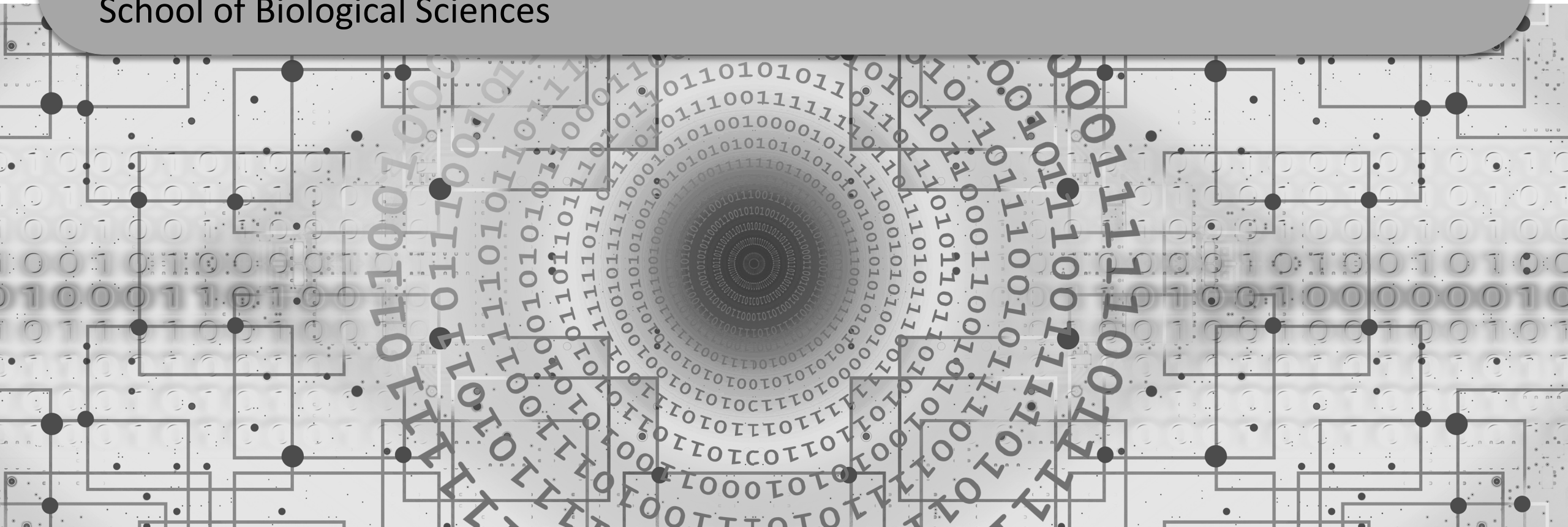


A Brief History of Data Science in Biology

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences

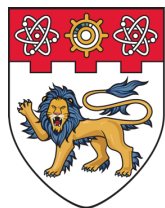


Learning Objectives

By the end of this topic, you should be able to:

- Describe the historical context and evolution of quantitative biology from bioinformatics to data science.
- Describe the specific applications of data science in biology.
- Describe the characteristics and applications of small, moderate and big data.
- Describe the future of biological data.





**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Historical Context

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



1800s: Earliest Instance of Biological “Big Data”

















Gregor Mendel
1822 - 1884

Established the power of “quantitative biology” (precursor of “biological data science”)

7 pea traits, or characters, studied by Mendel

1800s: Earliest Instance of Biological “Big Data”

7 pea traits, or characters, studied by

Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
						
Grey & Round	Yellow	White	Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
						
White & Wrinkled	Green	Violet	Constricted	Green	Terminal pods, Flowers top	Short ($\frac{3}{4}$ ft)
1	2	3	4	5	6	7

Source: By Mariana Ruiz LadyofHats [Public domain], via Wikimedia Commons

Established the power of “quantitative biology” (precursor of “biological data science”).

1800s: Earliest Instance of Biological “Big Data”

Data collection: Mendel's principles of inheritance was established through an analysis of some 30,000 pea plants.

Pattern recognition: Recognising the inheritance of certain traits could be explained by a few simple mathematical rules.

Pattern generalisation: Demonstrating that this observation also applies beyond peas for certain traits.

Data-centric Approach

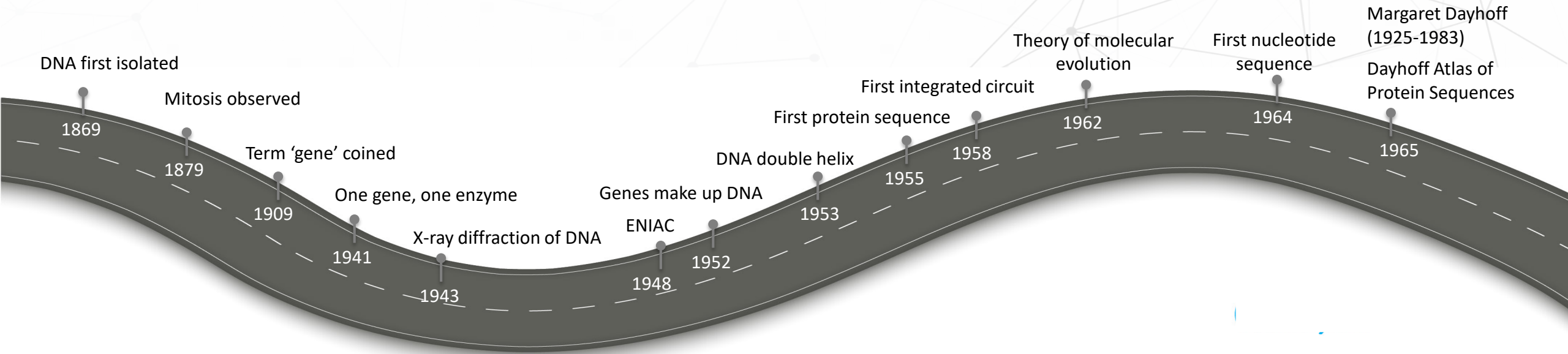
An **expanding collection of sequences** provided both a source of data and a set of interesting problems that were infeasible to solve without the number-crunching power of computers.

Sequence and structure is information and a central part of the conceptual framework of molecular biology.

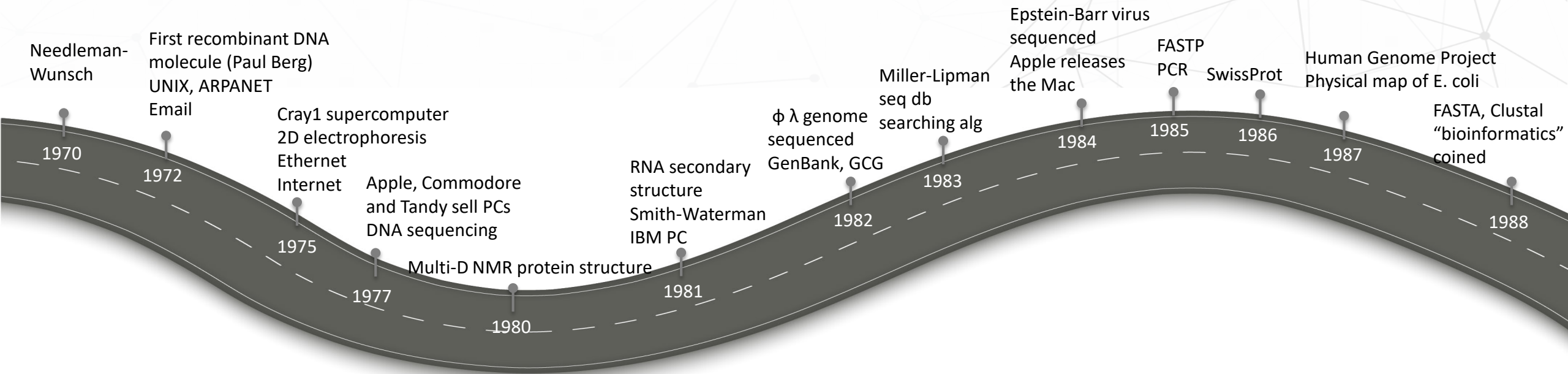
High-speed digital computers, which had developed from weapons research programmes during the Second World War, finally became widely available to academic biologists.

Why a data-centric approach became essential?

