



Data Science – Introduction and Basics

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

23-Feb-2023

What is Data Science?

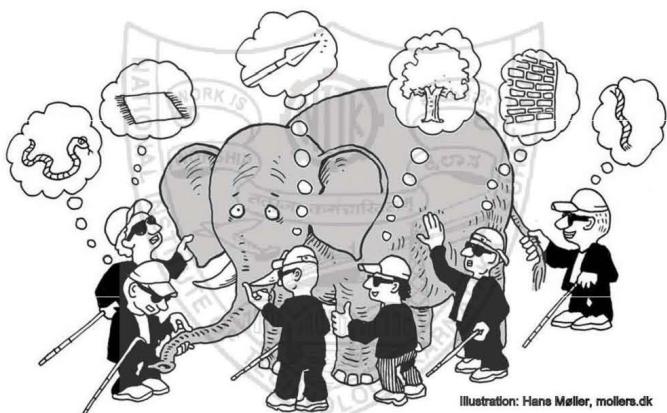


Illustration: Hans Moller, mollers.dk

What is Data Science?

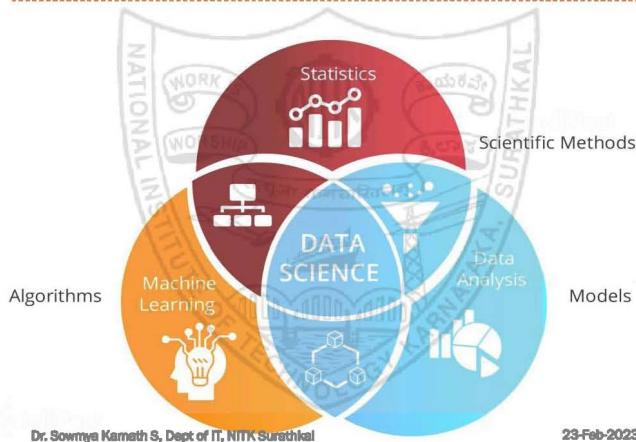
"the application of computational and statistical techniques to address or gain insight into some problem in the real world"

What is Data Science?

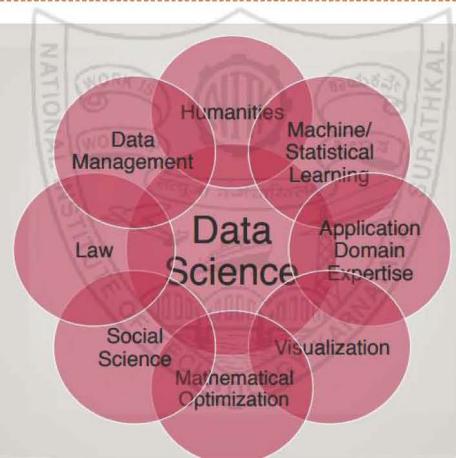
"is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured."

