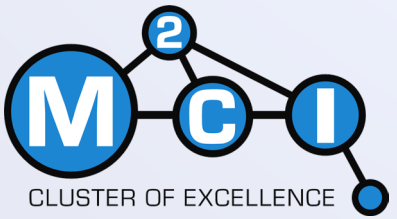
TADA!

Topics in Algorithmic Data Analysis

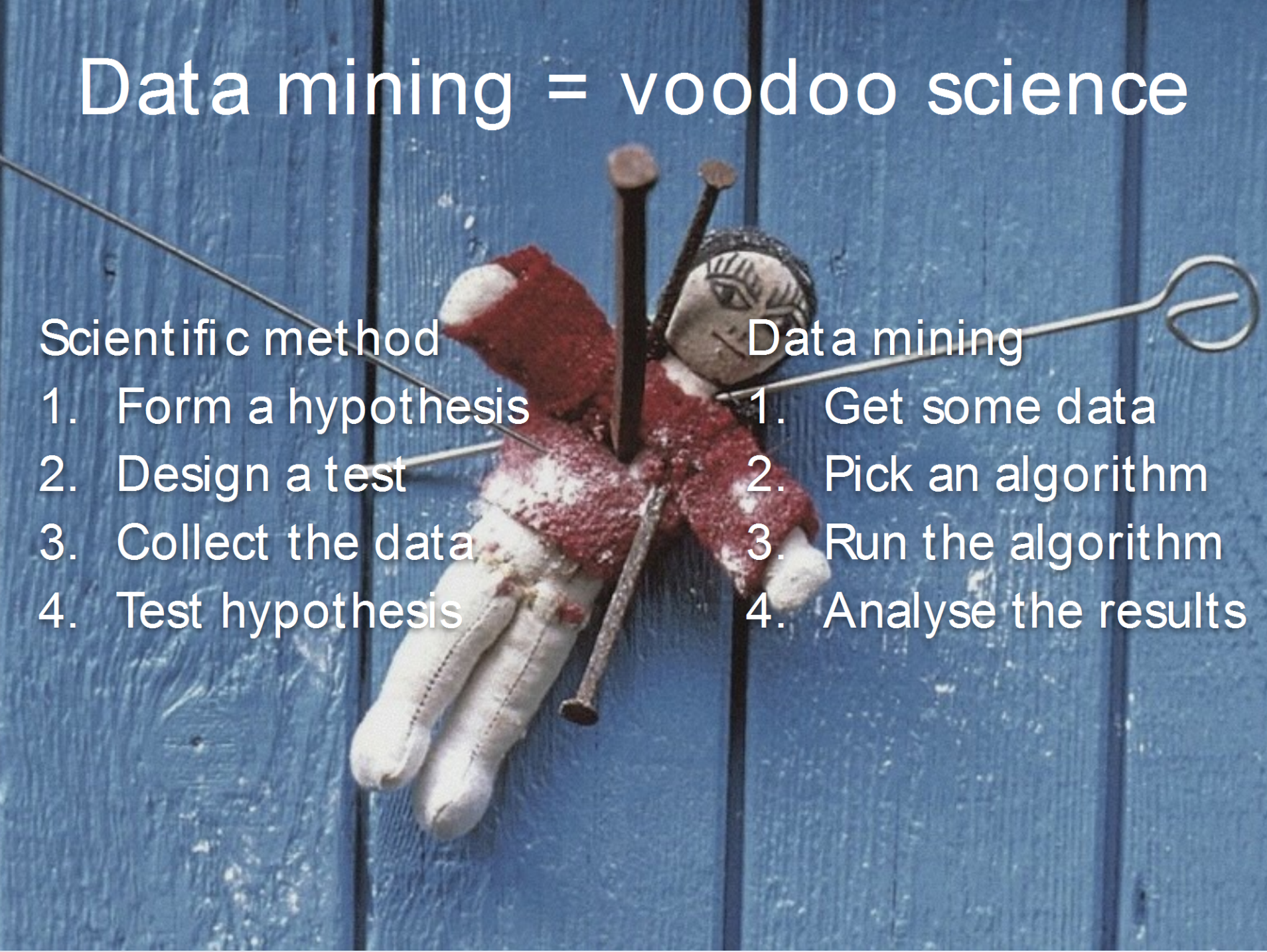
Pauli Miettinen & Jilles Vreeken



Organization

- 5 credit points
- Lectures 2 h per week (now, here)
- Divided into four topics
- No weekly tutorials
- Four written assignments + final exam
  - First given next week
- What/when/where/how/why explained next week

Data mining = voodoo science



Scientific method

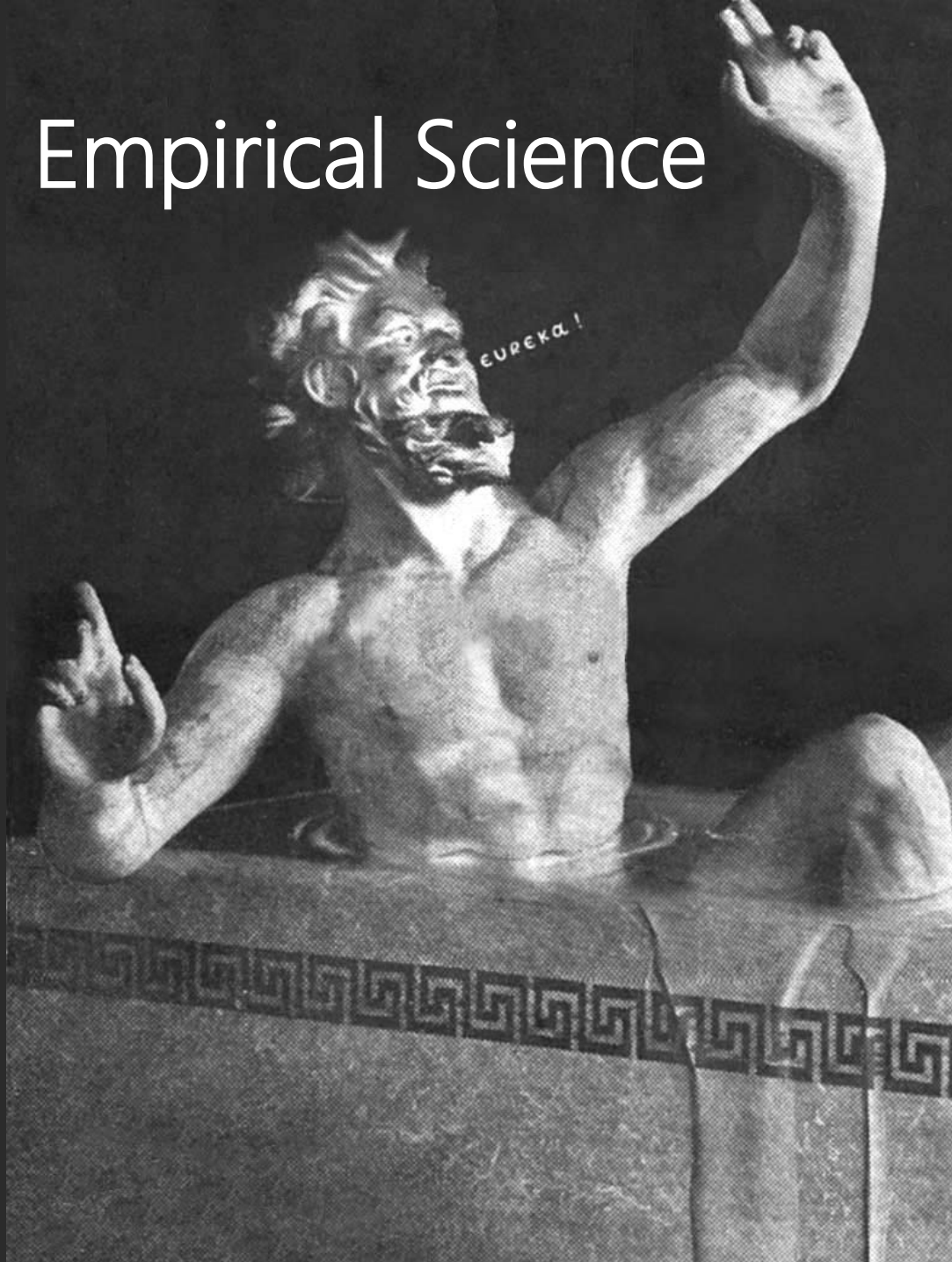
1. Form a hypothesis
2. Design a test
3. Collect the data
4. Test hypothesis

