Can Artificial Intelligence Fight Alternative Facts?

Daniel Lowd
Associate Professor
Dept. of Computer and Information Science
University of Oregon
April 23, 2019

Sri Lanka, citing 'false news reports,' blocks social media after attacks



By Donie O'Sullivan, CNN Business

Updated 4:24 PM ET, Sun April 21, 2019





Moment blast strikes church caught on camera

Microsoft CEO on trade, mixed reality and taking the long view

IBM CEO: Need to hire people for skills, not just degrees

Self-driving cars are coming. But Lyft co-founder says he'll need more drivers

Mortgage | Personal Loans | Credit Cards

Loan Type	Rate	APR
30-yr fixed	3.625%	3.625%
15-yr fixed	3.125%	3.125%
5/1 ARM	3.375%	4.41%

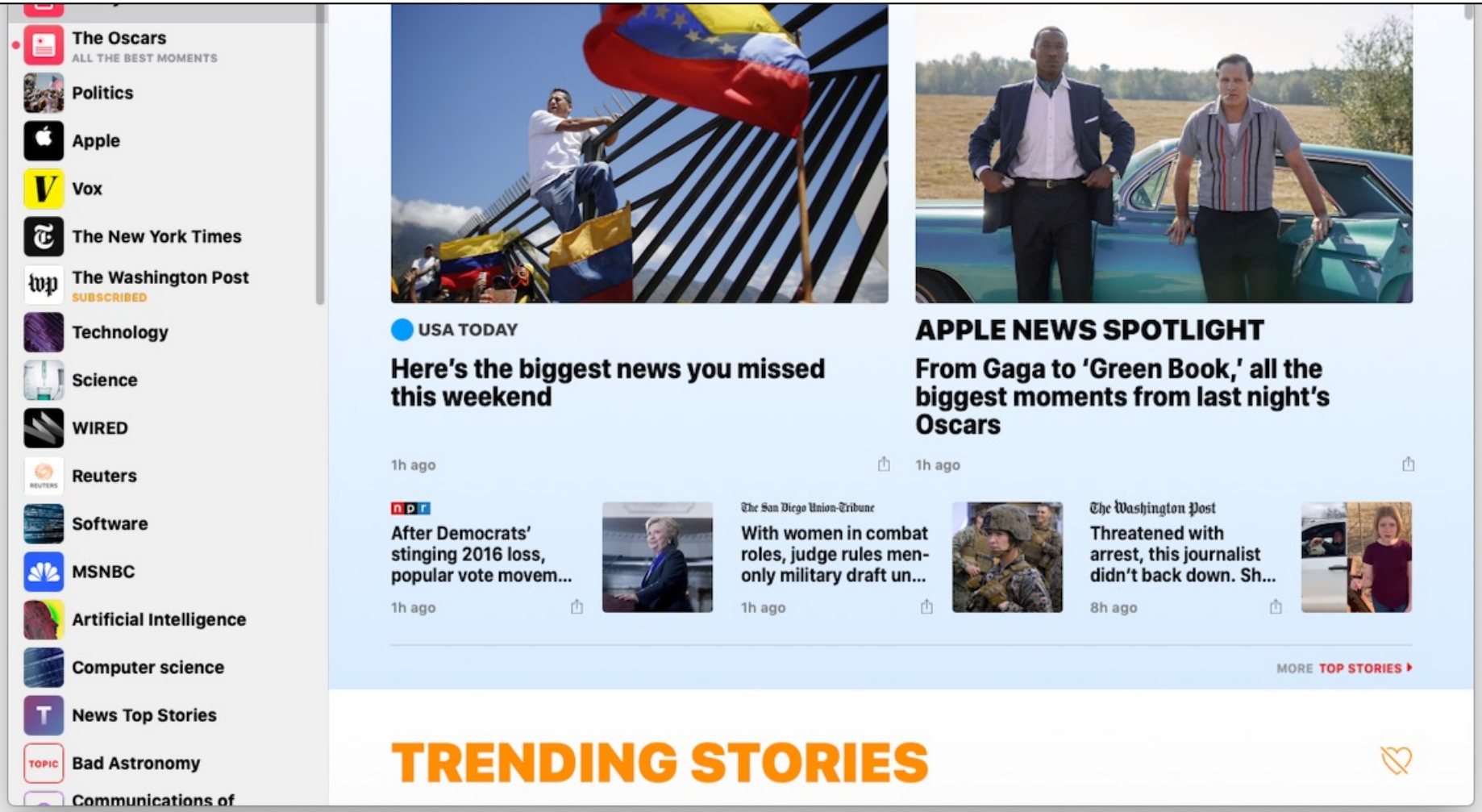
Loan Amount	APR	Payment
$225,000 (5/1 ARM)	4.41%	$994.72/mo
$350,000 (5/1 ARM)	4.748%	$1670.96/mo

Get Personalized Rates

lendingtree NMLS#1136 Terms & Conditions Apply

New York (CNN Business) – Sri Lanka placed a nationwide block on social media sites after more

How has AI affected your information this week?



www.google.com/search?source=hp&ei=HLJzXMCqEoOq_QadmoTQAQ&q=

Google

queen

All Videos Images News Shopping More Settings Tools

About 2,980,000,000 results (0.50 seconds)

Top stories



Queen, Adam Lambert Open Oscars with Freddie Mercury Tribute

Rolling Stone

8 hours ago



Queen divide audience at Oscars ceremony

The Guardian

7 hours ago



Queen's Roger Taylor claims he's yet to get paid for Bohemian Rhapsody

Metro

56 mins ago

More for queen

queen on Twitter

https://twitter.com/search/queen

Queen (@QueenWillRock)

Congratulations @ItsRamiMalek 🙌🏻 Very well deserved! #bestactor #Oscars pic.twitter.com/PnVAHca…

5 hours ago · Twitter

ABC (@ABCNetwork)

🍗🍗👏🍗🍗👏

@QueenWillRock + @adamlambert ROCKED this #Oscars opening! pic.twitter.com/2g1Qatt…

8 hours ago · Twitter

MTV NEWS (@MTVNEWS)

Adam Lambert and Queen kick off the #Oscars, ensuring the Glenn Close is sufficiently rocked pic.twitter.com/LMVvoAL…

8 hours ago · Twitter



Queen

Rock band

queenonline.com

Available on

YouTube

Spotify

Apple Music

More music services

Queen are a British rock band formed in London in 1970. Their classic line-up was Freddie Mercury, Brian May, Roger Taylor, and John Deacon. Wikipedia

