

Next Generation AI: Towards Widespread Enterprise Adoption

dean@deanwampler.com
@deanwampler
Domino Data Lab



Let your data science team use the tools they love.

And bring them together in an enterprise-strength platform, that enables them to spend more time solving critical business problems.

[Learn More](#)[Read Our Customer Stories »](#)

dominodatalab.com

System-of-Record for Enterprise Data Science Teams

Accelerate Research

Get self-serve access to the latest tools and scalable compute. Reuse past work and iterate more efficiently.

[Learn More »](#)

Centralize Infrastructure

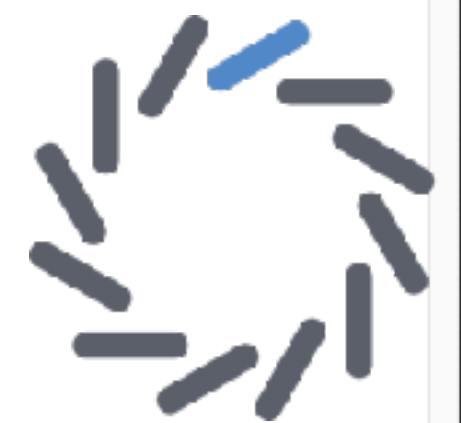
Manage the availability of powerful data science resources in a secure and governed system-of-record.

[Learn More »](#)

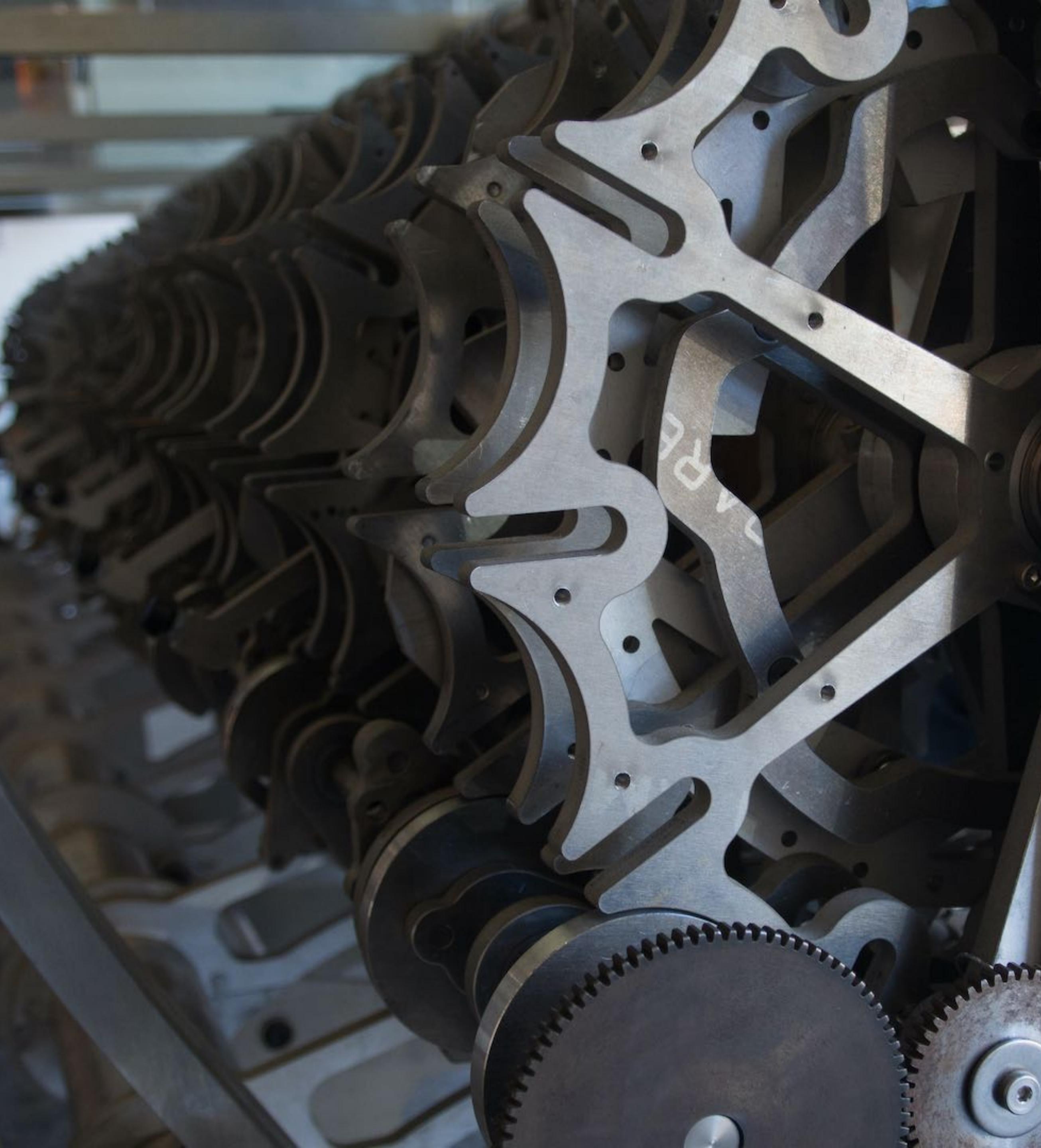
Deploy and Monitor Models

Expedite model consumption with apps, APIs and more – and ensure their

The screenshot shows a user interface for a data science platform. At the top, there's a navigation bar with tabs for 'Run', 'Jobs Timeline', 'Logs', 'Results', 'Details', 'Comments', and 'Resource Usage'. Below the navigation, there's a search bar and a download button. The main area displays several plots, including a line graph with red and yellow lines and a scatter plot with green dots. A table below the plots lists jobs with columns for 'No.', 'Title', 'Started', and 'AUC'. The first job listed is 'paramSearch.py -n 25 --loss exp'.



DOMINO



Outline

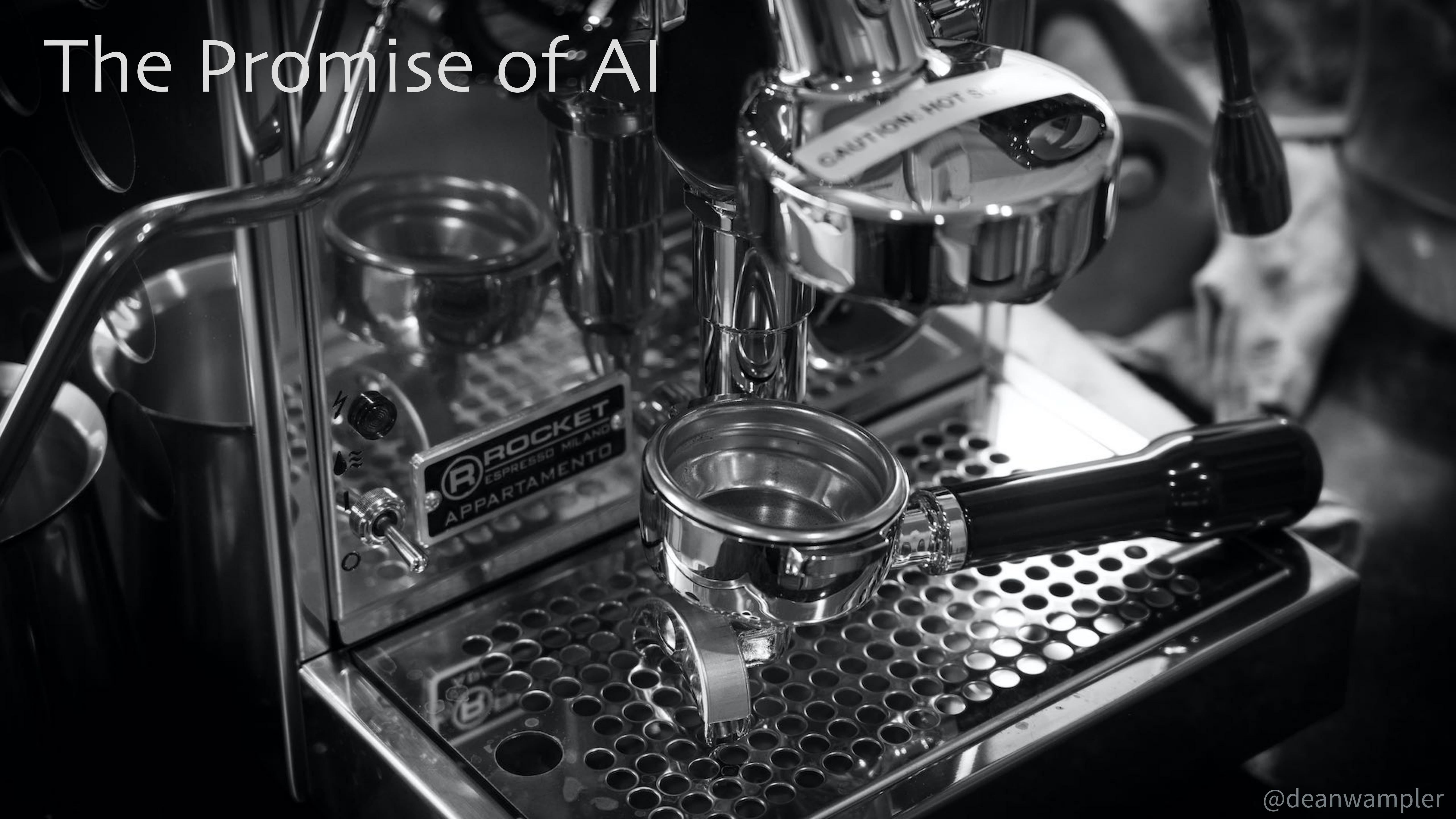
- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions



Outline

- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions

The Promise of AI



The Promise of AI

- Natural Language Processing
- Reinforcement Learning
- New applications of Deep Learning
- What Our Phones Are Telling Us...



Natural Language Processing



Applications

- Summarization
- Dialogues
- Naturalistic text to speech
- Translation
- Sentiment Analysis
- Fraud & Veracity Analysis
- Question Answering & Search



Summarization

- Legal documents
- Research papers
- News
- ...

Welcome to BBC.com

Vanguard Digital Advisor™
Customized financial guidance.
So you can build your future.

Take charge

+ Important information

Vanguard®
Digital Advisor

Sunday,

Trump says Biden won but again refuses to concede

Hamilton wins record seventh title



In “[Towards a Human-like Open-Domain Chatbot](#)”, we present Meena, a **2.6 billion parameter end-to-end trained neural conversational model**. We show that Meena can conduct conversations that are more sensible and specific than existing state-of-the-art chatbots.

- # Dialogs
- Chatbots
 - Human-computer dialogs

The screenshot shows a web browser displaying the Google AI Blog article titled "Towards a Conversational Agent that Can Chat About Anything". The article was posted on Tuesday, January 28, 2020, by Daniel Adiwardana and Thang Luong from Google Research, Brain Team. The text discusses the limitations of current chatbots, which are highly specialized and lack common sense, and how open-domain dialog research explores a complementary approach to create agents that are specialized but can still chat about virtually anything. The article also mentions that such agents could lead to better humanizing computer interactions, improving interactive movie and videogame characters. A red box highlights the first sentence of the text.

Google AI Blog
The latest news from Google AI

Towards a Conversational Agent that Can Chat About Anything

Tuesday, January 28, 2020

Posted by Daniel Adiwardana, Senior Research Engineer, and Thang Luong, Senior Research Scientist, Google Research, Brain Team

Modern conversational agents (chatbots) tend to be highly specialized – they perform well as long as users don't stray too far from their expected usage. To better handle a wide variety of conversational topics, open-domain dialog research explores a complementary approach to the one used by most existing systems. These agents are often very specialized but can still chat about virtually anything. This is a critical flaw – they often don't make sense. They may not have common sense or basic knowledge about the world. Moreover, chatbots often give responses that are not specific to the current context. For example, “I don't know.” is a sensible response to any question, but it's not specific. Current chatbots do this much more often than people because it covers many possible user inputs.

In “[Towards a Human-like Open-Domain Chatbot](#)”, we present Meena, a **2.6 billion parameter end-to-end trained neural conversational model**. We show that Meena can conduct conversations that are more sensible and specific than existing state-of-the-art chatbots. Such improvements are

A close-up photograph of a high-end espresso machine. The machine is made of polished stainless steel, reflecting the warm lighting of the environment. A black handle with a silver band is attached to a portafilter. A small white sticker on the handle reads "CAUTION: HOT SURFACE". In the background, a circular gauge or thermometer is visible on the machine's front panel. The overall aesthetic is clean and professional.

Naturalistic text to speech

- Needed for dialog generation



Translation

- Domain-specific languages
 - Medicine
 - Air traffic control
 - ...
 - “Rare” languages

A close-up photograph of a high-end espresso machine. The machine is made of polished stainless steel, reflecting the warm lighting of the environment. A black handle with a silver band is attached to the side. A small white sticker on the handle reads "CAUTION: HOT SURFACE". The machine has a circular portafilter holder and a steam wand. The background is blurred, showing a warm, golden light.

Sentiment Analysis

- Customer support
- Social media
- Public relations



Fraud & Veracity Analysis

- “Fake news”
- Better SPAM, Phishing, etc.
detection and mitigation.

The screenshot shows a PDF document titled "Fake News Detection on Social Media: A Data Mining Perspective" by Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. The document is from the KDD conference. The abstract discusses the use of social media for news consumption and the challenges of detecting fake news. The page number is 1 / 15.

**Fake News Detection on Social Media:
A Data Mining Perspective**

Kai Shu[†], Amy Sliva[‡], Suhang Wang[†], Jiliang Tang[‡], and Huan Liu[†]
[†]Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
[‡]Charles River Analytics, Cambridge, MA, USA
[‡]Computer Science & Engineering, Michigan State University, East Lansing, MI, USA
[†]{kai.shu,suhang.wang,huan.liu}@asu.edu,
[‡]asliva@cra.com, [‡]tangjili@msu.edu

ABSTRACT

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. On the other hand, it enables the

on, and discuss the news with friends or other readers on social media. For example, 62 percent of U.S. adults get news on social media in 2016, while in 2012, only 49 percent reported seeing news on social media¹. It was also found that social media now outperforms television as the



Question Answering & Search

- Customer support
- More advanced, targeted search results
- Support natural language queries
- Search legal docs, research papers, patents, ...



Images and Videos...

- Many of these same techniques and applications apply to image and video applications, too.

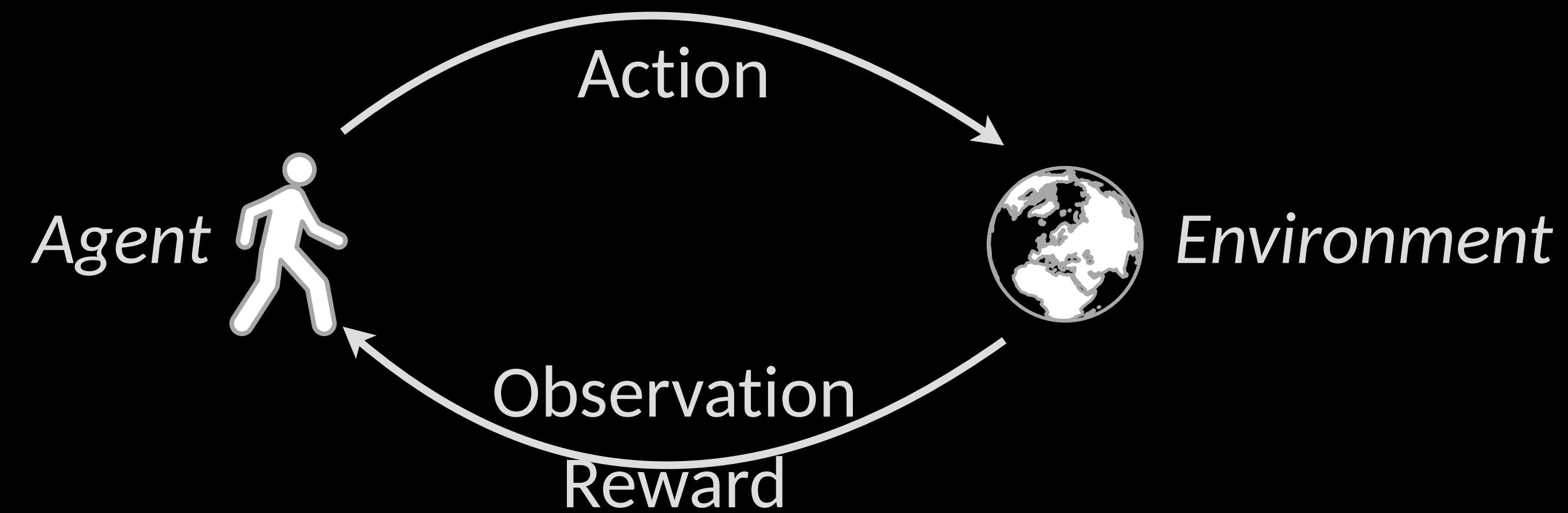


Reinforcement Learning



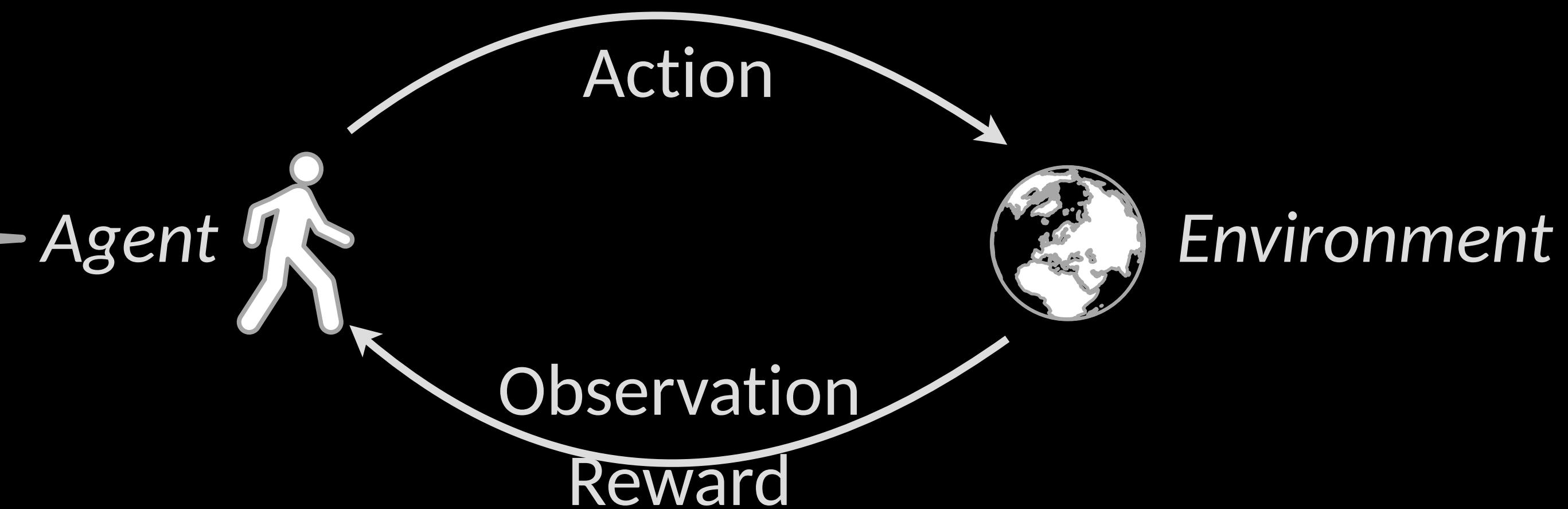
What Is RL?

- An agent observes an environment, takes a sequence of actions
- Goal: maximize the cumulative reward



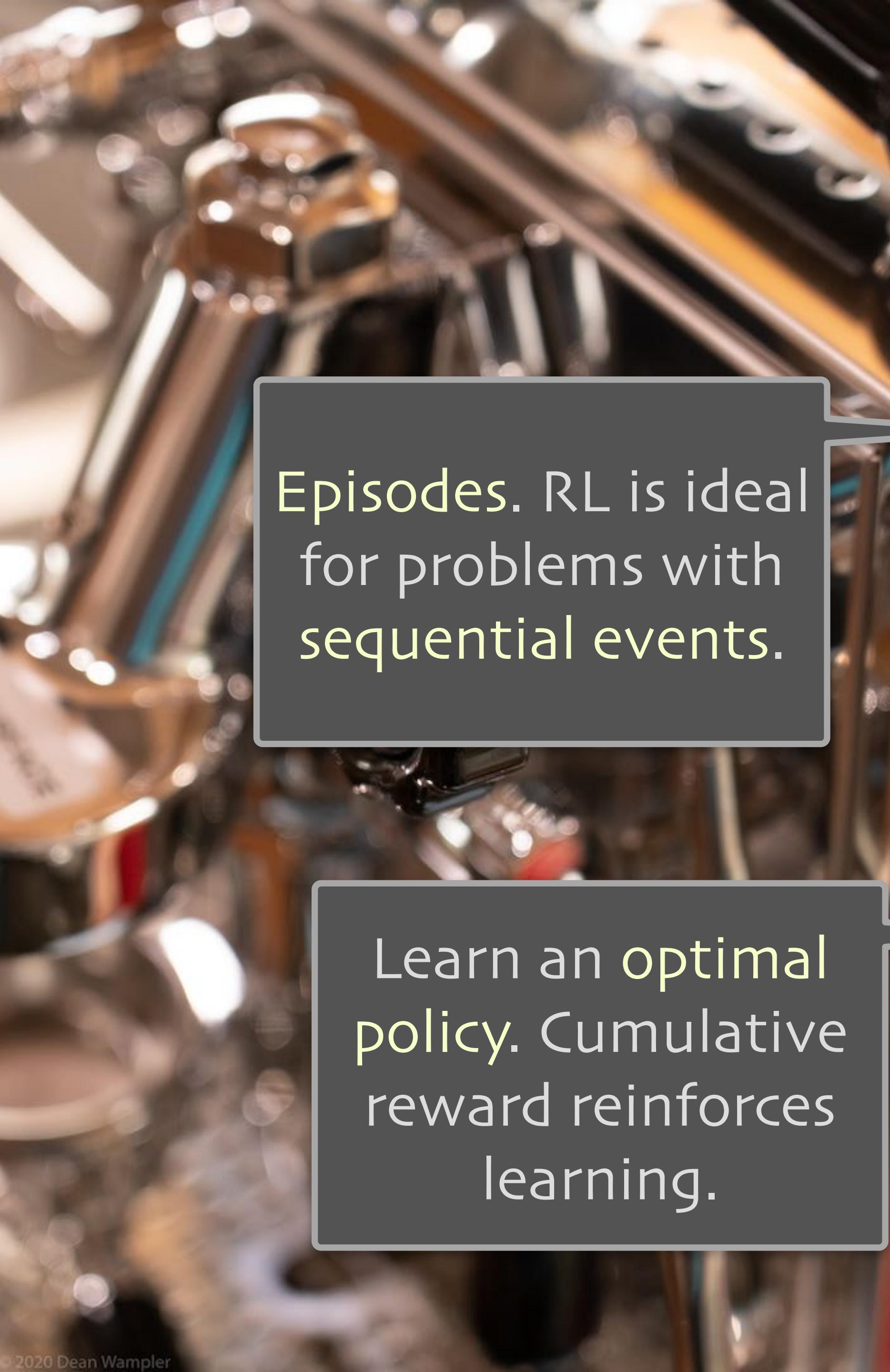
What Is RL?

- An agent observes an environment, takes a sequence of actions
- Goal: maximize the cumulative reward



Episodes. RL is ideal for problems with sequential events.

