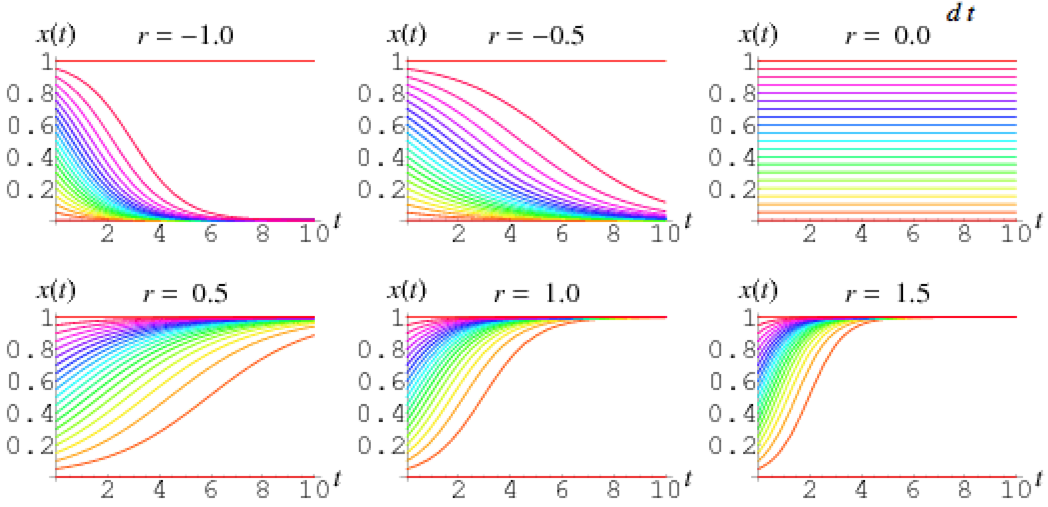
Informatics, algorithmics, and automation: machine learning in 21^{st} century biology

Pradipta Ray
UT Dallas

Mathematics in biology & medicine

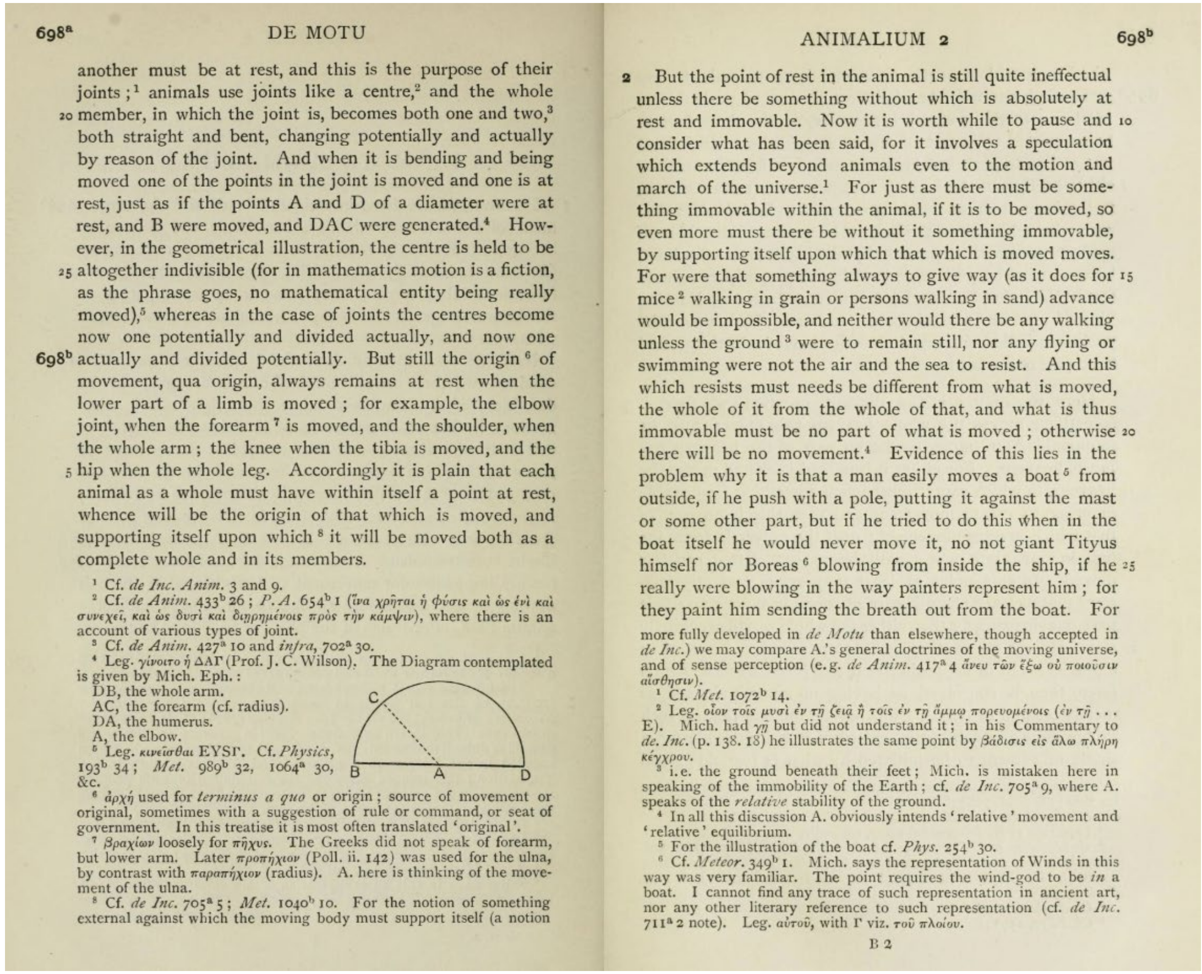
- A very long history – as early as the Hellenic civilization
- All branches of mathematics are harnessed – from classical geometry to calculus

$$\frac{dN}{dt} = \frac{rN\,(K-N)}{K},$$



Wolfram

Plots of the logistic equation modelling population growth



698a DE MOTU

another must be at rest, and this is the purpose of their joints;[1] animals use joints like a centre,[2] and the whole member, in which the joint is, becomes both one and two,[3] both straight and bent, changing potentially and actually by reason of the joint. And when it is bending and being moved one of the points in the joint is moved and one is at rest, just as if the points A and D of a diameter were at rest, and B were moved, and DAC were generated.[4] However, in the geometrical illustration, the centre is held to be altogether indivisible (for in mathematics motion is a fiction, as the phrase goes, no mathematical entity being really moved),[5] whereas in the case of joints the centres become now one potentially and divided actually, and now one
698b actually and divided potentially. But still the origin[6] of movement, qua origin, always remains at rest when the lower part of a limb is moved; for example, the elbow joint, when the forearm[7] is moved, and the shoulder, when the whole arm; the knee when the tibia is moved, and the hip when the whole leg. Accordingly it is plain that each animal as a whole must have within itself a point at rest, whence will be the origin of that which is moved, and supporting itself upon which[8] it will be moved both as a complete whole and in its members.

