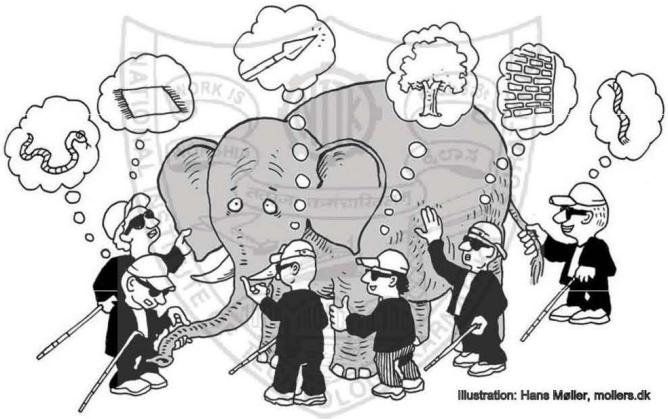
Feb – Jun 2023
IT258

Data Science – Introduction and Basics

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

What is Data Science?



What is Data Science?

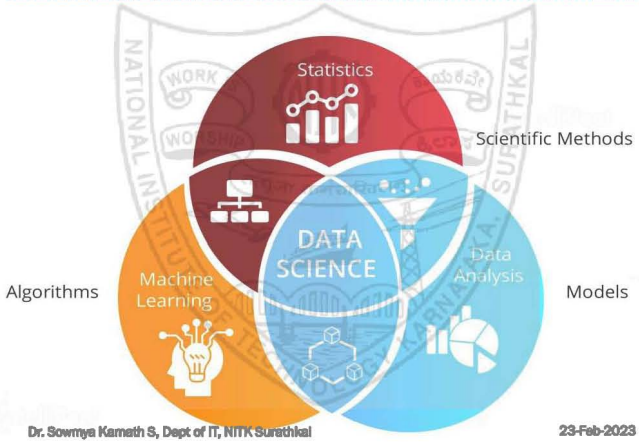
"the application of **computational** and **statistical** techniques to address or gain **insight** into some problem in the real world"

What is Data Science?

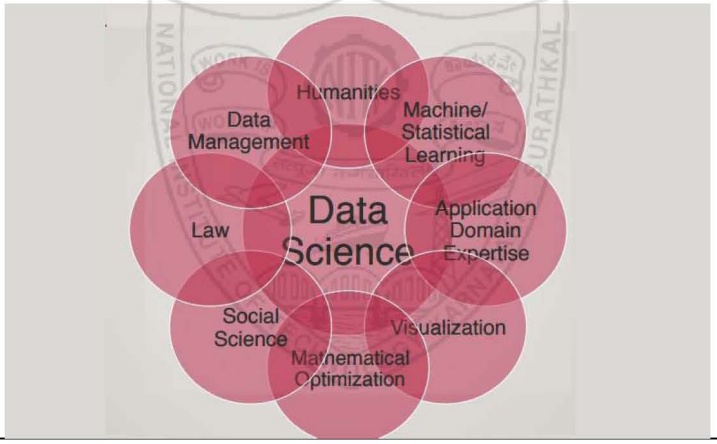
"is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract **knowledge** or insights **from data** in various forms, either **structured or unstructured**."

Data science = statistics +
data processing +
data modeling +
visualization +
machine learning +
analytics +
big data + …

