

# Social Media & Text Analysis

## lecture 1 - Introduction

**CSE 5539-0010 Ohio State University**  
**Instructor: @alan\_ritter**  
**Website: [socialmedia-class.org](http://socialmedia-class.org)**

# Course Website

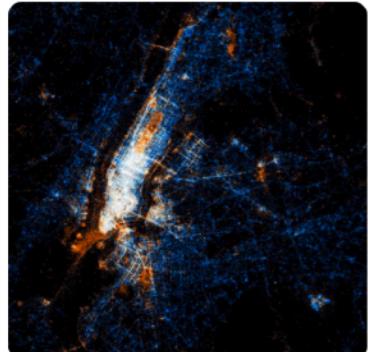
## <http://socialmedia-class.org/>

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](#)

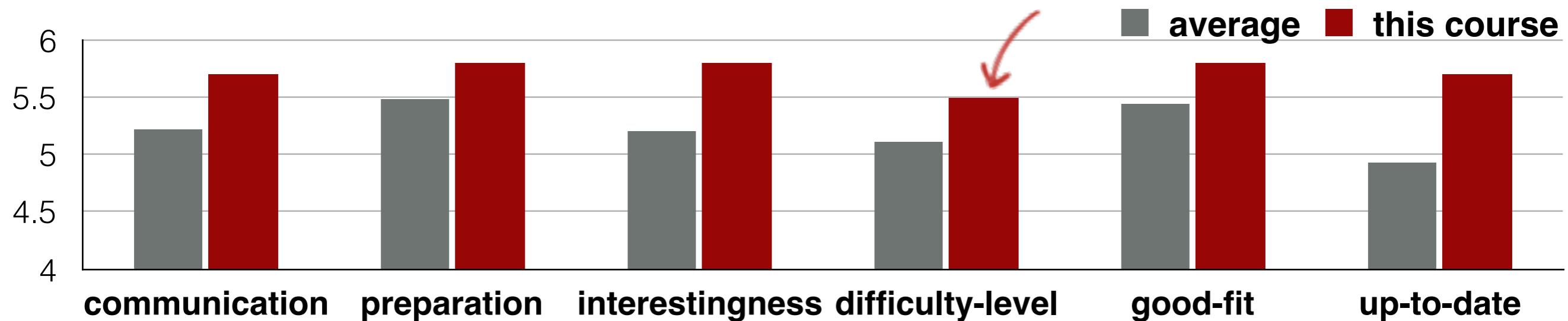
### Discussion Board

[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

**Teaching Evaluation @ NASSLLI 2016**



# 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

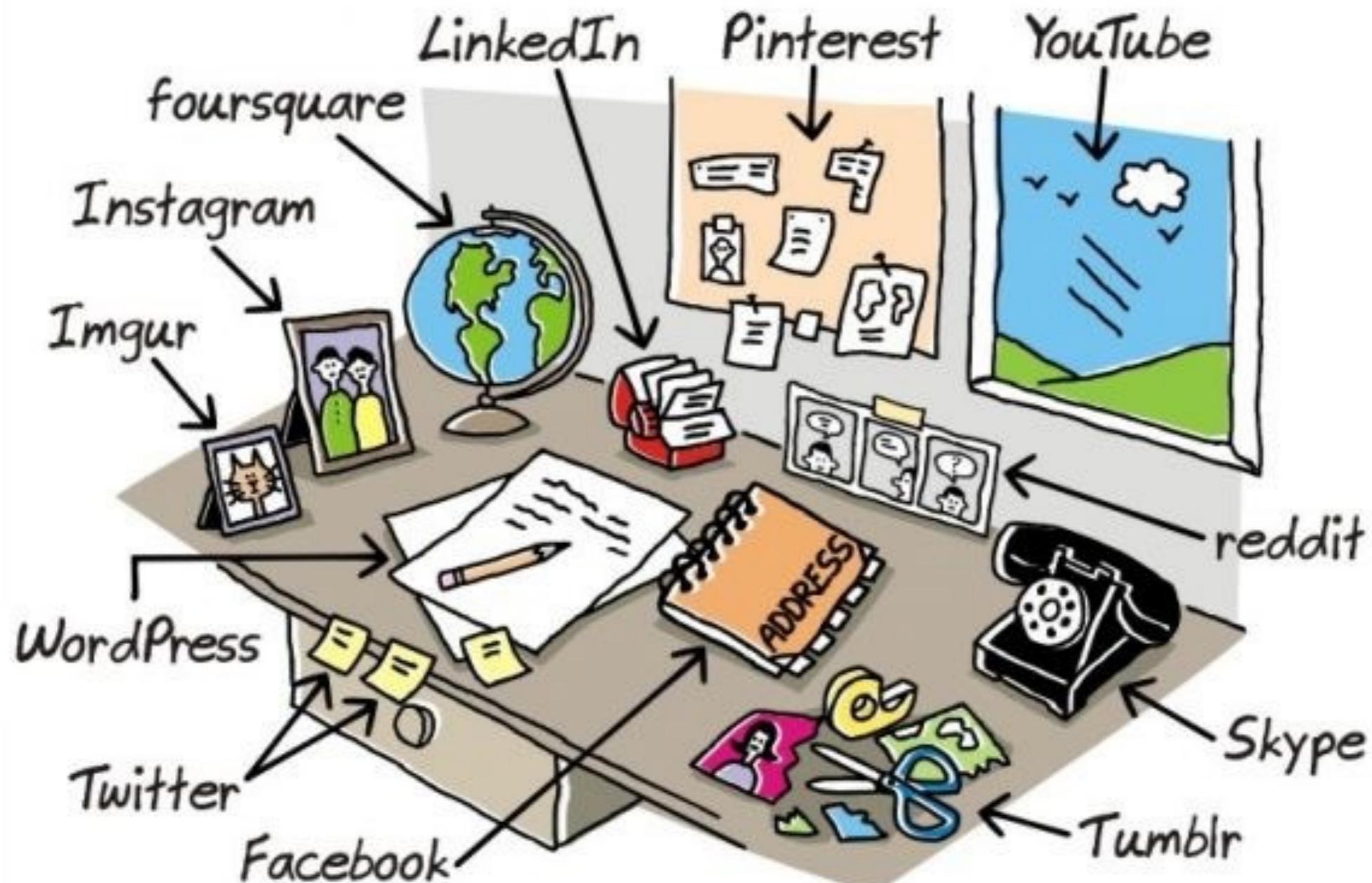
- Assistant Professor in CSE at the Ohio State University
- Postdoctoral researcher at Carnegie Mellon University Machine Learning Department
- PhD from University of Washington in Computer Science
- Research Areas:
  - Natural Language Processing
  - Machine Learning
  - Information Extraction
  - Social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

**TA: TBA**

(supported by my research fund)

