

# Challenges and Opportunities with Big Data



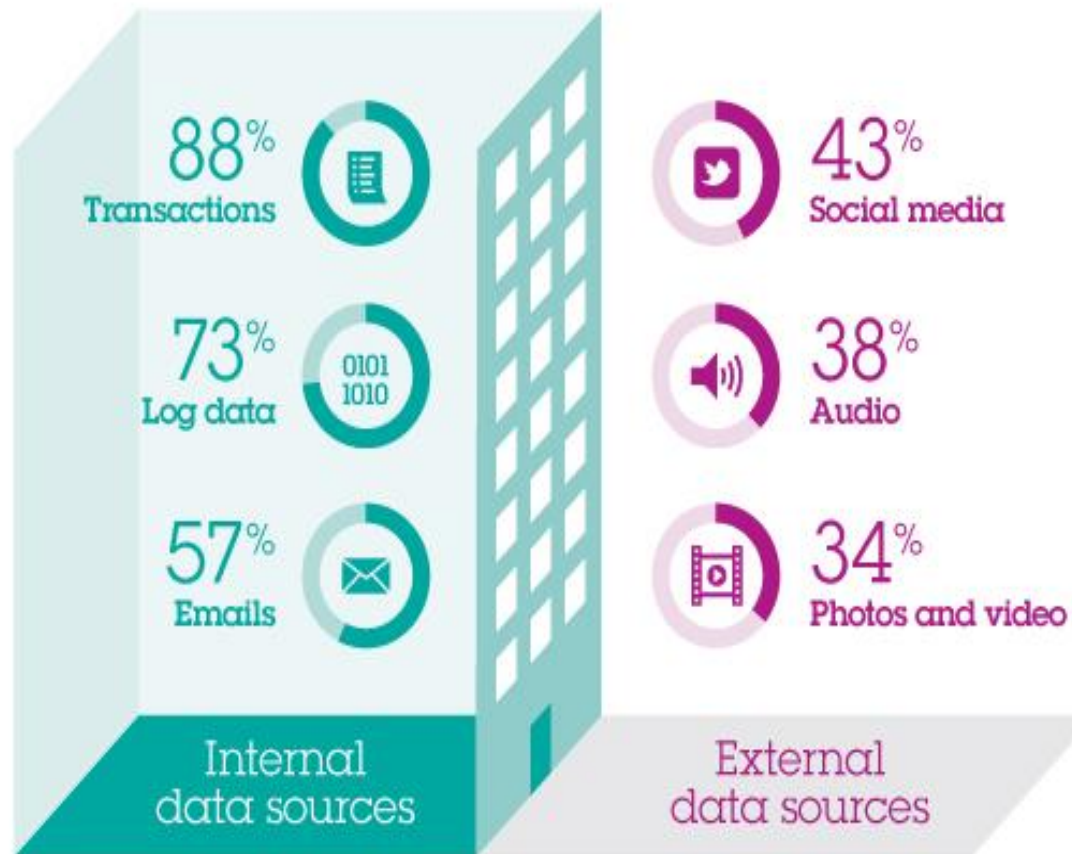
# Agenda

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



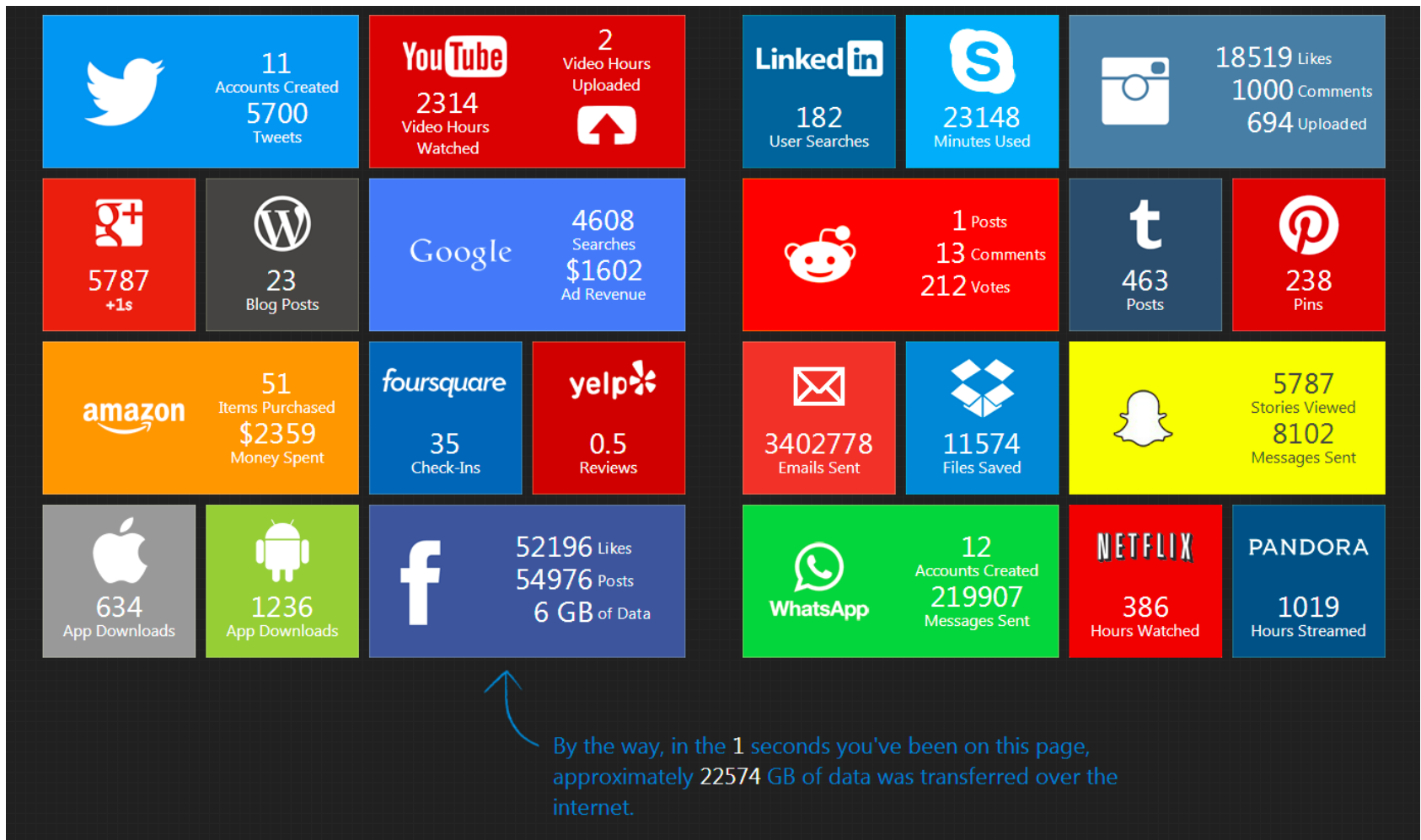
# Big Data

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

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