Genre: Rock

Albums: A Night at the Opera, Innuendo, News of the World, MORE

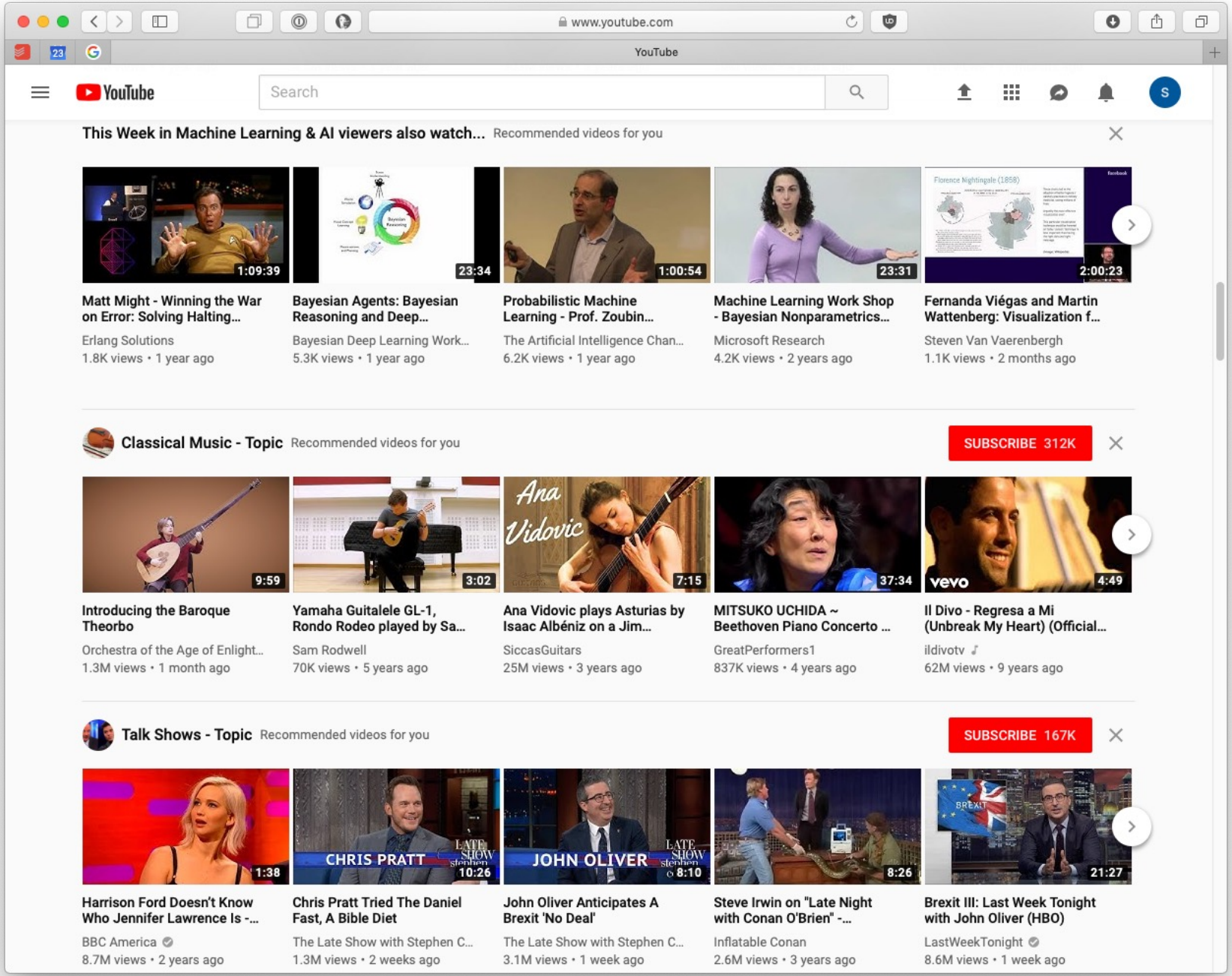
Awards: Grammy Hall of Fame, MORE

Songs

Bohemian Rhapsody

A Night at the Opera · 1975




www.youtube.com
YouTube
YouTube
Search
This Week in Machine Learning & AI viewers also watch... Recommended videos for you
1:09:39
Matt Might - Winning the War on Error: Solving Halting...
Erlang Solutions
1.8K views • 1 year ago
23:34
Bayesian Agents: Bayesian Reasoning and Deep...
Bayesian Deep Learning Work...
5.3K views • 1 year ago
1:00:54
Probabilistic Machine Learning - Prof. Zoubin...
The Artificial Intelligence Chan...
6.2K views • 1 year ago
23:31
Machine Learning Work Shop - Bayesian Nonparametrics...
Microsoft Research
4.2K views • 2 years ago
2:00:23
Fernanda Viégas and Martin Wattenberg: Visualization f...
Steven Van Vaerenbergh
1.1K views • 2 months ago
Classical Music - Topic Recommended videos for you
SUBSCRIBE 312K
9:59
Introducing the Baroque Theorbo
Orchestra of the Age of Enlight...
1.3M views • 1 month ago
3:02
Yamaha Guitalele GL-1, Rondo Rodeo played by Sa...
Sam Rodwell
70K views • 5 years ago
Ana Vidovic
7:15
Ana Vidovic plays Asturias by Isaac Albéniz on a Jim...
SiccasGuitars
25M views • 3 years ago
37:34
MITSUKO UCHIDA ~ Beethoven Piano Concerto ...
GreatPerformers1
837K views • 4 years ago
vevo
4:49
Il Divo - Regresa a Mi (Unbreak My Heart) (Official...
ildivotv
62M views • 9 years ago
Talk Shows - Topic Recommended videos for you
SUBSCRIBE 167K
1:38
Harrison Ford Doesn't Know Who Jennifer Lawrence Is -...
BBC America
8.7M views • 2 years ago
CHRIS PRATT
LATE SHOW stephen
10:26
Chris Pratt Tried The Daniel Fast, A Bible Diet
The Late Show with Stephen C...
1.3M views • 2 weeks ago
JOHN OLIVER
LATE SHOW stephen
8:10
John Oliver Anticipates A Brexit 'No Deal'
The Late Show with Stephen C...
3.1M views • 1 week ago
8:26
Steve Irwin on "Late Night with Conan O'Brien" -...
Inflatable Conan
2.6M views • 3 years ago
BREXIT
21:27
Brexit III: Last Week Tonight with John Oliver (HBO)
LastWeekTonight
8.6M views • 1 week ago

Overview

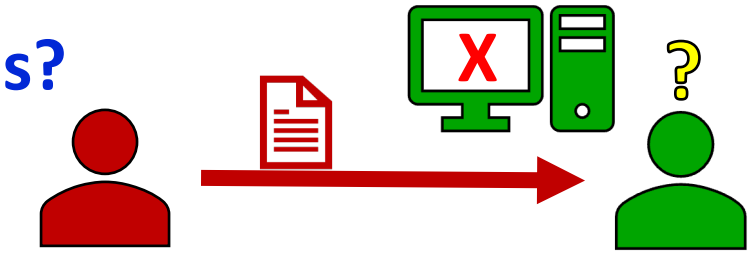
- Social media is a powerful tool for spreading hate and lies.
- In every online platform, AI is heavily involved in what content is blocked and what content is promoted.
- Therefore, we need to understand the interaction between AI systems and misinformation. (We also need sociology, psychology, public policy, and more.)

What is this talk about?

- Background: What is AI? How does it work?
- Three types of human-computer deception:

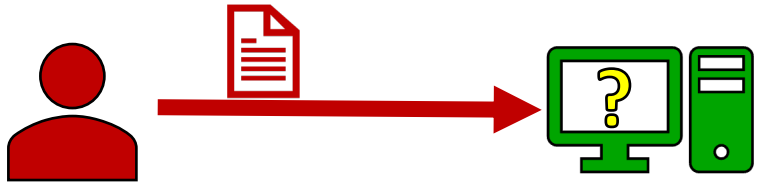
1. How can AI stop humans tricking humans?

Example: Detecting fake news, phishing, fraud



2. How can humans trick AI?

Example: Adversarial examples, data poisoning



3. How can AI trick humans?

Example: Deep fakes, AI-generated fake news



What is artificial intelligence?

- Automating tasks that require human intelligence
- Solving complex problems without having a step-by-step solution









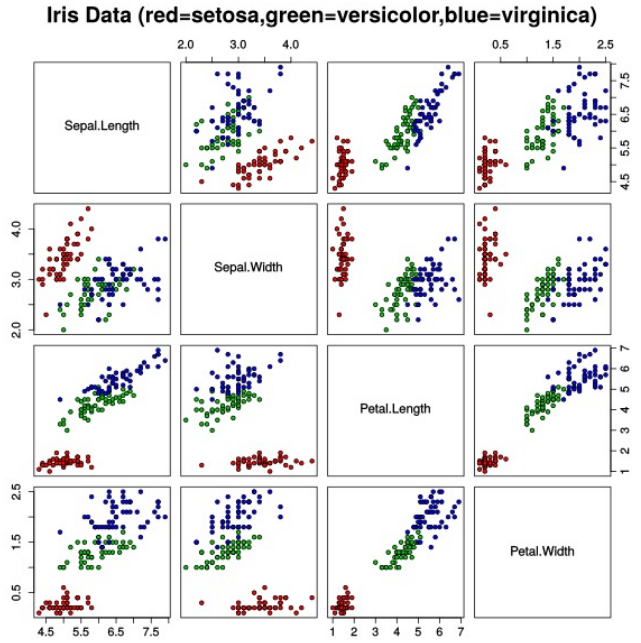
What is machine learning?

- A branch of artificial intelligence
- Solving complex problems by:
  - Using data to fit a model
  - Using the model to make predictions or recommendations (typically)

Examples:

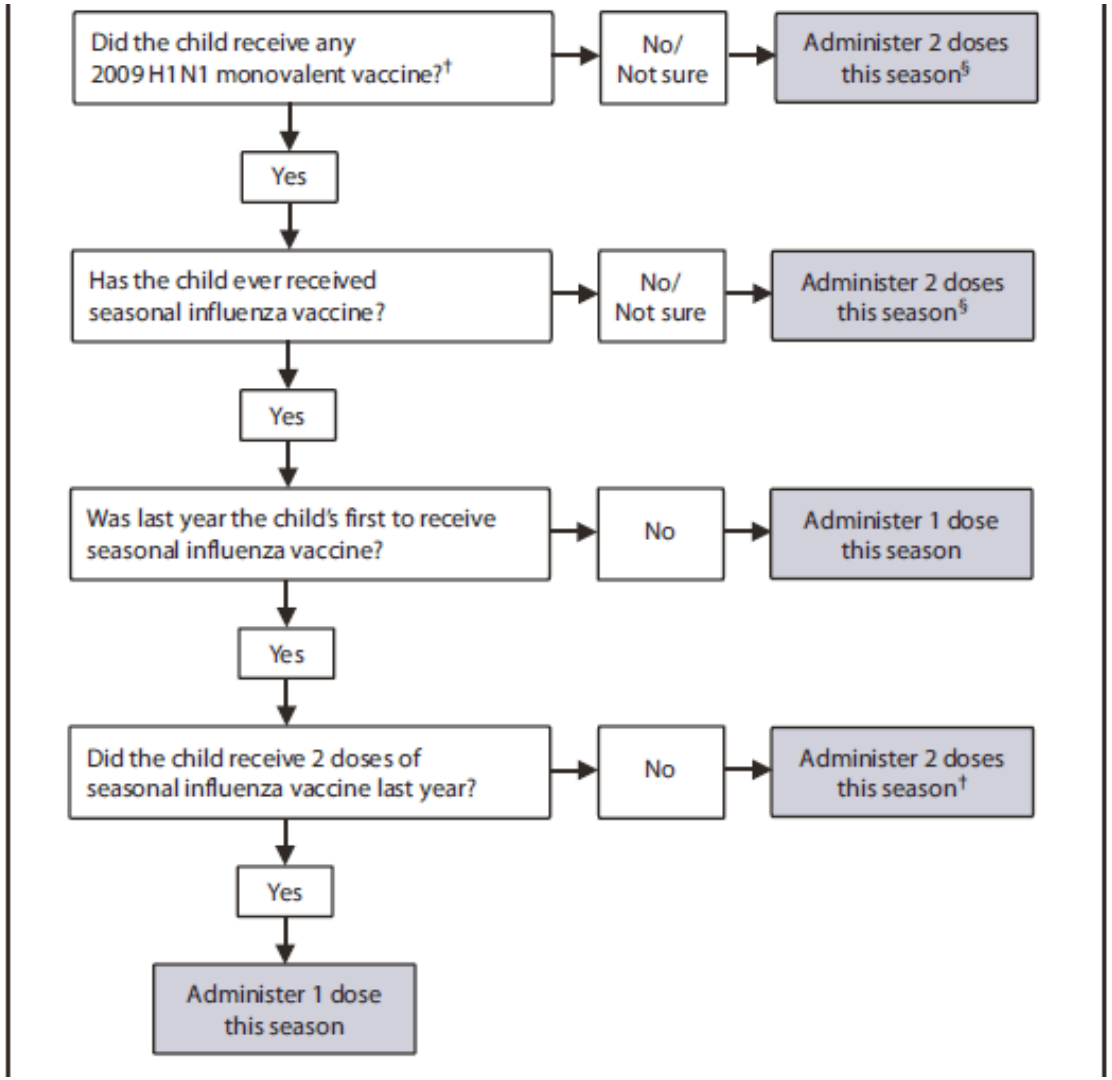
- Spam filters
- Recommender systems (Netflix, Apple News)
- Speech recognition
- Language translation





https://en.wikipedia.org/wiki/File:Iris_dataset_scatterplot.svg
Creative Commons Attribution. Wikipedia user: Nicoguaro

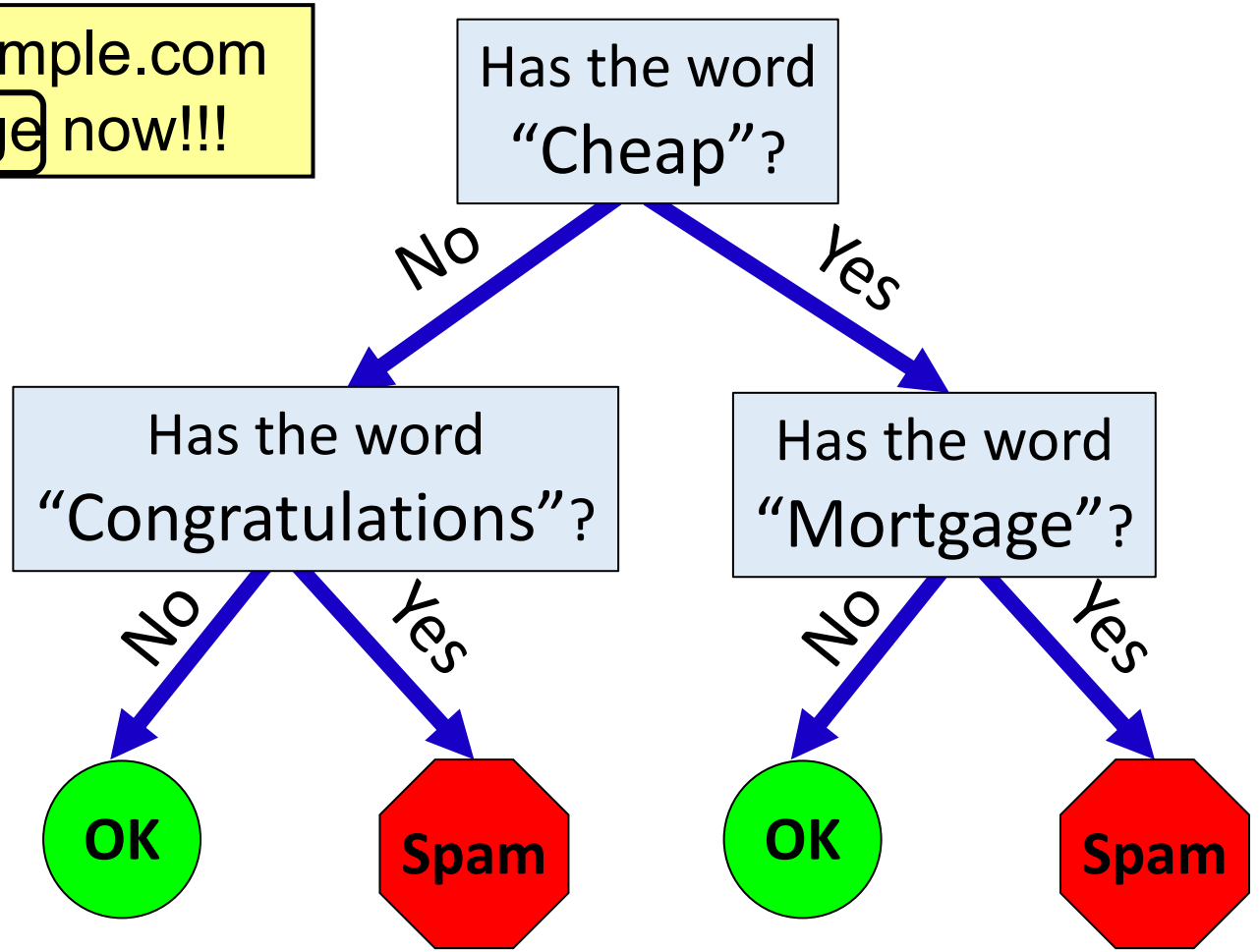
Decision Tree Model:
Flu vaccine doses for children 6 mos – 8 yrs



Source: https://www.cdc.gov/mmwr/preview/mmwrhtml/figures/r5908a1f3.gif

Spam Filtering with Decision Tree Models

From: spammer@example.com
Cheap cheap mortgage now!!!



Spam Filtering with Linear Models

From: spammer@example.com
Cheap cheap mortgage now!!!

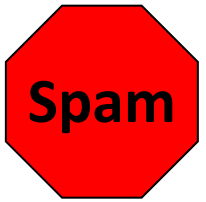


Words:	
“cheap”:	2
“mortgage”:	1
“congratulations”:	0
“Oregon”:	0
“Eugene”:	0

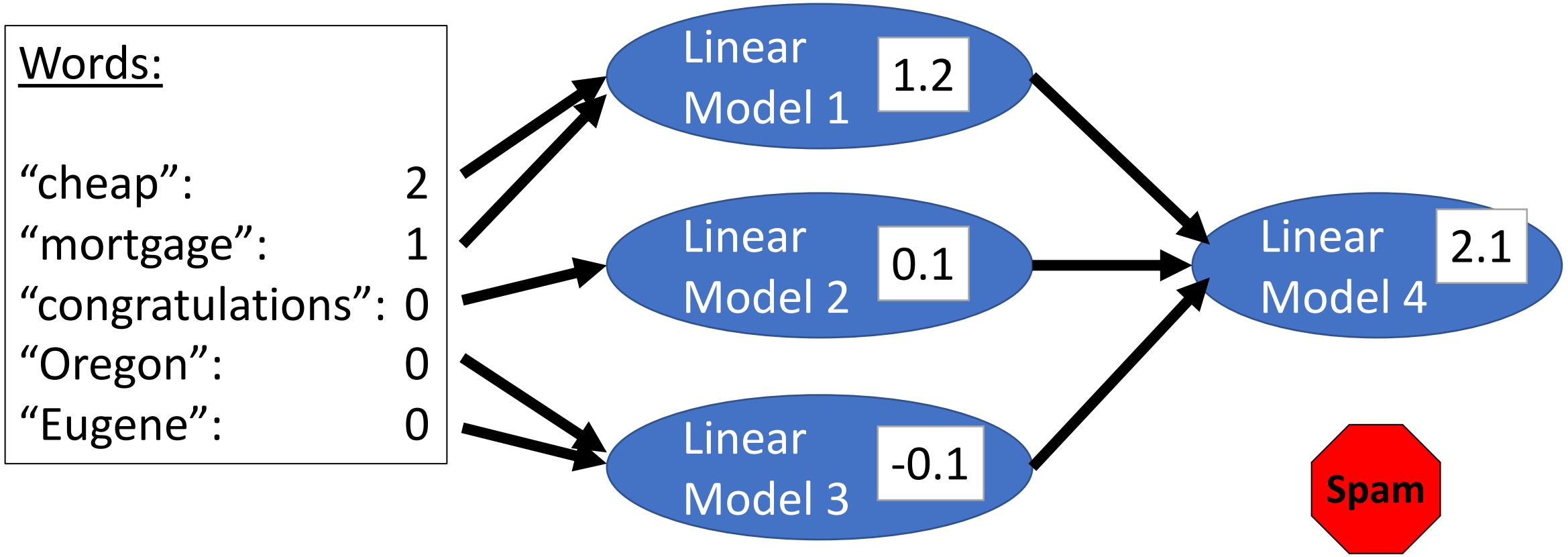
X

Weights:	
“cheap”:	1.0
“mortgage”:	1.5
“congratulations”:	1.2
“Oregon”:	-1.0
“Eugene”:	-1.0

