3050571 Practical Clin Data Sci

Session 2: Computational thinking

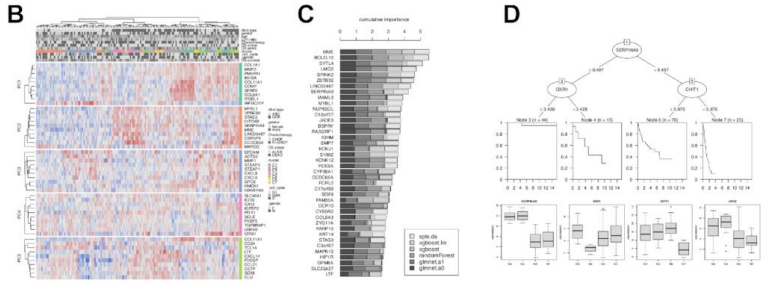
February 1, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

What can you do with computational thinking?

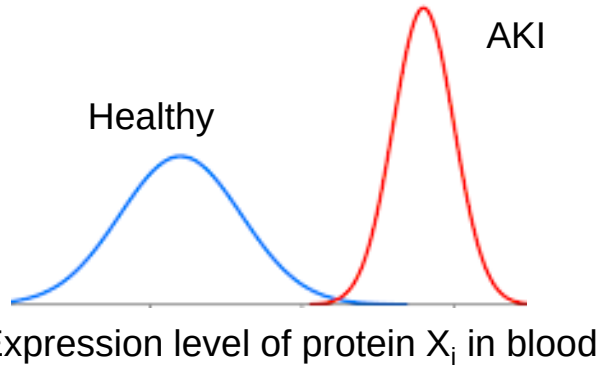
- **Data analysis**
  - Identify calculations that will support or disprove your hypothesis
  - Be aware of assumptions and limitations of each calculation
- Modeling and simulation
- Algorithm



Akhmedov, M. et al. NAR Genom and Bioinfor, 2(1):lqz019 (2019)

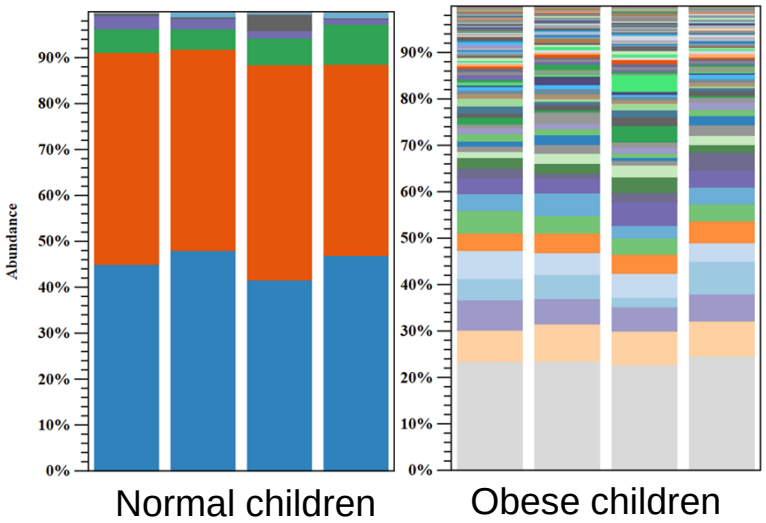
Gene ID	P61_2_C	P62_2_C	P63_2_C	P64_2_C	P68_2_C
ENSG00000000003.14	4.637576	6.183992	5.237635	2.372719	5.665966
ENSG00000000005.5	0	0	0	0	0
ENSG00000000419.12	11.22781	4.813792	2.99782	10.99452	10.7482
ENSG00000000457.13	7.656414	5.082675	7.710682	9.014404	8.488388
ENSG00000000460.16	3.172546	2.245954	5.974815	3.501081	4.162024
ENSG00000000938.12	0	0	0	0.042488	0
ENSG00000000971.15	6.626259	8.19511	5.904925	11.7748	2.050394
ENSG00000001036.13	1.790445	0.76823	3.670635	0.68115	1.894823
ENSG00000001084.11	19.53907	25.08378	11.04872	5.815902	20.23763
ENSG00000001167.14	15.34717	20.00867	17.10001	25.31168	27.41216
ENSG00000001460.17	0.889852	3.090642	0.744581	3.439525	2.417934
ENSG00000001461.16	3.771195	3.12468	1.385353	2.767444	2.973217
ENSG00000001497.16	16.75059	9.662455	15.4965	14.34071	10.62035
ENSG00000001617.11	2.998366	3.712208	3.885852	17.50663	3.019686

Quantifying the “goodness” of biomarkers



- Identify significantly differentially expressed proteins, X_1, X_2, …, X_{200} with *t*-test
- What if we want to identify the **best biomarkers**?
  - How to quantify the ability of a protein to distinguish healthy and AKI?
  - Independent two-sample *t*-test’s statistics $= \frac{\text{Mean}_1 - \text{Mean}_2}{\sqrt{\frac{1}{n}(\text{Variance}_1 + \text{Variance}_2)}}$

