Practical Systems-Level Development Challenges

Danna Gurari
University of Colorado Boulder
Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture
  - Josh Myers-Dean on “Tuning Foundation Models”
- Assignments (Canvas)
  - Lab assignment 3 due tomorrow (Friday) night
  - Final project outline due in 3 weeks
- Questions?

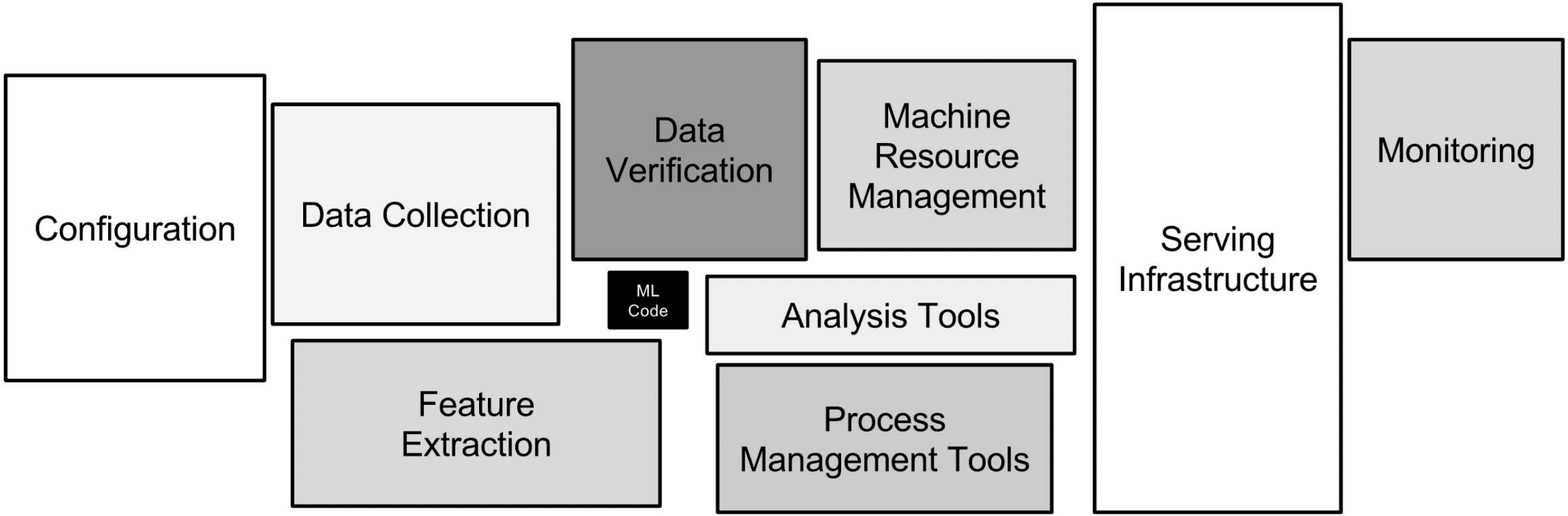
Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences

Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences

Most Code for a DL System Is Not for DL



Sculley et al. Neurips 2015

Researchers Have Examined Efforts to Build Software with AI Capabilities; e.g.,

(Neurips 2015)

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Researchers Have Examined Efforts to Build Software with AI Capabilities; e.g.,

(ICSE-SEIP 2019)

Software Engineering for Machine Learning: A Case Study

Saleema Amershi
Microsoft Research
Redmond, WA USA
samershi@microsoft.com

Andrew Begel
Microsoft Research
Redmond, WA USA
andrew.begel@microsoft.com

Christian Bird
Microsoft Research
Redmond, WA USA
cbird@microsoft.com

Robert DeLine
Microsoft Research
Redmond, WA USA
rdeline@microsoft.com

Harald Gall
University of Zurich
Zurich, Switzerland
gall@ifi.uzh.ch

Ece Kamar
Microsoft Research
Redmond, WA USA
eckamar@microsoft.com

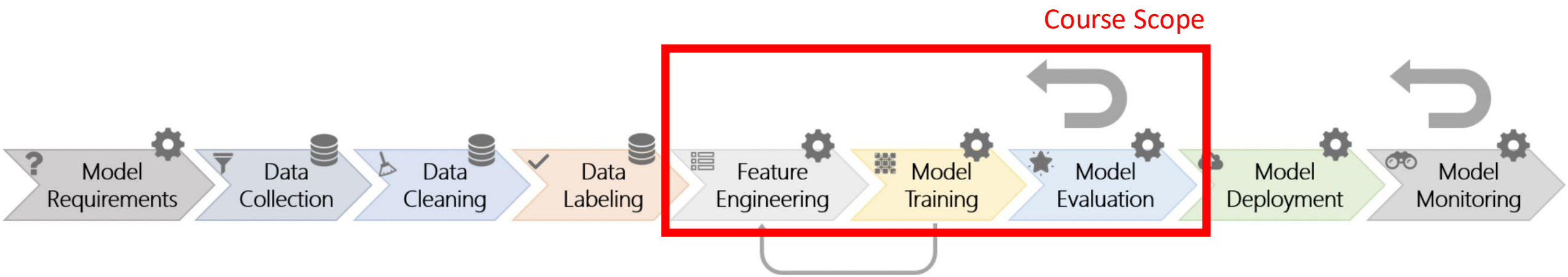
Nachiappan Nagappan
Microsoft Research
Redmond, WA USA
nachin@microsoft.com

Besmira Nushi
Microsoft Research
Redmond, WA USA
besmira.nushi@microsoft.com

Thomas Zimmermann
Microsoft Research
Redmond, WA USA
tzimmer@microsoft.com

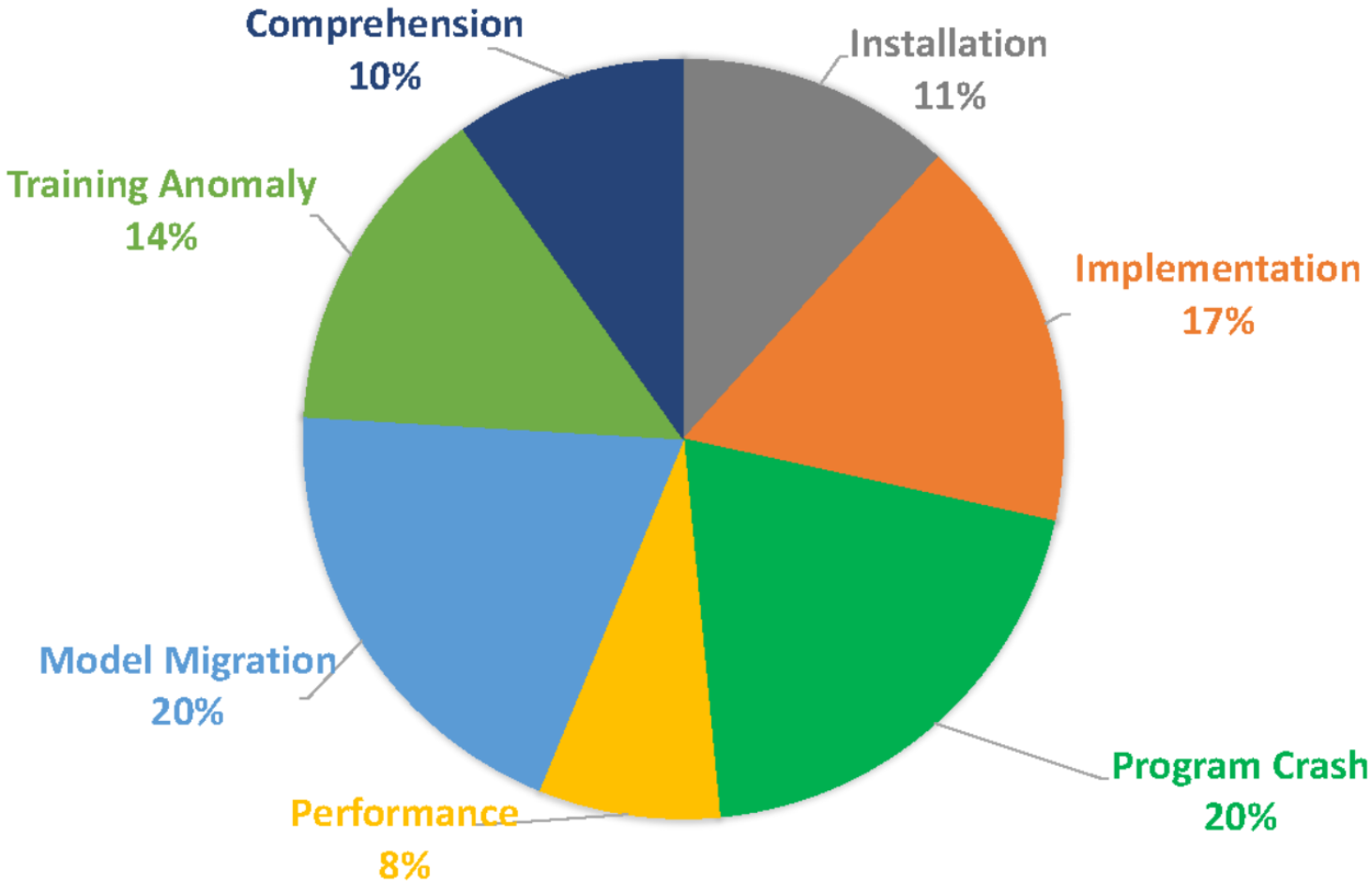
Researchers Have Examined Efforts to Build Software with AI Capabilities; e.g.,

