



**Hasso  
Plattner  
Institut**

IT Systems Engineering | Universität Potsdam

## Big Data

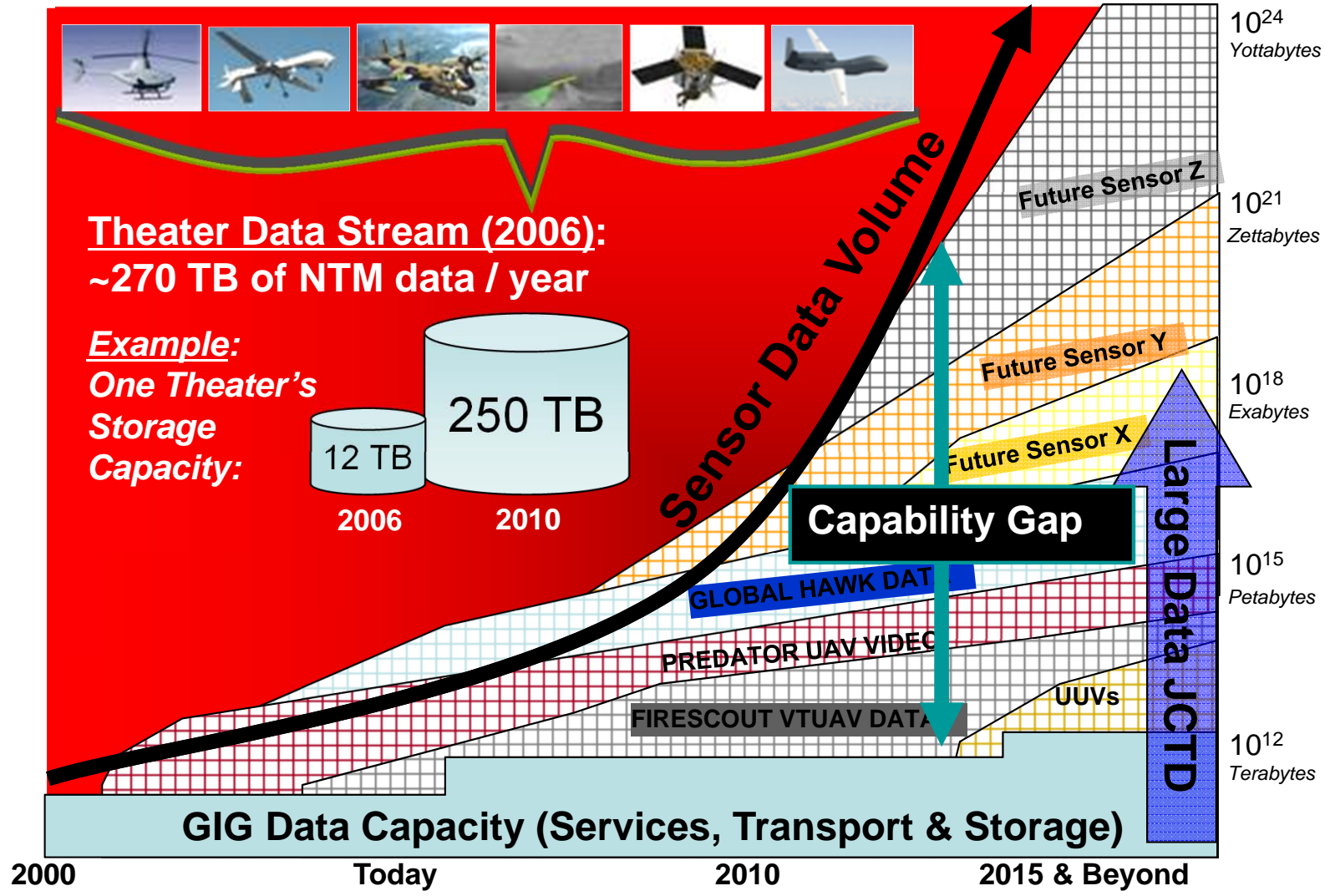
Panel at ICSOC 2013, Berlin

3.12.2013

Felix Naumann

# Military Projection of Sensor Data Volume (later refuted)

2



Using 1TB drives, this would require 1 trillion ( $10^{12}$ ) drives!

# Defining Big Data

3

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
  - Capture
  - Curation
  - Storage
  - Search
  - Sharing
  - Analysis
  - Visualization

If data is difficult to handle, it is Big Data.

# Gartner's 3 + 1 Vs

4

## ■ Volume

- Turn 12 terabytes of Tweets: product sentiment analysis
- 350 billion annual meter readings: predict power consumption

## ■ Velocity

- 5 million daily trade events: identify potential fraud
- 500 million daily call detail records: predict customer churn faster

## ■ Variety

- 100's of live video feeds from surveillance cameras
- 80% data growth in images, video and documents to improve customer satisfaction

## ■ Veracity

- 1 in 3 business leaders don't trust the information they use to make decisions.

