# Machine Learning isn’t Magic

David Adams

# Overview

1. Introduction and Context
2. Machine Learning (in 10 minutes)
3. Requirements for ML in medicine
4. Doing ML right is harder than it looks
5. Takeaways

# Introduction and Context

# I’m David Adams, background includes Physics, CS, Machine Learning



Bachelor’s **Physics**

Bachelor’s **Computer Science**

Ph.D. **Physics**

Master’s **Computer Science**

# I’m a Data Scientist at Verily (formerly at Google in Search Ads)

I’ve spent years seeing the work of, and learning from, some of the best minds in Machine Learning, like D. Sculley. These practitioners have been trailblazing and “making ML work” for well over a decade.

# I’m here to change the conversation about ML in healthcare


While at the AI in Healthcare Summit a few months ago there was a panel discussion where someone said, “*If my model has a better AUC*[1]*, the Dr. shouldn’t object*”

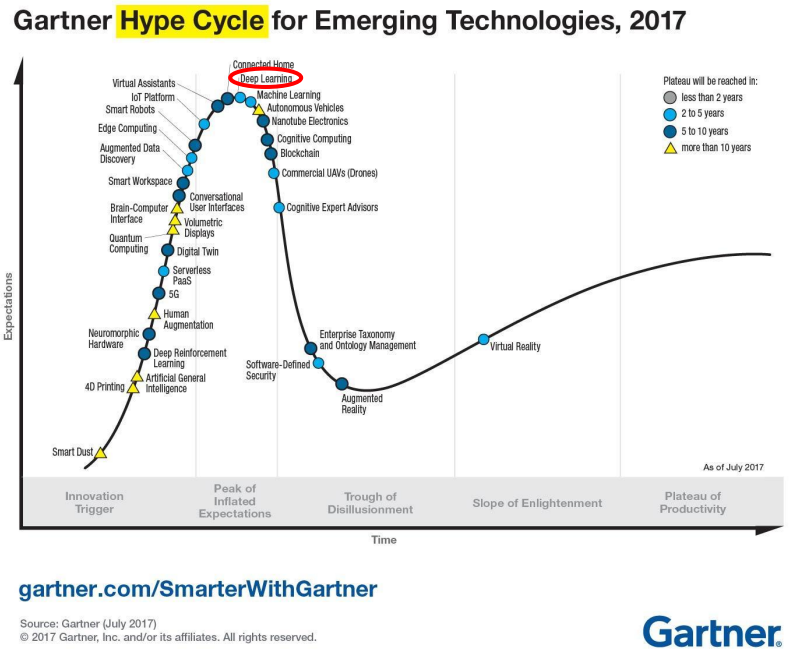
[1]AUC: Area Under the (Receiver Operator Characteristic) Curve

# I’m here to change the conversation about ML in healthcare


While at the AI in Healthcare Summit a few months ago there was a panel discussion where someone said, “*If my model has a better AUC*[1]*, the Dr. shouldn’t object*”

[1]AUC: Area Under the (Receiver Operator Characteristic) Curve

# I’m optimistic about ML’s future impact  (healthcare included)




gartner.com/SmarterWithGartner

Source: Gartner (July 2017)



Gartner

In some cases, I’m more positive than other people in the field: I worked on a project to show deep learning could beat the best humans in Go (dropped it once I learned DeepMind was going to put forward a serious effort).

# Physicians are right to be skeptical of Artificial Intelligence/ML

Physicians have well-honed skepticism when it comes evaluating safety and efficacy of drugs and devices. The same should be applied to ML because…

# Physicians are right to be skeptical of Artificial Intelligence/ML

Physicians have well-honed skepticism when it comes evaluating safety and efficacy of drugs and devices. The same should be applied to ML because…

