

Next Generation AI: Transitioning to the Continuous, Self-Learning Enterprise

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](#)Bristol Myers Squibb[®][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 »](#)

The screenshot shows the Domino platform's user interface. At the top, there are tabs for 'All', 'Last 100', and 'Last 50'. Below this is a plot area showing fluctuating lines in pink and yellow, labeled 'Acc: (0.95-0.98)' and 'AUC: (0.9-1)'. Below the plot is a 'Jobs Timeline' section with a table of active and completed jobs. One job, 'paramSearch.py -n 25 --loss exp', is highlighted. The table includes columns for 'No.', 'Title', 'Started', 'AUC', and 'Logs'. A preview of a plot titled 'results/AUC_ACC_exponential_5.png' is shown in the 'Logs' tab. The bottom of the interface has a search bar and a download button.



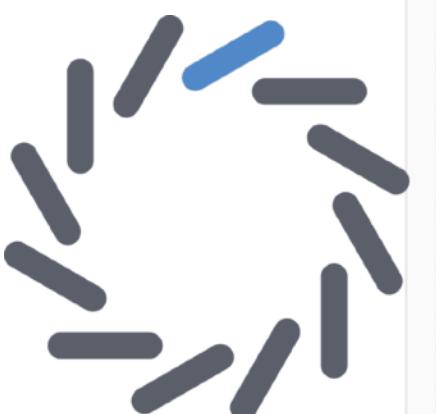
Centralize Infrastructure

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

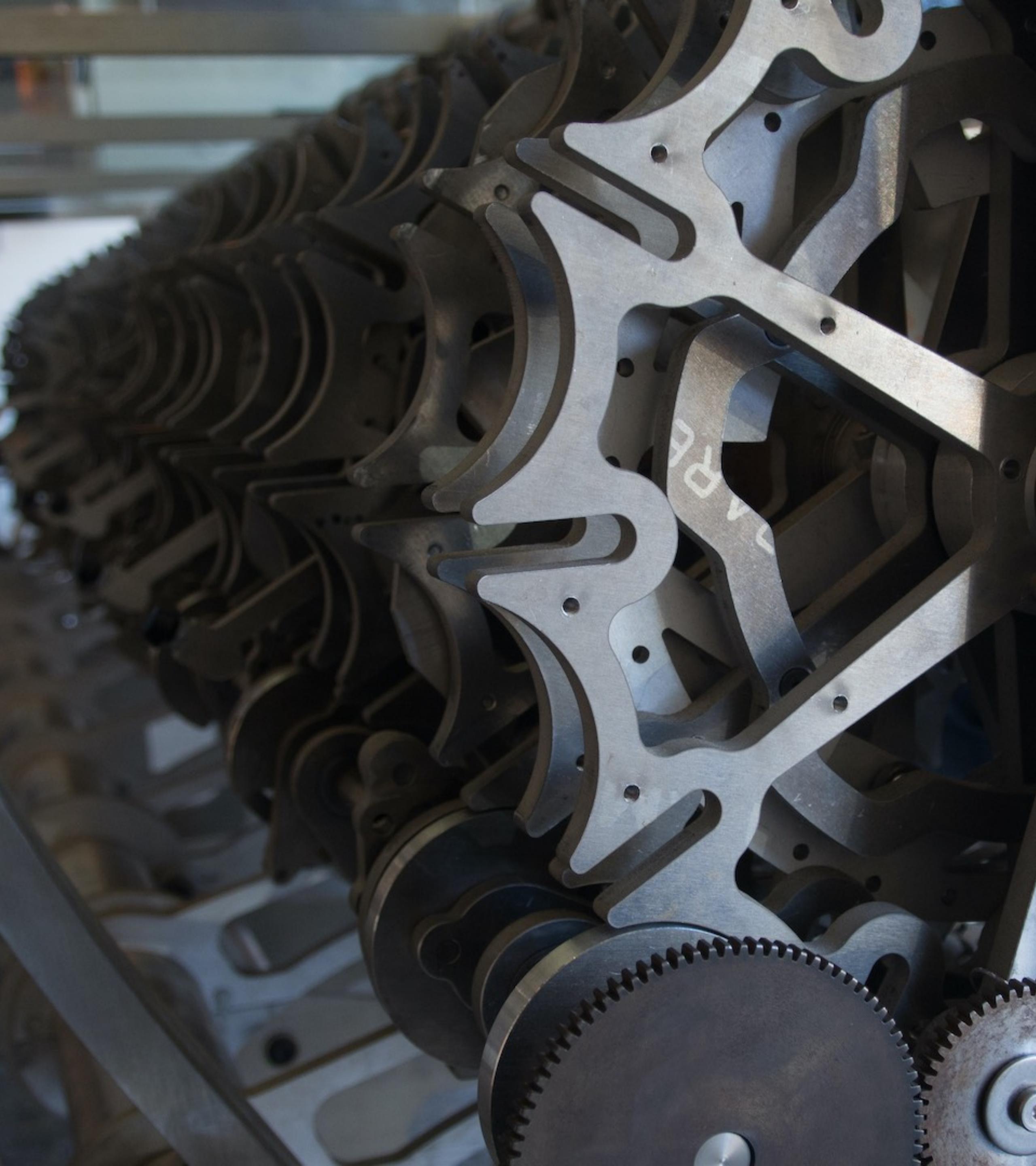
[Learn More »](#)

Deploy and Monitor Models

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



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



@deanwampler



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

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 window with the URL ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html. The page is from the Google AI Blog, featuring the Google logo and the text "The latest news from Google AI". The main article title is "Towards a Conversational Agent that Can Chat About... Anything", dated Tuesday, January 28, 2020. It is posted by Daniel Adiwardana, Senior Research Engineer, and Thang Luong, Senior Research Scientist, Google Research, Brain Team. The article discusses the limitations of modern chatbots and how open-domain dialog research explores a complementary approach to address them. On the right side of the page, there is a sidebar with a search bar, a "Labels" dropdown, an "Archive" dropdown, a "Feed" link, a "Follow @googleai" button, and a "Give us feedback in our Product Forums" link. A red box highlights the first sentence of the article's summary.

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.

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

Such improvements are a critical flaw – they often don't make sense. They with what has been said so far, or lack common sense

and 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



Naturalistic text to speech

- Needed for dialog generation



Translation

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



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. It includes author information from Arizona State University, Charles River Analytics, and Michigan State University, along with their email addresses. The abstract discusses the use of social media for news consumption and its double-edged nature.

**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



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.

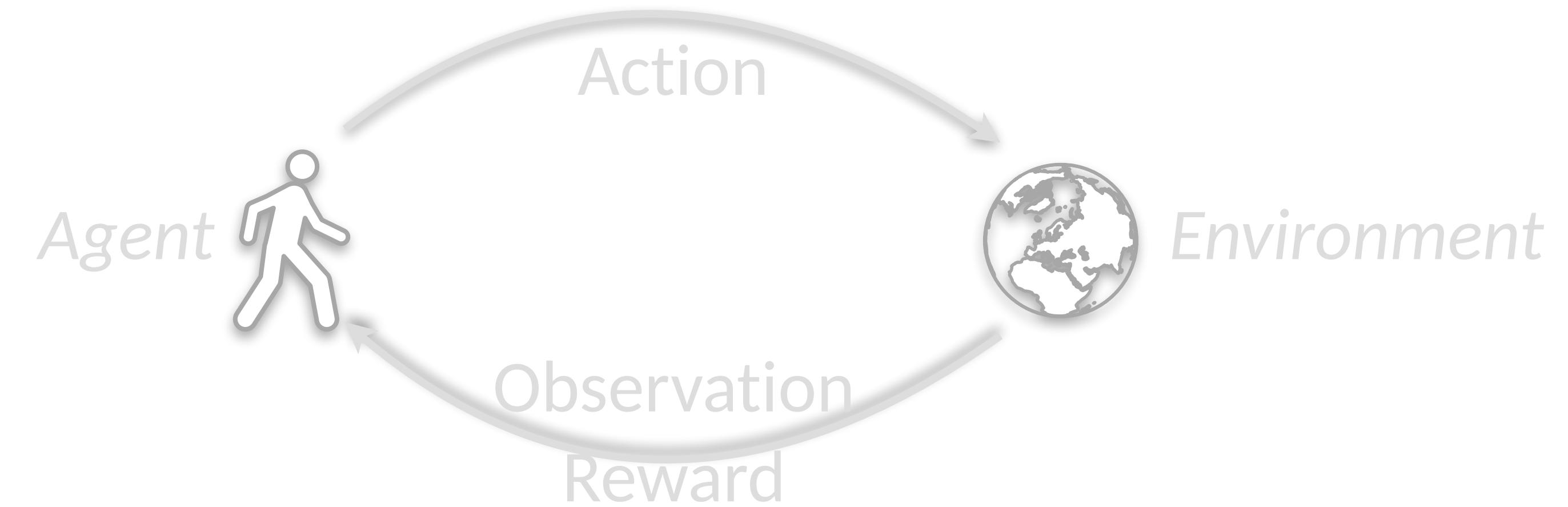


Reinforcement Learning



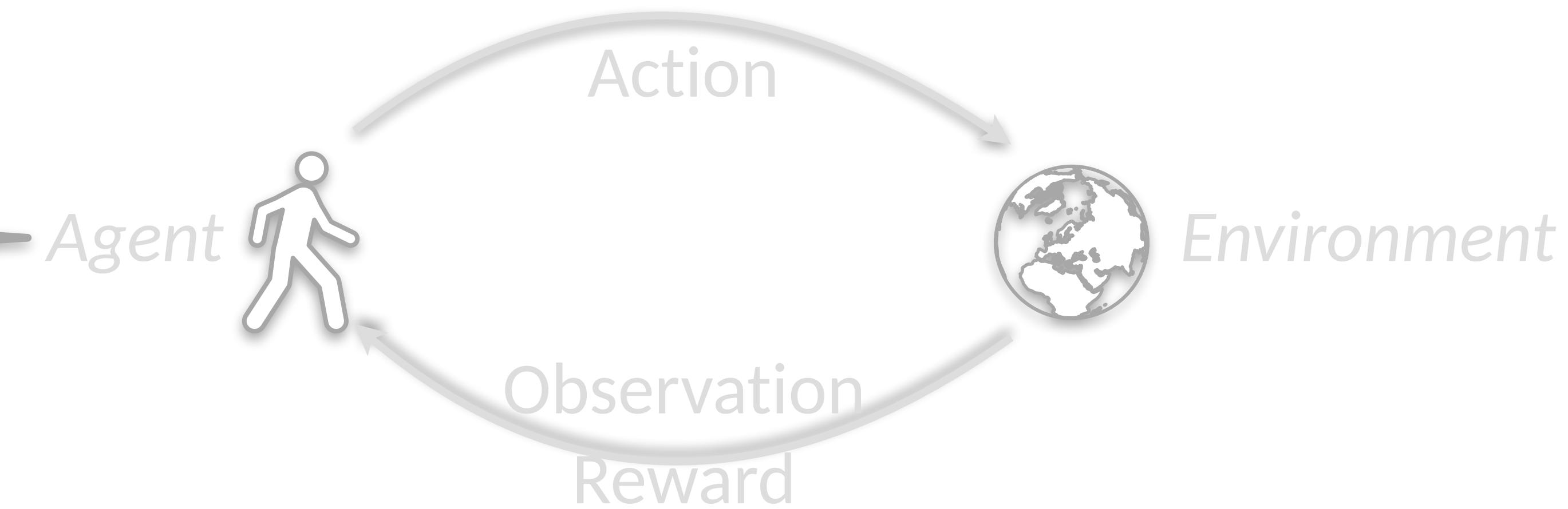
What Is RL?

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



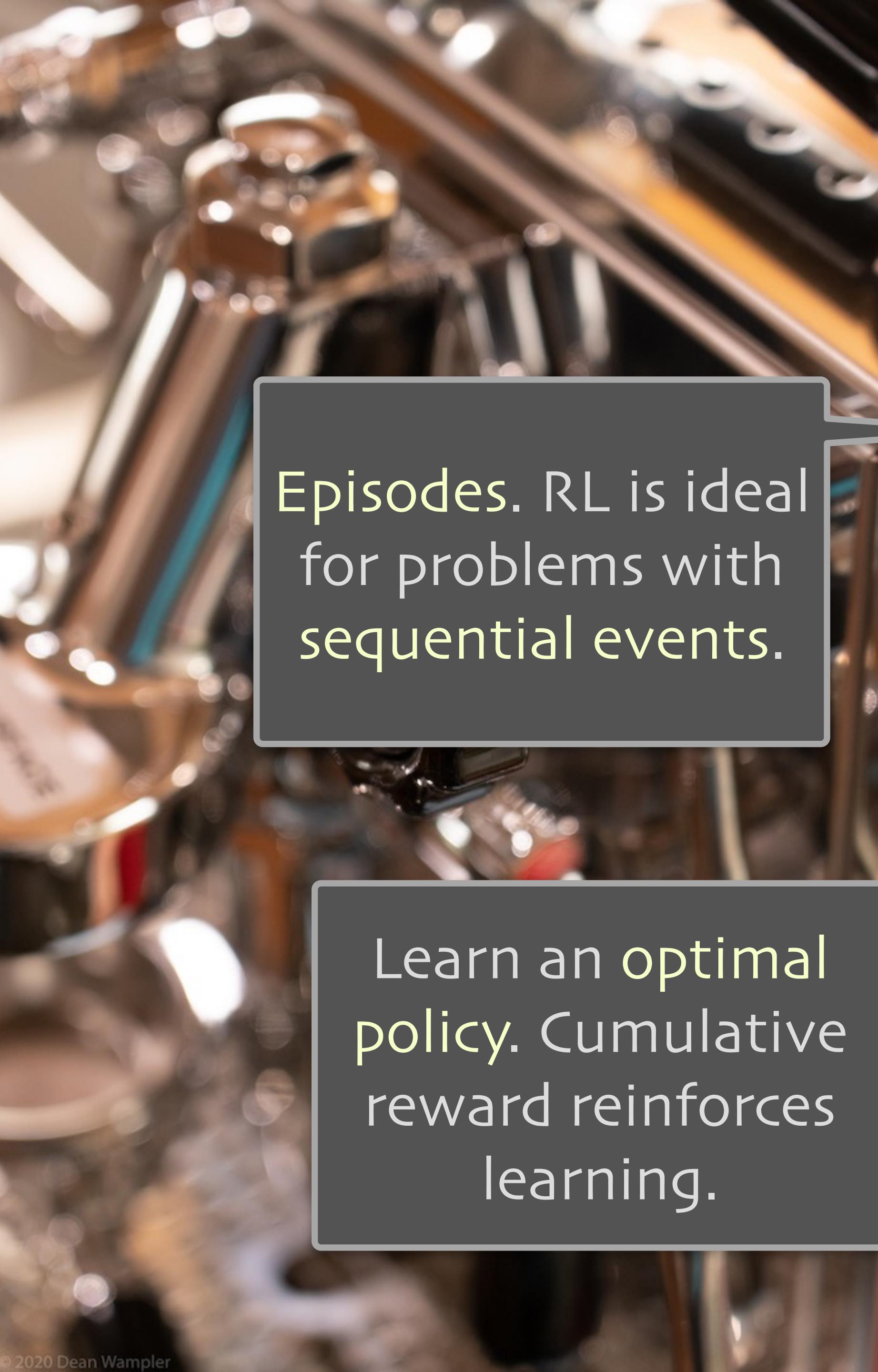
What Is RL?

- 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

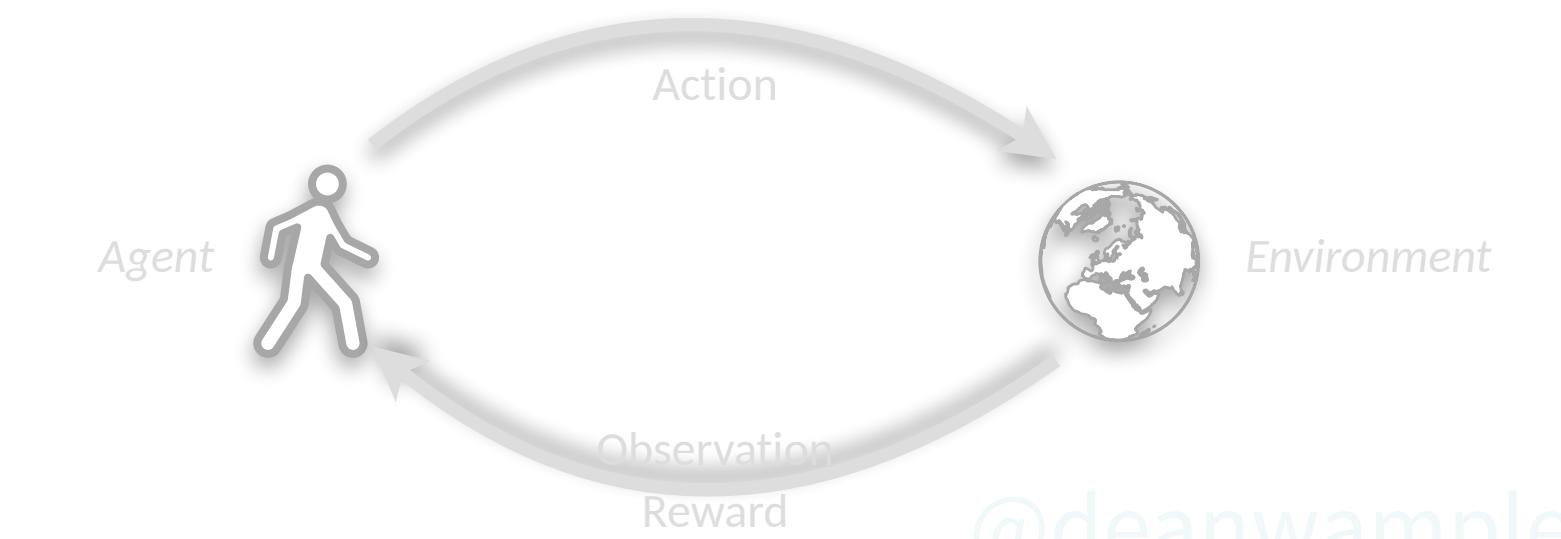
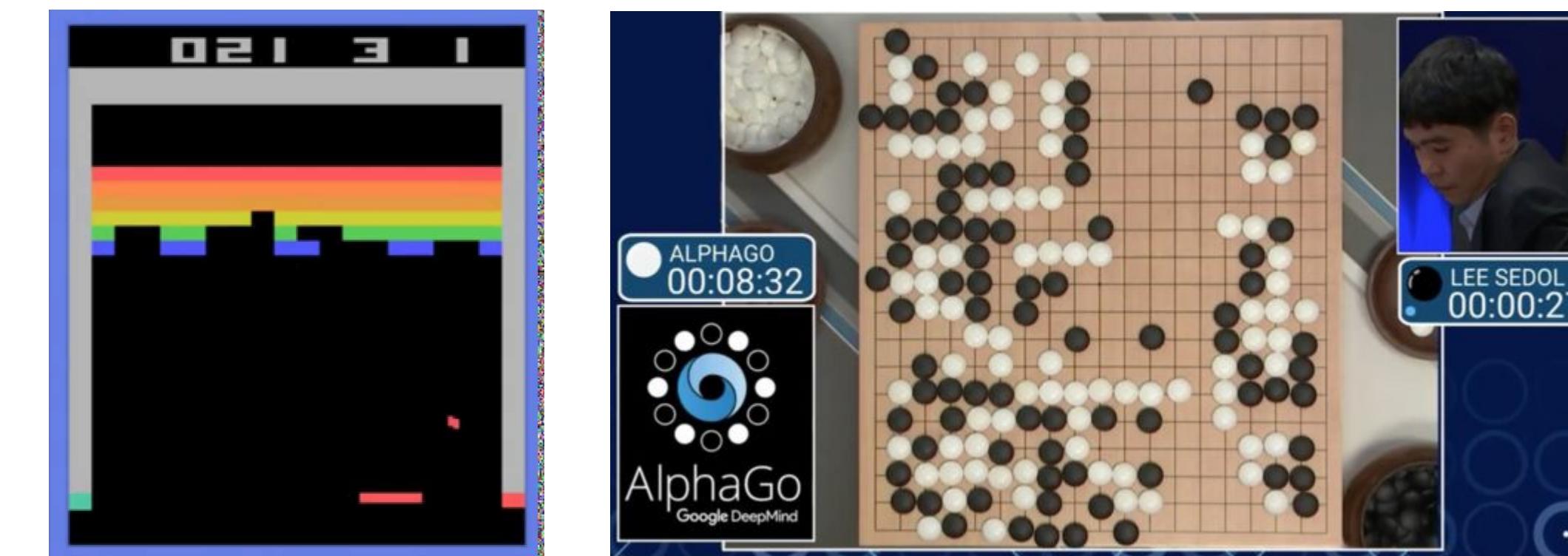
- Games
- Robots & Autonomous Vehicles
- Process Modeling & Automation
- System Optimization
- Advertising & Recommendation
- Markets





Games

- World's best expert game play in:

