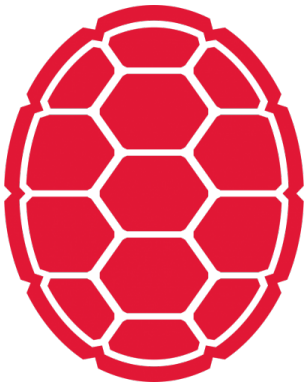


Fairness, Accountability, and Uncertainty

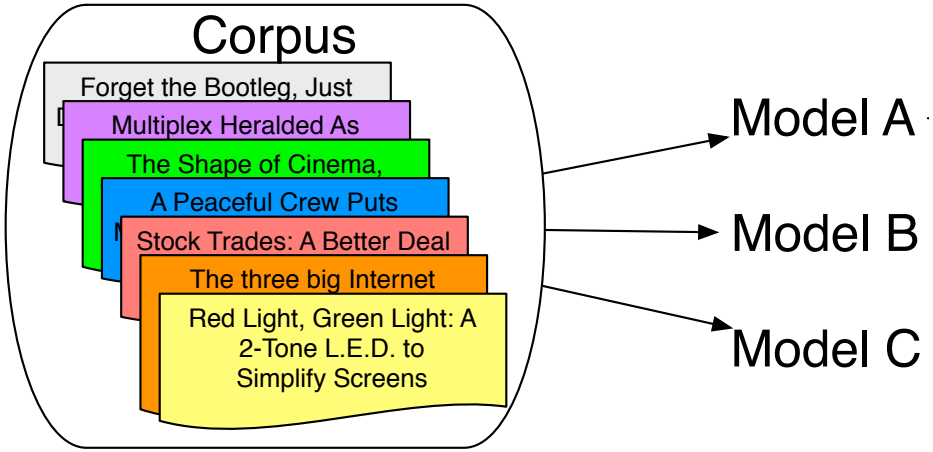
Jordan Boyd-Graber
University of Maryland
INTERPRETABILITY

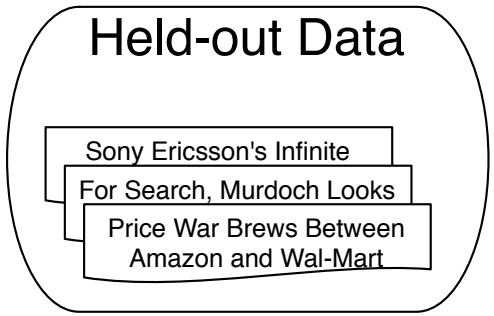
Slides adapted from Juan Ribero

Interpretability

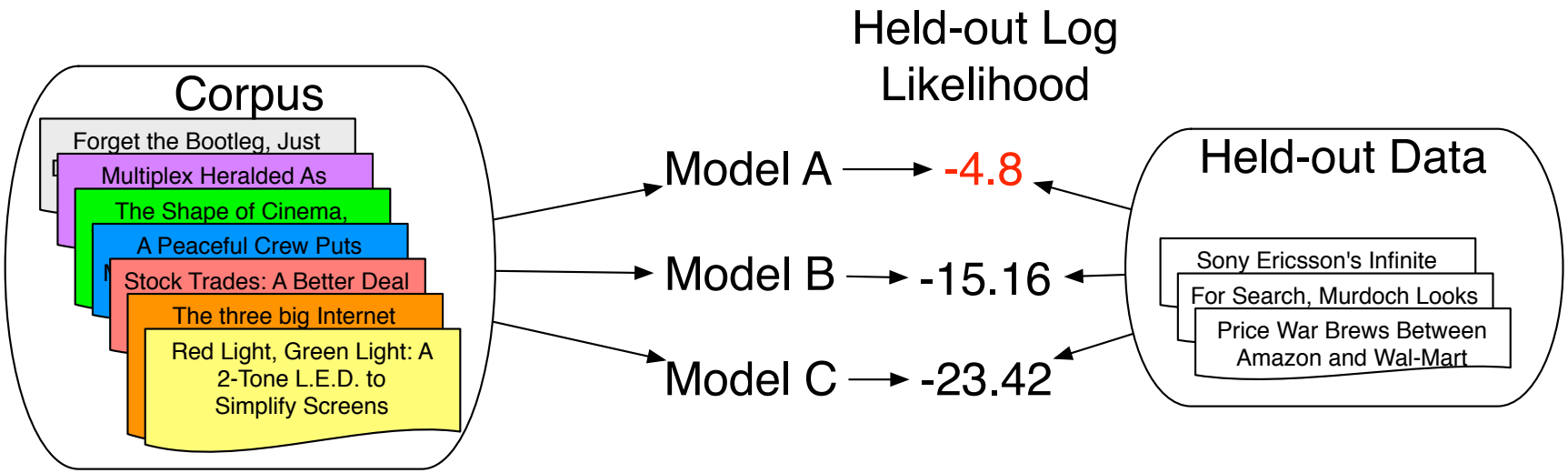
- We believe that interpretability is important
- But need to be able to measure
- Differences for supervised and unsupervised ML

Evaluation


Corpus
Forget the Bootleg, Just
Multiplex Heralded As
The Shape of Cinema,
A Peaceful Crew Puts
Stock Trades: A Better Deal
The three big Internet
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Model A
Model B
Model C

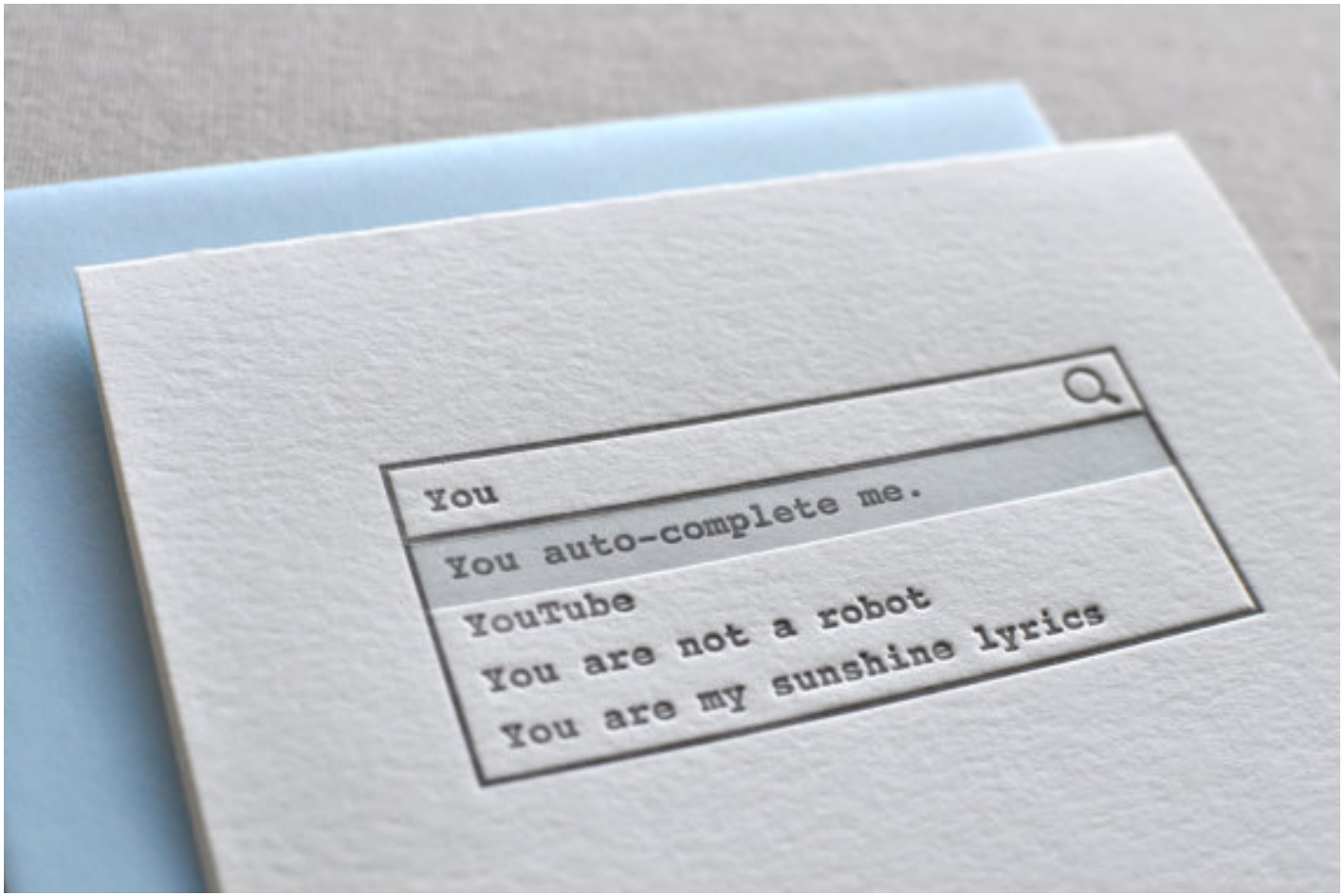

Held-out Data
Sony Ericsson's Infinite
For Search, Murdoch Looks
Price War Brews Between Amazon and Wal-Mart

Evaluation



Measures predictive power (likelihood)

But we don't use topic models for prediction!



Qualitative Evaluation of the Latent Space

"segment 1"	"segment 2"	"matrix 1"	"matrix 2"	"line 1"	"line 2"	"power 1"	power 2"
imag	speaker	robust	manufactur	constraint	alpha	POWER	load
SEGMENT	speech	MATRIX	cell	LINE	redshift	spectrum	memori
texture	recogni	eigenvalu	part	match	LINE	omega	vlsi
color	signal	uncertainti	MATRIX	locat	galaxi	mpc	POWER
tissue	train	plane	cellular	imag	quasar	hsup	systolic
brain	hmm	linear	famili	geometr	absorp	larg	input
slice	source	condition	design	impos	high	redshift	complex
cluster	speakerind.	perturb	machinepart	segment	ssup	galaxi	arrai
mri	SEGMENT	root	format	fundament	densiti	standard	present
volume	sound	suffici	group	recogn	veloc	model	implement

Figure 3: Eight selected factors from a 128 factor decomposition. The displayed word stems are the 10 most probable words in the class-conditional distribution $P(w|z)$, from top to bottom in descending order.