<http://www-01.ibm.com/software/data/bigdata/>

# Big and Small

5

- Big Data can be very small
  - Streaming data from aircraft sensors
  - Hundred thousand sensors on an aircraft is “big data”
  - Each producing an eight byte reading every second
  - Less than 3GB of data in an hour of flying
    - ◇ (100,000 sensors x 60 minutes x 60 seconds x 8 bytes).
- Not all large datasets are “big”.
  - Video streams plus metadata
  - Telco calls and internet connections
  - Can be parsed extremely quickly if content is well structured.
- The task at hand makes data “big”.

[http://mike2.openmethodology.org/wiki/Big\\_Data\\_Definition](http://mike2.openmethodology.org/wiki/Big_Data_Definition)

# Open vs. Closed Sources

6

## Open

- Open data
  - [linkeddata.org](http://linkeddata.org)
- Government data
  - [data.gov](http://data.gov), [data.gov.uk](http://data.gov.uk)
  - Eurostat
- Scientific data
  - Genes, proteins, chemicals
  - Scientific articles
  - Climate
  - Astronomy
- Published data
  - Tweet (limited)
  - Crawls
- Historical data
  - Stock prices

## Closed

- Transactional data
  - Music purchases
  - Retail-data
- Social networks
  - Tweets, Facebook data
  - Likes, ratings
- E-Mails
- Web logs
  - Per person
  - Per site
- Sensor data
- Military data

# „Big data“ in business

7

## ■ Amazon.com

- Millions of back-end operations every day
- Catalog, searches, clicks, wish lists, shopping carts, third-party sellers, ...



## ■ Walmart

- > 1 million customer transactions per hour
- 2.5 petabytes (2560 terabytes)



## ■ Facebook

- 250 PB, 600TB added daily (2013)
- 1 billion photos on one day (Halloween)



## ■ FICO Credit Card Fraud Detection

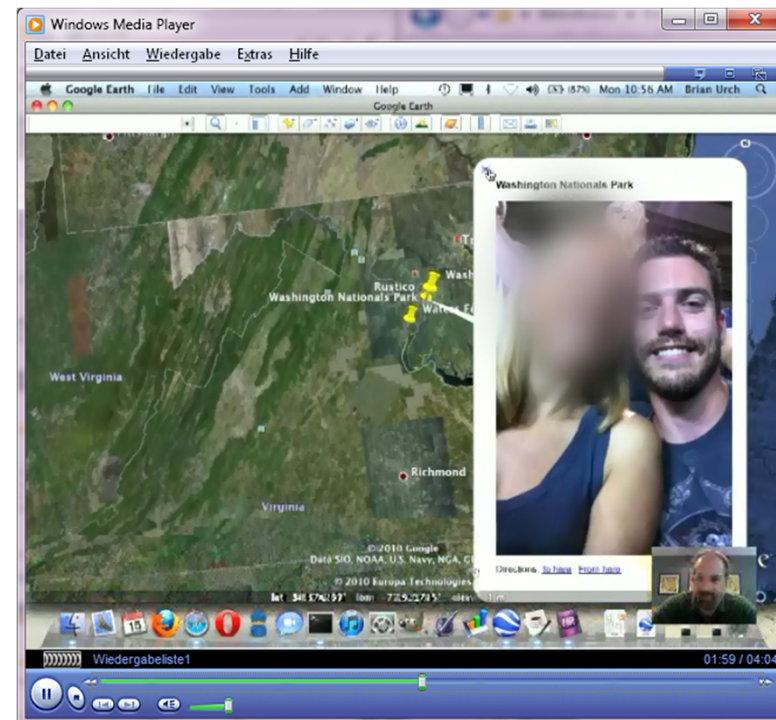
- Protects 2.1 billion active accounts



# „Big data“ in business

8

- Has been used to sell more hardware and software
- Has become a shallow buzzword.
  
- But: The actual big data is there, has added-value, and can be used effectively.
  - Data Mining
  - Marketing
  - Collaborative filtering
  - Raytheon's [RIOT software](#)
  - NSA, etc.
  - Kreditech