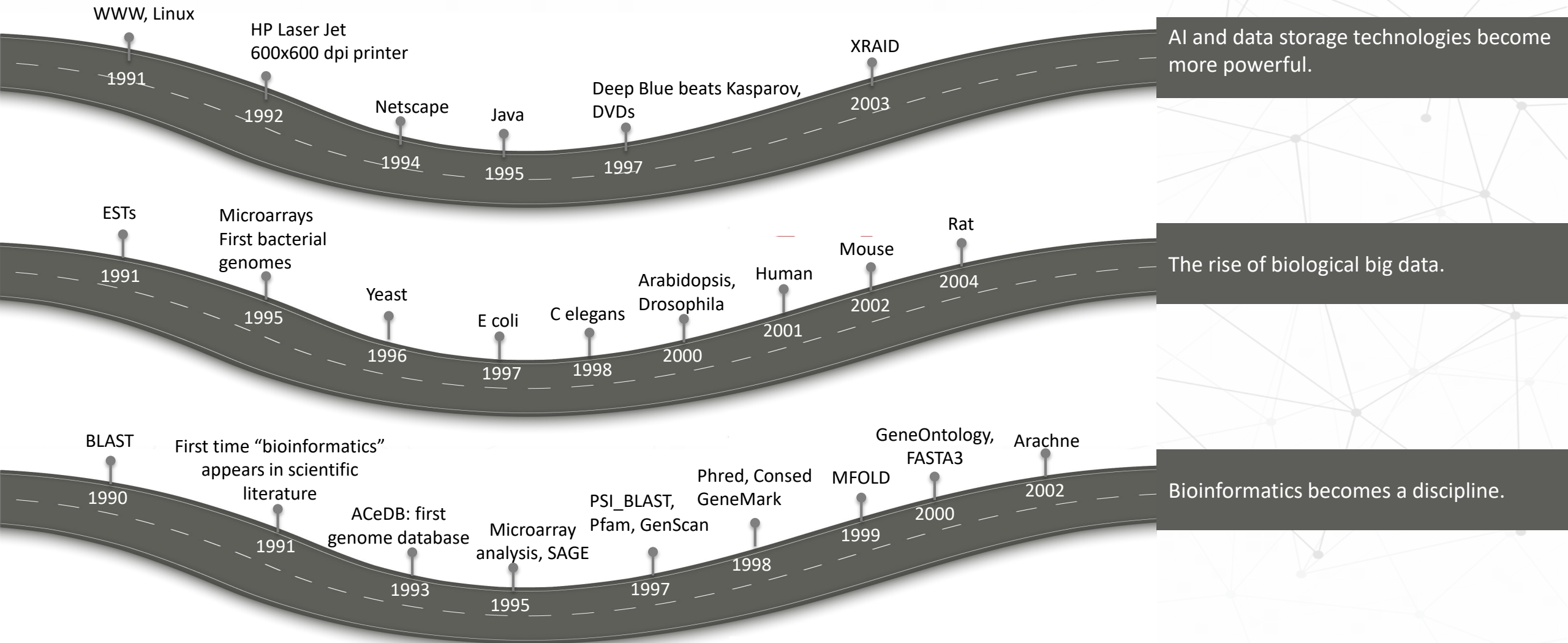
Rise of Big Data and Data Science



Rise of Big Data and Data Science



Rise of Big Data and Data Science

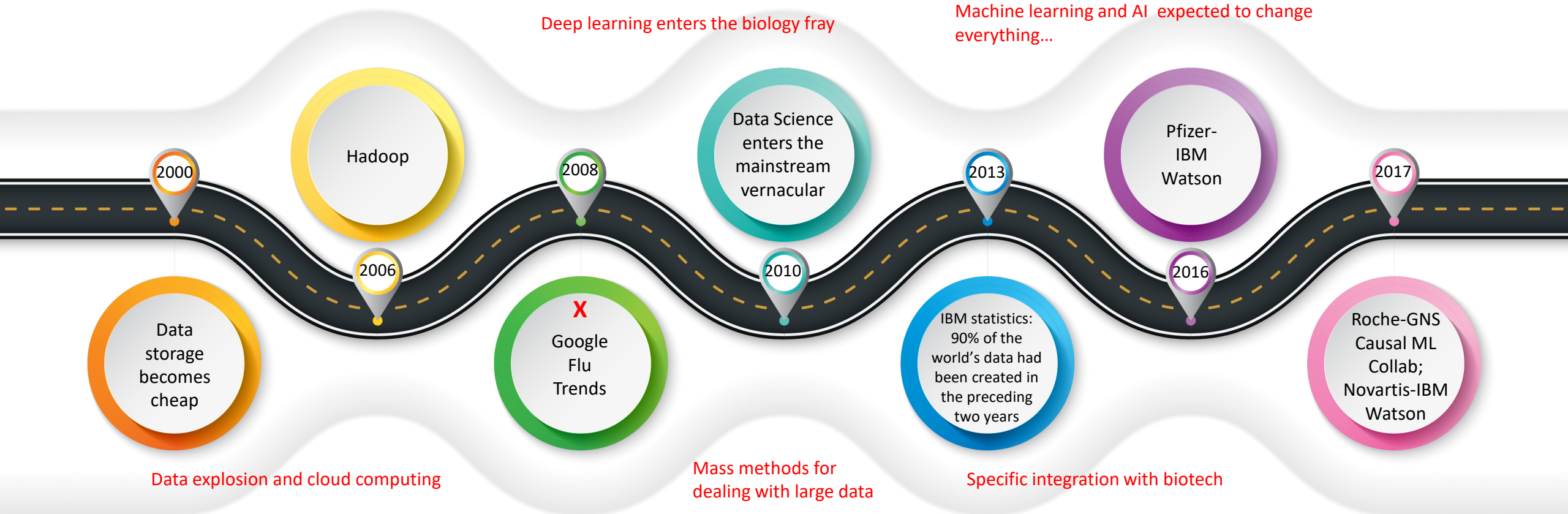


AI and data storage technologies become more powerful.

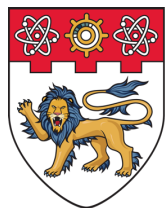
The rise of biological big data.

Bioinformatics becomes a discipline.

Age of Big Data and Data Science



Cheap Disks --> Big Data --> Cloud Computing --> Mass Analytic Tools --> Data Scientists --> Data Science Teams --> New Analytic Insights



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Biological Data Science

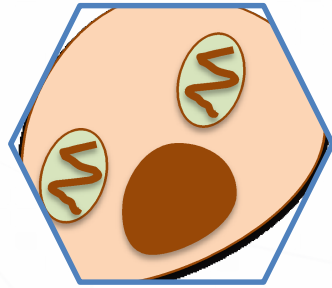
BS0004 Introduction to Data Science

Dr Wilson Goh

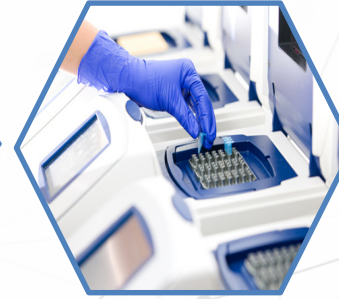
School of Biological Sciences



Biology as a Data-driven Science



DNA Sequencing Instruments



Super-resolution Digital Microscopy



Mass Spectrometer



Biology is becoming digitised.

Instruments produce a lot of raw data.

Greater throughput and resolution → Large Data

Instruments do not provide any meaningful interpretation on their own.

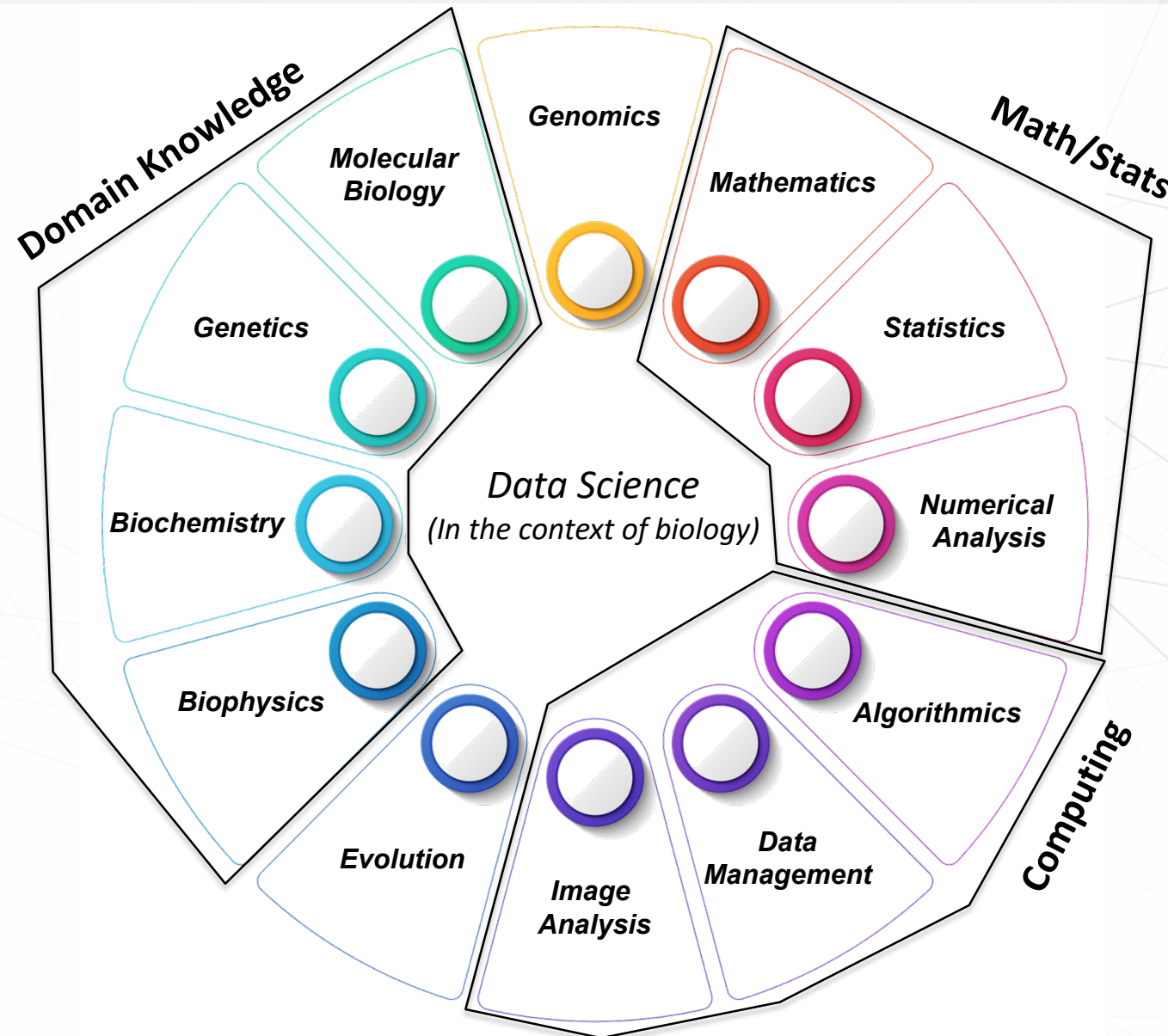
Data Science for Biology

Why data science for Biology will be challenging?

