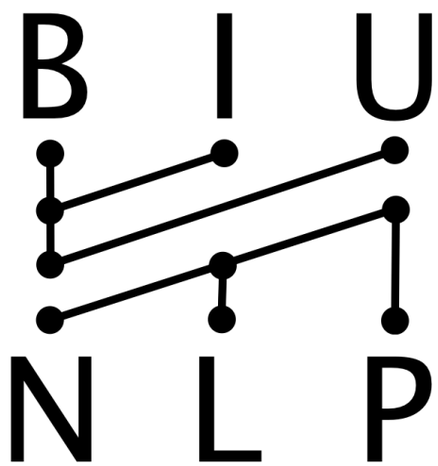
Embeddings Projections / alignment

Adversarial Examples
Adversarial Feature Learning

Ethics

Yoav Goldberg



About the exam

Exam

- 3 hours (maybe not all of it will be needed)

- Open materials

everything said in class is fair game.

and also elaborations of the ideas in class.

Exam

- You need to be able to read and write ML models as precise mathematical formulas.

- You need to be able to suggest modification to models, and express them in writing / math.

- Calculate sizes of networks, rough estimate on runtime (which operations are expensive which are cheap), number of parameters of different models.

Exam

- "Will this network run faster on a CPU or a GPU?"

- Which changes to this network will make it work faster without sacrificing much power?

Exam

- You need to be able to give names to equation.
  - "this is a differentiable gate".
  - "this is an MLP".
  - "this is attention".
  - "this is regularization".
  ...

- You need to be able to explain concepts ("what is dropout?")

- "here are the equations of an LSTM. Modify them to add regularization on the bias term of the input gate".

Exam

- What can and cannot be optimized with gradient descent?

Exam

- Think about loss functions
- Think about what a given architecture or representation can and cannot do
- "which of these systems produced this output"?

Exam

- Be able to think critically about design choices and their implications.

Exam

- Window vs RNN vs CNN vs Dilated CNN vs...
- Pre-trained embeddings, how to use them?
- How to represent words?
- LSTM Transducer vs Seq2Seq
- Attention vs. no Attention
- Softmax+cross-entropy vs margin loss
- Different pre-training architectures / choices

Exam

- "someone suggested to take architecture X and replace equation number (4) with ....

is it a good/bad idea? why?
what will happen?
do you need to do additional changes for this to work?"

Exam

- What do learning curves mean?
- How would you change a system to perform better? to do X?
- "Here is a system with some parameters. Here is a learning curve. Here is another learning curve. Which parameters did we change to get to the new curve?"

Embedding Projections

Motivation - Mikolov et al. 2013

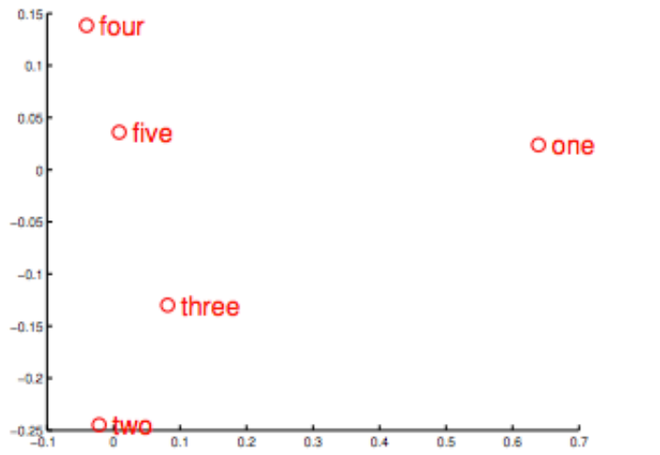
(also, Haghighi and Klein, 2008)

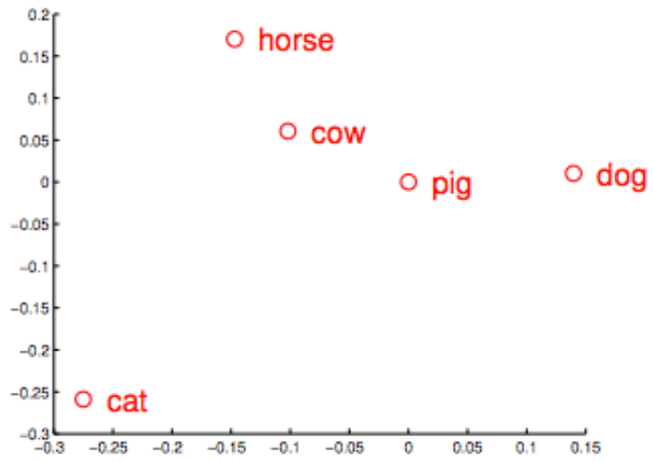
- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)





Motivation - Mikolov et al. 2013

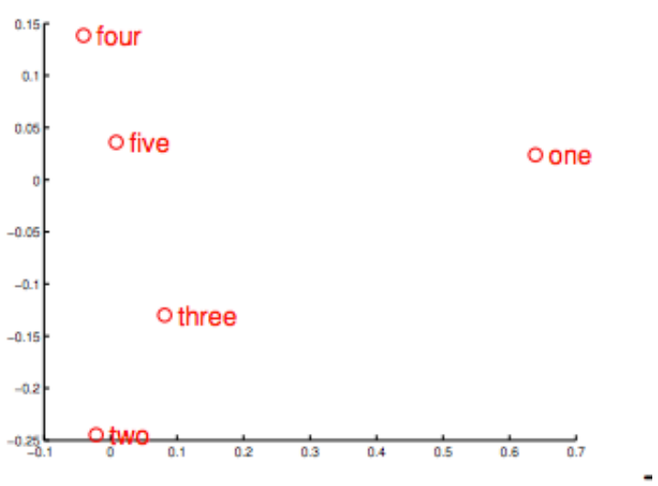
(also, Haghighi and Klein, 2008)

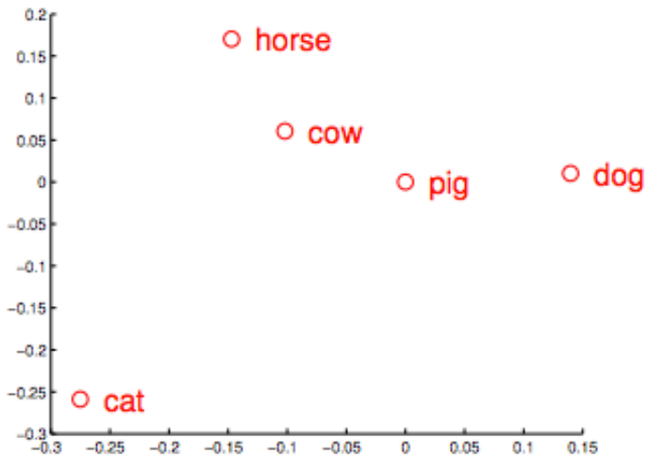
- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

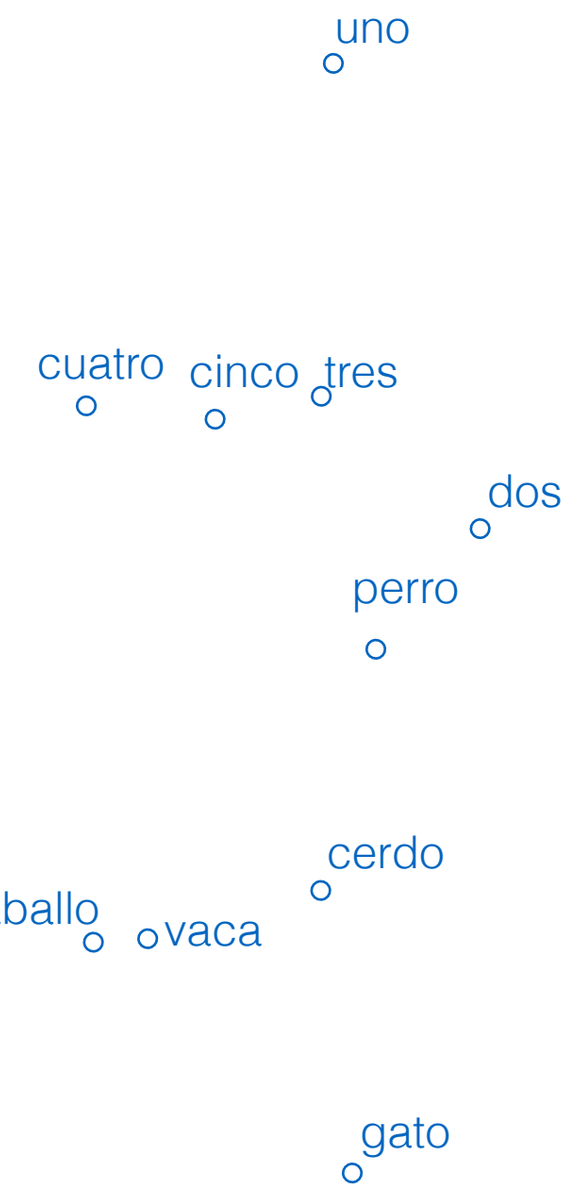
- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)







Motivation - Mikolov et al. 2013

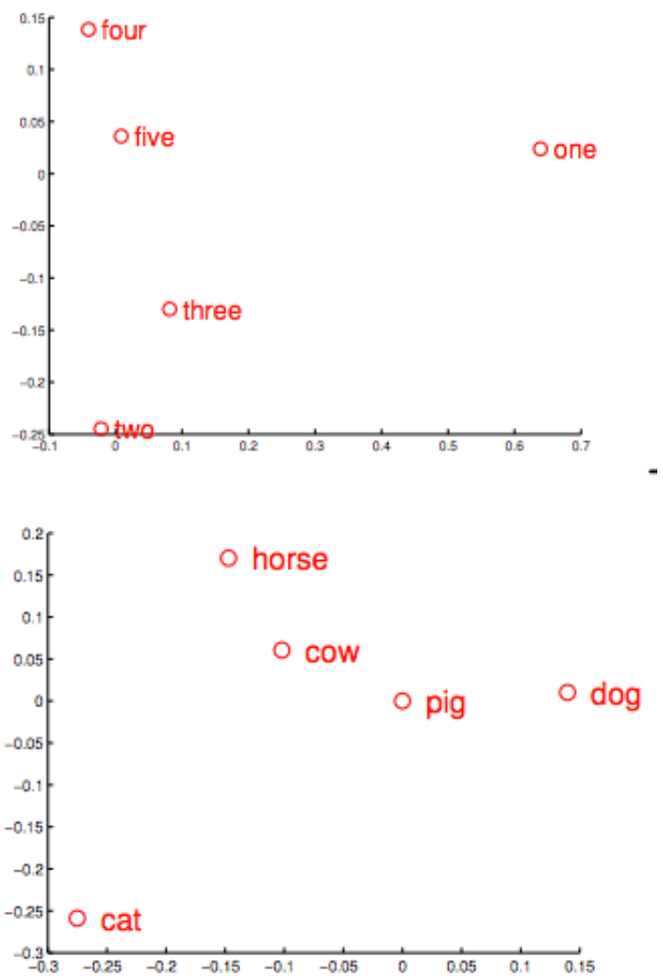
(also, Haghighi and Klein, 2008)

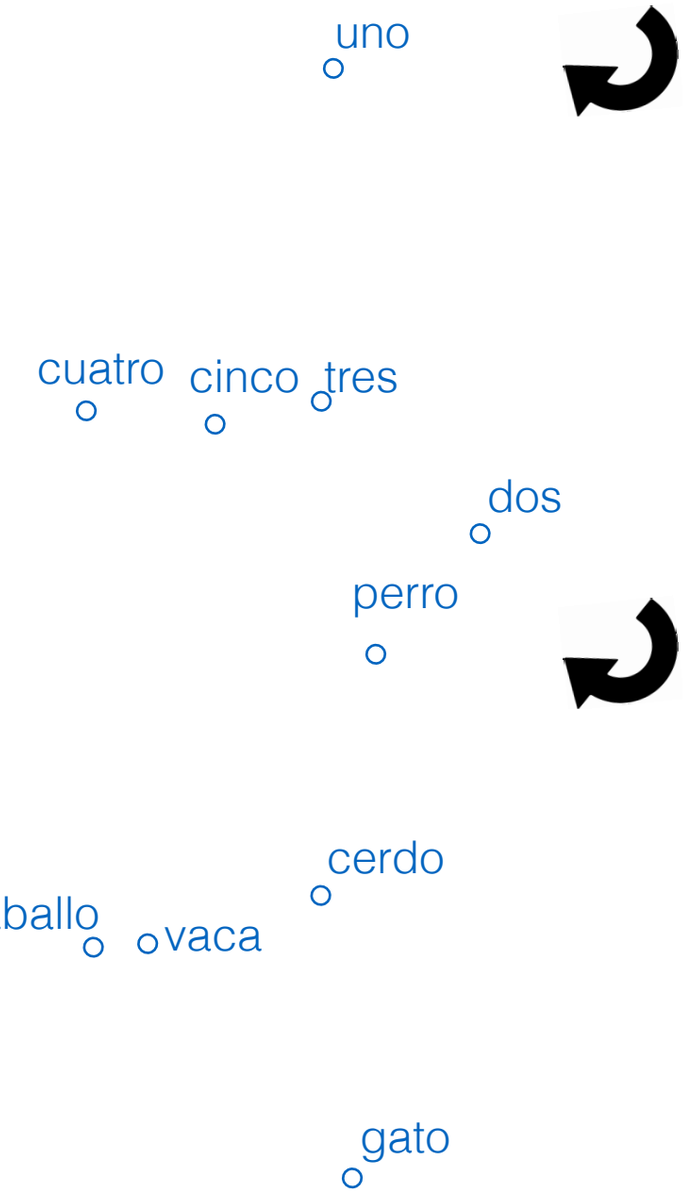
- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)





Motivation - Mikolov et al. 2013

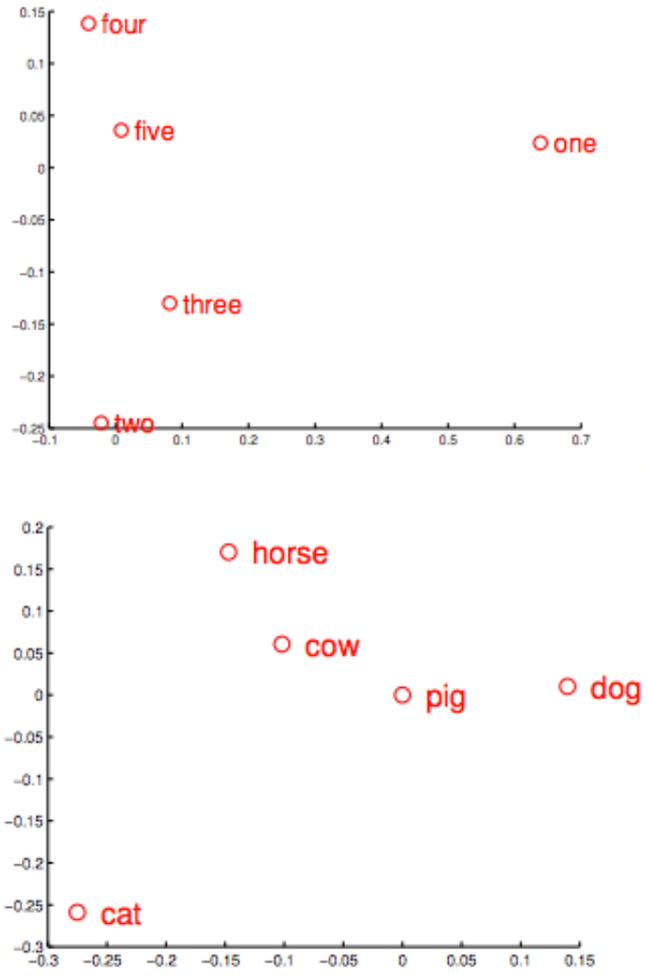
(also, Haghighi and Klein, 2008)

- “Exploiting Similarities among Languages for Machine Translation” - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)



Motivation - Mikolov et al. 2013

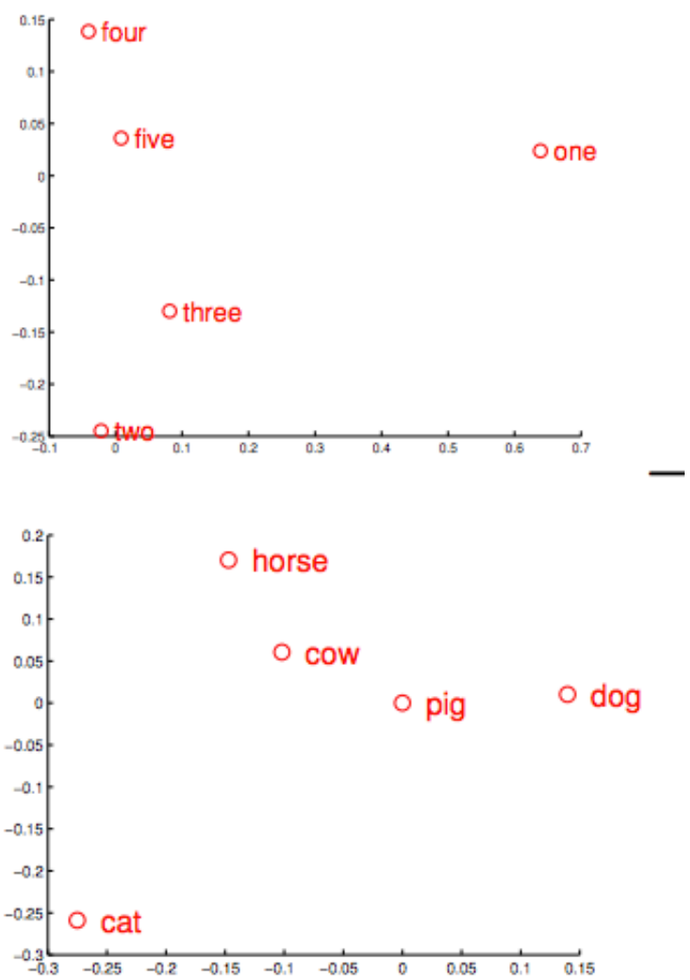
(also, Haghighi and Klein, 2008)

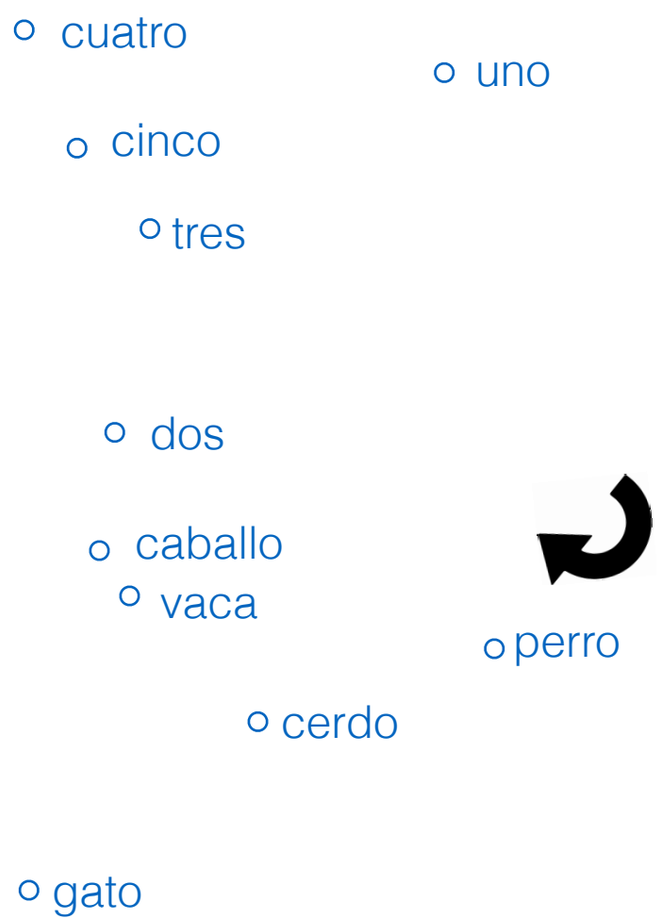
- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)





