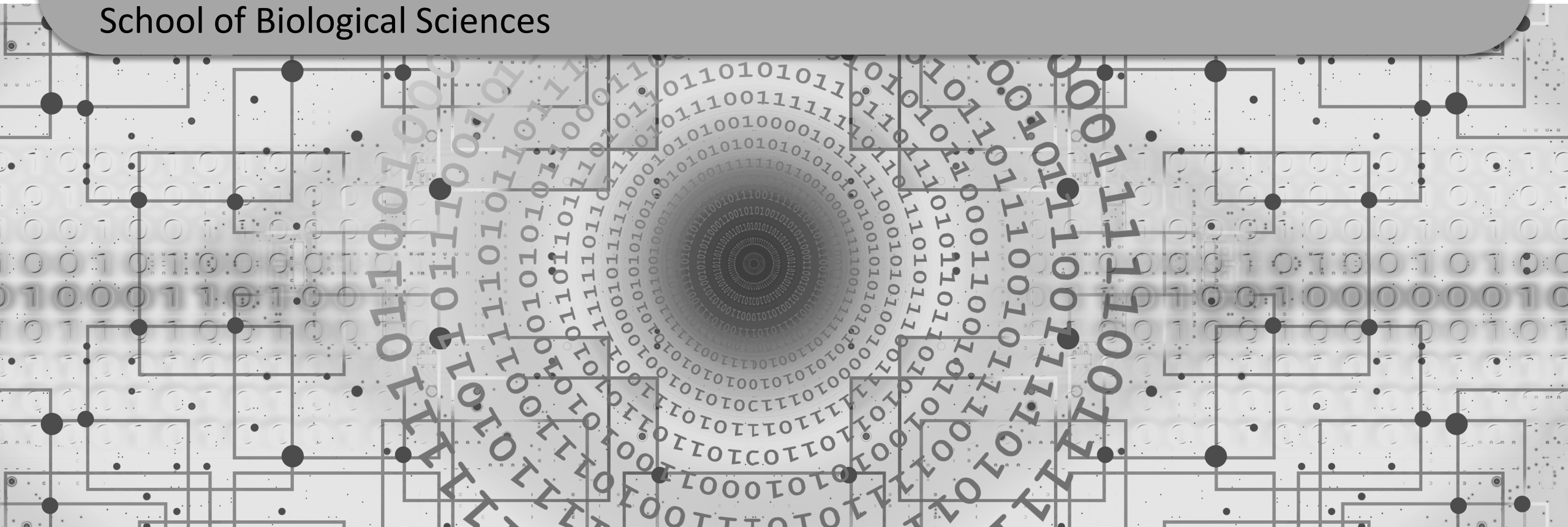
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Databases and Logic

BS0004 Introduction to Data Science

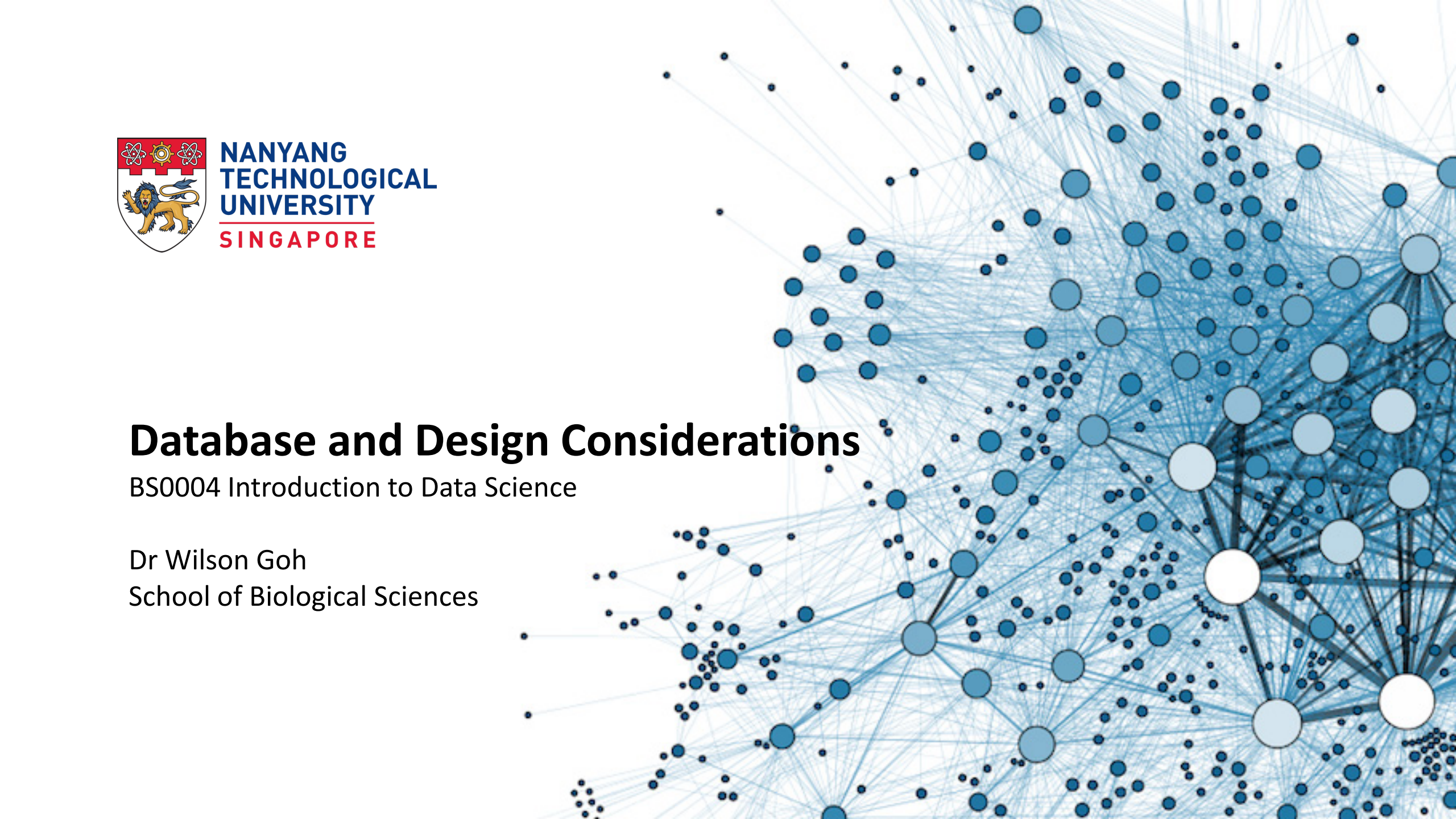
Dr Wilson Goh
School of Biological Sciences



Learning Objectives

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

- Explain the generic design considerations for databases.
- Explain database normalisation.
- Explain relational, flatfile and XML database models.
- Explain the different types of biological databases.
- Explain logic and the three major forms of logical reasoning.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Database and Design Considerations

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

What is a Database?

