Foundations of Big Data:

In digital world , there are vast amount of devices available and people use them and leave traces of data that we can collect.

Big Data Definition by Ernst and Young:

Big data refers to the dynamic , large and disparate volumes of data being created by people , tools and machines. It requires new innovative and scalable technology to collect , host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to customers, risk, profit, performance, productivity management and enhanced shareholder value.

Big data does not have a fixed definition but multiple . The common part about all of them are

The 5 Vs of big data

* Velocity
* Volume
* Variety
* Varacity
* Value

Velocity - its the the speed at which the data is accumulated. Data is generated extremely fast in a process that never stops. Could platforms can store such data. Near-time or real-time streaming, local, cloud-based tech can process very fast.

Volume – It’s the scale of the data .

Drivers are the increase in data sources, higher resolution sensors, and scalable infra.

Variety – Diversity of the data.

Structure data fits nicely into rows and columns in relational databases

Unstructured data is not organized in a pre-defined structure . Examples: tweets, blog posts, pictures and numbers and data.

Data can be come different sources. Both internal and external to the organization.

Driver for variety are different sources like mobile, social, wearable, geo, video, and more.

Varacity – quality of the data and the conformity of facts and accuracy.

The attributes are consistency, completeness, integrity and ambiguity.

Drivers include cost and need for traceability

With large amount of data availability , we need to make sure if the data is accurate,

Value – Its our ability convert data into something valueable,

Value doesn’t mean only profit. Depending on the purpose, there can customer, employee or personal satisfaction. We spend time to understand Big Data is to derive value from the data.

Examples of V’s in Action:

Every 60 seconds we someone uploads videos to youtube becomes data. Imagine how much data is accumulated over the years

The world population is 7 billion and the vast majority are using digital devices, latop, computers, wearable devices.They all generate data (approximately 2.5 quintillion bytes) every day.

Text, pics, vids, audio, health data, and many other types which are generated from different devices which are connected to the internet are called internet of things

80% of the data is unstructured and we must come up with a solution gets us Reliable and accurate insights. The data must be categorized , analyzed, and visualized.

Data Scientist today derive insights from big data and cope with the challenges that these massive datasets present.

The scale of the data means convensional data analsis tools might not be enough,.

We have other tools that makes use of distributed computing that helps us overcome this problem.

Apache Spark, Hadoop and its ecosystem provides ways to extract , load and analyze and process these massive datasets across distributed compute resources in providing new insights and knowledge.

This give organizations more ways to connect with customerts and enrich the services they offer,

What is Hadoop?

In big data , Data is sliced into small pieces and distributed to multiple computer(tens of thousands of them) then the same program is sent to this cluster. The program runs on the little piece of file and the results is sent back.

Then the results are sorted and then once again distributed to another process.

The first process is called mapper process. and the second one is called a reducer process.

Even though its a simple idea, we can do a lot of operations with very very very large dataset.

The advantage in big data is that they scale linearly. If you have twice the servers, then we have twice the performance.

Hadoop was originally a copy of the Google’s Big data Architeccture developed by Yahoo engineer Doug Cutting.

There are a lot of companies that have some footprint in big data hadoop.

Most of the components of data science existed for a decades. But are all coming together with some nuances.

At the bottom of Datascience we have Probability and stats, algebra , linear algebra, programming and databases.

Now we have the comutational power to apply new techniques. Like we can apply Ml where we can take very large datasets and instead of taking samples , we have take very large data to find patterns.

Once some of these ML techniques started emerging, we are able to analyze very large datasets.

What we see is the combination of the traditional with decision sciences.

How Big Data is Driving Digital Transformation?

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benifitsof new technologies.

This digital change integrates digital technology into all areas of an organization resulting in fundamental changes to how it operates and delivers value to customers.

It is an organizational and cultural change driven by Data Science, and especially Big Data.

The availability of vast data and competitive advantage of analyzing brings, triggered digital transformations throughout many industries.

Netflix changed from DVDs to streaming.

NBA teams data gathered from cameras to analyze the effective plays.

Lufthansa analyzed customer data to improve the experience.

Digital Transformation is not just duplicating the existing processes in digital form;

The in-depth analysis of how the business operates helps organization discover how to improve their proesses and operation and integrate DS in their workflow.

Most organiztions discover that the digital transformation needs fundametal changes in their approach to towards data, employees, and customers and it will affect their organizational culture.

As it affect everything and everyone in the organization, the deccisions are taken by high level executives.

The support of the CEO is crucial for digital transformation. Same goes for CIO and CDO. They also need support from people who control Budgetting , Personnel , and day- to day priorities.

The digital transformation needs new mindset. But is very important for originasational success.

# Data Science Skills & Big Data

Phd

Applied Physics

Economics

Programming

Unix , Linux

Jupyter notebook

regex

dbms

pandas

# Data Scientists at New York University

Most Know programming

Some have Ms

Some are MBA

How big data Cameup:

Started by Google.

