Semantic Data Management

ANNA QUERALT, OSCAR ROMERO

(FACULTAT D'INFORMÀTICA DE BARCELONA)

Introduction and Motivation

VARIETY IN COMPLEX DATA ECOSYSTEMS

“WITHOUT DATA,
YOU’RE JUST
ANOTHER PERSON
WITH AN OPINION”

W. Edwards Deming, American Statistician

New Business Model: Instagram's Fable

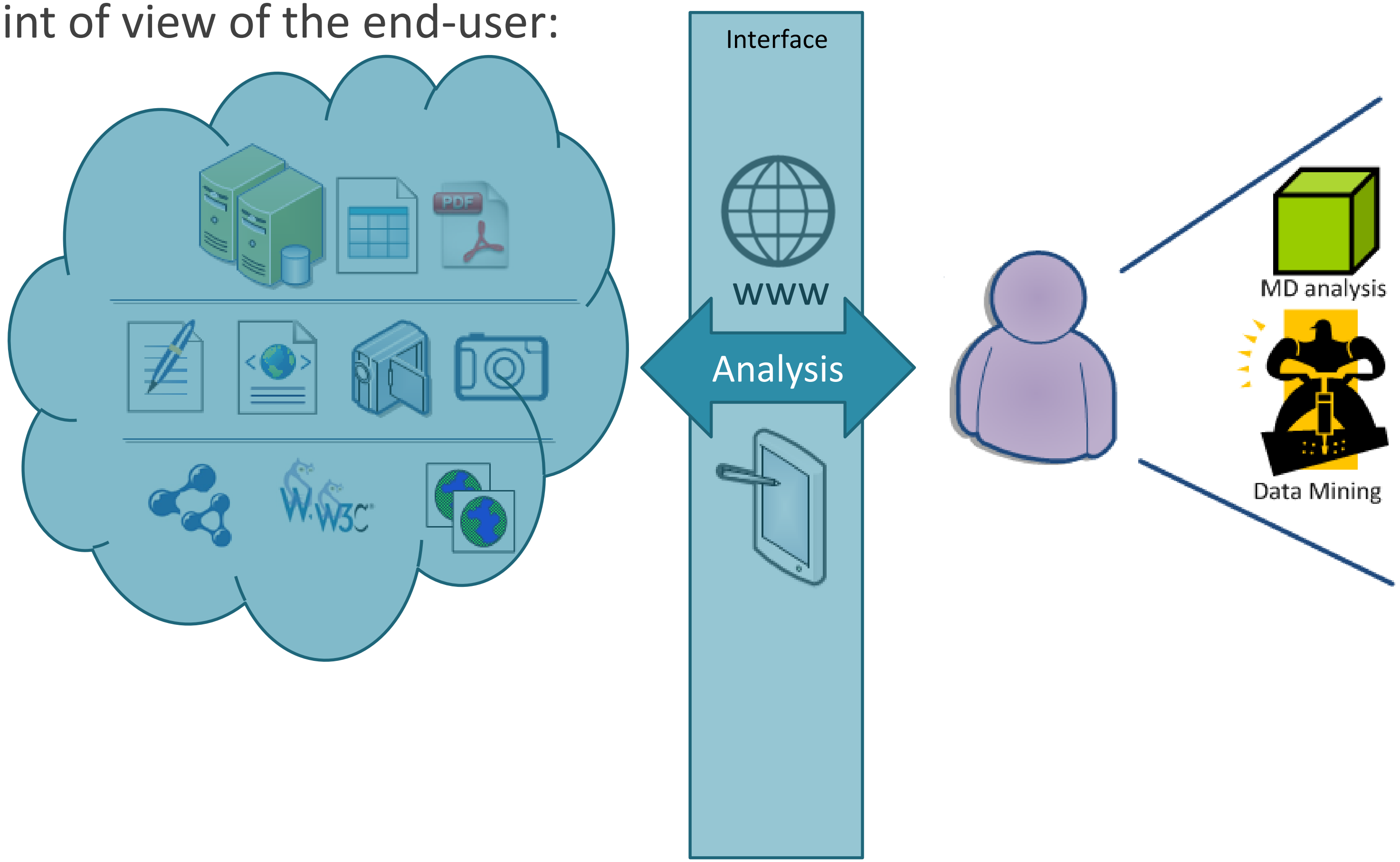
(xkcd.com)

Challenges of the Data-Driven Economy

FROM THE IT POINT OF VIEW

Data Analysis Democratisation

From the point of view of the end-user:



What is Big Data?

