



Data, Machine Learning, and AI: 2023 Opportunities and Trends

Ben Lorica, Mikio Braun, Jenn Webb



"Data, Machine Learning, and AI: 2023 Opportunities and Trends" is an annual comprehensive look at emerging developments in data infrastructure and engineering, machine learning (ML), and artificial intelligence. The report is divided into 10 sections, each focusing on a different aspect of AI and ML.

Section I: Generative AI is one of 2023's hottest areas for startup funding

Section II: Understanding, testing, and evaluating large (language) models

Section III: Model efficiency and sustainability

Section IV: Increased efforts to democratize machine learning and make it more accessible to non-experts

Section V: Data processing and data management tools for unstructured data including text, visual data, speech and audio

Section VI: Renewed focus on streaming (and data integration)

Section VII: Data engineers will focus more on operational tasks

Section VIII: The coming wave of regulations

Section IX: Profiling the next pegacorns

Section X: Other trends to watch

Overall, this report provides a comprehensive overview of the latest developments and trends in the field of artificial intelligence and machine learning, and offers insights into the opportunities and challenges that lie ahead in 2023 and beyond.

Section I: From research to real-world applications

Generative models → Generative AI

Generative models are used to learn to model the true data distribution $p(x)$ from observed samples x . Instead of traditional ML tasks like clustering, prediction, and classification, generative models with esoteric names like diffusion, VAE, GANs, and ELBO have recently been used to enable exploration, creation, and creative expression. Implementations of diffusion and other generative models are already becoming more widely available to developers and many startups are attempting to build products using these models.

	Text	Code	Images/Video	Speech/Audio
Examples				
Sample Users	Copywriter Marketer Support (technical/sales/customer) Sales Professional	Developer Engineer Analyst Data Scientist	Designer Content Creator Marketer	Call Center Agent Content Creator Gamer / Multimedia Product User

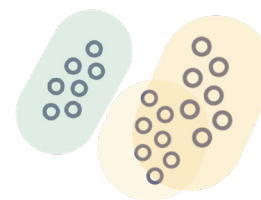
Categorizing machine learning models

Discriminative model: Focus on learning boundaries between classes



- Determine decision boundaries based on observed data
- Given an input x , what is the most likely output y
- Find y that maximizes $P(y | x)$

Generative model: Focus on learning classes



- What is the distribution of the input x ?
- What is the joint distribution of the input x and the output y ?
- Find $P(x)$ or $P(x, y)$

For more, see [this section on the generative model page](#) on Wikipedia.



Generative AI in the real world

Large language models (LLM): We continue to see many new startups that target copywriting and content marketing, general writing, and support (chatbots). We already know of startups with significant revenues that target copywriters and content creators. Current models are typically used to produce first drafts, but startups are actively working to produce better quality and longer-form text tuned for specific verticals.

Coding and programming assistants: We are happy users of GitHub Copilot, and a vast majority of its users say they feel a lot more productive while coding with it. But there are many other tools—LLMs have given rise to many AI tools that offer autocomplete-style suggestions. These tools lead to faster task completion times, help developers conserve mental energy, and allow them to focus on more satisfying tasks.

Image generation: Text-to-image generators include DALL-E, Imagen, Stable Diffusion, and more.

Speech synthesis: The artificial creation of human speech. Text-to-speech tools have been around for several years. This coming year, we expect new tools and startups focused on speech-to-speech (input and output are voice) synthesis. We are particularly keen on the potential applications of real-time speech-to-speech synthesis to such areas as health and medicine, customer support, media, and gaming. OpenAI Whisper hints at the potential of using massive amounts of data to train speech models that approach human-level robustness and accuracy.

Section II: Tools for understanding, testing, and evaluating large (language) models



AI INCIDENT DATABASE

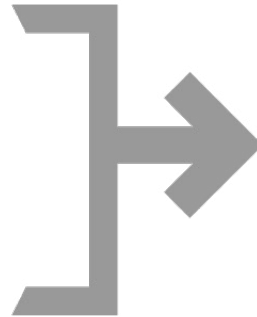
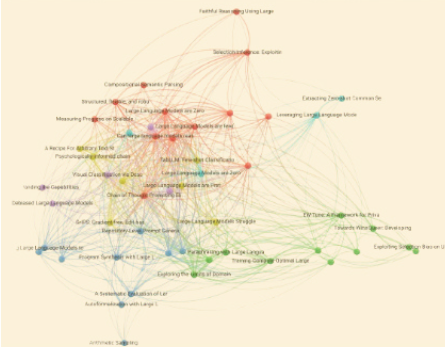
Why Meta's latest large language model survived only three days online Persistent Anti-Muslim Bias in Large Language Models

The AI oracle of Delphi uses the problems of Reddit to offer dubious moral advice Prompt injection attacks against GPT-3

arxiv.org papers on testing LLMs

[Submitted on 15 Sep 2021]

Challenges in Detoxifying Language Models



2022

Hugging Face **Evaluation on the Hub**



HELM

Center for Research on Foundation Models

Task	Scenarios				Language	Input perturbation	Metrics	
	What	Who	When	Language			Input	Output measure
Question answering	Wikipedia	Age groups	2018	English	Natural Questions	None	Accuracy	
Summation	Reviews	Gender	2011	French	SP08	Adaptation Type	Exact Match, F1, BLEU	
Text classification	News	RACE	2012	Chinese	?	Gender	F1, Accuracy, Toxicity	
Information retrieval	Twitter	Age, Occupation, Ethnicity	Pre-Internet	Swedish	?	Dialect	Efficiency, F1, Precision, Recall	

2023

More tools to come ...