Data mining

1. Get some data
2. Pick an algorithm
3. Run the algorithm
4. Analyse the results

1st Paradigm: Empirical Science

For thousands of years, science was **empirical**: describing natural **phenomena**

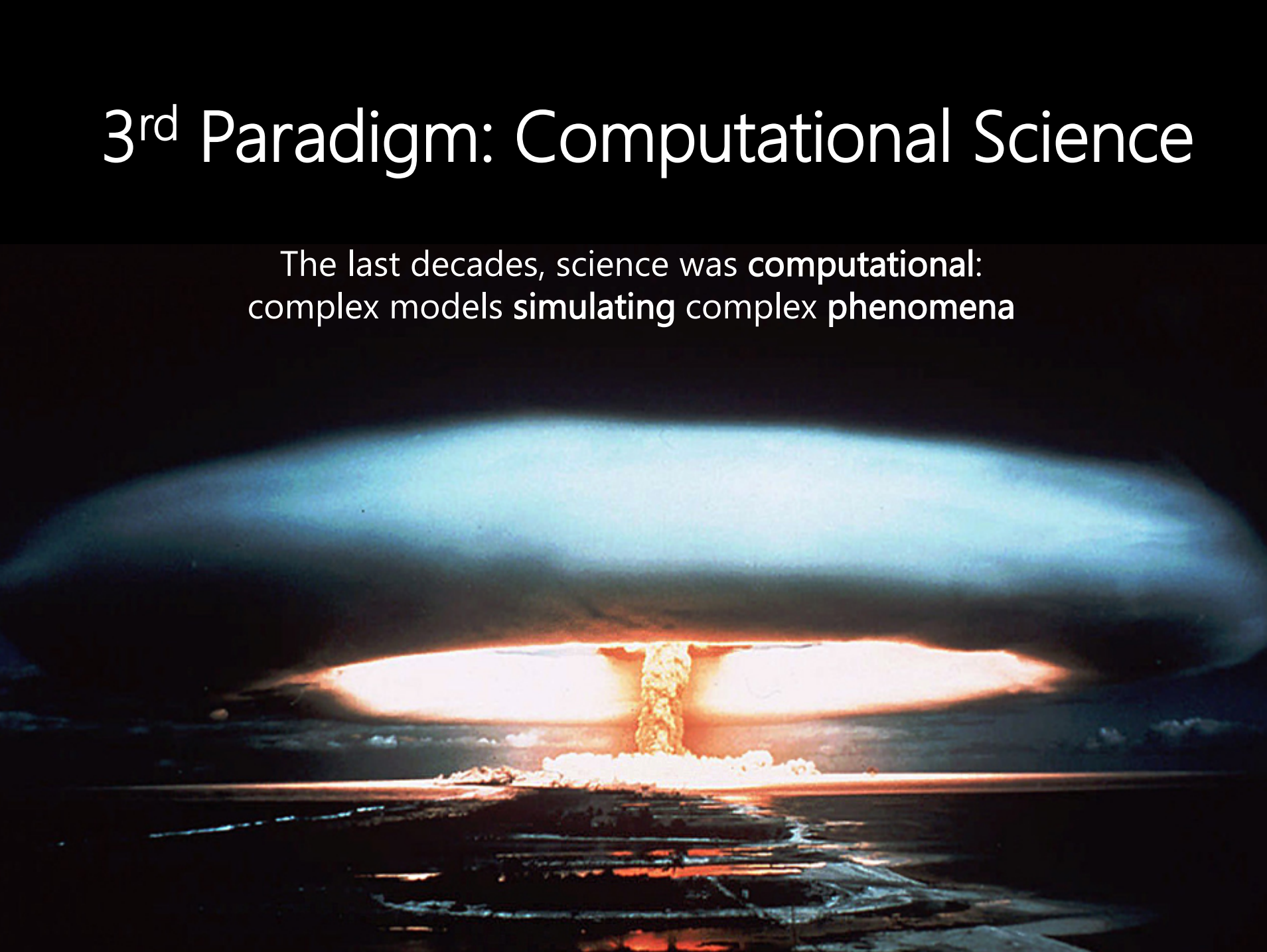


2nd Paradigm: Theoretical Science

The last few hundred years science was **theoretical**:
used models, generalizations, made **predictions**

3rd Paradigm: Computational Science

The last decades, science was **computational**:
complex models **simulating** complex **phenomena**



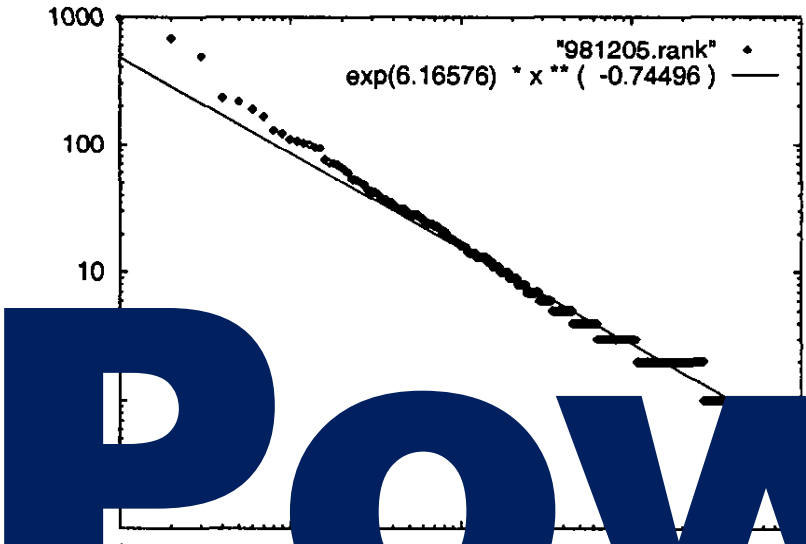
4th Paradigm: Data-Intensive Science

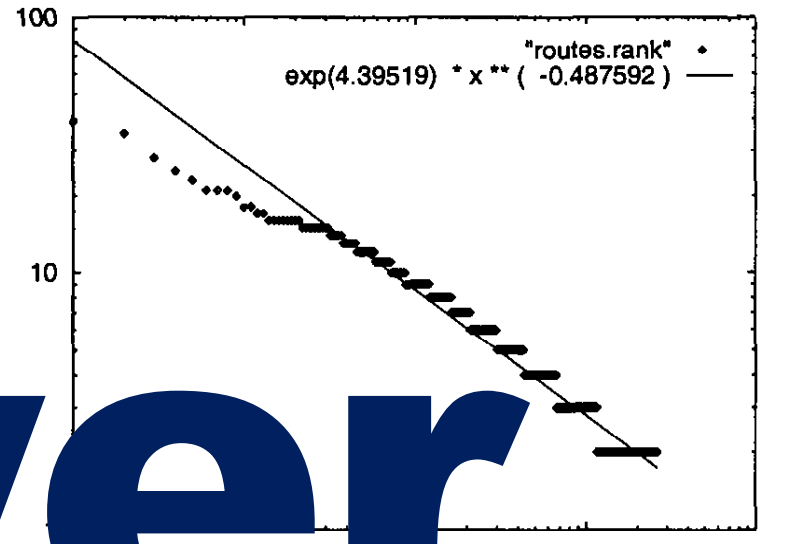
Interesting phenomena are **too complex**
to come up with good hypotheses.