# Big Science Data

9

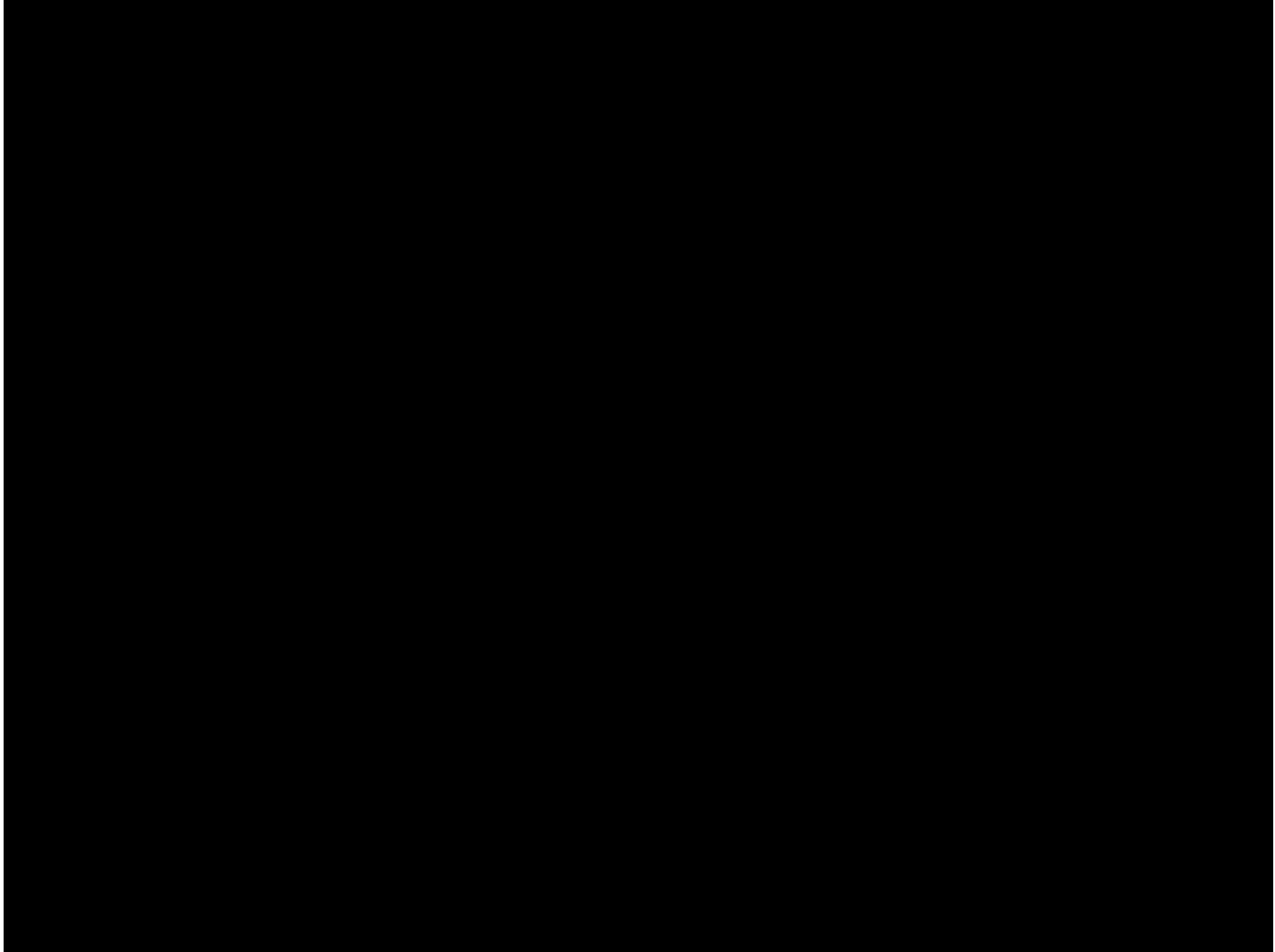
- Large Hadron Collider (LHC)
  - 150 million sensors; 40 million deliveries per second
  - 600 million collisions per second
  - Theoretically: 500 exabytes per day (500 quintillion bytes)
  - Filtering: 100 collisions of interest per second
  - 25 petabytes annual rate
- Sloan Digital Sky Survey (SDSS)
  - Amassed more data in first few weeks than all data collected in the history of astronomy.
  - 200 GB per night, 140 TB in all
  - Successor: Large Synoptic Survey Telescope will acquire that amount of data every five days.
- Human genome
  - Originally took 10 years to process; now one week.

# Big data = science

10

- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - All models are wrong, but some are useful. (George Box)
  - All models are wrong, and increasingly you can succeed without them. (Peter Norvig, Google)
- Before Big Data: Correlation is not causation!
- With Big Data: Who cares?
  - Traditional approach to science — hypothesize, model, test — is becoming obsolete.
- Petabytes allow us to say: "Correlation is enough."

[http://www.wired.com/science/discoveries/magazine/16-07/pb\\_theory](http://www.wired.com/science/discoveries/magazine/16-07/pb_theory)





