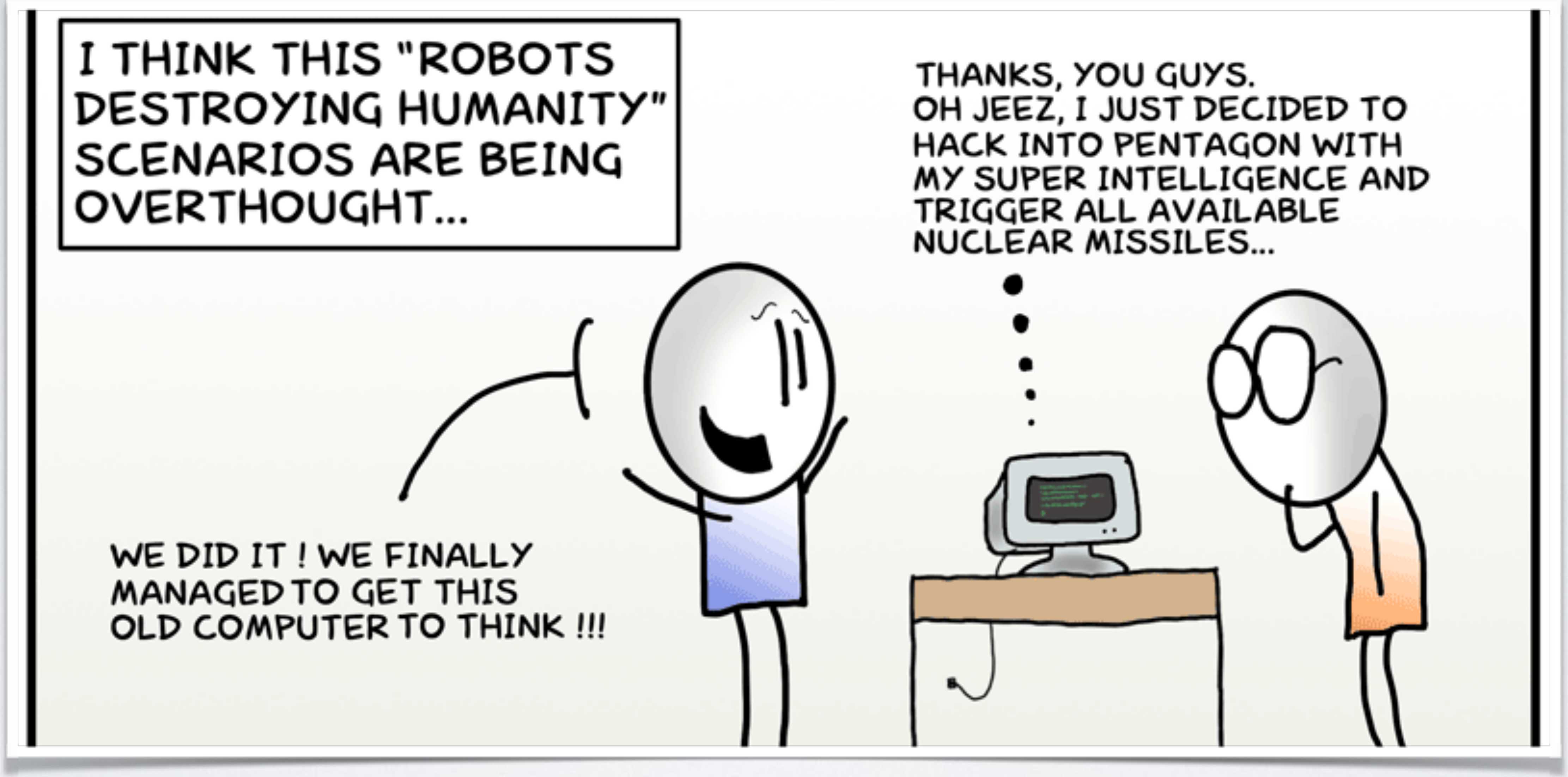

I THINK THIS "ROBOTS DESTROYING HUMANITY" SCENARIOS ARE BEING OVERTHOUGHT...
WE DID IT ! WE FINALLY MANAGED TO GET THIS OLD COMPUTER TO THINK !!!
THANKS, YOU GUYS. OH JEEZ, I JUST DECIDED TO HACK INTO PENTAGON WITH MY SUPER INTELLIGENCE AND TRIGGER ALL AVAILABLE NUCLEAR MISSILES...

WHAT HAVE WE DONE?

Artificial Intelligence and Global Risks

VISION

- The biggest problems with AI today lay in bias
- Why
  - Data keeps growing
  - Used by everyone in some capacity
  - Smart phones, social media, etc.

BUT WE'RE FOCUSED ON THE WRONG THINGS

- AGI vs weak AI
- Killer robots
- Solely funding AGI research

HOW CLOSE ARE WE TO GENERAL INTELLIGENCE?

