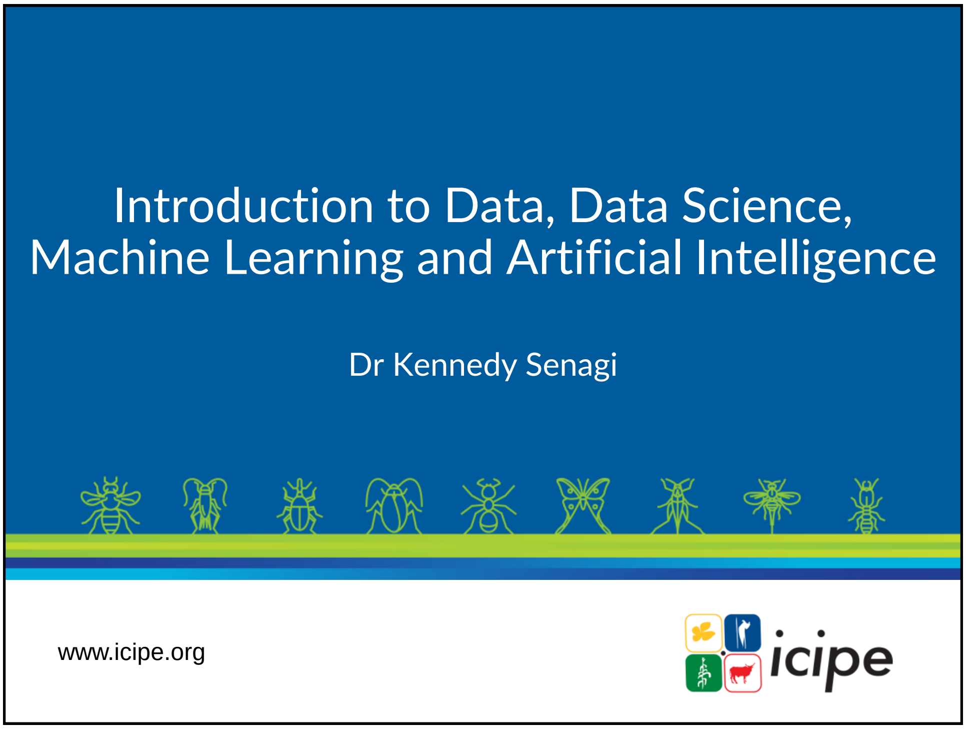
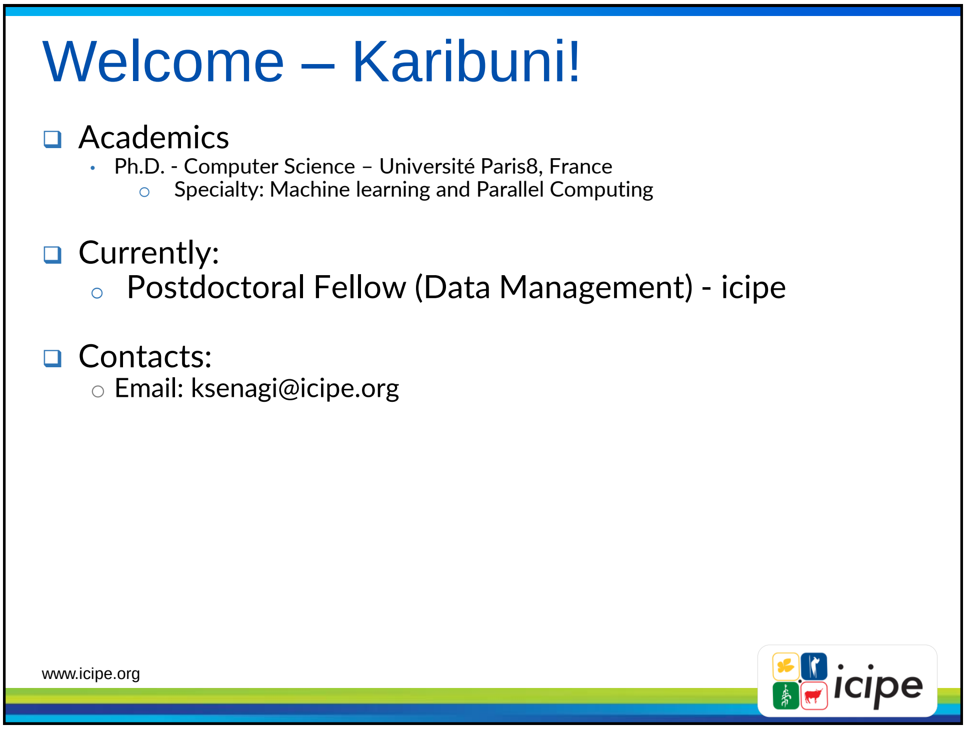


Introduction to Data, Data Science, Machine Learning and Artificial Intelligence

Dr Kennedy Senagi

www.icipe.org

1



Welcome – Karibuni!

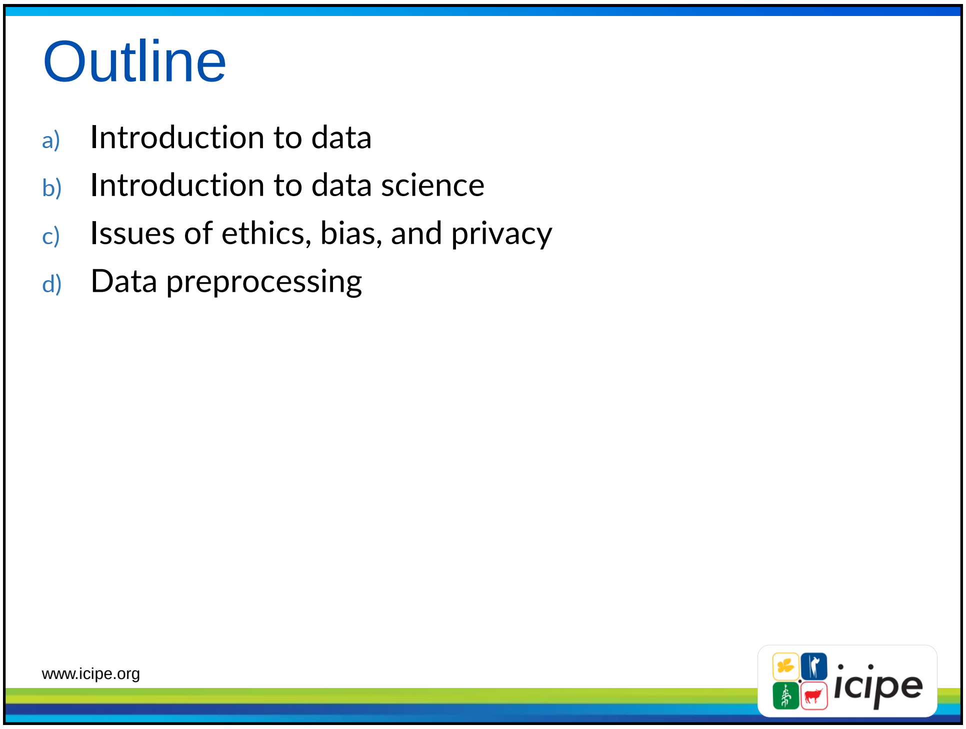
❑ Academics
- Ph.D. - Computer Science – Université Paris8, France
  - ○ Specialty: Machine learning and Parallel Computing

