Big Data for Radiation Analytics

Prof. Tim Weninger

Data Science Group

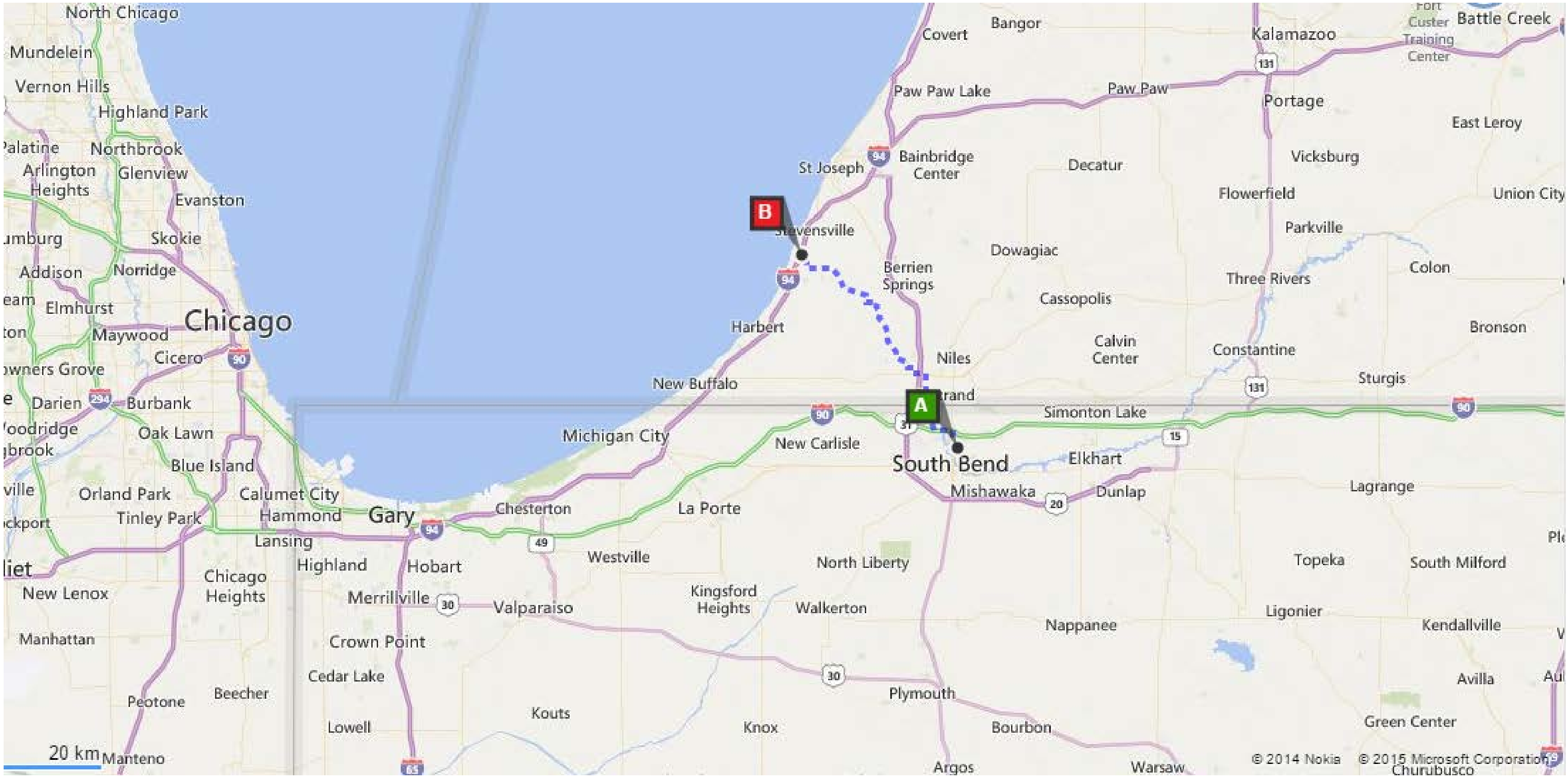
University of Notre Dame

ISOE Symposium
Urbana, IL, USA

Who am I

I am a professor of computer science and engineering

UNIVERSITY OF NOTRE DAME



What do I do?

I am a data scientist

- I find ways to use data to inform and improve tasks
  - I use copious amounts of data

