Computational Ethics for NLP

Information Retrieval and Natural Language Processing

[Prof. Yulia Tsvetkov](http://www.cs.cmu.edu/~ytsvetko/)
http://www.cs.cmu.edu/~ytsvetko/
http://demo.clab.cs.cmu.edu/ethical_nlp2020/

Computational Ethics for NLP

The common misconception is that
language has to do with **words** and what they mean.

It doesn't.

It has to do with **people** and what ***they*** mean.

Herbert H. Clark & Michael F. Schober, 1992

Decisions we make about our data, methods, and
tools are tied up with their impact on people and societies.

The Belmont Report
Three Basic Ethical Principles

1. Respect for Persons

- Individuals should be treated as autonomous agents
  - "Informed Consent"
- Persons with diminished autonomy are entitled to protection

The Belmont Report
Three Basic Ethical Principles

2. Benificence

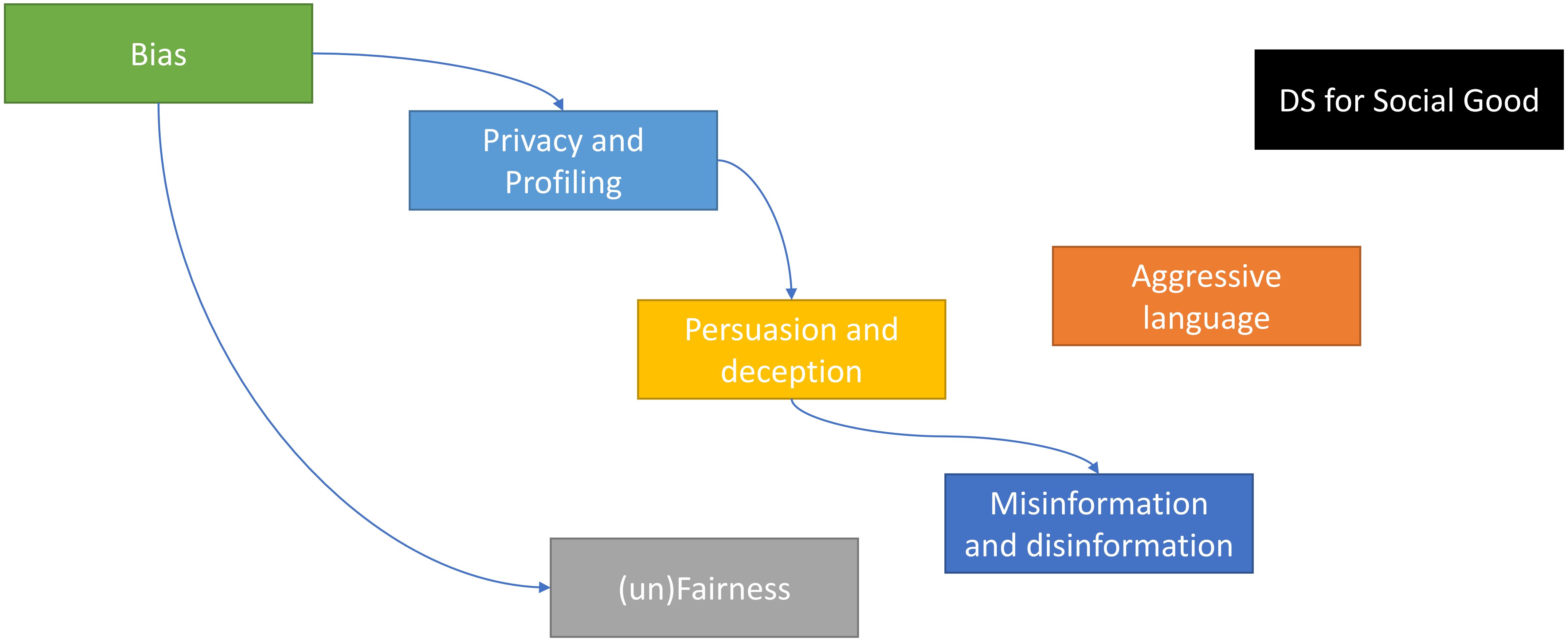
- Do no harm
- Maximize possible benefits and minimize possible harms.

The Belmont Report
Three Basic Ethical Principles

3. Justice

- Who ought to receive the benefits of research and bear its burdens?
  - Fair procedures and outcomes in the selection of research subjects
  - Advances should benefit all

Today’s lecture



Psychological perspective on cognitive bias

Biases inevitably form because of the innate tendency of the human mind to:

- Categorize the world to simplify processing
- Store learned information in mental representations (called schemas)
- Automatically and unconsciously activate stored information whenever one encounters a category member

Cognitive bias is a systematic pattern of deviation from rationality in judgement

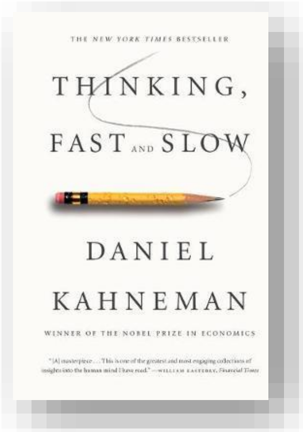
Common biases that affect how we make decisions

- **confirmation bias**: paying more attention to information that reinforces previously held beliefs and ignoring evidence to the contrary
- **ingroup favoritism**: when one favors in-group members over out-group members
- **group attribution error**: when one generalizes about a group based on a group of representatives
- **halo effect**: when overall impression of a person impacts evaluation of their specific traits
- **just-world hypothesis**: when one protects a desire for a just world by blaming the victims
- etc.

Thinking Fast, Thinking Slow

~10MP

~100 bytes



How Do We Make Decisions

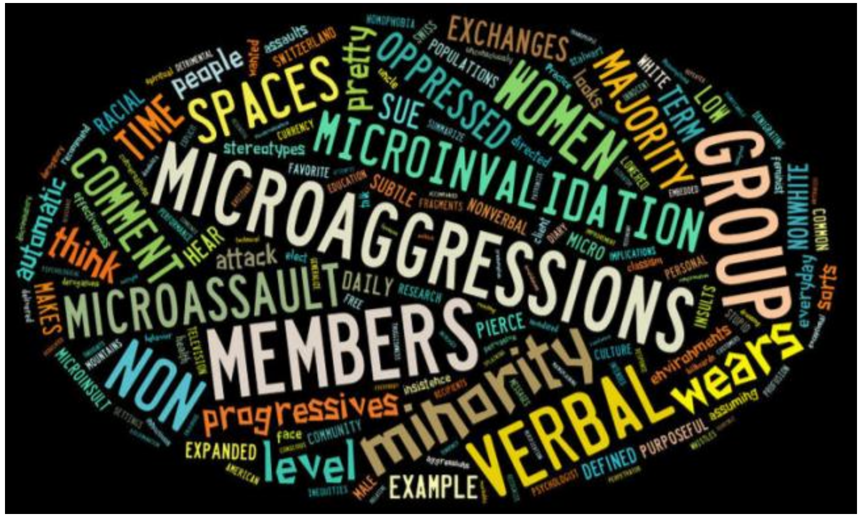
System 1
automatic

System 2
effortful

Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments. Over 95% of cognition is relegated to the System 1 “auto-pilot.”

