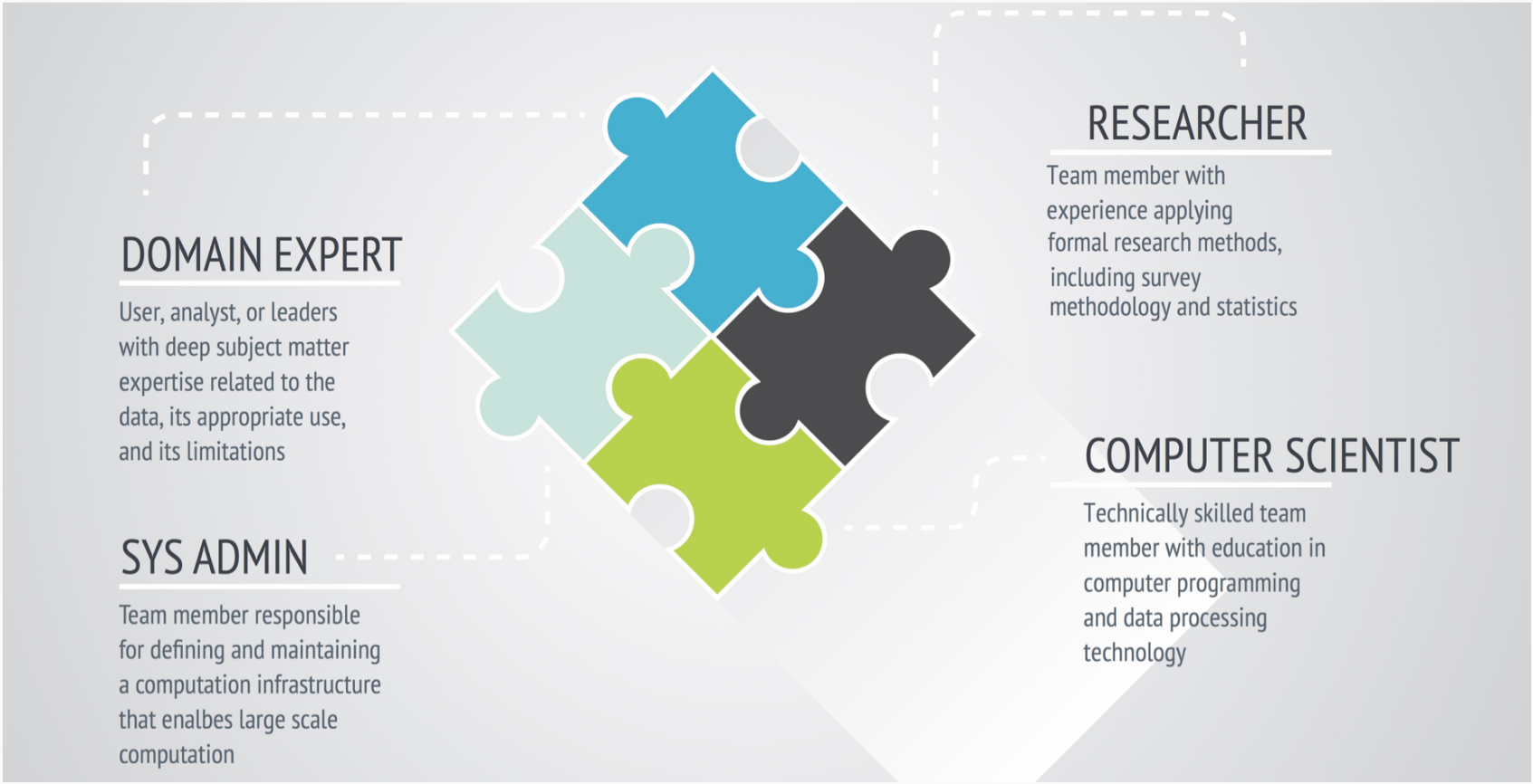
The BigData Toolbox

Introduction to Big Data for Social Science

Frauke Kreuter[1] ...

