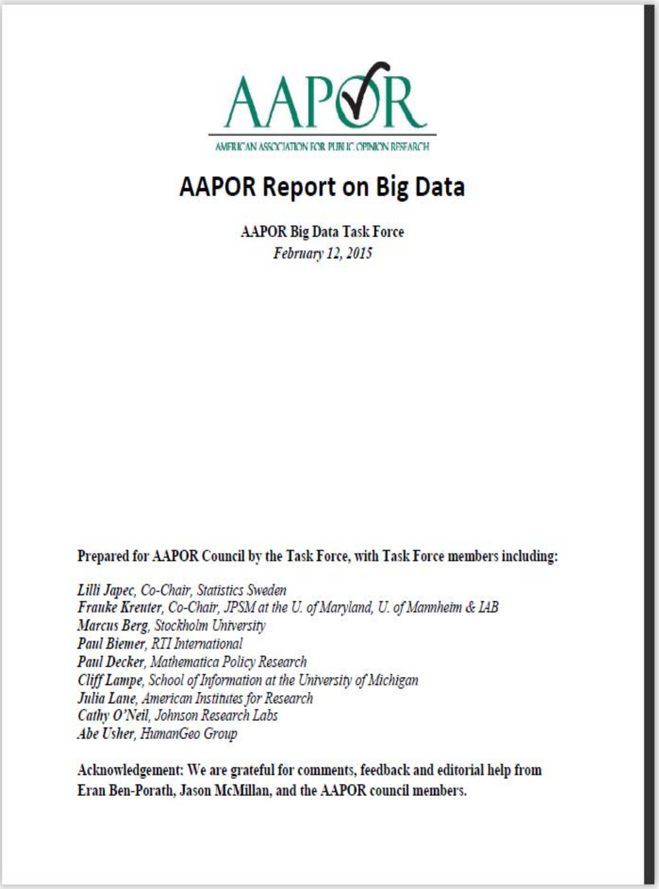
Summary

Big Data – Web Scraping and Machine Learning



AAPOR
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

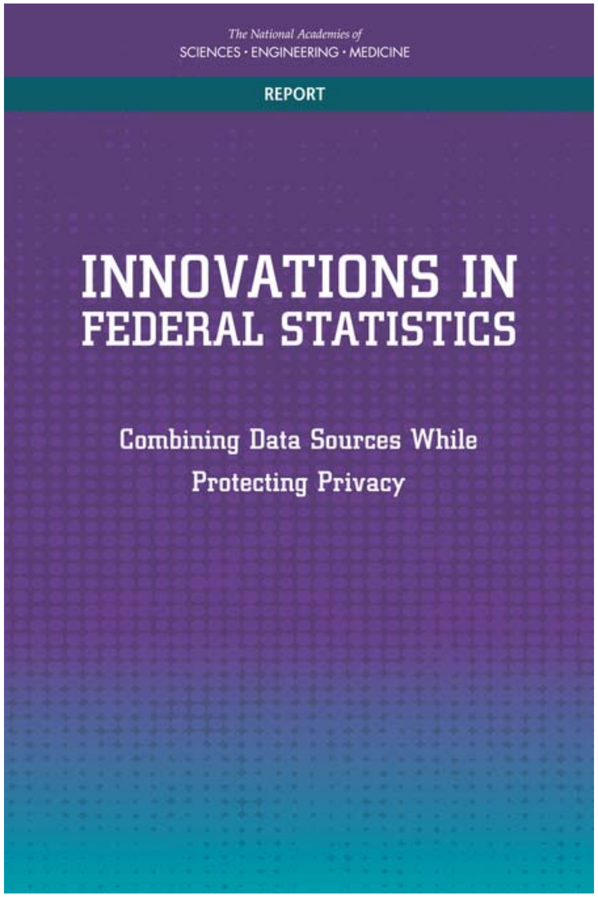
AAPOR Report on Big Data

AAPOR Big Data Task Force
February 12, 2015

Prepared for AAPOR Council by the Task Force, with Task Force members including:

Lilli Japec, Co-Chair, Statistics Sweden
Frauke Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB
Marcus Berg, Stockholm University
Paul Biemer, RTI International
Paul Decker, Mathematica Policy Research
Cliff Lampe, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Cathy O'Neil, Johnson Research Labs
Abe Usher, HumanGeo Group

Acknowledgement: We are grateful for comments, feedback and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members.

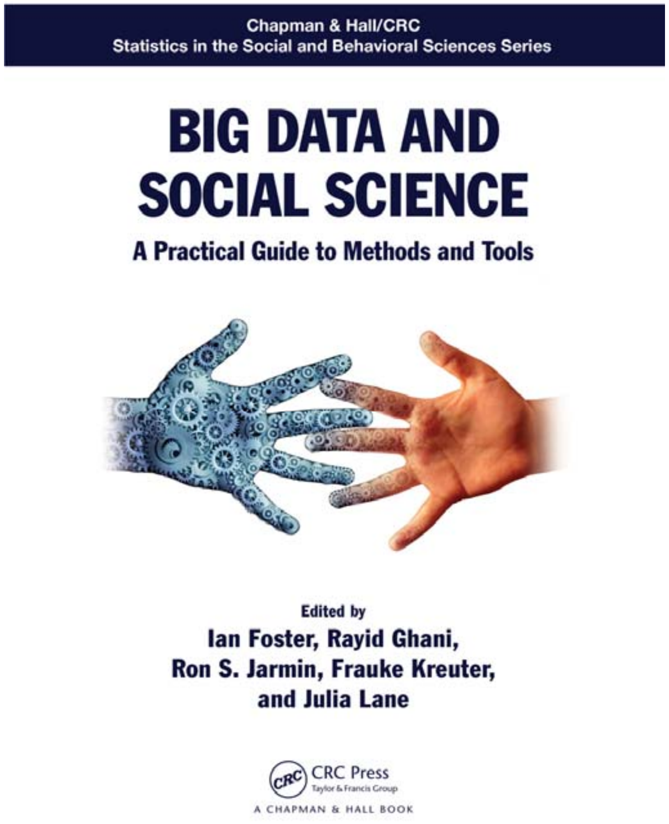


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

REPORT

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

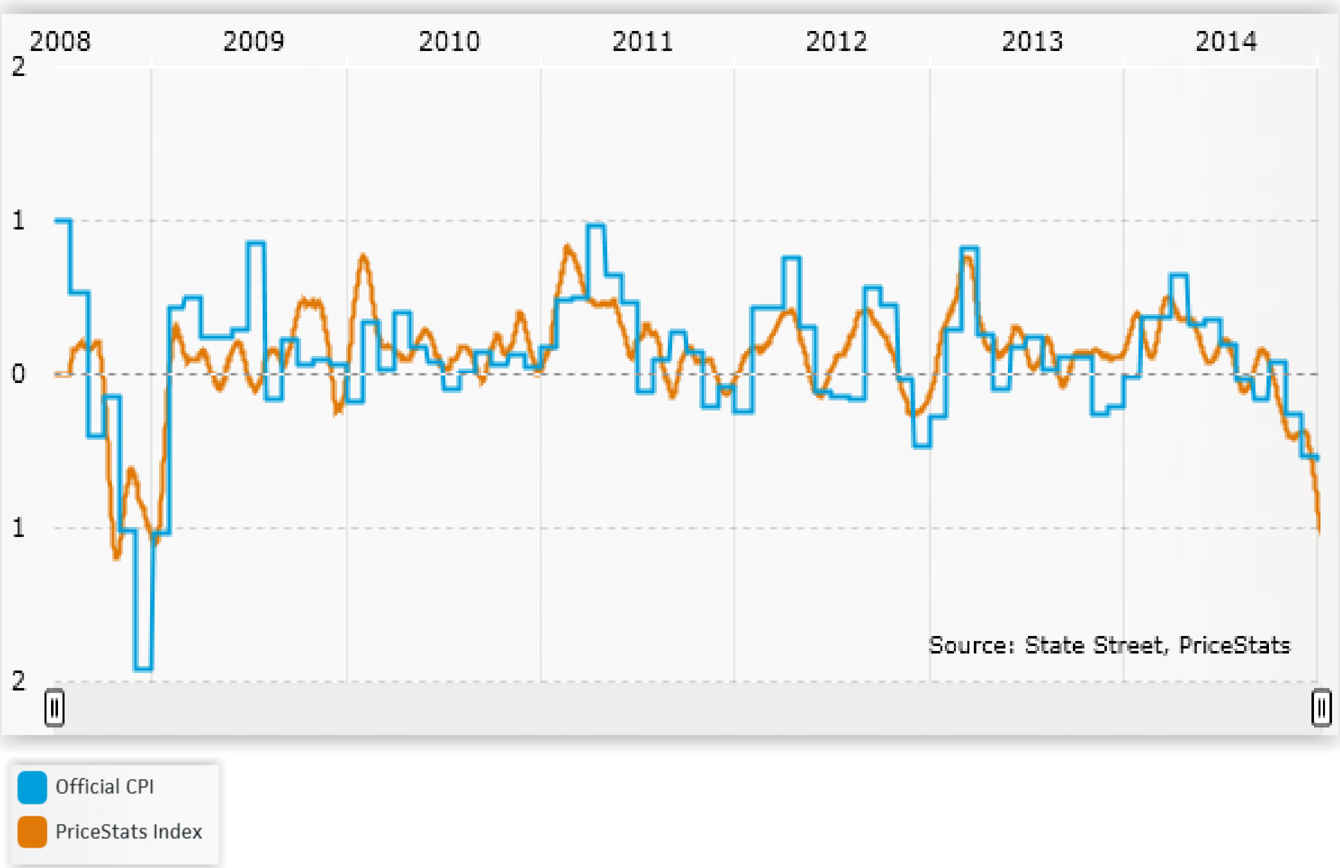
BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

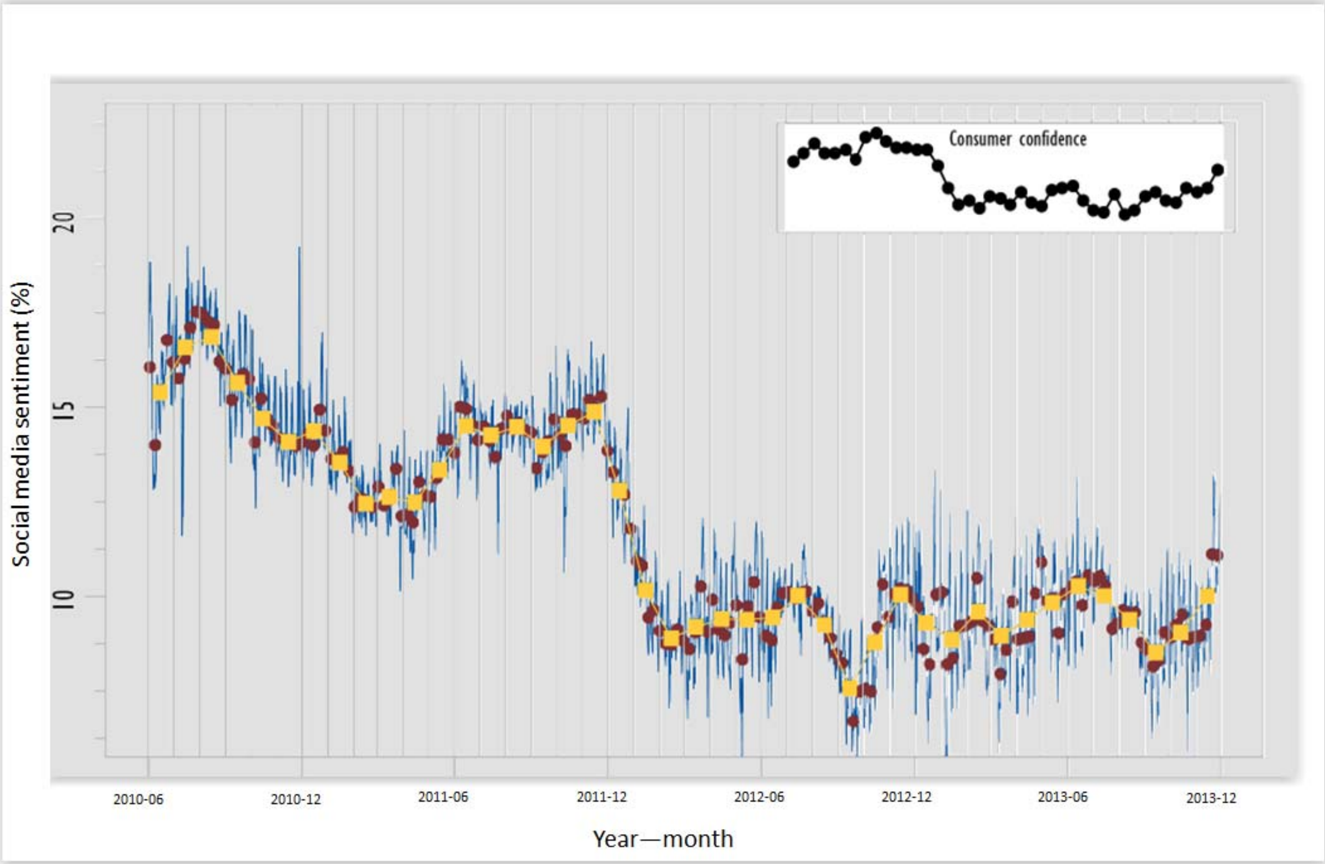
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

The Excitement

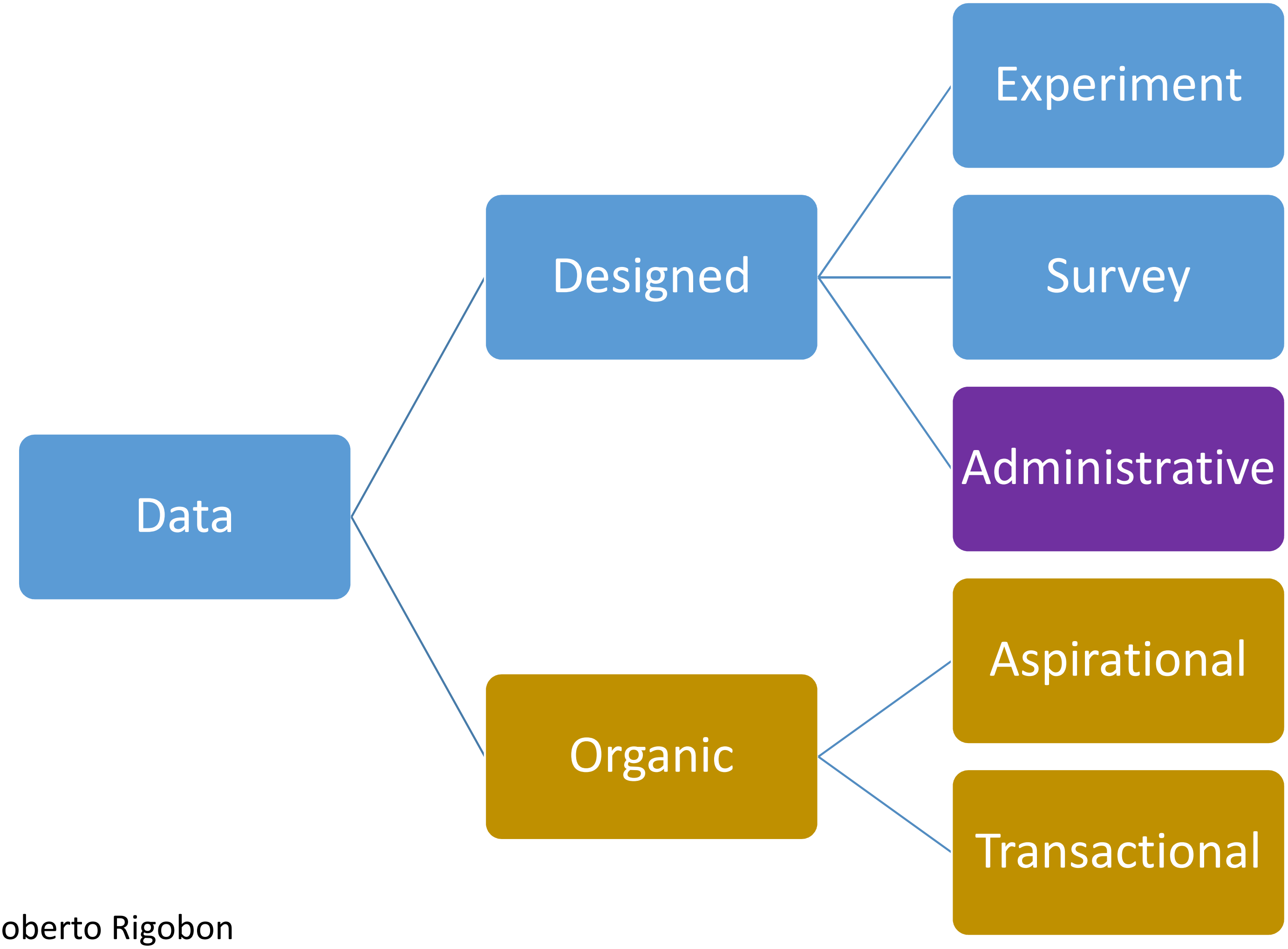


Source: State Street, PriceStats

US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI. Accessed January 18, 2015 from the PriceStats website.



Social media sentiment (daily, weekly and monthly) in the Netherlands, June 2010 - November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).

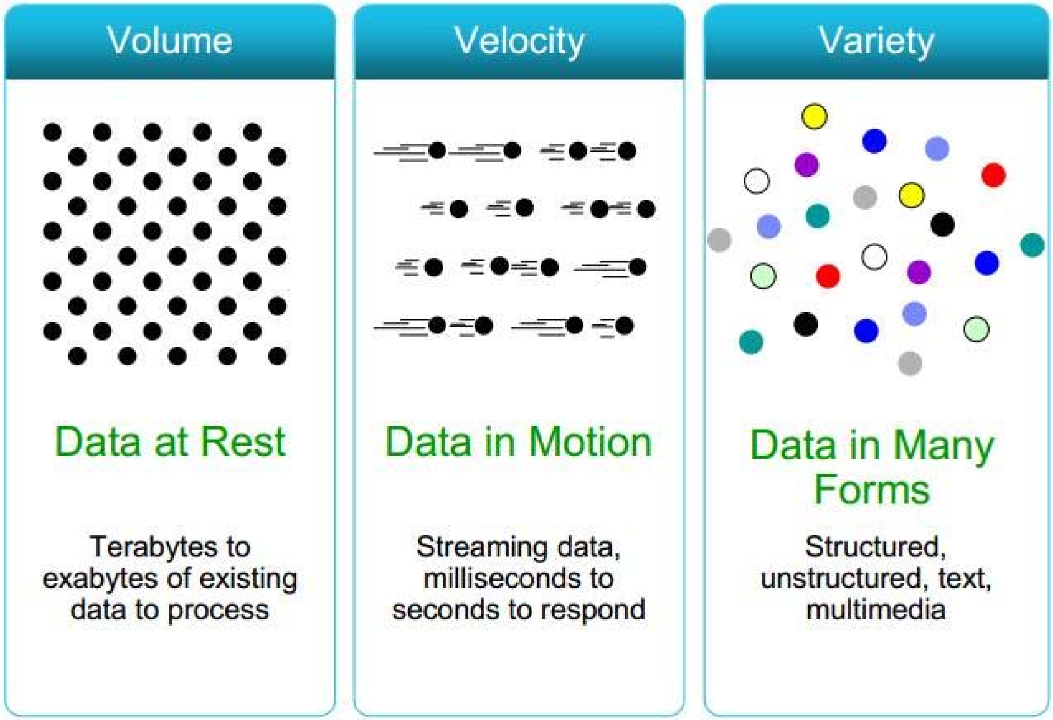


Source: Roberto Rigobon

Excitement over new data sources:

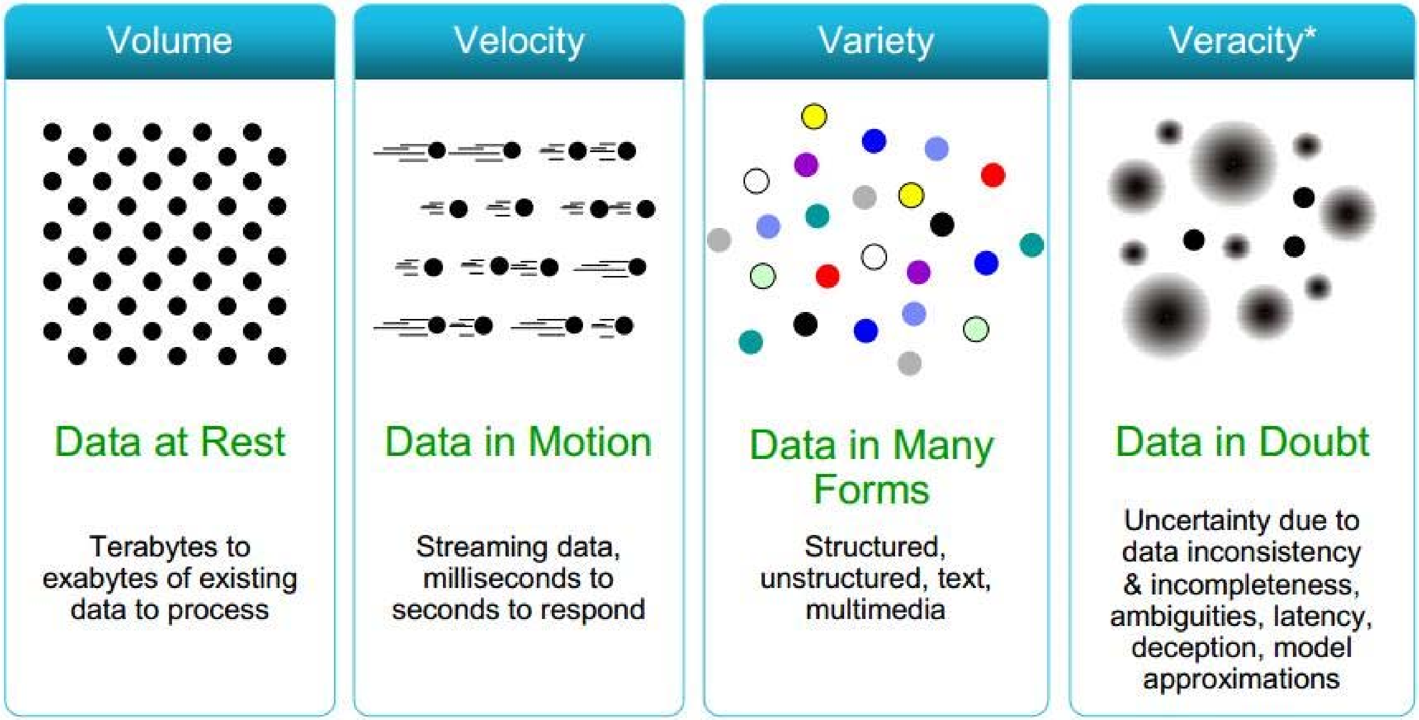
1. New research questions can be asked
   - spatial and temporal granularity
   - small parts of the populations
   - other form of data (text, visuals)
2. Reduced data collection costs
3. ‘Instant’ more timely availability

Vs as defining characteristic


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia

http://www.rosebt.com/blog/data-veracity

…one more V

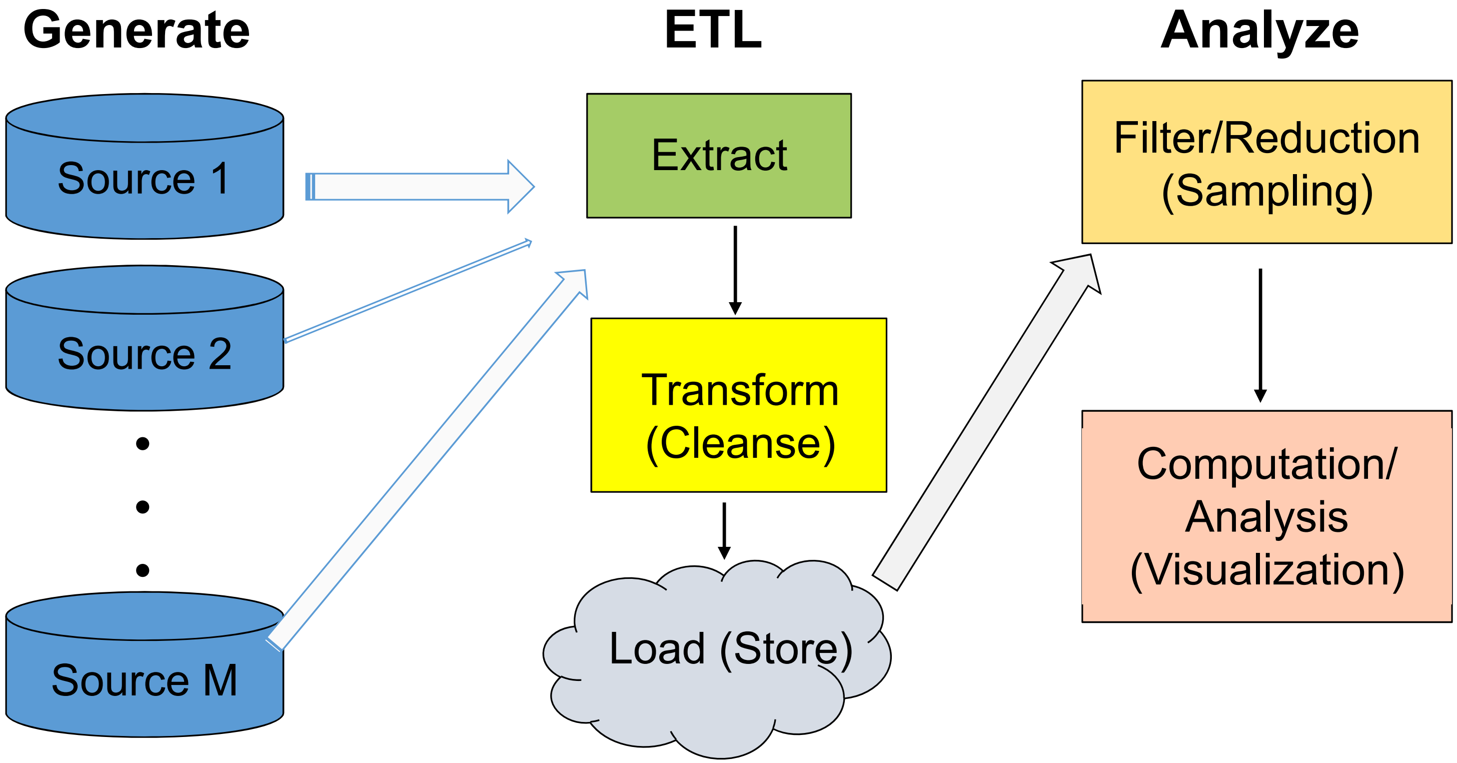


Volume	Velocity	Variety	Veracity*
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

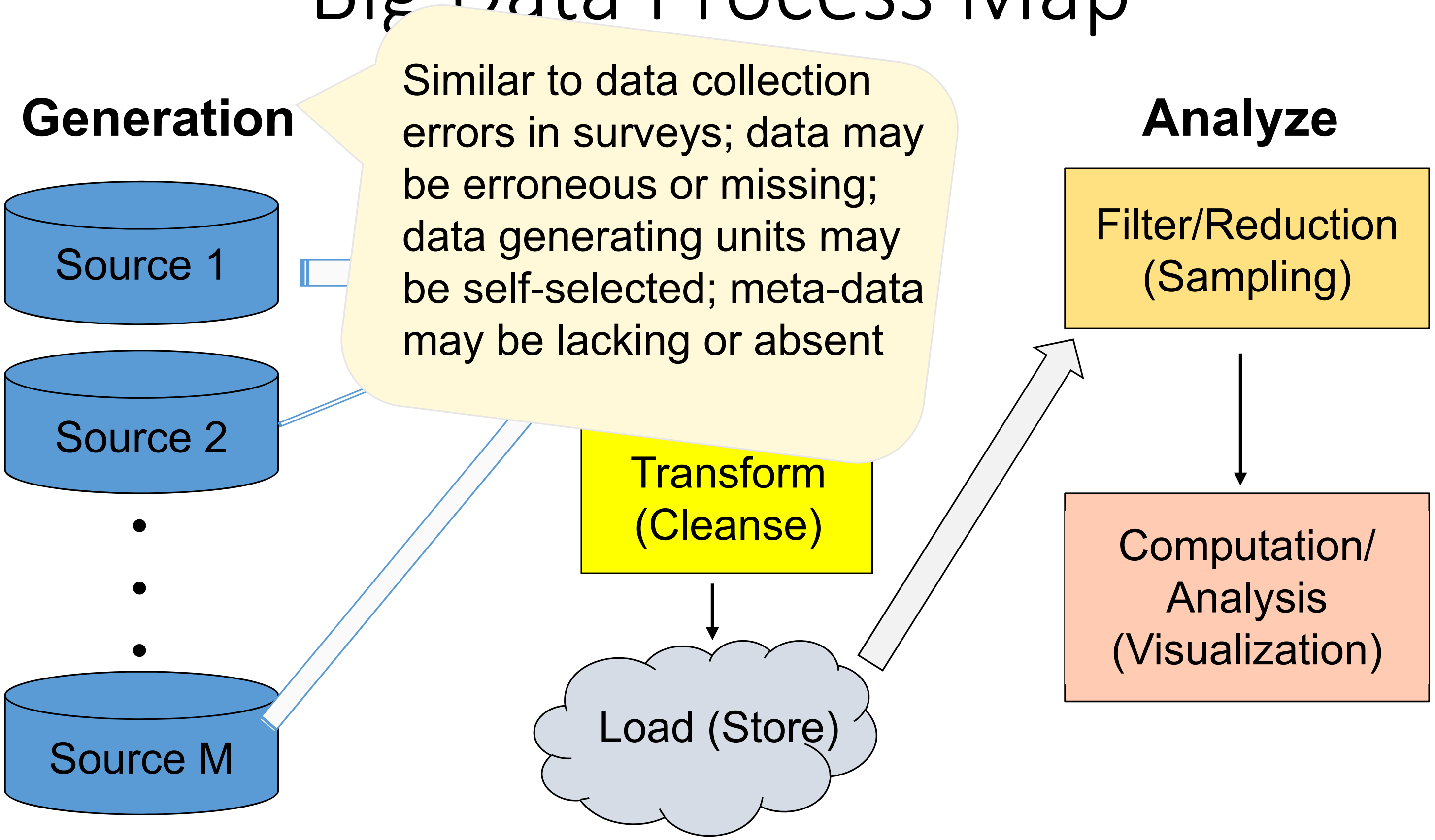
http://www.rosebt.com/blog/data-veracity

The Process

Big Data Process Map

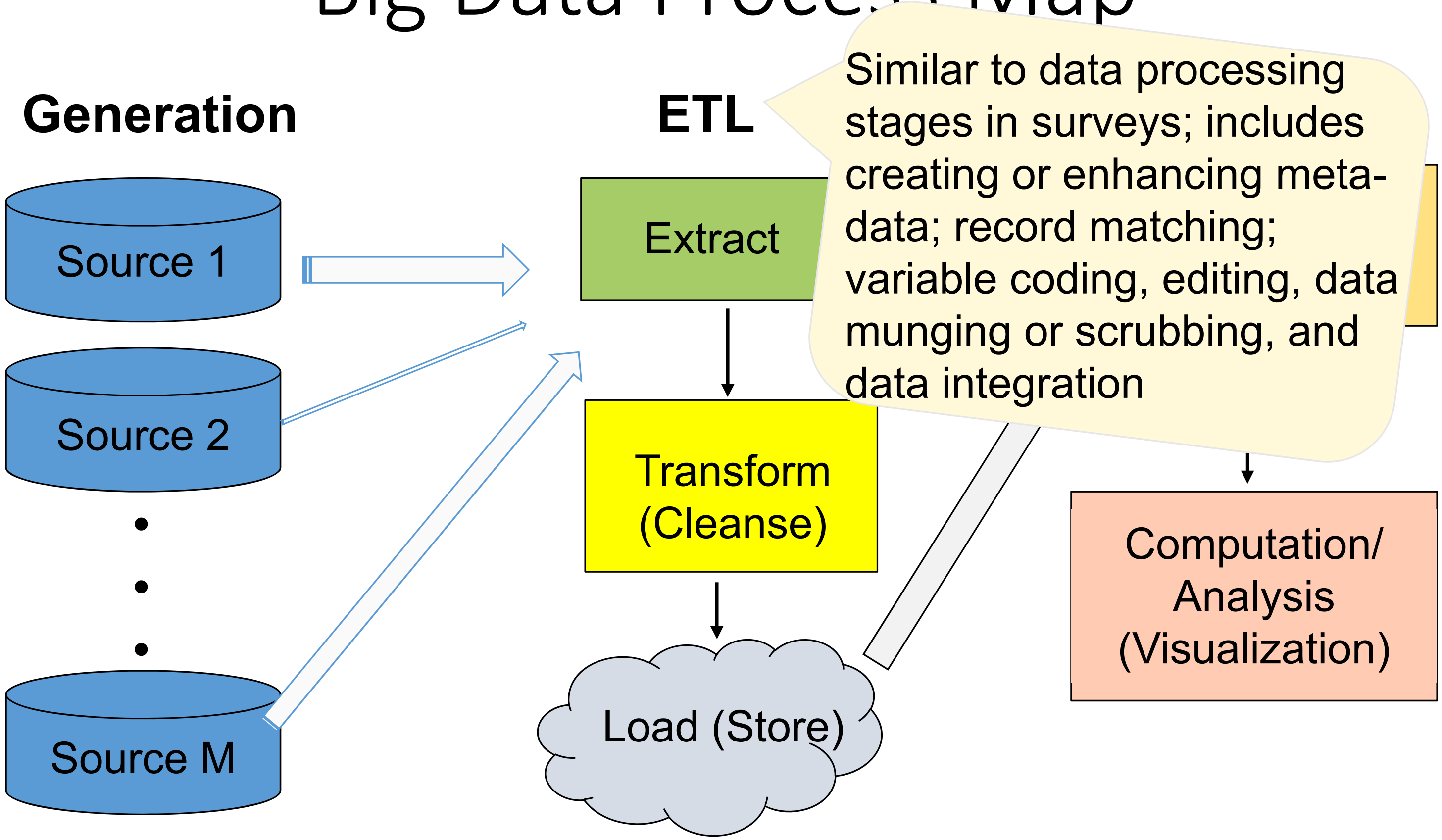


Big Data Process Map



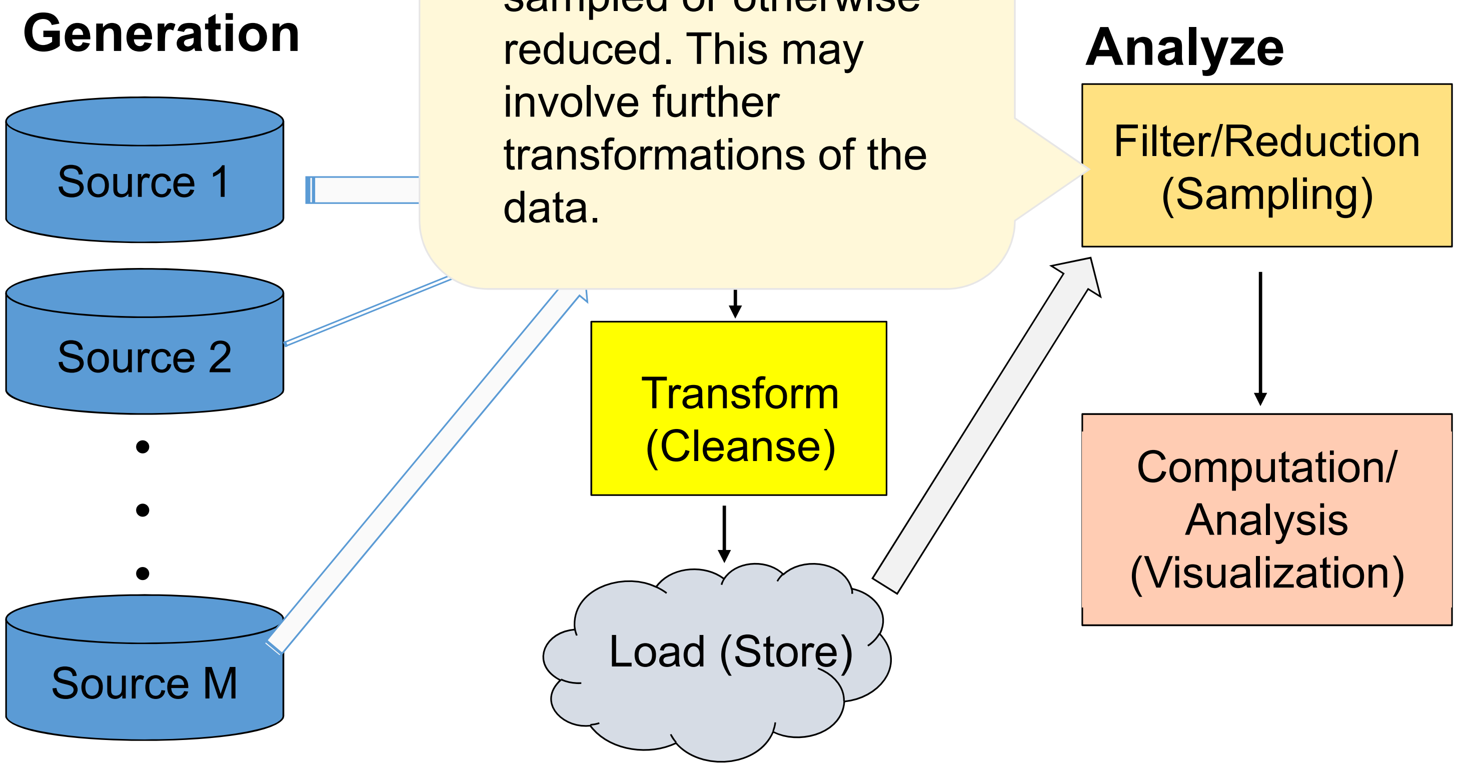
Similar to data collection errors in surveys; data may be erroneous or missing; data generating units may be self-selected; meta-data may be lacking or absent

Big Data Process Map



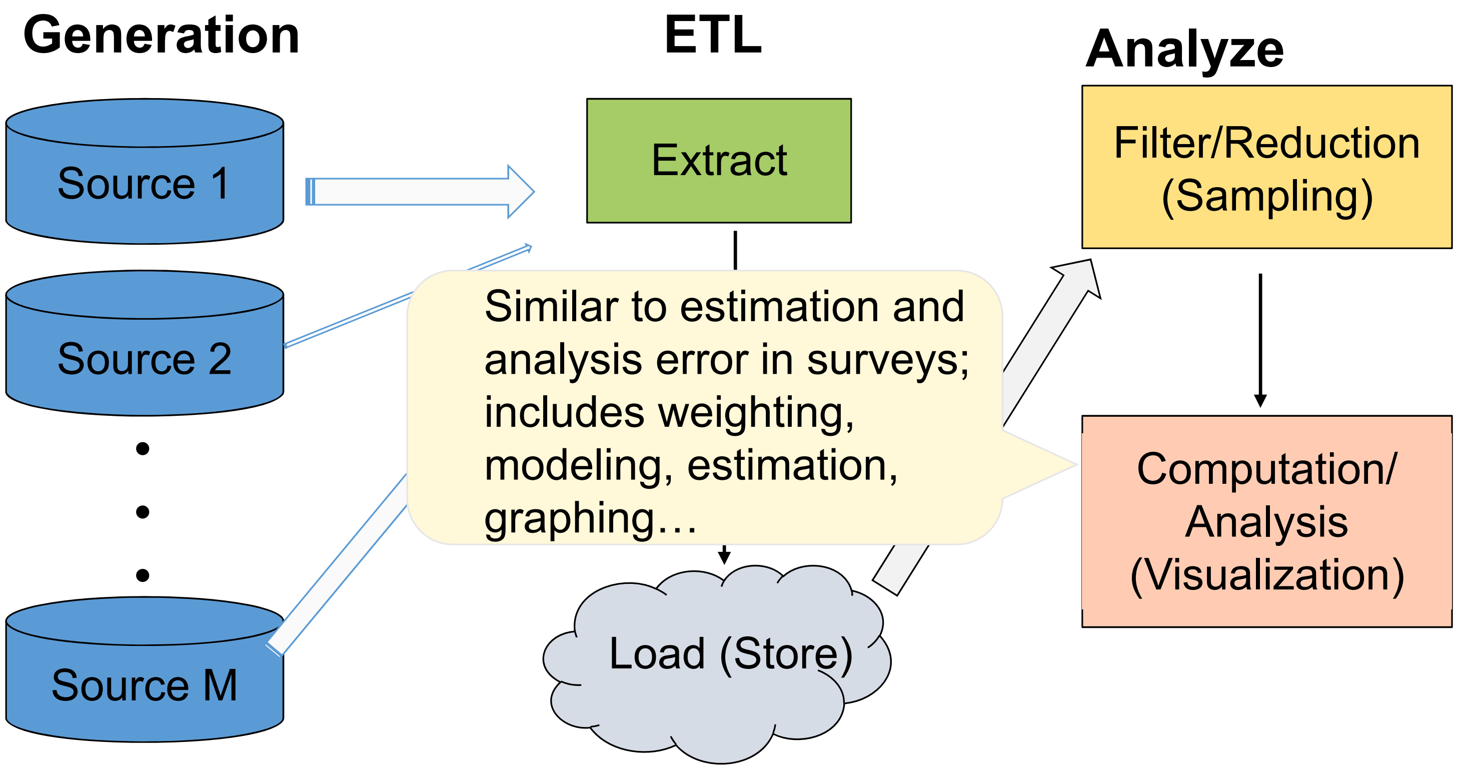
Similar to data processing stages in surveys; includes creating or enhancing meta-data; record matching; variable coding, editing, data munging or scrubbing, and data integration

Big Data Process Map



Data are filtered, sampled or otherwise reduced. This may involve further transformations of the data.

Big Data Process Map

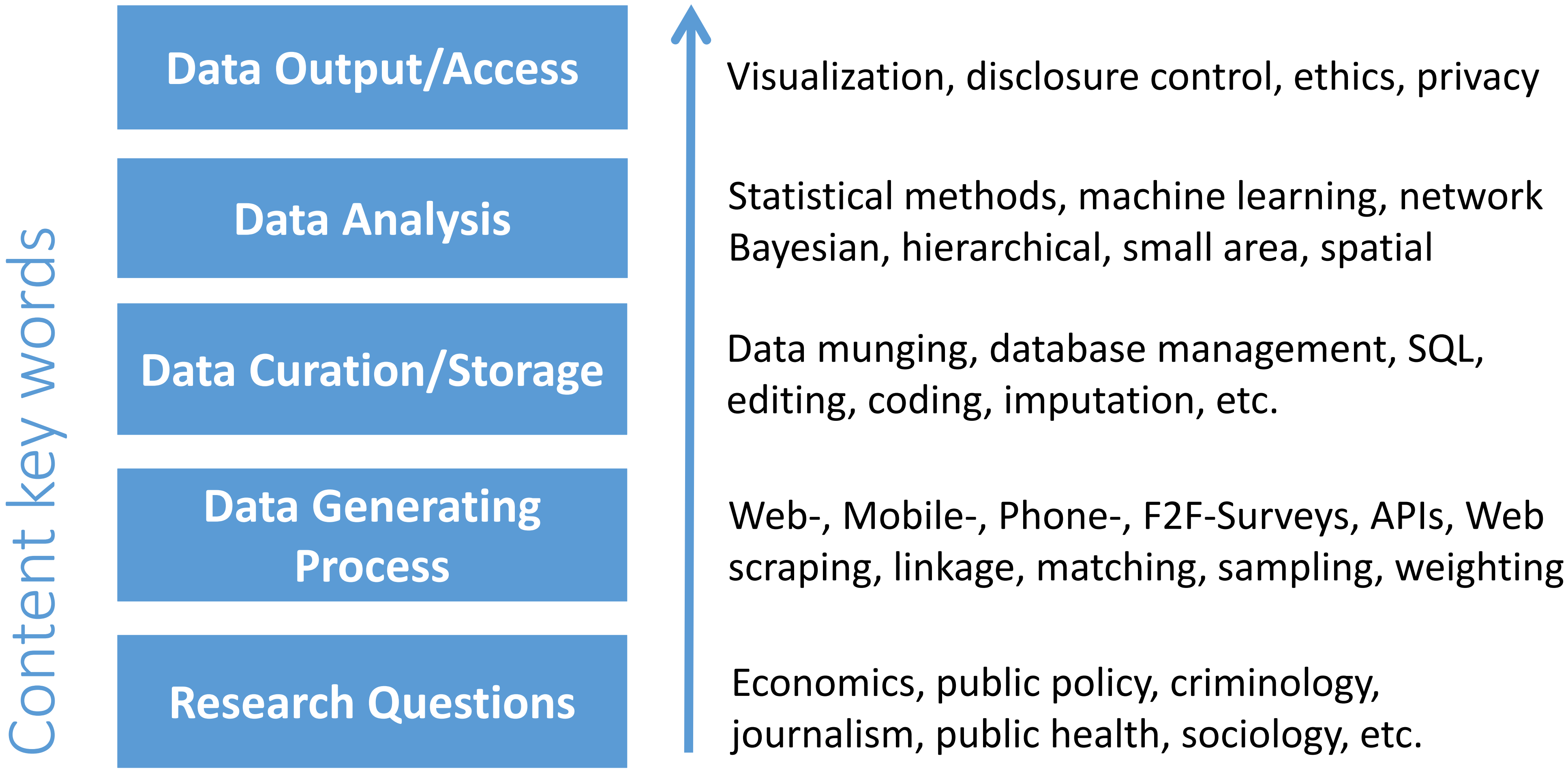


Similar to estimation and analysis error in surveys; includes weighting, modeling, estimation, graphing…

Key Ingredients for Valid Inference

1. **Data generating process needs to be known**
2. **Framework as tool to identify errors**
3. **Model or break confounders**
4. **Know your inferential goal**

The Skills



Content key words

Data Output/Access: Visualization, disclosure control, ethics, privacy

Data Analysis: Statistical methods, machine learning, network Bayesian, hierarchical, small area, spatial

Data Curation/Storage: Data munging, database management, SQL, editing, coding, imputation, etc.

Data Generating Process: Web-, Mobile-, Phone-, F2F-Surveys, APIs, Web scraping, linkage, matching, sampling, weighting

Research Questions: Economics, public policy, criminology, journalism, public health, sociology, etc.

Classical Statistical Approaches versus Statistical Machine Learning

Model Evaluation/Validation

Database Management

Programming with Big Data

When to use different data management and analysis technologies

Text files and scripting language	• Your data is small • Your analysis is simple • You do not expect to repeat analyses over time
Statistical packages	• Your data is modest in size • Your analysis maps well to your chosen statistical package
Relational database	• Your data is structured • You will be analyzing data repeatedly over time
NoSQL database	• Your data is unstructured • Your data is extremely large

Classical Statistical Approaches versus Statistical Machine Learning

Model Evaluation/Validation

Database Management

Programming with Big Data

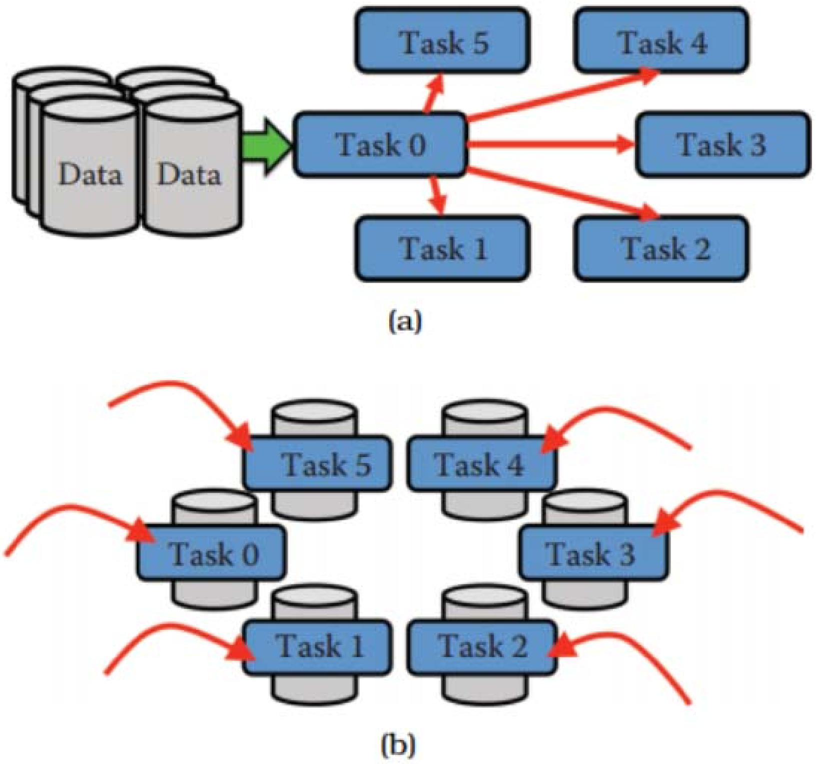


Figure 5.1. (a) The traditional parallel computing model where data is brought to the computing nodes. (b) Hadoop's parallel computing model: bringing compute to the data [241]

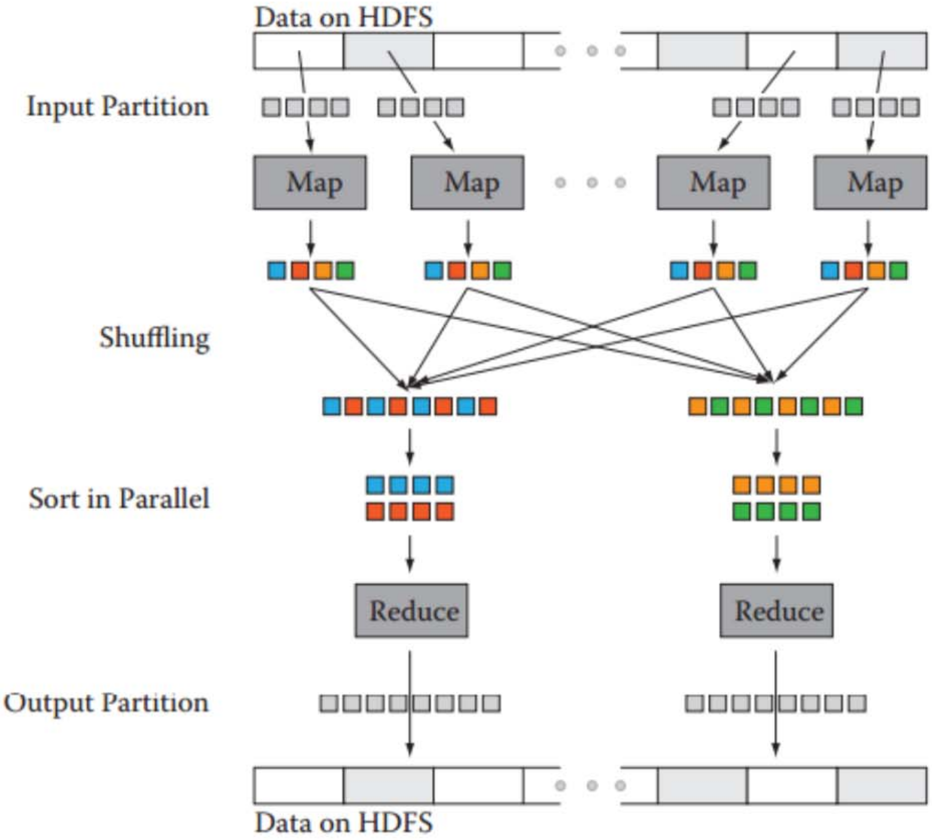
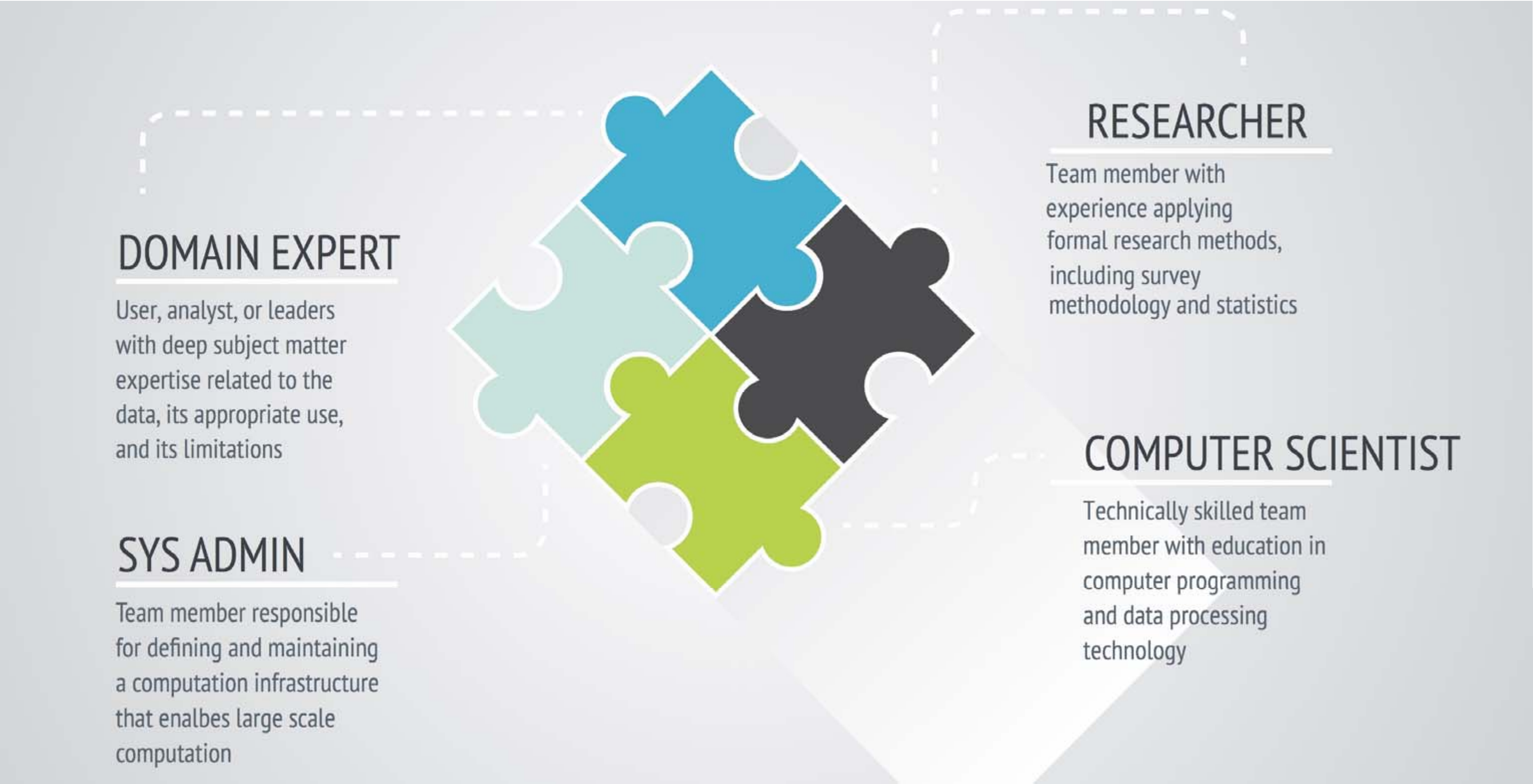


Figure 5.2. Data transfer and communication of a MapReduce job in Hadoop. Data blocks are assigned to several maps, which emit key–value pairs that are shuffled and sorted in parallel. The reduce step emits one or more pairs, with results stored on the HDFS



DOMAIN EXPERT

User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations

SYS ADMIN

Team member responsible for defining and maintaining a computation infrastructure that enalbes large scale computation

RESEARCHER

Team member with experience applying formal research methods, including survey methodology and statistics

COMPUTER SCIENTIST

Technically skilled team member with education in computer programming and data processing technology

Usher 2015

Learn More & Engage

http://coleridgeinitiative.org/

INTERNATIONAL PROGRAM IN SURVEY AND DATA SCIENCE

offered through the University of Mannheim and the Joint Program in Survey Methodology (Universities of Maryland and Michigan, Westat)

http://survey-data-science.net/

What Is ASA DataFest?

The American Statistical Association (ASA) DataFest is a
undergraduates work around the clock to find and shar
data set.

DataFest was founded at UCLA in 2011, when 30 studer
gathered for 48 intense hours to analyze five years of
Los
n and
hosted by several of the most prestigious colleges and
universities in the country. More than 500 students take

http://datafest.de

NSF

SPONSORED BY THE

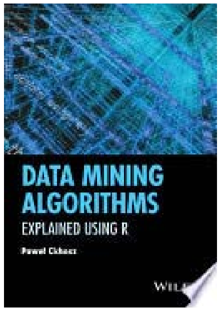
Federal Ministry
of Education
and Research

AUFSTIEG DURCH
BILDUNG >>
OFFENE HOCHSCHULEN

ASA

Data Mining/Machine Learning Resources

- http://www.dataminingconsultant.com/resources.htm
- Data Mining Algorithms Explained Using R (2015)
  - http://bit.ly/1yZYHjK



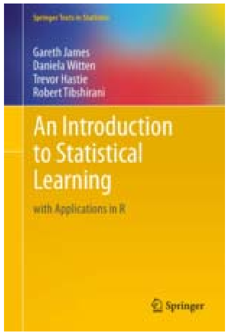
DATA MINING ALGORITHMS

EXPLAINED USING R

Paweł Cichosz

WILEY

- Data Mining for the Social Sciences (2015)
  - http://bit.ly/1DpPFC2
- An Introduction to Statistical Learning with Applications in R (2013)
  - Free PDF Version: http://bit.ly/1iUJso0
  - Online Resources for FREE lecture videos and labs in R
    - http://bit.ly/1snBMk5



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

- An overview of Machine Learning Functions available in R
  - http://cran.r-project.org/web/views/MachineLearning.html

DATA MINING for the SOCIAL SCIENCES

AN INTRODUCTION

PAUL ATTEWELL and DAVID MONAGHAN