# Why Social Media?

# Vintage Social Media



<http://wronghands1.wordpress.com>

© John Atkinson, Wrong Hands

# Broad Point of View



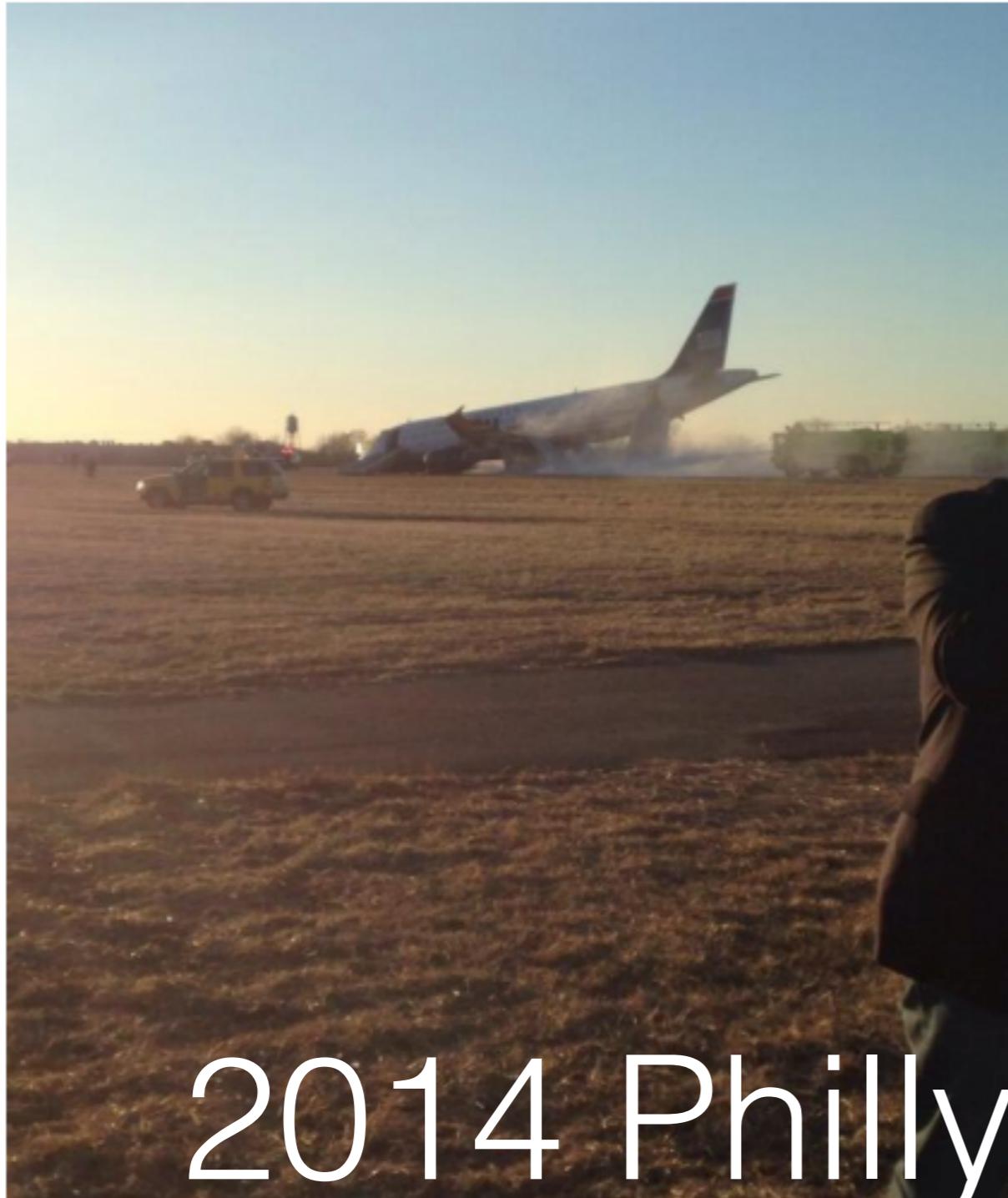


skip

@han\_horan

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

↪ Reply ⚡ Retweet ★ Favorite ... More

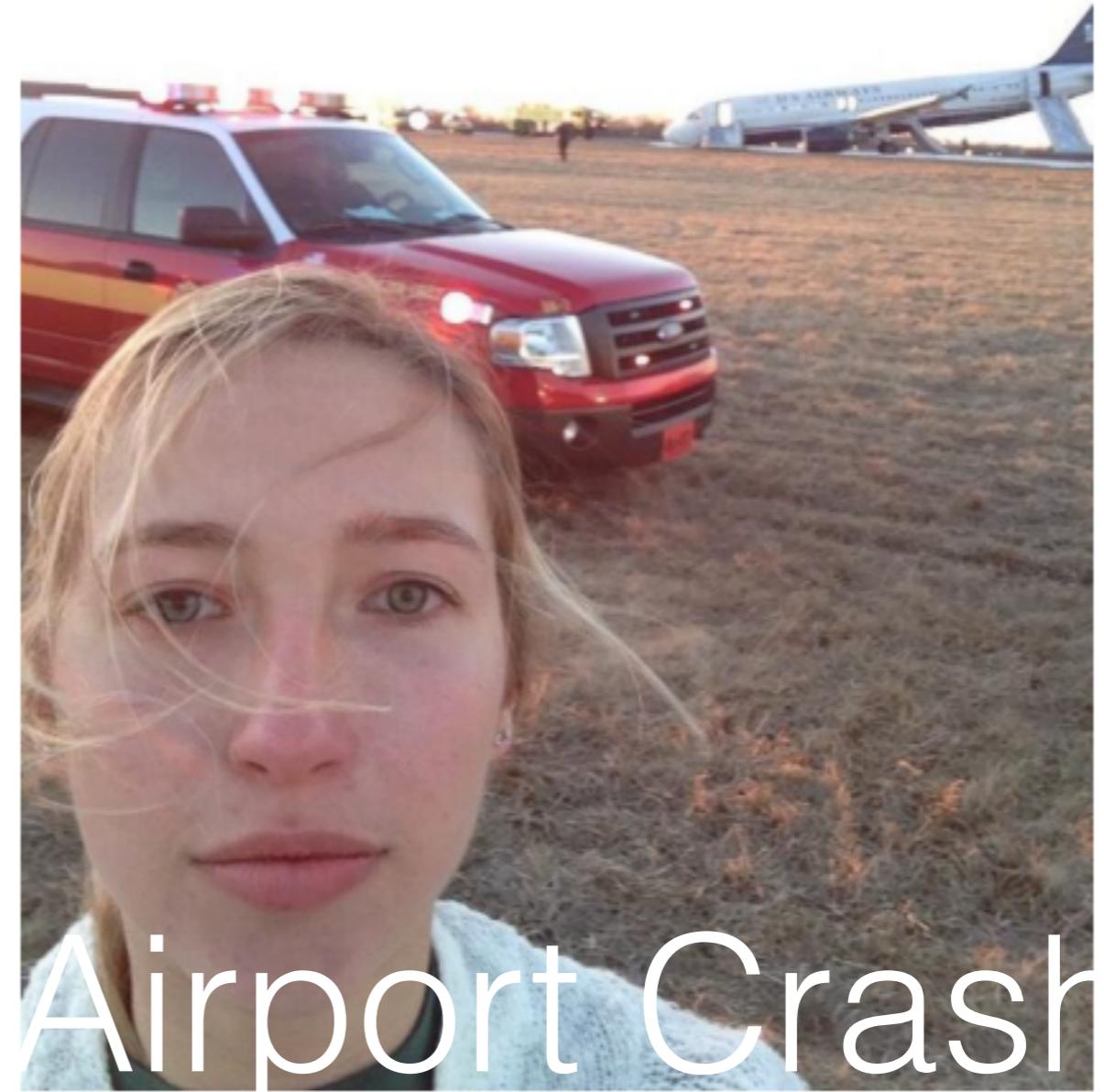


skip

@han\_horan

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

↪ Reply ⚡ Retweet ★ Favorite ... More



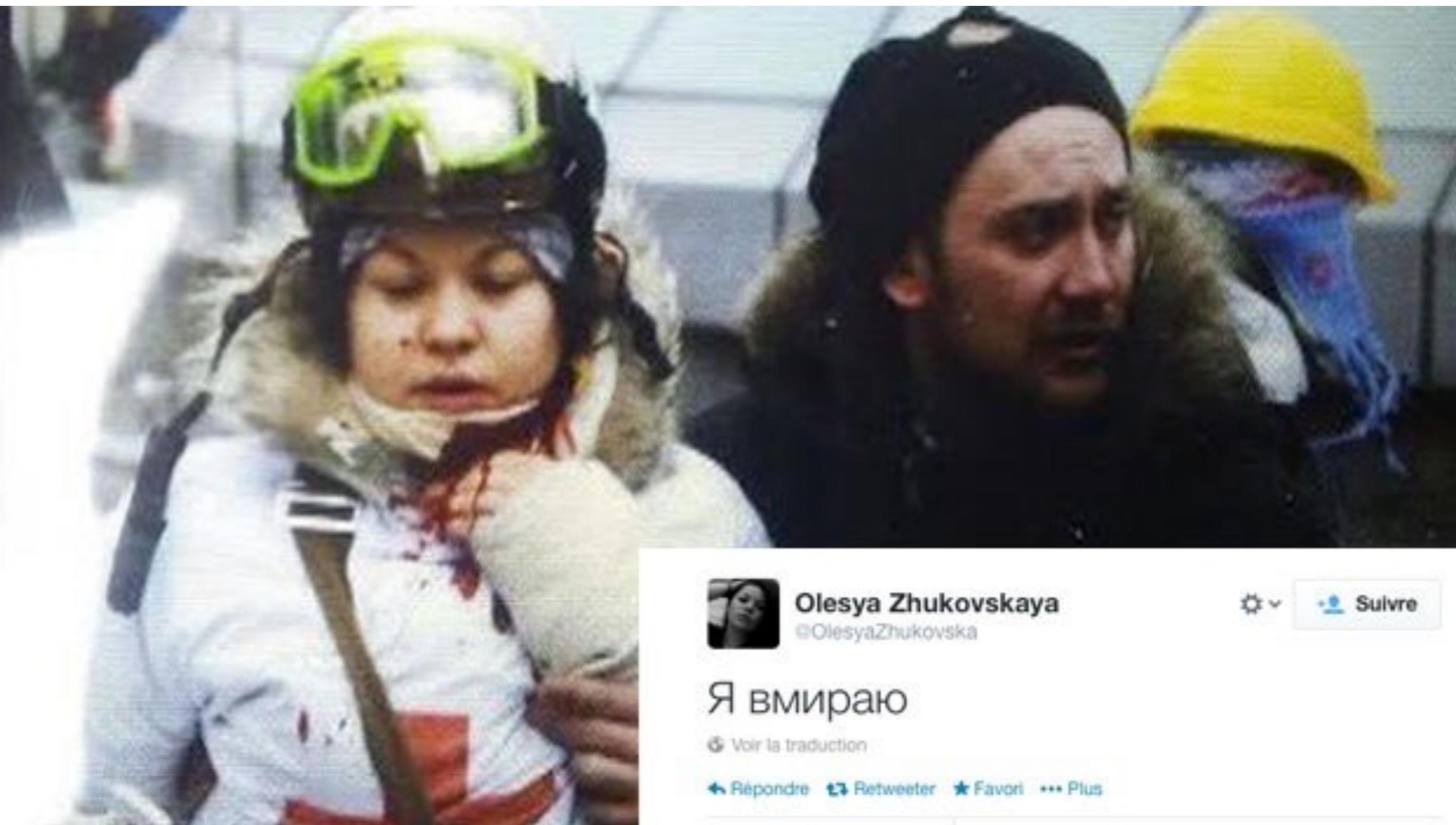