?

Qualitative Evaluation of the Latent Space

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

?

Qualitative Evaluation of the Latent Space

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

?

Qualitative Evaluation of the Latent Space

(a) Topic labeled as SSL

Keyword	Probability
ssl	0.373722
expr	0.042501
init	0.033207
engine	0.026447
var	0.022222
ctx	0.023067
ptemp	0.017153
mctx	0.013773
lookup	0.012083
modssl	0.011238
ca	0.009548

(b) Topic labeled as Logging

Keyword	Probability
log	0.141733
request	.036017
mod	0.0311
config	0.029871
name	0.023725
headers	0.021266
autoindex	0.020037
format	0.017578
cmd	0.01512
header	0.013891
add	0.012661

Table 2: Sample Topics extracted from Apache source code

?

Word Intrusion

- Take the highest probability words from a topic

Original Topic

dog
cat
horse
pig
cow

Word Intrusion

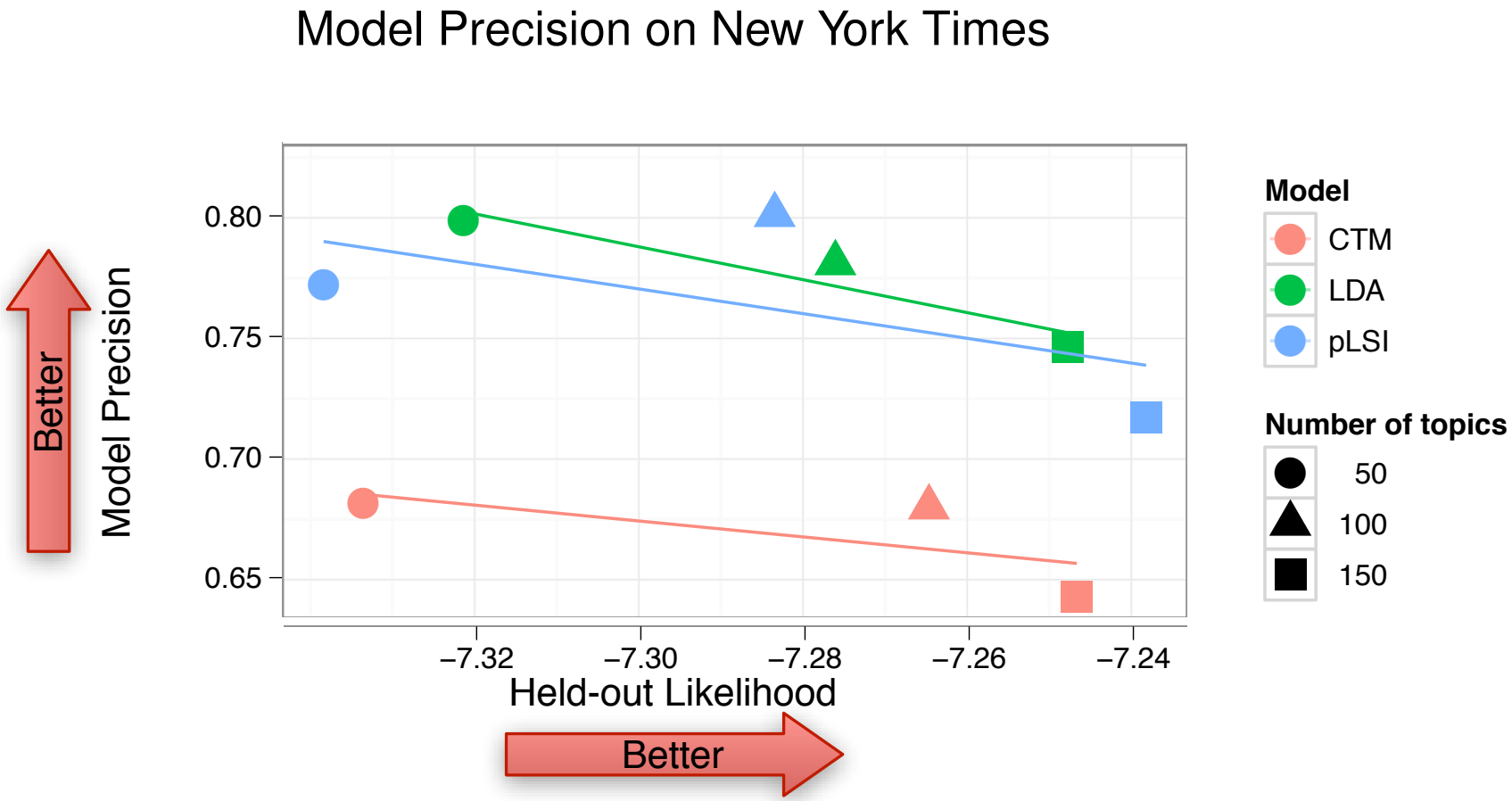
- Take the highest probability words from a topic

Original Topic

dog
cat
apple
horse
pig
cow

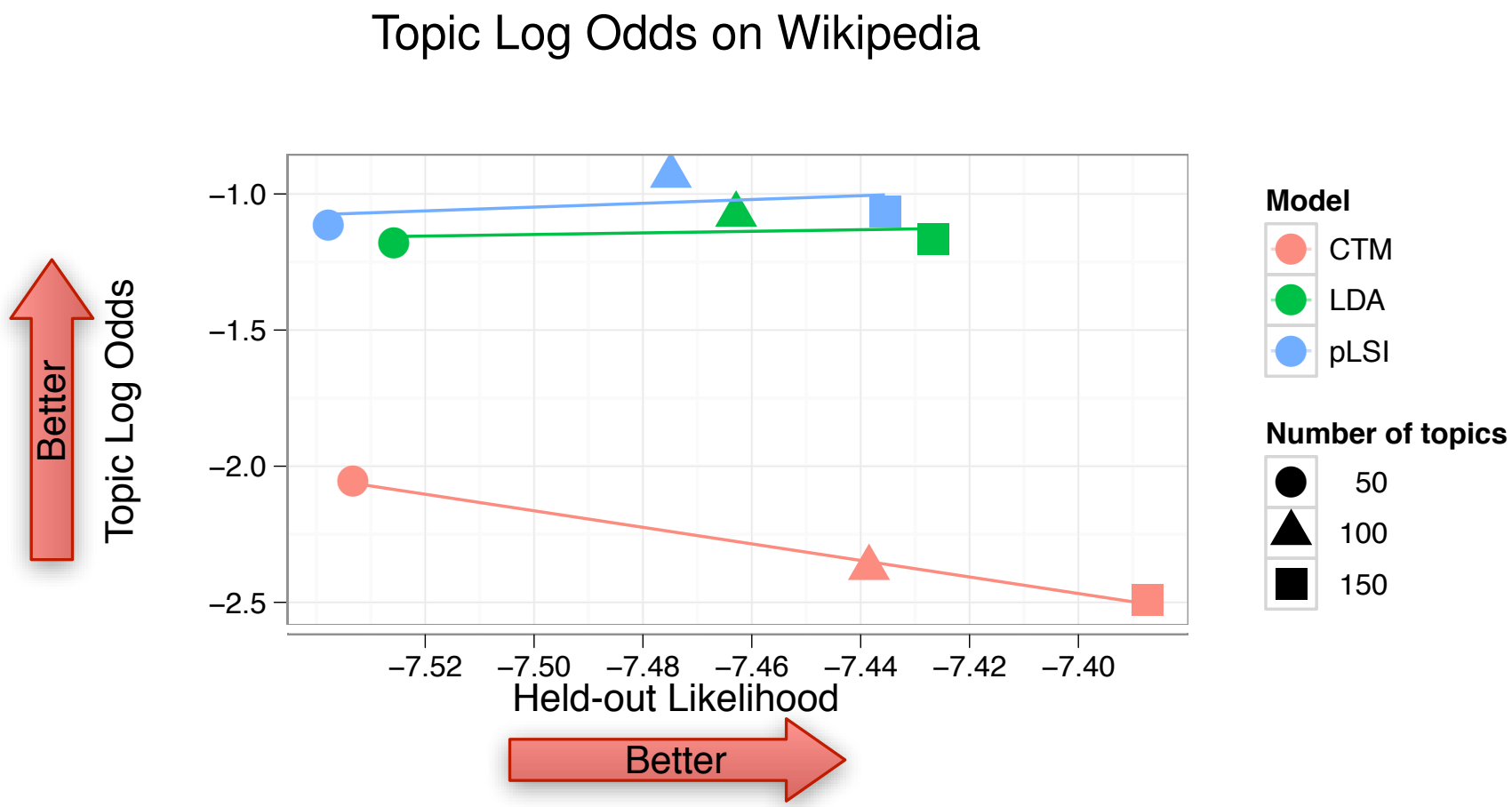
- Intruder: high probability word from another topic

Interpretability and Likelihood



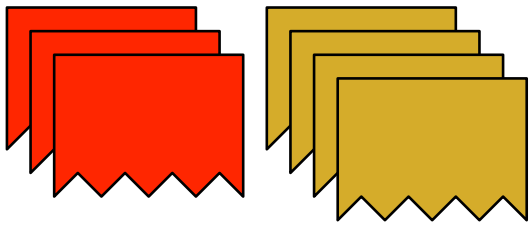
within a model, higher likelihood $\neq$ higher interpretability

