UC San Diego

CSE/DSC 234
Data-Centric AI and AI Engineering

Spring 2026

Arun Kumar

About Myself



Summers: 110F!

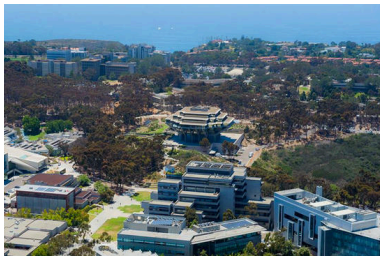
2009: Bachelors in CSE from IIT Madras, India



Winters: —40F!

2009—16: MS and PhD in CS from UW-Madison

PhD thesis area: Data systems for ML workloads



Ahh! :)

2016-: UC San Diego CSE

2019-: + UC San Diego HDSI

2024-: + RapidFire AI, Inc.

My Current Research

New abstractions, algorithms, and software systems
to **"democratize"** ML/AI from a
data management/systems standpoint

Democratization = **System Efficiency (Lower time/costs)** **+** **Human Efficiency (Higher productivity)**

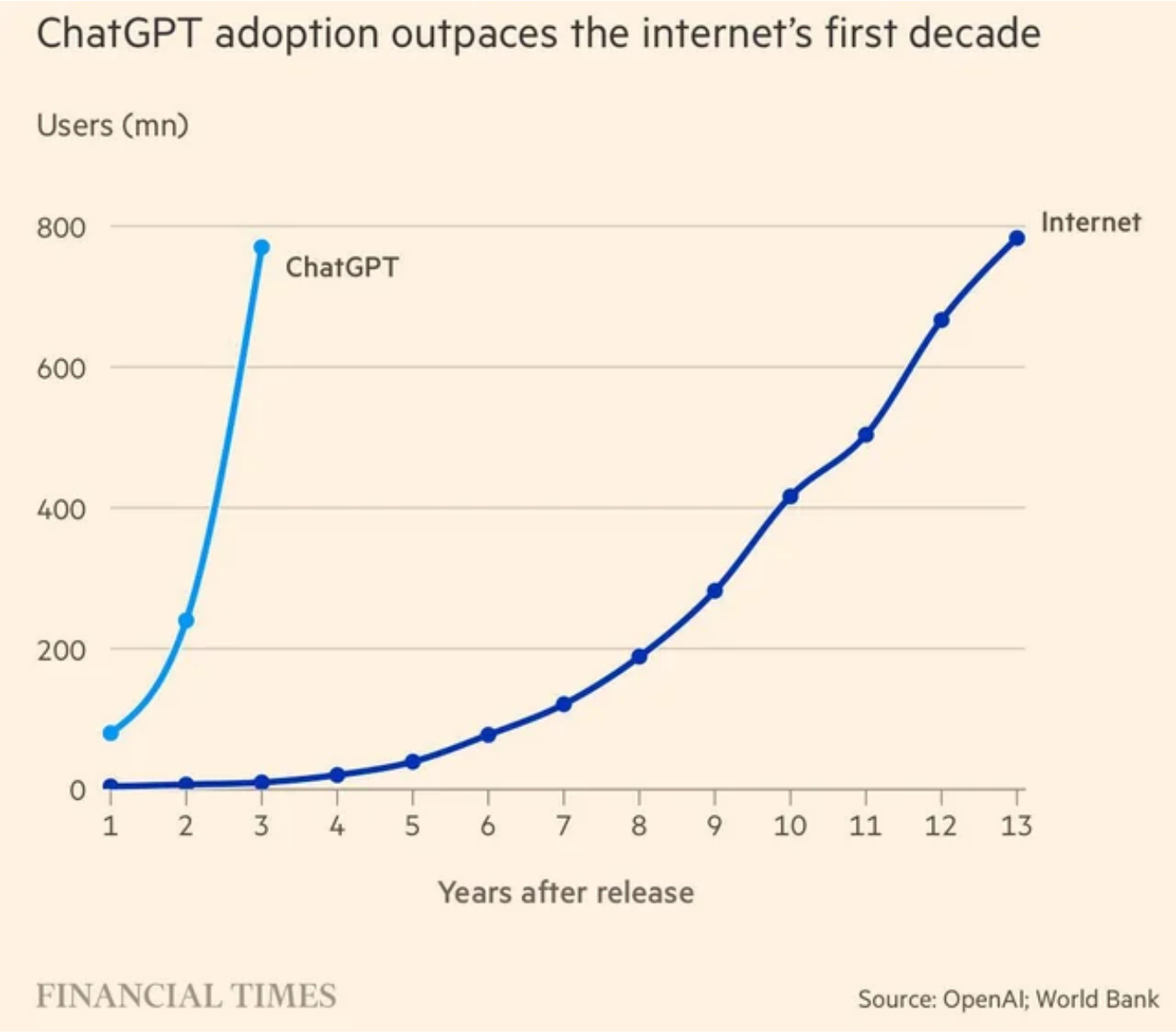
Practical and scalable data systems for AI

Inspired by principles of *databases and systems*

Exploit insights from *learning theory* and *optimization theory*

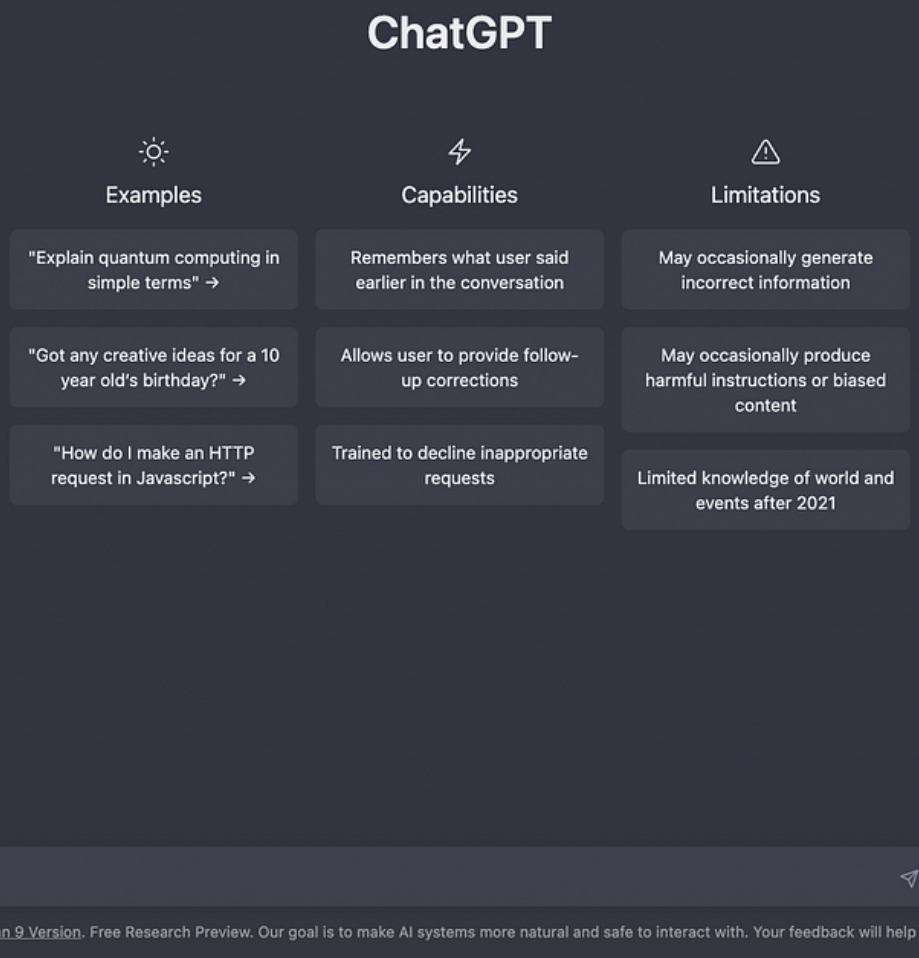
What is this course about? Why take it?