Structured collection of information.

Consists of basic units called **records** or entries.

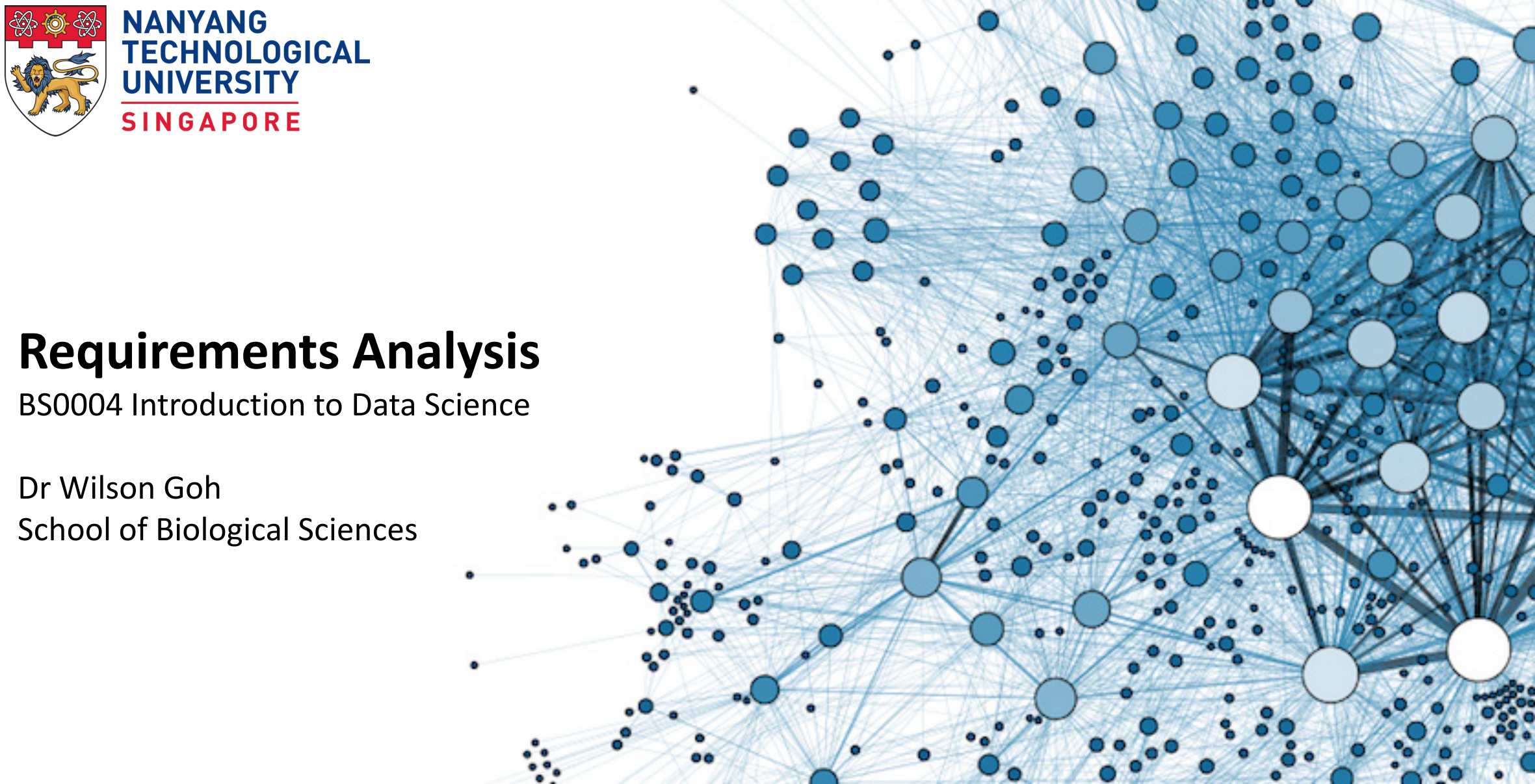
Each record consists of fields, which hold **pre-defined** data related to the record.

For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence).

Database Design

Designing an efficient, useful database is a matter of following the proper process, including these phases:

- Requirements analysis, or identifying the purpose of your database
- Organising data into tables
- Specifying primary keys and analysing relationships
- Normalising to standardise the tables



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Requirements Analysis

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Identifying the Purpose of the Database

- Understanding the purpose of your database will inform your choices throughout the design process.
- Make sure you consider the database from every perspective. For instance, if you were making a database for a public library, you'd want to consider the ways in which both patrons and librarians would need to access the data.

Identifying the Purpose of the Database

Here are some ways to gather information before creating the database:

- Interview the people who will use it.
- Analyse business forms, such as invoices, timesheets, surveys.
- Comb through any existing data systems (including physical and digital files).

Identifying the Purpose of the Database

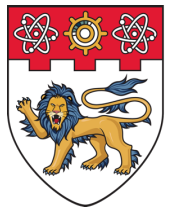
Start by gathering any existing data that will be included in the database. Then list the types of data you want to store and the entities, or people, things, locations, and events, that those data describe, like:

Customers
Name
Address
City, State, Zip
Email Address

Products
Name
Price
Quantity in Stock
Quantity on Order

Orders
Order ID
Sales Representative
Date
Product(s)
Quantity
Price
Total

Be sure to break down the information into the smallest useful pieces. For instance, consider separating the street address from the country so that you can later filter individuals by their country of residence. Also, avoid placing the same data point in more than one table, which adds unnecessary complexity.



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Organising Data

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Organising Data into Tables

Lay out a visual representation of your database.

To do that, you need to understand how relational databases are structured.

Within a database, related data are grouped into tables, each of which consists of rows (also called tuples) and columns, like a spreadsheet.

To convert your lists of data into tables, start by creating a table for each type of entity, such as products, sales, customers, and orders.

