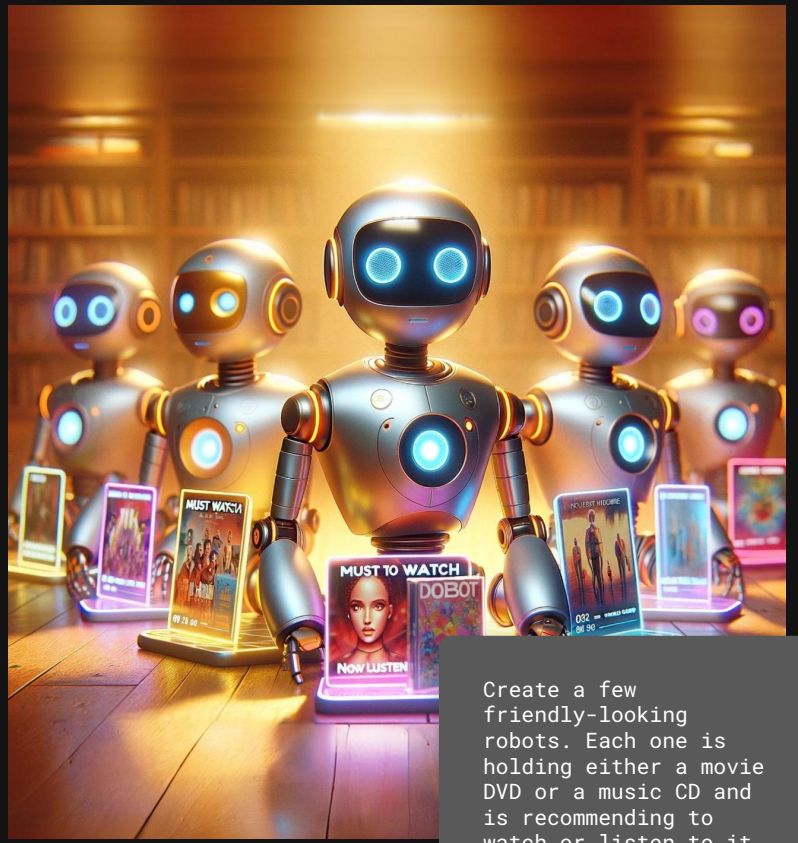
NVIDIA GTC

From Netflix Recommendations to Conversational Multi-agents:

The (R)Evolution of AI-Driven Product Innovation

Xavi Amatriain
VP of Product, Google Cloud



Create a few friendly-looking robots. Each one is holding either a movie DVD or a music CD and is recommending to watch or listen to it.
DALL-E 3

About me

Researcher in Recommender Systems

NETFLIX — Started and led ML Algorithms at Netflix

Quora — VP of Engineering at Quora

Curai Health — Co-founder/CTO at Curai

LinkedIn — VP of AI Product Strategy at LinkedIn

Google Cloud

VP of Product

Core ML/AI at Google

Outline

01

The past:

Data & Algorithms-driven product innovation

02

Yesterday:

ML -> Deep Learning

03

Today:

Product innovation in the Age of (Gen)AI

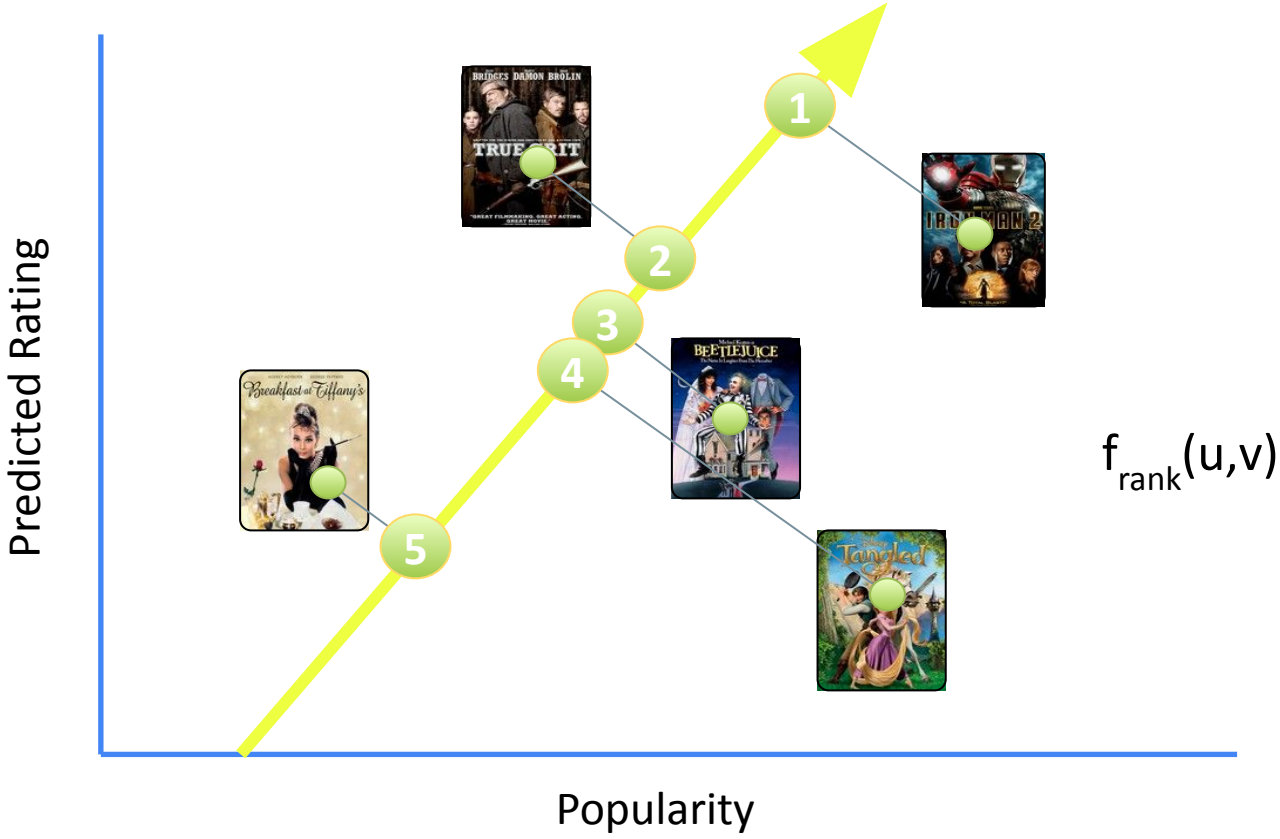
The Past:
Data & Algorithms-driven product innovation

Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)

The Huffington Post | By Dino Grandoni
Posted: 08/21/2013 1:44 pm EDT | Updated: 08/22/2013 8:31 am EDT

Like 55 people like this. Be the first of your friends.