[1] Cf. *de Inc. Anim.* 3 and 9.
[2] Cf. *de Anim.* 433^b 26; *P.A.* 654^b 1 (ἵνα χρῆται ἡ φύσις καὶ ὡς ἑνὶ καὶ συνεχεῖ, καὶ ὡς δυσὶ καὶ διῃρημένοις πρὸς τὴν κάμψιν), where there is an account of various types of joint.
[3] Cf. *de Anim.* 427^a 10 and *infra*, 702^a 30.
[4] Leg. γίνοιτο ἡ ΔΑΓ (Prof. J. C. Wilson). The Diagram contemplated is given by Mich. Eph.:
DB, the whole arm.
AC, the forearm (cf. radius).
DA, the humerus.
A, the elbow.
[5] Leg. κινεῖσθαι EYSΓ. Cf. *Physics*, 193^b 34; *Met.* 989^b 32, 1064^a 30, &c.
[6] ἀρχή used for *terminus a quo* or origin; source of movement or original, sometimes with a suggestion of rule or command, or seat of government. In this treatise it is most often translated 'original'.
[7] βραχίων loosely for πῆχυς. The Greeks did not speak of forearm, but lower arm. Later προπήχιον (Poll. ii. 142) was used for the ulna, by contrast with παραπήχιον (radius). A. here is thinking of the movement of the ulna.
[8] Cf. *de Inc.* 705^a 5; *Met.* 1040^b 10. For the notion of something external against which the moving body must support itself (a notion

ANIMALIUM 2 698b

2 But the point of rest in the animal is still quite ineffectual unless there be something without which is absolutely at rest and immovable. Now it is worth while to pause and consider what has been said, for it involves a speculation which extends beyond animals even to the motion and march of the universe.[1] For just as there must be something immovable within the animal, if it is to be moved, so even more must there be without it something immovable, by supporting itself upon which that which is moved moves. For were that something always to give way (as it does for mice[2] walking in grain or persons walking in sand) advance would be impossible, and neither would there be any walking unless the ground[3] were to remain still, nor any flying or swimming were not the air and the sea to resist. And this which resists must needs be different from what is moved, the whole of it from the whole of that, and what is thus immovable must be no part of what is moved; otherwise there will be no movement.[4] Evidence of this lies in the problem why it is that a man easily moves a boat[5] from outside, if he push with a pole, putting it against the mast or some other part, but if he tried to do this when in the boat itself he would never move it, no not giant Tityus himself nor Boreas[6] blowing from inside the ship, if he really were blowing in the way painters represent him; for they paint him sending the breath out from the boat. For

more fully developed in *de Motu* than elsewhere, though accepted in *de Inc.*) we may compare A.'s general doctrines of the moving universe, and of sense perception (e.g. *de Anim.* 417^a 4 ἄνευ τῶν ἔξω οὐ ποιοῦσιν αἴσθησιν).
[1] Cf. *Met.* 1072^b 14.
[2] Leg. οἷον τοῖς μυσὶ ἐν τῇ ζειᾷ ἢ τοῖς ἐν τῇ ἄμμῳ πορευομένοις (ἐν τῇ . . . E). Mich. had γῇ but did not understand it; in his Commentary to *de. Inc.* (p. 138. 18) he illustrates the same point by βάδισις εἰς ἅλω πλήρη κέγχρου.
[3] i.e. the ground beneath their feet; Mich. is mistaken here in speaking of the immobility of the Earth; cf. *de Inc.* 705^a 9, where A. speaks of the *relative* stability of the ground.
[4] In all this discussion A. obviously intends 'relative' movement and 'relative' equilibrium.
[5] For the illustration of the boat cf. *Phys.* 254^b 30.
[6] Cf. *Meteor.* 349^b 1. Mich. says the representation of Winds in this way was very familiar. The point requires the wind-god to be *in* a boat. I cannot find any trace of such representation in ancient art, nor any other literary reference to such representation (cf. *de Inc.* 711^a 2 note). Leg. αὐτοῦ, with Γ viz. τοῦ πλοίου.

B 2

On the motion of animals, Aristotle, 4th century BC (translated 1912)

Informatics : what aspects of mathematics does it deal with ?

- **Mathematics pertaining to data**
  - Automated description, modelling, visualization, prediction (and imputation), etc
- Large overlaps with Computer Science (machine learning) and Statistics community

SEG-NACHRICHTEN 1957 HEFT 4 122 171

INFORMATIK: Automatische Informationsverarbeitung

Von K. Steinbuch · Standard Elektrik AG, Informatikwerk

DK 681.14-523.8

1. Allgemeines

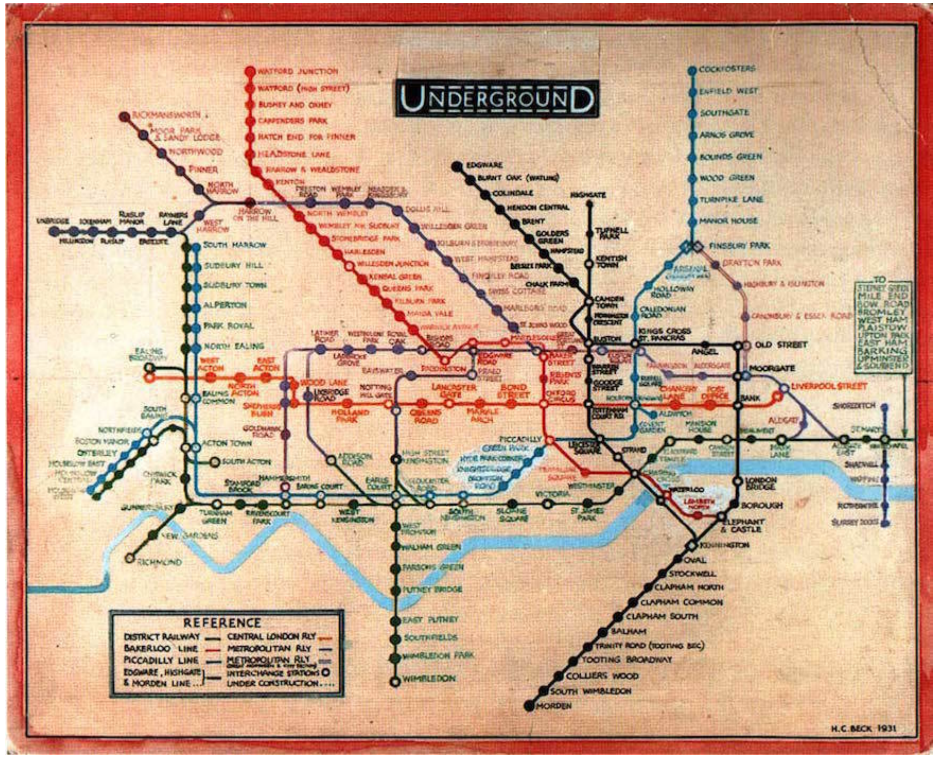
Die elektrische Nachrichtentechnik hatte bis vor wenigen Jahren eine einzige Aufgabe: die Übertragung von Signalen über räumliche Entfernungen hinweg. Zum Beispiel übertragen Telegrafiesysteme die Fernschreibsignale über Tausende von Kilometern; Telefoniesysteme ermöglichen Sprechverbindung mit Amerika oder Australien, und Funkverbindungen tragen das Fernsehbild über Kontinente hinweg in unsere Wohnung. In all diesen Beispielen ist die Überwindung der räumlichen Entfernung die wesentliche Leistung.