# Application Of Big Data

**Smarter  
Healthcare**



**Homeland  
Security**



**Traffic Control**



**Manufacturing**



**Multi-channel  
sales**



**Telecom**



**Trading  
Analytics**



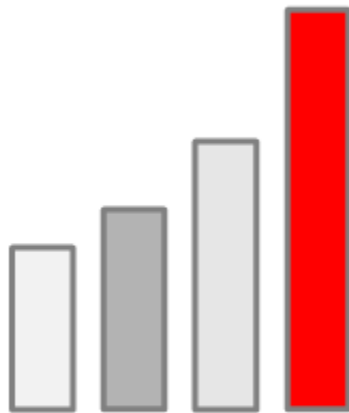
**Search  
Quality**



# Our future? ☺



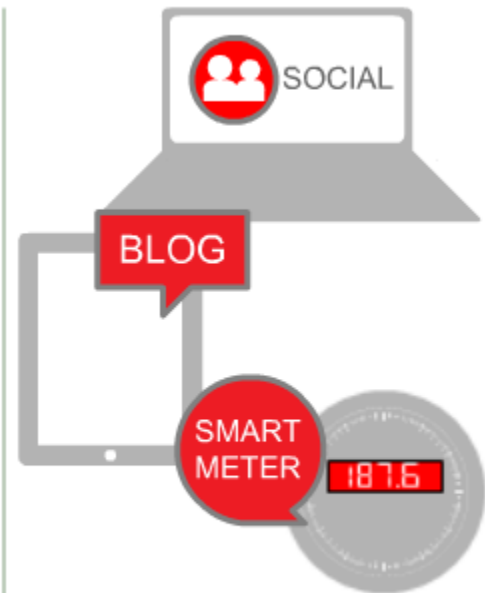
# Big data spans three dimensions:



VOLUME



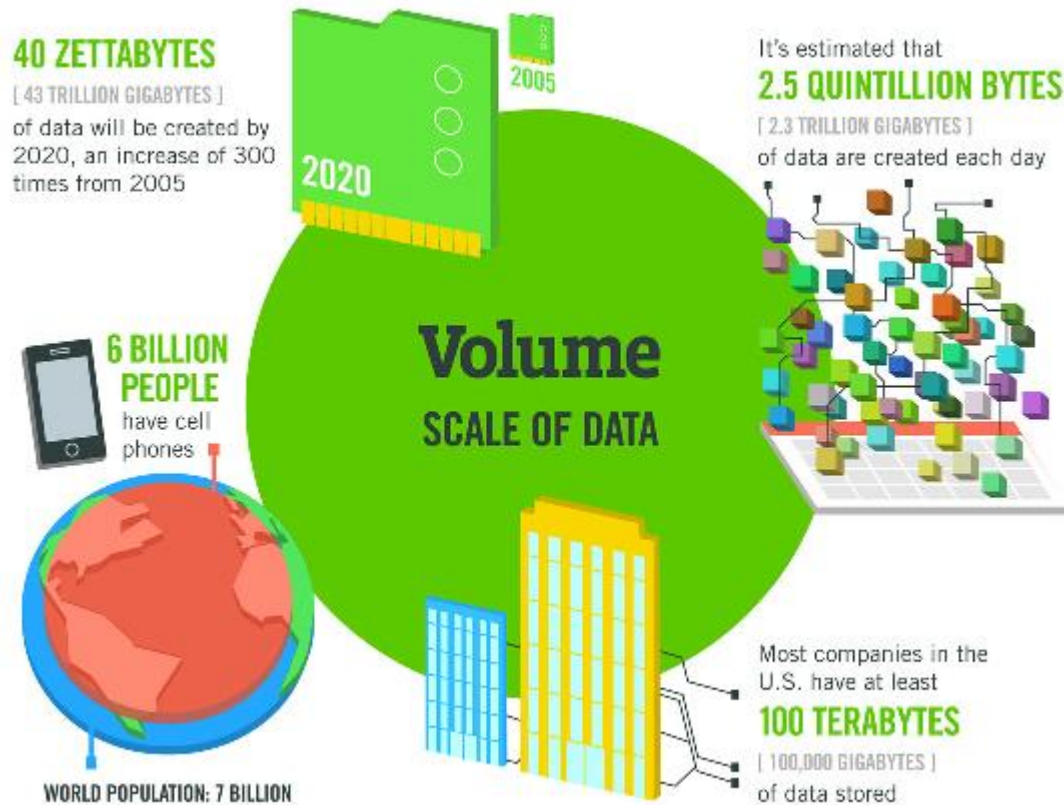
VELOCITY



VARIETY



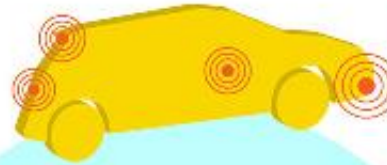
# Volume





# Velocity

The New York Stock Exchange captures  
**1 TB OF TRADE INFORMATION**  
during each trading session



Modern cars have close to  
**100 SENSORS**  
that monitor items such as  
fuel level and tire pressure

## Velocity ANALYSIS OF STREAMING DATA

By 2016, it is projected  
there will be  
**18.9 BILLION  
NETWORK  
CONNECTIONS**  
– almost 2.5 connections  
per person on earth



# Variety

As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]



