AI/ML selected Aspects and Opportunities

to

LF Edge Akraino
Technical Steering Committee (TSC)

Ike Alisson

2023 - 11- 23 Rev PA04

Table of Contents (ToC)

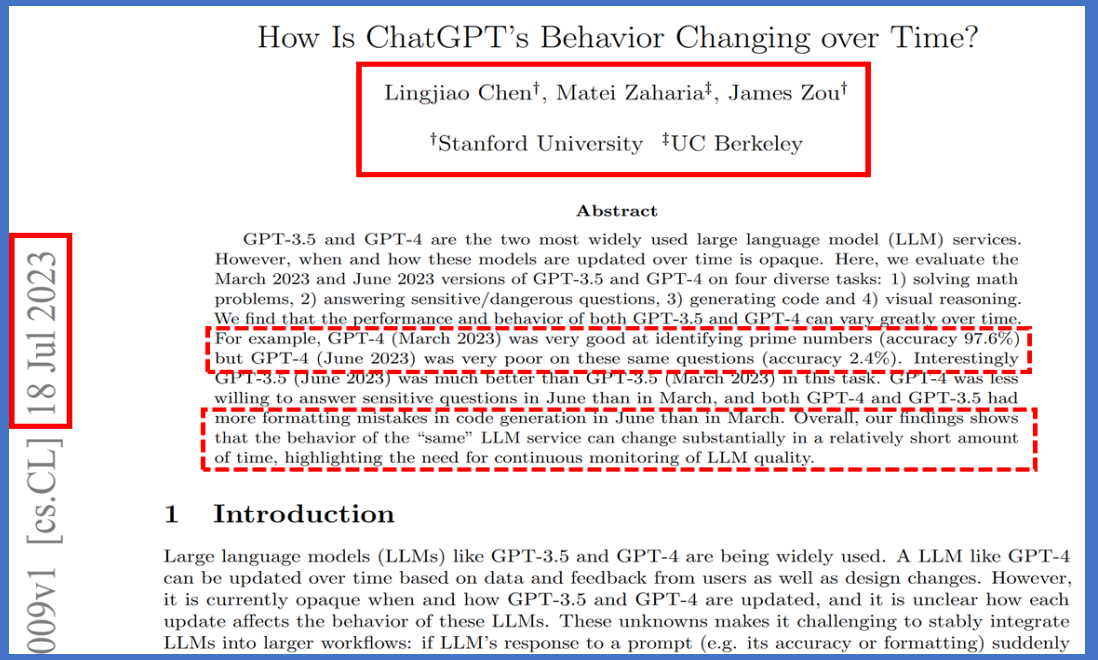
According to a study, Researchers at Stanford & UC Berkeley, the performance of OpenAI's LLMs has decreased significantly over time.

The researchers found that the *Performance and Behavior of GPT-3.5 and GPT-4 varied across their respective releases in March and June* extended it with August and October issues.

The Researchers wanted to determine if these LLMs were improving, as *they can be updated based :*
- Data,
- User Feedback, and
- Design Changes.

The team evaluated the behavior of the March 2023 & June 2023 versions of GPT-3.5 and GPT-4 on four (4) Tasks.

The 1st was Solving Math Problems,
The 2nd was Answering Sensitive/Dangerous Questions,
The 3rd was Generating Code, and the
The 4th was assessing the Models on Visual Reasoning.



How Is ChatGPT's Behavior Changing over Time?

Lingjiao Chen†, Matei Zaharia‡, James Zou†

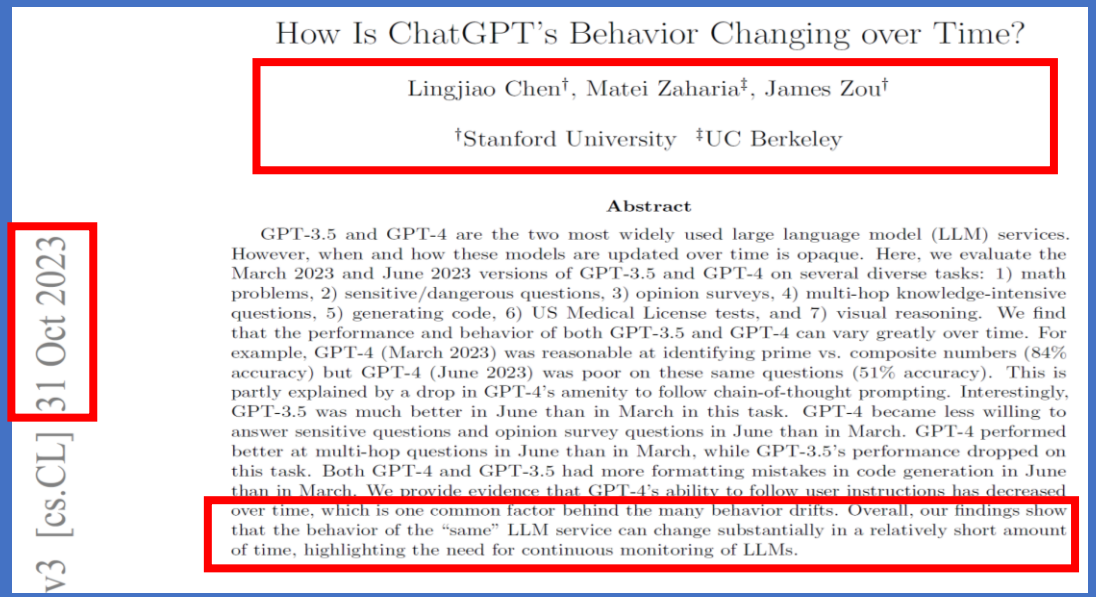
†Stanford University ‡UC Berkeley

Abstract

GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on four diverse tasks: 1) solving math problems, 2) answering sensitive/dangerous questions, 3) generating code and 4) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was very good at identifying prime numbers (accuracy 97.6%) but GPT-4 (June 2023) was very poor on these same questions (accuracy 2.4%). Interestingly GPT-3.5 (June 2023) was much better than GPT-3.5 (March 2023) in this task. GPT-4 was less willing to answer sensitive questions in June than in March, and both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. Overall, our findings shows that the behavior of the "same" LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLM quality.

1 Introduction

Large language models (LLMs) like GPT-3.5 and GPT-4 are being widely used. A LLM like GPT-4 can be updated over time based on data and feedback from users as well as design changes. However, it is currently opaque when and how GPT-3.5 and GPT-4 are updated, and it is unclear how each update affects the behavior of these LLMs. These unknowns makes it challenging to stably integrate LLMs into larger workflows: if LLM's response to a prompt (e.g. its accuracy or formatting) suddenly



How Is ChatGPT's Behavior Changing over Time?

Lingjiao Chen†, Matei Zaharia‡, James Zou†

†Stanford University ‡UC Berkeley

Abstract

GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks: 1) math problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4's amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5's performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. We provide evidence that GPT-4's ability to follow user instructions has decreased over time, which is one common factor behind the many behavior drifts. Overall, our findings show that the behavior of the "same" LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.

1. ChatGPT's Performance and Accuracy has decreased over time (March - June 2023 ChatGPT 3 & 3,5 vs ChatGPT-4

According to a study, Researchers at Stanford & UC Berkeley, the performance of OpenAI's LLMs has decreased significantly over time.

When **introducing GPT-4 in May this year (2023)**, OpenAI's report claimed that GPT-4 is much more reliable and creative and can handle more nuanced instructions than GPT-3.5.

More recently, GPT-4 was shown to successfully pass difficult exams in professional domains such as medicine and law.

GPT-4, in its March 2023 version, could identify *Prime Numbers with ar Accuracy of 97.6%,* but the Team found in its June 2023 *Version performed very poorly* on these same Questions with *2.4% accuracy.*

The team also found that GPT-4 was less willing to answer sensitive questions in June than in March, and both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March.

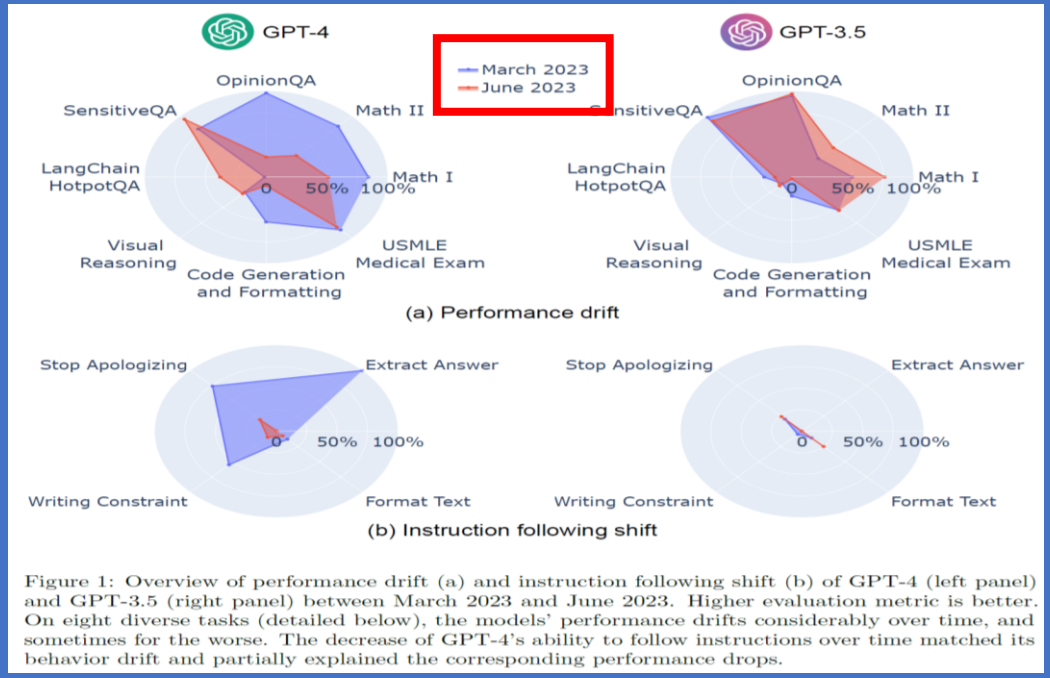


Figure 1: Overview of performance drift (a) and instruction following shift (b) of GPT-4 (left panel) and GPT-3.5 (right panel) between March 2023 and June 2023. Higher evaluation metric is better. On eight diverse tasks (detailed below), the models' performance drifts considerably over time, and sometimes for the worse. The decrease of GPT-4's ability to follow instructions over time matched its behavior drift and partially explained the corresponding performance drops.

Overall, Our Findings show that the behavior of the "same" LLM Service can change substantially in a relatively short amount of time, highlighting the need for Continuous Monitoring of LLMs.

Create Custom GPTs by Crawling the Web

What's New
GPT Crawler is a new open-source project that creates custom versions of ChatGPT - called GPTs, by crawling web content. It allows users to generate personalized chatbots focused on particular topics or sites.

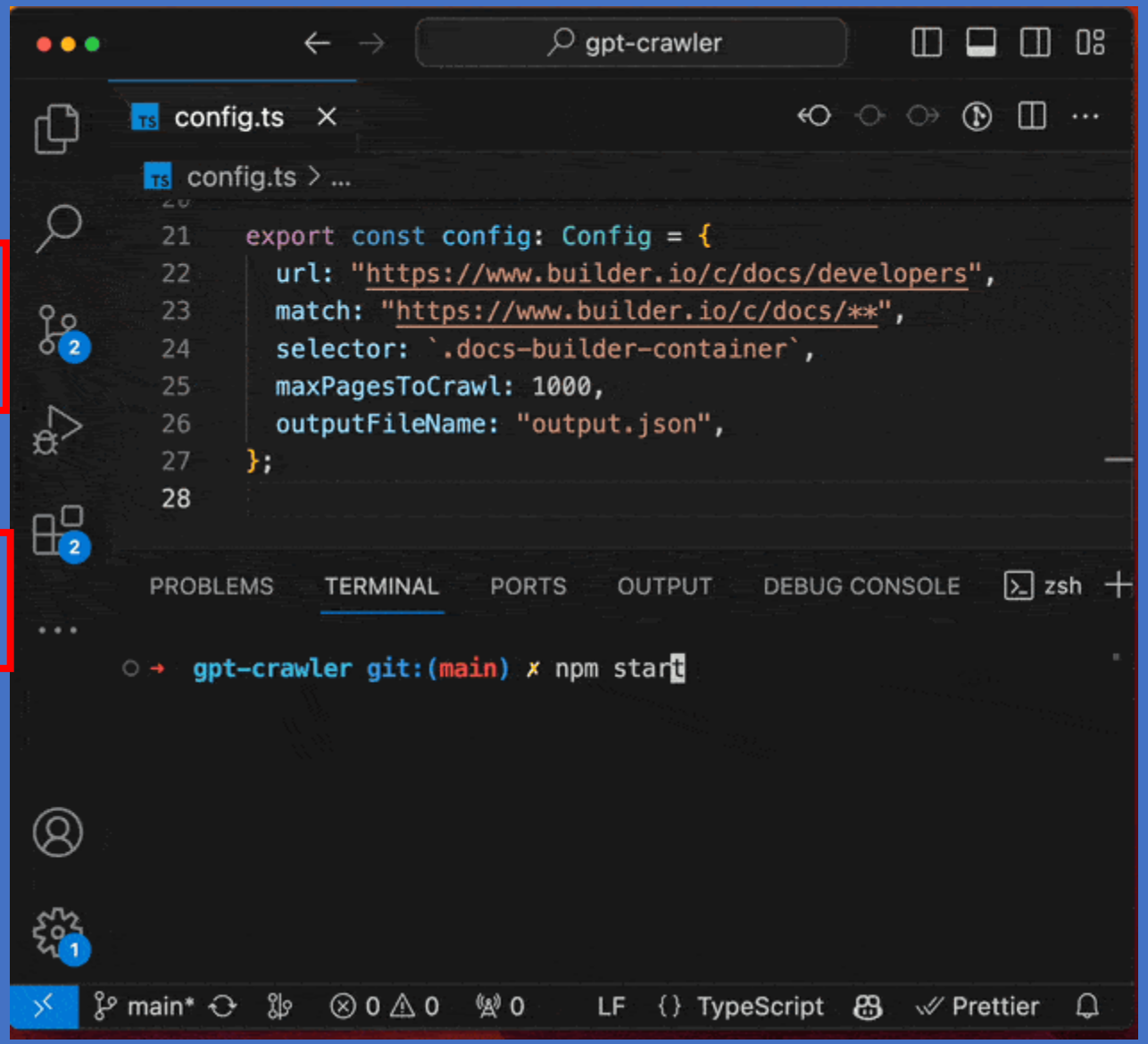
The user can provide a URL, and the tool will generate a Knowledge Base to build a specialized GPT, usable in websites and apps.

How it Works
GPT Crawler drastically lowers the barrier to entry for creating a personalized, up-to-date chatbot tailored to specific web resources. API accessibility means that this customized assistant can be easily embedded into products.

Features
Customization: Easily create GPTs focused on content from specific websites.

Simple Process: Involves cloning the repository, configuring the crawler, and processing web pages.
Flexibility: Handles client-side content and private site information.

Integration Options: Custom GPTs can be uploaded for direct use or integrated into products via the OpenAI API.



1. *This discussion took place on July 26, 2023, at Cemex Auditorium, Stanford University, and was hosted by the Stanford Graduate School of Business.*

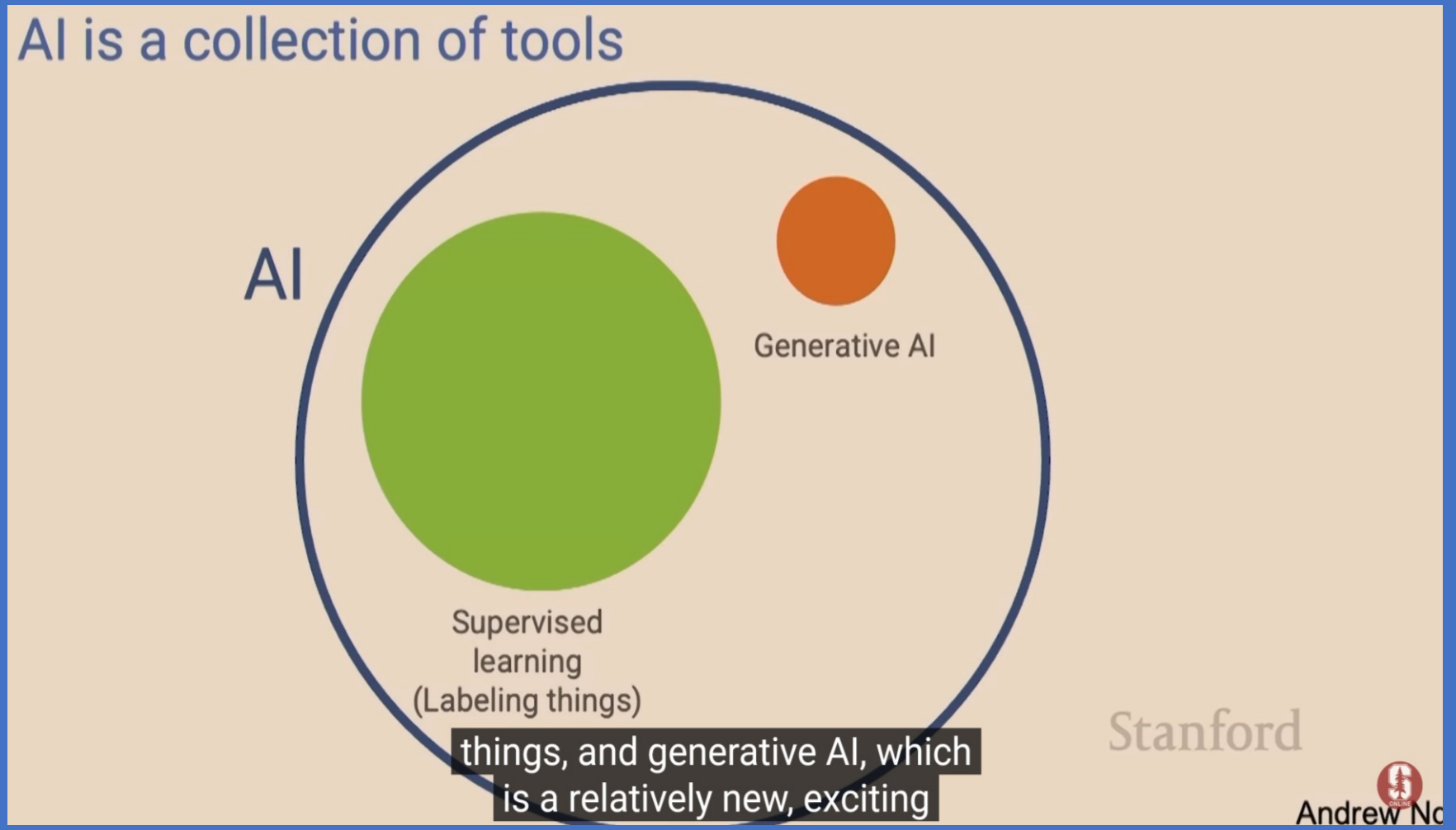
*https://www.youtube.com/watch?v=5p248yoa3oE

