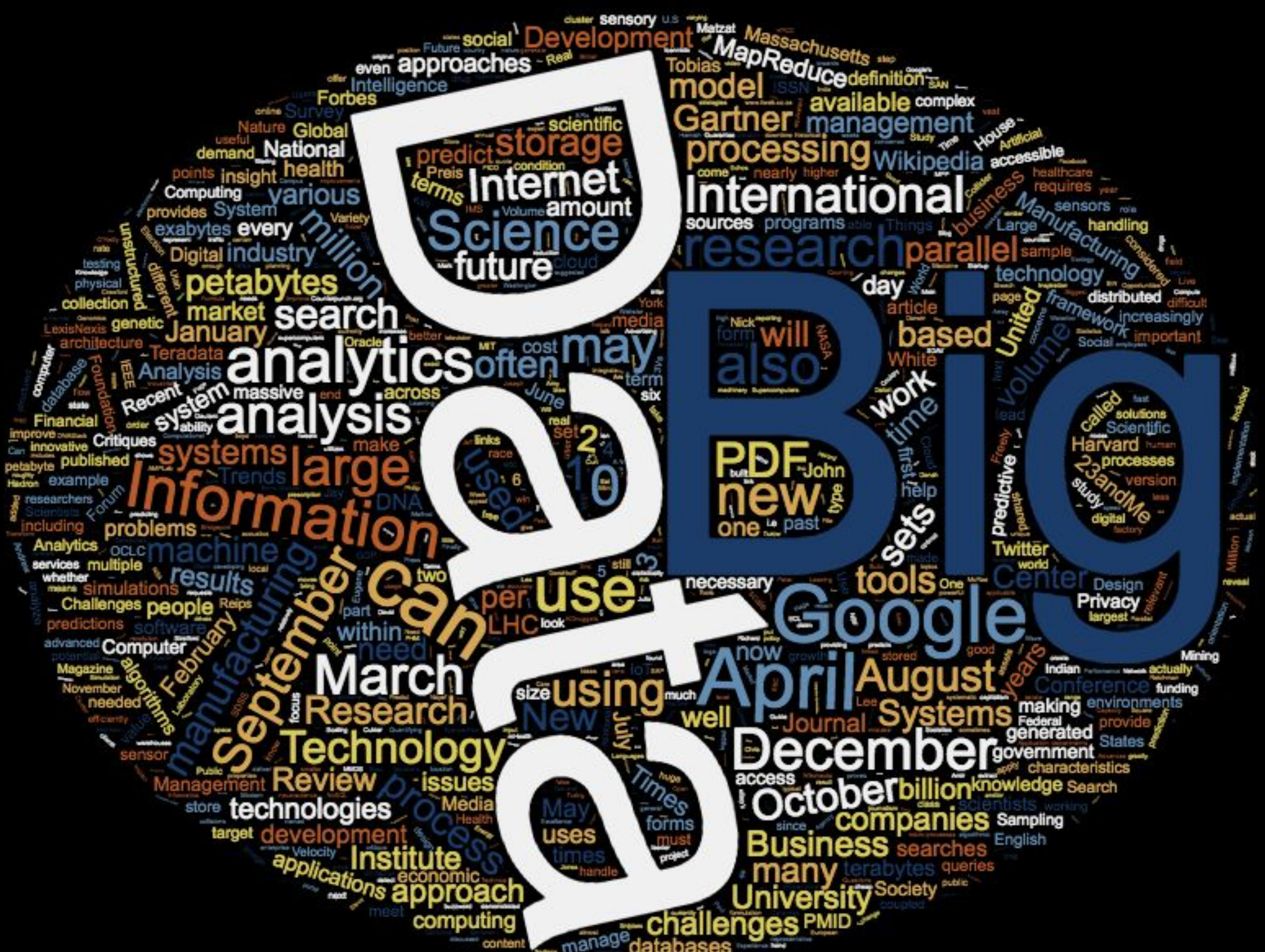
Week 1: From Relational Data to Big Data

Big Data Management – Fall 2017

Björn Þór Jónsson



Generated by http://www.wordclouds.com/ with text from https://en.wikipedia.org/wiki/Big_data

Relational Databases

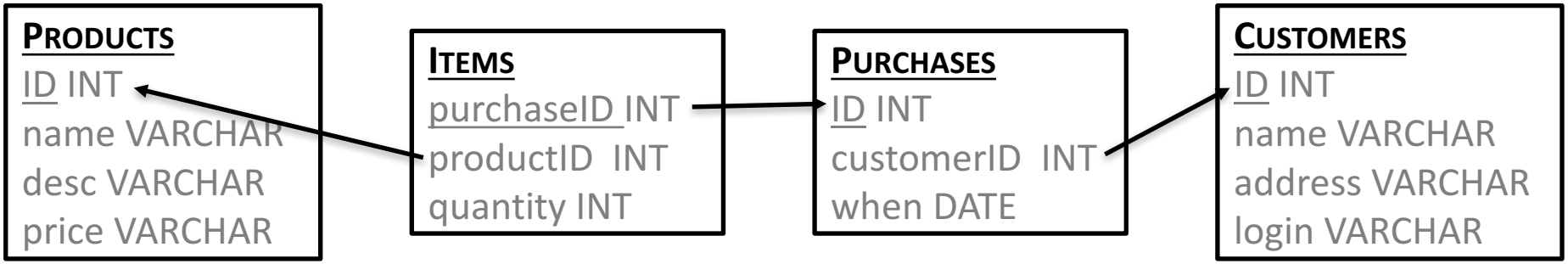
- What are the main techniques and concepts?

Relational Database Design

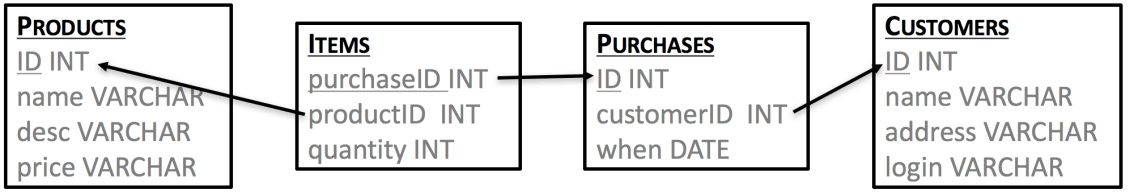
- How are databases traditionally designed?

For an online store we need to store information about customers, such as address, e-mail, and login information. We need to store information about our products, such as name, description, and price. Then we need to track all customer purchases, when they took place, which products were bought and how many copies of each.

Any problems with this design?



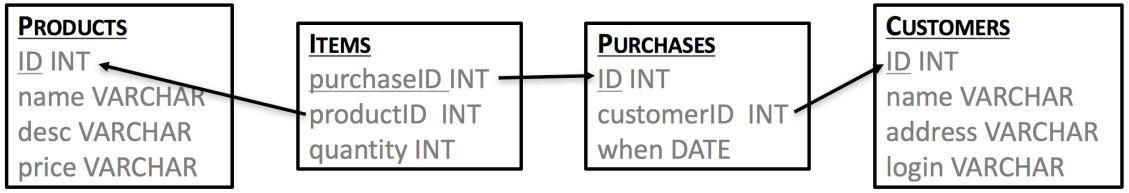
Problem 1: History



- What if…
  - Price changes?
  - Customer moves?

- Current state vs. Historical information
  - Traditional solution: Data warehouses
    - Snapshots of databases to keep track of history
  - Recent solution: Lambda architecture
    - Store everything and extract current state
  - Either way: Much more data!

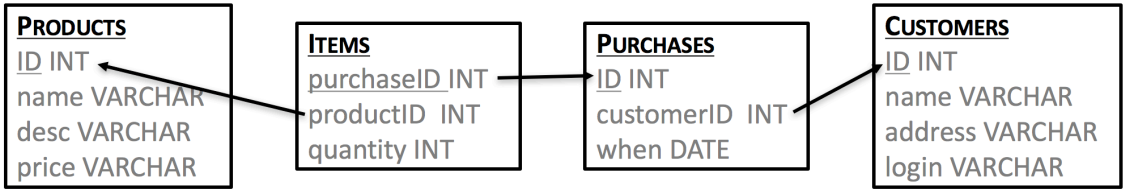
Problem 2: Coarseness



- What if…
  - Customer has problems with web-site?
  - Customer looks at items but does not purchase?
  - Do we want to know?

- Want much more detailed overview
  - Solution: Storing and analyzing logs
  - Logs are extensive: MUCH MORE data!

Problem 3: Simple Structure

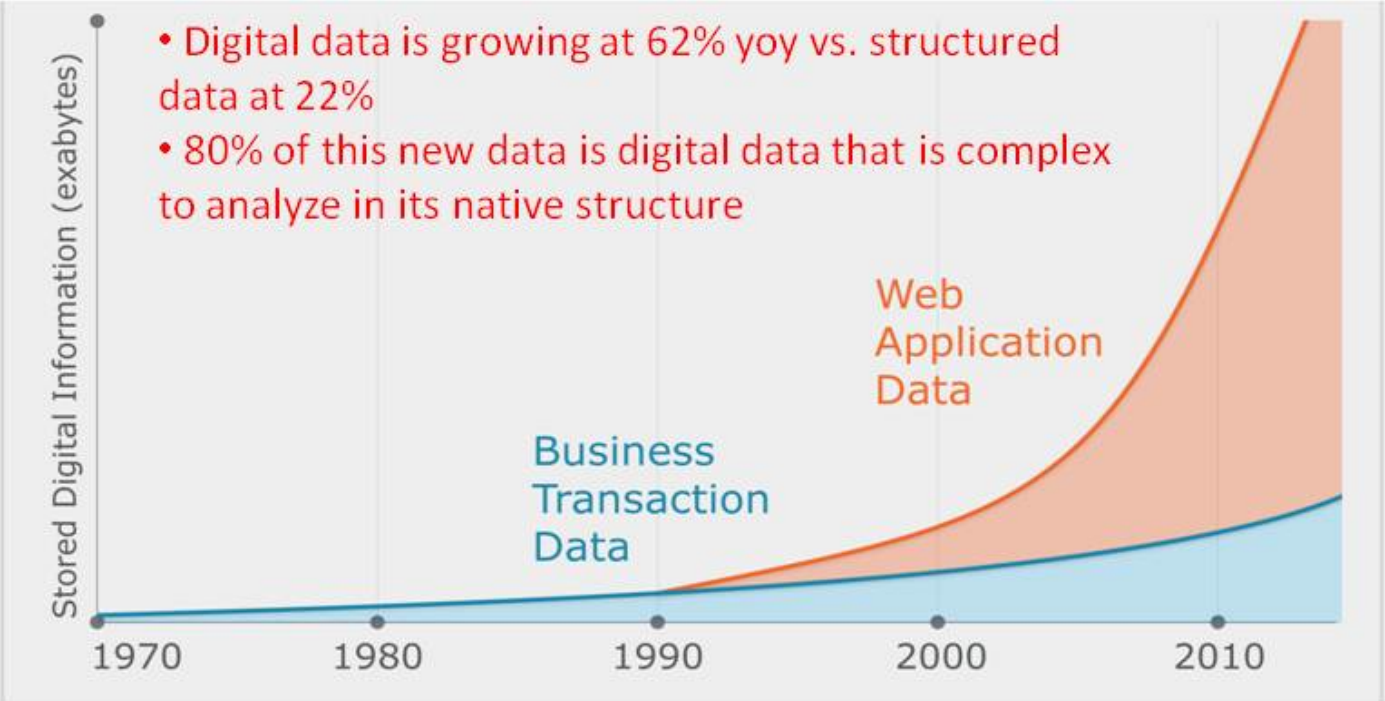


- What if…
  - We want to allow (and keep track of) comments?
  - Media (images/videos) is involved?

- Want to store any data type
  - Relational model forces you to **think inside the box**!
  - Unstructured and semi-structured data
  - Media is large: **MUCH MORE** data!

Beyond Relational Data!

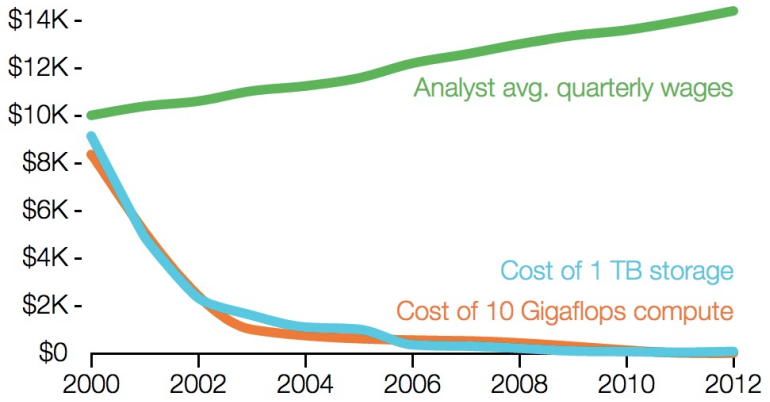
- Keep track of **all** the history
- Keep track of **all** interactions, also low-level
- Keep track of **all** data: media, text data, logs, …



Source: https://practicalanalytics.wordpress.com/2011/11/06/explaining-hadoop-to-management-whats-the-big-data-deal/

Handling Data Growth?

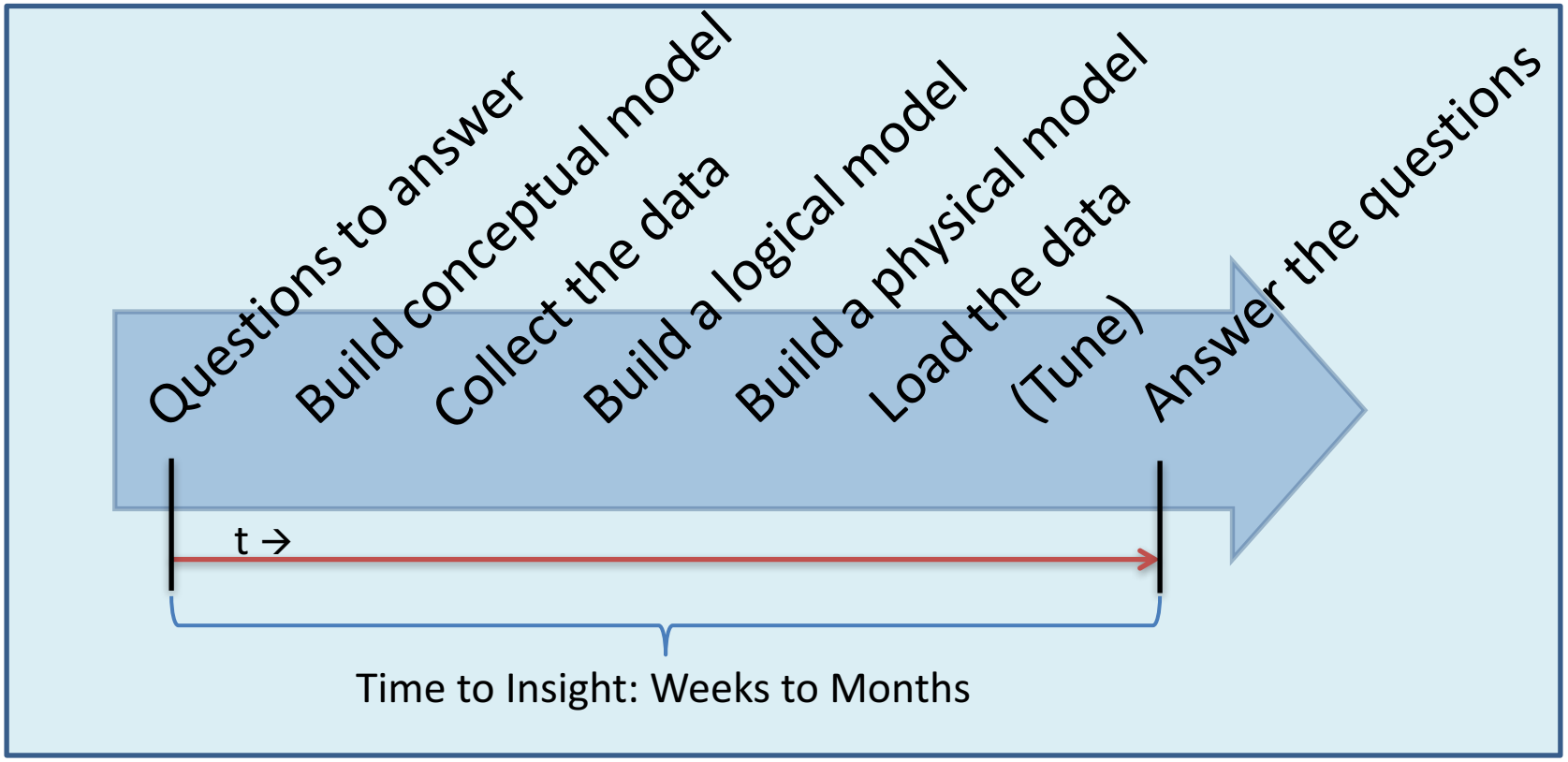
- Storage is cheap – people are not!