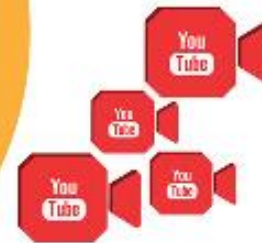
By 2016, it's anticipated there will be

**420 MILLION  
WEARABLE, WIRELESS  
HEALTH MONITORS**



**4 BILLION+  
HOURS OF VIDEO**

are watched on  
YouTube each month



**30 BILLION  
PIECES OF CONTENT**

are shared on Facebook  
every month



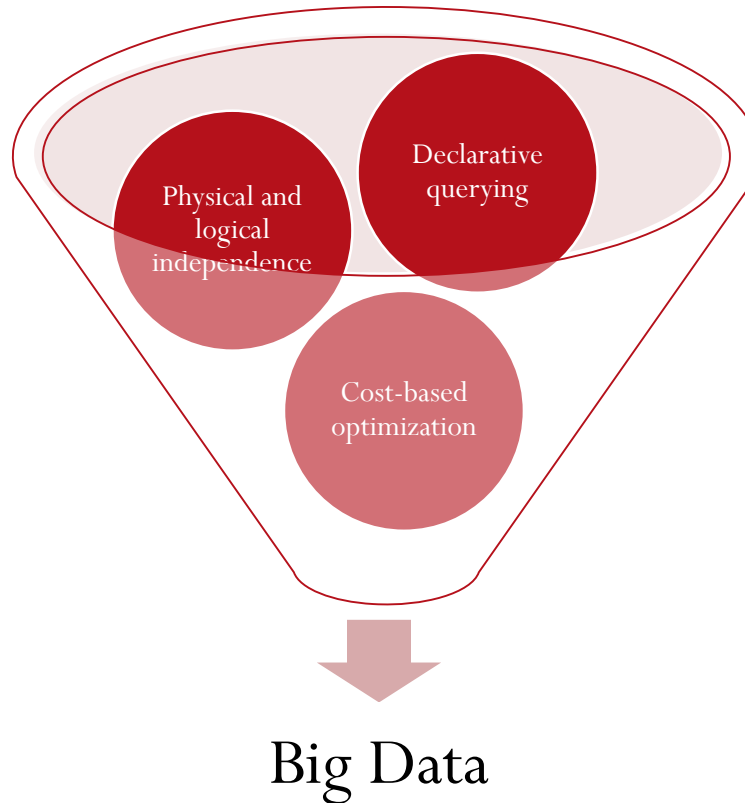
**Variety**  
DIFFERENT  
FORMS OF DATA

**400 MILLION TWEETS**

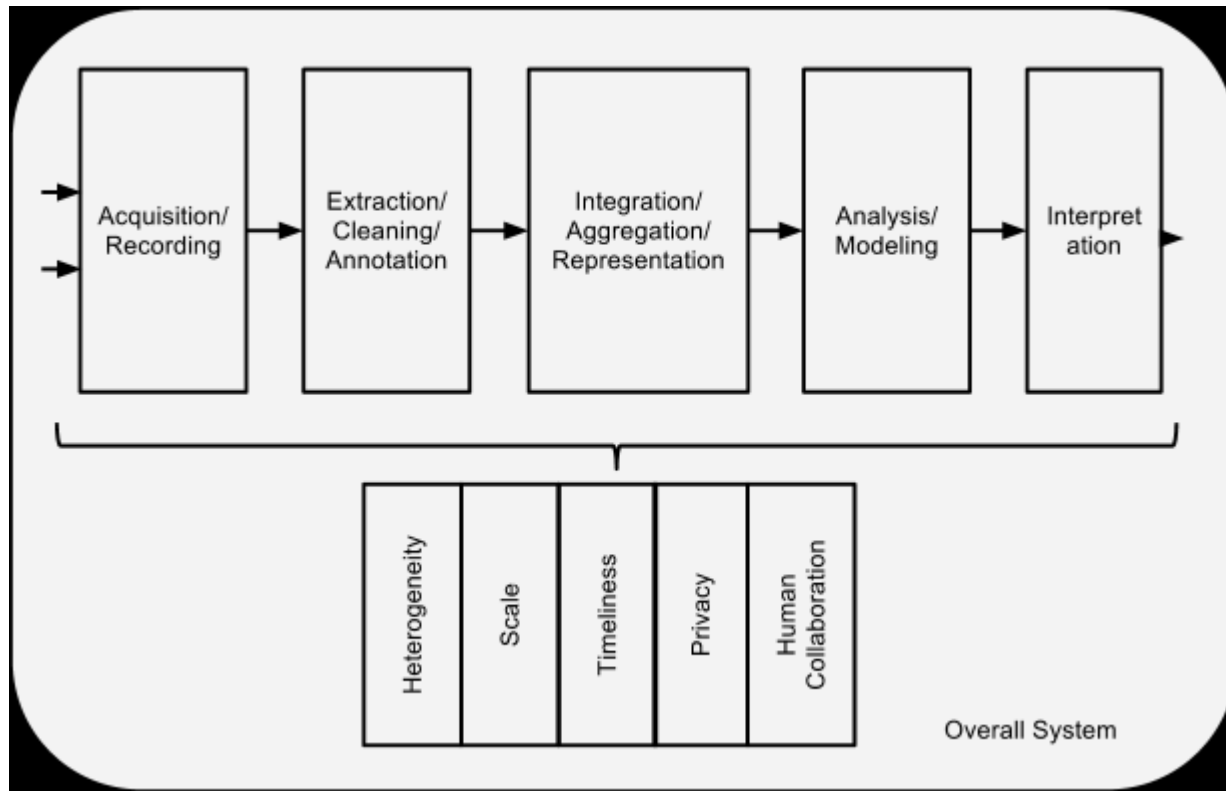
are sent per day by about 200  
million monthly active users



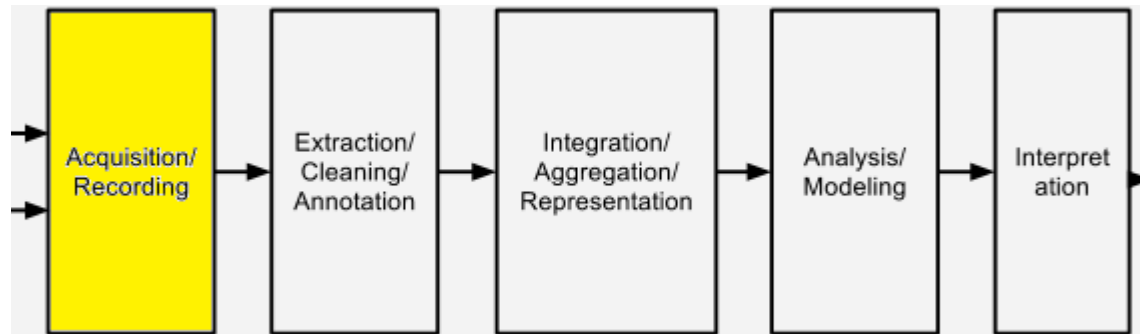
# What has been Achieved...