## Machine Learning isn’t Magic

# ML hype and misuse can lead to terrible accidents

**Serious expectations mismatch** has, so far, led to ~~two~~ three fatal accidents

# ML hype and misuse can lead to terrible accidents

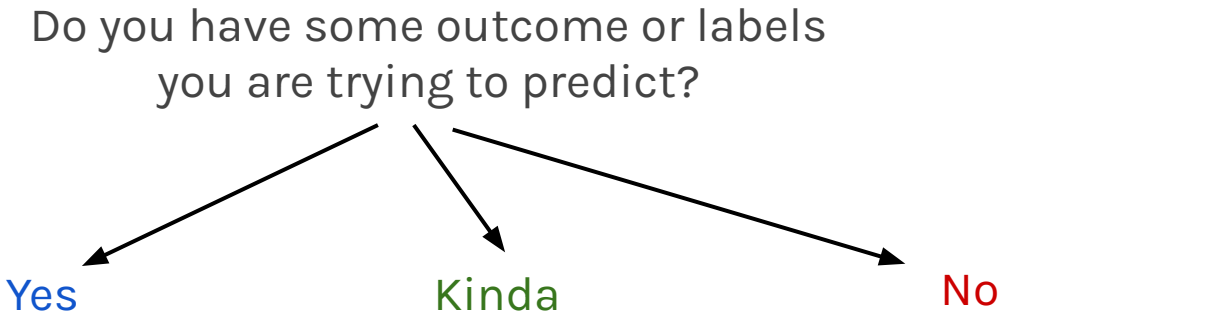
**Serious expectations mismatch** has, so far, led to ~~two~~ three fatal accidents

Appropriate and informed skepticism is what will prevent unsafe ML systems from harming patients

At this point, it's worth pointing out that this talk is self-serving. I want ML to succeed in medicine as soon as possible, which will be delayed if it seriously harms patients

# Machine Learning (in 10 minutes)

# ML is used to describe several types of problems



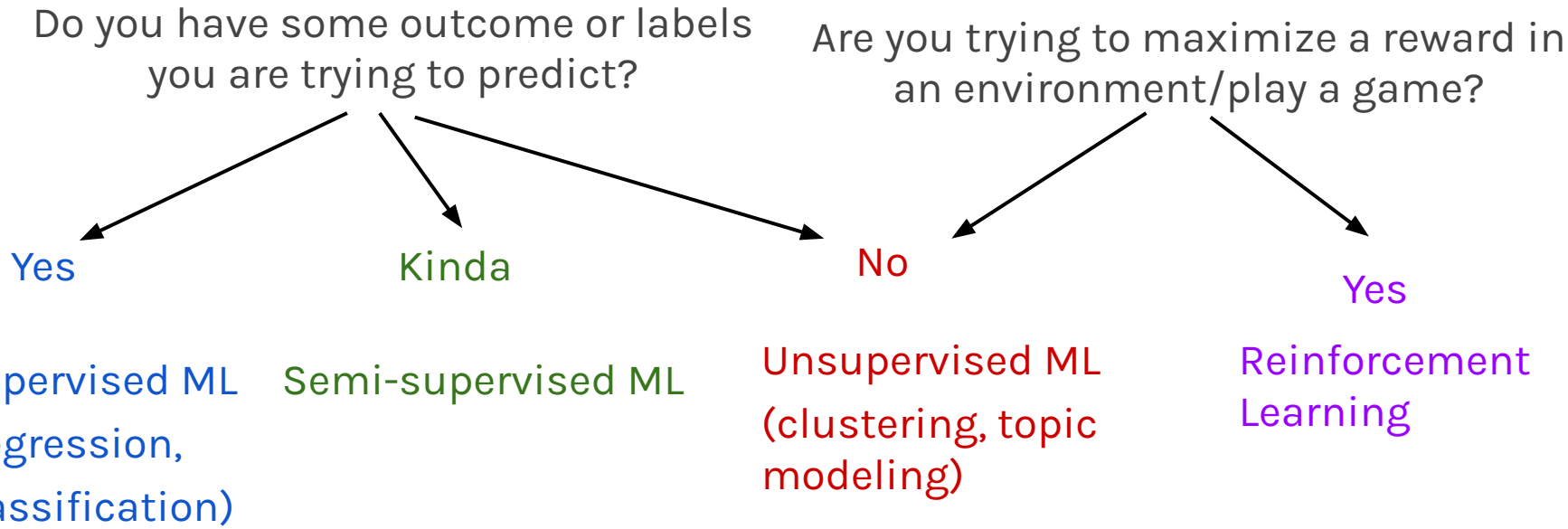

Supervised ML (regression, classification)

Semi-supervised ML

Unsupervised ML (clustering, topic modeling)

I'll primarily be talking about Supervised ML in the rest of this talk

# ML is used to describe several types of problems




I'll primarily be talking about Supervised ML in the rest of this talk

# The underpinning of (parametric) supervised ML is the stats. of regression analysis

$$Y \simeq f(\boldsymbol{X}, \boldsymbol{\theta})$$

$Y$: Independent variable (outcome/label)

$\boldsymbol{X}$: Dependent variables (features)

$\boldsymbol{\theta}$: Unknown parameters (model coefficients)

$f$: Model function (model)

The clearest distinction between stats and ML is the dimensionality of the features ($\boldsymbol{X}$)