As general-purpose models become more prevalent, there's a growing need for tools to help developers select models appropriate for their use case and, more importantly, to help them understand the limitations of these models. Along those lines, the startup Hugging Face recently [released low-code tools](#), which make it simple to assess the performance of a set of models along an axis such as FLOPS and model size, and to assess how well a set of models performs in comparison to another.

Researchers at Stanford's [Center for Research on Foundation Models](#) just unveiled the [results of a study](#) that evaluated the strengths and weaknesses of 30 well-known large language models. In the process, they developed a new benchmarking framework, [Holistic Evaluation of Language Models \(HELM\)](#), which can be described as follows:

- They organize the space of scenarios (use cases) and metrics (desiderata).

- They then select a subset of scenarios and metrics based on societal relevance (e.g., user-facing applications), coverage (e.g., different English dialects/varieties), and feasibility (i.e., amount of compute).

More broadly, we expect more tools for testing models prior to release:

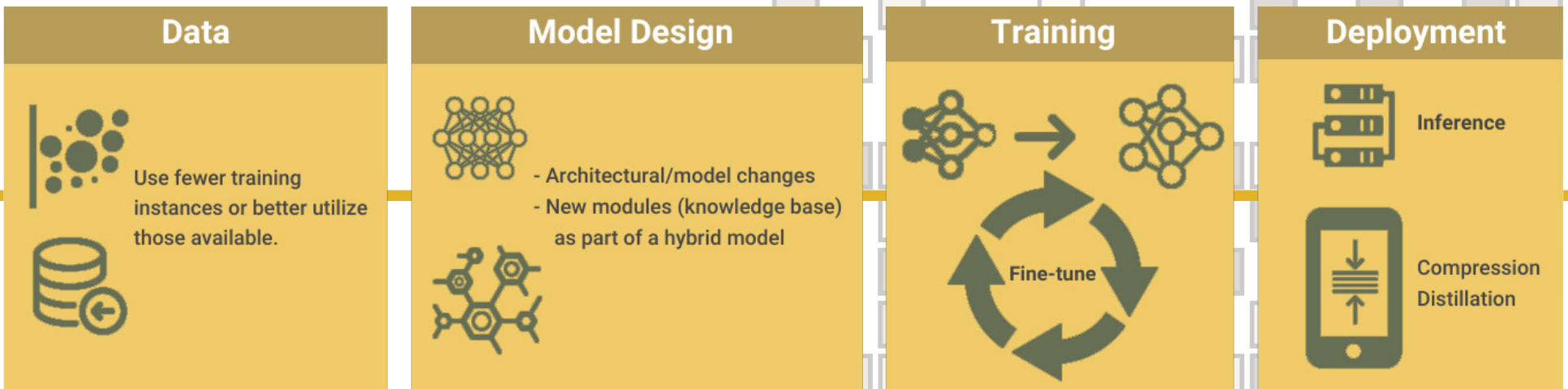
- [Why Meta's latest large language model survived only three days online](#)

Section III: Training *and* maintaining models puts the focus on efficiency and sustainability

As models (for speech, vision, and text) get more widely deployed and used, the seemingly positive correlation between model size and accuracy has prompted research into less resource-intensive methods that can produce comparable results. These research initiatives are beginning to inspire real-world deployments.

A recent [survey paper](#) provides a comprehensive overview of such initiatives in NLP:

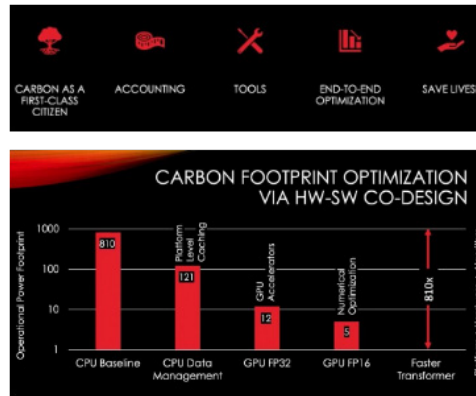
- **Data:** Using fewer training instances, or better utilizing those available, can increase efficiency.
- **Model design:** This pertains to architectural changes or new modules that accelerate the workflow of the main model. For example, a promising direction in text generation is to combine parametric models with retrieval mechanisms from a database or a knowledge graph.
- **Training:** Includes pre-training and fine-tuning.
- **Inference and compression**



Sustainable AI

Organizations like [Allen AI](#) and [Meta](#) are devoting resources to green/sustainable AI, a collection of tools and processes that explores the environmental impact of AI from a holistic perspective. The goal is to develop and deploy AI systems that yield novel results while considering computational and environmental costs, thereby reducing resource usage.

