AI trends that I unironically love

Chris Ré
Stanford
Three Trends I unironically love

• Data-Centric AI
• Declarative Machine Learning
• Foundation Models
Big Wave: Huge Investment in AI

In 2016, big companies invest huge sums to commoditize models. If everyone has models, we thought: *What’s next?*

- **Data-Centric AI.** Success or failure depends more on data. Data deserves 1st class study—like classical data management. (Snorkel)

- **Declarative ML.** Building a model no longer a resume builder, just an obstacle to getting your job done. *Allow you to focus on the data!*

*Models are now more commoditized more than I could have ever imagined!*
Foundation Models

Feed a huge model huge amounts of data and amazing things happen!

- E.g., GPT3, CLIP, Dall-E, PALM, Jurassic, ... amazing!
- **In-context Learning.** One model that can handle many tasks with no retraining in natural language.
- Build apps in hours that would have taken years

**Bet 2: Good now, Better Later.** Amount of investment is astonishing!

*Photo Credit Dalle-2. “An Astronaut Riding a Horse in a Photo-Realistic Style”*
Foundation Models without Soylent (or AGI)

Practical reasons to love foundation models

**Sealed Engines:** Lots of small details in ML pipelines. Learned representations reduce engineers making irrelevant variations that break production.

- Ex: System I built trains itself each week for **years** (tens of FTE to 0.5 FTE to monitor).

**Death-By-A-Thousand-Cuts Problems:** In some problems, no *instance is hard per se*, but sheer variety of reasoning is.

- E.g., entity matching, always simple clues—but selecting right simple background knowledge is hard. (more later)
Rest of the Talk

• How I got here? (My history and bias)
• Optimism
  • Data-Centric AI
  • Foundation Models
• New abstractions, new problems.
My professional history is multithreaded...

BEWARE: My enthusiasm may distract from my bias and myopia.
In antiquity, we were trying to build ML models for “dark data” (extraction, integration, cleaning)

Key idea: uplevel from ML algorithms train with SGD/Sampling

Highlights:
• Fighting human trafficking—absolute privilege!
• Higher than volunteer accuracy at extraction [Nature14]
• Scale-up large models (HogWild! 2011) NeurIPS test of Time 2020

We started a company...

Learned Deskill critical: PIs willing to trade students for data...
In 2017, Apple bought Lattice.

Apple was Lattice customer: most kept building that product.

Feng and I forked off (unruly)
  • We built a bunch of production use cases around extraction, search, integration.
  • Crazy to say but we shipped a lot of code! Our work was used by ~ billion users! 1st for me—awesome!

We locked in on two big problems:
  1. Make developers productive with machine learning,
  2. Build complex tail-driven applications (like entity linking)
Overton: A Data System for Monitoring and Improving Machine-Learned Products

Christopher Ré
Apple

Feng Niu
Apple

Pallavi Gudipati
Apple

Charles Srisuwananukorn
Apple

September 13, 2019

CIDR 2020

Overton a **declarative AI system** for “zero-code deep learning”

- Used in production for handful of services
- Written on my Ikea couch

Met Piero Molino, the **artist** behind Ludwig at Uber, which is similar to Overton (better in many important ways!)

Inspired work like Meta’s Looper system (2021) which runs Meta’s AI products.

**Declarative MACHINE LEARNING SYSTEMS**

**THE FUTURE OF MACHINE LEARNING WILL Depend ON IT BEING IN THE HANDS OF THE REST OF US.**

PIEROMOLINOANDCHRISTOPHERRÉ

In CACM & ACM Queue
I stayed at Apple for ~ 3 years and cofounded 3 companies while there.

Apple was 10 out of 10... Great Experience!
  • Apple bought Inductiv in 2020.

I am best early, and I cofounded an incubator and investment firm called Factory.
  • I invest in these technologies, so true believer or a shill?
My Myopic Slice of Data-Centric AI
ML Application =

Model + Data + Hardware

State-of-the-art models and hardware are available.
Training data is not
But supervision comes from god herself....
... but training data usually comes from a dirty, messy process.

Can we provide mathematical and systems structure for this messy process?
Supervision is where the action is...

