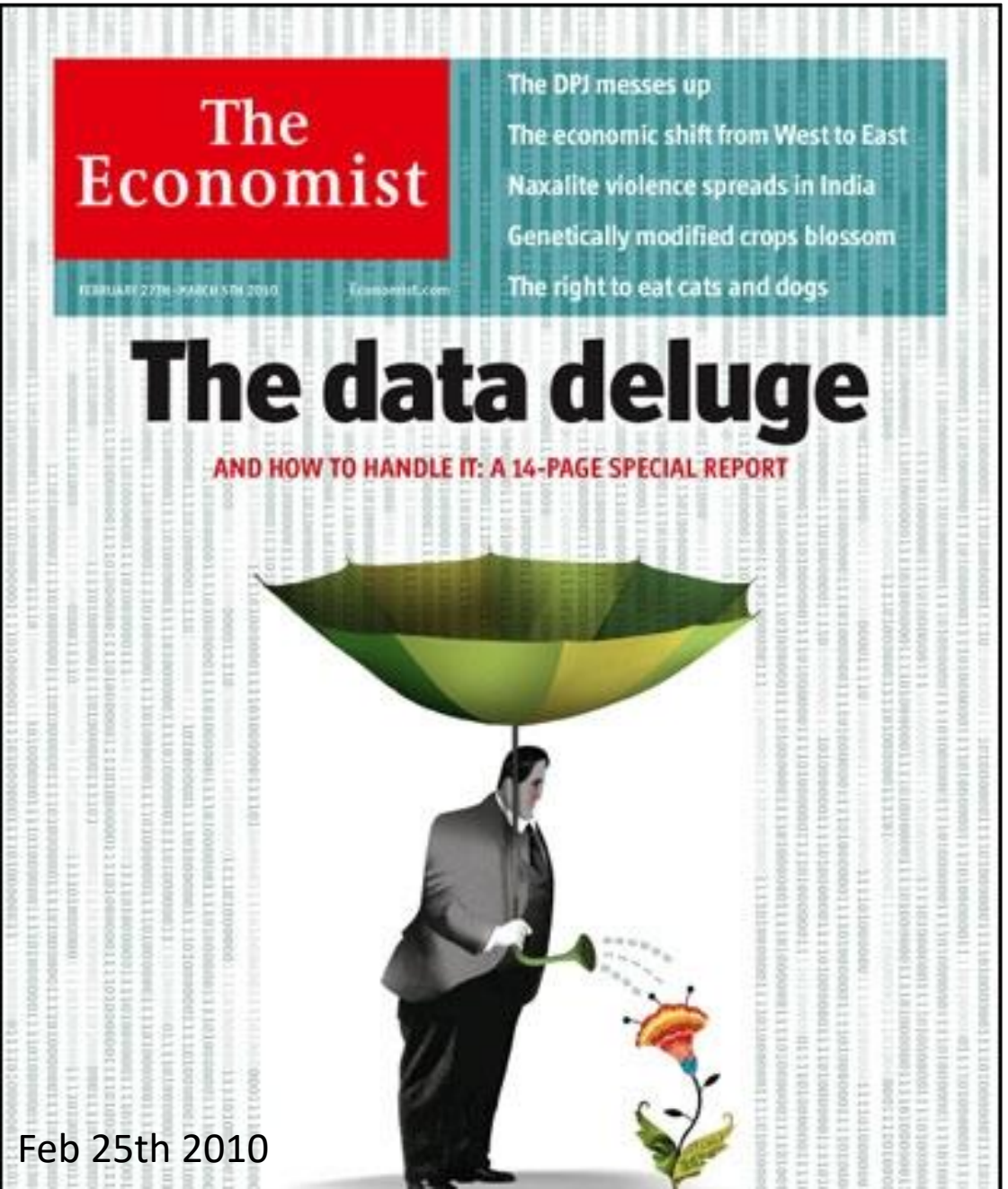
Big Data, Scientific Computing, & Data Analytics