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

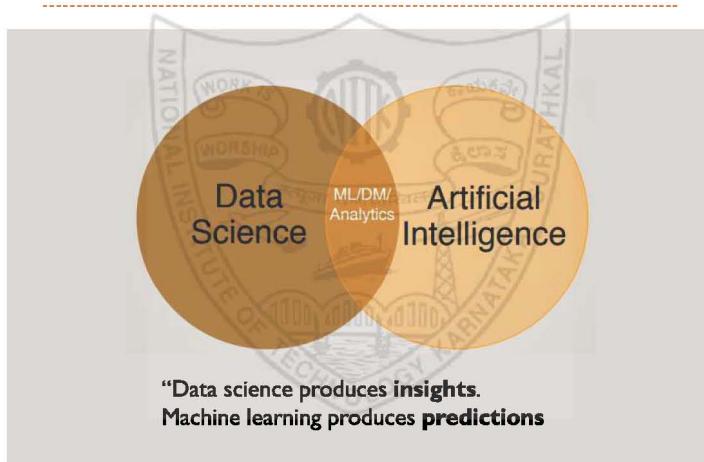
Data Science



The Data Science Confluence



Data Science and AI



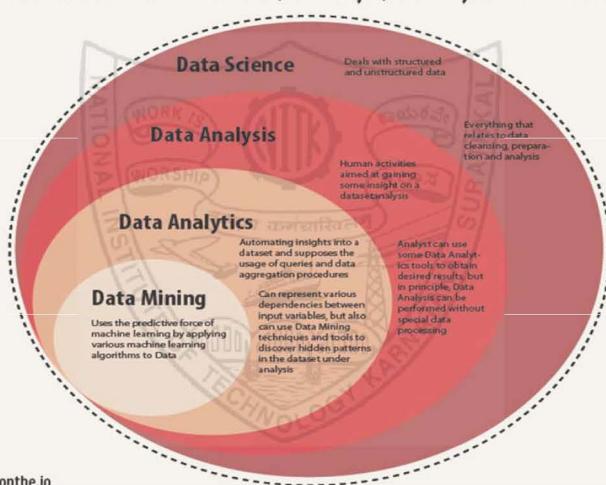
Data related disciplines..

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

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

23-Feb-2023

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



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

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

23-Feb-2023

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

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

23-Feb-2023

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

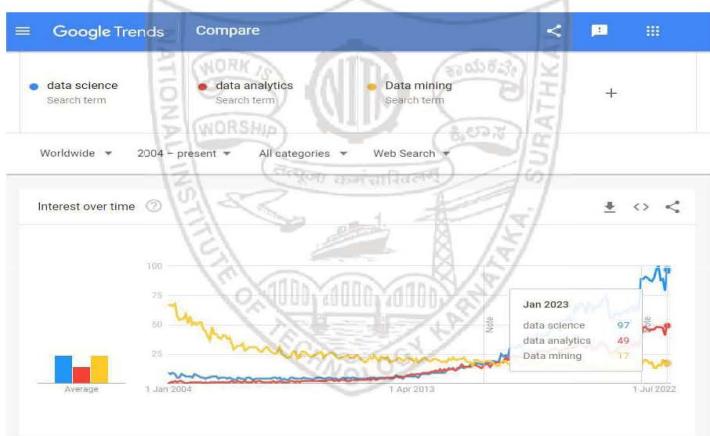
the ability to look at things 'differently' using a combination of

- mathematics,
- statistics,
- programming,
- in the context of the problem being solved,
+
the activities of cleansing, preparing and aligning the data.

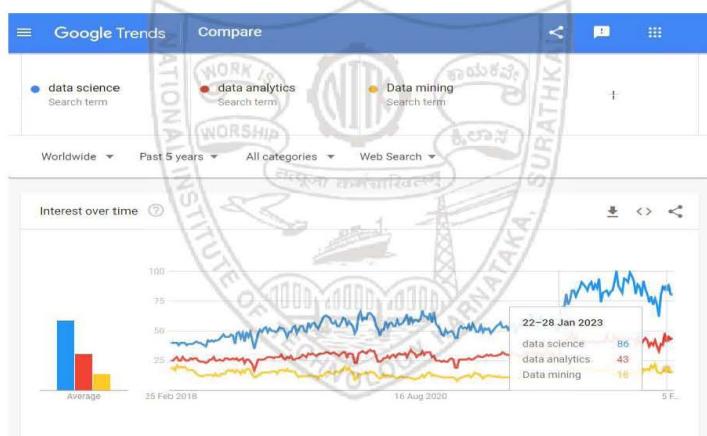
► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

