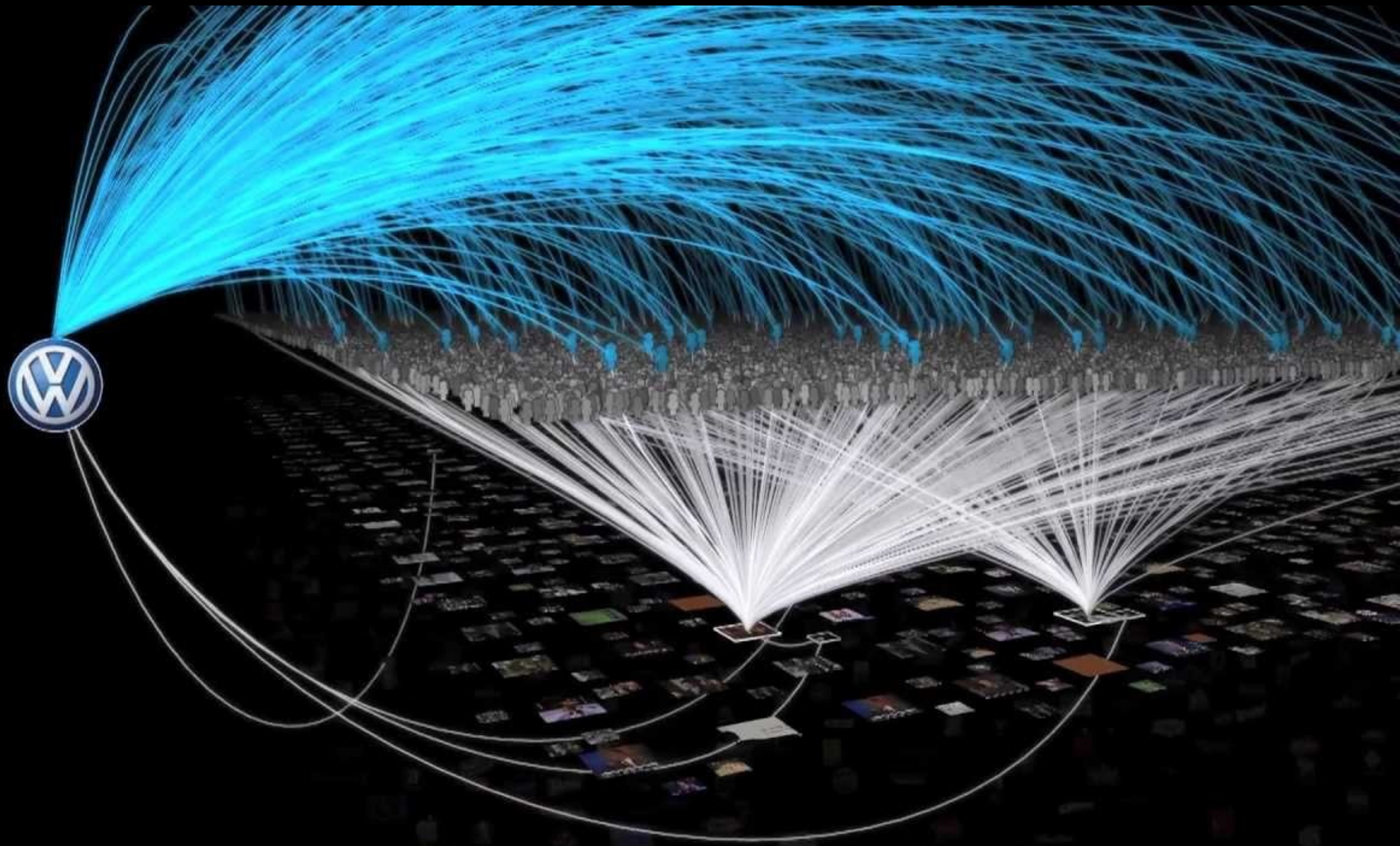
- is the **practice** of:
  - asking question (formulating hypothesis)
  - finding and collecting the data needed (often big data)
  - performing statistical and/or predictive analytics (often machine learning)
  - discovering important information and/or insights

Data Science

- the infamous definition:

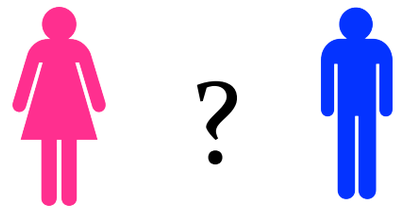


Marketing



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

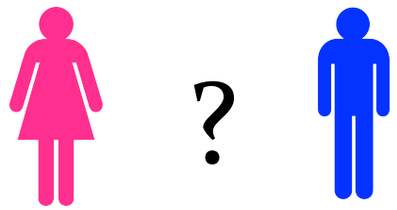
Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots

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

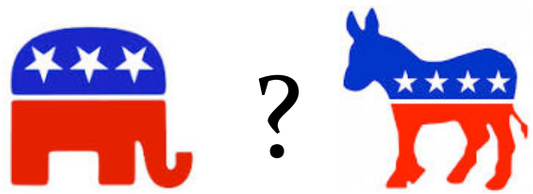


Follow

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.