# The underpinning of (parametric) supervised ML is the stats. of regression analysis

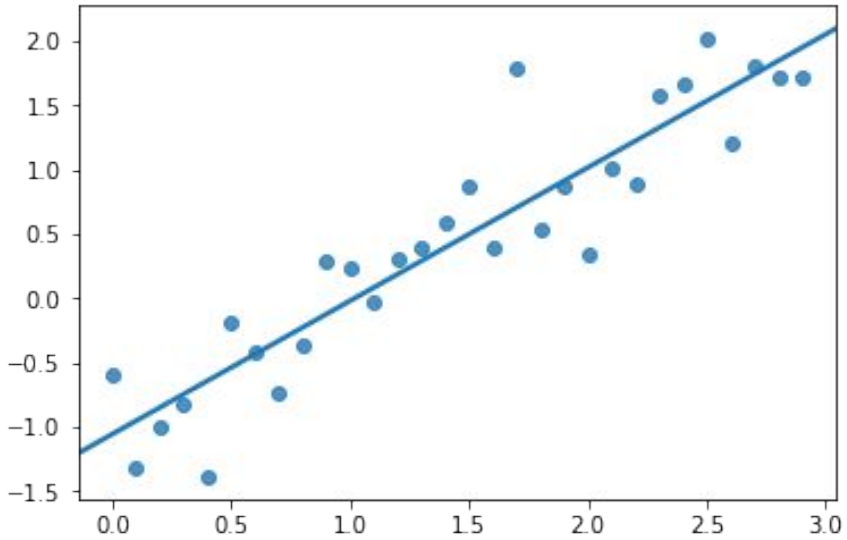
$$Y \simeq f(\boldsymbol{X}, \boldsymbol{\theta})$$

$Y$: Independent variable (outcome/label)

$\boldsymbol{X}$: Dependent variables (features)

$\boldsymbol{\theta}$: Unknown parameters (model coefficients)

$f$: Model function (model)




$$Y \simeq \boldsymbol{\theta}^{T}\boldsymbol{X}$$

$$Y \simeq \boldsymbol{\theta}_{0} + \boldsymbol{\theta}_{1}\boldsymbol{x}$$

The clearest distinction between stats and ML is the dimensionality of the features ($\boldsymbol{X}$)

# The underpinning of (parametric) supervised ML is the stats. of regression analysis

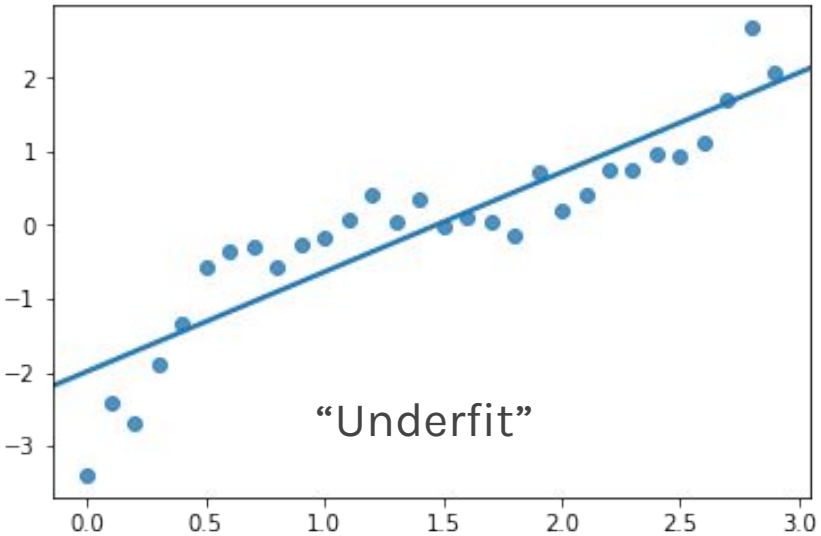
$$Y \simeq f(\boldsymbol{X}, \boldsymbol{\theta})$$

$Y$: Independent variable (outcome/label)

$\boldsymbol{X}$: Dependent variables (features)

$\boldsymbol{\theta}$: Unknown parameters (model coefficients)

$f$: Model function (model)




$$Y \simeq \boldsymbol{\theta}^T \boldsymbol{X}$$

$$Y \simeq \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 \boldsymbol{x}$$

The clearest distinction between stats and ML is the dimensionality of the features ($\boldsymbol{X}$)