Airport Crash

# Impact

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



# 2014 Ukrainian Revolution



Olesya Zhukovskaya

@OlesyaZhukovska



Suivre

Я вмираю

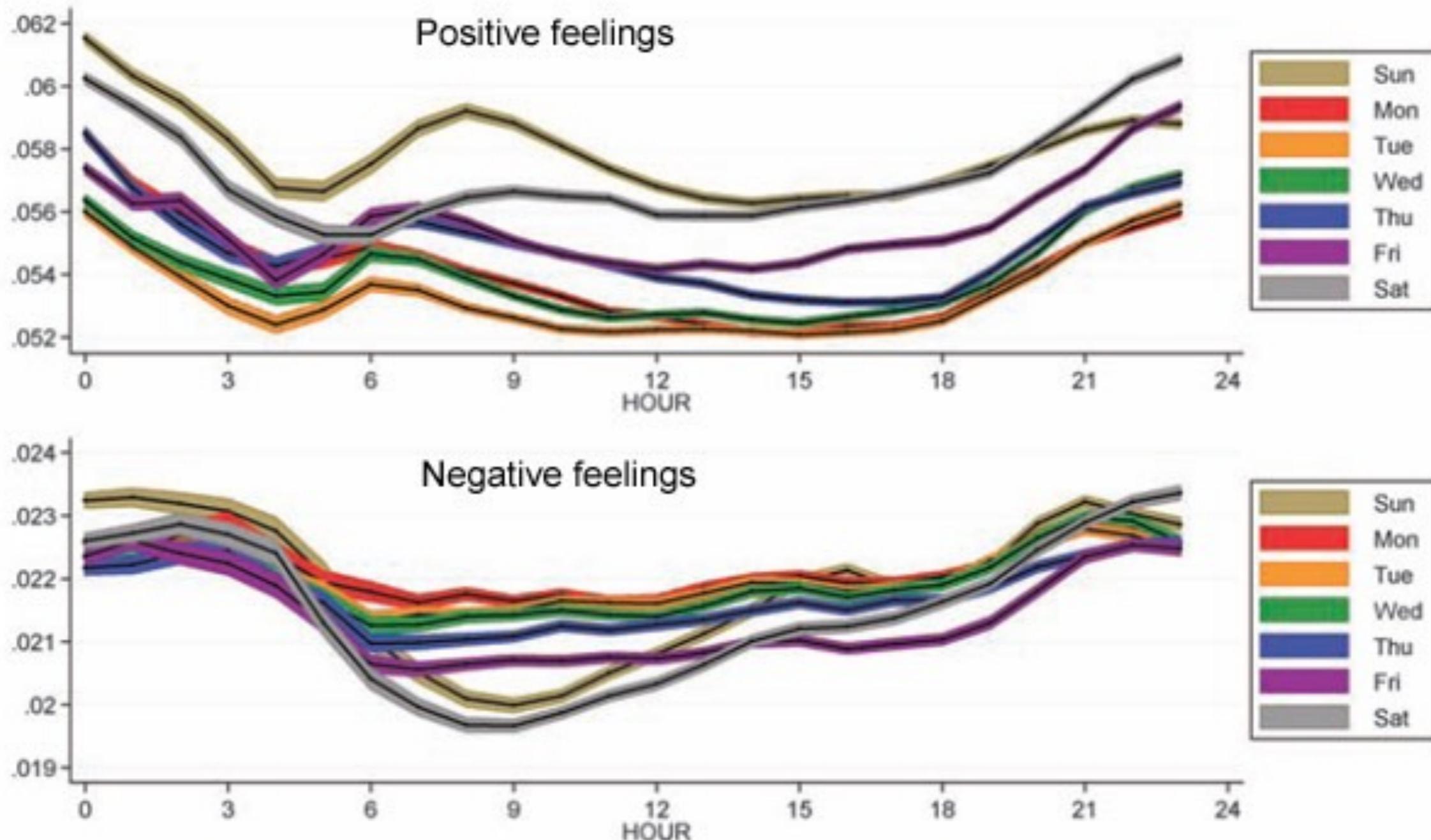
Voir la traduction

Repondre Retweeter Favori Plus

# Research Value

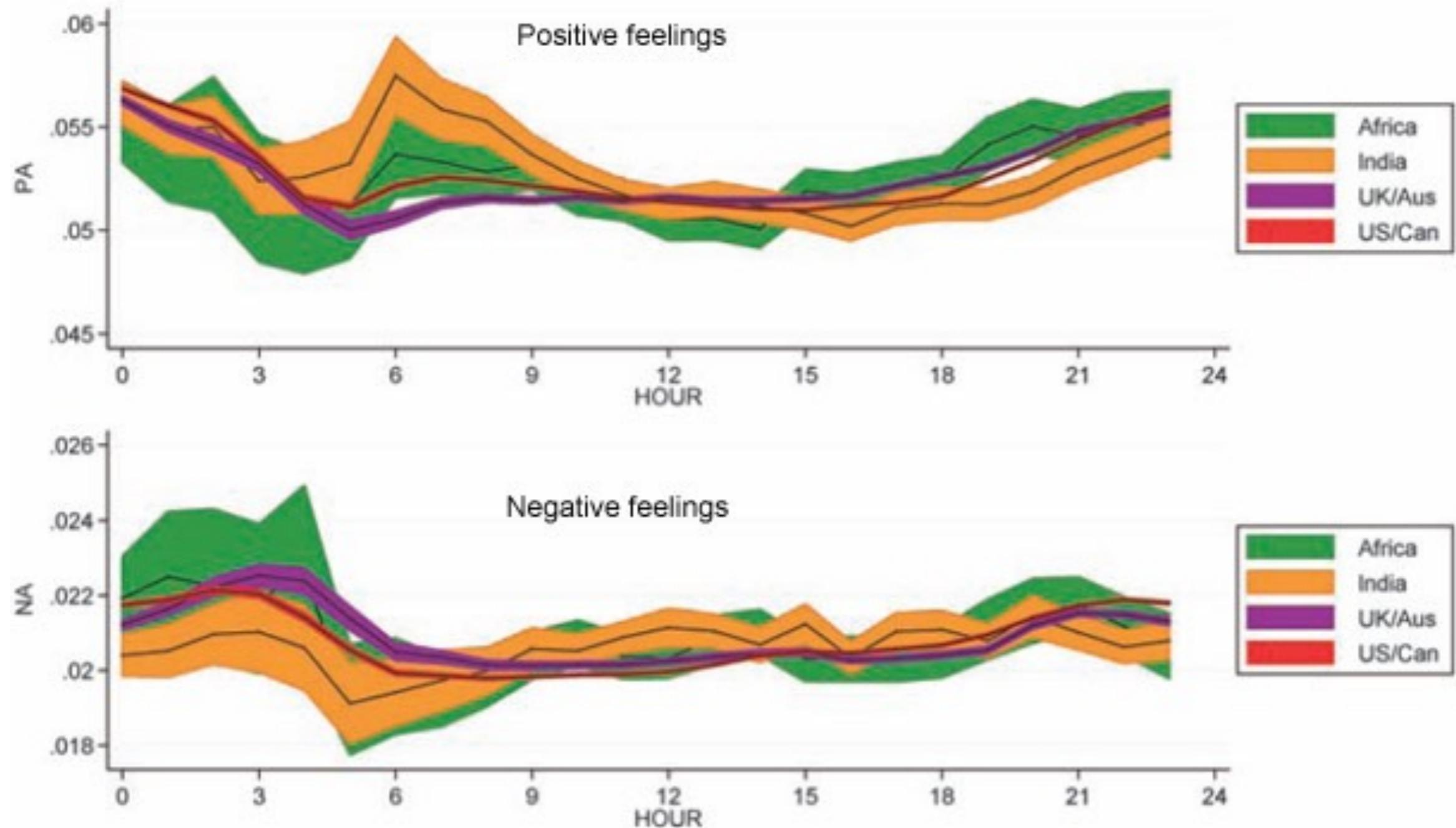
- ▶ In contrast to survey/self-report
- ▶ A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- ▶ Easy to access and aggregate **a lot** of data
- ▶ thus **a lot** of information