**Hasso  
Plattner  
Institut**

IT Systems Engineering | Universität Potsdam

## The Big Data Fallacy

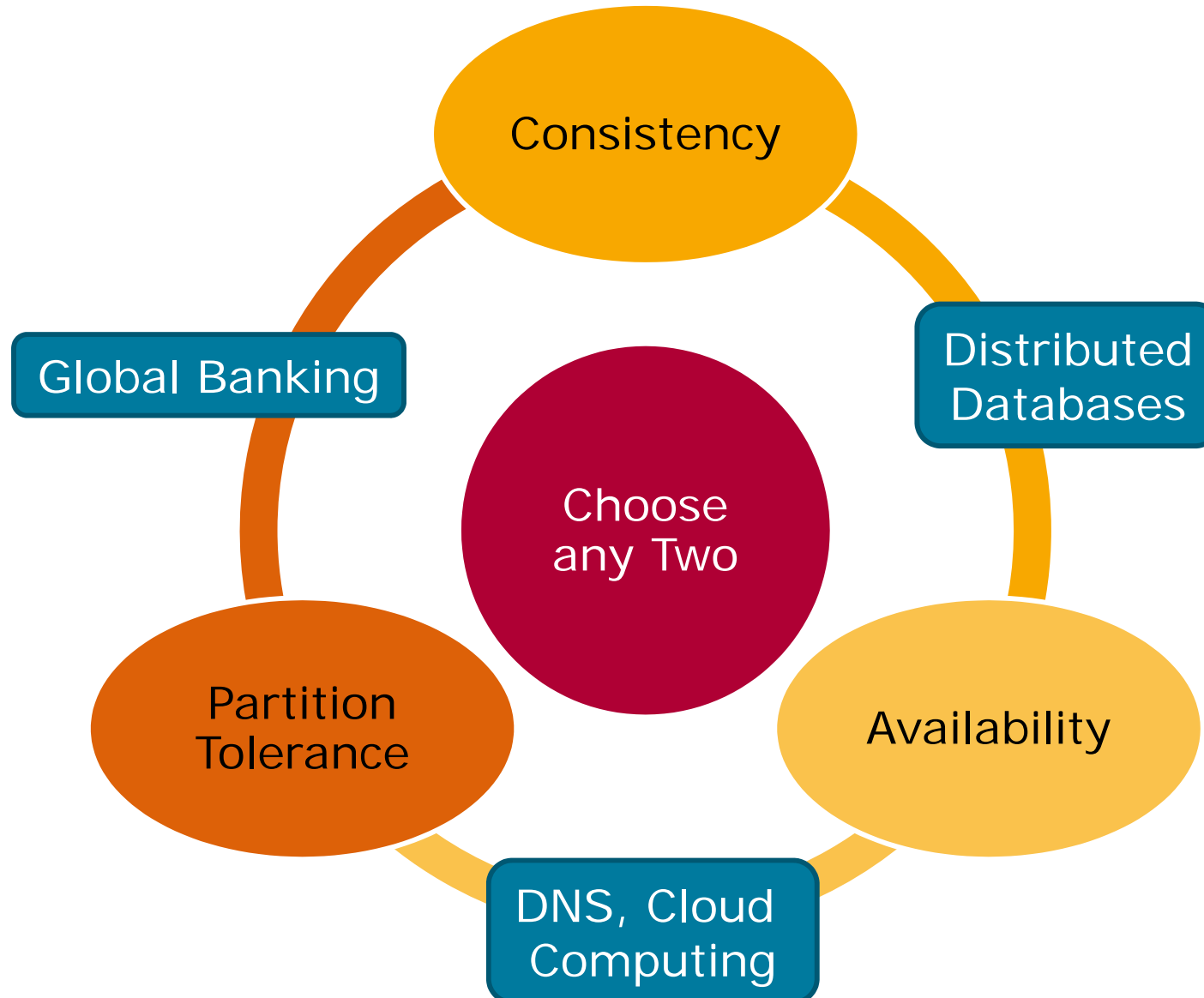
Panel at ICSOC 2013, Berlin

3.12.2013

Felix Naumann

# CAP Theorem for Distributed Systems

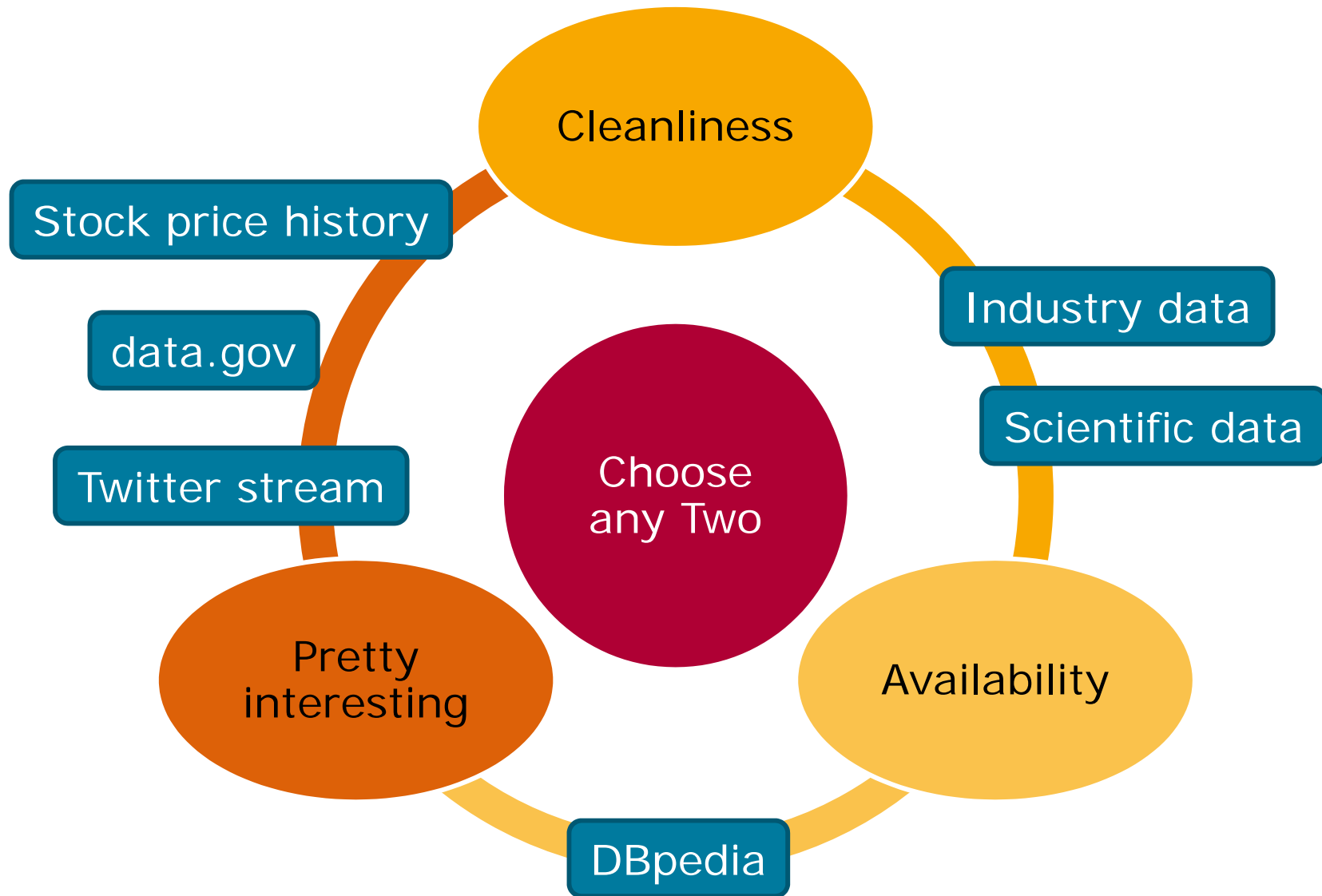
