

A logician-engineer's adventures in data science and analytics

Wong Limsoon



About Limsoon



Position

**Kwan-Im-Thong-Hood-Cho-Temple Chair Professor,
Dept of Computer Science, NUS**

Research

**database systems & theory, knowledge discovery,
bioinformatics & computational biology**

Honours

- **ACM Fellow**
- **FEER Asian Innovation Gold Award 2003**
- **ICDT Test of Time Award 2014**

Plan

- **Part 1: Helpful analytics**
 - Tactics to make data analysis more insightful
- **Part 2: Generating hypothesis**
 - Exploratory hypothesis testing and analysis as a datamining task
- **Part 3: Technical details of iDIG**
 - Under-the-hood look of the intelligent Data-driven Insight Generator, iDIG
- **Part 4: Art of data analysis**
- **Part 5: Science of data analysis**

Part 1: Helpful analytics



The gist of helpful analytics



- **Make it easy to formulate hypothesis**
 - Extraction from big, integrated databases
- **Make hypothesis testing sound**
 - Detection & correction of assumption violations
- **Find better hypothesis & explain why it is better**
 - E.g., “for men, taking A is better than B”

Example

REVEALING SAMPLE BIAS

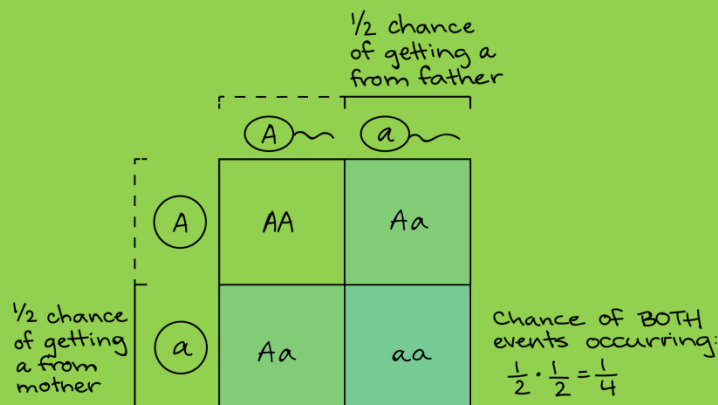
SNP	Genotypes	Group				χ^2	P value
		Controls [n(%)]		Cases [n(%)]			
rs123	AA	1	0.9%	0	0.0%	4.78E-21 ^b	
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

A seemingly
obvious
conclusion

- **A scientist claims the SNP rs123 is a great biomarker for a disease**
 - If rs123 is AA or GG, unlikely to get the disease
 - If rs123 is AG, a 3:1 odd of getting the disease
- **A straightforward χ^2 test. Anything more/wrong?**

Sample bias is revealed by domain logic



Basic rule of human genetics

		Group					
SNP	Genotypes	Controls [n(%)]		Cases [n(%)]		χ^2	P value
rs123	AA	1	0.9%	0	0.0%		4.78E-21 ^b
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

- **AG = 38 + 79 = 117, controls + cases = 189 \Rightarrow population is ~62% AG \Rightarrow population is >9% AA, unless AA is lethal**
- **“Big data check” shows AA is non-lethal for this SNP \Rightarrow sample is biased**

Example

FINDING EXCEPTIONS & CONTRADICTIONS

	Income >50K	Income \leq 50K	Total
Adm-clerical	439 (14.2%)	2645 (85.8%)	3084
Craft-repair	844 (22.8%)	2850 (77.2%)	3694
Total	1283	5495	6778

A seemingly
obvious
conclusion

- The data shows that, in Australia, craft repairers tend to earn more than administrative clerks
 - 23% of the former vs 14% of the latter has high income
- A straightforward χ^2 test. Anything more/wrong?

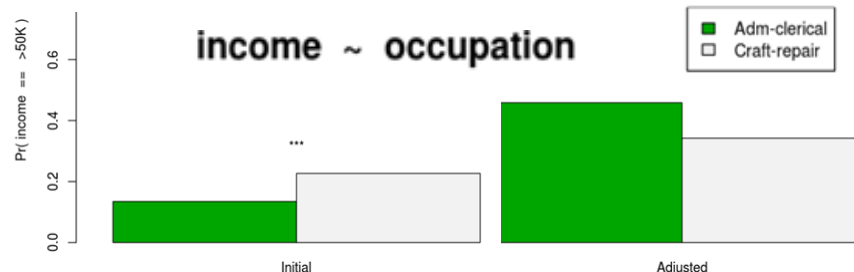
Exceptions & contradictions are found by context mining

{ <u>Workclass</u> = Self-emp-not-inc }	Income >50K	Income ≤50K	Total
<u>Adm-clerical</u>	16 (34.8%)	30 (65.2%)	46
<u>Craft-repair</u>	90 (18.0%)	409 (82.0%)	499
Total	106	439	545

- The “unincorporated self-employed” work class is an exception to the conclusion

- And the conclusion holds for neither male nor female!
- In fact, after detecting & adjusting for possible confounding factors, the conclusion is the opposite!

{Sex = Male}	Income >50K	Income ≤50K	Total
<u>Adm-clerical</u>	251 (24.2%)	787 (75.8%)	1038
<u>Craft-repair</u>	829 (23.5%)	2695 (76.5%)	3524
Total	1080	3482	4562



{Sex = Female}	Income >50K	Income ≤50K	Total
<u>Adm-clerical</u>	188 (9.2%)	1858 (90.8%)	2046
<u>Craft-repair</u>	15 (8.8%)	155 (91.2%)	170
Total	203	2013	2216

Example

EXTRACTING DEEPER INSIGHT

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
III	25	245	270
IV	48	212	260
V	57	233	290
Total	225	1125	1350

A seemingly
obvious
conclusion

- **Vaccines I-V are not equal in efficacy**
 - $0.001 < \chi^2 \text{ test p-value} < 0.01$ is significant
- **A straightforward χ^2 test. Anything more/wrong?**

Deeper insight can
also be dissected
without asking for
more data

Computation of the χ^2

Type of vaccines	Had flu	(O-E) ² /E	Avoided flu	(O-E) ² /E
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	25 (45.0)	8.889	245 (225.0)	1.778
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- Vaccine III contributes to the overall $\chi^2 = (8.889 + 1.778) / 16.564 = 64.4\%$



Vaccine III vs. rest

Type of vaccines	Had flu	Avoided flu	total
III	25	245	270
I, II, IV, V	200	880	1080
Total	225	1125	1350

- $\chi^2 = 12.7$ with 1 d.f.
- $P < 0.001$

χ^2 with Vaccine III removed

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
IV	48	212	260
V	57	233	290

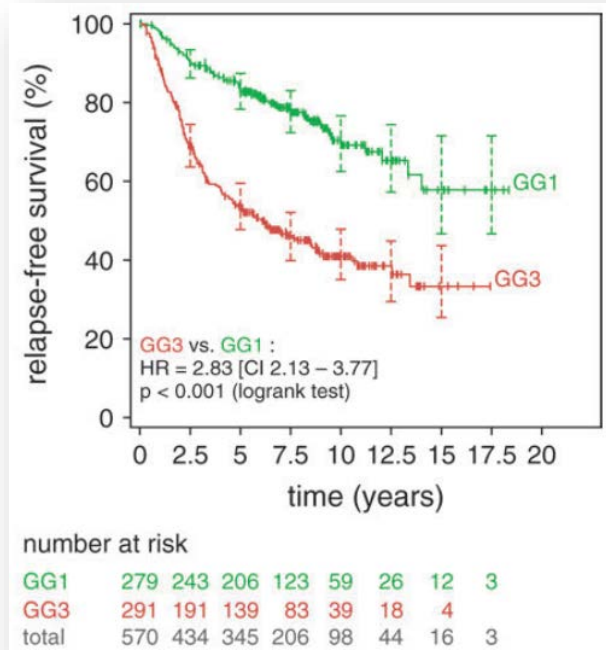
- $\chi^2 = 2.983$ with 3 d.f.
- $0.1 < p < 0.5$, not statistically significant



Vaccine III is different from / better than the rest

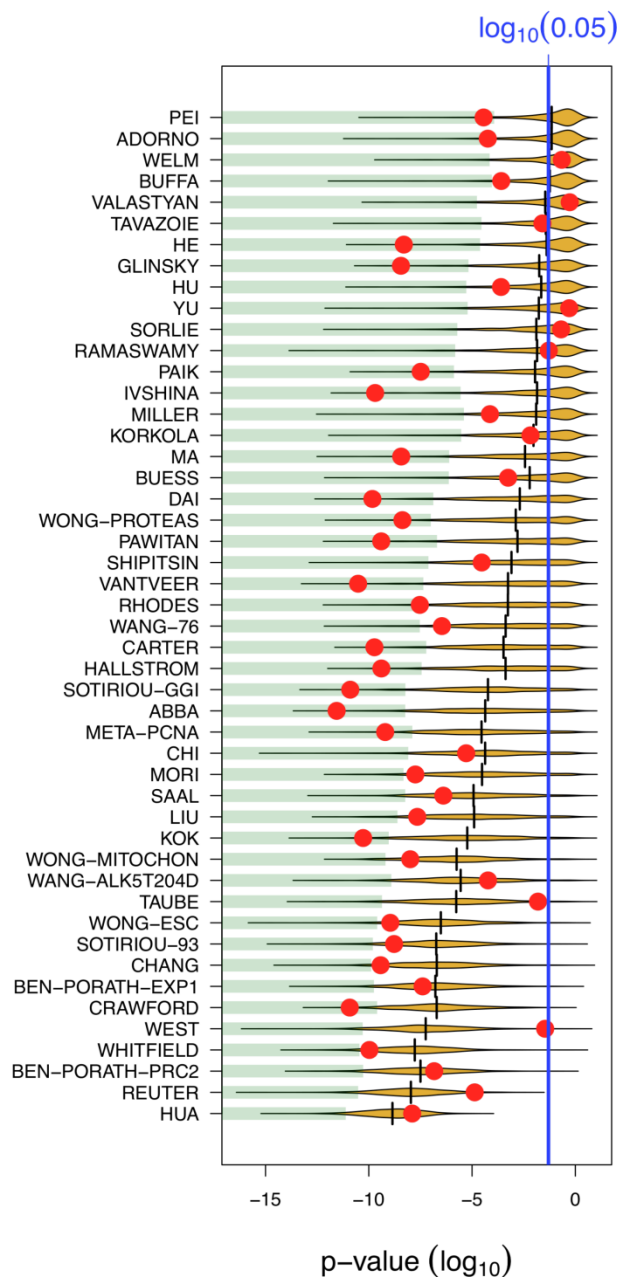
Example

DETECTING PROBLEMS IN NULL HYPOTHESIS



A seemingly
obvious conclusion

- A multi-gene signature is claimed as a good biomarker for breast cancer survival
 - Cox's survival model p-value << 0.05
- A straightforward Cox's proportional hazard analysis. Anything more/wrong?



Inappropriate null hypothesis detected by generating empirical null distribution

- Almost all random signatures also have p-value $\ll 0.05$
- ⇒ null model is confounded
- ⇒ significant signatures can't be trusted; they are no better than random ones!

Anna Karenina Principle

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

www.thequotes.in

Helpful analytics's version

- There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest
- A helpful analytics system helps the user find this one way

- So far, we have assumed there is an initial hypothesis
- If there is no initial hypothesis, how do we generate some?



**Time for a short break and
Some photos of Singapore**

Part 2: Generating hypotheses



The gist of hypothesis generation

- **Hypothesis**

- A comparison of two samples



- More informative than patterns and rules

- **Users not only get to know what is happening but also when or why it is happening**

- **Help users understand what is interesting about their data**

- Hypothesis mining algo

- GUI for visualization and summarization

Conventional hypothesis generation



- **How?**

- Collect data and eye ball a pattern!

