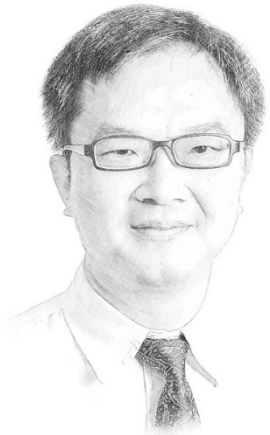
Big data and a bewildered lay analyst

Wong Limsoon

NUS
National University
of Singapore

About Limsoon



Position

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

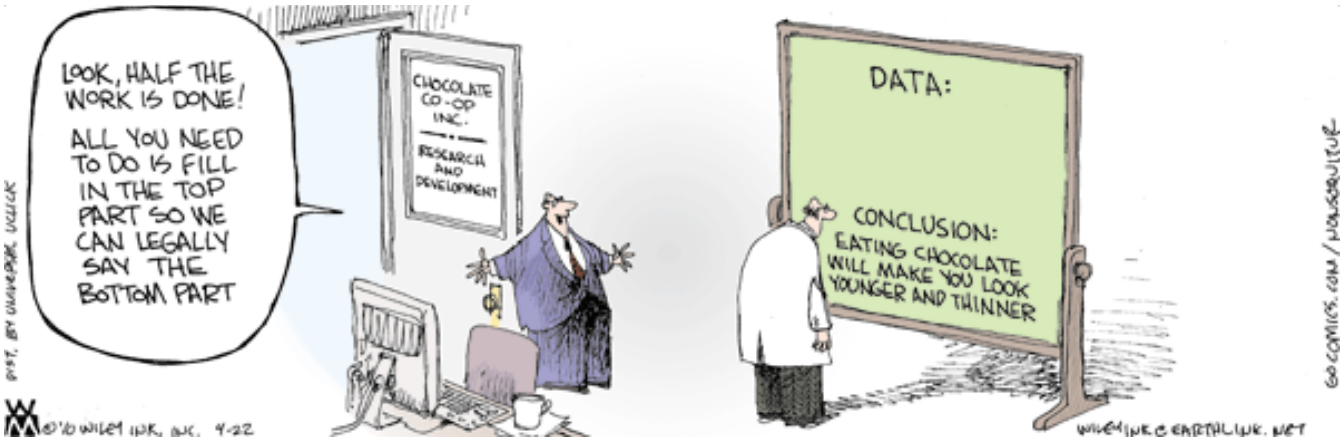
Research

Database systems & theory, knowledge discovery, bioinformatics & computational biology

Honours

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

Lecture plan



Make it easy to formulate hypothesis

Extraction from big, integrated databases

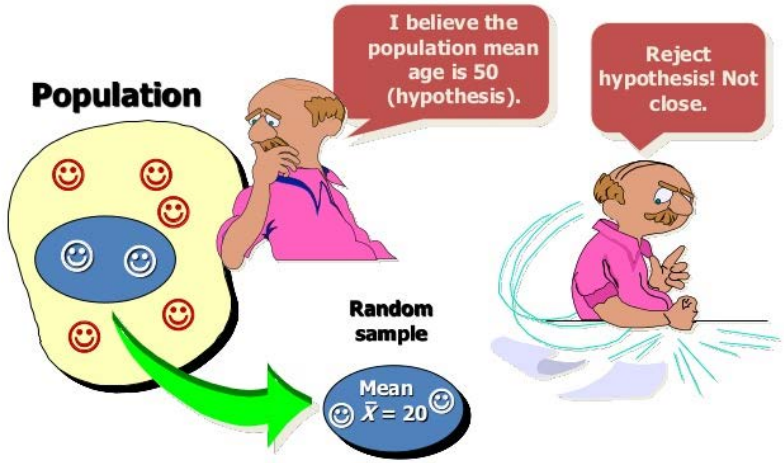
Make hypothesis testing sound

Detection & correction of assumption violations

Find better hypothesis & explain why it is better

E.g., “for men, taking A is better than B”

HYPOTHESIS TESTING



AM I TESTING THIS HYPOTHESIS CORRECTLY?

