Introduction to ‘Data Science’

Pedro Luis do Nascimento Silva

ENCE/IBGE

pedronsilva@gmail.com

‘Data’ + ‘Science’

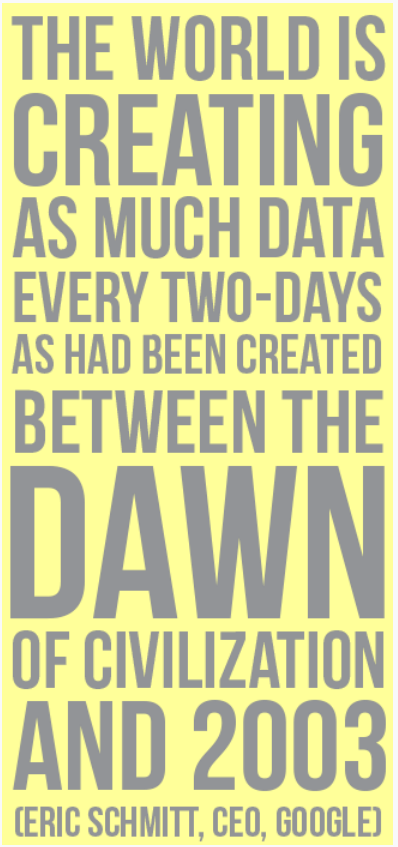
The ‘Data Era’ (Big or Small)

We live in an era with unprecedented availability and access to **data**.

The 'Data Era' (Big or Small)

We live in an era with unprecedented availability and access to **data**.

Global Partnership for Sustainable Development Data (GPSDD)
http://www.data4sdgs.org/#news



The 'Data Era' (Big or Small)

We live in an era with unprecedented availability and access to **data**.

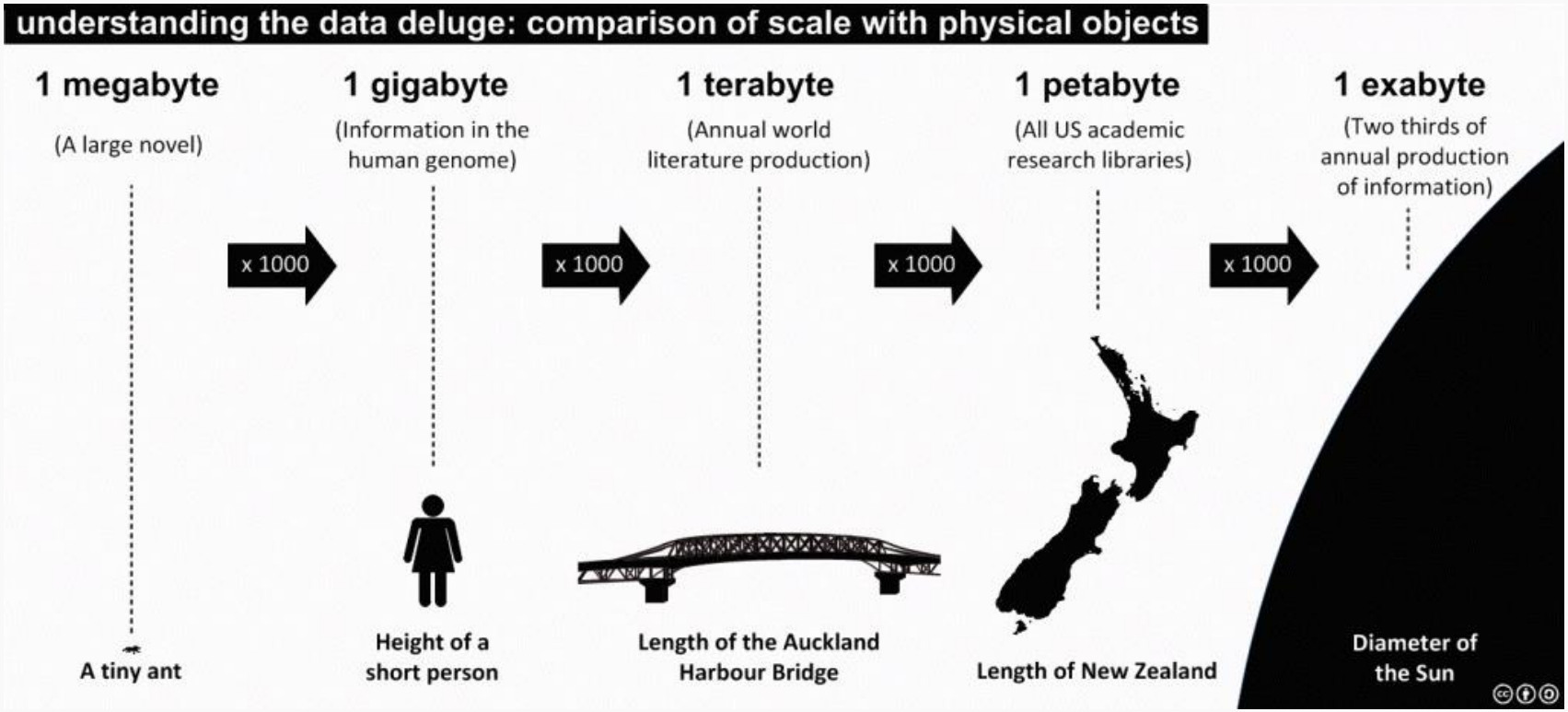
"Data in the world is doubling every 18 months."