We need to
unify theory, experimentation, and simulation

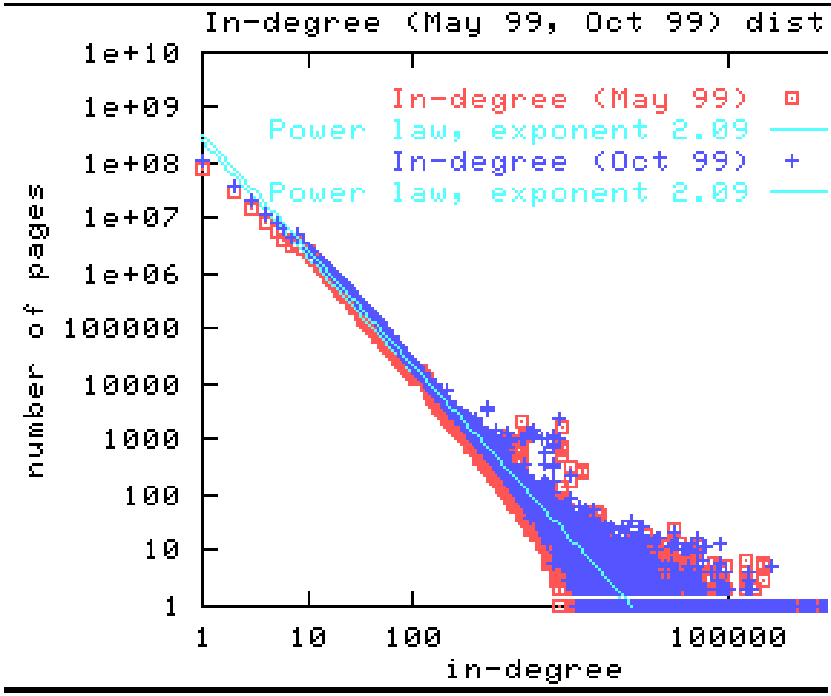
capture data, **mine** hypotheses, **inspect** and evaluate,
generate extra data to **select** the best ones, iterate

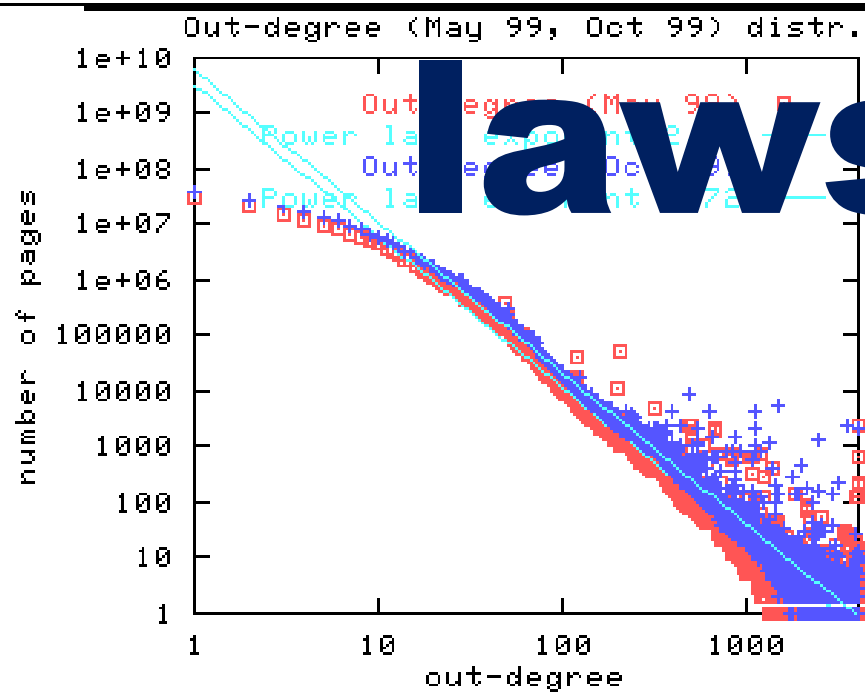
iterative procedure
between **world and model**,
scientist in the middle