Common ML Workflow (from Microsoft Study of Software Teams Building Software Applications with AI Capabilities)



Amershi et al. ICSE-SEIP 2019

Frequently Asked Questions in Deep Learning



Source: 39,628 Stack Overflow questions from a dump in December 2018 with one of these tags: tensorflow, pytorch, or deeplearning4j

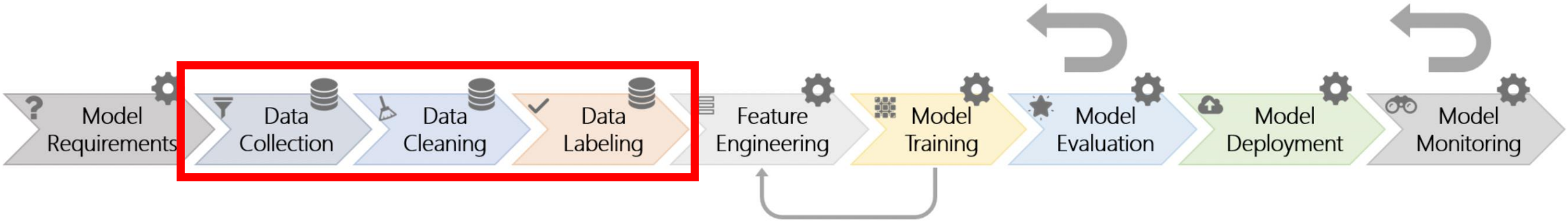
Approach: human labeling for initial set followed by automated labeling

Zhang et al. An Empirical Study of Common Challenges in Developing Deep Learning Applications. ISSRE 2019

Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences

Key Theme: Not Only Code, But Also Data Is Core to Software Engineering



Amershi et al. ICSE-SEIP 2019

(CHI 2021)

"*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan
nithyasamba@google.com
Google Research
Mountain View, CA

Shivani Kapania
kapania@google.com
Google Research
Mountain View, CA

Hannah Highfill
hhighfil@google.com
Google Research
Mountain View, CA

Diana Akrong
dakrong@google.com
Google Research
Mountain View, CA

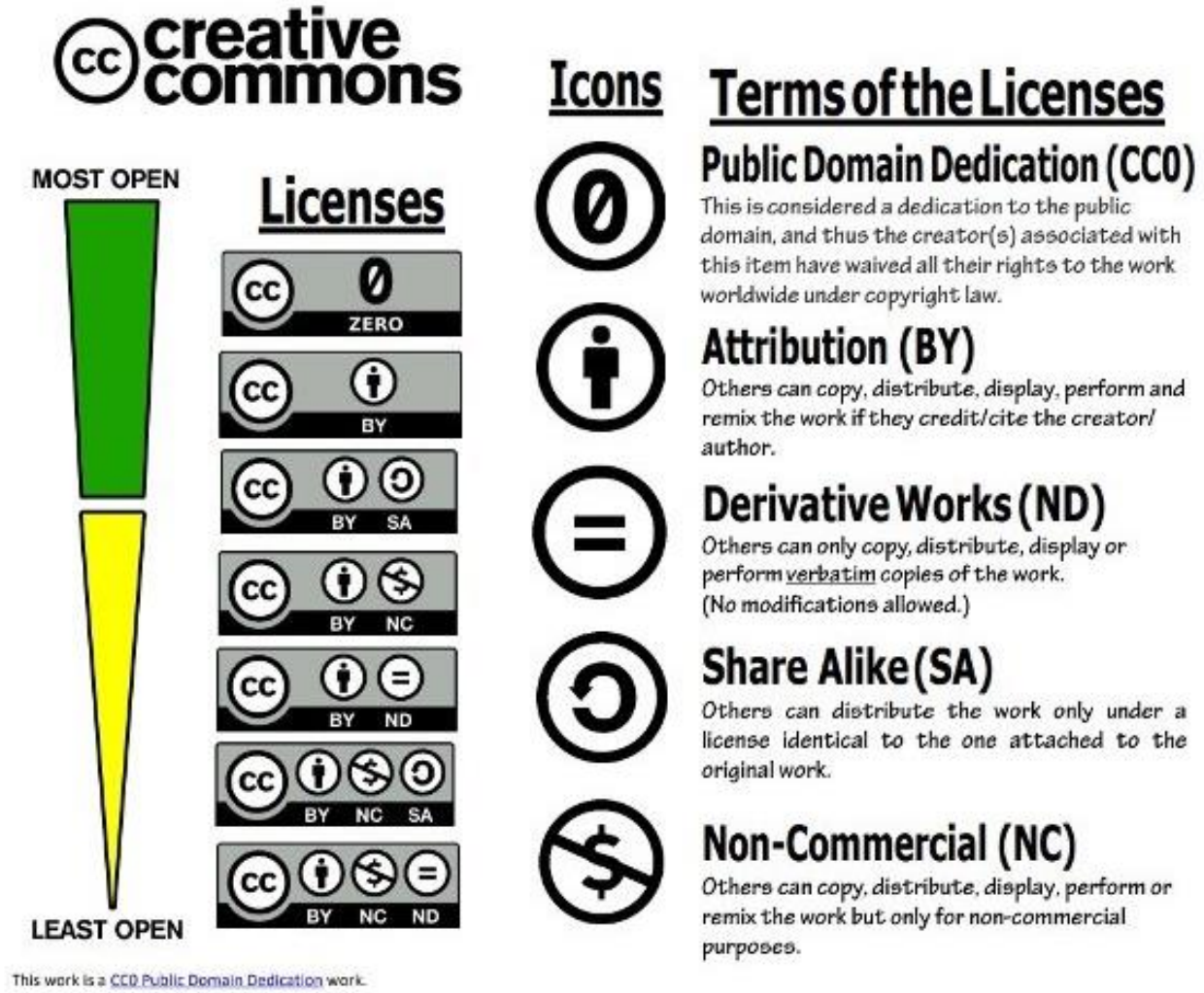
Praveen Paritosh
pkp@google.com
Google Research
Mountain View, CA

Lora Aroyo
loraa@google.com
Google Research
Mountain View, CA

Key Data Collection Approaches

- Human-labeled: e.g., ImageNet, VQA, COCO-Captions
  - What data?
  - What annotation protocol and interface?
  - Who annotates?
  - How to ensure high-quality results?
- Internet-curated: e.g., for GPT-3 and CLIP
  - How to crawl the dataset?
  - How to verify crawled data can be used (e.g., copyright laws)?
  - How to verify the data is sufficient quality for learning (e.g., discard pages with "lorem ipsum", few words, and dirty/obscene language)?
- Authentic use cases: e.g., VizWiz
  - How to obtain real users' data?
  - How to remove all private information?