IBM

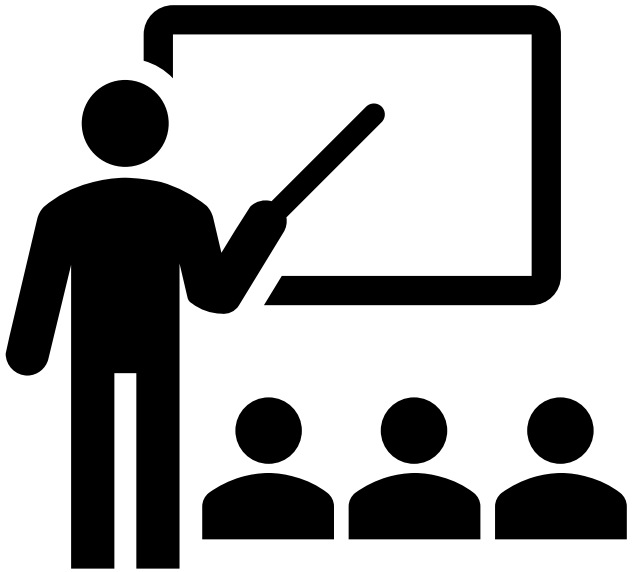
http://www-01.ibm.com/software/data/demystifying-big-data/

The 'Data Era' (Big or Small)

We live in an era with unprecedented availability and access to **data**.



© Barry Ritholtz, 2016



How we learn is also changing fast.

Science (= *Scientia* = Knowledge)

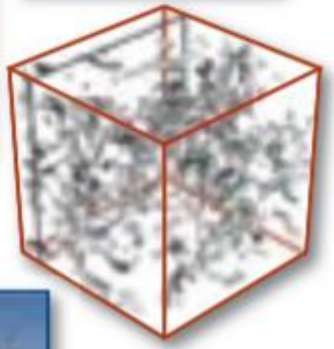
Science is a systematic enterprise that **builds and organizes knowledge** in the form of testable explanations and predictions about the universe.

https://en.wikipedia.org/wiki/Science

Science Paradigms

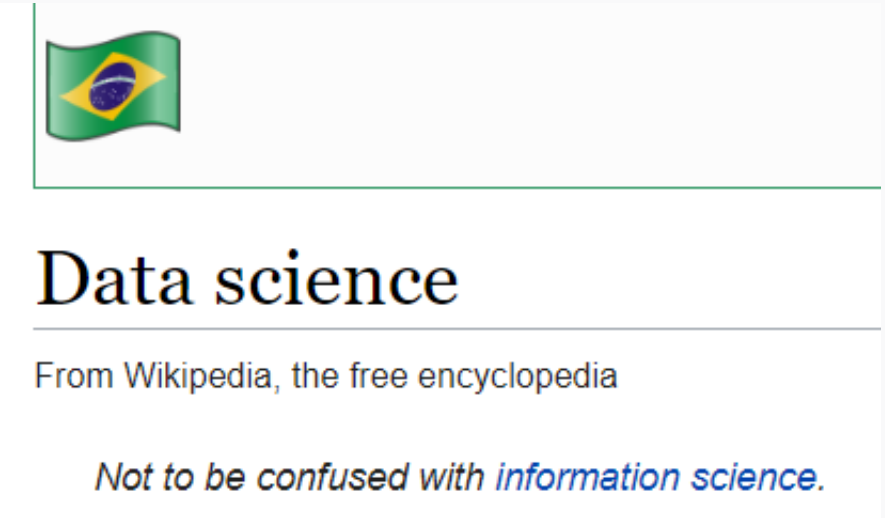
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Source: Gray, 2009.

