# Some simple tactics for deriving a deeper analysis of data

**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**

# Plan

Part 1: Helpful analytics

**Simple mechanical tactics to make data analysis more insightful**

Part 2: Exploratory hypothesis testing & analysis

**Translating these tactics into datamining tasks**

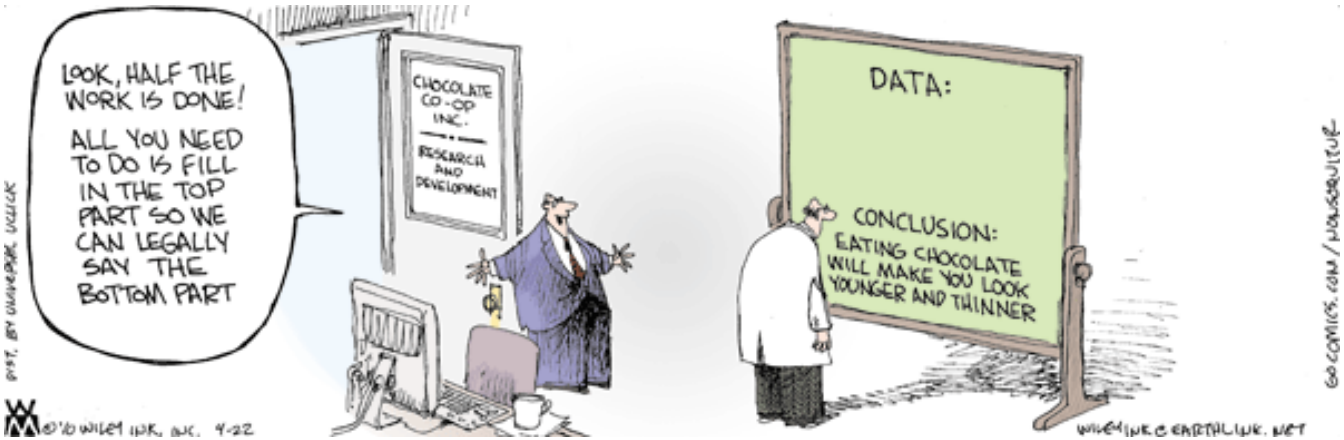
Part 3: Art & science of data analysis

**Beyond the mechanical**

# Part 1:
# Helpful analytics

NUS
National University
of Singapore

# 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”

# 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,<br>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**

NUS National University of Singapore

| 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

| SNP | Genotypes | Group: Controls [n(%)] | | Group: Cases [n(%)] | | $\chi^2$ | *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.

