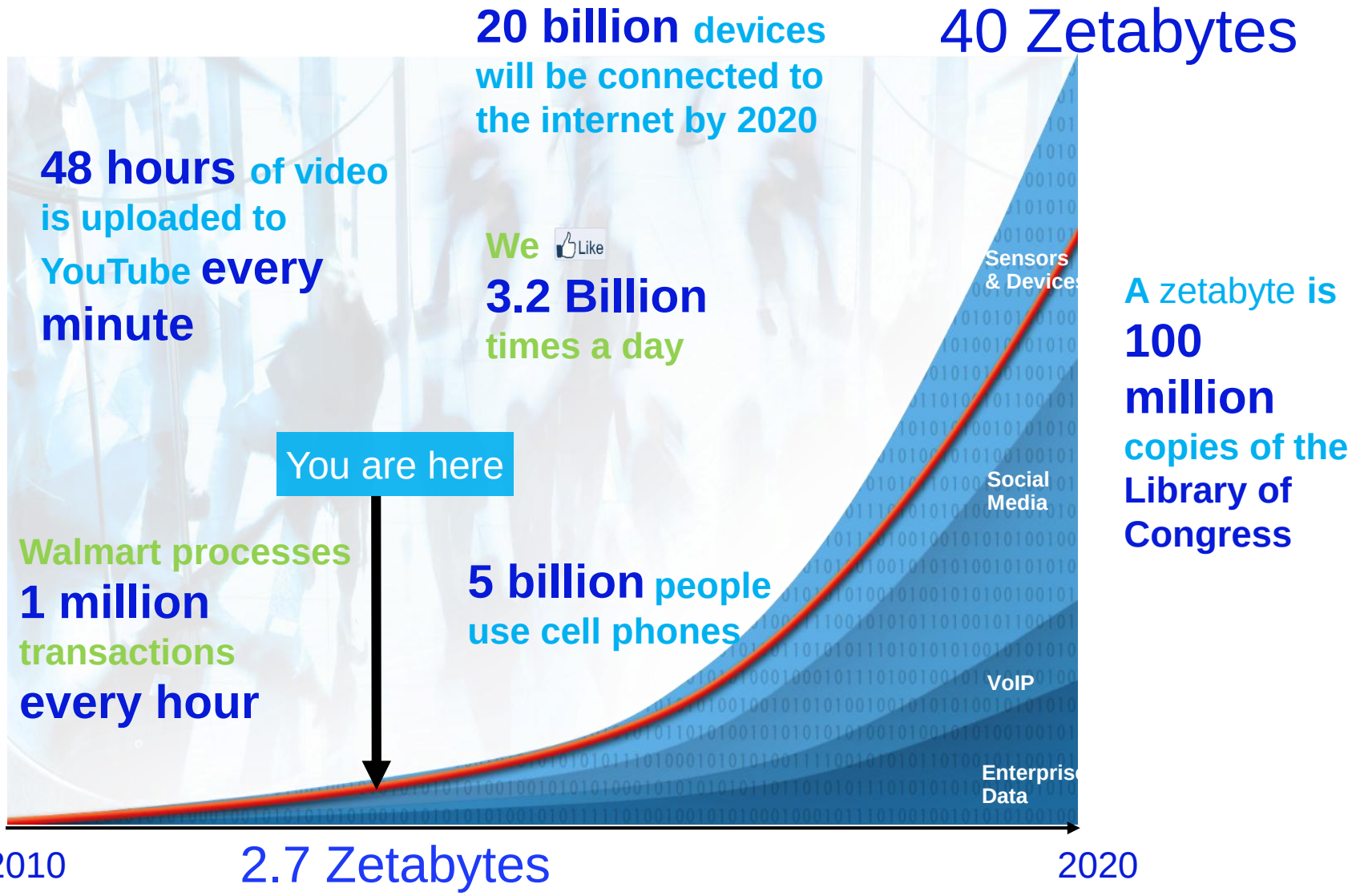
Data Wrangling: From the Wild to the Lake

Ignacio Terrizzano
Peter Schwarz
Mary Roth
John Colino

IBM Research - Almaden


20 billion devices will be connected to the internet by 2020
40 Zetabytes
48 hours of video is uploaded to YouTube every minute
We Like 3.2 Billion times a day
You are here
Walmart processes 1 million transactions every hour
5 billion people use cell phones
A zetabyte is 100 million copies of the Library of Congress
Sensors & Devices
Social Media
VoIP
Enterprise Data
2010
2.7 Zetabytes
2020