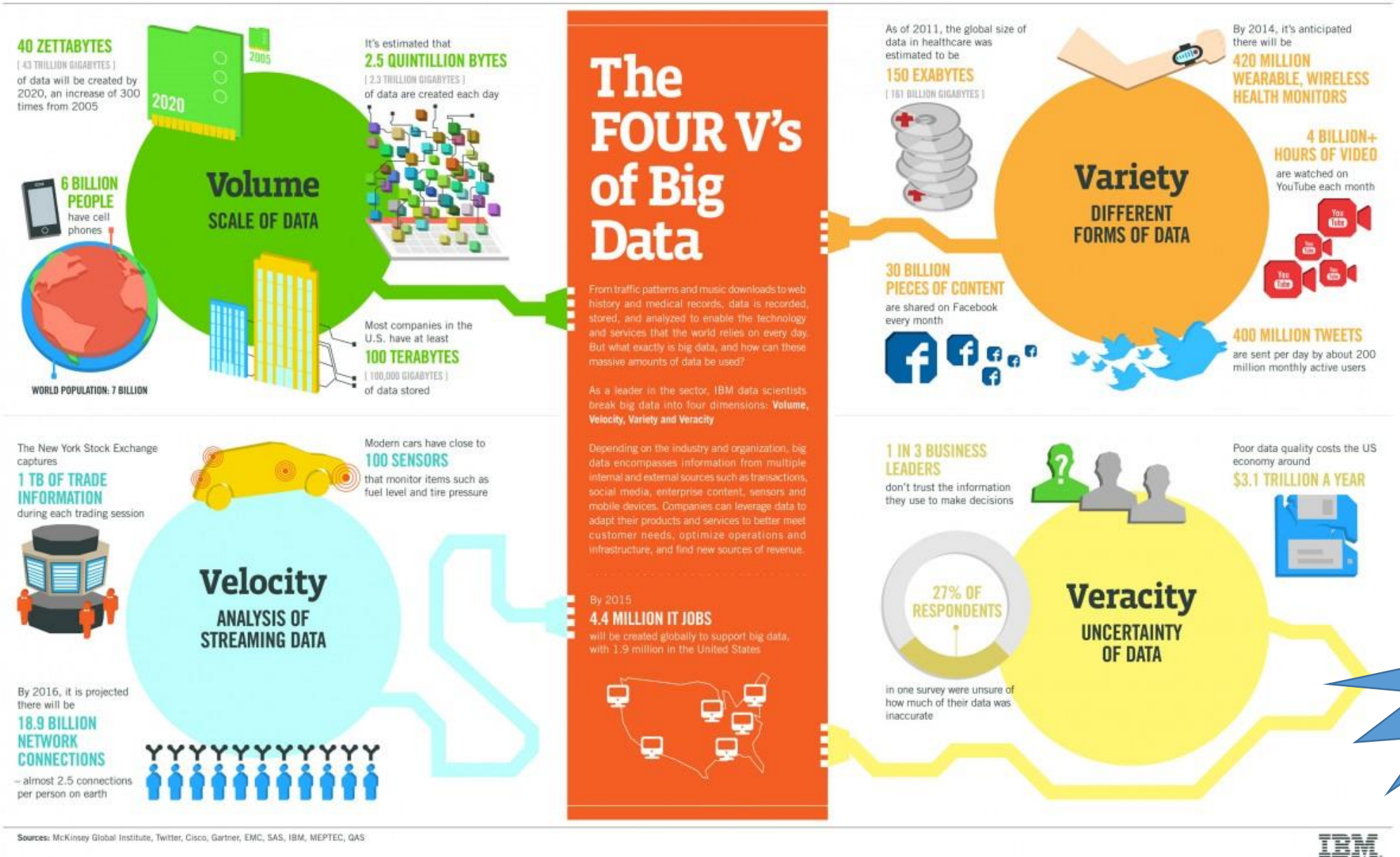
John Fay

23 February, 2018



Feb 25th 2010

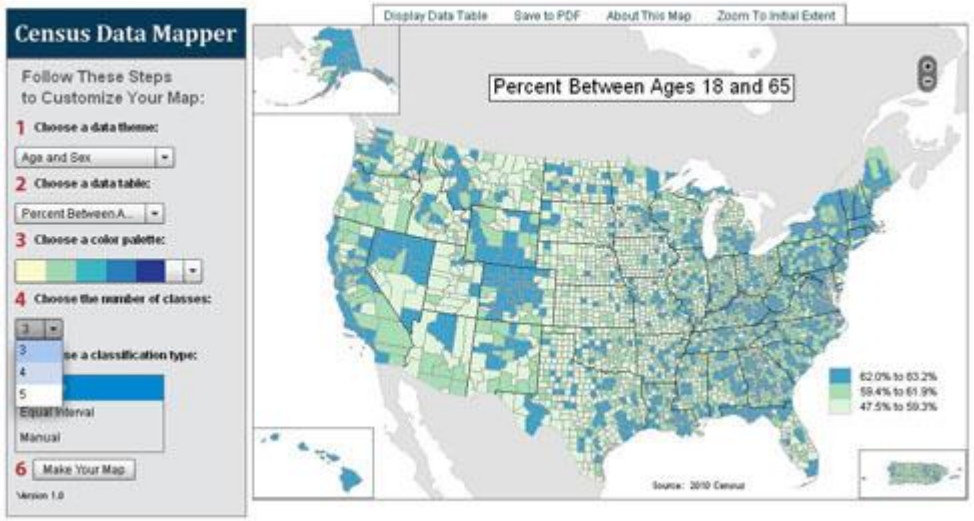
What is Big Data?