How Implicit Bias Manifests?

Micro-inequities



Micro-inequities: ephemeral, covert, unintentional, frequently unrecognized events that reinforce power dynamics or perceptions of “difference”

slights, exclusions, slips of the tongue, nonverbal signals, unchecked assumptions, unequal expectations, etc.

Slide credit: Geoff Kaufman

Microaggressions

"A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group"

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- "Girls just **aren't good** at math."
- "Don't you people **like** tamales?"
- "You're too **pretty** to be gay."

Bias in machine learning

- Bias of an estimator
  - the difference between this estimator's expected value and the true value of the parameter being estimated
- Inductive bias
  - assumptions made by the model to learn the target function and to generalize beyond training data

Discussion

- User-generated content represents “real world data”.

- Is it wrong to build models replicating real world data?

Data biases vs Ethical biases

Image search

- June 2017: image search query “Doctor”

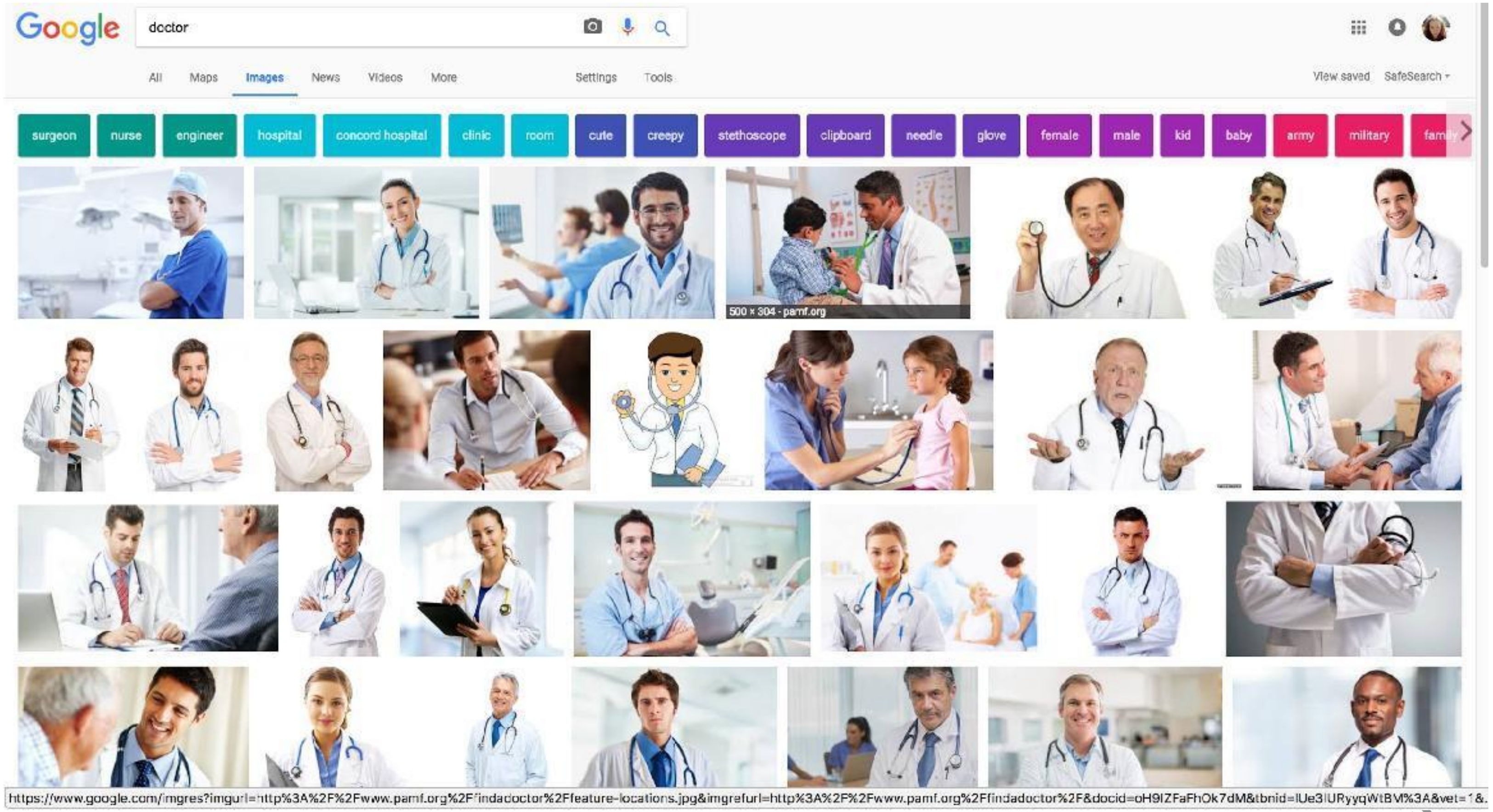


Image search

- June 2017: image search query “Nurse”