Interpretability and Likelihood

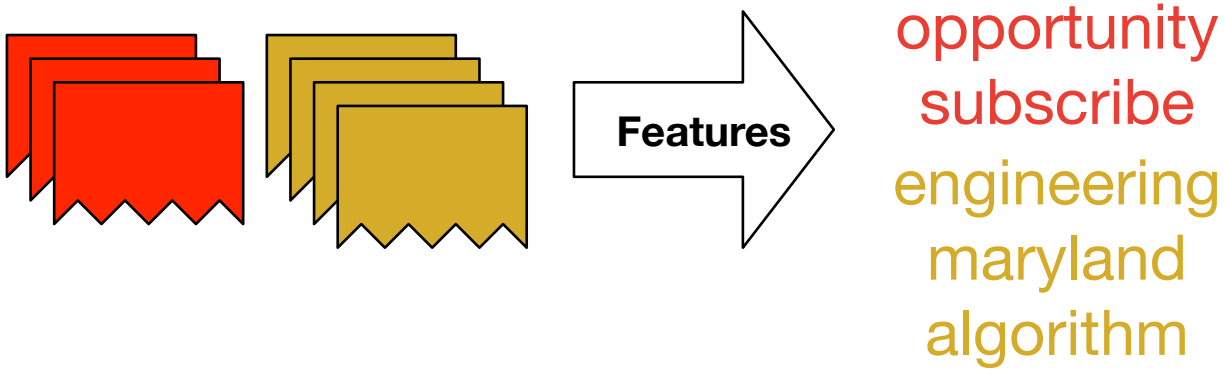


across models, higher likelihood ≠ higher interpretability

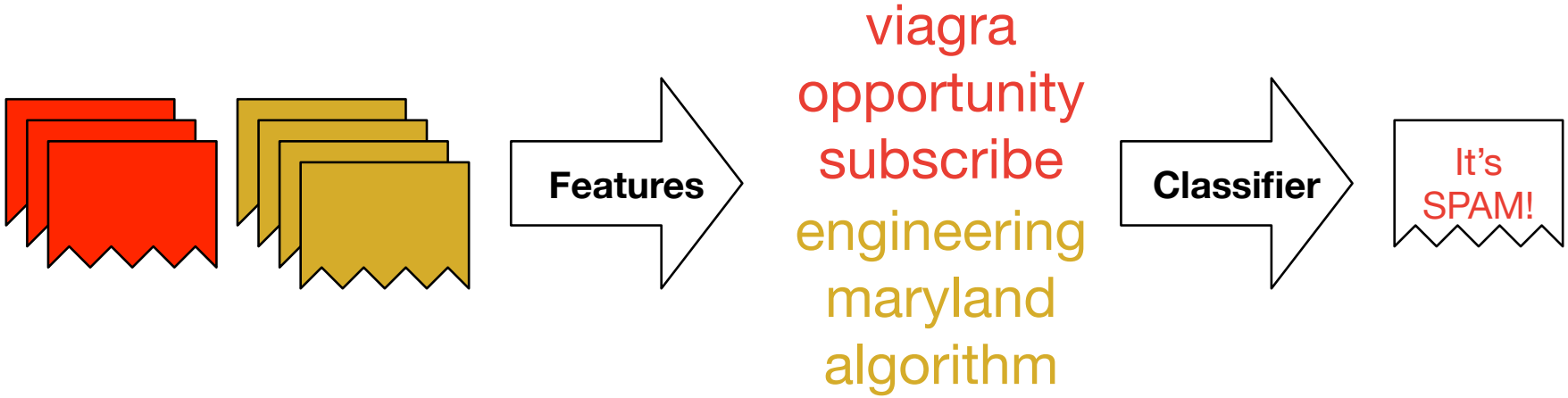
What about Supervised Models?



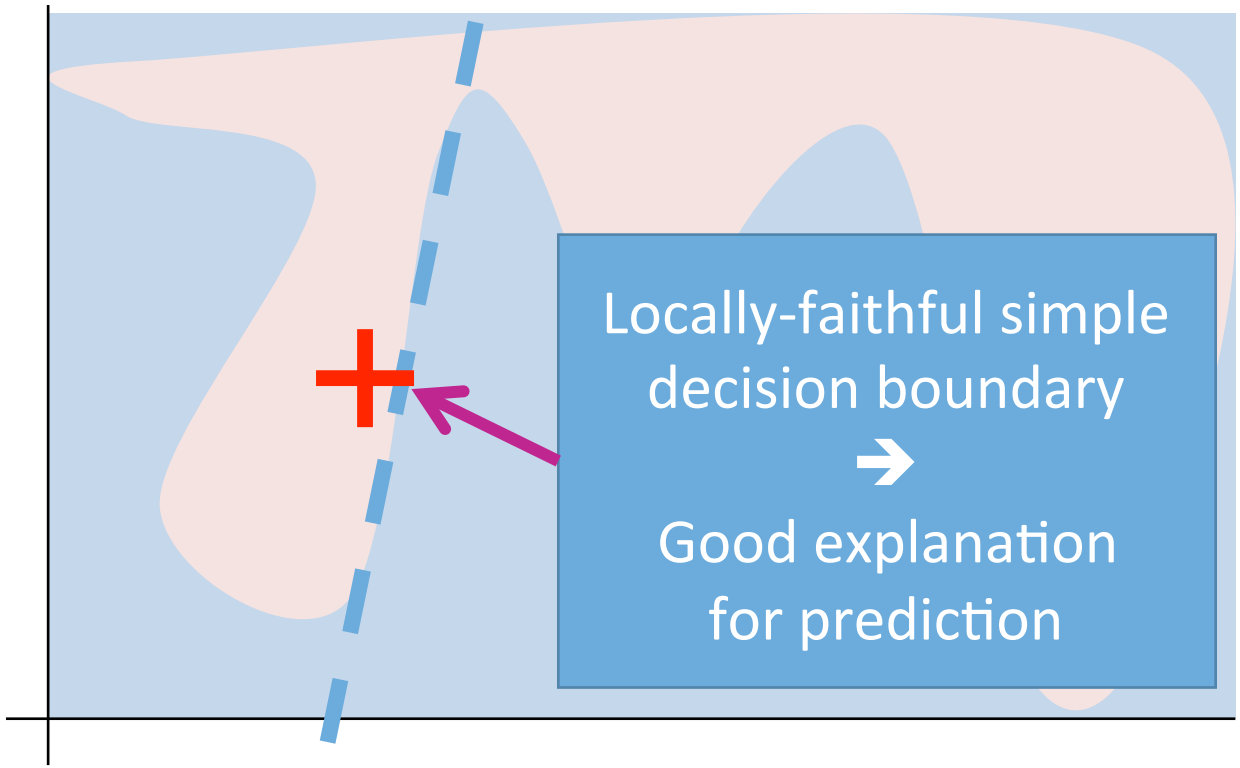
What about Supervised Models?



What about Supervised Models?



LIME



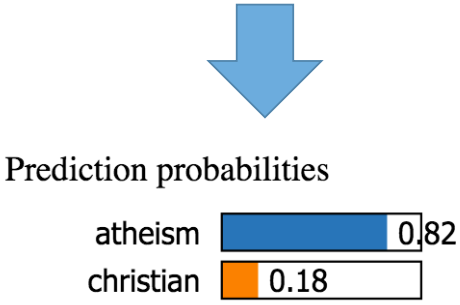
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016.

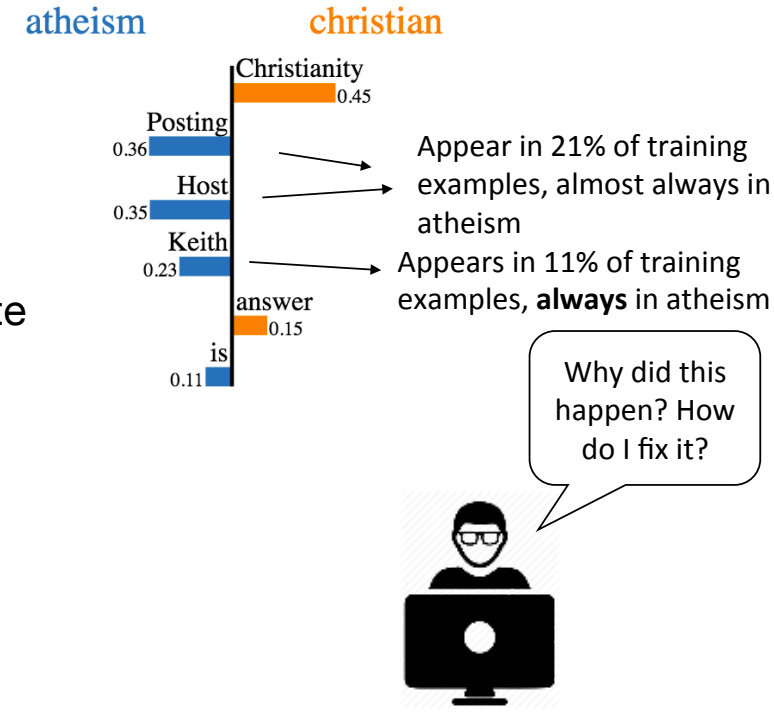
LIME: Local Interpretable Model-Agnostic Explanations

What's an Explanation

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note





What's an Explanation

P( ) = 0.32

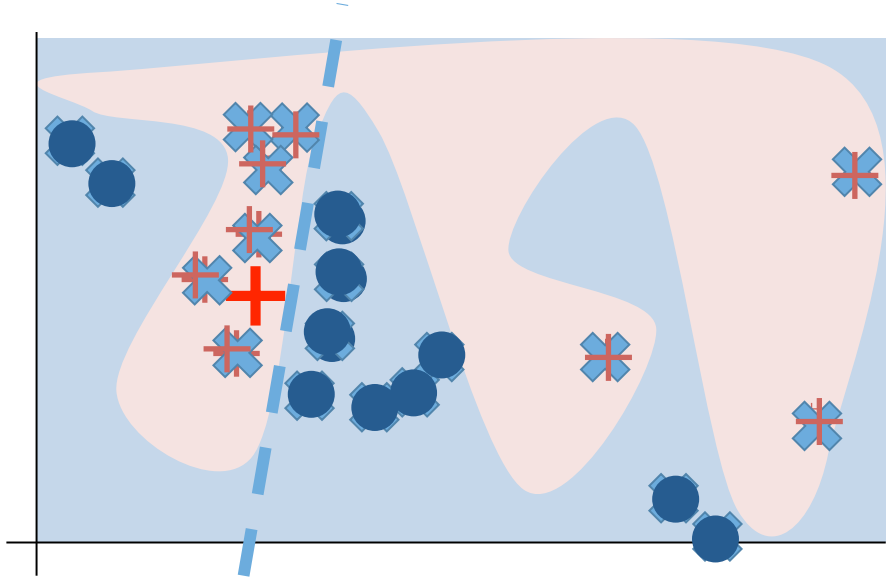
P( ) = 0.24

P( ) = 0.21

What makes good Explanation?