For a given particular use case, the goal is to measure the environmental impact of AI data, algorithms, and system hardware. Additionally, researchers in this area consider emissions across the life cycle of hardware systems, from manufacturing to operational use.



Internal Tools

- Reporting carbon emissions in production training workflows

External Tools

- Carbon Explorer

Meta AI

Source: Kim Hazelwood at the 2022 Ray Summit

Green AI **Ai2** Allen Institute for AI

By Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni

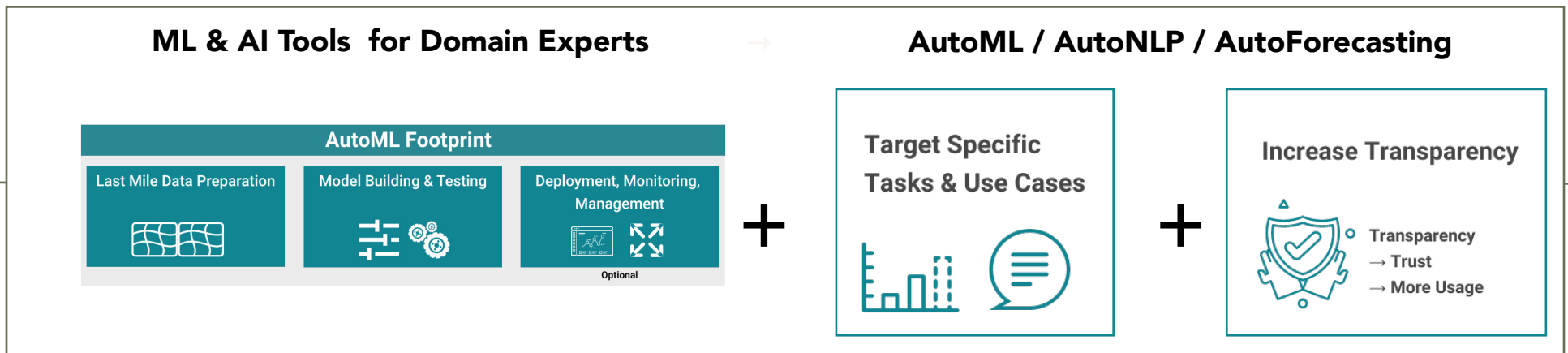
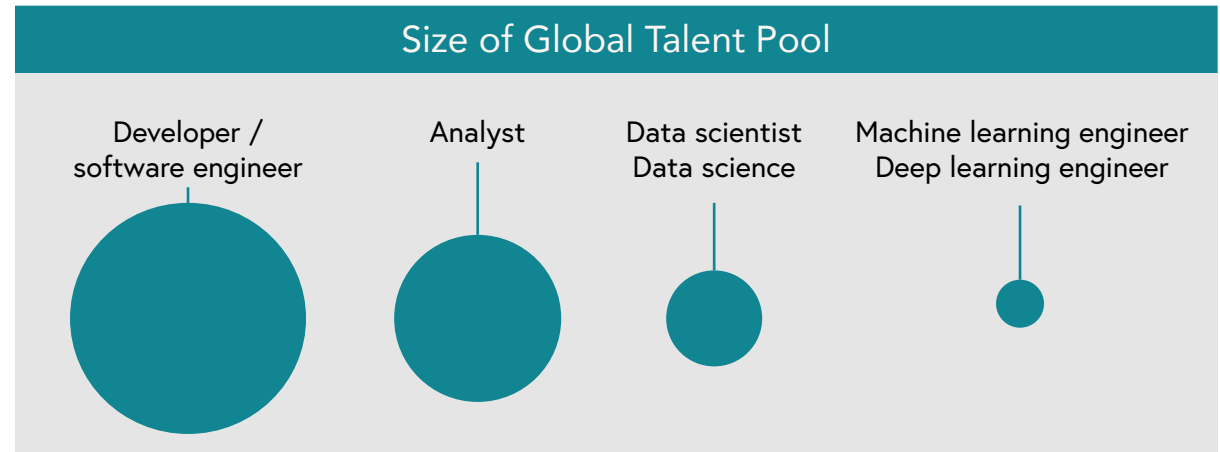


Section IV: Increased efforts to democratize machine learning and make it more accessible to non-experts

There are many startups using AI to build tools to boost developer productivity, and some are building low/no-code tools to open up programming tasks to non-coders.