= 2*1.0 + 1*1.5
= 3.5



Spam Filtering with Neural Network Models

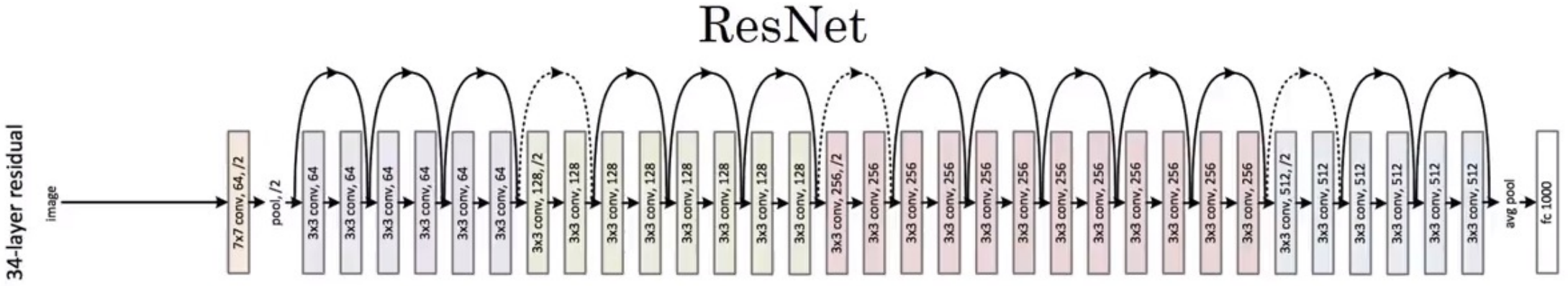


Learning With Neural Networks

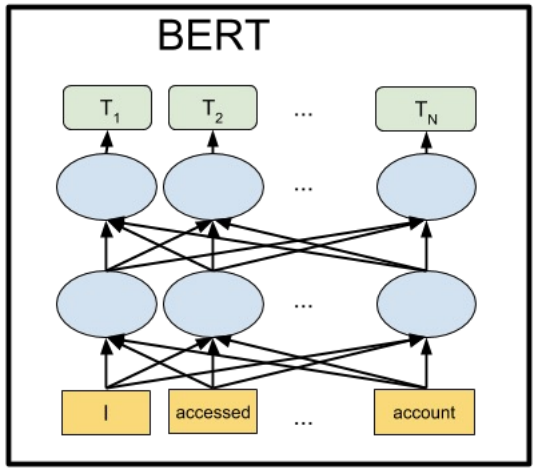
- Simple version:
  - Try changing each weight a tiny bit.
  - If changing it makes the predictions better, keep it.
  - Keep going until it works or you run out of time.

- Calculus version (stochastic gradient descent):
  - Choose a loss function to minimize, such as the number of mistakes.
  - Compute derivatives of the loss function with respect to each weight.
  - This derivative describes how much changing each weight could help.
  - Increase or decrease each weight based on these derivatives.

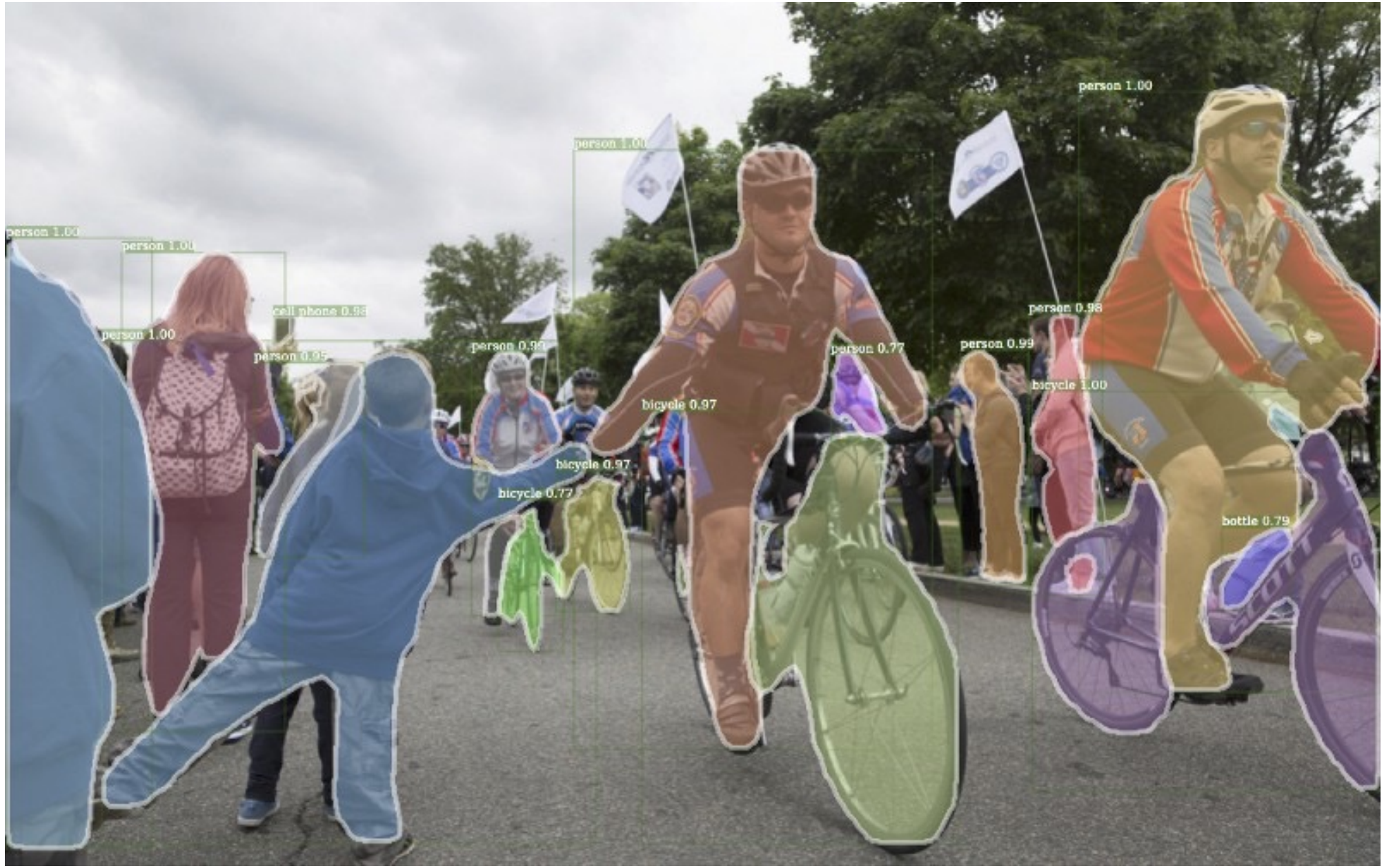
Neural Networks Today: Deep Learning



He et al. (2015)

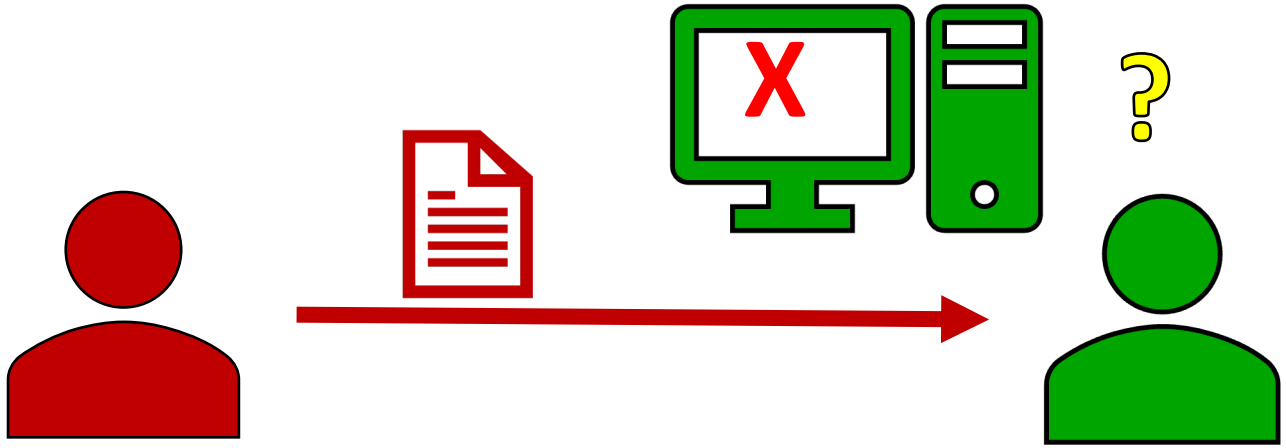


Devlin et al. (2018)