I specialize in

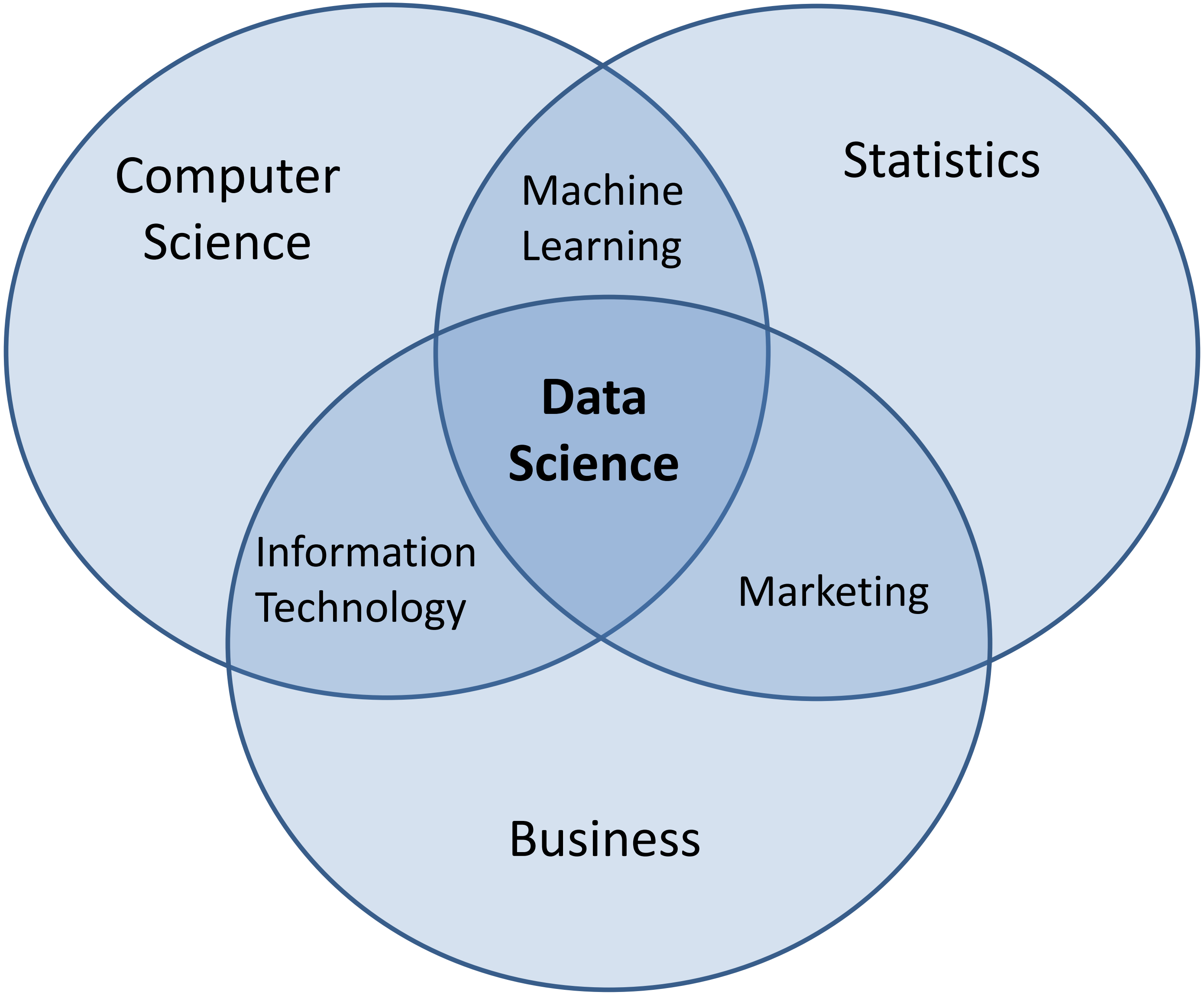
Large Scale Information Network Analysis

What am I doing here?

It’s -30 degrees in South Bend, IN

Convince you that data science may be able to help

First some background in data and info. science



Databases came first

Relational databases

Customer ID	Tax ID	Name	Address	[More fields]
1111111	441-1122	Smith, John Jr.	501 Sunnyvale	…
2222222	551-2211	Hite, Robert	401 W. 1st St	…

Tax ID	Year	Total kWh
441-1122	2011	13050
441-1122	2012	14010

Databases came first

Transactional databases

Customer ID	Acct No	Name	Address	[More fields]
1111111	626-11-2402	Smith, John Jr.	501 Sunnyvale	…
2222222	727-44-9080	Hite, Robert	401 W. 1st St	…

Acct No	TransactionID	Time	Amount
626-11-2402	00001	0630 12252012	+1000.00
626-11-2402	00002	0631 12252012	-5.00
626-11-2402	00003	1410 12262012	-15.00

Enabled easy accounting

- Lots of accountants were laid off
- Lots of IT guys were hired

We started to analyze the data

Find patterns, trends in the data

Data Cube

- Slice, Dice, Rollup, Drilldown on data

Association Rule Mining

- Find dependencies between transactions

Clustering

- Group similar items together

Classification

- Determine which labeled class an item belongs to

Together this is generally referred to as **Data Mining**

- Originally called Knowledge Discovery in Databases

Association Rule Mining

Find patterns, trends in the data

Some examples:

Supermarket –

{Onions, Ketchup, Buns} -> {Hamburger}

{Diapers} ->

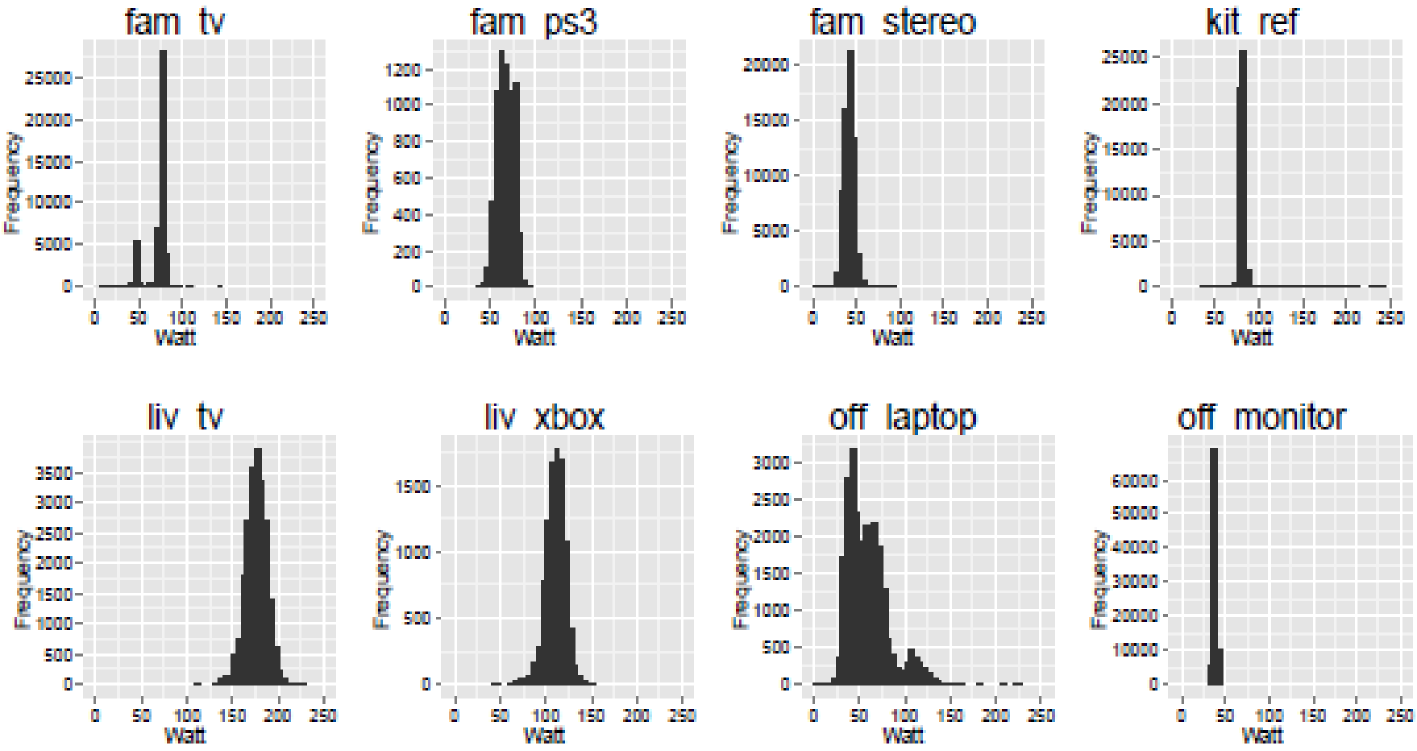
Something to think about:

Why are bread and milk in the back of the store?