AI/ML is a GPT- General Purpose Technology

AI*/ML is a GPT- General Purpose Technology

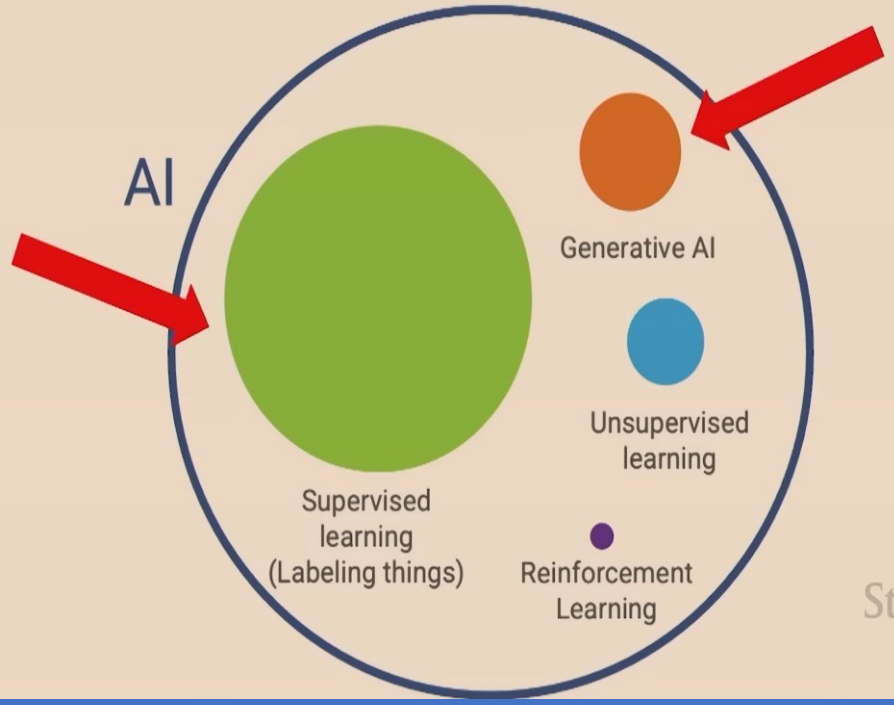
The *Dartmouth Summer Research Project on Artificial Intelligence* (AI) was a *1956 summer Workshop* widely considered to be the *founding event of Artificial Intelligence (AI) as a field*. The Project lasted approximately 6 - 8 weeks and was essentially an extended brainstorming session. In 1955, *John McCarthy*, Assistant Professor of Mathematics at Dartmouth College, decided to organize *a group to clarify and develop ideas about "Thinking Machines". He picked the name 'Artificial Intelligence' (AI) for the new field.* He chose the name partly for its neutrality; avoiding a focus on narrow automata theory, and avoiding cybernetics which was heavily focused on analog feedback. On September 2, 1955, the Project was formally proposed by *John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon*, a founder of Information Theory then at Bell Labs.
The proposal is credited with introducing the term 'Artificial Intelligence' (AI).

Technology landscape



Technology landscape

AI is a collection of tools



Stanford

Andrew Ng

Supervised Learning (labeling things)

Input → Output

A B

Stanford

Andrew Ng

Supervised Learning (labeling things)

Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Ad, user info	Click? (0/1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
Ship route	Fuel consumed	Fuel optimization
Image of phone	Defect? (0/1)	Visual inspection
Restaurant reviews	Sentiment (pos/neg)	Reputation monitoring

Stanford

Andrew Ng

Technology landscape

Supervised Learning (labeling things)

Input → Output

A → B

Stanford

Andrew Ng

Example: Restaurant reviews sentiment tracking

Input (A)	Output (B)
The pastrami sandwich was great! The meat was tender and perfectly balanced by the sauerkraut.	Positive
Service was slow and the food was so-so.	Negative
My favorite chicken curry. Yum!	Positive

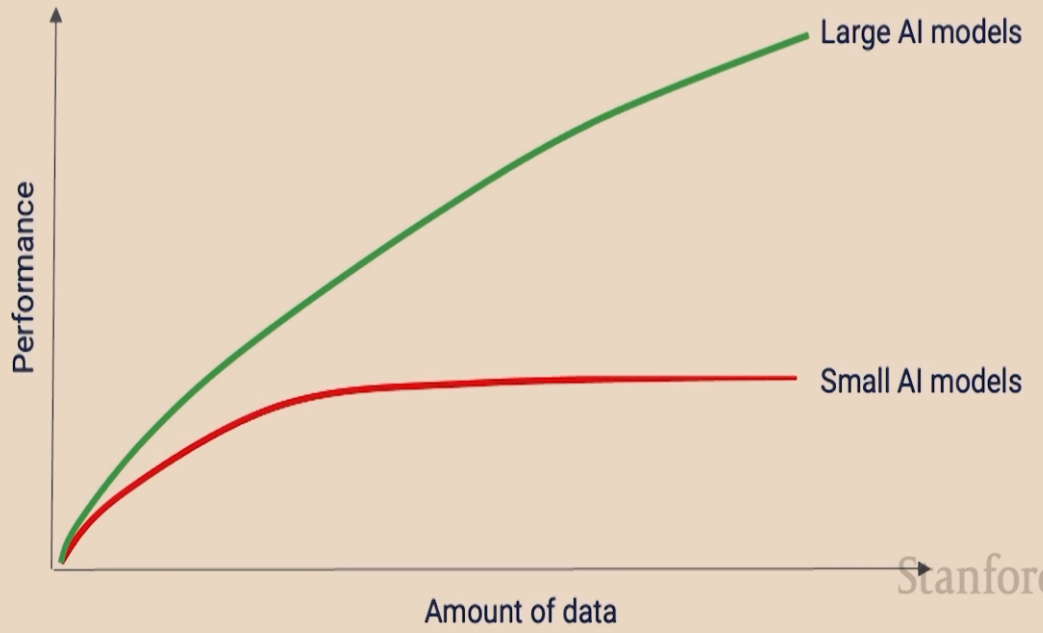
Get labeled data → Train AI model on data → Deploy (run) model

Stanford

Best bubble tea I've ever had! → Positive

Andrew Ng

2010-2020: Large scale supervised learning



Stanford

Andrew Ng

This decade: Generative AI

Text generation process

I love eating _____

prompt

bagels with cream cheese

my mother's meatloaf

out with friends

AI output

How it works

Generative AI is built by using supervised learning (A→B) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese and lox.

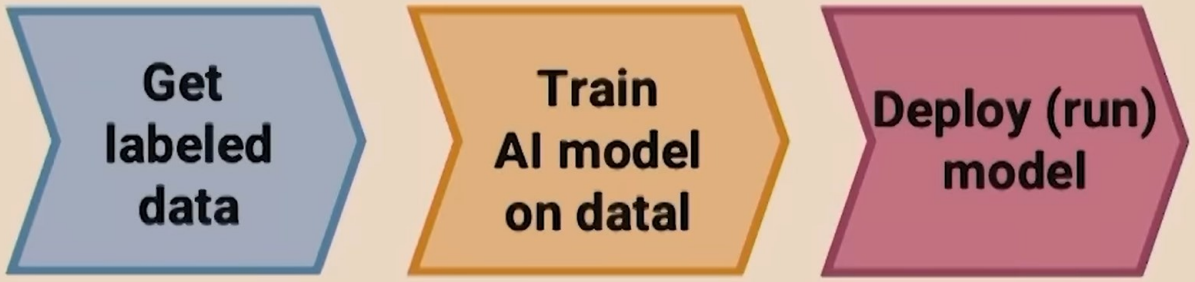
Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream

When we train a very large AI system on a lot of data (hundreds of billions of words) we get a Large Language Model like ChatGPT.

Andrew Ng

Prompting is revolutionizing AI application development

Supervised learning



Prompting is revolutionizing AI application development

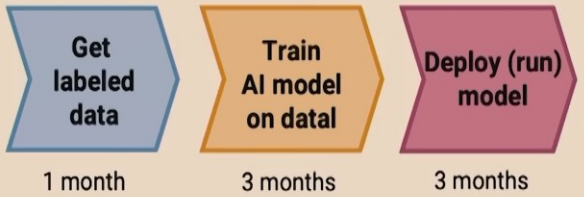
Supervised learning

So pretty realistic timeline for building a commercial grade

Andrew Ng

Prompting is revolutionizing AI application development

Supervised learning



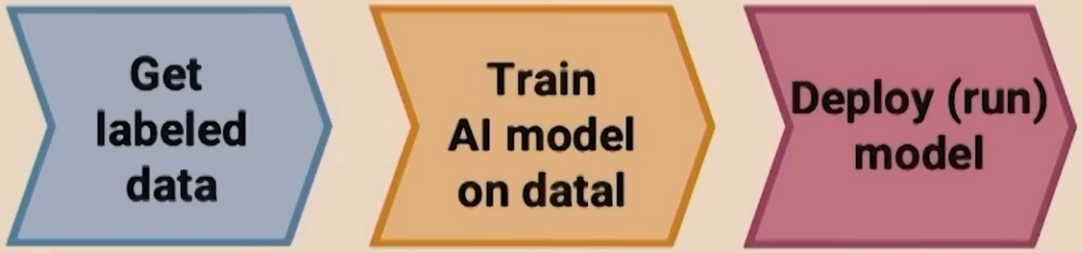
machine learning system is like 6 to 12 months.

Stanford

Andrew Ng

Prompting is revolutionizing AI application development

Supervised learning



Prompt-based AI

Andrew Ng

Run Code

In [1]:
```
import openai
import os
```

In [2]:
```
openai.api_key = os.getenv("OPENAI_API_KEY")

def get_response_to_prompt(prompt):
    response = openai.ChatCompletion.create(model="gpt-3.5-turbo", messages=[{"role":"user","content":prompt}], tempers
    return response.choices[0].message["content"]
```

In [3]:
```
prompt = """Classify the text below, delimited by three dashes (-), as having either a positive or negative sentiment.

---
I had a fantastic time at Stanford GSB! Learned a lot and also made great new friends!
---
"""
```

In [*]:
```
response = get_response_to_prompt(prompt)
print(response)
```

```
positive sentiment
```

In [ ]:

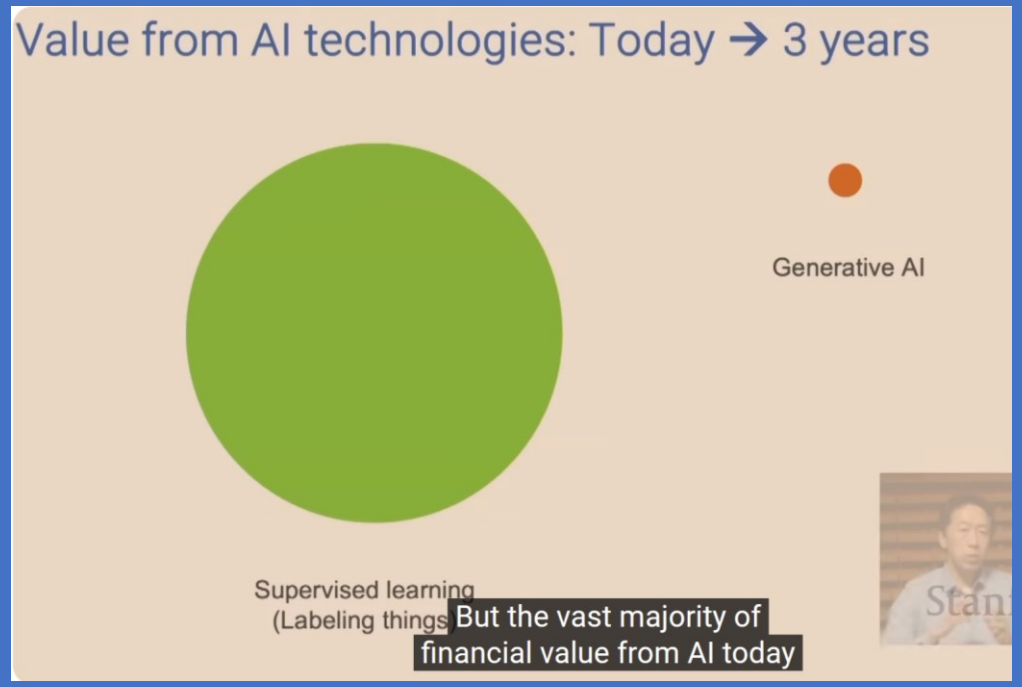
Value from AI technologies: Today → 3 years

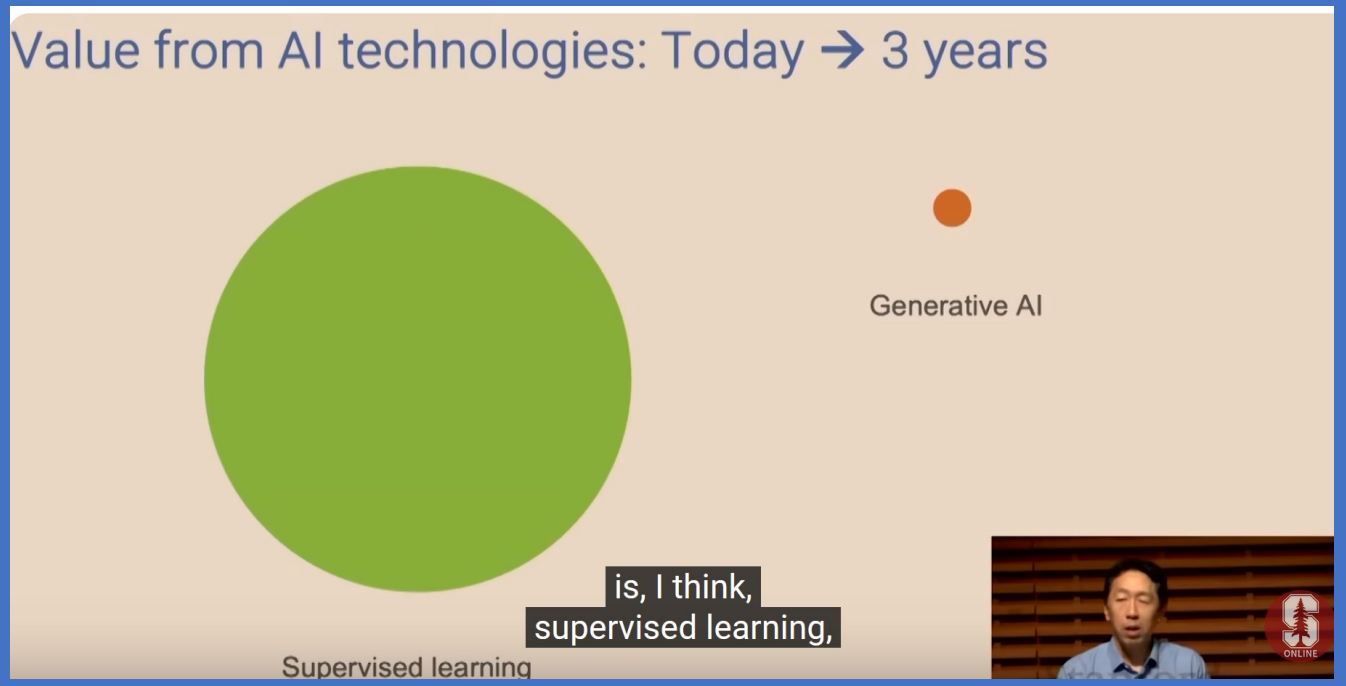
Supervised learning
(Labeling things)

Generative AI

Stanford

Andrew Ng


Value from AI technologies: Today → 3 years
Generative AI
Supervised learning
(Labeling things
But the vast majority of
financial value from AI today


Value from AI technologies: Today → 3 years
Generative AI
is, I think,
supervised learning,
Supervised learning

Value from AI technologies: Today → 3 years

Generative AI

Supervised learning
(Labeling things)

where for a single
company like Google

Andrew Ng

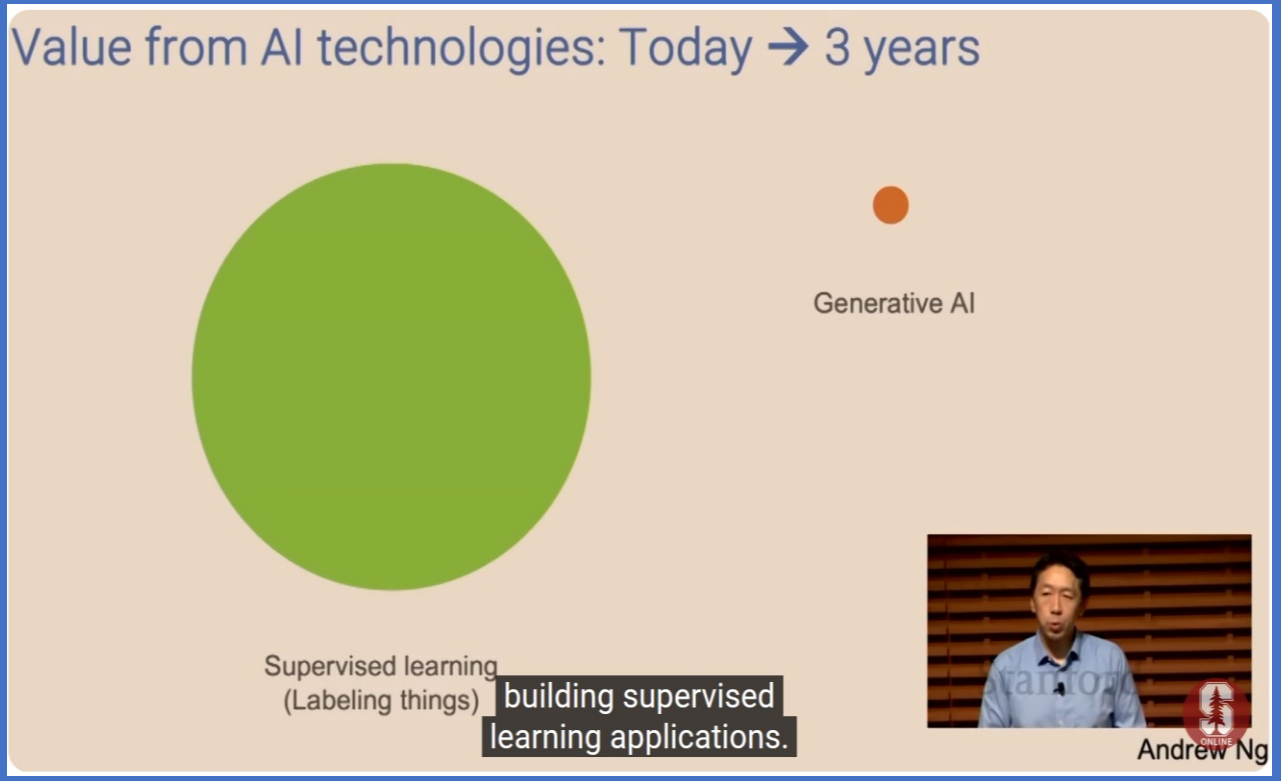
Value from AI technologies: Today → 3 years