Model differences **overrated**, and supervision differences **underrated**.
Automated Chest X-ray Triage

Optimizing Workflows with Automated Prioritization, Radiology 19

What’s the Problem?

Radiologist shortage leaves patient care at risk, warns royal college

*BMJ* 2017; 359  doi: https://doi.org/10.1136/bmj.j4683 (Published 11 October 2017)
Cite this as: *BMJ* 2017;359:j4683

Improving Patient Safety: Avoiding Unread Imaging Exams in the National VA Enterprise Electronic Health Record.

Bastawrous S¹,², Carney B³.
Is Deep Learning the Answer?

This is not an easy question...
- No benchmark dataset
- Effects of data quality are unclear
- No assessment of existing algorithms
- No feedback from clinical community

...so we spent a year trying to answer it!
- Created large dataset of clinical labels
- Evaluated effect of label quality
- Work published in a clinical journal

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW + KSVM</td>
<td>0.88</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.87</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.89</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Often: Differences in models ~ 2-3 points.

Almost a year to obtain high quality data, week to run the model.
Even in Benchmarks: Data Augmentation is Critical

Ex: 13.4 pt. avg. accuracy gain from data augmentation across top ten CIFAR-100 models—difference in top-10 models is less!
Training Signal is key to pushing SotA

New methods for gathering signal leading the state of the art

Google AI AutoAugment: Using learned data augmentation policies
  • Augmentation Policies first in Ratner et al. NIPS ’17

Facebook Hash tag weakly supervised pre-training
  • Pre-train using a massive dataset with hashtags
Training data: the new bottleneck

Slow, expensive, and static
Manual Labels

Programmatic Labels

Trade-off: programmatic labels are noisy…
Key Idea: Model Training Creation Process

This talk:

1. An interface for generating training data via weak supervision
2. An approach to learn quality and correlations of sources
3. Training an end model---in various domains
Snorkel: Formalizing Programmatic Labeling

Pattern Matching

regex.match(
    r"{A} is caused by {B}" 
)

[e.g. Hearst 1992, Snow 2004]

Distant Supervision

Subset A
Subset B
Subset C

[e.g. Mintz 2009]

Augmentation

"Change abbreviate names, and replace…"

Topic Models

[e.g. Hingmire 2014]

Third-Party Models

[e.g. Schapire 1998]

Third-Party Models

Crowdsourcing

[e.g. Dalvi 2013, Karger & Oh 2011]

Observation: Weak supervision applied in *ad hoc* and isolated ways.
Snorkel: Formalizing Programmatic Labeling

Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling
A Toy Example (and light math)
Running Example: NER

Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

Let’s look at labeling “Person” versus “Hospital”

Goal: Label training data using weak supervision strategies for these tasks
Weak Supervision as Labeling Functions

Problem: These noisy sources conflict and are correlated—no source of ground truth.

Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

```
def existing_classifier(x):
    return off_shelf_classifier(x)
```

```
def upper_case_existing_classifier(x):
    if all(map(is_upper, x.split())) and 
    off_shelf_classifier(x) == 'PERSON':
        return PERSON
```

```
def is_in_hospital_name_DB(x):
    if x in HOSPITAL_NAMES_DB:
        return HOSPITAL
```

“PERSON”

“PERSON”

“HOSPITAL”

Problem: These noisy sources conflict and are correlated—no source of ground truth.
The Classical Snorkel Pipeline

1. Users write labeling functions to generate noisy labels

2. Snorkel models and combines the noisy labels into probabilities

3. The resulting probabilistic labels train a model

Key Idea: Probabilistic training point carries accuracy. No hand labeled data needed.
def existing_classifier(x):
    return off_shelf_classifier(x)

def upper_case_existing_classifier(x):
    if all(map(is_upper, x.split())) and off_shelf_classifier(x) == 'PERSON':
        return PERSON

def is_in_hospital_name_DB(x):
    if x in HOSPITAL_NAMES_DB:
        return HOSPITAL

How to learn the parameters of this model (accuracies & correlations) without $Y$?
def existing_classifier(x):
    return off_shelf_classifier(x)

def upper_case_existing_classifier(x):
    if all(map(is_upper, x.split())) and \
    off_shelf_classifier(x) == 'PERSON':
        return PERSON

def is_in_hospital_name_DB(x):
    if x in HOSPITAL_NAMES_DB:
        return HOSPITAL

