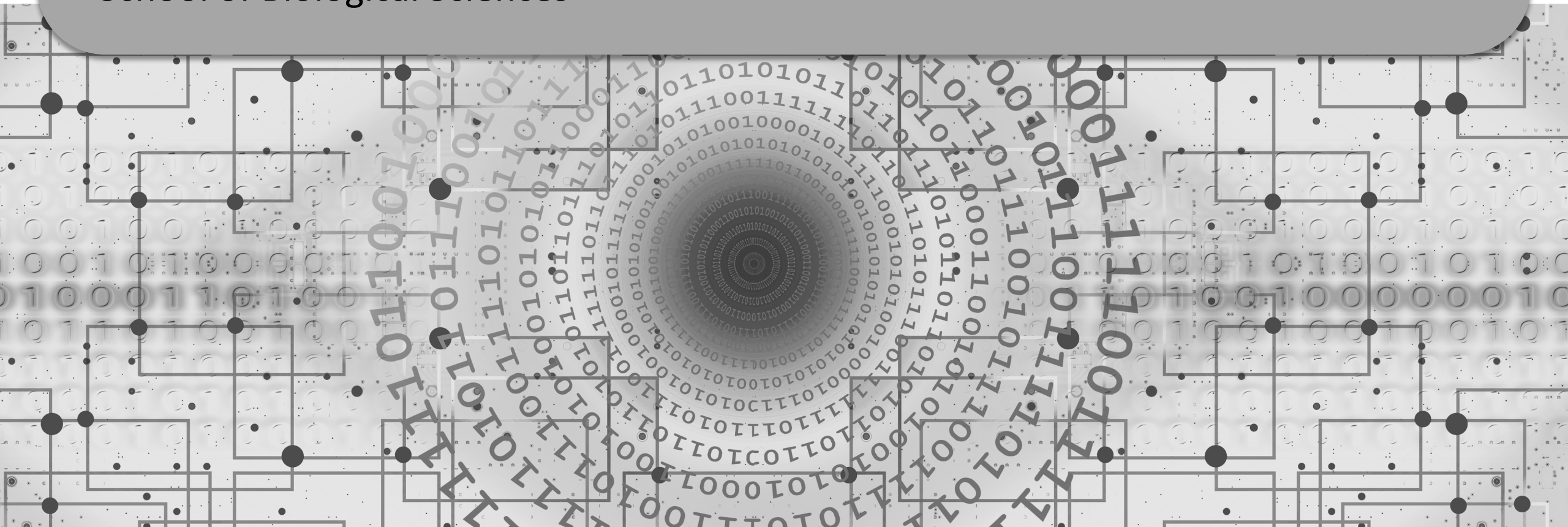
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

How to Become a Data Scientist – Essential Qualities and Skills

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 data science.
- Describe the different levels of data analytics.
- Describe the three components of data science.
- Explain the steps involved in data science investigation.
- Explain the risks involved in data analytics.

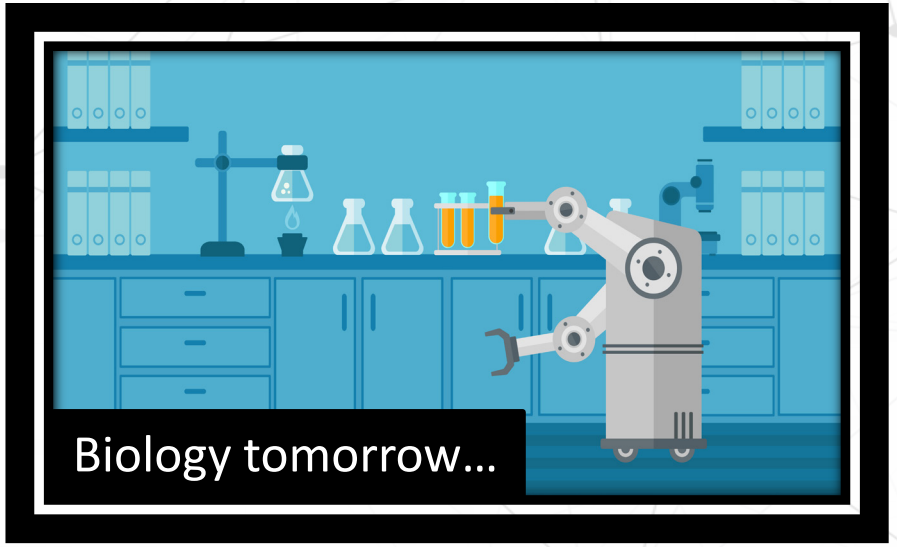
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

What is Data Science?

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Biology is becoming increasingly digitised…



We live in interesting times…

Biology outstrips Moore's Law…

Personalised Genomes

Between 1990 and 2003, unravelling the human genome – with more than 3 billion building blocks – cost approximately $2.7 billion, but in 2014 the costs for unravelling the same genome were barely $4,000.

Rapid Assembly of Organisms

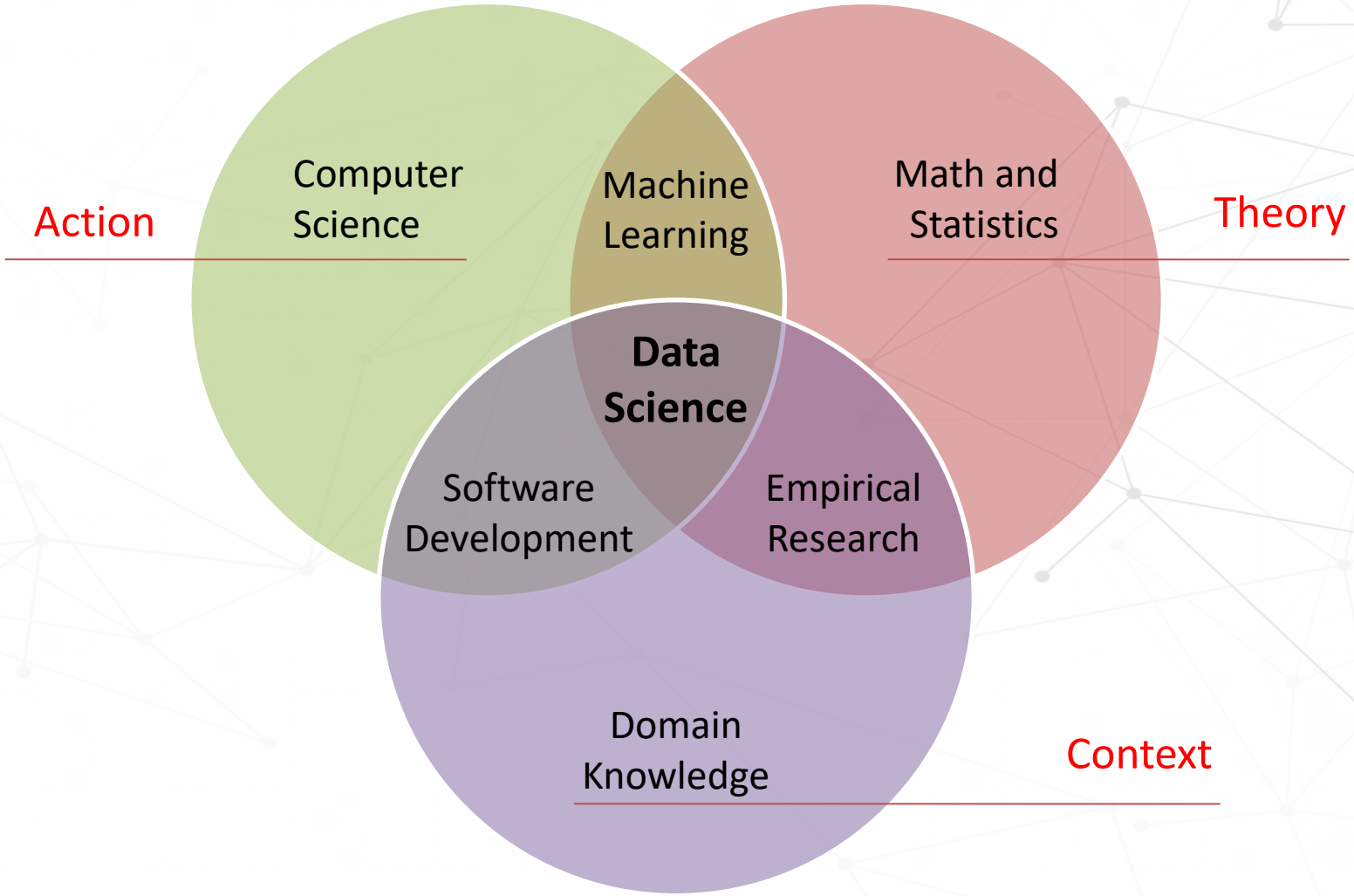
Mapping out the complex tomato genome took an international consortium five years to accomplish, but today we can read the genomes of 150 different tomatoes in one year.