13



# CAP Observation for Big Data

(from a computer scientist's point of view)

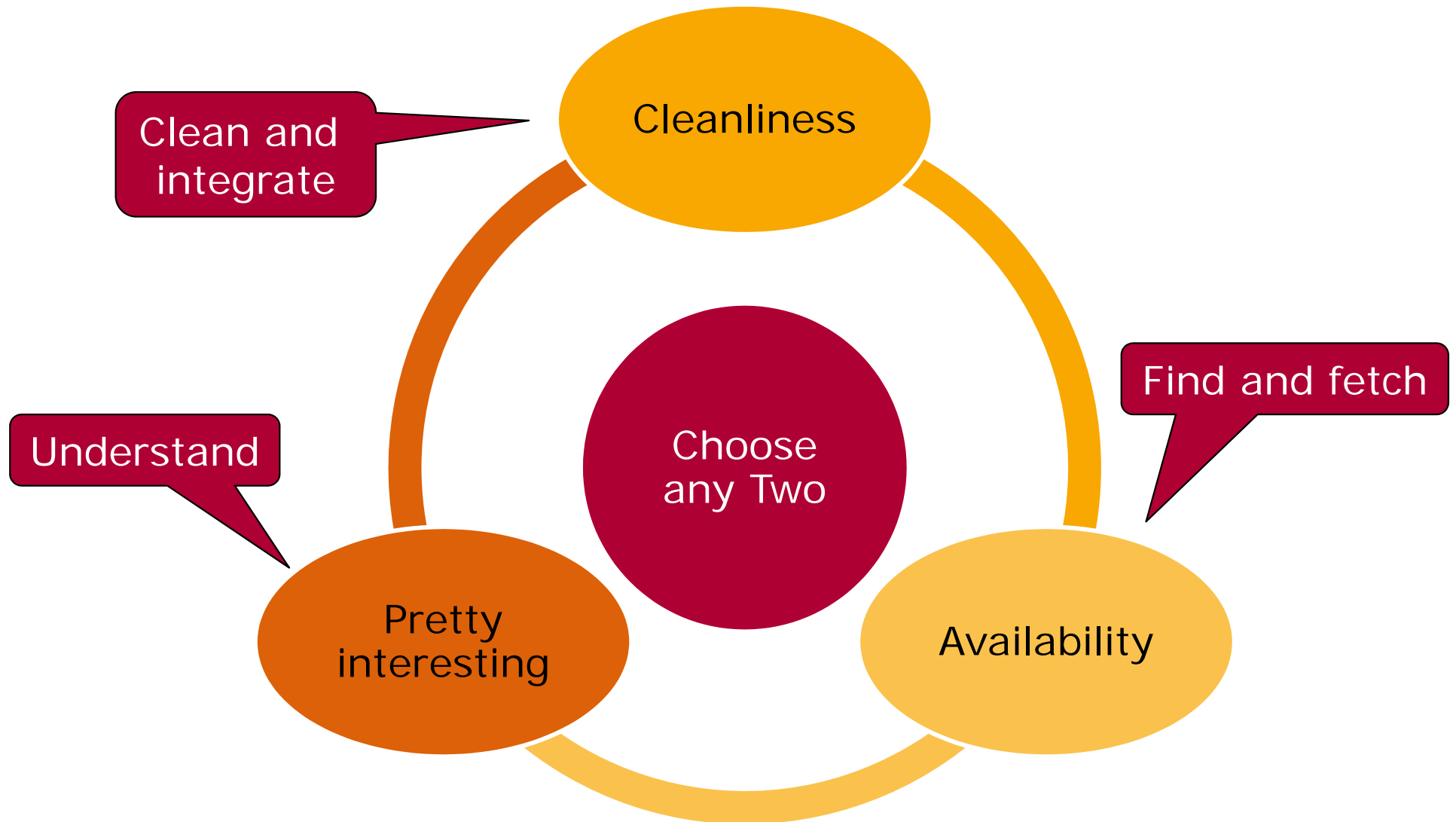
14



# CAP Observation for Big Data

(from a computer scientist's point of view)