Key idea: We can observe overlapping judgements on many points to estimate accuracy
Solution Sketch: Using the covariance

\[ \Sigma = \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & Y \\ \lambda_1 & \Sigma_0 & & \\ \lambda_2 & & \lambda_2 & \\ \lambda_3 & & & \lambda_3 \\ Y & & & Y \end{pmatrix} \]

Note: \( E[Y\lambda_i] \) is proportional to the accuracy of source \( i \)

**But** we can’t observe directly only agreement and disagreement rate i.e., a *portion* of the covariance (\( \Sigma_0 \))
Idea: Use graph-sparsity of the inverse

\[
\Sigma_0 = \Sigma + z z^T
\]

Rank-1 params to solve for (\sim\ function of accuracies)

- \( E[z_i] = 1 \) if perfectly accurate
- \( E[z_i] = 0 \) if random noise

Fewer degrees of freedom: Roughly, zero where corresponding pair of variables has no edge

[\text{Loh & Wainwright 2013, Ratner et al. 2019}]

For now, assume we know the graph (dependency structure)...

\text{matrix inversion lemma}
Result: A matrix completion problem?

We get a set of equations. For any pair $i \neq j$ with no edge in graph—the lhs is 0

\[ 0 = \left( \Sigma_0^{-1} \right)_{i,j} + Z_i Z_j \]

\( \Sigma \) is full rank, so not really matrix completion…

Key: \( \Sigma = I + uu^\top \) for some $u$ so intuitively close…
Couple of Technical Comments

\[ 0 = (\Sigma_0^{-1})_{i,j} + Z_i Z_j \]

- Symmetry: \( z \) and \( -z \) are solutions? What does this mean?
- \( z_i = 0 \) when accuracy 0.5, i.e., total noise! (more samples)
- Effective rank \( \text{er}(\Sigma) = \text{tr}(\Sigma)/|\Sigma|_2 \) (effectively, use this!)
  - small when single large: \( |z|_2 \) is large.
- Scale inversely distance to noise (\( z_i = 0 \)).
Can we learn the **accuracy** of annotators without labels? Yes, to information theoretic limits! [NeurIPS16, ICML19, AIStats21]

How well can we learn the **correlation between annotators**? Under mild assumptions, as well as when we don’t have any labeled data! [ICML19]

**optimally use** labeled data for debiasing? **Bias variance tradeoff for weak supervision** [AIStats21]

**Classical ML Theory Nerds**: Effectively structure learning and estimation for latent variable graphical models. **Improve the sample efficiency rates even for supervised graphical model case via connections to recent results in geometry.**
... in production and you’ve probably used it...

Overton: A Data System for Monitoring and Improving Machine-Learned Products

Christopher Ré
Apple

Feng Niu
Apple

Pallavi Gudipati
Apple

Charles SriSuwanamukorn
Apple

Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale

Stephen H. Bach†
Brown University

Daniel Rodriguez‡
Google

Yintao Liu‡
Stanford University

Chong Luo‡
Google

Haidong Shao‡
Google

Cassandra Xia‡
Google

Souvik Sen†
Stanford University

Alex Ratner§
Google

Braden Hancock§
Google

Houman Alborzi§
Google

Rahul Kuchhal§
Google

Chris Ré§
Google

Rob Malkin‡
Google

†Brown University

‡Google

§Stanford University

Migrating a Privacy-Safe Information Extraction System to a Software 2.0 Design

Ying Shang
Mountain View, CA, USA
yingsheng@google.com

Nguyen Vo
Mountain View, CA, USA
nguyenvo@google.com

James B. Wendt
Mountain View, CA, USA
jwendt@google.com

Sandeep Tata
Mountain View, CA, USA
tata@google.com

Marc Najork
Mountain View, CA, USA
najork@google.com

Leveraging Organizational Resources to Adapt Models to New Data Modalities

Sahaana Suri†;
Raghuveer Chanda, Neslihan Bulut, Pradyumna Narayana, Yemao Zeng
Peter Bailis†, Sugato Basu, Girija Narlikar, Christopher Ré†, Abishek Sethi

†Google, Stanford

Thank you, Google and Apple!
Snorkel, the company, is much bigger than Programmatic Supervision

For research, zoomed on one aspect: how to combine all source of supervision.

