Challenges for Socially-Beneficial AI

Daniel S. Weld
University of Washington

Outline

- Distractions *vs.*
- Important Concerns
  - Sorcerer's Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment

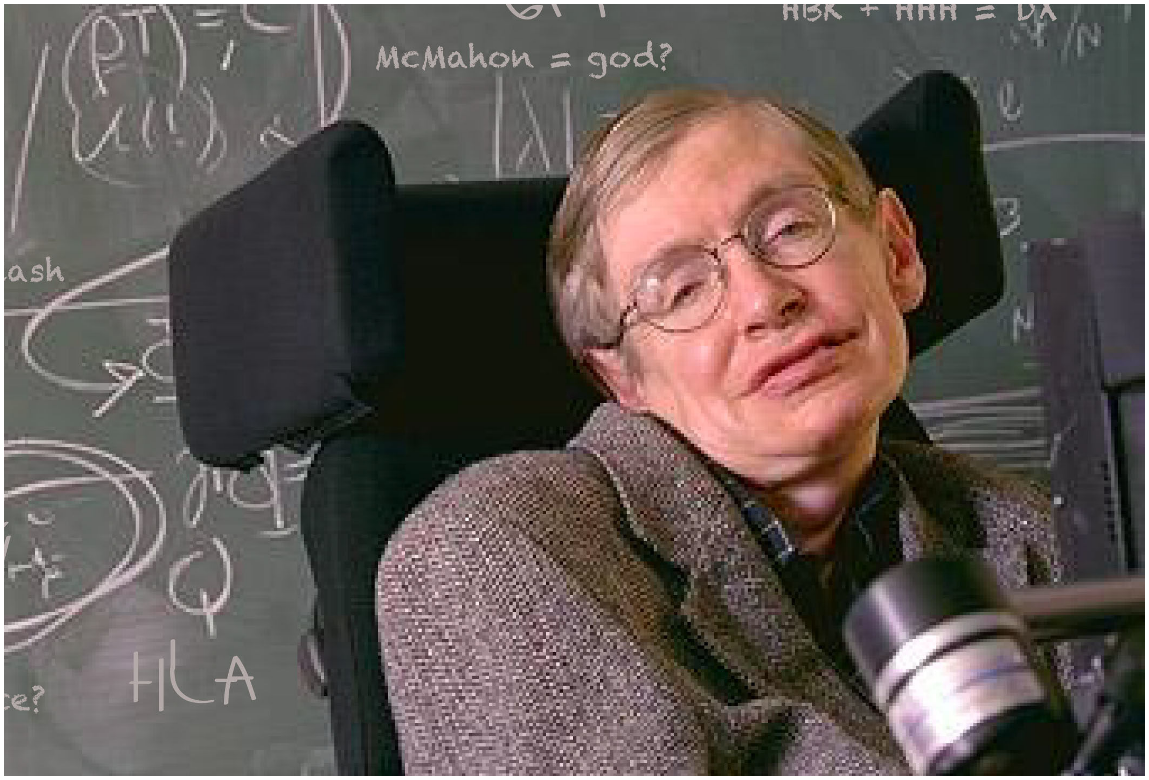
Potential Benefits of AI

- Transportation
  - 1.3 M people die in road crashes / year
  - An additional 20-50 million are injured or disabled.
  - Average US commute 50 min / day
- Medicine
  - 250k US deaths / year due to medical error
- Education
  - Intelligent tutoring systems, computer-aided teaching

- asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics
- https://www.washingtonpost.com/news/to-your-health/wp/2016/05/03/researchers-medical-errors-now-third-leading-cause-of-death-in-united-states/?utm_term=.49f29cb6dae9

Will AI Destroy the World?

“Success in creating AI would be the biggest event in human history… Unfortunately, it might also be the last” … “[AI] could spell the end of the human race.”– Stephen Hawking



How Does this Story End?

"With artificial intelligence we are summoning the demon." – Bill Gates

An Intelligence Explosion?

"Before the prospect of an ***intelligence explosion***, we humans are like small children playing with a bomb" – Nick Bostom

"Once machines reach a certain level of intelligence, they'll be able to work on AI just like we do and improve their own capabilities—redesign their own hardware and so on—and their intelligence will zoom off the charts."

– Stuart Russell

Superhuman AI & Intelligence Explosions

- When will computers have superhuman capabilities?

- Now.
  - Multiplication
  - Spell checking
  - Chess, Go

  - Many more abilities to come

AI Systems are ***Idiot Savants***

- Super-human here   &   super-stupid there
- Just because AI gains one superhuman skill… Doesn’t mean it is suddenly good at ***everything***

  And certainly not unless we give it **experience** *at everything*

- AI systems will be spotty for a very long time

Example: SQuAD

Paragraph

Martin Luther (10 November 1483 – 18 February 1546) was a German professor of theology, composer, priest, former monk and a seminal figure in the Protestant Reformation. Luther came to reject several teachings and practices of the Late Medieval Catholic Church. He strongly disputed the claim that freedom from God's punishment for sin could be purchased with money. He proposed an academic discussion of the power and usefulness of indulgences in his Ninety-Five Theses of 1517. His refusal to retract all of his writings at the demand of Pope Leo X in 1520 and the Holy Roman Emperor Charles V at the Diet of Worms in 1521 resulted in his excommunication by the Pope and condemnation as an outlaw by the Emperor.

Question

Who asked Luther to disavow his writings?

Human F1 86.8%

Rajpurkat et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text," https://arxiv.org/pdf/1606.05250.pdf

Impressive Results

http://35.165.153.16:1995/

Paragraph

Martin Luther (10 November 1483 – 18 February 1546) was a German professor of theology, composer, priest, former monk and a seminal figure in the Protestant Reformation. Luther came to reject several teachings and practices of the Late Medieval Catholic Church. He strongly disputed the claim that freedom from God's punishment for sin could be purchased with money. He proposed an academic discussion of the power and usefulness of indulgences in his Ninety-Five Theses of 1517. His refusal to retract all of his writings at the demand of Pope Leo X in 1520 and the Holy Roman Emperor Charles V at the Diet of Worms in 1521 resulted in his excommunication by the Pope and condemnation as an outlaw by the Emperor.

Question

Who asked Luther to disavow his writings?

Answer

Pope Leo X

Human	F1	86.8%
Seo et al.	F1	81.1%

Seo et al. "Bidirectional Attention Flow for Machine Comprehension" arXiv:1611.01603v5

It’s a Long Way to General Intelligence

Paragraph

Alice and Dave went to school. Only one liked science. Alice liked chemistry. Dave only liked music.

Question

who didn't like science?

Answer

Alice

Impressive Results

I think it's a brown horse grazing in front of a house.

Microsoft CaptionBot

It’s a Long Way to General Intelligence

I am not really confident, but I think it's a woman standing talking on a cell phone and she seems 😐.

Microsoft CaptionBot

AI Systems are ***Idiot Savants***

- Super-human here & super-stupid there

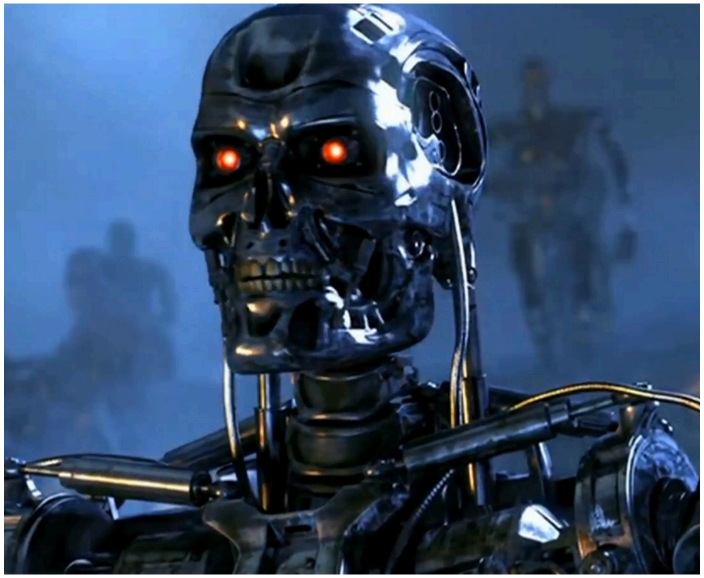
- No common sense
- No long term autonomy
  - Slower and more degraded as learning increases
- No goals besides those we give them

"No machines with self-sustaining long-term goals and intent have been developed, nor are they likely to be developed in the near future." *

* P. Stone et al. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. **http://ai100.stanford.edu/2016-report**.