❑ Currently:
- ○ Postdoctoral Fellow (Data Management) - icipe

❑ Contacts:
- ○ Email: ksenagi@icipe.org

www.icipe.org

2

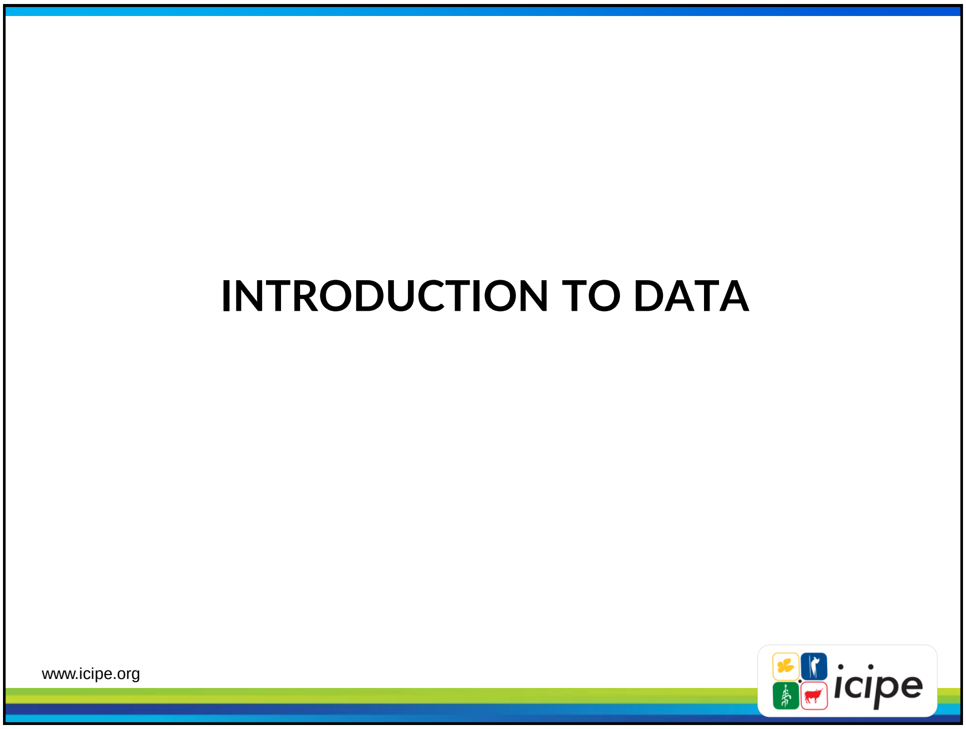


Outline

a) Introduction to data
b) Introduction to data science
c) Issues of ethics, bias, and privacy
d) Data preprocessing

www.icipe.org

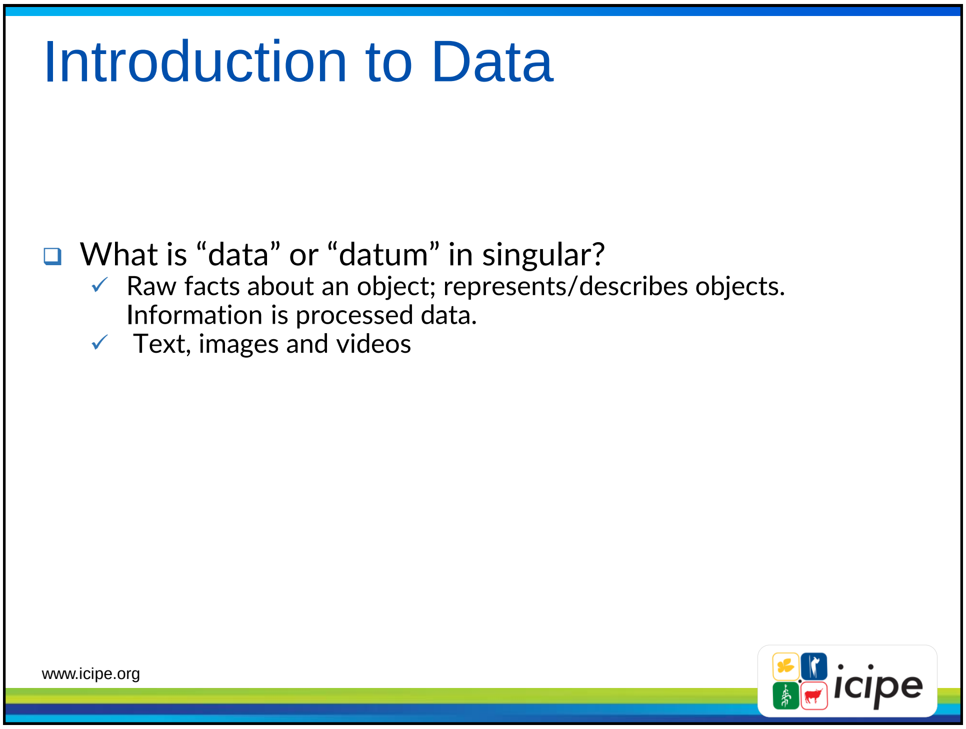
3



INTRODUCTION TO DATA

www.icipe.org

4



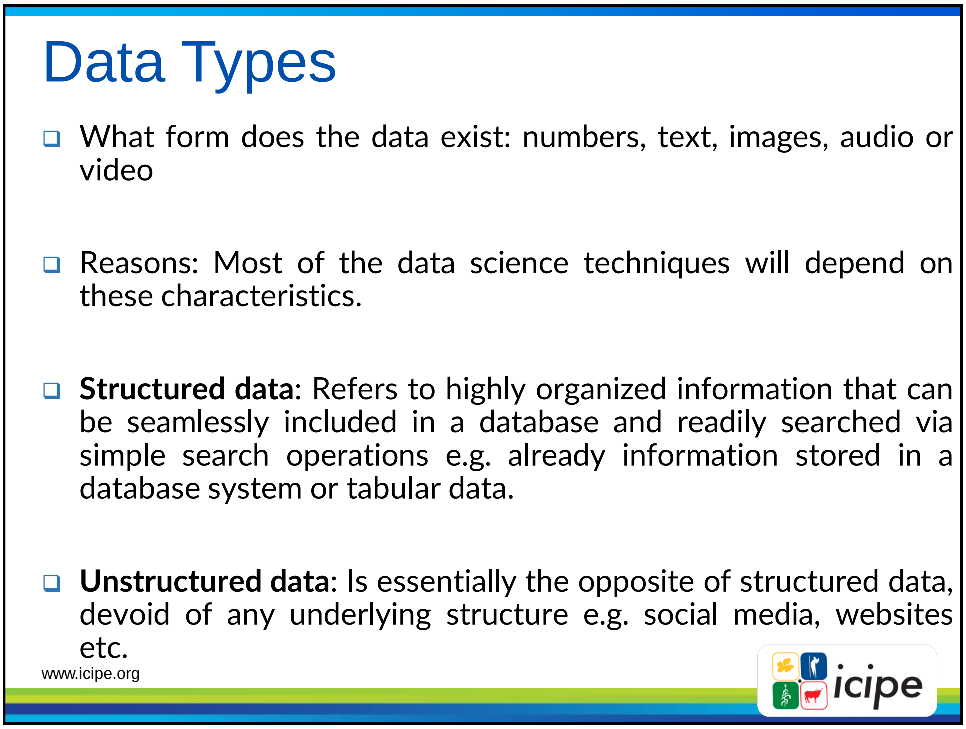
Introduction to Data

- What is “data” or “datum” in singular?
  - ✓ Raw facts about an object; represents/describes objects. Information is processed data.
  - ✓ Text, images and videos

www.icipe.org

icipe

5



Data Types

- What form does the data exist: numbers, text, images, audio or video

- Reasons: Most of the data science techniques will depend on these characteristics.

- **Structured data**: Refers to highly organized information that can be seamlessly included in a database and readily searched via simple search operations e.g. already information stored in a database system or tabular data.