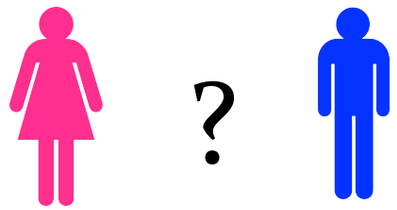


Source: Volkova, Van Durme, Yarowsky, Bachrach "Tutorial on Social Media Predictive Analytics" NAACL 2015

User Profiling

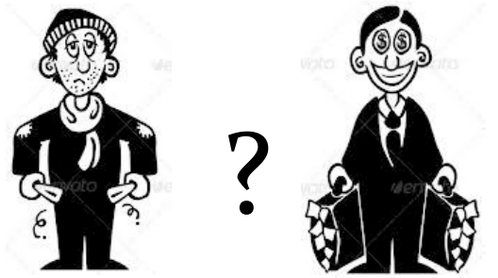
Follow

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



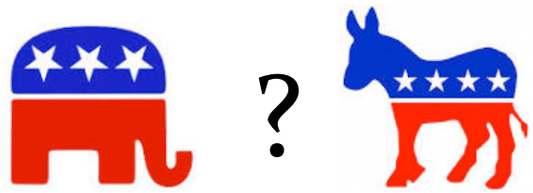
Follow

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



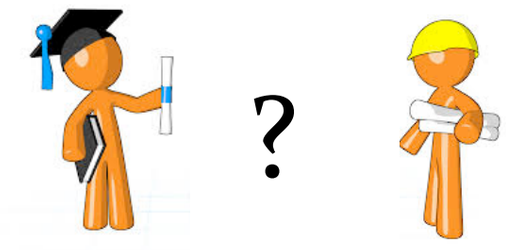
Follow

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



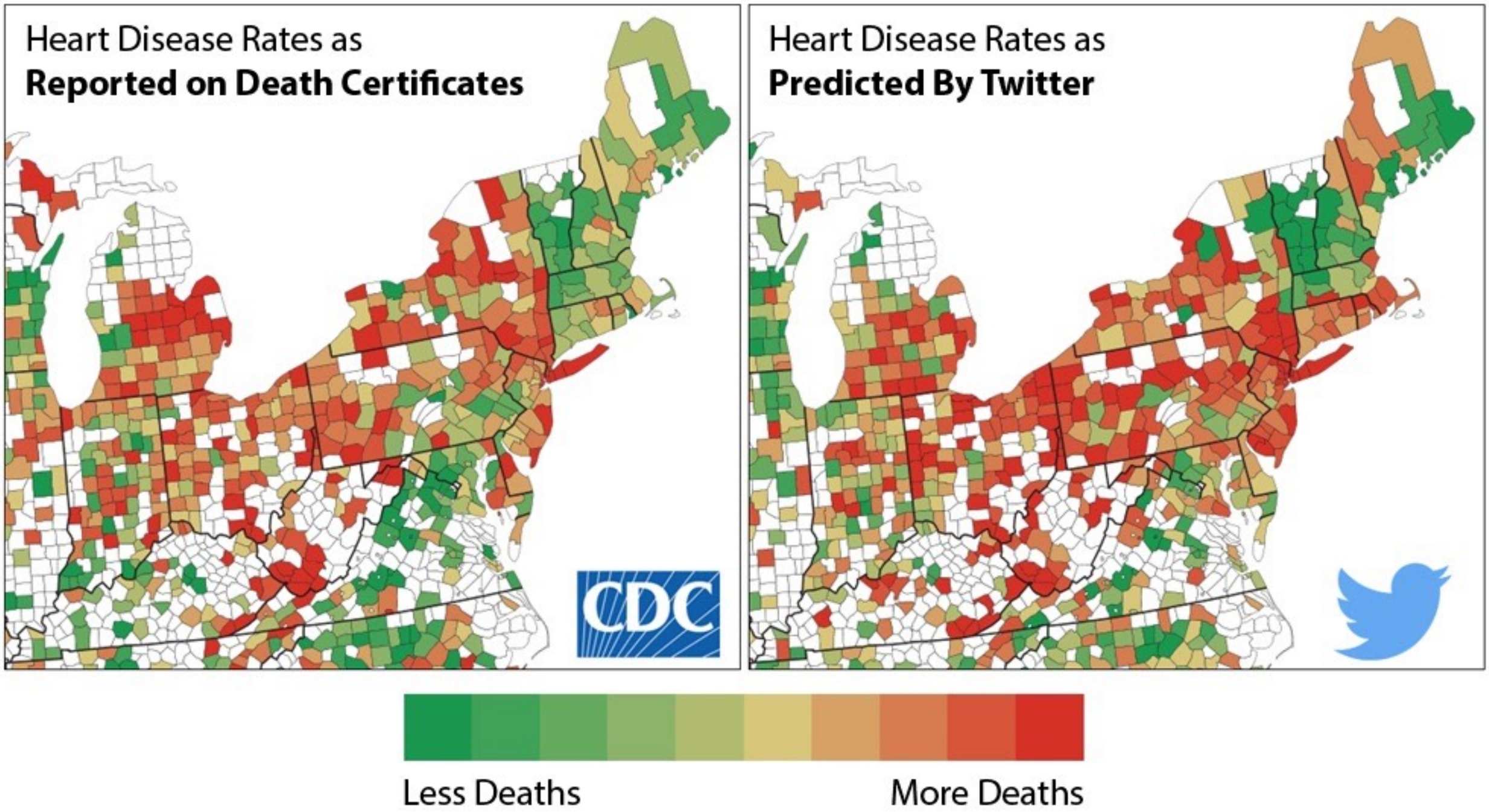