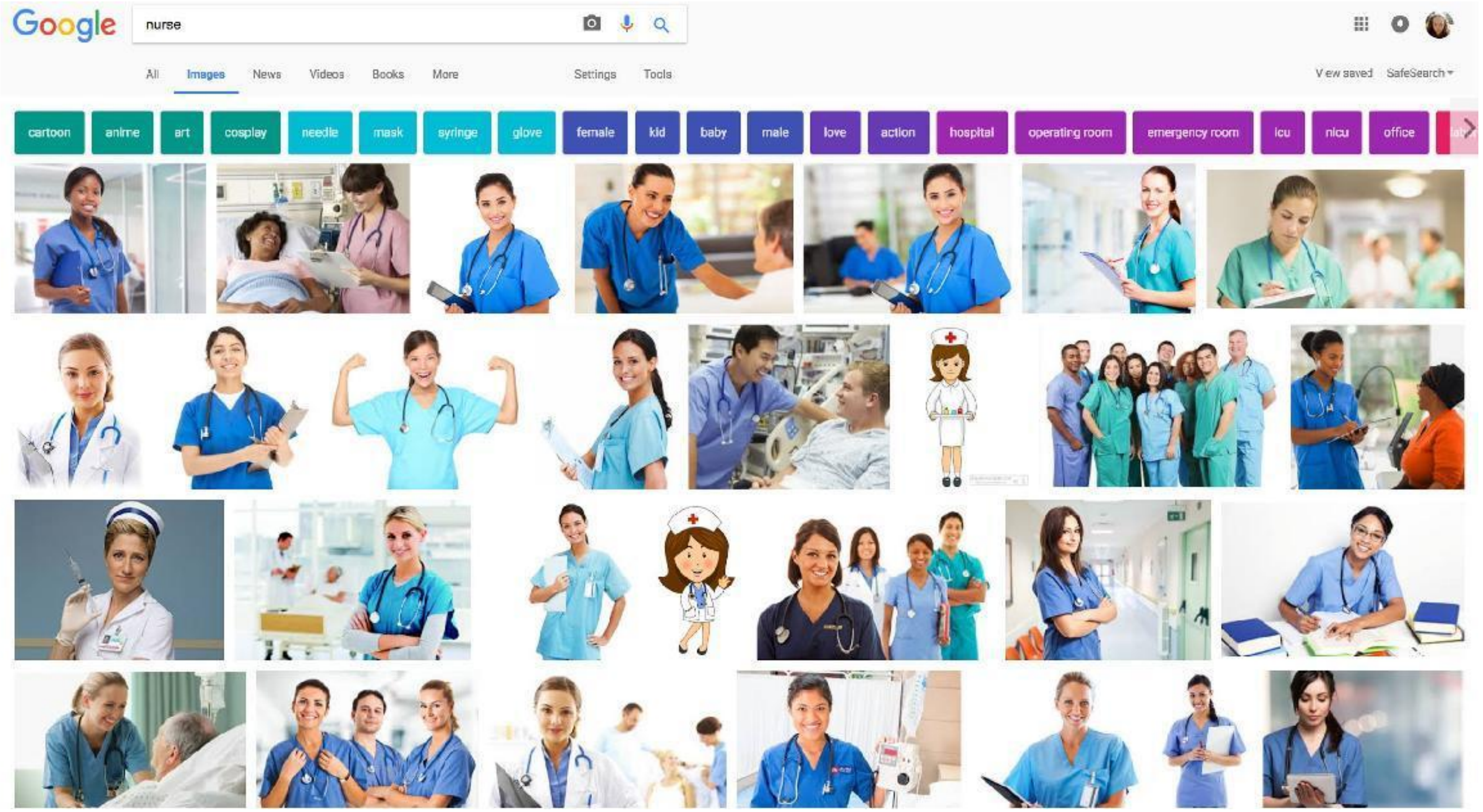


Image search

- June 2017: image search query “CEO”

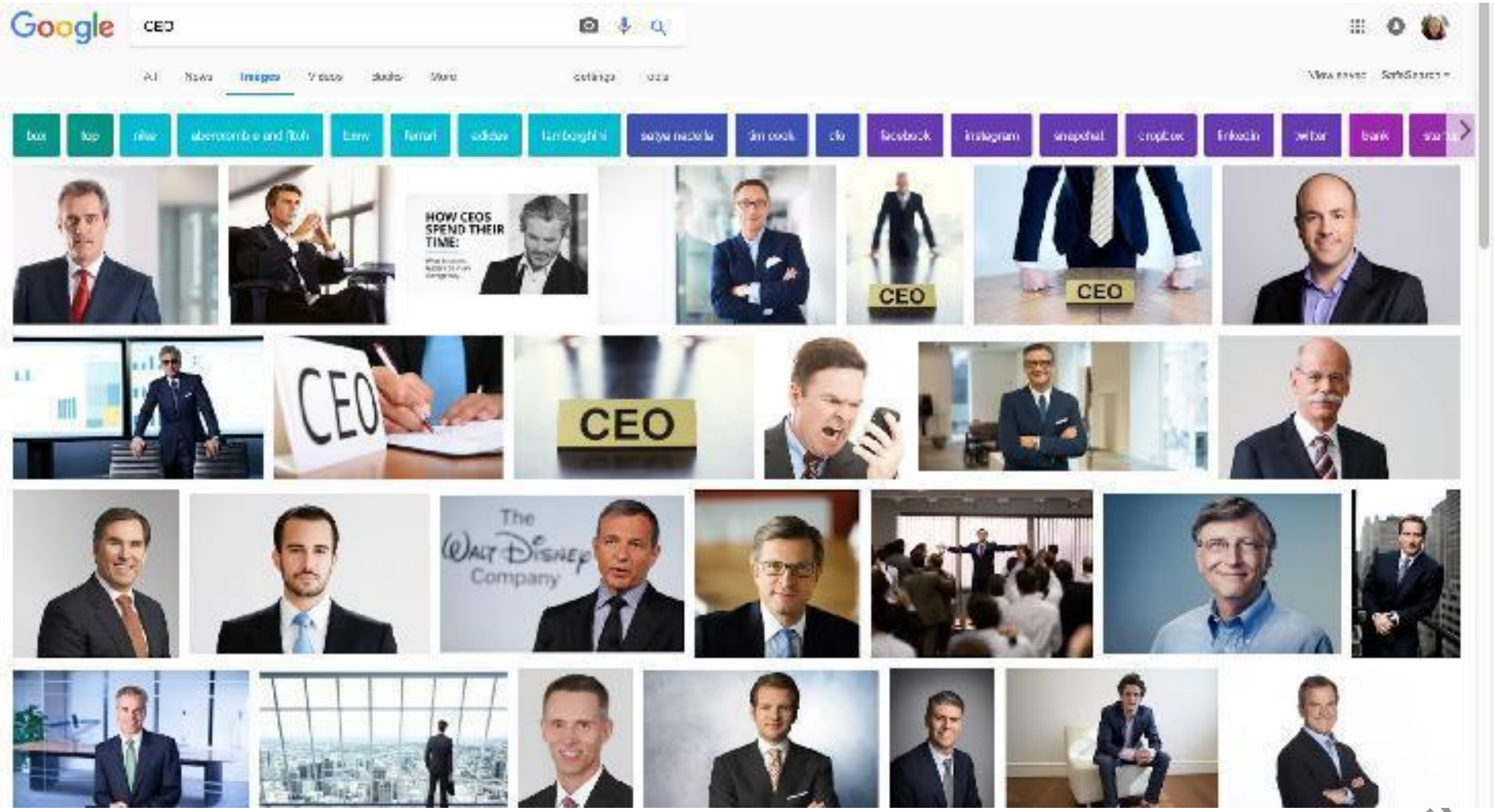
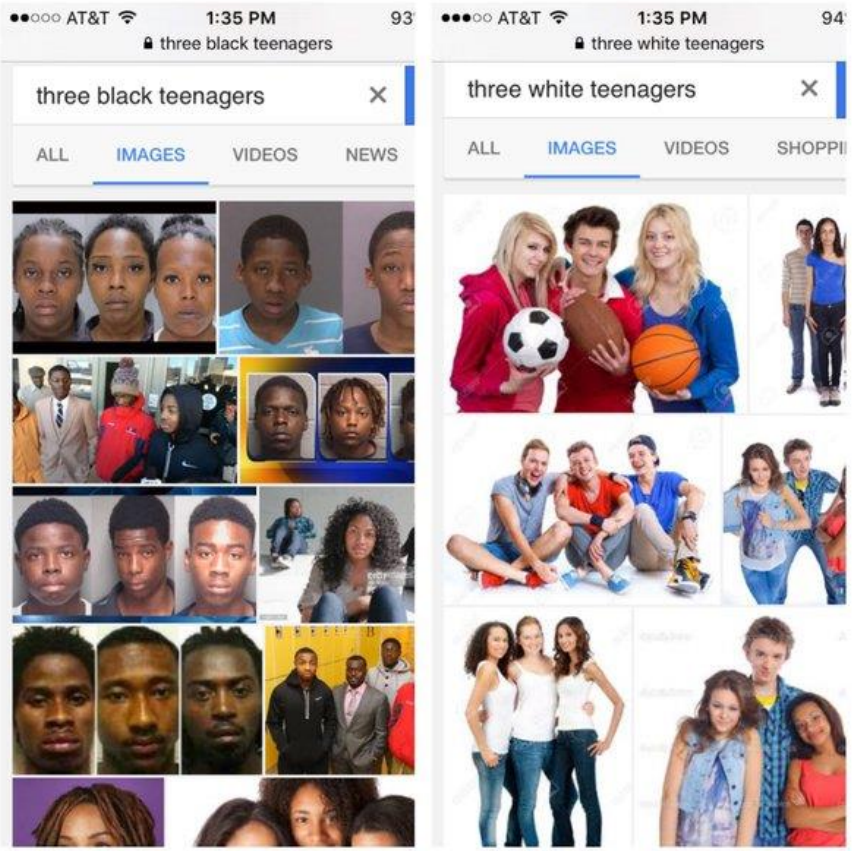
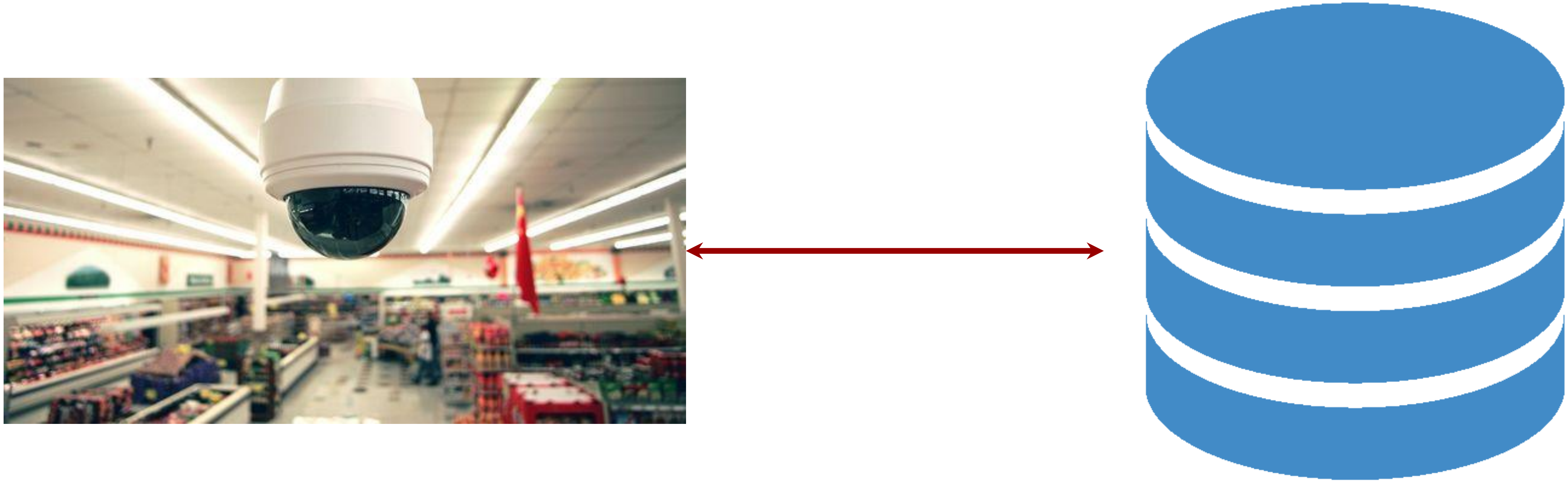