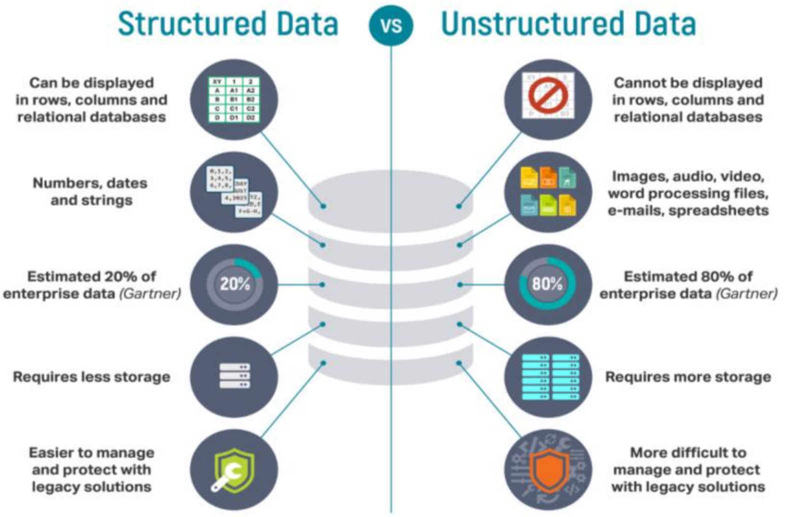
- **Unstructured data**: Is essentially the opposite of structured data, devoid of any underlying structure e.g. social media, websites etc.

www.icipe.org

icipe

6

Structured and Unstructured Data



Structured Data VS Unstructured Data

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data *(Gartner)*

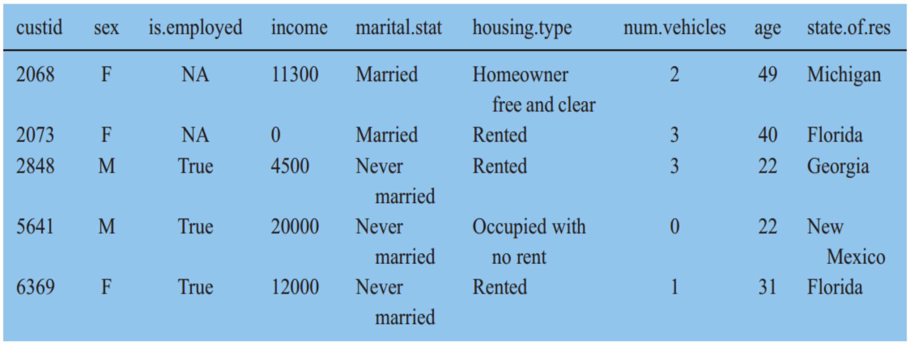
Requires more storage

More difficult to manage and protect with legacy solutions

www.icipe.org

7

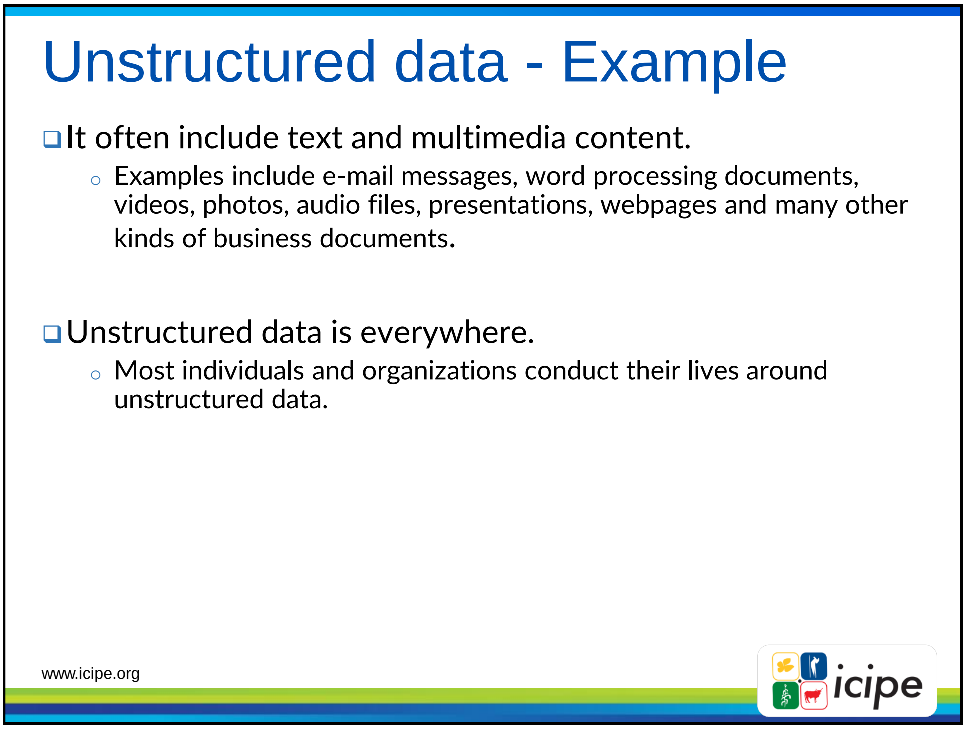
Structured Data - Example



custid	sex	is.employed	income	marital.stat	housing.type	num.vehicles	age	state.of.res
2068	F	NA	11300	Married	Homeowner free and clear	2	49	Michigan
2073	F	NA	0	Married	Rented	3	40	Florida
2848	M	True	4500	Never married	Rented	3	22	Georgia
5641	M	True	20000	Never married	Occupied with no rent	0	22	New Mexico
6369	F	True	12000	Never married	Rented	1	31	Florida

www.icipe.org

8



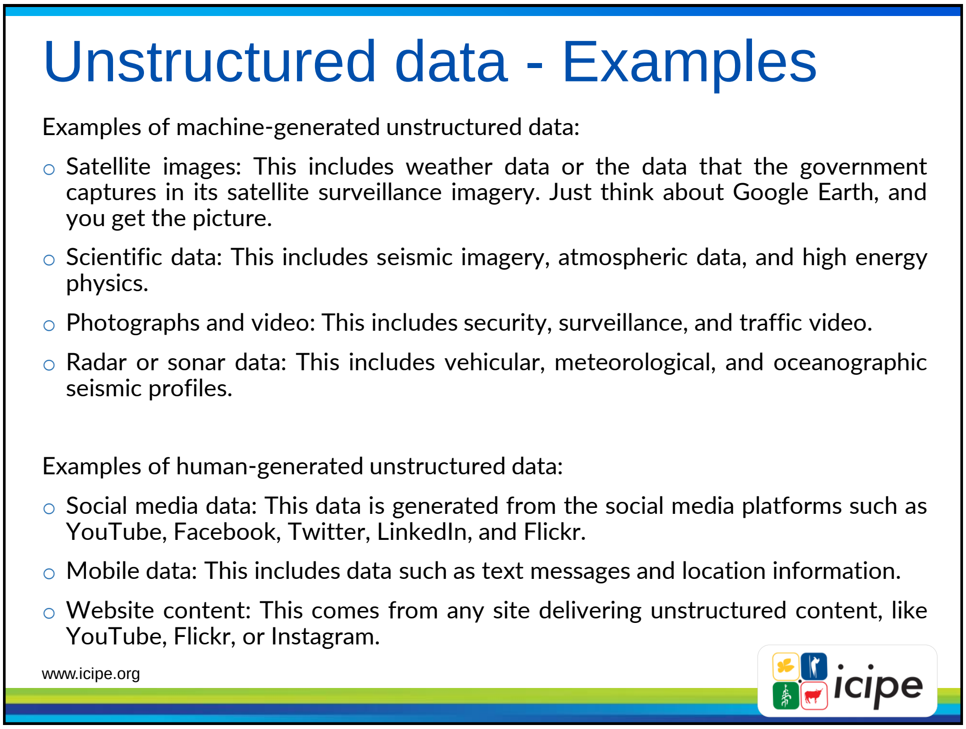
Unstructured data - Example

- ❑ It often include text and multimedia content.
  - ○ Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents.

- ❑ Unstructured data is everywhere.
  - ○ Most individuals and organizations conduct their lives around unstructured data.

www.icipe.org

9



Unstructured data - Examples

Examples of machine-generated unstructured data:

- ○ Satellite images: This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
- ○ Scientific data: This includes seismic imagery, atmospheric data, and high energy physics.
- ○ Photographs and video: This includes security, surveillance, and traffic video.
- ○ Radar or sonar data: This includes vehicular, meteorological, and oceanographic seismic profiles.

