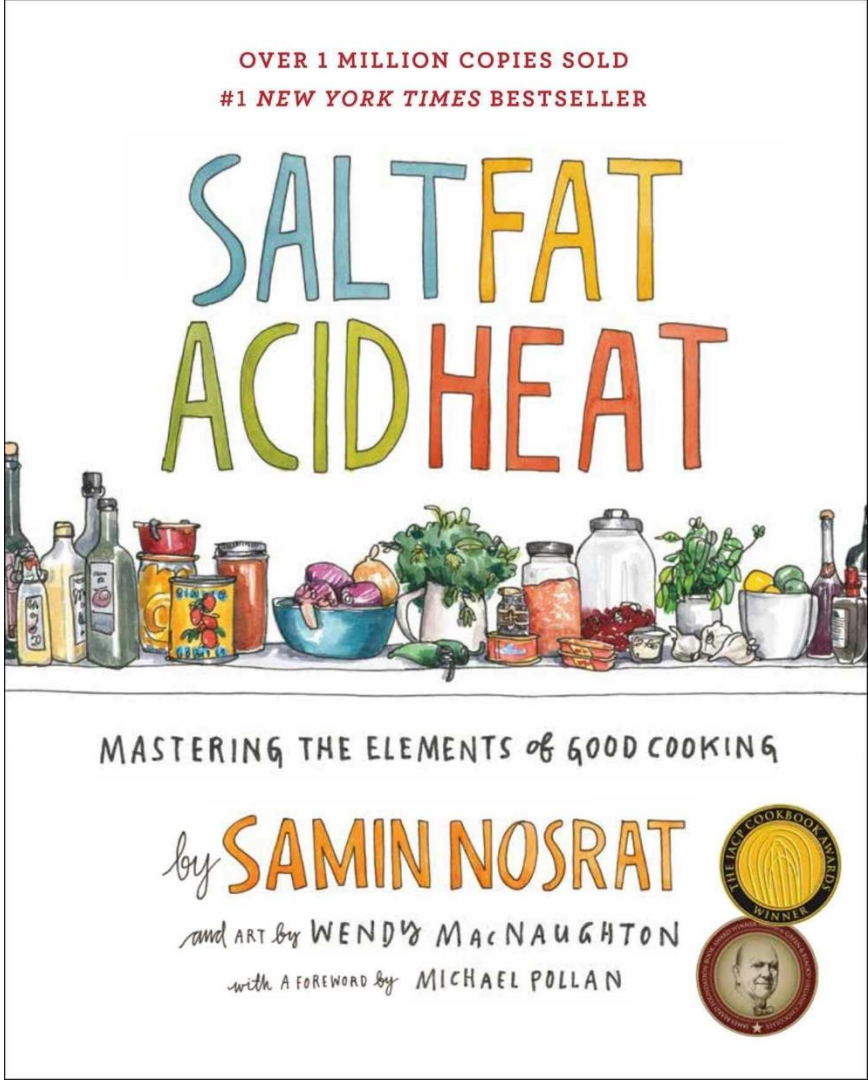
Bioinformatics

Like cooking, but with data!

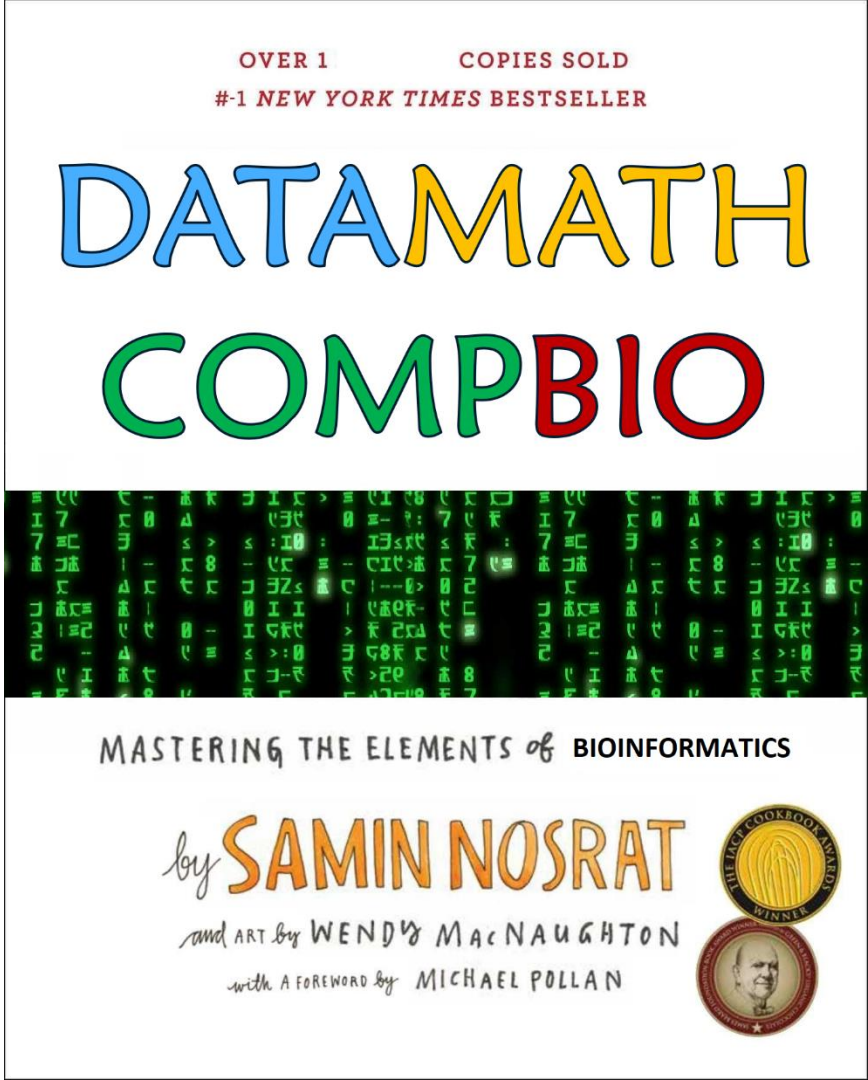
Yi Jin LIEW | 22 Jul 2022



Bioinformatics

Like cooking, but with data!

Yi Jin LIEW | 22 Jul 2022



STARTER

My journey

MAIN

Data

Comp

Math

Bio

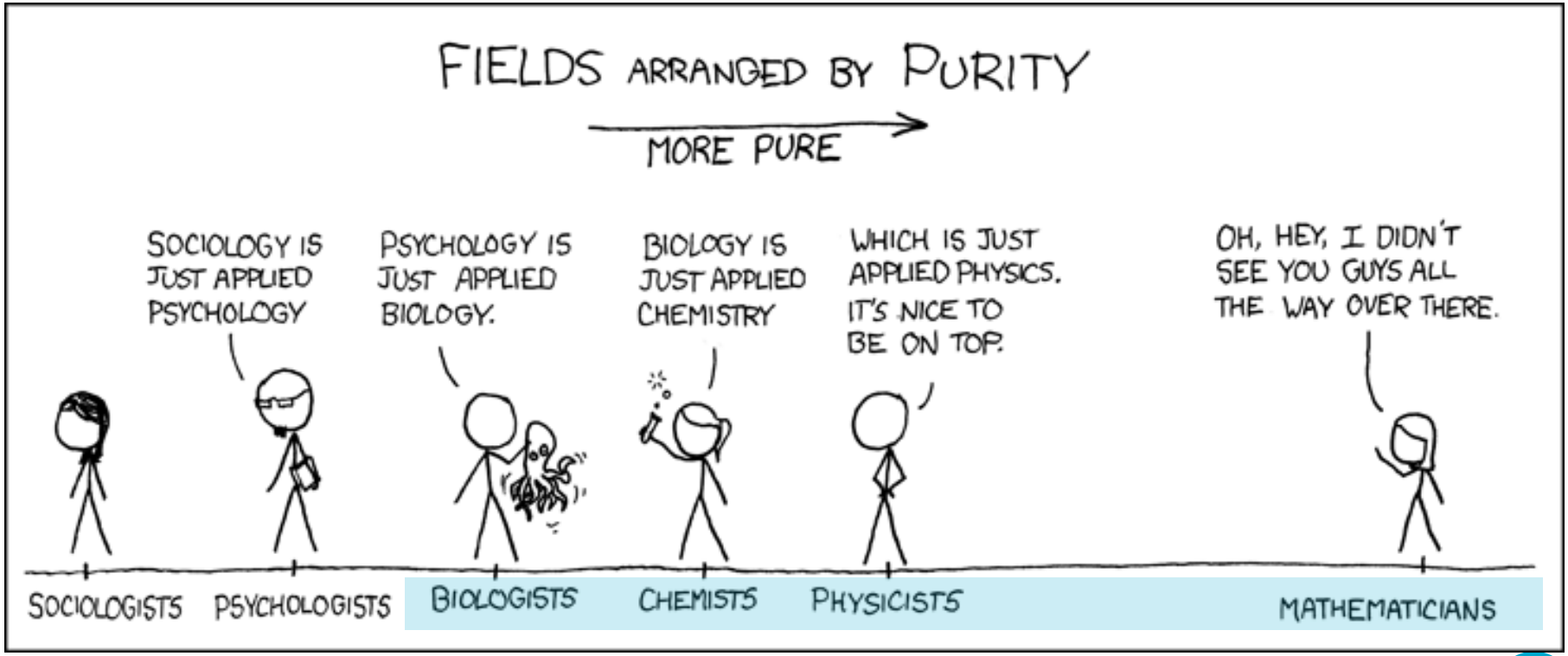
DESSERT

Questions/Demo/Freestyle

Starter

My bioinformatics journey

High school: Singapore



Uni: University of Cambridge (UK)

- **Undergrad:** Natural Sciences (Genetics)
- **PhD:** Genetics

Jardine Foundation

CAMBRIDGE TRUST

COMMONWEALTH EUROPEAN AND INTERNATIONAL

CSIRO

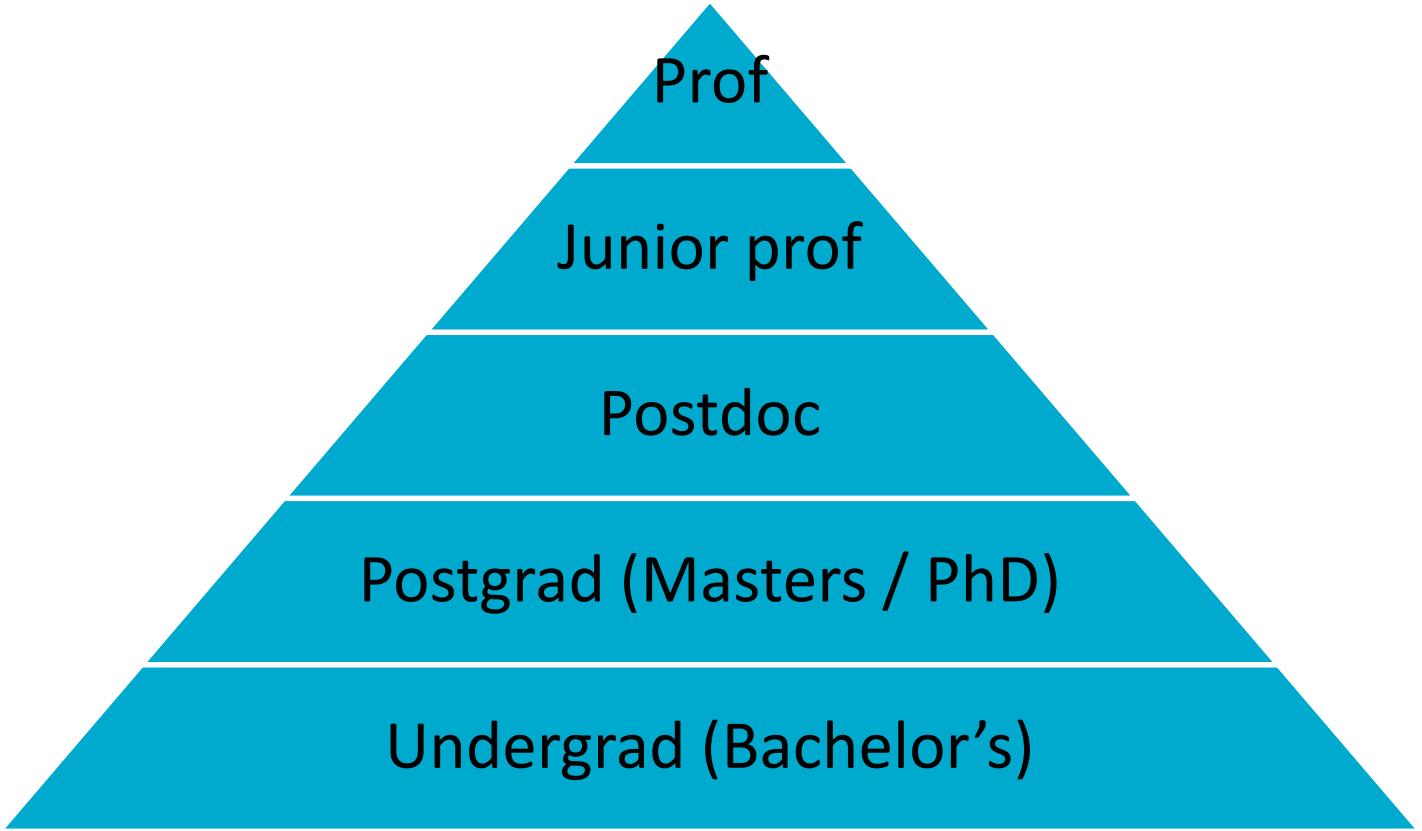
PhD? What's that?

