Big Data
Graph Data
Incomplete Information

Leonid Libkin

A bit about the group

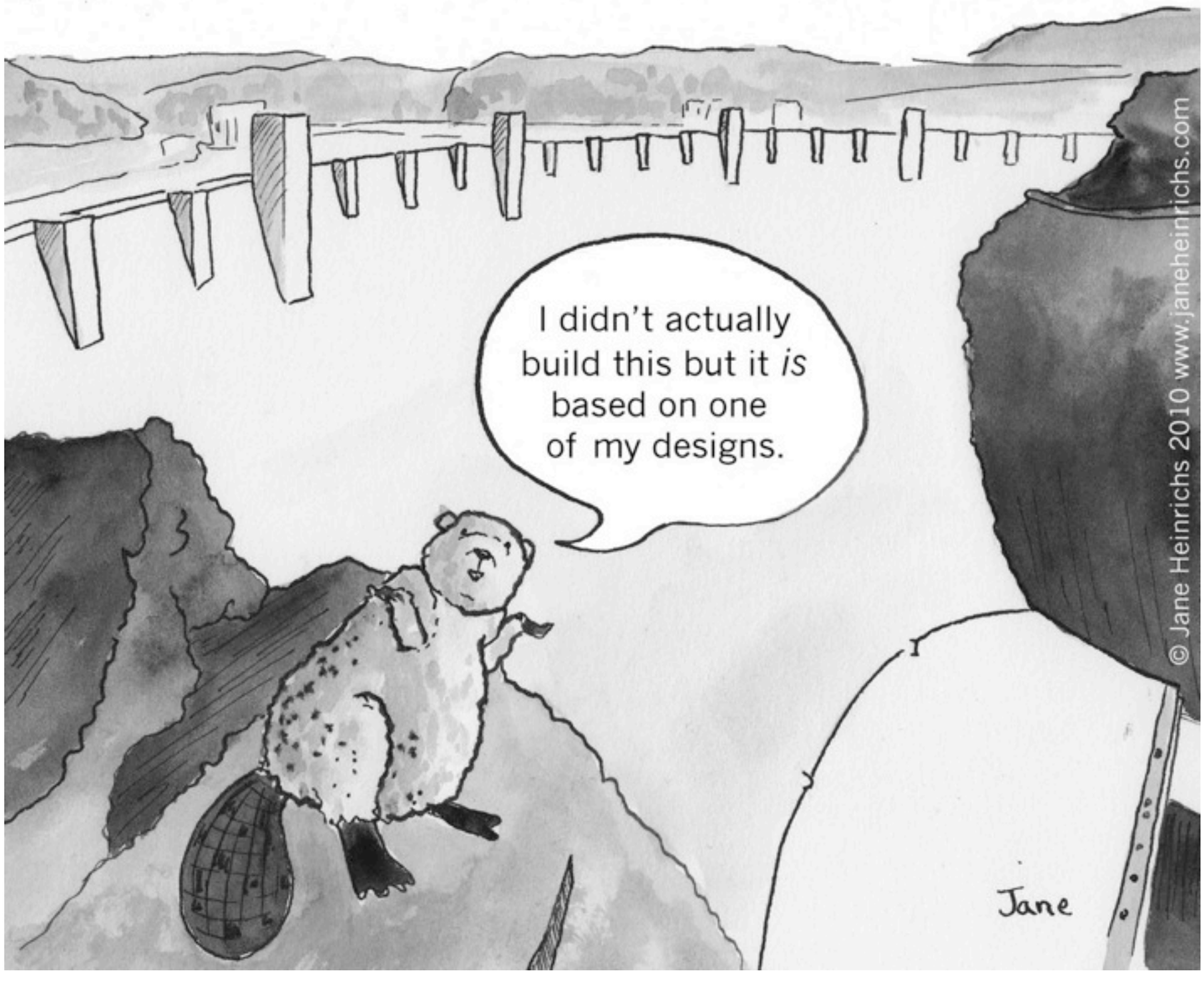
- Ranges from 6 (never again) to one
- Right now, 2 postdocs, soon 3
- Looking for new student (one, at most two)
- Key themes: data management (3 Vs of big data - volume, variety, verasity: scalability; relational, XML, graph data; incompleteness and inconsistency), foundations (as they are needed to handle those questions)

Past students/postdocs

- Mainly academic jobs (12 out of 14 have academic positions in places such as Paris, Singapore, Santiago, Warsaw, Bordeaux; one at IBM, one at Oracle)
- Several notable awards by students:
  - BCS Distinguished Dissertation Award
  - Cor Bayeen Award
  - EPSRC postdoctoral fellowship
  - ACM SIGMOD Honorable mention (2nd prize)
  - 8 (or more) best paper awards

- Main demands to students:
  - very good background
  - interest in what they are doing
- Flexibility with projects: there is always a choice, nothing is ever imposed

Foundational work



If not...