1. ChatGPT Takes the World by Storm!

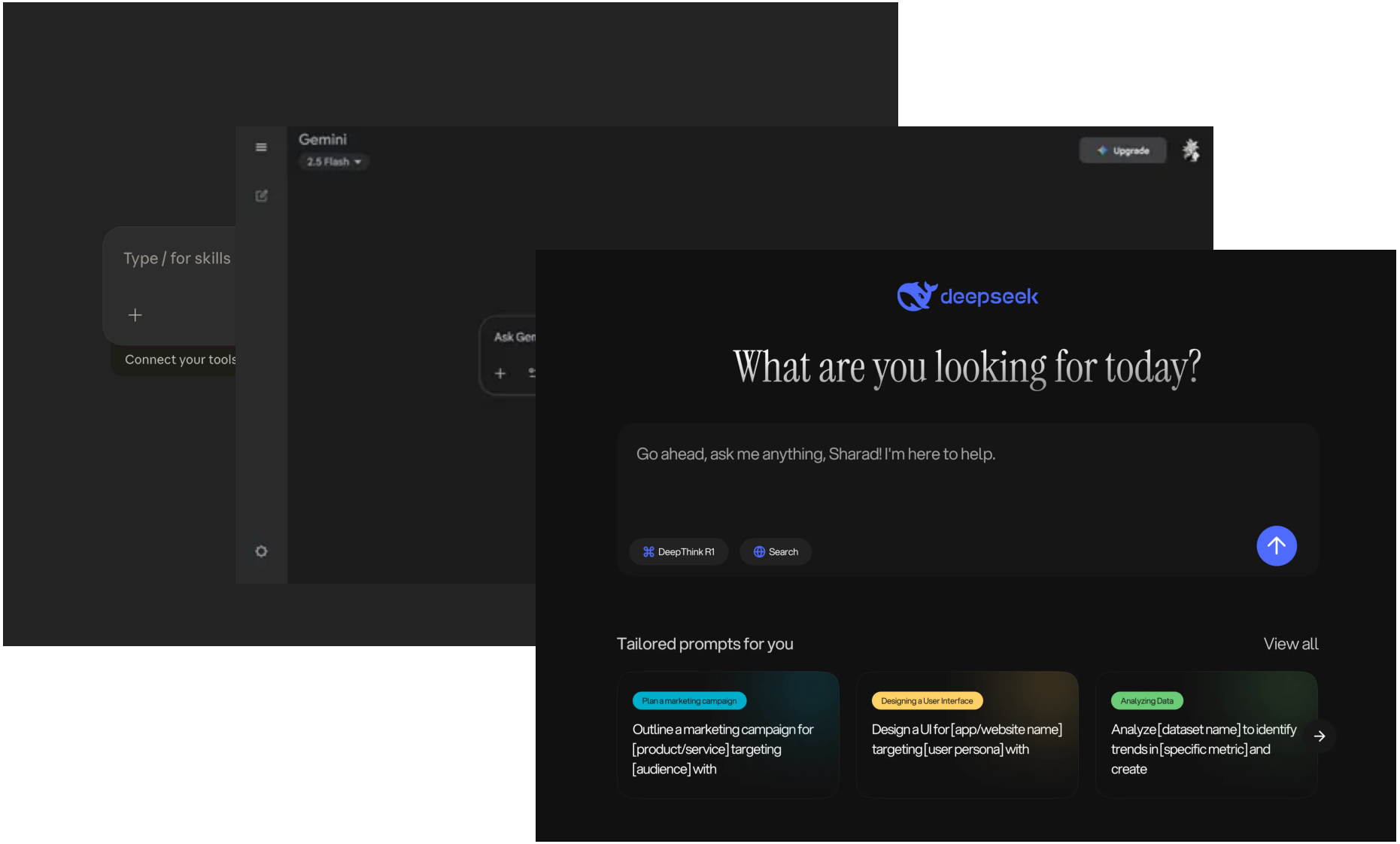


Why was ChatGPT a pivot point in AI?

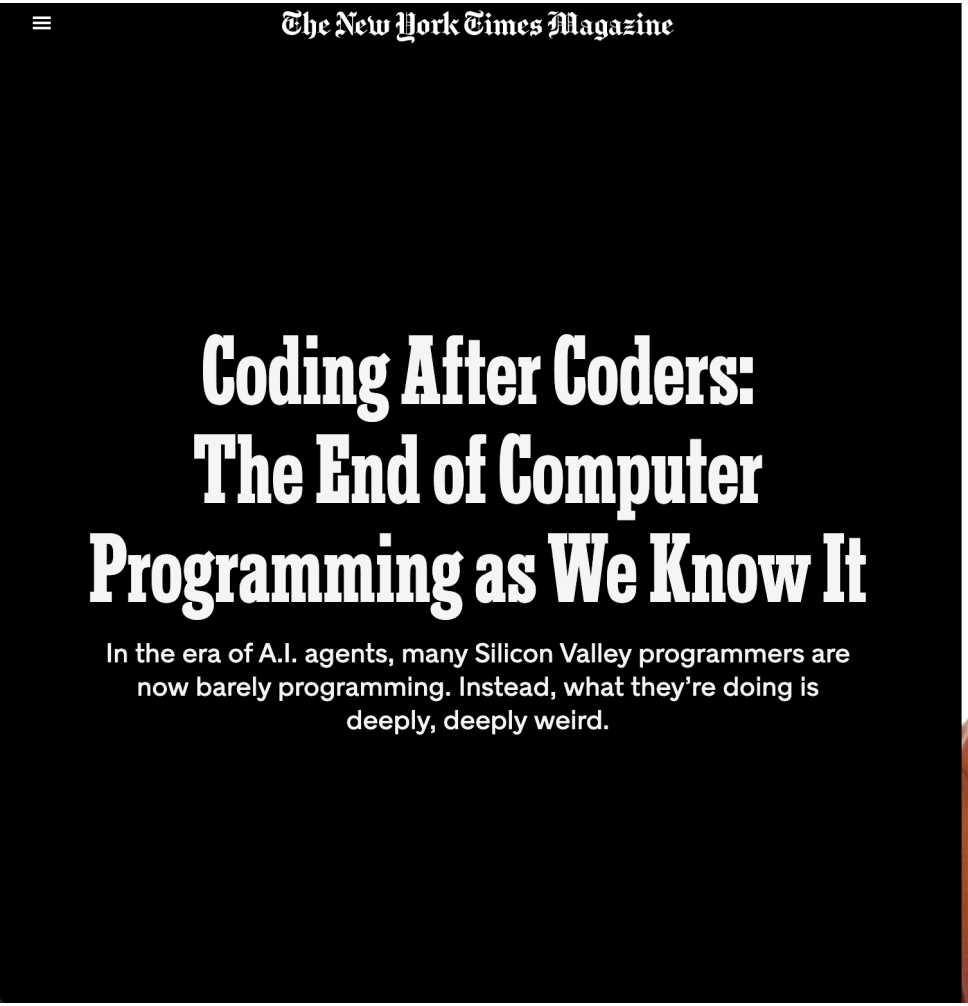
From Task-Specific AI to General AI (AGI)



