Introduction to Data Science and Analytics

Solomon Teferra Abate

@

SIS, AAU

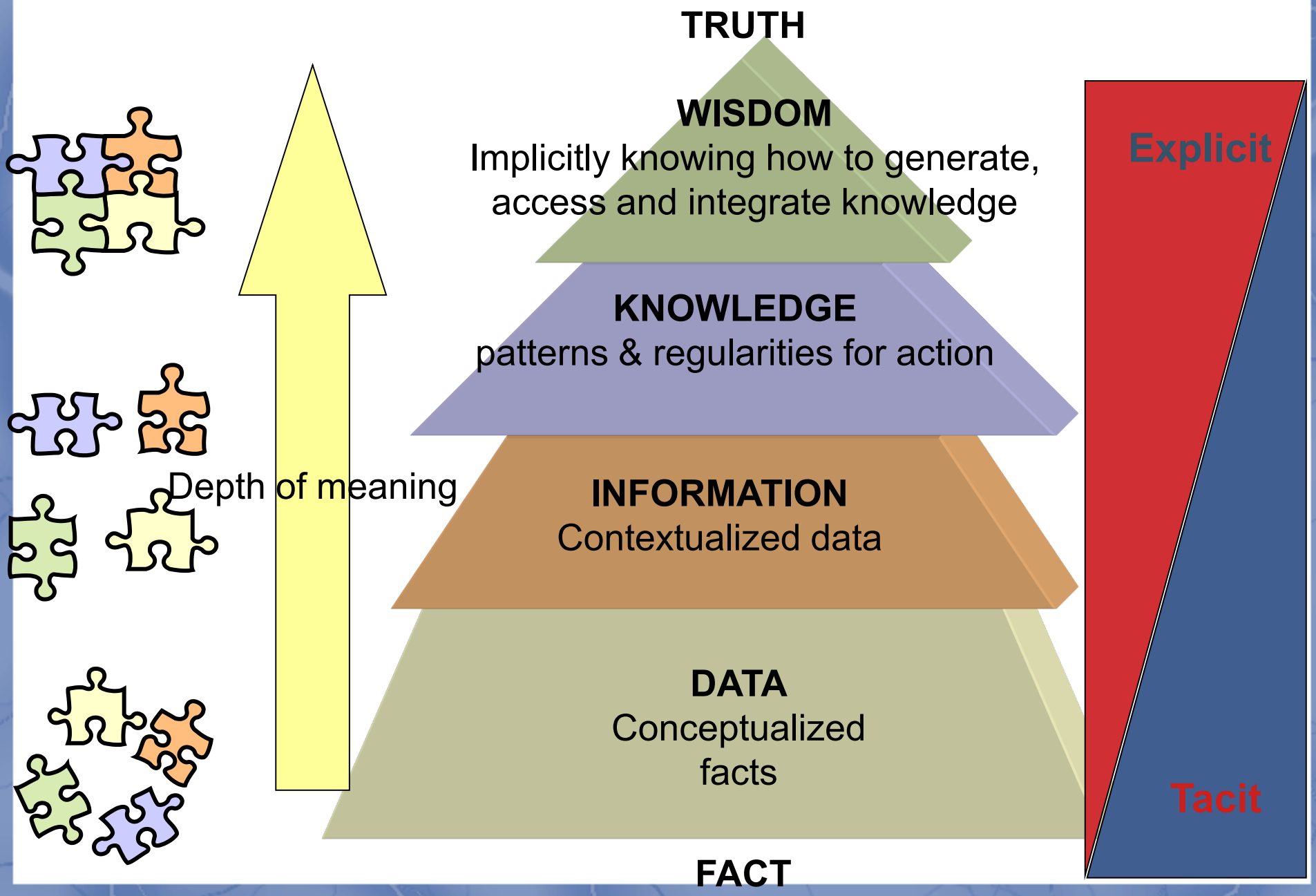
Introduction to Data Science

- Overview of the data Revolution
- The demand for data Professionals
- Defining the Data Science Discipline
- Data Science and other Data Related Disciplines
- Knowledge areas of the Data Science Discipline.