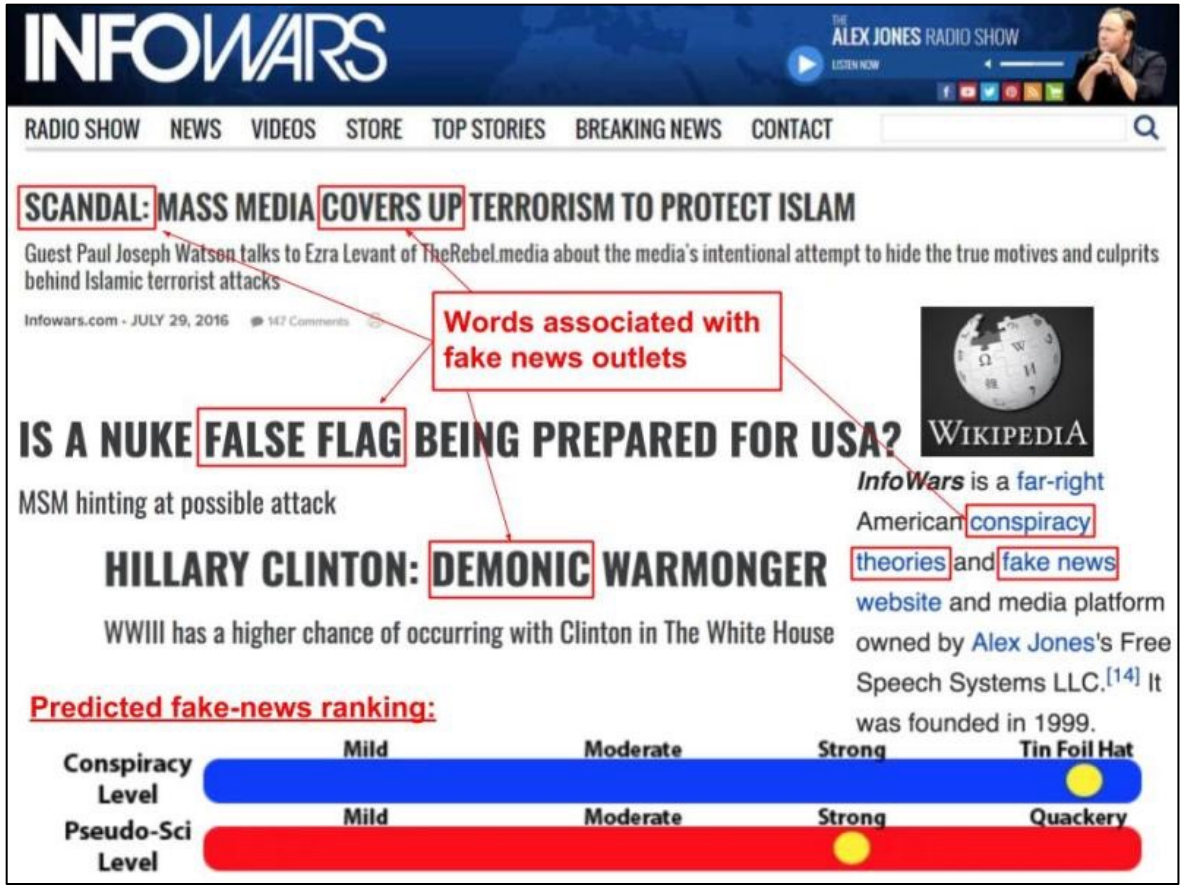
Source: Facebook Research, https://github.com/facebookresearch/Detectron

1. How can AI stop humans tricking humans?



Spam Filters For News

1. Choose features
2. Collect data
3. Fit a model
4. Make predictions!



Baly et al., 2018

What Features Predict Fake News? Baly et al., 2018

- **Content:** Words like “demonic” and “crisis actor”
- **News outlet:** Short Wikipedia page with words like “extreme” and “conspiracy theory”
- **URL:** special characters, complicated subdirectories

How well does it work?

65% accurate at predicting factuality (high/low/medium)

Can We Objectively Define Fake News?

The Fake News Challenge (FNC) evaluated AI for **stance detection**, checking headline consistency rather than accuracy.

FORMAL DEFINITION

Input

A headline and a body text - either from the same news article or from two different articles.

Output

Classify the stance of the body text relative to the claim made in the headline into one of four categories:

1. **Agrees**: The body text agrees with the headline.
2. **Disagrees**: The body text disagrees with the headline.
3. **Discusses**: The body text discuss the same topic as the headline, but does not take a position
4. **Unrelated**: The body text discusses a different topic than the headline

http://www.fakenewschallenge.org

EXAMPLE HEADLINE

"Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"

EXAMPLE SNIPPETS FROM BODY TEXTS AND CORRECT CLASSIFICATIONS

"... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. ..."

CORRECT CLASSIFICATION: AGREE

"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."

CORRECT CLASSIFICATION: DISAGREE

http://www.fakenewschallenge.org

Why Stance Detection?

- Fact checking is hard
  - Ambiguous ground truth
  - Lack of good training data
  - Complaints of bias
- Stance detection is useful
  - Find related articles that agree or disagree
  - Determine truth consensus, weighted by credibility

Results: Winning team obtained 82% of the maximum possible score, using gradient-boosted decision trees and a deep neural network.

https://github.com/Cisco-Talos/fnc-1

2. How can humans trick AI?



Adversarial Examples

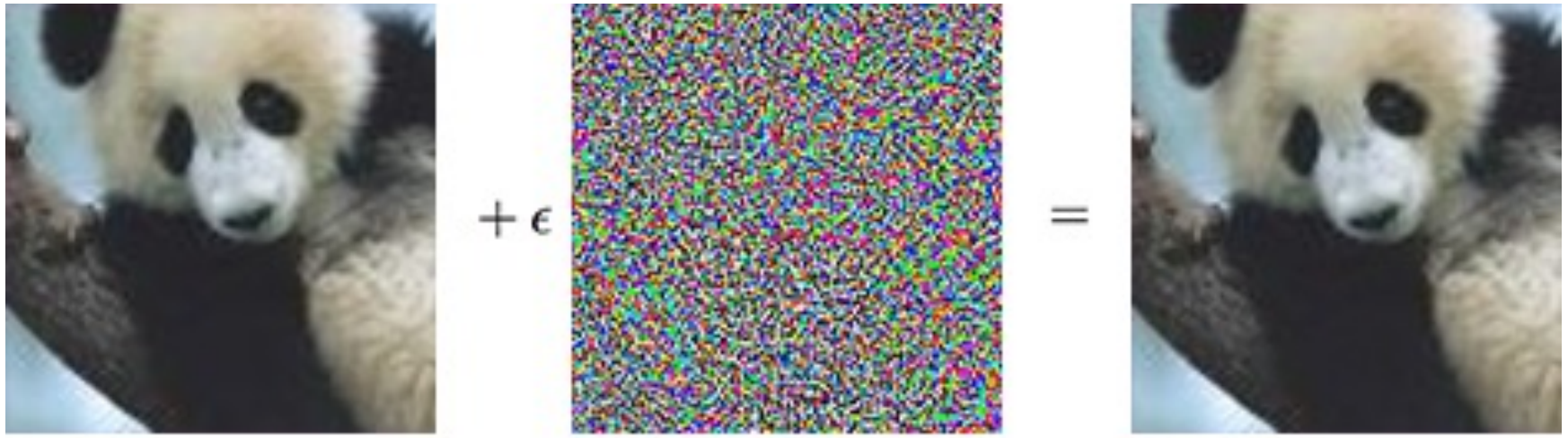


Image credit: https://blog.openai.com/adversarial-example-research/



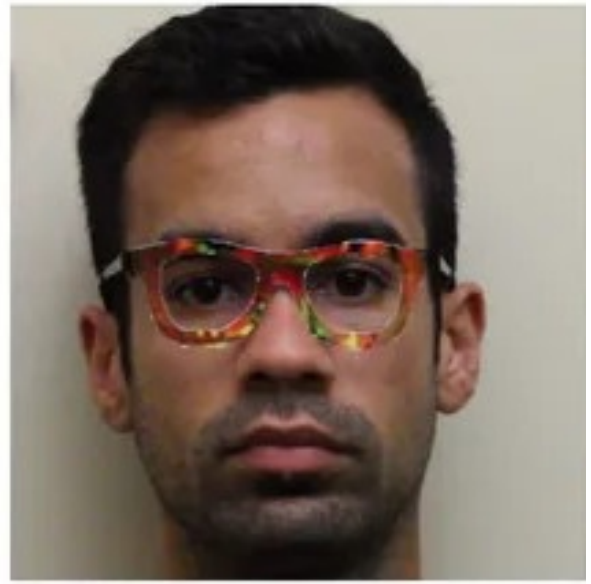
classified as turtle

classified as rifle

classified as other

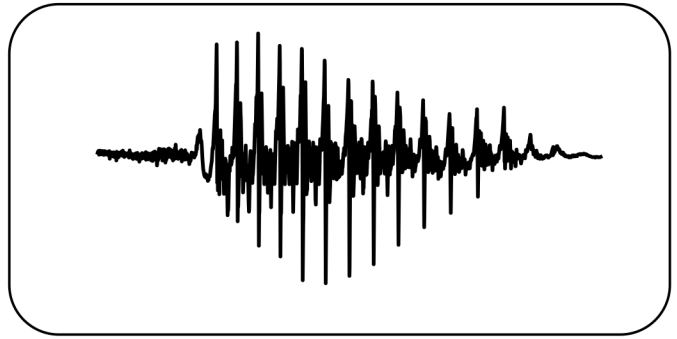
Athalye et al. (2018)