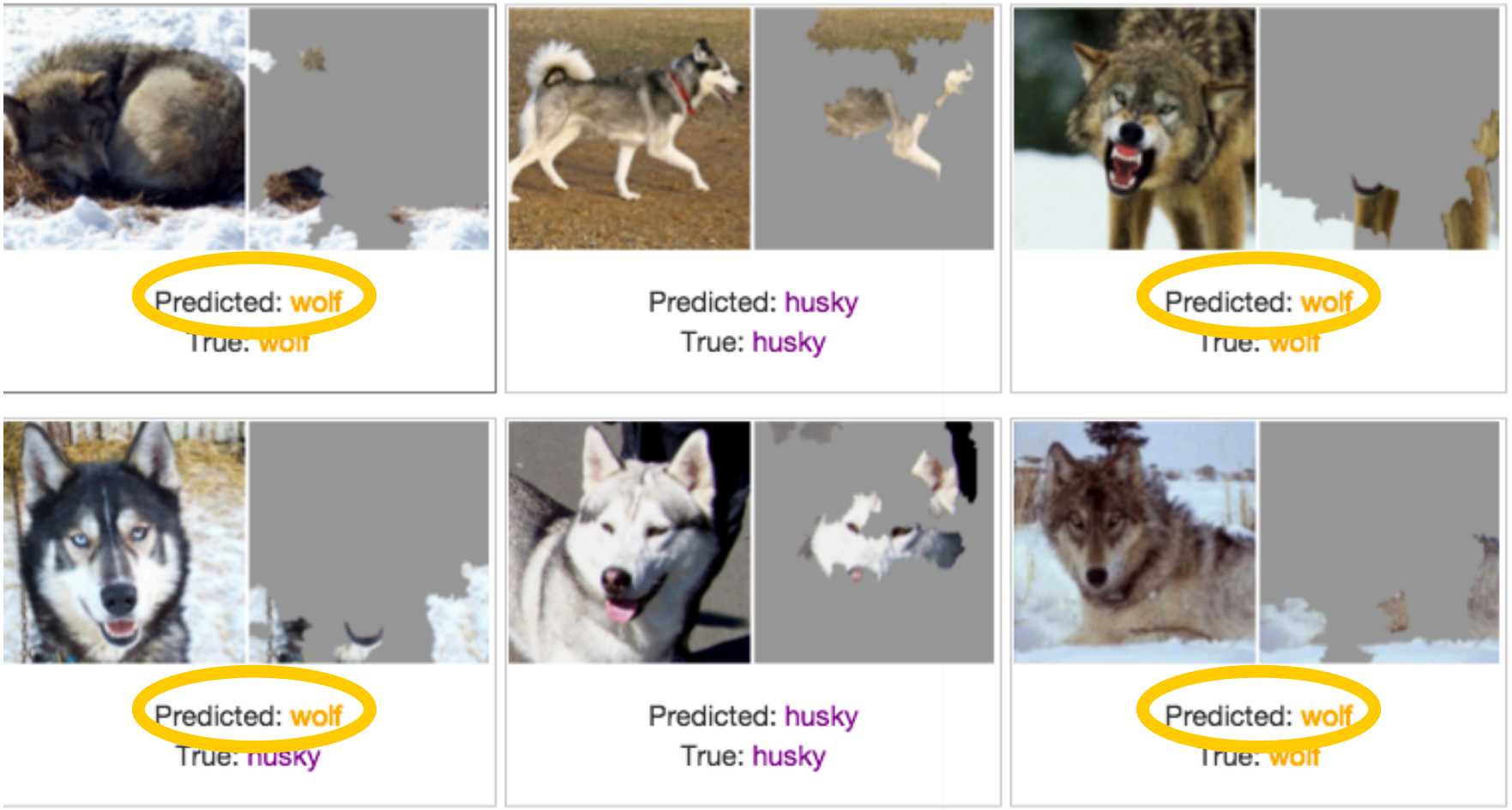
- Interpretable: Humans can Understand
- Faithful: Describes Model
- Model Agnostic: Generalize to Many Models

Method

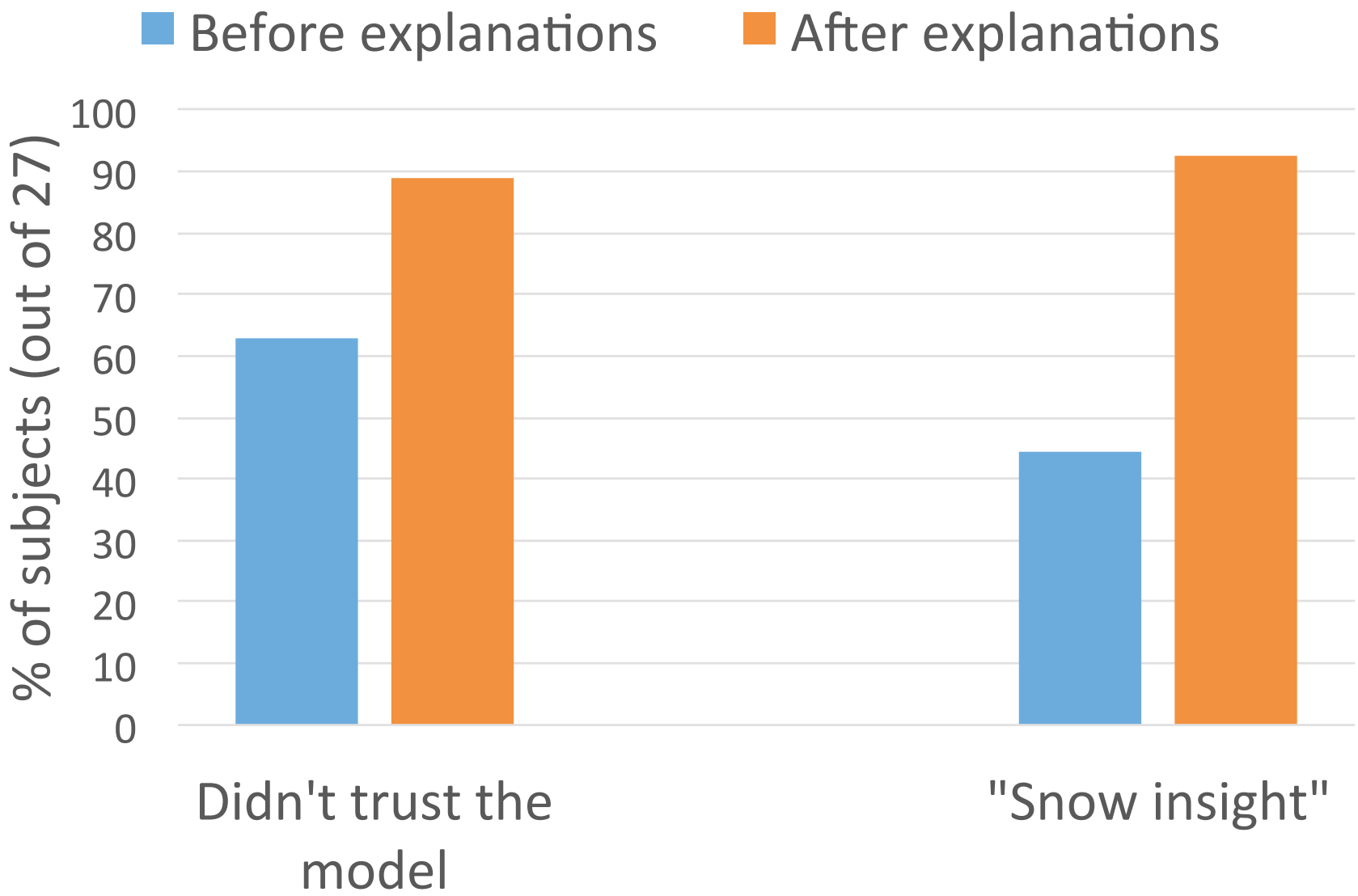


- Complicated model predicts “near” example
- Simple model explains **local variation**
- **Explains what complicated model focused on**

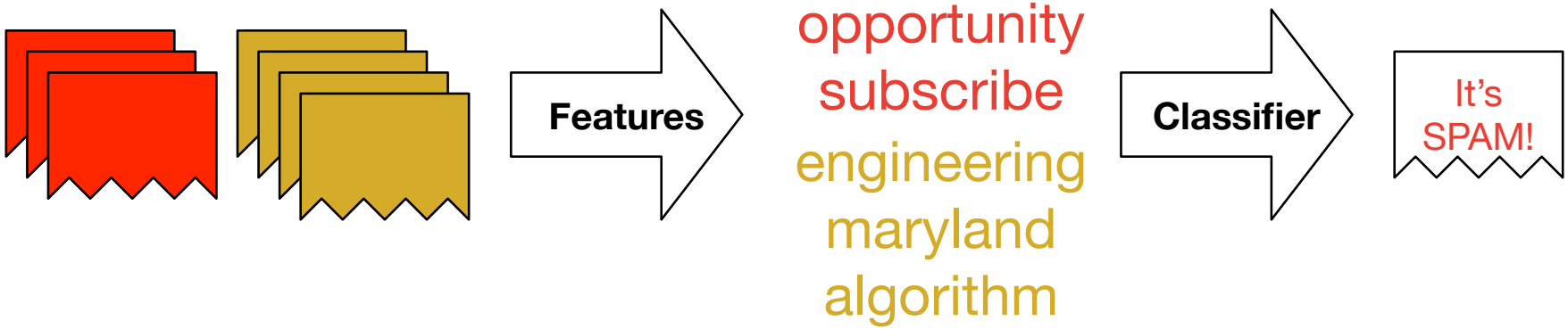
Is this a good Classifier?