# The underpinning of (parametric) supervised ML is the stats. of regression analysis

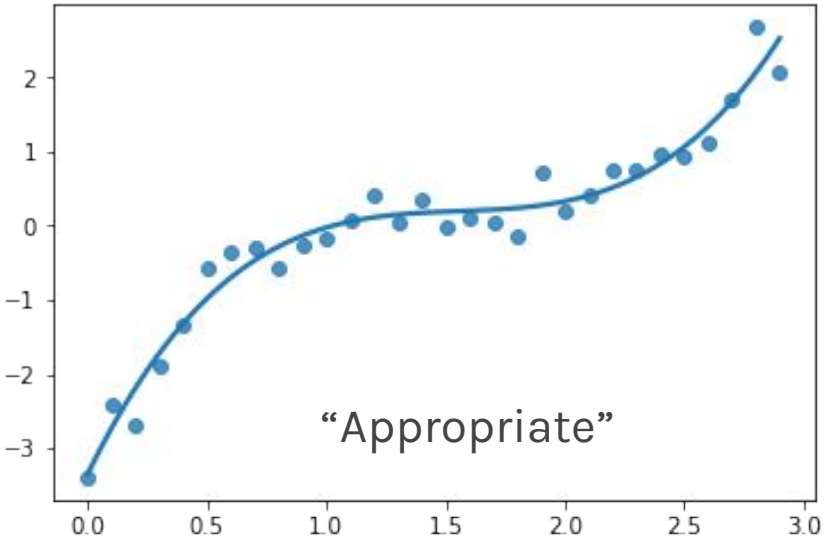
$$Y \simeq f(\boldsymbol{X}, \boldsymbol{\theta})$$

$Y$: Independent variable (outcome/label)

$\boldsymbol{X}$: Dependent variables (features)

$\boldsymbol{\theta}$: Unknown parameters (model coefficients)

$f$: Model function (model)




$$Y \simeq \boldsymbol{\theta}^T \boldsymbol{X}$$

$$Y \simeq \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 x + \boldsymbol{\theta}_2 x^2 + \boldsymbol{\theta}_3 x^3$$

The clearest distinction between stats and ML is the dimensionality of the features ($\boldsymbol{X}$)

# The underpinning of (parametric) supervised ML is the stats. of regression analysis

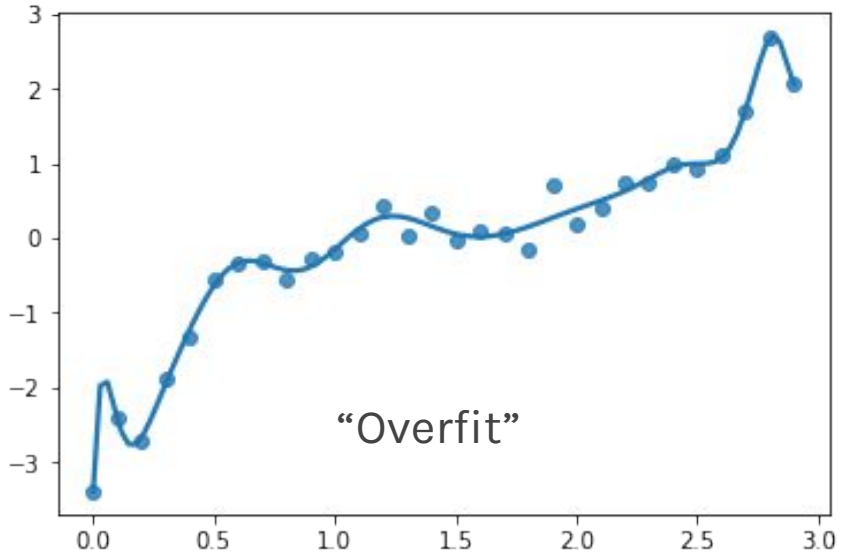
$$Y \simeq f(\boldsymbol{X}, \boldsymbol{\theta})$$

$Y$: Independent variable (outcome/label)

$\boldsymbol{X}$: Dependent variables (features)

$\boldsymbol{\theta}$: Unknown parameters (model coefficients)

$f$: Model function (model)