Examples of human-generated unstructured data:

- ○ Social media data: This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- ○ Mobile data: This includes data such as text messages and location information.
- ○ Website content: This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

www.icipe.org

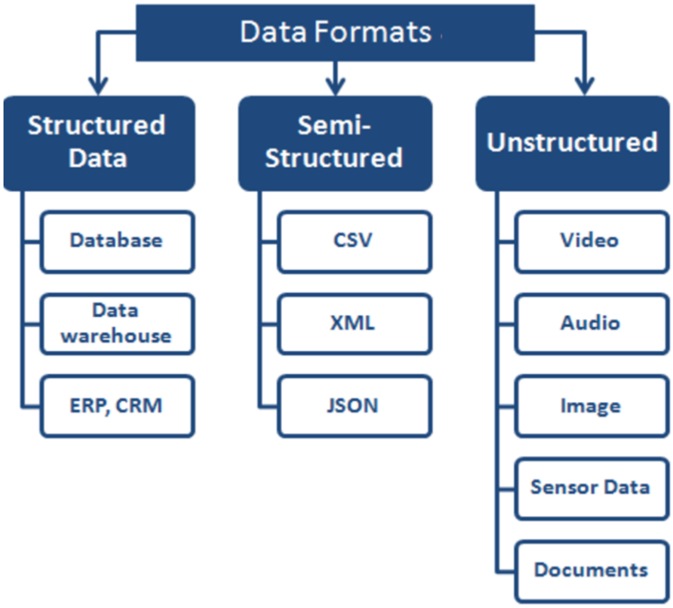
10

Source of data

- ❑ Human and/or machines – Primary or secondary
- ❑ Primary data - data collection –fieldwork or lab work
  - ❑ Other: legacy (manual/automated) data, hospitals, transport, banks, telecommunications, finance, agriculture, metrology, space, education/institutions, M&E surveys, customer logs,
- ❑ Secondary data
  - ❑ Open-data:
    - ○ Freely available in a public domain and can be used by anyone as they wish, without restrictions from copyright, patents, or other mechanisms of control. But acknowledge source/creator.
  - ○ Social media:
    - ○ Social media has become a gold mine for collecting data to analyze for research or marketing purposes.

www.icipe.org

11



Data Formats

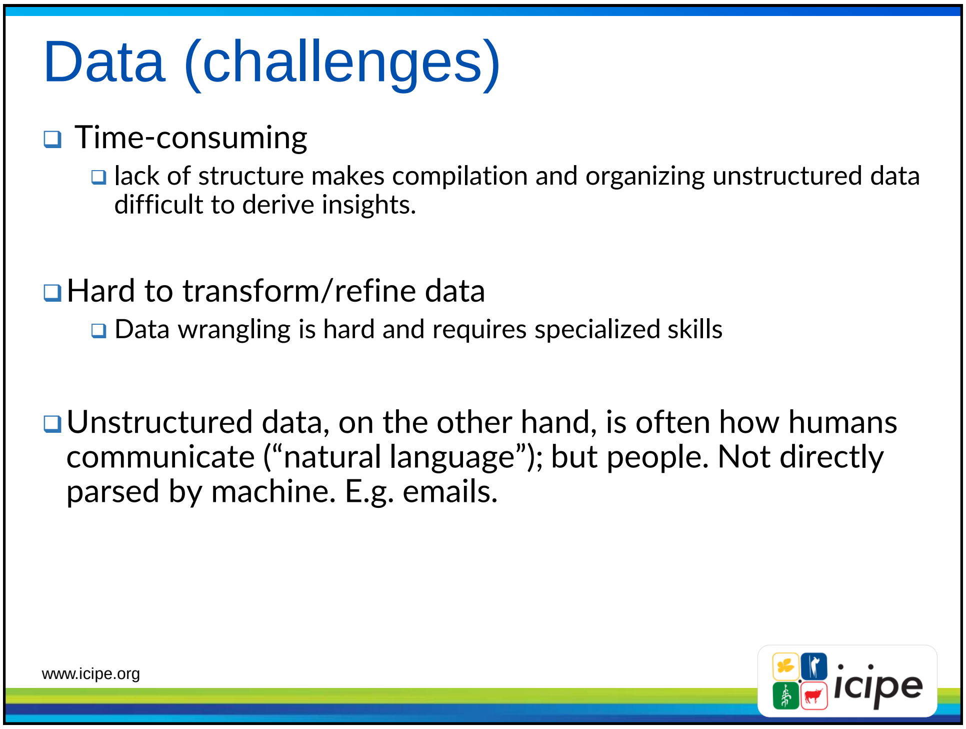
Structured Data
- Database
- Data warehouse
- ERP, CRM

Semi-Structured
- CSV
- XML
- JSON

Unstructured
- Video
- Audio
- Image
- Sensor Data
- Documents

www.icipe.org

12

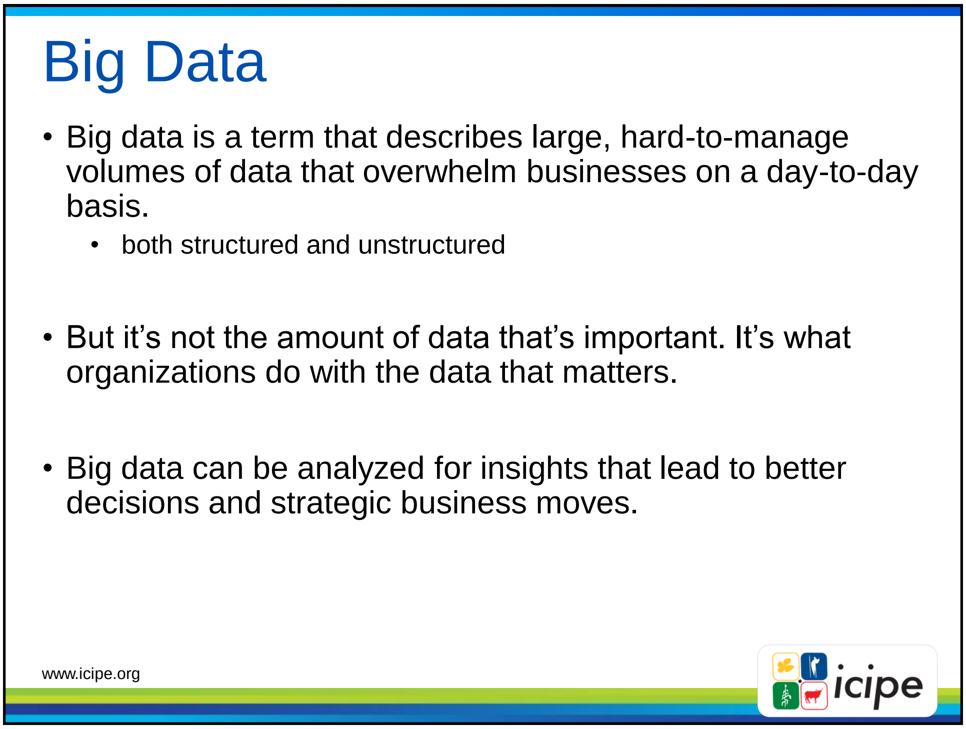


Data (challenges)

- ❑ Time-consuming
  - ❑ lack of structure makes compilation and organizing unstructured data difficult to derive insights.
- ❑ Hard to transform/refine data
  - ❑ Data wrangling is hard and requires specialized skills
- ❑ Unstructured data, on the other hand, is often how humans communicate ("natural language"); but people. Not directly parsed by machine. E.g. emails.