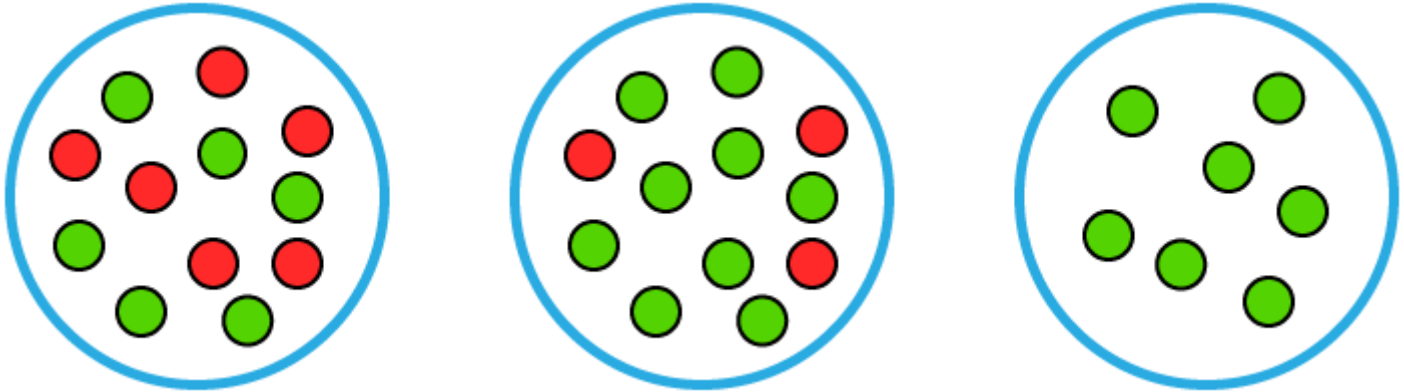
Quantifying diversity of gut microbiomes



Colors indicate different microbial taxa

- Visually, microbial taxa distributions in obese children are clearly more diverse
- Can we quantify this pattern?
  - Number of taxa
  - **Entropy** = $-\text{Frequency}_1 \, \mathbf{log_2}(\text{Frequency}_1) - \cdots - \text{Frequency}_n \, \mathbf{log_2}(\text{Frequency}_n)$

Entropy quantifies purity of a mixture

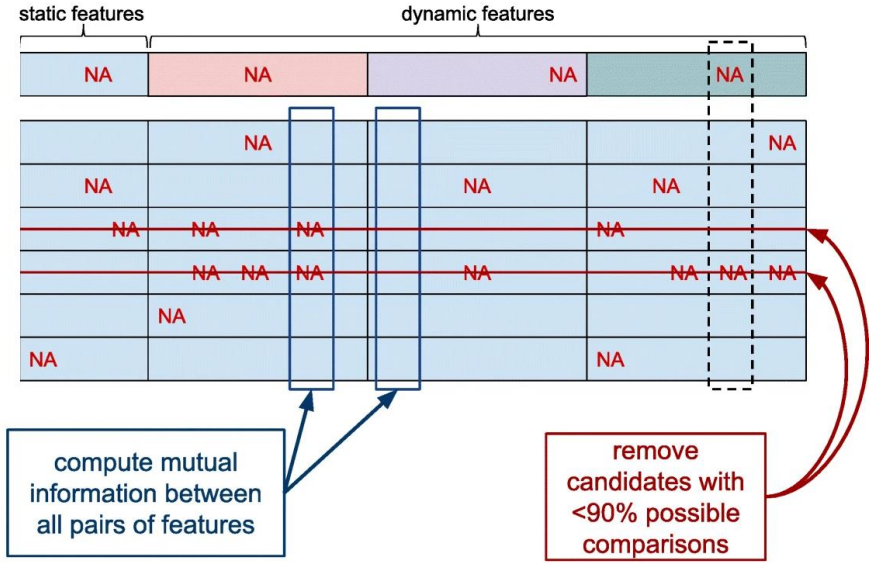


https://www.javatpoint.com/entropy-in-machine-learning

- **Entropy** = $-\text{Freq}_1 \log_2(\text{Freq}_1) - \cdots - \text{Freq}_n \log_2(\text{Freq}_n)$
  - Left sample: $-\frac{6}{12}\log_2\left(\frac{6}{12}\right) - \frac{6}{12}\log_2\left(\frac{6}{12}\right) = \log_2(2) = 1$
  - Middle sample: $-\frac{3}{12}\log_2\left(\frac{3}{12}\right) - \frac{9}{12}\log_2\left(\frac{9}{12}\right) = 0.811278$
  - Right sample: $-\frac{0}{12}\log_2\left(\frac{0}{12}\right) - \frac{12}{12}\log_2\left(\frac{12}{12}\right) = 0$