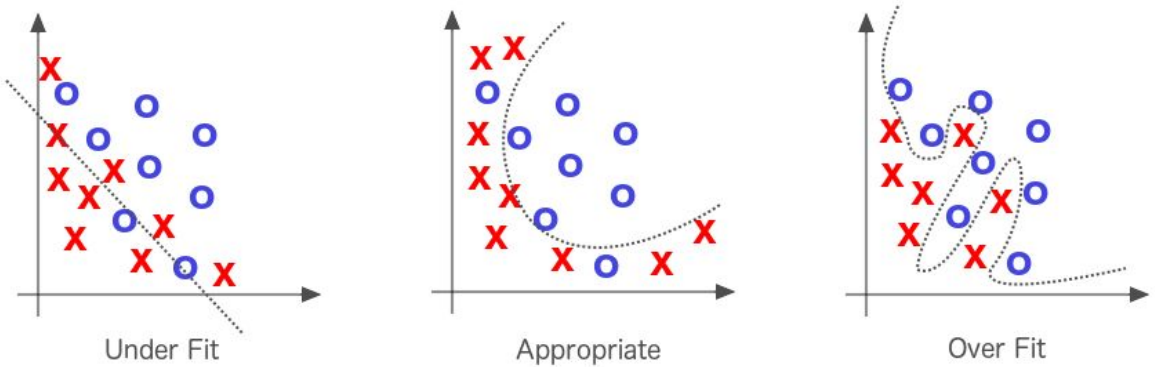



$$Y \simeq \boldsymbol{\theta}^T \boldsymbol{X}$$

$$Y \simeq \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 \boldsymbol{x} + \boldsymbol{\theta}_2 \boldsymbol{x}^2 + \boldsymbol{\theta}_3 \boldsymbol{x}^3 + \boldsymbol{\theta}_4 \boldsymbol{x}^4 + \ldots$$

Model form, capacity, amount & variety of data, and regularization impact fit

# The model predicts outcome given features using training data




Training Data ← Subset of “real-world” data used for model training

Features ← Parts of the data used to predict (correlate with) the outcome

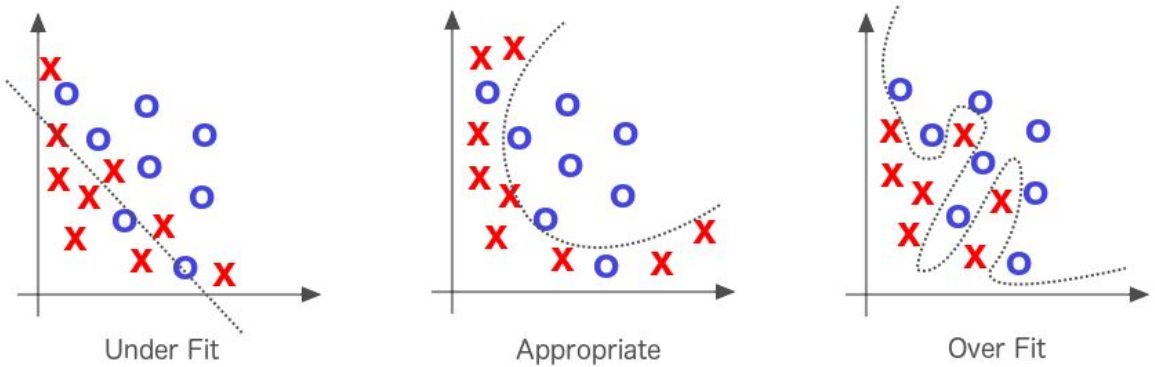
Outcome/label ← What should be predicted (may require hand-labeling)

Model ← some function that minimizes[1] your chosen error

Training/Fit ← Finding the function/model that minimizes[1] the error

[1] The error minimization is, in practice, approximate, i.e., a local minima

# The model predicts outcome given features using training data




**Training Data ← Subset of "real-world" data used for model training**

Features ← Parts of the data used to predict (correlate with) the outcome

Outcome/label ← What should be predicted (may require hand-labeling)

Model ← some function that minimizes[1] your chosen error

Training/Fit ← Finding the function/model that minimizes[1] the error

[1] The error minimization is, in practice, approximate, i.e., a local minima

# Requirements for ML in medicine

# There are a few key Ingredients to make supervised ML ***work***

- **A well-defined, unambiguous label,** e.g., IP LOS (*not* should I give this pill)
- Representative input data (without outcome/label leakage)
- High quality label (may require paying physicians to label data)

"Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy…", V. Gulshan, et al. JAMA 2016

"Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy", J. Krause, et al. Ophthalmology (In press)

# There are additional requirements for supervised ML to be ***useful*** in med


- The outcome is useful and does not suffer from **bias** (selection bias, etc.)
- The prediction is **actionable** (leads to an intervention)
- Accurate, **calibrated**, non-inferior predictions for relevant sub-populations

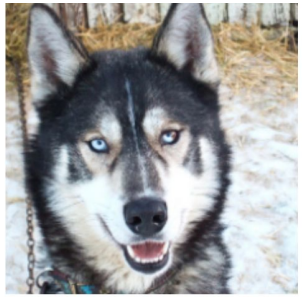
# There are add. requirements for ML to be ***trustworthy*** in medicine