A 3+ year training course to "git gud" in doing science

- **Pros**
  - Ask exciting scientific questions
  - Make reasonable assumptions
  - Design experiments to answer those questions
  - Learn new lab techniques, new ways to look at data
  - CHANGE THE WORLD, YO
- **Cons**
  - Can be hard to enter (usually need Bachelor's + honours)
  - Min 3 years, max...
  - Usually encouraged to go abroad for new perspectives
  - You don't end up earning more $ (really)
  - Deal with routine failure

CSIRO

Academia, i.e. how to upgrade “Dr.” to “Prof.”



Postdoc: King Abdullah Univ of Sci and Tech (KSA)

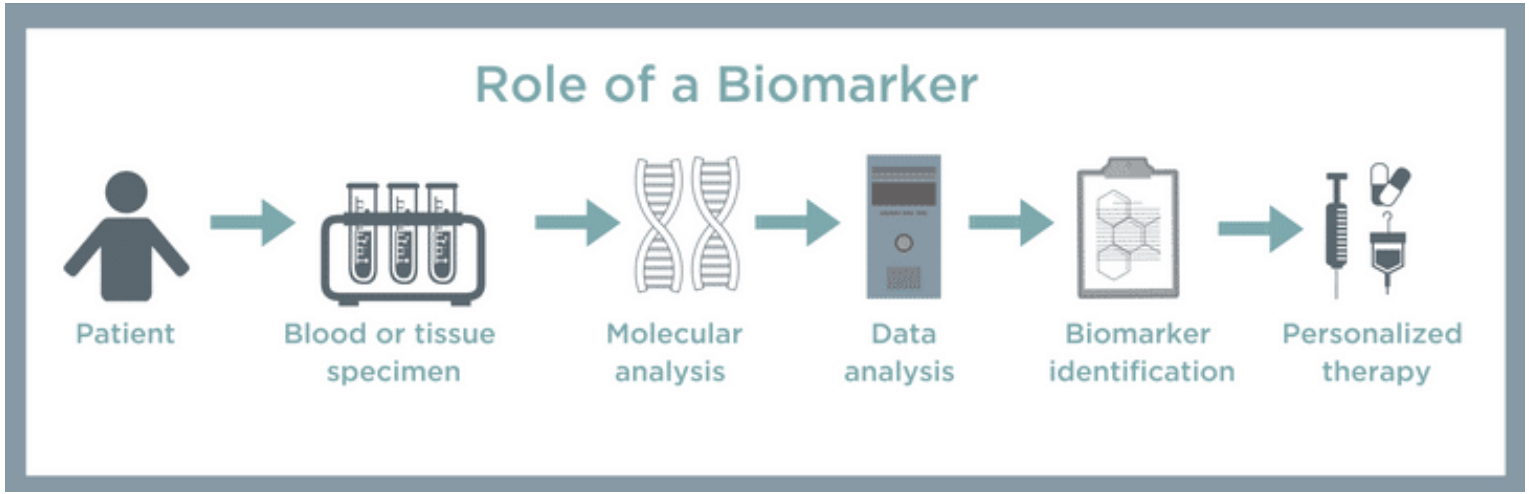
LETTERS
https://doi.org/10.1038/s41558-019-0687-2
nature climate change

Intergenerational epigenetic inheritance in reef-building corals

Yi Jin Liew[1,3,7], Emily J. Howells[2,4,7], Xin Wang[1,5], Craig T. Michell[1,6], John A. Burt[2], Youssef Idaghdour[2] and Manuel Aranda[1*]

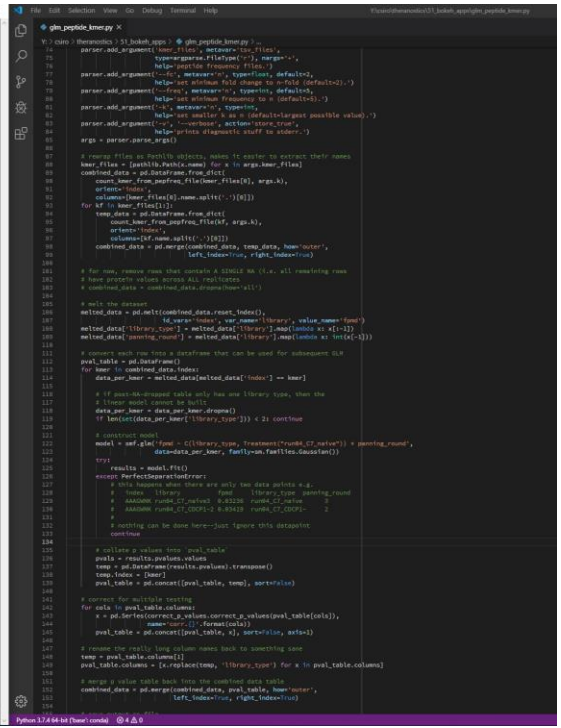
Research Scientist: CSIRO

- I am a **bioinformatician** discovering **biomarkers** for disease (Diagnostics group within Nutrition & Health program)



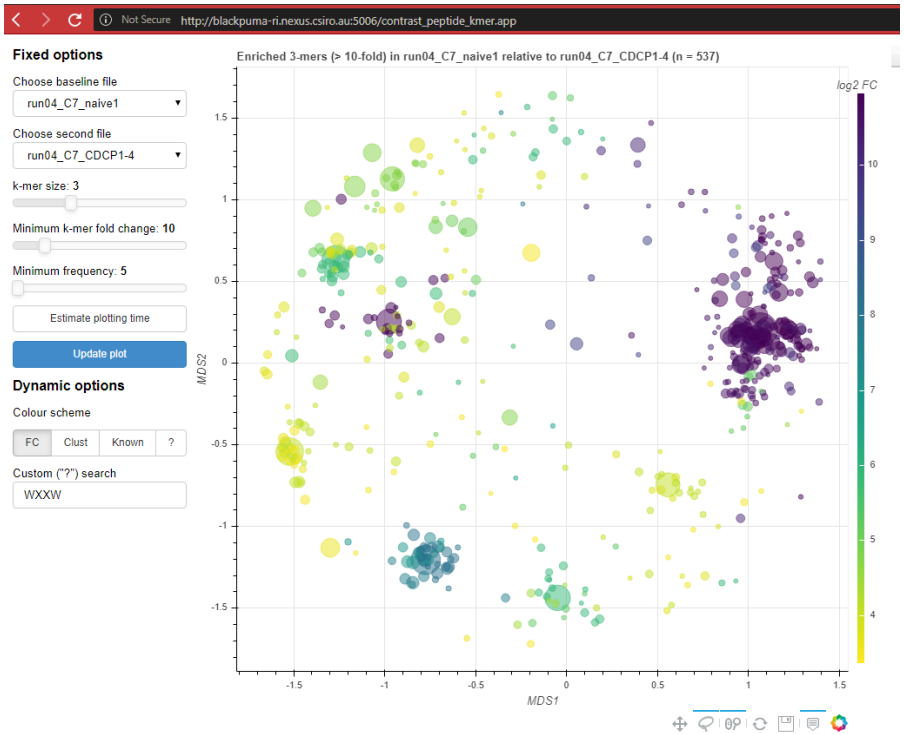
My day-to-day stuff

- I **write code** (after staring at the screen for a long time)



My day-to-day stuff

- I **make pretty pictures** to identify patterns in the data



My day-to-day stuff

- I write long-winded **reports on my observations**

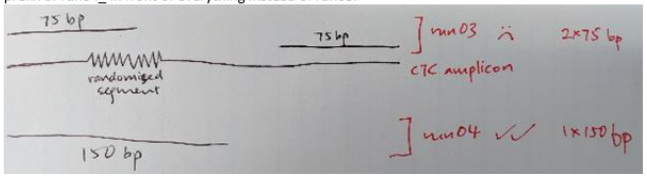
multiqc_report.0304.html 1 MB | run04_C7_CDCP1-4.stats.tsv 16 KB | cdcp1-4.sanger.peptide_freq.tsv 4 KB | cdcp1-3.sanger.peptide_freq.tsv 5 KB | run04_C7_CDCP1-3.stats.tsv 16 KB

hey all,

quick email to summarise next gen sequencing runs “run03” and “run04”.

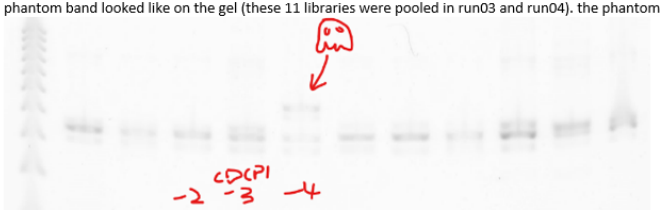
(for non-NR people) why there are two runs

after the library was prepped and loaded into the machine, the 2 x 75 sequencing option was erroneously selected instead of 1 x 150. the 75 bp forward run covers 80% of the randomised segment, making it worthless, unfortunately. we noticed the error early on, called illumina up, but they said that the settings can’t be changed once the run has started. oh well. run04 was done the way we envisioned, hence the prefix of run04_ in front of everything instead of run03.



unmasking phantoms bands with run03

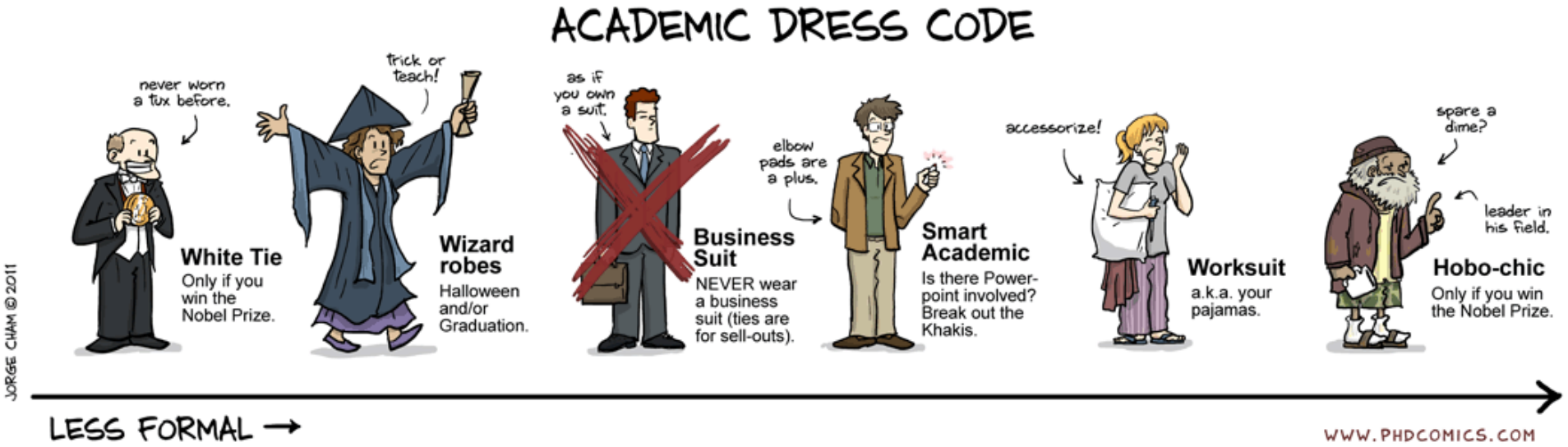
as halloween’s just around the corner, jason mentioned that we might be able to make some headway into understanding the identity of the phantom bands from run03 results. for context, this is what the phantom band looked like on the gel (these 11 libraries were pooled in run03 and run04). the phantom band is approx 100 bp longer than the intended band.



What's needed to be a bioinformatician?