SNP	Genotypes	Group: Controls [n(%)]		Group: 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.

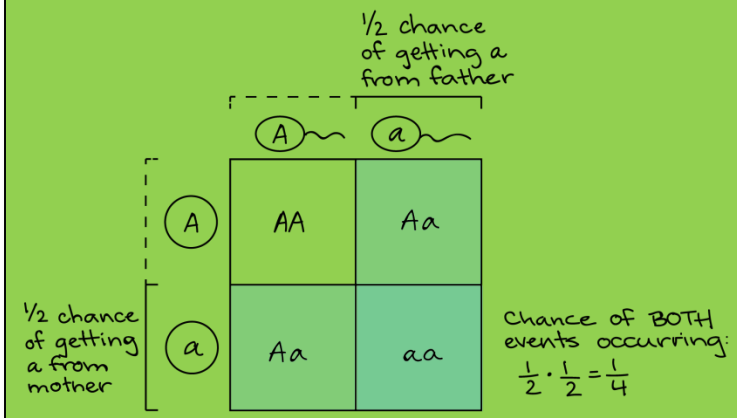
A seemingly obvious conclusion

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 χ2 test. Anything more/wrong?

Sample may not be fidel to real-world population



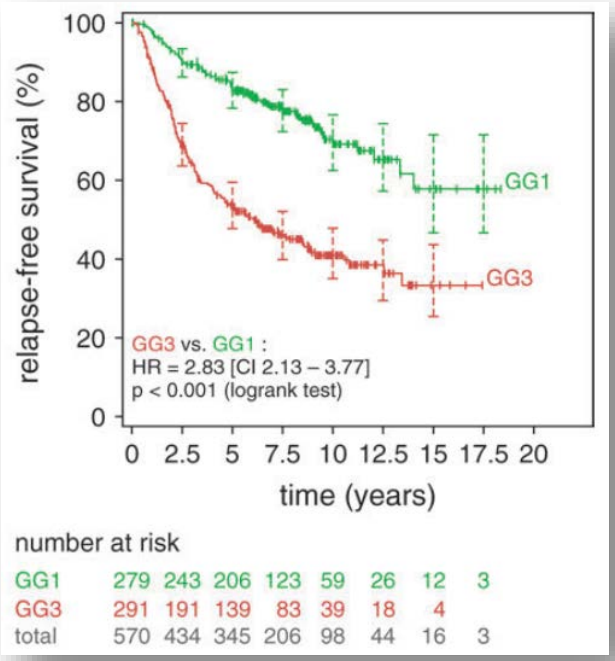
Basic rule of human genetics

SNP	Genotypes	Group: Controls [n(%)]		Group: 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.

