AI/ML selected Aspects and Opportunities
to
LF Edge Akraino
Technical Steering Committee (TSC)

Ike Alisson

2023 - 11- 23     Rev PA04
1. Introduction - scope of the presentation with an example from the Report on How is ChatGPT Behaviour Changing over time?
2. Overview of Andrew NG presentation on Opportunities in AI at Stanford
3. Prof. Yann LeCun on the difference between ML Programming and Computer Science Programming
4. Overview (selected parts) from Richard Feynman's lecture titled Can Machines think?
5. Supporting slides
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.
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 an 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.

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.

[Image: Andrew Ng: Opportunities in AI - 2023](https://www.youtube.com/watch?v=5p248yoa3oE)

Opportunities in AI

by Andrew Ng
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).
So this includes, a technique called supervised learning, things, and generative AI, which is a relatively new, exciting
AI is a collection of tools

Supervised Learning (labeling things)

<table>
<thead>
<tr>
<th>Input (A)</th>
<th>Output (B)</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>Spam? (0/1)</td>
<td>Spam filtering</td>
</tr>
<tr>
<td>Ad, user info</td>
<td>Click? (0/1)</td>
<td>Online advertising</td>
</tr>
<tr>
<td>Image, radar info</td>
<td>Position of other cars</td>
<td>Self-driving car</td>
</tr>
<tr>
<td>Ship route</td>
<td>Fuel consumed</td>
<td>Fuel optimization</td>
</tr>
<tr>
<td>Image of phone</td>
<td>Defect? (0/1)</td>
<td>Visual inspection</td>
</tr>
<tr>
<td>Restaurant reviews</td>
<td>Sentiment (pos/neg)</td>
<td>Reputation monitoring</td>
</tr>
</tbody>
</table>
### Example: Restaurant reviews sentiment tracking

<table>
<thead>
<tr>
<th>Input (A)</th>
<th>Output (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The pastrami sandwich was great! The meat was tender and perfectly balanced by the sauerkraut.</td>
<td>Positive</td>
</tr>
<tr>
<td>Service was slow and the food was so-so.</td>
<td>Negative</td>
</tr>
<tr>
<td>My favorite chicken curry. Yum!</td>
<td>Positive</td>
</tr>
</tbody>
</table>

#### 2010-2020: Large scale supervised learning

- **Large AI models**
- **Small AI models**

![Performance vs. Amount of data graph](image)
This decade: Generative AI

Text generation process

I love eating

bagels with cream cheese
my mother’s meatloaf
out with friends

prompt

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.

<table>
<thead>
<tr>
<th>Input (A)</th>
<th>Output (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>My favorite food is a bagel</td>
<td>bagel</td>
</tr>
<tr>
<td>My favorite food is a bagel</td>
<td>with</td>
</tr>
<tr>
<td>My favorite food is a bagel</td>
<td>cream</td>
</tr>
</tbody>
</table>

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.
Prompting is revolutionizing AI application development

Supervised learning

- **Get labeled data**: 1 month
- **Train AI model on data**: 3 months
- **Deploy (run) model**: 3 months

Image: Andrew Ng
Prompting is revolutionizing AI application development.

- Supervised learning
  - Get labeled data: 1 month
  - Train AI model on data: 3 months
  - Deploy (run) model: 3 months

So pretty realistic timeline for building a commercial grade.

Andrew Ng

Prompting is revolutionizing AI application development.

- Supervised learning
  - Get labeled data: 1 month
  - Train AI model on data: 3 months
  - Deploy (run) model: 3 months

Machine learning system is like 6 to 12 months.

Andrew Ng
Prompting is revolutionizing AI application development

Supervised learning
- Get labeled data: 1 month
- Train AI model on data: 3 months
- Deploy (run) model: 3 months