Image search

- June 2016: web search query "three black teenagers"



Privacy and Profiling



- Being seen vs being identified

Three aspects of privacy

- **Territorial privacy**: Public vs private space
- **Personal privacy**: Being seen vs being watched
- **Informational privacy**: Being seen vs being watched; Being seen vs being tracked

(Holvast, 1993, Rosenberg, 1992)

$\boxed{\rightarrow}$ Privacy issues arise due to personalization and aggregation of data

Do People REALLY Care About Protecting Their Privacy?

- **Information privacy paradox**: privacy attitudes vs privacy behaviors (Kokolakis ‘17)
  - Surveys of internet users' attitudes show that users are highly concerned about their privacy and the collection and use of their personal information (TRUSTe, 2014, Pew Research Center, 2014)
  - But easily trade their personal data
    - Revealing personal details to a shopping bot (Spiekermann et al. ‘01)
    - Trading online history for ~7 Euros (Carrascal et al. ‘13)

Dangers in Misusing Private Information

Examples of scenarios how people can be harmed

- Identity fraud with stolen SSN
- Medical records
- Private vs public accounts on social media: “People You May Know“
- Phone number, call history
- Location history
- Profile pictures across communities and social circles

Cambridge
Analytica

What Can We Reveal?

"Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including:

- sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. "

Kosinski M., Stillwell D., and Graepel T. (2013) **Private traits and attributes are predictable from digital records of human behavior.** *PNAS*

What Can We Reveal Without User's Data/Language?

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky (2017) **Writer Profiling Without the Writer's Text.** *SocInfo*