# Phases and challenges in the Big Data Analysis Pipeline



# The Big Data Analysis Pipeline



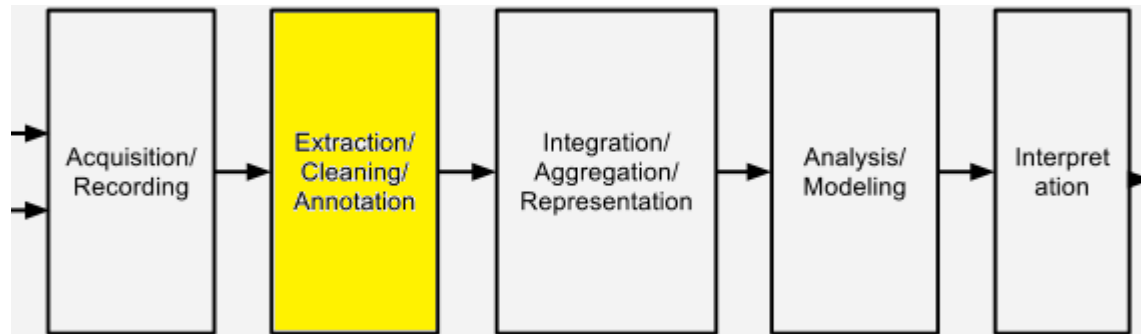
# Data Acquisition and Recording

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



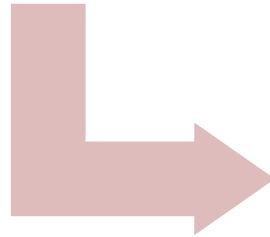


# The Big Data Analysis Pipeline

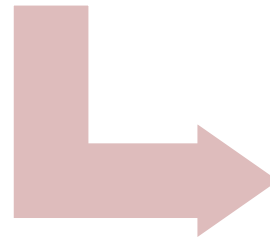


# Information Extraction and Cleaning

Raw unformatted data

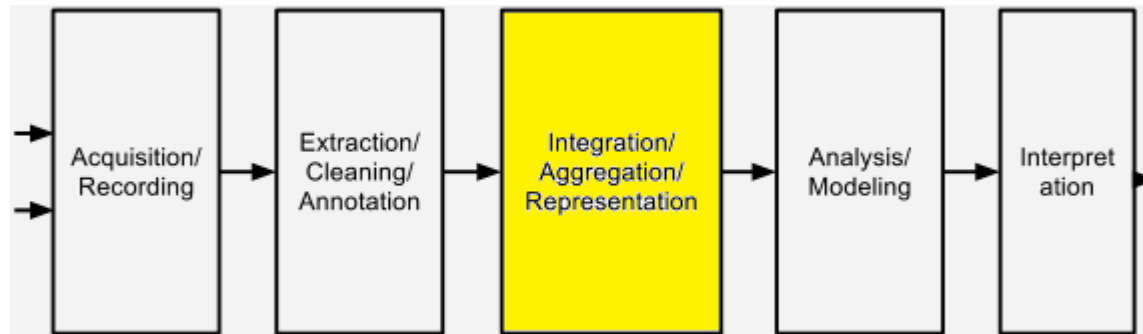


Structured and suitable data for analysis



**Technical  
challenge**

# The Big Data Analysis Pipeline

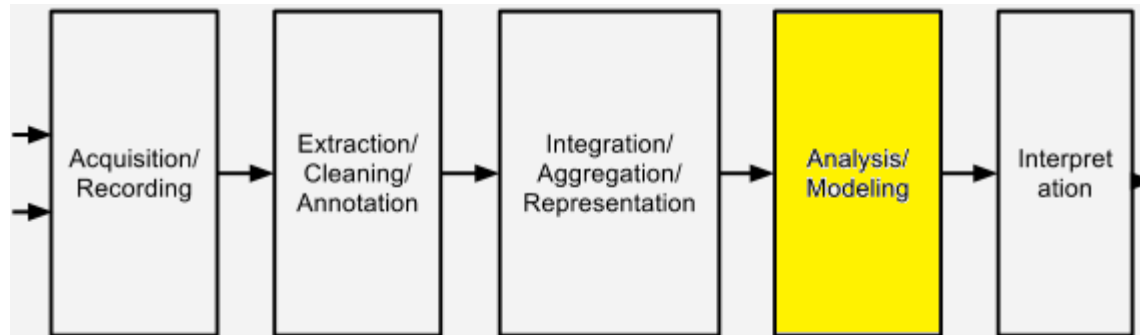


# Data Integration, Aggregation, and Representation

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



# The Big Data Analysis Pipeline

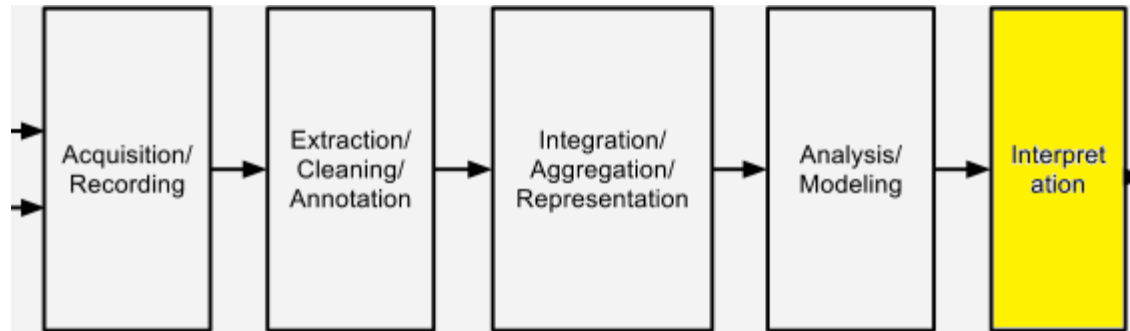


# Query Processing, Data Modeling, and Analysis

- 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).