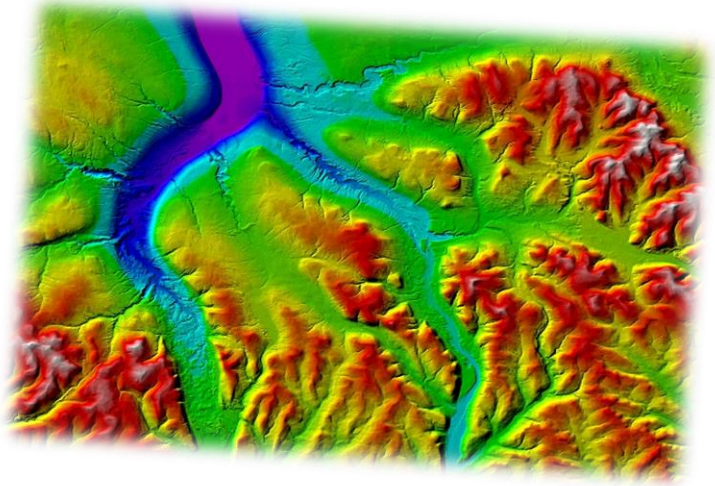


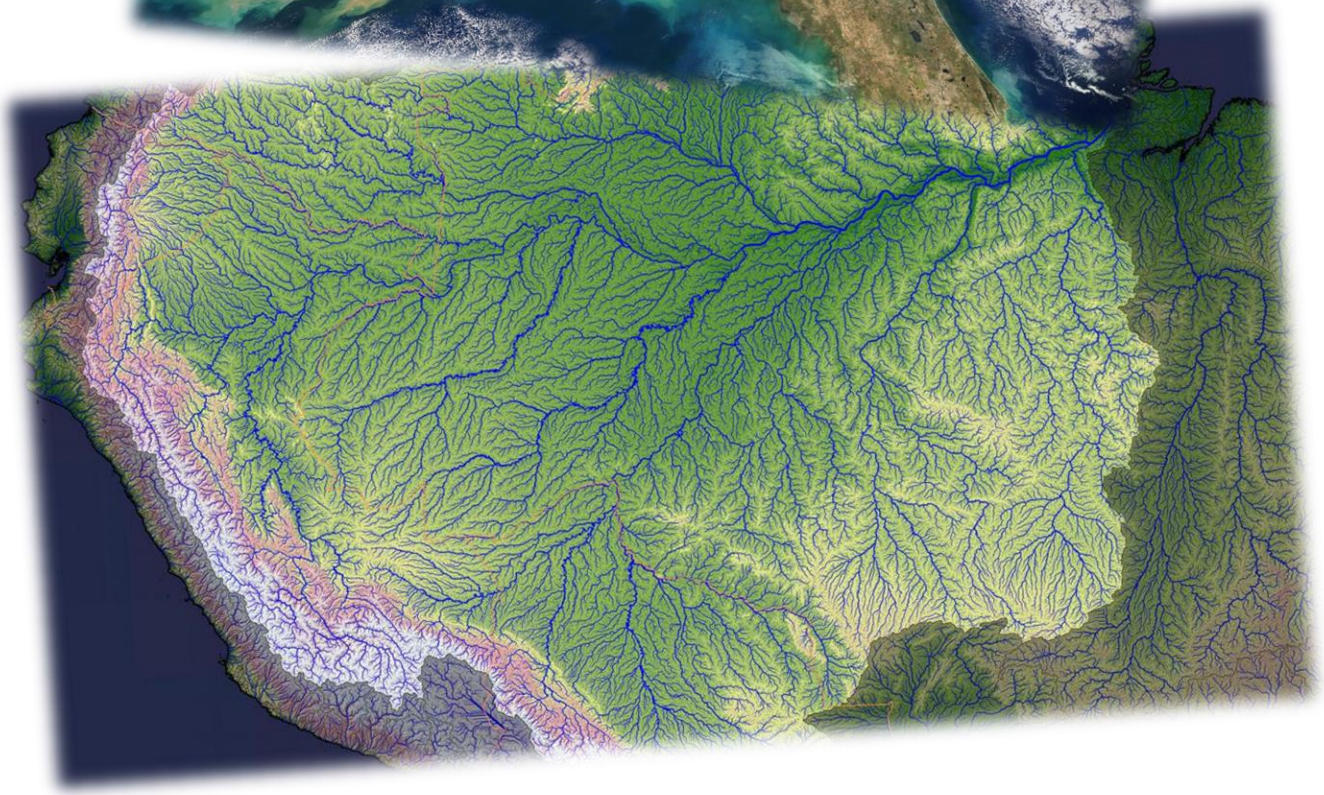
& Value

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Big Data *(is not new…)*







THE CLIMATE CORPORATION

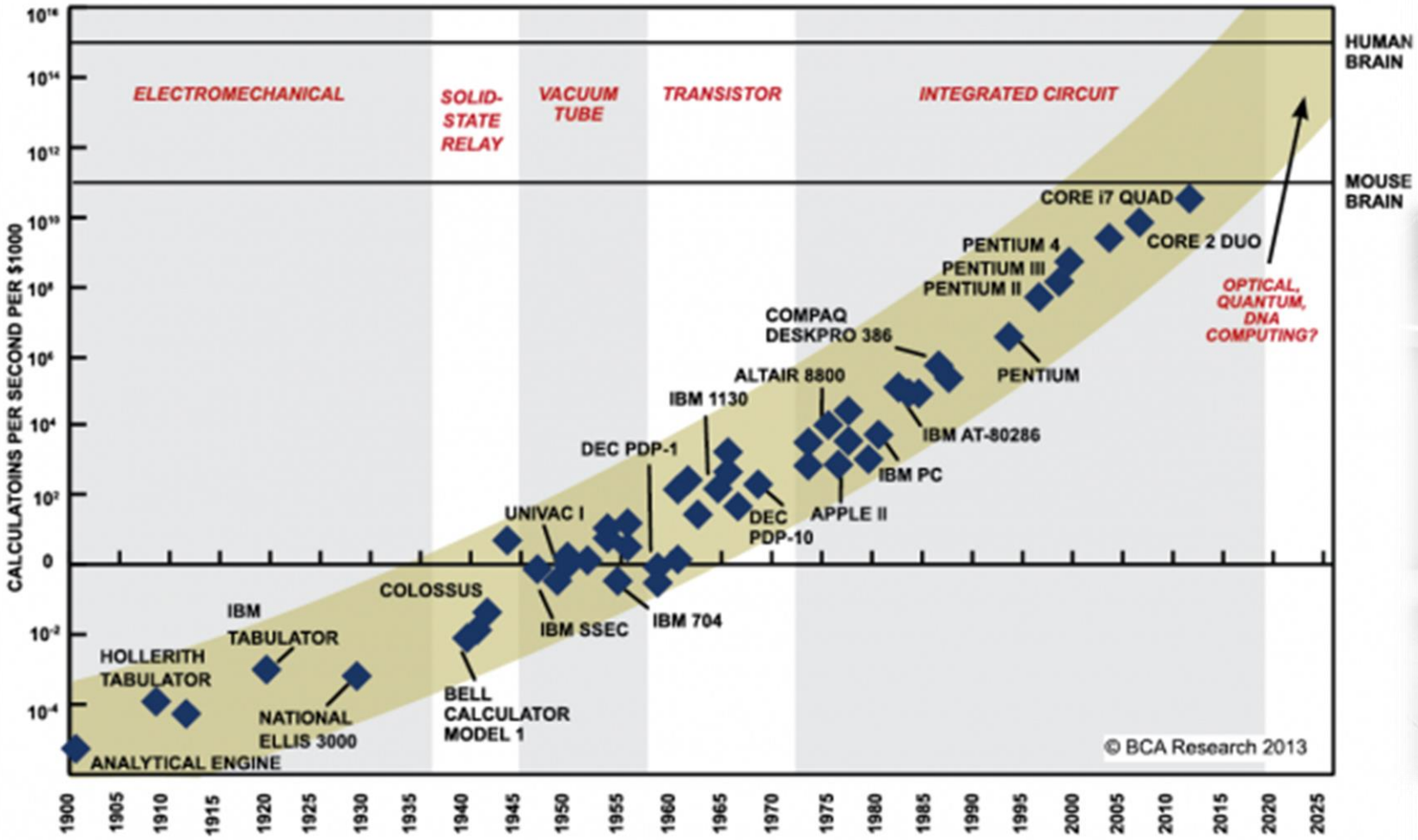
https://en.wikipedia.org/wiki/The_Climate_Corporation

Big Data…

What are some examples of "Big Data" that might be useful to business & environment issues?

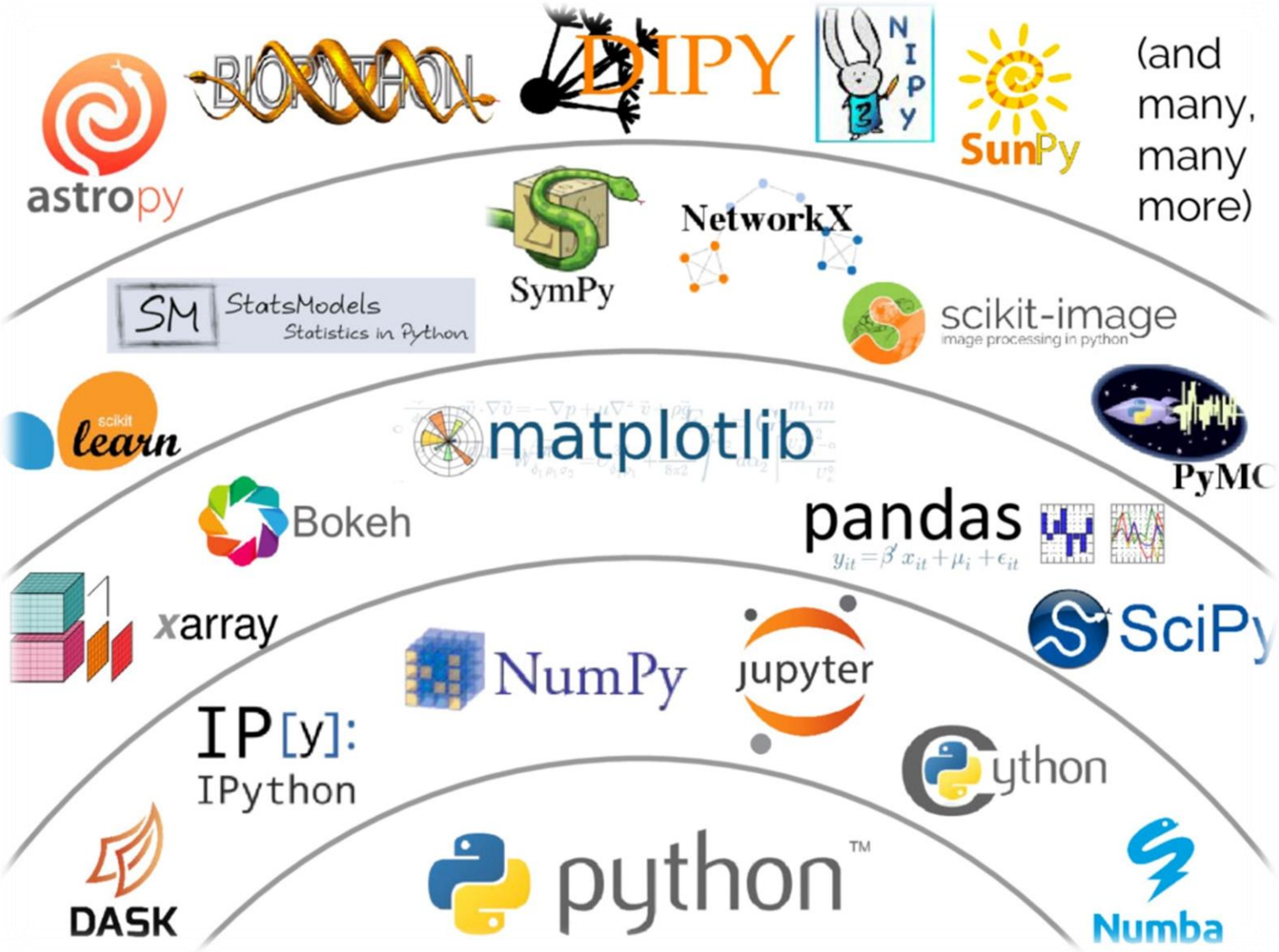
What is "Scientific Computing"?