PID	Race	Sex	Age	Smoke	Stage	Drug	Response
1	Caucasian	M	45	Yes	1	A	positive
2	Chinese	M	40	No	2	A	positive
3	African	F	50	Yes	2	B	negative
...
N	Caucasian	M	60	No	2	B	negative

- **Limitation**

- Scientist has to think of a hypothesis first
 - Allow just a few hypotheses to be tested at a time

- **So much data have been collected ...**

- No clue on what to look for

Exploratory hypothesis testing



- **Data-driven hypothesis generation**
 - Have a dataset but dunno what hypotheses to test
 - Use computational methods to automatically formulate and test hypotheses from data
- **Problems to be solved**
 - How to formulate hypotheses?
 - How to automatically generate & test hypotheses?

Formulation of a hypothesis

- “For Chinese, is drug A better than drug B?”
- **Three components of a hypothesis:**
 - Context (under which the hypothesis is tested)
 - **Race: Chinese**
 - Comparing attribute
 - **Drug: A or B**
 - Target attribute/target value
 - **Response: positive**
- **$\langle \{\text{Race=Chinese}\}, \text{Drug=A|B}, \text{Response=positive} \rangle$**

Generating a hypothesis

- $\langle \{\text{Race=Chinese}\}, \text{Drug=A|B}, \text{Response=positive} \rangle$
- To test this hypothesis we need info:
 - $N^A = \text{support}(\{\text{Race=Chinese}, \text{Drug=A}\})$
 - $N^A_{\text{pos}} = \text{support}(\{\text{Race=Chinese}, \text{Drug=A}, \text{Res=positive}\})$
 - $N^B = \text{support}(\{\text{Race=Chinese}, \text{Drug=B}\})$
 - $N^B_{\text{pos}} = \text{support}(\{\text{Race=Chinese}, \text{Drug=B}, \text{Res=positive}\})$

context	Comparing attribute	response= positive	response= negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{\text{pos}}$
	Drug=B	N^B_{pos}	$N^B - N^B_{\text{pos}}$

⇒ Frequent pattern mining

More formally

- **Given**
 - Dataset D , min_sup , max_pvalue , min_diff
 - $A_{\text{target}} = v_{\text{target}}$
 - $\mathcal{A}_{\text{grouping}}$: context/comparing attributes
- **Find all $H = \langle P, A_{\text{diff}} = v_1 | v_2, A_{\text{target}} = v_{\text{target}} \rangle$**
 - $A_{\text{diff}} \in \mathcal{A}_{\text{grouping}}$ & $\forall (A=v)$ in P , $A \in \mathcal{A}_{\text{grouping}}$
 - $\text{sup}(P_i) \geq \text{min_sup}$, where $P_i = P \cup \{A_{\text{diff}} = v_i\}$, $i=1, 2$
 - $\text{p-value}(H) \leq \text{max_pvalue}$
 - $|p_1 - p_2| \geq \text{min_diff}$, where p_i is proportion of v_{target} in sub-population P_i , $i=1, 2$

Need for hypothesis analysis



- **Exploration is not guided by domain knowledge**
⇒ Spurious hypotheses has to be eliminated
- **Reasons behind significant hypotheses**
 - Find attribute-value pairs that affect the test statistic a lot

Spurious hypotheses

	response= positive	response= negative	proportion of positive response
Drug=A	890	110	89.0%
Drug=B	830	170	83.0%
Drug=A, Stage=1	800	80	90.9%
Drug=B, Stage=1	190	10	95%
Drug=A, Stage=2	90	30	75%
Drug=B, Stage=2	640	160	80%

- Simpson's Paradox**

- “Stage” has assoc w/ both “drug” & “response”:
 - Doc's tend to give drug A to patients at stage 1, & drug B to patients at stage 2
 - Patients at stage 1 are easier to cure than patients at stage 2
- Attribute “stage” is called a confounding factor

Reasons behind significant hypotheses

	Failure rates
Product A	4%
Product B	2%
Product A, time-of-failure=loading	6.0%
Product B, time-of-failure=loading	1.9%
Product A, time-of-failure=in-operation	2.1%
Product B, time-of-failure=in-operation	2.1%
Product A, time-of-failure=output	2.0%
Product B, time-of-failure=output	1.9%

- **Problem is narrowed down**
 - Product A has exceptionally higher failure rate than product B only at the loading phase

Algo for hypothesis generation



- **A hypothesis is a comparison betw two or more sub-populations, and each sub-populations is defined by a pattern**
- **Step 1: Use freq pattern mining to enumerate large sub-populations and collect their statistics**
 - Stored in the CFP-tree structure, which supports efficient subset/superset/exact search
- **Step 2: Pair sub-populations up to form hypotheses, and then calculate their p-values**
 - Use each freq pattern as a context
 - Search for immediate supersets of the context patterns, and then pair these supersets up to form hypotheses

Algo for rough hypothesis analysis

- **Given a hypothesis H**
 - To check whether H forms a Simpson's Paradox with an attribute A,
 - **add values of A to context of H**
 - **re-calculate the diff betw the two sub-populations**
 - To calculate DiffLift and Contribution of an attribute-value pair $A=v$,
 - **add $A=v$ to context of H**
 - **re-calculate the diff**
- **All done via immediate superset search on frequent patterns**

Other aspects

- **Controlling false-positive rate**
 - Bonferroni's correction
 - Benjamini and Hochberg's method
 - Permutation test
- **Concise representations of hypotheses**
 - freq patterns & hypotheses have lots of redundancy
- **Organization & presentation of hypotheses**
 - Visualization
 - Summarization

Liu, et al. "Supporting exploratory hypothesis testing and analysis". *ACM Transactions on Knowledge Discovery from Data*, 9(4):Article 31, 2015



Uncovering Hidden Insights with
Data-Driven Hypothesis Testing

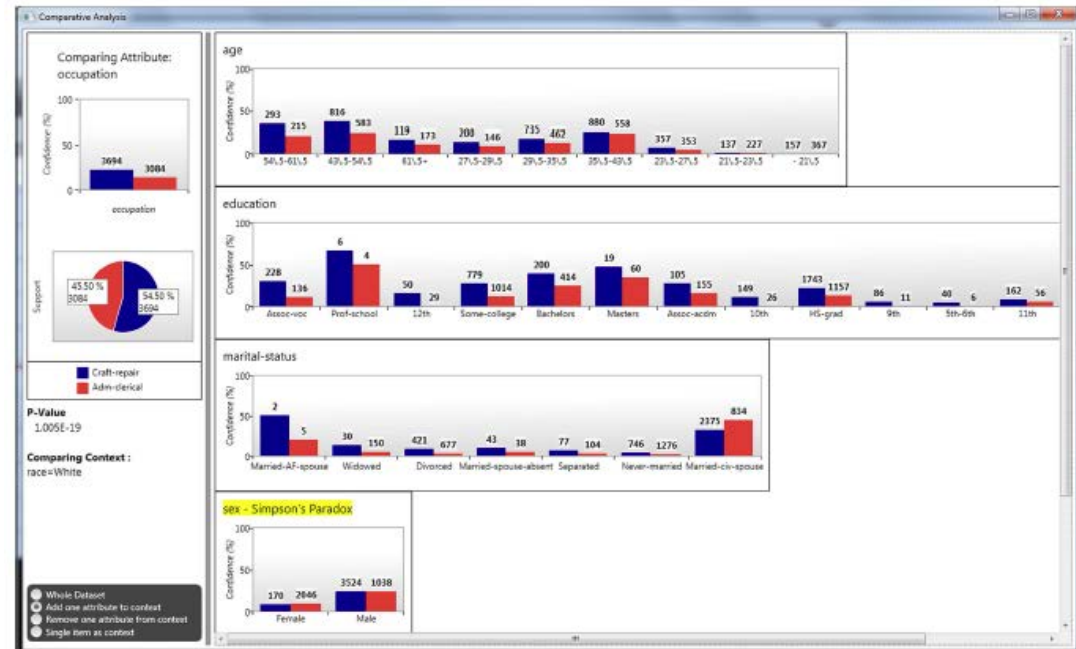
Examples

ID	Gender	Education	Occupation	Income
1	F	Bachelor	Adm-clerical	>50K
2	M	High-School	Sales	≤50K
...

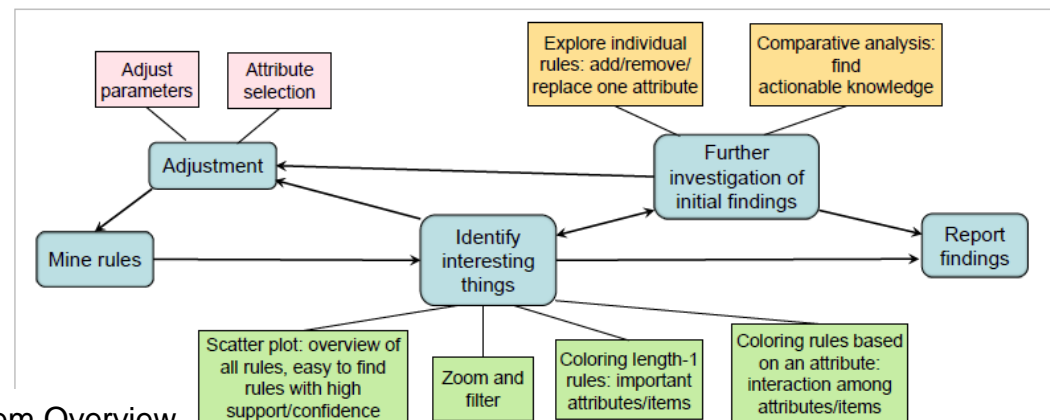
An example dataset

Typical questions:

1. Which groups of people are more likely to have a high income?
2. Which attributes are important to income?
3. What is the effect of "Education" on income with respect to other attributes?
4. Women earn less than men in general. How can women have a high income?



Comparative analysis



System Overview

Experiment settings

- **PC configurations**
 - 2.33Ghz CPU, 3.25GB memory, Windows XP
- **Datasets:**
 - mushroom, adult: UCI repository
 - DrugTestI, DrugTestII: study assoc betw SNPs in several genes & drug responses.

Datasets	#instances	#continuous attributes	#categorical attributes	$A_{\text{target}}/V_{\text{target}}$
adult	48842	6	9	class=>50K (nominal)
mushroom	8124	0	23	class=poisonous (nominal)
DrugTestI	141	13	74	logAUCT (continuous)
DrugTestII	138	13	74	logAUCT (continuous)

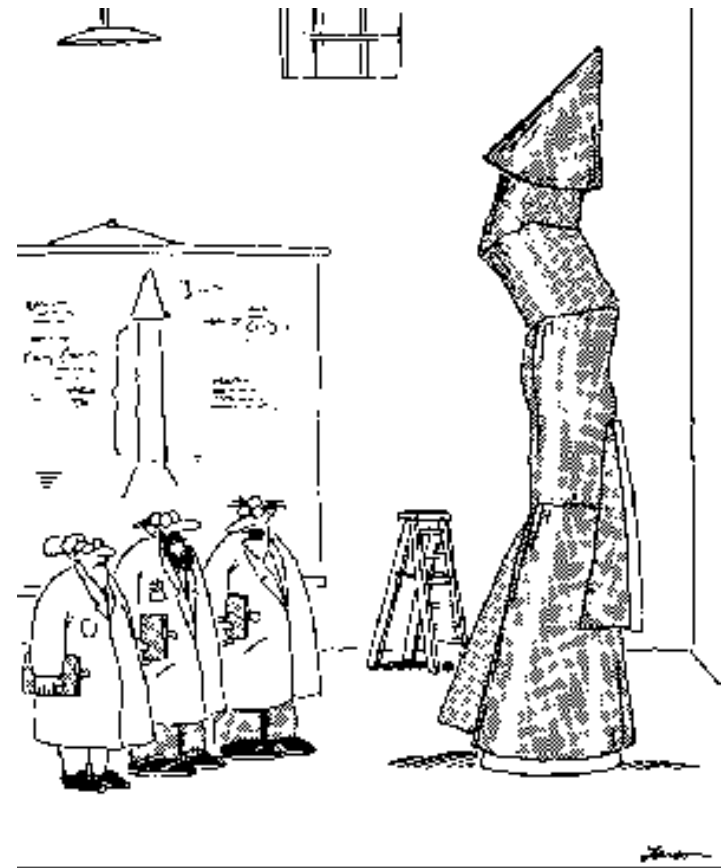
Running time

- **Three phases**
 - Frequent pattern mining
 - Hypothesis generation
 - Hypothesis analysis

Datasets	min_sup	min_diff	GenH	AnalyzeH	AvgAnalyzeT	#tests	#signH
adult	500	0.05	0.42 s	6.30 s	0.0015 s	5593	4258
adult	100	0.05	2.69 s	37.39 s	0.0014 s	41738	26095
mushroom	500	0.1	0.67 s	19.00 s	0.0020 s	16400	9323
mushroom	200	0.1	5.45 s	123.47 s	0.0020 s	103025	61429
DrugTestI	20	0.5	0.06 s	0.06 s	0.0031 s	3627	20
DrugTestII	20	0.5	0.08 s	0.30 s	0.0031 s	4441	97

max_pvalue = 0.05

- We have now seen that filling in contingency tables, looking for contradictions, etc. can be cast as frequent pattern mining tasks
- How is this done efficiently?



