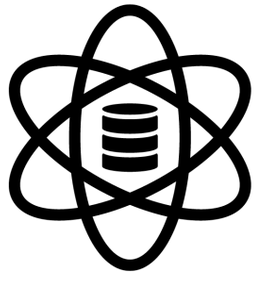
Introduction to Data Science

A Course for Natural Science Students at the Graduate Level



Raghunathan Ramakrishnan

Tata Institute of Fundamental Research Hyderabad, INDIA

https://www.tifrh.res.in/~ramakrishnan/

ramakrishnan@tifrh.res.in

Spring Term 2021

Content

Bibliography

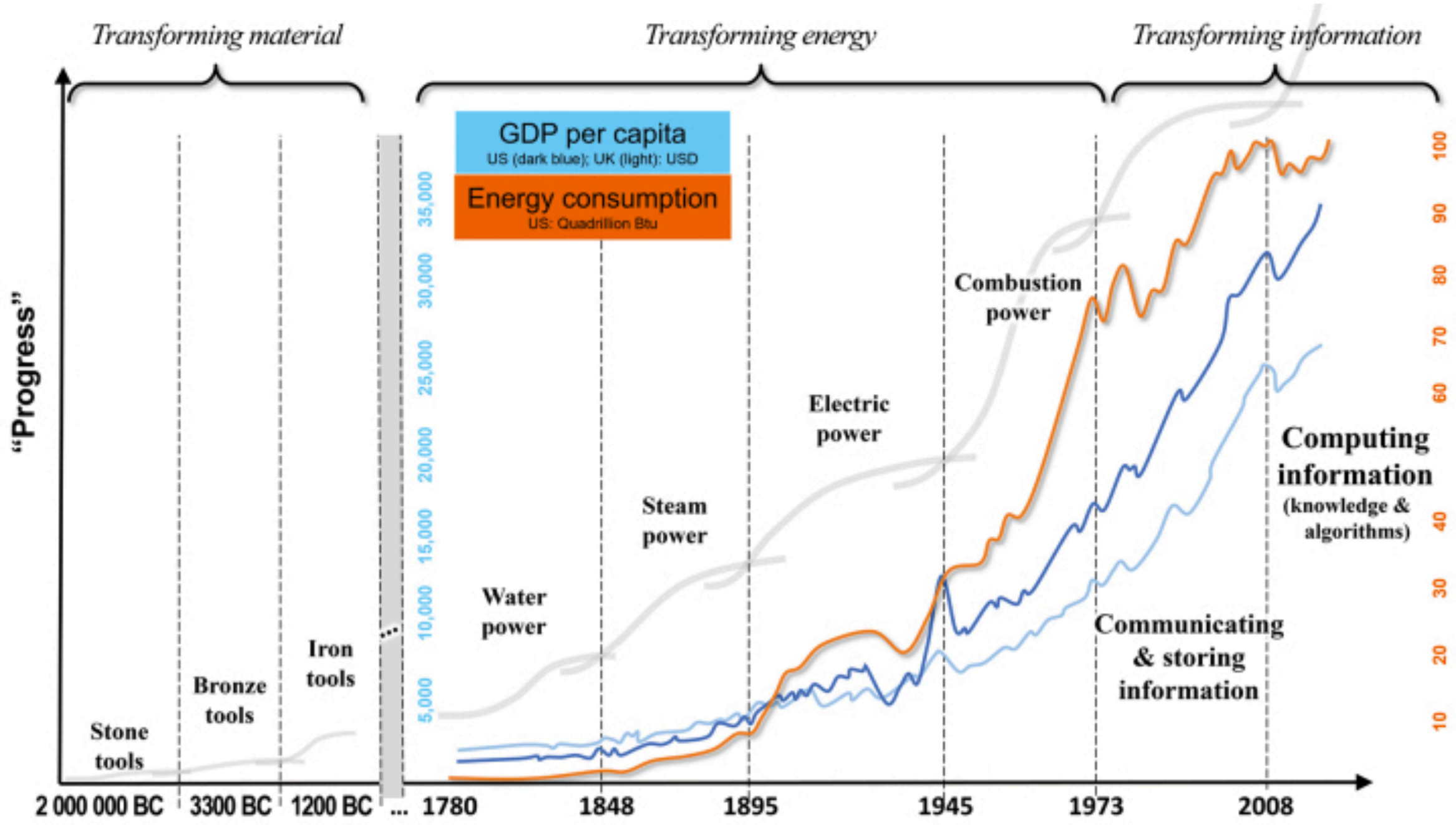
1. *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*, Laura Igual, Santi Segu, Springer (2017).
2. *Introducing Data Science*, Davy Cielen, Arno D. B. Meysman, Mohamed Ali, Manning (2016).
3. *Principles of Data Science: Learn the techniques and math you need to start making sense of your data*, Sinan Ozdemir, Packt (2016).
4. *Data Mining, Introductory and Advanced Concepts*, Margaret H. Dunham, Pearson (2003).
5. *Lecture Notes: Fundamentals of Big Data Analytics*, Rudolf Mathar (2019).
6. *Python Data Science Essentials*, Alberto Boschetti, Luca Boschetti, Packt (2015).
7. Data Science for Dummies, Lillian Pierson, John Wiley & Sons (2015).