Organising Data into Tables

An example of the **Customers Table**

First Name	Last Name	Age	Zip Code
Roger	Williams	43	34760
Jerrica	Jorgensen	32	97453
Samantha	Hopkins	56	64829

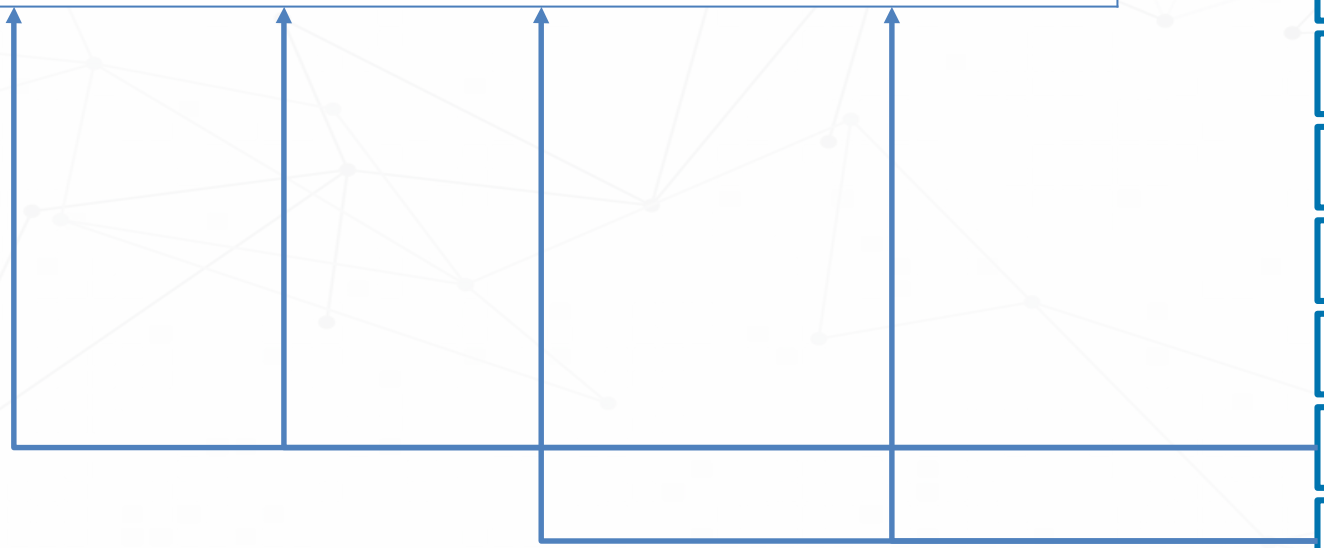
Each row of a table is called a record. Records include data about something or someone, such as a particular customer.

Columns (also known as fields or attributes) contain a single type of information that appears in each record.

Organising Data into Tables

An example of the **Customers Table**

First Name	Last Name	Age	Zip Code
Roger	Williams	43	34760
Jerrica	Jorgensen	32	97453
Samantha	Hopkins	56	64829



To enforce consistency, each column can be assigned to only hold data of a particular type. Other data types:

- CHAR: a specific length of text
- TEXT: large amounts of text
- FLOAT, DOUBLE: floating point numbers
- BLOB: binary data
- VARCHAR: text of variable lengths
- INT: positive or negative whole number

Organising Data into Tables

An example of the **Customers Table**

First Name	Last Name	Age	Zip Code
Roger	Williams	43	34760
Jerrica	Jorgensen	32	97453
Samantha	Hopkins	56	64829

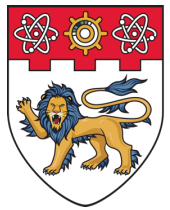
Not sufficiently unique. Assigning a unique user identifier or username would be better. A primary key (PK) is a **unique identifier** for a given entity (Table), meaning that you could pick out an exact customer even if you only knew that value.

Decide which **attribute** or **attributes** will serve as the **primary key** for each table, if any. Attributes chosen as primary keys should be **unique, unchanging, and always present** (never NULL or empty).

Organising Data into Tables

When it comes time to create the actual database, you'll put both the logical data structure and the physical data structure into the data definition language supported by your database management system.

At that point, you should also estimate the size of the database to be sure you can get the performance level and storage space it will require.



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Creating Relationships between Data Tables

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Creating Relationships between Tables/ Entities

With your data now broken down into tables, you're ready to analyse the relationships between those tables.

Customers
Name
Address
City, State, Zip
Email Address

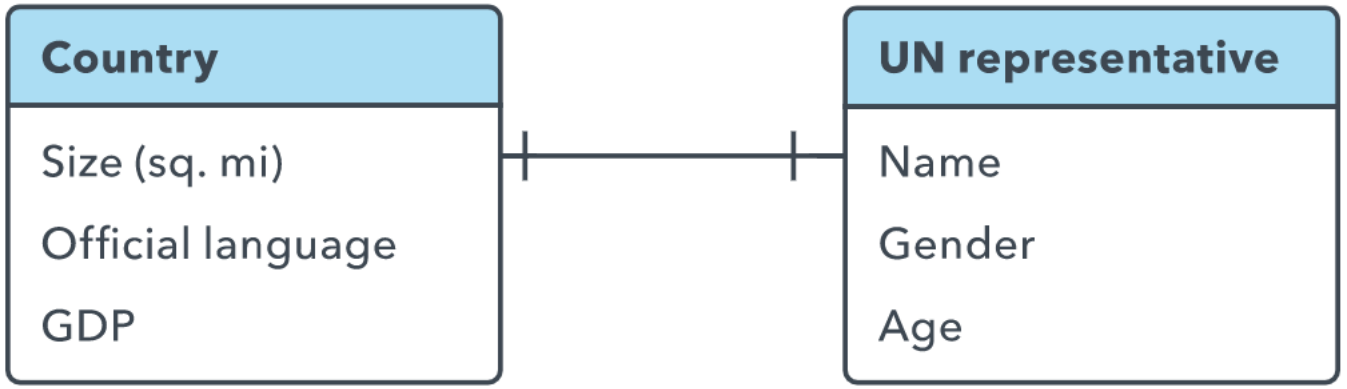
Products
Name
Price
Quantity in Stock
Quantity on Order

Orders
Order ID
Sales Representative
Date
Product(s)
Quantity
Price
Total

Cardinality refers to the quantity of elements that interact between two related tables. Identifying the cardinality helps make sure you've divided the data into tables most efficiently.

