Some issues that are often overlooked in big data analytics

Wong Limsoon

What is big data and why

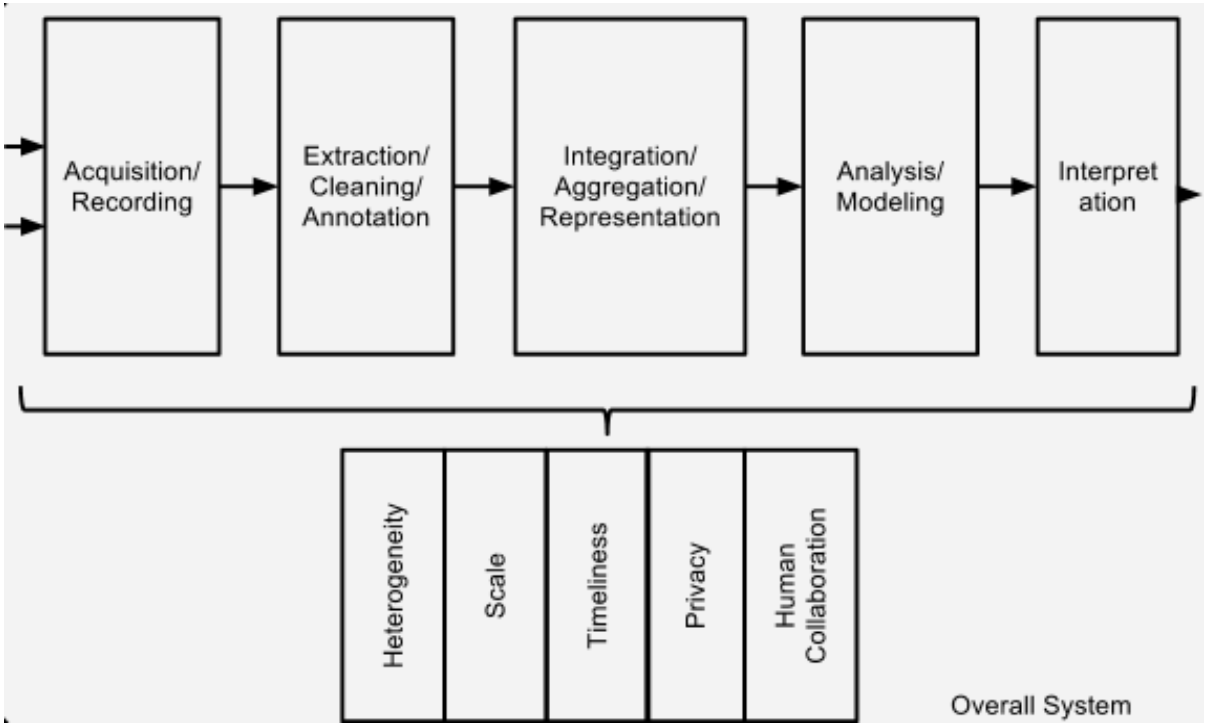
- **Big data *a la* Gartner**
  - Volume, velocity, variety
- **Other characteristics**
  - Veracity, v...

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 analysis methods & systems**

Talk outline

- **Forgotten assumptions**
  - Normal distribution
  - I.I.D.
  - Proper design of experiment
  - Domain-specific laws
- **Overlooked information**
  - Non-associations
  - Context
- **More may not be better**
  - Protein complexes

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

		Wisdom-of-crowd aggregation		
Question	True value	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.

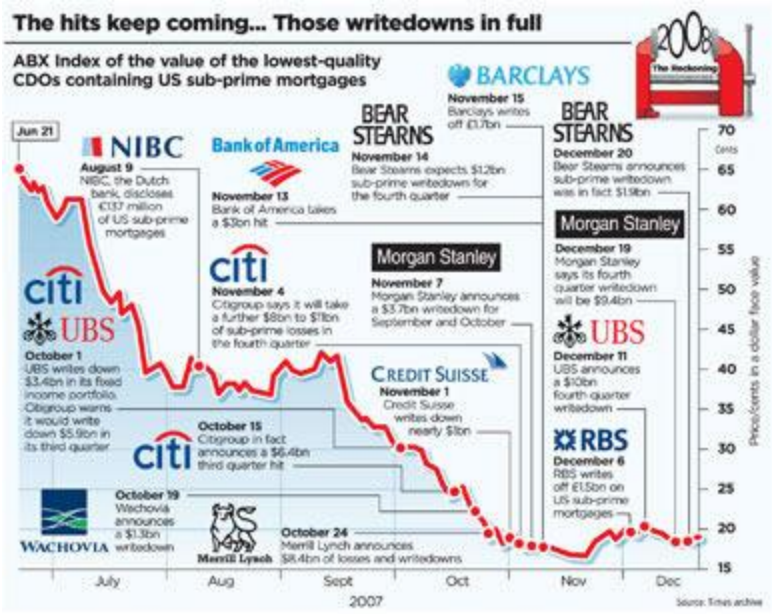
- **Estimates not normally distributed**
- **They are lognormally distributed**