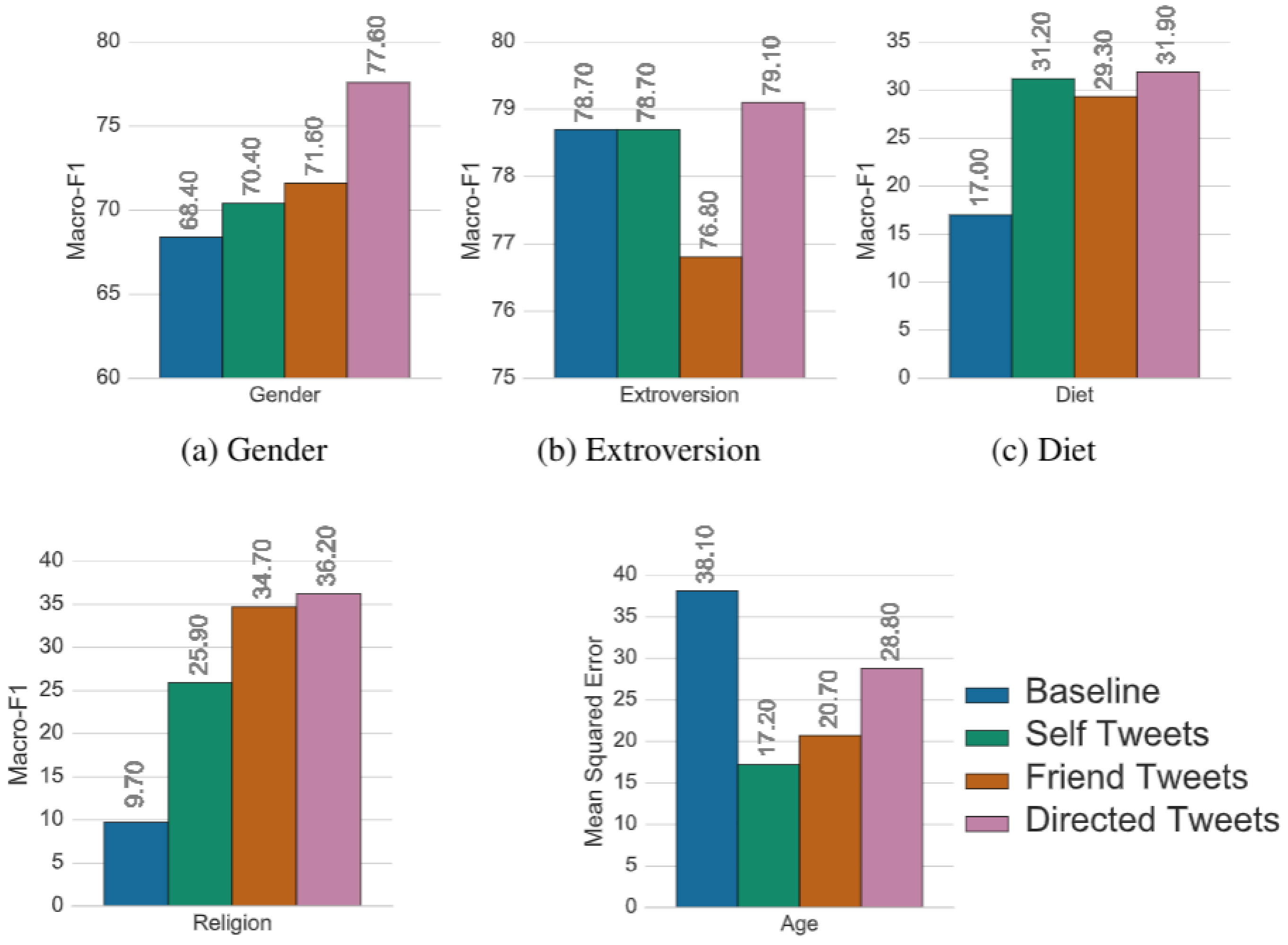
Data

- Self-identified labels plus heuristics on user names plus aggressive filtering

Attribute	# of Tweets	Majority Class	%
Gender	59800	Male	52.5
Religion	19940	Christian	65.8
Extroversion	24576	Introvert	63.0
Diet	9001	Unrestricted	41.0
Age	38134	21.3 (mean)	5.7 (s.d.)

Table 1. Dataset sizes for each demographic attribute and frequencies of the majority classes.

Results

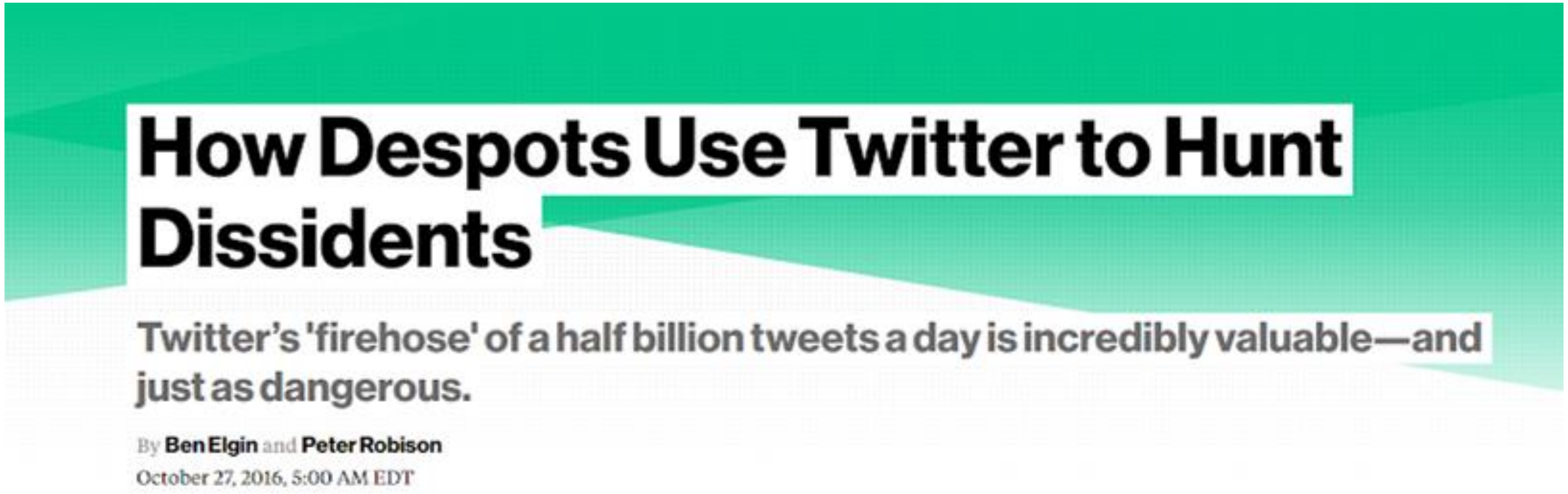


(a) Gender (b) Extroversion (c) Diet

(d) Religion (e) Age

Fig. 2. Predictive accuracy for each attribute, reported as Macro-F1 and Mean Squared Error.

The Dual Use of Profiling Techniques



How Despots Use Twitter to Hunt Dissidents

Twitter's 'firehose' of a half billion tweets a day is incredibly valuable—and just as dangerous.

By Ben Elgin and Peter Robison

October 27, 2016, 5:00 AM EDT

https://www.bloomberg.com/news/articles/2016-10-27/twitter-s-firehose-of-tweets-is-incredibly-valuable-and-just-as-dangerous

Discussion

- Big trend in NLP: Generating **polite answers** and summaries.
- Big trend in IR/NLP: Asking **clarifying questions** in chatbots

Eliza: Weizenbaum (1966)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

...

WHO ELSE IN YOUR FAMILY TAKES CARE 0F YOU?
My father
YOUR FATHER

Ethical implications of ELIZA

- People became deeply emotionally involved with the program
- Weizenbaum's secretary asked him to leave the room when she talked with ELIZA
- When he suggested that he might want to store all the ELIZA conversations for later analysis, **people immediately pointed out the privacy implications**
    - Suggesting that they were having quite private conversations with ELIZA

Barbie Wants to Get to Know Your Child

Barbara Grosz, NYT 2015

"Hey, new question," Barbie said. "Do you have any sisters?"

"Yeah," Tiara said. "I only have one."

"What's something nice that your sister does for you?" Barbie asked.

"She does nothing nice to me," Tiara said tensely.

Barbie forged ahead. "Well, what is the last nice thing your sister did?"

"She helped me with my project — and then she *destroyed* it."

"Oh, yeah, tell me more!" Barbie said, oblivious to Tiara's unhappiness.