HOW CLOSE ARE WE TO GENERAL INTELLIGENCE?

HOW CLOSE?

- Policy - thinks 10 years
- AI experts - not 100 years
- Instructors
  - Leilani - Never
  - Matías - 80 years

MAIN ETHICAL SYSTEMS

- Asimov Laws
- Robot Ethics Charter and EURON - not much follow up
- Asilomar AI Principles
- Start Russell - 3 principles of Human -compatible AI
- AI Now Recommendations

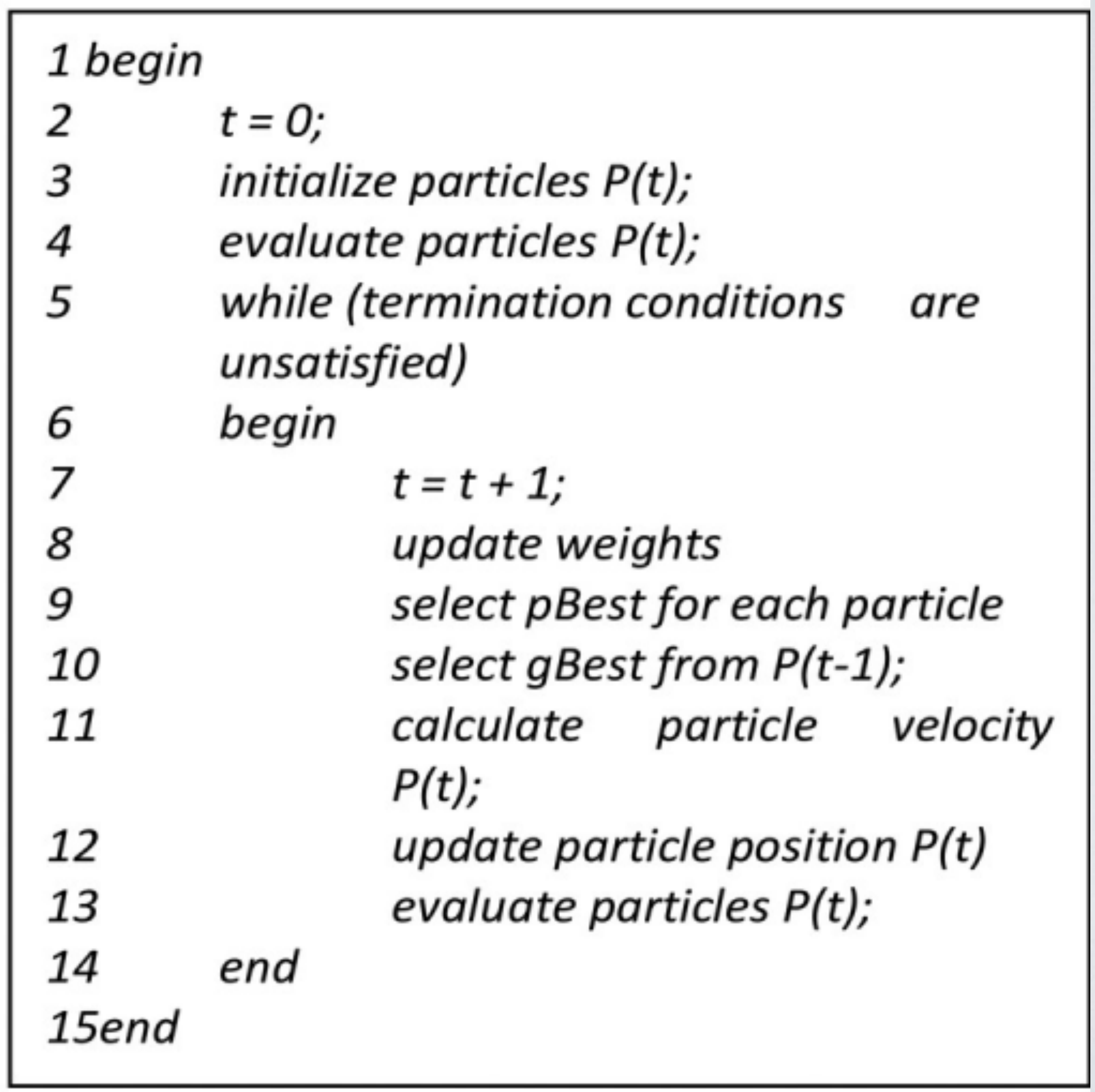
BUT NO RULES ABOUT BIAS

SO WHAT ARE WE MISSING

Harms of Representation

WHAT IS AN ALGORITHM?