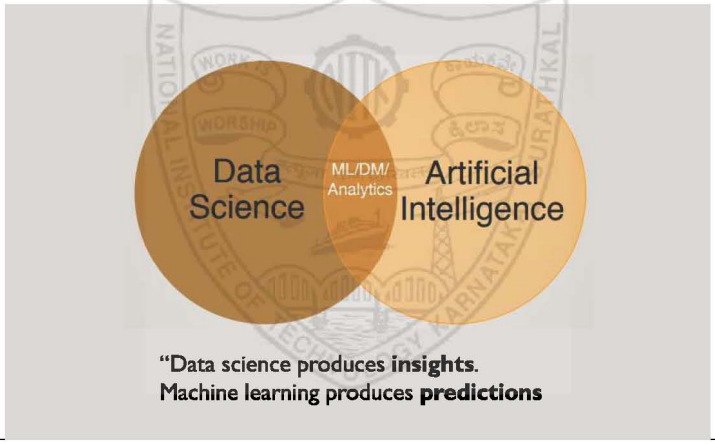
Data Science



The Data Science Confluence



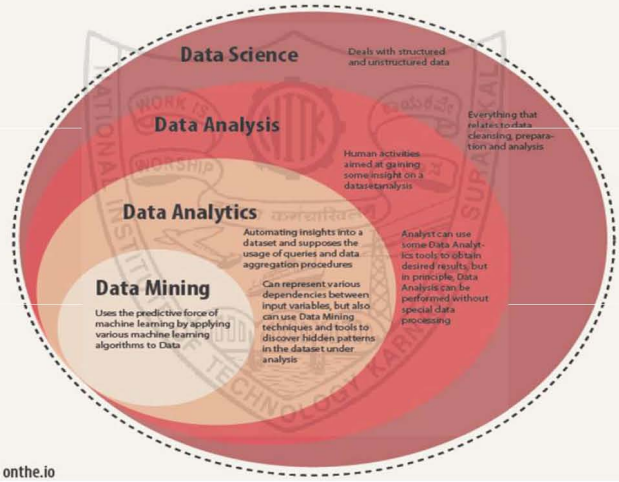
Data Science and AI



Data related disciplines..

- **Data Mining**
- **Data Analytics**
- **Data Analysis**
- **Data Science**

What is the difference between Data Science, Data Analysis, Data Analytics and Data Mining ?



Source: onthe.io

Data related disciplines..

- **Data Mining**

Given lots of data,

Discover patterns and models that are:

Valid: some findings on new data with some certainty
Useful: should be possible to act on the item
Unexpected: non-obvious to the system
Understandable: humans should be able to interpret the pattern

Data related disciplines..

- **Data Mining**
  - **Descriptive methods**
    - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
  - **Predictive methods**
    - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Data related disciplines..

- **Data Analytics**
  - applying a computational/algorithmic process to derive meaningful, previously unknown insights into the various aspects of the dataset.
  - E.g. Predictive analytics, preventive analytics
- **Data Analysis**
  - the process of compiling and analyzing data to support decision making.
  - E.g. For exploratory data analysis, data mining etc..

Data related disciplines..

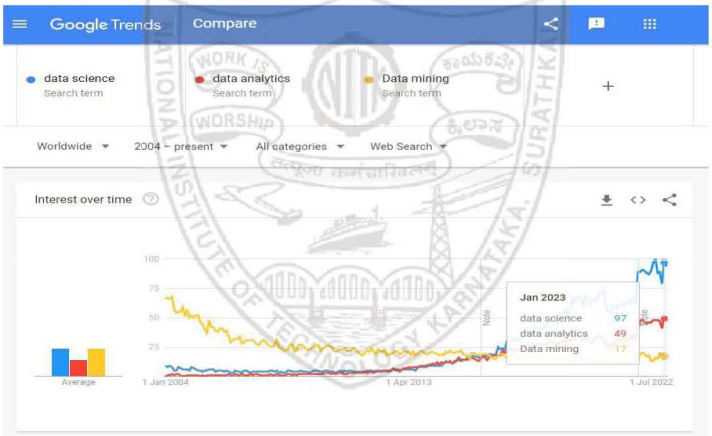
- **Data Science**

  the ability to look at things 'differently' using a combination of
  - mathematics,
  - statistics,
  - programming,
  - in the context of the problem being solved,

  +

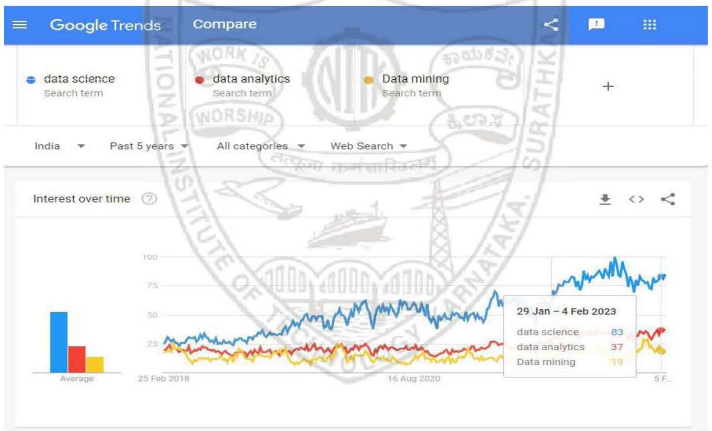
  the activities of cleansing, preparing and aligning the data.