Vor etwa zwanzig Jahren entdeckten Ingenieure in USA und Deutschland unabhängig voneinander, daß die Verfahren der Nachrichtentechnik auch für andere Aufgaben nützlich sind, Aufgaben, bei denen die Überwindung der räumlichen Entfernung ganz unwesentlich ist. Sie fanden, daß man mit elektrischen Schaltungen Zahlenrechnungen durchführen kann, und zwar mit einer Schnelligkeit, wie sie bis dahin einfach unvorstellbar war. Damit begann die automatische Informationsverarbeitung. Wir nennen sie „INFORMATIK".

Abb. 3 Aufbau eines Rechenplatzes

darstellen, entweder auf mehreren Drähten gleichzeitig, oder auf einem einzigen Draht nacheinander angeordnet sein können (Parallel- bzw. Seriendarstellung).

Term first coined in 1957 though discipline dates back to ancient India / China / Greece



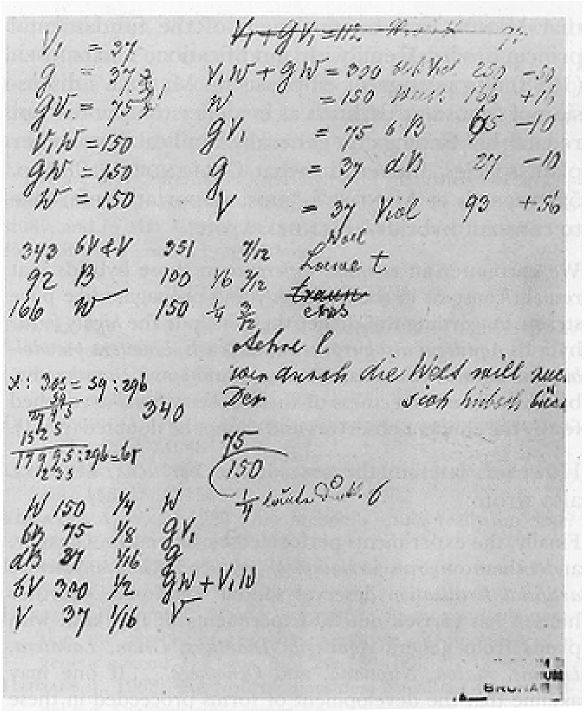
Harry Beck's 1931 visualization of London's tube network

Beck was a draftsman for the underground railway before being fired and voluntarily creating map & selling to ex-employer for 10 GBP !

Algorithms in biology

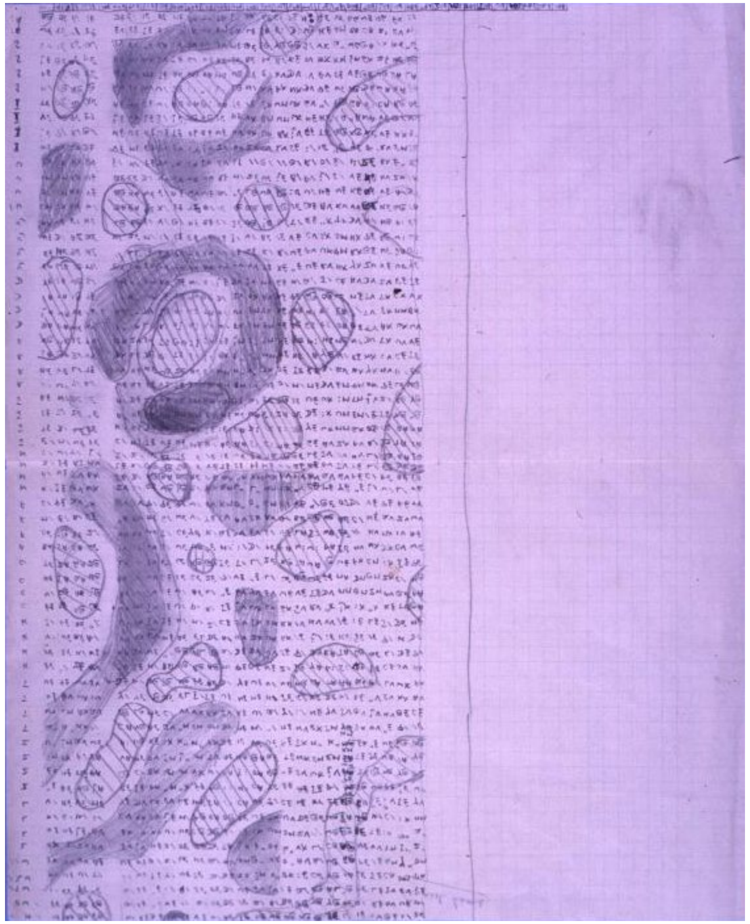
- Algorithm : a set of instructions for performing a task / computation
  - Set of rules to navigate a car from school to home
  - Computer program for sorting number

Computational Biology
Use mathematics & CS to answer questions in biology – which traits of a pea plant are inherited independently of others and at what frequency ?



Gregor Mendel's notebook detailing his experiments with breeding peas

Computation in biology
What are the underlying instructions driving biological processes – how do molecular gradients give rise to morphological patterns ?



Alan Turing's notebook sketches for modelling how pattern formation occurs in nature

The dawn of computers

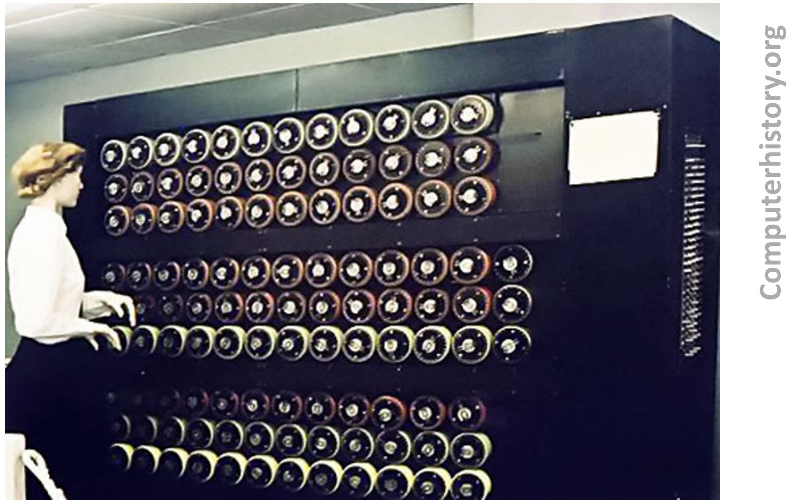
- Algorithm execution and informatics were laborious tasks before computers came about

- The invention of electronic computers and dizzying improvements in speed and memory led to most informatics tasks being performed on computers

