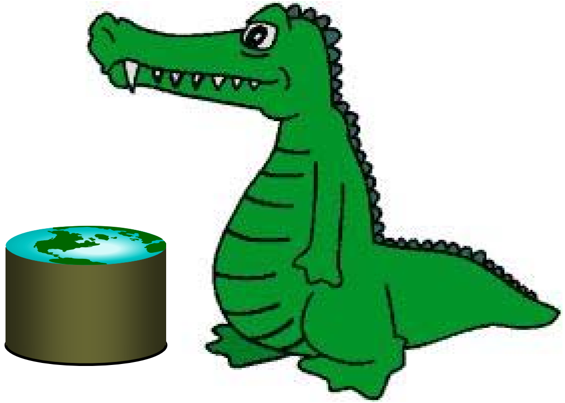
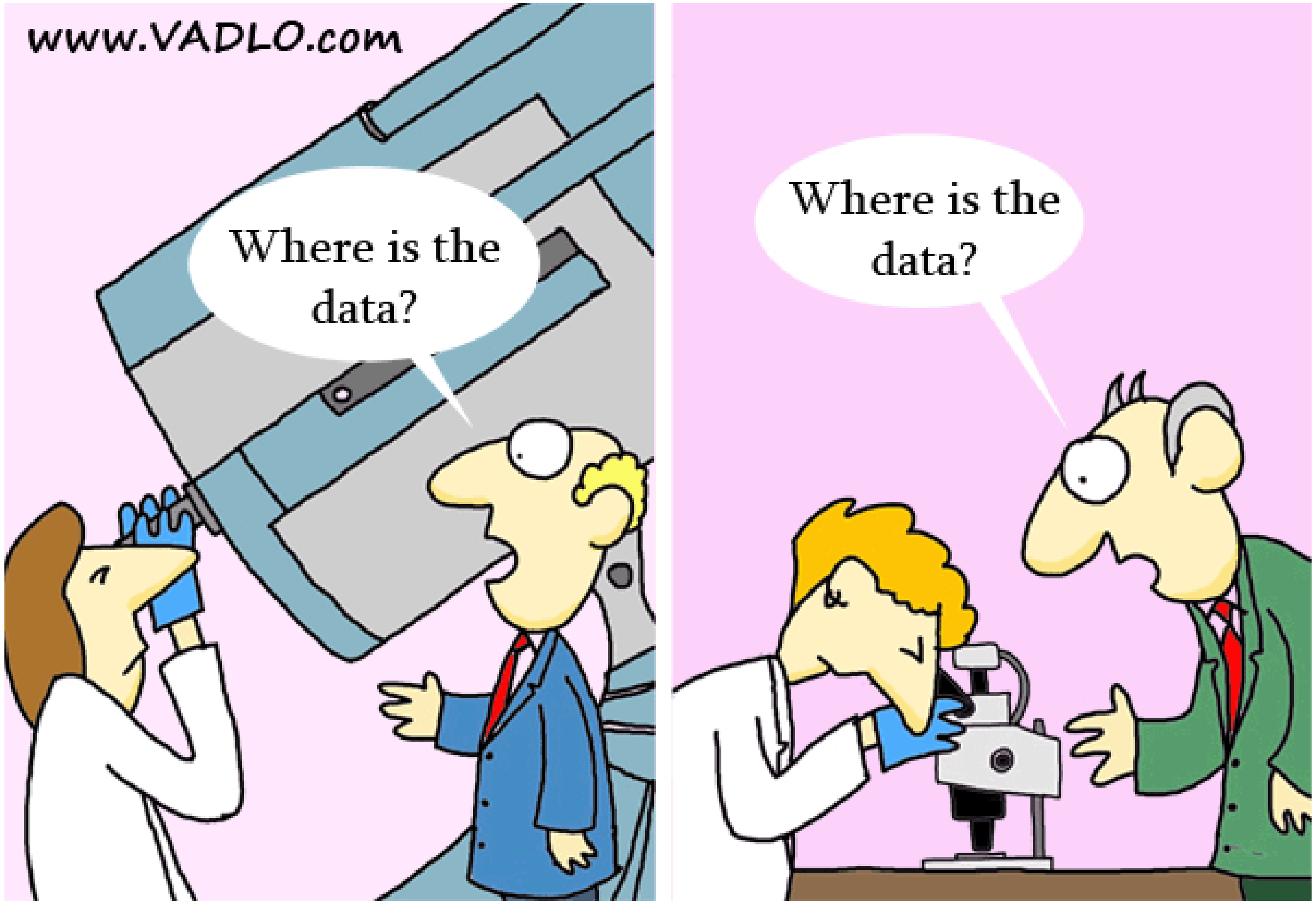
DATA SCIENCE: APLICACIONES A LA BIOLOGIA Y A LA MEDICINA CON PYTHON Y R