Moore's Law



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

“Scientific Computing”…



“Scientific Computing”…

GitHub

PostgreSQL

Microsoft SQL Server

ORACLE

tableau

D3

R

python™

What is “Data Science?”

DATA

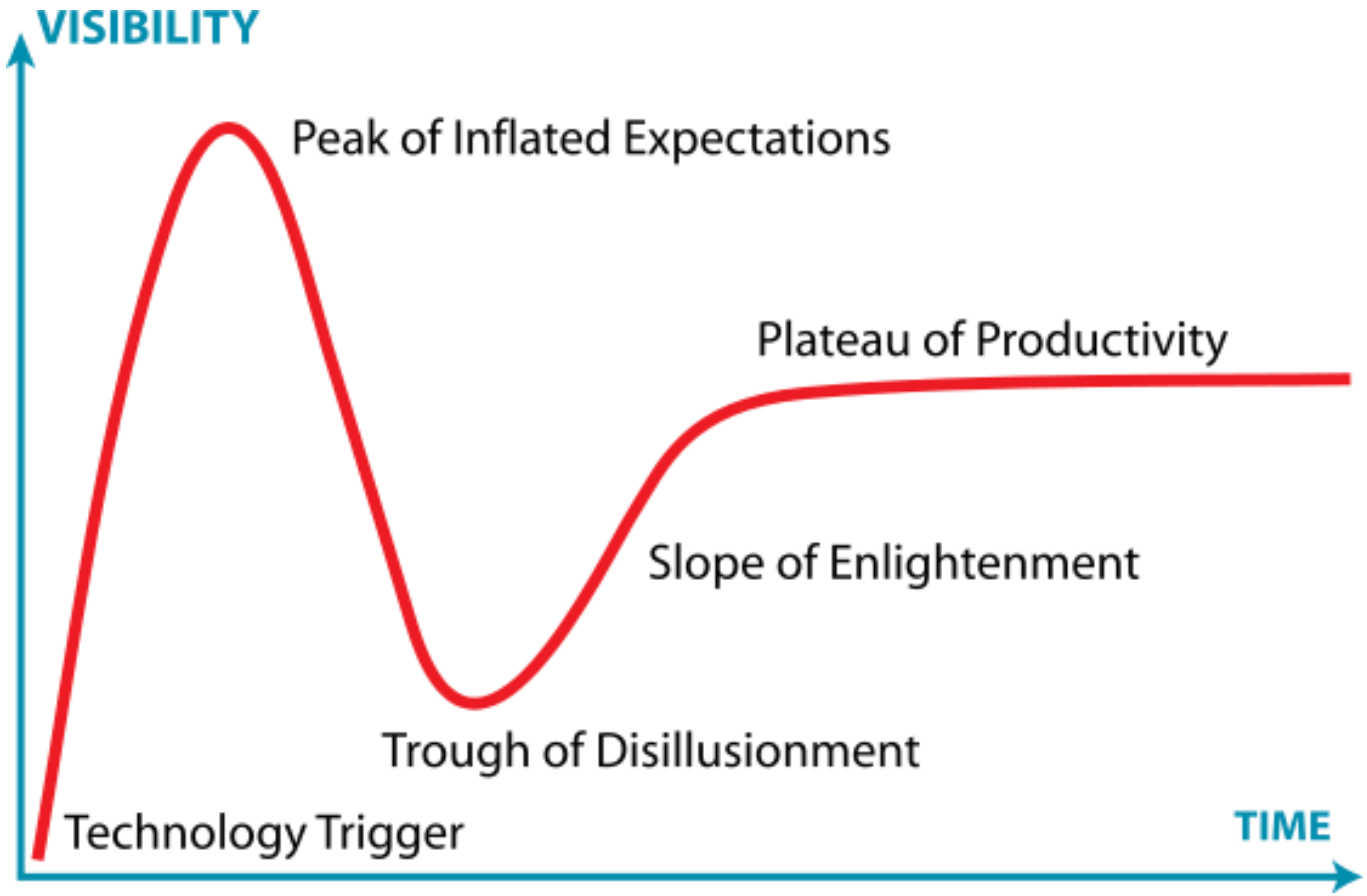
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

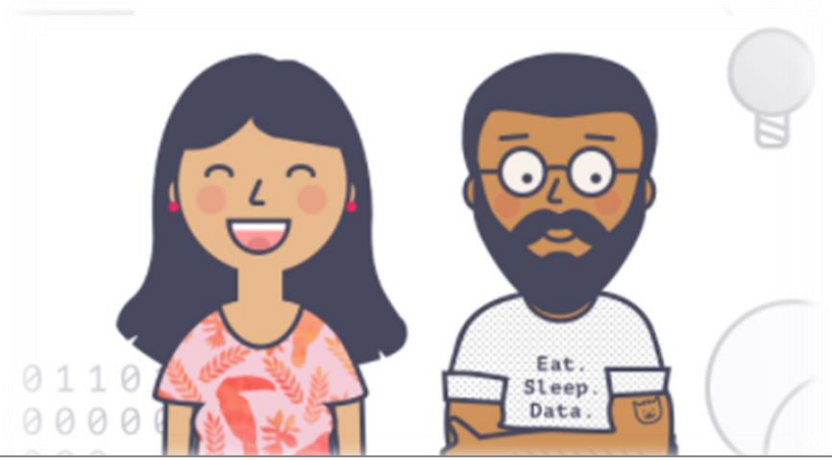
FROM THE OCTOBER 2012 ISSUE

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

What is "Data Science?"