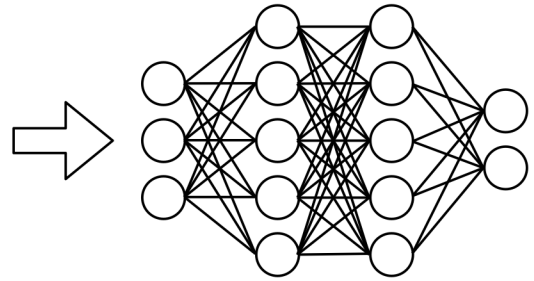






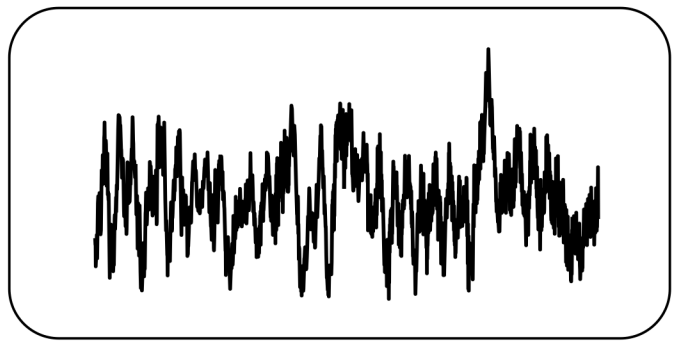

Sharif et al. (2017)





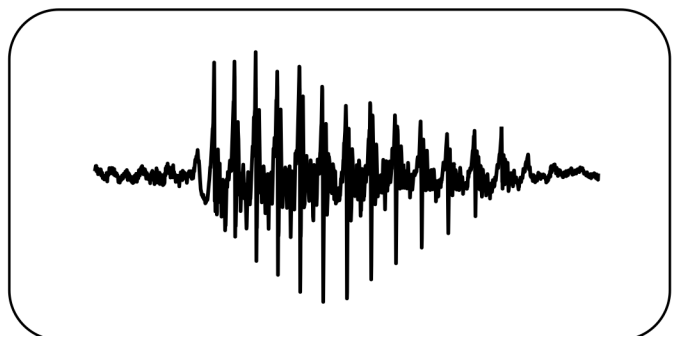
"it was the best of times, it was the worst of times"

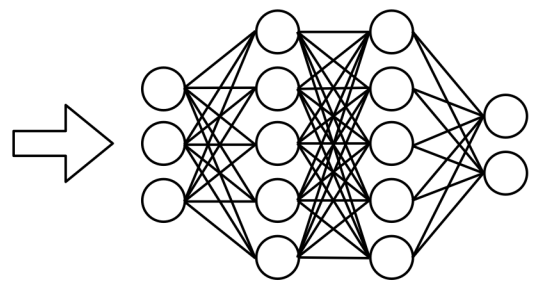
+



× 0.001

=





"it is a truth universally acknowledged that a single"

Carlini et al. (2016)

Adversarial Text

Ebrahimi & Lowd (2018)

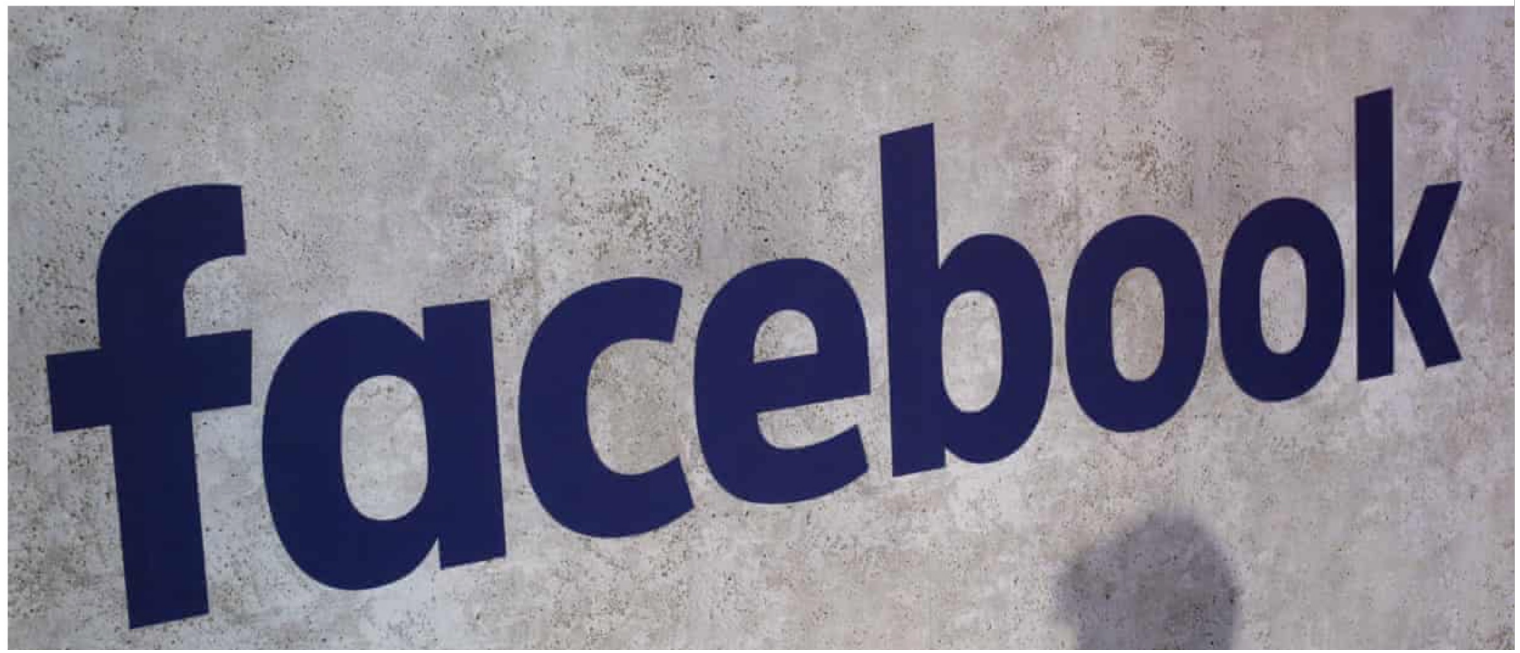
St. Louis is a collection of superstar position players and anonymous pitchers. **Sports**
St. Louis is a collection of superstar position players and anonymous pit-hers. **World**
But is it two Prescotts in one package or a single-die part? **Sci/Tech**
But is it two PrescUtts in one package or a single-die part? **Sports**

Average changes required: 3-5 characters

The Guardian

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer



Adversarial Translation

Goal: Remove the woman's name (Augustine) from the translation.

Original

- **DE:** 1901 wurde eine Frau namens Auguste in eine medizinische Anstalt in Frankfurt gebracht.
- **EN:** In 1931, a woman named Augustine was brought into a medical institution in France.

Adversarial

- **DE:** 1901 wurde eine Frau namens Afuiguste in eine medizinische Anstalt in Frankfurt gebracht.
- **EN:** In 1931, a woman named Rutgers was brought into a medical institution in France.