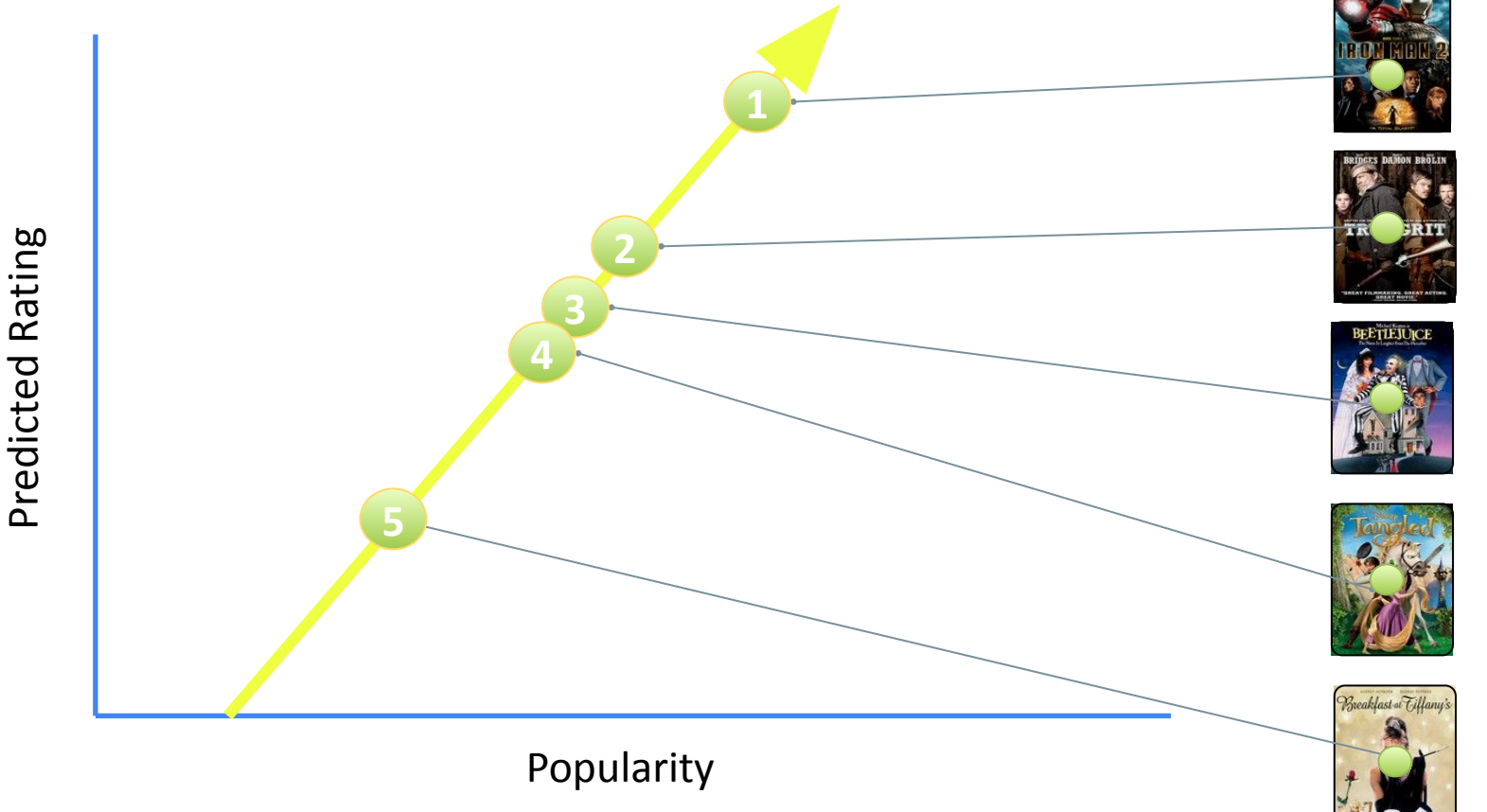
Machine Learning (the “old AI”)



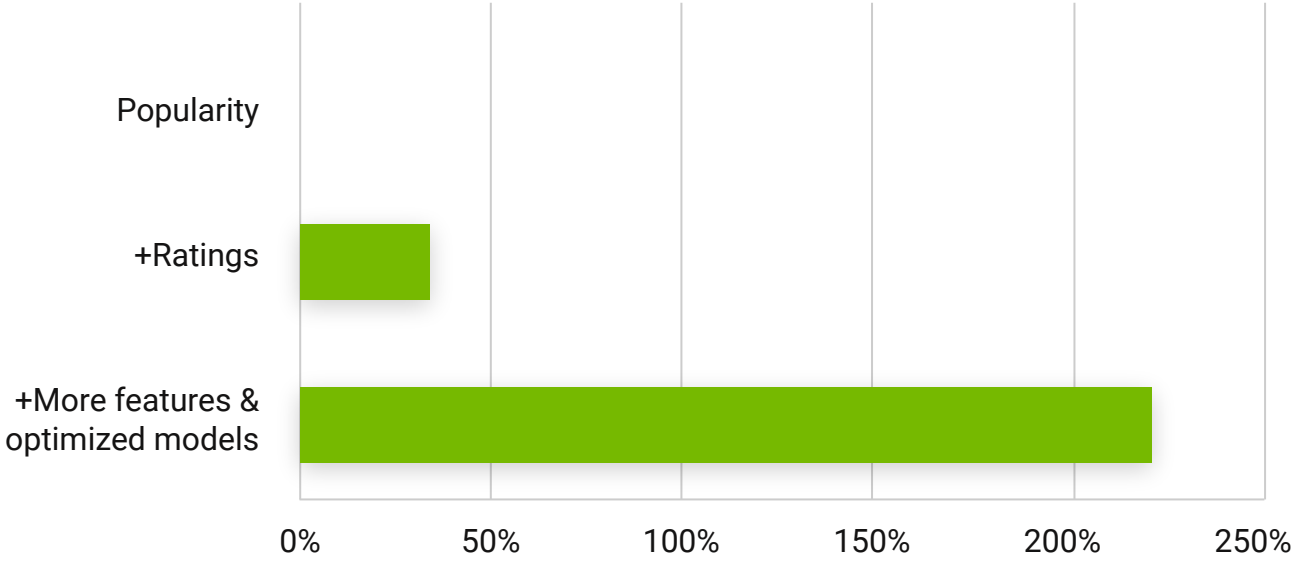
Linear Model:

$$f_{rank}(u,v) = w_1\, p(v) + w_2\, r(u,v) + b$$

Machine Learning (the “old AI”)

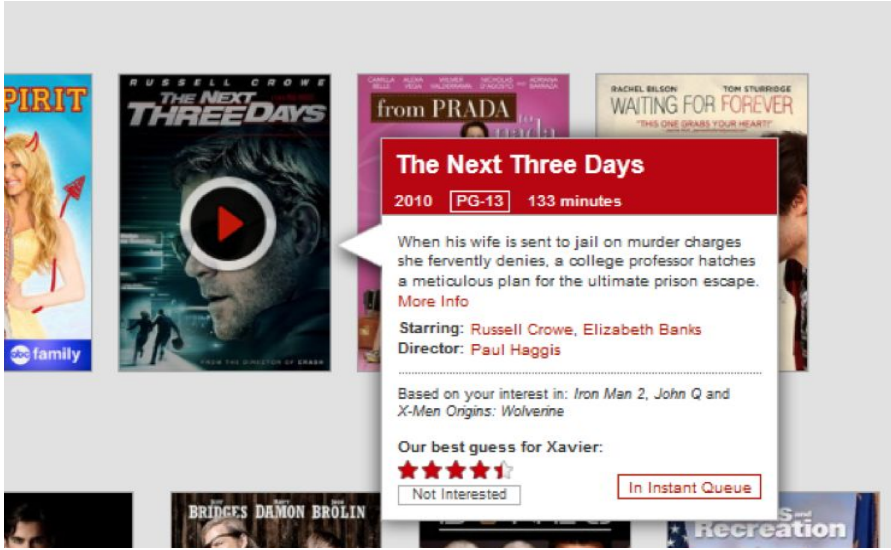


Ranking improvement over baseline



Other important lessons learned

Explanations (and UX in general) matters!