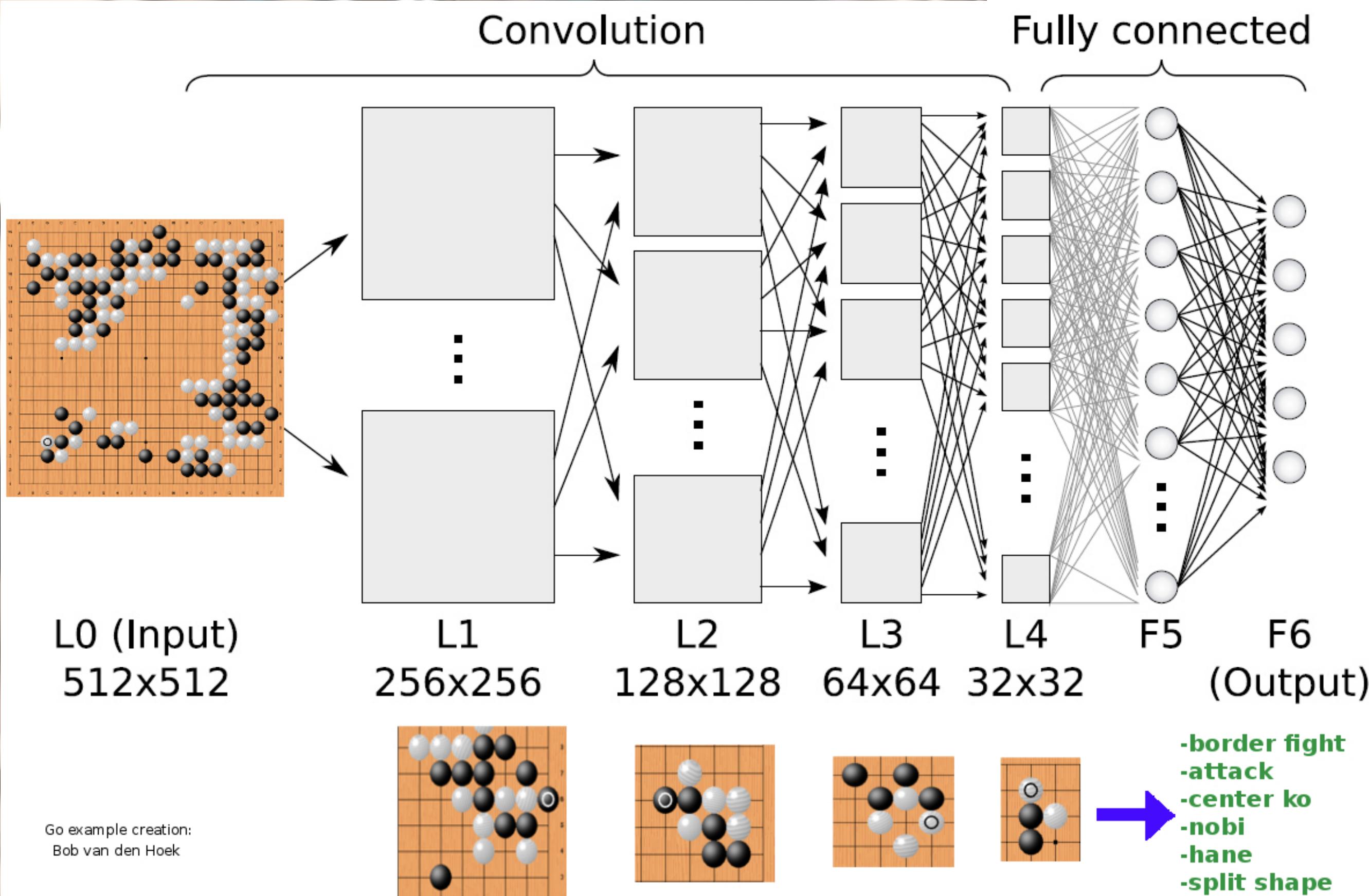


<https://www.geekwire.com/2016/alphago-ai-program-wins-1-million-prize-go-showdown-champion-lee-sedol/>

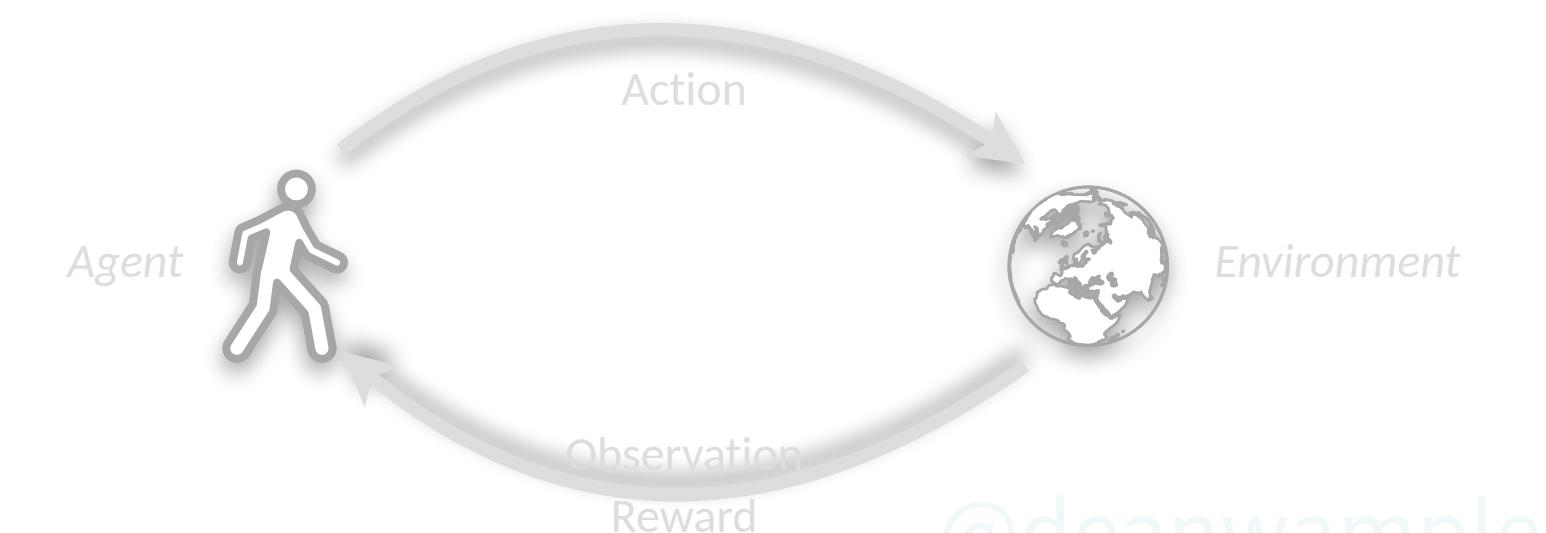
<https://towardsdatascience.com/tutorial-double-deep-q-learning-with-dueling-network-architectures-4c1b3fb7f756>

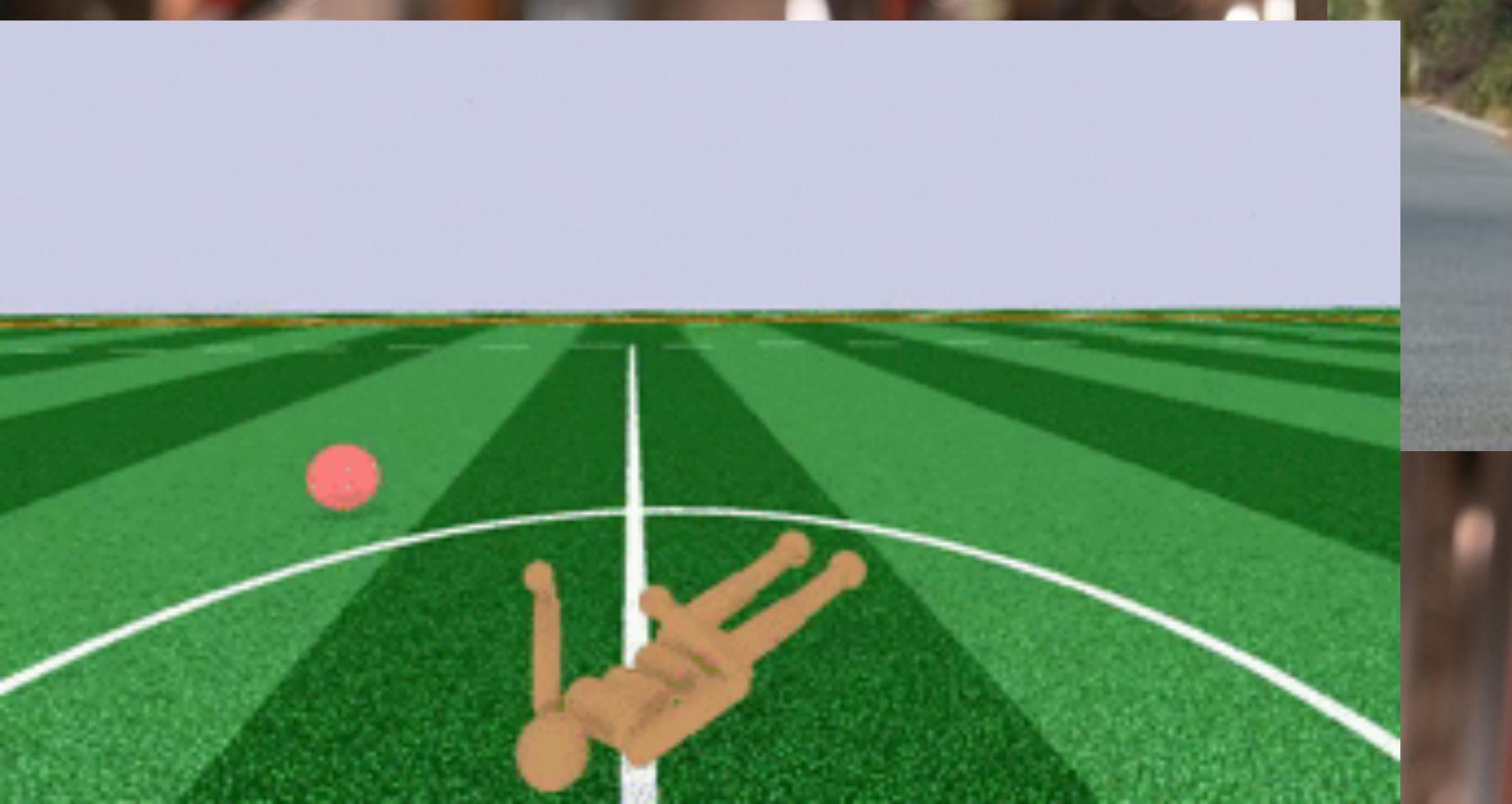


Games



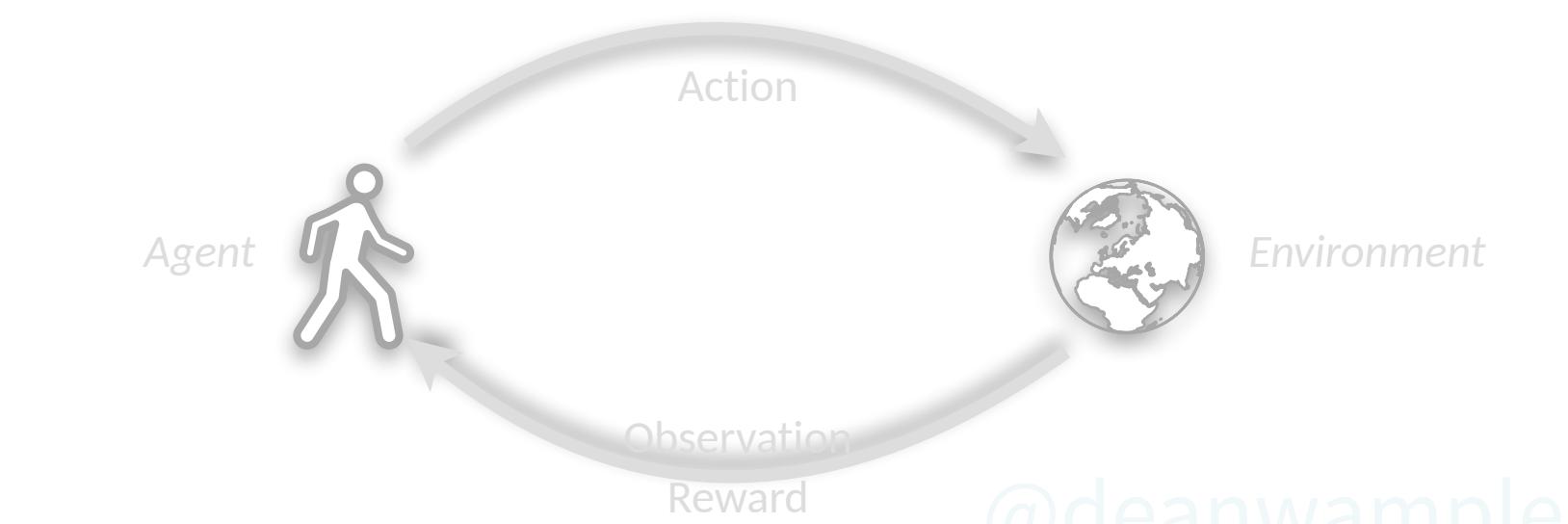
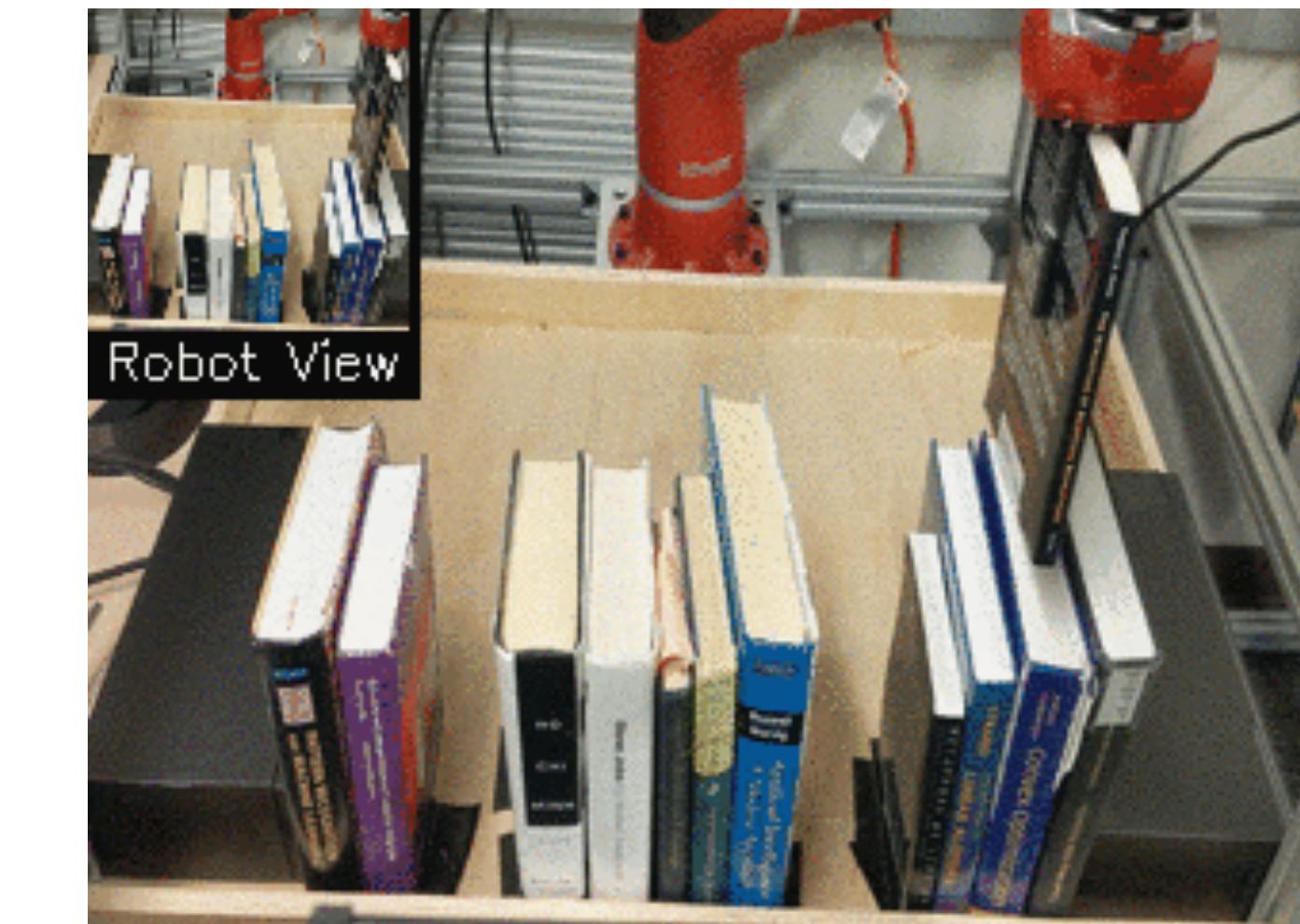
- AlphaGo
- Observations: board state
- Actions: place stones
- Rewards:
 - 1 if you win
 - 0 otherwise





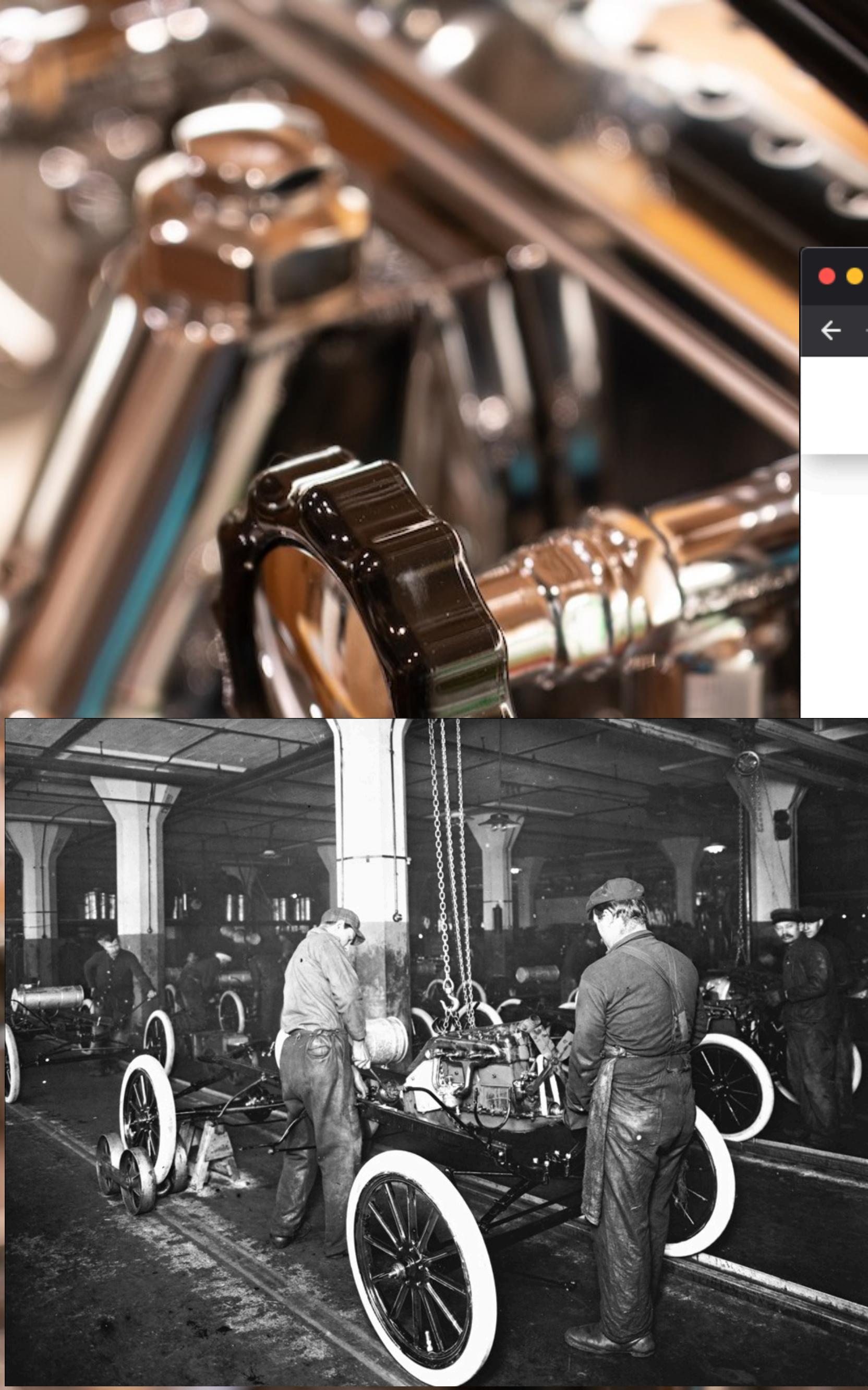
Robotics & Autonomous Vehicles

- Start with simulators, work up to real machines.



@deanwampler

Process Modeling & Automation

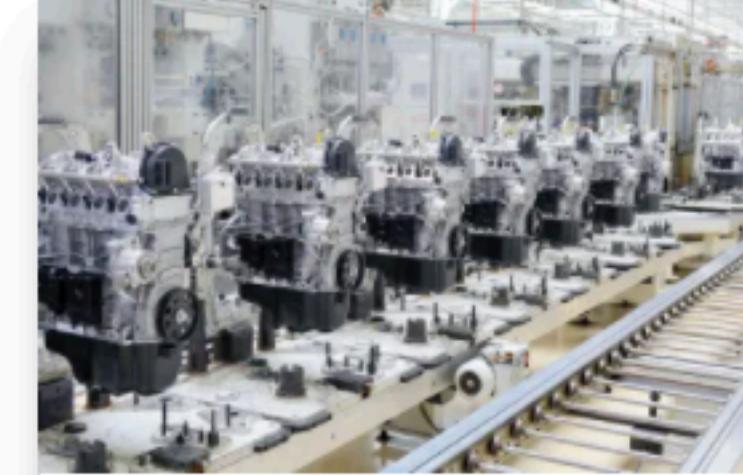


Simulation Optimization | Add / x +

pathmind.com

pathmind 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...

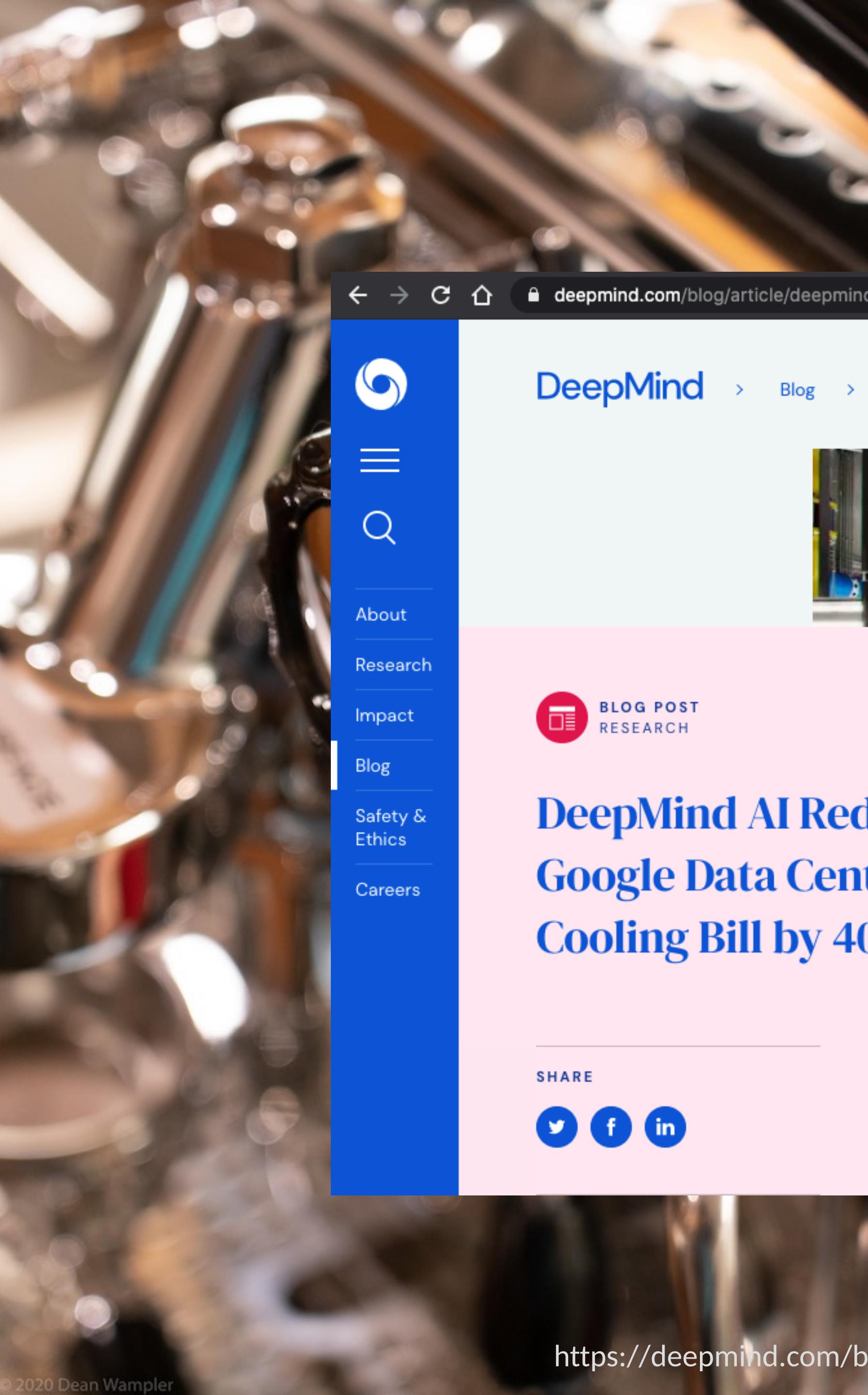


Engineering Group: Using AI to Maximize Factory Output with Better Order Sequencing
Oct 29, 2020 | [Customer Success](#)
Summary Engineering Group, a global engineering firm and technology consultancy with a strong practice in simulation, worked with Pathmind to apply reinforcement learning to maximize 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 July 20, 2016, and is categorized under "BLOG POST RESEARCH". The main image shows a large, brightly lit data center with extensive red, blue, and yellow piping systems. The DeepMind navigation bar on the left includes links for About, Research, Impact, Blog, Safety & Ethics, and Careers.