Introduction to Data Science

- Overview of the data Revolution
- The demand for data Professionals
- Defining the Data Science Discipline
- Data Science and other Data Related Disciplines
- Knowledge areas of the Data Science Discipline.

Data, Information, Knowledge, Wisdom & Truth

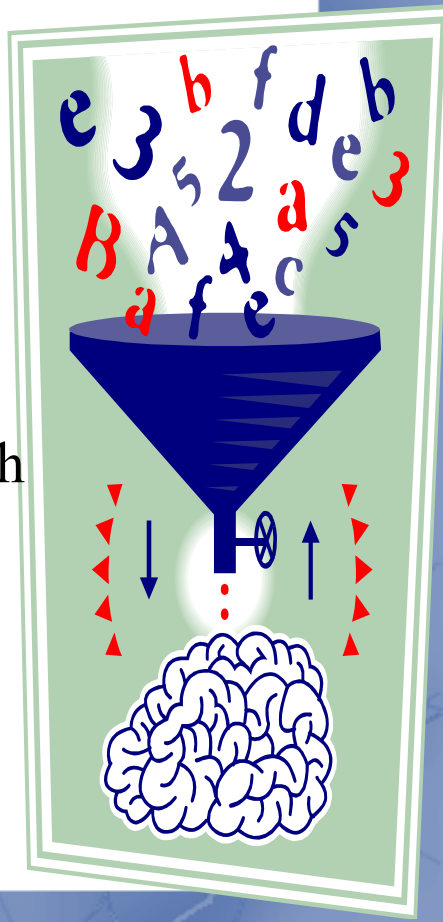


Data, Information, Knowledge, Wisdom & Truth

➔What is Data and Information? Are they different from Knowledge? Wisdom? Truth?

- fact != data != information != knowledge != wisdom != truth

- **Data:** Conceptualizing a set of discrete facts about events
  - No meaning attached to it as a result of which it may have multiple meaning
  - Example: what does "Alex" mean?
- **Information:** Aggregation of data as per the context that makes decision making easier.
  - ✓ Meaning is attached and contextualized
  - ✓ Answers questions: what, who, when, where
- **Knowledge**: includes facts about the real world entities and the relationship between them. It is an Understanding gained through experience
  - ✓ Answer 'how' question
- **Wisdom**: embodies principles, insight and moral by integrating knowledge
  - ✓ Answer 'why' question
- **Truth**: making



Two types of data

- Data is divided into **data at rest** and **data in motion**.
- Data at rest:
  - This refers to data that has been collected from various sources and is then analyzed after the event occurs.
  - The point where the data is analyzed and the point where action is taken on it occur at two separate times.
- Data in motion:
  - The collection process for data in motion is similar to that of data at rest; however, the difference lies in the analytics.
  - In this case, the analytics occur in real-time as the event happens.