VOLUME

Veracity

Velocity

Value

vArIaBiLiTy

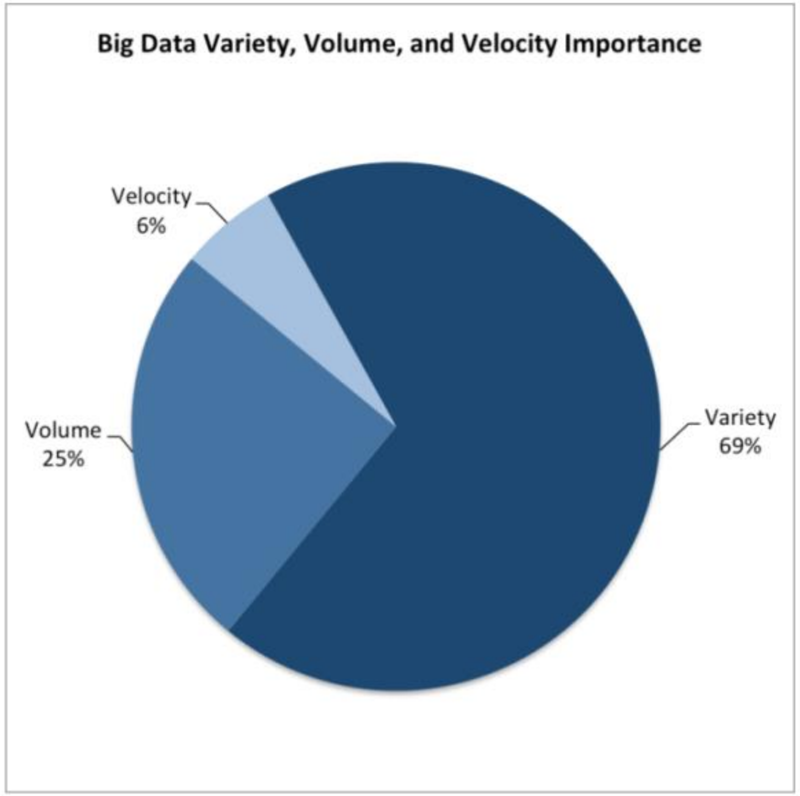
Variety

Today, the Focus is on Variety

That Big Data is synonymous with large volumes of data is a **myth**

"Rather, it is the ability to ***integrate*** *more sources of data than ever before — new data, old data, big data, small data, structured data, unstructured data, social media data, behavioral data, and legacy data"*

The Variety Challenge



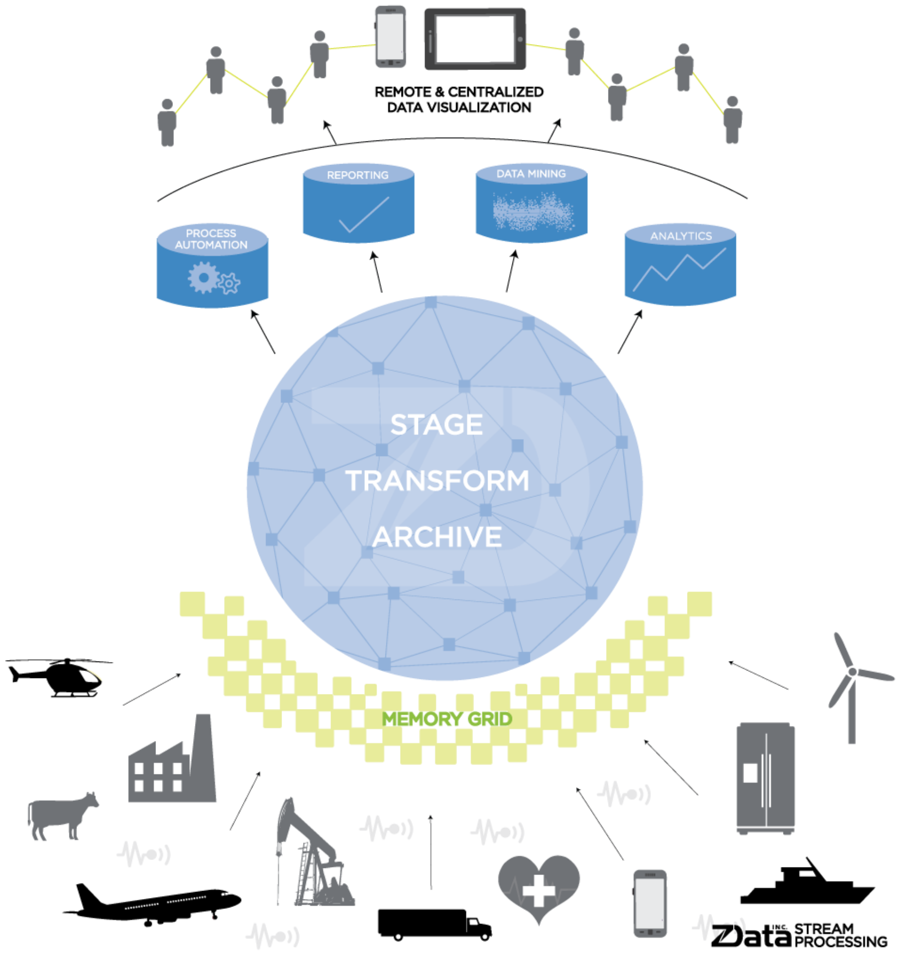
MIT Sloan Management Review (2016): http://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/