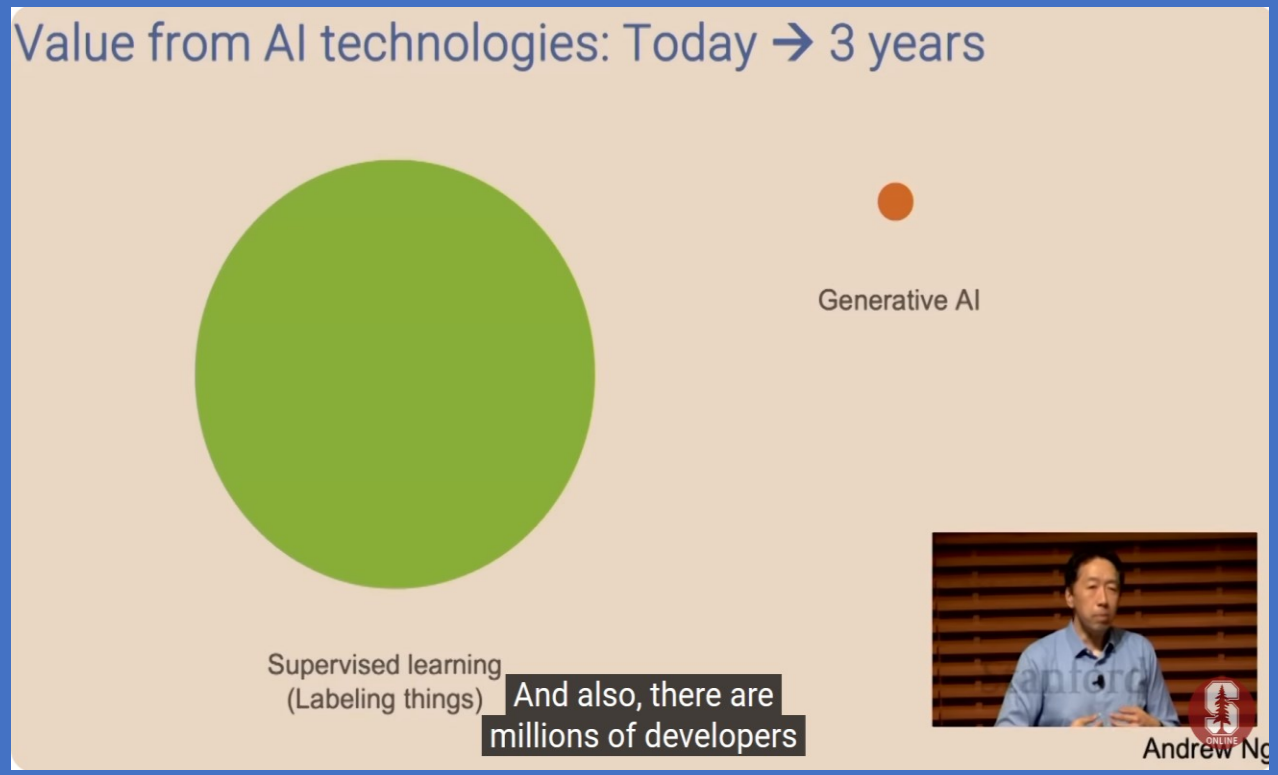
Generative AI

Supervised learning
(Labeling things)

can be worth more than
$100 billion US a year.

Andrew Ng


Value from AI technologies: Today → 3 years
Generative AI
Supervised learning
(Labeling things)
building supervised learning applications.
Andrew Ng


Value from AI technologies: Today → 3 years
Generative AI
Supervised learning
(Labeling things)
And also, there are millions of developers
Andrew Ng

Value from AI technologies: Today → 3 years

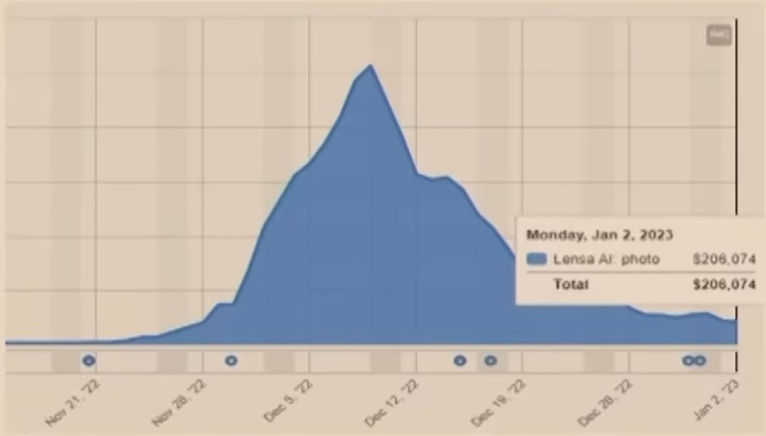
Stanford

Andrew Ng

AI technologies are general purpose technologies

- General purpose technologies are useful for many tasks.
  - Massive value remains to be created using supervised learning (labeling things).
  - Generative AI is another major tool, creating even more opportunities.

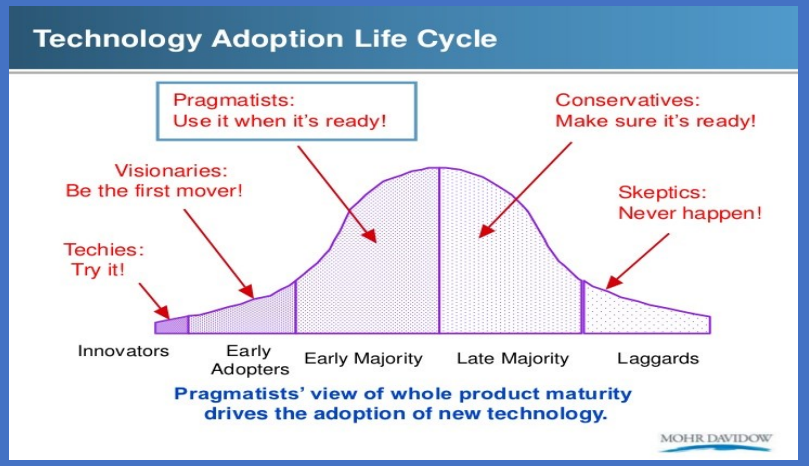
- There will be fads along the way. Lensa revenue:

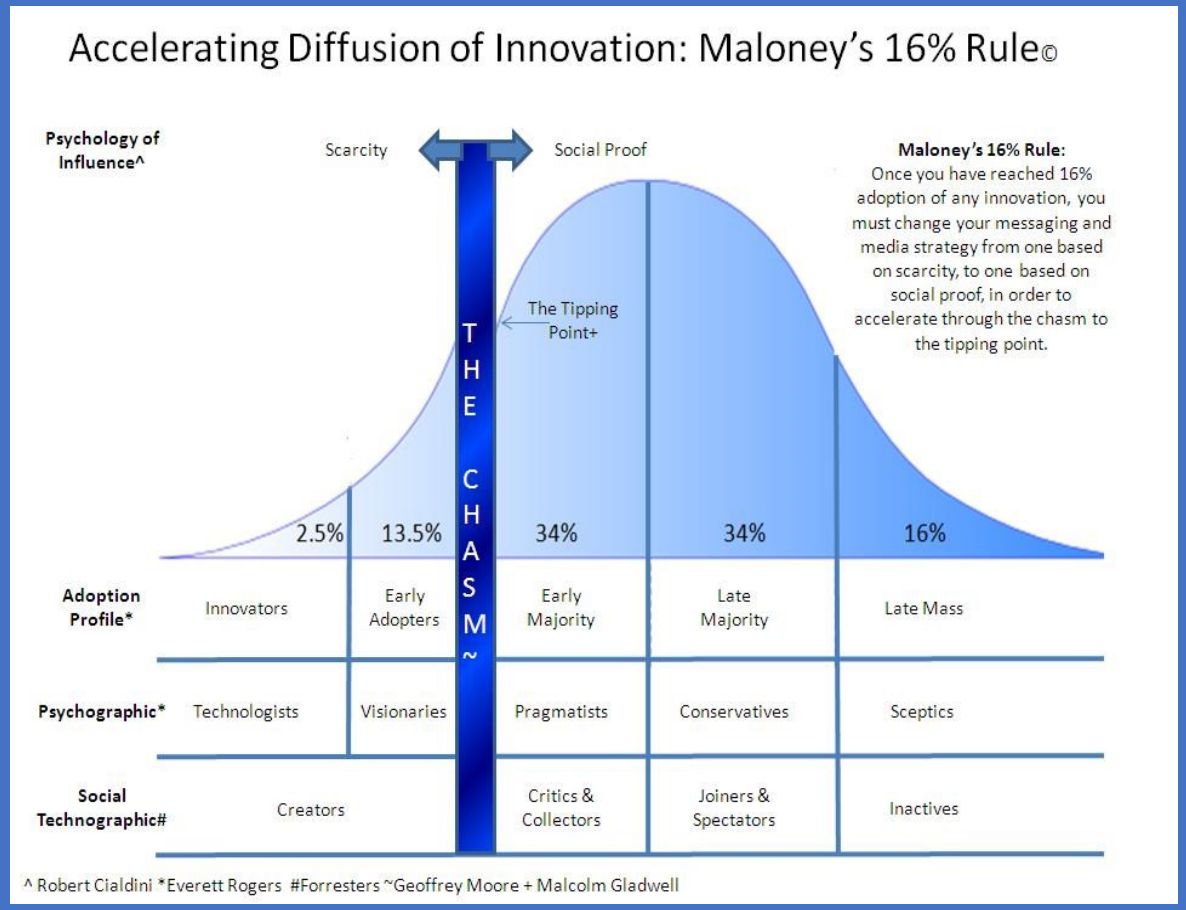


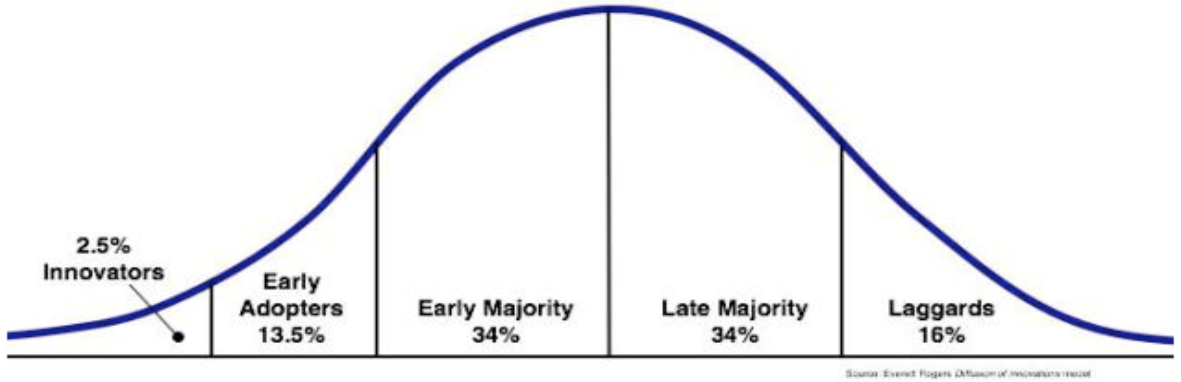
[Twitter @SashaKaletsky]

Stanford

Andrew Ng







Diffusion of Innovation Adoption Curve

The theory is that each category of adopters acts as an influencer and reference group for the next.

But there is a problem with this theory, and it lies between the Early Adopters and the Early Majority.

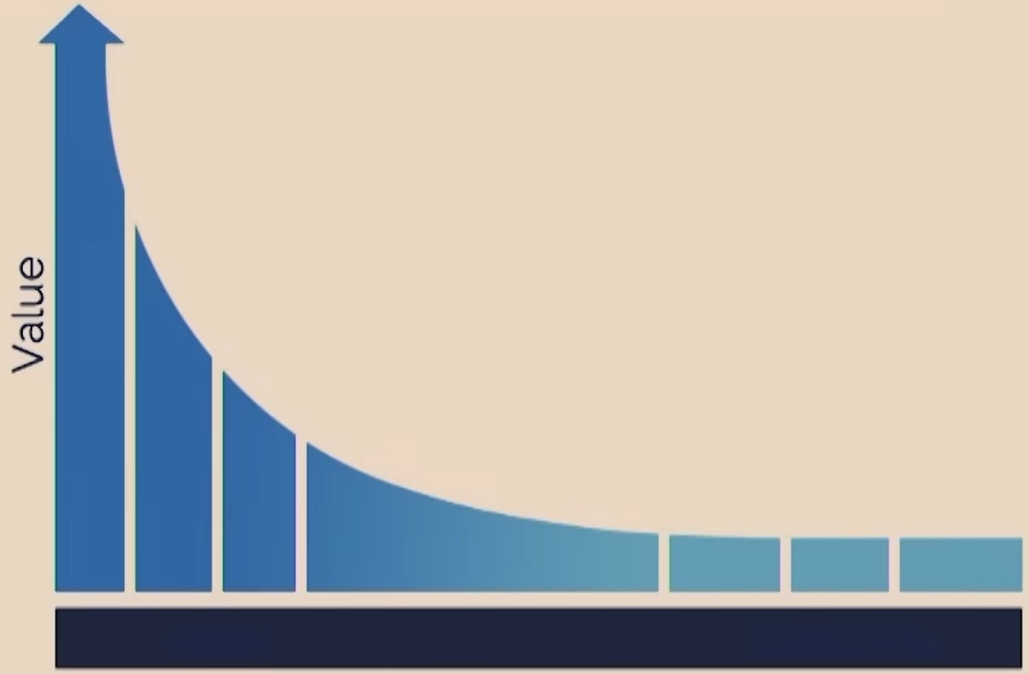
These groups don't reference each other because their Psychographics are very different.

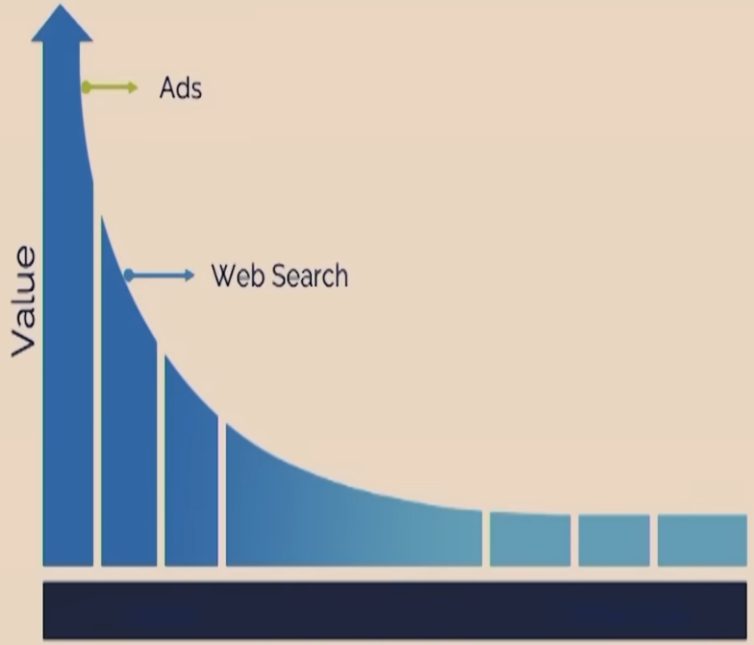
According to Rogers, Early Adopters are *"Visionaries"* and the Early Majority are *"Pragmatists"*

It is like putting a marketer and a lawyer in the same room. They are unlikely to get along, or listen to each other

Why isn't AI widely adopted yet?