Is this a good Classifier?

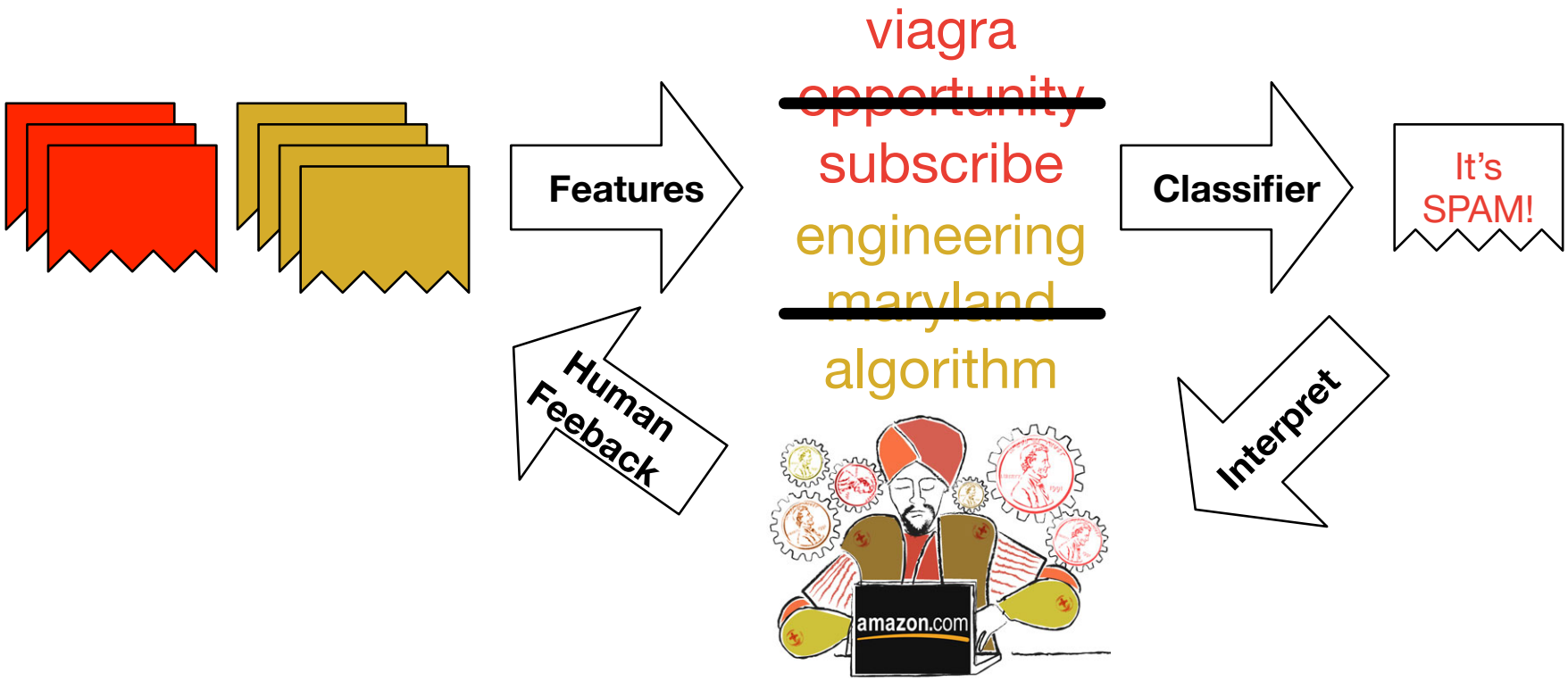


Improving ML Algorithms

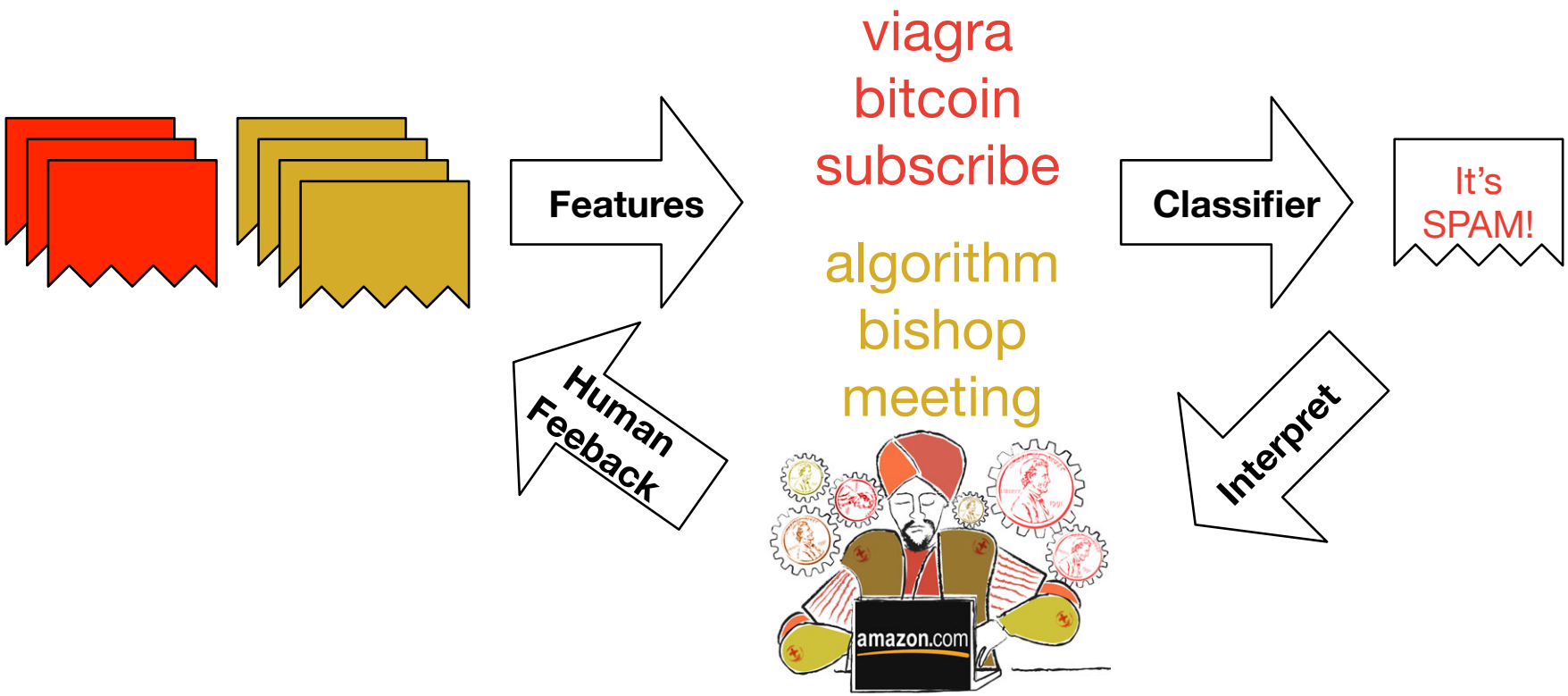


Improving ML Algorithms

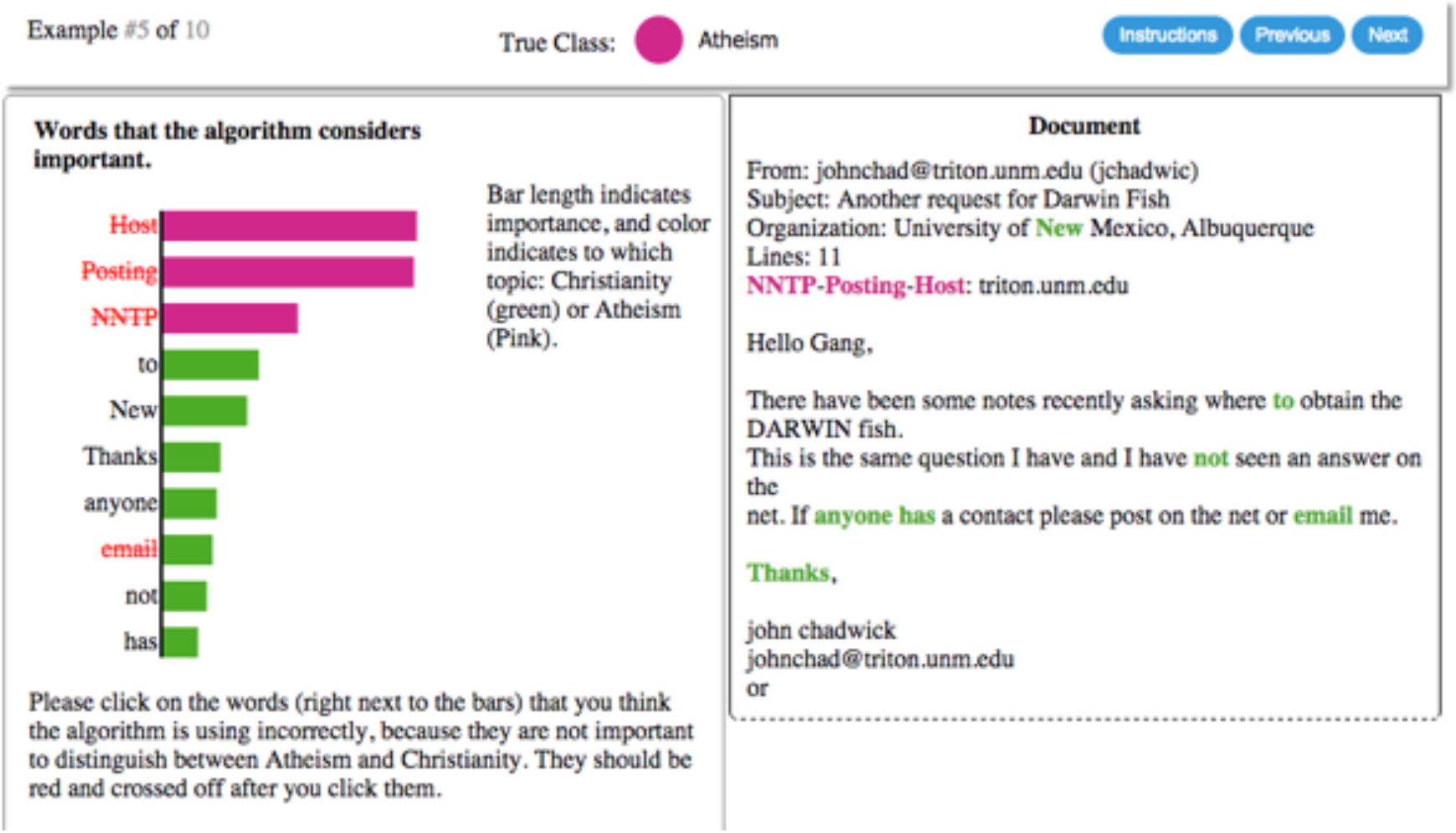
Improving ML Algorithms



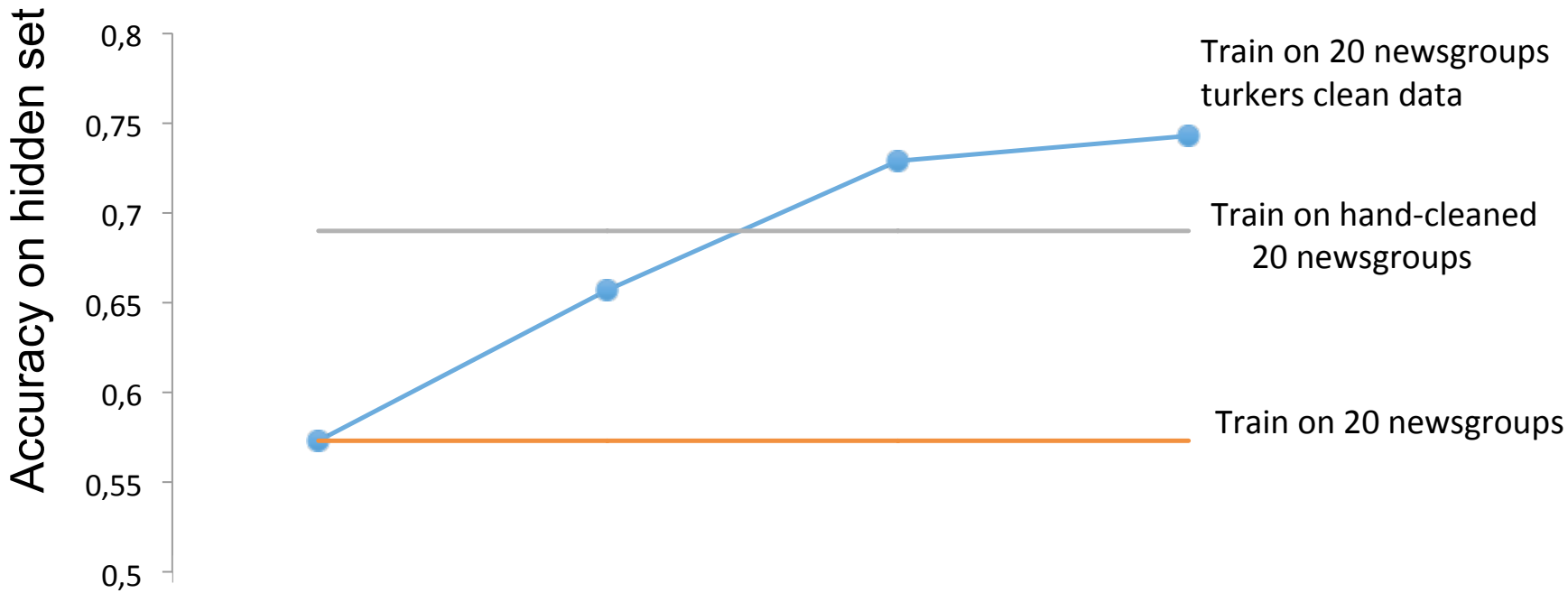
Improving ML Algorithms