Robbers?

Association Rule Mining (Example)

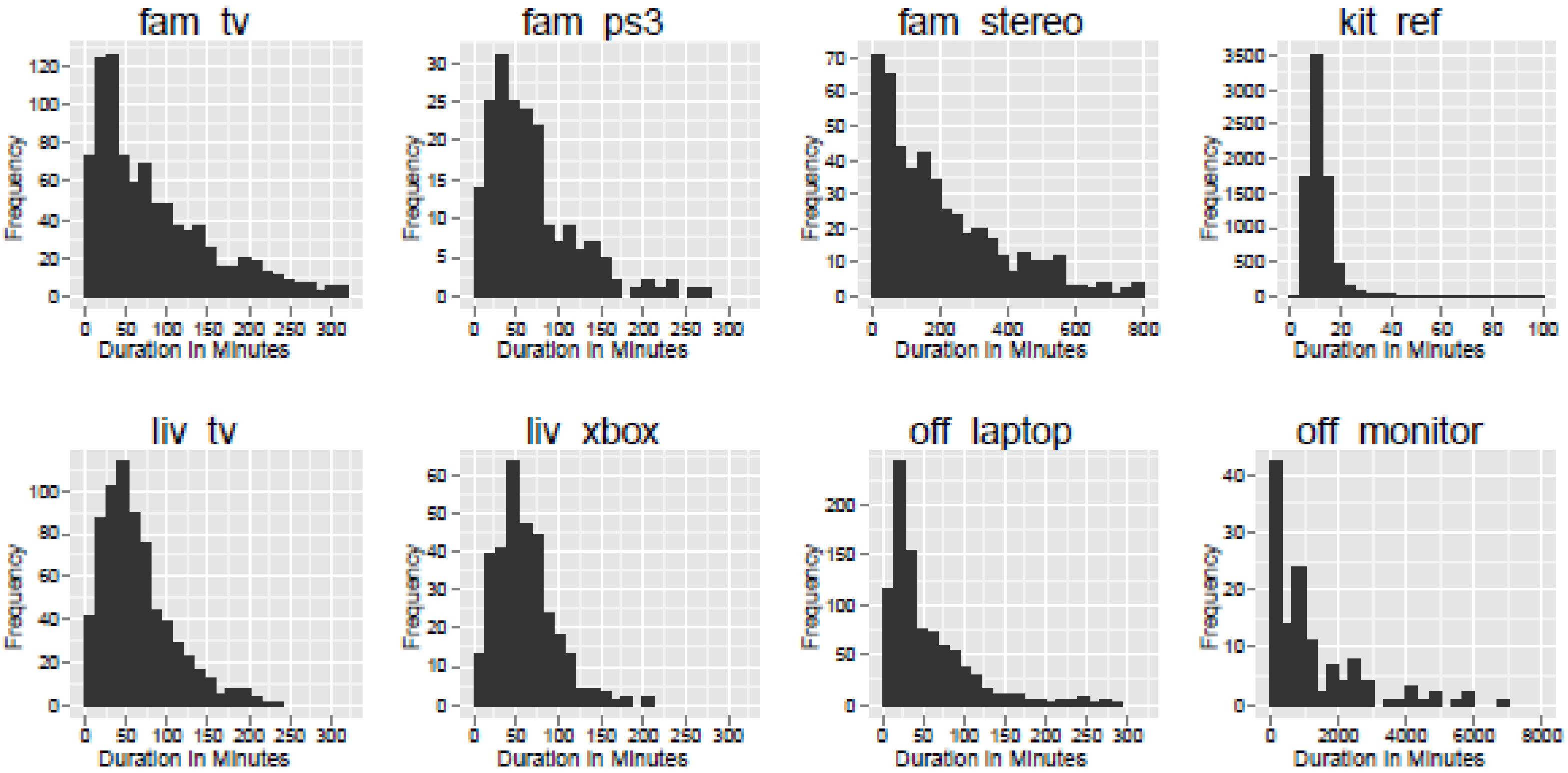
Residential Power Disaggregation



Histogram of Power Consumption

Association Rule Mining