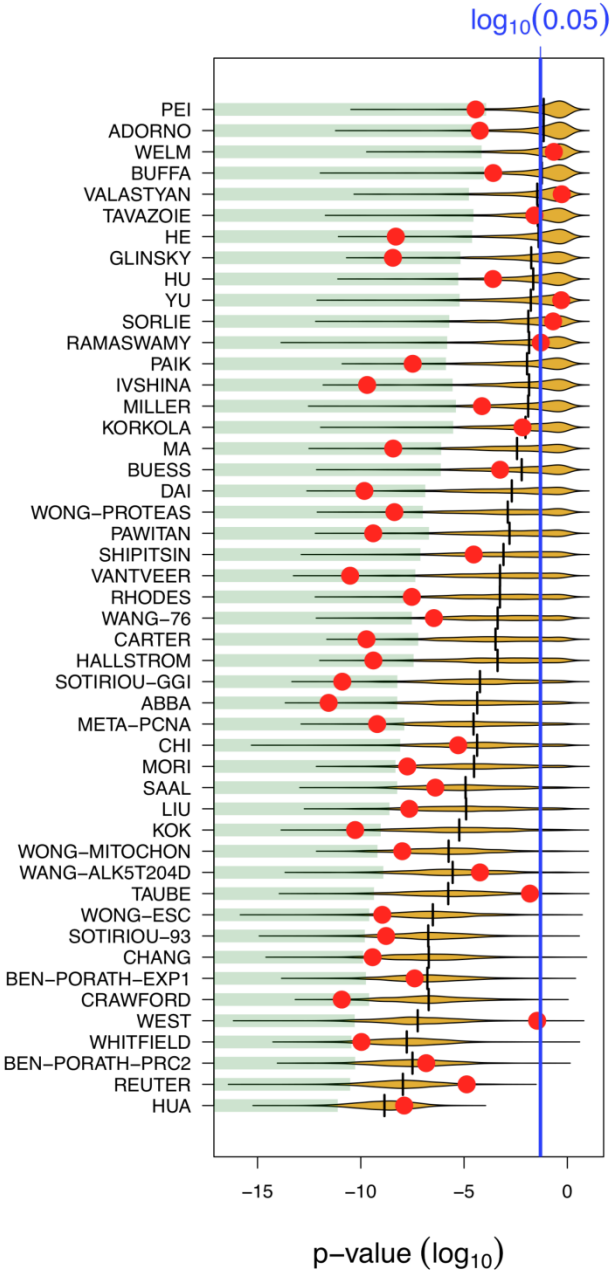
AG = 38 + 79 = 117, controls + cases = 189 ⇒ population is ~62% AG ⇒ population is >9% AA, unless AA is lethal

"Big data check" shows AA is non-lethal for this SNP ⇒ sample is biased



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?**



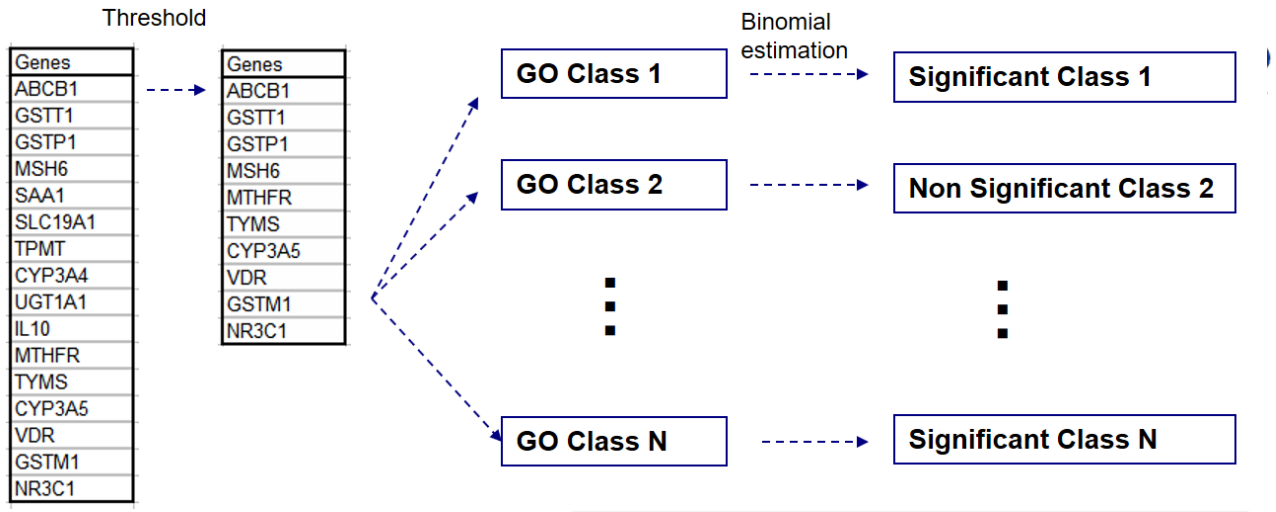
Null distribution may not be appropriate

- **Almost all random signatures also have p-value << 0.05**

⇒ **null model is confounded**

⇒ **significant signatures can't be trusted; they are no better than random ones!**

