SFU

Slides: Tatyana Mozgacheva
tmozgach@sfu.ca

Challenges and Opportunities with Big Data

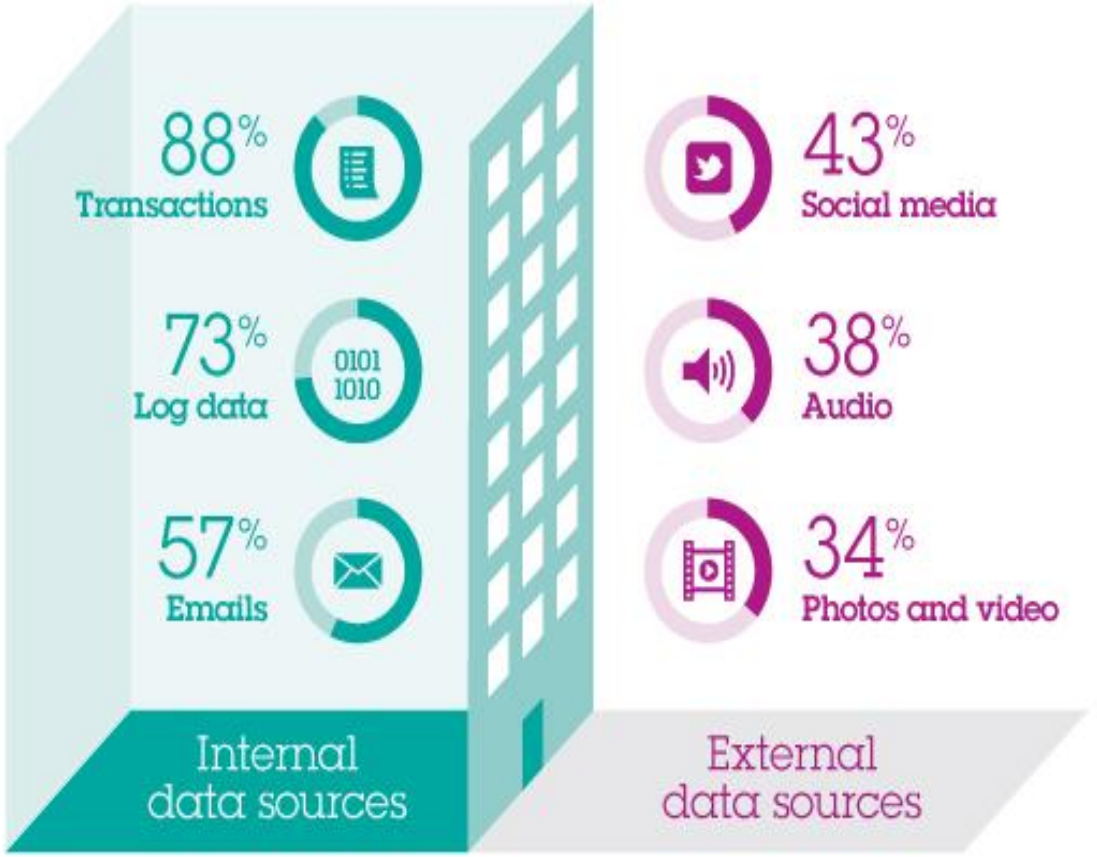
Agenda

- ✓ What is the **Big Data**?
- ✓ **Application** of the Big Data
- ✓ **3 Vs** of the Big Data
- ✓ The Big Data analysis **pipeline**
- ✓ **Challenges**

SFU

Big Data

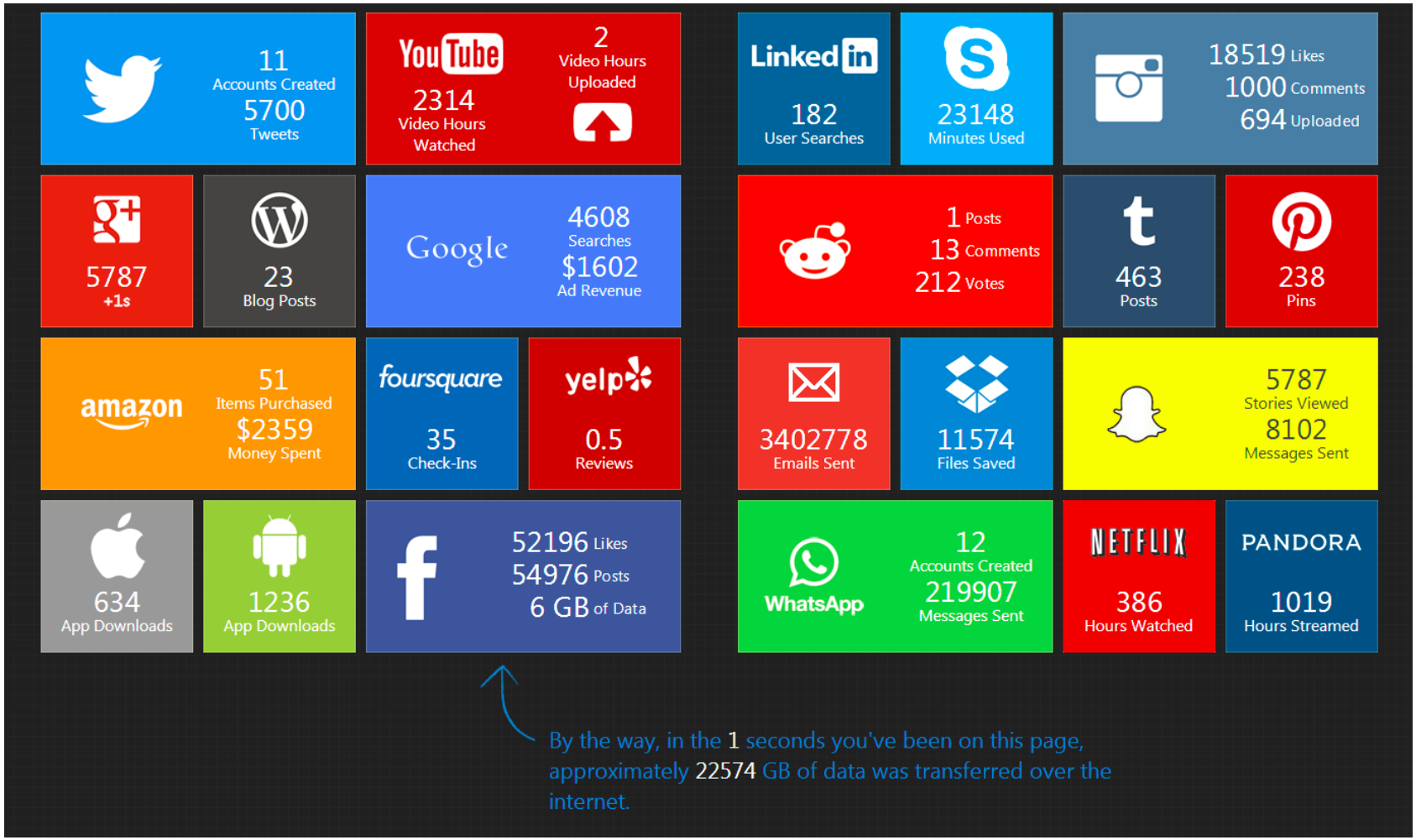
Every day, we create **2.5 quintillion bytes of data** — so much that 90% of the data in the world today has been created in the **last two years alone.**



Source: IBM Big Data And Analytics Hub

SFU

How much data is generated in social media in 1 second ?