- Need better technology to handle the data

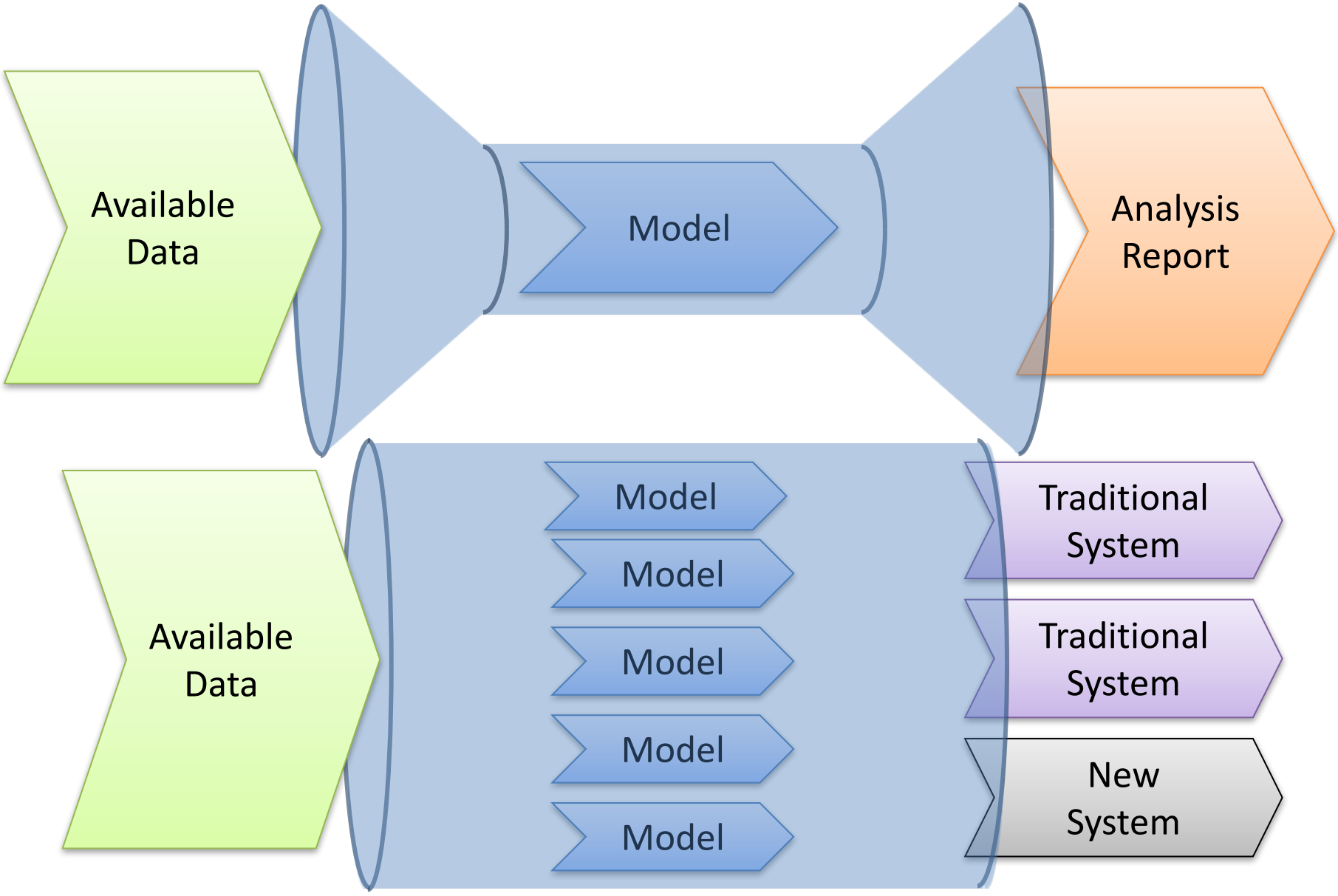
➔ *Business Push*

Traditional "Sense Making"

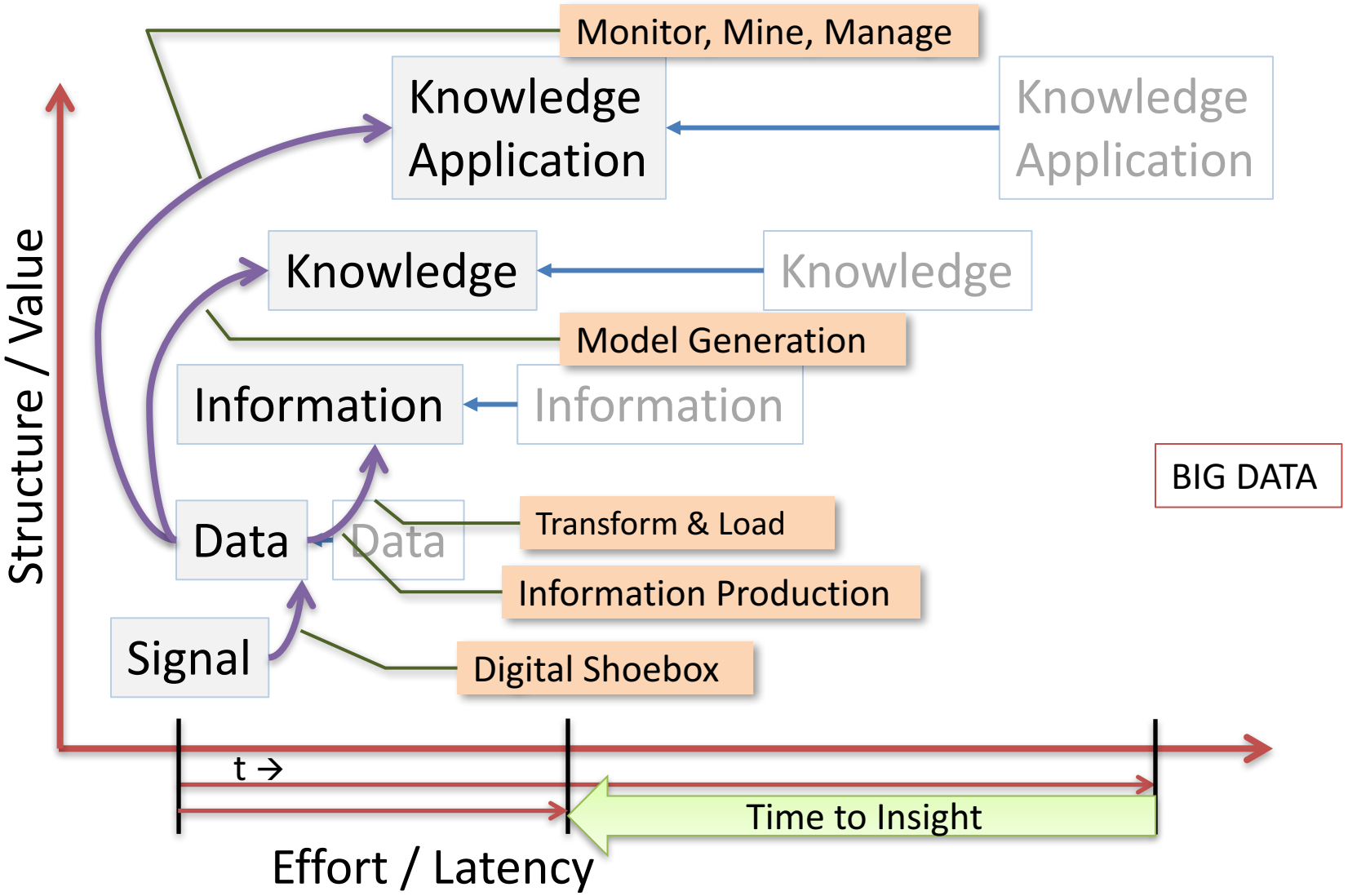


OLD SCHOOL

Source - http://www.youtube.com/watch?v=2YWvmBtEymE



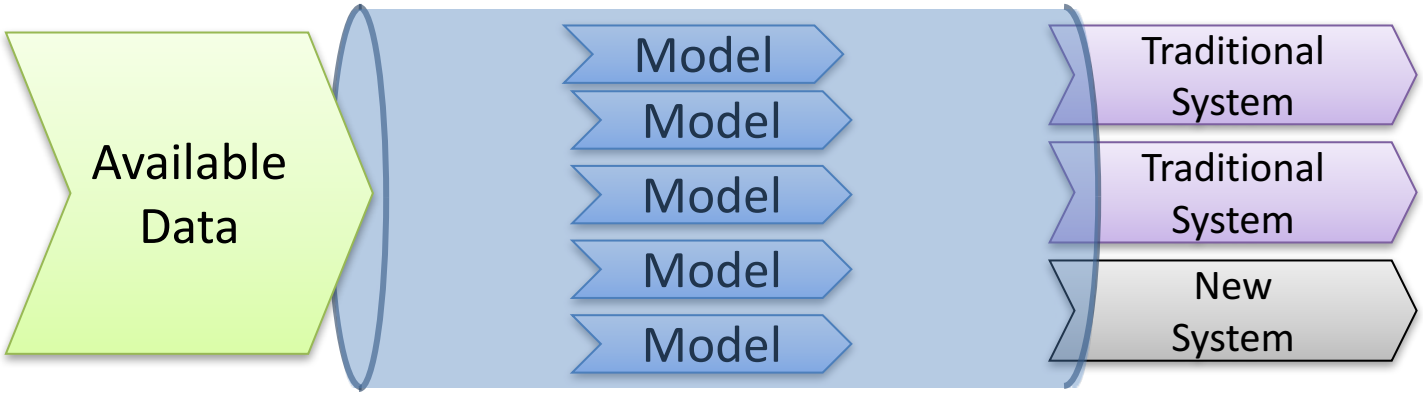
Source – http://www.vldb.org/2011/files/slides/keynotes/**campbell**_keynote.pptx



Source: http://www.vldb.org/pvldb/vol4/p694-campbell.pdf

Coping with a Faster World?

- Decisions must be made ever faster!

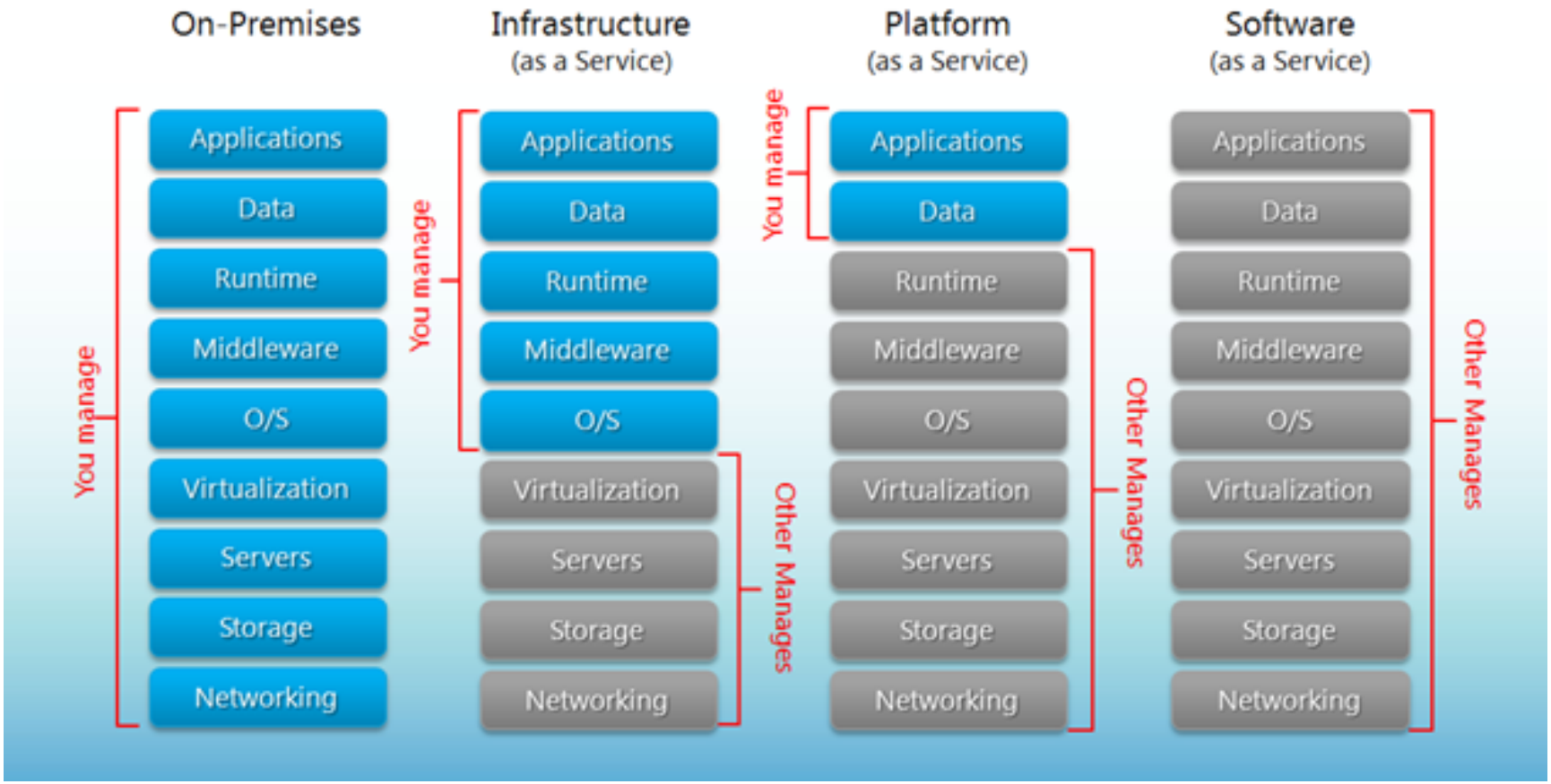


- Need better technology to use data **now**!

➔ *Application Pull*

Technology Trends

Separation of Responsibilities



Source: http://blogs.technet.com/b/kevinremde/archive/2011/04/03/saas-paas-and-iaas-oh-my-quot-cloudy-april-quot-part-3.aspx/

Warehouse-Scale Computer

THE DATACENTER AS A COMPUTER

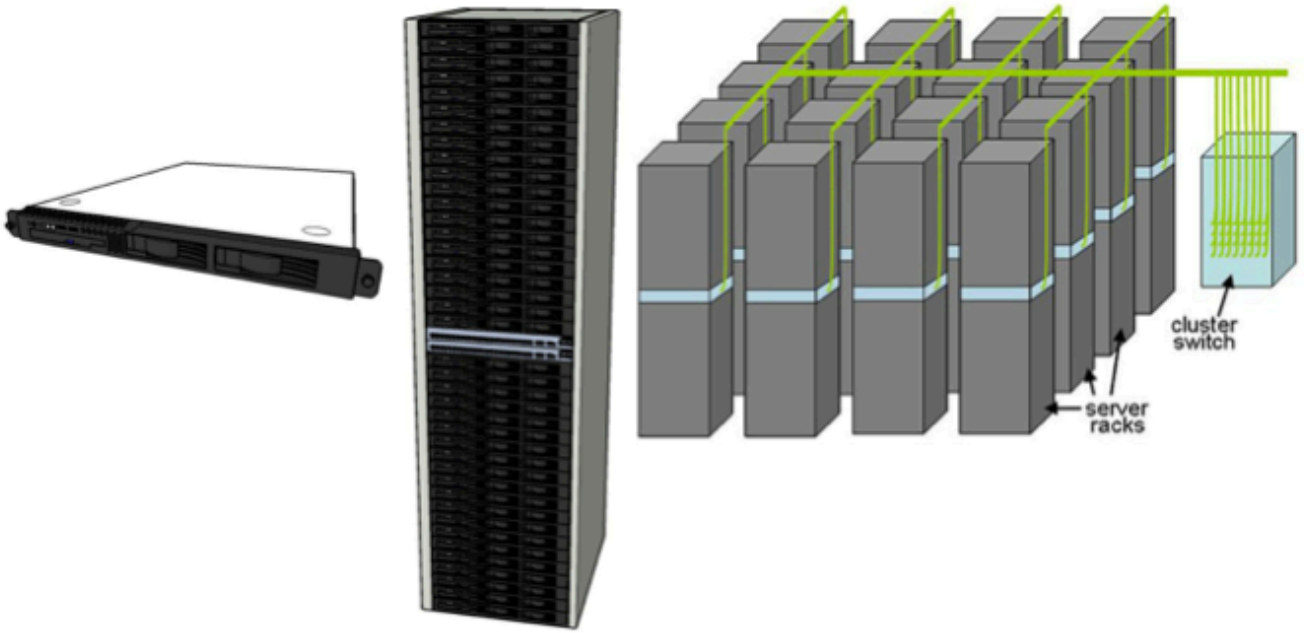


FIGURE 1.1: Typical elements in warehouse-scale systems: 1U server (left), 7′ rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).