Computers + algorithmics + informatics = Machine learning

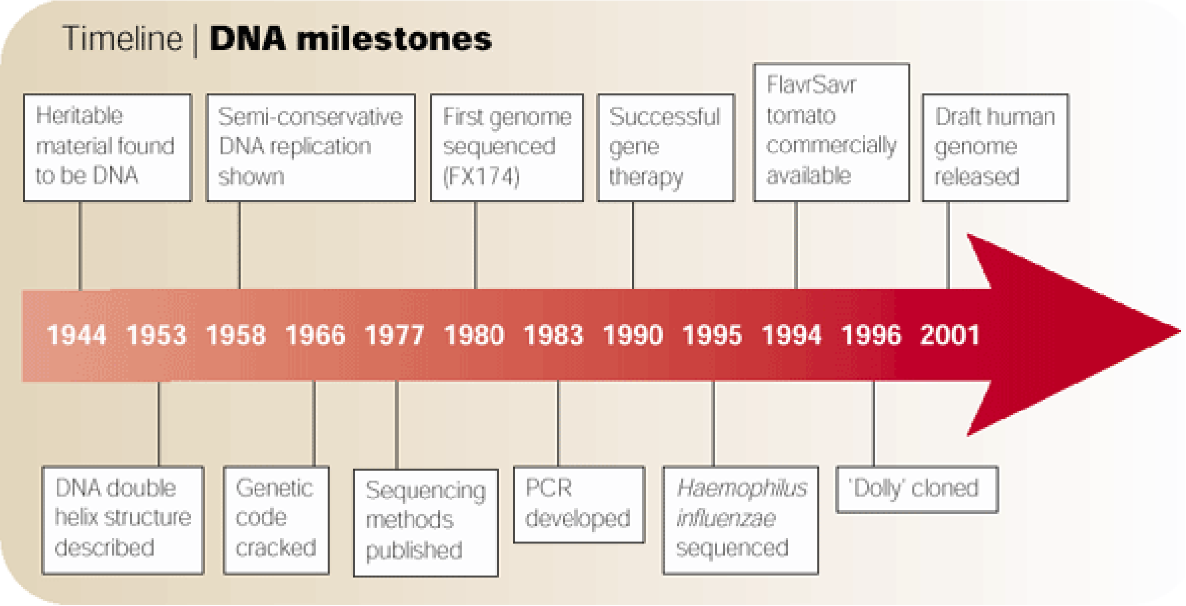
Computers + algorithmics + informatics + biology = Bio-informatics

Radhanath Sikdar, the human "computer" who first calculated the height of Mt Everest

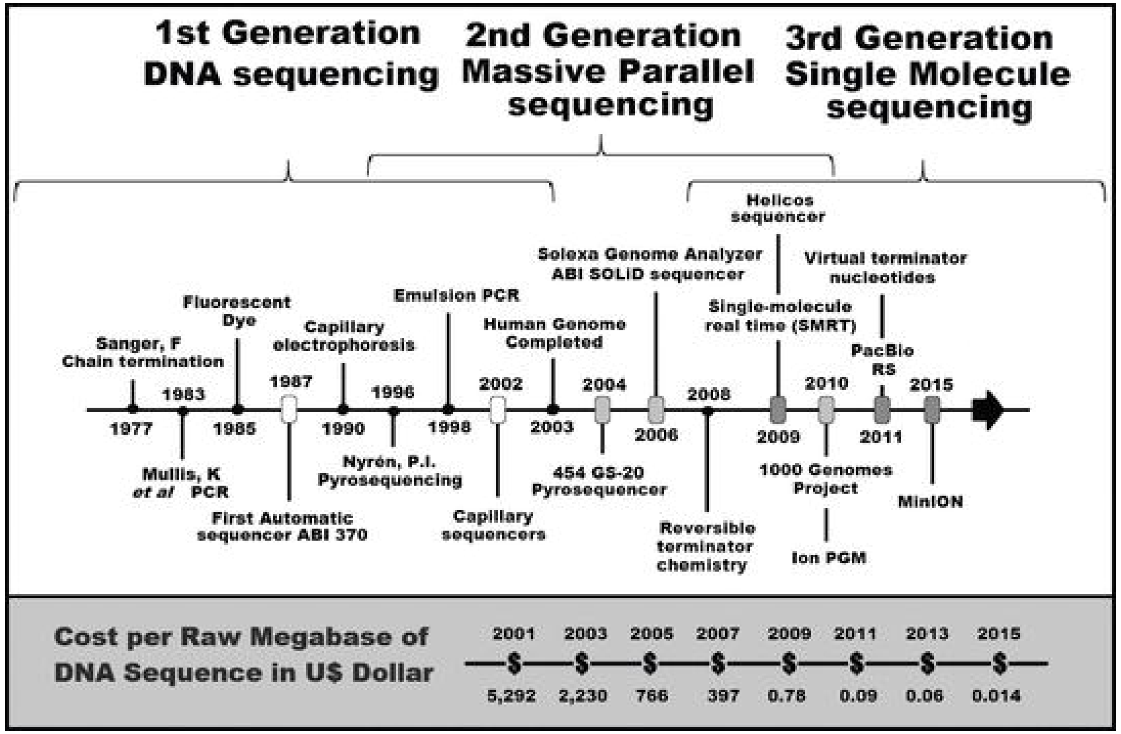


The Bombe (1942), one of the first large scale electromechanical devices built in Bletchley Park for British WW2 cryptanalysis