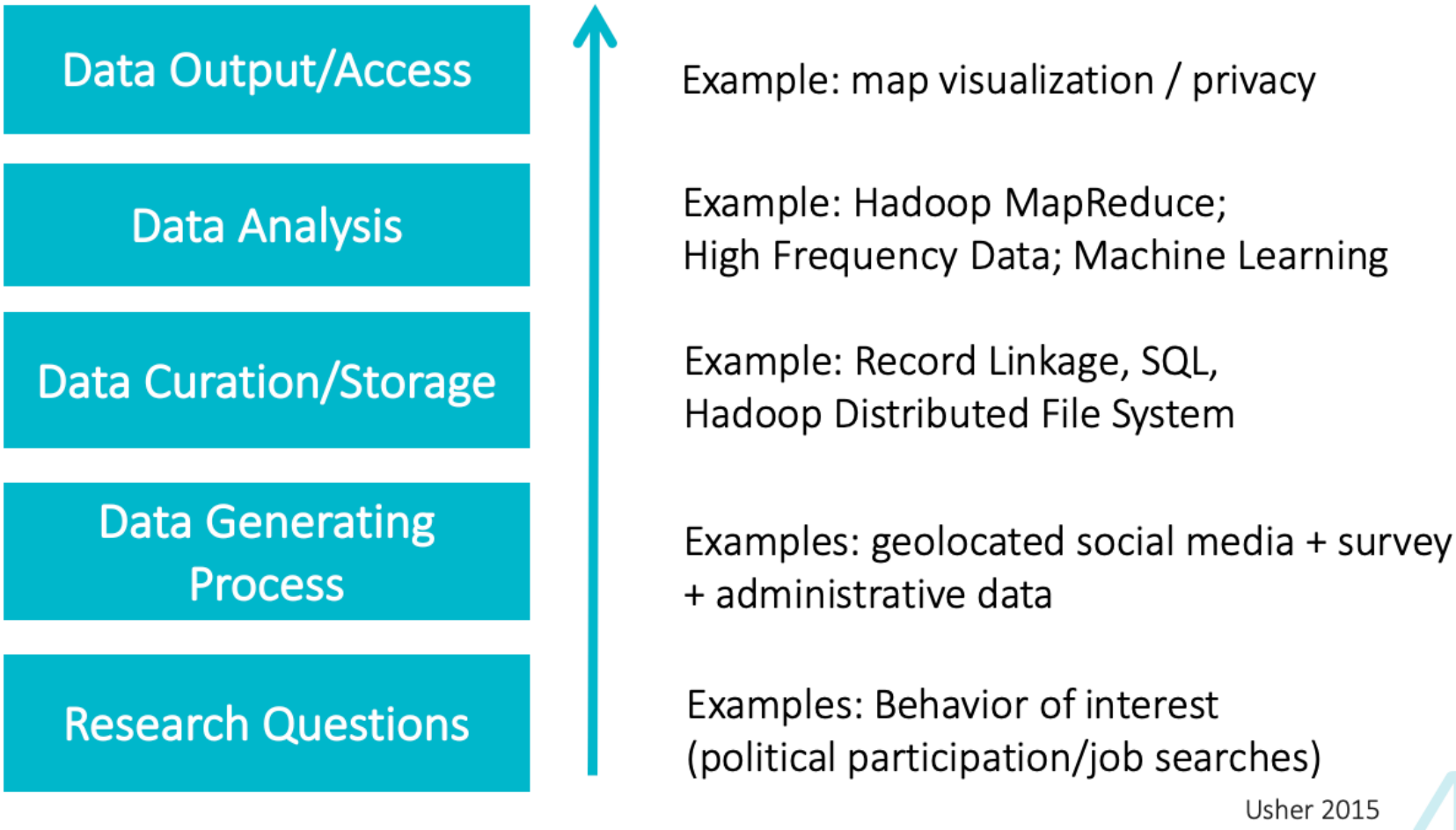
June 3–4, 2019

[1]fkreuter@umd.edu

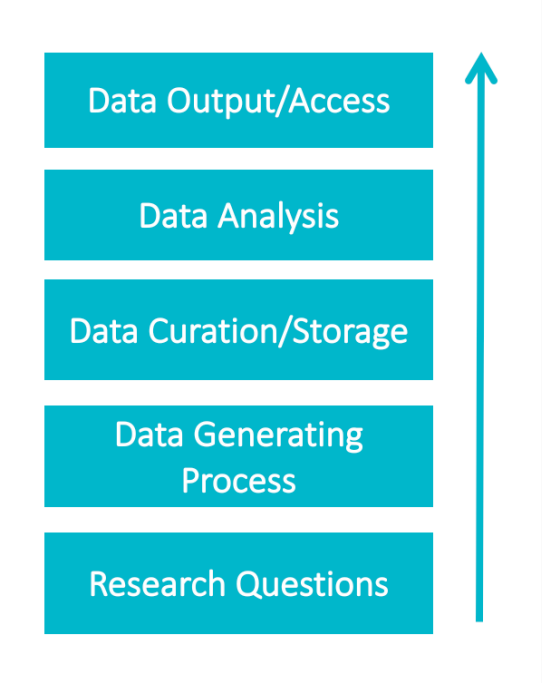

DOMAIN EXPERT
User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations
SYS ADMIN
Team member responsible for defining and maintaining a computation infrastructure that enalbes large scale computation
RESEARCHER
Team member with experience applying formal research methods, including survey methodology and statistics
COMPUTER SCIENTIST
Technically skilled team member with education in computer programming and data processing technology

The BigData Toolbox

What tasks are required to get there?

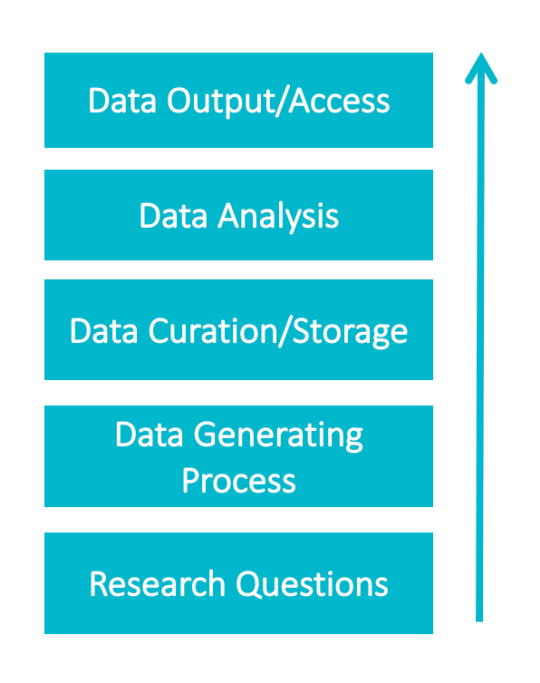


What tasks are required to get there?