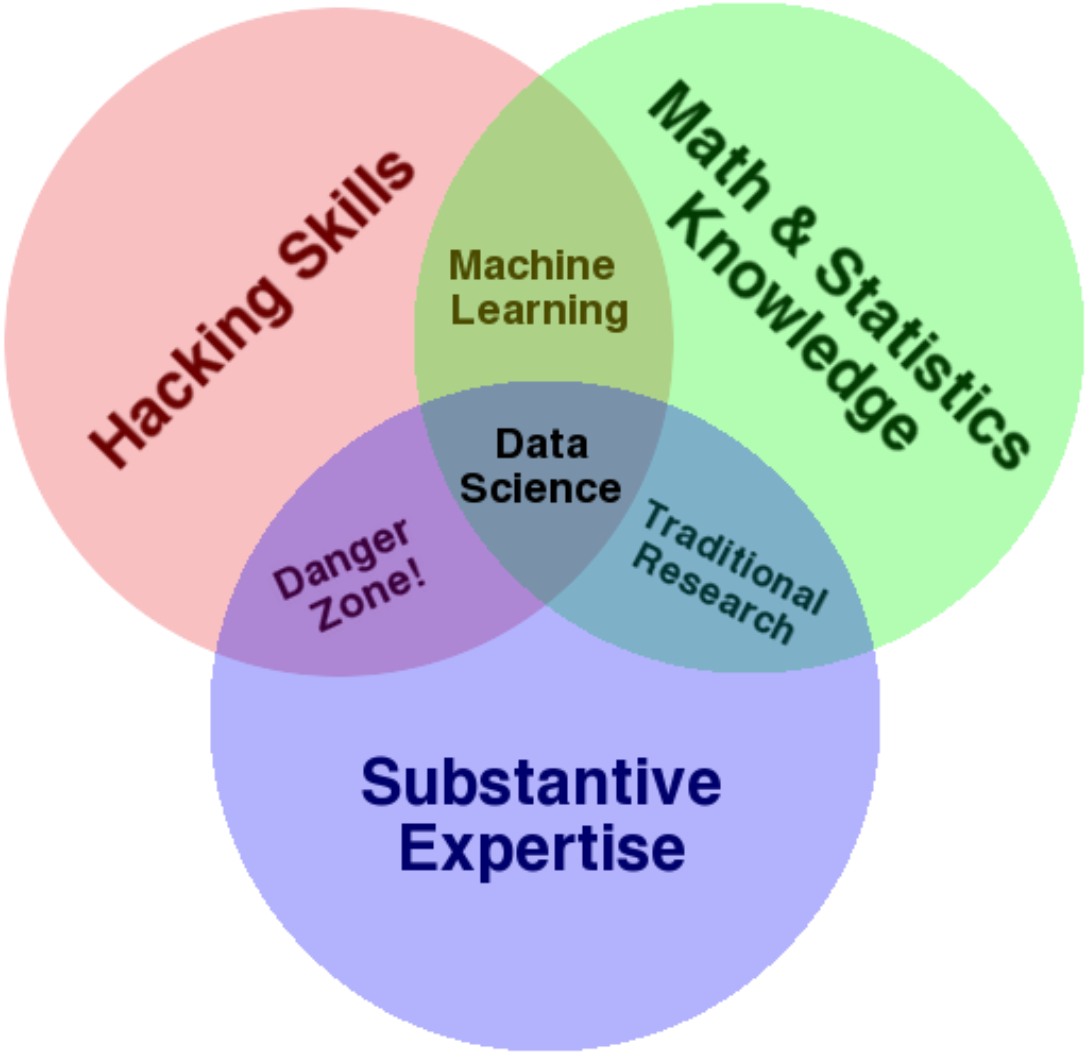


What is "Data Science??"



A data scientist:

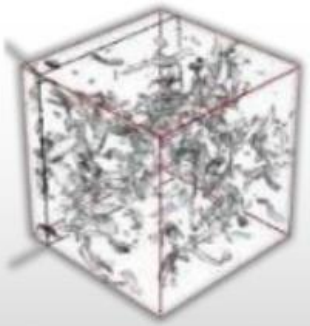
"Someone who knows more stats than a computer scientist and more computer science than a statistician"



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The 4th Paradigm for Scientific Discovery

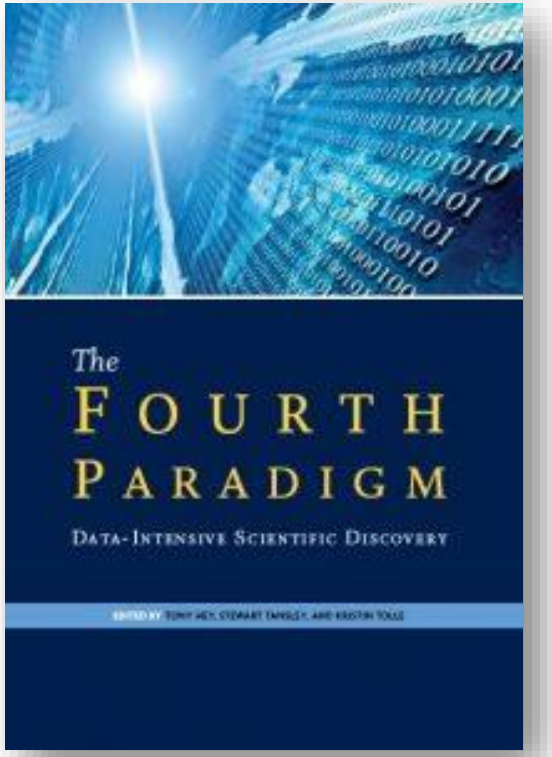
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton's laws, Maxwell's equations...*	*Simulation of complex phenomena*	*Unify theory, experiment and simulation with large multidisciplinary Data* *Using data exploration and data mining (from instruments, sensors, humans...)* *Distributed Communities*