Source: "digitalinformationworld.com – Social Media in Real-Time

SFU

Application Of Big Data

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

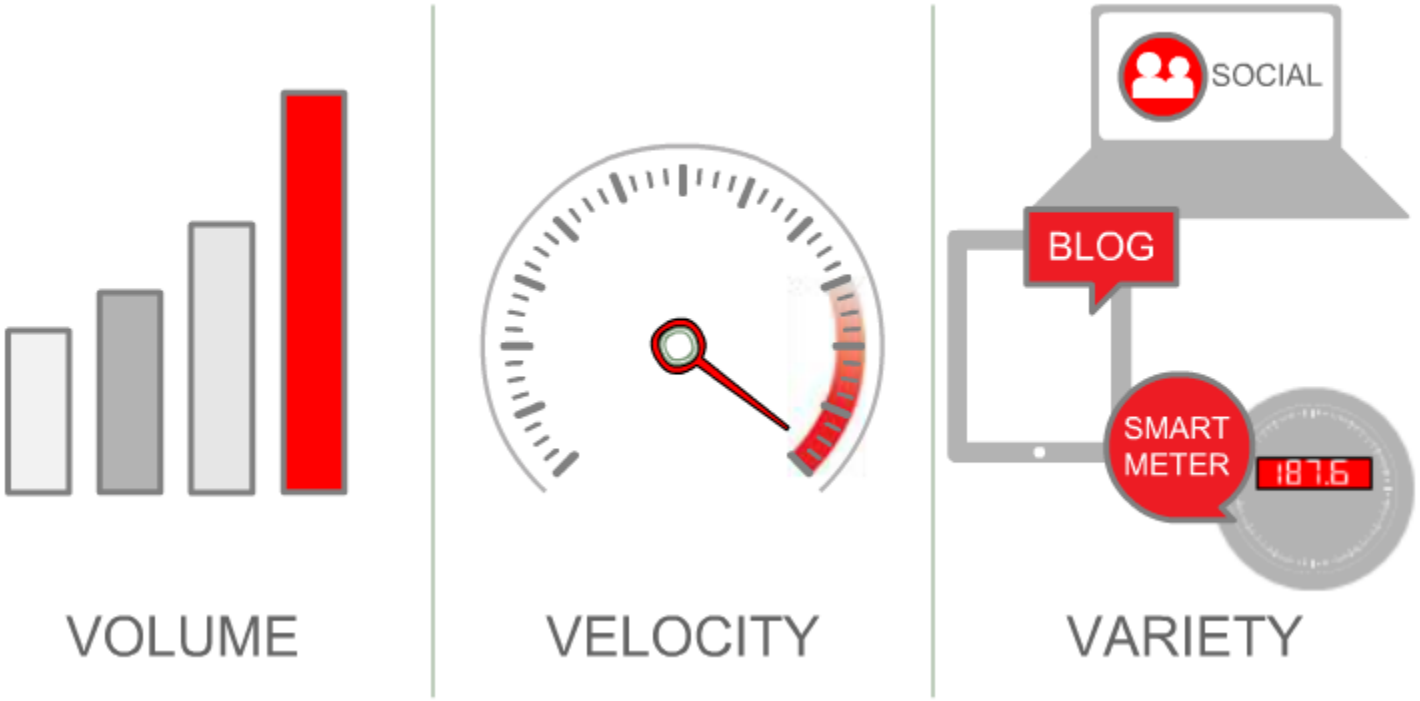
Search Quality

SFU

Our future? ☺

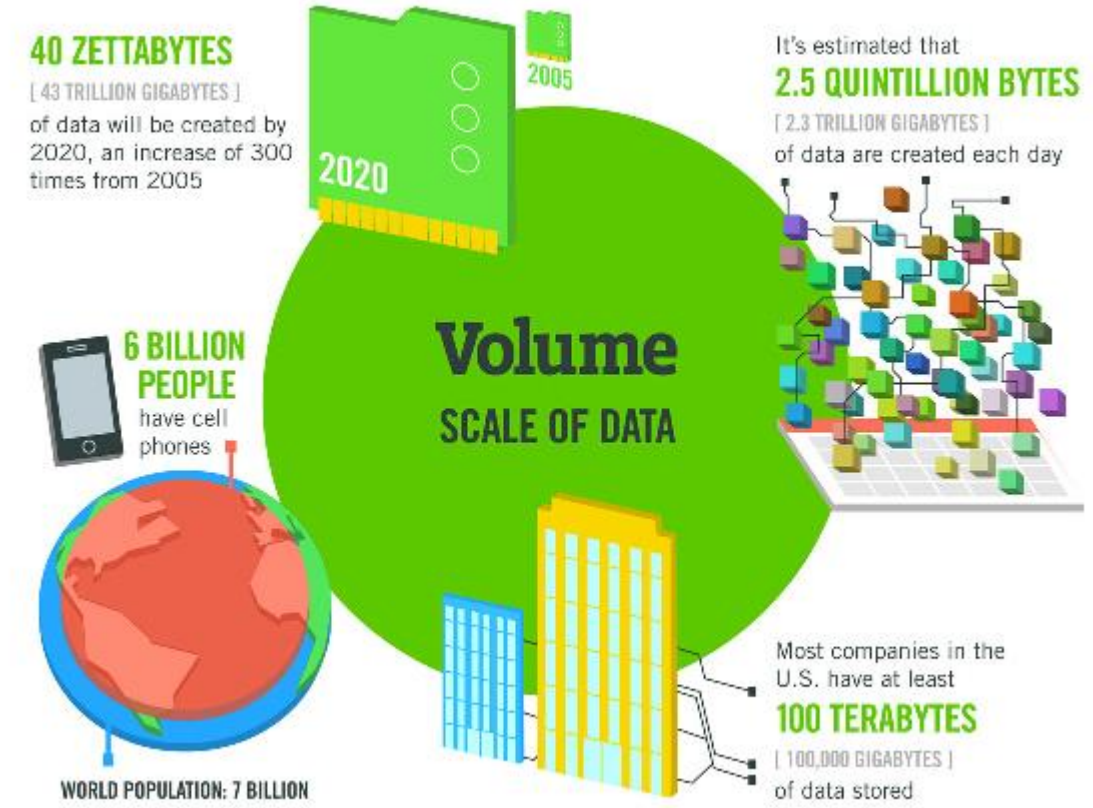
SFU

Big data spans three dimensions:



SFU

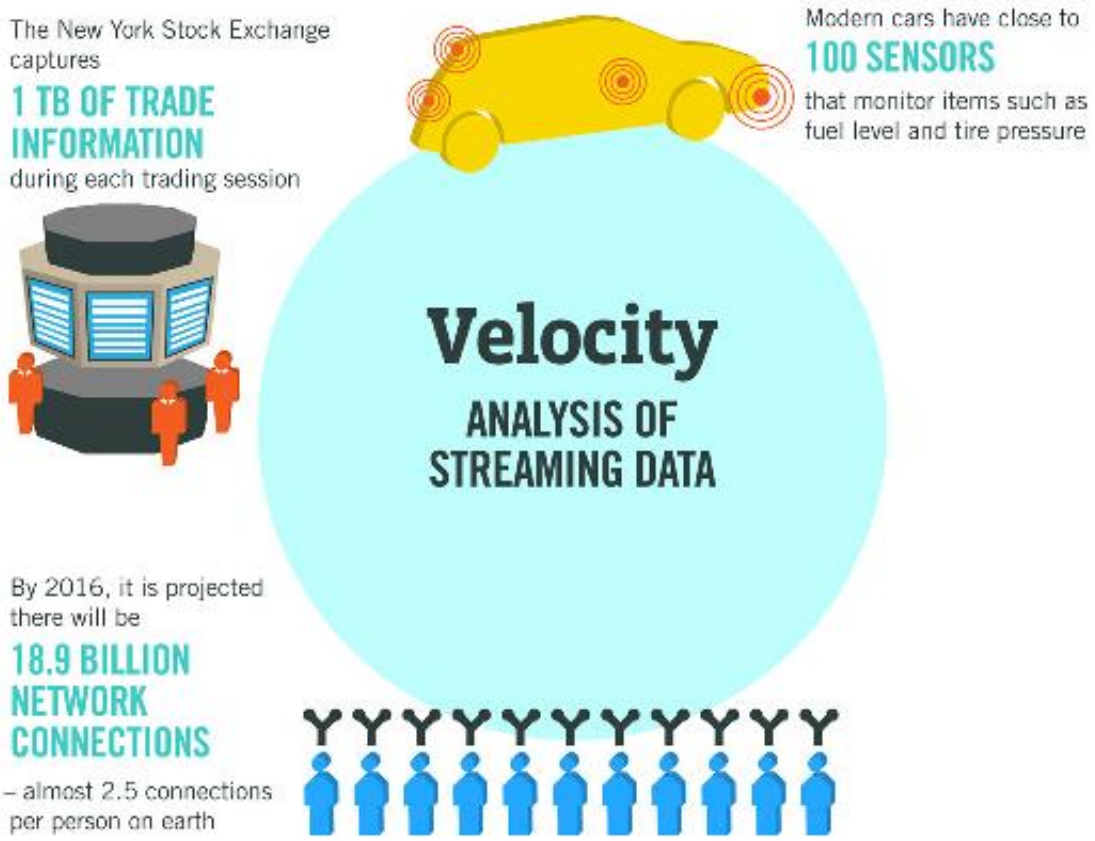
Volume



Source: McKinsey Global Institute, Twitter, CISCO, Gartner, EMC, SAS, IBM

SFU

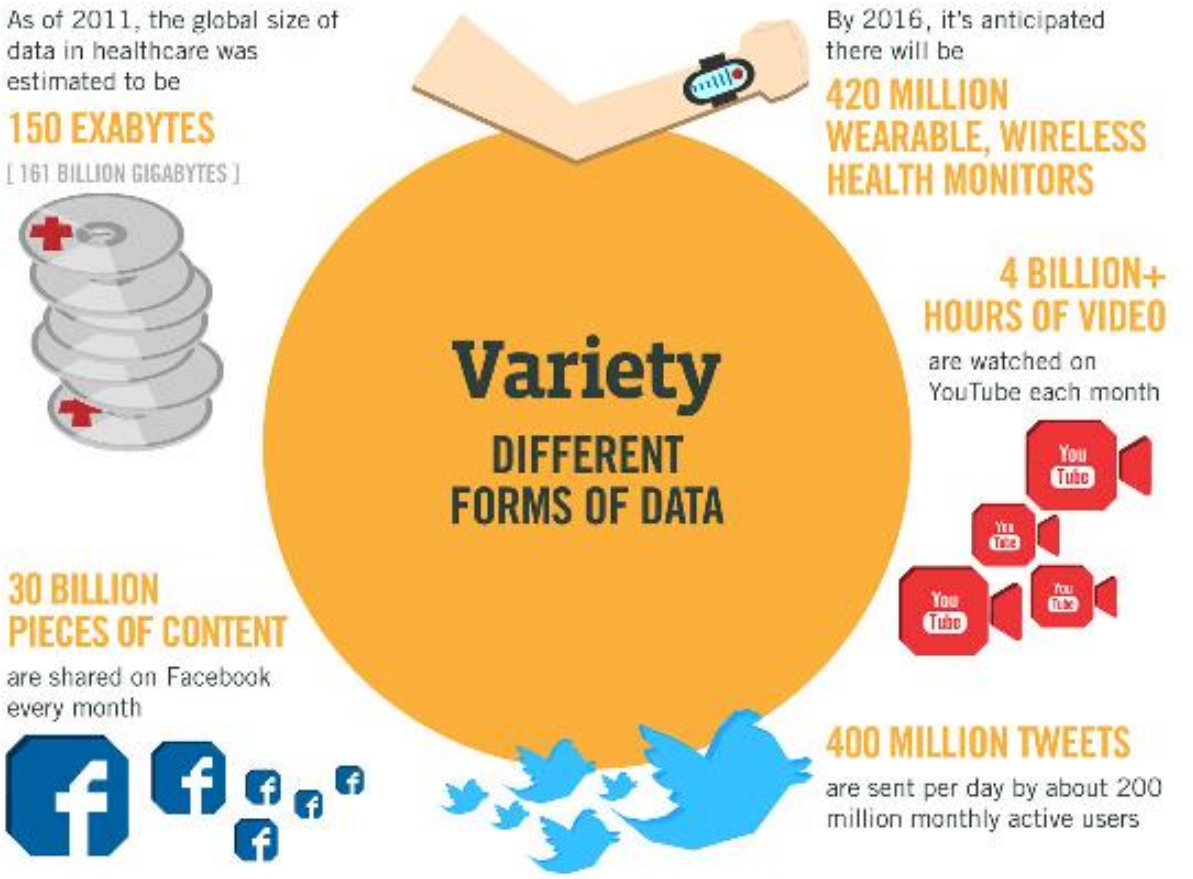
Velocity



Source: McKinsey Global Institute, Twitter, CISCO, Gartner, EMC, SAS, IBM

SFU

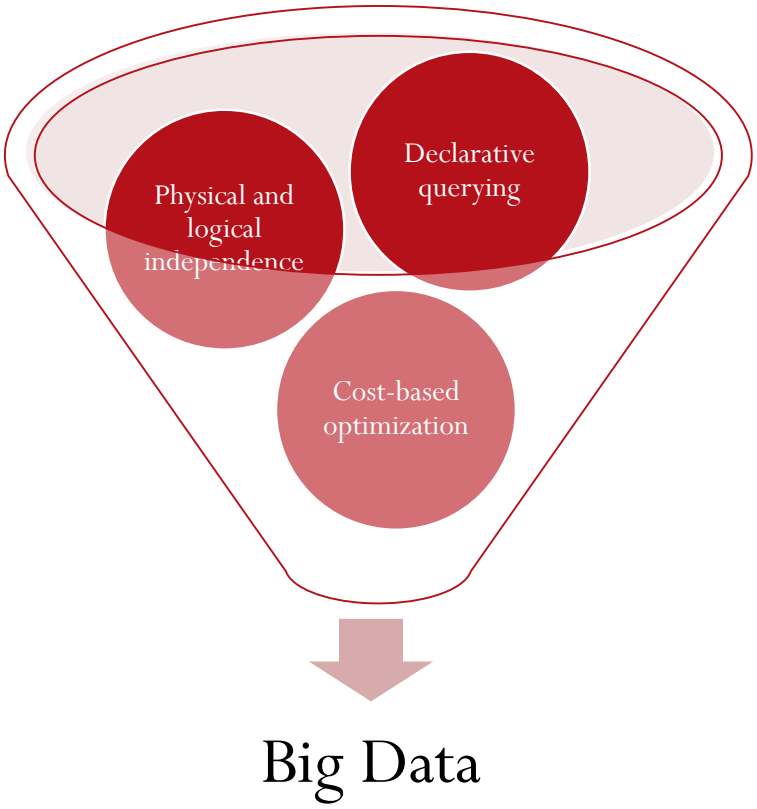
Variety



Source: McKinsey Global Institute, Twitter, CISCO, Gartner, EMC, SAS, IBM