Comparing Interest in Data-related disciplines
(Worldwide, All time)



Comparing Interest in Data-related disciplines
(Worldwide, Last 5 years)

Comparing Interest in Data-related disciplines
(India, last 5 years)



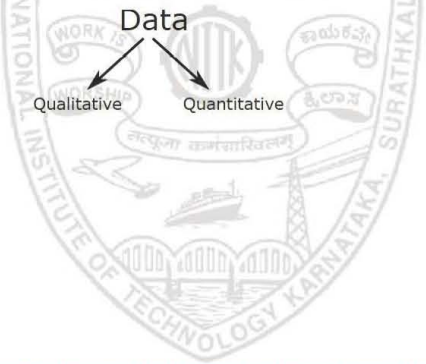
The most fundamental question...

What is data???

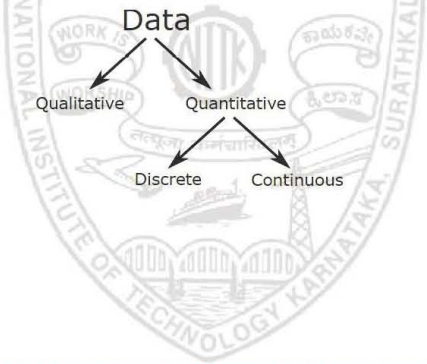
Trying to define data ...

- "facts and statistics collected together for reference or analysis."
- "things known or assumed as facts, forming the basis for reasoning or calculation."
- "used to describe things by assigning a value to them."

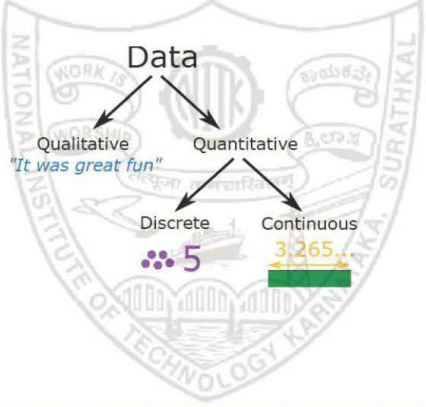
What is data?



What is data?



What is data?

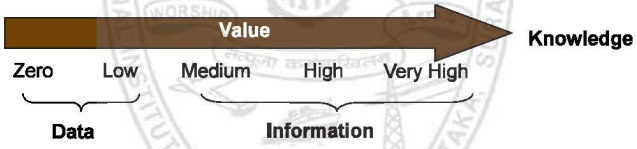


Data v/s. Information

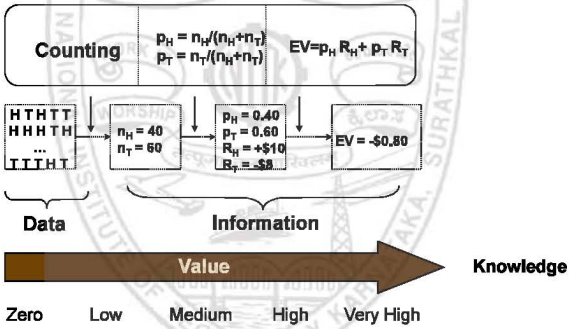
- **Data**: a set of facts.
- **Information**: data that are processed to useful; provides answers to "who", "what", "where", and "when" questions
- **Knowledge**: application of data and information; answers "how" questions.
- **Understanding**: an appreciation of "why"
- **Wisdom**: evaluated understanding.

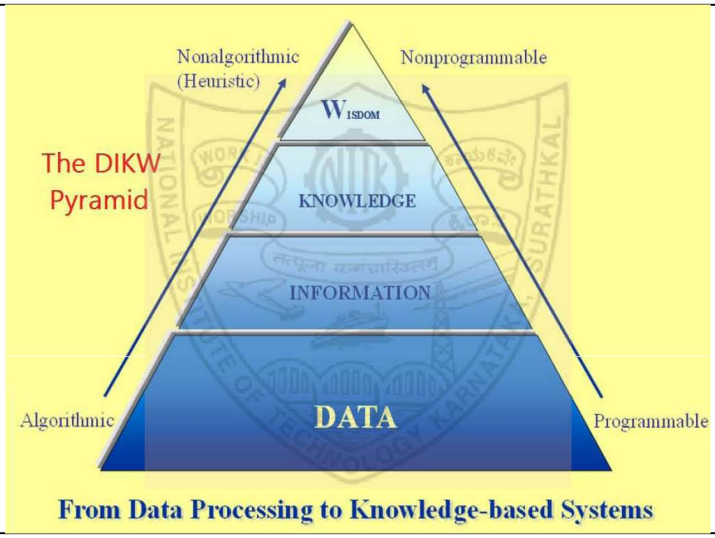
Ref: Ackoff, "From Data to Wisdom", 1989

Relationship between D, I and K



DIKW in data analysis...