Toni Monleón-Getino
Section of Statistics. Fac Biology



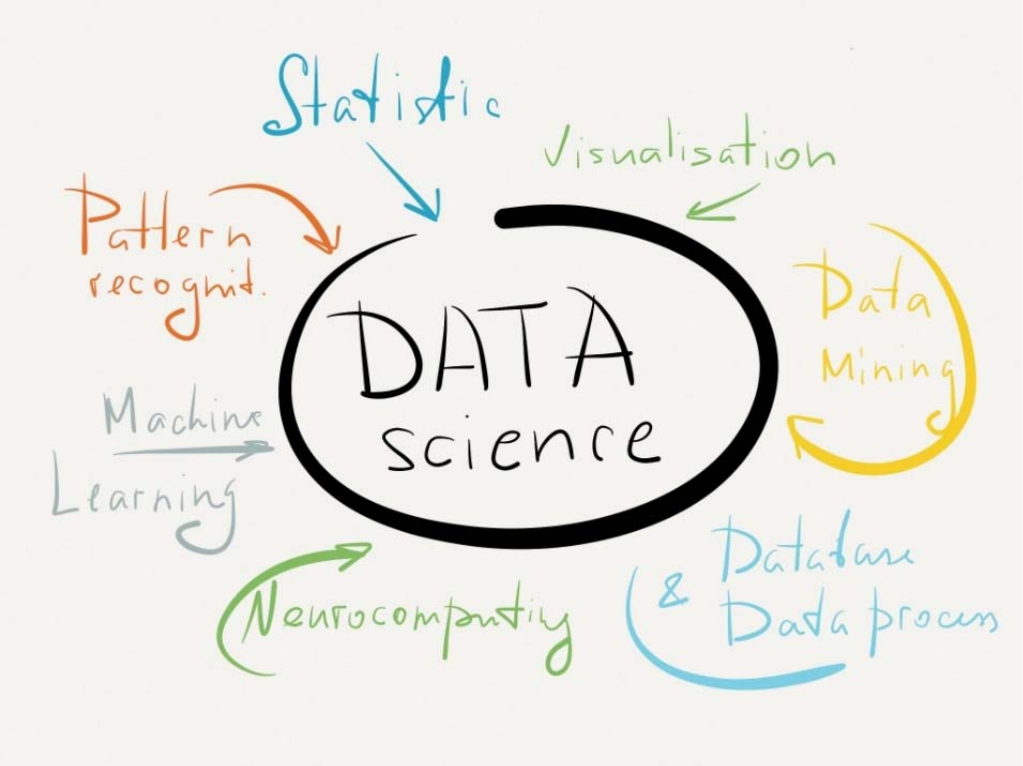
Based on notes from CS194 at UC Berkeley by Michael Franklin, John Canny, and Jeff Hammerbacher / University of Florida, CISE Department
Prof. Daisy Zhe Wang

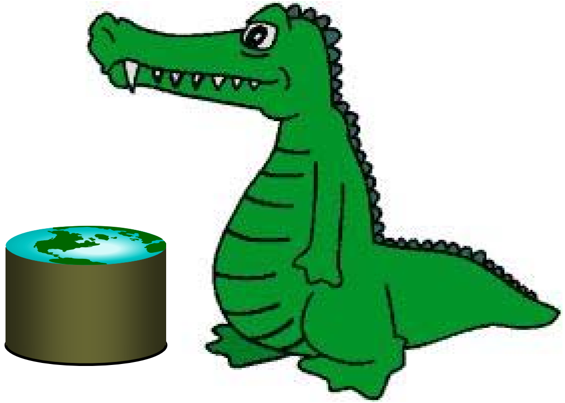
Grand Unification of Sciences



Grand Unification of Sciences

What is Data Science?





Why, Where, What, How, Who

“Data Science” an Emerging Field

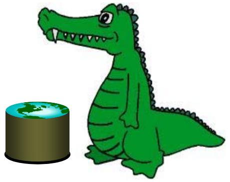
What is Data Science?