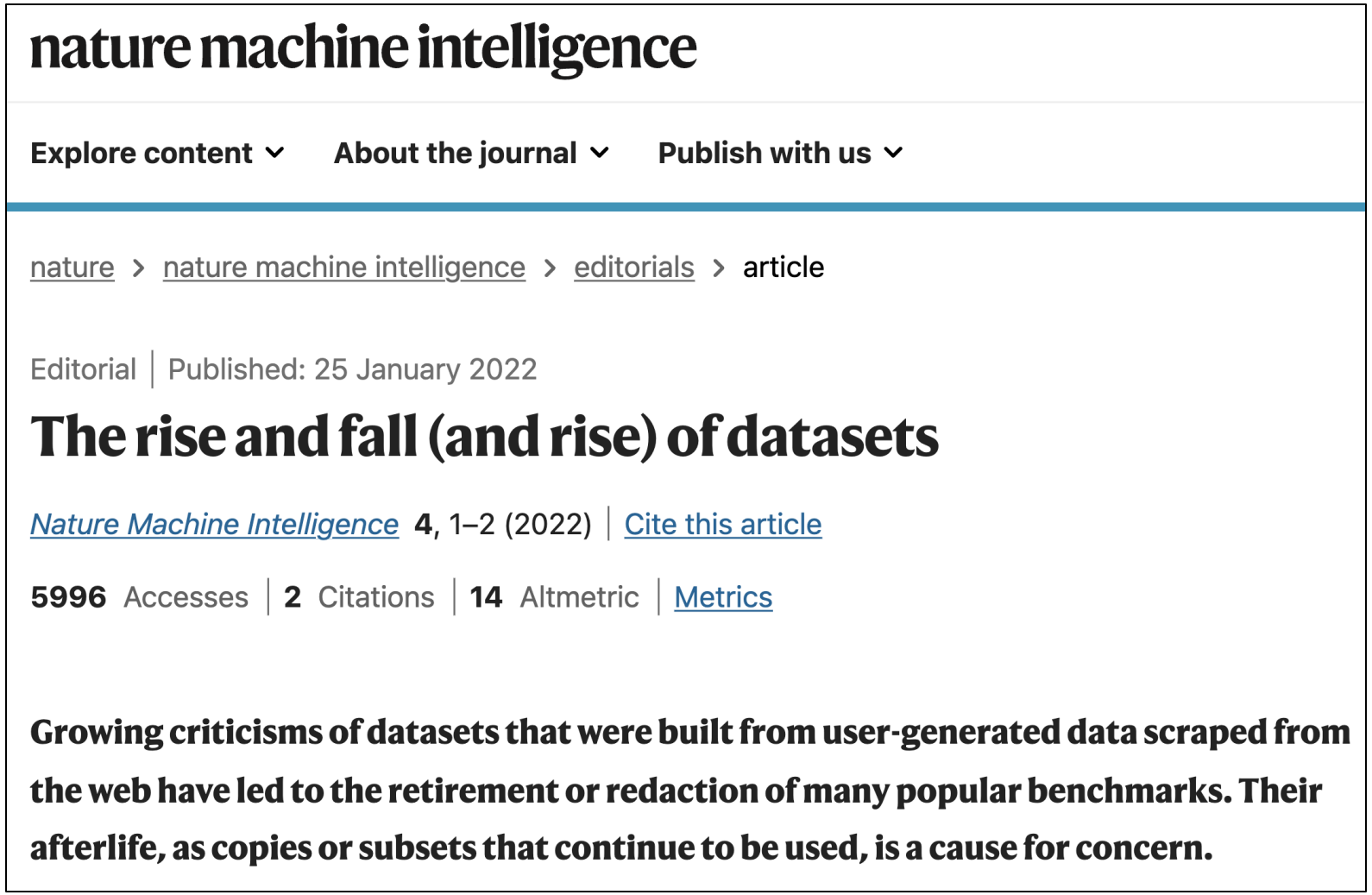
Common Challenge: Suitable License; e.g.,



https://teaching.resources.osu.edu/teaching-topics/simple-guide-creative-commons

Common Challenge: Suitable License; e.g.,

ImageNet redacted:



nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › editorials › article

Editorial | Published: 25 January 2022

The rise and fall (and rise) of datasets

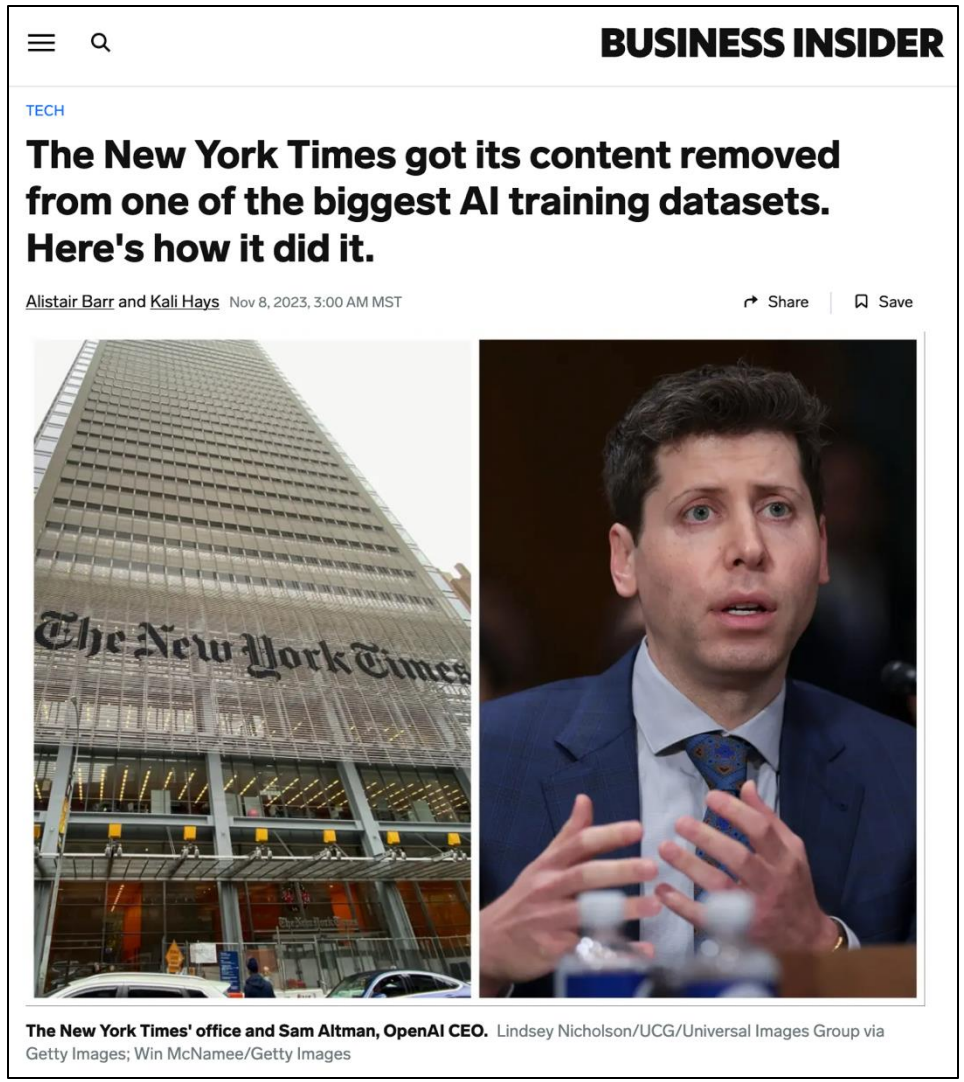
Nature Machine Intelligence **4**, 1–2 (2022) | Cite this article

5996 Accesses | **2** Citations | **14** Altmetric | Metrics

Growing criticisms of datasets that were built from user-generated data scraped from the web have led to the retirement or redaction of many popular benchmarks. Their afterlife, as copies or subsets that continue to be used, is a cause for concern.

https://www.nature.com/articles/s42256-022-00442-2

Common Challenge: Suitable License; e.g.,



BUSINESS INSIDER

TECH

The New York Times got its content removed from one of the biggest AI training datasets. Here's how it did it.