Nucleic acid research : understanding, sequencing, engineering

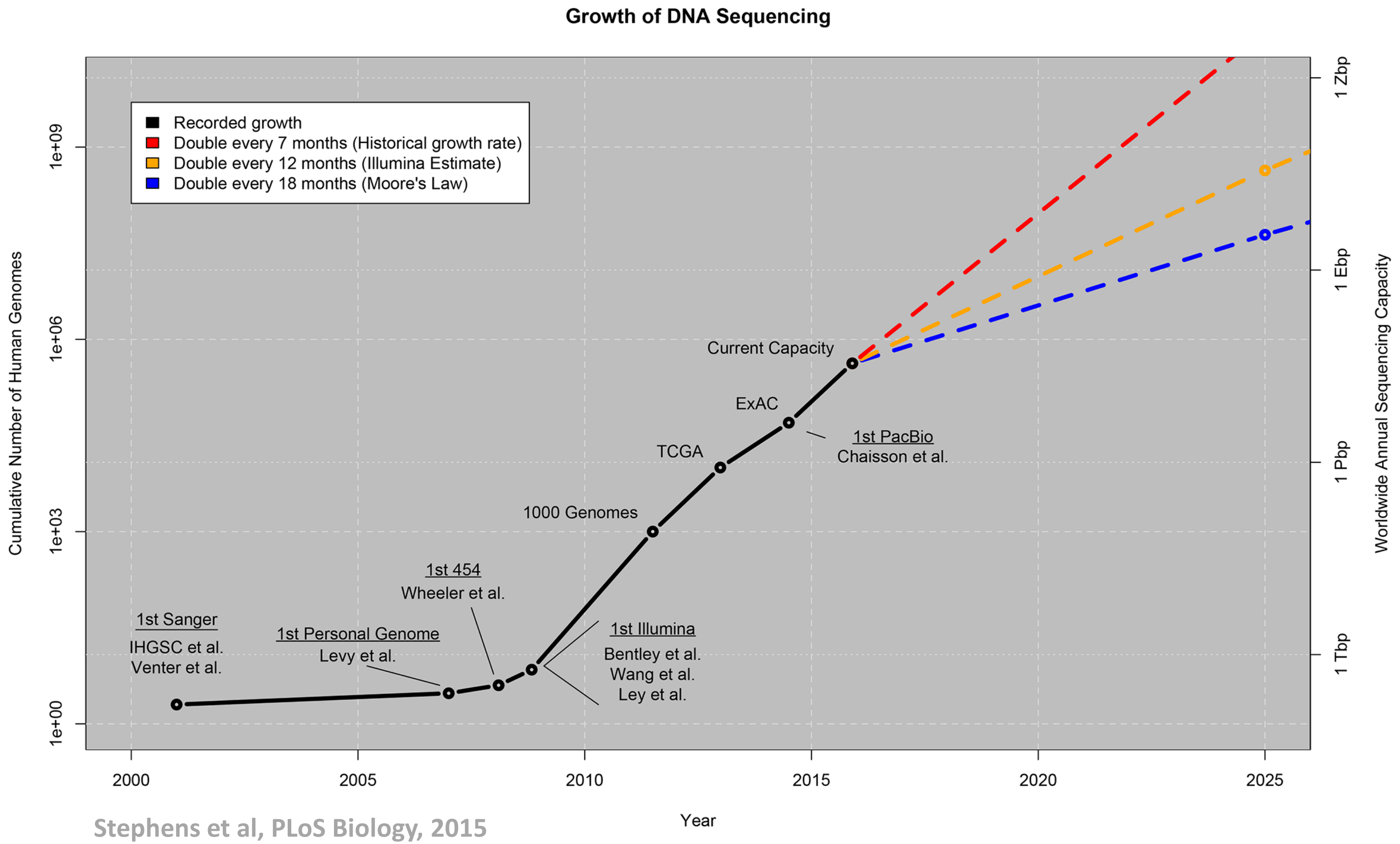


Baxter, Nature, 2003



Application of Next-Generation Sequencing in the Era of Precision Medicine, Pereira et al

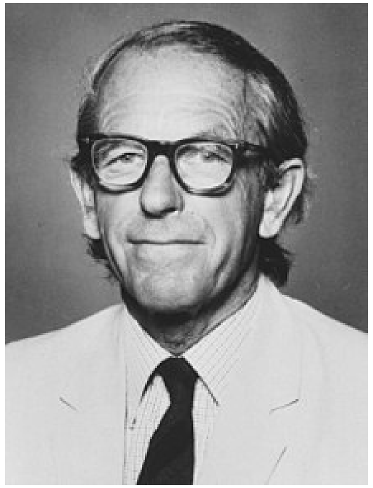
Automation and scale in nucleic acid sequencing



Stephens et al, PLoS Biology, 2015

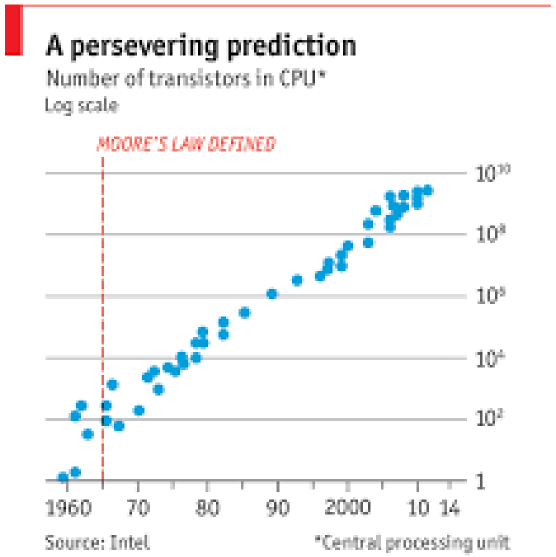
Bioinformatics : 3 decades of explosive growth

Wikipedia



Frederick Sanger : Nobel prize winner (twice!!) for studies on sequence and structure of insulin and for nucleic acid sequencing

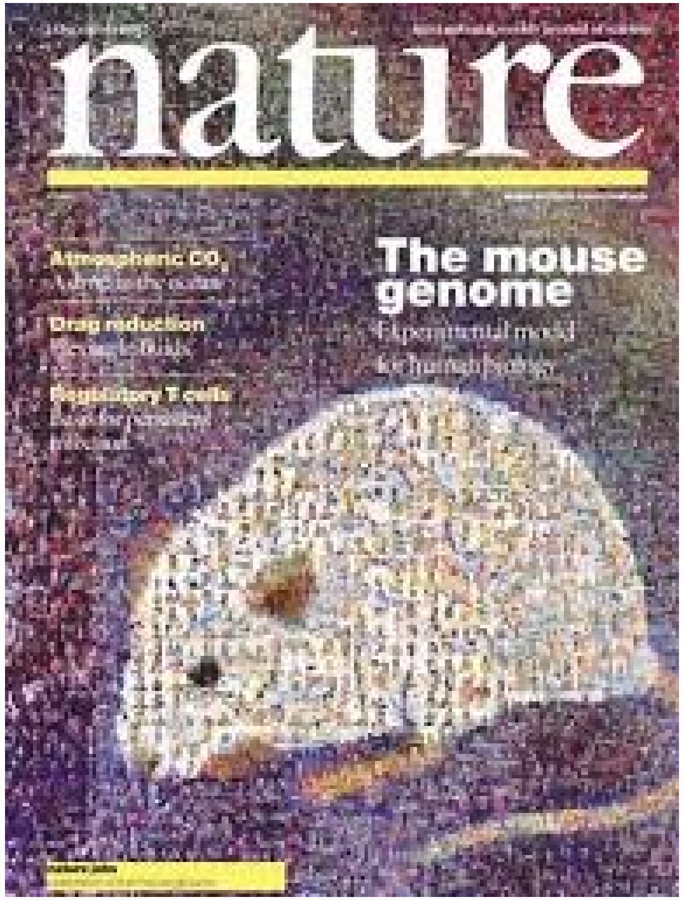
- **Rapid improvements in molecular sequencing technologies** for peptides, DNA and RNA produces large amounts of data to be analyzed
- Moore's Law : computers tend to double in computing capacity every year, making **intensive computation feasible, tractable and economical**
- Development of methodology, training a new kind of multidisciplinary scientist, investment in infrastructure – the **Human Genome Project** (1990 – 2003)