Follow

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

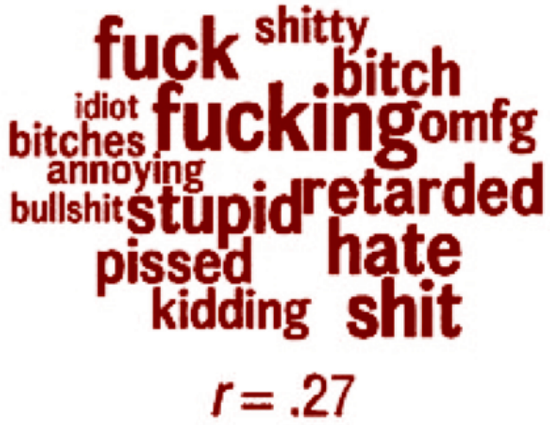
Health



Source: World Well-Being Project @ University of Pennsylvania

Health

Hostility, Aggression

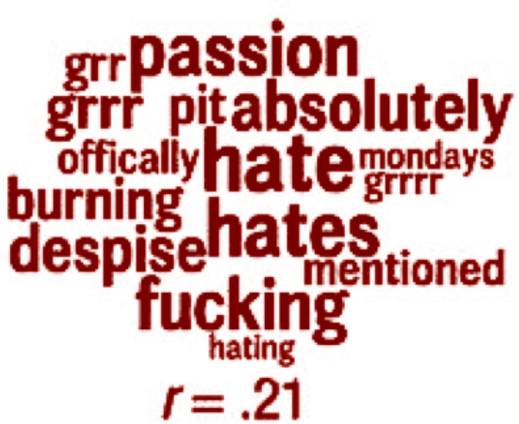


$r = .27$

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 @ University of Pennsylvania

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*

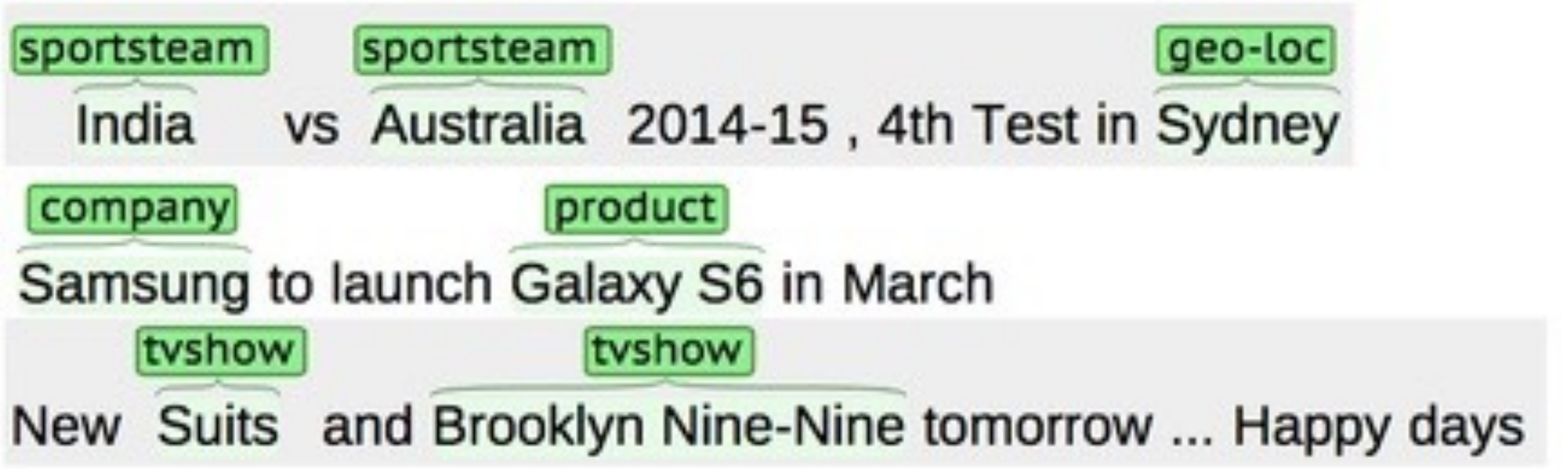
*this Nets vs Bulls game is **too live***