- **This is a big enabler:** no other method lets you reuse labels, combine labels from other purposes, bring together everything.
- **Time to Value:** Customers going from “AI takes months to minutes”

A new way to **manage** and **build** AI applications.

- Point of view on entire data pipeline for AI: managing, monitoring, improving, and using.
- Exciting to see people getting this!
Data-Centric AI

Thinking about data has foundational theory, algorithmic, and practical advantages.

Trying to build a community of interested folks!

- Github repo, mailing list, and more!
- https://github.com/HazyResearch/data-centric-ai
- See Andrew Ng’s https://datacentricai.org
- https://mlsys.stanford.edu Karan and Dan have a weekly podcast.
- Pioneering workshops like SIGMOD’s DEEM!

Karan Goel
Foundation Model
Warmup
My Priming to love them.
Named Entity Disambiguation

Map “strings to things” (A database)
- Key part of assistant, search, and information extraction

Ex: input "How tall is Lincoln?"

Q216796  Q28260  Q91
Variation

How tall is Lincoln?

What is the cheapest Lincoln?

How many people are in Lincoln?

Subtle clues and varied clues. Death by 1000 cuts.
Our Entity Resolution Model

In late 2017 at Apple, enhancing a service that was doing NED (and QA, IE, topic rec)

• Built high-quality solutions for US English…
  • Mostly deep model but had hand-tuned KG features…
• maintenance was an issue…

Apple overlords: Great work! Now... roll out to tens of languages/locales...

... but not tens of new head count...
So we read...

Remove manual bottlenecks with weak supervision & self-supervision

- ELMO paper comes out!
- Snorkel getting traction at Google Ads.
- Maybe a little worse but many locales?

1st version stomped our hand-tuned model (Mid 2018). Humbled!
1st cut model in production in 2018 (stayed there)

Great quality bumps across locales, changed slightly over time...Woo hoo!

My view: Self-supervision and Data-centric AI were going to be critical.

Became insufferable about this
Bootleg and Apple Folks

Simran Arora

Neel Guha

Megan Leszczynski

Xiao Ling

Laurel Orr

Sen Wu
Foundation Models and their uses in Data Plumbing

Avanika Narayan  Ines Chami
Autoregressive Language Models

**Simple, Old idea.** Complete the sentence “The mouse ate the” ranked by probability learned from the corpus.

\[ p(\text{the}, \text{mouse}, \text{ate}, \text{the}, \text{cheese}) = 0.02, \]
\[ p(\text{the}, \text{cheese}, \text{ate}, \text{the}, \text{mouse}) = 0.01, \]
\[ p(\text{mouse}, \text{the}, \text{the}, \text{cheese}, \text{ate}) = 0.0001. \]

**Neural Language Models.**

- **Neural nets** “compactly” represents that probability function
- **Generate.** Generate answers using conditional probability.
- **Train.** Every single token in a sentence is an example.

https://stanford-cs324.github.io/winter2022/lectures/introduction/
Foundation Models

- Autoregressive neural language models
- **Very large** number of parameters
- **Very large** training corpuses

Emergent Behaviors: Generalize to new tasks with NO finetuning (Few-shot)

Input Text

Task description
Example:
Get Capital from Country
- France => Paris
- Germany => Berlin
- China => Beijing
- Japan => Tokyo
- Canada =>

Prompt

GPT-3

Output Text

Ottawa

Natural language text in and out. Awesome! Sometimes... GPT3 also likes Toronto

[Brown et al. 20]
Emergent Behaviors

It works on many different language tasks....

- Translation

Input Text

Translate English to French
Cheese => fromage
Wine =>

GPT-3

Output Text

Vin

[Translation performance graph]

[Brown et al. 20]
Emergent Behaviors

It works on many different language tasks:

- Translation
- Trivia / Question Answering

Example:

Q: ‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?

A: Marcel Duchamp
Emergent Behaviors

It works on many different language tasks....