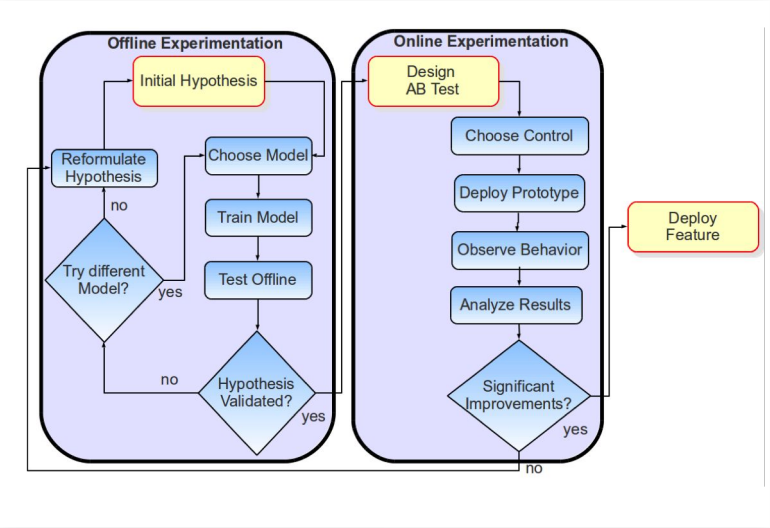
The importance of the experimentation framework

Offline

- Prefer ranking or UX-related metrics
- Measure other aspects such as diversity or novelty
- Keep bank of metrics that you can map to online results with post-hoc analysis

Online

- Measure differences in metrics across statistically identical populations exposed to different algorithms.
- Overall Evaluation Criteria (OEC)
  - Use **long-term metrics** whenever possible
  - **Short-term metrics** can be informative and allow faster decisions. But, not always aligned with OEC



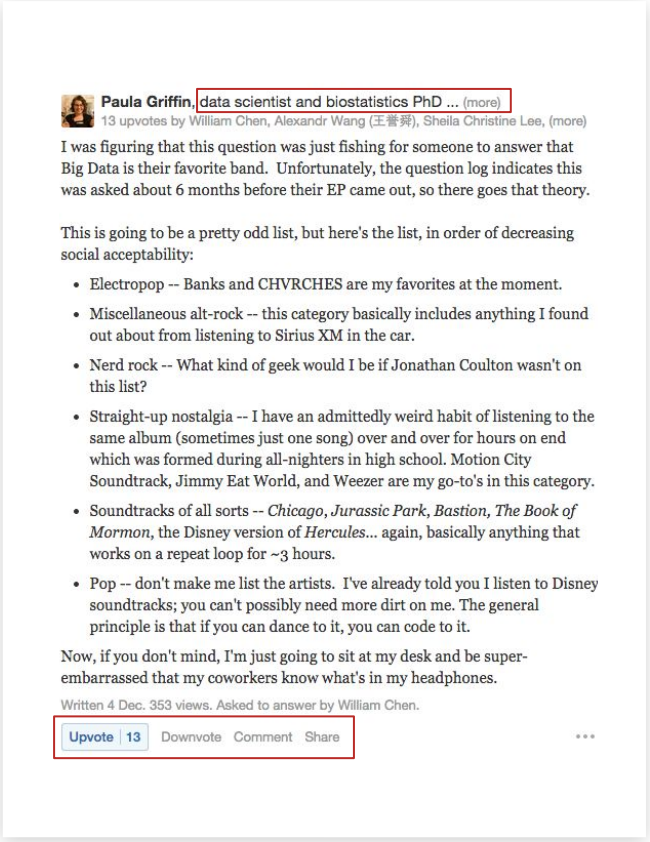
Domain knowledge

What is a good Quora answer?

Truthful	Reusable	Provides explanation	Well formatted	...

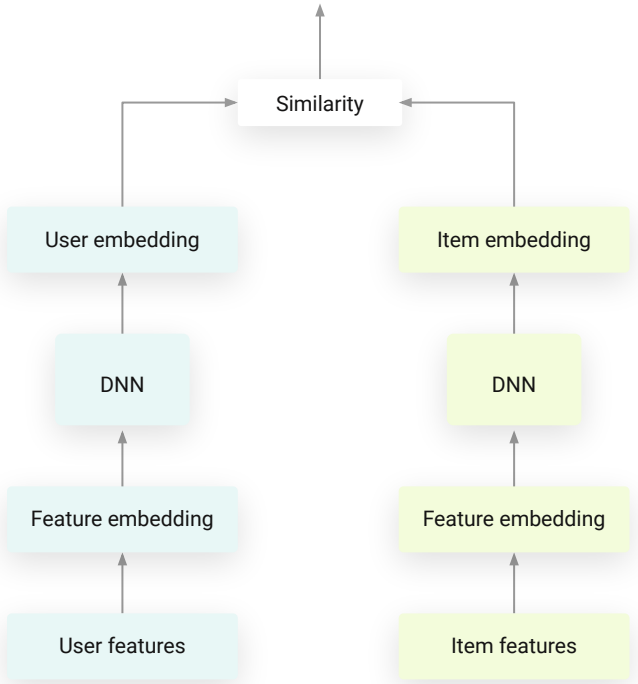
How are those dimensions translated into features?

- Features that relate to the answer
- Quality itself
- Interaction features (upvotes/downvotes, clicks, comments…)
- User features (e.g. expertise in topic)