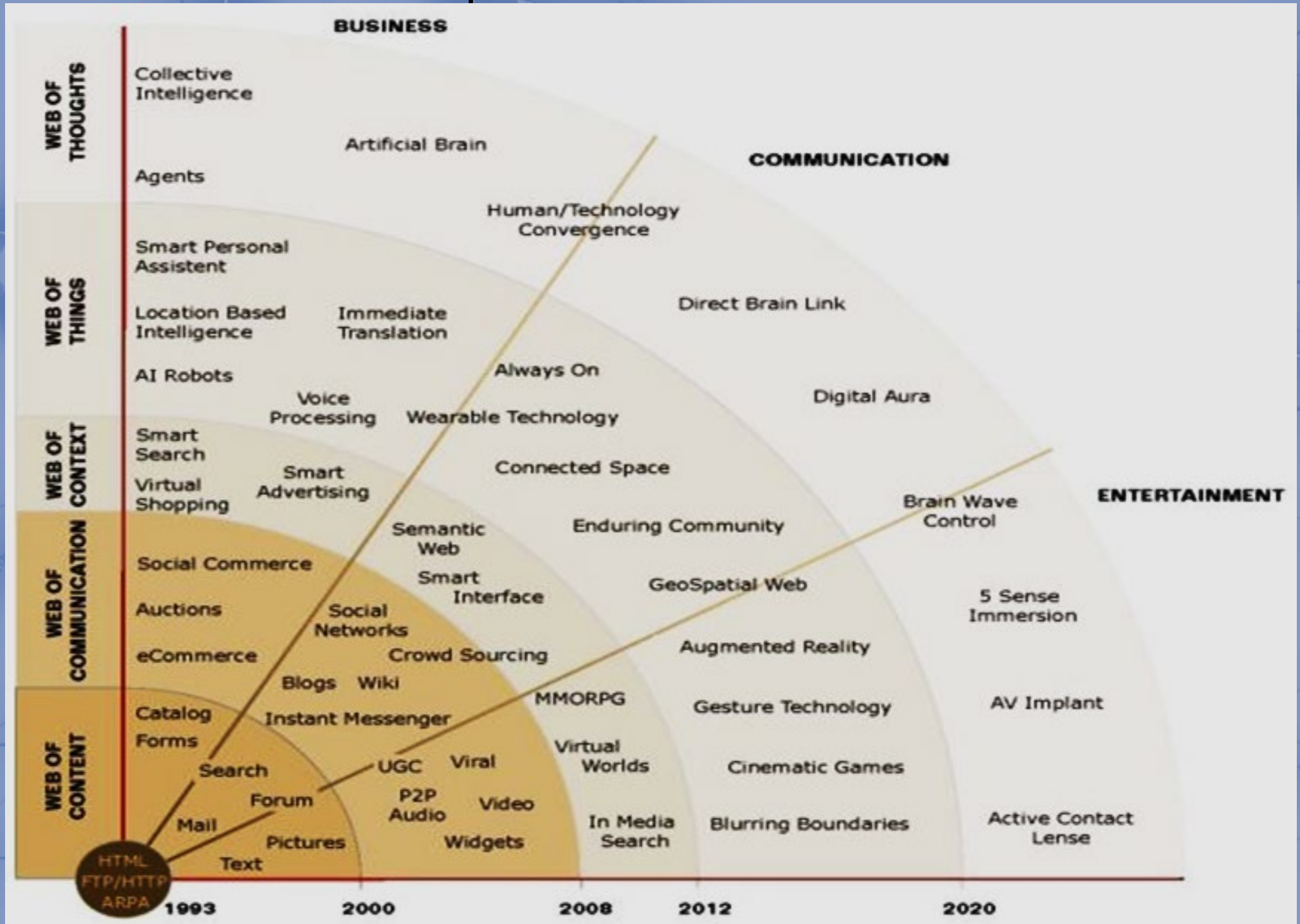
Types of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Multimedia data (Image, Audio, Video)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

Data Revolution

- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data allowing correlations to be found
  - to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.
- 2.5 quintillion (or 10^{18}) bytes of data are generated every day!
  - Data come from many quarters: Social media sites, Sensors, Digital photos, **Business transactions** and Location-based data
- Business transactions outside of the web???

The Web Expansion: Web 0.0 to Web 5.0



The Web Expansion: Web 0.0 to Web 5.0

- **Web 0.0 – The Development of the Web**
  - The World Wide Web as we know it was invented in 1989 at CERN by Tim Berners-Lee, a British scientist who reimagined the user-side functionality of the early Internet.
- **Web 1.0 – The Read-Only Web**
  - The World Wide Web was introduced to the public in 1990/1991. In those early public internet days, the World Wide Web was a place to seek and find things. Hence it is web 1.0 is the “Read-Only Web,”. The Web was not largely interactive at this time. Instead, it was a place where we were largely content consumers.
  - By 1999, there were approximately 3 million web sites (and yes, these were the days where “website”—single word—wasn’t yet the norm). And because of this massive amount of information online, the Read-Only Web brought about the explosion of “web browsers”—think Mosaic, Netscape Navigator, Opera, Internet Explorer—and, of course “search engines.” Hello, Yahoo, founded in 1995, and Google, founded in 1998, among many others.
- **Web 2.0 – The Social (Read-Write) Web**
  - Where Web 1.0 connected people with information, Web 2.0 connected people with people. Internet users became participants in the Web, interacting and bringing their own value, rather than just acting as content consumers.
  - In 1999, LiveJournal and Blogger launched as blogging platforms. The crowd-sourced encyclopedia Wikipedia launched in 2001. MySpace launched in 2003. Facebook launched in 2004. YouTube launched in 2005. The list of examples of this Web 2.0 shift is too long to name them all, but the short story of this era was that the Web was no longer just a collection of things to read. It was now a way to connect, and to connect in more ways than ever before.