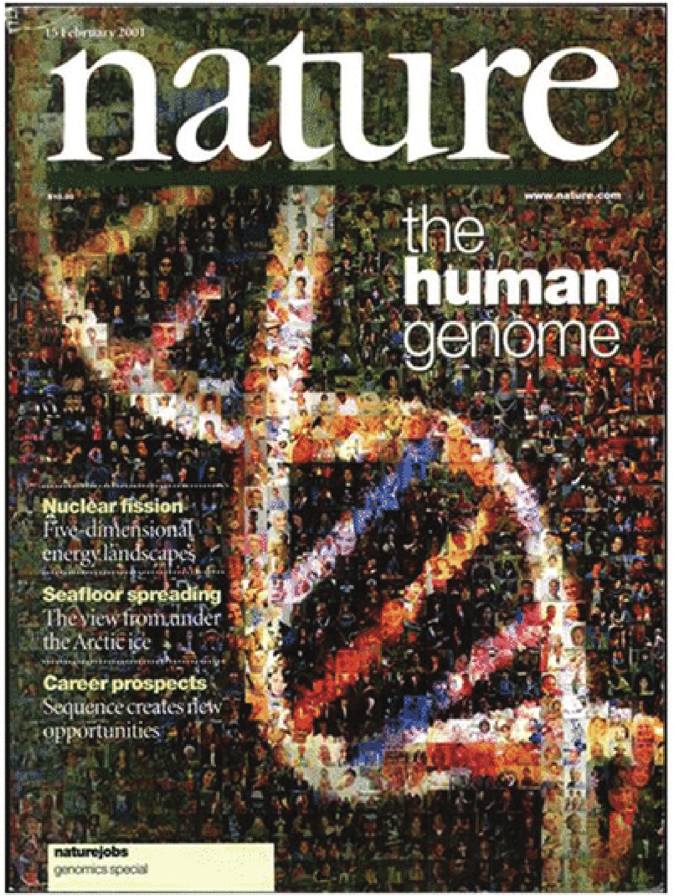


Economist.com

The mouse & human genome projects

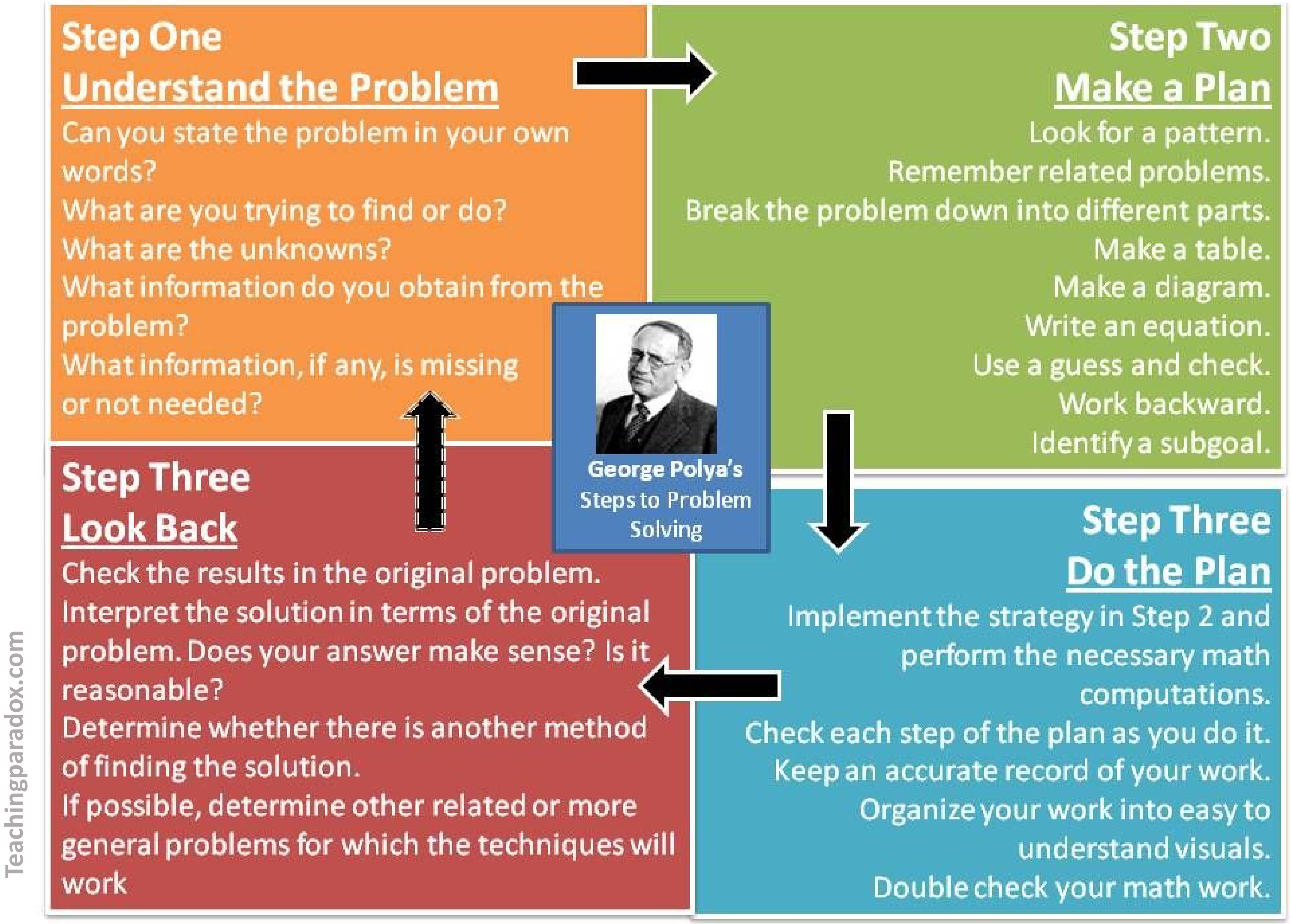
- Sequenced and assembled reference drafts of the human and mouse genome
- Mapped and predicted genes in the reference sequence
- Laid the groundwork for genomics (including computational genomics) as a discipline





Asking the right questions

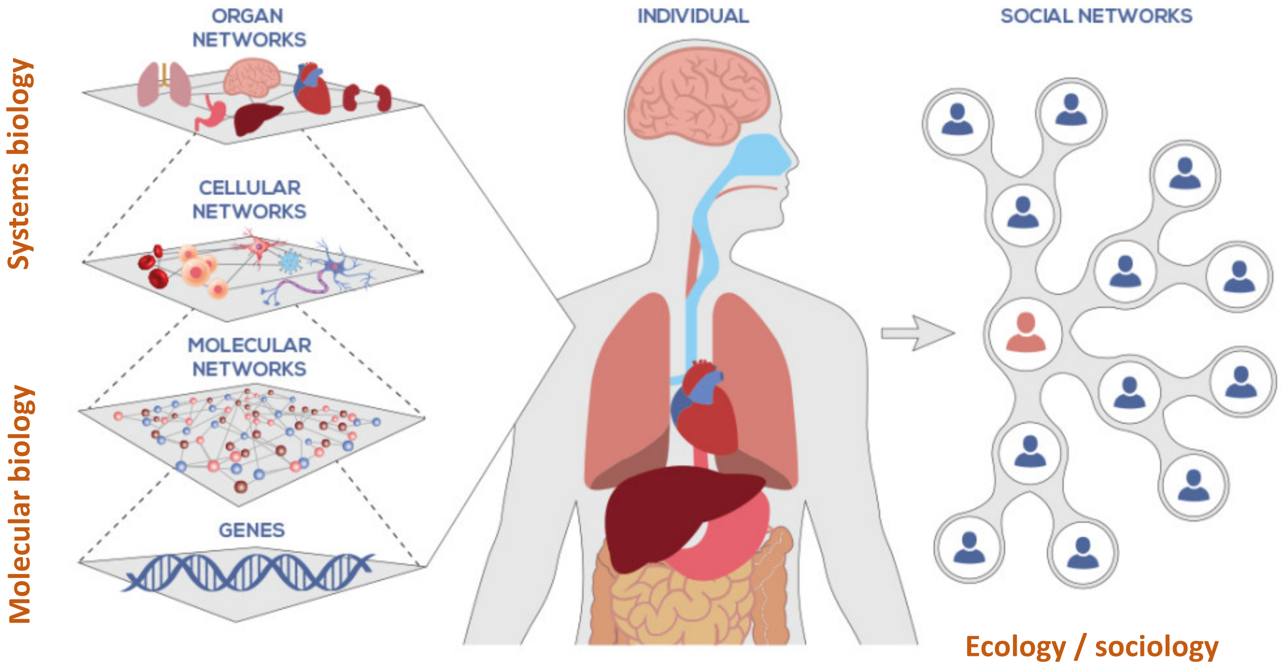
- How to solve it, written by George Polya
  - Strongly encouraged to read this book !
- Problem solving is a cyclic process : true for bioinformatics as well