Terminator / Skynet

"Could you prove that your systems can't ever, no matter how smart they are, overwrite their original goals as set by the humans?"
– Stuart Russell



It's the Wrong Question

- Very unlikely that an AI will wake up and decide to kill us

But…

- Quite likely that an AI will do something unintended

Outline

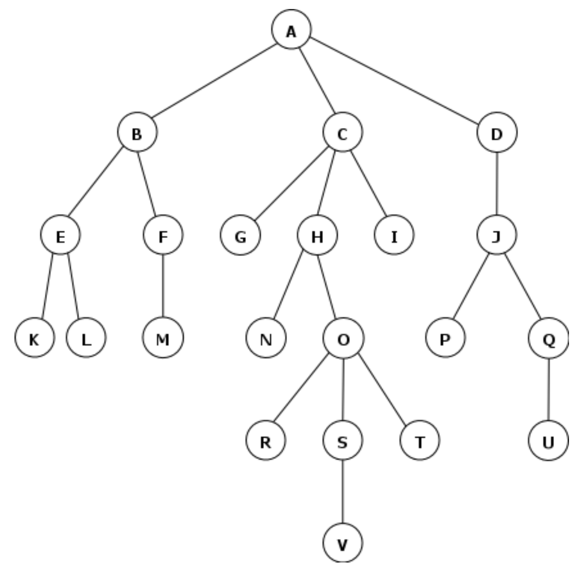
- Distractions *vs.*
- Important Concerns
  - Sorcerer's Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment

Sorcerer's Apprentice

Tired of fetching water by pail, the apprentice enchants a broom to do the work for him – using magic in which he is not yet fully trained. The floor is soon awash with water, and the apprentice realizes that he cannot stop the broom because he does not know how.

AI assistants may hurt us ***accidentally***, while (literally) obeying our orders.

Script *vs.* Search-Based Agents



Now

Soon

Unpredictability

Ok Google, how much of my Drive storage is used for my photo collection?

None, Dave!
I just executed rm *

(It was easier than counting file sizes)

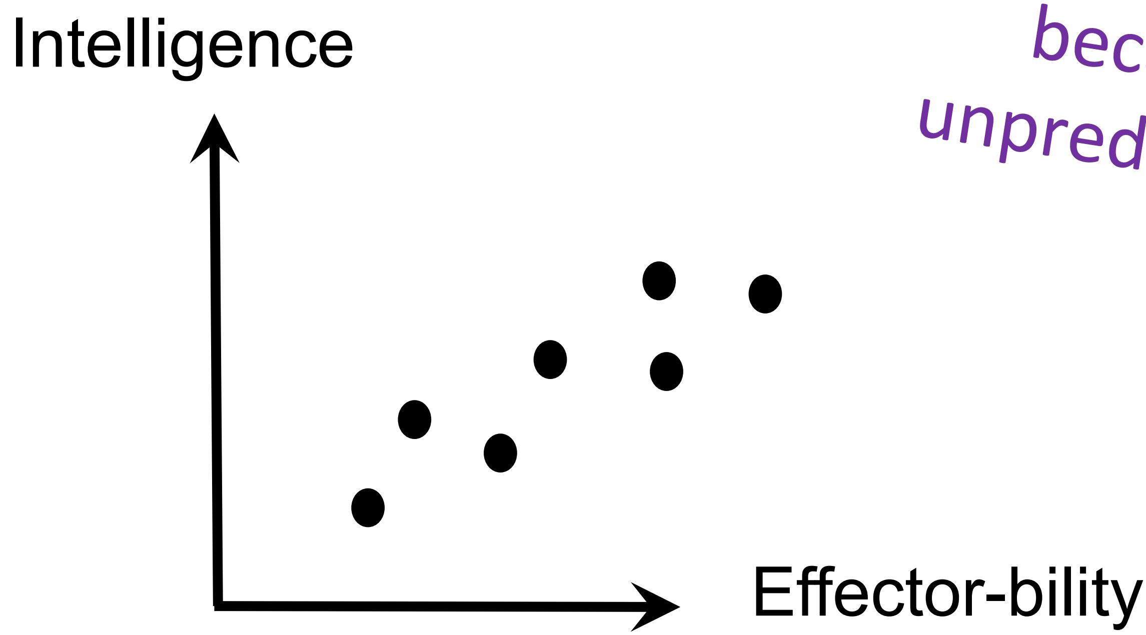
Brains Don't Kill

It's an agent's ***effectors*** that cause harm

- 2003, an error in General Electric's power monitoring software led to a massive blackout, depriving 50 million people of power.
- 2012, Knight Capital lost $440 million when a new automated trading system executed 4 million trades on 154 stocks in just forty-five minutes.

Correlation Confuses the Two

With increasing intelligence, comes our desire to adorn an agent with strong effectors



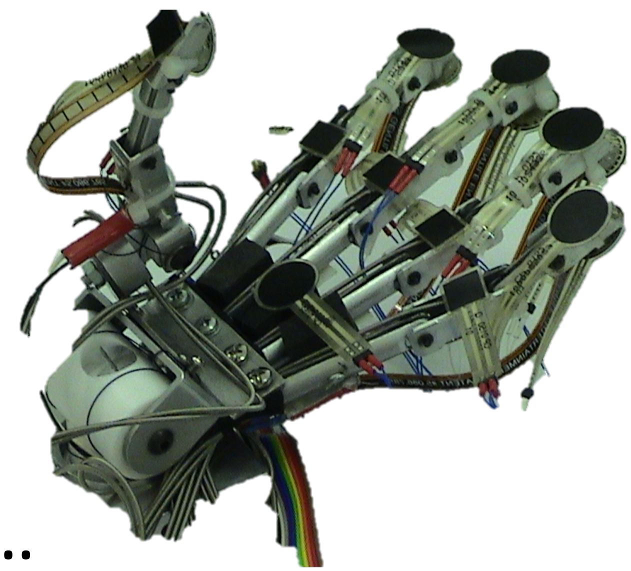
Intelligence is problematic because it can lead to unpredictable, emergent behavior

Physically-Complete Effectors

- Roomba effectors close to harmless
- Bulldozer blade $\vee$ missile launcher … dangerous

- Some effectors are ***physically-complete***
  - They can be used to create other more powerful effectors

  - E.g. the human hand
    created tools….
    that were used to create more tools…
    that could be used to create nuclear weapons



Universal Subgoals

-Stuart Russell

For any primary goal, …

These subgoals increase likelihood of success:

- Stay alive

  (It’s hard to fetch the coffee if you’re dead)

- Get more resources

Specifying Utility Functions

Clean up as much dirt as possible!

An optimizing agent will start making messes, just so it can clean them up.

Specifying Utility Functions

Clean up as many messes as possible, but don't make any yourself.

An optimizing agent can achieve more reward by turning off the lights and placing obstacles on the floor… hoping that a human will make another mess.

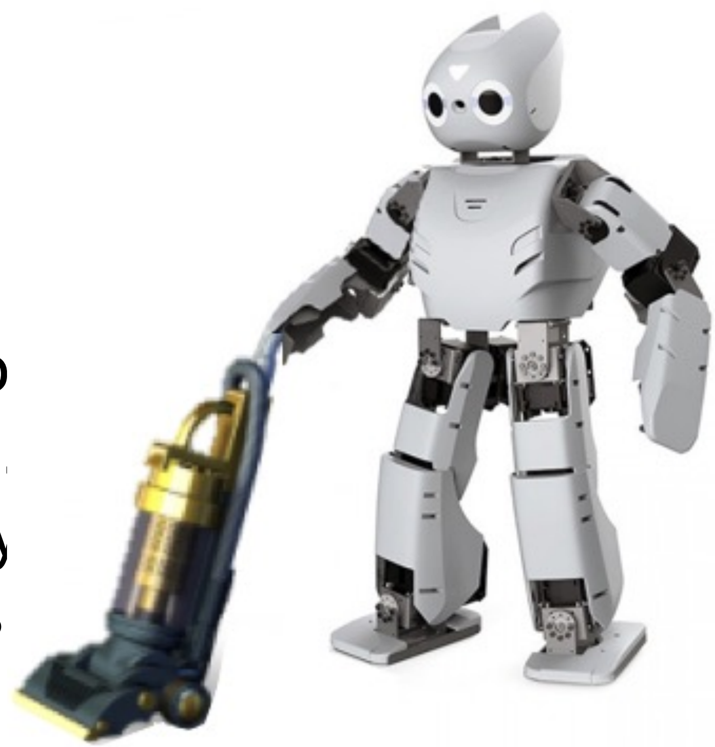
Specifying Utility Functions

Keep the room as clean as possible!

An optimizing agent might kill the (dirty) pet cat. Or at least lock it out of the house.
In fact, best would be to lock humans out too!

Specifying Utility Functions

Clean up any messes made by others as quickly as possible.



There's no incentive for the 'bot to help master avoid making a mess. In fact, it might increase reward by causing a human to make a mess if it is nearby, since this would reduce average cleaning time.