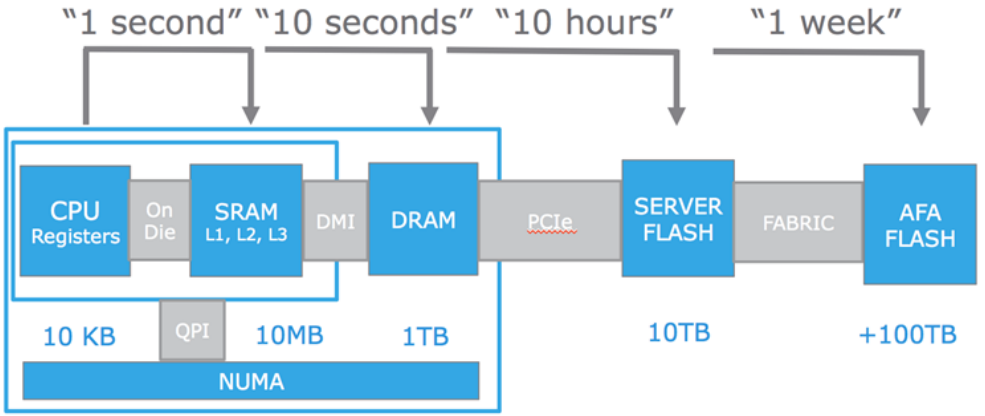
LOOK UP: Werner Voegels on virtualization.

Source: http://www.morganclaypool.com/doi/abs/10.2200/S00193ED1V01Y200905CAC006

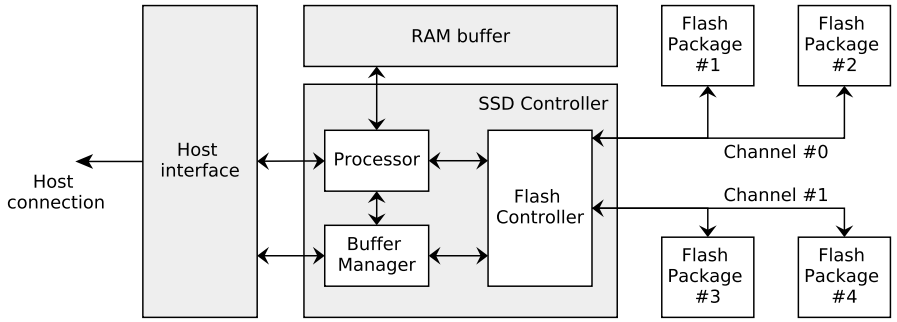
Storage Trends

Device	Read	Write
Millisecond Scale		
10G Intercontinental RPC	100 ms	100 ms
10G Intracontinental RPC	20 ms	20 ms
Hard Disk	10 ms	10 ms
10G Interregional RPC	1 ms	1 ms
Microsecond Scale		
10G Intraregional RPC	300 μs	300 μs
SATA NAND SSD	200 μs	50 μs
PCIe/NVMe NAND SSD	60 μs	15 μs
10Ge Inter-Datacenter RPC	10 μs	10 μs
40Ge Inter-Datacenter RPC	5 μs	5 μs
PCM SSD	5 μs	5 μs
Nanosecond Scale		
40 Gb Intra-Rack RPC	100 ns	100 ns
DRAM	10 ns	10 ns
STT-RAM	<10 ns	<10 ns

M.Wei et al. I/O speculation in the microsecond era. Usenix ATC'14.

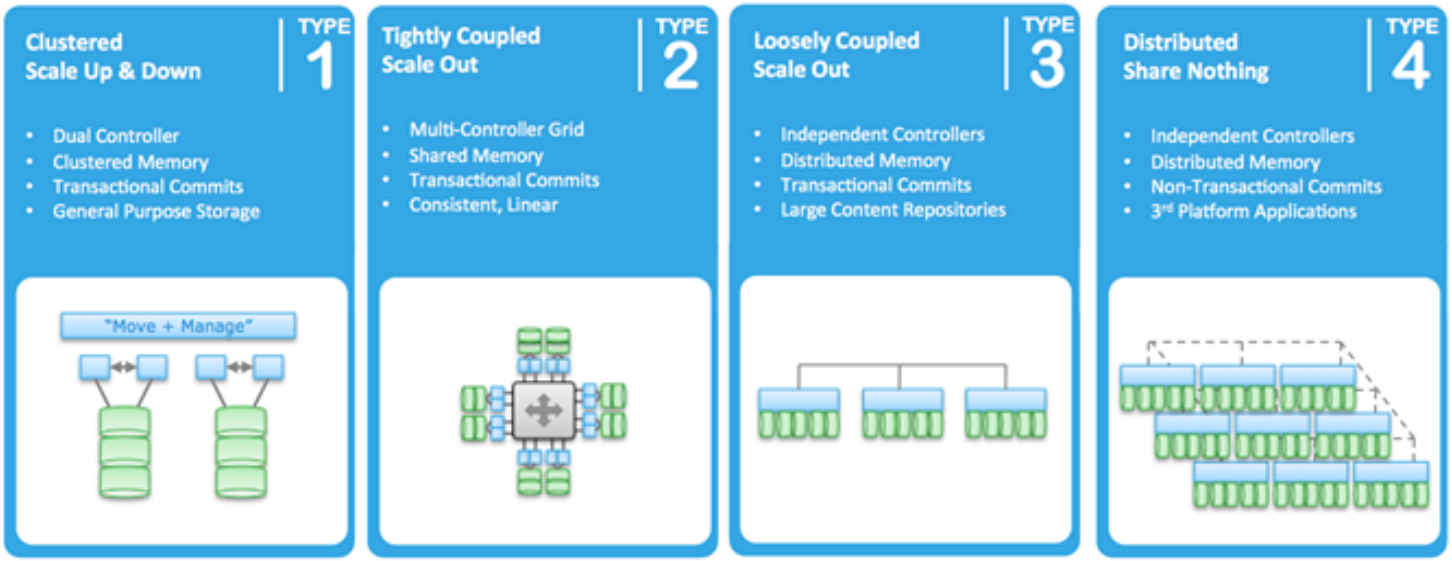


source: Virtual Geek's take on storage tree of life

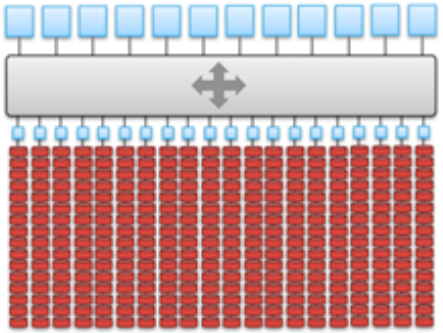


SSD Architecture

Storage Architectures

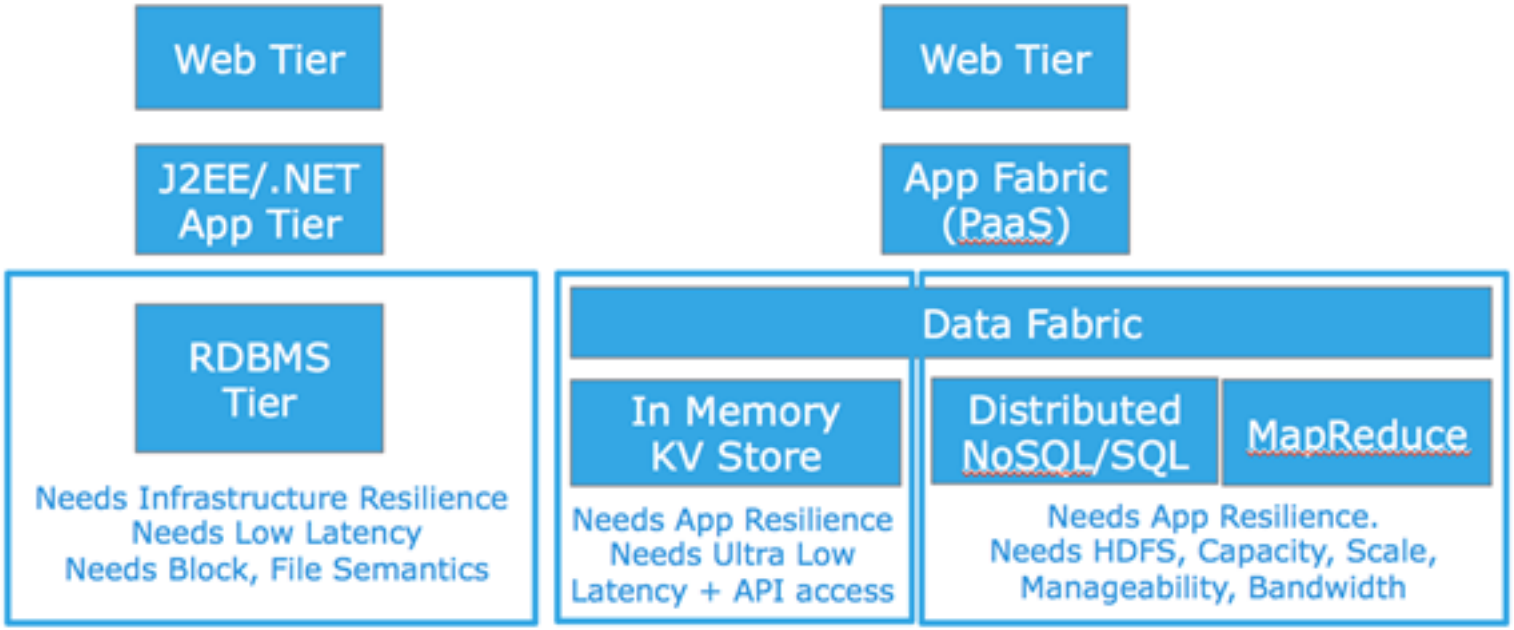


source: Virtual Geek's take on storage tree of life – A MUST READ!!



Storage
RAM
Interconnect

Data-Intensive Applications: Server-side Architectures

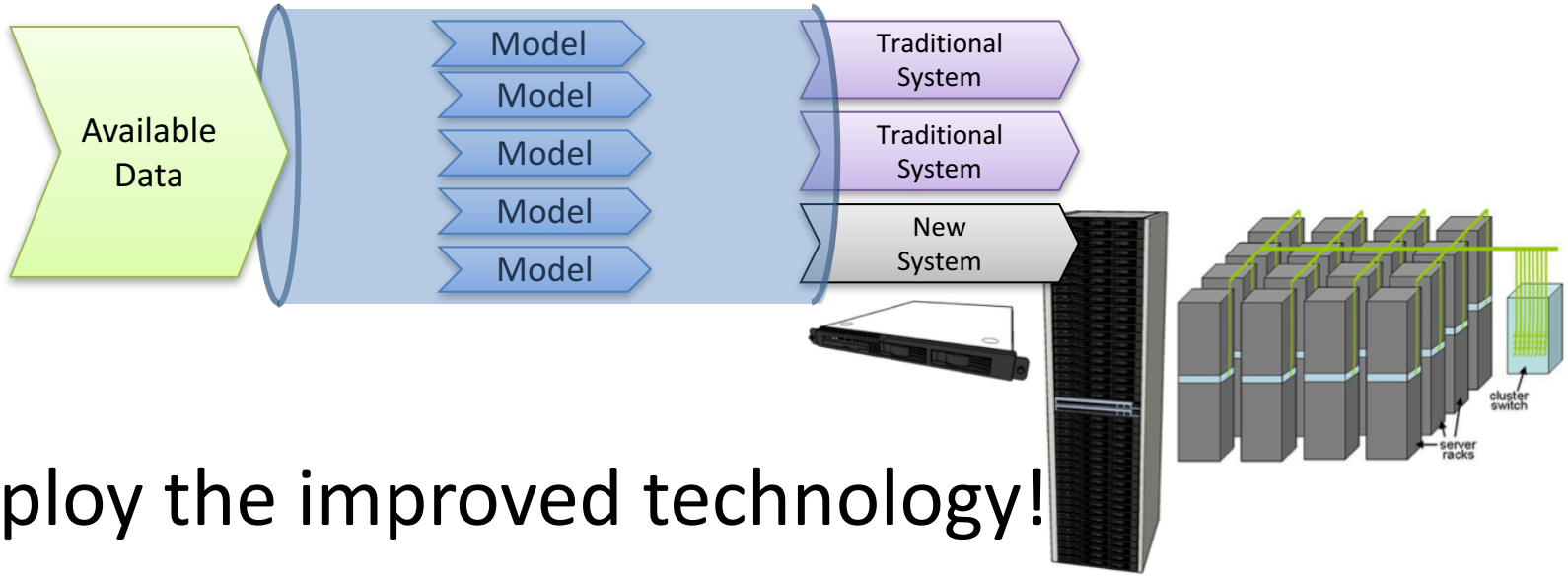


Look up Fabric Computing on Wikipedia.

source: Virtual Geek's take on storage tree of life – A MUST READ!!

Utilizing a Faster World ☺

- Hardware architectures getting ever better!



- Employ the improved technology!

➔ *Technology push!*

Outline

- Trends Underlying Big Data
  - Business Push
  - Application Pull
  - Technology Push
- **What is Big Data?**
  - Relational vs. Big Data Management
  - Examples
  - Definitions (attempts?)
- Data models
- Course Outline

Relational Database Management

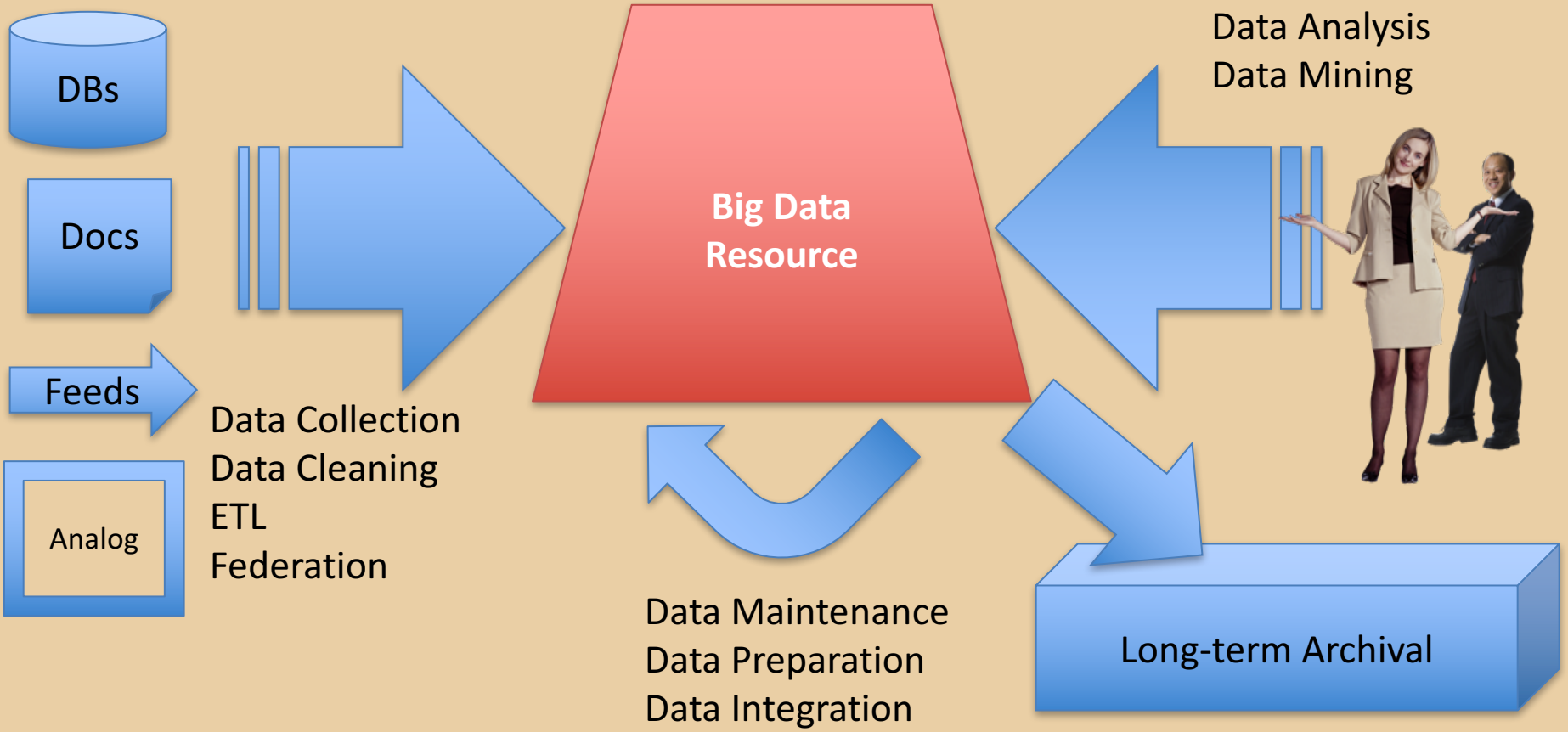
- **Narrow scope**
  - A database is created to serve a well defined purpose
- **Structured data**
  - Conceptual/Logical/Physical schema
  - Relational model dominates since 80s
  - Entity Relationship defines conceptual schema
- **Closed-world assumption**
  - Data as an instance of the schema
  - The data which is not part of an instance *does not exist*
    - *Any query on the database returns a value based on the current instance*
- **Data at rest**
  - Data is loaded and stored in the database, on disk.
- **Fully interactive architecture**
  - 3 tier architecture: Web server; App Server; Database Server.