- Knowledge: solid grounding in
  - Biology
  - Mathematics (!!!)
  - English (!!!!!)
- Soft skills: like all other jobs, try to be good at
  - Interpersonal skills
  - Explaining stuff
  - Work hard
- Above all, stay **CURIOUS**

Hidden scientific perk



Main

Four course meal

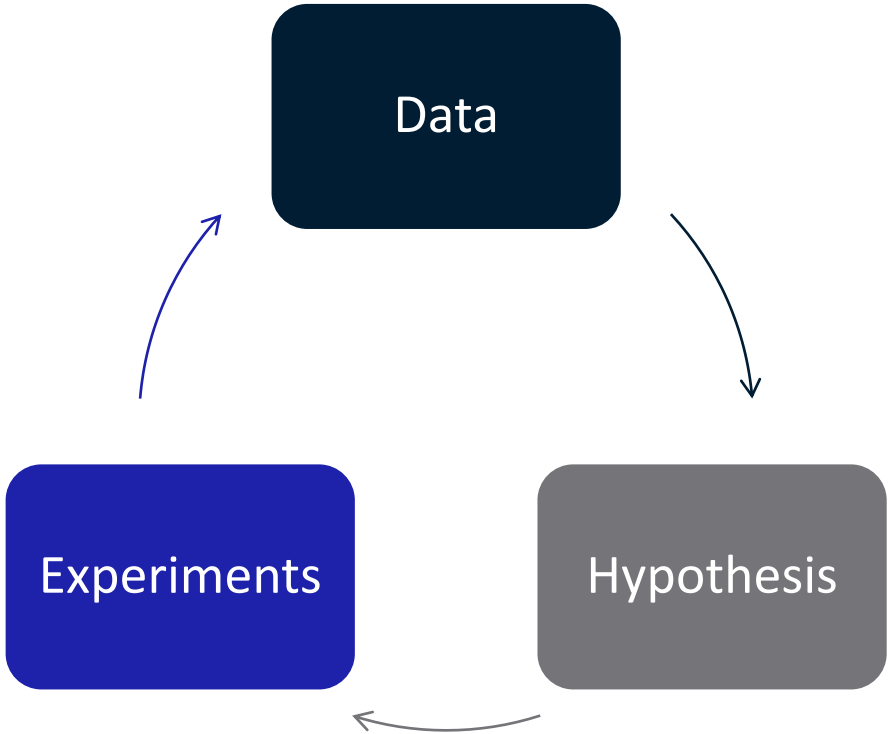
Main course

- DATA: the base of the dish
  - Importance of great ingredients
- COMPUTING: processing the ingredients
  - Cut? Garotte? Slice?
- MATHEMATICS: taste-tasting
  - Is it good, or SIGNIFICANTLY good?
- BIOLOGY: presenting the dish
  - A5 Wagyu sucks, if customer wanted scrambled eggs

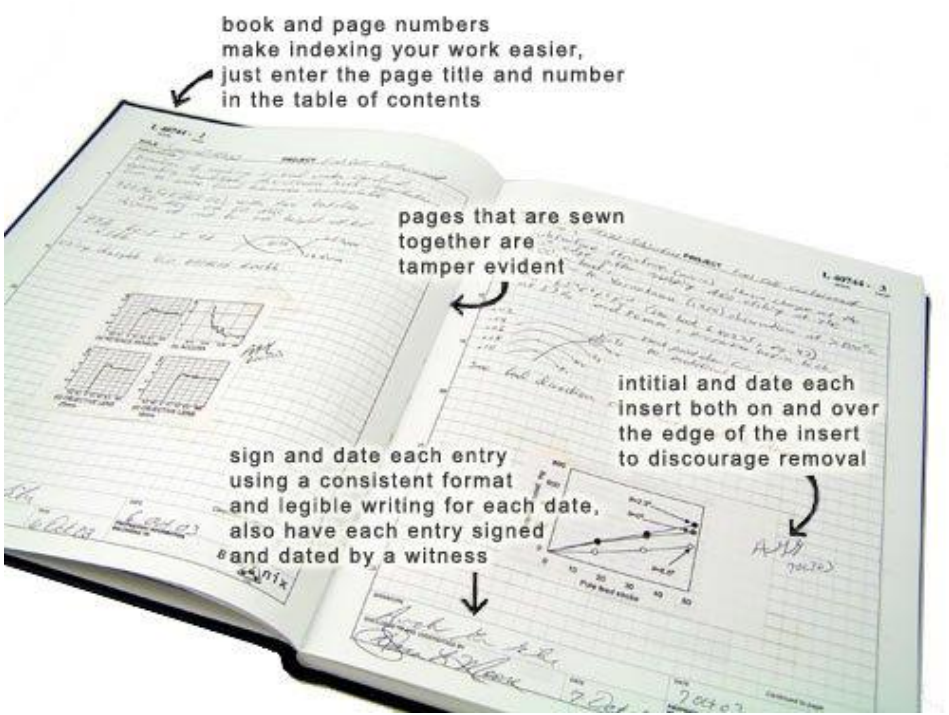
DATA

Garbage in, garbage out

Data-driven hypotheses

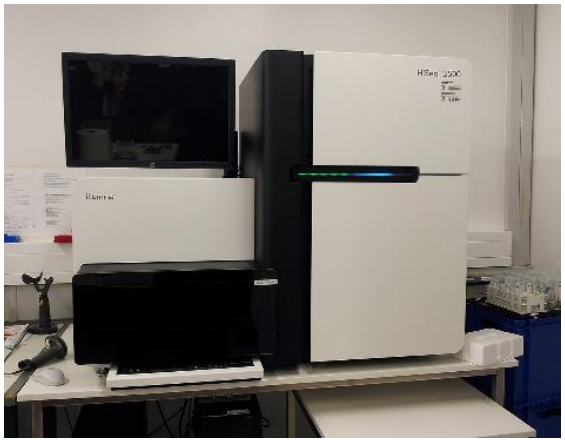


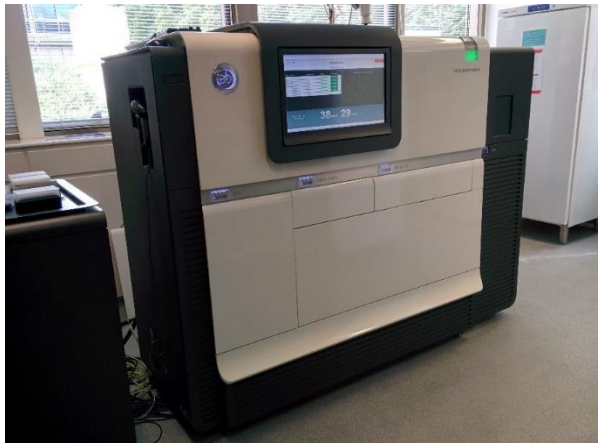
Data generated from experiments




```
@A00121:320:HWKW3DRXX:1:1101:2094:1016 1:N:0:CACTGTAG+AGTCGCTT
NAAGTAAGTTTATTTATTTAATTAAATGTAGTTTTATTTTAATTAATAATATTAATGAAATTATATATTTGATGTGATGATTATATTTTTTTTTTAATAGATTTTAAAATAAATATAATTA
+
#:FFFFFF,:FF:FFF:FFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF::FFFFFFFFFFFFFFFFFF:::FFFFF:FFFFF:FFFFF:FFFFFFFFFFFFFFFFFFFFFFF
@A00121:320:HWKW3DRXX:1:1101:2528:1016 1:N:0:CACTGTAA+AGTCGCTT
NGGAAGAATTTTTTTTGAAAATTAGTATAAGATAAGGATGTTTTTTTTTTATTATTTTTATTTAATATAGTATTGAAAATTTTGGTTATAGTAATTAGTTAAGAGAAAAAAAATAAAGGGTA
+
#FFFFFFFFF:F:FF,FFFFFFFF,FFFFF:FFFFF:FFF:FFFF:,,FF::FFFFF:FFFFFFFFFFF,FF:FFFFFFFFFFF,:F,,FFFFF:F,FFF,,,FFFFF:FFFF,FF,FF
@A00121:320:HWKW3DRXX:1:1101:2546:1016 1:N:0:CACTGTAG+AGTCGCTT
NGAGAGGTTGGGTTTATAGAGTTTAGGTTATTTTAGATTTTTTGTTAAGAATTTTAAGTTATAAATGTTAGTTTGTATTTTTAAGTGTTTTTTTTTTTTCTCTTTCTGTTATTTTTTTTTA
+
#F:FFFF:FF,FFF,FFFFFFFFFF:FFFFFFFFFF,FFF,:FFF,FFFFFFFFFFFFFFF::FFFFF:FFFFFFFFF,F,F:FF,FFFFFFFF:F,FF,,F,,F,F,F,,F,,
@A00121:320:HWKW3DRXX:1:1101:2781:1016 1:N:0:CACTGTAG+AGTCGCTT
NTATAAATAAATGATAAATGTATGAGGTGATGGATATGTTGAATATTTTAATTTGATTTTTTTATATAATATATATATGTATTGAAATATTATATTATATTTATAAATATGTATAATTATT
+
#FFFFF:FF,FFFFFFF:F,FFFFFFFFFF:F:FFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFF:FF,FFF:FF:FFFFF:FFFFFFFFFFFF:FFF,F,F:FFF,FF::F:FFF:FFF
@A00121:320:HWKW3DRXX:1:1101:2871:1016 1:N:0:CACTGTAG+AGTCGCTT
NATGTTGTTTAGGTTGGTTTGAGTTTTGGGTTTAAGCGATTTTTTGTTTTTGTTTTTTAAAGTGTTGGGATTATAGGTATGAGTTATAGTGTTTTATTATTGATTGTTTTTTAATTTA
+
#FFF:,FF,FFF,FF:FFFF,FFFFFFFFF,FFFFFF,,FFF:FF:FF:F,FF,FFFFFFFFFF,FF,FFFFFFFFF:FFFFFFFFFF:FFFF:,FFFF,FFFF:,:FF,FFFFFF,F
@A00121:320:HWKW3DRXX:1:1101:3016:1016 1:N:0:CACTGTAG+AGTCGCTT
NTTATTTGAAGTTTTTTAGAAAAGTATTAAGTTTTTTTAAATATATTTTTTATATTAAAATGAGAGTTTTTTTTTGTTATAAGGAGGTTTTTTTTTTTTTTTTTTTTTAATTTTATTGTTTA
+
#FFFF:F,:FFFF:FFFFFFFF,F::FFFFFFFFFFFFF:FFFFFFFFF,FFFF,FF:F:FF,F:FFF,FFF,FFFFFFF,,FFFFF,,F:,FFFFFFFFFFFF,,F,,:,FF,,:FFF,F
@A00121:320:HWKW3DRXX:1:1101:3034:1016 1:N:0:CACTGTAG+AGTCGCTT
NTTAAAGATATTTTAATTATTAAAGGTTTATTATTAGAGGTGTGTGTGGATTAGCGAGTTTATTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCACTGTAGATCTCGTATGCCGTC
+
#FFFFFFFFFFFFFFFF:,:FFF::FFFFFFFFFFFFFFF:,F,FFFF:FFFF:,,F:FFFFF:FFFF,FFFF:FF,F,F,FF,FFFF:F:,FFF,F,,F,FFFFFFFF:FFFF:FFFF,F
@A00121:320:HWKW3DRXX:1:1101:3125:1016 1:N:0:CACTGTAG+AGTCGCTT
NTTATTTATTTTATTAATCGAATATTTTTAATTTTGTTTAGAGTTTTAAATGTTAGGTATTTTTGATAAATGTTTTATATTAAAGATTTTTTTTGATTTTGTATATTTTTAGATTAATGTC
+
#F:FF,:F:FF:FFFFF,,F,FFFFFFFFFF,FFFF,FF:FFFFFFFF::F:,FF:,FFFF,FFFFFFFF::,F:FFFF,FF,:FFFF:FFF:FFF,F,FF:FFF:,,FF:F,F:::FFF
@A00121:320:HWKW3DRXX:1:1101:3341:1016 1:N:0:CACTGTAG+AGTCGCTT
NAGTTTTGGGATATTTTTTAATGGAATAATAAATAATGATAAAAATGAATGAATTATAGTTATGTAAAGTATGATTTAATTTTAAAAAATTTTTATAGAAAAAAGAAAAATATGTAGAAAA
+
#FFFF,:F:FF:F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F:FFFFFF:FFFFFF:F:FFFFFFFFFFFFF:FFFFFFFFFFFFFFFFF:F:F,FFFFFFFF:FFFFFF
@A00121:320:HWKW3DRXX:1:1101:3414:1016 1:N:0:CACTGTAG+AGTCGCTT
NTTTAGTTAATTTTTGTATTTTTAGTAGAGATGGAGTTTGTGTATGTTATTTAGGTTGTTTTTGAATTTTTGGGTTTAAGAGATTTGTTTGTTTTAGTTTATTAATGTGTTGGGATTATAA
+
#F,FFFF,FFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFF:FFF:FFFFFFFFFFFFFFFF:FFF:FFFFFFFFF:FFF:FF:FFFFF,FFFFFFFF:F,FFFFFFFFFFF
@A00121:320:HWKW3DRXX:1:1101:3902:1016 1:N:0:CACTGTAG+AGTCGCTT
NGGTGTTTATTGATATGTTTGTGGGTTTAGTTTAGTGTTGGGGGTTGAAGGTTTGGTATATAGGTTTAGTTGAGTTTGTAGTGGAGTATTATAGGTGATATTTGTGTAAGGAAGGGGTAGT
+
#FFFF:FFFFFFFFFFFF,FFFF:FFFFFFFF,FFFFF,FFFFFFFFFFF,FFFF,FFFFF:FF:FFF,FFFFFFF:FFFFFFFFF:F,F,FFFF,FF,F:FFF,,FF,:FFFFF,
@A00121:320:HWKW3DRXX:1:1101:4047:1016 1:N:0:CACTGTAG+AGTCGCTT
NTTGTAATTTTAGTATTTTGGGAGATTAAGGACGGTAGATTGTTTGAGGTTAGGAGTTTGAGATTAGTTTGGGTAATATGGTGAAATTTCGTTTTTATTAAAAATAGAAAAATTATTCGGG
+
#F,FF:FF:FFFFFFFFFFFFFFFFFFFFFFF:FFF:FFF::F:FFFFFFFFF,FFFFF,FFFF:FFF::FFFFFFF:FFFFFF:FF,FF,FFFFFFFFFFFFFFFFFFFFF,FFF:F
@A00121:320:HWKW3DRXX:1:1101:4065:1016 1:N:0:CACTGTAG+AGTCGCTT
```