# Mood



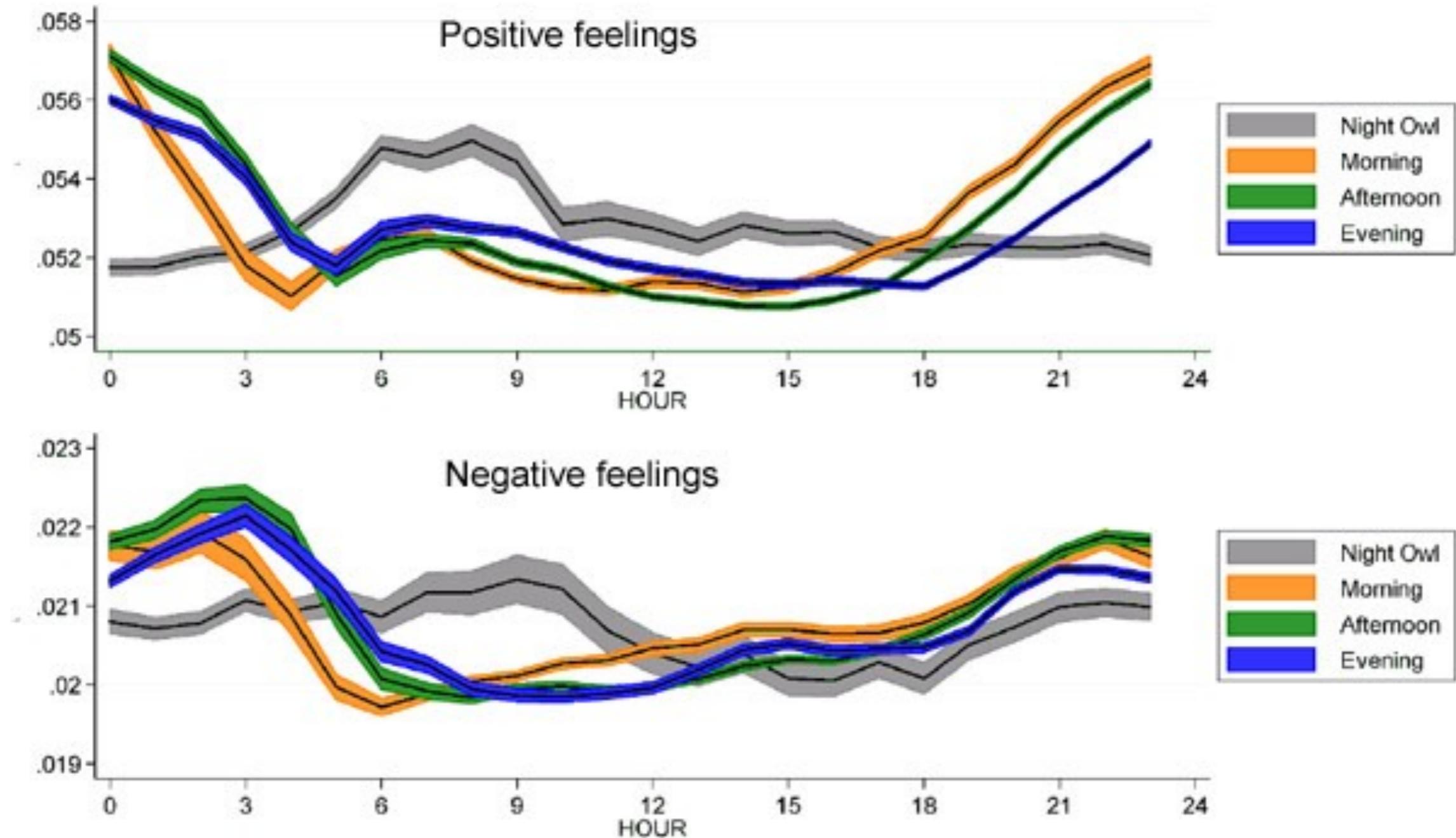
Source: Golder & Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures" Science 2011

# Mood



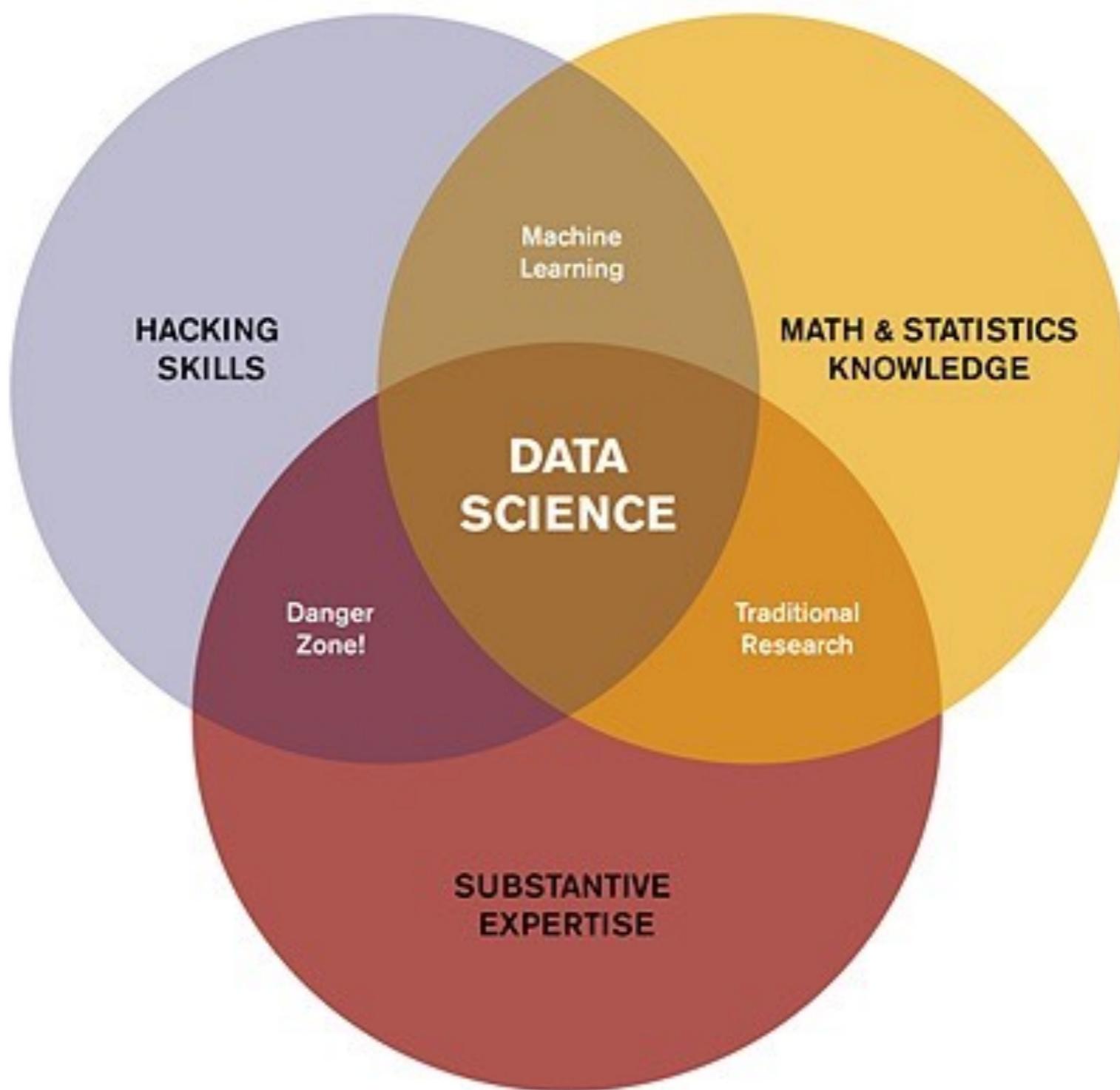
Source: Golder & Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures" Science 2011