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

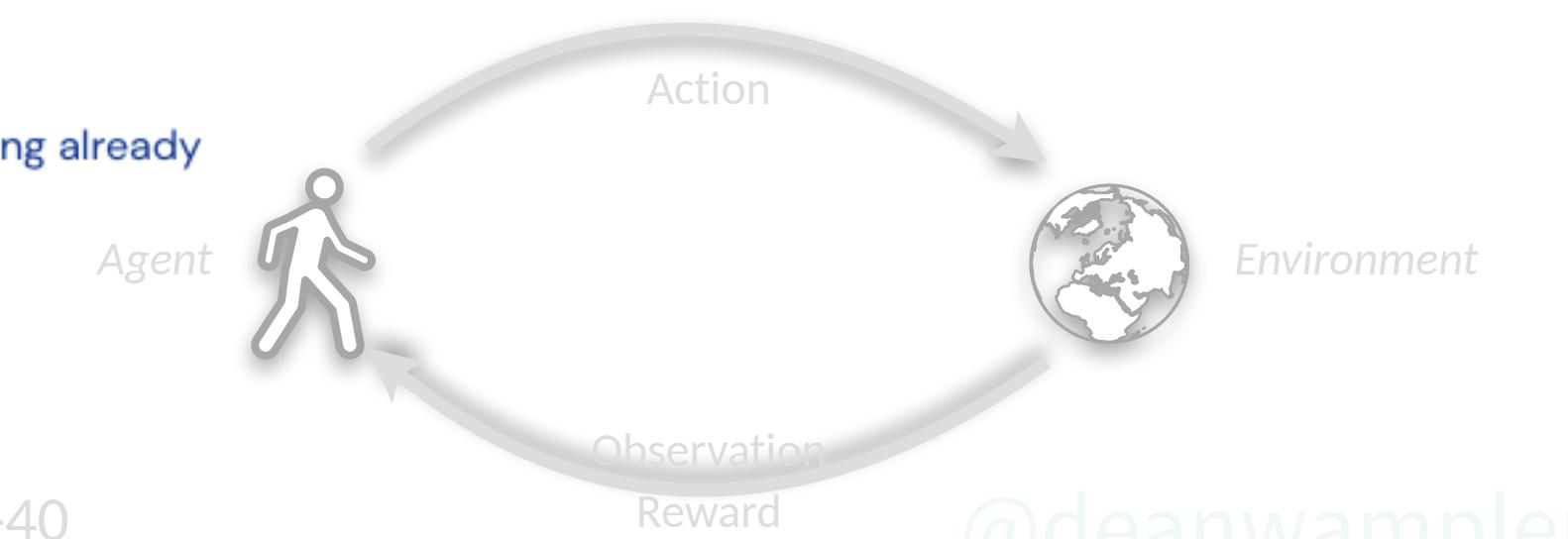
20 JUL 2016

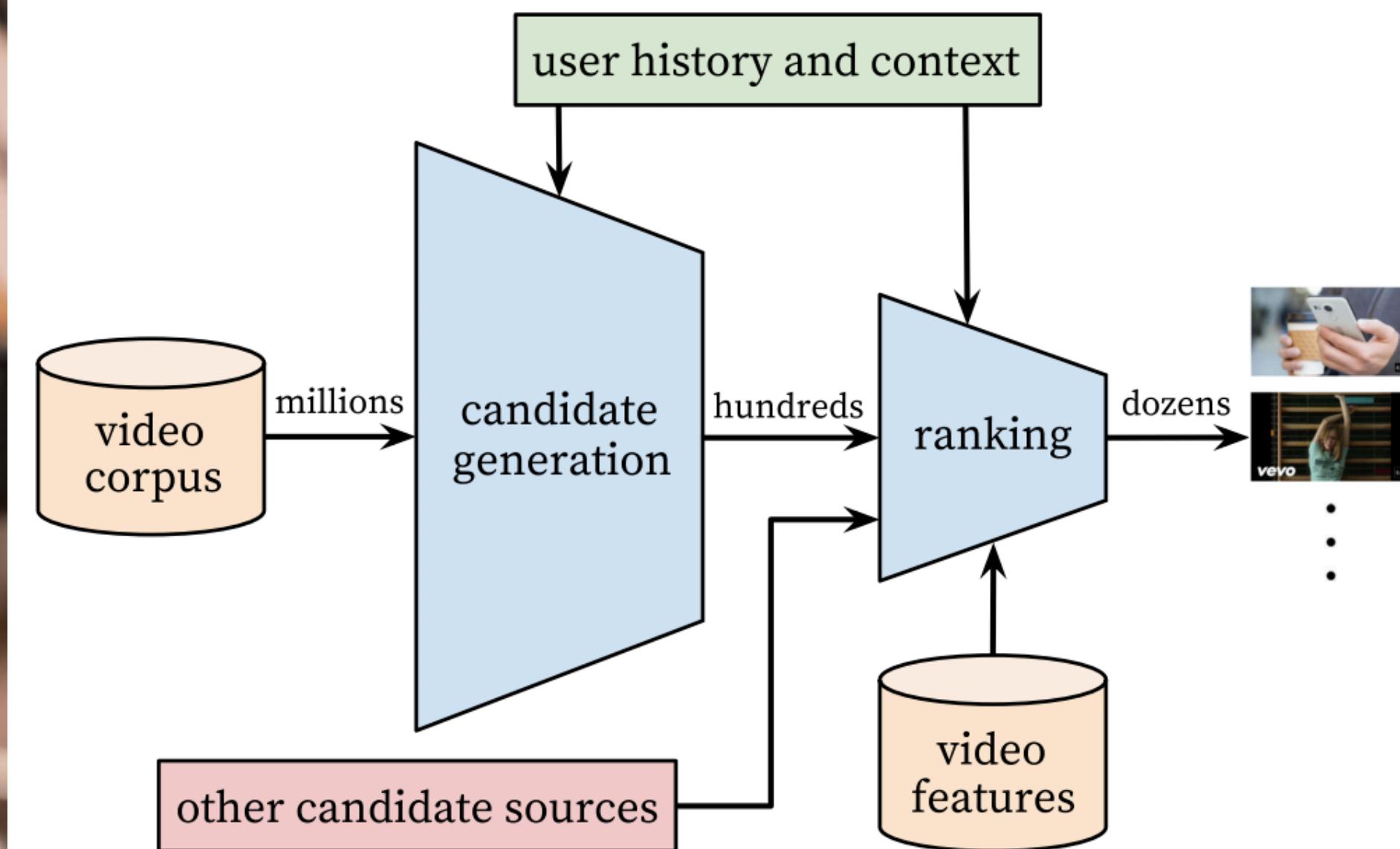
BLOG POST RESEARCH

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>

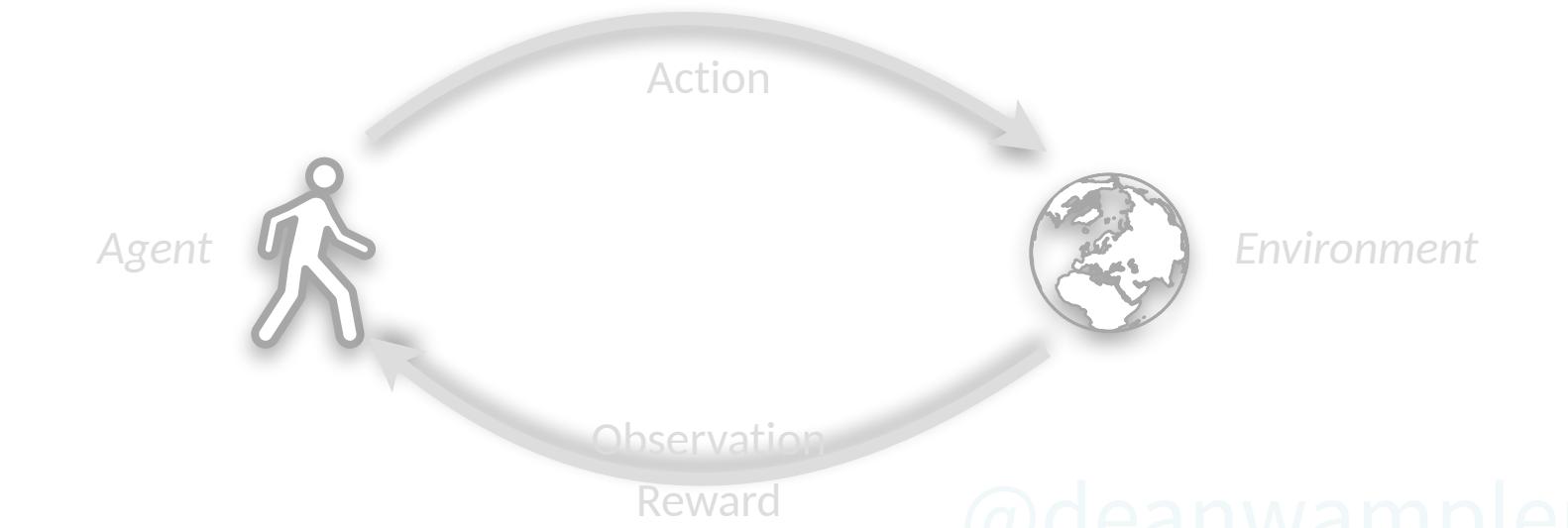




<https://research.google/pubs/pub45530/>

Advertising & Recommendation

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

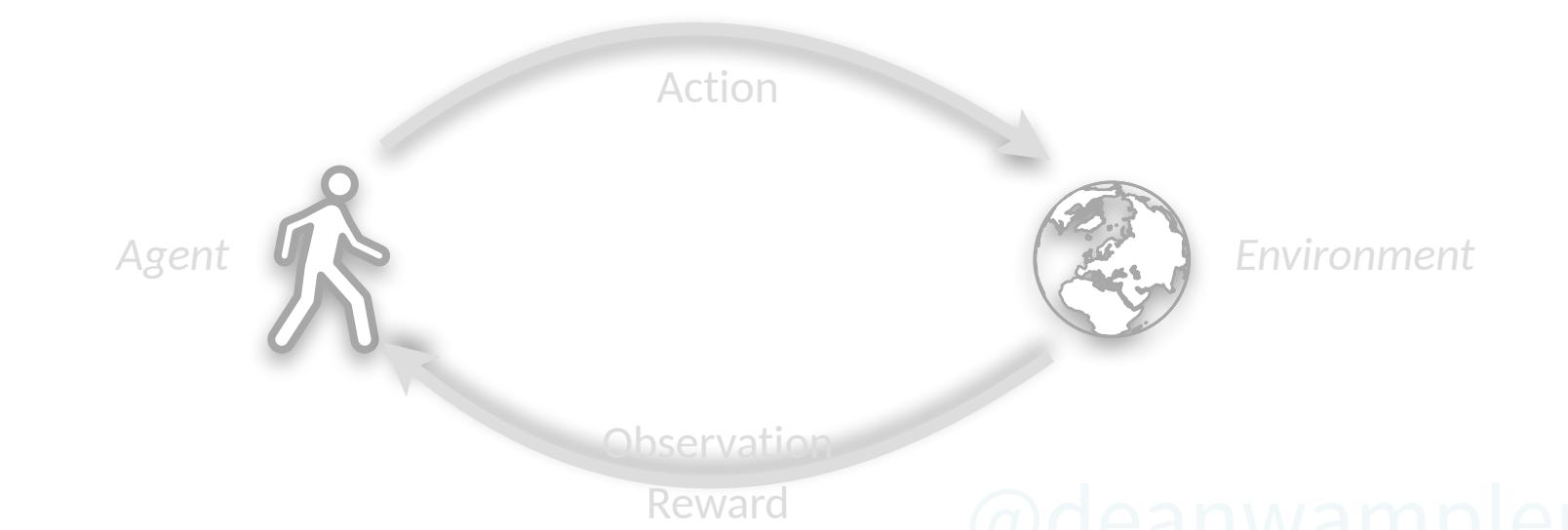


@deanwampler



Markets

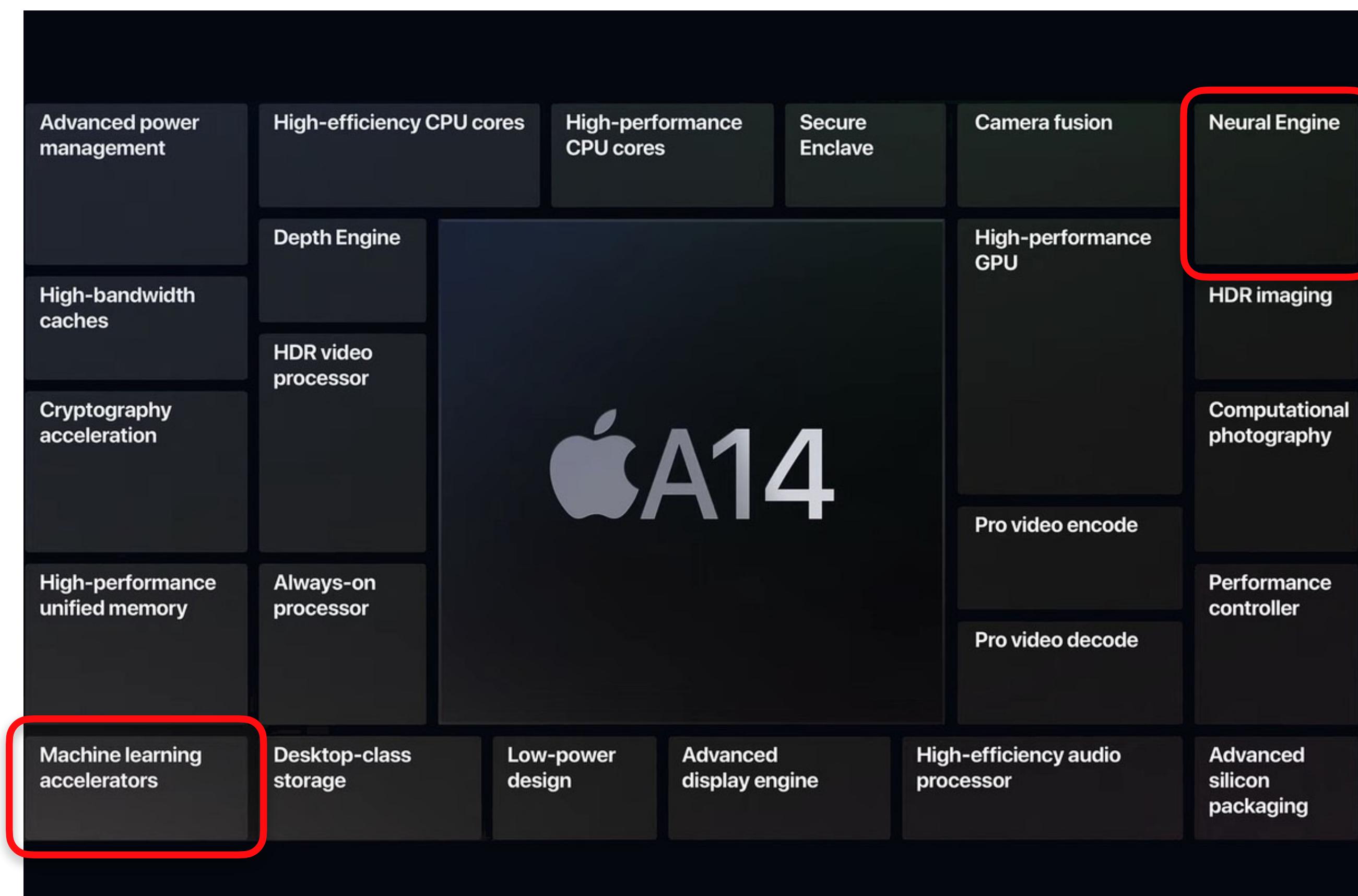
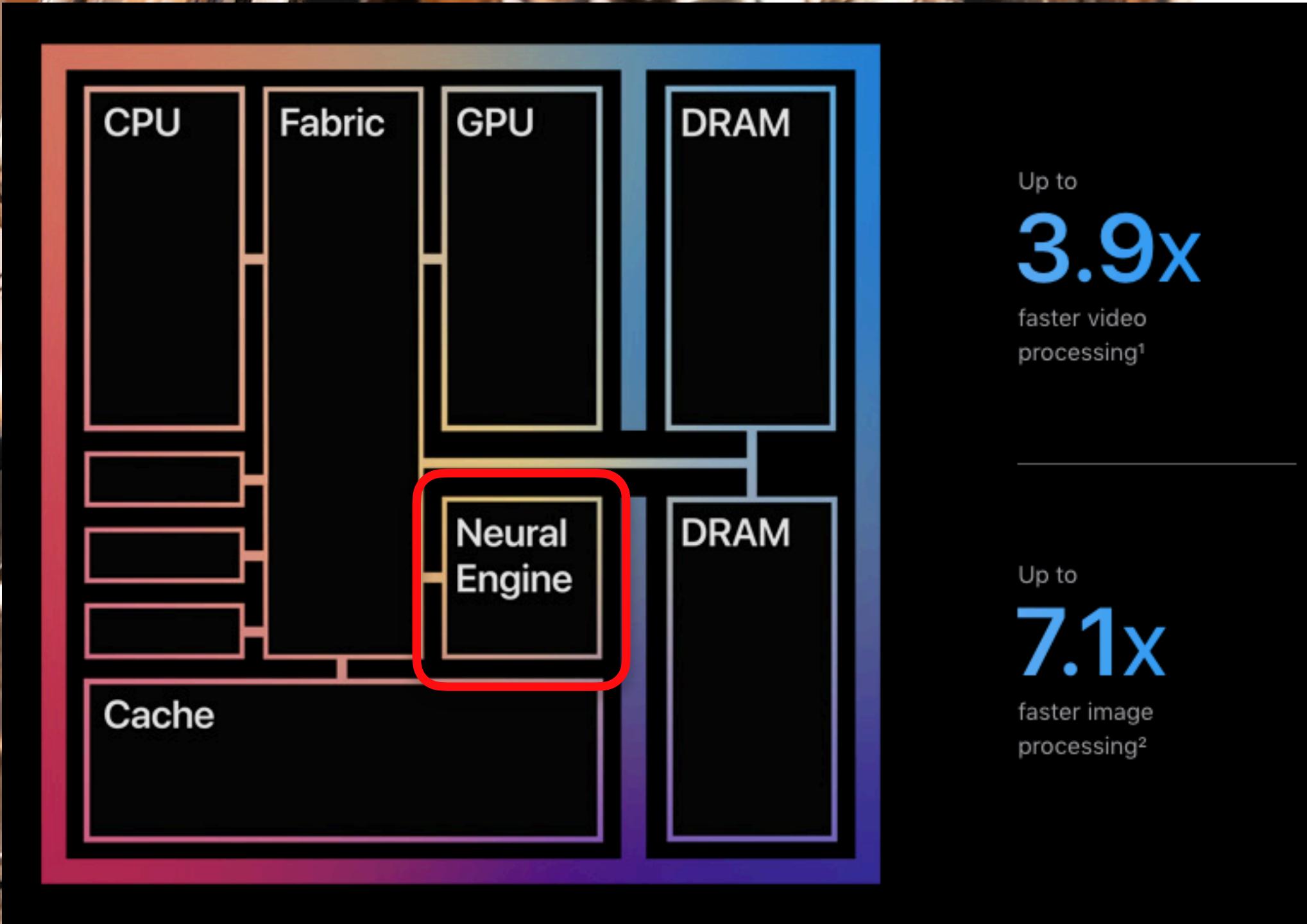
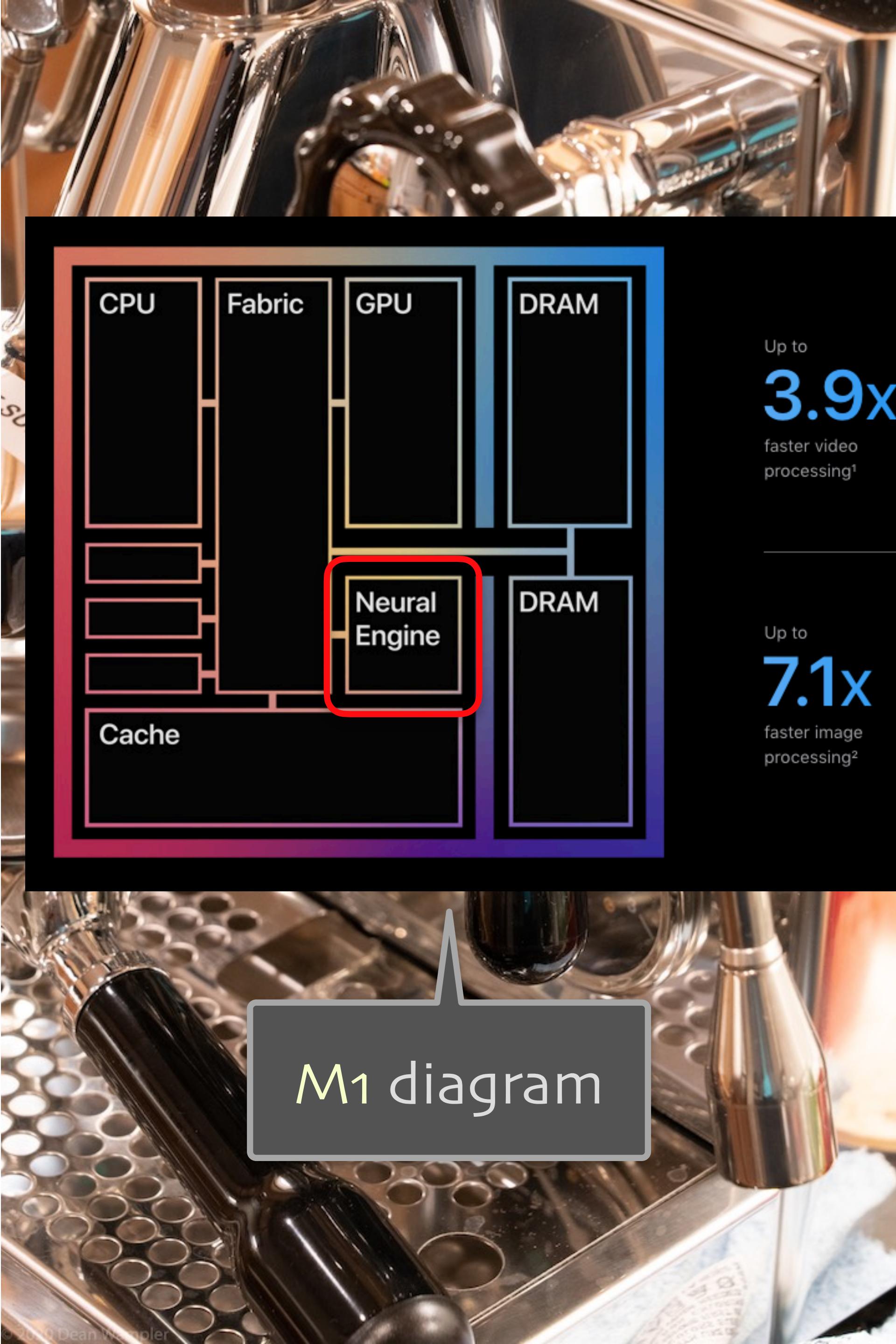
- Inherently time-ordered
- Lots of different “signals”
- **Contextual, multi-armed bandits**





What Our Phones
Are Telling Us...