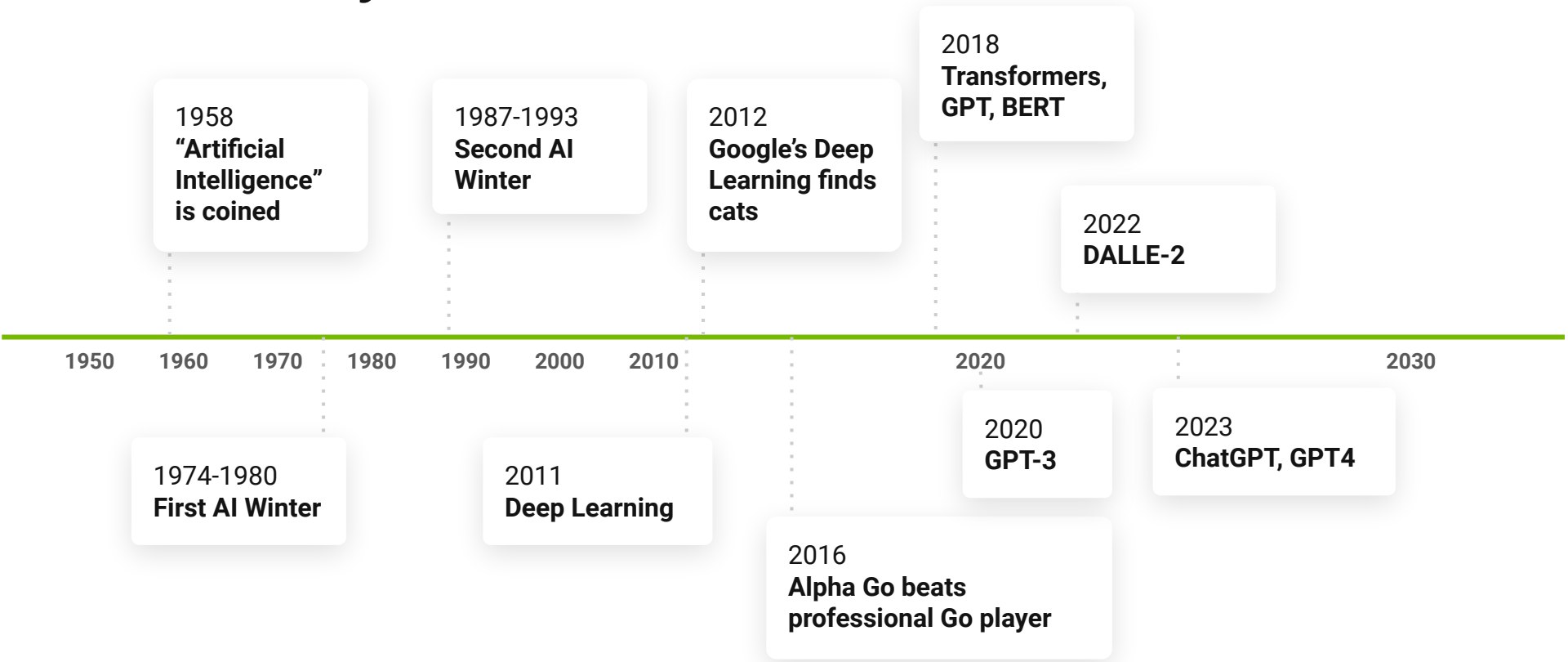


01

Yesterday:
ML -> Deep Learning



The history of AI



Deep learning

Applying deep learning to AirBnB search

Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C. Turnbull, Brendan M. Collins and Thomas Legrand
Airbnb Inc.
malay.haldar@airbnb.com

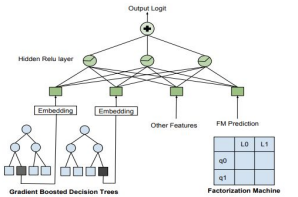
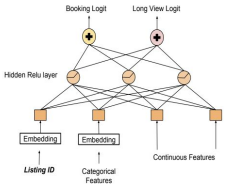


Figure 3: NN with GBDT tree nodes and FM prediction as features



Wide and deep learning for recommender systems

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushare Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichen Hong, Vihan Jain, Xiaobing Liu, Hemal Shah
Google Inc

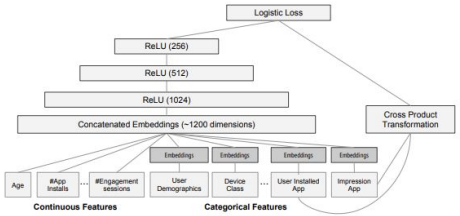
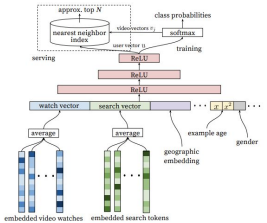


Figure 4: Wide & Deep model structure for apps recommendation.

Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google
Mountain View, CA
{pcovington, jka, msargin}@google.com



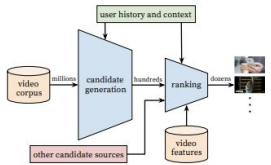


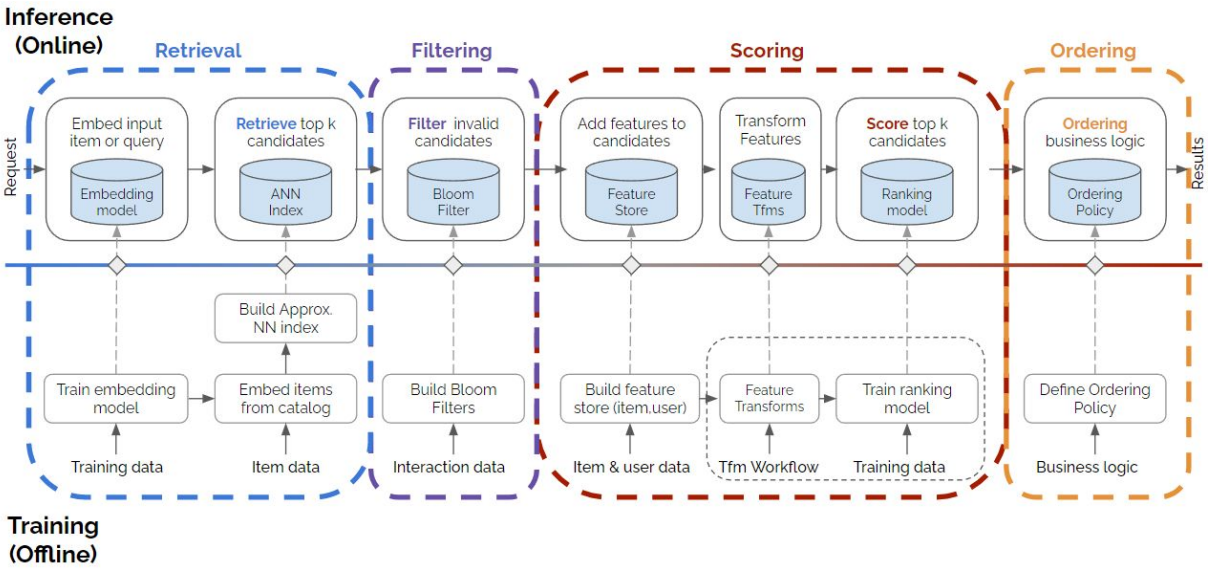
Figure 2: Recommendation system architecture demonstrating the "funnel" where candidate videos are retrieved and ranked before presenting only a few to the user.