The Web Expansion: Web 0.0 to Web 5.0

- **Web 3.0 – The Semantic (Read-Write-Execute) Web**
  - The shift from Web 1.0 to Web 2.0 was huge, but the shift from 2.0 to 3.0 is even more of a paradigm shift. With the rise of data, not only can people consume information and connect with each other, but applications can connect with other applications independently to execute functions on their own. “Big data” is a popular term for a reason.
  - The knowledge stored on the Web is now better connected than ever before, and this information has become enriched. The IoT, AI, ML, Augmented Reality, and Virtual Reality are no longer terms of the future. Web applications can interpret information for humans, creating efficiencies, analyses, and possibilities like never before.
  - However in the era of Web 3.0, while the applications can connect data and execute functions with that data, these applications cannot yet provide context to data, understand relevance, or make more complex decisions in regard to this data.
- **Web 4.0 – The Mobile Web**
  - The definitions of Web 1.0, 2.0, and 3.0 are fairly standardized; however, it’s when we move beyond this point, different sources suggest different divisions. For our purposes, we will define Web. 4.0 as the “Mobile Web,” and yes, one could argue that in many ways it is occurring simultaneously with Web 3.0.
  - Once the Web was a place where one might “surf,” sitting back and relaxing. This concept is the epitome of Web 1.0 or 2.0. However, now the Web is always in action. Chances are, it’s in your pocket or otherwise on your person. Today, there are more mobile connections than there are people. We are a mobile, connected society, and from smart watches to smart phones, the Mobile Web era is changing the dynamic of how we interact with the Web.

- **Web 5.0 – The Intelligent / Emotional (Symbiotic) Web**
  - The rise of virtual assistants that predict your needs from your behaviors, without many cues, is a hint at the Intelligent Web to come. Web 5.0 will see applications able to interpret information on more complex levels, emotionally as well as logically. This is the Web that acts in true symbiosis with daily life, without a thought, organically intertwined with what we do.
  - AI enables computers to communicate like a person, but the technology that enables them to think, reason, and respond on their own, in a human way, is not as far away as you might guess.
  - Web 5.0 will also focus on the individual, perhaps allowing a website to convey a different experience for each different person. It could perceive the emotions of an individual and respond appropriately, and it could detect subtleties that enable more powerful interactions.
  - Right now, Web 5.0 is a vision of the not-so-far-off future, but time will tell what this new technology will truly bring.

Introduction to Data Science

- Overview of the data Revolution
- The demand for data Scientist/Professionals
- Defining the Data Science Discipline
- Data Science and other Data Related Disciplines
- Knowledge areas of the Data Science Discipline.

The need for data science

- No matter how extremely efficient your algorithm is, they can often be beaten simply by having more data.

The need for data scientists

- There is a high demand for Data Science and data analytics
  - databases, warehousing, data architectures
  - data analytics – statistics, data mining, machine learning

- Data science is the ability to store large amounts of data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it for decision making and problem solving in such dynamic world

- It supports “Business Intelligence”
  - for smart decision-making and problem solving
  - for predicting potential market, potential product, potential customers
  - need data for identifying risks, opportunities, conducting “what-if” analyses

The need for data scientists

➔Data Science career is the hottest profession in the new era

- ✓ According to the US Bureau of Labor Statistics, the number of jobs requiring Data Science skills is expected to grow by **27.9 percent by 2026**.
- ✓ The data science market size is expected to grow from USD 95.3.9 billion in 2021 to USD 322.9 billion by 2026.
- ✓ According to latest reports, data engineering is one of the fastest-growing domains in technology, with over **88.3 percent growth** in job postings.
- ✓ The average salaries in data science and machine learning jobs are surging sky-high with a data scientist getting USD 176,213, and a data engineer taking home USD 166,992.
- ✓ According to latest reports by Forbes, 5 percent of all data analytics and data science jobs can be found in IT, finance, insurance and related professional services.
- ✓ About 80 percent of the firms across the globe are investing a large part of their earnings into creating a skillful data analytics division, thus hiring the smartest of people in the industry domain

The need for data scientists

➔ Why???