What is a Big Data Resource?

- A big data resource is a **collection of data which is made available for analysis**
  - Analysis of data is a **process** *of inspecting, cleaning, transforming, and modeling* data with the goal of discovering useful information, suggesting conclusions, and supporting decision making.
- For example:
  - The databases that underly the learnit blog or FM's Building Management System are not big data resources
  - The data made available by Eurostat on unemployment in Europe can be considered a big data resource.
    - More examples of this kind at Google Public Data Explorer

What is Big Data Analysis?

- Need for insight based on data not currently available for analysis
  - How can we set goals for energy management at the IT University?
    - We do not know how electricity is consumed.
    - We do not know where electricity is consumed.
    - We do not know how to link electricity consumption and people's practices.
  - How long does it take you to answer your mails?



Big data is not a product, but a collection of **processes**

Big Data Management

- **Wide scope**
  - Data is made available for yet-to-be-defined analysis
- **Data Variety**
  - Time series are highly structured; Text is not
- **Open world assumption**
  - Data sources might be added or removed
  - So any analysis is only valid based on the current state of the big data resource
- **Data in movement, and at rest**
  - Data streams complements stored data
  - Some data streams are stored, others are not
- **Lambda Architecture (or variant thereof)**
  - Batch layer; serving layer; speed layer; Analytics

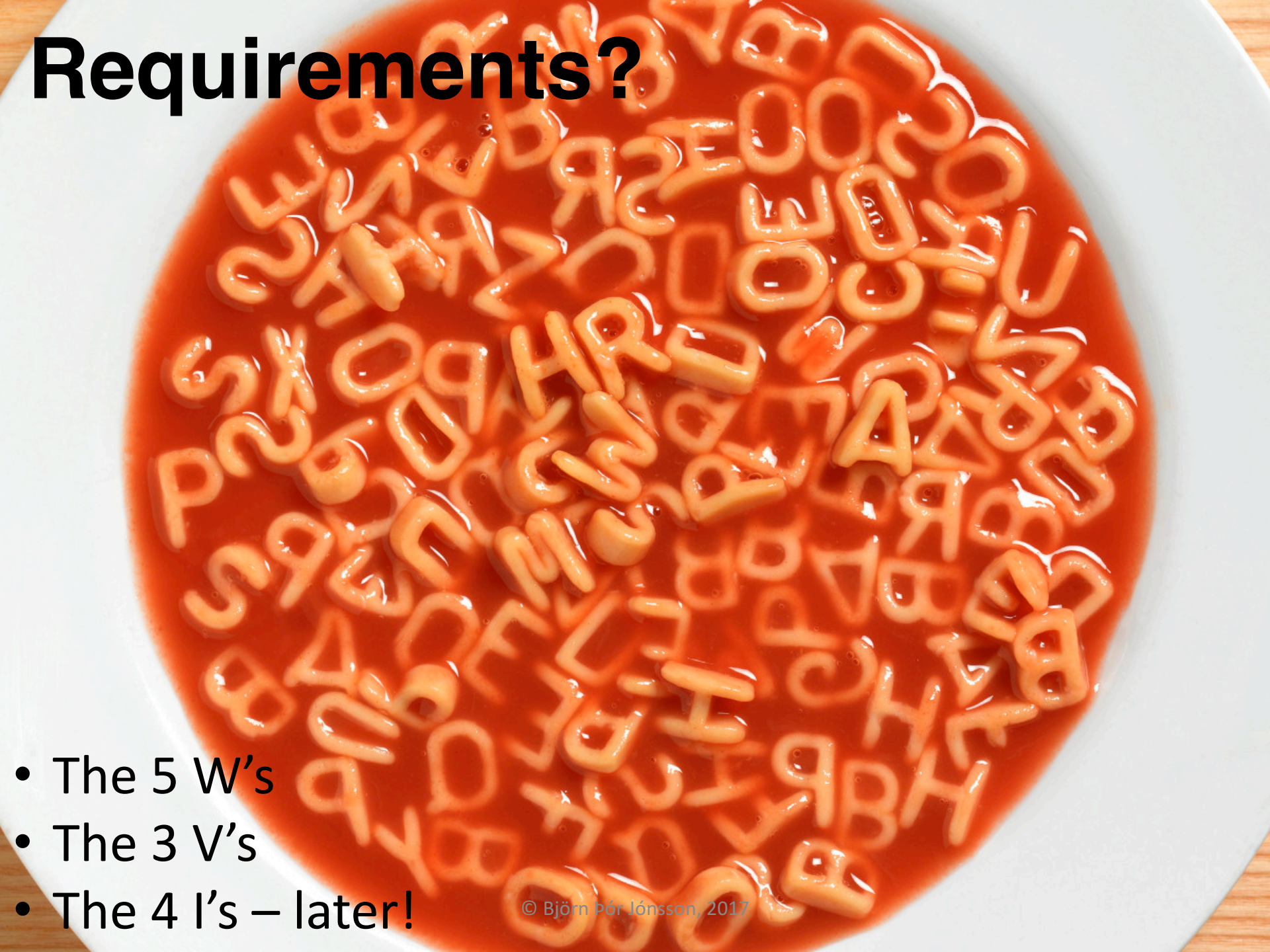
SMALL DATA		BIG DATA
Specific questions	***GOAL***	Broad concerns
One location	***LOCATION***	Many locations
Structured	***STRUCTURE***	Varied, unstructured
Single user	***SOURCE***	Many providers
Transient	***LONGEVITY***	Durable
Focused	***MEASUREMENTS***	Broad
Can be recreated	***REPRODUCIBILITY***	Gone if not captured
Small risk	***STAKES***	Big risk
Simple	***INTROSPECTION***	Metadata is vital
Complete	***ANALYSIS***	Incremental

Source: Jules J. Berman, Big Data Principles

Why Big Data

- “THE MOST COMMON PURPOSE OF BIG DATA IS TO PRODUCE SMALL DATA”
  - Berman, page xxiv

Requirements?



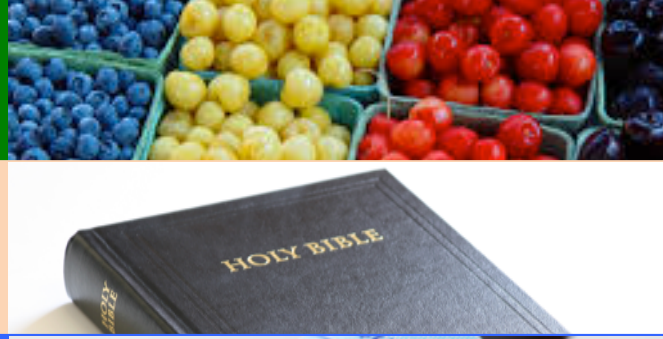
- The 5 W's
- The 3 V's
- The 4 I's – later!

The Three “V”s

Volume

Velocity

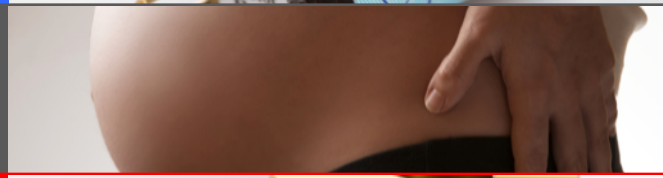
Variety



Veracity

Validity

Viability



Value

The 3 Vs

- Volume, Velocity, Variety: What does it mean?

- At best, the Vs are dimensions to structure:
  - Non-functional requirements
  - Capacity sizing
  - Performance evaluation


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
100+ 40+ Answers.com YAHOO! ANSWERS
QUESTIONS ASKED ON THE INTERNET...
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
Google
Google Search
GO-Globe.com

Big Data Analytics: Making Government Data Work

"Big data" comes with many promises, but the data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data is nothing but a whole lot of dark figures.


CSC

A Briefing from GBC: Industry Insights
November 2012

Making Big Data Work for Government

"Big data" comes with many promises. Not the least of these is access to previously unaccounted information, "dark data," and large volumes of unstructured information. Many estimates place the amount of unstructured data worldwide at 90 percent of all data; and as the average federal agency stores 1.4 petabytes of data,[1] this represents a potentially vast store of previously untapped information for analysts.[2] But big data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data will be nothing but a whole lot of dark figures.

Currently, federal agencies cannot make use of all their data because they do not (or cannot afford to) employ enough data scientists – that is, experts who possess domain knowledge and can use big data technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

"Gartner estimates that big data initiatives in 2013 will total $34 billion," and big data will drive $232 billion in spending over the next five years."

Subduing Big Data

The technological challenges of capturing, securing and managing the worldwide explosion of data are not insignificant (the amount of worldwide data currently measures about 2.7 zettabytes and is projected to double every two years[3]. Yet, corralling big data is as much an organizational challenge as a technological one. Industry development of big data technologies has largely bifurcated into two main functions: data discovery and the more traditional business intelligence.[4] Companies have been investing much time and money in the former.[5] Gartner estimates that big data initiatives in 2013 will total $34 billion,[6] and big data will drive $232 billion in spending over the next five years.[7]

Many organizations are trying to address the technology challenge of analyzing big data. The Technology Foundation's Big Data Commission, for example, recently released a report that gives organizations guidelines for an effective big data program. Examination of these steps shows a decided focus not just on technical capabilities, but on organizational policies as well. The Commission recommends that CIOs take a "holistic approach to help guide the agency from an information management perspective and follow these key steps:

- Identify data and content that are critical to its mission
- Identify how, when, where and to whom information should be made available
- Determine appropriate data management, governance and security practices
- Identify and prioritize the information projects that deliver the most value.[8]

CSC

Government Industry Insights

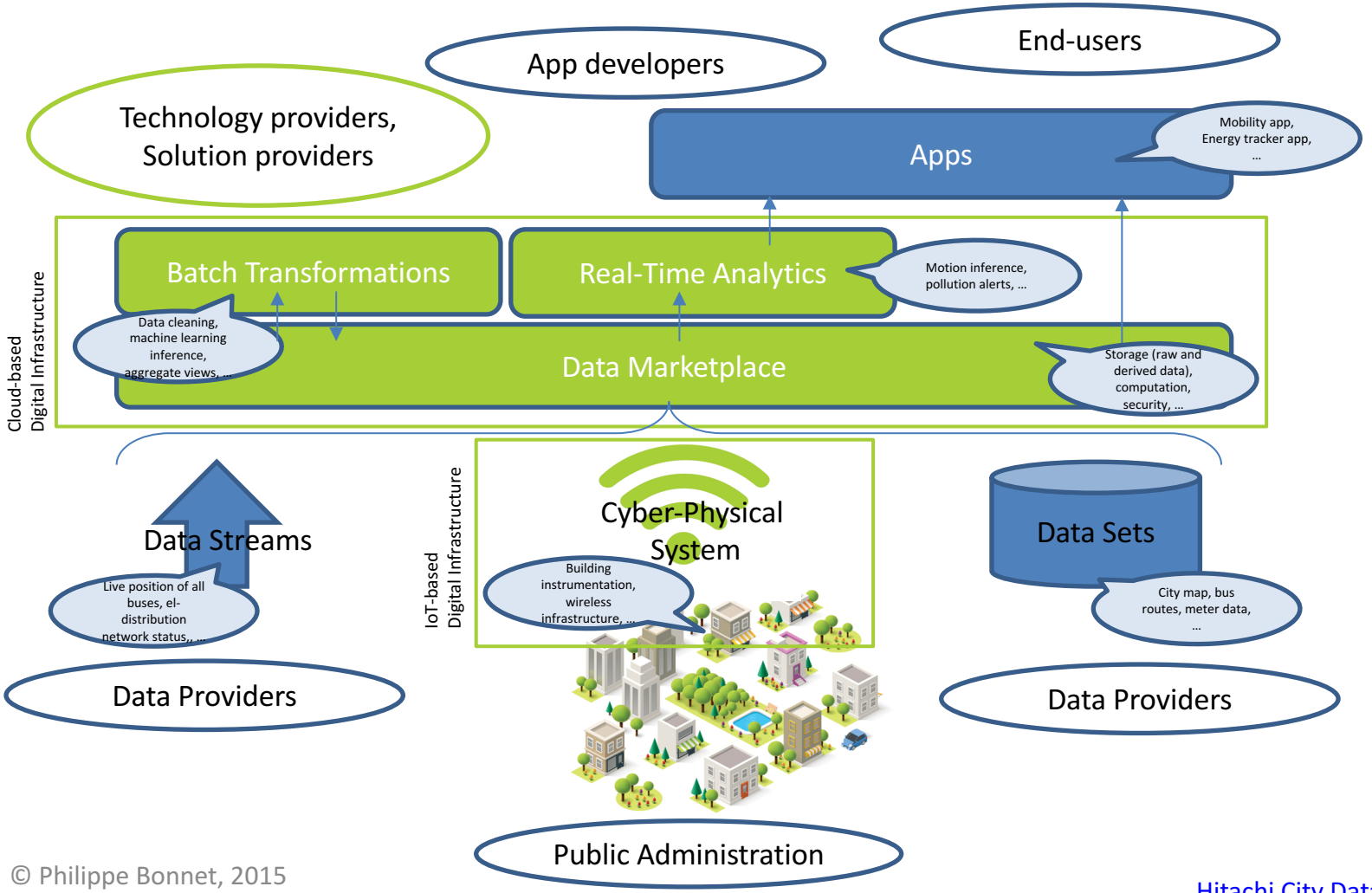
Government Big Data

Currently, federal agencies cannot make use of all their government data because they do not (or cannot afford to) employ enough data scientists—that is, experts who possess domain knowledge and can use government big data analytic technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

Download the Big Data brief (PDF, 267KB)

Contact Us

Big Data at City Scale



Hitachi City Data Market

Big Data at Building Scale

Light on/off events not logged

Wireless Router Monitoring

IT Dept

Building Management System

Facility Management

Learnit Log

Admin
Teachers
Students?

Big Data at Building Scale

Light on/off events

Wireless Router Monitoring

Building Management System

Learnit Log

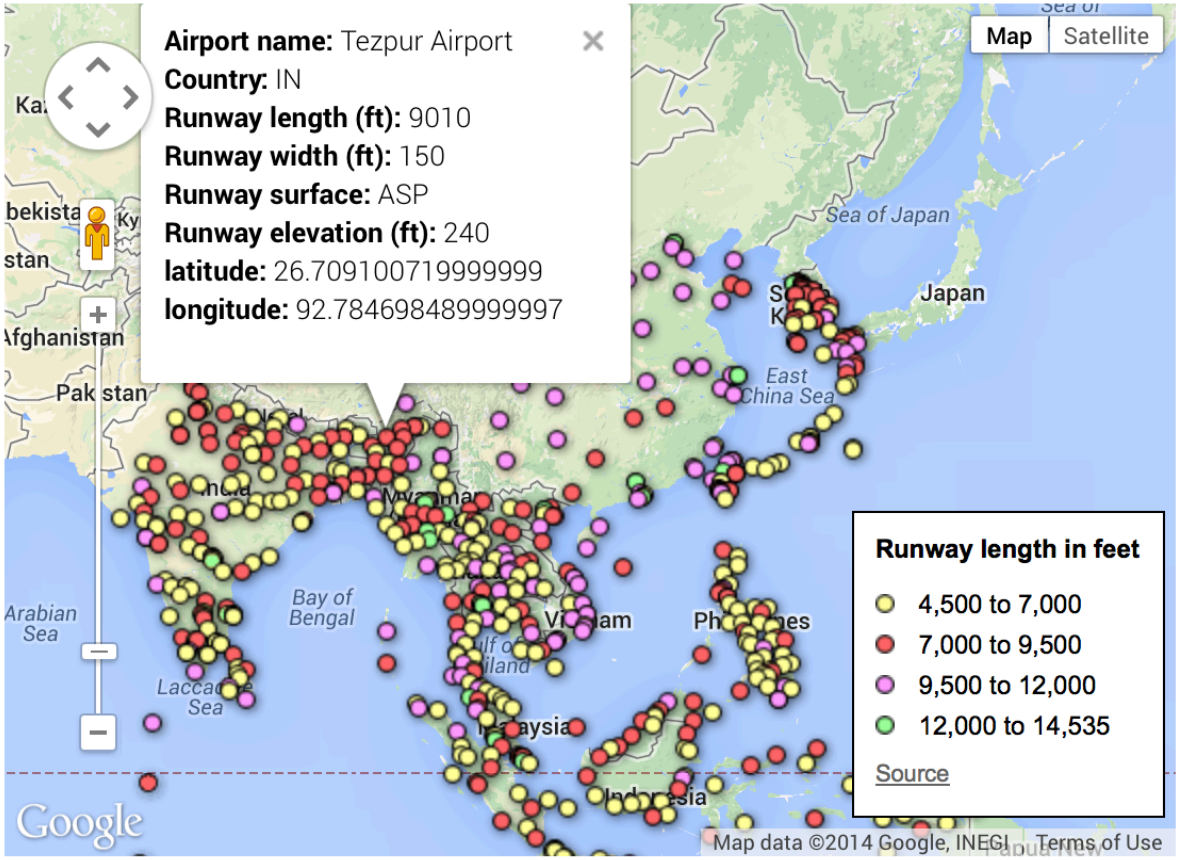
EF

FM

Mgt

ITU Big Data Resources available for Analytics within and outside the IT University