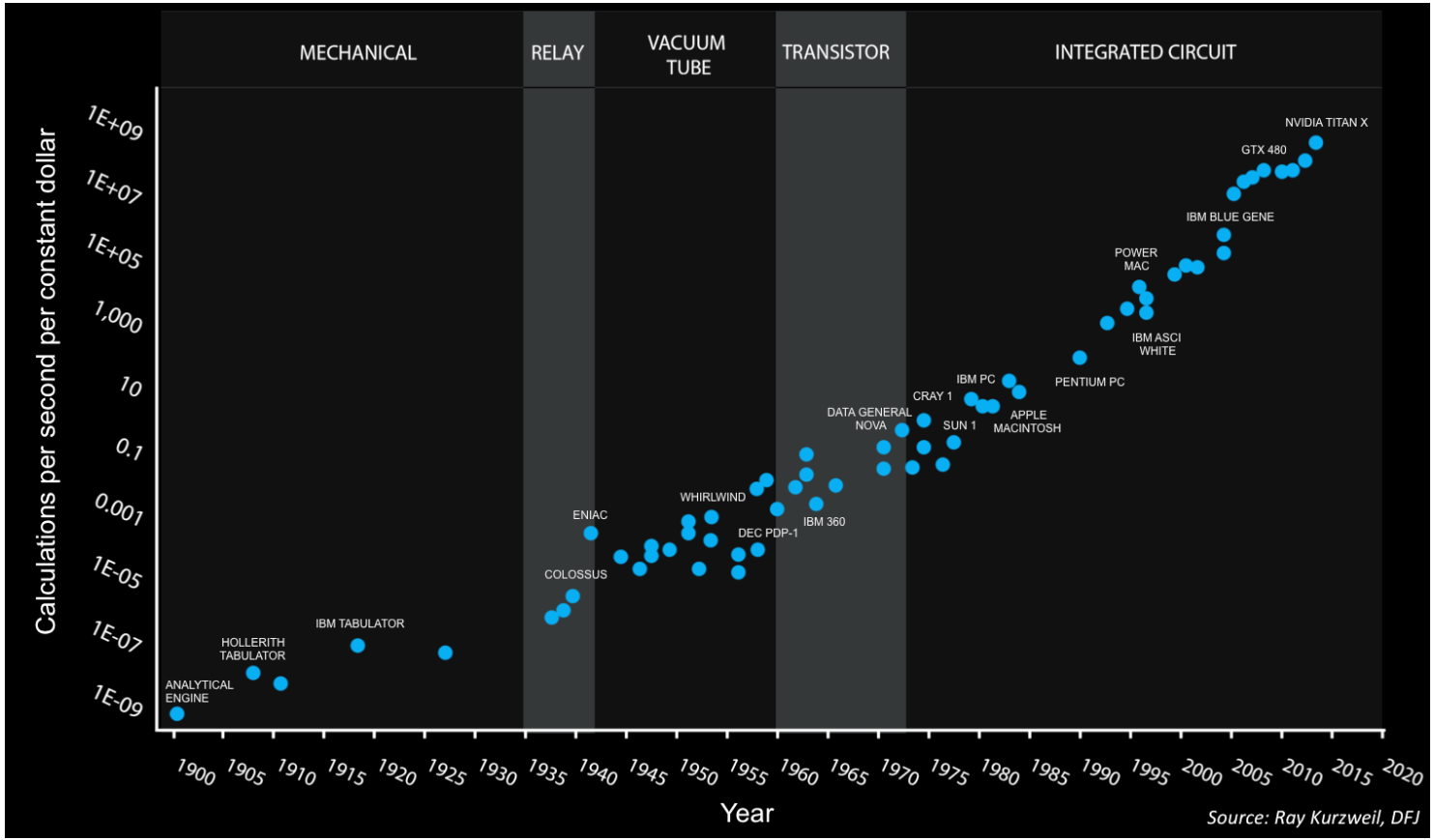
Case study: DNA sequencing



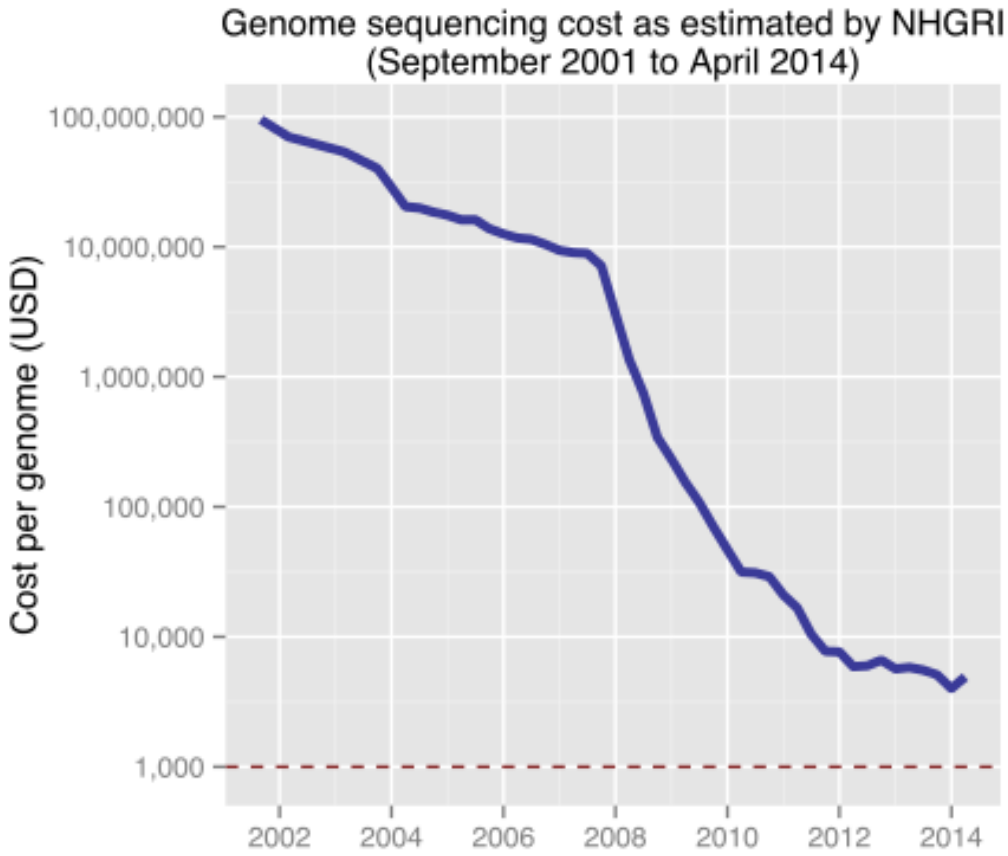


1. Sanger sequencer
2. Next-generation sequencing (Illumina / SOLiD / 454)
3. Single molecule sequencing (PacBio / ONT)

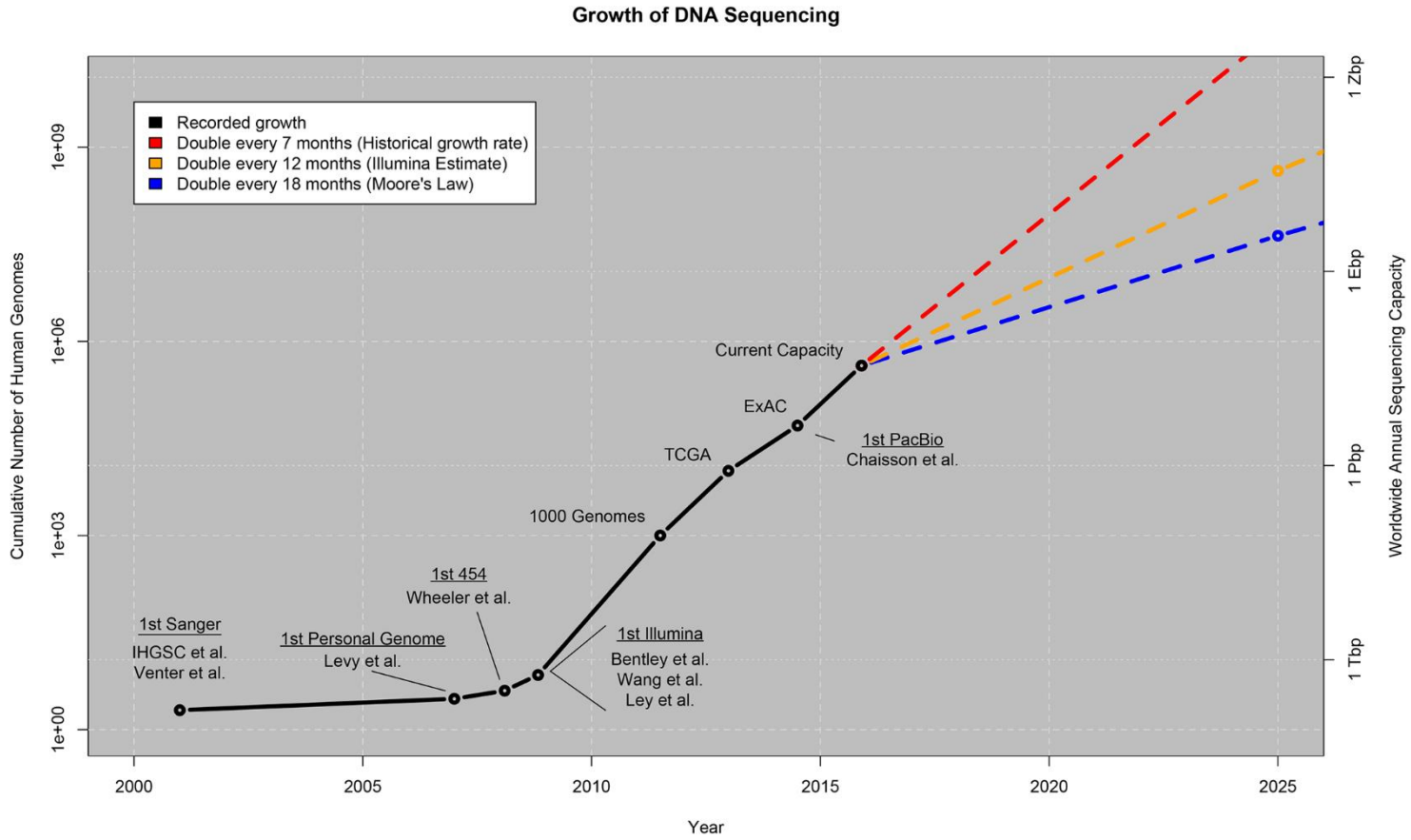
Case study: exponential growth in computing power



Case study: overall DNA sequencing cost



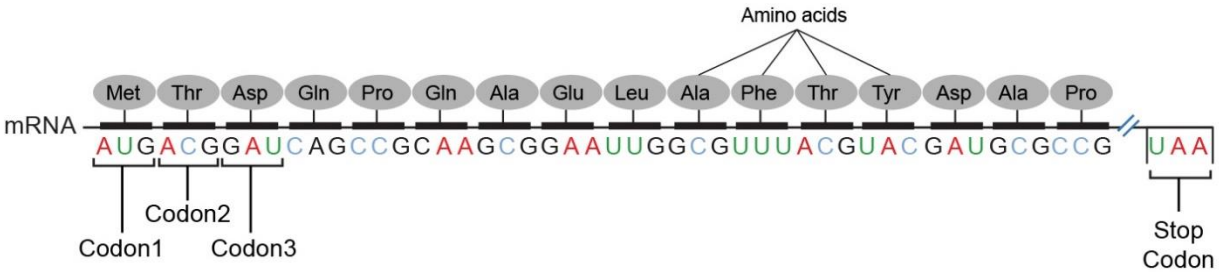
Case study: Explosive growth in DNA data

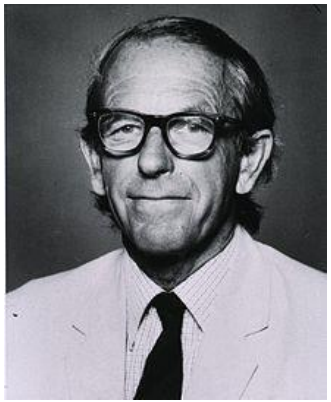


COMP

Bioinformatics (= computational biology)

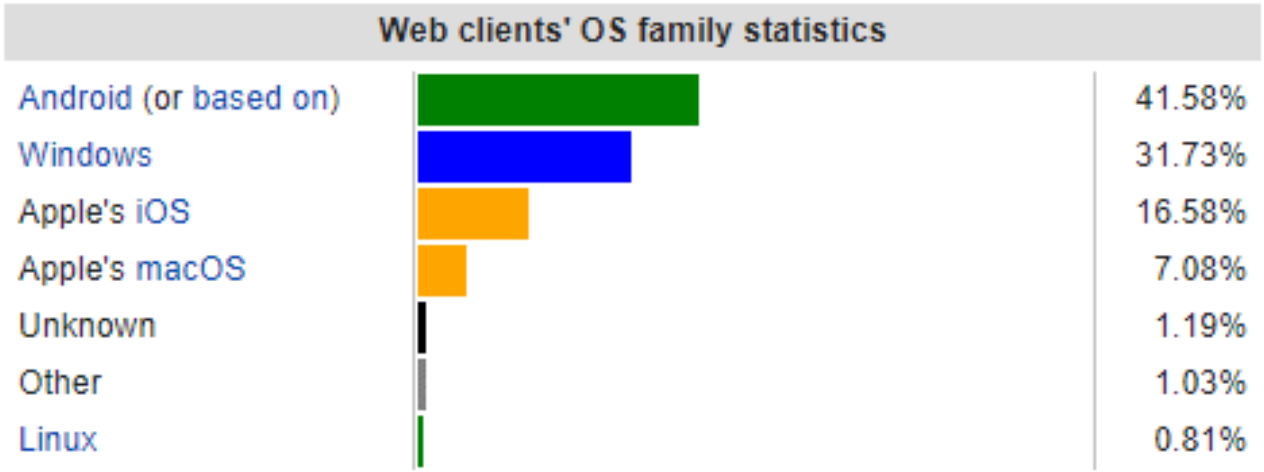
- Bioinformatics = **bio**logy + computing (**informatics**)
- Earliest bioinformatics problems were:
  - Aligning protein sequences (protein sequencing pioneered 1950s)
  - Studying bacteriophage genomes (DNA sequencing pioneered 1970s), reinforced concepts of codons, open reading frames etc.