The Data Lake

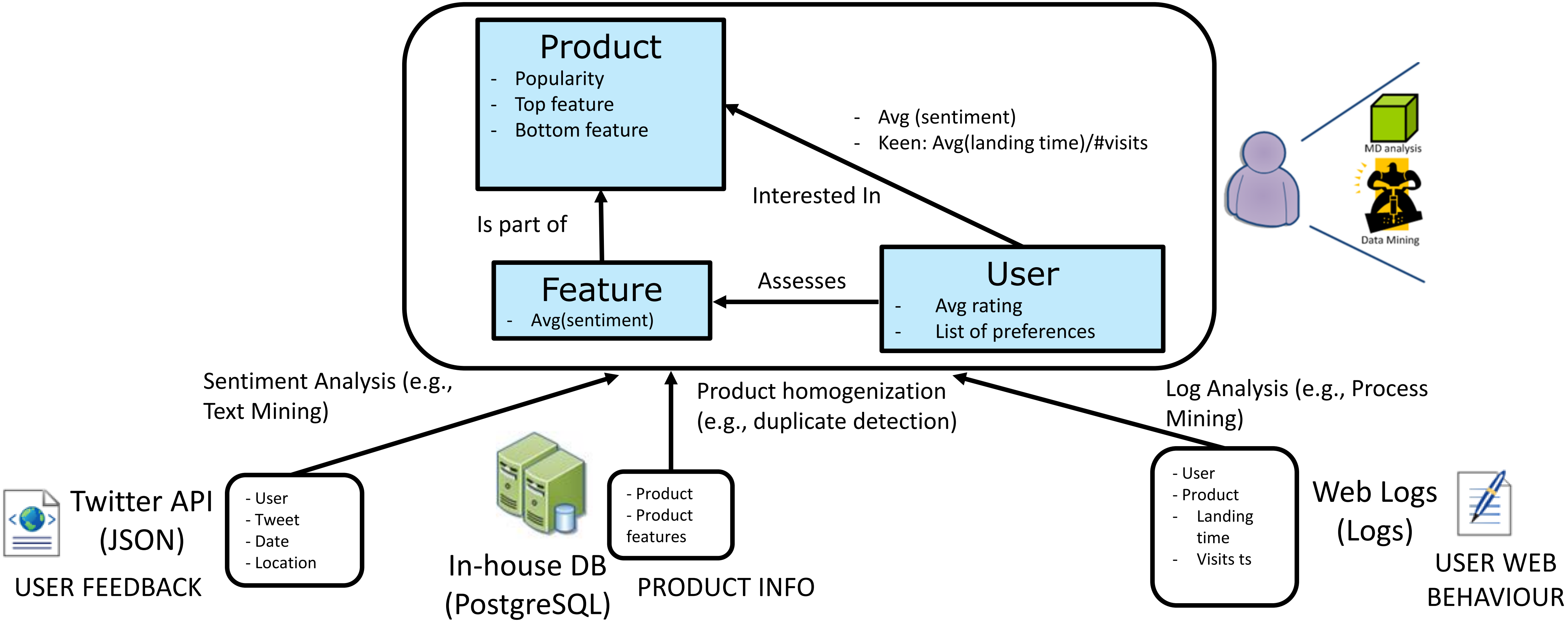
IDEA: Load-first, Model-Later

Modeling at load time restricts the potential analysis that can be done later (Big Analytics)

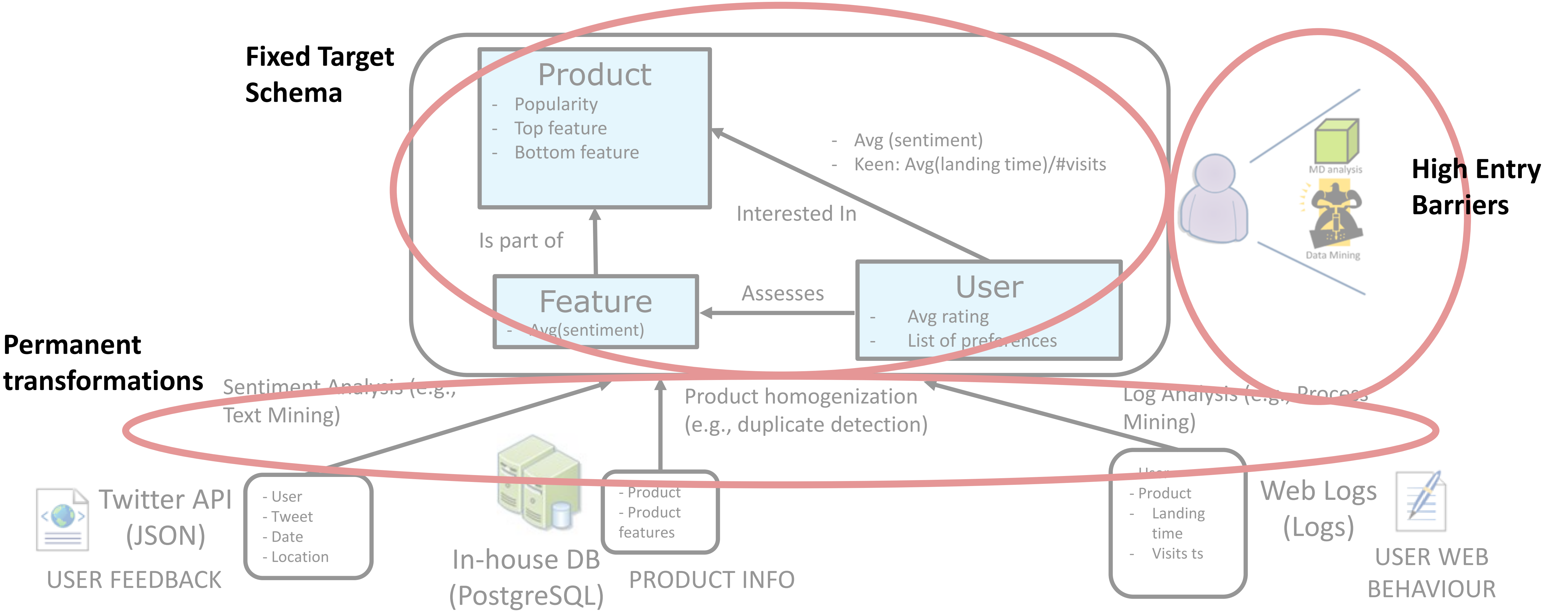
Store raw data and create on-demand views to handle with precise analysis needs



Model-First (Load-Later)