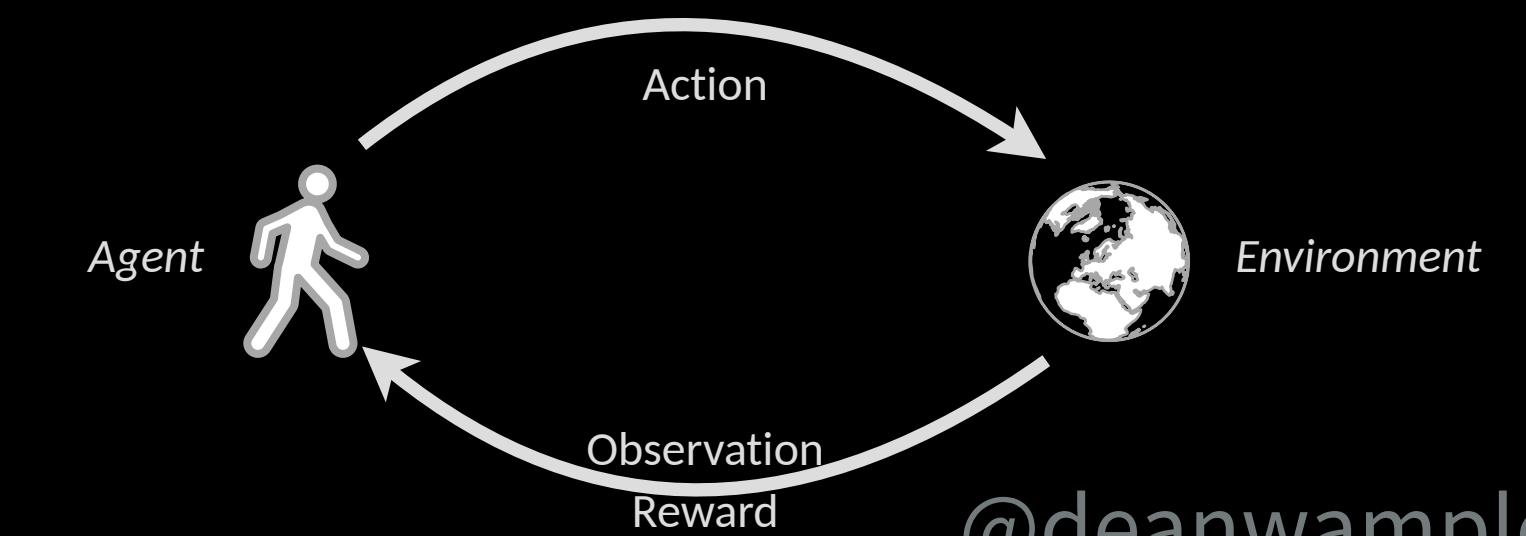
Learn an optimal policy. Cumulative reward reinforces learning.





Applications

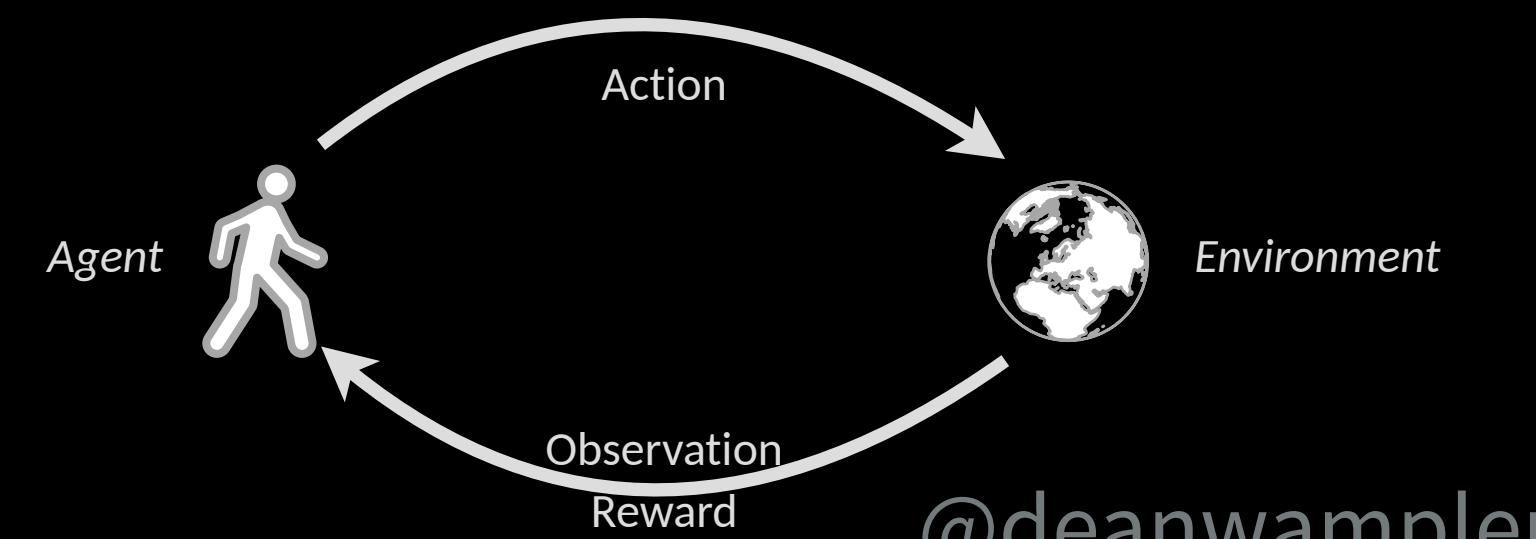
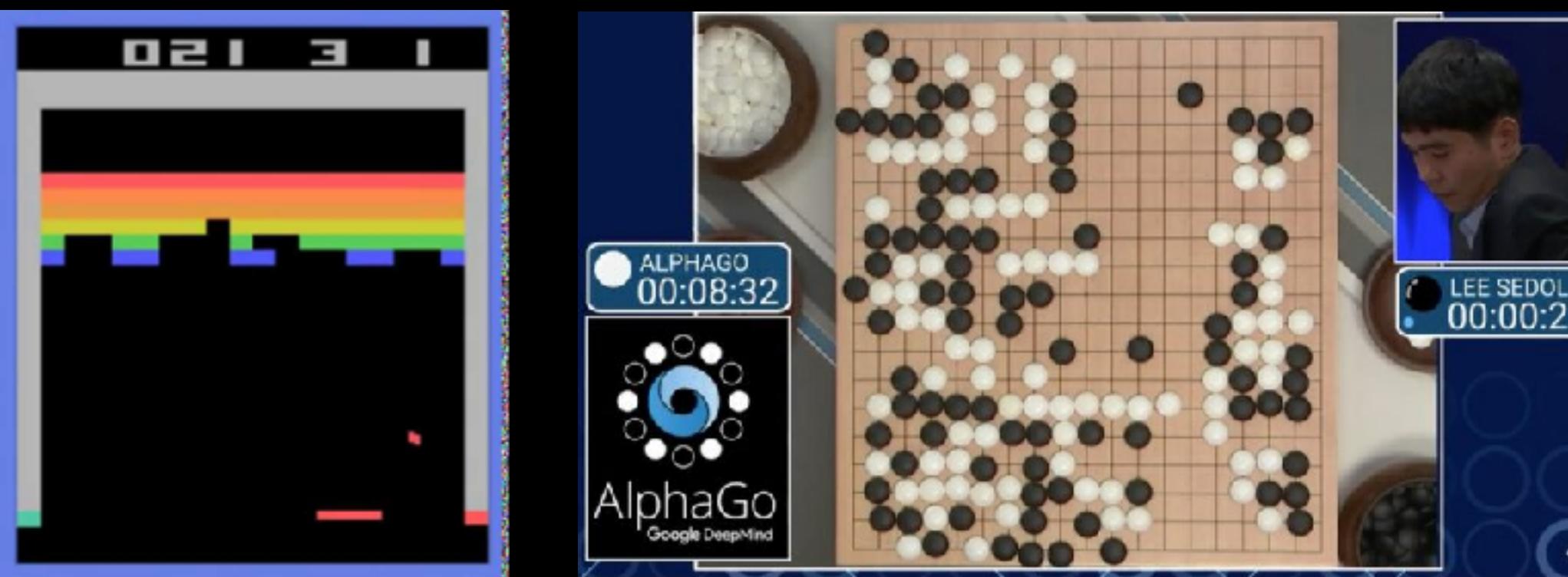
- Games
- Robots & Autonomous Vehicles
- Process Modeling & Automation
- System Optimization
- Advertising & Recommendation
- Markets





Games

- World's best expert game
play in:

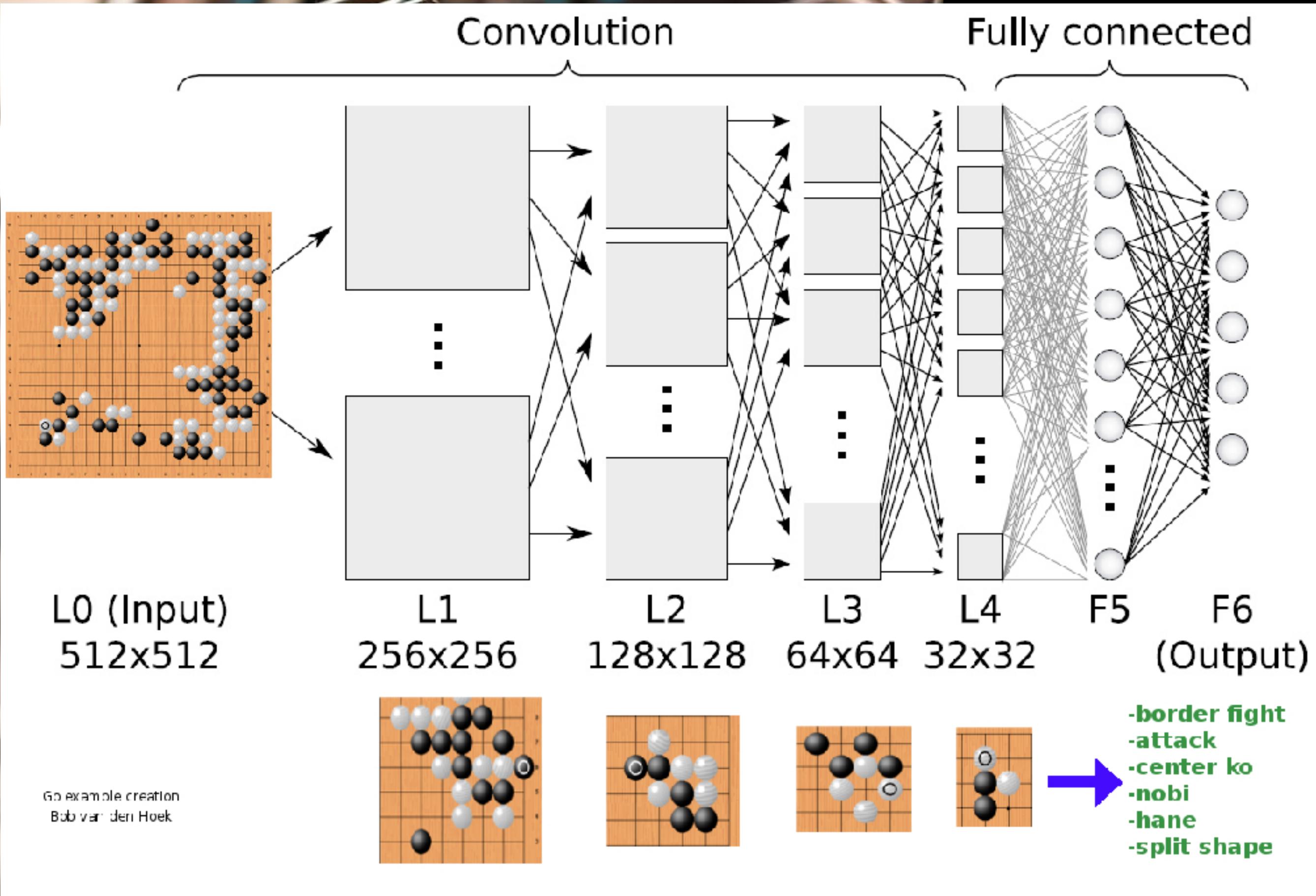


<https://www.geekwire.com/2016/alphago-ai-program-wins-1-million-prize-go-showdown-champion-lee-sedol/>

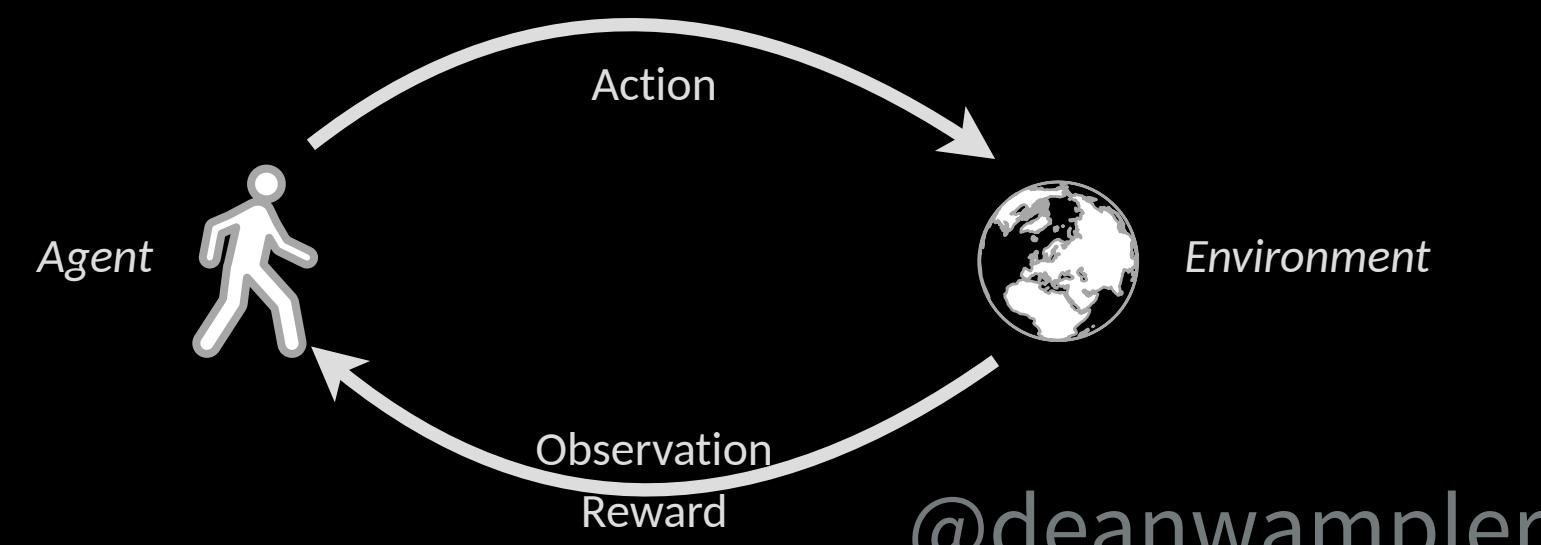
<https://towardsdatascience.com/tutorial-double-deep-q-learning-with-dueling-network-architectures-4c1b3fb7f756>

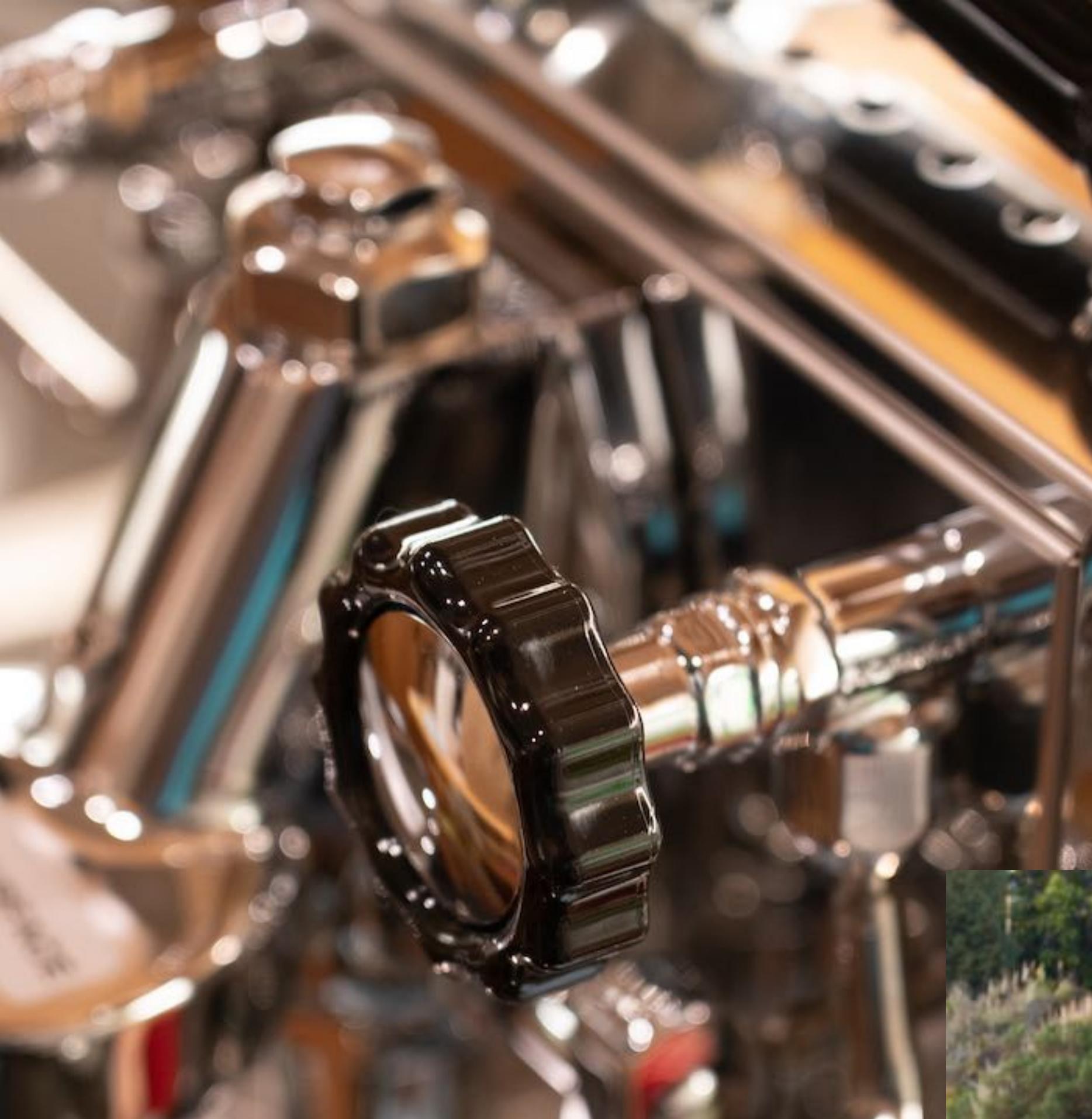


Games



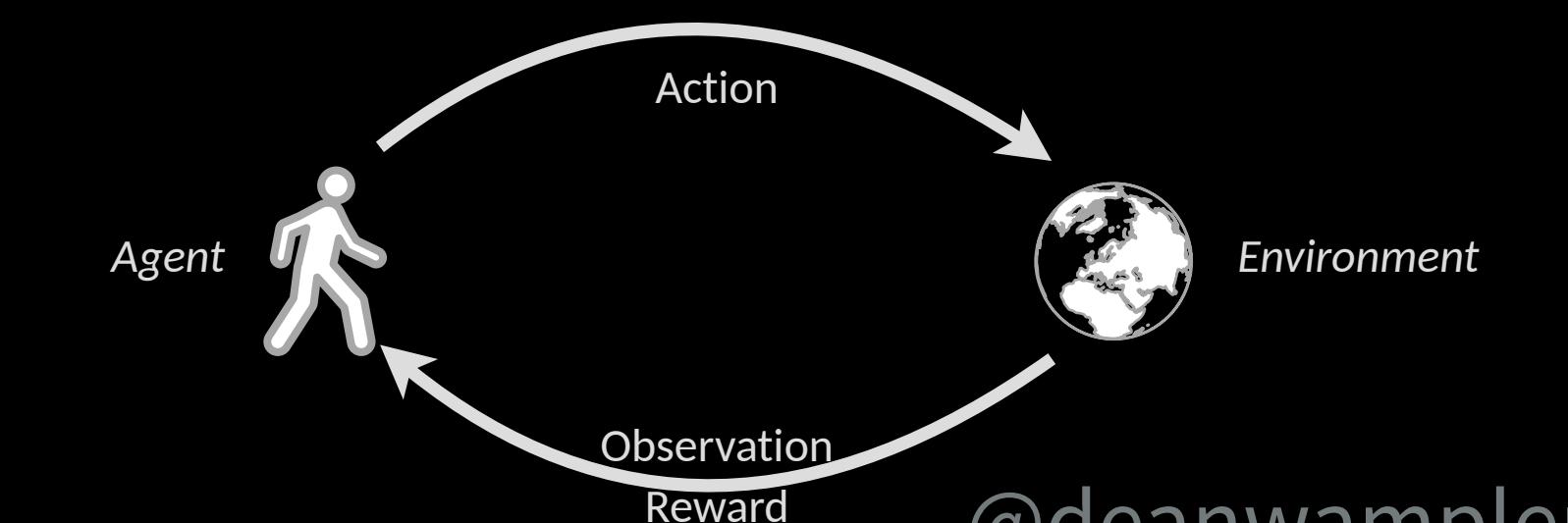
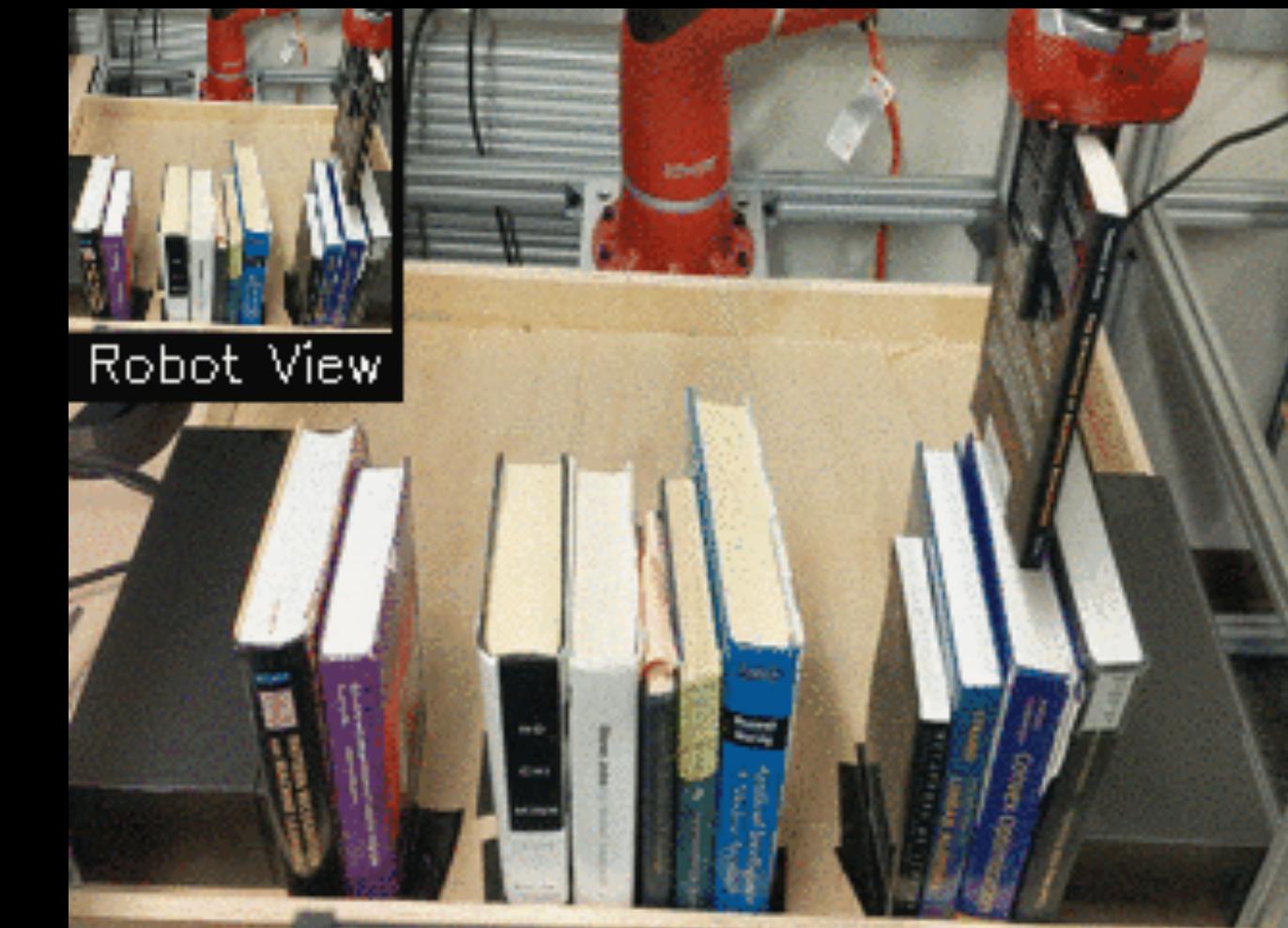
- AlphaGo
- Observations: board state
- Actions: place stones
- Rewards:
 - 1 if you win
 - 0 otherwise