With demand for AI and machine learning rising, an encouraging sign is that tools for building, deploying and maintaining models continue to get better. However, many of these tools require experimenting with different models, hyperparameter tuning, as well as the judgment of data scientists who have some familiarity with the domain and the underlying data.

As we noted in a recent [post on AutoML](#) and [time series](#), there is an enormous pool of potential contributors (developers and analysts) with limited backgrounds in machine learning and statistics. Thankfully, there are startups, companies, and open source projects focused on making ML more accessible to non-expert users.



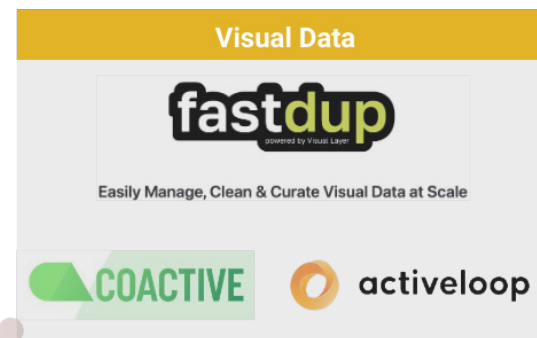
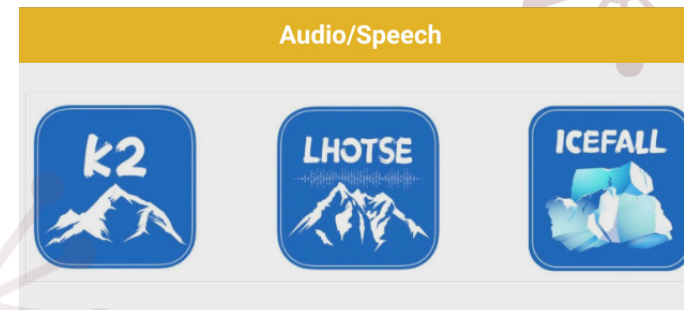
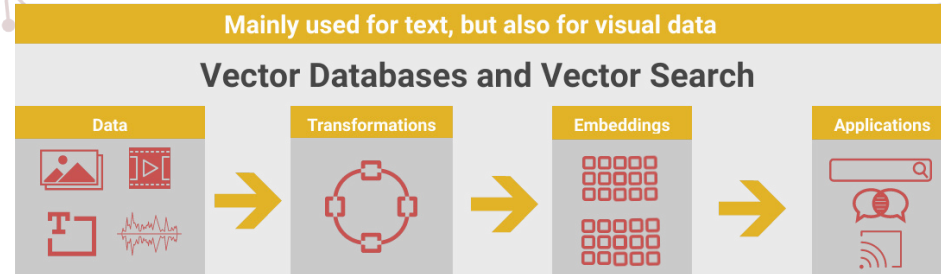
Section V: Tools for unstructured data

Computer vision and speech technologies may have spurred the resurgence in interest in AI (deep learning), but tools for processing, wrangling, storing, and analyzing visual and audio data are lagging. Over the past few months we've come across teams that are focused on making these important data types much more accessible to developers and data science teams. We suspect there'll be more teams focused on tools for visual and audio data over the next year.

- [Introducing a free tool for curating image data sets at scale](#)
- [New open source tools to unlock speech and audio data](#)
- [Deep Lake: A lakehouse for deep Learning](#)

A related set of tools leverages developments in nearest neighbor algorithms and neural models. Vector databases and vector search are on the radar of a growing number of technical teams. Advances in neural networks have made dense vector representations of data more common. Embeddings are common in organizations using neural networks. You can think of vector databases as what you get when you embed your entire database. Hence, vector databases are a family of database-managed architectures that allow you to integrate AI into your data management system.

- [The Vector Database Index](#)



Section VI: Renewed focus on streaming (and data integration)