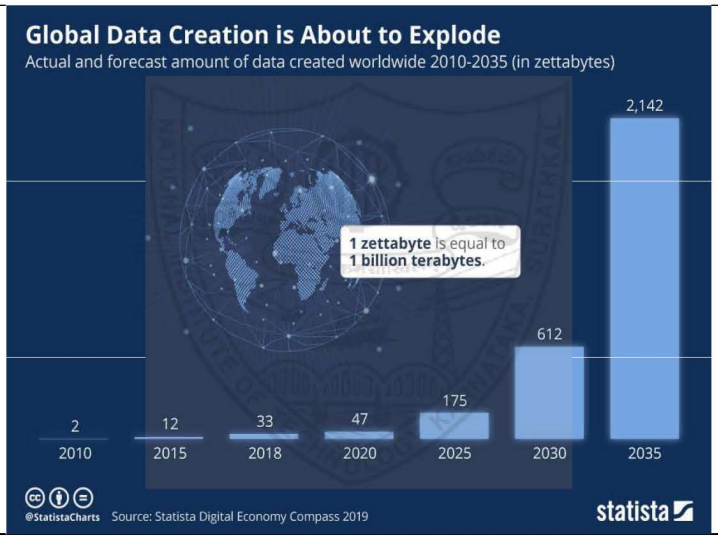




From Data to Knowledge and beyond...

"Datafication"

- "Datafication"
  - taking all typical aspects of life and turning them into data.
  - Term coined by Kenneth Cukier and Victor Mayer-Schöenberger in 2013.

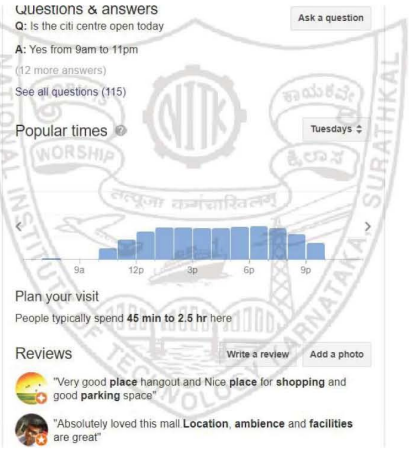


Datafication (contd.)

- Data examples →
  - What you like/enjoy has been turned into a stream of your "likes"
  - What you read and share is considered your opinion on certain topics (particularly in politics)
  - Datafying cities, workplaces, traffic systems, processes, ….