15



# Find and fetch

16

The screenshot shows the Data.gov website interface. At the top, there is a search bar with the text 'Tornado' and a search icon. To the right of the search bar is a 'Login' link. Below the search bar is a navigation menu with items: HOME, ABOUT, DATA, METRICS, OPEN GOVERNMENT, BLOGS, and COMMUNITIES. A large blue banner below the navigation menu reads 'DATA CATALOG'. Underneath the banner, there is a breadcrumb trail 'Home / Datasets' and two buttons: 'Organizations' and 'Interactive Datasets'. A paragraph of text explains that only datasets marked as 'Federal' are subject to the U.S. Federal Government and Data.gov's Data Policy. Below this text is a 'Filter by location' section with a 'Clear' link and a search box containing 'Tornado'. To the left of the search box is a map of North America with a search box 'Enter location...'. Below the map are links for 'Map data CC-BY-SA by OpenStreetMap' and 'Tiles by MapQuest'. At the bottom left, there is a 'Dataset Type' section with 'A-Z' and '1-9' filters and a 'Clear All' link. The main content area displays '26 datasets found for "Tornado"' and 'Order by: Relevance'. Two dataset entries are visible, both marked as 'Federal' with a blue diagonal banner. The first entry is 'FEMA Hazard Mitigation Program Summary' by the Federal Emergency Management Agency, Department of Homeland Security. The second entry is 'FEMA Public Assistance Funded Projects Summary' by the Federal Emergency Management Agency, Department of Homeland Security.



# More Tornado Sources on data.gov

17

- Federal Emergency Management Agency, Department of Homeland Security (FEMA)
- Federal Geographic Data Committee
- National Oceanic and Atmospheric Administration, Department of Commerce (NOAA)
- National Weather Service (NWS) Storm Prediction Center (SPC)
- Severe Weather Data Inventory
- U.S. Geological Survey, Department of the Interior
- Fire and EMS Districts
- ... and probably many more.

# Find and fetch

18

- Download
  - Data volumes make downloading increasingly infeasible
  - Fedex HDDs
  - Fedex tissue samples instead of sequence data
  
- Generating big (but synthetic) data
  1. Automatically insert interesting features and properties
  2. Then „magically“ detect them
  
- Sharing data
  - Repeatability of experiments
  - Not possible for commercial organizations

# Understand: Schemata

19

```

{{Infobox Company
| name           = The Corporation Company
| logo           = [[Image:Example.png|160px]]
| type           = [[Public company|Public]] {{{nyse|TCC1}}, {{{tyo|TCC1}}}
| genre          = Corporate histories
| predecessor    = The Wikitory Company
| foundation     = [[New York City]], [[United States|U.S.]] {{{Start date|1900}}}
| founder        = Wikiped Wikiad
| location_city  = [[Seattle]], [[Washington]]
| location_country = [[United States|U.S.]]
| location       =
| locations      = 300 stores (2000) at [[2000-12-31]]
| area_served    = [[North America]]
| key_people     = Wikiped Wikiad <small>[[Entrepreneur|Founder]]</small> <br />
                 Waldo Wikiad <small>[[Chief executive officer|CEO]]</small>
| industry       = [[Publishing]]
| products       = [[Book]]s, [[magazine]]s
| services       = Literary restoration, literary archiving
| revenue        = US$500,000,000 (2000), {{{increase}} 5% from 1999
| operating_income = US$350,000,000 (2000) {{{steady}} from 1999
| net_income     = US$50,000,000 (2000) {{{decrease}} 12% from 1999
| assets         = US$1,500,000,000 at [[2000-12-31]] {{{decrease}} 9% from year earlier
| equity         = US$950,000,000 at [[2000-12-31]] {{{increase}} 6% from year earlier
| owner          = Wikiped Wikiad
| num_employees  = 1,500 (2000)
| parent         = Mega Corporation Inc.
| divisions      = TCC Company Histories, TCC Magazine Services
| subsid         = Restored Book Company, Super Archives, Ltd.
| homepage       = [http://www.thecorporationcompany.com/ TheCorporationCompany.com]
| footnotes     =
| intl          =
}}
    
```