Suggested Reading

1. Gewin, Virginia. *Data sharing: An open mind on open data*. *Nature* 529.7584 (2016): 117-119.

1. Basic concepts

From material age to industrial age to information age



Digital transformation of society

[Ref. https://doi.org/10.31887/dcns.2020.22.2/mhilbert]

Enter the data age

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m
tweets are sent every day
Twitter

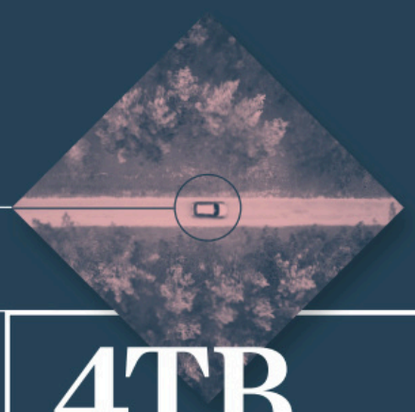
4PB
of data created by Facebook, including
350m photos
100m hours of video watch time
Facebook Research

320bn
emails to be sent each day by 2021

306bn
emails to be sent each day by 2020

294bn
billion emails are sent
Radicati Group

3.9bn
people use emails



4TB
of data produced by a connected car
Intel

65bn
messages sent over WhatsApp and two billion minutes of voice and video calls made
Facebook

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

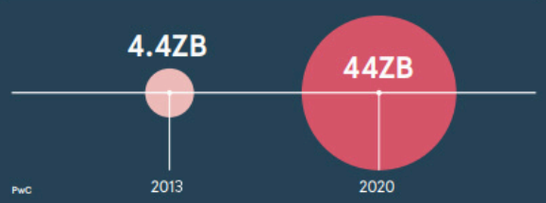
Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

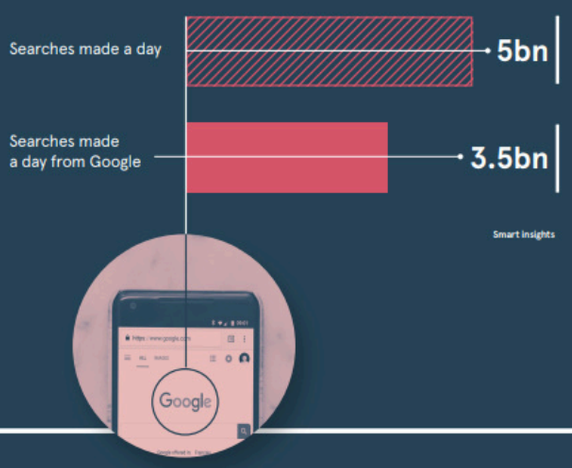
463EB
of data will be created every day by 2025
IDC

95m
photos and videos are shared on Instagram
Instagram Business

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC



Smart Insights

28PB
to be generated from wearable devices by 2020
Statista

RACONTEUR

[Ref. How much data is generated each day?]