Some Definitions



Data science

From Wikipedia, the free encyclopedia

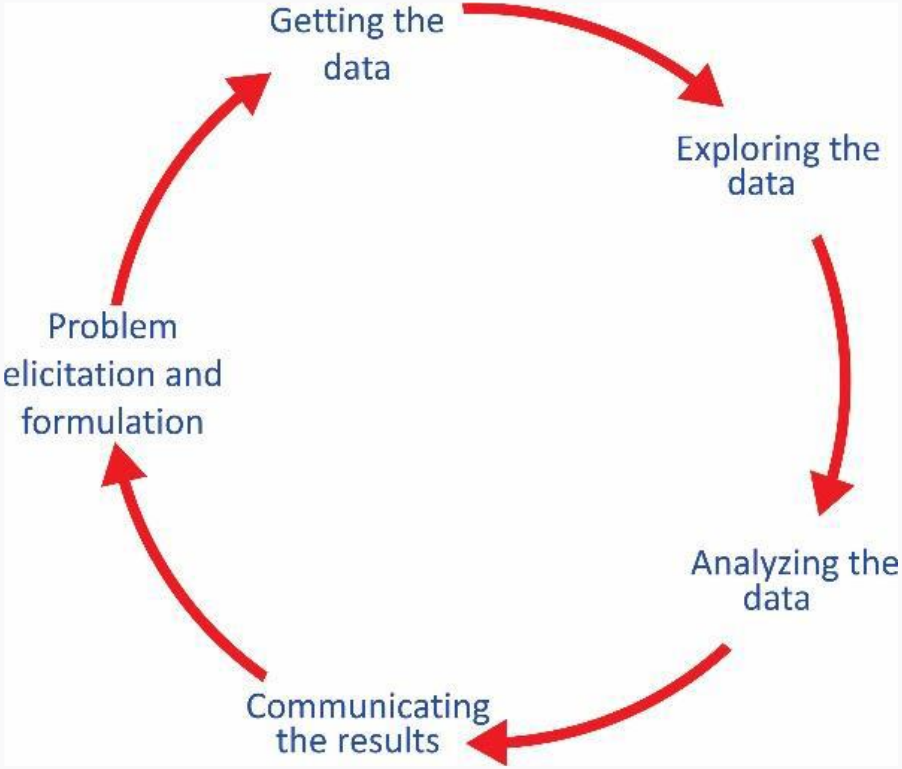
Not to be confused with information science.