Fred Sanger
(1918 – 2013)

Modern bioinformatics run on Linux



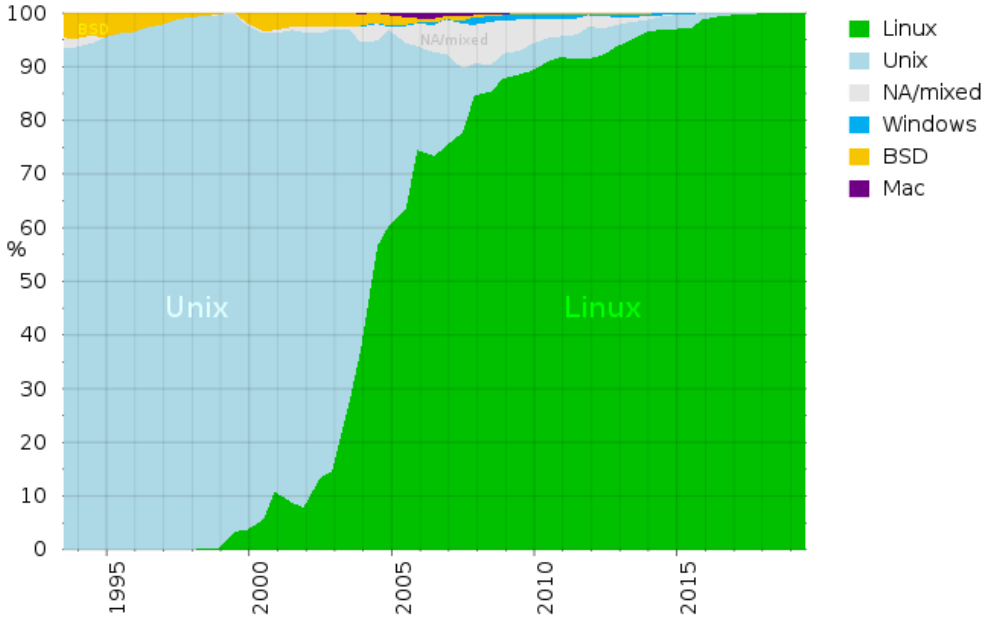
Web clients' OS family market share according to StatCounter for December 2020.[38] The information on web clients is obtained from user agent information

Bioinformaticians are the 1%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Openness: anyone can read source code of bioinformatics software

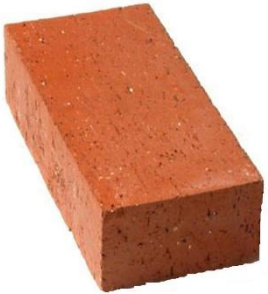
Bioinformaticians are the 1%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Supercomputers worldwide run Linux



Bioinformaticians are the 1%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Ease at manipulating large files: nowadays, files > 1 GB are extremely common. Handling large files in Windows/OS X is extremely clunky!
  - Programming philosophy: **modular** vs **monolithic**

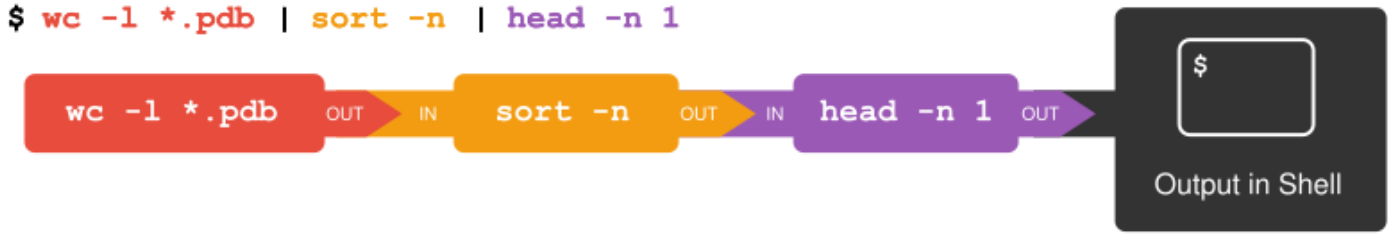


Linux

Windows/OS X

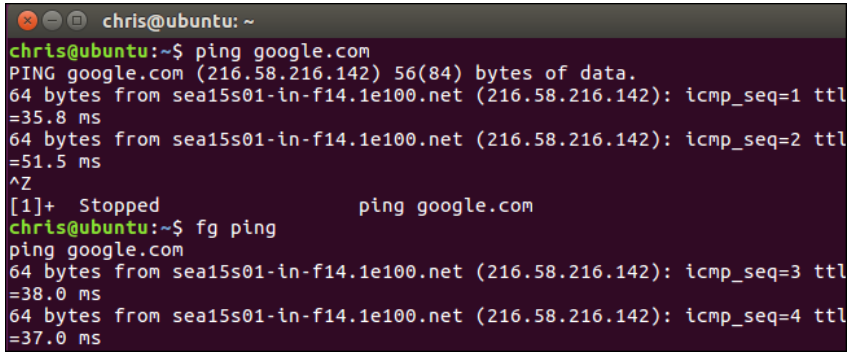
Bioinformaticians are the 1%

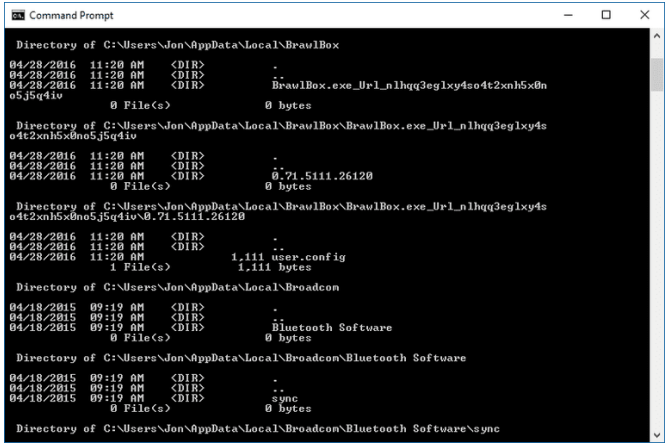
- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Text input/output, not silly proprietary filetypes (try opening a Word document in TextEdit/Notepad)
  - Piping: output of a tool can be “piped” as the input into another tool

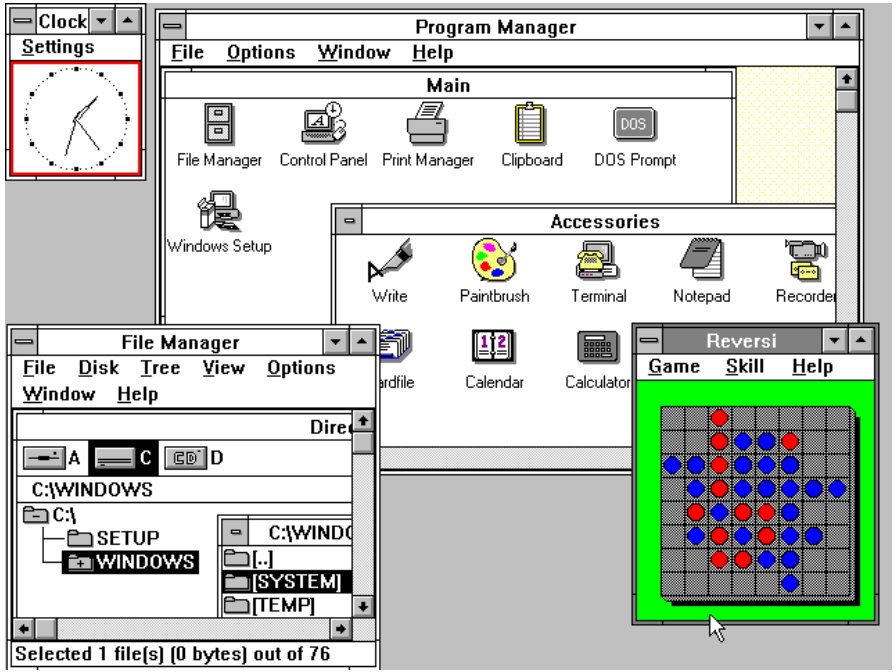


When bioinformaticians put several programs together to produce the desired output, they say they’ve built a “**pipeline**”

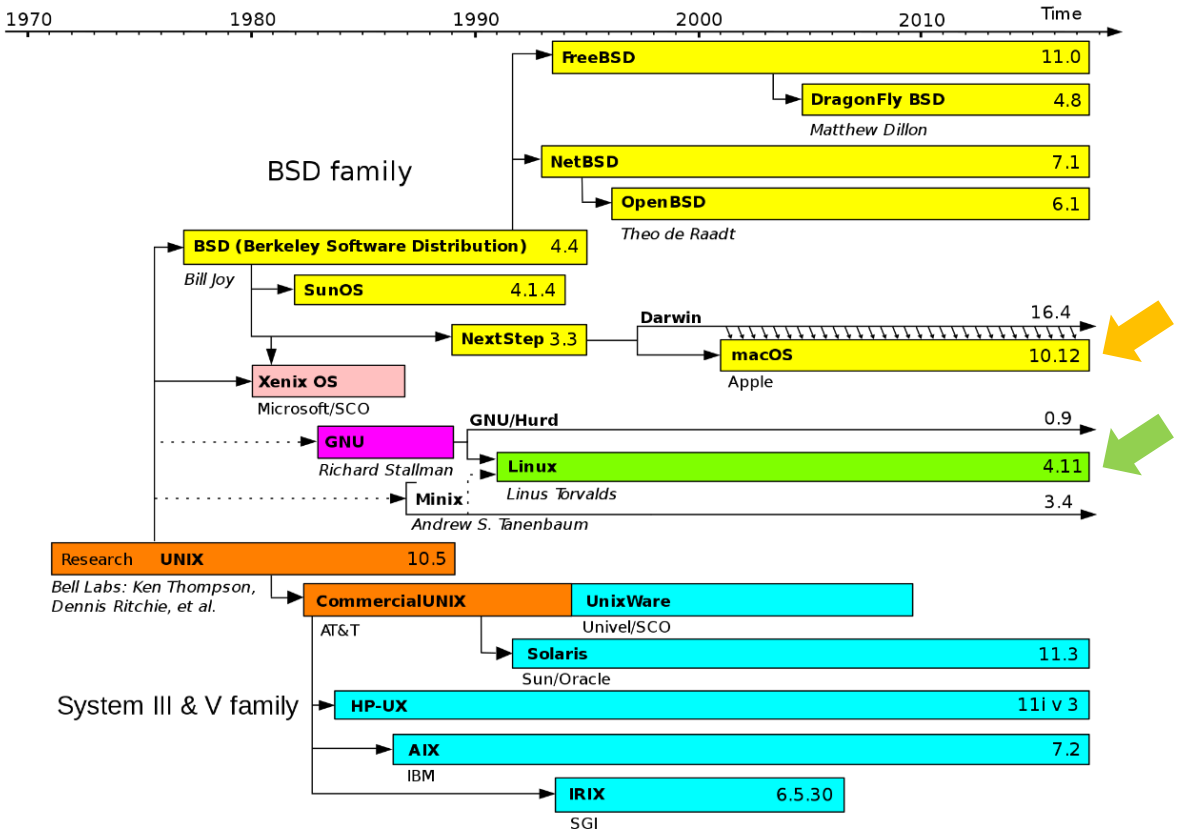
Heavy lifting done in command-line, not GUI





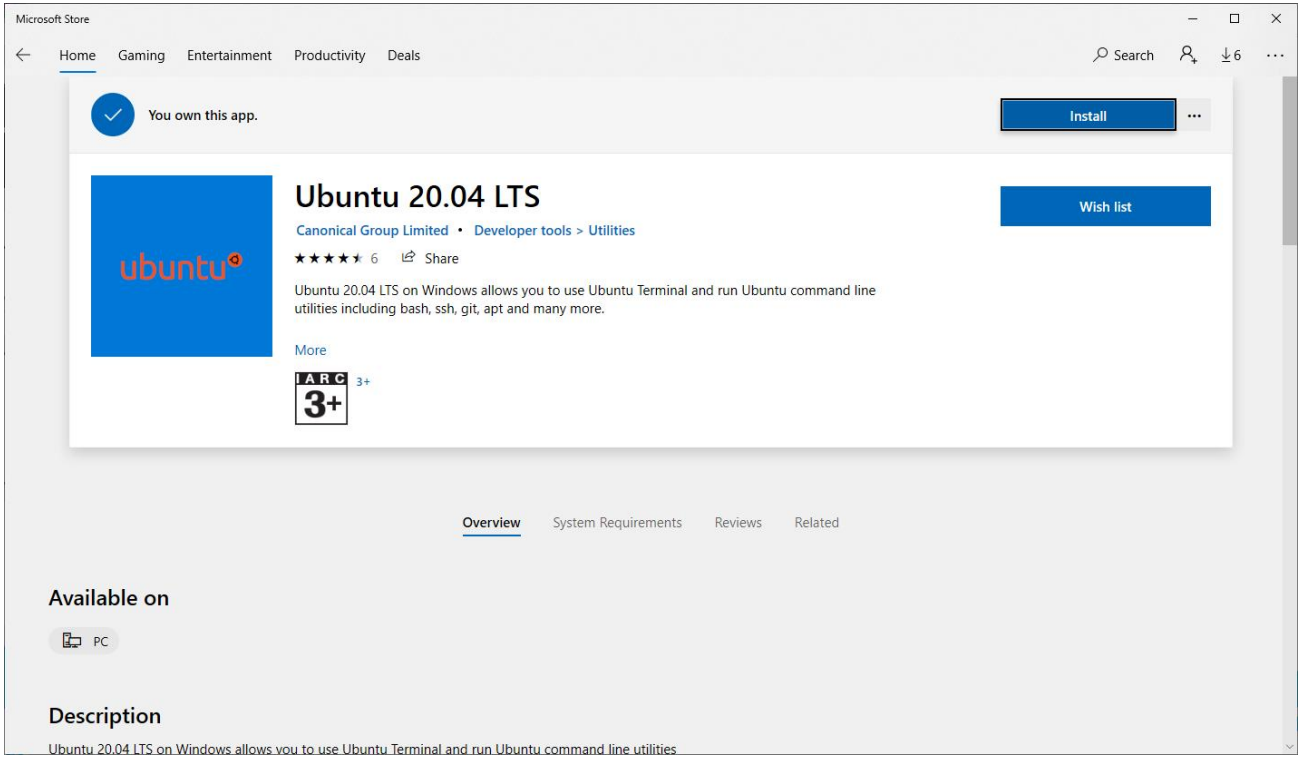


Command-lines are different!