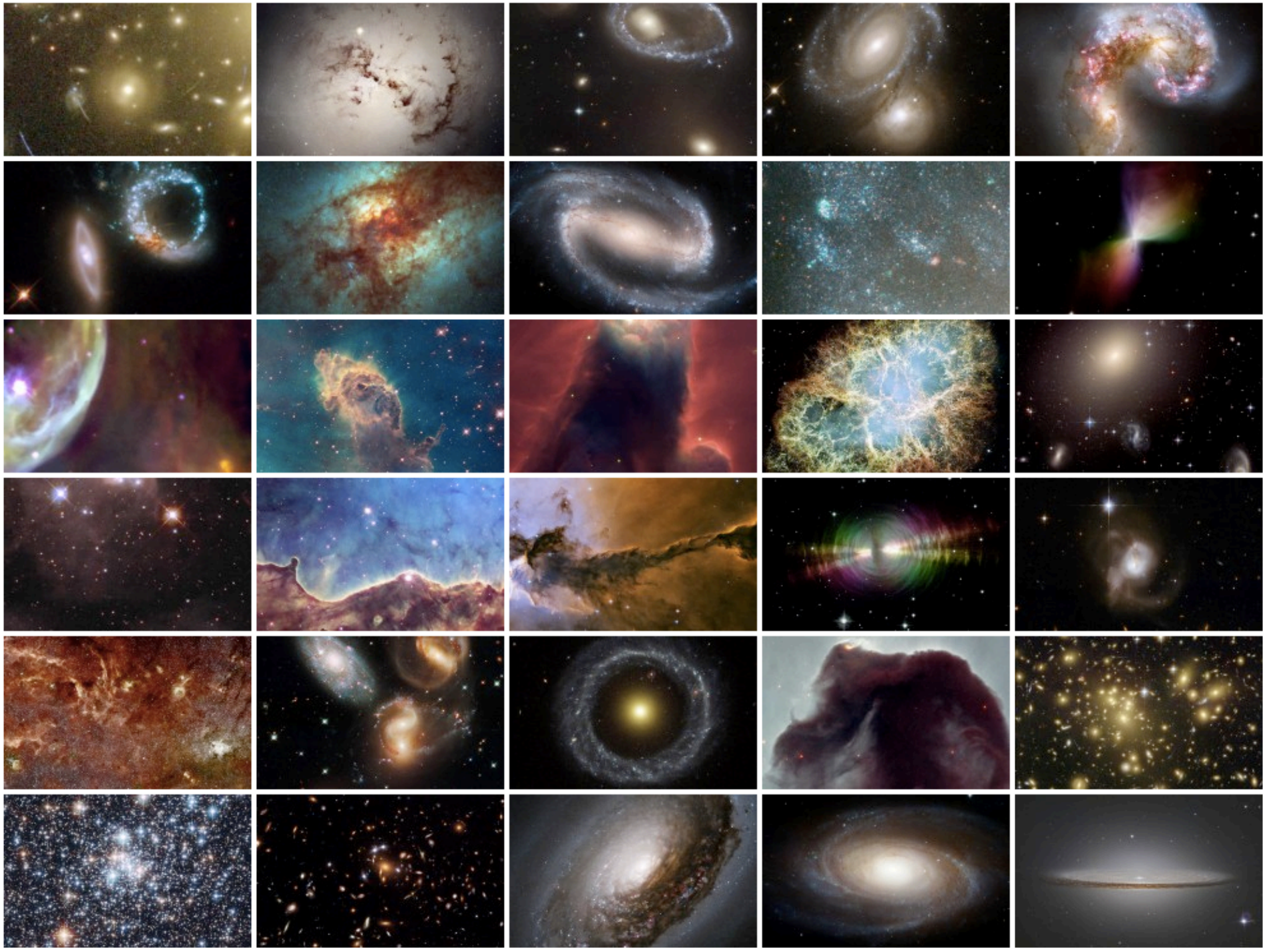
Search and Rescue Example



Map by David Strip based on Google maps and OurAirports.com

See Jame Fallows'article at the Atlantic.


© Björn Þór Jónsson, 2017



The Five “W”s

What?

Who?

Where?

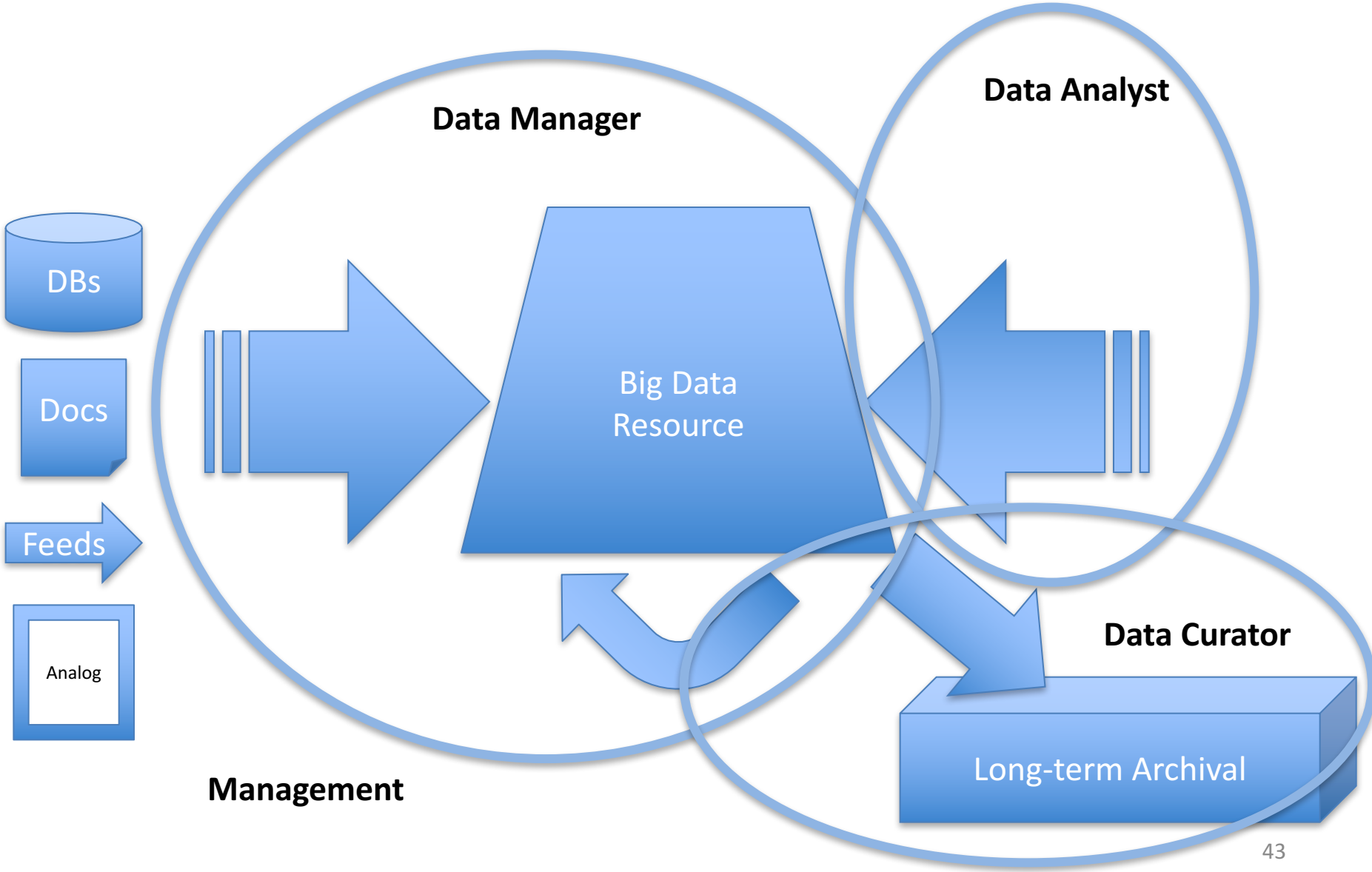
When?

Why?

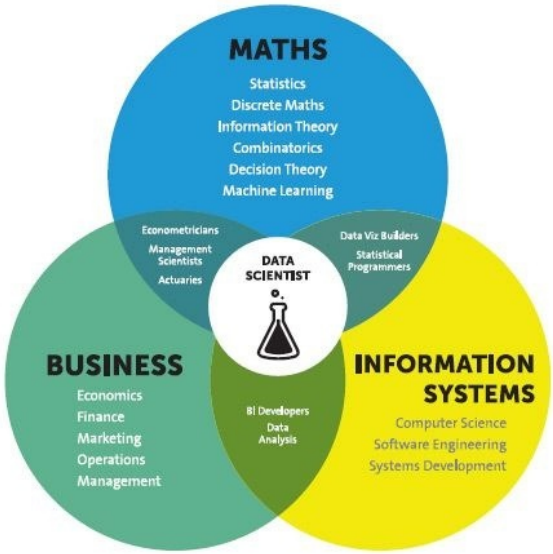
When does Big Data Make Sense?

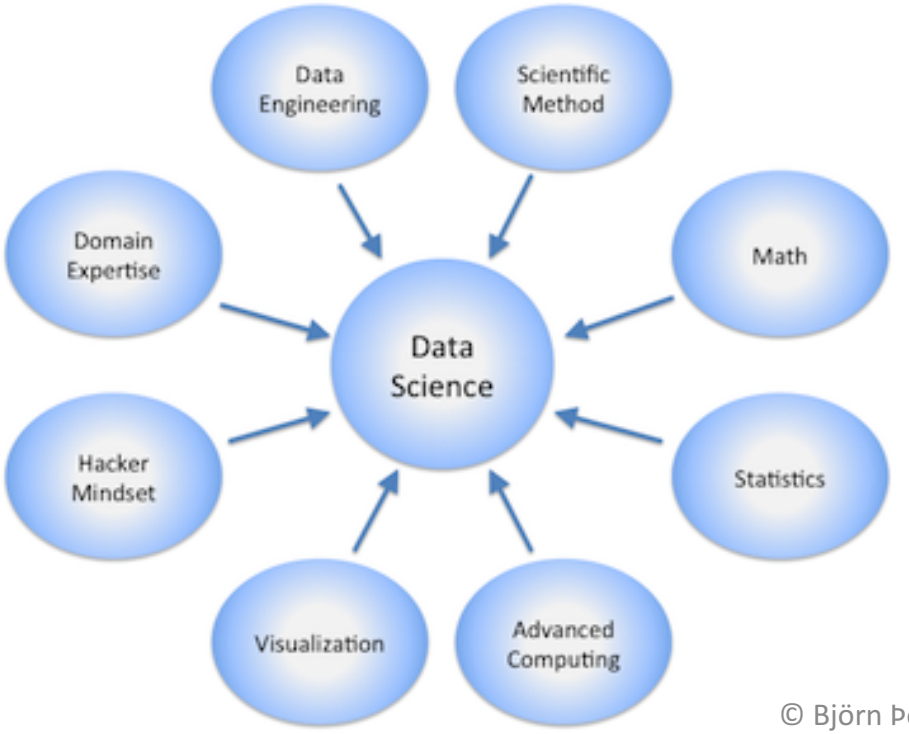
- Need for insight based on data not currently available for analysis

Who is involved?



Data Science





Outline

- Trends Underlying Big Data
- What is Big Data?
- **Data Models**
  - Concepts
  - Naming/Identification/Ontologies
  - Structured/Semi-structured/Unstructured
- Course Outline

Conceptual Model

- Entities
  - An entity is a “thing” or object
  - An entitiy set is a collection of similar entities
  - An attribute is a property of an entity
- Relationships
  - A relationship connects two or more entity sets
  - Relationship sets are collection of similar relationships
  - An attribute is a property of a relationship

Logical Data Model

1. Mathematical representation of data.
   - Examples: set/multiset (relations); trees/graphs (linked data).
2. Operations on data.
   - Example: Select all items whose acceleration, defined as (sqrt(x^2+y^2+z^2)) is greater than 1,5
3. Constraints
   - Example: All valid items should have an acceleration, defined as (sqrt(x^2+y^2+z^2)) greater than 1,5

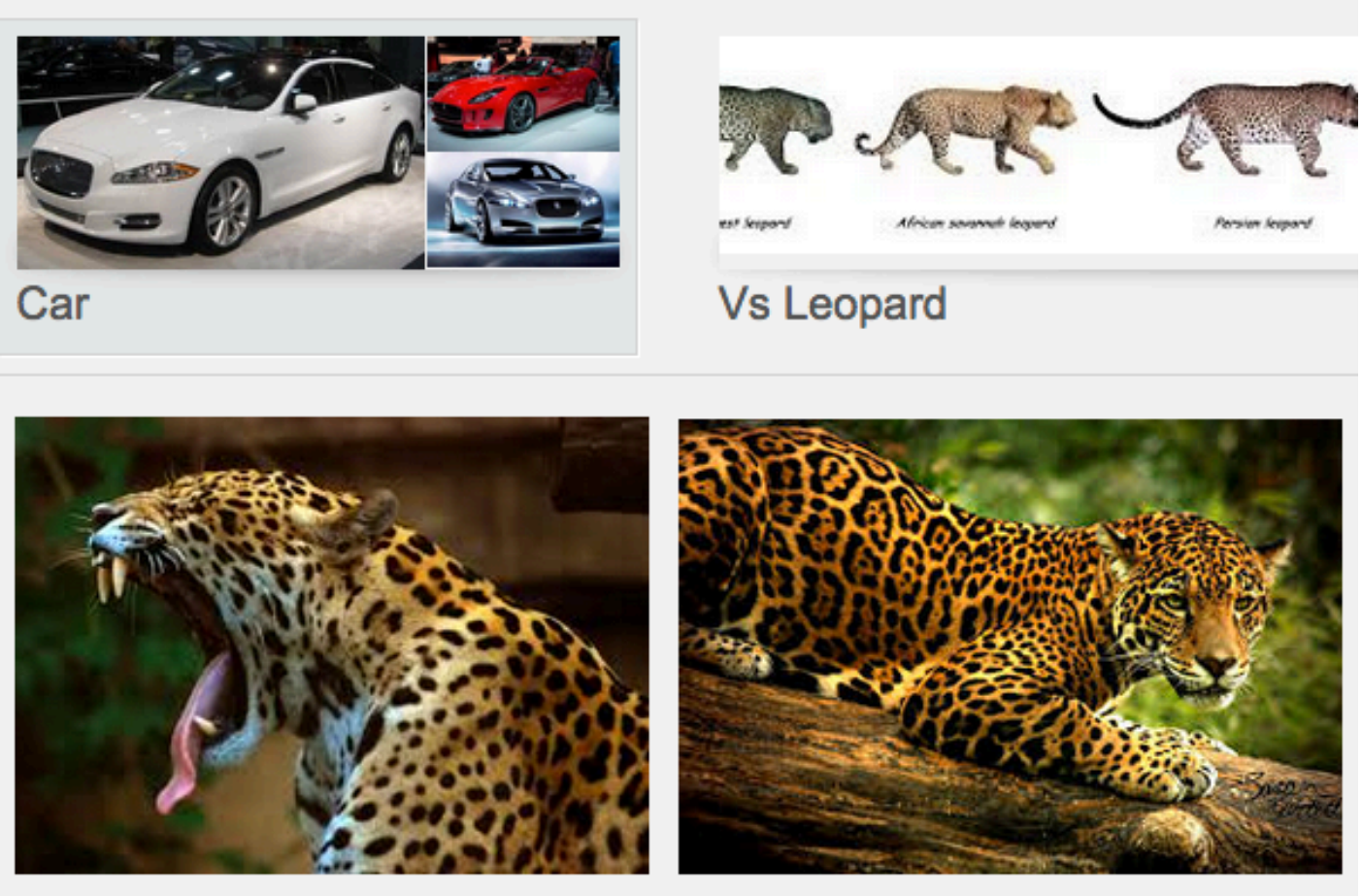
Physical Data Model

- Relational databases
  - SQL tables; indices; partitionning
  - Row or column store mapped onto files
- NoSQL systems
  - Columnar representation of nested data structures
    - Parquet
  - Serialisation/deserialisation (SerDe) of linked data structures onto files
    - Avro, thrift, protocol buffers
    - Includes compression

What is Hard?