Important concepts in modern AI production systems

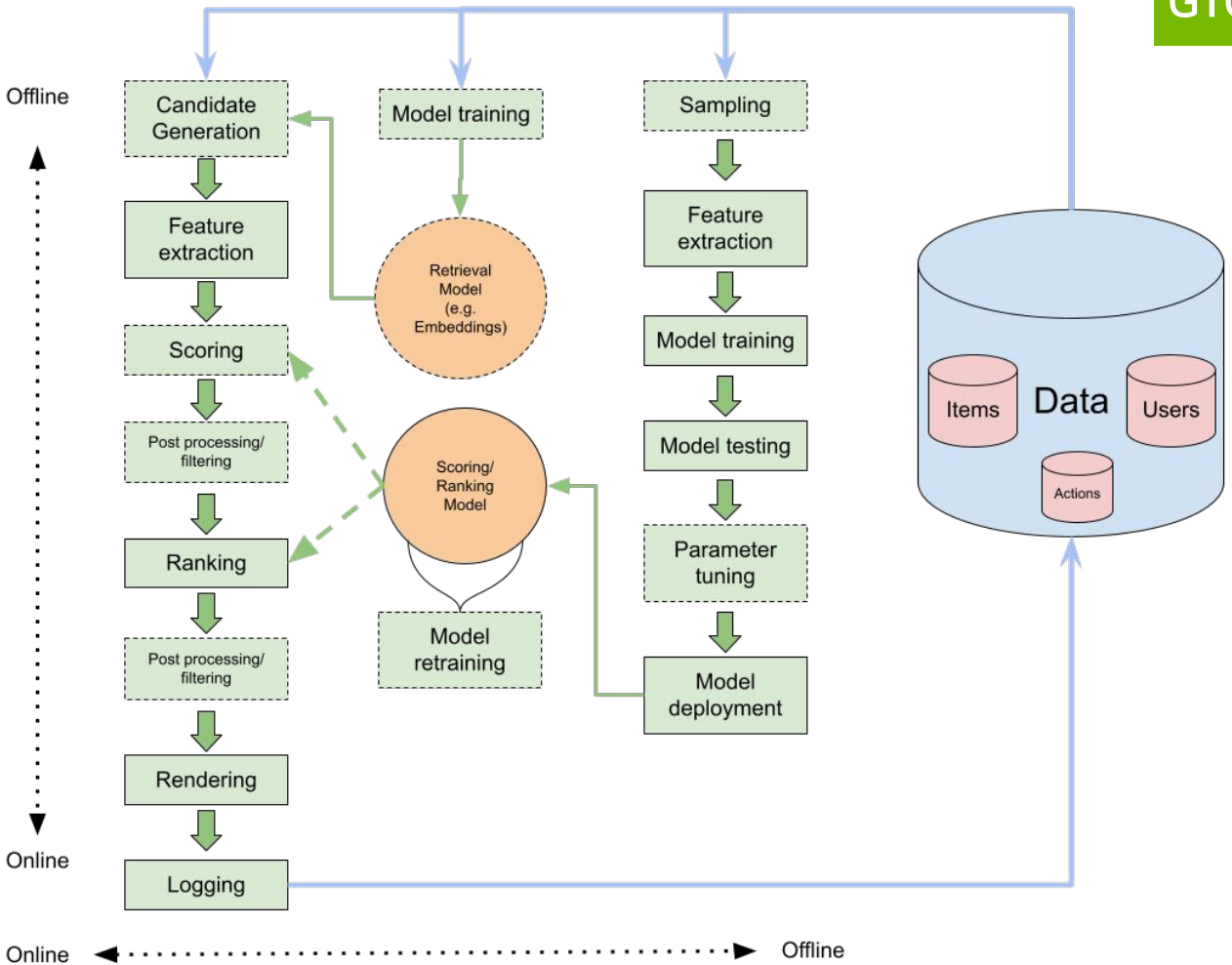
01
Multi-stage systems

02
Embeddings

Multi-stage systems
(NVIDIA version)

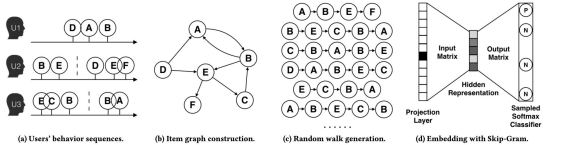


Multi-stage systems
(My version)



Embeddings

Billion-scale commodity embedding for e-commerce recommendation in Alibaba

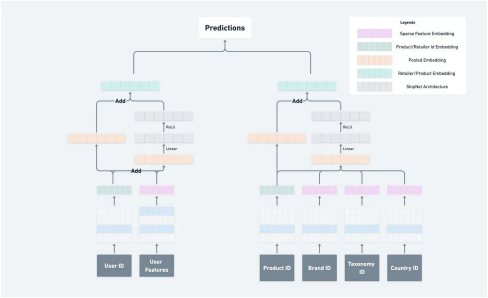


(a) Users' behavior sequences. (b) Item graph construction. (c) Random walk generation. (d) Embedding with Skip-Gram.

Figure 2: Overview of graph embedding in Taobao: (a) Users' behavior sequences: One session for user u1, two sessions for user u2 and u3; these sequences are used to construct the item graph; (b) The weighted directed item graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; (c) The sequences generated by random walk in the item graph; (d) Embedding with Skip-Gram.

Figure 5: Similar items for cold start items. Top 4 similar items are shown. Note that "cat" means category.

How we use engagement-based embeddings to improve search and recommendation on Faire



DECEMBER 1, 2022

Introducing DreamShard: A reinforcement learning approach for embedding table sharding

By: Daochen Zha, Bhargav Bhushanam, **Louis Feng**, Liang Luo, Yusuo Hu, **Yuandong Tian**, Jade Nie, Dima Ivashchenko, Mykola Lukashenko, Yuzhen Huang

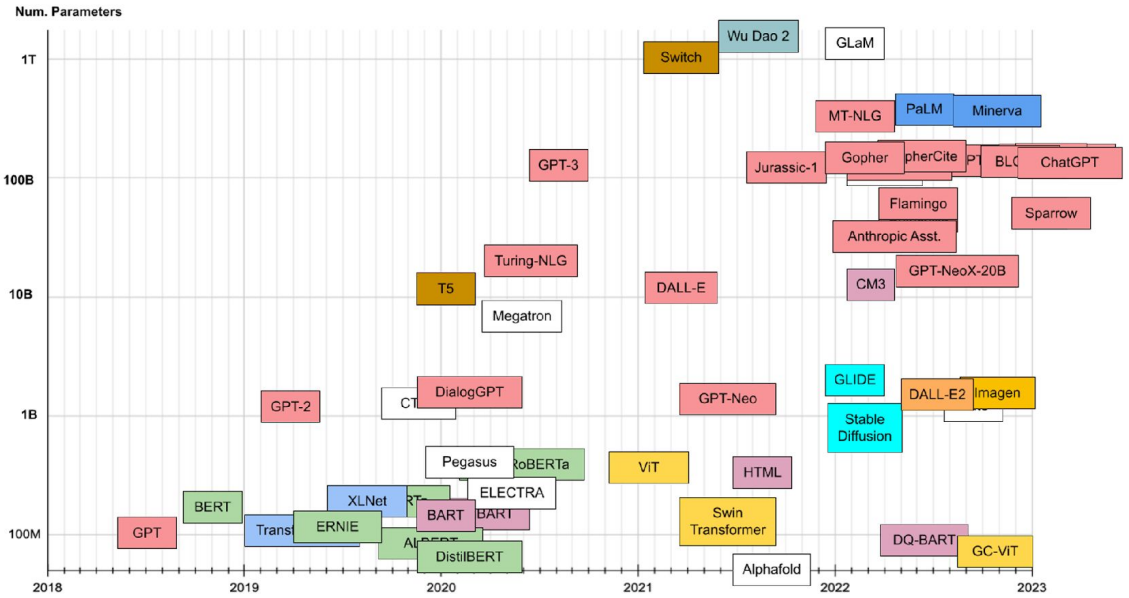
02

Today:
The Age of (Gen)AI

Generate an impressionist style painting of a robot reading a book
Gemini Advanced

The history of AI (part II)

In just the last five years…



GenAI components

GPT = **Generative**

Generate new content based on a natural language input (prompt)

\+ **Pretrained**

Trained on the whole internet and more

Can learn on the fly

Can be fine-tuned

\+ **Transformer**

Just a different kind of Deep Learning architecture using “attention”

Human-level AI

- GPT-4 out of the box:
  - Scores in the 90th percentile for the BAR exam to become a lawyer
  - Scores in the 90th percentile for the SAT both in Math and Reading and Comprehension
  - Exceeds USMLE score required to graduate in medicine in the US by 20 points
  - It is much better at coding than many software engineers
- Gemini Ultra and Claude 3 beat it at many of these tasks