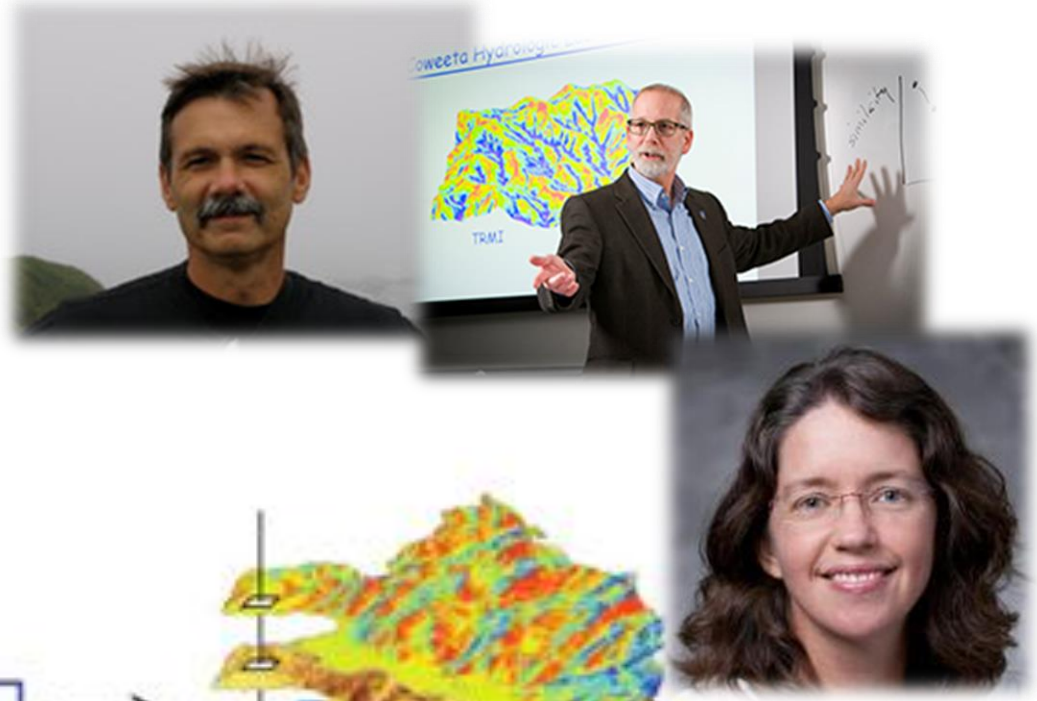
Credit: Dennis Gannon

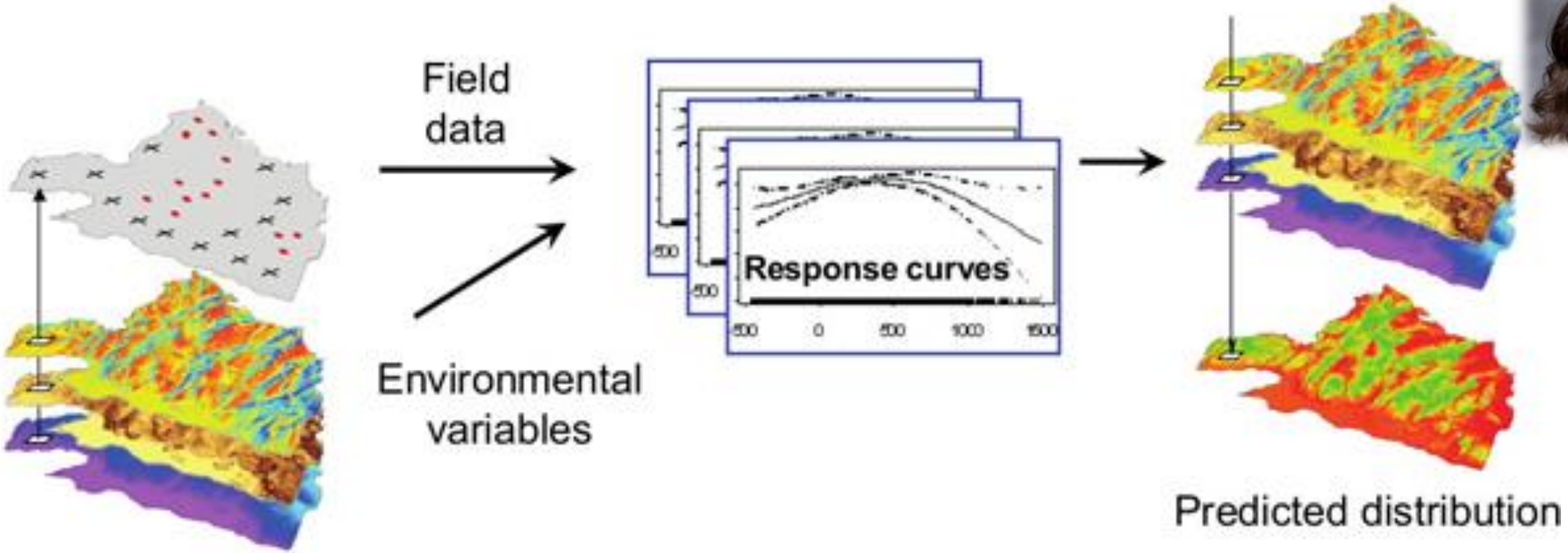
What is "Data Science?!"