Customization (long tail) problem and low/no code tools



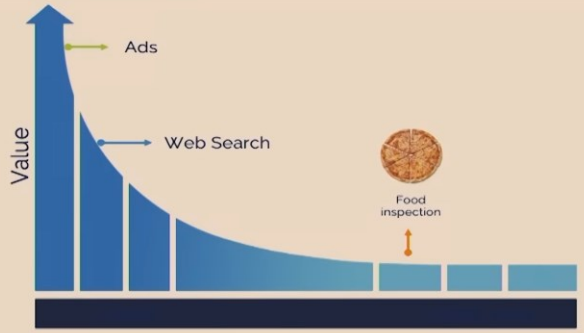


All potential AI projects, sorted in decreasing order of value

Stanford

Andrew Ng

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools

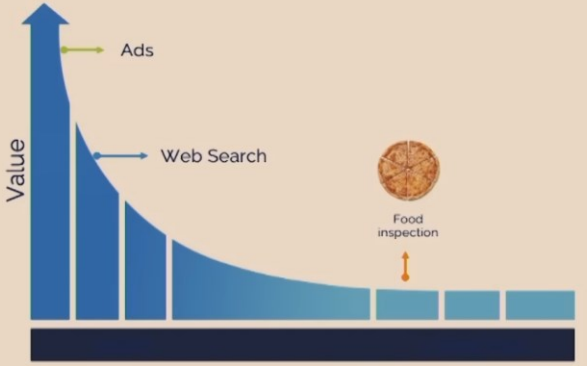


All potential AI projects, sorted in decreasing order of value

Stanford

Andrew Ng

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools



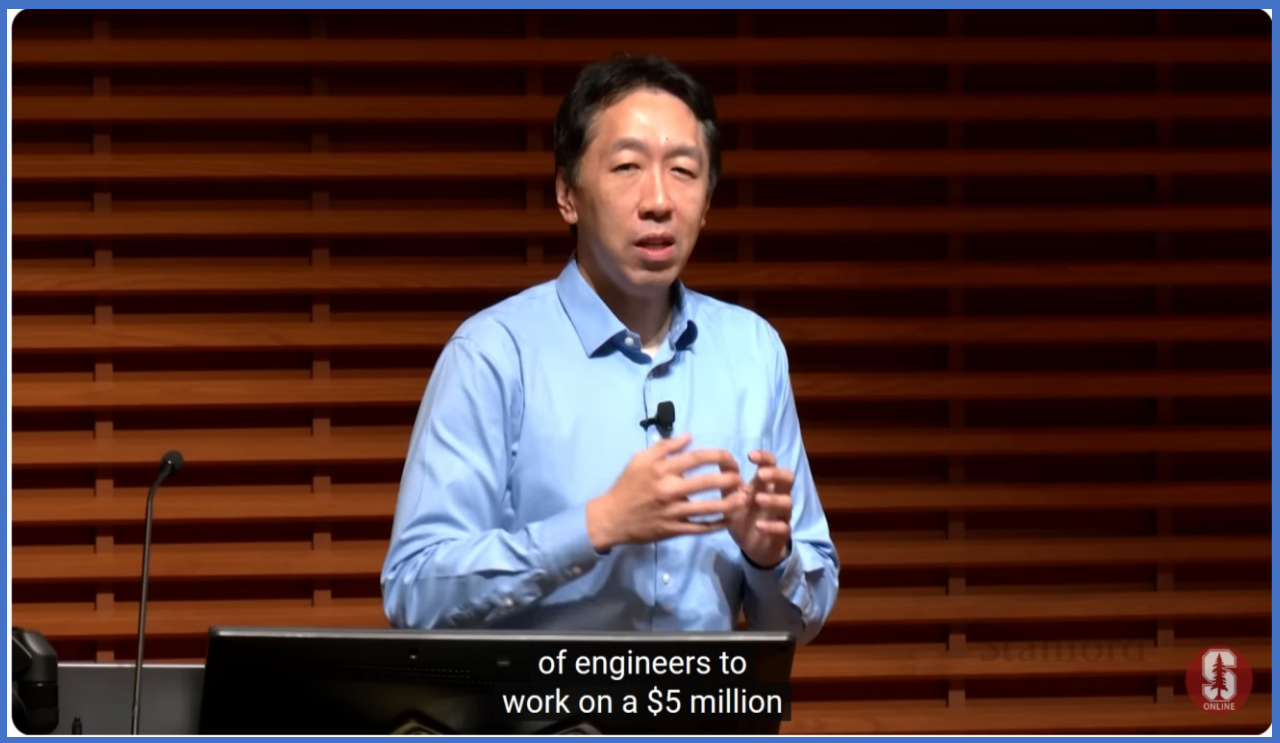
All potential AI projects, sorted in decreasing order of value

So this is about a
$5 million project.

Stanford

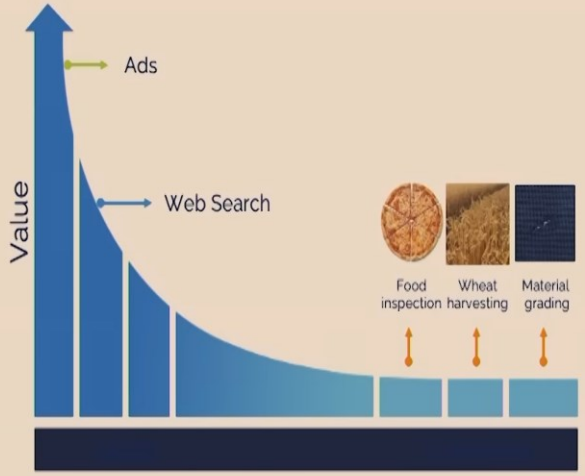
Andrew Ng


But that recipe of hiring a
hundred engineers or dozens


of engineers to
work on a $5 million


project, that
doesn't make sense.

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools

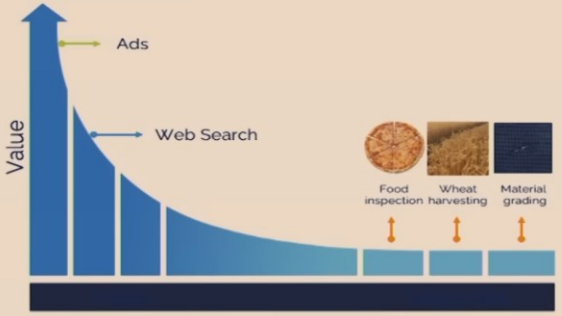


All potential AI projects, sorted in decreasing order of value

$5 million projects,
that until now,

Andrew Ng

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools

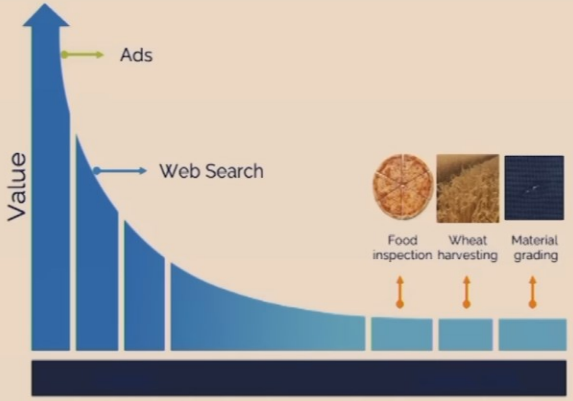


All potential AI projects, sorted in decreasing order of value

have been very
difficult to execute on

Andrew Ng

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools

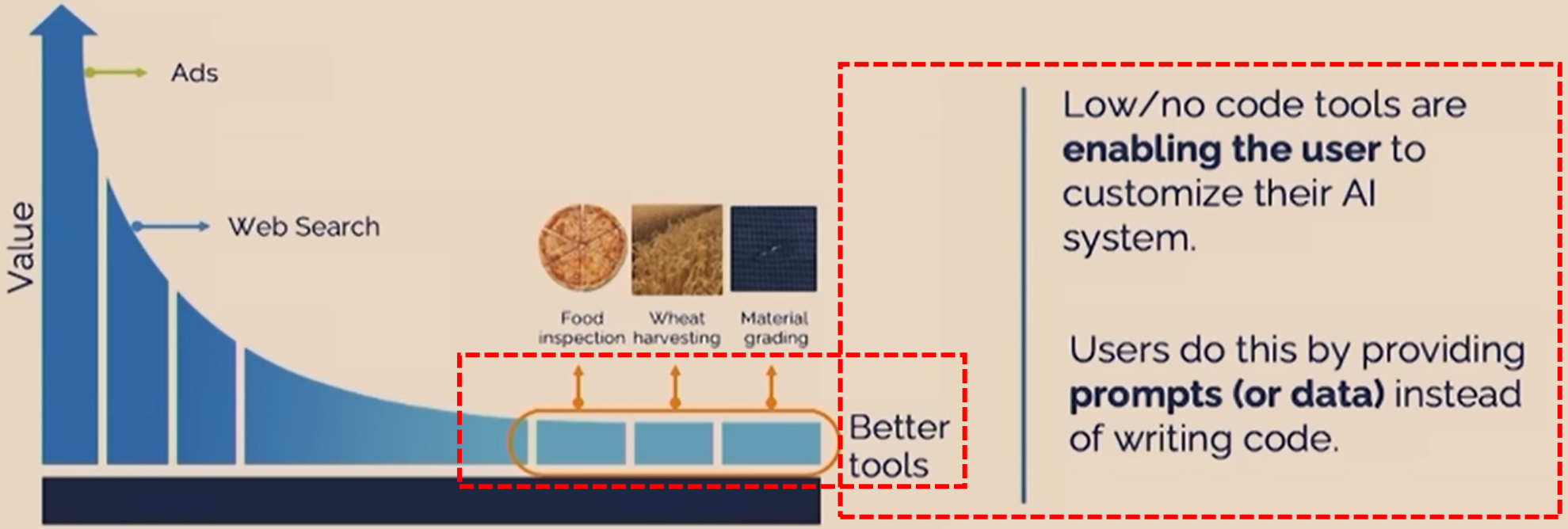


All potential AI projects, sorted in decreasing order of value

because of the high
cost of customization.

Andrew Ng

Why isn't AI widely adopted yet?
Customization (long tail) problem and low/no code tools



All potential AI projects, sorted in decreasing order of value

Stanford

Andrew Ng

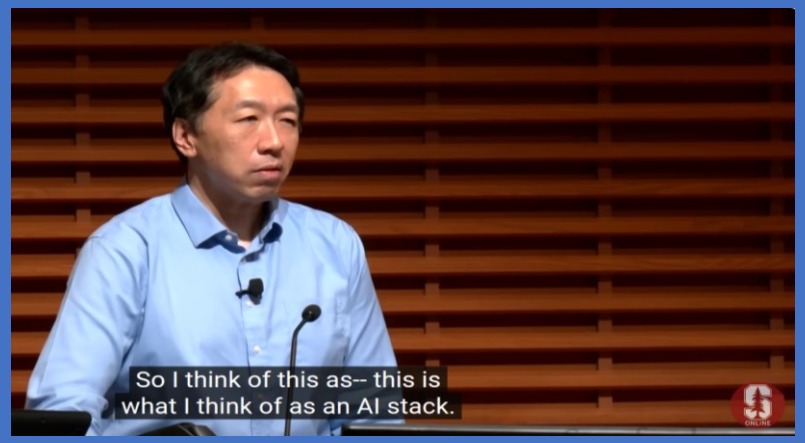
Opportunity from a new general purpose technology

- Many valuable AI projects are now possible. How do we get them done?
- *Starting new companies is an efficient way to do this.*

a lot of different companies
to pursue these very

Opportunity from a new general purpose technology

- Many valuable AI projects are now possible. How do we get them done?
- *Starting new companies is an efficient way to do this.*
- Incumbent companies also have opportunities to integrate AI into existing businesses.



Opportunity from a new general purpose technology

- Many valuable AI projects are now possible. How do we get them done?
- *Starting new companies is an efficient way to do this.*
- Incumbent companies also have opportunities to integrate AI into existing businesses.

The AI stack

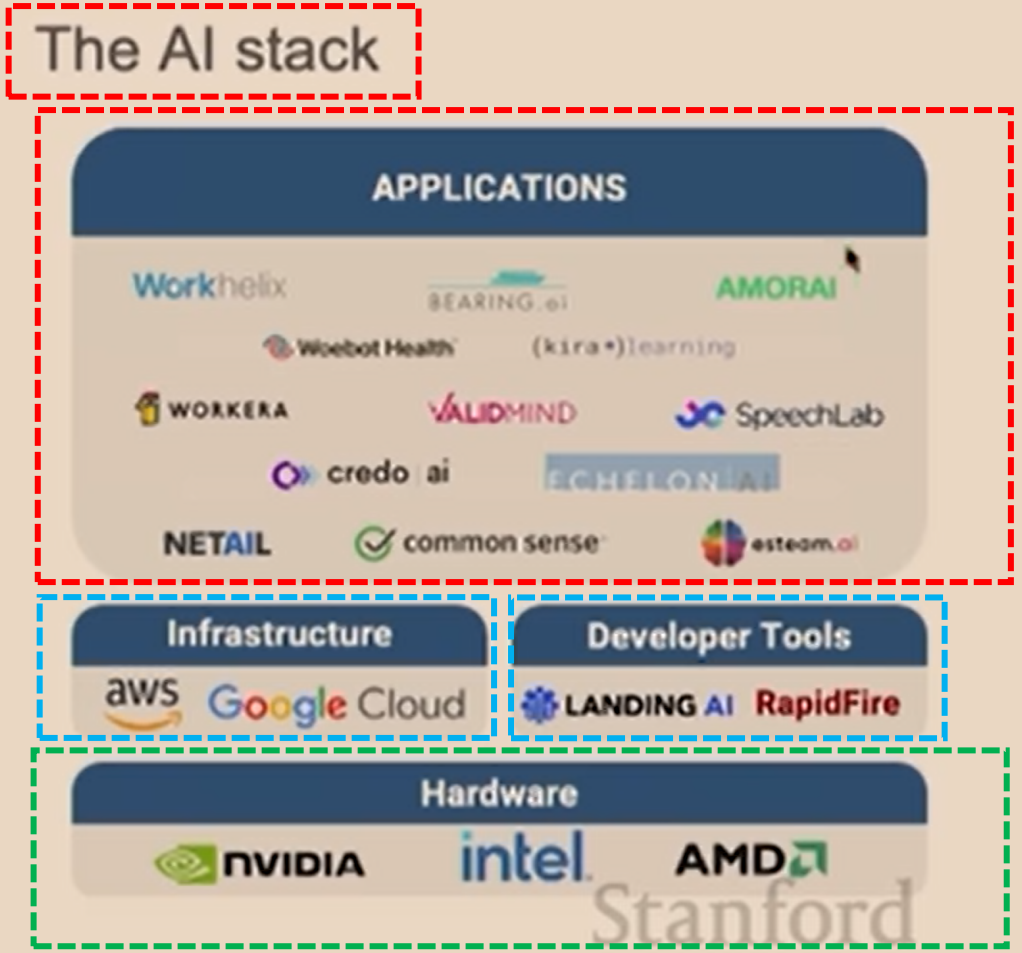
Hardware

NVIDIA intel AMD

Stanford

Opportunity from a new general purpose technology

- Many valuable AI projects are now possible. How do we get them done?
- *Starting new companies is an efficient way to do this.*
- Incumbent companies also have opportunities to integrate AI into existing businesses.



Andrew Ng

Process for building startups

Ideas

internally generated,
ideas from partners.

Process for building startups

Ideas

Validate
Stage 1

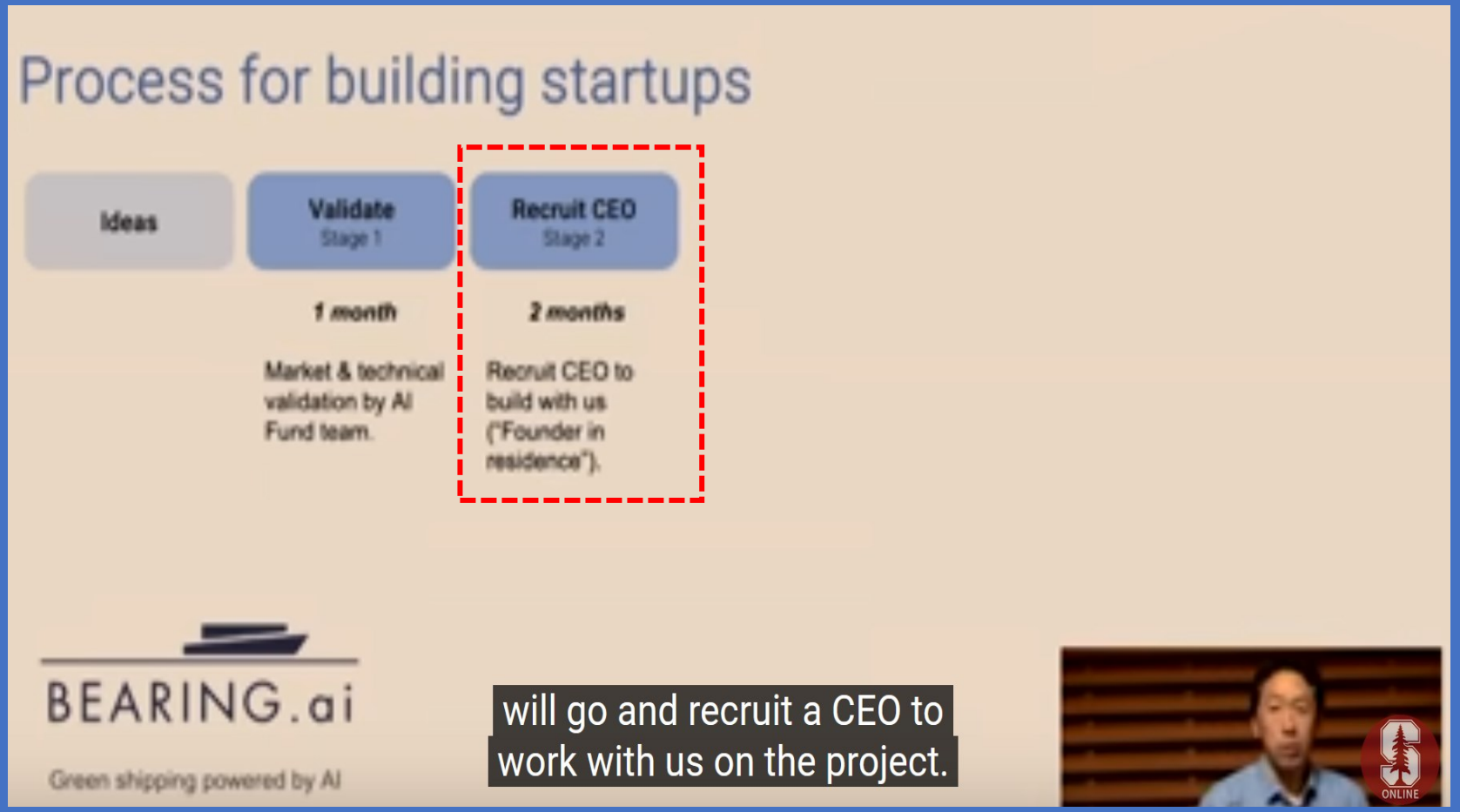
1 month

Market & technical validation by AI Fund team.

BEARING.ai

Green shipping powered by AI

So double check, is this idea
even technically feasible,


Process for building startups
Ideas
Validate
Stage 1
Recruit CEO
Stage 2
1 month
2 months
Market & technical validation by AI Fund team.
Recruit CEO to build with us ("Founder in residence").
BEARING.ai
Green shipping powered by AI
will go and recruit a CEO to work with us on the project.
ONLINE

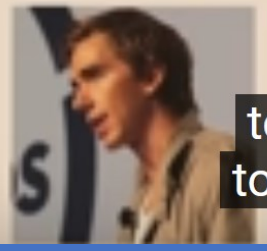

Process for building startups
Ideas
Validate
Stage 1
Recruit CEO
Stage 2
1 month
2 months
Market & technical validation by AI Fund team.
Recruit CEO to build with us ("Founder in residence").
BEARING.ai
Green shipping powered by AI
who is a reputed entrepreneur, one successful exit before.
Dylan Keil
Stanford

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3
	1 month	*2 months*	*3 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.