Data initiatives in Chemistry/Materials Science

BERLIN BIG DATA CENTER

About Us / Mission

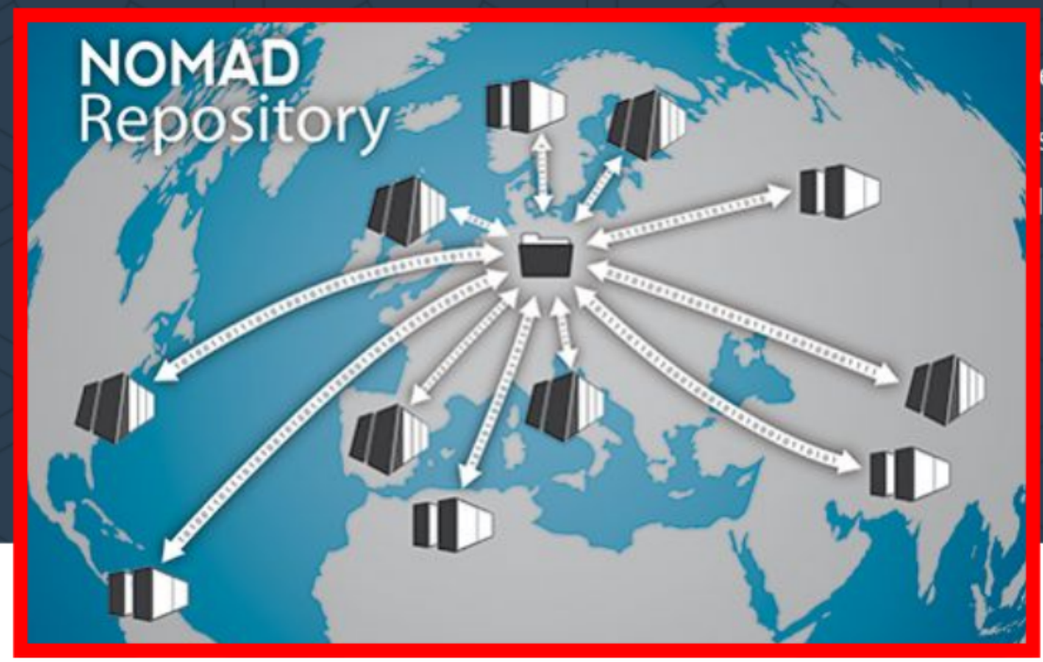
Mission

Goals

Consortium

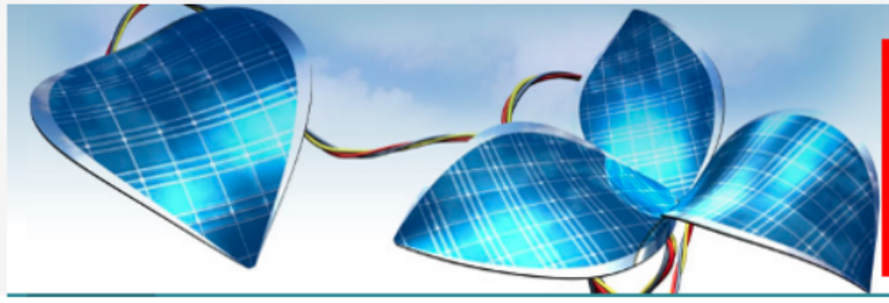
Funding