Mutual Information ~ correlation for categorical data

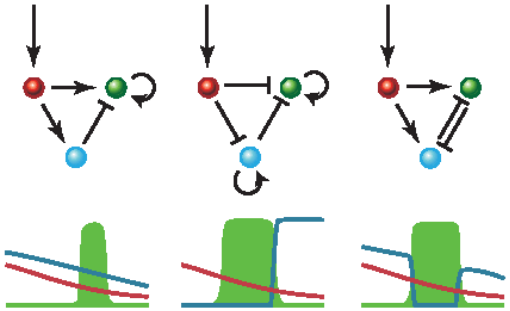
- MI = difference between P(X, Y) and P(X) P(Y)
- If X and Y are statistically independent
  - P(X, Y) = P(X) P(Y)
  - MI = 0
- $\mathrm{MI}(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

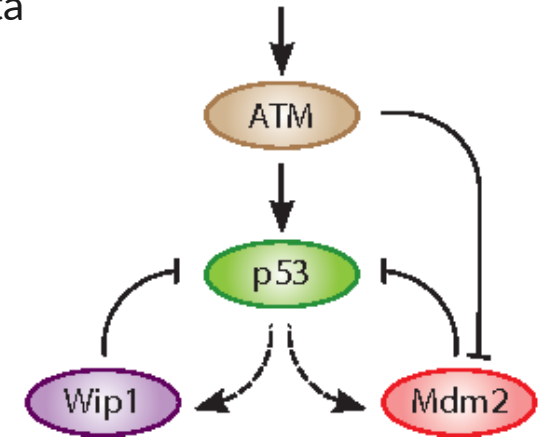


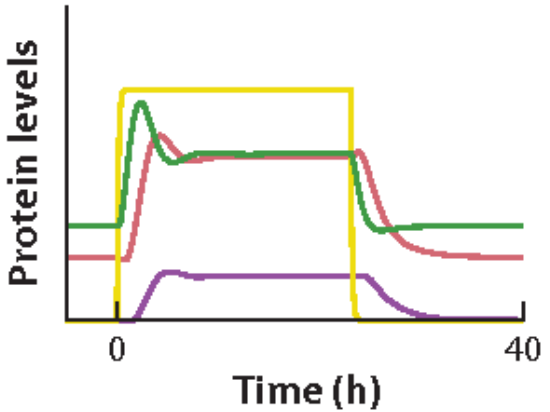
Tavazzi, E. et al. BMC Medical Informatics and Decision Making 2020

What can you do with computational thinking?

- Data analysis
- **Modeling and simulation**
  - Identify mechanisms that underlie the phenomenon or system of interest
  - Develop models
  - Study the behavior of the models / systems
  - Synthesize new data
- Algorithm

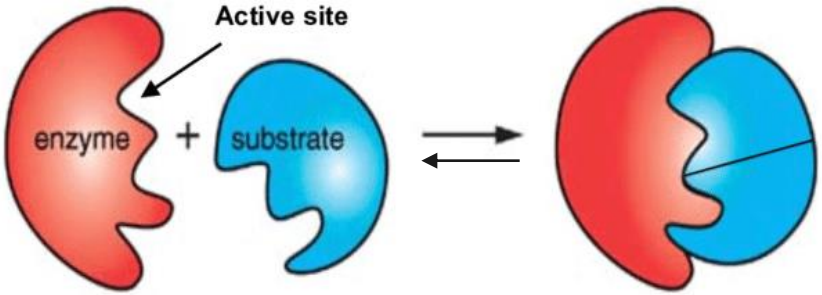






Rue and Garcia-Ojalvo. Annu Rev Biophys 42:605-627 (2013)

Enzyme substrate binding



Graham Hutchings. “Development of new highly active nano gold catalysts for selective oxidation reactions” (2014)

- $K_{dissociation}$ = [E][S] / [E-S] at equilibrium

- Association = P(E meeting S) x $P_{binding}$ = k_1 [E][S]
- Dissociation = Number of E-S molecules x $P_{dissociating}$ = k_2 [E-S]
- At equilibrium, Association = Dissociation
  - k_1 [E][S] = k_2 [E-S] → $K_{dissociation}$ = k_2 / k_1