BEARING.ai

Green shipping powered by AI



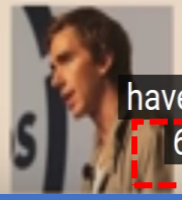
to work with them
to build a prototype

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3
	1 month	*2 months*	*3 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.

BEARING.ai

Green shipping powered by AI



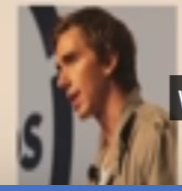
have about a two thirds, 66% survival rate,

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4
	1 month	*2 months*	*3 months*	*~12 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

BEARING.ai

Green shipping powered by AI



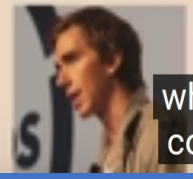
we then write the first check in,

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4
	1 month	*2 months*	*3 months*	*~12 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

BEARING.ai

Green shipping powered by AI



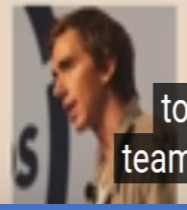
which then gives the company resources

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4
	1 month	*2 months*	*3 months*	*~12 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

BEARING.ai

Green shipping powered by AI



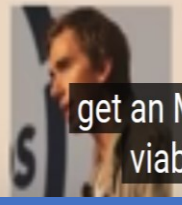
to hire an executive team, build the key team,

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4
	1 month	*2 months*	*3 months*	*~12 months*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

BEARING.ai

Green shipping powered by AI



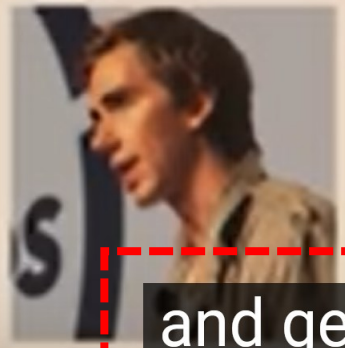
get an MVP working, minimum viable product working,

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5
	1 month	*2 months*	*3 months*	*~12 months*	*indefinite*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.	~$2-5M seed funding. Startup graduates and is well on its way.

BEARING.ai

Green shipping powered by AI



Dylan Keil

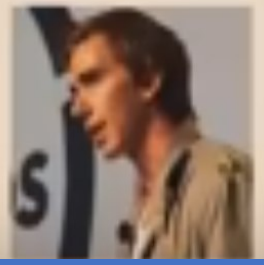
and get some real customers.

Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5
	1 month	*2 months*	*3 months*	*~12 months*	*indefinite*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.	~$2-5M seed funding. Startup graduates and is well on its way.

BEARING.ai

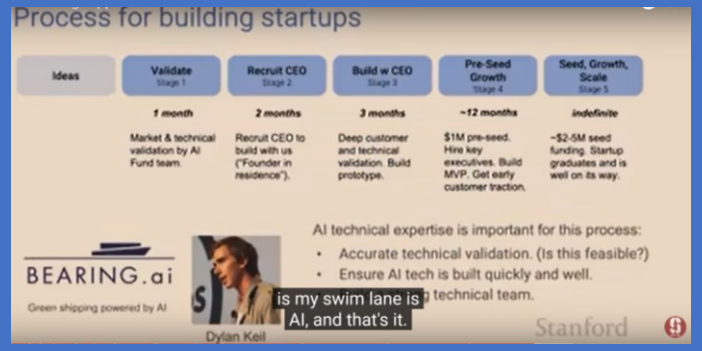
Green shipping powered by AI



AI technical expertise is important for this process:

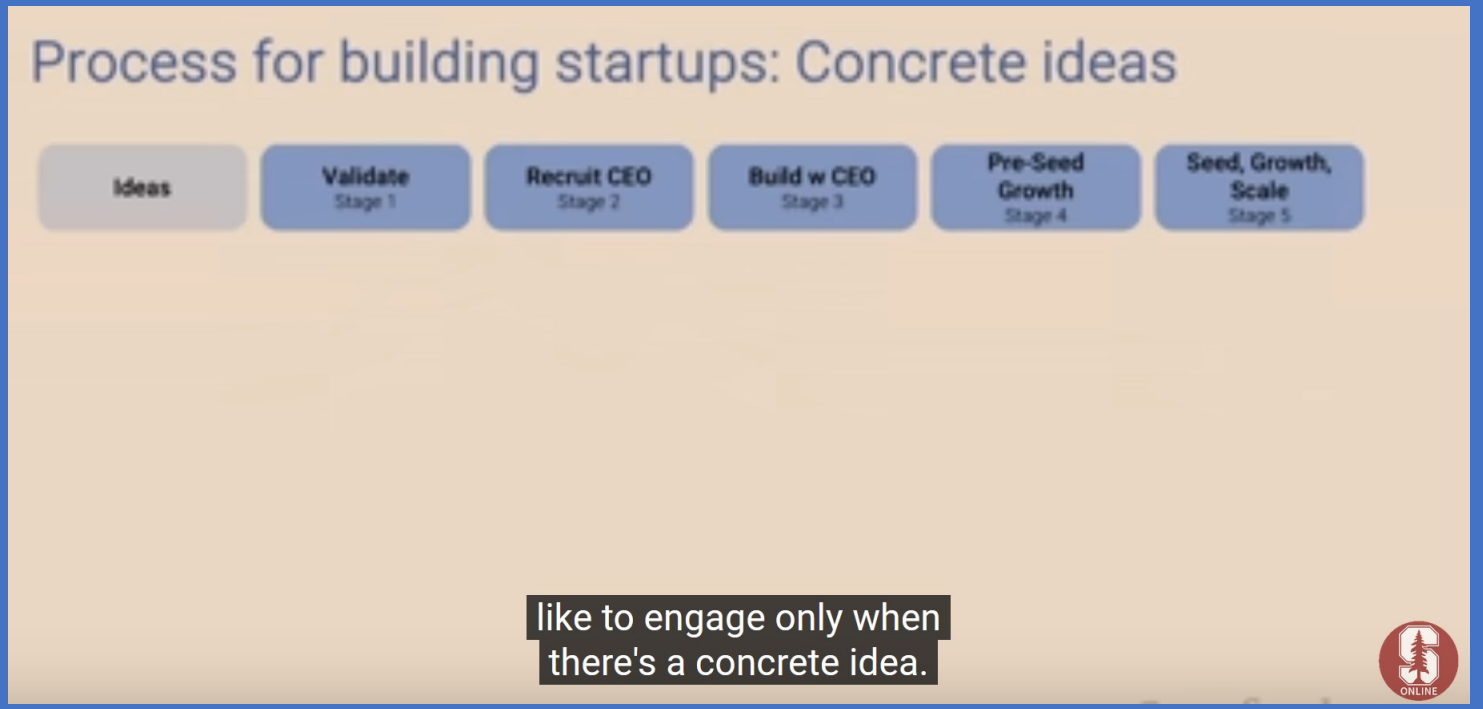
- Accurate technical validation. (Is this feasible?)
- Ensure AI tech is built quickly and well.
- Build a strong technical team.

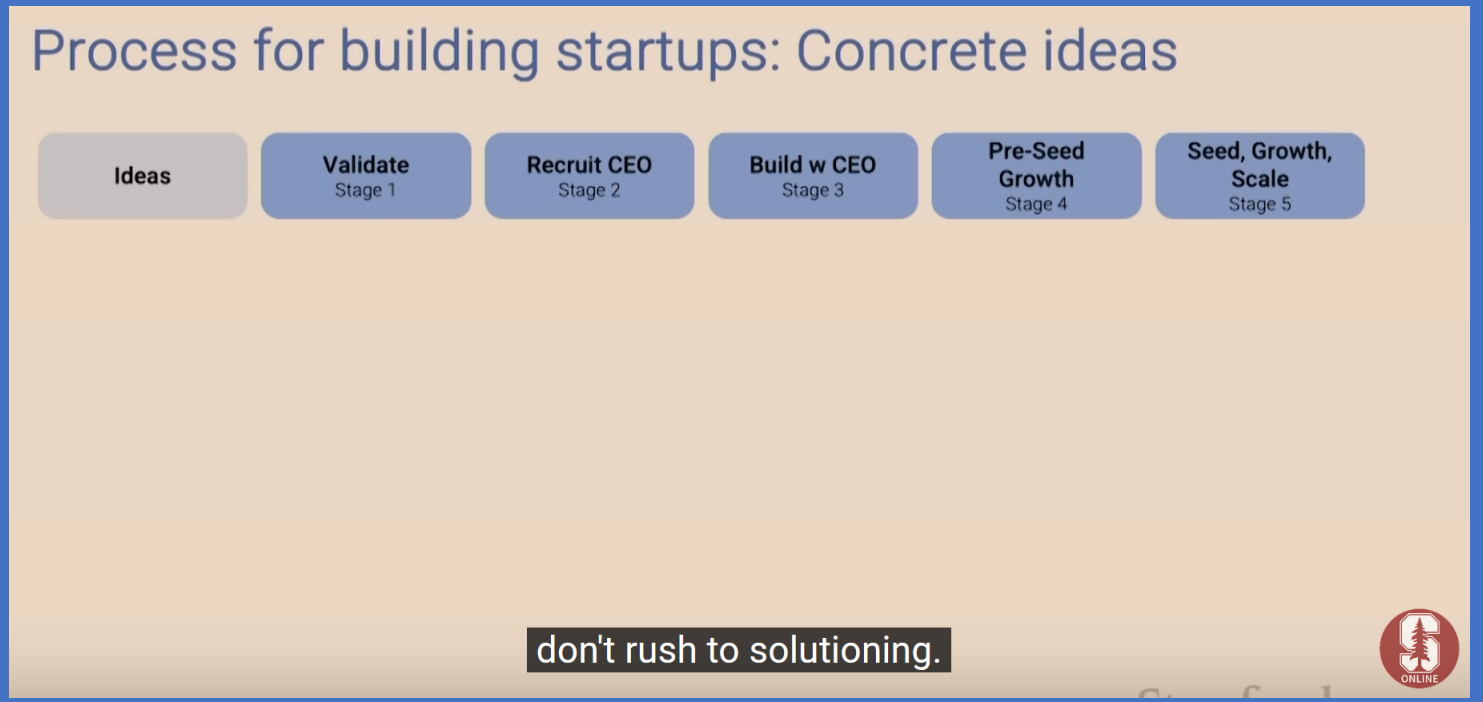
ONLINE




one other weird
thing I've learned


about building
startups, which is I


Process for building startups: Concrete ideas
Ideas
Validate
Stage 1
Recruit CEO
Stage 2
Build w CEO
Stage 3
Pre-Seed
Growth
Stage 4
Seed, Growth,
Scale
Stage 5
like to engage only when
there's a concrete idea.


Process for building startups: Concrete ideas
Ideas
Validate
Stage 1
Recruit CEO
Stage 2
Build w CEO
Stage 3
Pre-Seed
Growth
Stage 4
Seed, Growth,
Scale
Stage 5
don't rush to solutioning.

Process for building startups: Concrete ideas

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5

Explore a lot of alternatives before you do a solution.

Process for building startups: Concrete ideas

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5

Honestly, we tried that, it was very slow.

Process for building startups: Concrete ideas

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5

Not concrete:

Apply AI to financial services. (Or logistics, supply chain…).

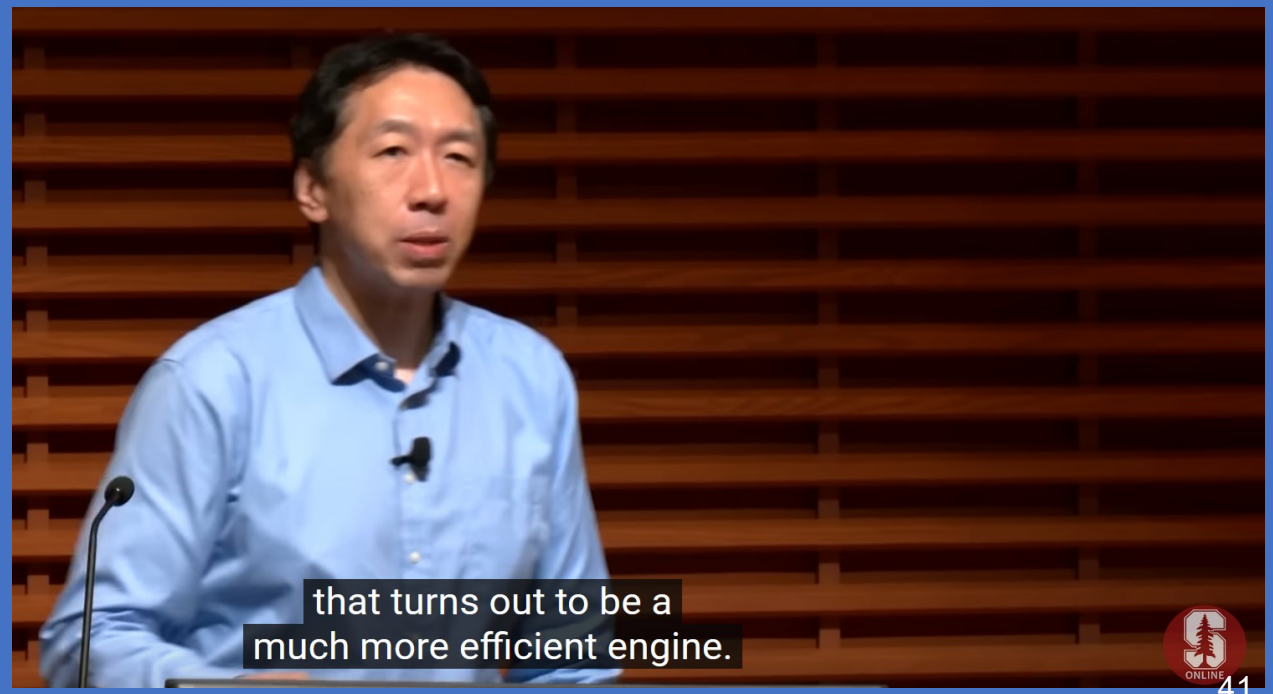
Concrete idea:

BuyGPT eliminates commercials by automatically buying every product in every ad, in exchange for not having to see any ads.

Concrete ideas:

- Can be validated or falsified efficiently
- Gives clear direction to execute
- Often from a subject matter expert who's deeply thought about a problem

thinking thing of exploring
a lot of ideas and winnowing



Responsible AI

- My teams work only on projects that move humanity forward.
- For example, we kill projects that are otherwise financially sound on ethical grounds.

Risks of AI

- AI today has problems with bias, fairness, accuracy, …. But the technology is improving quickly.

Stanford

Andrew Ng

Risks of AI

- AI today has problems with bias, fairness, accuracy, …. But the technology is improving quickly.
- AI will disrupt many occupations.

[Credit: Eloundou et al., 2023]

Stanford

Andrew Ng

Risks of AI

- Artificial General Intelligence (AI that can do anything a human can do) is still decades away.

Stanford

Andrew Ng

Risks of AI

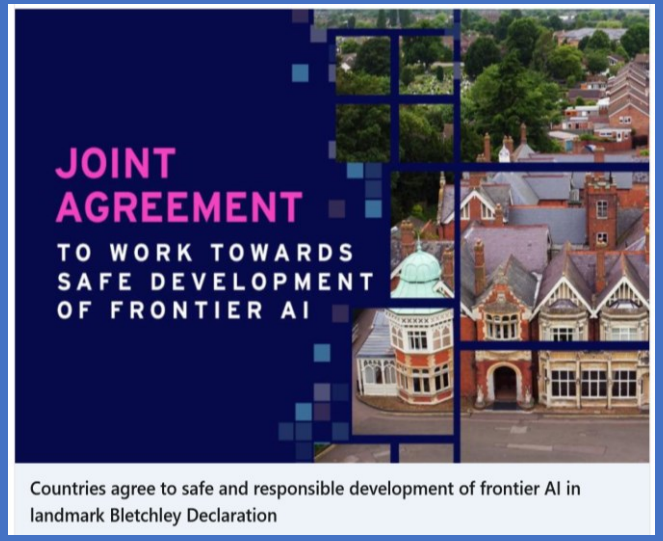
- Artificial General Intelligence (AI that can do anything a human can do) is still decades away.
- AI creating extinction risk for humanity is wildly overhyped.
  - Human society has ample experience steering very powerful entities (such as corporations and nation states).
  - AI develops gradually, and the "hard take off" scenario (where AI suddenly achieves superintelligence overnight) is not realistic.

Risks of AI

- Artificial General Intelligence (AI that can do anything a human can do) is still decades away.
- AI creating extinction risk for humanity is wildly overhyped.
  - Human society has ample experience steering very powerful entities (such as corporations and nation states).
  - AI develops gradually, and the "hard take off" scenario (where AI suddenly achieves superintelligence overnight) is not realistic.
- AI is an important piece of the solution to the real existential risks to humanity (the next pandemic, climate change, ...).