# Mood



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

# Data Science

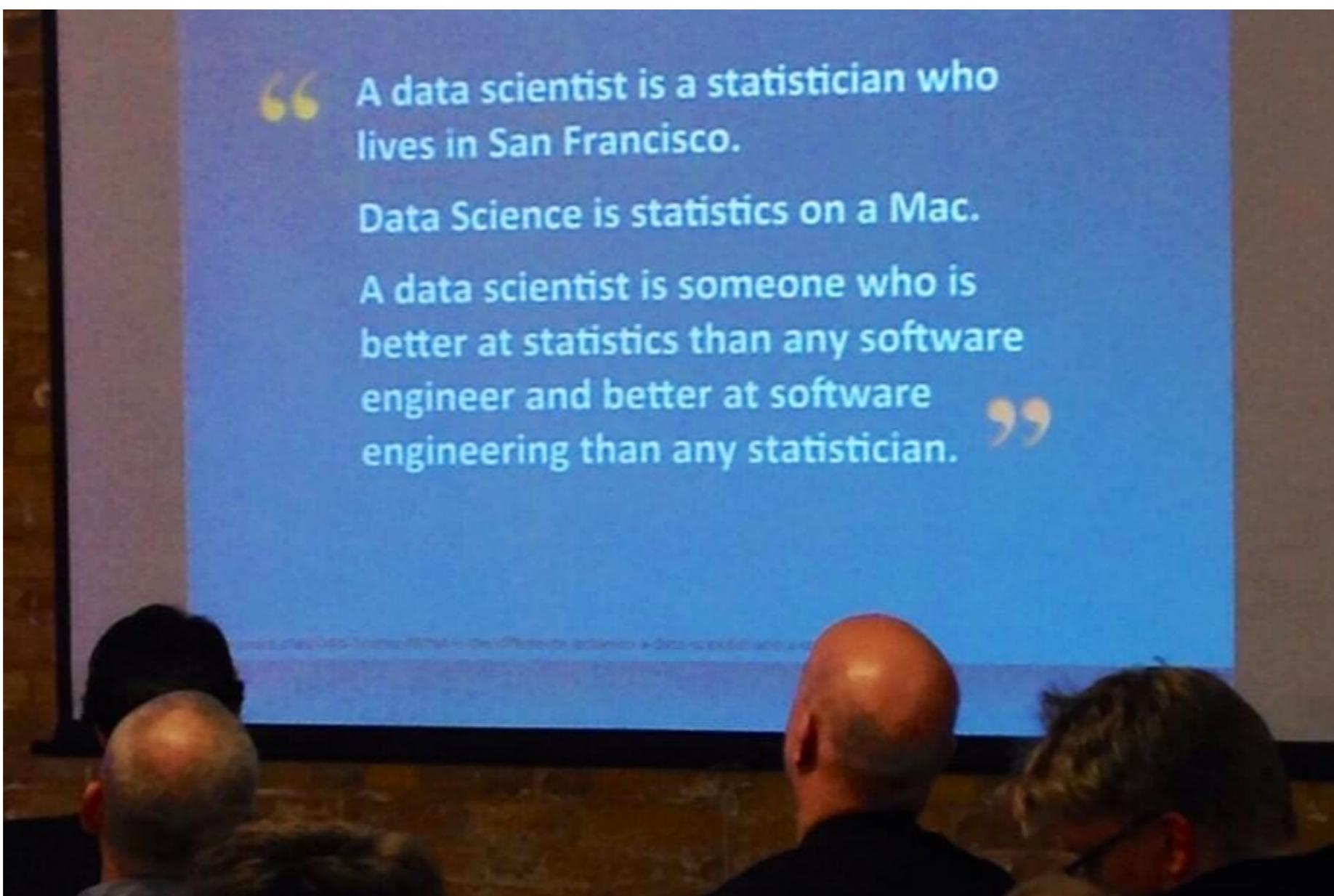


# Data Science

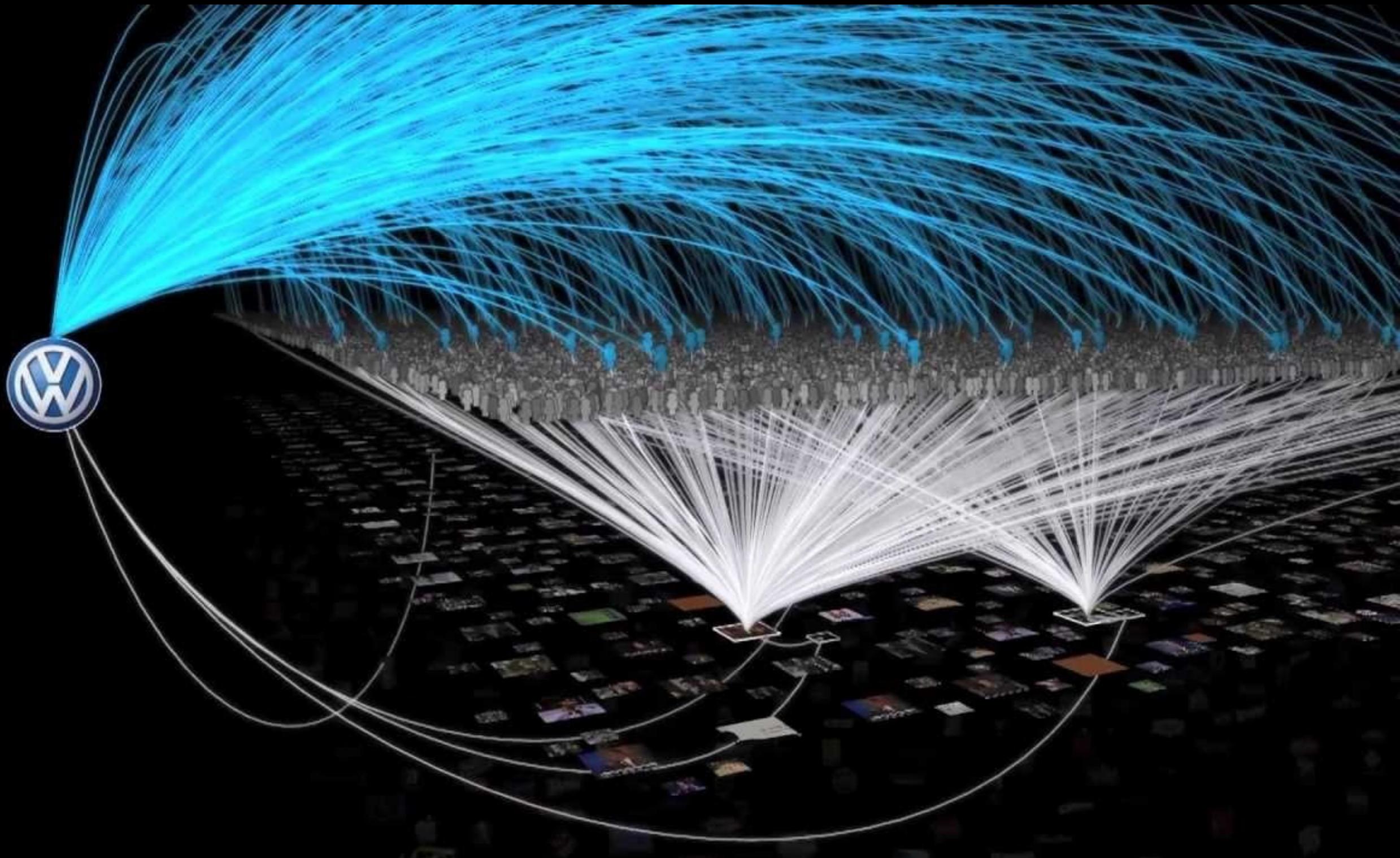
- ▶ is the **practice** of:
  - asking question (formulating hypothesis)
  - finding and collecting the data needed  
(often big data)
  - performing statistical and/or predictive analytics  
(often machine learning)
  - discovering important information and/or insights

# Data Science

- the infamous definition:



# Marketing



Source: Twitter Ads [https://www.youtube.com/watch?v=K8KJWoNk\\_Rg](https://www.youtube.com/watch?v=K8KJWoNk_Rg)

# User Profiling



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



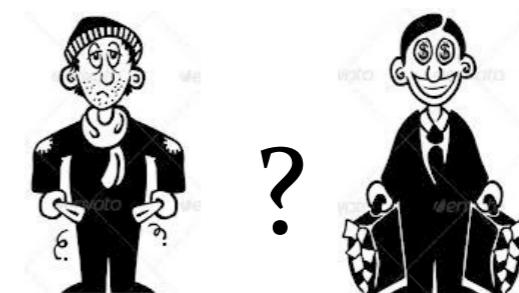
# User Profiling



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



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



# User Profiling



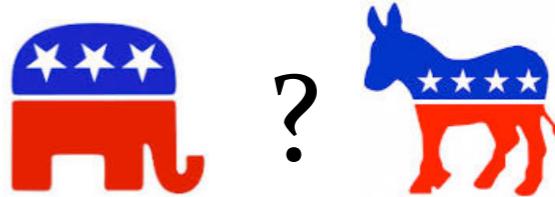
Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping



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



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



# User Profiling



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



?



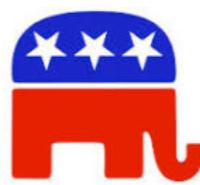
Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots



?



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



?



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



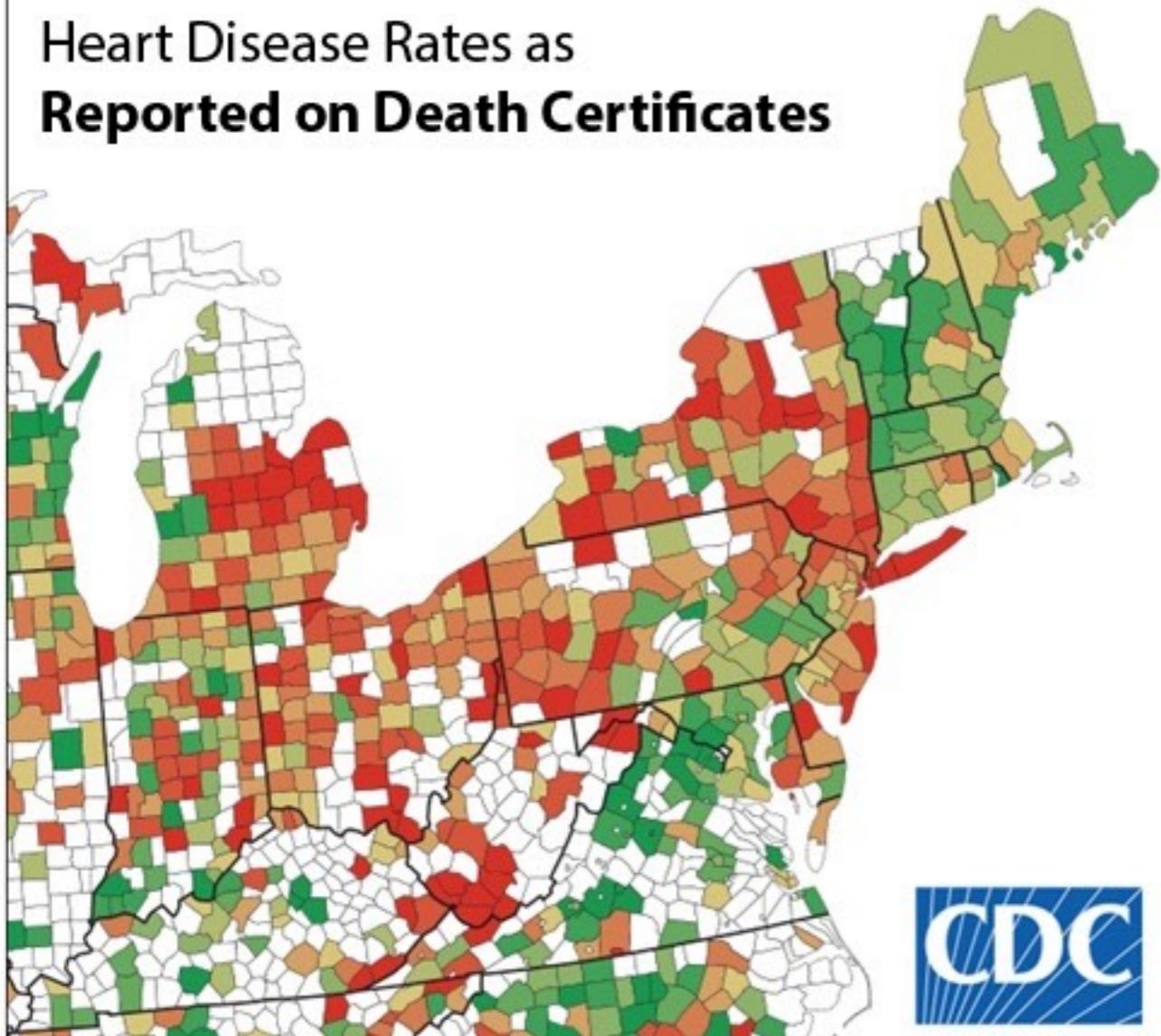
?