The New York Times

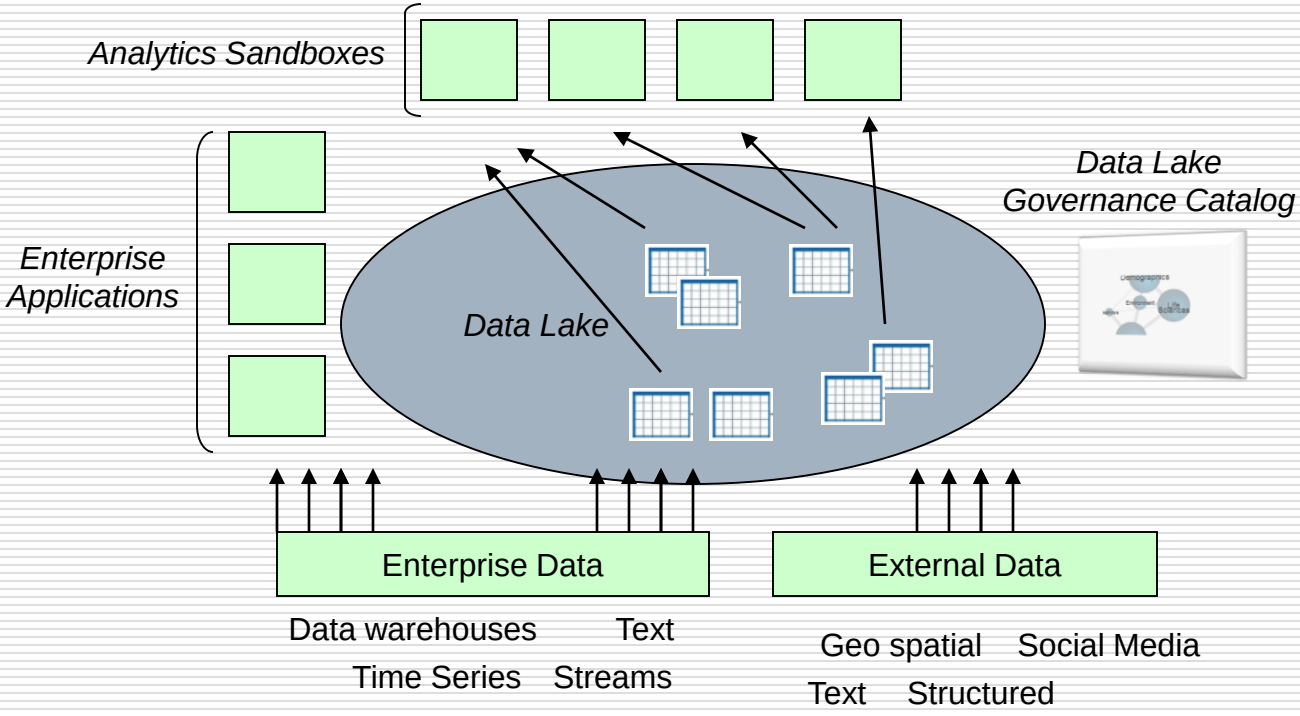
TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

- Where can I find the data I need?
- How do I obtain the data from the source?
- How can I relate this data to other data?
- How do I get the data to where I need it?
- How do I keep the data current?

A New Approach: The Enterprise "Data Lake"