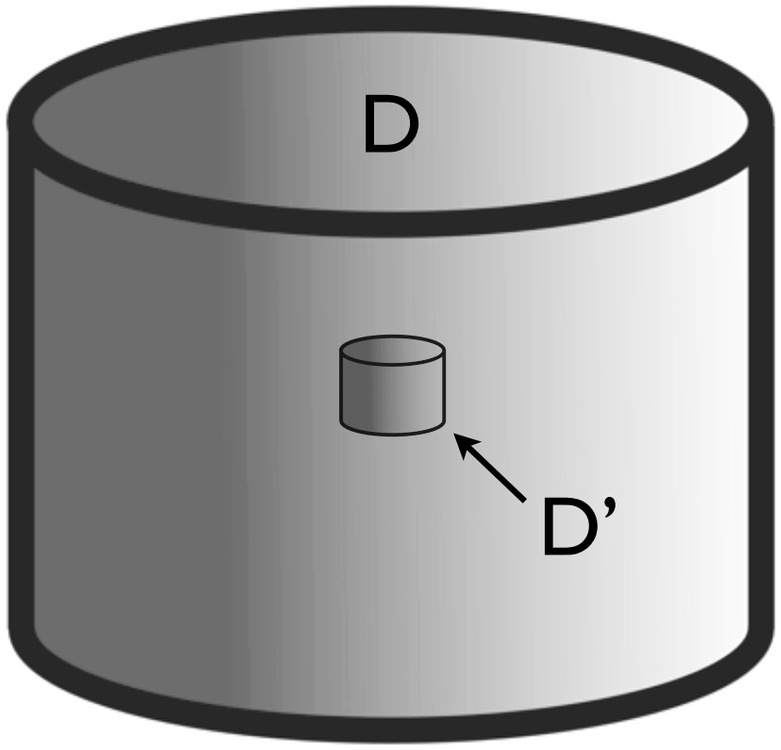
Big data and data management

- Everyone is talking about untapped value of big data
- but data analytics only account for a small fraction of time invested in big data processing!
- Data wrangling (handling data before analyses can begin) can take up to 80% of the effort.
- But data management tasks need to be adapted.

The 4 Vs

- Volume, Velocity, Variety, Veracity
- Volume - scalability
- Variety - graph data (XML is done and gone)
- Veracity - handling uncertainty

Scalability



Q(D)=Q(D’)