Cambrian Explosion of (Multimodal) LLMs



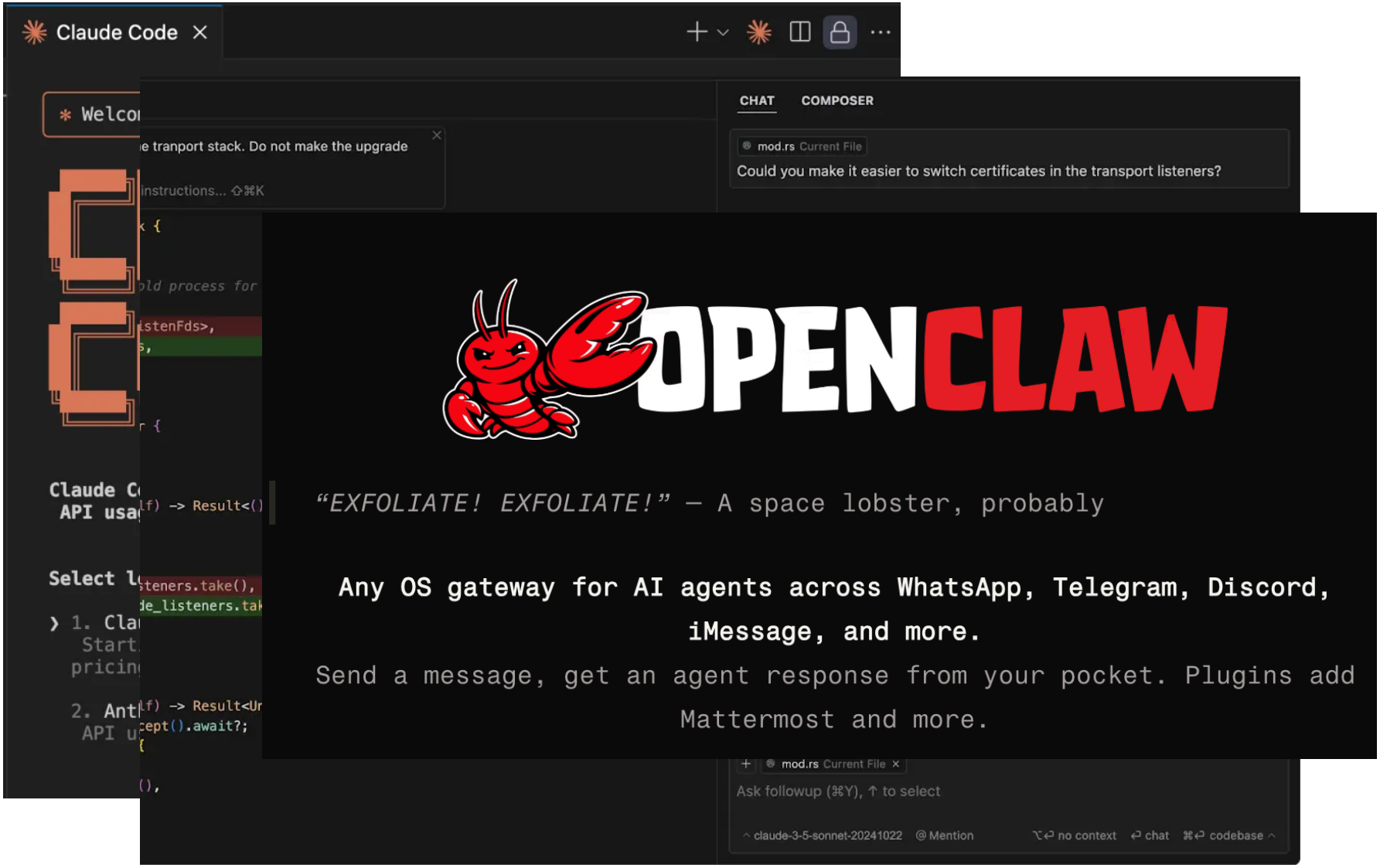
2. The End of Coding as We Know/Knew It





Who will write and test code now?

AGI as Pair Programmer/Coding Team!

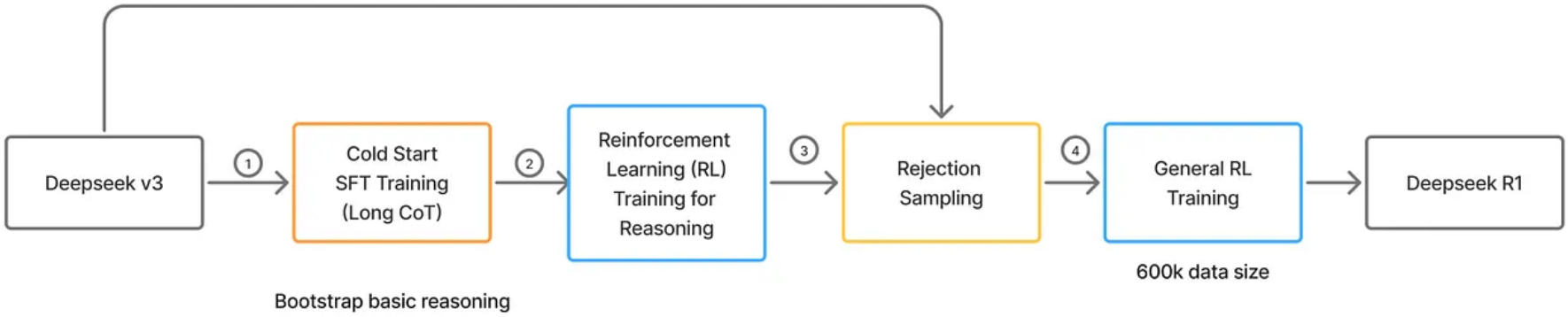


3a. DeepSeek R1 Wipes Out 1 Trillion in US Stocks



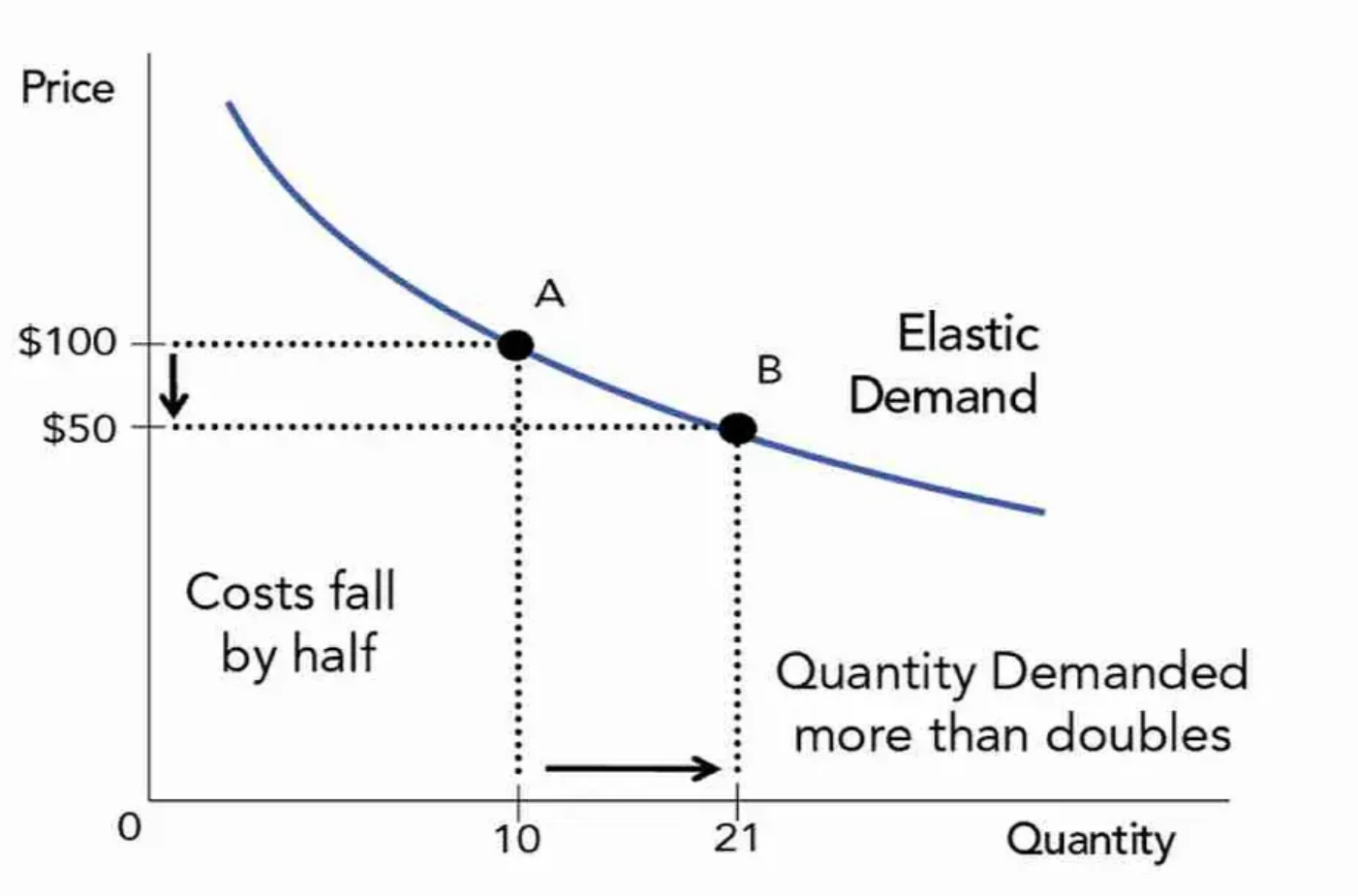
Why did the markets freak out?

DeepSeek Moment: Beyond More Data + Compute



https://rajeshparikh.substack.com/p/qwen-3-and-deepseek-r1s-post-training
https://arxiv.org/abs/2501.12948

But … Jevons Paradox is Real



3b. Claude Triggers Collapse in SaaS Stocks

FEATURES

Anthropic: AI triggers the 'SaaSpocalypse'