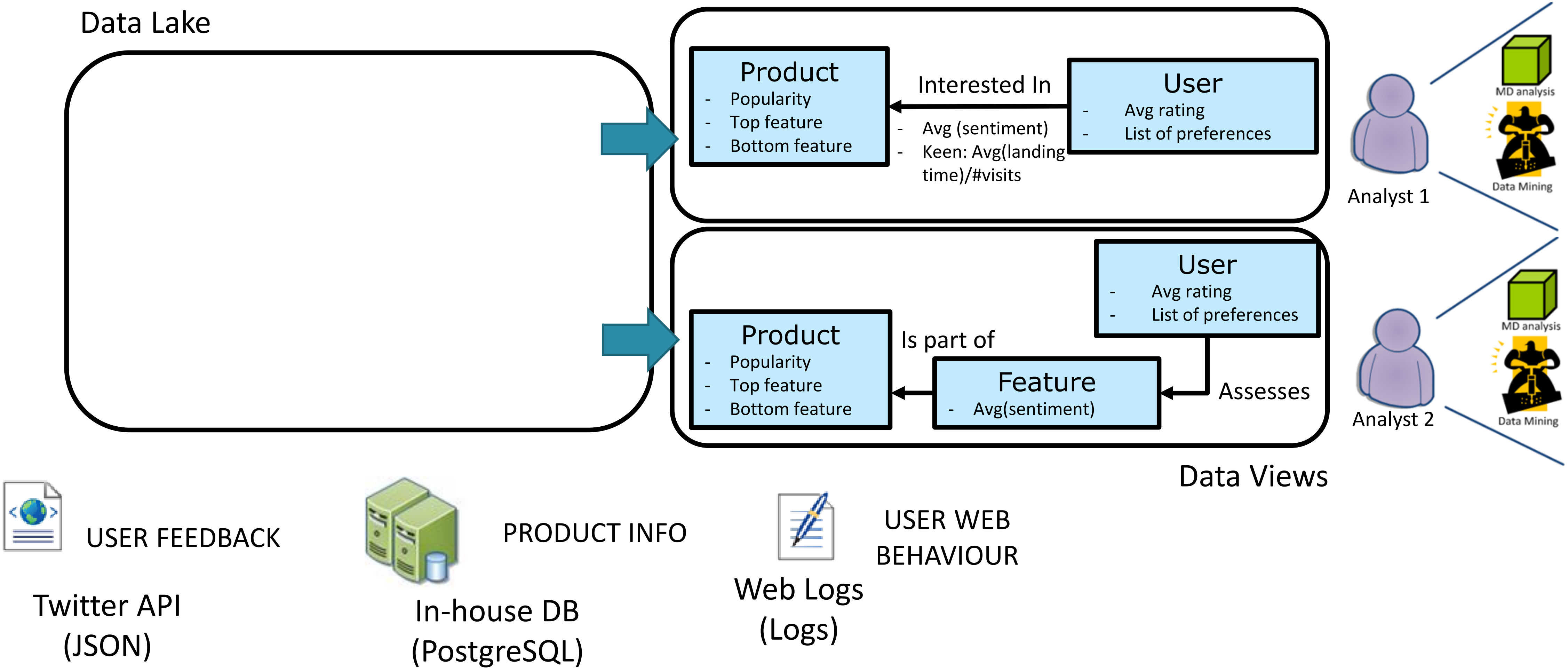


Drawbacks

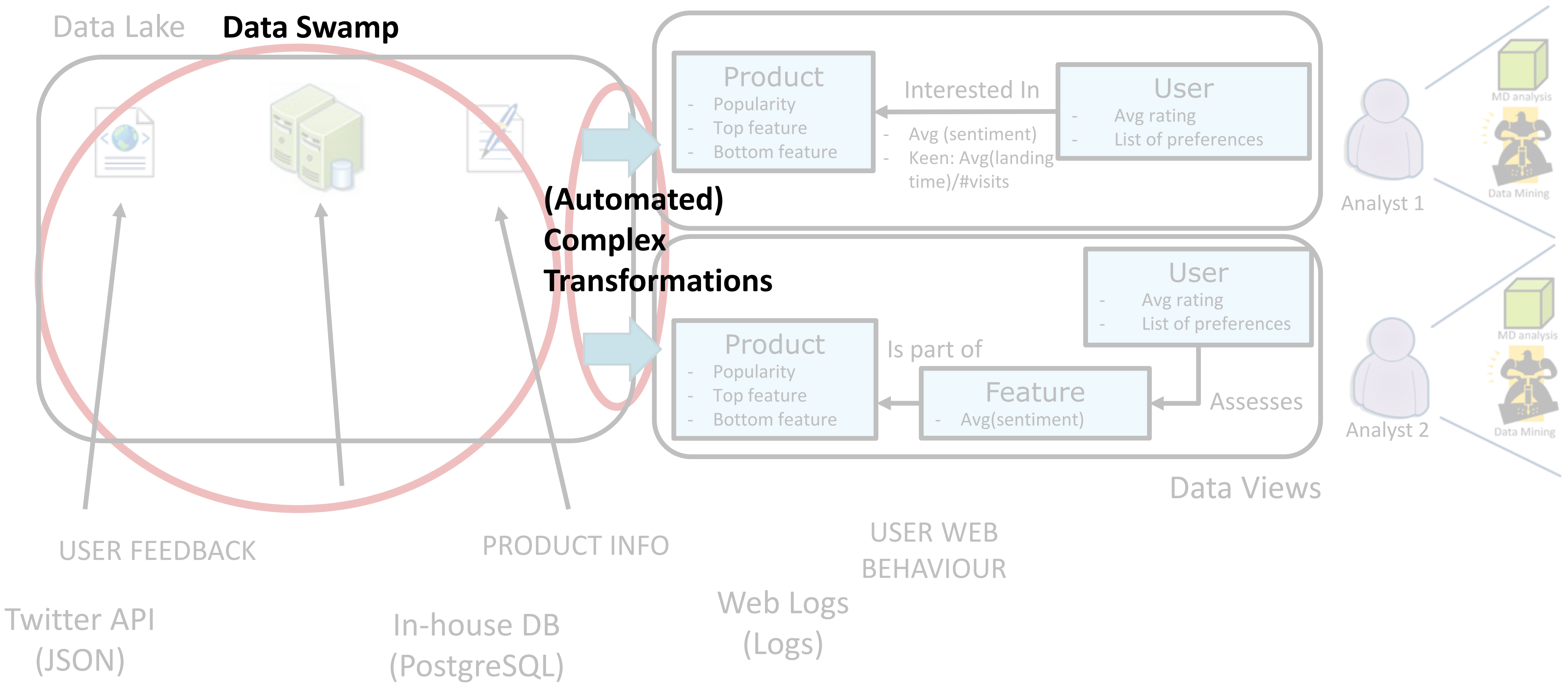


Fixed Target Schema
Product
- Popularity
- Top feature
- Bottom feature
- Avg (sentiment)
- Keen: Avg(landing time)/#visits
Interested In
Is part of
Feature
- Avg(sentiment)
Assesses
User
- Avg rating
- List of preferences
MD analysis
Data Mining
High Entry Barriers
Permanent transformations
Sentiment Analysis (e.g., Text Mining)
Product homogenization (e.g., duplicate detection)
Log Analysis (e.g., Process Mining)
Twitter API (JSON)
- User
- Tweet
- Date
- Location
USER FEEDBACK
In-house DB (PostgreSQL)
- Product
- Product features
PRODUCT INFO
- Product
- Landing time
- Visits ts
Web Logs (Logs)
USER WEB BEHAVIOUR