- ✓ ...ready made tools available
- ✓ ...some new IF different inferential goals
- ✓ ...caveat: silos, historic systems
- ✓ ...very good sense of quality
- ✓ ...new units of analyses

What tasks are required to get there?



✓ ...ready made tools available

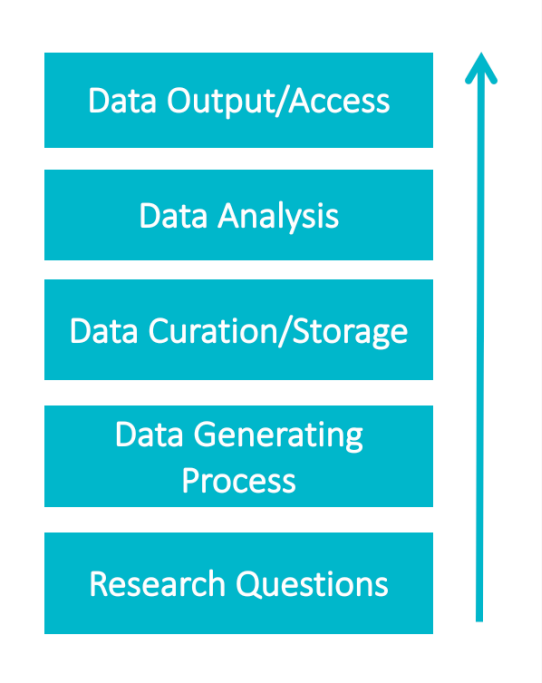
✓ ...some new IF different inferential goals

✓ ...caveat: silos, historic systems

✓ ...very good sense of quality

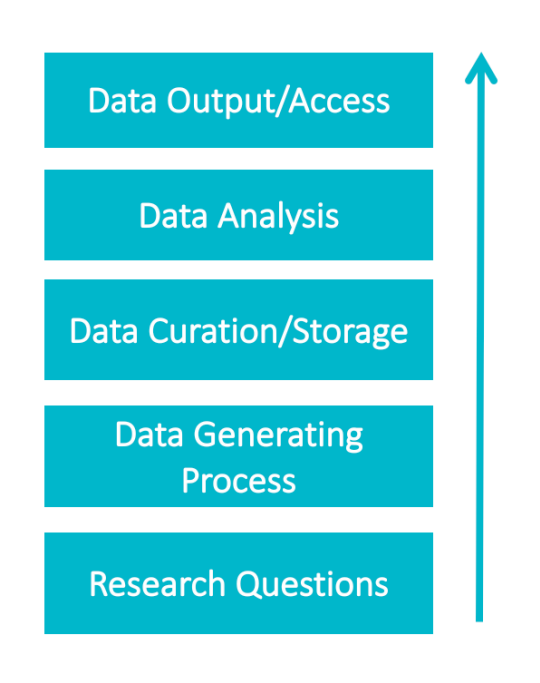
✓ ...new units of analyses

What tasks are required to get there?



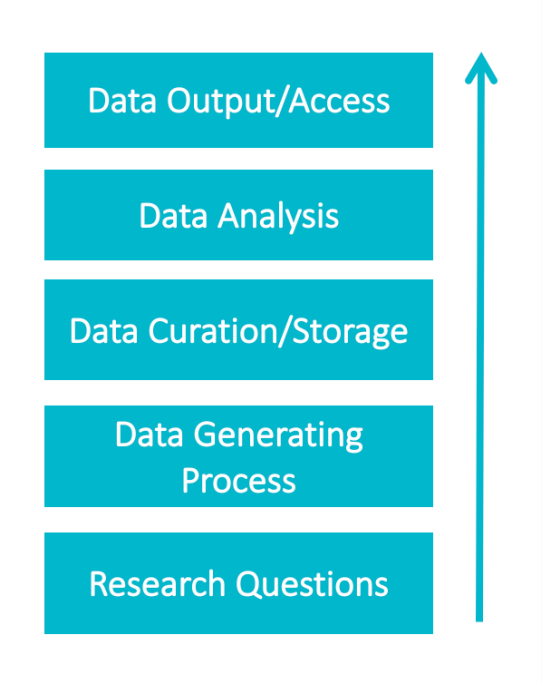
- ✓ ...ready made tools available
- ✓ ...some new IF different inferential goals
- ✓ ...caveat: silos, historic systems
- ✓ ...very good sense of quality
- ✓ ...new units of analyses

What tasks are required to get there?