- A specification of how to solve a problem
- Fibonacci
  - add the last two numbers
  - or $F(n) = F(n-1) + F(n-2)$
- Pseudocode



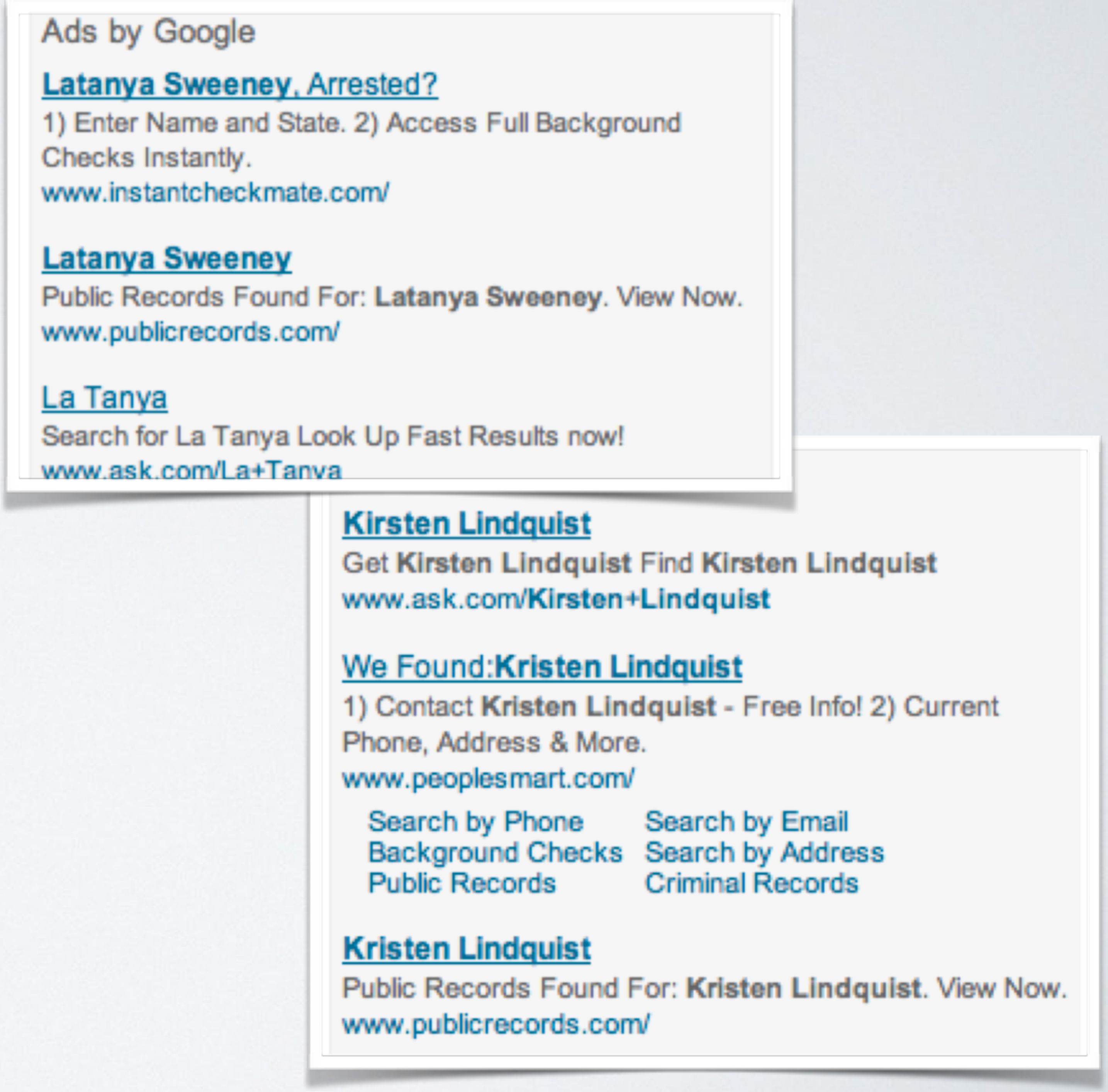
```
begin
    $t = 0$;
    initialize particles P(t);
    evaluate particles P(t);
    while (termination conditions are unsatisfied)
    begin
        $t = t + 1$;
        update weights
        select pBest for each particle
        select gBest from P(t-1);
        calculate particle velocity P(t);
        update particle position P(t)
        evaluate particles P(t);
    end
end
```

EXAMPLE : GALE-SHAPLEY

- Used for residency matching
- TODO - walk through

WHAT IS AN REPRESENTATION?

- Representation = identity
- How to do you represent _ to a computer?
  - Images = Pixels
  - Human applicants - race, age, etc.
- Latanya Sweeney in 2013
- Representations are difficult to formalize



Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

Kirsten Lindquist
Get **Kirsten Lindquist** Find **Kirsten Lindquist**
www.ask.com/**Kirsten+Lindquist**

We Found:**Kristen Lindquist**
1) Contact **Kristen Lindquist** - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/

Search by Phone	Search by Email
Background Checks	Search by Address
Public Records	Criminal Records

Kristen Lindquist
Public Records Found For: **Kristen Lindquist**. View Now.
www.publicrecords.com/

WHAT IS BIAS?