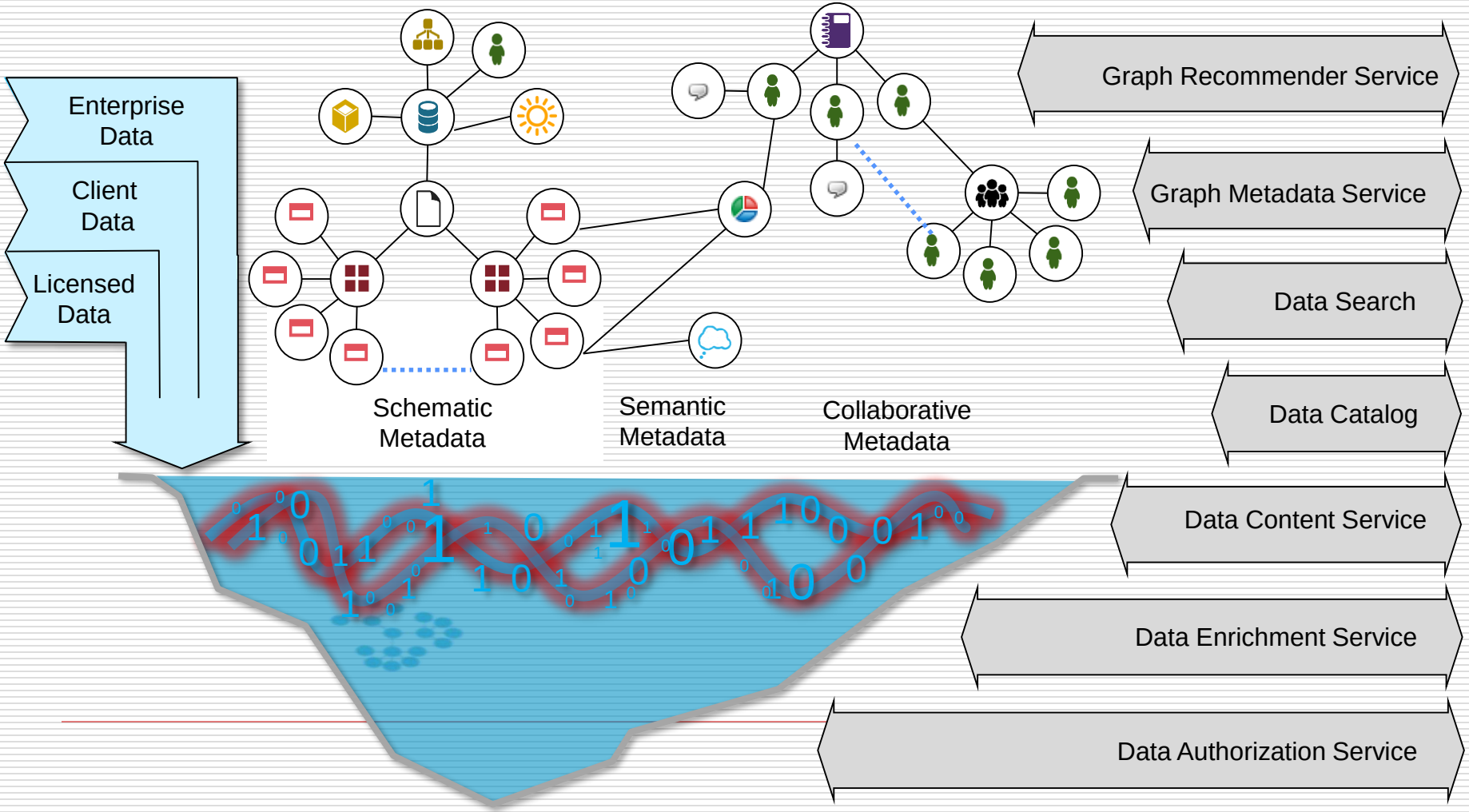
Data Lake or Data Swamp?

Data lakes therefore carry substantial risks. The most important is the inability to determine data quality or the lineage of findings by other analysts or users that have found value, previously, in using the same data in the lake … Without descriptive metadata and a mechanism to maintain it, the data lake risks turning into a data swamp. And without metadata, every subsequent use of data means analysts start from scratch.