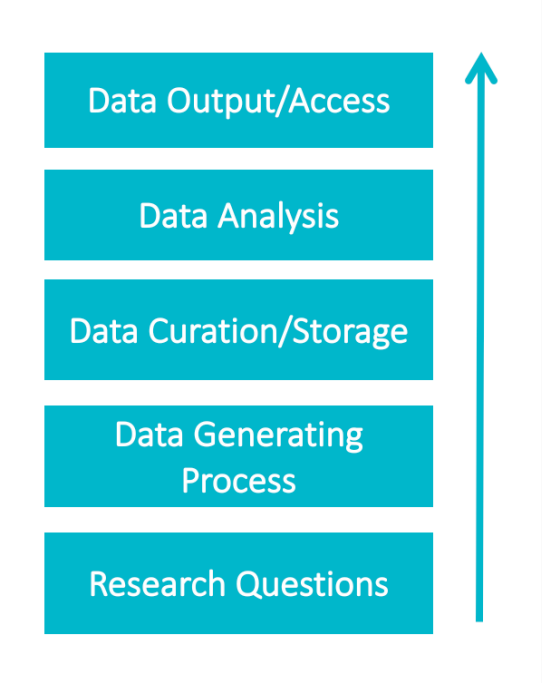
- ✓ ...ready made tools available
- ✓ ...some new IF different inferential goals
- ✓ ...caveat: silos, historic systems
- ✓ ...very good sense of quality
- ✓ ...new units of analyses

What tasks are required to get there?



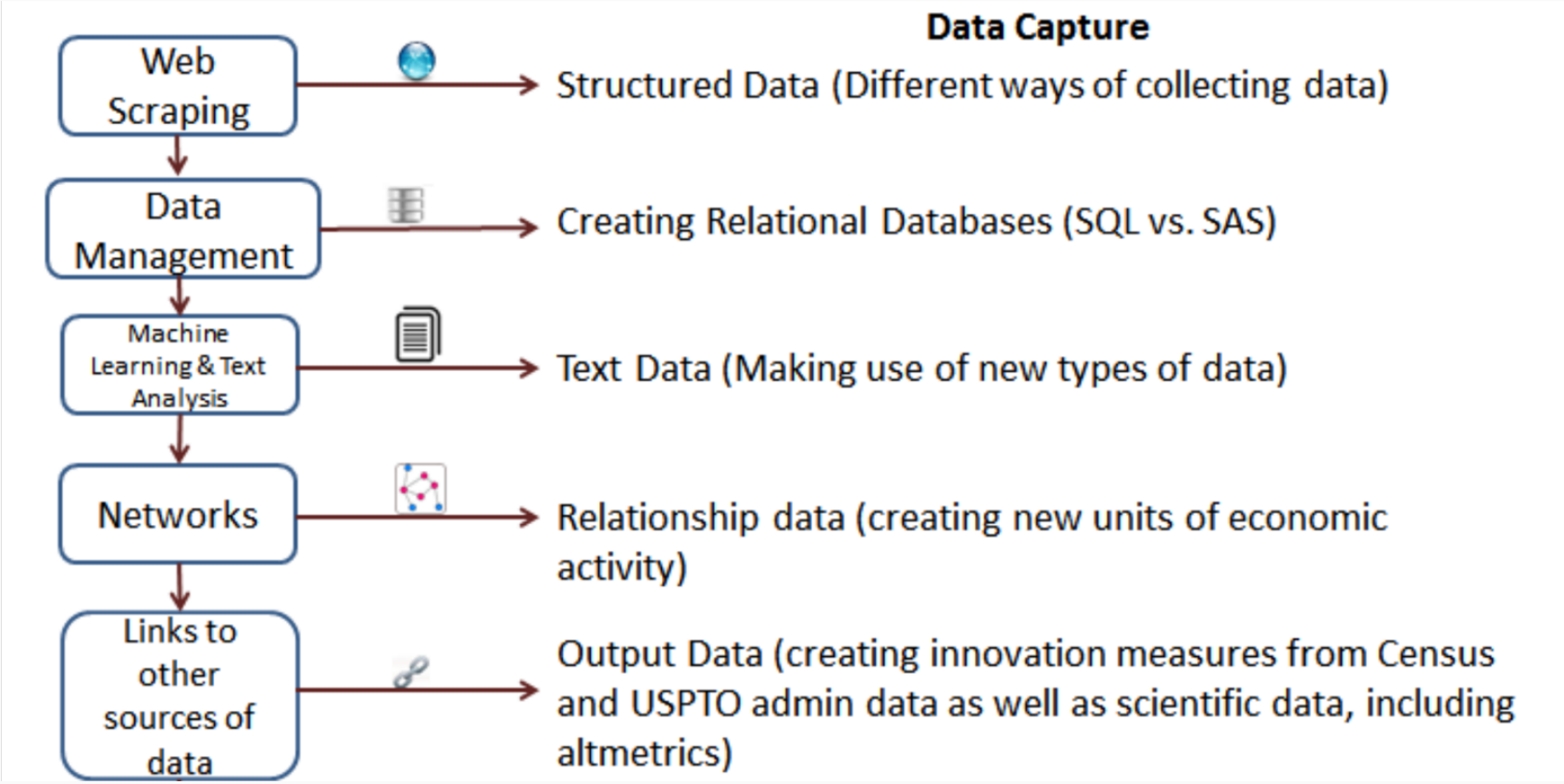
- ✓ ...ready made tools available
- ✓ ...some new IF different inferential goals
- ✓ ...caveat: silos, historic systems
- ✓ ...very good sense of quality
- ✓ ...new units of analyses

What tasks are required to get there?