The Materials Project



ectronic
sed access to
l as powerful

Open *Materials Database*



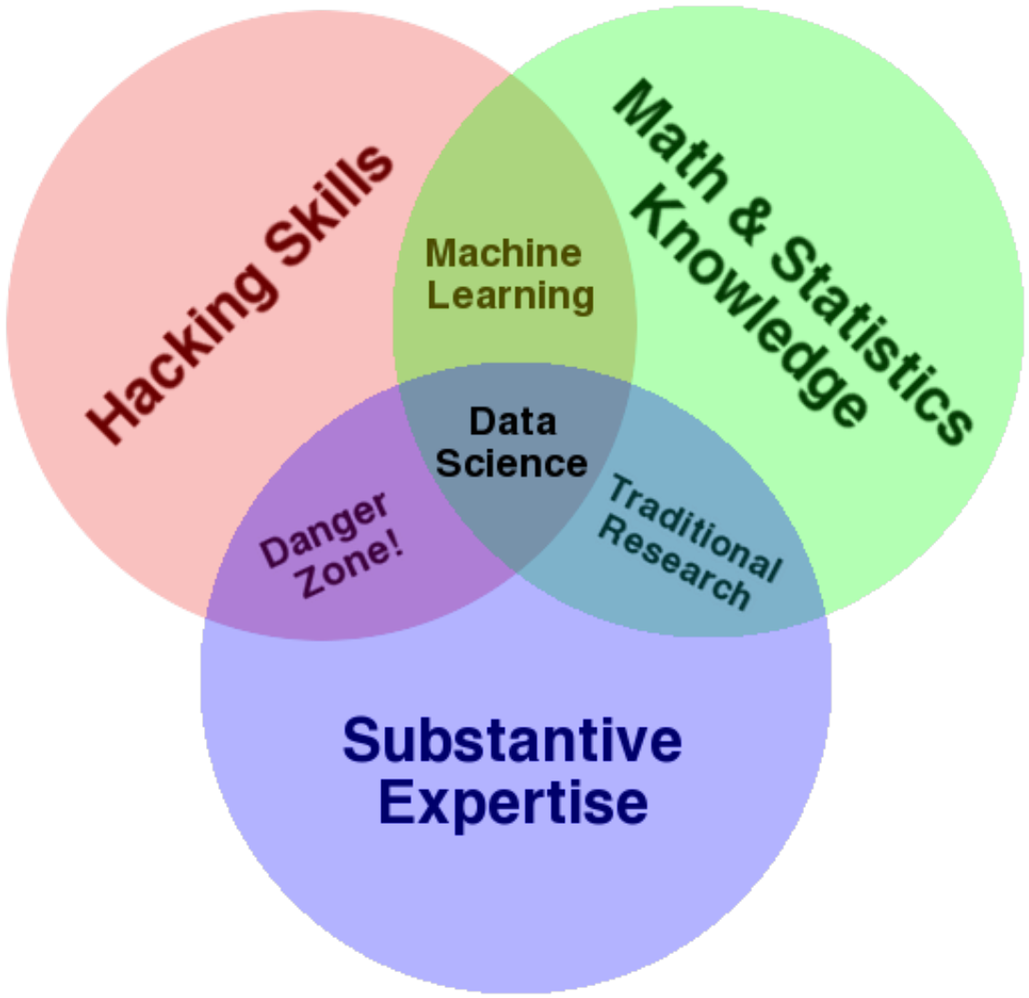
CEP – the Harvard Clean Energy Project

powered by world community grid.