The future belongs to the companies and people that turn data into products

O’Reilly Radar report, 2011

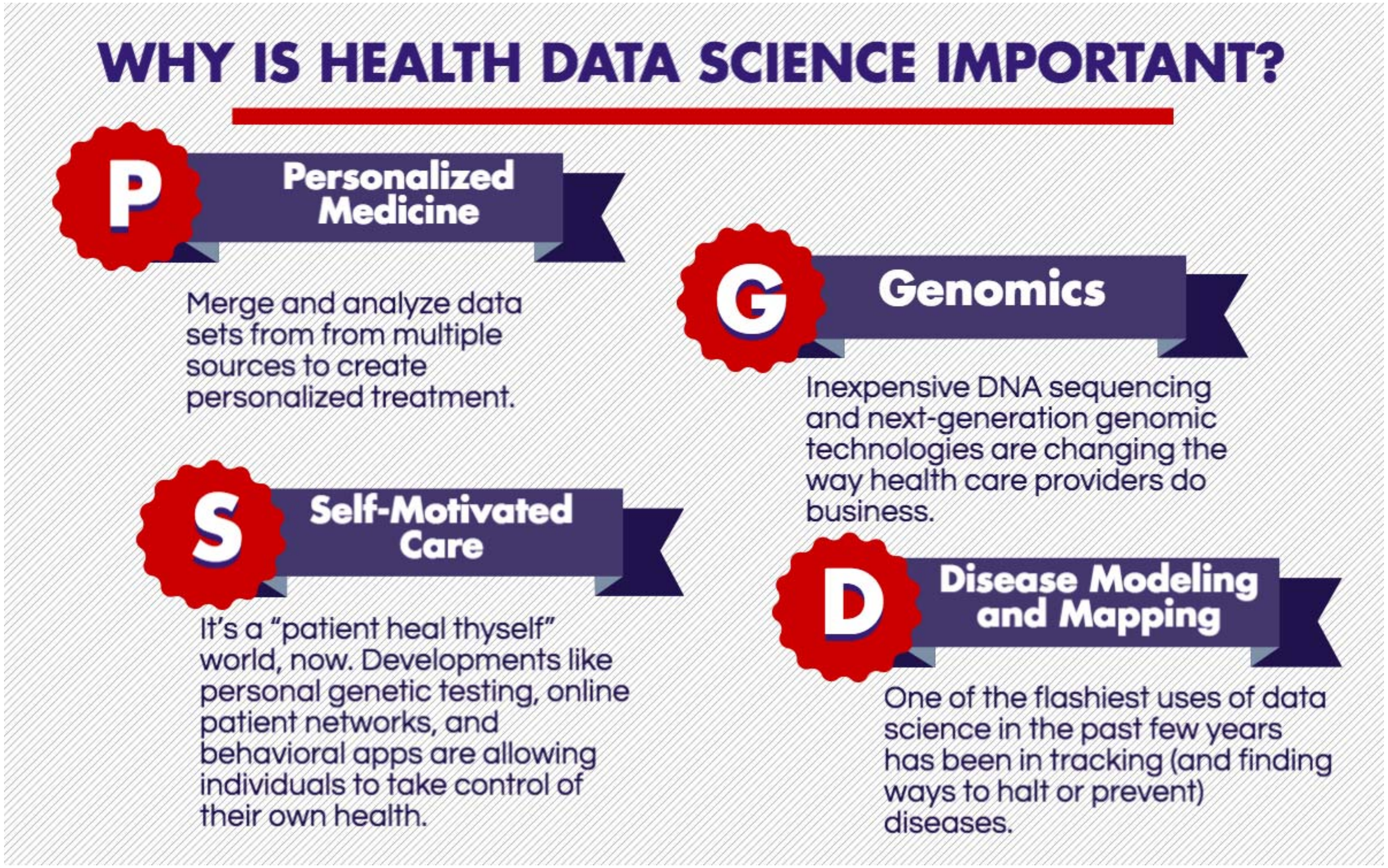
Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products and science.



Goal of Data Science

Turn data into data products and science.



WHY IS HEALTH DATA SCIENCE IMPORTANT?

P **Personalized Medicine**

Merge and analyze data sets from from multiple sources to create personalized treatment.

G **Genomics**

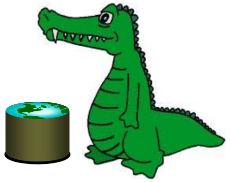
Inexpensive DNA sequencing and next-generation genomic technologies are changing the way health care providers do business.