Datafying cities

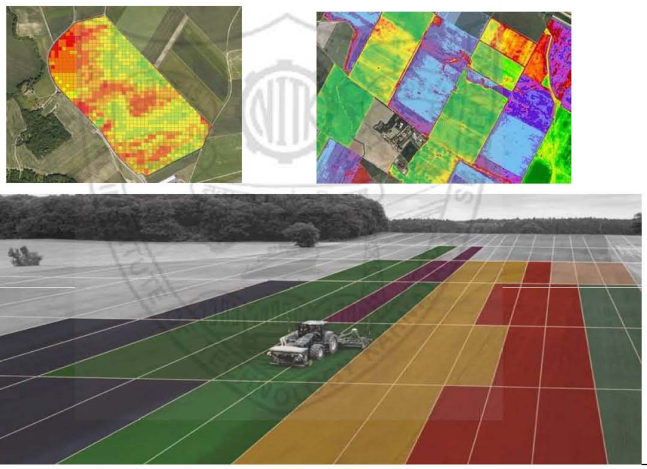
Datafying public spaces



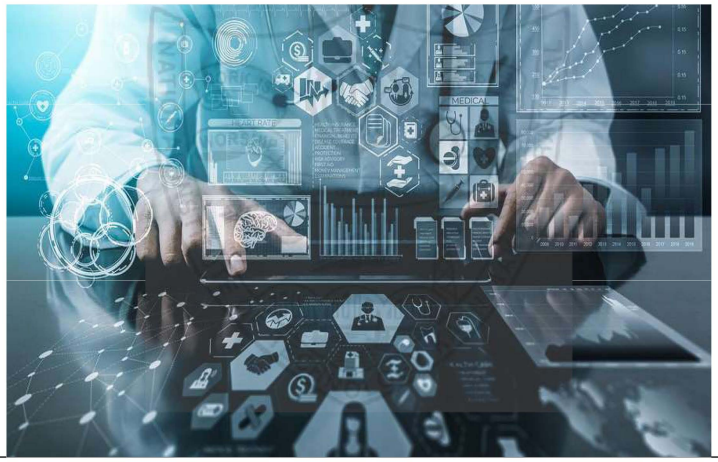
Datafying systems

Datafying processes

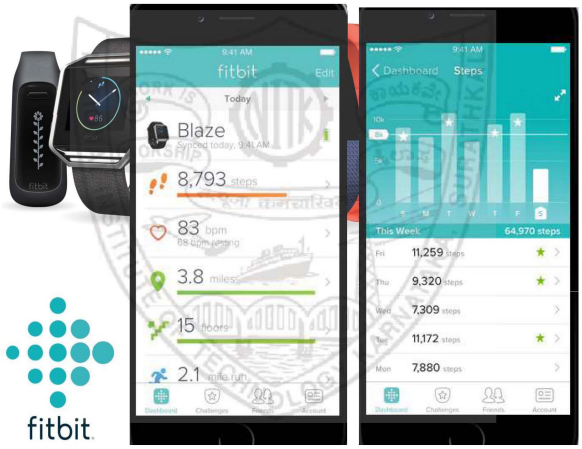
Datafying processes



Datafying domains



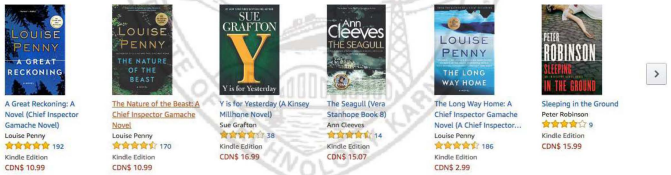
Datafying activities



amazon

Customers who bought this item also bought

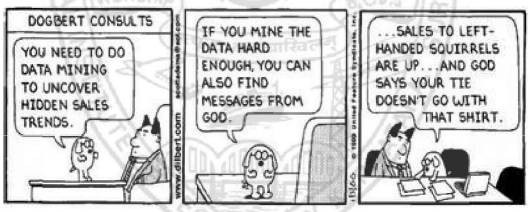
Page 1 of 17



So, is DATA all we need?

Is Data all we need?

- **Meaningfulness** of Analytical Answers is important!
  - Significant risk of "**discovering**" patterns that are **meaningless**.



Is Data all we need?

- **Bonferroni's principle:**

  "… if you look for interesting patterns in more places than your amount of data will support, you are bound to find crap!"

Is Data all we need?

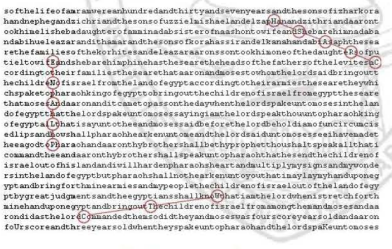
- **Bonferroni's principle:**
  - a statistical method for accounting for random events.
  - Process:
    - determine the number of expected random events of interest in the dataset
    - if the observed number is significantly greater than this number, the chances of *any* observations providing useful insight are almost nonexistent.

Example: Michael Drosnin's bestseller book

Proposed that purportedly meaningful messages can be extracted using the *Equidistant Letter Sequence* (ELS) method, where, letters are selected based on a starting point and counting every nth letter based on a given 'skip number' in a given direction.

Example: Michael Drosnin's bestseller book

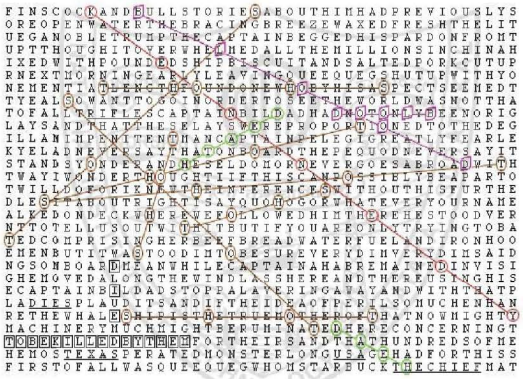
Proposed that purportedly meaningful messages can be extracted using the *Equidistant Letter Sequence* (ELS) method, where, letters are selected based on a starting point and counting every nth letter based on a given 'skip number' in a given direction.