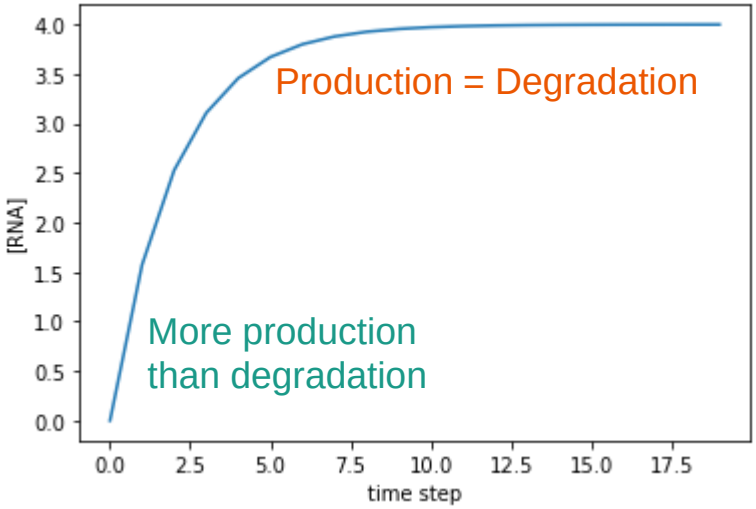
A simple gene expression model

$k_{transcription}$

$$DNA \rightarrow DNA + RNA$$

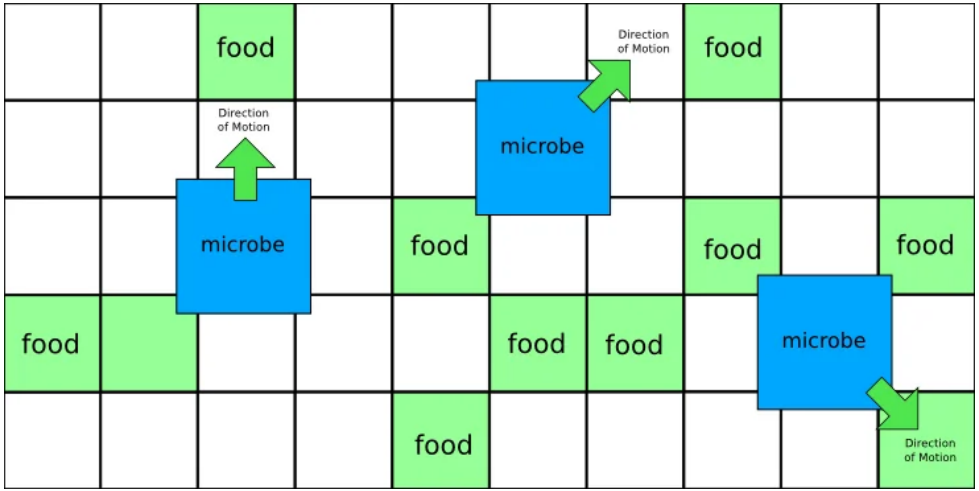
$$RNA \rightarrow \emptyset$$

$k_{degradation}$

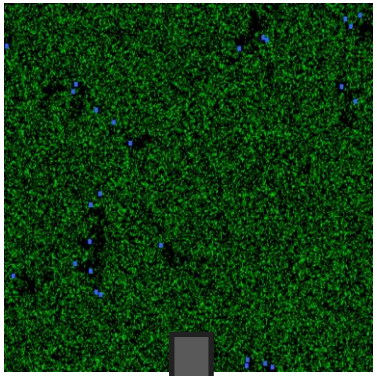


- $\frac{d[RNA]}{dt} = k_{transcription} - k_{degradation}[RNA]$
- This is called an **ordinary differential equation**