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~82nd	149 ~41st
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~54th
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th–100th	5 86th–100th	5 86th–100th
AP Biology	5 85th–100th	5 85th–100th	4 62nd–85th
AP Calculus BC	4 43rd–59th	4 43rd–59th	1 0th–7th
AP Chemistry	4 71st–88th	4 71st–88th	2 22nd–46th
AP English Language and Composition	2 14th–44th	2 14th–44th	2 14th–44th
AP English Literature and Composition	2 8th–22nd	2 8th–22nd	2 8th–22nd
AP Environmental Science	5 91st–100th	5 91st–100th	5 91st–100th
AP Macroeconomics	5 84th–100th	5 84th–100th	2 33rd–48th
AP Microeconomics	5 82nd–100th	4 60th–82nd	4 60th–82nd
AP Physics 2	4 66th–84th	4 66th–84th	3 30th–66th
AP Psychology	5 83rd–100th	5 83rd–100th	5 83rd–100th
AP Statistics	5 85th–100th	5 85th–100th	3 40th–63rd
AP US Government	5 88th–100th	5 88th–100th	4 77th–88th
AP US History	5 89th–100th	4 74th–89th	4 74th–89th
AP World History	4 65th–87th	4 65th–87th	4 65th–87th
AMC 10	30/150 6th–12th	36/150 10th–19th	36/150 10th–19th
AMC 12	60/150 45th–66th	48/150 19th–40th	30/150 4th–8th
Intro Sommelier (theory knowledge)	92%	92%	80%
Certified Sommelier (theory knowledge)	86%	86%	58%
Advanced Sommelier (theory knowledge)	77%	77%	46%
Leetcode (easy)	31/41	31/41	12/41
Leetcode (medium)	21/80	21/80	8/80
Leetcode (hard)	3/45	3/45	0/45

GenAI is already re-defining most companies' roadmaps

Forbes

FORBES > INNOVATION > ENTERPRISE TECH

Microsoft's Plan To Infuse AI And ChatGPT Into Everything

Bernard Marr Contributor

Follow

0

Mar 6, 2023, 02:22am EST

Listen to article 9 minutes

Microsoft has big plans for artificial intelligence (AI), and it's becoming clear that it believes that ChatGPT – and natural language technology in general - will play a big part.

What if ChatGPT was trained on decades of financial news and data? BloombergGPT aims to be a domain-specific AI for business news

The news and data giant has — with a relatively small team — built a generative AI that it says outperforms the competition on its own specific information needs.

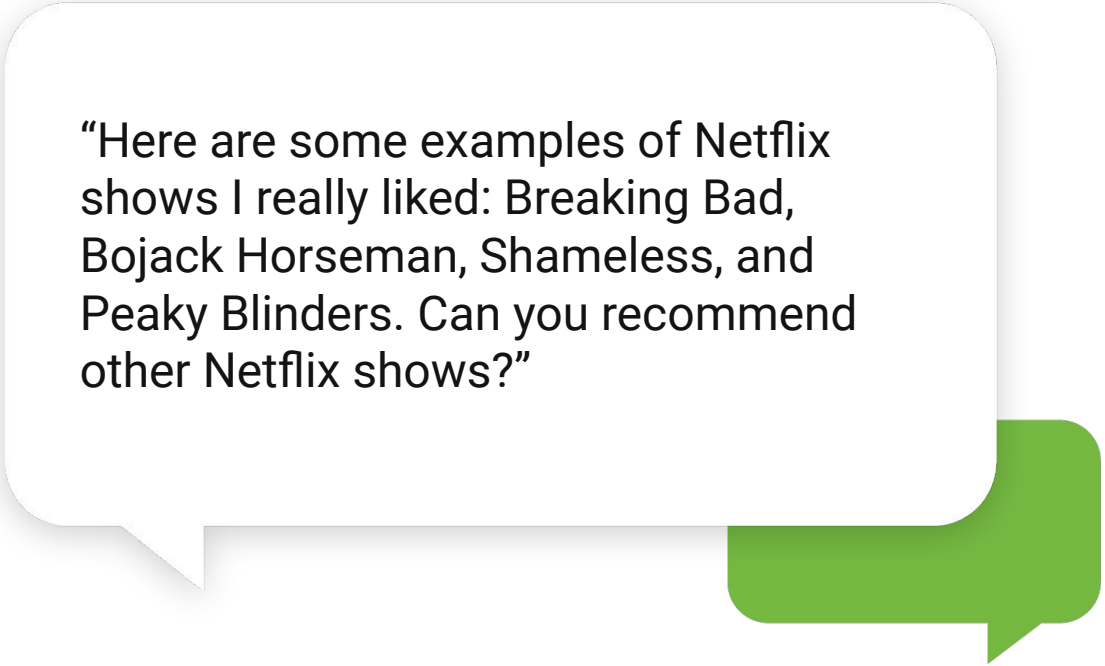
Businessweek
Technology

Google's Plan to Catch ChatGPT Is to Stuff AI Into Everything

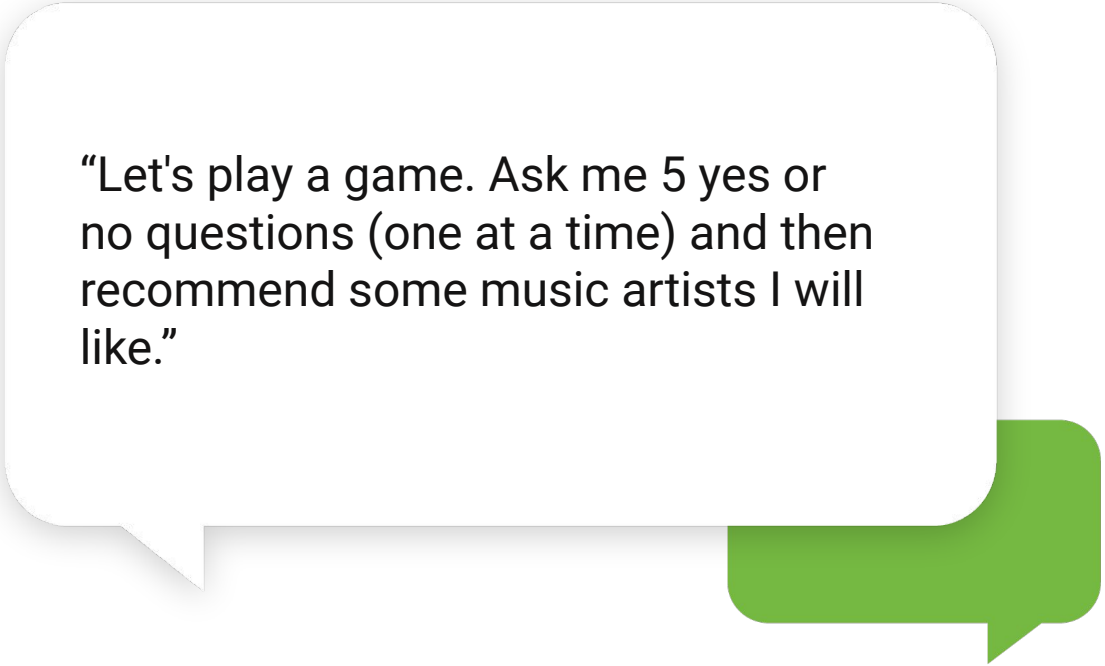
A new internal directive requires "generative artificial intelligence" to be incorporated into all of its biggest products within months.

Let’s try LLMs for recommendations!