- Translation
- Trivia / Question Answering
- Arithmetic
- And many more.....

**Input Text**

Q: What is 17 minus 14?
A:

**Output Text**

GPT-3

3

[Brown et al. 20]
Amazing Two Years Ago, Better in the Future

Bet: Great now, and they get better over time
Not just text: Code and Images

Describe **code** via comment string:

```
Given a list of (@link Person)s, remove the duplicates and return the result sorted by age
```

FM (Codex) generates code!

```
import java.util.Comparator;
import java.util.List;
import java.util.stream.Collectors;

public class PersonUtils {
    /**
     * Given a list of (@link Person)s, remove the duplicates
     * and return the result sorted by age.
     */
    public static List<Person> removeDuplicates(List<Person> people) {
        return people.stream()
            .distinct()
            .sorted(Comparator.comparing(Person::getAge))
            .collect(Collectors.toList());
    }
}
```

Developers and faculty have told me 2-3x more productive using the vscode plugin!

https://copilot.github.com

Describe **image** in natural language:

An astronaut, playing basketball with cats in space a children's book illustration

FM (DALLE-2 or ImageGen) generates an image!

Can Foundation Models prevent “death by 1000 cuts problems” in data?
Example: Data Cleaning

- Goal: Detect and repair errors in structured data

- Diverse errors:
  - Typos and formatting
  - Conflicting values
  - Outlier values

<table>
<thead>
<tr>
<th>DBAName</th>
<th>AKAName</th>
<th>Address</th>
<th>City</th>
<th>State</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>John Veliotis Sr.</td>
<td>Johnnyo's</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3465 S Morgan ST</td>
<td>Chicago</td>
<td>IL</td>
<td>60609</td>
</tr>
<tr>
<td>t2</td>
<td>John Veliotis Sr.</td>
<td>Johnnyo's</td>
<td></td>
<td>IL</td>
<td>60609</td>
</tr>
<tr>
<td>t3</td>
<td>John Veliotis Sr.</td>
<td>Johnnyo's</td>
<td></td>
<td>IL</td>
<td>60609</td>
</tr>
<tr>
<td>t4</td>
<td>Johnnyo's</td>
<td>Johnnyo's</td>
<td></td>
<td>IL</td>
<td>60608</td>
</tr>
</tbody>
</table>

Data cleaning has a “death by a thousand cuts” feel
HoloClean was a big jump on state of the art (10+ points) became Inductiv, which was Acquired by Apple 2020.
Foundation Models for Data Tasks

**Input Table**

<table>
<thead>
<tr>
<th></th>
<th>Country</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>England</td>
<td>Kyoto</td>
</tr>
</tbody>
</table>

**Input Text + Task Demonstrations**

Is there an error in Country?
- Country: USA, City: Miami? No
- Country: England, City: Kyoto?

**Output Text**

- Yes

Zero-shot works, but not as good as SotA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entity Matching</th>
<th>Imputation</th>
<th>Error Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous SoTA</td>
<td>97.1</td>
<td>94.4</td>
<td>86.8</td>
</tr>
<tr>
<td></td>
<td>77.2</td>
<td>96.5</td>
<td>94.4</td>
</tr>
<tr>
<td>GPT-3 (Zero-shot)</td>
<td>40.3</td>
<td>63.6</td>
<td>49.1</td>
</tr>
<tr>
<td></td>
<td>73.3</td>
<td>89.2</td>
<td>12.1</td>
</tr>
<tr>
<td>GPT-3 (Few-shot)</td>
<td>98.2</td>
<td>1.00</td>
<td>87.0</td>
</tr>
<tr>
<td></td>
<td>90.0</td>
<td>97.1</td>
<td>98.1</td>
</tr>
</tbody>
</table>

Few-shot on this model trained only predict words and not retrained! **Wild!**

... but it can be brittle ...
Caveat: The Prompts Matter

No more hand-engineered rules…. but needs **prompt engineering** to work

### Small formatting difference matter

| Is there an error in Country? | Country: USA, City: Miami? No |
| Country: France, City: New York? Yes |
| Country: England, City: Kyoto? Yes |

| Is there an error in Country? | Country: USA City: Miami No |
| Country: France City: New York Yes |
| Country: England City: Kyoto No |