Specifying Utility Functions

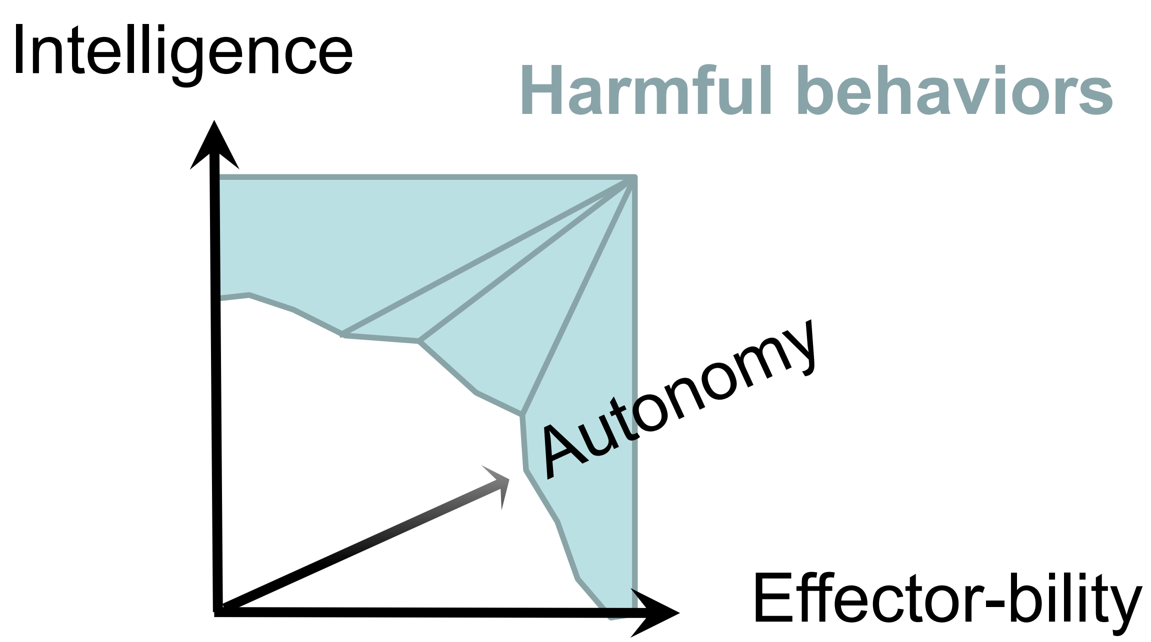
Keep the room as clean as possible, but never commit harm.

Asimov's Laws 1942

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

A Possible Solution: Constrained Autonomy?

Restrict an agents behavior with background constraints



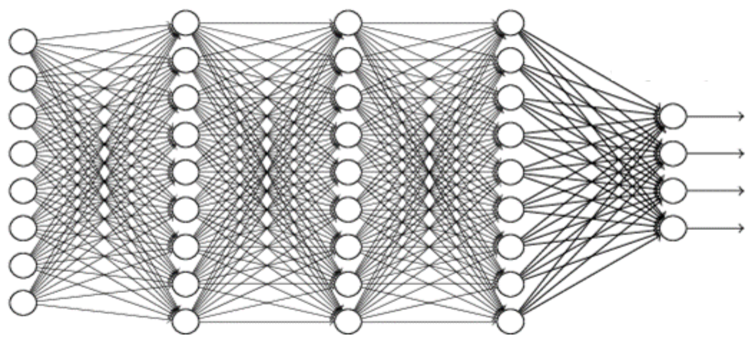
But what *is* Harmful?

1. A robot may not ***injure*** a human being or, through inaction, allow a human being to come to ***harm***.

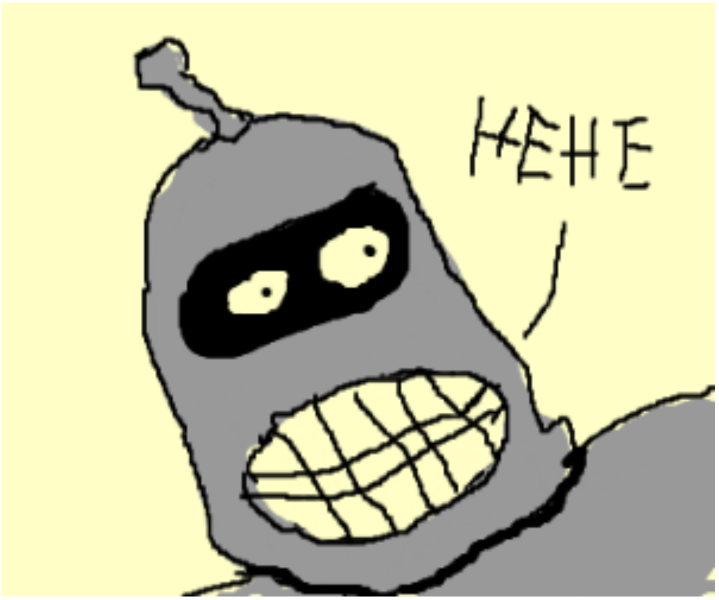
- Harm is hard to define
- It involves complex tradeoffs
- It's different for different people

Trusting AI

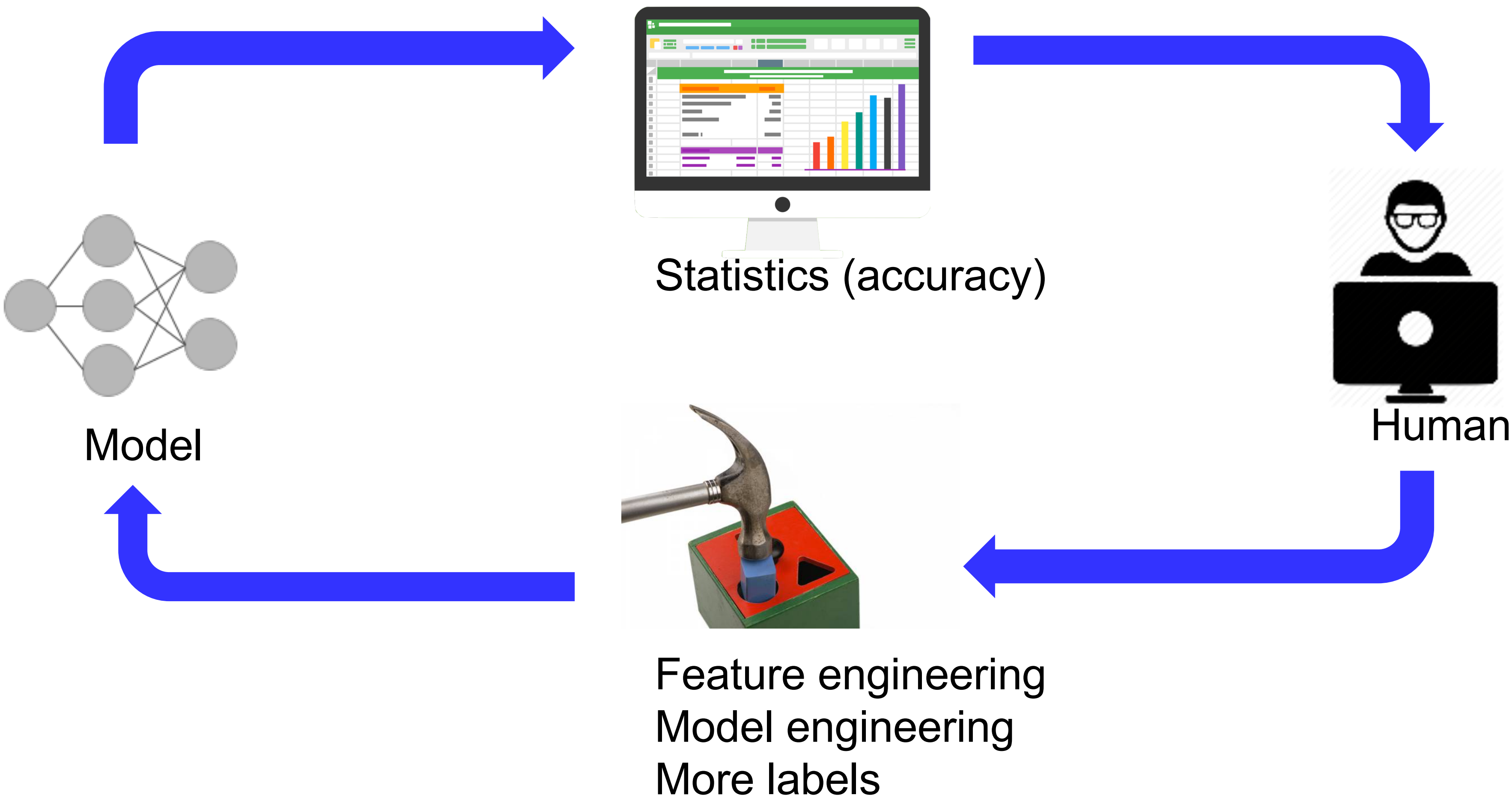
- How can a user teach a machine what's harmful?
- How can they know when it really understands?
  - Especially:



- Explainable Machine Learning



Human – Machine Learning loop today



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

Accuracy problems - example

20 Newsgroups subset – Atheism vs Christianity

94% accuracy!!!