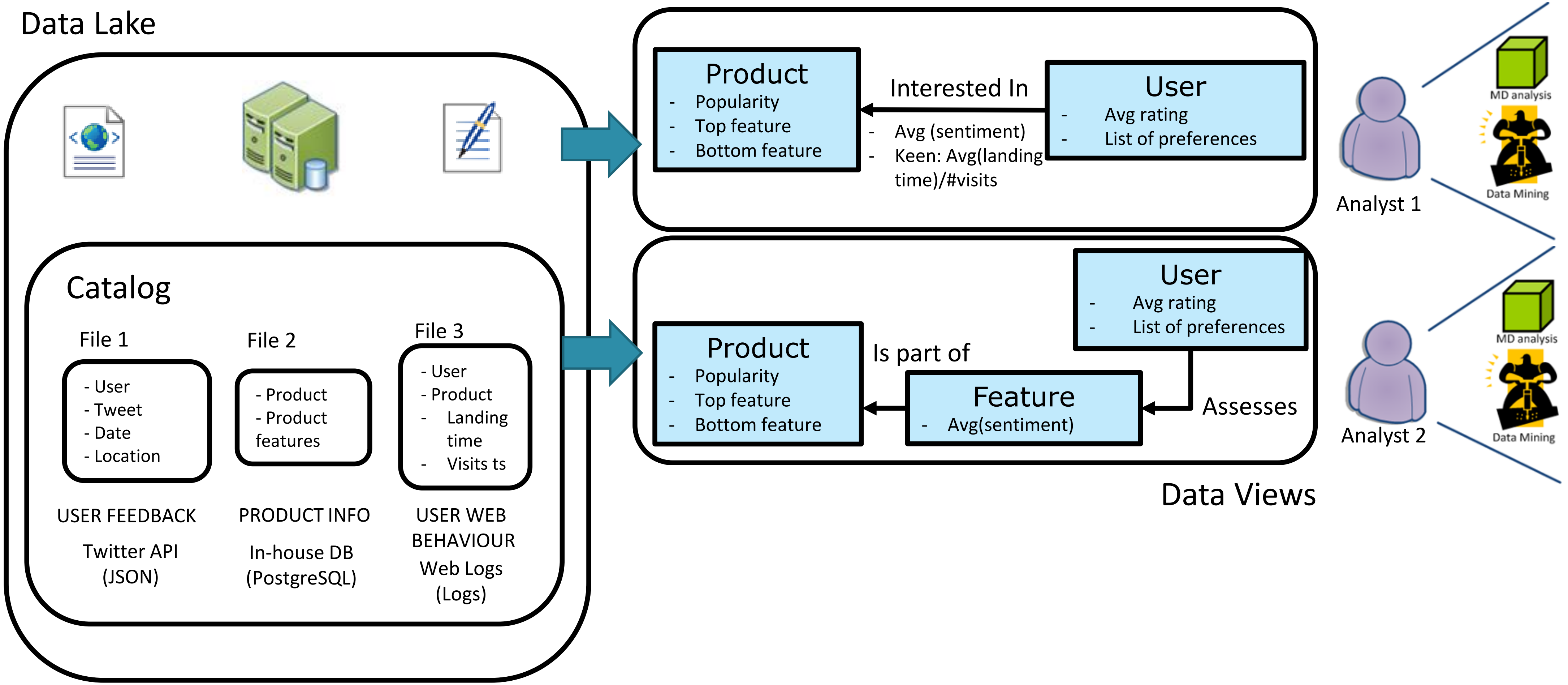
Load-First Model-Later



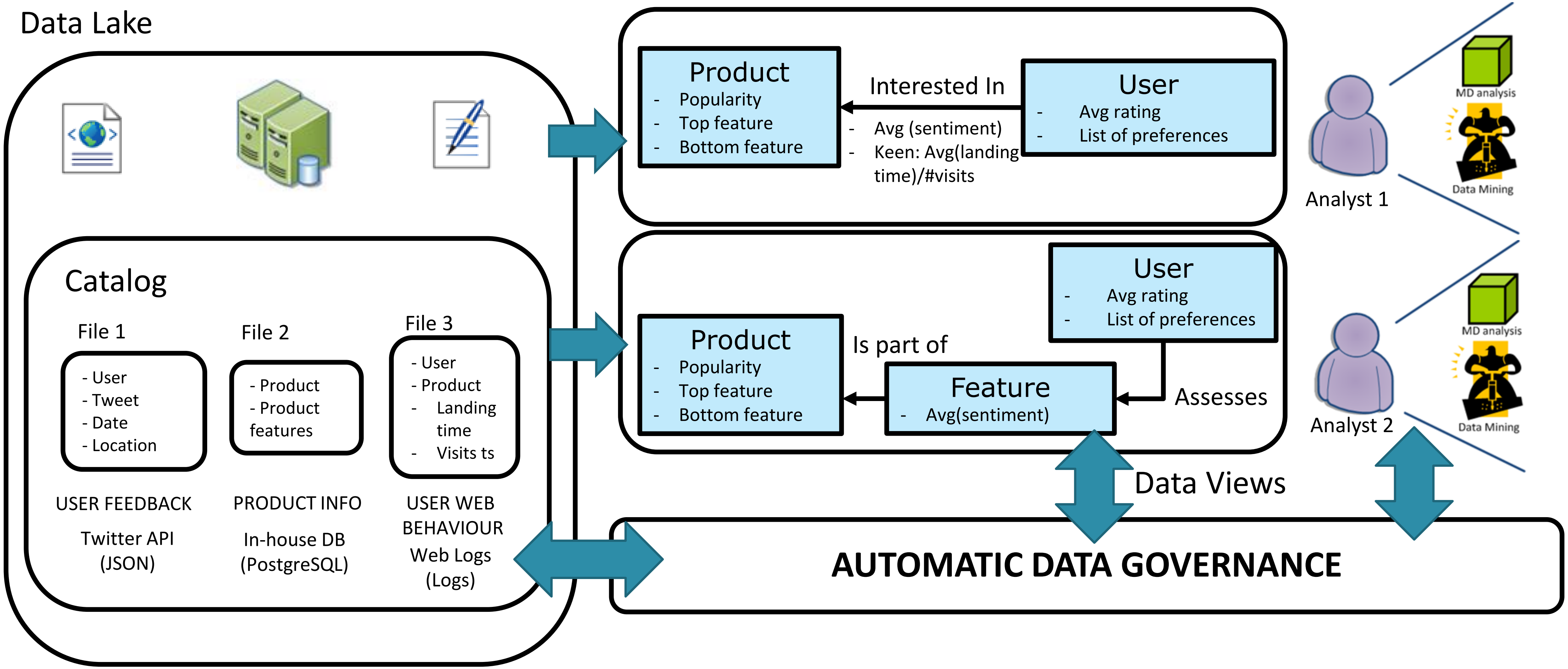
Drawbacks


Data Lake
Data Swamp
(Automated) Complex Transformations
Product
- Popularity
- Top feature
- Bottom feature
Interested In
- Avg (sentiment)
- Keen: Avg(landing time)/#visits
User
- Avg rating
- List of preferences
Analyst 1
MD analysis
Data Mining
User
- Avg rating
- List of preferences
Product
- Popularity
- Top feature
- Bottom feature
Is part of
Feature
- Avg(sentiment)
Assesses
Analyst 2
MD analysis
Data Mining
Data Views
USER FEEDBACK
PRODUCT INFO
USER WEB BEHAVIOUR
Twitter API (JSON)
In-house DB (PostgreSQL)
Web Logs (Logs)

From Data Swarms to Semantic Data Lakes



From IT-Centered to User-Centered



Data Variety:
Graphs to the Rescue

Graph Data Model in a Nutshell

Occurrence-oriented

- It is a schemaless data model
  - There is no explicit schema
  - Data (and its relationships) may quickly vary
- Objects and relationships as first-class citizens
  - *An object o relates (through a relationship r) to another object o'*
    - *Such relationship is often known as a triple (o r o')*
  - Both objects and relationships may contain properties
- Built on top of the graph theory
  - Euler (18th century)
  - More natural and intuitive than the relational model to deal with relationships

Notation (I)

A **graph** *G* is a set of nodes and edges: *G (N, E)*

N - **Nodes** (or vertices): *n1, n2, … Nm*

E - **Edges** are represented as pairs of nodes: (*n1, n2*)

- An edge is said to be **incident** to *n1* and *n2* (also, *n1* and *n2* are said to be **adjacent)**
- An edge is drawn as a line between *n1 and n2*
- **Directed edges** entail direction: <u>**from**</u> *n1* <u>**to**</u> *n2*
- An edge is said to be **multiple** if there is another edge exactly relating the same nodes
- An **hyperedge** is an edge inciding in more than 2 nodes

Types of graphs:

- **Multigraph**: If it contains at least one multiple Edge
- **Simple graph**: If it does not contain multiple edges
- **Hypergraph**: A graph allowing hyperedges

Notation (II)

Size (of a graph): #edges

Degree (of a node): #(incident edges)