The power of data science comes from its ability to find relationships over very large numbers of observations, commonly stored in terabytes or petabytes of data.

However, given the size and complexities of these relationships, an exhaustive analytical pipeline requires an end-to-end integration of approaches, forming an analysis stack starting with data collection and continuing through computational and statistical evaluations toward higher-level biological interpretations and insights.

Highly Multidisciplinary



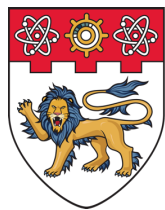
Problem with Multidisciplinarity

Scientists can not be experts in all the domains.

Solution is multidisciplinary teams and/ or multi-lab projects.

Problems:

- Biologists (generally) hate statistics and computers.
- Computer scientists (generally) ignore statistics and biology.
- Statisticians and mathematicians (generally):
 - Speak a strange language for any other human being.
 - Spend their time writing formula everywhere.
- Complexity of the biological domain:
 - Each time you try to formulate a rule, there is a possible counter-example.
 - Even the definition of a single word requires a book rather than a sentence (Exercise: find a consensual definition of "*gene*").



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Small Data Applications

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Small Data

- Small data is data that is 'small' enough for human comprehension.



- It is data in a volume and format that makes it accessible, informative and actionable.



- In today's "big data" world however, a third dimension meaning for small data is now coined:
Small data connects people with timely, meaningful insights (derived from big data and/or "local" sources), organised and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks.

- In other words, small data is the "purified gold" (insights) mined from the large mass of big data.



- What do you think? Go check out the comments section at <https://smalldatagroup.com/2013/10/18/defining-small-data/>

Characteristics of Small Data



In the hundreds of KB to MB range.



Limited samples --- ~1 to 10 range normally.



Can be analysed manually.



Can be analysed on a regular computer.

Small Data in Biology

Biology has traditionally been an observational rather than a deductive science. Although recent developments have not altered this basic orientation, the nature of the data has radically changed. It is arguable that until recently all biological observations were fundamentally anecdotal - admittedly with varying degrees of precision, some very high indeed.

--- Arthur Lesk



By ismb - 122-ISMBECCB15-TuesAM, CC BY 2.0,
<https://commons.wikimedia.org/w/index.php?curid=46139453>

Small Data in Biology

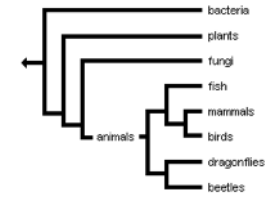
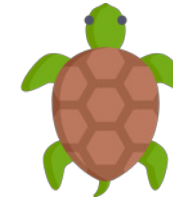
From its humble beginnings, biology is mostly about small data:

- Inference from relatively small numbers of observations.
 - Observation of Wildlife in the Galapagos -> Theory of evolution
 - Groupings of species by general characteristics -> Phylogeny
 - Understanding how disease occurs due to mutation by comparing sequences -> e.g. Sickle Cell Anemia, and many other examples
- Biology was limited by technology and availability of samples.
- And digital biological data is pretty much a new thing.

1953 Watson-Crick structure of DNA published.

1975 F. Sanger, and independently A. Maxam and W. Gilbert, develop methods for sequencing DNA.

1977 Bacteriophage ϕ X-174 sequenced: First 'complete genome.'



Small Data in Biology

- Biology is rife with sequence information (e.g. a sequence may correspond to a gene, an mRNA or a protein).
- Sequence is correlated with function (e.g. the p53 gene sequence corresponds to an oncogene which drives cancer).
- Sequence is also data --- when we are dealing with a small number of sequences, this is a small data problem.
- There are many useful things we can do with small data:
 - One of the most obvious being to compare 2 strings to see how similar they are (with similarity being a proxy for evolutionary relationships) --- this is also known as pairwise sequence comparison (e.g. BLAST).
 - Pairwise sequence comparisons may be generalised towards simultaneous comparisons of > 2 sequences at once --- this is known as multiple sequence comparison (e.g. T-COFFEE and MUSCLE/ Multiple Sequence Comparison by Log-Expectation).
 - Note: Sequence comparison of biological data is essentially an application of the string matching problem in CS to biology.

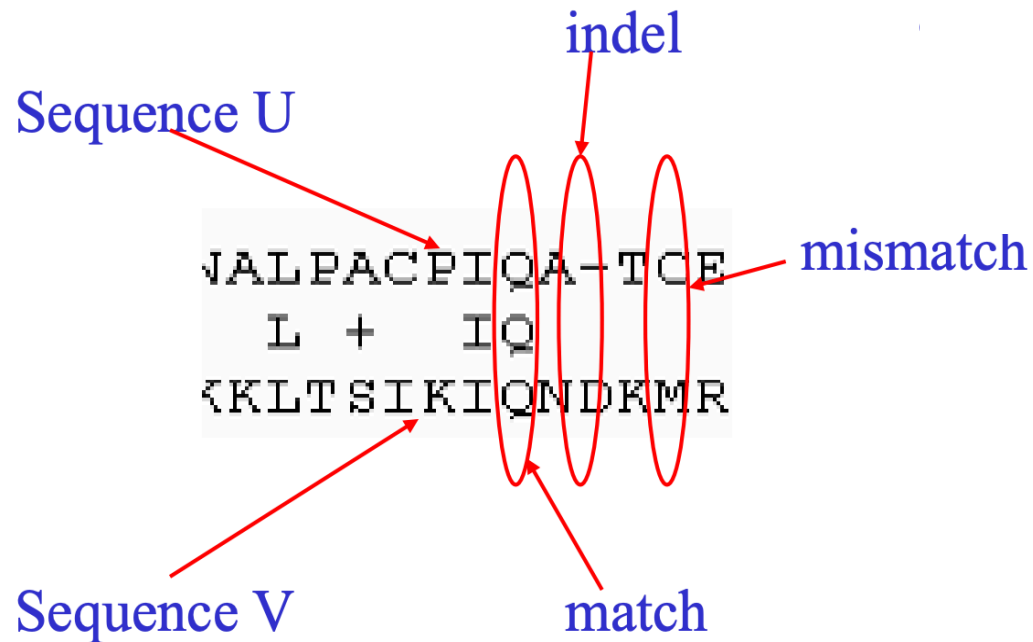
Sequence Alignment

An early example:

Doolittle et al. (Science, July 1983) searched for platelet-derived growth factor (PDGF) in his own DB. He found that PDGF is similar to v-sis oncogene

```
PDGF-2  1      SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34
p28sis 61  LARGKRSLGSLVAEPAMIAECKTRTEVFESRRLIDRTN 100
```

Sequence Alignment



- Key aspect of sequence comparison is sequence alignment.
- A sequence alignment maximises the number of positions that are in agreement in two sequences.

Sequence Alignment

- **Infer protein function**

When two protein look similar, we conjecture they come from the same ancestor and inherit the ancestor's function (i.e. they are homologous).

- **Find evolution distance between two species**

Evolution modifies the DNA of species -> Similarity of their genome correlates with their evolutionary distance.

- **Help genome assembly**

Human genome project reconstructs the whole genome based on overlapping info of a huge amount of short DNA pieces.

Global and Local Sequence Alignments

There are two types of pairwise alignments, *local* and *global* alignments.

A **local alignment** is an alignment of two sub-regions of a pair of sequences. This type of alignment is appropriate when aligning two segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence.



A **global alignment** is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment, the sequences are assumed to be homologous along their entire length.

Scoring Systems in Pairwise Alignments

Scoring systems in pairwise alignments

In order to align a pair of sequences, a scoring system is required to score matches and mismatches. The scoring system can be as simple as “+1” for a match and “-1” for a mismatch between the pair of sequences at any given site of comparison. However substitutions, insertions and deletions occur at different rates over evolutionary time.

This variation in rates is the result of a large number of factors, including the mutation process, genetic drift and natural selection. For protein sequences, the relative rates of different substitutions can be empirically determined by comparing a large number of related sequences. These empirical measurements can then form the basis of a scoring system for aligning subsequent sequences. Many scoring systems have been developed in this way. These matrices incorporate the evolutionary preferences for certain substitutions over other kinds of substitutions in the form of log-odd scores. Popular matrices used for protein alignments are [BLOSUM](#) and PAM matrices.

Note: The BLOSUM and PAM matrices are substitution matrices. The number of a BLOSUM matrix indicates the threshold (%) similarity between the sequences originally used to create the matrix. BLOSUM matrices with higher numbers are more suitable for aligning closely related sequences. For PAM, the lower numbered tables are for closely related sequences and higher numbered PAMs are for more distant groups.

