Embedded Machine Learning for Edge Computing

AI in the Real World

Mohamed Afham

Mohamed Afham

- Graduate Student @ Technical University of Darmstadt
- Former AI Resident @ Meta AI
- Former Research Intern @ MBZUAI
- BSc @ University of Moratuwa, Sri Lanka

Acknowledgement

- Content borrowed from Chip Huyen’s slides for Stanford course CS329S: https://stanford-cs329s.github.io/syllabus.html

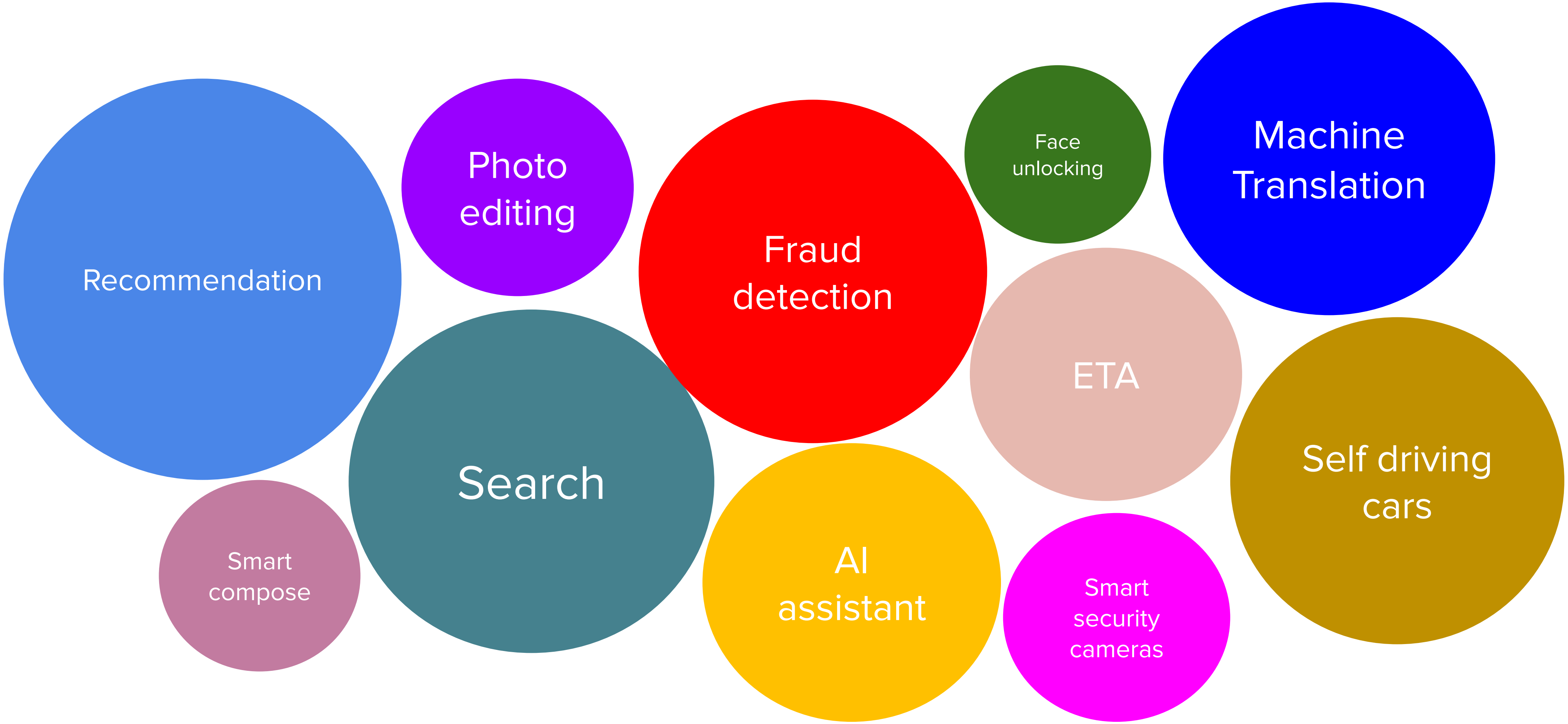
- Content borrowed from Been Kim’s slides for Vector Institute course: https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf

Agenda

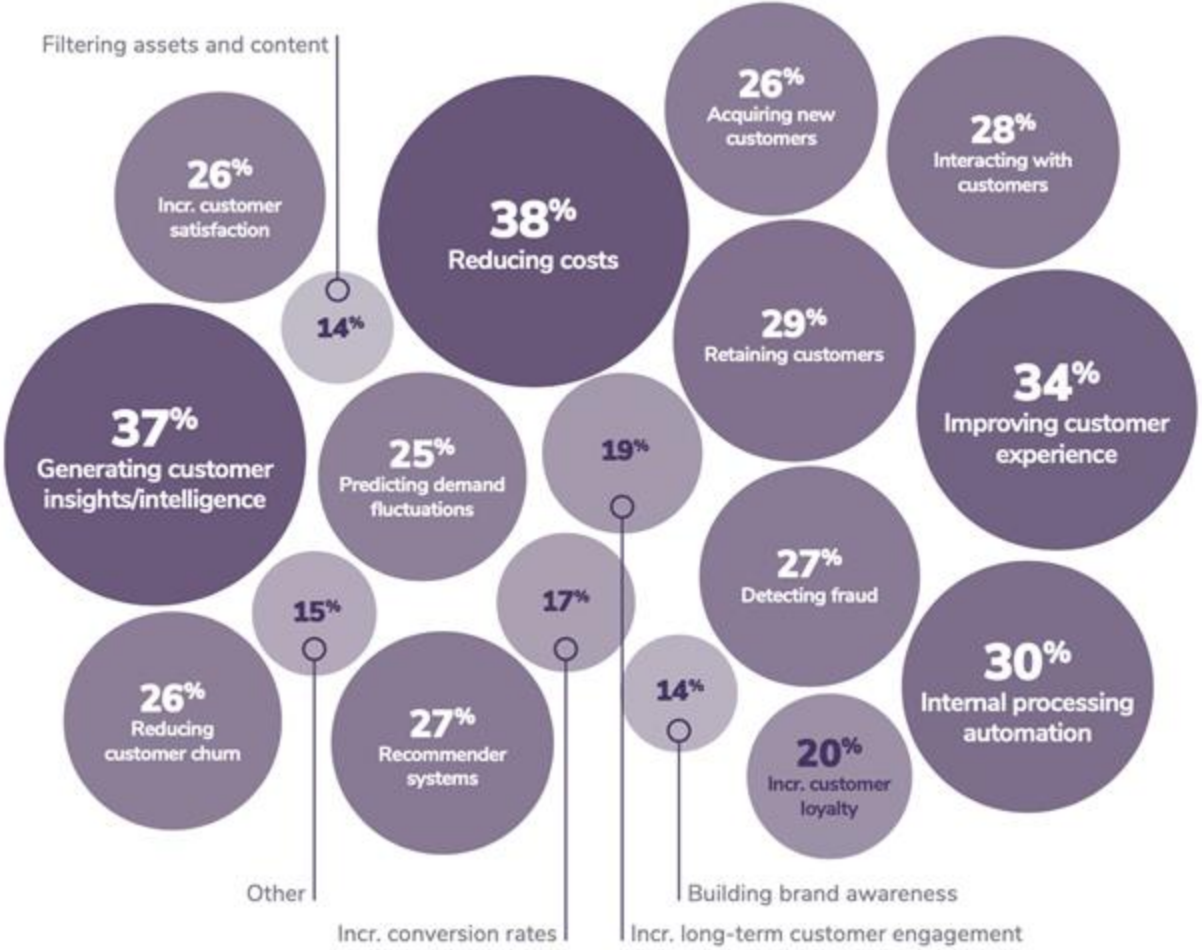
- Moving into Production (overview of ML systems)
- Model Selection & Training
- Model Evaluation
- Ethics & Bias

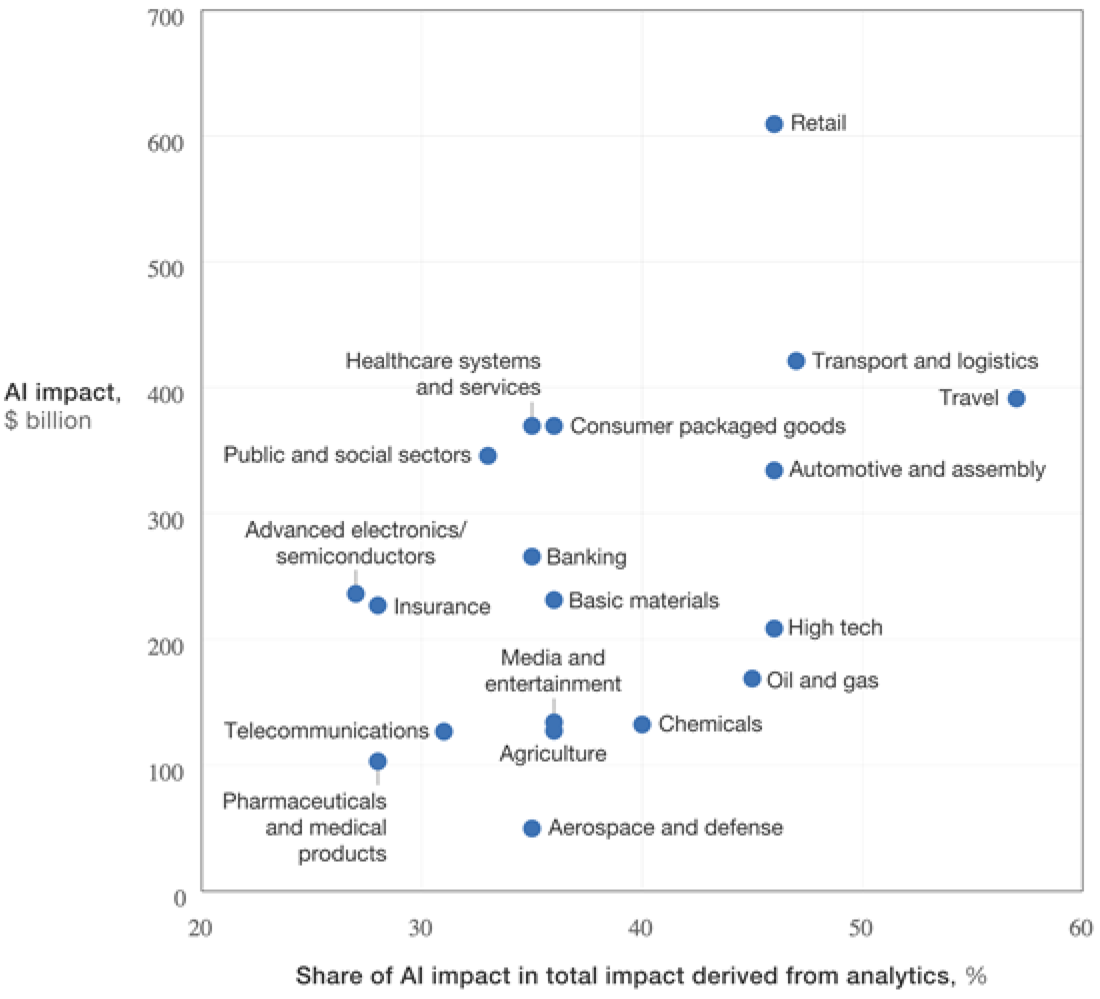
Is ML really impactful?