- OS X and Linux have similar command lines (arrows)
- Windows command lines are very different
  - **cmd**: more UNIX-like
  - **PowerShell**: manipulates objects instead of text

2016 onwards: Linux on Windows!



Low-level programming languages

- **Machine code**
  - Processor-specific instructions read by microprocessors

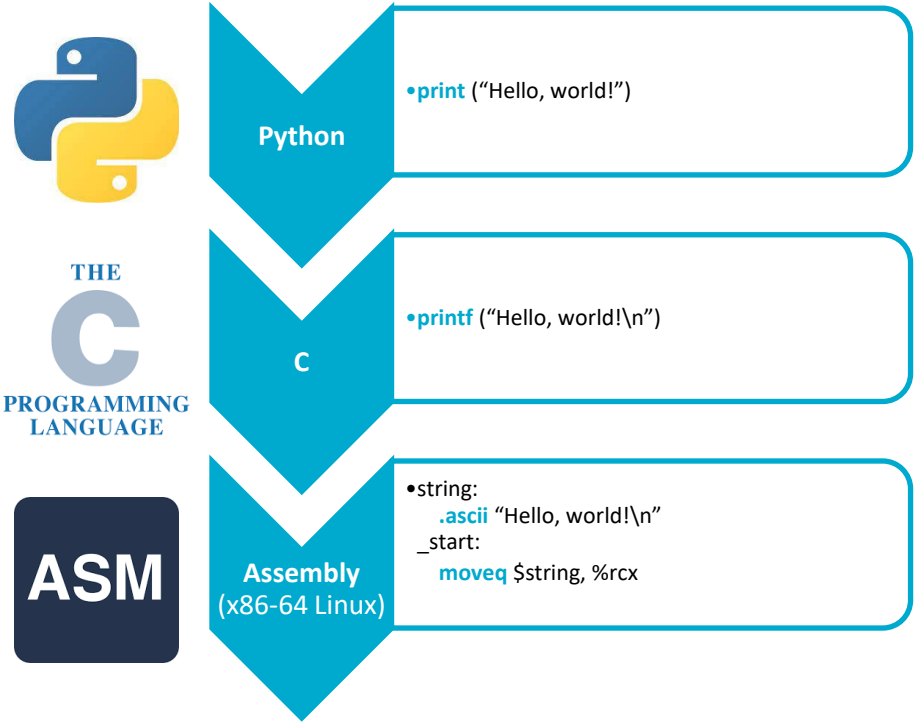
    8B542408 83FA0077 06B80000 0000C383
    FA027706 B8010000 00C353BB 01000000
    B9010000 008D0419 83FA0376 078BD98B
    C84AEBF1

- **Assembly**
  - Still processor-specific, BUT easier to write!

```
mov edx, [esp+8]
cmp edx, 0
ja @f
mov eax, 0
ret
```

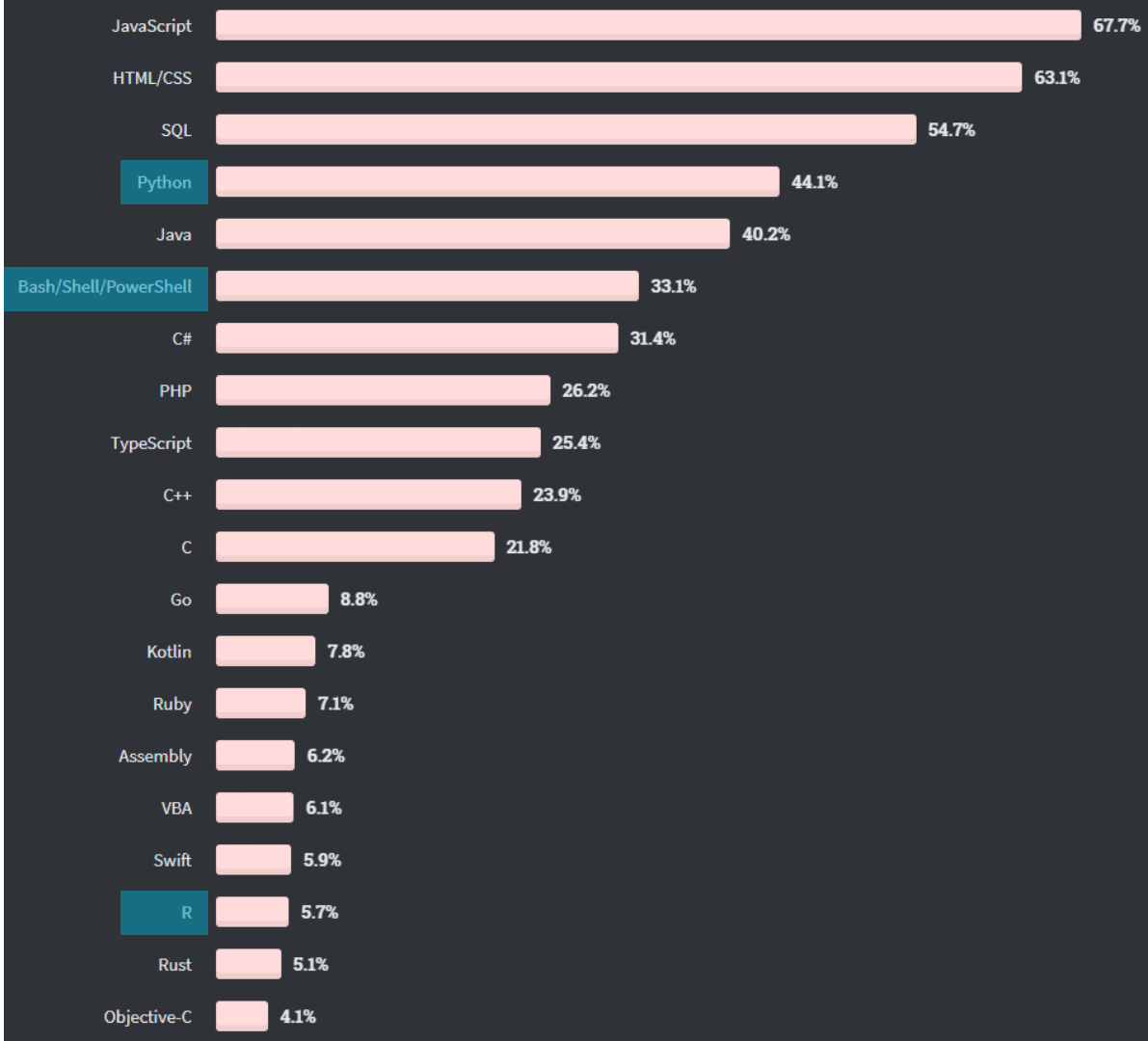
High-level programming languages

- Modern programming languages: ideas written in English, then **“reverse translated” (compiled)** back into machine code



Which lang?

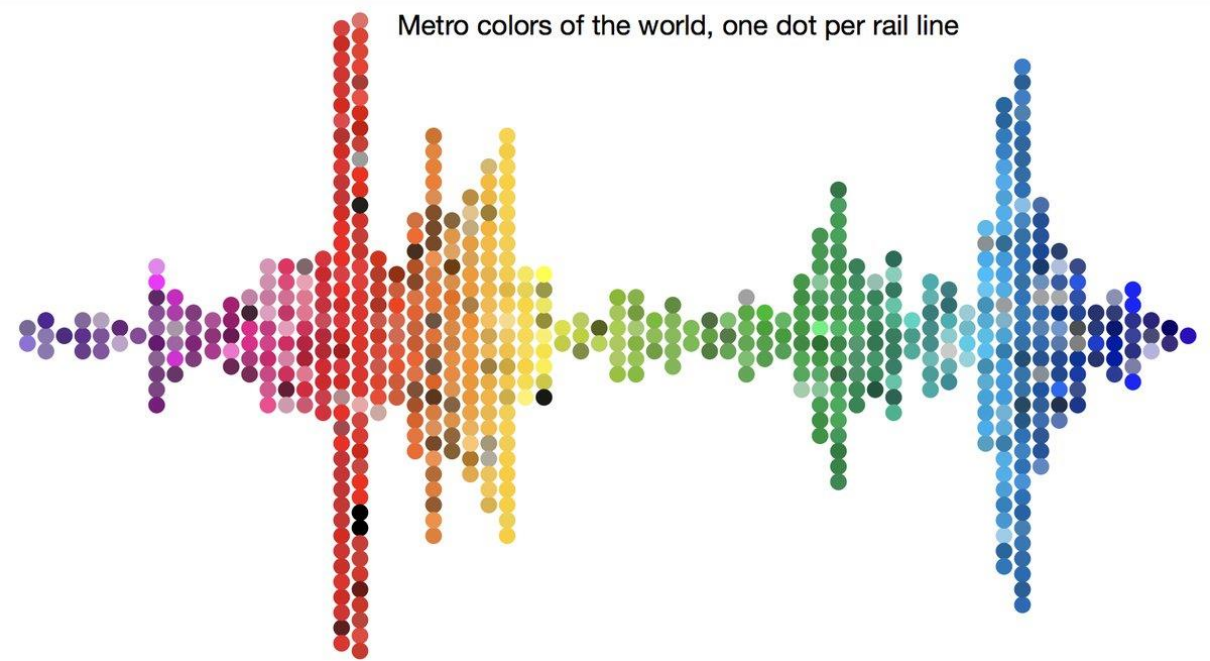
- StackOverflow runs annual survey, 2020 results shown here
- Highlighted are languages **most relevant to bioinformatics**
  - (learning more never hurts!)



MATH

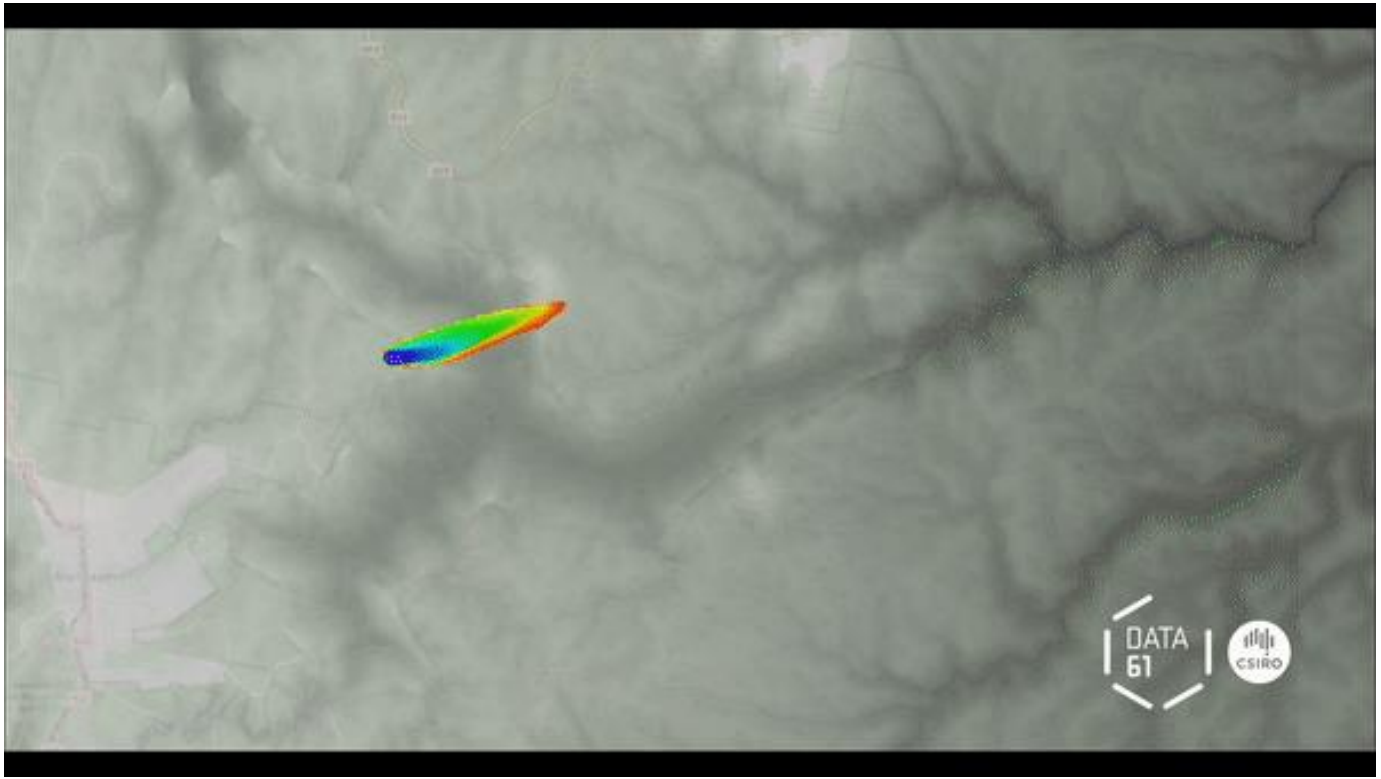
Statistics: why learn it?