Motivation - Mikolov et al. 2013

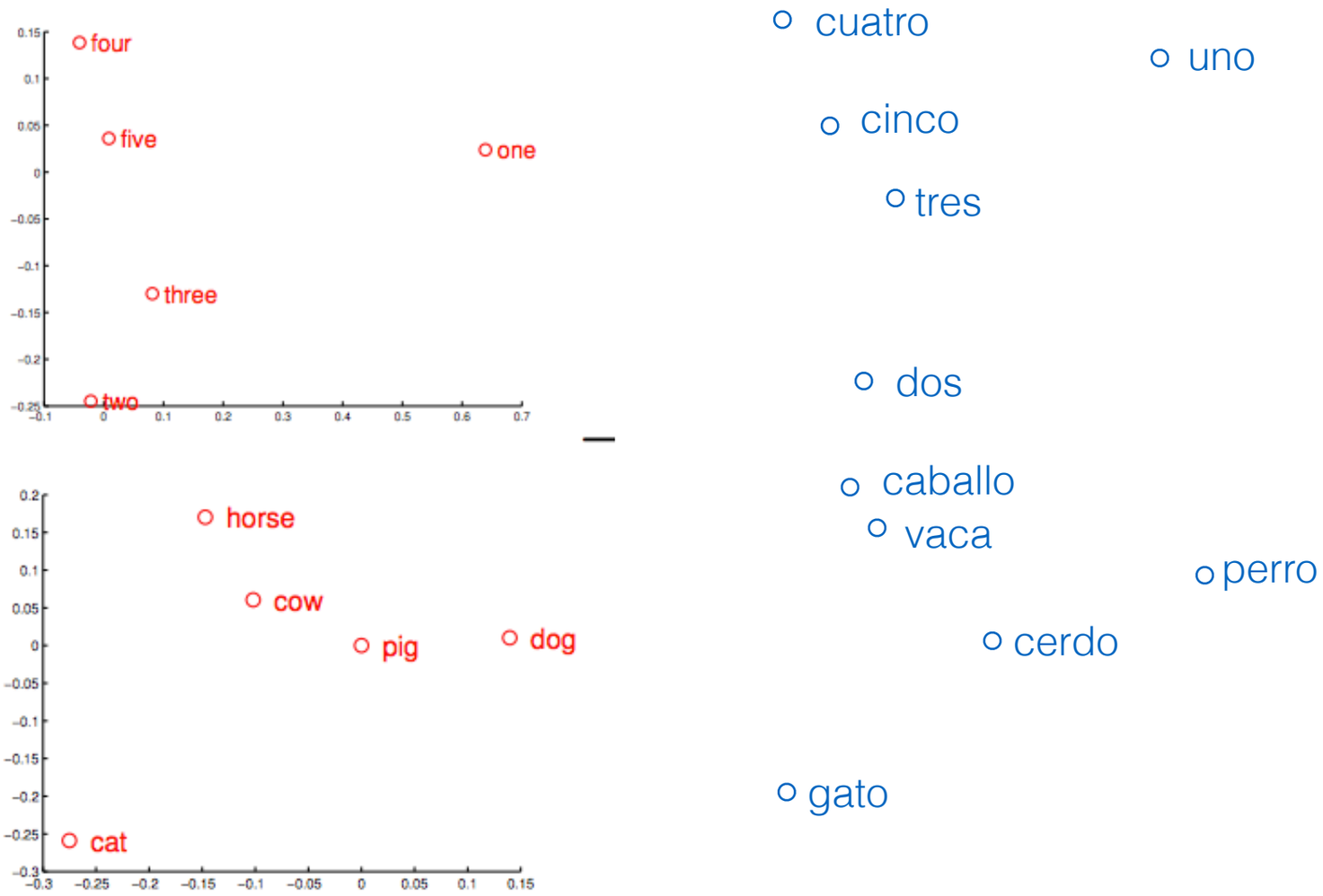
(also, Haghighi and Klein, 2008)

- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)



Motivation - Mikolov et al. 2013

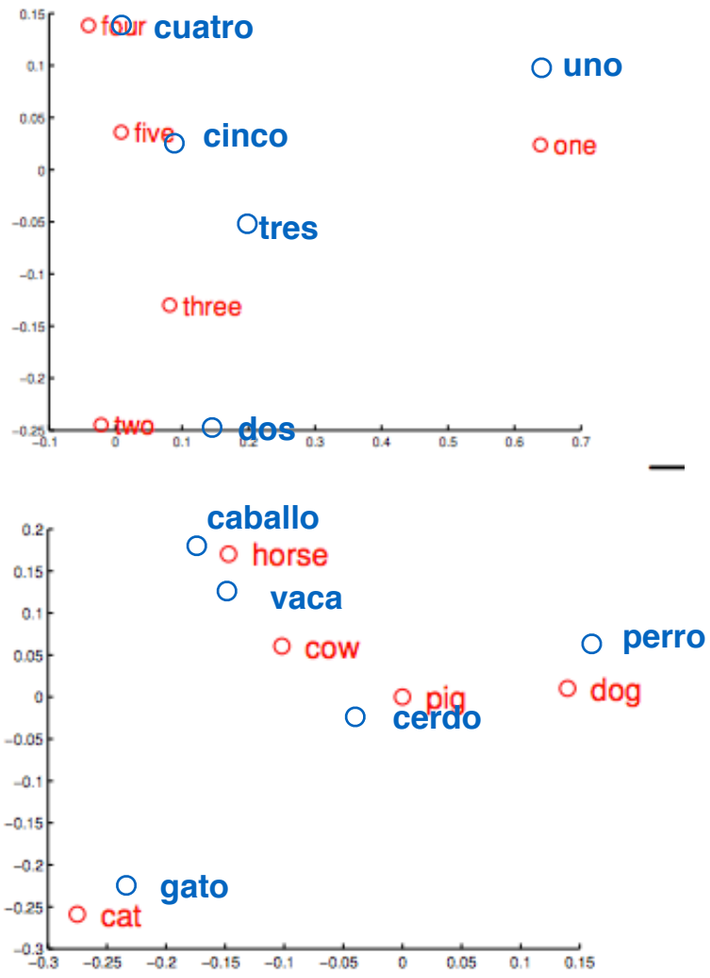
(also, Haghighi and Klein, 2008)

- "Exploiting Similarities among Languages for Machine Translation" - Mikolov, Le & Sutskever, 2013

- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation

- Learned a rotation matrix to translate words from one embedding space to another with some success

- Weakly supervised - requires a small dictionary (5000 entries)



Motivation - Mikolov et al. 2013

- Learned a rotation matrix to translate words from one embedding space to another with some success

$$\arg\min_{M} \sum_{x_i, y_i \in pairs} ||Mx_i - y_i||_2^2$$

Motivation - Mikolov et al. 2013

- Learned a rotation matrix to translate words from one embedding space to another with some success

$$\arg\min_{M} \sum_{x_i, y_i \in pairs} ||Mx_i - y_i||_2^2$$

$$= \arg\min_{M} ||ME^1 - E^2||_2^2$$

- Artetxe, Labake & Agirre, ACL 2017:
  - Use **numbers** as the initial pivot items.
  - Do it an an iterative procedure.

$$\arg\min_{M} \sum_{x_i, y_i \in pairs} ||Mx_i - y_i||_2^2$$

$$= \arg\min_{M} ||ME^1 - E^2||_2^2$$

- Artetxe, Labake & Agirre, ACL 2017:
  - Use **numbers** as the initial pivot items.
  - Do it an an iterative procedure.

$$\arg\min_{M} \sum_{x_i, y_i \in pairs} ||Mx_i - y_i||_2^2$$

$$= \arg\min_{M} ||ME^1 - E^2||_2^2$$

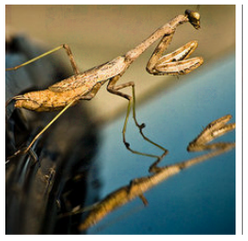
(also, can be solved exactly with SVD)

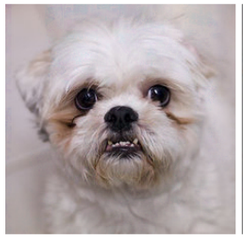
Beyond language-to-language

- Words from 1900 to words in 1990
- Words from young speakers to old speakers
- Words from left-wing to right-wing writers
- ....

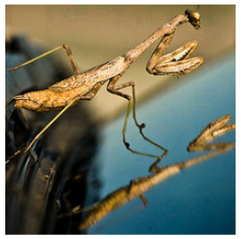
and now for something completely different

Adversarial Examples

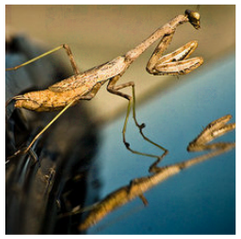




speakers

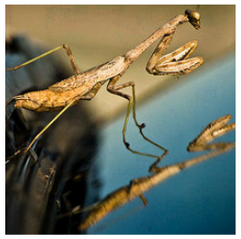


speakers



grasshopper

speakers



grasshopper

poodle

school-bus

...

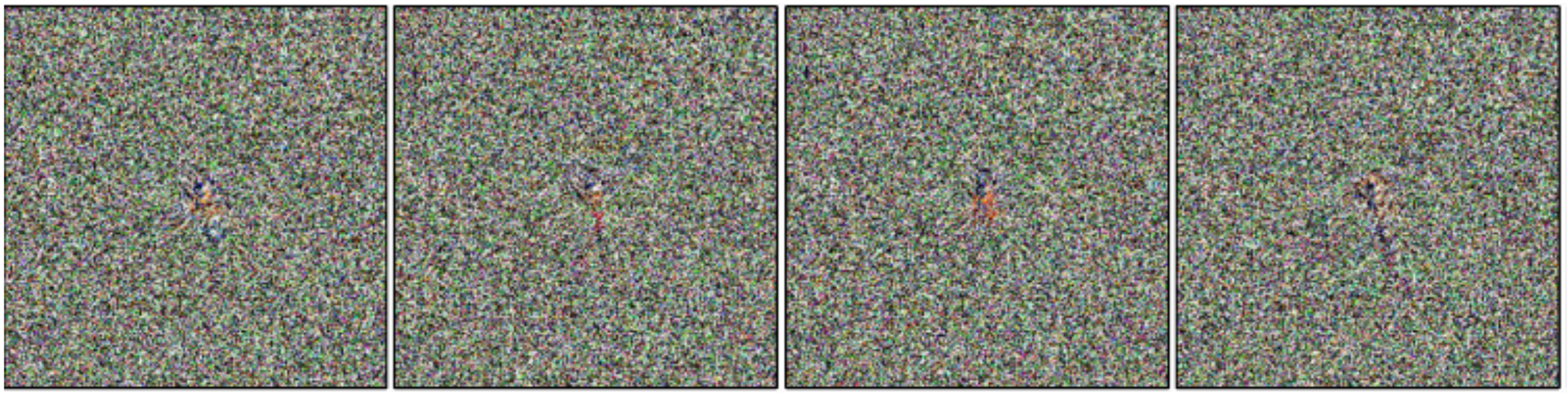
but also

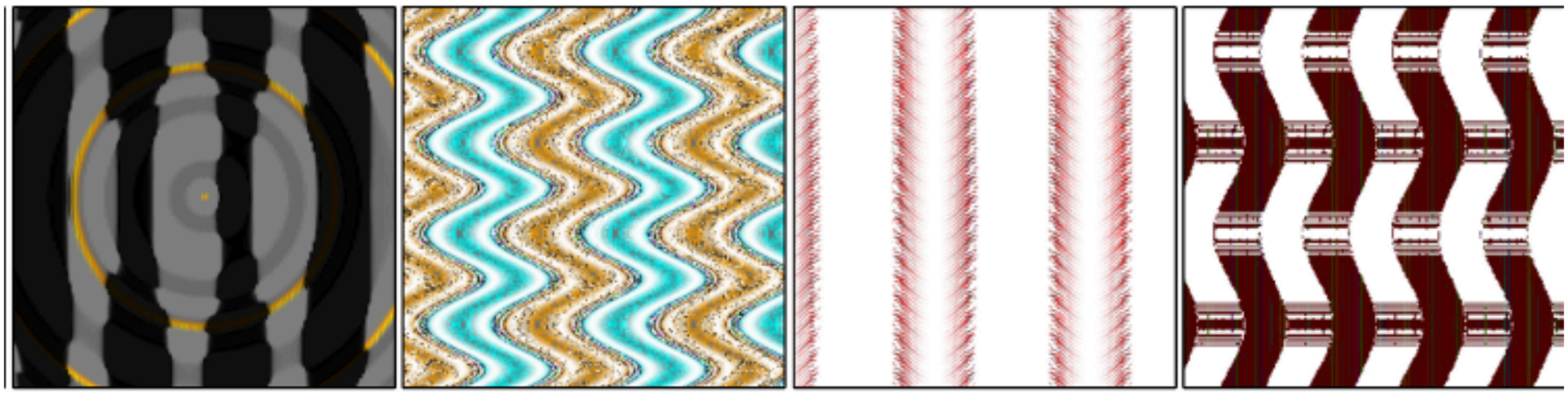
Deep Neural Networks are Easily Fooled:
High Confidence Predictions for Unrecognizable Images

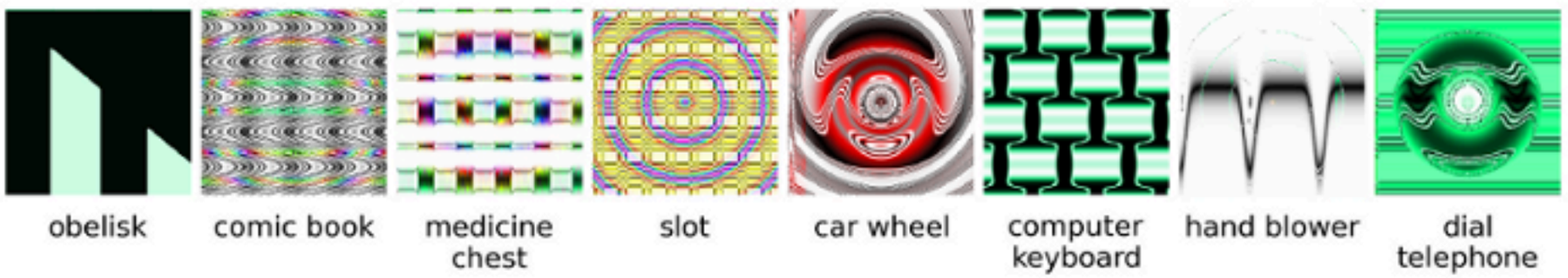
Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

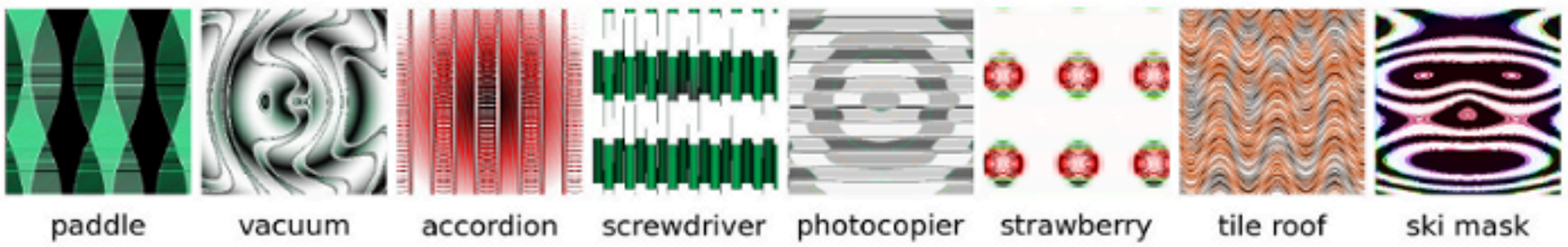
Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

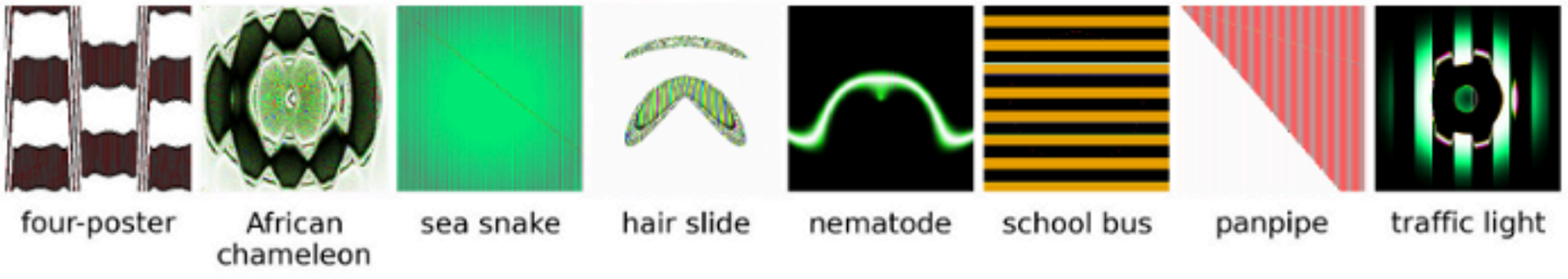
Jeff Clune
University of Wyoming
jeffclune@uwyo.edu


brambling
redshank
robin
cheetah
armadillo
lesser panda
centipede
jackfruit


king penguin
starfish
baseball
electric guitar
freight car
remote control
peacock
African grey


obelisk
comic book
medicine chest
slot
car wheel
computer keyboard
hand blower
dial telephone
assault rifle
stethoscope
digital clock
soccer ball
bagel
pinwheel
crossword puzzle
punching bag

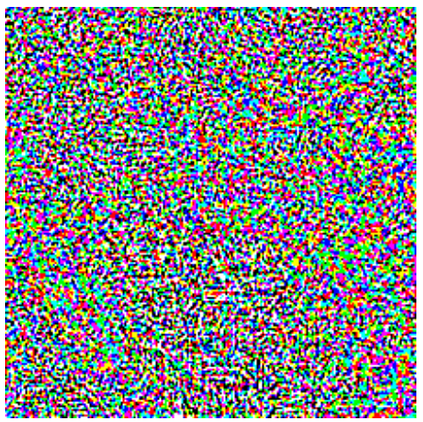

paddle
vacuum
accordion
screwdriver
photocopier
strawberry
tile roof
ski mask


four-poster
African chameleon
sea snake
hair slide
nematode
school bus
panpipe
traffic light
projector
pole
spotlight
green snake
trifle
volcano
chainlink fence
monarch

$\boldsymbol{x}$

“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”
99.3 % confidence

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

Intriguing properties of neural networks

Christian Szegedy
Google Inc.