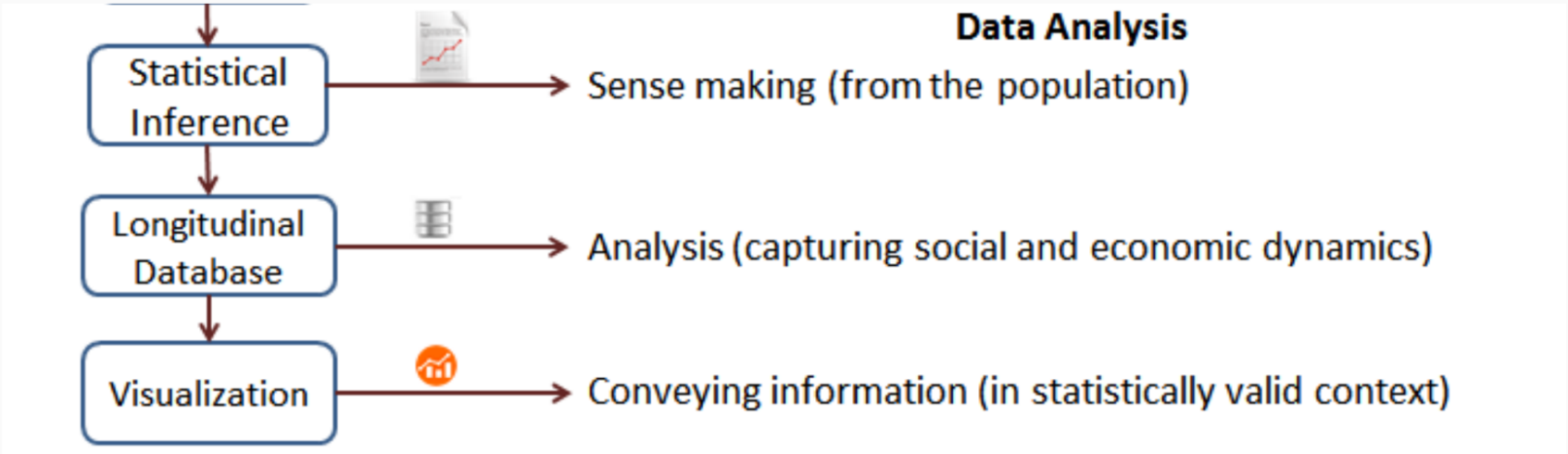


- ✓ ...ready made tools available
- ✓ ...some new IF different inferential goals
- ✓ ...caveat: silos, historic systems
- ✓ ...very good sense of quality
- ✓ ...new units of analyses

Big Data for Federal Agencies (Lane 2014)



Big Data for Federal Agencies (Lane 2014)

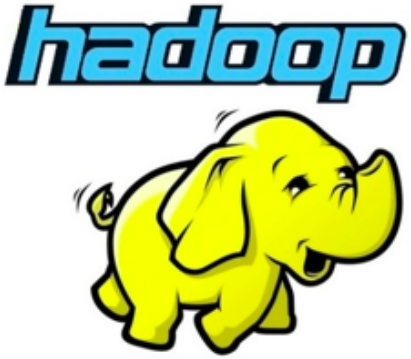


Computer scientist

- Data preparation
- MapReduce algorithms
- Python/R programming
- Hadoop ecosystem

System Administrator

- Storage systems
  (MySQL, Hbase, Spark)
- Cloud computing:
  - Amazon Web Services (AWS)
  - Google Compute Engine
- Hadoop ecosystem



Database Management

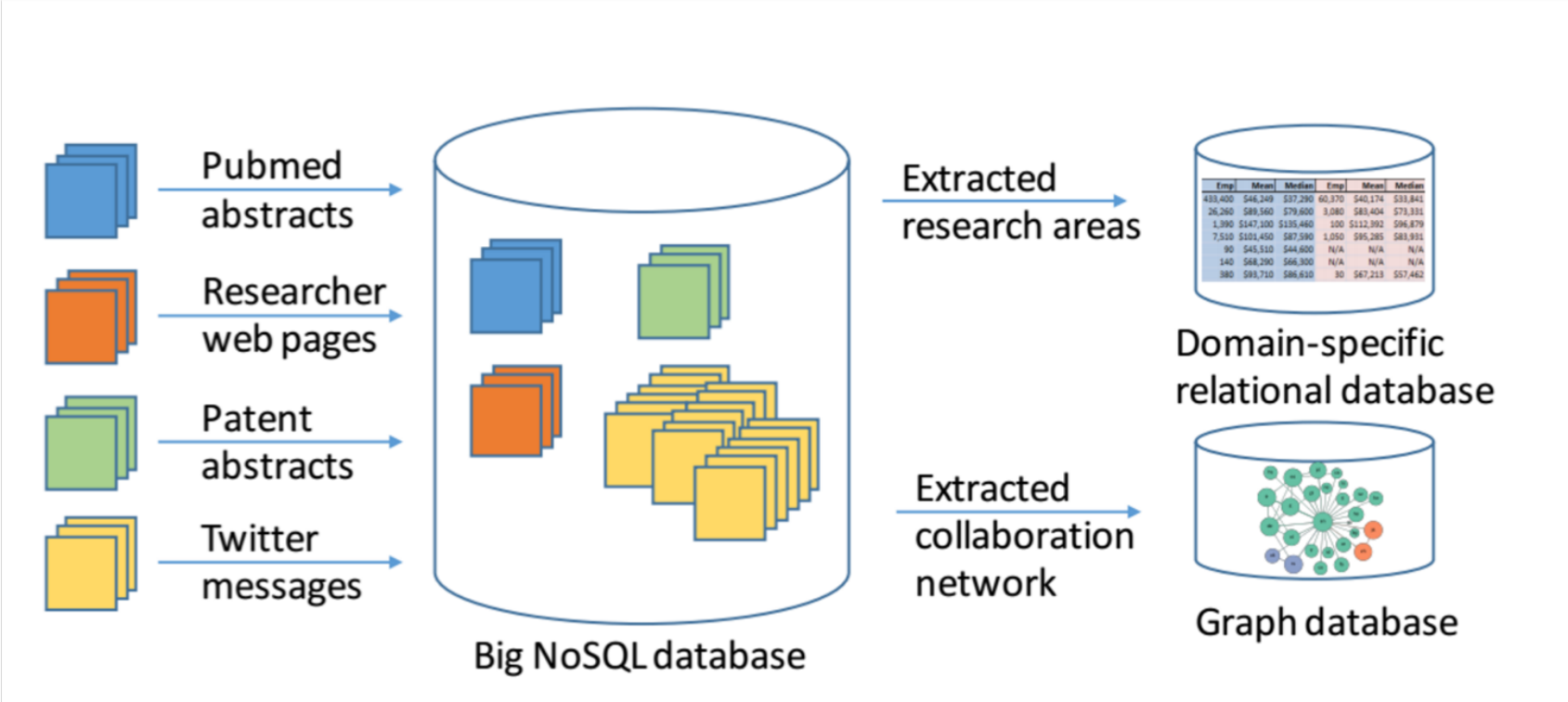
When to use different data management and analysis technologies