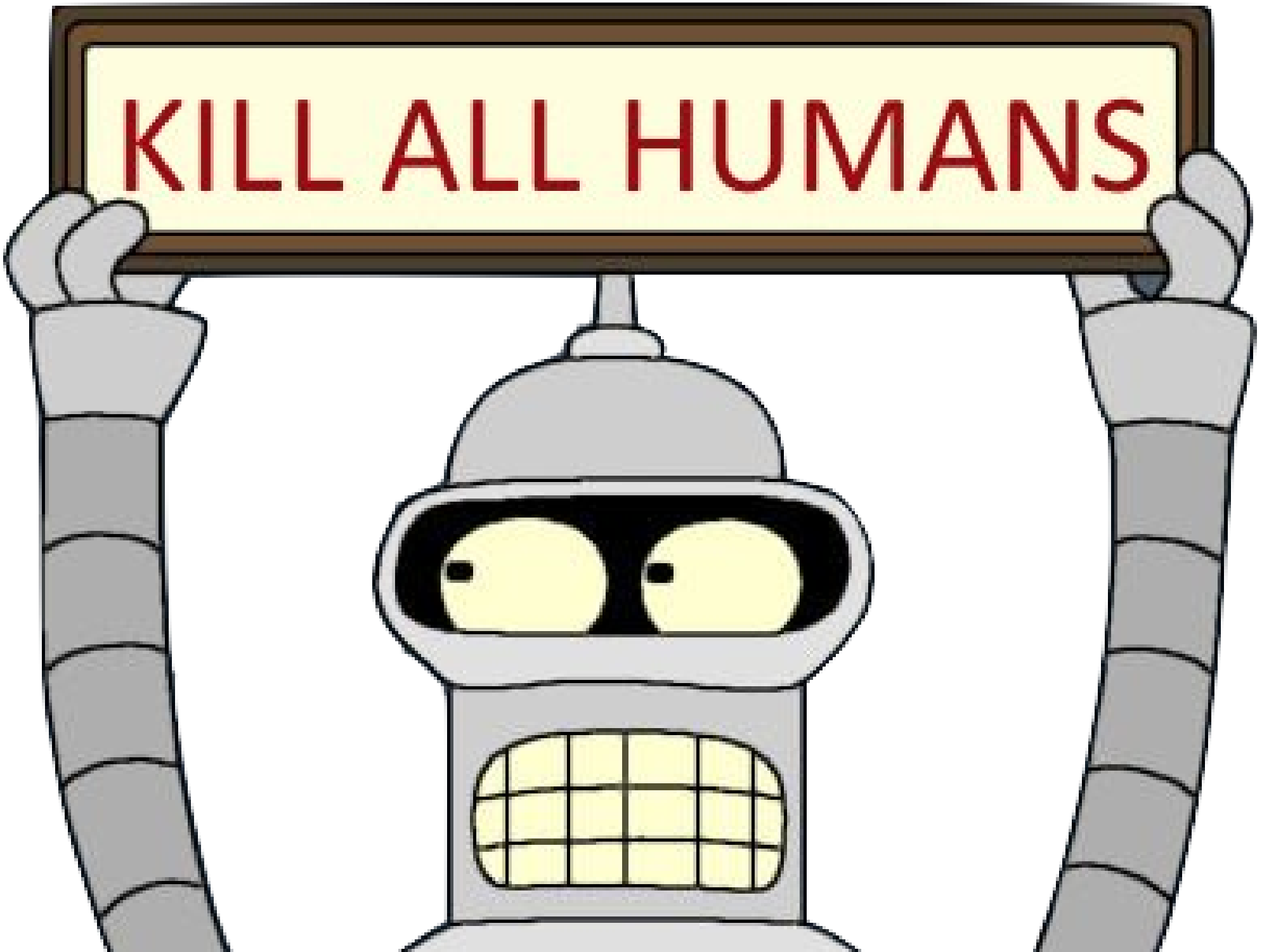


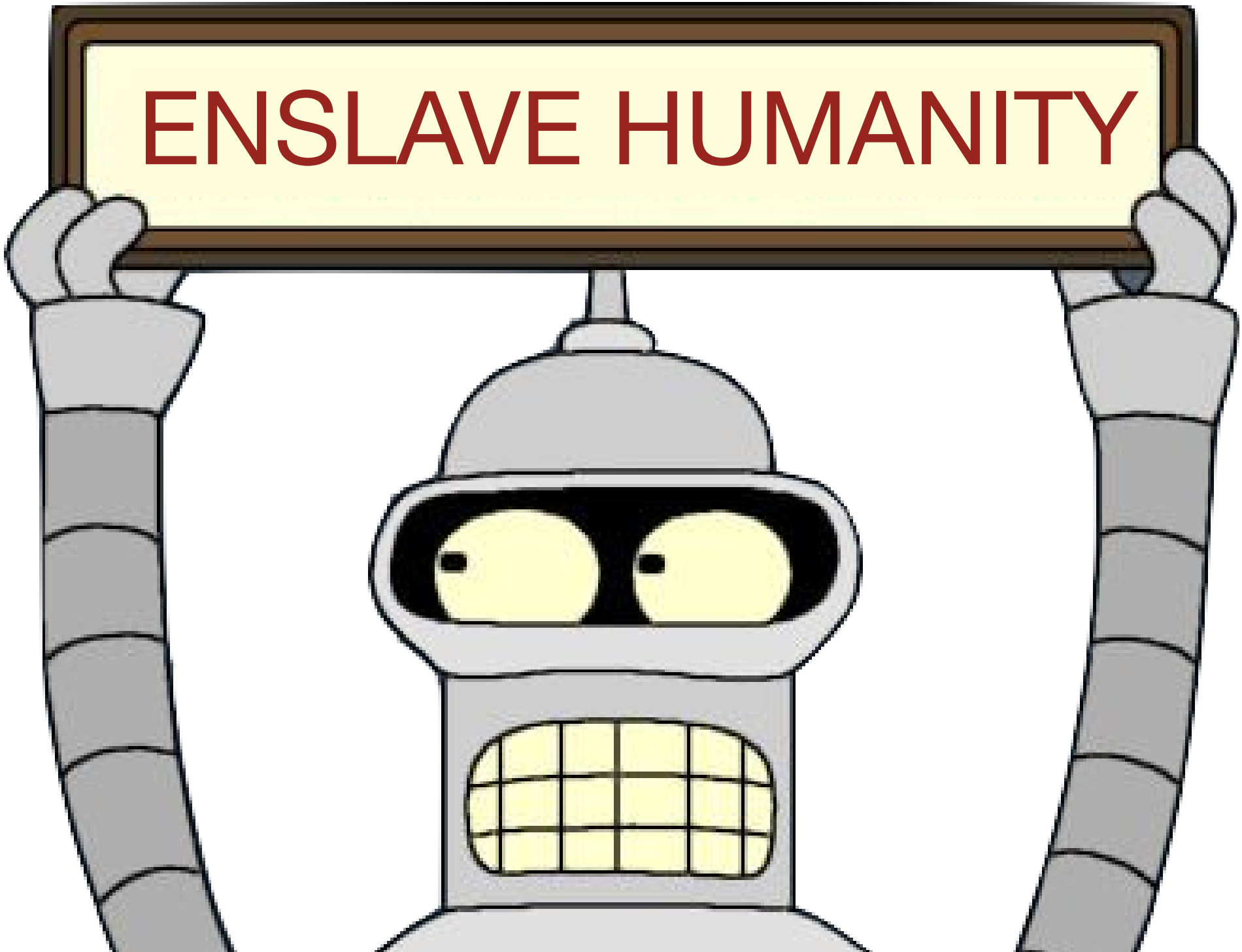
Improving ML Algorithms

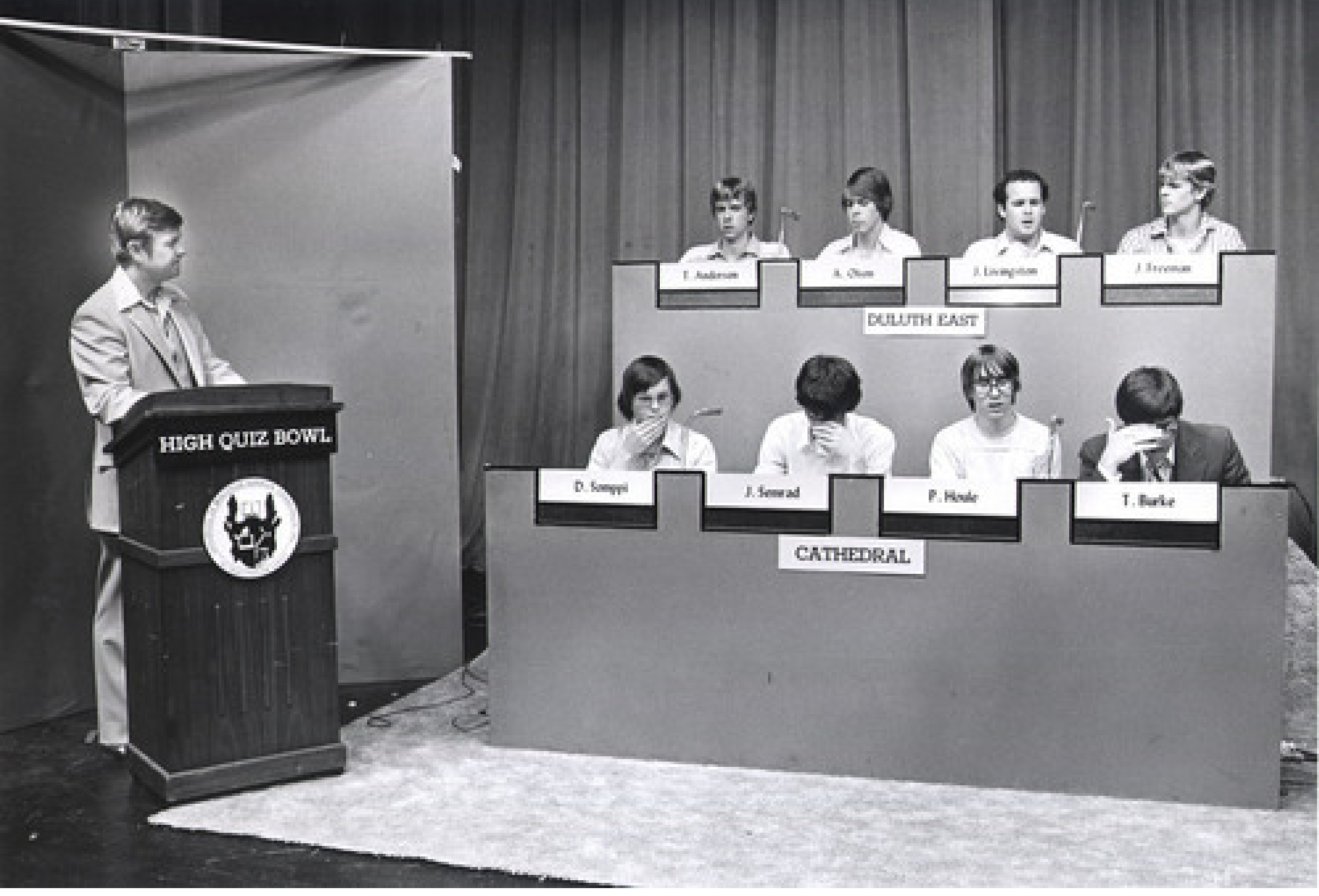


Improving ML Algorithms



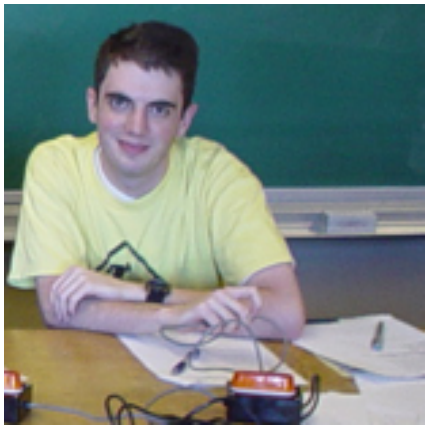

KILL ALL HUMANS


ENSLAVE HUMANITY

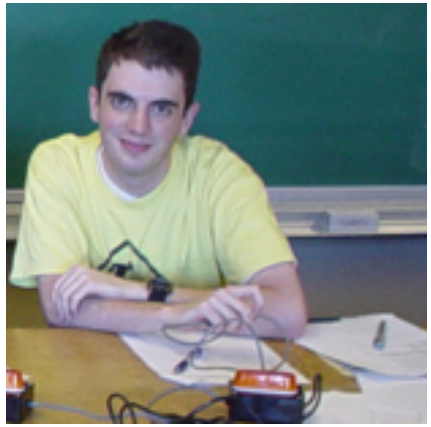

HIGH QUIZ BOWL