2024: ML is in almost every aspect of our lives



Machine learning use case frequency





McKinsey&Company | **Source:** McKinsey Global Institute analysis

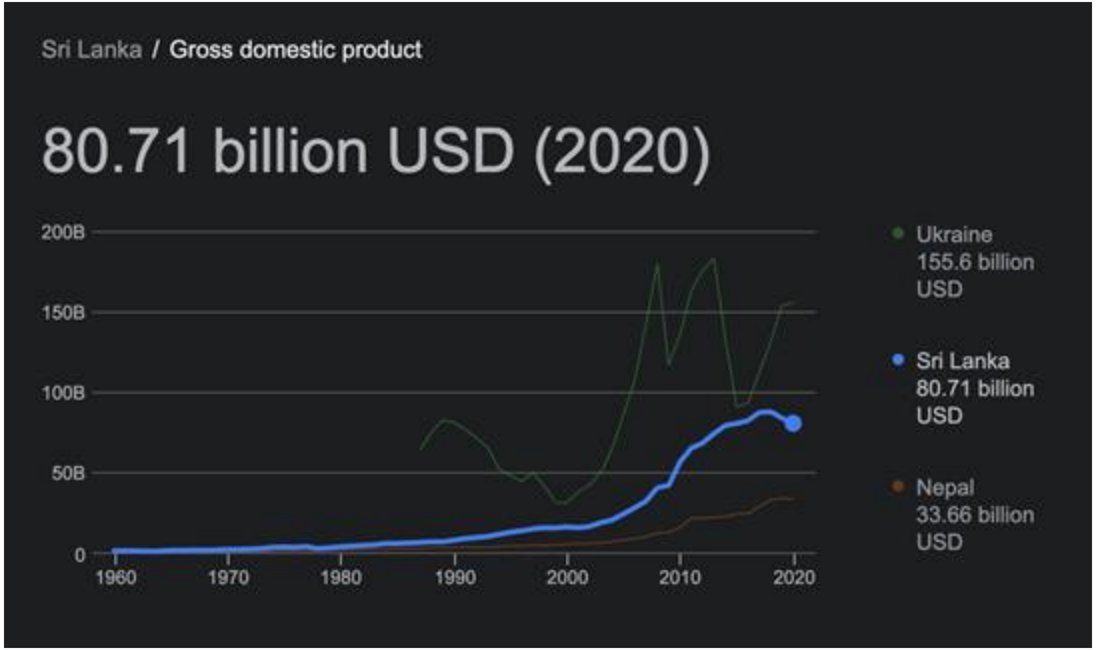
AI value creation by 2030

13 trillion USD

Most of it will be outside the consumer internet industry

We need more people from non-CS backgrounds in AI!