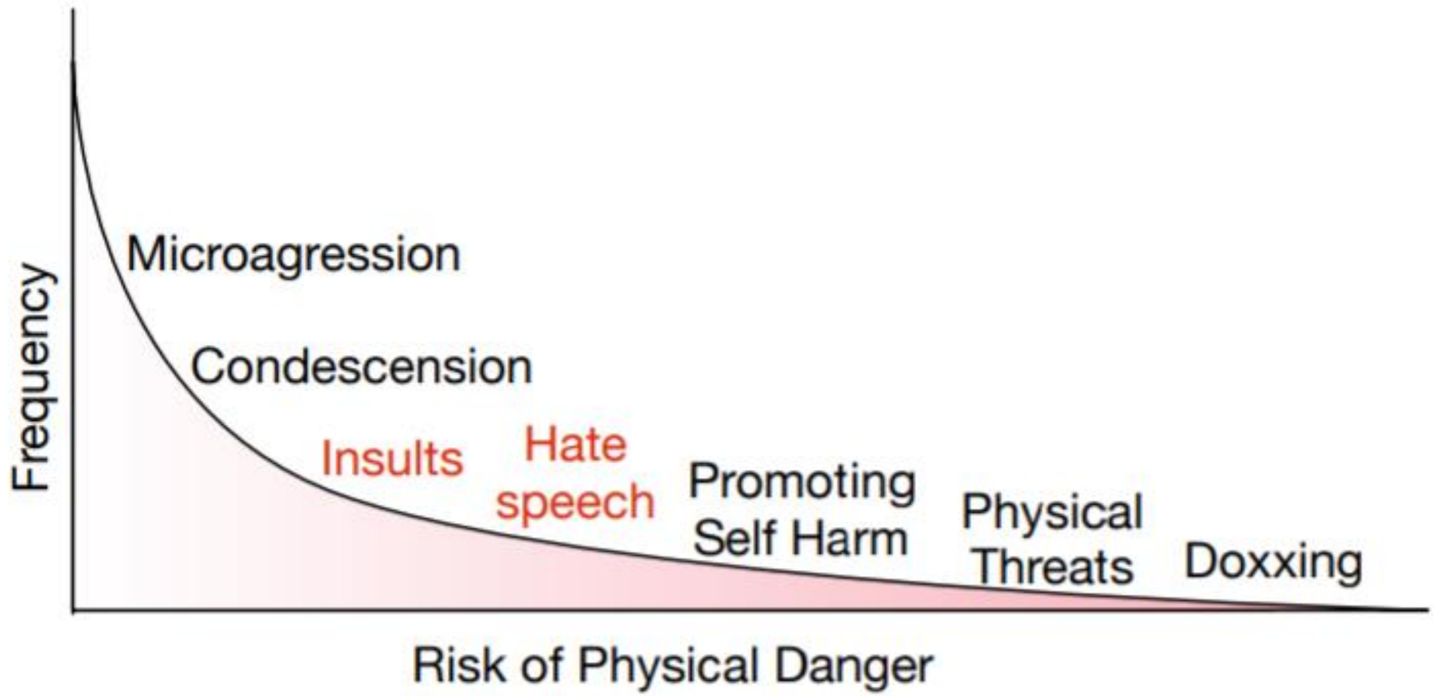
"That's it, Barbie," Tiara said.

"Have you told your sister lately how cool she is?"

"No. She is *not* cool," Tiara said, gritting her teeth.

"You never know, she might appreciate hearing it," Barbie said.

The Spectrum of Toxic Language



Jurgens D., Chandrasekharan E., and Hemphill L. (2019) **A Just and Comprehensive Strategy for Using NLP to Address Online Abuse.** *ACL*

Hate Speech has Many Shades

- Umbrella term: Abuse
- Hate speech
- Offensive language
- Sexist and racist
- Aggression
- Profanity
- Cyberbullying
- Harassment
- Trolling
- Anti-social behavior
- Toxic language
- …

What is Hate Speech?

"any communication that disparages **a person or a group** on the basis of some characteristic such as **race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic**"

(Nockleby, J. Encyclopedia of the American Constitution 2000)

TARGET

What is Hate Speech?

*“language that is used **to expresses hatred** towards a targeted group or is **intended to be derogatory, to humiliate, or to insult** the members of the group”*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

INTENT

What is Hate Speech?

*"language that **threatens** or **incites violence**"*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

EFFECT

What is Hate Speech?

"any offense motivated, in whole or in a part,
by the offender's ***bias*** *against an aspect of a group of people"*

(Silva et al., Analyzing the Targets of Hate
in Online Social Media, *ICWSM* 2016)

THE CAUSE

Who Are Target of Hate Speech?

I **<intensity> <userintent>** <hatetarget>

*"I f*cking hate white people"*

Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

(Silva et al., Analyzing the Targets of Hate in Online Social Media, ICWSM 2016)

Who Are Target of Hate Speech?

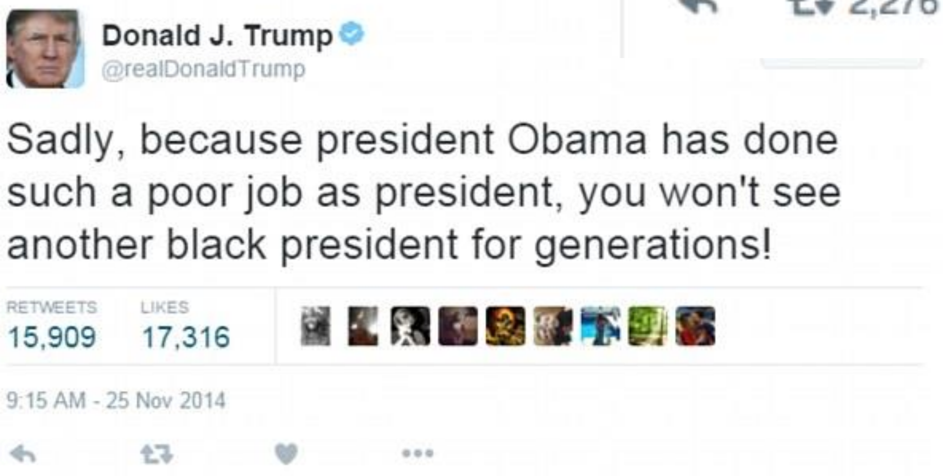
I <intensity> <userintent> **<hatetarget>**

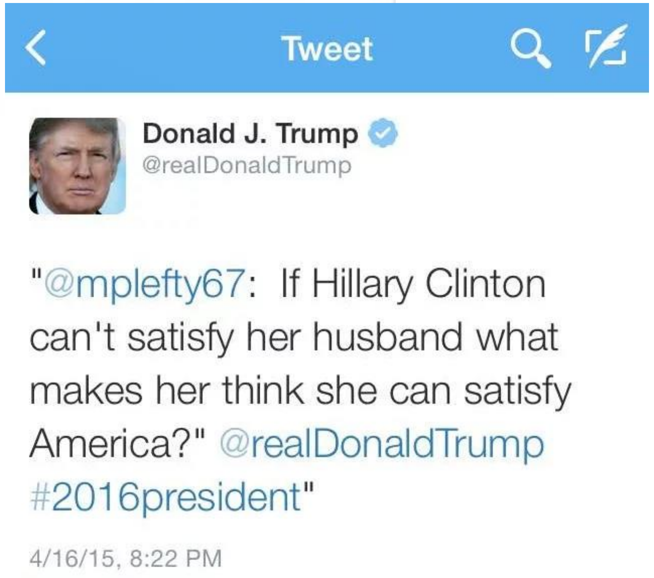
*"I f*cking hate white people"*

Twitter		***Whisper***	
Hate target	**% posts**	**Hate target**	**% posts**
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

(Silva et al., Analyzing the Targets of Hate in Online Social Media, ICWSM 2016)

Stereotypes in user-generated content





Discussion

- Who should fix this?
  - The researcher/developer?
  - The user of the technology?
  - Paper reviewers?
  - The IRB? The University?
  - Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

Misinformation

- This is a widely studied subject in Communication and Social Sciences.
  - **Misinformation** is false or inaccurate information that is communicated regardless of an intention to deceive
  - **Disinformation** is deliberately misleading or biased information; manipulated narrative or facts
  - **Propaganda** is information spread to make someone or something look bad or good. Propaganda is designed to influence people emotionally.