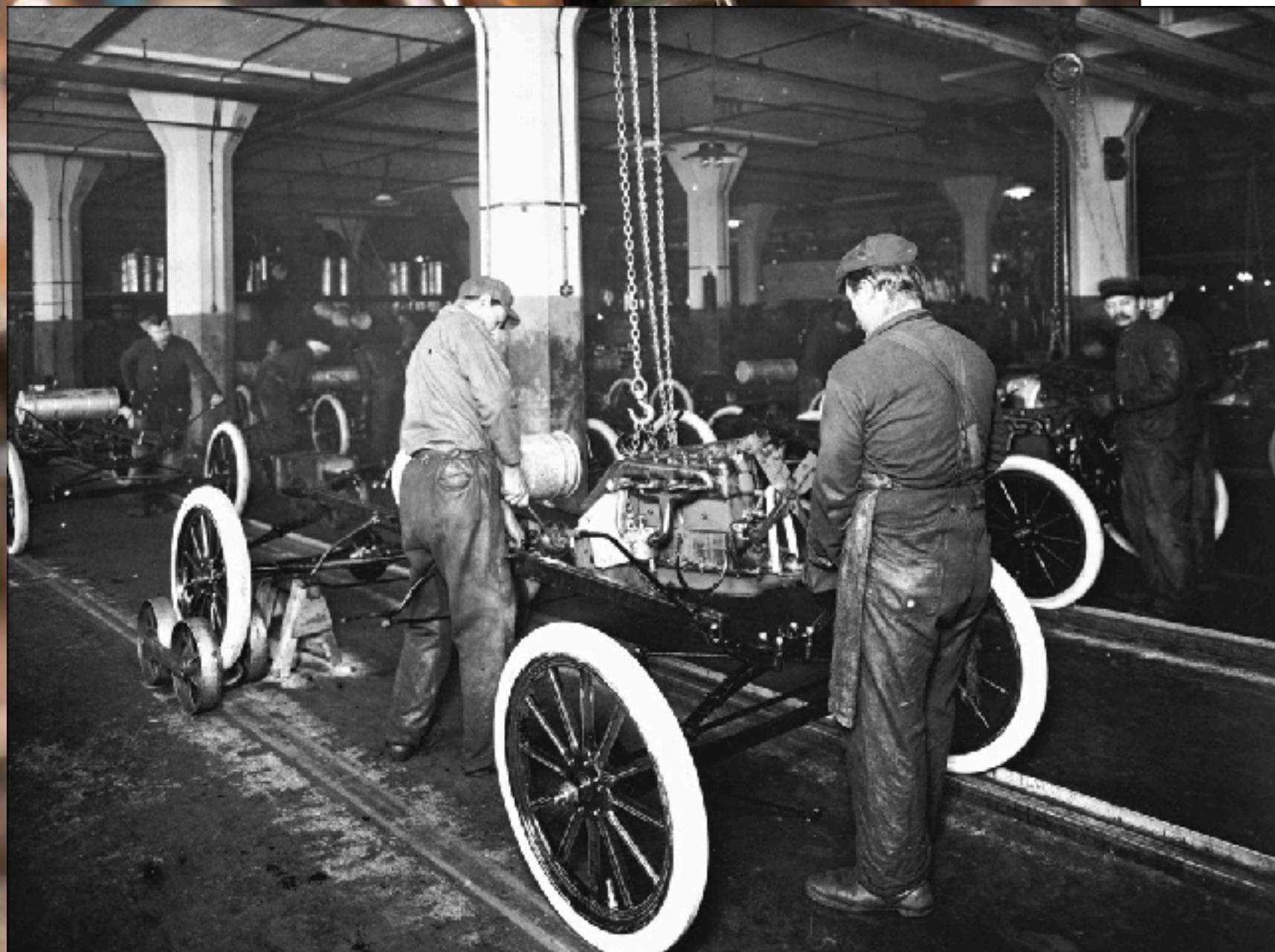


Robotics & Autonomous Vehicles

- Start with simulators, work up to real machines.



Process Modeling & Automation



Simulation Optimization | Add / x +

pathmind.com

Products Services Industries Resources Company SIGN IN REQUEST DEMO

Recent Updates



Engineering Group: Manufacturing Optimization with AI Minimizes Factory Flow Bottlenecks

Oct 30, 2020 | [Customer Success](#)

Summary Engineering Group, a global engineering firm and technology consultancy with a strong practice in simulation, worked with Pathmind to apply reinforcement learning to intelligently route heavy industrial parts over a complex assembly line in...



Engineering Group: Using AI to Maximize Factory Output with Better Order Sequencing

Oct 29, 2020 | [Customer Success](#)

Summary Engineering Group, a global engineering firm and technology consultancy with a strong practice in simulation, worked with Pathmind to apply reinforcement learning to maximize factory output by making smarter decisions about order...



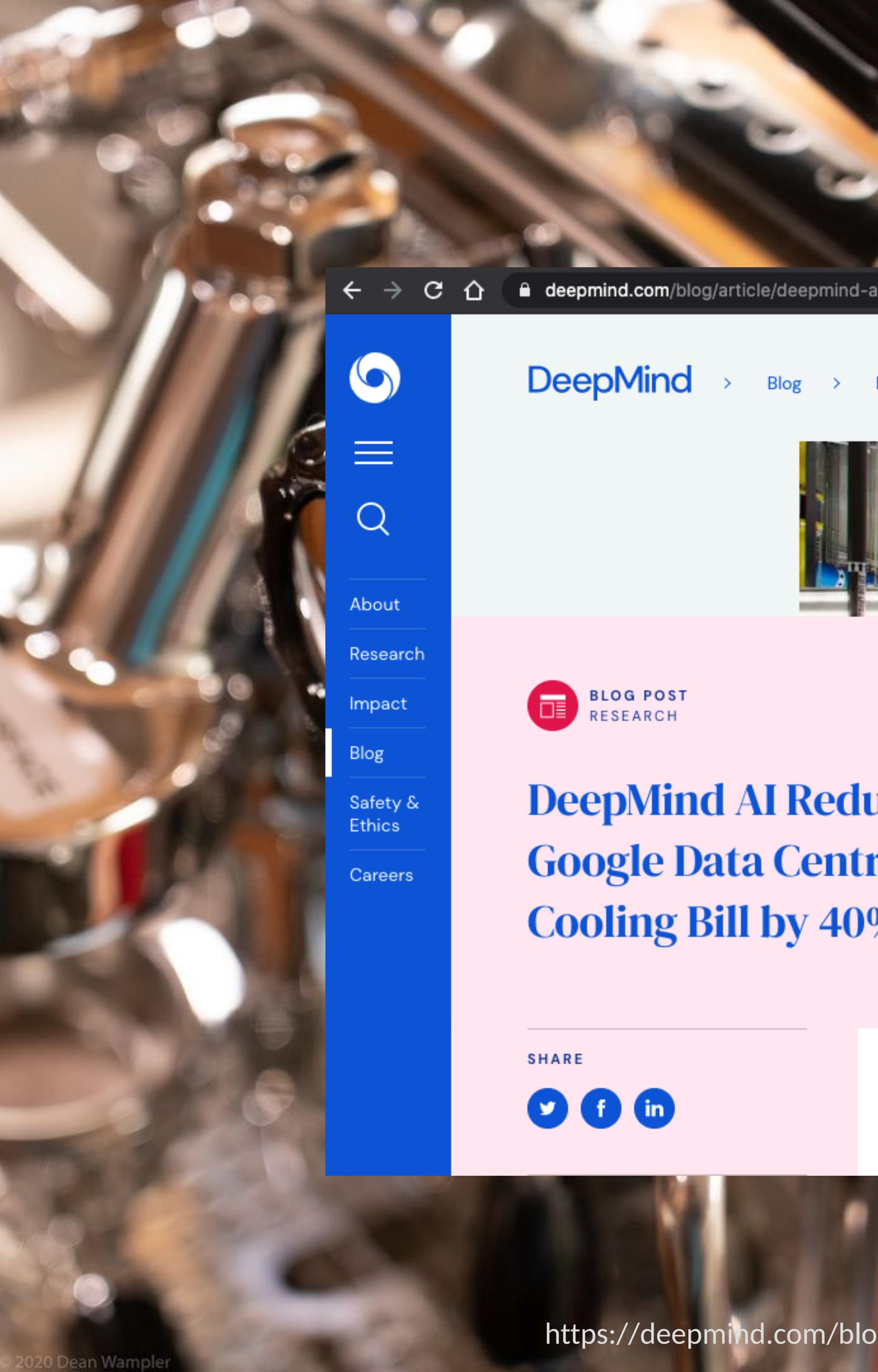
Princeton Consultants: Using AI to Maximize Efficiency of Machine Scheduling

Oct 13, 2020 | [Customer Success](#)

Summary Princeton Consultants, a simulation consulting firm, serves a manufacturing client with a hard machine scheduling problem. Its optimizer had difficulty scheduling machines for new types of items that needed to be processed; it was not able...



System Optimization



A screenshot of a DeepMind blog post titled "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%". The post is dated 20 JUL 2016 and is categorized under "BLOG POST RESEARCH". The main image shows a long corridor in a data center with numerous colorful pipes and machinery. The URL of the post is visible at the bottom.

DeepMind > Blog > DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

BLOG POST
RESEARCH

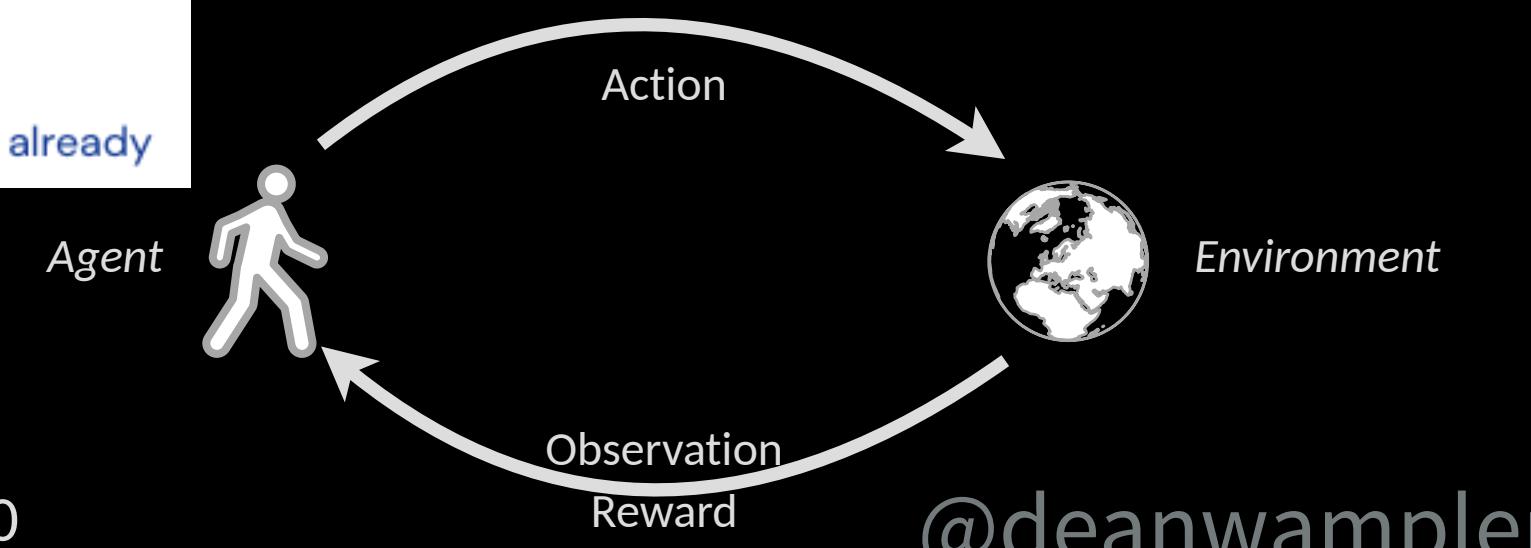
20 JUL 2016

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

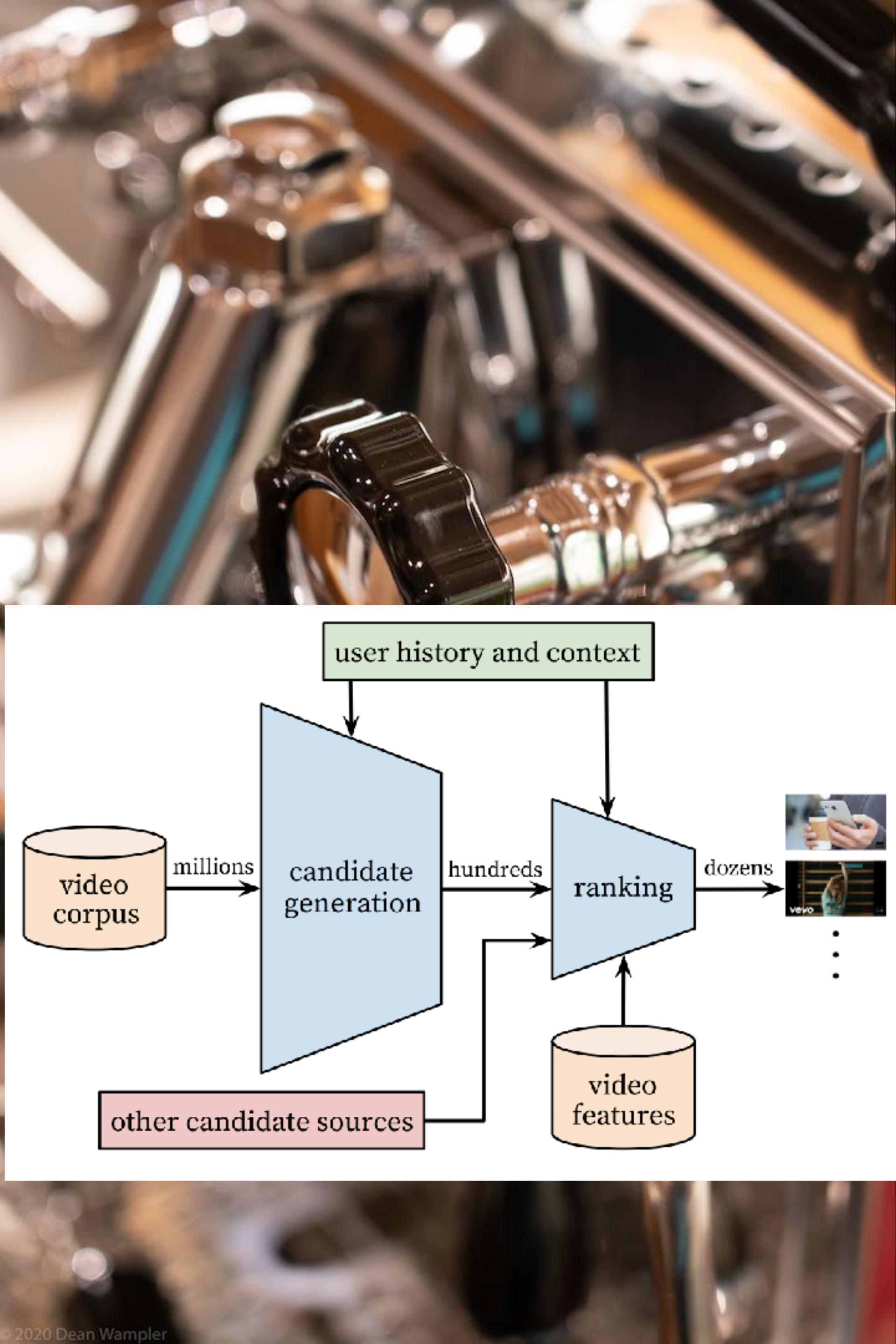
SHARE

From smartphone assistants to image recognition and translation, machine learning already

<https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40>

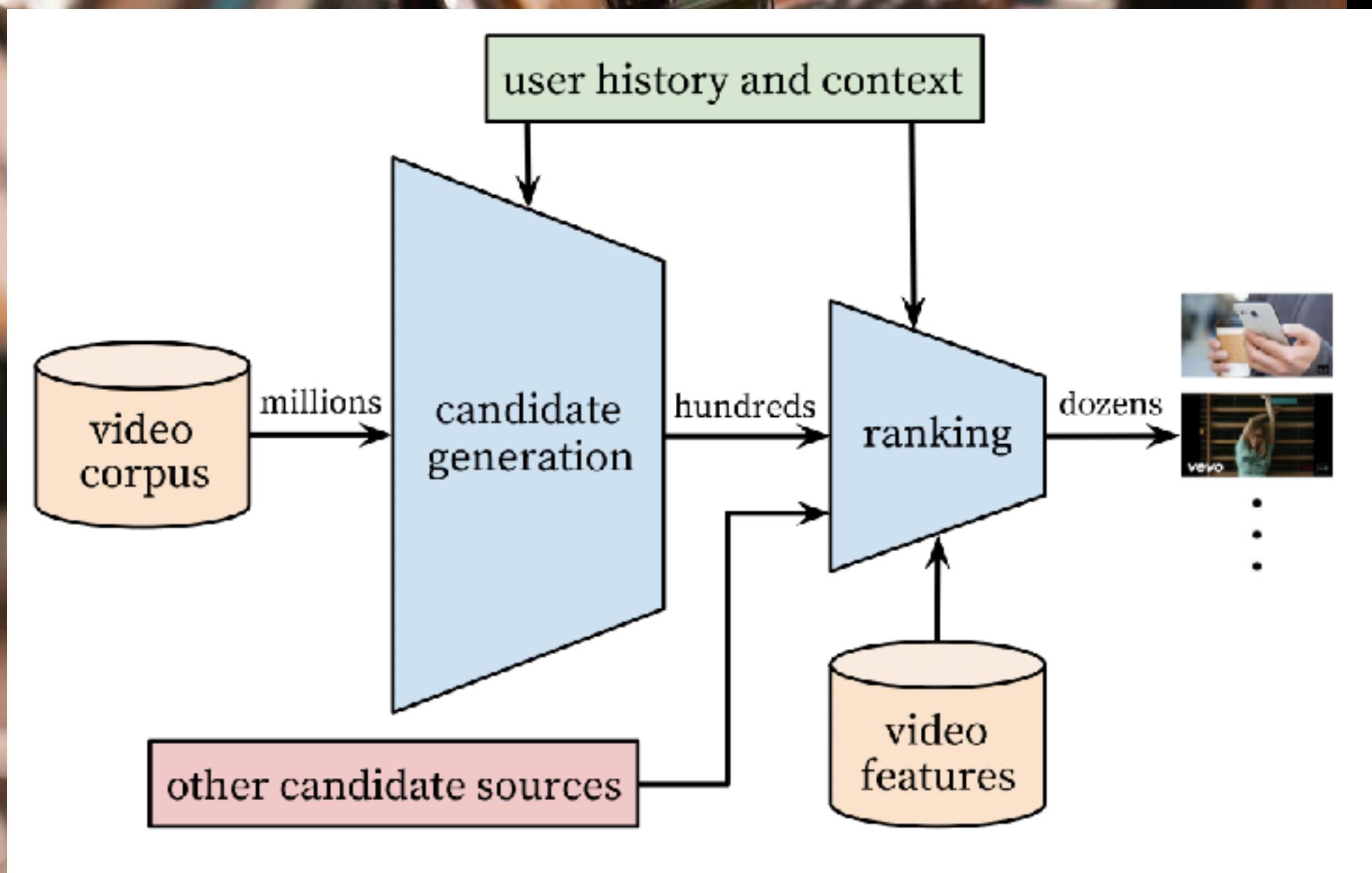


@deanwampler

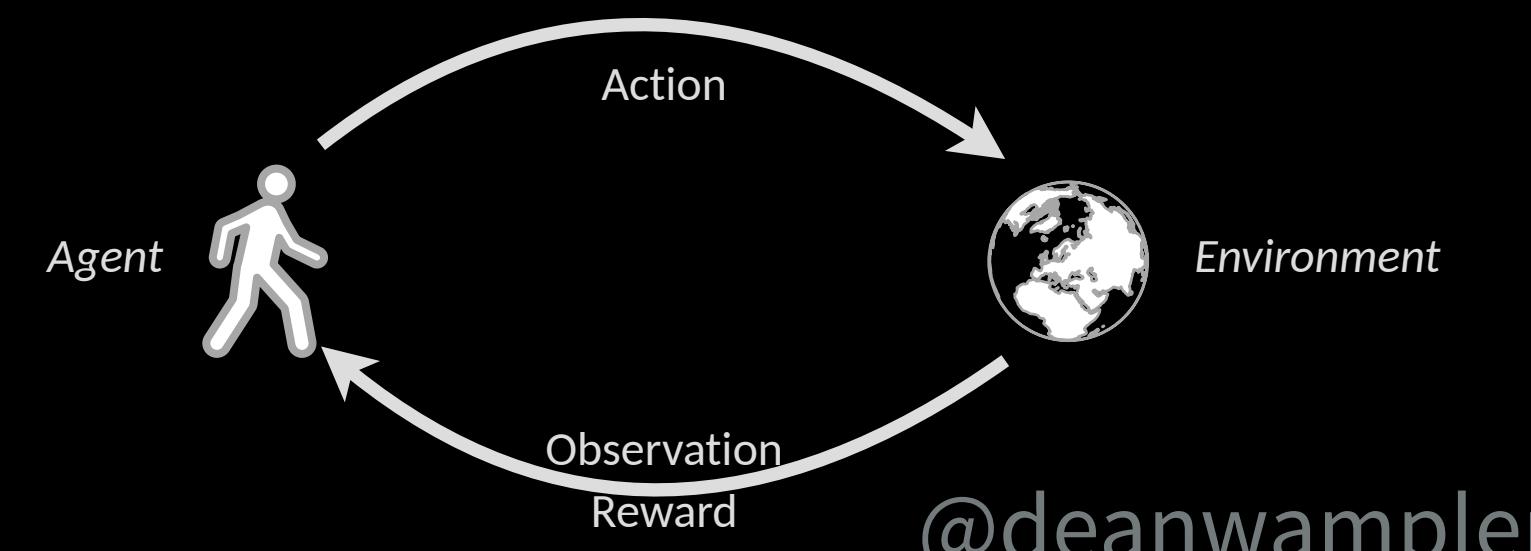


Advertising & Recommendation

- A “mature” problem, yet RL is providing a new approach.
- Better modeling of evolving preferences.
- Better scalability than collaborative filtering, etc.