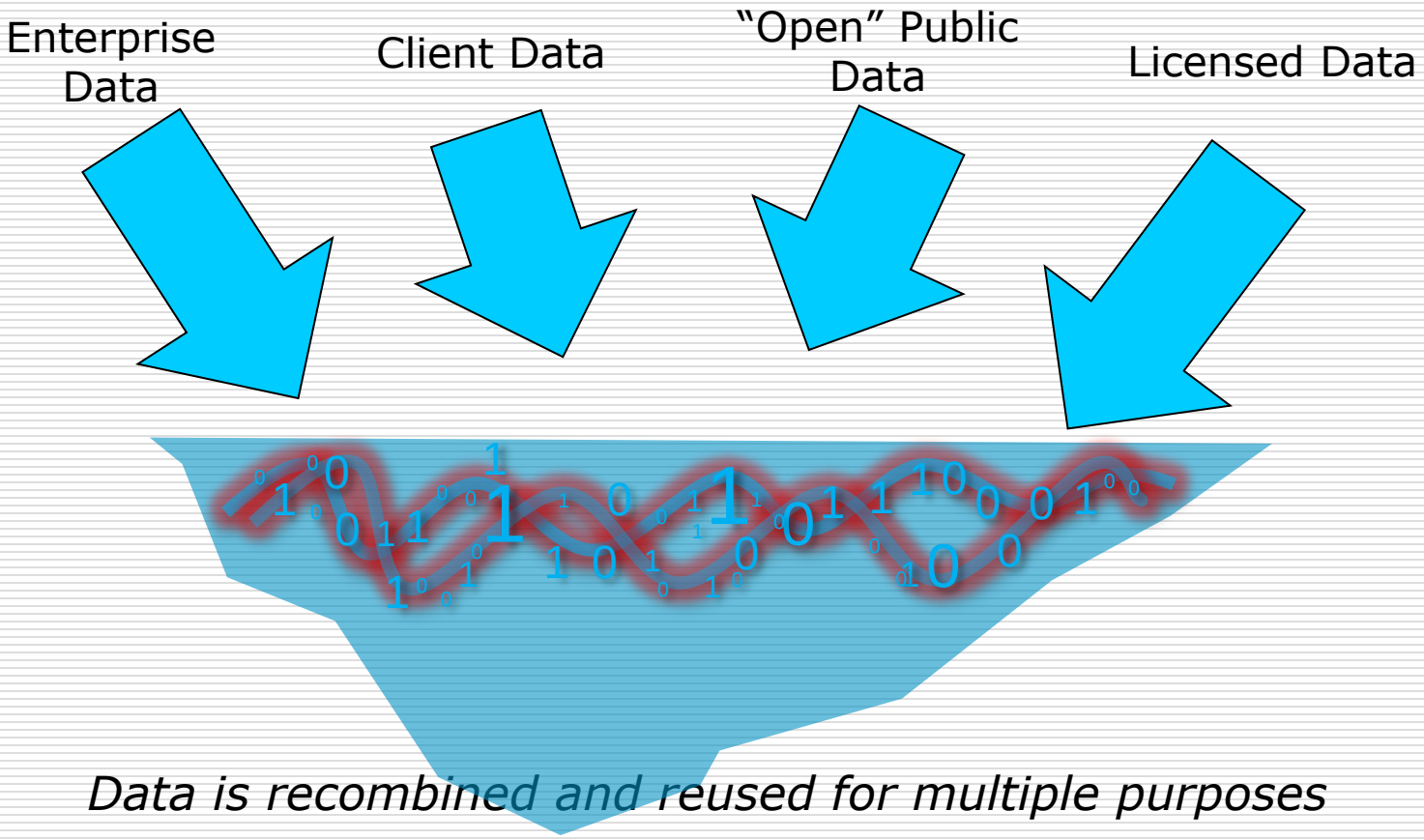
- Gartner analyst Nick Heudecker, July 2014