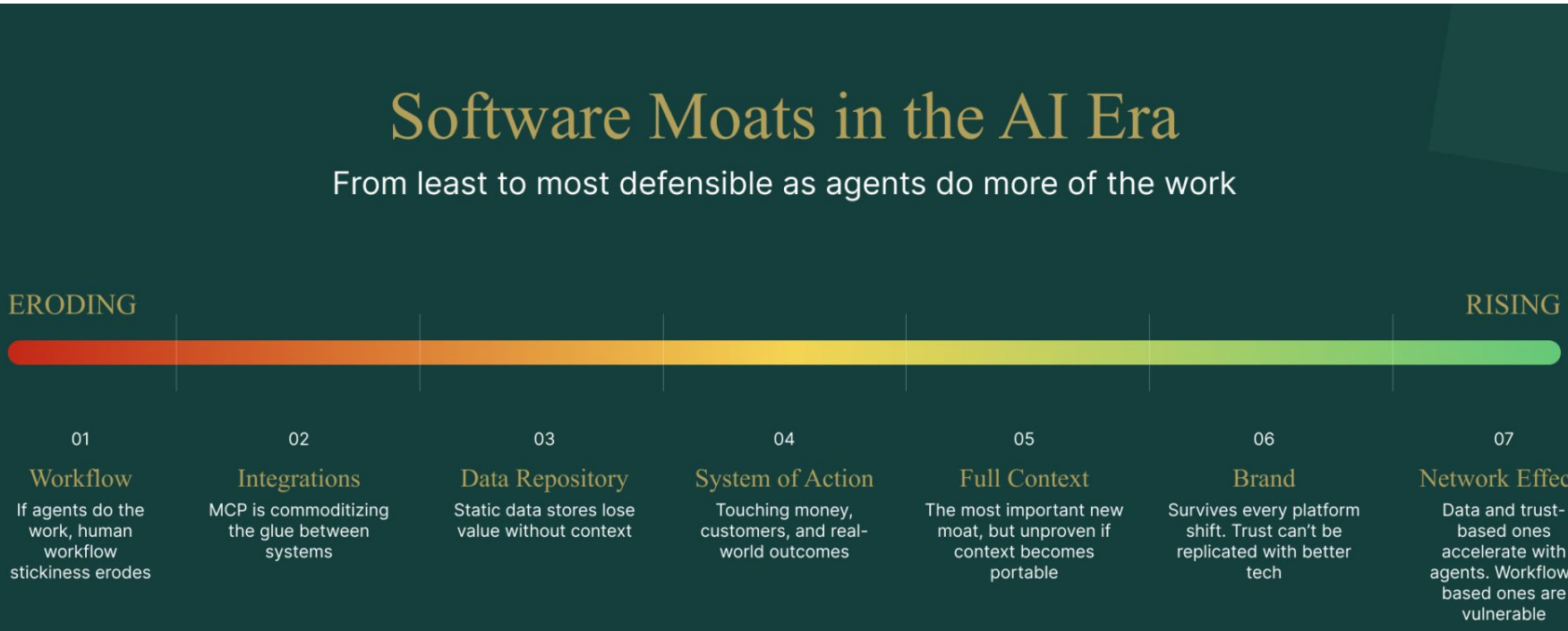
A grim reaper for software services?

BY THE WEEK US PUBLISHED FEBRUARY 19, 2026



Why did the markets freak out, again?!

Software Moats are Indeed Under Threat



But … SaaS as a Category is Not Dead

Wired for Scale: Sid Rao's Musings

The Software Apocalypse That Isn't

As Usual, We Believe That AI Agents and SaaS are in a Zero Sum Game. This Investor Behavior Really Foreshadows the Coming AI Bubble, Doesn't It?



SID RAO
FEB 24, 2026

9 1 2 Share

Let me tell you something about apocalypses. They're terrific for selling newspapers. They're less terrific at predicting the future.

Right now, there's a fever running through Sand Hill Road and the analyst desks of every major bank. The thesis goes like this: agentic coding can build software applications in microseconds, therefore no enterprise customer will ever buy a SaaS subscription again, therefore every SaaS company is dead, therefore sell everything, therefore the sky is falling. It's a neat little syllogism. It's also wrong. Not entirely wrong — which is what makes it dangerous — but wrong in the ways that matter.

https://www.srao.blog/p/the-software-apocalypse-that-isnt

Innumerable enterprise applications are being disrupted by AGI


indiamart


DASHBOARD





Domain sciences also being disrupted*

Pre-LLM era Deep Learning (DL) already disrupted most natural sciences!

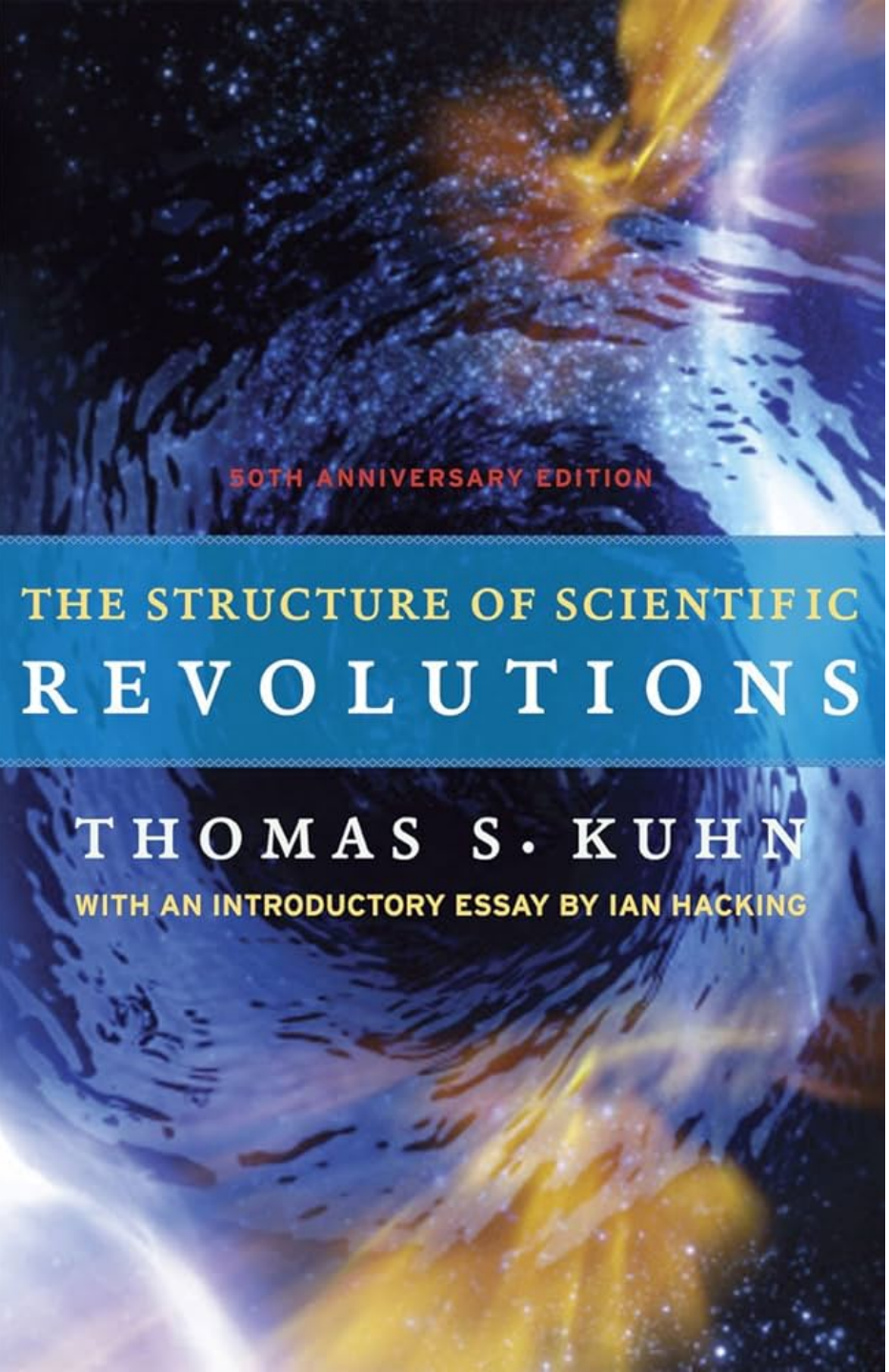

SOCIAL MEDIA ANALYSIS

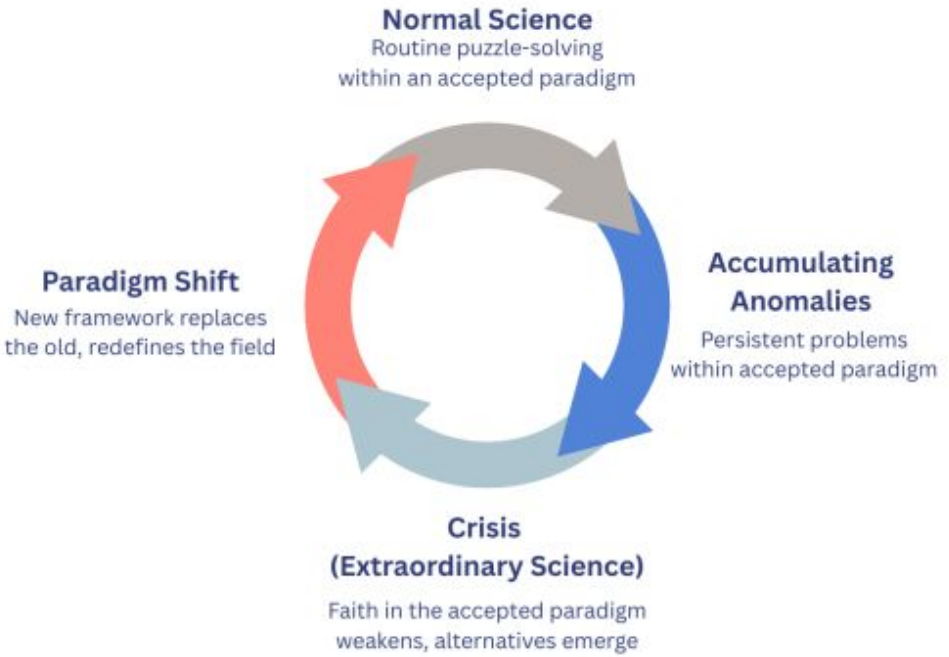





Edtech

We are not living in “normal” times. :)


50TH ANNIVERSARY EDITION
THE STRUCTURE OF SCIENTIFIC
REVOLUTIONS
THOMAS S. KUHN
WITH AN INTRODUCTORY ESSAY BY IAN HACKING


Normal Science
Routine puzzle-solving
within an accepted paradigm
Accumulating
Anomalies
Persistent problems
within accepted paradigm
Crisis
(Extraordinary Science)
Faith in the accepted paradigm
weakens, alternatives emerge
Paradigm Shift
New framework replaces
the old, redefines the field

But what more is there to all this than just taking a bunch of ML/AI courses?