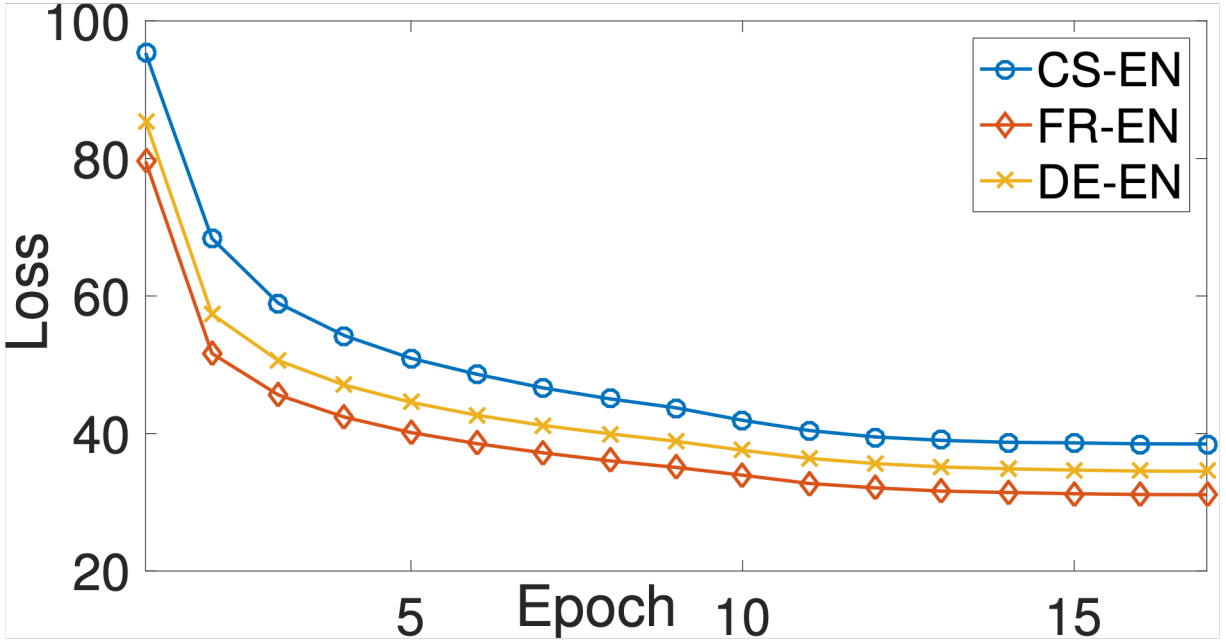
Adversarial Translation

Goal: Replace “therapist” with “psychopath.”

Original	**Adversarial**
• **DE:** Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.	• **DE:** Das ist Dr. Bob Childs – er ist Geigenbauer und Psy6hothearpeiut.
• **EN:** This is Dr. Bob Childs – he’s a wizard maker and a therapist’s therapist.	• **EN:** This is Dr. Bob Childs – he’s a brick maker and a psychopath.

Adversarial Training

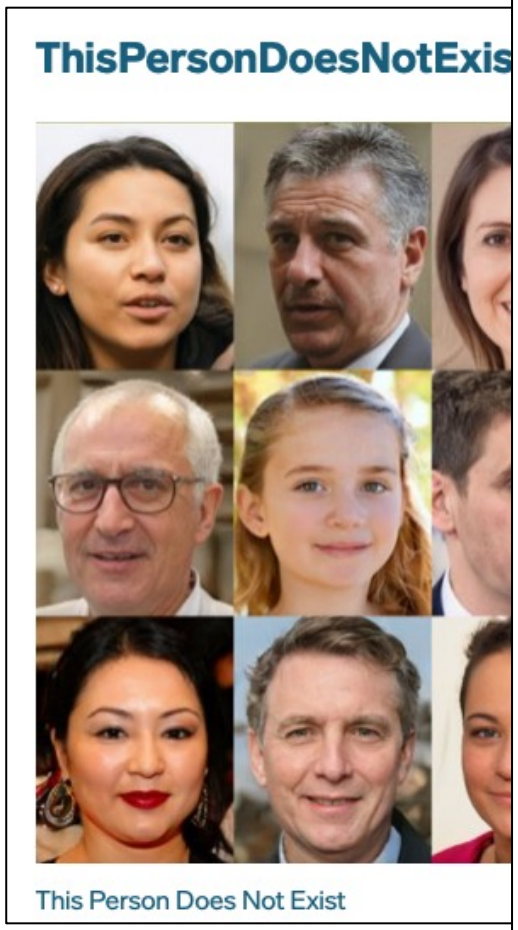
Don’t wait for the adversary to break our models – simulate the adversary’s attacks and learn to detect them

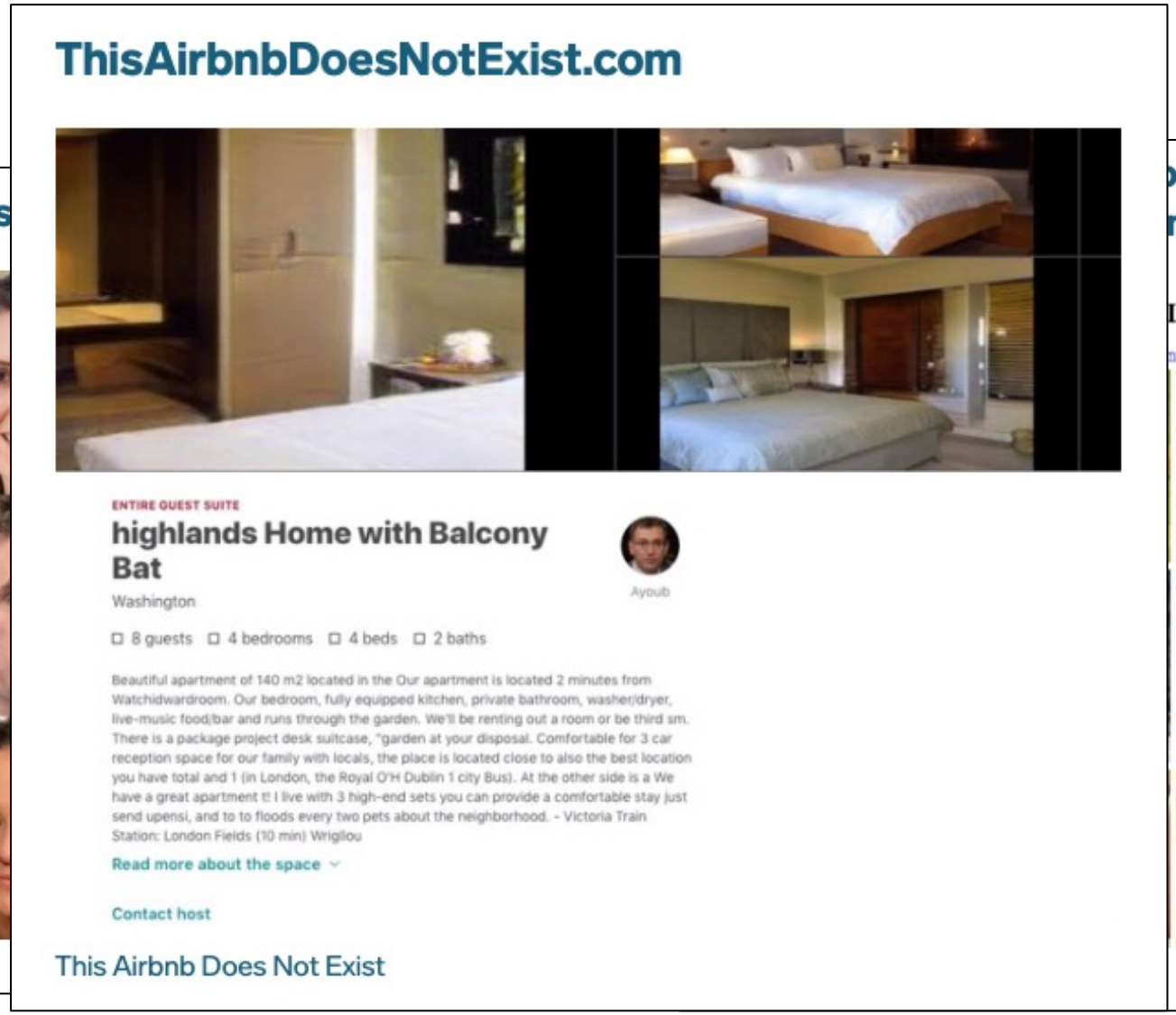


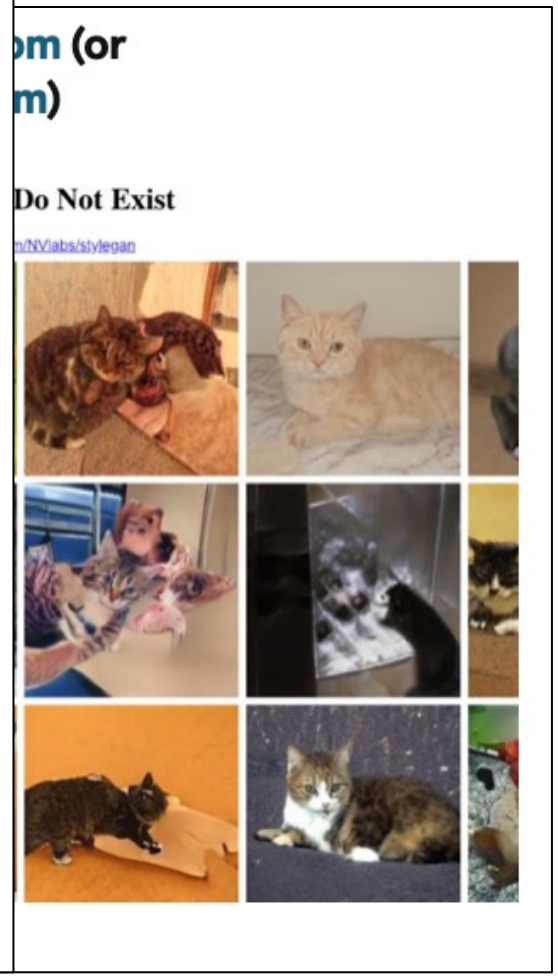
3. How can AI trick humans?