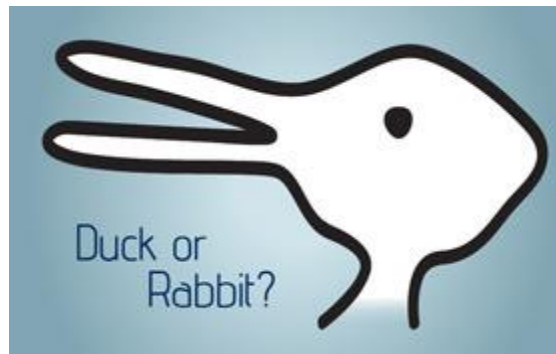


# The Big Data Analysis Pipeline

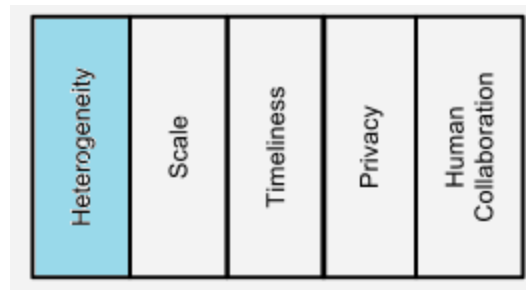


# Interpretation

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

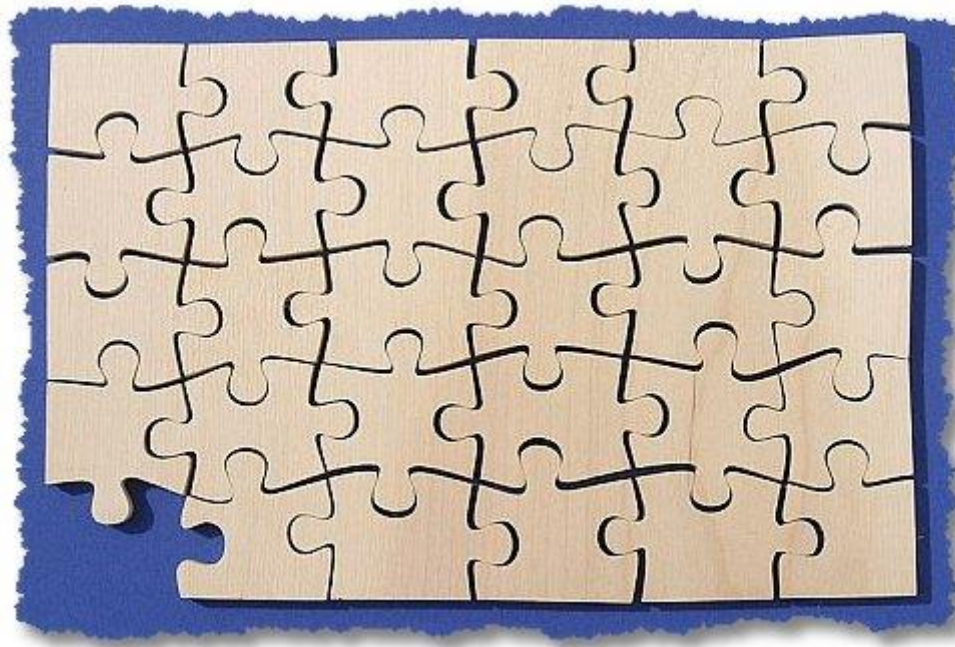


# Challenges in the Big Data Analysis Pipeline



# 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



# Scale

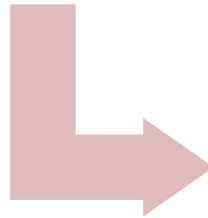
Data volume is **scaling faster** than *compute resources*.

**Dramatic shift:**

Increasing numbers of cores



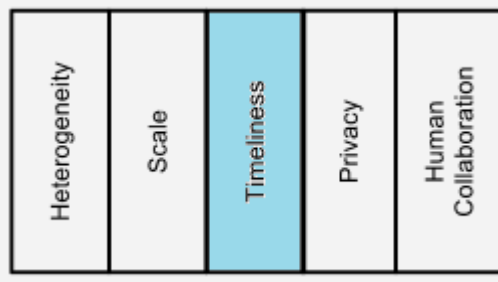
Parallelism within a single node



Parallel data processing techniques that were applied in the past for processing data across nodes **don't directly apply** for intra-node parallelism



# Challenges in the Big Data Analysis Pipeline



# Timeliness

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



The average merchant experiences **156 successful fraudulent transactions** per month.



The value of an average **fraudulent transaction** is \$114.