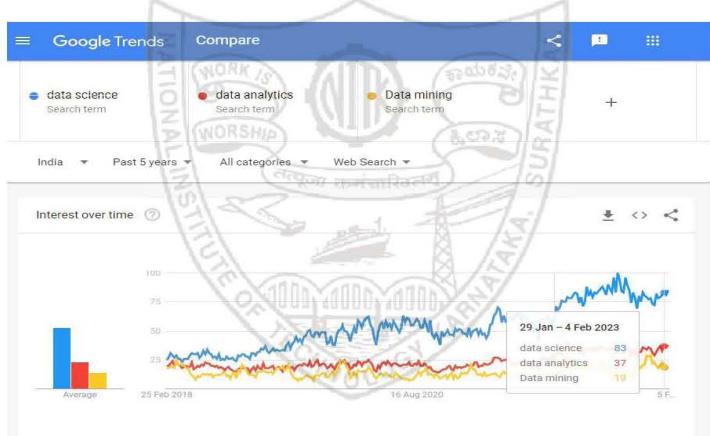
Comparing Interest in Data-related disciplines (Worldwide, All time)

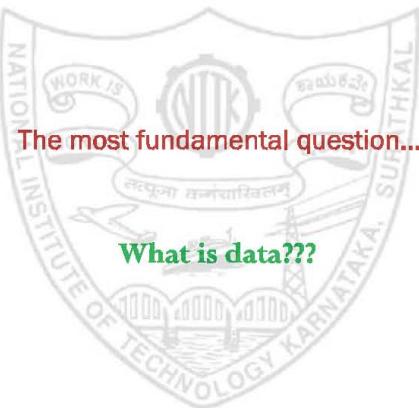


Comparing Interest in Data-related disciplines (Worldwide, Last 5 years)



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





The most fundamental question...

What is data???

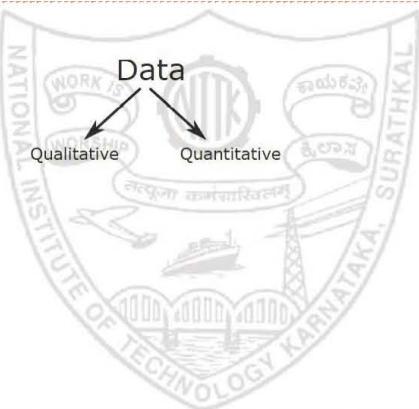
Trying to define data ...

- ▶ “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.”

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

23-Feb-2023

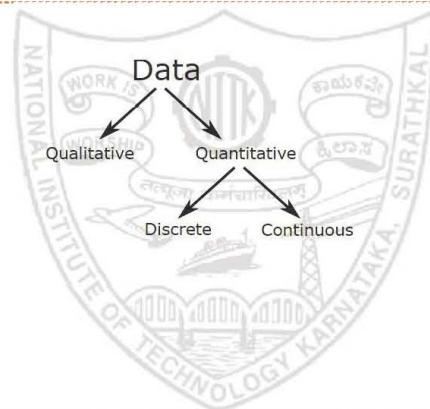
What is data?



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

23-Feb-2023

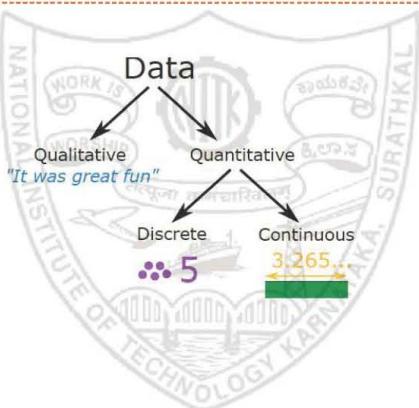
What is data?



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

23-Feb-2023

What is data?