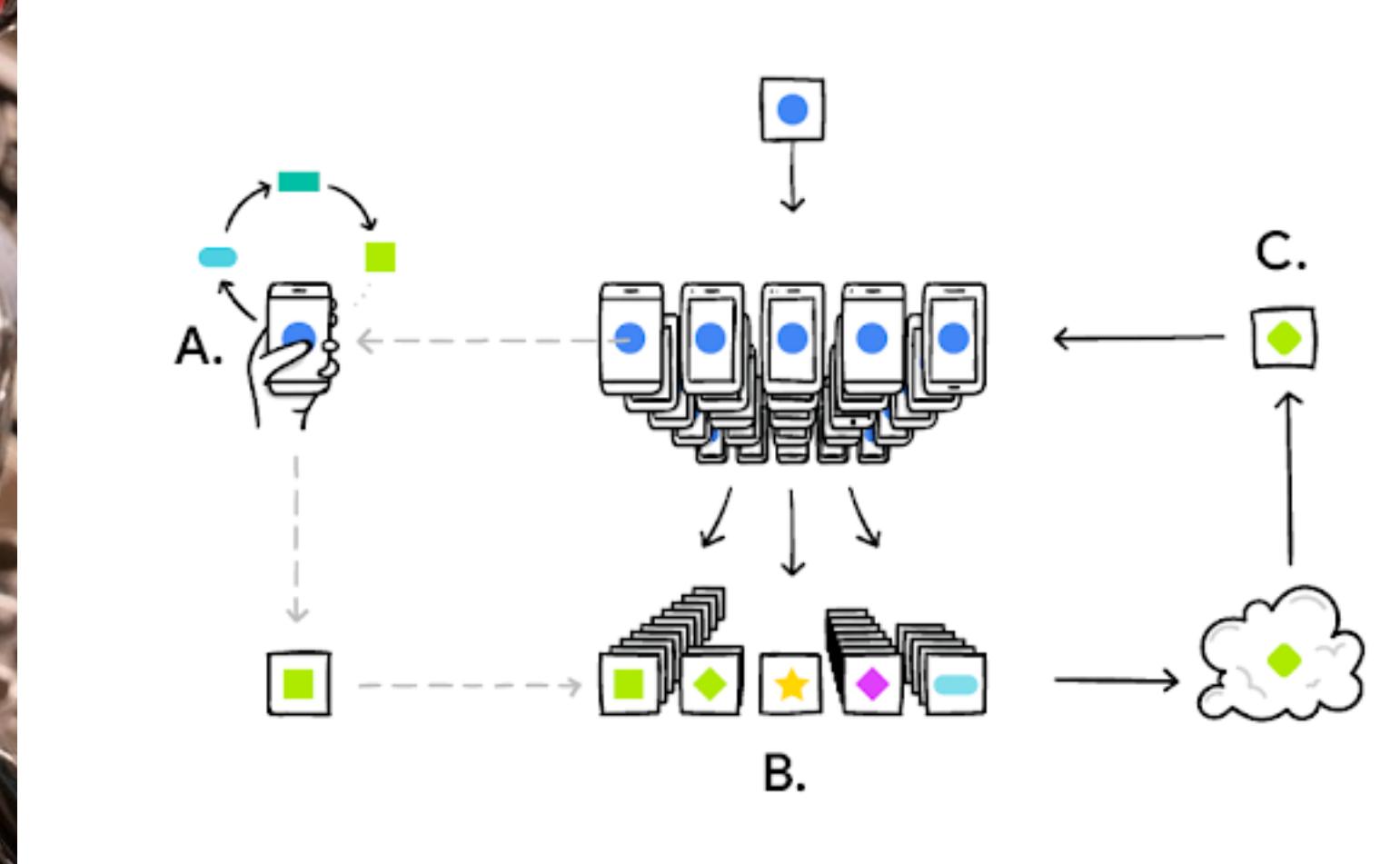
Apple Silicon





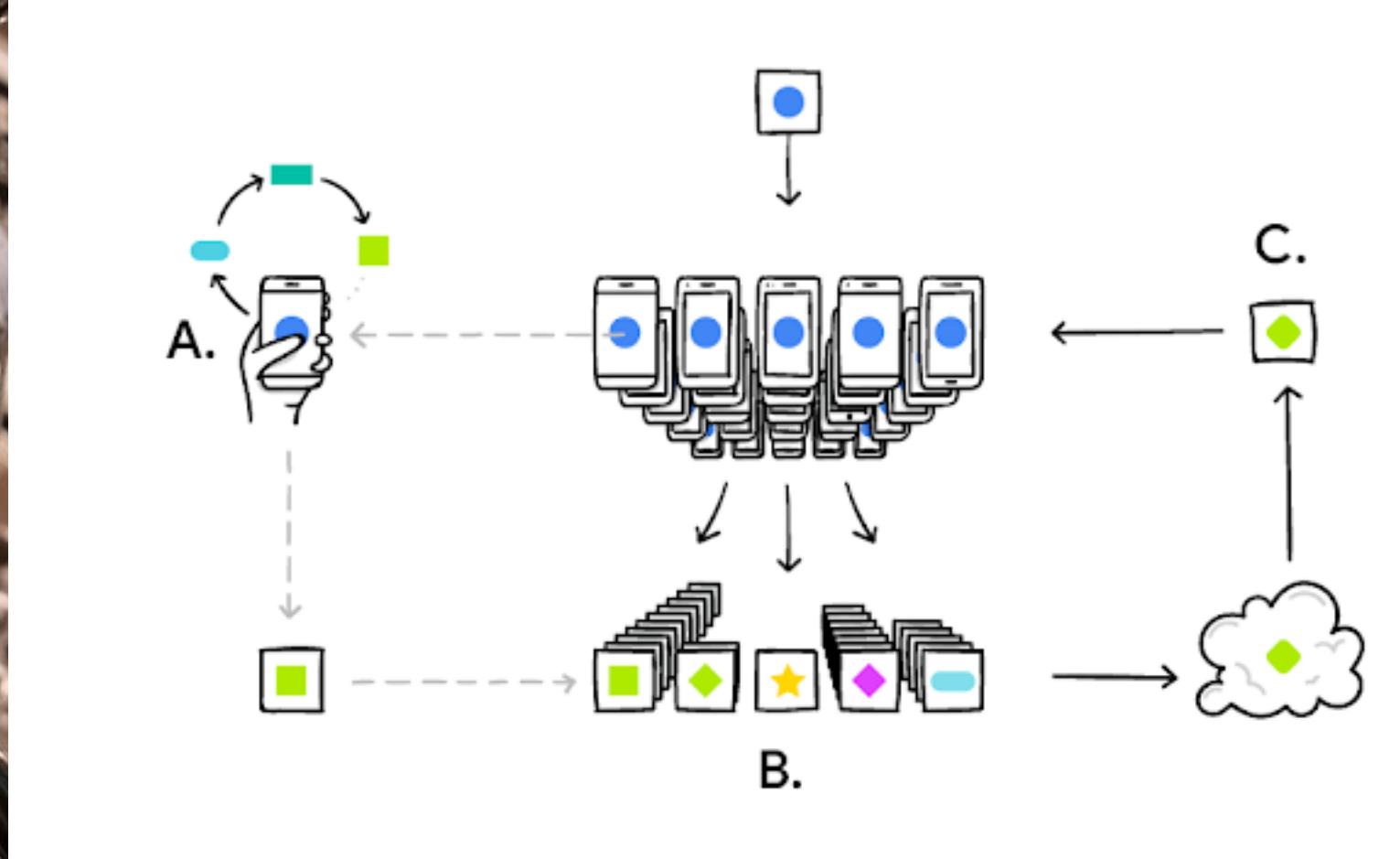
Applications that Exploit ML/AI

- Unlocking: finger and face ID
- Predictive typing
- Siri
- Recommendations
- ...
- Probably most will use it in one or another, sooner or later!



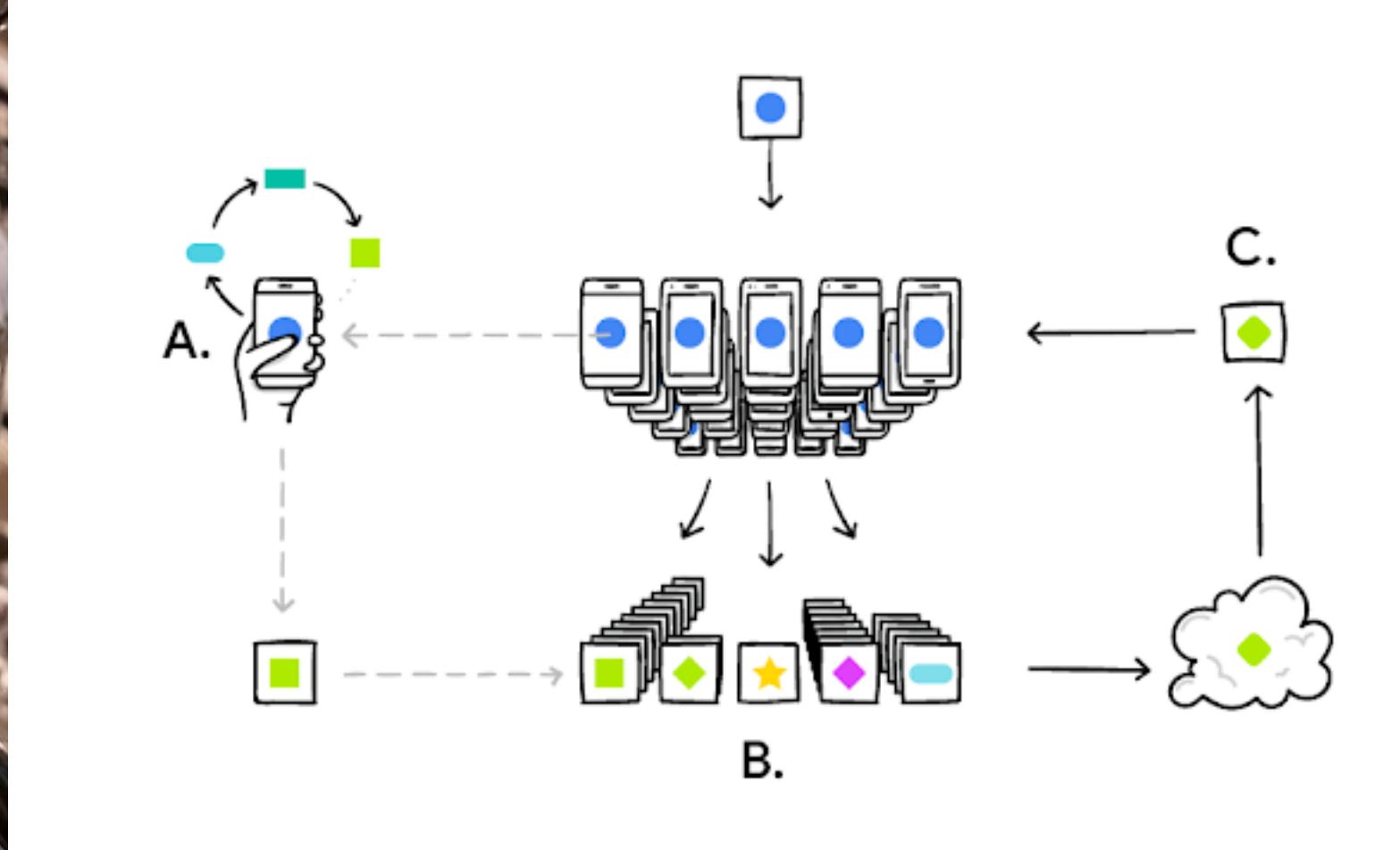
Technologies that Make this Possible

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



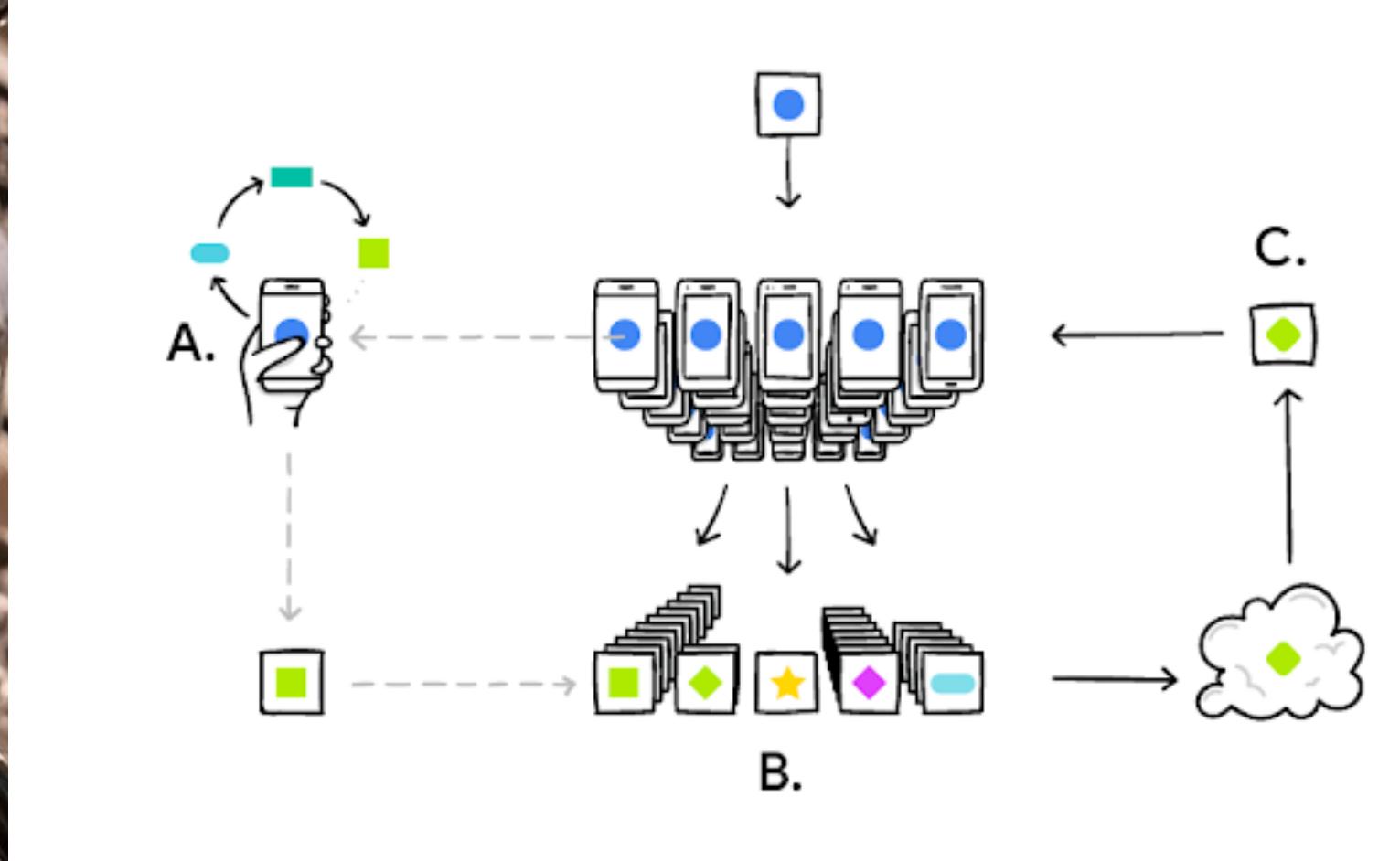
Technologies that Make this Possible

- Federated Learning **Advantages**
 - Your private data stays local
 - Local model is fine tuned for you
 - No central data storage required
 - Central processing is minimized
 - Instead, all our phones do most of the training



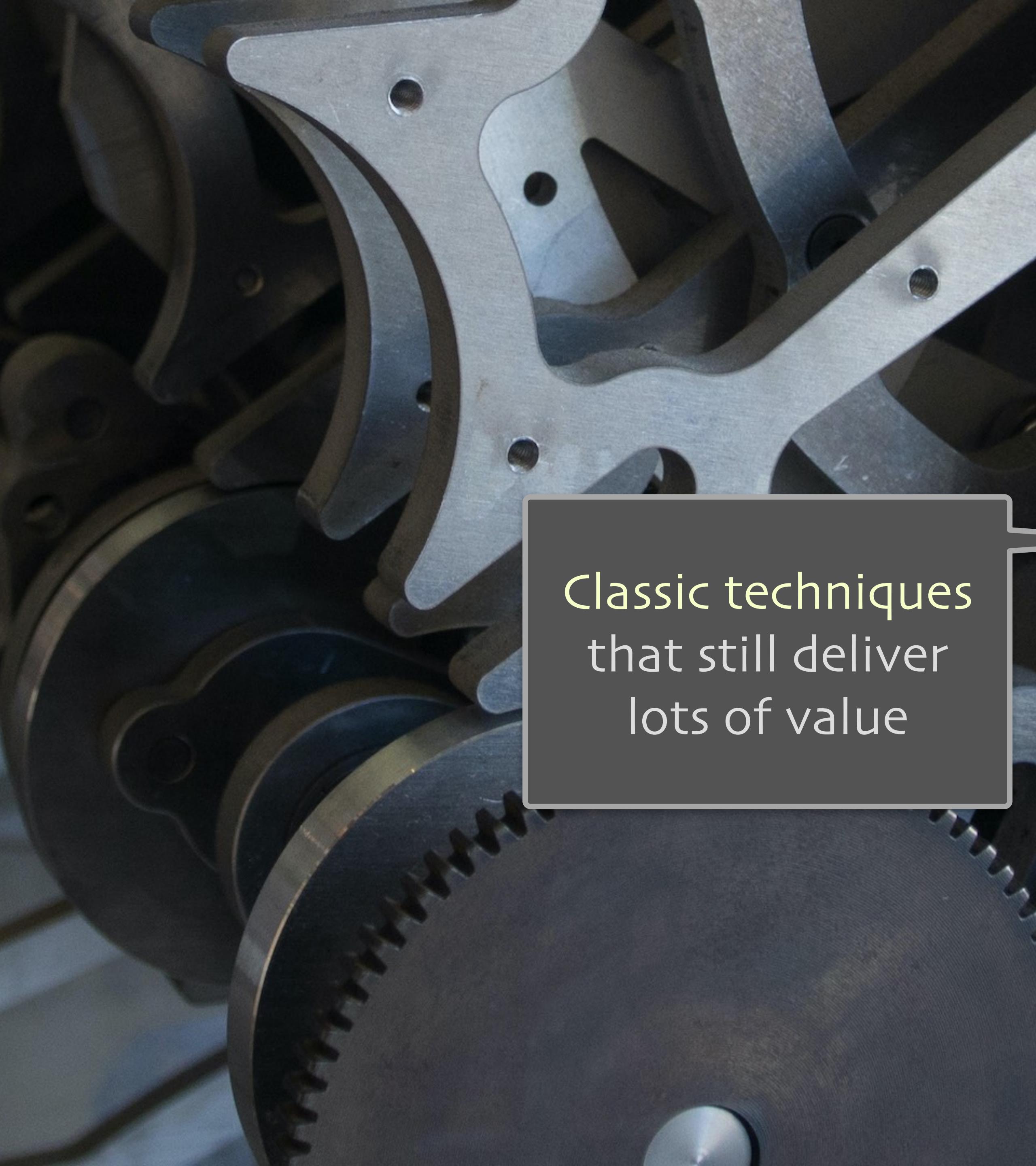
Technologies that Make this Possible

- **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



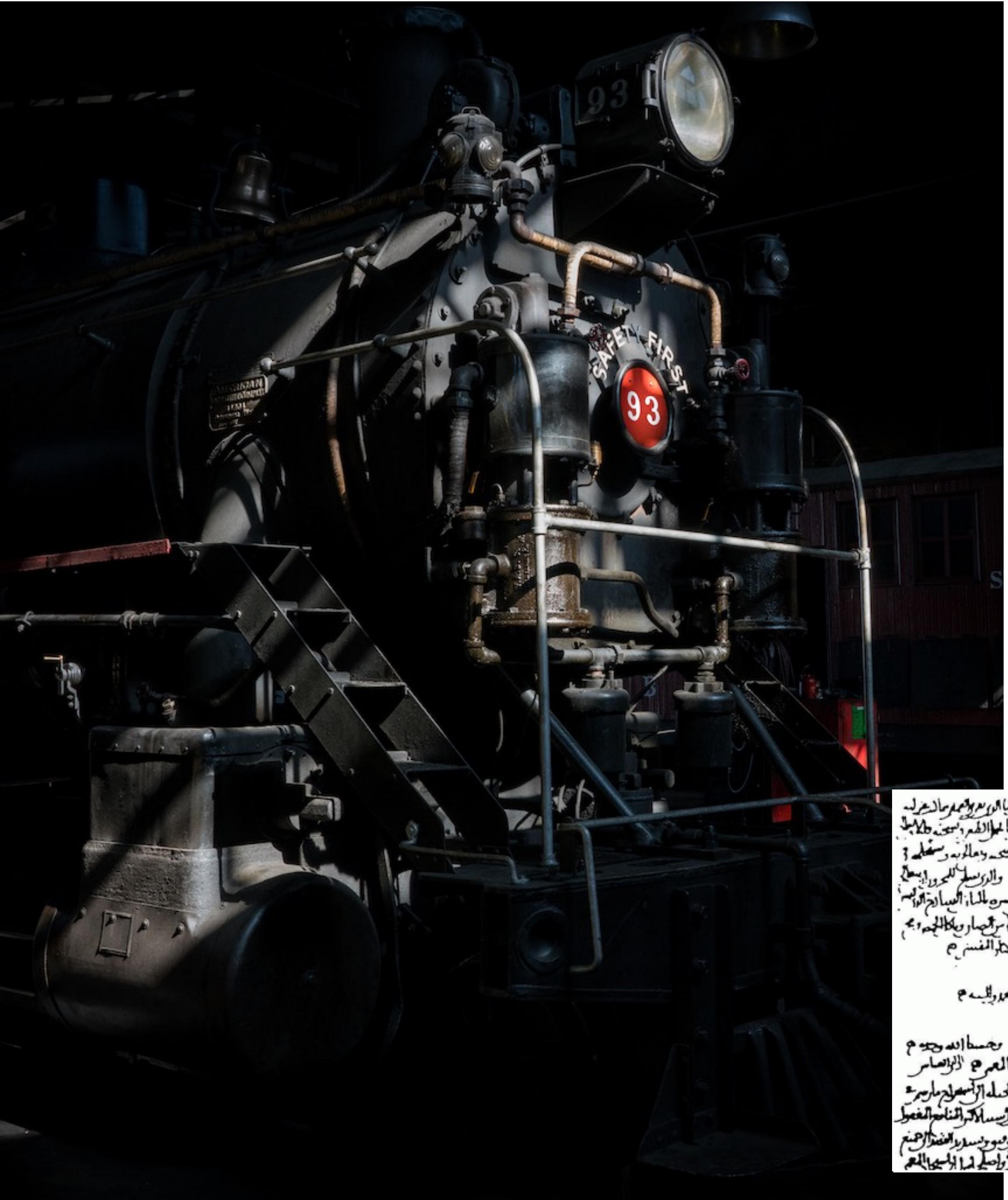
Enterprise Applications?