⇒ **Subjects had problems choosing the right order of magnitude**

NUS National University of Singapore

and what held yesterday may not hold today

2007 Financial Crisis

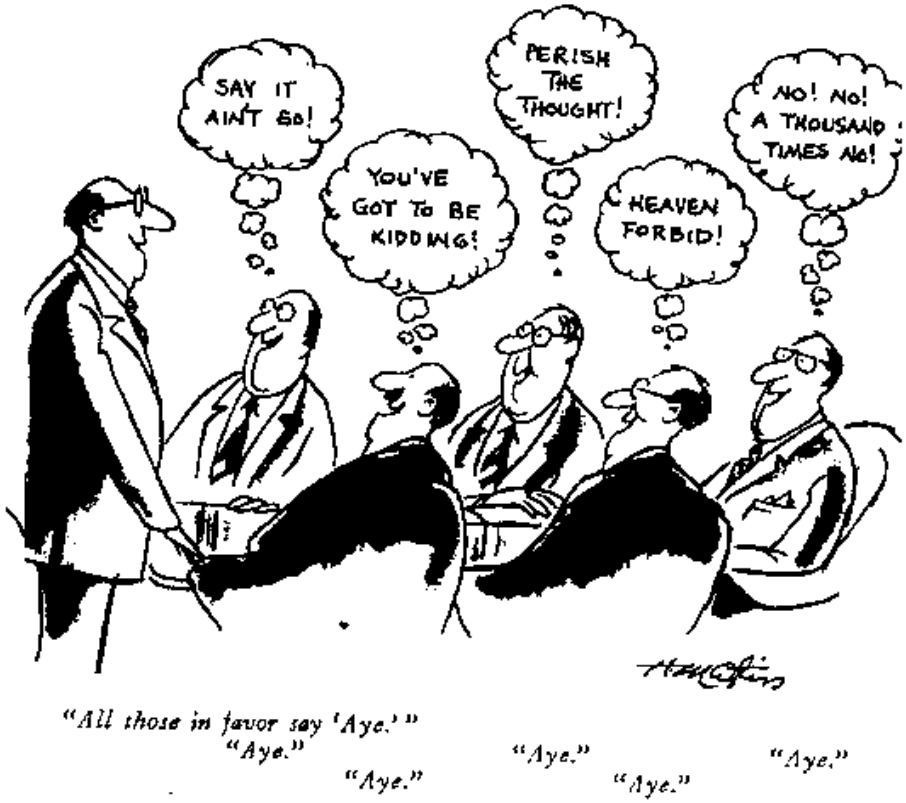


- **All of them religiously check VaR (Value at Risk) everyday**

- **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



Forgotten assumptions

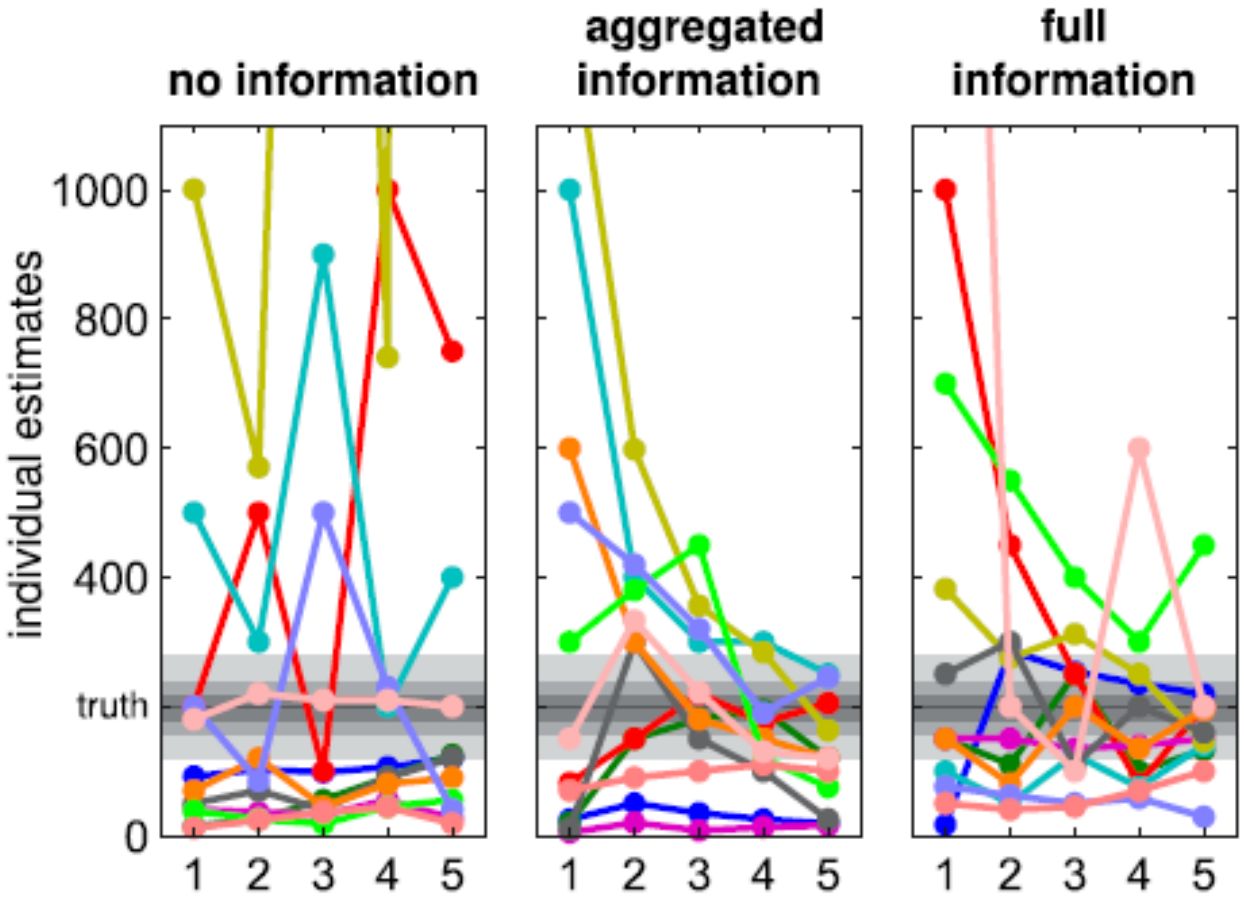
I.I.D.

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**

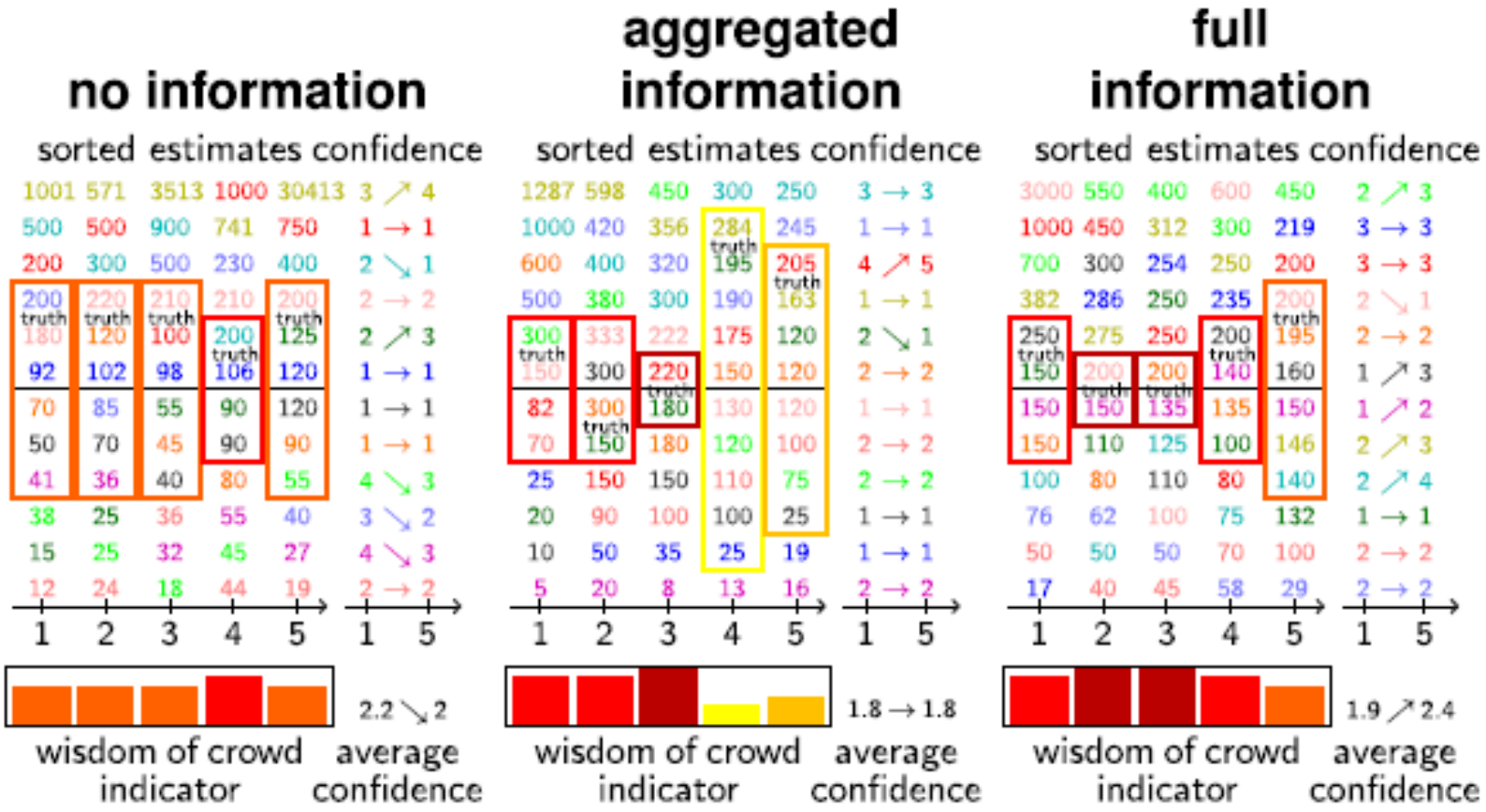
Social influence effect



- **Social influence diminishes diversity in groups**

⇒ **Groups potentially get into "group think"!**

Range reduction effect



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

Social influence diminishes wisdom of the crowd

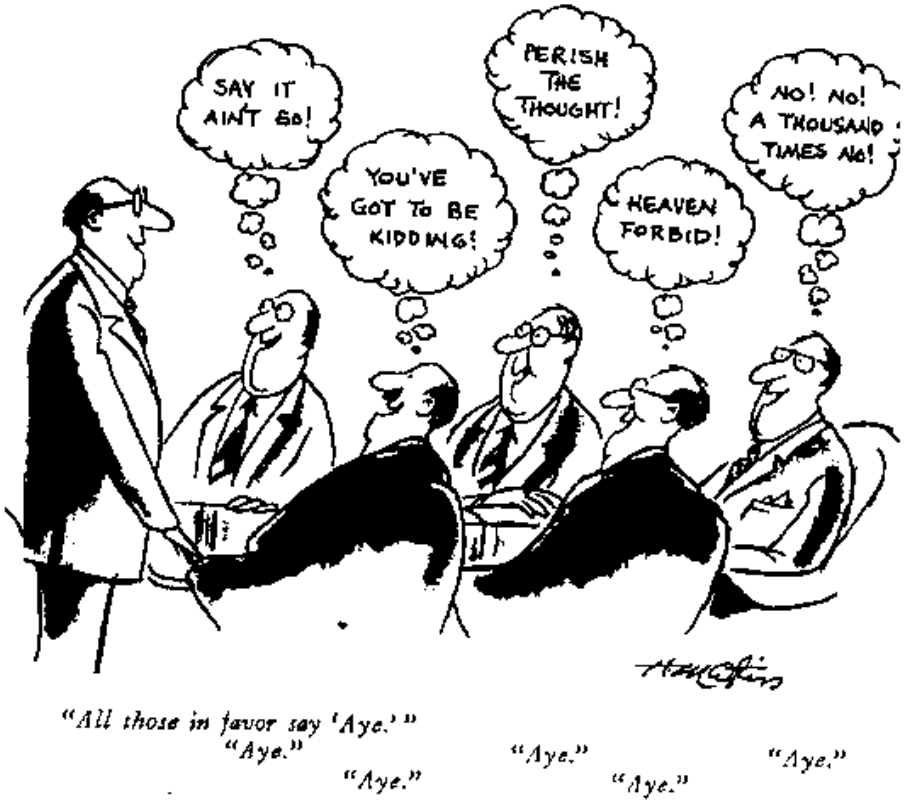
- **Social influence triggers convergence of individual estimates**

- **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

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

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%

- **Craft-repair/Adm-clerical sample not identical in other aspects**

Stratification

- **Cannot test "Men earn more than women" directly**
  - 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.

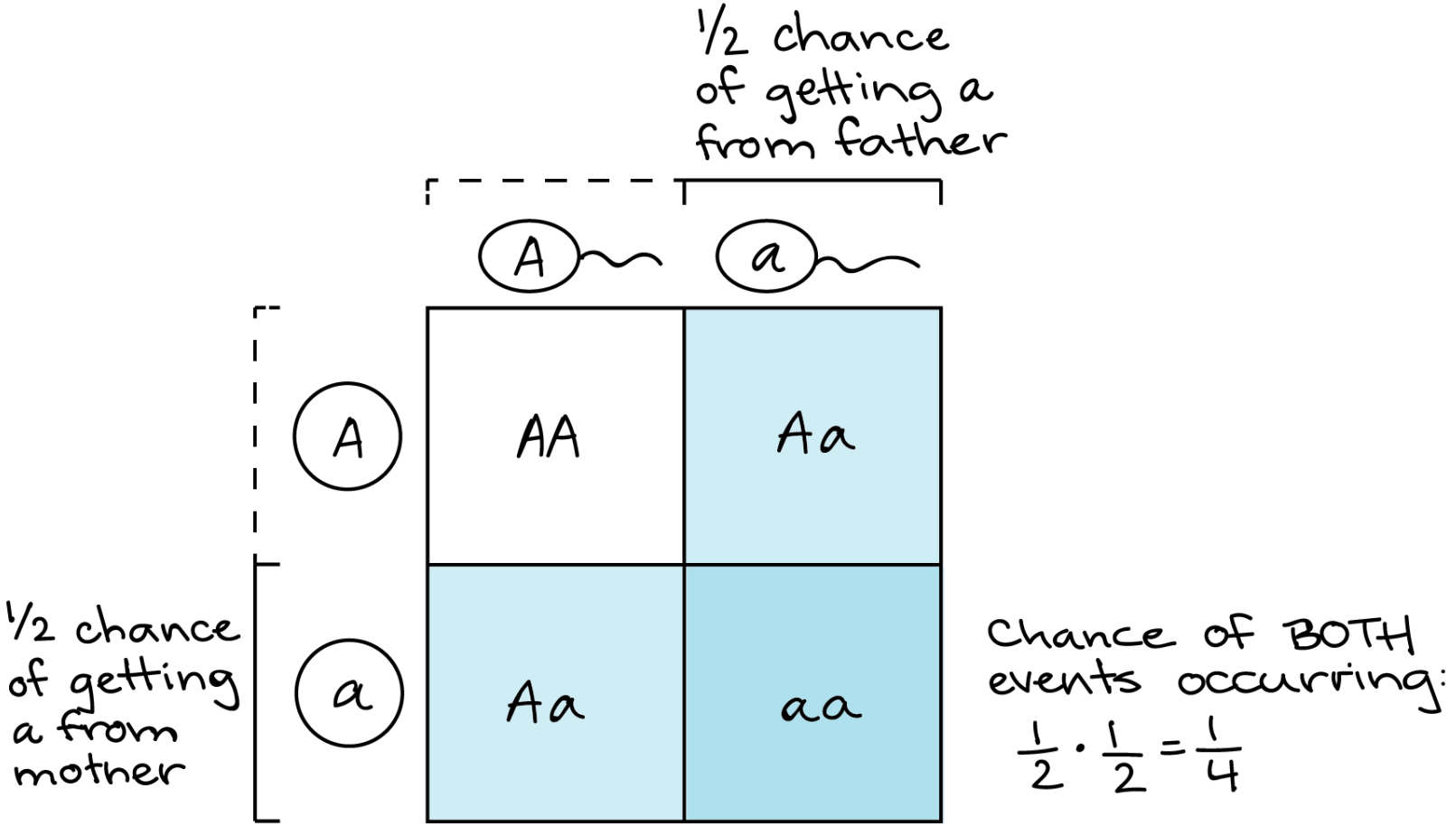
President-elect Truman holding the infamous issue of the *Chicago Tribune*, telling the press, "That ain't the way I heard it!"

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.

Forgotten assumptions

DOMAIN-SPECIFIC LAWS

A basic rule of human genetics



A suspicious contingency table

SNP	Genotypes	Group: Controls [n(%)]		Group: Cases [n(%)]		χ^2	*P* value
rs??????	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; so ~62% of population is AG**

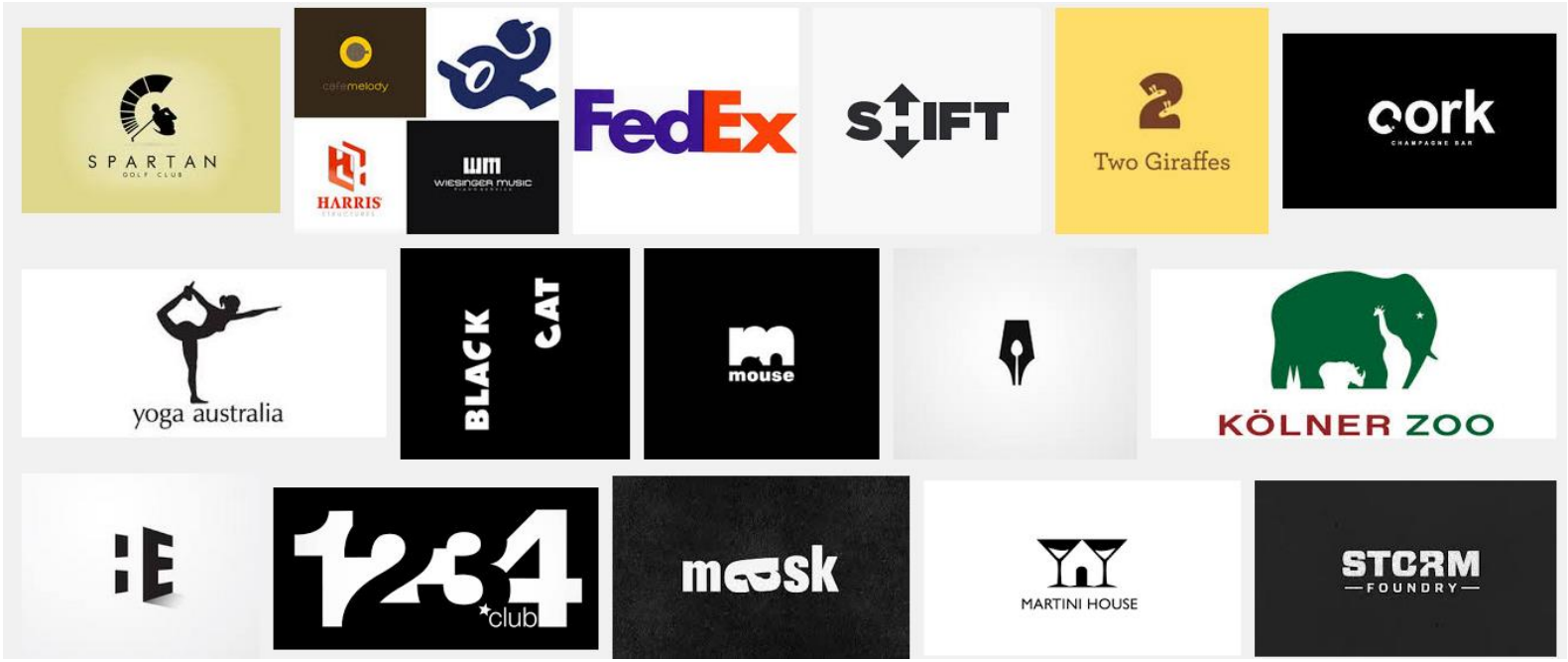
⇒ **~10% of population is AA, unless AA is fatal**

- **Big data check shows AA is non-fatal for this SNP**

⇒ **Sample is biased**

Lessons learned

- **Need to check for domain-specific rules to ensure test validity**
- **Big data can be helpful**



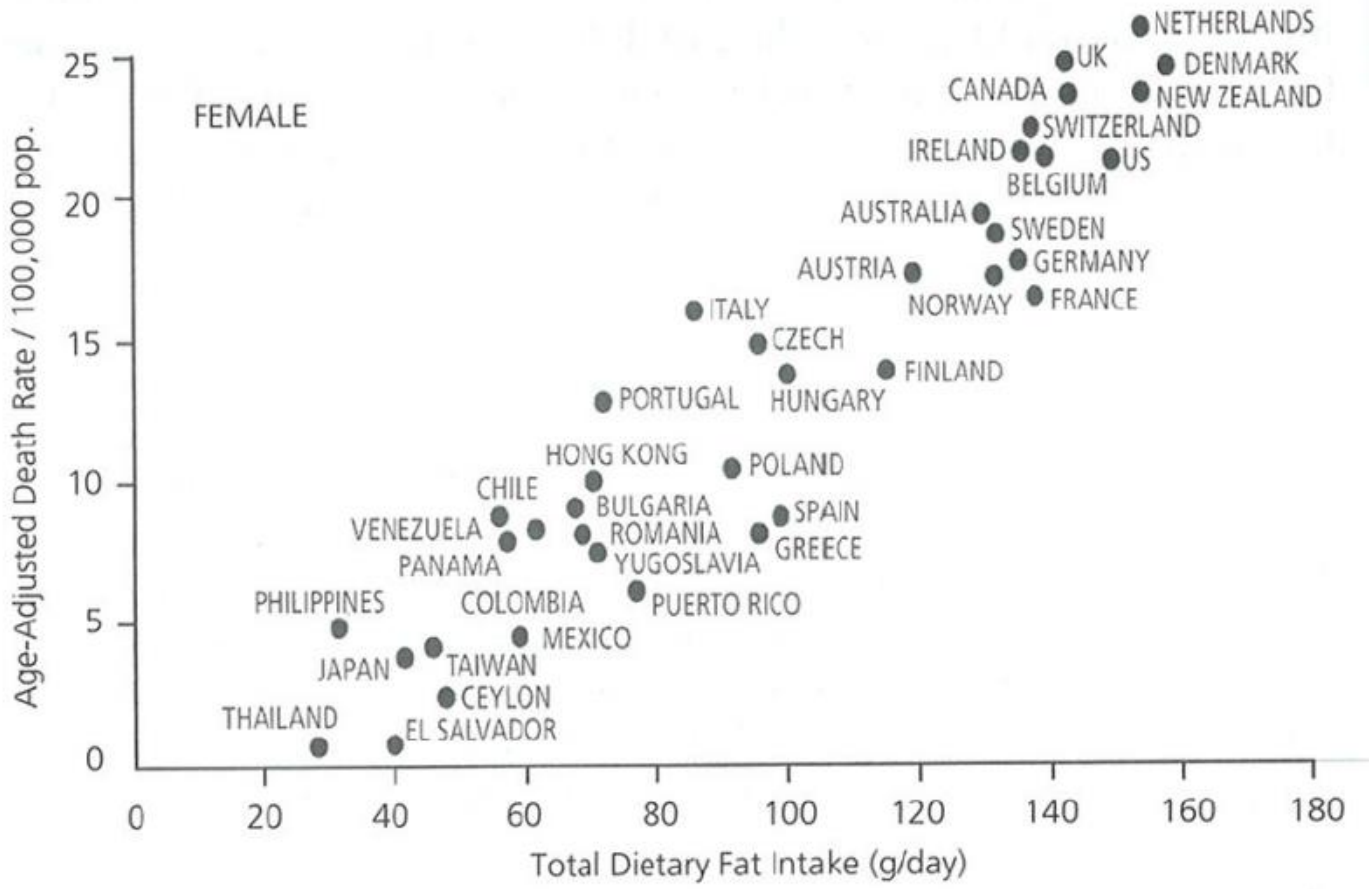
Overlooked information

NON-ASSOCIATIONS

We tend to ignore non-associations

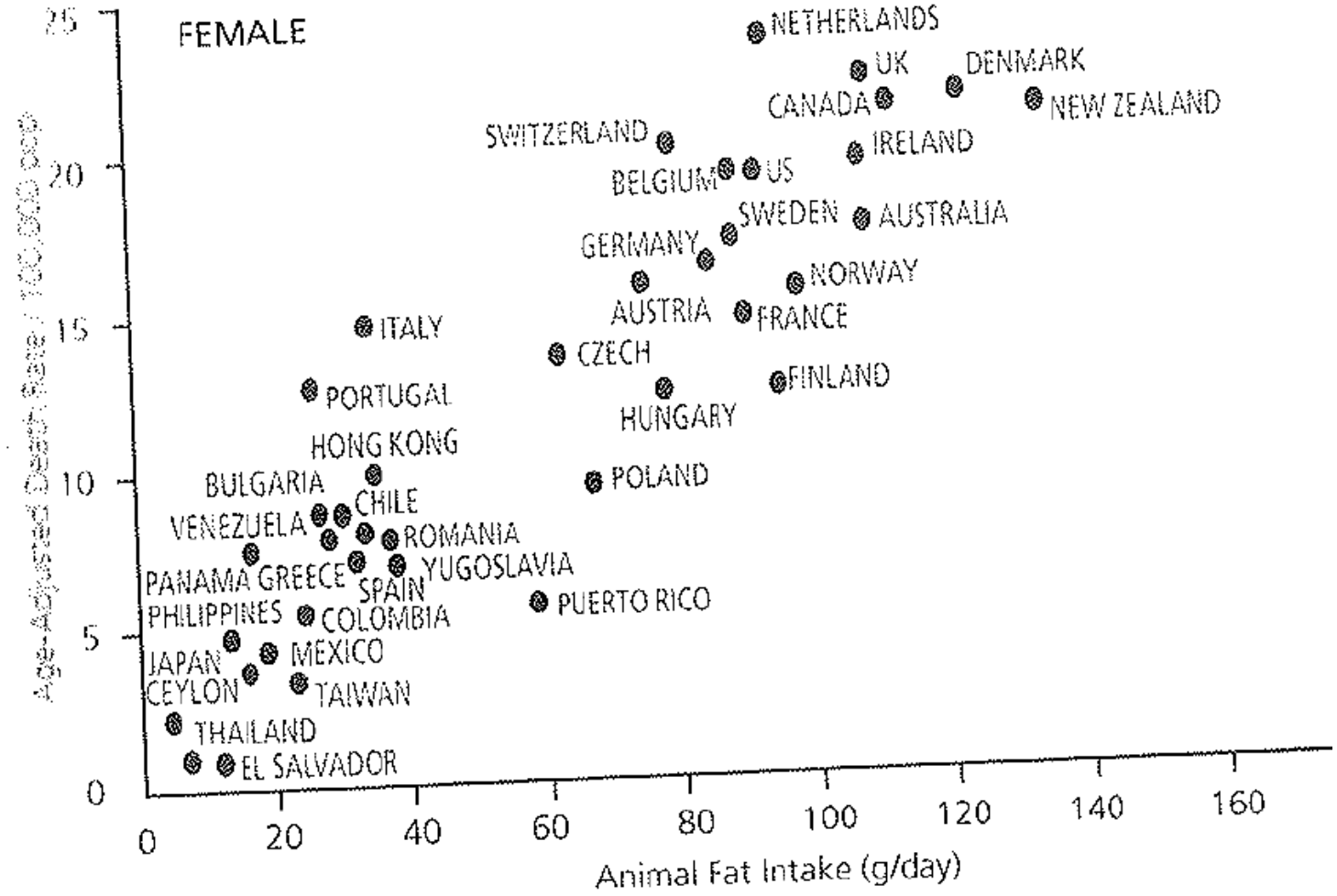
- **We have many technologies to look for associations and correlations**
  - Frequent patterns
  - Association rules
  - …
- **We tend to ignore non-associations**
  - We think they are not interesting / informative
  - There are too many of them
- **We also tend to ignore relationship between associations**

We love to find correlations like this...



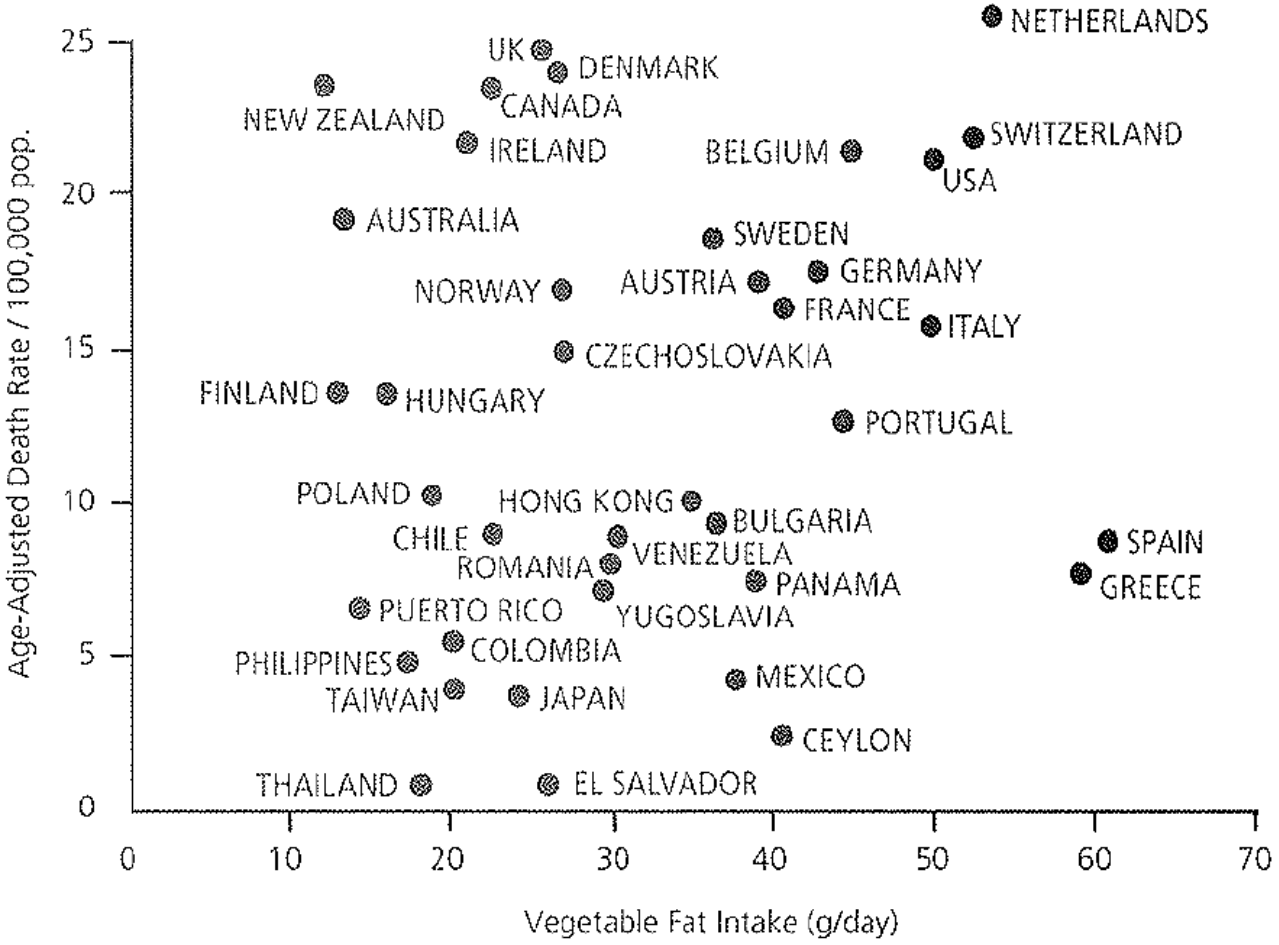
- **Dietary fat intake correlates with breast cancer**

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

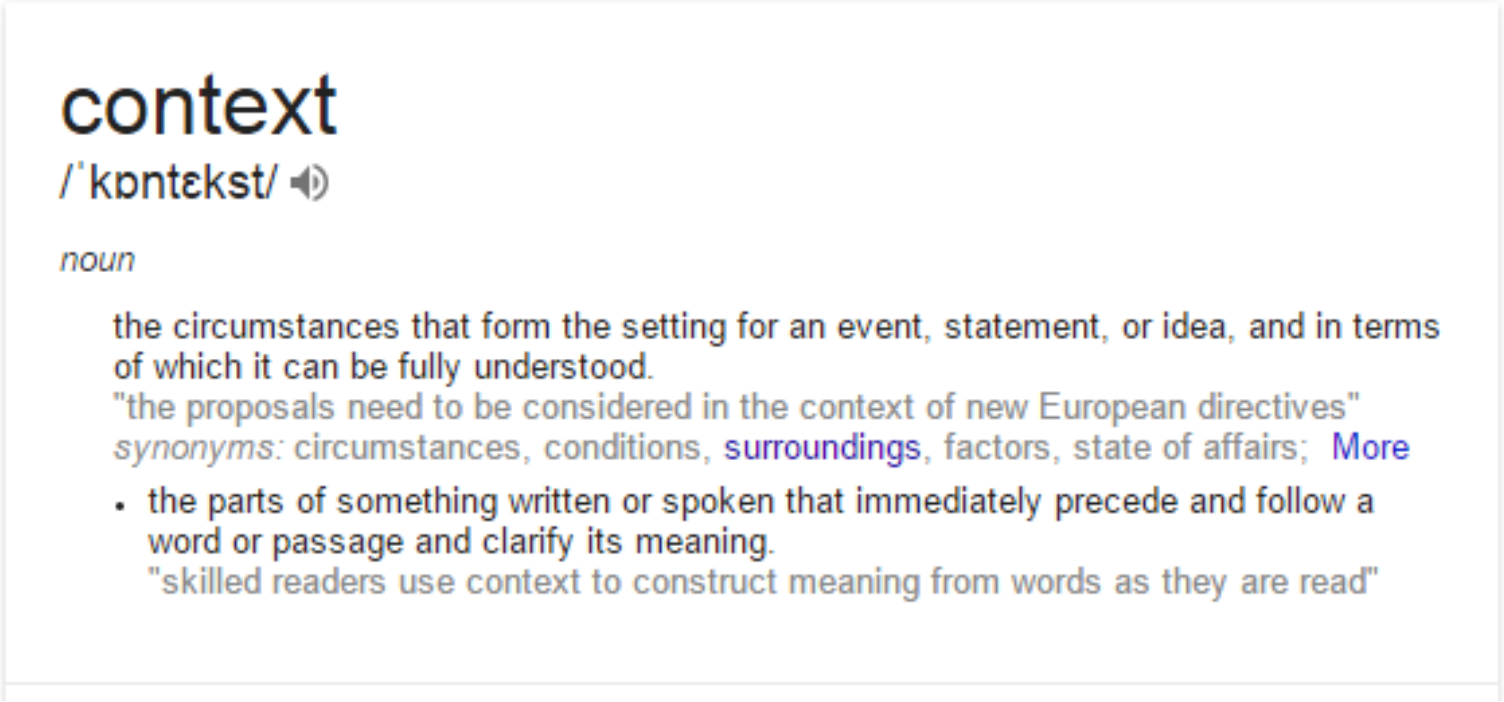
A: Dietary fat intake correlates with breast cancer

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!



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%**

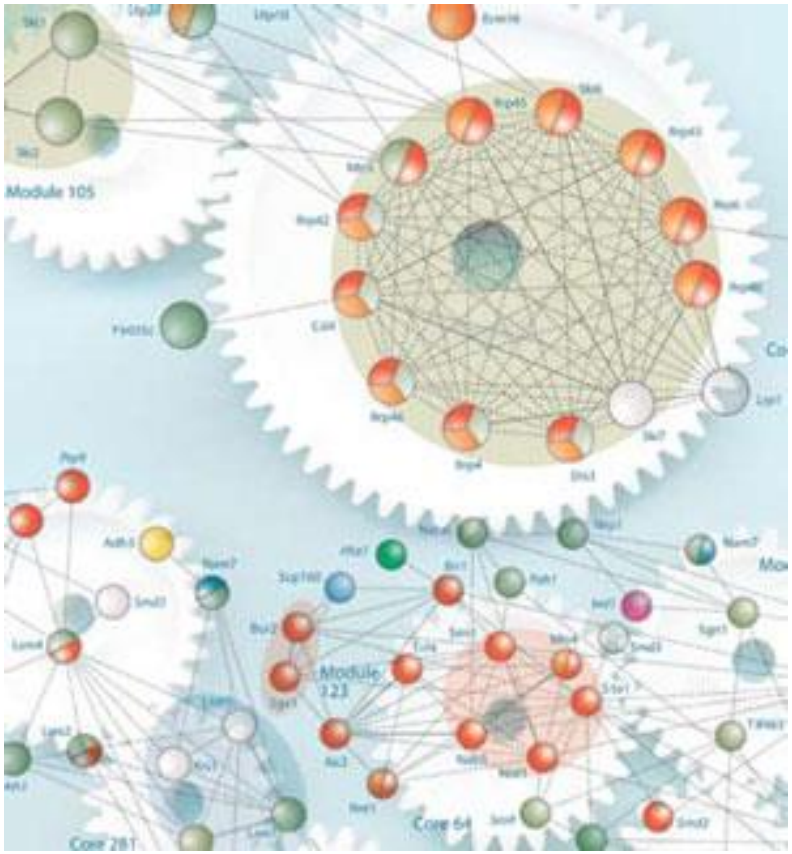
- **I.e., the answer is 2%**
- **Did you guess 95% as the answer?**

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

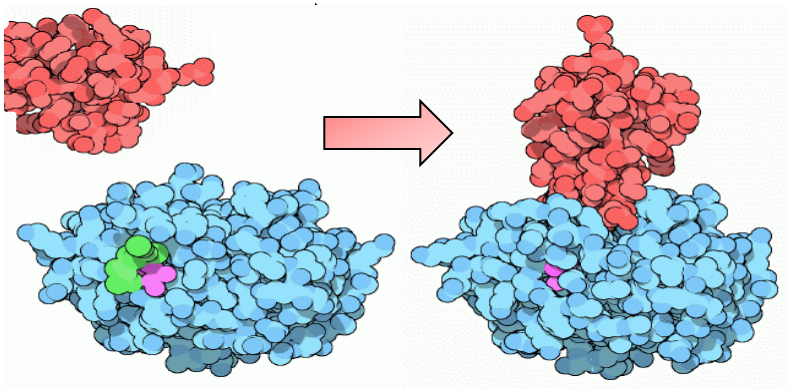
- If A/B treat the same single disease, this is ok
- If B treats two diseases, this is not sensible
- The disease has to go into the context



More may not be better

PROTEIN COMPLEXES

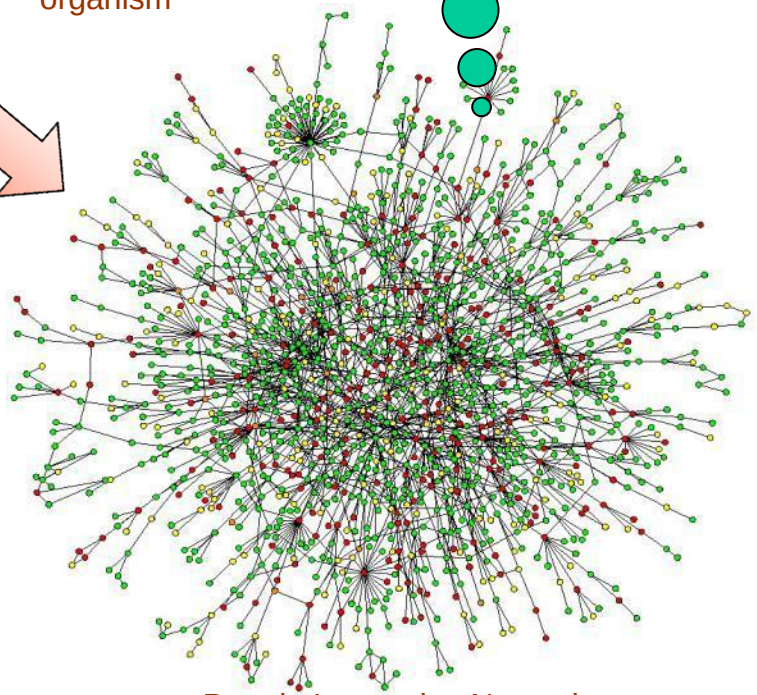
Protein-protein interaction networks



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

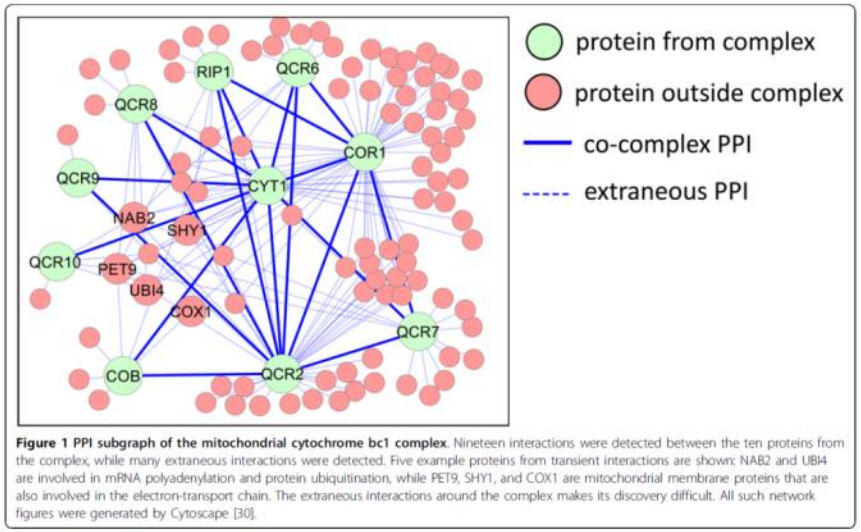


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation 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

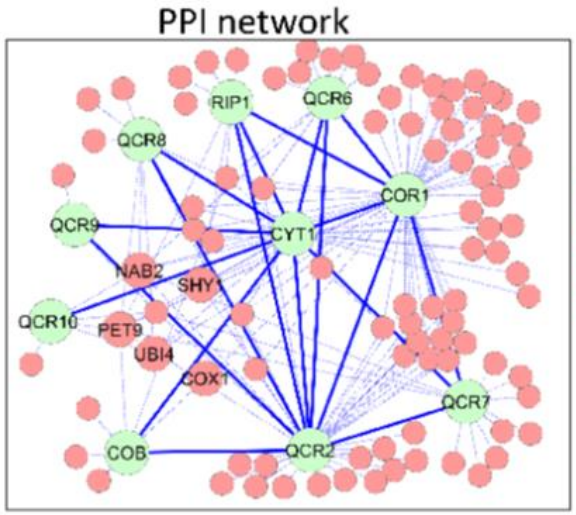
Data source	Database	Scoring method
PPI	BioGRID, IntACT, MINT	Iterative AdjustCD.
L2-PPI (indirect PPI)	BioGRID, IntACT, MINT	Iterative AdjustCD
Functional association	STRING	STRING
Literature co-occurrence	PubMed	Jaccard coefficient

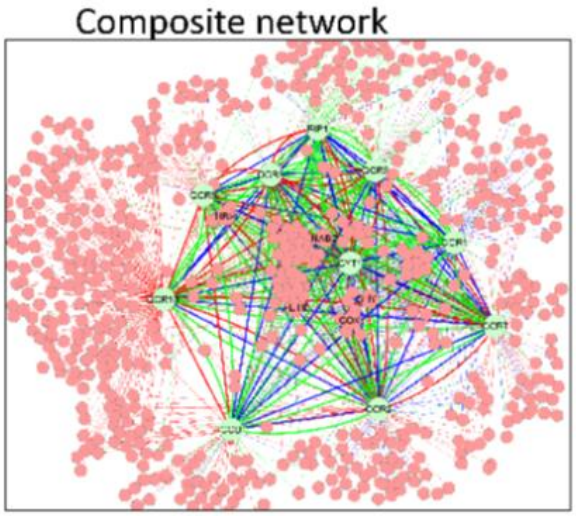
	Yeast			Human		
	# Pairs	% co-complex	coverage	# Pairs	% co-complex	coverage
PPI	106328	**5.8%**	**55%**	48098	10%	14%
L2-PPI	181175	1.1%	18%	131705	5.5%	20%
STRING	175712	5.7%	89%	311435	3.1%	27%
PubMed	161213	4.9%	70%	91751	4.3%	11%
All	531800	**2.1%**	**98%**	522668	3.4%	49%

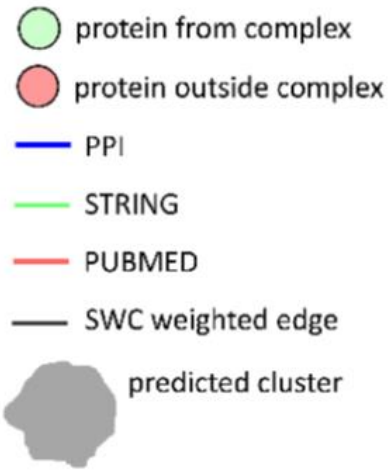
Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

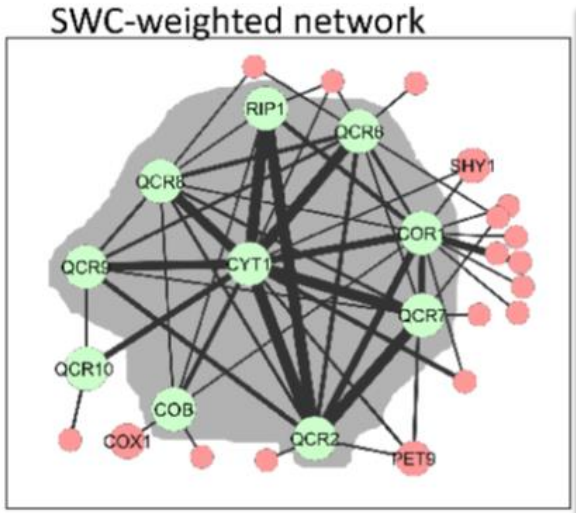
NUS National University of Singapore

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

NUS National University of Singapore

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