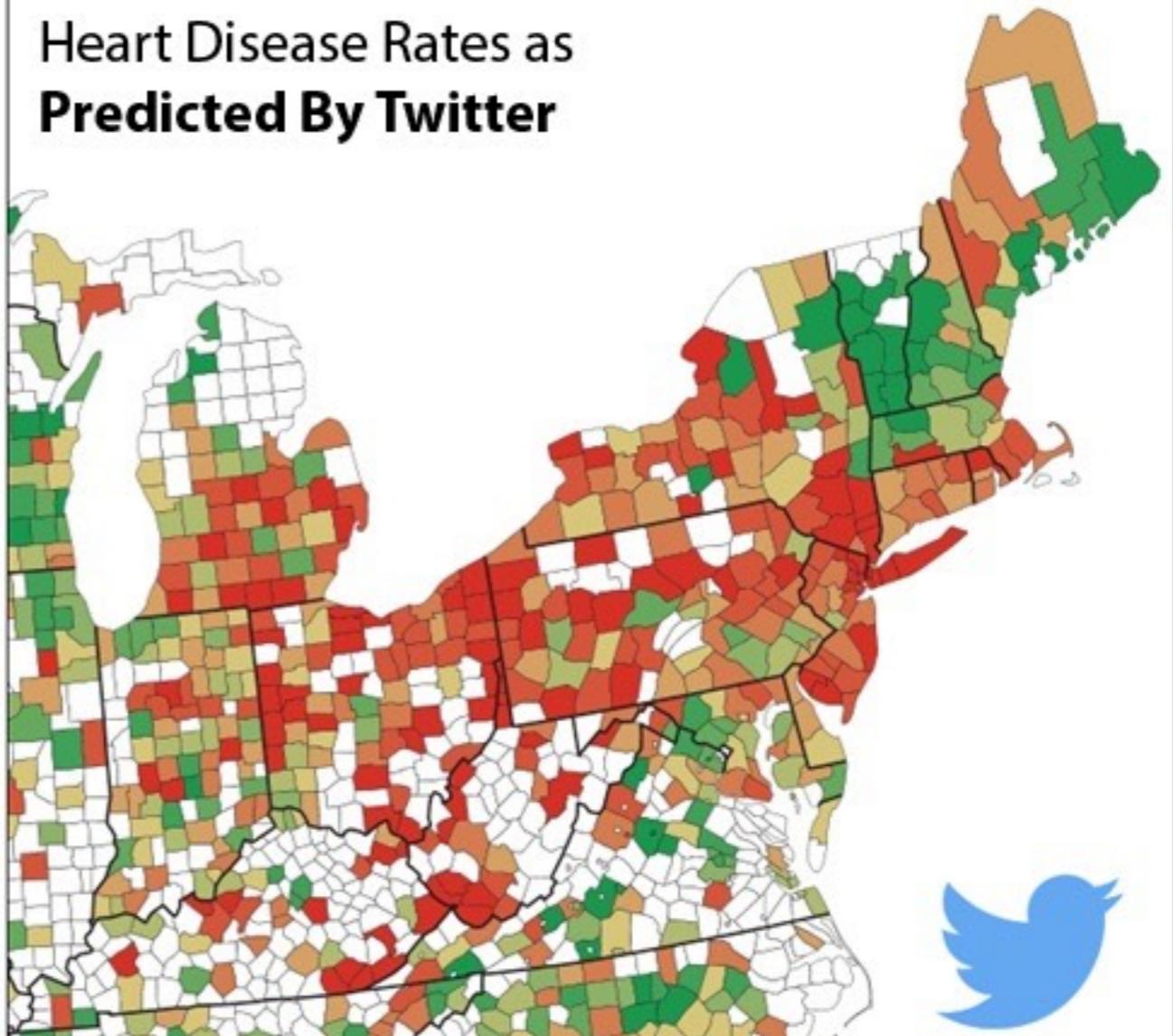
Source: Volkova, Van Durme, Yarowsky, Bachrach  
"Tutorial on Social Media Predictive Analytics" NAACL 2015

# Health

**Heart Disease Rates as  
Reported on Death Certificates**



**Heart Disease Rates as  
Predicted By Twitter**



# Health

Hostility,  
Aggression

A word cloud centered around negative language, including words like 'fuck', 'shit', 'bitch', 'idiot', 'bitches', 'annoying', 'bullshit', 'stupid', 'retarded', 'pissed', 'hate', 'kidding', and 'shit'. The word 'fuck' is the largest and most prominent.

$r = .27$

Hate,  
Interpersonal  
Tension

A word cloud centered around negative emotions and actions, including 'grr', 'passion', 'grrr', 'pit', 'absolutely', 'officially', 'burning', 'despise', 'hates', 'mention', 'fucking', and 'hating'. The word 'hate' is large and central.

$r = .21$

Boredom,  
Fatigue

A word cloud centered around sleep and relaxation, including 'bed', 'bath', 'goodnight', 'tired', 'curl', 'sleepy', 'outta', 'ready', 'crawl', 'layin', 'exhausted', 'shower', and 'cuddle'. The word 'sleep' is the largest and most central.

$r = .20$

A word cloud centered around professional and educational activities, including 'group', 'leadership', 'attend', 'conference', 'council', 'board', 'meeting', 'meetings', 'youth', 'staff', 'student', 'center', 'members', and 'convention'. The word 'conference' is the largest.

Skilled  
Occupations

$$r = -.17$$

A word cloud centered around positive experiences and feelings, including 'fabulous', 'hope', 'safe', 'fantastic', 'holiday', 'enjoyed', 'wonderful', 'hopes', 'weekend', 'peeps', 'enjoy', 'great', 'tgif', 'awsome', and 'hoped'. The word 'weekend' is the largest.

Positive  
Experiences

$$r = -.15$$

A word cloud centered around resilience and personal growth, including 'power', 'strong', 'overcome', 'struggles', 'strength', 'courage', 'challenge', 'greater', 'peace', 'obstacles', 'faith', 'trials', 'stronger', and 'endure'. The word 'overcome' is the largest.

Optimism

$$r = -.13$$

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

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

Samsung to launch Galaxy S6 in March

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

The diagram shows three sentences with their entities highlighted and categorized:

- India** and **Australia** are categorized as **sportsteam**.
- Sydney** is categorized as **geo-loc**.
- Samsung** is categorized as **company**.
- Galaxy S6** is categorized as **product**.
- Suits** and **Brooklyn Nine-Nine** are categorized as **tvshow**.