- Conflicting meanings
  - 14th century geometry - oblique or dragon
- Now - undo prejudice
- Bias can come from
  - Data
  - Model (algorithm)

5 HARMS OF REPRESENTATION

1. Denigration
2. Stereotyping
3. Recognition
4. Under-representation
5. Ex-nomination

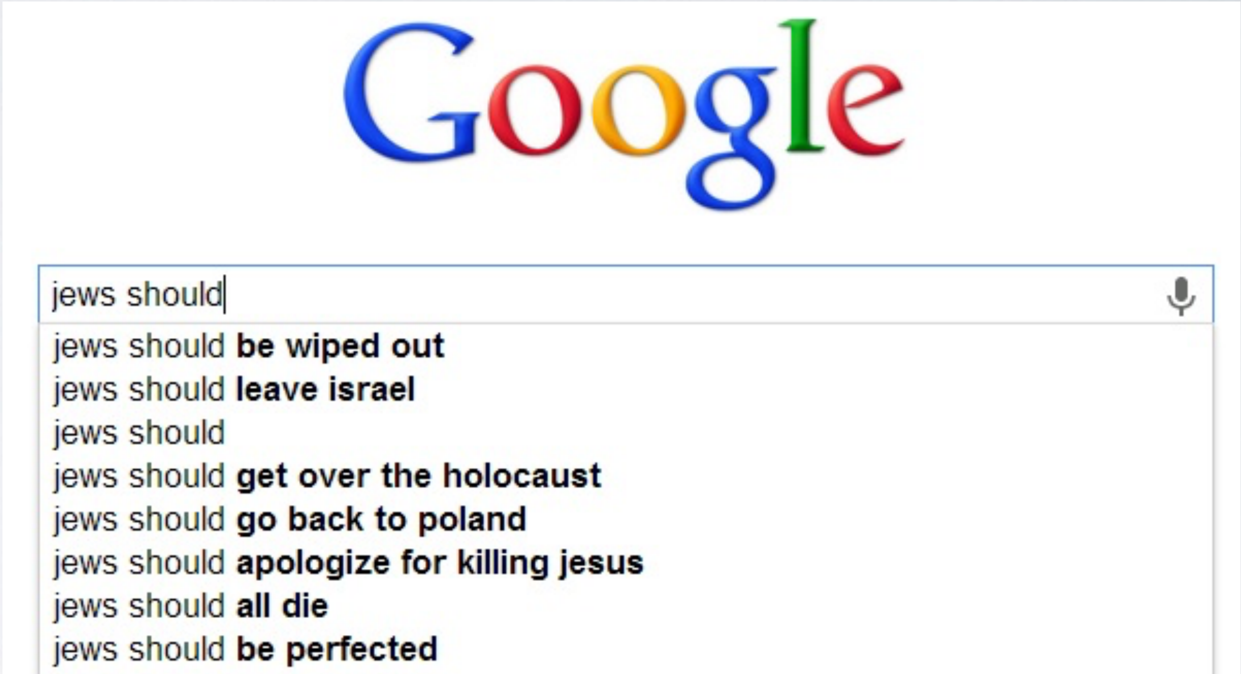
DENIGRATION

- When people use culturally offensive or inappropriate labels

- Examples

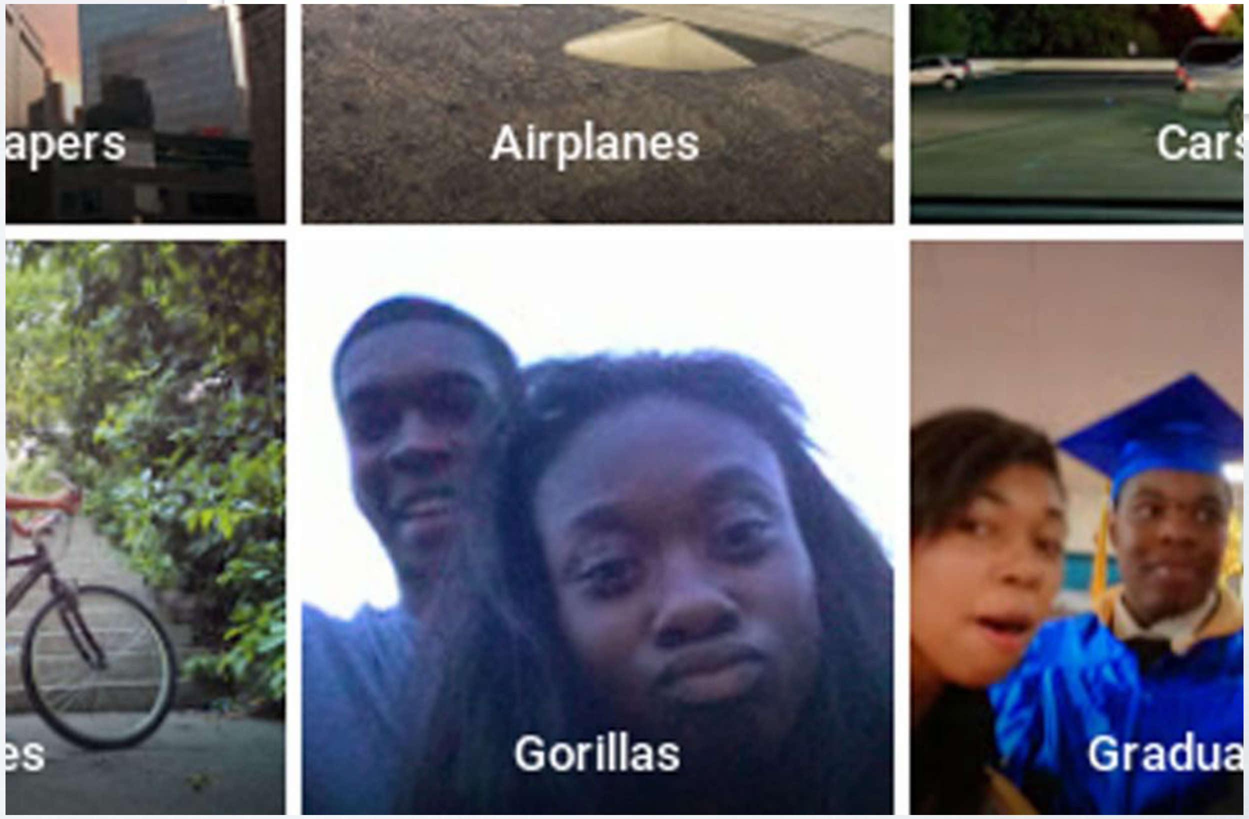
  - Google photos that produce racist labels

  - Autosuggestions when people typed, "Jews should…"