Algorithms for Pairwise Alignments

Algorithms for pairwise alignments

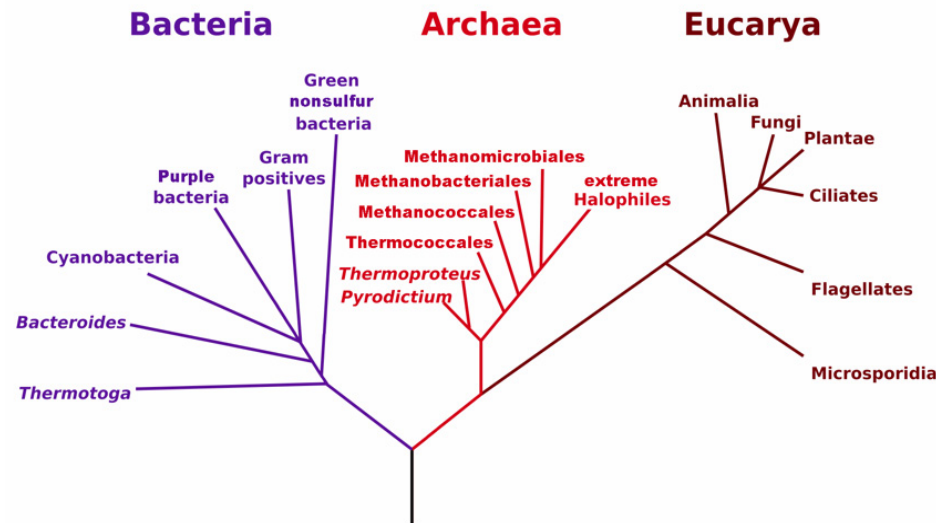
Once a scoring system has been chosen, we need an algorithm to find the optimal alignment of two sequences. This is done by inserting gaps in order to maximise the alignment score. If the sequences are related along their entire sequence, a global alignment is appropriate. However, if the relatedness of the sequences is unknown or they are expected to share only small regions of similarity, (such as a common domain) then a local alignment is more appropriate.

An efficient algorithm for global alignment was described by [Needleman and Wunsch 1970](#), and their algorithm was later extended by [Gotoh 1982](#) to model gaps more accurately. For local alignments, the [Smith-Waterman](#) algorithm is the most commonly used.

Can use this to Model Evolutionary Relationships

Ribosomal RNAs turned out to have the essential feature of being present in all organisms, with the right degree of divergence. (Too much or too little divergence and relationships become invisible.) On the basis of 16S ribosomal RNAs, C. Woese divided living things most fundamentally into three Domains (a level above Kingdom in the hierarchy): Bacteria, Archaea and Eukarya.

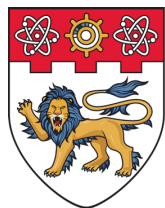
Phylogenetic Tree of Life



Major divisions of living things, derived by C. Woese on the basis of 16S RNA sequences.

Trying them out for your own

- Pairwise Comparison:
 - Global
 - EMBOSS Needle (<https://www.ebi.ac.uk/Tools/psa/>)
 - EMBOSS Stretcher (<https://www.ebi.ac.uk/Tools/psa/>)
 - Local
 - EMBOSS Water (<https://www.ebi.ac.uk/Tools/psa/>)
 - EMBOSS Matcher (<https://www.ebi.ac.uk/Tools/psa/>)
- Multiple Comparison
 - T-COFFEE (<http://tcoffee.crg.cat/>)
 - MUSCLE (<https://www.ebi.ac.uk/Tools/msa/muscle/>)
- Heuristics
 - BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Moderate Data Applications

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Moderate Data

- Moderate data is data that forms the current readouts from most moderate day instruments.
 - This includes the high-throughput reads/assaying of biological entities (the genome, transcriptome and proteome).
 - Although these platforms can mine deeply for all genes, they are ultimately limited in sample size and/or resolution and does not qualify as big data.
- It is data in a volume and format that while accessible and informative, requires some degree of work and downstream analysis to make it actionable.
 - Downstream analysis involves a philosophy known as comparative study (aka comparative analysis).
- **Moderate data analysis currently dominates the –omics landscape in biological research.**

Characteristics of Moderate Data



In the hundreds of MB to GB range.



Limited to almost generous samples ~5 to 100.



Difficult to analyse manually.



Can be analysed on a regular computer.

What is comparative analysis?

- Comparative studies in biology use an investigative philosophy that many scientists identify as the “comparative method.”
- In one sense, for those concerned with evolutionary history, the comparative method provides insights into adaptation by correlating differences among species with ecological factors (Futuyma 1986).
- In another sense, biologists often study the particular features of one species to learn about some aspect of a second species.
- **In other words, you need 2 things: Factors, which are variables you can measure, such as weight, height, etc. And Classes, groups which you can compare against, such as gender (male vs female), species (man vs chimp), etc.**

https://openlibrary.org/books/OL2722009M/Evolutionary_biology

Comparative Analysis and its Ancient Roots

The idea that the systems of an organism may be better understood by the use of comparison and contrast among organisms is an ancient one:

- Aristotle (384–322 BC) sought common characters of organisms as a means of classification and explanation.
- Cole (1944) cites the writings of the Hippocratic School (4th century BC) concerning an attempt to compare the human skeleton to that of other vertebrates.
- Gardner (1965) states that Galen (AD 130–200) based his textbook of human anatomy, *On Anatomical Preparations*, on “dissections of such animals as sheep, oxen, dogs, bears, and apes.”
- Cole (1944) also cites Crie (1882) who refers to Belon as “the father of comparative anatomy.”
- Belon's work (1555), *L'Histoire de la Nature des Oyseaux*, in which the skeletal structures of birds are compared to those of humans, was one of the first explicit applications of the comparative method in biology.

The new tech heavy -omics sciences is based on the old scientific tradition!

Comparative Analysis Revisited

In recent years, the term *comparative method/analysis/study* is increasingly used to refer to a set of statistical procedures for achieving various purposes:

- Reconstructing phylogenies and for controlling for phylogenetic effects during inter- and intrataxon comparisons ([Harvey and Pagel 1991](#)).
- The major concern with statistical comparisons across taxa is the failure to account for the role of identity by descent in producing shared characteristics ([Felsenstein 1985](#)) -> convergence without true relationships e.g. birds= insects because they both have wings!
- In the case of taxa analysis, features are “constructs” that are engineered. For example, you may choose to measure the length of the wings, or the width of the legs. Whatever you choose to measure, is a variable that you have constructed/engineered.
 - Constructed/engineered features may or may not be informative.
 - There is also the element of choice/design (you may choose however you wished to analyse some anatomical feature).

Comparative Analysis Revisited

- Biological features have now changed with the advancement of technology... from anatomical features -> gene sequences and copy number -> gene expression -> protein expression.
 - In -omics analysis, the engineered feature is determined by the technology
 - In DNA chip, the feature is DNA copy number
 - In microarray, the feature is gene expression
 - In proteomics, the feature is protein expression
- In such cases, you **do not get a choice** on what variables you want to measure. They are predefined.
- Additionally, because there are so many, some of these variables could be potentially informative.
- And so, what you want to do, is to hone-in on some of these informative signals out of a sea of no signal or sea of noise.

Comparative Analysis using Gene Expression Microarrays

Pause the video and read this.

Comparative genomic analysis of primary tumors and metastases in breast cancer.

Personalised medicine uses genomic information for selecting therapy in patients with metastatic cancer. An issue is the optimal tissue source (primary tumor or metastasis) for testing. We compared the **DNA copy number** and **mutational profiles** of **primary breast cancers and paired metastases** from **23 patients** using **whole-genome array-comparative genomic hybridisation** and **next-generation sequencing of 365 “cancer-associated” genes**. Primary tumors and metastases harbored copy number alterations (CNAs) and mutations common in breast cancer and showed **concordant profiles**. The global concordance regarding CNAs was shown by **clustering** and **correlation matrix**, which showed that **each metastasis correlated more strongly with its paired tumor than with other samples**. Genes with recurrent amplifications in breast cancer showed 100% (*ERBB2*, *FGFR1*), 96% (*CCND1*), and 88% (*MYC*) concordance for the amplified/non-amplified status. **Among all samples, 499 mutations were identified**, including 39 recurrent (*AKT1*, *ERBB2*, *PIK3CA*, *TP53*) and 460 non-recurrent variants. The tumors/metastases concordance of variants was 75%, higher for recurrent (92%) than for non-recurrent (73%) variants. Further mutational discordance came from very different variant allele frequencies for some variants. **We showed that the chosen targeted therapy in two clinical trials of personalised medicine would be concordant in all but one patient (96%) when based on the molecular profiling of tumor and paired metastasis**. Our results suggest that the genotyping of primary tumor may be acceptable to guide systemic treatment if the metastatic sample is not obtainable. **However, given the rare but potentially relevant divergences for some actionable driver genes, the profiling of metastatic sample is recommended.**

Keywords: array-CGH, breast cancer, genomics, metastasis, sequencing

<https://www.ncbi.nlm.nih.gov/pubmed/27028851>

Comparative Analysis using Gene Expression Microarrays

Try filling this in yourself first

Data Summary	Attributes
Classes	
Sample size	
What is measured?	
What kind of metrics are being used?	
Are all genes being monitored? (Is this a global screen for all genes?)	
Data analysis plan	
Question/Hypothesis	