Creating Relationships between Tables/ Entities

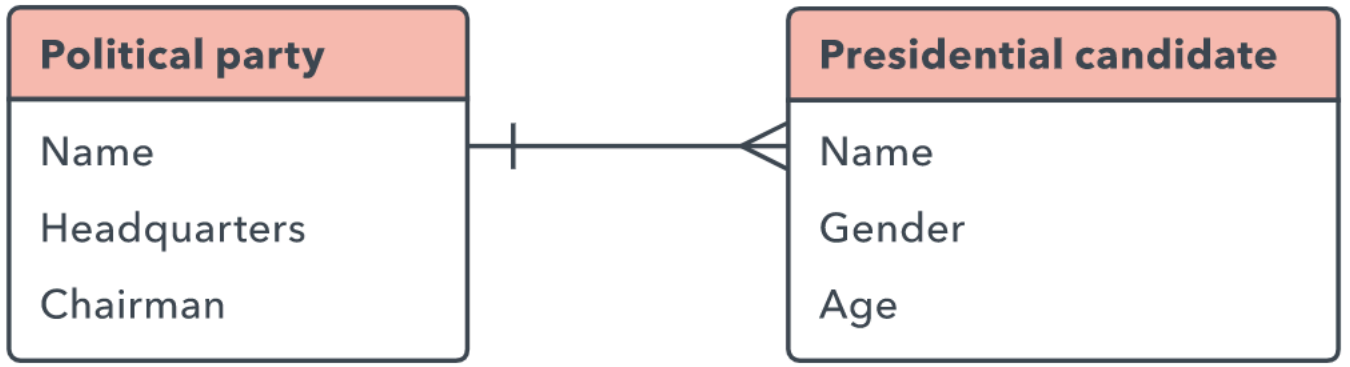
One-to-one Relationships



- When there's only one instance of Entity A for every instance of Entity B, they are said to have a one-to-one relationship (often written 1:1).
- A 1:1 relationship usually indicates that you'd be better off combining the two tables' data into a single table.
- To guarantee that the data matches up correctly, you'd then have to include at least one identical column in each table, most likely the primary key.

Creating Relationships between Tables/ Entities

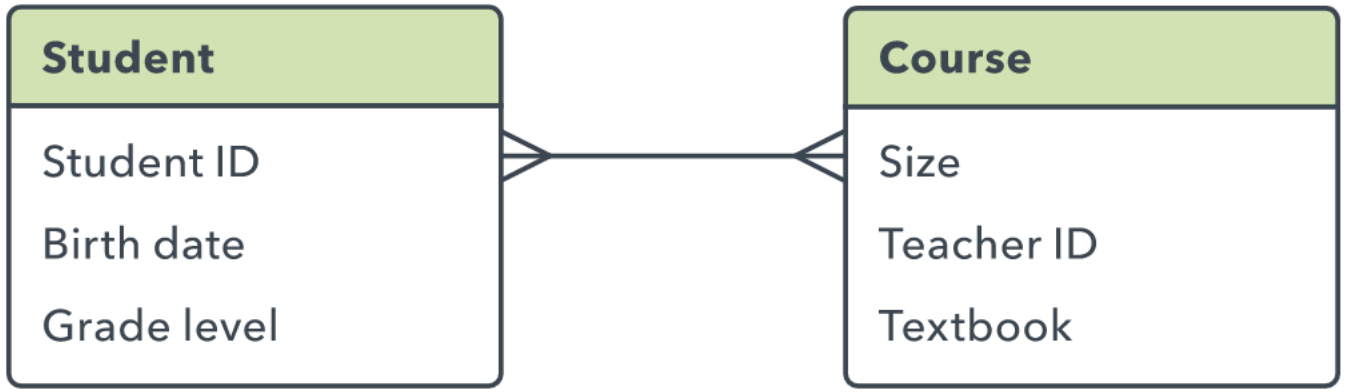
One-to-many Relationships



- To implement a 1:M relationship as you set up a database, simply add the primary key from the “one” side of the relationship as an attribute in the other table.
- When a primary key is listed in another table in this manner, it’s called a foreign key.
- The table on the “1” side of the relationship is a considered a parent table to the child table on the other side.

Creating Relationships between Tables/ Entities

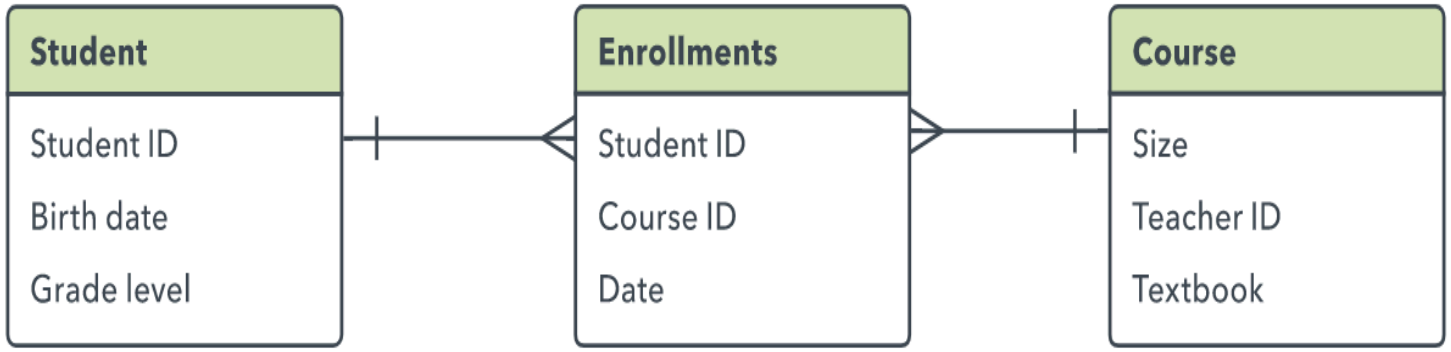
Many-to-many Relationships



- It's not directly possible to implement this kind of relationship in a database. Instead, you have to break it up into two one-to-many relationships.

Creating Relationships between Tables/ Entities

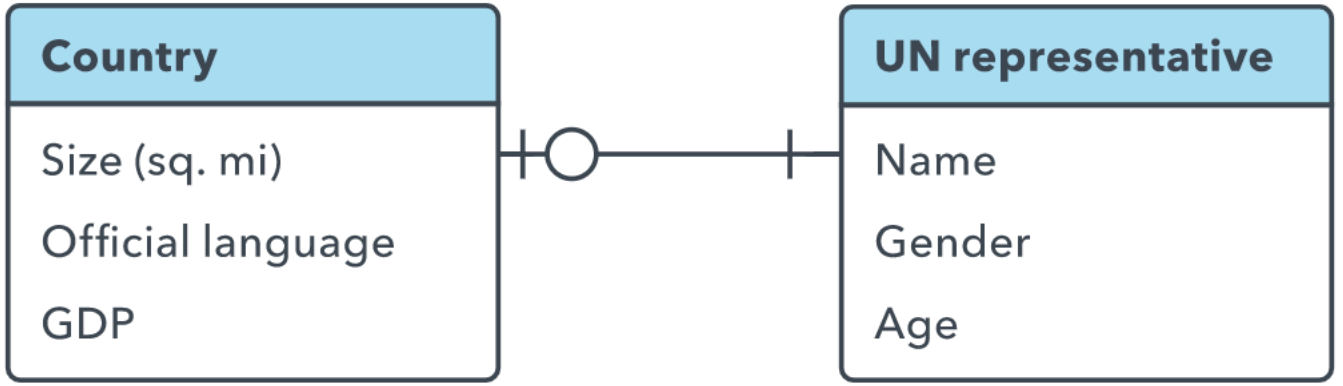
Many-to-many Relationships



- Each record in the link table would match together two of the entities in the neighboring tables (it may include supplemental information as well).
- For instance, a link table between students and classes might look like the above.