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

https://en.wikipedia.org/wiki/Data_science

Data Science

Data Science *is the science of learning from data.*

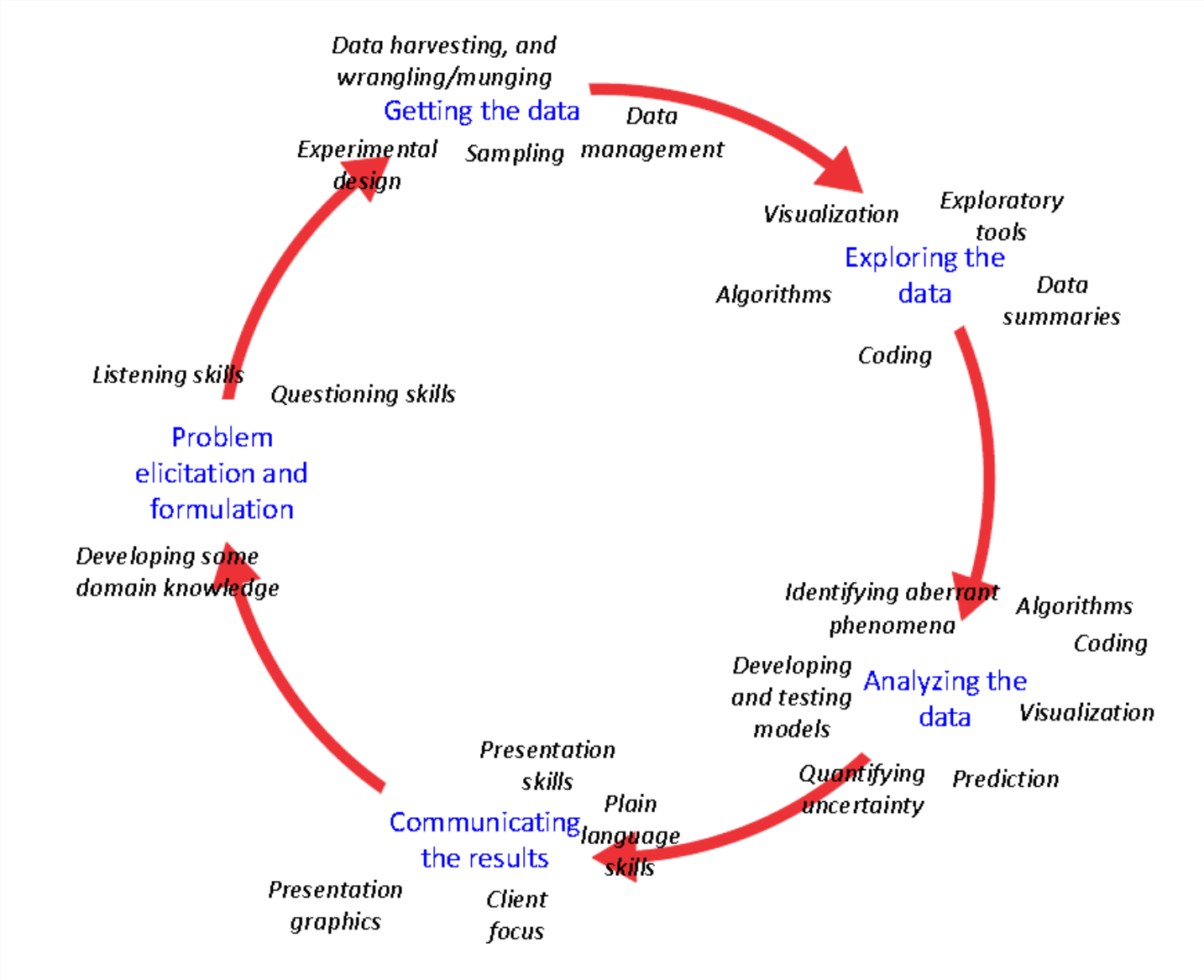


It draws on several disciplines, including Computer Science, Mathematics and Statistics, together with areas such as problem elicitation and formulation, collaboration and communication skills.

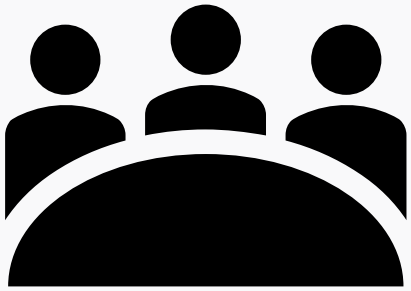
IDSSP(2009).

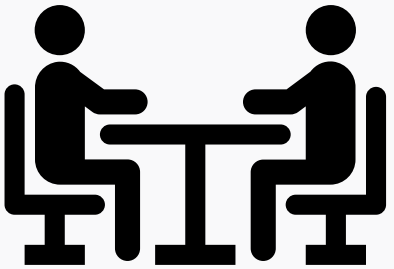
Data Science

Data Science *is the science of learning with data.*



IDSSP(2009).





Defining a new science.

Emergence of a ‘Data Science’

Donoho (2017) credits Tukey (1962) for defining and anticipating what we would now call ‘Data Science’.

According to Tukey, a ‘science’ requires three constituents:

a) “Intellectual content;

b) Organization in an understandable form;

c) Reliance upon the test of experience as the ultimate standard of validity.”