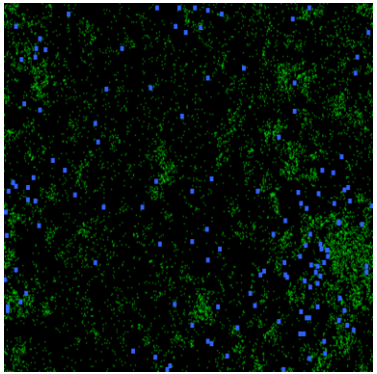
Microbial growth



Time = 0



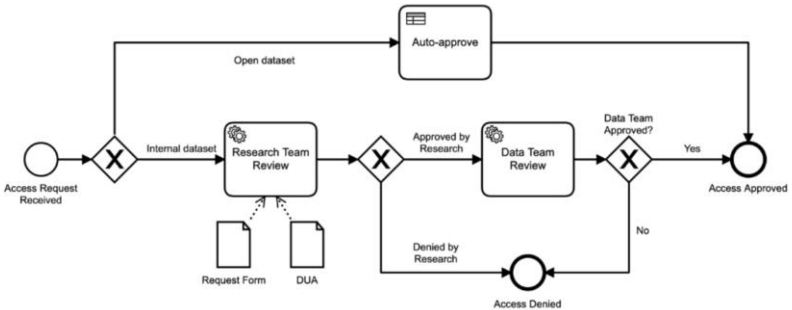
Time = 100



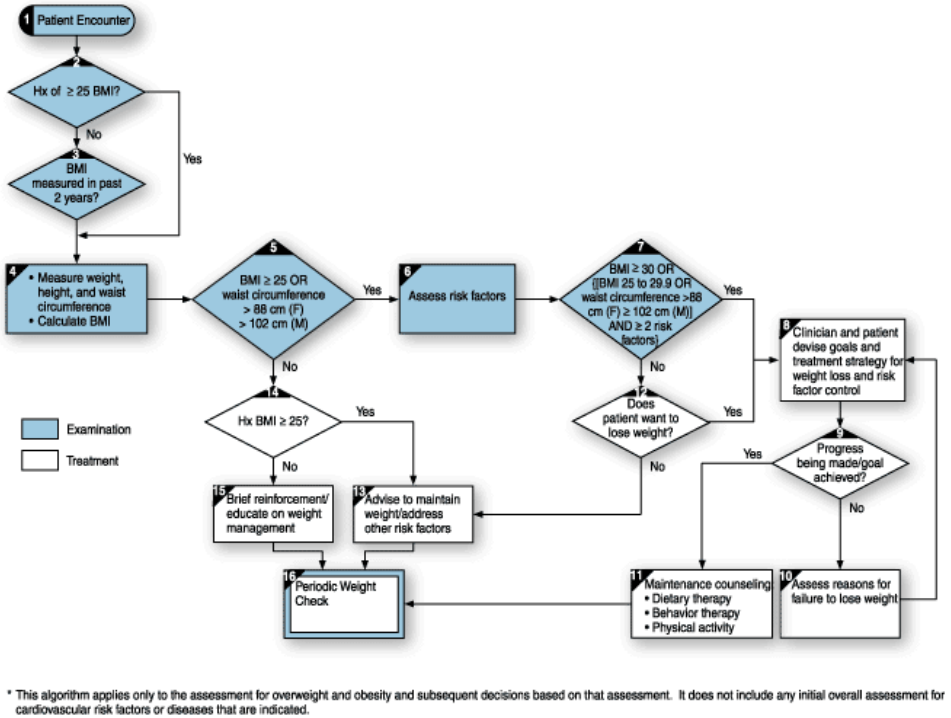
- Model different movement behaviors (probability of moving in a certain direction) with different genes
- See more at **https://beltoforion.de/en/simulated_evolution/**

What can you do with computational thinking?

- Data analysis
- Modeling and simulation
- **Algorithm**
  - Formulating step-by-step instructions
  - Identifying weak points



https://www.aridhia.com/blog/accessibility/



https://en.wikipedia.org/wiki/Medical_algorithm

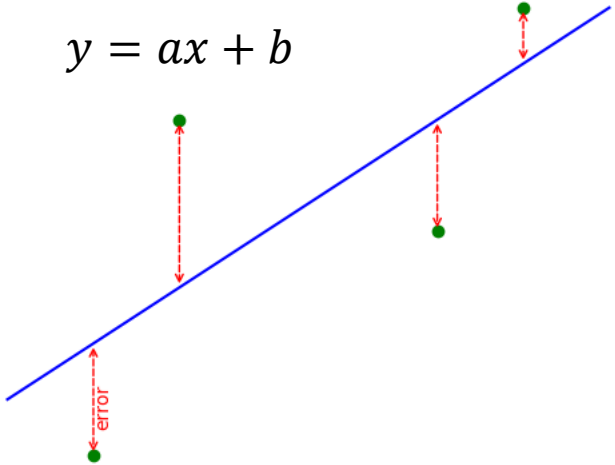
Dynamic programming

No. of rocks	1	2	3	4	5	6	7	8	9	10
Winner										
No. of rocks	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Winner										?

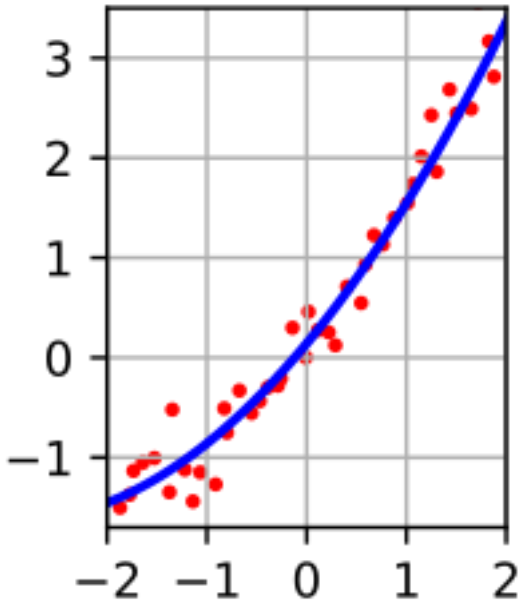
- Dynamic programming build the solution of complex problem using on the solutions of simpler ones
- **There is a pile of 20 rocks. Two players take turns by removing 1 or 2 rocks from the pile. Whoever removes the last rock(s) win.** Who is the winner?