Academic ML/AI

Machine Learning Algorithms

Deep Learning Algorithms

Large Language Models

Natural Language Processing

Information Retrieval

Computer Vision

…

Real-World ML/AI 101

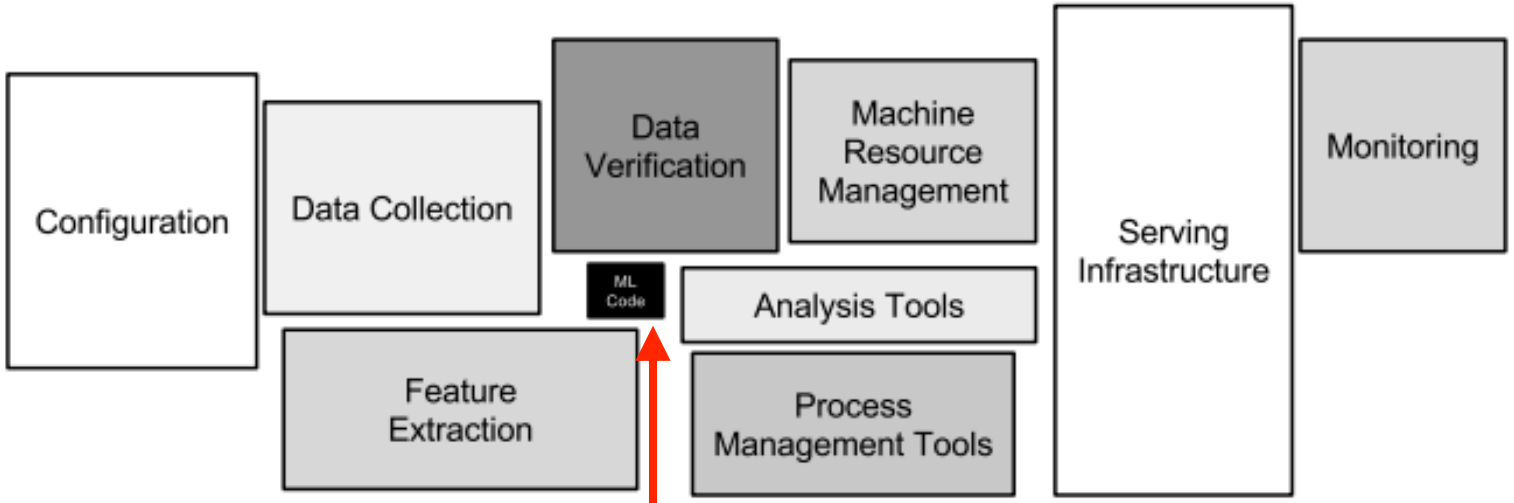
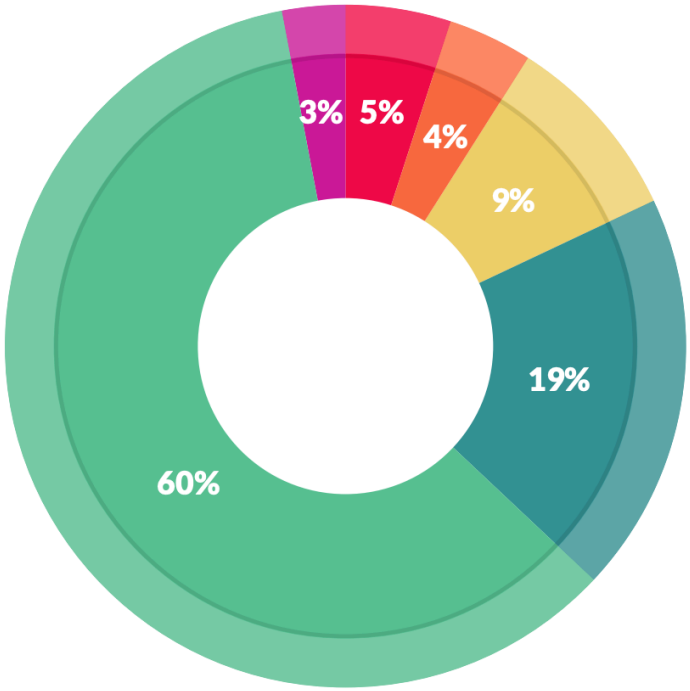


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Almost all of your ML / AI courses put together! :)

Real-World ML/AI 101



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

This was in 2016!
If anything, "mining/algorithms" parts have shrunk further. :)

Real-World ML/AI 101

The Rise of “Data-Centric” AI



Real-World ML/AI 101



Tools & methodologies for accelerating open-source dataset iteration:

- Tools that quantify and accelerate time to source and prepare high quality data
- Tools that ensure that the data is labeled consistently, such as label consensus
- Tools that make improving data quality more systematic
- Tools that automate the creation of high quality supervised learning training data from low quality resources, such as forced alignment in speech recognition
- Tools that produce consistent and low noise data samples, or remove labeling noise or inconsistencies from existing data
- Tools for controlling what goes into the dataset and for making high level edits efficiently to very large datasets, e.g. adding new words, languages, or accents to speech datasets with thousands of hours
- Search methods for finding suitably licensed datasets based on public resources
- Tools for creating training datasets for small data problems, or for rare classes in the long tail of big data problems
- Tools for timely incorporation of feedback from production systems into datasets
- Tools for understanding dataset coverage of important classes, and editing them to cover newly identified important cases
- Dataset importers that allow easy combination and composition of existing datasets
- Dataset exporters that make the data consumable for models and interface with model training and inference systems such as webdataset.
- System architectures and interfaces that enable composition of dataset tools such as, MLCube, Docker, Airflow

https://datacentricai.org/

Real-World ML/AI 101

Transitioning from Academic Machine Learning to AI in Industry



Jeremy Karnowski Follow
Aug 9, 2017 · 6 min read

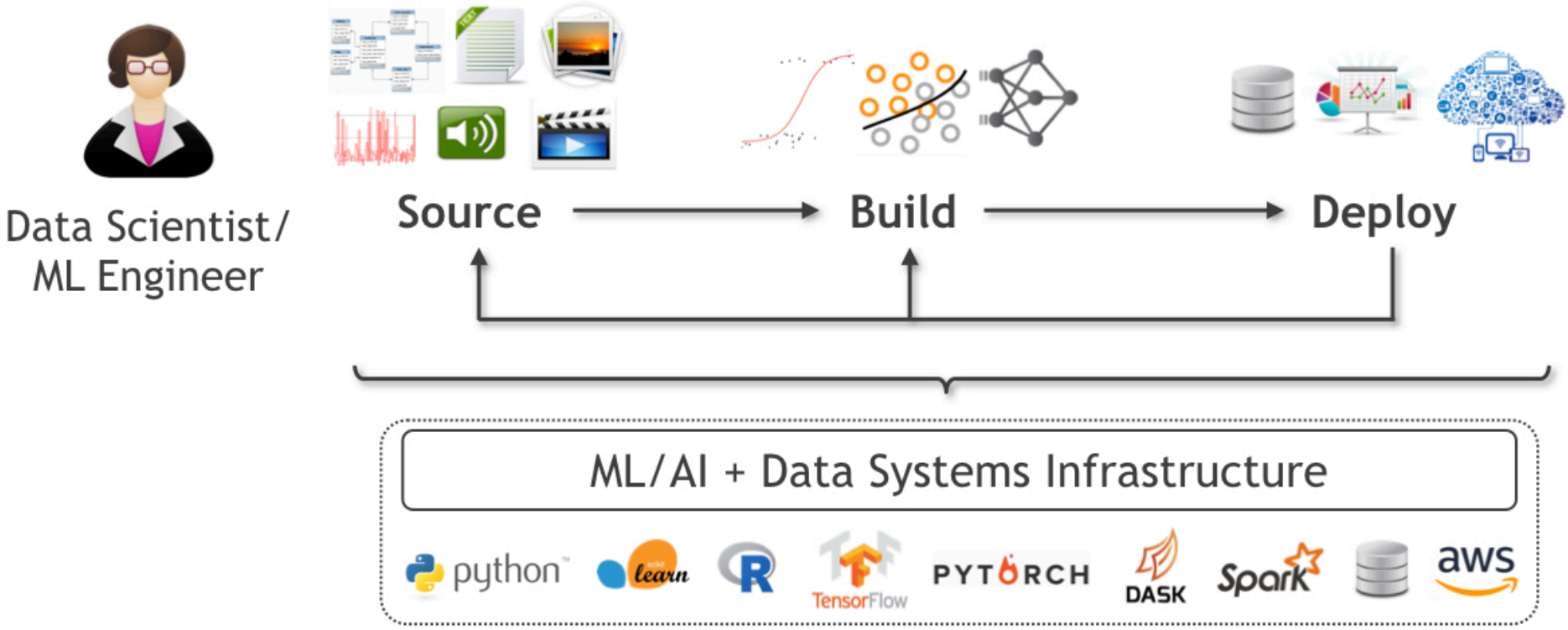
Jeremy Karnowski and Emmanuel Ameisen, Insight AI