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

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

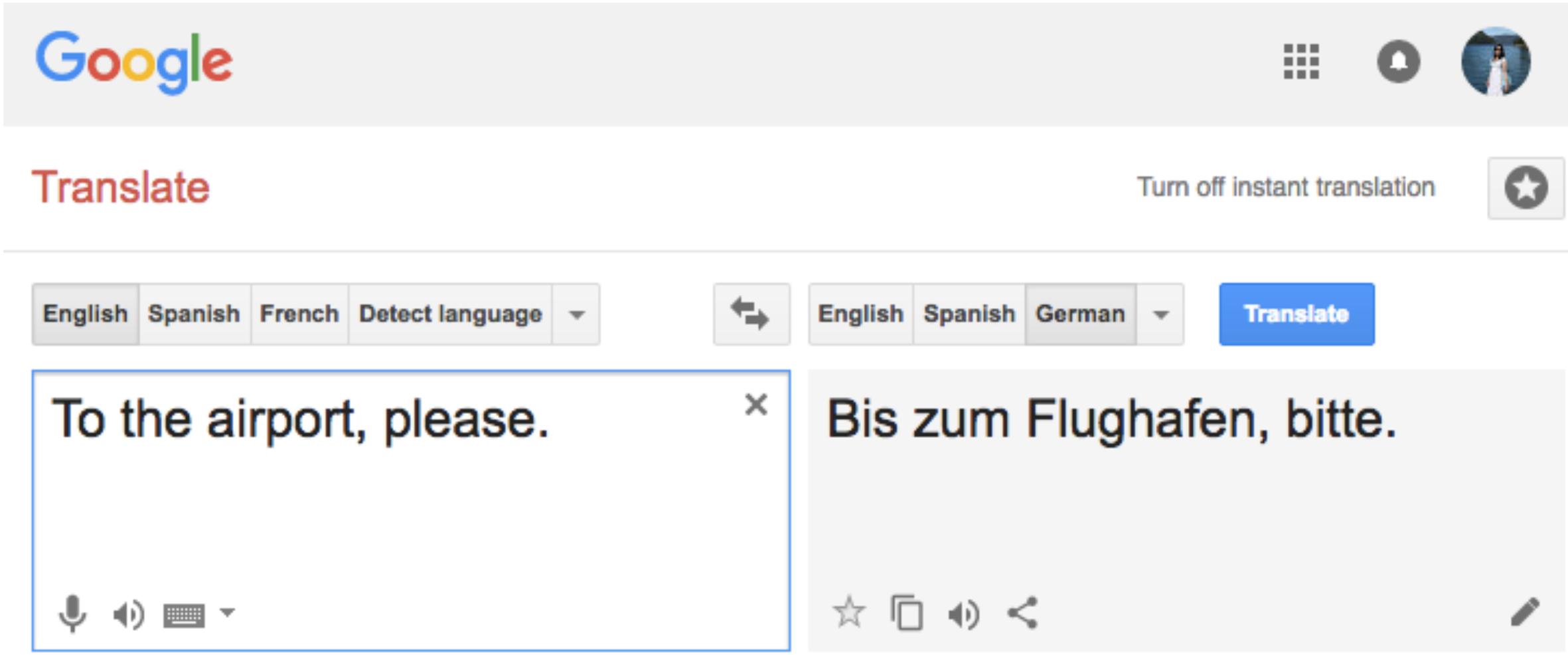
*This NetsBulls series is **intense***

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

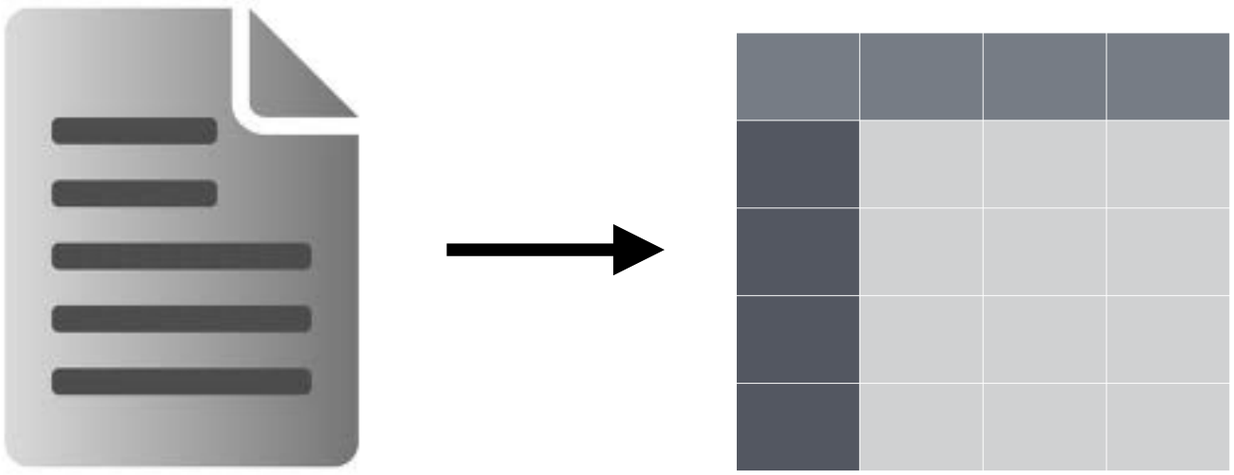
Machine Translation



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