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



Outline

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



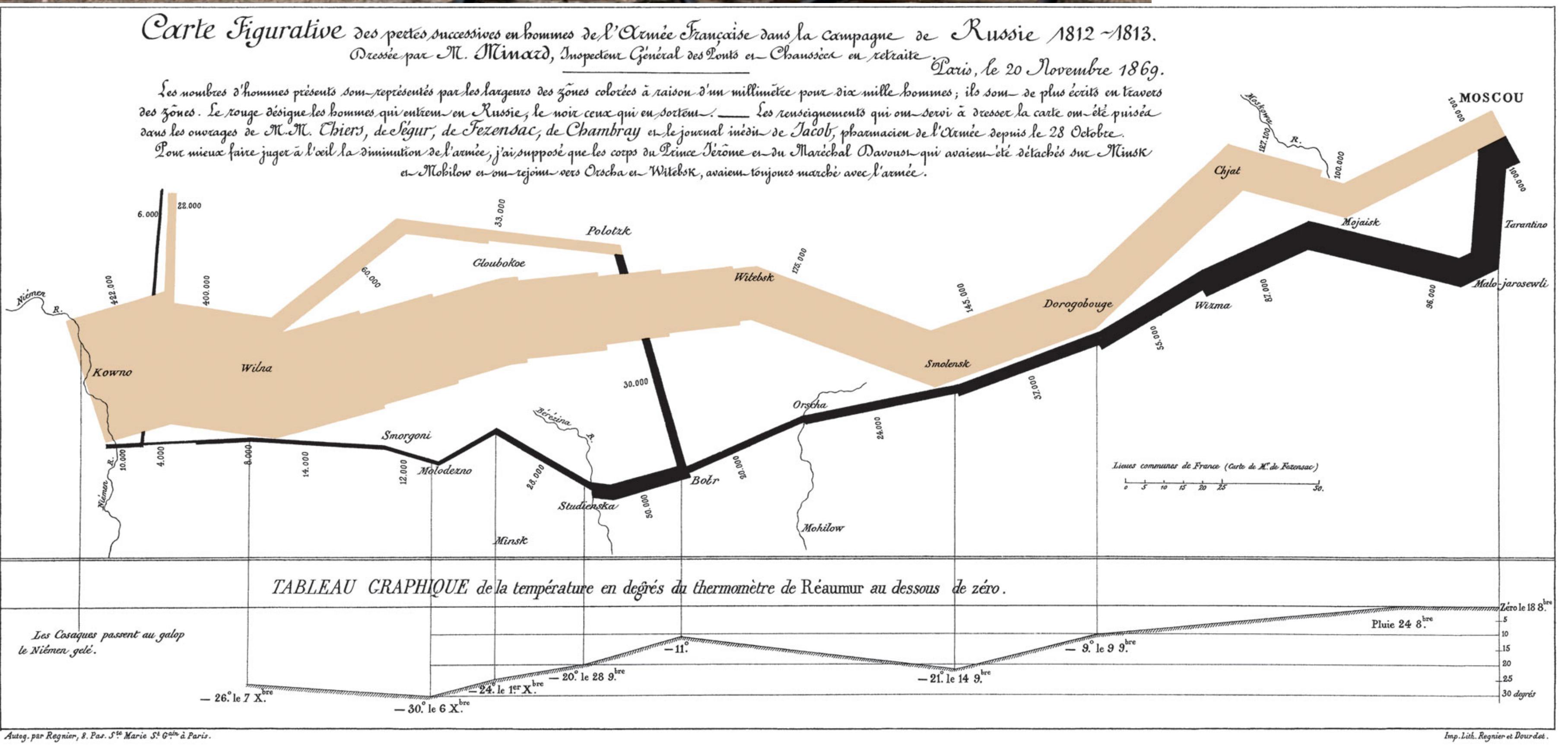
Statistical Inference

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





<https://datavizblog.com/2013/05/26/dataviz-history-charles-minards-flow-map-of-napoleons-rus>



Visualization

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



Visualization

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



Neural Nets

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

Handwritten Zip Code Recognition with Multilayer Networks

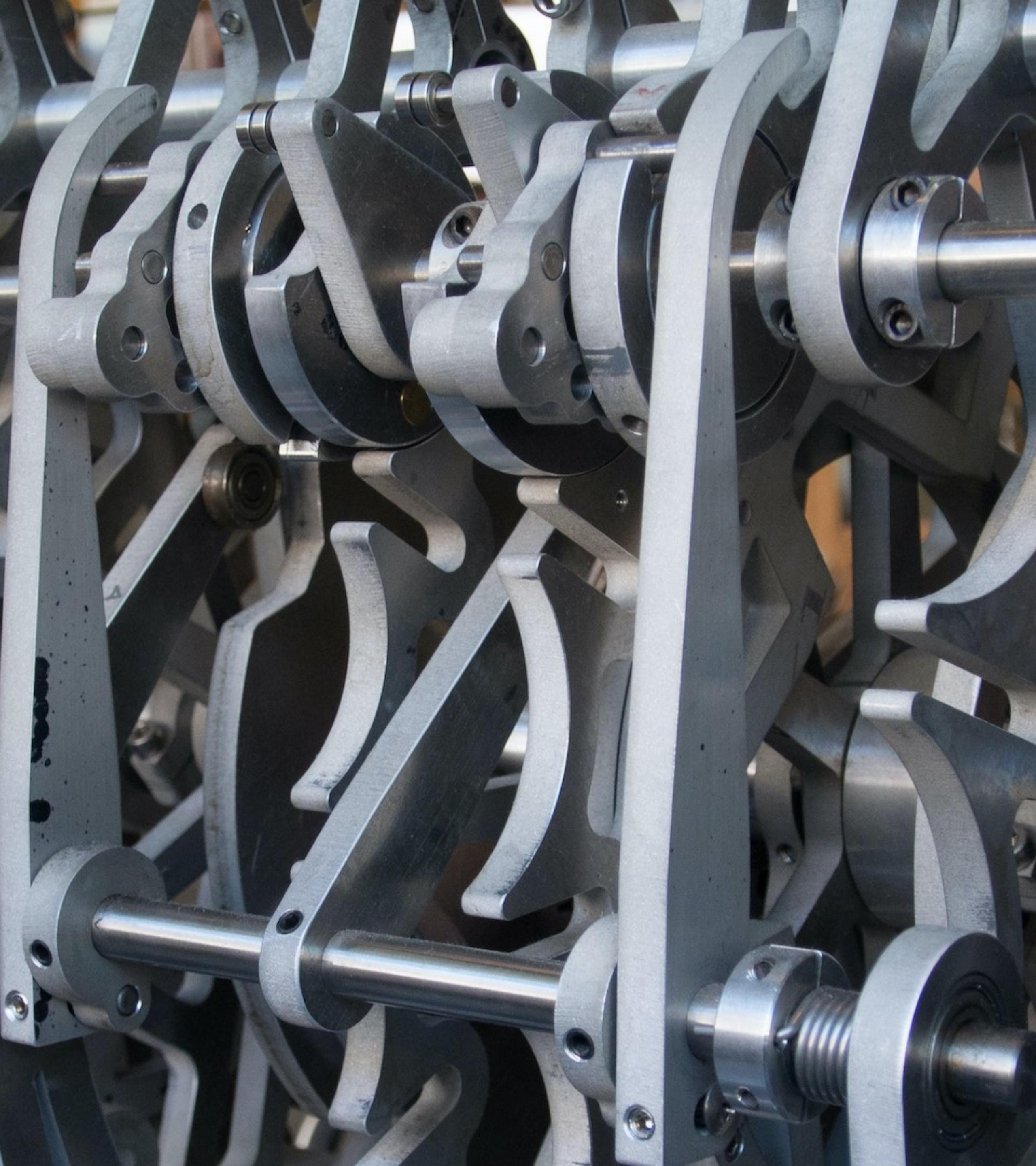
**Y. Le Cun, O. Matan, B. Boser, J. S. Denker, D. Henderson,
R. E. Howard, W. Hubbard, L. D. Jackel and H. S. Baird**
AT&T Bell Laboratories, Holmdel, N.J. 07733

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

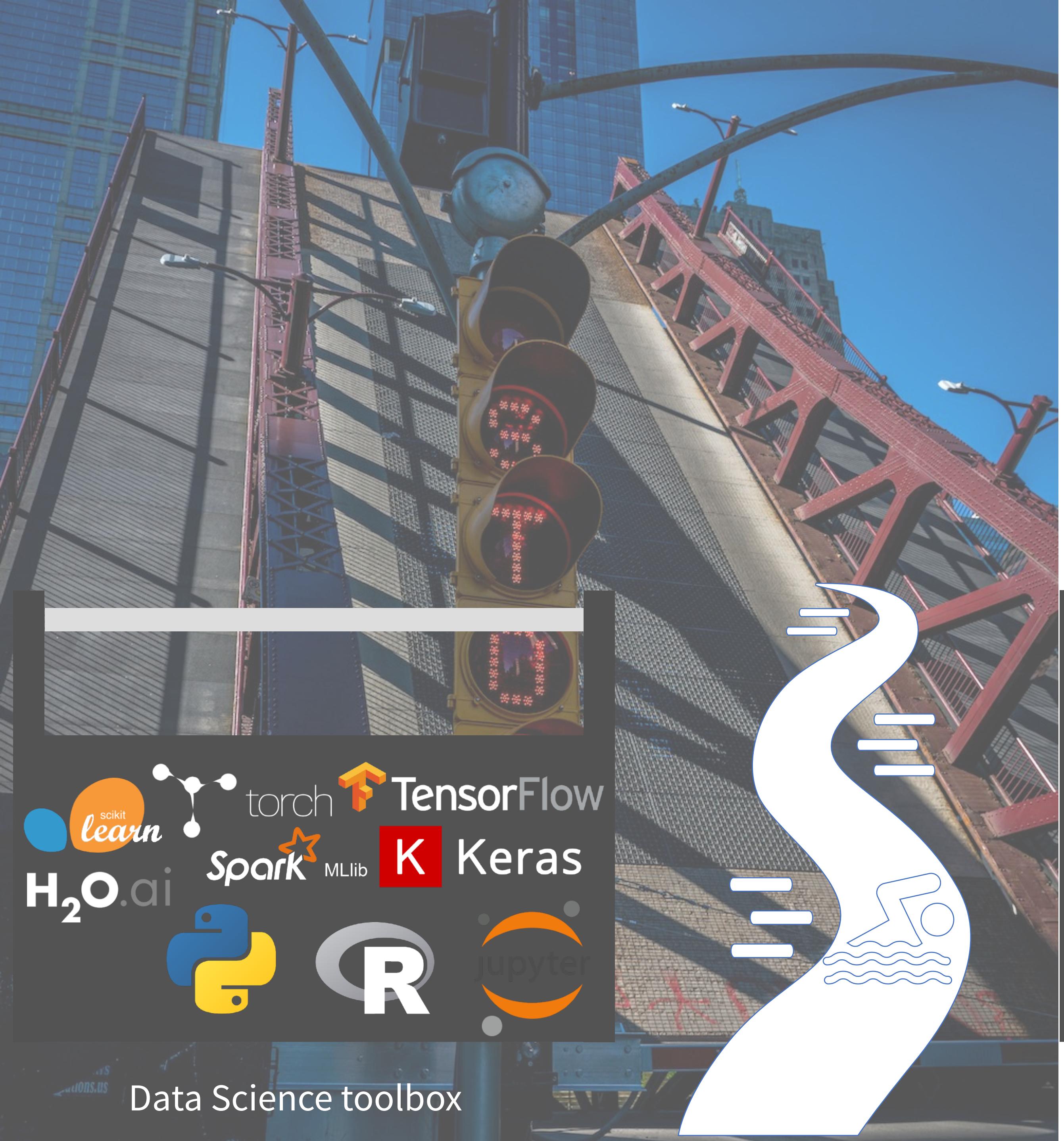
- The Promise of AI
- AI in the Enterprise
 - The Past
 - The Present
 - The Future
- Conclusions

A photograph of the Chicago Riverwalk. In the foreground, the calm water of the Chicago River reflects the surrounding city skyline. On the left, the iconic Marina City complex with its two distinctive spiral towers rises above the river. To the right, a modern residential building with a grid-like facade stands along the waterfront. The sky is clear and blue.

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

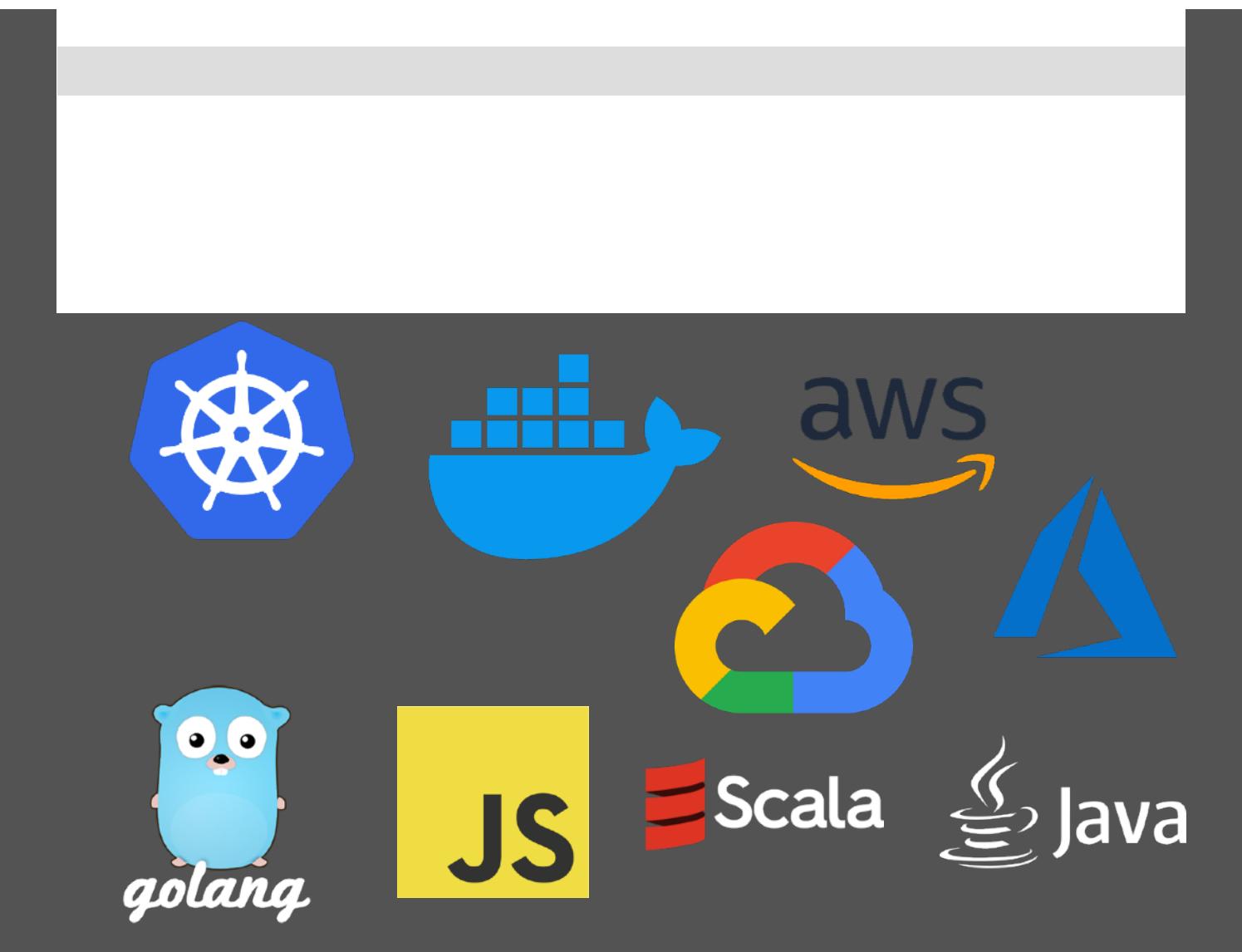


Data Science vs. Data Engineering



Data Science vs. Data Engineering

- A cultural and technical divide



Data Science toolbox

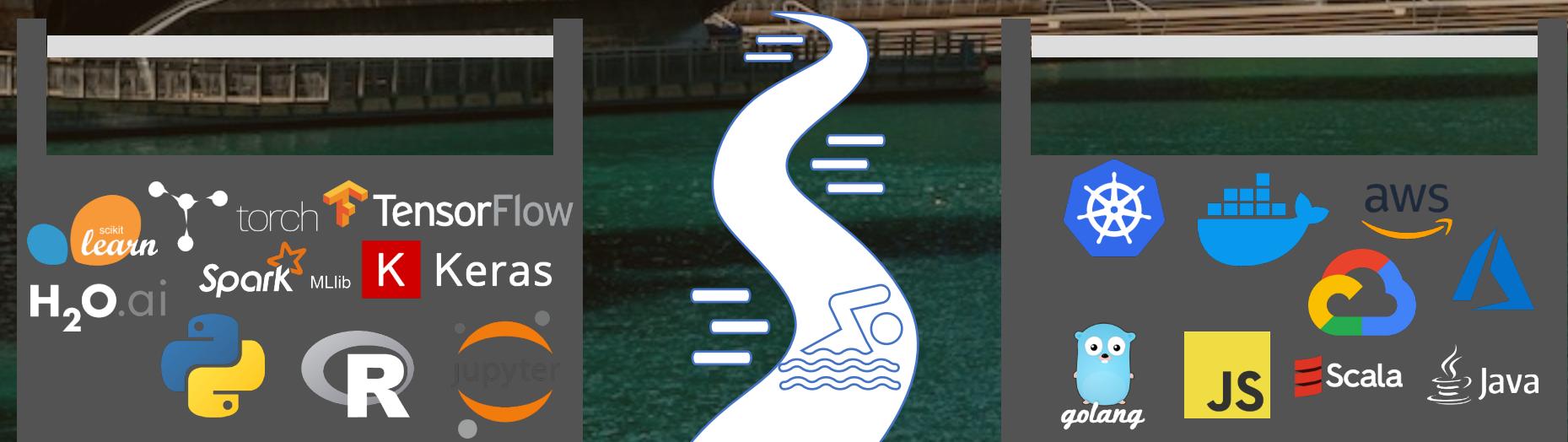
Data Scientists

- 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>



@deanwampler

Bridging the Divide



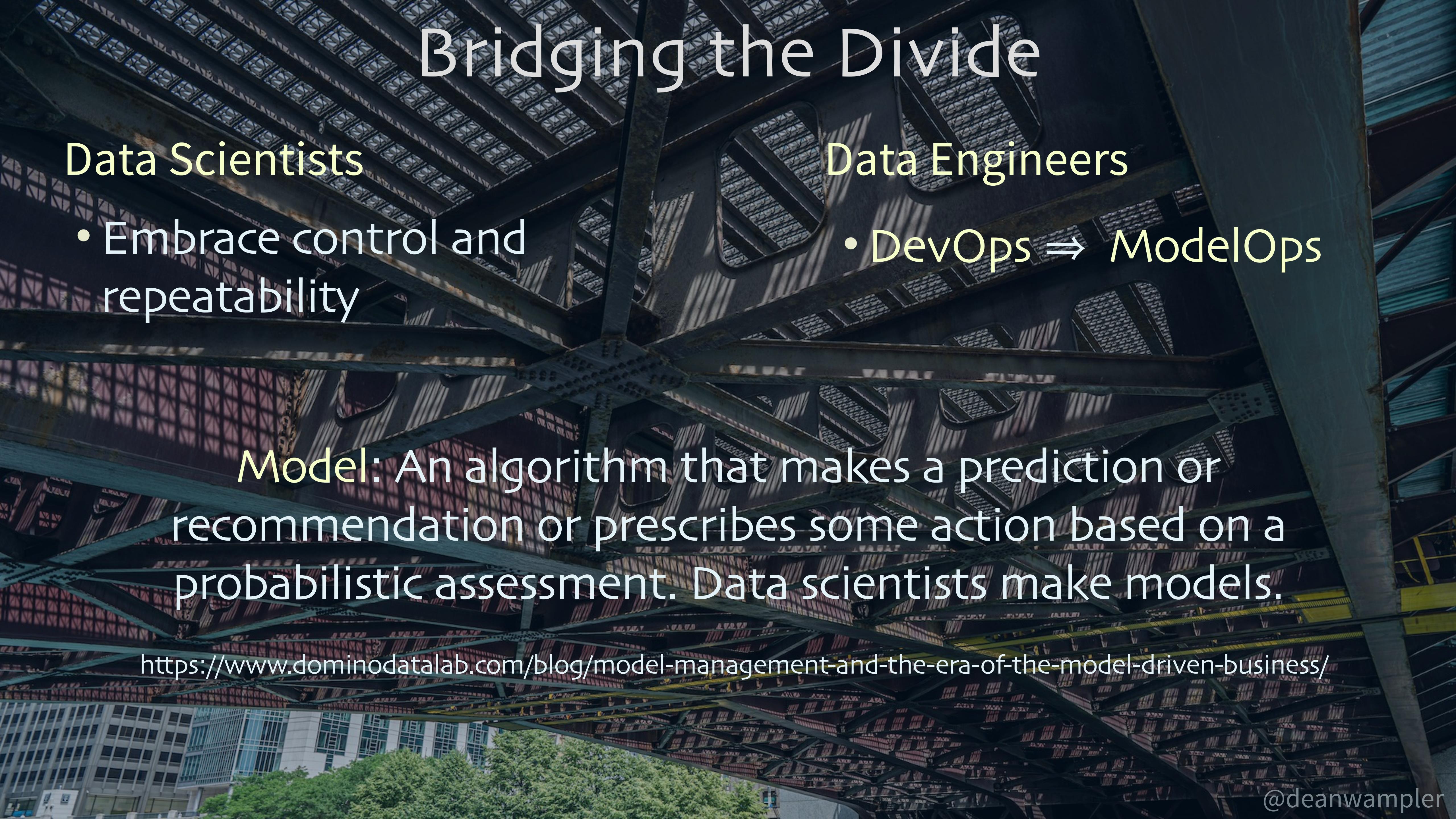
Data Scientists

- 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
 - Clothing recommendations for customers

ModelOps

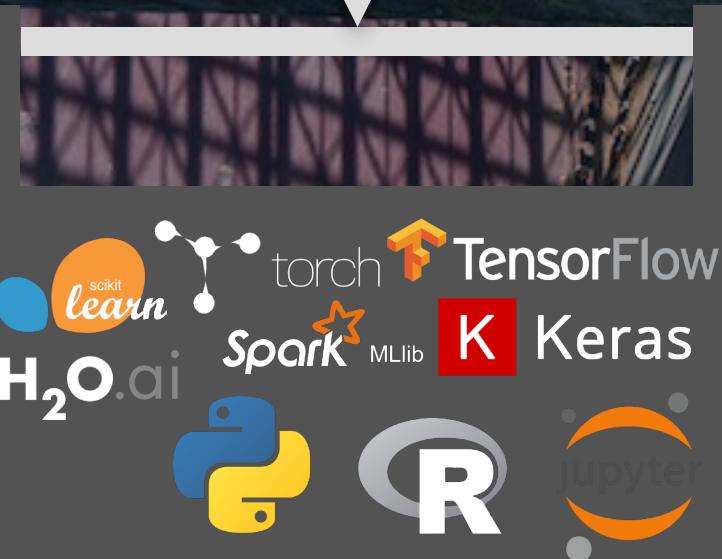
And if you look at the most successful companies in the world, you'll find models at the heart of their business driving that success.