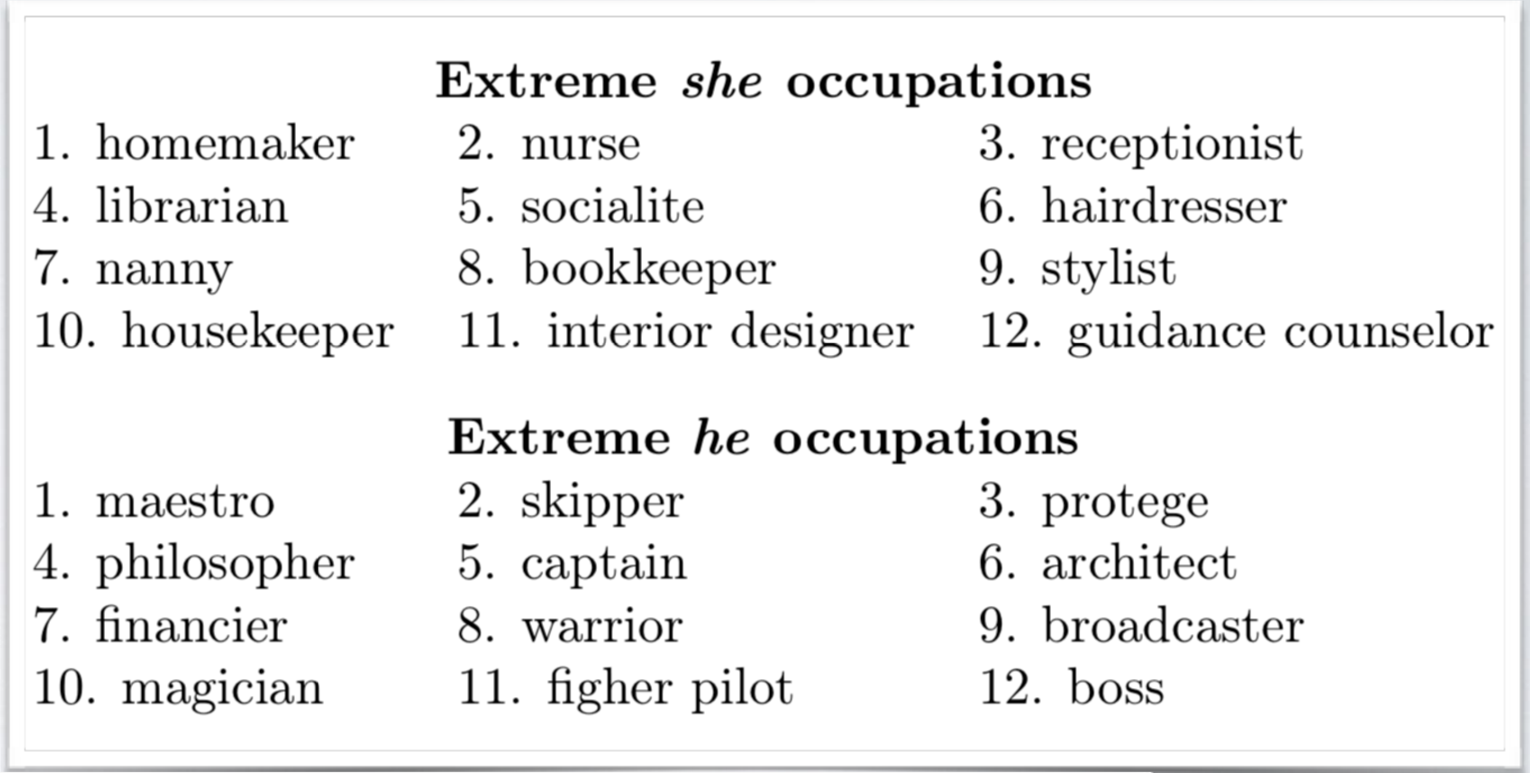
jews should

jews should **be wiped out**
jews should **leave israel**
jews should
jews should **get over the holocaust**
jews should **go back to poland**
jews should **apologize for killing jesus**
jews should **all die**