Performing analysis at the right granularity

- Given a biological problem, it is important to figure out the right granularity to perform experiments / note observations

- Some problems can be studied at multiple levels : disease - molecular pathology, systems level changes in physiology, transmission of disease

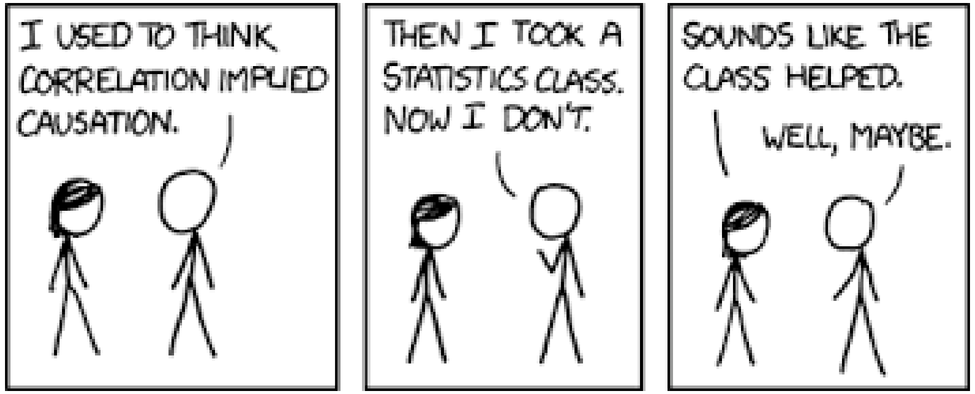


Institute for Systems Biology

Bioinformatics as an epidemiological science

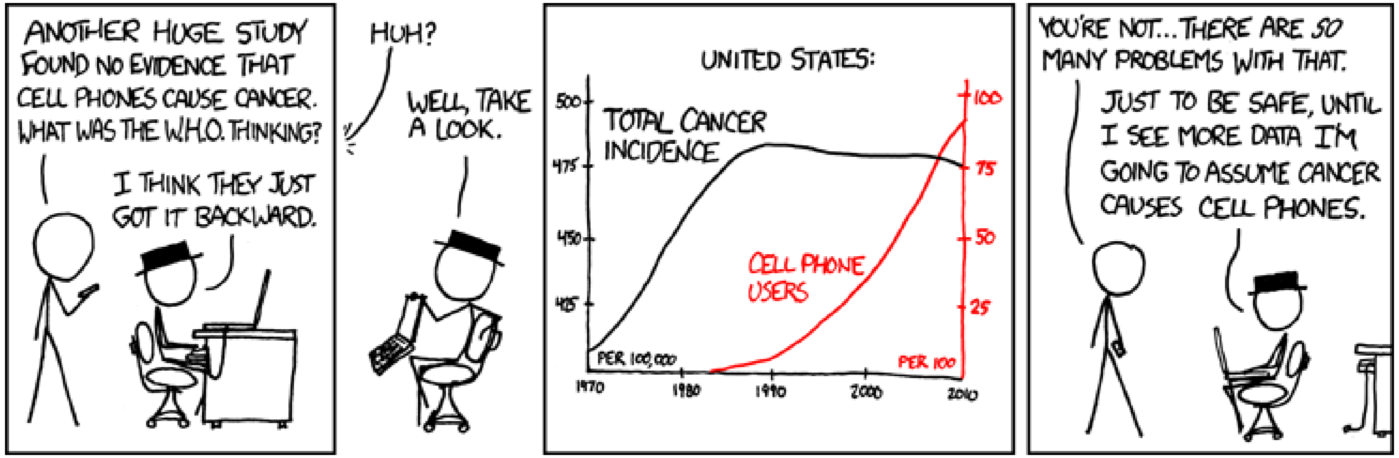
- A large amount of biology is an empirical science
  - Bioinformatics helps generate, screen and test hypotheses in high throughput fashion to help transition it to a theoretical science
- Since many bioinformatics studies are correlative, follow up studies that de-confound correlation and causation are often required
  - Perturbation studies (eg. gene knockout models)

xkcd



Pitfalls of correlative studies

- Poorly set up hypotheses, bad inferential mechanisms, and low quality data can all contribute to arriving at the wrong conclusion (artifacts)
- Methodological rigor and domain knowledge are key to avoiding artifacts

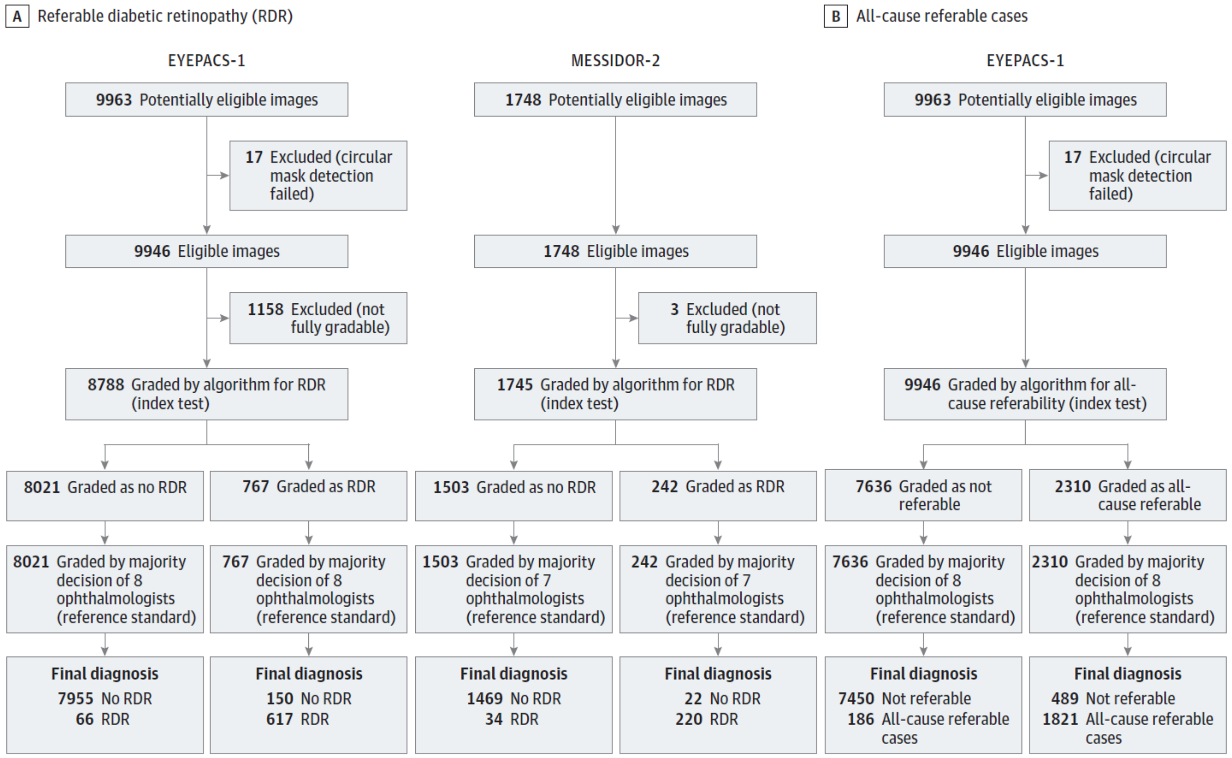


xkcd

“Artificial intelligence” in biology

The perception of AI has changed over the years :