Stanford

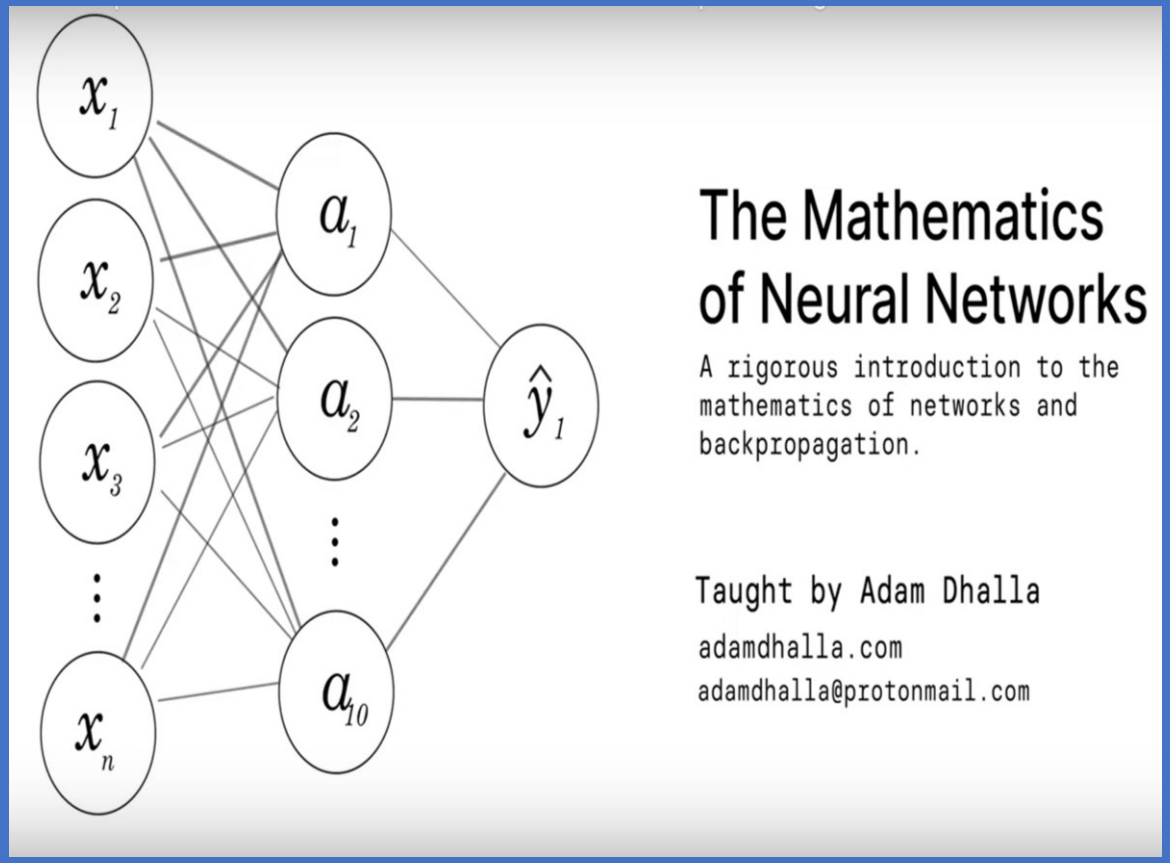
Andrew Ng



Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration

5. Yann leCun comparison between Computer/IT Science Programming and ML Science Programming from his interview on Deep Learning, ConvNets, and Self-Supervised Learning - https://www.youtube.com/watch?v=SGSOCuByo24 11:45 - 12:30

The Complete Mathematics of Neural Networks and Deep Learning



Syllabus

Part I: Introduction
40min

1.1 Prerequisites
1.2 Agenda
1.3 Notation
1.4 Big Picture
1.5 Matrix Calculus Review
  1.5.1 Gradients
  1.5.2 Jacobians
  1.5.3 Scalar Chain Rule
  1.5.4 Jacobian Chain Rule

Part II: Forward Propagation
30min

2.1 The Neuron Function
2.2 Weight and Bias Indexing
2.3 A Layer of Neurons

Part III: Derivatives of Neural Networks & Gradient Descent
2hr 20min

3.1 Motivation & Cost Function
3.2 Differentiating a Neuron's Operations
  3.2.1 Binary Elementwise Funcs
  3.2.2 Hadamard Product
  3.2.3 Scalar Expansion
  3.2.4 Sum
3.3 Derivative of an activation
3.4 Derivative of the cost
3.5 Intuition: derivative of the cost w.r.t the weights
3.6 Differentiating w.r.t the bias
3.7 Gradient Descent Intuition
3.8 The Gradient Descent Algorithm & Stochastic Gradient Descent
3.9 Derivatives of a layer (and why it doesn't work)

Part IV: Backpropagation
1hr 30min

4.1 The Error of a Node
4.2 The Four Equations of Backpropagation
  4.2.1 E1: Error of a^L
  4.2.2 E2: Error of a^l
  4.2.3 E3: Cost w.r.t biases
  4.2.4 E4: Cost w.r.t weights
  4.2.5 Equation 4 vectorized
4.3 Tying Part III and Part IV together
4.4 The Backpropagation Algorithm
4.5 Looking forward

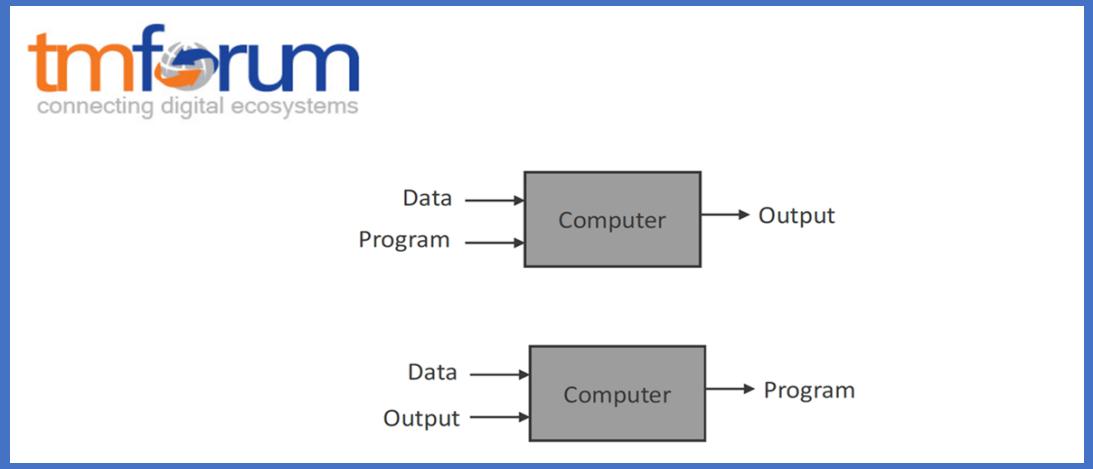
timestamps in description

https://www.youtube.com/watch?v=lxl3nykKG9M

2. 5G System use of AI/ML

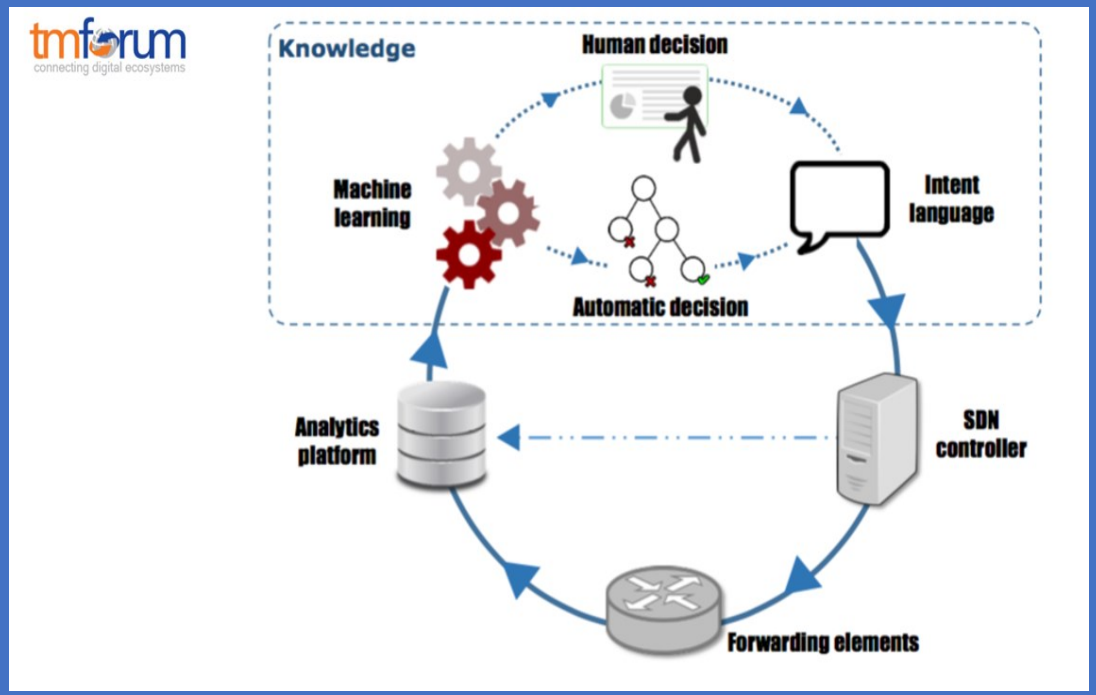
*In **Programming**,* a Human writes a Computer Program and provides the Data, which the Computer processes to create the Output.

In Machine Learning (ML), Humans provide the Data along with the Desired Output, Rules and Constraints, and the Computer (Algorithms with trained Models) writes the Program to deliver this.



A *Knowledge-defined Network (**KDN**)* operates by means of a Control Loop to provide:
- Automation,
- Recommendation,
- Optimization,
- Validation and
- Estimation.

CSPs are beginning to *use AI and Machine Learning (ML) in three (3) Key Areas*:
1. Customer Experience Management
2. Service Management and Optimization
3. Network Management and Optimization

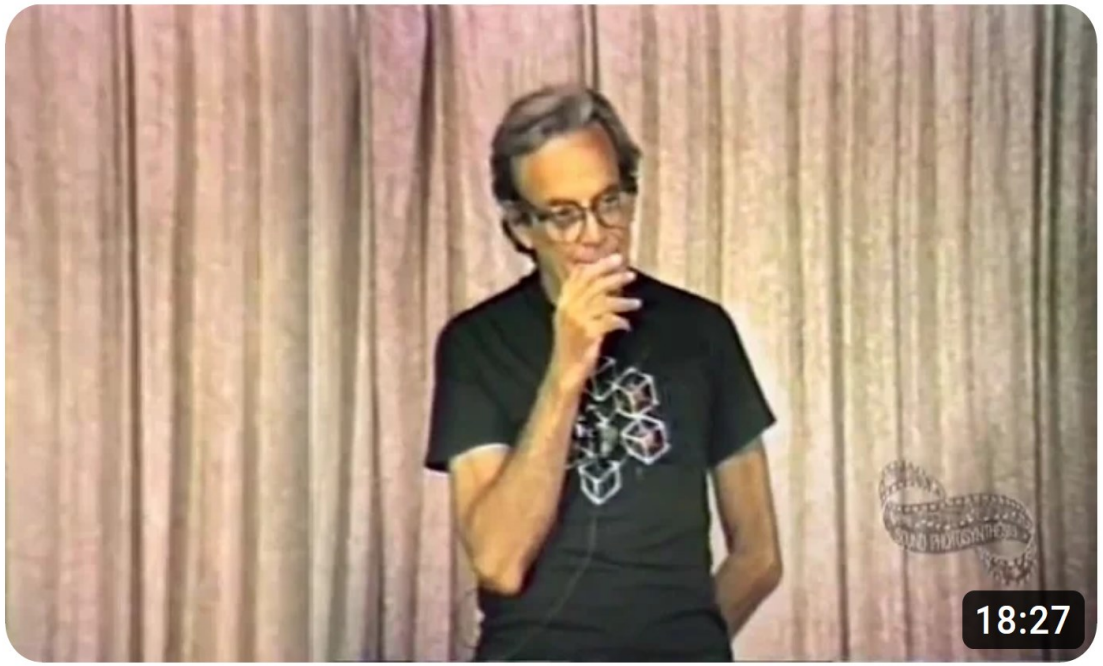


The Knowledge Plane (KP) is a distributed & decentralized construct within the Network that
- Gathers,
- Aggregates, and
- Manages

Information about Network behavior and Operation, and provides an integrated view to all parties (Operators, Users, and the Network itself). The Goal is to enlarge our view of what constitutes ***the Network to match the intuition of a User,*** and to enhance our ability to manage the network intelligently, without disturbing the open and unknowing forwarding plane (Ref. D.C., KP for I., v4.6 05/03).

2. Richard Feynman: Can Machines Think?

03:55-07:33 and 10:55- 11:20 11:40-18:00



Richard Feynman: Can Machines Think?

Audience Question:

Can computers discover new ideas
and relationships by themselves?

Audience Question:

Do you think there will ever be
a machine that will think like human beings
and be more intelligent than human beings?

Ref: https://www.youtube.com/watch?v=ipRvjS7q1DI

1. Cloud & Communications Systems' (current) Challenges & Issues

Today's Cloud and Communications Systems are NOT CAPABLE of

- *Capturing,*
- *Transmitting,*
- *Storing, and*
- *Analysing*

the Petabytes of Data generated by the soon-to-be trillions of Sensors operating 24/7.

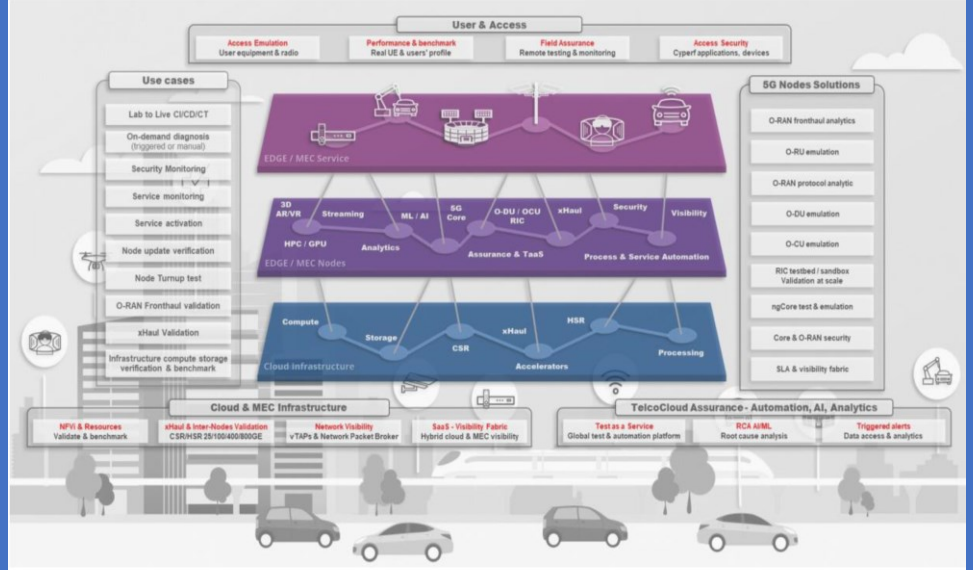
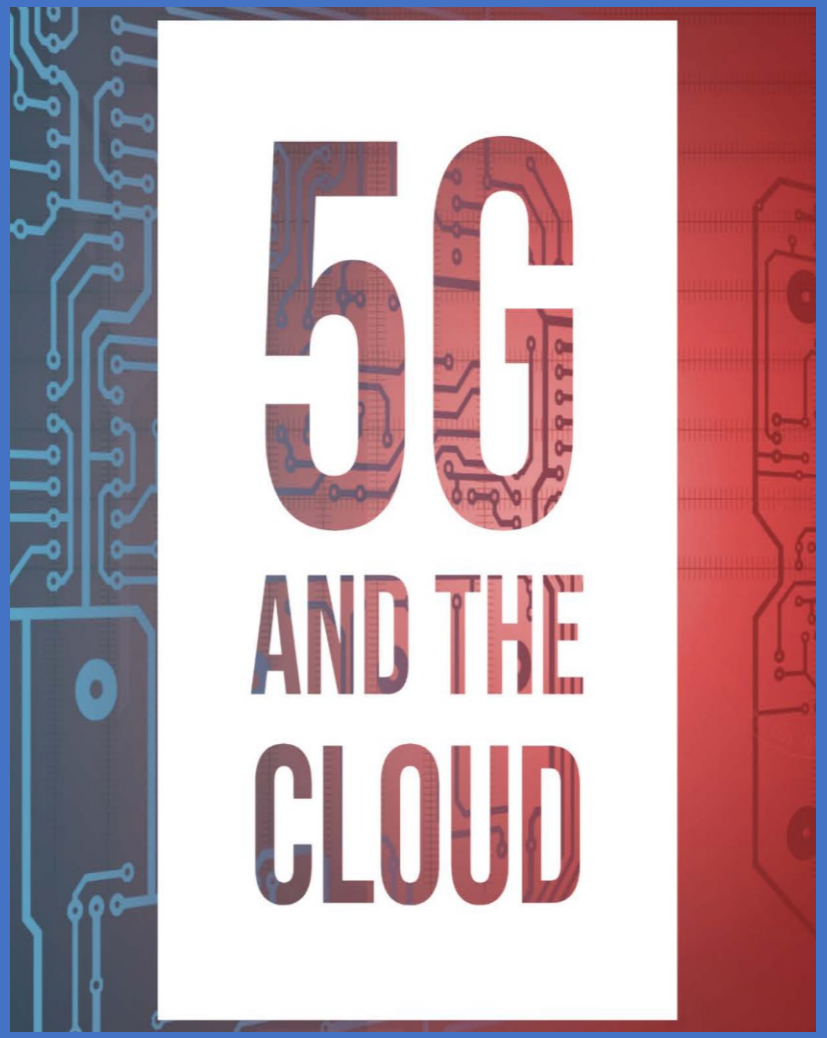


Figure : Telco Edge Cloud, Next-Gen Service Assurance at Scale

They are also NOT PREPARED to deliver the Compute needed for Real-Time AI/ML Inferencing required to drive such demands that we anticipate will come from:

- *FoF (Factory of the Future)*
- *VR/XR/MR (Virtual, Extended, Mixed Reality and Extended Reality) with Haptic Interactions,*
- *NPNs/SNPNs Non Public Network/Stand-alone NPNs*
- *PINs and CPNs (Personal IoT Network/Customer Premises Networks)*
- *(V2X) Connected Vehicles,*
- *Assisted living, or*
- *Merging of Physical & Digital worlds with 5G & B5G*

1. CLOUD NATIVE PHILOSOPHY-RELATED ISSUES



The Cloud Native issues appear because the whole of the Cloud Native Development Philosophy has been applied:

- *without consideration* of the *Actual Deployment and Operational Environments.*

In brief, **the *Positive and Negative Aspects of Cloud Native from a 5G SA/SBA Network Function (NF) Perspective* are summarized as follows:**

POSITIVE (+)	NEGATIVE (-)
Cloud Native has undeniably improved: - Development, - Delivery and Test, - In-Service Upgrades - Improved Version Management	The Context in which *Cloud Native* was designed is being *misrepresented or abused in two (2) senses:* 1. Cloud Native was *designed for People who write & operate the Applications*. In today's Cellular Network, this clearly is not the case 2. ***Cloud Native* was designed for *Applications* in which long *interruptions* are *tolerable***, therefore, *good Reliability is measured* ***in minutes of outage per month.*** This is also clearly not the case for **(2G, 3G, 4G, 5G) Cellular Communication Networks** where the expectation is that **outages last less than 5.26 minutes per year.**

50

1. The Cloud is "Changing"

1st - Applications want to be deployed anywhere & change deployment anytime.

The focus moves from "Sharing Resources" to "Composing Dynamic Capabilities, in Real-time, even after Deployment.