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

23-Feb-2023

Data v/s. Information

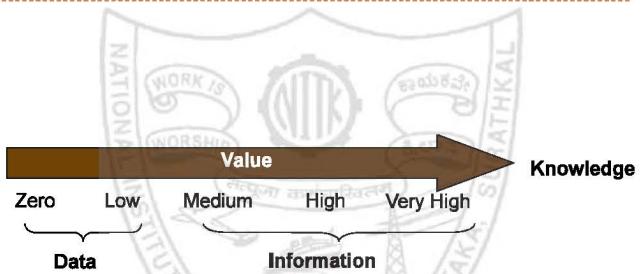
- ▶ **Data:** a set of facts.
- ▶ **Information:** data that are processed to useful; provides answers to "who", "what", "where", and "when" questions
- ▶ **Knowledge:** application of data and information; answers "how" questions.
- ▶ **Understanding:** an appreciation of "why"
- ▶ **Wisdom:** evaluated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

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

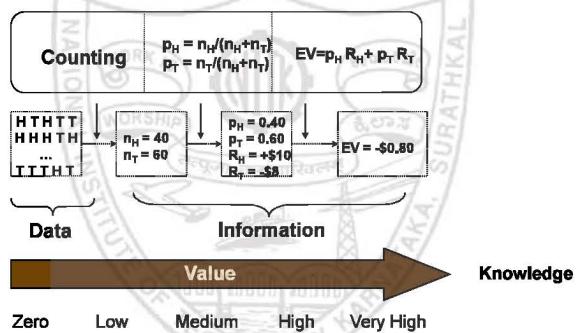
► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

Relationship between D, I and K



DIKW in data analysis...

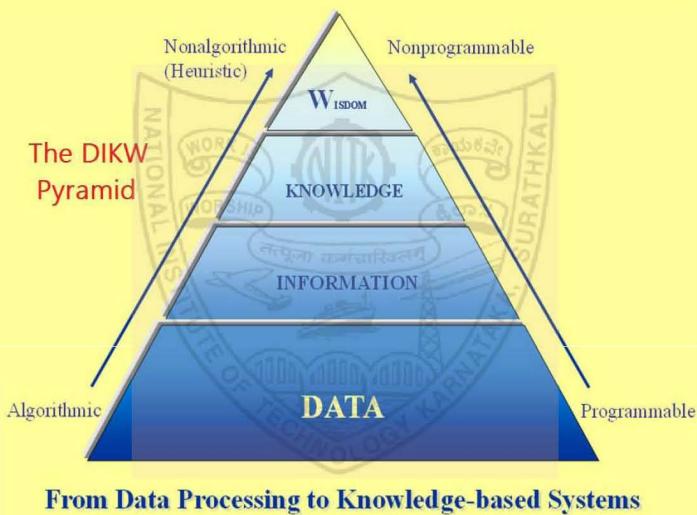


► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

23-Feb-2023

The DIKW Pyramid



From Data to Knowledge and beyond...

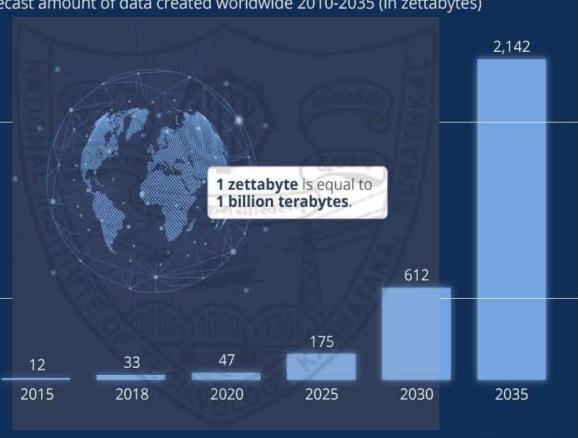
Datafication

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



Global Data Creation is About to Explode

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)



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

23-Feb-2023

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,

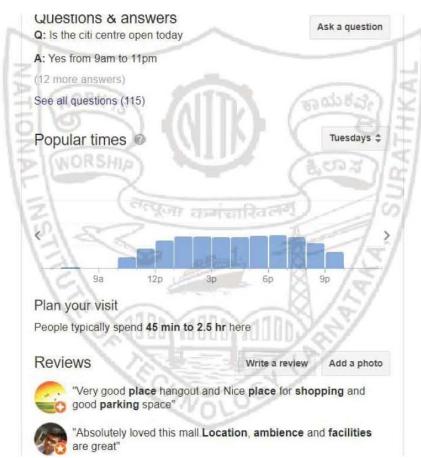
► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