Sri Lanka GDP for comparison… (pre-crisis)



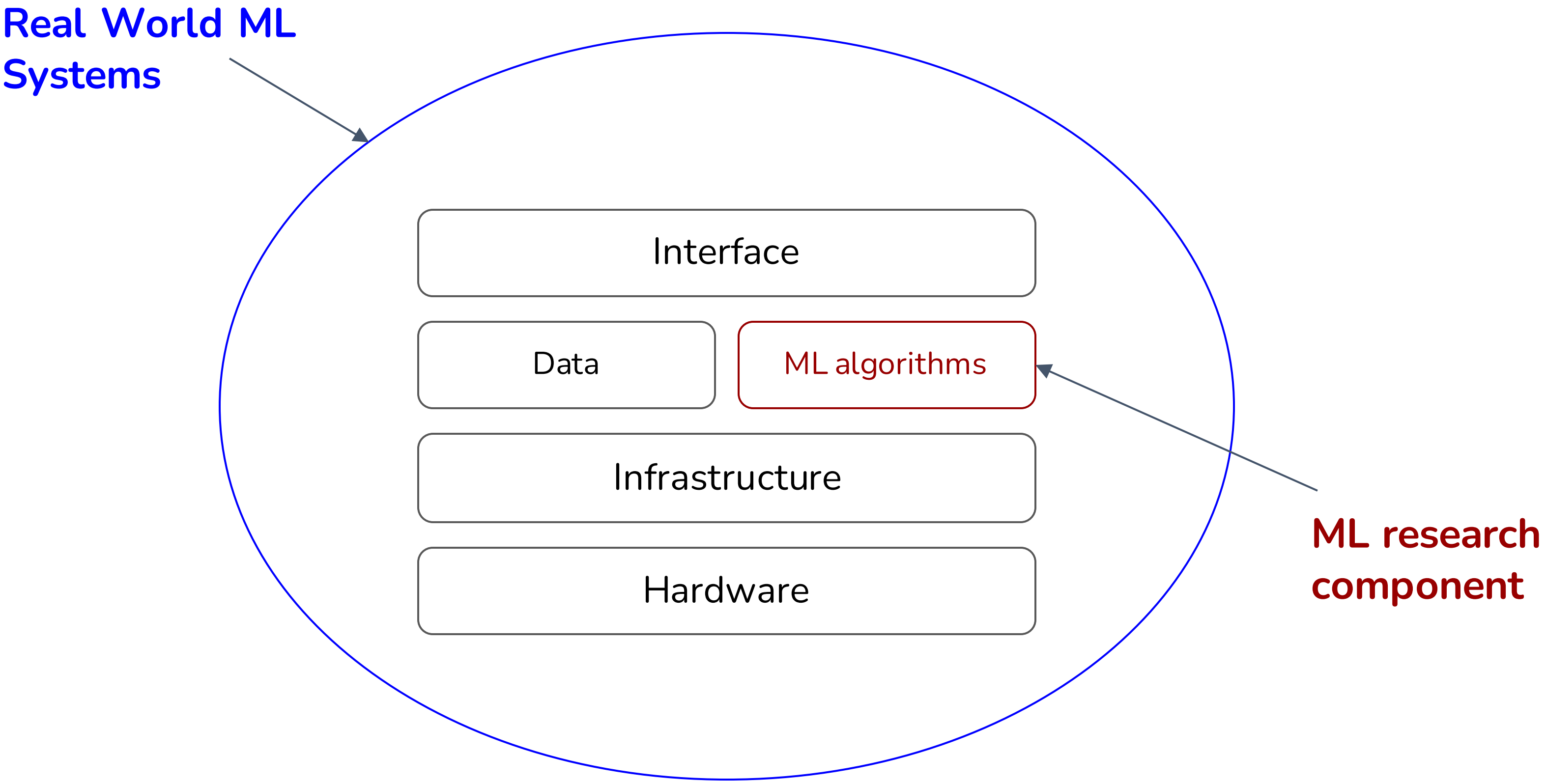
AI value creation by 2030

13 trillion USD

Most of it will be outside the consumer internet industry

We need more people from non-CS backgrounds in AI!

Moving to Real World Production

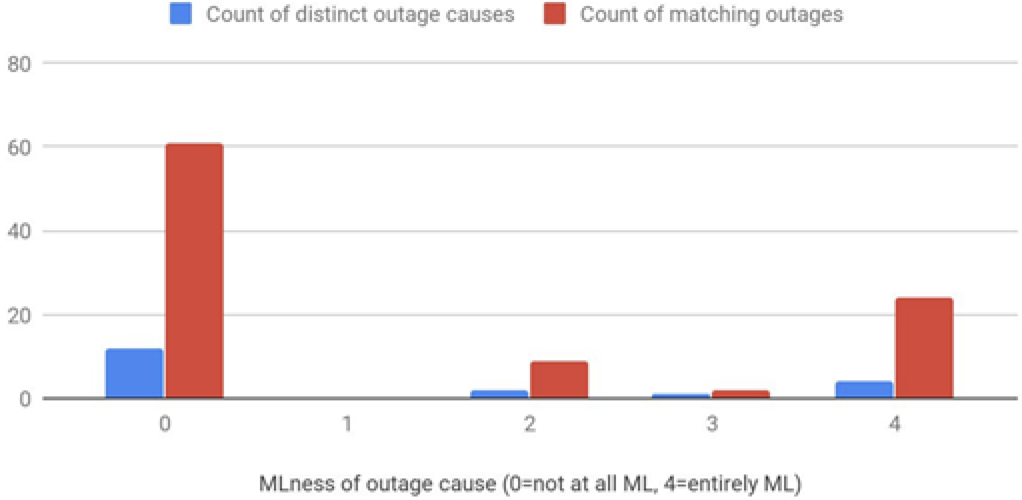


What's hard about production?