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.*** *Carriers Beginning* ***April 11th***

COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
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 …*

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)

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

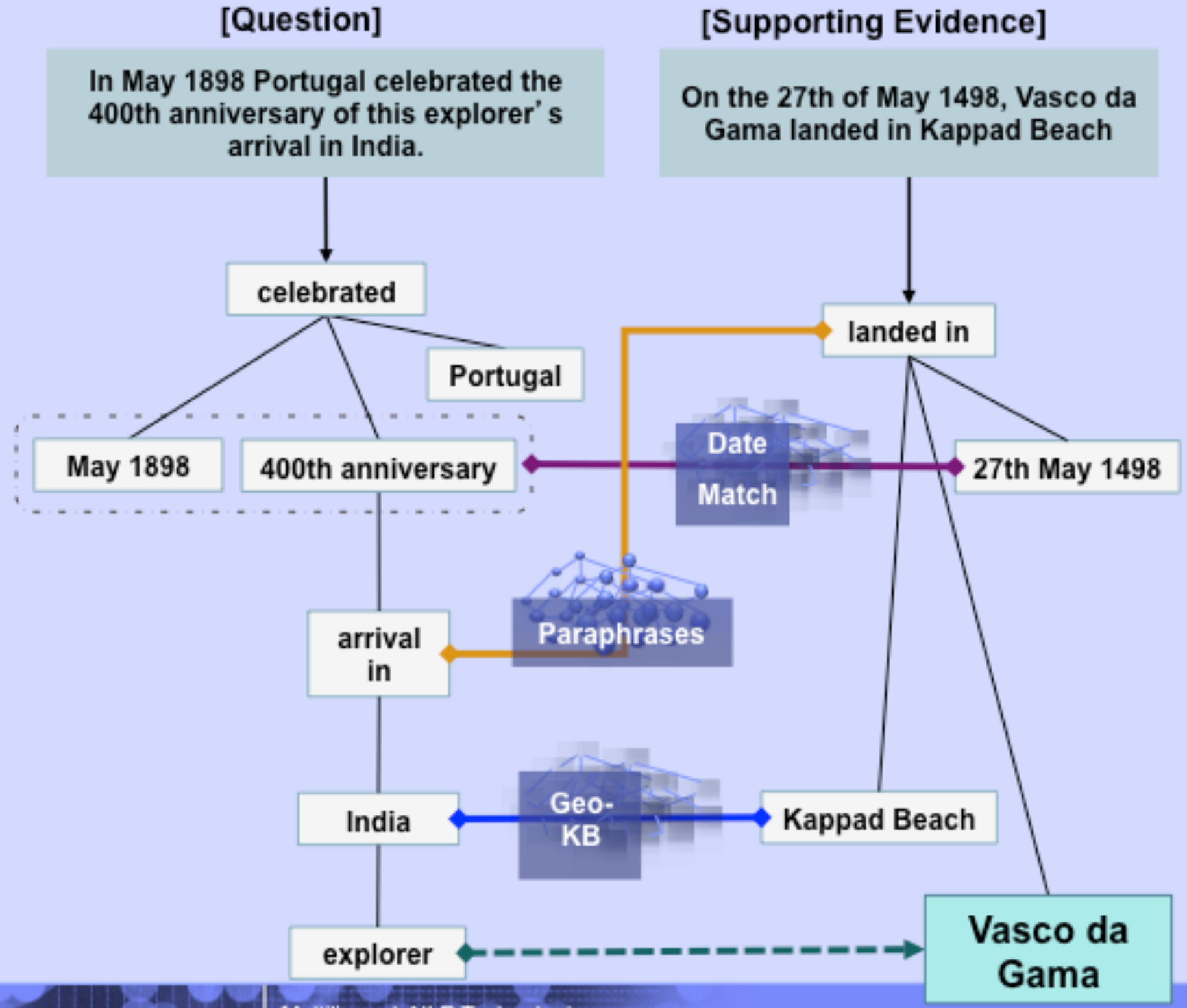
Who is the CEO stepping down from Boeing?