OR

Metadata Repository and Services



Not Just More Data, More *Kinds* of Data



Data is recombined and reused for multiple purposes

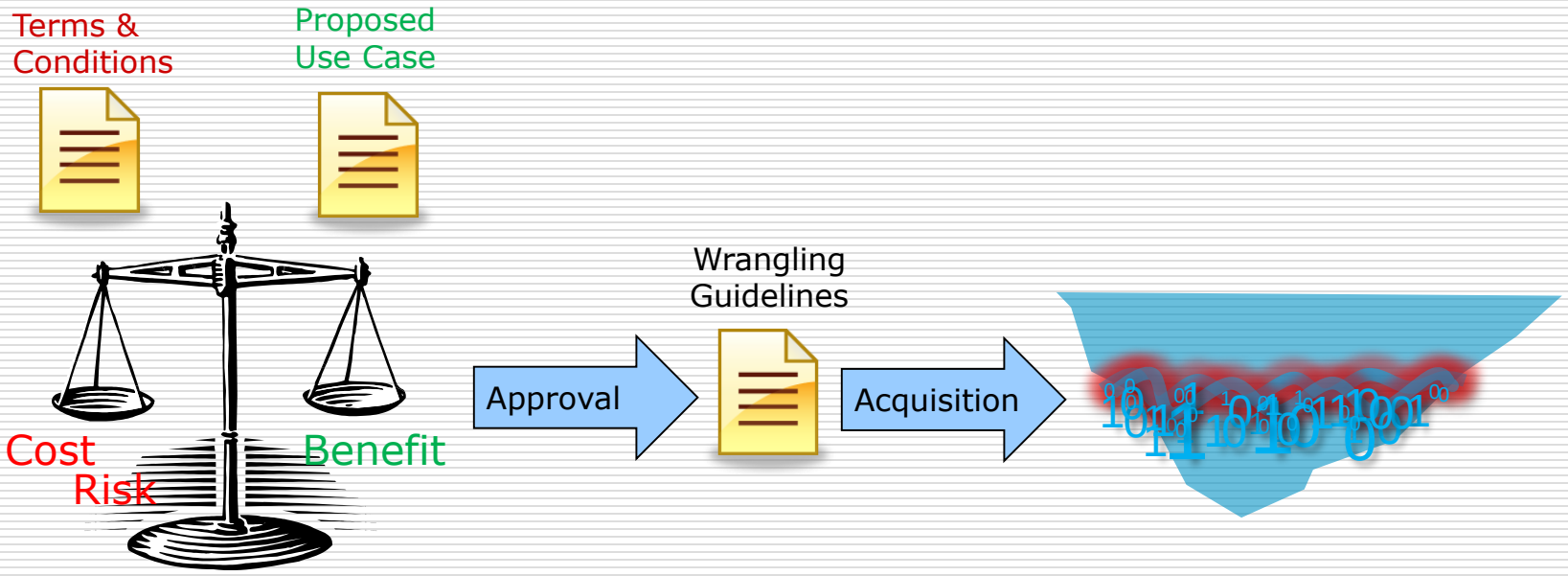
Data Licensing

- Very complicated!
  - Like open source, but with few industry-friendly standards
  - Terms & Conditions in legalese
- Issues include:
  - Commercial use
  - Cost
  - Copyright
  - Indemnity for errors
  - Derived works
  - Redistribution rules
  - Retention/expiration rules
  - Export regulations
  - Privacy considerations
  - Citation requirements
  - Wrangling rules