"WTC plane crash"

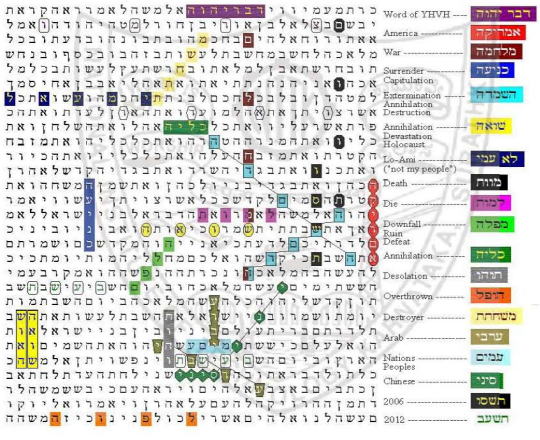
"Kennedy shot"

Reaction to Michael Drosnin's hidden codes book



Ref: Assassinations Foretold in Moby Dick! https://users.cecs.anu.edu.au/~bdm/dilugim/moby.html

Reaction to Michael Drosnin's hidden codes book

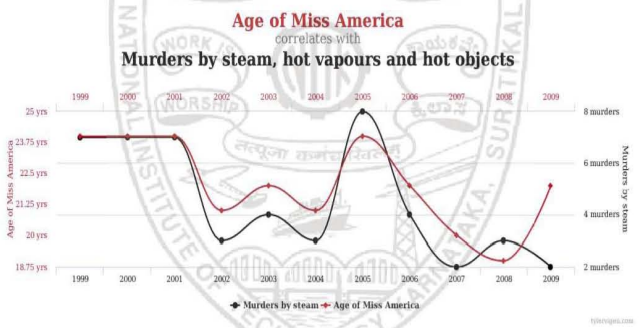


Reaction to Michael Drosnin's hidden codes book

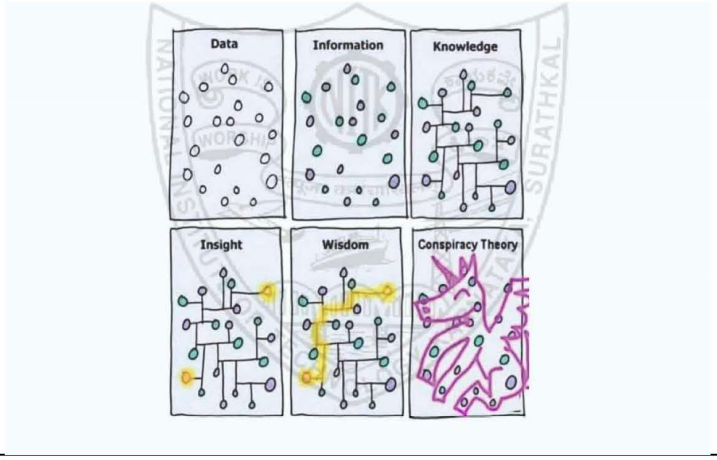
The Demise of Drosnin: A **shocking discovery** has been made deep within the text of **Moby Dick**.

Ref: https://users.cecs.anu.edu.au/~bdm/dilugim/drosnin.html

Bonferroni's Principle - example



DIKW exploited to the extreme ?



Rhine Paradox

Joseph Banks Rhine (September 29, 1895 – February 20, 1980), usually known as **J. B. Rhine**, was an American botanist who founded parapsychology as a branch of psychology, founding the parapsychology lab at Duke University, the *Journal of Parapsychology*, the Foundation for Research on the Nature of Man, and the Parapsychological Association. Rhine wrote the books *Extrasensory Perception* and *Parapsychology: Frontier Science of the Mind*.

Parapsychology is the study of alleged psychic phenomena (extrasensory perception, telepathy, precognition, clairvoyance, psychokinesis (also called telekinesis), and psychometry) and other paranormal claims, for example, those related to near-death experiences, synchronicity, apparitional experiences, etc.[1]

Joseph Banks Rhine; https://en.wikipedia.org/wiki/Joseph_Banks_Rhine

Rhine Paradox

- Joseph Rhine's Extra-sensory Perception experiments.
  - a great example of how not to conduct scientific research!