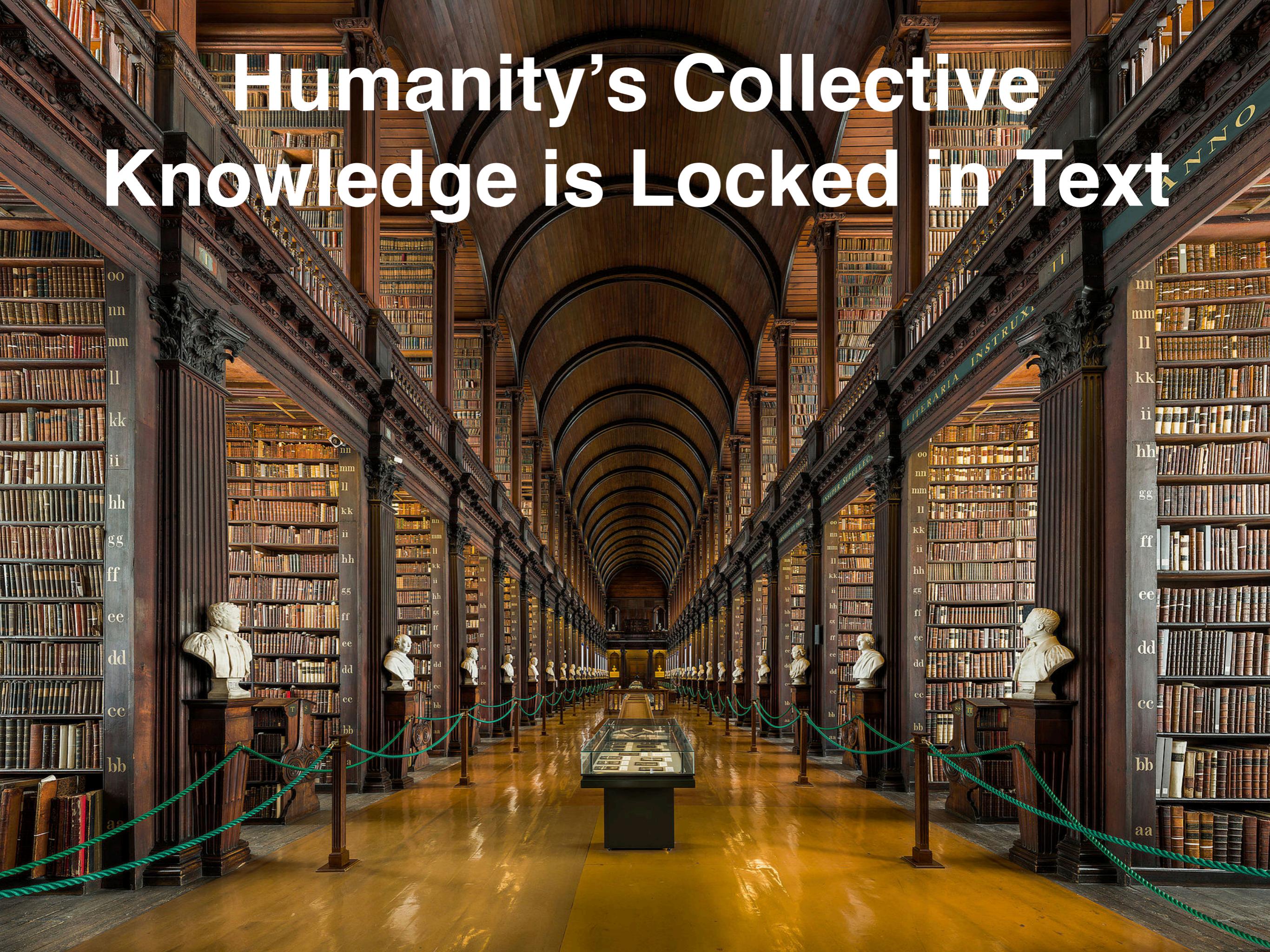
Tim Baldwin, **Marie-Catherine de Marneffe**, Bo Han, Young-Bum Kim, **Ritter, Wei Xu**

Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition

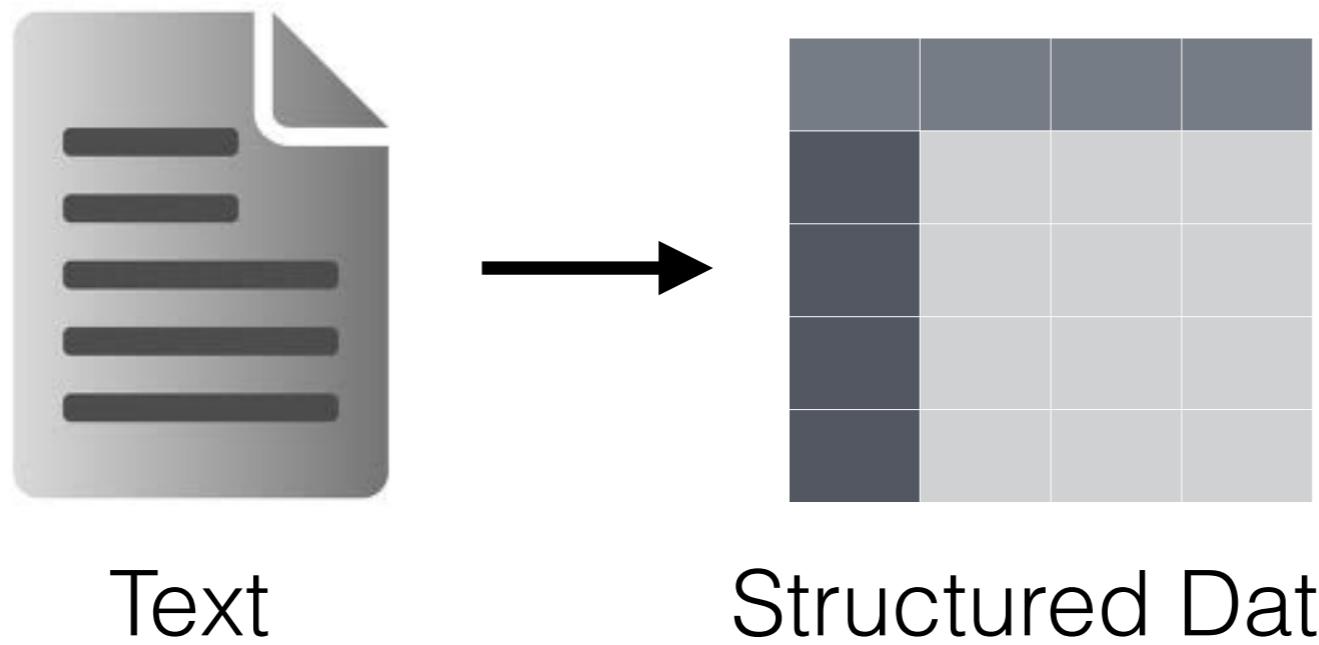
# Machine Translation

The screenshot shows the Google Translate interface. At the top, there's a navigation bar with the Google logo, a grid icon, a bell icon, and a user profile picture. Below it, the word "Translate" is written in red, with a "Turn off instant translation" link and a star icon next to it. The main area has two language selection bars: one for the source language (English) and one for the target language (German). Between them is a double arrow icon. The source text "To the airport, please." is in the English input field, and the translated text "Bis zum Flughafen, bitte." is in the German output field. Both fields have small icons for microphone, speaker, keyboard, and a close button. The German output field also includes a star icon, a copy icon, a speaker icon, a left arrow icon, and a pencil icon.

# Humanity's Collective Knowledge is Locked in Text



# Information Extraction



# 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

Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis. (2015)  
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TAC 2014  
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" TAC 2014

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. "Paraphrasing for Style" In COLING (2012) BUCC (2013)

# Question Answering

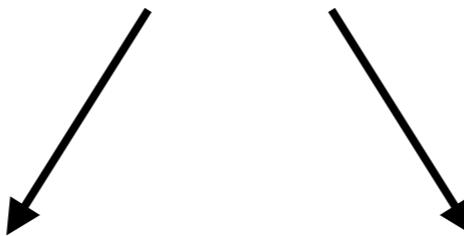
Who is the CEO stepping down from Boeing?

*... 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 ...*



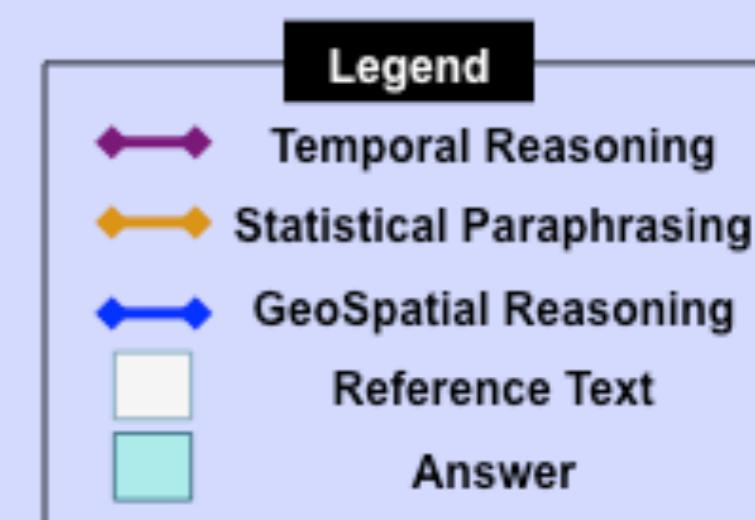
# Watson leverages multiple algorithms to perform deeper analysis