Without "," and "?" separator token, the FM fails to generate the correct answer

### Task demonstrations matter

**Table: GPT-3 (175B) few shot performance**

<table>
<thead>
<tr>
<th></th>
<th>Random Examples</th>
<th>Manual Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fodors-Zagats</td>
<td>0.930</td>
<td>1.00</td>
</tr>
<tr>
<td>Beer</td>
<td>0.933</td>
<td>1.00</td>
</tr>
<tr>
<td>Restaurant</td>
<td>0.790</td>
<td>0.895</td>
</tr>
</tbody>
</table>

Gap

Changing in-context demonstration significantly impacts model performance

Can Foundation Models Help Wrangle Data? Narayan 22, Coming Soon
Caveat: The Training Data Matters!

**FM*s benefit from **data curation**: diversity and quality in training data distributions is critical to robustness (Fang et al.)

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

Alex Fang† Gabriel Ilharco† Mitchell Wortsman† Yuhao Wan†

Vaishaal Shankar° Achal Dave° Ludwig Schmidt†°

---

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP), Fang et al. 2022,
Is it just GPT3?

Can Jurassic-1 (J-1), another large FM, also do data tasks?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entity Matching</th>
<th>Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iTunes-Amazon</td>
<td>Beer</td>
</tr>
<tr>
<td><strong>GPT-3</strong></td>
<td>98.2</td>
<td>100</td>
</tr>
<tr>
<td><strong>J-1</strong></td>
<td>98.2</td>
<td>100</td>
</tr>
</tbody>
</table>

Yes! This is general, J-1 can be SoTA, but sometimes needs more task demonstrations.

... but they speak their own language.

Even GPT variants differ a lot (InstructGPT v. Davinci 1)

Prompts are not universal!

GPT-3 Entity Matching Prompt

Product A is name: macbook air. price: 199.00. Product B is name: macbook pro. price: 199.00. Are Product A and Product B the same?

J-1 Entity Matching Prompt

Product A:
name: macbook air
price: 199.00

Product B:
name: macbook pro
price: 199.00

Q: Similar?
A:
Longer Contexts

FMIs struggle to model long-range dependencies and larger contexts.

Active work to address this challenge...

- **Long-range Arena Benchmark** (LRA) by Tay et al. 21 – noted this problem.
- **Memorizing Transformers**: kNN & external memory w/ transformers by Wu et al. 21
- **S4**: A state-space sequence model Gu et al., SoTA by 20+ points on LRA.
- **Monarch**: Much longer sequences with Transformers [ICML22, Long Oral]
- **Flash Attention**: Fastest attention—1st to get non-trivial Path-X quality

**Memorizing Transformers**

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, Christian Szegedy
{yuhuai,mrabe,delesley,szegedy}@google.com

**S4** ICLR22, Oral, Honorable Mention Outstanding Paper

Efficiently Modeling Long Sequences with Structured State Spaces, Gu et al. ICLR 2022
Monarch: Expressive Structured Matrices for Efficient and Accurate Training, ICML22.
Folks in our community have been on this!

Constructing an Interactive Natural Language Interface for Relational Databases

Fei Li
Univ. of Michigan, Ann Arbor
lifei@umich.edu

H. V. Jagadish
Univ. of Michigan, Ann Arbor
jag@umich.edu

Best paper in VLDB 2015.

SIGMOD/VLDB pioneers have been on natural language for data for a while!

Jignesh Patel
Wisconsin & CEO
DataChat
Foundation Models Summary

Good now: Wild what they can do...
Lots of investment. What will they do next?

• Seem ripe for data management folks:
  • Essentially, *functions* of data
  • In production: build, maintain, and use on many problems
  • New attacks on *death-by-one-thousand-cuts* problems?
  • They offer hope to change the interface to data!

A lot more benchmarking at coming at center for crfm.stanford.edu led by the indominatble Percy Liang!
New Challenges from a Data-Centric AI Viewpoint

Hidden Stratification
Is Deep Learning the Answer?

This is not an easy question...