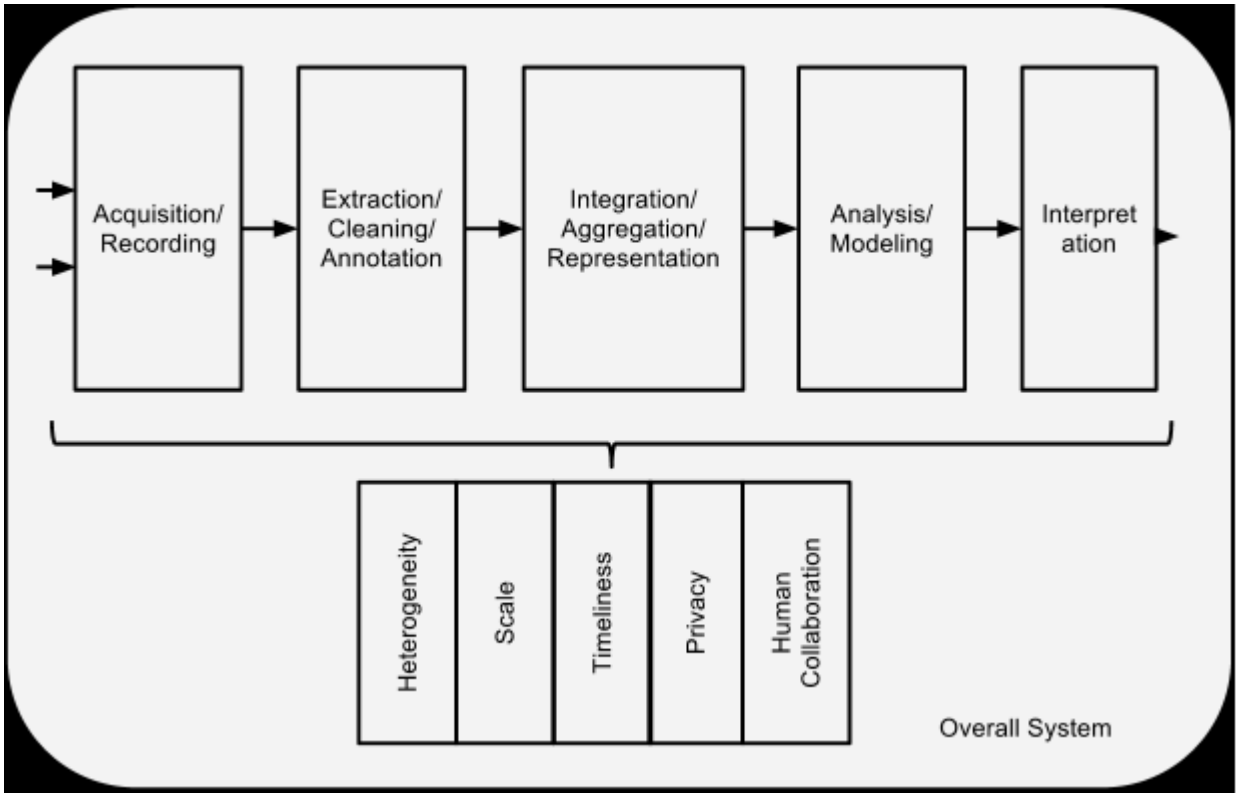
SFU

What has been Achieved…



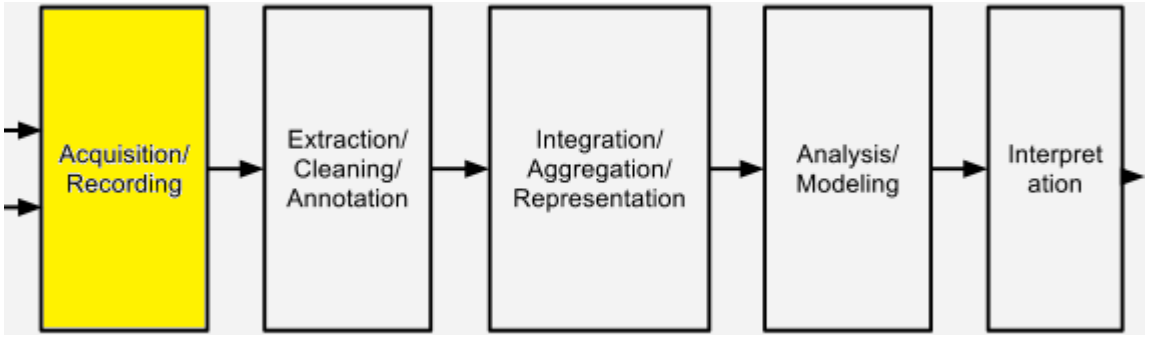
SFU

Phases and challenges in the Big Data Analysis Pipeline



SFU

The Big Data Analysis Pipeline



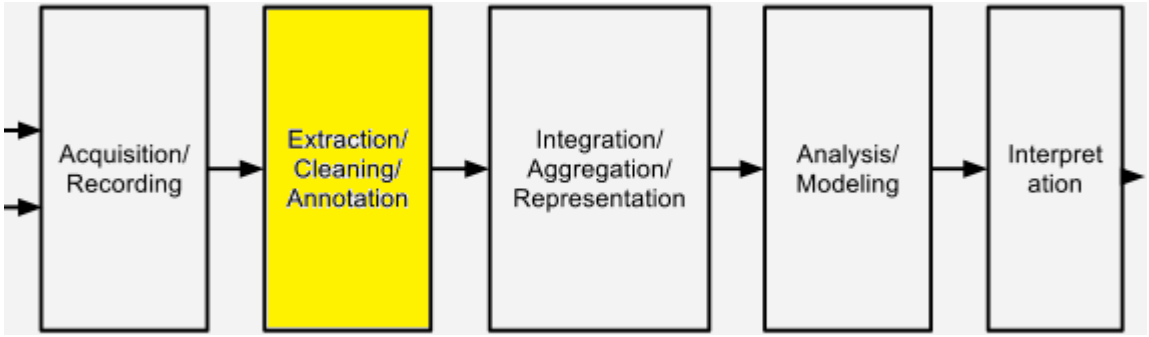
SFU

Data Acquisition and Recording

- What data to **keep**?
- What to **discard**?
- How to filter out the data **on the fly**?
- What is right **metadata**?