- Social media platforms amplify this problem.

How to tell the truth in a persuasive manner

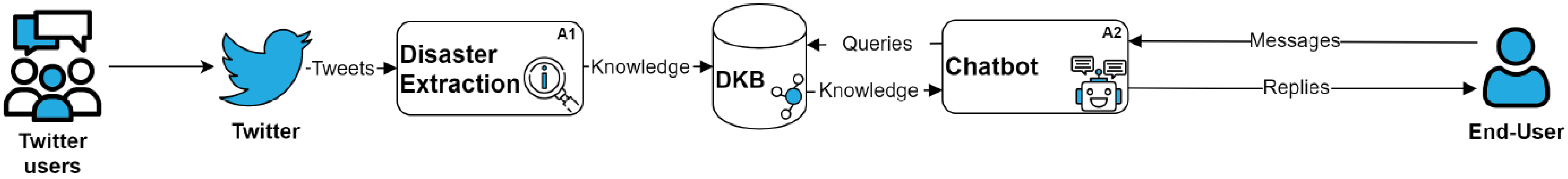
We have … Army, navy and air force Reporting guidelines Press briefings	They have … A war machine Censorship Propaganda
We … Take out Suppress Dig in	They … Destroy Kill Cower in their fox holes
Our men are … Boys Lads	Their men are … Troops Hordes

The Guardian 1990

DS for Social Good

• Education	QA / chatbots
• Psychological counselling	Chatbots
• Disaster response	Twitter / chatbot
• Depression detection	Slack
• Legal tasks	100% recall
• Clinical decision support	NovaMedSearch

Chatbot for disasters



Chatbots for citizen feedback

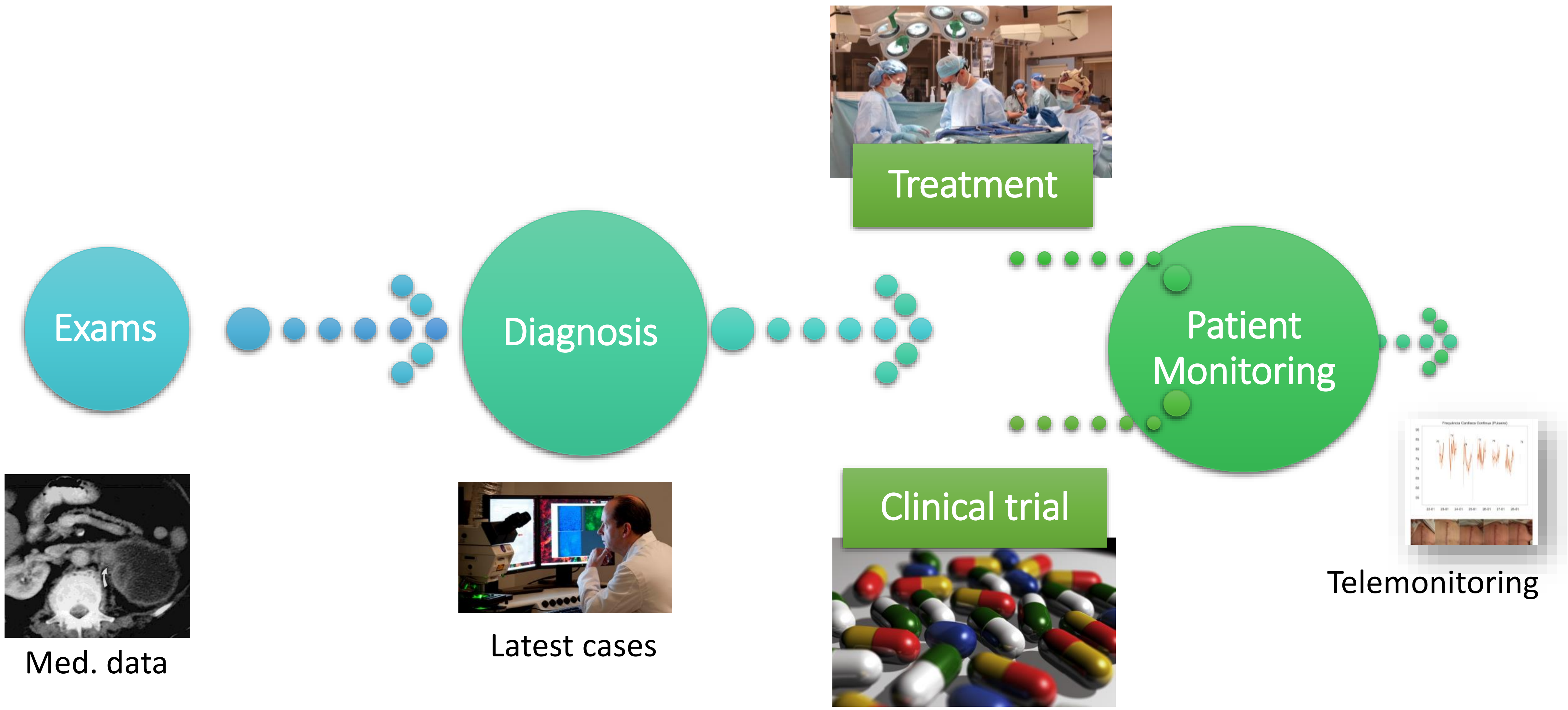
Clinical Workflow

- Most work in Precision Medicine aims to find the treatment based on the **individual’s health records**.

- Data-driven approaches allow the discovery of critical information that can support clinicians in their decisions.

- This support can happen in all stages of the clinical workflow.