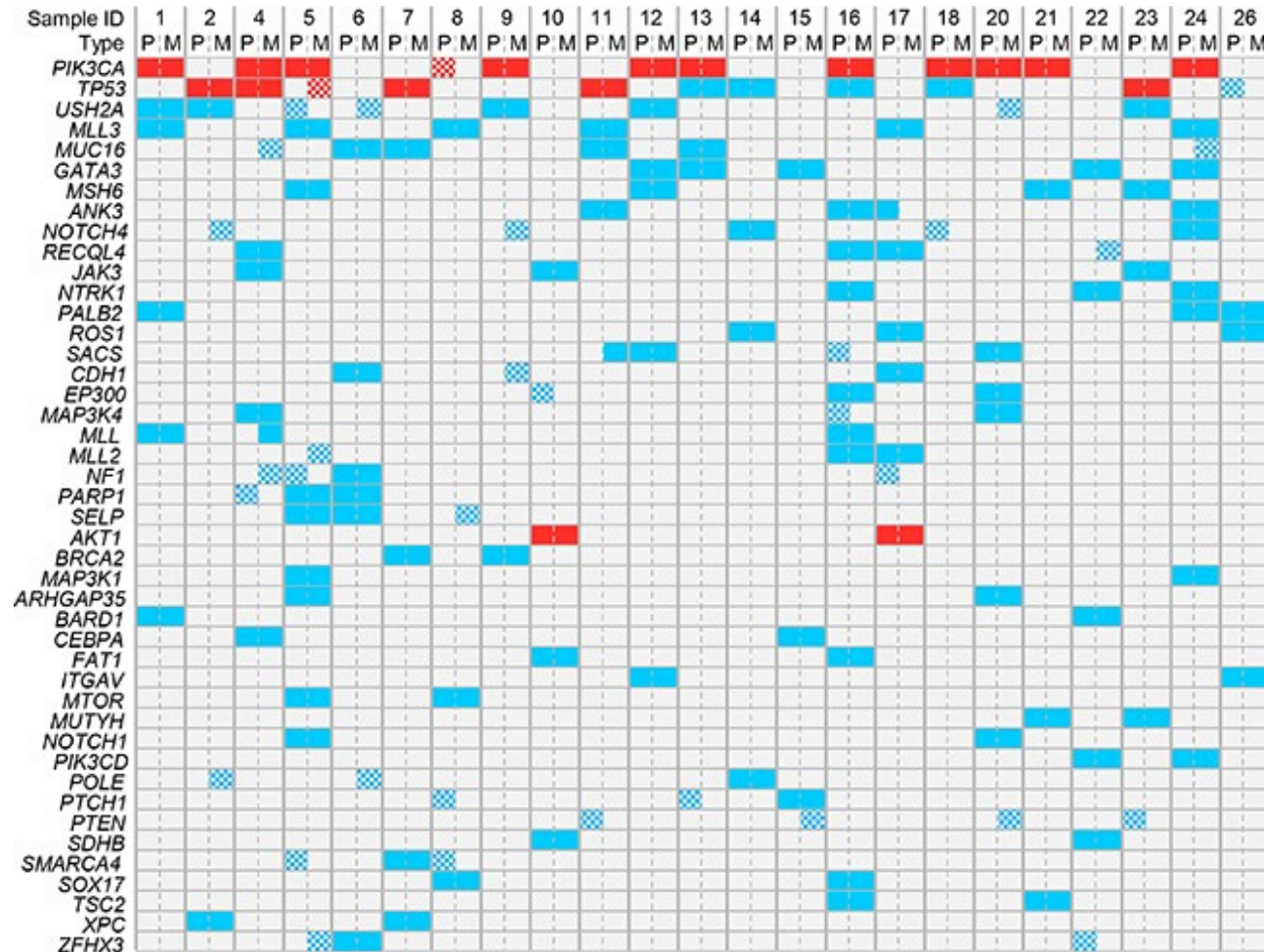
<https://www.ncbi.nlm.nih.gov/pubmed/27028851>

Comparative Analysis using Gene Expression Microarrays

Data Summary	Attributes
Classes	Primary tumor vs metastatic
Sample size	23 (2 samples per patient)
What is measured?	DNA copy number changes (DNA copy number) and gene expression (NGS)
What kind of metrics are being used?	Correlation (to measure similarity between samples) and Clustering (to see which samples are most similar to each other)
Are all genes being monitored? (Is this a global screen for all genes?)	No. Only profiles of 365 cancer genes are looked at. This is a targeted screen. Also, out of 499 mutations being monitored, 39 are recurrent.
Data analysis plan	Since each patient yields 2 sets of samples. It should be logically a “paired” setup involving comparisons of 2 samples against each other, per patient. (Look at independent vs paired testing)
Question/Hypothesis	If there is a metastasis, can we use the primary tumor to guide personalised treatment (Esp if the metastases is inaccessible). In other words, should we worry about high divergence between the primary tumor and metastases?

<https://www.ncbi.nlm.nih.gov/pubmed/27028851>

Comparative Analysis using Gene Expression Microarrays



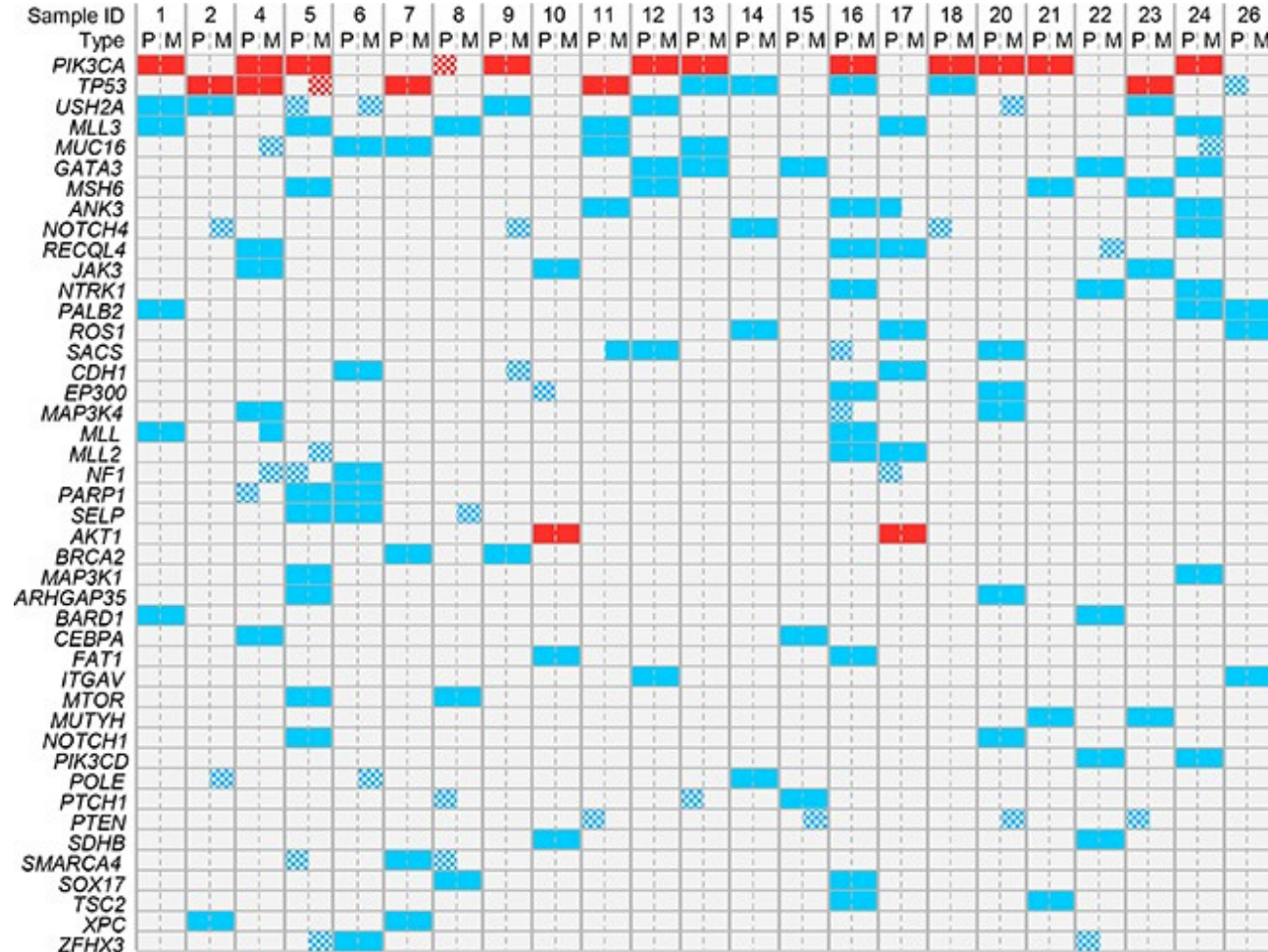
Distribution of mutations in all samples

The mutations present in at least 4 out of 46 samples are shown. Genes are ordered from top to bottom by decreasing frequency of mutations. Samples are ordered by patient number. Recurrent mutations are in red and non-recurrent mutations are in blue. The checkerboard pattern indicates the discordant mutations between primary tumors (P) and paired metastases (M).

Take a while to look and analyse this plot. And answer the following questions.