Everything is optimization

Curve fitting with least square



https://en.wikipedia.org/wiki/Least_squares

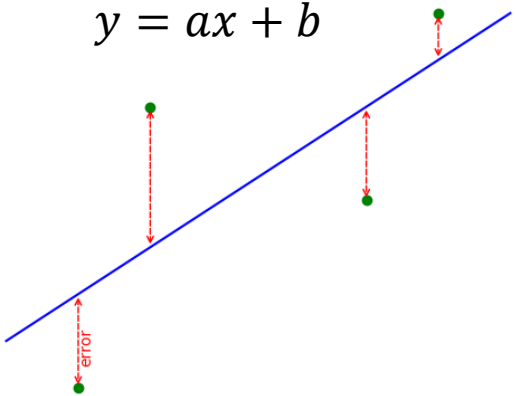


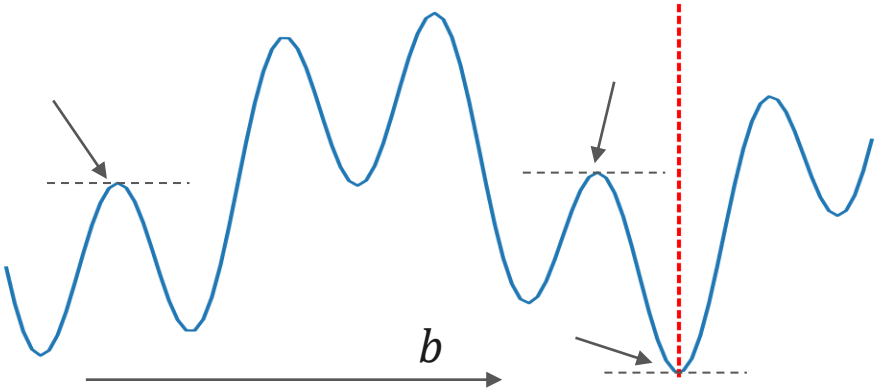
$y = ax^2 + bx + c$

- Finding the **best** *a*, *b*, and *c* that make the curve fit to the observations
- Minimize least square error $\frac{1}{n}\left((\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2\right)$

How can we find the best parameters

- Minimize $\mathrm{L}(a,b) = \frac{1}{n}\left((\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2\right)$
- Randomly search for (a, b)
  - Heuristic guess: $a = \frac{\sum y_i}{\sum x_i}$
- Use calculus
  - At optimal, slopes are zero
  - Solve $\frac{\partial \mathrm{L}(a,b)}{\partial a} = 0$ and $\frac{\partial \mathrm{L}(a,b)}{\partial b} = 0$





Newton-Ralphson method

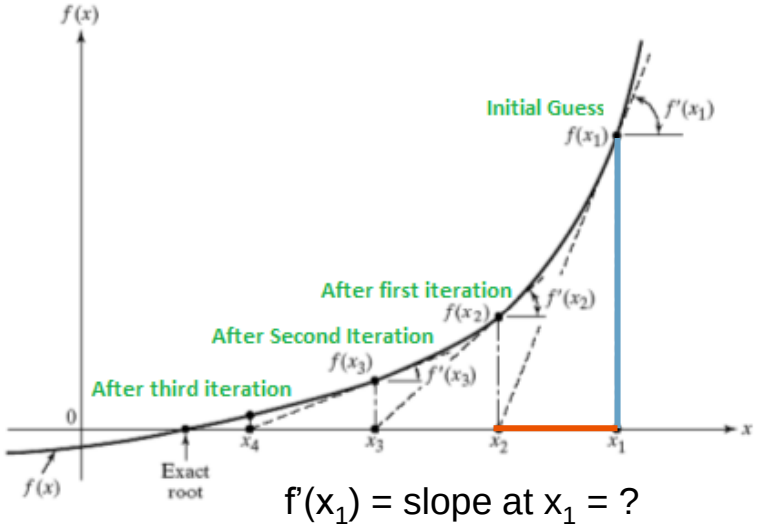
- Want to maximize L(x) by solving

$$f(x) = \frac{dL(x)}{dx} = 0$$

- Start with an initial guess x_1
- Calculate $f(x)$ and $f'(x)$ at x_1
- Define $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$

- Repeat the process for x_3, x_4, …
- Stop when x_i converges to a value

f(x) = the first derivative of the objective L(x)



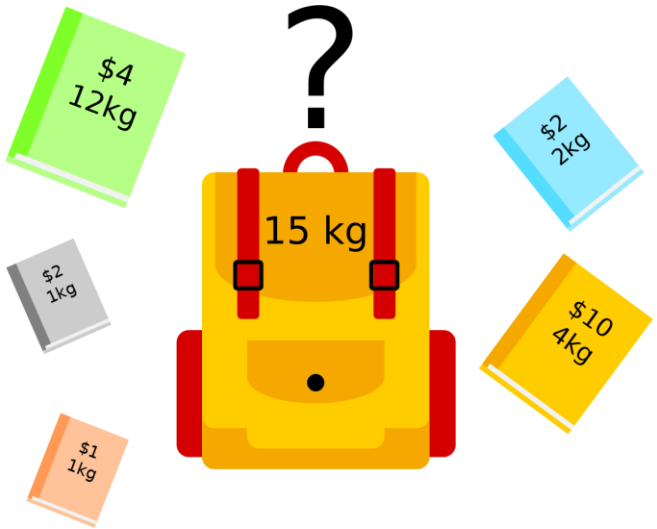
f'(x_1) = slope at x_1 = ?

https://predictivehacks.com/newton-raphson-method-in-python/

(Almost) all algorithms involve optimization

- **Curve fitting** = minimize square error
- **Drug-protein docking** = minimize energy
- **Hospital queuing** = minimize wait time
- **Diagnosis** = maximize accuracy
- **Triaging** = maximize number of severe patients at the top of the list