A seemingly obvious conclusion



ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (e.g., genes whose t-statistic meets the 5% significance threshold of t-test), and checking the significance of the size of the intersection using the hypergeometric test

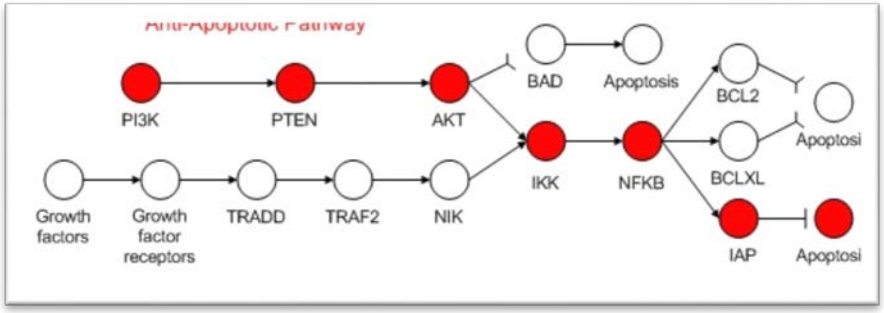
- **A pathway is claimed as an explanation for a disease phenotype as it is enriched with differentially expressed genes**
  - ORA p-value << 0.05
- **A straightforward hypergeometric test. Anything more/wrong?**

Null hypothesis may be inappropriate

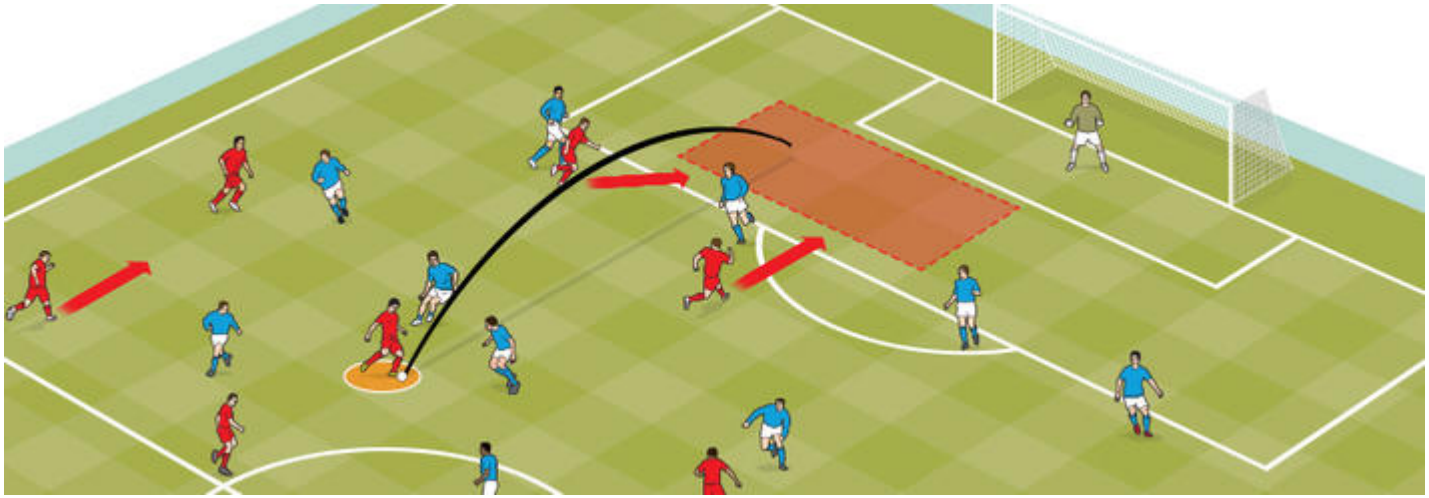
- **The null hypothesis basically says "Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size"**

- **This null hypothesis is obviously false**

$\Rightarrow$ **Lots of false positives**



- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behavour of genes in a pathway is more coordinated than random ones



ARE THERE TACTICS FOR DERIVING DEEPER INSIGHT FROM DATA?