<https://research.google/pubs/pub45530/>

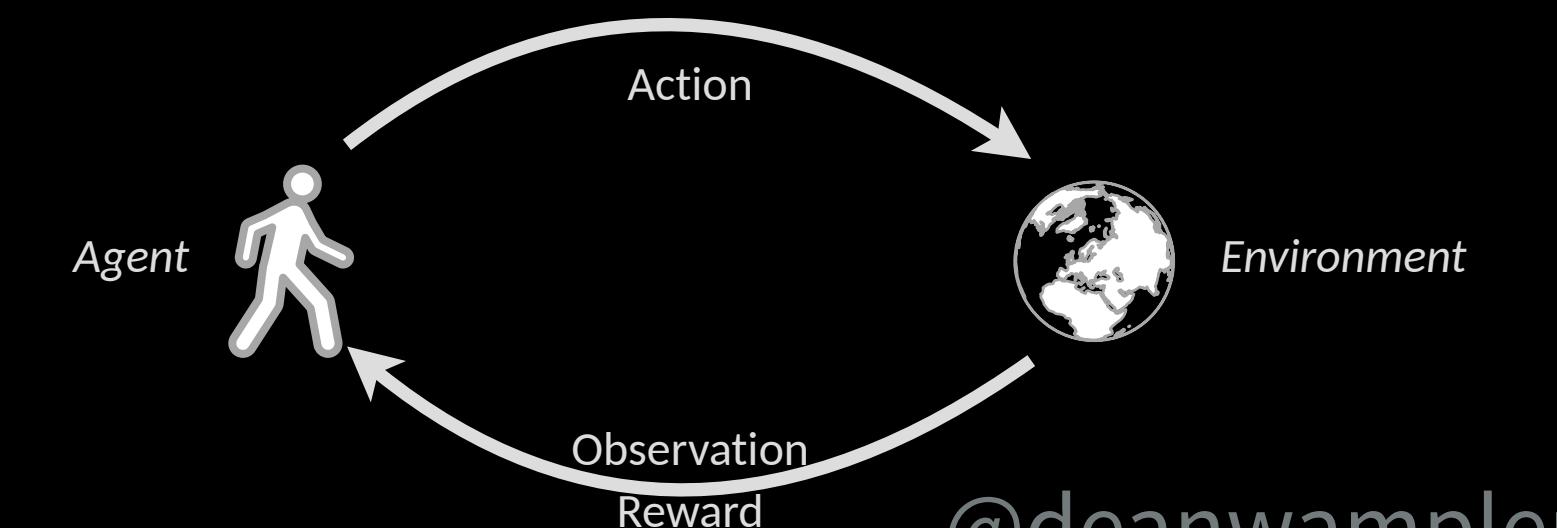


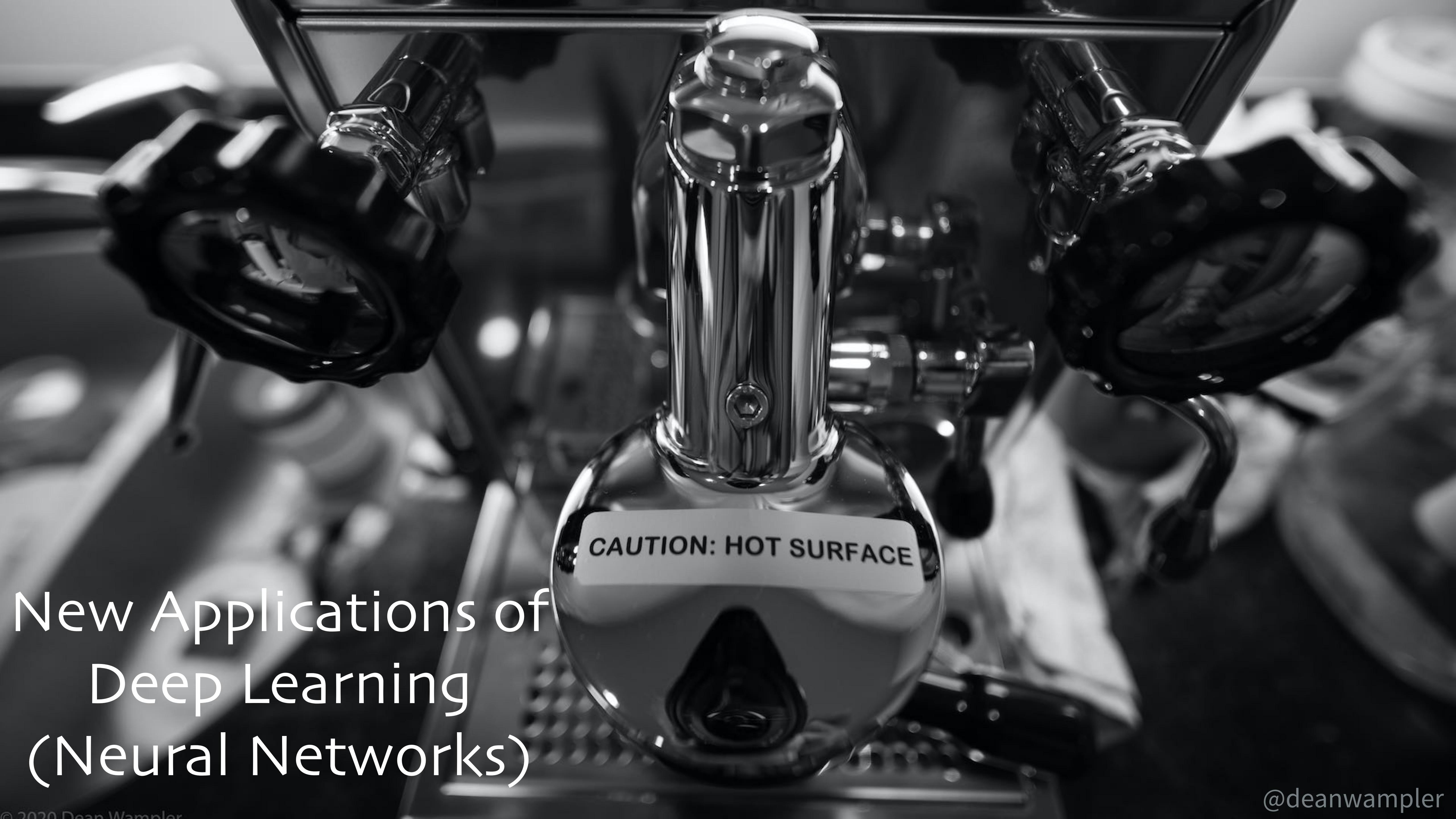
@deanwampler



Markets

- Inherently time-ordered
- Lots of different “signals”
- Contextual, multi-armed bandits



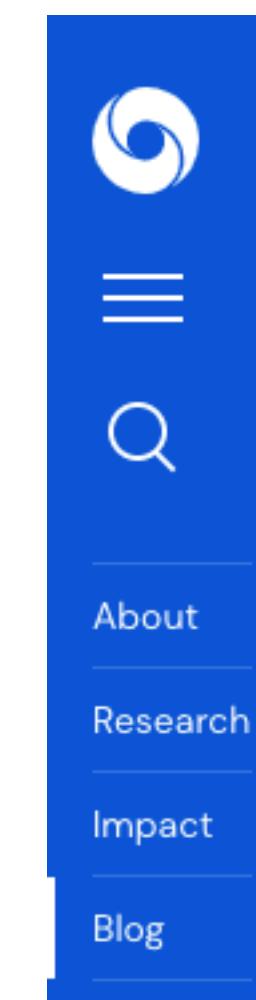


New Applications of Deep Learning (Neural Networks)

@deanwampler

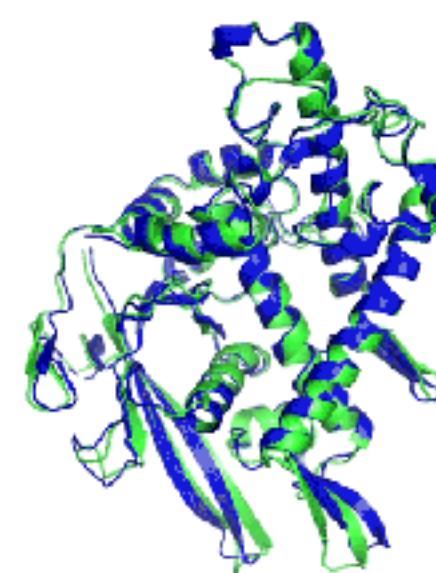


Biology, Medicine

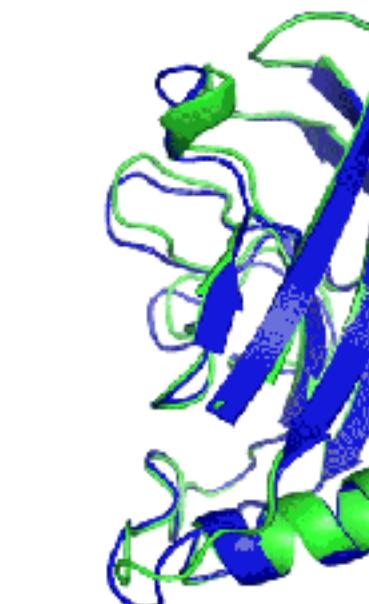


30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

- Experimental result
- Computational prediction



Biology, Medicine

nature reviews cancer

[View all |](#)

[Explore our content ▾](#) [Journal information ▾](#)

[nature](#) > [nature reviews cancer](#) > [perspectives](#) > [article](#)

Perspective | Published: 17 May 2018

OPINION

Artificial intelligence in radiology

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz &
Hugo J. W. L. Aerts [✉](#)

[Nature Reviews Cancer](#) **18**, 500–510(2018) | [Cite this article](#)

15k Accesses | **317** Citations | **311** Altmetric | [Metrics](#)

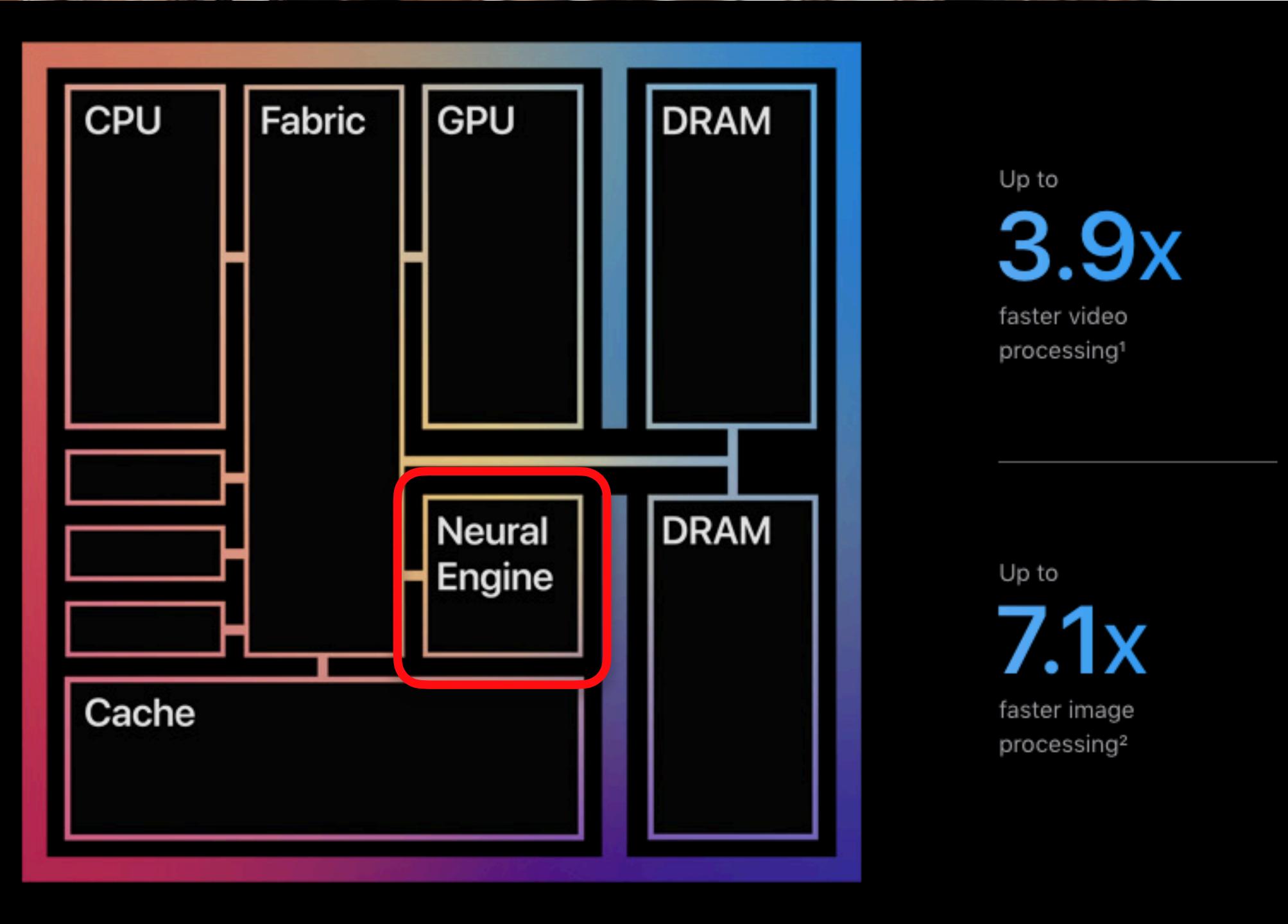
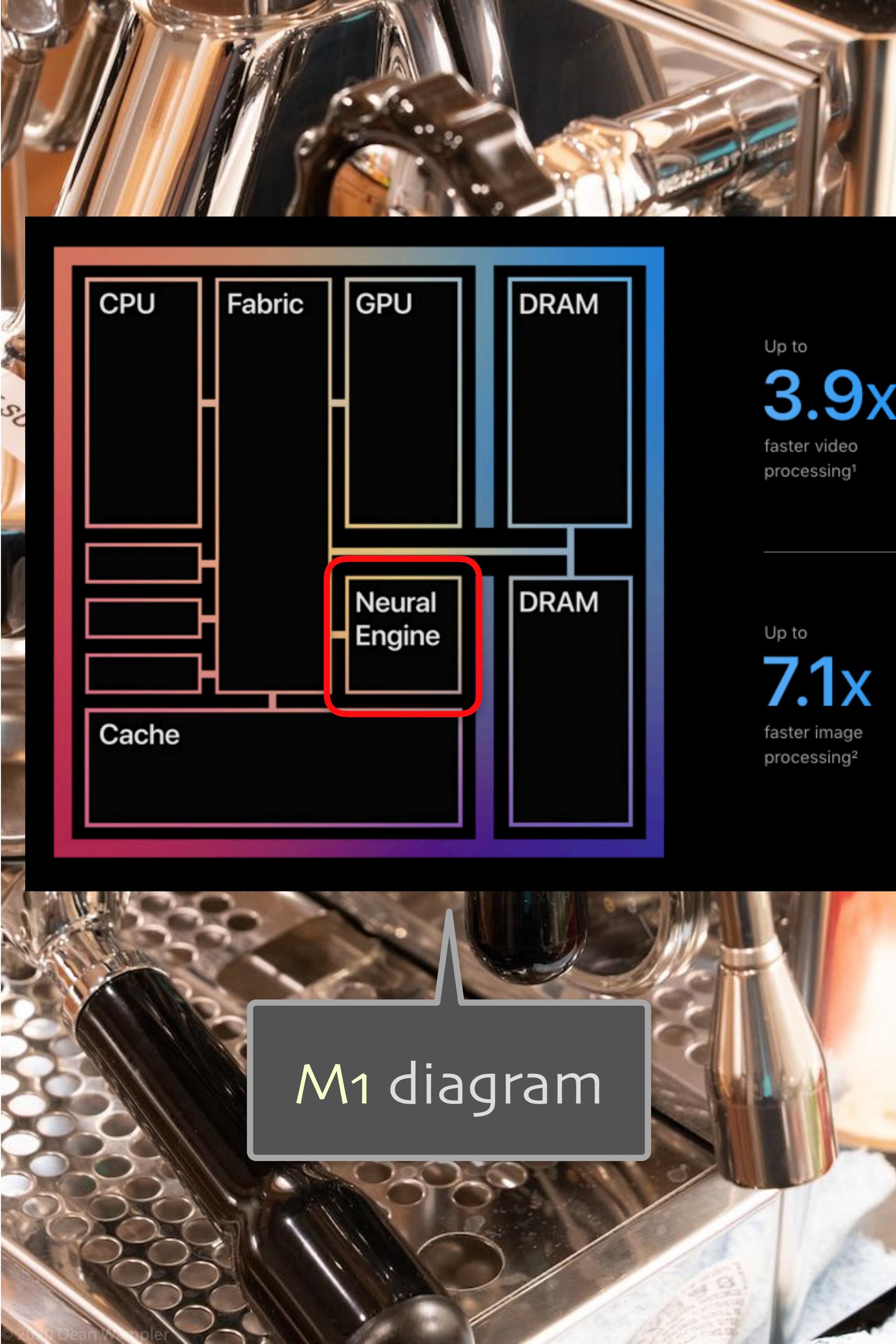
Abstract

Artificial intelligence (AI) algorithms, particularly deep learning, have demonstrated remarkable progress in image-recognition tasks. Methods ranging from convolutional neural networks to variational autoencoders have found myriad applications in the medical image analysis field, propelling it forward at a rapid pace. Historically, in radiology practice, trained physicians visually assessed medical images for the detection



What Our Phones
Are Telling Us...

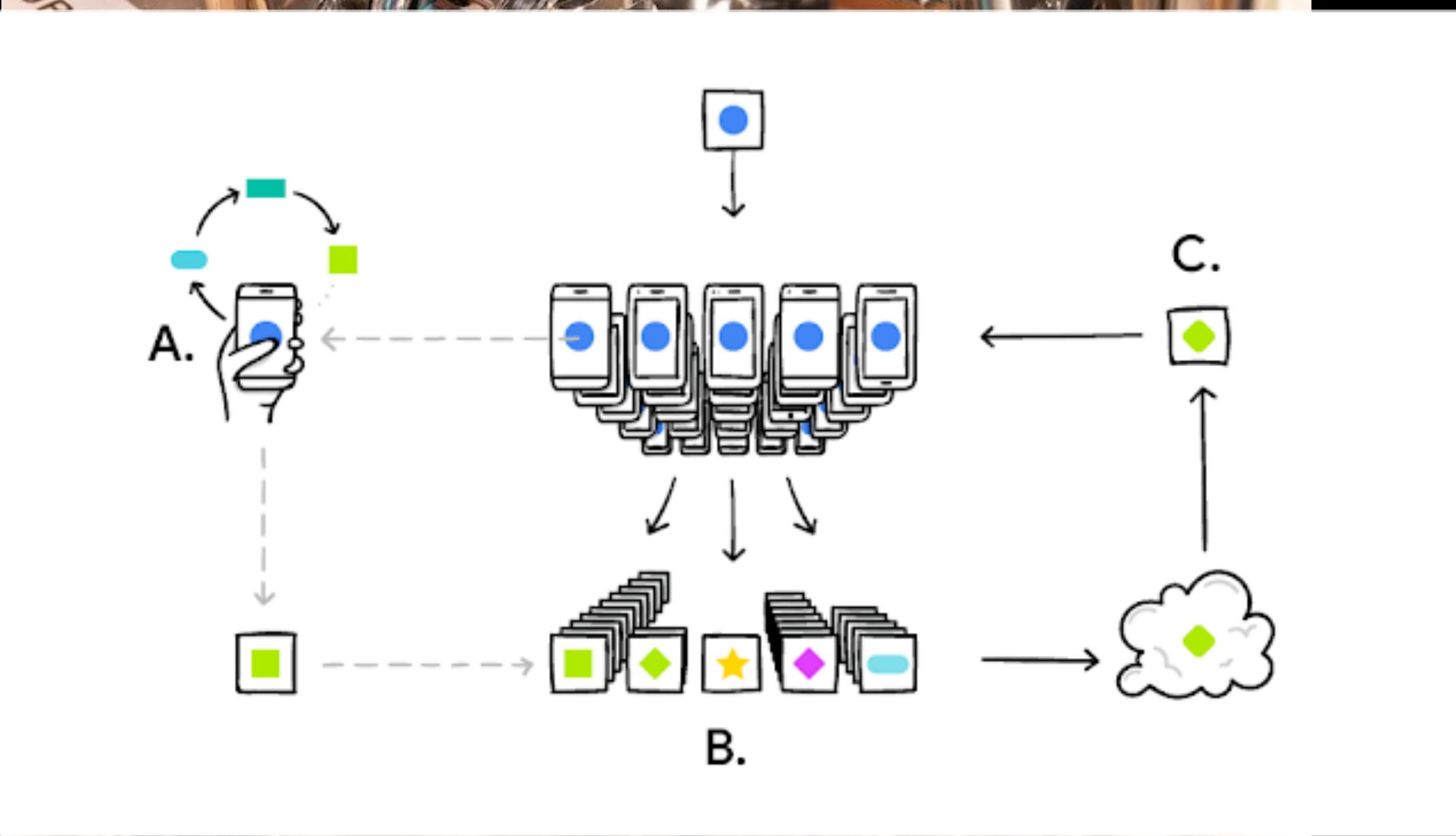
Apple Silicon





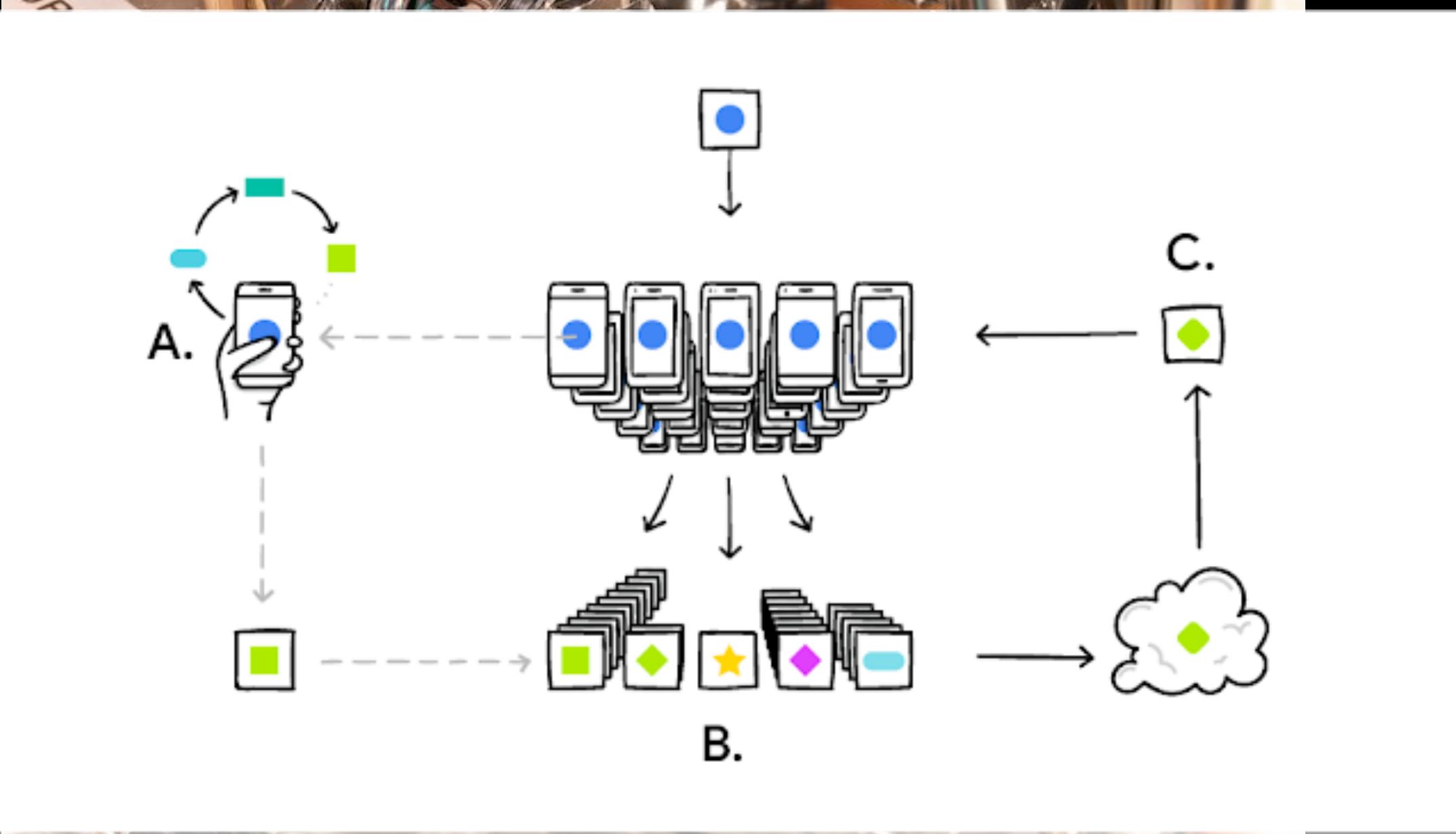
Applications that Exploit ML/AI

- Unlocking: finger and face ID
- Predictive typing
- Voice assist - Siri
- Health monitoring
- Security (beyond passwords...)
- Recommendations
- ...
- Probably most apps will use it in one way or another, eventually!