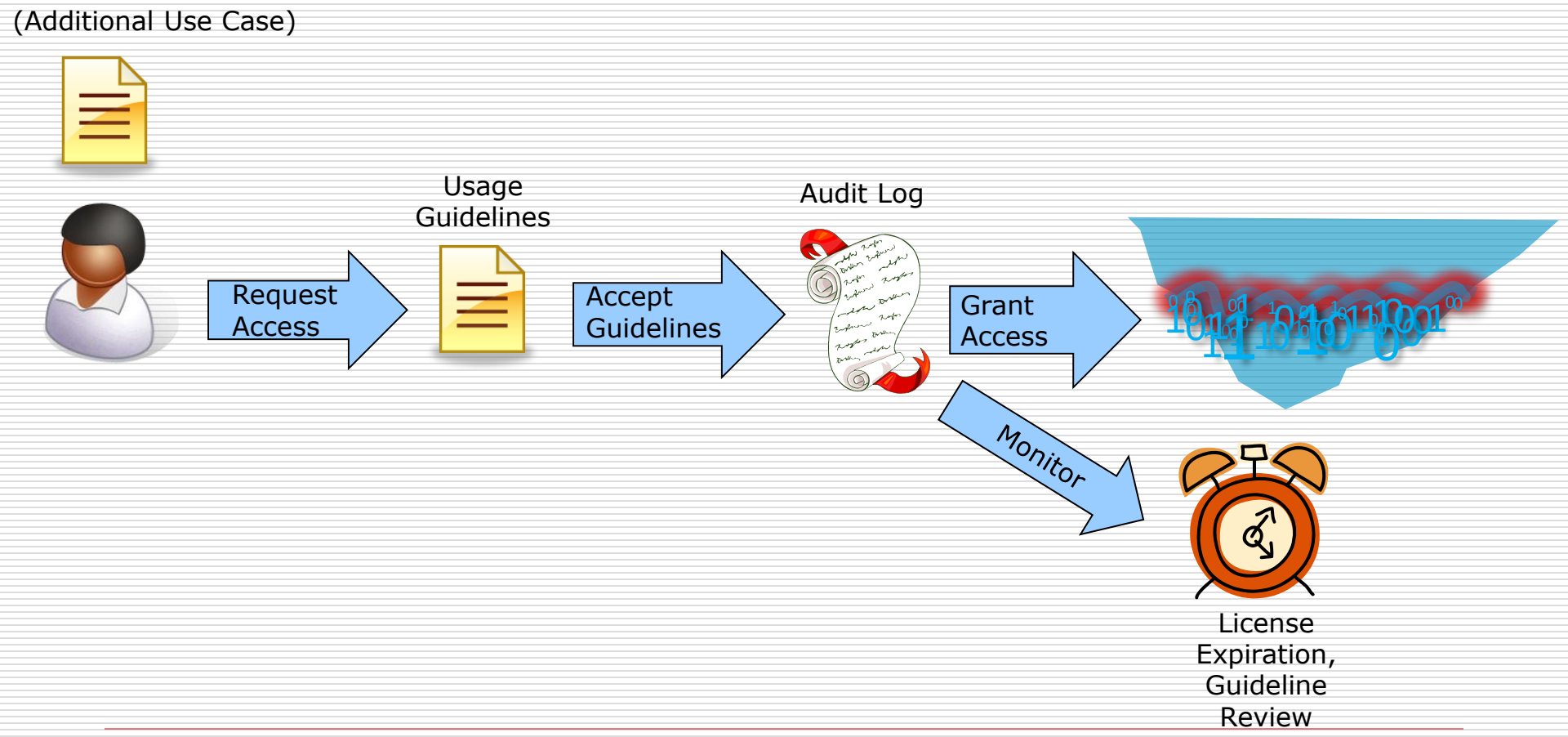
Data Challenges for the Enterprise

- Who is using what data?
- For what purpose?
- What restrictions are there on its use?
- Where did it come from?
- Does someone already have what we need?
- Is it up-to-date?

Data Governance Process: Acquisition



Data Governance Process: Access & Monitoring

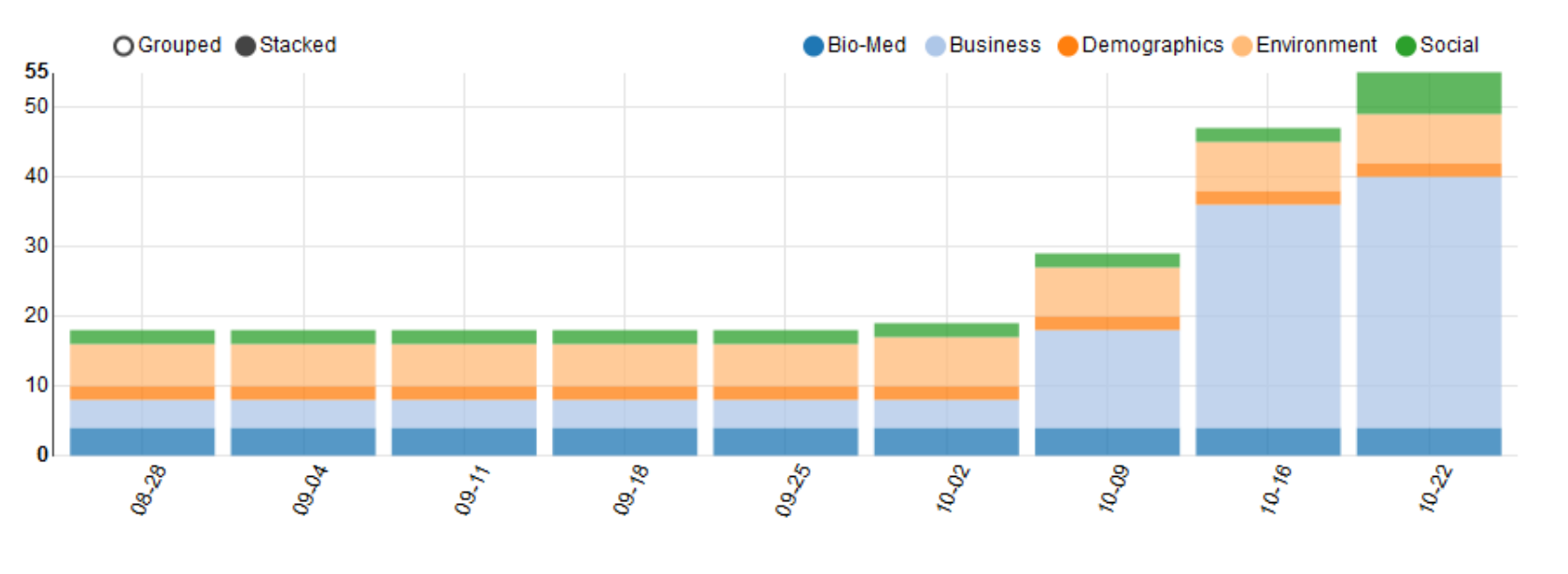


The IBM Research Data Lake

- A place for researchers to find the data they need
  - Reliably sourced
  - Legally sound
  - Enriched with metadata
- A place to contribute data for other researchers to use
  - (see all of the above)
- Usage is governed
  - Comprehensible end-user guidelines
  - Well-defined processes for acquisition, access & contribution
  - Auditable records maintained of all documents and agreements