Prompt-based AI
- Specify prompt: minutes/hours
- Deploy model: hours/days
```python
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}], temperature=0)
       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

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

Supervised learning (Labeling things) can be worth more than $100 billion US a year.

Generative AI

where for a single company like Google
Value from AI technologies: Today → 3 years

Supervised learning (Labeling things) → building supervised learning applications.

And also, there are millions of developers.
Value from AI technologies: Today → 3 years

- Supervised learning (Labeling things)
- Generative AI
- Unsupervised learning
- Reinforcement Learning

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:
Accelerating Diffusion of Innovation: Maloney’s 16% Rule

Maloney’s 16% Rule: Once you have reached 16% adoption of any innovation, you must change your messaging and media strategy from one based on scarcity, to one based on social proof. In order to accelerate through the chasm to the tipping point.

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

Value

All potential AI projects, sorted in decreasing order of value

Stanford
So once you go to other industries,
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,

All potential AI projects, sorted in decreasing order of value

have been very difficult to execute on

because of the high cost of customization.
Why isn’t AI widely adopted yet? Customize (long tail) problem and low/no code tools

Low/no code tools are enabling the user to customize their AI system.

Users do this by providing prompts (or data) instead of writing code.

All potential AI projects, sorted in decreasing order of value
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.
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.
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.
Process for building startups

- Ideas
- Validate
  - Stage 1
    - 1 month
    - Market & technical validation by AI Fund team.

So double check, is this idea even technically feasible.
Process for building startups

- Ideas
  - Stage 1: Validate
    - 1 month
    - Market & technical validation by AI Fund team.

- Recruit CEO
  - Stage 2: 2 months
  - Recruit CEO to build with us (“Founder in residence”).

will go and recruit a CEO to work with us on the project.
Process for building startups

1 month
Market & technical validation by Al Fund team.

2 months
Recruit CEO to build with us (“Founder in residence”).

3 months
Deep customer and technical validation. Build prototype.

to work with them to build a prototype
Process for building startups

1. Ideas
   - 1 month
   - Market & technical validation by AI Fund team.

2. Validate
   - Stage 1
   - 2 months
   - Recruit CEO to build with us ("Founder in residence").

3. Recruit CEO
   - Stage 2
   - 3 months
   - Deep customer and technical validation. Build prototype.

4. Build
   - CEO
   - Stage 3

5. Pre-Seed Growth
   - Stage 4
   - ~12 months
   - $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, which then gives the company resources, have about a two thirds, 66% survival rate,
### Process for building startups

<table>
<thead>
<tr>
<th>Stage</th>
<th>Time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validate</td>
<td>1 month</td>
<td>Market &amp; technical validation by AI Fund team.</td>
</tr>
<tr>
<td>Recruit CEO</td>
<td>2 months</td>
<td>Recruit CEO to build with us (&quot;Founder in residence&quot;).</td>
</tr>
<tr>
<td>Build w CEO</td>
<td>3 months</td>
<td>Deep customer and technical validation. Build prototype.</td>
</tr>
<tr>
<td>Pre-Seed Growth</td>
<td>~12 months</td>
<td>$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.</td>
</tr>
<tr>
<td>Build w CEO</td>
<td>3 months</td>
<td>Deep customer and technical validation. Build prototype.</td>
</tr>
<tr>
<td>Pre-Seed Growth</td>
<td>~12 months</td>
<td>$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.</td>
</tr>
</tbody>
</table>

**BEARING.ai**

*Green shipping powered by AI*

- to hire an executive team, build the key team,
- get an MVP working, minimum viable product working,
Process for building startups

- **Ideas**
  - Stage 1
  - 1 month
  - Market & technical validation by AI Fund team.

- **Validate CEO**
  - Stage 2
  - 2 months
  - Recruit CEO to build with us ("Founder in residence").

- **Build with CEO**
  - Stage 3
  - 3 months
  - Deep customer and technical validation. Build prototype.

- **Pre-Seed Growth**
  - Stage 4
  - ~12 months
  - $1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

- **Seed, Growth, Scale**
  - Stage 5
  - Indefinite
  - ~$2-5M seed funding. Start up graduates and is well on its way.

and get some real customers.
Process for building startups

**Ideas**
1 month
Market & technical validation by AI Fund team.

**Validate Stage 1**
Recruit CEO to build with us ("Founder in residence").

**Recruit CEO Stage 2**
Deep customer and technical validation. Build prototype.

**Build w CEO Stage 3**
$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.

**Pre-Seed Growth Stage 4**
~$2-5M seed funding. Startup graduates and is well on its way.

**Seed, Growth, Scale Stage 5**
Indefinite

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.**
### Process for building startups: Concrete ideas

<table>
<thead>
<tr>
<th>Ideas</th>
<th>Validate</th>
<th>Recruit CEO</th>
<th>Build w CEO</th>
<th>Pre-Seed</th>
<th>Seed, Growth, Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Stage 2</td>
<td>Stage 3</td>
<td>Stage 4</td>
<td>Stage 5</td>
<td></td>
</tr>
</tbody>
</table>

- **Idea**: Like to engage only when there's a concrete idea.
- **Idea**: Don't rush to solutioning.
Explore a lot of alternatives before you do a solution.

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

**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

**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.

**thinking thing of exploring a lot of ideas and winnowing**

---

That finding those good ideas that someone has already

that turns out to be a much more efficient engine.
Risks and social impact

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.
- AI will disrupt many occupations.

(Credit: Eloundou et al., 2023)
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, ...).
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=SGSOCubyo2t 11:45 - 12:30
The Complete Mathematics of Neural Networks and Deep Learning

The Mathematics of Neural Networks

A rigorous introduction to the mathematics of networks and backpropagation.

Taught by Adam Dhalla
adamdhalla.com
adamdhalla@protonmail.com

Syllabus

Part I: Introduction
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
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
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

Part IV: Backpropagation
4.1 The Error of a Node
4.2 The Four Equations of Backpropagation
4.2.1 E1: Error of a
4.2.2 E2: Error of a1
4.2.3 E3: Cost w.r.t biases
4.2.4 E4: Cost w.r.t weights
4.2.5 Equation 4 vectorized
4.2.6 E5: Error of a2
4.2.7 E6: Error of a3
4.2.8 E7: Error of a4
4.2.9 E8: Error of a5
4.2.10 E9: Error of a6
4.2.11 E10: Error of a7
4.2.12 E11: Error of a8
4.2.13 E12: Error of a9
4.2.14 E13: Error of a10

Part V: Backpropagation together
4.4 The Backpropagation Algorithm
4.5 Looking forward

timestamps in description

https://www.youtube.com/watch?v=Ixl3nykKG9M
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?

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

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?
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.

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
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:

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

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

*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.**
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 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.
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:

<table>
<thead>
<tr>
<th>Category of Inference</th>
<th>Supervised learning</th>
<th>Semi-supervised learning</th>
<th>Unsupervised learning</th>
<th>Reinforcement learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (numeric), classification</td>
<td>Regression (numeric), classification</td>
<td>Association, Clustering</td>
<td>Reward-based behaviour</td>
<td></td>
</tr>
<tr>
<td>Labelling data (Note)</td>
<td>Labelling data (Note), and Unlabelled data</td>
<td>Unlabelled data</td>
<td>Non-pre-defined</td>
<td></td>
</tr>
</tbody>
</table>

**Note**: The labelled data means the input and output parameters are explicitly labelled for each training data example.

- **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.

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 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.

**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/Unavailability 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.
Remarks & Questions?