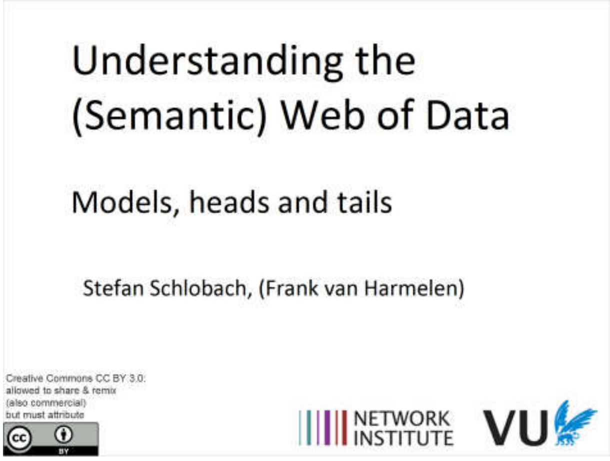


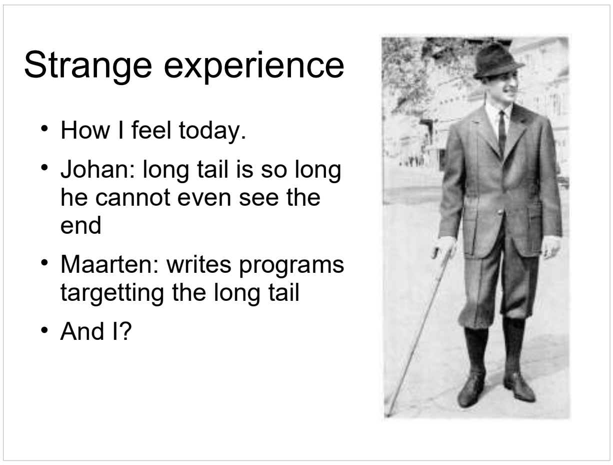
Understanding the (Semantic) Web of Data

Models, heads and tails

Stefan Schlobach, (Frank van Harmelen)

NETWORK INSTITUTE

VU

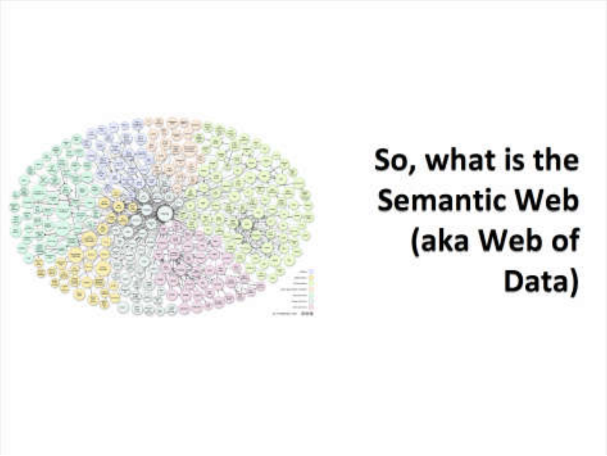


Strange experience

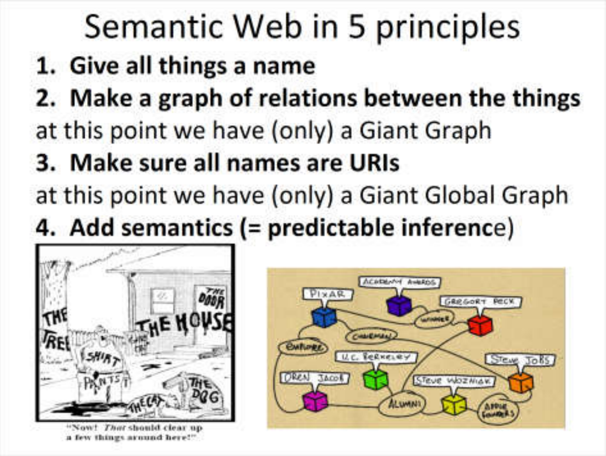
- How I feel today.
- Johan: long tail is so long he cannot even see the end
- Maarten: writes programs targetting the long tail
- And I?

The Semantic Web has a long tail!