Alistair Barr and Kali Hays Nov 8, 2023, 3:00 AM MST

Share | Save

The New York Times' office and Sam Altman, OpenAI CEO. Lindsey Nicholson/UCG/Universal Images Group via Getty Images; Win McNamee/Getty Images

Common Crawl no longer includes millions of URLs to copyrighted content!

https://www.businessinsider.com/new-york-times-content-removed-common-crawl-ai-training-dataset-2023-11

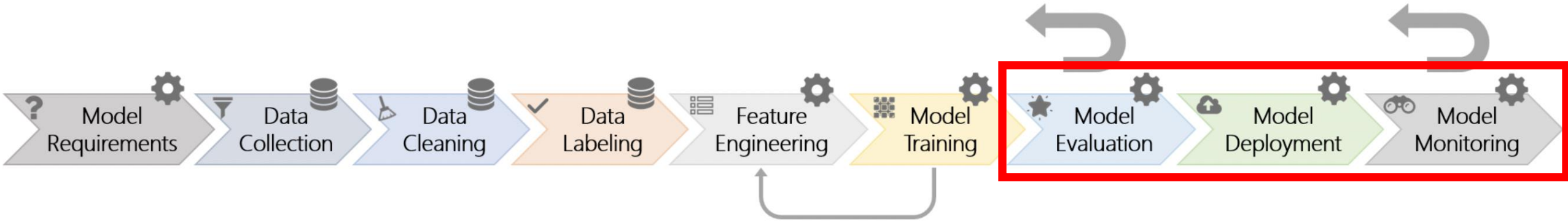
Other Considerations

- Storage costs
- Efficient visualization/search tools
- Data scale (e.g., how to collect rare content; e.g., medical/satellite-based)
- Data distribution (e.g., what’s the skew of representation across contents)

Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences

Model Maintenance

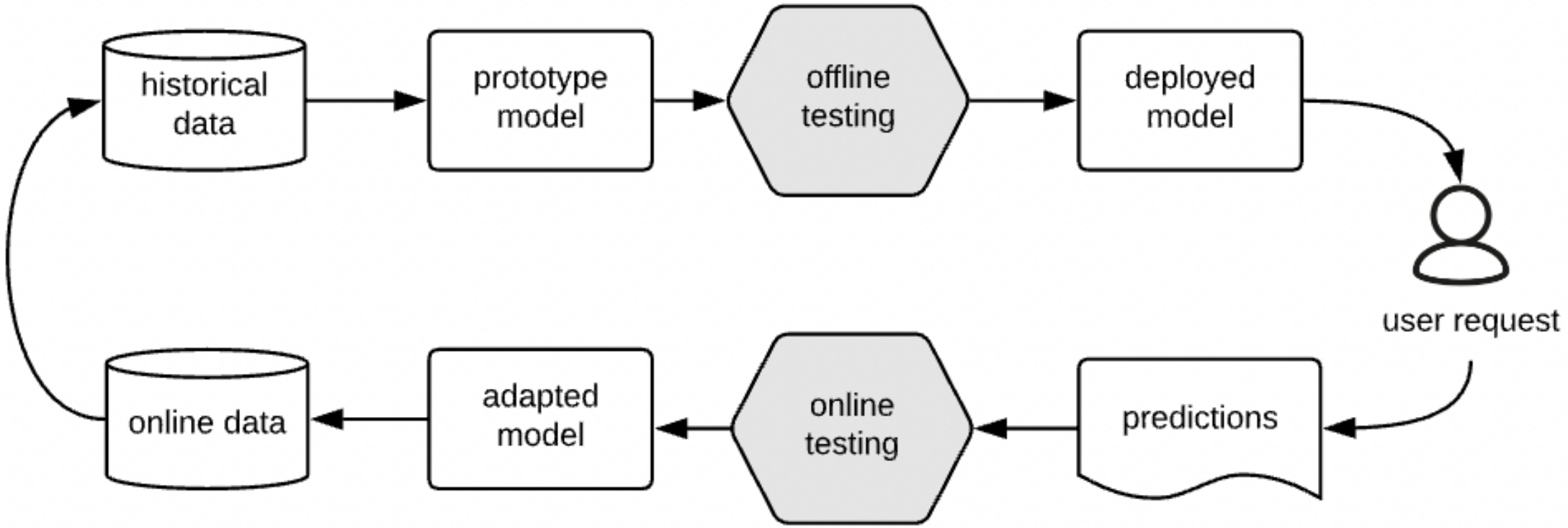


Amershi et al. ICSE-SEIP 2019

Testing Frameworks

Often need rapid release cycles of new models to support current trends, given how quickly data can become outdated; e.g., evolving

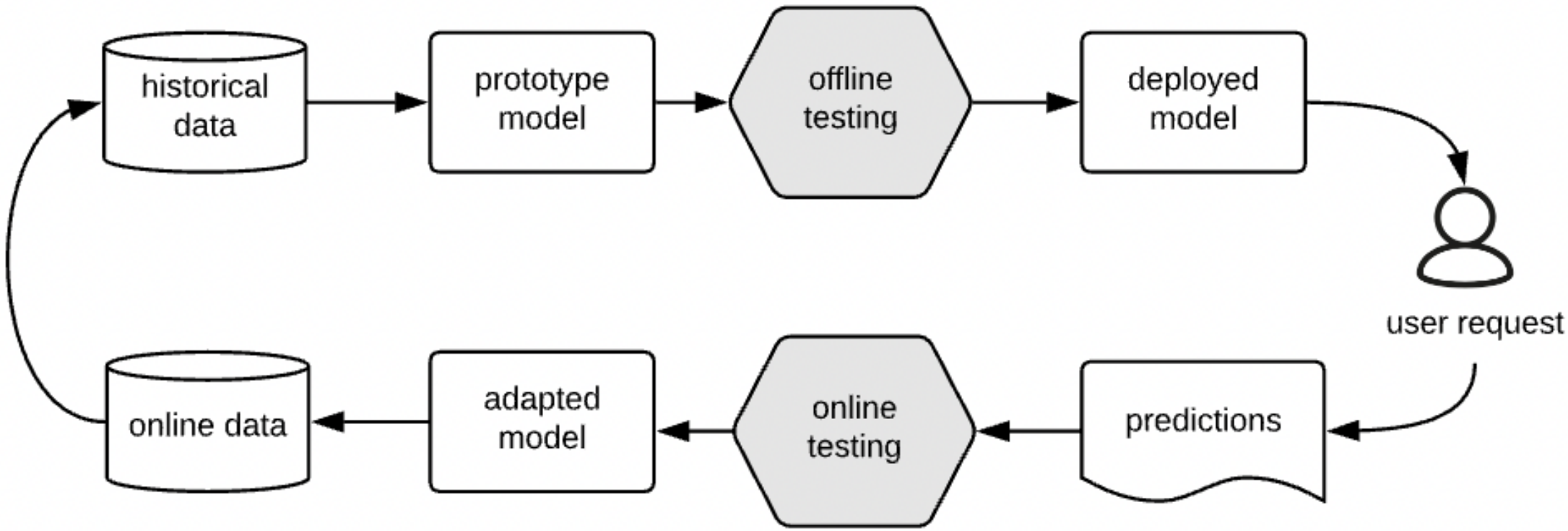
- shopping products
- politics (e.g., who is current president)
- building/product designs (e.g., for blind people's visual interpretation services)



Zhang et al. Machine Learning Testing: Survey, Landscapes and Horizons. IEEE Transactions on Software Engineering 2020

Testing Frameworks

Data suites often are created to test hard cases and/or a model's "ethics"



Zhang et al. Machine Learning Testing: Survey, Landscapes and Horizons. IEEE Transactions on Software Engineering 2020

Unique Software Engineering Requirements

- Unlike traditional software development:
  - Modular design is challenging for DL components (we often choose DL *because* it is impractical to identify human-based rules)
  - Data management and versioning is critical

Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences

Neelima

Everley

Nick

What are you currently working on related to DL, and how do you spend your time when developing DL methods?