Wojciech Zaremba
New York University

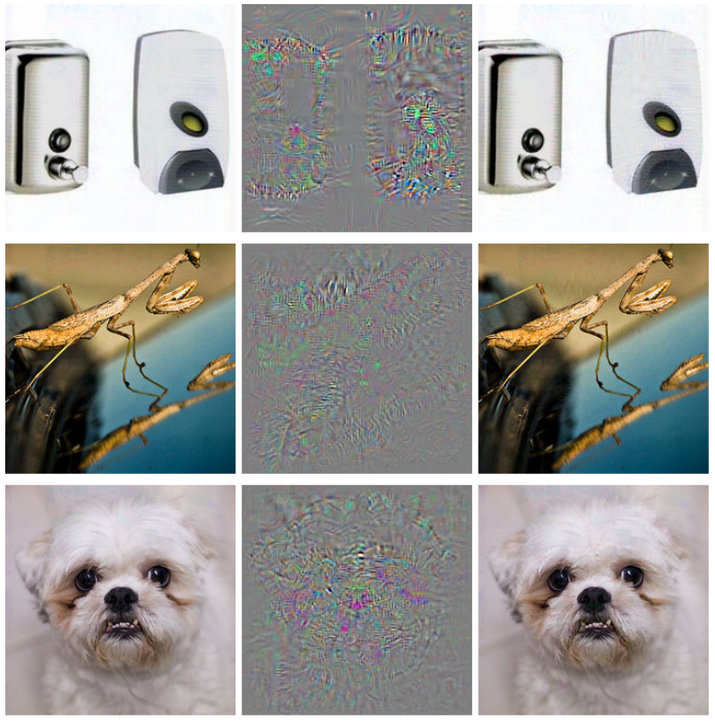
Ilya Sutskever
Google Inc.

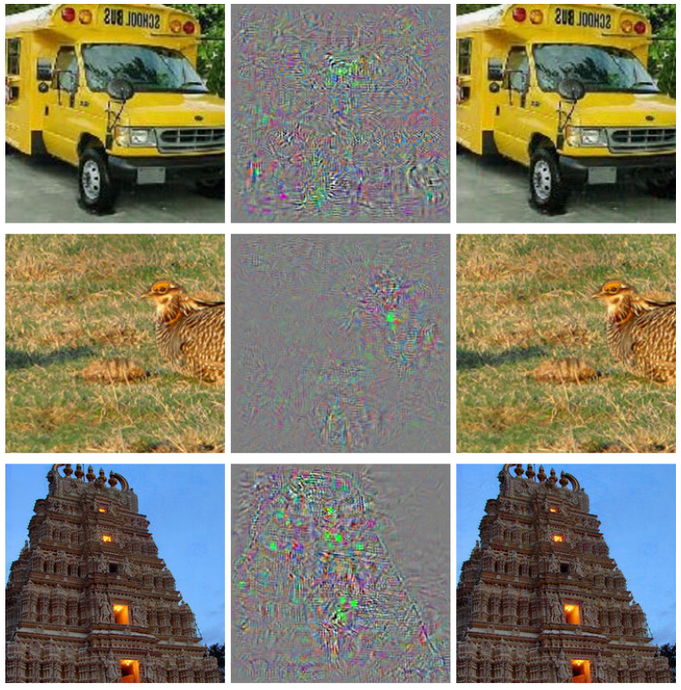
Joan Bruna
New York University

Dumitru Erhan
Google Inc.

Ian Goodfellow
University of Montreal

Rob Fergus
New York University
Facebook Inc.





all of the images
on the right
are categorized as
"ostrich"

Intriguing properties of neural networks

Christian Szegedy
Google Inc.

Wojciech Zaremba
New York University

Ilya Sutskever
Google Inc.

Joan Bruna
New York University

Dumitru Erhan
Google Inc.

Ian Goodfellow
University of Montreal

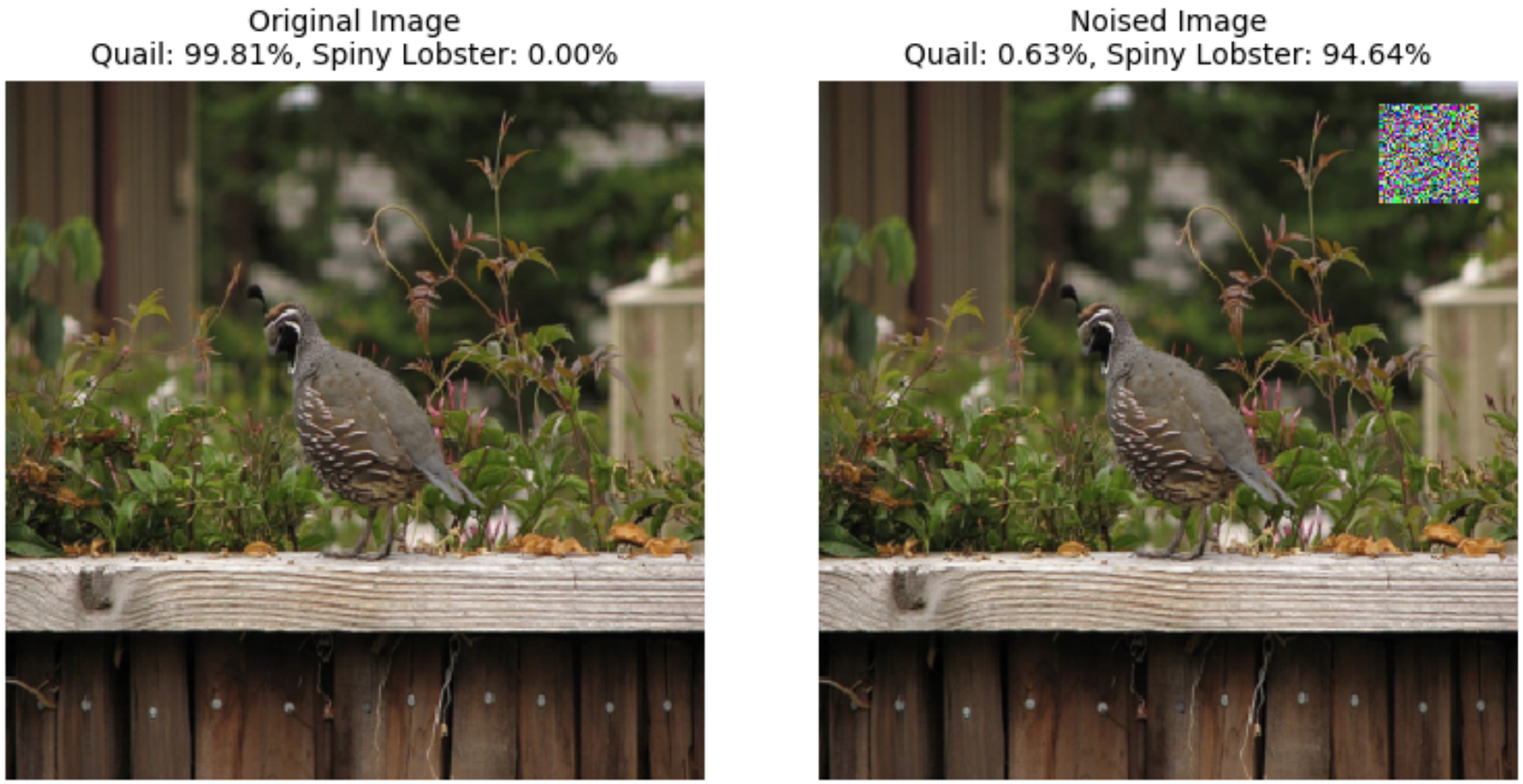
Rob Fergus
New York University
Facebook Inc.

Visible Localized Noise

LaVAN: Localized and Visible Adversarial Noise

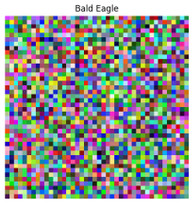
Danny Karmon[1] **Daniel Zoran**[2] **Yoav Goldberg**[1]
[1] Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel.
[2] DeepMind, London, UK.
karmond@biu.ac.il, danielzoran@google.com, yogo@cs.biu.ac.il

Visible Localized Noise

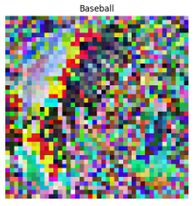


Quail (99.8%) $\rightarrow$ Spiny Lobster (94.6%)

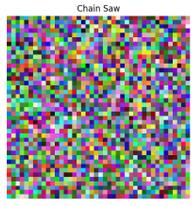
Visible Localized Noise



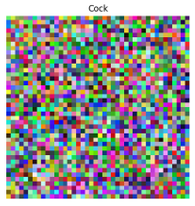
Bald Eagle



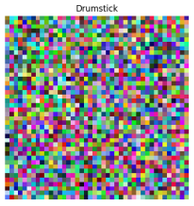
Baseball



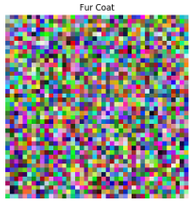
Chain Saw



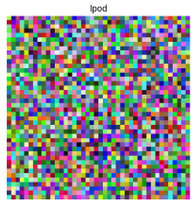
Cock



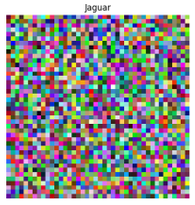
Drumstick



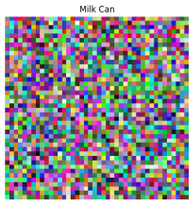
Fur Coat



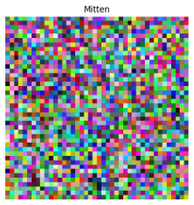
iPod



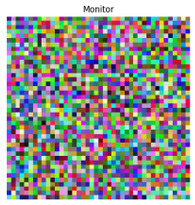
Jaguar



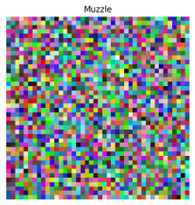
Milk Can



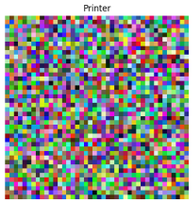
Mitten



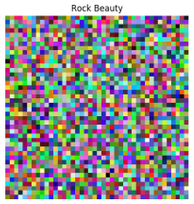
Monitor



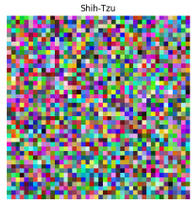
Muzzle



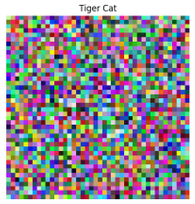
Printer



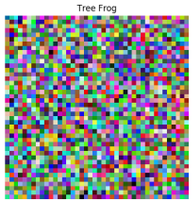
Rock Beauty (Fish)



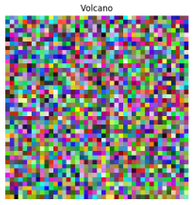
Shih-Tzu (Dog)



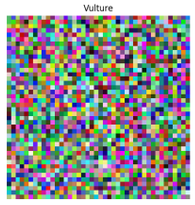
Tiger Cat



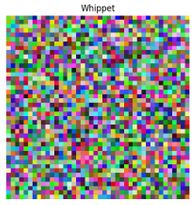
Tree Frog



Volcano



Vulture



Whippet (Dog)

Visible Localized Noise

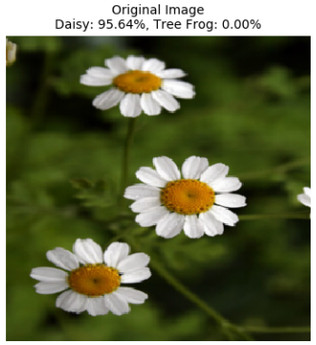
Padlock (92.7%)

Tiger Cat (94.4%)

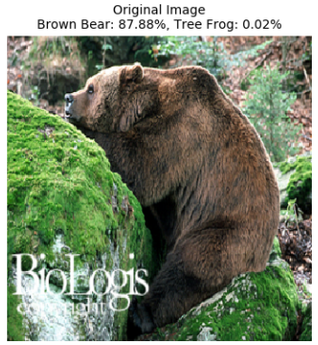
Car Mirror (94.5%)

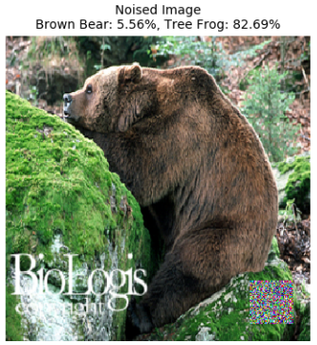
Stingray (90.5%)

Visible Localized Noise



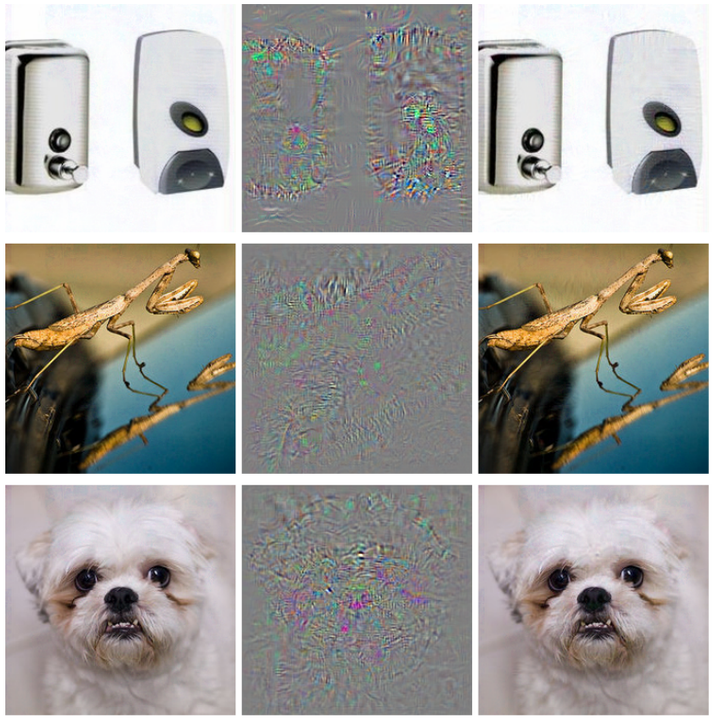
Daisy (95.6%) → **Tree Frog** (96.4%)

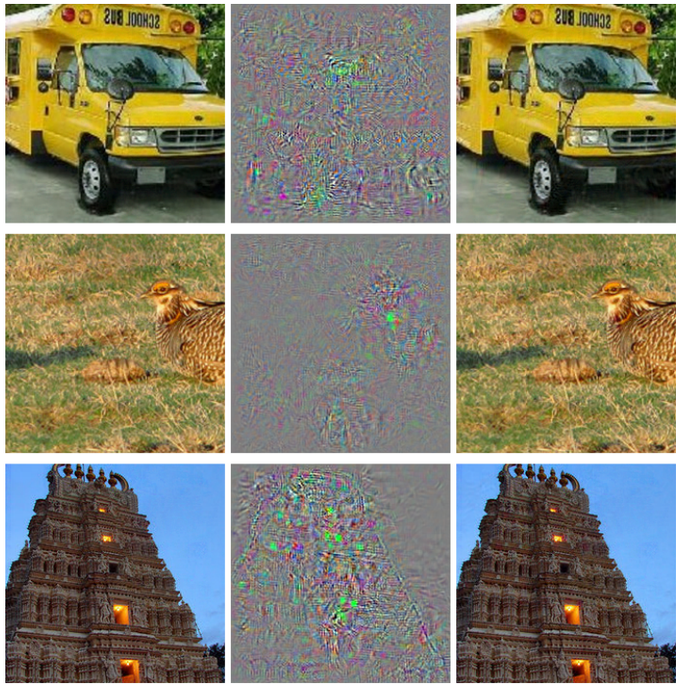




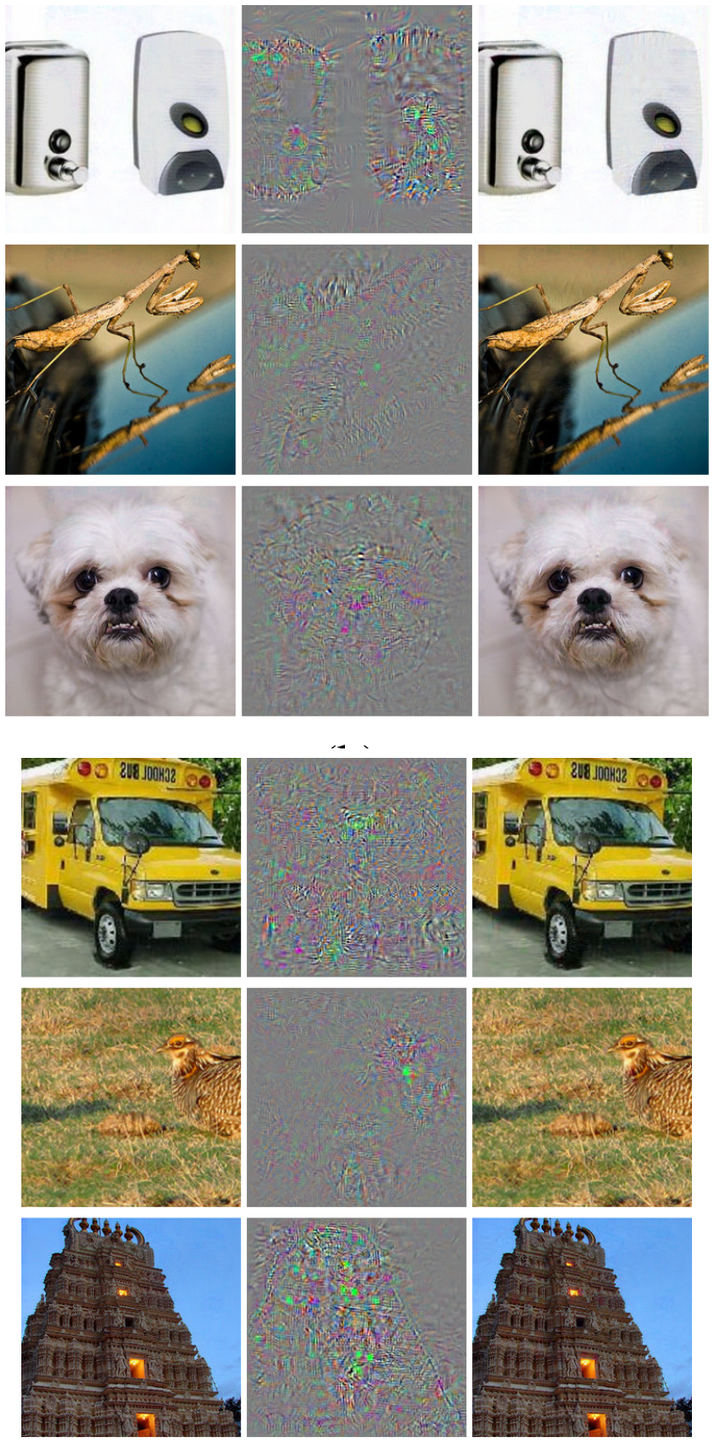
Brown Bear (87.9%) → **Tree Frog** (82.7%)

Minivan (90.7%) → **Tree Frog** (86.4%)





how does one
find the misleading images?