- ✓ **Abundance of data:** Organizations around the world are finding it a big challenge to handle the enormous amounts of data at their disposal, and an even bigger challenge is how to manage the future datasets that will be exponentially larger.
- ✓ **Talent deficit:** Finding an equipped talent in the data science domain is tough. People adept at understanding and using data to drive business benefits are a rare gem to find. The demand for data analysts and scientists is like gushing water that is unstoppable and the supply is like a trickle. A 2021 report by McKinsey stated that the U.S. alone has got a shortage of more than 190,000 data science professionals. Since then, the demand has grown manifolds.
- ✓ **Diverse and long skillset required:** Being a data science professional requires so much than having an ordinary knowledge of programming, or coding. You must be proficient in applying tools such as Spark, Hadoop, and NoSQL. Besides, you must be well trained in machine learning, programming, and statistical modeling. It's really hard to find all these skills in one person.
- ✓ **No entry for professionals with zero knowledge of related subjects:** Entry is almost banned for professionals or students with no connection with computer science, engineering, mathematics/statistics, and general science. Data Science is a multidisciplinary field and requires expertise in either one of the above-mentioned fields.
- ✓ Handsome salary: There's no doubt about that! The pay is simply great! But so is the work that goes into being a data science professional in your organization.

Introduction to Data Science

- Overview of the data Revolution
- The demand for data Professionals
- Defining the Data Science Discipline
- Data Science and other Data Related Disciplines
- Knowledge areas of the Data Science Discipline.

Data Science

- **Data science**, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to **extract information or knowledge or insights** from **data** in various forms, either structured or unstructured, **similar to data mining**.
- Data science is a field of study and practice that involves the collection, storage, and processing of data in order to derive important insights to solve a problem or understand a phenomenon. Such data may be generated by humans (surveys, logs, etc.) or machines (weather data, road vision, etc.), and could be in different formats (text, audio, video, augmented or virtual reality, etc.).

Data Science

- Just as natural science focuses on understanding the characteristics and laws that govern natural phenomena, data scientists are interested in investigating the characteristics of data looking for patterns that reveal how people and society can benefit from data.

Data Science

- Data science is a concept to unify statistics, data analysis and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.
- Jim Gray imagined data science as a **"fourth paradigm"** of science (empirical, theoretical, computational and now data-driven)