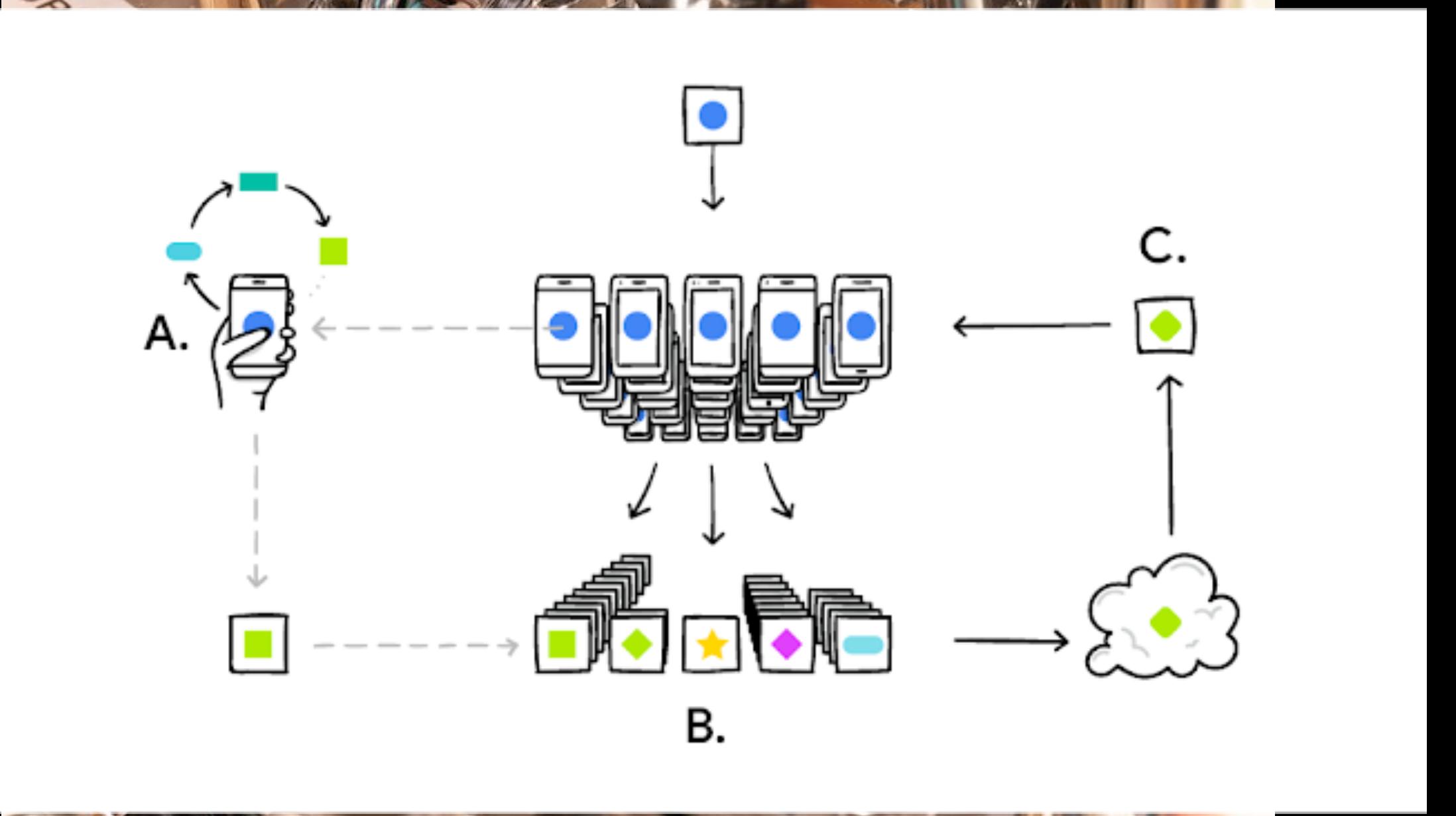
Technologies that Make this Possible

- Federated Learning
 - A. Your local usage trains a model
 - B. Model updates from many users are aggregated to form a consensus update
 - C. Updated model propagated to all users.
 - D. Repeat...



Technologies that Make this Possible

- Federated Learning Advantages
 - Your private data stays local
 - Local model is fine tuned for you
 - Less central data storage required
 - Central processing is minimized
 - Instead, our phones do most of the training



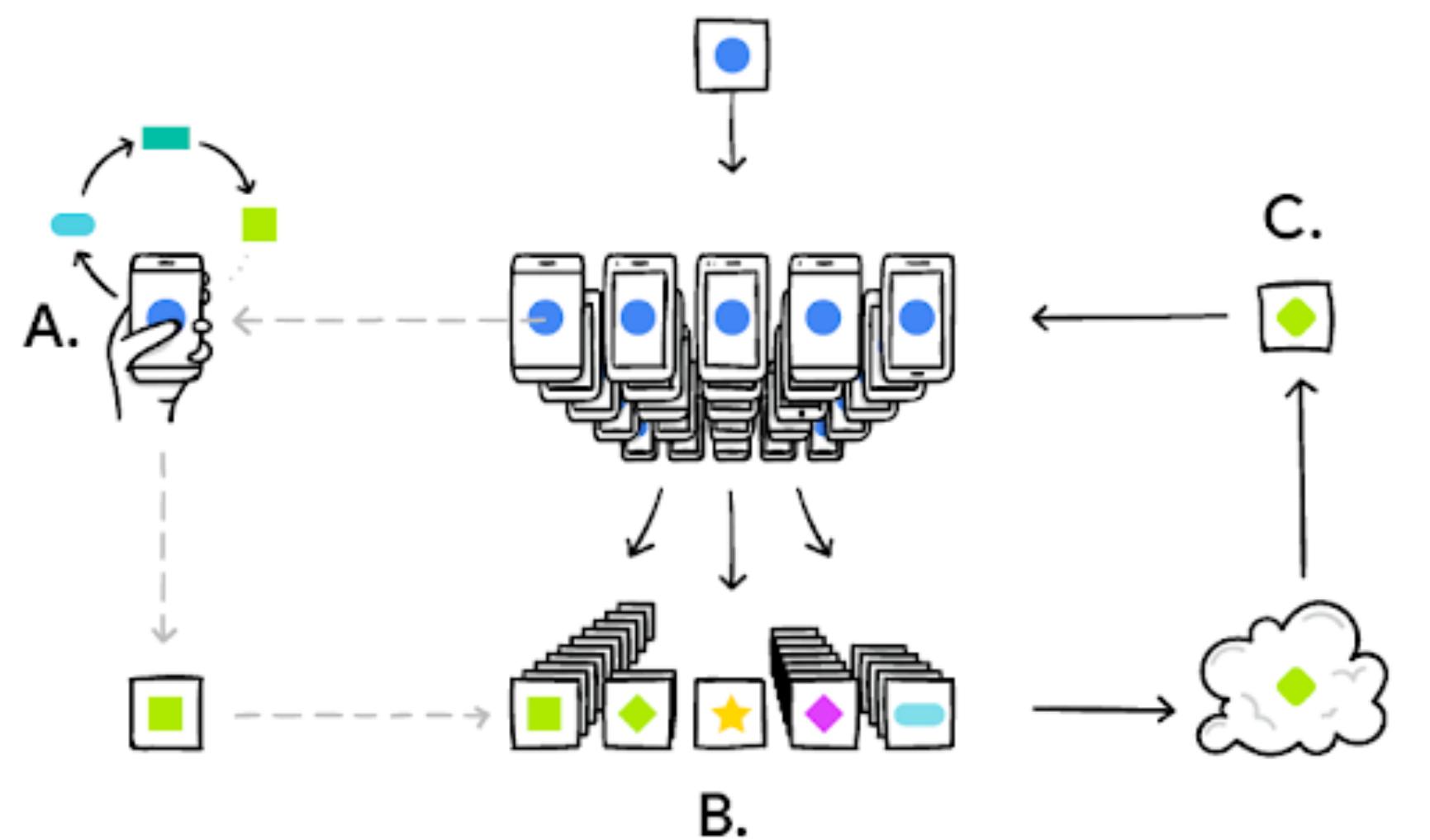
Technologies that Make this Possible

- Differential Privacy
- “Differential” - If I run a query without your record, then with your record, what can I learn about you from the difference??
- Introduce “noise” into the data so that:
 - Private data is obscured
 - Introduced error is bounded



Enterprise Applications?

- What services would your customers reject now, but accept if you offered the services using Federated Learning & Differential Privacy??





Classic techniques
that still deliver
lots of value

Outline

- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions

Statistical Inference

- Al Kindi (801-873):
 - On Deciphering Cryptographic Messages
 - Creator of cryptanalysis
 - Earliest known use of statistical inference



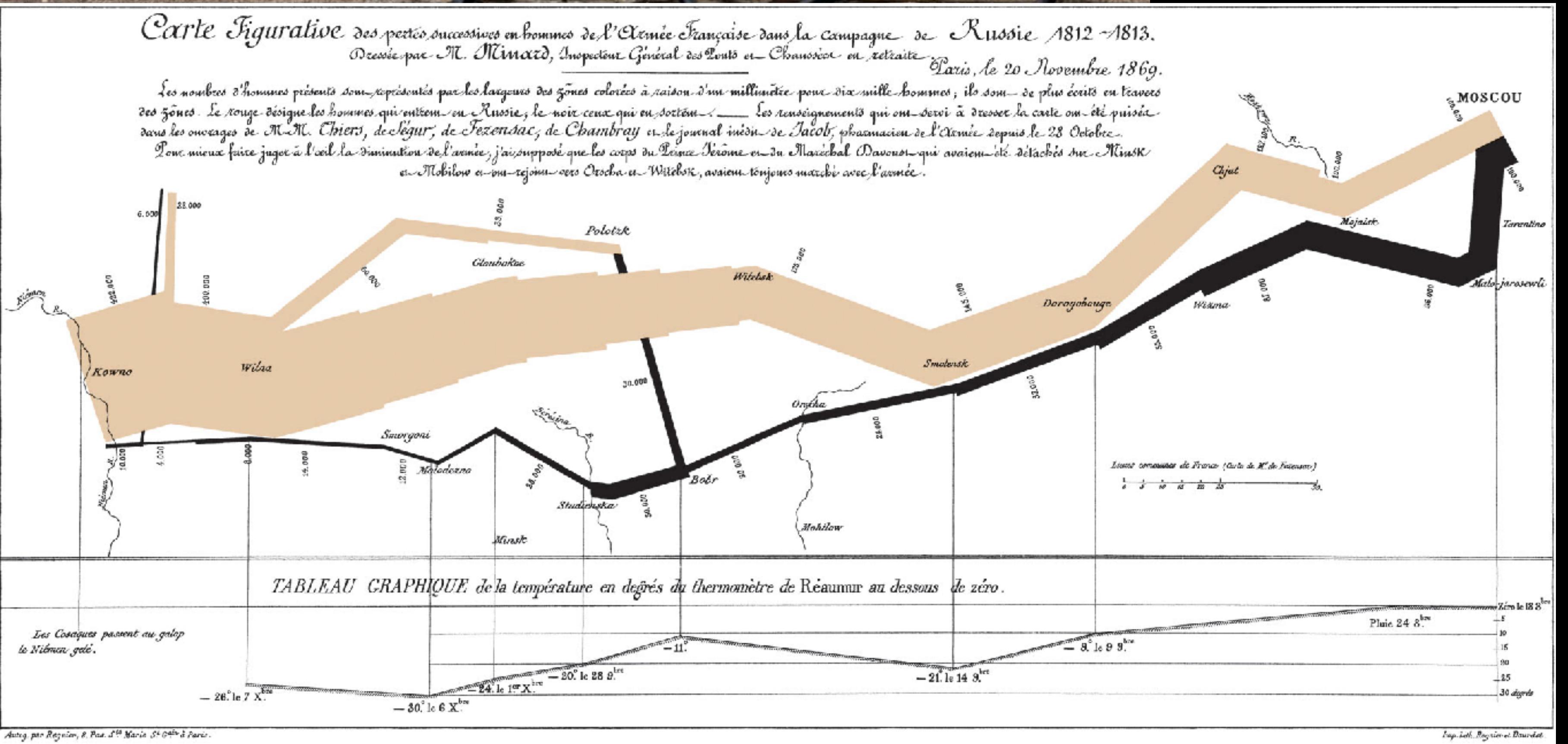
ذات يوم دخلوا على المأمور بالاستخبارات سرطان إلى المدير وقدموا معرضاً عرضوا عليه إبراهيم بيطاطاً يحيط به طبلة من ملوكهم، فلما سمع ذلك طلب المأمور عرضه، وعندما وصل إليه طبلة، قال لهم المأمور: «إنكم لا تؤدونني بطبلكم، بل أؤدوكم بطبلي». فلما سمعوا ذلك، أخذوا طبلة المأمور ووضعوها على طبلتهم، ولهذا سمي طبلة المأمور طبلة المأمور. ثم أخذوا طبلة المأمور ووضعوها على طبلة طبلة المأمور، ولذلك سمي طبلة طبلة المأمور. ثم أخذوا طبلة طبلة المأمور ووضعوها على طبلة طبلة طبلة المأمور، ولذلك سمي طبلة طبلة طبلة المأمور. ثم أخذوا طبلة طبلة طبلة المأمور ووضعوها على طبلة طبلة طبلة طبلة المأمور، ولذلك سمي طبلة طبلة طبلة طبلة المأمور. ثم أخذوا طبلة طبلة طبلة طبلة المأمور ووضعوها على طبلة طبلة طبلة طبلة طبلة المأمور، ولذلك سمي طبلة طبلة طبلة طبلة طبلة المأمور. ثم أخذوا طبلة طبلة طبلة طبلة طبلة المأمور ووضعوها على طبلة طبلة طبلة طبلة طبلة طبلة المأمور، ولذلك سمي طبلة طبلة طبلة طبلة طبلة طبلة المأمور. ثم أخذوا طبلة طبلة طبلة طبلة طبلة طبلة المأمور ووضعوها على طبلة طبلة طبلة طبلة طبلة طبلة طبلة المأمور، ولذلك سمي طبلة طبلة طبلة طبلة طبلة طبلة طبلة المأمور.



<https://en.wikipedia.org/wiki/Al-Kindi>



<https://datavizblog.com/2013/05/26/dataviz-history-charles-minards-flow-map-of-napoleons-russian-campaign-of-1812-part-5/>



Visualization

- Charles Minard's visualization of Napoleon's Russia Campaign (drawn 1861)



Visualization

- “On the Mode of Communication of Cholera”; by John Snow (1854)

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak



Neural Nets

- 1943 - McCulloch and Pitts - single layer
- ...
- Le Cun, et al.
(1989-1990)

Handwritten Zip Code Recognition with Multilayer Networks

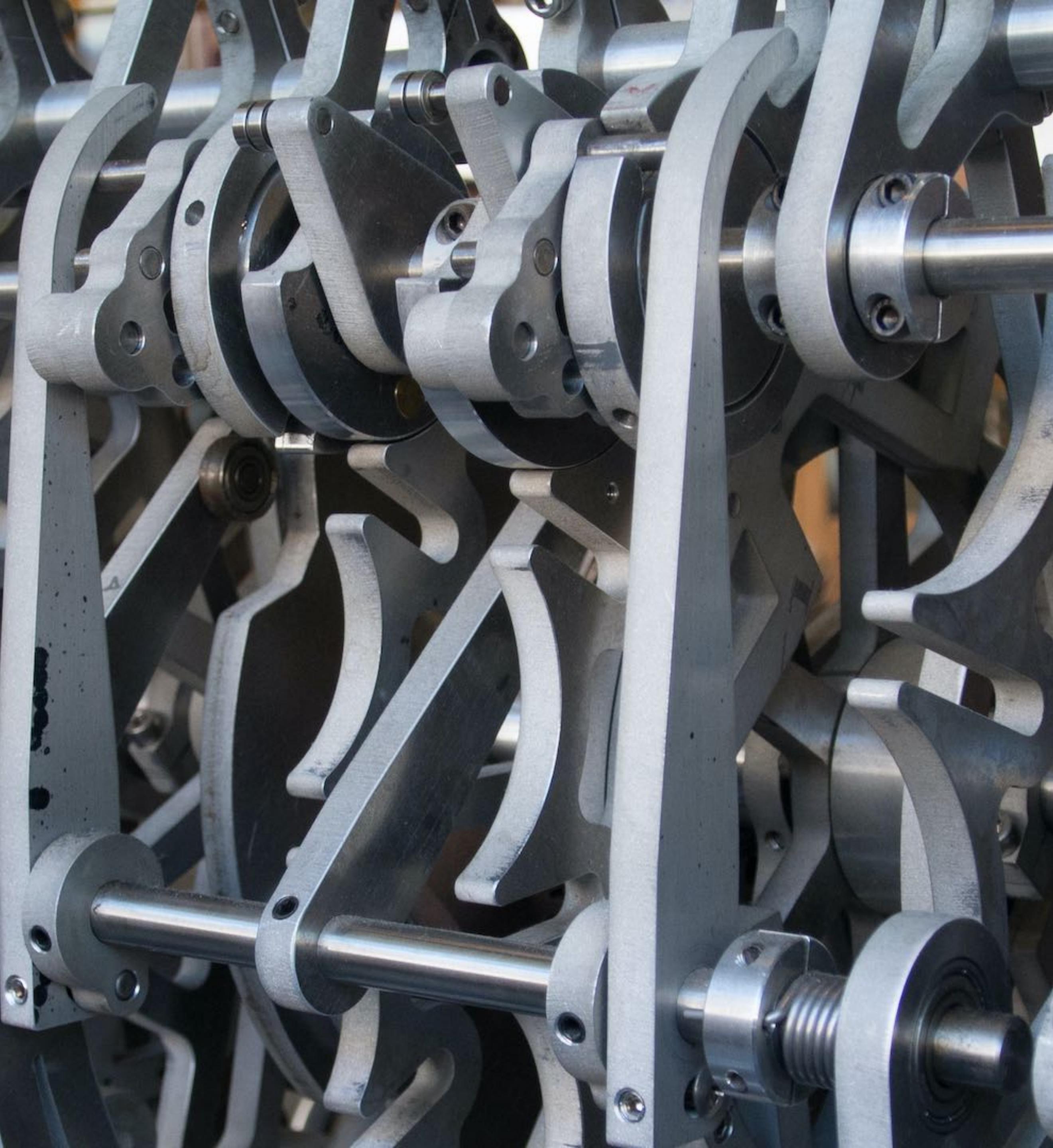
**Y. Le Cun, O. Matan, B. Boser, J. S. Denker, D. Henderson,
R. E. Howard, W. Hubbard, L. D. Jackel and H. S. Baird**
AT&T Bell Laboratories, Holmdel, N.J. 07733

A zip code

Abstract

We present an application of backpropagation networks to handwritten zip

only be obtained by designing a network architecture that contains a certain amount of *a priori* knowledge about the problem. The basic design



Outline

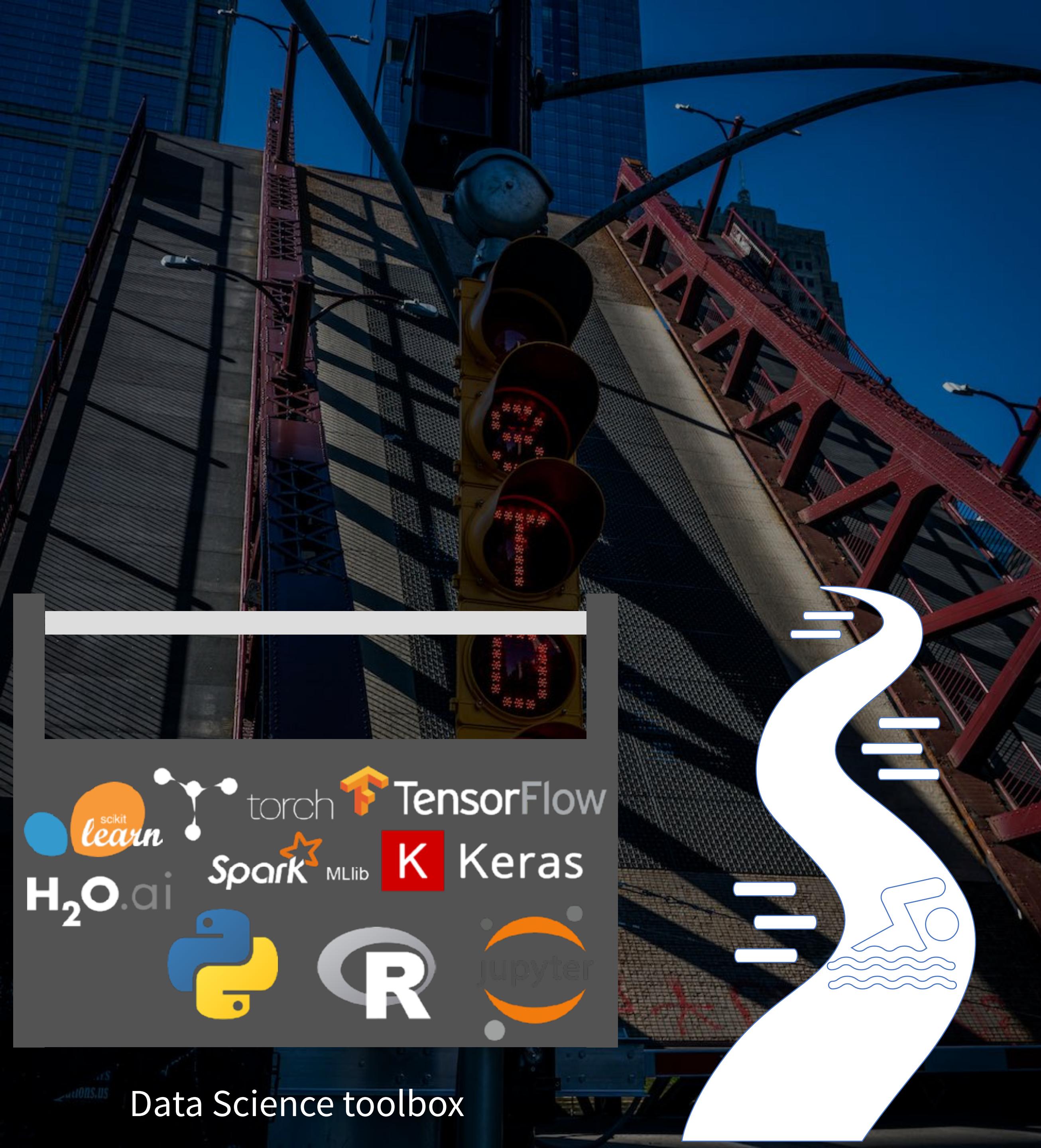
- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions

A photograph of the Chicago skyline, viewed from across the Chicago River. The foreground shows the river with a small boat. In the background, the iconic Marina City towers rise on the left, and the Michigan Avenue Bridge spans the river. Other skyscrapers of the Loop are visible against a clear blue sky.

All the current capabilities of the
Promise of AI section are
available now, but they are hard
to build and use.