1. System design
2. Structured ML modules
3. Software testing
4. Integrating with data infrastructure
5. Model serving

All of these are still relevant but in a new form!

https://blog.insightdatascience.com/preparing-for-the-transition-to-applied-ai-8eaf53624079

The Old Lifecycle: ML/DL Engineering

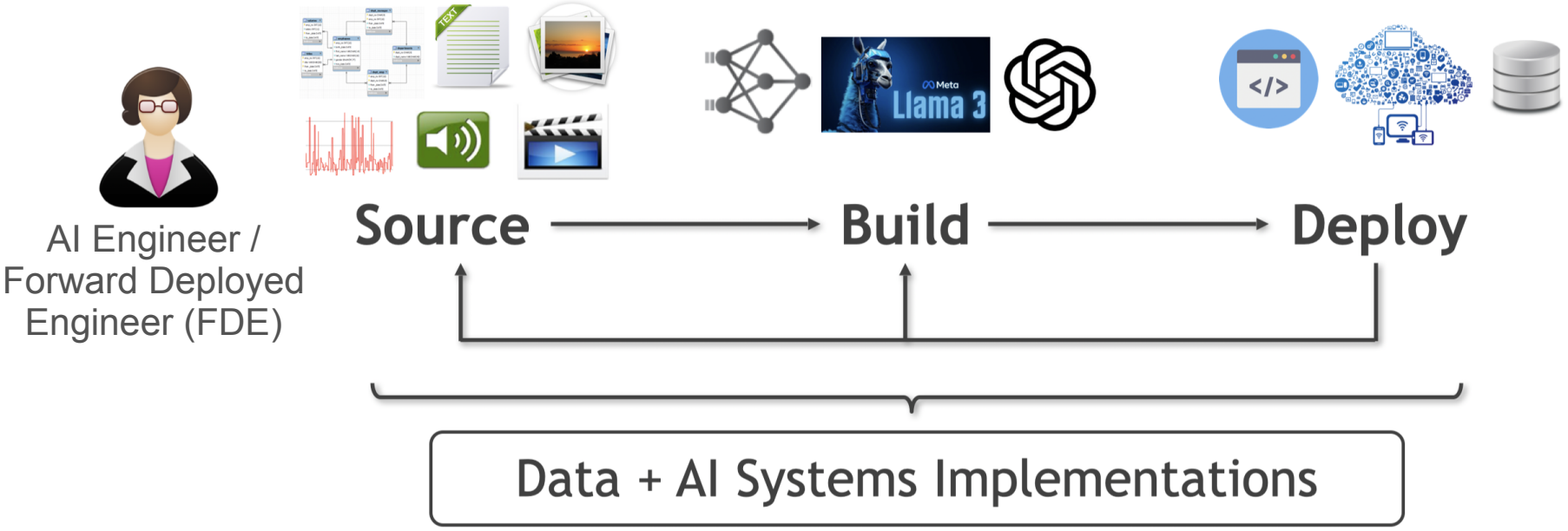


Data Acquisition
Data Preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

The New Lifecycle: AI Engineering



Data Acquisition
Data Preparation
Create Evals

Context Engineering
Agent Engineering
LLM Customization

Guardrails
Monitoring

Out with the Old, in with the New!



Arun Kumar · You
Cofounder and CTO of RapidFire AI; Associate Professor at the Universit...
6d ·

Forward Deployed Engineers are the new DL Engineers,
who were the new ML Engineers,
who were the new Data Scientists,
who were the new Statisticians.

What goes around comes around. 🤭

Okay TBF the evolution of methods and professional roles has been rapid. But the utility of data + algorithms + systems for applications is eternal. Evolve accordingly. 😊

Out with the Old, in with the New!



Arun Kumar • You
Cofounder and CTO of RapidFire AI; Associate Professor at the Universit...
3w •

Agent engineering is the new context engineering,
which was the new model engineering,
which was the new feature engineering,
which was the new knowledge engineering.

The more things change, the more they stay the same. 😘

Okay TBF each of those leaps was a major upgrade in the complexity of tasks possible with AI but the hedonic treadmill means we never stop. 😄

Out with the Old, in with the New!