- Model applicability understood (survival bias, little data for some pop., etc.)
- **Reliable,** including frequent evidence that it's performing well
- **Systems engineering** around it to improve safety
- *Interpretability (nice-to-have at this point)

"Towards A Rigorous Science of Interpretable Machine Learning", Finale Doshi-Velez and Been Kim, ArXiv (2017)

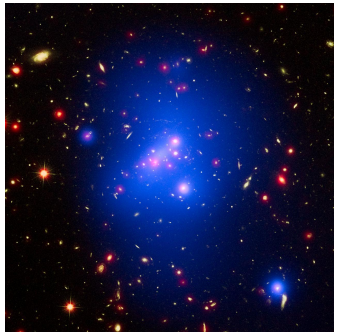
# Doing ML *right* is harder than it looks

# ML **is a correlation tool** that generalizes from examples, as naively as possible



(a) Husky classified as wolf

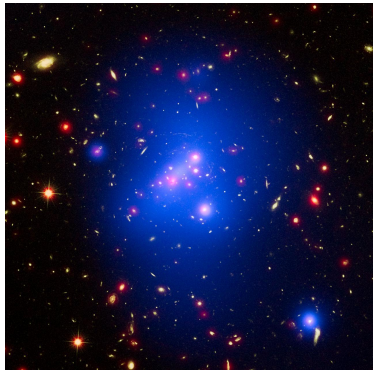
(b) Explanation

It's extremely important to bring informed skepticism to models and look for evidence that something is fishy.

Ribeiro et al. (2016) https://arxiv.org/pdf/1602.04938.pdf

# **Data/Label leakage** may be the most pernicious problem in ML




**Label leakage:** Model features have useful information about the label that won't be available when deployed.

The primary ways label leaking is detected are:

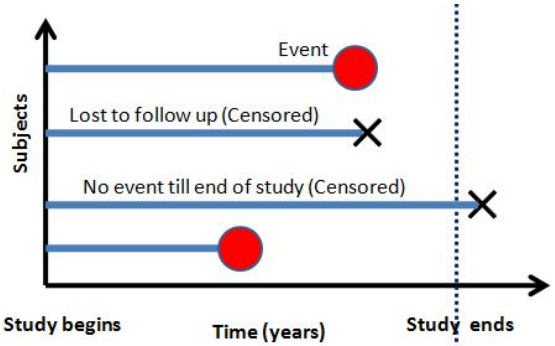
- Building a small, interpretable baseline model
- Careful evaluation of features
- **Unbelievably good performance (or feature importance)**

# *Statistical* bias can quietly come from many sources

Population Bias

Sample Population

Survival Bias



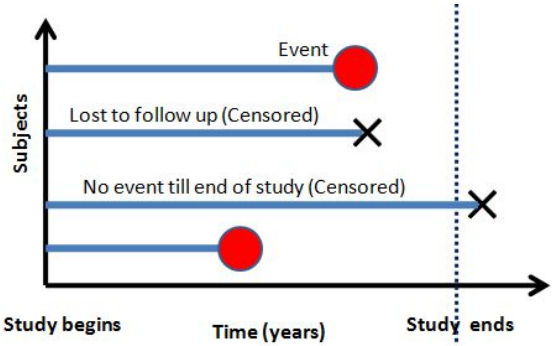

- Analysis of impact of bias
- Selection of datasets with less bias
- Modify sampling from dataset

Also, ***treatment*** (prescribe a painkiller) and ***response*** (doesn't take the pill) **bias.**

# *Statistical* bias can quietly come from many sources

Population Bias


Survival Bias




Data Cleaning

- Analysis of impact of bias
- Selection of datasets with less bias
- Modify sampling from dataset

- Minimal & documented data cleaning
- Model data missingness
- Use a model designed for missingness[1]

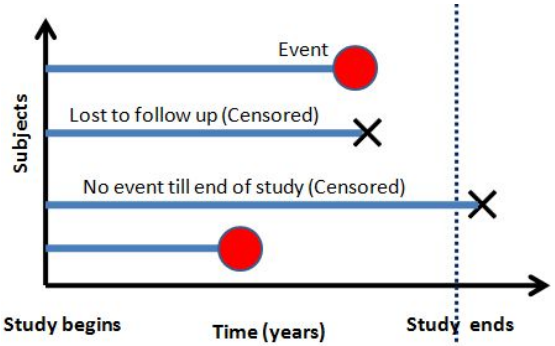
Also, ***treatment*** (prescribe a painkiller) and ***response*** (doesn't take the pill) **bias.**

[1]"Semi-supervised … Cycle Wasserstein Regression GANs" M. B.A. McDermott et al. 2018 AAAI