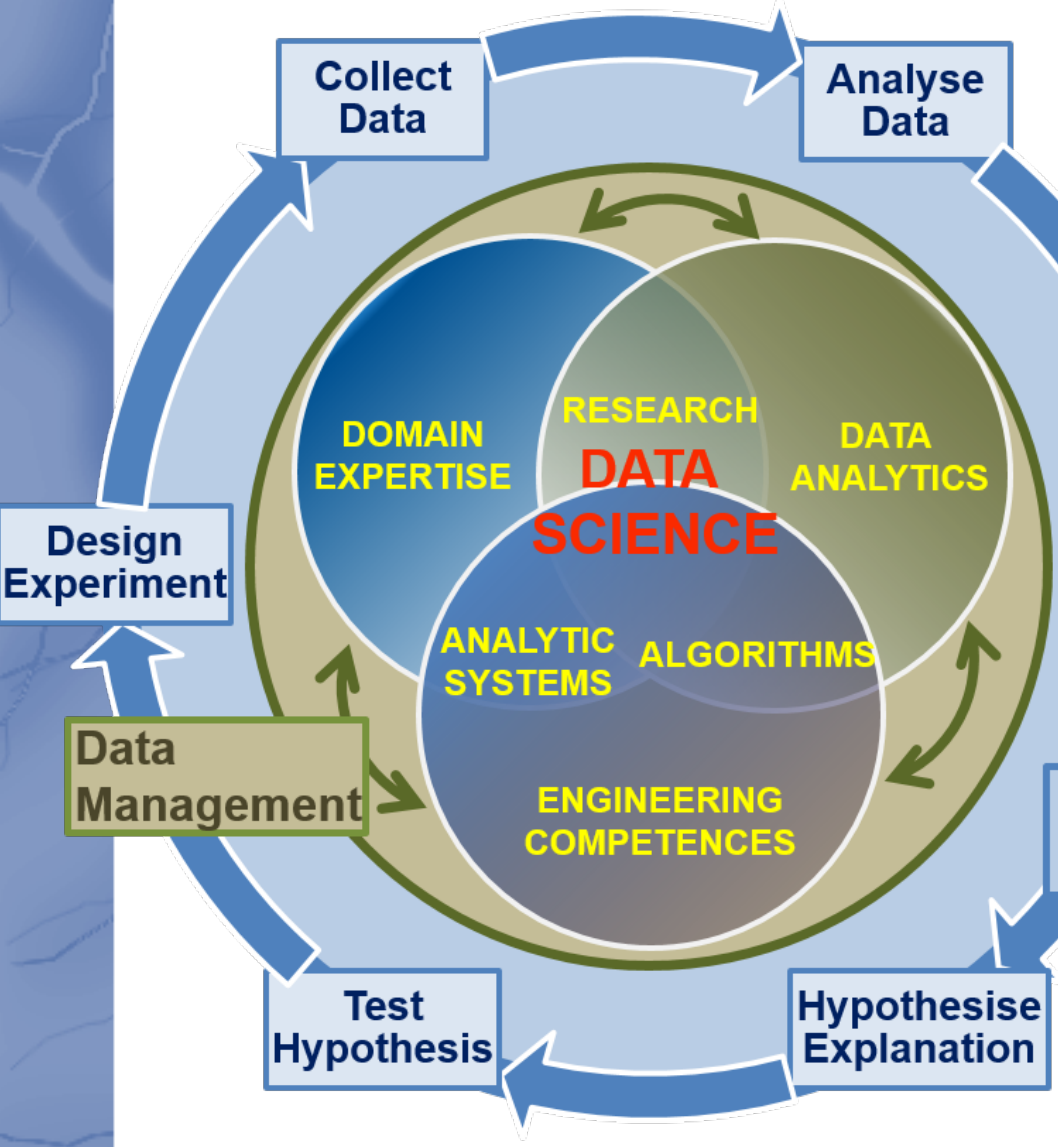
Introduction to Data Science

- Overview of the data Revolution
- The demand for data Professionals
- Defining the Data Science Discipline
- Data Science and other Data Related Disciplines
- Knowledge areas of the Data Science Discipline.

Data Science Competence Groups - Research

Data Science Competence: 5 areas/groups

- Data **Analytics**
- Data Science **Engineering**
- Domain **Expertise**
- Data Management
- Scientific Methods



Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesize Explanation
- Test Hypothesis

Business Process Management (for biz competences)