What people think data science is?

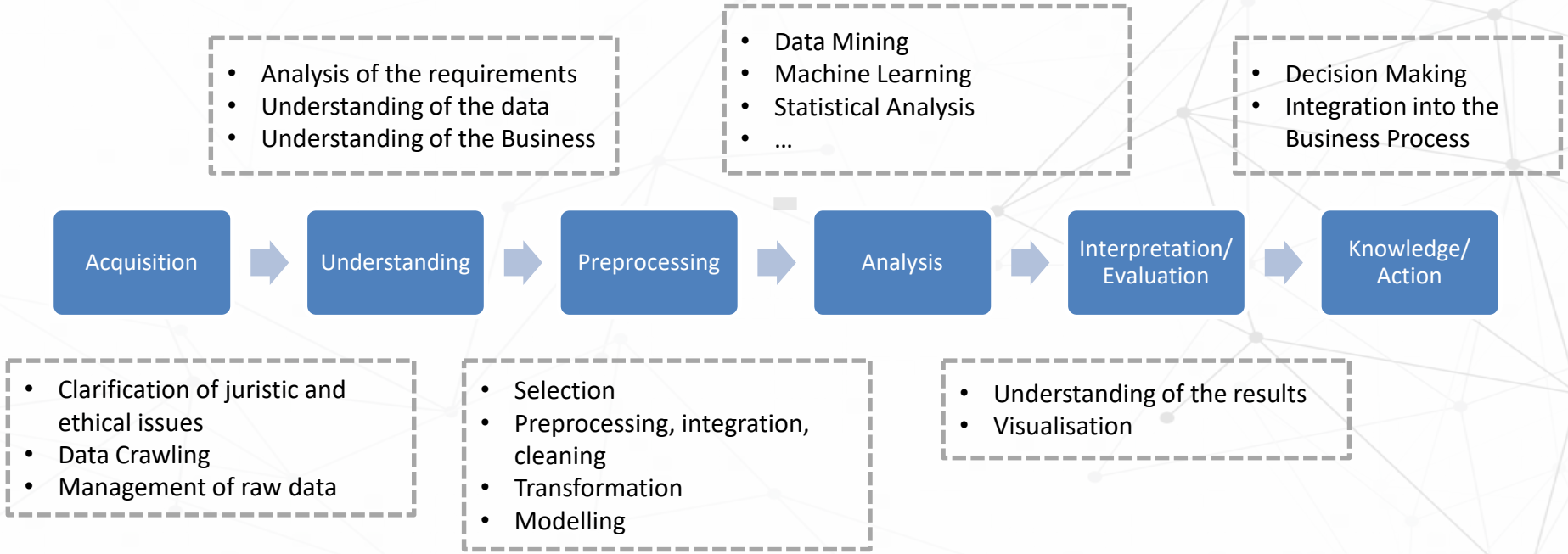
What is Data Science?

Data Science is the “art” of converting raw data to useful information that can be used to draw conclusions and make decisions.

What is Data Science?

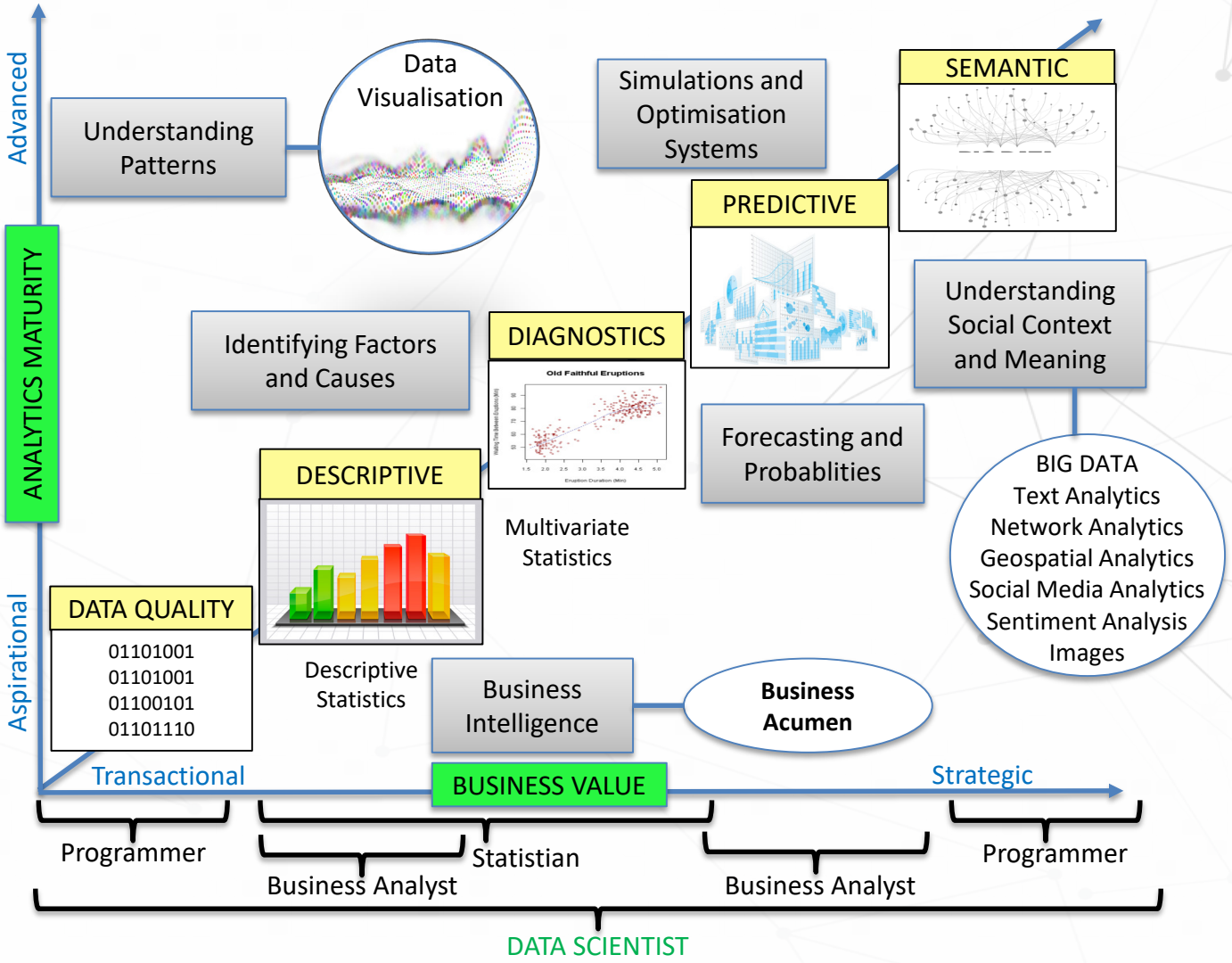


What is Data Science?

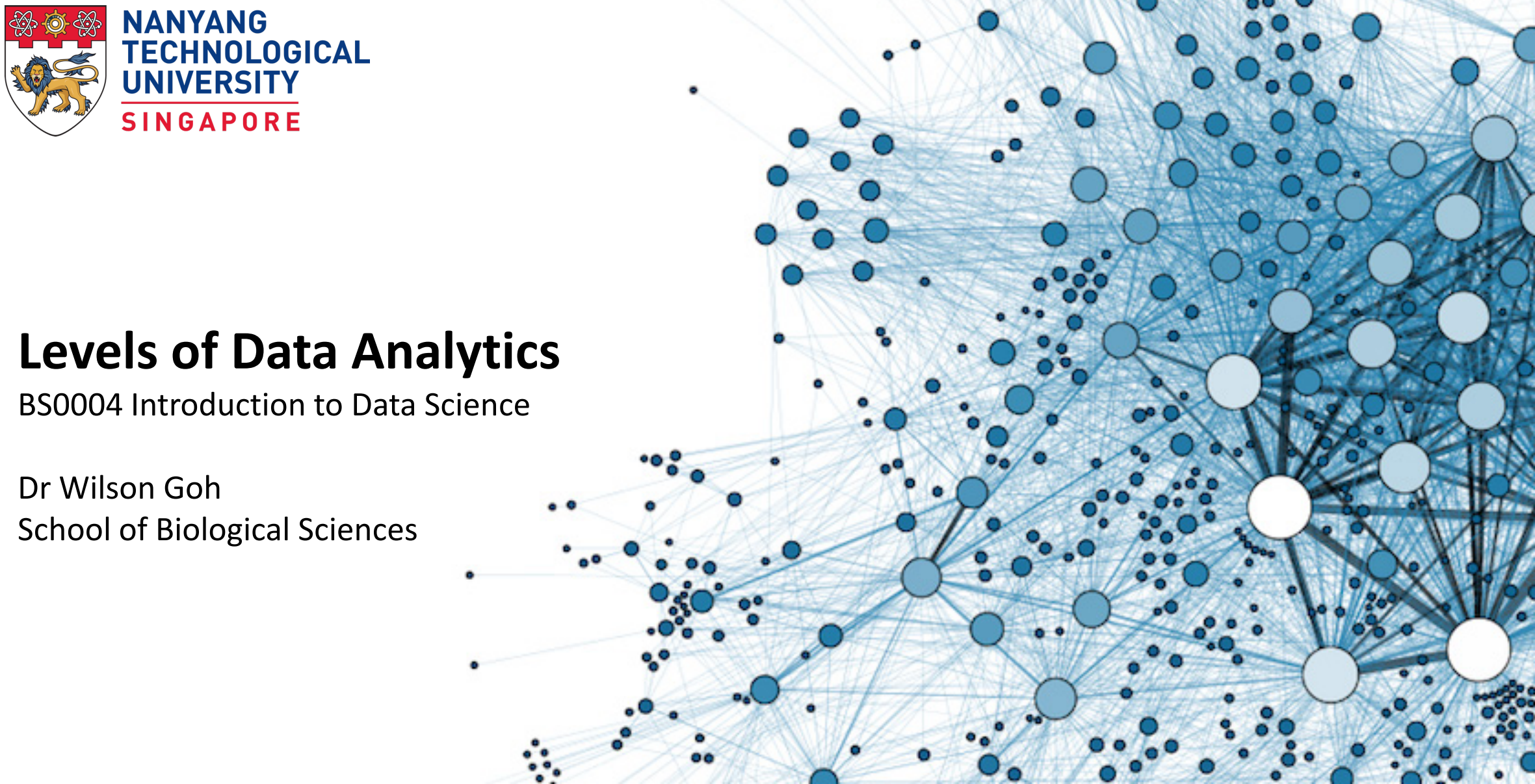


Just swap “Business” with “Biology”

Different Degrees of Data Science



Source: http://www.datasciencecentral.com/



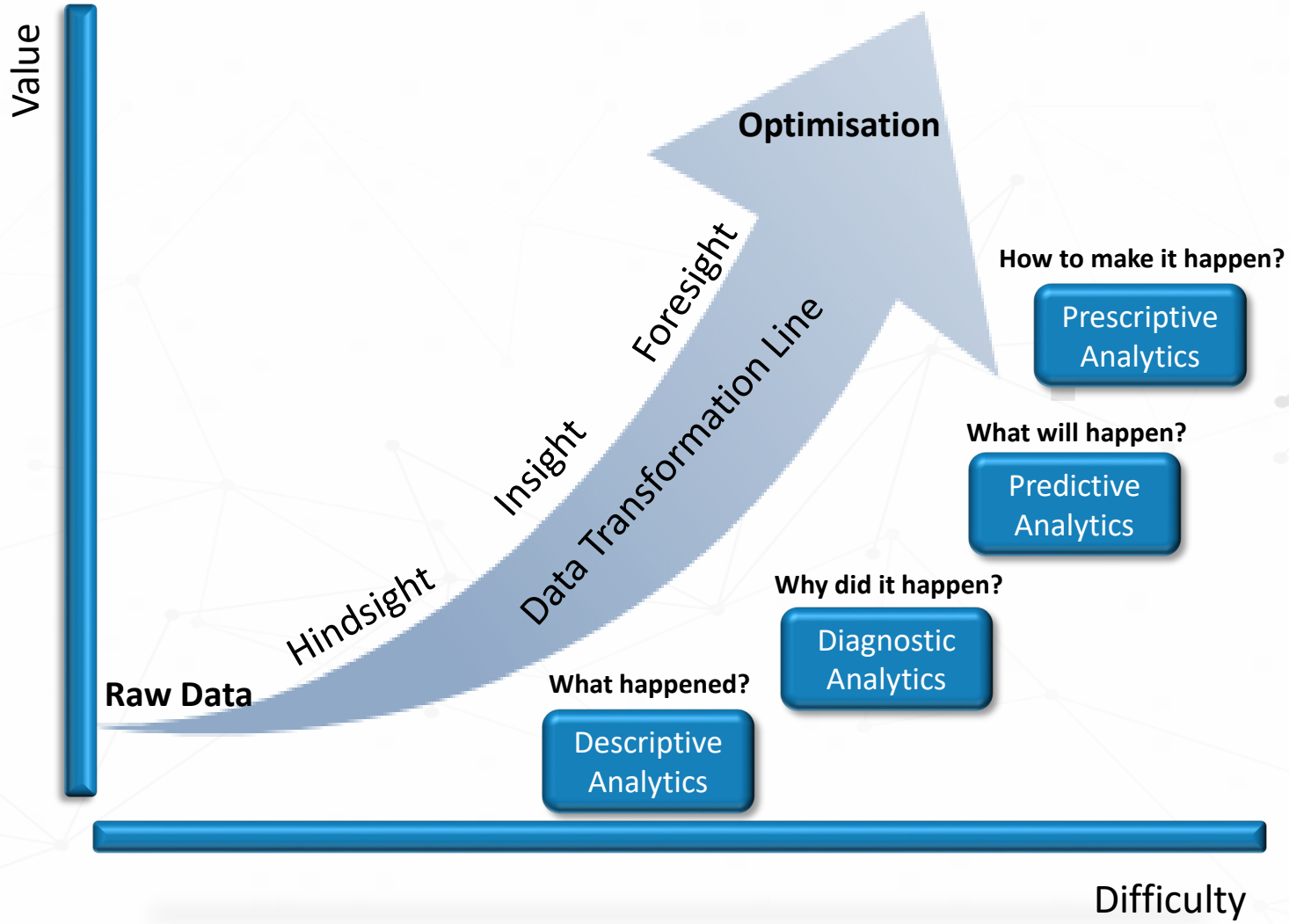
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Levels of Data Analytics

BS0004 Introduction to Data Science

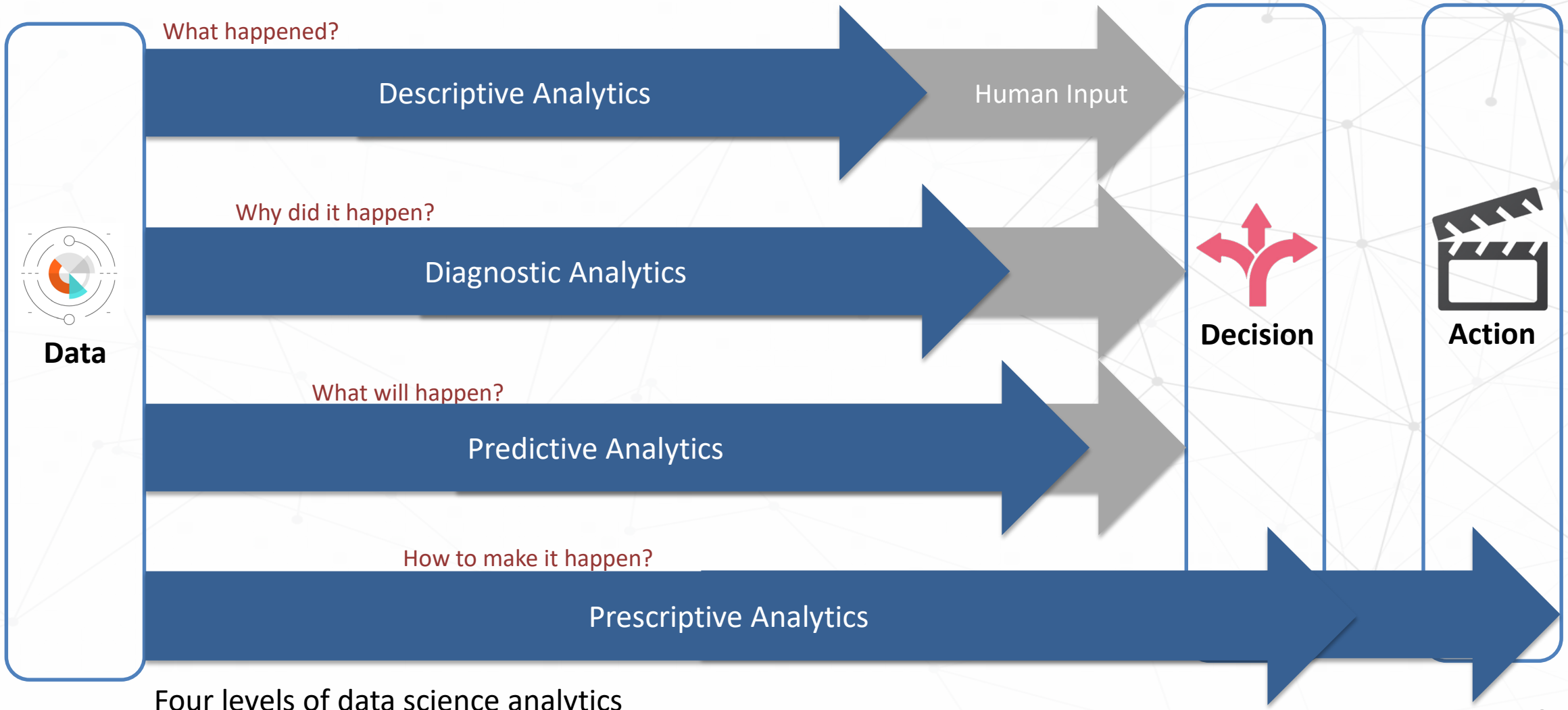
Dr Wilson Goh
School of Biological Sciences

Value and Difficulty

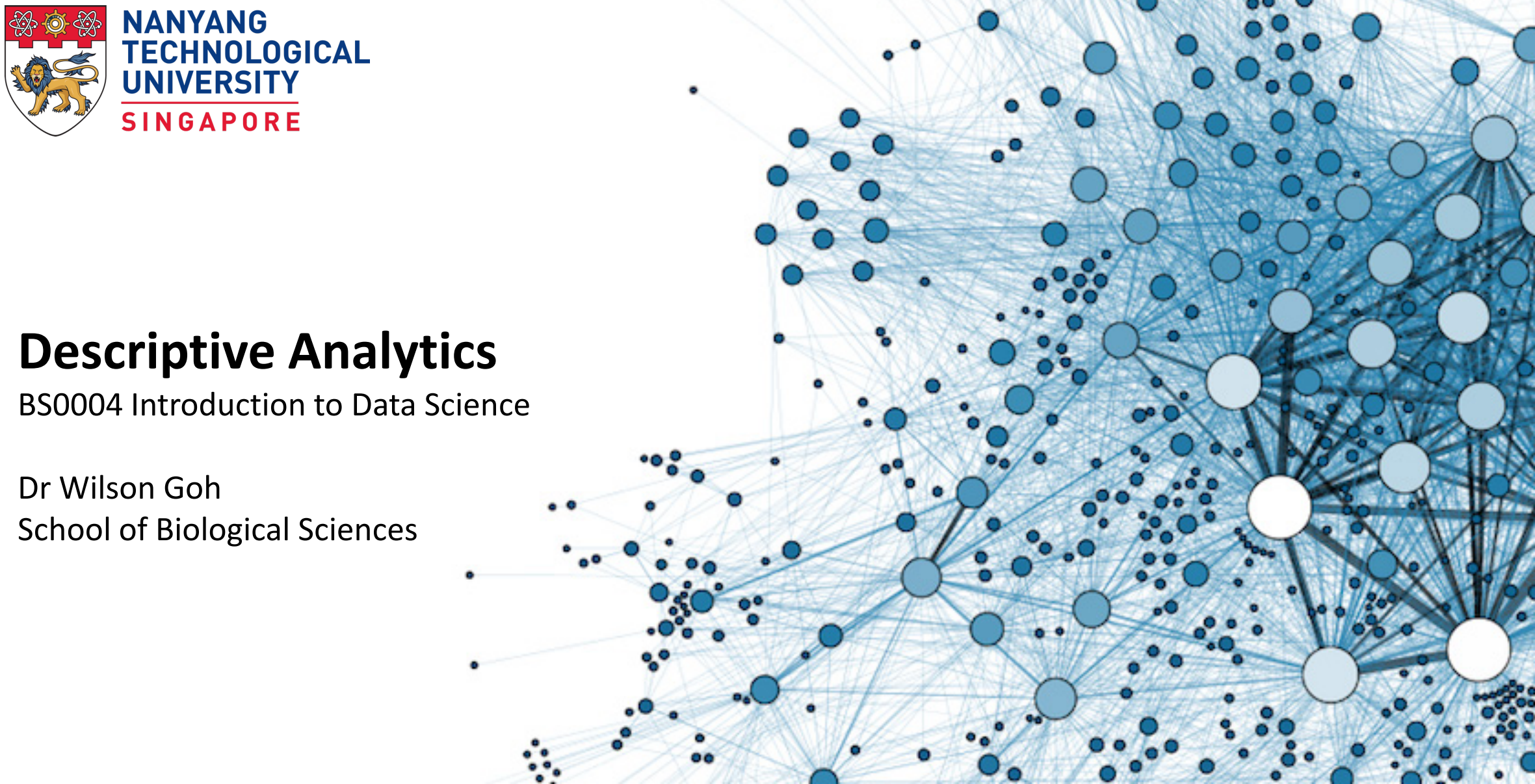


Gartner analytics value-difficulty chart

From Data Science to Action



Four levels of data science analytics



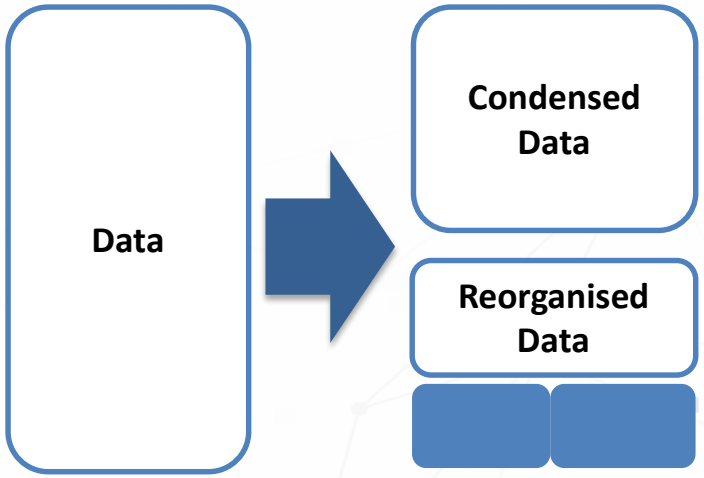
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

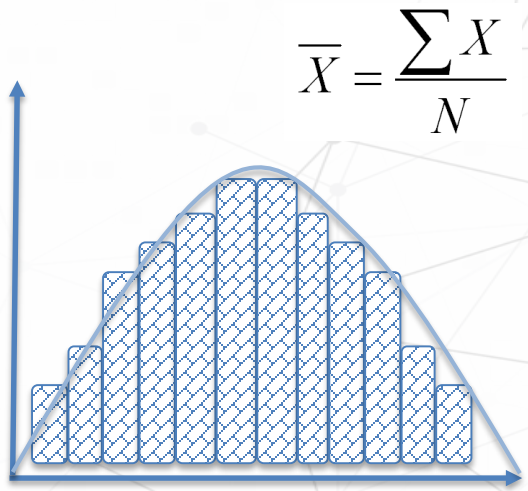
Descriptive Analytics

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

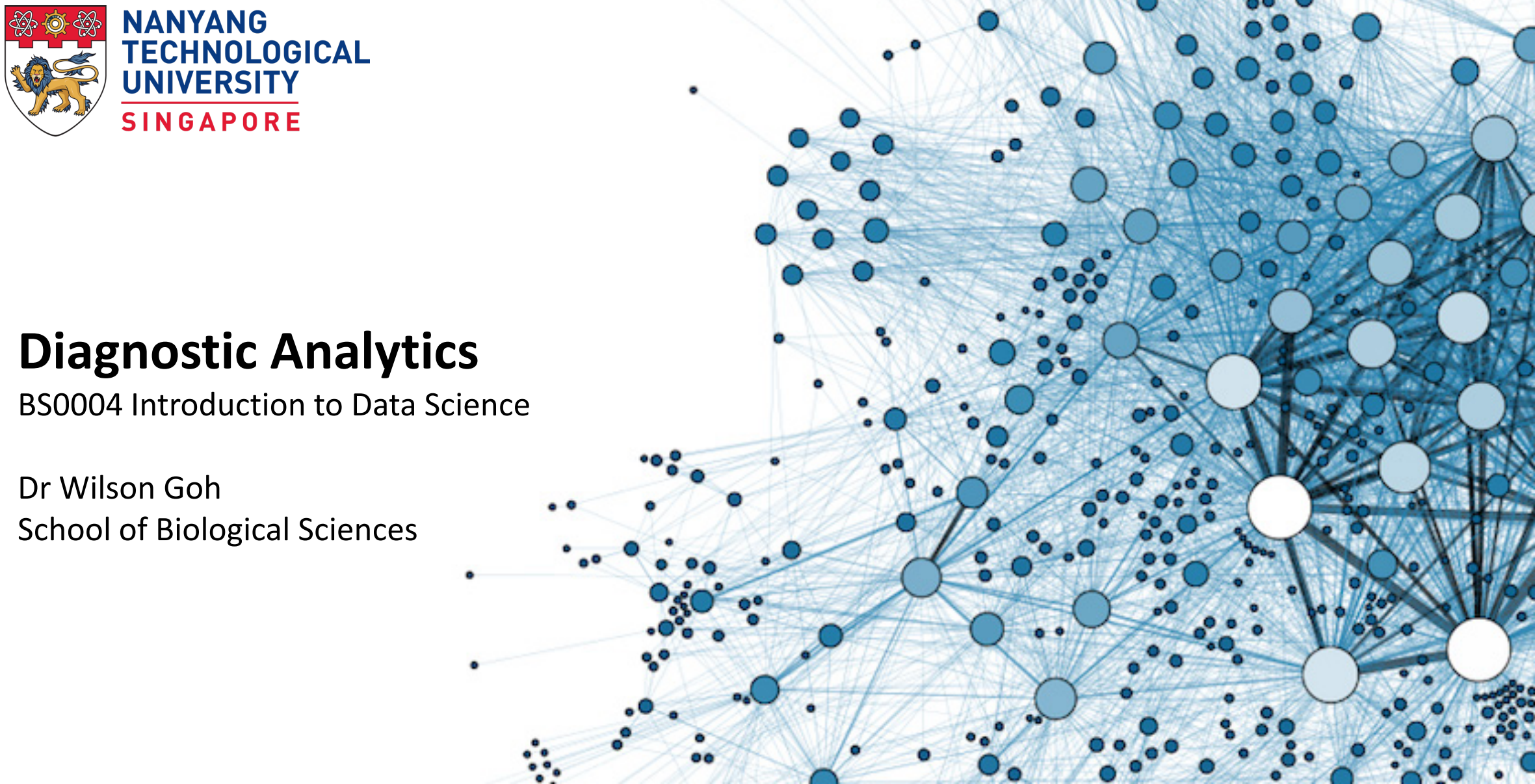
Descriptive Analytics





Summary Statistics

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



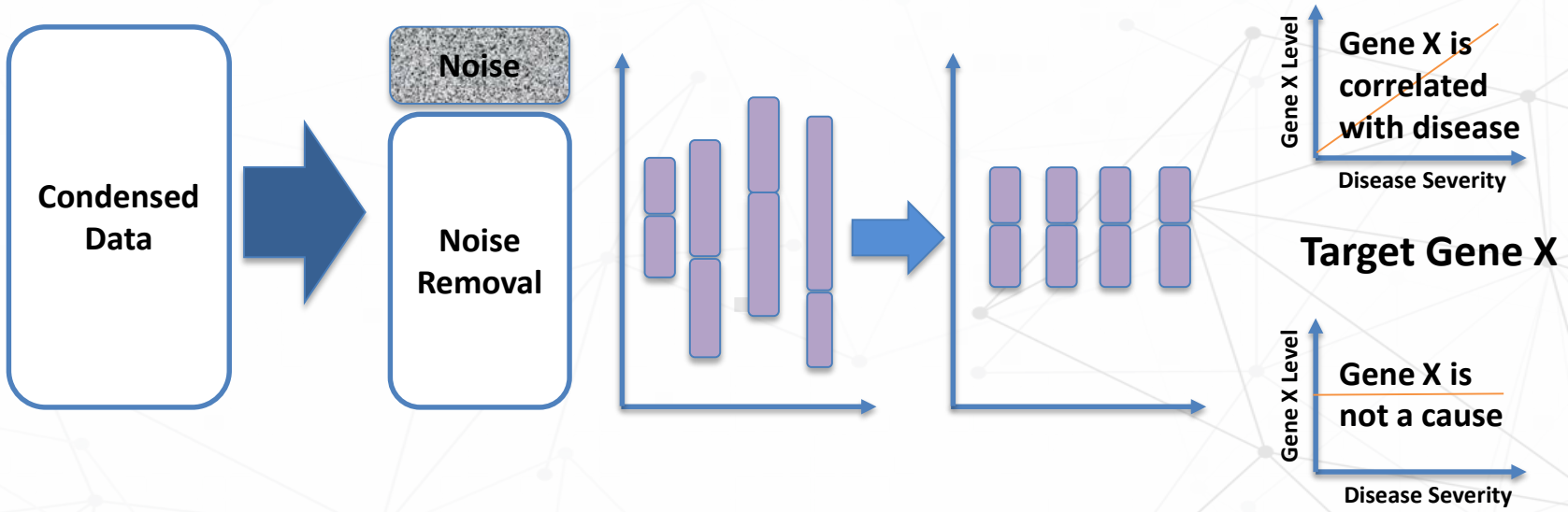
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Diagnostic Analytics

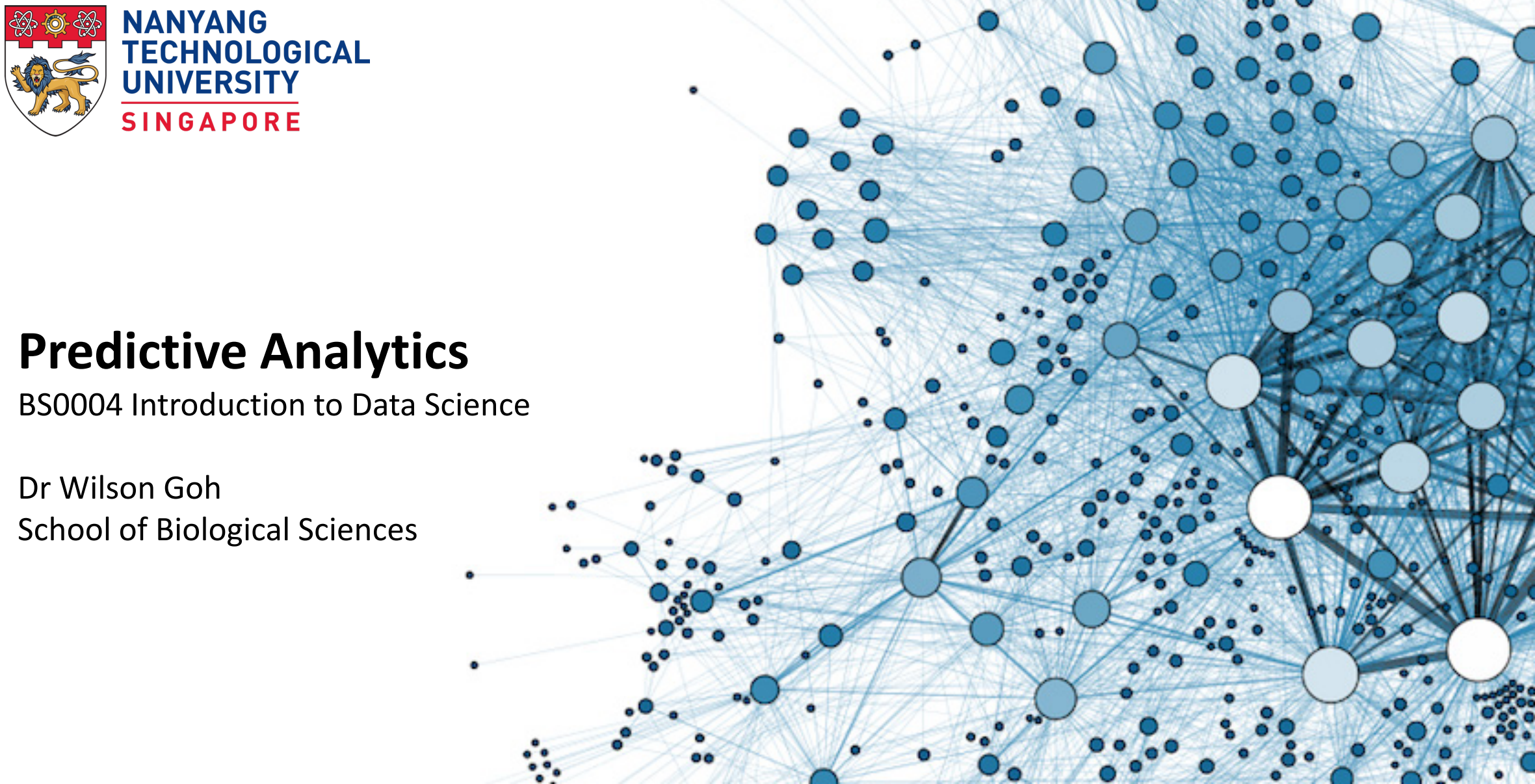
BS0004 Introduction to Data Science

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.



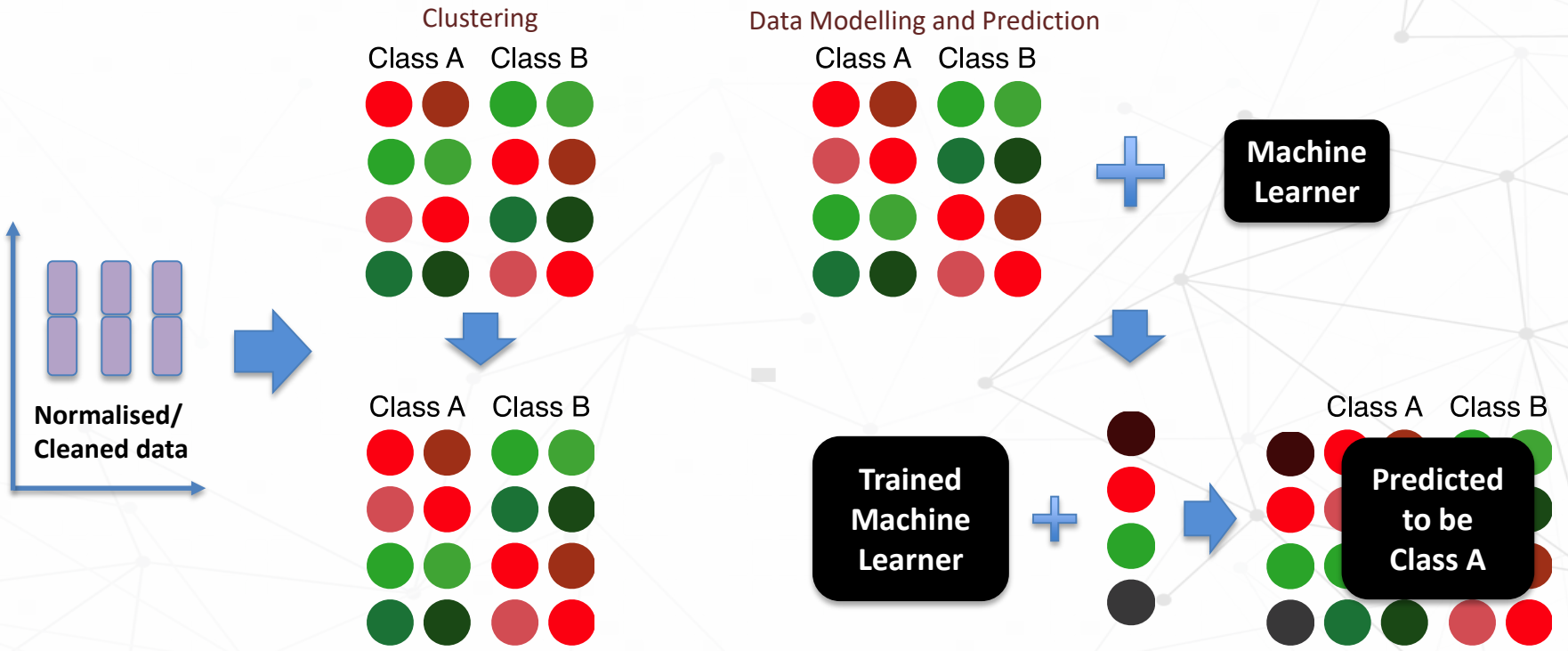
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Predictive Analytics

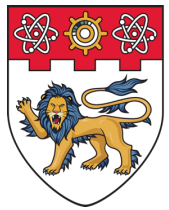
BS0004 Introduction to Data Science

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).



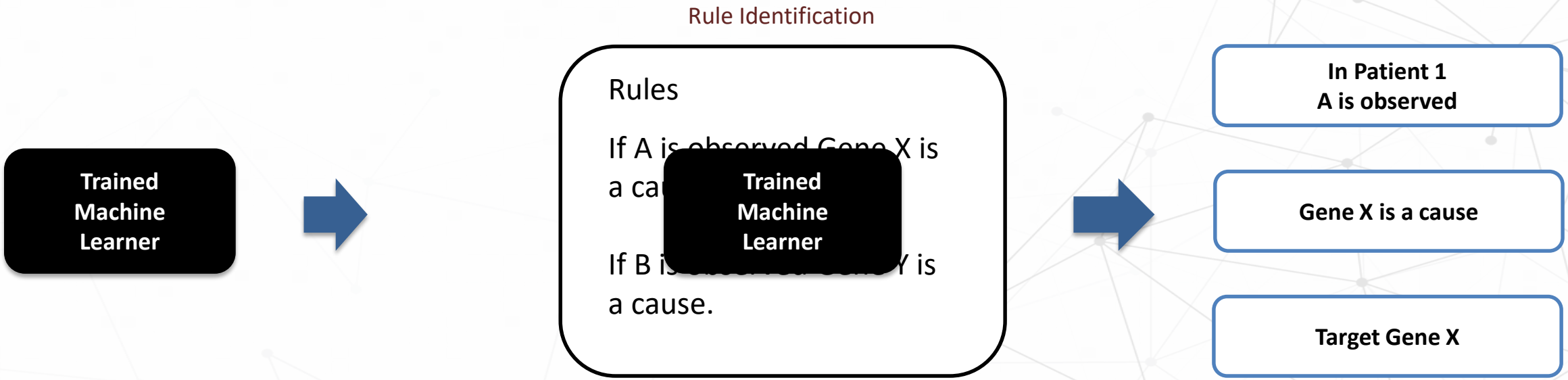
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

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.

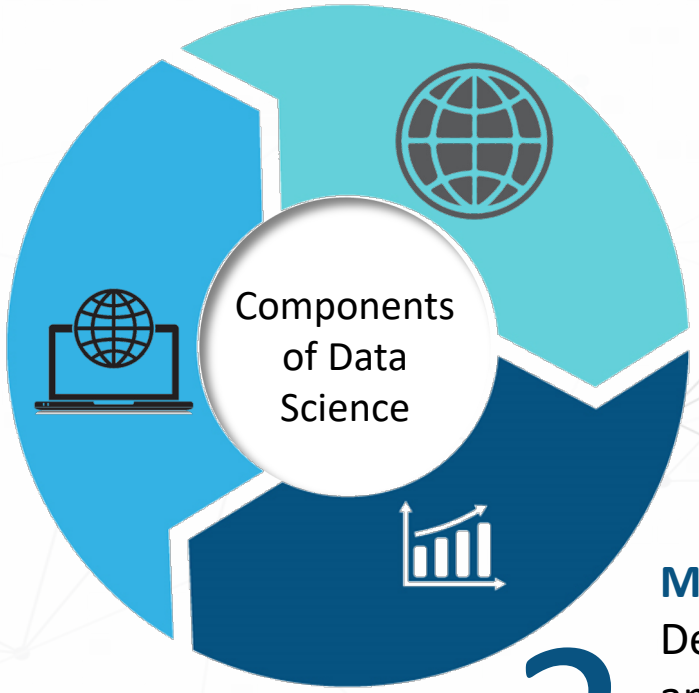
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Components of Data Science