Neelima: Background

Graduated from UC San Diego in 2021 with a degree in Mathematics and CS

Worked for the Department of Defense for around 2 years

- **Project War-Gaming**:
  - Leveraged Deep Reinforcement Learning
- **Project: Object (Target) Tracking**
  - Training an RNN on our data to be able to track a target without human input

Wanted to explore more problems in computer vision, especially object tracking

Neelima: Current Research

Leveraged contrastive learning models to understand how data augmentation can improve small object detection

Benchmarking and evaluating state of the art foundation models (like SAM2) for novel tasks

Fine-tuning state of the art models on our custom datasets

Working to create a new tracking model to simultaneously track an object and its respective parts across a video.

Neelima: Evaluating Models for a Novel Task

1. Develop a general idea for what a successful model would be able to accomplish
2. Find existing models that solve a task that is very similar to your novel task, and note the evaluation metrics used to evaluate those models
3. Adapt those evaluation metrics to your novel task, and note how the metrics might change given your current problem.
4. Analyze where the evaluation metrics show the weakness of the model
5. Analyze the qualitative aspects of your results. Are the quantitative results reflecting what you observe visually?

Nick: Background

- Rutgers class of 2021 (math and EE)
  - Various DL research projects including audio processing and music composition
- Software engineer (low level C/Cuda) 2021-23
  - Real time ultra high-res video processing (10-12k)
- PhD program at CU starting F23
  - Focus on efficient DL and the foundations of learning

Nick: Current Research

- Common goal: obtain good performance with fewer computational resources
- Knowledge Distillation: an architectural perspective
  - Extract “dark knowledge” from large teacher models
  - Instill the knowledge into small student models to improve efficiency
  - Logits and hidden layer “latent representations” are the most common bridges between teacher and student
- Core-set Selection: a dataset-centric perspective
  - Extract the “most informative” samples from large scale datasets
- Dataset Distillation: a dataset-centric perspective
  - Reverse engineer synthetic training samples from model gradients
- Real world training time is a function of architectural and dataset perspectives!

Nick: Current Research

- Some interesting questions for DL:
  - Do all layers learn to do the same thing?
  - What is generalization? Can we explicitly model it?
  - Is backpropagation necessary?
  - Can a single theory describe *all* architectures? Could we use such a theory to guide the creation of new architectures?
  - What are the limitations of “low-order” linear algebra?

Nick: Practical Challenges in DL

- Breakdown of time spent working on DL projects/ideas
  - Ideation and theory: 40%
  - Codebase dev and building tools: 30%
  - Running experiments and analyses: 20%
  - Debugging: 10%
- Code quality and reproducability
  - Many research code-bases are poorly written, buggy, and hard to use
  - Code is rarely optimized
  - Codebases often reveal details that were “swept under the carpet” by the original papers
- Managing hyper-parameter explosions
  - Keep close track of experiments!
  - It’s worth it to invest in scripts and tools to organize results

Nick: Tips & Suggestions

- Don’t be afraid to re-implement models/techniques/ideas yourself!
  - Valuable learning experience that helps find bugs or errors in existing code. You might even make things run faster!
- Pay attention to the details!
  - Most deep learning libraries are very high level, and don’t expose many details of function implementations. It’s easy to not notice subtle bugs as a result.
- Be careful with library versions, CUDA driver versions, etc…
  - Sometimes “bugs” can simply be the result of version mis-matches

Everley Tseng

4th year PhD student advised by Dr. Danna Gurari

2017 — **BS** - Engineering and System Science (Electronic Engineering)

2019 — **Industry** - Computer Vision Engineer

2021 — **MS** - Computer Science (Computational Intelligence)