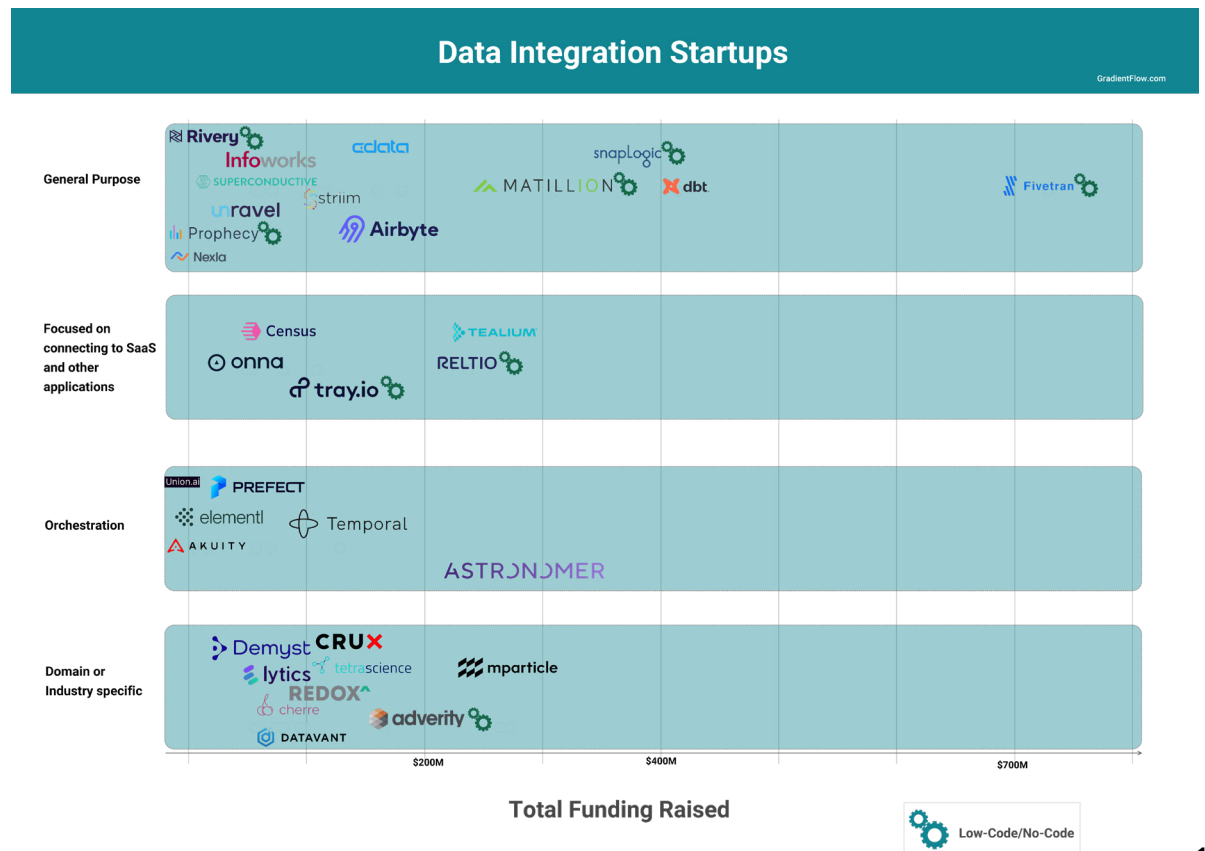
Data integration will continue to be a very active and exciting area, with new tools forging ahead toward higher reliability, ease of use, more connectors, improved orchestration, monitoring, and observability. Some of the best solutions come from companies that offer broad platforms (e.g., Databricks has outstanding [offerings in this area](#)).

[Project Lightspeed](#) (next-gen Spark Streaming) is just the most recent indicator of renewed interest in streaming and streaming applications. Our belief is that companies that do streaming first will gain a decisive advantage over the next few years.

- [The Stream Processing Index](#)
- [The Data Integration Market](#)
- [Summer of Orchestration](#)

Companies that master streaming will have a decisive edge.

One interesting trend is the rise of low-code/no-code tools for data integration and data pipelines.



Section VII: Data engineers are focusing more on operational tasks

The rise of cloud warehouses and [lakehouses](#) means data engineers and data platform teams can get more done—at scale—compared to a few years ago, when teams had to piece together and manage a variety of tools. Their role is shifting from infrastructure development to operational tasks.

Emerging areas of focus include:



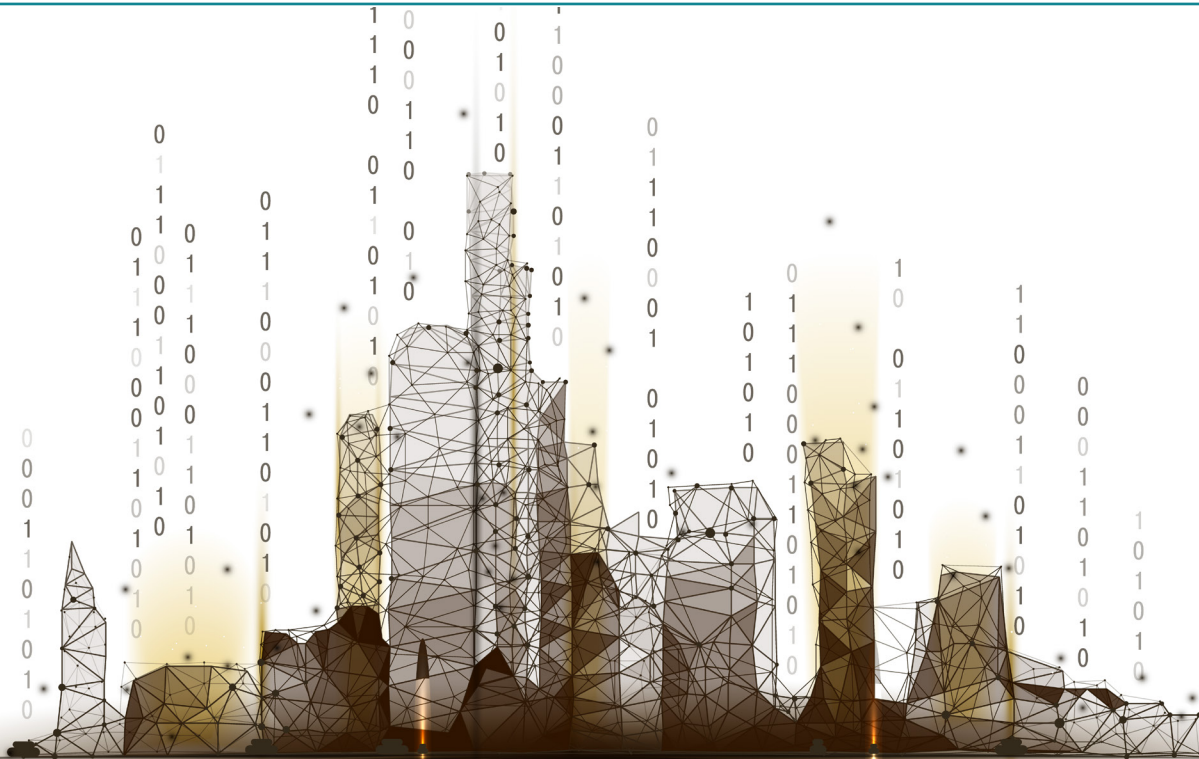
Ops: reliability, automation, monitoring and observability, and incident response. A new wave of orchestration solutions are helping.