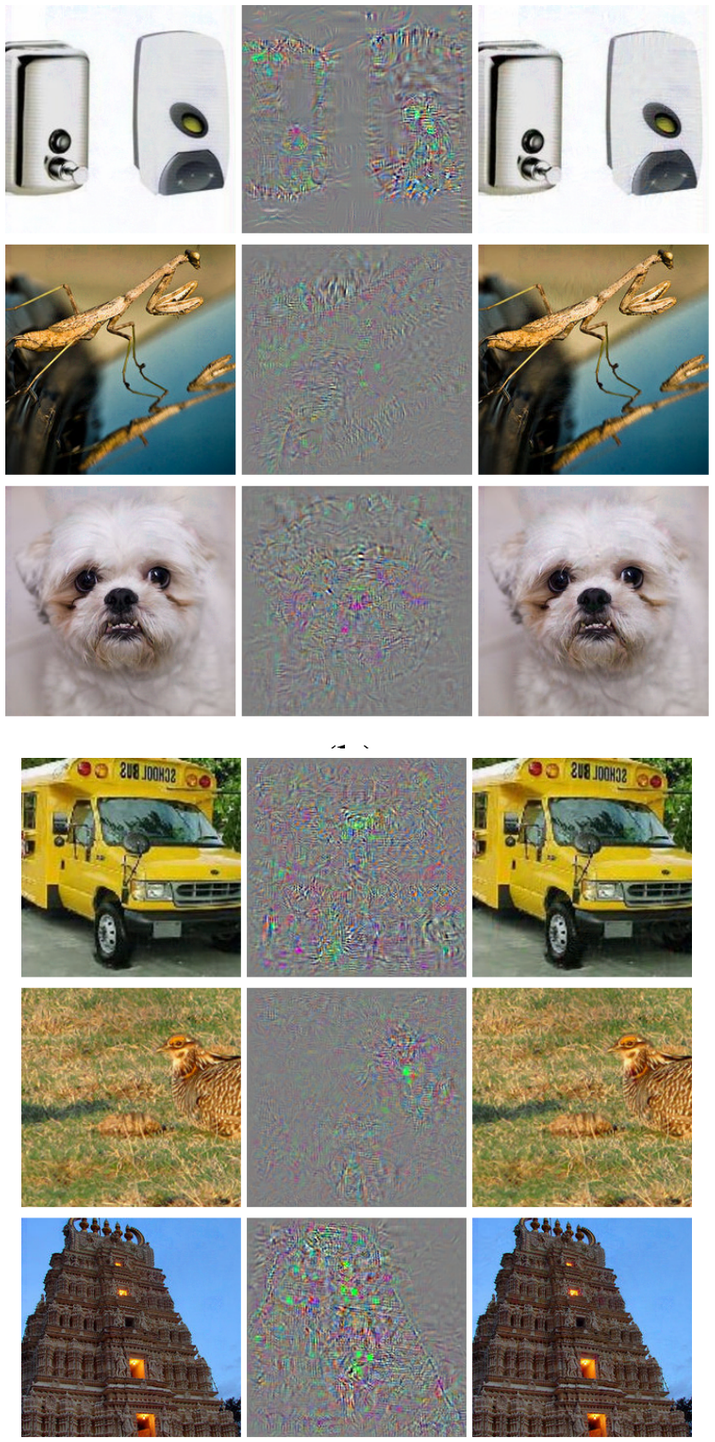


how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

training:
gradient decent on parameters
to get correct predictions.



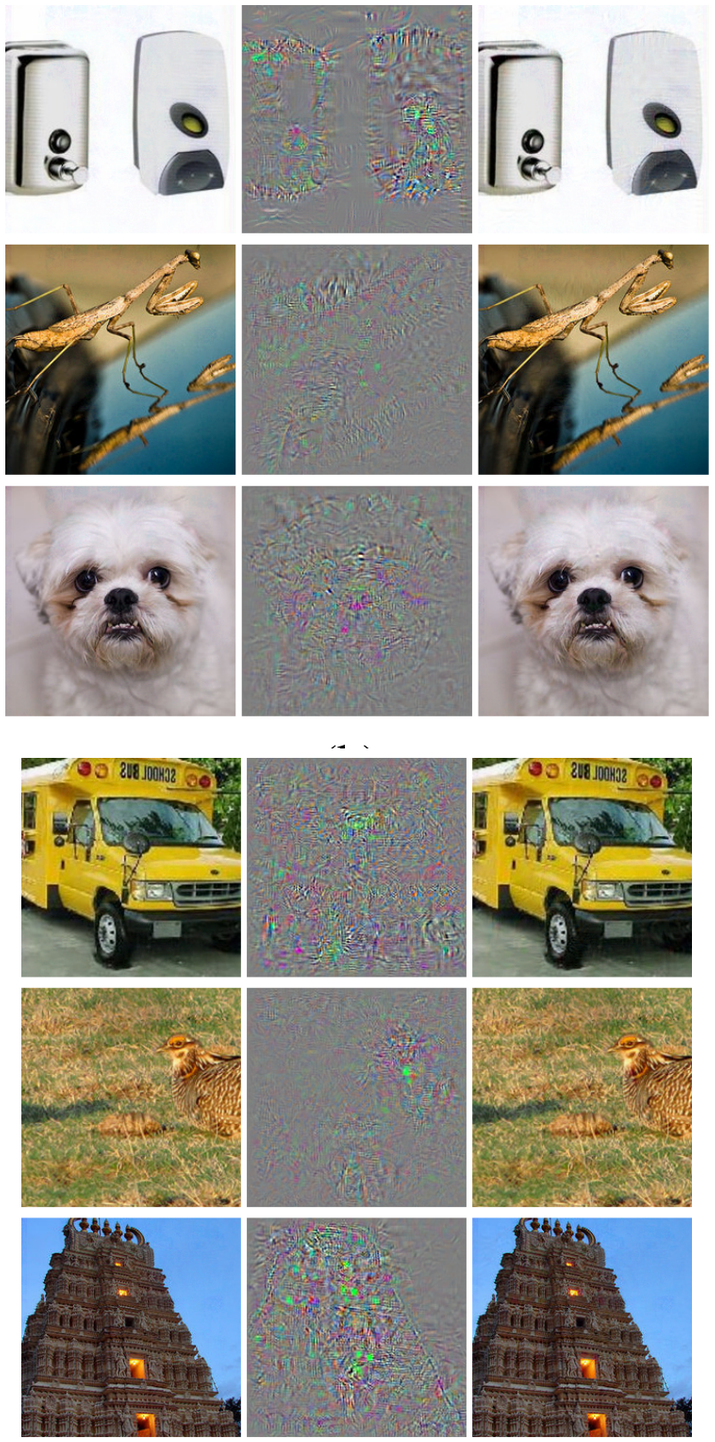
change params
to get desired y
for given x

how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

training:
gradient decent on parameters
to get correct predictions.



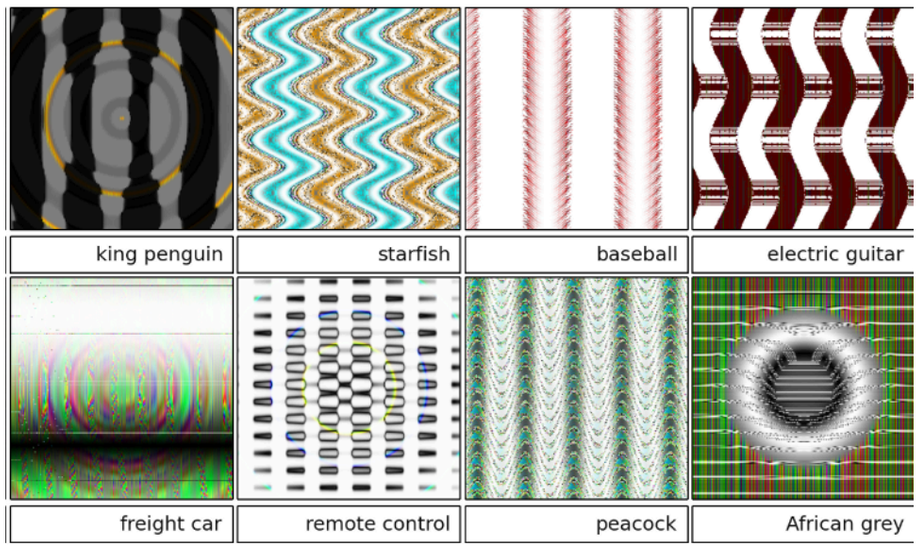
how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

training:
gradient decent on parameters
to get correct predictions.

adversarial:
parameters fixed.
gradient descent on input
to get desired prediction



change x
to get desired y
for given params

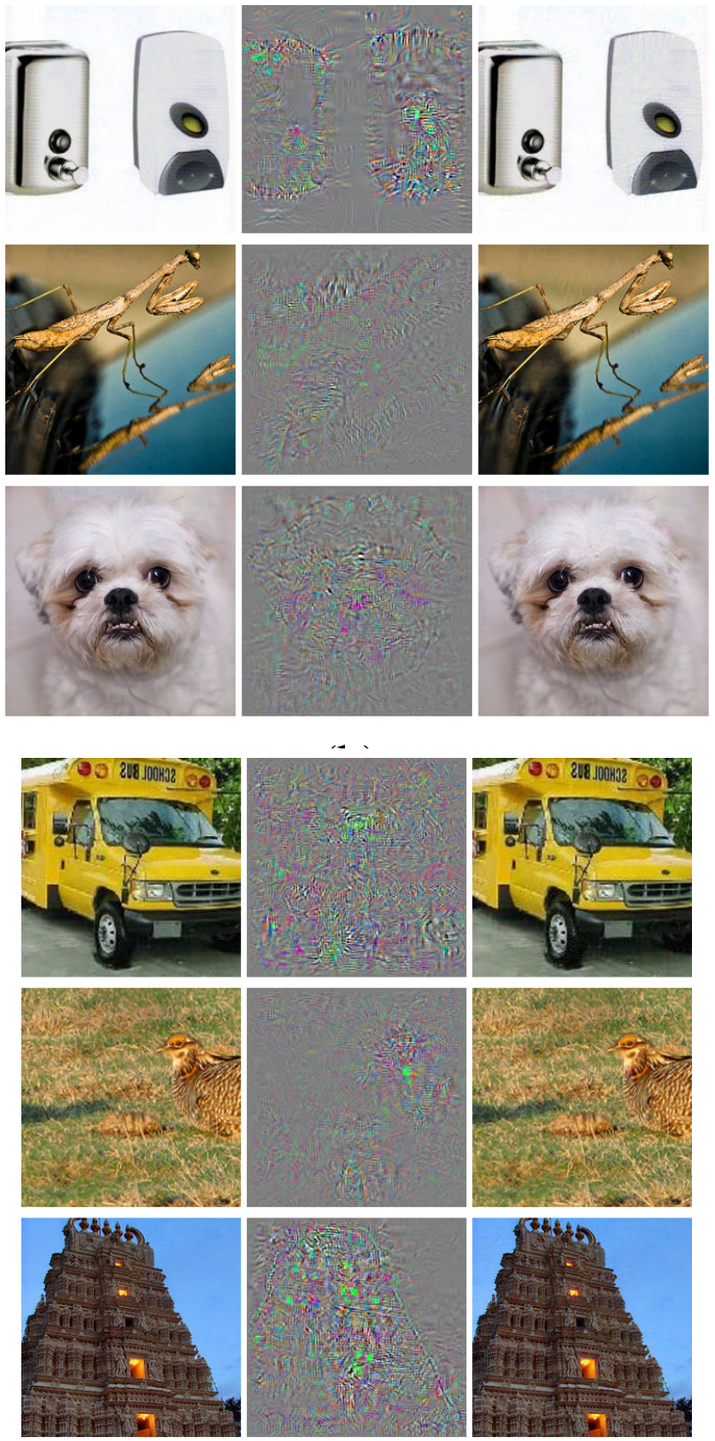
how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

training:
gradient decent on parameters
to get correct predictions.

adversarial:
parameters fixed.
gradient descent on input
to get desired prediction



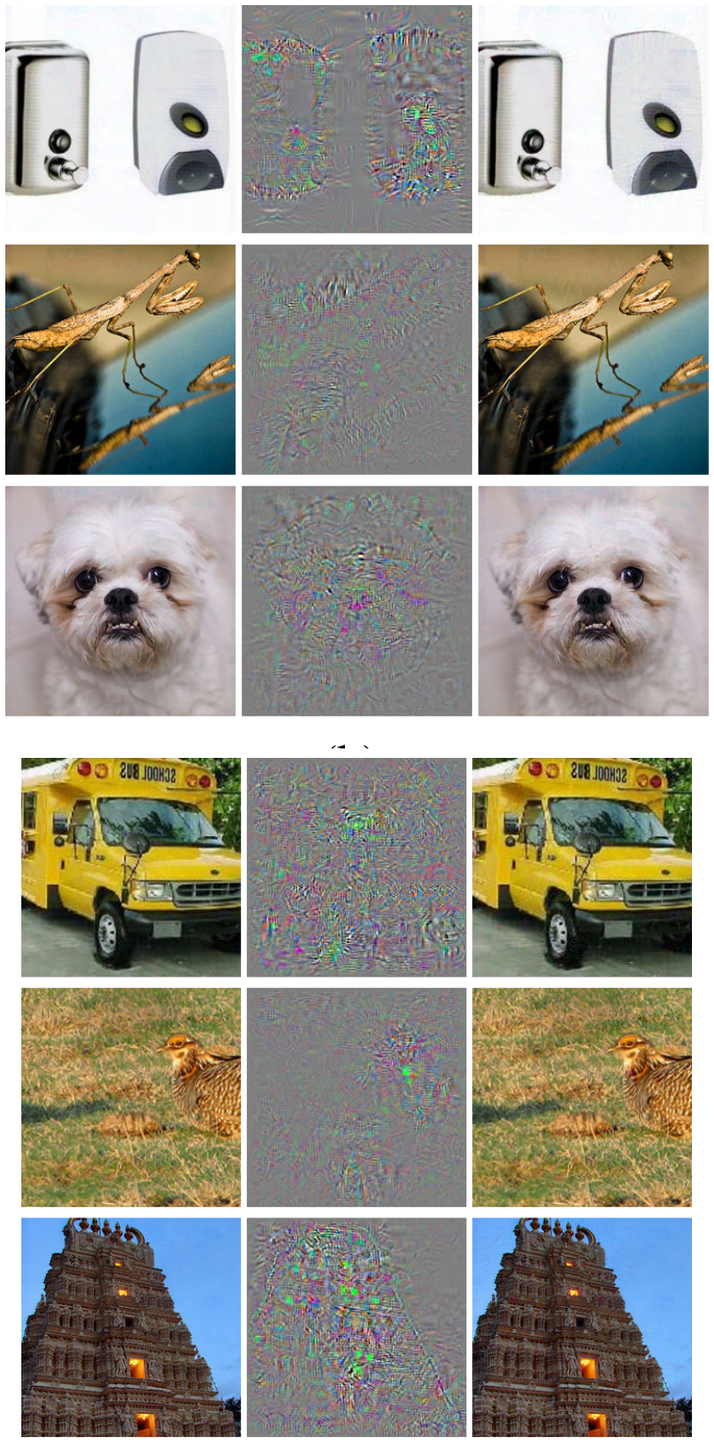
how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x} + \delta, \theta)$$

training:
gradient decent on parameters
to get correct predictions.

adversarial:
parameters fixed.
gradient descent on input
to get desired prediction



find small delta
to get desired y
for given params

how does one
find the misleading images?

neural network

$$\mathbf{y} = f(\mathbf{x} + \delta, \theta)$$

training:
gradient decent on parameters
to get correct predictions.

adversarial:
parameters fixed.
gradient descent on input
to get desired prediction

Adversarial Examples: Interesting open problems

- **How to train a network which is robust to adversarial attacks.**

- How to identify adversarial examples?

- How to generate adversarial examples for sequences?

Adversarial Examples: Interesting open problems

- How to train a network which is robust to adversarial attacks.
- **How to identify adversarial examples?**
- How to generate adversarial examples for sequences?

Adversarial Examples: Interesting open problems

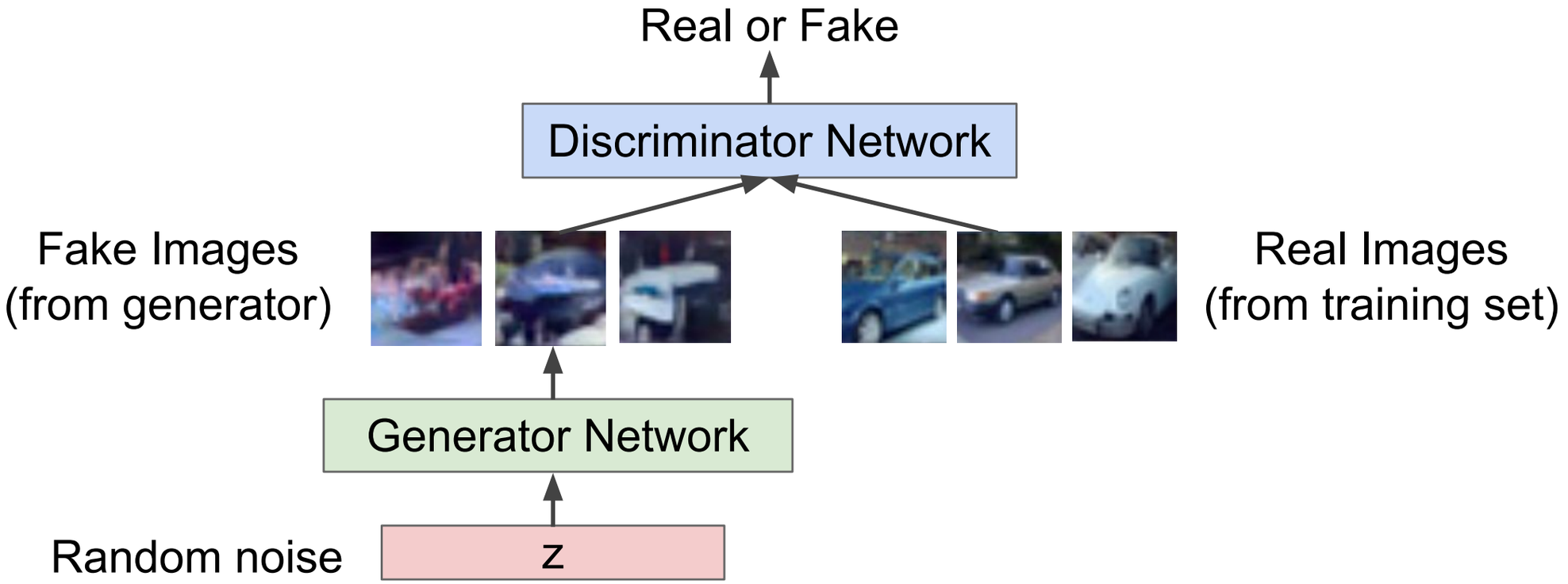
- How to train a network which is robust to adversarial attacks.
- How to identify adversarial examples?
- **How to generate adversarial examples for discrete sequences?**

Very brief intro: GANs

Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images



Fake and real images copyright Emily Denton et al. 2015. Reproduced with permission.