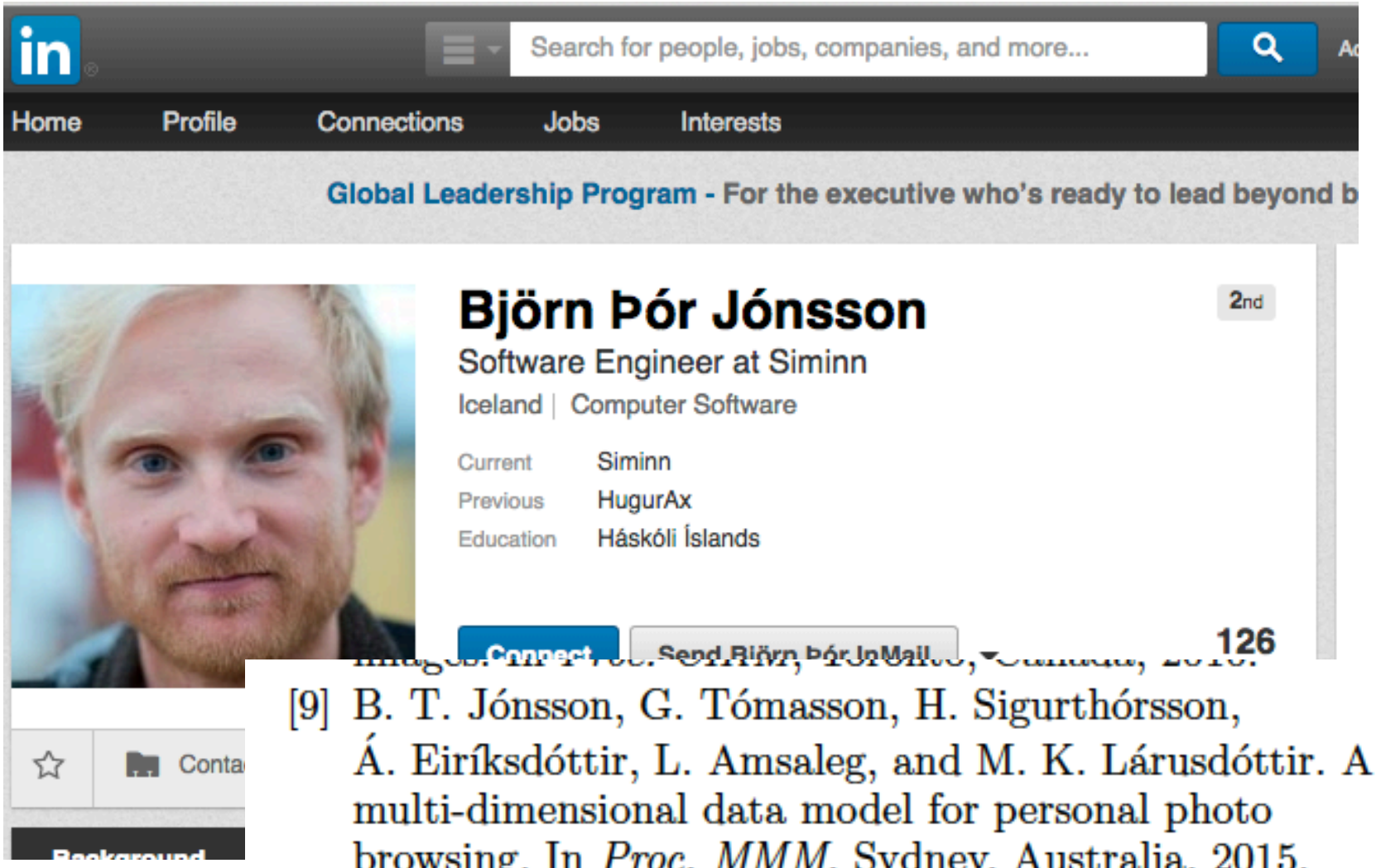
- Scoping
  - What is an entity, what is an attribute?
- Naming
  - How to name entities and properties
  - Identification
- Dealing with change
  - Integration across data sources: ontologies
  - In time: Immutability

What's in a name?



https://www.google.com/search?site=&tbm=isch&source=hp&biw=1440&bih=802&q=jaguar&oq=jaguar&gs_l=img.3..0l10.1339.2325.0.2503.6.4.0.2.2.0.108.298.1j2.3.0.msedr...0...1ac.1.61.img..1.5.307.a-kqsMXizB4

What's in a name?



orcid.org

Search

English

ORCID

Connecting Research and Researchers

FOR RESEARCHERS | FOR ORGANIZATIONS | ABOUT | HELP | SIGN IN

DISTINGUISH YOURSELF IN THREE EASY STEPS

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized. Find out more.

1 **REGISTER** Get your unique ORCID identifier Register now! Registration takes 30 seconds.

2 **ADD YOUR INFO** Enhance your ORCID record with your professional information and link to your other identifiers (such as Scopus or ResearcherID or LinkedIn).

3 **USE YOUR ORCID ID** Include your ORCID identifier on your Webpage, when you submit publications, apply for grants, and in any research workflow to ensure you get credit for your work.

LATEST NEWS

Tue 01/13/2015
New webinar: The metadata round trip

Mon 01/12/2015
ORCID Partners with Hypothes.is and NIF on Helmsley Trust-Supported Open Annotation Project

Thu 01/08/2015

1930's - 1950's

1930's - 1950's, Private Vehicle

1930's - 1950's

1950 - 1958, Private Vehicle
Jim Fox/Picture by Greg Gibson

1958 - 1967, Private Vehicle

1976 Series, Private Vehicle

2008 Series, Private Vehicle

Civil Defense (1)

1953 Series, Navy

1953 Series, Army

1976 Series, Army

1953 Series, Airforce

Airforce Home Guard (1)

2008, Unified Defense Force (2)

Before 2012, N.A.T.O. Headquarters (77000 - 77999 Block)

1930's, Motorcycle Dealer

1950's. Dealer (1)

1950's, Commercial, Partially Tax Exempt (*)

1950's, Temporary

1965, Temporary

2003, Temporary

Temporary Test Plate

Driver Education

Mid-1990's, Airport Vehicle

1970's, Trailer

Diplomatic Corps (1)

Vehicle Paying Intermediate Tax Rate

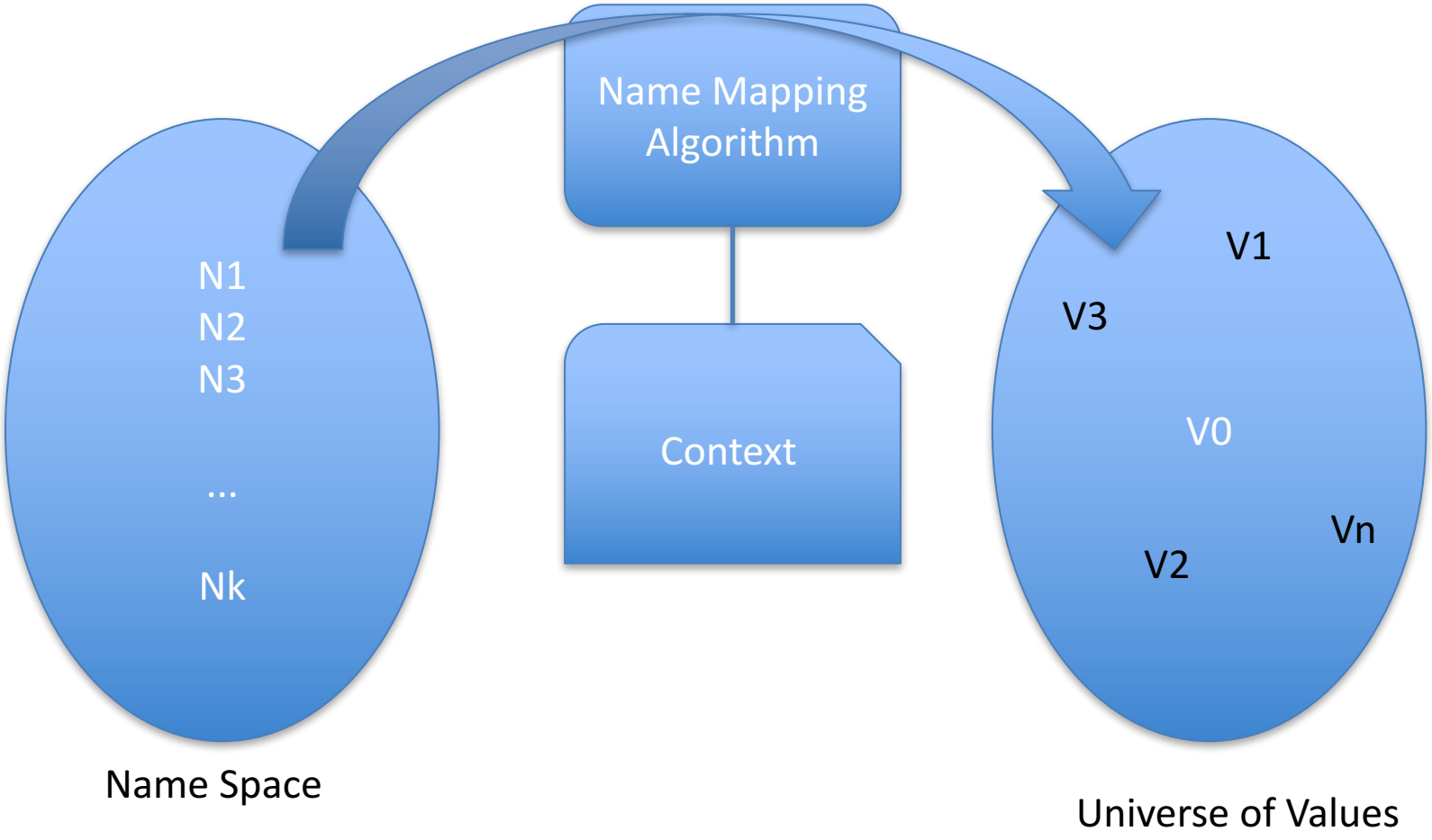
Emergency Services
Ambulance

Emergency Services

Emergency Services
Copenhagen Firebrigade (1)

http://www.worldlicenseplates.com/world/EU_DENM.html

Name Mapping



Examples (Database)

- Key constraint
  - Name space: key attribute values
  - Universe of values: tuples in a table
  - Name mapping algorithm: Index
  - Context: A given table
- Sequence Id
  - Name space: sequence of integers
  - Universe of values: tuples in a table
  - Name mapping algorithm: Index
  - Context: A given table

Example (File System)

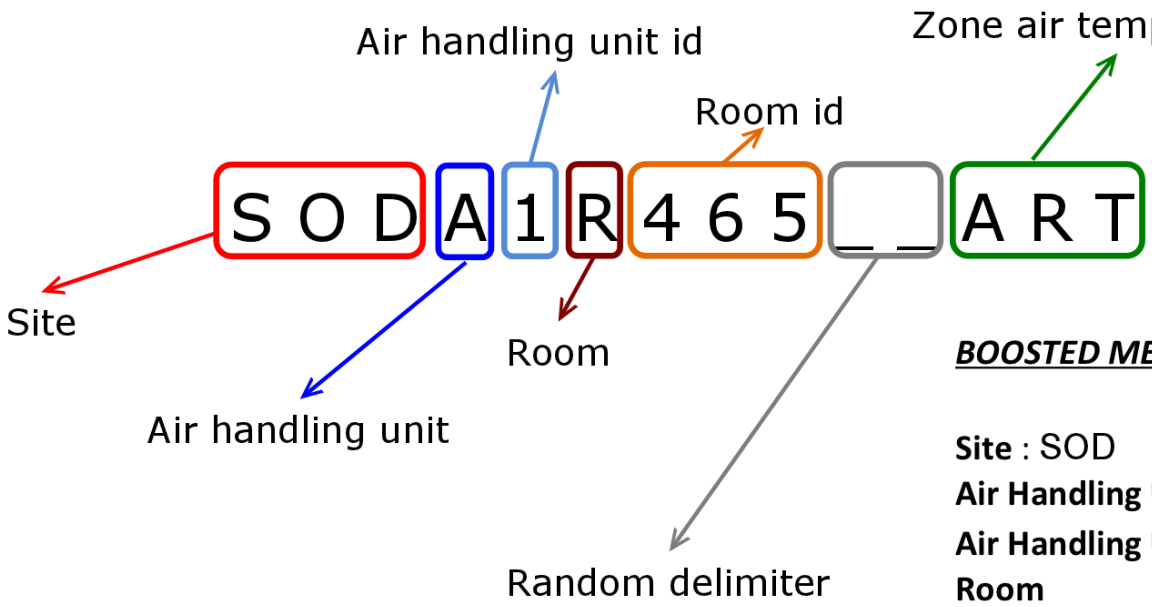
- File System Name
  - Name space: Path name
  - Universe of values: File Identifiers
  - Name mapping algorithm: recursive mapping
  - Context: Root directory

Example (Unix File System)

Layer	Names	Values	Context	Name-mapping algorithm	
Symbolic link	Path names	Path names	The directory hierarchy	PATHNAME_TO_GENERAL_PATH	↑
Absolute path name	Absolute path names	Inode numbers	The root directory	GENERALPATH_TO_INODE_NUMBER	user-oriented names
Path name	Relative path names	Inode numbers	The working directory	PATH_TO_INODE_NUMBER	↓
File name	File names	Inode numbers	A directory	NAME_TO_INODE_NUMBER	machine-user interface
Inode number	Inode numbers	Inodes	The inode table	INODE_NUMBER_TO_INODE	↑
File	Index numbers	Block numbers	An inode	INDEX_TO_BLOCK_NUMBER	machine-oriented names
Block	Block numbers	Blocks	The disk drive	BLOCK_NUMBER_TO_BLOCK	↓

Figure from Saltzer and Kaashoek

Logical Identifiers



BOOSTED METADATA:

Site : SOD
Air Handling Unit
Air Handling Unit id : 1
Room
Room id : 465
Zone air temp sensor : ART

Slide courtesy of Arkha Batthacharya

Identifiers

Properties

- Size – possible values
  - N
  - Ty
- Uniq
  - D
- Ranc
  - N
  - U
  - D

Generation

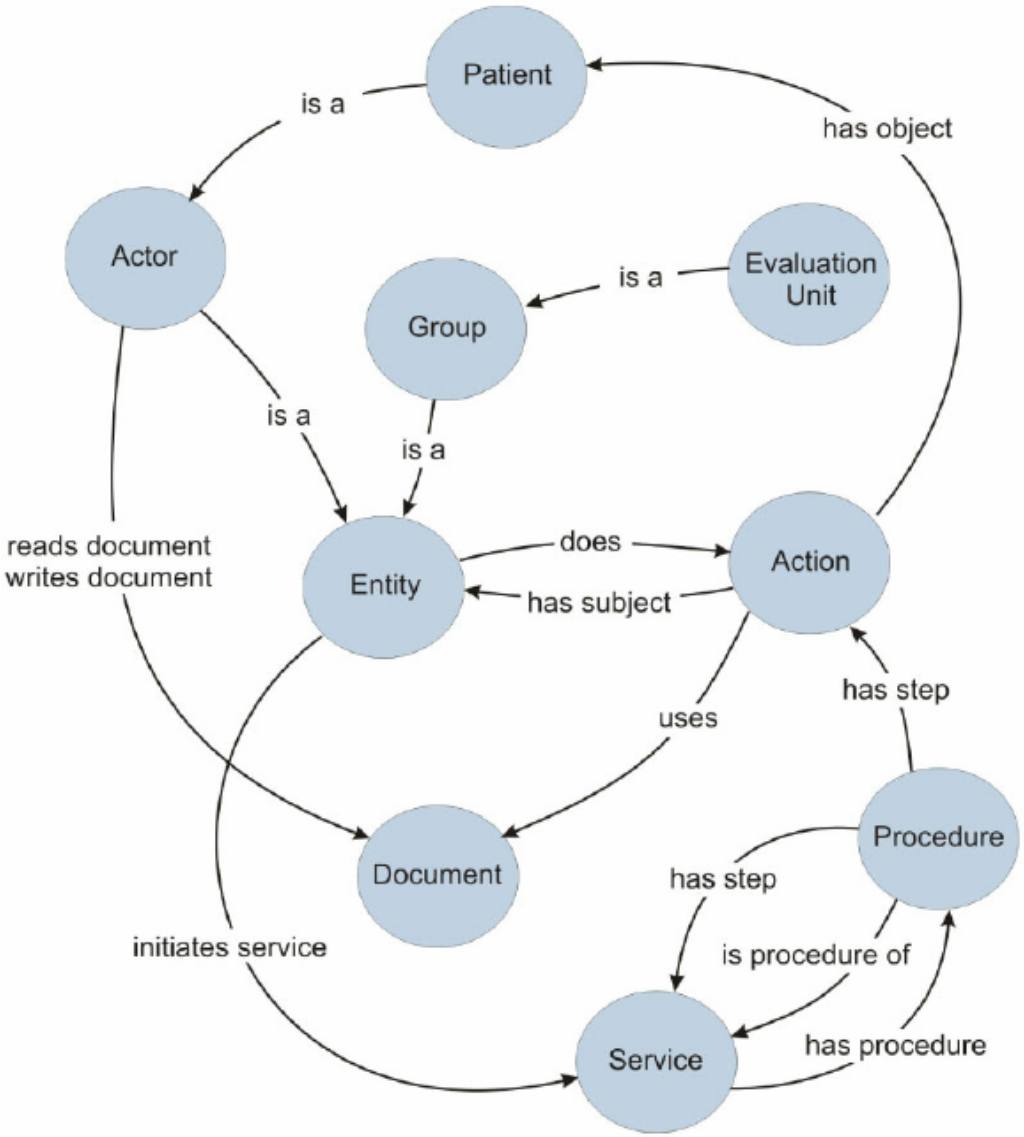
- 1 place of generation

http://becuo.com/despicable-me-minions-whaaaat

Ontology

- Facts:
  - Concepts (entity sets) and individuals (entities)
    - attributes or properties
      - changed by some event
- Axioms:
  - Named relationships between concepts
    - Hierarchy
    - Direct Acyclic Graph
    - Arbitrary Graph
- Rules:
  - if-then rules to derive new facts

