# Some often-overlooked issues in analytics

**Wong Limsoon**

National University of Singapore

# What is big data and why

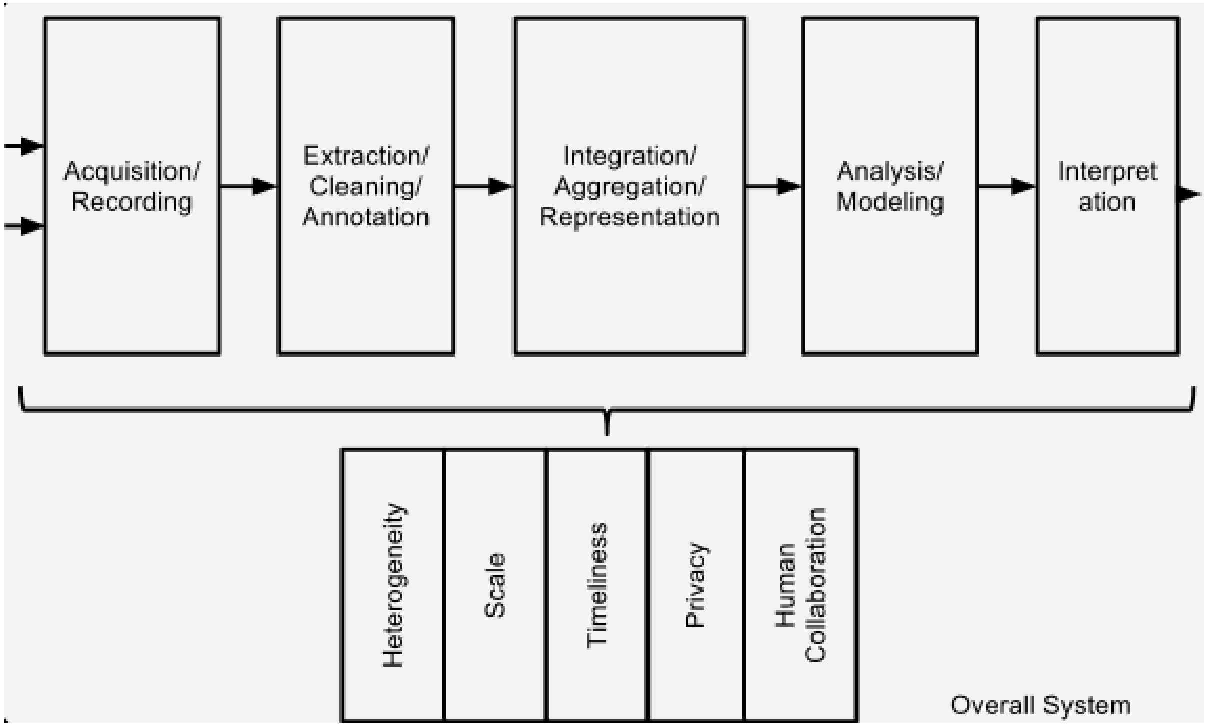
- **Big data *a la* Gartner**
  - Volume, velocity, variety
- **Other characteristics**
  - Veracity, v...

**A practical definition**

**“More than you know how to handle”**

- **Why big data?**
  - Can collect cheaply, due to automation
  - Can store cheaply, due to falling media prices
  - Many success stories, where useful predictions were made with the data

# Challenges in big data



- **Much emphasis is on scaling issues**
- **But there are non-scaling-related issues that affect fundamental assumptions in current bioinformatics and statistical analysis**
  – Big data may break analysis procedures in fundamental ways

# Talk outline

- **Forgotten assumptions**
  - Normal distribution
  - The 1st “I” in I.I.D.
  - The 2nd “I” in I.I.D.
- **Overlooked information**
  - Non-associations
  - Context
- **More may not be better**
  - Protein complexes
  - Causal genes

I purchased a stock which my broker says can give me 50% capital gains in 12 months
But did he tell you how much value this stock can lose in a given month?

## Forgotten assumptions

# NORMAL DISTRIBUTION

# Wisdom of the crowd

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

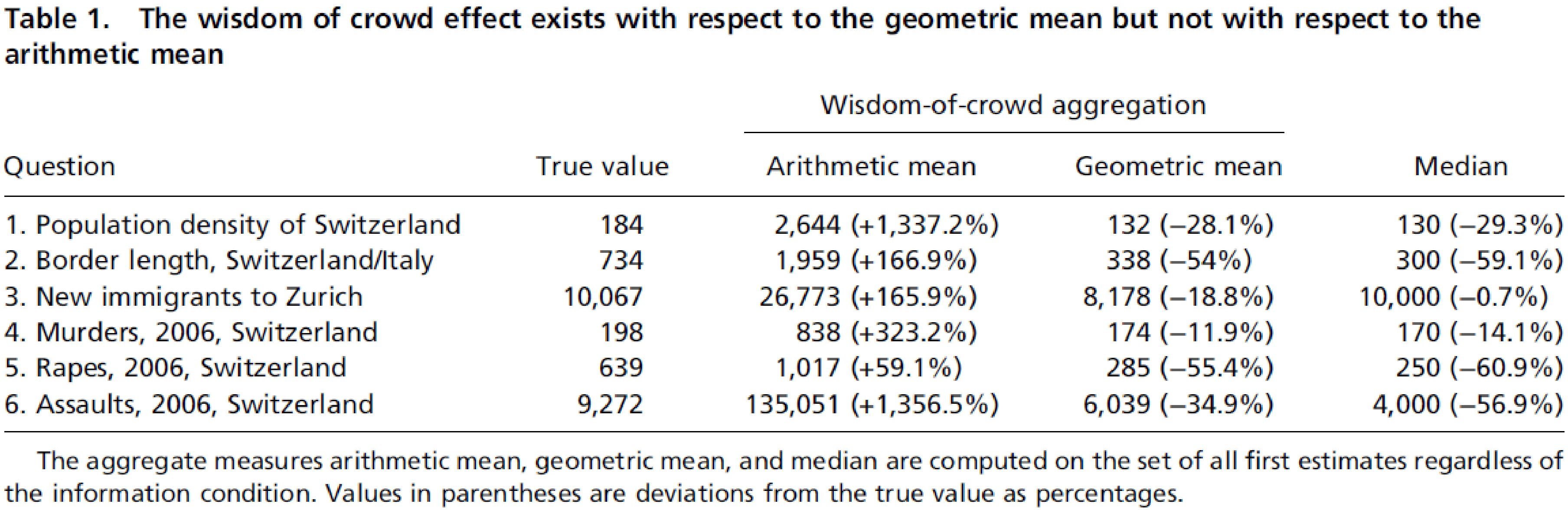


**Table 1. The wisdom of crowd effect exists with respect to the geometric mean but not with respect to the arithmetic mean**

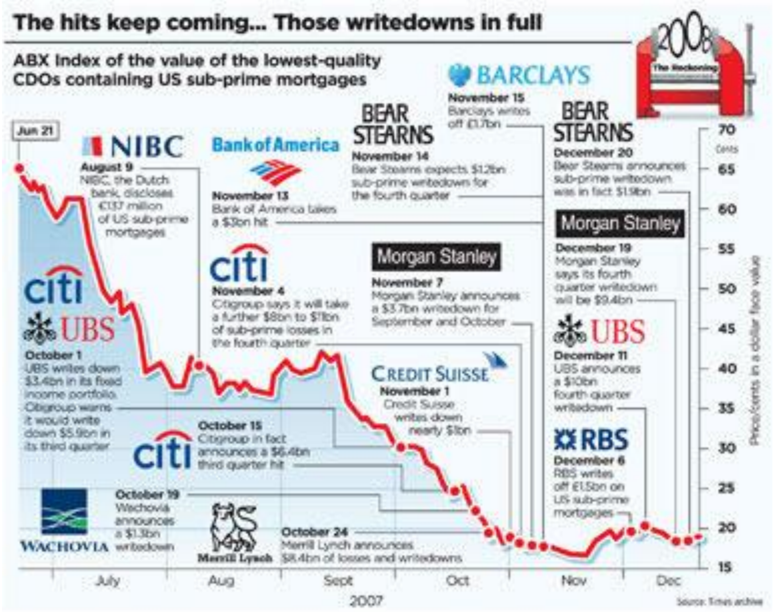
| Question | True value | Wisdom-of-crowd aggregation<br>Arithmetic mean | Geometric mean | Median |
|---|---|---|---|---|
| 1. Population density of Switzerland | 184 | 2,644 (+1,337.2%) | 132 (–28.1%) | 130 (–29.3%) |
| 2. Border length, Switzerland/Italy | 734 | 1,959 (+166.9%) | 338 (–54%) | 300 (–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**

# 2007 Financial Crisis



The hits keep coming... Those writedowns in full

ABX Index of the value of the lowest-quality CDOs containing US sub-prime mortgages

August 9 NIBC, the Dutch bank, discloses €137 million of US sub-prime mortgages

October 1 UBS writes down $3.4bn in its fixed income portfolio. Citigroup warns it would write down $5.9bn in its third quarter

October 15 Citigroup in fact announces a $6.4bn third quarter hit

October 19 Wachovia announces a $1.3bn writedown

October 24 Merrill Lynch announces $8.4bn of losses and writedowns

November 1 Credit Suisse writes down nearly $1bn

November 4 Citigroup says it will take a further $8bn to $11bn of sub-prime losses in the fourth quarter

November 7 Morgan Stanley announces a $3.7bn writedown for September and October

November 13 Bank of America takes a $3bn hit

November 14 Bear Stearns expects $1.2bn sub-prime writedown for the fourth quarter

November 15 Barclays writes off £1.7bn

December 6 RBS writes off £1.5bn on US sub-prime mortgages

December 11 UBS announces a $10bn fourth quarter writedown

December 19 Morgan Stanley says its fourth quarter writedown will be $9.4bn

December 20 Bear Stearns announces sub-prime writedown was in fact $1.9bn

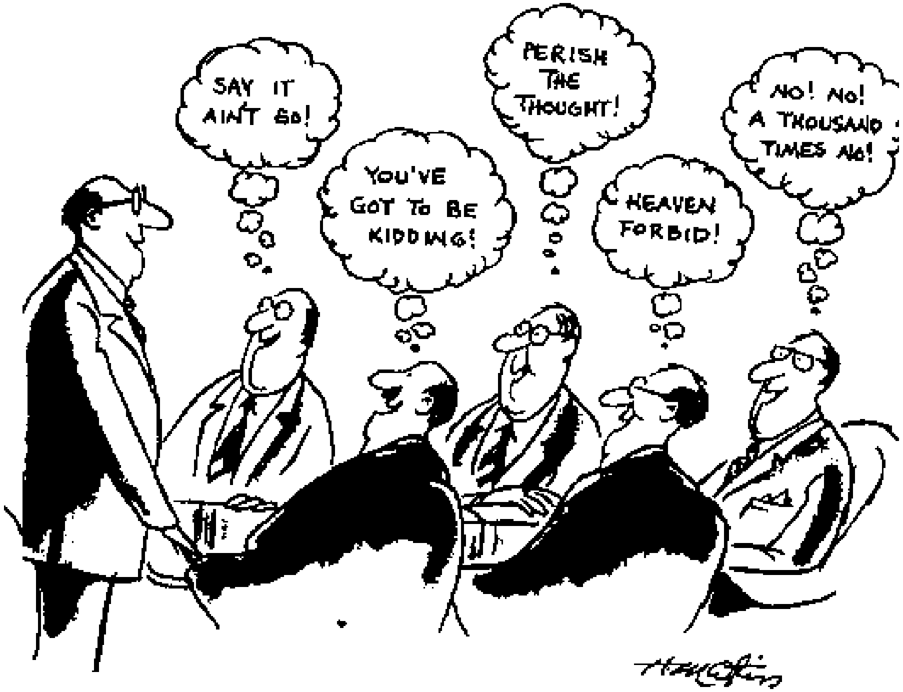
Source: Times archive

- **All of them religiously check VaR (Value at Risk) everyday**

- **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 those in favor say 'Aye.'"*
*"Aye."*
*"Aye."*
*"Aye."*
*"Aye."*
*"Aye."*

## Forgotten assumptions

# THE 1ST "I" IN I.I.D.

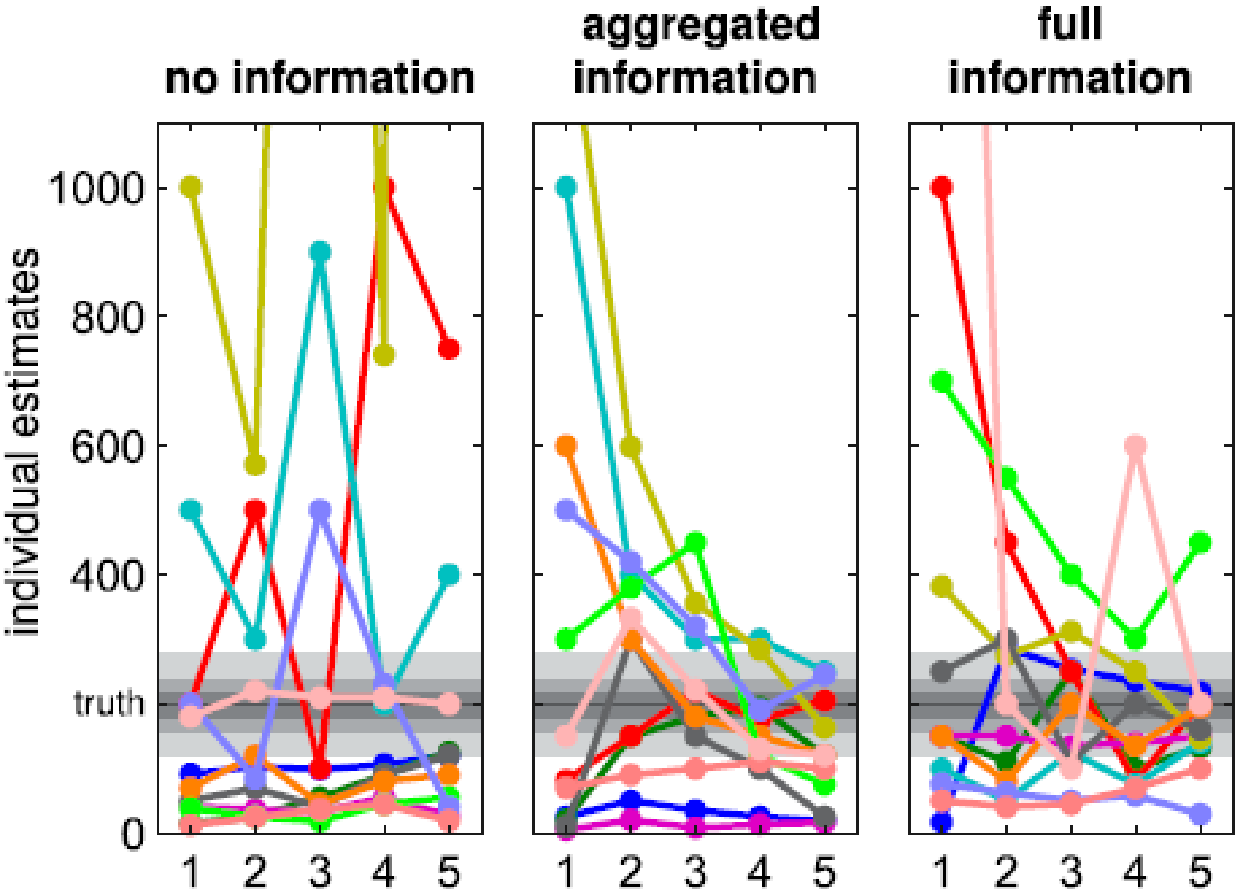
# Experiments on social influence

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**

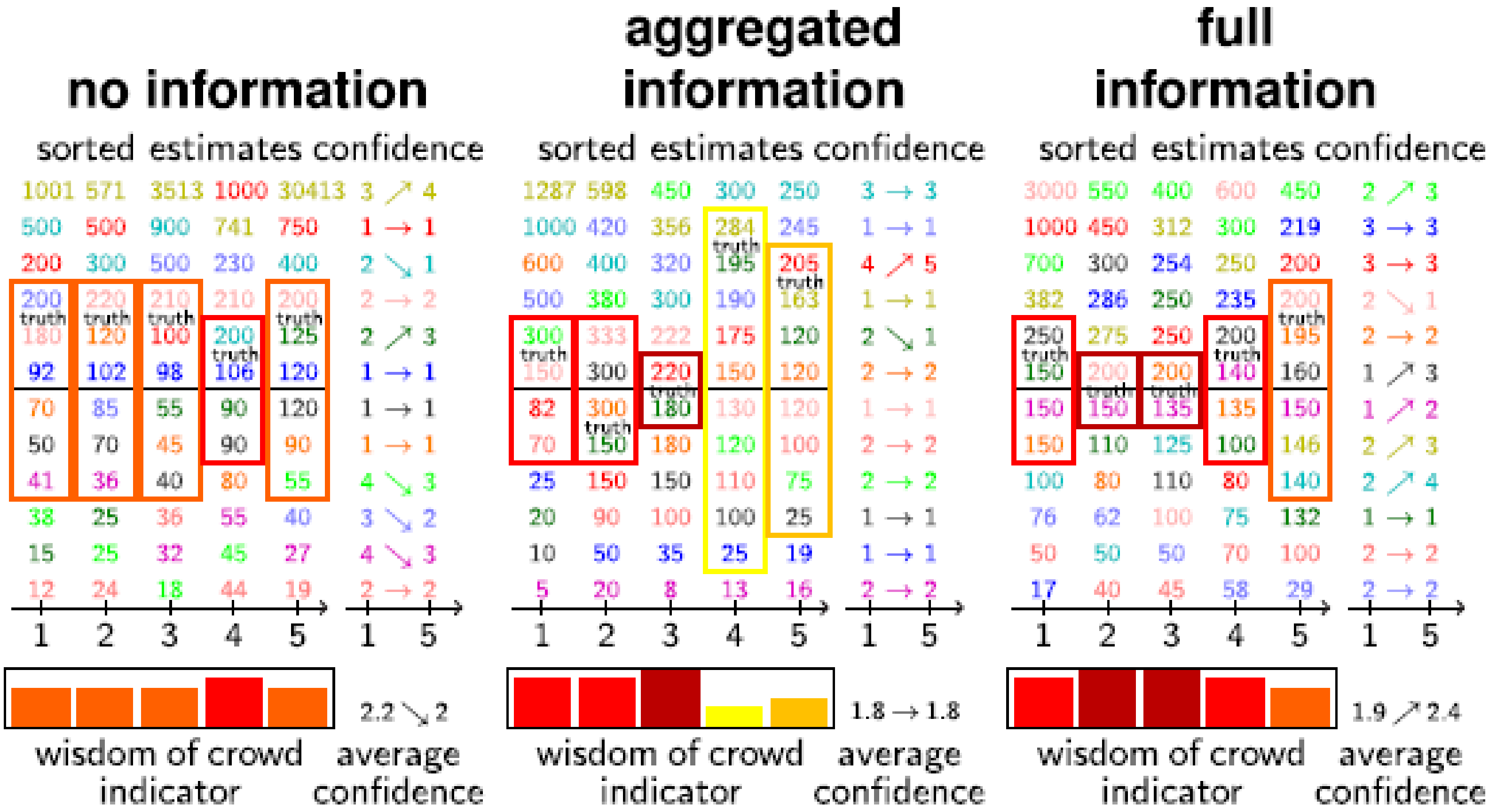
# Social influence effect



- **Social influence diminishes diversity in groups**

⇒ **Groups potentially get into “group think”!**

# Range reduction effect



no information

| sorted estimates | | | | | confidence |
|---|---|---|---|---|---|
| 1001 | 571 | 3513 | 1000 | 30413 | 3 ↗ 4 |
| 500 | 500 | 900 | 741 | 750 | 1 → 1 |
| 200 | 300 | 500 | 230 | 400 | 2 ↘ 1 |
| 200 | 220 | 210 | 210 | 200 | 2 → 2 |
| truth 180 | truth 120 | truth 100 | 200 | truth 125 | 2 ↗ 3 |
| 92 | 102 | 98 | truth 106 | 120 | 1 → 1 |
| 70 | 85 | 55 | 90 | 120 | 1 → 1 |
| 50 | 70 | 45 | 90 | 90 | 1 → 1 |
| 41 | 36 | 40 | 80 | 55 | 4 ↘ 3 |
| 38 | 25 | 36 | 55 | 40 | 3 ↘ 2 |
| 15 | 25 | 32 | 45 | 27 | 4 ↘ 3 |
| 12 | 24 | 18 | 44 | 19 | 2 → 2 |
| 1 | 2 | 3 | 4 | 5 | 1 5 |

2.2 ↘ 2

wisdom of crowd indicator

average confidence

aggregated information

| sorted estimates | | | | | confidence |
|---|---|---|---|---|---|
| 1287 | 598 | 450 | 300 | 250 | 3 → 3 |
| 1000 | 420 | 356 | 284 | 245 | 1 → 1 |
| 600 | 400 | 320 | truth 195 | 205 | 4 ↗ 5 |
| 500 | 380 | 300 | 190 | truth 163 | 1 → 1 |
| 300 | 333 | 222 | 175 | 120 | 2 ↘ 1 |
| truth 150 | 300 | 220 | 150 | 120 | 2 → 2 |
| 82 | 300 | truth 180 | 130 | 120 | 1 → 1 |
| 70 | truth 150 | 180 | 120 | 100 | 2 → 2 |
| 25 | 150 | 150 | 110 | 75 | 2 → 2 |
| 20 | 90 | 100 | 100 | 25 | 1 → 1 |
| 10 | 50 | 35 | 25 | 19 | 1 → 1 |
| 5 | 20 | 8 | 13 | 16 | 2 → 2 |
| 1 | 2 | 3 | 4 | 5 | 1 5 |

1.8 → 1.8

wisdom of crowd indicator

average confidence

full information

| sorted estimates | | | | | confidence |
|---|---|---|---|---|---|
| 3000 | 550 | 400 | 600 | 450 | 2 ↗ 3 |
| 1000 | 450 | 312 | 300 | 219 | 3 → 3 |
| 700 | 300 | 254 | 250 | 200 | 3 → 3 |
| 382 | 286 | 250 | 235 | 200 | 2 ↘ 1 |
| 250 | 275 | 250 | 200 | truth 195 | 2 → 2 |
| truth 150 | 200 | 200 | truth 140 | 160 | 1 ↗ 3 |
| 150 | truth 150 | truth 135 | 135 | 150 | 1 ↗ 2 |
| 150 | 110 | 125 | 100 | 146 | 2 ↗ 3 |
| 100 | 80 | 110 | 80 | 140 | 2 ↗ 4 |
| 76 | 62 | 100 | 75 | 132 | 1 → 1 |
| 50 | 50 | 50 | 70 | 100 | 2 → 2 |
| 17 | 40 | 45 | 58 | 29 | 2 → 2 |
| 1 | 2 | 3 | 4 | 5 | 1 5 |

1.9 ↗ 2.4

wisdom of crowd indicator

average confidence

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

# Social influence diminishes wisdom of the crowd

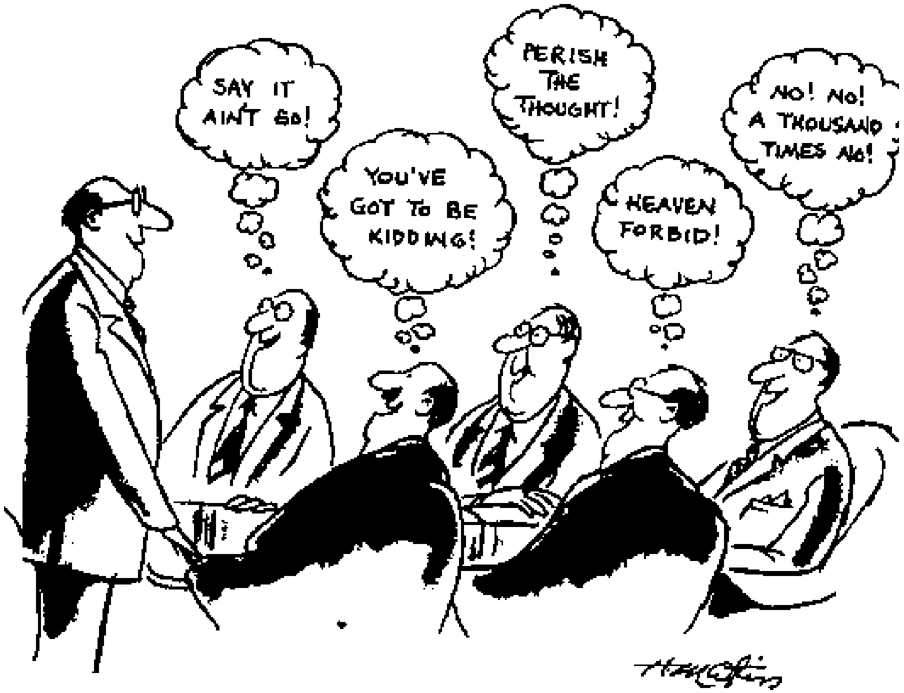
- Social influence triggers convergence of individual estimates

- 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

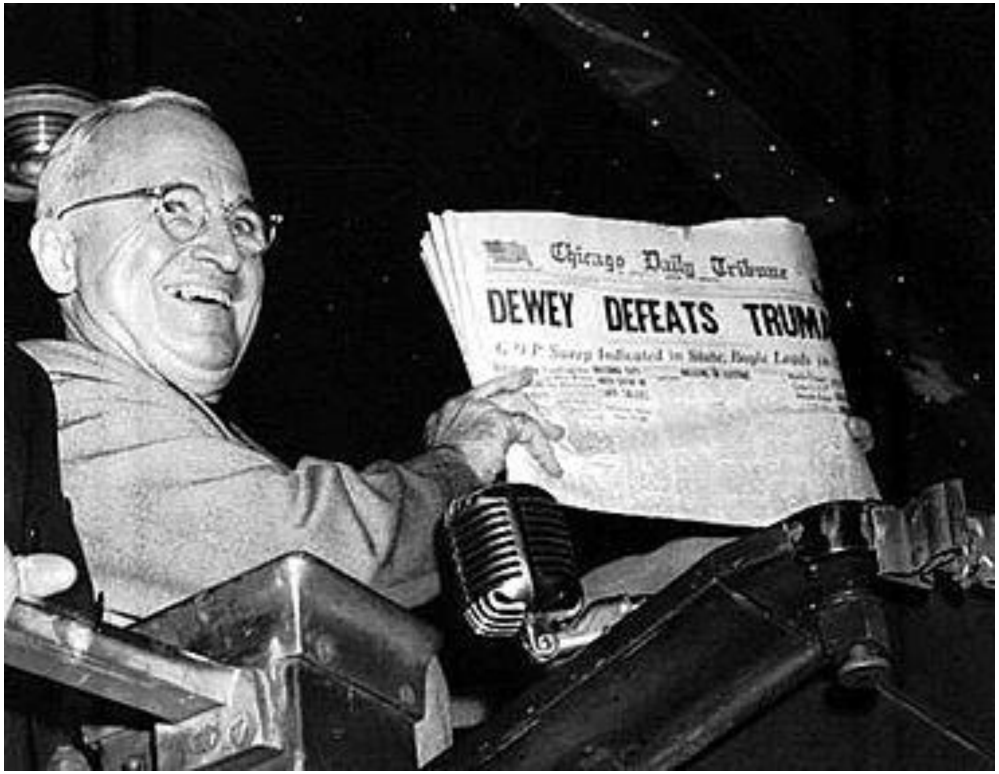
# Related issue: People do not say what they really want to say



"All those in favor say 'Aye.'"
"Aye."
"Aye."
"Aye."
"Aye."
"Aye."

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

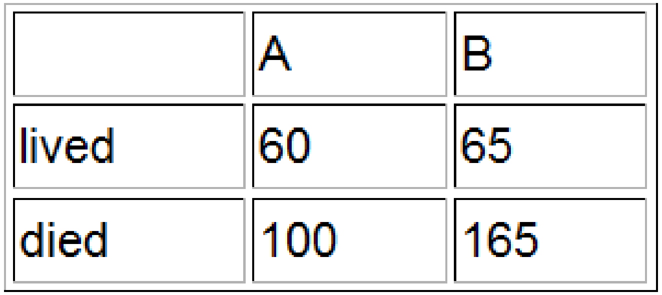
# THE 2$^{ND}$ "I" IN I.I.D.

# Statistical tests

- **Commonly used statistical tests (T-test, $\chi$2 test, Wilcoxon rank-sum test, …) all assume samples are drawn from independent identical distributions (I.I.D.)**

# How to ensure I.I.D.?

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

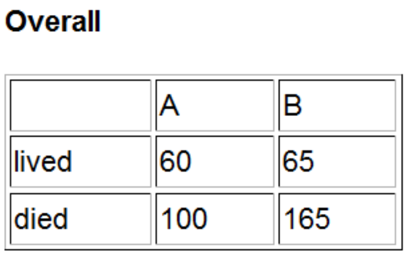


| | A | B |
|---|---|---|
| lived | 60 | 65 |
| died | 100 | 165 |

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?

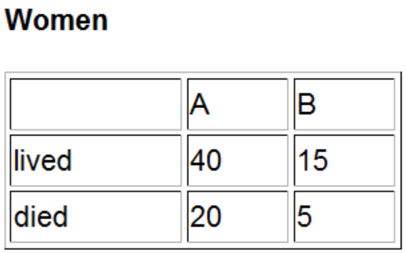
# What is happening here?



Overall

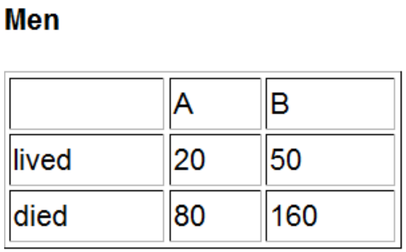
| | A | B |
|---|---|---|
| lived | 60 | 65 |
| died | 100 | 165 |

Looks like treatment A is better



Women

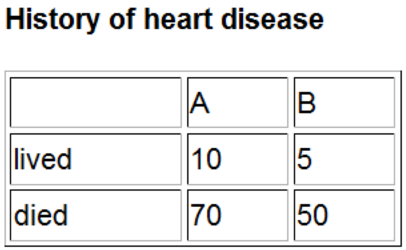
| | A | B |
|---|---|---|
| lived | 40 | 15 |
| died | 20 | 5 |



Men

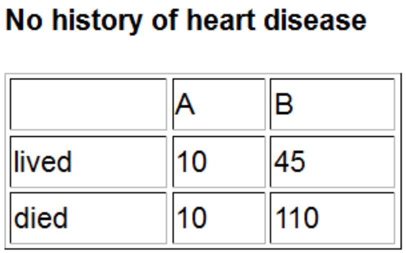
| | A | B |
|---|---|---|
| lived | 20 | 50 |
| died | 80 | 160 |

Looks like treatment B is better



History of heart disease

| | A | B |
|---|---|---|
| lived | 10 | 5 |
| died | 70 | 50 |



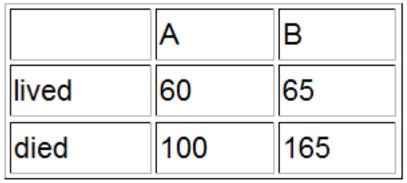
No history of heart disease

| | A | B |
|---|---|---|
| lived | 10 | 45 |
| died | 10 | 110 |

Looks like treatment A is better

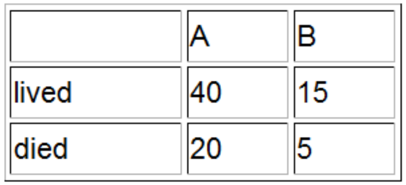
# Sample not identically distributed

Overall



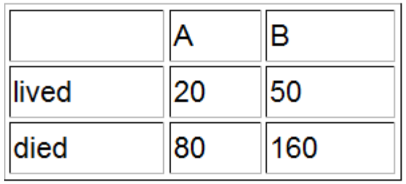
| | A | B |
|---|---|---|
| lived | 60 | 65 |
| died | 100 | 165 |

Women



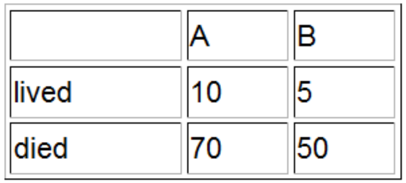
| | A | B |
|---|---|---|
| lived | 40 | 15 |
| died | 20 | 5 |

Men



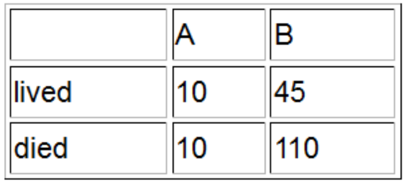
| | A | B |
|---|---|---|
| lived | 20 | 50 |
| died | 80 | 160 |

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%)

# Simpson's paradox in an Australian population census

| Context | Comparing Groups | sup | $P_{class=>50K}$ | p-value |
|---|---|---|---|---|
| Race =White | Occupation = Craft-repair | 3694 | 22.84% | 1.00 × $10^{-19}$ |
| | Occupation = Adm-clerical | 3084 | 14.23% | |

| Context | Extra attribute | Comparing Groups | sup | $P_{class=>50K}$ |
|---|---|---|---|---|
| Race =White | Sex = Male | Occupation = Craft-repair | 3524 | 23.5% |
| | | Occupation = Adm-clerical | 1038 | 24.2% |
| | Sex = Female | Occupation = Craft-repair | 107 | 8.8% |
| | | Occupation = Adm-clerical | 2046 | 9.2% |

- **Violation of the 2nd "I" of I.I.D.**
- **Btw,"men earn more than women" also violates the 2nd "I" in I.I.D.**

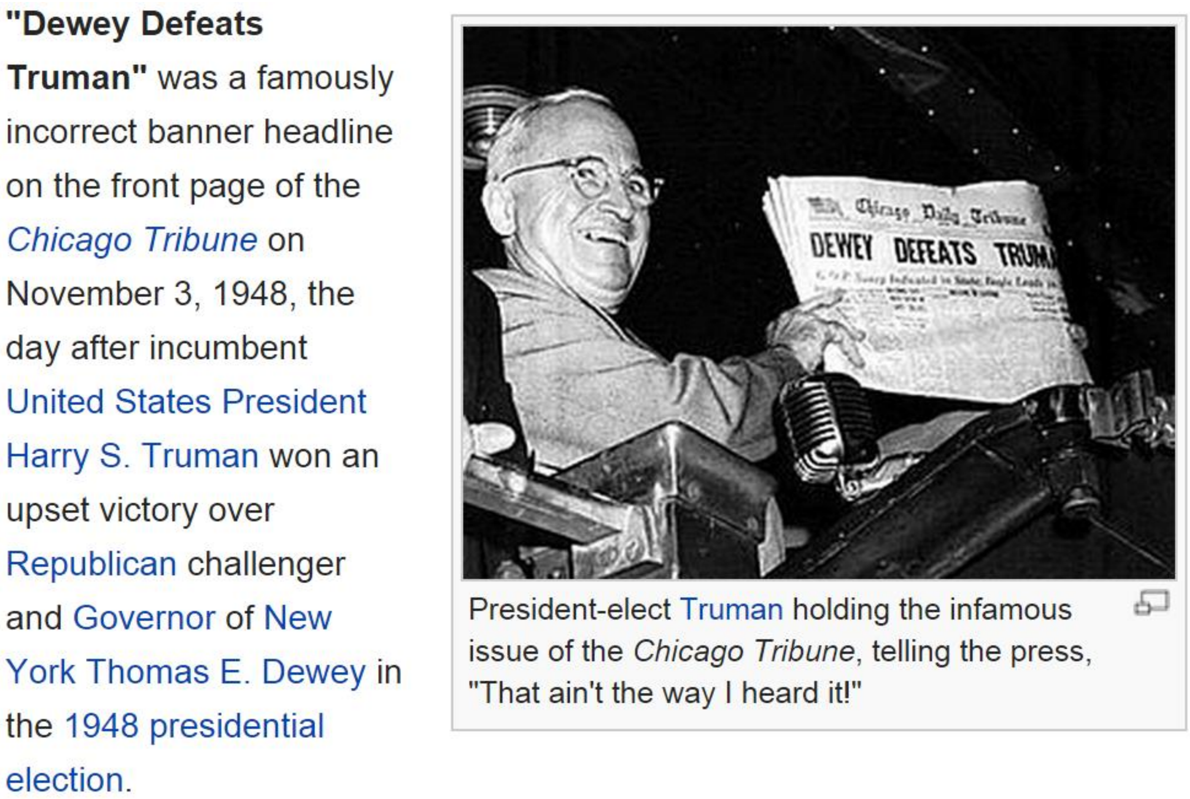
# Stratification

- **Cannot test "Men earn more than women" directly because I.I.D. is violated**
  - Different distributions of men & women wrt occupation

- **Test instead**
  - "$S_1$: For craftsmen, men earn more than women"
  - "$S_2$: For admin clerks, men earn more than women"
  - …

  **where craftsmen, admin clerks, … form an exhaustive list of disjoint occupations, provided each of $S_1$, $S_2$, … is valid**

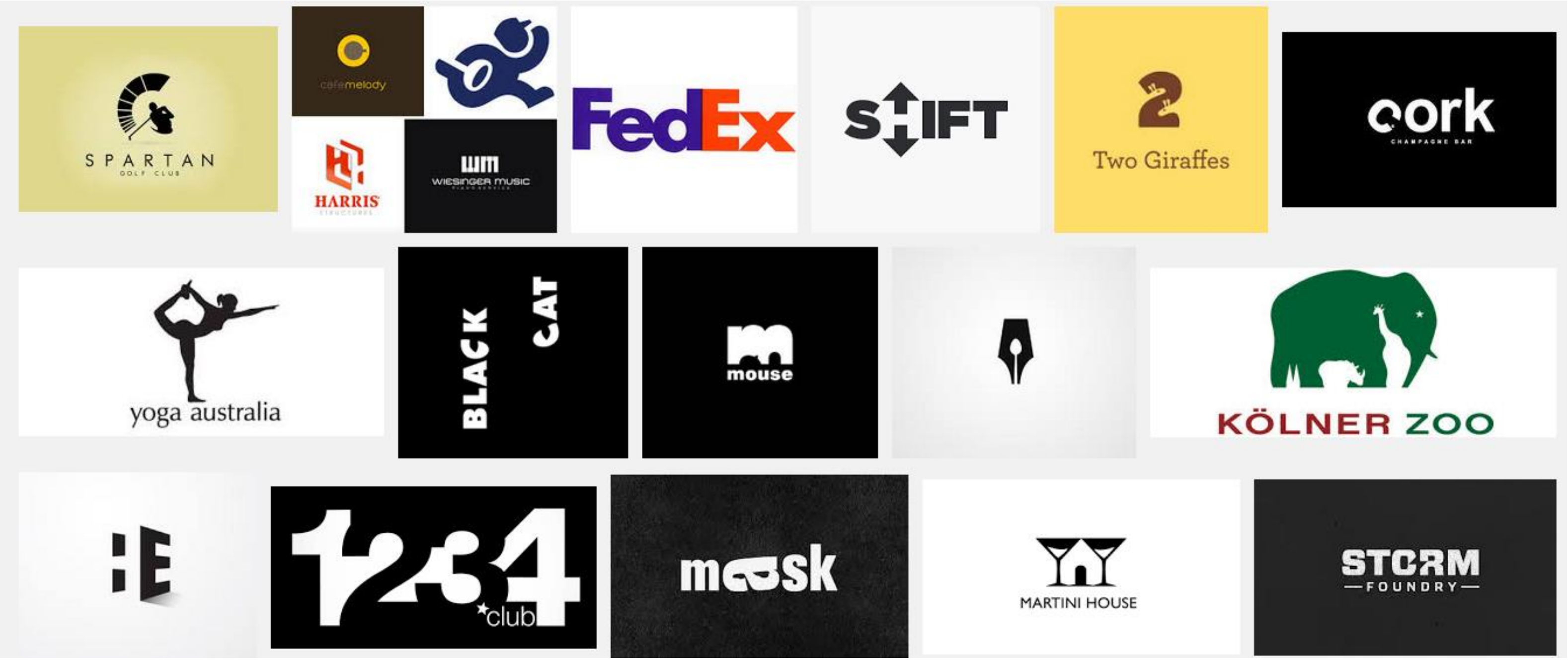
# Related issue: Sampling bias



**"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.



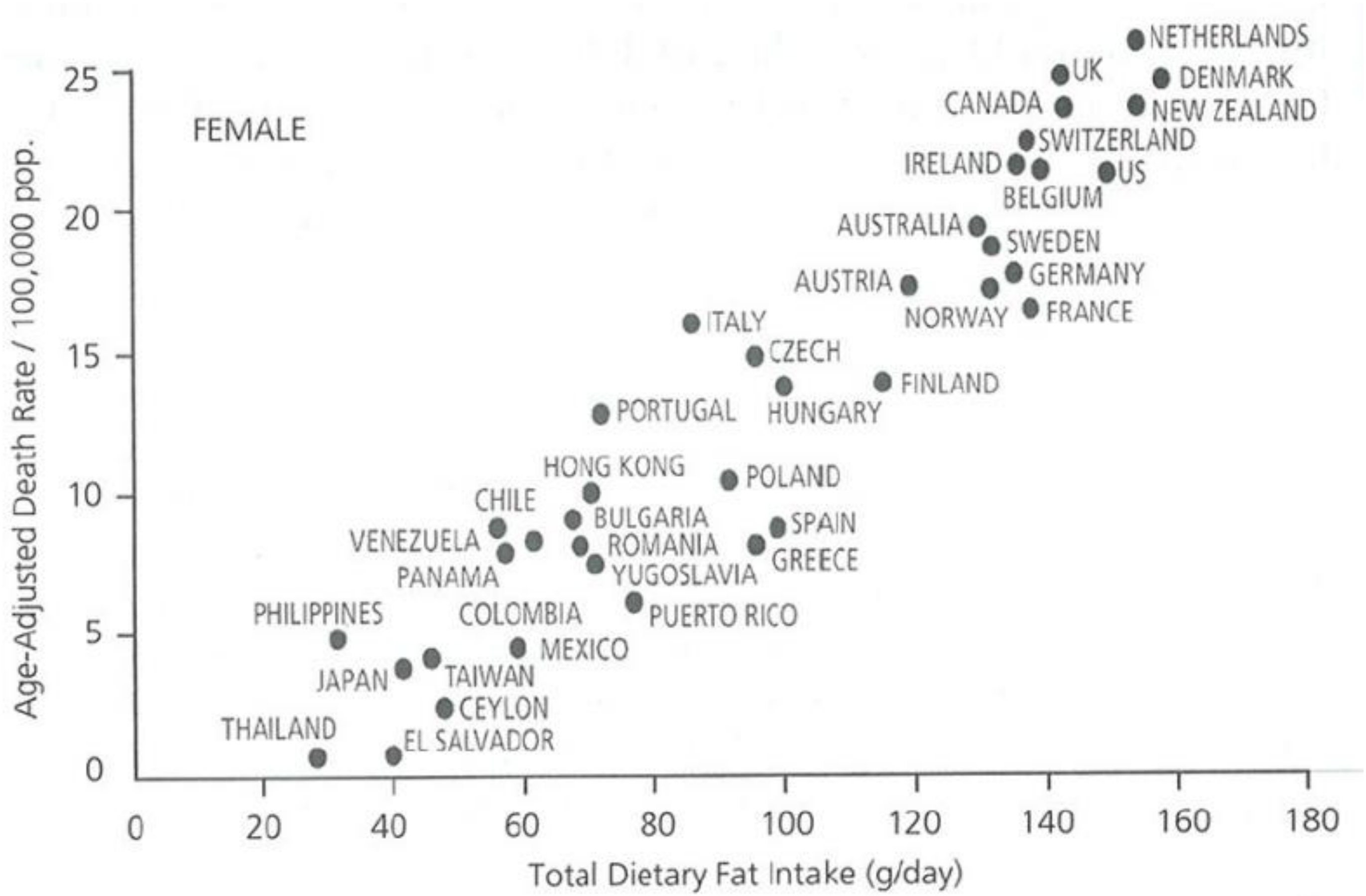
## Overlooked information

# NON-ASSOCIATIONS

# We tend to ignore non-associations

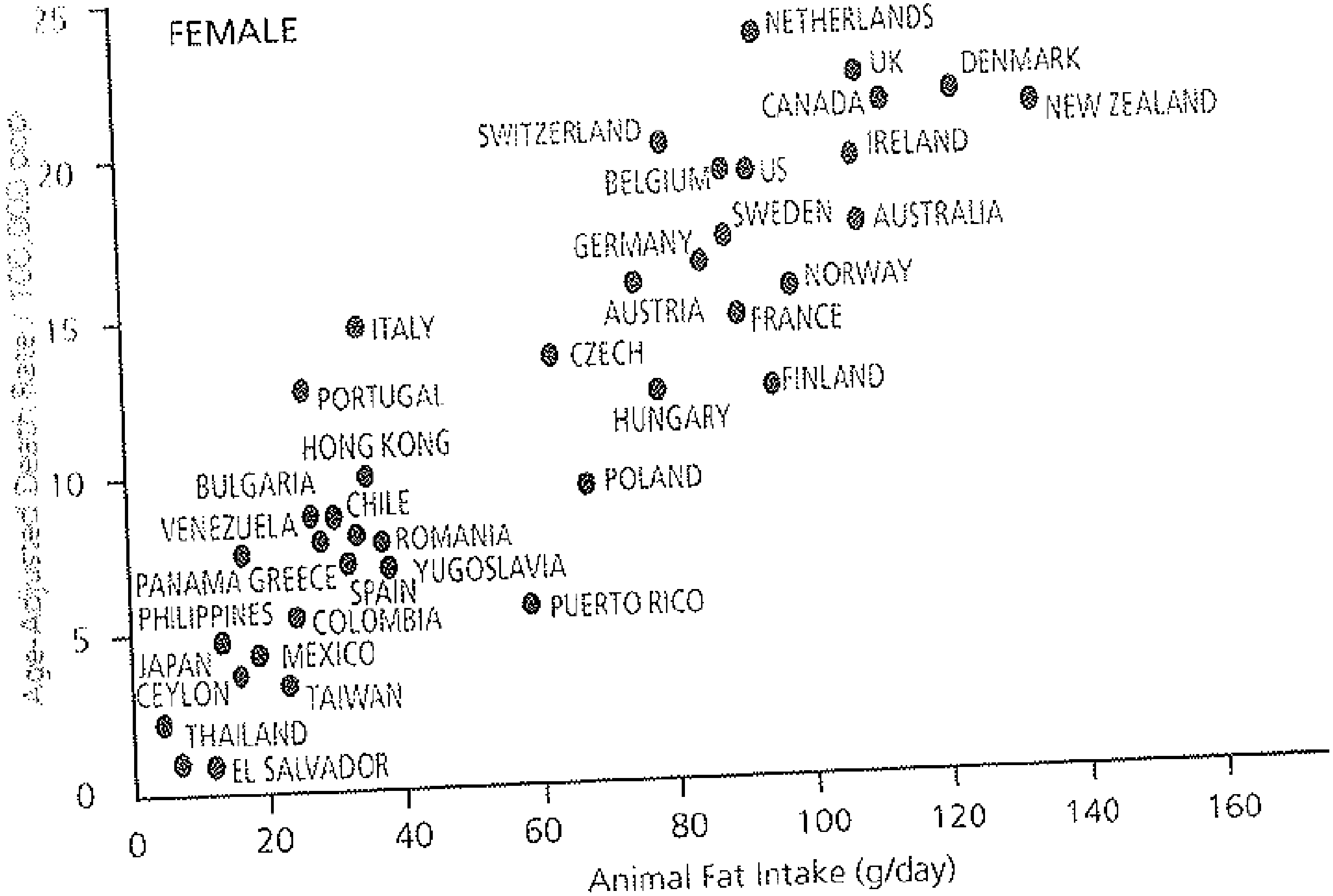
- **We have many technologies to look for associations and correlations**
  - Frequent patterns
  - Association rules
  - …
- **We tend to ignore non-associations**
  - We think they are not interesting / informative
  - There are too many of them
- **We also tend to ignore relationship between associations**

# We love to find correlations like this…



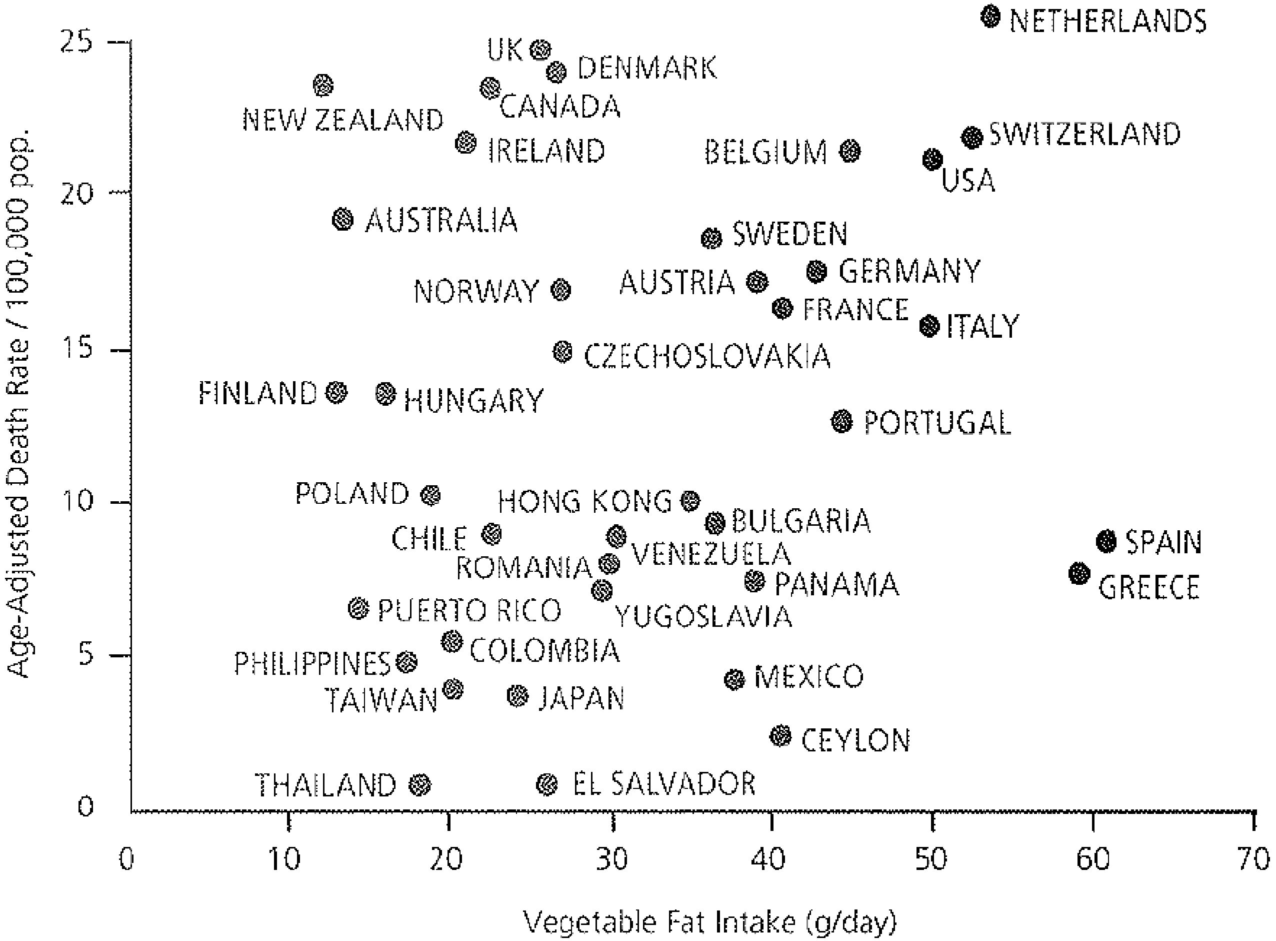
- **Dietary fat intake correlates with breast cancer**

# 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**

# Yet 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!

# Back to the Simpson's paradox

| Context | Comparing Groups | sup | $P_{class=>50K}$ | p-value |
|---|---|---|---|---|
| Race =White | Occupation = Craft-repair | 3694 | 22.84% | $1.00 \times 10^{-19}$ |
| | Occupation = Adm-clerical | 3084 | 14.23% | |

| Context | Extra attribute | Comparing Groups | sup | $P_{class=>50K}$ |
|---|---|---|---|---|
| Race =White | Sex = Male | Occupation = Craft-repair | 3524 | 23.5% |
| | | Occupation = Adm-clerical | 1038 | 24.2% |
| | Sex = Female | Occupation = Craft-repair | 107 | 8.8% |
| | | Occupation = Adm-clerical | 2046 | 9.2% |

- **2nd "I" in I.I.D. is violated**
- **Btw,"men earn more than women" also violates the 2nd "I" in I.I.D.**

# It pays to look at relationship betw associations & non-associations

**A. Wrt craftsmen / admin clerks, there are more / less men than women**

**B. Wrt men / women, craftsmen earn similar to admin clerks**

**C. Wrt craftsmen / admin clerks, men earn more than women**

- $P(m|c) > P(w|c) \Rightarrow P(m|c) > 50\%$
- $P(w|a) > P(m|a) \Rightarrow P(m|a) < 50\%$

i.e. $P(m|c) > P(m|a)$

- $P(\$|m,c) \approx P(\$|m,a)$
- $P(\$|w,c) \approx P(\$|w,a)$

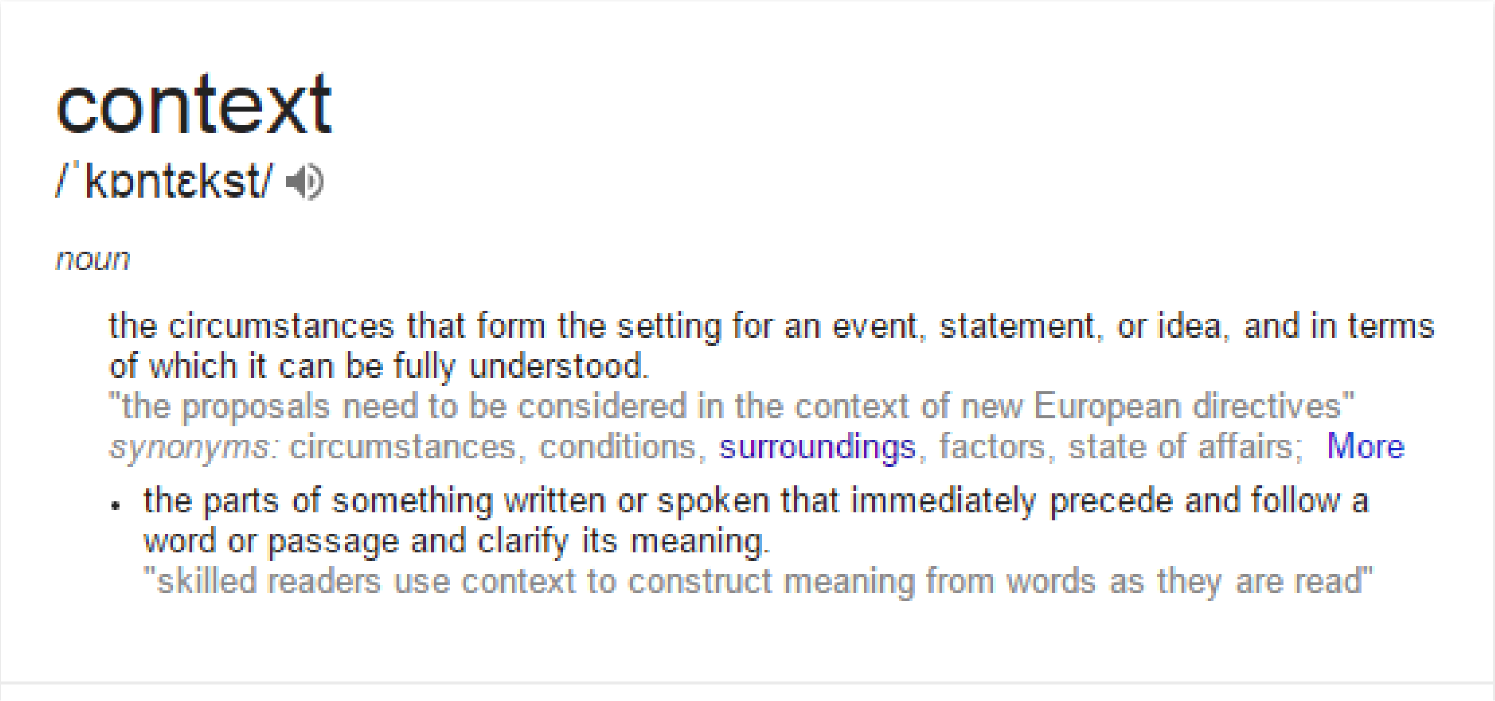
- $P(\$|m,c) > P(\$|w,c)$
- $P(\$|m,a) > P(\$|w,a)$

$$
\begin{aligned}
& P(\$|c) \\
&= P(\$,m|c) + P(\$,w|c) \\
&= P(\$|m,c)\,P(m|c) + P(\$|w,c)\,P(w|c) \\
&= [P(\$|m,c) - P(\$|w,c)]\,P(m|c) + P(\$|w,c) \\
&> [P(\$|m,a) - P(\$|w,a)]\,P(m|a) + P(\$|w,a) \\
&= P(\$|m,a)\,P(m|a) + P(\$|w,a)\,P(w|a) \\
&= P(\$,m|a) + P(\$,w|a) \\
&= P(\$|a)
\end{aligned}
$$

i.e., $P(\$|c) \geq P(\$|a)$

**“Craftsmen earn more than admin clerks”** is an artefact

i.e., even if “craftsmen earn more than admin clerks” passes a valid statistical test, it is a derivative of A, B, C



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

# CONTEXT

# 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

# 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⟩**

# The right support threshold

- ⟨{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 | $N^B_{pos}$ | $N^B - N^B_{pos}$ |

- **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})

⇒ **Frequent pattern mining, but be careful with support threshold, need to relativize to context**

# Relativizing to context

- **Most people cannot set support threshold correctly when relativizing to context**

## A quick test!

- **Suppose a test of a disease presents a rate of 5% false positives, and the disease strikes 1/1000 of the population**

- **Let’s say people are tested randomly and a particular patient’s test is positive**

- **What’s the probability that he is stricken with the disease?**

# Answer

- P(d) = 0.1%
- P(pos| ~d) = 5%
- P(pos| d) = 100%, assuming 100% sensitivity

- P(pos) = P(pos| d) P(d) + P(pos| ~d) P(~d) ≈ 5%

- P(d| pos) = P(pos| d) P(d) / P(pos) = 0.1% / 5% = 2%

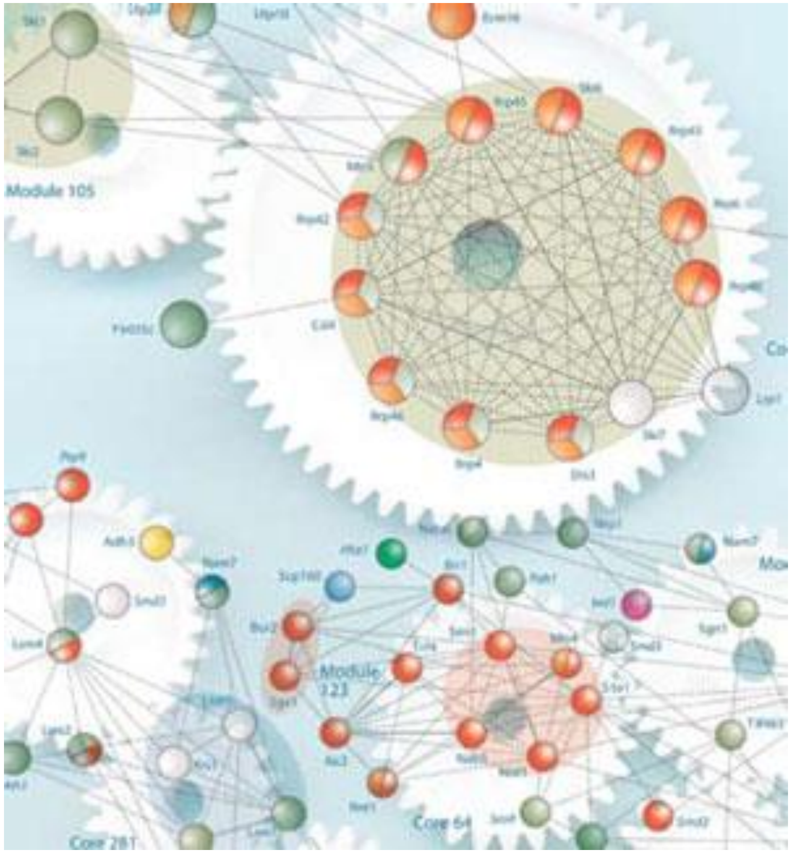
- I.e., the answer is 2%
- Did you guess 95% as the answer?

# 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 | $N^B{}_{pos}$ | $N^B - N^B{}_{pos}$ |

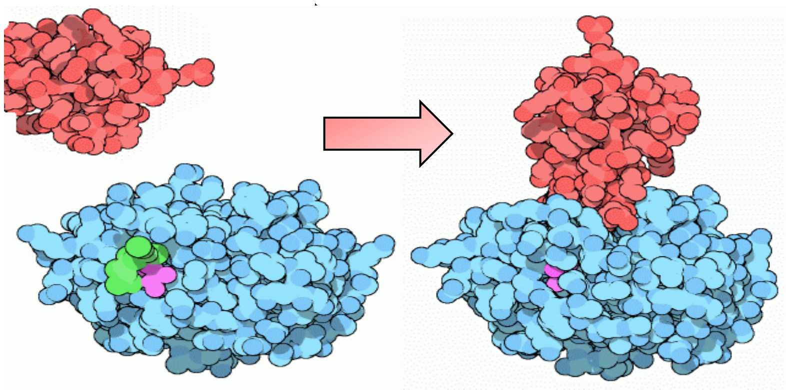
- If A/B treat the same single disease, this is ok
- If B treats two diseases, this is not sensible
- The disease has to go into the context



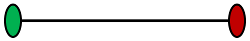
More may not be better

# PROTEIN COMPLEXES

# Protein-protein interaction networks



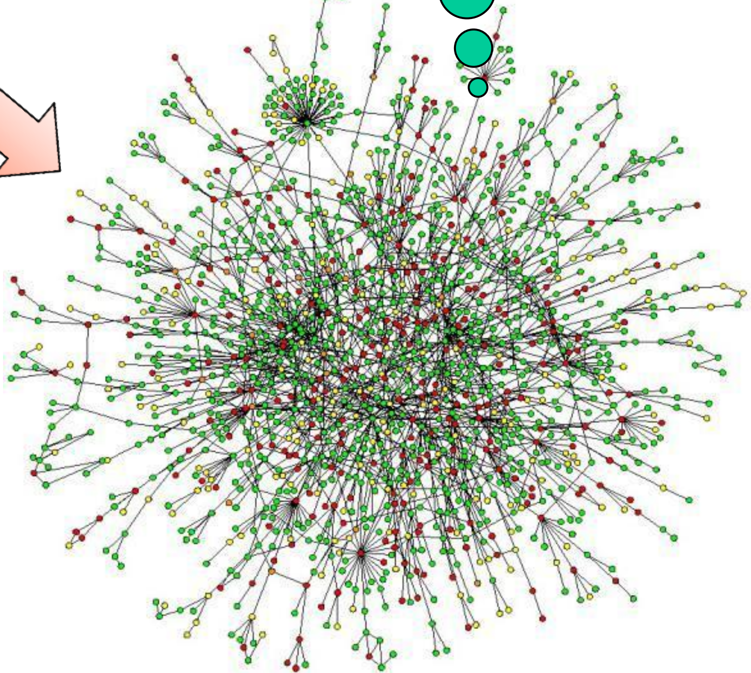
Individual proteins come together and interact



Collection of such interactions in an organism

**PPIN**

Valuable source of knowledge



Protein Interaction Network

- **Proteins come together & interact**
- **The collection of these interactions form a Protein Interaction Network or PPIN**

# Difficulties

- **Cytochrome BC1 complex**
  - Involved in electron-transport chain in mitochondrial inner membrane

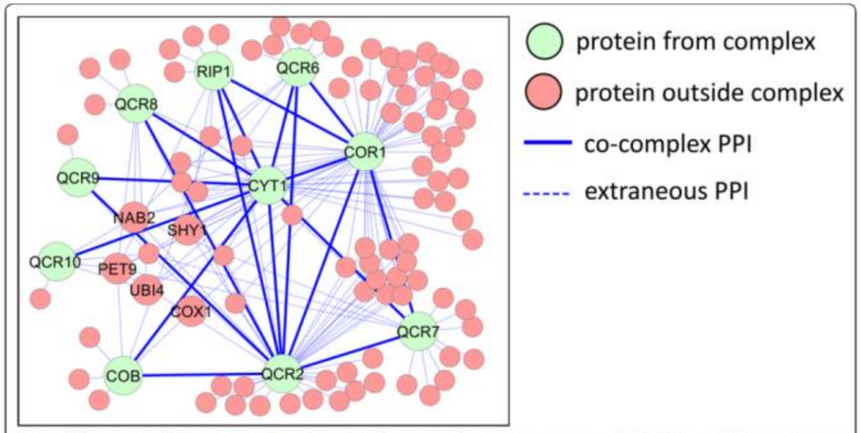


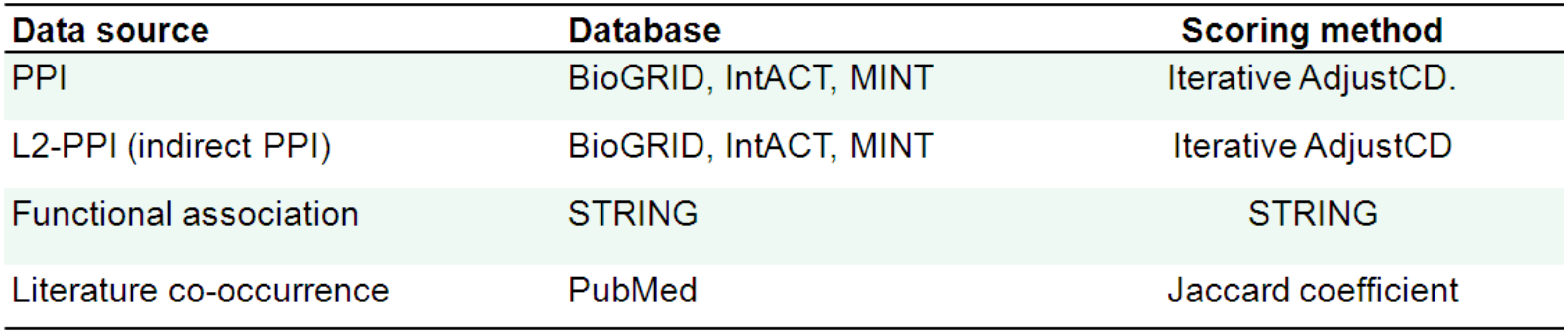
**Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex**. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [30].

- **Discovery of BC1 from PPI data is difficult**
  - Sparseness of its PPI subnetwork
    - **Only 19 out of 45 possible interactions were detected between the complex's proteins**
  - Extraneous interactions with other proteins outside the complex
    - **E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function**

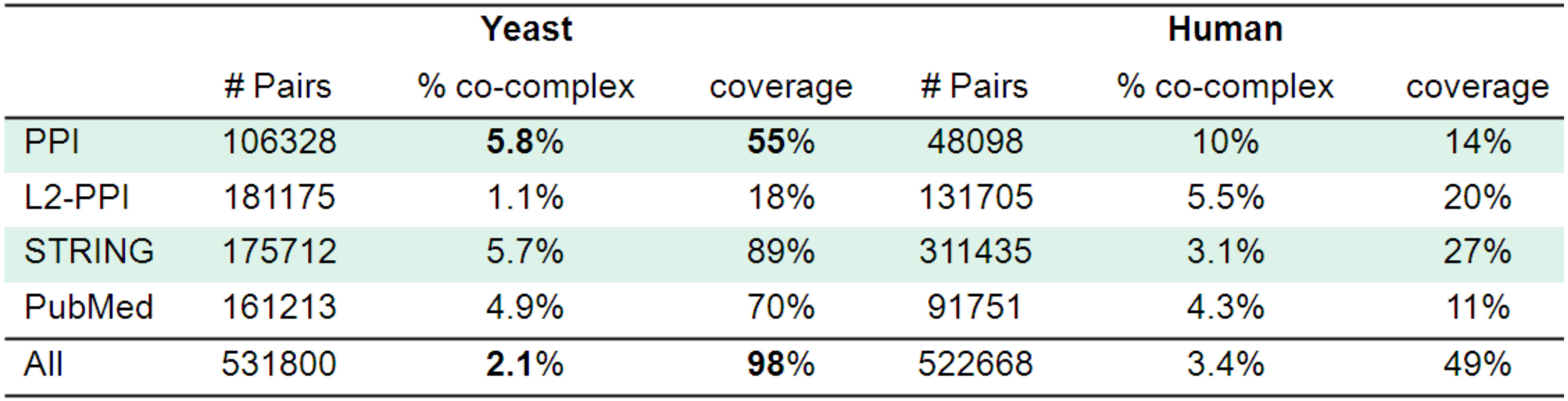
Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

# Perhaps "big data" can help?

- **Composite network**
  - Vertices represent proteins, edges represent relationships between proteins. Put an edge betw proteins u, v, iff u and v are related according to any of the data sources



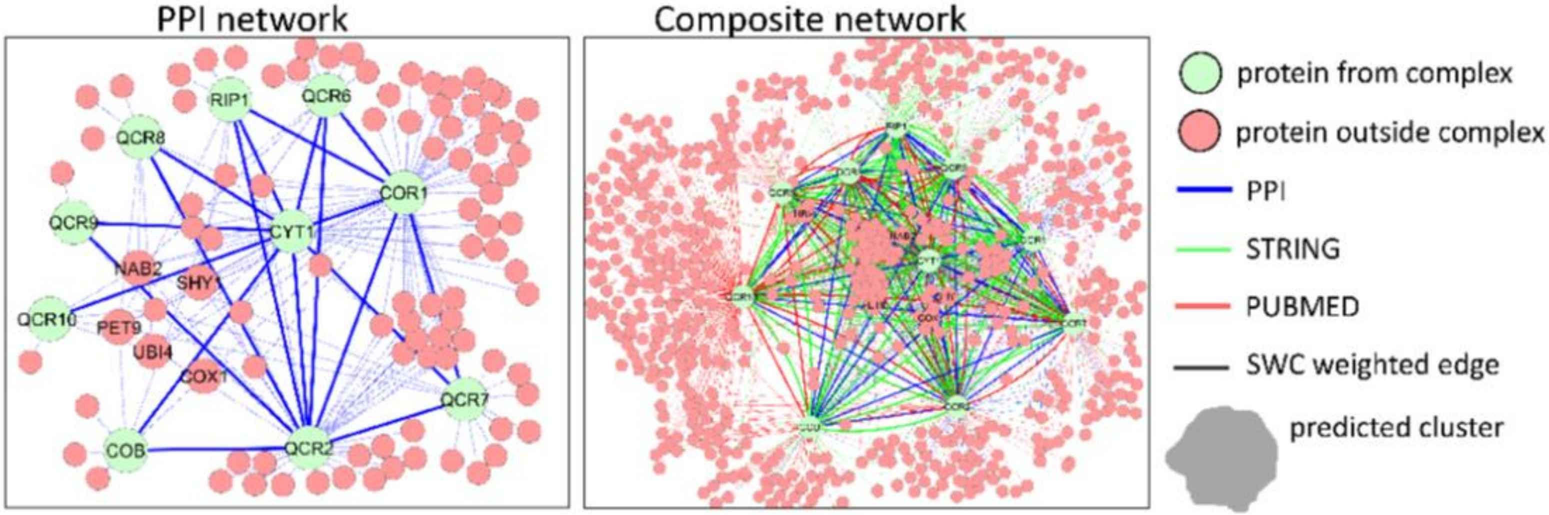
| Data source | Database | Scoring method |
|---|---|---|
| PPI | BioGRID, IntACT, MINT | Iterative AdjustCD. |
| L2-PPI (indirect PPI) | BioGRID, IntACT, MINT | Iterative AdjustCD |
| Functional association | STRING | STRING |
| Literature co-occurrence | PubMed | Jaccard coefficient |

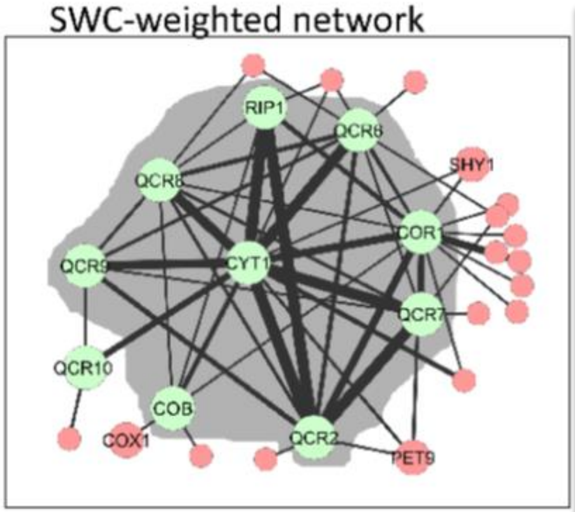


| | Yeast | | | Human | | |
|---|---|---|---|---|---|---|
| | # Pairs | % co-complex | coverage | # Pairs | % co-complex | coverage |
| PPI | 106328 | **5.8**% | **55**% | 48098 | 10% | 14% |
| L2-PPI | 181175 | 1.1% | 18% | 131705 | 5.5% | 20% |
| STRING | 175712 | 5.7% | 89% | 311435 | 3.1% | 27% |
| PubMed | 161213 | 4.9% | 70% | 91751 | 4.3% | 11% |
| All | 531800 | **2.1**% | **98**% | 522668 | 3.4% | 49% |

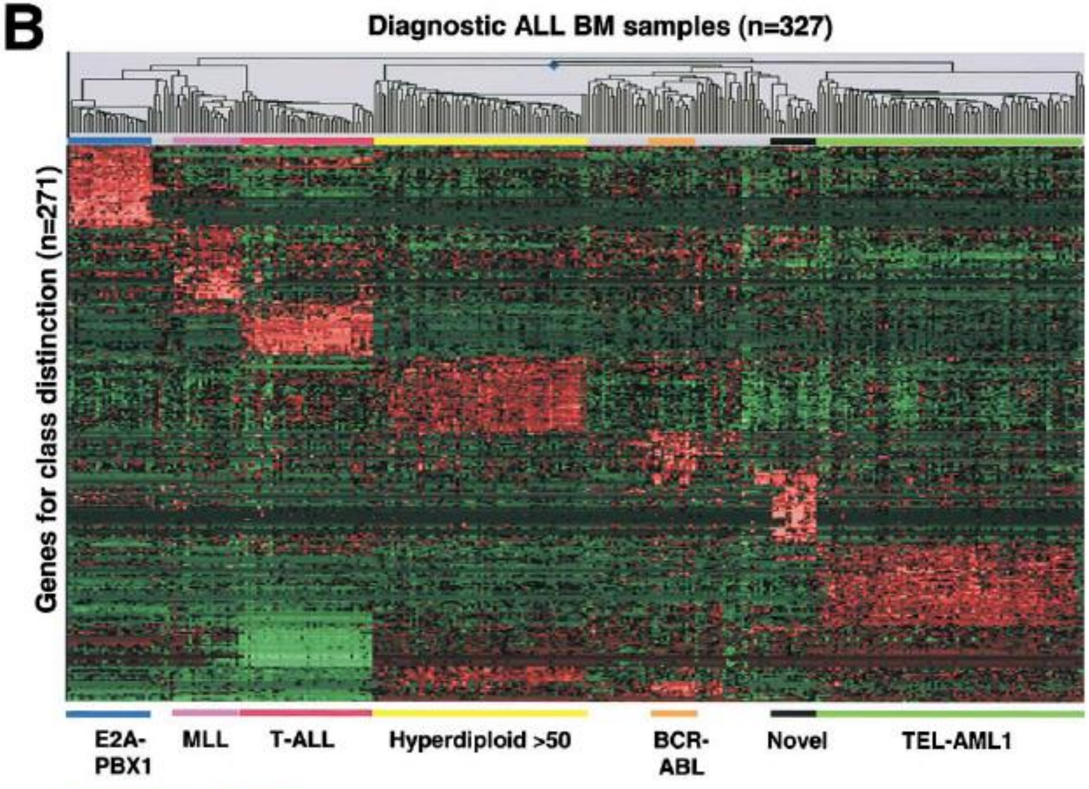
Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

# More is not always better, unless..





While proteins in BC1 become fully connected in the composite network, there is a blow-up in extraneous proteins. So clustering won’t discover the complex, unless you know how to remove the extraneous proteins



More may not be better

# CAUSAL GENES

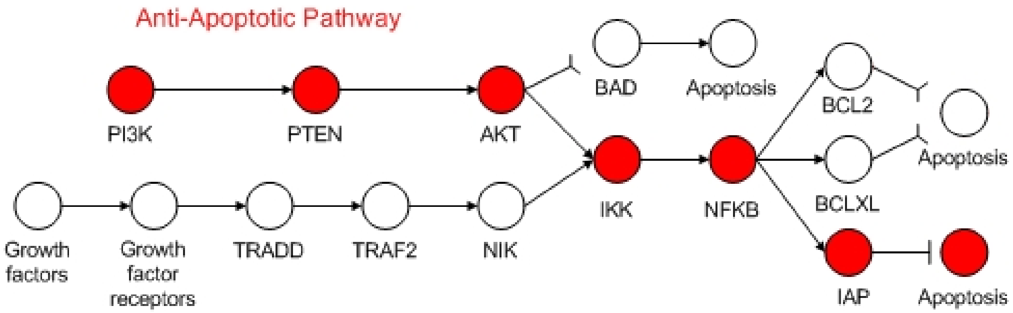
# Gene expression analysis challenge

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

| Datasets | DEG | POG |
|---|---|---|
| Prostate Cancer | | |
| | Top 10 | 0.30 |
| | Top 50 | 0.14 |
| | Top100 | 0.15 |
| Lung Cancer | | |
| | Top 10 | 0.00 |
| | Top 50 | 0.20 |
| | Top100 | 0.31 |
| DMD | | |
| | Top 10 | 0.20 |
| | Top 50 | 0.42 |
| | Top100 | 0.54 |

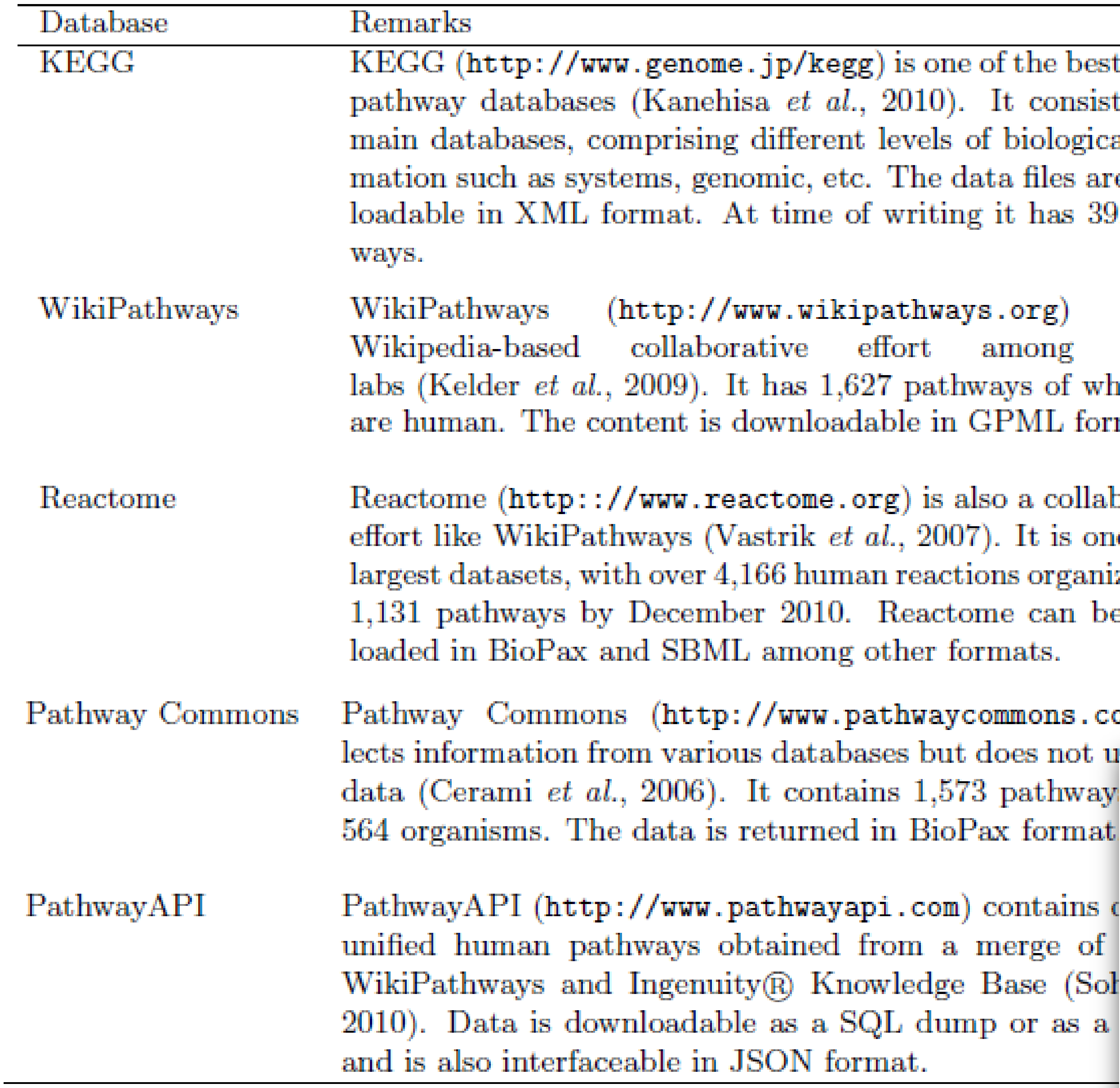
Zhang et al, *Bioinformatics*, 2009

# Biology to the rescue?



- **Each disease phenotype has some underlying cause**
- **There is some unifying biological theme for genes that are truly associated with a disease subtype**

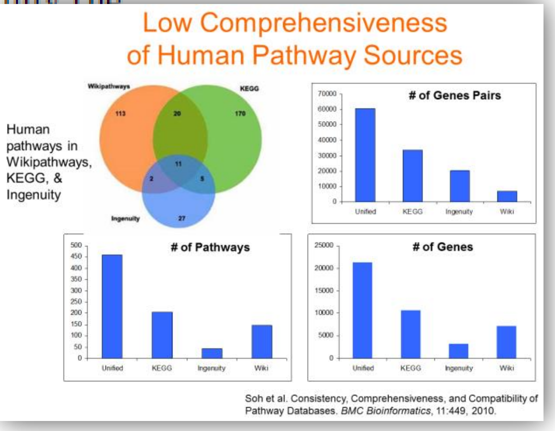
- **Uncertainty in selected genes can be reduced by considering biological processes of the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**



| Database | Remarks |
|---|---|
| KEGG | KEGG (http://www.genome.jp/kegg) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways. |
| WikiPathways | WikiPathways (http://www.wikipathways.org) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format. |
| Reactome | Reactome (http:://www.reactome.org) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats. |
| Pathway Commons | Pathway Commons (http://www.pathwaycommons.com) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathway 564 organisms. The data is returned in BioPax format |
| PathwayAPI | PathwayAPI (http://www.pathwayapi.com) contains unified human pathways obtained from a merge of WikiPathways and Ingenuity® Knowledge Base (Soh 2010). Data is downloadable as a SQL dump or as a and is also interfaceable in JSON format. |

Goh, et al. *Proteomics*, 12(4-5):550-563, 2012.

## Big data of biological pathways

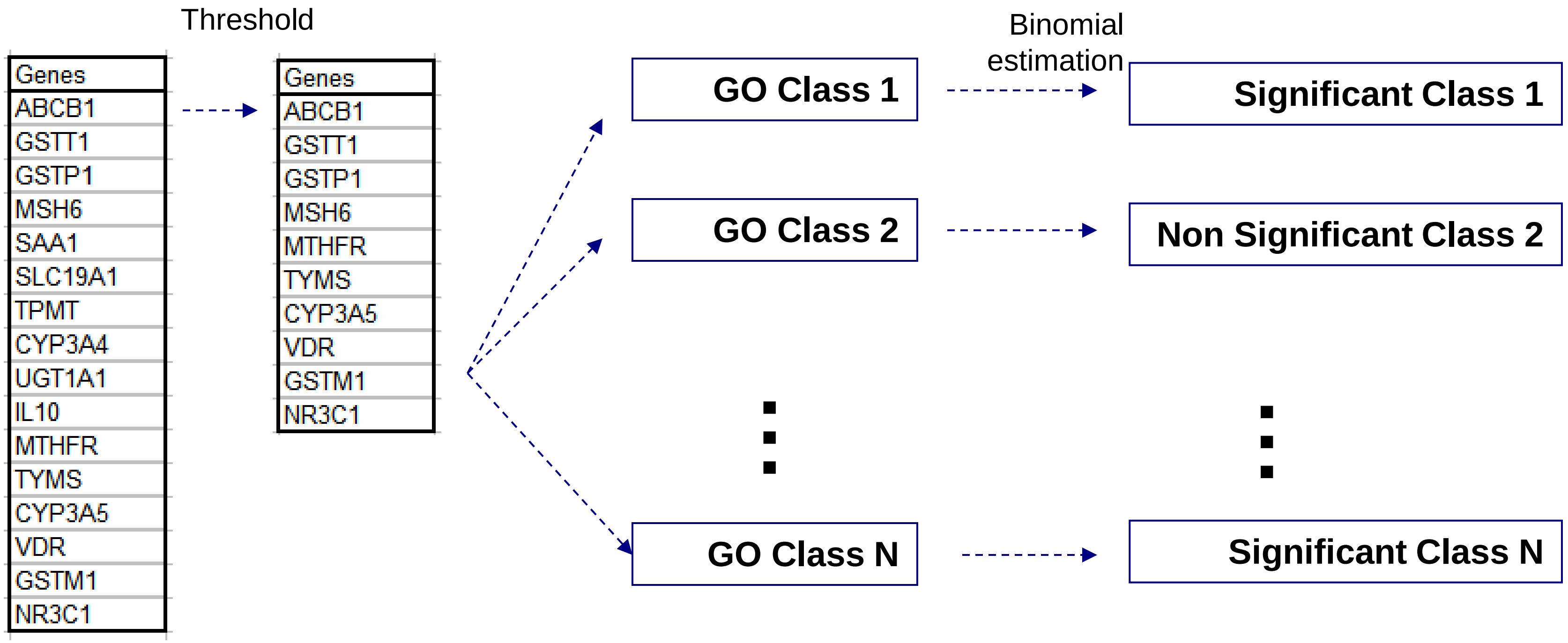


Low Comprehensiveness of Human Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity

Soh et al. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics*, 11:449, 2010.

# Overlap Analysis: ORA



Threshold

| Genes |
|---|
| ABCB1 |
| GSTT1 |
| GSTP1 |
| MSH6 |
| SAA1 |
| SLC19A1 |
| TPMT |
| CYP3A4 |
| UGT1A1 |
| IL10 |
| MTHFR |
| TYMS |
| CYP3A5 |
| VDR |
| GSTM1 |
| NR3C1 |

| Genes |
|---|
| ABCB1 |
| GSTT1 |
| GSTP1 |
| MSH6 |
| MTHFR |
| TYMS |
| CYP3A5 |
| VDR |
| GSTM1 |
| NR3C1 |

Binomial estimation

GO Class 1 → Significant Class 1

GO Class 2 → Non Significant Class 2

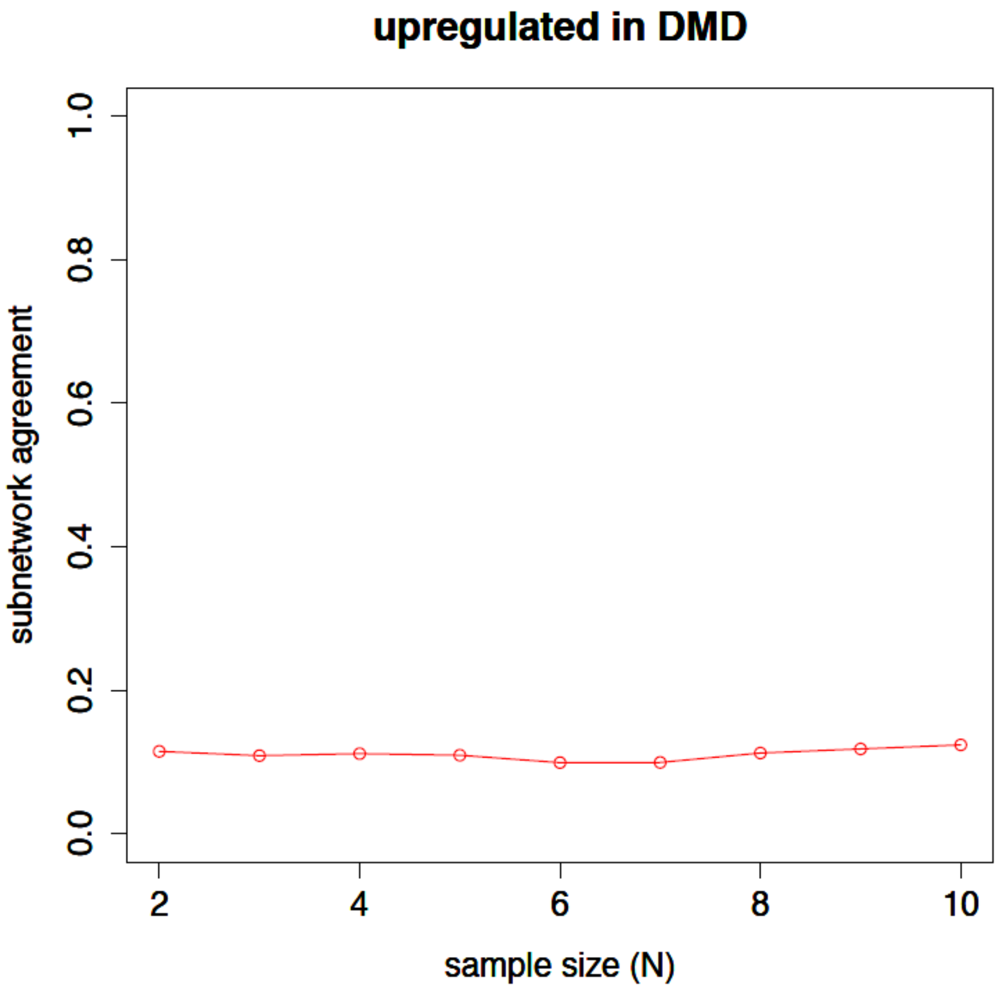
⋮

GO Class N → Significant Class N

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

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

# Disappointing Performance



DMD gene expression data

- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data

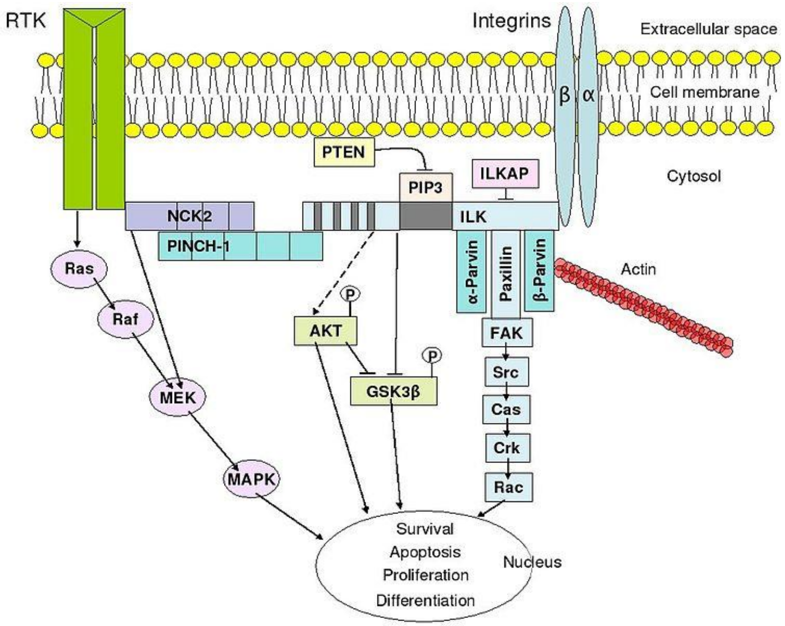
- PathwayAPI, Soh et al., 2010

# Issue #1 with ORA

- **Its 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**

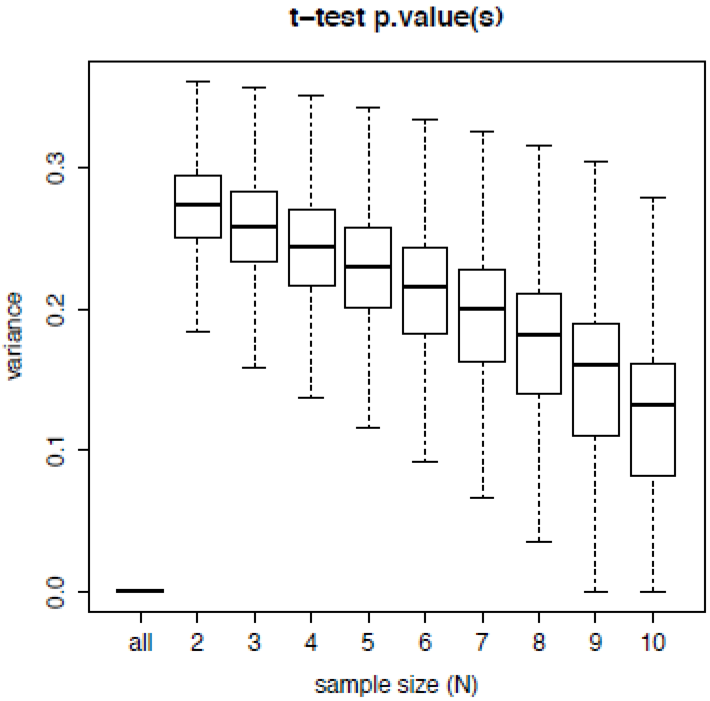
⇒ **Lots of false positives**



- 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

# Issue #2 with ORA

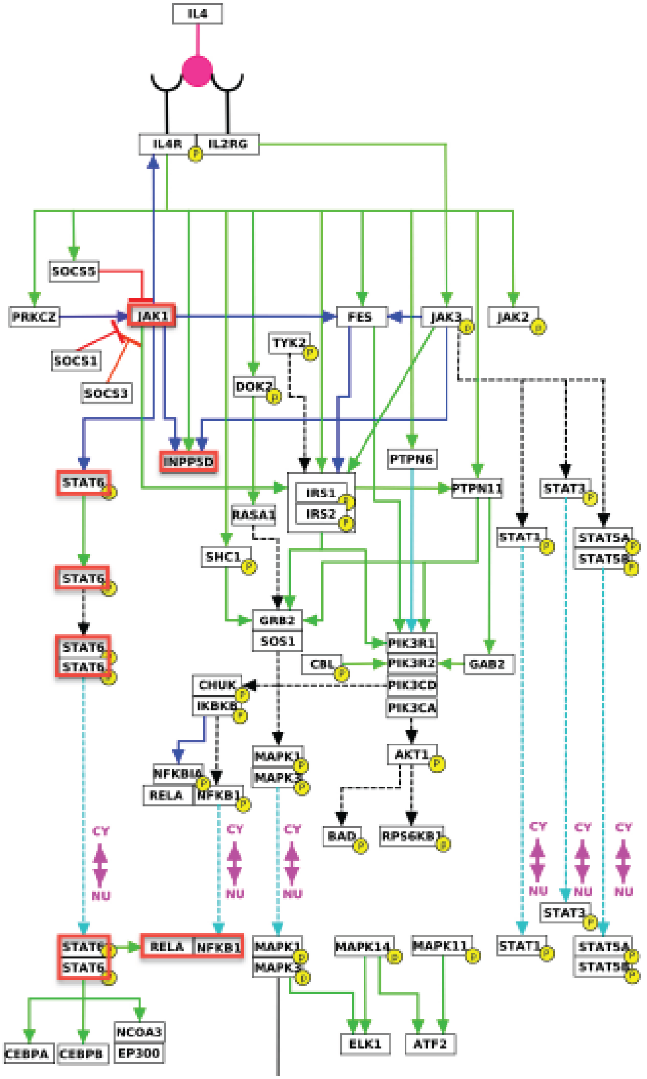
- **It relies on a pre-determined list of DE genes**
- **This list is sensitive to the test statistic used and to the significance threshold used**
- **This list is unstable regardless of the threshold used when sample size is small**



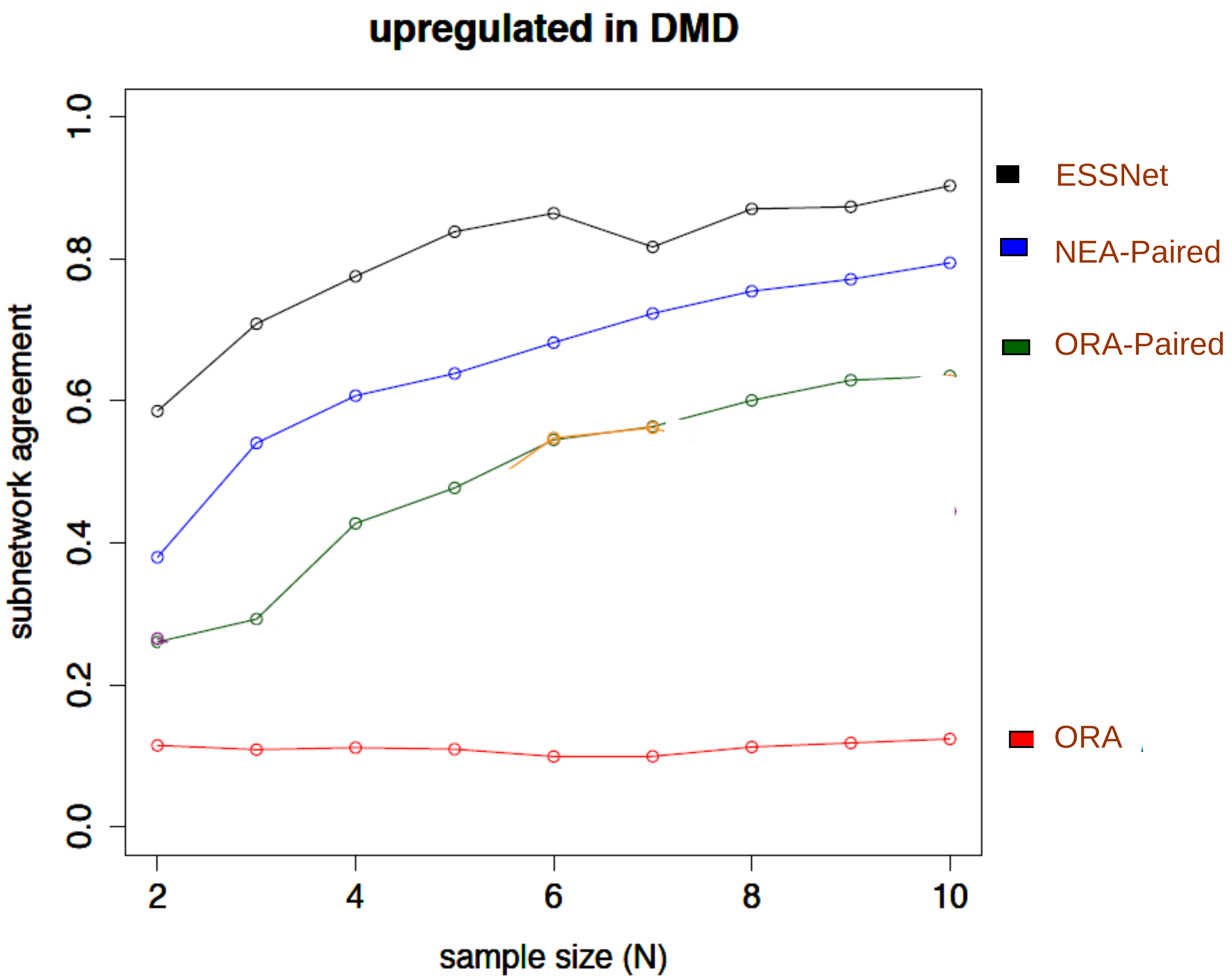
# Issue #3 with ORA

- **It tests whether the entire pathway is significantly differentially expressed**

- **If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch**



# As we address the three issues, performance improves



# What have we learned?

- **More data can offer a more complete picture, fill in gaps, etc.**
- **More data can also introduce noise into an analysis**
- **Unless you know how to tame this noise, more data may not lead to a better analysis**
- **Mechanical application of statistical and data mining techniques often does not work**
- **Must understand statistical and data mining tools & the problem domain**
  - Must know how to logically exploit both