Larry Page and Sergey Brin wanted to know how to Solve their PageRank Allgorithm

Basically to crawl all the webpages.

There was nothing like that.

Theu developed the Big data Cluters and then Hadoop was a copy of the same.

New Analytical techniques and algorithms were introduced to handle these huge clusters.

https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTAxRU4tU2tpbGxzTmV0d29yay9sYWJzL01vZHVsZSUyMDIvRGF0YV9NaW5pbmcubWQiLCJ0b29sX3R5cGUiOiJpbnN0cnVjdGlvbmFsLWxhYiIsImFkbWluIjpmYWxzZSwiaWF0IjoxNjg0MzQxMDE4fQ.qyRXZBzGK-yLOSBh8ofvpfl9htbLl4yiYPAJY9QBdZM

Deep Learning Vs Machine Learning:

What is the Difference?

Big Data - Data that is so big and so quickly built so they defy traditional analysis methods

With distributed comuting and new techiniques and algorithms these vast data sets can be analyzed.

Bid Data 5 Vs

Data mining is the process of automatically searching and analyzing data to discover new patterns.

We need to preprocess the data to preare for transformation.

Then insights and patterns are mined using visualization , Machine Learning and Statistical Models

Machine Learning:

A sunset of AI. Uses compter algorithms to analyze data and make intelligent decisions based on what it has learnt without explicitly being programmed.

Machine Learning algorithms are traned with large amounts of datasets and they learn from examples

They dont follow rules-based algorithms.

It allows computer to make preditiction and with the new results make itself better for future usage.

Deep Learning:  
A specialized subset if ML that uses layered neural networks to simulate human decision-making.

Deep Learning Algorithms Can label and categorize information and identify patterns,

It enables AI systems to continuously learn on the job and improve the quality and accuracu of the results by determining whether decisions were correct.

Aritificial neural Networks:

Takes inspiration from biological neural networks . Although they work quite a bit differently.

A neural network in AI is a small set of compute units call neurons that take in data and transforms it depending on the algorithm used.

there can be a number of layers.

Each layer can have differnt algorithm that focusses on differnt stuff.

Example: A newral network can scan a pixelated drawing and figure out which character it is by smoothing out the corners and comparing with differnt shapes.

neural networks become more efficient as the data increases.

But for the other algorithms the efficiency increases only untill a certain extent. After that, the efficiency more or less stays the same

AI Vs Data Science:

Data Science:

Process of extracting knowledge and insights from large volumes of disparate data.

It an iterdeciplinary field involving math, stats, visualization. ML and more

Help us to gain insigits and dive businesses.

Data Science can use many AI tech niques to derive insights from data.

for example it could use ML algorithms and deep learning model to gain insights and draw inferences.

There is some interaction between AI and data science.

One is not the subset of the other.

Data Science is a broad term that encompasses the entire data processing methodology while AI includes includes everything that allows a computer to learn how to solve problemss and make intelligent decisions.

Both can use Big Data.

Neural Networks and Deep Learning

An computer’s attempt to mimic neurons and how the brain functions

Takes inputs into processing nodes , does some transformation and aggregate them into something and then maybe go to another level of nodes.

Finally some output wil be given.

How neural Network works?

Uses computer program that tries to mimic human brain like a network of neurons.

They vere computationally intensive and they wen out of phase .

Then deeplearning cameup wich is prettymuch newral networks on steroids. Containing the neural network in layers which can work on very large datasets. It also uses a lot of computing power to solve them.

These needs GPUs.

We have computing power available to do this.

Lot of servers now has GPUs and TPUs

Usecases of Deep Learning:

Speech Recognistion.

recognizing images and object detection

genrates speech after recognizing speech.

How can One get stated with neural networks?

Linear algebra, matrix,

packages are available in python

Needs high power computational ability . Basically servers with GPUs

Applicationsof Machine Learning:

Recommender system

Classification

Cluster analysis

market basket analysis

Predictive analytics -- Uses Decision Trees, bayesian Analysis, naive Bayes calssification.

Need to know how, when and where the techiniques not how they work. Precision versus recall, problem of oversampling and overfitting.

Can be use in fintect

recommendations.

Retail and Banking

Fraud detection in realtime.

Regressio reading:

<https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTAxRU4tU2tpbGxzTmV0d29yay9sYWJzL01vZHVsZSUyMDIvUmVhZGluZ19SZWdyZXNzaW9uLm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTY4Mzc0NDk2OX0.HDVV6NTc9EyZroSFfHz7t5S_NPHA629jNiYkNYhJQSM>

https://docs.google.com/presentation/d/1JQIwqNHL0UPZFoMJ37JsPva1FmAC2P5N/edit?usp=drive\_web&ouid=112477175007979309082&rtpof=true