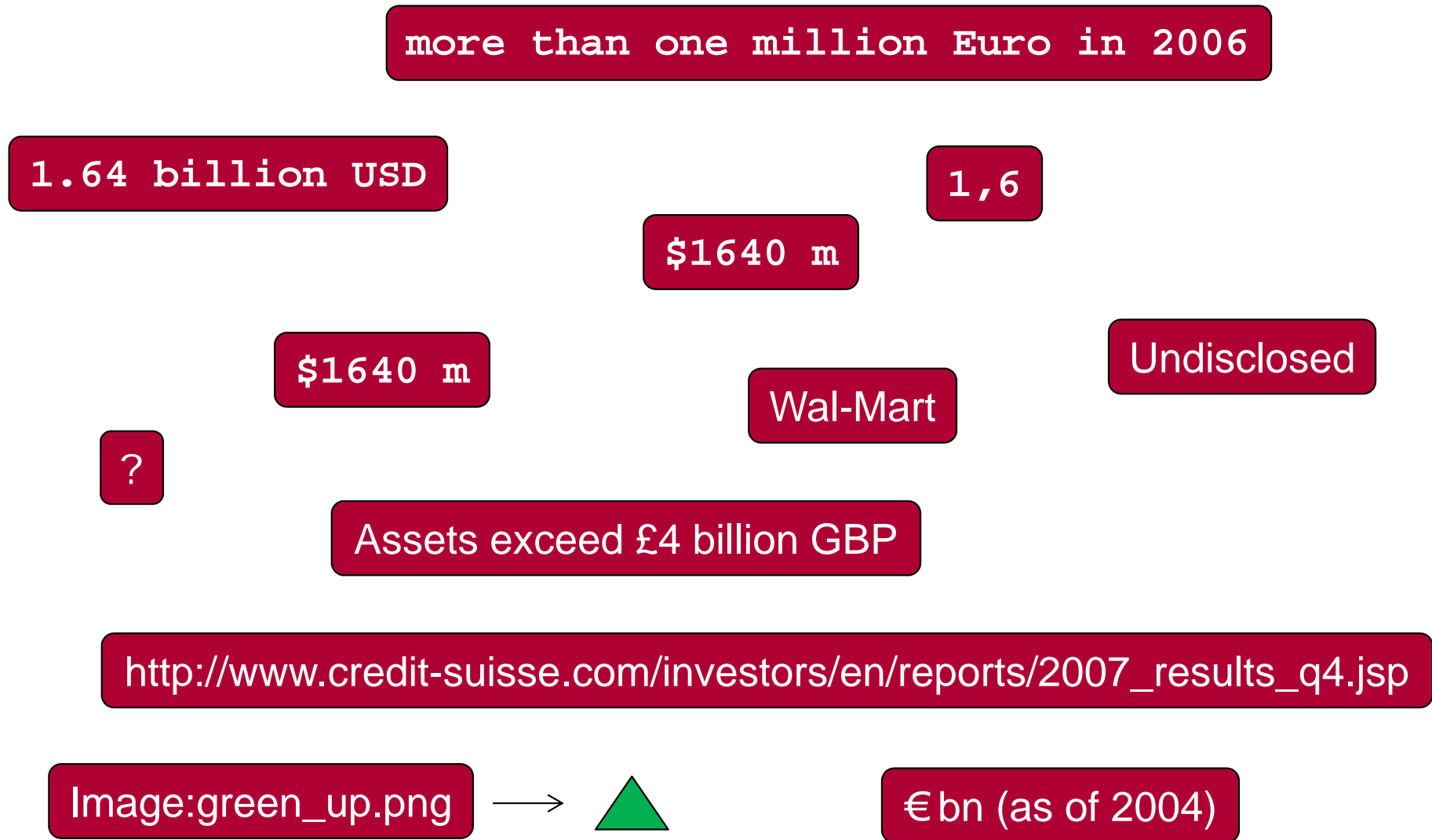
Vertical list	Requirements
<pre> {{Infobox Company   name           =   logo           =   type           =   genre          =   fate           =   predecessor    =   successor      =   foundation     =   founder        =   defunct        =   location_city  =   location_country =   location       =   locations      =   area_served    =   key_people     =   industry       =   products       =   services       =   revenue        =   operating_income =   net_income     =   aum            =   assets         =   equity         =   owner          =   num_employees  =   parent         =   divisions      =   subsid         =   homepage       =   footnotes     =   intl          = }}                 </pre>	<p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p> <p>REQUIRED</p>

fieldName	<info>	Dollars Obligated	Current Contract Value	Ultimate Contract Value	Major Agency	Modified Contracting Agency	Contracting Agency	Contracting Office	Program / Funding Agency	Program / Funding Office	Reason For Purchase For DoD
example1		\$220,989,132	\$220,989,132	\$220,989,132	Dept. of Defense	97AS: Defense Logistics Agency	Defense Logistics Agency	SP0600	Defense Logistics Agency	SP0600	Invalid code
example2		\$33,710,000	\$33,710,000	\$33,710,000	Dept. of Defense	1700: NAVY, Department of the	NAVY, Department of the	N00024	NAVY, Department of the	N00024	Convenience and Economy
info		add?			kind of category for subagency						
info2		never null	never null	never null	never null, use standardized from modified	never null			Contracting Agency, one contract might have several funding agencies		
scrubbing						split			use Contracting Agency if left blank		
map to LegalEntity as recipient											
map to LegalEntity as Parent recipient											
	subject = "USSpending",		amount.curr	amount.ulti							




# Clean and Integrate: "Revenue" values

21



# Clean and integrate: GovWILD

22

- 150,000 persons
  - 270,000 legal entities
  - 1,100,000 funds
  - 43,000,000 triples
- 
- Government Web Integration for Linked Data
- 
- 
- Keyword Queries
  - Linked Data Interface (dereference URIs)
  - Exploration of entities mentioned in New York Times articles
  - Data Download (RDF, SQL Dump, JSON files)