The IBM Research Data Lake

- Small now, but growing
- Domain experts identifying key datasets
- Volunteer "cowboys" wrangling approved data
- Governance process in place



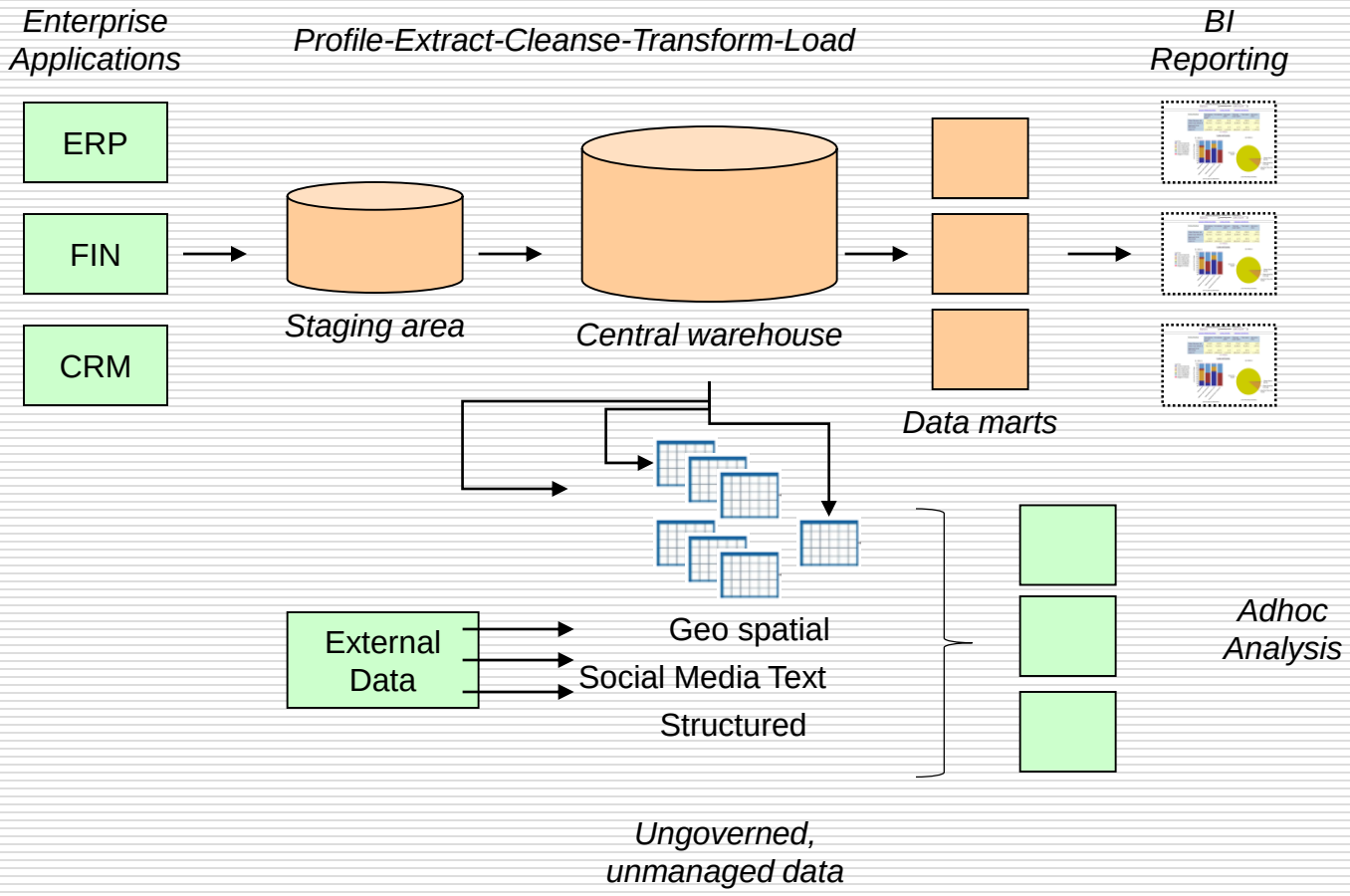
The IBM Research Data Lake

- Part of a complete collaborative data exploration environment
- Hosted in the Accelerated Discovery Lab at IBM Research – Almaden
  - Presence in both internal and external clound environments
- Pilot engagement with a Metagenomics project, starting now
  - Consortium with members from IBM, a university and industry
  - Secure data management and governance a must!
- Will serve 500+ researchers, starting this year!