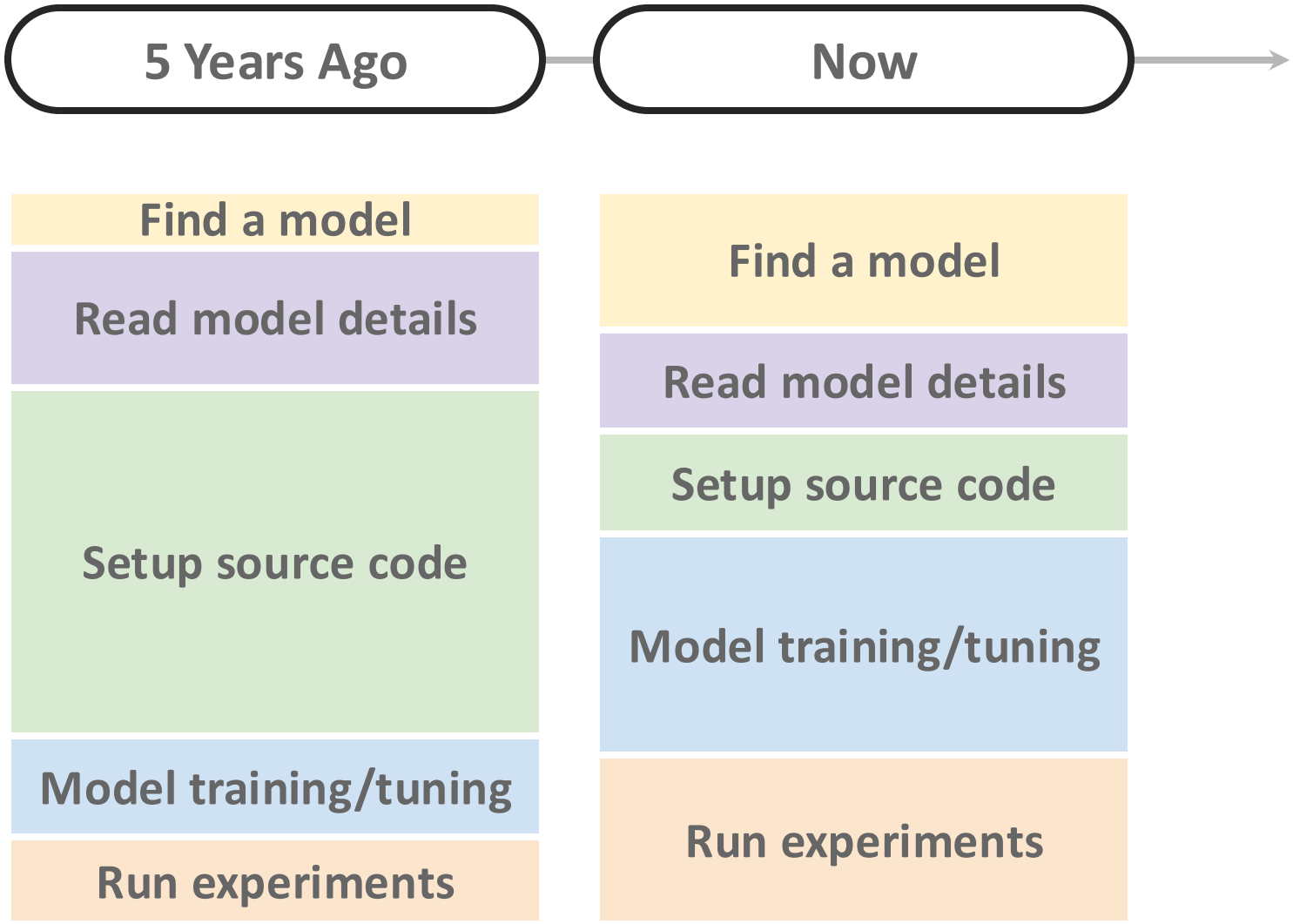
Present — **PhD** - Computer Science

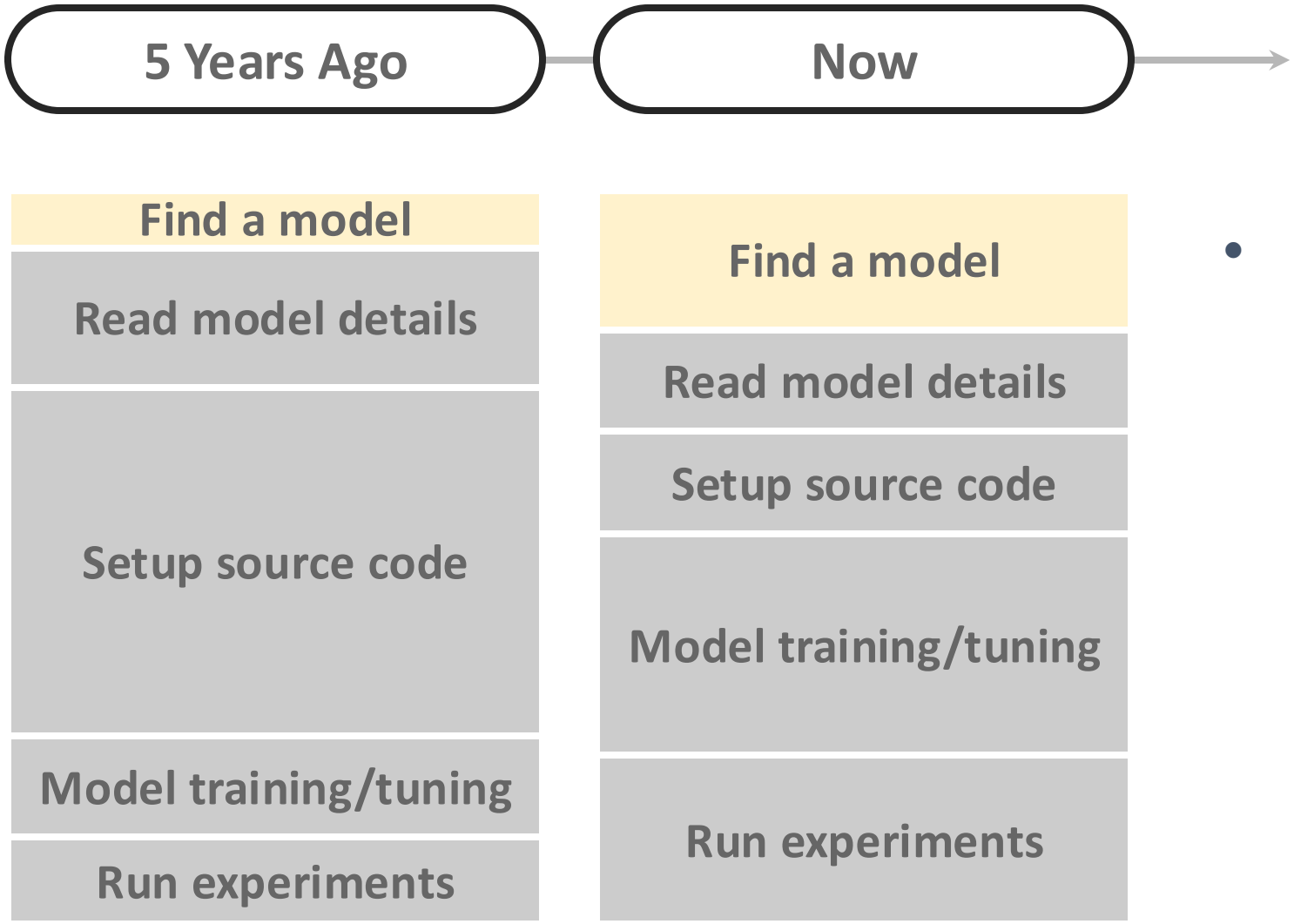
DL Research Areas

- Computer Vision
- Dataset Curation
- Model Benchmarking

Application Aspects

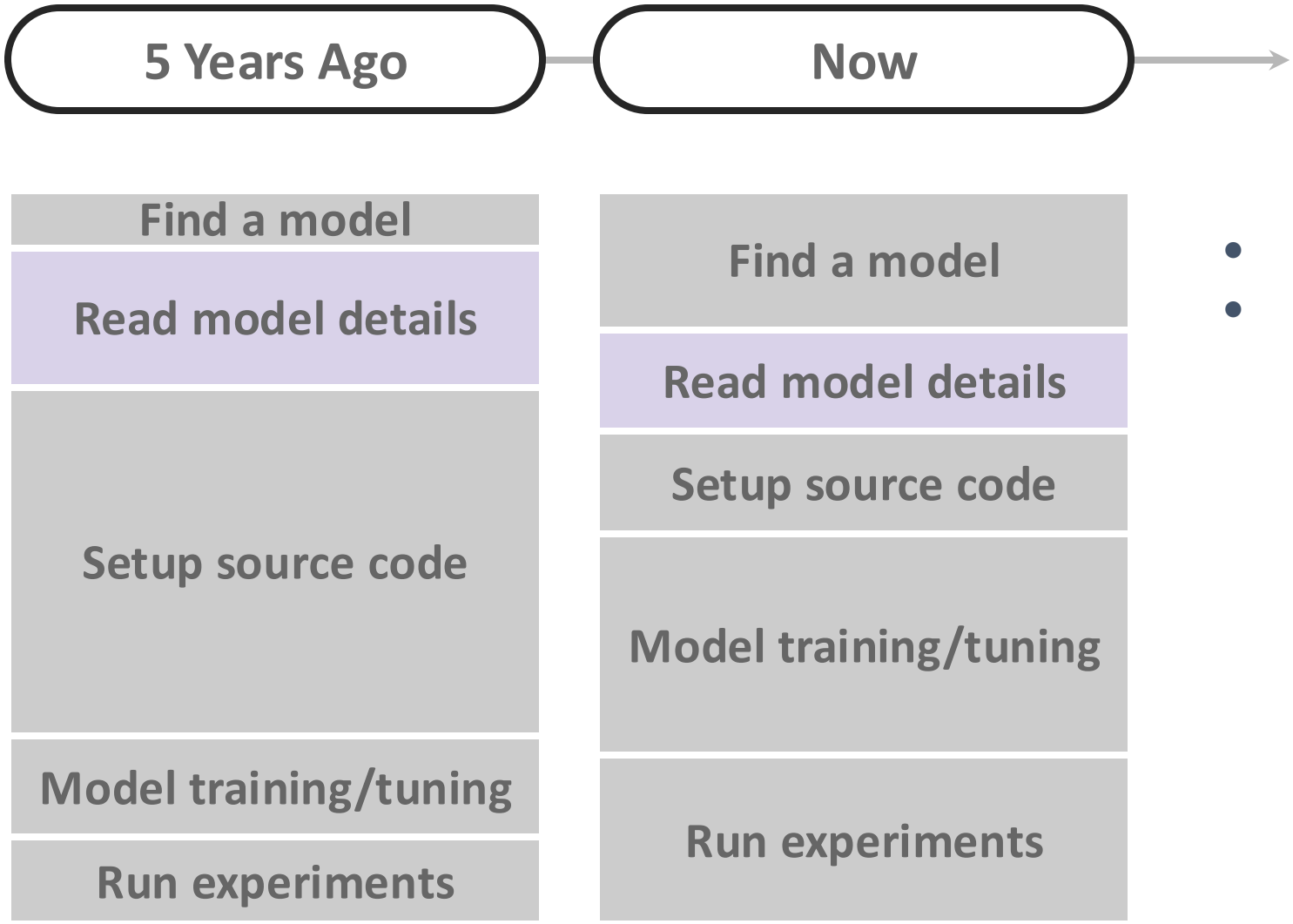
- Accessibility
- Data Privacy


5 Years Ago
Now
What changed?
Find a model
Read model details
Setup source code
Model training/tuning
Run experiments
Find a model
Read model details
Setup source code
Model training/tuning
Run experiments