# *Statistical* bias can quietly come from many sources

Population Bias


Survival Bias




Data Cleaning

- Analysis of impact of bias
- Selection of datasets with less bias
- Modify sampling from dataset

+

- Minimal & documented data cleaning
- Model data missingness
- Use a model designed for missingness[1]

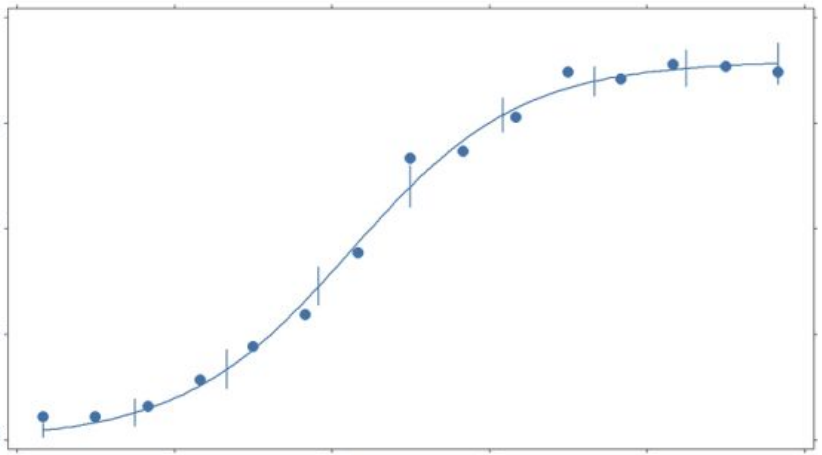
Also, ***treatment*** (prescribe a painkiller) and ***response*** (doesn't take the pill) **bias.**

**Solution:**

- Dialog with experts who know the data
- Analysis to determine impact of the bias
- (Possibly) attempt to mitigate

[1]"Semi-supervised … Cycle Wasserstein Regression GANs" M. B.A. McDermott et al. 2018 AAAI

# Don’t expect ML to figure out what problem to solve



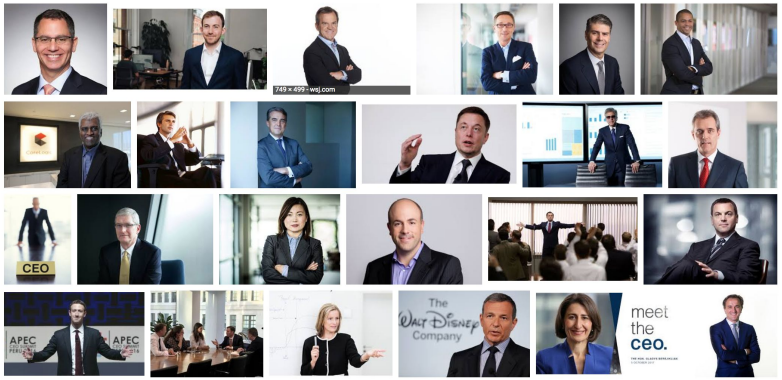
ML is a tool for *solving problems*, not for *finding useful problems*.

ML practitioners naturally gravitate towards well-posed, easy to answer questions, not the important ones!

**Solution**: Project selection and design should be a conversation between ML practitioners and the subject matter experts

# ML community is *starting to* understand the impact of other bias

## Underrepresentation




**Recognition**: Does not “work for” a group

**Denigration**: Use of culturally disparaging terms

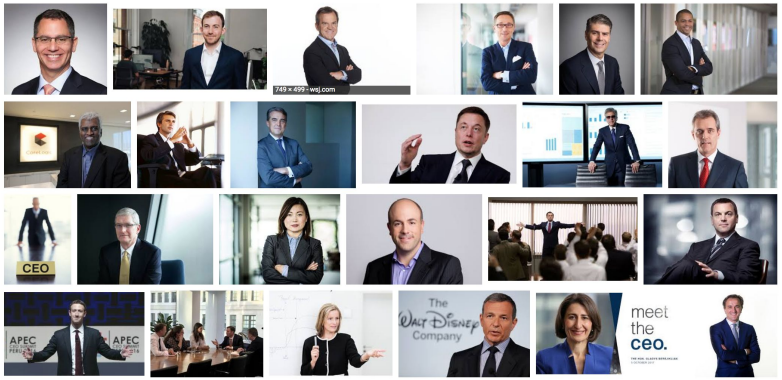
**Stereotype**: Oversimplified image or idea of a particular type of person

“The trouble with bias”, NIPS 2017 (Kate Crawford)