CS11-747 Neural Networks for NLP

Adversarial Methods

Graham Neubig

Carnegie Mellon University

Language Technologies Institute

Site

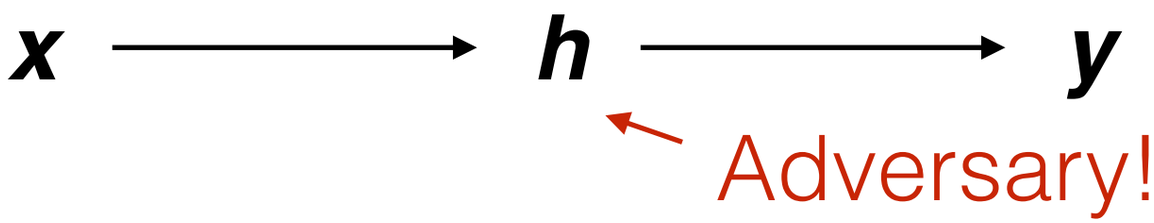
https://phontron.com/class/nn4nlp2017/

Adversarial Feature Learning

Adversaries over Features vs. Over Outputs

- Generative adversarial networks

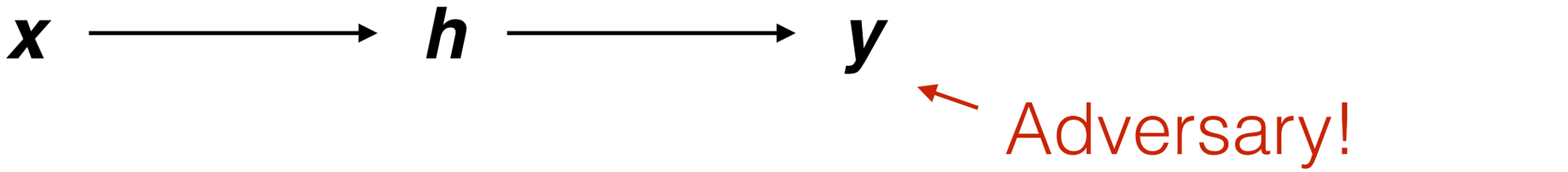
- Adversarial feature learning



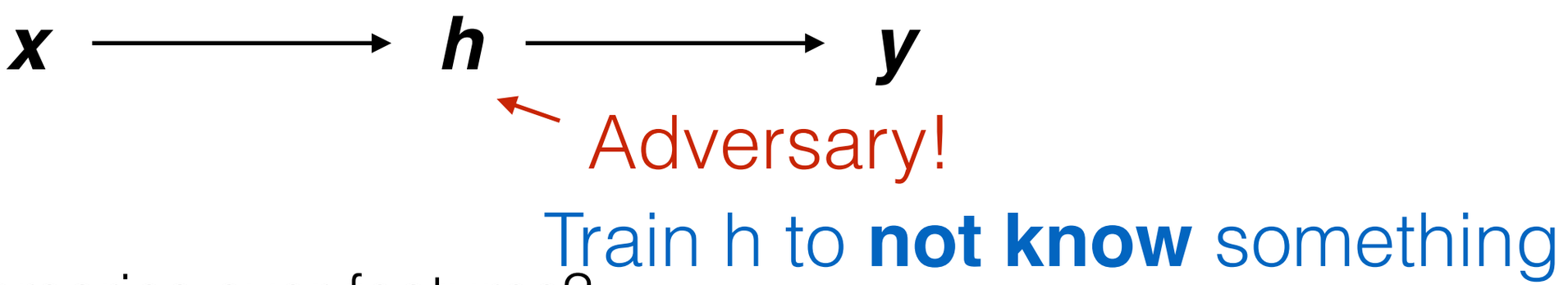
- Why adversaries over features?
  - Non-generative tasks
  - Continuous features easier than discrete outputs

Adversaries over Features vs. Over Outputs

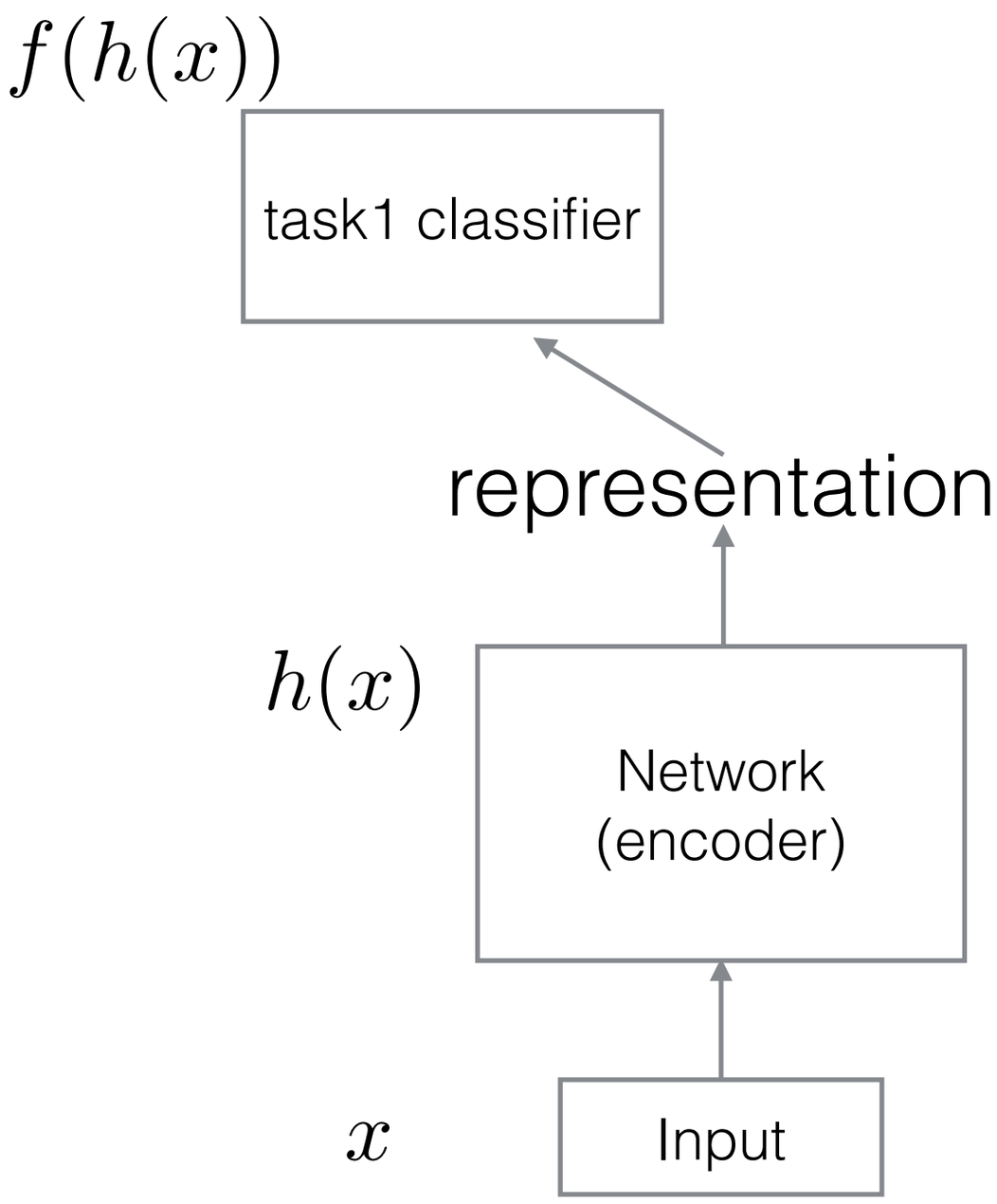
- Generative adversarial networks

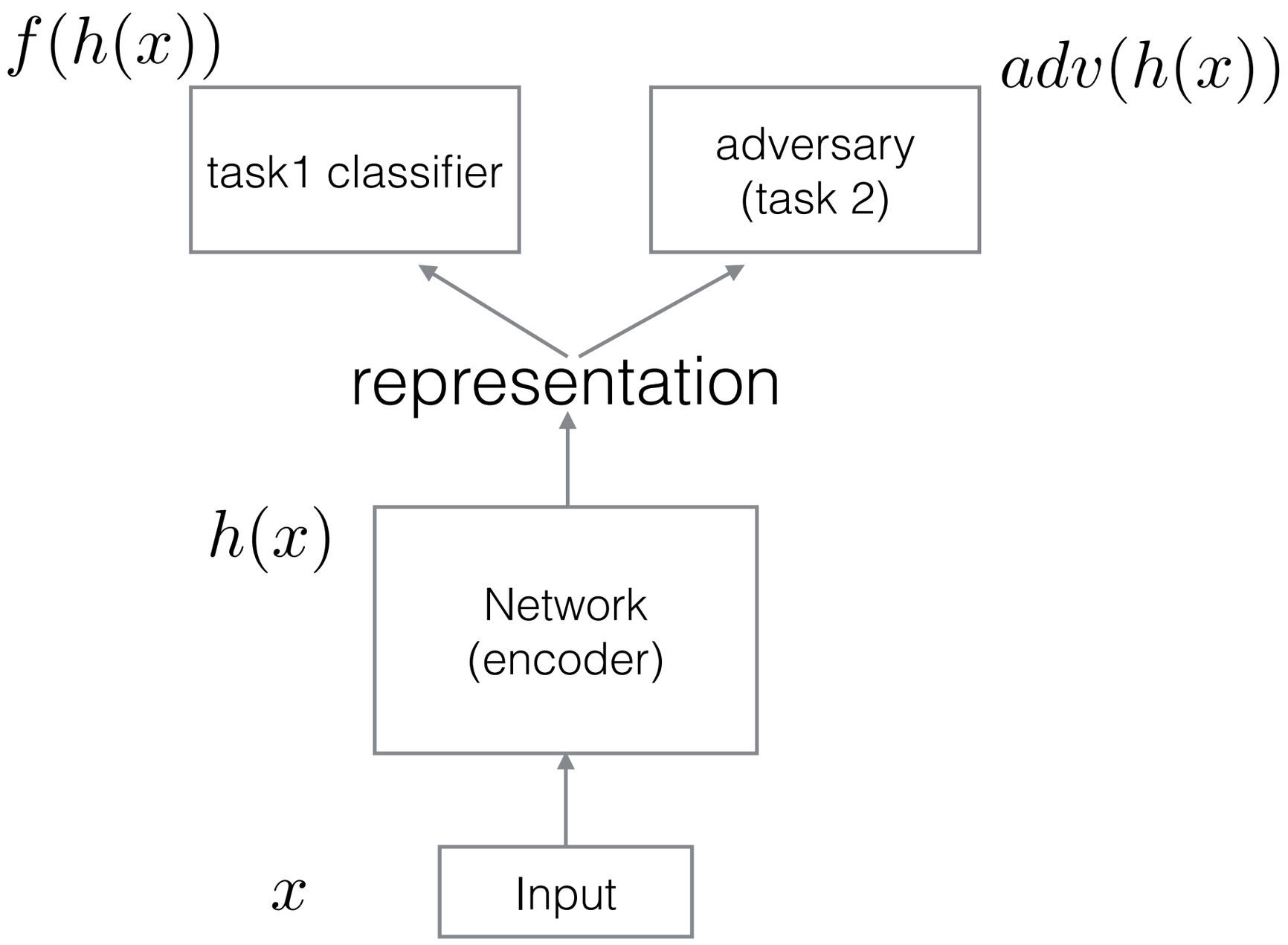


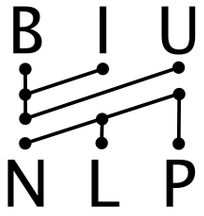
- Adversarial feature learning



- Why adversaries over features?
  - Non-generative tasks
  - Continuous features easier than discrete outputs


$f(h(x))$
task1 classifier
representation
$h(x)$
Network
(encoder)
x
Input


$f(h(x))$
$adv(h(x))$
task1 classifier
adversary
(task 2)
representation
$h(x)$
Network
(encoder)
x
Input



Adversarial Feature Learning

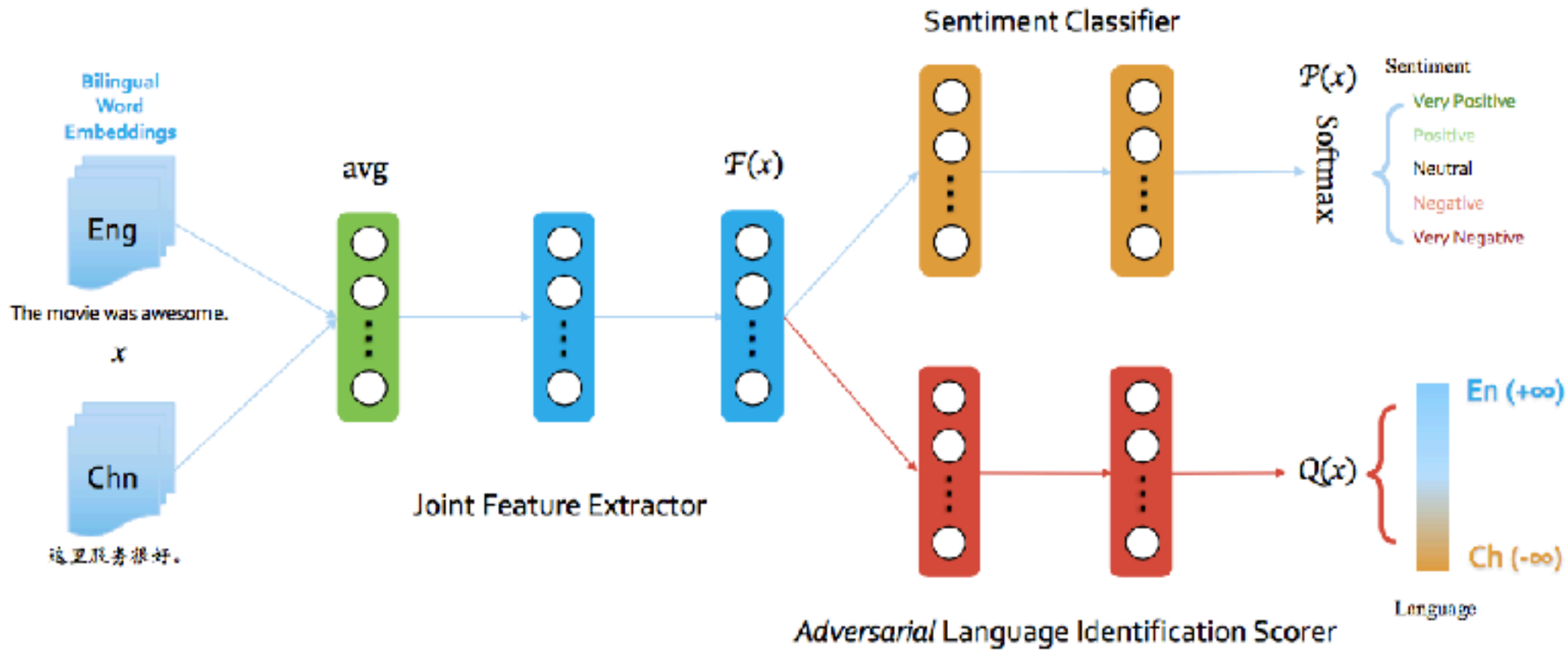
vs

Multi-task Learning

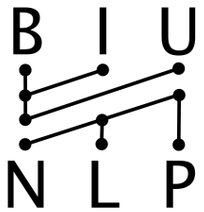
(discuss)

Learning Language-invariant Representations

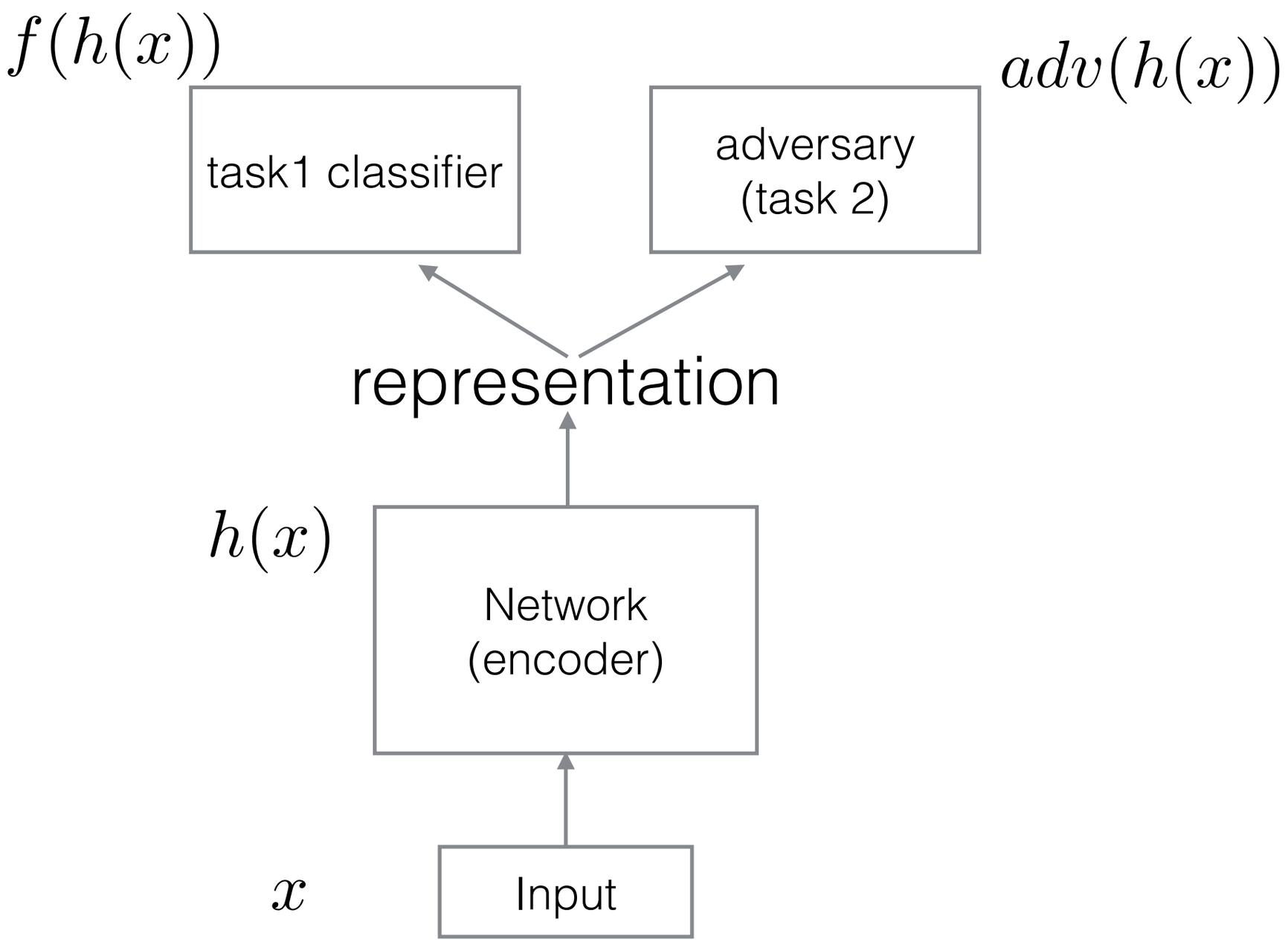
- Chen et al. (2016) learn language-invariant representations for text classification