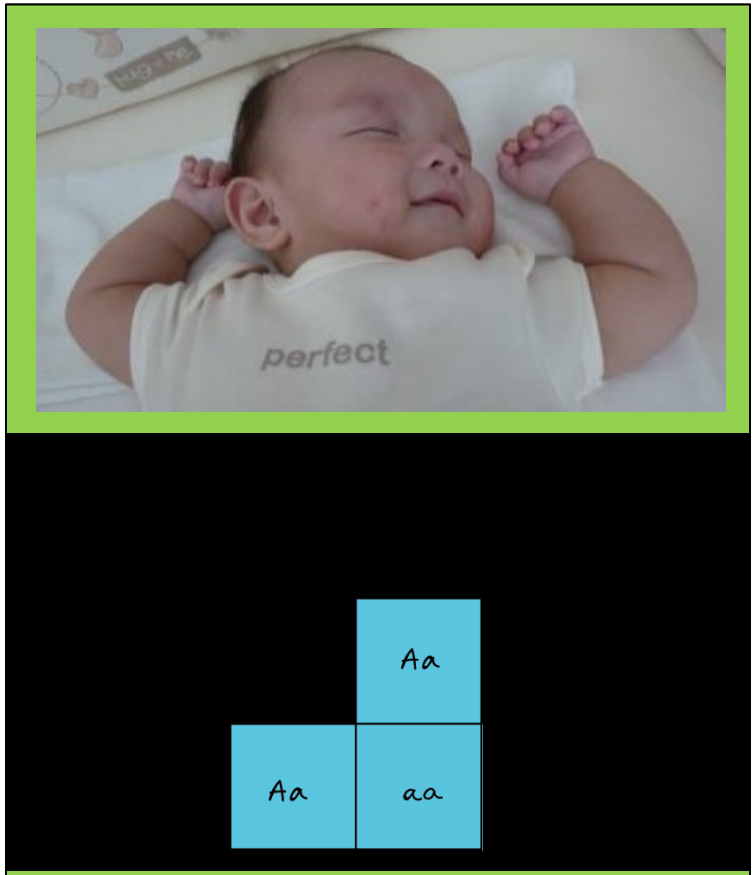
A seemingly obvious conclusion

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

# Sample bias is revealed by domain logic




**Basic rule of human genetics**

| SNP | Genotypes | Group: Controls [n(%)] | | Cases [n(%)] | | $\chi^2$ | *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**

NUS National University of Singapore

What have we learned?

**Think in terms of contingency tables**

**Look for subpopulations causing exception, contradiction, and/or trend strengthening**

**Some times must also use simple domain logic to detect problems**

Can we do these automatically and efficiently?

# Part 2: Exploratory hypothesis testing & analysis


NUS
National University
of Singapore

# 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 algorithms

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

Just a few hypotheses got tested at a time

**So much data have been collected …**

No clue on what to look for

# Exploratory hypothesis testing

NUS National University of Singapore

**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**

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

# Generating a hypothesis: Think in terms of contingency tables

NUS National University of Singapore

**⟨{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**

# Need for hypothesis analysis

**Lots of contingency tables (i.e. hypotheses) can be generated quickly …**

- Exploration is not guided by domain knowledge

⇒ Spurious hypotheses have to be eliminated

- Reasons behind significant hypotheses

⇒ Find attribute-value pairs that affect the test statistic a lot

**Alternatively, generate & explore hypotheses incrementally, starting from the most general?**

# Spurious hypotheses … detected by looking at subpopulations

| | 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 for significant hypotheses … found by looking at subpopulations

| | 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

# Algorithm 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


Uncovering Hidden Insights with Data-Driven Hypothesis Testing

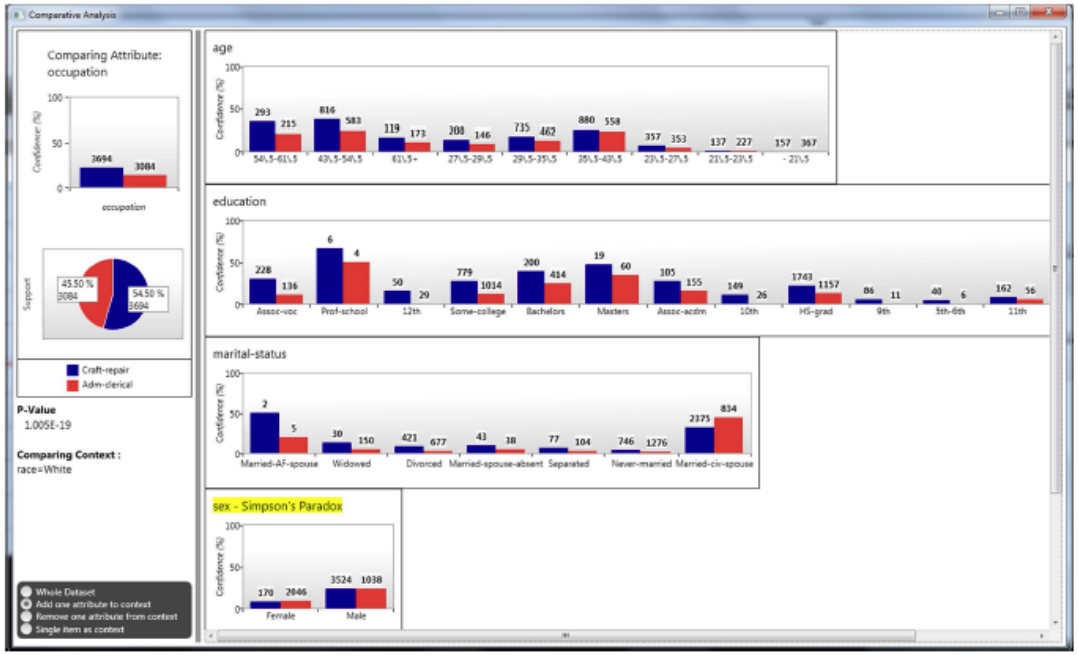
## Examples

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

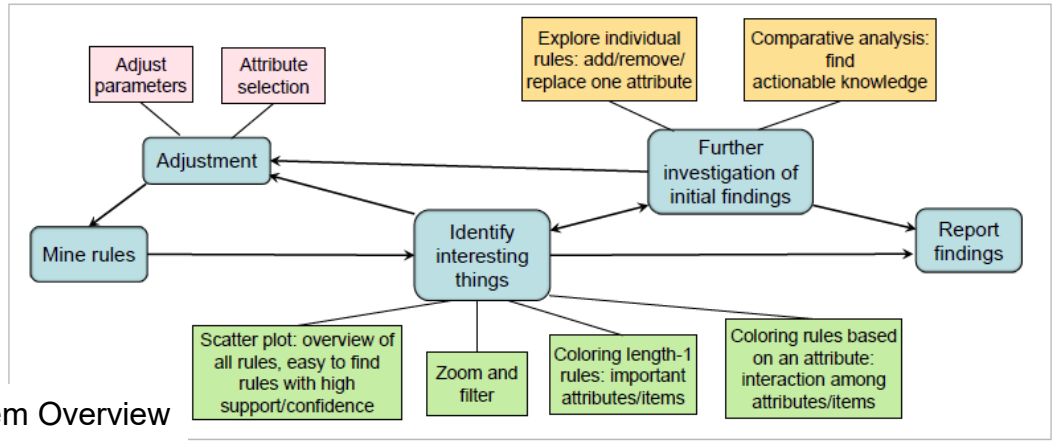
An example dataset

**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 | $A_{target}/v_{target}$ |
|---|---|---|---|---|
| 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) |

# 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

# Part 3:
# Art & science of data analysis

NUS
National University
of Singapore

NUS National University of Singapore

**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**

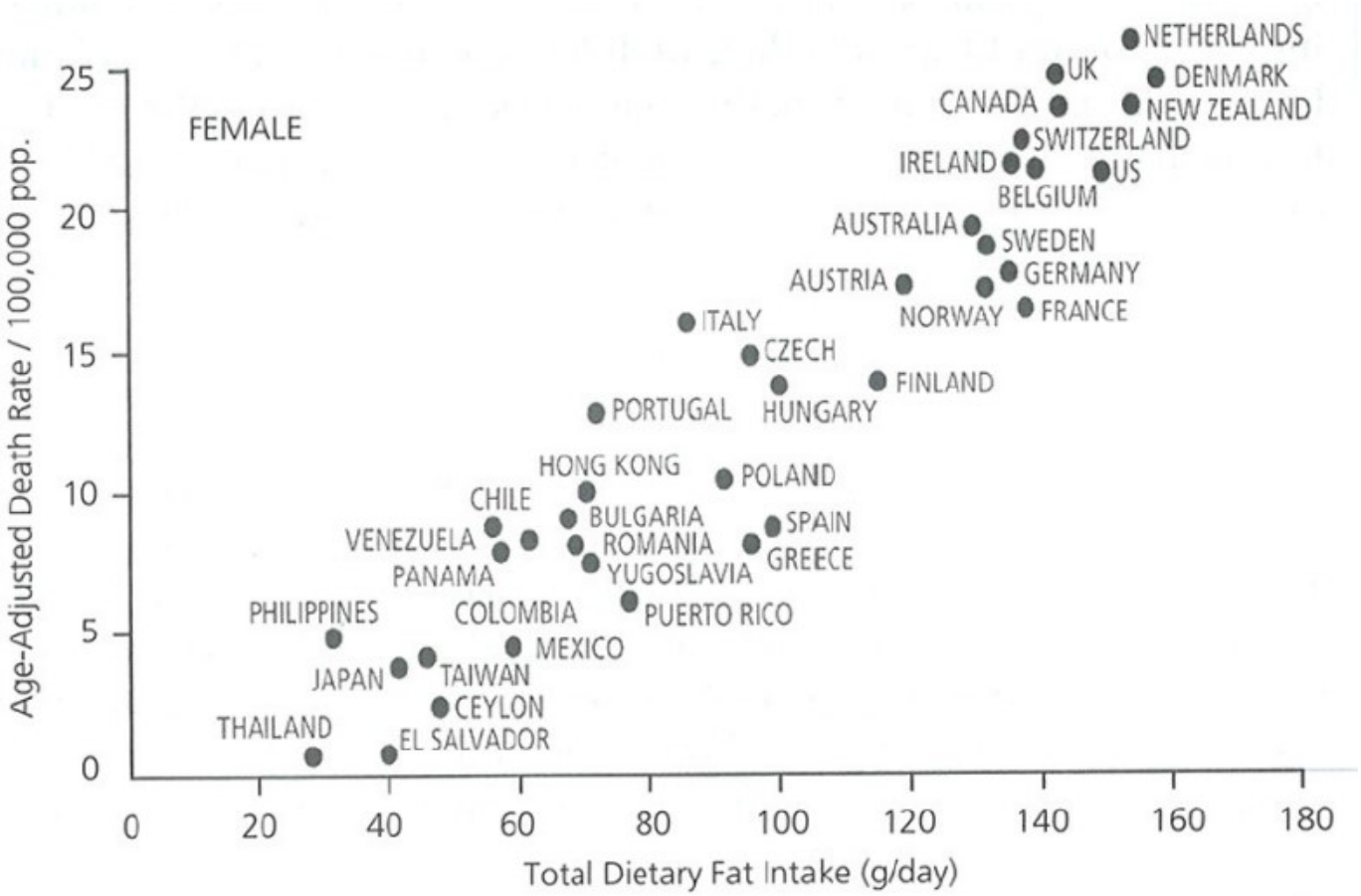
- Not interesting
- Too many of them

**Is this a good thing?**


**The power of negative space!**

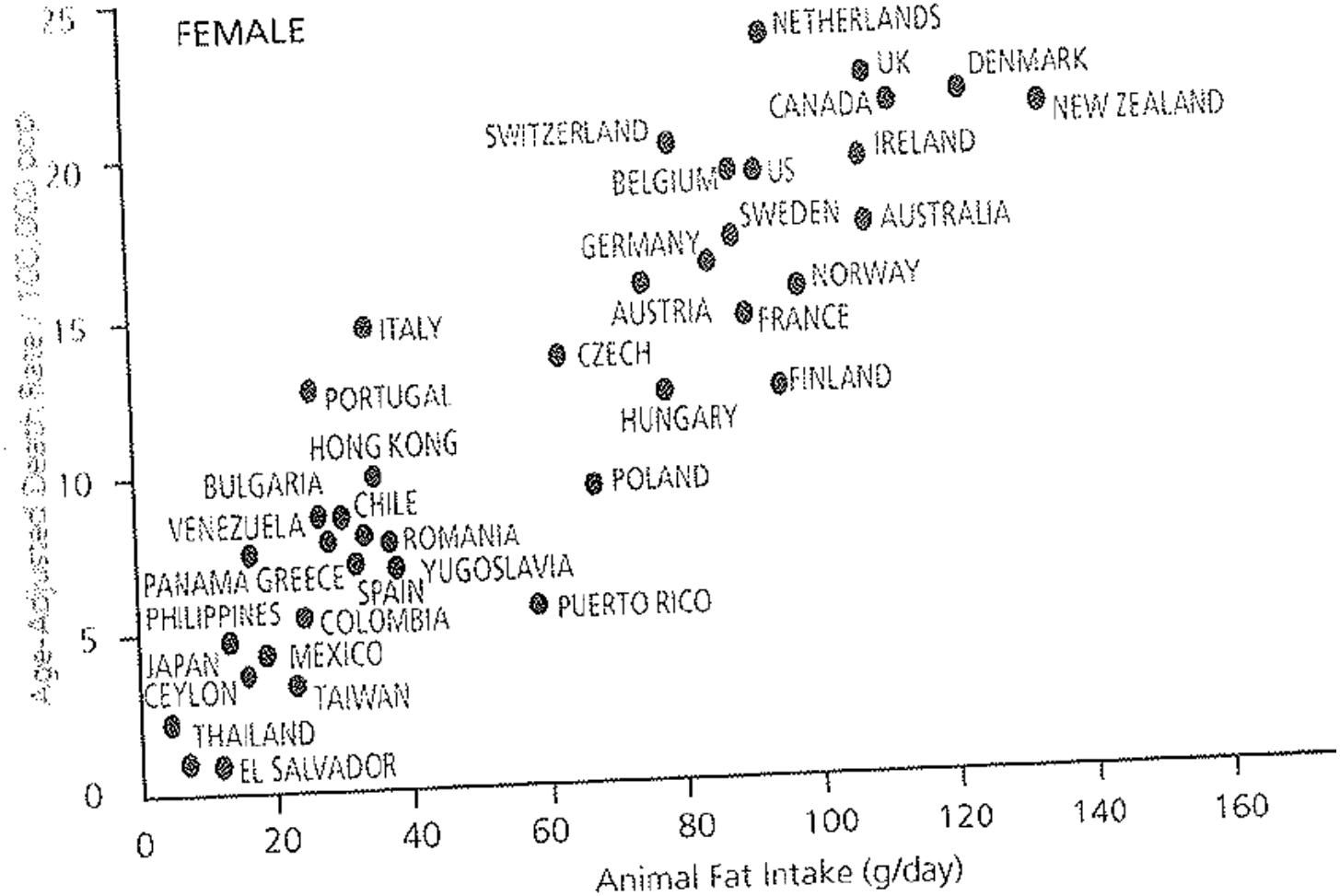
- **How many animals do you see?**

# We love to find correlations like this…



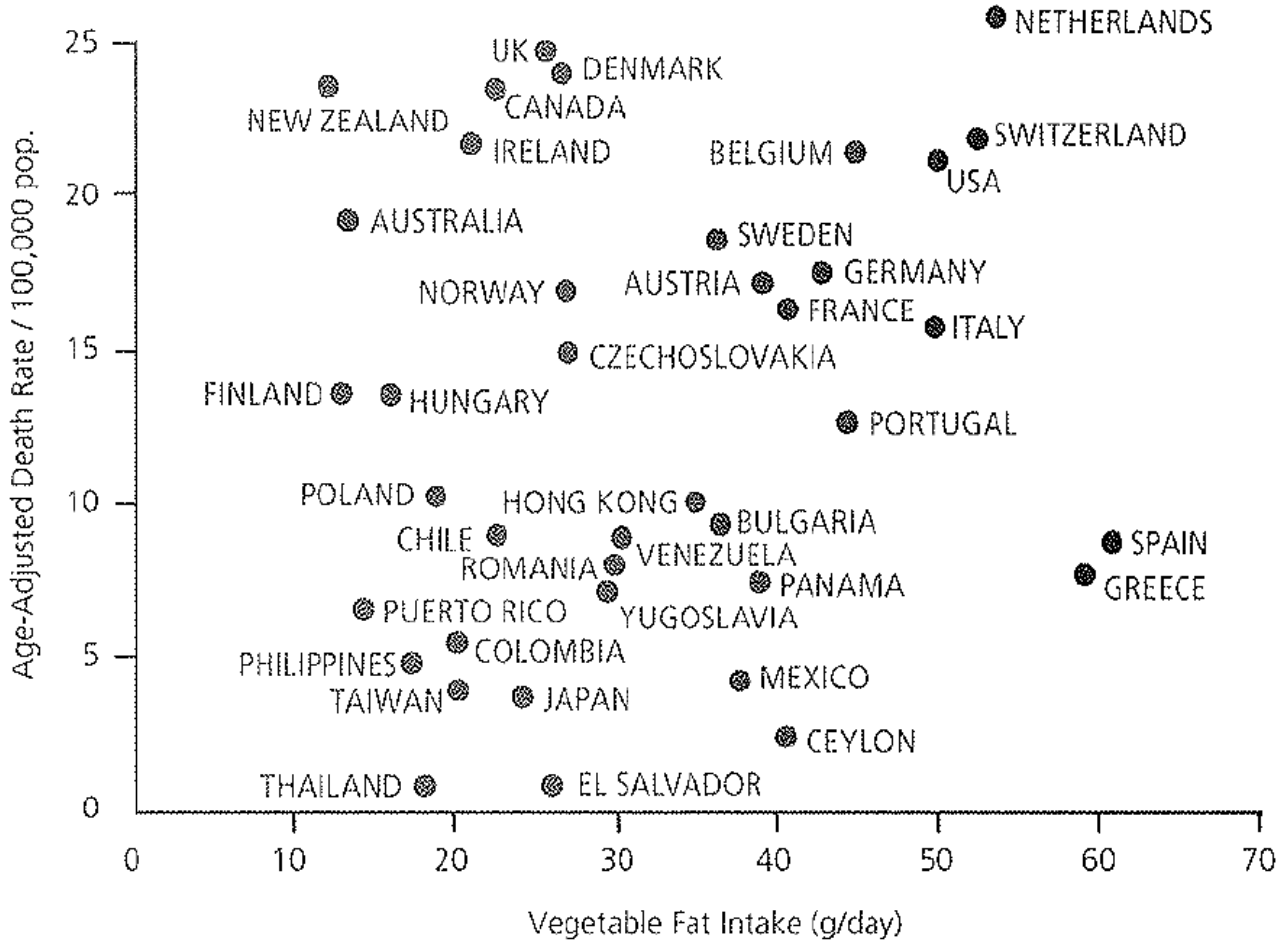

**Dietary fat intake correlates with breast cancer**

# 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**

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

NUS National University of Singapore

**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**

$\Rightarrow$ **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**

# We don't check independence

**In clinical testing, we carefully choose the sample to ensure the test is independent of other factors**

- Patients are not related
- 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!**

# 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 |

**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 |

## Taking A

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

## Taking B

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

## Men taking A

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

## Men taking B

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

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**

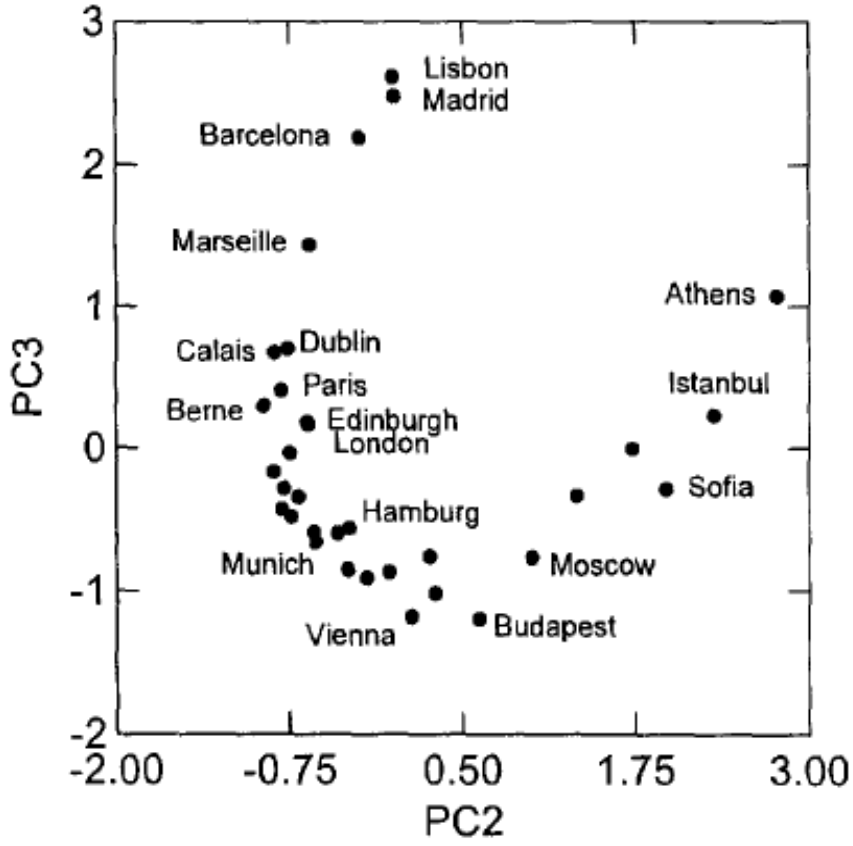
PC1 correlates with distance of European cities to Latium cities

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

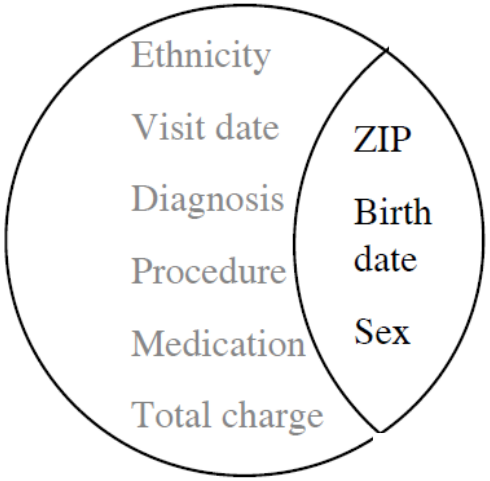
**Are PC2, PC3, … useless / non-informative?**

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

So you can tell Madrid is not near Warsaw



# Is anonymized data really anonymous?

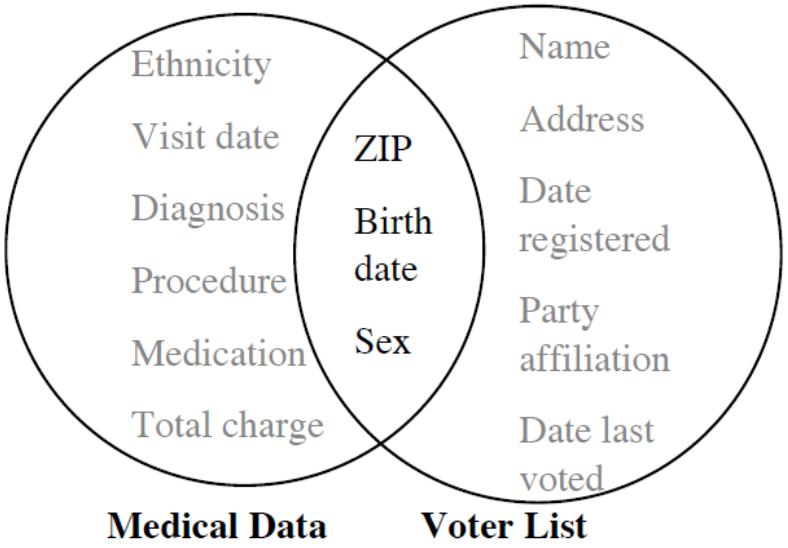



**Medical Data**

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!

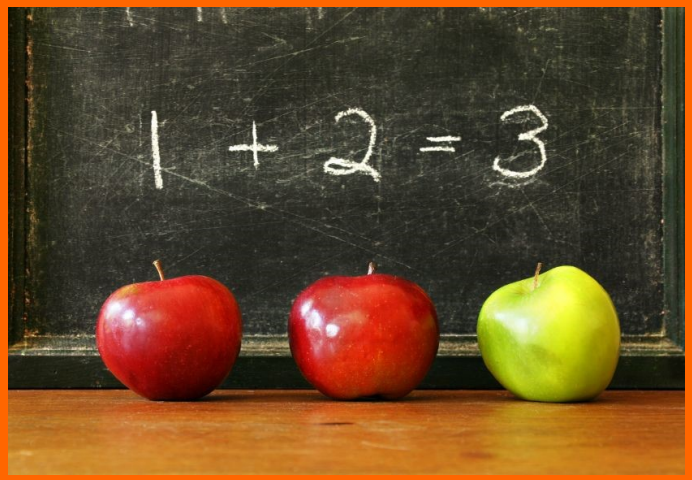



For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes [4]. The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using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

# Summary




NUS
National University
of Singapore

# 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: **It is often logic that triumphs in data analysis, not mechanical use of datamining, machine learning, and statistical methods**

# Good to read

Guimei Liu, Haojun Zhang, Mengling Feng, Limsoon Wong, See-Kiong Ng. **Supporting exploratory hypothesis testing and analysis**. *ACM Transactions on Knowledge Discovery from Data*, 9(4):Article 31, April 2015

Wei Zhong Toh, Kwok Pui Choi, Limsoon Wong. **Redhyte: A self-diagnosing, self-correcting, and helpful hypothesis analysis platform**. *Journal of Information and Telecommunication*, 1(3):241--258, July 2017

Limsoon Wong. **Big data and a bewildered lay analyst**. *Statistics & Probability Letters*, to appear