



---

UNIVERSITÀ DEGLI STUDI DI MILANO  
DIPARTIMENTO DI INFORMATICA

# BIG DATA AND DIGITAL METHODS

## LECTURE 4

# DATA SCIENCE

Dario Malchiodi



# SOURCES

- C. Somahano: Big Data [sorry] & Data Science: what does a data scientist do?
- R Orban: How to become a data scientist



# DATA TO BE ANALYZED



1969, man on the Moon: 64 KB

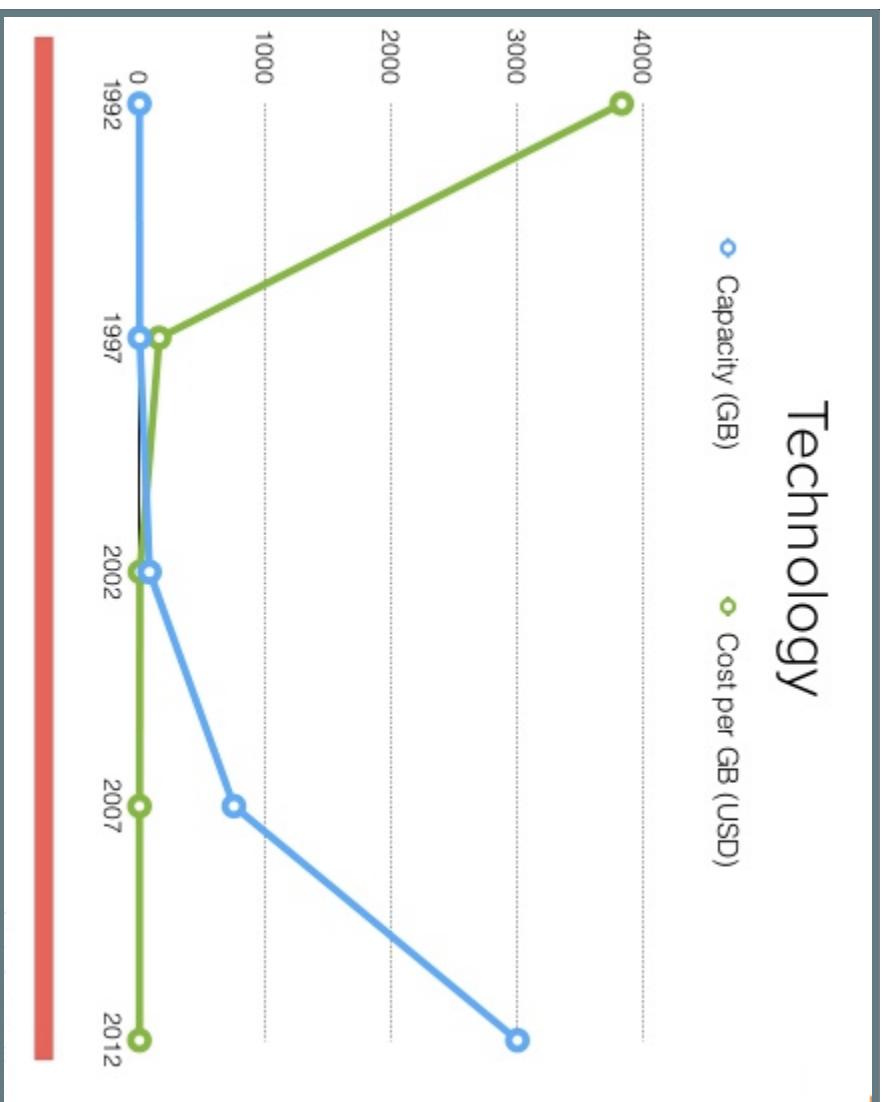
---



2012, Red Bull skydive stratos: tens of GB!



