Computing for Medicine

Data Science: Modeling

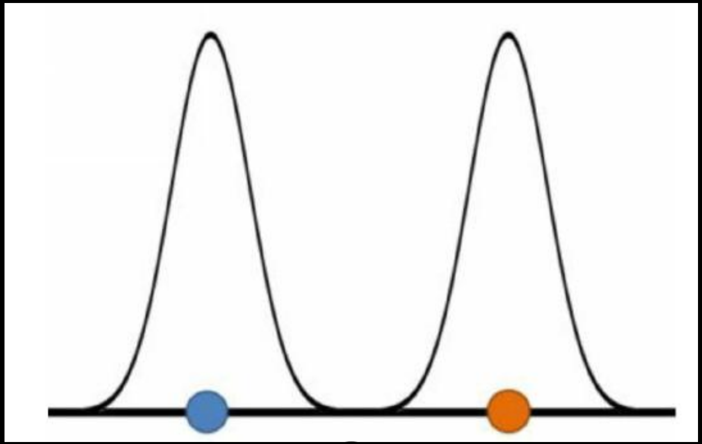
Tavpritesh Sethi

Statistical Inference

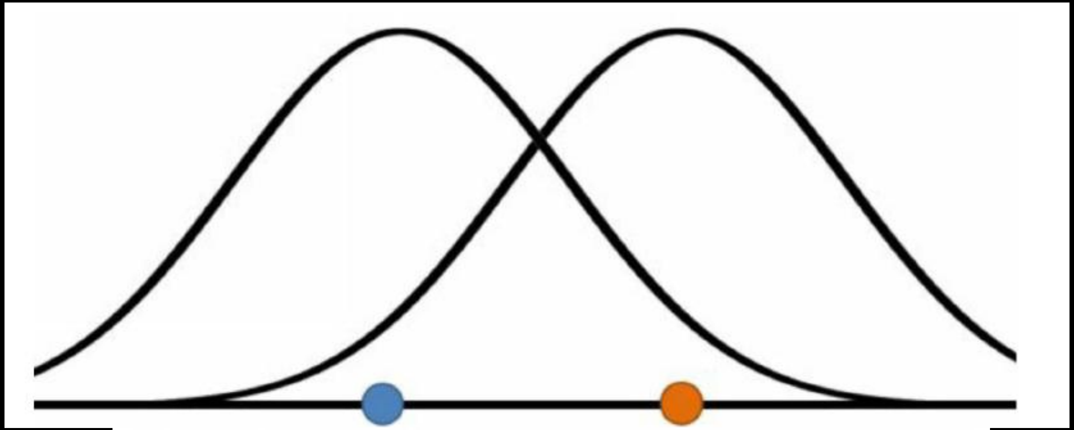
- Hypothesis Testing
  - T test, ANOVA, Wilcoxon Rank sum test etc

- Parameter Estimation
  - Simple Linear Models, Generalized Linear Models etc

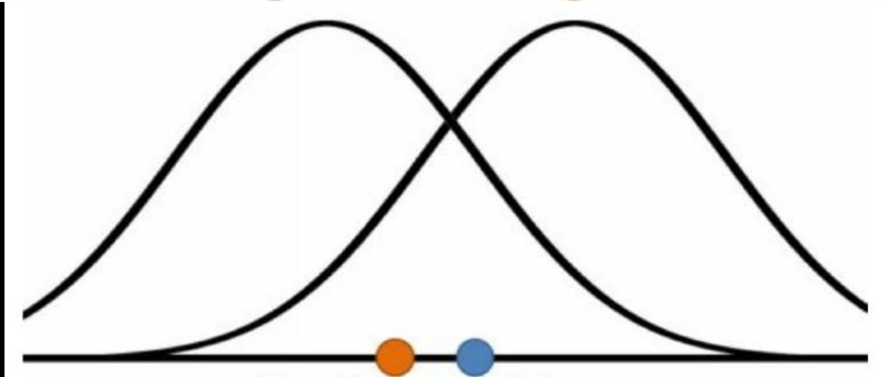
Example: Use of Normal distribution in Inference



1

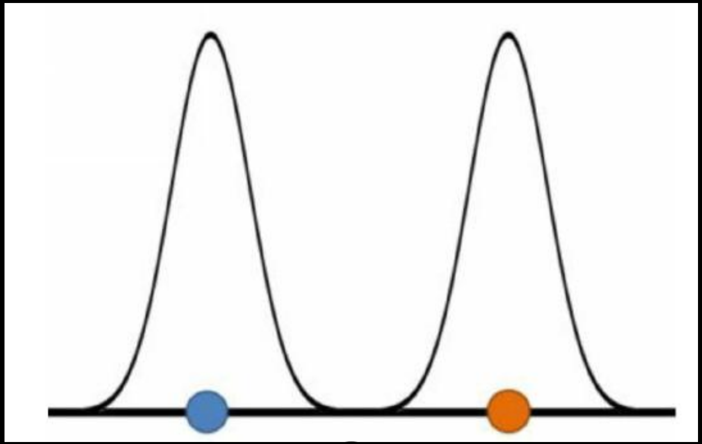


2

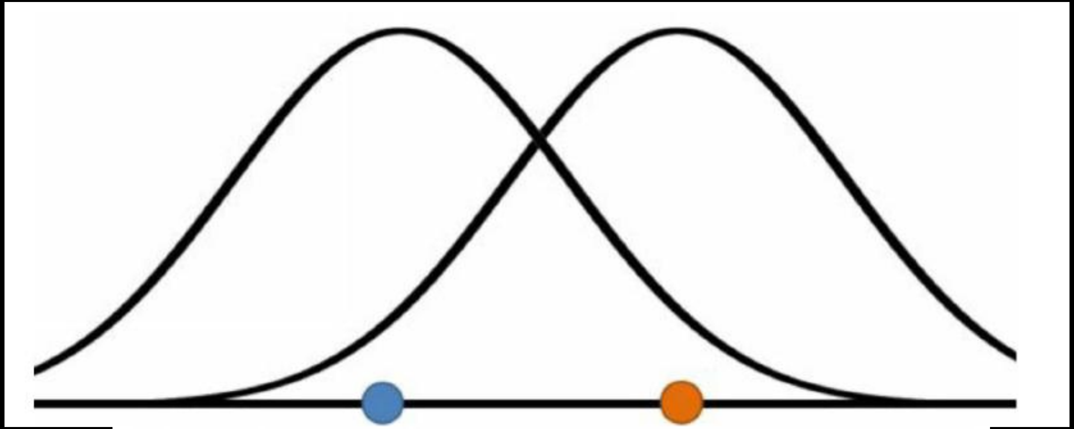


3

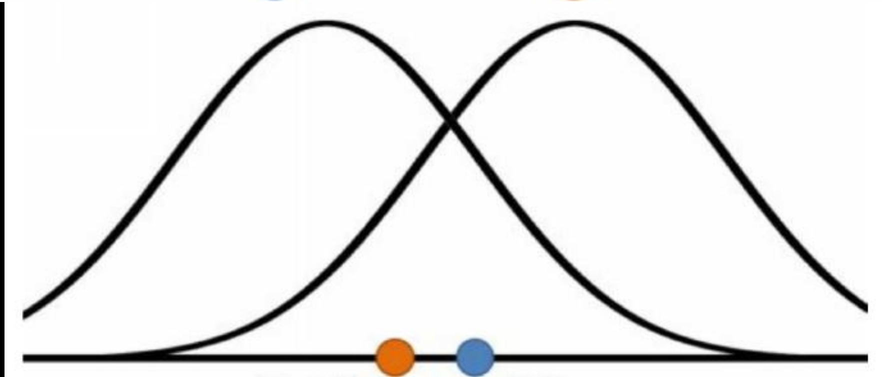
Hypothesis testing



1



2



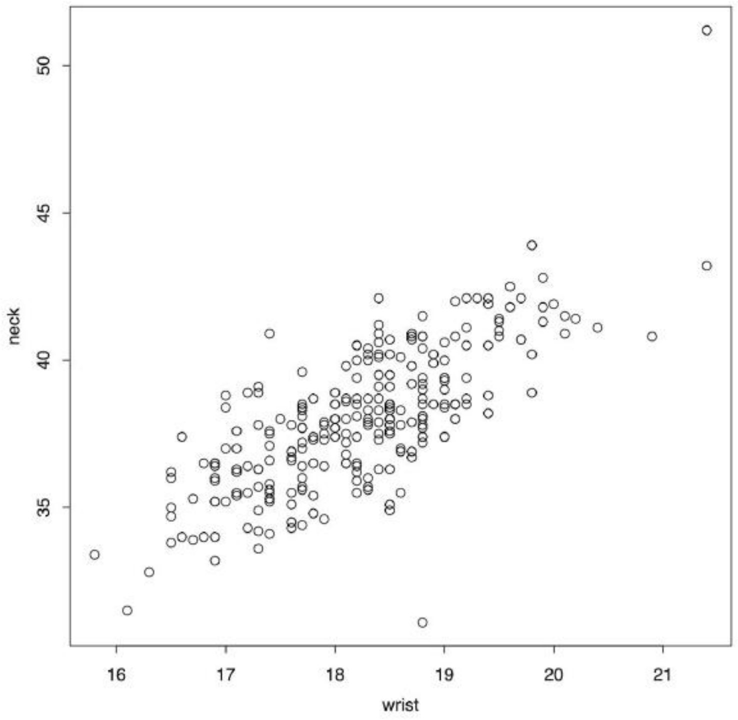
3

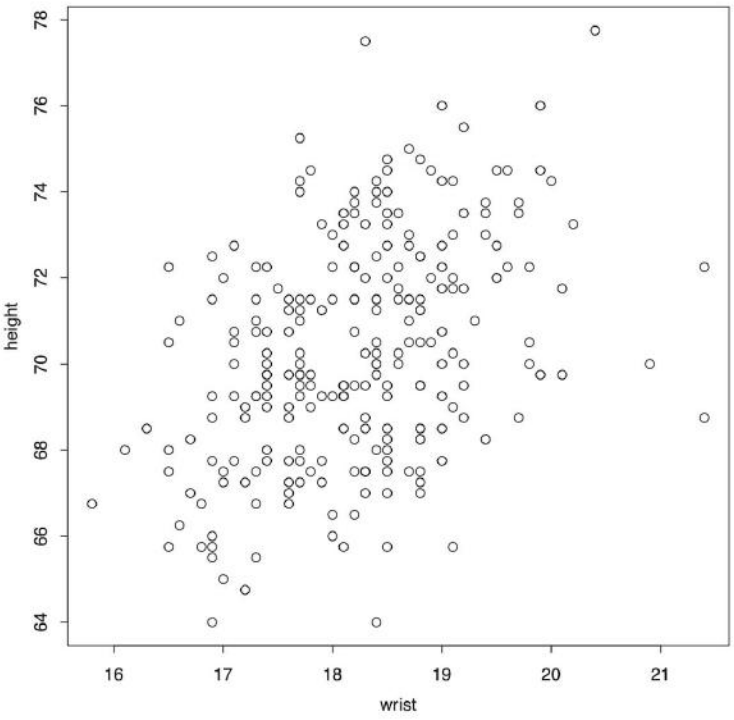
Association: Thinking with Correlations

$$\text{cov}(x,y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}).$$

$$\text{cor}(x,y) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \text{cov}(x,y)/(s_x s_y).$$

Correlations is not Causation





Anscombe's Quartet

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

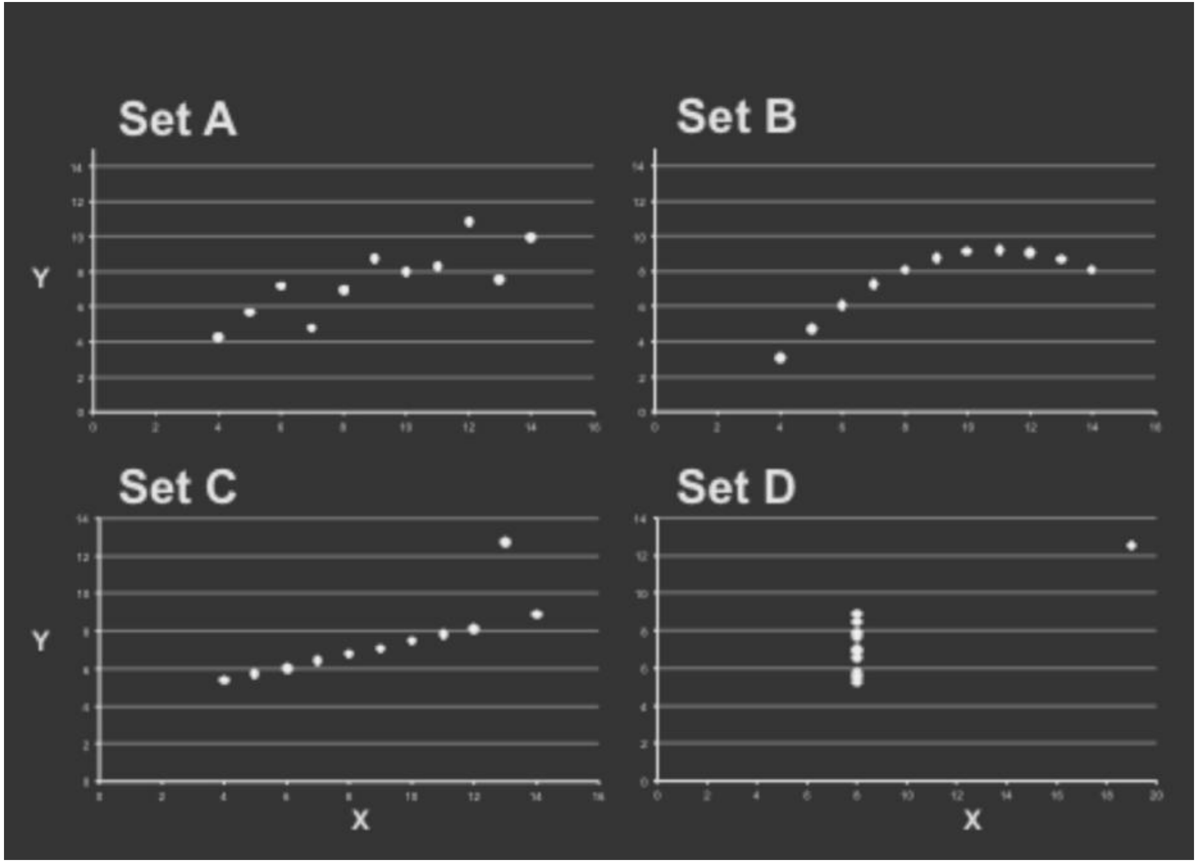
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\,X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

Anscombe's Quartet



Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada
{*first.last*}@autodesk.com

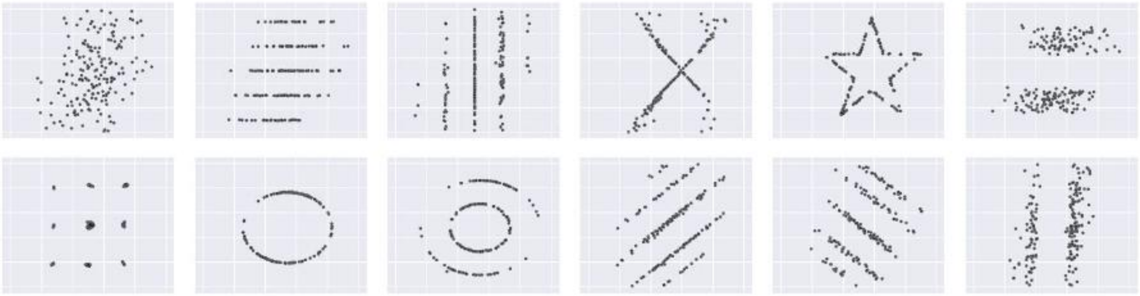
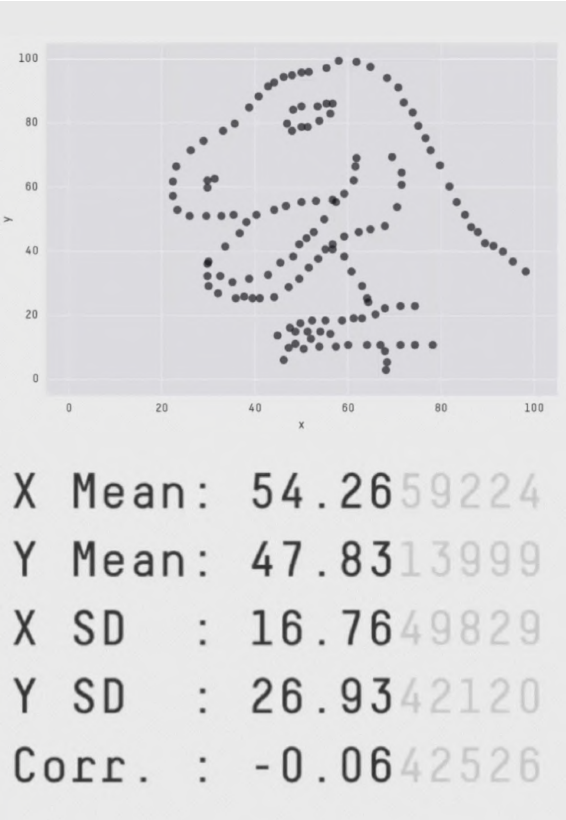
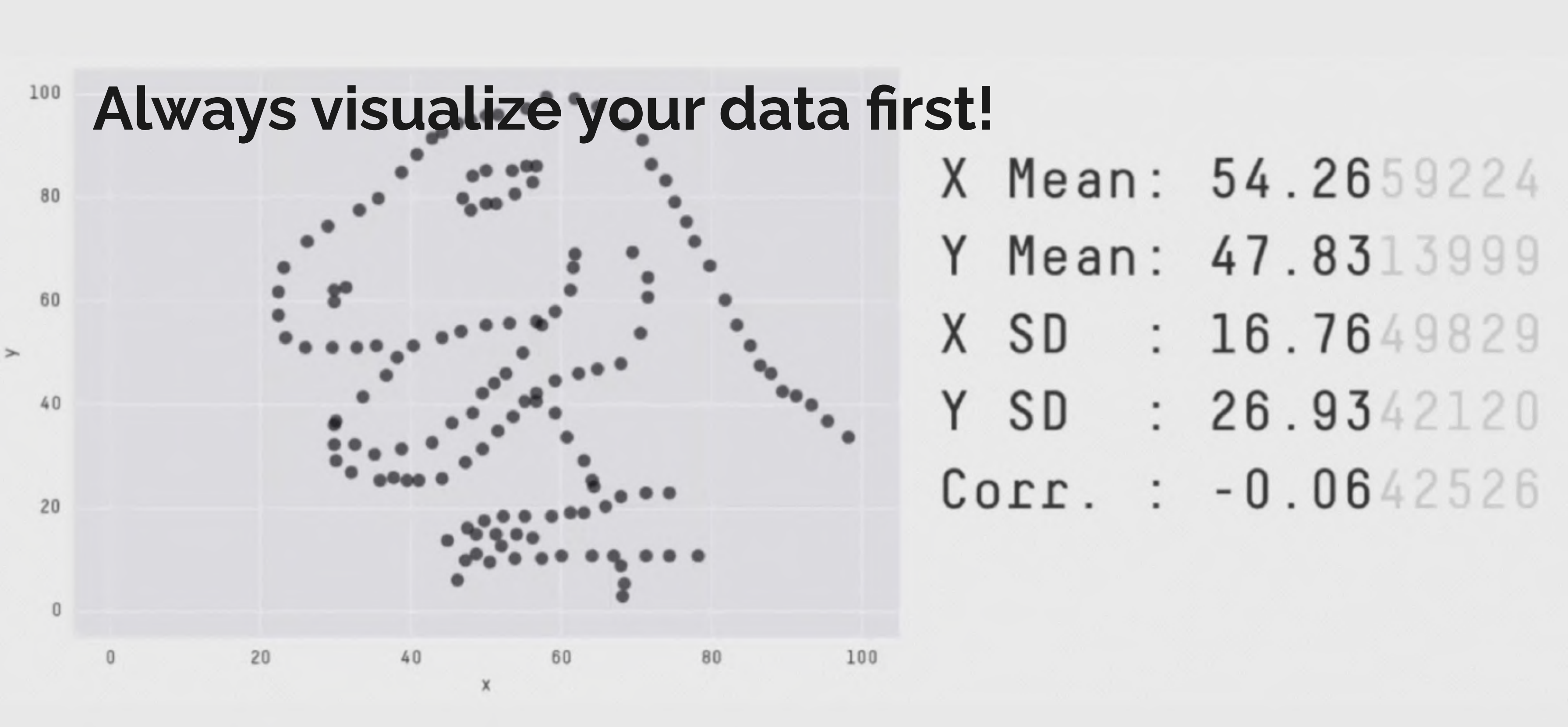


Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. (*$\bar{x}$ =54.02, $\bar{y}$ = 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = +0.32)*

https://www.autodesk.com/research/publications/same-stats-different-graphs

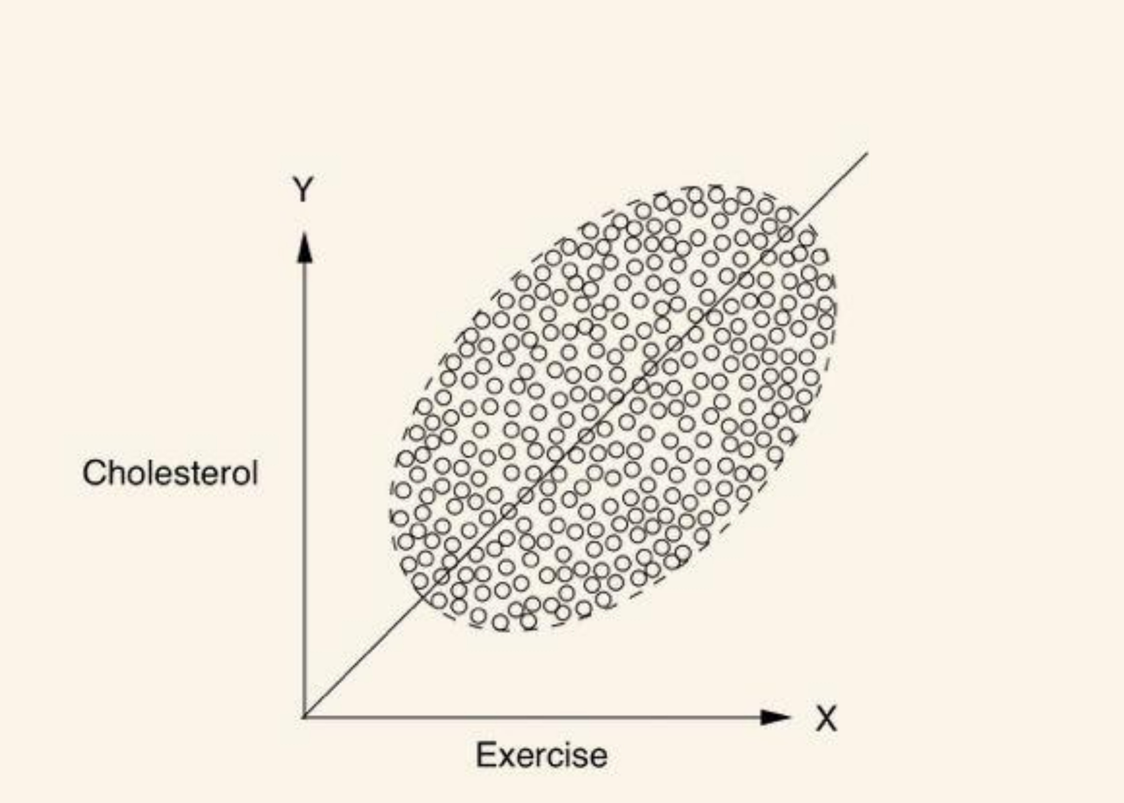


Always visualize your data first!

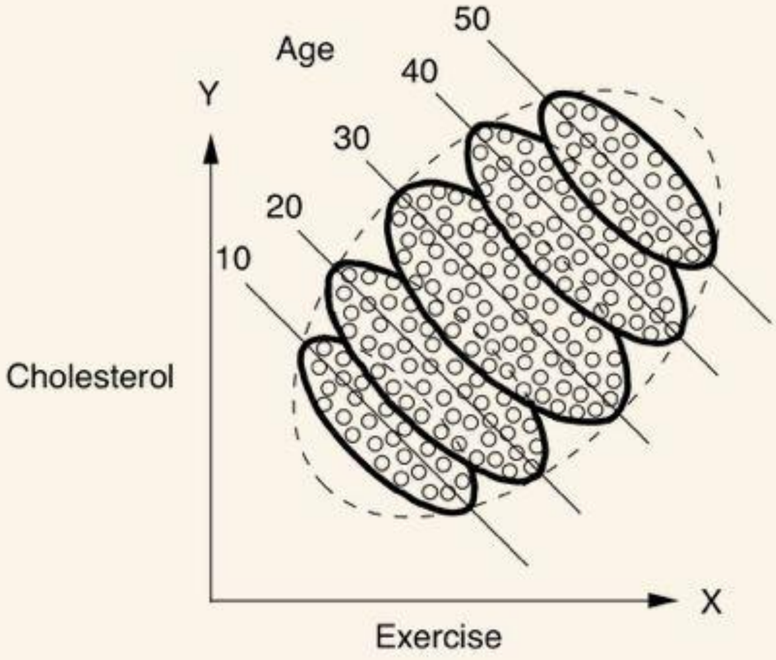


Confounding

Correlations and Confounding



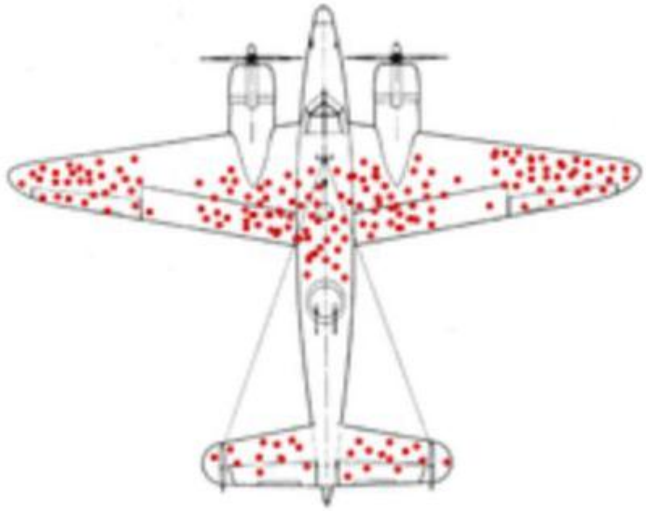
Hidden Confounding Factors



http://bayes.cs.ucla.edu/PRIMER/

Correlation is not causation!

Where to put the armor?



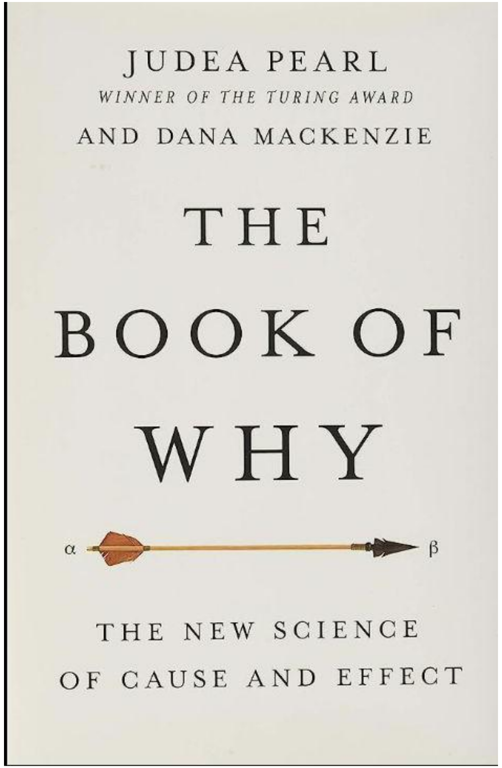
Israeli Data: August 15,2021

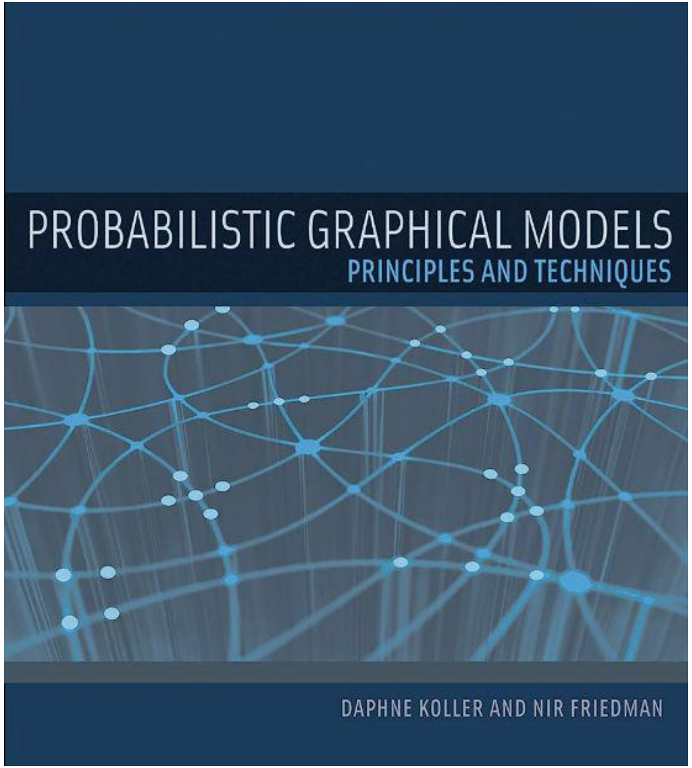
From: https://datadashboard.health.gov.il/COVID-19/general

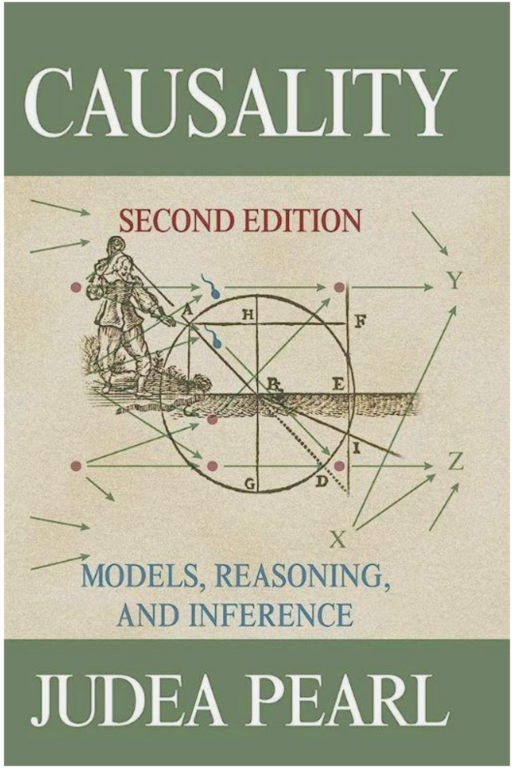
Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,170,563 90.4%	171 90.9	290 13.6	**85.2%**

This strange result illustrates something called **Simpson's Paradox**, in this case meaning you can have **very high efficacy in each group**, but the **overall efficacy looks much lower** because one group (older people) is ***more vaccinated*** and have a ***much higher risk of severe disease***.

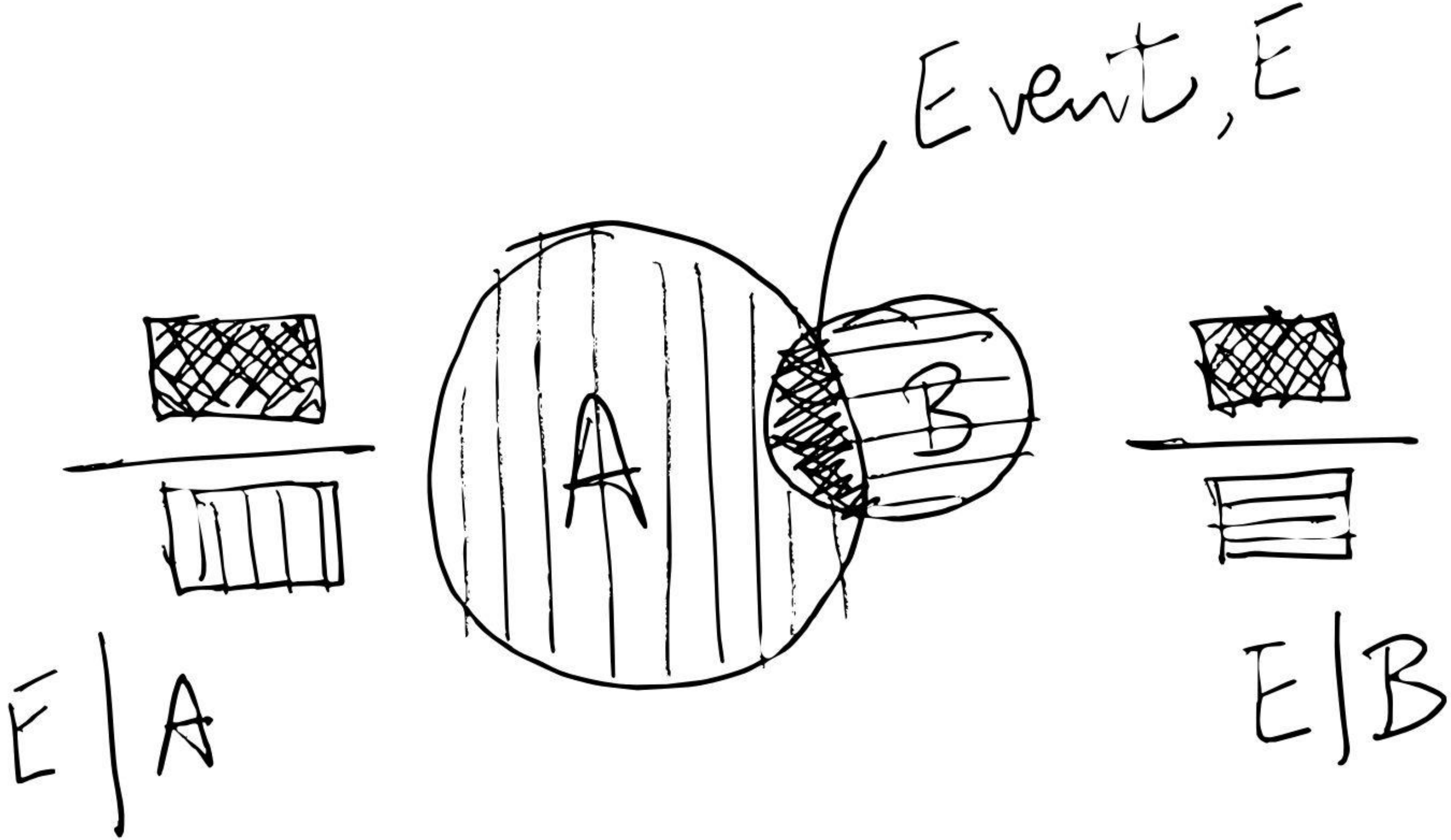
Recommended Reading



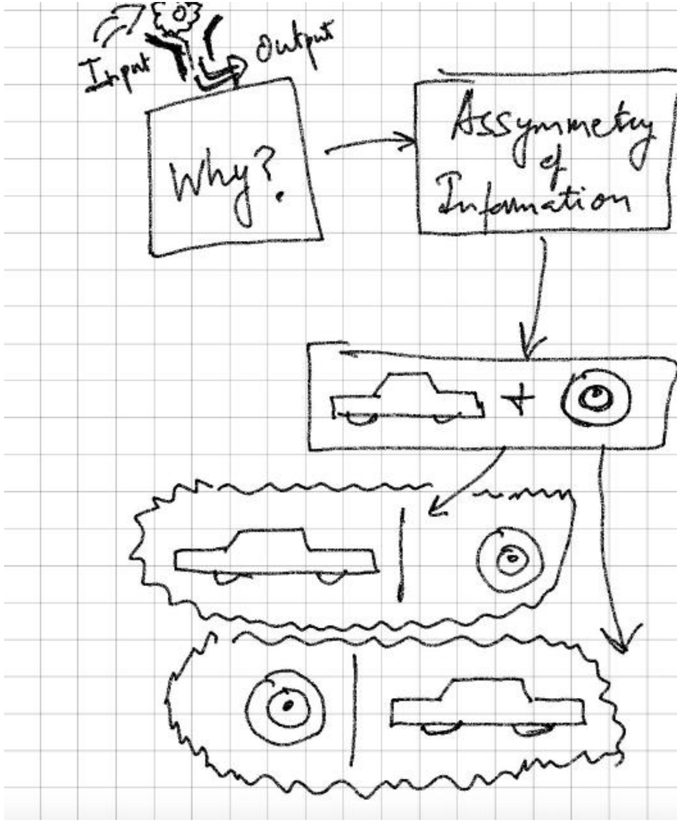


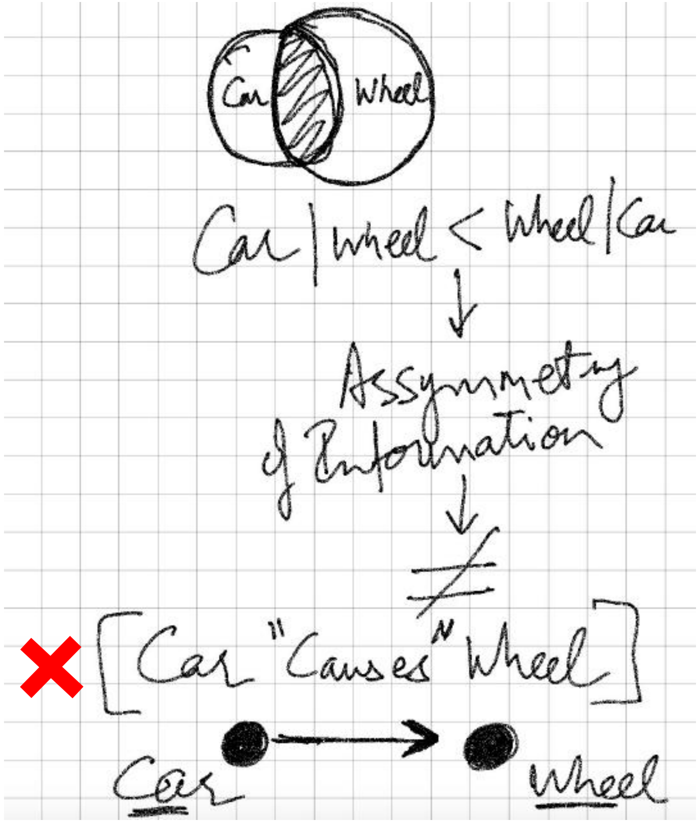


Conditional Probability

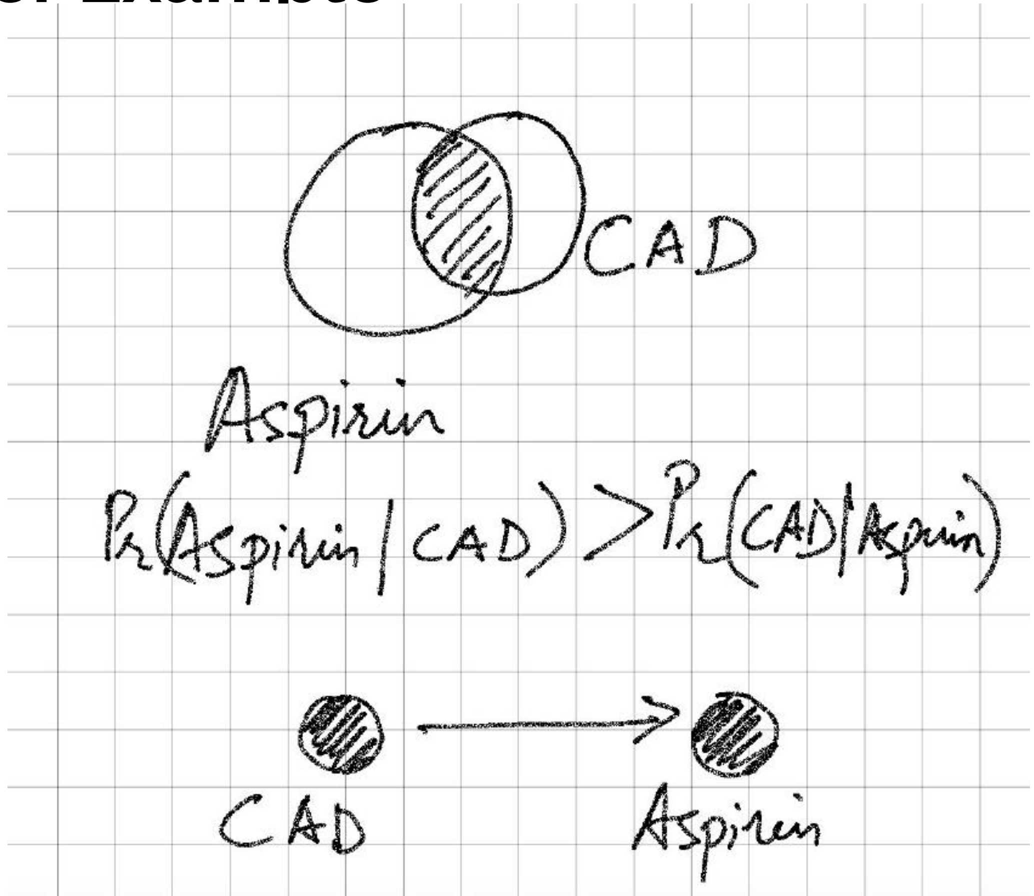


Bayes Rule Exploits Real World Asymmetry

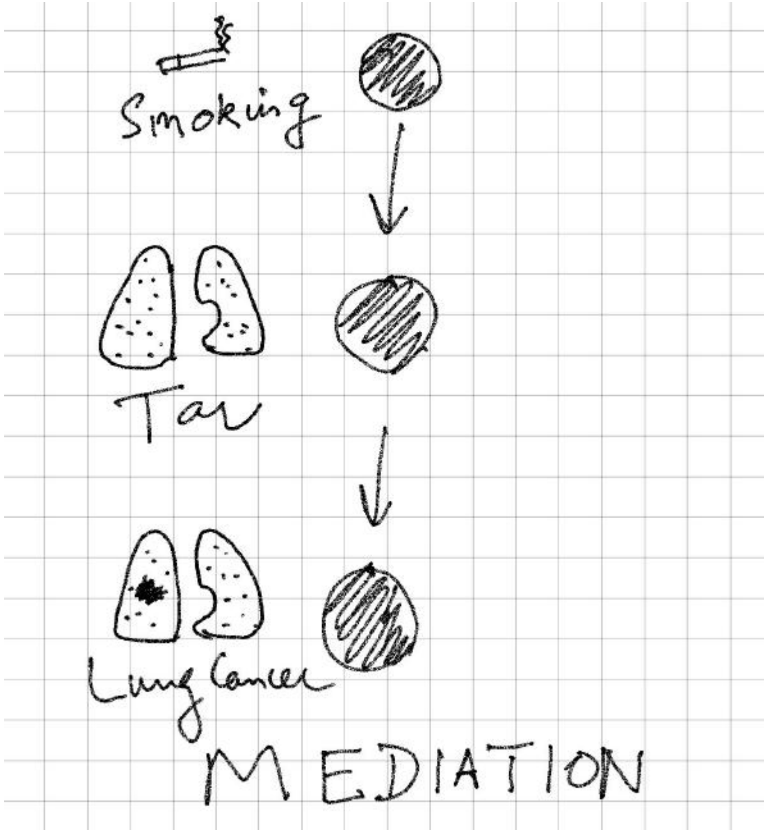


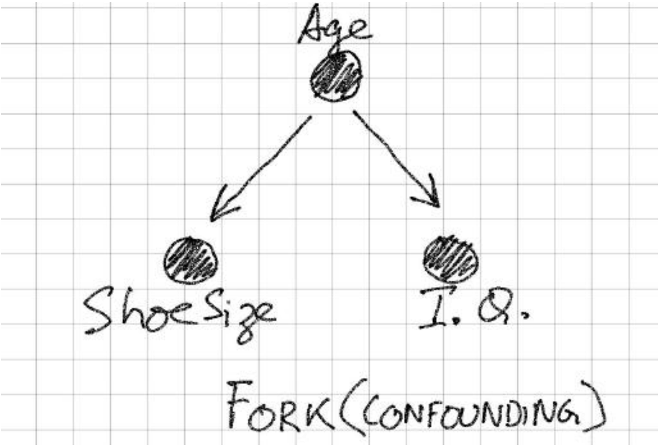


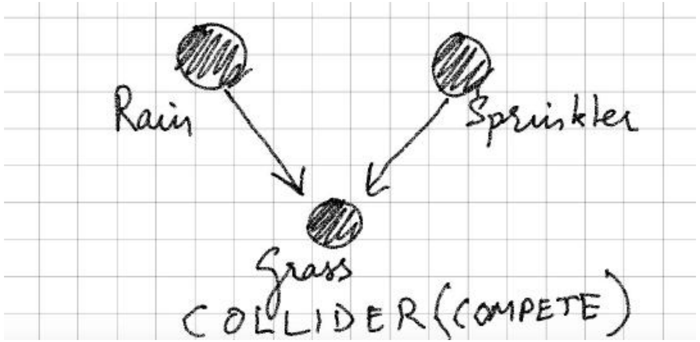
Another Example



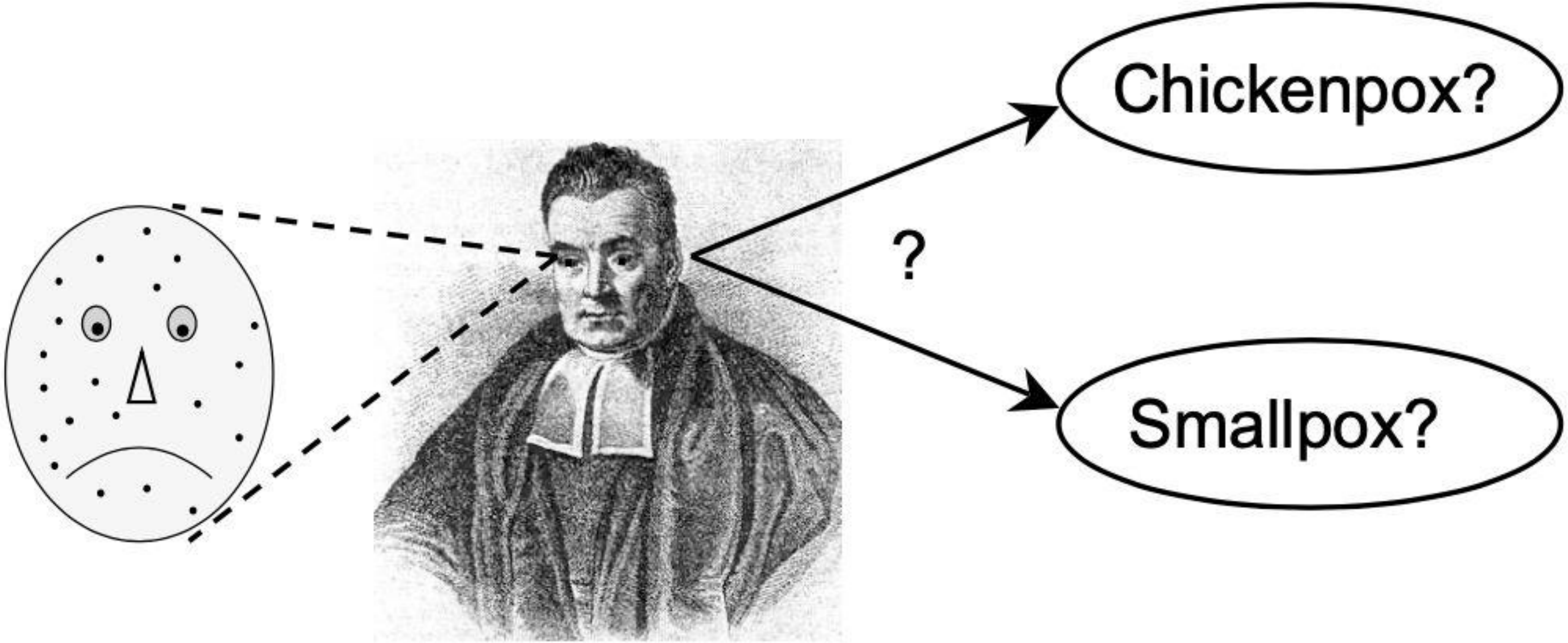
World of Conditional Probabilities







Bayes Rule



We know,

$$p(\text{spots}|\text{smallpox}) = 0.9.$$

$$p(\text{spots}|\text{chickenpox}) = 0.8.$$


When You Hear Hooves, Think Horse,
Not Zebra

Likelihood

$$p(\text{spots}|\text{smallpox}) = 0.9.$$

Likelihood of smallpox = probability of spots given smallpox

$$p(\text{spots}|\text{chickenpox}) = 0.8.$$

Likelihood of chickenpox = probability of spots given smallpox

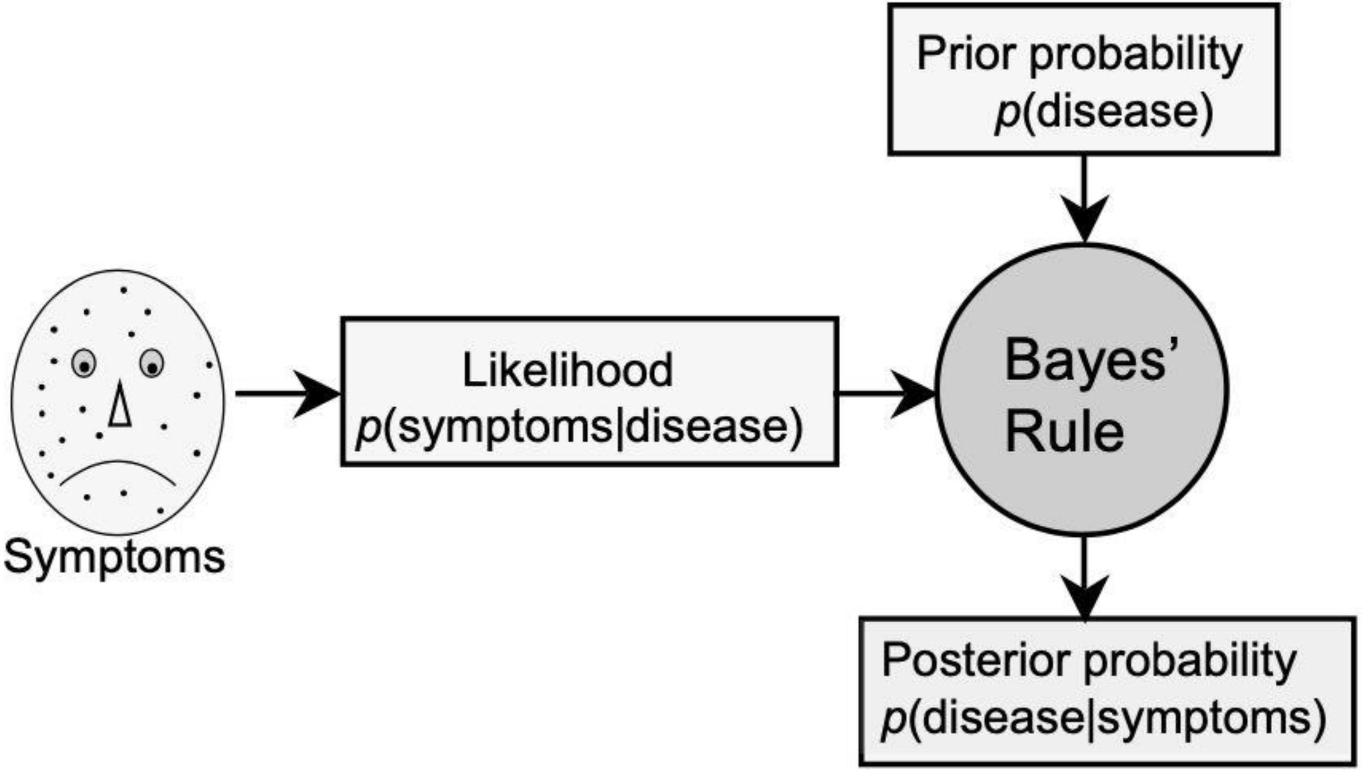
Beware the confusion in language!

Maximum Likelihood Estimate

$$p(\text{spots}|\text{smallpox}) = 0.9. \quad > \quad p(\text{spots}|\text{chickenpox}) = 0.8.$$

Statistical models that work by maximizing the value of likelihood is known as maximum likelihood estimate (MLE)

Bayes Rule in Machine Learning

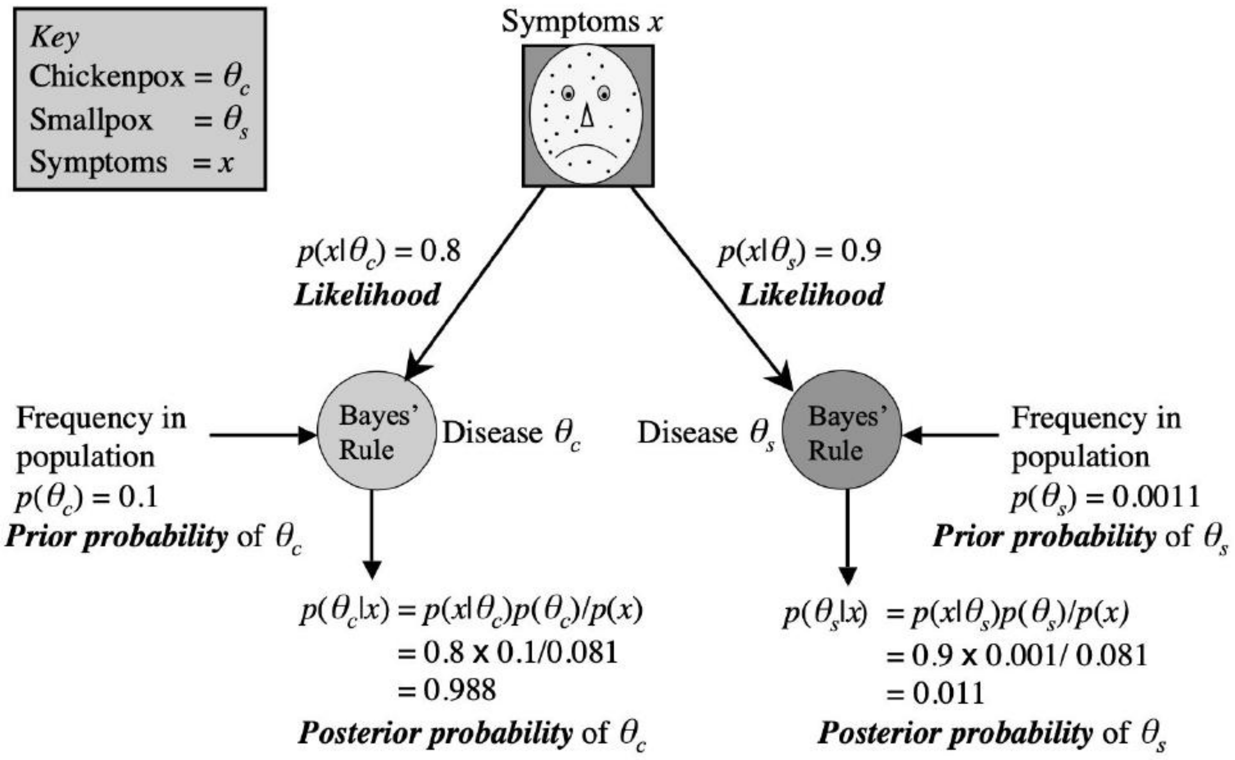


Bayes Rule

$$p(\text{smallpox}|\text{spots}) = \frac{p(\text{spots}|\text{smallpox}) \times p(\text{smallpox})}{p(\text{spots})}.$$

```
# likelihood = prob of spots given smallpox.
pSpotsGSmallpox <- 0.9;
# prior = prob of smallpox.
pSmallpox <- 0.001;
# marginal likelihood = prob of spots.
pSpots <- 0.081;
# find posterior = prob of smallpox given spots.
pSmallpoxGSpots = pSpotsGSmallpox * pSmallpox / pSpots;
#print
s <- sprintf("pSmallpoxGSpots = %.3f", pSmallpoxGSpots)
print(s)
# Output:   pSmallpoxGSpots = 0.011
```

Bayesian Inference



Maximum a Posteriori (MAP) Estimate

Smallpox $$p(\theta_s|x) = \frac{p(x|\theta_s) \times p(\theta_s)}{p(x)}.$$

Chickenpox $$p(\theta_c|x) = \frac{p(x|\theta_c) \times p(\theta_c)}{p(x)}.$$

Succinct Notation

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

$$p(\text{hypothesis}|\text{data}) \quad = \quad \frac{p(\text{data}|\text{hypothesis}) \times p(\text{hypothesis})}{p(\text{data})}$$

Maximum a Posteriori (MAP) Estimate

Smallpox

$$p(\theta_s|x) \quad = \quad \frac{p(x|\theta_s) \times p(\theta_s)}{p(x)}.$$

Chickenpox

$$p(\theta_c|x) \quad = \quad \frac{p(x|\theta_c) \times p(\theta_c)}{p(x)}.$$

$$R_{post} \quad = \quad \frac{p(x|\theta_c)}{p(x|\theta_s)} \times \frac{p(\theta_c)}{p(\theta_s)}.$$

What has cancelled out?

Bayes Factor

$$R_{post} = \frac{p(x|\theta_c)}{p(x|\theta_s)} \times \frac{p(\theta_c)}{p(\theta_s)}.$$

Posterior Odds

Likelihood Ratio
(Bayes Factor)

Prior Odds

The cancelled out term is also called Marginal Likelihood or Evidence

Model Selection

$$posterior\ \ odds = Bayes\ \ factor \times\ prior\ \ odds.$$

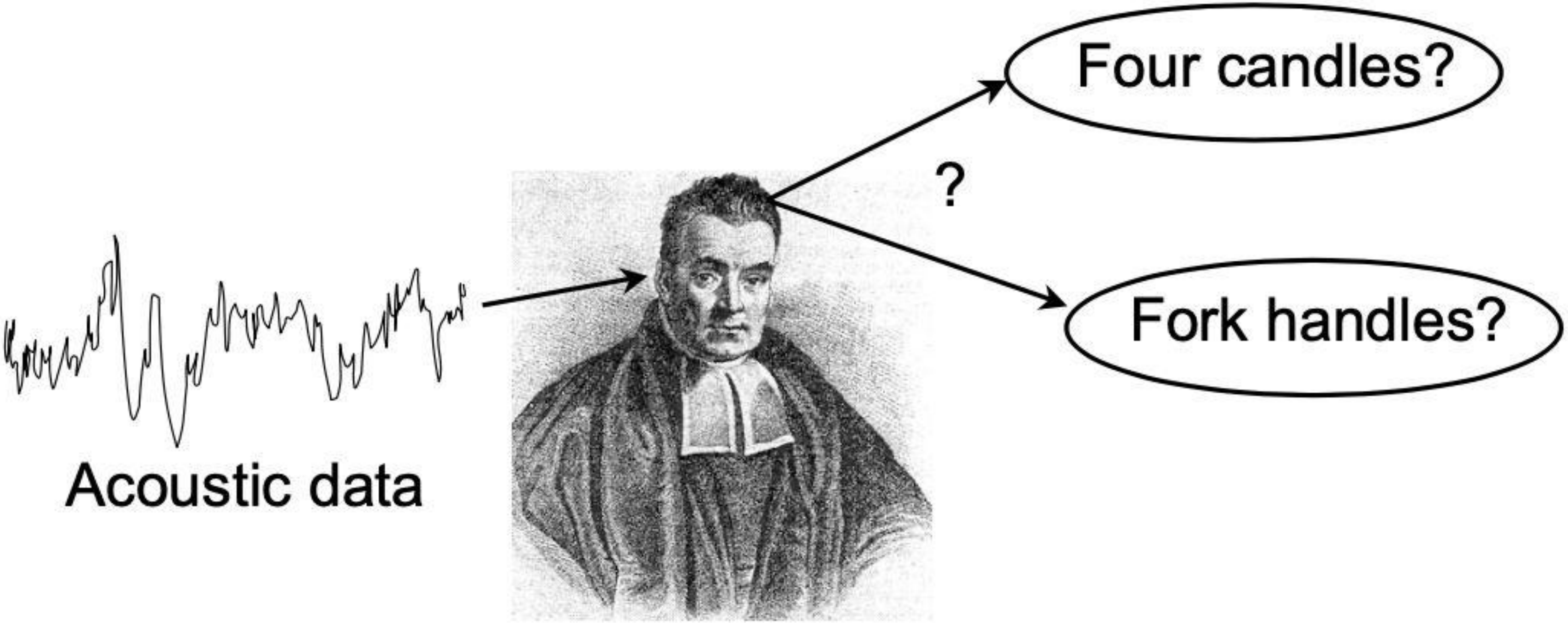
$$R_{post} = \frac{0.80}{0.90} \times \frac{0.1}{0.001} = 88.9.$$

Thumb rule: A posterior odds greater than 3 or less than 1/3 is considered substantial difference between the probabilities of the two models.

When is Bayes Rule Useful?

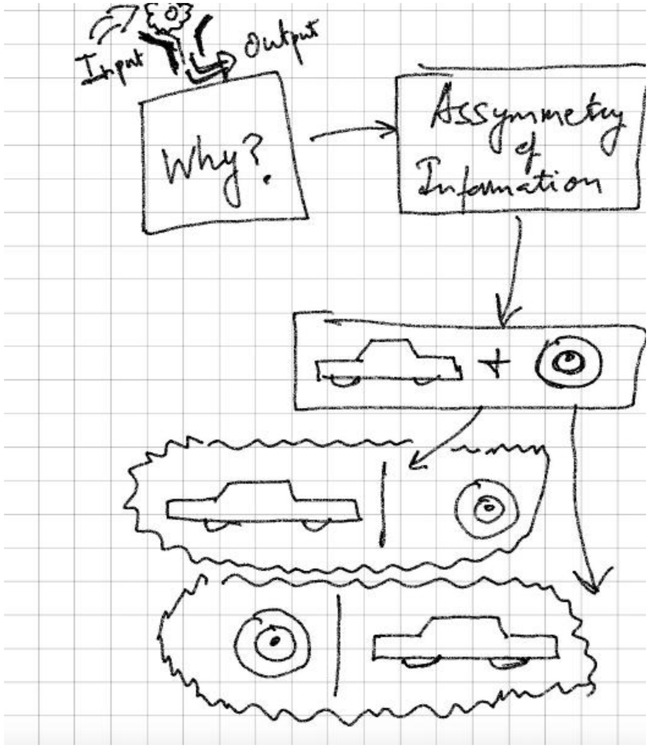
https://www.youtube.com/watch?v=pV1IP4N9ajg

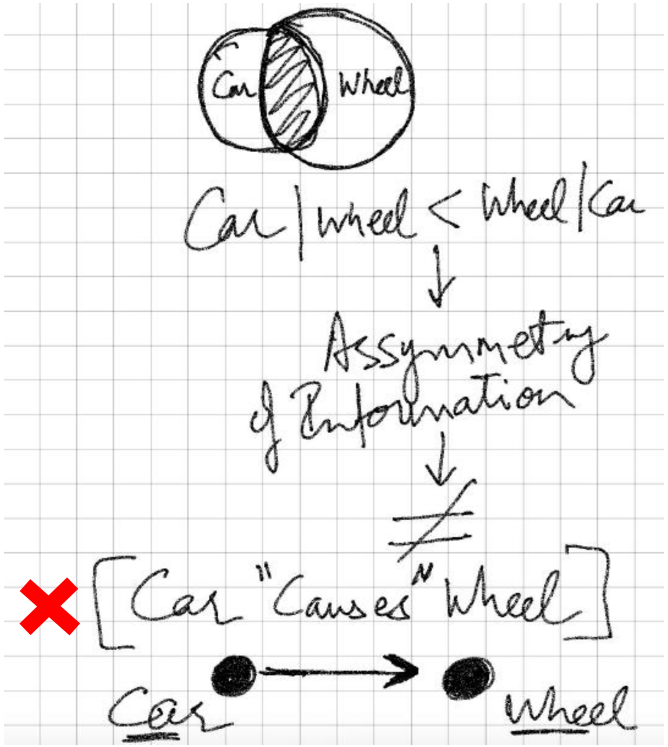
ML with Bayesian Models

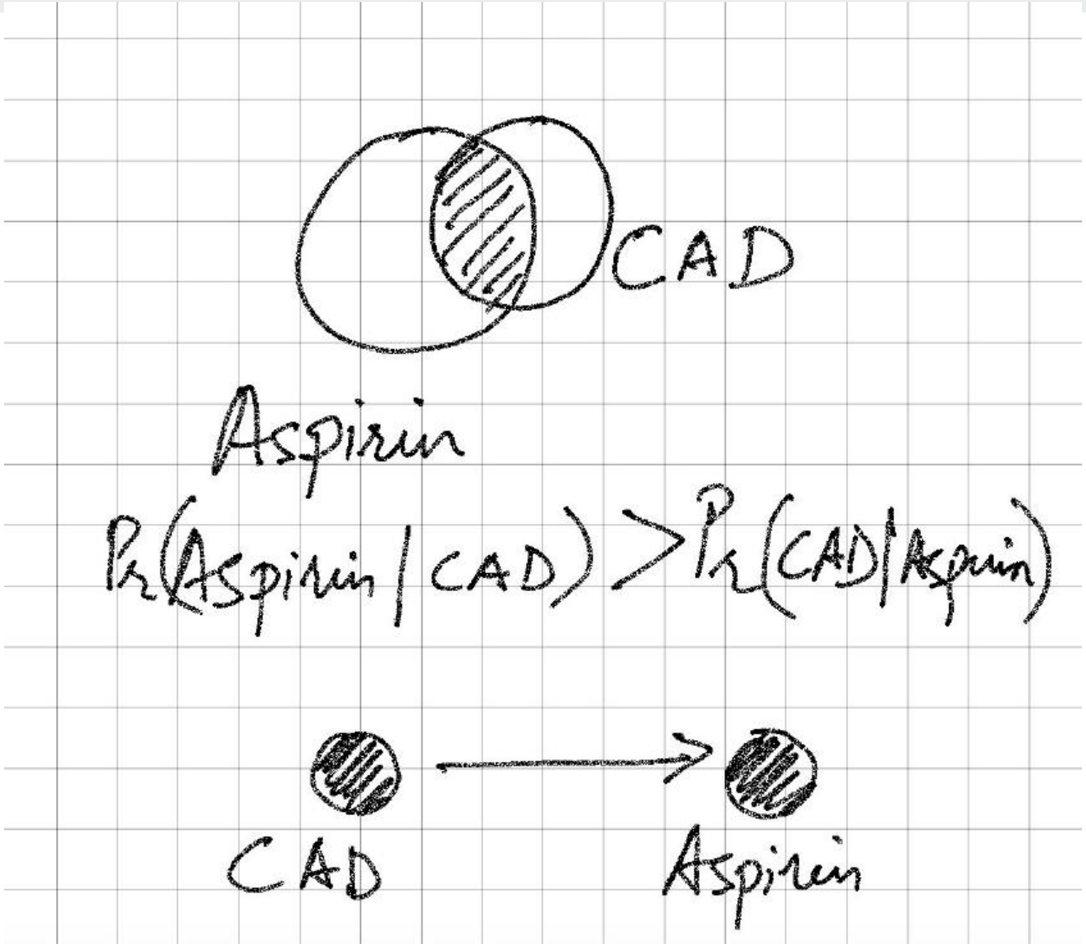


BAYESIAN NETWORKS

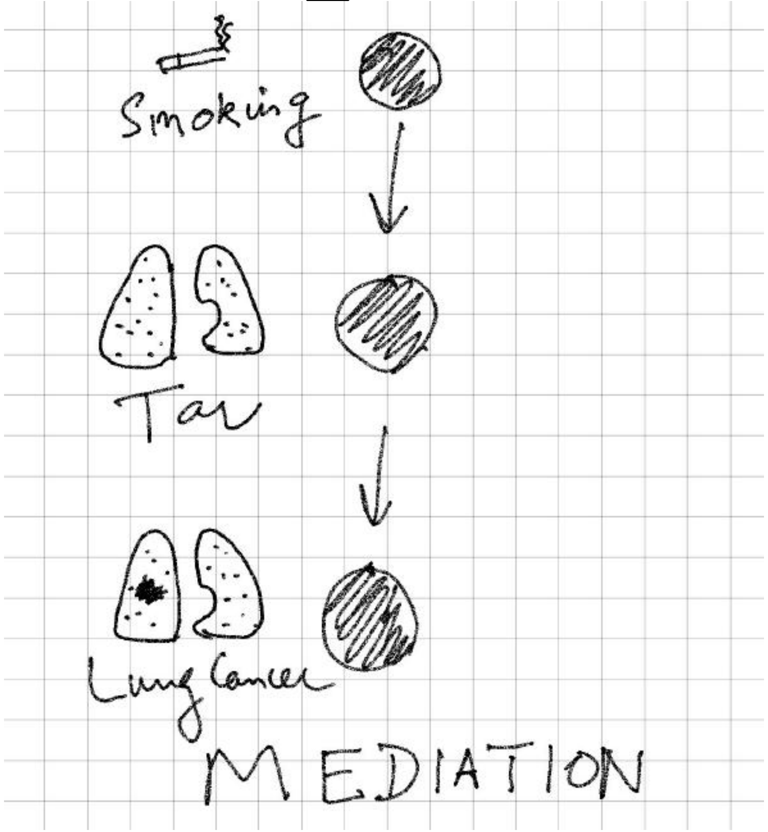
Bayesian Networks

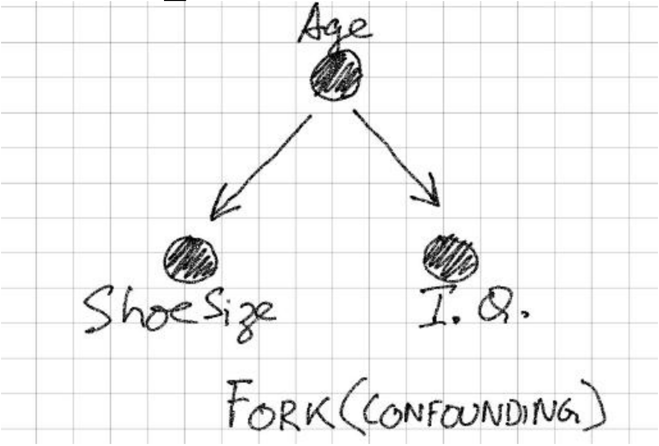


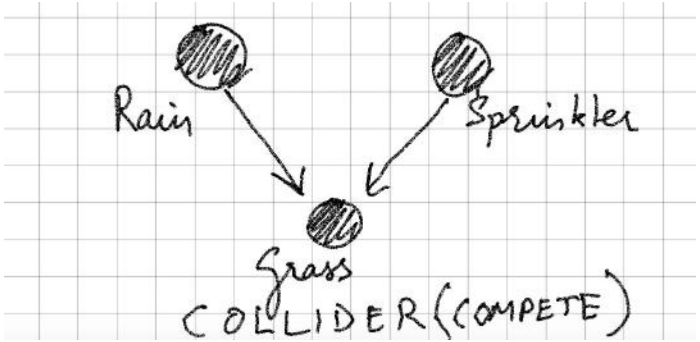



CAD
Aspirin
Pr(Aspirin | CAD) > Pr(CAD|Aspirin)
CAD
Aspirin

Building Blocks of Bayesian Nets



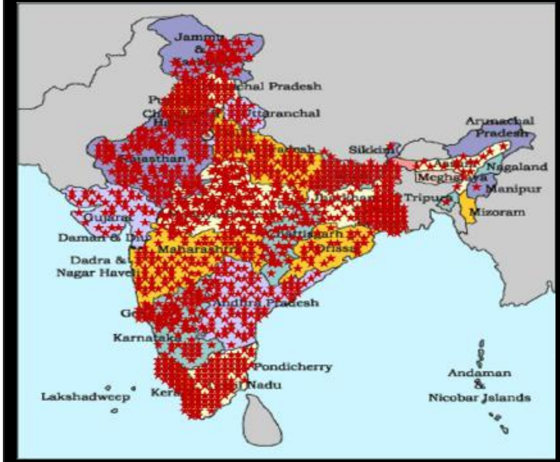


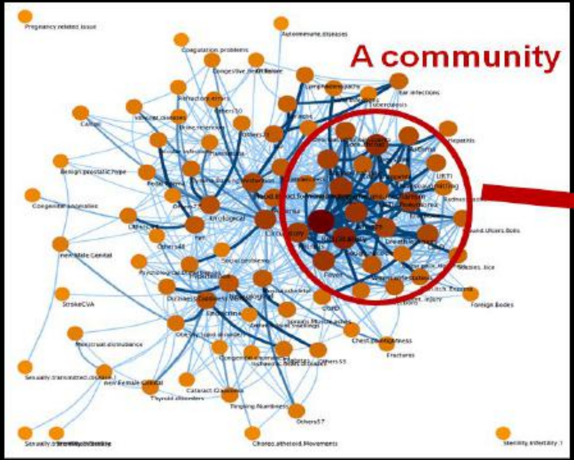


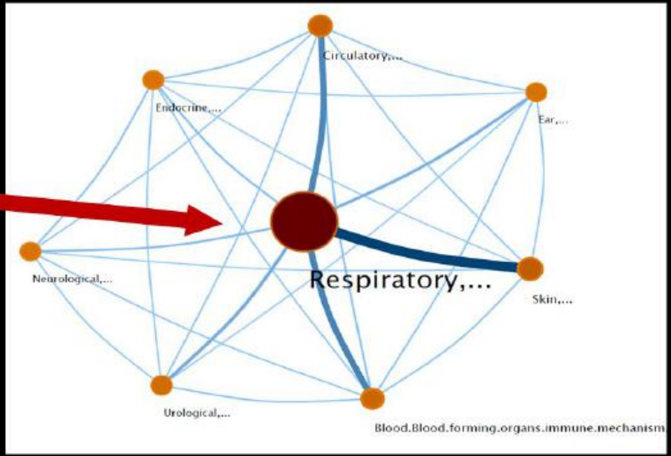
Prevalence of Symptoms in a Single Indian Healthcare Day on a Nationwide Scale

One day point prevalence study of symptoms in 2,04,912 patients across India

Chest Research Foundation, Pune, India





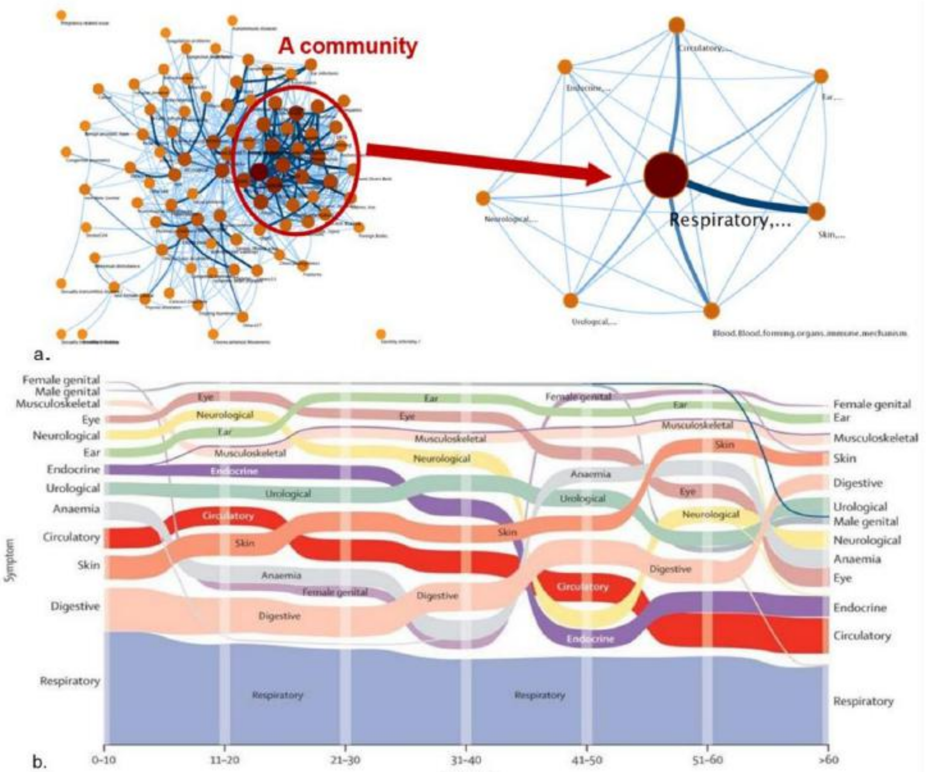


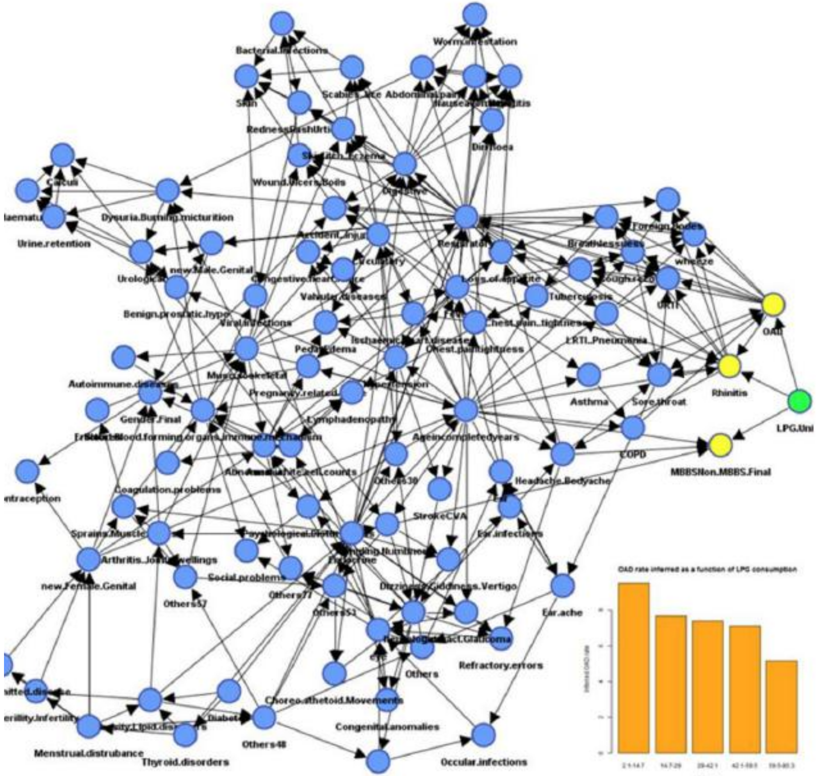
Networks approach: many diseases happen together, connectivity differs across age

Lancet Global Health, Dec 2015

Sundeep Salvi et al. Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: a 1-day point prevalence study (the POSEIDON study), *The Lancet Global Health,* Volume 3, Issue 12,
2015, Pages e776-e784, https://doi.org/10.1016/S2214-109X(15)00152-7.

Associations to Decisions





Sundeep Salvi et al. Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: a 1-day point prevalence study (the POSEIDON study), *The Lancet Global Health,* Volume 3, Issue 12, 2015, Pages e776-e784, https://doi.org/10.1016/S2214-109X(15)00152-7.

Case Study II: AI for Reducing Health Inequities

- The richest American men live **15 years** longer than the poorest American men[1]
- The richest American women live **10 years** longer than the poorest American women[1]
- **Healthcare inequities** impose an estimated burden of **$300 billion per year** in the United States
- Longevity is the **sum-total of influences** on the healthcare
- Hence **longevity-gap** is a complex **socio-demographic** challenge
- Key motivation: **learn policy** for **mitigating** the longevity-gap using explainable AI and release it to public, policymakers

Sethi T, S. Maheshwari, A. Mittal, S. Chugh. Learning to Address Health Inequality in the United States with a Bayesian Decision Network. Proceedings of the AAAI Conference on Artificial Intelligence 33, 710-717. DOI: https://doi.org/10.1609/aaai.v33i01.3301710 c

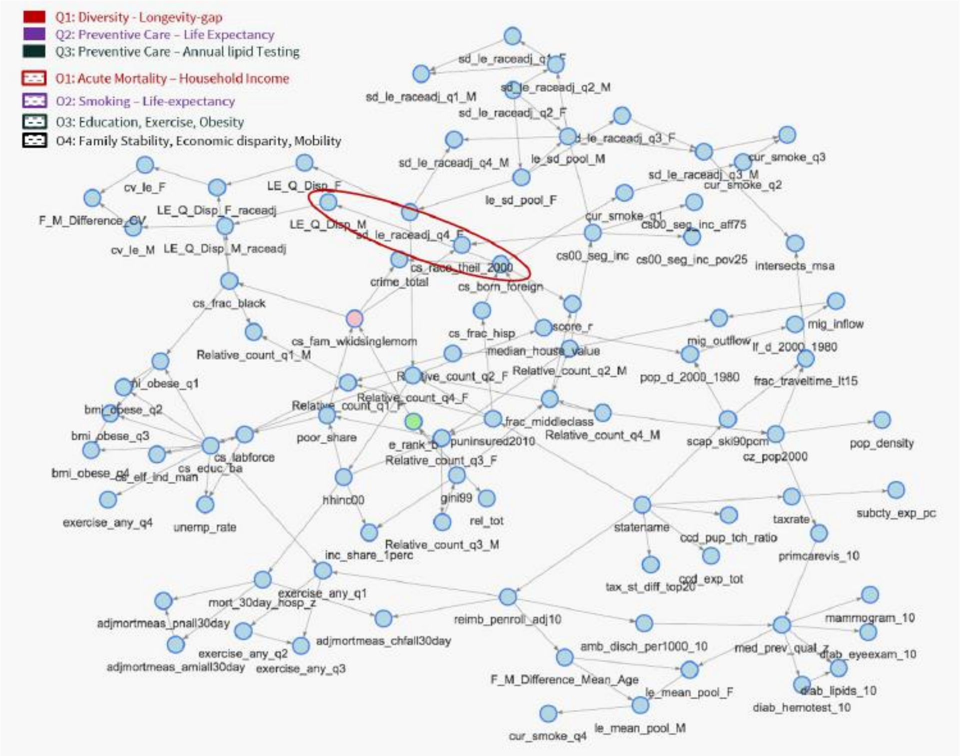
Key Messages

We used data from: **Mortality** (Census), **Healthcare** Indices (Dartmouth Atlas), **Health-behaviors** (CDC, BRFSS), **Education** (K-12 and Post Secondary), **Demographics** (e.g. race, ethnicity, diversity, gender ratio etc), **Socioeconomics** (e.g. Gini index, Poverty rate, Income segregation, Social Mobility), **Social Cohesion.** (e.g. Social Capital Index, Religious adherents)**, Labor market conditions and Taxation.** (e.g. unemployment, manufacturing sector).

Key Messages

1. **Social.** Diversity mitigates health inequality in the US. Counties with higher diversity are 38% less likely to have a longevity gap between the rich and the poor.
2. **Preventive Care**. Counties with high quality primary care services (not just visits but investigations) have a 43% increase in the probability of living beyond 85 years in females (corresponding 30% increase for males.)
3. **Clinical**. Acute mortality (30-day Hospital Mortality Index) is 30% less in Counties with household income in the highest segment.
4. **Usual suspects**. Smoking, Education, Exercise as expected to be key influencers of longevity.
5. **Socio-demographic**. Family stability decreases crime rate, increases upward social mobility across economic tiers and indirectly decreases Gini disparity in counties.

Diversity Mitigates Longevity Gap



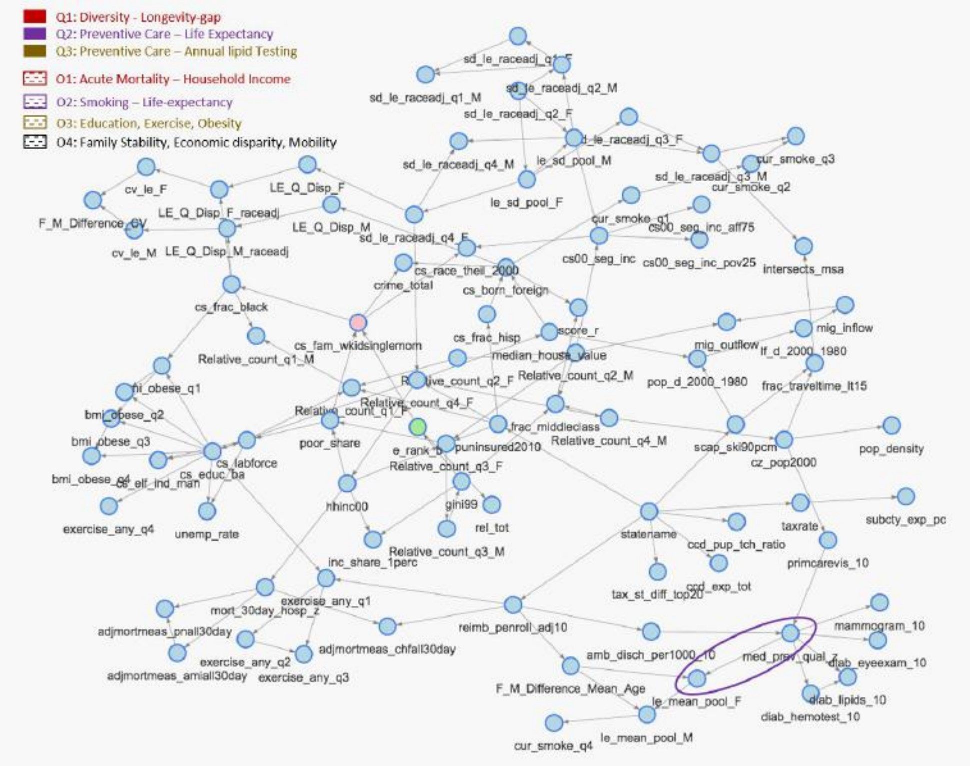
A 1. Diversity

Counties with higher diversity are 38% less likely to have a longevity gap between the rich and the poor.

Likely explanation: Structure indicates that Higher Diversity is Associated with Higher Income, especially in Females, thus improving healthcare services.

Sethi T, S. Maheshwari, A. Mittal, S. Chugh. Learning to Address Health Inequality in the United States with a Bayesian Decision Network. Proceedings of the AAAI Conference on Artificial Intelligence 33, 710-717. DOI: https://doi.org/10.1609/aaai.v33i01.3301710 c

The Impact of Primary Care



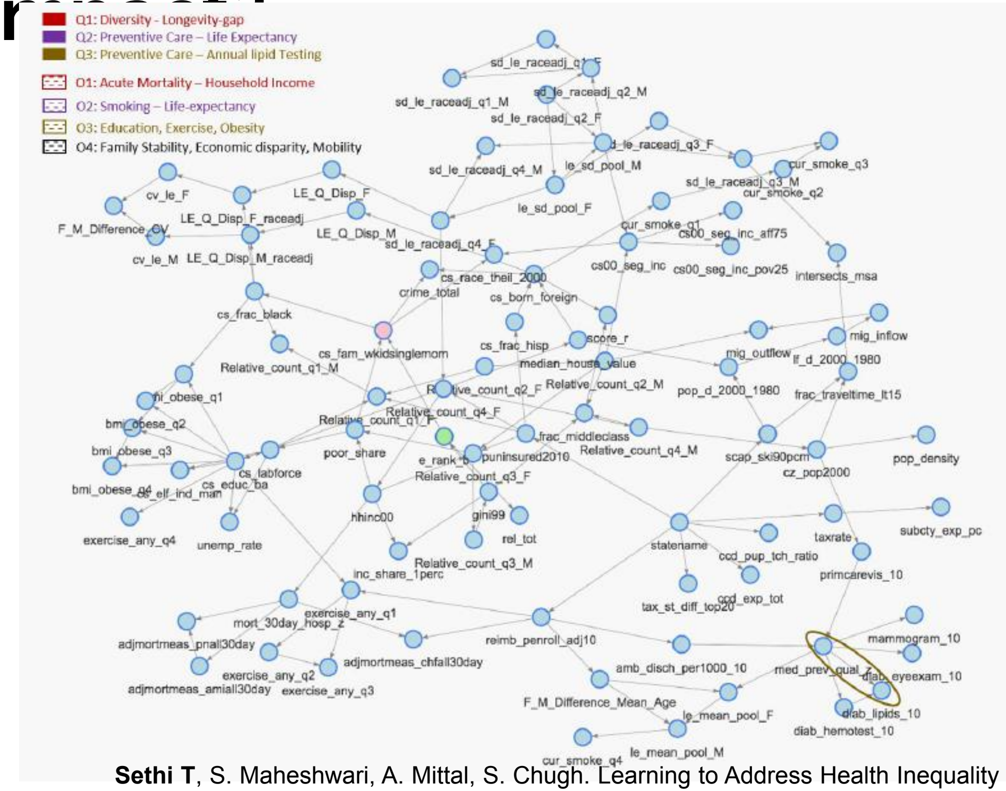
A 2. Quality of Preventive Care

Counties with high quality primary care services have a 43% increase in the probability of living beyond 85 years in females (corresponding 30% increase for males.)

Likely explanation: Self Evident, but previously unquantified in a Joint Model

Sethi T, S. Maheshwari, A. Mittal, S. Chugh. Learning to Address Health Inequality in the United States with a Bayesian Decision Network. Proceedings of the AAAI Conference on Artificial Intelligence 33, 710-717. DOI: https://doi.org/10.1609/aaai.v33i01.3301710 c

Which primary care investigation has most impact?



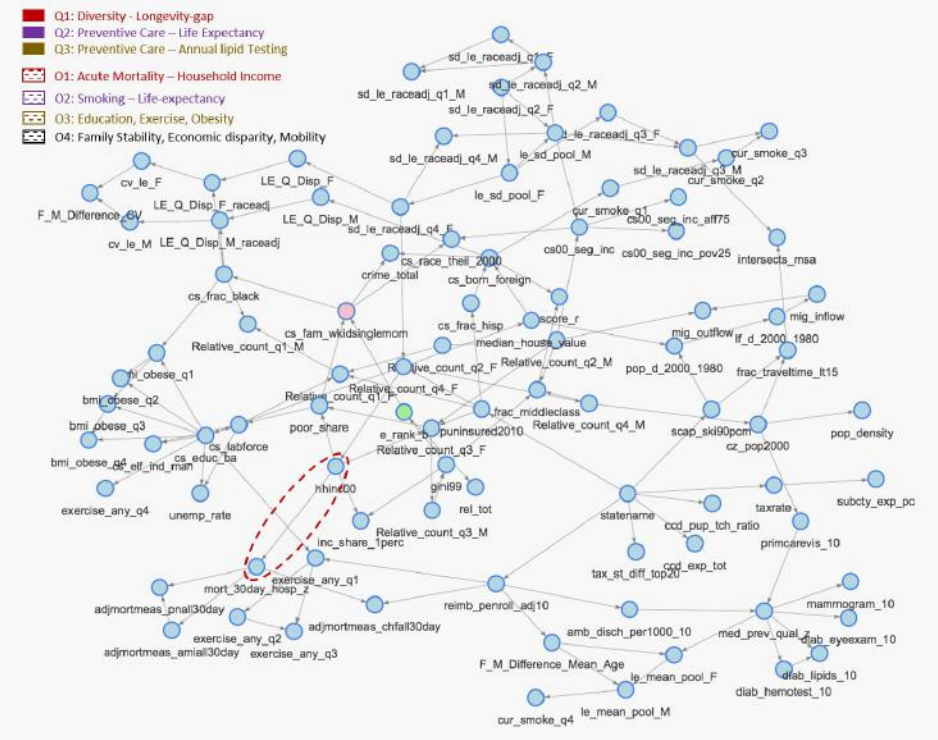
A 3. Annual Lipid Testing esp. in the diabetic population

diab_eyeexam_10	mammogram_10	diab_lipids_10	payoff
[70.2,85.6]	[68.2,86.1]	[79.3,92.9]	0.52
[70.2,85.6]	[68.2,86.1]	[65.6,79.3)	0.48
[62.2,70.2)	[68.2,86.1]	[79.3,92.9]	0.44
[70.2,85.6]	[59.5,68.2)	[79.3,92.9]	0.40
[62.2,70.2)	[68.2,86.1]	[65.6,79.3)	0.26
[70.2,85.6]	[59.5,68.2)	[65.6,79.3)	0.22
[42.4,62.2)	[68.2,86.1]	[79.3,92.9]	0.20
[62.2,70.2)	[59.5,68.2)	[79.3,92.9]	0.12
[70.2,85.6]	[31.1,59.5)	[79.3,92.9]	0.11
[42.4,62.2)	[68.2,86.1]	[65.6,79.3)	0.08

Likely explanation: Diabetics are at highest risk of cardiovascular mortality

Sethi T, S. Maheshwari, A. Mittal, S. Chugh. Learning to Address Health Inequality in the United States with a Bayesian Decision Network. Proceedings of the AAAI Conference on Artificial Intelligence 33, 710-717. DOI: https://doi.org/10.1609/aaai.v33i01.3301710 c

Income and Acute Mortality

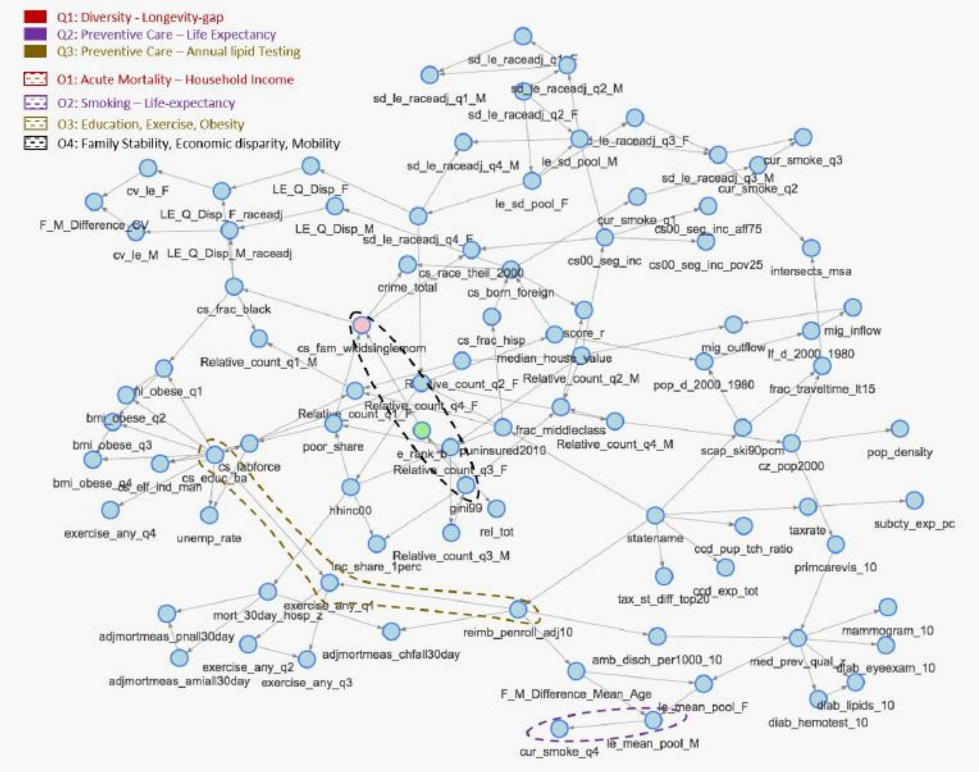


O 1.

ACUTE MORTALITY (30-DAY HOSPITAL MORTALITY INDEX > 0.92) IS 30% LESS IN COUNTIES WITH HOUSEHOLD INCOME IN THE HIGHEST SEGMENT.

HIGHEST CONTRIBUTOR TO ACUTE MORTALITY IN LOWER INCOME HOUSEHOLDS IS PNEUMONIA

Smoking, Exercise, Health Behaviors

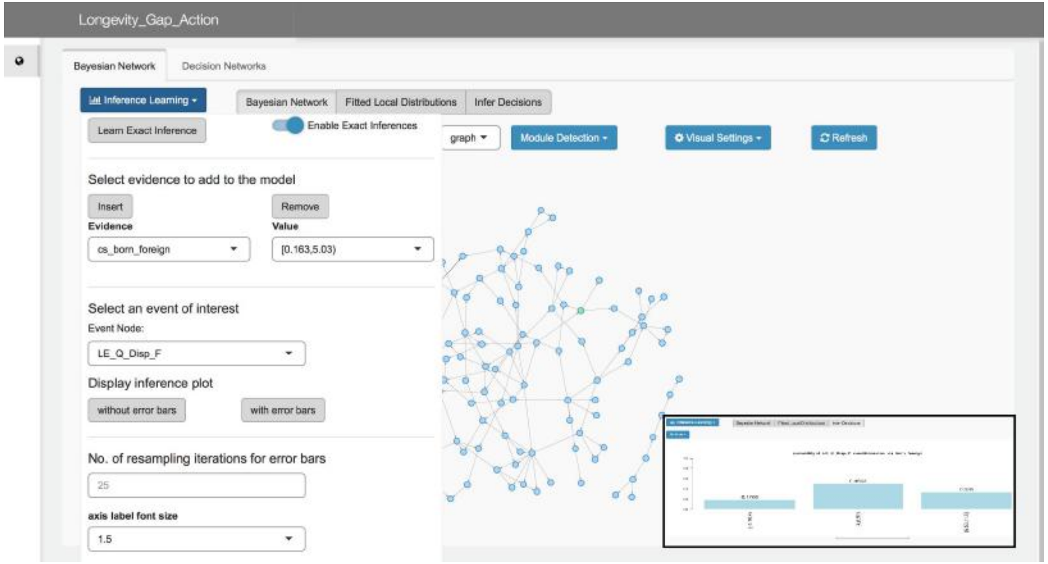


Usual suspects (O2-O4)

- High life-expectancy in males (81.9 - 85 years) makes it 33% more likely for **smoking** to be in the lowest stratum in the county.
- Counties with high proportion of exercise have 19% less hospitalization rates
- Counties with lower family stability are 40% more likely to have lower social mobility

Tavpritesh Sethi, Anant Mittal, Shubham Maheshwari, Samarth Chugh. *Learning to Address Health Inequality in the United States with a Bayesian Decision Network.* https://arxiv.org/abs/1809.09215 **Accepted for publication in the Thirty-third AAAI conference in Artificial Intelligence**, AAAI-2019

Deploy your XAI models as Web applications



https://github.com/SAFE-ICU/Longevity_Gap_Action

Key Messages

1. **Social.** Diversity mitigates health inequality in the US. Counties with higher diversity are 38% less likely to have a longevity gap between the rich and the poor.
2. **Preventive Care**. Counties with high quality primary care services (not just visits but investigations) have a 43% increase in the probability of living beyond 85 years in females (corresponding 30% increase for males.)
3. **Clinical**. Acute mortality (30-day Hospital Mortality Index) is 30% less in Counties with household income in the highest segment.
4. **Usual suspects**. Smoking, Education, Exercise as expected to be key influencers of longevity.
5. **Socio-demographic**. Family stability decreases crime rate, increases upward social mobility across economic tiers and indirectly decreases Gini disparity in counties.

Tavpritesh Sethi, Anant Mittal, Shubham Maheshwari, Samarth Chugh. *Learning to Address Health Inequality in the United States with a Bayesian Decision Network.* https://arxiv.org/abs/1809.09215 **Accepted for publication in the Thirty-third AAAI conference in Artificial Intelligence**, AAAI-2019

India in AMR

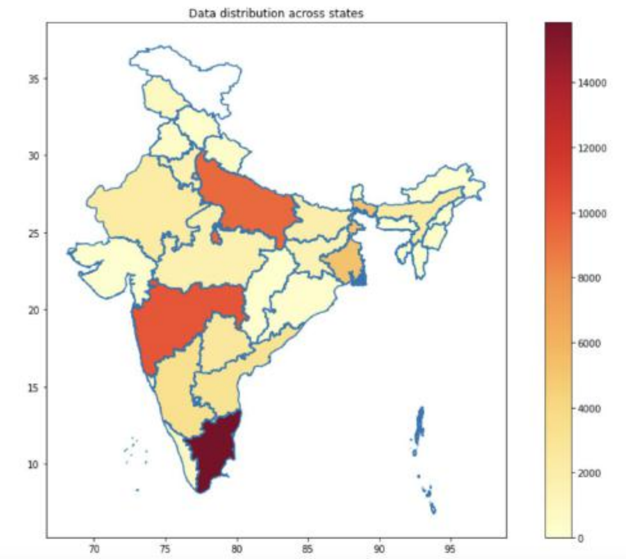
THE LANCET Regional Health
Southeast Asia

This journal · Journals · Publish · Clinical · Global health · Multimedia · Events · About

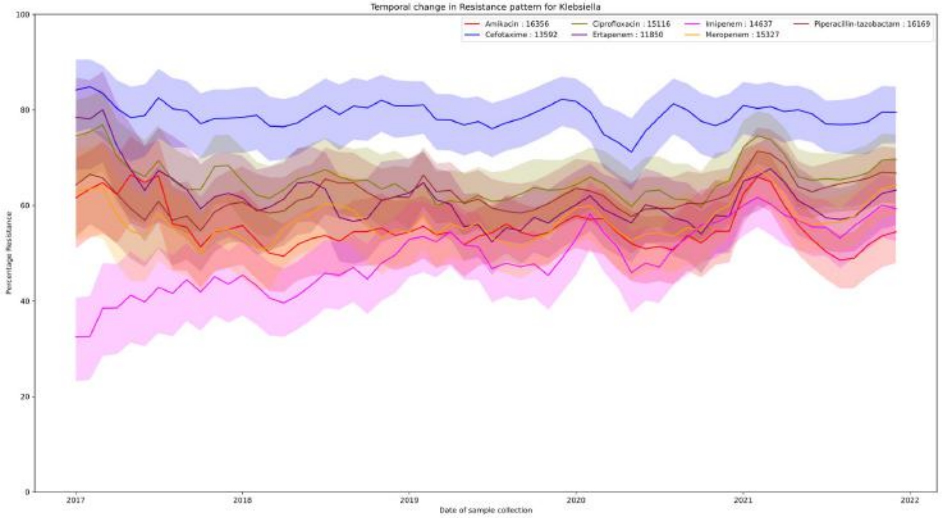
ARTICLES · Volume 26, 100412, July 2024 · *Open Access* · Download Full Issue

Emerging trends in antimicrobial resistance in bloodstream infections: multicentric longitudinal study in India (2017–2022)

Jasmine Kaur [a,b,c] · Harpreet Singh [c] · Tavpritesh Sethi [a,b]



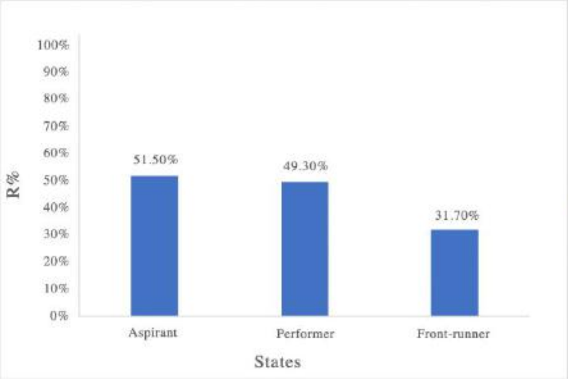
Klebsiella Sepsis

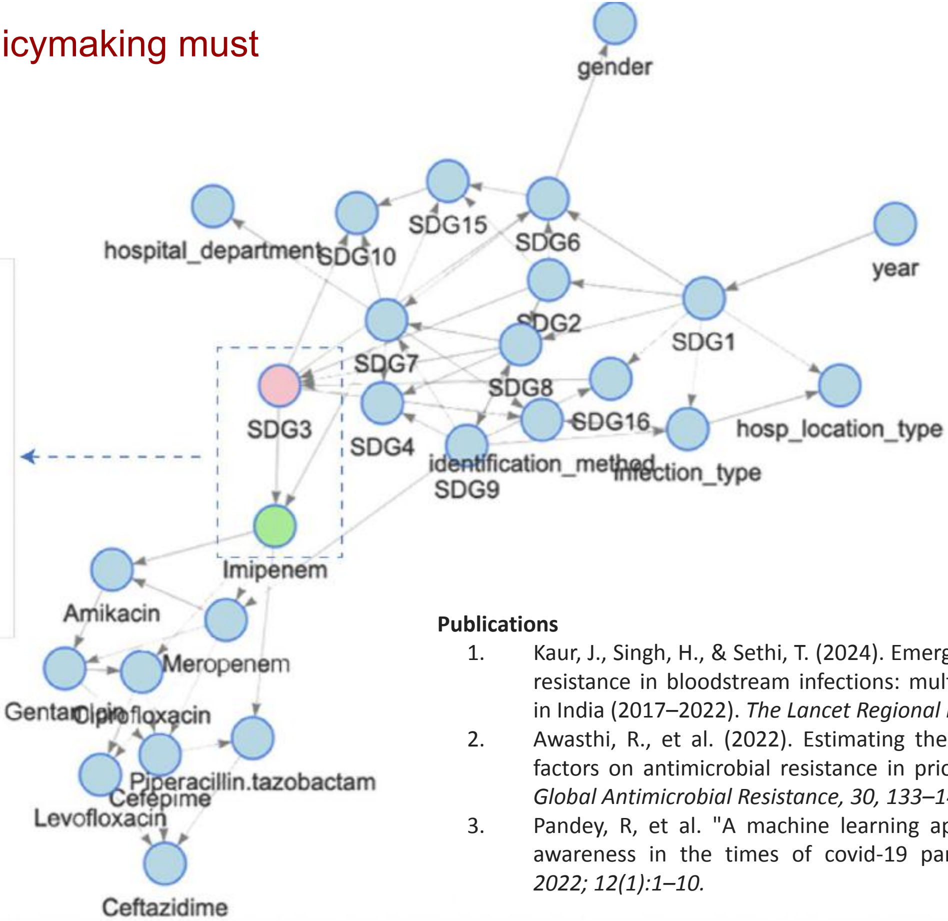


0.25% monthly average for increase in Imipenem resistance

AI for Understanding Impact on SDG Achievement

Key Takeaway: Evidence-based policymaking must leverage AI for tracking progress.





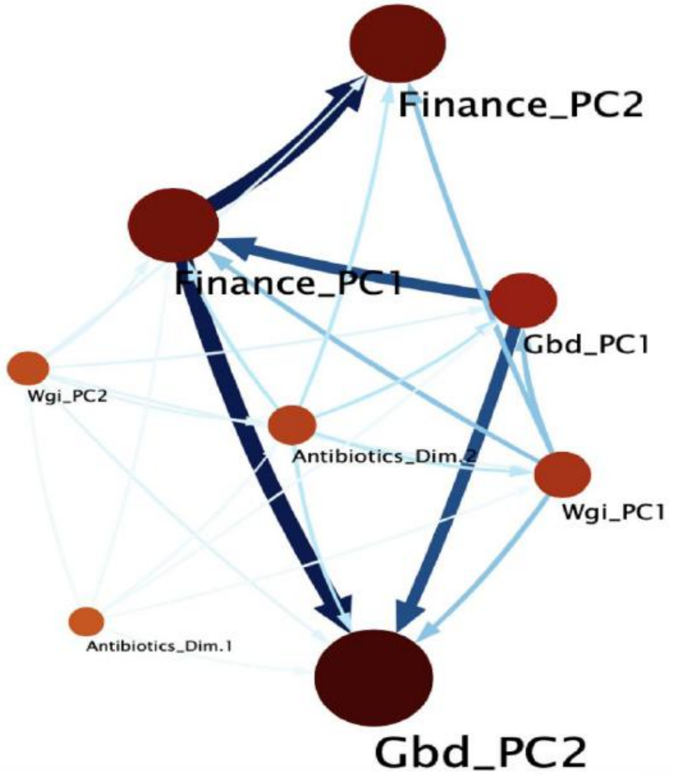
Publications

1. Kaur, J., Singh, H., & Sethi, T. (2024). Emerging trends in antimicrobial resistance in bloodstream infections: multicentric longitudinal study in India (2017–2022). *The Lancet Regional Health-Southeast Asia*, *26*.
2. Awasthi, R., et al. (2022). Estimating the impact of health systems factors on antimicrobial resistance in priority pathogens. *Journal of Global Antimicrobial Resistance, 30, 133–142*.
3. Pandey, R, et al. "A machine learning application for raising wash awareness in the times of covid-19 pandemic". *Scientific reports 2022; 12(1):1–10.*

Systems Indicators and AMR

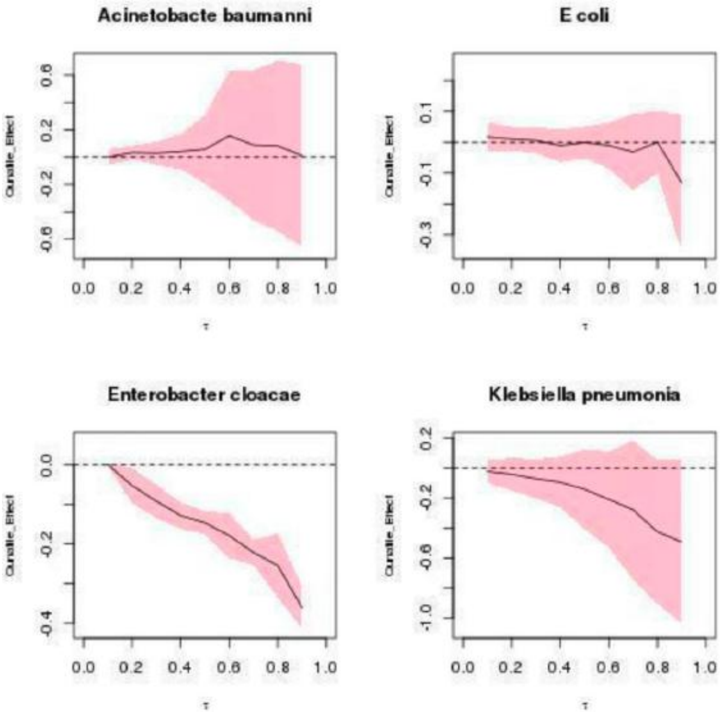
Age and temporal distribution of data

	Women	Men	Overall
Sex	No. (%)	No. (%)	No. (%)
	278 128 (43.88%)	348 136 (54.93%)	633 820
Age group			
0 to 2 years	15 329 (42.83%)	20 012 (55.91%)	35 793
13 to 18 years	6348 (47.06%)	7043 (52.21%)	13 490
19 to 64 years	130 162 (44.04%)	163 201 (55.22%)	295 537
3 to 12 years	12 613 (46.88%)	14 052 (52.22%)	26 907
65 to 84 years	86 561 (41.63%)	119 801 (57.62%)	207 922
85 and over	23 406 (54.29%)	19 364 (44.91%)	43 114
Year			
2004	9433 (46.93%)	10 655 (53.01%)	20 101
2005	10 473 (48.04%)	11 274 (51.71%)	21 801
2006	13 967 (46.65%)	15 876 (53.03%)	29 940
2007	18 264 (45.70%)	21 409 (53.57%)	39 964
2008	16 303 (44.33%)	19 925 (54.18%)	36 773
2009	18 575 (44.55%)	22 514 (54.00%)	41 692
2010	14 476 (44.59%)	17 341 (53.42%)	32 462
2011	11 622 (44.66%)	13 931 (53.53%)	26 023
2012	22 432 (42.16%)	29 477 (55.40%)	53 206
2013	29 962 (42.73%)	38 850 (55.40%)	70 125
2014	30 492 (43.23%)	39 482 (55.98%)	70 529
2015	27 992 (43.21%)	36 104 (55.73%)	64 785
2016	29 744 (42.74%)	39 290 (56.45%)	69 598
2017	24 393 (42.93%)	32 008 (56.33%)	56 821

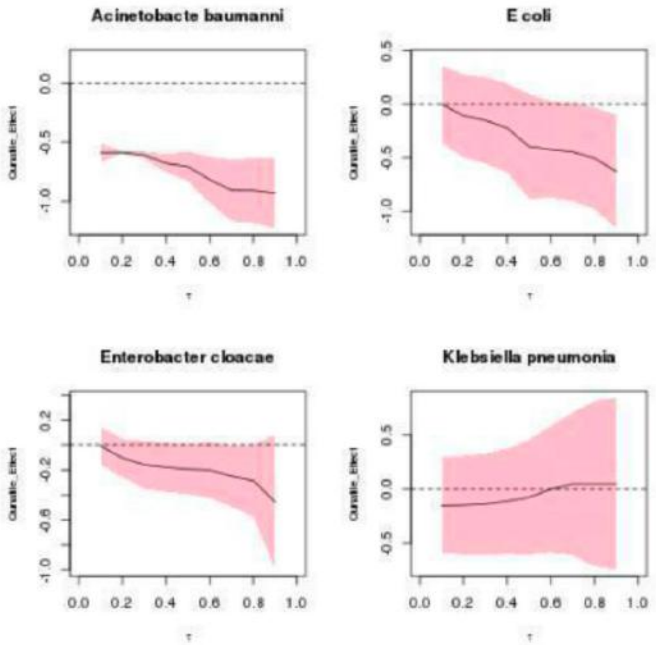


Awasthi R, Rakholia V, Agrawal S, Dhingra LS, Nagori A, Kaur H, Sethi T. Estimating the impact of health systems factors on antimicrobial resistance in priority pathogens. J Glob Antimicrob Resist. 2022 Sep;30:133-142.

Impact Calculation: Counterfactual Analysis



Ceftriaxone High Income



Ceftriaxone Middle Income

Key Steps in Building a BN Model

- Learn Structure
  - Score Based
  - Constraint Based
- Validate Structure
  - Bootstrapping
  - Domain Based Sanitization
- Conduct Inference
  - Exact Inference
  - Approximate Inference