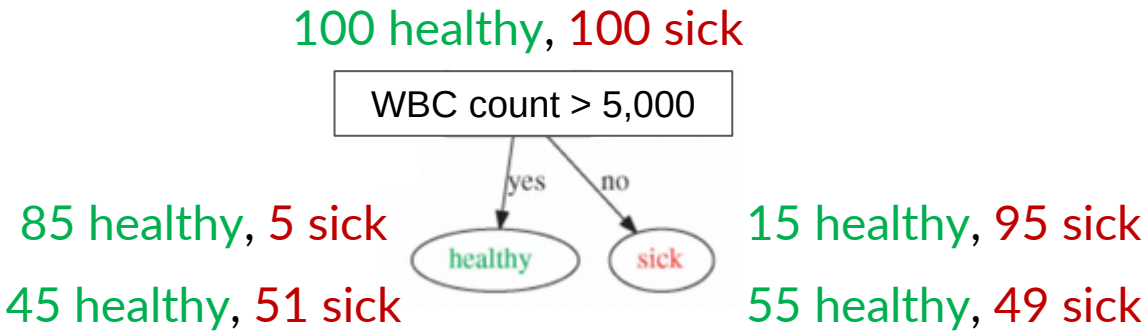
Knapsack problem



https://en.wikipedia.org/wiki/Knapsack_problem

- Limited capacity
- Want to pick as much value as possible while not exceeding the capacity
- General optimization setting
  - **Objective** = Total value
  - **Constraint** = Total weight ≤ W
- Paying **X** baht with the minimal number of coins and bank notes

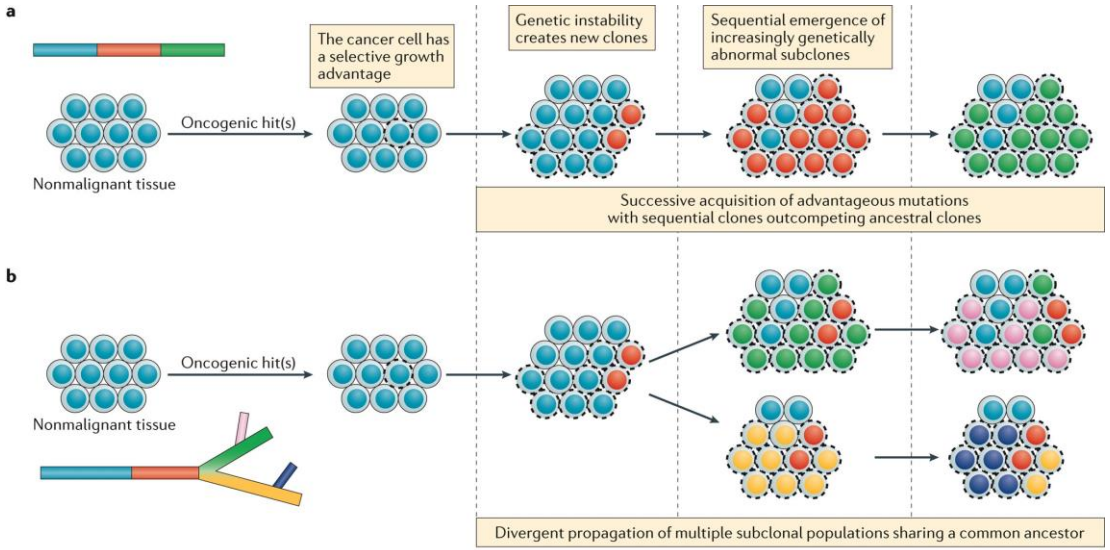
Constructing a decision point



- Gini impurity: $\sum p(1-p)$
- Entropy: $-\sum p \ln(p)$
  - Minimal at p = 0 or 1 → Perfect split
  - Maximal at p = 0.5 → 50-50 split
- **Search for variable and cutoff** that strongly reduce impurity