Creating Relationships between Tables/ Entities

Mandatory or Not?



- Another way to analyse relationships is to consider which side of the relationship has to exist for the other to exist. The mandatory side can be marked with a circle on the line where a dash would be.
- For instance, a country has to exist for it to have a representative in the United Nations, but the opposite is not true.
- Two entities can be mutually dependent (one could not exist without the other).

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Normalisation of Data

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Normalisation

Once you have a preliminary design for your database, you can apply normalisation rules to make sure the tables are structured correctly.

- Think of these rules as industry standards
- Normalisation in tiers (levels 1, 2, 3)
- Inheritance: each form, or level of normalisation, includes the rules associated with the lower forms

Database Normalisation

First Normal Form (1NF)

Rule: Each cell in the table can have only one value, never a list of values.

Product ID	Colour	Price
1	Brown, Yellow	$15
2	Red, Green	$13
3	Blue, Orange	$11

Database Normalisation

First Normal Form (1NF)

Product ID	Colour	Price
1	Brown, Yellow	$15
2	Red, Green	$13
3	Blue, Orange	$11

Try splitting this up?

X

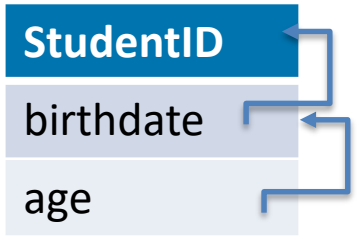
Products
Color1
Color2
Color3
Price

A table with groups of repeated or closely related attributes does not meet the first normal form.

Database Normalisation

Second Normal Form (2NF)

Rule: Each attribute should be fully dependent on the primary key.

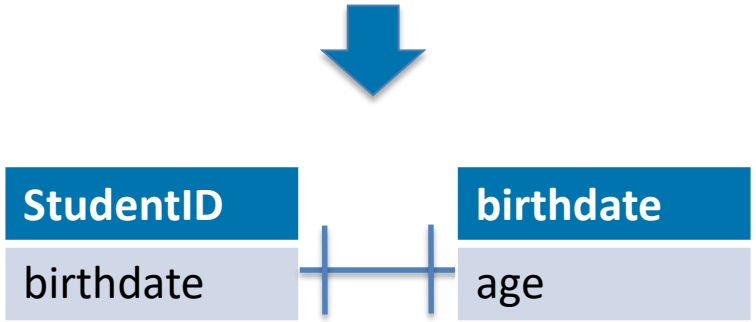


"birthdate" which in turn depends on "studentID"

"age" depends on "birthdate"

X 2NF

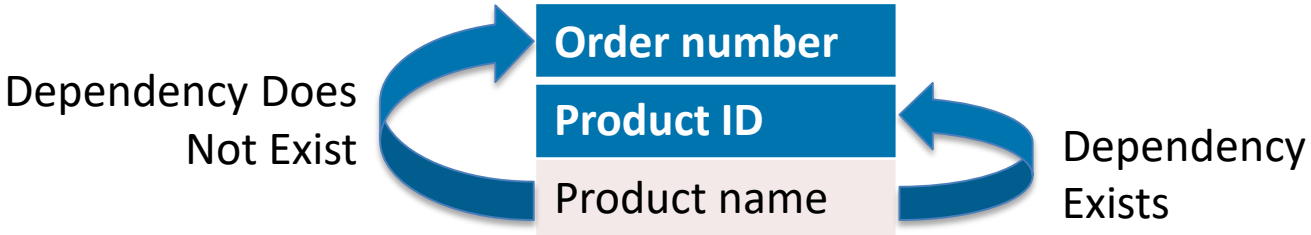
Age does not depend directly on studentID.



Database Normalisation

Second Normal Form (2NF)

Rule: A table with a primary key made up of multiple fields violates the second normal form if one or more of the other fields do not depend on every part of the key. So be careful of using multi-attribute keys.



X 2NF

Database Normalisation

Third Normal Form (3NF)

Rule: Every non-key column be independent of every other column. This keeps you from storing any derived data in the table.

Order	Price	Tax
14325	$40.99	$2.05
14326	$13.73	$0.69
14327	$24.15	$1.21

The “tax” column directly depends on the total price of the order.

Database Normalisation

While these forms explain the best practices to follow generally, the degree of normalisation depends on the context of the database.

Online Transaction Processing (OLTP)

- Users are concerned with creating, reading, updating, and deleting records, should be normalised.
- Ease of updating and changing.

Online Analytical Processing (OLAP)

- Favour analysis and reporting might fare better with a degree of denormalisation, since the emphasis is on speed of calculation. These include decision support applications in which data needs to be analysed quickly but not changed.
- Ease of analysis.

Data Integrity Rules

Entity Integrity Rule: Primary key must be unique.