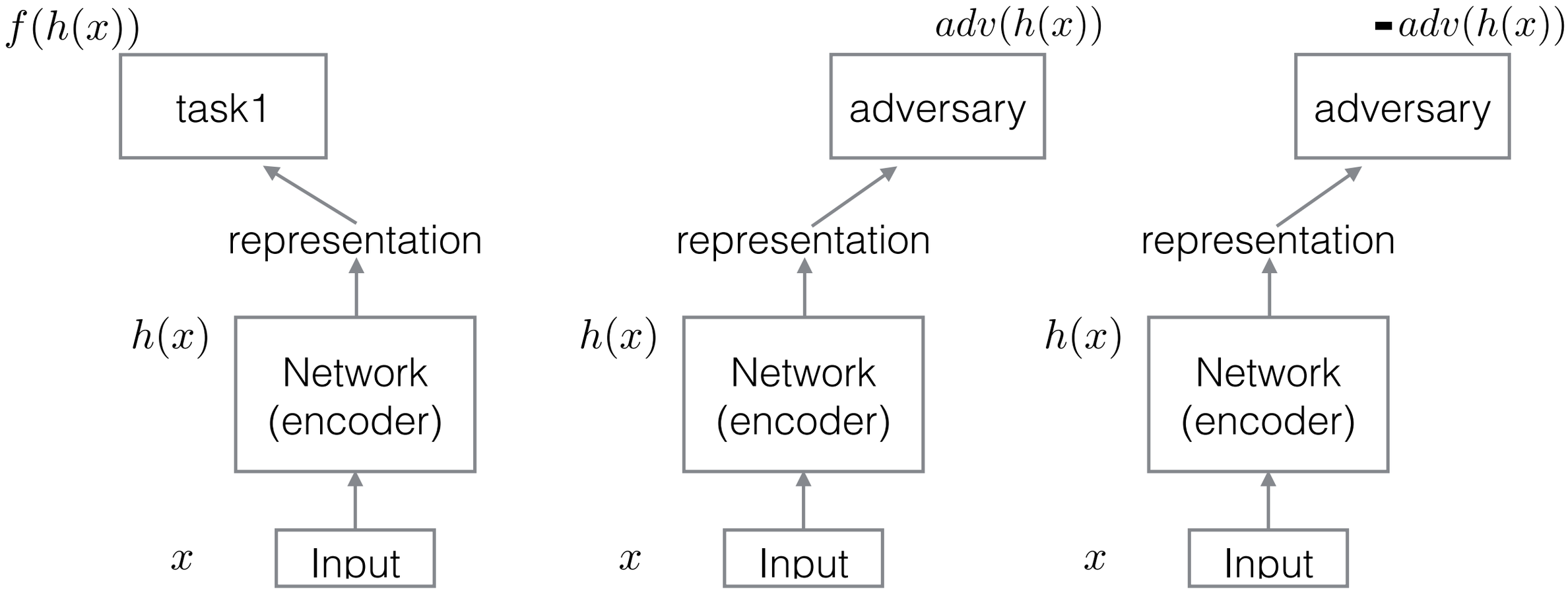
- Also on multi-lingual machine translation (Xie et al. 2017)



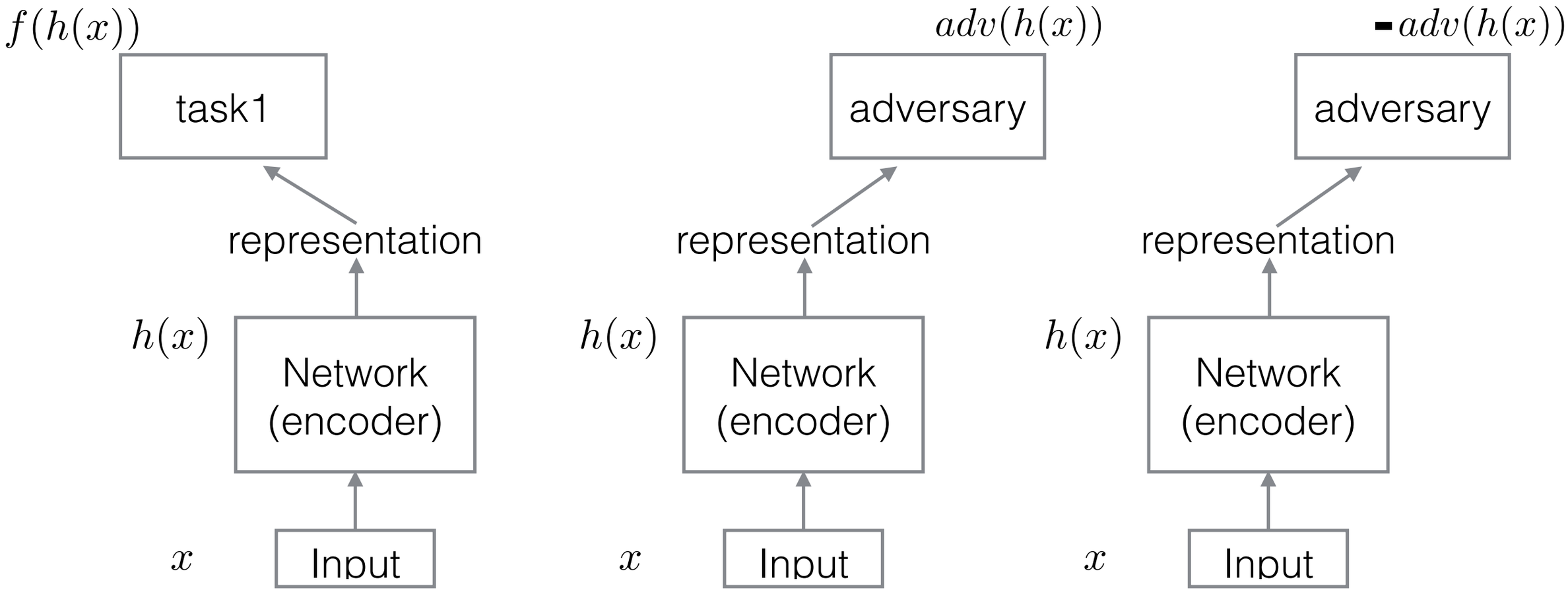
How do we train this in practice?


$f(h(x))$
$adv(h(x))$
task1 classifier
adversary
(task 2)
representation
$h(x)$
Network
(encoder)
x
Input