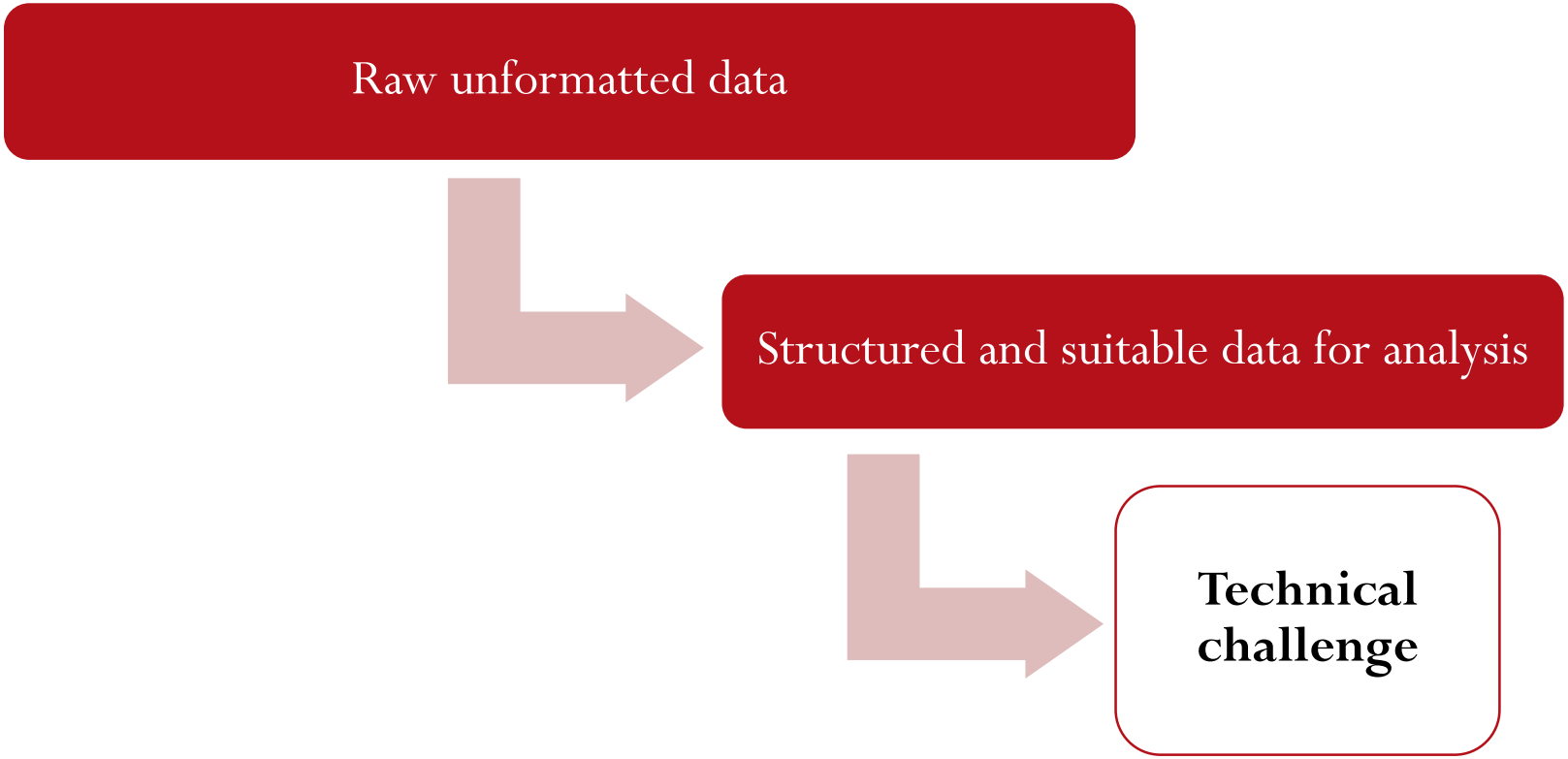
SFU

The Big Data Analysis Pipeline



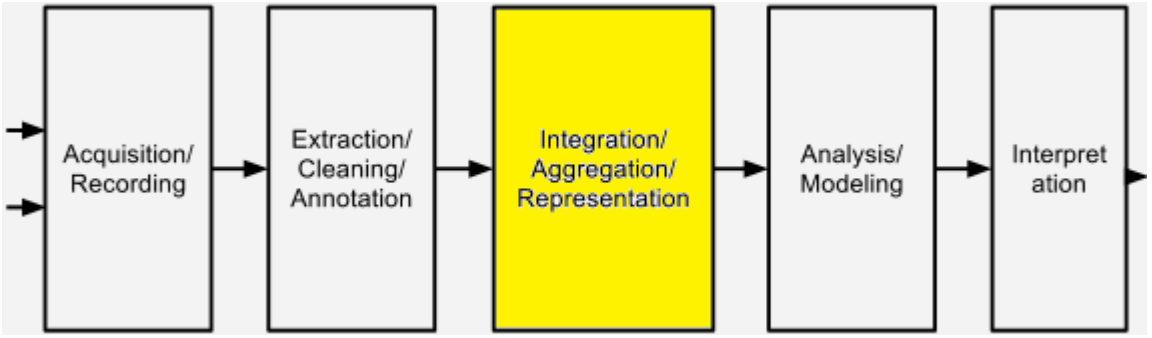
SFU

Information Extraction and Cleaning



SFU

The Big Data Analysis Pipeline



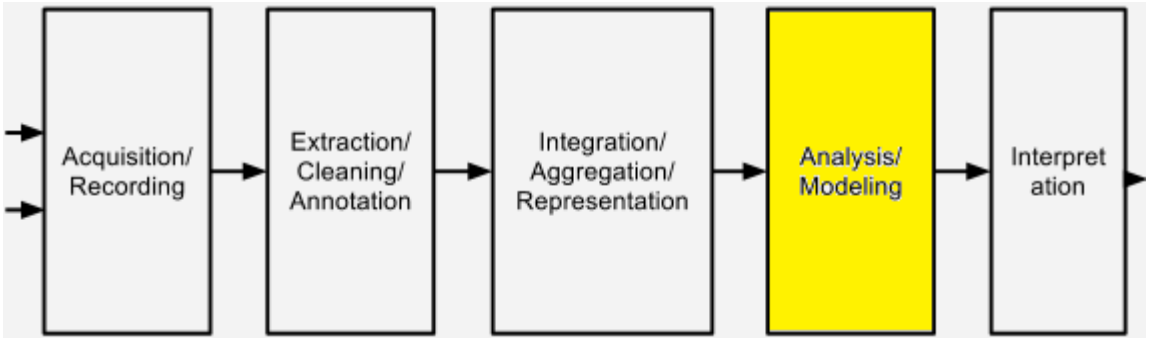
SFU

Data Integration, Aggregation, and Representation

- How to combine **heterogeneous** data?
- How to select a **suitable database** design?

SFU

The Big Data Analysis Pipeline



SFU

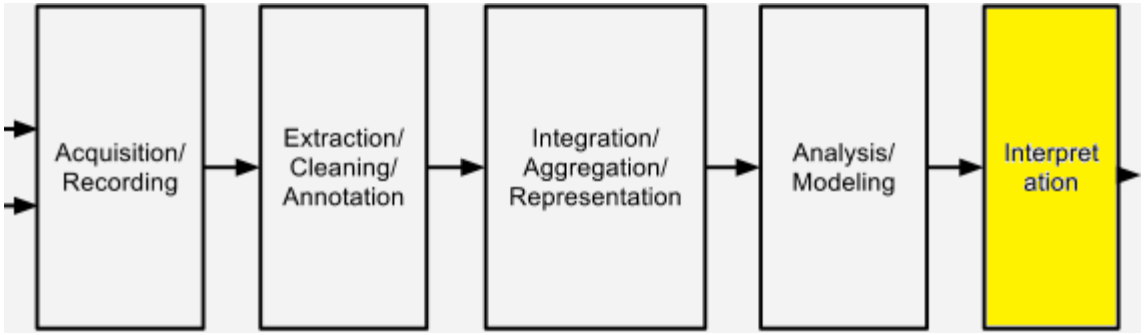
Query Processing, Data Modeling, and Analysis

- Querying and mining Big Data are fundamentally **different** from traditional statistical analysis.

- **Information redundancy** can be explored for:
  - missing data,
  - to crosscheck conflicting cases,
  - to validate trustworthy relationships,
  - to uncover hidden relationships and models.

- Lack of coordination between database systems with analytics packages (e.g. statistical analyses).

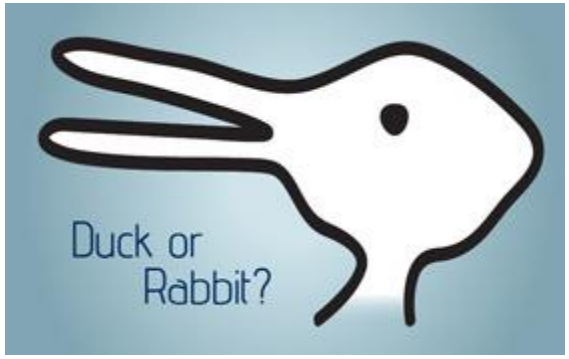
The Big Data Analysis Pipeline



SFU

Interpretation

- **Simplify** life of analyst.
- Necessity of **provenance** data to repeat the analysis with different assumptions, parameters, or data sets.