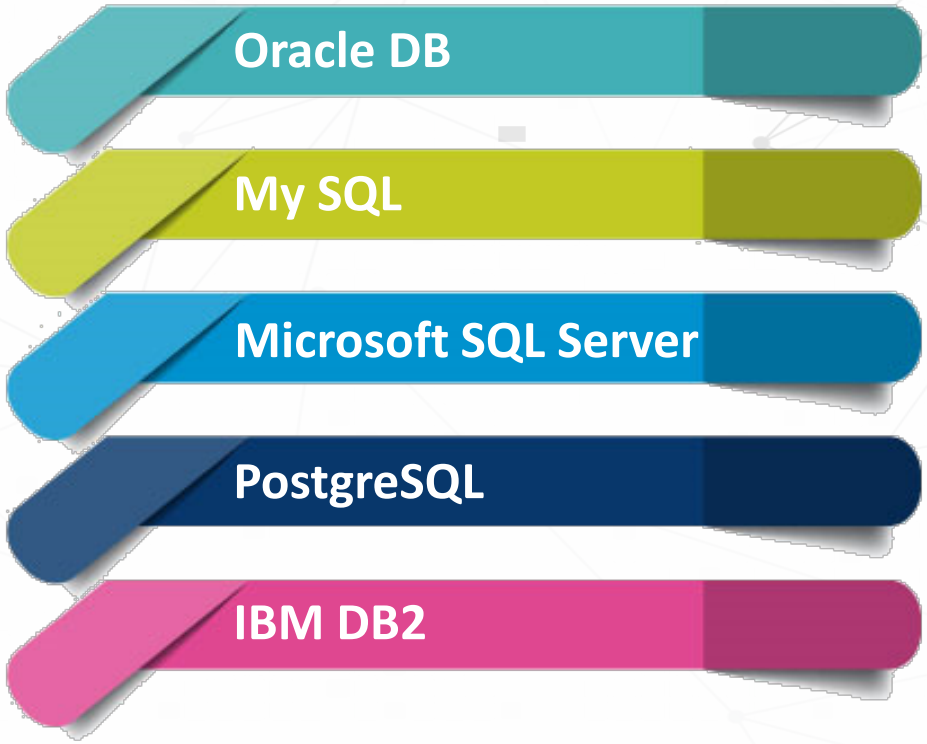
The primary key can never be NULL. If the key is made up of multiple columns, none of them can be NULL. Otherwise, it could fail to uniquely identify the record.

Referential Integrity Rule: Each foreign key must be matched with 1 primary key.

If the primary key changes or is deleted, those changes will need to be implemented wherever that key is referenced throughout the database.

Database Management Systems

Many of the design choices you will make also depend on which database management system you use. Some of the most common systems include:



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Other Database Models

BS0004 Introduction to Data Science

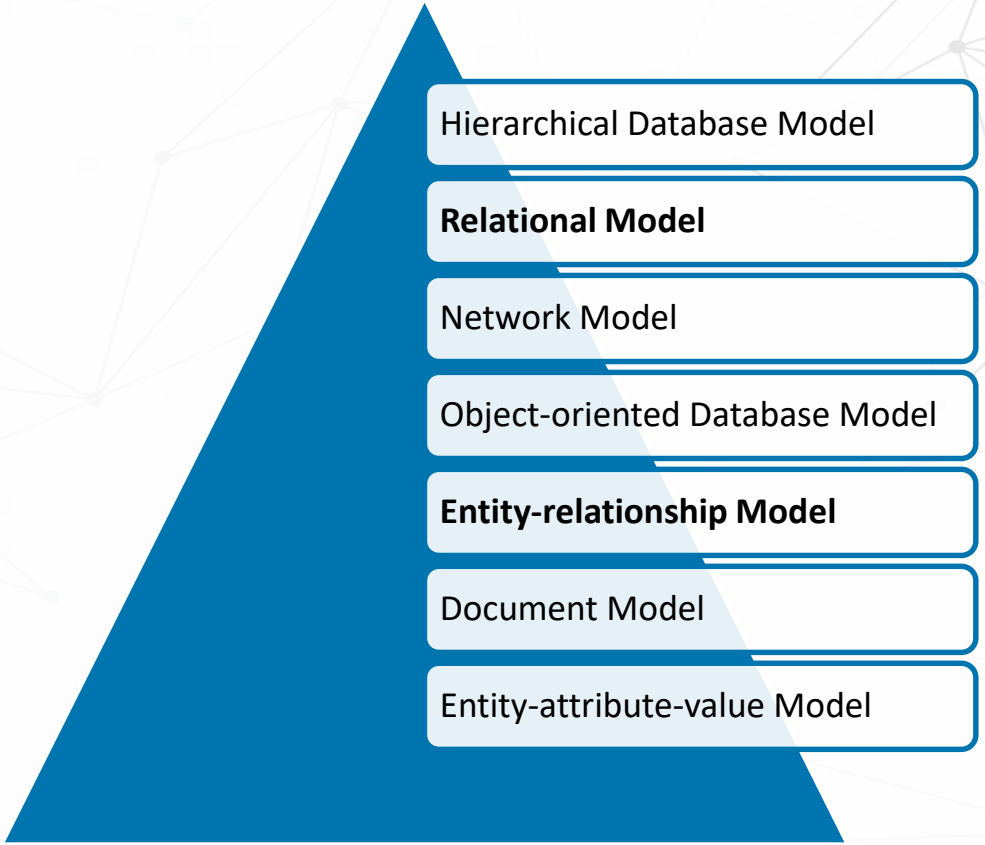
Dr Wilson Goh
School of Biological Sciences

What is a Database Model?