Some examples…

Species distribution modeling

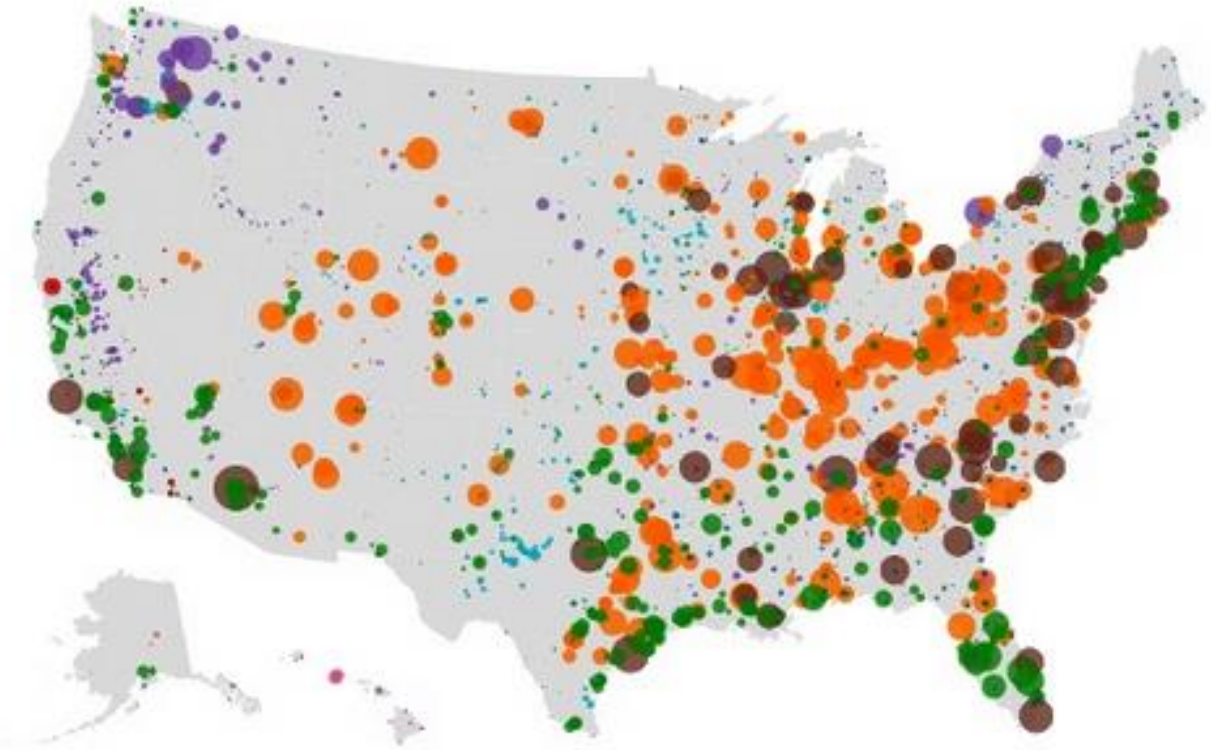




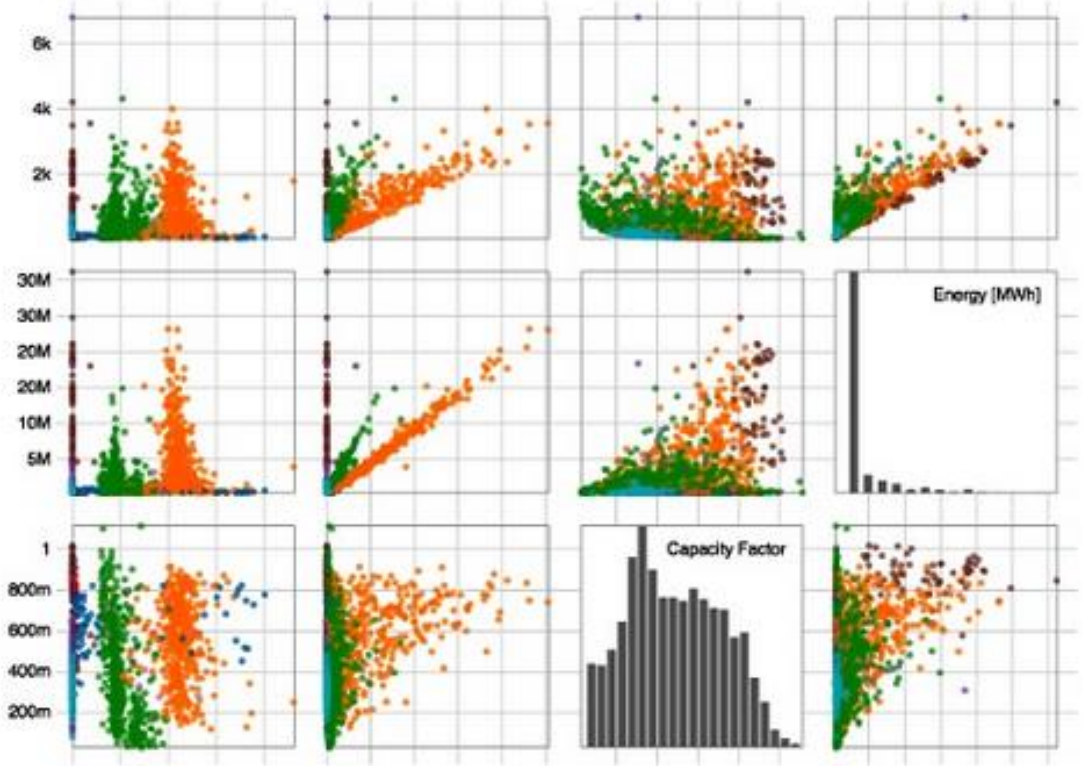
Some examples…

Energy Data Analytics

U.S. Electricity Generation: An Interactive Map



U.S. Electricity Generation: Comparing Resources



Some examples…

The internet of water

WATER DATA

INTERNET OF WATER

A network of interconnected data shared between different water sectors and regions will enable the real-time transmission of water-related data and information to more efficiently and sustainably manage water resources.

ENABLE OPEN WATER

Quantify, document and communicate the value of open, shared and integrated water data to build the business case for investing in making water data open and shareable.

INTEGRATE EXISTING PUBLIC WATER DATA

There are already some water data sharing communities integrating existing public water data; these efforts should be further supported with lessons and tools shared between these (and new) communities.

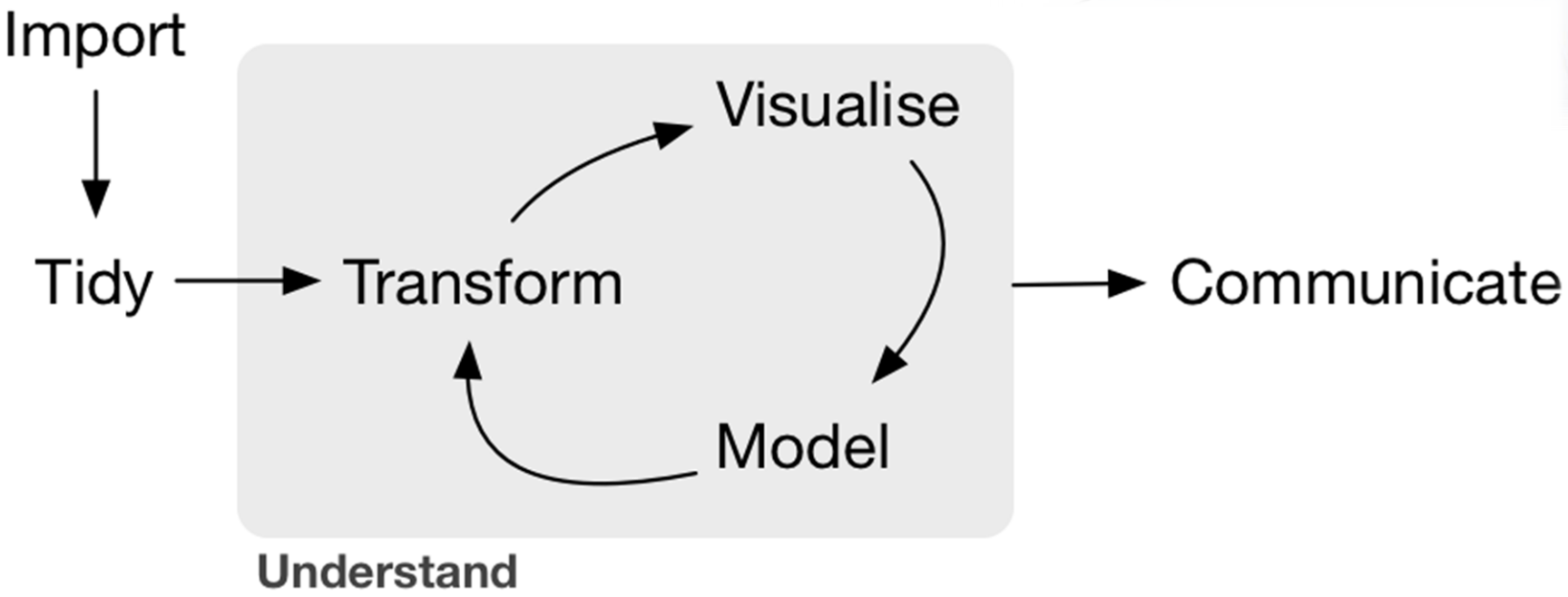
CONNECT REGIONAL DATA SHARING COMMUNITIES

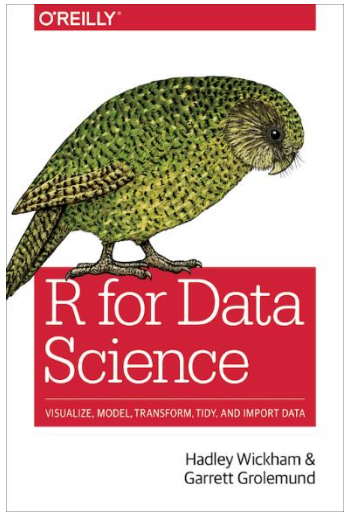
Similar to the Internet, the IOW will also require the development of a governance structure to connect regional data sharing communities, reducing redundancy and gaining efficiencies.