Datafying cities



Datafying public spaces



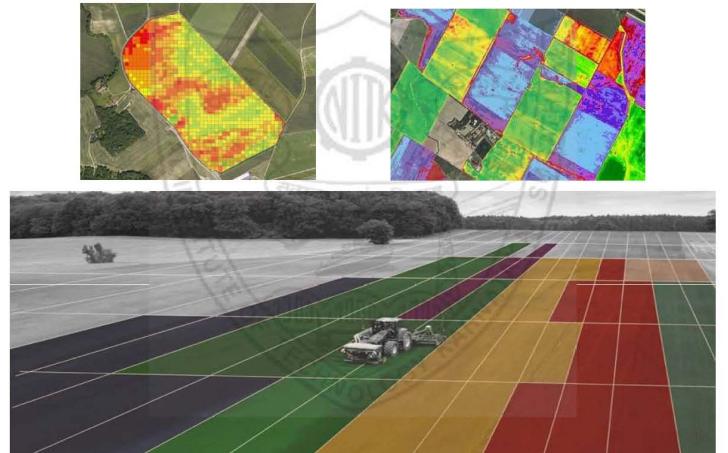
Datafying systems



Datafying processes



Datafying processes



Datafying domains



Datafying activities



amazon

Customers who bought this item also bought



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.

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

23-Feb-2023

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 n^{th} letter based on a given 'skip number' in a given direction.

Reaction to Michael Drosnin's hidden codes book

FINN COCKLAND BULL STORIES ABOUT HIM HAD PREVIOUSLY BEEN
OPENED IN PENNSYLVANIA AND THE FACING BREWERY MAXED FRESH HELIOTROPE
UEGAN NOBLE TRUMPETTE CAPTAIN IN BEING GED HIS FARM ONE FROM
UTTHOUGH IT WASN'T HE MEDAL THEM INSIGNIA CHINA
IXED WITH THE GROUND SHIPS CULTURE AND SALTED FRESH CUT
RNEXT HORSES IN LEAVING QUEENSHIP PRACTICALLY WITH THO
NEHEMIAH TENDERLY SAW HEFFECTIVE HEAD
TOTAL SIGHTS OF THE COUNTRY IN THE ORIG
LAYS AND RAPTURED BY THE SIGHTS REPROBATION TO THE DEED
ILLIAN IMPONENT HANG CAPTAIN IN PLEASANTLY EARLIE
KYELED AND NEVER SAY IT STANDS ON AND SELDOM HORSEY SABRE WITH
STANDS SYNDICATE AND SELDOM HORSEY SABRE WITH
TWAY WINE EARTH THIS CAN'T STYLISHLY BEAR TO
TWILL DOF OF THE EARTH THIS CAN'T STYLISHLY BEAR TO
THE DELEGATION TABORITE UNISQUAGUE FOR EVERY YOUR NAME
ADONAK DODGE C KEEPS LOWDOWN HISTORICAL
MPTURE OF THE EARTH THIS CAN'T STYLISHLY BEAR TO
DEDCOPERS OF THE EARTH THIS CAN'T STYLISHLY BEAR TO
EME BUT HAS TOOK INTENSE REVERSE BY DIVE IN
NGSOMEBOARD AN AVAILABLE CAPTAIN HABITAINED LYVISI
GHEIOED A LONG TIME WIND LASS SHE AND THERE IS USUALLY HIS
ECAPTAINS A DASTOPALAYRINGA YAN AND THERE IS USUALLY HIS
LADIES PLAUDITS AND IF THE IDEA OF PERILSONUCHEN
RETHE WHALESHIRE IN THE THUNDERHEAD HATNIGHT
HACHINER TUCHICHTHURHINIAED HEREFCONERNING
THE DEEDS OF THE EARTH THIS CAN'T STYLISHLY BEAR TO
HEART FEARS FOR DOMESTICITY HUMBLE MASTERS
BOSTONIAN HUMBLE MASTERS