three different sub-objectives



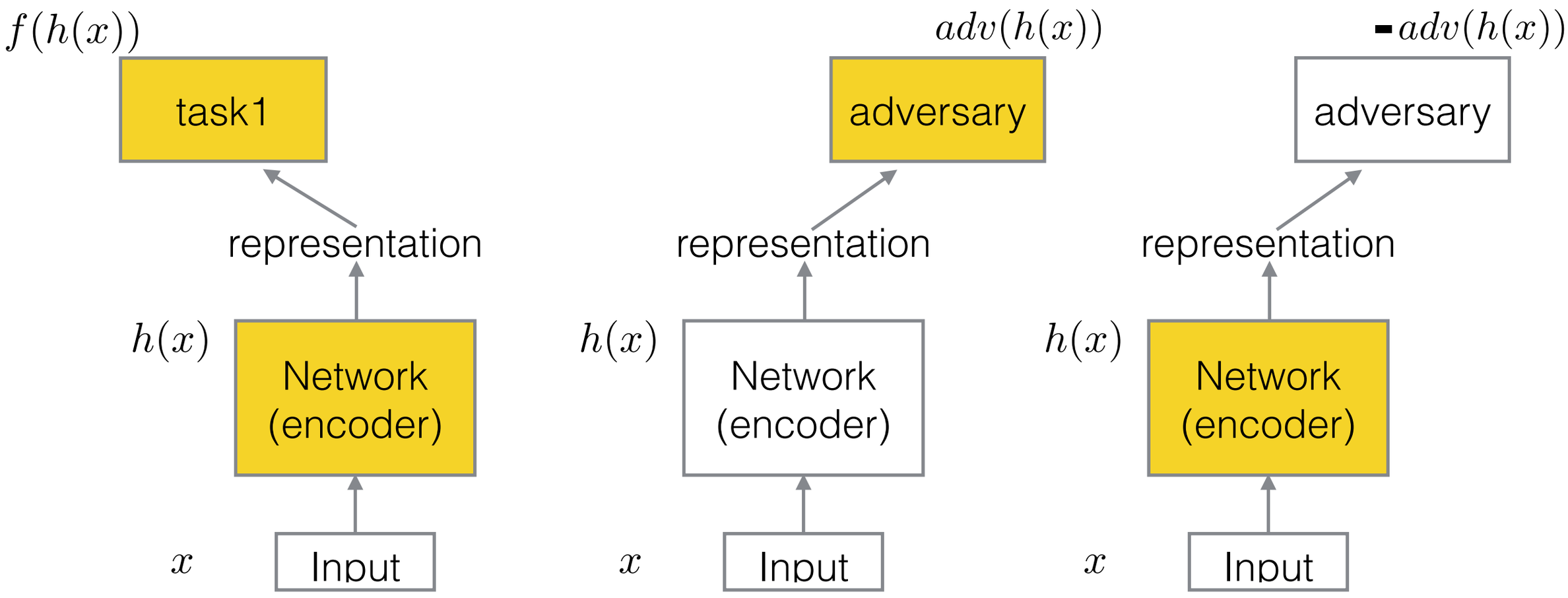
three different sub-objectives



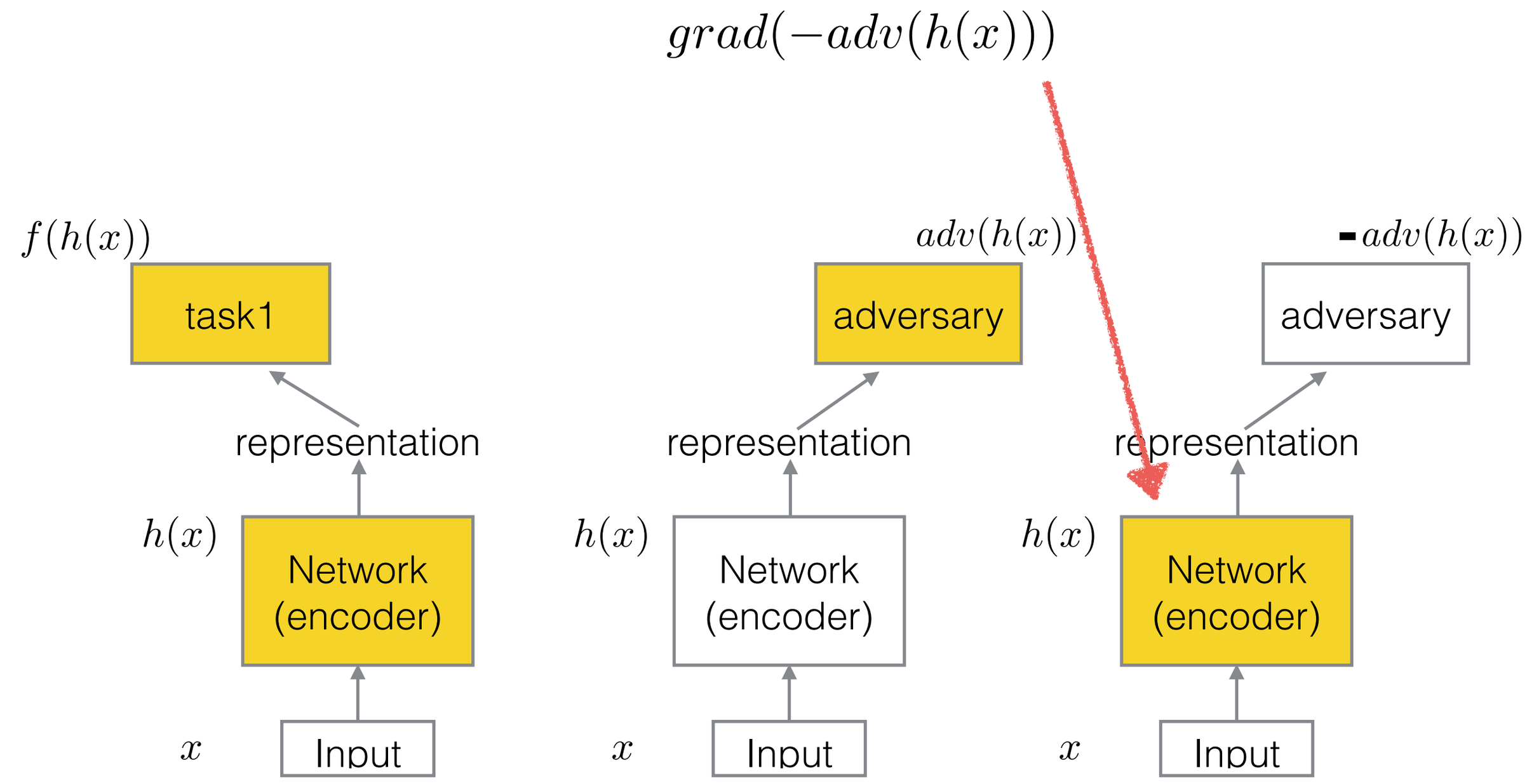
classify well

adversary should succeed

encoder should make adversary fail

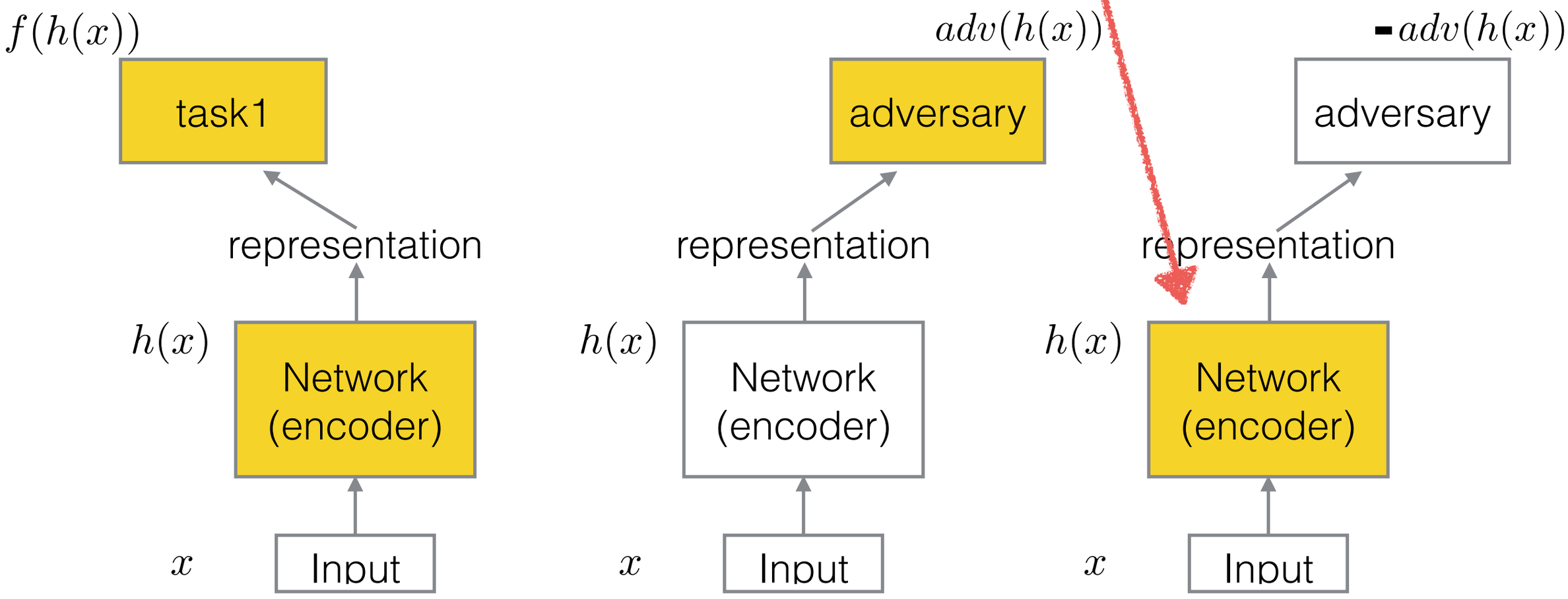


yellow: update parameters
white: don't update

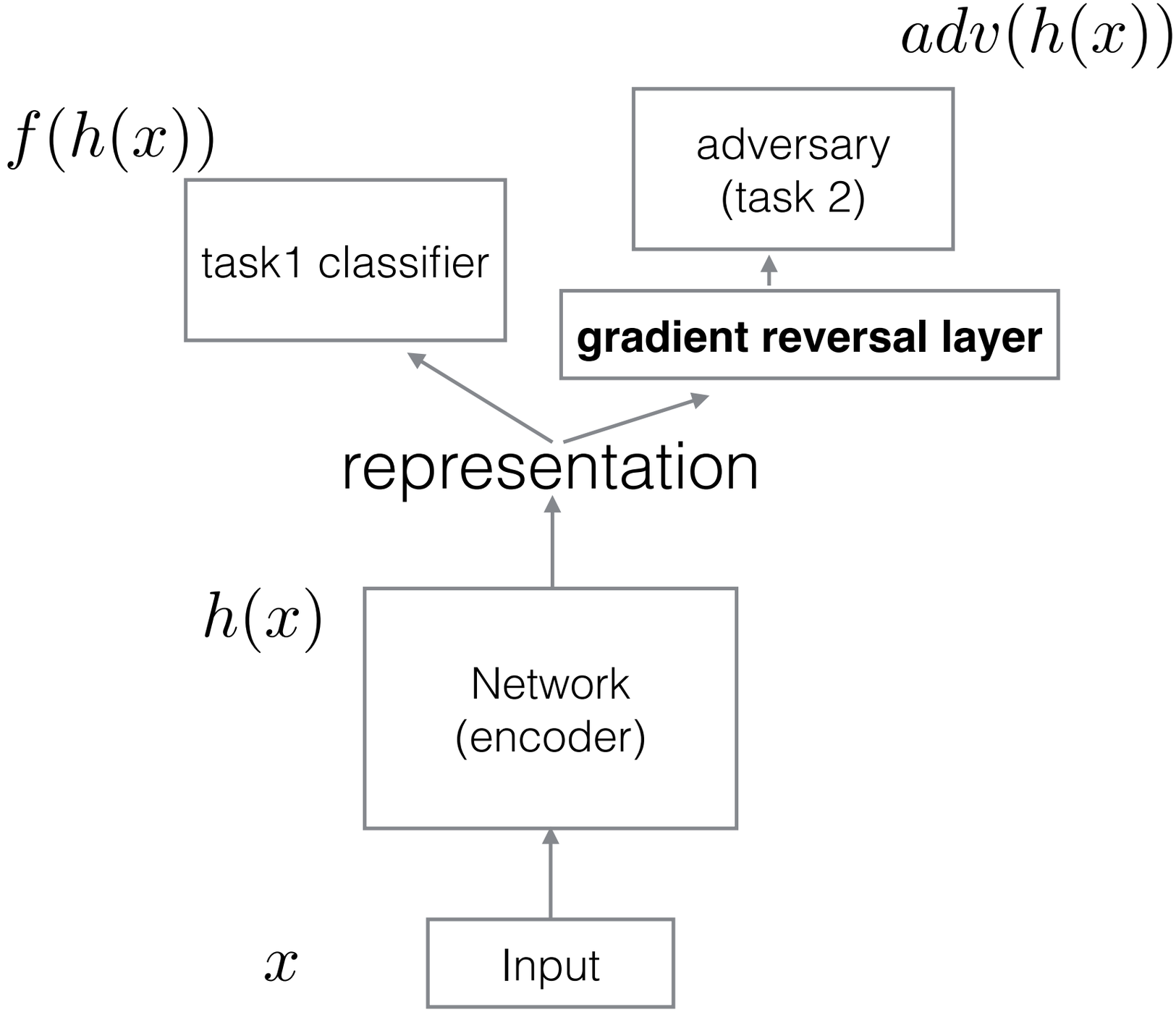


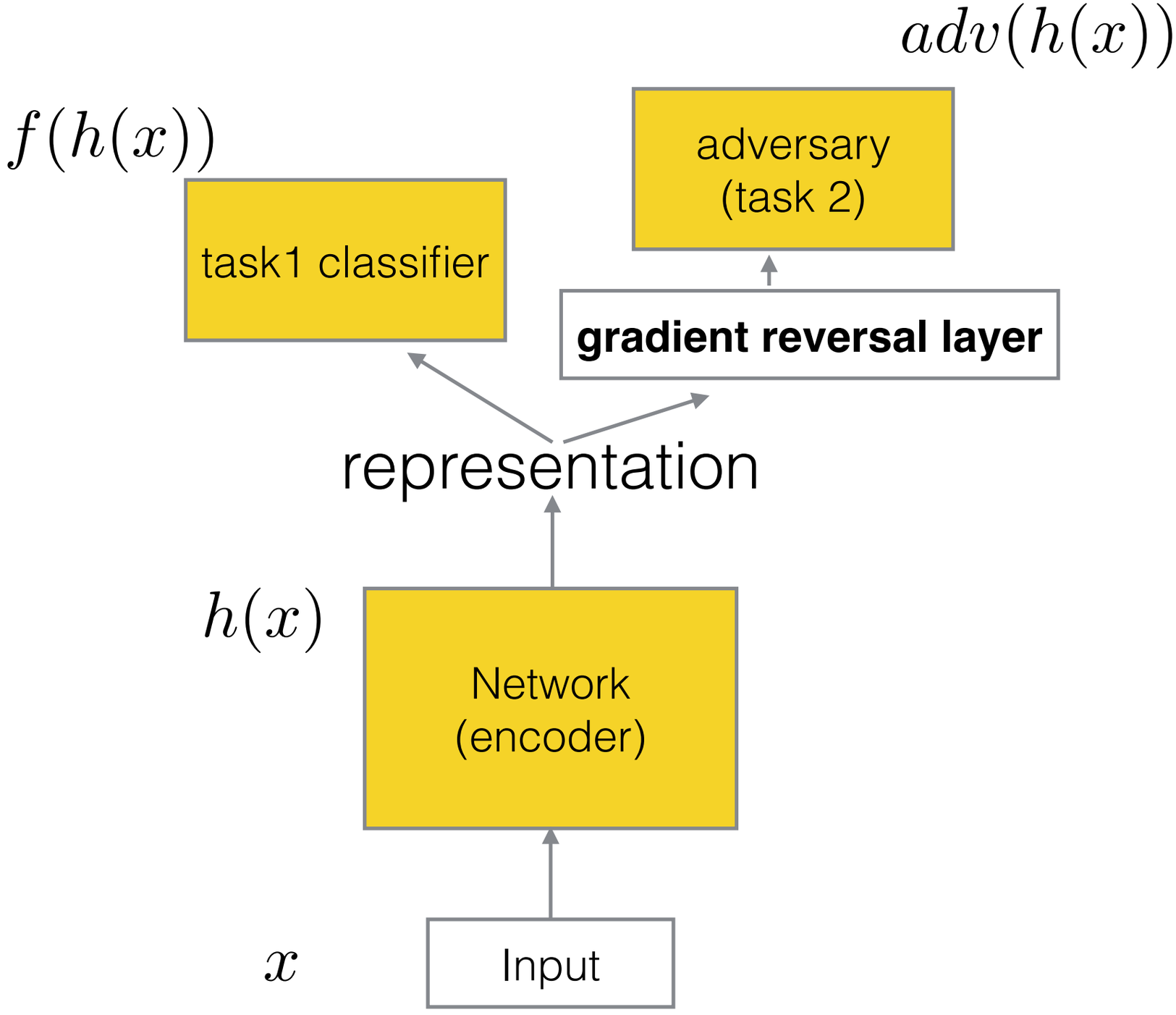
yellow: update parameters
white: don't update

$$grad(-adv(h(x))) = -grad(adv(h(x)))$$



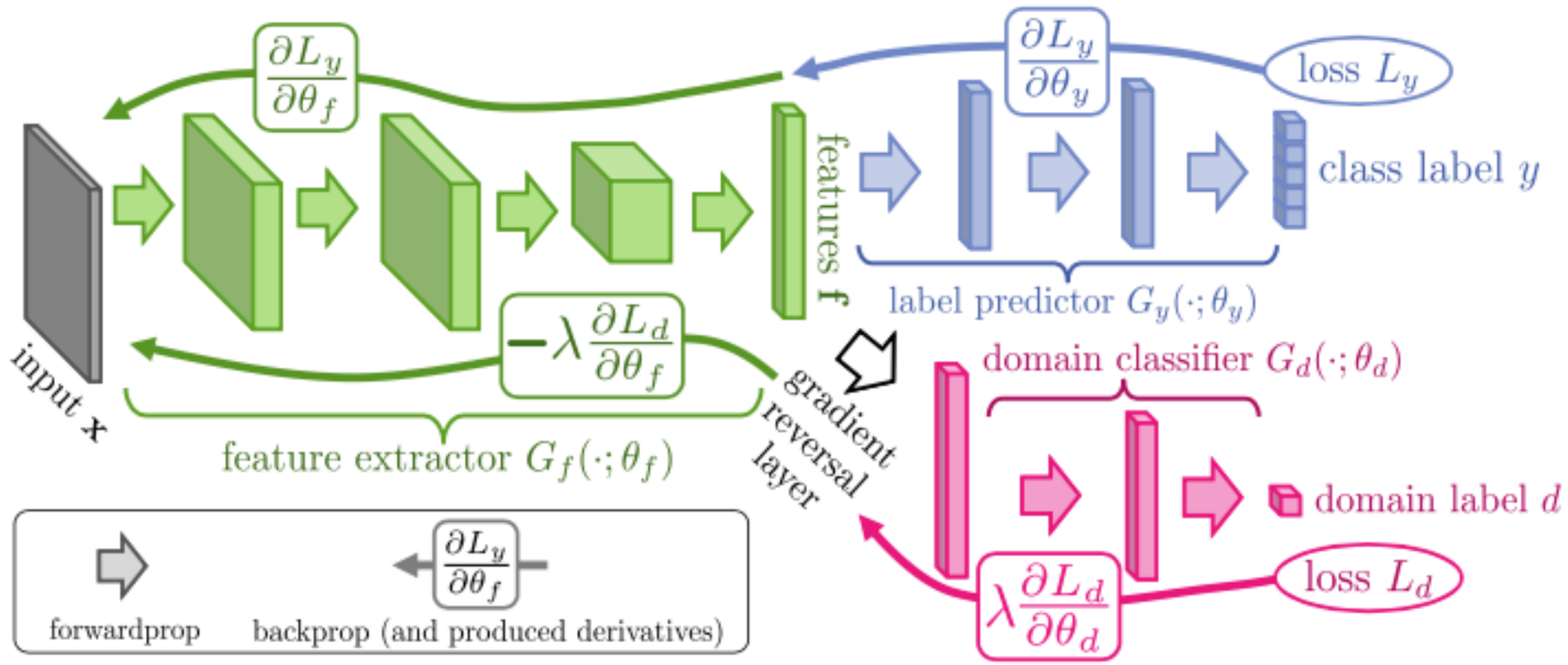
yellow: update parameters
white: don't update


$adv(h(x))$
$f(h(x))$
adversary
(task 2)
task1 classifier
gradient reversal layer
representation
$h(x)$
Network
(encoder)
x
Input


$adv(h(x))$
$f(h(x))$
adversary
(task 2)
task1 classifier
gradient reversal layer
representation
$h(x)$
Network
(encoder)
x
Input

Learning Domain-invariant Representations (Ganin et al. 2016)

- Learn features that cannot be distinguished by domain



- Interesting application to synthetically generated or stale data (Kim et al. 2017)

Another information removal method: INLP

- "Iterative Null-space Projection"
- Goal: a representation h' such that **no linear classifier can predict Z from h'**.

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] **Yanai Elazar**[1,2] **Hila Gonen**[1] **Michael Twiton**[3] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher

Another information removal method: INLP

- "Iterative Null-space Projection"

- Goal: a representation h' such that **no linear classifier can predict Z from h'**.

more stable than gradient reversal

less heuristic

Another information removal method: INLP

- "Iterative Null-space Projection"

- Goal: a representation h' such that **no linear classifier can predict Z from h'**.

- **Main idea:**
  train classifier c1.
  find its null-space.
  project rep representation to null-space.
  train classifier c2 on result.
  find its null-space...

Another information removal method: INLP

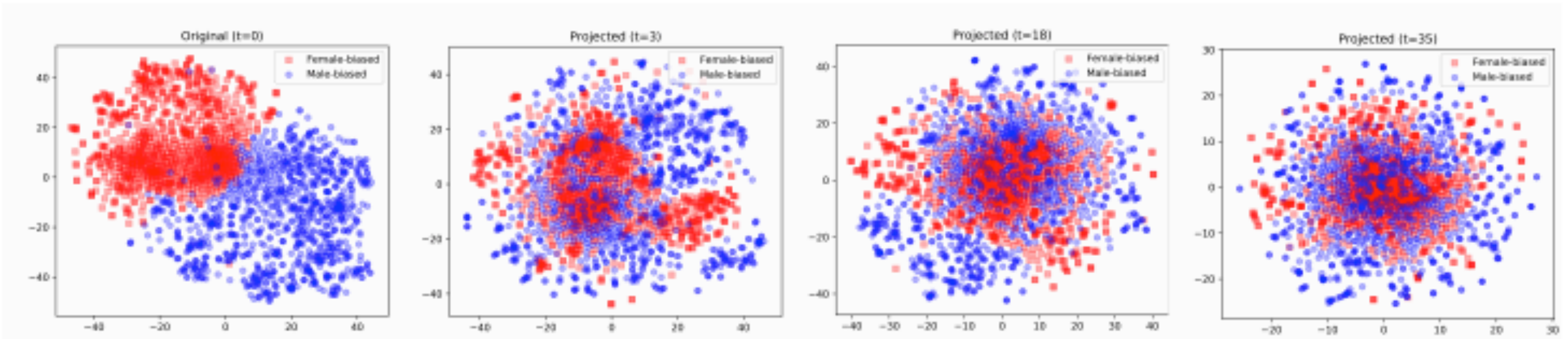


Figure 1: t-SNE projection of GloVe vectors of the most gender-biased words after t=0, 3, 18, and 35 iterations of INLP. Words are colored according to being male-biased or female-biased.

Ethics

"think about what you create and how."

B I U
N L P

should we built this?

should we built this?

"online gaming"?

user tracking for oppressive governments?

user tracking for ads?

user tracking for IDF?

should we built this?

"online gaming"?

user tracking for oppressive governments?

user tracking for ads?

user tracking for IDF?

use your own judgement

B I U
N L P

FACEPTION
Facial Personality Analytics

BIU NLP

FACEPTION

Facial Personality Analytics

OUR CLASSIFIERS

High IQ

Academic Researcher

Professional Poker Player

Terrorist

Utilizing advanced machine learning techniques we developed and continue to evolve an array of classifiers. These classifiers represent a certain persona, with a unique personality type, a collection of personality traits or behaviors. Our algorithms can score an individual according to their fit to these classifiers.

score an individual according to their fit to these classifiers.

Who will gain from this?

Who will suffer from this?

Ethics

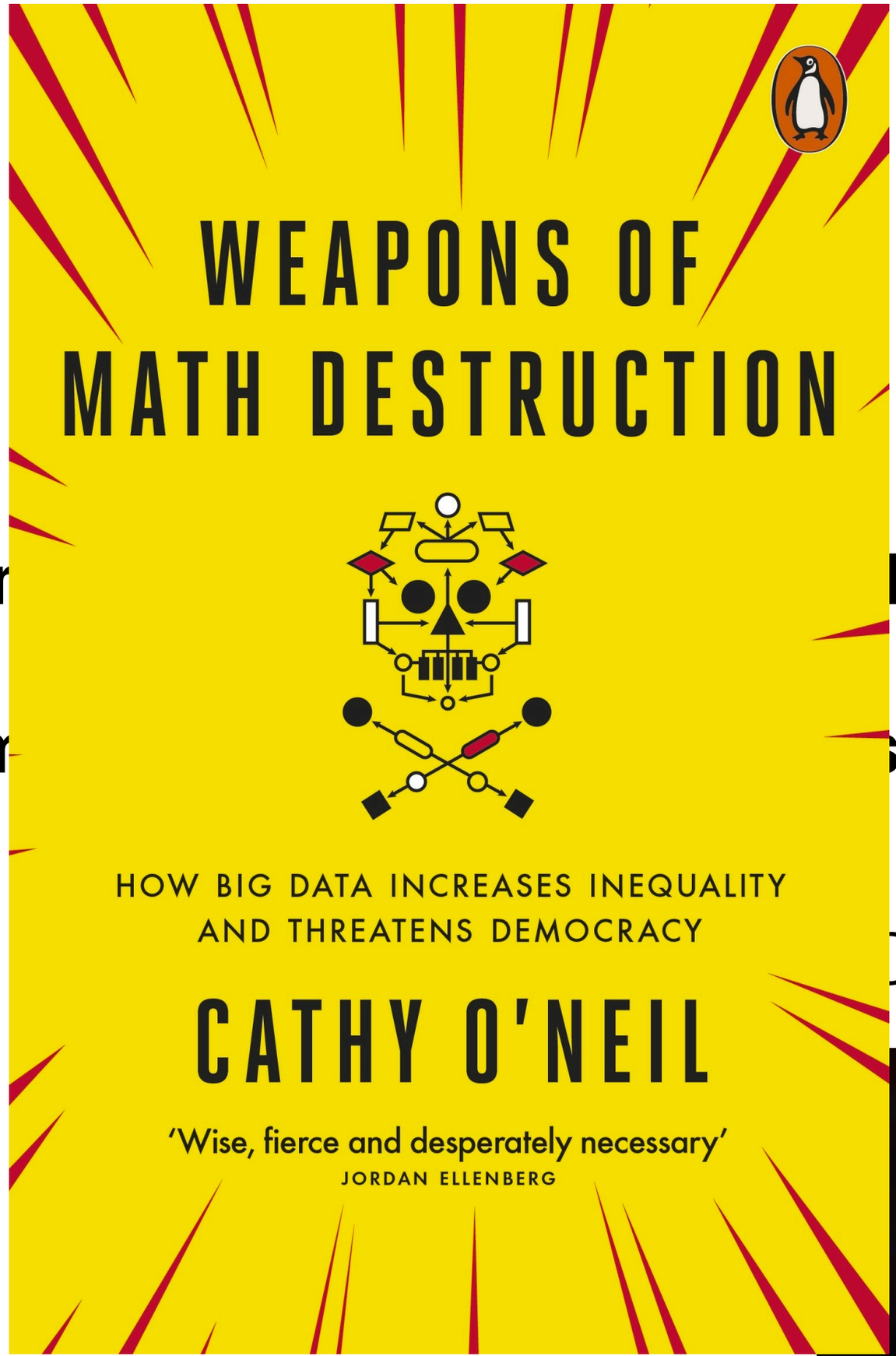
- Your system will operate in the real world.
- It's decisions will affect real people.
- The fact that something is automated/algorithmic/ learned does not mean it is objective/clean/ harmless/safe
- Someone **designed** the algorithm.
  Someone **created** the training data.

Ethics

algorithms are opinions embedded in code

ML models automate the status quo

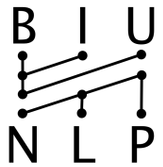
- Cathy O'neil



algorithm … l in code

ML m… s quo

… Cathy O'neil



Ethics

some things are obvious, other are more subtle

Ethics

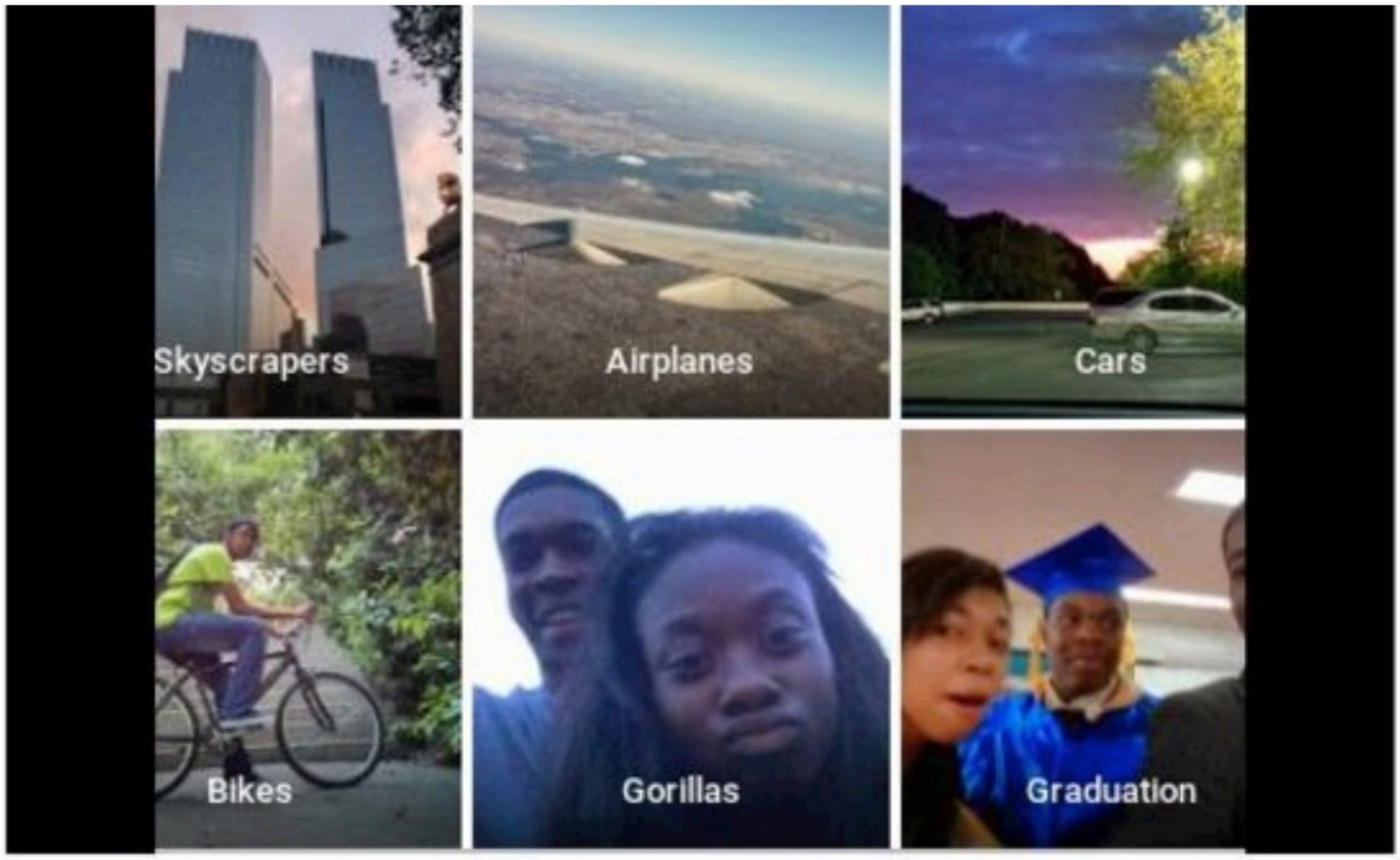
- The "Gorilla" case.