#ai #artificialintelligence #forwarddeployedengineer #fde

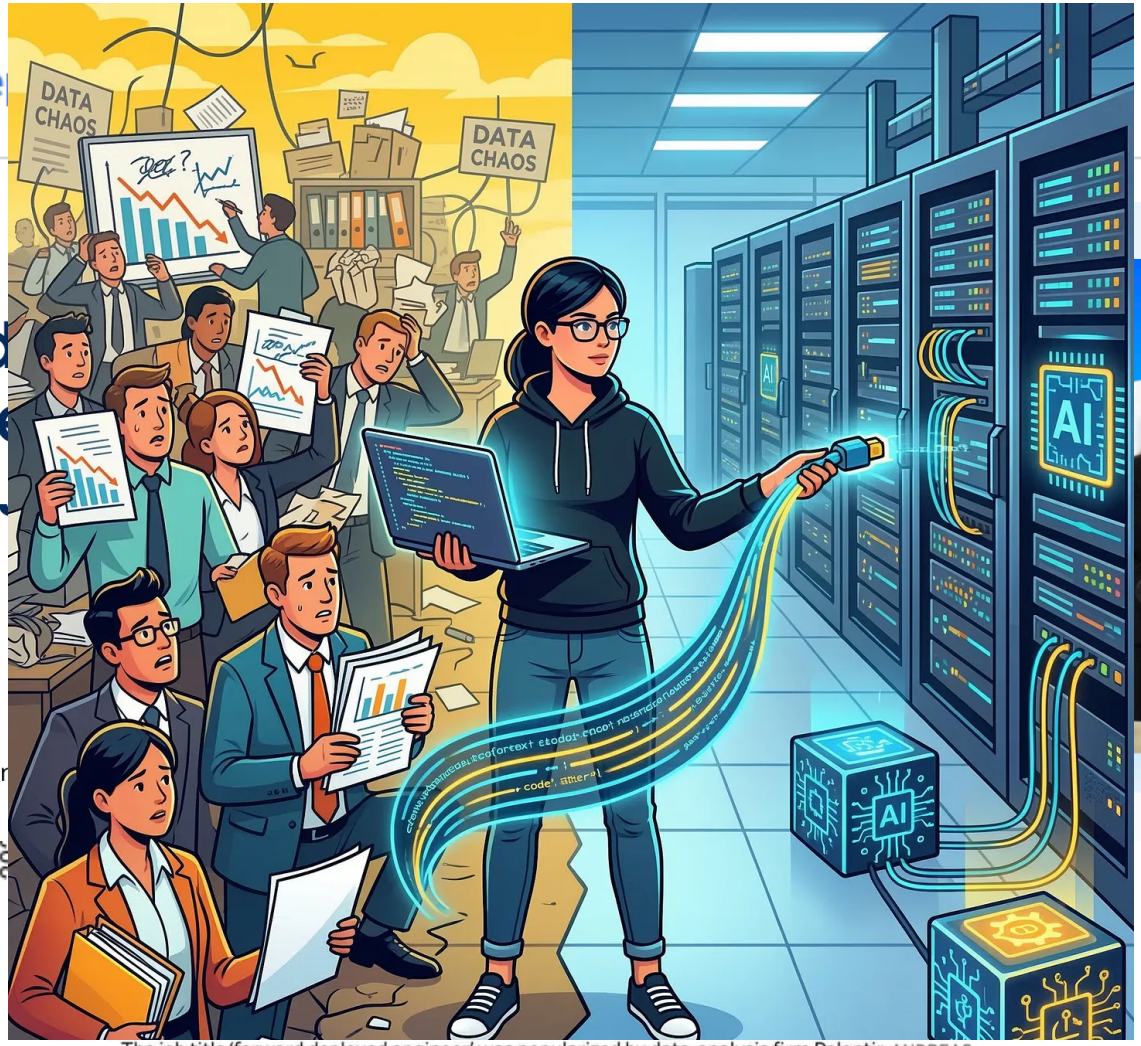
Forward de

50

Future of Work

How Forward
Engineers Ar
Makes Tech
Human

Lydia Dishman
Contributing Author
March 18, 2026 • 5 mi

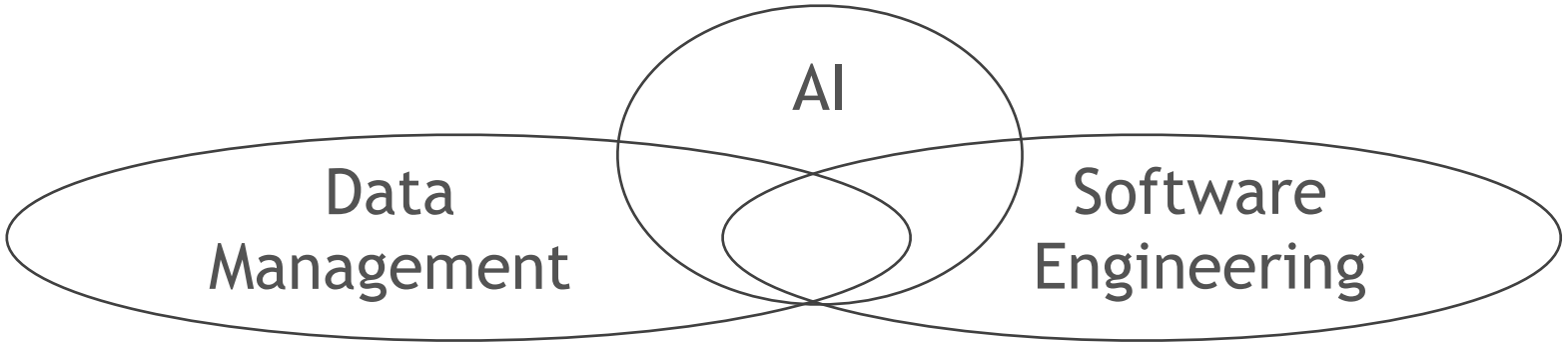




The job title 'forward deployed engineer' was popularized by data-analysis firm Palantir. ANDREAS BECKER/EPA/SHUTTERSTOCK

FDEs are basically “AGI babysitters” for a product team. :)

CSE/DSC 234 will get you to think about the data and engineering aspects of A(G)I applications



1. **"Data-Centric AI …":** How to acquire, annotate, synthesize, and curate data and data-informed evals?
2. **"… AI Engineering":** How to go from data + LLMs/AGI APIs to shippable software systems?
3. **3 Main Stages:**
   - 3.1. **Source**: Raw application -> AI-ready data + evals
   - 3.2. **Build**: AI-ready data + evals -> A(G)I pipelines
   - 3.3. **Deploy**: Productionize A(G)I pipelines

Focus on Practical Concerns

Key concerns in AI algorithmics:

Accuracy/Quality Metrics

Runtime efficiency (sometimes)

Additional key *practical* concerns in AI engineering:

Usability + Reliability

Scalability + Efficiency at scale

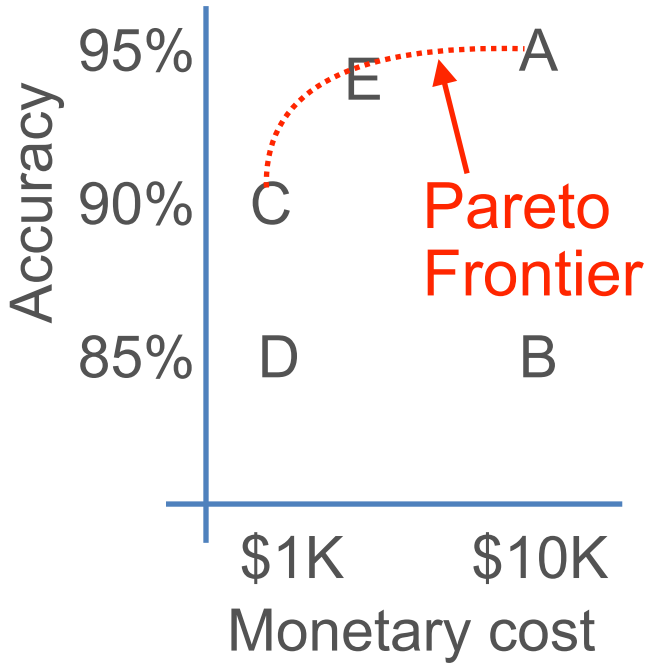
Manageability

Total Cost of Ownership (TCO)

Real-World AI: Pareto Surfaces

Q: *Suppose you are given 4 LLM pipelines A, B, C, and D with accuracies of 95%, 85%, 90%, and 85%, resp. for a spam detection task. Which one will you pick?*

Q: *What about now?*



- Real-world AI applications grapple with multi-dimensional *Pareto surfaces*: accuracy/reliability metrics, TCO, inference latency, tool availability, explainability, compliance metrics, etc.
- *Multi-objective optimization* criteria set by application needs and business policies.

Learning Outcomes of this Course

- Understand the data-centric engineering challenges across the AI application lifecycle—from creating data and evals to retrieval, customization, and deployment—and how to handle them in practice.
- Build and evaluate LLM-powered applications (RAG, SFT, agents, etc.) using systematic experimentation rather than ad hoc prompting.
- Reason about practical tradeoffs between accuracy/eval metrics, cost, latency, reliability, etc. in real AI applications.
- Think critically about the fast-evolving landscape of AI engineering tools, methodologies, and R&D to discern from the hype what actually matters in production.

What this course is NOT about

- NOT a course on basics of AI or databases
  - Sanity check! You must know what all these terms mean: LLM, prompt, tokenization, embedding, transformer, gradient descent, fine-tuning, GPU
- NOT a course on ML/AI *algorithmics*; we focus on the data and engineering systems *around* LLMs
- NOT a course on how to just use/prompt LLM APIs
- NOT a course on ML systems *internals*; take Hao's DSC/CSE 291 (Deep Learning Systems) for that

Now for the logistics …

Prerequisites

- A full course on **ML algorithms** (e.g., CSE 151 or 258) is absolutely necessary. It could have been taken at UCSD or elsewhere.
- **Python programming** knowhow is also necessary.
- Introductory courses on NLP/LLMs and on databases/data management are also highly recommended but not strictly required.
- Substantial project or industrial experience on relevant topics can be substituted for prior coursework and Python experience, subject to the instructor's consent. Email the instructor if you would like to enroll but are unsure if you satisfy the prerequisites.

https://cseweb.ucsd.edu/classes/sp26/cse234-a/index.html

Components and Grading

- Midterm Exam: 15%
- Cumulative Final Exam: 35%
- AI Engineering Project 1: 20%
- AI Engineering Project 2: 20%
- Peer Instruction (PI) Activities: 10% (8 x 1.25%)
- See course website for more details on A(G)I use for Projects, Academic Integrity policies, Exam logistics, Midterm "Second Chance", and PI Activities logistics

https://cseweb.ucsd.edu/classes/sp26/cse234-a/index.html

Grading Scheme

Hybrid of absolute and relative; grade is the **better** of the two

Grade	Absolute Cutoff (>=)	Relative Bin (Use strictest)
A+	95	Highest 5%
A	90	Next 15% (5-20)
A-	85	Next 15% (20-35)
B+	80	Next 15% (35-50)
B	75	Next 15% (50-65)
B-	70	Next 10% (65-75)
C+	65	Next 5% (75-80)
C	60	Next 5% (80-85)
C-	55	Next 5% (85-90)
D	50	Next 5% (90-95)
F	< 50	Lowest 5%