S **Self-Motivated Care**

It's a "patient heal thyself" world, now. Developments like personal genetic testing, online patient networks, and behavioral apps are allowing individuals to take control of their own health.

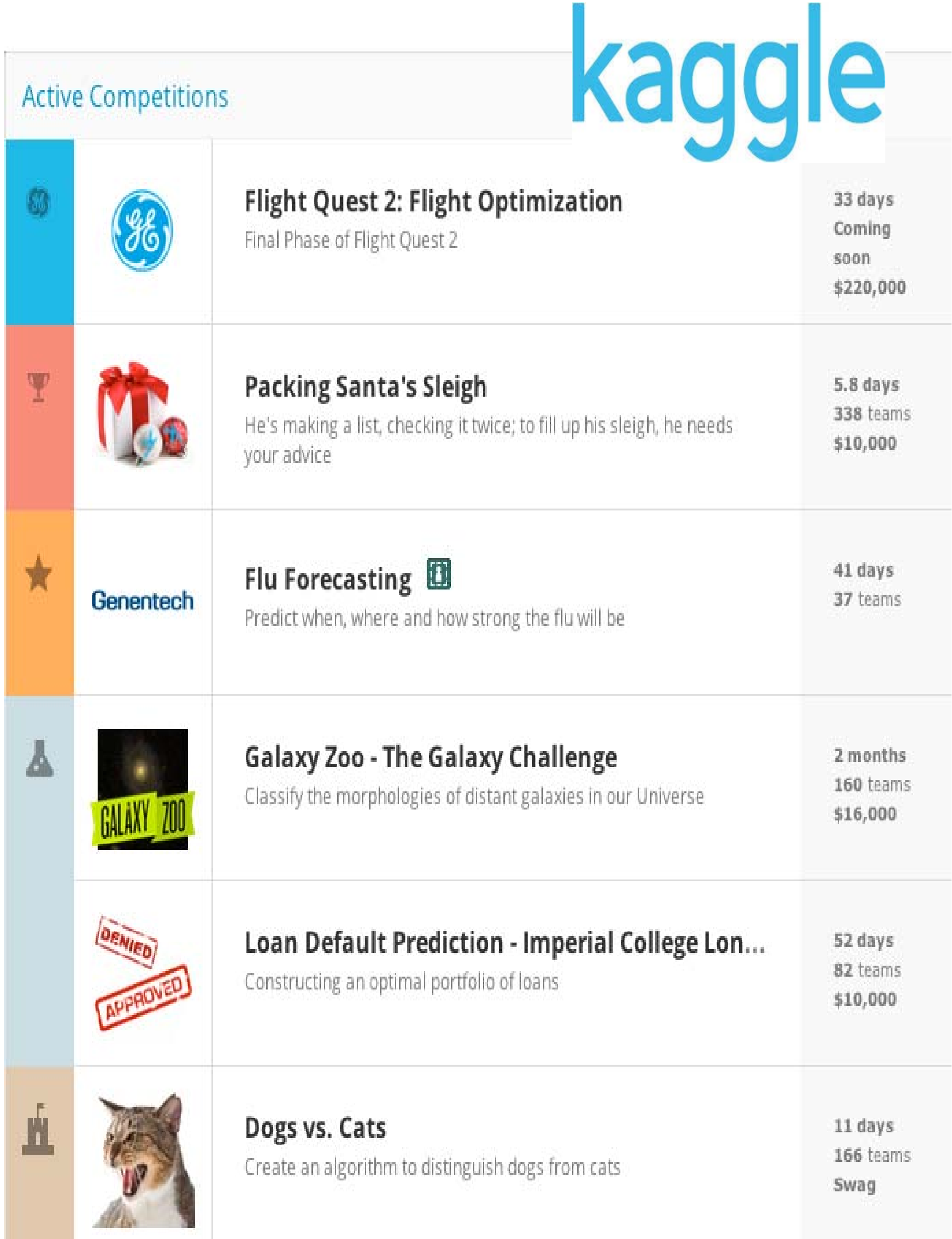
D **Disease Modeling and Mapping**

One of the flashiest uses of data science in the past few years has been in tracking (and finding ways to halt or prevent) diseases.