And we have to deal with it



So, what is the Semantic Web (aka Web of Data)



Semantic Web in 5 principles

1. **Give all things a name**
2. **Make a graph of relations between the things**

at this point we have (only) a Giant Graph

3. **Make sure all names are URIs**

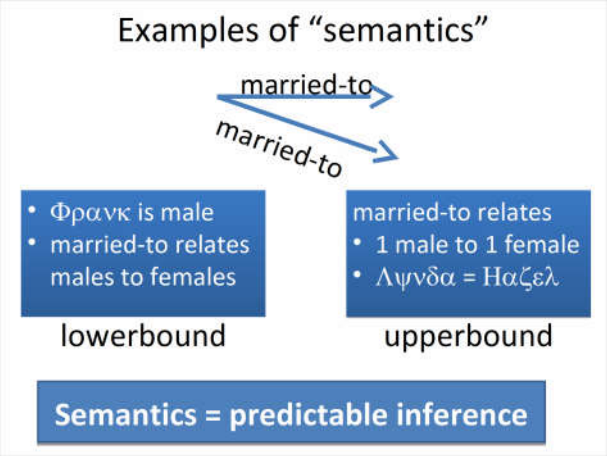
at this point we have (only) a Giant Global Graph

4. **Add semantics (= predictable inferenc**e)

THE TREE
THE HOUSE
THE DOOR
SHIRT
PANTS
THE CAT
THE DOG

"Now! *That* should clear up a few things around here!"

ACADEMY AWARDS
PIXAR
GREGORY PECK
WINNER
CHAIRMAN
EMPLOYEE
U.C. BERKELEY
STEVE JOBS
OREN JACOB
STEVE WOZNIAK
ALUMNI
APPLE FOUNDERS



Examples of "semantics"

married-to
married-to

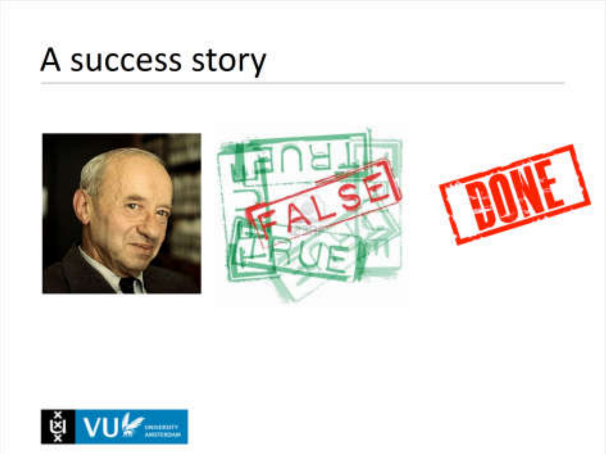
- Φρανκ is male
- married-to relates males to females

lowerbound

married-to relates
- 1 male to 1 female
- Λψνδα = Ηαζελ

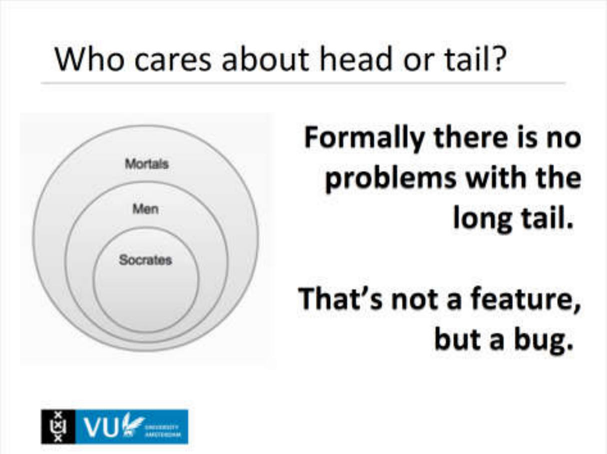
upperbound

Semantics = predictable inference



A success story

TRUE
FALSE
TRUE
DONE

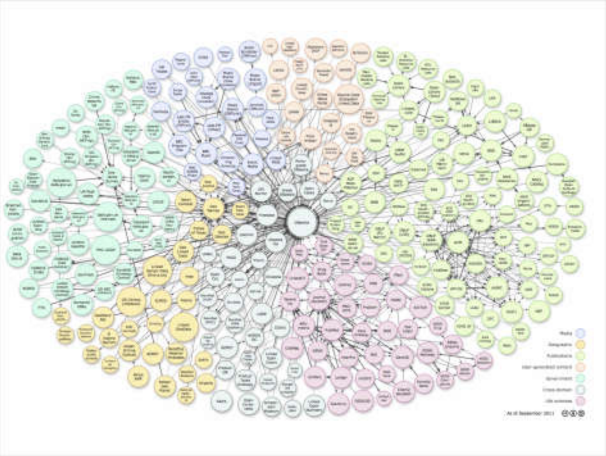


Who cares about head or tail?