Data Science vs. Data Engineering



Data Science vs. Data Engineering

- A cultural and technical divide



Data Science toolbox

Software Engineering toolbox

@deanwampler

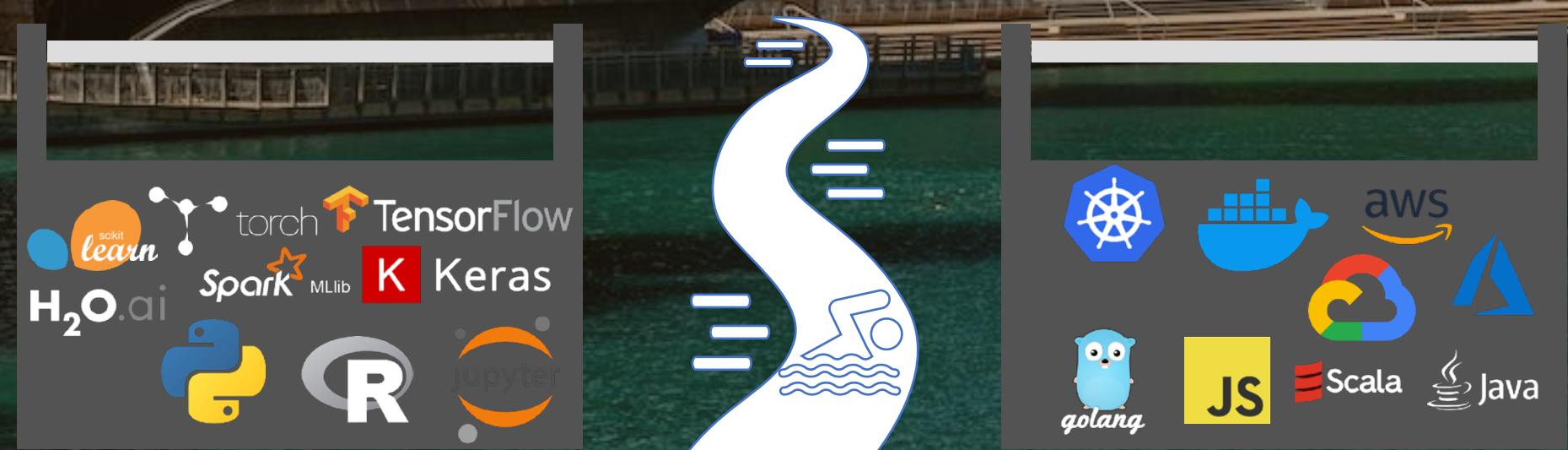
Data Scientists

- Comfortable with uncertainty
- Less process oriented
 - Iterative, experimental

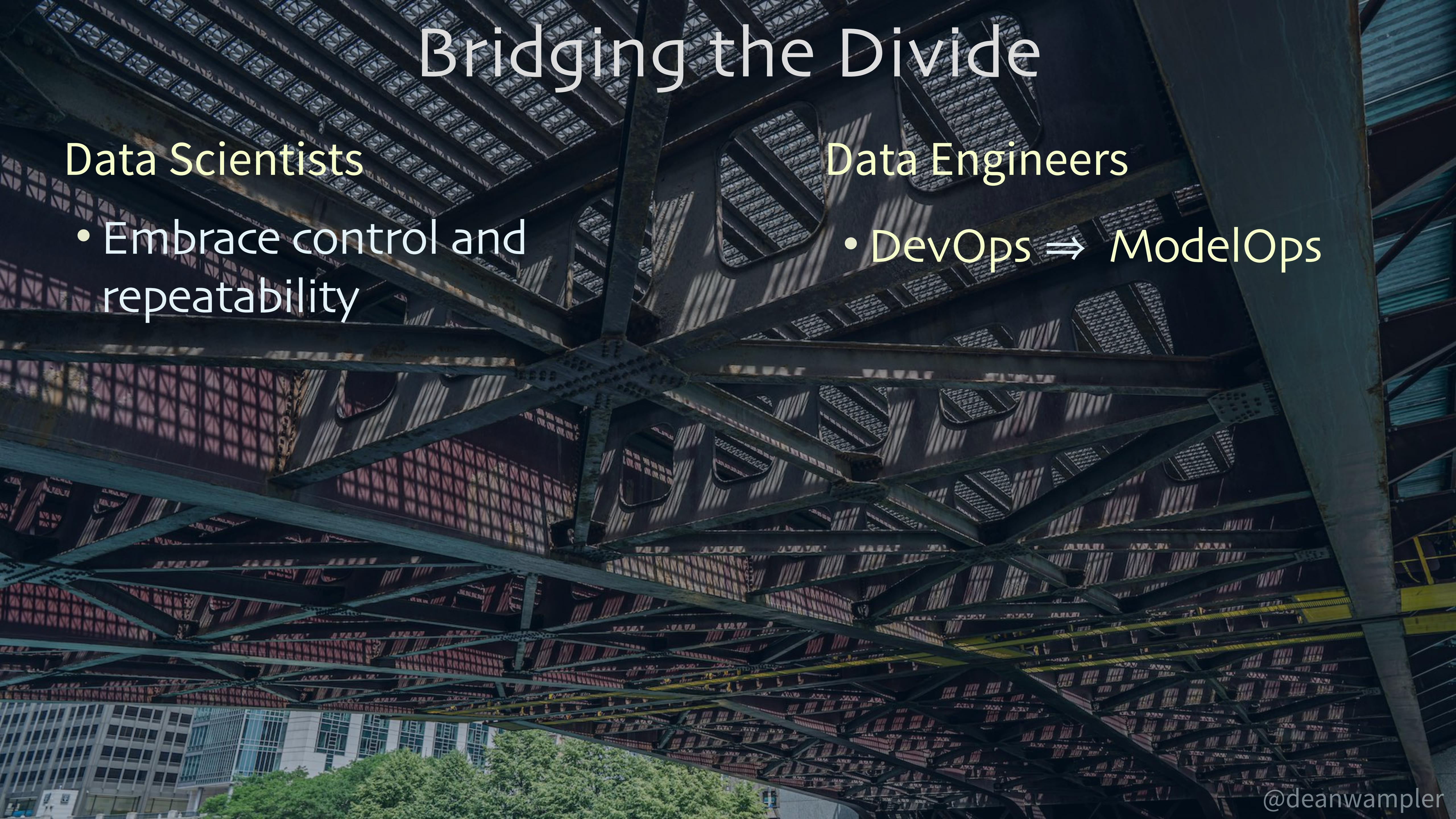
Data Engineers

- Uncomfortable with uncertainty
- Process oriented
 - Agile Manifesto
 - ... which does not mention data!

<https://derwen.ai/s/6fqt>



Bridging the Divide



Data Scientists

- Embrace control and repeatability

Data Engineers

- DevOps → ModelOps

Bridging the Divide

Data Scientists

- Embrace control and repeatability

Data Engineers

- DevOps → ModelOps

Model: An algorithm that makes a prediction or recommendation or prescribes some action based on a probabilistic assessment. Data scientists make models.

<https://www.dominodatalab.com/blog/model-management-and-the-era-of-the-model-driven-business/>



ModelOps

“ModelOps is a principled approach to operationalizing a model in apps. ModelOps synchronizes cadences between the application and model pipelines. ... you can optimize your data science and AI investments using data, models, and resources from edge to core to cloud.”

<https://www.ibm.com/cloud/machine-learning/modelops>

ModelOps

And if you look at the most successful companies in the world, you'll find models at the heart of their business driving that success.

- Example: Netflix recommendation model
 - Drives subscriber engagement, retention, and operational efficiency.
 - Their recommendation model is worth more than \$1B per year (2016).

ModelOps

And if you look at the most successful companies in the world, you'll find models at the heart of their business driving that success.

- Example: Coca-Cola
 - Optimizes orange juice production, ...
- Example: Stitch Fix and Trunk Club
 - Clothing recommendations for customers

ModelOps

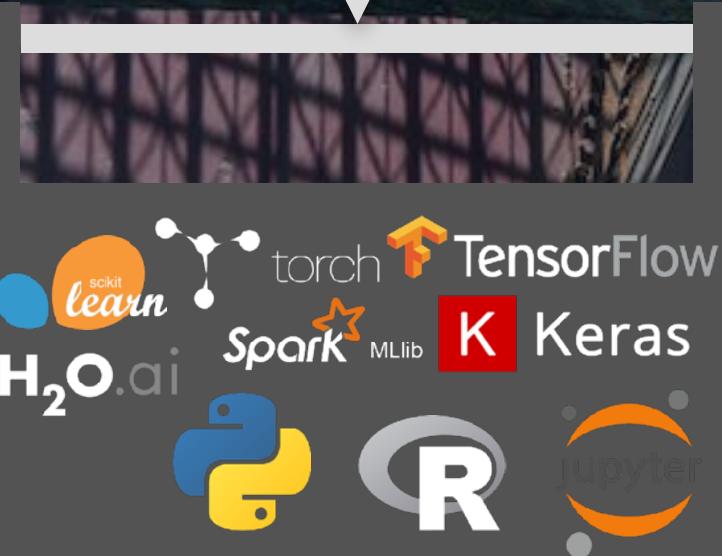
And if you look at the most successful companies in the world, you'll find models at the heart of their business driving that success.

- Example: Insurance companies
 - Actuarial models (very old technique...)
 - Now using models to make automated damage estimates from accident photos, reducing dependence on claims adjusters.

ModelOps



Data



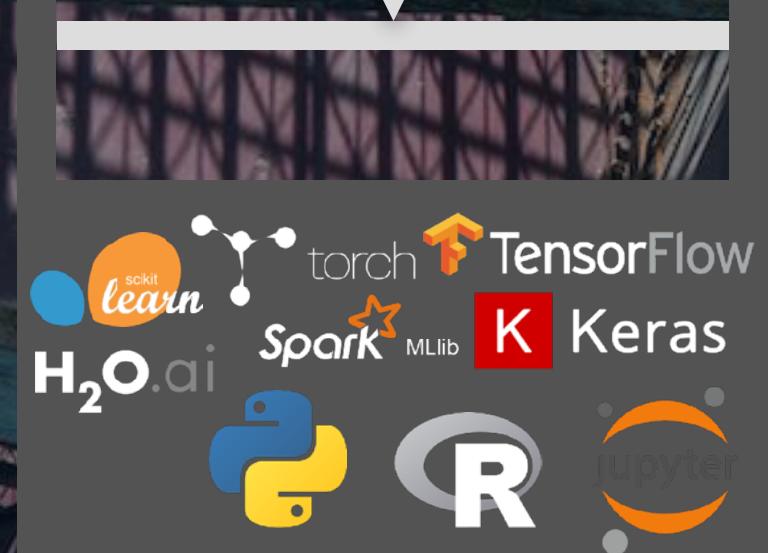
Model Development

Research
new models

ModelOps



Data



Model Development

Research
new models



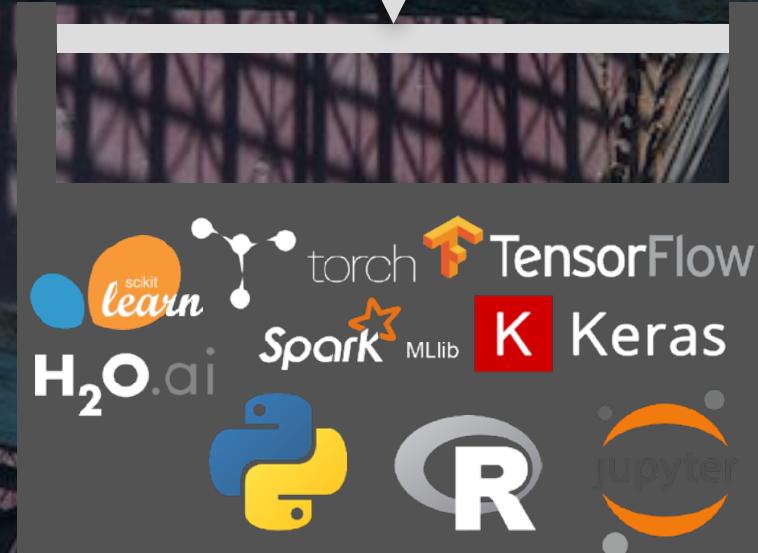
Model CI/CD Pipeline

Versioning, traceability,
reproducibility, automation

ModelOps



Data

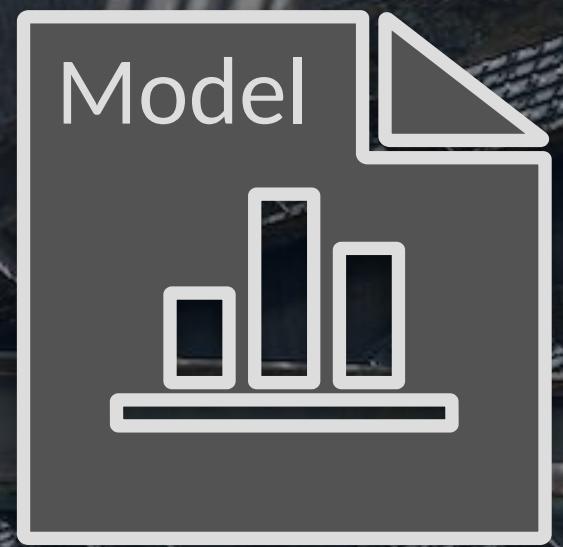


Model Development

Research
new models



Model CI/CD Pipeline

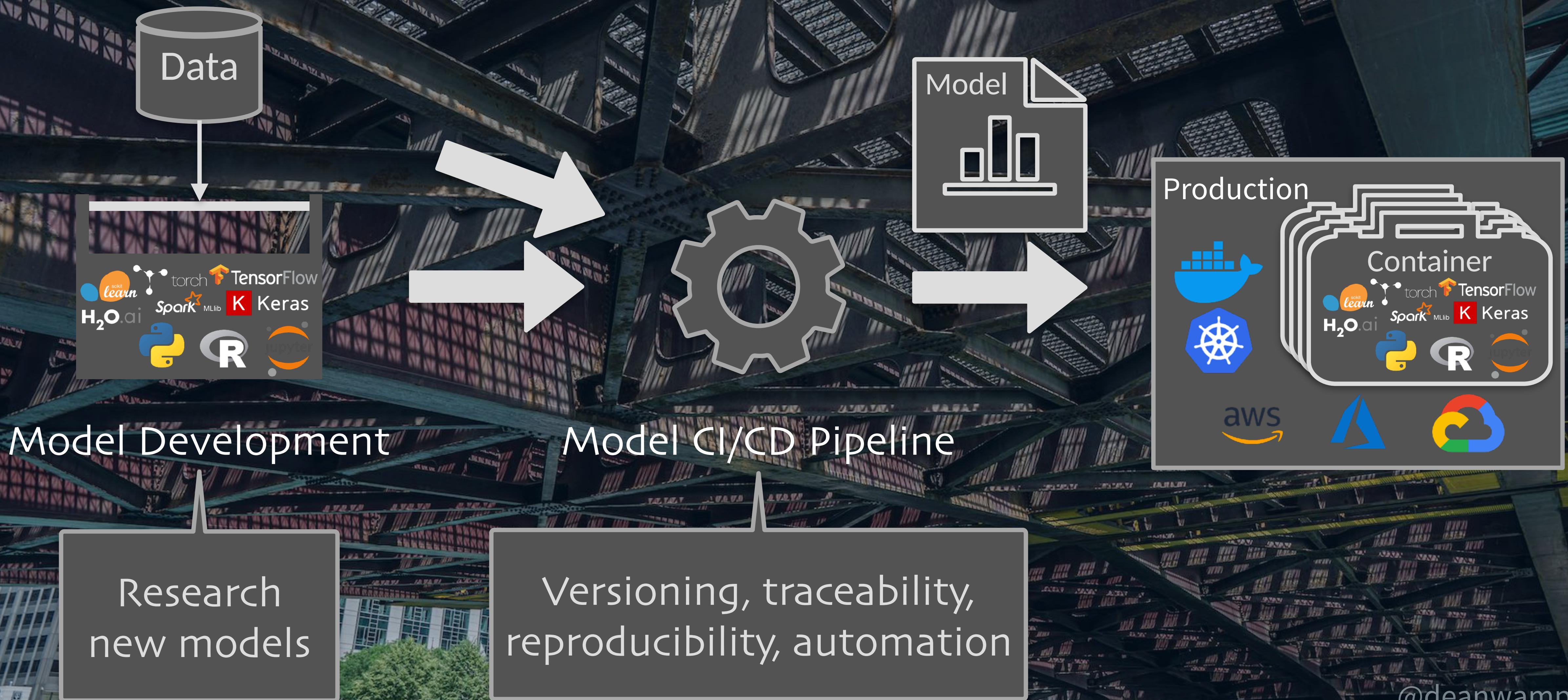


Model

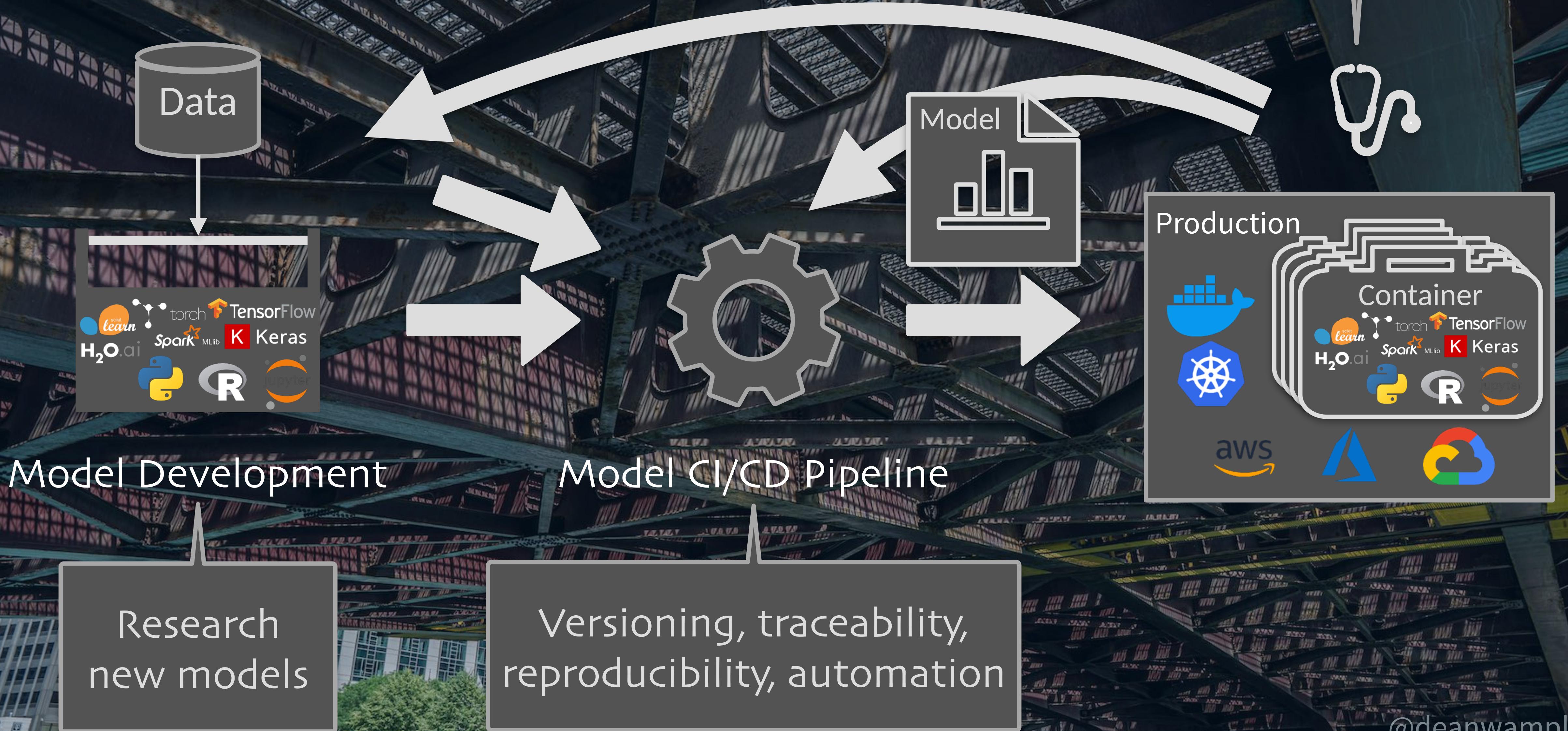


Versioning, traceability,
reproducibility, automation

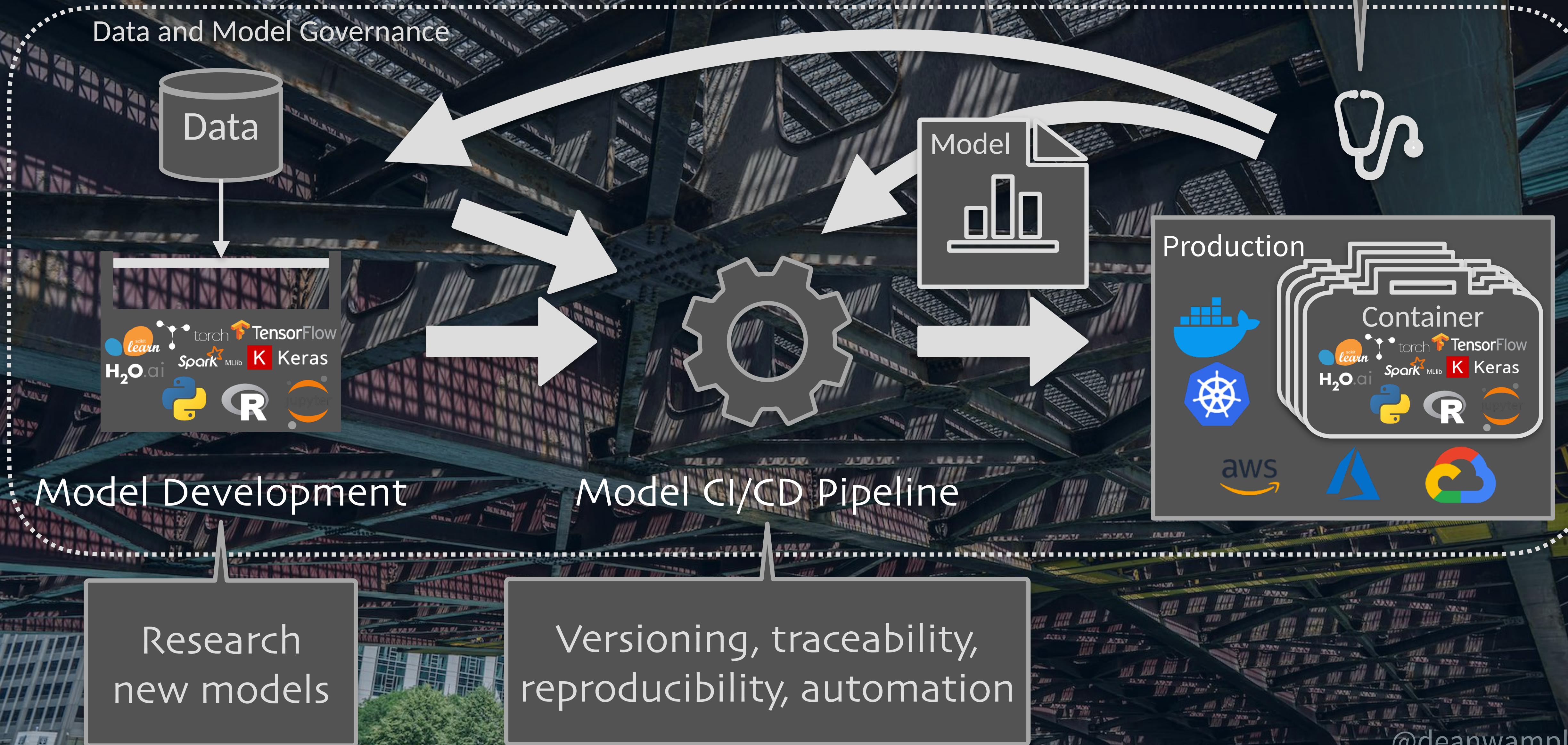
ModelOps



ModelOps



ModelOps



ModelOps

Monitor

This is shown as a batch process, but expect these processes to evolve into streaming pipelines, with continuous training.