Emergence of a ‘Data Science’

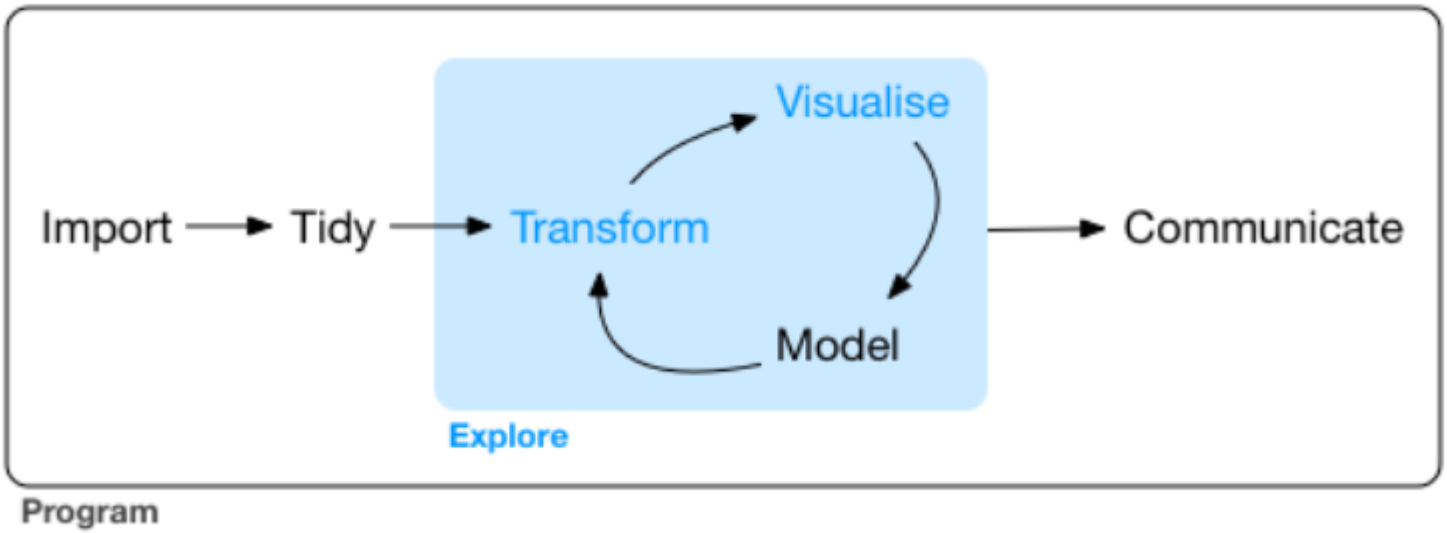
“... a science does not just spring into existence simply because a **deluge of data** will soon be filling telecom servers, and because some administrators think they can sense the resulting trends in hiring and government funding.”

“Fortunately, there is a solid case for some entity called ‘data science’ to be created, which would be a **true science**: facing essential questions of a lasting nature and using scientifically rigorous techniques to attack those questions.”

Donoho (2017)

Data Science Activities

The activities of DS according to Wickham & Grolemund (2017).



Data Science Activities

The activities of DS are classified into six divisions:

1. Data Gathering, Preparation, and Exploration;
2. Data Representation and Transformation;
3. Computing with Data;
4. Data Modeling;
5. Data Visualization and Presentation;
6. Science about Data Science.

Donoho (2017)

Science about Data Science

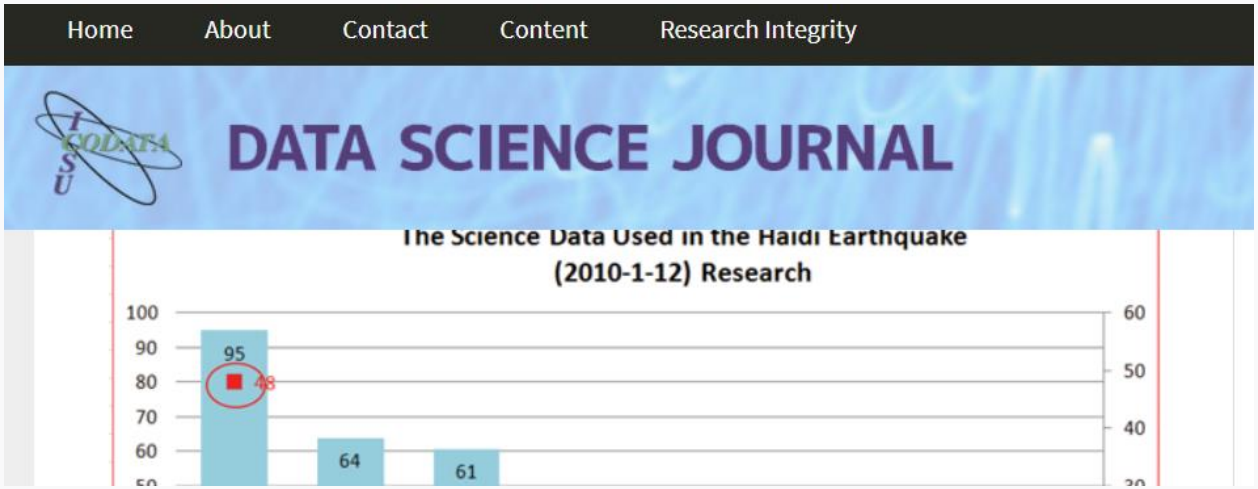
Data scientists are doing science about data science when they:

- Identify commonly occurring analysis / processing workflows;
- Measure the effectiveness of standard workflows in terms of the human time, the computing resource, the analysis validity, or other performance metric;
- Uncover emergent phenomena in data analysis, for example, new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results;
- Work to make future such science possible — such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta-analysis.

Donoho (2017)

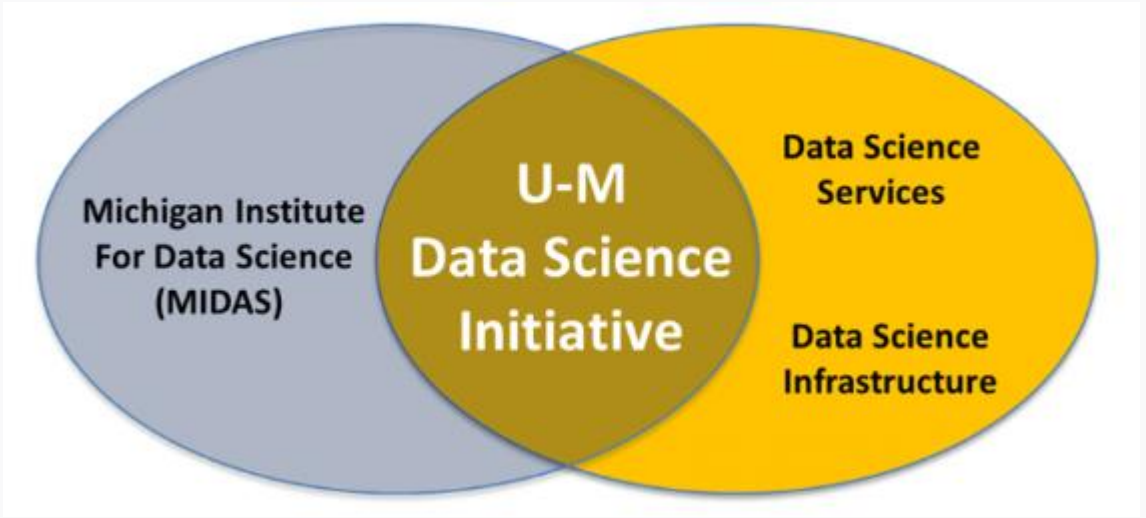
Discussion

Data Science has carved its place in the **scientific community**.



Discussion

Data Science has carved its place in the **educational arena**.



Discussion

Data Science has carved its place in the **job market**.

Search key	Hits (millions)
business analyst	443
business analyst jobs	416
data scientist	288
data scientist jobs	200
statistical analyst	157
statistical analyst jobs	64
predictive analytics	78
predictive analytics jobs	38
statistician	10
statistician jobs	6

Summing up

Data science is here to stay.

We must **engage** to benefit from emerging **opportunities** for:

- Research;
- Education;
- Applications & jobs

and to help meet the many **challenges** posed by the 'Data Era' and the emerging scientific paradigms of eScience or 'data centric' discovery.

Thanks for your attention!

pedronsilva@gmail.com

References

1. DONOHO, David. 50 Years of Data Science. Journal of Computational and Graphical Statistics v. 26, n. 4, p. 745–766 , 2017. Disponível em: <https://doi.org/10.1080/10618600.2017.1384734>.0672326965.
2. IDSSP: the International Data Science in Schools Project. Draft Curriculum Framework. 2019.
3. Jim Gray on eScience: A Transformed Scientific Method. Edited by HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. In: The Fourth Paradigm: Data-Intensive Scientific Discovery. REDMOND, WASHINGTON: Microsoft Research, 2009. 252 p.
4. TUKEY, J.W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.
5. WICKHAM, Hadley; GROLEMUND, Garrett. R for Data Science: Import, Tidy, Transform, Visualize and Model Data. Boston: O'Reilly, 2017. 492 p.