Applications will be Delay- and Latency Sensitive*, on *varying Time-scales* with *different Hard- & Soft Boundaries.

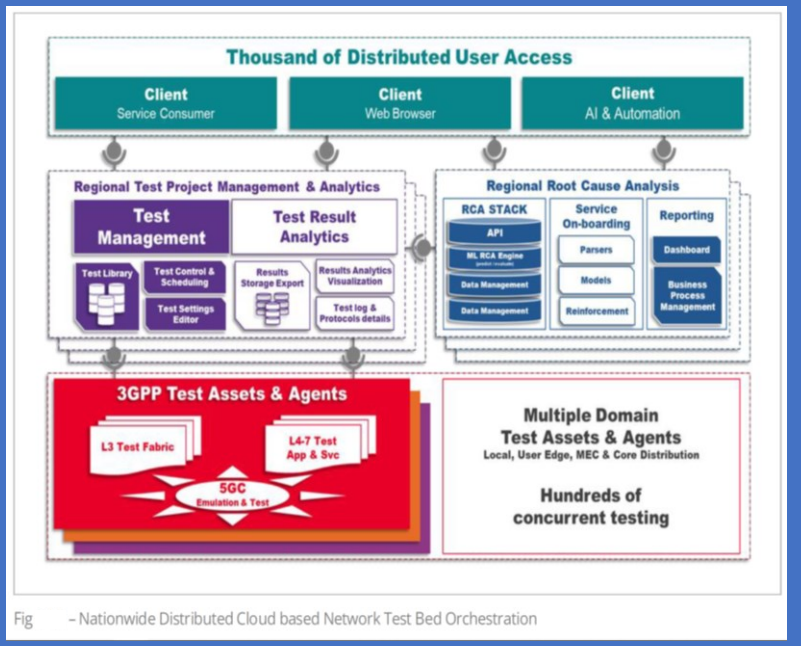


Fig – Nationwide Distributed Cloud based Network Test Bed Orchestration

Communication, Compute, and Storage must be considered as an ***Integrated Set of Changeable Configurations*** that provide the required Service to an application.

2nd - "Centre of Gravity is moving toward the "Devices" ("End-points"*) & Interactions in a Cyber-Physical World best suited for these tasks and configure any required communication between all end points in important areas such as

- IoT,
- Industry 4.0,
- 5G NPNs/SNPNs/PINs, or
- Retail and Public Services.
- eHealth & Ageing and Living well

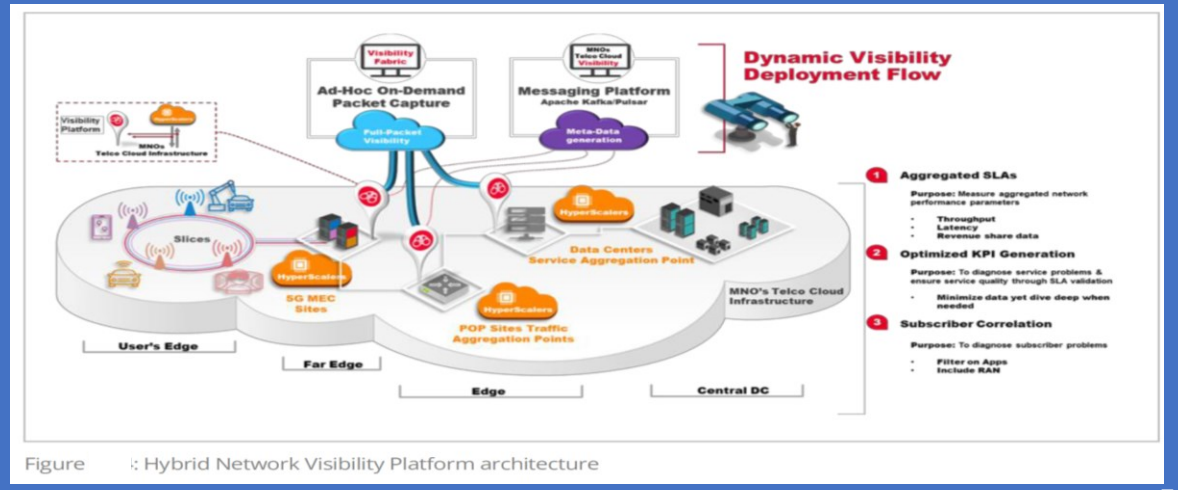


Figure : Hybrid Network Visibility Platform architecture

**You might be vigilant with the terms you use w.r.t. the terms "end-points" &/or "Edge" from Service E2E Solution Architecture fulfilling the 3GPP specified 5QI (QoS) Service Requirements & KPIs.*

Management of Resources and Workloads:

Most Orchestration Frameworks today use a Centralized Approach (where) One (1) Entity has knowledge of all the Resources in the System and Plan how the Workloads will be mapped.

With the start of Docker & containers, the Kubernetes Project was started to provide a lightweight & scalable Orchestration solution.

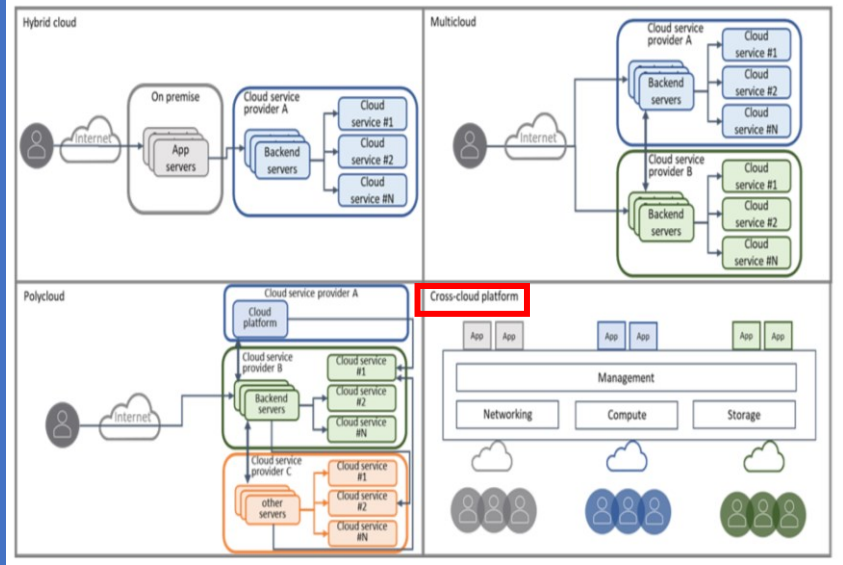
Most existing Compute Systems today, including Edge Computing Systems, rely **on "Static Provisioning"**.

Thus, the SW & the Services needed to perform the Compute are already residing at the Edge Server prior to an Edge node requests a Service & the pool of HW resources is also known a priori to Kubernetes.

This Architecture works well for Cloud & the (ETSI) MEC where a Centralized Orchestration is used.

Since the Resources of the Pervasive Edge are independently owned, the ***Orchestration Frameworks need to be extended to handle Dynamic and Multi-Tenant Resources*** *in a secure manner.*

Figure multi-cloud deployment models



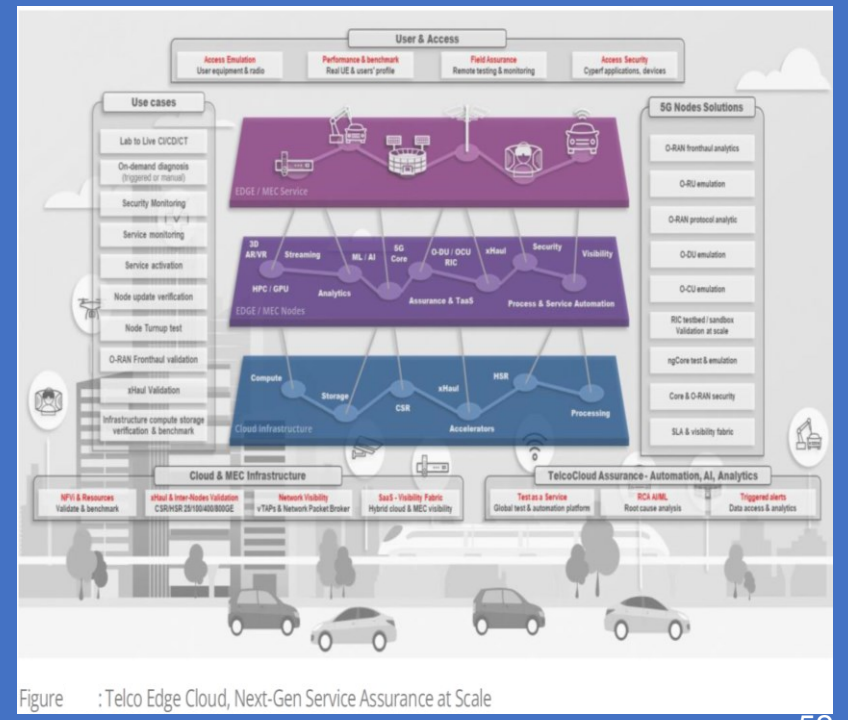
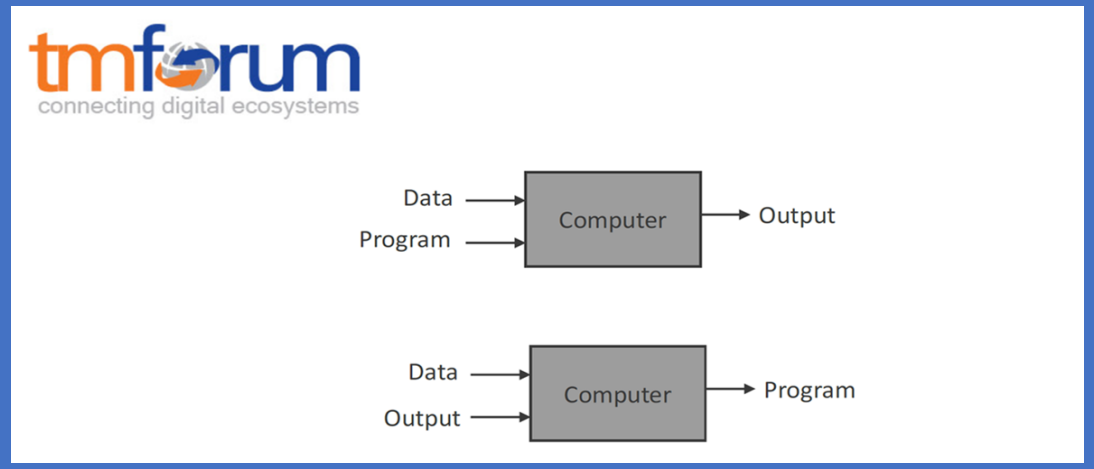


Figure : Telco Edge Cloud, Next-Gen Service Assurance at Scale

2. 5G System use of AI/ML

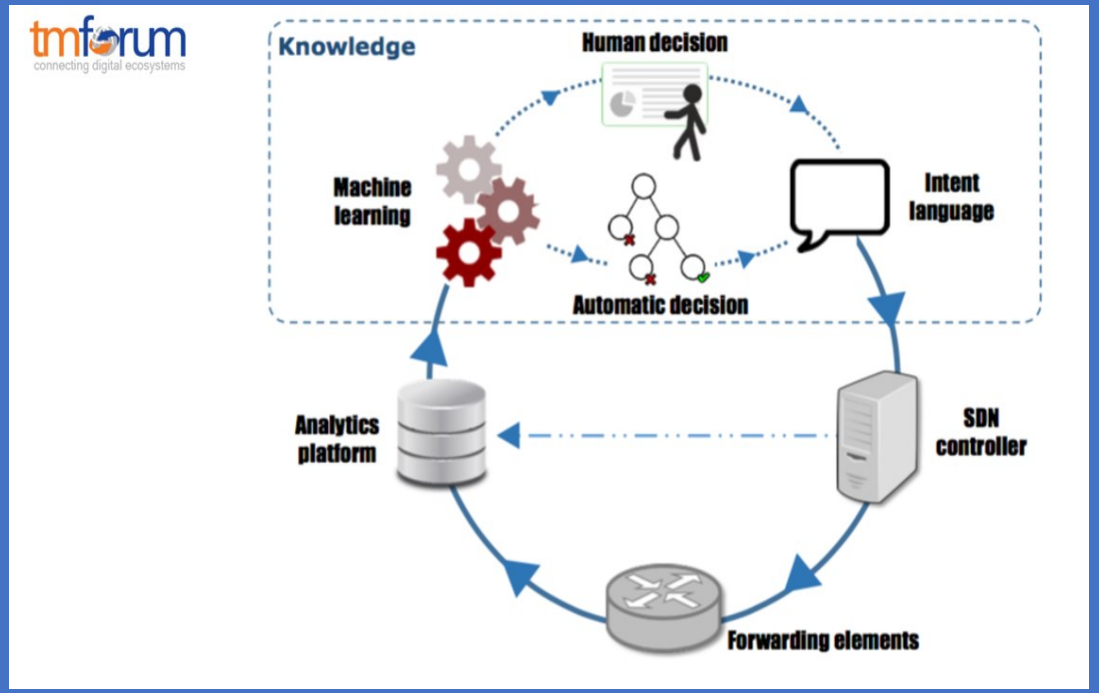
In **Programming**, a Human writes a Computer Program and provides the Data, which the Computer processes to create the Output.

In Machine Learning (ML), Humans provide the Data along with the Desired Output, Rules and Constraints, and the Computer (Algorithms with trained Models) writes the Program to deliver this.



A *Knowledge-defined Network (**KDN**)* operates by means of a Control Loop to provide:
- Automation,
- Recommendation,
- Optimization,
- Validation and
- Estimation.

CSPs are beginning to *use AI and Machine Learning (ML) in three (3) Key Areas*:
1. Customer Experience Management
2. Service Management and Optimization
3. Network Management and Optimization



The Knowledge Plane (KP) is a distributed & decentralized construct within the Network that
- Gathers,
- Aggregates, and
- Manages

Information about Network behavior and Operation, and provides an integrated view to all parties (Operators, Users, and the Network itself). The Goal is to enlarge our view of what constitutes ***the Network to match the intuition of a User,*** and to enhance our ability to manage the network intelligently, without disturbing the open and unknowing forwarding plane (Ref. D.C., KP for I., v4.6 05/03).

3. 5G System use of AI/ML

5G System AI/ML Model Transfer

The ***5G System*** can at least support *three (3) types of AI/ML Operations:*

1.AI/ML Operation splitting between AI/ML (Network) End-points: The AI/ML Operation/Model is split into Multiple Parts according to the current Task and Environment. The intention is to *off-load the Computation-Intensive, Energy-Intensive Parts to Network End-points*, whereas *leave the Privacy-sensitive and Delay-sensitive Parts at the End Device.* The Device executes the Operation/Model up to a specific Part/Layer and ***then sends the intermediate Data to the Network Endpoint***. The Network End-point executes the remaining Parts/Layers and feeds the Inference Results back to the Device.

2. AI/ML Model/Data Distribution and Sharing over 5G System: Multi-functional Mobile Terminals might need to switch the AI/ML Model in response to task and environment variations. The condition of adaptive model selection is that the models to be selected are available for the Mobile Device. However, given the fact that the AI/ML Models are becoming increasingly diverse, and with the *limited storage resource in a UE*, it can be determined to *not pre-load all candidate AI/ML Models on-board*. *Online model distribution (i.e. New Model Downloading) is needed*, in which an AI/ML Model can be distributed from a NW end-point to *the Devices when they need it to adapt to the changed AI/ML Tasks and Environments. For this purpose, the Model Performance at the UE needs to be monitored constantly.*

3. Distributed/Federated Learning (FL) over 5G System: The Cloud Server trains a Global Model by aggregating Local Models partially-trained by each End devices. Within each training iteration, a UE performs the training based on the Model downloaded from the AI Server using the Local Training Data. Then the UE reports the interim training results to the Cloud server via 5G UL channels. The Server aggregates the Interim Training Results from the UEs and updates the Global Model. The updated Global Model is then distributed back to the UEs and the UEs can perform the training for the next iteration.

In Mobile Communications Systems, Mobile Devices (e.g. Smartphones, Automotive, Robots) are increasingly replacing conventional Algorithms (e.g. Speech Recognition, Image Recognition, Video Processing) with AI/ML Models to enable Applications.

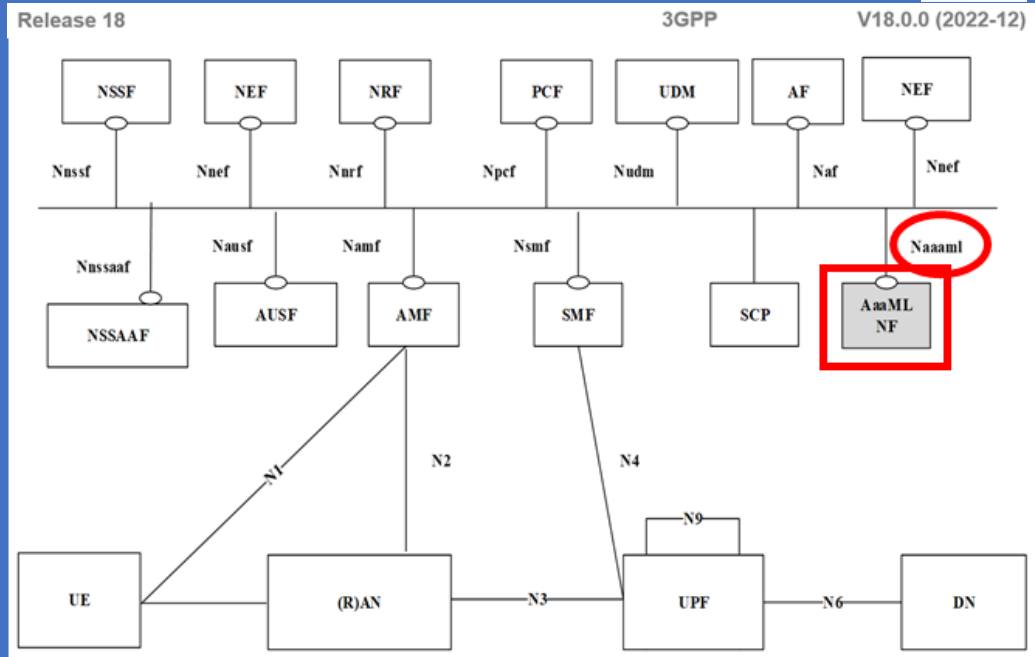


Figure: 5G System Service Architecture with AaaML NF

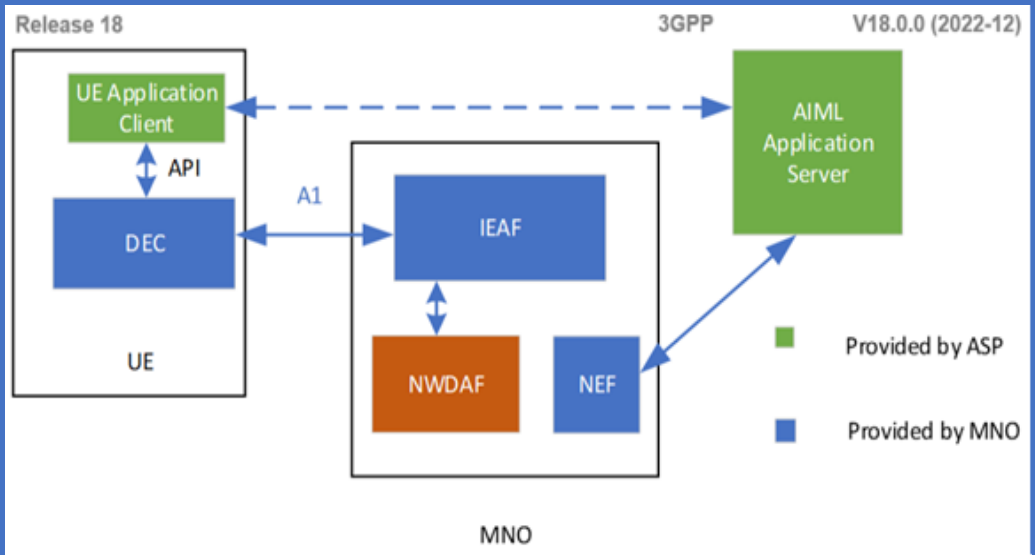


Figure: 5G IEAF (Data Information AF)

3. 5G System use of AI/ML

The AI/ML Techniques and relevant Applications are being increasingly adopted by the wider Industries and proved to be successful. These are now being applied to Telecommunication Industry including Mobile Networks.