www.icipe.org

13



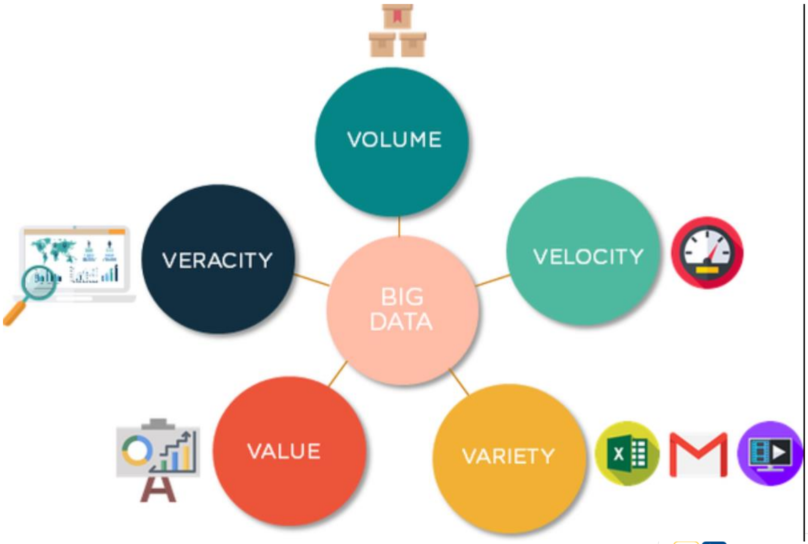
Big Data

- Big data is a term that describes large, hard-to-manage volumes of data that overwhelm businesses on a day-to-day basis.
  - both structured and unstructured
- But it's not the amount of data that's important. It's what organizations do with the data that matters.
- Big data can be analyzed for insights that lead to better decisions and strategic business moves.

www.icipe.org

14

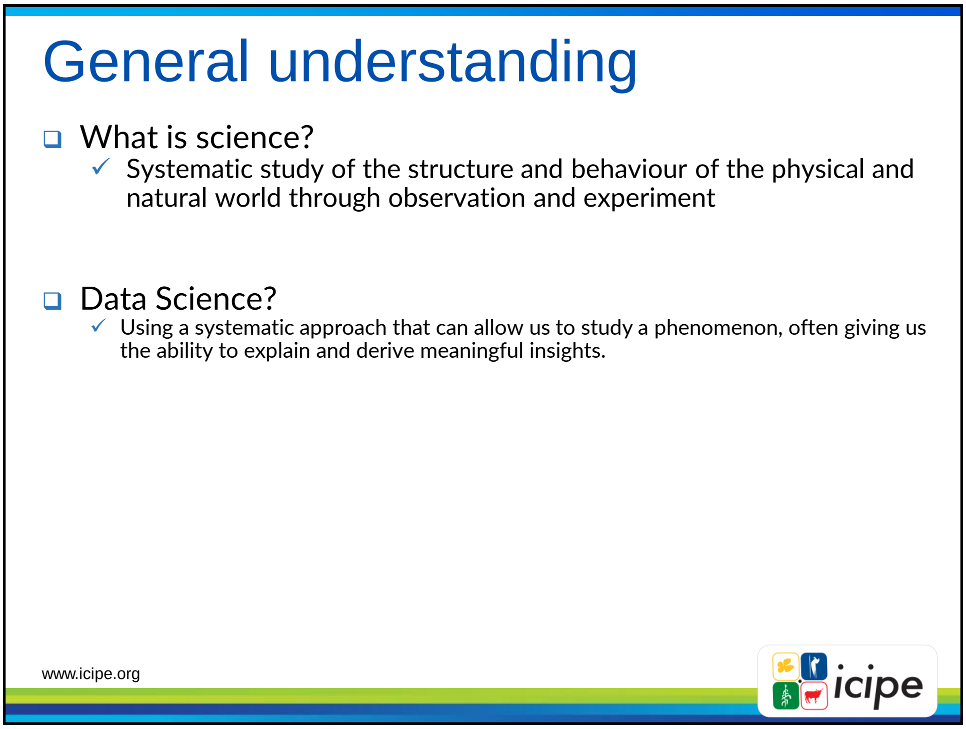
Big Data (challenges)



15

INTRODUCTION TO DATA SCIENCE

16



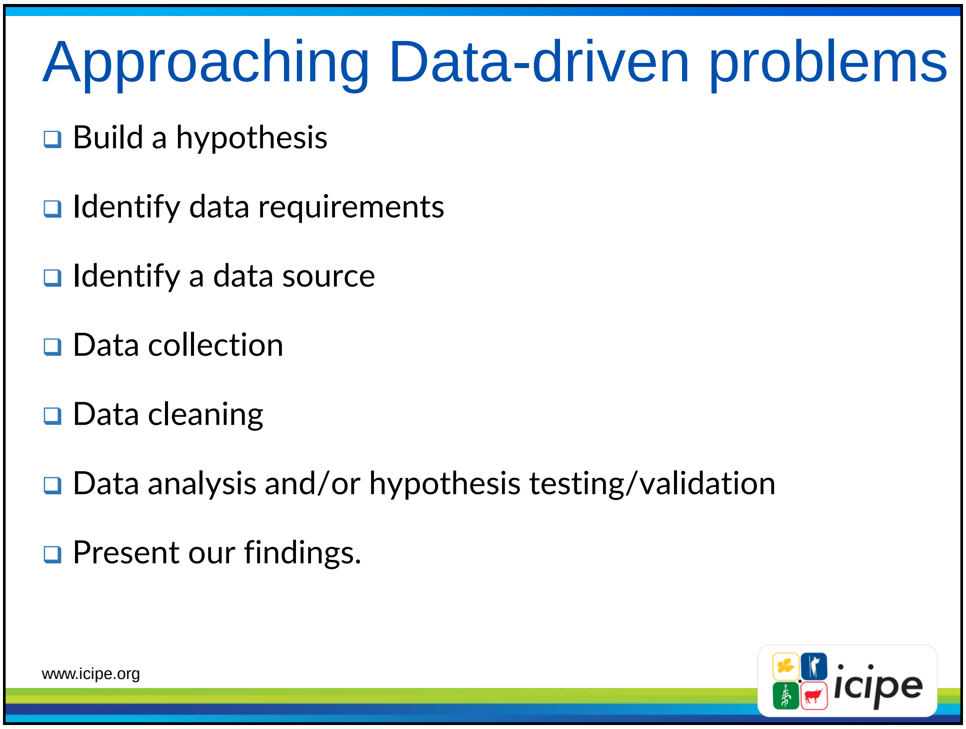
General understanding

- What is science?
  - Systematic study of the structure and behaviour of the physical and natural world through observation and experiment
- Data Science?
  - Using a systematic approach that can allow us to study a phenomenon, often giving us the ability to explain and derive meaningful insights.

www.icipe.org

icipe

17



Approaching Data-driven problems

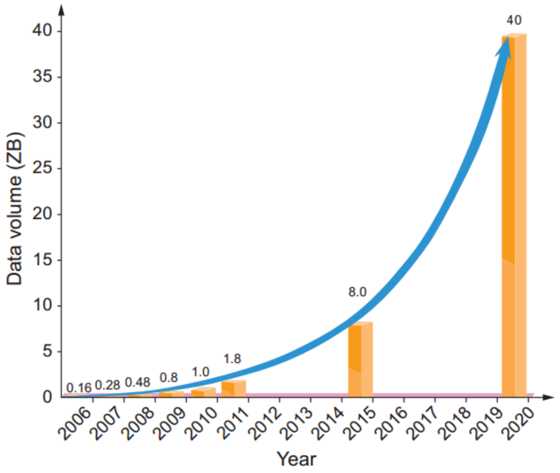
- Build a hypothesis
- Identify data requirements
- Identify a data source
- Data collection
- Data cleaning
- Data analysis and/or hypothesis testing/validation
- Present our findings.

www.icipe.org

icipe

18

Data Volume in the Last 15 Years



Increase of data volume in last 15 years. (Source: IDC's Digital Universe Study, December 2012.[5])

www.icipe.org

19

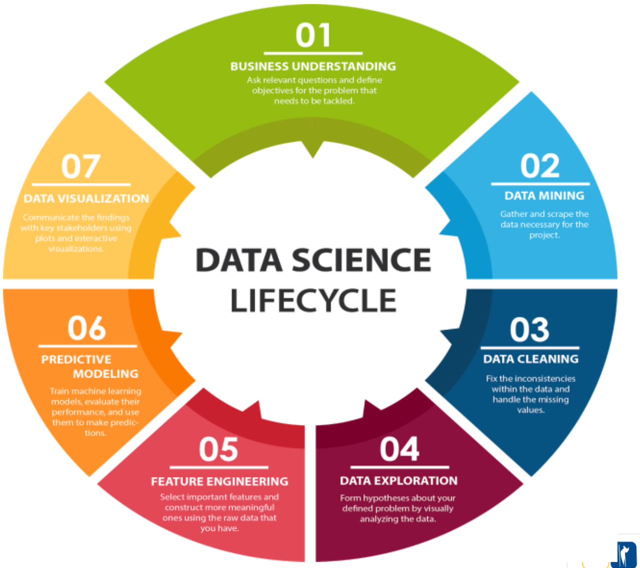
What do Data Scientists do?