Data asset governance and lineage



Cloud computing cost management and optimization (sometimes referred to as [FinOps](#))



Unlike data privacy where companies could organize their initiatives around a few major frameworks (GDR, CCPA), AI compliance involves many more regulatory frameworks that cover disparate business areas. According to our friends at [BNH](#)—the first U.S. law firm focused on AI and analytics—the best step companies can take over the next year is to operationalize the [NIST AI Risk Management Framework](#). Operationalizing the NIST framework, demonstrates good faith and helps reduce risks associated with AI.



Partial list of regulations on the brink of passing or already in place:

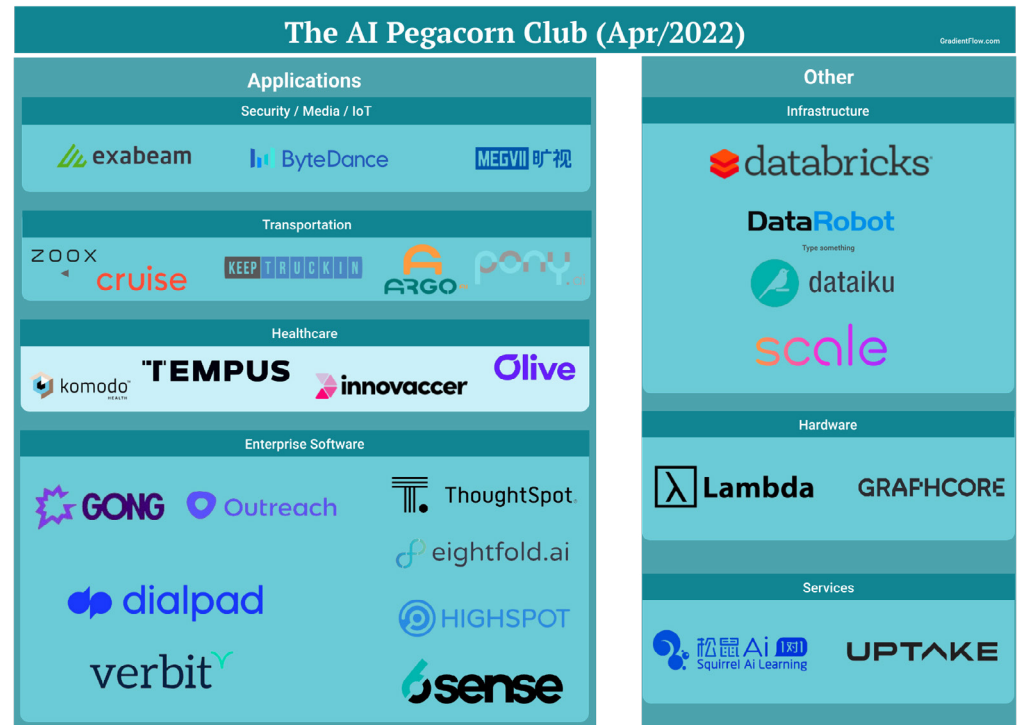
- **EU AI Act**: Establishes a process for self-certification and government oversight of high-risk AI systems, transparency requirements for AI systems interacting with people, and seeks to ban a few "unacceptable" qualities of AI systems.
- **NYC automated employment decision tools**: Pertain to software systems "used to substantially assist or replace discretionary decision-making for making employment decisions that impact natural persons".
 - **New York City delays enforcement of AI bias law**
- **District of Columbia's Stop Discrimination by Algorithms Act of 2021**: Establishes a framework for restricting and requiring businesses that use algorithms to make credit and eligibility decisions, including those directing advertising and marketing solicitations.
- **California's Fair Employment & Housing Council**: A major step toward regulating the use of artificial intelligence and machine learning in conjunction with employment decisions.
- **Blueprint for an AI Bill of Rights (White House)**: Though not a proposed legislation, the White House Office of Science and Technology Policy has identified five principles for designing, using, and deploying automated systems.