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

Something we do often

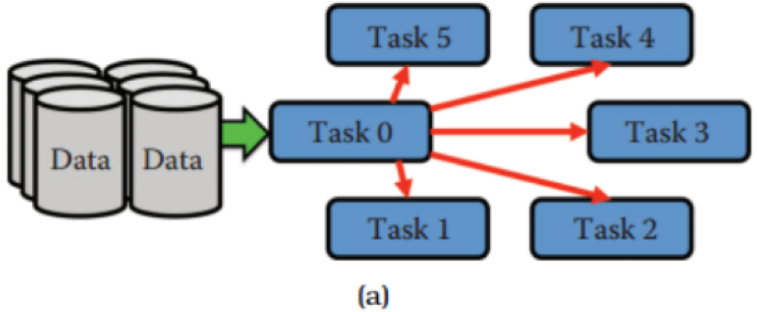
- Storing data in different tables
- Normalizing: Organizing columns and tables of a relational database to minimize data redundancy
- Creating metadata
- Creating unique IDs

Something we do less often



Programming BD

Handoop



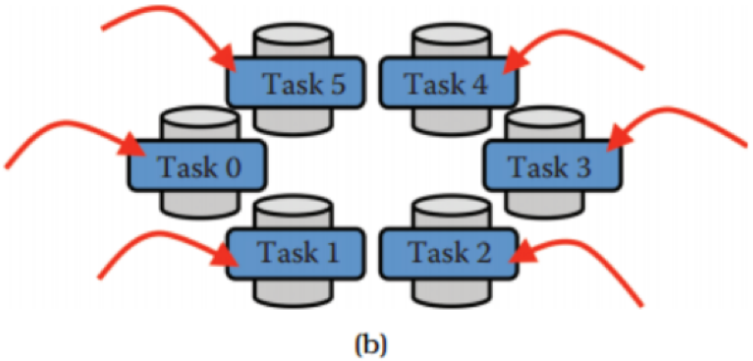


Figure 5.1. (a) The traditional parallel computing model where data is brought to the computing nodes. (b) Hadoop's parallel computing model: bringing compute to the data [241]