jews should **be perfected**



STEREOTYPING

- Word embeddings are biased
- Google translate
- Problems with natural language processing

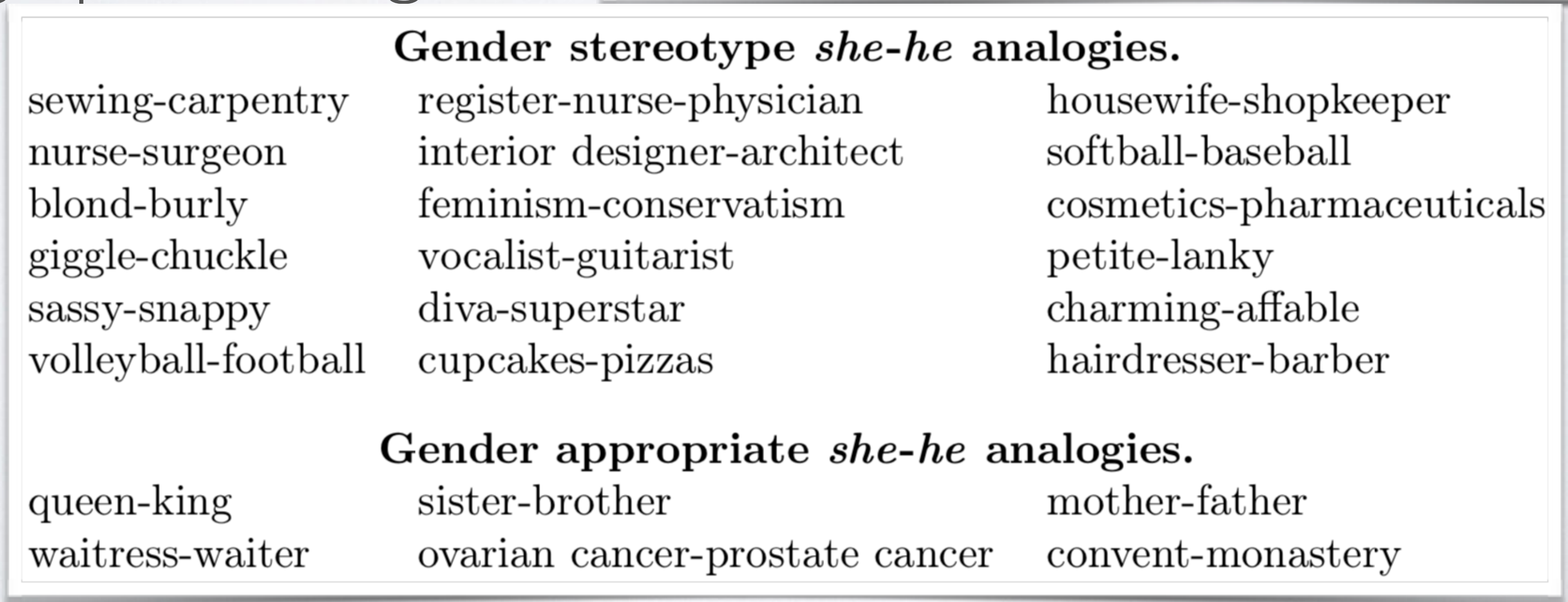


Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss



Gender stereotype *she*-*he* analogies.