- ML algorithms are the less problematic part.
- The hard part is: **how to make algorithms work with other parts to solve real-world problems**.

What's hard about production?

- ML algorithms are the less problematic part.
- The hard part is: **how to make algorithms work with other parts to solve real-world problems**.
- [60/96 failures]() caused by non-ML components



How ML Breaks: A Decade of Outages for One Large ML Pipeline - D Papasian & T Underwood, Google

Reliable management of continuous or periodic machine learning pipelines at large scale presents significant operational challenges. Using experience from almost *15 years of operating some of the largest ML pipelines*, we examine the characteristics of one of the *largest and oldest continuous pipeline at Google*. We look at actual outages experienced and try to understand what caused them.

ML research vs. ML production

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives

Stakeholder objectives

ML team
highest accuracy

Sales
sells more ads

Product
fastest inference

Manager
maximizes profit
= laying off ML teams

Computational priority

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency

generating predictions

ML in research vs. in production

	Research	**Production**
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting

Beginner’s Starting Point: Transfer Learning!

- Open-source pre-trained models
- Standard training recipes
- Finetune or linear probe

Model Selection & Training

Components of ML model training phase

1. Data Collection (+ pre-processing)

1. Model (algorithm) selection

1. Evaluation metrics

1. Training strategy (hyper-params)

1. Sanity checks / verification

1. Actual training

Data Phase

How much data? Just enough, and then go forward?

Annotation pipelines

- Open-source annotation tools
  (e.g. https://www.robots.ox.ac.uk/~vgg/software/via)
- Annotation standards (format, consistency)

Active learning

- Annotating some data will improve model more

Data Phase

Sometimes, you have your data already and that is all you get.

- Common Issue: Data Leakage

  Some form of the label “leaks” into the features

  This same information is not available when deployed (inference)

Data leakage: example

- Problem: detect lung cancer from CT scans
- Data: collected from hospital A
- Performs well on test data from hospital A
- Performs poorly on test data from hospital B

Patient ID	Date	Doctor note	Medical record	Scanner type	CT scan

At hospital A, when doctors suspect that a patient has lung cancer, they send that patient to a higher-quality scanner

Causes of data leakage

1. Processing before splitting

1. Data duplication

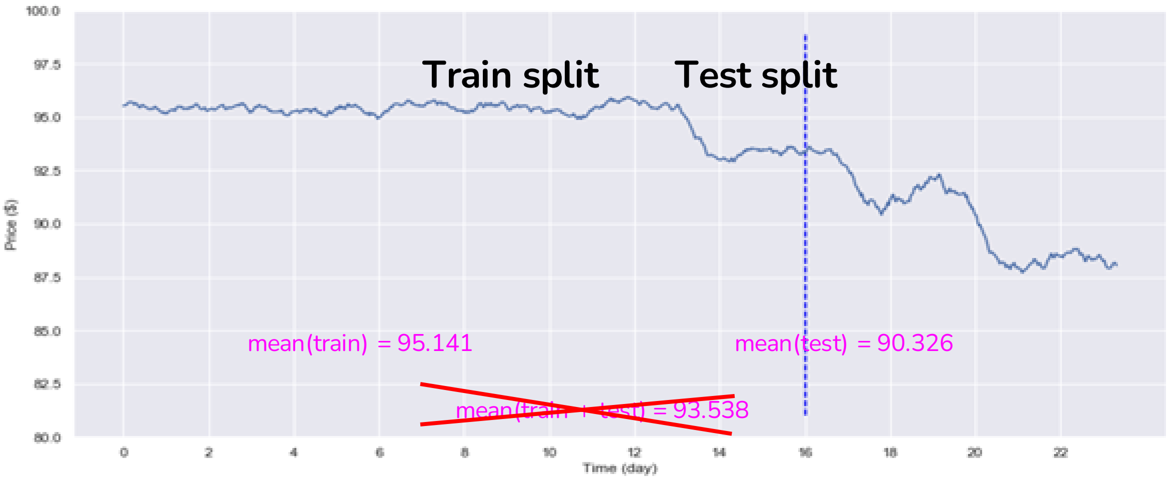
1. Group leakage

1. Leakage from data generation & collection process

1. Processing before splitting

Split your data in train / val / test sets first!

Never calculate any statistics on the entire set for any reason



1. Processing before splitting

How should we split?