"981205.rank"
exp(6.16576) * x ** (-0.74496)
1000
100
10
1


"routes.rank"
exp(4.39519) * x ** (-0.487592)
100
10
10
1000
10000

Power laws

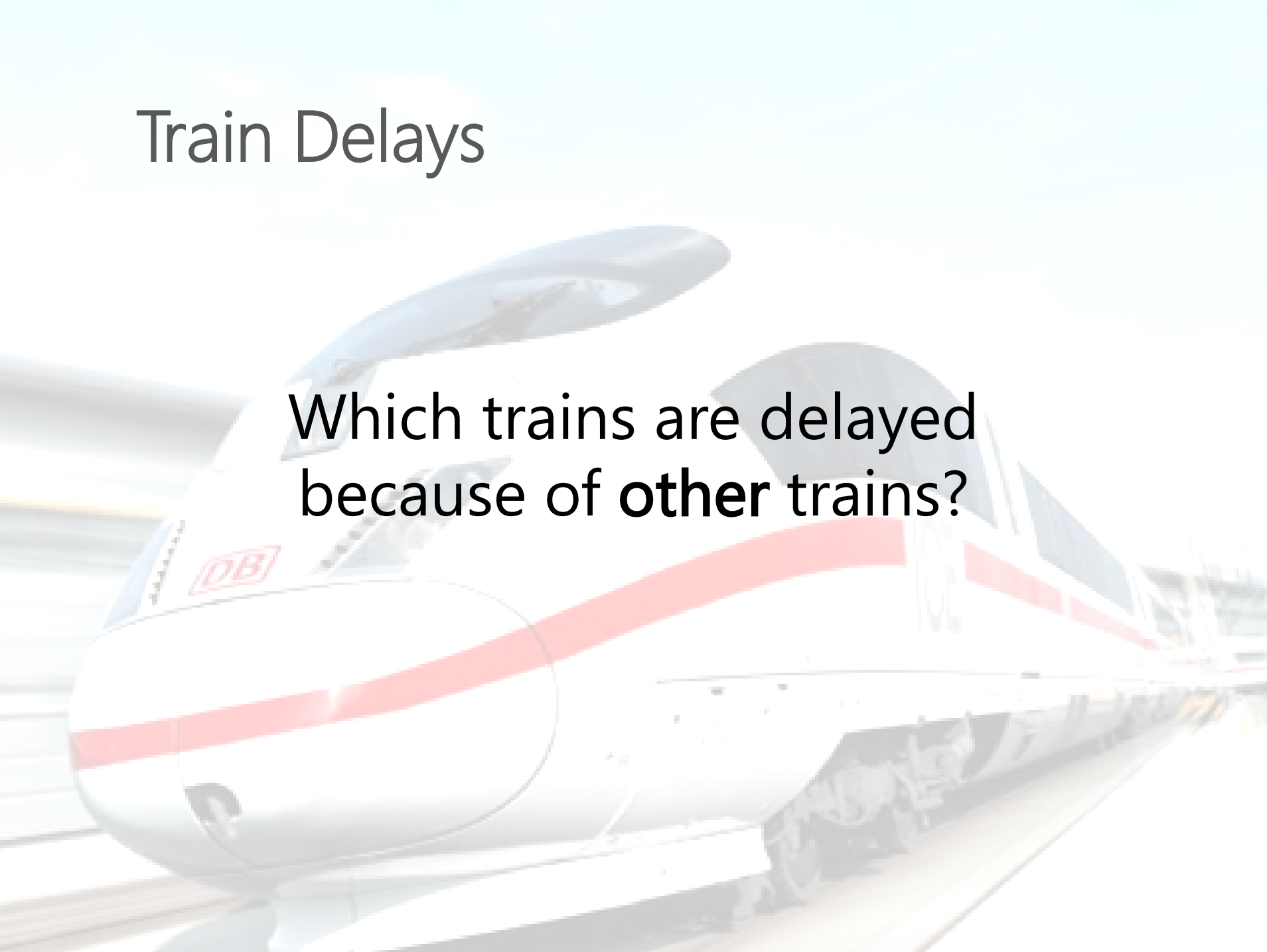

In-degree (May 99, Oct 99) dist
In-degree (May 99)
Power law, exponent 2.09
In-degree (Oct 99)
Power law, exponent 2.09
number of pages
in-degree
1e+10
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
100000


Out-degree (May 99, Oct 99) distr.
number of pages
out-degree
1e+10
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
1000

Shopping Data

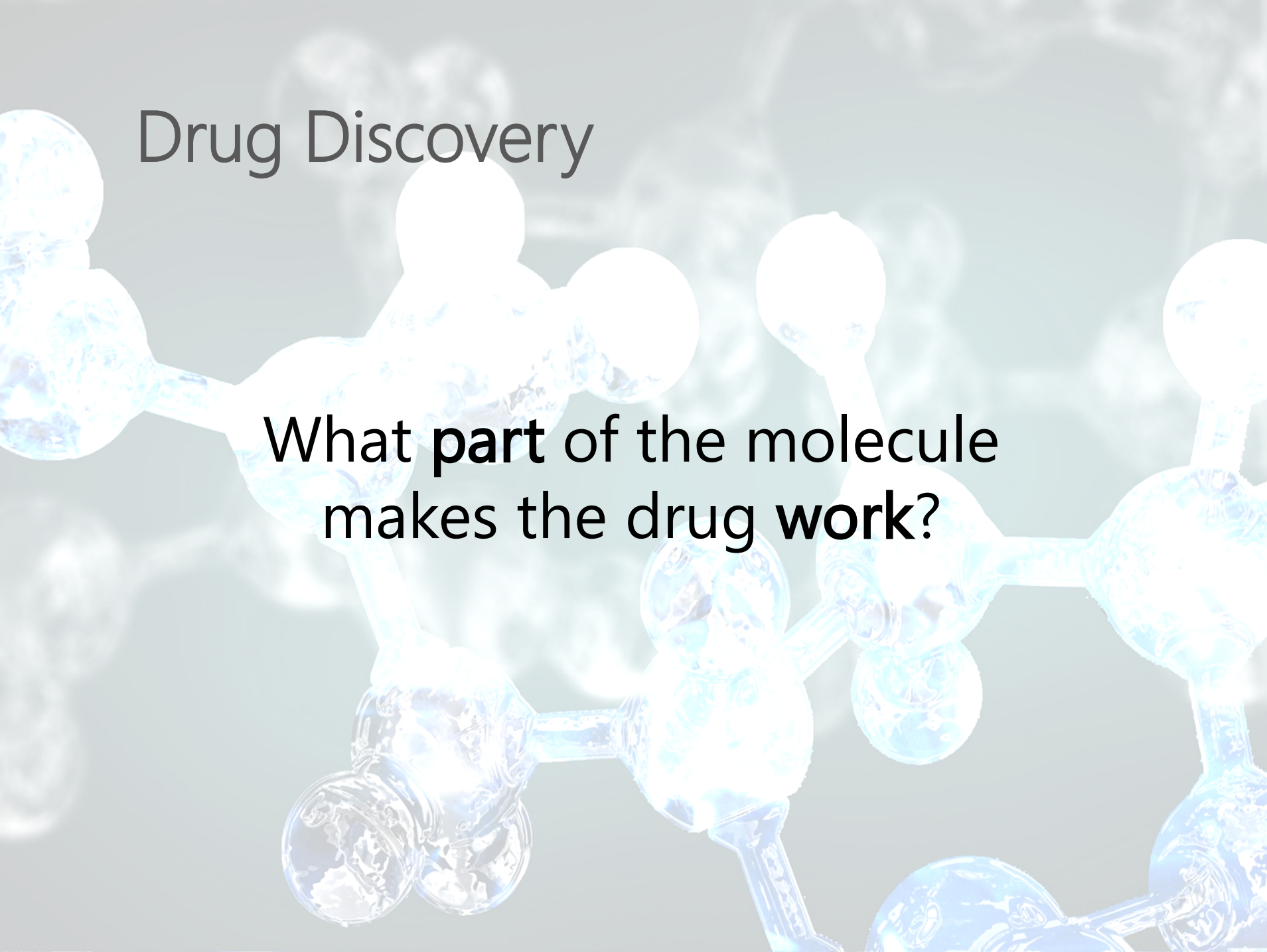
Which products are often bought **together**?

Train Delays



Which trains are delayed because of **other** trains?

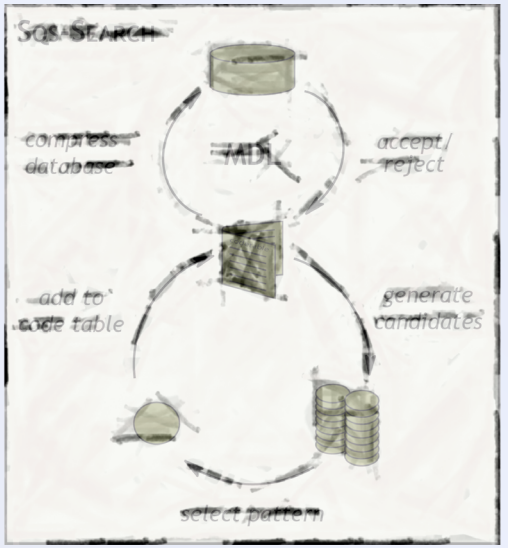
Drug Discovery



What **part** of the molecule makes the drug **work**?