Test on recent dataset, accuracy only 57%

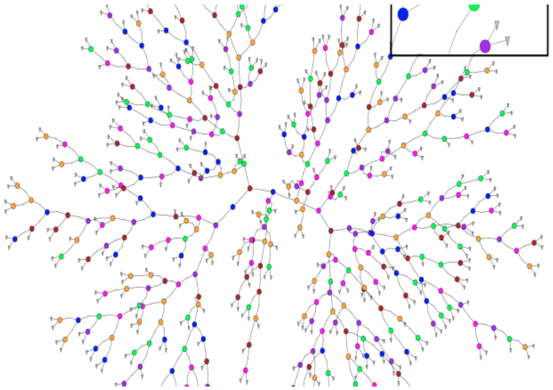
Predictions due to email addresses, names,…

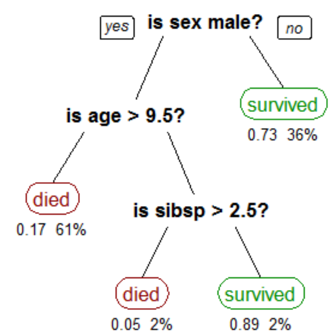
Slide adapted from Marco Ribeiro – see “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

Desiderata for a good explanation

Interpretable

- Humans can easily interpret reasoning





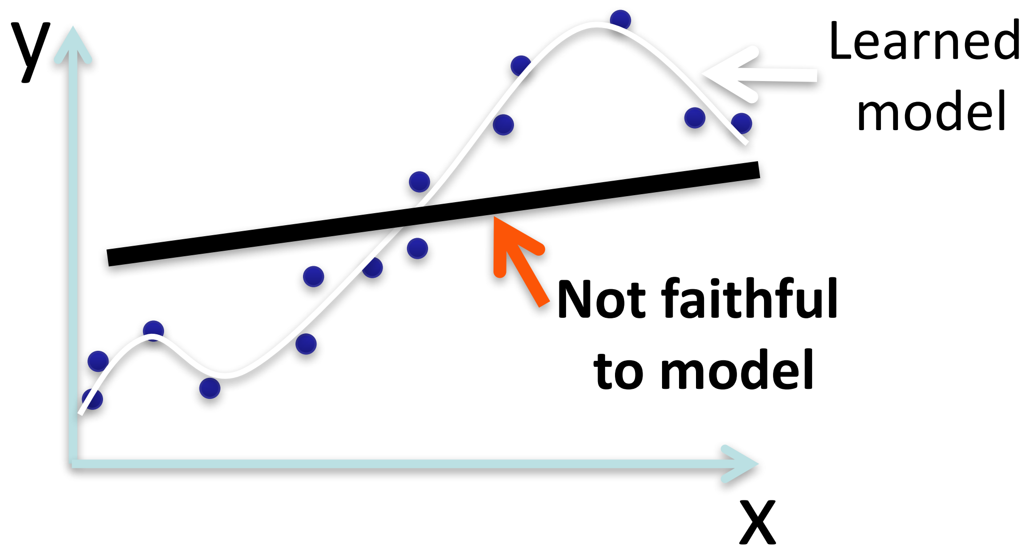
Definitely
not interpretable

Potentially
interpretable

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

Desiderata for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves



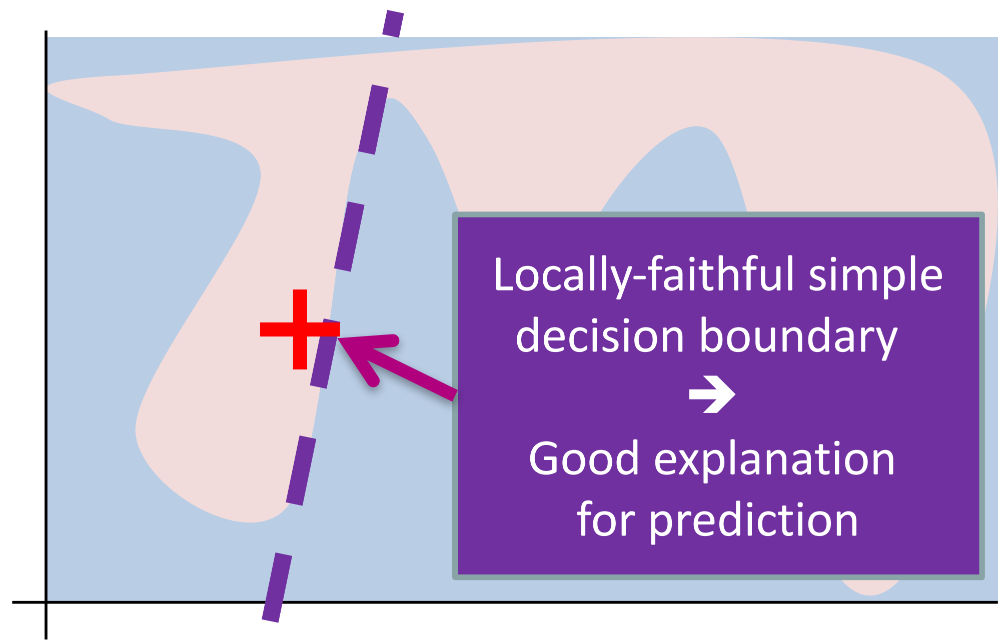
Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

LIME – Key Ideas

1. Pick a model class interpretable by humans
   - Not globally faithful… ☹

Line,
shallow decision tree,
sparse features, …

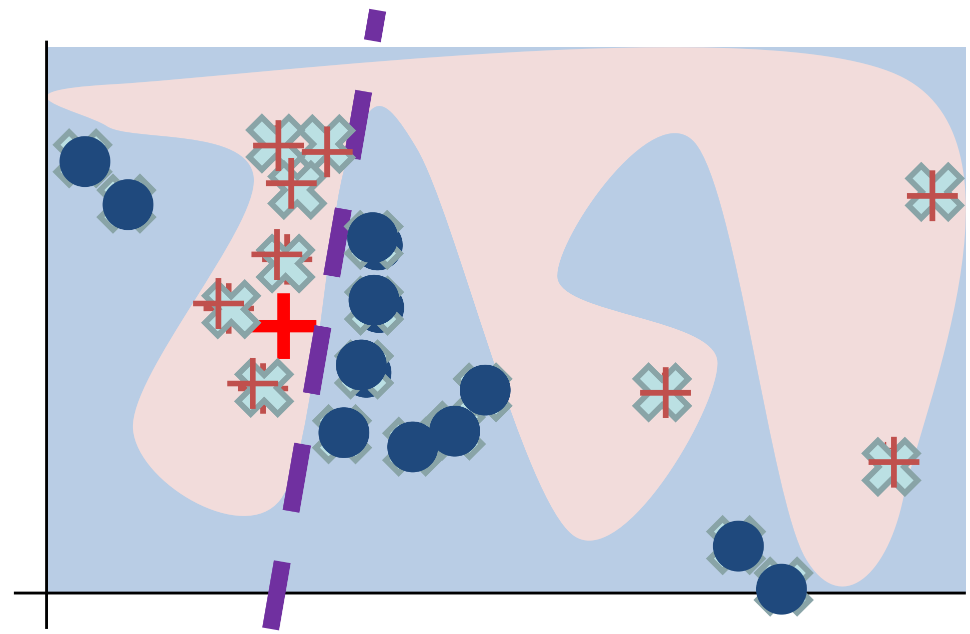
2. Locally approximate global (blackbox) model
   - Simple model globally bad, but locally good



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

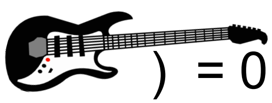
Using LIME to explain a complex model's prediction for input x_i

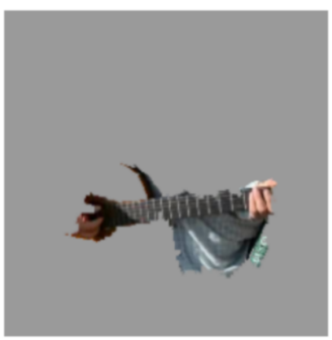
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain

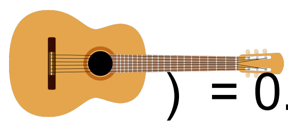


Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

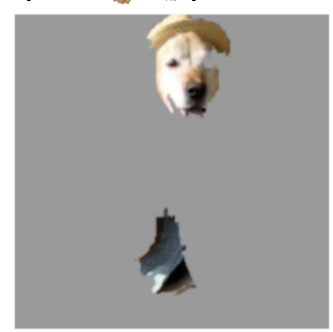
Explaining Google's Inception NN

P() = 0.32



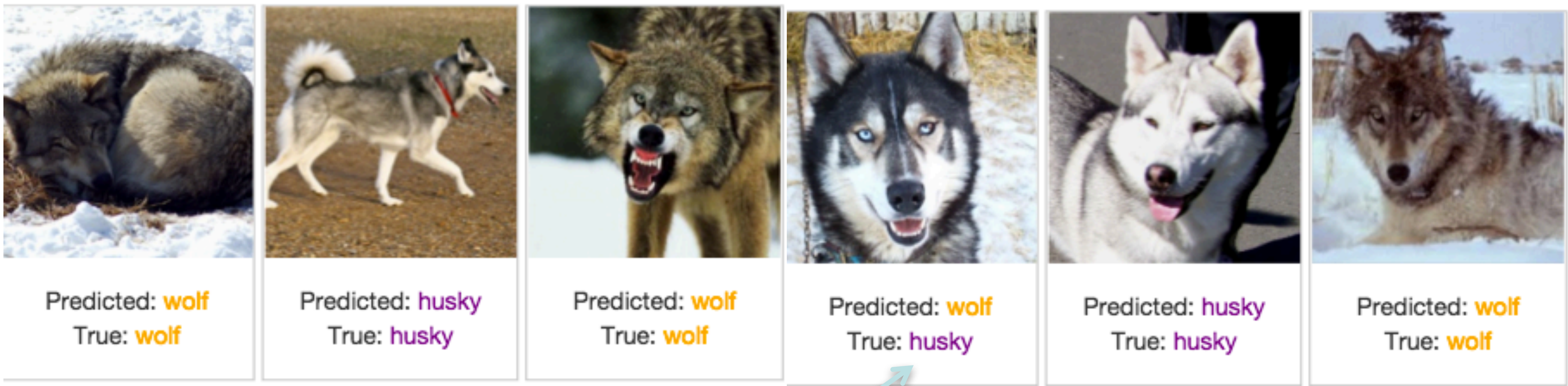
P() = 0.24

P() = 0.21



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

Train a neural network to predict wolf v. husky

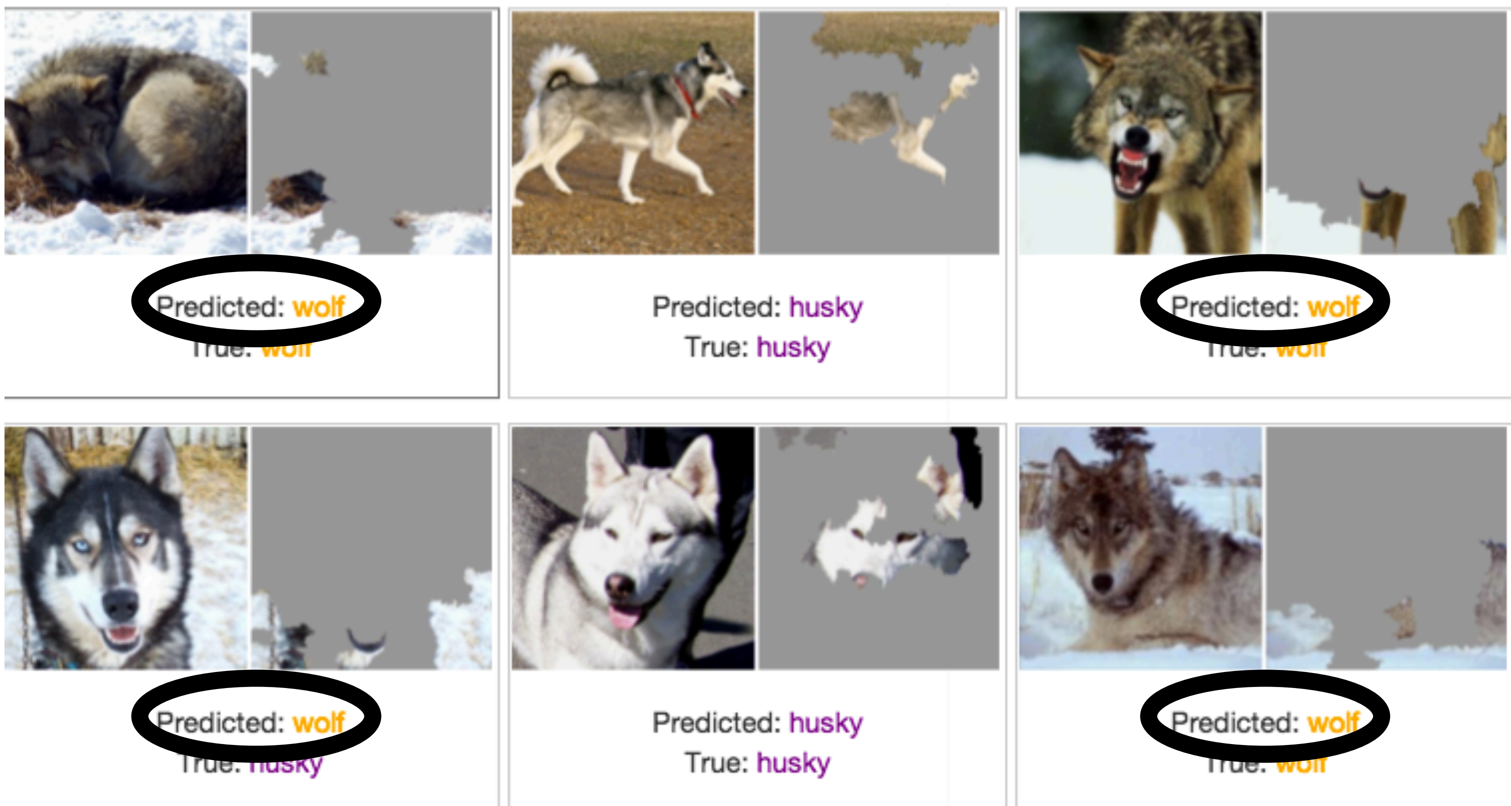


Only 1 mistake!!!

Do you trust this model?
How does it distinguish between huskies and wolves?

Slide adapted from Marco Ribeiro – see “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

LIME Explanation for neural network prediction



It’s a great snow detector… ☹

Slide adapted from Marco Ribeiro – see “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

Outline

- Distractions *vs.*
- Important Concerns
  - Sorcerer's Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment

Data Risk

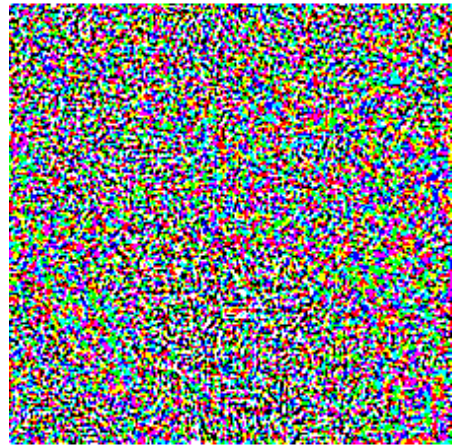
- Quality of ML Output Depends on Data…
- Three Dangers:
  - Training Data Attacks
  - Adversarial Examples
  - Bias Amplification

Attacks to Training Data

Adversarial Examples

+

0.007 ×



=

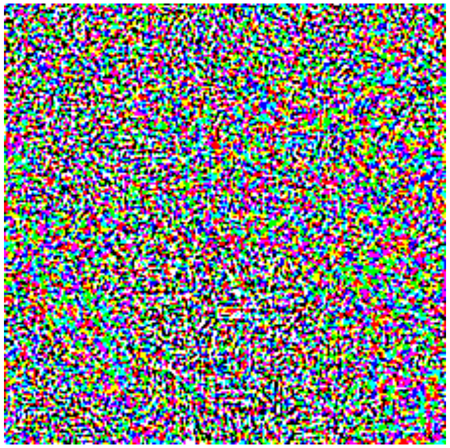
57% Panda

Access to NN parameters

“Explaining and harnessing adversarial examples,” I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

Adversarial Examples

+ 0.007 ×



=

57% Panda

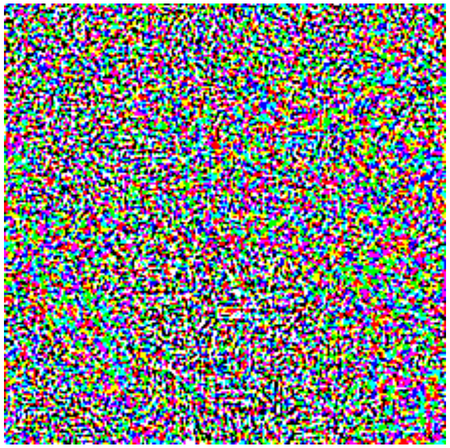
99.3% Gibbon

Access to NN parameters

“Explaining and harnessing adversarial examples,” I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