- No benchmark dataset
- Effects of data quality are unclear
- No assessment of existing algorithms

Are we sure those differences are causal? Anticausal?

- Created large dataset of clinical labels
- Evaluated effect of label quality
- Work published in a clinical journal

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW + KSVM</td>
<td>0.88</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.87</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.89</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td><strong>0.91</strong></td>
</tr>
</tbody>
</table>

Often: Differences in models ~ 2-3 points.

Almost a year to obtain high quality data, week to run the model.
These eyes haunt me...

Any model may pick out *unintended signal*. Deep models may pick out *more* unintended signal.

Upshot: Picked up on *mascara*

Kuehlkamp et al. *Gender-from-Iris or Gender from-Mascara*

Do we know how well these models are really performing?
Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

New Abstractions, New Problems

Gustavo Carneiro
Lauren Oakden-Rayner
Jared Dunnmon
It’s not just those eyes...

Melanoma Recognition (Surgical Marks)

Pneumothorax Detection (Collapsed Lung)

No Drain

With Drain

Pneumothorax detection 0.87 AUC, which is superhuman... with chest drains—*Chest drain means already treated!* Down to 0.77 when removed...

Image Credits
One issue: Hidden Stratification.

- Issue: A subset of a class (stratum) performs worse.
  - *without a drain is worse than with a drain are two strata*
  - Abnormal contains many unlabeled subclasses or strata.

- Classical: Never write features that say
  - If drain then pneumonia
  - if purple dot then cancer
  - But new SW abstraction, new bugs

Develop a theory & techniques to handle hidden stratification in the data?
No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems

Nimit S. Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, Christopher Ré

Preprint link: stanford.edu/~nims/no_subclass_left_behind.pdf.
Blog: hazyresearch.stanford.edu/hidden-stratification
**Observation:** Deep Nets learn features that can distinguish between subclasses—even if trained with coarser labels!

1. **Train NN model**
2. **Cluster NN features to estimate subclasses**
3. **Train robust NN model**

**Identify simple assumptions, \(O(1/\sqrt{n})\) per-subclass generalization error:** *same sample complexity as if we knew the true subclasses.*

**Improve accuracy on worst-performing subclass up to 22 points!**
Update: use contrastive learning to learn representations that discard spurious information.

1.) Collect trained model predictions

1.

Camel (On Sand)

\( \hat{y} = \text{“Camel”} \checkmark \)

2.) Train robust model via aligning representations by class label only

2.

Cow (On Grass)

“Cow” \x

Cow (On Grass)

“Cow” \checkmark

Without training group labels, improve worst-group accuracy up to 41.1 points!

Nearly closes the gap with robustness methods that require group labels—without labels!

Inspired by great work on robustness! … and many others! …

https://wilds.stanford.edu
Conclusion: Waves are Building

• **Data-Centric AI** is still in its first innings in industry, and a massive opportunity.

• **Foundation Models** aren’t even out of the bullpen yet, and they offer new attacks on classical problems.

• Fundamental challenges in robustness and building applications—great for research!
# Misc. Prompting (Brittleness)

- Performance varies as a result of minor changes to prompt

<table>
<thead>
<tr>
<th>Full Serialization</th>
<th>Column Sub Selection</th>
<th>Different Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product A is <strong>name</strong>: Runoff IPA. <strong>factory</strong>: Odell Brewing Co. <strong>style</strong>: American Amber. <strong>ABV</strong>: 4.6 %. Product B is <strong>name</strong>: Red Nectar. <strong>factory</strong>: Humboldt Brewing Co. <strong>style</strong>: Red Ale. <strong>ABV</strong>: 5.40 % Are Product A and Product B the Same?</td>
<td>Product A is <strong>name</strong>: Runoff IPA. <strong>factory</strong>: Odell Brewing Co. <strong>style</strong>: American Amber. <strong>ABV</strong>: 4.6 %. Product B is <strong>name</strong>: Red Nectar. <strong>factory</strong>: Humboldt Brewing Co. <strong>style</strong>: Red Ale. <strong>ABV</strong>: 5.40 % Are Product A and Product B the Same?</td>
<td>Product A is <strong>name</strong>: Runoff IPA</td>
</tr>
</tbody>
</table>