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

# ML community is *starting to* understand the impact of other bias

## Underrepresentation




**Recognition**: Does not "work for" a group

**Denigration**: Use of culturally disparaging terms

**Stereotype**: Oversimplified image or idea of a particular type of person


"*I tested Clips with a half-dozen kids who were Asian, African American and white, and the software seemed equally interested in them all.*" (Washington Post)

https://www.washingtonpost.com/news/the-switch/wp/2018/02/27/googles-first-camera-isnt-an-evil-all-seeing-eye-yet

"The trouble with bias", NIPS 2017 (Kate Crawford)

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

# ML in med. *in production* can learn from Google-like companies


Some of the lessons learned:

- Have a model retraining (motivation: coding changes), release process
- Monitor inputs and model behavior
- Be ready for rollback (systems engineering thought process)

“Hidden Technical Debt in Machine Learning Systems”, NIPS 2013 (D. Sculley et al.)

“Ad Click Prediction: a View from the Trenches” KDD 2013 (H. Brendan McMahan et al.)

# ML in med. *in production* can learn from Google-like companies


Some of the lessons learned:

- Have a model retraining (motivation: coding changes), release process
- Monitor inputs and model behavior
- Be ready for rollback (systems engineering thought process)
- Features (even countries) come, go, and change definition (ICD-9/10)
- If running online (dynamic), it's best to have stability guarantees
- **Beware of feedback loops (this is critical!)**

"Hidden Technical Debt in Machine Learning Systems", NIPS 2013 (D. Sculley et al.)

"Ad Click Prediction: a View from the Trenches" KDD 2013 (H. Brendan McMahan et al.)

# ML code is a tiny fraction the entire production ML system

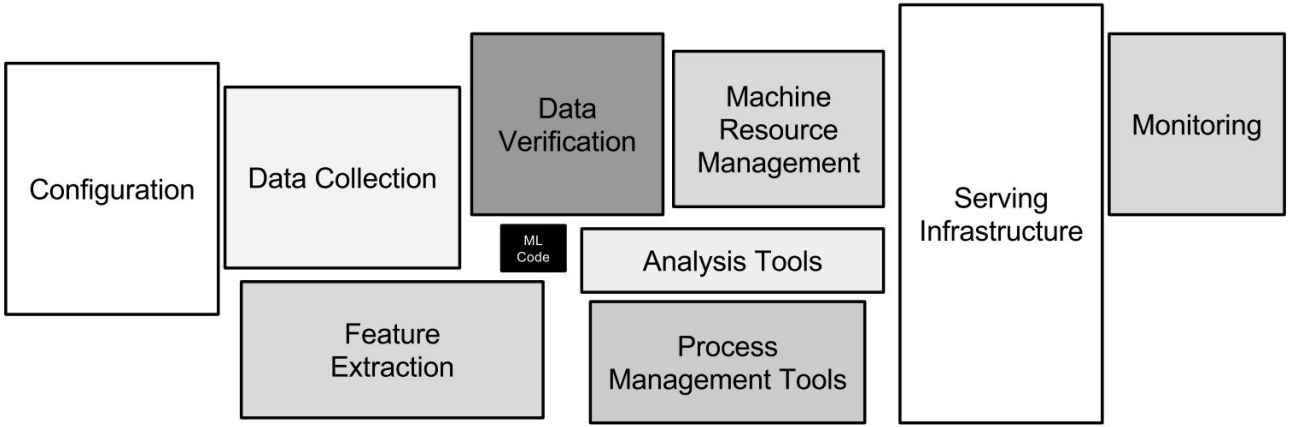



Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

The vast majority of the work required in these systems comes **after** having a high quality research-level system and result

"Hidden Technical Debt in Machine Learning Systems", NIPS 2013 (D. Sculley et al.)

# Takeaways

# The “So What” I want you to take home is ML is only a tool, not magic


- “A Good AUC” isn’t enough, need details of data used, model explanation, etc.
- A lack of skepticism/paranoia can harm
- It will be real work to get ML safely in medicine

# The “So What” I want you to take home is ML is only a tool, not magic


- “A Good AUC” isn’t enough, need details of data used, model explanation, etc.
- A lack of skepticism/paranoia can harm
- It will be real work to get ML safely in medicine
- Medicine is different than playing Go
- ML is only applicable to some problems
- First go for “easy wins”, avoid “sexy” but more dangerous problems

# Informed skepticism could prevent ML from harming patients in rare events


Some questions to ask:

*"…And how did you test this?"*

*"Why I should care/Why this is a clinically relevant endpoint?"*

*"Why should I believe this model will work in my facility?"*

# Thank You!