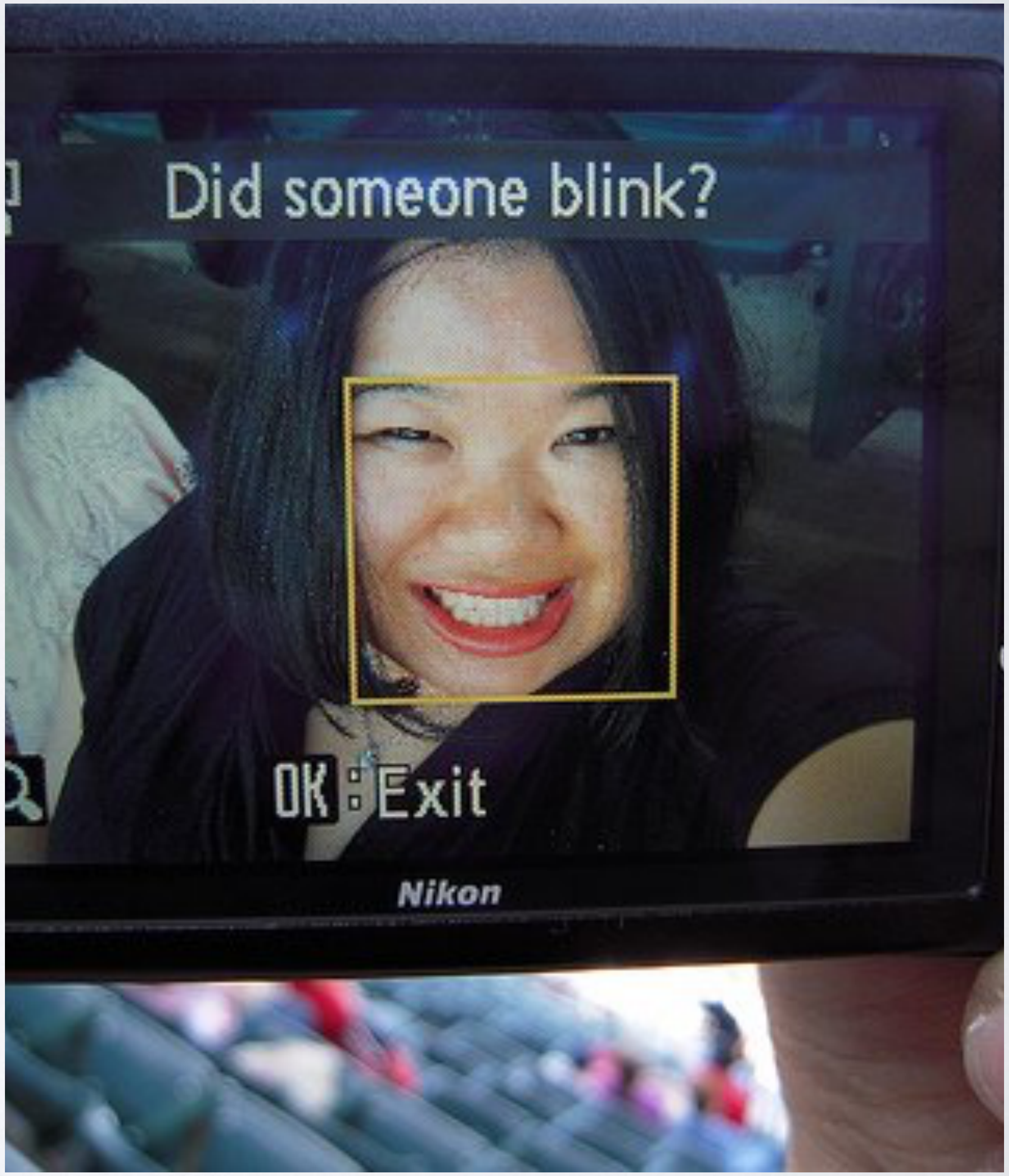
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she*-*he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