<https://www.ncbi.nlm.nih.gov/pubmed/27028851>

Comparative Analysis using Gene Expression Microarrays

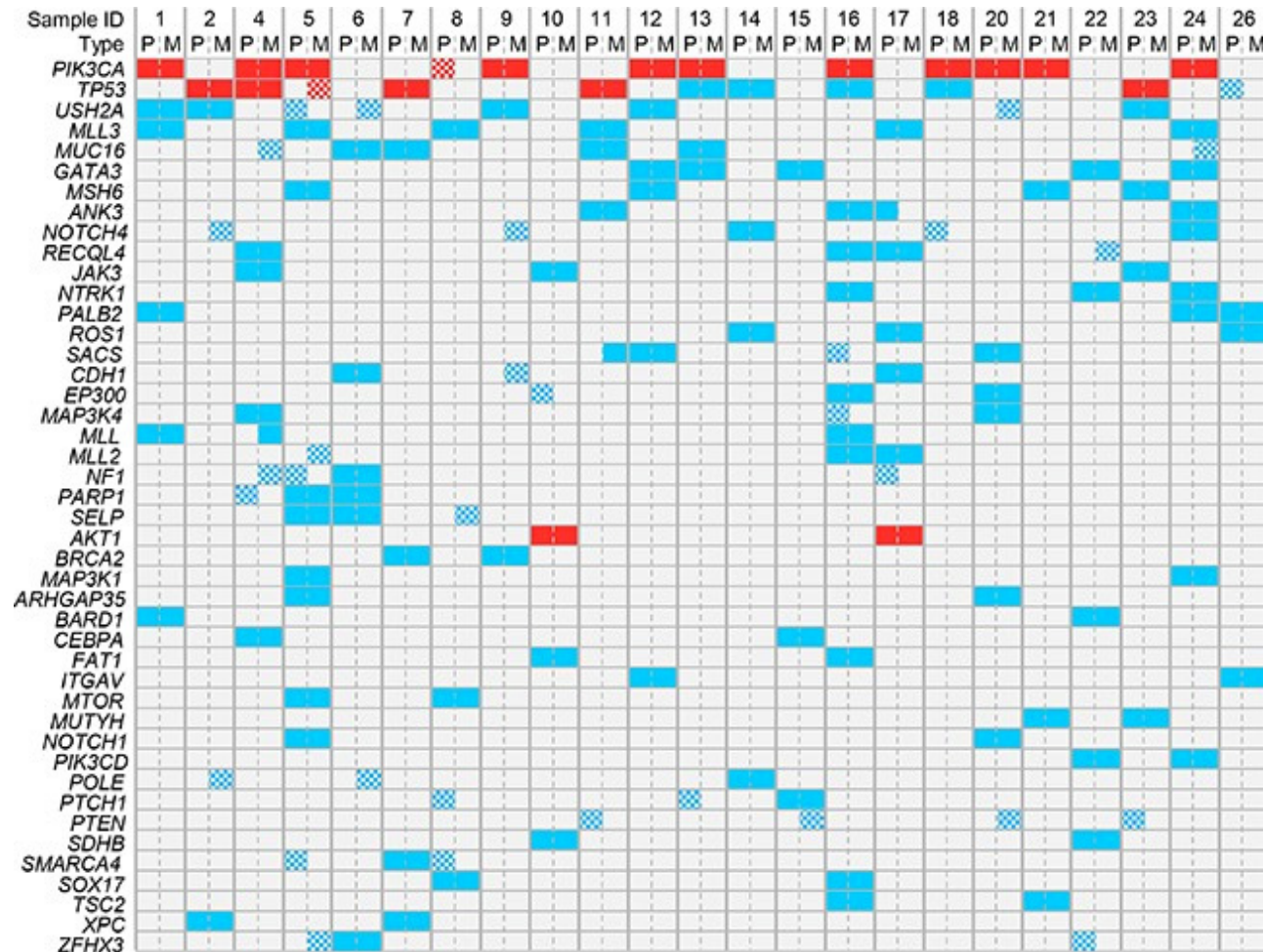


Qn 1: Based only on this plot, are recurrent mutations associated with higher prevalence amongst the 23 patients?

Ans: Yes. This applies specifically to PIK3CA. However, it could also be said that PIK3CA is strongly enriched for recurrent mutations.

<https://www.ncbi.nlm.nih.gov/pubmed/27028851>

Comparative Analysis using Gene Expression Microarrays

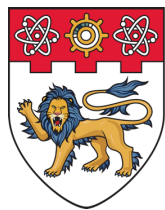


Qn 2: Based only on this plot, would you conclude primary tumour are similar to metastases?

Ans: Yes.

Checkerboards (signifying discordance) are relatively rare. Primary tumours between samples differ greatly from each other (see ID1 to 23). However, they are most similar to themselves, including the spawned metastases (see ID1 within and compare against ID2 for a start). In other words, we can use the primary tumour to guide treatment strategy usually. You may also notice that discordant events are not randomly distributed. Those with discordant events may have worse prognosis. What should we do?

<https://www.ncbi.nlm.nih.gov/pubmed/27028851>



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Big Data Applications

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Big Data

- Big data is data that are stored and accessible from cloud-computing platforms and massive data warehouses.
 - Examples in biology include PRIDE (PRoteomics IDentifications database) and GEO (Gene Expression Omnibus).
- It is data in a volume and format that is not easily accessible due to its size and requires extensive mining to extract insight, requires a lot of degree of work, specialised downstream expertise to make it actionable.
 - Comparative analysis is still possible, but may be too simplistic to make full use of the data.
- **Big data analysis is the way to look towards as biological becomes increasingly digitised and open access databases get larger and larger.**

Characteristics of Big Data



In the hundreds of GB to PB range.



Large number of samples 100 upwards.

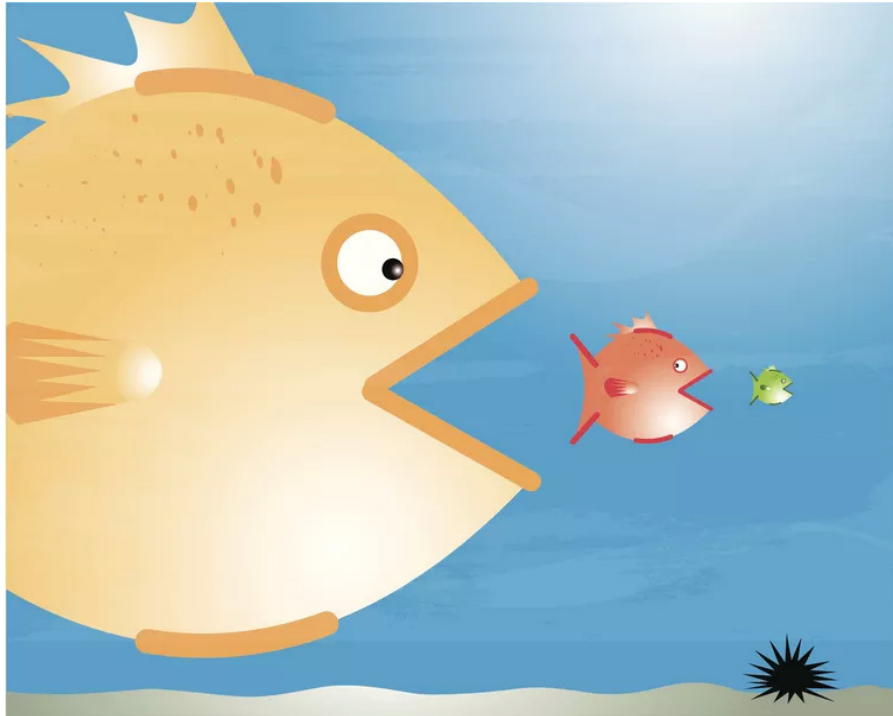


Cannot be analysed manually.



Cannot be analysed on a regular computer – requires novel solutions e.g. cloud-based computing, parallel processing, etc.

Big Data



All these computer technology storage units of measurement are based on the *byte*, which is the amount of storage required to store a single character of text.

- **petabyte** (PB), which is larger than a,
- **terabyte** (TB), which is larger than a,
- **gigabyte** (GB), which is larger than a,
- **megabyte** (MB), which is larger than a,
- **kilobyte** (KB), which is larger than a,
- **byte** (B).

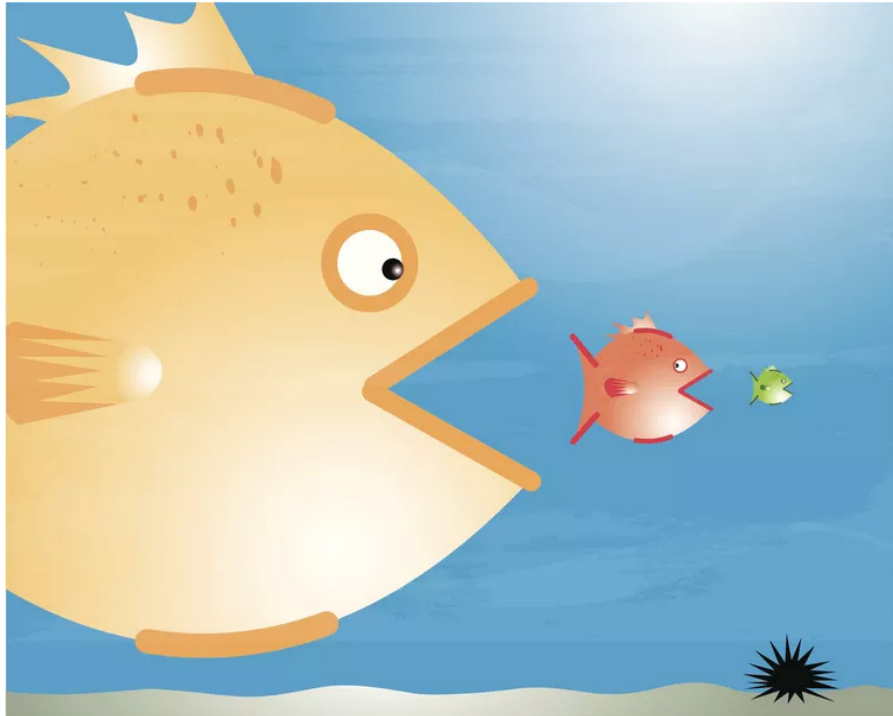
There are 1,024 MB in 1 GB.

There are 1,024 GB in 1 TB.

There are 1,024 TB in 1 PB.

As of 2018, most new, average priced computer hard drives are in the **1 to 3 TB range**.