A seemingly obvious conclusion

Context
Race = White

Occupation	Income>50K	Income<50K
Adm-clerical	439 (14%)	2,645 (86%)
Craft-repair	844 (23%)	2,850 (77%)

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 $\chi 2$ test. Anything more/wrong?

Exception as deeper insight

Context
Race = White, Workclass = Self-emp-not-inc

Occupation	Income>50K	Income<50K
Adm-clerical	16 (35%)	30 (65%)
Craft-repair	90 (18%)	409 (82%)

The "unincorporated self-employed" work class is an exception to the conclusion that "craft repairers tend to earn more than administrative clerks"

Contradictions as deeper insight

Context	Occupation	Income>50K	Income<50K
Race = White, Sex = Male	Adm-clerical	251 (24%)	787 (76%)
	Craft-repair	829 (24%)	2,695 (76%)

Context	Occupation	Income>50K	Income<50K
Race = White, Sex = Female	Adm-clerical	188 (9%)	1,858 (91%)
	Craft-repair	15 (9%)	155 (91%)

The conclusion "craft repairers tend to earn more than administrative clerks" holds for neither male nor female

The conclusion is an artefact of male earning more than female

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

- $0.001 < \chi 2$ test p-value < 0.01 is significant

A straightforward $\chi 2$ test. Anything more/wrong?

Trend-strengthening subpopulation as deeper insight

Computation of the $\chi 2$

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**

$\chi 2$ 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

CAN THESE TACTICS BE AUTOMATED?

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**

⟨{Race=Chinese}, Drug=A|B, Response=positive⟩

Generating a hypothesis: Think in terms of contingency tables

⟨{Race=Chinese}, Drug=A|B, Response=positive⟩

To test this hypothesis we need info:

- N^A =support({Race=Chinese, Drug=A})
- N^A_{pos} =support({Race=Chinese, Drug=A, Res=positive})
- N^B =support({Race=Chinese, Drug=B})
- N^B_{pos} =support({Race=Chinese, Drug=B , Res=positive})

Context	Comparing Attribute	response= positive	response= negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{pos}$
	Drug=B	N^B_{pos}	$N^B - N^B_{pos}$

⇒ **Frequent pattern mining**

Algo for hypothesis generation

A hypothesis is a comparison between two or more sub-populations, and each sub-population 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

Add values of an extra attribute A to context of H

Re-calculate test statistic

- **Test statistic is reversed ➔ Exception?**
- **Test statistic becomes insignificant ➔ Contradiction?**
- **Test statistic is strengthened ➔ Better explanation?**

All done via immediate superset search on frequent patterns

- **A frequent pattern ≈ a population**
- **A superset of a frequent pattern ≈ a subpopulation**

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

NUS National University of Singapore

iDIG
Intelligent Data-Driven Insights Generator

Uncovering Hidden Insights with Data-Driven Hypothesis Testing

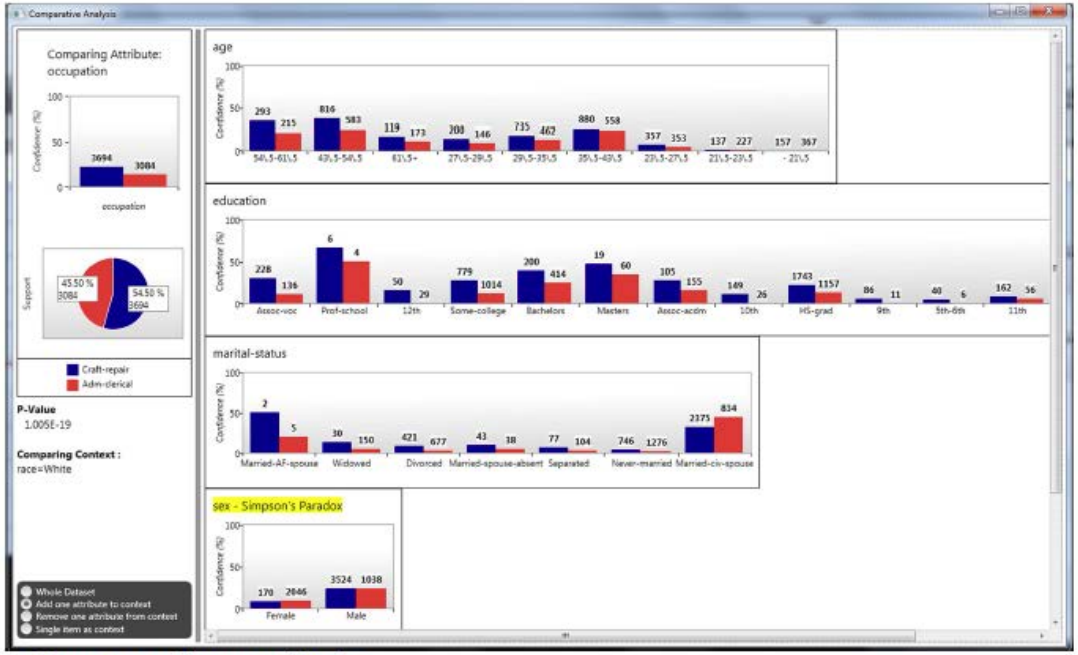
Examples

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

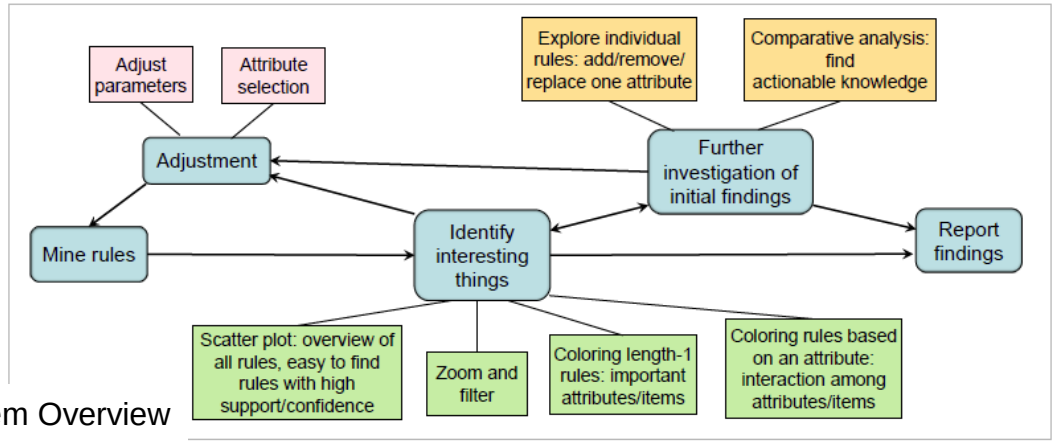
An example dataset

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

Running time

Three phases

Frequent pattern mining

Hypothesis generation

Hypothesis analysis

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_{target}/V_{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)

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

ART OF DATA ANALYSIS

There is only so much a data mining or hypothesis exploration system can do for you automatically

You need to do some logical thinking when using these systems or looking at their outputs

- Don’t ignore non-associations
- Don’t ignore context
- Ensure a conclusion is independent of other factors

And your data may be telling more than you think

We tend to ignore non-associations

Many technologies for association and correlation mining

- Frequent patterns
- Association rules
- …

But ignore non-associations

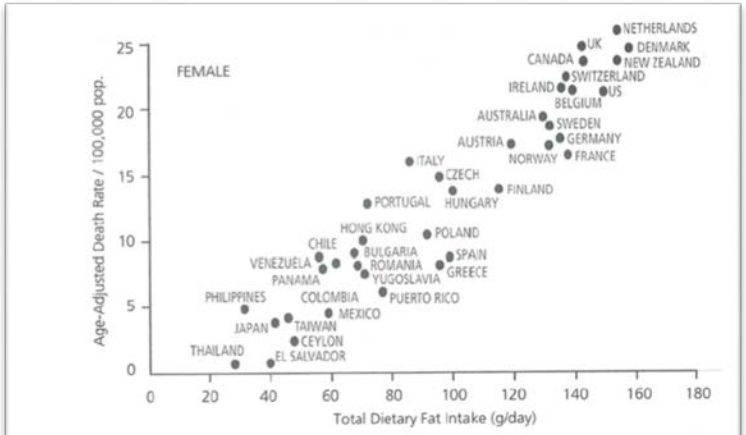
- Not interesting
- Too many of them

Is this a good thing?

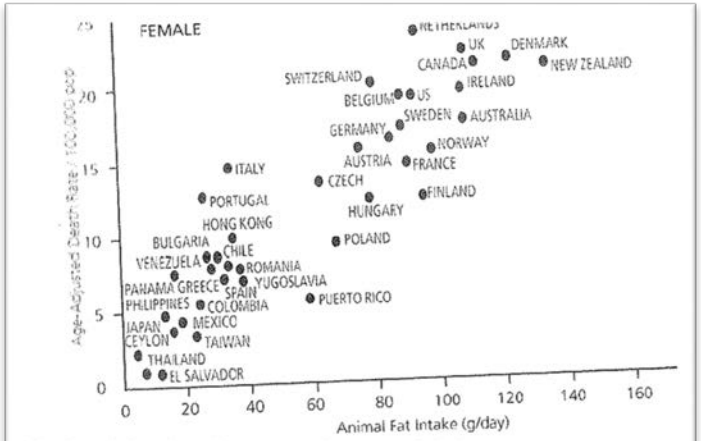
The power of negative space!

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

We love to find correlations like these…



Dietary fat intake correlates with breast cancer



Animal fat intake correlates with breast cancer

But not non-correlations like this…

Plant fat intake doesn't correlate with breast cancer

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!**

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

- ⟨{Race=Chinese}, Drug=A|B, Response=positive⟩

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

If A/B treat the same single disease, it is ok

If B treats two diseases, but A one, it is not sensible

⇒ The disease has to go into the context

Madrid and Warsaw are at almost the same distance to Latium cities

Are Madrid and Warsaw near each other?

Distances of European cities (km) from the main cities of Latium

	Rome	Latina	Frosinone	Viterbo	Rieti
Amsterdam	430	447	449	415	409
Athens	347	321	331	346	364
Barcelona	283	305	293	292	271
Beograd	227	222	236	220	238
Berlin	393	400	409	374	373
Bern	227	249	247	220	205
Bonn	353	370	372	339	330
Bruselles	388	406	406	371	365
Bucharest	364	355	368	359	378
Budapest	268	261	274	246	259
Calais	418	448	446	418	405
Copenhagen	510	522	527	492	491
Dublin	622	645	641	615	600
Edinburgh	637	655	655	625	615
Frankfurt	318	333	336	302	295
Hamburg	435	448	453	417	414
Helsinki	727	729	739	706	713
Istanbul	452	430	443	443	464
Lisbon	615	637	622	624	604
London	474	494	493	464	456
Luxembourg	325	346	346	315	307
Madrid	449	470	458	460	440
Marseille	200	223	213	202	183
Moscow	782	773	785	759	774
Munich	230	245	250	216	213
Oslo	664	675	682	646	645
Paris	365	386	383	357	343
Prague	305	313	320	286	290
Sofia	294	273	286	280	301
Stockholm	653	658	668	632	636
Warsaw	435	433	444	413	421
Vienna	255	254	265	233	240
Zurich	227	246	246	214	205

Giuliani et al., Physics Letters A, 247:47-52, 1998

PCA of distance matrix of European cities to Latium cities

Factor loadings and proportions of explained variance

Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569	0.000043	0.000005

PC1 accounts for >99% of variance

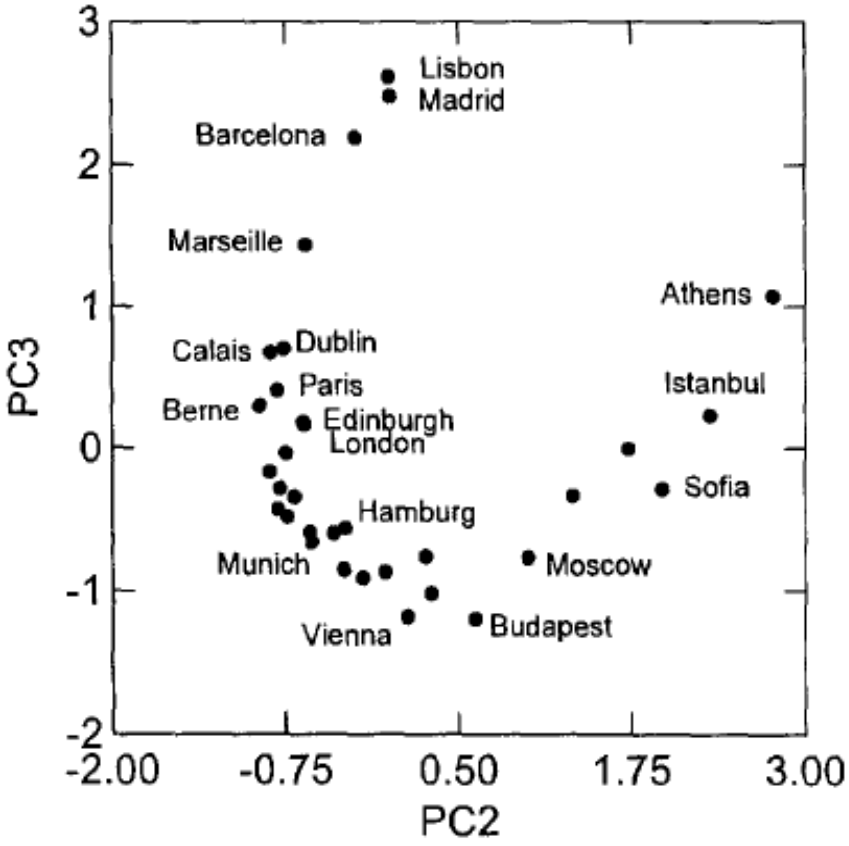
PC1 correlates with distance of European cities to Latium cities

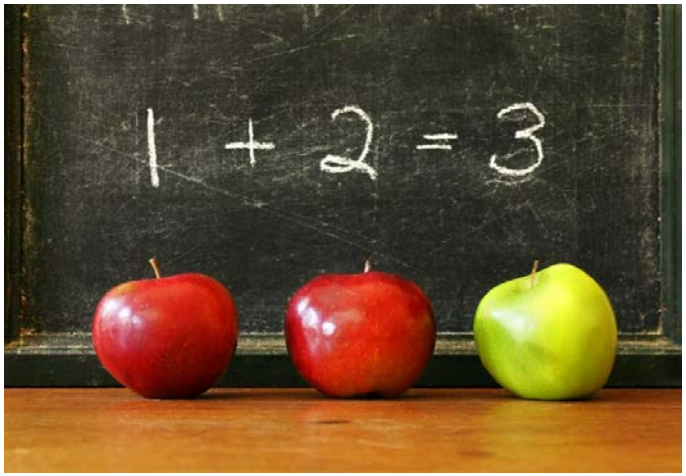
PC2, PC3, … account for < 1% of variance

Are PC2, PC3, … useless / non-informative?

PC2 & PC3 are the angular orientation of European cities centered on Latium

So you can tell Madrid is not near Warsaw





SUMMARY

NUS National University of Singapore

What have we learned?

It is easy to make mistakes when analyzing data

Think in terms of contingency tables

Look for subpopulations causing exception, contradiction, & trend strengthening

Mechanical use of data mining, statistical test, etc. can only take you so far

and we can do these automatically and efficiently

Limsoon Wong, "Big data and a bewildered lay analyst", *Statistics & Probability Letters*, 2018