**55% of fraud** is related to ecommerce, as reported by multi-channel merchants.



**1.32% of revenue** is lost to fraud, a **94% increase** from 2014.



**29% of merchants** feel it is too expensive to control fraud.

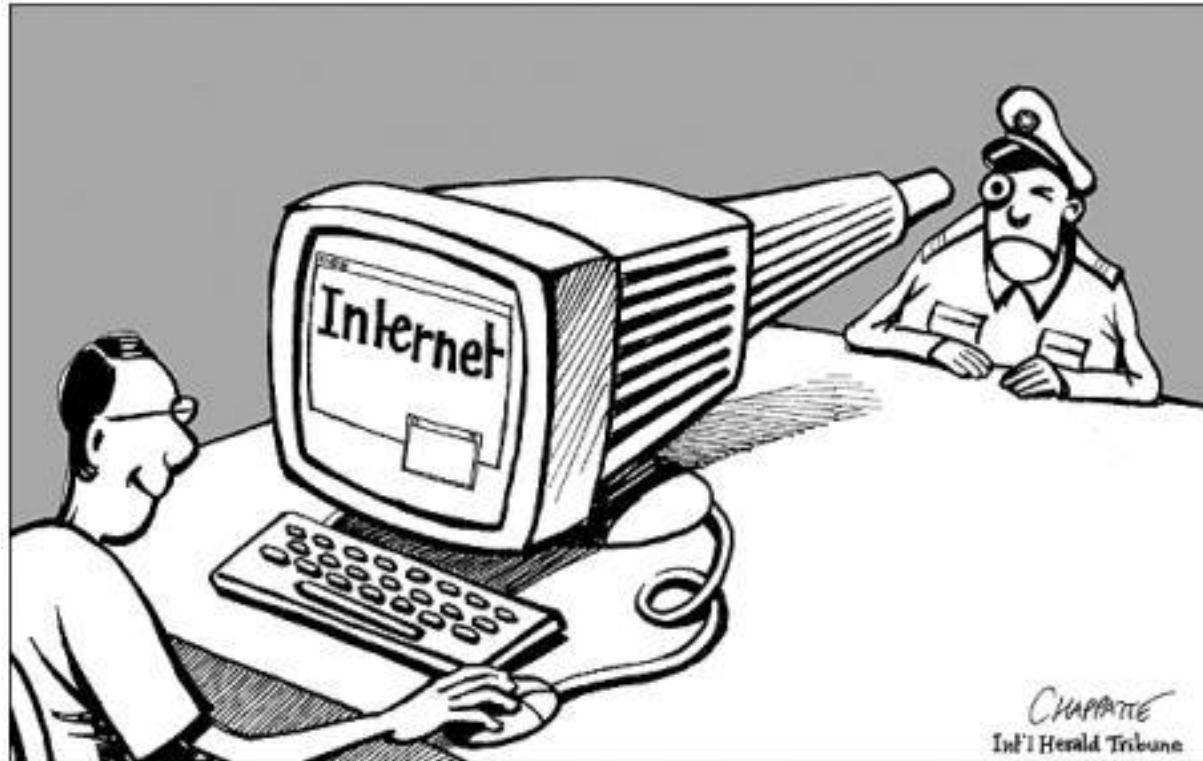


**25% of declined** potentially fraudulent transactions are false positives.