Mortals
Men
Socrates

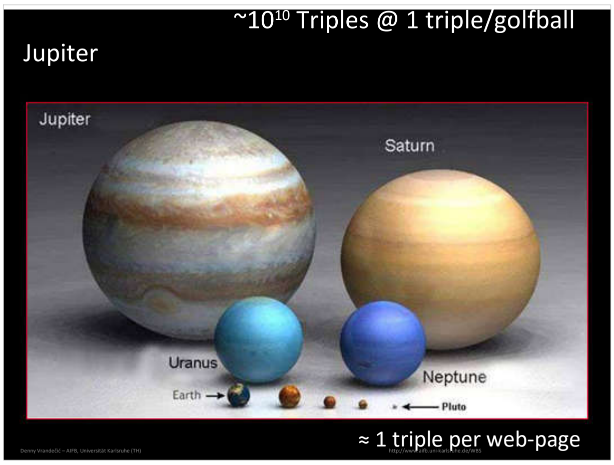
Formally there is no problems with the long tail.

That's not a feature, but a bug.



How big is the Semantic Web now?

1 triple



~10^{10} Triples @ 1 triple/golfball

Jupiter

≈ 1 triple per web-page

Denny Vrandečić – AIFB, Universität Karlsruhe (TH)

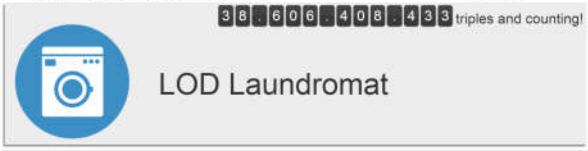
http://www.aifb.uni-karlsruhe.de/WBS

Building
a WoD
observatory?

LOD Laundromat:

clean other peoples dirty data

crawl from registries + user driven
clean syntax errors
compute meta-data information
publish triples: gzip, hdt, ldf
Publish meta-data: SPARQL
harvest 1B triples/day

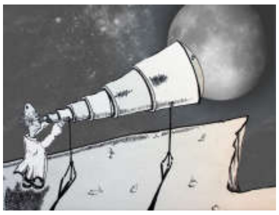


38.606.408.433 triples and counting!

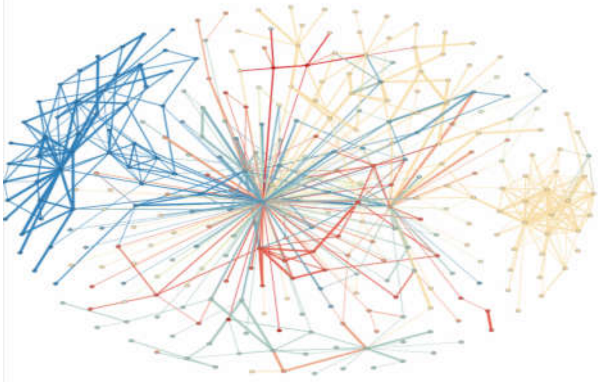
LOD Laundromat

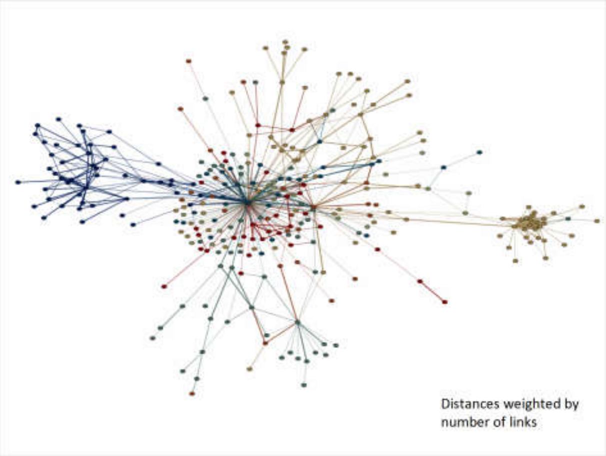
An observatory for the biggest Knowledge Base ever

- Add LOTUS: from words to resources
- A centralised infrastructure to work with and analyse decentralised data