More patterns
than you
can shake
a stick
at

Pattern-based Modelling



Mining Algorithm

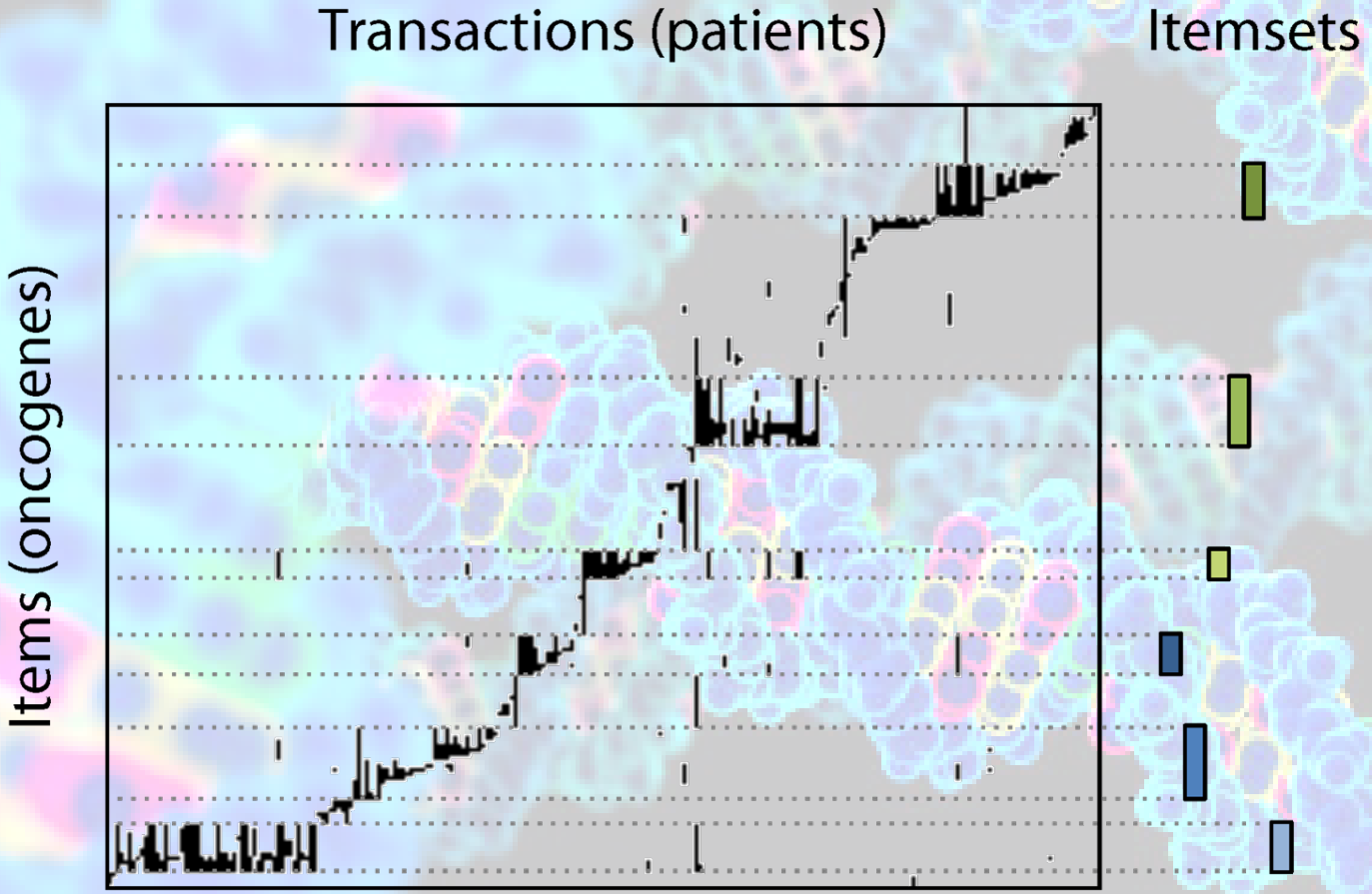
support vector machin svm
associ rule mine
nearest neighbor
frequent itemset mine
naïv bay
linear discrimin analysi lda
cluster high dimension
state art
frequent pattern mine
algorithm
synthet real

summary of JMLR abstract database

Summarising

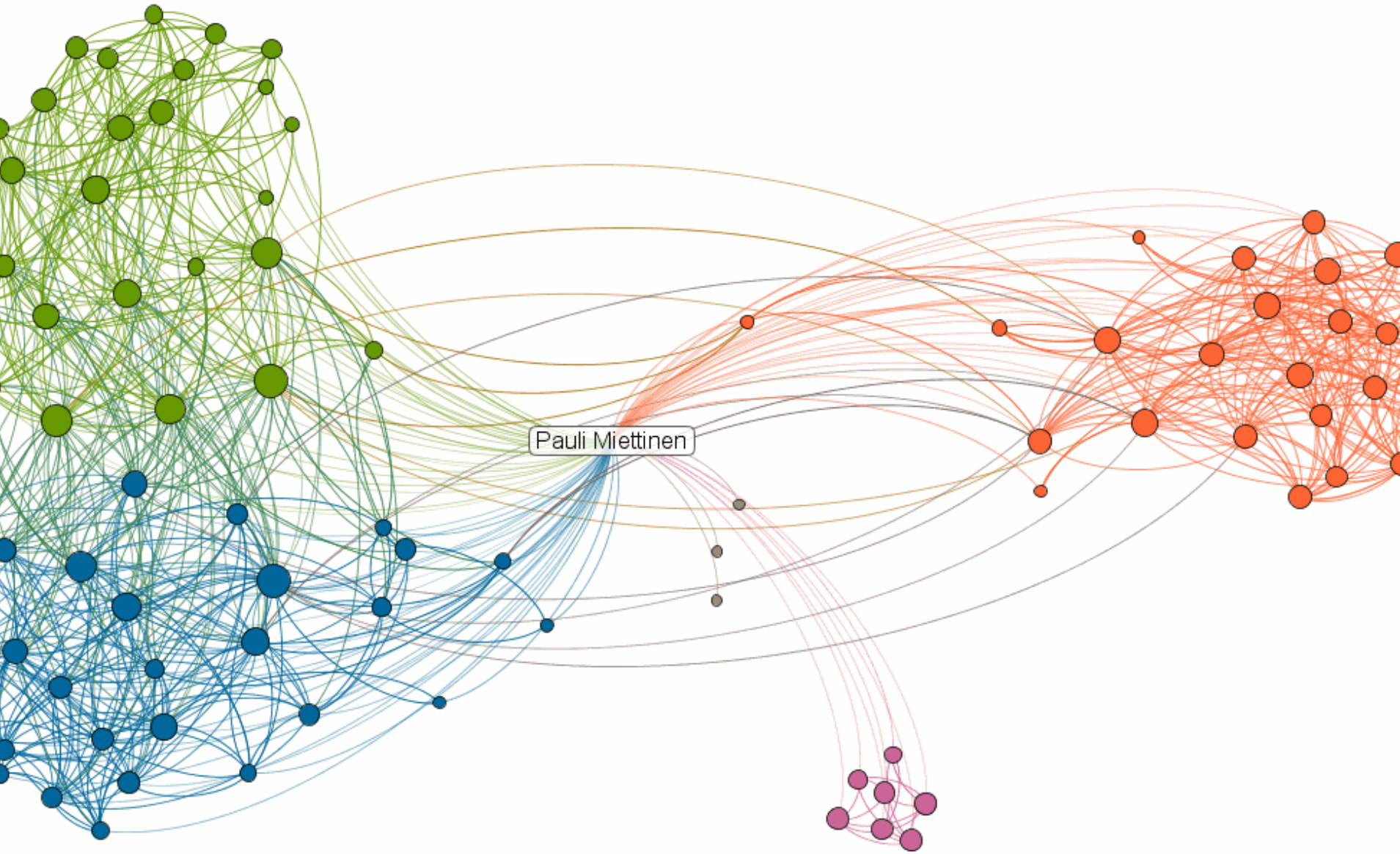
Which sales **characterise** your customers?