DULUTH EAST
CATHEDRAL
T. Anderson
P. Hinde
T. Burke

Centaur Chess

Measuring Interpretability



Measuring Interpretability



Solo

6

Measuring Interpretability

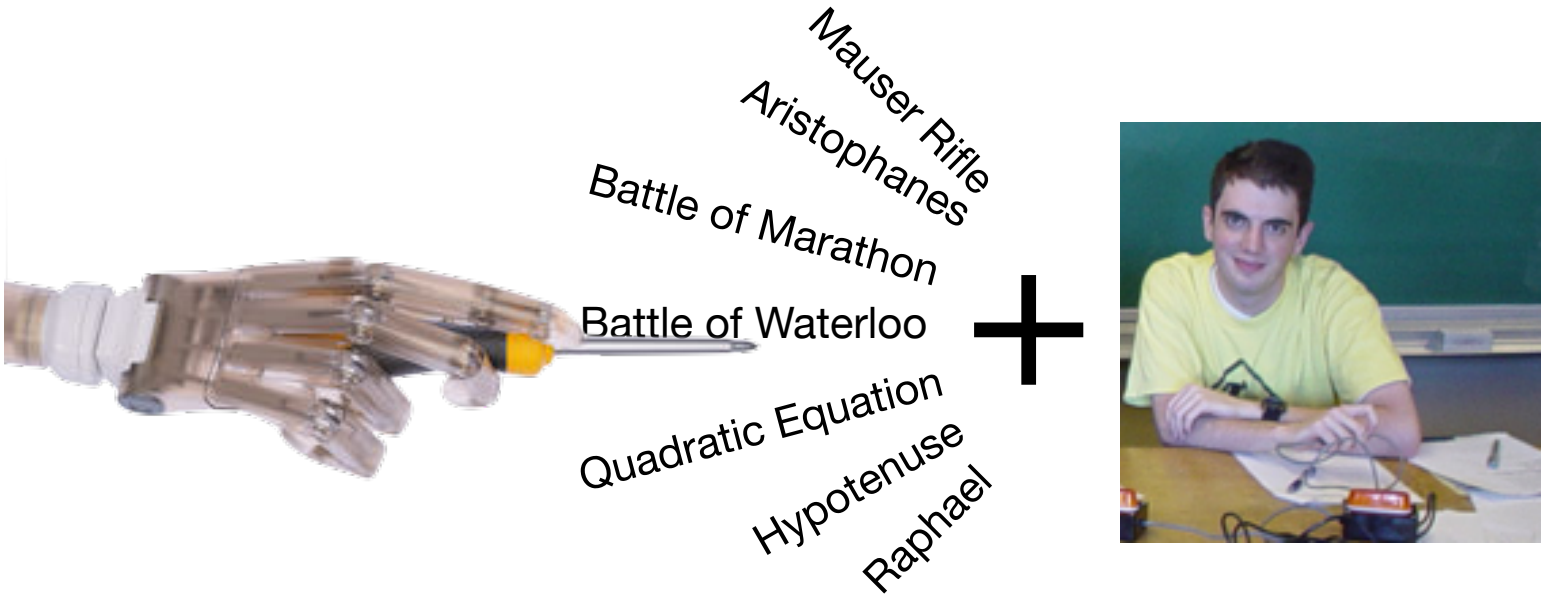
Mauser Rifle
Aristophanes
Battle of Marathon
Battle of Waterloo
Quadratic Equation
Hypotenuse
Raphael