MOLDIS

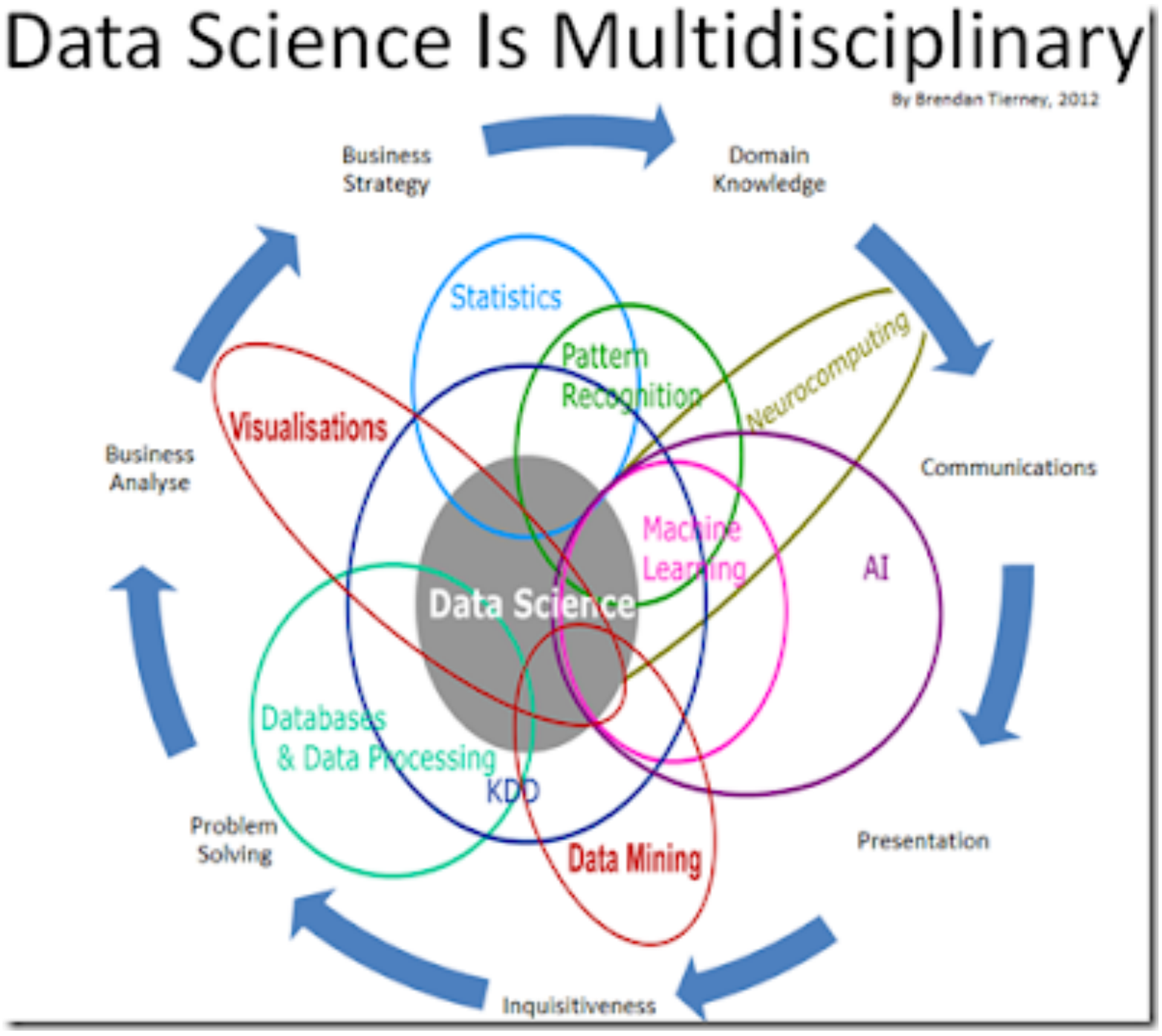
A big data analytics platform for molecular discovery

What is data science?



- In the danger zone: people "know enough to be dangerous"
- In this area people who are perfectly capable of extracting and structuring data, likely related to a field they know quite a bit about, and probably even know enough R to run a linear regression and report the coefficients; but they lack any understanding of what those coefficients mean.
- It is from this part of the diagram that the phrase "lies, damned lies, and statistics" emanates, because either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created.
- [Ref: The data science Venn diagram, 2010]

What is data science?