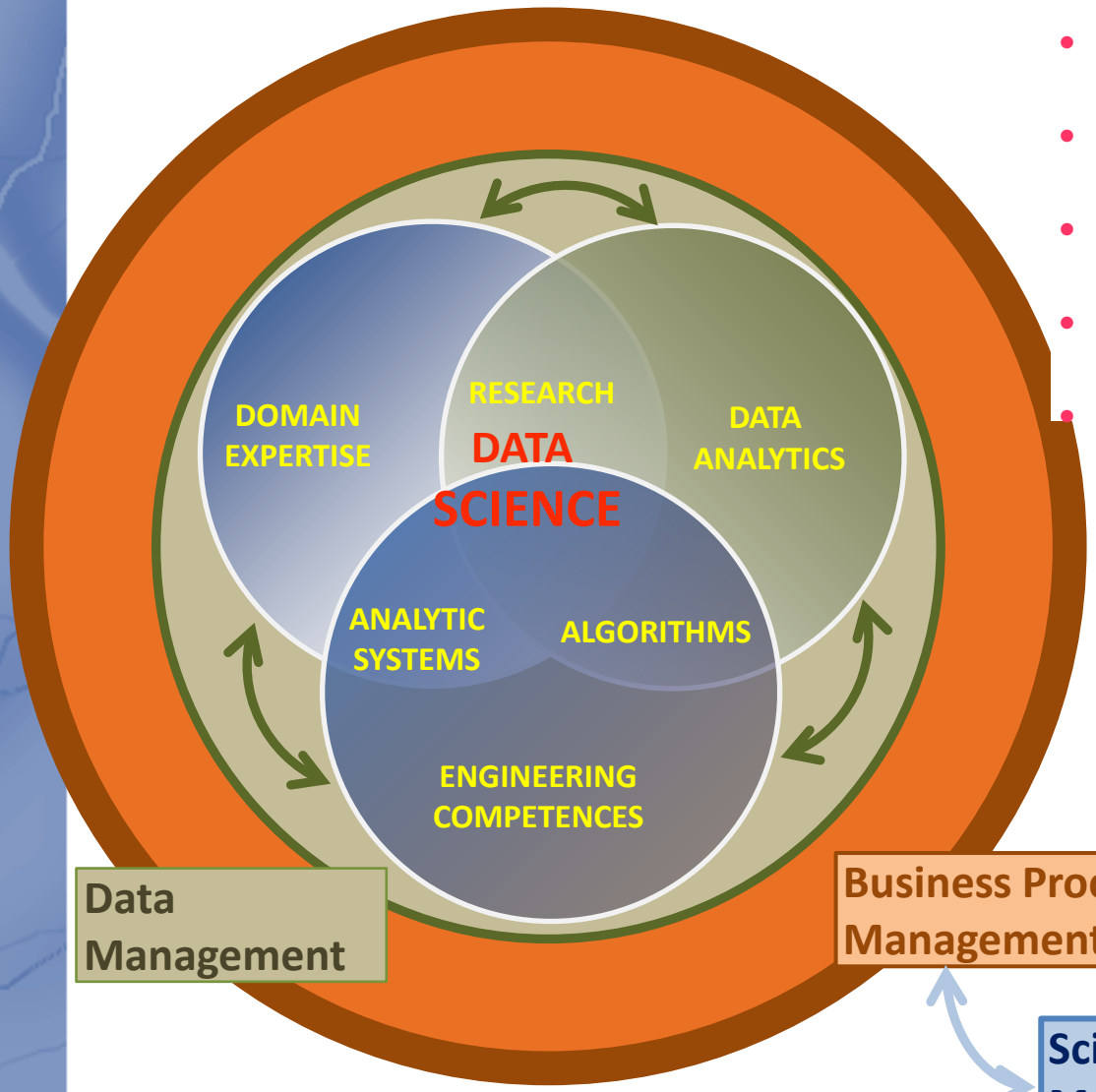
Business Process Operations

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimize & Re-design

Data Science Competences Groups – Business

Data Science Competence: 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods



Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesize Explanation
- Test Hypothesis

Business Process Operations

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimize & Re-design

Identified Data Science Competence Groups

Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
	Collect and manage different source of data	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
	Visualise complex and variable data.	Develop algorithms to analyse multiple source of data	Influence the development of organizational objectives	Analyse multiple data sources for marketing purposes
		Prototype new data analytics applications		Analyse customer data to identify/optimise customer relations actions

Identified Data Science Competence Groups

Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
	Collect and manage different source of data	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
	Visualise complex and variable data.	Develop algorithms to analyse multiple source of data	Influence the development of organizational objectives	Analyse multiple data sources for marketing purposes
		Prototype new data analytics application		Analyse customer data to identify/optimise customer

Identified Data Science Competence Groups

Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
	Collect and manage different source of data	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
	Visualise complex and variable data.	Develop algorithms to analyse multiple source of data	Influence the development of organizational objectives	Analyse multiple data sources for marketing purposes
		Prototype new data analytics applications		Analyse customer data to

Identified Data Science Competence Groups

Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
	Collect and manage different source of data	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
	Visualise	Develop algorithms to		Analyse multiple

Identified Data Science Competence Groups

Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Scientific/Research Methods (DSRM)	DS Domain Knowledge (including Business Apps)
Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
	Collect and manage different source of data	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organizational roles
	Visualise complex and variable data.	Develop algorithms to analyse multiple source of data	Influence the development of organizational objectives	Analyse multiple data sources for marketing purposes
		Prototype new data analytics applications		Analyse customer data to identify/optimize customer relations actions