## [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



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

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



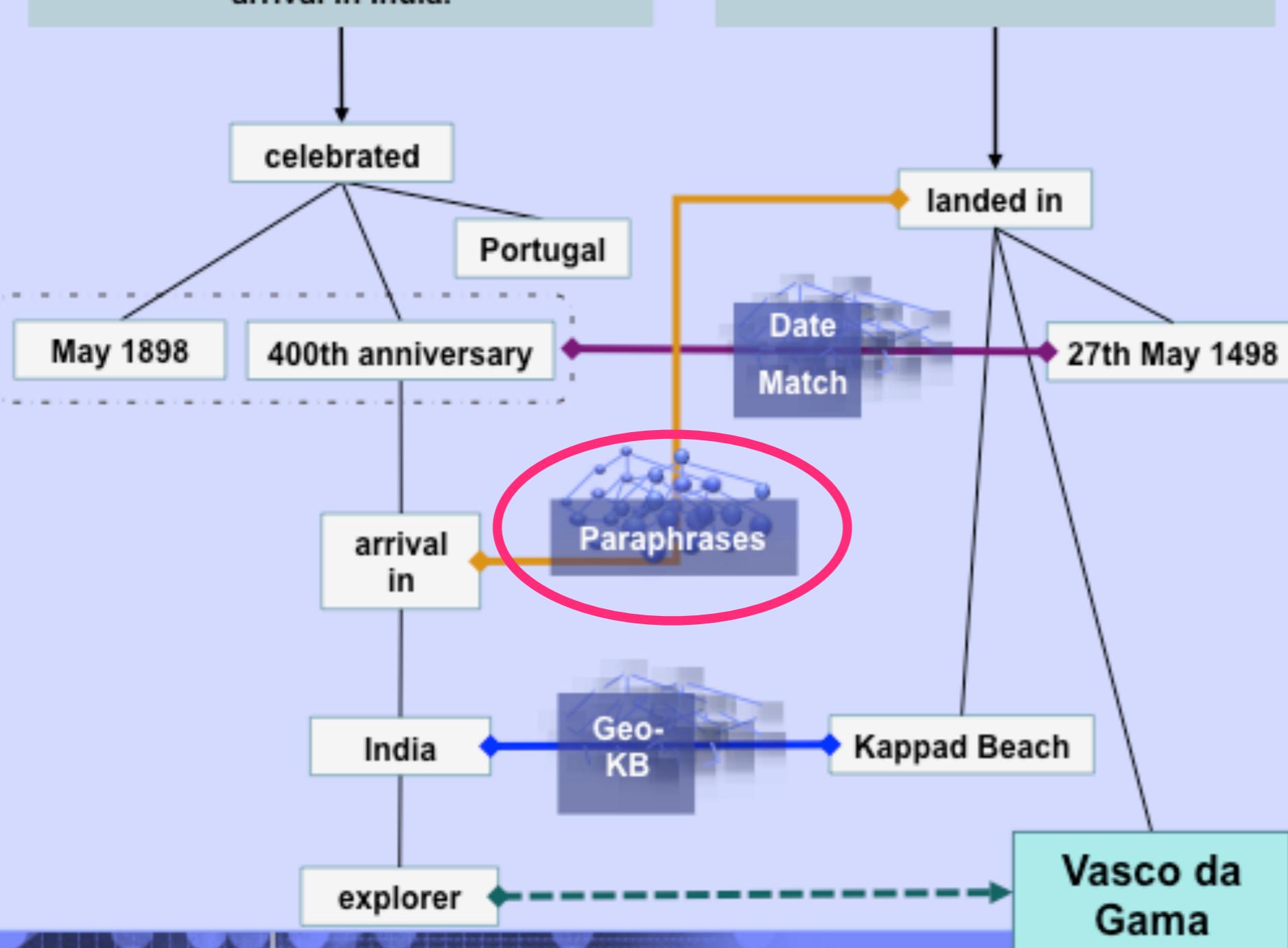
# Watson leverages multiple algorithms to perform deeper analysis

## [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



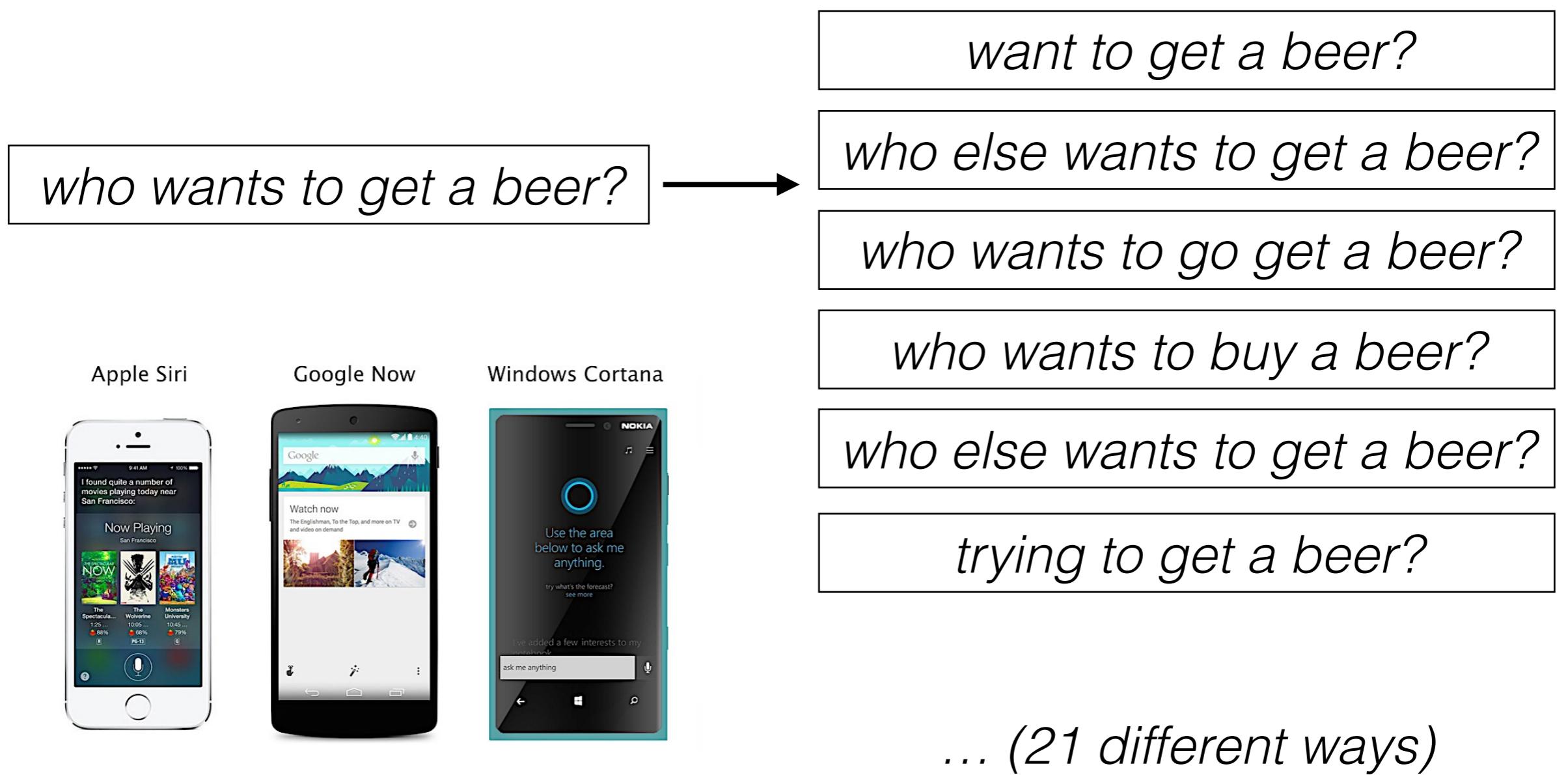
## Legend

- 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





# 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

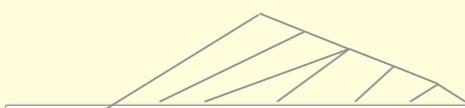
Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my **mouse**.



## Parsing



I can see Alcatraz from the window!

## Machine translation (MT)

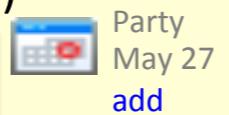
第13届上海国际电影节开幕...



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 (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 programming 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!*





# Stylistic Language Generation

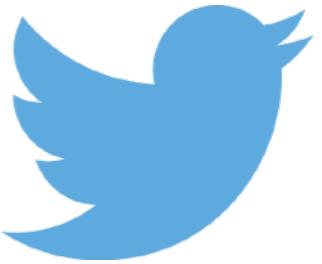
- Data and code:

<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



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 board
- Office hour — Mondays 4-5pm (Dreese 595)
  - No office hours on the 22nd

# Piazza Discussion Board

The screenshot shows the Piazza platform interface for a class named "CSE 5539 AU2017 (35985)". The left sidebar lists various course modules, with "Modules" currently selected. The main content area displays a feed of posts. At the top of the feed is a pinned post titled "Search for Teammates!". Below it are three posts from yesterday: "Introduce Piazza to your stu...", "Get familiar with Piazza", and "Tips & Tricks for a successful Piazza". A "Welcome to Piazza!" message follows. On the right side, there are sections for "Read tips and tricks for a successful Piazza", "Enroll your students" (with a text input field containing "john@email.com, smith@email.com" and a "Enroll Students" button), and "Student Enrollment" (showing "0 enrolled").

CSE 5539 AU2017 (35985) > Modules > Piazza

Autumn 2017

Home

Assignments

Grades

People

**Modules**

Files

Collaborations

Chat

Announcements

Syllabus

Conferences

Discussions

Outcomes

Quizzes

Pages

LockDown Browser

PIAZZA CSE 5539 AU2017 (35985) ▾ Q & A Resources Statistics Manage Class

polls hw1 hw2 hw3 hw4

Unread Updated Unresolved Following

New Post Search or add a post...

PINNED

■ Private Search for Teammates! 8/21/17

YESTERDAY

■ Private Introduce Piazza to your stu... 11:46PM

■ Private Get familiar with Piazza 11:46PM

■ Private Tips & Tricks for a successf... 11:46PM

Welcome to Piazza!

Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together thi

**Read tips and tricks for a successful Piazza**

**Private** Tips & Tricks for a successful class

read now

**Enroll your students**

Paste email addresses below in any format. Or visit Manage Class page to upload yo  
Class Signup Link.

↳ Each will receive a welcome email.

john@email.com, smith@email.com

Enroll Students

**Student Enrollment**

0 enrolled

**Are there TAs/other instructors in your cou**

# By Next Class:

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

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 for research. This page gives an overview of prominent research findings and introduces 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. Her research interests lie at the intersection of machine learning, natural language processing, and social media. She holds a Ph.D. from the University of Pennsylvania. Prior to joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the [ACL 2017](#), serving as a workshop co-chair for [ACL 2017](#), an area chair for [EMNLP 2016](#) and the public relations chair for [NAACL 2016](#).

**Homework**

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

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

**socialmedia-class.org**