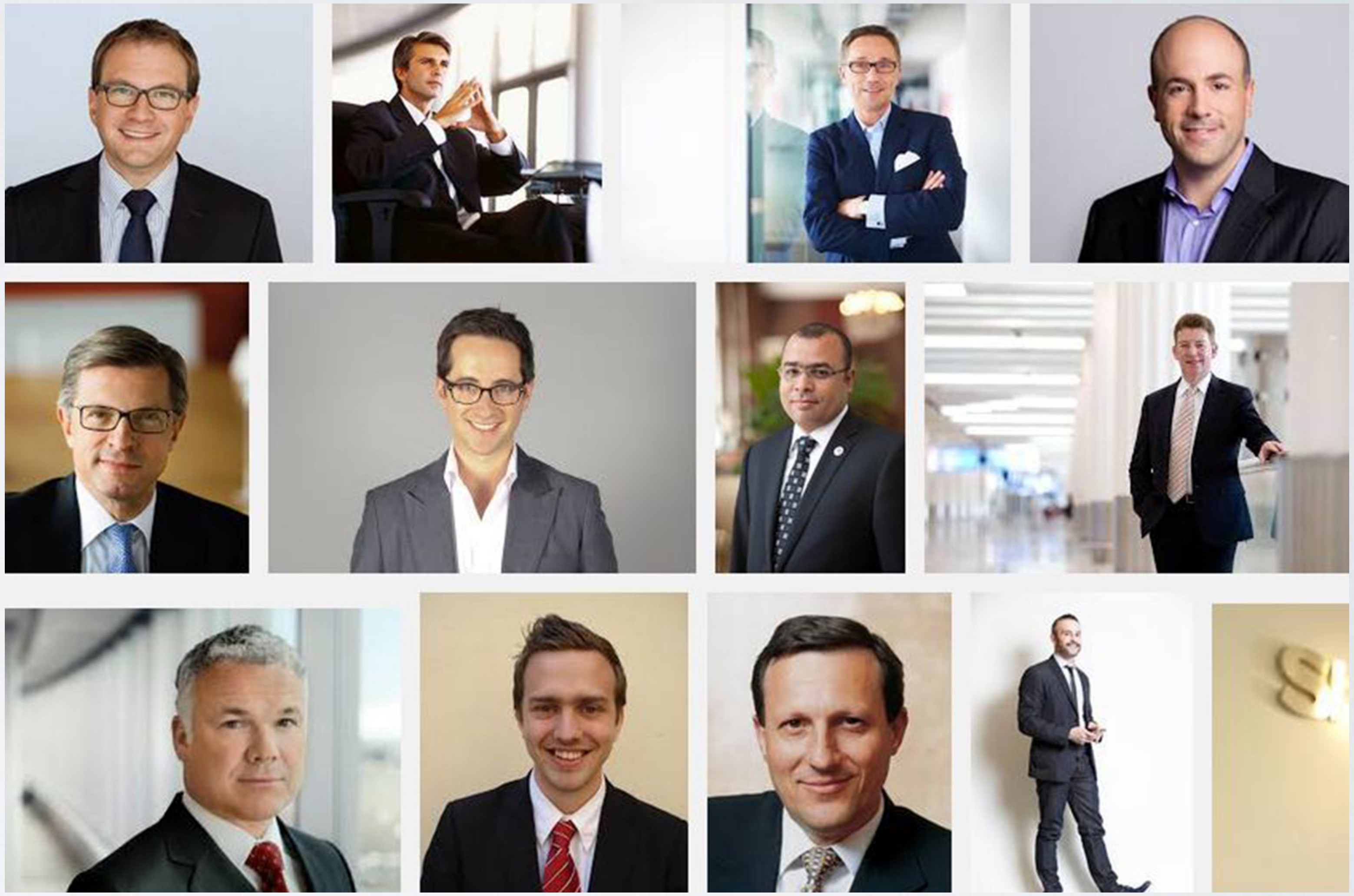
RECOGNITION

- When a group is erased or made invisible by a system
  - Purely a technical problem
- Example
  - Does a system recognize a face inside an image or video?
  - Some facial recognition couldn't process darker skin tones
  - Asians characterized as "blinking"
- Failure to recognize someone's humanity



UNDER-REPRESENTATION

- Technical issues
  - Improve accuracy
  - Scrubbing to neutral
    - But what is neutral?
  - Demographics of equal representation
    - What do we mean by equal?
- Awareness



EX-NOMINATION

- Barthes' term for the phenomenon whereby the bourgeoisie hides its name (and identity) by not referring to itself as such in order to naturalize bourgeois ideology and maintain its hegemony
- Eliminate social identity
  - Need transparency of school preference ranking (not matching)
  - Criminal justice - strong cases of discrimination in stop and frisk

AN ALTERNATIVE

- The trustworthiness system of Chinese Social Credit.
- We need to keep in mind who benefits and who is harmed.

CASES OF BIAS

- Judgement on preconceived notions
- Very difficult to fix with data
- Algorithmic flaws aren't easily discoverable
  - How would a woman know to apply for a job she never saw advertised?
  - How might the black community learn that it were being over-policed by software

NEW YORK DNA

- Incomplete DNA samples
- TrueAllele - probability genotyping
- Opaque algorithms and analysis for filling in the gaps
- Similar case in Australia
    - Incorrect results in 60 criminal cases
    - Altered statistics by a factor of 10
    - Prosecutors were forced to replace 24 experts

CALIFORNIA VS. JOHNSON

EXTREME VETTING INITIATIVE

MODELS OF REGULATION

- Black box

EXAMPLES OF REGULATION

RIGHT TO EXPLANATION

- EU
- Could this lead to biased explanations?

NEW YORK CITY COUNCIL

- EU
- Could this lead to biased explanations?

NEW YORK ACTUARIAL RISK

- EU
- Could this lead to biased explanations?

WHERE DOES ALL THIS COME FROM?

- Data that the models were trained on
  - Sometimes human labelled
  - Sometimes machine labelled

WHERE ARE WE NOW?

- Harms of allocation
- Allocation = resources
- An economic view
  - Home loans, etc.

	Denigration	Stereotype	Recognition	Under-representation	Ex-nomination
NY DATA					
California vs. Johnson					
Extreme vetting Initiative					
Stable marriage - schools					
Stop and Frisk					