The Byte Table



Metric	Value	Bytes
Byte (B)	1	1
Kilobyte (KB)	$1,024^1$	1,024
Megabyte (MB)	$1,024^2$	1,048,576
Gigabyte (GB)	$1,024^3$	1,073,741,824
Terabyte (TB)	$1,024^4$	1,099,511,627,776
Petabyte (PB)	$1,024^5$	1,125,899,906,842,624
Exabyte (EB)	$1,024^6$	1,152,921,504,606,846,976
Zettabyte (ZB)	$1,024^7$	1,180,591,620,717,411,303,424
Yottabyte (YB)	$1,024^8$	1,208,925,819,614,629,174,706,176

This is just for reference. Please do not memorise.

Our Current Lives in the GB Age

Talking about the GB is a bit more commonplace—we see GBs everywhere, from memory cards, to movie downloads, smartphone [data plans](#), and more. A single GB is equivalent to **a little over 700 floppy disks** or **just over a single CD**.

A GB is not a small number by any means, but these days it's a level of data we use up quickly, sometimes several times over each day. It's a number we very much run up against on a regular basis.

- 1 GB can store **almost 300 songs** in [MP3](#) format.
- A single HD [Netflix](#) movie might **gobble up over 4 GB** as you watch. A 4K version might run **over 20 GB!**
- A DVD movie disc **holds about 9.4 GB**.
- Most [smartphones](#) store **64 GB or 128 GB of data** (your apps, music downloads, etc.).
- Your smartphone data plan, which you use when you're away from your wireless network at home, might be capped at **5 GB, 10 GB, or a bit more** per month.

Like we showed in the MB to GB conversion a few sections above, 1 GB is the same as **over one billion bytes**. That's no small number, but it's not nearly as impressive of an amount as it once was.

We are just pushing into the TB world

- A single TB is a *lot* of space. It would take **1,498 CD-ROM discs** to store just 1 TB worth of information.
- As of 2018, most new, average priced computer hard drives are in the **1 to 3 TB range**.
- Many [ISPs](#) cap **monthly data usage at 1 TB**.
- An 4th generation Playstation (or game console) ships with 1 TB hard drive (and a current generation video game is about 10-50 GB).
- Around **130,000 digital photos would require 1 TB** of space...close to 400 photos every day for a year!
- IBM's famous Watson game-playing **supercomputer has 16 TB of RAM**.
- We are still not yet seeing the Peta-byte in our everyday lives.

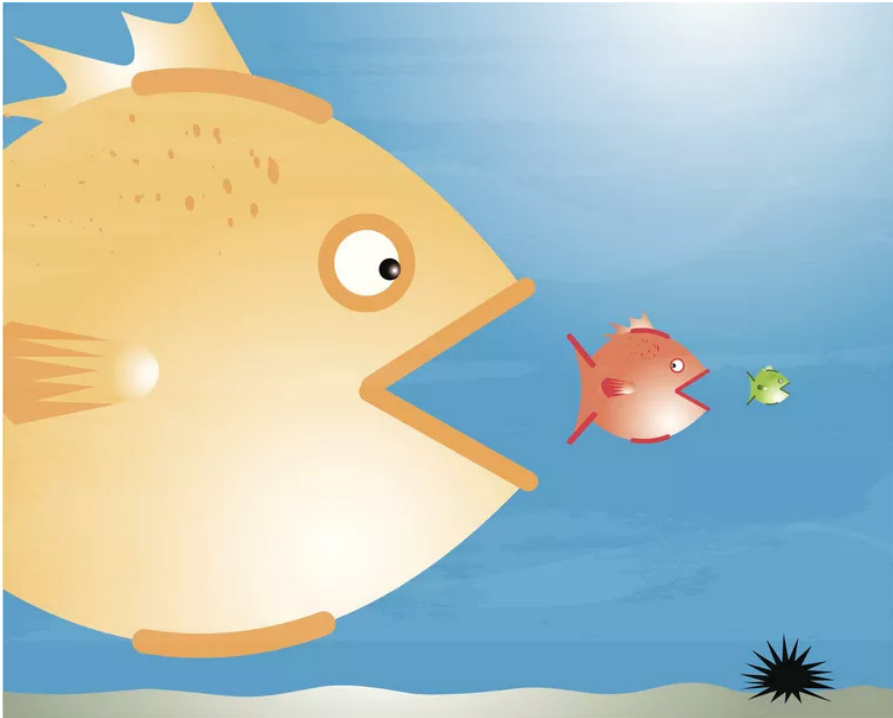
Real big data in our every day lives (behind the scenes)

Organisation (2014 statistics)	Est. amount of data processed per day	Source
eBay	100 pb	http://www-conf.slac.stanford.edu/xldb11/talks/xldb2011_tue_1055_TomFastner.pdf
Google	100 pb	http://www.slideshare.net/kmstechnology/big-data-overview-2013-2014
Baidu	10-100 pb	http://on-demand.gputechconf.com/gtc/2014/presentations/S4651-deep-learning-meets-heterogeneous-computing.pdf
NSA	29 pb	http://arstechnica.com/information-technology/2013/08/the-1-6-percent-of-the-internet-that-nsa-touches-is-bigger-than-it-seems/
Facebook	600 Tb	https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/
Twitter	100 Tb	http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-02-Lin.pdf
Spotify	2.2 Tb (compressed; becomes 64 Tb in Hadoop)	http://www.slideshare.net/AdamKawa/hadoop-operations-powered-by-hadoop-hadoop-summit-2014-amsterdam
Sanger Institute	1.7 Tb (DNA sequencing data only)	http://www.slideshare.net/insideHPC/cutts

Real big data in our every day lives (behind the scenes)

Organisation	Est. amount of data stored	Source
Google	15,000 pb (=15 exabytes)	https://what-if.xkcd.com/63/
NSA	10,000 pb (possibly overestimated, see source)	http://www.forbes.com/sites/netapp/2013/07/26/nsa-utah-datacenter/
Baidu	2,000 pb	http://on-demand.gputechconf.com/gtc/2014/presentations/S4651-deep-learning-meets-heterogeneous-computing.pdf
Facebook	300 pb	https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/
Ebay	90 pb	http://www.itnews.com.au/News/342615,inside-ebay8217s-90pb-data-warehouse.aspx
Sanger (sequencing equipment)	22 pb (for DNA sequencing data only; ~45 pb for everything per Ewan Birney May 2014)	http://insidehpc.com/2013/10/07/sanger-institute-deploys-22-petabytes-lustre-powered-ddn-storage/
Spotify	10 pb	http://www.slideshare.net/AdamKawa/hadoop-operations-powered-by-hadoop-hadoop-summit-2014-amsterdam

Big Data (and outside the realm of our daily lives)



1 TB seems like a lot of data (and for daily lives, is more than enough). In biology,

- There are more than 2.7 million samples are now available from the Gene Expression Omnibus database (Last check: **19 Nov 2018**).
- Assuming each file is about 1 GB (very modest estimation), this is already easily in the range of 2.7 Petabytes.
- Biology is entering the digital era.

Why bother with big data?

While naturally subject to signal-to-noise challenges, big data may potentially compensate for the noisiness of each individual data set precisely because of its scale.

Intuitively, signals that occur independently in multiple data sets are more likely to be “real”; for example, genes identified as cell-cycle regulated in multiple genome-scale studies are more likely to be truly cell-cycle regulated.

But this is provided that a common signal “exists” and is “detectable”.

Simply identifying repeating signals can also zero in on **common technical and biological artefacts** or **very broad (and thus often less interesting) biological signals**, such as the general stress response that *S. cerevisiae* exhibit across essentially all treatments or broad growth regulators in human cell culture data.

How big data can be used in biology

Big Data also has the potential of revolutionising our use of model organisms, enabling accurate, less-biased, molecular-level identification of the most informative model for genes and diseases in the least expensive and most tractable experimental system.

The key advantage is the ability to go beyond sequence-based orthology to systematically assess functional conservation, promising a functional mapping of proteins, pathways, and phenotypes across organisms.

For example, biologists can use a method based on probabilistically mapping protein networks from a large compendium of high-throughput expression data across organisms to systematically predict which genes are most likely to participate in the same biological process and thus have analogous function in different organisms.

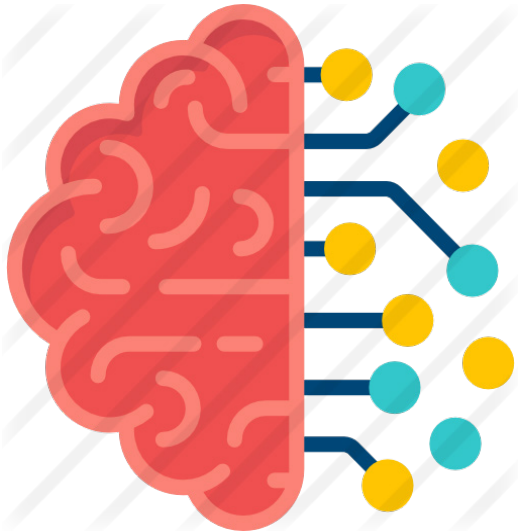
In other words, massive data integration.

User-friendly Systems for Big Data Analysis in Biology

galaxyproject.org	Platform for genome-scale biomedical research
imp.princeton.edu	Functional networks in model organisms and humans
giant.princeton.edu	Tissue-specific networks and genome-wide association studies in humans
thebiogrid.org	Database of protein and genetic interactions
seek.princeton.edu	Cross-platform search engine for expression data
genomespace.org	Framework for integrative genomics analysis
cbiportal.org	Visualisation and analysis of cancer genomic data

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501356/>

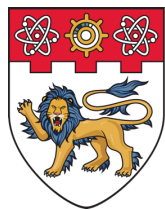
Biologists are not Replaceable by Machines



The wealth that Big Data brings will enable cell biologists to better design and focus their experimental programs with the expectation that biological insights will come faster and more efficiently.