"It's time we face reality, my friends. ... We're not exactly rocket scientists."

**Time for a second short break and
Some photos around NUS**

Part 3: Technical details of iDIG



Generating a hypothesis



- $\langle \{ \text{Race}=\text{Chinese} \}, \text{Drug}=\text{A|B}, \text{Response}=\text{positive} \rangle$
- To test this hypothesis we need info:
 - $N^A = \text{support}(\{ \text{Race}=\text{Chinese}, \text{Drug}=\text{A} \})$
 - $N^A_{\text{pos}} = \text{support}(\{ \text{Race}=\text{Chinese}, \text{Drug}=\text{A}, \text{Res}=\text{positive} \})$
 - $N^B = \text{support}(\{ \text{Race}=\text{Chinese}, \text{Drug}=\text{B} \})$
 - $N^B_{\text{pos}} = \text{support}(\{ \text{Race}=\text{Chinese}, \text{Drug}=\text{B}, \text{Res}=\text{positive} \})$

context	Comparing attribute	response=positive	response=negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{\text{pos}}$
	Drug=B	N^B_{pos}	$N^B - N^B_{\text{pos}}$

⇒ Frequent pattern mining

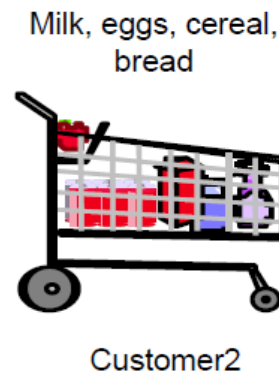


Copyright 2017 © Wong Limsoon

MINING FREQUENT PATTERNS

Market basket analysis

- What do my customers buy?
- Which products are bought together?



- Find correlations between the different items that customers buy

Source: A. Puig

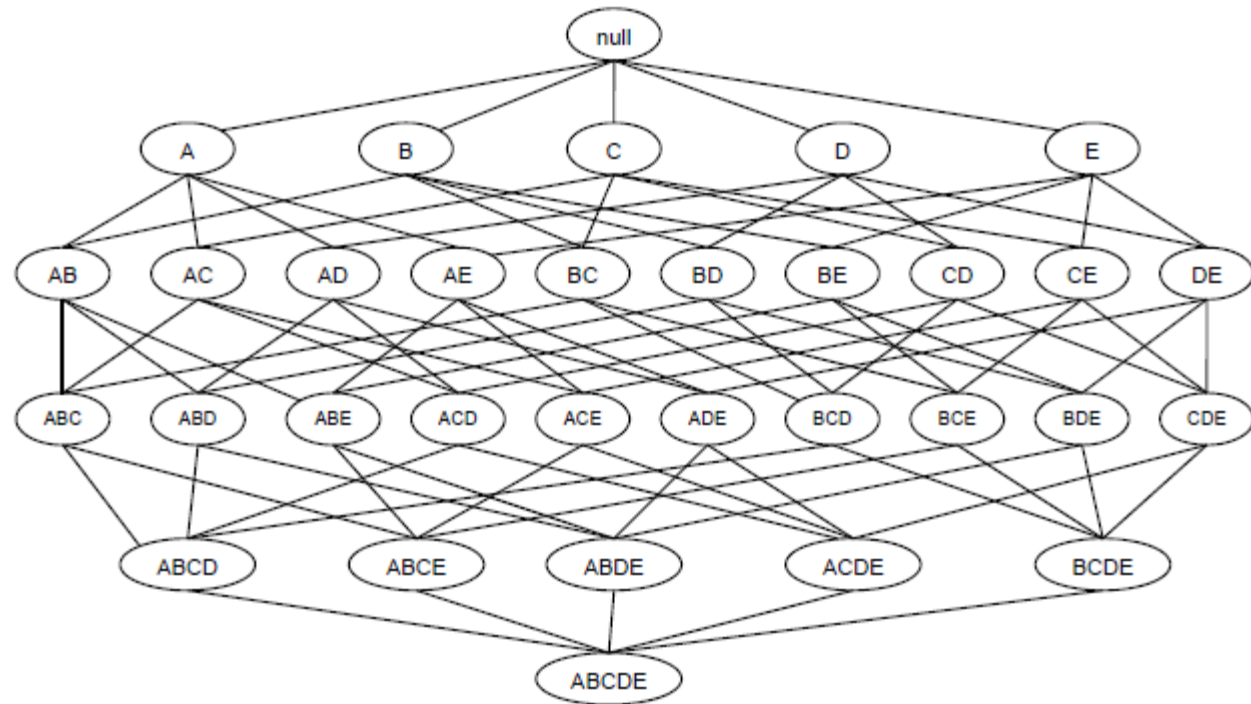
Frequent pattern mining

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

- **Frequent itemsets**
 - Items that often appear together
 - {bread, peanut-butter}

- Transaction db $T = \{t_1, \dots, t_n\}$ is a set of trans
- Each trans t_k is an **itemset** $I = \{i_1, \dots, i_m\}$
- Find **freq patterns** among sets of items in T

Generate freq itemsets with $\text{support} \geq \text{minsup}$



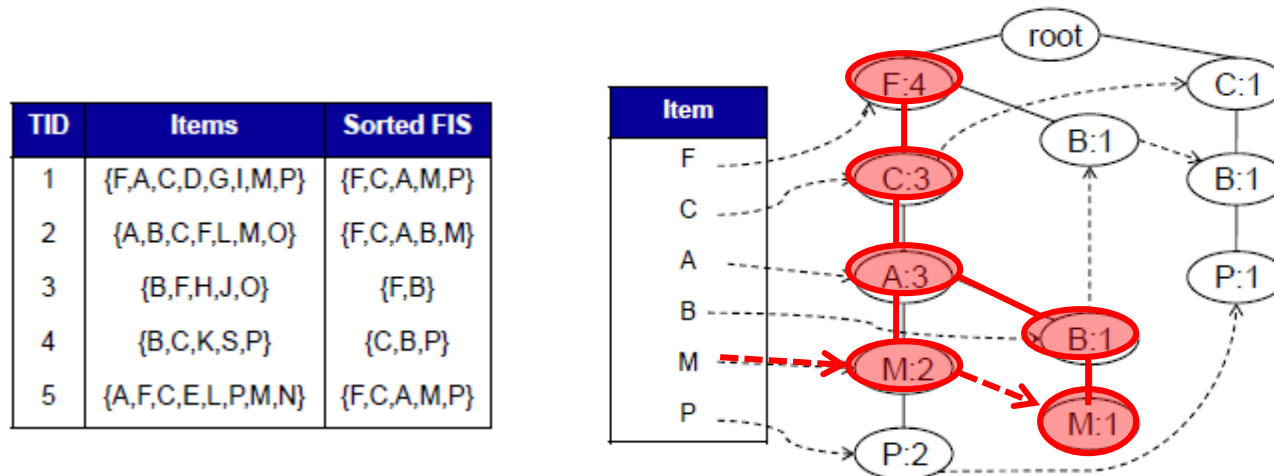
- Given d items. There are 2^d possible itemsets
- Do we need to generate them all?

Han et al. “Mining frequent patterns without candidate generation”.
SIGMOD 2000, pp.1–12



FP-Tree: Counting itemset occurrences

- Build in one scan a data structure, FP-Tree



- Use it for fast support counting
 - To count the support of an itemset {FCM}, follow the “dotted” links on M. At each node M:*n*, note its support *n* & visit its prefix chain; if FCM is found in the prefix, add *n* to the support

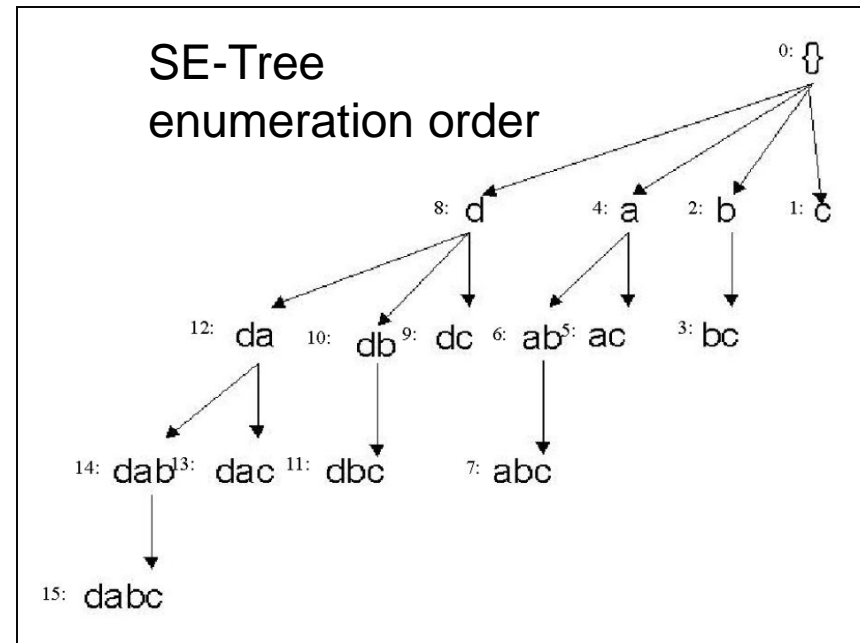
Source: A. Puig

Li et al. "Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns". *KDD 2007*, pp. 430--439



SE-Tree: Mining freq itemsets

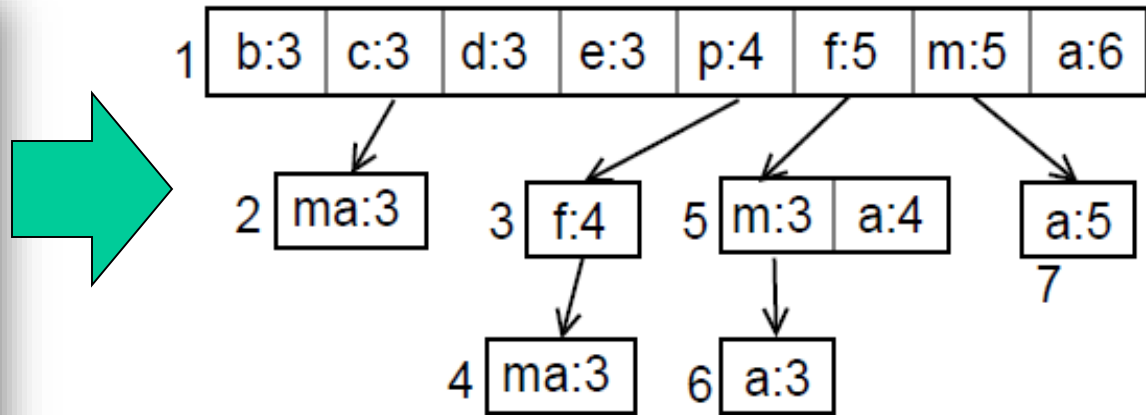
- **Build FP-Tree on the db**
- **Visit itemsets non-redundantly by following the right-to-left top-to-bottom SE-Tree order**
- **When visiting an itemset**
 - Use the FP-tree to count its support efficiently
 - If it is frequent, output it, & visit its supersets
 - Otherwise skip visiting its supersets



Liu et al. “CFP-Tree: A compact disk-based structure for storing and querying frequent itemsets”. *Information Systems*, 32(2):295-319, 2007

CFP-Tree: Storing freq itemsets

TID	Transactions
1	a, c, e, f, m, p
2	b, e, v
3	a, b, f, m, p
4	d, e, f, h, p
5	a, c, d, m, v
6	a, c, h, m, s
7	a, f, m, p, u
8	a, b, d, f, g



Every entry in a CFP-tree represents one or more itemsets with the same support, and these itemsets contain the items on the path from the root to the entry. For instance, node 2 represents 3 itemsets: $\{c, m\}$, $\{c, a\}$ and $\{c, m, a\}$, and these three itemsets have the same support of 3. Let E be an entry and X be an itemset represented by E . Entry E stores three pieces of information: (1) m items ($m \geq 1$), (2) the support of X , and (3) a pointer pointing to the child node of E .