Summarising

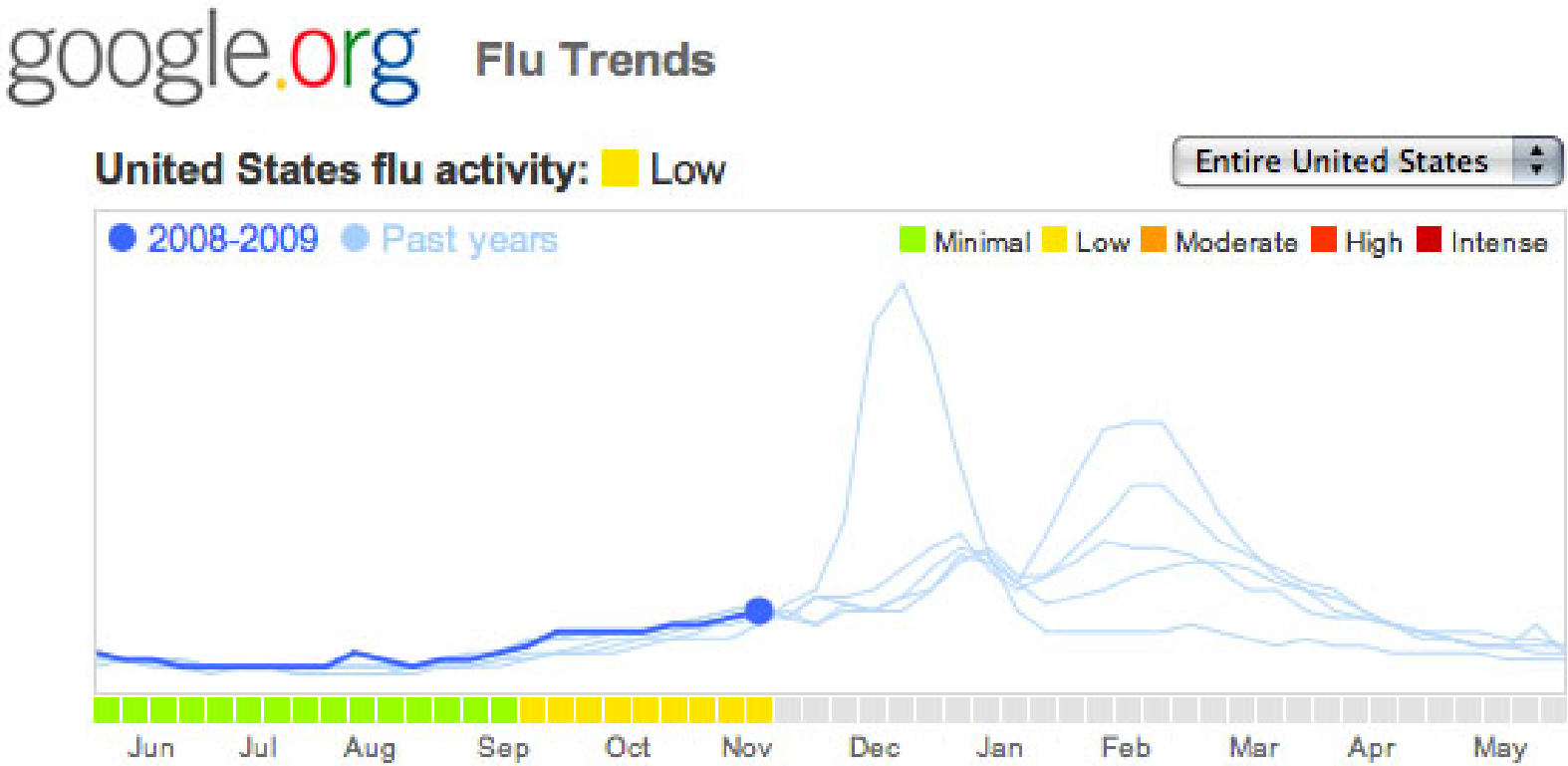


Pauli Miettinen's Professional Network

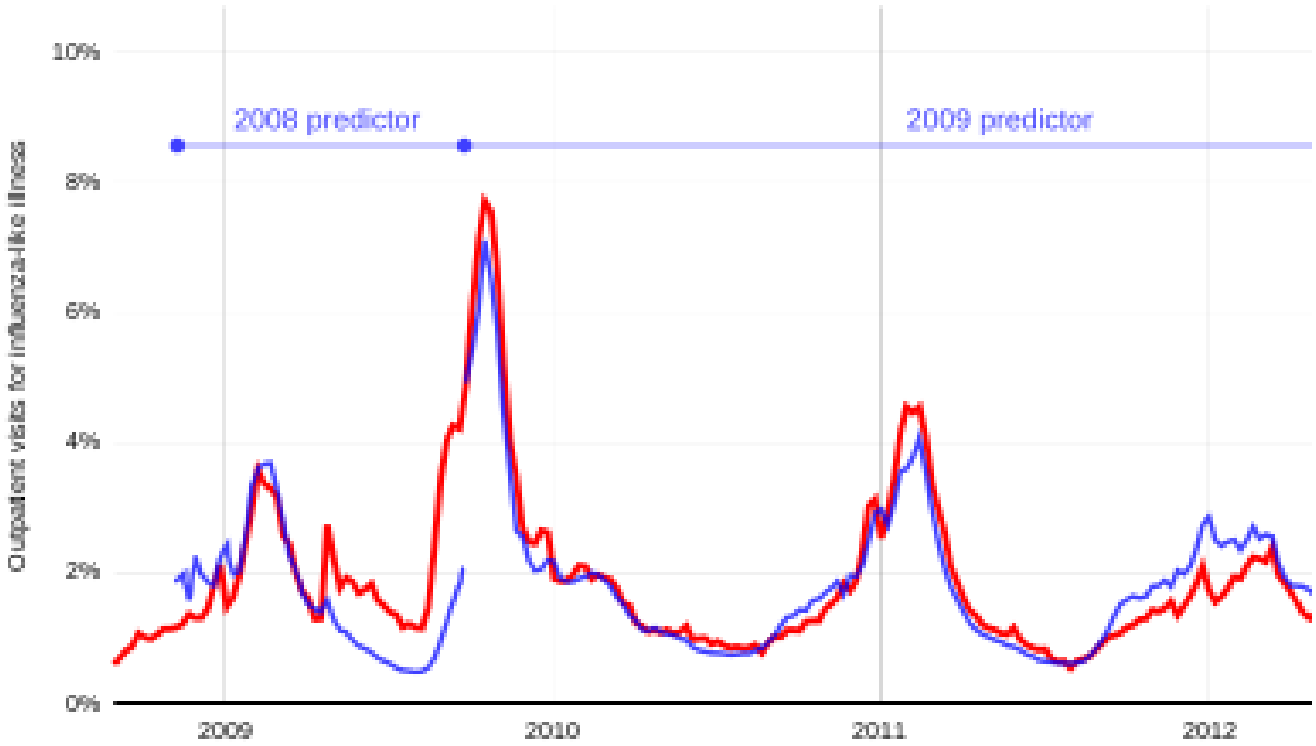
as of January 12, 2014



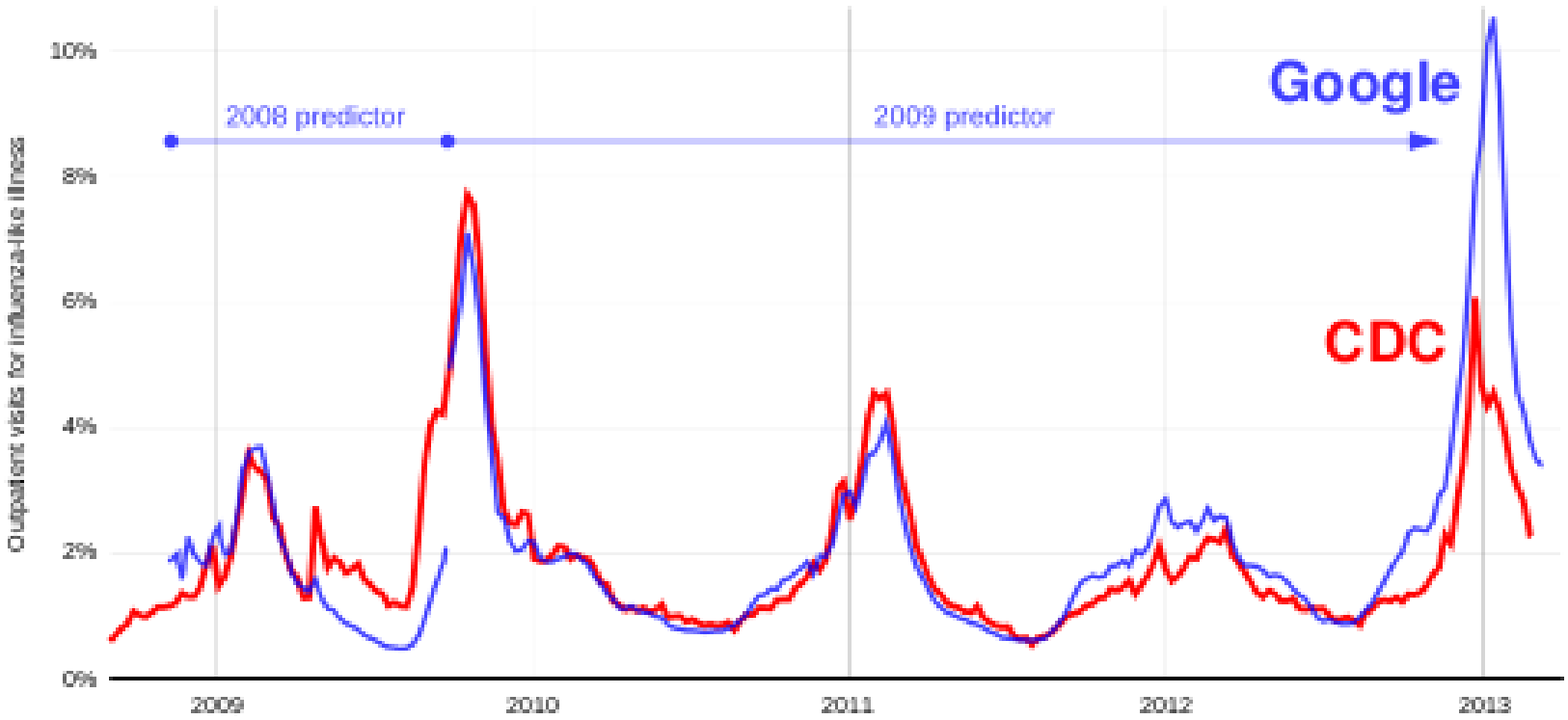
Google Flu


google.org
Flu Trends
United States flu activity: Low
Entire United States
2008-2009
Past years
Minimal
Low
Moderate
High
Intense
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May