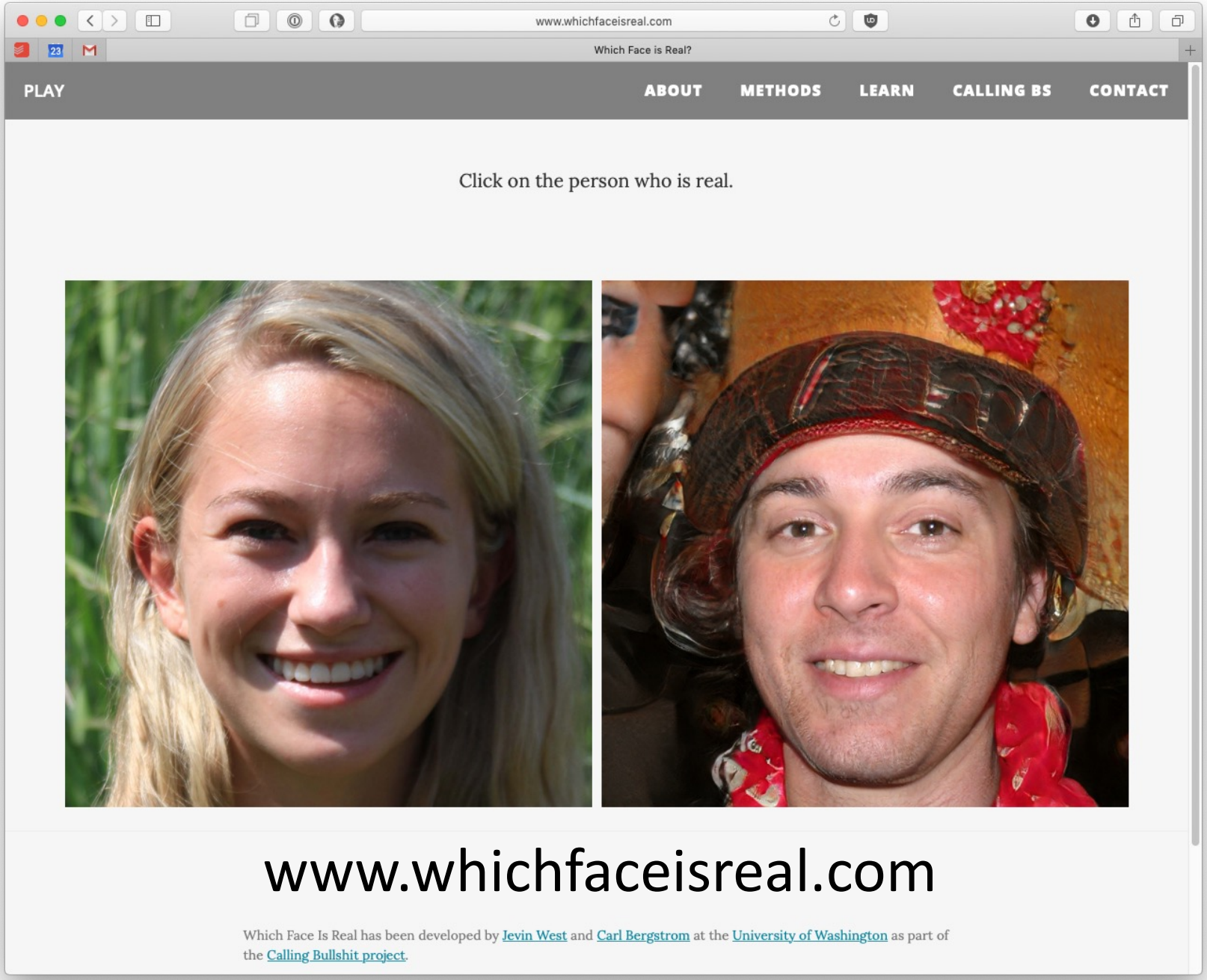


AI-Generated Images









www.whichfaceisreal.com

Deep Fakes

- Train a neural network to produce variants of one person's face
- Automatically map that onto videos with another person's face

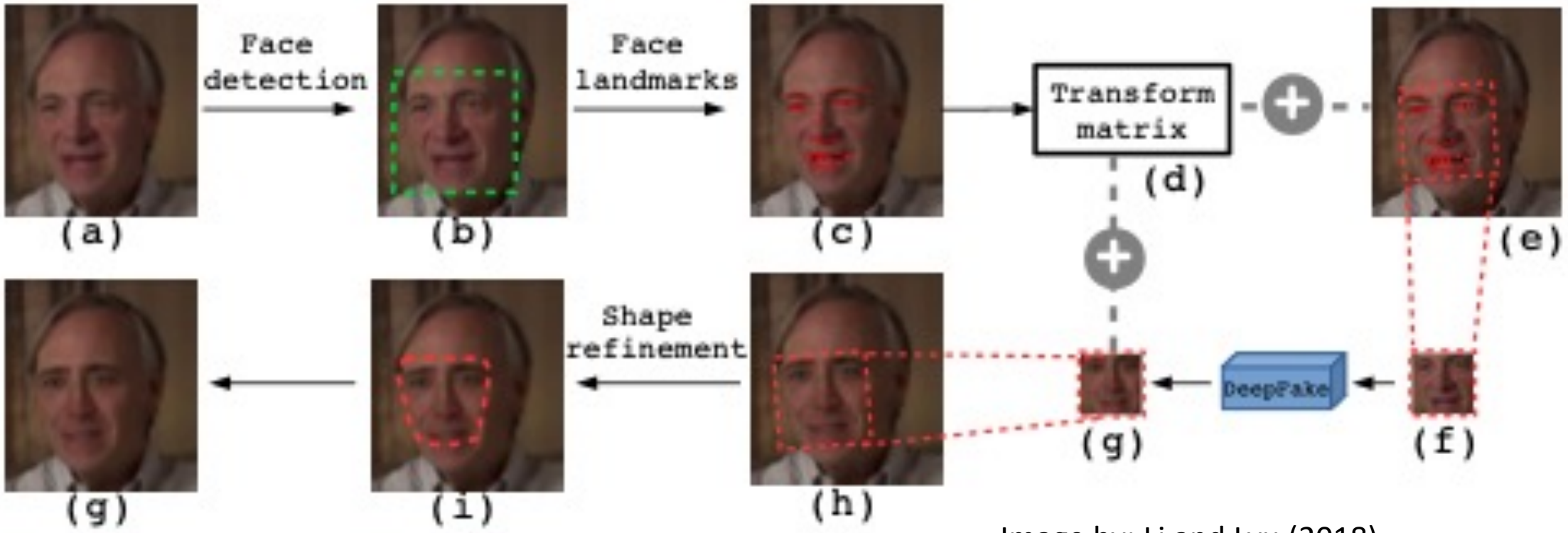


Image by: Li and Lyu (2018)

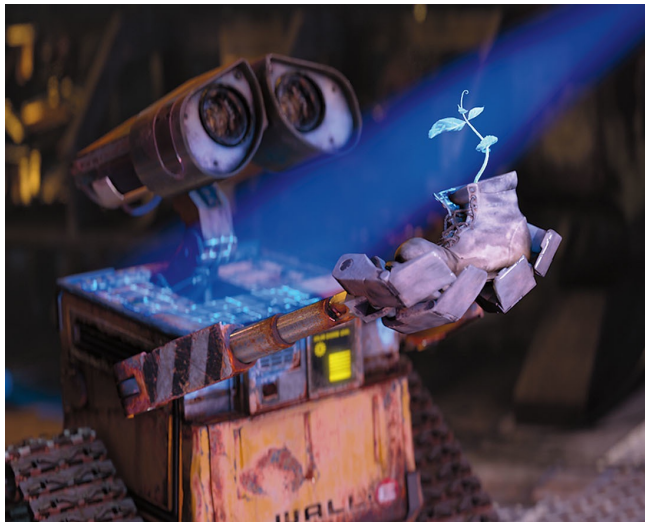
Detecting Deep Fakes

- Deconvolutions produce checkerboard patterns. (https://distill.pub/2016/deconv-checkerboard/)
- Deep fake videos often lack key human characteristics, such as a pulse or blinking. (Li and Lyu, 2018a)
- Deep fake videos have image warping artifacts. (Li and Lyu, 2018b)

...but will future Deep Fake variants have these same limitations?



Conclusion: Will A Human l Intelligence Save Us … OR … Kill Us All?





"I'm sorry, Dave."

- **AI is a tool:** AI is made by humans, used by humans, and works with humans.
- **Policy is essential:** At best, AI will follow human policy (for better or worse).
- **Understanding is essential:** AI is complex, but there are ways we can make it more robust and more intelligible, so we know when to trust it.