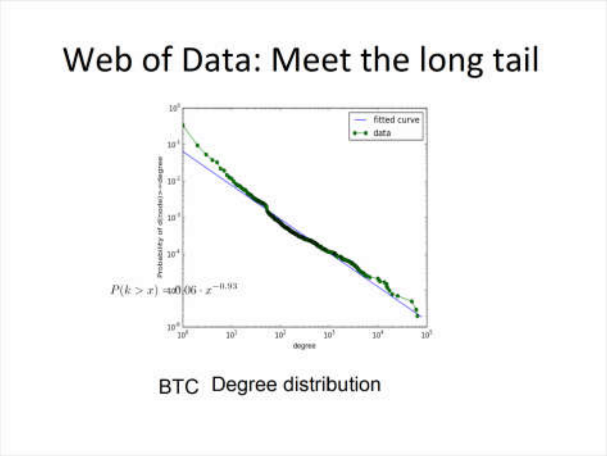


Models, Heads and Tails





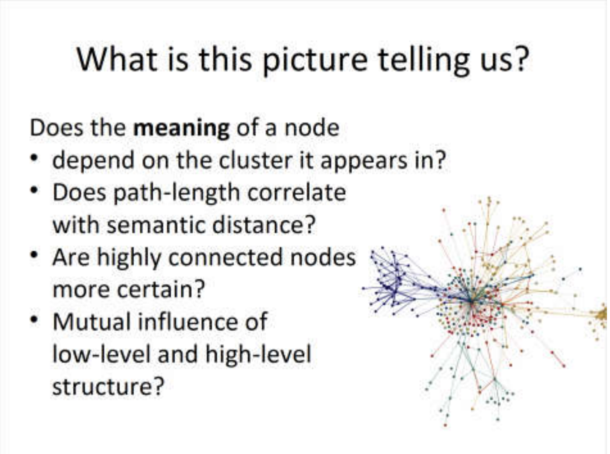
Distances weighted by number of links



Web of Data: Meet the long tail

$P(k > x) = 0.06 \cdot x^{-0.93}$

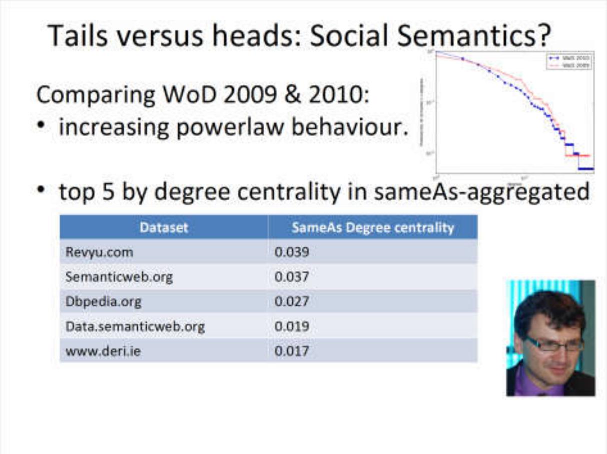
BTC Degree distribution



What is this picture telling us?

Does the **meaning** of a node

- depend on the cluster it appears in?
- Does path-length correlate with semantic distance?
- Are highly connected nodes more certain?
- Mutual influence of low-level and high-level structure?



Tails versus heads: Social Semantics?

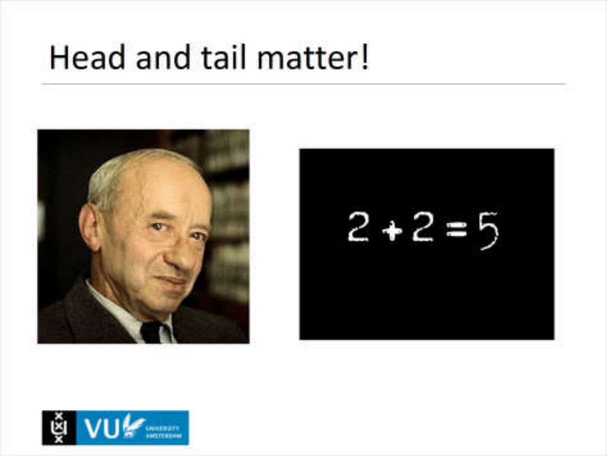
Comparing WoD 2009 & 2010:

- increasing powerlaw behaviour.

WoD 2010
WoD 2009

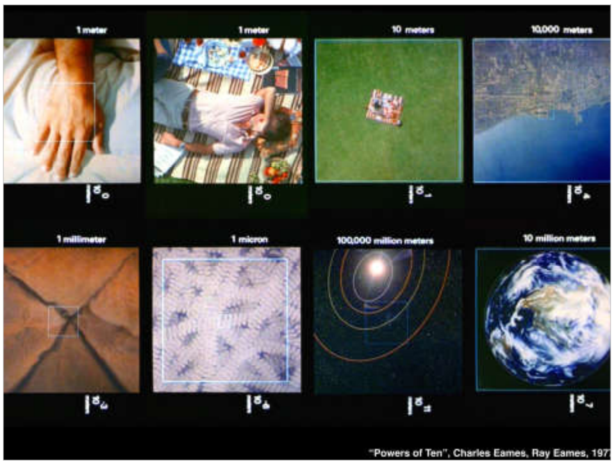
- top 5 by degree centrality in sameAs-aggregated

