CS4220 Knowledge Discovery Methods for Bioinformatics

Unit1b: Essence of Knowledge Discovery (Part B: Art of Statistical Analysis)

Wong Limsoon

National University of Singapore

Outline

- **Forgotten assumptions**
  - Normal distribution
  - I.I.D.
  - Proper design of experiment
  - Domain-specific laws
- **Overlooked information**
  - Non-associations
  - Context

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		
		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**

$\Rightarrow$ **Subjects had problems choosing the right order of magnitude**

Time for Exercise #1

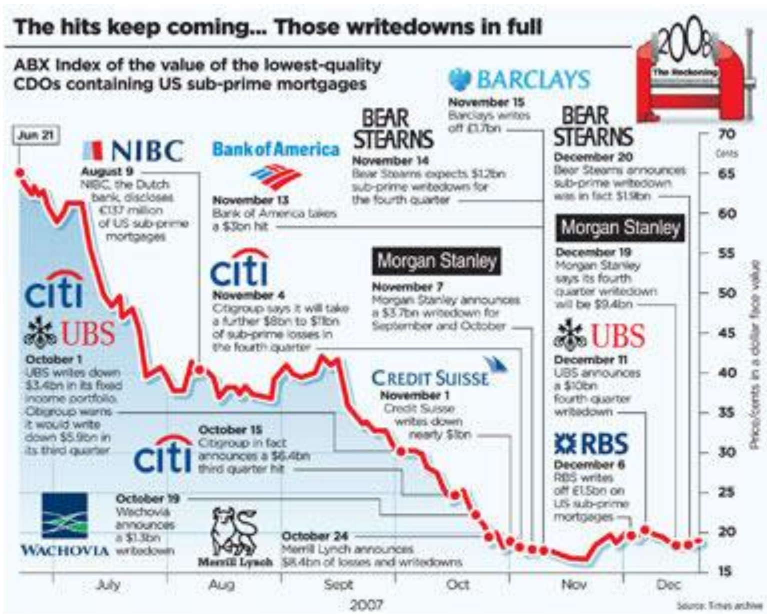
- **Suppose you are given a set S of values (e.g. the age of a group of people). Choose a number or value x so that x would be a good representative of the values in S when**
  - S is normally distributed
  - S is log-normally distributed
  - S has some arbitrary distribution

- **What is the general principle underlying your choices?**

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

and what held yesterday may not hold today

2007 Financial Crisis

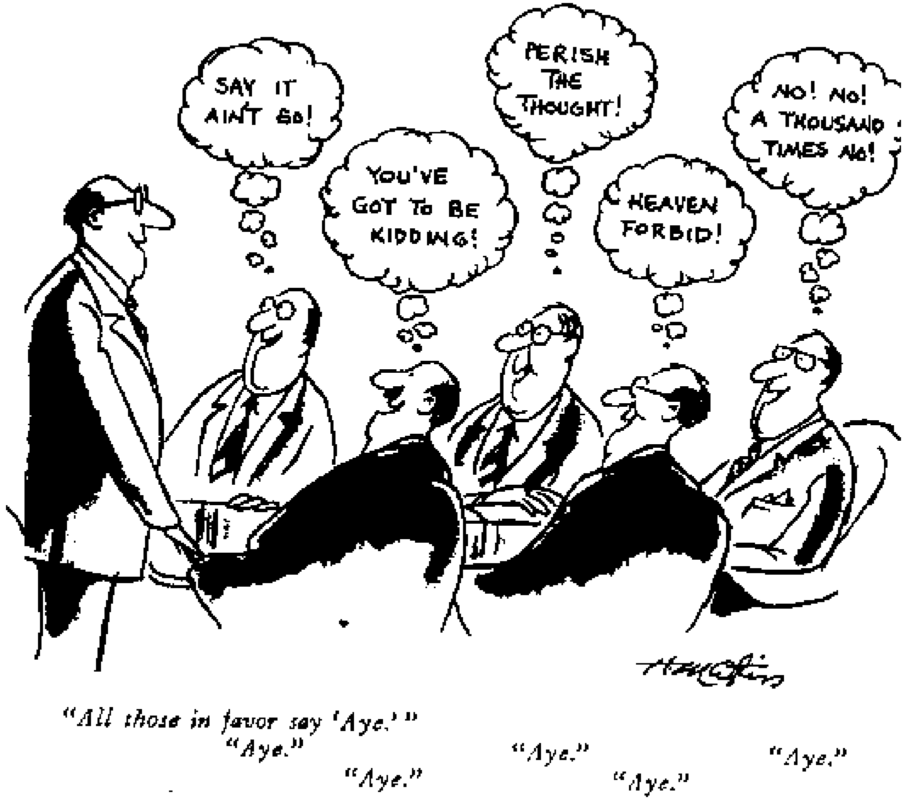


- **All of them religiously check VaR (Value at Risk) everyday**

- **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

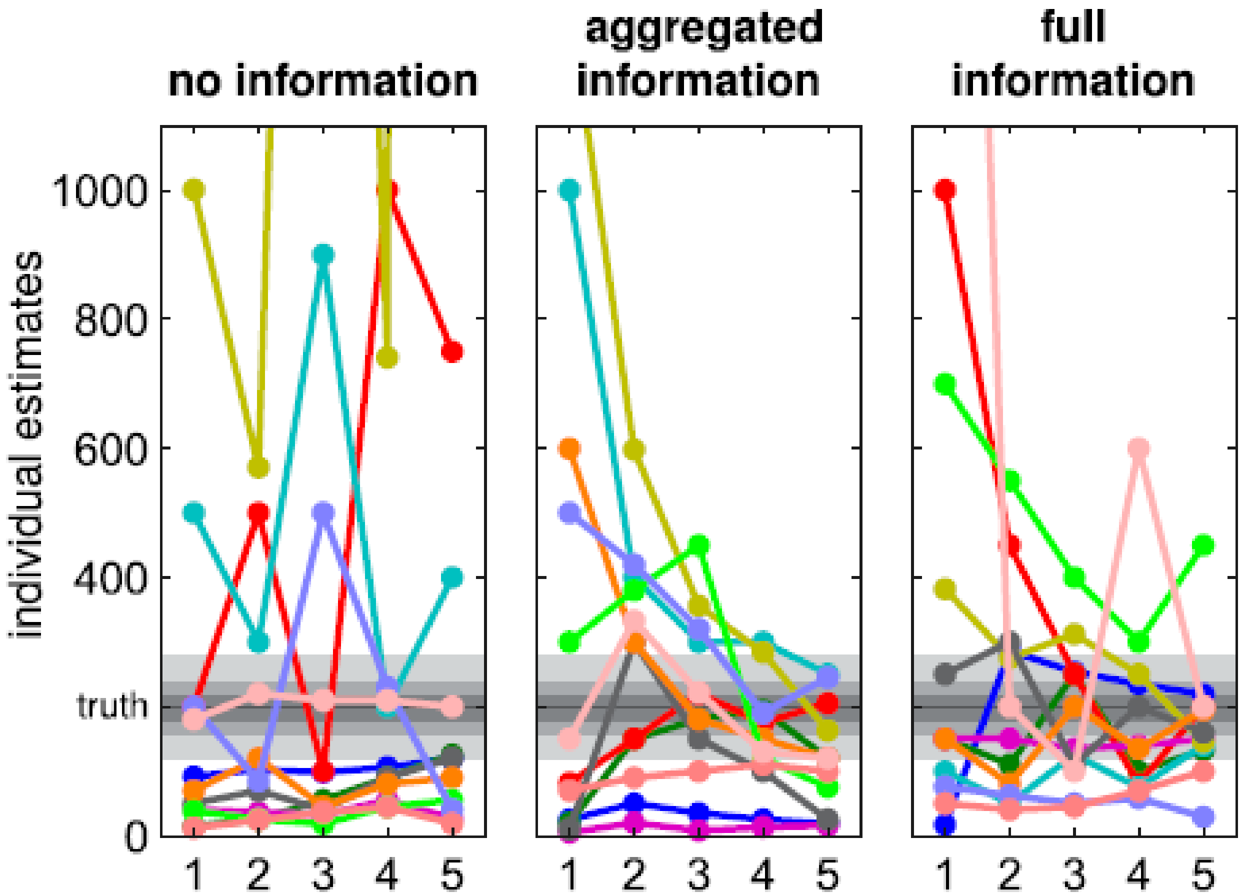
I.I.D.

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**

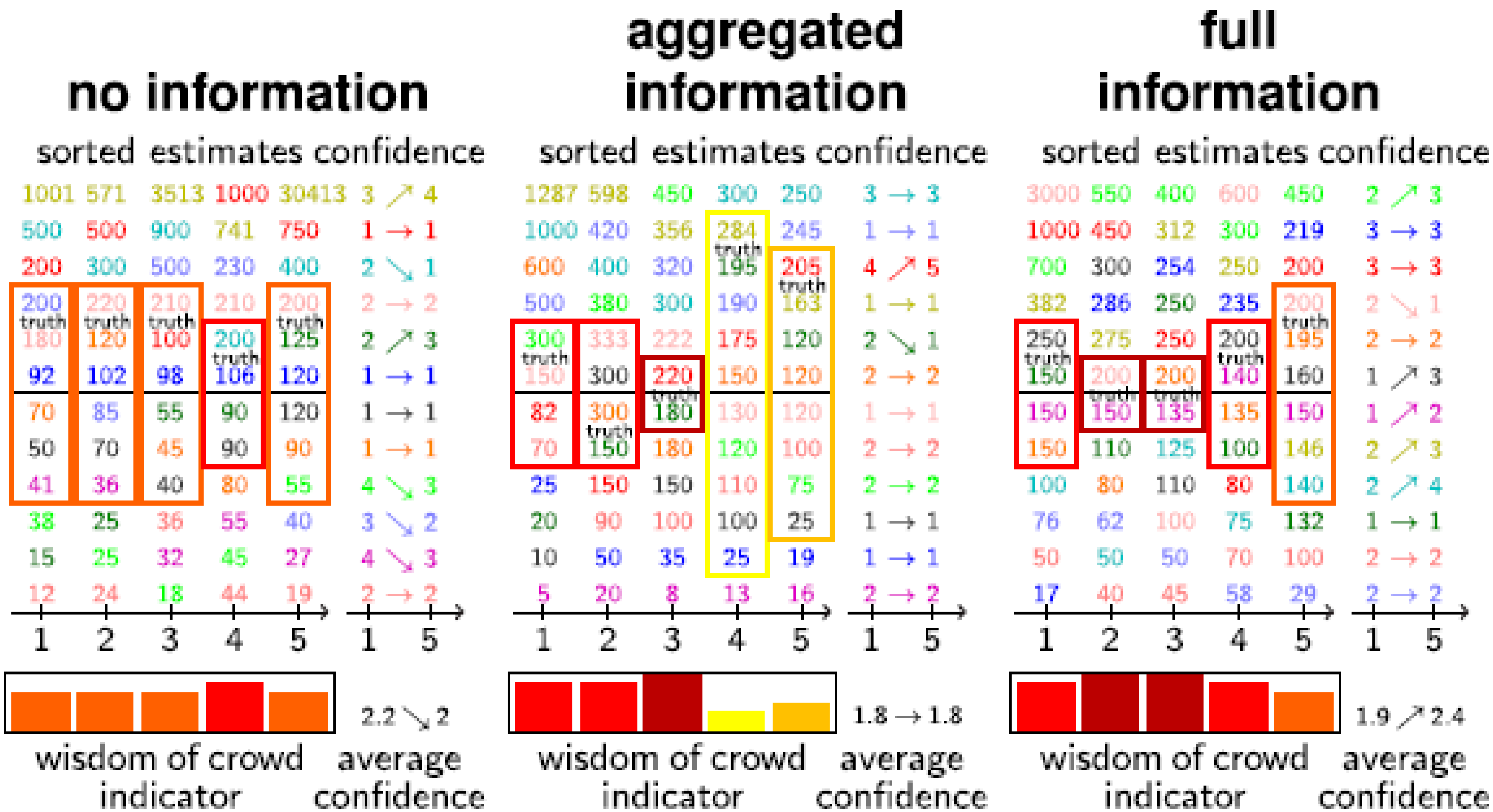
Social influence effect



- **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

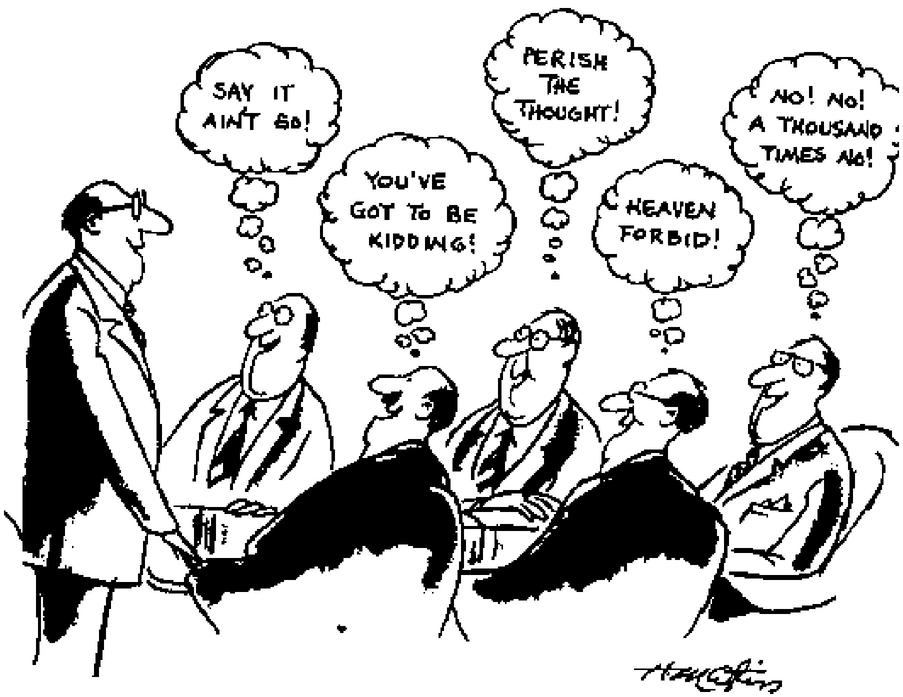
- **Social influence triggers convergence of individual estimates**

- **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**

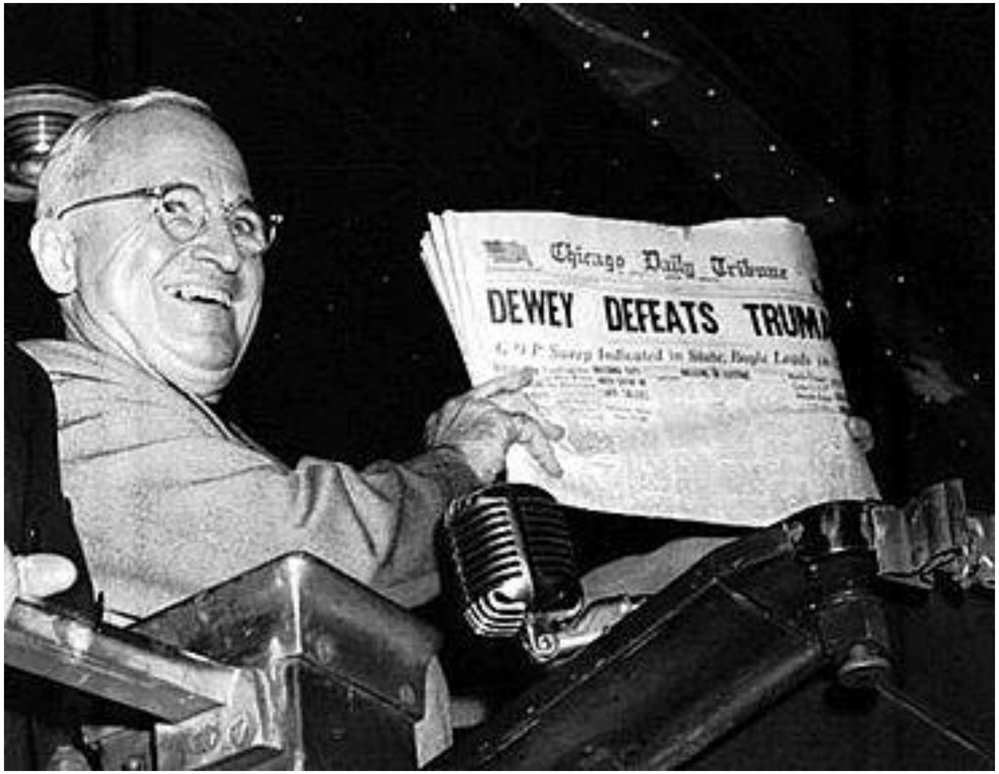
Related issue: People do not say what they really want to say



"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

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

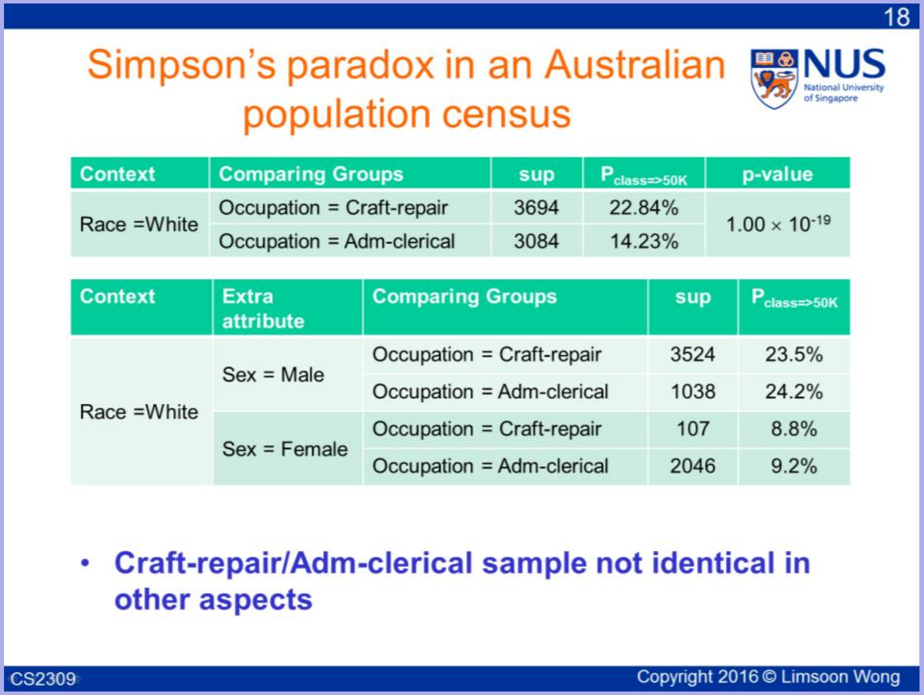
Context	Comparing Groups	sup	$P_{class=>50K}$	p-value
Race =White	Occupation = Craft-repair	3694	22.84%	1.00×10^{-19}
	Occupation = Adm-clerical	3084	14.23%	

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**

Time for Exercise #2

- **Slide #18 suggests that men earn more than women. How would you verify this hypothesis? Should you do a chi-square test using the table shown below?**



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**

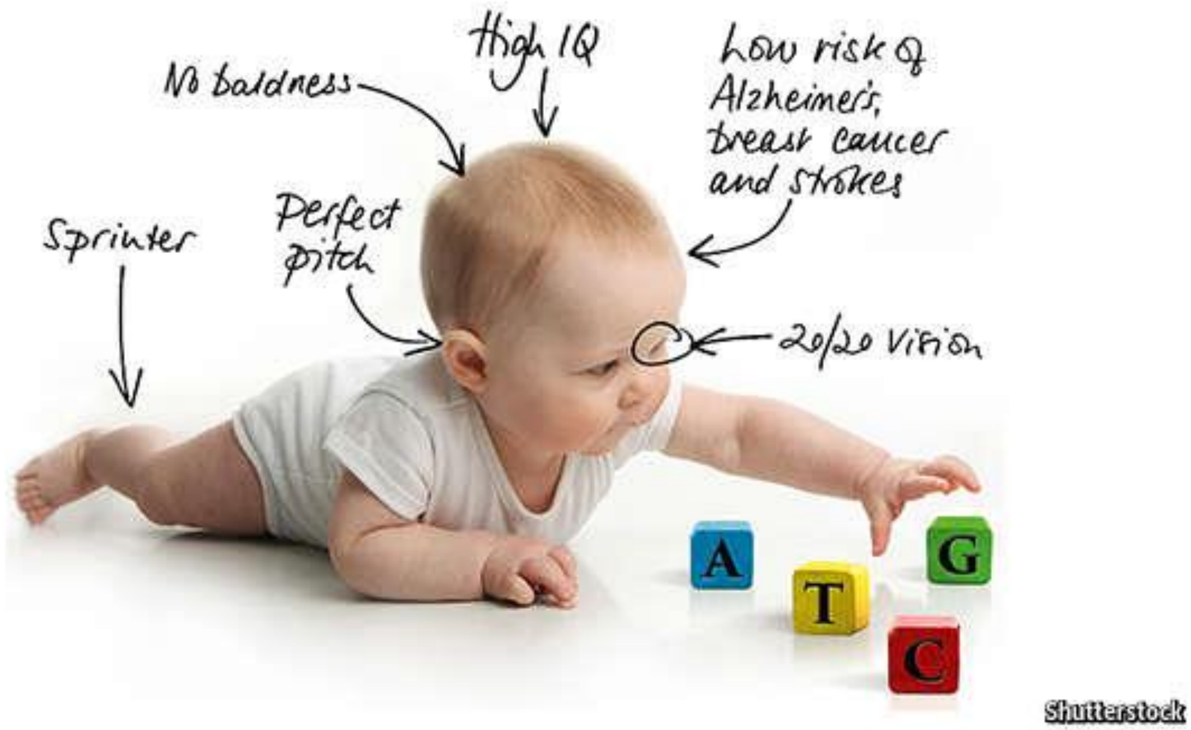
	Earn <50k	Earn ≥50k
Sex = Male	3483 (76%)	1079 (24%)
Sex = Female	1955 (91%)	198 (9%)

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!"

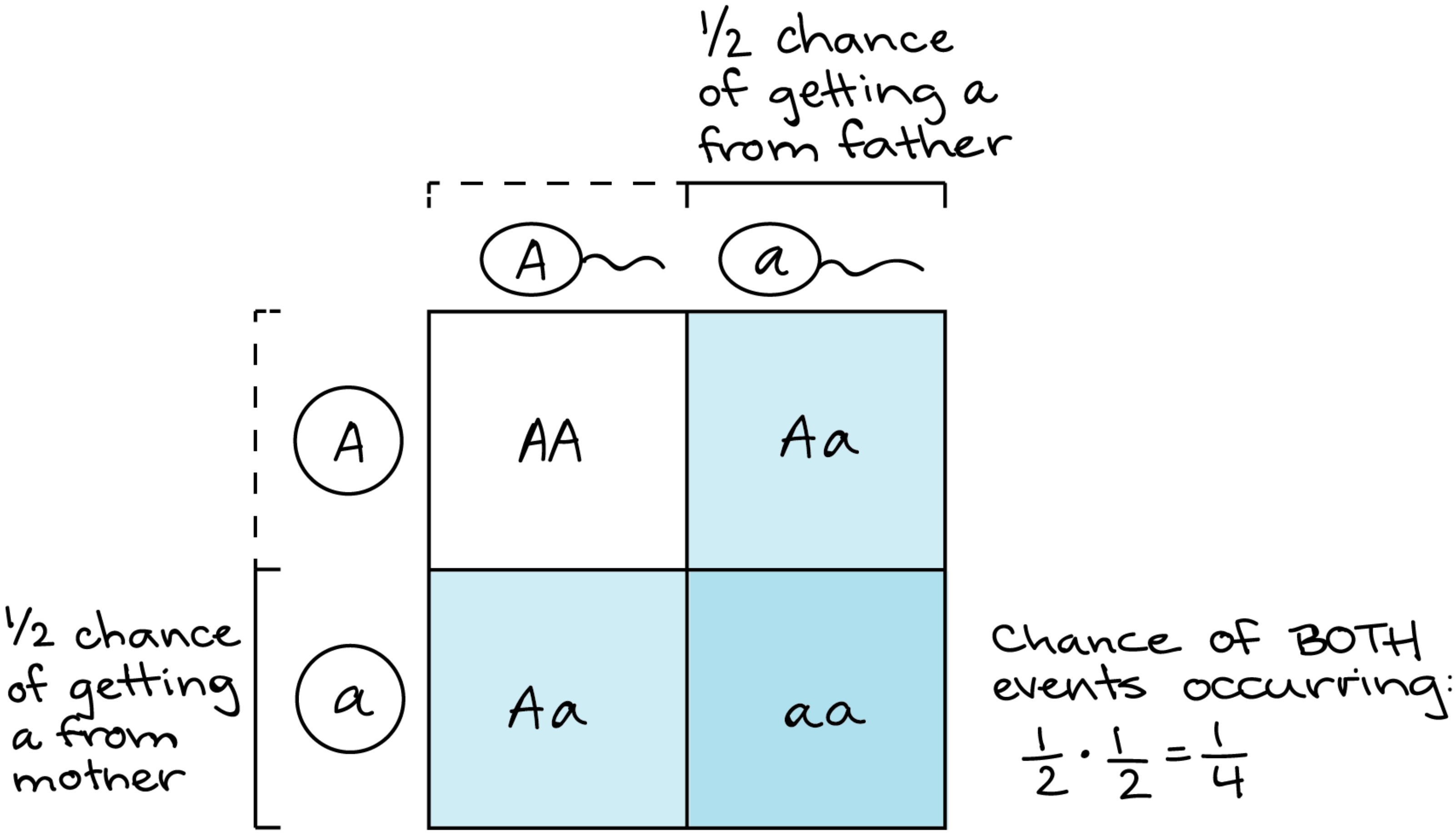
The reason the Tribune was mistaken is that their editor trusted the results of a phone survey… Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses.



Forgotten assumptions

DOMAIN-SPECIFIC LAWS

A basic rule of human genetics



A suspicious contingency table

SNP	Genotypes	Group				χ^2	*P* value
		Controls [n(%)]		Cases [n(%)]			
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.

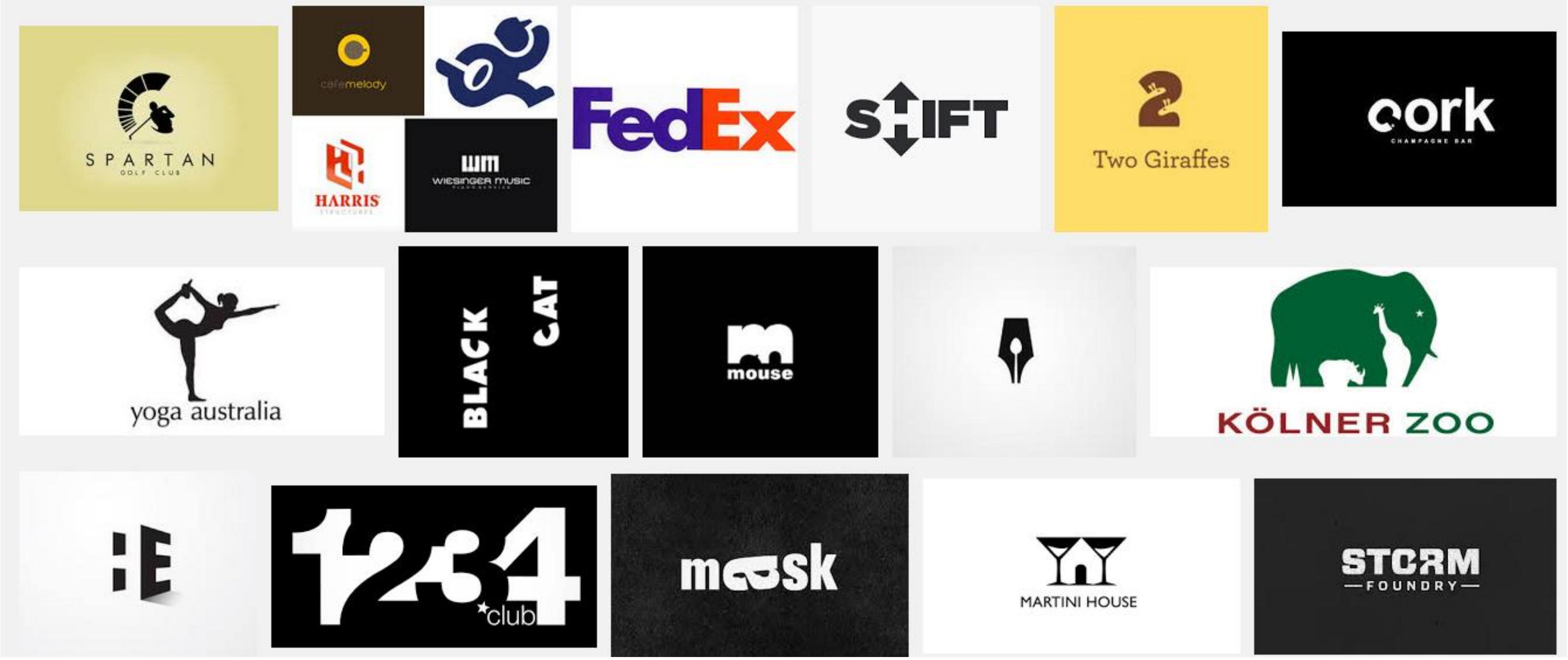
Time for Exercise #3

- **Slide #24 says the contingency table looks suspicious. Why?**

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.



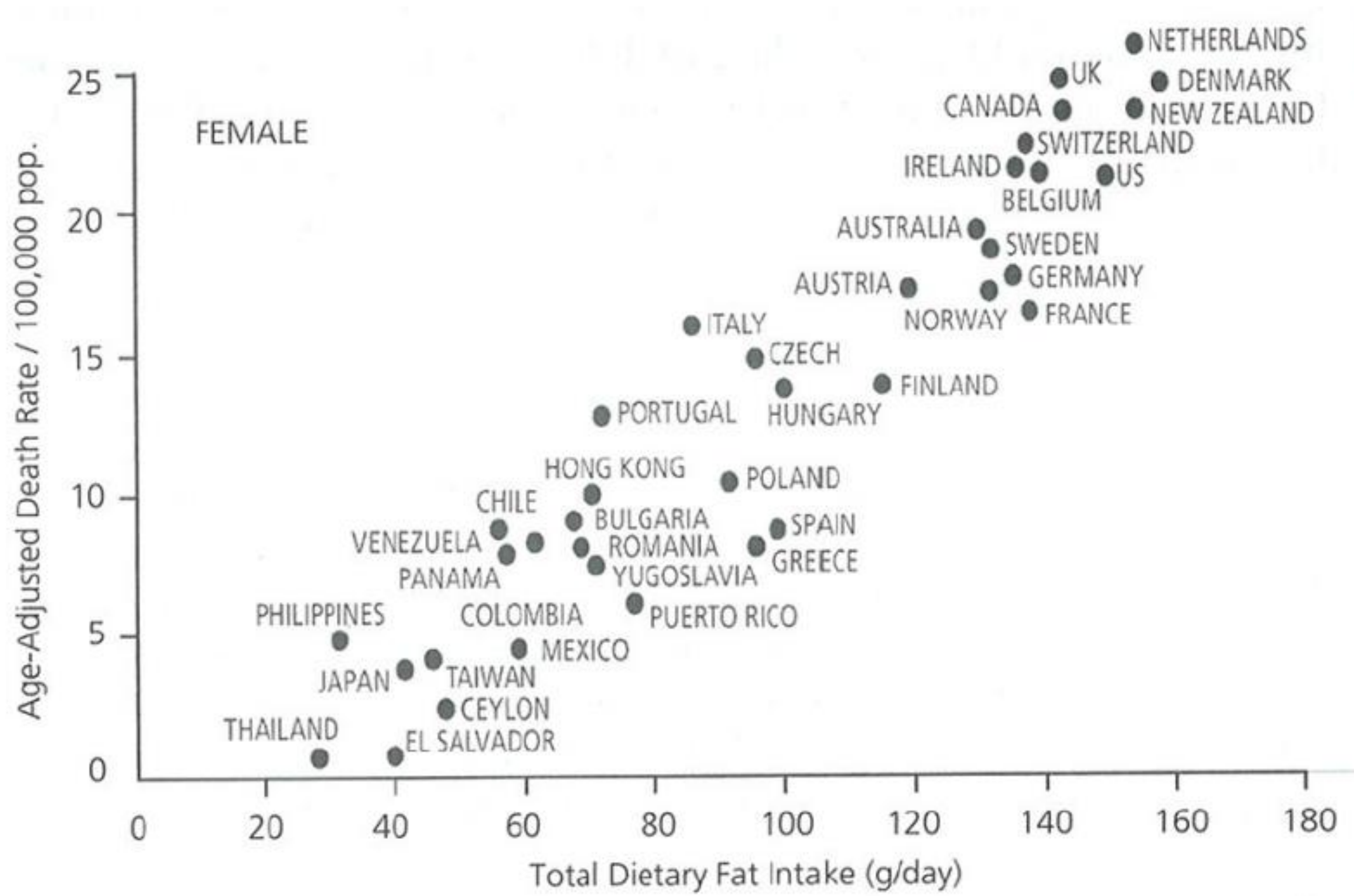
Overlooked information

NON-ASSOCIATIONS

We tend to ignore non-associations

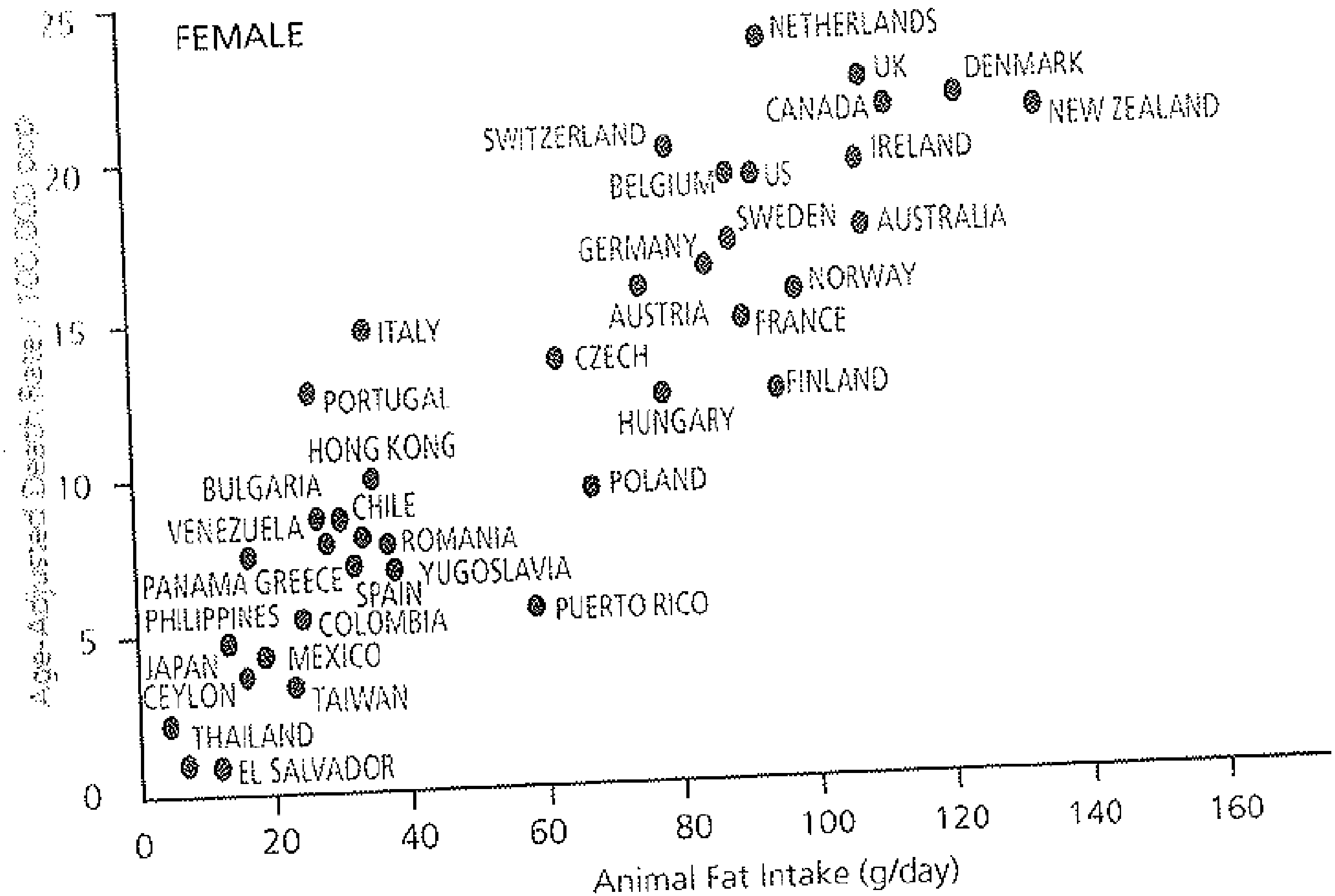
- **We have many technologies to look for associations and correlations**
  – Frequent patterns
  – Association rules
  – …
- **We tend to ignore non-associations**
  – We think they are not interesting / informative
  – There are too many of them
- **We also tend to ignore relationship between associations**

We love to find correlations like this…



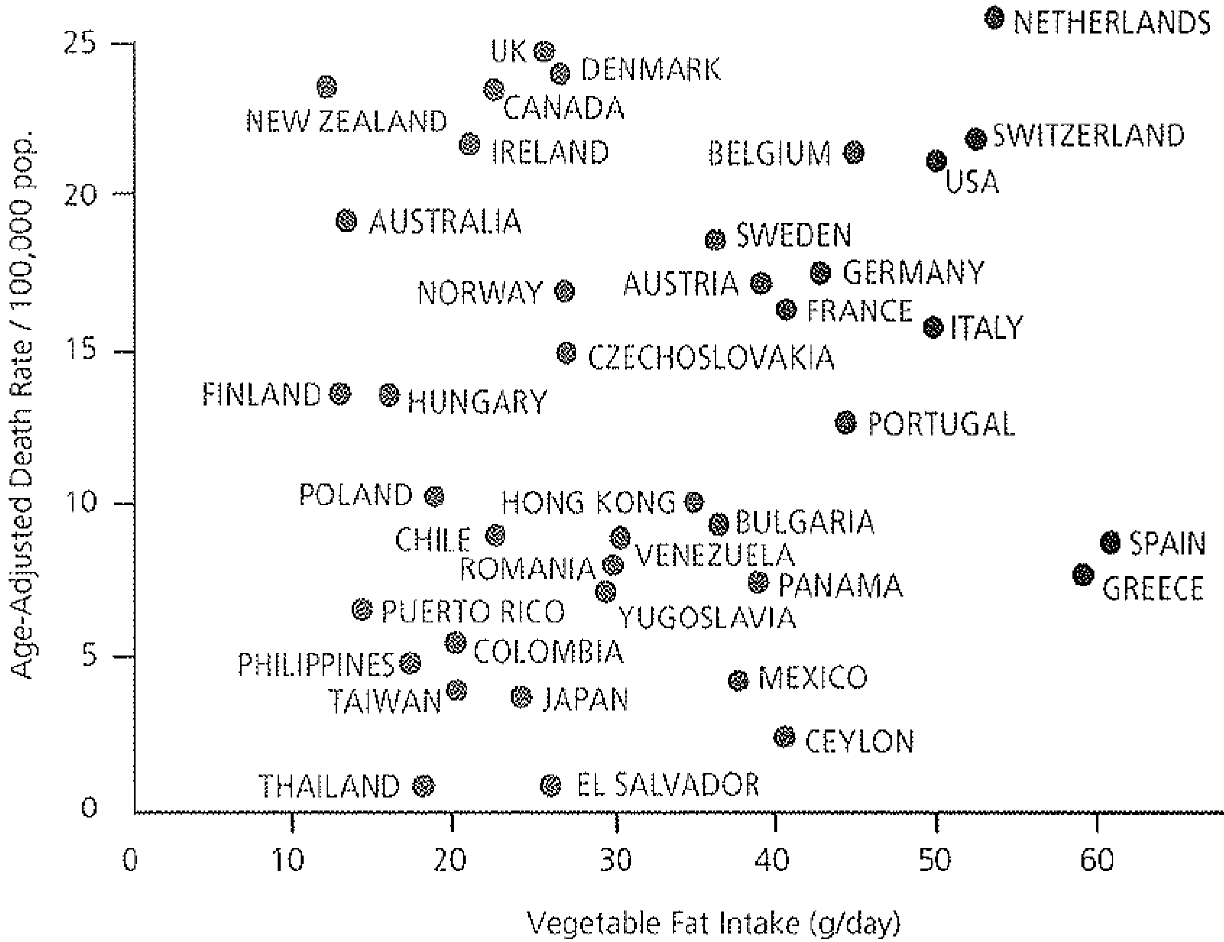
- **Dietary fat intake correlates with breast cancer**

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

$\Rightarrow$ 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})

$\Rightarrow$ **Frequent pattern mining, but be careful with support threshold, need to relativize to context**

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

Time for Exercise #4

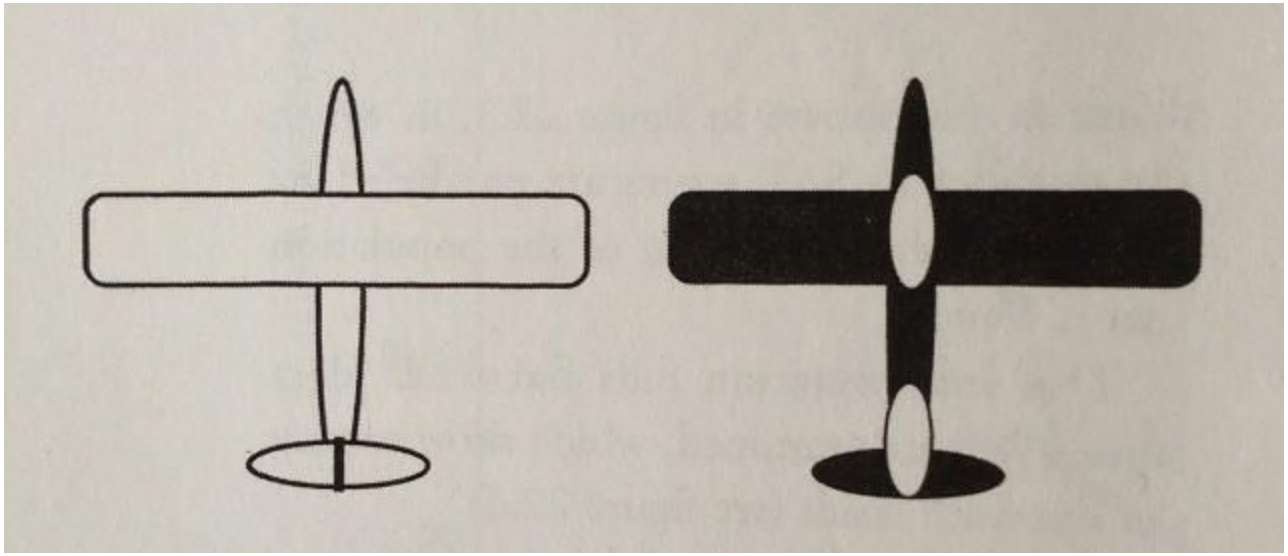
- **Suppose a test of a disease presents a rate of 5% false positives, and the disease strikes 1/1000 of the population. Let's say people are tested randomly and a particular patient's test is positive. What is the probability that he is stricken with the disease?**

What have we learned?

- **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

Abraham Wald's analysis of survivability of bombers in WWII

Look this story up



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

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