- ❑ Collect, clean, retrieve, analyze and store data
  - ✓ All for the purpose of deriving meaningful insights toward making decisions and solving problems
- ❑ Apply scientific approaches and techniques
  - ✓ Use systematic, verifiable, and repeatable processes
- ❑ Uncover insights from mining data
  - ✓ Through exploration of the data using various tools and techniques, testing hypotheses, and creating conclusions with data and analyses as evidence.
- ❑ Data Visualization
  - ✓ Human to see underlying data patterns and insights
- ❑ Data inference, algorithm development, and technology
  - ✓ To solve analytically complex problems

www.icipe.org

20

Data Science Lifecycle



DATA SCIENCE LIFECYCLE

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

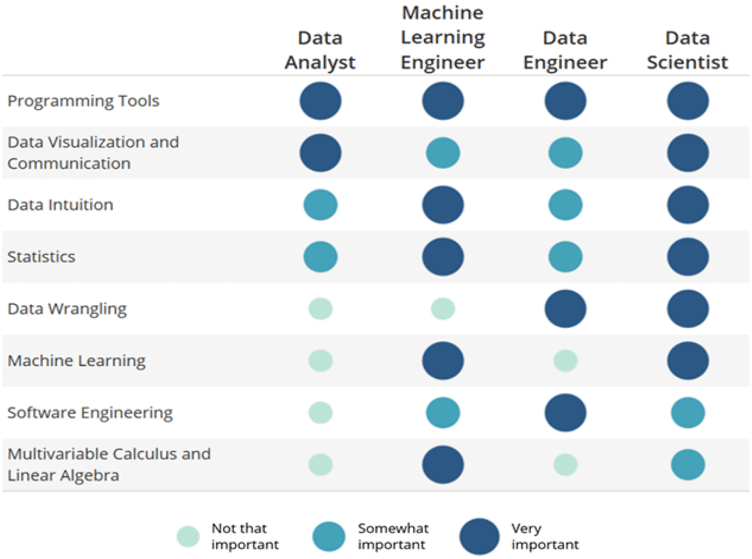
06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

www.icipe.org

21

Skills for Data Scientists



	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◐	◐	●
Data Intuition	◐	●	◐	●
Statistics	◐	●	◐	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◐	●	◐
Multivariable Calculus and Linear Algebra	○	●	○	◐

○ Not that important ◐ Somewhat important ● Very important

www.icipe.org

22

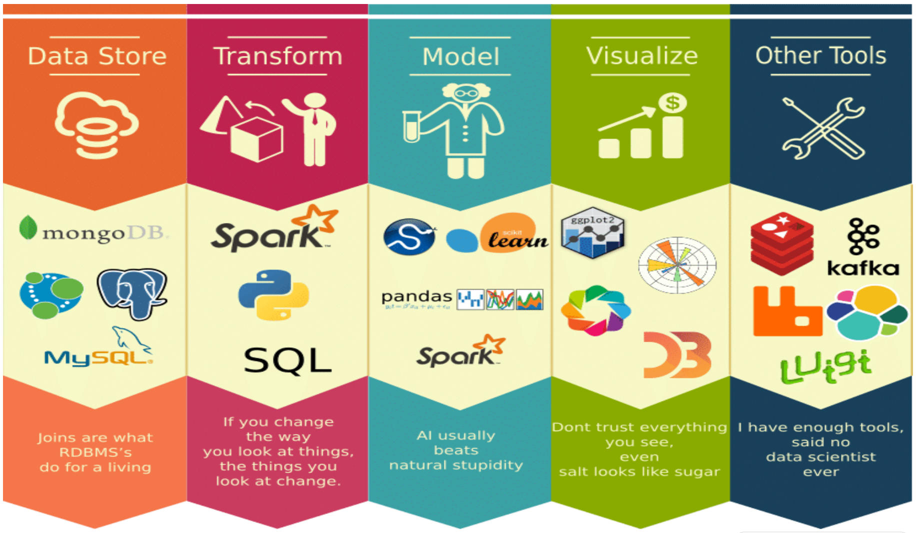
Other Skills for Data Scientists

- ❑ Willing to Experiment:
  - ○ Needs to have the drive, intuition, and curiosity not only to solve problems as they are presented, but also to identify and articulate problems on her own.
- ❑ Proficiency in Mathematical Reasoning:
  - ○ Have a strong grasp on the basic statistical methods and how to employ them.
- ❑ Data Literacy:
  - ○ This is the ability to extract meaningful information from a dataset. Its important to assess a dataset for relevance and suitability for the purpose of interpretation, to perform analysis, and create meaningful visualizations to tell valuable data stories.
- ❑ Story narration:
  - ○ A structured approach for communicating data insights. It involves a combination of three key elements: data, visuals, and narrative / interpretation. Domain experts can/should be involved in coming up with meaningful stories.

www.icipe.org

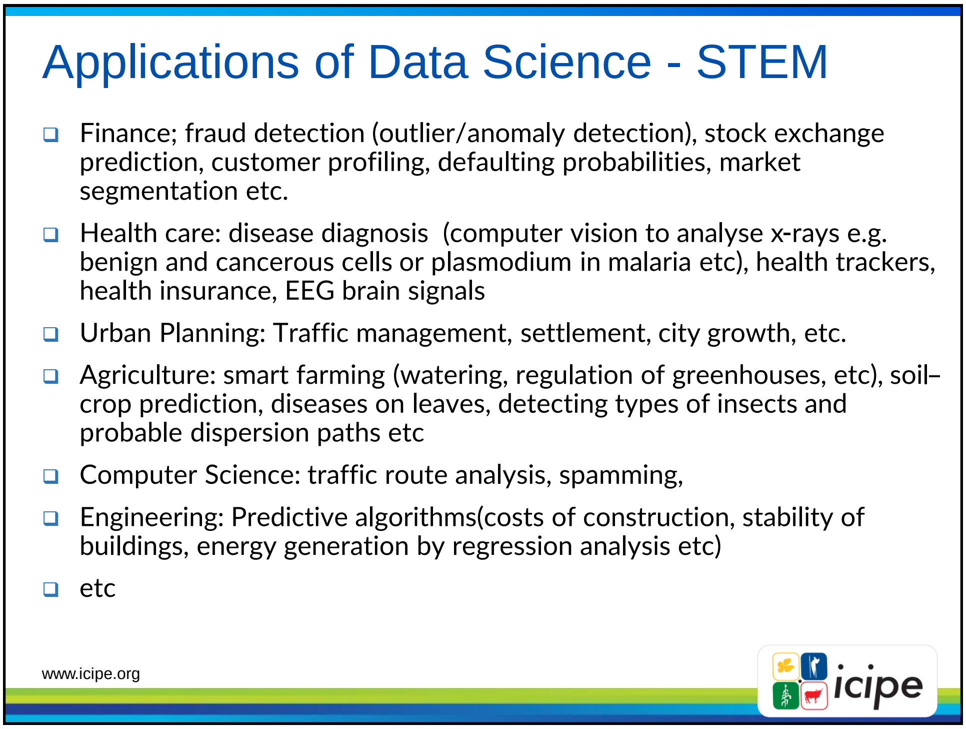
23

Tool for Data Science


Data Store
Transform
Model
Visualize
Other Tools
mongoDB
MySQL
Spark
SQL
scikit learn
pandas
Spark
ggplot2
D3
kafka
Luigi
Joins are what RDBMS's do for a living
If you change the way you look at things, the things you look at change.
AI usually beats natural stupidity
Dont trust everything you see, even salt looks like sugar
I have enough tools, said no data scientist ever