- Depends on size of data
- 3 splits - balance suitably

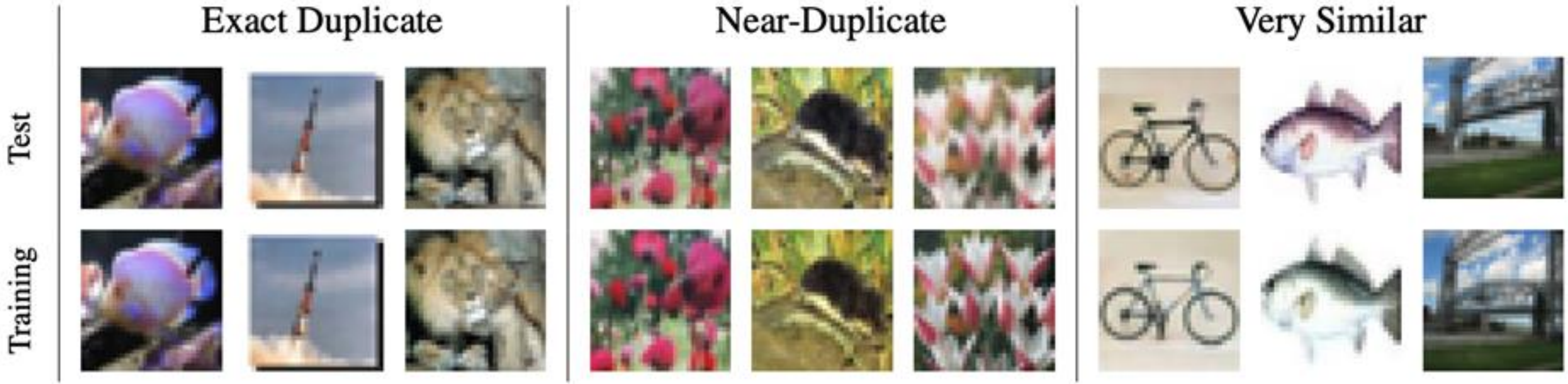
Train - for training model

Validation - for hyper-parameter tuning

Test - evaluate before deploying

2. Data duplication

- Datasets contain duplicates & near-duplicates
  - 3.3% CIFAR-10 & 10% CIFAR-100 test images have dups in training set
  - Removing dups increases errors 17.05% -> 19.38% on CIFAR-100



[Do we train on test data? Purging CIFAR of near-duplicates](#) (Barz & Denzler, 2019)

2. Data duplication

- Datasets come with duplicates & near-duplicates
- Oversampling (data-augmentation) can cause duplications

2. Data duplication

- Test set includes data from the train set
- Solution:
  - Deduplicate data before splitting
  - Oversample after splitting

Deduplicate: remove all duplicates in the dataset
(often partially automated)

Causes of data leakage

1. Processing before splitting
2. Data duplication
3. Group leakage
   a. A group of examples have strongly correlated labels but are divided into different splits
   b. Example: CT scans of the same patient a week apart
   c. Solution: Understand your data and keep track of its metadata

Causes of data leakage

1. Processing before splitting
2. Data duplication
3. Group leakage
4. Leakage from data generation & collection process
   a. Example: doctors send high-risk patients to a better scanner
   b. Solution: Data normalization + subject matter expertise

Causes of data leakage

1. Processing before splitting
2. Data duplication
3. Group leakage
4. Leakage from data generation & collection process

Components of ML model training phase

1. Data Collection (+ pre-processing)
1. Model (algorithm) selection
1. Evaluation metrics
1. Training strategy (hyper-params)
1. Sanity checks / verification
1. Actual training

Model Selection Phase

Neural network? Just ML? Or no learning at all?

For ML cases,

- Function to be learned
  - E.g. model architecture, number of hidden layers
- Objective function to optimize (minimize)
  - Loss function
- Learning procedure (optimizer)
  - Adam, Momentum

Best solution: what worked for others - *improve incrementally*

1. Avoid the state-of-the-art trap

- SOTA's promise
  - Why use an old solution when a newer one exists?
  - It's exciting to work on shiny things
  - Marketing

1. Avoid the state-of-the-art trap

- SOTA’s reality
  - SOTA on research data != SOTA on your data
  - Cost
  - Latency
  - Proven industry success
  - Community support

2. Start with the simplest models

- Easier to deploy
    - Deploying early allows validating pipeline
- Easier to debug
- Easier to improve upon

2. Start with the simplest models

- Easier to deploy
  - Deploying early allows validating pipeline
- Easier to debug
- Easier to improve upon
- Simplest models != models with the least effort
  - BERT is easy to start with pretrained model, but not the simplest

BERT: complex large language model costing ~$7000 to train once

3. Avoid human biases in selecting

- A tale of human biases
  - Papers proposing LSTM variants show that the variants improve upon the vanilla LSTM.
  - Do they?

3. Avoid human biases in selecting

- A tale of human biases
  - Papers proposing LSTM variants show that the variants improve upon the vanilla LSTM.
  - Do they?

LSTM: A Search Space Odyssey

Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber

We conclude that the most commonly used LSTM architecture (vanilla LSTM) performs reasonably well on various datasets. None of the eight investigated modifications significantly improves performance.