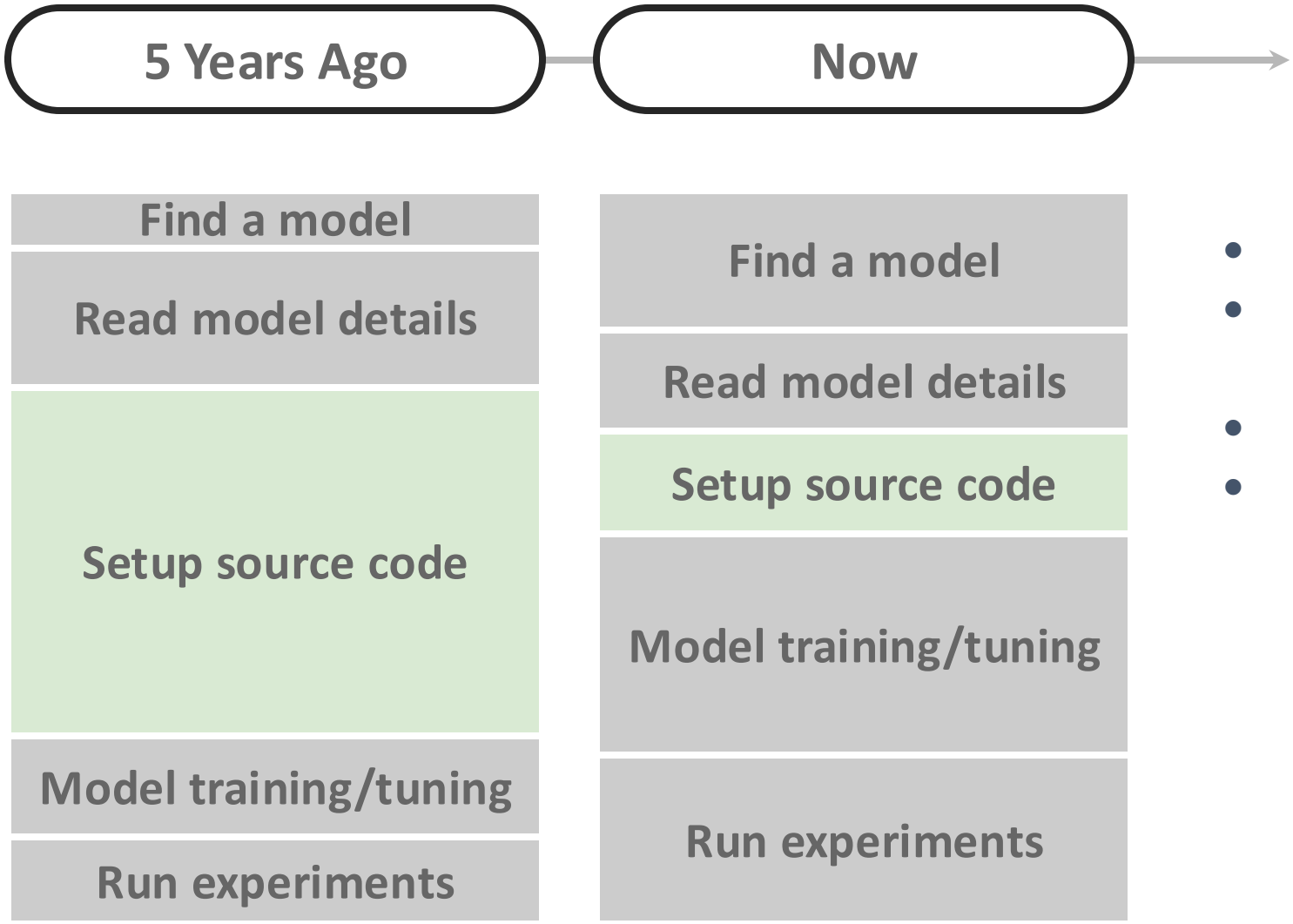
What changed?

- **Growing diversity of approaches**



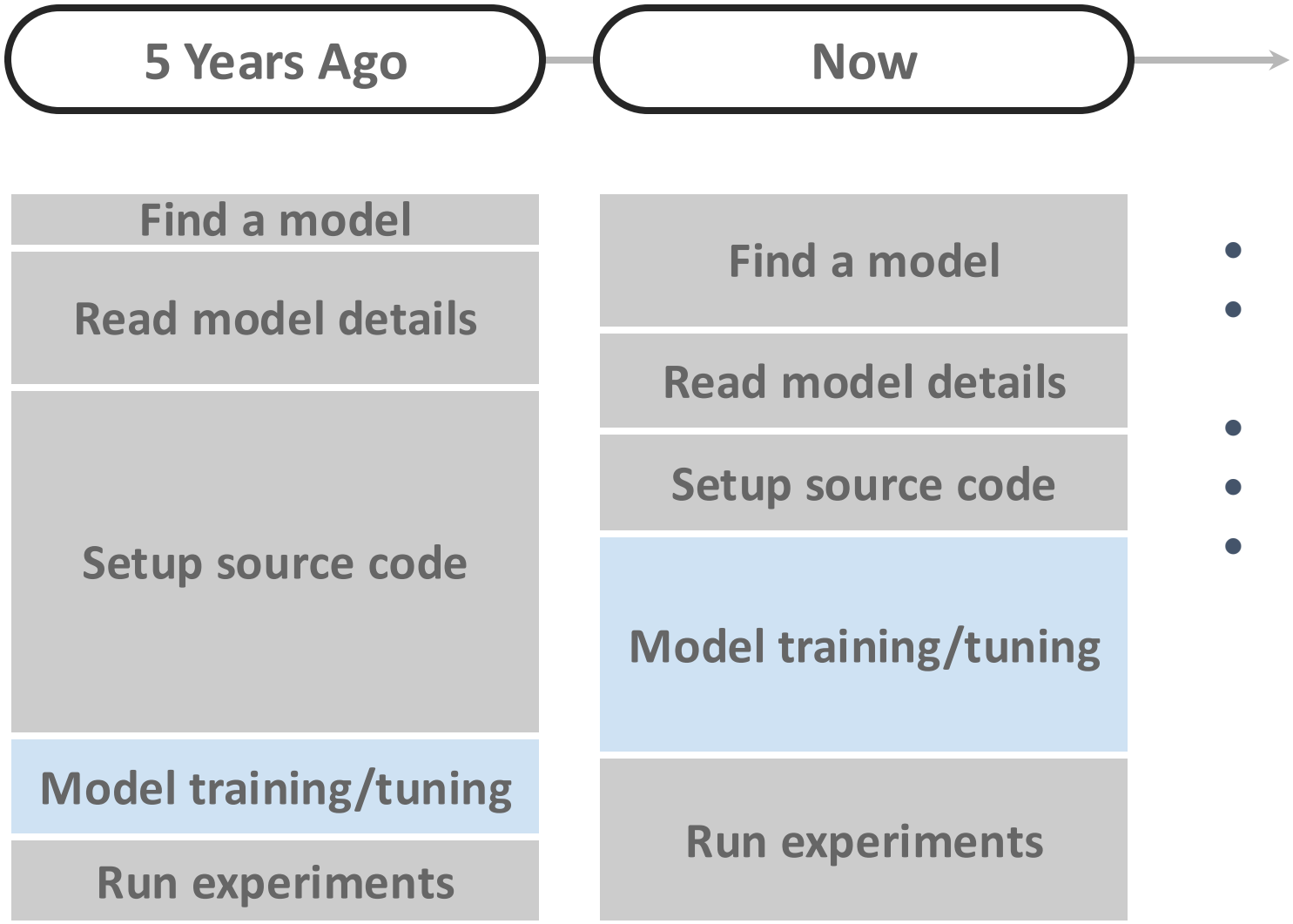
What changed?