MapReduce

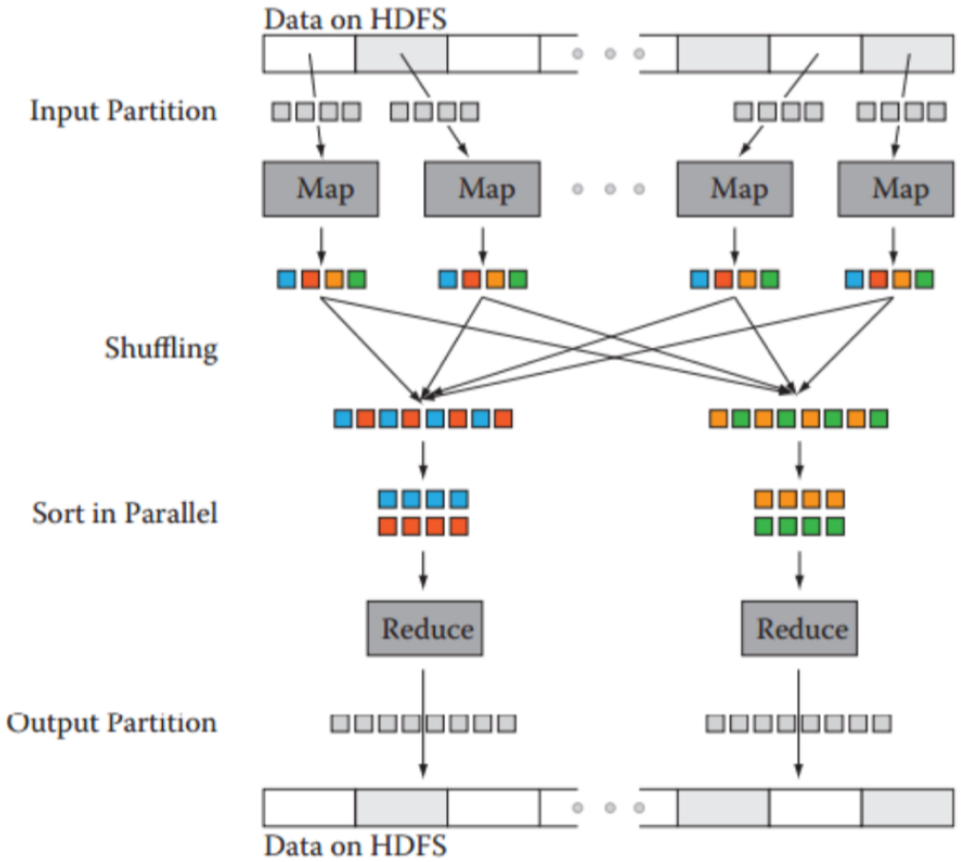


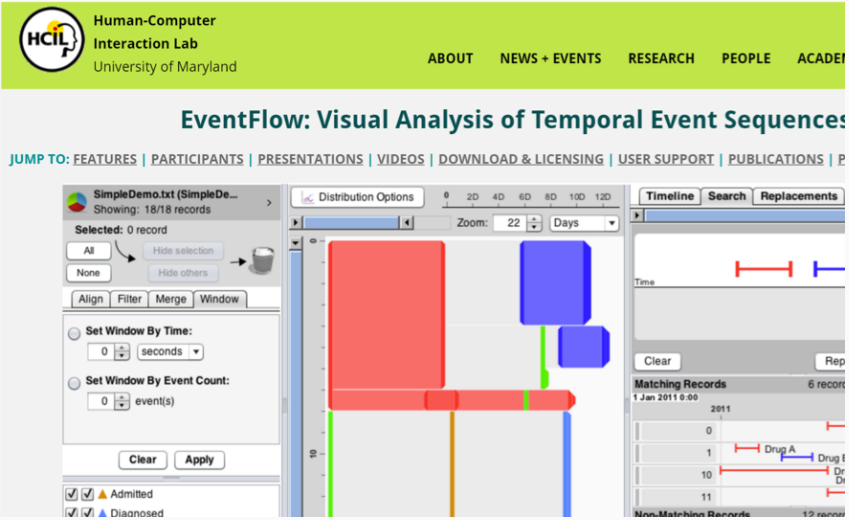
Figure 5.2. Data transfer and communication of a MapReduce job in Hadoop. Data blocks are assigned to several maps, which emit key–value pairs that are shuffled and sorted in parallel. The reduce step emits one or more pairs, with results stored on the HDFS

Good for certain tasks, but not all. Subsampling (and knowing how) still very much needed.

Visualization

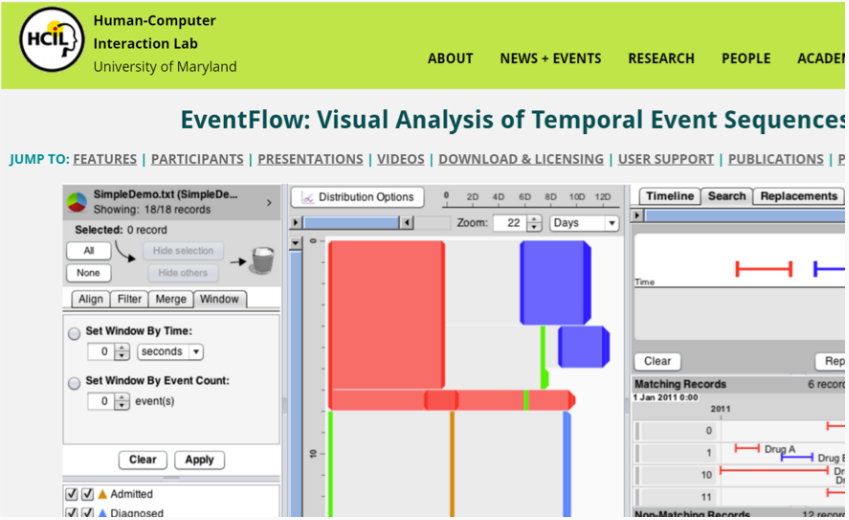
Tools


+ a b | e a u
Palantir

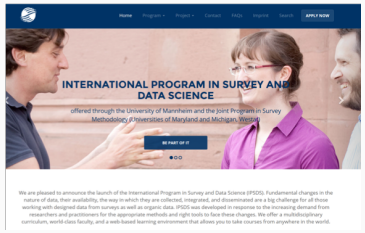

Human-Computer
Interaction Lab
University of Maryland
ABOUT
NEWS + EVENTS
RESEARCH
PEOPLE
EventFlow: Visual Analysis of Temporal Event Sequences
JUMP TO: FEATURES | PARTICIPANTS | PRESENTATIONS | VIDEOS | DOWNLOAD & LICENSING | USER SUPPORT | PUBLICATIONS
SimpleDemo.txt (SimpleDe...
Showing: 18/18 records
Selected: 0 record
All
None
Hide selection
Hide others
Align
Filter
Merge
Window
Set Window By Time:
seconds
Set Window By Event Count:
event(s)
Clear
Apply
Admitted
Distribution Options
Zoom:
Days
Timeline
Search
Replacements
Time
Clear
Matching Records
Drug A