Thank You!

Questions?

Typical Enterprise ETL Architecture



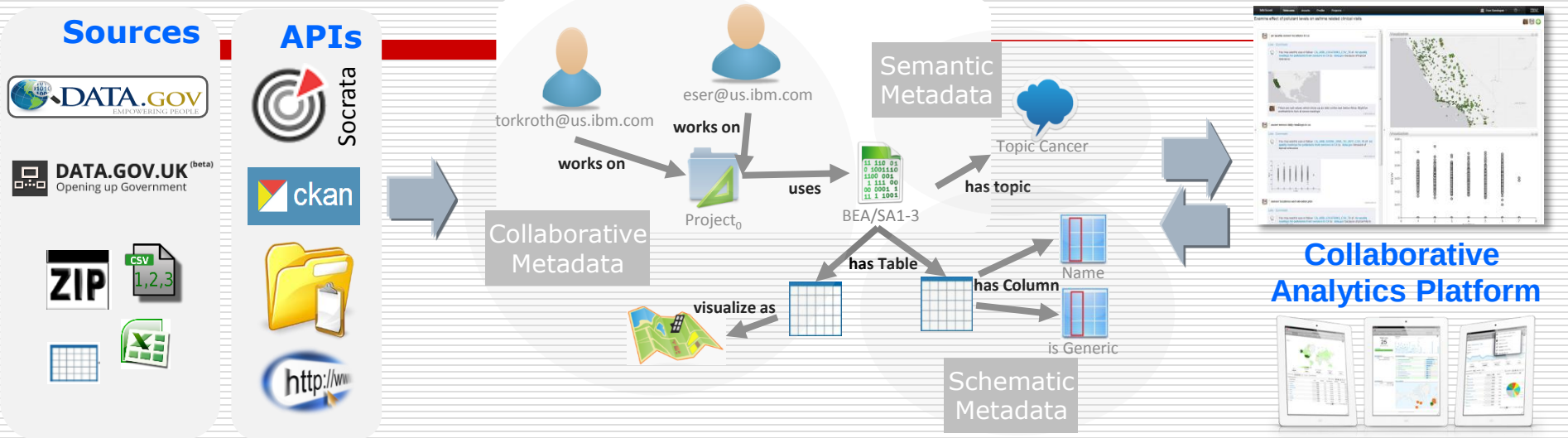
Enriching the Data Lake With Metadata

- Schematic metadata - How is this data represented?
  - Delimiters, formats, datatypes…
- Semantic metadata - What does this data mean and how is it related to other data?
  - Annotate objects (data sets, tables, columns, etc.) with text, tags and provenance (creator, publisher, contact information, etc.)
  - Link objects to standard ontologies (e.g. DBPedia) and business terms
  - Exploit algorithms that link objects to similar objects
- Collaborative metadata - How has this data been used?
  - Track relationships among people, organizations, data sets

Acquire

Enrich

Provision



Graph Metadata Store

Analytics Sandboxes

Automated and semi-automated tools to harvest and curate data (public and private), making it easier to use and govern

At the core is semantic, collaborative, and schematic graph metadata store that connects data to users in a meaningful way.

Users provision and contribute data and metadata via a collaborative analytics platform that serves analytics sandboxes in a contextually-ware, social, visual, and interactive way.

Data Governance Process

- Identify the dataset to be obtained
  - Identify owner, source
  - Locate licensing terms and conditions, determine cost
- Determine the requestor's use case
  - Research or commercial?
  - Worldwide or local use?
  - Will the data be displayed?
  - Will the data be redistributed?
  - Will derived works be created? Distributed?
- Weigh:
  - Potential Risk
  - Potential Benefit
  - Cost
- Data wrangler obtains the data and makes it available
  - Subject to specific Wrangling and User Guidelines

If Acquisition is Approved….

- Data wrangler adds data to the lake
  - Subject to specific Wrangling Guidelines
- Users “check out” data
  - Subject to specific Usage Guidelines
  - User’s acceptance of guidelines logged and retained
  - Periodic re-acceptance required (e.g. annually)
- Each additional use case requires separate approval