Section IX: Applications will continue to lead to more pegacorns

We are in a challenging economic environment for startups. Given the importance of AI and data intelligence for most companies, AI and data startups will continue to thrive even in these tough economic times. In fact, we know of some new AI companies that have achieved pegacorn status (\$100M in annual revenue). As with our [original list](#), these are primarily companies focused on applications. The reason AI application companies tend to do better is that there are more use cases at the application layer than at the infrastructure layer where a single company can serve the same purpose across multiple companies (e.g., data management or data integration).

[The AI \\$100M revenue club](#)

[The data pegacorns](#)



How to get into the AI pegacorn club

Revenue

\$100M in annual revenue pieced together from various data sources including Crunchbase, Zoominfo, public announcements, and other media sources

Independent

Privately held standalone company and/or was recently acquired by a larger public company within the past year and operates as a standalone company

Founding Date

Founded around 2012 and later, this is when deep learning breakthroughs started

ML

Has machine learning as a key component of their product offering

Tools for multi-cloud

- Terraform: [An introduction](#)
- New tools like [Skyplane](#) point toward a future when multi-cloud data platforms are more common.
 - Sky Lab: [Skyplane](#), [SkyPilot](#)
 - Sky computing is a new computing model where resources from multiple cloud providers are utilized to create large-scale distributed platforms.

AI and geopolitics

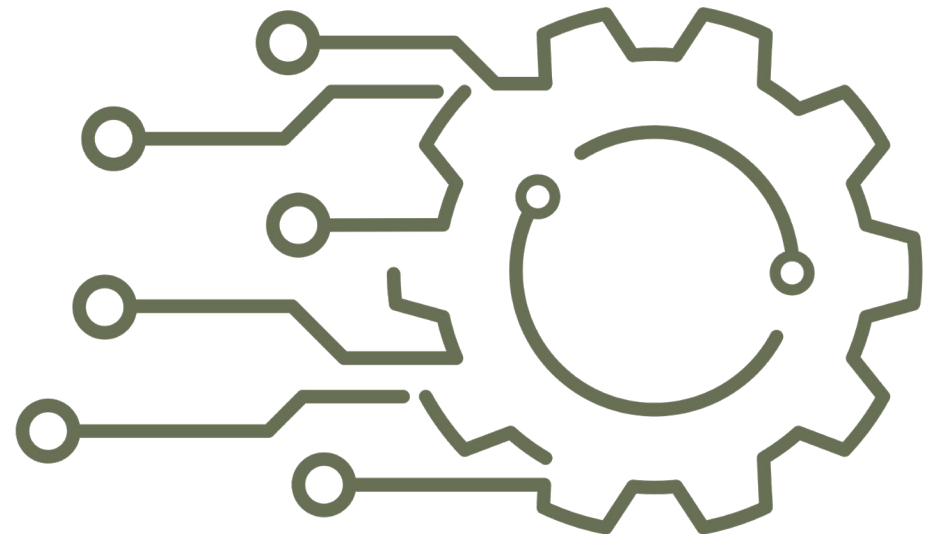
The Biden administration's export controls for "[Certain advanced computing and semiconductor manufacturing items; supercomputer and semiconductor end use](#)," will have an impact on the data and AI communities. In recent years, Chinese companies have ramped up their contributions in research (number of publications at top-tier conferences), open source software projects, and startups. Decoupling China from the rest of the world will have repercussions for AI.