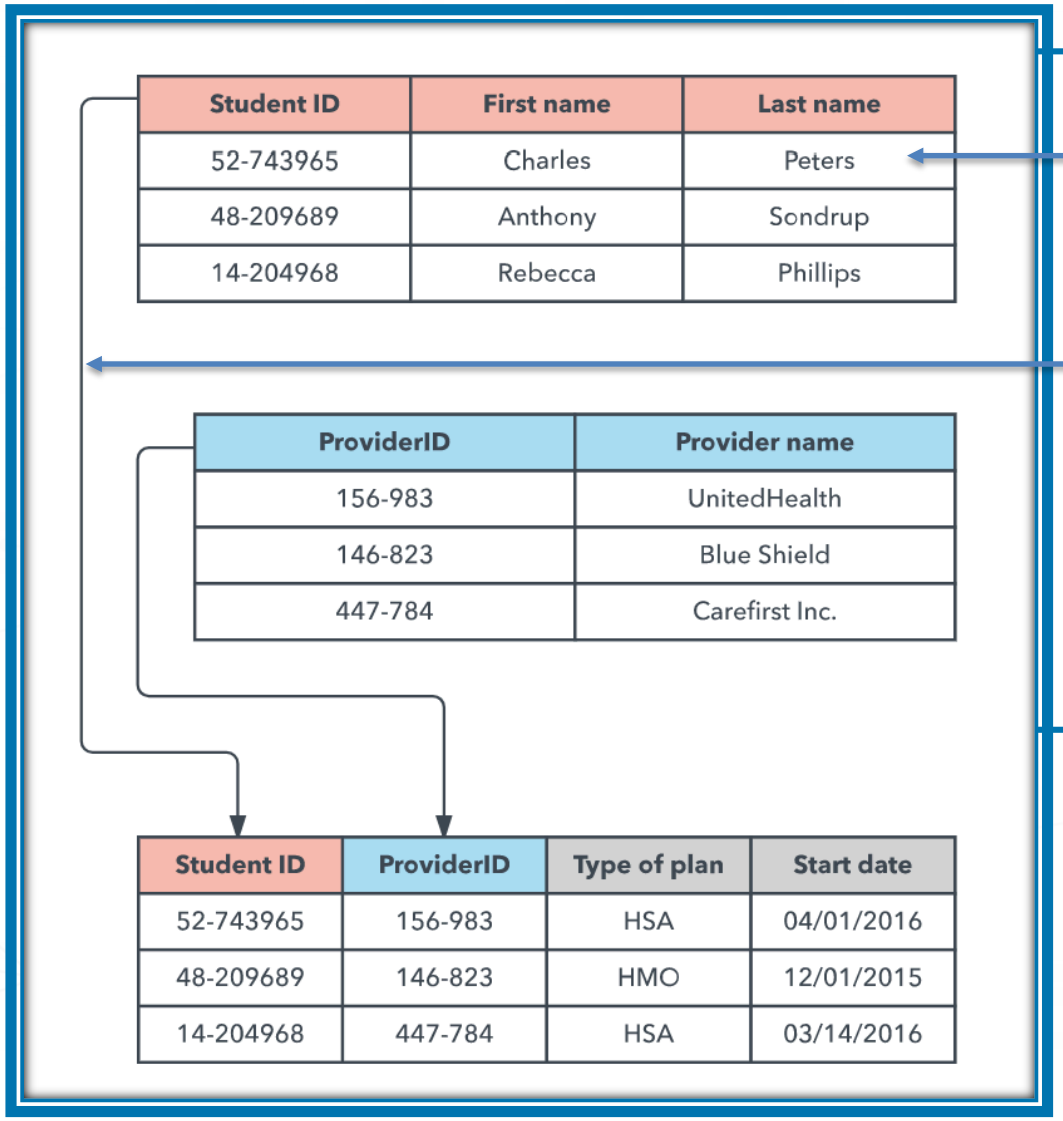
A database model shows the logical structure of a database, including the relationships and constraints that determine how data can be stored and accessed.

Types of Database Models

You may choose to describe a database with any one of these depending on several factors. The biggest factor is whether the database management system you are using supports a particular model.



Relational Model



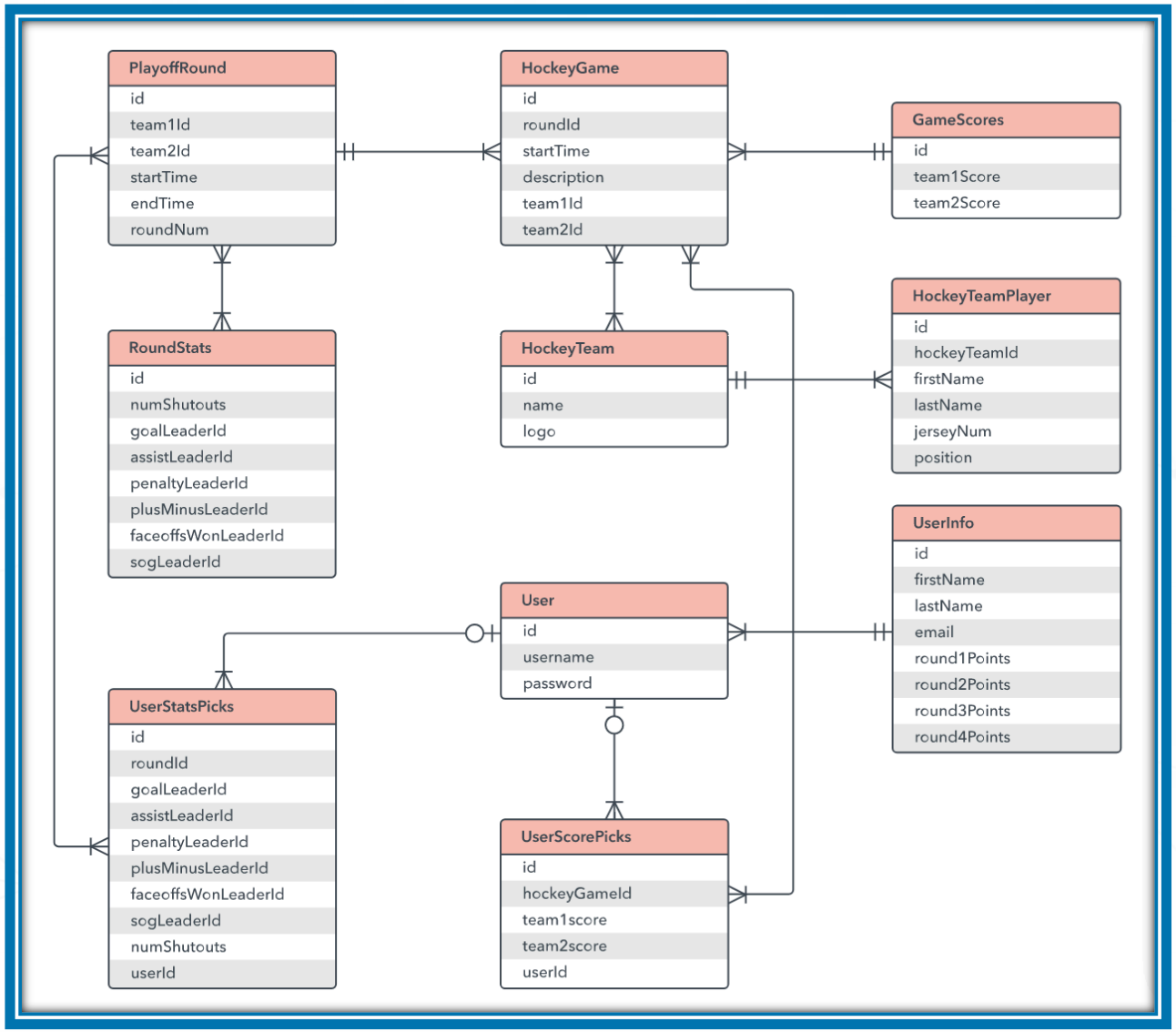
Student ID	First name	Last name
52-743965	Charles	Peters
48-209689	Anthony	Sondrup
14-204968	Rebecca	Phillips

ProviderID	Provider name
156-983	UnitedHealth
146-823	Blue Shield
447-784	Carefirst Inc.