- Allows one to **quantify** whether data is interesting!



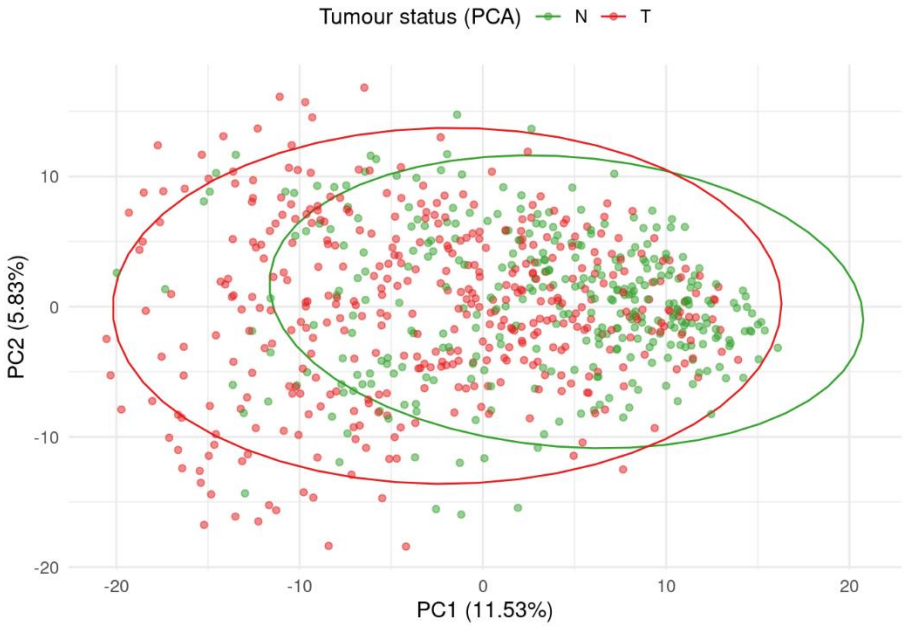
Statistics: basis of machine learning / AI

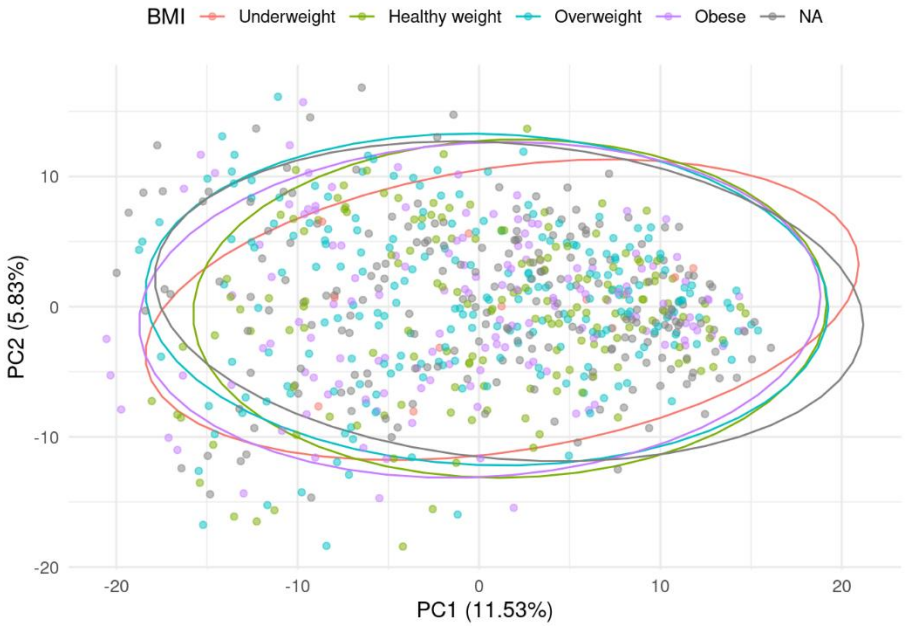
- CSIRO's Data61 models **spread of bushfire** with AI



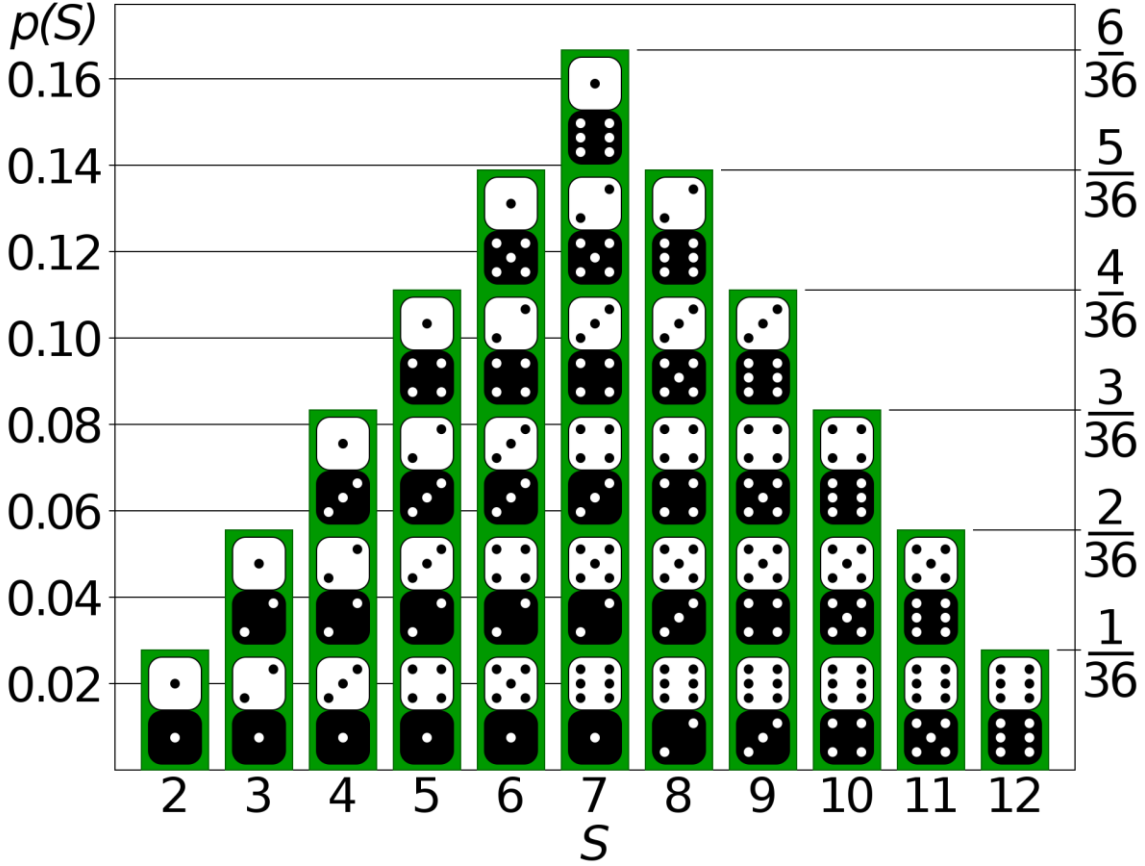
Examples from my work

- Visualising **microbiome patterns** vs. tumour status, or vs. weight



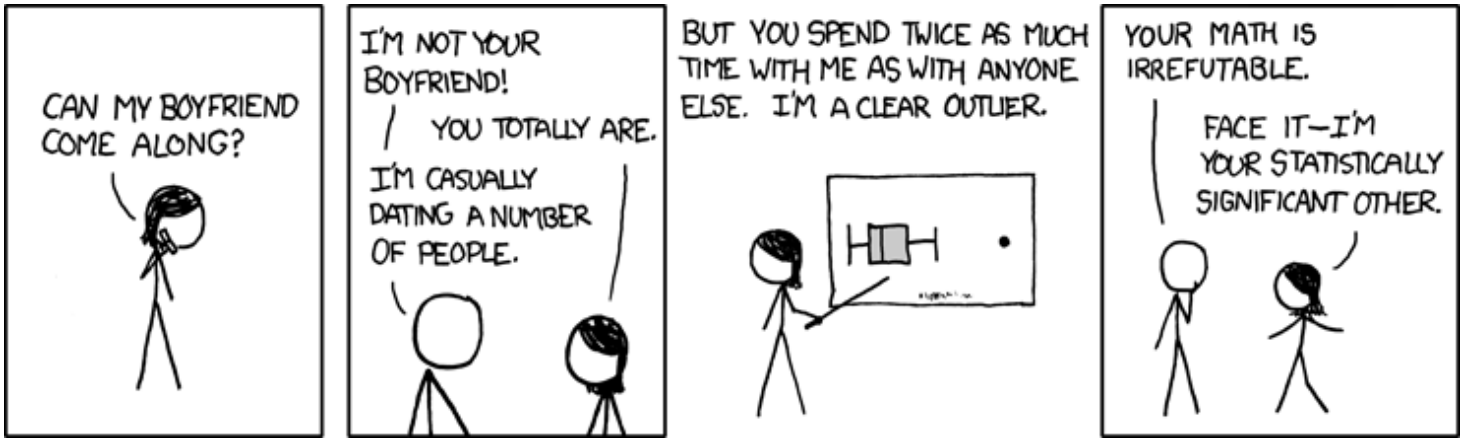


Statistics: understanding probabilities

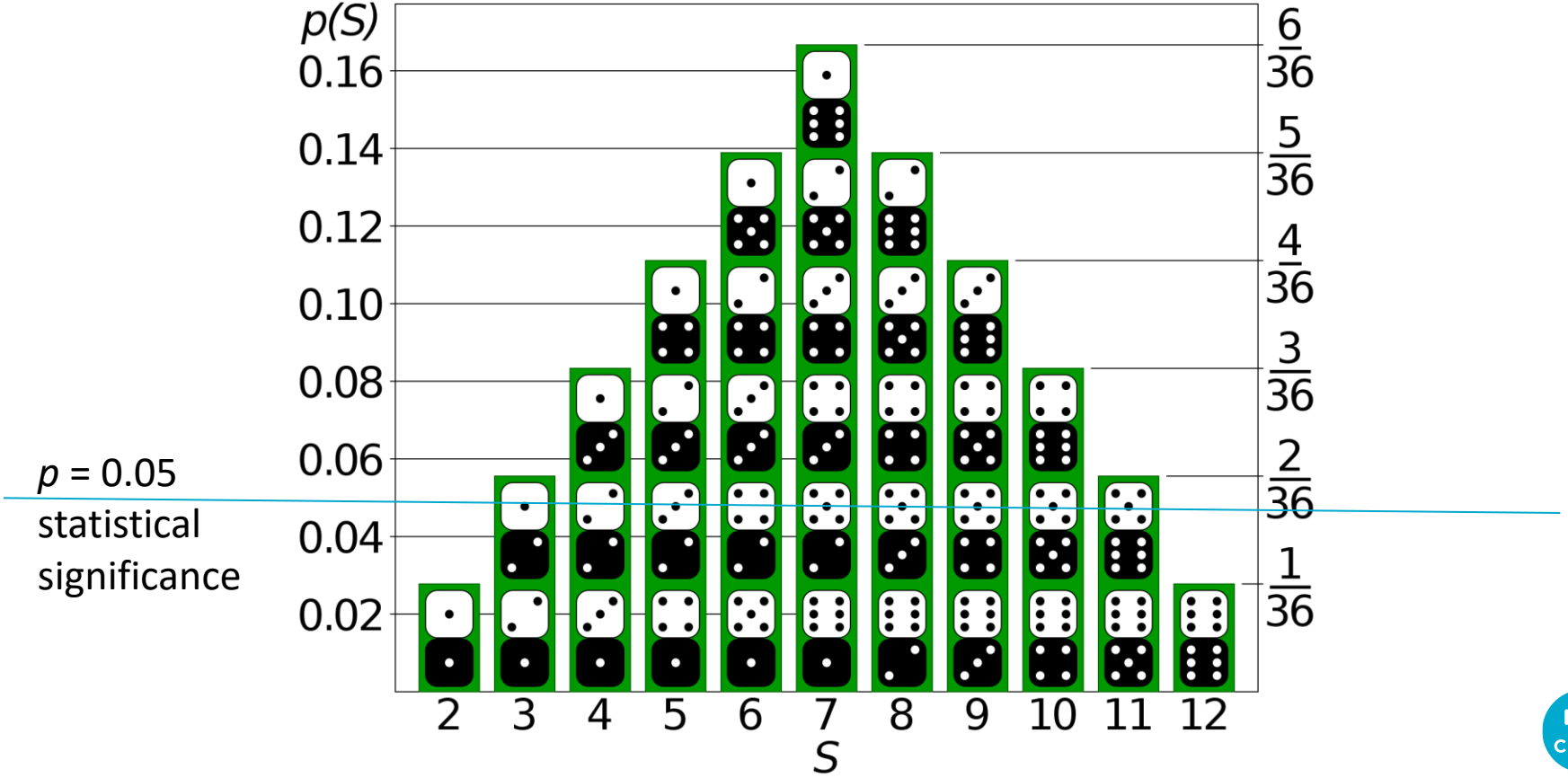


Statistical significance

- “p < 0.05”: what do the *p* values **mean**?