Some recent ML Competitions at https://www.kaggle.com/

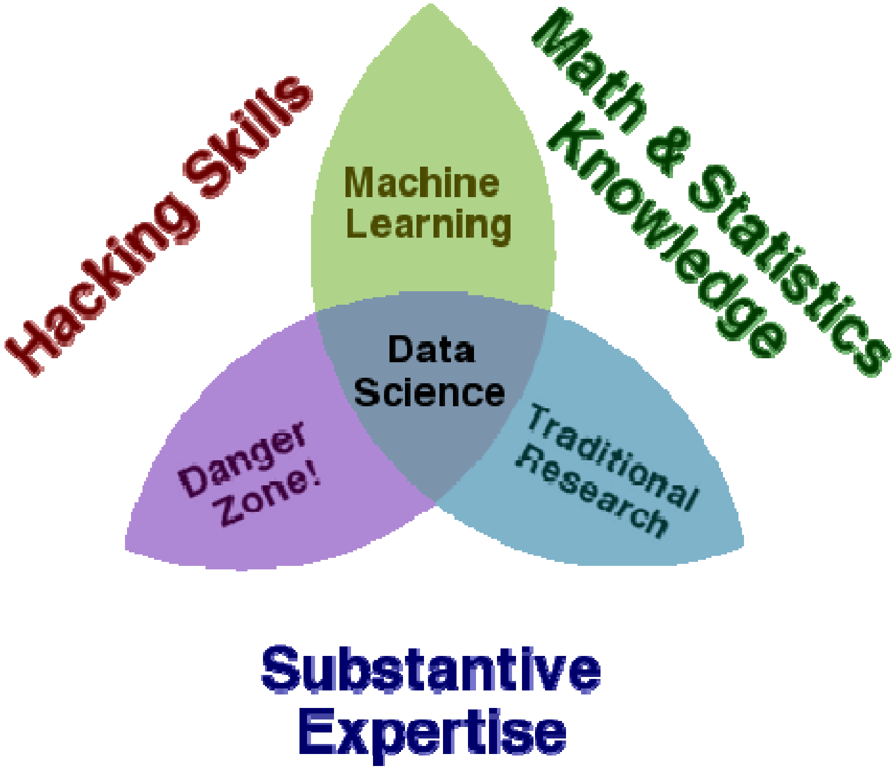
NIST Pre-Pilot Data Science Evaluation – likely to be incorporated to be part of Labs/Final project



Active Competitions

Genentech	Flight Quest 2: Flight Optimization Final Phase of Flight Quest 2	33 days Coming soon $220,000
	Packing Santa's Sleigh He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams $10,000
Genentech	Flu Forecasting Predict when, where and how strong the flu will be	41 days 37 teams
GALAXY ZOO	Galaxy Zoo - The Galaxy Challenge Classify the morphologies of distant galaxies in our Universe	2 months 160 teams $16,000
DENIED APPROVED	Loan Default Prediction - Imperial College Lon... Constructing an optimal portfolio of loans	52 days 82 teams $10,000
	Dogs vs. Cats Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag

Data Science – A Visual Definition



Contrast: Databases

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra, Riak, Memcached, Apache River, …

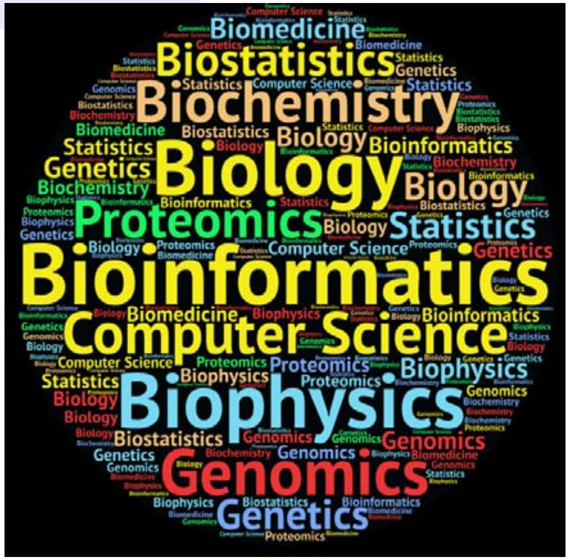
ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