- Growing diversity of approaches
- **Less architecture revolution; more training technique revolution**



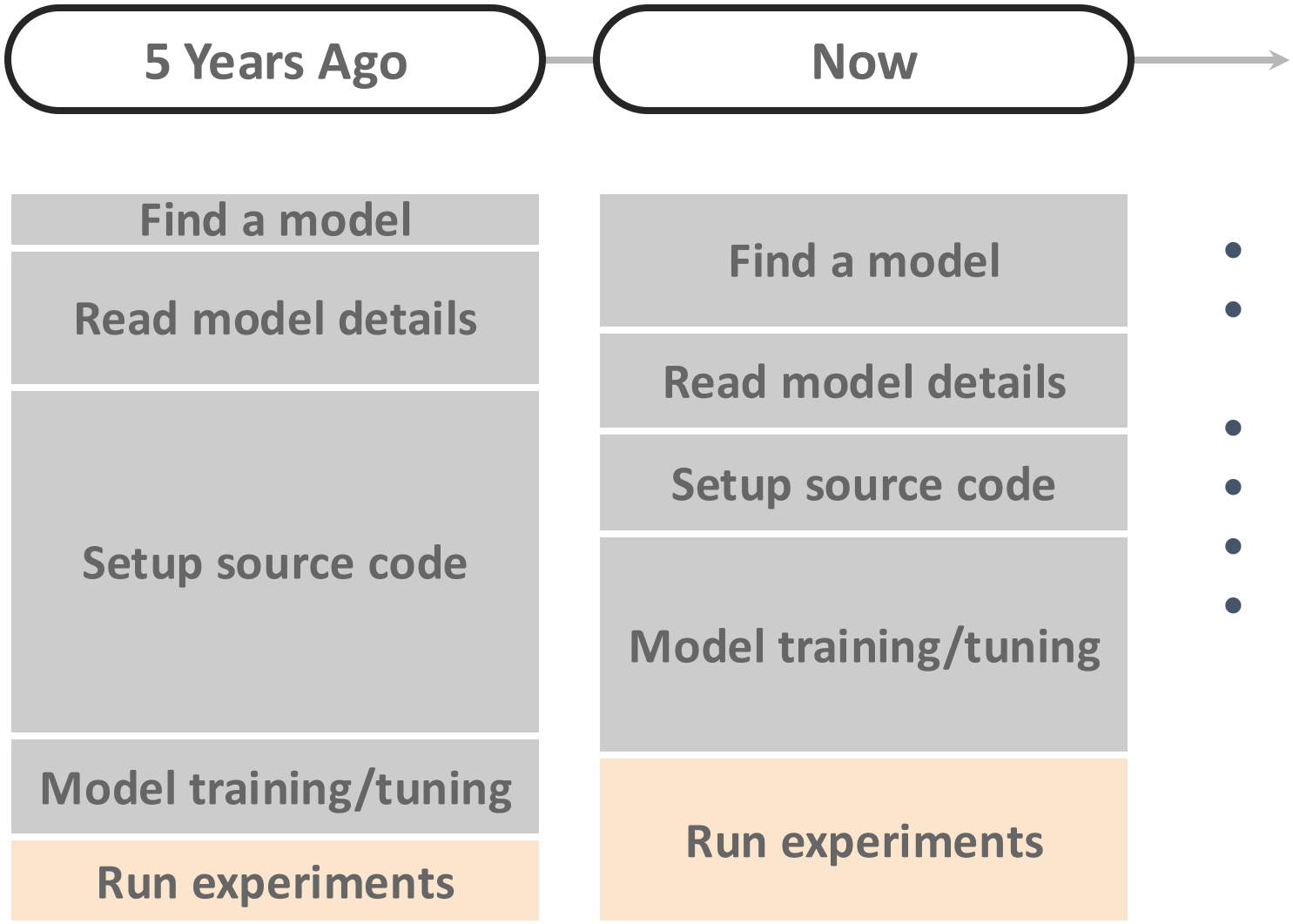
What changed?

- Growing diversity of approaches
- Less architecture revolution; more training technique revolution
- **Hugging Face**
- **Open source & trained weights**



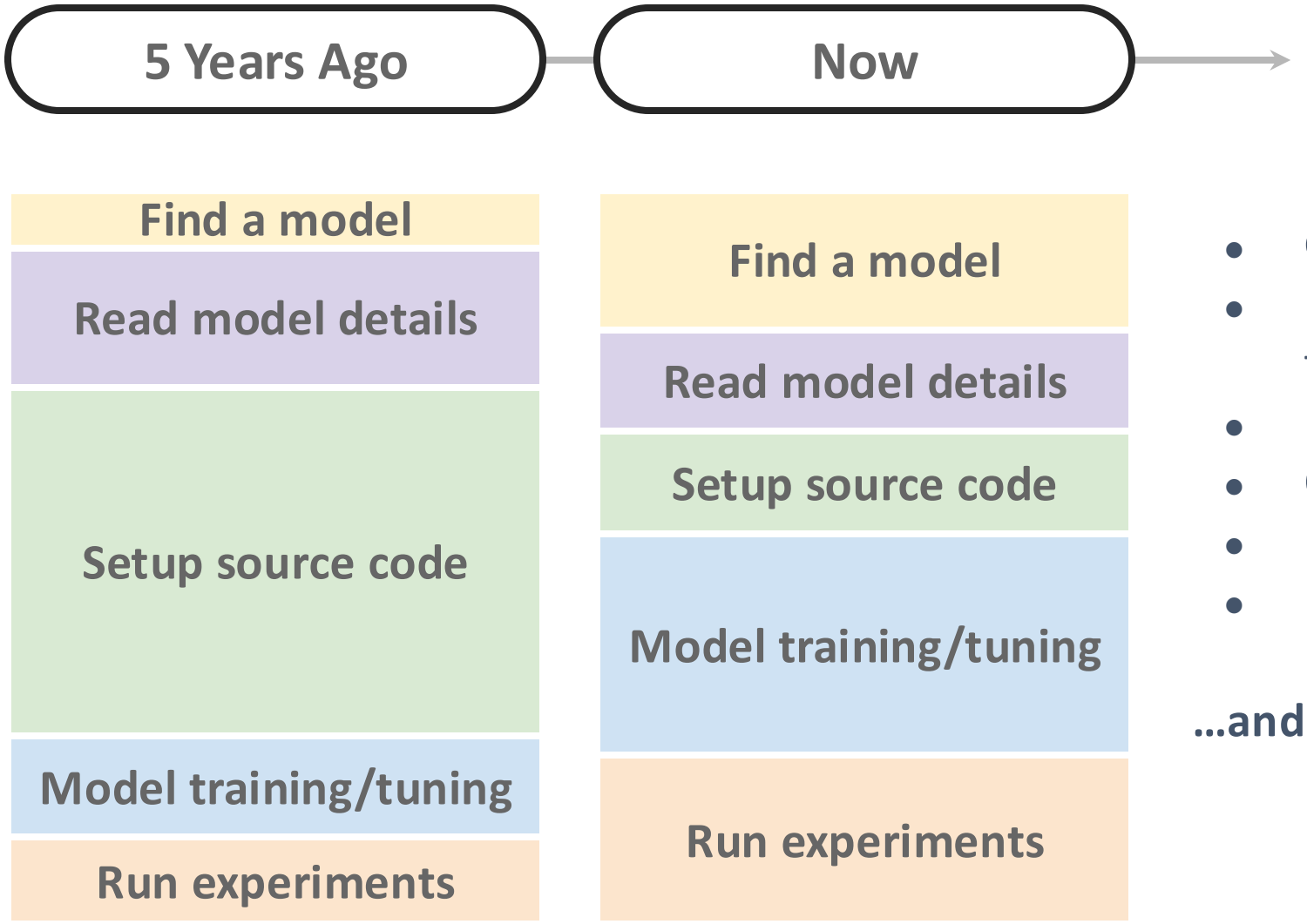
What changed?

- Growing diversity of approaches
- Less architecture revolution; more training technique revolution
- Hugging Face
- Open source & trained weights
- **Foundation models**



What changed?

- Growing diversity of approaches
- Less architecture revolution; more training technique revolution
- Hugging Face
- Open source & trained weights
- Foundation models
- **Prompting variants**



What changed?

- Growing diversity of approaches
- Less architecture revolution; more training technique revolution
- Hugging Face
- Open source & trained weights
- Foundation models
- Prompting variants

…and accumulated DL experience!

Future Challenges

1. Too many new methods and algorithms to follow
2. Frequent SOTA revolutions
3. Too many variants to experiment
4. Lack of computing resource

What changed?

- Growing diversity of approaches
- Less architecture revolution; more training technique revolution
- Hugging Face
- Open source & trained weights
- Foundation models
- Prompting variants

…and accumulated DL experience!

Future Challenges

1. Too many new methods and algorithms to follow
2. Frequent SOTA revolutions
3. Too many variants to experiment
4. Lack of computing resource

Gain more experience.

Know what you need, and learn what you can!

Neelima

Everley

Nick

Students’ questions?

Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ experiences


The End