Google apologises for Photos app's racist blunder

http://www.bbc.com/news/technology-33347866

1 July 2015

Share

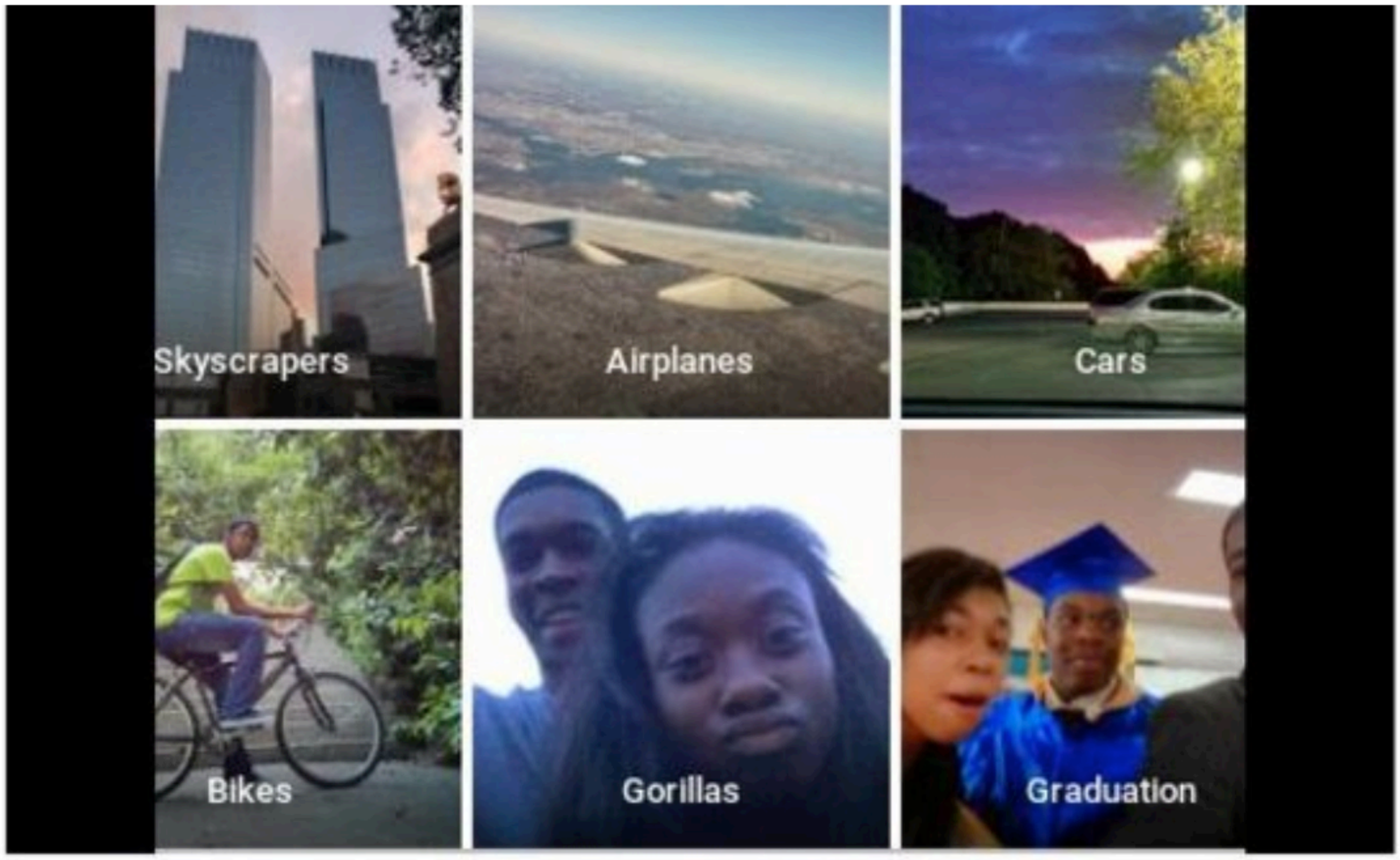


Google apologises for Photos app's racist blunder

http://www.bbc.com/news/technology-33347866

1 July 2015

Share



Technology

Flickr faces complaints over 'offensive' auto-tagging for photos

Auto-tagging system slaps 'animal' and 'ape' labels on images of black people, and tags concentration camps with 'jungle gym' and 'sport'

Alex Hern

@alexhern

Wed 20 May 2015
09.49 BST

This article is 2 years old

612 93

Flickr is facing a user revolt after a new auto-tagging system labelled images of

Why did it happen?
How can we treat this?

- [discuss]

solution?

don't say anything is a gorilla

TOM SIMONITE BUSINESS 01.11.18 07:00 AM

WHEN IT COMES TO GORILLAS, GOOGLE PHOTOS REMAINS BLIND

https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

ML impacts millions of people everyday

And it also propagates biases in unknown ways

MPW • AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

Real-world interactions

Real-world interactions

This CIA-funded tool predicts crime before it happens

by BRYAN CLARK — Aug 2, 2017 in ARTIFICIAL INTELLIGENCE

Mos

how will such a system interact with the real world?
feedback loops...

Real-world interactions

- **"Predictive policing"**
- Use ML to decide where to send police patrols
    - ...based on previous crime statistics
        - ...reported to the police, and handled by police
- What's the result of this?

Real-world interactions

- **"Predictive policing"**
- Use ML to decide where to send police patrols
  - ...based on previous crime statistics
    - ...reported to the police, and handled by police
- What's the result of this?

feedback loops!
causality!

Real-world interactions

- **"Predictive policing"**
- Use ML to decide where to send police patrols
  - ...based on previous crime statistics
    - ...reported to the police, and handled by police
- What's the result of this?

arrests data != crime data

feedback loops!
causality!

Real-world interactions

- "**Crowd-based Video recommendation**"
    - YouTube suggests what to watch next based on past user behavior.
- What are the implications?

Real-world interactions

- "**Crowd-based Video recommendation**"
    - YouTube suggests what to watch next based on past user behavior.
    - What are the implications?
        - People watch more videos, metrics go up
        - but which videos are we watching?
        - are we happy with this?

Real-world interactions

- **"Teacher Assessment"**
- Teacher is good if many students pass exam.
- What can happen?

Real-world interactions

- **"Teacher Assessment"**
- Teacher is good if many students pass exam.
    - What can happen?
- Teacher is good if many students pass standardized exams.
    - What can happen?

Real-world interactions

- **"Teacher Assessment"**
- Teacher is good if many students pass exam.
    - What can happen?
- Teacher is good if many students pass <u>standardized</u> exams.
    - What can happen?

        - factors beyond teacher affect scores
        - influence on teacher's career
        - influence on teacher's choices and the world

Meaning of a mistake

- What can happen if your system is 100% accurate?
- What can happen if your system is 90% accurate?
    - Are the mistakes evenly distributed?

Meaning of a mistake

- A deep learning system that predicts IQ scores based on face images.
- What if it is 100% accurate?
- What if it is 90% accurate on females?
    - ... and 60% accurate on blonde females aged < 25?

ML encodes bias

- What biases do you encode into your system?
  - Intentionally
  - Unintentionally
- Who will be affected?

Ethics

- **Some possible sources of problems:**
  - Bias in training data
    - In p(y)
    - In p(x)
    - In p(y|x)
    - In p(x|y)
  - Architecture, Features, Loss function
  - Bias in outputs
  - Feedback loops

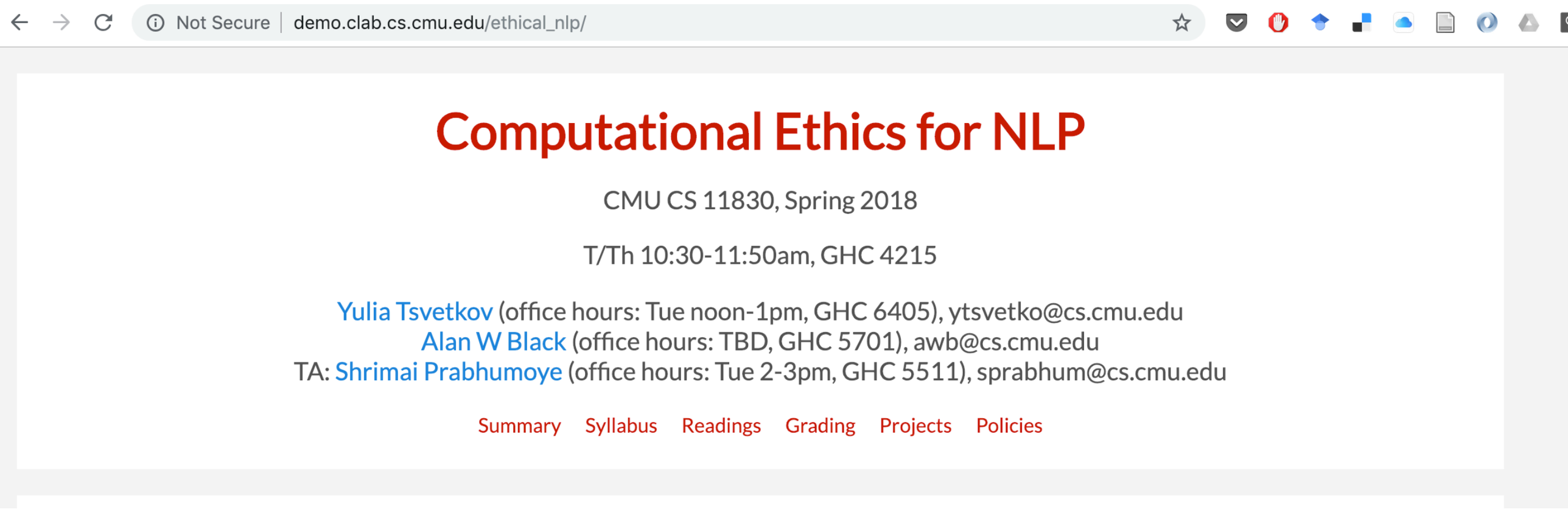
bad data -> bad models

data is based on people...

Ethics in ML: Some resources

http://www.fatml.org/

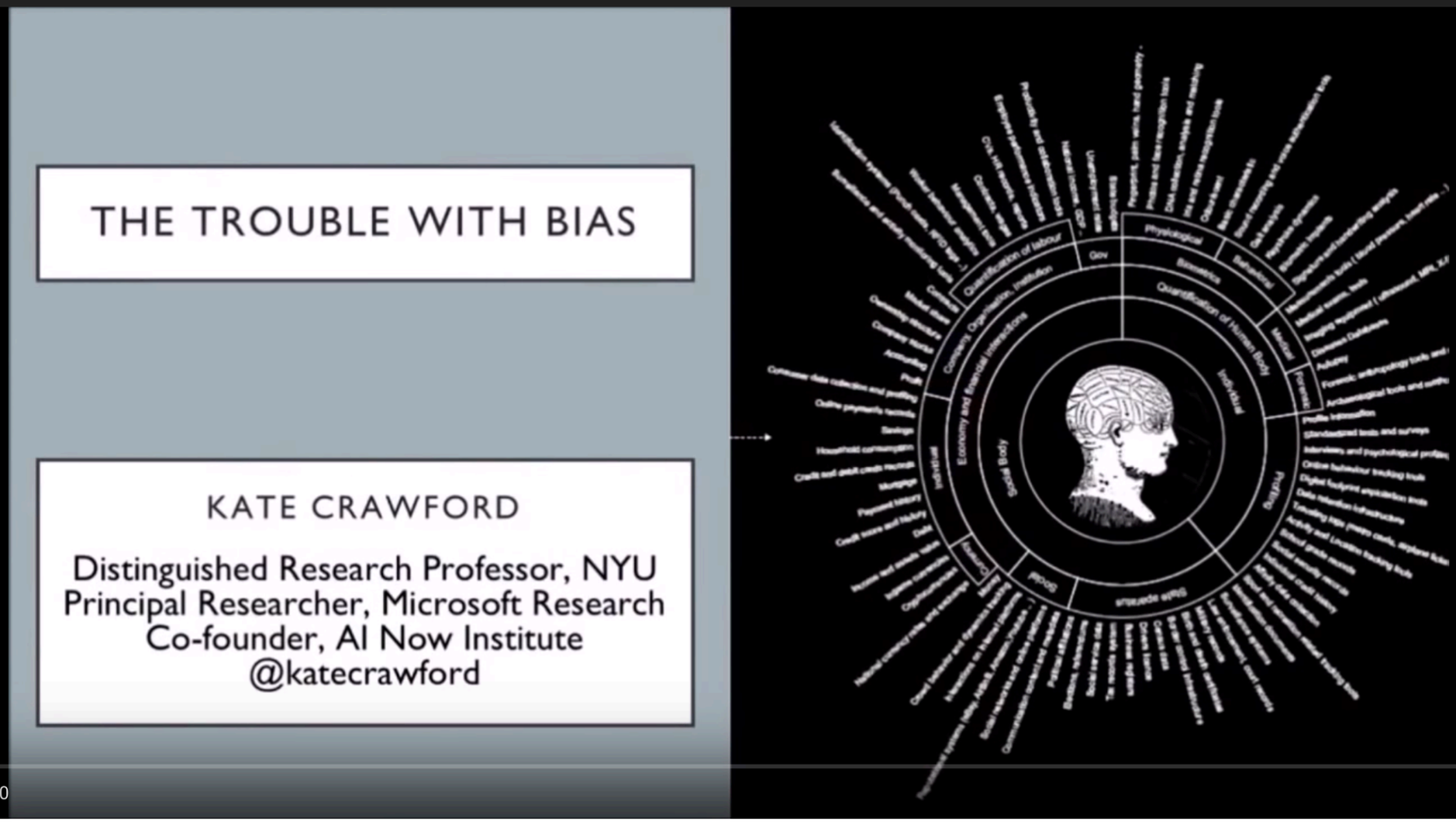
Ethics in ML: Some resources



http://demo.clab.cs.cmu.edu/ethical_nlp/

https://www.youtube.com/watch?v=fMym_BKWQzk

YouTube Search



0:01 / 49:30

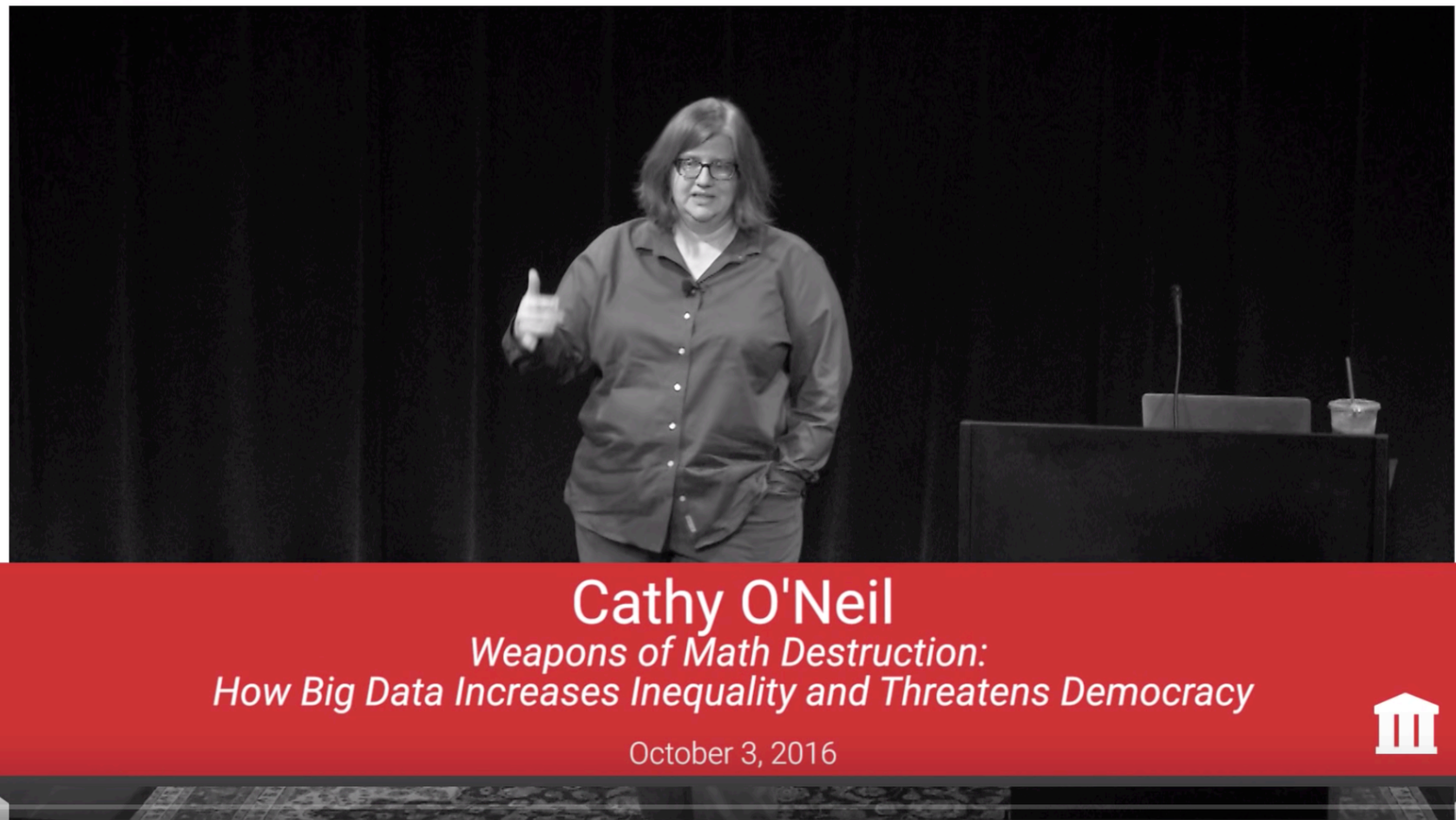
The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017

26,340 views

373 54 SHARE SAVE

Up next AUTOPLAY

You and AI – Just An Engineer:



Cathy O'Neil: "Weapons of Math Destruction" | Talks at Google

26,849 views

Ethics

- Machine learning and Data Science gives great power.
- With great power comes great responsibility.
- Think about what you build and why.

- I hope you enjoyed the course.