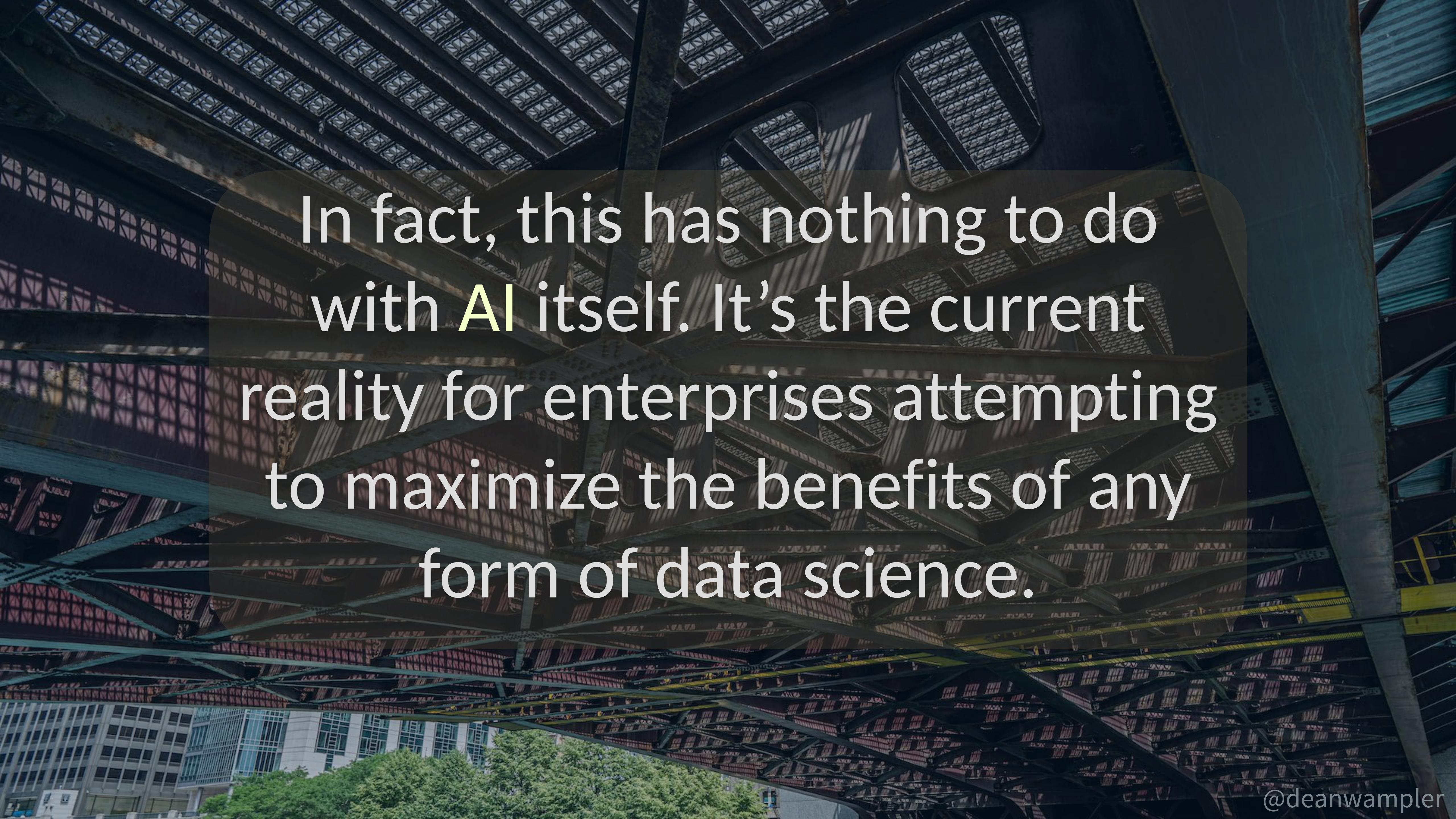
Mode

R

new models

reproducibility, automation



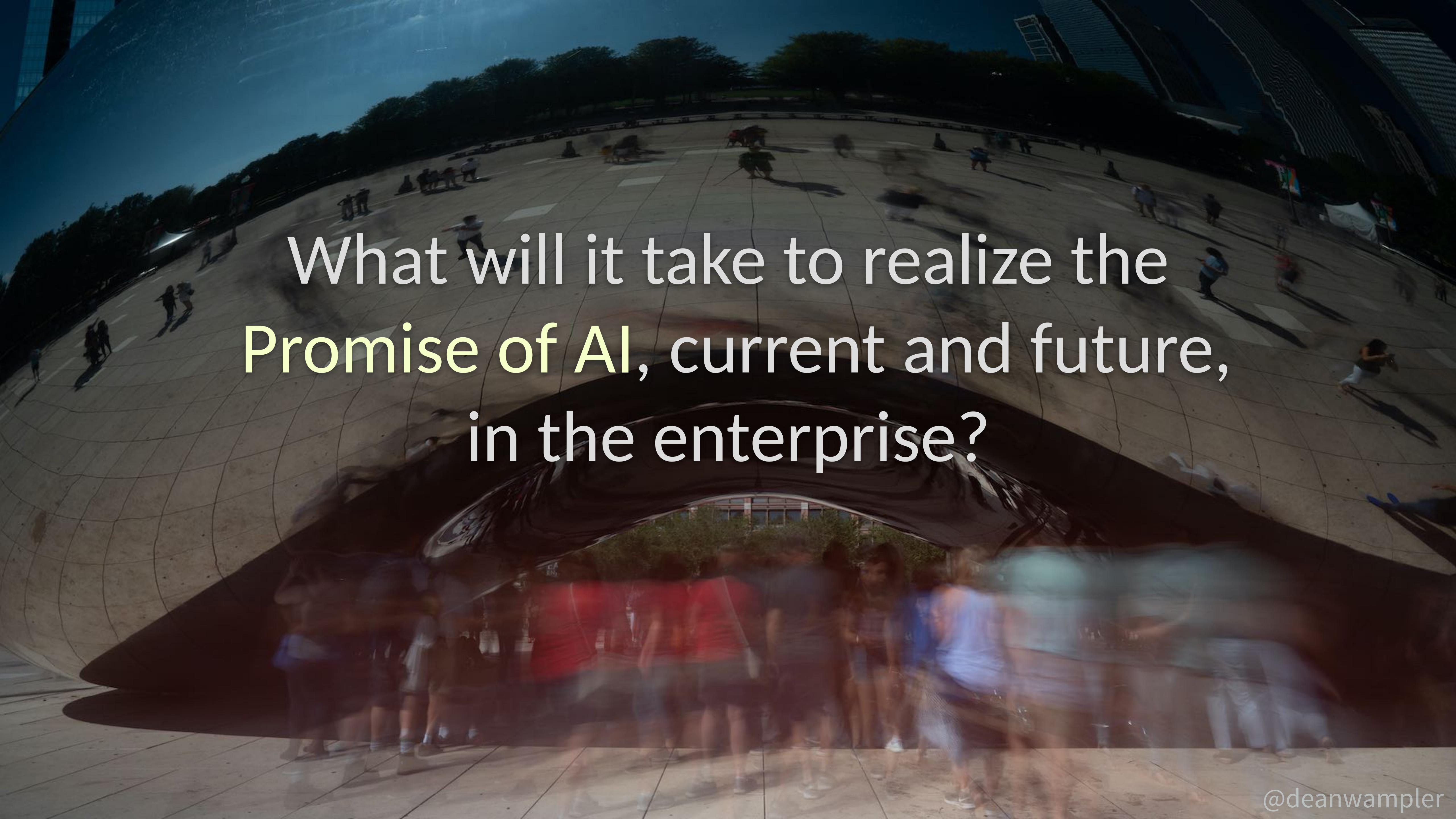


In fact, this has nothing to do
with AI itself. It's the current
reality for enterprises attempting
to maximize the benefits of any
form of data science.



Outline

- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions



What will it take to realize the
Promise of AI, current and future,
in the enterprise?



AI in the Enterprise

- Fully adopting:
 - Natural Language Processing
 - Reinforcement Learning
 - Ubiquitous AI in Applications



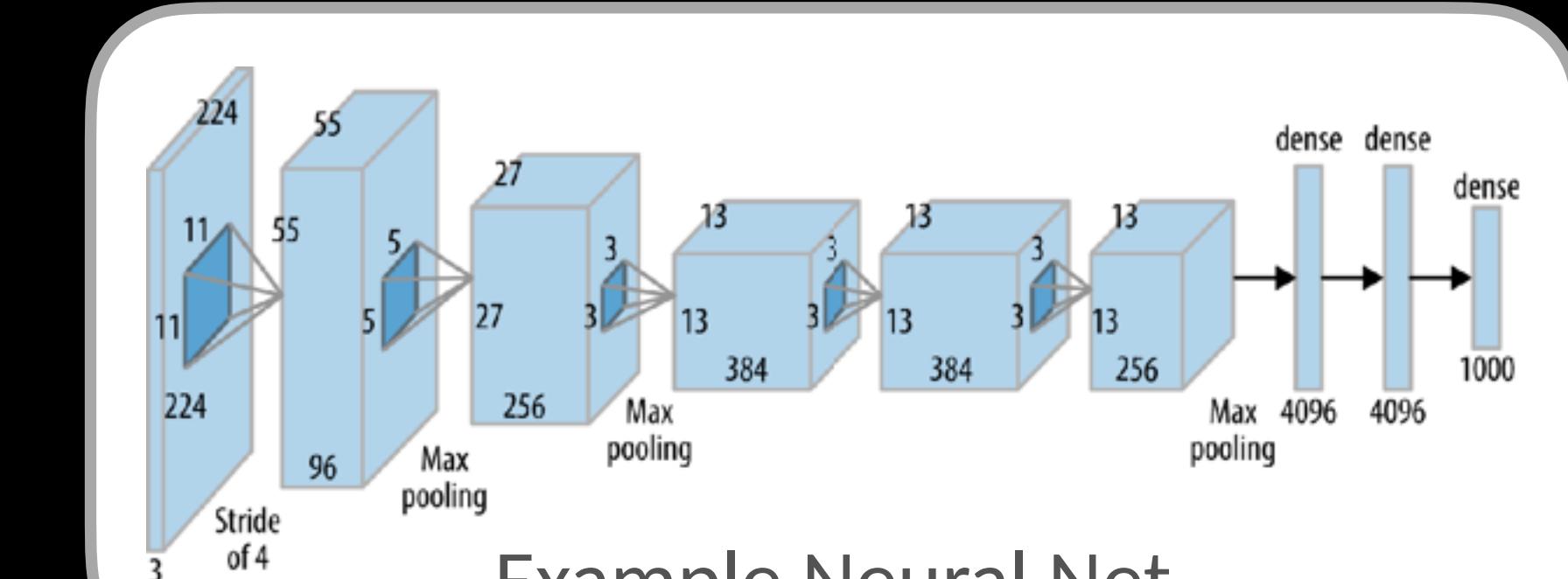
Along the way...

- Infrastructure Changes
 - Cloud
 - Scaling computation
 - Diff. Privacy & Fed. Learning
- Software Development

“The largest version GPT-3 175B or ‘GPT-3’ has 175 B Parameters, 96 attention layers and 3.2 M batch size.”

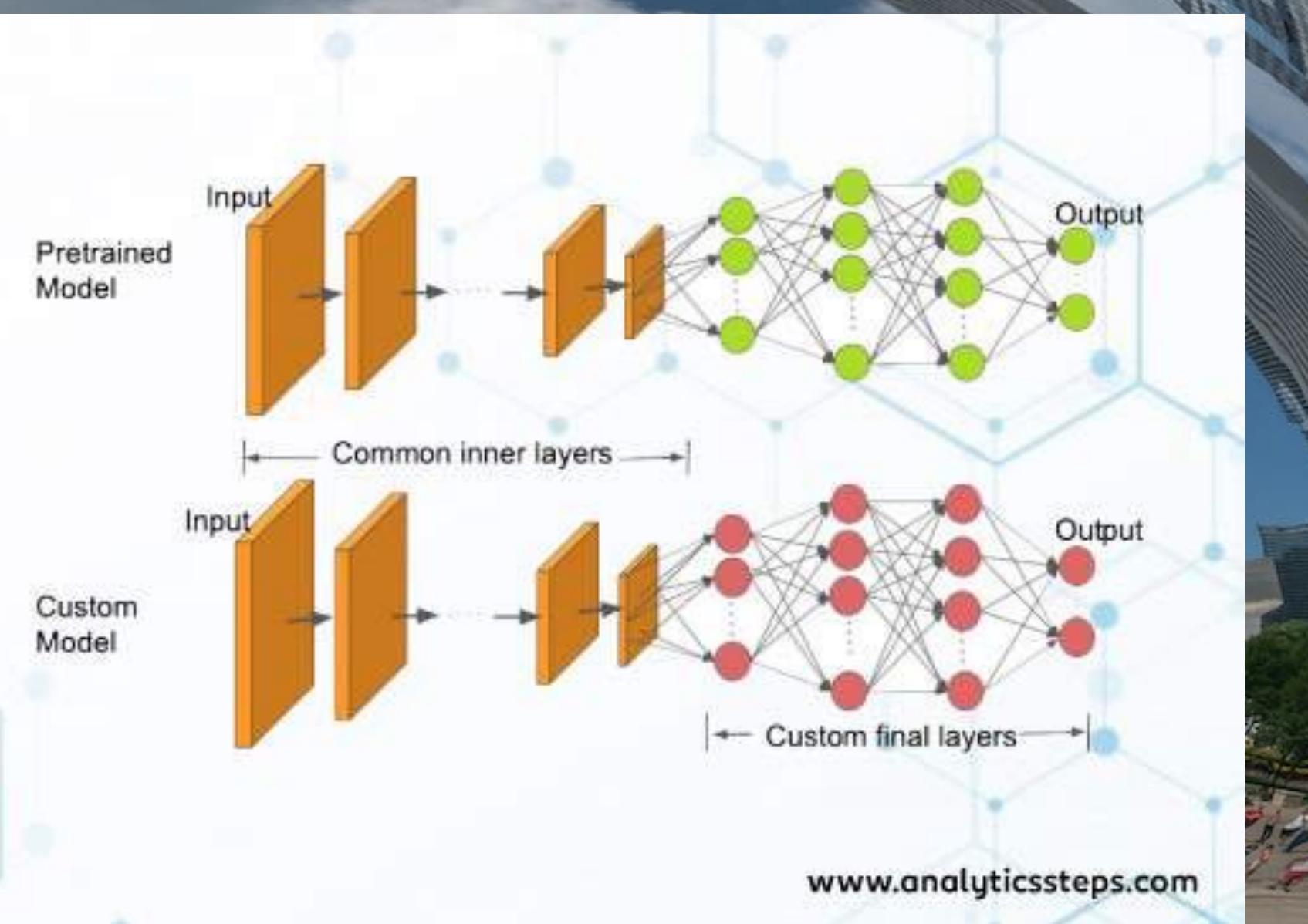
Transfer Learning

- NLP: The world's largest neural networks



Example Neural Net

Transfer Learning



<https://analyticssteps.com/blogs/how-transfer-learning-done-neural-networks-and-convolutional-neural-networks>

Transfer Learning

- Fortunately, you can start with a trained model and further refine it for your problem.



Reinforcement Learning

- While “classic” RL uses a simulator, you can also train on historical (“offline”) data.
- Use when a good simulator doesn’t exist or is too hard to create.

<https://arxiv.org/abs/2005.01643>

@deanwampler



Infrastructure

- Model training, especially NNs, is very expensive.
 - Burst to the cloud
 - Or have lots of in-house compute available!



Infrastructure

- A hybrid-cloud model balances:
- Security & regulatory benefits of on-premise cluster
- Burst of resources when you need them.



Infrastructure

- But, don't forget the cost of moving data between on-premise clusters and the cloud, as well as between clouds!



Infrastructure

- Leverage federated learning and differential privacy.
- Offload some computation!
- Meet data privacy objectives.

<https://openmined.org/>

@deanwampler

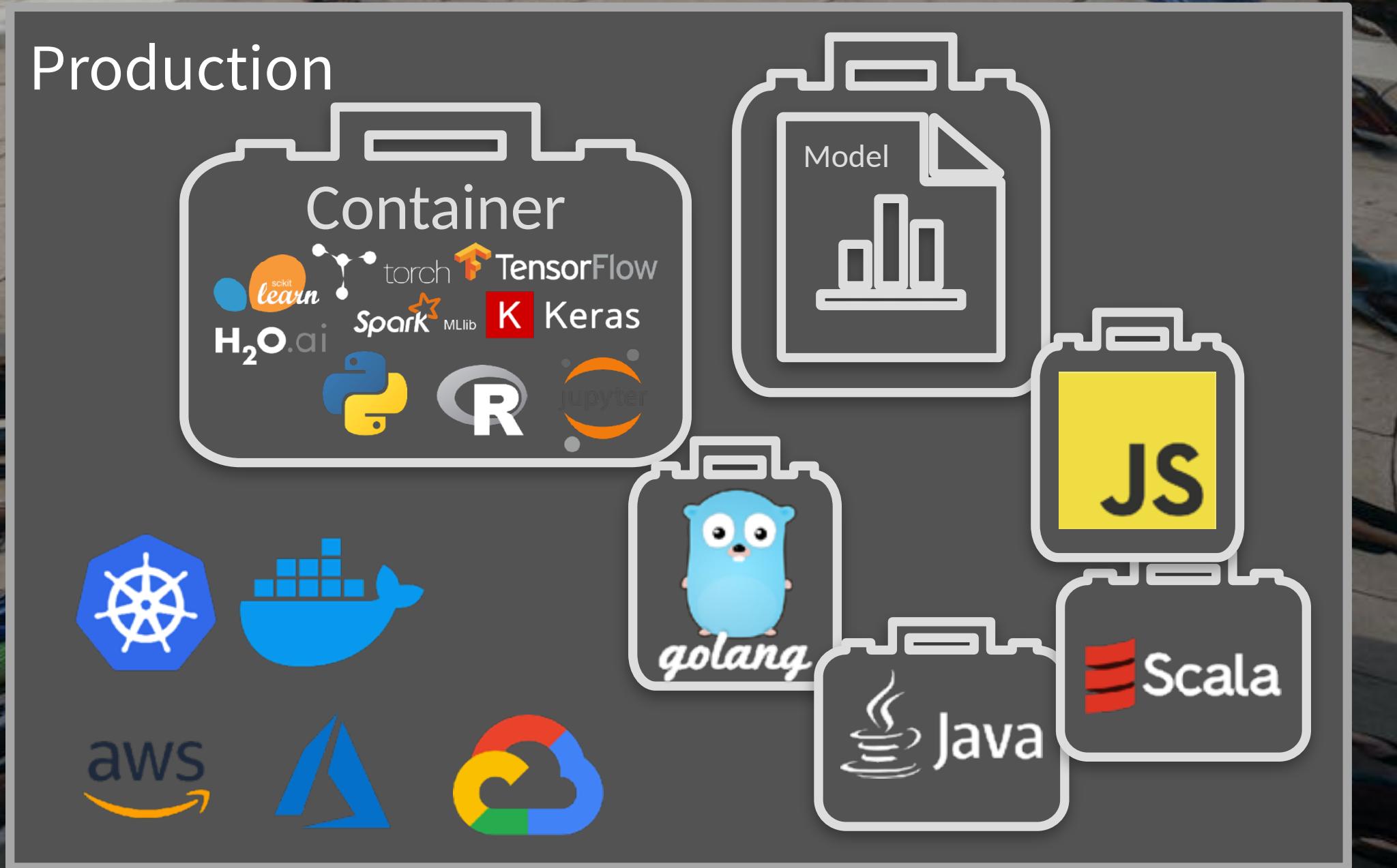


Software Development Impacts

- Ubiquitous AI requires:
 - Heterogeneous tools
 - Batch and stream data processing
- Statistical & probabilistic thinking

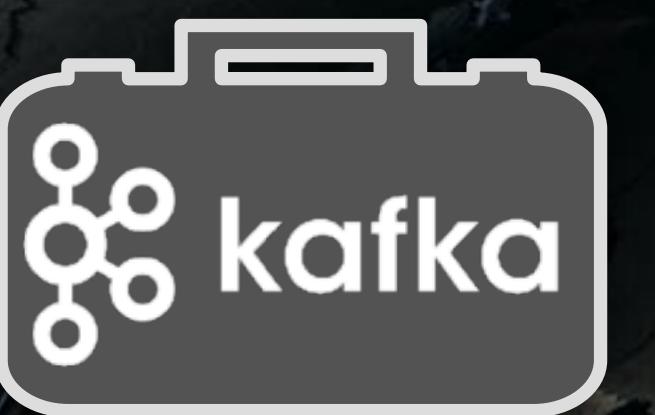
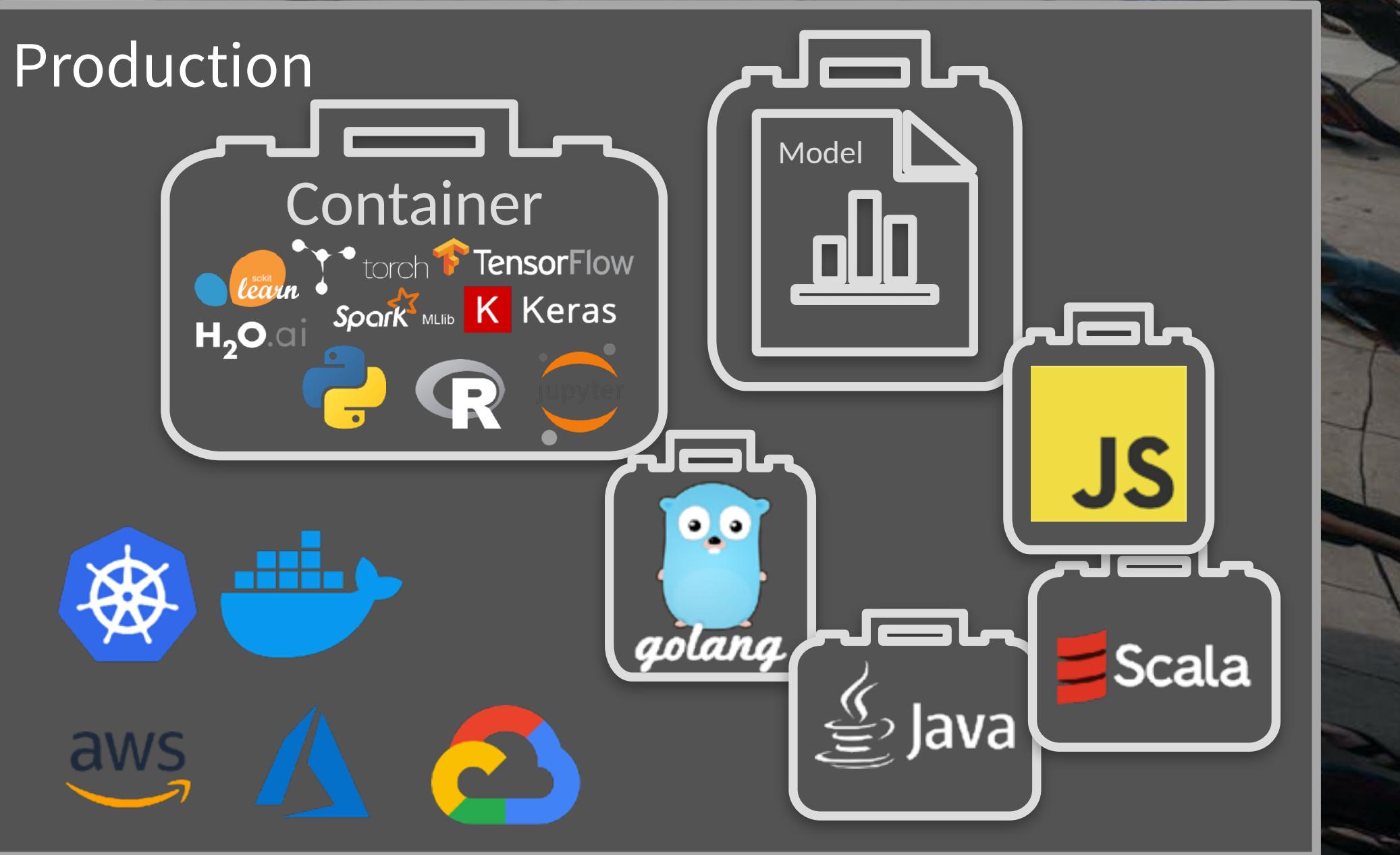
Software Development Impacts

- Ubiquitous AI requires:
 - Heterogeneous tools
 - Batch and streaming data processing
 - Statistical & probabilistic thinking



Software Development Impacts

- Ubiquitous AI requires:
 - Heterogeneous tools
 - Batch and streaming data processing
 - Statistical & probabilistic thinking