2012

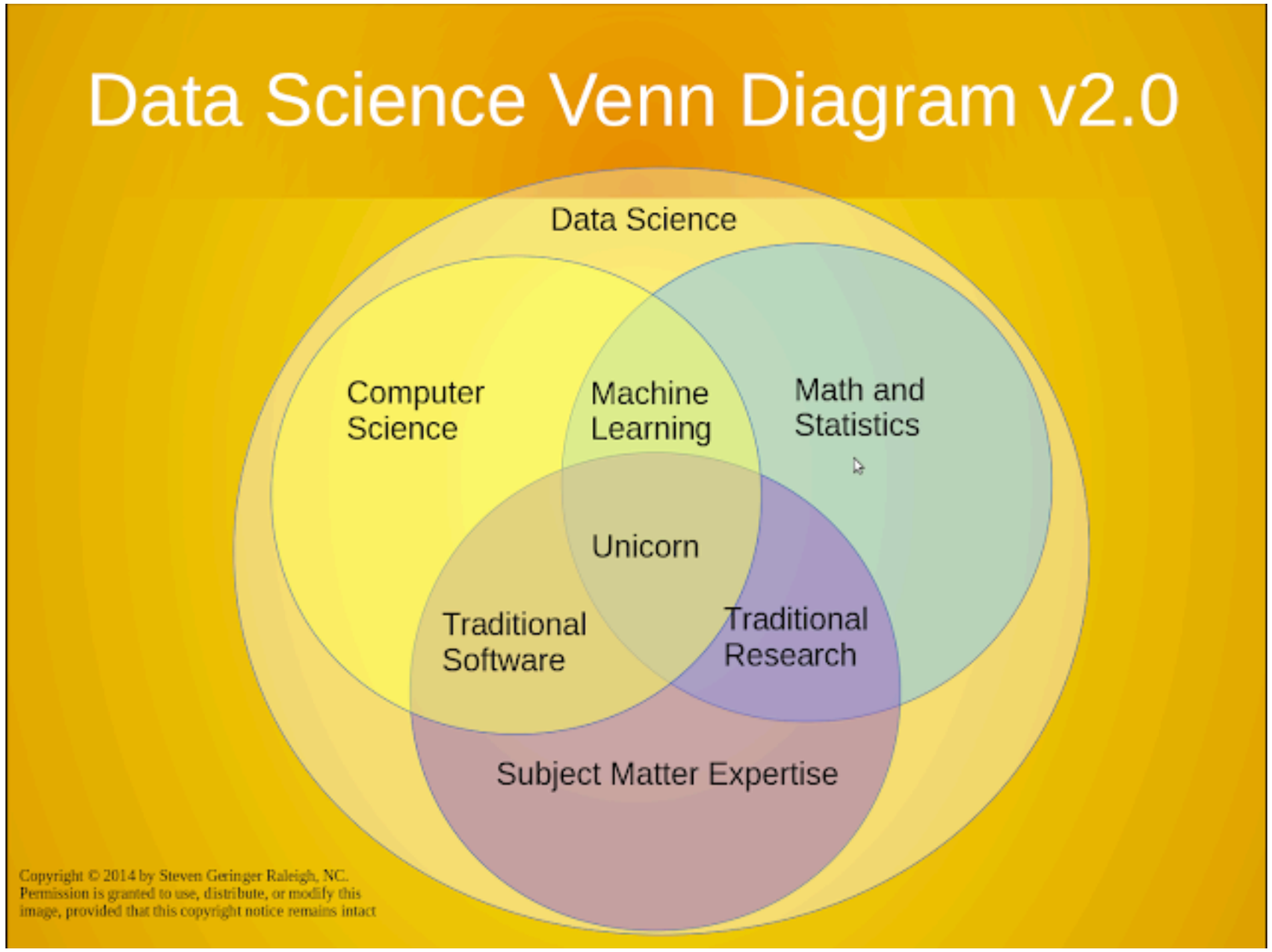
- [Ref: Battle of the data science Venn diagrams]

What is data science?

2013

- [Ref: Battle of the data science Venn diagrams]

What is data science?