For every entry E in a multiple-entry node, only the items after E can appear in the subtree pointed by E

The CFP-tree nodes are stored on pages of 4096 bytes in depth-first order. A node is stored before its child nodes. If the size of a node is larger than page size, then it is stored on several consecutive pages. Nodes that are no larger than one page are not allowed to span two pages.

Liu et al. "A performance study of three disk-based structures for indexing and querying frequent itemsets". *PVLDB*, 6(7):505-516, 2013.



Querying
stored freq
itemsets
from CFP-
Tree

Is Q freq?
and what is
its support?

Algorithm 5 CFP-tree_Exact_Match Algorithm

Input:

$cnode$ is a CFP-tree node;

Q' is the set of items in Q that have not be covered yet;

Description:

```

1: if  $cnode$  contains only one entry  $E$  then
2:   if  $Q' - E.items = \emptyset$  then
3:     output  $E.support$ ;
4:   else if  $E.child \neq NULL$  then
5:     CFP-tree_Exact_Match( $E.child$ ,  $Q' - E.items$ );
6:   else
7:     Output "not found";
8: else if  $cnode$  contains multiple entries then
9:   if all items in  $Q'$  occur in  $cnode$  then
10:     $E$  = the first entry in  $cnode$  such that  $E.items \in Q'$ ;
11:    if  $Q' - E.items = \emptyset$  then
12:      Output  $E.support$ ;
13:    else if  $E.child \neq NULL$  then
14:      CFP-tree_Exact_Match( $E.child$ ,  $Q' - E.items$ );
15:    else
16:      Output "not found";

```

Algo for rough hypothesis analysis

- **Given a hypothesis H**
 - To check whether H forms a Simpson's Paradox with an attribute A,
 - add values of A to context of H
 - re-calculate the diff betw the two sub-populations
 - To calculate DiffLift and Contribution of an attribute-value pair $A=v$,
 - add $A=v$ to context of H
 - re-calculate the diff
- All done via immediate superset search

IMMEDIATE SUPERSET SEARCH

Simple immediate subset search

- Given Q , use exact match to search $Q - \{q\}$ for each q in Q

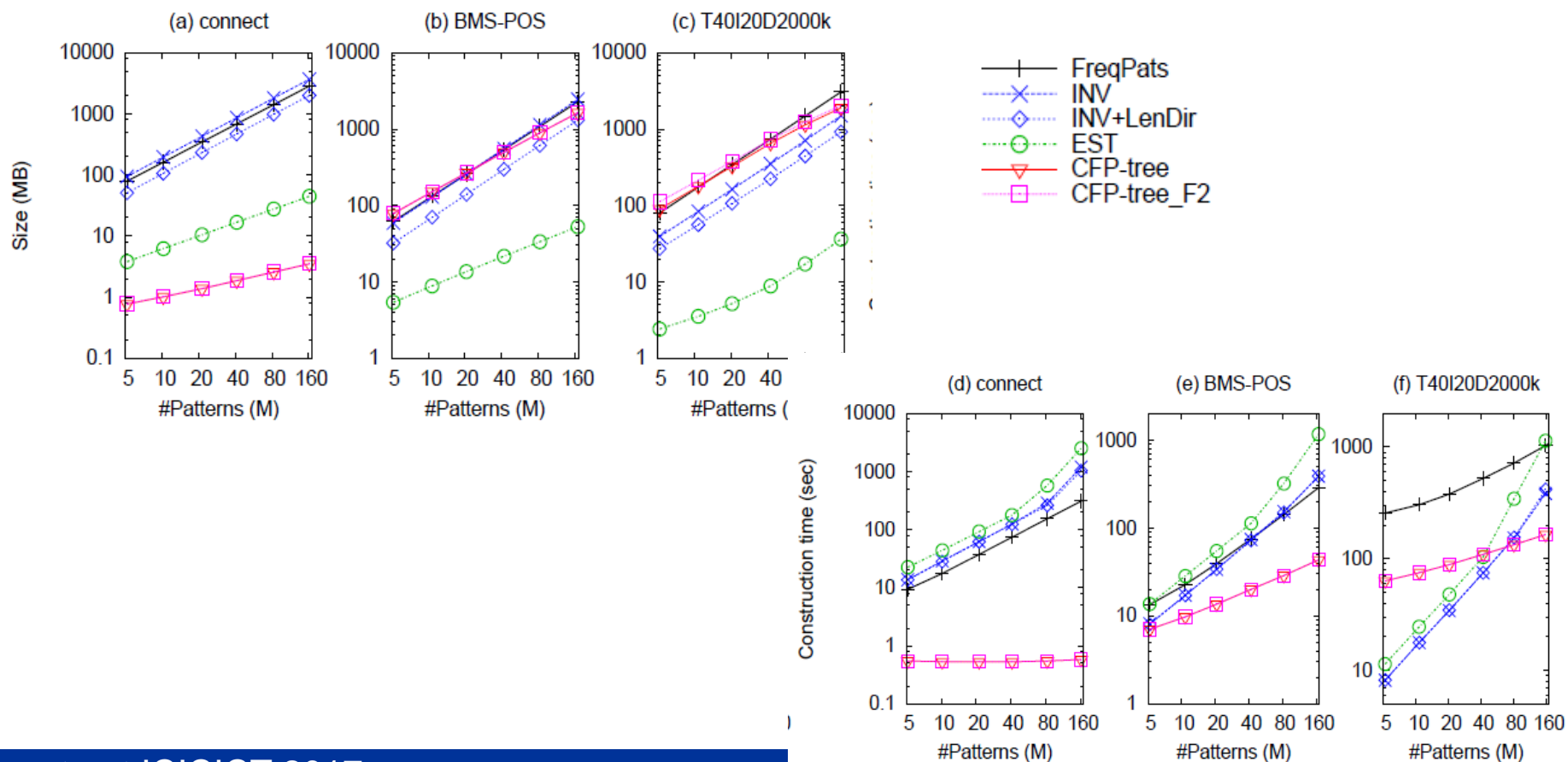
Simple immediate superset search

- Given Q , use exact match to search $Q \cup \{q\}$ for each q not in Q
- Can we do better? Figure out which q does not need checking! Details in paper below:

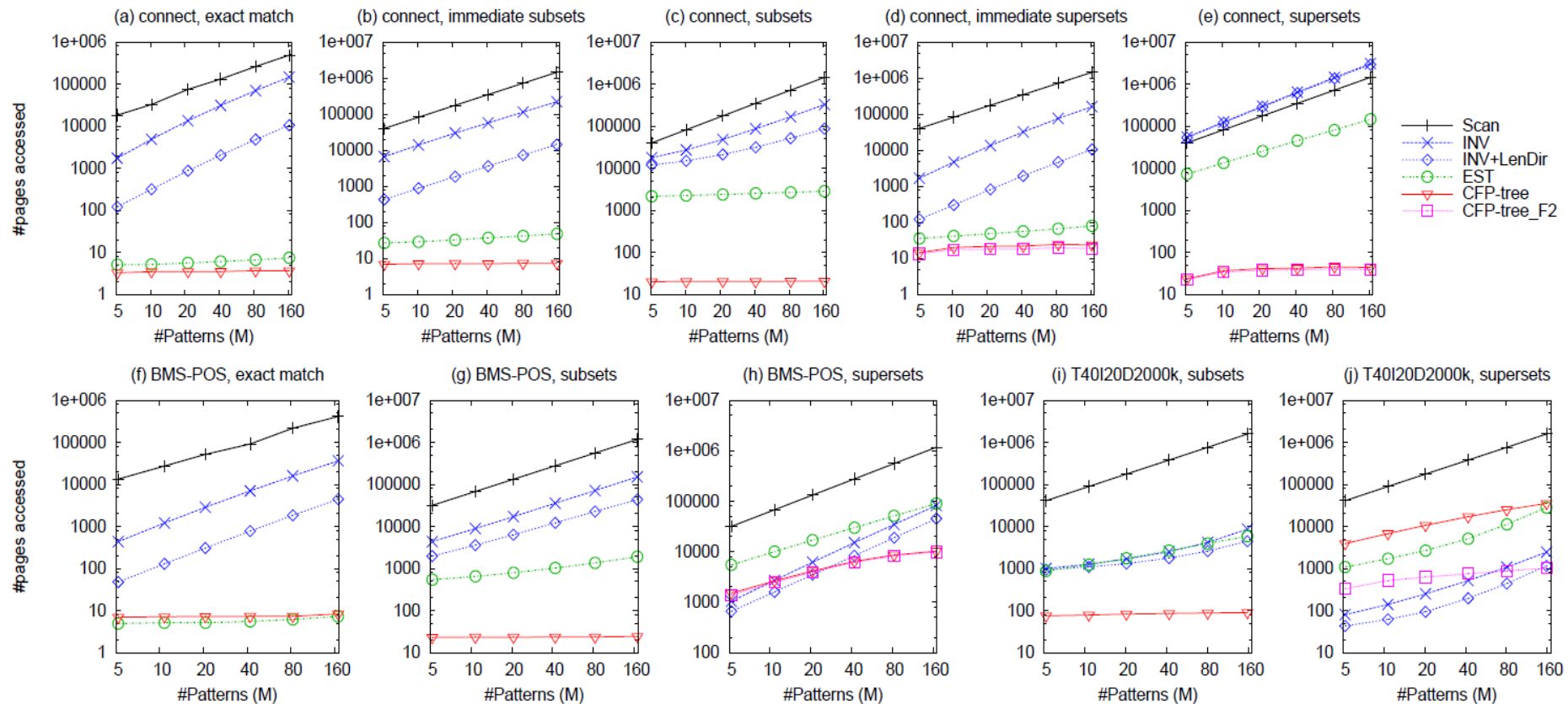
Liu et al. "A performance study of three disk-based structures for indexing and querying frequent itemsets". *PVLDB*, 6(7):505-516, 2013.

Size and construction time

Datasets	#transactions	#items	avg_tlen	max_tlen	size
connect	67,557	129	43	43	0.34MB
BMS-POS	515,597	1657	6.5	164	11.1MB
T40I20D2000k	2,000,000	12,228	40.0	78	427.9MB

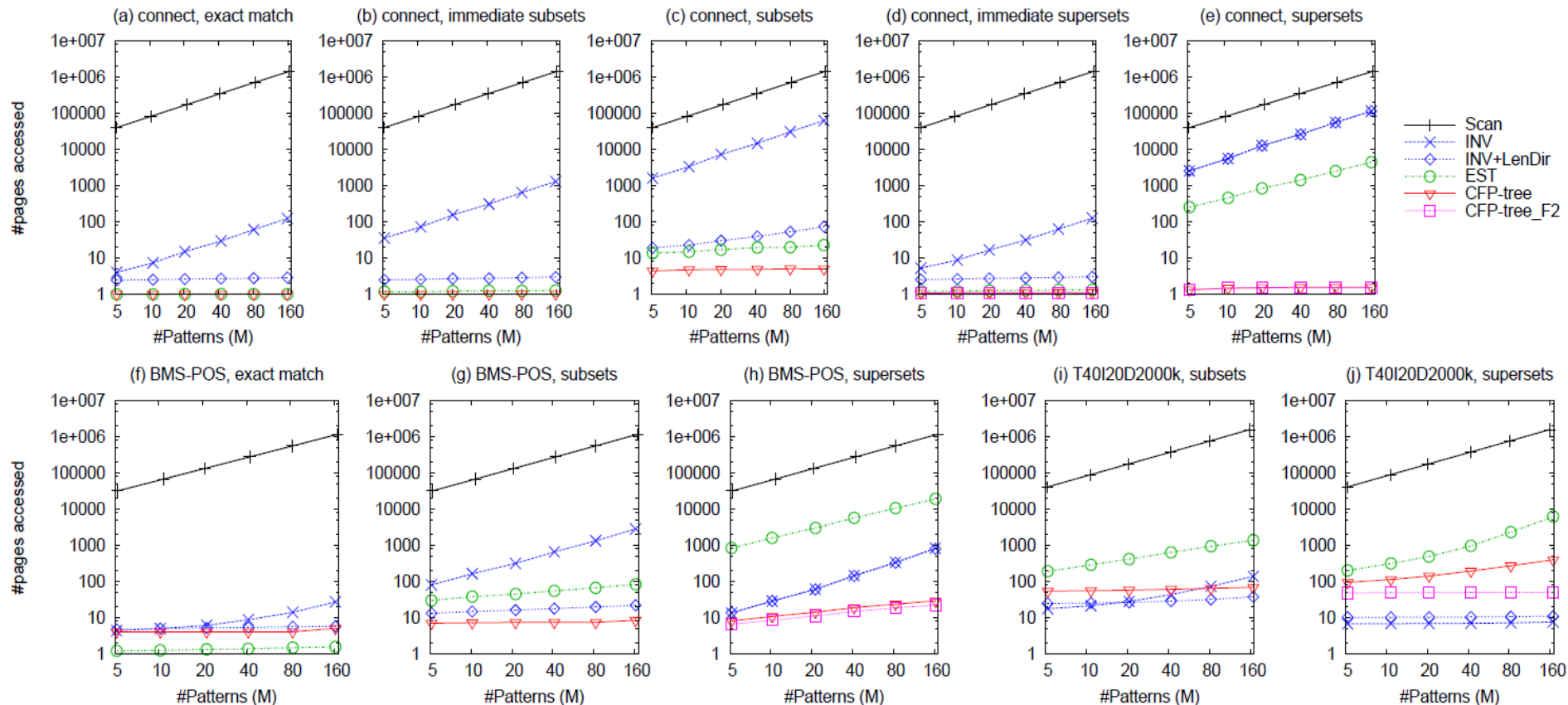


Scalability: 1000 queries w/ hits



- **CFP-Tree scales sub-linearly for queries with hits!**

Scalability: 1000 queries w/o hits



- **CFP has near constant cost for queries w/o hits**

**Time for a third short break and
Some info about data science & analytics research
in NUS School of Computing**

Part 4: Art of data analysis



NUS
National University
of Singapore

Triumph of logic

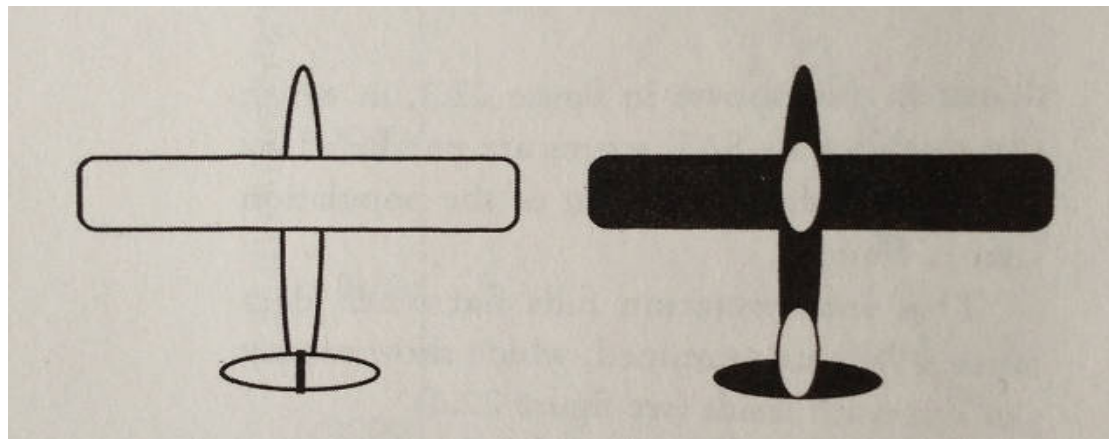
THE PRESENCE OF ABSENCE

Where should
extra armour be
put on a bomber
plane to improve
its survival?



- **US General's solution**

- Collect statistics on which parts of bomber planes got shot how many times
- Put armour on the hot spots



Undamaged plane (left). A plane shaded everywhere bullets struck returning aircraft (right).

- **Is this a good solution? Why?**

- **Abraham Wald's analysis**
 - Data were collected from planes that survived
 - The more bullet holes seen in a part, the more hits that part could take
 - Thus the parts that were unscathed would need more armour

Abraham Wald (1902-1950) was a Hungarian mathematician. He invented sequential hypothesis testing from his work on bomber planes. He died in a plane crash in the Nilgiri mountains.

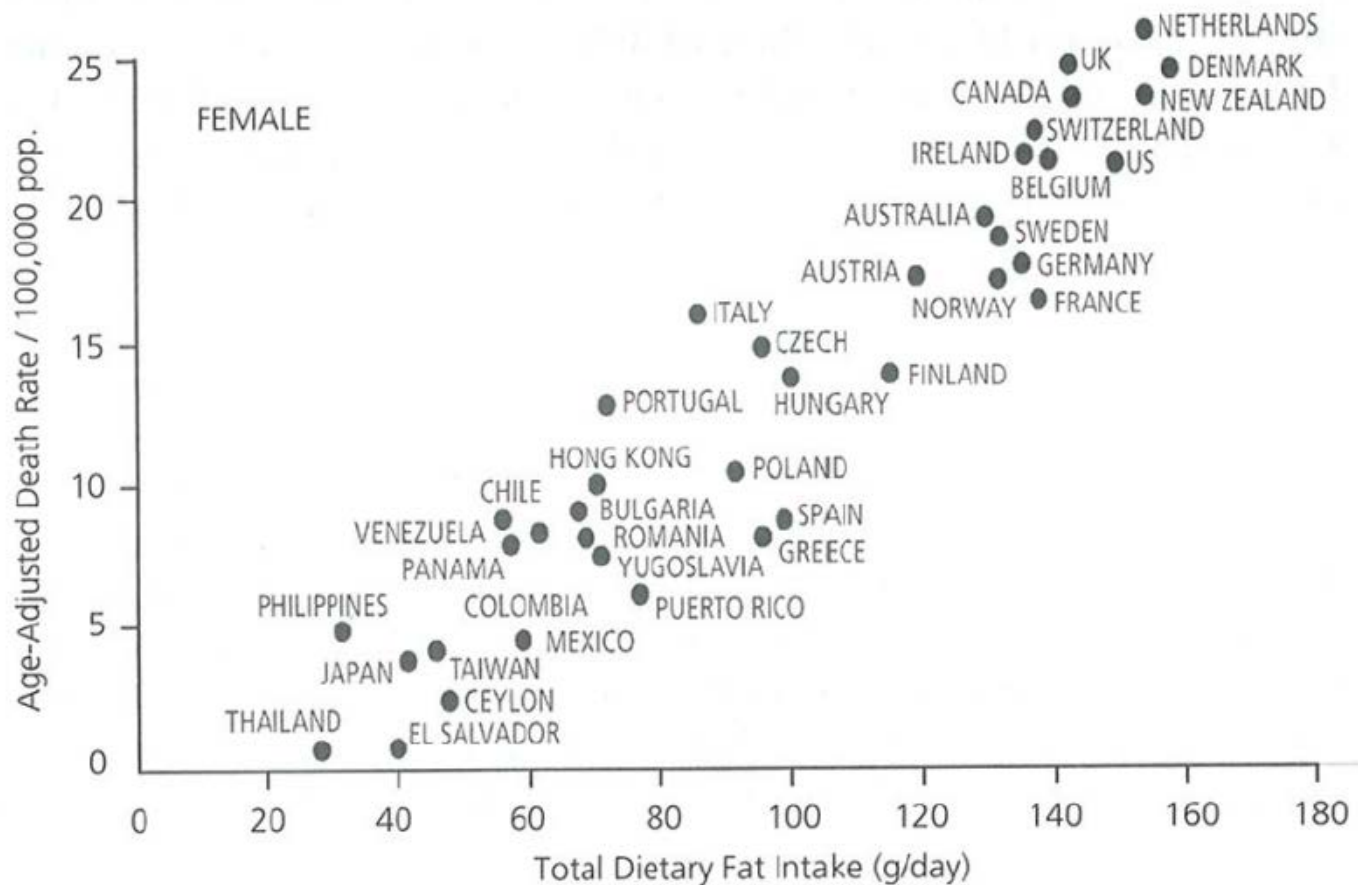


- “It is so easy to make bad inferences with data... there’s a creative part of understanding quantitative data that requires a sort of artistic or creative approach to research.” ---Nate Bolt
- <http://www.fastcodesign.com/1671172/how-a-story-from-world-war-ii-shapes-facebook-today>

Triumph of logic

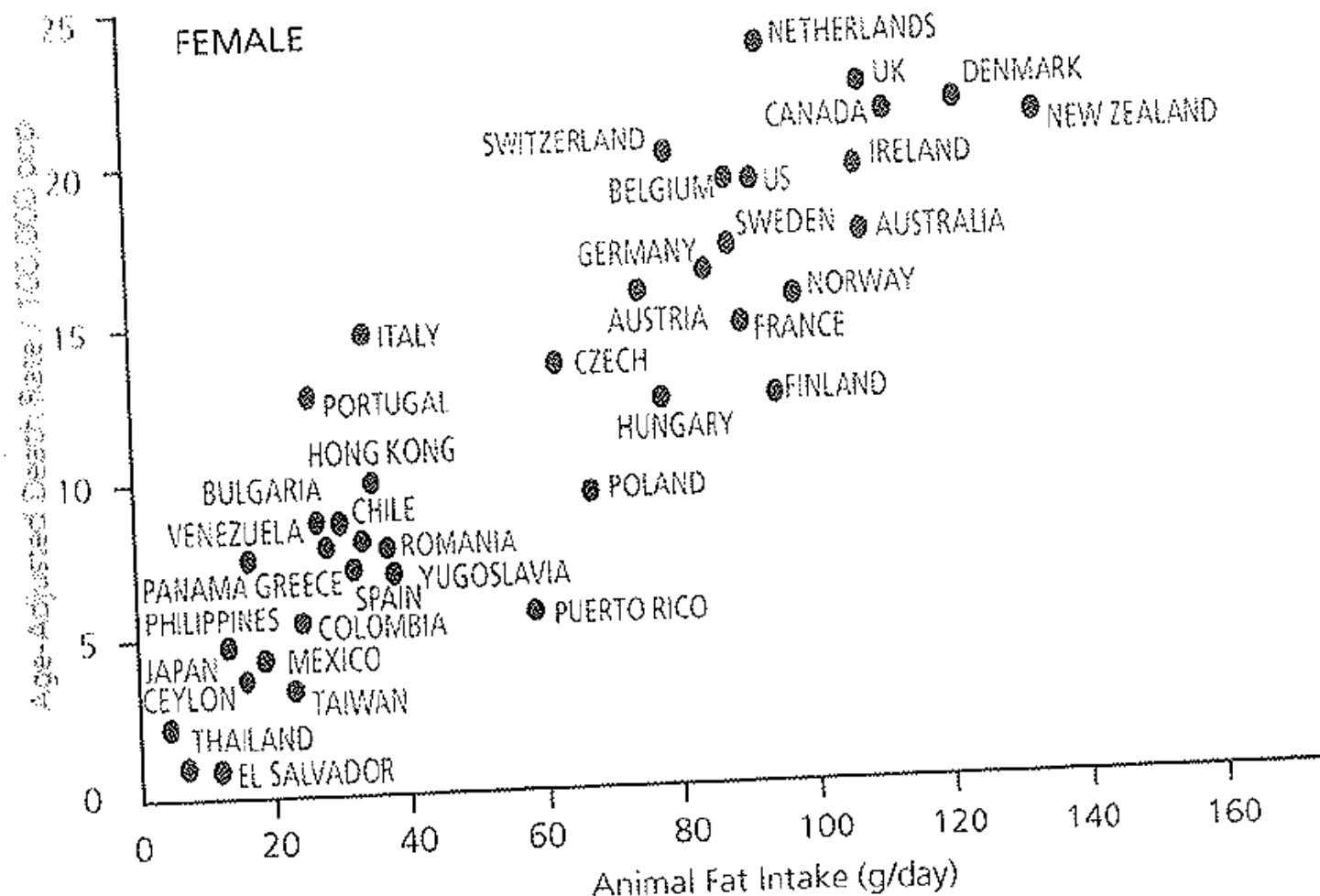
ABSENCE IS PRESENCE

We love to find correlations like this?



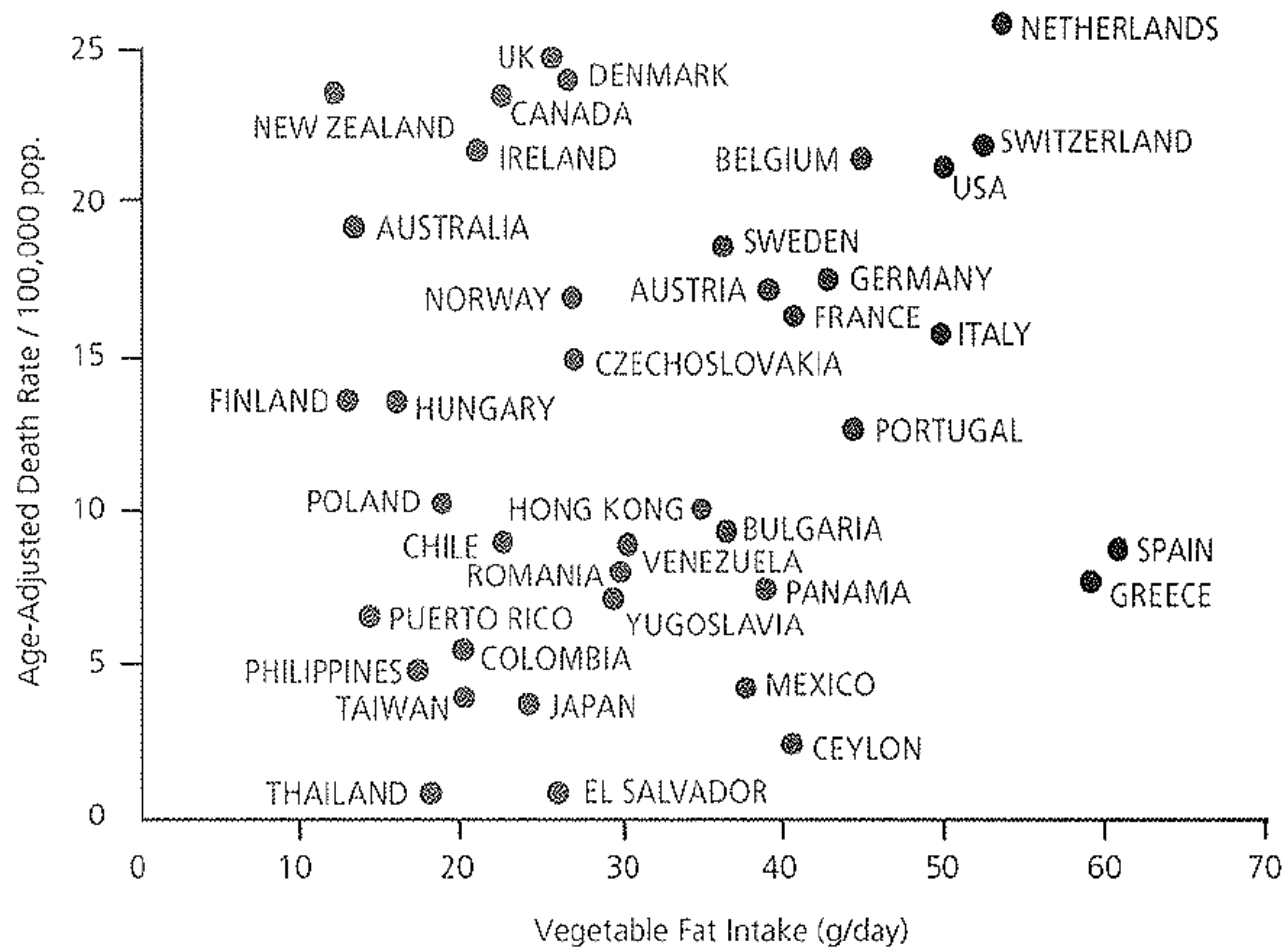
- **Dietary fat intake correlates with breast cancer**

And like this...



- **Animal fat intake correlates with breast cancer**

But not non-correlations like this..



- **Plant fat intake doesn't correlate with breast cancer**

We tend to ignore non-associations

- **We have many technologies to look for associations and correlations**
 - Frequent patterns
 - Association rules
 - ...
- **We tend to ignore non-associations**
 - We think they are not interesting / informative
 - There are too many of them
- **Is this a good thing to do? Why?**

There is much to be gained when we take both into our analysis



**A: Dietary fat intake
correlates with breast
cancer**

**B: Animal fat intake
correlates with breast
cancer**

**C: Plant fat intake
doesn't correlate with
breast cancer**

**⇒ Given C, we can
eliminate A from
consideration, and
focus on B!**



**The power of
negative space!**

- **How many animals do you see?**

Triumph of logic

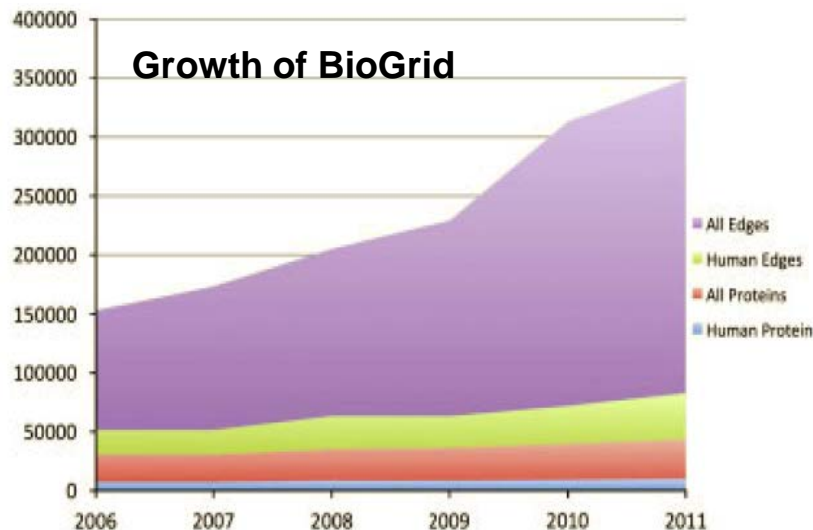
**THE DATA IS TELLING YOU
MORE THAN WHAT IT IS
TELLING YOU**

Protein-protein interaction detection

- Many high-throughput assays for PPIs

Generating large amounts of expt data on PPIs can be done with ease

- But ...



High-throughput approaches sacrifice quality for **quantity**:
(a) limited or biased coverage:
false negatives, &
(b) high error rates:
false positives

Noise in PPI networks

Experimental method category ^a	Number of interacting pairs	Co-localization ^b (%)	Co-cellular-role ^b (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz <i>et al.</i> (published results)	956	66	45
A1: GY2H Uetz <i>et al.</i> (unpublished results)	516	53	33
A2: GY2H Ito <i>et al.</i> (core)	798	64	40
A3: GY2H Ito <i>et al.</i> (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, <i>in vitro</i>	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

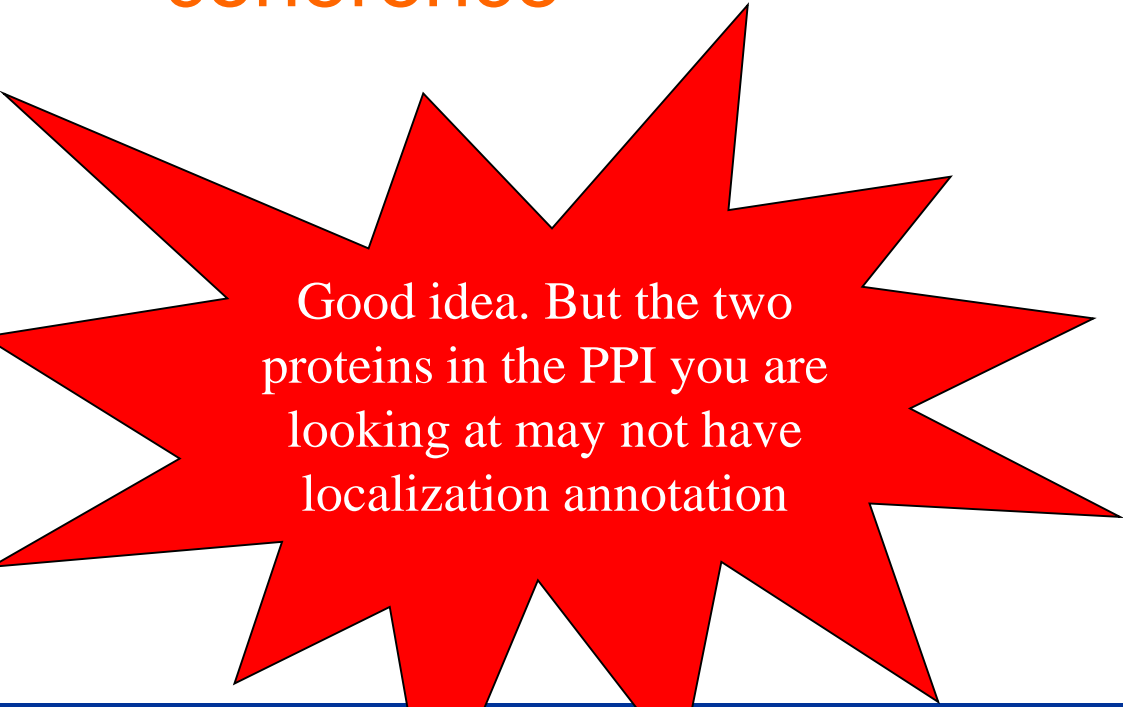
Sprinzak *et al.*, *JMB*, 327:919-923, 2003

Large disagreement betw methods

- Can you think of things a biologist can do to remove PPIs that are likely to be noise?

De-noising PPI networks using localization coherence

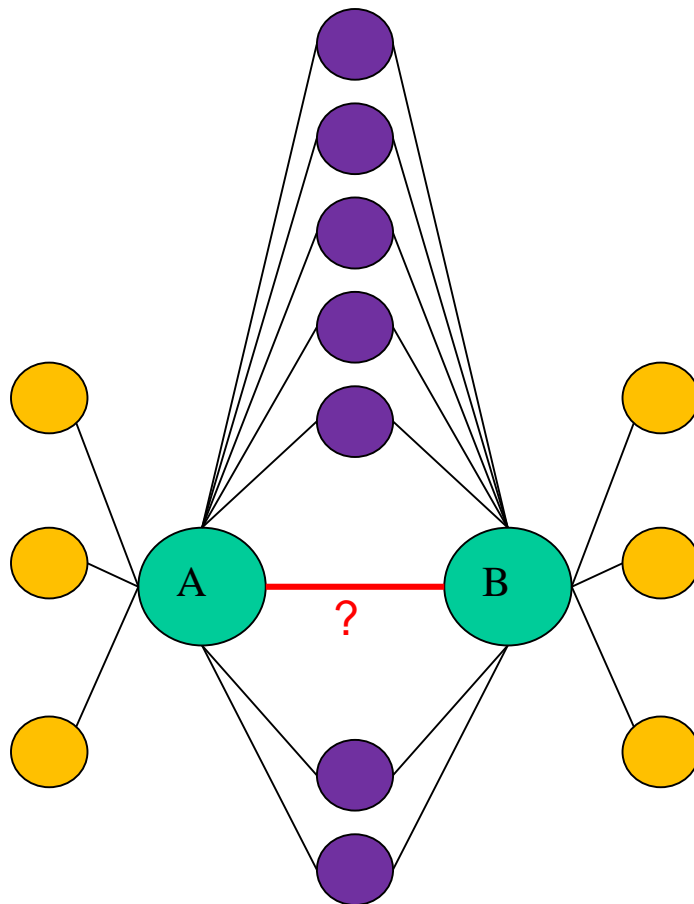
- Two proteins should be in the same place to interact. Agree?



Good idea. But the two proteins in the PPI you are looking at may not have localization annotation

- **Do you really need to know where two proteins are, in order to know whether they are in the same place?**

Topology of neighbourhood of real PPIs

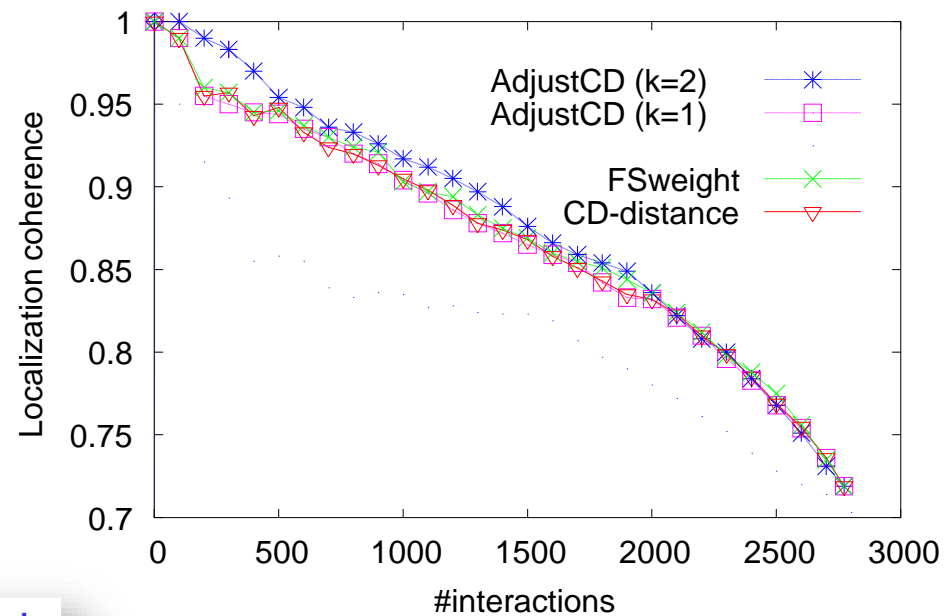


- Suppose 20% of putative PPIs are noise
- ⇒ ≥ 3 purple proteins are real partners of both A and B
- ⇒ A and B are likely localized to the same cellular compartment (Why?)
- ⇒ A and B are likely PPI

Liu et al. Complex discovery from weighted PPI networks.
Bioinformatics, 25(15):1891-1897, 2009

It works!

Cf. ave localization coherence of protein pairs in DIP < 5%
 ave localization coherence of PPI in DIP < 55%



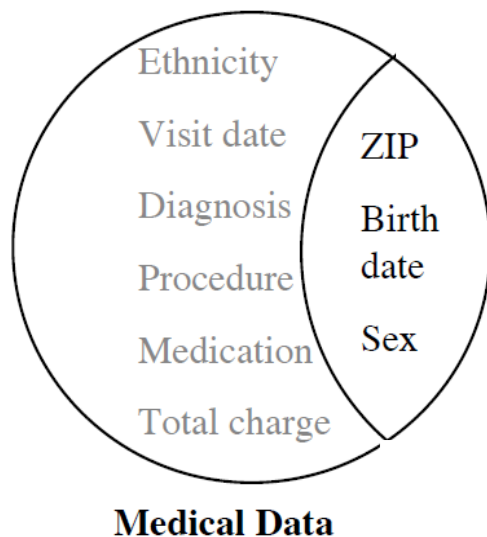
- **Given a pair of proteins (u, v) in a PPI network**
 - N_u = the set of neighbors of u
 - N_v = the set of neighbors of v

- $$CD(u,v) = \frac{2 | N_u \cap N_v |}{| N_u | + | N_v |}$$

Triumph of logic

**THE DATA TELLS YOU MORE
THAN WHAT IT DOESN'T
WANT TO TELL YOU**

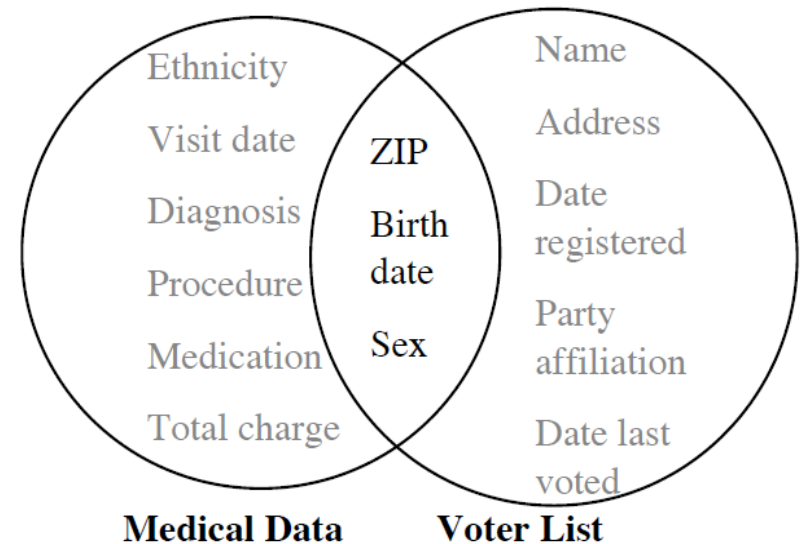
Is anonymized data really anonymous?



The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth [2]. The leftmost circle in Figure 1 contains a subset of the fields of information, or *attributes*, that NAHDO recommends these states collect; these attributes include the patient's ZIP code, birth date, gender, and ethnicity.

In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patient-specific data with nearly one hundred attributes per encounter along the lines of the those shown in the leftmost circle of Figure 1 for approximately 135,000 state employees and their families. Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry [3].

Latanya Sweeney inferred the governor's medical record by linking the GIC record to Voter list!



For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes [4]. The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals.

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

Sweeney, "k-anonymity: A model for protecting privacy", *Int J Unc Fuzz Knowl Based Syst*, 10:557-570, 2002

- Oftentimes it is logic that triumphs in data analysis, not mechanical use of datamining, machine learning, and statistical methods

Part 5: Science of data analysis



The science exists but ...

- **Current data analysis methods have clear assumptions & practices**
 - Normal distribution
 - I.I.D.
 - Proper design of experiment
 - Domain-specific laws
 - Proper context
- **Analysis outcome is valid when assumptions hold & practices followed**
- **But often assumptions not checked & practices not followed when people run these methods!**



Forgotten assumptions

**NORMAL DISTRIBUTION,
BUT REAL WORLD IS OFTEN NOT
NORMALLY DISTRIBUTED**

Wisdom of the crowd

Lorenz et al., *PNAS*, 108(22):9020-9025, 2011



Table 1. The wisdom of crowd effect exists with respect to the geometric mean but not with respect to the arithmetic mean

Question	True value	Wisdom-of-crowd aggregation		
		Arithmetic mean	Geometric mean	Median
1. Population density of Switzerland	184	2,644 (+1,337.2%)	132 (−28.1%)	130 (−29.3%)
2. Border length, Switzerland/Italy	734	1,959 (+166.9%)	338 (−54%)	300 (−59.1%)
3. New immigrants to Zurich	10,067	26,773 (+165.9%)	8,178 (−18.8%)	10,000 (−0.7%)
4. Murders, 2006, Switzerland	198	838 (+323.2%)	174 (−11.9%)	170 (−14.1%)
5. Rapes, 2006, Switzerland	639	1,017 (+59.1%)	285 (−55.4%)	250 (−60.9%)
6. Assaults, 2006, Switzerland	9,272	135,051 (+1,356.5%)	6,039 (−34.9%)	4,000 (−56.9%)

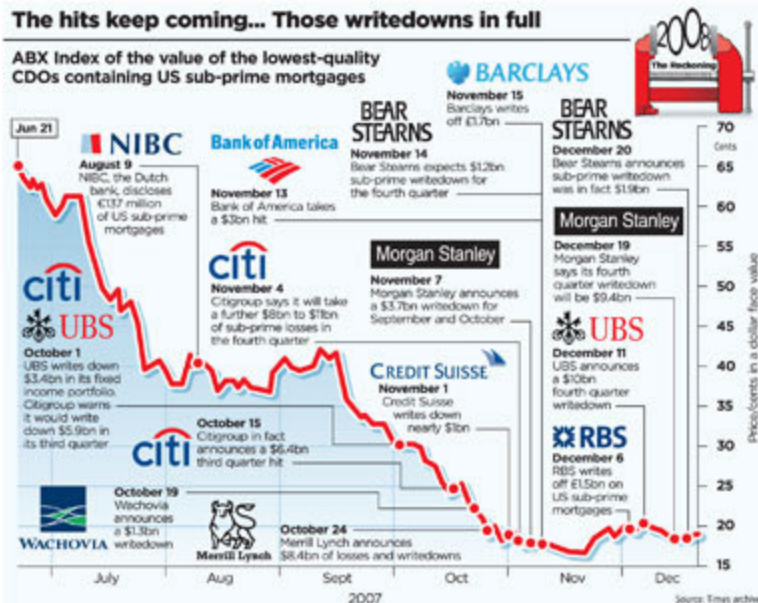
The aggregate measures arithmetic mean, geometric mean, and median are computed on the set of all first estimates regardless of the information condition. Values in parentheses are deviations from the true value as percentages.

- Estimates not normally distributed
 - They are lognormally distributed
- ⇒ Subjects had problems choosing the right order of magnitude

**Me: I'm
finally happy.
Life: Lol,
wait a sec.**

and what held yesterday may not hold today

2007 financial crisis



- VaR measures the expected loss over a horizon **assuming normality**
- “When you realize that VaR is using tame historical data to model a wildly different environment, the total losses of Bear Stearns’ hedge funds become easier to understand. It’s like the historic data only has rainstorms and then a tornado hits.” – New York Times, 2 Jan 2009
- You can still turn things into your advantage if you are alert: When VaR numbers start to miss, either there is something wrong with the way VaR is being calculated, or the market is no longer normal
- All of them religiously check VaR (Value at Risk) everyday



"All those in favor say 'Aye.'"

"Aye."

"Aye."

"Aye."

"Aye."

"Aye."

Forgotten assumptions

**I.I.D.,
BUT REAL WORLD IS OFTEN NOT
INDEPENDENTLY DISTRIBUTED**

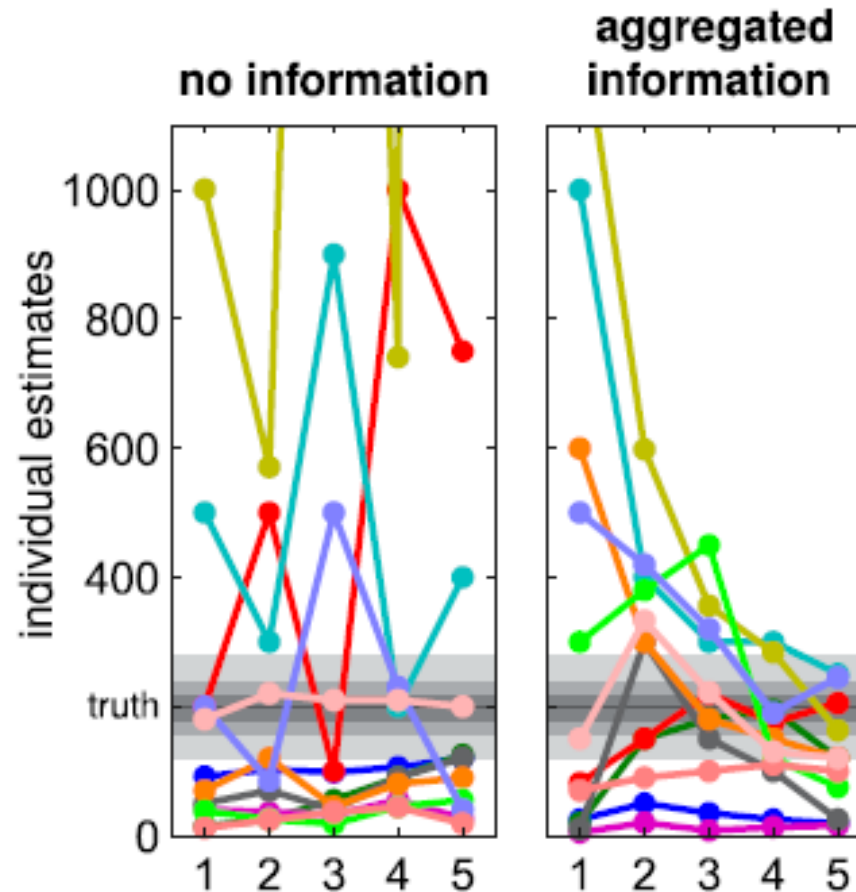
Experiments on social influence

Lorenz et al., *PNAS*, 108(22):9020-9025, 2011



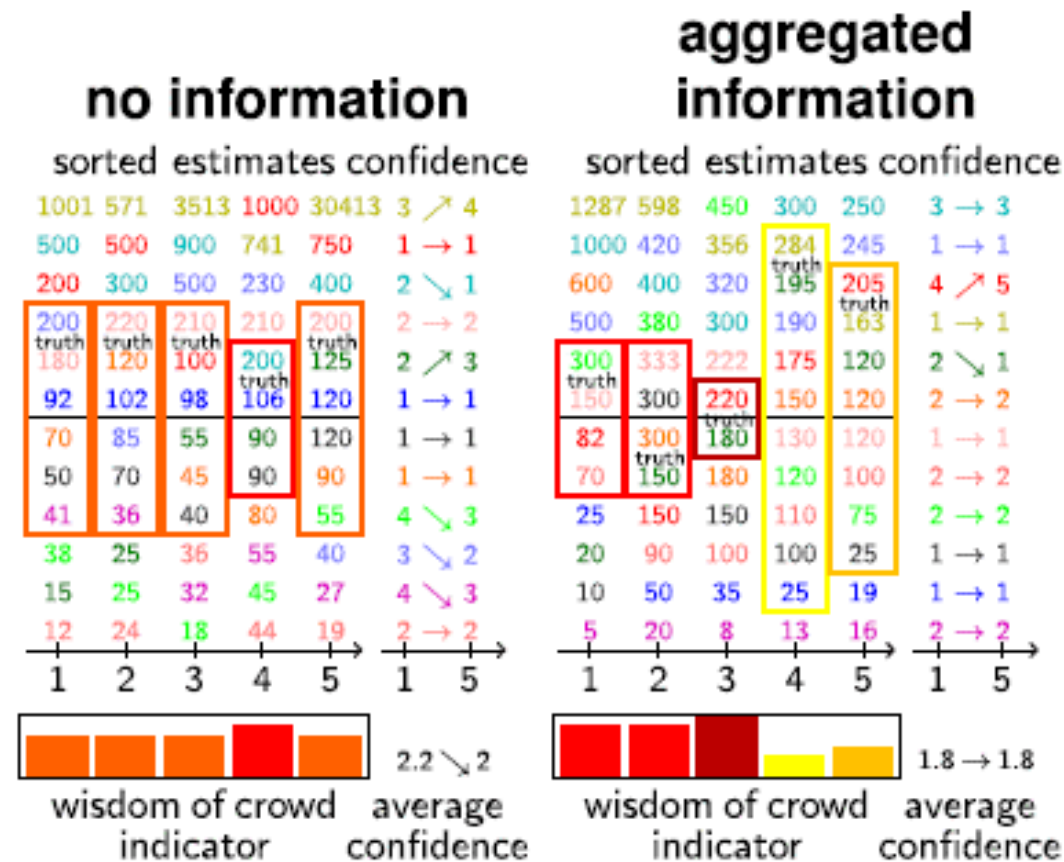
- 12 groups, 12 subjects each
- Each subject solves 6 different estimation tasks regarding geographical facts and crime statistics
- Each subject responds to 1st question on his own
- After all 12 group members made estimates, everyone gives another estimate, 5 consecutive times
- Different groups based their 2nd, 3rd, 4th, 5th estimates on
 - Aggregated info of others' from the previous round
 - Full info of others' estimates from all earlier rounds
 - Control, i.e. no info
- Two questions posed for each of the three treatments
- Each declares his confidence after the 1st and final estimates

Social influence effect



- **Social influence diminishes diversity in groups**
 ⇒ **Groups potentially get into “group think”!**

Range reduction effect



- Group zooms into wrong estimate
- Truth may even be outside all estimates

Social influence diminishes wisdom of the crowd

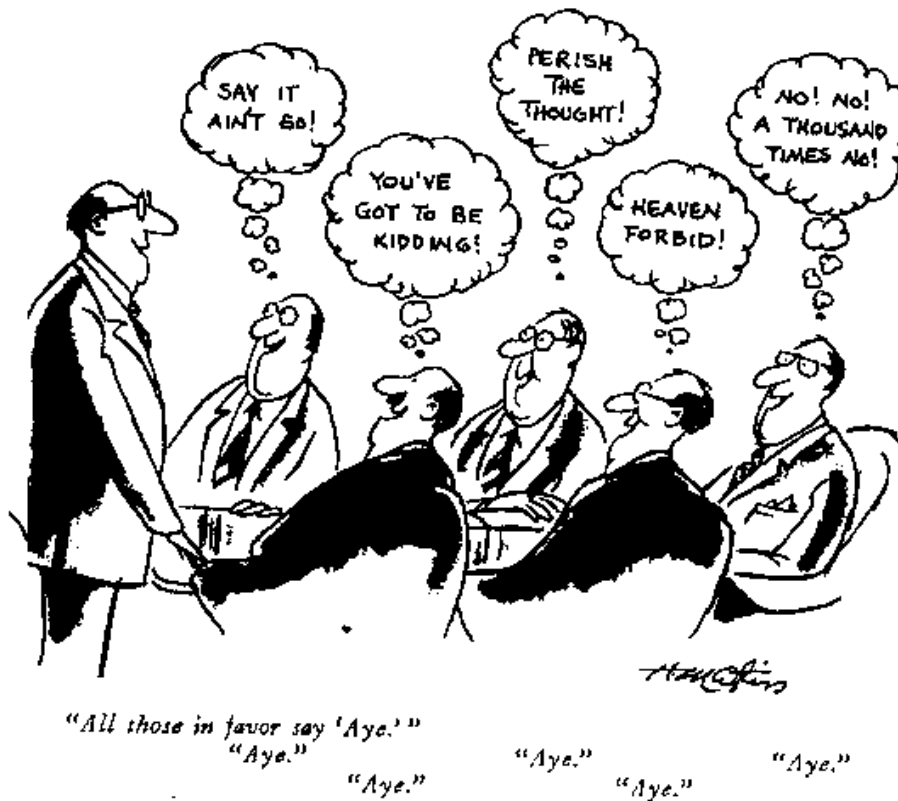


- Social influence triggers convergence of individual estimates
 - The remaining diversity is so small that the correct value shifts from the center to the outer range of estimates
- ⇒ An expert group exposed to social influence may result in a set of predictions that does not even enclose the correct value any more!
- Conjecture: Negative effect of social influence is more severe for difficult questions

Related issue: People do not say what they really want to say

Stephen King, "Conflict between public and private opinion", *Long Range Planning*, 14(4):90-105, August 1981

"In fact, the evidence is very strong that there is a genuine difference between people's private opinions and their public opinions."





Forgotten assumptions

**PROPER DESIGN OF EXPT,
BUT REAL WORLD BIG DATA IS NOT
“DESIGNED”**

Design of experiments

- In clinical testing, we **carefully choose the sample to ensure the test is valid**
 - Independent: Patients are not related
 - Identical: Similar # of male/female, young/old, ... in cases and controls

	A	B
lived	60	65
died	100	165

Note that sex, age, ... don't need to appear in the contingency table

- In big data analysis, and in many datamining works, people hardly ever do this!
 - Is this sound?

What is happening here?



Overall

	A	B
lived	60	65
died	100	165

Looks like treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Looks like treatment B is better

History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

Looks like treatment A is better

A/B sample not identical in other attributes



Overall

	A	B
lived	60	65
died	100	165

- **Taking A**

- Men = 100 (63%)
- Women = 60 (37%)

- **Taking B**

- Men = 210 (91%)
- Women = 20 (9%)

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

- **Men taking A**

- History = 80 (80%)
- No history = 20 (20%)

- **Men taking B**

- History = 55 (26%)
- No history = 155 (74%)

Related issue: Sampling bias

"Dewey Defeats Truman" was a famously incorrect banner headline on the front page of the *Chicago Tribune* on November 3, 1948, the day after incumbent United States President Harry S. Truman won an upset victory over Republican challenger and Governor of New York Thomas E. Dewey in the 1948 presidential election.



President-elect Truman holding the infamous issue of the *Chicago Tribune*, telling the press, "That ain't the way I heard it!"

The reason the Tribune was mistaken is that their editor trusted the results of a phone survey... Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses.

context

/ˈkɒntɛkst/ 

noun

the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.

"the proposals need to be considered in the context of new European directives"

synonyms: circumstances, conditions, [surroundings](#), factors, state of affairs; [More](#)

- the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning.

"skilled readers use context to construct meaning from words as they are read"

Overlooked information

**CLEAR CONTEXT,
BUT REAL WORLD BIG DATA OFTEN HAS
CONFOUNDED CONTEXTS**

... And worse, we tend to ignore context!



- **We have many technologies to look for associations and correlations**
 - Frequent patterns
 - Association rules
 - ...
- **We tend to assume the same context for all patterns and set the same global threshold**
 - This works for a focused dataset
 - But for big data where you union many things, this spells trouble

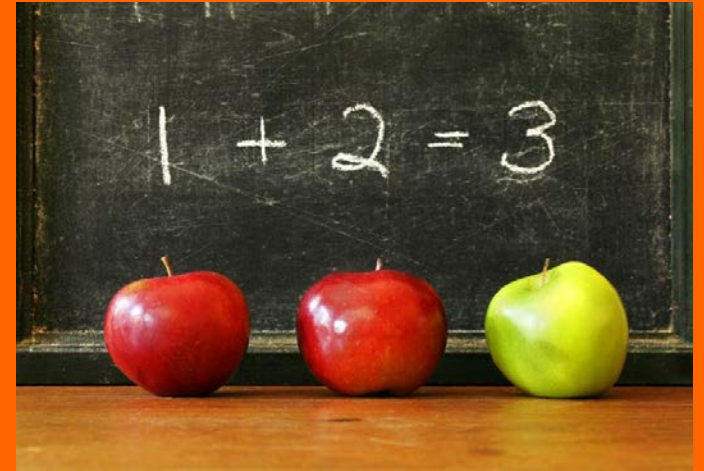
The right context

- $\langle \{\text{Race=Chinese}\}, \text{Drug=A|B}, \text{Response=positive} \rangle$

Context	Comparing attribute	response=positive	response=negative
{Race=Chinese}	Drug=A	N_{pos}^A	$N^A - N_{\text{pos}}^A$
	Drug=B	N_{pos}^B	$N^B - N_{\text{pos}}^B$

- If A/B treat the same single disease, this is ok
- If B treats two diseases, this is not sensible
- The disease has to go into the context

Summary



What have we learned?

- **Part 1: Simple tactics to get deeper insight from data**
- **Part 2: These tactics can be realized using frequent pattern mining**
- **Part 3: Data structures and algorithms for efficient frequent pattern mining and querying**
- **Part 4: It is often logic that triumphs in data analysis, not mechanical use of datamining, machine learning, and statistical methods**
- **Part 5: Science of data analysis exists, but:**
 - Assumptions often don't hold and not checked
 - Practices often not followed