Example: Score 89 and 60%le; Rel: B+; Abs: A-; so, A-

Tentative Course Schedule

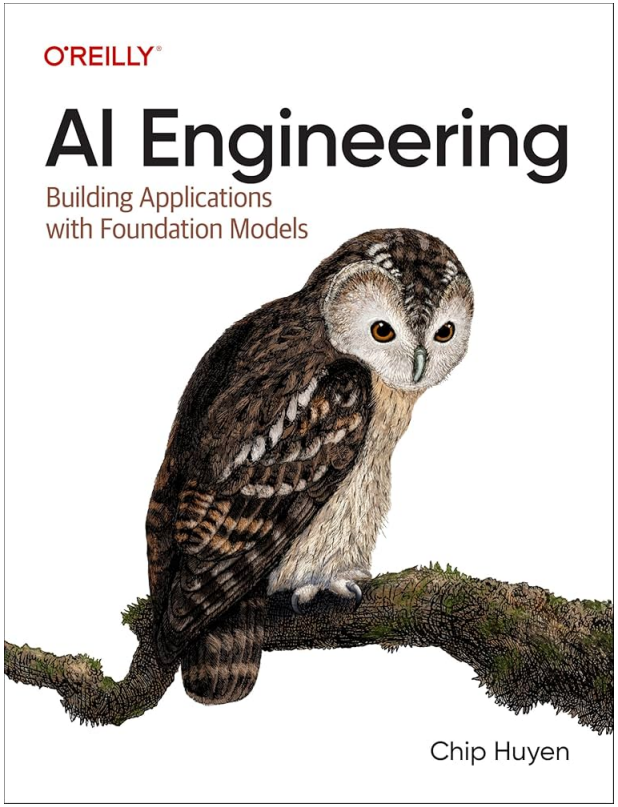
Week	Topic	Slides
1	**Introduction; From AI to AGI** (Task-Agnosticism, Domain-Generality, Promptability)	-
1-2	**Topic 1: Creating Data and Evals** (Feedback, Annotation, Structure, Curation, Synthesis, Multimodality, LLM-as-Judge, Code/SLMs for Evals)	-
2-3	**Topic 2: Retrieval and Context Engineering** (Prompt Eng., VectorRAG, GraphRAG, Search, Retrieval vs. Long Context)	-
4-5	**Topic 3: Agent Engineering** (Tool Use, Memory, Reasoning, MCP, A2A, Multi-Agent Workflows, Observability)	
6	**Invited Guest Lecture: Ajay Rayasam (Osage University Partners) on Tue, May 5**	-
6	**Review Discussion** on Topics 0-3 on **Wed, May 6**	-
6	**Midterm Exam** on Topics 0-3 on **Thu, May 7**	-
7	**Invited Guest Lecture: Lakshya Agrawal (UC Berkeley) on Tue, May 12**	-
7-8	**Topic 4: Model Customization** (SFT, RFT, Continued Pretraining, Transfer Learning, Training vs. Inference-Only Customization)	-
8	**Invited Guest Lecture: Kabir Nagrecha (Tessera Labs) on Thu, May 21**	-
9	**Topic 5: AutoML in the AGI Era** (Optuna/HyperOpt to AFlow/MASS/GEPA/etc.)	-
9-10	**Topic 6: From MLOps to LLMOps** (Guardrails, Red-teaming, Routing, Caching, Monitoring)	
10	**Topic 7: AGI in Practice** (Software Engineering, Business Intelligence, Higher Education, Sociotechnical Issues)	-
11	**Final Exam on Thu, Jun 11**	-

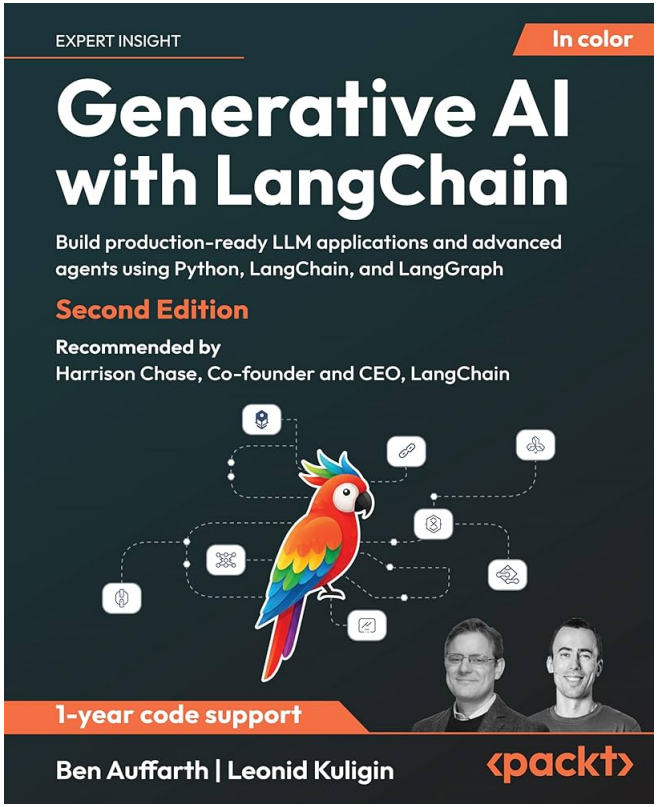
Source

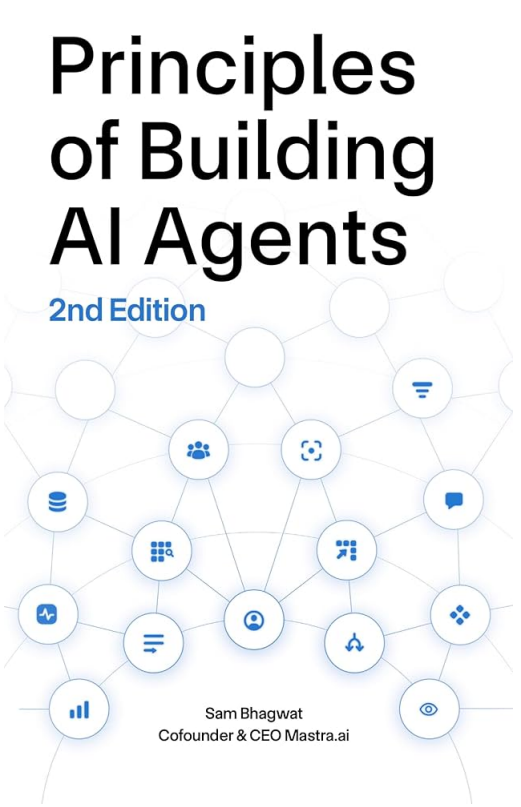
Build

Deploy

Suggested Textbooks







Day-to-day Logistics

- Lectures, Exams, Peer Instruction (PI) Activities are **in-person only**
  - Automatically **podcast** to podcast.ucsd.edu
  - Standard UCSD policy on makeup for exams / PI activities
  - Special circumstances: contact me upfront to discuss
- Discussion slot will likely be via **Zoom** only
- Attending Lectures/Discussions live is **optional**
- **Google Forms** for PI Activities
- **Canvas** for submitting AI Engineering Projects, all grade releases, all Exercise and Exam documents
- **Piazza** for all announcements and asynchronous Q&A
  - Do help out your peers by discussing questions

https://cseweb.ucsd.edu/classes/sp26/cse234-a/index.html

Course Administrivia

- **Lectures**: **TuTh 5-6:20pm PT @ WLH 2001**
  - Podcast always available for async. viewing
- **Instructor**: Arun Kumar; akk018 [at] ucsd.edu
  - Office hours: **Thu 3:30-4:30 pm PT** @ 3218 CSE & Zoom
- **Piazza**: https://piazza.com/ucsd/spring2026/csedsc234
- **Teaching Assistants**:
  - Ruobing Han, Manas Jain, Raghav Jain, and Har Simrat Singh
- See course website for TA contact info. and their OHs

https://cseweb.ucsd.edu/classes/sp26/cse234-a/index.html

Triton AI Student Assistant Pilot

Instructor AI Assistant

Accessible only to instructors and their designated collaborators (like TA/IAs), this assistant helps streamline course development workflows. It can be used to:

- **Generate instructional content:** Create quizzes, lecture summaries, and study guides.
- **Repurpose materials:** Instantly transform video transcripts into structured notes or outlines.
- **Brainstorm ideas:** Explore new ways to structure their course or explain complex topics.

Student AI Assistant

Available to all enrolled students in the course, this assistant helps them engage more deeply with instructor-selected materials. It is designed to:

- **Act as a Socratic tutor:** Guide students toward answers and encourage critical thinking rather than simply providing solutions.
- **Offer instant clarification:** Answer student questions based strictly on the approved course content.
- **Increase engagement:** Provide a safe, interactive space for students to explore topics at their own pace.

https://tritonai.ucsd.edu/tritongpt/instruction.html

General Dos and Do NOTs

Do:

- Come to lectures in person as much as possible, ask questions, and participate actively in class discussions and PI Activities
- Review slides and podcast videos by yourself
- Follow all announcements and participate in discussions on Piazza; pitch in to answer peers' questions
- Use "CSE234"/"DSC234" as subject prefix for all emails to me/TAs

Do NOT:

- Record anything in class without explicit permission from me and all other participants
- Harass, intimidate, or intentionally talk over other students
- Violate academic integrity on any graded components