Total Information Awareness

- Four states of Information Awareness -
  - KNOW WHAT YOU KNOW
  - DON'T KNOW WHAT YOU KNOW
  - KNOW WHAT YOU DON'T KNOW
  - DON'T KNOW WHAT YOU DON'T KNOW

Total Information Awareness

- Four states of Information Awareness -
  - KNOW WHAT YOU KNOW—*Information you **know** and **have**.*
  - DON'T KNOW WHAT YOU KNOW—*Information you **don't know** you **have***
  - KNOW WHAT YOU DON'T KNOW—*Information **you know you need** but **don't have.***
  - DON'T KNOW WHAT YOU DON'T KNOW—*Information you **don't even know exists**.*

Total Information Awareness

- TIA - More of an desire than a reality
  - Impossible to achieve as we cannot know everything, in every context.

Total Information Awareness

- TIA - More of an desire than a reality
  - Impossible to achieve as we cannot know everything, in every context.
  - If you are looking for so many vague connections, then you will find things that were bogus.
    - Violation of innocents' privacy or worse!

Is Data all we need?

- It is easy to find "interesting" patterns where none exist!
- How should we judge whether a "pattern" is interesting?
- When should we worry about falsely labeling patterns "interesting"? (E.g. Google mistranslates a sentence vs. incorrect cancer diagnosis. . .)
- ....

Is data all we need?

- Major Challenges –
  - How to store data?
  - How to process this data?
  - How to analyse this data?
  - How to model this data?
  - .....
  - How to transform into **interesting** and **meaningful actionable** insights / knowledge?

Data + __?__ = value + insights = profits/power/upper hand!

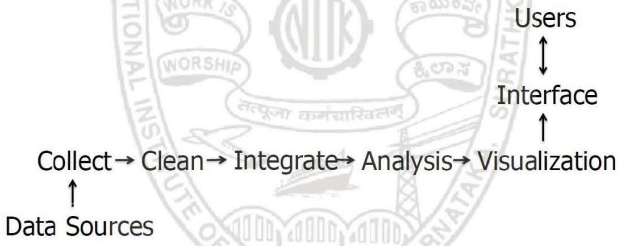
Goal of Data Science

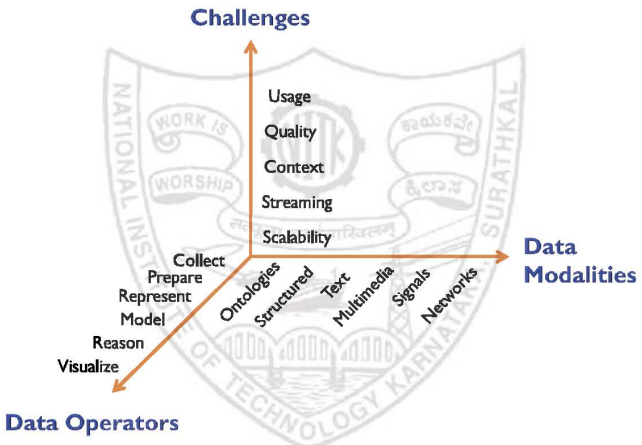
Turn raw data into data products

Data to Data products....

- Transaction Databases → *Fraud Detection, purchase trend analysis ..*
- Wireless Sensor Data → *Smart Home, automated monitoring ..*
- Text Data, Social Media Data → *Opinion Mining, Product Reviews, Consumer Satisfaction*
- Software Log Data → *Automatic Troubleshooting, Bug tracking ..*
- Genotype and Phenotype Data → *Disease prediction, new treatments, patient modeling …*

The Life of Data





Data Scientist/Analyst - Responsibilities

- Data Management
  - Data collection, storage, cleaning, filtering, integration …
- Large-scale Parallel Data Processing
  - Parallel computing, incremental processing…
- Statistics and Machine Learning
  - Data modeling, inference, prediction, pattern recognition …
- Interface and Data Visualization
  - HCI design, visualization, story-telling…

More reading...

- Bell, Gordon, Tony Hey, and Alex Szalay. "*Beyond the data deluge.*" *Science* 323.5919 (2009): 1297-1298.
- O'Neil, Cathy, and Rachel Schutt. *Doing data science: Straight talk from the frontline.* " O'Reilly Media, Inc.", 2013.
- Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons.