Data Science in Biology

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Learning Objectives

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

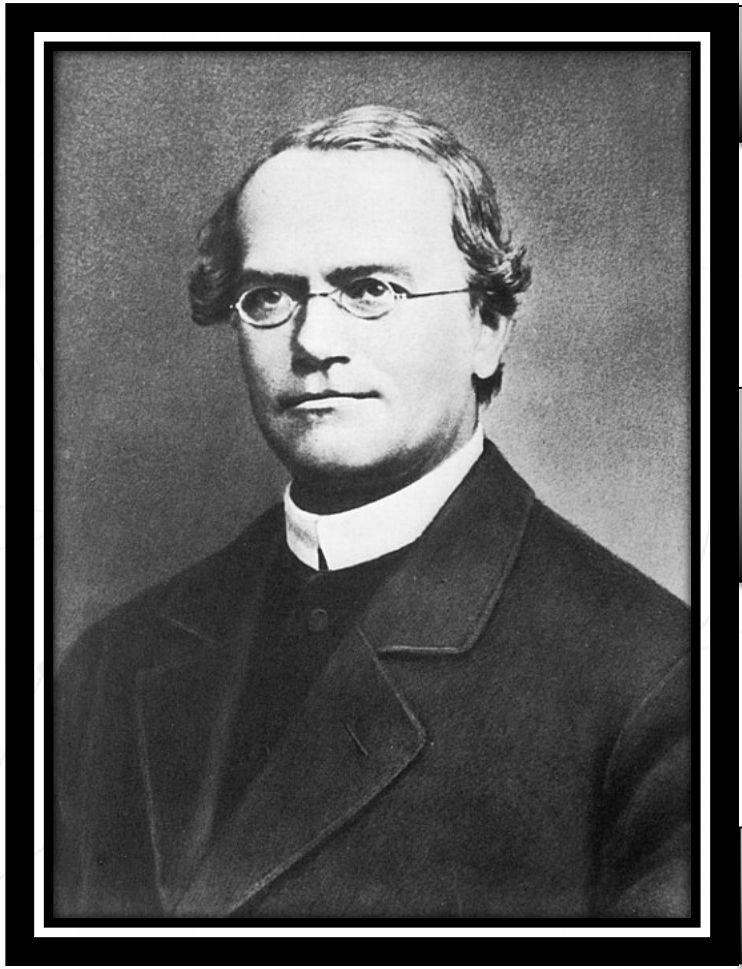
- Describe the historical context and evolution of quantitative biology from bioinformatics to data science.
- Describe the different levels of data analytics.
- Describe the three components of data science.
- Explain the steps involved in data science investigation.
- Describe the specific applications of data science in biology.
- Explain the risks involved in data analytics.

Historical Context

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

1800s: Earliest Instance of Biological “Big Data”



Gregor Mendel
1822 - 1884

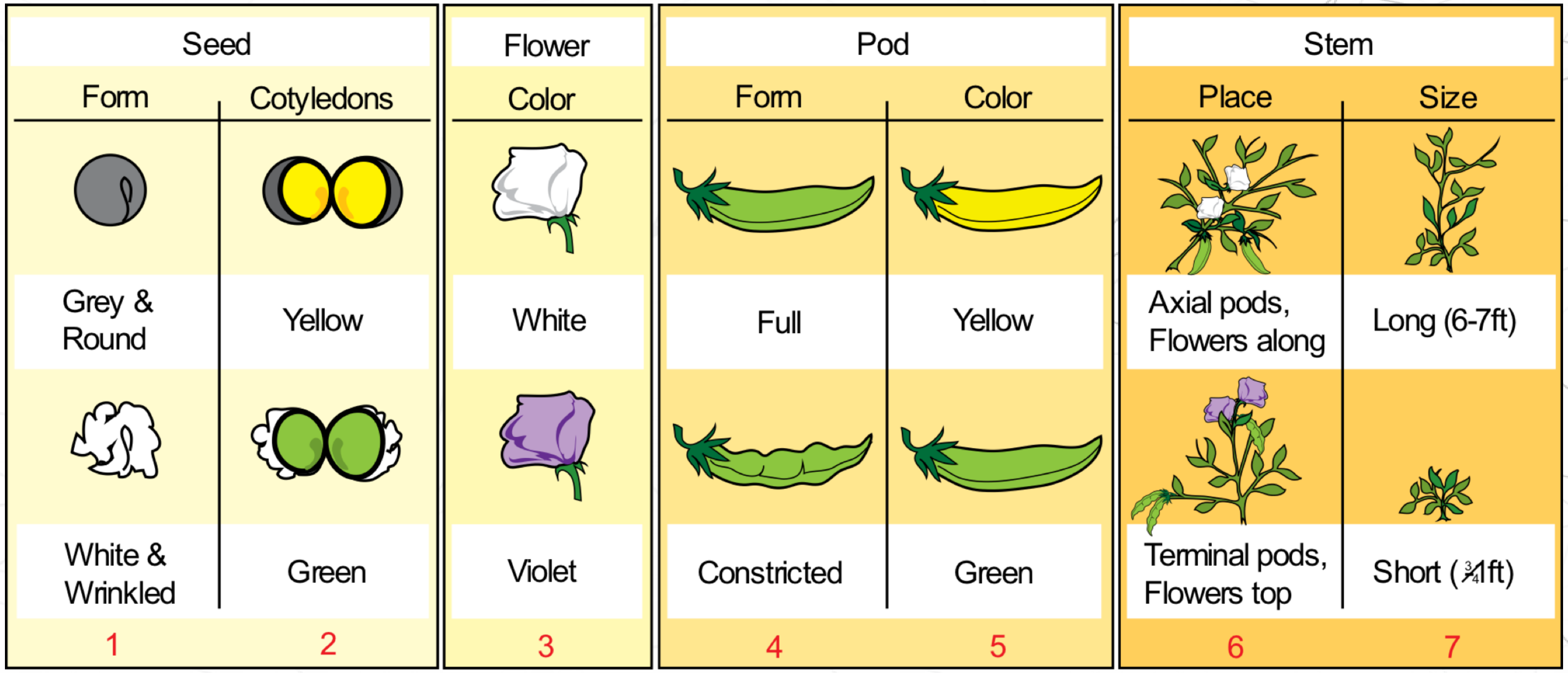
Established the power of ”quantitative biology” (precursor of “biological data science”)

7 pea traits, or characters, studied by Mendel

Source: By Unknown - http://0.tqn.com/d/biology/1/0/l/e/3244238.jpg, Public Domain, https://commons.wikimedia.org/w/index.php?curid=33347279

1800s: Earliest Instance of Biological "Big Data"

7 pea traits, or characters, studied by Mendel



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

Why a data-centric approach became essential?

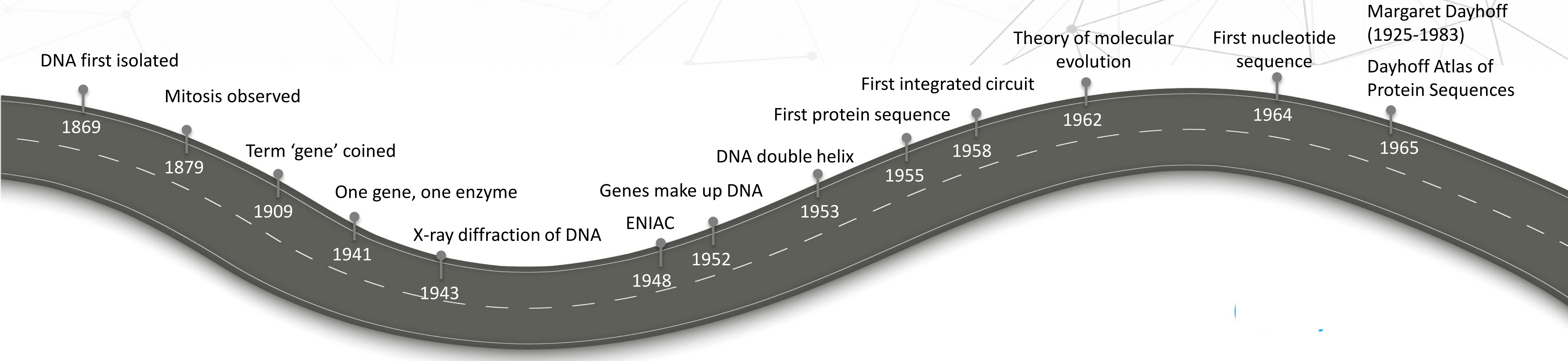
An **expanding collection of sequences** provided both a source of data and a set of interesting problems that were infeasible to solve without the number-crunching power of computers.

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

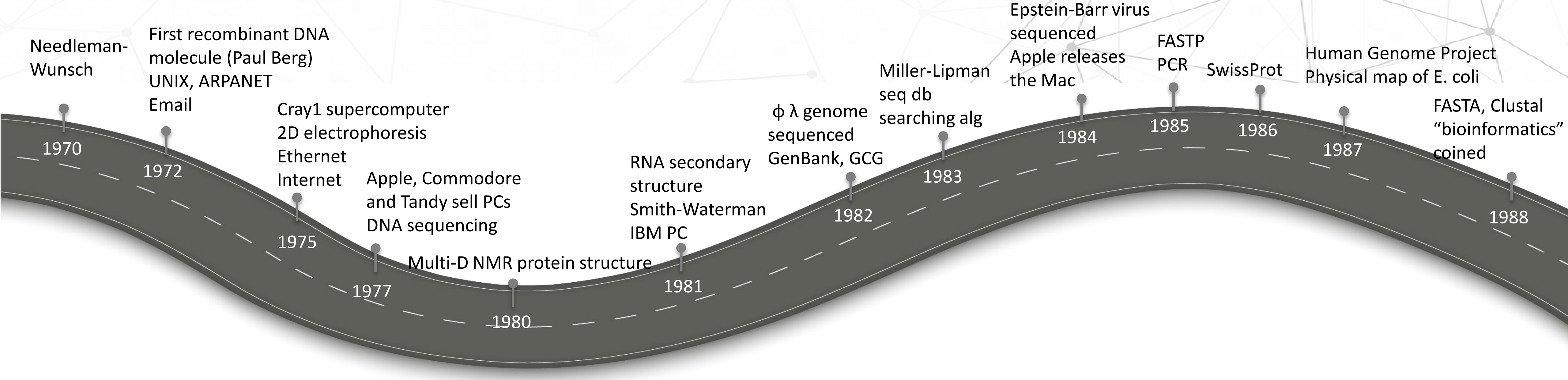
High-speed digital computers, which had developed from weapons research programmes during the Second World War, finally became widely available to academic biologists.

Source: Hagen, Nature Reviews Genetics 1, 231–236 (2000)

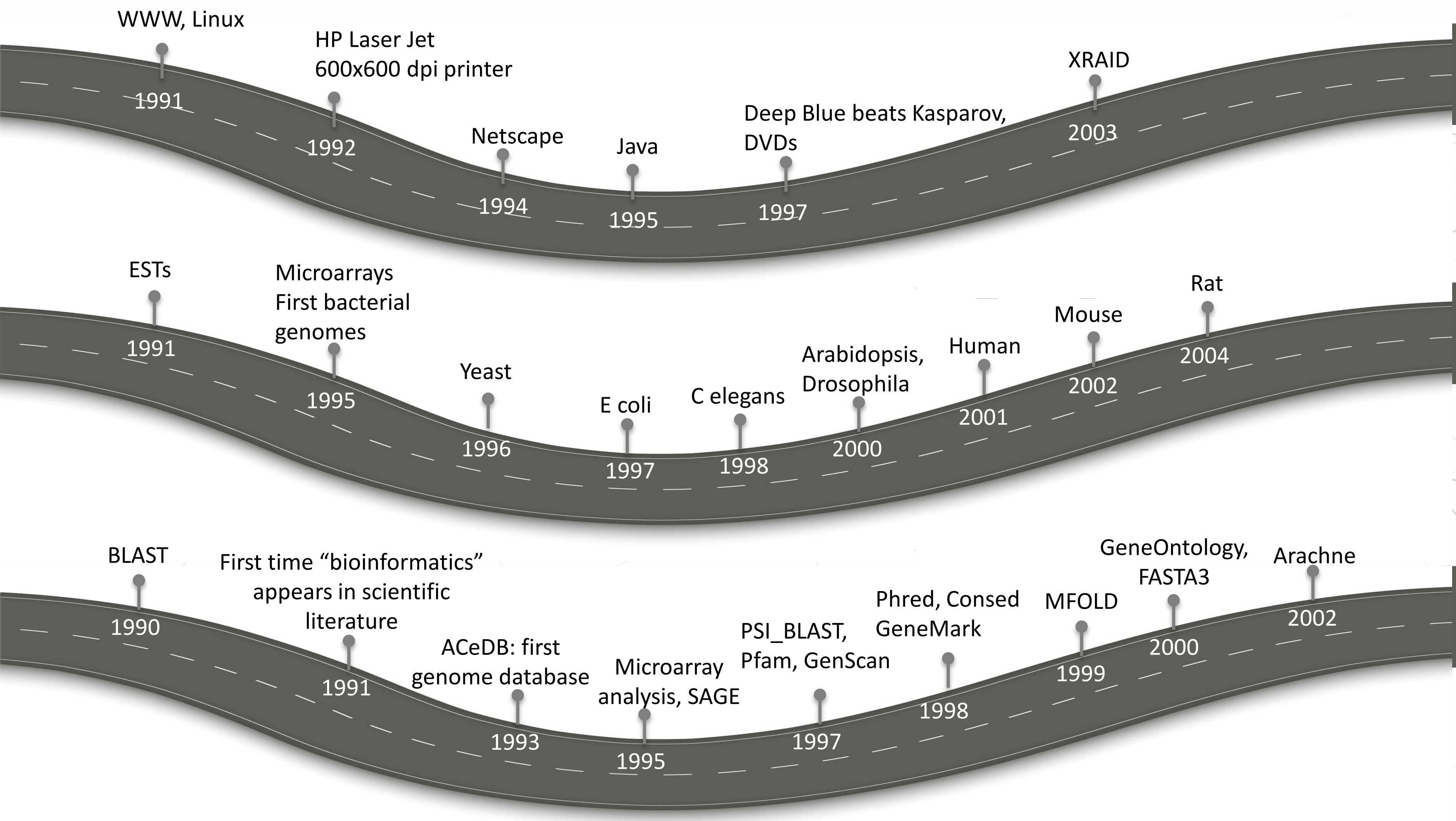
Rise of Big Data and Data Science



Rise of Big Data and Data Science



Rise of Big Data and Data Science

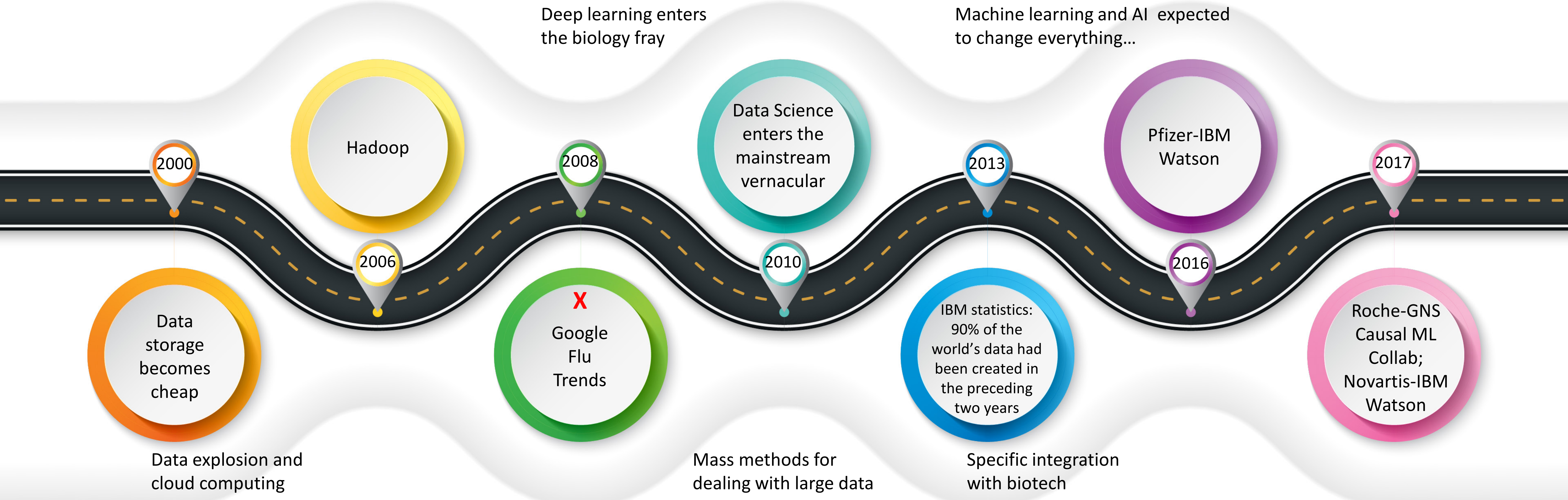


AI and data storage technologies become more powerful.

The rise of biological big data.

Bioinformatics becomes a discipline.

Age of Big Data and Data Science



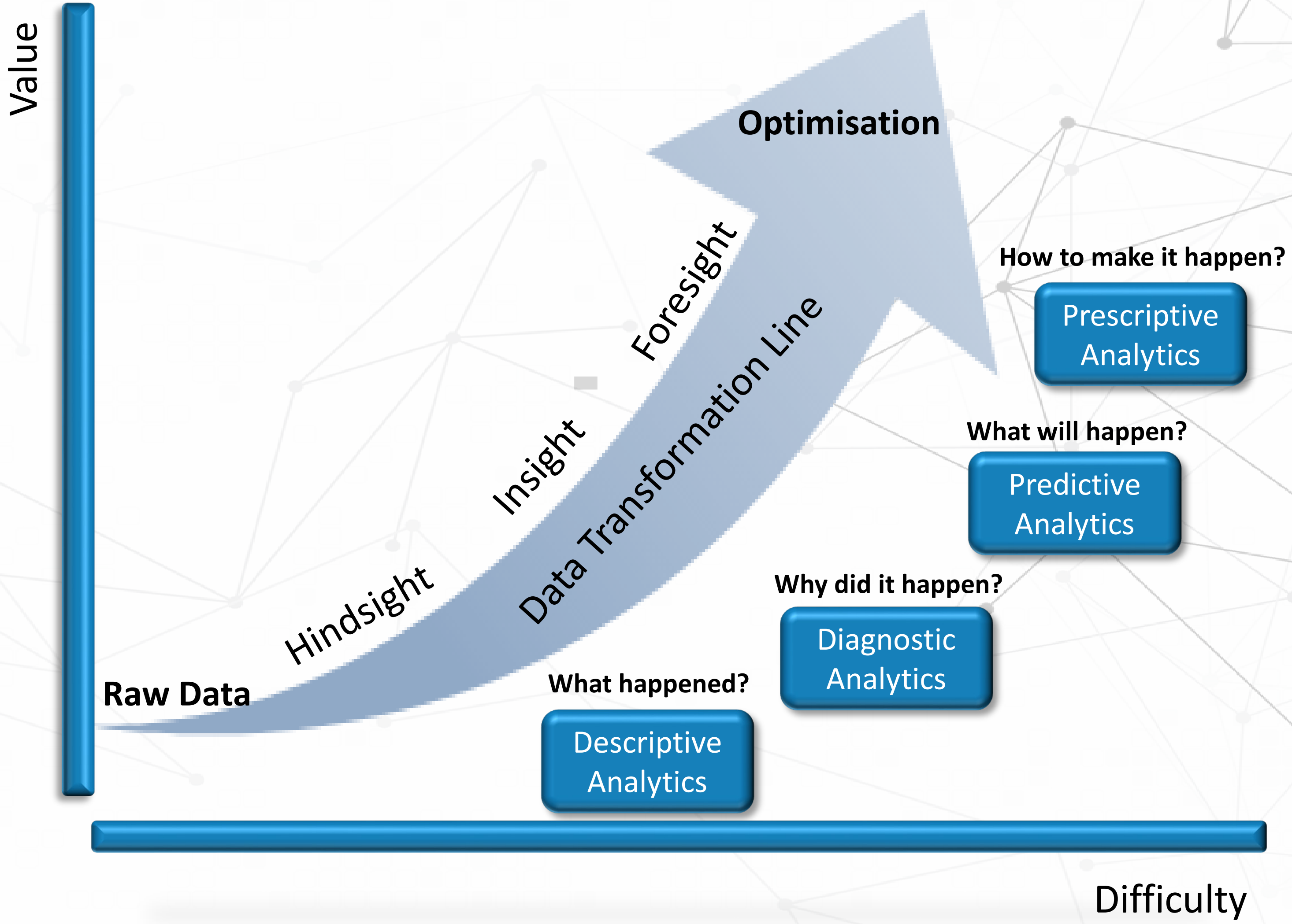
Cheap Disks --> Big Data --> Cloud Computing --> Mass Analytic Tools --> Data Scientists --> Data Science Teams --> New Analytic Insights

Levels of Data Analytics

BS3033 Data Science for Biologists

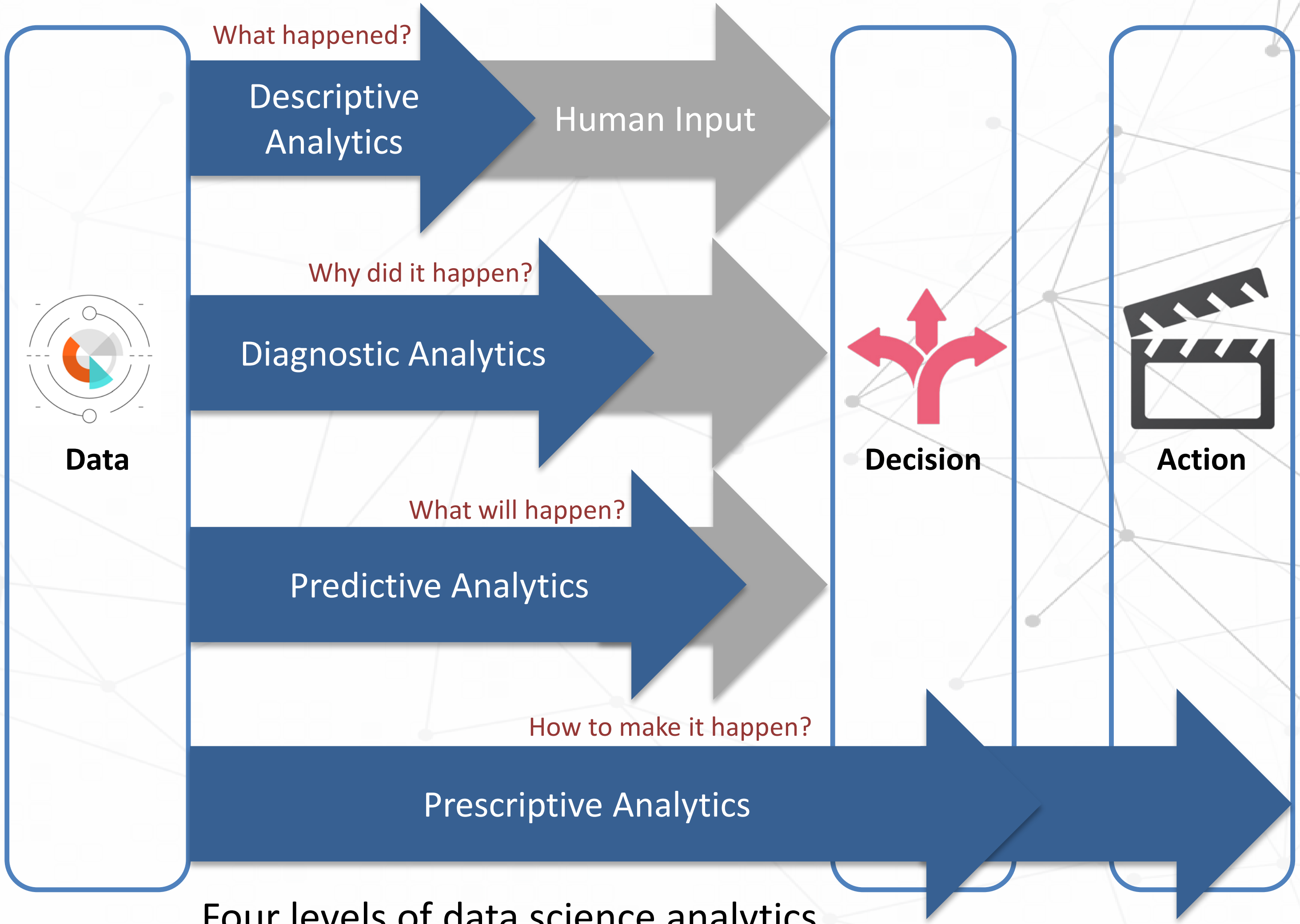
Dr Wilson Goh
School of Biological Sciences

Value and Difficulty



Gartner analytics value-difficulty chart

From Data Science to Action



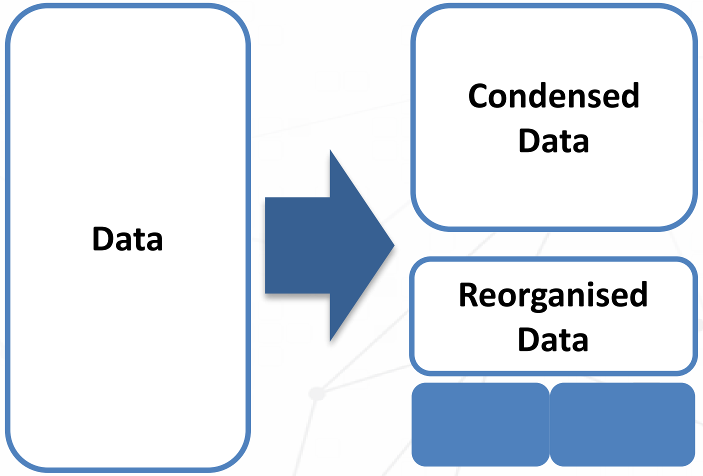
Four levels of data science analytics

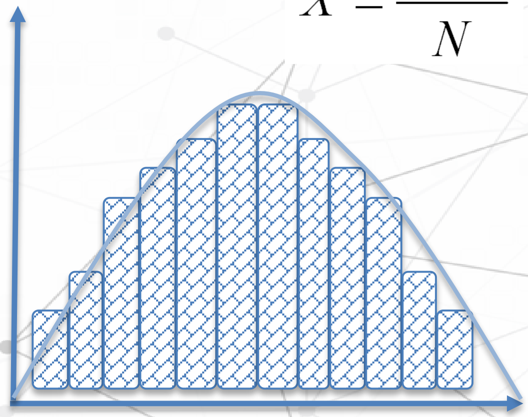
Descriptive Analytics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Descriptive Analytics





$$\overline{X} = \frac{\sum X}{N}$$

Summary Statistics

- It is the simplest form of analytics.
- It involves reorganisation and condensation of data.
- It uses summary statistics to "summarise" the data.

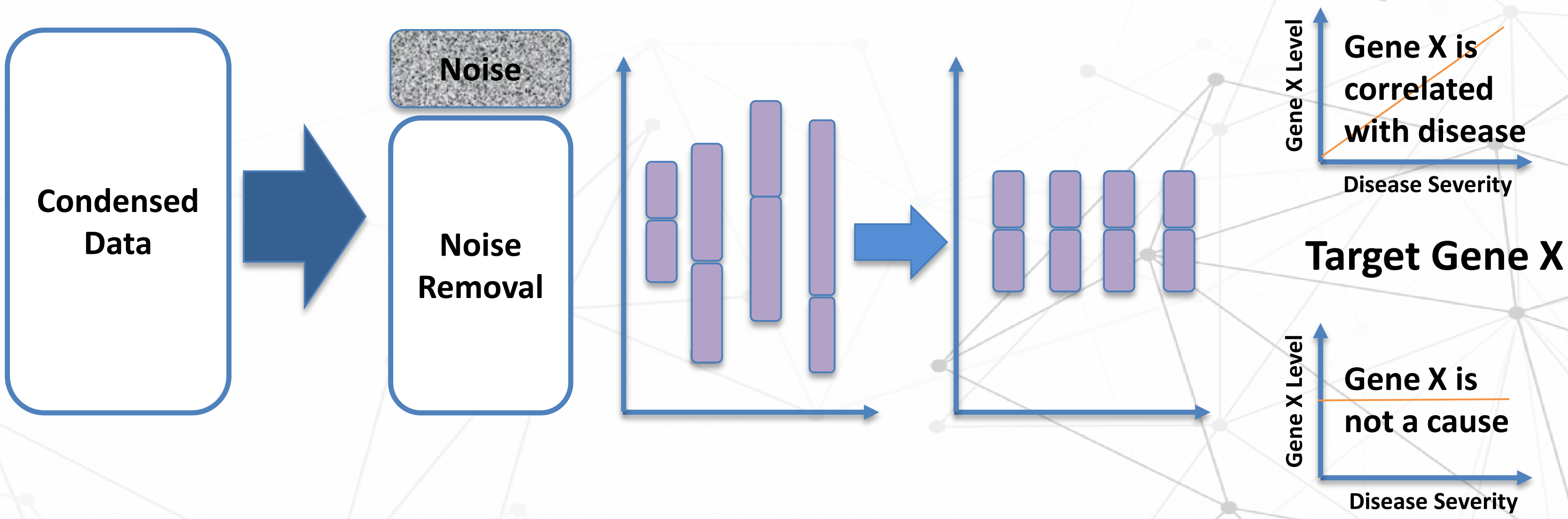
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Diagnostic Analytics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Diagnostic Analytics



- It is built on top of descriptive analytics.
- It may involve denoising, renormalisation and bias correction.
- It infers relationships in data and aims to identify key causes.

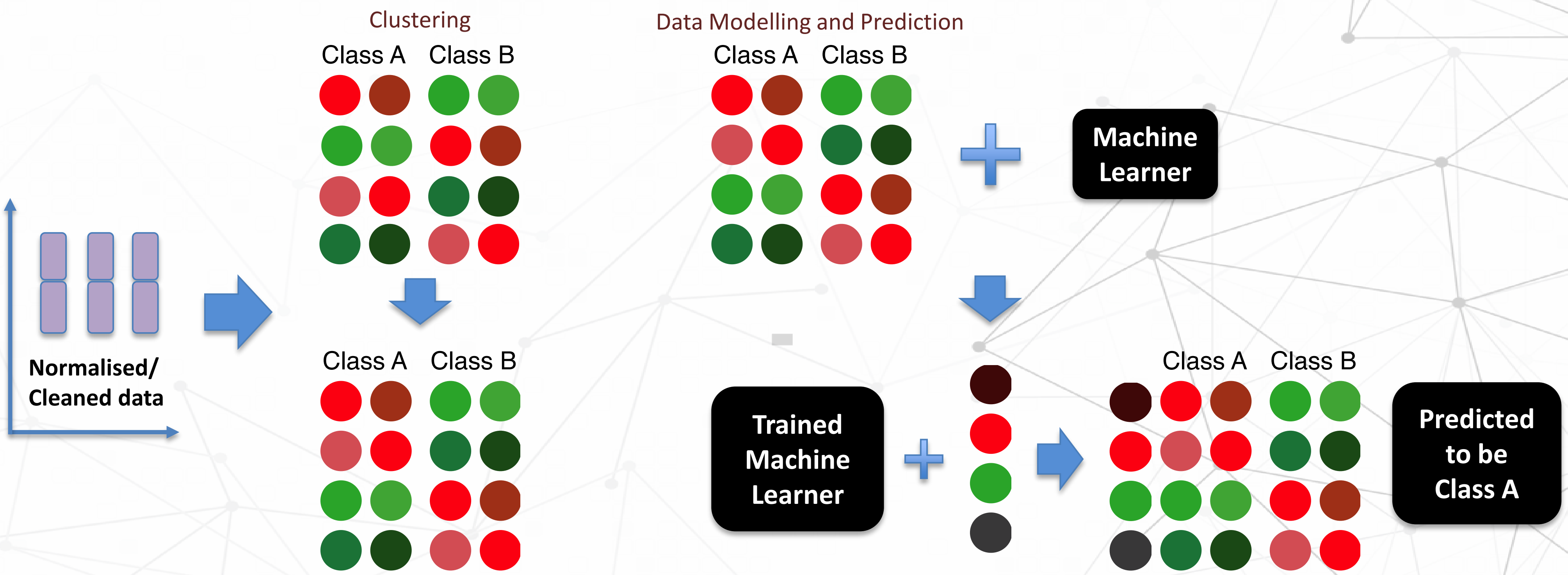
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Predictive Analytics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Predictive Analytics



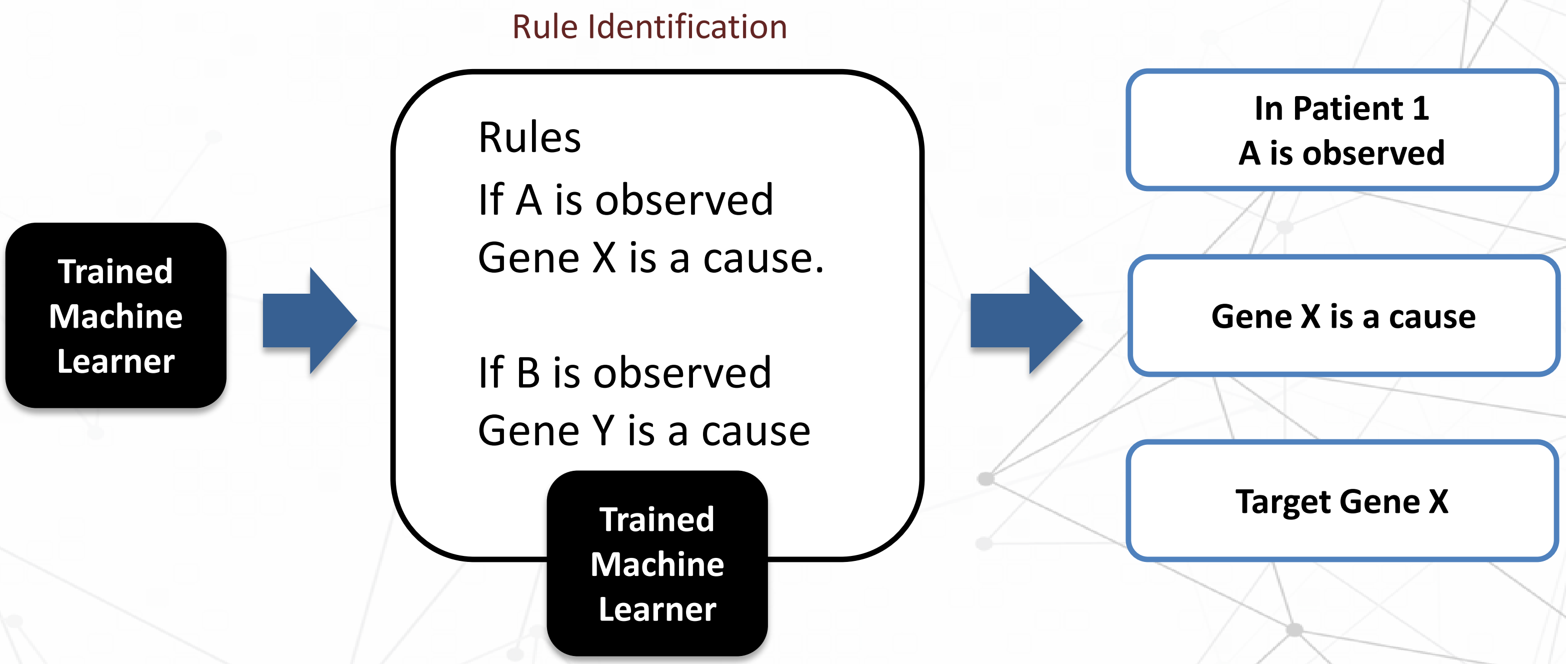
- It is built on top of descriptive and diagnostic analytics.
- It may involve the use of clustering and machine learning techniques (data modelling).
- The goal is to predict the identify of an unknown entity, or determine when a phenomenon will happen (for example, cancer relapse).

Prescriptive Analytics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Prescriptive Analytics



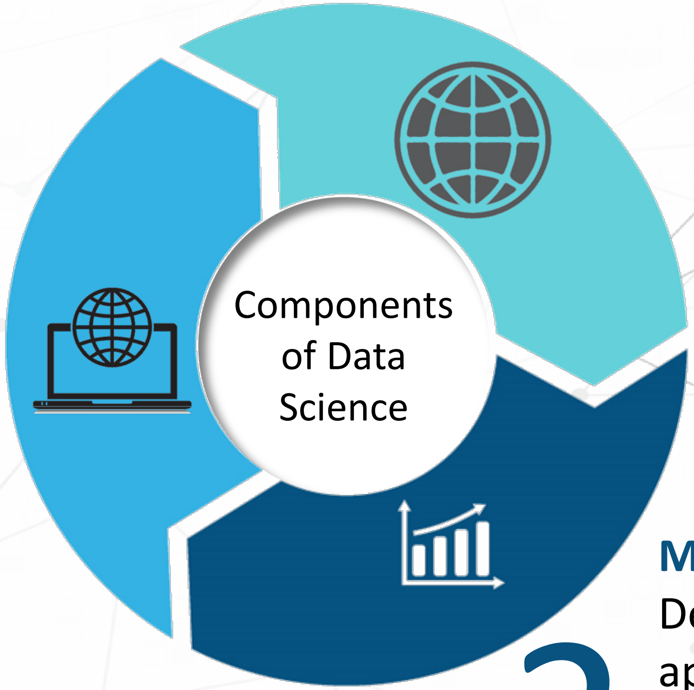
- It is built on top of descriptive, diagnostic and predictive analytics.
- It involves advanced machine learning and artificial intelligence techniques (cause-effect modelling).
- The goal is to influence the occurrence of a phenomenon (If I do this, this will/will not happen).
- The rule of identification is usually not straightforward.

Components of Data Science

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Components of Data Science



1 **Domain**
Business, Biology, Sociology

2 **Math/ Stats**
Deriving appropriate metrics for summarising, normalising data

- **Theoretical Statistics**
- **Applied Statistics**
- **Research Design**
- **Network Modeling**

3 **Computing**
Implementing algorithms, repetitive analysis, storing data

- **Programming**
- **Algorithms**
- **Machine learning and AI**
- **Computer Graphics**
- **Databases**

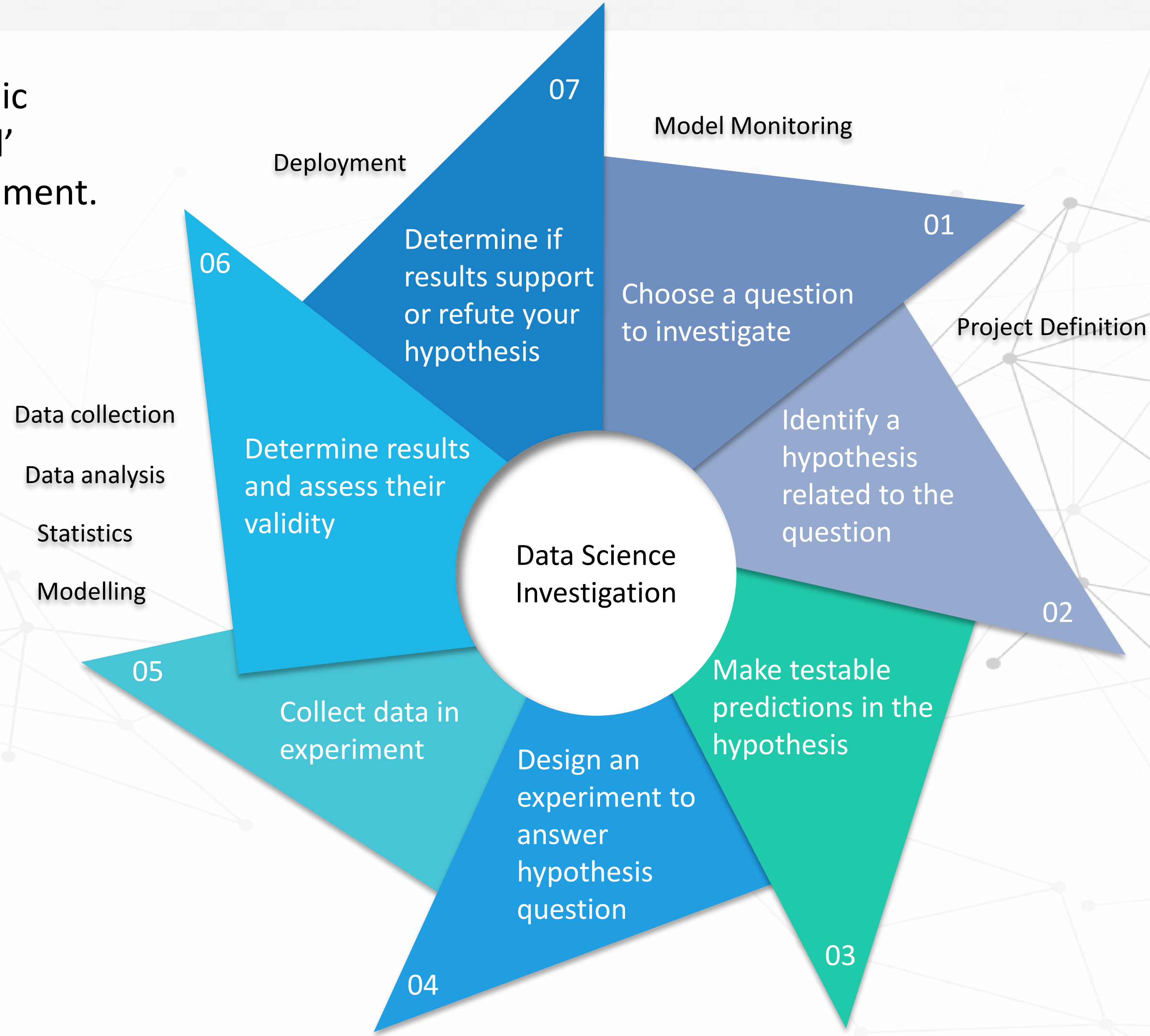
Steps of Data Science Investigation

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

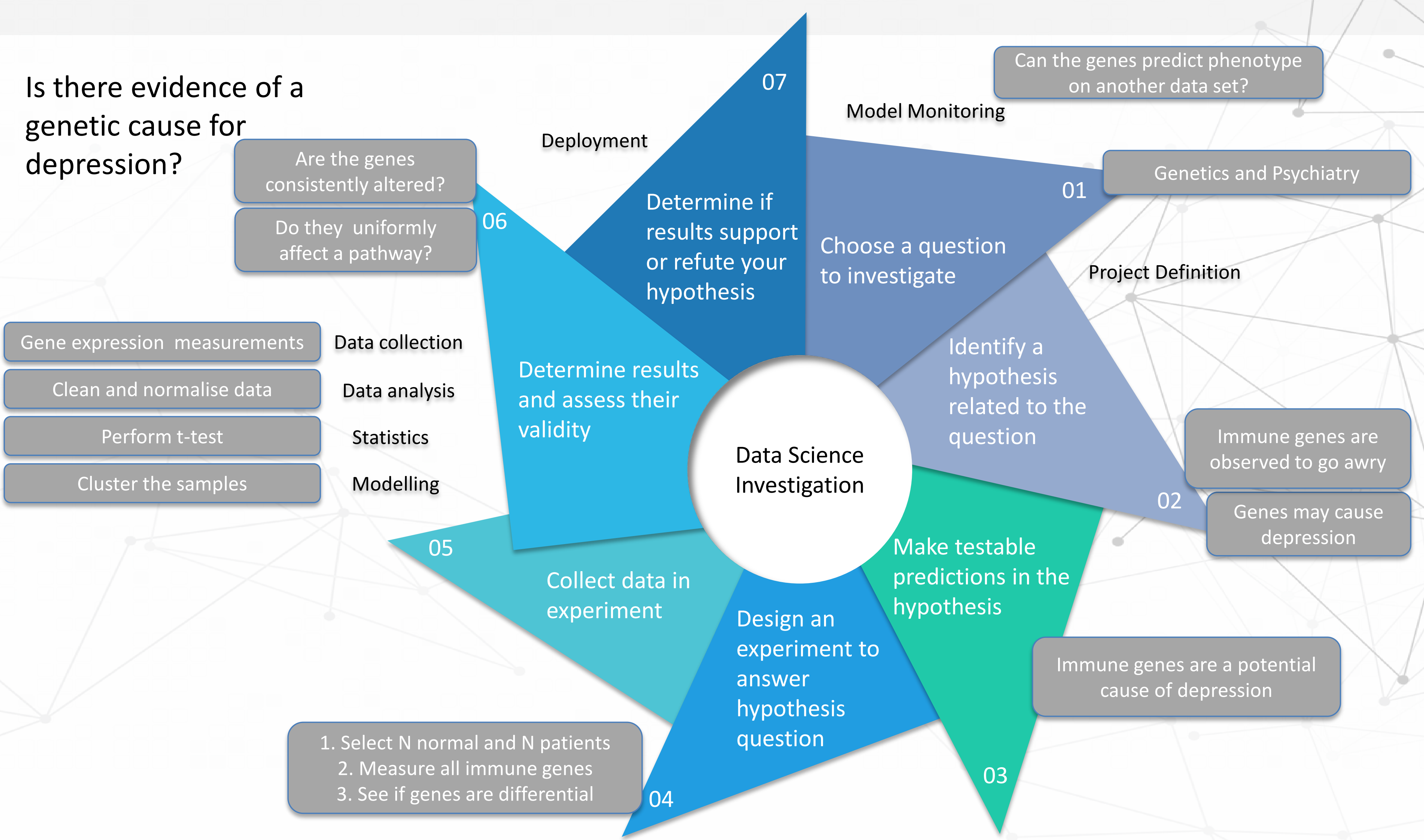
Data Science for Scientific Investigation

It follows the same basic procedure as a 'normal' wetlab scientific experiment.



Data Science for Scientific Investigation

Is there evidence of a genetic cause for depression?



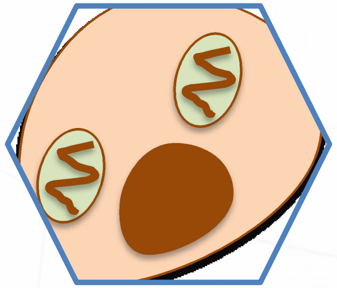
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

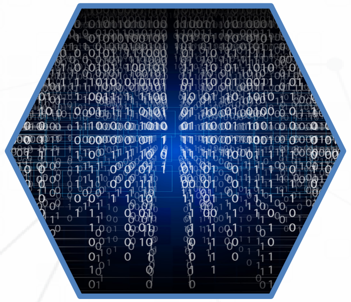
Biological Data Science

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Biology as a Data-driven Science

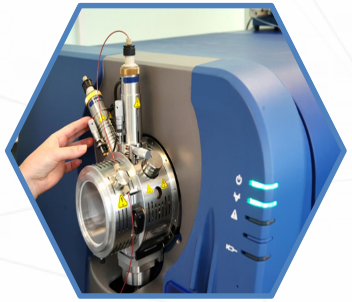




DNA Sequencing Instruments

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.

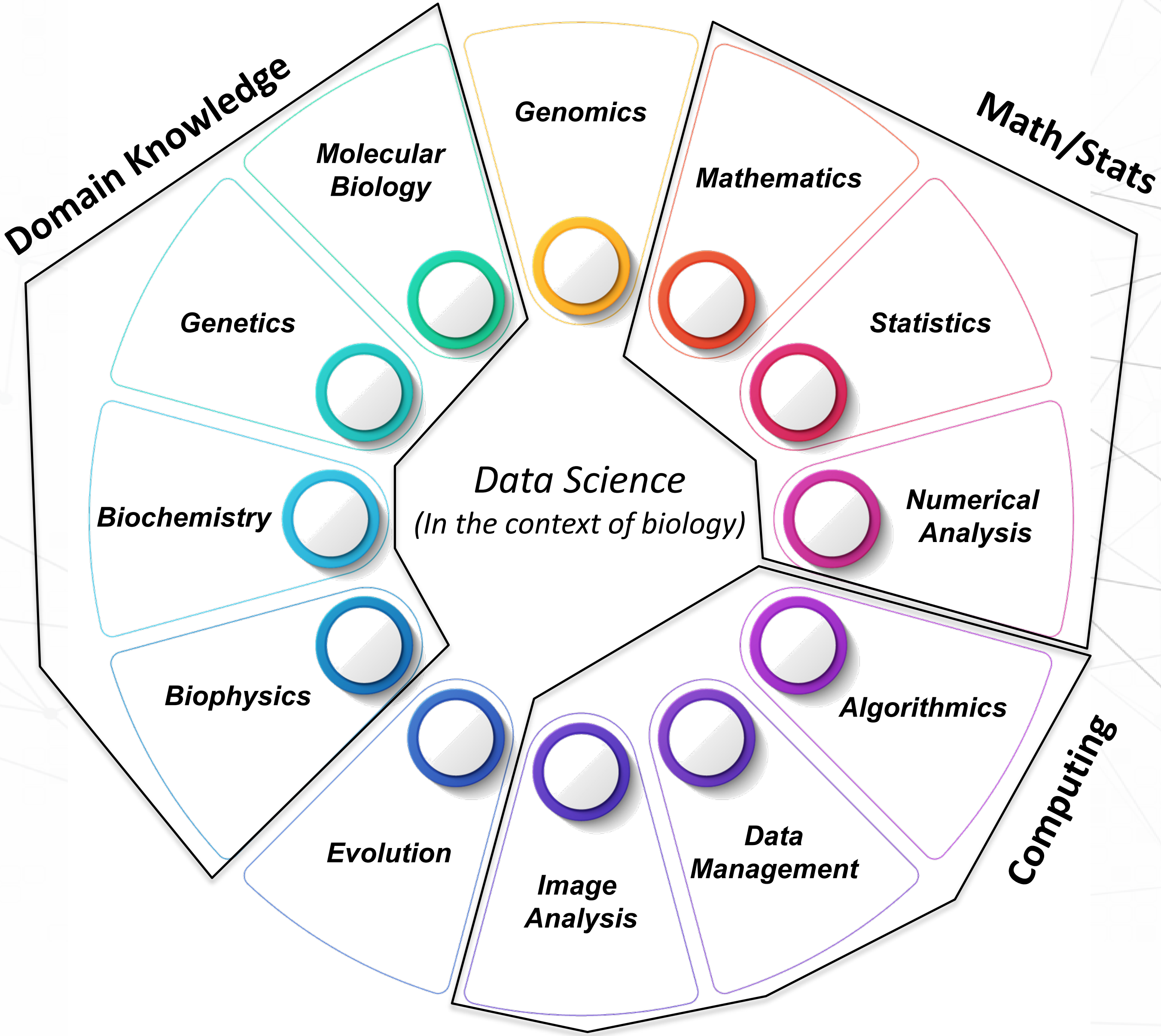
Data Science for Biology

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*").

Risks of Data Analytics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Risks of Data Analytics

Data Science is essentially a science of inference (prediction).

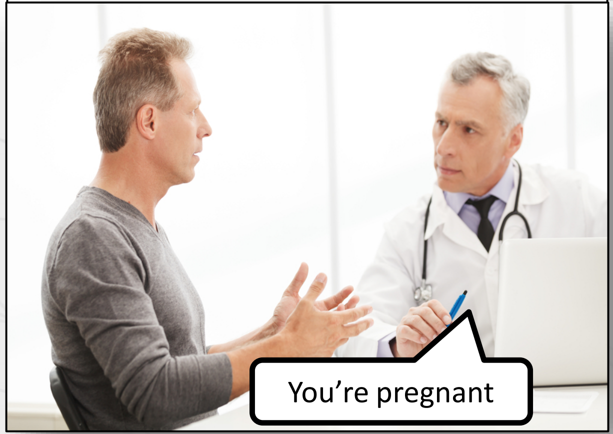
What you infer:
A young beautiful princess.

Reality:
An old wrinkled woman.

Risks of Data Analytics

Any analysis of massive data will unavoidably generate a certain rate of errors (***false positives*** and ***false negatives***).

Type I Error
(False Positive)



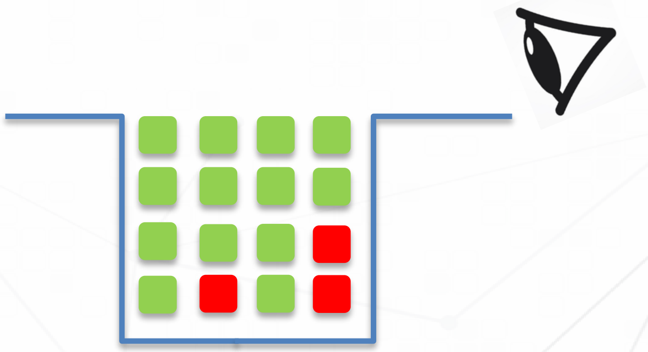
Type II Error
(False Negative)

Risks of Data Analytics

Risks

- Good research and development will include an evaluation of the error rates.
- Good methods should minimise the error rate where practical.
- However, there is always a trade-off between getting only correct answers (higher false negatives) and getting all the correct answers (higher false positives).

Risks of Data Analytics



Analogy

Imagine that you have a bag of cubes.

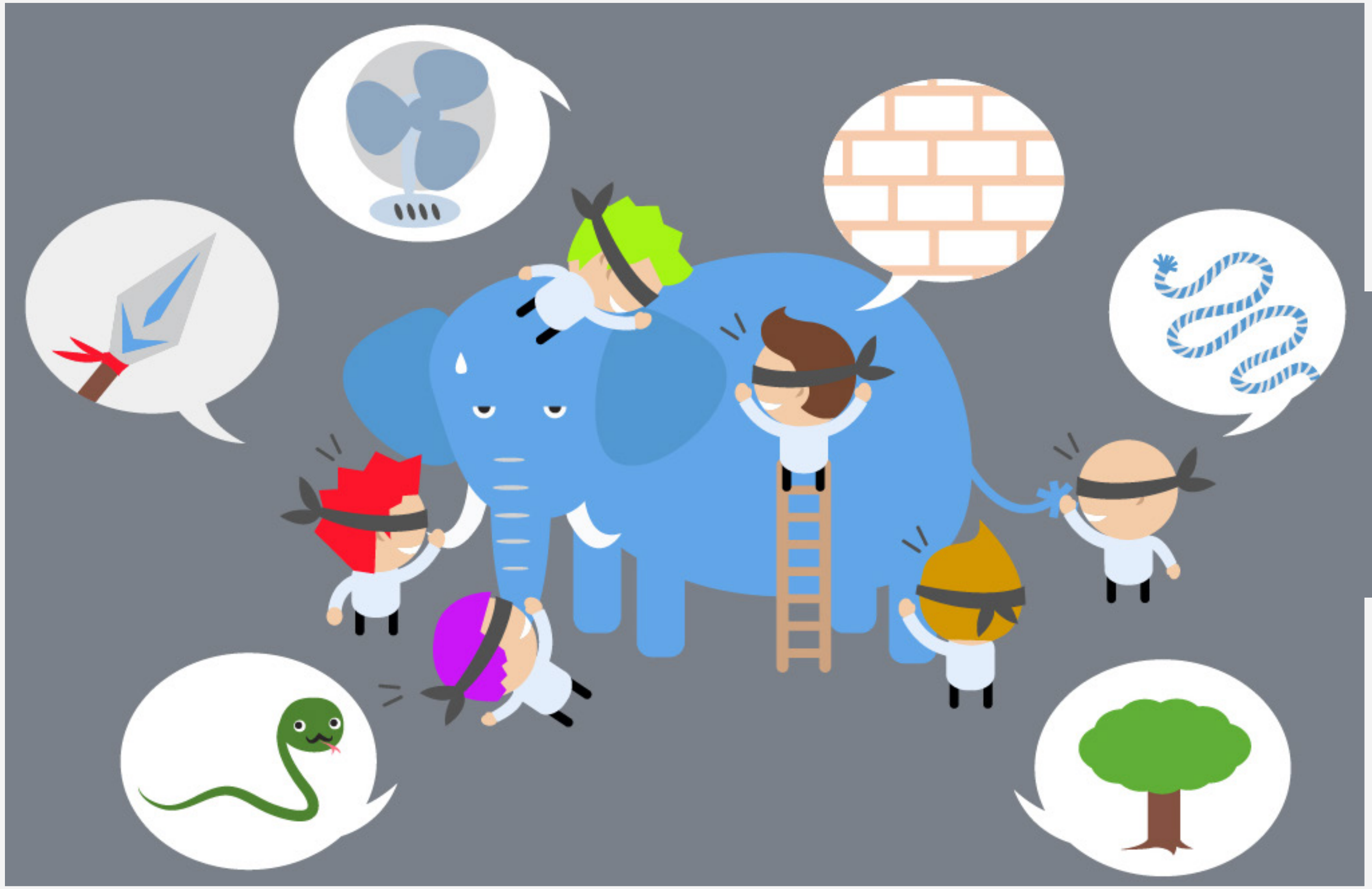
Most are green and a few are red.

Let us also assume that the cubes are arranged in rows such that at eye level, you can only see the green cubes.

If you want to guarantee that you only get green cubes, you take the top where you are confident (**no mistakes, but miss out some**).

However, if you need to get all the green cubes, you will have to tolerate getting some reds (**get all, but make some mistakes**).

Risks of Data Analytics



It is naïve to only want few but correct answers as you can get “blind-sided”. For building robust models for understanding a phenomenon, we need more data, even if it means tolerating some errors!

Data Science Gone Wrong

Data Science can go wrong badly but hopefully, we learn from mistakes. Let's take the example of **the spectacular failure of Google Flu Trends (GFT).**

Reasoning: No smoke without fire. People's Google search behavior reflects their situation, and needs.

Intuition: We can predict flu areas by flu keyword search.

Initial Success: GFT could produce accurate estimates of flu prevalence two weeks earlier than the CDC's data – turning the digital refuse of people's searches into potentially life-saving insights**.**

Subsequent Failure: GFT failed spectacularly and missed predicting the peak of the 2013 flu season**.**

So, what happened?: Overfitting and confounding. Irrelevant terms like "High school basketball" got picked up. Also people's search behaviour changed over time or can be influenced. For example, when younger people in Singapore watch news about bird flu in HK, they go online and search.

Success Stories

IBM Watson

- What it is: It is an AI meant for natural language processing.
- Achievements: Won a $1 million prize in Jeopardy.
- Uses:
  - Provides healthcare instructions for nurses at Sloan-Kettering cancer center.
  - Seeking immuno-oncology targets (with Pfizer)
  - Personalised consumer-interfacing (with GSK)

Amazon Predictive Dispatch

- What it is: Amazon's system for shipping us goods before we have even made a decision to buy it, purely based on prediction
- Uses:
  - Helps streamline logistics.
  - Amazon is now selling their predictive services and data to other global corporations.

Summary

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. Descriptive, Diagnostic, Predictive and Prescriptive Analytics are the four levels of data science analytics. The first three levels guide you in decision making and the fourth level guides you in taking action.

2. Any level of analytics involves three components – the domain knowledge, math and statistics, and computing.

3. Data science investigation follows the same basic procedure as a ‘normal’ wetlab scientific experiment.

4. Biological Data Science acknowledges that computer science, mathematics, physics, statistics, and other quantitative fields have developed advanced techniques that can be applied toward understanding biological data.

5. Any analysis of massive data will unavoidably generate a certain rate of errors. Good research and development will include an evaluation of the error rates and food methods will minimize the error rate. However, there is always a tradeoff between specificity and sensitivity.