BS0004 Introduction to Data Science

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

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

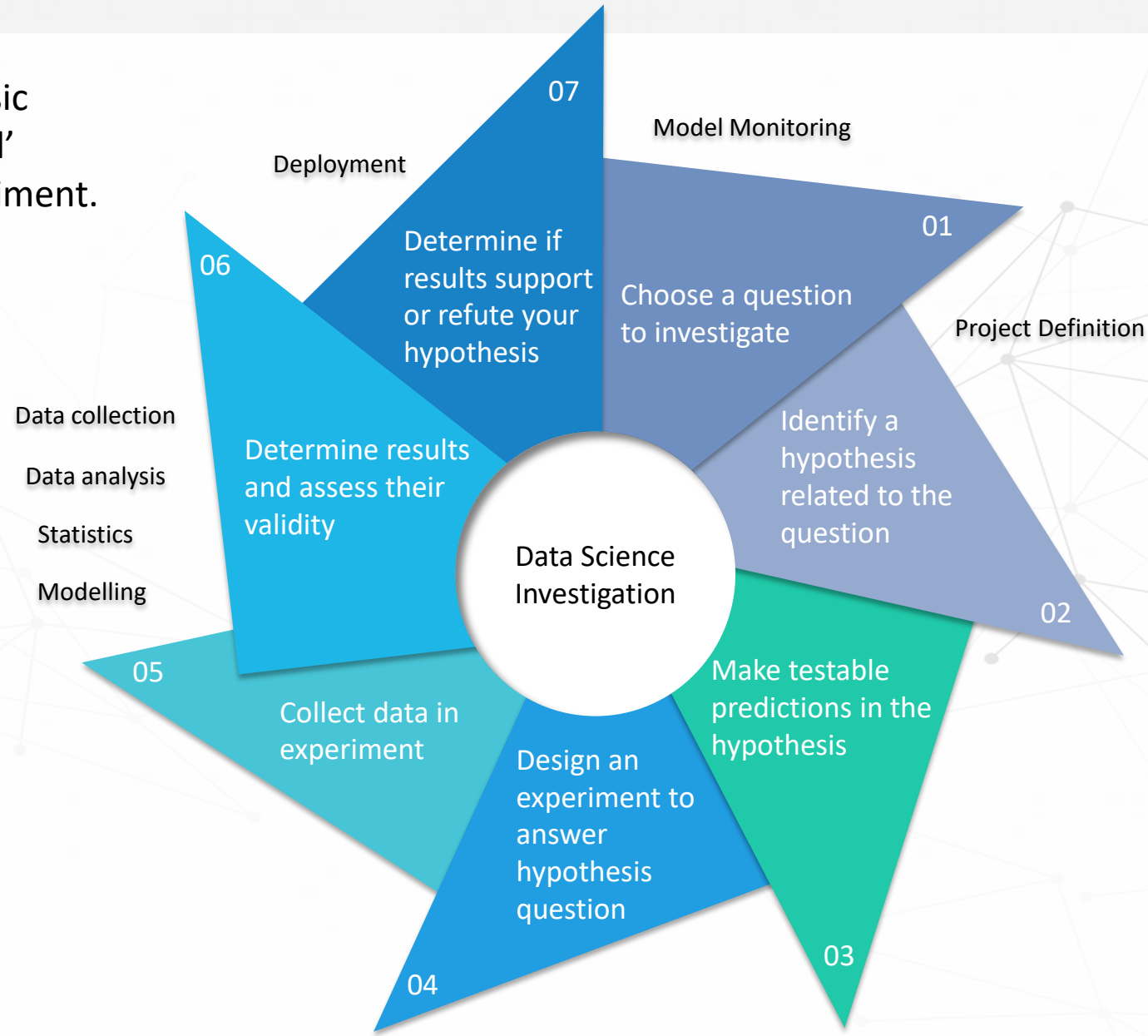
Steps of Data Science Investigation

BS0004 Introduction to Data Science

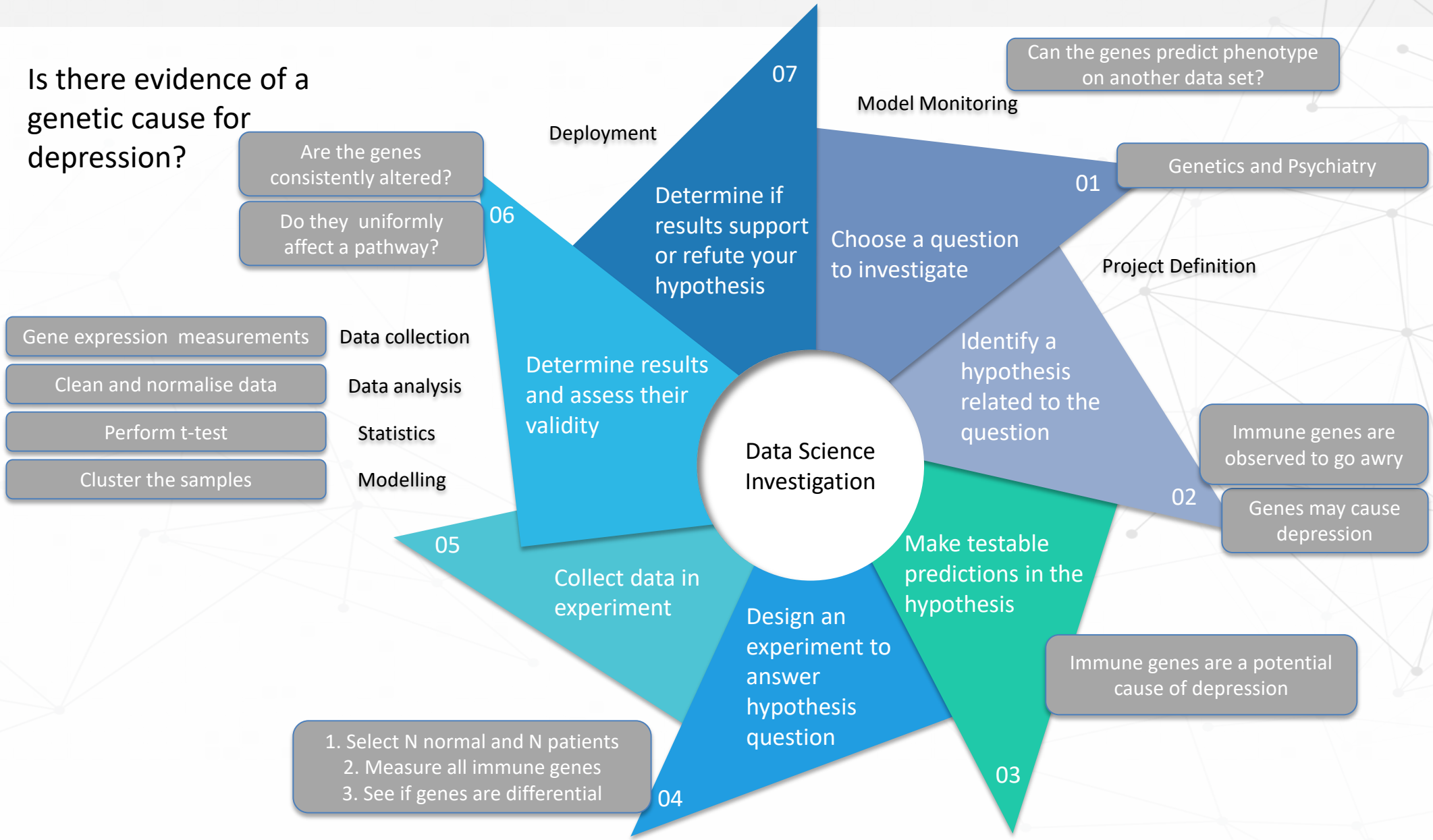
Dr Wilson Goh
School of Biological Sciences

Data Science for Scientific Investigation

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



Data Science for Scientific Investigation



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Risks of Data Analytics

BS0004 Introduction to Data Science

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)

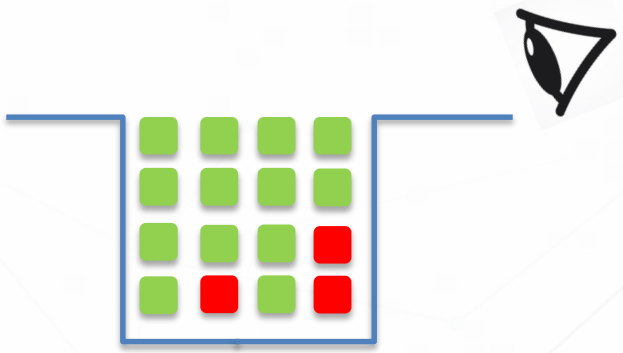
Risks of Data Analytics

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.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Summary

BS0004 Introduction to Data Science

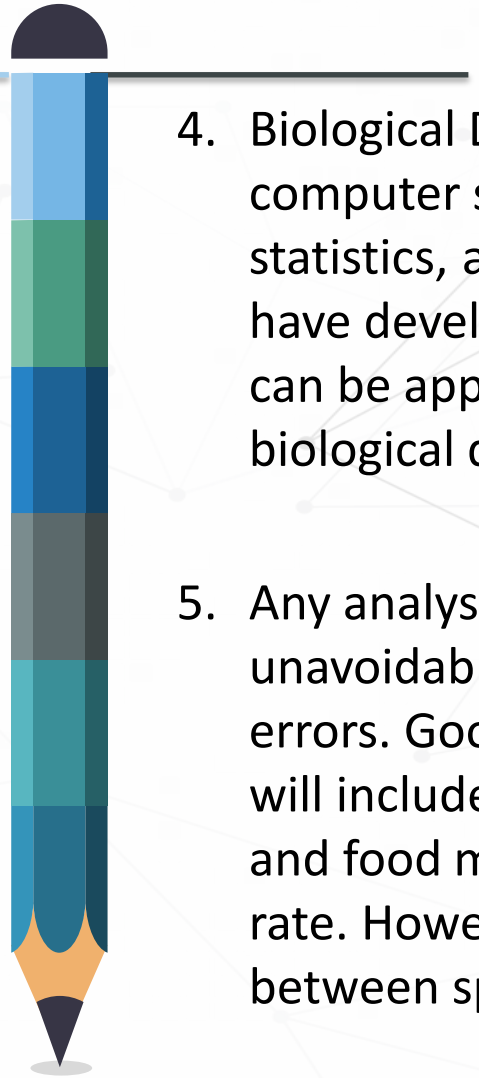
Dr Wilson Goh
School of Biological Sciences

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 the required 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 minimise the error rate. However, there is always a tradeoff between specificity and sensitivity.