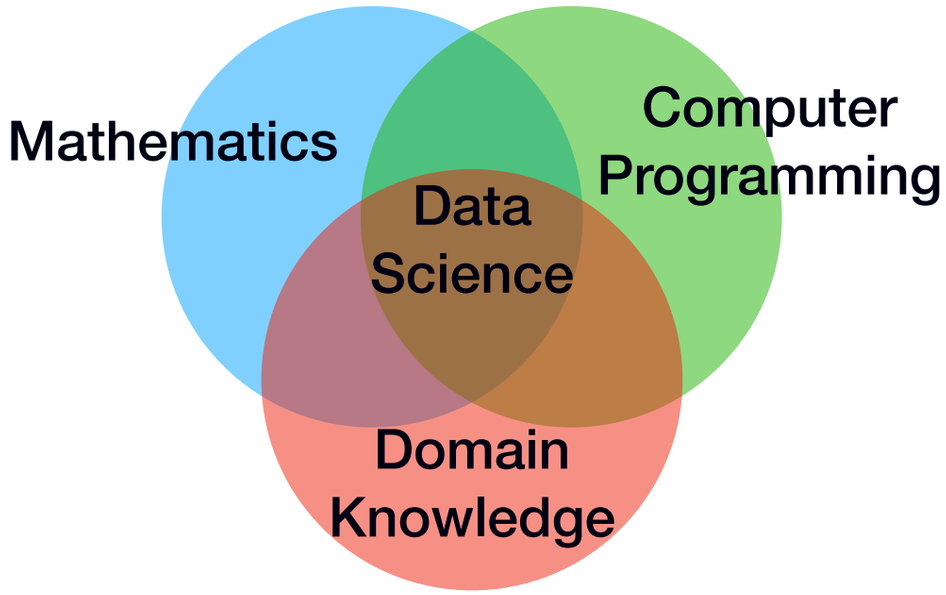


2014

- "Aren't all scientists, by definition, data scientists?" There are no sciences that do not deal in data
- [Ref: Battle of the data science Venn diagrams]

What data science really is

- Data Science is the science of acquiring knowledge from data
- Data Science is at the intersection of the three major areas: mathematics, computer programming, and domain/expert knowledge
- Domain/expert knowledge is the knowledge of the intended area of application of data science techniques. E.g. finance, weather, materials science, etc.



- Some popular applications of data science include: e-mail spam filtering, autocomplete, autocorrect, smart face lock, virtual assistant, chatbots, character recognition, finance, healthcare, robotics [Ref: tinyurl.com/sk3cbe4e] .
- Data Science is required in any fields with a surplus of data. The amount of data makes it impossible to extract meaningful information using conventional analysis.
- This course aims to present skills required for efficient use of data. Recall: In the 90's people have to take courses to learn how to use computers and word processors!
- Communicating data insights is crucial.
    1. Data visualisation techniques
    2. Data-driven story telling (case studies)
- This couse is for those who want to understand and utilize the basic practices of data science for any domain.

Mathematics

- Mathematical topics relevant to data science deal with data analysis and data modelling.
- These modelling techniques are founded upon statistical/probabilistic principles.
- For solving the equations involved, one require linear algebra concepts.
- The following topics will be discussed in this course

1. Statistics and Probability (e.g. measures, hypothesis tests, distribution functions)
2. Linear algebra (e.g. linear equations, eigenvalue problems)
3. Machine learning methods for data analysis and data modelling
    - Dimensionality reduction techniques (e.g. PCA)
    - Classification and clustering (e.g. *k*-means)
    - Regression methods (e.g. SVM, kernel-based, deep learning)

Computer programming

- Coding is a part of the game
- Proficiency in a computer programming language is important for handling the data and for building models.
- Python, R and Julia are preferred languages because of the availability of pre-built modules for data analysis, linear algebra, machine learning methods and visualisation.
- The following Python modules will be used in this course:
    1. Numpy and Scipy: Numerical and Scientific computation in Python
    2. SCIKIT-Learn: Machine learning in Python
    3. PANDAS: Python data analysis library
    4. Matplotlib: Python plotting

Role of domain knowledge

- Every area of human endeavour is forced to adopt data science
- Some of the titles of data scientists with expert knowledge:
    1. Ad Tech Data Scientist
    2. Banking Digital Analyst
    3. Clinical Data Scientist
    4. Geo-Engineer Data Scientist
    5. Retail Personalization Data Scientist
    6. Applied Materials Data Scientist
    7. Chemometrician/Chemical Data Scientist

Talk like a data scientist

Google and learn what the following terms mean

360 process, algorithms, analysts, analytics base table, anomaly detection, API, association-rule mining, attribute, Azure, back-propagation, bias, Big Data, bootstrap, captured data, classification, cloud services, clustering, combinatorics, confusion matrix, correlation, CRISP-DM, CSV, data, database, datasets, data warehouse, decision tree, deep learning, derived attribute, DIKW pyramid, e-commerce, exhaust data, ETL, feature selection, GitHub, Hadoop, HPC, instance, Jupyter, Kaggle, linear regression, ML, MOOCs, MPP, metadata, neural network, ODS, OLAP, PCA, p-value, Python, quants, R, RDBMS, regression, selection bias, SQL, test set, training set