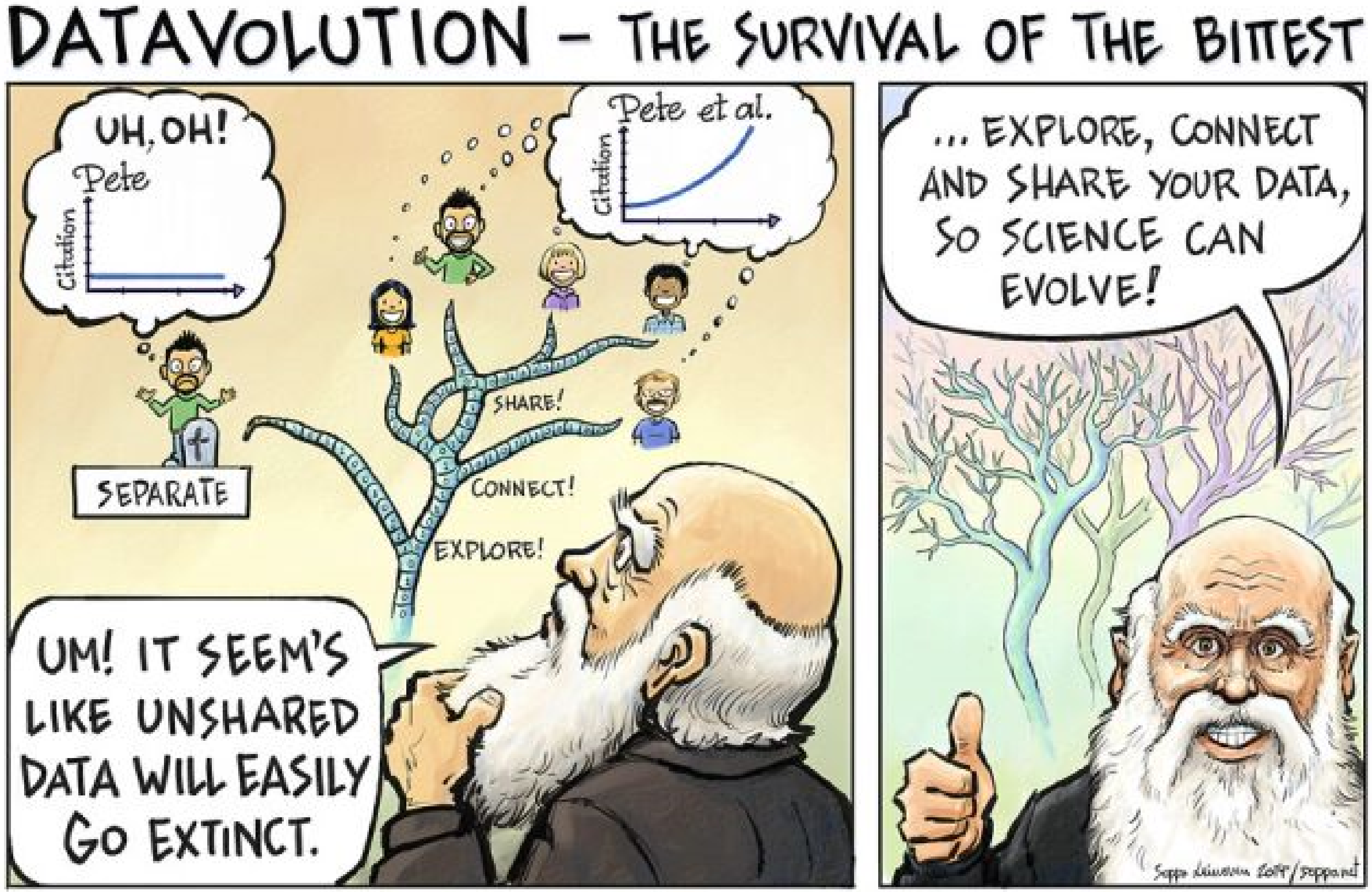
Contrast: Machine Learning

Machine Learning
Develop new (individual) models
Prove mathematical properties of models
Improve/validate on a few, relatively clean, small datasets
Publish a paper

Data Science
Explore many models, build and tune hybrids
Understand empirical properties of models
Develop/use tools that can handle massive datasets
Take action!



datavolution


DATAVOLUTION – THE SURVIVAL OF THE BITTEST
UH, OH!
Pete
Citation
SEPARATE
Pete et al.
Citation
SHARE!
CONNECT!
EXPLORE!
UM! IT SEEM'S LIKE UNSHARED DATA WILL EASILY GO EXTINCT.
... EXPLORE, CONNECT AND SHARE YOUR DATA, SO SCIENCE CAN EVOLVE!
Seppo Leinonen 2014 / seppo.net