- 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



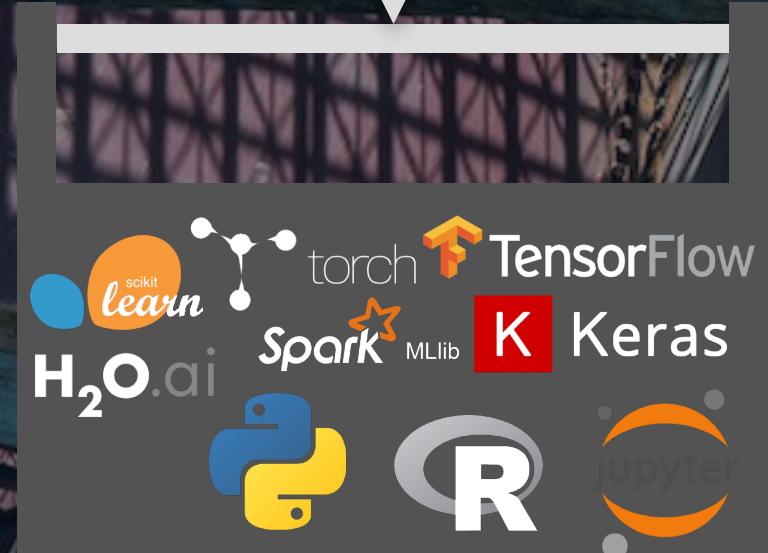
Model Development

Research
new models

ModelOps



Data



Model Development

Model CI/CD Pipeline

Research
new models

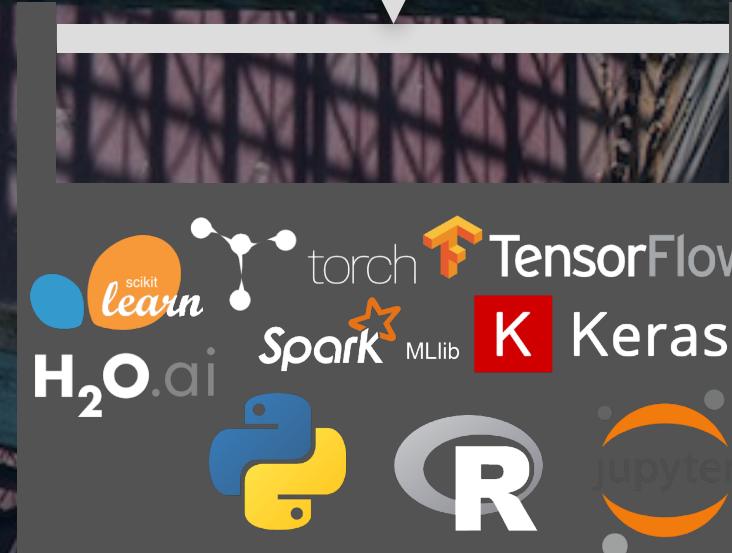
Versioning, traceability,
reproducibility, automation



ModelOps

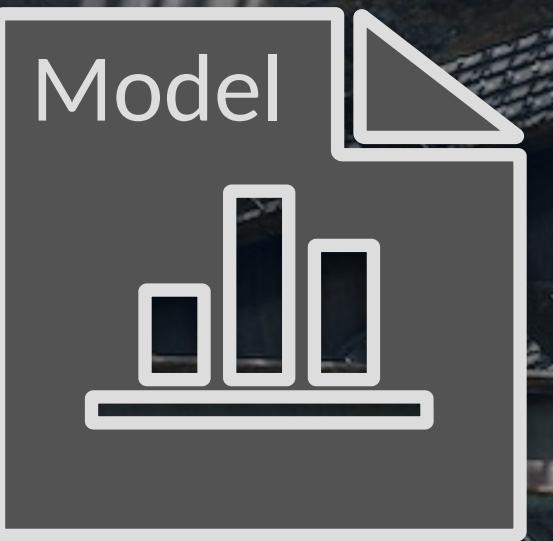


Data



Model Development

Model CI/CD Pipeline

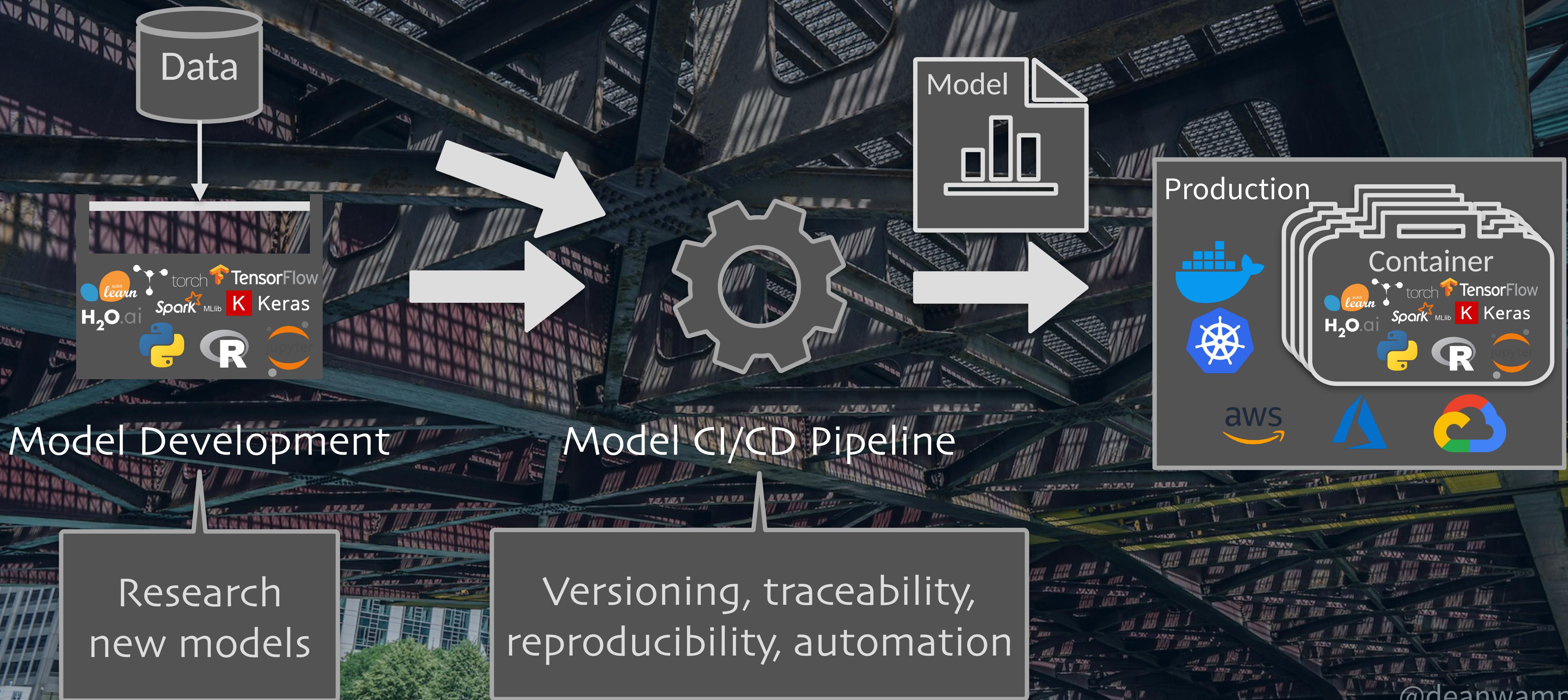


Model

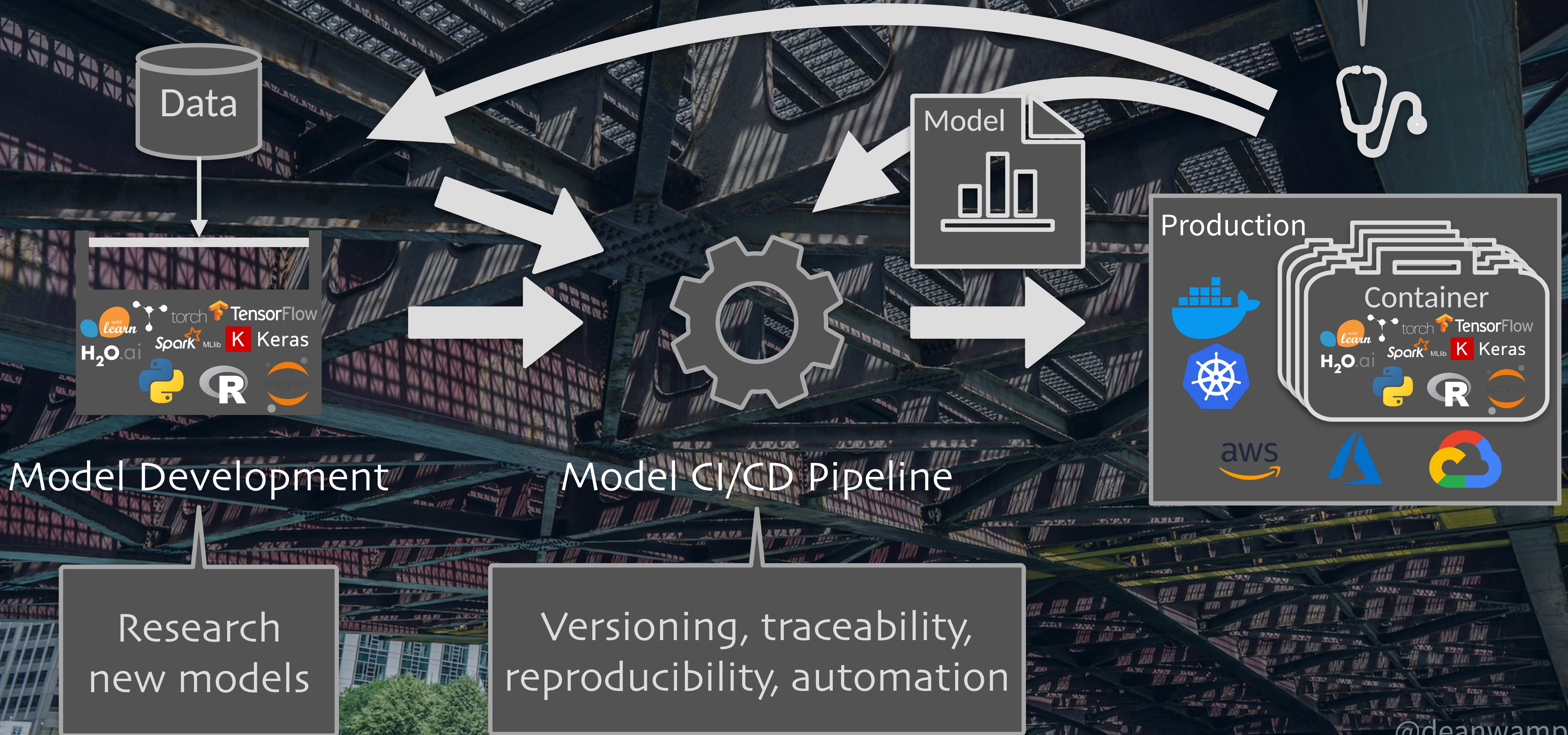
Research
new models

Versioning, traceability,
reproducibility, automation

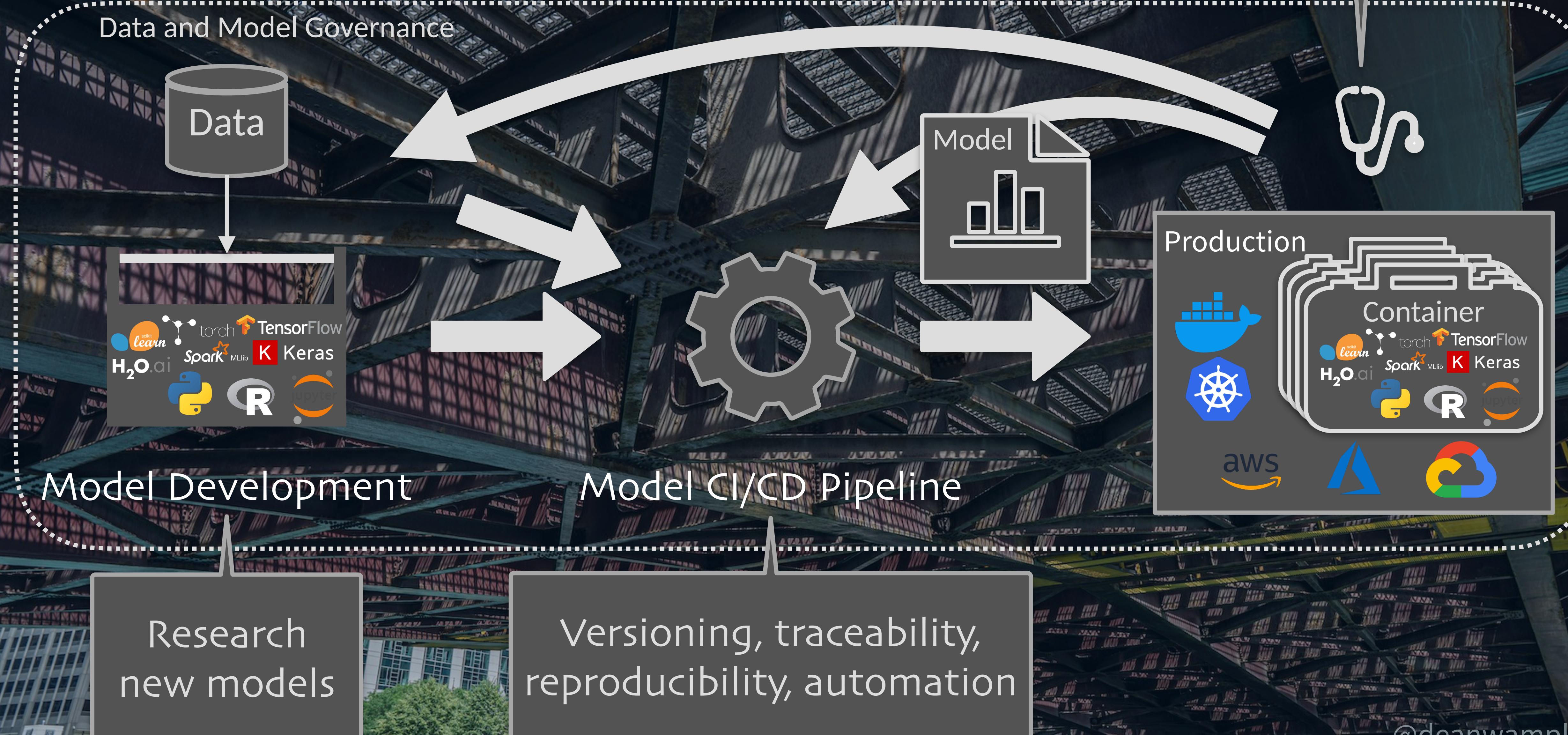
ModelOps



ModelOps



ModelOps



ModelOps

Monitor

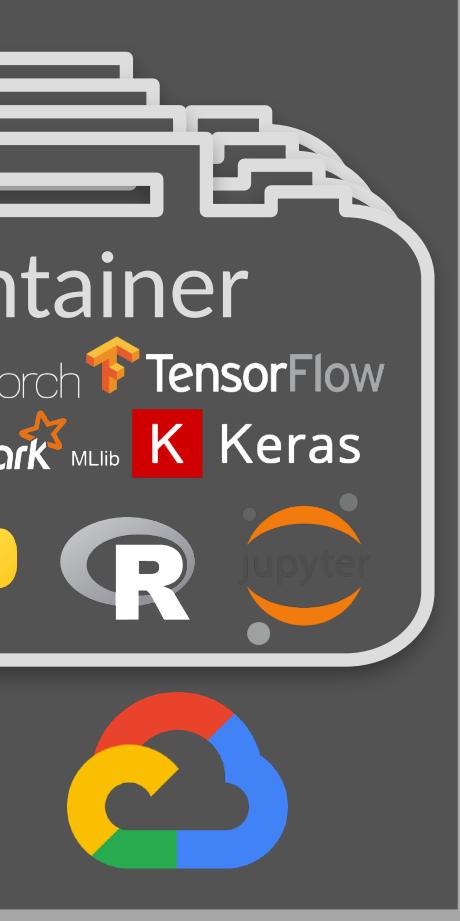
This is shown as a batch process, but expect these processes to evolve into streaming pipelines, with continuous training.