www.icipe.org

24

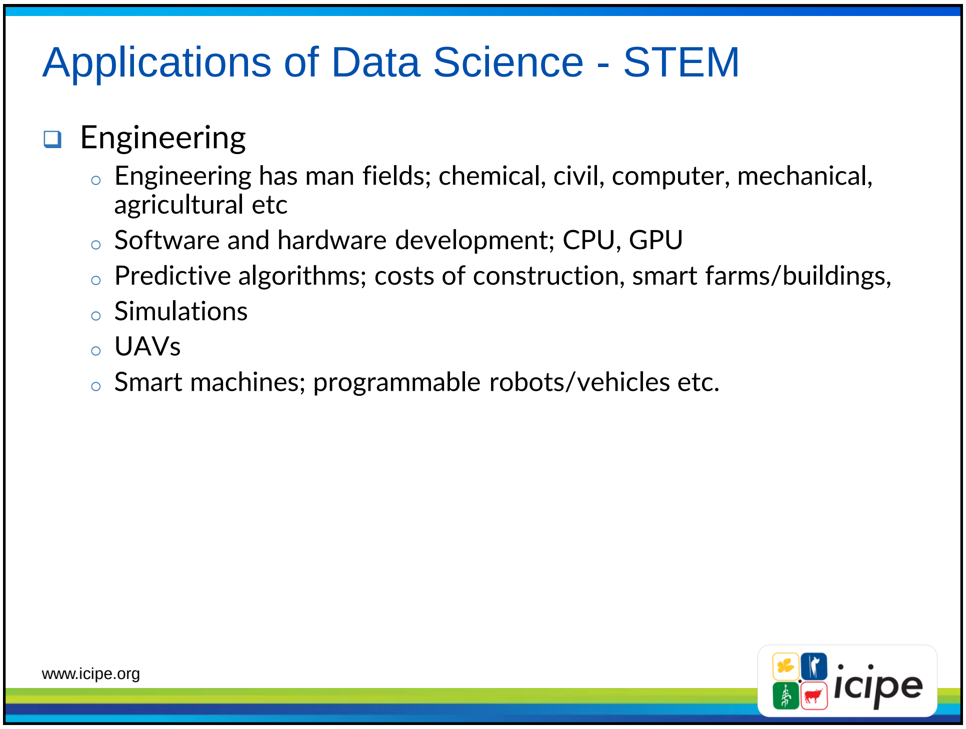


Applications of Data Science - STEM

- ❑ Finance; fraud detection (outlier/anomaly detection), stock exchange prediction, customer profiling, defaulting probabilities, market segmentation etc.
- ❑ Health care: disease diagnosis (computer vision to analyse x-rays e.g. benign and cancerous cells or plasmodium in malaria etc), health trackers, health insurance, EEG brain signals
- ❑ Urban Planning: Traffic management, settlement, city growth, etc.
- ❑ Agriculture: smart farming (watering, regulation of greenhouses, etc), soil–crop prediction, diseases on leaves, detecting types of insects and probable dispersion paths etc
- ❑ Computer Science: traffic route analysis, spamming,
- ❑ Engineering: Predictive algorithms(costs of construction, stability of buildings, energy generation by regression analysis etc)
- ❑ etc

www.icipe.org

25



Applications of Data Science - STEM

- ❑ Engineering
  - ○ Engineering has man fields; chemical, civil, computer, mechanical, agricultural etc
  - ○ Software and hardware development; CPU, GPU
  - ○ Predictive algorithms; costs of construction, smart farms/buildings,
  - ○ Simulations
  - ○ UAVs
  - ○ Smart machines; programmable robots/vehicles etc.

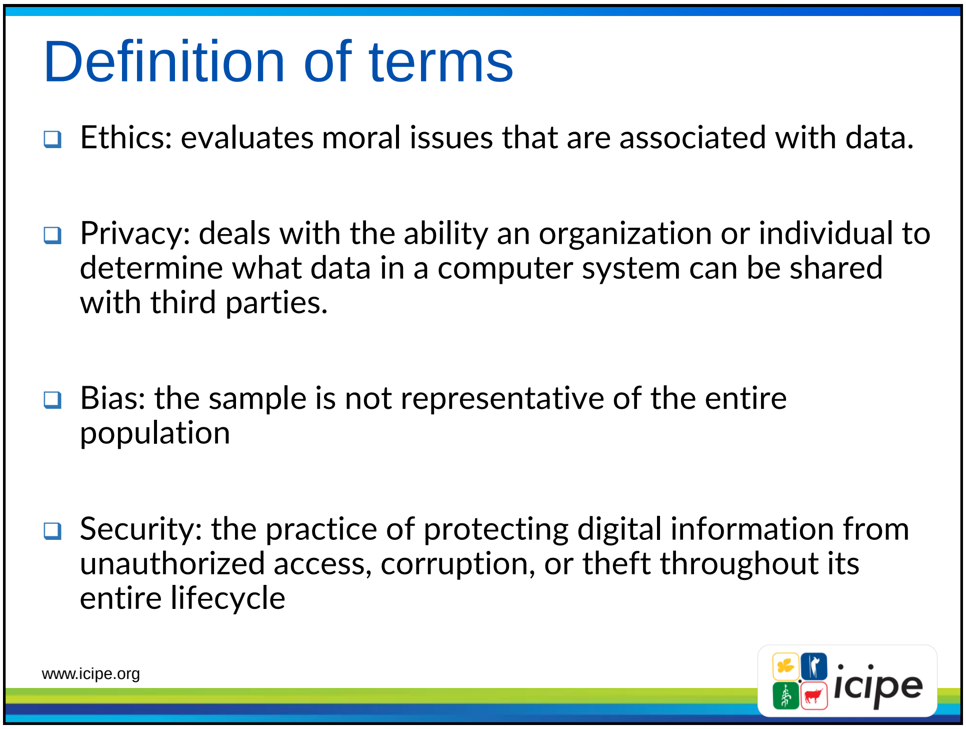
www.icipe.org

26

ISSUES OF ETHICS, BIAS AND PRIVACY IN DATA SCIENCE

www.icipe.org

27



Definition of terms

- ❑ Ethics: evaluates moral issues that are associated with data.
- ❑ Privacy: deals with the ability an organization or individual to determine what data in a computer system can be shared with third parties.
- ❑ Bias: the sample is not representative of the entire population
- ❑ Security: the practice of protecting digital information from unauthorized access, corruption, or theft throughout its entire lifecycle

www.icipe.org

28

Data Privacy and Security: Why?

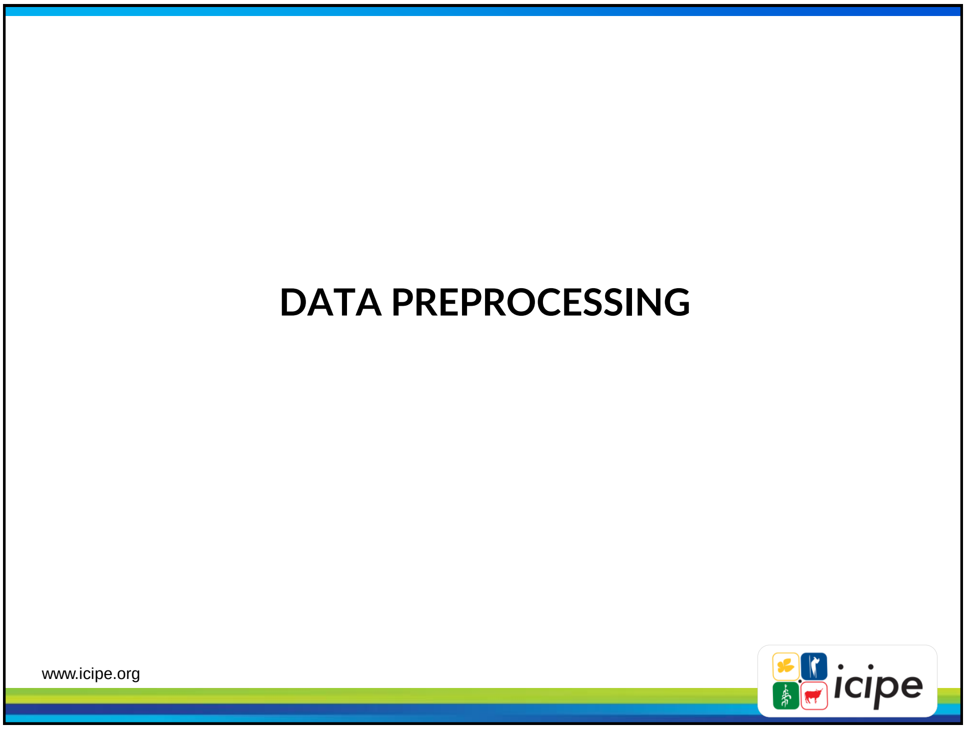
- ❑ State/organization/vendor laws
- ❑ Increasing penalties
- ❑ Theft of consumer information increasing
- ❑ Increased government investigations
- ❑ Private consumer litigation
- ❑ Bad image on brands
- ❑ Attacks on systems increasing