Step	Time		Input size on master node and element count			Details
	Jaql 0.4	Jaql 0.5	LegalEntity.json	Person.json	Fund.json	
<b>Scrubbing</b>						
Scrub and map 15 files	1h 15min	1h 15min	Start with 11 GB size			<ul style="list-style-type: none"> <li>- map and normalize attributes</li> <li>- set references within a source (includes many joins)</li> <li>- group entities / match entities of the same source</li> <li>- use dictionaries for enrichment</li> </ul>
Merging Scrubbed Files	3 min	3 min				<ul style="list-style-type: none"> <li>- concatenate files in HDFS to achieve 3 files containing persons, legal entities, and funds</li> </ul>
			162 MB - 217 087 entities	544 MB - 1 357 810 entities	471 MB - 998 150 ent.	
<b>Matching of LegalEntity</b>						
Write from HDFS to master	6 min	7 sec	-  -			
Find similar entities on workstation	30 min	24 min	-  -			<ul style="list-style-type: none"> <li>- computes duplicates in pairs of 2, <b>non-parallel</b></li> </ul>
Write back to HDFS	7 sec	6 sec	44 MB - 7530 pairs			
Fuse similar objects	10 min	10 min	-  -			<ul style="list-style-type: none"> <li>- compute transitive closure of IDs (transform and combine with UDF)</li> <li>- join clustered IDs with objects (2 minutes)</li> <li>- group by cluster_ID</li> <li>- split large clusters (transform with UDF)</li> <li>- fuse these clusters (transform with UDF)</li> </ul>
Update fused IDs in all files (merge new IDs from Legal Entity into Person, Fund and LegalEntity)	10 min	10 min	211 362 entities	1 357 810 entities	998 150 ent.	<ul style="list-style-type: none"> <li>- transform on source file to find all ID changes</li> <li>- transform on target file to find all possibly old references</li> <li>- join both</li> <li>- group by target ID</li> <li>- join this with target file (3 min for merging from LegalEntity to Person)</li> <li>- transform this to set new IDs</li> </ul>
<b>Matching of Person</b>						
Write from HDFS to master	18 min	20 sec		544 MB		
Find similar entities on workstation	44 min	48 min		-  -		<ul style="list-style-type: none"> <li>- as above, <b>non-parallel</b></li> </ul>
Write back to HDFS	8 sec	12 sec		79 MB - 51 634 pairs		
Fuse similar objects	11 min	10 min		Join 35 744 fused with all Persons		<ul style="list-style-type: none"> <li>- as above</li> </ul>
Remove irrelevant Freebase Persons	1 min	1 min		filter 328 889 out of 1 323 112		<ul style="list-style-type: none"> <li>- remove freebase persons without references (filter)</li> </ul>
Update fused IDs in all files	10 min	9 min	211 362 entities	328 889 entities	998 150 ent.	<ul style="list-style-type: none"> <li>- as above, from Person file to all others</li> </ul>
<b>Finalize data</b>						
Precanned Query for US states	9 min	10 min	-  -	-  -	-  -	<ul style="list-style-type: none"> <li>- for every object create stateEntities array with connected state names (transform on LegalEntity, Person, Fund)</li> <li>- filter US states from legal entities to create US states file</li> <li>- replace state names with state IDs (similar to updating IDs before)</li> </ul>
Clean up attributes	2 min	1.5 min	-  -	-  -	-  -	<ul style="list-style-type: none"> <li>- remove empty arrays</li> </ul>
Write JSON from HDFS to master	40 min	1min	175 MB	428 MB	524 MB	
<b>Prepare for RDF export</b>						
Add attributes	1 min	2 min	-  -	-  -	-  -	<ul style="list-style-type: none"> <li>- add „label“ and „uri“ fields (transform with UDF)</li> </ul>
Replace ID references by URI references	23 min	19 min	-  -	-  -	-  -	<ul style="list-style-type: none"> <li>- as update IDs above, for most combinations of LegalEntity, Person, and Fund (Funds are never referenced)</li> </ul>
Write from HDFS to master	46 min	1 min	185 MB - 211 362 entities	453 MB - 328 889 entities	689 MB - 998 150 ent.	
	sum: 5h 39 min	Sum: 3h 45min				

# Interesting queries

24

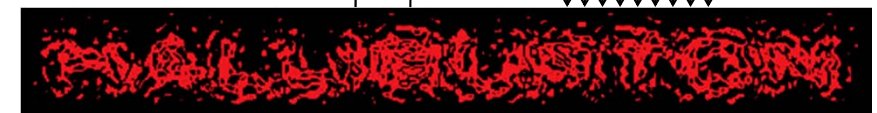
- Find all *classmates* of George W. Bush who, during his term, have worked at a company that has received government funding.
- For each member of congress, find all earmarks awarded to organizations that have *employed a relative* of that member of congress.
- For each government employees, find all companies that have received funding supported by that member and have *employed him after/before their term in congress*.



Chairman of the board

Funds

CEO





## A final word: Ethics of big data

25



- Industry keynote speakers on credit ratings using big data
  - „If the data is out there, we will find it.“
  - „... and that is why I closed my Twitter account.
  - ... and that is why I had my son close his Twitter account.“