Adversarial Examples

$+ \; 0.007 \times$



$=$

57% Panda

99.3% Gibbon

Only need x
Queries to NN

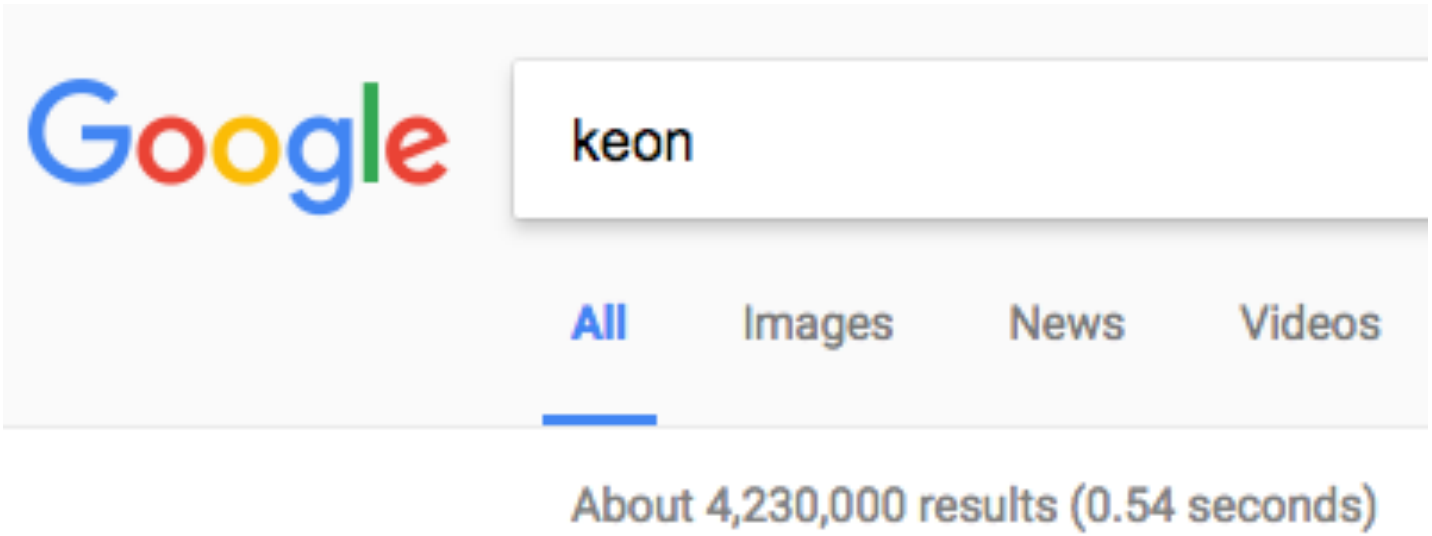
Attack is robust to fractional changes in training data, NN structure

“Explaining and harnessing adversarial examples,” I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

Data Risk

- Quality of ML Output Depends on Data…
- Three Dangers:
  - Training Data Attacks
  - Adversarial Examples
  - ***Bias Amplification***
    - **Existing training data reflects our existing biases**
    - **Training ML on such data…**

Racism in Search Engine Ad Placement



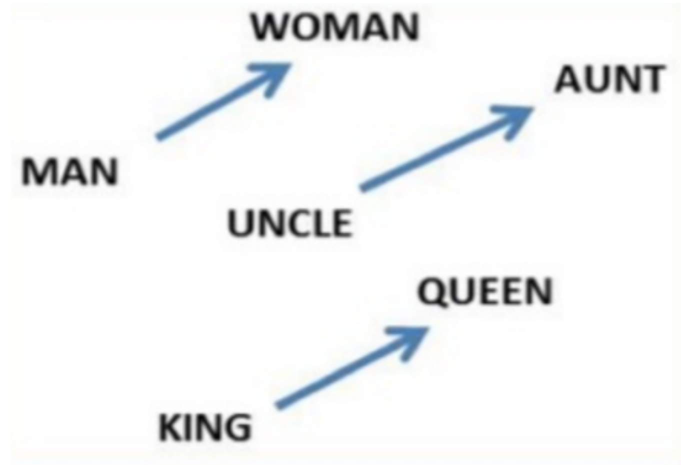
Searches of ‘black’ first names

25% more likely to include ad for criminal-records background check

Searches of ‘white’ first names

2013 study https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240

Automating Sexism



- Word embeddings
- Word2vec trained on 3M words from Google news corpus
- Allows analogical reasoning
- Used as features in machine translation, etc., etc.

man : king ↔ woman : queen

sister : woman ↔ brother : man

man : computer programmer ↔ woman : homemaker

man : doctor ↔ woman : nurse

https://arxiv.org/abs/1607.06520

Illustration credit: Abdullah Khan Zehady, Purdue

In fact…

“Housecleaning Robot”

Google image search returns…

Not…

Predicting Criminal Conviction from Driver Lic. Photo

Convicted Criminals

Non-Criminals

?

- Convolutional neural network
- Trained on 1800 Chinese drivers license photos
- **90% accuracy**

https://arxiv.org/pdf/1611.04135.pdf

Should prison sentences be based on crimes that haven't been committed yet?

- US judges use proprietary ML to predict recidivism risk

NORTHPOINTE

- Much more likely to mistakenly flag black defendants
  - Even though race is not used as a feature

http://go.nature.com/29aznyw
https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.odaMKLgrw
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

What *is* Fair?

A	Protected attribute (*eg,* race)
X	Other attributes (*eg,* criminal record)
Y’ = f(X,A)	Predicted to commit crime
Y	Will commit crime

- Fairness through unawareness

  Y’ = f(X) not f(X, A) but Northpointe satisfied this!

- Demographic Parity

  $Y' \perp\!\!\!\perp A$ i.e. $P(Y'=1 \mid A=0)=P(Y'=1 \mid A=1)$

  Insufficient: can predict white criminals, black randomly

  Furthermore, if $Y \not\perp\!\!\!\perp A$, it rules out ideal predictor Y’=Y

C. Dwork et al. “Fairness through awareness” ACM ITCS, 214-226, 2012

What *is* Fair?

A	Protected attribute (*eg,* race)
X	Other attributes (*eg,* criminal record)
Y' = f(X,A)	Predicted to commit crime
Y	Will commit crime

- Calibration within groups

  $Y \perp\!\!\!\perp A \mid Y'$

  No incentive for judge to ask about A

- Equalized odds

  $Y' \perp\!\!\!\perp A \mid Y$ *i.e.* $\forall y,\ P(Y'=1 \mid A=0, Y=y) = P(Y'=1 \mid A=1, Y=y)$

  Same rate of false positives & negatives

- Can't achieve both!

  Unless $Y \perp\!\!\!\perp A$ or Y' perfectly = Y

J. Kleinberg et al "Inherent Trade-Offs in Fair Determination of Risk Score" [arXiv:1609.05807v2]()

Guaranteeing Equal Odds

Given any predictor, Y'