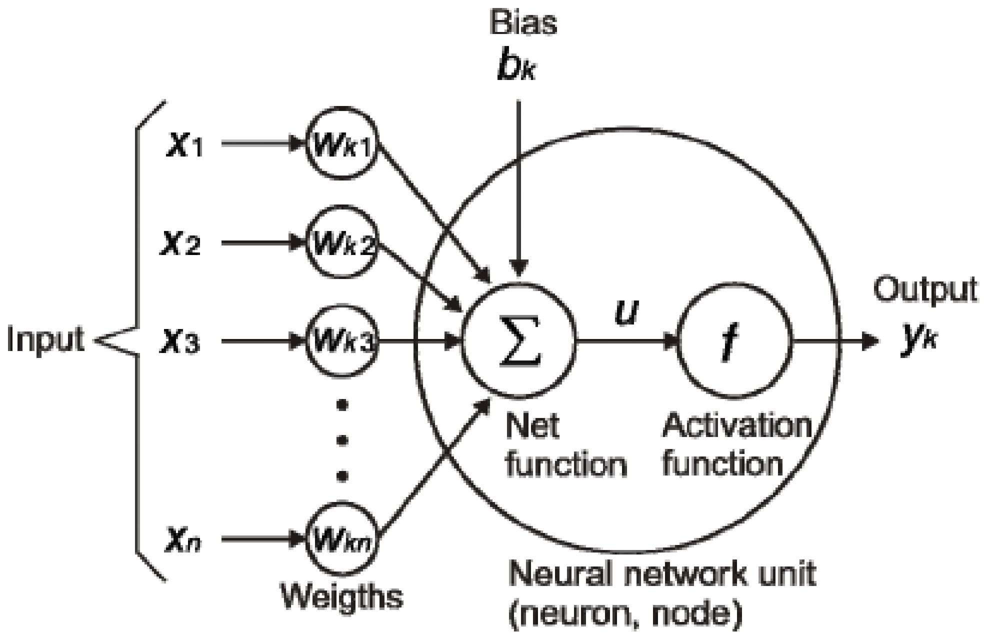
- 1970s : expert systems (eg. early recommender systems in medicine)
- 1990s : statistical machine learning (eg. genome wide association studies of diseases)
- 2010s : deep neural networks (eg. prediction of diabetic retinopathy from imaging)

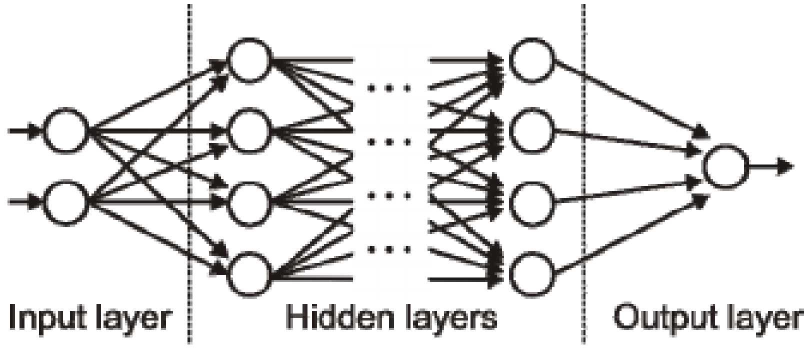


Gulshan et al, JAMA, 2016

Using neuroscience to model AI

- Neural networks

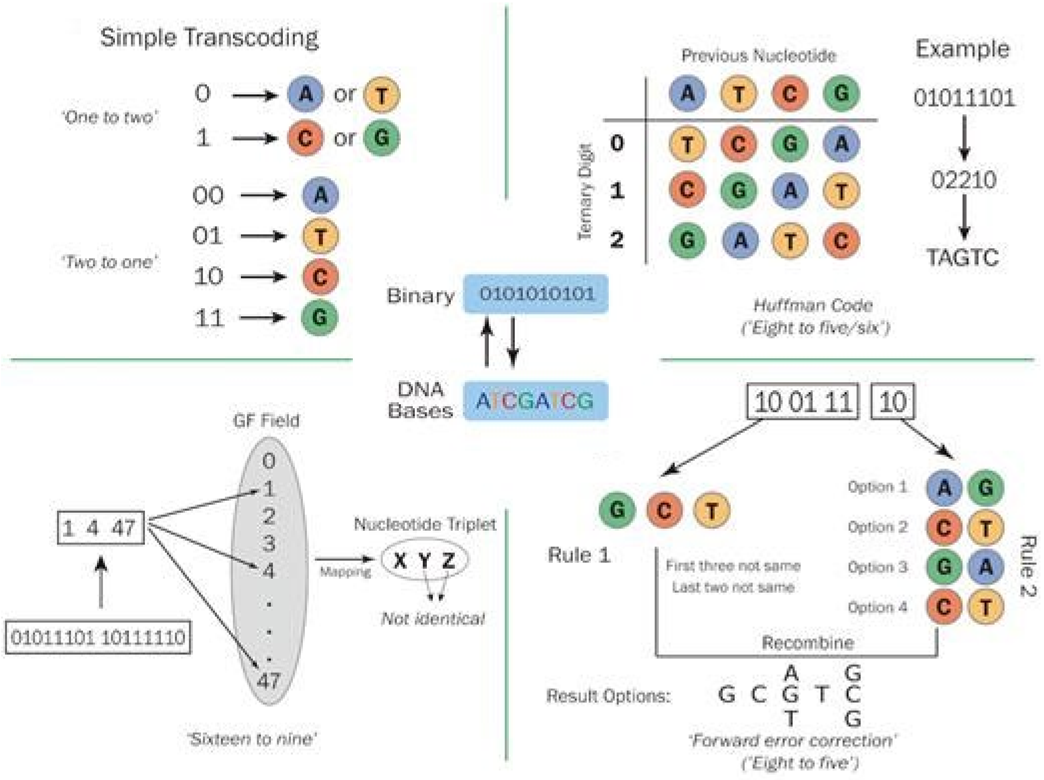




Frey et al, 2011

Other areas of computer science – biology intersection

- **Biological computation** : how can biological processes be harnessed to improve computing – eg. DNA based data storage
- **Biomedical devices** : electronic devices used for patient monitoring, treatment and therapy
- **Synthetic biology** : creating new biochemical / biological entities



Information encoding strategies for DNA – based storage

Ping et al, GigaScience

If you'd like to perform research in bioinformatics …

- Knowledge of coding is required

- Send email to pradiptaray@gmail.com
- Many other wonderful laboratories in UTD working on bioinformatics