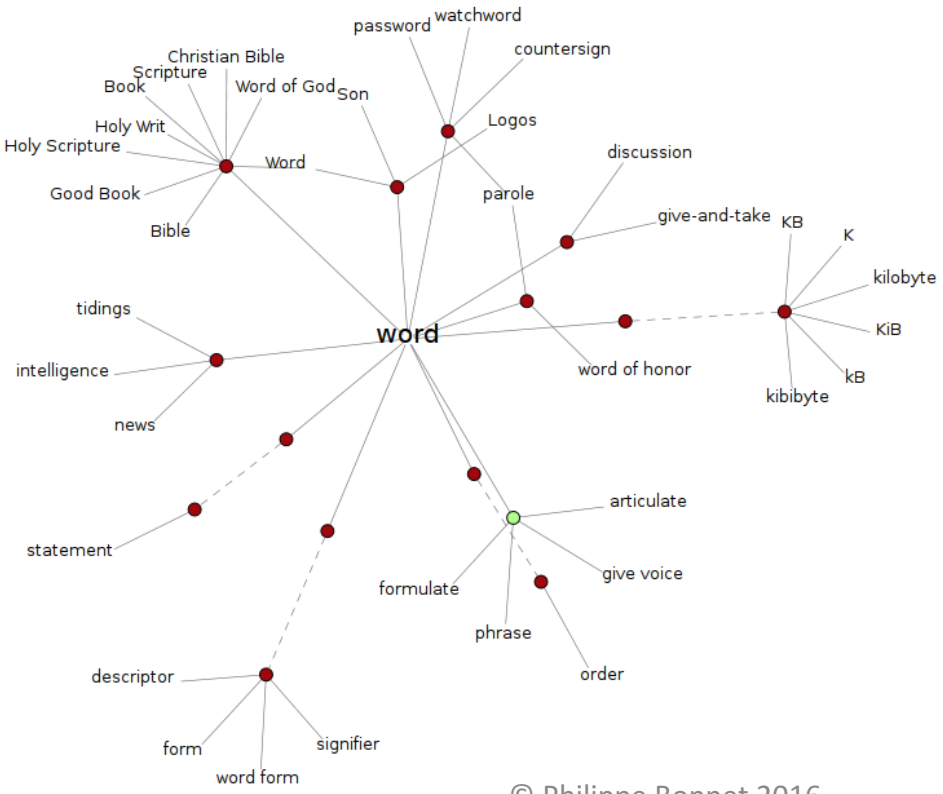
Healthcare Ontology: K4Care



http://wordnet.princeton.edu/

Wordnet

- Synset (synonym ring)
  - Group of entities that are semantically equivalent



Structured Data

- Instance and schema are clearly separated
  - relational, object-oriented

Data part:
(#123, [{[“John”, s111111111, [123,”Main St”]],
[“Joe”, s222222222, [321, “Pine St”]] }
])

Schema part:
PersonList[*ListName*: String,
Contents: [*Name*: String,
Id: String,
Address: [*Number*: Integer, *Street*: String]]
]

Semi-Structured Data

- *Self-describing*:
  - Attribute names embedded in the data itself, *but are distinguished* from values
  - Doesn't need schema to figure out what is what (but schema might be useful nonetheless)

(#12345,
[*ListName*: "Students",
Contents: { [*Name*: "John Doe",
Id: "s111111111",
Address: [*Number*: 123, *Street*: "Main St."]] ,
[*Name*: "Joe Public",
Id: "s222222222",
Address: [*Number*: 321, *Street*: "Pine St."]] }
])

XML (https://www.w3.org/XML/)

- Must have a *root element*
- Every *opening tag* must have matching *closing tag*
- Elements must be *properly nested*

  <foo><bar></foo></bar> is a no-no
- An *attribute* name can occur *at most once* in an opening tag. If it occurs,
  - It *must have an explicitly specified value* (Boolean attrs, like in HTML, are not allowed)
  - The value *must be quoted* (with " or ')
- *XML processors are not supposed to try and fix ill-formed documents (unlike HTML browsers)*

JSON (http://json.org)

- A JSON object is an unordered collection of name-value pairs.
  - Each name is followed by :
  - A value is a string (in between quotes), a number, true/false, null, an object or an array; values can be nested
  - Name-value pairs are separated by ,
  - An object starts with { and ends with }
- A JSON array is an ordered collection of values
  - An array starts with [and ends with]


```
{
 "/sensor0" : {
  "Metadata" : {
   "SourceName" : "Test Source",
   "Location" : { "City" : "Berkeley" }
  },
  "Properties": {
   "Timezone": "America/Los_Angeles",
   "UnitofMeasure": "Watt",
   "ReadingType": "double"
  },
  "Readings" : [[1351043674000, 0], [1351043675000, 1]],
  "uuid" : "d24325e6-1d7d-11e2-ad69-a7c2fa8dba61"
 }
}
```

Representing Documents

- BLOB: unstructured
  - Need to extract structure for further processing
  - Term extraction, NPL
- Entity with text attributes
  - Structured data
  - Semi-structured data

Representing Media

- BLOB: unstructured
  - Need to extract structure for further processing
  - Concept extraction, deep learning
- Entity with text attributes
  - Structured data
  - Semi-structured data
  - Always need the BLOB!?

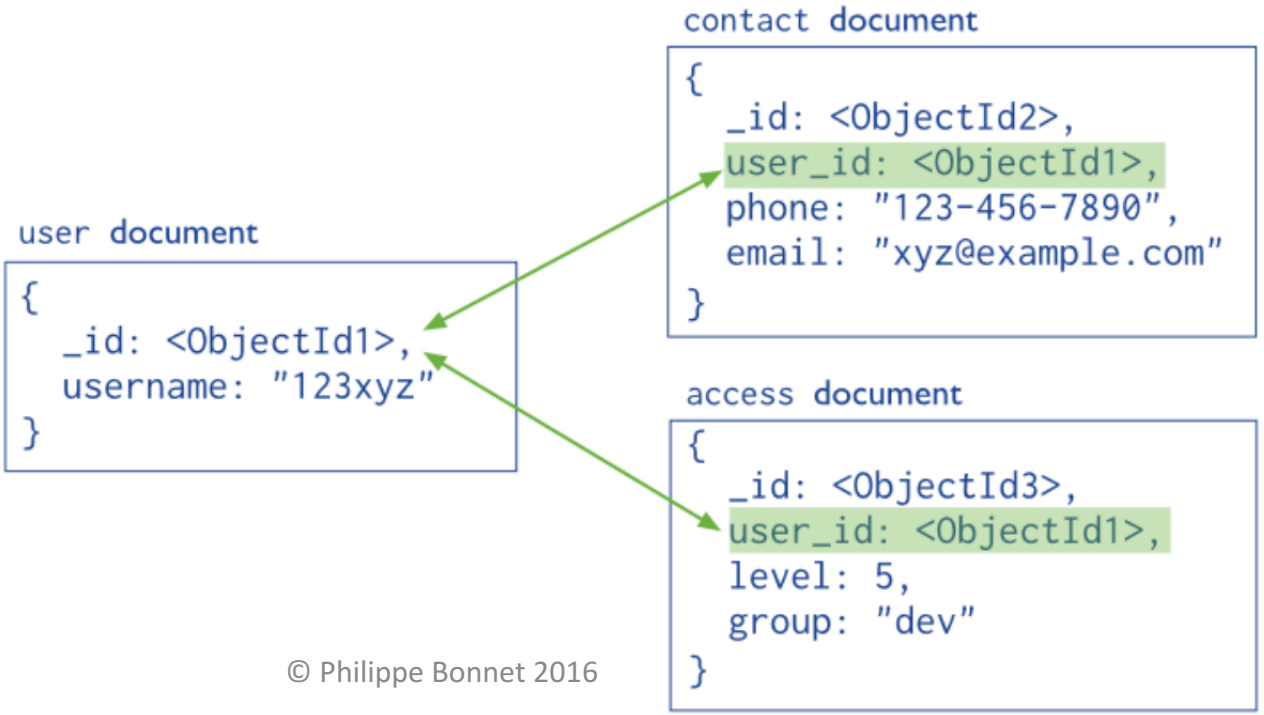
Linked Data (Tree, Graph) in SQL

- Adjacency List R(N1, N2)
- Denormalization R(N1, ListOfPaths)
- Preorder traversal R(N, Lft, Rgt)

https://www.mongodb.com

Linked Data in MongoDB

- Entities represented as Json objects, called documents
  - Identification through _id attribute
- Relationships represented as references



Linked Data in Key-Value Stores

- Entities and relationships represented as associative arrays (or maps)
  - (Ordered) collection of (key,value) pairs
  - Each key occurs only once in the collection
  - Interface
    - add/remove (key,value) pairs
    - modify value for a given key (or a range of keys)
    - Look up a value with a given key
- Systems:
  - Cassandra, LevelDB, RocksDB, Aerospike ...

Linked Data in Big Table

- A Bigtable is a sparse, distributed, persistent multidimensional sorted map.
  - The map is indexed by a dimension key, and a timestamp; each value in the map is an uninterpreted array of bytes.
- Well-suited for
  - Relationships that can be denormalized onto a multidimensional space

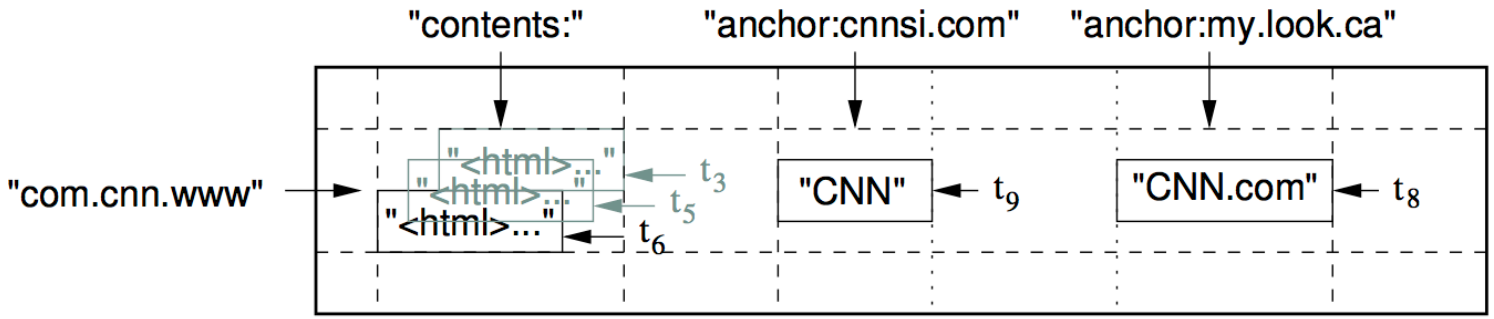
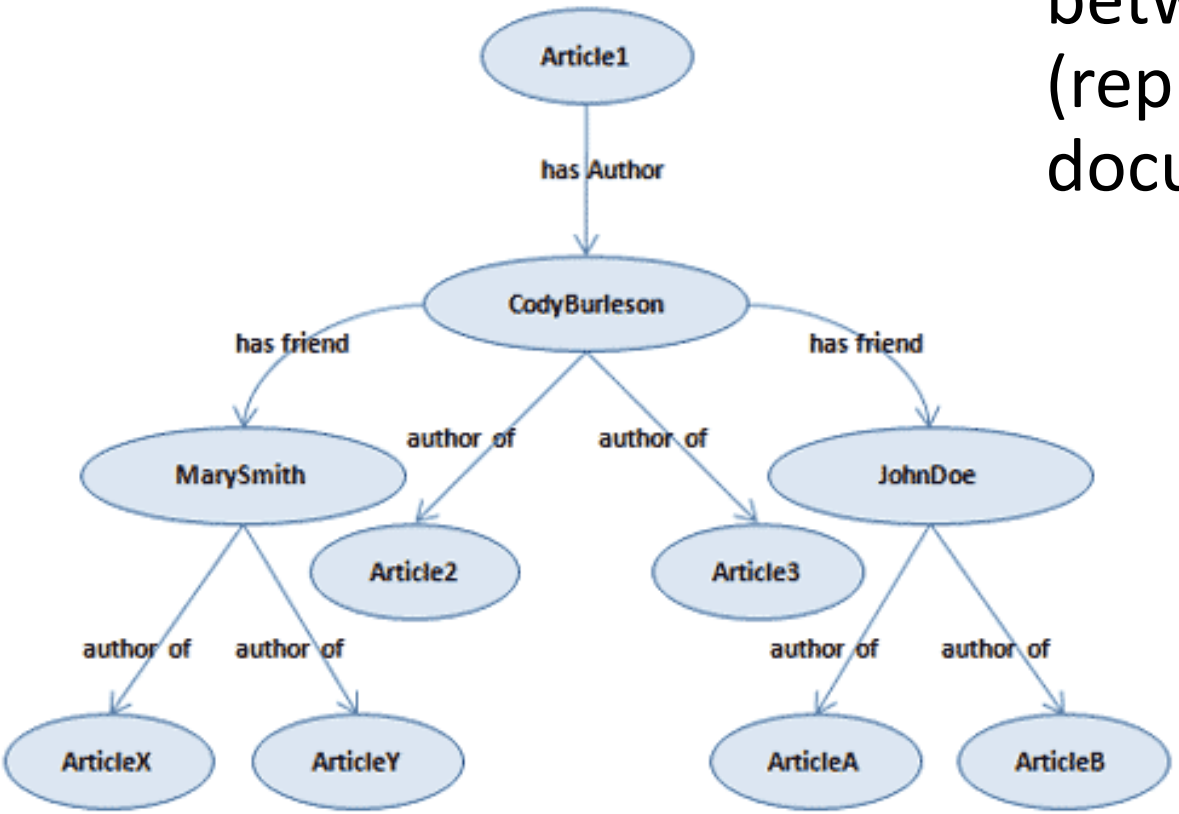


Figure 1: A slice of an example table that stores Web pages. The row name is a reversed URL. The contents column family contains the page contents, and the anchor column family contains the text of any anchors that reference the page. CNN's home page is referenced by both the Sports Illustrated and the MY-look home pages, so the row contains columns named anchor:cnnsi.com and anchor:my.look.ca. Each anchor cell has one version; the contents column has three versions, at timestamps t_3, t_5, and t_6.

http://www.w3.org/RDF/

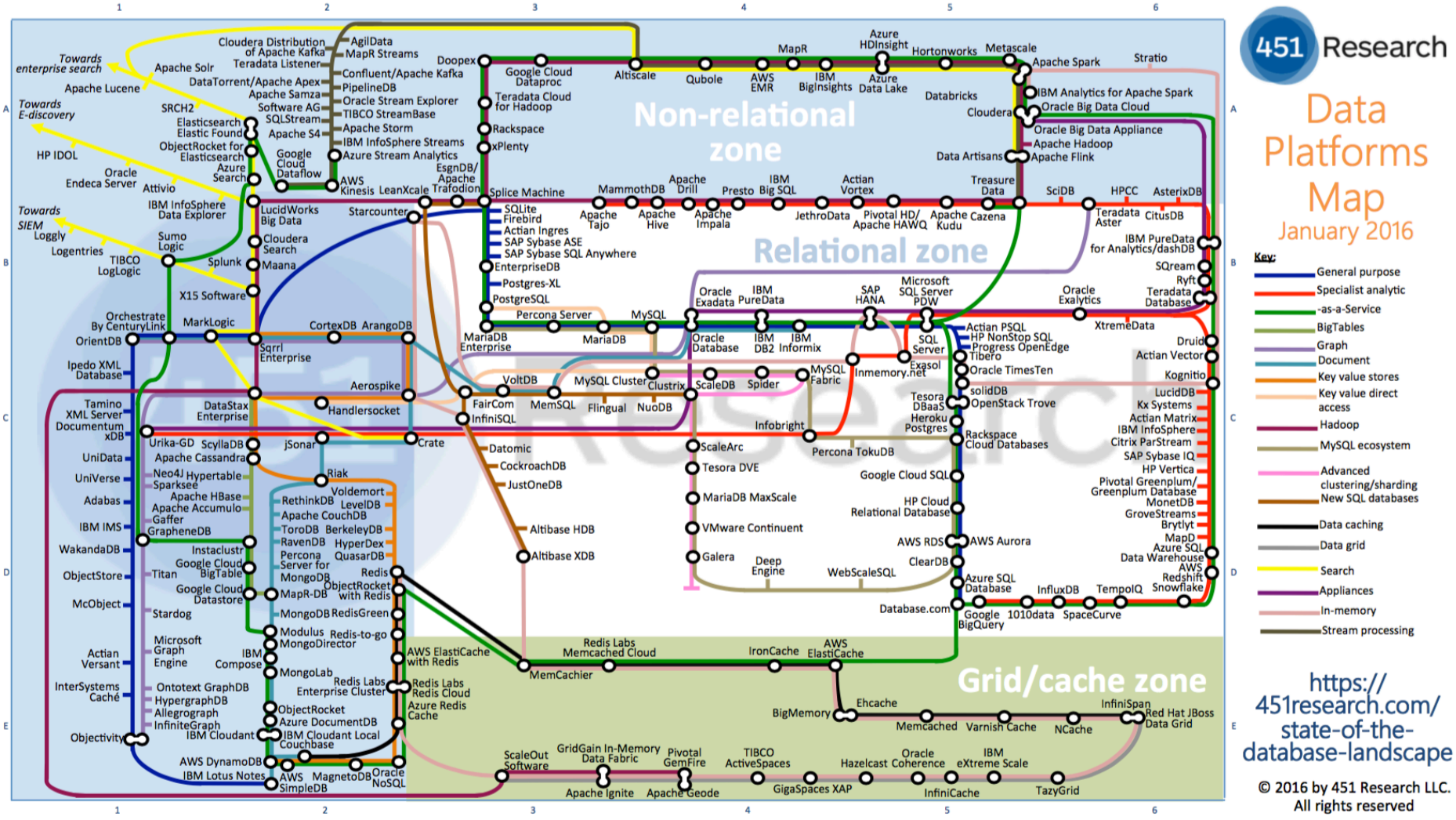
Linked Data in RDF

Directed, labeled graph, where the edges represent relationships between two entities (represented as XML documents).