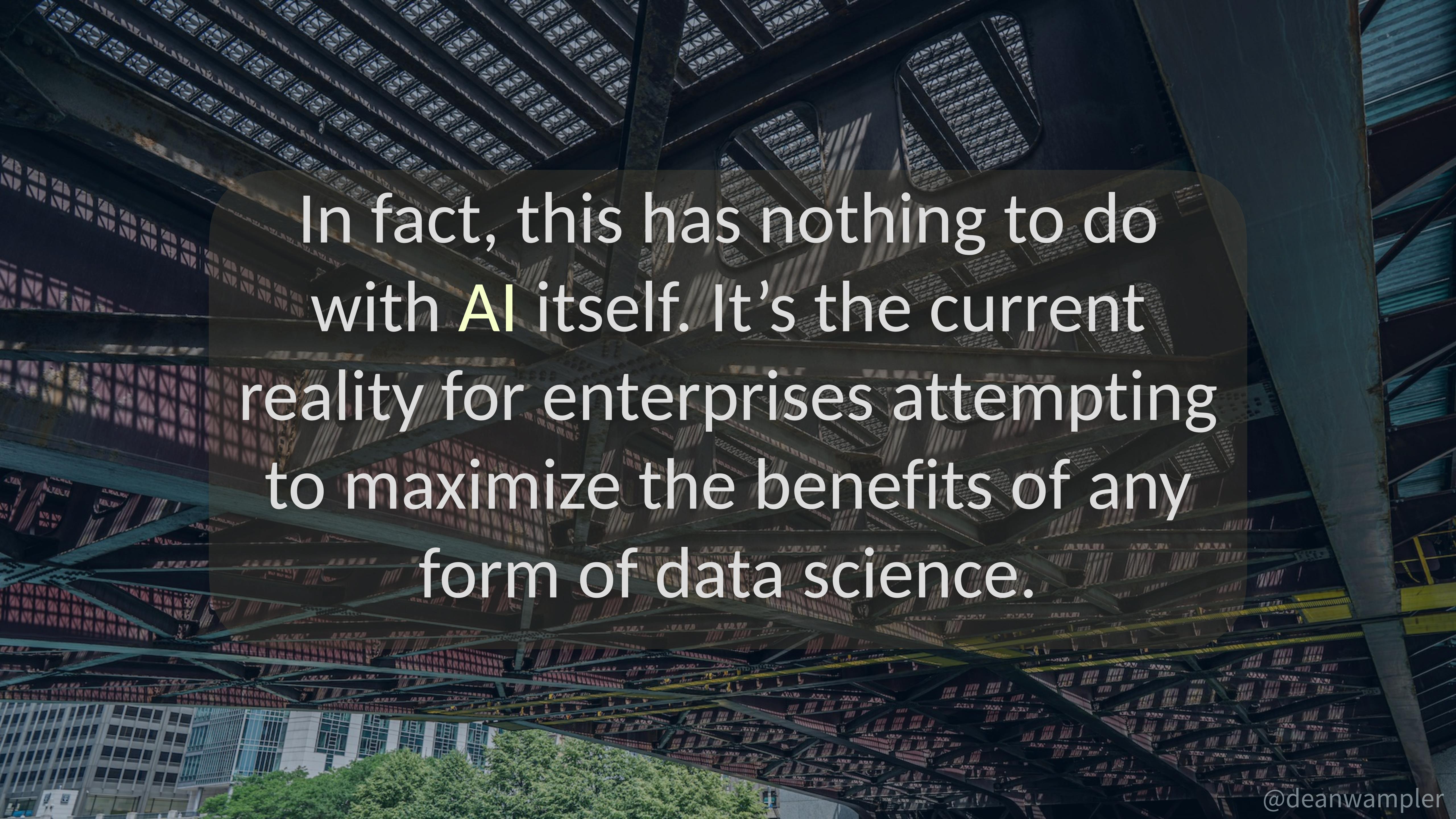
Mode

R

new models

reproducibility, automation



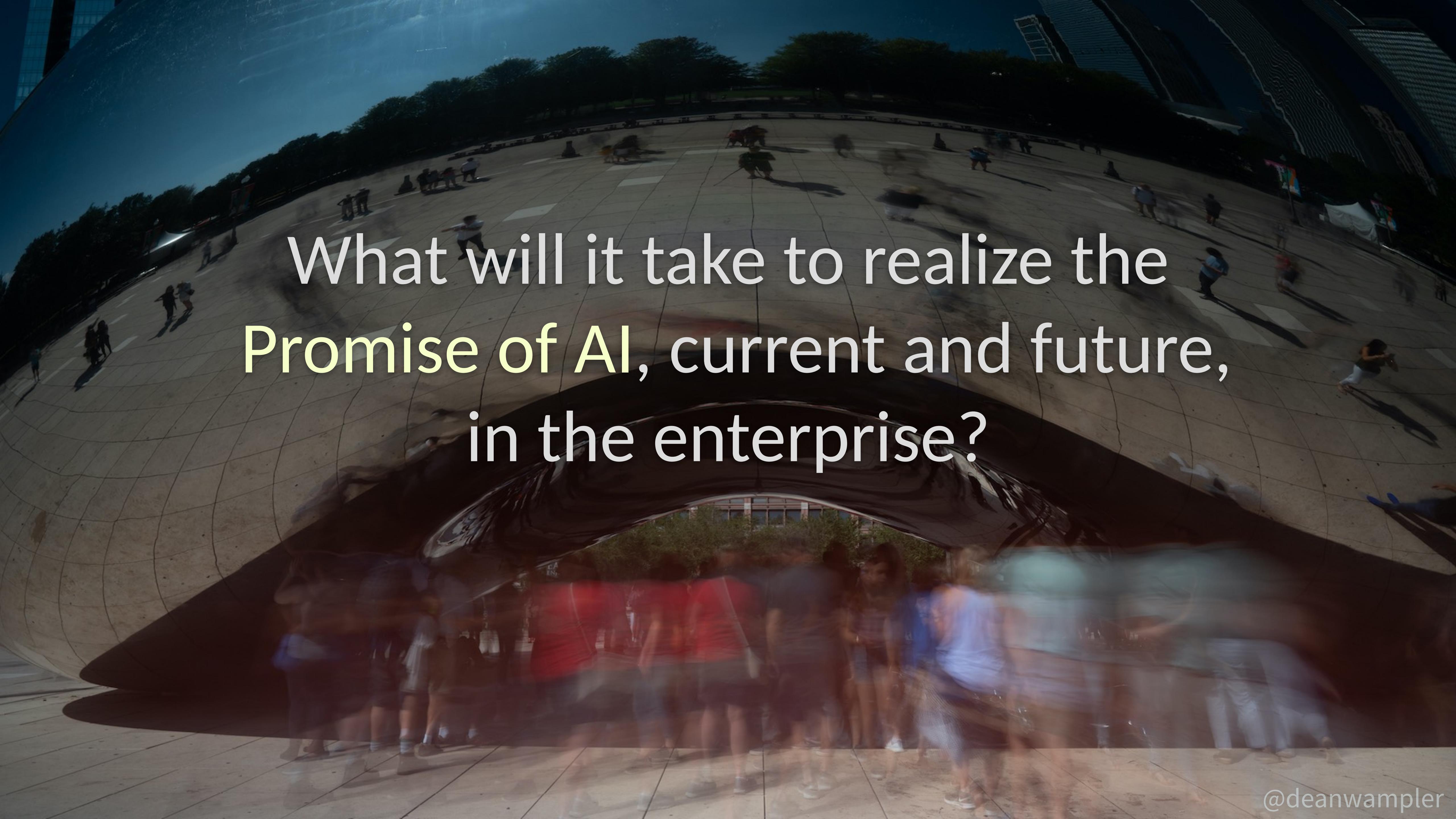


In fact, this has nothing to do
with AI itself. It's the current
reality for enterprises attempting
to maximize the benefits of any
form of data science.



Outline

- The Promise of AI
- - 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

-
-
-
-



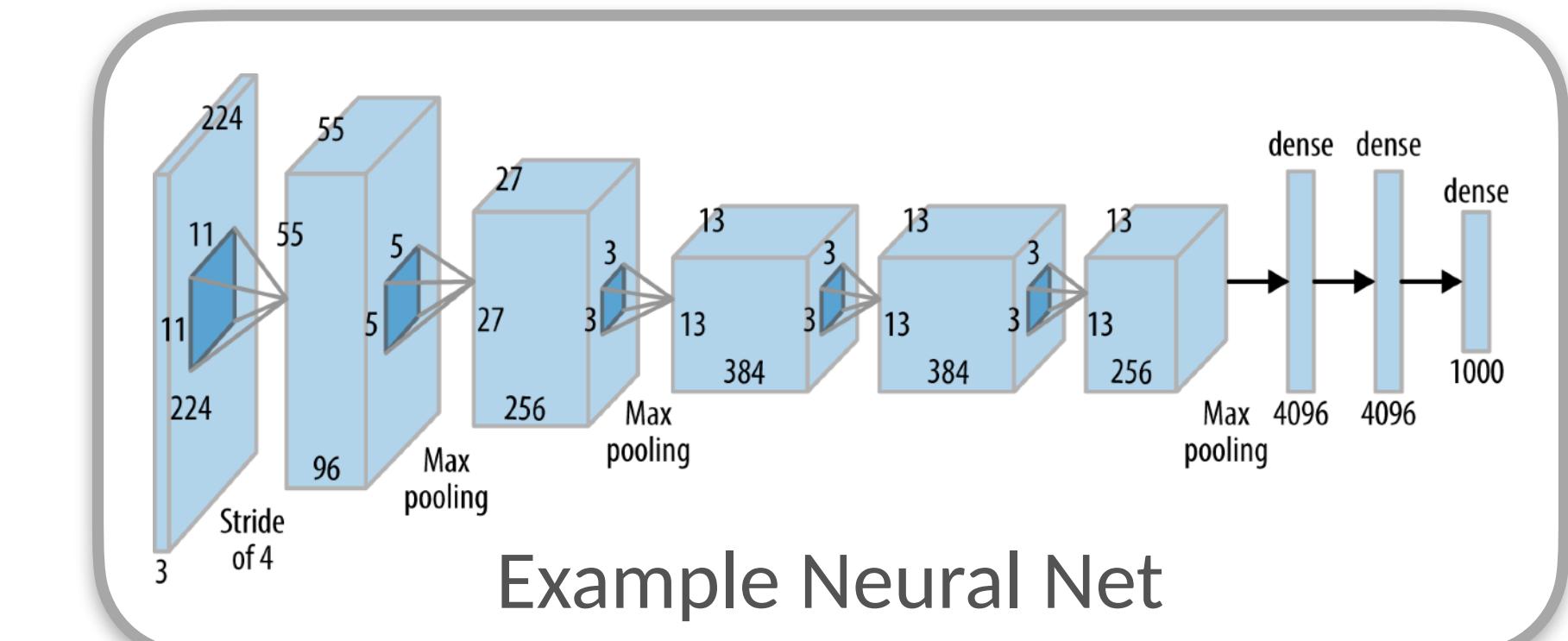
Along the way...



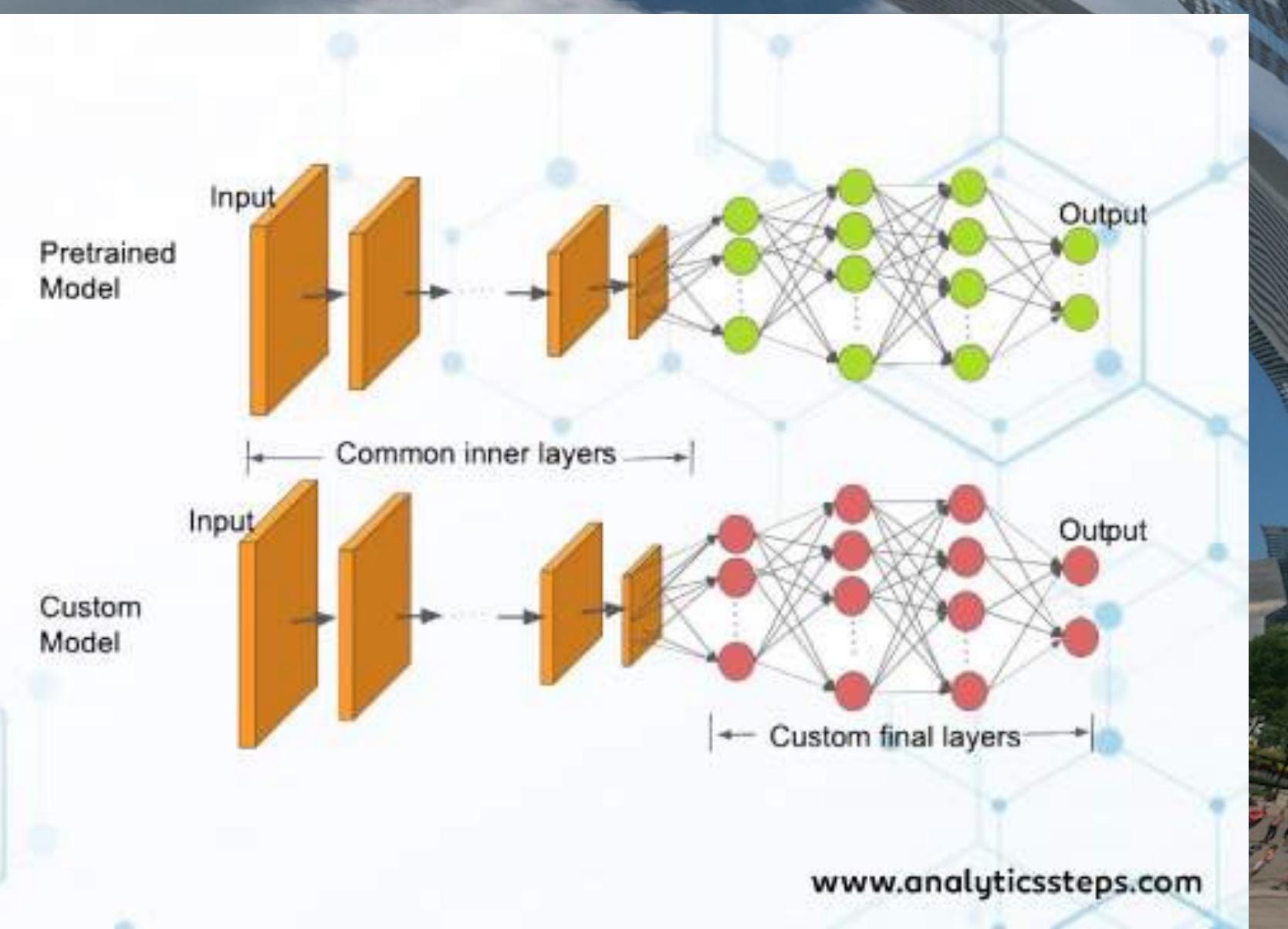
“The largest version GPT-3
175B or ‘GPT-3’ has 175 B
Parameters, 96 attention
layers and 3.2 M batch size.”

NLP

-



Transfer Learning



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

Transfer Learning



Reinforcement Learning

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

@deanwampler



Infrastructure



expensive
Burst





Infrastructure

- hybrid-cloud



Infrastructure

- cost
of moving data



Infrastructure

- federated learning
- differential privacy

<https://openmined.org/>

@deanwampler

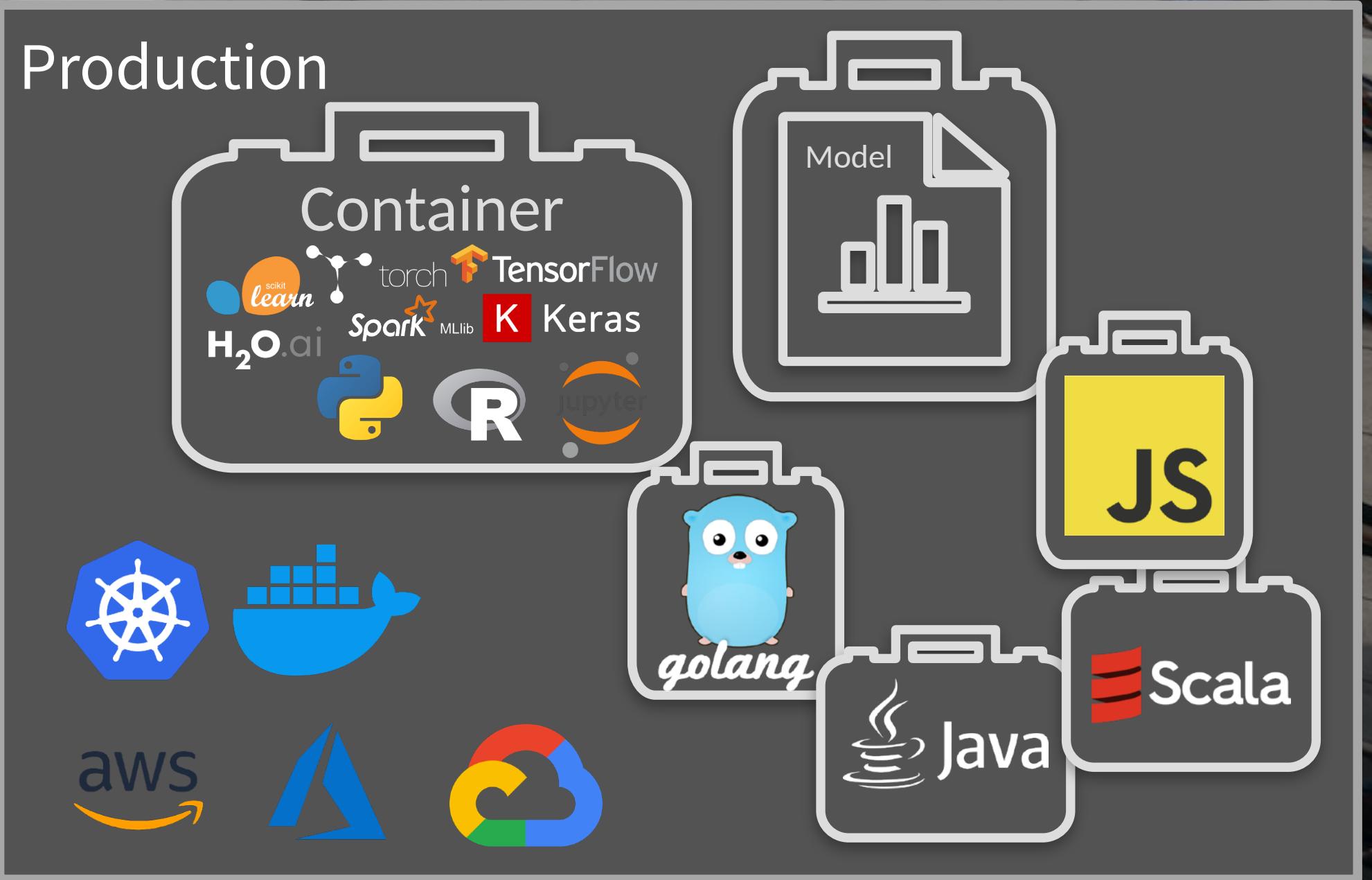


Software Development Impacts

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

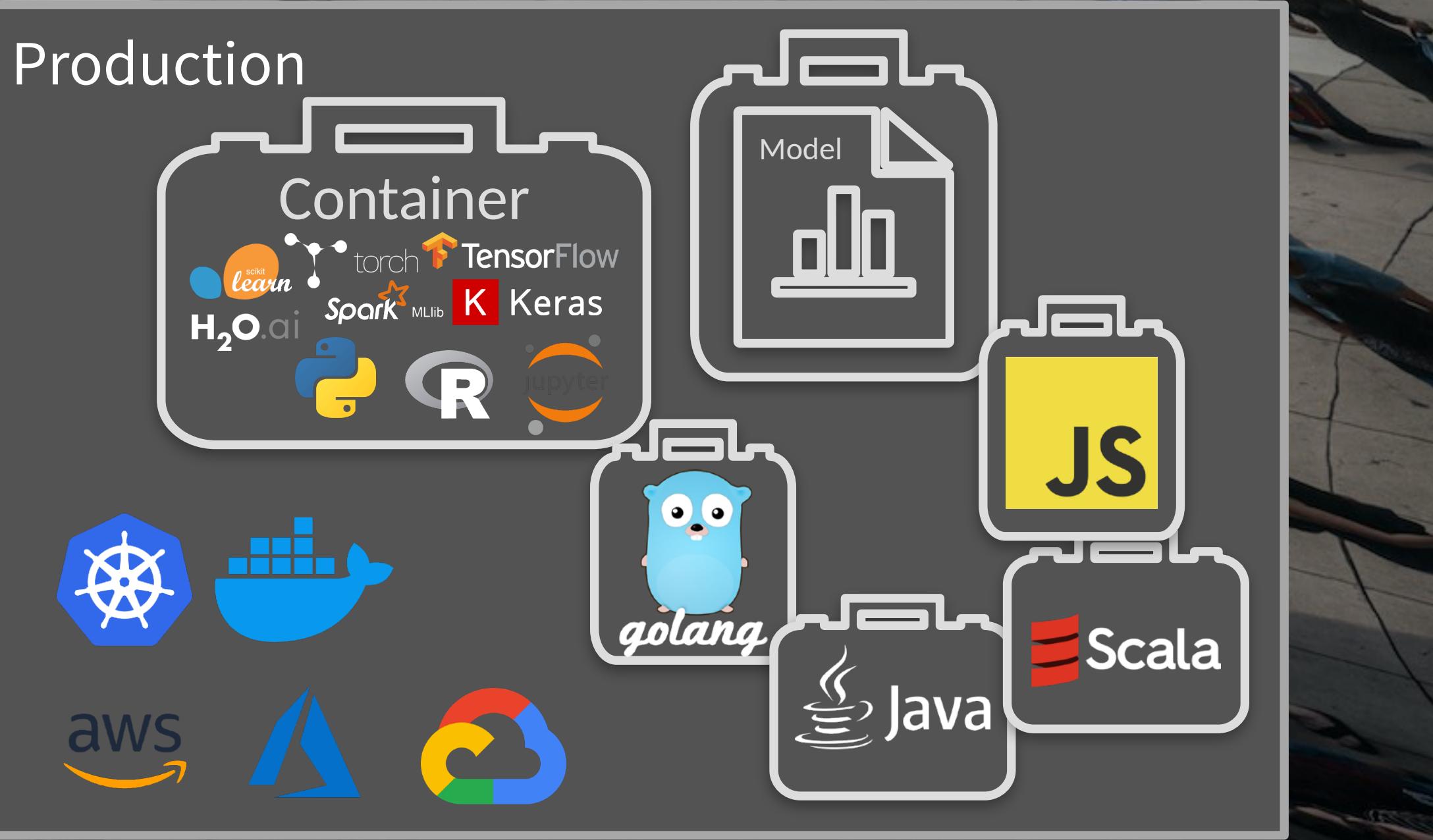
Software Development Impacts

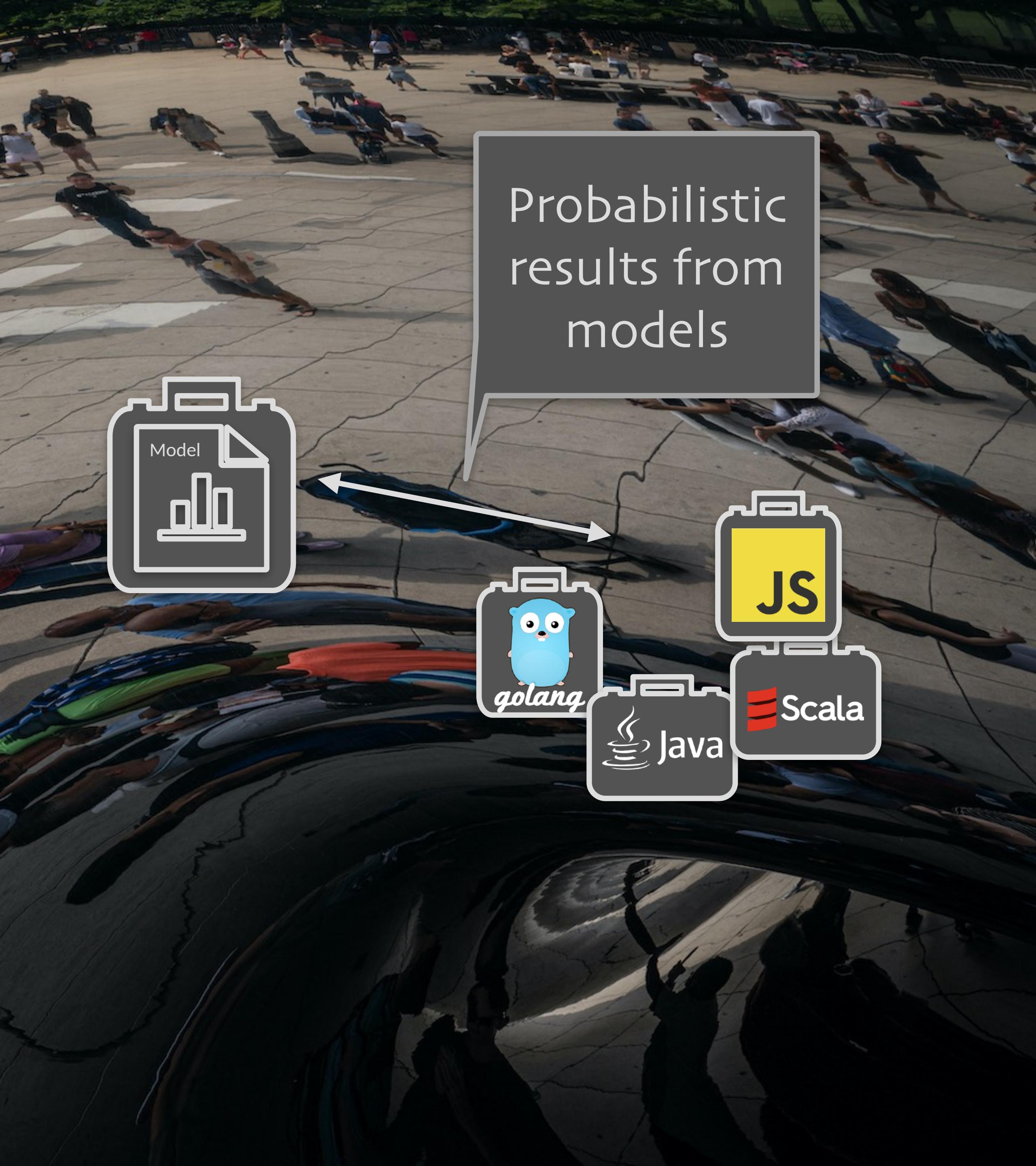
- Ubiquitous AI requires:
 - Heterogeneous tools
 - Batch and streaming data processing
 - Statistical & probabilistic thinking