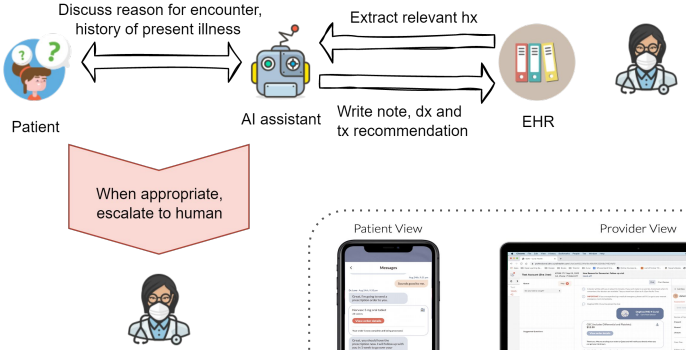
Basic LLM recommendations

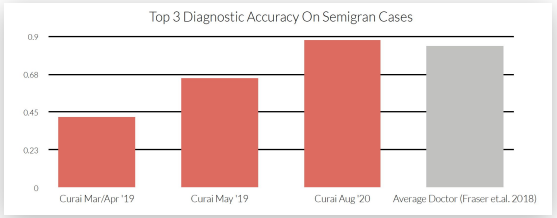


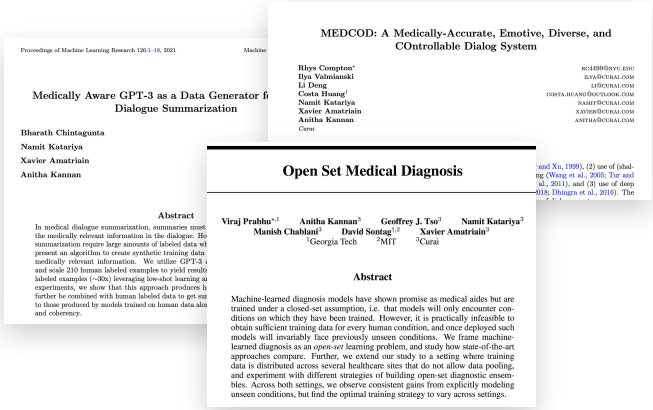
Beyond basic LLM recommendations



Everything will be disrupted by AI. Very quickly

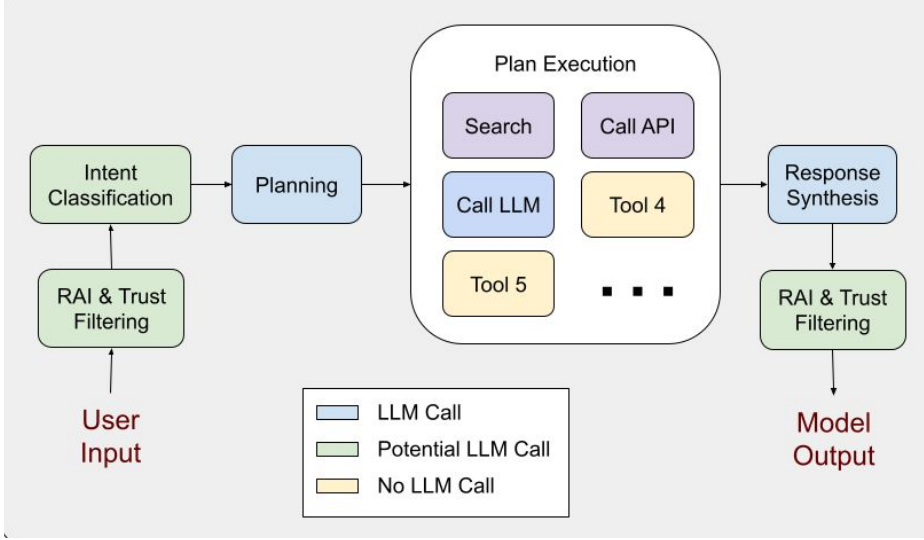




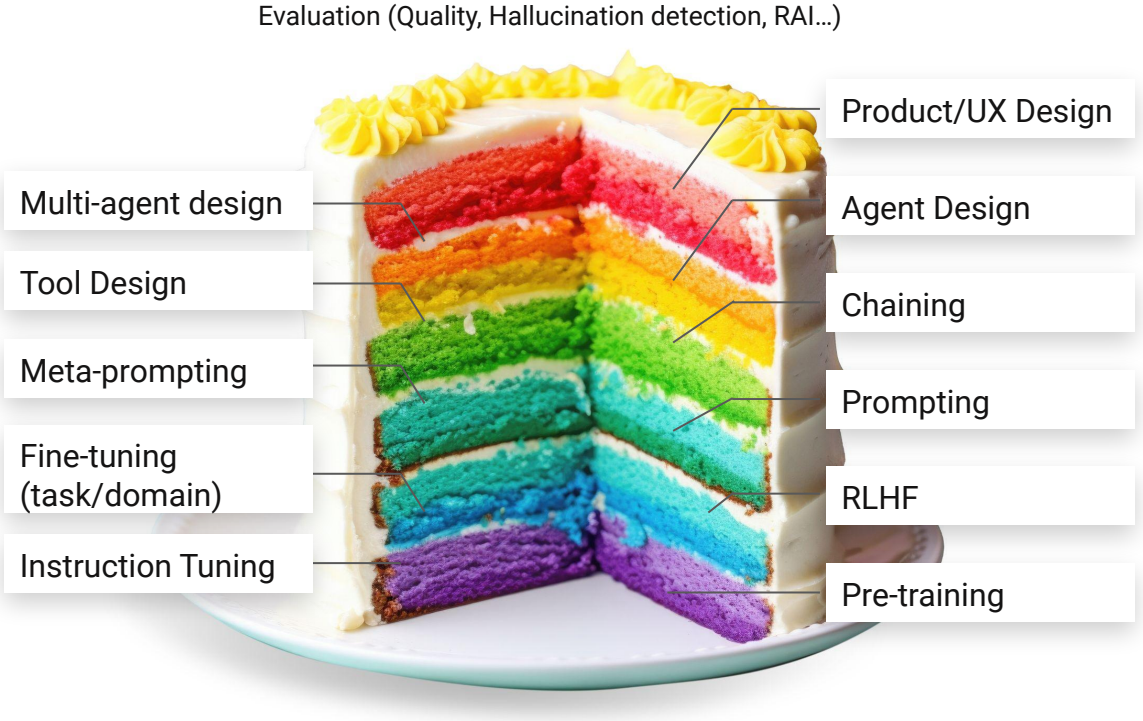


Curai Health

What's behind an LLM application



The multilayer cake of GenAI



How is this similar to the “old AI”

1. Importance of Product and UX design
2. Importance of evaluation and metrics
3. Importance of domain knowledge (e.g. healthcare)

How is this different to the “old AI”

1. Now the UX **is** the AI
2. We need new evaluation metrics and frameworks (e.g. hallucination)
3. Domain knowledge is needed but is also present in the models

Let's focus on Agents and Multi-agents

Agents

Agent = LLM based system that has access to tools and can decide how to use them

- Agents are hard to maintain and implement and tools/libraries like LangChain, Auto-GPT, or AutoGen come to the rescue