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

Reaction to Michael Drosnin's hidden codes book

בר ארה	Word of YHVH →
אמריקת	America →
ספרא	Sabre →
טביעה	Submersion →
השכלה	Education →
שוואה	Slaughter →
לא פניו	"not my people" →
מוות	Death →
ריהוי	Revelation →
פְּקָדָה	Prediction →
כלדה	Overthrow →
בדל	Destroyer →
משבר	Shattered →
טביהם	Nations-Peoples →
סיני	Chinese →
השך	2006 →
שׁוֹבֵב	2012 →

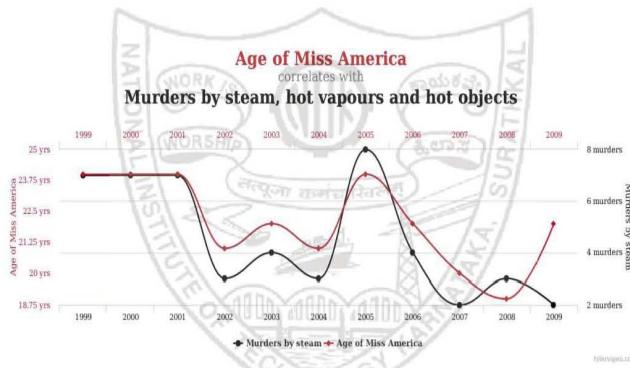
Reaction to Michael Drosnin's hidden codes book

L Y S I C E S O U T A C O N S I D E R A B L E H O L E I N T H E
M W H A L E S H E A D H O I S T E D O N H E R S T A B B O A R D S
D F U L L Y L I K E A T R E A S U R E H U N T E R I N S O M E O L
U N D E R T H E M O U N T A I N S O F T H E S E A T O H I D E H I
E P R O V E D T H A T T H I S I S F O R T H E P U R P O S E O F
H O R S E S I N A R O N G A N D S O C L O S E L Y S H O U L D E R
Y E S O L D F A T P E C U L I A R A N D N O T V E R Y P L E A S A
N G H I M T O O H A Y E B E F K I N N E D T H E F I R S T D A Y A
I G N F O R T H E R E C I T I S R I G H T O P P O S I T E T H E G O
O I V I S I T E D T H I S W O N D E R F U L Q U A S H W A L E A N D S A W T
O R D I N A R Y D U T I E S R E P A I R I N G S T O V E B O A T S
A L W A T E R S A N D G L I C I N G T O W A R D S T H E J A P A N
T H E I N T R I C A T E H A M P T E R T H E F E W H I C E S U B B
H T H E S E G R E E Y H A I S O F N I N E I S N O T W O R T H T H W
U N P R O V I D E D W I T H C O N S T A N T A T C H M A N T H
E A W A Y N O W T H E D E C K I S T H E M S I R A N D S O S A Y
M E D D R I V I N G N A A I D I N T O H I S J E A R T H B U T H E R

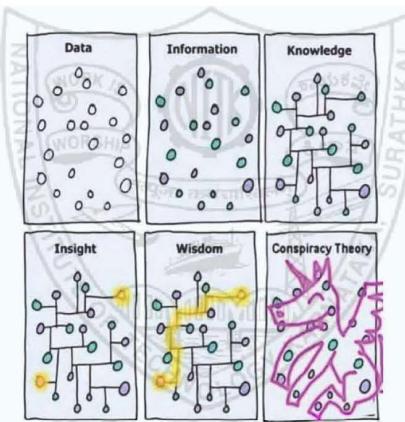
The Demise of Drosnin: A shocking discovery has been made deep within the text of **Moby Dick**.

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

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

23-Feb-2023

Rhine Paradox

► Joseph Rhine's Extra-sensory Perception experiments.