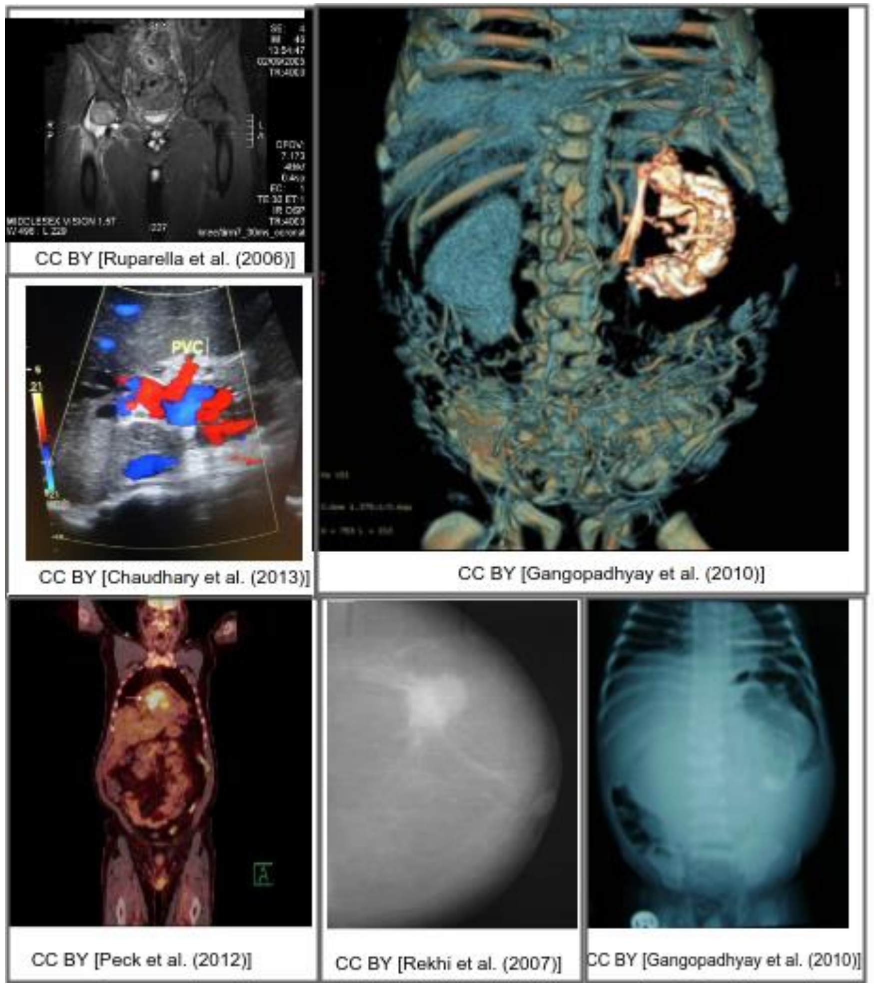
DS for social good



Medical ImageNet

- Aims to create a peta-byte collection of medical images with extensive annotations.

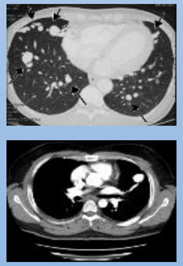
- This resources promises to significantly boost the accuracy of medical diagnosis systems.



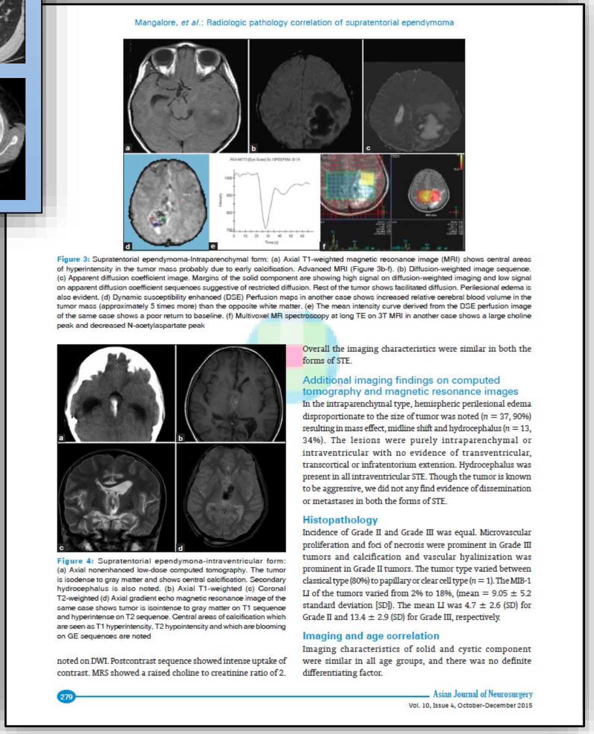
Understand patient data

Patient clinical history*

A woman in **her mid-30s**.
CT scan revealed a cystic mass in the right lower lobe.
She later developed **right arm weakness** and **aphasia**.
She was treated, but **four years later** suffered **another stroke**.
Follow-up CT scan showed multiple new cystic lesions.



Electronic Health Records



Mangalore, *et al.*: Radiologic pathology correlation of supratentorial ependymoma

Figure 3: Supratentorial ependymoma-Intraparenchymal form: (a) Axial T1-weighted magnetic resonance image (MRI) shows central areas of hyperintensity in the tumor mass probably due to early calcification. Advanced MRI (Figure 3b-f). (b) Diffusion-weighted image sequence. (c) Apparent diffusion coefficient image. Margins of the solid component are showing high signal on diffusion-weighted imaging and low signal on apparent diffusion coefficient sequences suggestive of restricted diffusion. Rest of the tumor shows facilitated diffusion. Perilesional edema is also evident. (d) Dynamic susceptibility enhanced (DSE) Perfusion maps in another case shows increased relative cerebral blood volume in the tumor mass (approximately 5 times more) than the opposite white matter. (e) The mean intensity curve derived from the DSE perfusion image of the same case shows a poor return to baseline. (f) Multivoxel MR spectroscopy at long TE on 3T MRI in another case shows a large choline peak and decreased N-acetylaspartate peak

Figure 4: Supratentorial ependymona-intraventricular form: (a) Axial nonenhanced low-dose computed tomography. The tumor is isodense to gray matter and shows central calcification. Secondary hydrocephalus is also noted. (b) Axial T1-weighted (c) Coronal T2-weighted (d) Axial gradient echo magnetic resonance image of the same case shows tumor is isointense to gray matter on T1 sequence and hyperintense on T2 sequence. Central areas of calcification which are seen as T1 hyperintensity, T2 hypointensity and which are blooming on GE sequences are noted

noted on DWI. Postcontrast sequence showed intense uptake of contrast. MRS showed a raised choline to creatinine ratio of 2.

Overall the imaging characteristics were similar in both the forms of STE.

Additional imaging findings on computed tomography and magnetic resonance images

In the intraparenchymal type, hemispheric perilesional edema disproportionate to the size of tumor was noted ($n = 37$, 90%) resulting in mass effect, midline shift and hydrocephalus ($n = 13$, 34%). The lesions were purely intraparenchymal or intraventricular with no evidence of transventricular, transcortical or infratentorium extension. Hydrocephalus was present in all intraventricular STE. Though the tumor is known to be aggressive, we did not any find evidence of dissemination or metastases in both the forms of STE.

Histopathology

Incidence of Grade II and Grade III was equal. Microvascular proliferation and foci of necrosis were prominent in Grade III tumors and calcification and vascular hyalinization was prominent in Grade II tumors. The tumor type varied between classical type (80%) to papillary or clear cell type ($n = 1$). The MIB-1 LI of the tumors varied from 2% to 18%, (mean = 9.05 ± 5.2 standard deviation [SD]). The mean LI was 4.7 ± 2.6 (SD) for Grade II and 13.4 ± 2.9 (SD) for Grade III, respectively.

Imaging and age correlation

Imaging characteristics of solid and cystic component were similar in all age groups, and there was no definite differentiating factor.

279

Asian Journal of Neurosurgery
Vol. 10, Issue 4, October-December 2015

* This is a query of the ImageCLEF Medical Case-based Search dataset (it has 28.000 query-document annotations provided by MDs).

Summary

- Computational Ethics
  - Biases
  - Privacy and Profiling
  - Persuasion and deception
  - Aggressive language (hate speech, trolling)
  - Misinformation and disinformation

- DS for Social Good
  - Education
  - Natural disasters
  - Depression
  - Clinical