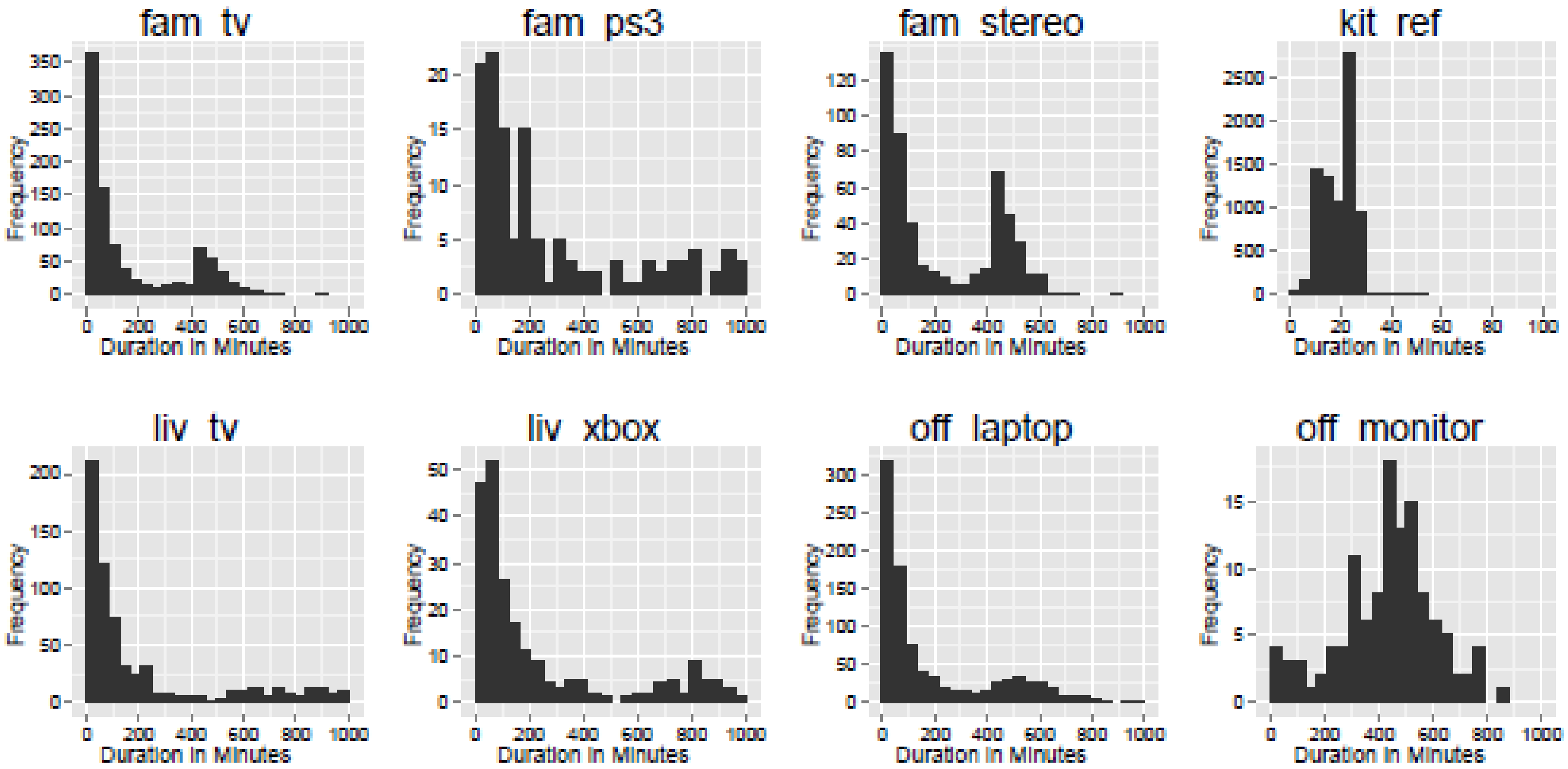
Transaction database (ON)



Histogram of ON durations

Association Rule Mining

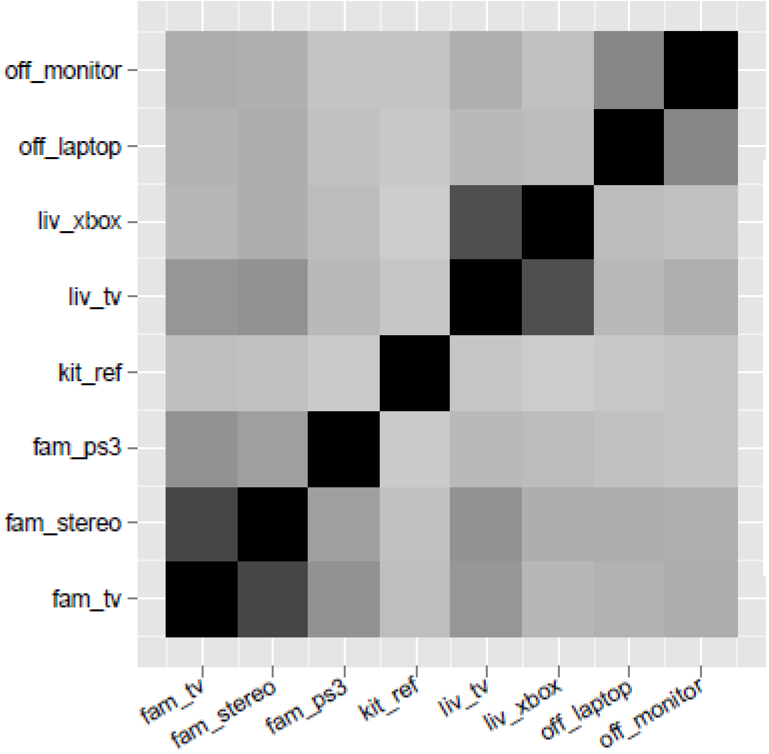
Transaction database (OFF)



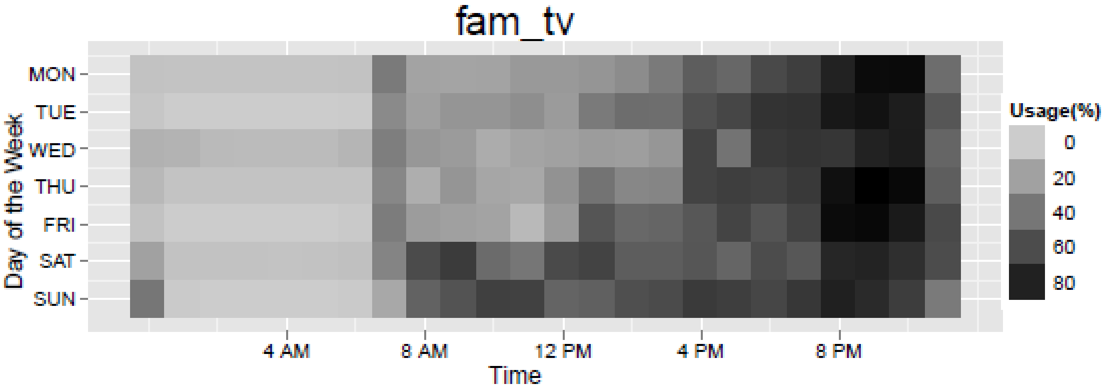
Histogram of OFF durations

Association Rule Mining

Correlation and usage pattern mining



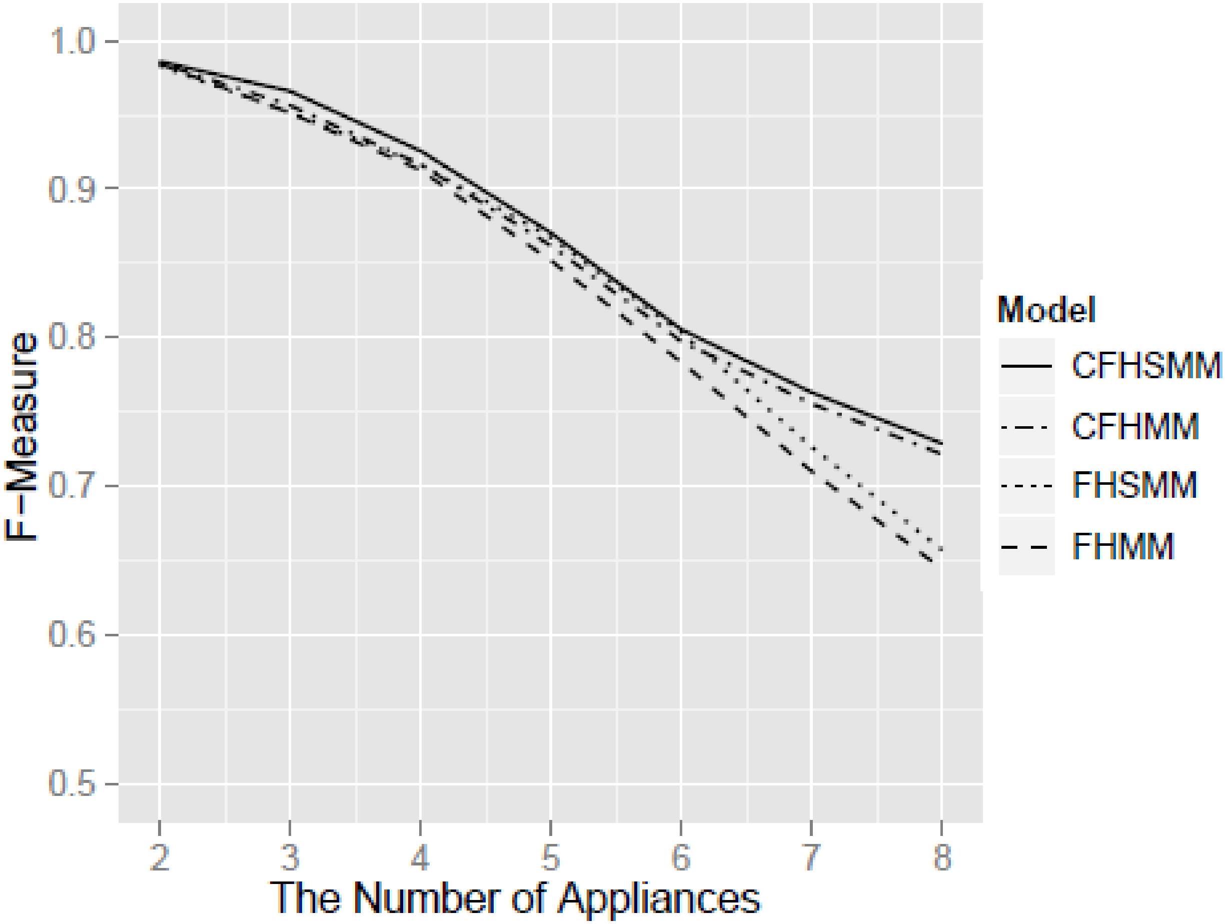
Appliance Correlation Matrix



Usage patterns

Association Rule Mining

Can we determine which appliances are running?



Clustering

Group similar items together

Based on a similarity measure

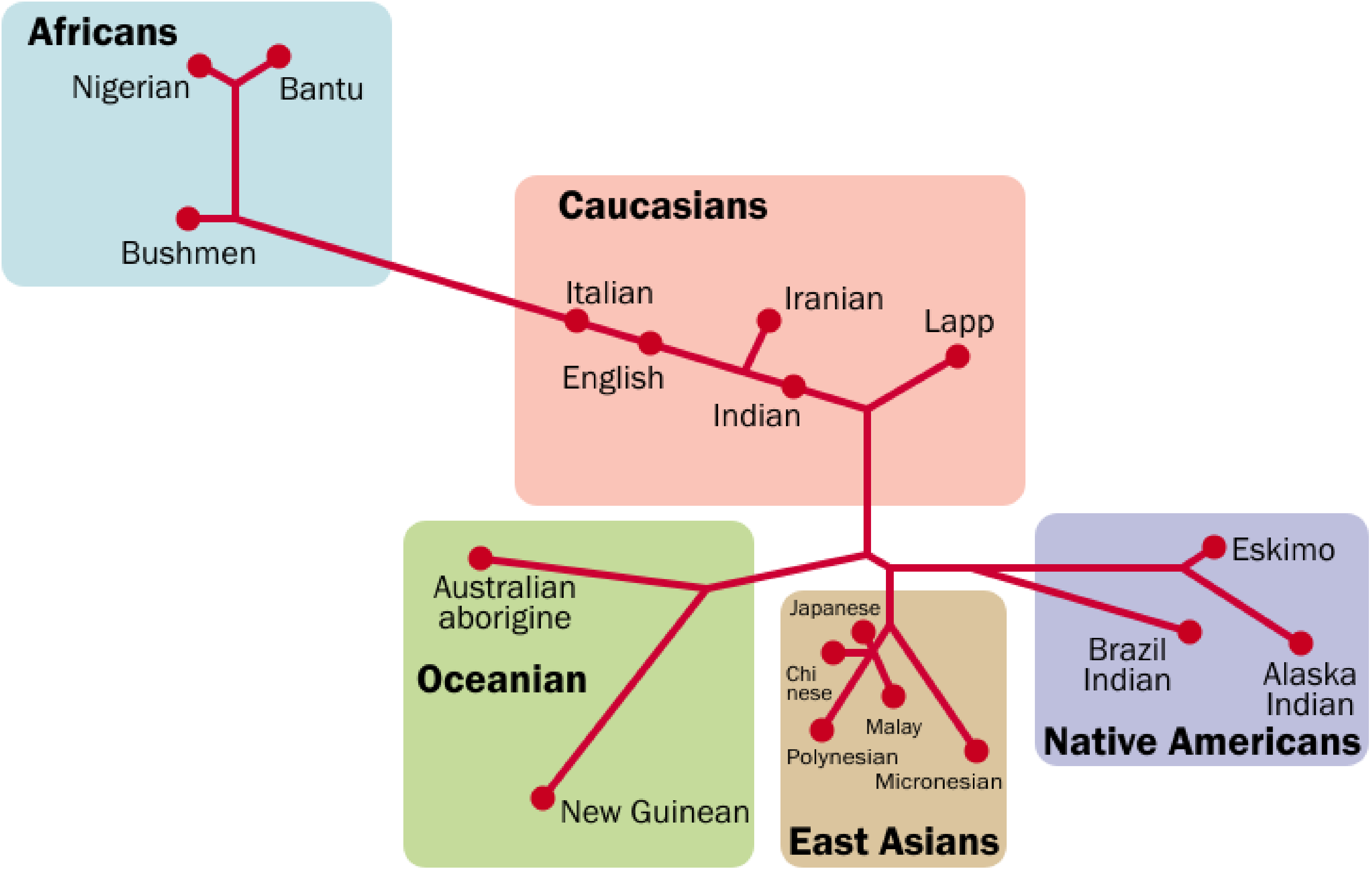
- Time, frequency, geography, anything else, combination of anything.
  - Literally countless similarity measures

The “Google Algorithm” is a similarity measure

How similar is the query terms to the Web page?

Thousands of clustering algorithms

Clusters of DNA Sequences



Occupational Exposure Clusters

Cancer clusters

- Determined when a greater-than-expected number of cancer cases are found in a region, occupation, etc.

How do we find cancer clusters?

- Erin Brockovich
  - Hexavalent Chromium
- CDC, EPA, HHS
  - They get lots and lots of cancer data points
  - Analysts use clustering tools to wrangle the statistics

Epidemiology

- Integrated Information Analytics Center (IIAC)

Classification

Given a set of examples, find the class/group to which a new item belongs

Also a thousand different classification algorithms

- No free lunch theorem

Based on features!

- Humans have to tell the program what to look for

What are features?

Customer ID	Acct No	Name	Address	[More fields]

Differential Diagnosis in Medicine

Complaint	Complaint #2	Body Temp.	Area	Duration	Diagnosis
Runny Nose	Coughing	101.6	Head	3 days	Cold
Aching	Nausea	103.2	Body	4 days	Flu
Runny Nose	Coughing	101.5	Head	3 days	Cold
Runny Nose	Coughing	102.1	Head	6 days	?? **Cold**

Runny Nose	Coughing	98.4	Head	6 days	Allergies

Not enough training data

Spam Filtering

From	Subject	Text	Spam
john@gmail	We need to talk	Give me a call sometime and we can	No
dave@yahoo	Enlarge your penis	Cheap viagra…	Yes
george@im.x	I’d like to meet	Give me a call sometime and we can	??
tim@nd.edu	ISOE talk	Hi, I am planning to give a talk at t…	??

Features are so very important

From	Have I emailed sender?	Private Account?	Similar emails in system from same sender?	Subject	Text	Spam
george@im.x	No	No	Yes Clustering	We need to talk	Give me a call sometime and we can…	Yes
tim@illinois	Yes	Yes	No	ISOE talk	Hi, I am planning to…	No

Data Driven Business Processes

Companies often have lots of data

- Companies rightfully guard their data as trade secrets.

But they often ask

“I have all this data, but I don’t know what to do with it?”

CEO reads a magazine or a case study

…and begins making mistakes

Let’s talk about what data science can (and cannot) do

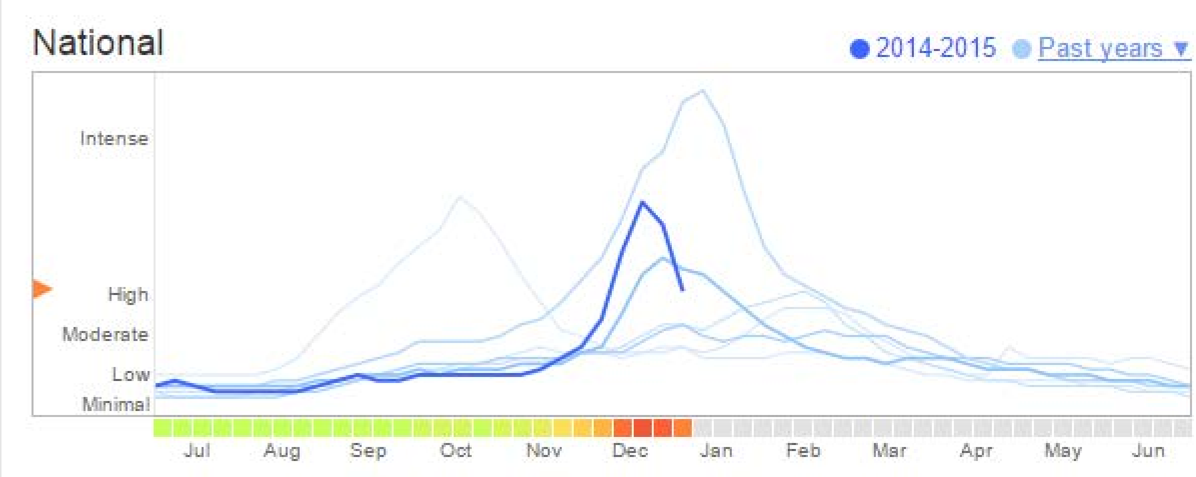
Danger!

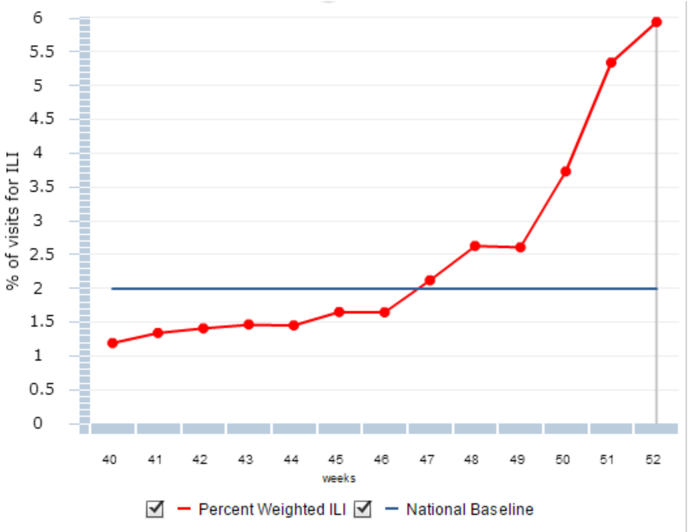
“My responses are limited, you must ask the right question”

Dr. Alfred Lanning - iRobot – 20th Century Fox – 2004

Case study in not doing the right thing

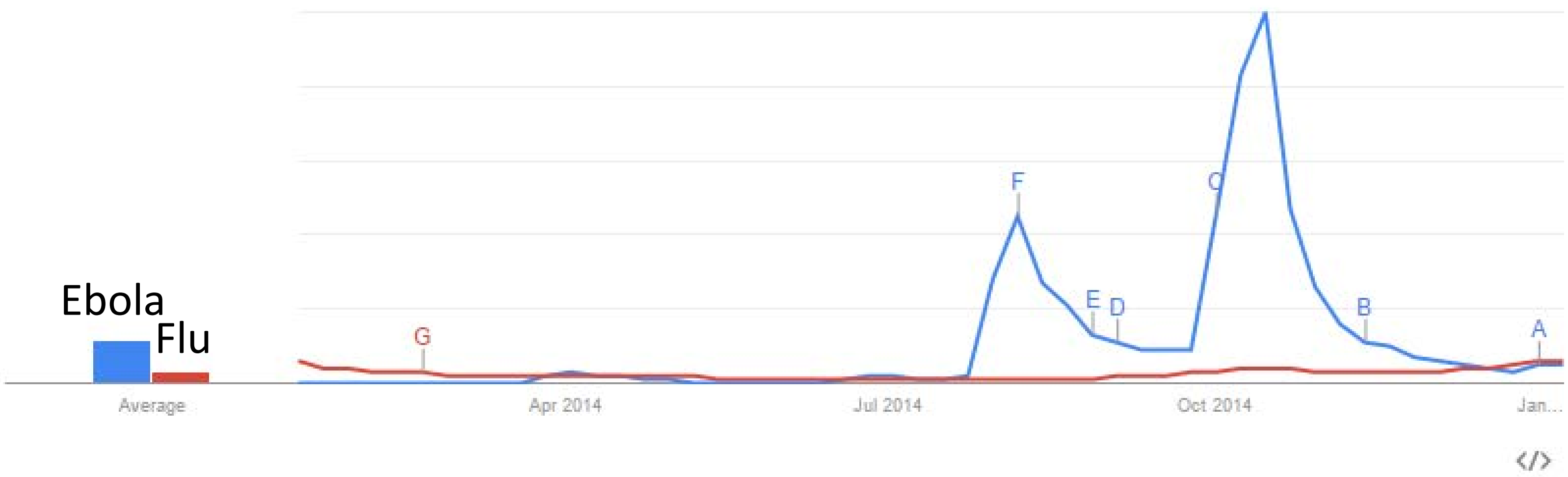
Google Flu Trends





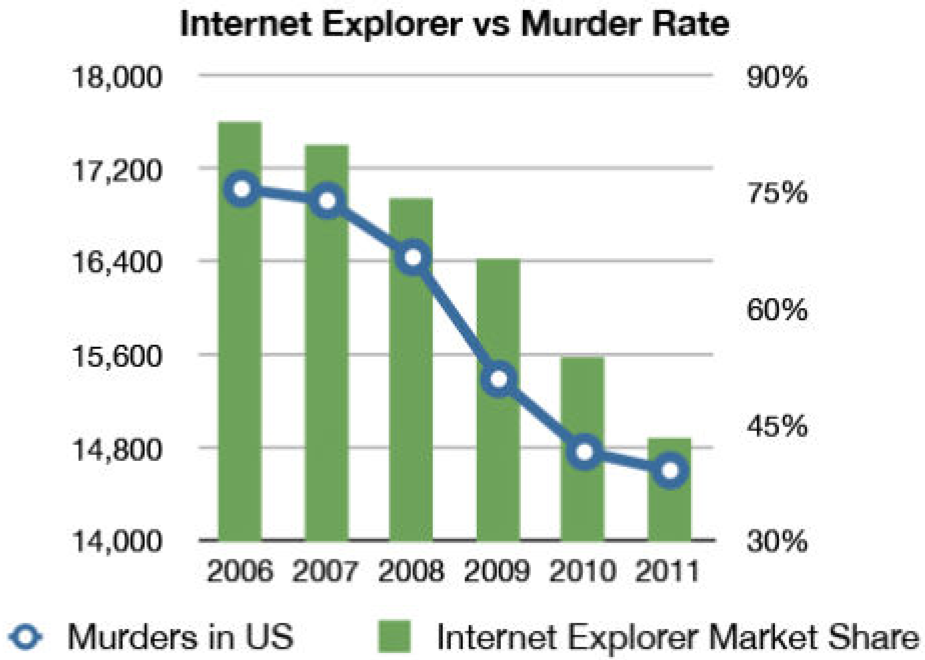
Case study in not doing the right thing

Ebola in America



Case study in not doing the right thing

Correlation != Causation



In general…

Hiring a good Data Scientist
is like hiring an Electrician:

Good Practices

Capture as much data as you can.

- Everything
  - Disk storage is cheap (and getting cheaper)
- When in doubt, write it down

Do not just report aggregate statistics.

- Averages can’t be un-averaged
  - Data is lost

Big Data

Can we leverage all of the data from all of the Nuclear Generating Stations to lower exposure and decease outage time.

This is the promise of BIG Data

ISOE members can’t do this yet.

Why can’t Nuclear Stations use big data yet?

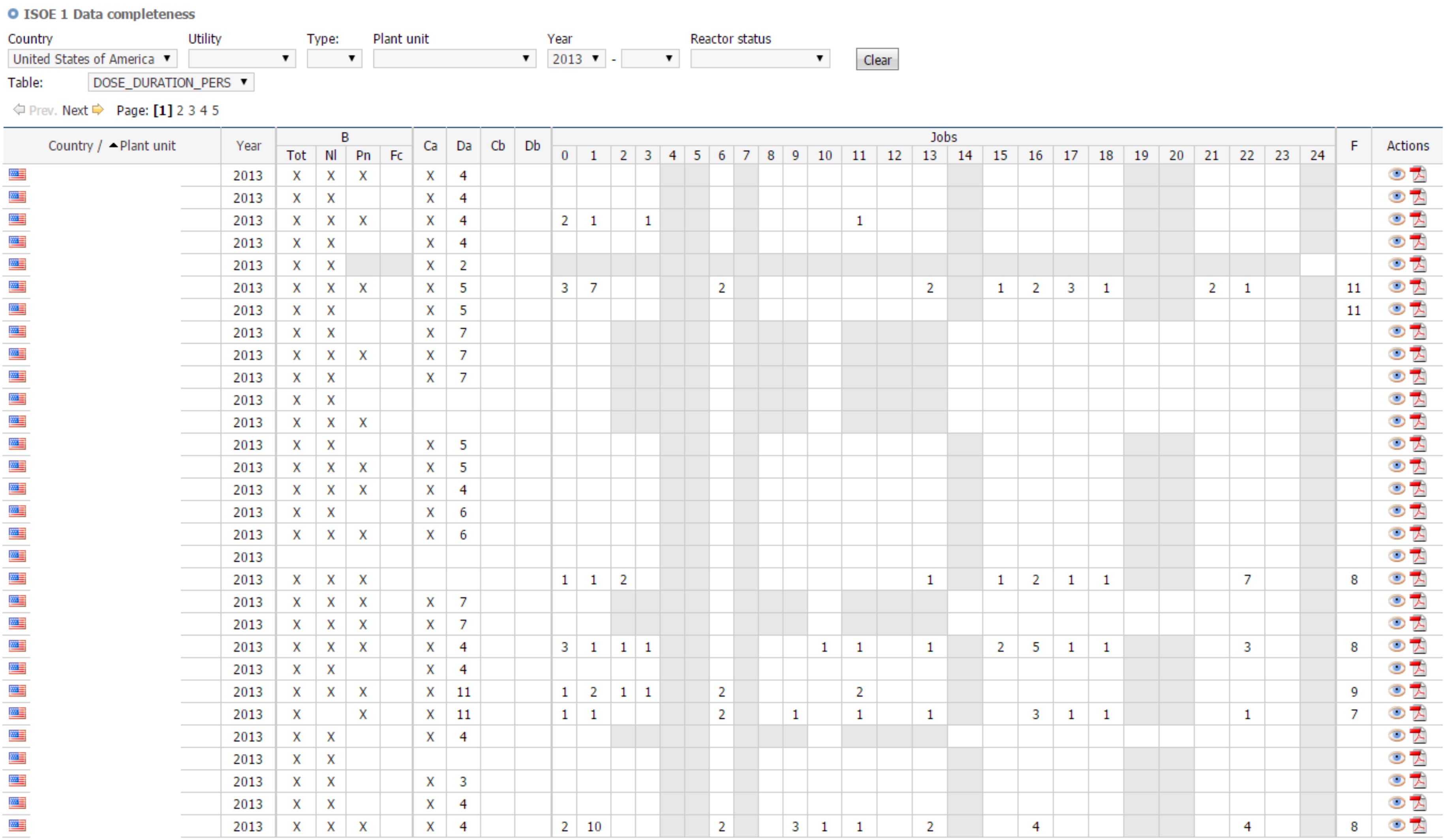
Consider CDC, Gmail spam filters, etc.

Why are they successful?

Because their data is in the same format,
in the same place.

Why can’t Nuclear Stations use big data yet?

ISOE Database is a good start… but incomplete



ISOE 1 Data completeness

Country: United States of America ▾

Utility: ▾

Type: ▾

Plant unit: ▾

Year: 2013 ▾ - ▾

Reactor status: ▾

Clear

Table: DOSE_DURATION_PERS ▾

Prev. Next Page: **[1]** 2 3 4 5

Country / ▲Plant unit	Year	B				Ca	Da	Cb	Db	Jobs																									F	Actions
		Tot	NI	Pn	Fc					0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24		
	2013	X	X	X		X	4																													
	2013	X	X			X	4																													
	2013	X	X	X		X	4			2	1		1								1															
	2013	X	X			X	4																													
	2013	X	X			X	2																													
	2013	X	X	X		X	5			3	7					2							2		1	2	3	1			2	1			11	
	2013	X	X			X	5																												11	
	2013	X	X			X	7																													
	2013	X	X	X		X	7																													
	2013	X	X			X	7																													
	2013	X	X																																	
	2013	X	X	X																																
	2013	X	X			X	5																													
	2013	X	X	X		X	5																													
	2013	X	X	X		X	4																													
	2013	X	X			X	6																													
	2013	X	X	X		X	6																													
	2013																																			
	2013	X	X	X						1	1	2											1		1	2	1	1				7			8	
	2013	X	X	X		X	7																													
	2013	X	X	X		X	7																													
	2013	X	X	X		X	4			3	1	1	1							1	1		1		2	5	1	1				3			8	
	2013	X	X			X	4																													
	2013	X	X	X		X	11			1	2	1	1			2					2														9	
	2013	X		X		X	11			1	1					2			1		1		1			3	1	1				1			7	
	2013	X	X			X	4																													
	2013	X	X																																	
	2013	X	X			X	3																													
	2013	X	X			X	4																													
	2013	X	X	X		X	4			2	10					2			3	1	1		2			4						4			8	

This is difficult

CDC has federal law requiring reporting

Spam filtering is very expensive for Google.

I challenge you to report as much data as you can

Imagine if we had all the data…

Occupational Exposure

- Can we predict the mRem exposure of a task?
- Is a repetitive task chronically over/under the estimate exposure budget?
- Extra Credit - Can we reduce the individual and overall mRem exposure?

Outage Management

- Can we predict the estimated duration of a task?
- Which tasks are chronically over/under the estimated duration ?
- Extra Credit – Can we reduce the total outage duration?

Thank you