Software Development Impacts

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



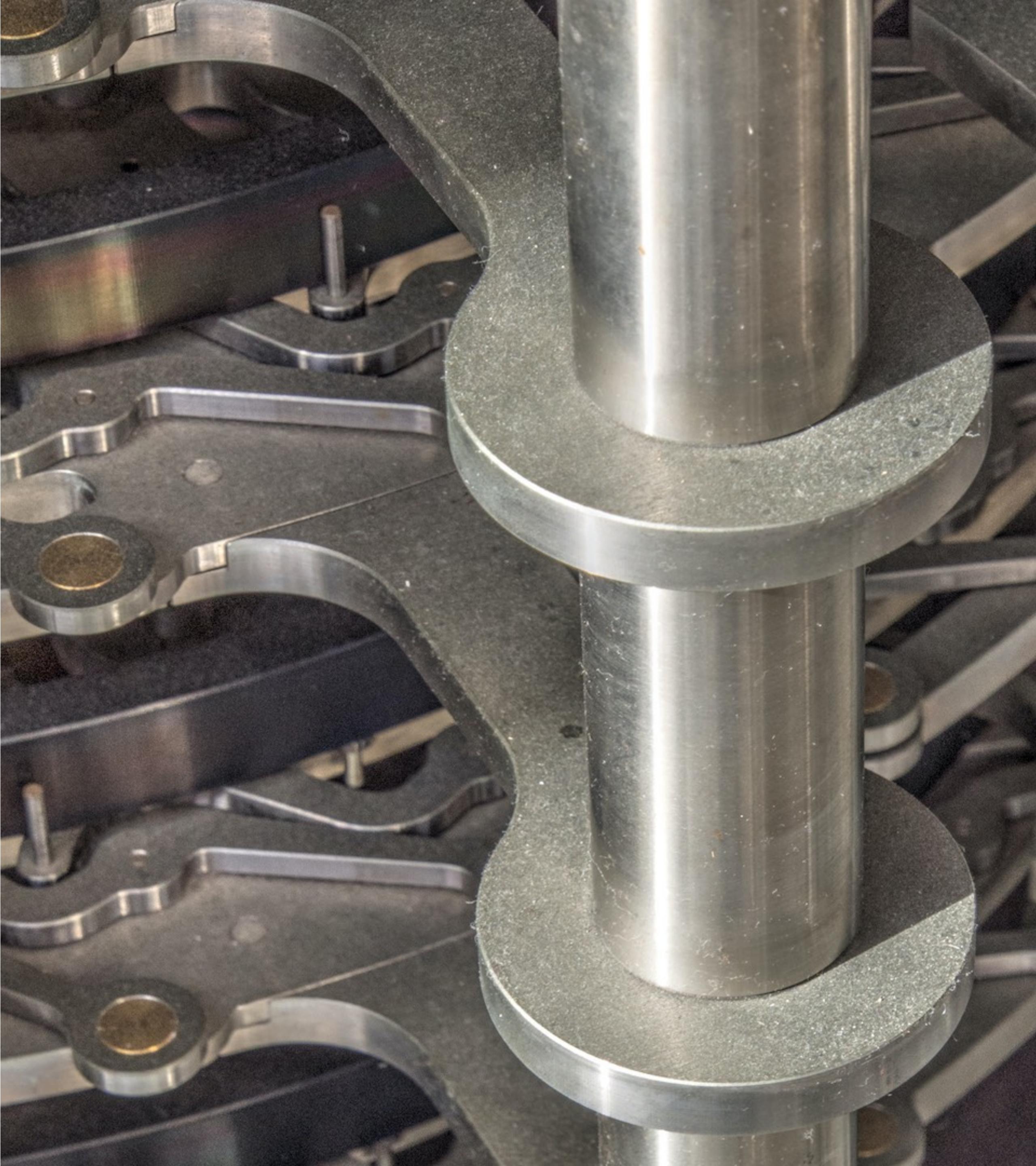
A wide-angle photograph of a large group of people performing synchronized stretching exercises on a paved surface. They are arranged in several rows, stretching their arms and legs in unison. The scene is outdoors, likely in a park or public square, with trees and a fence visible in the background.

Software Development Impacts

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



Probabilistic
results from
models

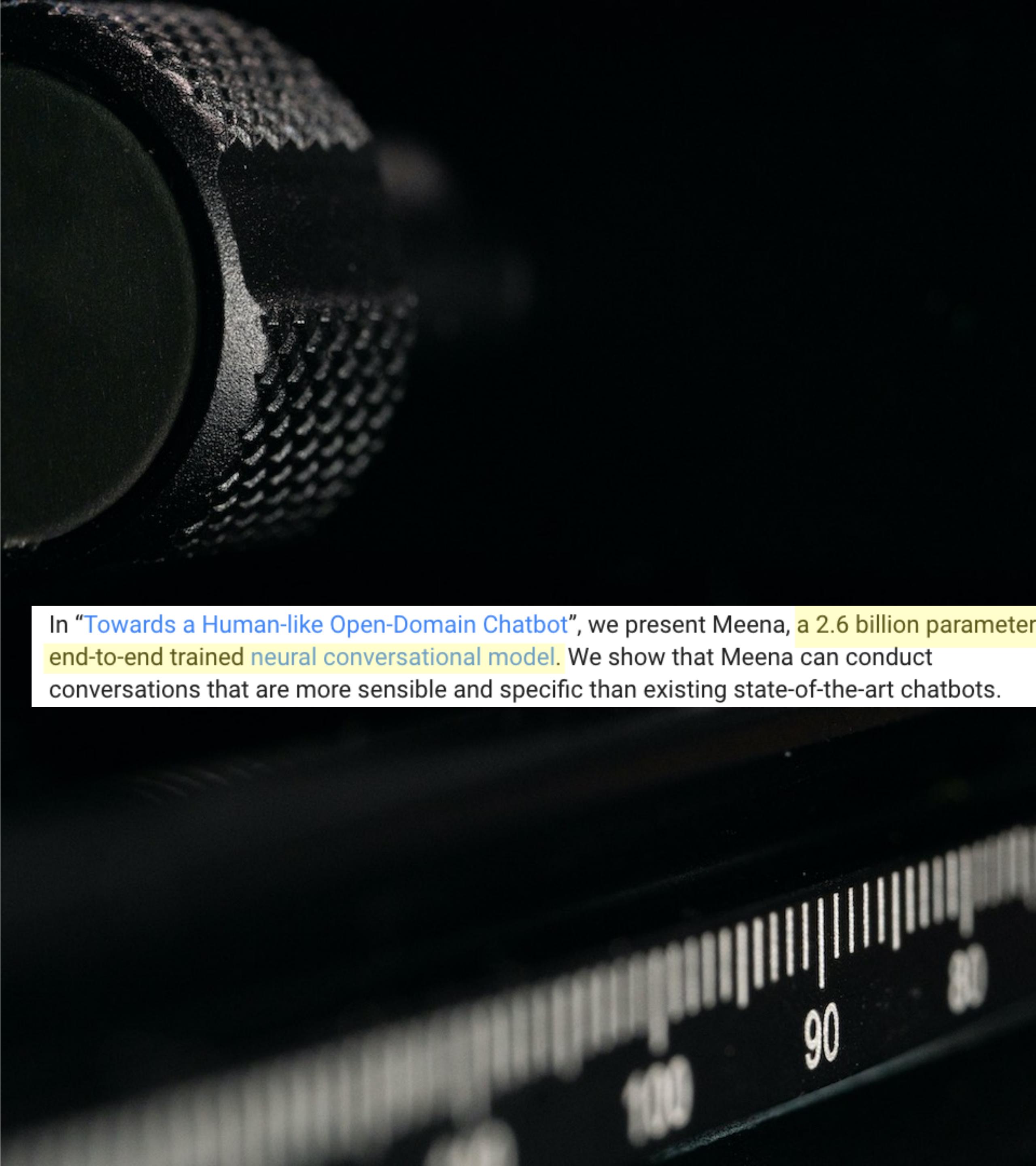


Outline

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



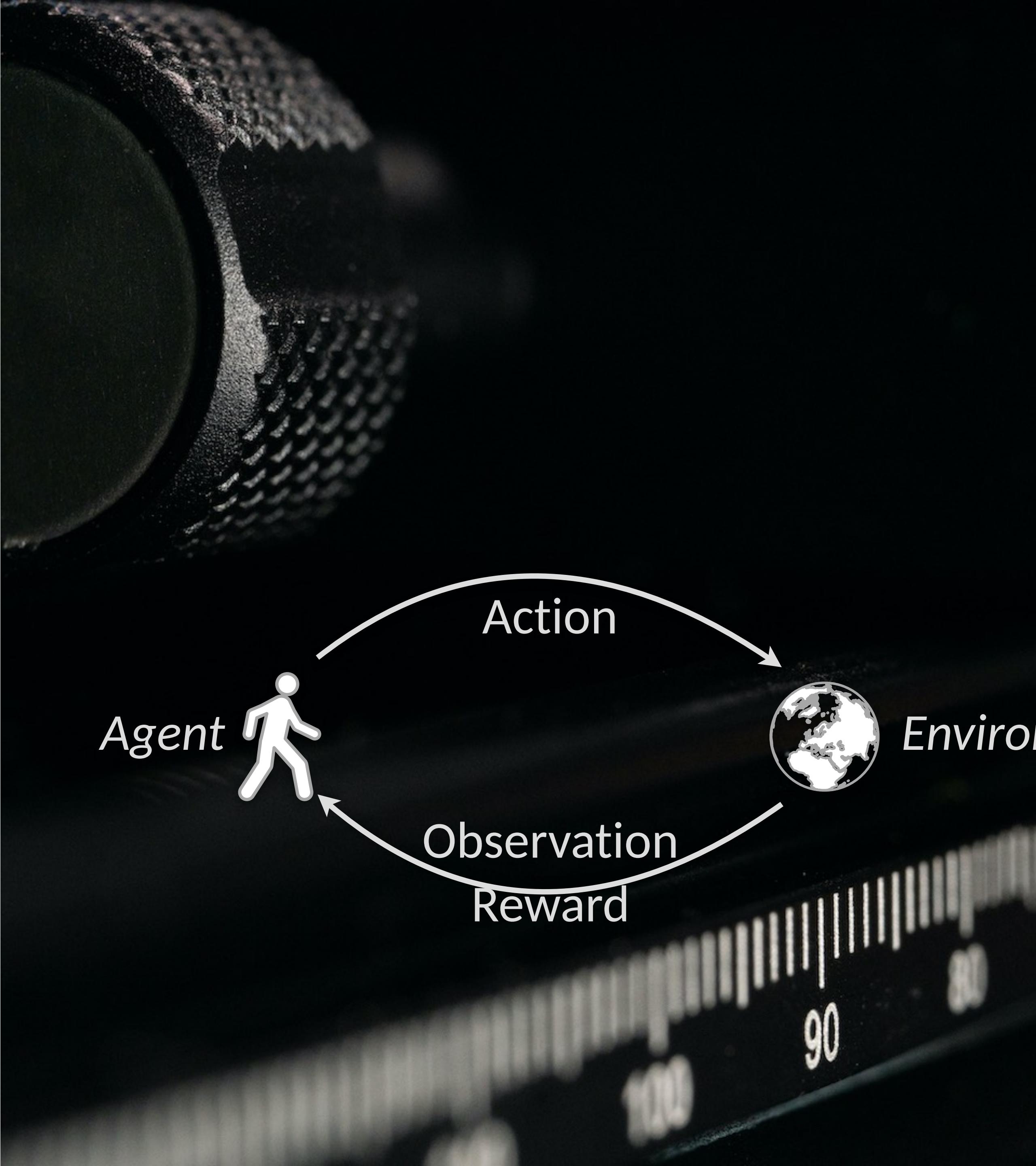
We can expect AI to become ubiquitous in the coming years, providing competitive advantages to enterprises that learn how to use it.



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.



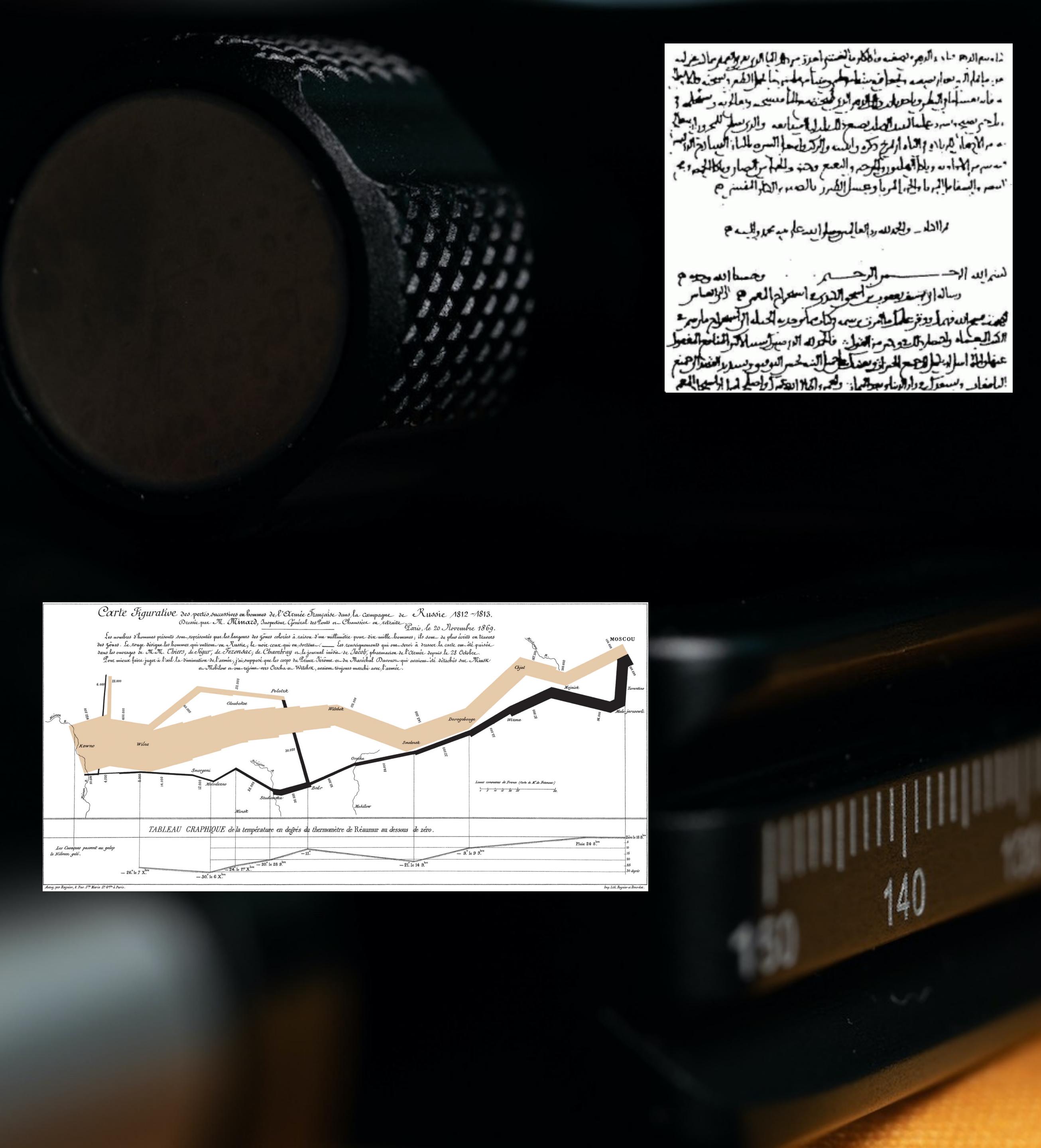
AI's Promise

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



AI's Promise

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

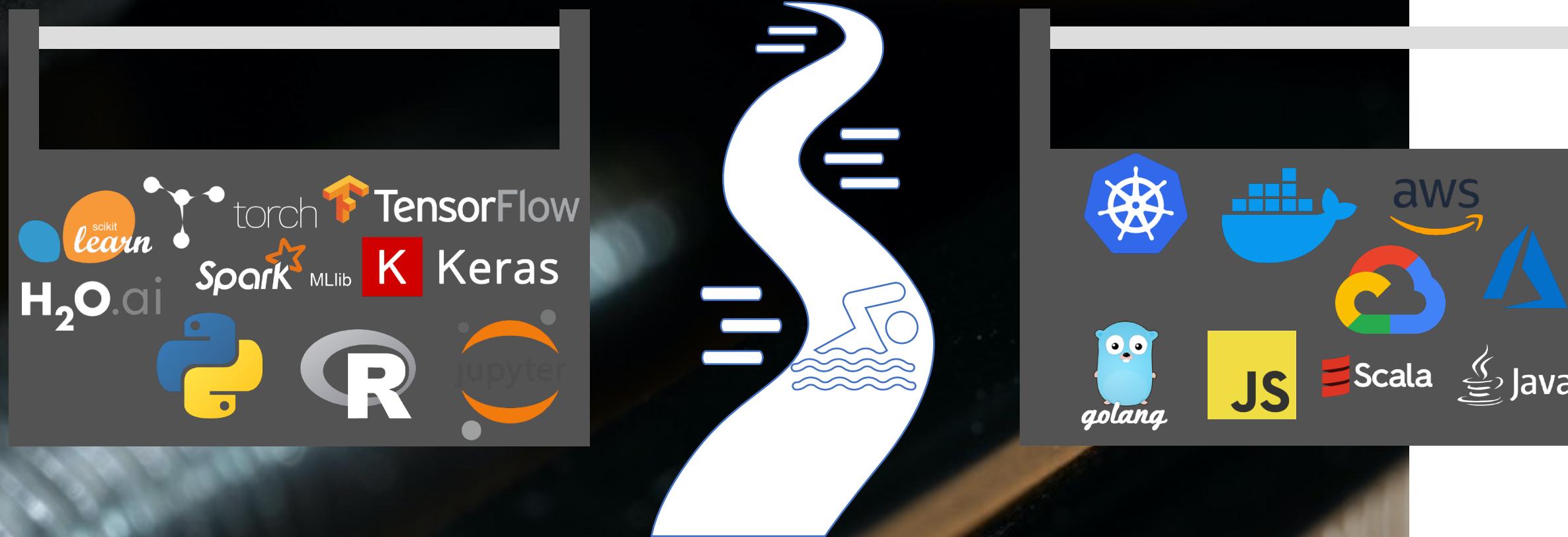


The Past

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

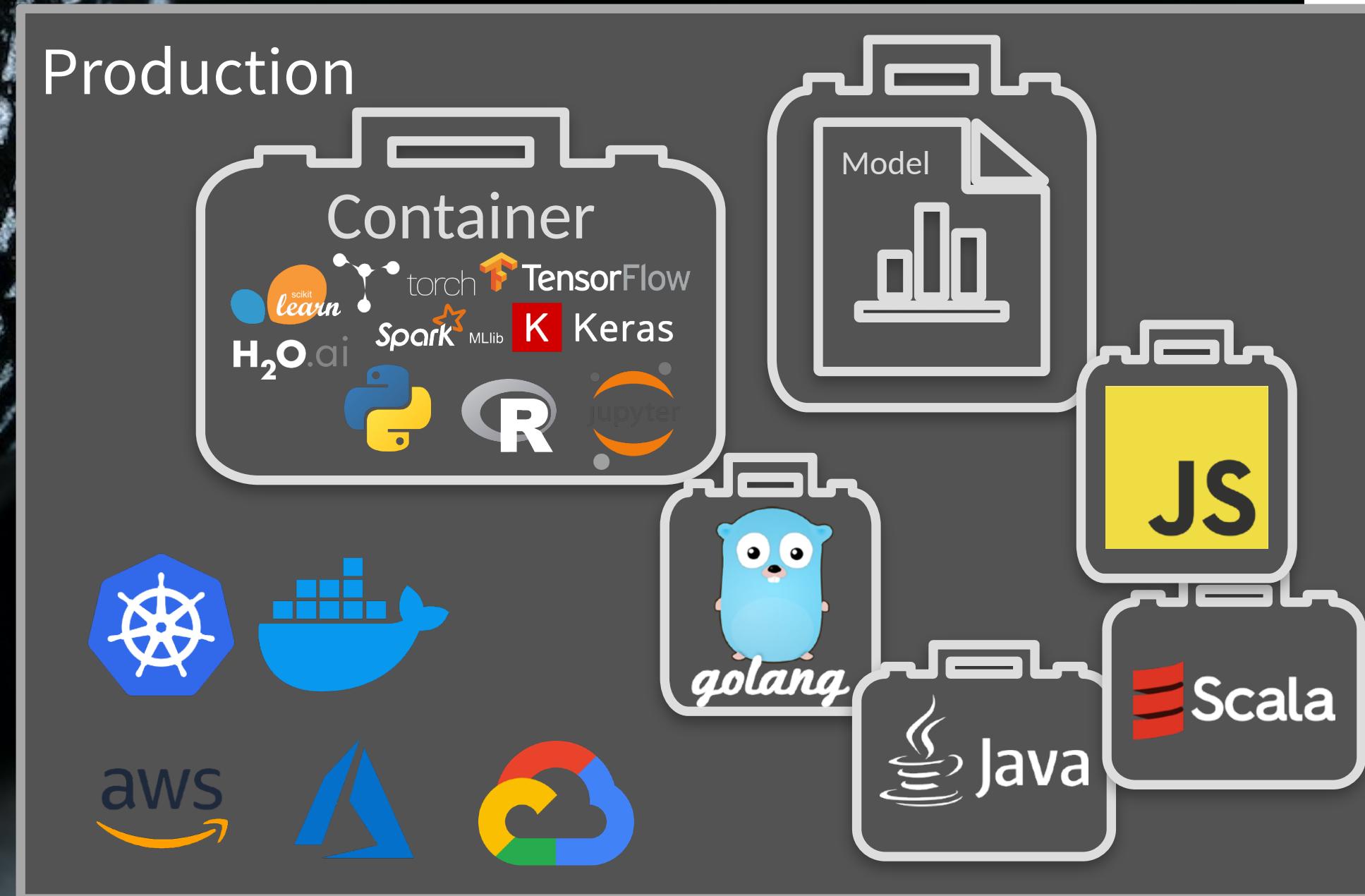
The Present

- 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



Thank You!

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