29

Privacy, bias, and ethics

- ❑ What, how, where, and why was the data collected?
- ❑ Who collected it?
- ❑ What did they intend to use it for?
- ❑ If the data was collected from people, did these people know that:
  - ○ Such data was being collected about them
  - ○ How the data would be used? Under what circumstances it can be shared/disclosed
- ❑ How data is legally collected or stored or used for other purposes!
- ❑ Often those collecting data mistake availability of data as the right to use that data!
- ❑ Whether or how data is shared with third parties!
- ❑ Regulatory restrictions!

30

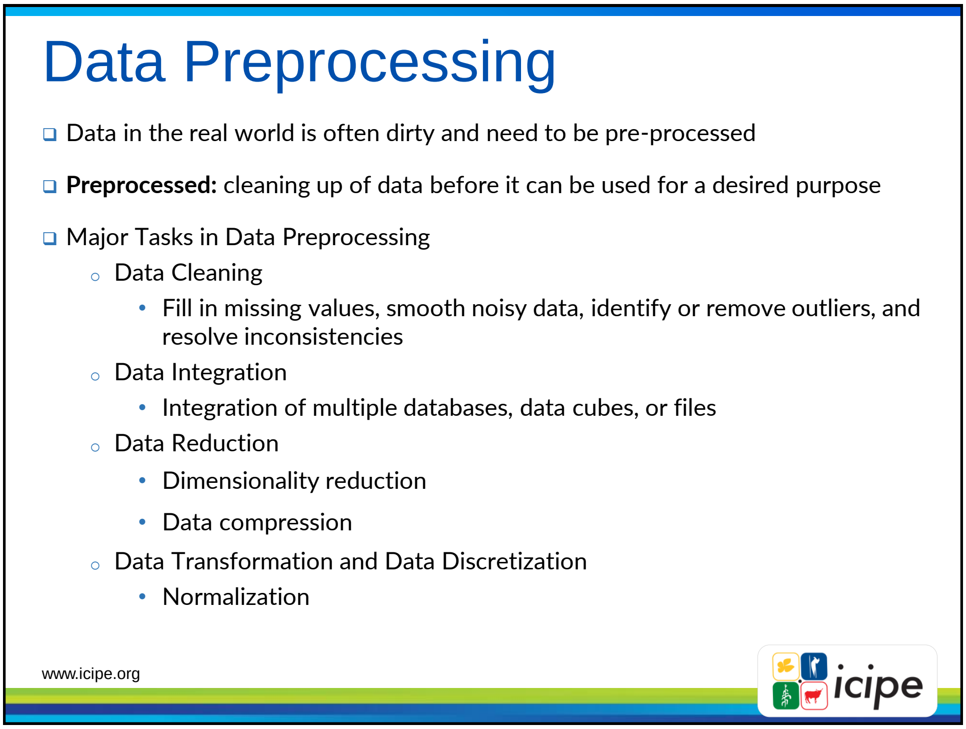


DATA PREPROCESSING

www.icipe.org

icipe

31



Data Preprocessing

- ❑ Data in the real world is often dirty and need to be pre-processed
- ❑ **Preprocessed:** cleaning up of data before it can be used for a desired purpose
- ❑ Major Tasks in Data Preprocessing
  - Data Cleaning
    - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
  - Data Integration
    - Integration of multiple databases, data cubes, or files
  - Data Reduction
    - Dimensionality reduction
    - Data compression
  - Data Transformation and Data Discretization
    - Normalization

www.icipe.org

icipe

32

Measure of data quality

- Accuracy: correct or wrong, accurate or not
- Completeness:
  - Incomplete: When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.
- Consistency: some modified but some not, dangling, …
  - Inconsistent. Data contains discrepancies in codes or names. E.g., if the “Name” column for registration records of employees contains values other than alphabetical letters, or if records do not start with a capital letter, discrepancies are present.
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?
- Noisy: When data contains errors or outliers. E.g. some of the data points in a dataset may contain extreme values that can severely affect the dataset’s range.
- Etc etc.

www.icipe.org

icipe

33

Direct Financial Support to *icipe* from:-

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Agency for Development and Cooperation SDC
Sida
Foreign, Commonwealth & Development Office
Government of Kenya
Federal Democratic Republic of Ethiopia
AFRICAN UNION
ASI
AWARD
Australian Government Australian Centre for International Agricultural Research
Austraining International
BAYER
BERTHA FOUNDATION
BILL & MELINDA GATES foundation
BBSRC bioscience for the future
Government of the Republic of Benin
Government of Burkina Faso
INSTITUTIONAL LINKS
Newton Fund
BioInnovate Africa
biovision
Biovision Africa Trust
CAMBRIDGE AFRICA
BMZ Federal Ministry for Economic Cooperation and Development
cim
cirad
Government of Côte d'Ivoire
CONSERVATION INTERNATIONAL
CS&S Code for Science & Society
DAAD
DFG Deutsche Forschungsgemeinschaft
Embrapa
ETH zürich
FiBL
FORD FOUNDATION
FNSNF
giz
Government of Ghana
Google.org
Grand Challenges Canada
IAEA
IDRC CRDI
IFAD Investing in rural people
IFS International Foundation for Science
iie
impaxio
Innovate UK
IVCC
JRS Biodiversity Foundation
FONDATION L'ORÉAL
Government of South Korea
JIRCAS
Keele UNIVERSITY
mastercard foundation
Max Planck Institute for Chemical Ecology
THE McKNIGHT FOUNDATION
MINISTRY FOR FOREIGN AFFAIRS OF FINLAND
moz://a
Government of Mozambique
NATIONAL ACADEMY OF SCIENCES
NATIONAL GEOGRAPHIC SOCIETY
NIH National Institutes of Health
NRF
NATIONAL RESEARCH FUND
NSF
Netherlands Directorate-General for International Cooperation
NWO
Open Philanthropy
R. Geigy-Stiftung
Russell IPM
Norad
ROCKEFELLER FOUNDATION
THE ROYAL SOCIETY
Government of Rwanda
rss
Scottish Funding Council
Government of Senegal
switchafrica GREEN
THE WORLD BANK
twas
UNEP
University of Glasgow
USAID FROM THE AMERICAN PEOPLE
USDA
Vetenskapsrådet
VolkswagenStiftung
wellcome
World Health Organization
WORLD TRADE ORGANIZATION
WSD

www.icipe.org

icipe

34

Thank you

icipe

International Centre of Insect Physiology and Ecology

P.O. Box 30772-00100, Nairobi, Kenya
Tel: +254 (20) 8632000
E-mail: icipe@icipe.org
Website: www.icipe.org

Support *icipe*: www.icipe.org/support-icipe

facebook.com/icipe.insects/icipe
twitter.com/icipe
linkedin.com/company/icipe