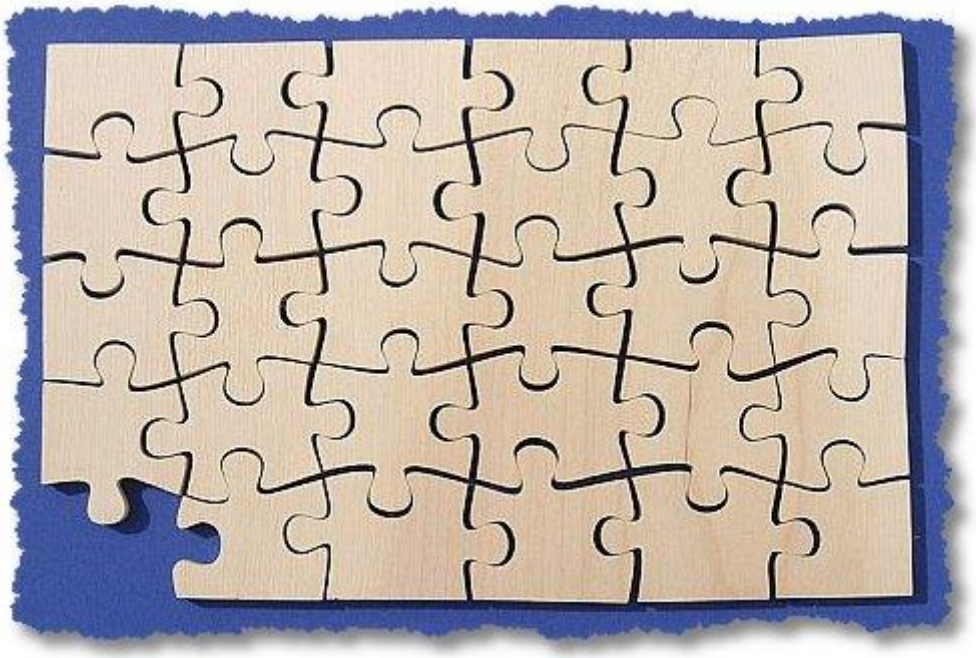
SFU

Challenges in the Big Data Analysis Pipeline

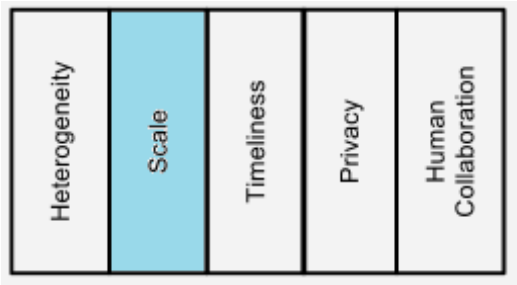
SFU

Heterogeneity and Incompleteness

- Data from **different sources**/platforms.
- Data **formats** are **different.**
- Data **missing** due to security, privacy, or other reasons.



Challenges in the Big Data Analysis Pipeline

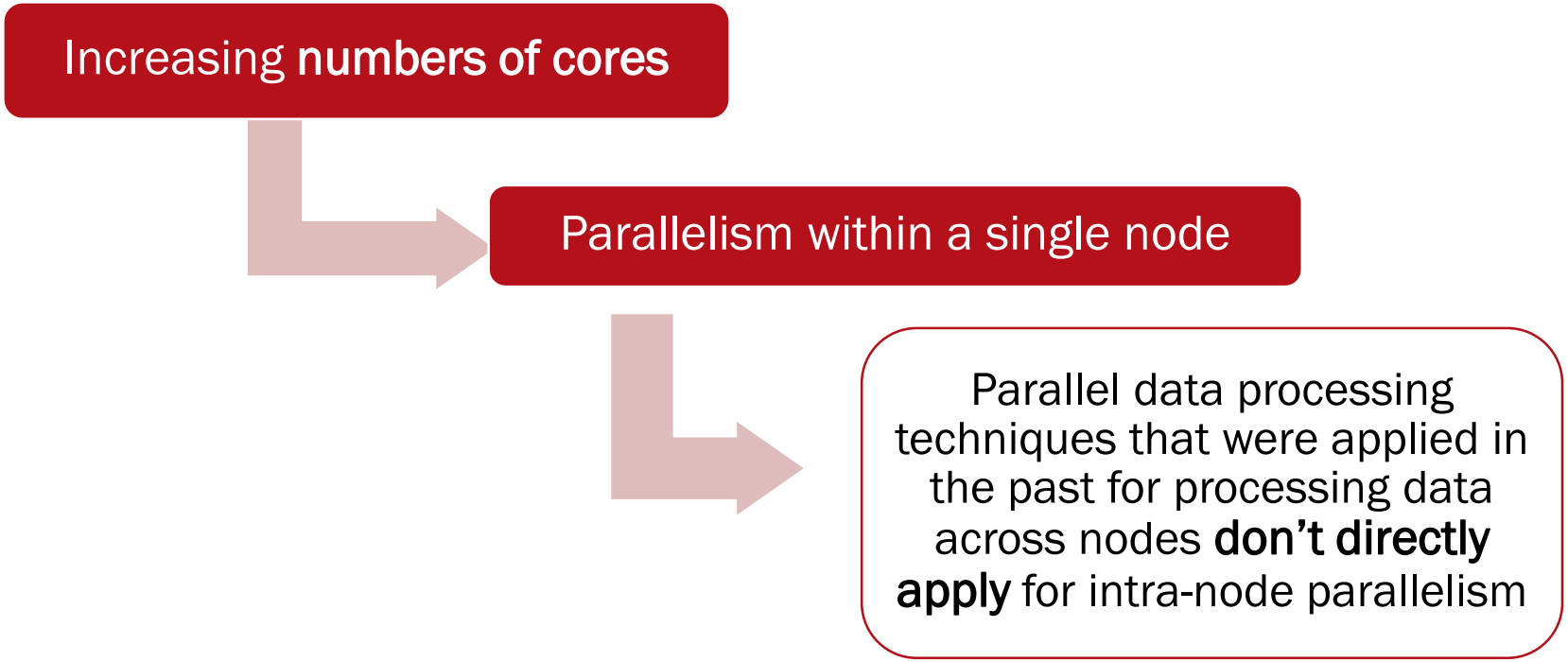


SFU

Scale

Data *volume* is **scaling faster** than *compute resources*.

Dramatic shift:



SFU

Challenges in the Big Data Analysis Pipeline

SFU

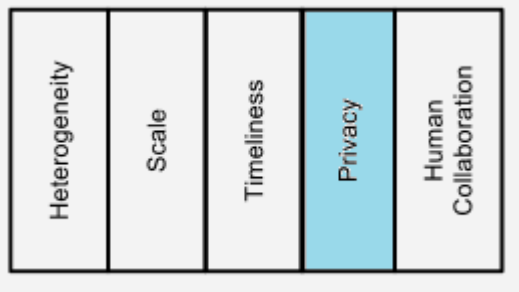
Timeliness

A **full analysis** of data is **not feasible** in real-time.

Source: http://monicaec.com/red-flags-warning-signs-of-potential-credit-card-fraud/