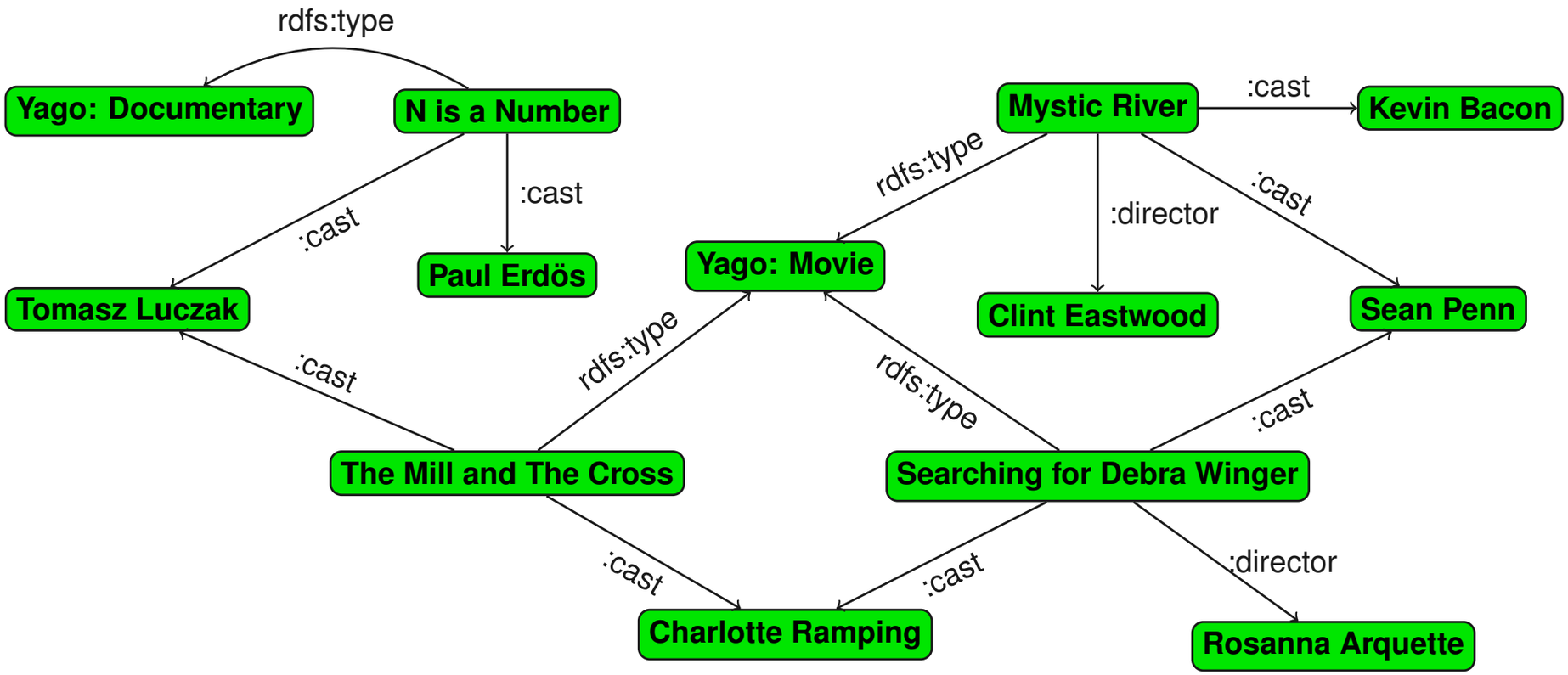
Scale-independence: answering queries regardless of size

People do it in ad hoc ways (eg Facebook), we study it

Scalability

- Need sublinear algorithms
- Massive literature, but mainly concentrates on summary properties
  - find average number of friends in a social network
    - can be found in $O(\sqrt{n})$ up to a factor close to 2
- But no good algorithms for typical data management queries
  - Find friends of John who live in Edinburgh

Graph Databases



Old techniques do not work.

New issues: combining data and topology

Graph data querying

- Data in graphs: standard techniques (SQL)
- Topology of graphs: specialized queries
  - reachability + regular expressions
- Combining is nontrivial - but necessary

Incomplete Information

- Practice: incorrect answers (your laptop thinks that $|X|>|Y|$ and $X-Y=\varnothing$ are consistent!)
- Theory: computationally expensive notions of correctness

- It has been like that for 30+ years, until very recently
- Trying to break the curse of incomplete information

Why N-1=0 for all N

built into SQL standards, every one from the 1980s - hence you have it on you laptop!

R
1
2
3
...
N

S
-

Why N-1=0 for all N

built into SQL standards, every one from the 1980s - hence you have it on you laptop!

R	S	Difference R-S
1	-	
2		
3		
...		
N		

Why N-1=0 for all N

built into SQL standards, every one from the 1980s - hence you have it on you laptop!

R	S	Difference R-S
1	-	
2		
3		
...		
N		

```
SELECT R.A FROM R
WHERE R.A NOT IN
        (SELECT S.A FROM S)
```

Why N-1=0 for all N

built into SQL standards, every one from the 1980s - hence you have it on you laptop!

R

1
2
3
...
N

S

-

Difference R-S

```
SELECT R.A FROM R
WHERE R.A NOT IN
        (SELECT S.A FROM S)
```

Answer: EMPTY for all N

Why N-1=0 for all N

built into SQL standards, every one from the 1980s - hence you have it on you laptop!

R

1
2
3
...
N

S

-

Difference R-S

```
SELECT R.A FROM R
WHERE R.A NOT IN
        (SELECT S.A FROM S)
```

Answer: EMPTY for all N

So N-1=0 after all!

If interested...

- please come and talk to me
- libkin@inf.ed.ac.uk
- https://www.google.co.uk/#q=libkin