# Challenges in the Big Data Analysis Pipeline



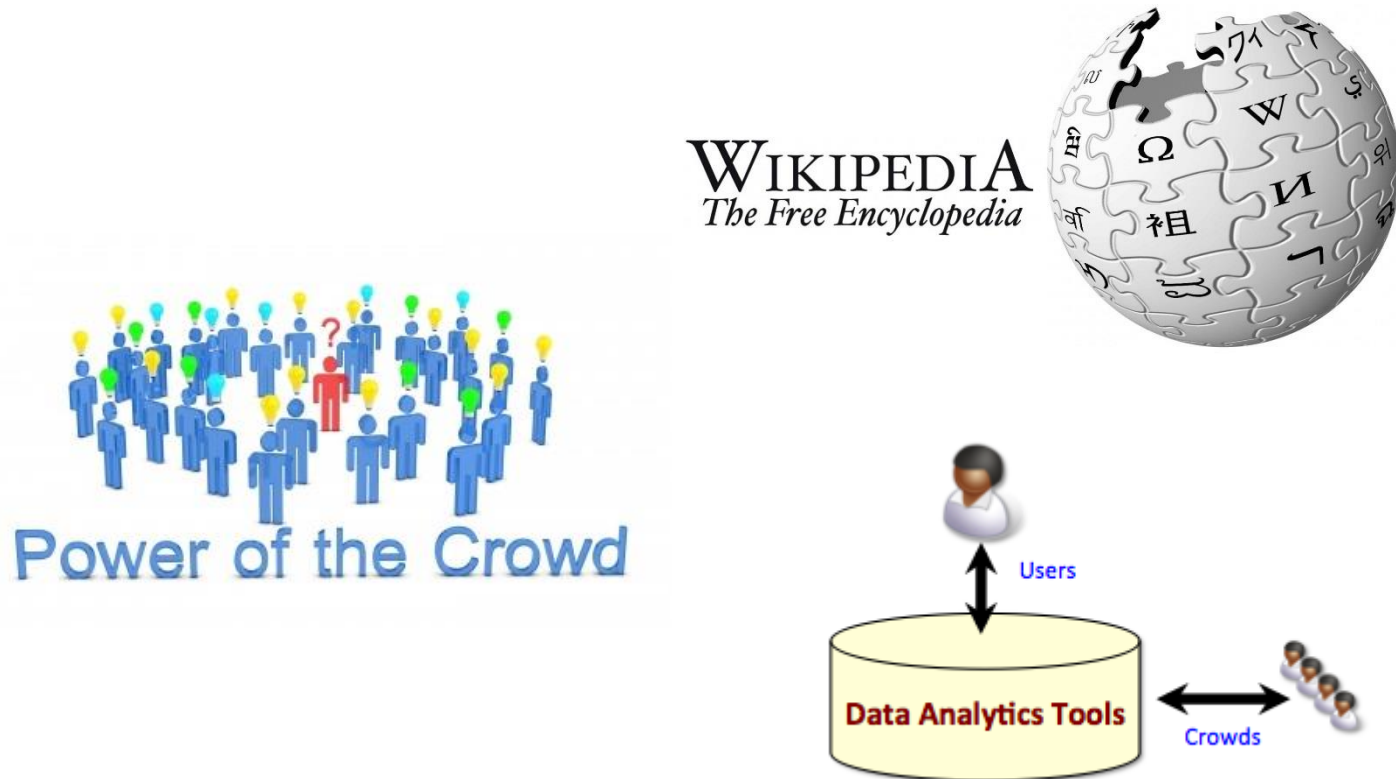
# Privacy



# Challenges in the Big Data Analysis Pipeline



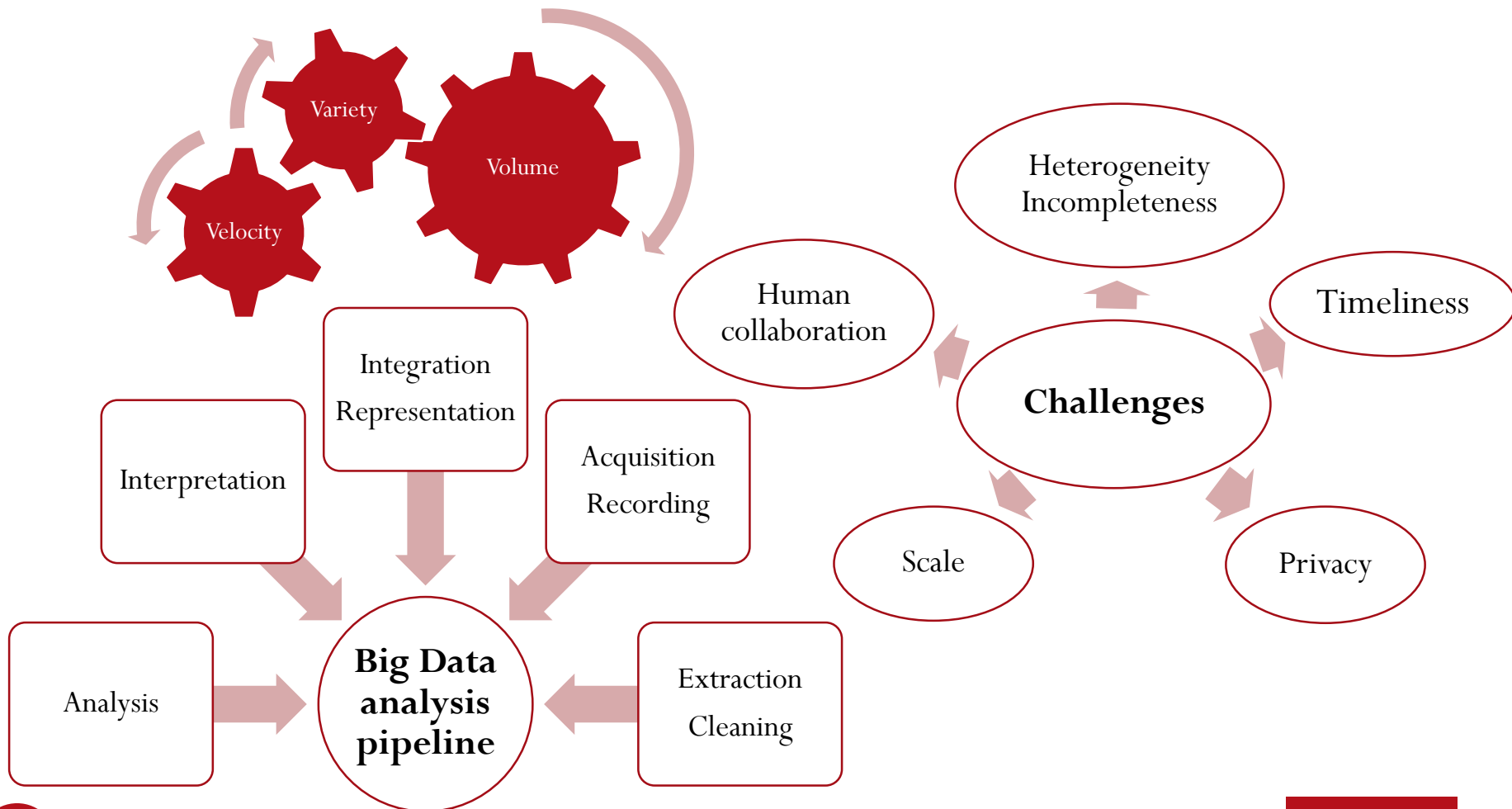
# Human collaboration





# Conclusion

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



**THANK YOU  
FOR  
YOUR ATTENTION  
ANY  
QUESTIONS**