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

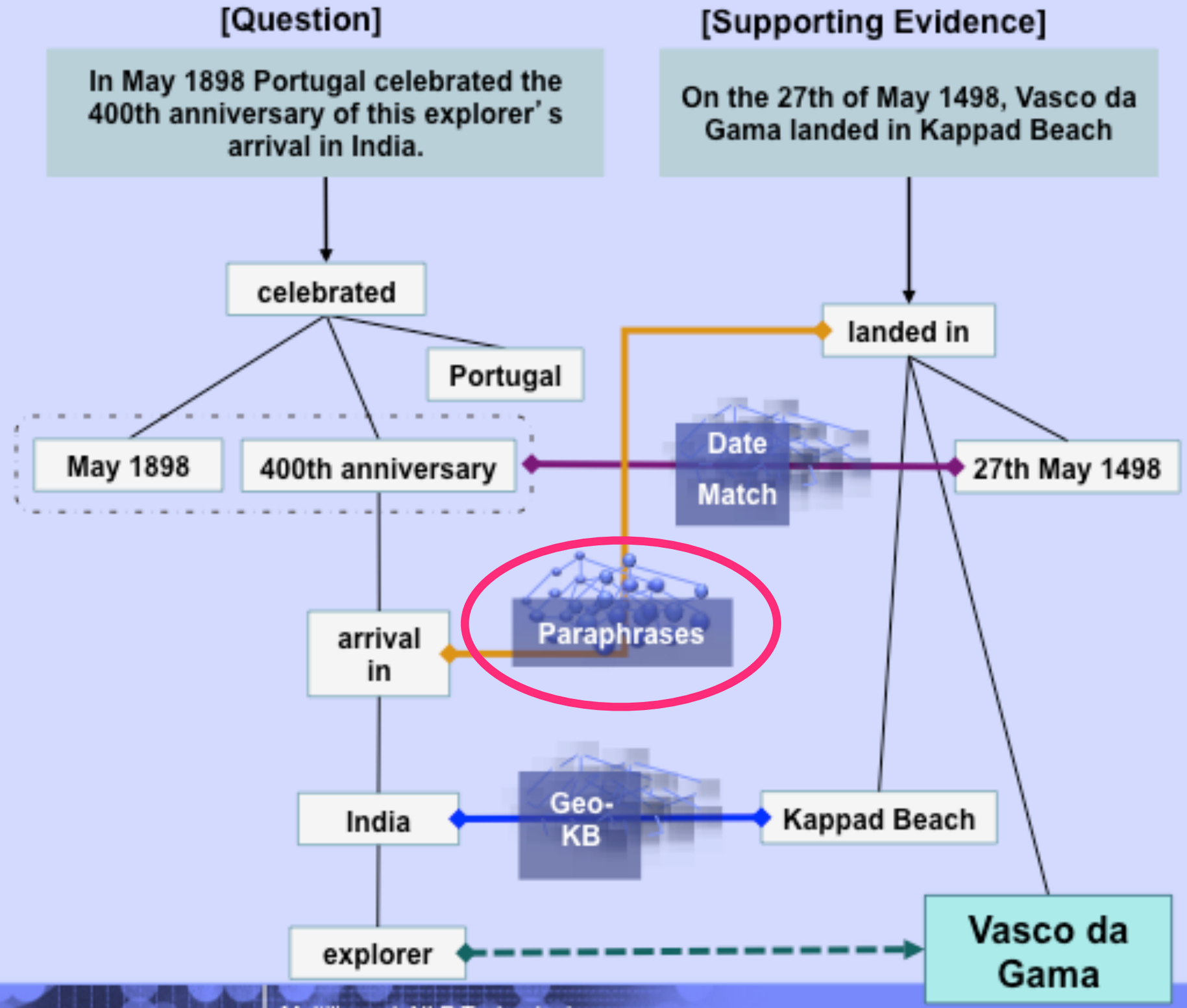


Stronger evidence can be much harder to find and score…

- 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)

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?

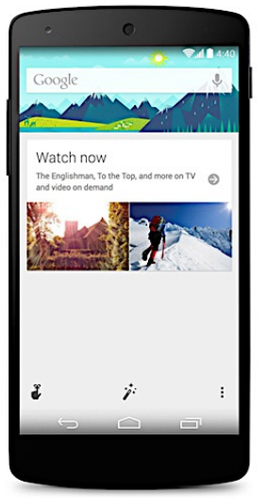
who else wants to get a beer?