- The degree of a node denotes the node adjacency
- The neighbourhood of a node are all its adjacent nodes

Out-degree (of a node): #(edges leaving the node)

- Sink node: A node with 0 out-degree

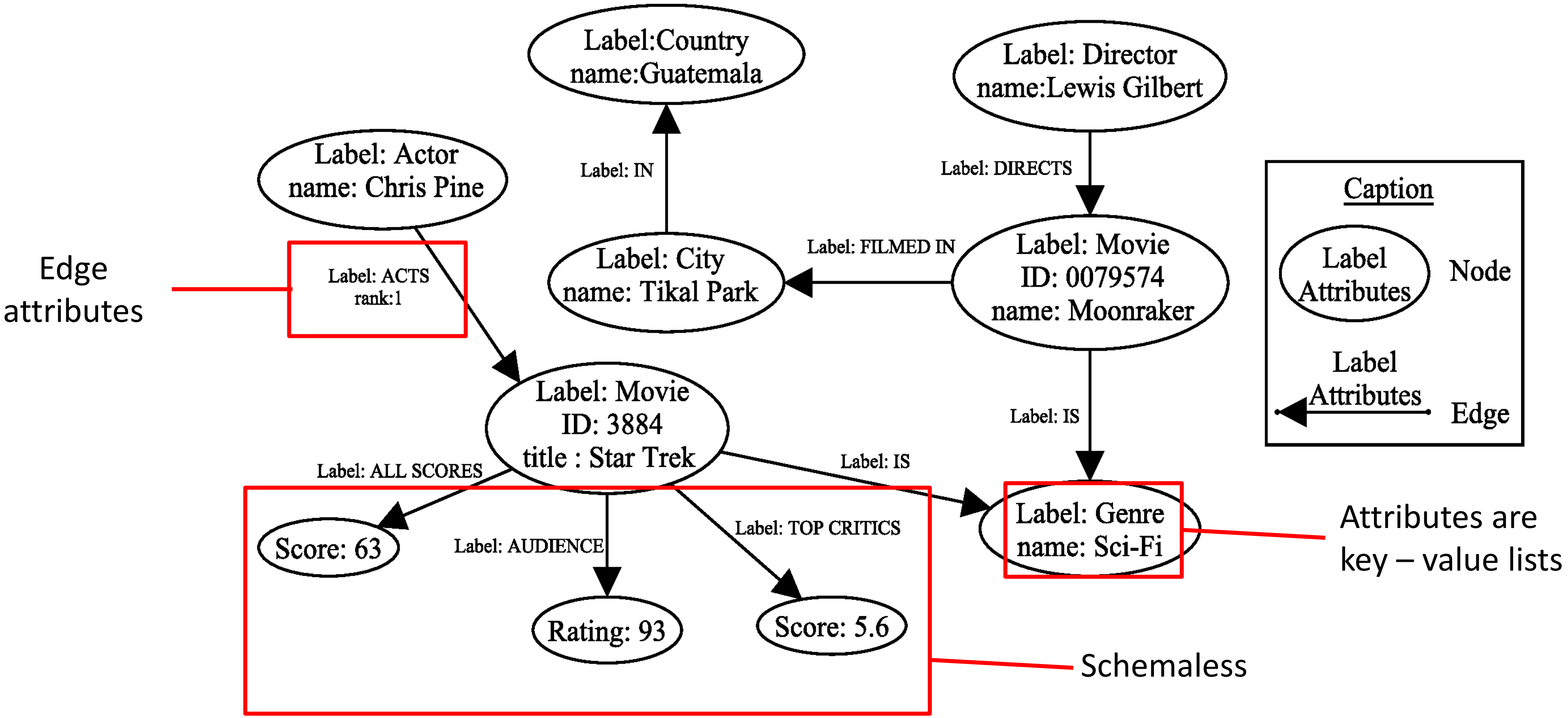
In-degree (of a node): #(incoming edges reaching the node)

- Source node: A node with 0 in-degree

Cliques and trees are specific kinds of graphs

- **Clique**: Every node is adjacent to every other node
- **Tree**: A connected acyclic simple graph

Example



Label:Country
name:Guatemala
Label: Director
name:Lewis Gilbert
Label: Actor
name: Chris Pine
Label: IN
Label: DIRECTS
Caption
Edge
attributes
Label: ACTS
rank:1
Label: City
name: Tikal Park
Label: FILMED IN
Label: Movie
ID: 0079574
name: Moonraker
Label
Attributes
Node
Label
Attributes
Edge
Label: IS
Label: Movie
ID: 3884
title : Star Trek
Label: ALL SCORES
Label: IS
Score: 63
Label: AUDIENCE
Label: TOP CRITICS
Label: Genre
name: Sci-Fi
Attributes are
key – value lists
Rating: 93
Score: 5.6
Schemaless

Showcasing Graphs

Crossing data from social networks it is possible to identify a graph like the one that follows:

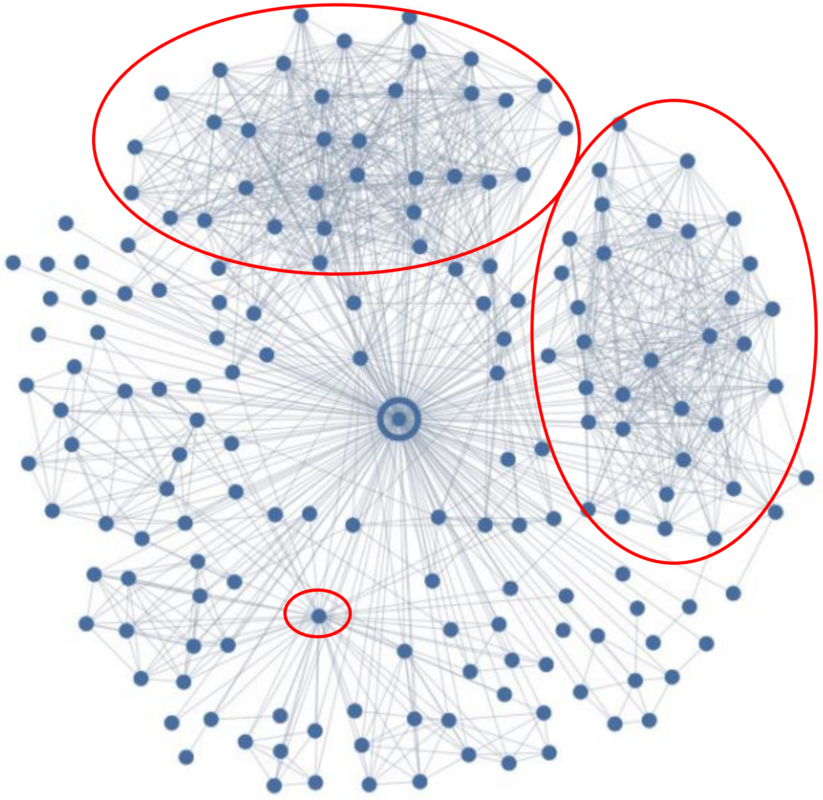
- In the centre there is a specific person *P*
- The rest are *P* connections and connections among them

Using sociology techniques…

- We can identify *P social foci*:
  - Dense clusters of connections, representing relationships
  - Typically, college friends, coworkers, relatives, etc.
- The *significant other* can be identified by a high *dispersion* rate
  - Highly connected with *P* connections,
  - But with a high dispersion degree wrt *P* social foci

Hypothesis: when the node with higher dispersion degree Identified is not the partner, this couple is likely to split up in a period of 60 days

L. Backstrom, J. Kleinberg. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook https://arxiv.org/pdf/1310.6753v1.pdf



Graph Data Models and Data Analytics

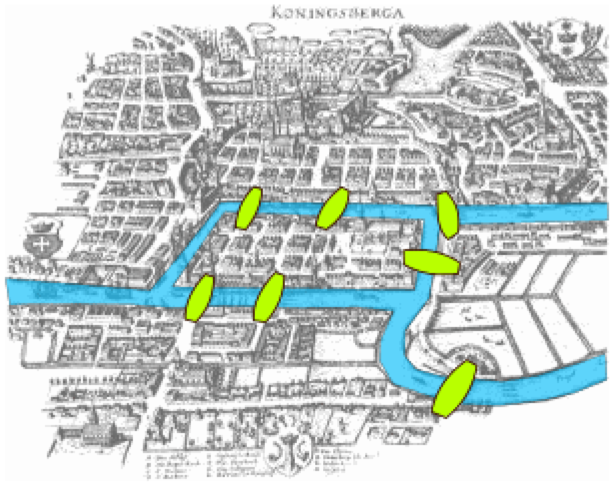
From a data management point of view:

- They are extremely flexible
- Schemaless by definition
- **Data and metadata are stored together** (i.e., data with annotations)
  - *Thus, we say that they store semantic (i.e., together with its meaning) data*
- Custom annotations facilitate data governance

Graphs are not only about data variety

From a data analytics point of view:

- Allow to exploit the data structure topology
  - Shortest path, centrality measures, community detection, etc.
- Graph data analytics is deterministic (i.e., by default non-probabilistic)
- Plenty of advances to enable probabilistic analysis on top of graphs
  - For example. graph embeddings and Graph Neural Networks (GNNs)



Seven Bridges of Koningsberg
(the born of graph theory)

Graph Data Models

What is a Graph Data Model?

Graph data models are composed of data structures, constraints and operators:

Data Structures

- Nodes
- Edges
- Attributes
- Etc.

Constraints

- From a data structure point of view: nodes and edges are disjoint
- From a schema point of view: schemaless

Operators

- Graph operators (grounded in the graph theory): pattern matching, reachability, neighbourhood, etc.
  - For these operations: graphs are translated into mathematical structures (!)
- Algebraic operators (coming from databases): selection, projection, join, union, aggregation, etc.
- Probabilistic operators (ML-based operators): prior a transformation of the graph into a vectorial representation

Graph Data Models

Two main families:

- **<u>Property Graphs</u>**
  - Born in the database field
  - Not predefined semantics
  - Follow a Closed-World assumption
  - Generate data silos
  - Algebraic operations on top of traditional graph operations
- **<u>Knowledge Graphs</u>**
  - Born in the knowledge representation field
  - May assume the Open-World assumption
  - Facilitate data sharing and linking
  - Two main families
    - RDF and RDF(S)
      - Born in the semantic web field
      - Vocabulary-based pre-defined semantics
      - Combine traditional graph operations, algebraic operations and simple reasoning operations
    - Description Logics (DL)-based languages (e.g., OWL)
      - Representation of (subsets of) first-order logic
      - Pre-defined semantics based on logics
      - Reasoning operations grounded in logics

Summary

Graphs are the perfect canonical data model to tackle data variety:

- Semantic expressiveness,
- Semantic relativeness

As a result, data and metadata (semantic annotations on data) are stored together

- Data is stored with its meaning
- Machine-readable metadata opens the door to automatic data management

Main graph families

- Property graphs
- Knowledge graphs

Thanks! Any Question?