Quite Healthy



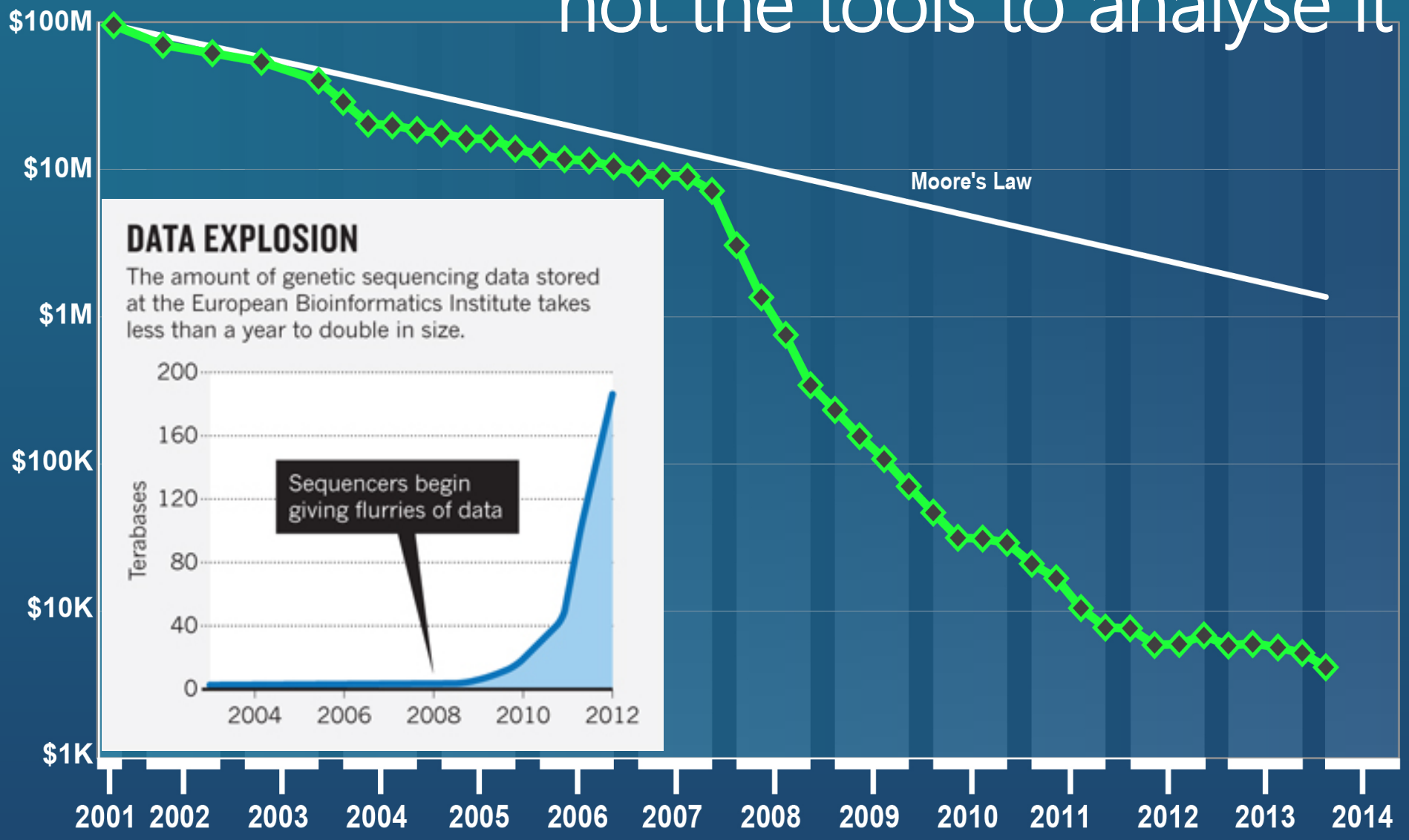
Patient Deceased



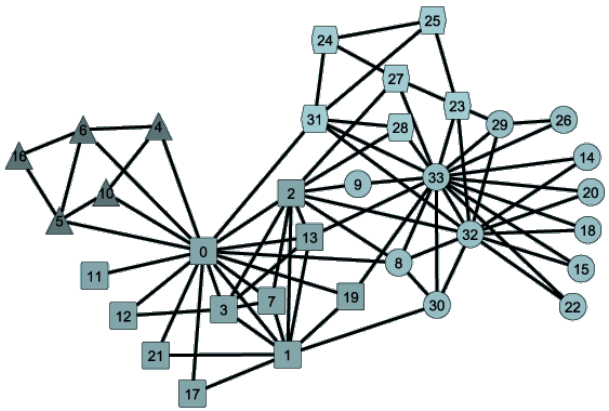

BIG
DATA

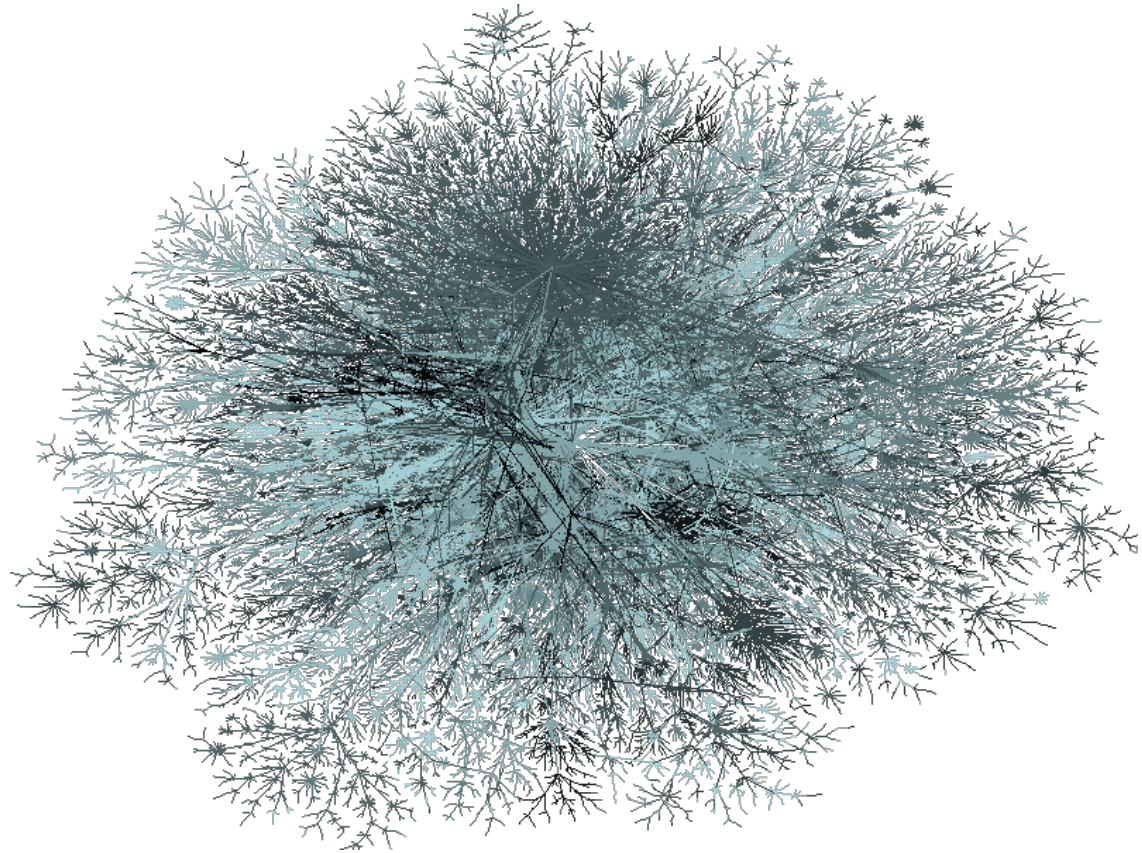
No model is perfect

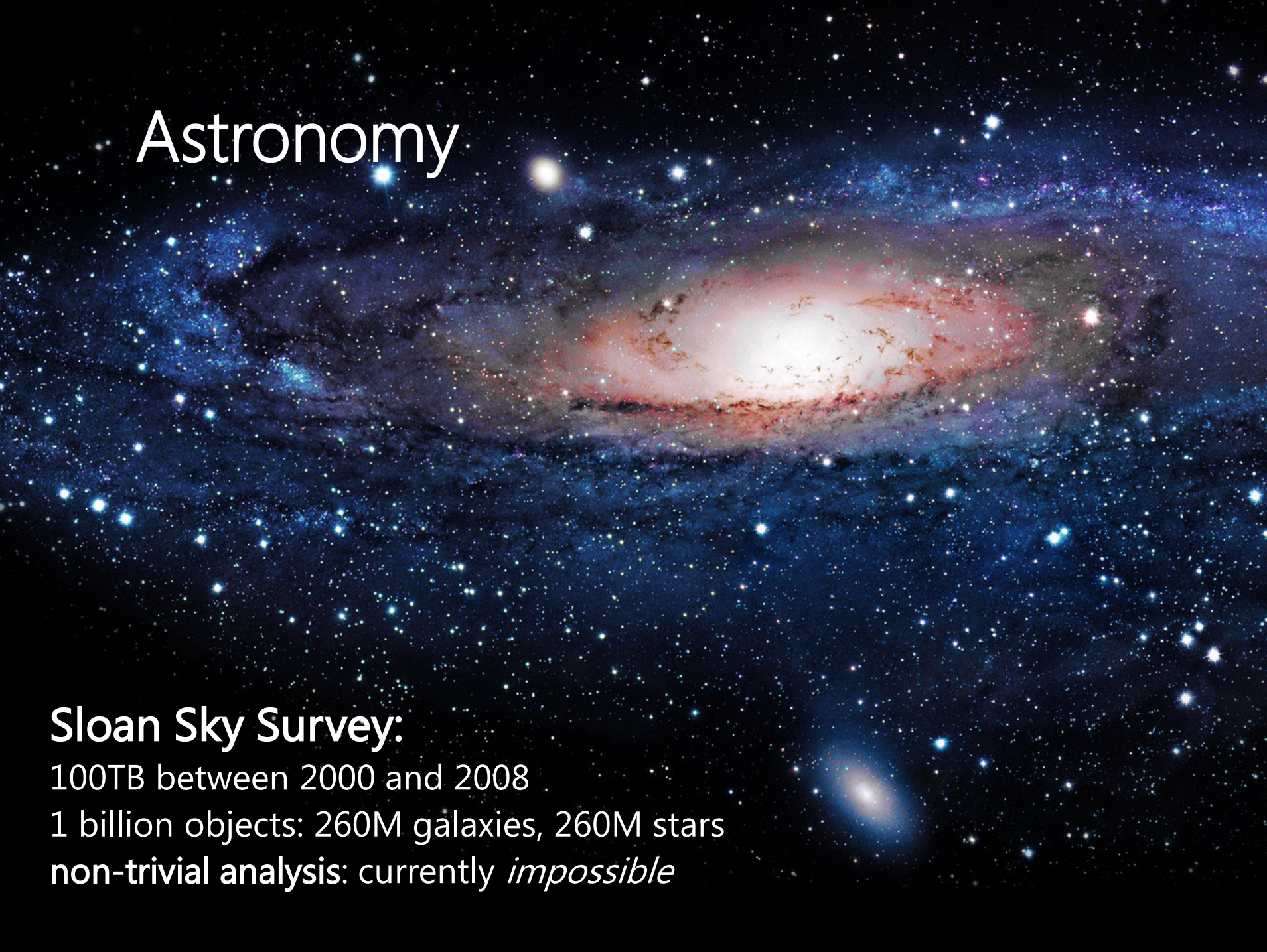
Science has lots of data, not the tools to analyse it



Social Science & the Web






Astronomy
Sloan Sky Survey:
100TB between 2000 and 2008
1 billion objects: 260M galaxies, 260M stars
non-trivial analysis: currently *impossible*


We Can Do It!