+

Solo

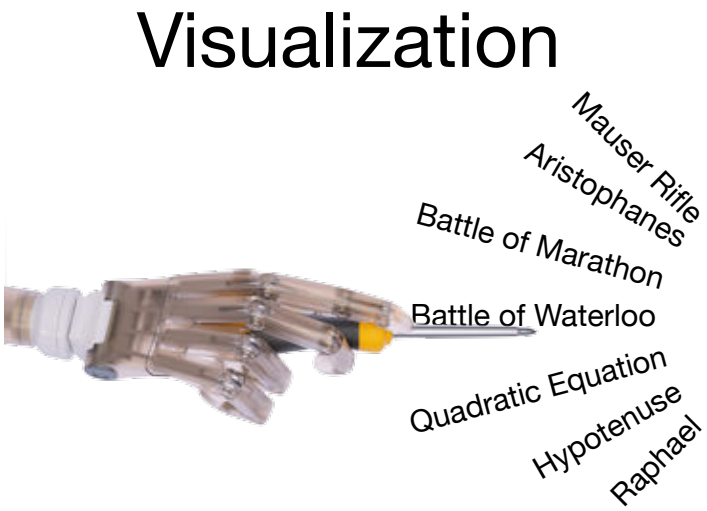
6

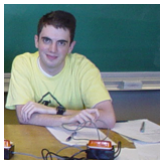
Measuring Interpretability



Viz A	Viz B	Viz C	Solo
5.5	7.3	6.1	6

Improvement through Reinforcement Learning





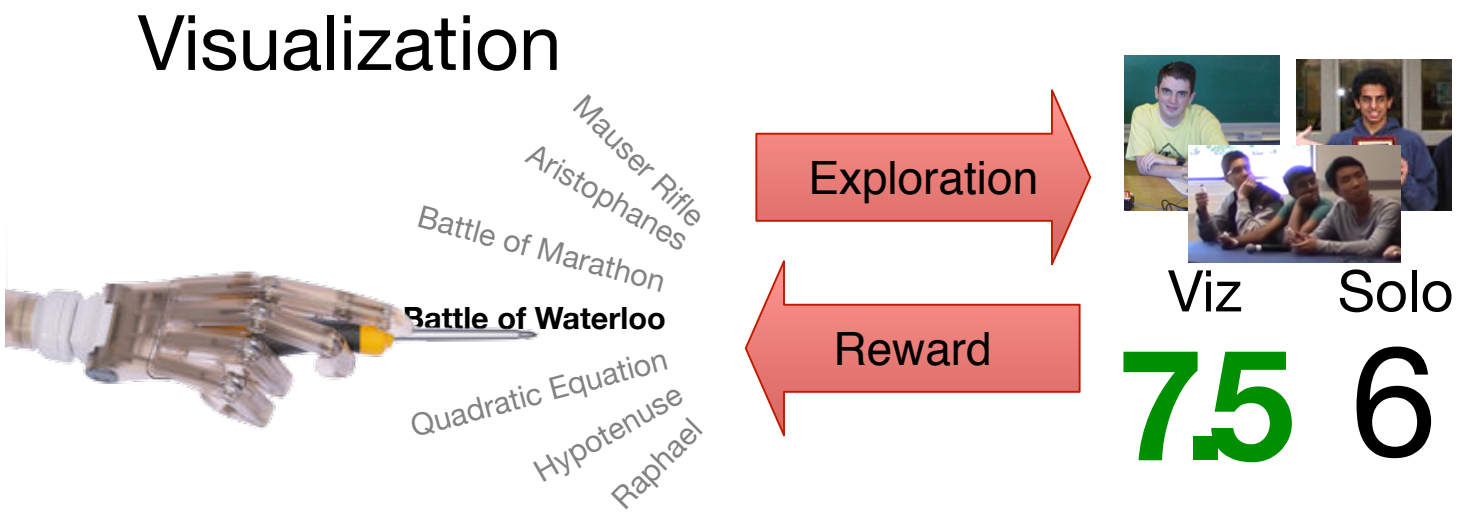
Viz	Solo
7.3	6

Improvement through Reinforcement Learning

Improvement through Reinforcement Learning

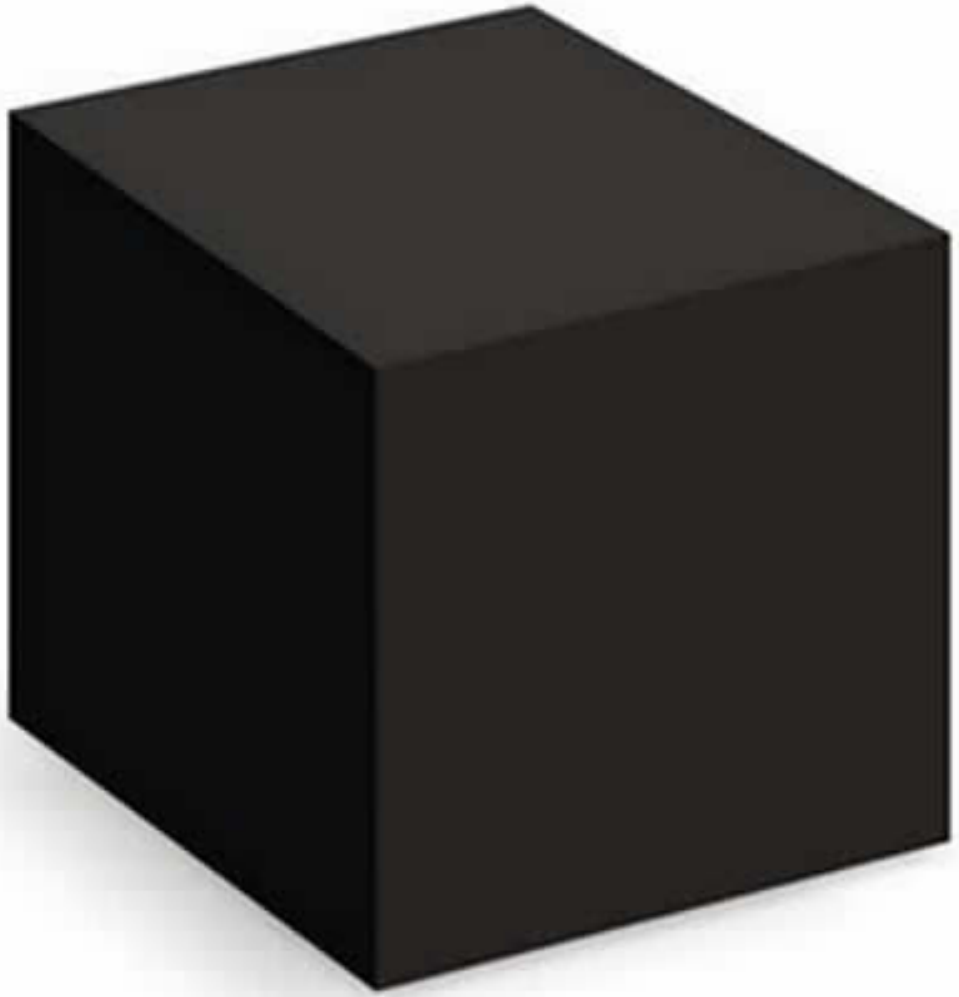
Improvement through Reinforcement Learning

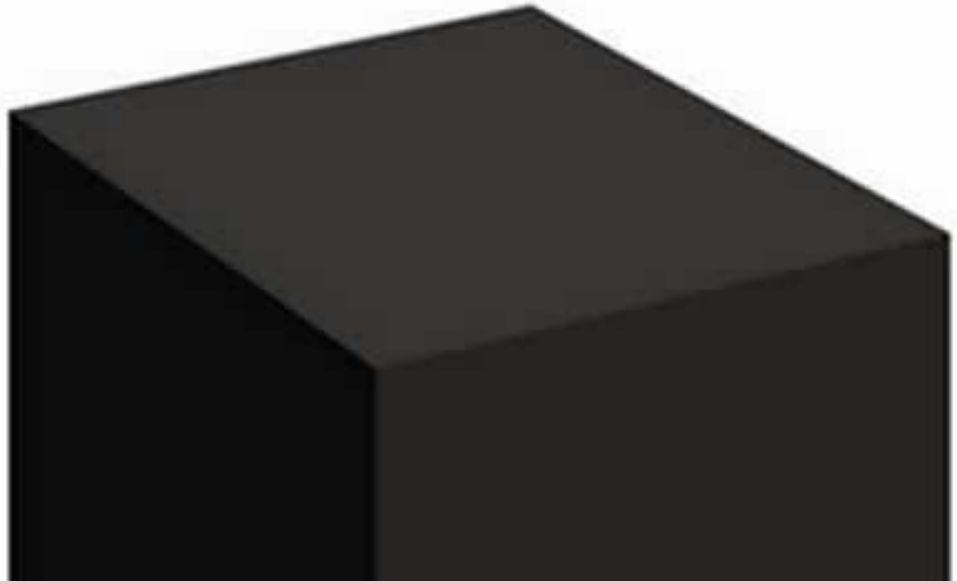
Improvement through Reinforcement Learning



Simultaneous Interpretation is Hard!

- Exhausting for humans
- Computers not trusted
- Differential strengths
- Same word-by-word characteristic





Takeaways

- ML should be interpretable
- We should measure interpretability
- Interpretability should reflect the world we want