Simulation

Tumor growth

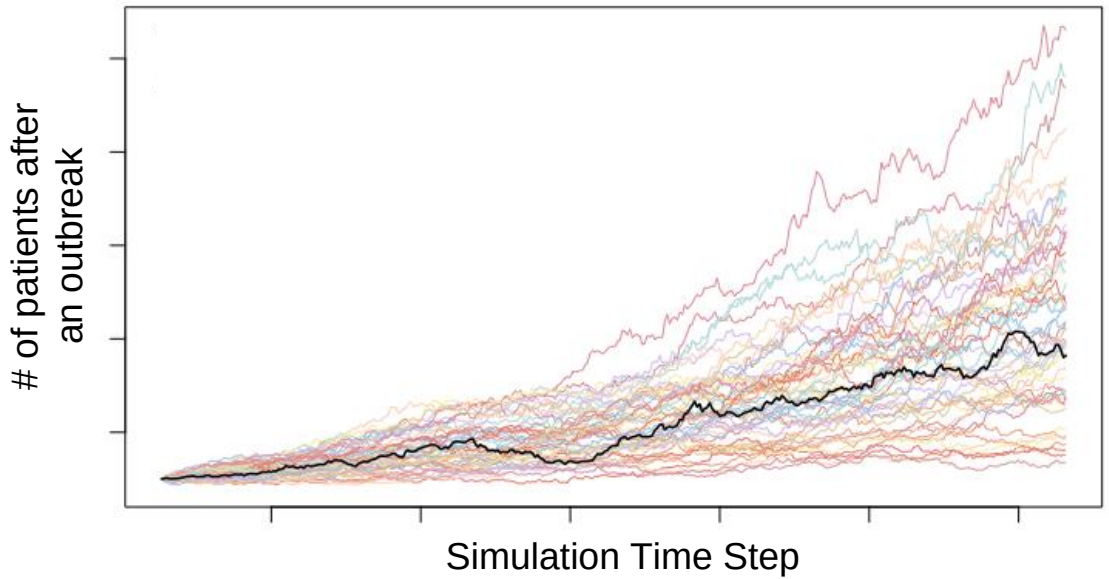


Dagogo-Jack and Shaw, Nat Rev Clin Oncol (2017)

Nature Reviews | Clinical Oncology

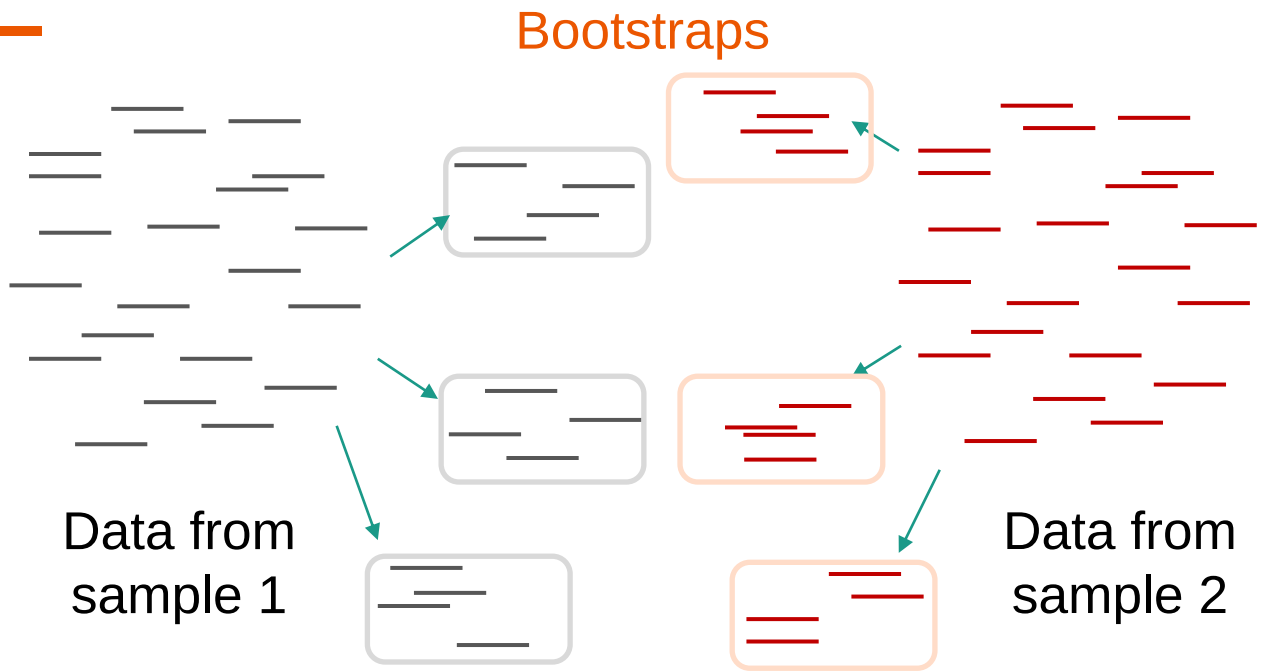
- Cell's actions: Gain mutation, Divide, Die
  - What are the parameters influencing these actions?

Monte Carlo technique



- Perform repeated sampling to explore broad parameter scenarios
- Provide estimates for the probability of different scenarios and outcomes
- May require millions of repetition (can utilize multiple CPUs)

Bootstrapping



- Instead of using the whole data to perform one calculation
- Sampling from data to perform multiple calculations → Estimate variance

Summary

- Computational thinking is about **transforming abstract hypothesis into a quantitative statement** that can be tested with data
- **Identify relevant components and mechanisms** that underlie the phenomenon and **develop the models**
- **Fit the models or simulate the phenomenon** to derive knowledge
- Every algorithm boils down to **an optimization problem**

Let’s do some practices

Apply computational thinking principles

- Reduce hospital deficit
- Reduce patient wait time
- Predict future bone fracture in elderly patients
- Predict future admission of home isolation patients
- Improve OR scheduling
- Reduce prescription error in pharmacy
- Alert doctors when patients visit other clinics
- Identify potential viral outbreak
- Triaging patients for ICU
- Manage mobile beds and wheelchairs

Any questions?

See you on February 2nd