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 ACL (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!

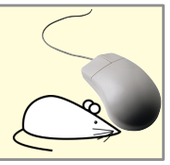
The waiter ignored us for 20 minutes.

Coreference resolution

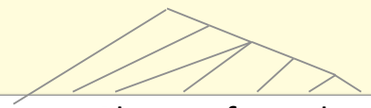
Carter told Mubarak he shouldn't run again.

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
  - Some deep learning basics

Guest Lectures

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

Grading

- two programing assignments (45 pts/individual)
- A 3rd assignment/research project (**optional**, 20 bonus pts)
- in-class presentation (20 pts/group of two)
- paper summaries (20 points/individual, about 10 papers)
- several take-home Quizzes (10 points/individual)
- participation in class discussions (5 pts)

Programming Assignments

- All in Python
- two programing assignments (45 points — 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 10 minute presentation (20 points)
  - A Social Media Platform
  - Or a NLP Researcher

Quizzes

- several simple take-home quizzes (about 5 or 6)
- hard-copy on paper
- will not be graded; but count for 10 points
- We have **Quiz #1 today** on pre-requirements!

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!

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

Stylistic Language Generation

- Data and code:

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

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

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
  - CMU students: https://arxiv.org/abs/1707.01161

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

Language Styles

wonderfully delightfully beautifully fine well good nicely superbly

she says

he says

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

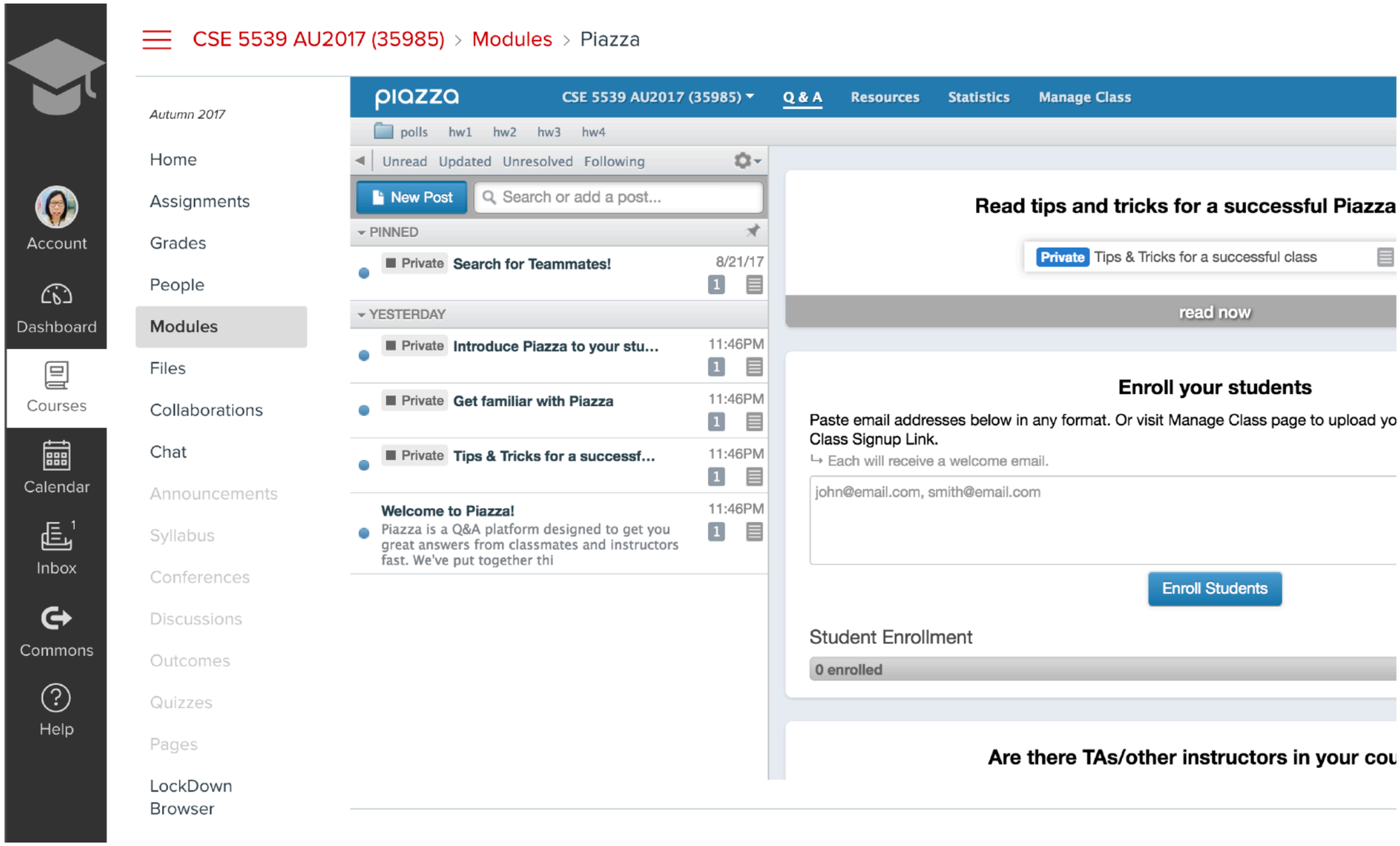
What will you get out of this class?