Recent developments in Data Science

- Tidy Data & the “HadleyVerse”
- Collaborative Pipelines & Reproducible Research

Tidy Data





[PDF] Tidy Data - Journal of Statistical Software
https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf
by H Wickham - Cited by 171 - Related articles
Aug 20, 2014 - **Tidy Data**. **Hadley Wickham** ... The principles of **tidy data** are closely tied to those of relational databases and Codd's rela- ..... 20Traditions.**pdf** ...

Tidy Data Concept…

- Each **variable** forms a *column*;
- Each **observation** forms a *row*; and
- The collection of **observational units** forms a *table*.

Count of individuals observed each day

	day	wolf	hare	fox
1	Monday	2	20	4
2	Tuesday	1	25	4
3	Wednesday	3	30	4

Is this tidy?

Defining Tidy Data

Messy…

	day	wolf	hare	fox
1	Monday	2	20	4
2	Tuesday	1	25	4
3	Wednesday	3	30	4

Tidy!

	day	species	count
1	Monday	wolf	2
2	Tuesday	wolf	1
3	Wednesday	wolf	3
4	Monday	hare	20
5	Tuesday	hare	25
6	Wednesday	hare	30
7	Monday	fox	4
8	Tuesday	fox	4
9	Wednesday	fox	4

Why tidy??

Easy manipulation of the data…

- Filtering rows (observations)
- Transforming data (derived columns)
- Aggregating
- Sorting

Plotting…

Modeling…

	day	wolf	hare	fox
1	Monday	2	20	4
2	Tuesday	1	25	4
3	Wednesday	3	30	4

	day	species	count
1	Monday	wolf	2
2	Tuesday	wolf	1
3	Wednesday	wolf	3
4	Monday	hare	20
5	Tuesday	hare	25
6	Wednesday	hare	30
7	Monday	fox	4
8	Tuesday	fox	4
9	Wednesday	fox	4

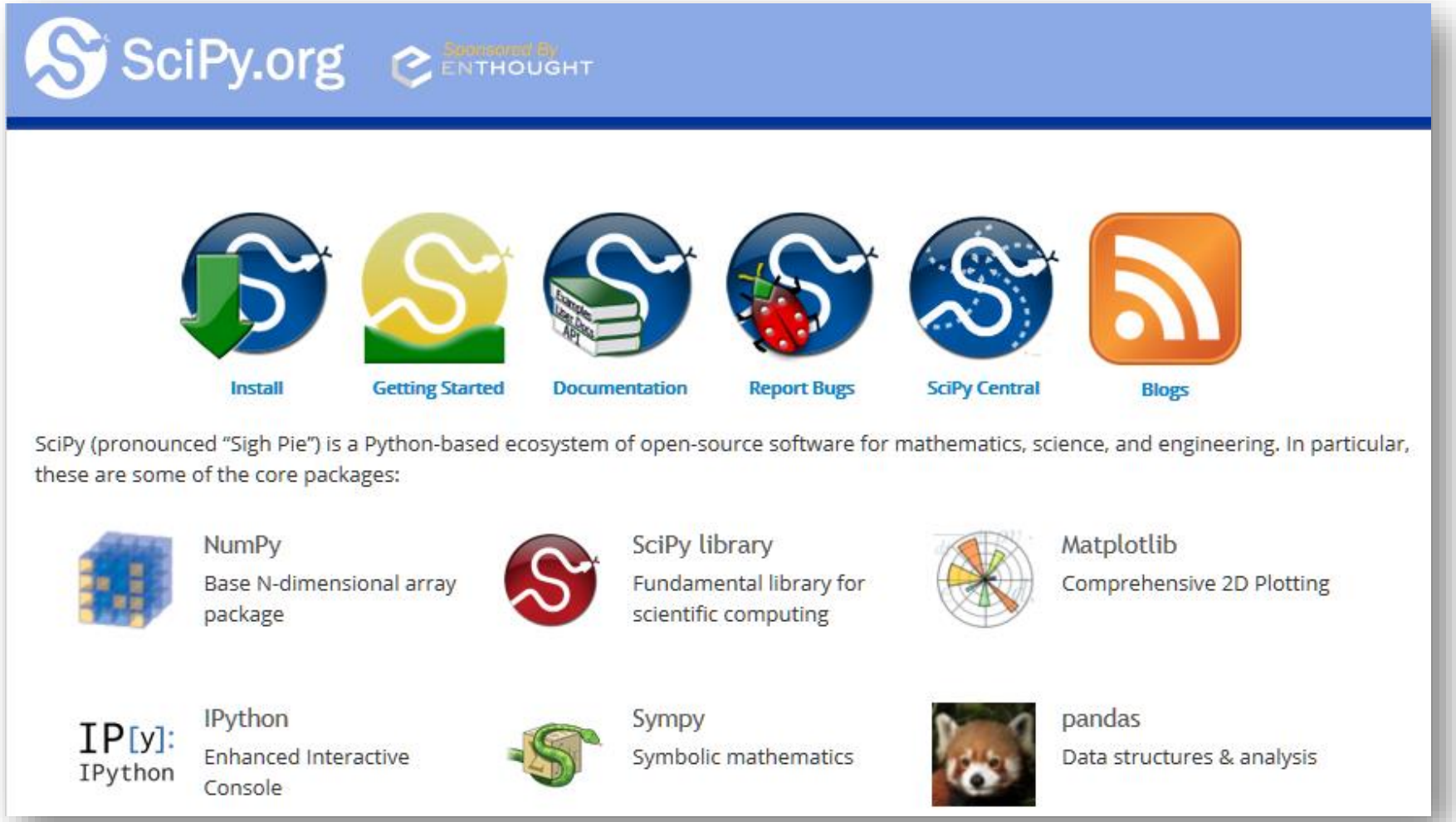
Data science – in R

- TidyVerse
  Set of R Tools for tidying data and working with tidy data
  - https://www.tidyverse.org/packages/

- Tools are designed to string – or “pipe” – commands together
  - Output of one tool becomes the input of another…

```
the_data <-
  read.csv('/path/to/data/file.csv') %>%
  subset(variable_a > x) %>%
  transform(variable_c = variable_a/variable_b) %>%
  head(100)
```

Data science – in Python



Typical Data Analytic Operations

- **Acquiring data**

- **Cleaning and "wrangling" data**
  - Filtering/sorting
  - Aggregating/summarizing

- **Visualizing data**
  - Plotting
  - Interactive apps

Acquiring [online] data

Tiers of access to online data:

- **Manual download** (e.g. via links)
- ***Scraping* data**: grabbing data shown on a web page
- **APIs** (**A**pplication **P**rogramming **I**nterfaces): specialized web addresses
- **Specialized Packages**: Programming “wrappers” for data access

[Python] DEMOs

Data wrangling

- Excel: Sort, Filter, & Pivot Tables
- Databases & SQL (e.g. *MS Access*)
- R/Python & DataFrames

Data Visualization

- Plotting in Excel
- Plotting in R/Python
  - ggplot
  - matplotlib
- Tableau
- D3/JavaScript