Statistics: understanding probabilities



Fisher’s exact test

- Is this observation **statistically significant**?

Fisher’s exact test

- ... depends on which bag of M&Ms you used!

$p < 0.05$

$p > 0.05$

Fisher's exact test

- For the heck of it, let's calculate some *p* values:
  - M&M website says that pack has ~50 candies
  - There are 5 colours, R G Y B O
  - We expect 10 green per packet
  - Cookie has 10 green M&Ms
- Set up Fisher's exact table

	Green M&M	Non-green M&M
Cookie	10	0
Not-on-cookie	0	40

Fisher’s exact test

https://www.graphpad.com/quickcalcs/contingency1/

GraphPad Software

Scientific Software Data Analysi

QuickCalcs

1. Select category 2. Choose calculator **3. Enter data** 4. View results

Analyze a 2x2 contingency table

Enter your data

Enter the number of subjects actually observed. Don't enter proportions, percentages or means.
Learn how to create a contingency table.

	green	non-green
cookie	10	0
not-on-cookie	0	40

https://www.graphpad.com/quickcalcs/contingency2/

GraphPad Software

Scientific Software Data Analysis

QuickCalcs

1. Select category 2. Choose calculator 3. Enter data **4. View results**

Analyze a 2x2 contingency table

	green	non-green	Total
cookie	10	0	10
not-on-cookie	0	40	40
Total	10	40	50

Fisher's exact test

The two-tailed P value is less than 0.0001
The association between rows (groups) and columns (outcomes) is considered to be extremely statistically significant.

Learn how to interpret the P value.

CSIRO

Fisher's exact test

- If we want precise *p* values, we can use R:

R Console

```
> fishers_matrix <- matrix(c(10,0,0,40), nrow=2)
> fishers_matrix
     [,1] [,2]
[1,]   10    0
[2,]    0   40
> fisher.test(fishers_matrix)

        Fisher's Exact Test for Count Data

data:  fishers_matrix
p-value = 9.735e-11
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 47.75873      Inf
sample estimates:
odds ratio
       Inf

>
```

BIO

Fisher's exact test, feat. biology

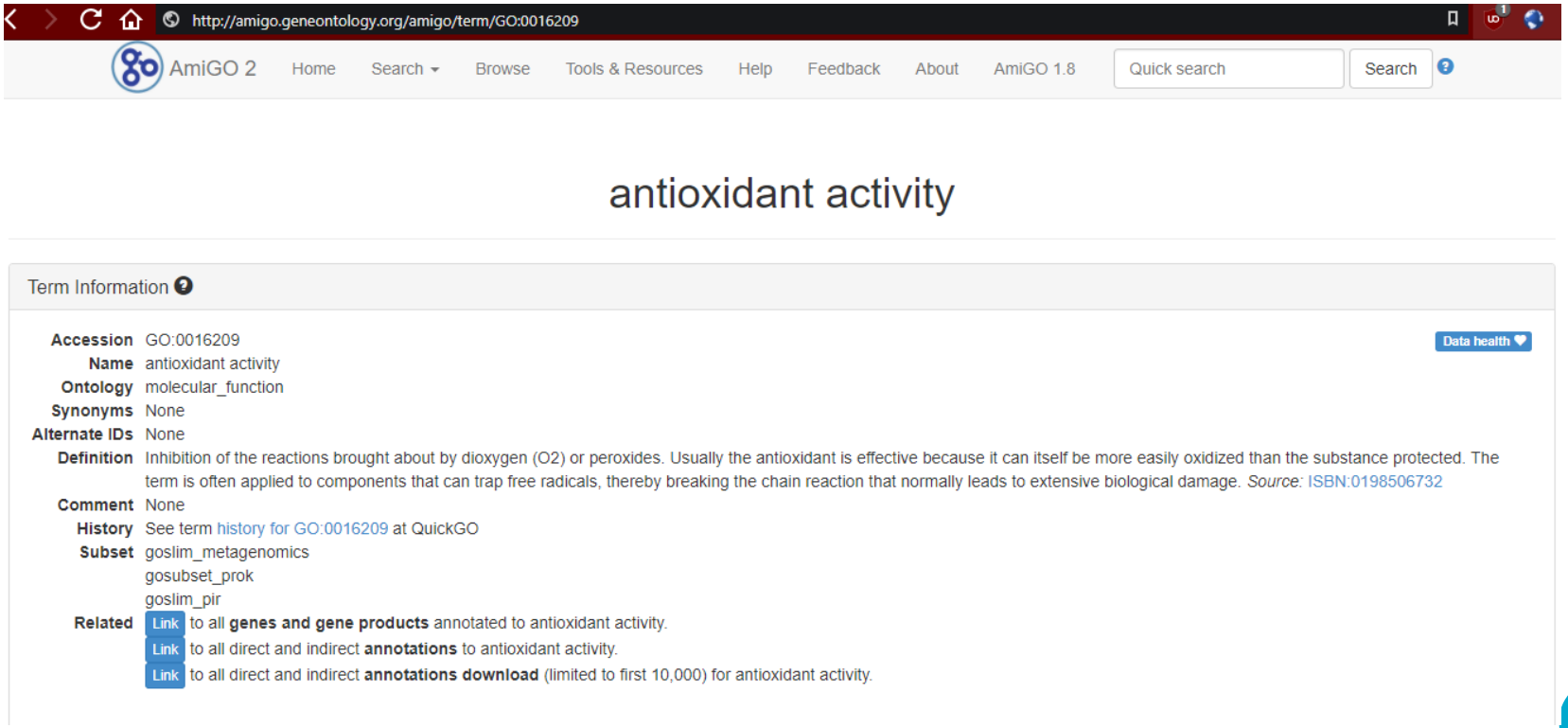
- Let's look at a more biological example
  - Perform a heat stress experiment
  - Obtain genes that were upregulated under stress
  - Check GO terms associated with these genes

	Genes with GO:0016209	Genes without GO:0016209
Upregulated	10	0
Not upregulated	0	40

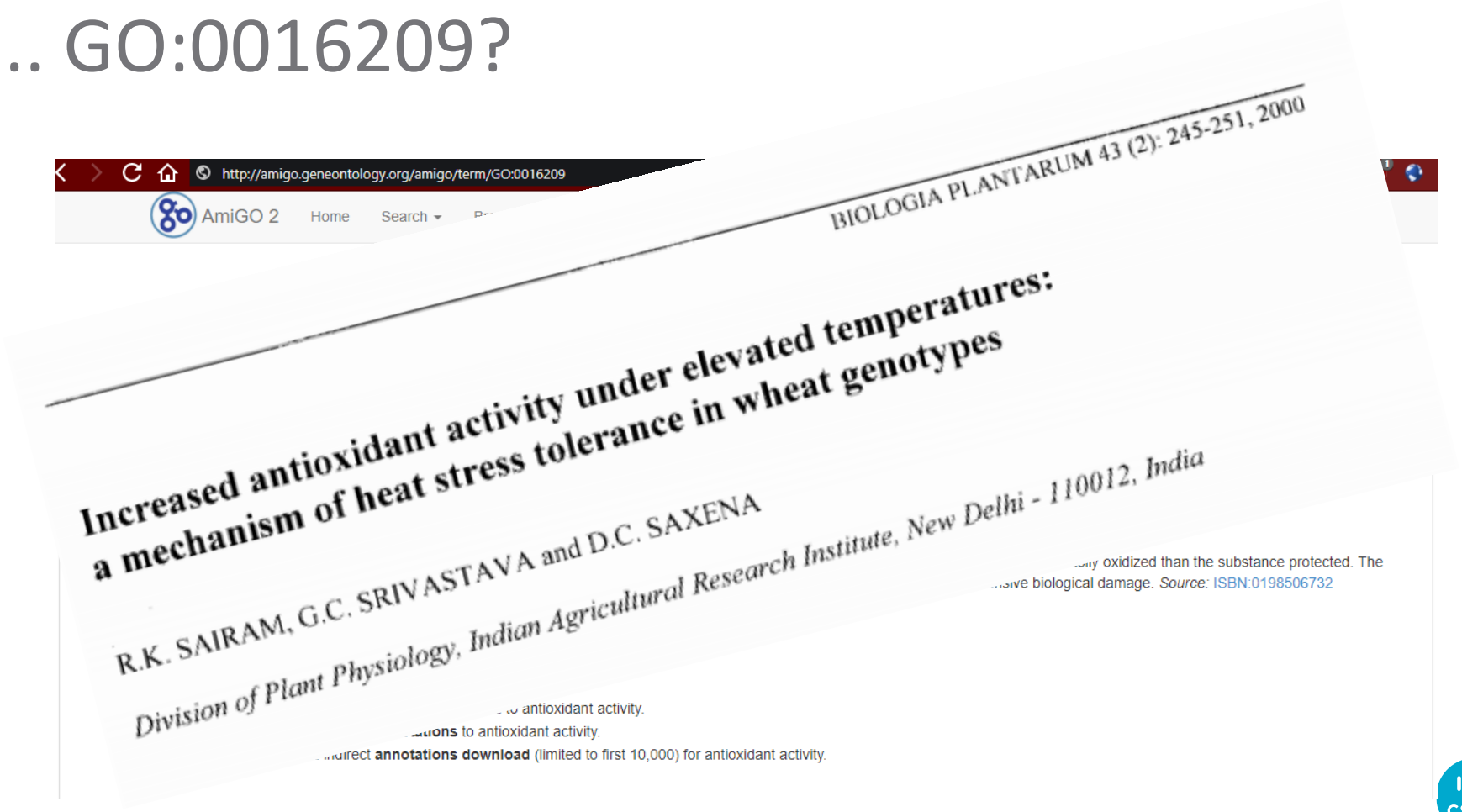
- $p = 9.735 \times 10^{-11}$

  = 0.00000000009735 (In English, this is "statistically very significant")

… GO:0016209?


http://amigo.geneontology.org/amigo/term/GO:0016209
AmiGO 2
Home
Search
Browse
Tools & Resources
Help
Feedback
About
AmiGO 1.8
Quick search
Search
antioxidant activity
Term Information
Accession GO:0016209
Name antioxidant activity
Ontology molecular_function
Synonyms None
Alternate IDs None
Definition Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can itself be more easily oxidized than the substance protected. The term is often applied to components that can trap free radicals, thereby breaking the chain reaction that normally leads to extensive biological damage. Source: ISBN:0198506732
Comment None
History See term history for GO:0016209 at QuickGO
Subset goslim_metagenomics
gosubset_prok
goslim_pir
Related Link to all genes and gene products annotated to antioxidant activity.
Link to all direct and indirect annotations to antioxidant activity.
Link to all direct and indirect annotations download (limited to first 10,000) for antioxidant activity.
Data health

… GO:0016209?


http://amigo.geneontology.org/amigo/term/GO:0016209
AmiGO 2
Home
Search
BIOLOGIA PLANTARUM 43 (2): 245-251, 2000
Increased antioxidant activity under elevated temperatures: a mechanism of heat stress tolerance in wheat genotypes
R.K. SAIRAM, G.C. SRIVASTAVA and D.C. SAXENA
Division of Plant Physiology, Indian Agricultural Research Institute, New Delhi - 110012, India
oxidized than the substance protected. The
ive biological damage. Source: ISBN:0198506732
antioxidant activity.
ations to antioxidant activity.
direct annotations download (limited to first 10,000) for antioxidant activity.

Knowledge of bio gives context to data


A CARTOON NETWORK ORIGINAL
the POWERPUFF GIRLS
POWER OF FOUR
CN
CARTOON NETWORK

Modern examples of bioinformatics

- Sequence analysis
  - Sequence searches: infer **function** of unknown DNA sequence
  - Comparative genomics: infer **evolutionary trees** from conserved proteins
  - Evolutionary biology: detect **gene duplication / horizontal gene transfer**
  - Mutational analysis: detect **predisposition to diseases** via SNP patterns
- Expression studies
  - Microarrays / RNA-seq: detect **upregulated or downregulated** genes
  - Protein mass spectrometry: deduce **function**, quantify **expression**
- Structural studies
  - Protein X-ray crystallography: calculate **most likely structure** of enzymes

Modern examples of bioinformatics

- Systems biology
  - “Interactome”: deduce **key proteins** from map of protein interactions
  - Pathway analysis: deduce **presence/absence of conserved pathways** in new genomes
- Image analysis
  - **Track movement** of cells, flies, fish, humans…
- Data mining
  - IBM’s “Watson” supercomputer chomps thru medical literature, helps provide **diagnosis** and detect whether **drug combinations have bad side effects**

Dessert

Questions / Demo / Freestyle

Thank you

Health and Biosecurity
Yi Jin Liew
Research Scientist

yijin.liew@csiro.au