# COST VS. CAPACITY





# AN ANALOGY

Think data as «crude oil» to be

- extracted
- transported in (big) «tankers»
- siphoned through (capable) «pipelines»
- stored in (massive) «silos»
- refined through «data science»
- in order to get valuable «data products»



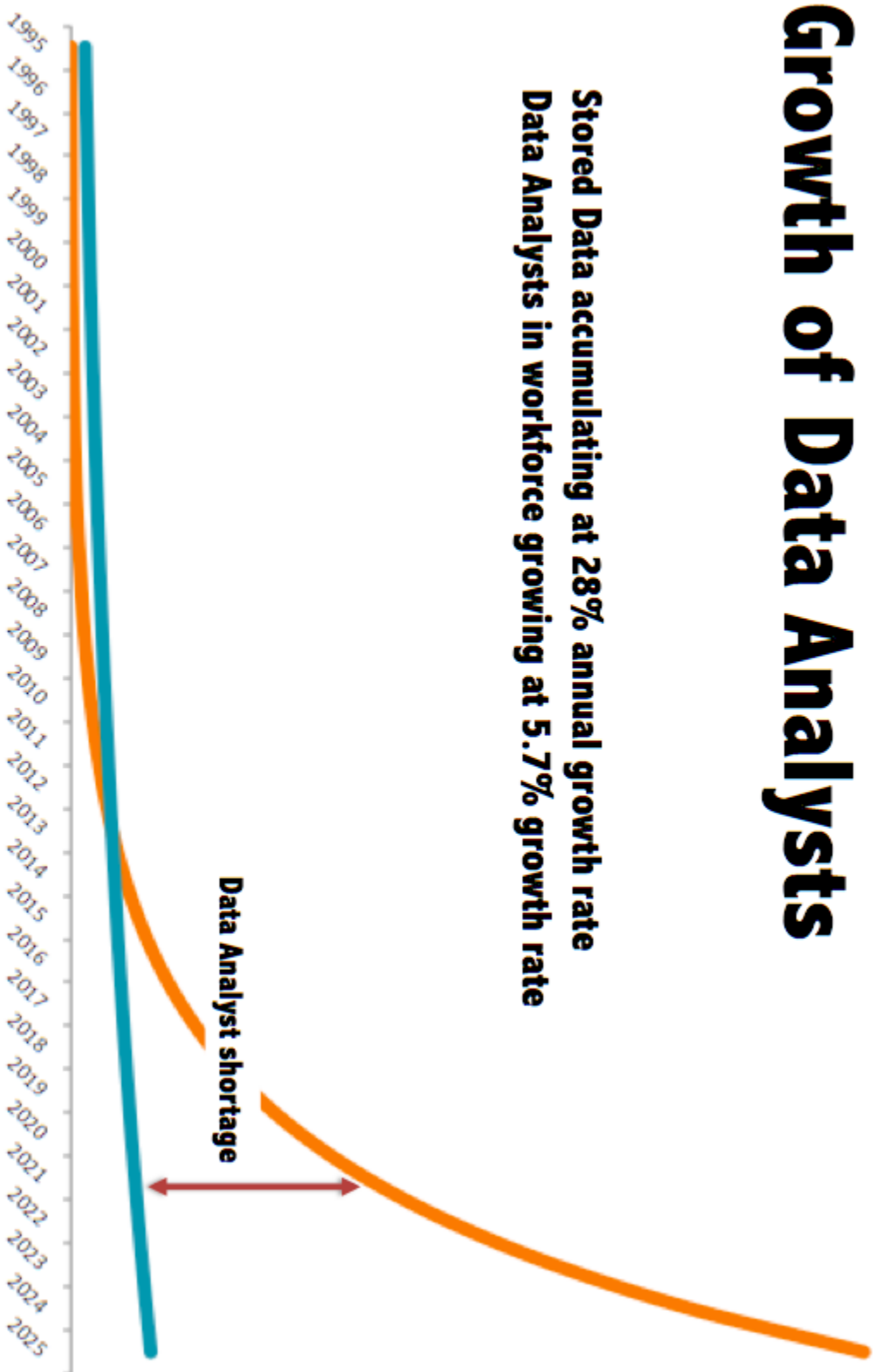
# DATA SCIENCE: INFORMAL DEFINITIONS

- Discovering what we don't know from data
- Obtaining predictive, actionable insight from data
- Creating Data Products that have business impact now
- Communicating relevant business stories from data
- Building confidence in decisions that drive business value



# Growth of Data vs. Growth of Data Analysts

**Stored Data accumulating at 28% annual growth rate**  
**Data Analysts in workforce growing at 5.7% growth rate**



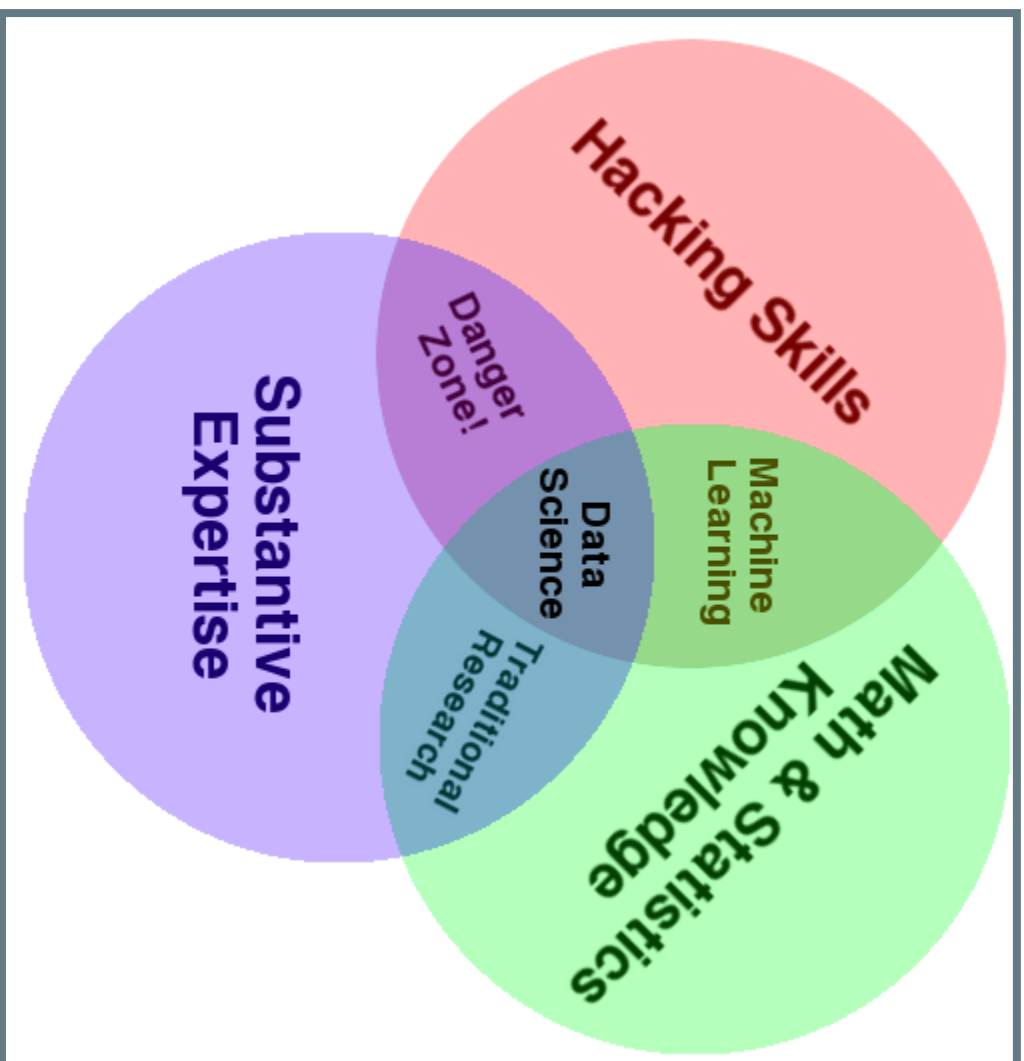
[source: <http://www.delphianalytics.net/wp-content/uploads/2013/04/GrowthOfDataVsDataAnalysts.png>]

# **H. MASON AND C. WIGGINS, 2010**

- Data science is clearly a blend of the hackers' arts, statistics and machine learning...
- and the expertise in mathematics and the domain of the data for the analysis to be interpretable...
- It requires creative decisions and open-mindedness in a scientific context.



# D. CONWAY, 2010



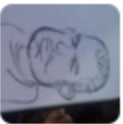


# D. J. PATIL, 2011

- We realized that as our organizations grew, we both had to figure out what to call the people on our teams. «Business analyst» and «Data analyst» seemed too limiting
- The term that seemed to fit best was «data scientist»: those who use both data and science to create something new



# J. WILLS, 2012



**Josh Wills**  
@josh\_wills



**Follow**

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS

**1,172**

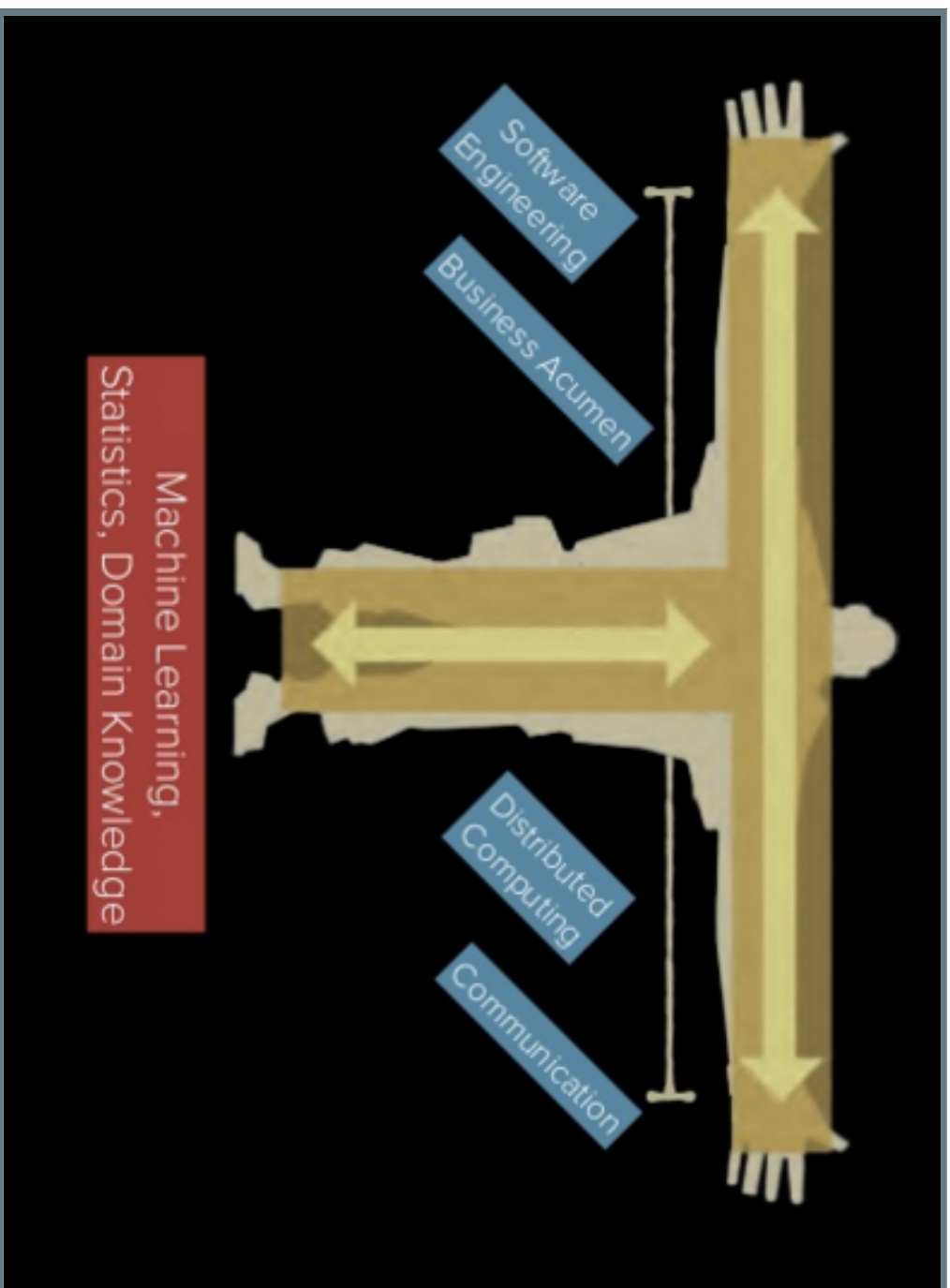
FAVORITES

**640**

6:55 PM - 3 May 2012



# T-SHAPED SKILLSET





# C. SOMOHANO, 2013

- Is skeptical, curious. Has inquisitive mind
- Knows Machine Learning, Statistics, Probability
- Applies Scientific Method. Runs Experiments
- Is good at Coding & Hacking
- Able to deal with IT Data Engineering
- Knows how to build data products
- Able to find answers to known unknowns
- Tells relevant business stories from data
- Has Domain Knowledge



the sexiest job of the 21st century

### Degreed In Geek

It doesn't hurt to have degrees from top operations research, statistics or mathematics programs,

[such as those at MIT, Stanford](#)

### Problem-Solving Prowess

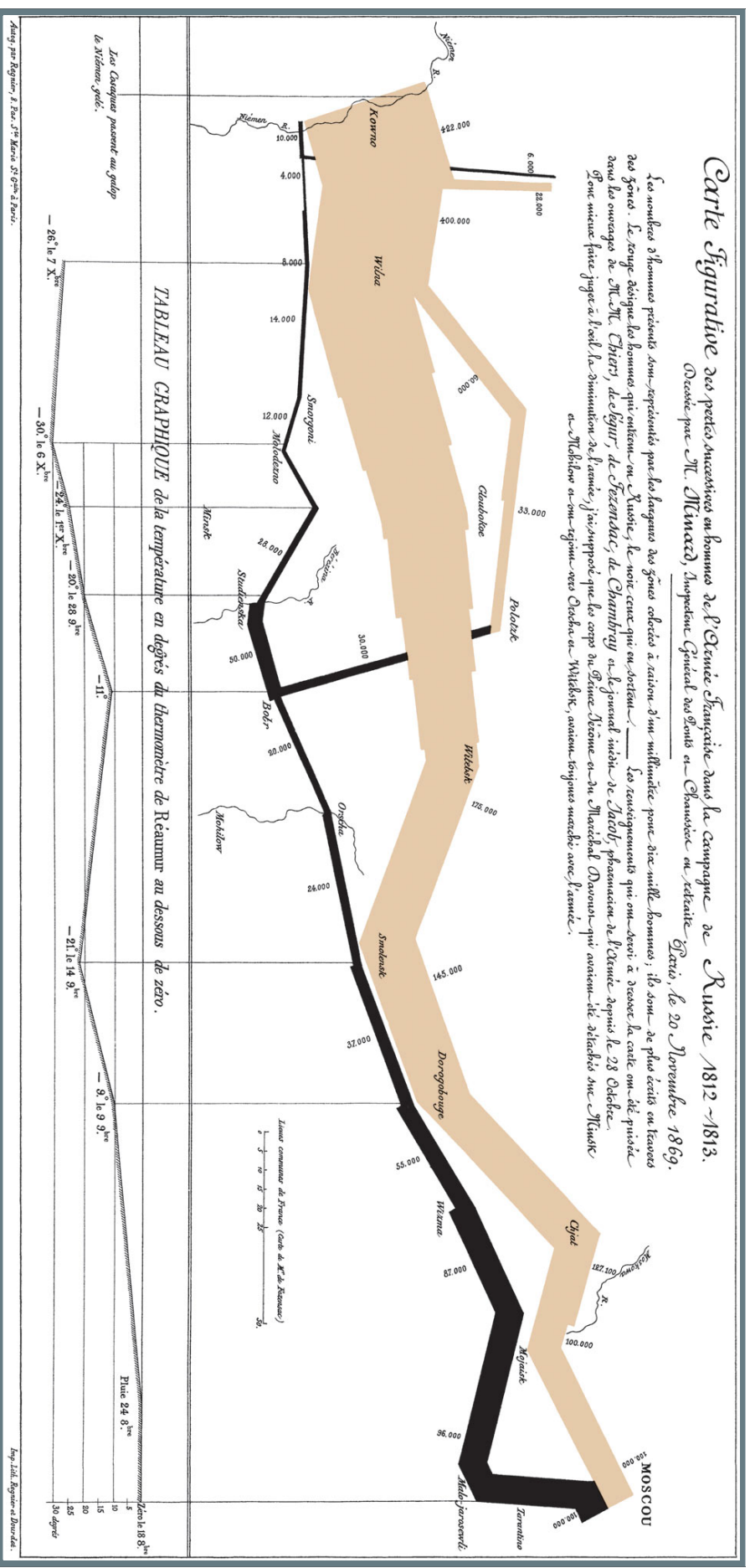
A problem solver at heart who's able to

[find solutions to even the most complex](#)

**IT'S NOT AT ALL A  
«MODERN» JOB!**



# CHARLES JOSEPH MINARD, 1851



number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates



# FLORENCE NIGHTINGALE, 1858

2.

DIAGRAM OF THE CAUSES OF MORTALITY

IN THE ADVANCE OF THE EAST

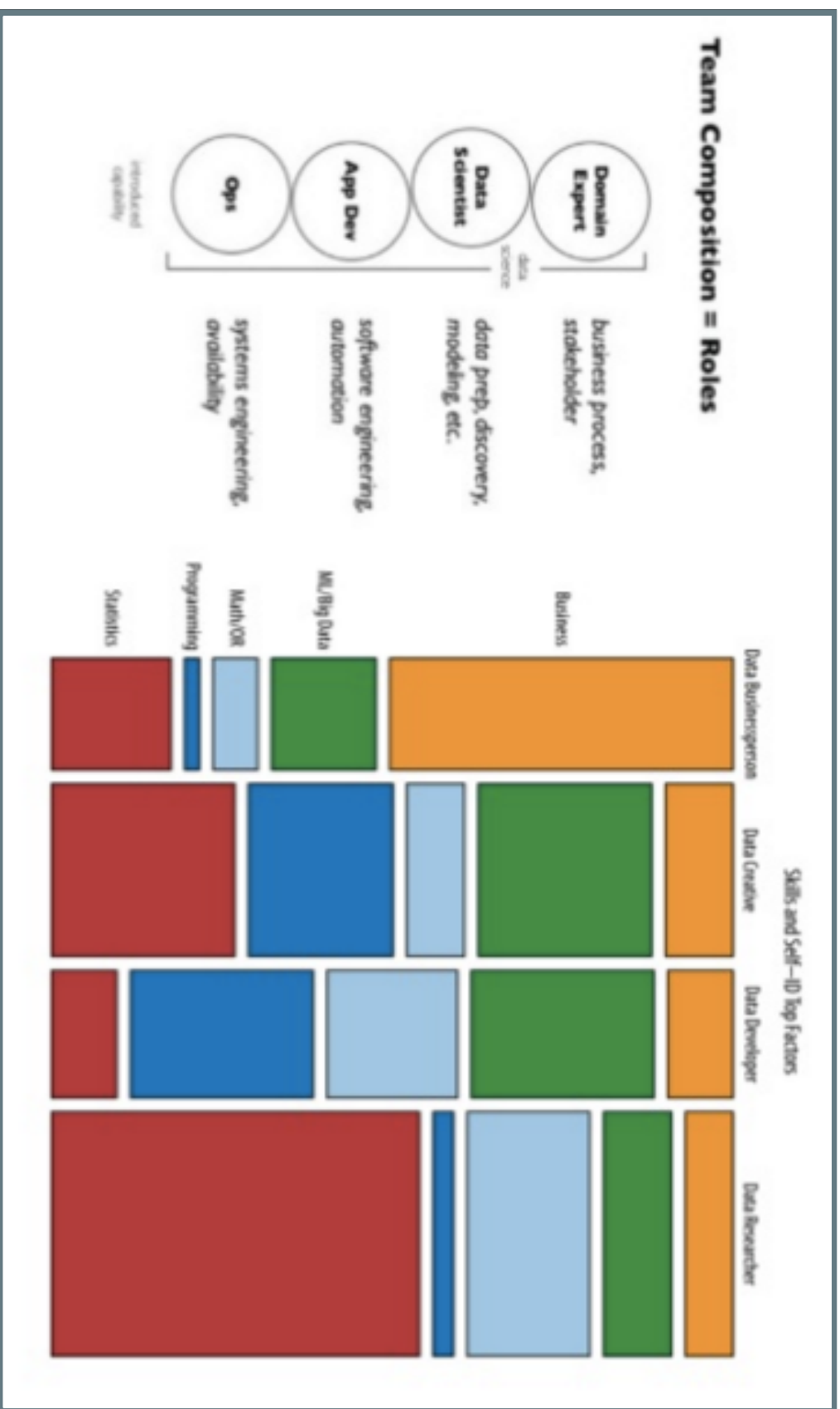
1.

# TOP 10 THINGS DATA SCIENTISTS DO

1. Ask Good Questions: What is what we don't know? And what we'd like to know?
2. Define and test an hypothesis, run experiments
3. Scoop, scrap, sink, & sample business relevant data
4. Munge and wrestle data, tame data
5. Explore data, discover unknowns
6. Model data and algorithms
7. Understand data relationships
8. Let a computer learn from data
9. Create data products that deliver actionable insight
10. Tell relevant business stories from data



# DATA SCIENCE ROLES





# A DATA SCIENTIST TOOLKIT 1

- Different modern programming languages (Java, R, Python, Scala)
- Frameworks for distributed storage and computation (Hadoop, Spark, Storm)
- Querying languages for distributed information (HBase, Pig & Hive)
- Software enabling efficient data collection and transfer (Web scrapers, Flume, Sqoop)



# A DATA SCIENTIST TOOLKIT 2

- Frameworks for accessing structured data (SQL, RDBMS, OLAP)
- Frameworks for accessing non-structured data (NoSQL, Mongo DB, Couchbase, Cassandra...)
- Machine learning and data-crunching libraries (Weka, SciPy, NumPy, scikit-learn, pandas)
- Visualization libraries (D3.js, Gephi, ggplot2, Tableau, Flare, Shiny)
- Systems for «heavy» symbolic/numerical computations (Wolfram Mathematica, SPSS, Matlab, SAS)



# DATA SCIENCE METRO MAP

ing  
e Extraction  
Sparse Values  
ed Estimators  
g Missing Values  
rubbing  
lization  
nsionality & Numerosity  
Reduction

# PRINCIPLES

1. Socio-Technical Systems (STS) are complex!
2. Data is never at rest
3. Data is dirty, deal with it
4. Data munging & data wrestling > 70% time
5. Simplification, reduction, distillation
6. Curiosity. Empiricism. Skepticism

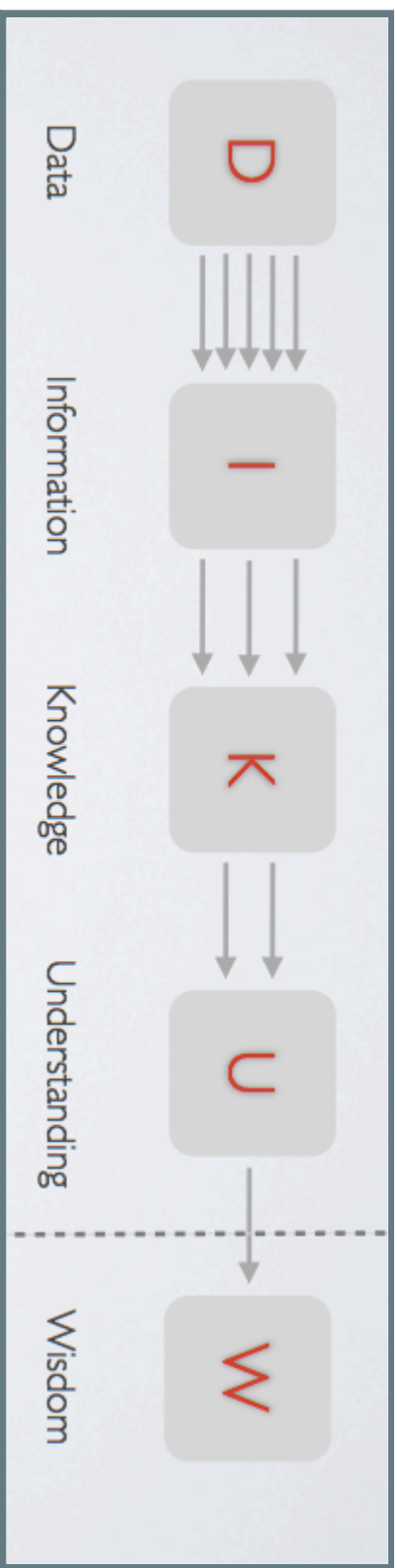


# KNOWN & UNKNOWN (D. RUMSFELD)

- There are known knowns. These are things we know that we know
- There are known unknowns. That is to say, there are things that we know we don't know
- But there are also unknown unknowns. There are things we don't know we don't know



# DIKUW





# DATA DISCOVERY

Things you know

# DATA MODELING

$F(x, \text{parameters, random noise}) \mapsto y$

We understand the world

- How well «my data model» works
- Statisticians, Data Analysts, Data Miners
- Linear Regression, Logistic Regression
- Known Distributions
- Confidence Intervals
- Predictor Variables & Goodness of Fit



# ALGORITHMIC MODELING

$x \mapsto \text{black box} \mapsto y$

We don't understand the world

- The world produces data in a black-box
- Data Scientists
- Machine Learning, AI & Neural Nets
- Random Forests, SVM, ...
- Iterative
- Predictive Accuracy



# LEARNING FROM DATA IS TRICKY

- Statistical vs. Machine learning
- Supervised vs. unsupervised Learning
- Induction vs. deduction
- Sampling & Confidence Intervals, Probability & Distribution, ...
- Correlation vs. Causation



# DATA SCIENCE PROCESS 1





# DATA SCIENCE PROCESS 2





# A DATA PRODUCT IS...

- Curated and crafted from raw data
- A result of exploration and iterations
- A machine that learns from data
- An answer to known unknowns or unknown unknowns
- A mechanism that triggers immediate business value
- A probabilistic window of future events or behavior