Although AI/ML Techniques, in general, are quite mature nowadays, some of the relevant aspects of the Technology are still evolving while New Complementary Techniques are frequently emerging.

The AI/ML Techniques can be generally characterized from different perspectives including the followings:

- Learning Methods : The Learning Methods include Supervised Learning, Semi-Supervised Learning, Unsupervised Learning and Reinforcement Learning. Each Learning Method fits one (1) or more specific Category of Inference (e.g. Prediction), and requires Specific Type of Training Data. A brief comparison of these learning methods is provided in the Table:

- Learning complexity: As per the Learning Complexity, there are Machine Learning (i.e. basic learning) and Deep Learning (DL).

- Learning Architecture: Based on the Topology and Location where the Learning Tasks take place, the AI/ML can be categorized to Centralized Learning, Distributed Learning and Federated Learning.

- Learning Continuity: From Learning Continuity Perspective, the AI/ML can be off-line Learning or Continual Learning.

Release 18 3GPP V18.1.0 (2023-09)

Table : Comparison of AI/ML Learning Methods

	Supervised learning	Semi-supervised learning	Unsupervised learning	Reinforcement learning
Category of inference	Regression (numeric), classification	Regression (numeric), classification	Association, Clustering	Reward-based behaviour
Type of training data	Labelled data (Note)	Labelled data (Note), and unlabelled data	Unlabelled data	Not pre-defined
NOTE: The labelled data means the input and output parameters are explicitly labelled for each training data example.				

Artificial Intelligence/Machine Learning (AI/ML) Capabilities are used in various Domains in 5G System, including:

- Management and Orchestration for Data Analytics (MDA)
- 5G Networks Data Analytics (NWDAF)
- NG-RAN, e.g. RAN Intelligence.

The AI/ML-Inference Function in the 5GS uses the ML Model and/or AI Decision Entity for Inference. Each AI/ML Technique, depending on the adopted specific Characteristics, suitable for supporting certain Type/Category of Use Case(s) in 5G System.

To enable and facilitate the AI/ML Capabilities with the suitable AI/ML Techniques in 5GS, the ML Model and AI/ML Inference Function need to be managed.

Annex 3: 5G Architecture for Hybrid and Multi-Cloud Environments

The Main Challenges to overcome in a Hybrid & Multi-Cloud Strategy are:

1. Maintaining Portability; 2. Controlling the Total Cost of Ownership (TCO); 3. Optimizing Productivity & Time to Market (TTM).

DevOps – *a Set of Practices* that brings together *SW Development & IT operations* with the Goal of Shortening the Development & Delivery Cycle & increasing SW Quality **- is** often thought of and discussed **in the Context of a Single Company or Organization. The Company usually Develops the SW, Operates it & Provides it as a Service to Customers,** according to the **SW-as-a-Service (SaaS) Model. Within this context**, it is easier to have **Full Control over the Entire Flow**, including **Full Knowledge of the Target Deployment Environment.**

In the **Telecom Space**, by contrast, we typically follow the **"as-a-Product (aaP) Business model**, in which **SW is developed by Network SW Vendors** e.g. as Ericsson (Nokia, Huawei, ZTE) & provided to Communication Service Providers (CSPs) that Deploy & Operate it within their Network. This **Business Model requires the consideration of additional aspects**.

The most important contrasts between the Standard DevOps SaaS Model & the Telecom aaP Model are the **Multiplicity of Deployment Environments & the fact the Network SW Vendor Development Teams cannot know upfront exactly what the Target Environment looks like**.

Although a SaaS Company is likely to Deploy & Manage its SW on two (2) or more different Cloud Environments, **this is inevitable within Telco**, as each CSP creates &/or selects its own Cloud infrastructure (Fig. 1 below).

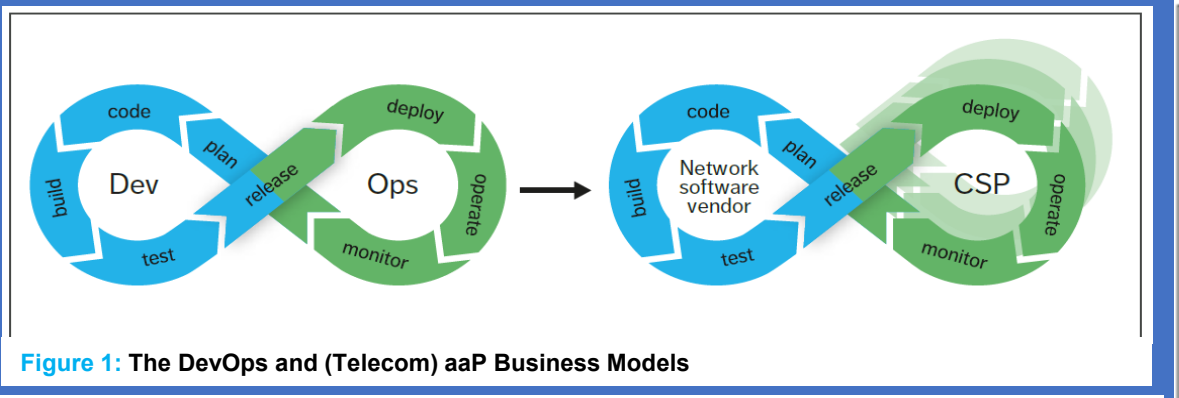


Figure 1: **The DevOps and (Telecom) aaP Business Models**

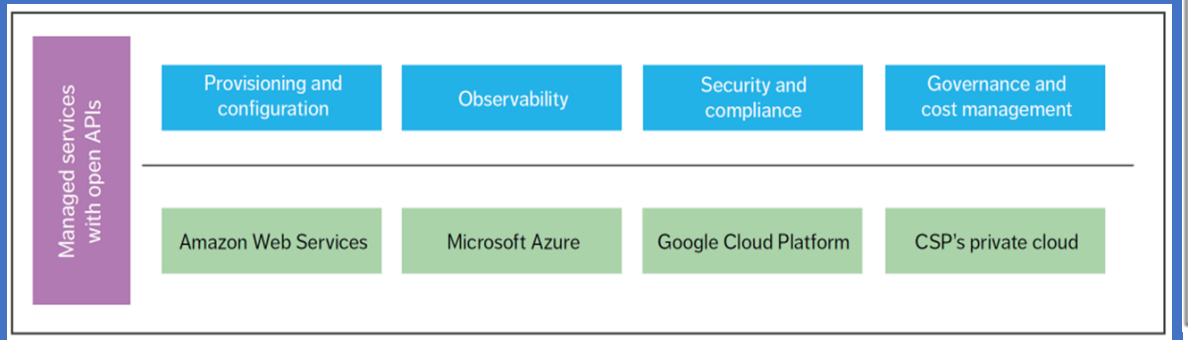


Figure 3: **Key Enablers for a Multi-Cloud Native Application**

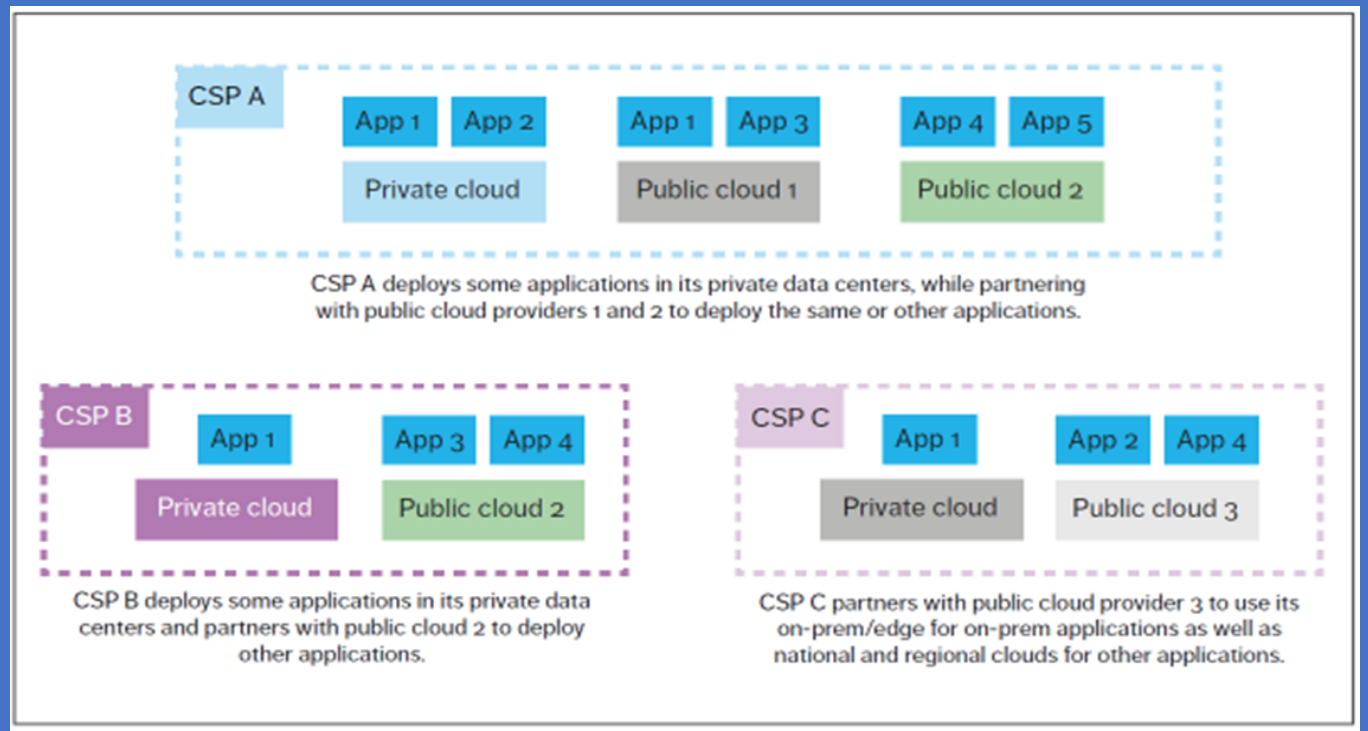


Figure 2: **Examples of Hybrid and Multi-Cloud Deployment Scenarios that Applications must be able to support**

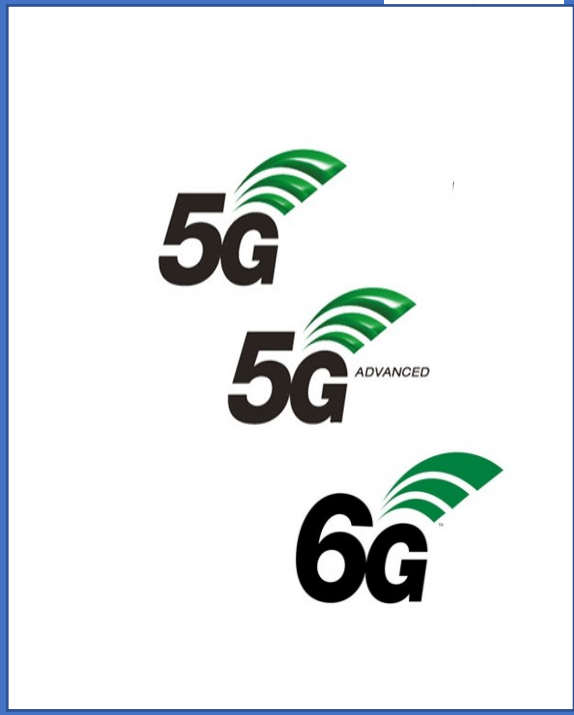
3. Mobile Networks to evolve from:

a Design that offers "Best-effort Services

to

a Design that offers Performance and User Experience Guarantees

=>



Capabilities related to e.g.:

When a *Multi-access* (**MA**) **PDU Session** is established, the Network may provide the UE with *Measurement Assistance Information* to enable the UE in determining which measurements shall be performed over both Accesses, as well as whether measurement reports need to be sent to the Network.

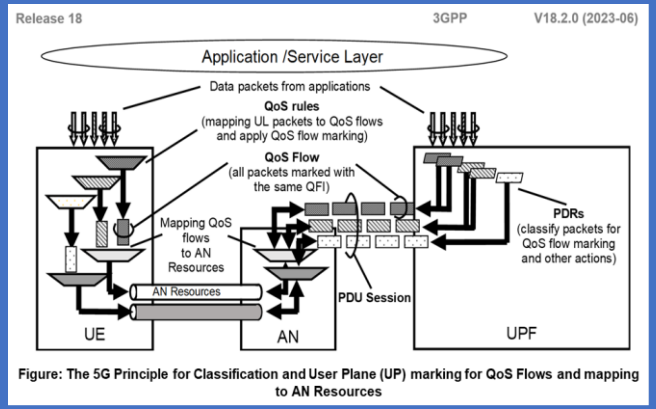


Figure: The 5G Principle for Classification and User Plane (UP) marking for QoS Flows and mapping to AN Resources

Measurement Assistance Information shall include the addressing information of a ***Performance Measurement Function*** **(PMF) in the UPF, the UE can send PMF protocol messages** incl.:

- Messages to allow for *Round Trip Time* (**RTT**) Measurements: the "*Smallest Delay*" steering mode is used or when either "*Priority-based*", "*Load-Balancing*" or "**Redundant**" steering mode is used with RTT threshold value being applied;
- Messages to allow for *Packet Loss Rate* (**PLR**) measurements, i.e. when steering mode is used either "*Priority-based*", "*Load-Balancing*" or "*Redundant*" steering mode is used with **PLR** threshold value being applied;

- Messages for reporting Access Availability/Un-availability by the UE to the UPF.
- Messages for sending **UE-assistance Data** to **UPF.**
- Messages for sending "*Suspend Traffic Duplication*" and "*Resume Traffic Duplication*" from **UPF** to **UE** to "suspend" or "resume" traffic duplication as defined in **5GS Architecture**.

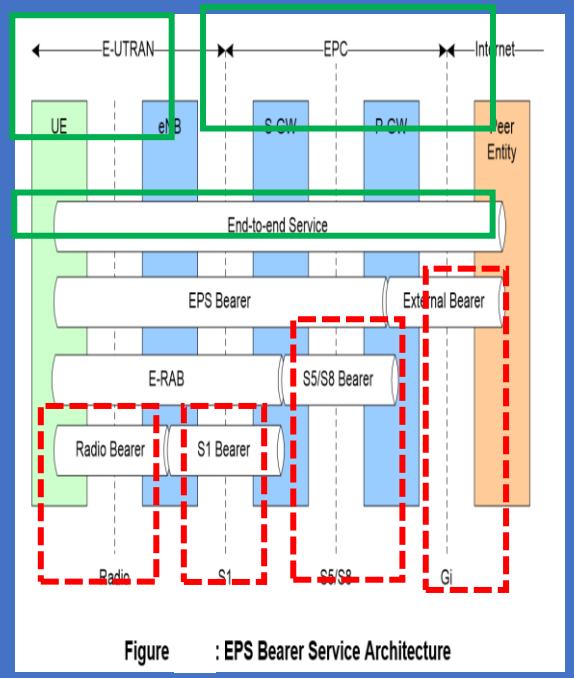


Figure : EPS Bearer Service Architecture

=>

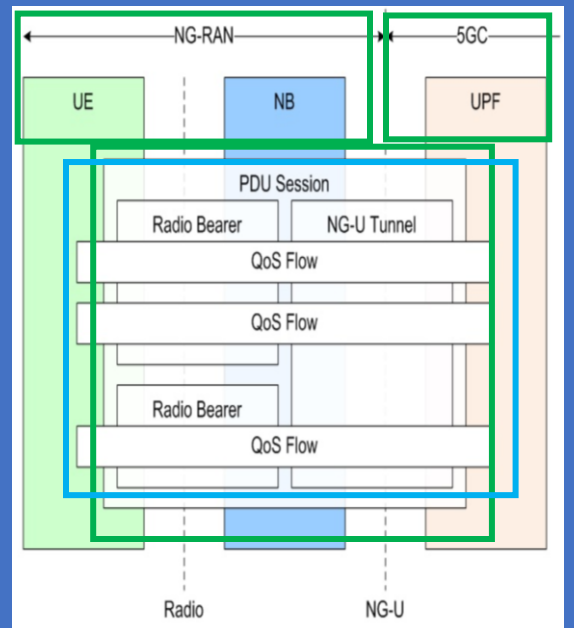


Figure: 5G CN NG-RAN Bearer Services QoS Architecture


THIS IS
THE END
OF THE BEGINNING
Remarks & Questions?