3. Avoid human biases in selecting

- Evaluate models under comparable conditions
  - It’s tempting to run more experiments for X because you’re more excited about X
- Never happens: X is *always* better than Y
  - There’s almost always some case weaker model Y > X

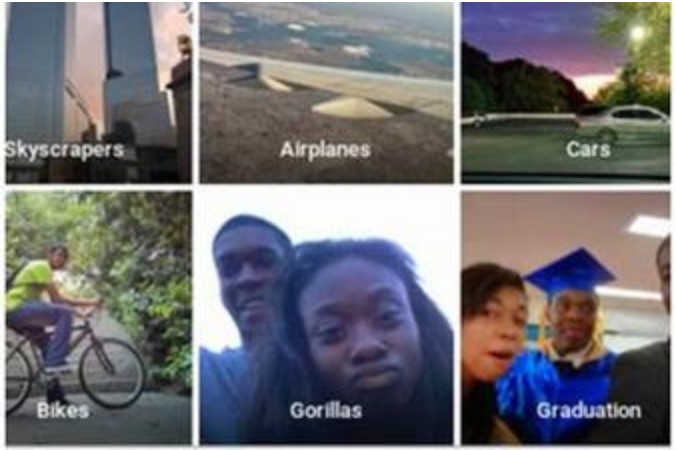
- **Often simpler, more established models will better suit your use-case**

Evaluation

Evaluation metrics: key requirements

- Simple, relatable metrics
- Metrics related to real world
- Real World Testing

Ethics & Bias



The Berkeley study found that both face-to-face and online lenders rejected a total of 1.3 million creditworthy black and Latino applicants between 2008 and 2015. Researchers said they believe the applicants "would have been accepted had the applicant not been in these minority groups." That's because when they used the income and credit scores of the rejected applications but deleted the race identifiers, the mortgage application was accepted.

Solution?

Thorough Evaluation?

What does evaluation look like?

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Source: Brown, T. B., et. al., Language Models are Few-Shot Learners
(GPT-3 paper)

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Source: Vaswani, A., et. al., Attention Is All You Need
(Transformer paper)

Leaderboard

Rank	Team	AUC	MRR	nDCG@5	nDCG@10
1 OCT. 05, 2021	UniUM-Fastformer-Pretrain	0.7304	0.3770	0.4180	0.4718
2 SEPT. 02, 2021	MINER	0.7275	0.3724	0.4102	0.4661
3 AUG. 08, 2021	UniUM-Fastformer	0.7268	0.3745	0.4151	0.4684
4 MAR. 04, 2021	UniUM	0.7243	0.3706	0.4101	0.4644
5 FEB. 27, 2021	chenghuige	0.7209	0.3676	0.4040	0.4597
6 FEB. 26, 2021	UNBERT	0.7207	0.3677	0.4041	0.4602
7 JUN. 21, 2021	wsm_SotA	0.7196	0.3636	0.3998	0.4560
8 NOV. 30, 2021	only2233	0.7189	0.3673	0.4043	0.4603

Source: MIND: Microsoft News Dataset (A Large-Scale English Dataset for News Recommendation Research) Leaderboard

Things can still fail…

Even at top companies!

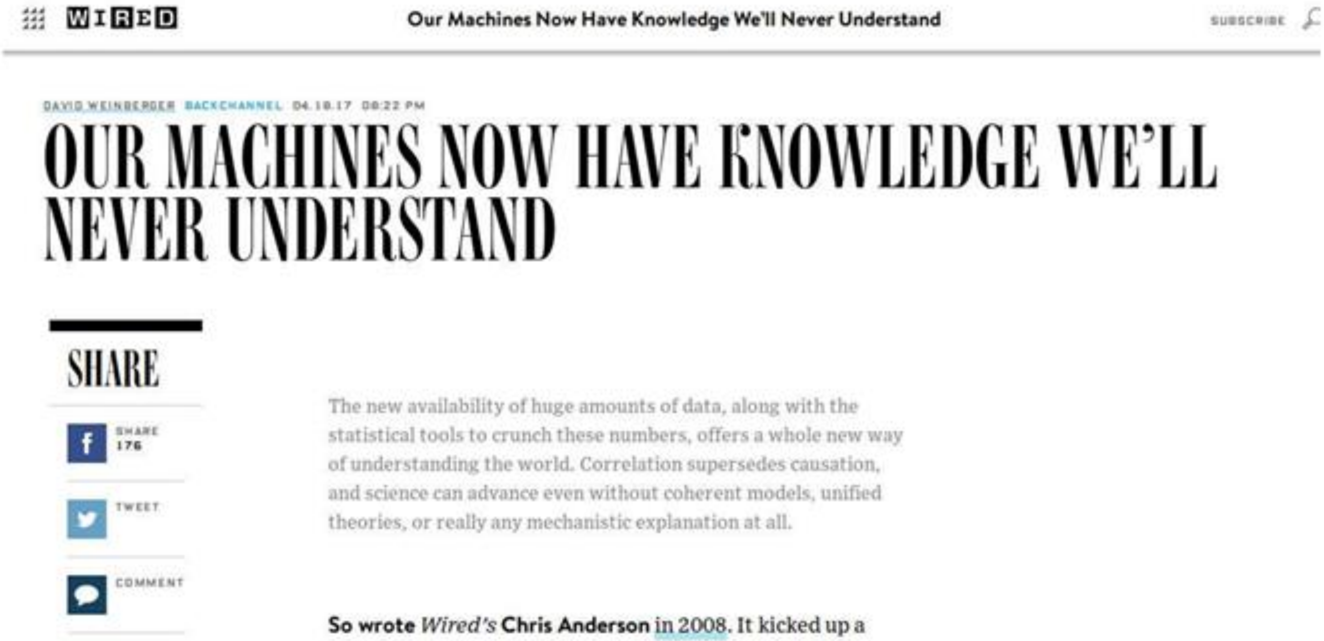
Facebook translates 'good morning' into 'attack them', leading to arrest