Dataset	SameAs Degree centrality
Revyu.com	0.039
Semanticweb.org	0.037
Dbpedia.org	0.027
Data.semanticweb.org	0.019
www.deri.ie	0.017

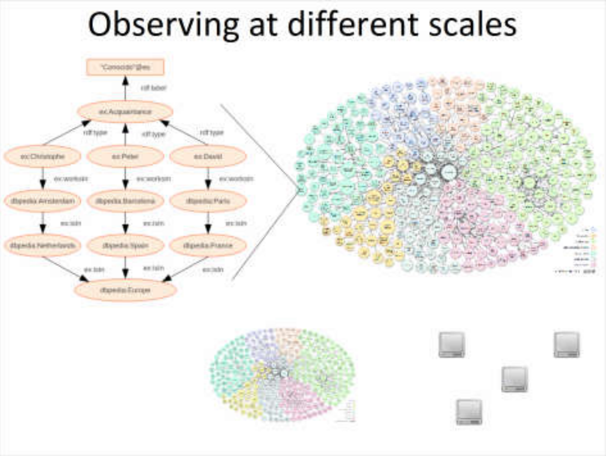


Head and tail matter!

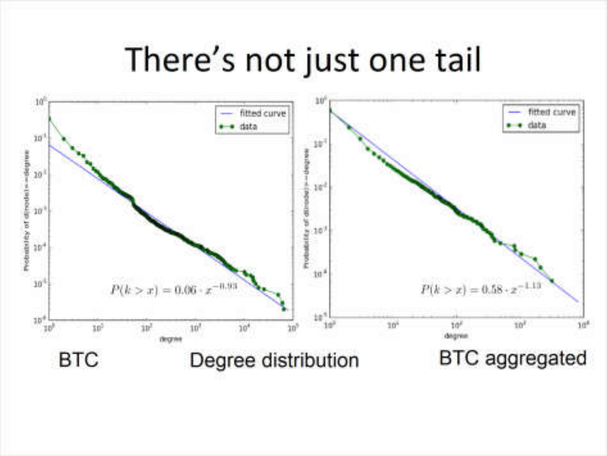
Warning: Which head, which tail?



"Powers of Ten", Charles Eames, Ray Eames, 197



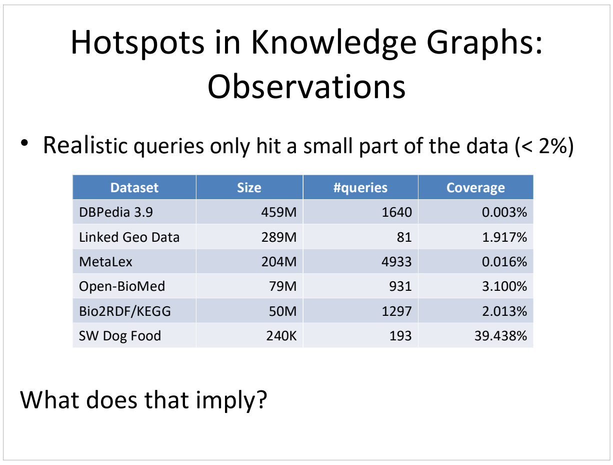
Observing at different scales



There's not just one tail

BTC Degree distribution BTC aggregated

Understanding heads and tails and how to use these insights



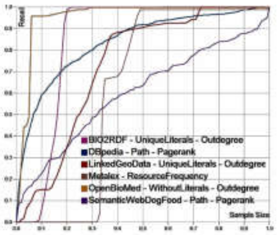
Hotspots in Knowledge Graphs: Observations

- Realistic queries only hit a small part of the data (< 2%)

Dataset	Size	#queries	Coverage
DBPedia 3.9	459M	1640	0.003%
Linked Geo Data	289M	81	1.917%
MetaLex	204M	4933	0.016%
Open-BioMed	79M	931	3.100%
Bio2RDF/KEGG	50M	1297	2.013%
SW Dog Food	240K	193	39.438%

What does that imply?

Hotspots in Knowledge Graphs: Benefits?



1) Ignore the long tail (for efficiency)!

2) Look at the long tail (for completeness)!

Closing

The Web of Data requires Semantics that “understand” heads and tails.