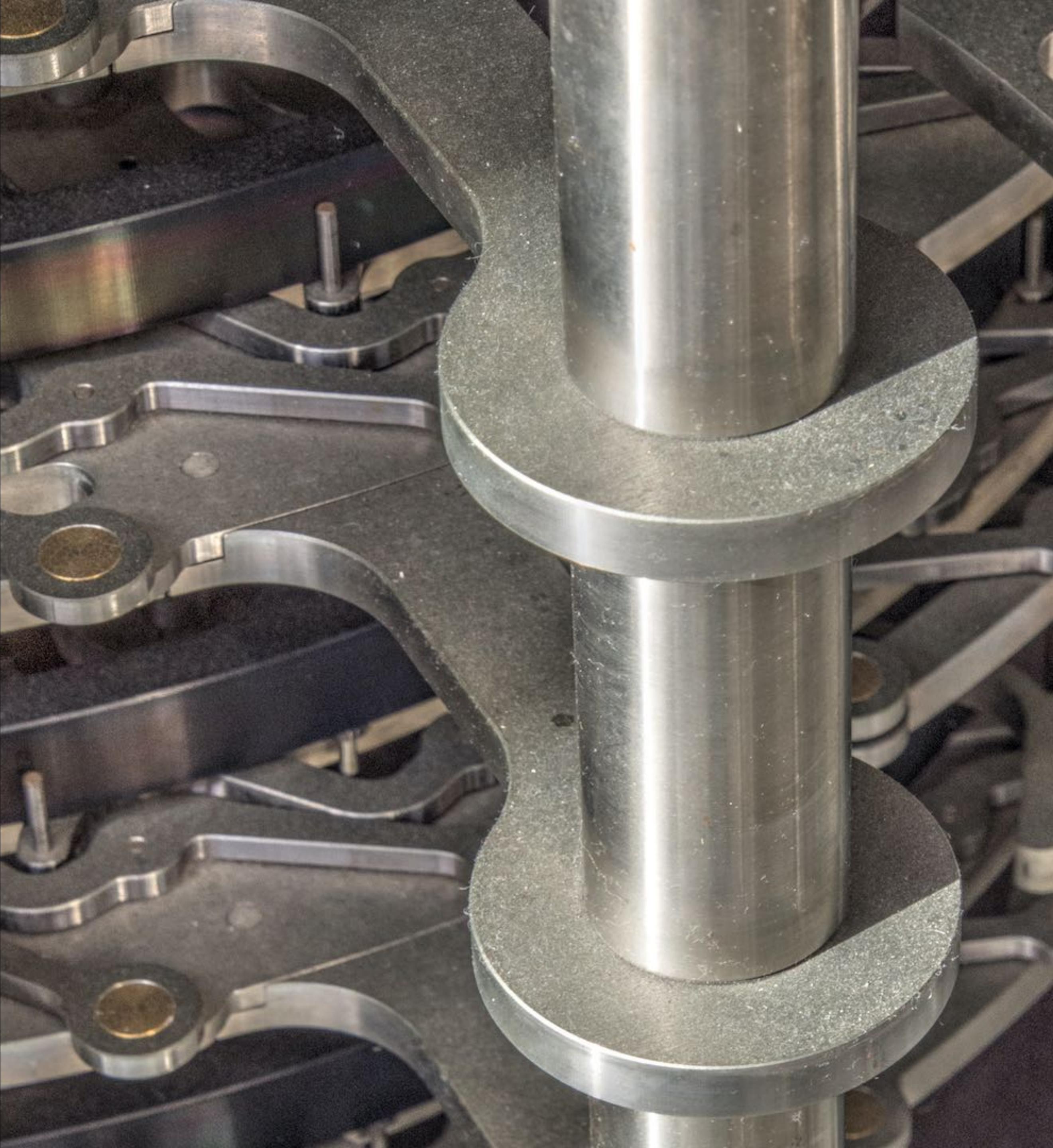


Probabilistic
results from
models



Software Development Impacts

- Ubiquitous AI requires:
 - Heterogeneous tools
 - Batch and streaming data processing
- Statistical & probabilistic thinking

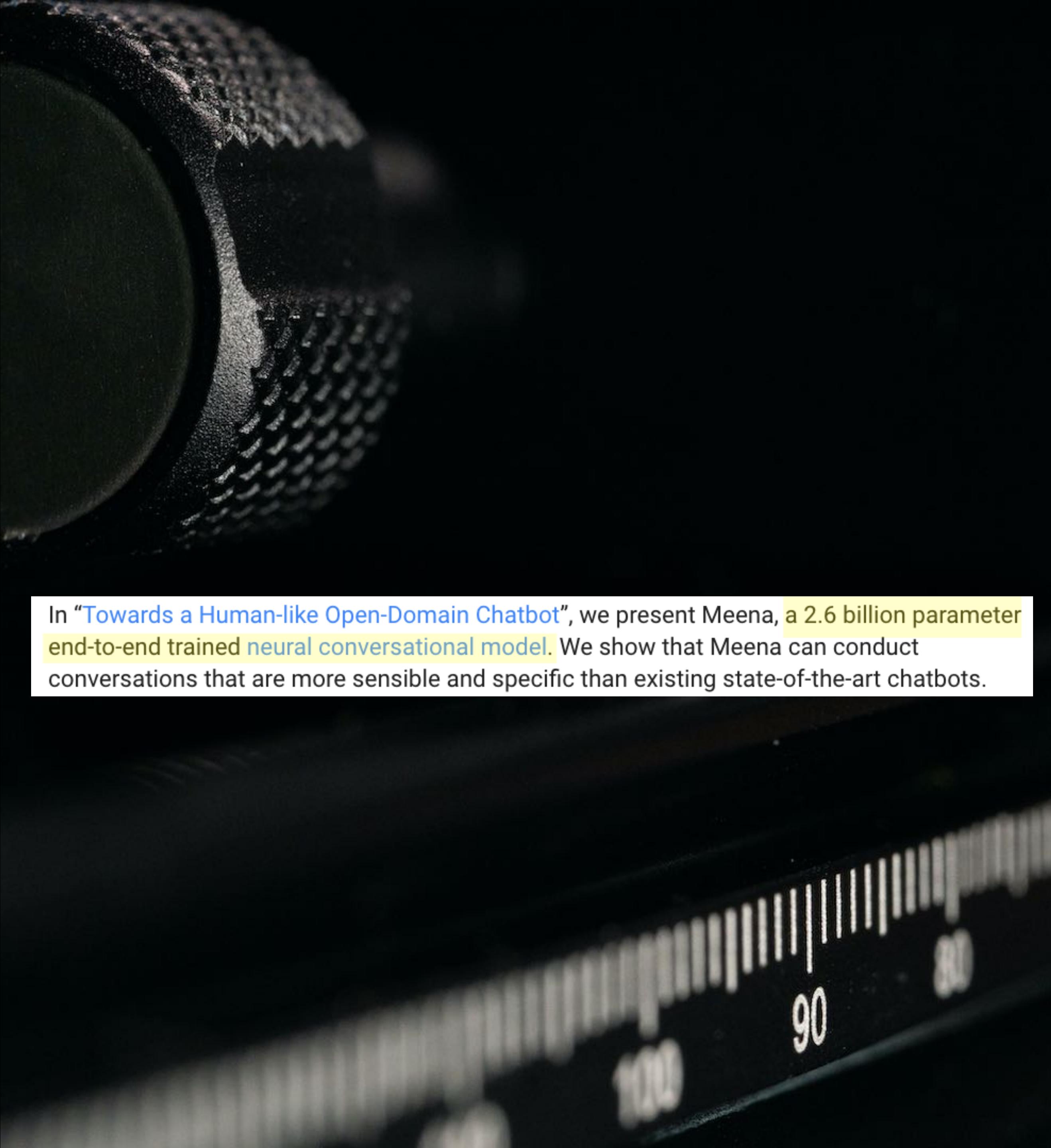


Outline

- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions



We can expect AI to become ubiquitous in the coming years, providing competitive advantages for enterprises that learn how to use it.



AI's Promise

- Natural Language Processing has become very capable, with wide applications

In "Towards a Human-like Open-Domain Chatbot", we present Meena, a 2.6 billion parameter end-to-end trained neural conversational model. We show that Meena can conduct conversations that are more sensible and specific than existing state-of-the-art chatbots.

AI's Promise



- Reinforcement Learning is being applied to many enterprise problems where sequential activity is central.



nature reviews cancer

View all

Explore our content ▾ Journal information ▾

nature > nature reviews cancer > perspectives > article

Perspective | Published: 17 May 2018

OPINION

Artificial intelligence in radiology

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz & Hugo J. W. L. Aerts✉

Nature Reviews Cancer 18, 500–510(2018) | Cite this article

15k Accesses | 317 Citations | 311 Altmetric | Metrics

Abstract

Artificial intelligence (AI) algorithms, particularly deep learning, have

ognition tasks. Methods
rialization autoencoders
age analysis field,
in radiology practice,
ges for the detection

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

AI's Promise

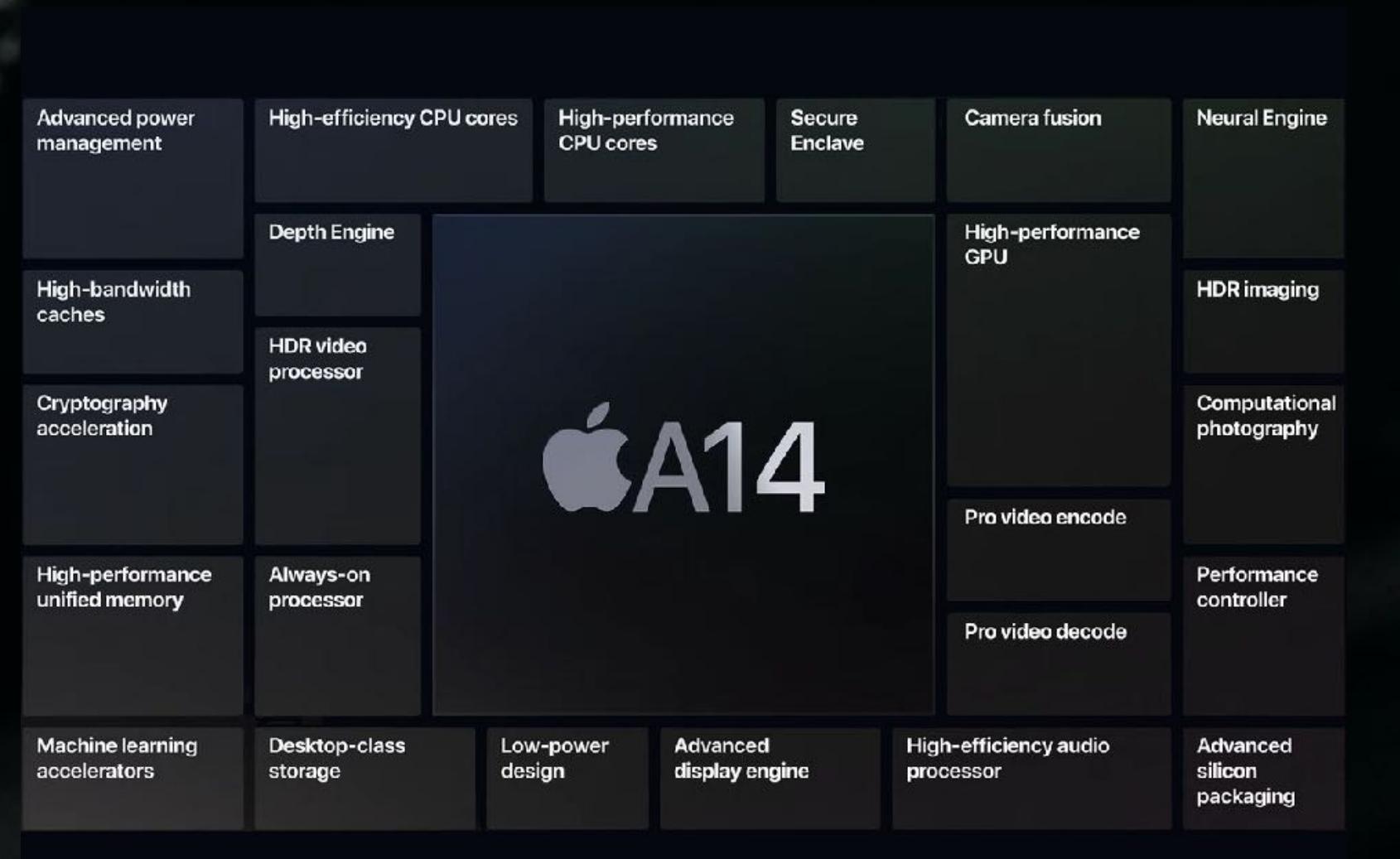
- New sciences and industries are benefiting from AI



@deanwampler

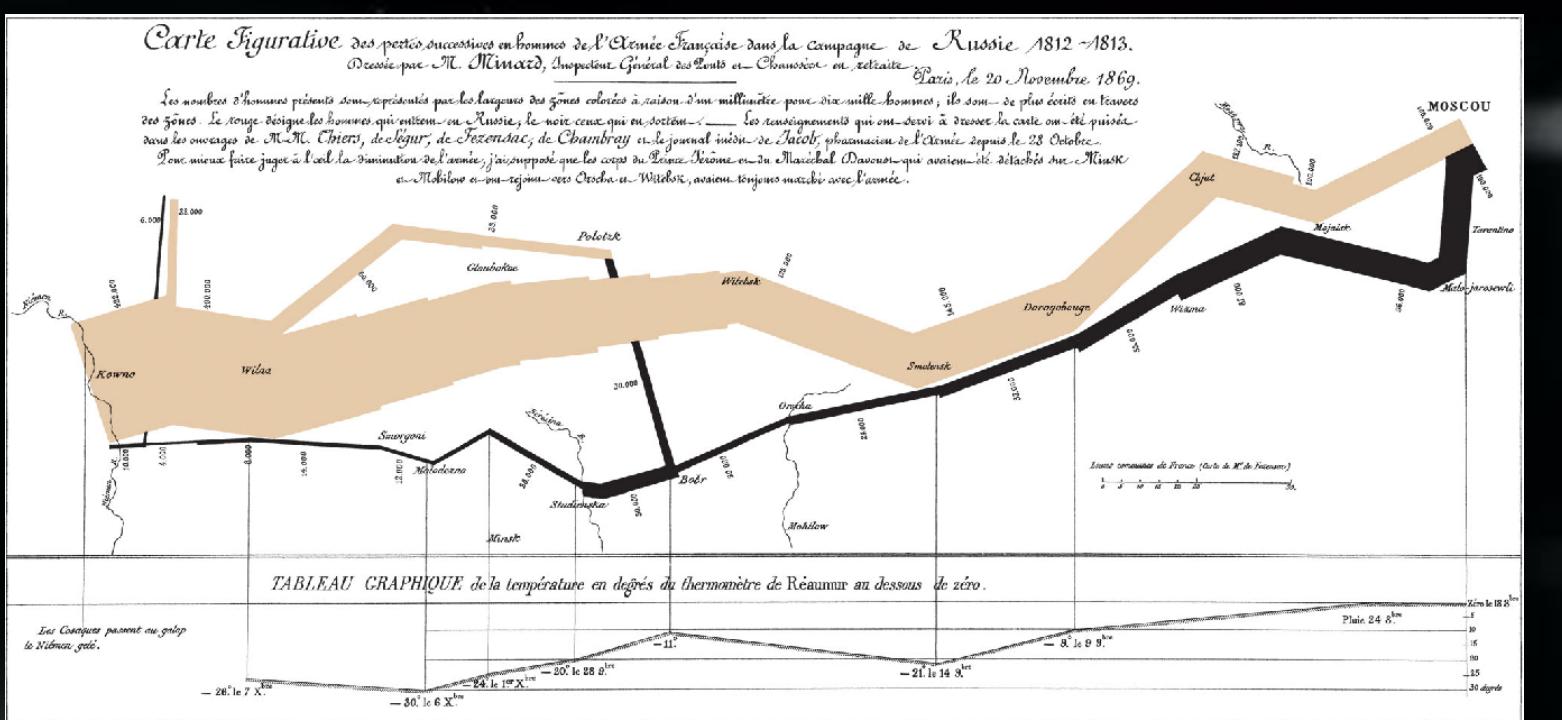
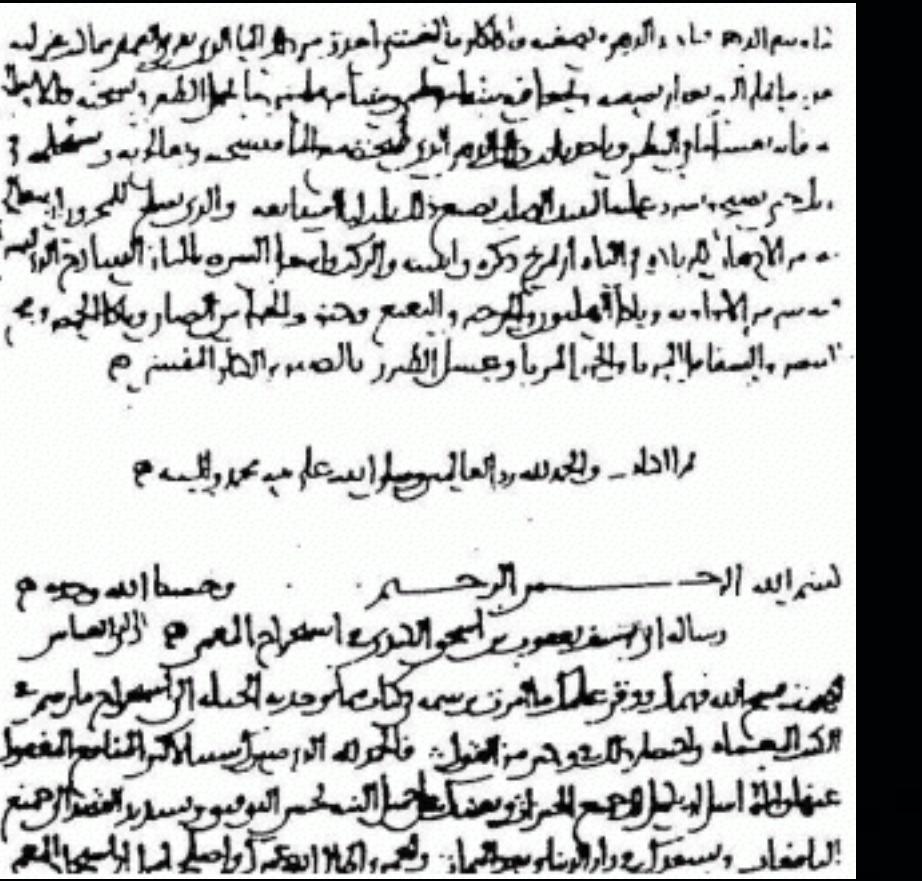
AI's Promise

- Mobile phones are showing us how AI is enabling new system features and enhancing capabilities in applications



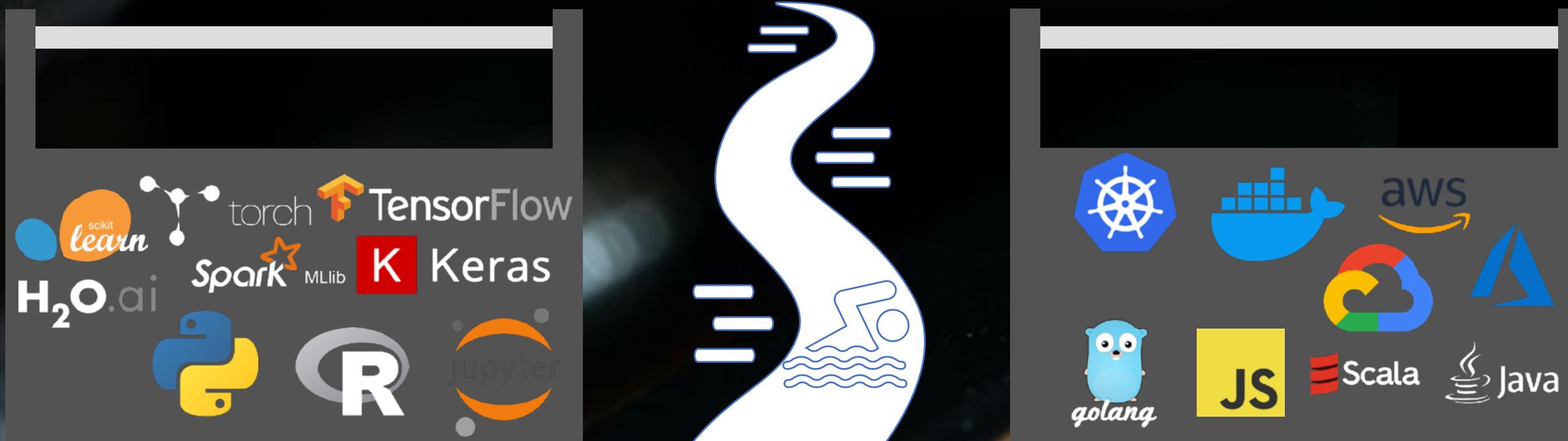
The Past

- Traditional data science tools still provide important benefits:
 - Proven Maturity
 - Explainability
 - Cheap to use!



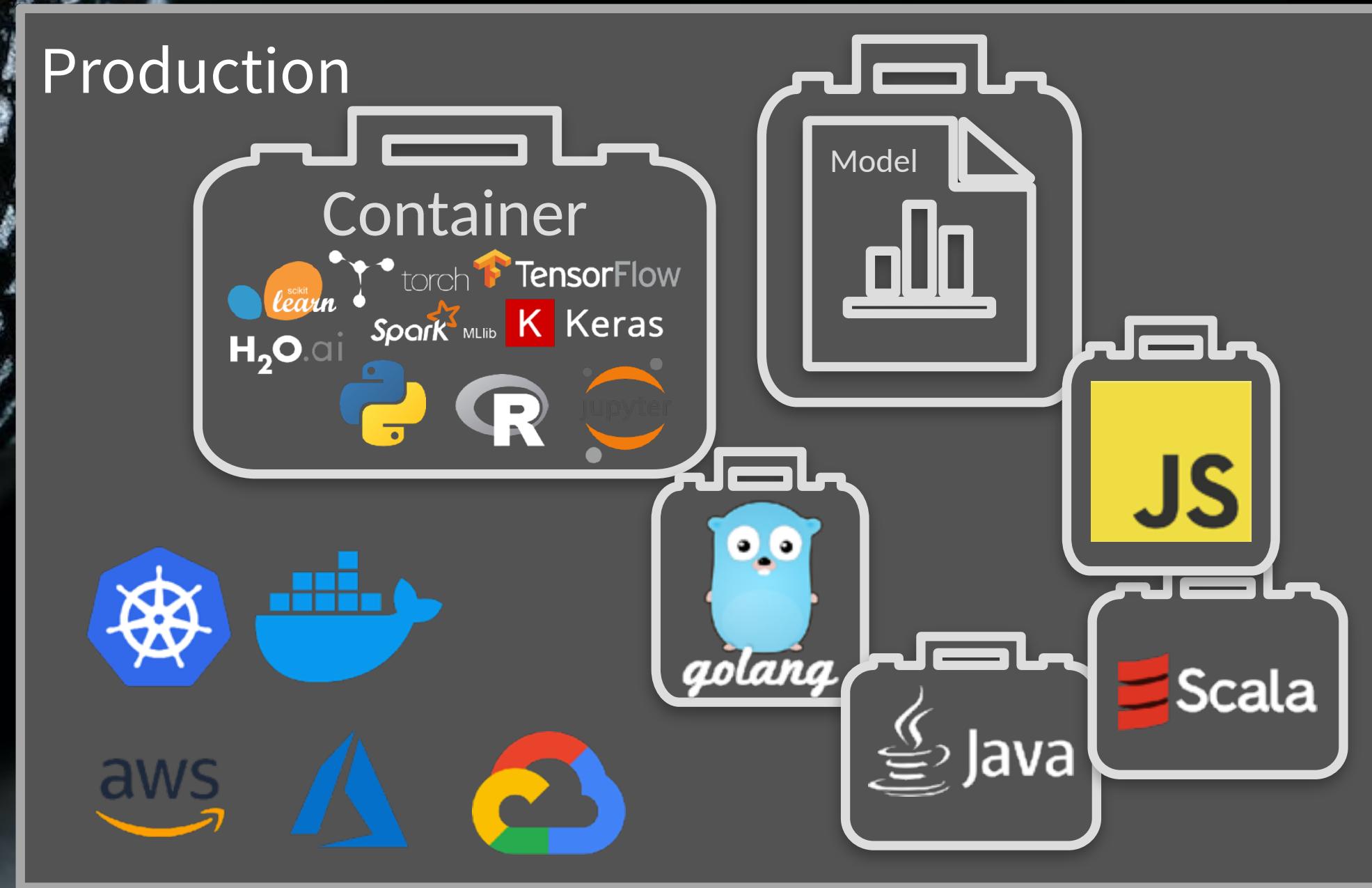
The Present

- We have to bridge the divide between data science and data engineering now.
- Or AI won't be an option.



The Future

- To fully benefit, we need to embrace:
 - Scalable compute
 - Hybrid cloud
 - Kubernetes & containers
 - New SW design and implementation tools and techniques



Questions?

dean@deanwampler.com
@deanwampler
polyglotprogramming.com/talks