- 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

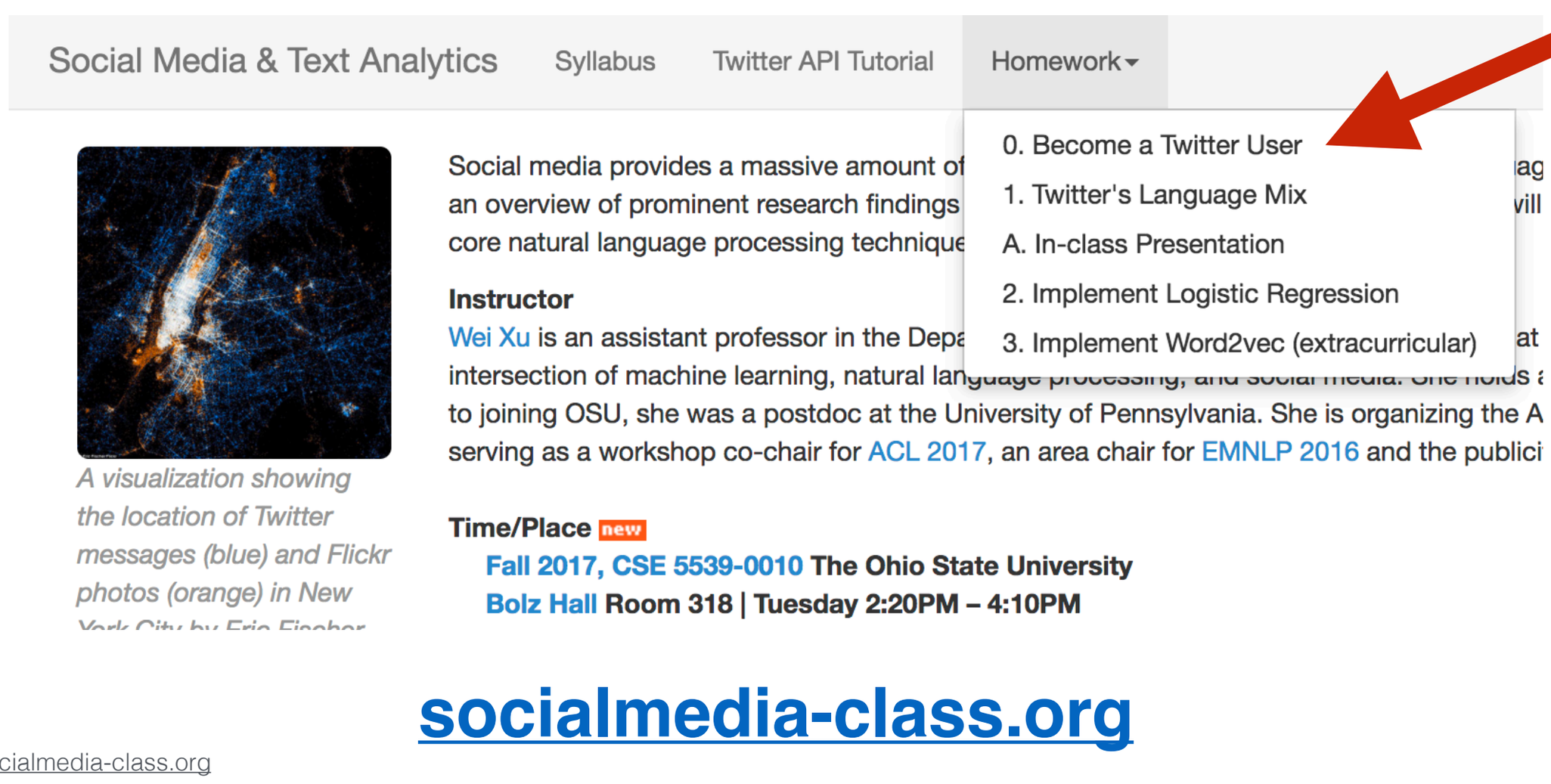
- Have a question? Ask in/after class
- Or ask on Piazza discussion broad
- Office hour — Mondays 4-5pm (Dreese 595)
    - No office hours on the 22nd

Piazza Discussion Broad



By Next Class:

- Hand in Quiz #1
- HW#0 Become a Twitter User



socialmedia-class.org