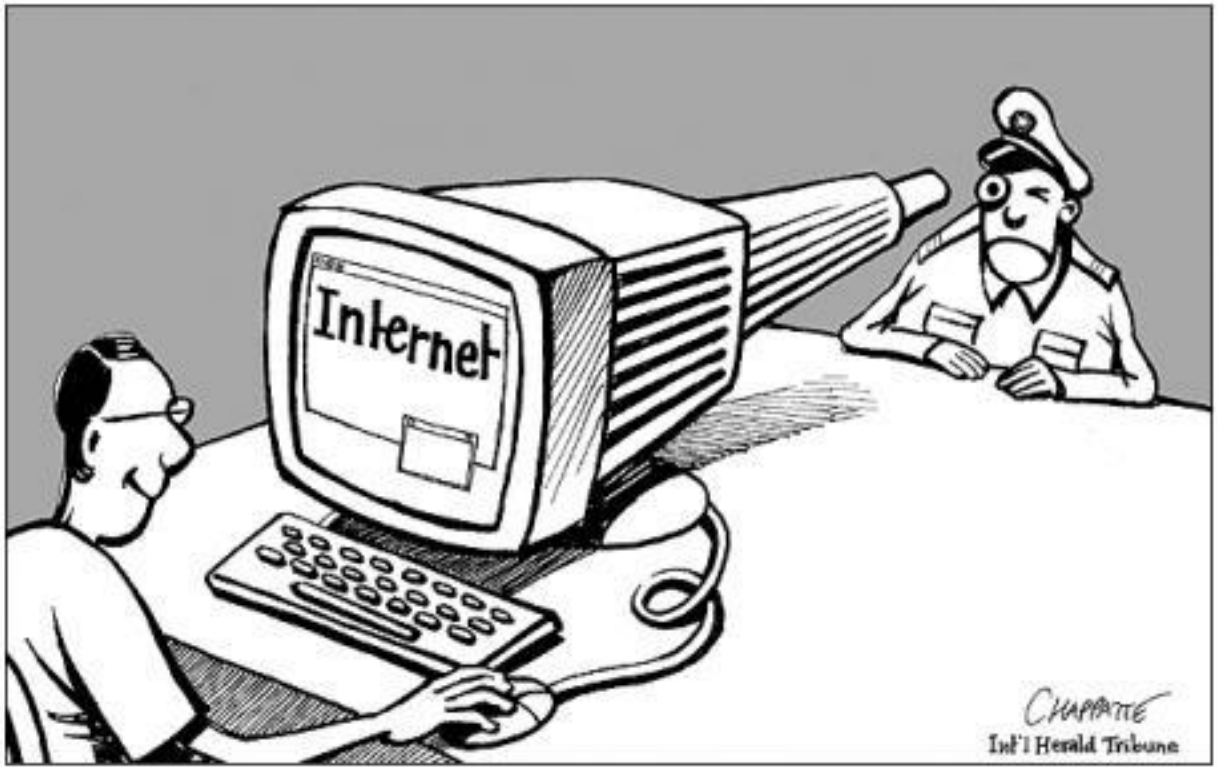
SFU

Challenges in the Big Data Analysis Pipeline



SFU

Privacy



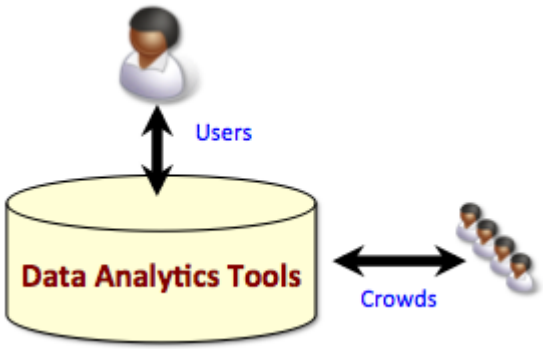
SFU

Challenges in the Big Data Analysis Pipeline

SFU

Human collaboration

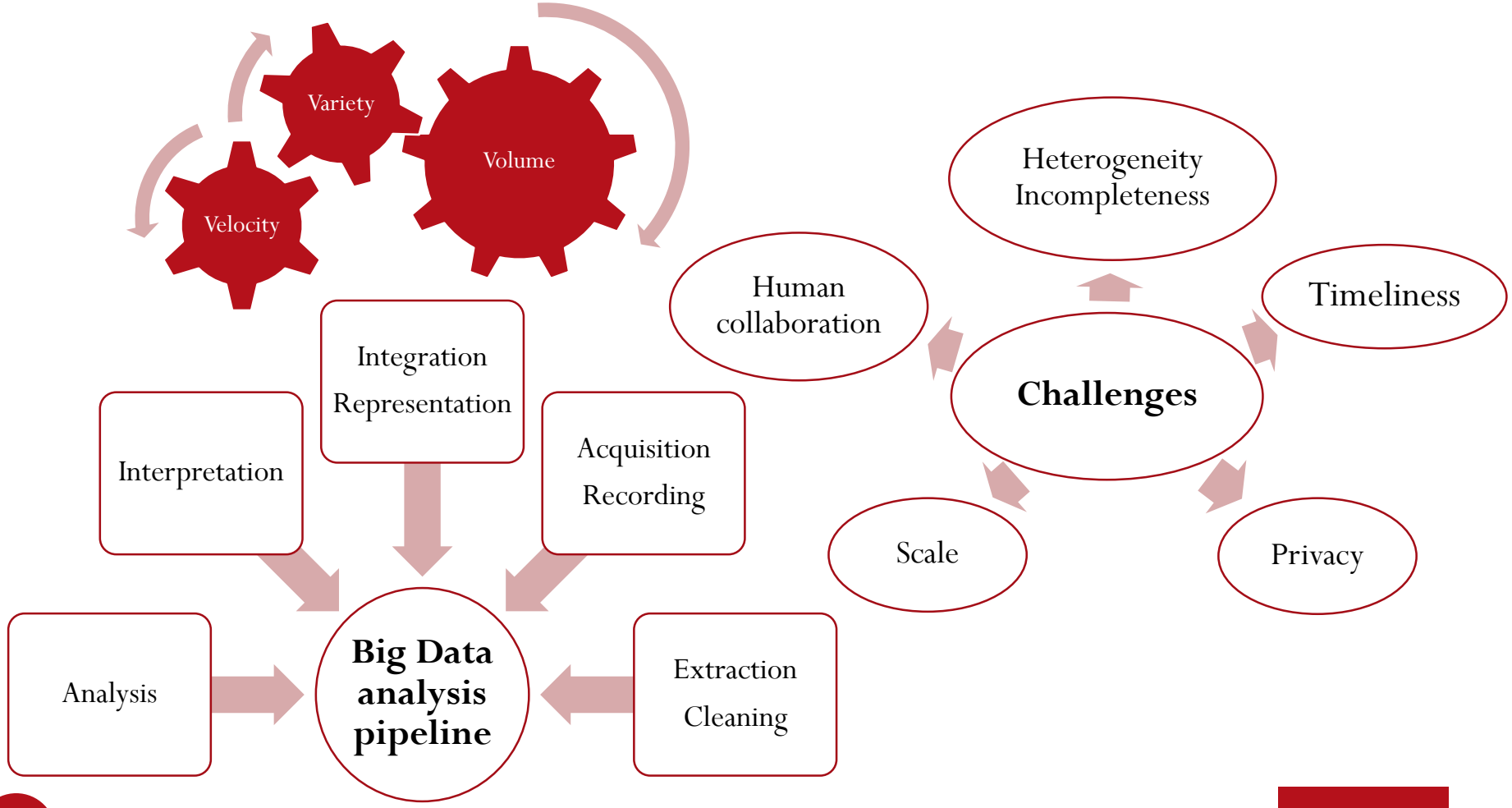
WIKIPEDIA
The Free Encyclopedia



SFU

Conclusion

Objective: many technical and ethnical **challenges** must be **addressed before** this potential can be realized fully.



SFU

THANK YOU FOR YOUR ATTENTION ANY QUESTIONS