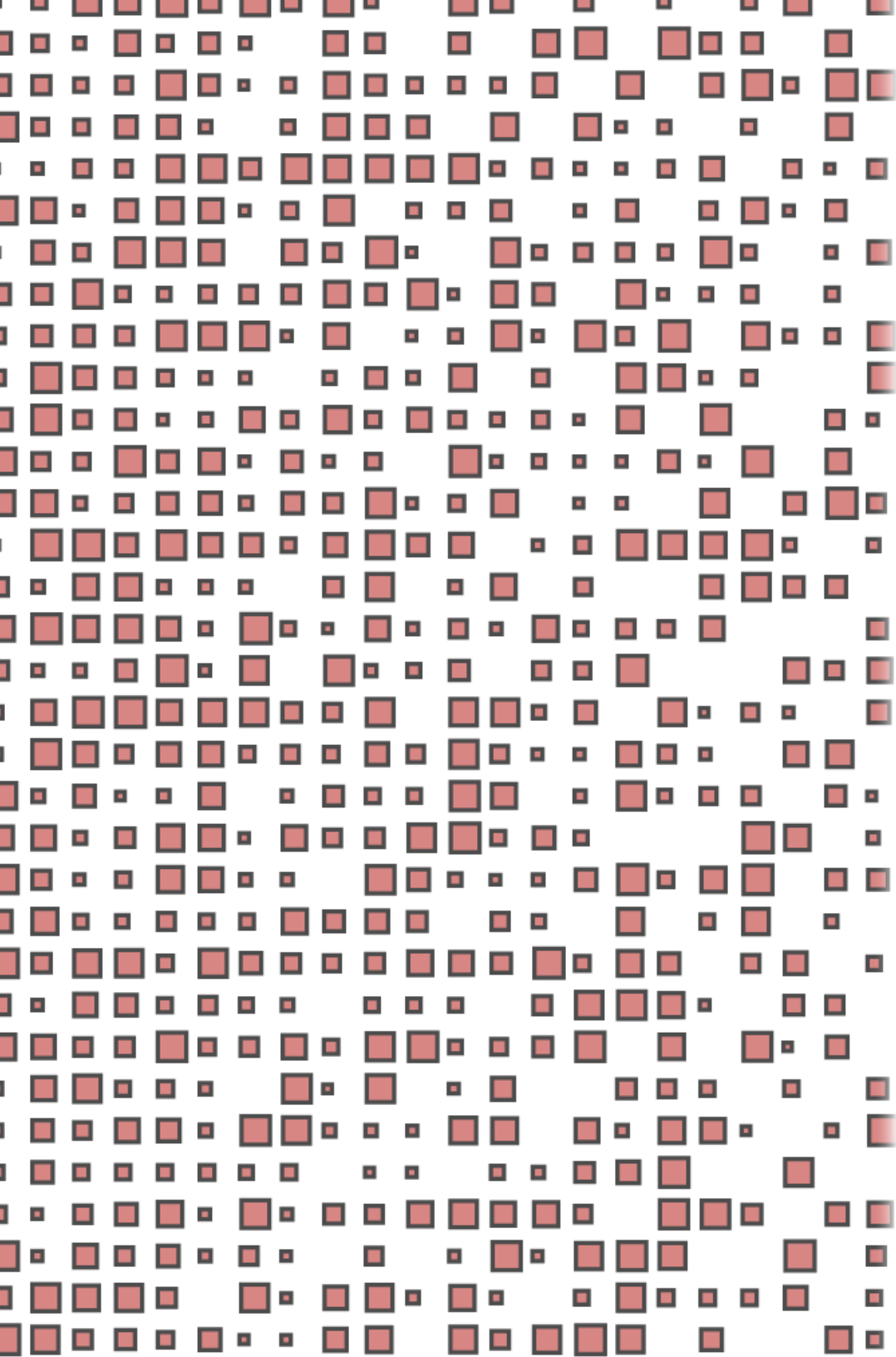
Data and Cybersecurity

Privacy and machine learning: [Measuring the popularity and exploring the readiness of confidential computing tools](#)

Ransomware threats: [Ransomware accounted for \\$20 billion in global losses in 2021](#). According to estimates, that number will grow to \$265 billion by 2031.

Securing your software supply chain: (PyPI) [attacks grew 41%](#) in 2022. Here's a [recent example](#) of such an attack against PyTorch. Given how critical Python is to data engineering, machine learning, and AI teams, here are some tips on [how to secure your Python supply chain](#).



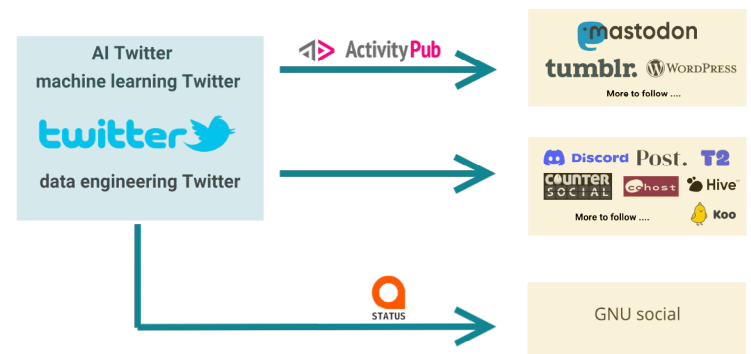


The emergence of Twitter alternatives

With the chaos taking place on Twitter, where will the ML, AI, and engineering communities migrate? There are three emerging venues.

We think a combination of 1 and 2, with a higher proportion going to 1, will be most likely:

1. Federated communities built around [ActivityPub](#), an open, decentralized social networking protocol.
2. New (centralized) platforms.
3. Federated communities based on [OStatus](#), an open standard for federated microblogging.





GRADIENT FLOW

Subscribe to the Gradient
Flow Newsletter to stay up
to date on emerging trends