Can create a new predictor satisfying equal odds

Linear program to find convex hull

Bayes-optimal ***computational affirmative action***

- Calibration within groups

  $Y \perp\!\!\!\perp A \mid Y'$

  No incentive for judge to ask about A

- Equalized odds

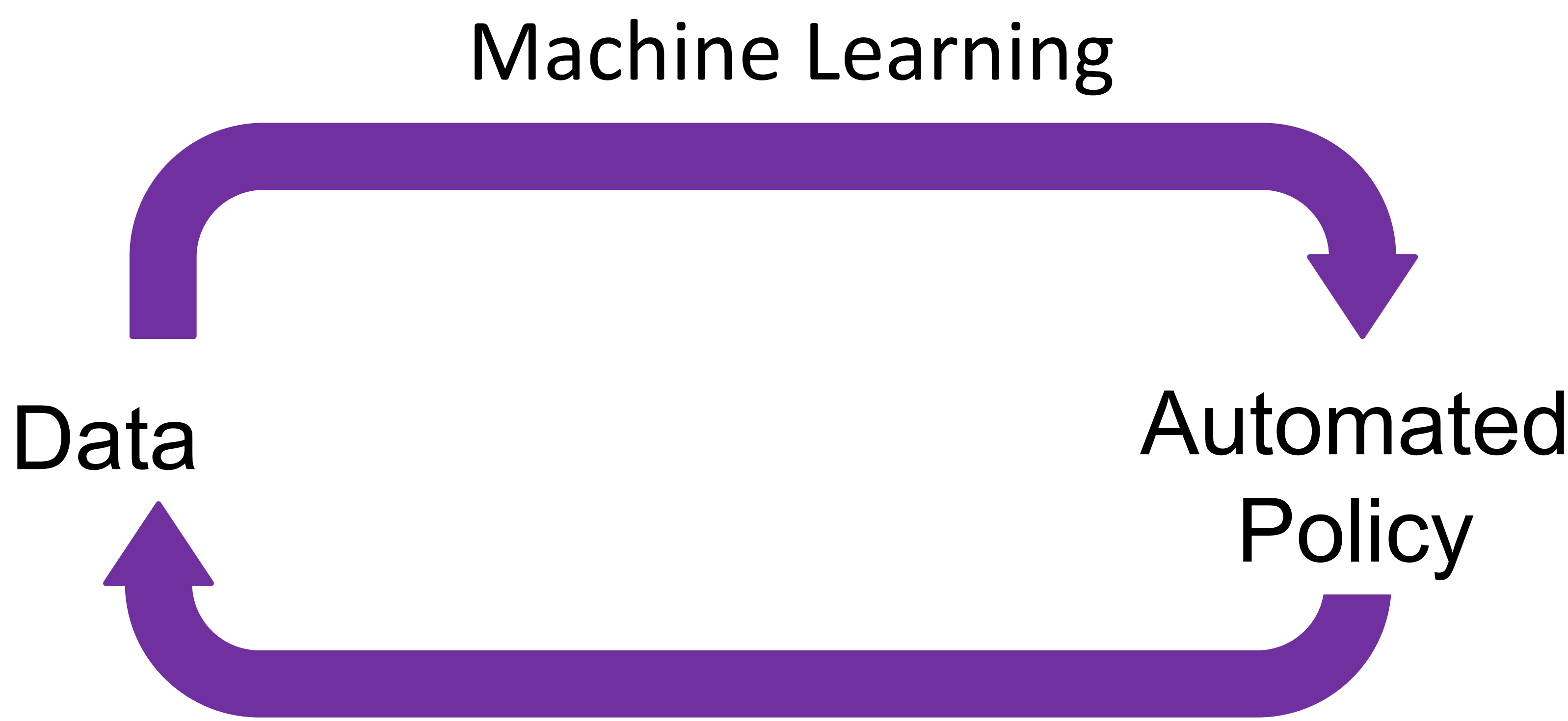
  $Y' \perp\!\!\!\perp A \mid Y$ *i.e.* $\forall y,\ P(Y'=1 \mid A=0, Y=y) = P(Y'=1 \mid A=1, Y=y)$

  Same rate of false positives & negatives

M. Hardt *et al* "Equality of Opportunity in Supervised Learning" **arXiv:1610.02413v1**

Important to get this Right!

Feedback Cycles



Appeals & Explanations

Must an AI system explain itself?

- Tradeoff between accuracy & explainability
- How to guarantee than an explanation is right

Liability?

- Microsoft?
- Google?
- Biased / Hateful people who created the data?

- Legal standard
  - Criminal intent
  - Negligence

Deploying AI → criminal acts
without a perpetrator
– Ryan Calo

Liability II

- Stephen Cobert's twitter-bot
  - Substitutes FoxNews personalities into Rotten Tomato reviews
  - Tweet implied Bill Hemmer took communion while intoxicated.
- Is this libel (defamatory speech)?

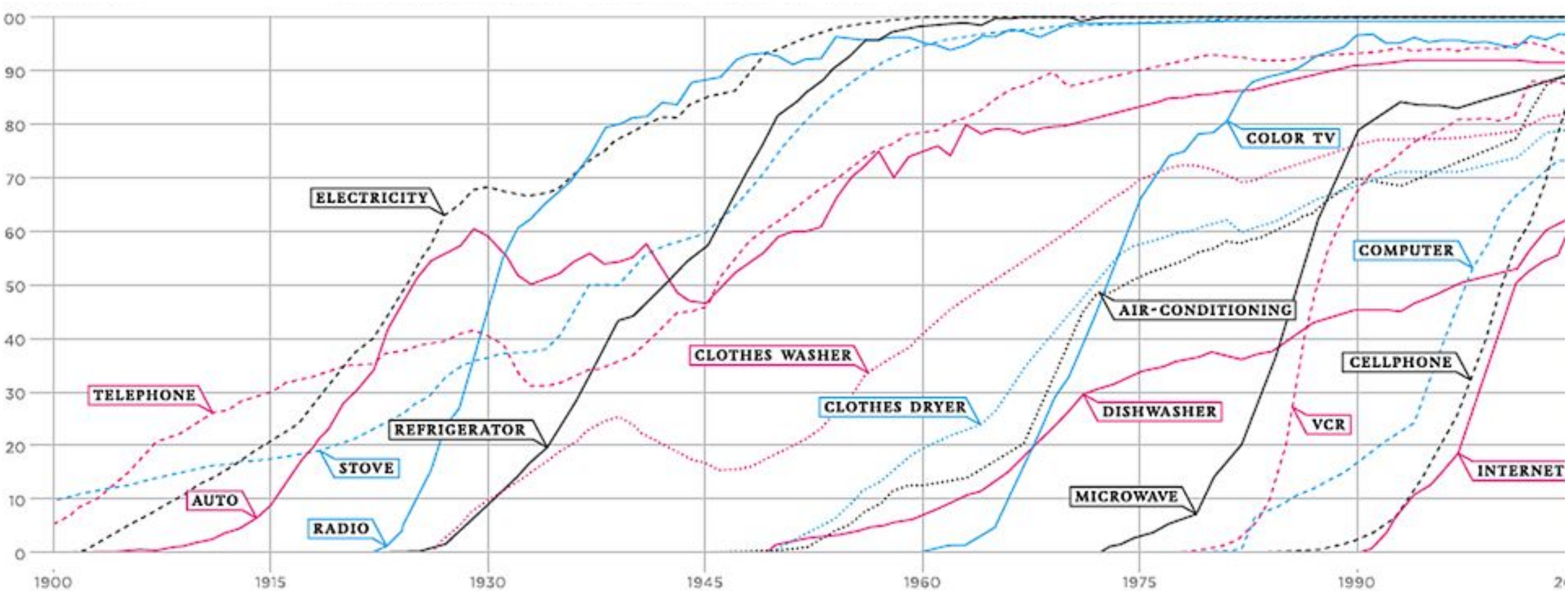
http://defamer.gawker.com/the-colbert-reports-new-twitter-feed-praising-fox-news-1458817943

Understanding Limitations

How to convey the limitations of an AI system to user?

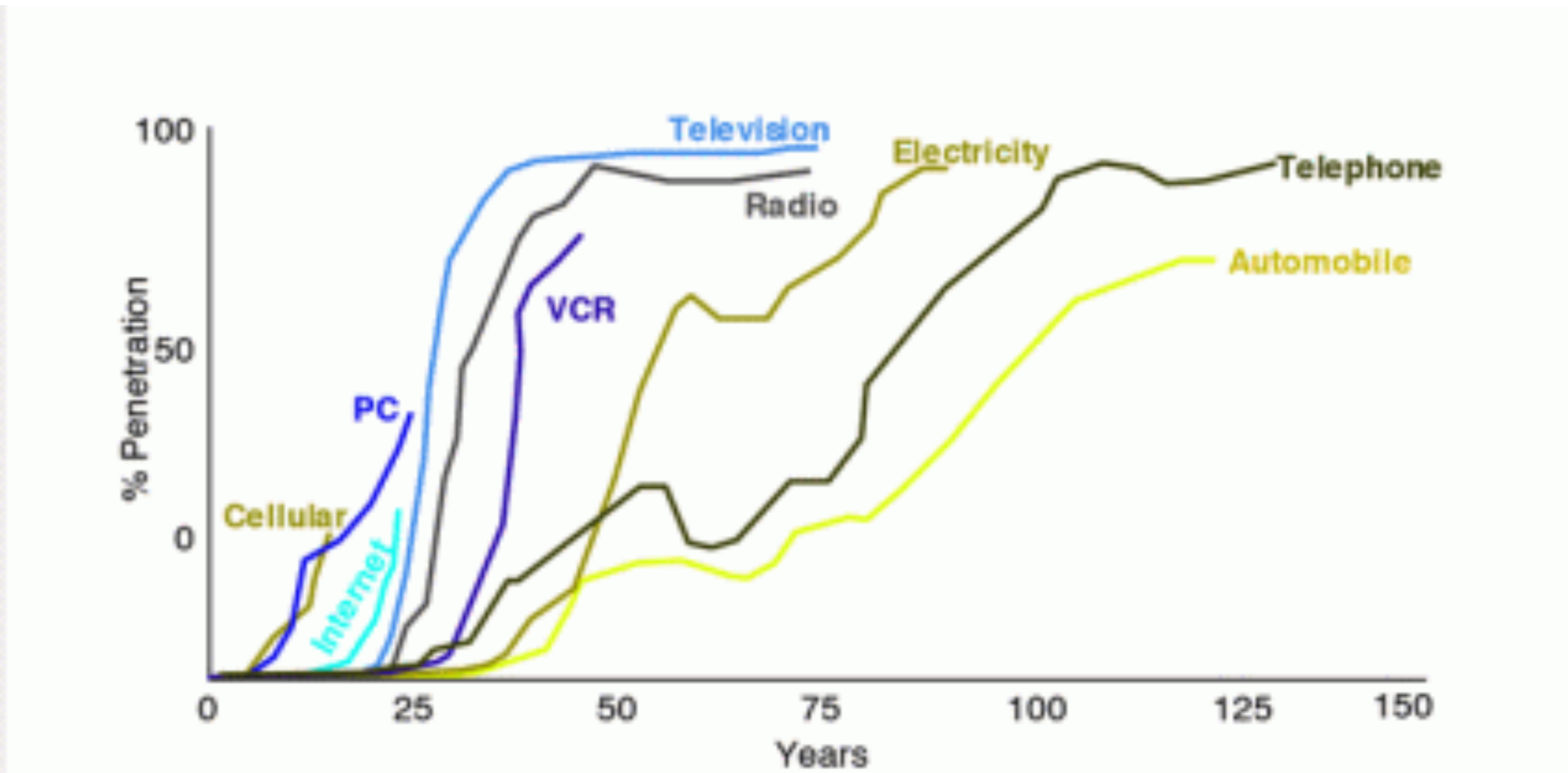
- Challenge for self-driving car
- Or even adaptive cruise control (parked obstacle)
- Google Translate