We are not even close to replacing individual experiments (and the cell biologists who do them!) with computers, but instead are in the midst of an exciting time when we are just beginning to tap the major effect of Big Data on the world of cell biology.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501356/>



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

The Future of Biological Data

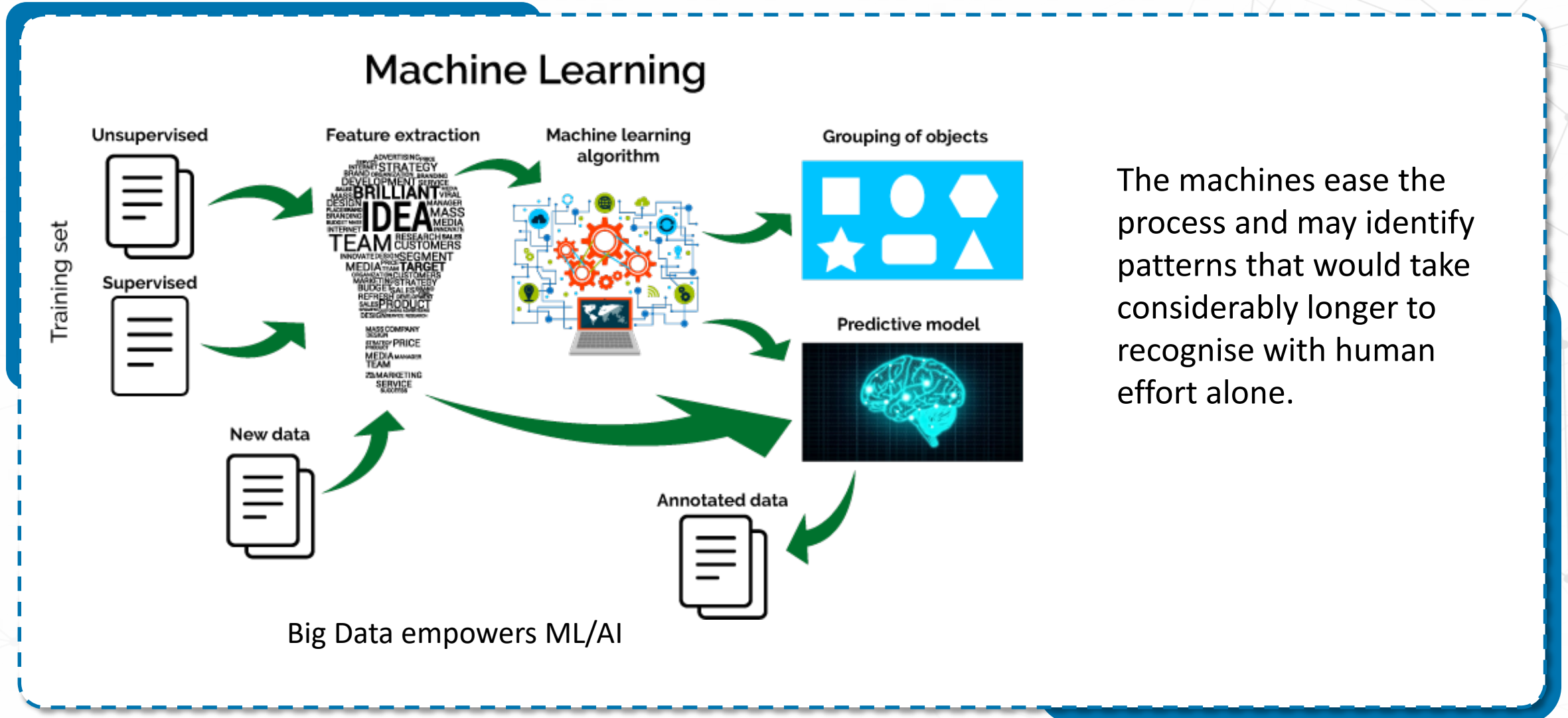
BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



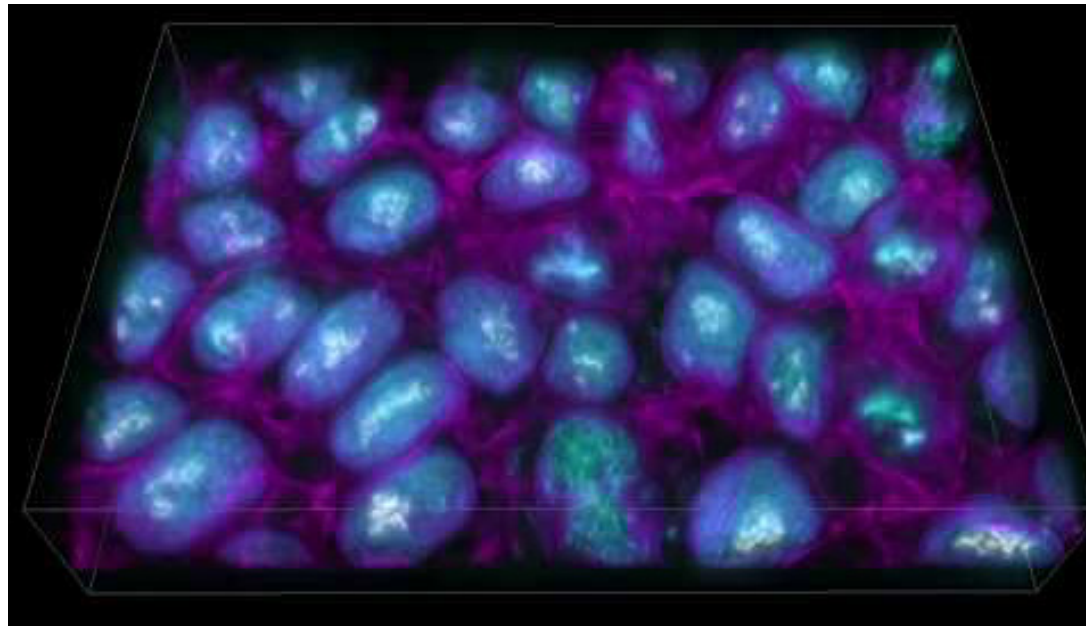
The Future of Biotechnology Lies in Machine Learning and AI



The machines ease the process and may identify patterns that would take considerably longer to recognise with human effort alone.

Allen Cell Explorer

Scientists at the Allen Institute have used machine learning to train computers to see parts of the cell the human eye cannot easily distinguish. Using 3D images of fluorescently labeled cells, the research team taught computers to find structures inside living cells without fluorescent labels, using only black and white images generated by an inexpensive technique known as brightfield microscopy.



<https://www.allencell.org/>

Benevolent AI

- A learning algorithm that processes natural language and formulate new ideas from what it reads, sifts through vast chemical libraries, medical databases and conventionally presented scientific papers, looking for potential drug molecules (with particular focus on Motor Neuron Disease, Parkinson's Disease, Glioblastoma and Sarcopenia).
- April 2018 – Raised USD\$150 Million (with most backers from US e.g. Woodford Investment Management despite being UK-based).



<https://benevolent.ai/>

XtalPi

- XtalPi is a pharmaceutical technology company that is reinventing the industry's approach towards drug research and development with its Intelligent Digital Drug Discovery and Development (ID4) platform which integrates quantum mechanics, AI, and cloud computing, allowing pharmaceutical companies to increase efficiency, accuracy, and success rates at critical stages of drug R&D.
- USD \$66 million investment from Google, Tencent, Sequoia China.

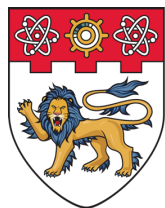


<http://www.xtalpi.com/>

Big Pharma and IBM Watson

- In late 2016, pharmaceutical giant Pfizer announced a collaboration with IBM, involving the use of the latter's Watson AI for immuno-oncological research.
- In June 2017, Novartis also announced a collaboration with IBM Watson to use AI for improving health outcomes in breast cancer patients (Clinical Trial Matching).





**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Summary


BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Key Takeaways from this Topic

- 
1. The distinction between small to big data lies in its actual size in bytes, the number of samples, and/or the numbers of variables covered.
 2. Small data has taken on new meaning --- it may refer to a small dataset, or to the extract “gold” or insight from big data.
 3. Current –omics analysis from most experiments are not as big as they claim to be. They fall in the realm of moderate data.
 4. Comparative analysis involves comparison of samples between different classes benchmarked on a common set of variables.
 5. Although we live in the TB age, big data in the PB and even the EB range govern many aspects of our lives.
 6. Biology is becoming increasingly digitised. As we enter the age of AI, it is important to understand how these new technologies may help us derive novel insight and therapies given heaps of stored data.

Recommended Readings

[Sequence Alignment] Lesk A.
2002. Introduction to
Bioinformatics. Oxford
University Press, Inc. New York,
NY, USA.

[Comparative Analysis] Sanford
et al. 2002. The Comparative
Method Revisited. BioScience,
52(9):830–836.

[Case study in Gene Expression]
Bertucci F et al. 2016.
Comparative genomic analysis
of primary tumors and
metastases in breast cancer.
Oncotarget, 7(19):27208-19.

[Implications of big data in
Biology] Dolinskia K and
Troyanskayaa OG. Implications
of Big Data for cell biology.
2015. Mol Biol Cell, 26(14):
2575–2578.