Linked Data as Graphs

- Graphs can be represented with adjacency lists
  - Indexed adjacency list (relations)
    - Stored as a map or a relation
  - Index-free adjacency list (graphs)
    - Pointers from one node to the other
    - Data structure underlying graph databases
- Systems: Neo4J, GraphX (Spark), GraphLab, …



451 Research

Data Platforms Map

January 2016

Key:

- General purpose
- Specialist analytic
- -as-a-Service
- BigTables
- Graph
- Document
- Key value stores
- Key value direct access
- Hadoop
- MySQL ecosystem
- Advanced clustering/sharding
- New SQL databases
- Data caching
- Data grid
- Search
- Appliances
- In-memory
- Stream processing

https://451research.com/state-of-the-database-landscape

Metadata

Metadata is 'data about data' and can be divided into four basic classifications:

- **Business Metadata** - the business meaning of data. It includes business definitions of the objects and metrics, hierarchies, business rules, and aggregation rules.
- **Operational Metadata** - Operational metadata stores information about who accessed what and when.
- **Technical Metadata** - Technical Metadata describes the data structures and formats such as table types, data types, indexes, and partitioning method.
- **Process Metadata** - Process Metadata describes the data input process.

Typical Use Cases

- **Introspection** - discovering and harvesting of information into a metadata repository
- **Impact Analysis** - analyse dependencies across the architecture. Requires end to end view of interdependencies
- **Data Lineage** - description of the origins of a piece of data and the proves by which it arrived in the database. The 'provenance' or 'pedigree.'

https://blogs.oracle.com/IanT/entry/is_metadata_important

Metadata

- Only if every user has a common and exact understanding of the data can it be exchanged trouble-free.
  - ISO/IEC 11179 Metadata Registry Specification

- “Microsoft says nobody uses metadata!”
  - https://social.technet.microsoft.com/Forums/en-US/1fb1adc5-b7ec-48e3-bb1a-ebeff70e4c0e/microsoft-says-nobody-used-metadata-so-we-don?forum=sharepointgeneralprevious

Why Metadata?

Small Data		Big Data
• Specific questions	*GOAL*	• Broad concerns
• One location	*LOCATION*	• **Many locations**
• Structured	*STRUCTURE*	• **Varied, unstructured**
• Single user	*SOURCE*	• **Many providers**
• Transient	*LONGEVITY*	• **Durable**
• Focused	*MEASUREMENTS*	• Broad
• Can be recreated	*REPRODUCIBILITY*	• Gone if not captured
• Small risk	*STAKES*	• Big risk
• Simple	*INTROSPECTION*	• **Metadata is vital**
• Complete	*ANALYSIS*	• Incremental

Take Away Points

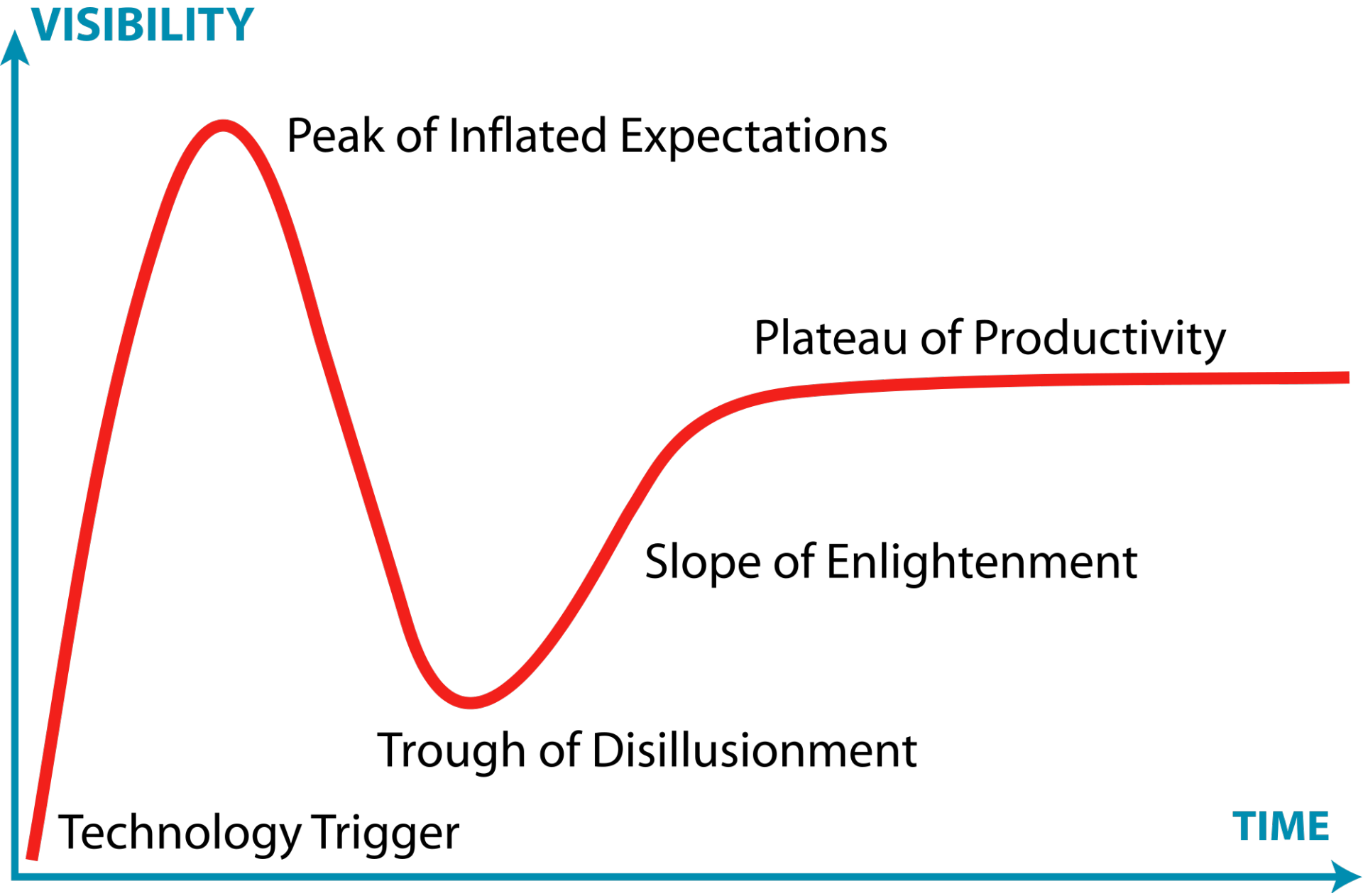
- Conceptual model based on entities and relationships
  - Focus on naming, identification, versioning (changes in time) and ontologies (integration)
- Structured data maps ER onto relations
- Semi-structured data maps ER onto linked data represented with MongodDB, KV store, BigTable, RDF, Graphs

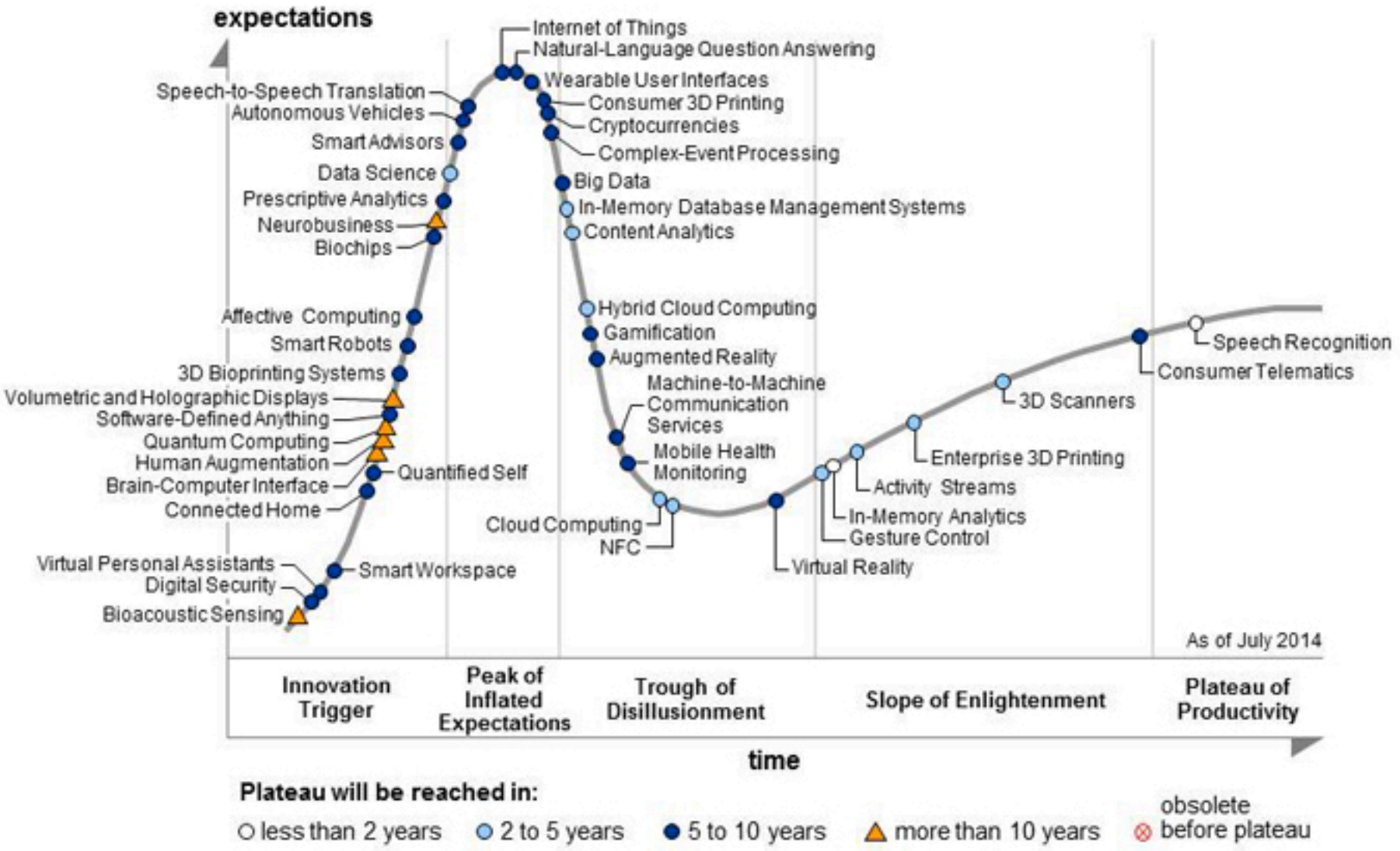
Big Data: Who's doing it?

Gartner: "64% of enterprises surveyed indicate that they're deploying or planning Big Data projects. Yet even more acknowledge that they still don't know what to do with Big Data."

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it"

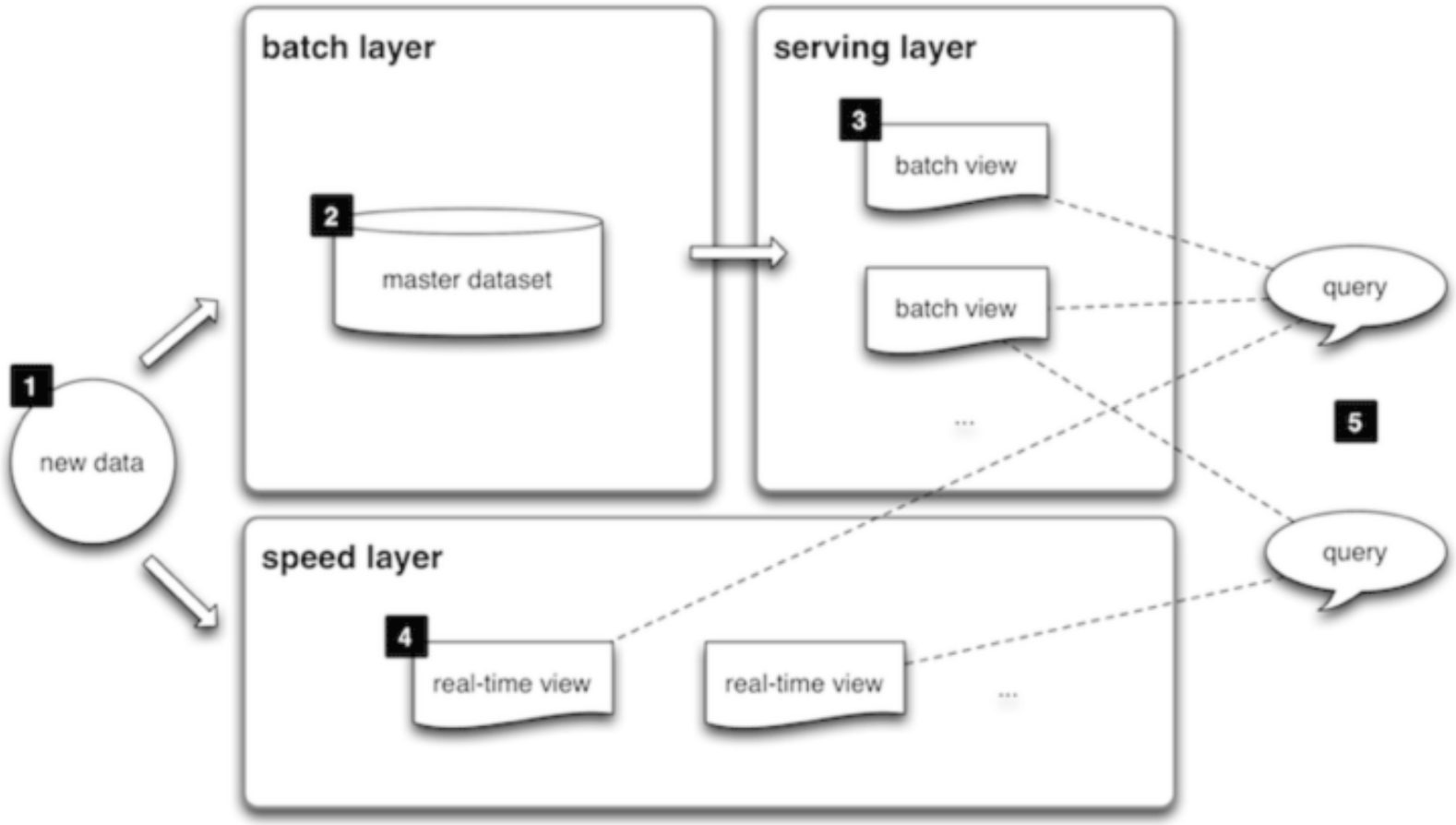
Popular Saying


VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME



http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html

Lambda Architecture



http://www.drdobbs.com/database/applying-the-big-data-lambda-architectur/240162604

“teenage sex” analogy cont’d…

- From LinkedIn:
  - After the initial excitement you wonder what all the fuss was about?
  - The positive side of this analogy, those who really know what they are doing, do fall in love and conceive some great ideas and make their data exploration a wonderful experience.

Take Away Points

1. Big data is not a product but a collection of processes centered around big data resources
   - Collections of data made available for analysis
2. Lambda Architecture is a good way to organize Big Data management
3. Joint focus on data manager and data analyst
   - Not a data mining class, however!

Outline

- Trends Underlying Big Data
- What is Big Data?
- Data Models
- **Course Outline**