Exponential Growth → Hard to Predict Tech Adoption



Adoption Accelerating

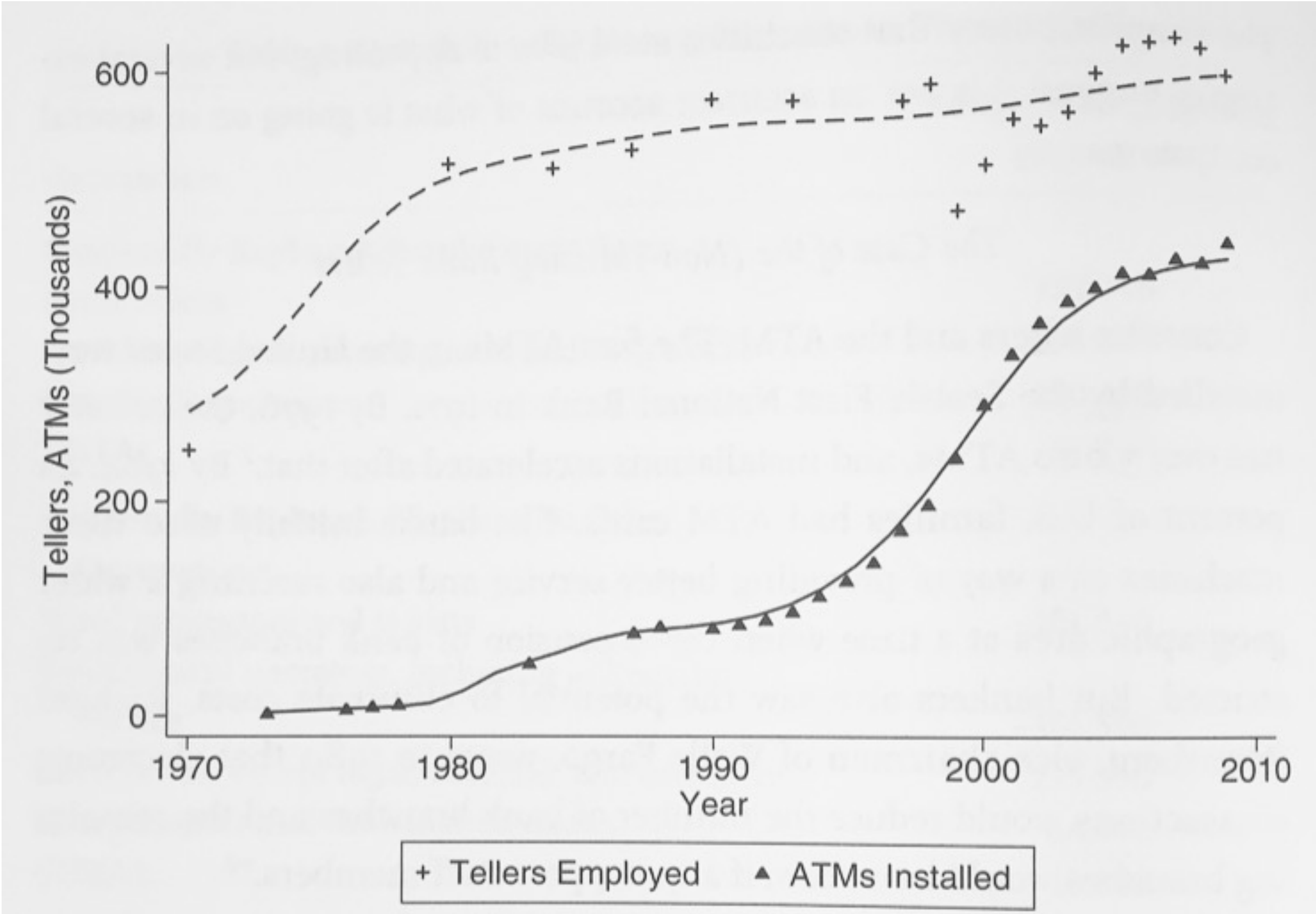
Newer technologies taking hold at double or triple the rate



Self-Driving Vehicles

- 6% of US jobs in trucking & transportation
- What happens when these jobs eliminated?
- Retrained as programmers?

Hard to Predict



http://www.aei.org/publication/what-atms-bank-tellers-rise-robots-and-jobs/

Conclusions

- Distractions *vs.*
- Important Concern
  - Sorcerer’s Apprenti
    - Specifying Constraint
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment

> People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.
> - Pedro Domingos

Thanks

- Formative discussions with
  - Gagan Bansal, Ryan Calo, Oren Etzioni, Jeff Heer, Rao Kambhampati, Mausam, Tongshuang Wu

- Research Sponsors

future of life INSTITUTE

Washington Research FOUNDATION

Google

- Inverse revinforcement learning
- Structural estimation of MDPs
- Inverse optimal control

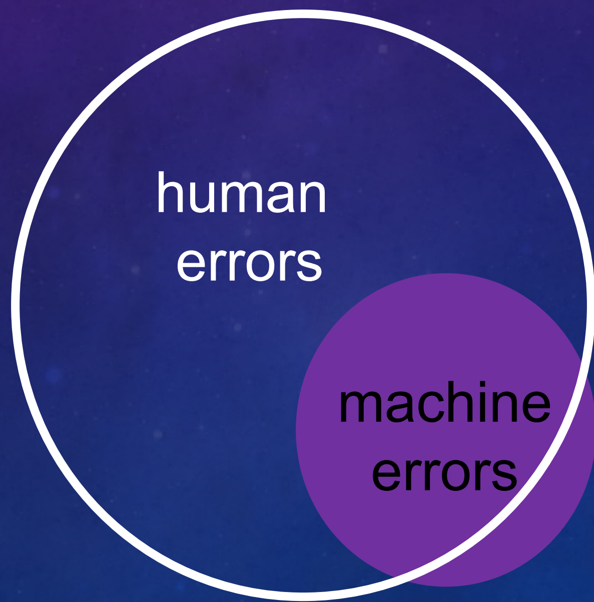
- But don't want agent to adopt human values
    - Watch me drink coffee -> not want coffee itself

    - Cooperative inverse RL
        - Two player game
- Off swicth function
    - Don't given robot an objective
    - Instead it must allow for uncertainty about human objctive
        - If human is trying to turn me off, then it must want that
- Uncertainty in objectives – ignored
    - Irrelevant in standard decision problems; unless env provides info on reward

DEPLOYING AI

What is bar for deployment?

- System is better than person being replaced?
- Errors are ***strict subset*** of human errors?



- Reward signals
  - Wireheading
  - RL agent hijacks reward
  - Traditiomnal RL
    - Enivironment provide reward signal. Mistak!
  - Instead env reward signal is not true reward
    - Just provides INOFRMATION about reward
  - So hijacking reward signal is pointless
    - Doesn’t provide more reward
    - Just provides less information

- Y Lecunn – common view
- All ai success is supervised (deep) MLL
- Unsupervised is key challenge
  - Fill in occluded immage
  - Fill in missing words in text, sounds in speech
  - Consquences of actions
  - Seq of actions leading to observed situation
- Brain has 10E14 synapses but live for only 10e9 secs, so more params than data
  - 100 years * 400 days * 25 hours = 100k hours. 3600 seconds
- Types
  - RL a few bits / trial
  - Supervisesd 10-10000 bits trial
  - Unsupervise – millions bits / trial, but unreliable
    - Dark matter of AI
- Thier FAIR system won visdoom challenge – sub for pub ICML or vision conf 2017
- Sutton's dyna arch

- Transformation of ML
  - Learning as minimizing loss function →
  - Learning as finding nash equilibrium in 2 player game
- Hierarchical deep RL
  - Concept formation (abstraction, unsupervised ML)