► a great example of how not to conduct scientific research!

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

23-Feb-2023

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

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

23-Feb-2023

Total Information Awareness

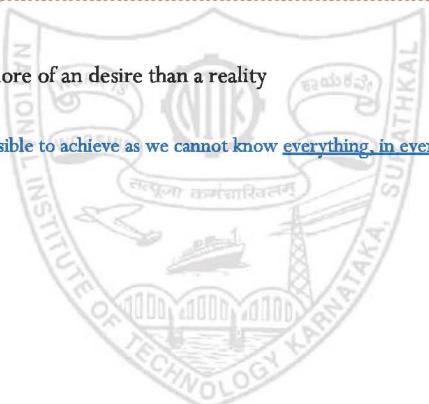
- Four states of Information Awareness -
- KNOW WHAT YOU KNOW — *Information you know and have.*
- DON'T KNOW WHAT YOU KNOW — *Information you don't know you have*
- KNOW WHAT YOU DON'T KNOW — *Information you know you need but don't have.*
- DON'T KNOW WHAT YOU DON'T KNOW — *Information you don't even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

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

23-Feb-2023

Total Information Awareness

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

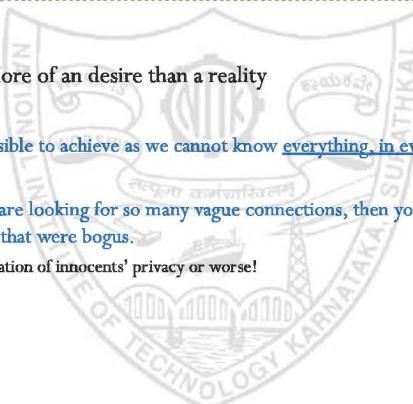


► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

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!

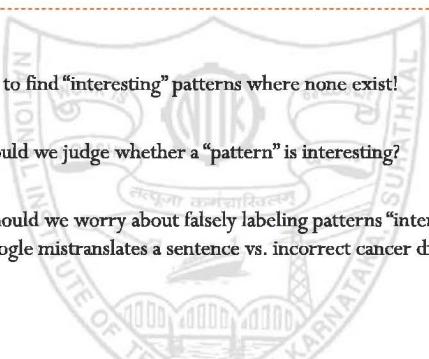


► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

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. . .)
- ▶

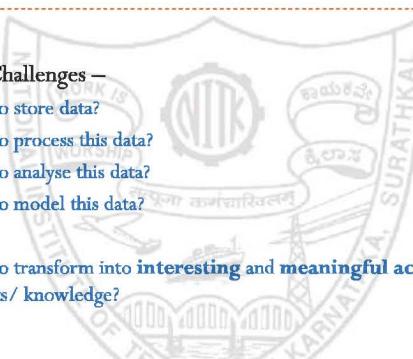


► Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

23-Feb-2023

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?**



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

23-Feb-2023



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 ...*

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

23-Feb-2023

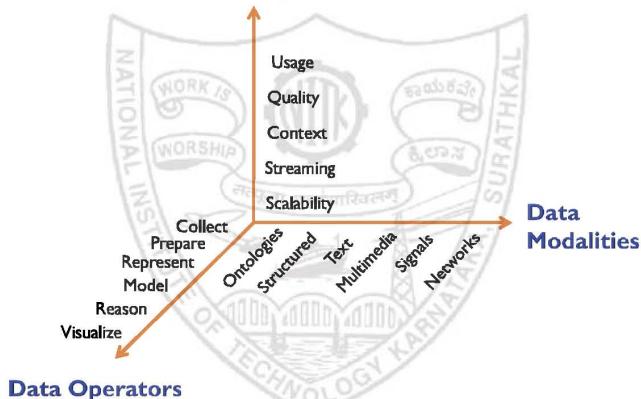
The Life of Data



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

23-Feb-2023

Challenges



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...

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

23-Feb-2023

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.

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

23-Feb-2023