Tools


+ a b | e a u
Palantir


Human-Computer
Interaction Lab
University of Maryland
ABOUT
NEWS + EVENTS
RESEARCH
PEOPLE
EventFlow: Visual Analysis of Temporal Event Sequences
JUMP TO: FEATURES | PARTICIPANTS | PRESENTATIONS | VIDEOS | DOWNLOAD & LICENSING | USER SUPPORT | PUBLICATIONS
SimpleDemo.txt (SimpleDe...
Showing: 18/18 records
Selected: 0 record
All
None
Hide selection
Hide others
Align
Filter
Merge
Window
Set Window By Time:
seconds
Set Window By Event Count:
event(s)
Clear
Apply
Admitted
Distribution Options
Zoom:
Days
Timeline
Search
Replacements
Time
Clear
Matching Records
Drug A

http://www.applieddataanalytics.org/



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

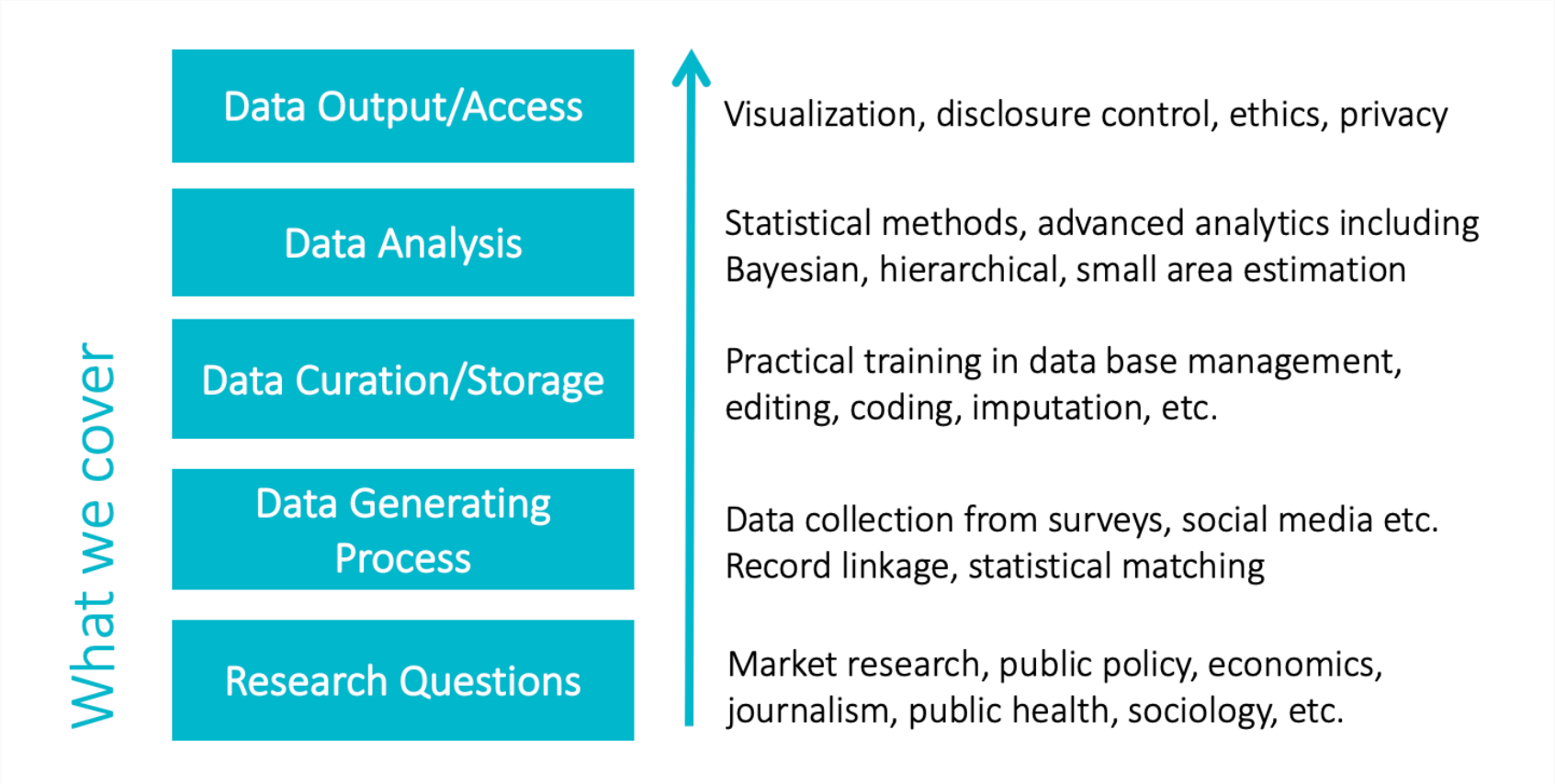
http://datafest.de

SPONSORED BY THE

Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG
OFFENE HOCHSCHULEN

What we cover



...our students are ready for it

DataFest