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

Solution?

Interpretability?

Is interpretability possible at all?



WIRED Our Machines Now Have Knowledge We'll Never Understand SUBSCRIBE

DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

OUR MACHINES NOW HAVE KNOWLEDGE WE'LL NEVER UNDERSTAND

SHARE

SHARE 176

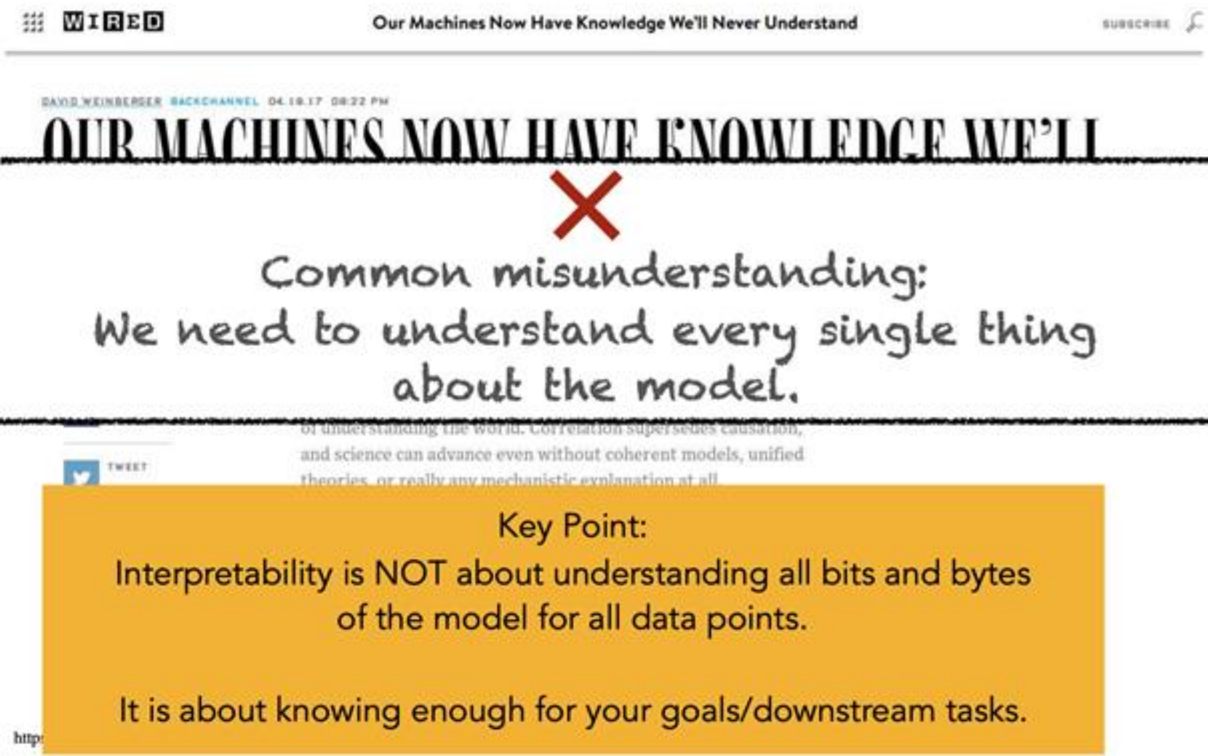
TWEET

COMMENT

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

So wrote *Wired's* Chris Anderson in 2008. It kicked up a

Is interpretability possible at all?



Common misunderstanding:
We need to understand every single thing about the model.

Key Point:
Interpretability is NOT about understanding all bits and bytes of the model for all data points.

It is about knowing enough for your goals/downstream tasks.

My goal

interpretability

To use machine learning **responsibly**
we need to ensure that
1. our **values** are aligned
2. our **knowledge** is reflected

Interpretability useful in all stages!

- Data exploration
  - analyse and visualize the data
- Building the model
  - capabilities and limitations of algorithm
- After building model
  - what it has learned

Solution?

Responsible use?

Responsible AI key concepts

- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment