Attendance

CS365
Foundations of Data Science

Lecture 1
(1/18/23)

Charalampos E. Tsourakakis
ctsourak@bu.edu

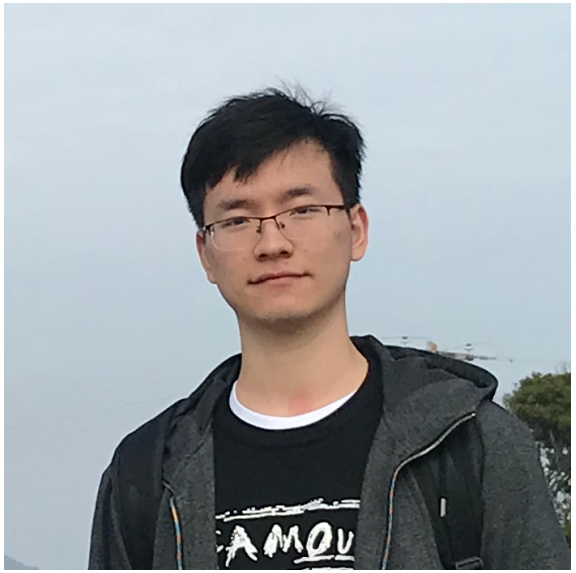
Logistics

- When: Tue, Thu 5pm-6.15pm
- Where: CAS 211
- Prof: Babis Tsourakakis
- Email: ctsourak@bu.edu
- Office hours: Tuesday 11-noon (CDS 912)
  Thursday 10.30-11.30am

Teaching Fellow

- Mr. Tianyi Chen
- Email: ctony@bu.edu
- Teaching Lab on each Monday
  - Attend your own session
- Office hours
  a) Wed 2:00-3:30 pm (CDS 908)

  b) Fri 10:00-11:30 am



Logistics

- Web site: https://tsourakakis.com/cs365-foundations-of-data-science-spring24/
- Piazza: https://piazza.com/class/lrgnqll42av781
  - If you are not already registered, send me an email.
- Gradescope: https://www.gradescope.com/courses/714478

Sign up on gradescope - Access code will become available on Piazza

Logistics

- **Project (30%)**
- **Homeworks (30%)**
- **Test 1 (20%)**
  - Tentative Tuesday 2/22
- **Test 2 (20%)**
  - Tentative 4/2

Logistics - Project

5 ~ 6 questions to Answer from Dataset

Project (30%) - groups up to 2 people

1. Project proposal (10%) **Deadline 2/6**
   - Define the problem you will focus on
   - What methods do you plan to explore
   - Datasets you will use
   - If it is a group of two (2) students, explain the work of each.
2. Milestone (10%) **Deadline 3/5**
   - 3-page report (mini-version of the final report)
   - Code in Colab

Logistics - Project

3. Final report (80%) **Deadline 4/23**

- Delivarable 1:
  Written in Latex using a template we will provide you (8 page limit)
  - Introduction
  - Related work
  - Resources you used, including generative AI
  - Methods
  - Experimental results
  - Conclusion
- Deliverable 2:
  - Code the reproduces your findings
- Delivarable 3:
  - Assigned mini kaggle-like project (details to follow)

Project ideas

- Kaggle https://www.kaggle.com/
- UCI datasets https://archive.ics.uci.edu/datasets
- Neurips
- KDD

Other conferences of interest include ICML, ICLR, AAAI, ECML-PKDD and WebConf.

Programming Languages

- I will assume that you are familiar with Python.

https://docs.python.org/3/

- Another language I recommend for this class is Julia:

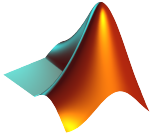
http://julialang.org/

Programming Languages

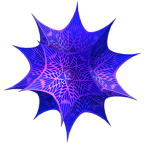
Other widely used languages in data science that I won't be using, but you are welcome to use include:

- R https://www.r-project.org/

- Matlab https://www.mathworks.com/products/matlab.html (BU offers a licence)



- Mathematica https://www.wolfram.com/mathematica/ (BU offers a licence)



Overview of Course

- We focus on the *mathematical* and *algorithmic* foundations of data science:
  - PART 1 : Theoretical Background
    - Probability
    - Information theory
    - Linear algebra
    - Calculus
    - Algorithm design
    - Optimization
  - PART 2: Data Science
    - Dimensionality reduction
    - Time-series analysis
    - Machine Learning and Data Analysis (clustering, classification)
    - Network analysis

Prerequisites

Projection
vector space

Students taking this class must have taken:

- CS 112
- CS 131 (MA293)
- CS 132 (MA242)
- and CS 237 (MA581) or equivalent.

The consent of the instructor is necessary to take the class. Otherwise, you won't get a grade during HWs, exams etc.

- CS 330 is *highly* recommended but not mandatory requirements as the previous.

I assume that you are familiar with at least one programming language.

Learning outcomes

- Students who successfully complete this course will have built a *theoretical* background that will allow them to take more advanced data-intensive classes, such as
  - Data Science,
  - Machine Learning,
  - Data Mining
  - Deep Learning

- Homeworks will improve your proficiency in data acquisition, manipulation and analysis as well.

Academic conduct

- Academic standards and the code of academic conduct are taken very seriously by our university, by the College of Arts and Sciences, and by the Department of Computer Science.

- Course participants must adhere to the CAS Academic Conduct Code – please take the time to review this document if you are unfamiliar with its contents.

- https://www.bu.edu/academics/policies/

Collaboration policy

- You are encouraged to collaborate with one another in studying the textbook and lecture material.

- However, homeworks should be worked out and written by yourself.
  - You may get help from the Professor and the TF for a specific problems, but do not expect them to do it for you, however.

- **Piazza**: you are **encouraged** to use Piazza to ask questions on any topic; course material, homework and lab problems, logistics.

  - We strongly prefer communicating via Piazza over emails to the course staff, unless it is a sensitive personal issue.

Textbook

- There will be assigned readings from the following books that are available online (click for the pdf)

  - Machine Learning: A Probabilistic Perspective by Kevin Murphy
  - Mathematics for Machine Learning by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong.
  - Foundations of Data Science by Avrim Blum, John Hopcroft, Ravi Kannan
  - Understanding Machine Learning: From theory to algorithms by Shai Shalev-Shwartz and Shai Ben-David
  - Introduction to Probability for Data Science by Stanley Chan

Collaboration and Attendance

- It’s OK to tell people where to look to get answers, or to correct mistakes.

Collaboration of any form on exams is strictly forbidden and will be punished

- If in doubt about posting a certain content publicly, you can create a private post only visible to you and the instructors.

- Attending lectures is not mandatory, but it is ***highly*** recommended.
  - Lab attendance will be taken by your TF.

Data Science

Datum: A piece of information (Definition from Oxford Languages)

“This 5,000-Year-Old Rock

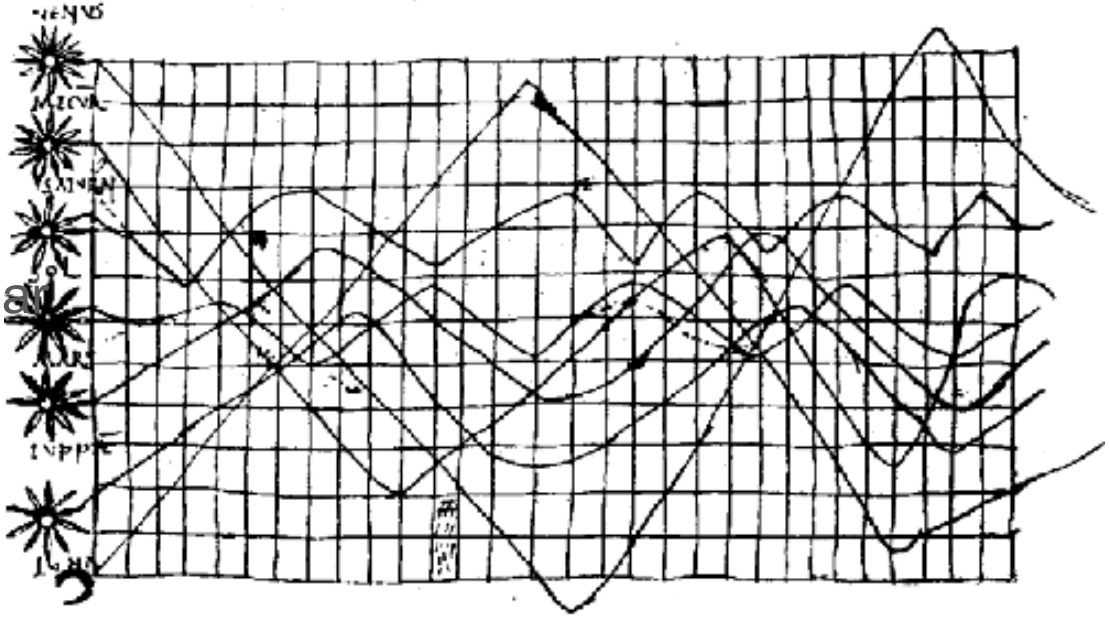
Could Be One of The World's

Earliest Maps”

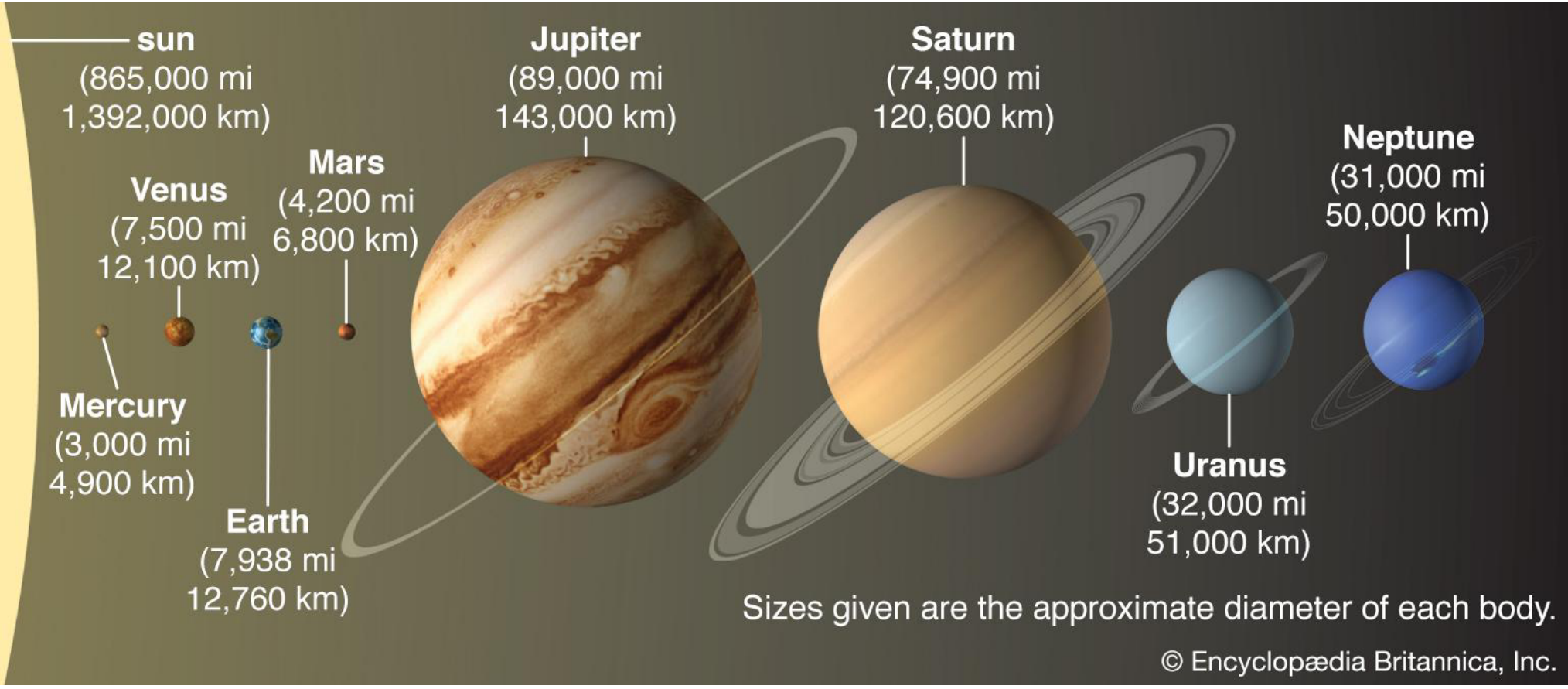
Source: Sciencealert

Data visualization

- Positions of the Sun, Moon, and Planets throughout the year (Europe, 950 AD)



- Galileo's visualization of Sunspots


sun
(865,000 mi
1,392,000 km)
Mercury
(3,000 mi
4,900 km)
Venus
(7,500 mi
12,100 km)
Earth
(7,938 mi
12,760 km)
Mars
(4,200 mi
6,800 km)
Jupiter
(89,000 mi
143,000 km)
Saturn
(74,900 mi
120,600 km)
Uranus
(32,000 mi
51,000 km)
Neptune
(31,000 mi
50,000 km)
Sizes given are the approximate diameter of each body.
© Encyclopædia Britannica, Inc.

Periodic table

Periodic Table of Elements

	1 IA												13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 0
1	1 H Hydrogen 1.008	2 IIA																2 He Helium 4.003
2	3 Li Lithium 6.941	4 Be Beryllium 9.012											5 B Boron 10.812	6 C Carbon 12.011	7 N Nitrogen 14.007	8 O Oxygen 15.999	9 F Fluorine 18.998	10 Ne Neon 20.178
3	11 Na Sodium 22.99	12 Mg Magnesium 24.305	3 IIIB	4 IVB	5 VB	6 VIB	7 VIIB	8 VIIIB	9 VIIIB	10 VIIIB	11 IB	12 IIB	13 Al Aluminum 26.982	14 Si Silicon 28.086	15 P Phosphorus 30.974	16 S Sulfur 32.066	17 Cl Chlorine 35.453	18 Ar Argon 39.948
4	19 K Potassium 39.098	20 Ca Calcium 40.078	21 Sc Scandium 44.956	22 Ti Titanium 47.867	23 V Vanadium 50.942	24 Cr Chromium 51.996	25 Mn Manganese 54.938	26 Fe Iron 55.845	27 Co Cobalt 58.933	28 Ni Nickel 58.693	29 Cu Copper 63.546	30 Zn Zinc 65.38	31 Ga Gallium 69.723	32 Ge Germanium 72.64	33 As Arsenic 74.922	34 Se Selenium 78.96	35 Br Bromine 79.904	36 Kr Krypton 83.798
5	37 Rb Rubidium 85.468	38 Sr Strontium 87.62	39 Y Yttrium 88.906	40 Zr Zirconium 91.224	41 Nb Niobium 92.906	42 Mo Molybdenum 95.96	43 Tc Technetium 97.907	44 Ru Ruthenium 101.07	45 Rh Rhodium 102.906	46 Pd Palladium 106.42	47 Ag Silver 107.868	48 Cd Cadmium 112.412	49 In Indium 114.818	50 Sn Tin 118.711	51 Sb Antimony 121.760	52 Te Tellurium 127.60	53 I Iodine 126.904	54 Xe Xenon 131.294
6	55 Cs Cesium 132.905	56 Ba Barium 137.327		72 Hf Hafnium 178.49	73 Ta Tantalum 180.948	74 W Tungsten 183.84	75 Re Rhenium 86.207	76 Os Osmium 190.23	77 Ir Iridium 192.217	78 Pt Platinum 195.085	79 Au Gold 196.967	80 Hg Mercury 200.59	81 Tl Thallium 204.383	82 Pb Lead 207.2	83 Bi Bismuth 208.980	84 Po Polonium (209)	85 At Astatine (210)	86 Rn Radon (222)
7	87 Fr Francium (223)	88 Ra Radium (226)		104 Rf Rutherfordium (267)	105 Db Dubnium (268)	106 Sg Seaborgium (271)	107 Bh Bohrium (272)	108 Hs Hassium (277)	109 Mt Meitnerium (276)	110 Ds Damstadtium (281)	111 Rg Roentgenium (280)	112 Cn Copernicium (285)	113 Nh Nihonium (286)	114 Fl Flerovium (289)	115 Mc Moscovium (289)	116 Lv Livermorium (293)	117 Ts Tennessine (294)	118 Og Oganesson (294)

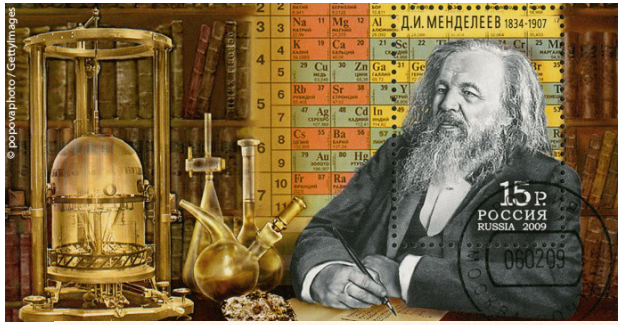
1 — Atomic number
H — Element symbol
Hydrogen — Element name
1.008 — Atomic mass

57 La Lanthanum 138.905	58 Ce Cerium 140.116	59 Pr Praseodymium 140.908	60 Nd Neodymium 144.242	61 Pm Promethium (145)	62 Sm Samarium 150.36	63 Eu Europium 151.964	64 Gd Gadolinium 157.25	65 Tb Terbium 158.925	66 Dy Dysprosium 162.500	67 Ho Holmium 164.930	68 Er Erbium 167.259	69 Tm Thulium 168.934	70 Yb Ytterbium 173.054	71 Lu Lutetium 174.967
89 Ac Actinium (227)	90 Th Thorium 232.038	91 Pa Protactinium 231.036	92 U Uranium 238.029	93 Np Neptunium (237)	94 Pu Plutonium (244)	95 Am Americium (243)	96 Cm Curium (247)	97 Bk Berkelium (247)	98 Cf Californium (251)	99 Es Einsteinium (252)	100 Fm Fermium (257)	101 Md Mendelevium (258)	102 No Nobelium (259)	103 Lr Lawrencium (262)

Metals
- Alkali Metals
- Alkaline Earth Metals
- Transition Metals
- Post-transition Metals
- Actinides
- Lanthanides

Nonmetals
- Noble Gases
- Halogens
- Other Nonmetals

Black = Solid at 25°C
Blue = Gas at 25°C
Red = Liquid at 25°C
Gray = Synthetic



ОПЫТЪ СИСТЕМЫ ЭЛЕМЕНТОВЪ.

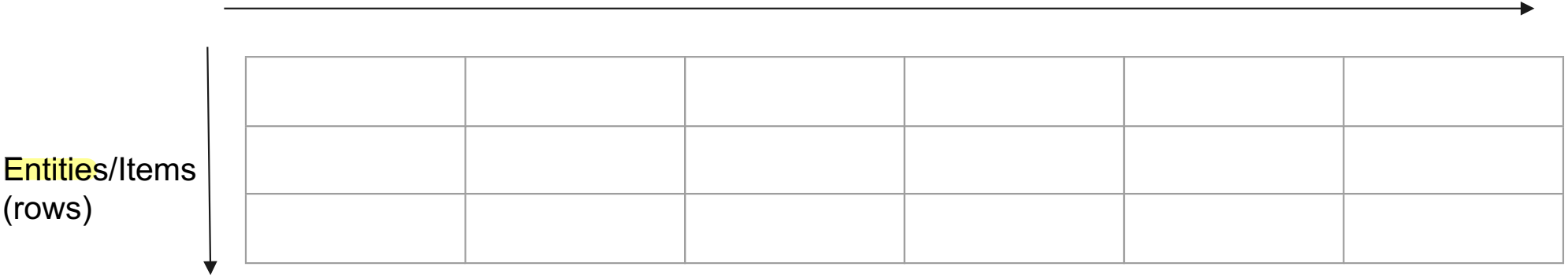
ОСНОВАННОЙ НА ИХЪ АТОМНОМЪ ВѢСѢ И ХИМИЧЕСКОМЪ СХОДСТВѢ.

			Ti = 50	Zr = 90	? = 180.	
			V = 51	Nb = 94	Ta = 182.	
			Cr = 52	Mo = 96	W = 186.	
			Mn = 55	Rh = 104,4	Pt = 197,1.	
			Fe = 56	Ru = 104,4	Ir = 198.	
			Ni = Co = 59	Pl = 106,6	O = 199.	
H = 1			Cu = 63,4	Ag = 108	Hg = 200.	
	Be = 9,4	Mg = 24	Zn = 65,2	Cd = 112		
	B = 11	Al = 27,1	? = 68	Ur = 116	Au = 197?	
	C = 12	Si = 28	? = 70	Sn = 118		
	N = 14	P = 31	As = 75	Sb = 122	Bi = 210?	
	O = 16	S = 32	Se = 79,4	Te = 128?		
	F = 19	Cl = 35,6	Br = 80	I = 127		
Li = 7	Na = 23	K = 39	Rb = 85,4	Cs = 133	Tl = 204.	
		Ca = 40	Sr = 87,6	Ba = 137	Pb = 207.	
		? = 45	Ce = 92			
		?Er = 56	La = 94			
		?Yt = 60	Di = 95			
		?In = 75,6	Th = 118?			

Д. Менделѣевъ

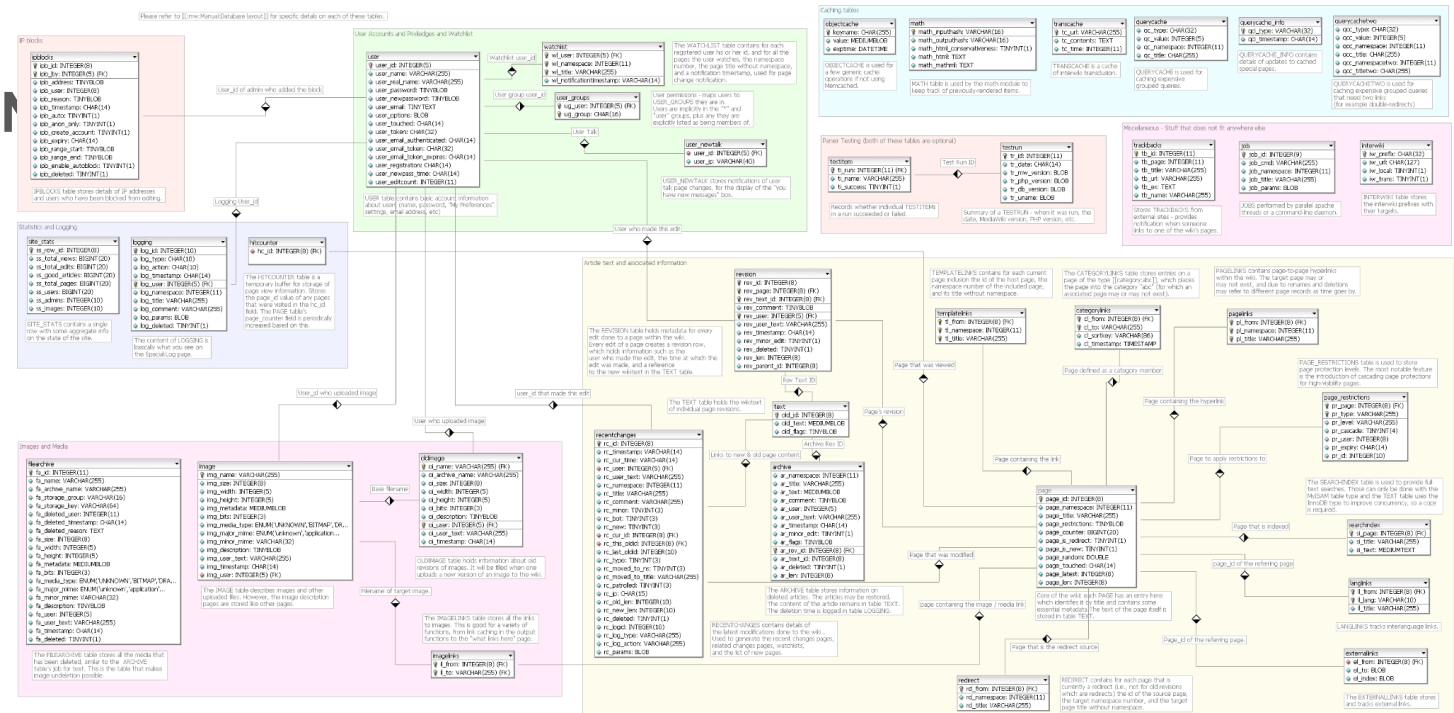
Types of data: Tables

Attributes (columns)

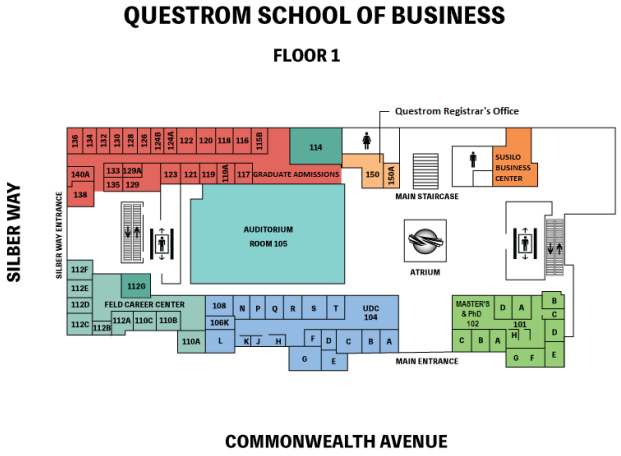


Databases and tables

Relational model, pioneered by E. Codd.



Multidimensional tables



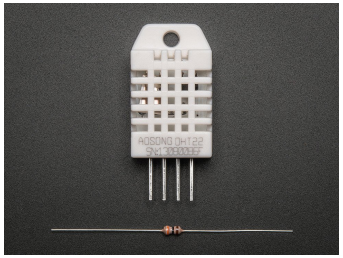
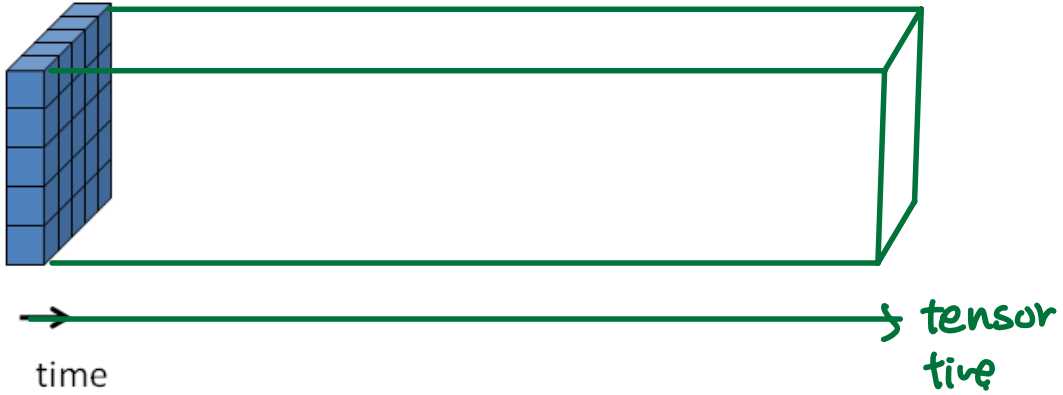


Fig sensor
- Temp
- Hum



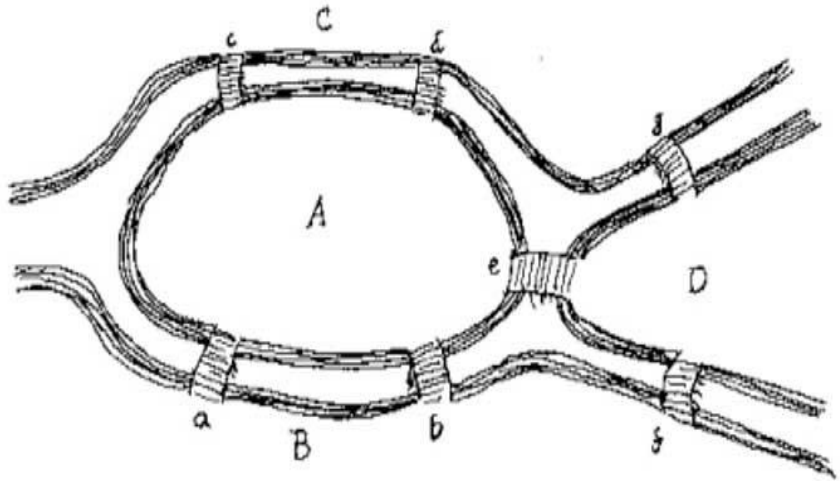
t=1
time

Sensor id	Temp	Hum.
 1		
 2		

t=2
…

What is a graph?

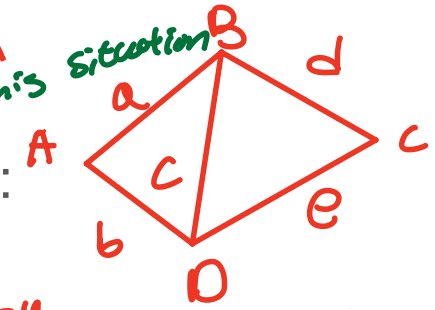
- Königsberg was a city located on the Pregel river in Prussia.
- Two islands, areas in both banks connected by 7 bridges.
- **Problem**: Could one wake up in their place, cross every bridge *exactly once*, and return to their place?



What is a graph?

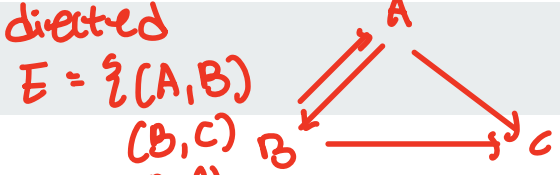
What is a graph?

- A graph **G** is a triple (V_G, E_G, ψ_G) consisting of:
  1. A set of vertices/nodes V(G)
  2. A set of edges E(G), disjoint from V(G)
  3. Incidence function ψ that associates with each edge a pair (not necessarily distinct) nodes:
     a. Unordered pair → undirected graph
     b. Ordered pair → directed graph
- If ψ(e)={u,v}:=*uv*, u,v are the *endpoints* of e that *joins* them. Nodes u,v are called *adjacent*.

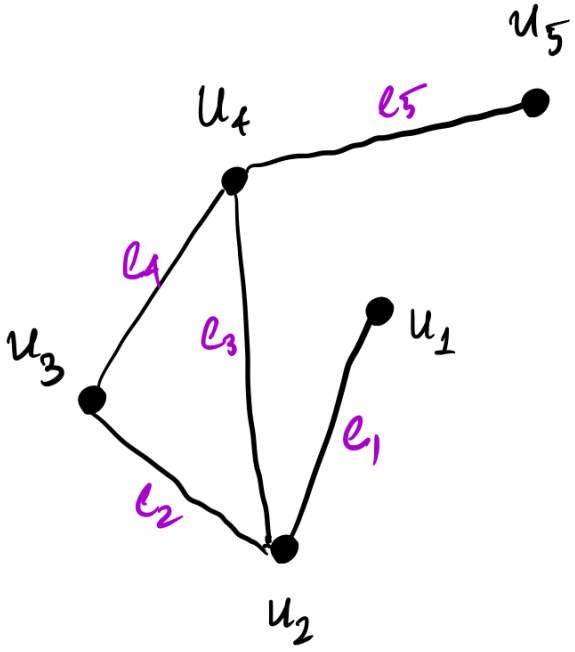


directed

$E = \{(A,B)$
(B,C)
(B,A)
(A,C)



What is a graph? Example



$V = \{u_1, u_2, u_3, u_4, u_5\}$

$E = \{e_1, e_2, e_3, e_4, e_5\}$

$\psi(e_1) = u_1u_2$, $\psi(e_2) = u_2u_3$

$\psi(e_3) = u_2u_4$, $\psi(e_4) = u_3u_4$

$\psi(e_5) = u_4u_5$.

What is a graph?

- Undirected graphs
- Directed graphs
- Degree=#neighbors

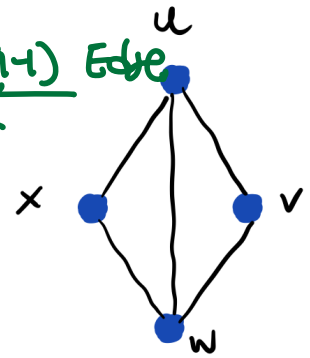
Directed graph

- In-degree=#incoming edges

  out-degree=#outcoming edges

n vertex

undirected : $\binom{n}{2}$ $\frac{n(n-1)}{2}$ Edge

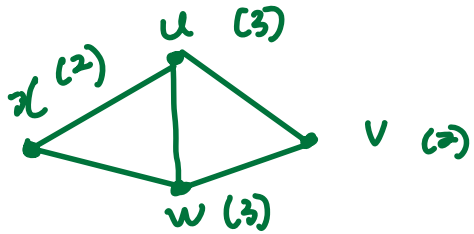


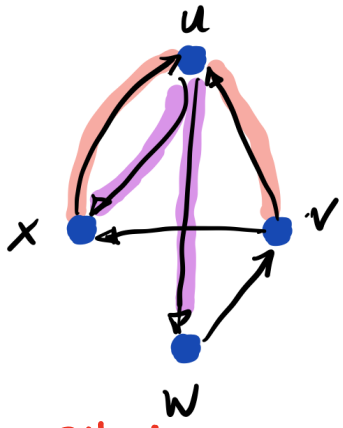
undirected

$N(u) = \{x, v, w\}$

$\deg(u) = |N(u)| = 3$

# of Edges





outgoing

$N^+(u) = \{x, w\}$

$d^+(u) = |N^+(u)| = 2$

incoming

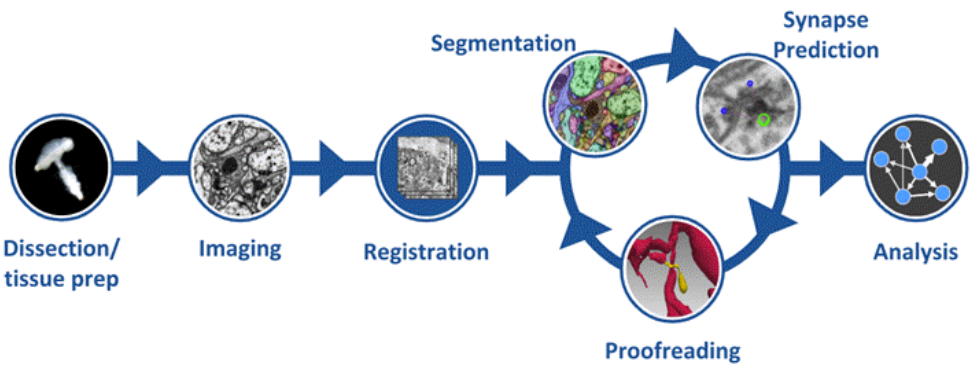
$N^-(u) = \{x, v\}$

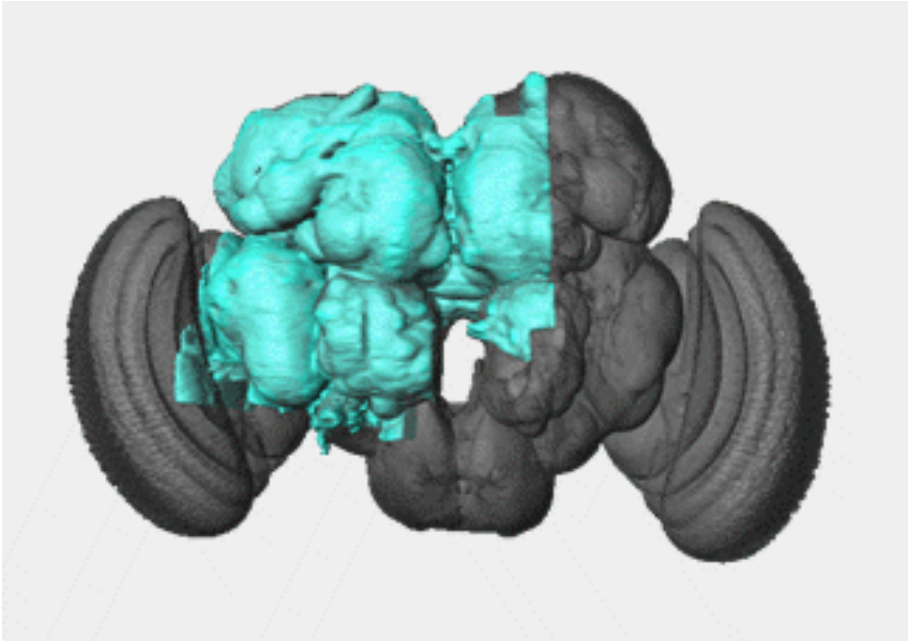
$d^-(u) = |N^-(u)| = 2$

3+2+3+2 = 10 (2 × # of Edges) Prove: Every Edge contributes two Edges.

Graphs model a wide variety of datasets

Drosophila hemibrain networks





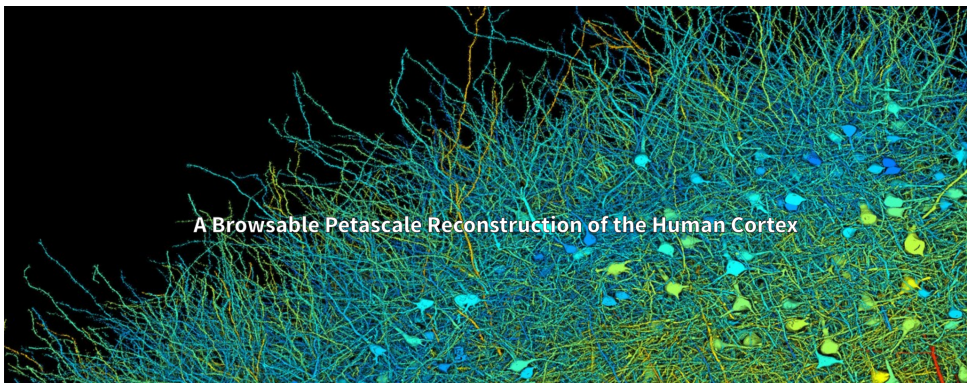
Source:

https://www.janelia.org/project-team/flyem/hemibrain

Graphs model a wide variety of datasets

- Human brain as G(V,E)
- V: neurons
- E: synapses

https://h01-release.storage.googleapis.com/landing.html



Graphs model a wide variety of datasets

Graphs model a wide variety of datasets

Computer networks

- V: computers

- E: connection links

Image source

Graphs model a wide variety of datasets

Zhou et al. *Cell Discovery* (2020)6:14
https://doi.org/10.1038/s41421-020-0153-3

Cell Discovery
www.nature.com/celldisc

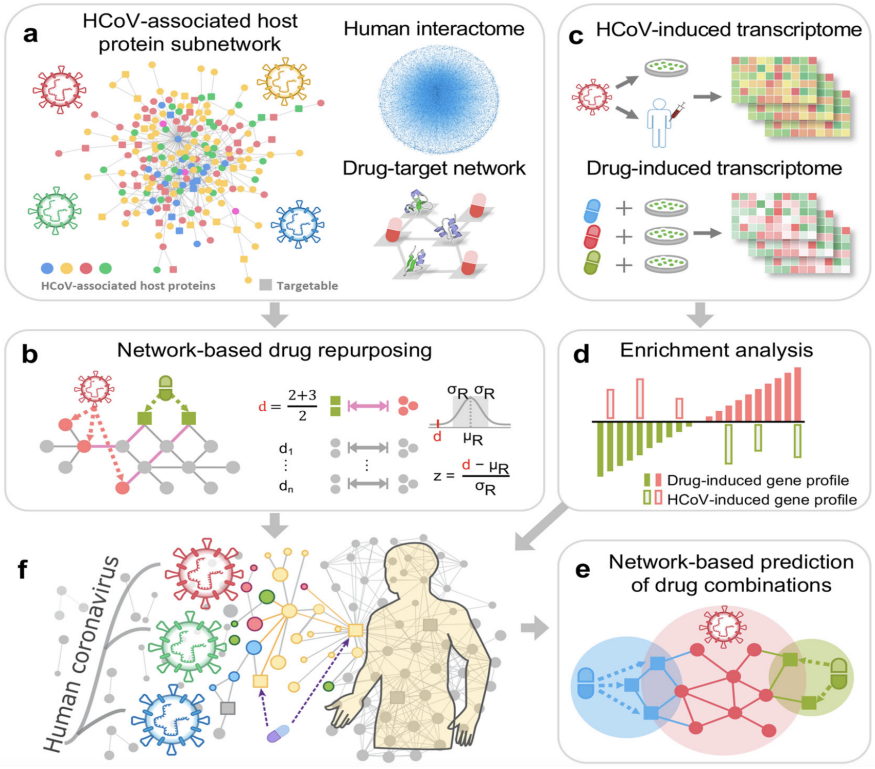
ARTICLE Open Access

Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2

Yadi Zhou[1], Yuan Hou[1], Jiayu Shen[1], Yin Huang[1], William Martin[1] and Feixiong Cheng[1,2,3]

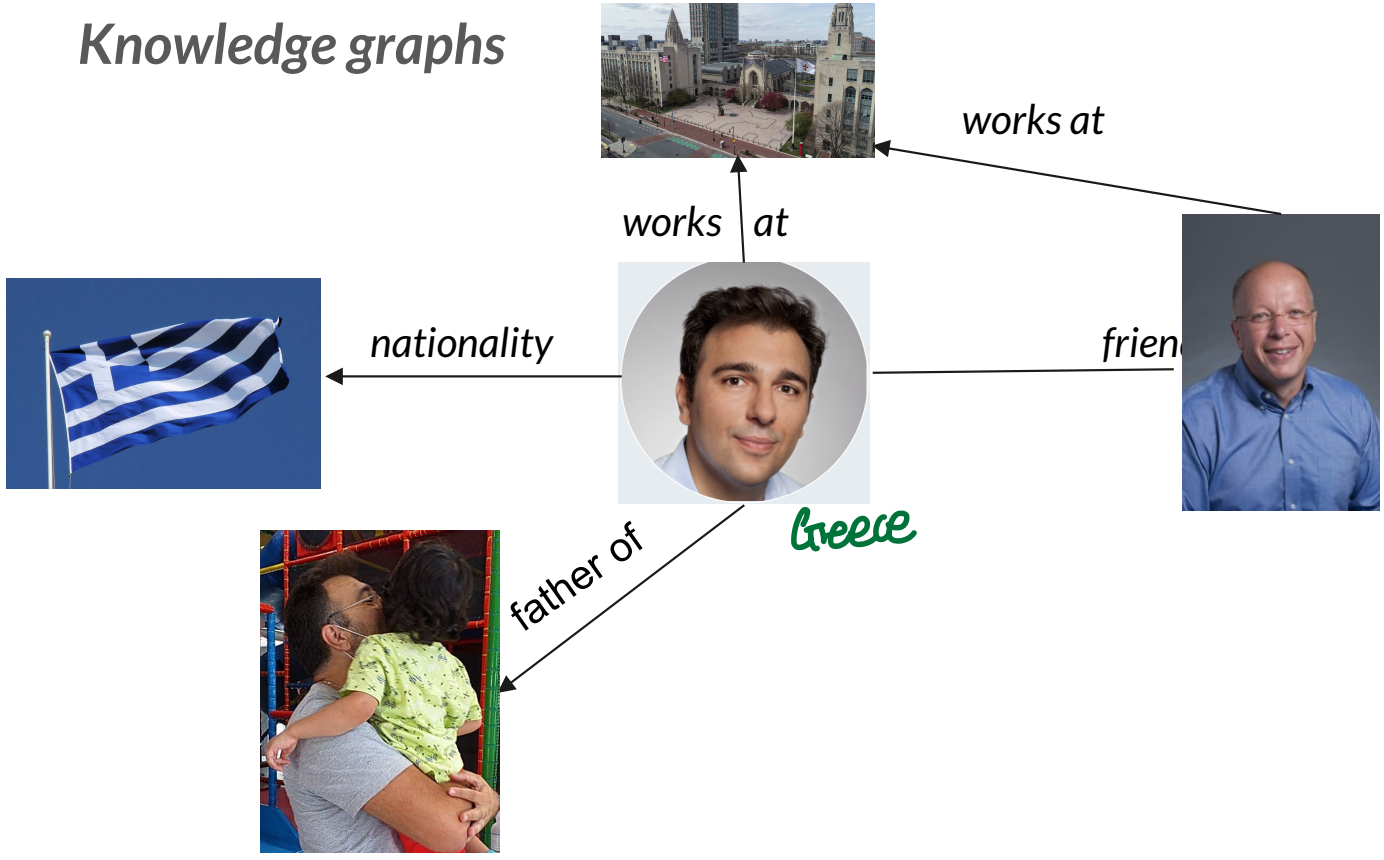
Abstract

Human coronaviruses (HCoVs), including severe acute respiratory syndrome coronavirus (SARS-CoV) and 2019 novel coronavirus (2019-nCoV, also known as SARS-CoV-2), lead global epidemics with high morbidity and mortality. However, there are currently no effective drugs targeting 2019-nCoV/SARS-CoV-2. Drug repurposing, representing as an effective drug discovery strategy from existing drugs, could shorten the time and reduce the cost compared to de novo drug discovery. In this study, we present an integrative, antiviral drug repurposing methodology implementing a systems pharmacology-based network medicine platform, quantifying the interplay between the HCoV–host interactome and drug targets in the human protein–protein interaction network. Phylogenetic analyses of 15 HCoV whole genomes reveal that 2019-nCoV/SARS-CoV-2 shares the highest nucleotide sequence identity with SARS-CoV (79.7%). Specifically, the envelope and nucleocapsid proteins of 2019-nCoV/SARS-CoV-2 are two evolutionarily conserved regions, having the sequence identities of 96% and 89.6%, respectively, compared to SARS-CoV. Using network proximity analyses of drug targets and HCoV–host interactions in the human interactome, we prioritize 16 potential anti-HCoV repurposable drugs (e.g., melatonin, mercaptopurine, and sirolimus) that are further validated by enrichment analyses of drug-gene signatures and HCoV-induced transcriptomics data in human cell lines. We further identify three potential drug combinations (e.g., sirolimus plus dactinomycin, mercaptopurine plus melatonin, and toremifene plus emodin) captured by the "*Complementary Exposure*" pattern: the targets of the drugs both hit the



Graphs model a wide variety of datasets

Knowledge graphs



Graphs model a wide variety of datasets

Collaboration networks

- V: scientists
- E: collaborations

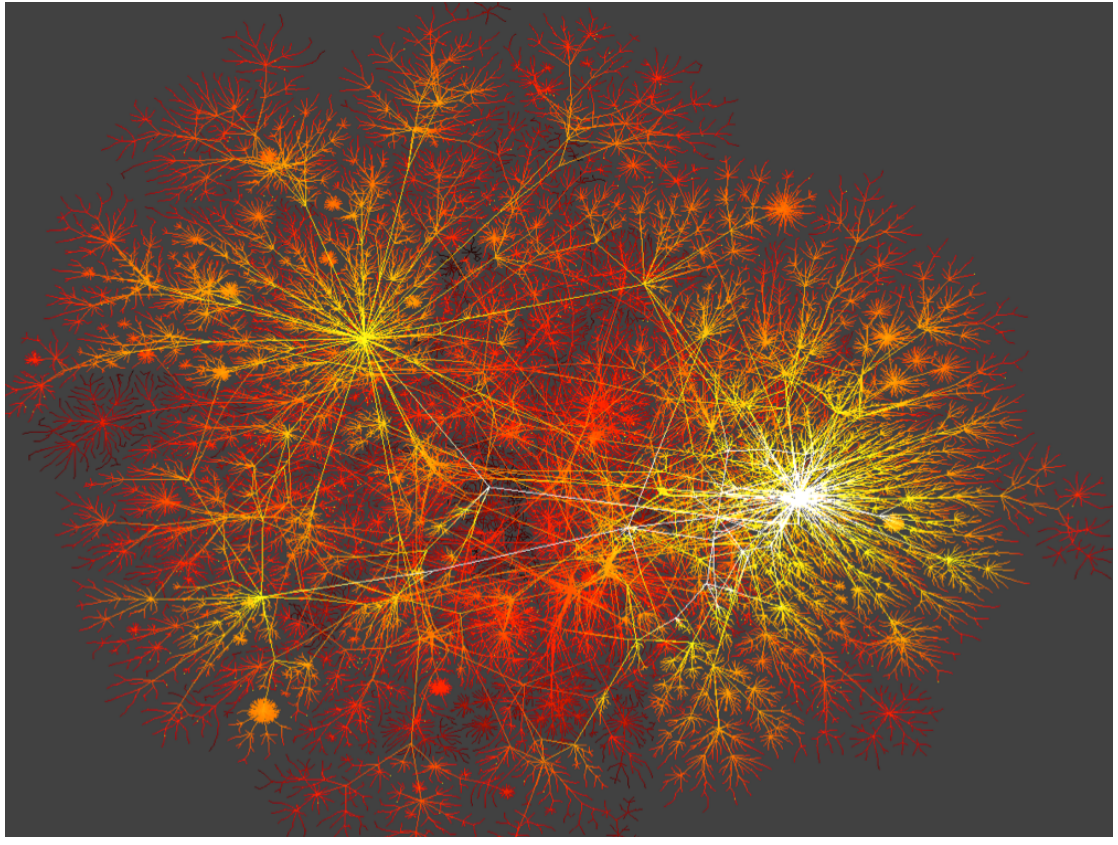
Image

Graphs model a wide variety of datasets

Transportation networks

Graphs model a wide variety of datasets

Internet



Graphs model a wide variety of datasets

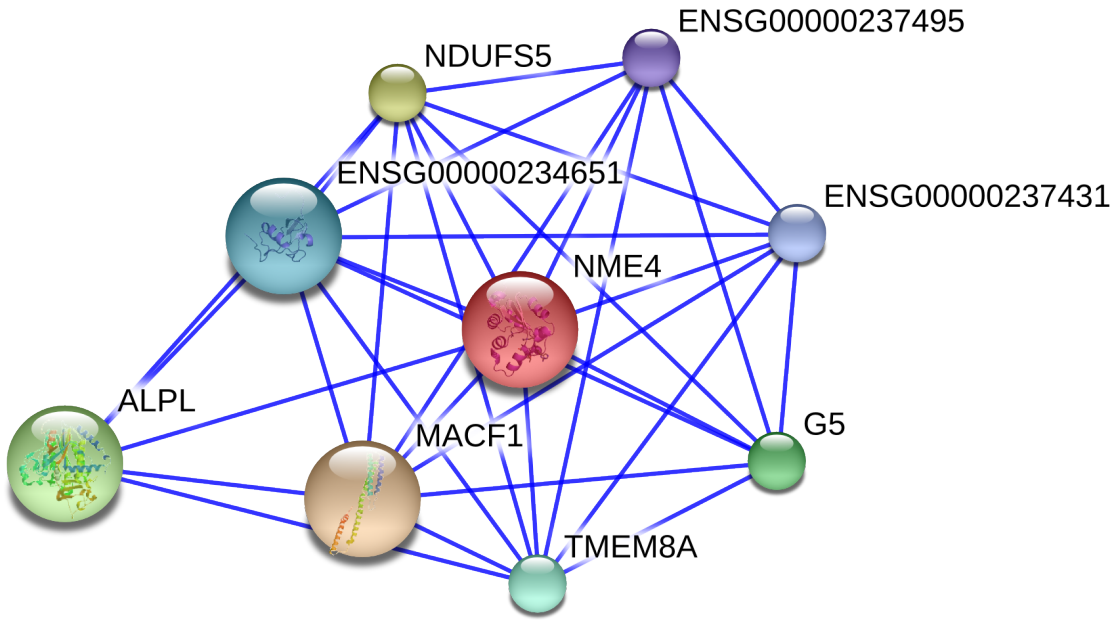
Transaction networks
- V: bank account
- E: transaction from-to

Can we spot money launderers and other fraudsters?

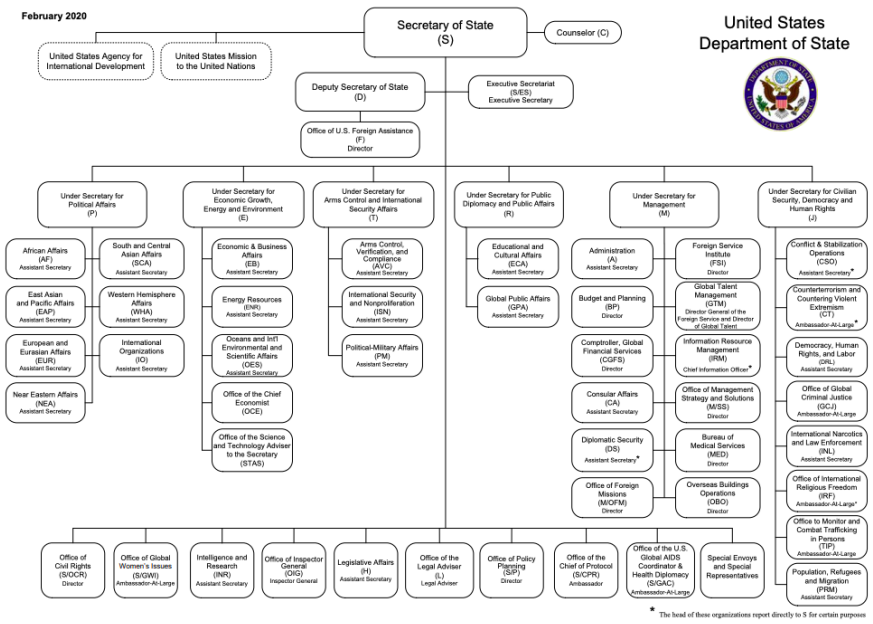
Graphs model a wide variety of datasets

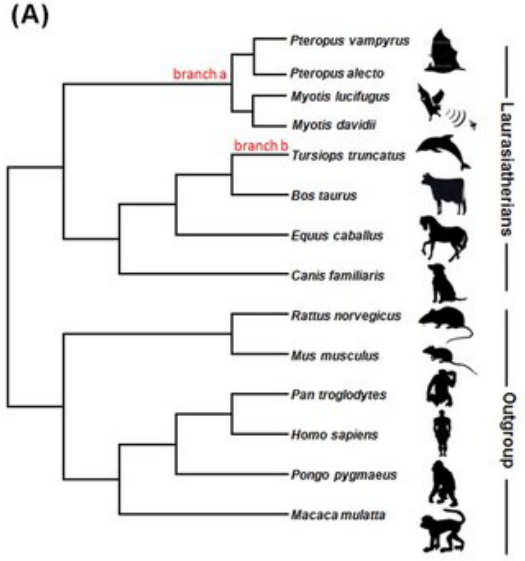
PPI networks

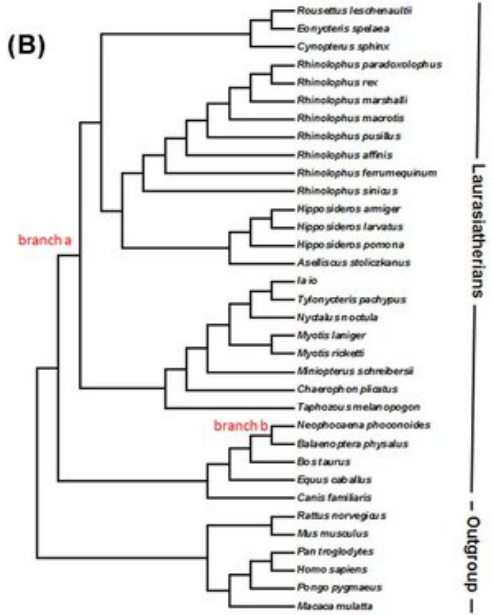
Image credit



Trees: (connected) graph with no cycles



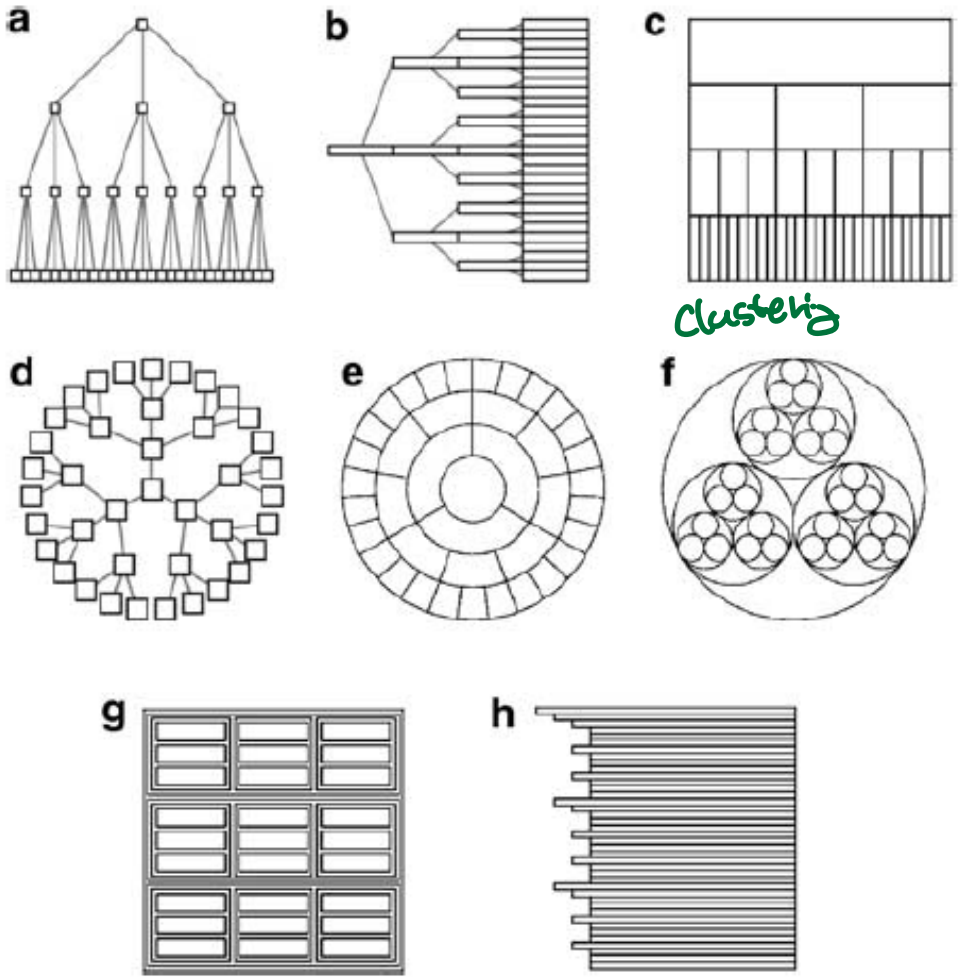




Remark

- Datasets can be visualized in a variety of ways.
- For example, a complete 3-ary tree of depth 3 can be visualized as follows:

Source: Quantifying the space-efficiency of 2D graphical representations of trees by McGuffin & Robert



Geography and Geometry data

gadm.org/download_country.html

GADM Maps Data About

Download GADM data (version 3.6)

Country

United States

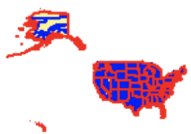
Geopackage

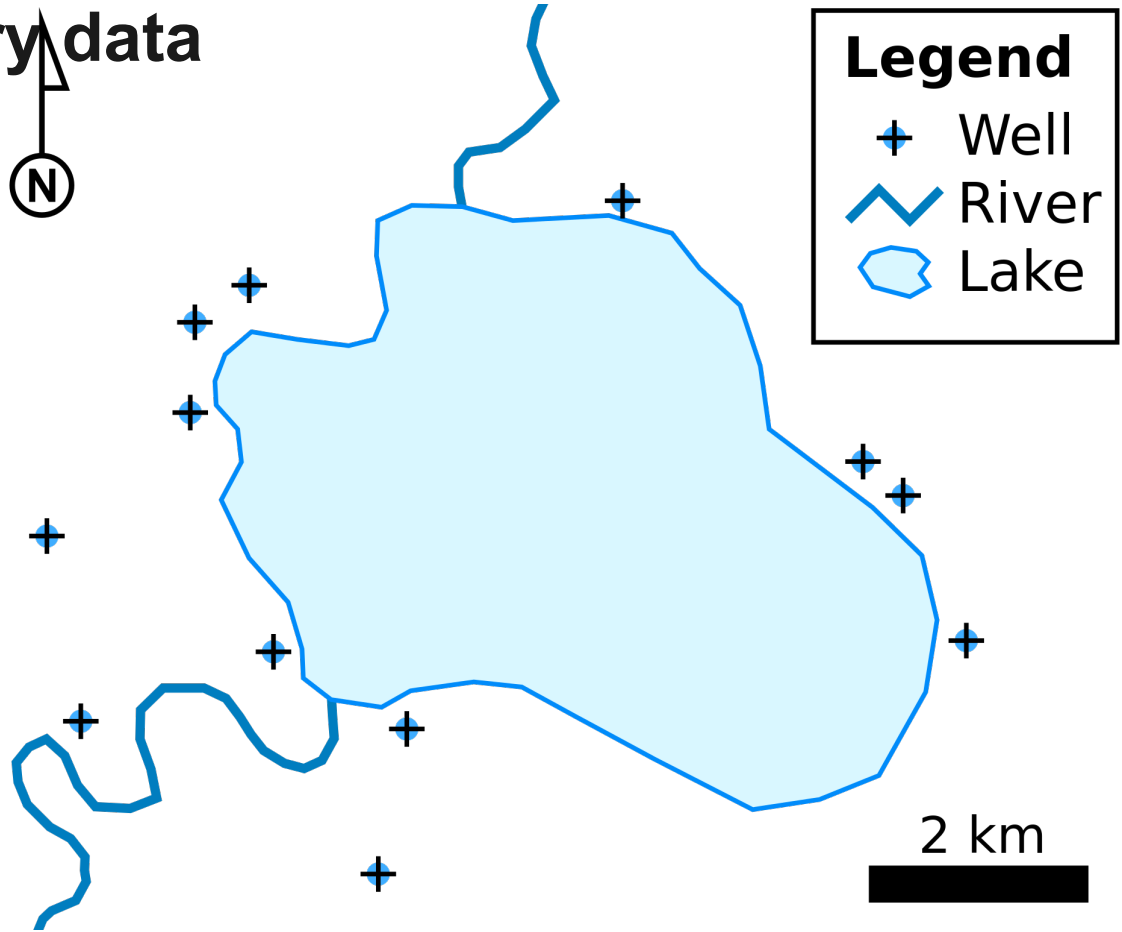
Shapefile

R (sp): level-0, level1, level2

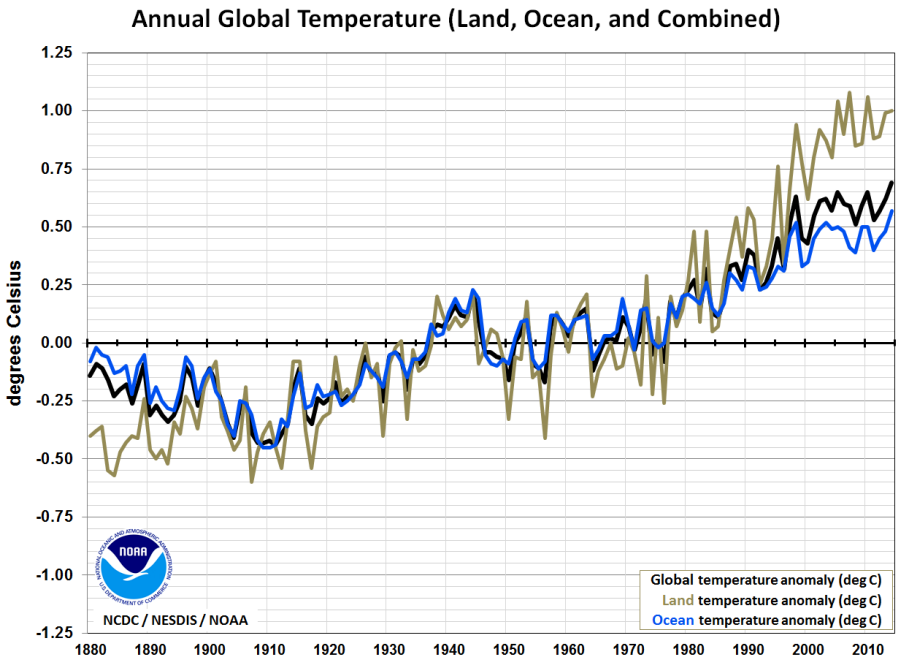
R (sf): level-0, level1, level2

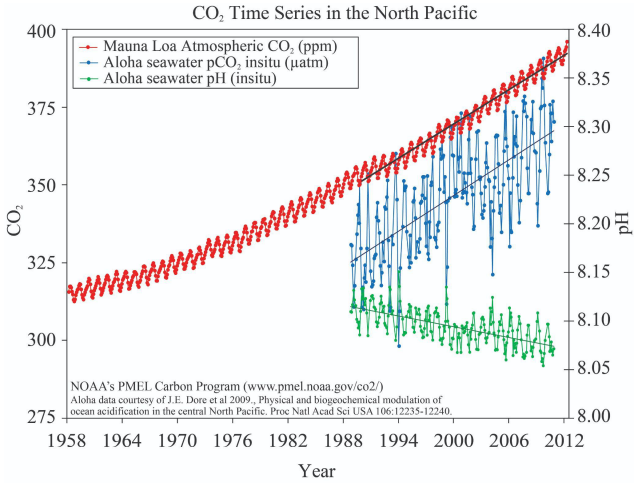
KMZ: level-0, level1, level2

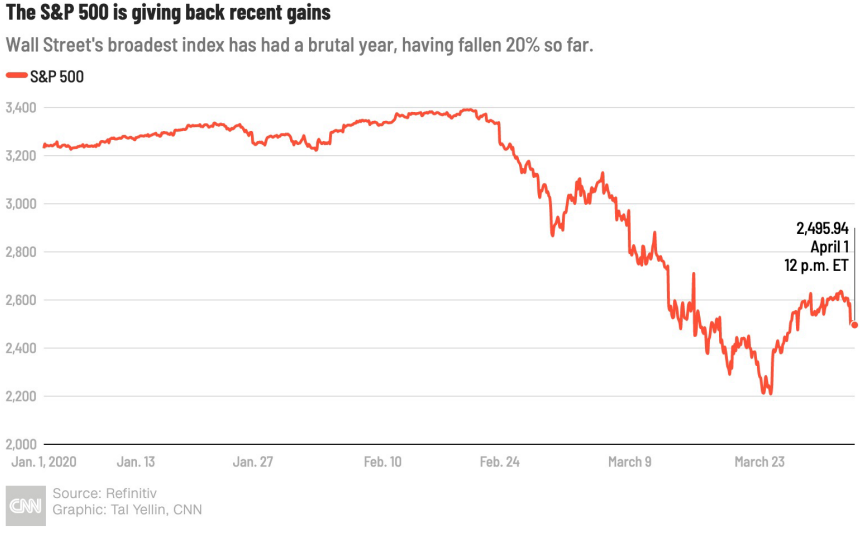




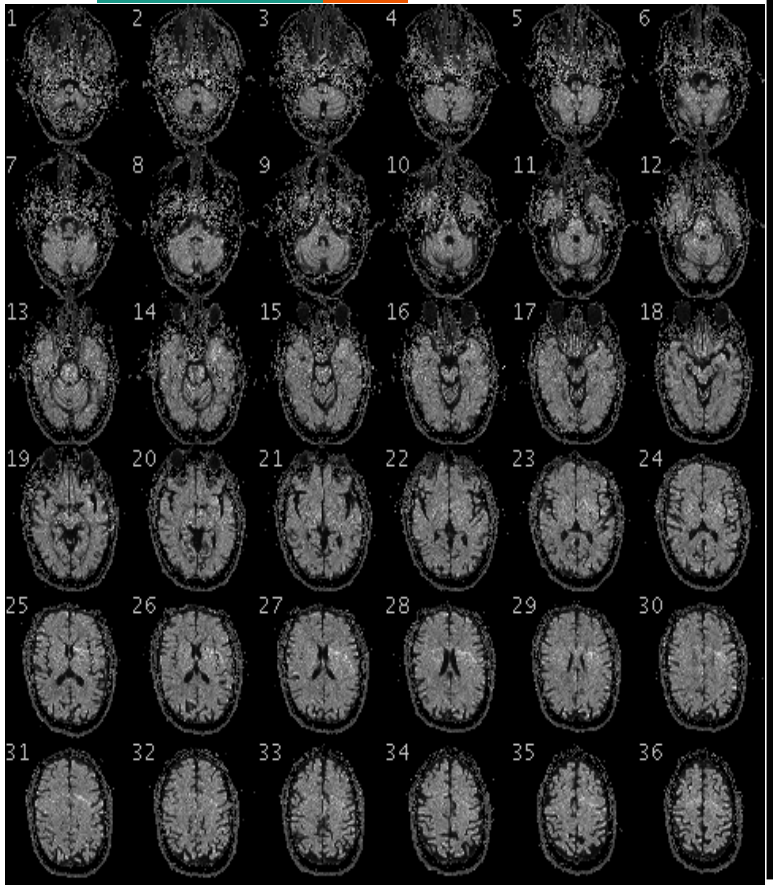
Time-series data

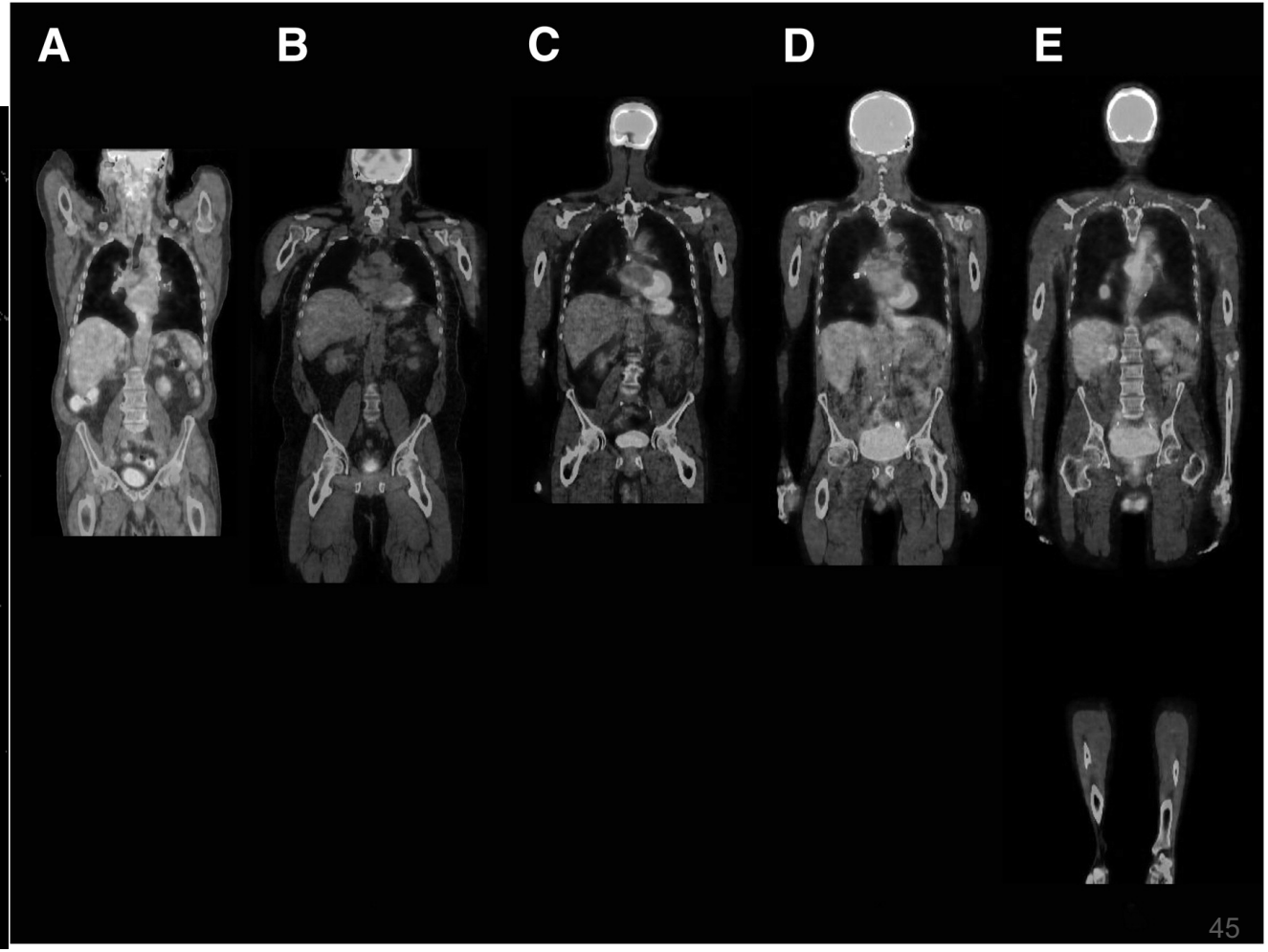






Complex data



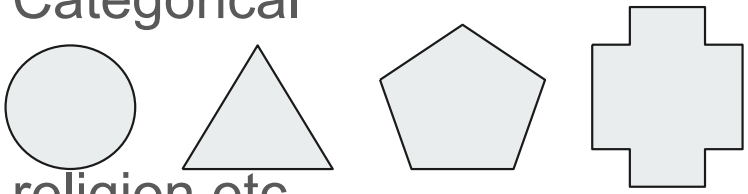


Dataset availability and types of attributes

- **Static vs. dynamic**
  - A dynamic dataset is a stream of updates. E.g., consider a table where items/tuples can be
    - inserted,
    - deleted,
    - or updated
- **Attribute Types**
  - Categorical
  - Quantitative

Attribute types

- Categorical



religion etc.

, marital status, gender,

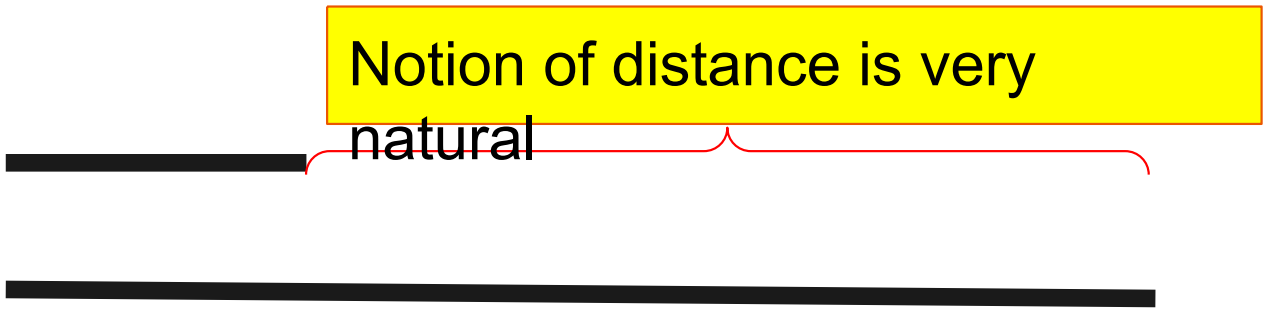
- Ordinal (sometimes also classified under quantitative)

Like	Neutral	Dislike

XL L M S (

Attribute types

- Quantitative data is information that can quantified; measured or counted, and thus be given a numerical value.



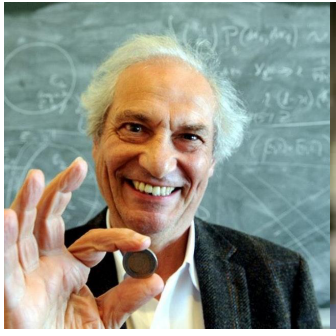
Probability

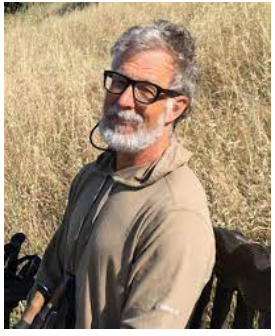
What is a fair coin?

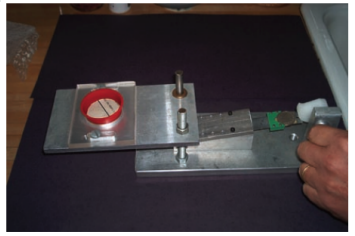
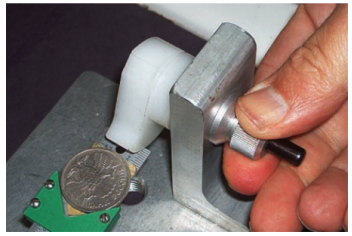
Are all coin flips random?

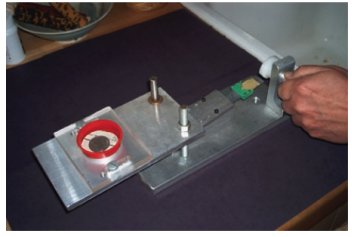
- Can a coin flip be “rigged”?
  - Yes!

- Dynamical Bias in the Coin Toss
  by
  P Diaconis, S Holmes, R Montgomery



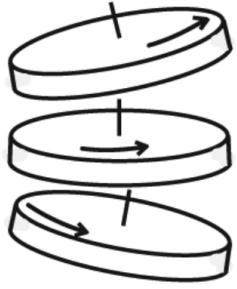




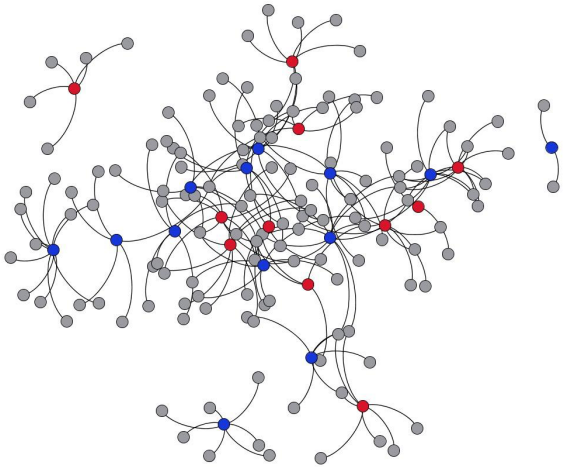
(a) (b) (c) (d)

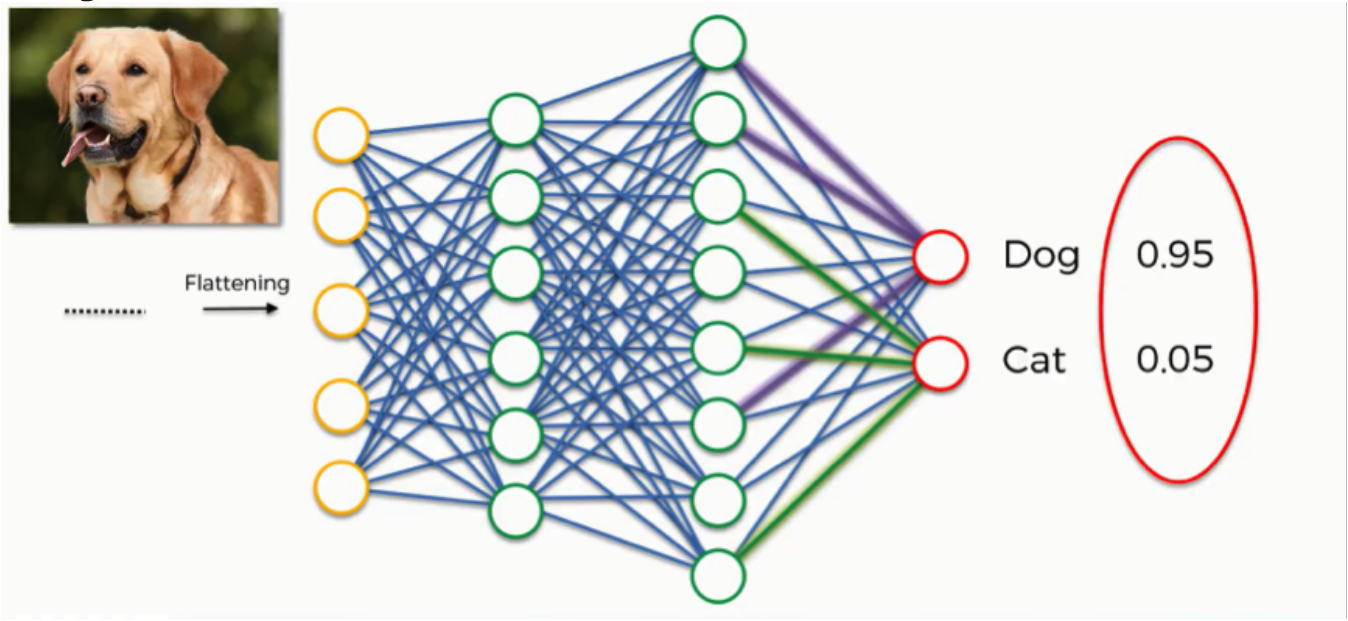
Fig. I



(a) (b)

Modeling uncertainty

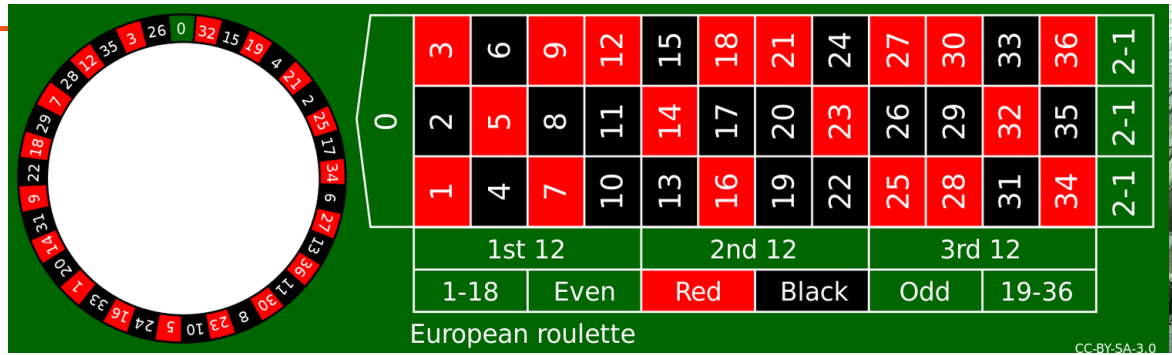




Disease spreading

Source: Blog

Modeling uncertainty



- Information theory, modeling the reliability of numerous complex systems, insurance companies, investments etc.
- **Today's agenda**: reminders of prereq probability material through problem solving.

Roulette

Odds & Payouts at European & American Roulette			
Roulette Bet	**Payout**	**European Roulette Odds**	**American Roulette Odds**
Single Number	35 to 1	2.70%	2.60%
2 Number Combination	17 to 1	5.4%	5.3%
3 Number Combination	11 to 1	8.1%	7.9%
4 Number Combination	8 to 1	10.8%	10.5%
5 Number Combination	6 to 1	13.5%	13.2%
6 Number Combination	5 to 1	16.2%	15.8%
Column	2 to 1	32.40%	31.6%
Dozen	2 to 1	32.40%	31.6%
Even/Odd	1 to 1	48.60%	47.4%
Red/Black	1 to 1	48.60%	47.4%
Low/High	1 to 1	48.60%	47.4%

- Even American: {2,4,6,8,...,34,36}, hence Pr(even)=18/38=0.47368
- Even European: same favorable outcomes {2,4,6,8,...,34,36}, but Pr(even)=18/37=0.4864

Probability space

$$(\Omega, \mathcal{F}, \mathbb{P})$$

Sample space

event space

probability measure

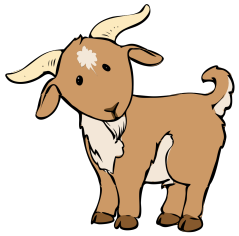
$$\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$$

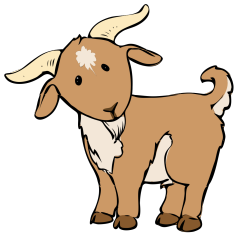
Questions: what is a random variable? What is the difference between continuous and discrete random variables?

Monty-hall problem

Suppose you're on a game show, and you're given the choice of three doors:

- Behind one door is a car; behind the others, goats.
- You pick a door, say No. A, and the host, who knows what's behind the doors, opens another door, say No. C, which has a goat.
- He then says to you, "Do you want to pick door No. B?" Is it to your advantage to switch your choice?





Assumptions

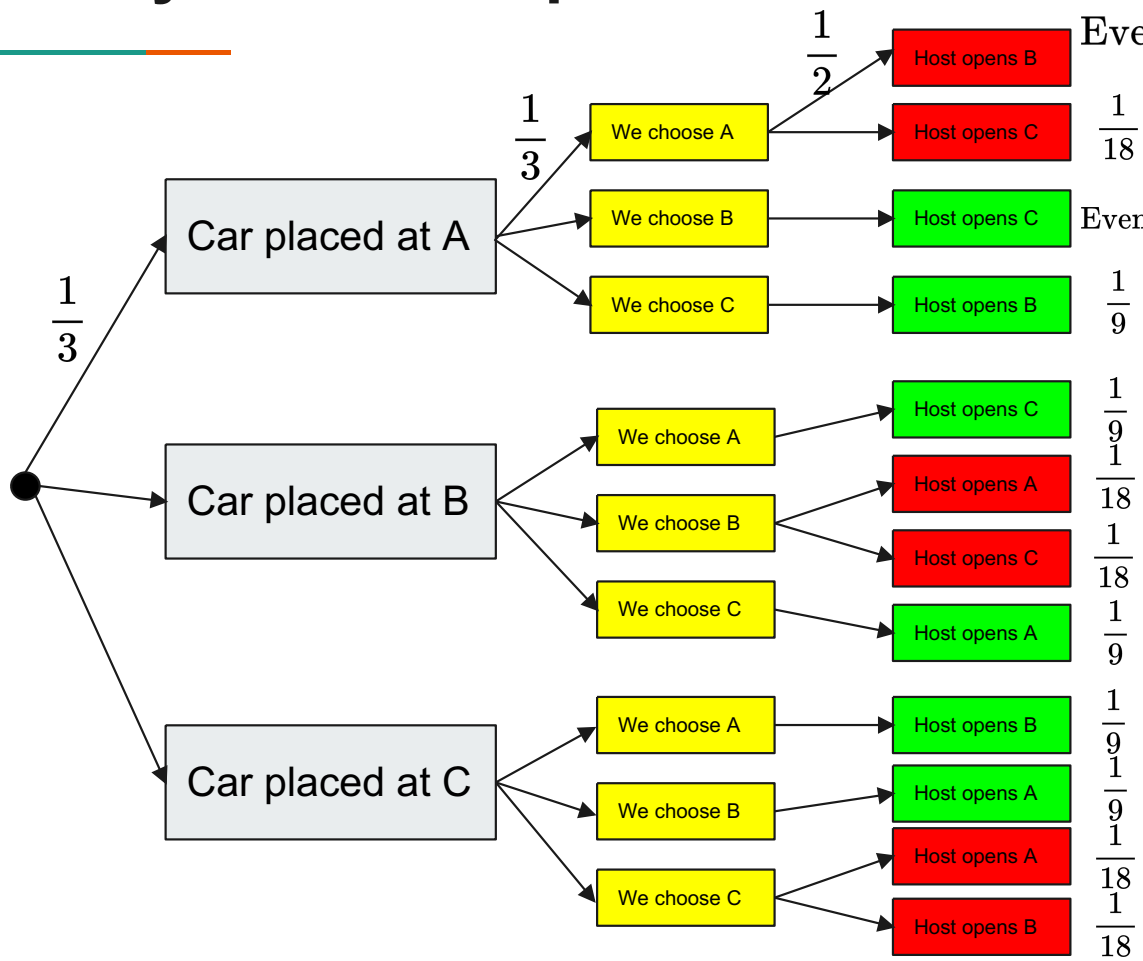
Let’s make the problem concrete by specifying certain assumptions.

- Let’s say the car is placed uniformly at random (**uar**) behind a door.
- Our initial guess is also **uar**
- The host opens a door with a goat. When there exist two such doors, i.e., our guess is the car, he chooses **uar**.

CS131 Reminder: Four step-method

1. Find the sample space
2. Define events of interest
3. Determine outcome probabilities
4. Compute event probabilities

Monty hall - 4 steps in one slide

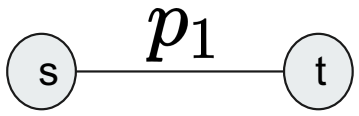


$$\text{Event (AAB)} : \frac{1}{3} \times \frac{1}{3} \times \frac{1}{2} = \frac{1}{18}$$

$$\text{Event (ABC)}: \frac{1}{3} \times \frac{1}{3} \times 1 = \frac{1}{9}$$

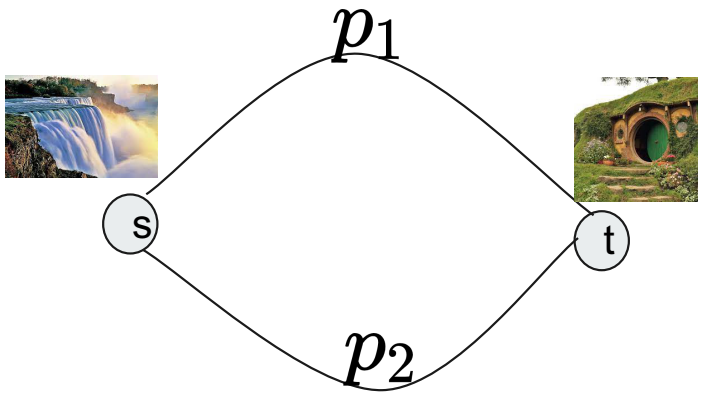
Therefore, switching yields
Probability of winning 6/9=⅔
Vs not switching 6/18=⅓

Transfer water



- Consider a water source **s** and a destination village **t**.
- Each pipe *i* has probability of failure p_i. Pipes fail ***independently***.
- **Question**: What is the probability we cannot get water from s to t? In other words:
  - when is the village **t** not reachable from the water source **s**?

Exercise 1



- Clearly, there is no path if both pipes fail
- Since they are independent, the probability of this event is the product of the probabilities of the individual events

Thus, failure probability is p_1p_2

Reminders: Independent events, conditional probability

Intuitive two events A,B are dependent if A's occurrence or non-occurrence provides us with some information about event B.

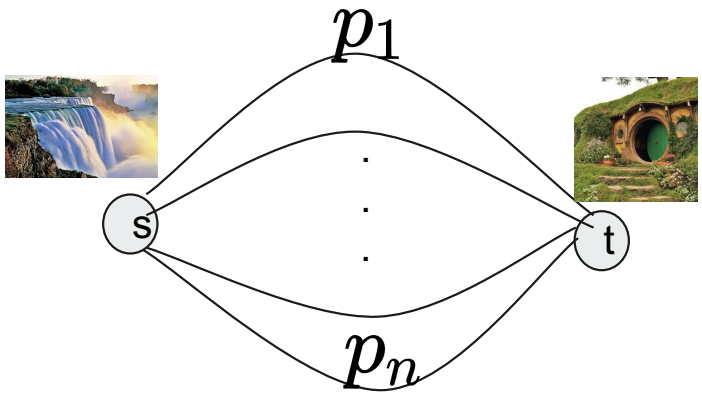
Formally, A,B are independent if and only iff $\boxed{\Pr(A \cap B) = \Pr(A)\Pr(B)}$

By rearranging we get $\boxed{\Pr(A) = \frac{\Pr(A \cap B)}{\Pr(B)}}$

Recall that by the law of conditional probability $\boxed{\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}}$

Therefore, when A,B are independent Pr(A)=Pr(A|B) and of course Pr(B)=Pr(B|A).

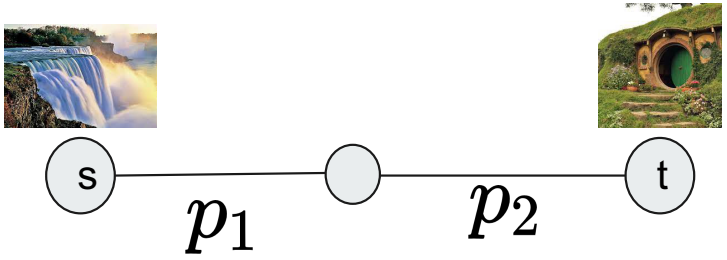
Exercise 2



- Clearly, there is no path if **all** pipes fail
- Since they are independent, the probability of this event is the product of the probabilities of the individual events

Thus, failure probability is $p_1 p_2 \ldots p_n$

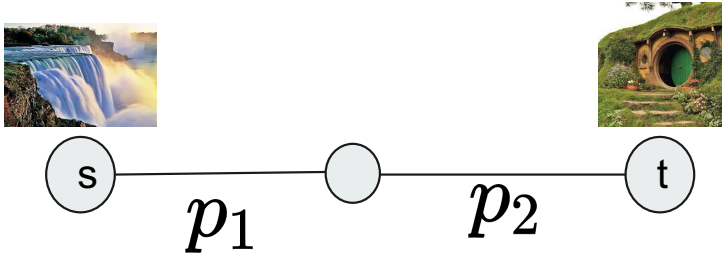
Exercise 3



- Clearly, there is no path if **at least one of** the pipes fail
- Let A_i be the event that pipe i fails.
- Then,

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\ &= p_1 + p_2 - \Pr(A_1)\Pr(A_2) \\ &= p_1 + p_2 - p_1 p_2\end{aligned}$$

Exercise 3



- **Using conditional probability**

- We condition on whether the one of the two pipes (say the first) is broken or not.

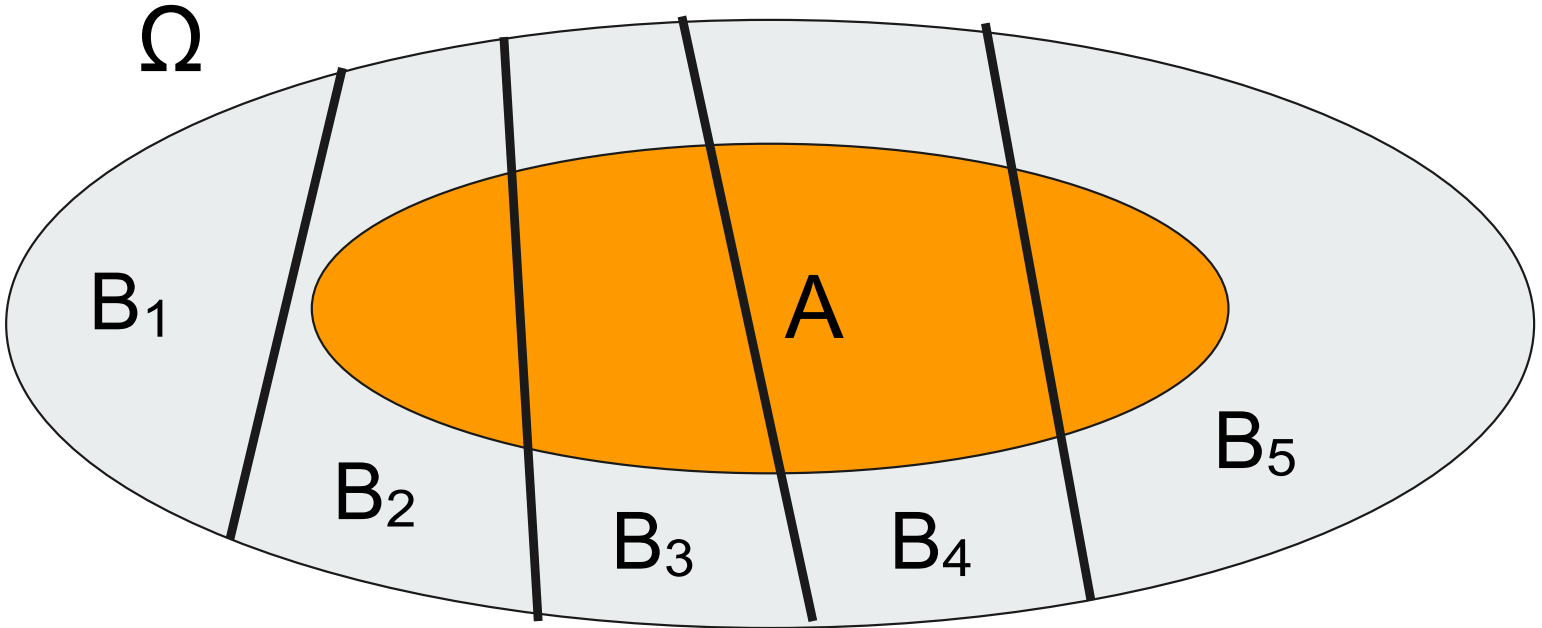
- Let A_i be the event that pipe i fails.

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + (1 - \Pr(A_1))\Pr(A_2) \\ &= p_1 + (1 - p_1)p_2 \\ &= p_1 + p_2 - p_1 p_2\end{aligned}$$

We used the law of total probability

Let Ω be a probability space. Let $B_1,\ldots,B_m$ be a partition of Ω. Then,

$$\Pr(A) = \sum_{i=1}^{m} \Pr(A \cap B_i) = \sum_{i=1}^{m} \Pr(B_i)\Pr(A|B_i)$$



Exercise 4

- Instead of thinking the probability that t will not be reachable from s, we think of the probability that it is. **Reminder**: $\Pr\left(\bar{A}\right) = 1 - \Pr(A)$
  - Let $\bar{A}_i$ be the event that pipe i does not fail.
- The probability of not failing is $\Pr\left(\cap_{i=1}^{n} \bar{A}_i\right) = \prod_{i=1}^{n} \Pr\left(\bar{A}_i\right) = \prod_{i=1}^{n} (1 - p_i)$

Therefore, the right answer is $1 - \prod_{i=1}^{n} (1 - p_i)$.

Reminder: chain rule

Chain rule:

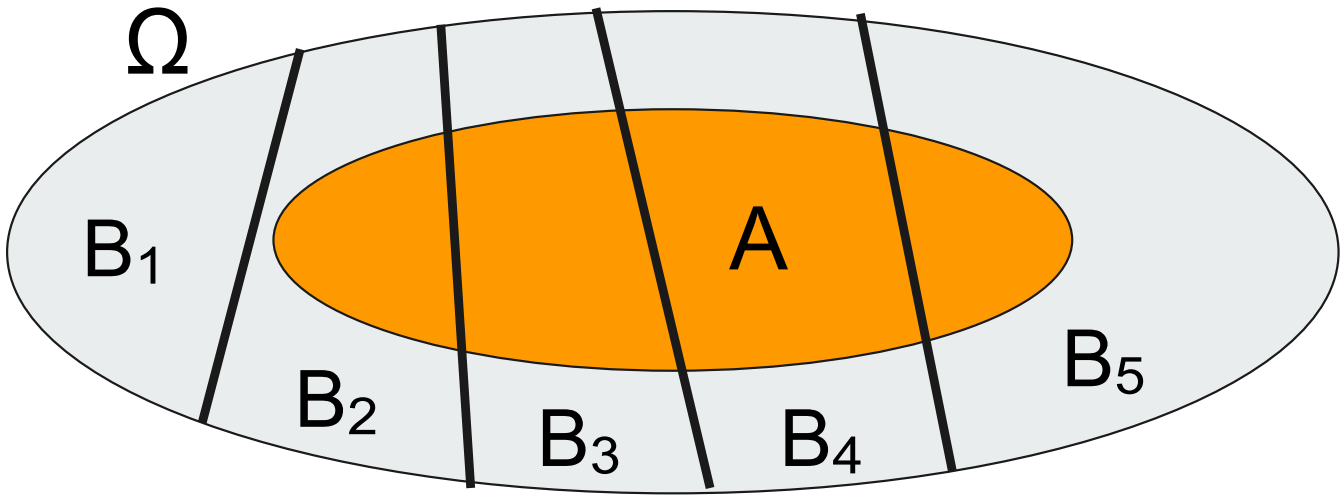
$$\Pr(A_1 \cap \ldots \cap A_n) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_2A_1)\ldots\Pr(A_n|A_{n-1}..A_1)$$

In our case the events are mutually independent, so this simplifies to the product of the individual probabilities of the events A_i.

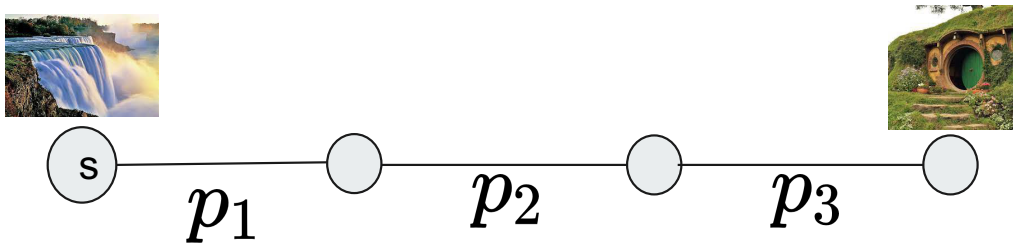
Question: what is the difference between pairwise and mutually independent events?

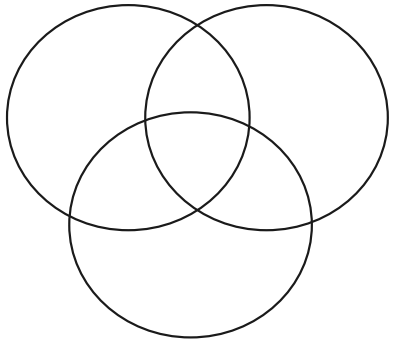
Conditional probability + Law of total probability → Bayes rule

$$\Pr(B_i|A) = \frac{\Pr(B_i \cap A)}{\Pr(A)} = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{n} \Pr(B_j)\Pr(B_j|A)}$$



Exercise n=3





$$1-(1-p_1)(1-p_2)(1-p_3) = 1-(1-p_1)(1-p_2-p_3+p_2p_3)$$

$$= 1-(1-p_2-p_3+p_2p_3-p_1+p_1p_2+p_1p_3-p_1p_2p_3)$$

$$= p_1+p_2+p_3-p_1p_2-p_1p_3-p_2p_3+p_1p_2p_3$$

Does this remind of something from CS131?

Exercise 4

- We condition on whether the one of the two pipes (say the first) is broken or not.
- Let A_i be the event that pipe i fails.
- We are interested in $\Pr(A_1 \cup \ldots \cup A_n)$.

Inclusion-exclusion

$$\Pr\left(\cup_{i=1}^{n} A_i\right) = \sum_{\mathcal{J}\subseteq\{1,\ldots,n\};|\mathcal{J}|=k} (-1)^{k+1} P\left(\bigcap_{i\in\mathcal{J}} A_i\right)$$

Proof sketch (inductive proof)

When n=1 the statement is obvious. Use the IH and the fact that

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) + P\left(A_{n+1} \setminus \bigcup_{i=1}^{n} A_i\right)$$

$$= P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) - P\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right).$$

Inclusion exclusion

Another convenient way to write the IE formula is the following

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) = S_1 - S_2 + S_3 - \ldots + (-1)^{n-1} S_n$$

where $$S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}).$$

In our setting, due to the independence of the events A_i we can write the following expression

$$\Pr(\cup A_i) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq [n], |I|=k} \prod_{i \in I} \Pr(A_i)$$

let's write down some terms

$$\begin{aligned} \Pr(\cup A_i) &= p_1 + \ldots + p_n \\ &\quad -(p_1 p_2 + \ldots + p_{n-1} p_n) \\ &\quad +(p_1 p_2 p_3 + \ldots + p_{n-2} p_{n-1} p_n) \\ &\quad - \ldots \end{aligned}$$

Union bound

Let $A_1,..,A_n$ be events in a probability space. Then, we get the following upper bound on the probability of their union.

$$\Pr(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} \Pr(A_i)$$