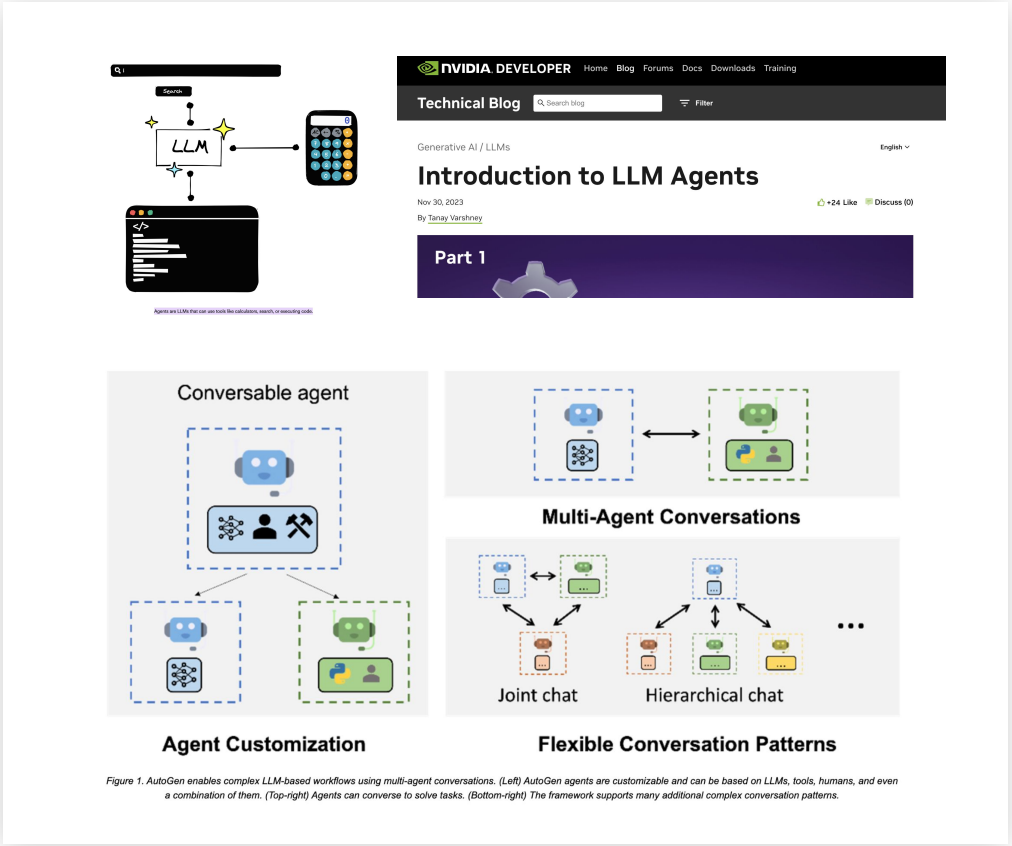
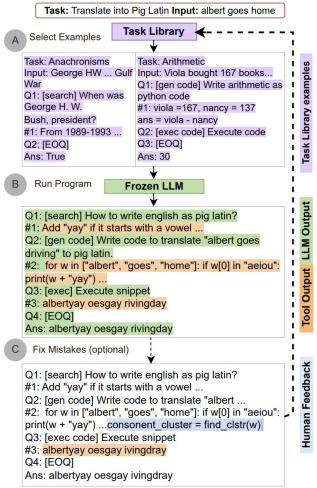


Figure 1. AutoGen enables complex LLM-based workflows using multi-agent conversations. (Left) AutoGen agents are customizable and can be based on LLMs, tools, humans, and even a combination of them. (Top-right) Agents can converse to solve tasks. (Bottom-right) The framework supports many additional complex conversation patterns.

Implementing multi-agent systems

ART
(Automatic multi-step reasoning and tool-use)



REACT
(Reason and Act)

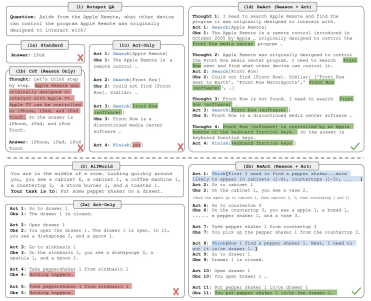


Figure 1: (1) Comparison of 4 prompting methods, (a) Standard, (b) Chain-of-thought (CoT, Reason Only), (c) Act-only, and (d) ReAct (Reason+Act), solving a HotpotQA (Yang et al., 2018) question. (2) Comparison of (a) Act-only and (b) ReAct prompting to solve an AlfWorld (Shridhar et al., 2020b) game. In both domains, we omit in-context examples in the prompt, and only show task solving trajectories generated by the model (Act, Thought) and the environment (Obs).

DERA
(Dialog-Enabled Resolving Agents)

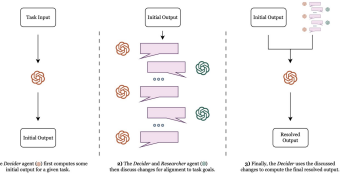
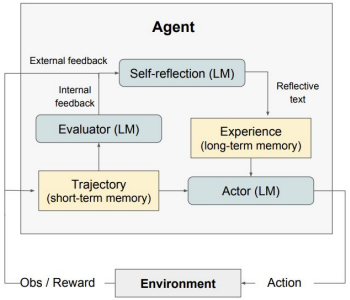


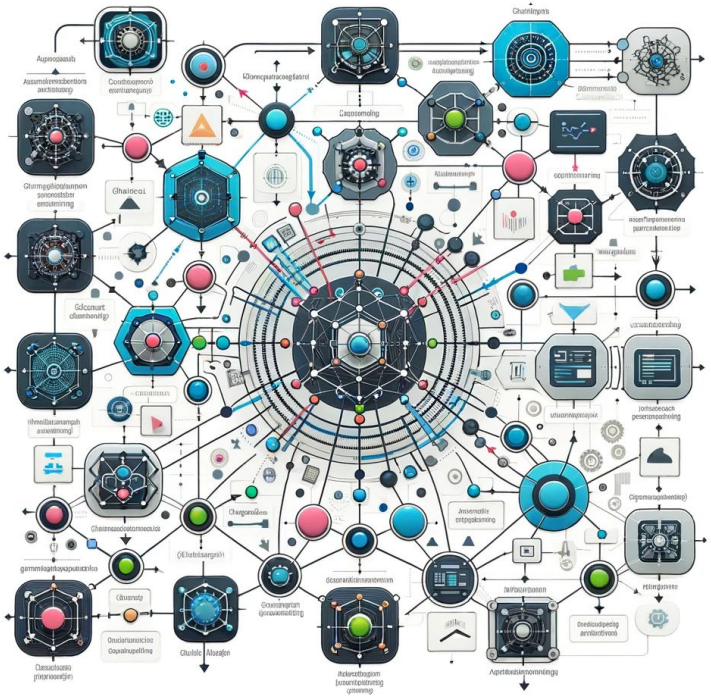
Figure 1: Overview of DERA. The method consists of two agents–a *Researcher* and a *Decider*. The *Decider* generates an initial output for the task (step 1). Then, the *Decider* and *Researcher* work through the problem via conversation (step 2), with the *Researcher* tasked to help identify crucial problem components. The *Decider* has the autonomy to integrate the *Researcher* 's inputs and makes judgments on the final output (step 3). Neither agent has knowledge of the ideal final output.

Reflection



Multi-agent systems and AGI

- AGI is a misnomer commonly used to talk about human-level intelligence
- Human level intelligence and beyond is already being accomplished by combining specialized agents
- Specialized multi-agent systems increase our chances of taking all the upsides of AI while minimizing the risks!



Conclusions

GatesNotes THE BLOG OF BILL GATES

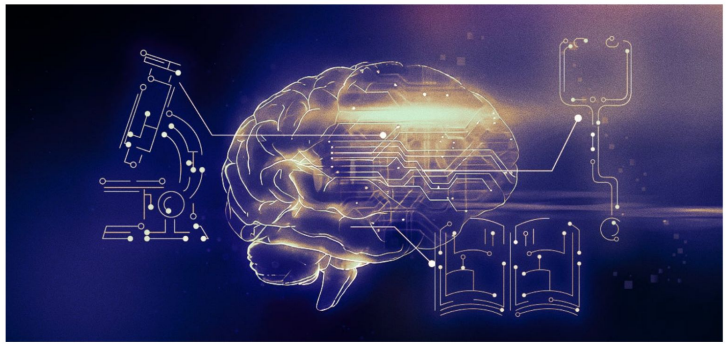
LOG IN

A NEW ERA

The Age of AI has begun

Artificial intelligence is as revolutionary as mobile phones and the Internet.

By **Bill Gates** | March 21, 2023 • 14 minute read



Q&A