Student ID	ProviderID	Type of plan	Start date
52-743965	156-983	HSA	04/01/2016
48-209689	146-823	HMO	12/01/2015
14-204968	447-784	HSA	03/14/2016

- Data is organised into tables.
- Each row is called a tuple.
- Types of relationships between tables.
- Within the database, tables can be normalised.
- Relational databases are typically written in Structured Query Language (SQL).

Entity Relational Models



Other Database Model

Flat Model

```
LOCUS       eIF4E          2881 bp    DNA             INV       31-AUG-1999
DEFINITION  Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
            gene, alternative splice products, complete cds.
ACCESSION   eIF4E
VERSION
KEYWORDS    .
SOURCE      fruit fly.
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
            Pterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea;
            Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 2881)
  AUTHORS   Lavoie,C.A., Lachance,P.E.D., Sonenberg,N. and Lasko,P.
  TITLE     Alternatively spliced transcripts from the Drosophila eIF4E gene
            produce two different Cap-binding proteins
  JOURNAL   J. Biol. Chem. 271 (27), 16393-16398 (1996)
  MEDLINE   96279193
REFERENCE   2  (bases 1 to 2881)
  AUTHORS   Darwin,C.R.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-AUG-1999) Evolutionary Biology Department, Oxbridge
            University, 1859 Tennis Court Lane, Camford OX1 2BH, England
FEATURES             Location/Qualifiers
     source          1..2881
                     /organism="Drosophila melanogaster"
                     /strain="Oregon R"
                     /chromosome="3"
                     /map="67A8-B2"
     gene            80..2881
                     /gene="eIF4E"
     CDS             join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
                     /gene="eIF4E"
                     /codon_start=1
                     /product="eukaryotic initiation factor 4E-II"
                     /translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETG
                     EPAGNTATTTAPAGDDAVRTEHLVKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTV
                     EDFWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDN
                     LWLDVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDAL
                     RLGRNNSLQYQLHKDTMVKQGSNVKSIYTL"
```

Header

Feature

Sequence

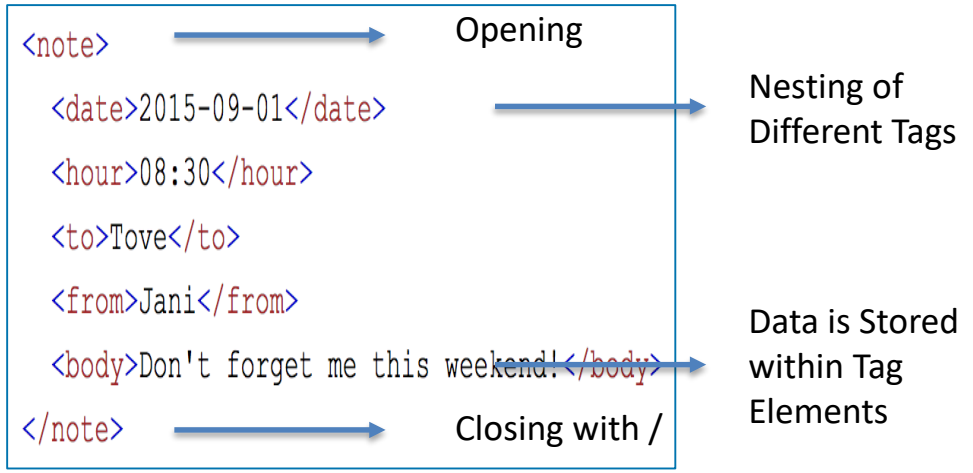
Flat model is the earliest, simplest data model. It lists all the data in a single table. In order to access or manipulate the data, the entire flat file must be read into memory (inefficient). **GenBank**, which stores info on biological sequences, is based on flat model.

Source: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Other Database Model

XML

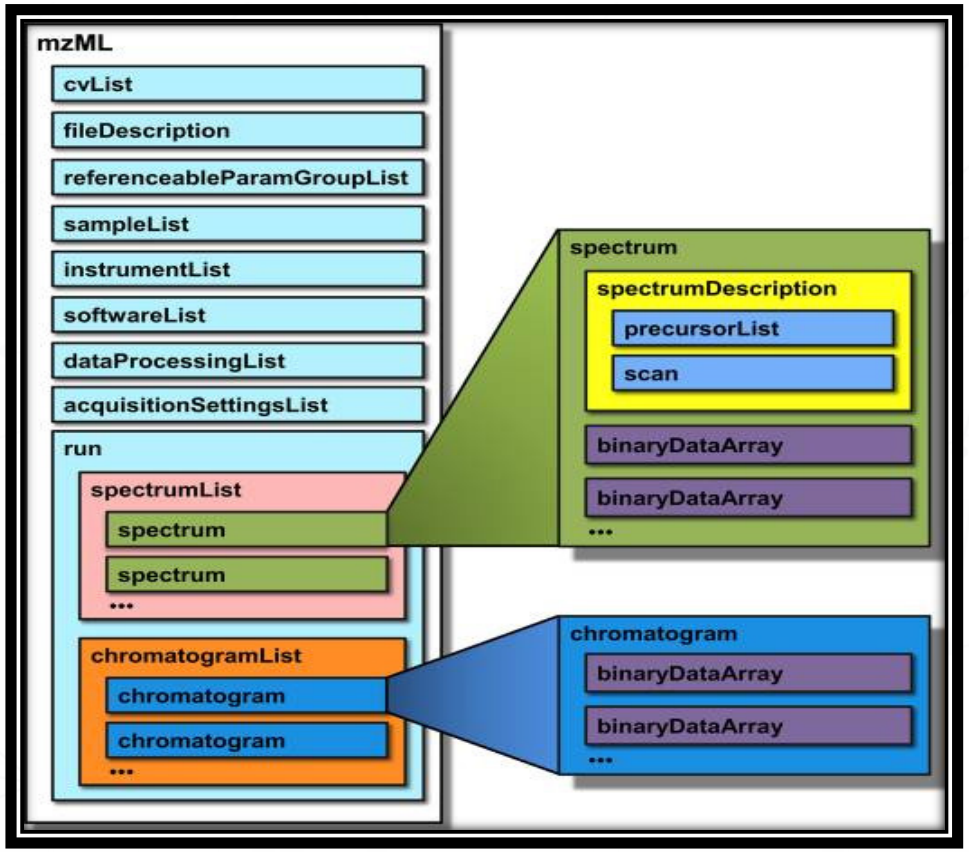
- XML stands for eXtensible Markup Language
- XML is a markup language (like HTML)
- Use for storing and transport data
- Self-descriptive; flexible and expandable



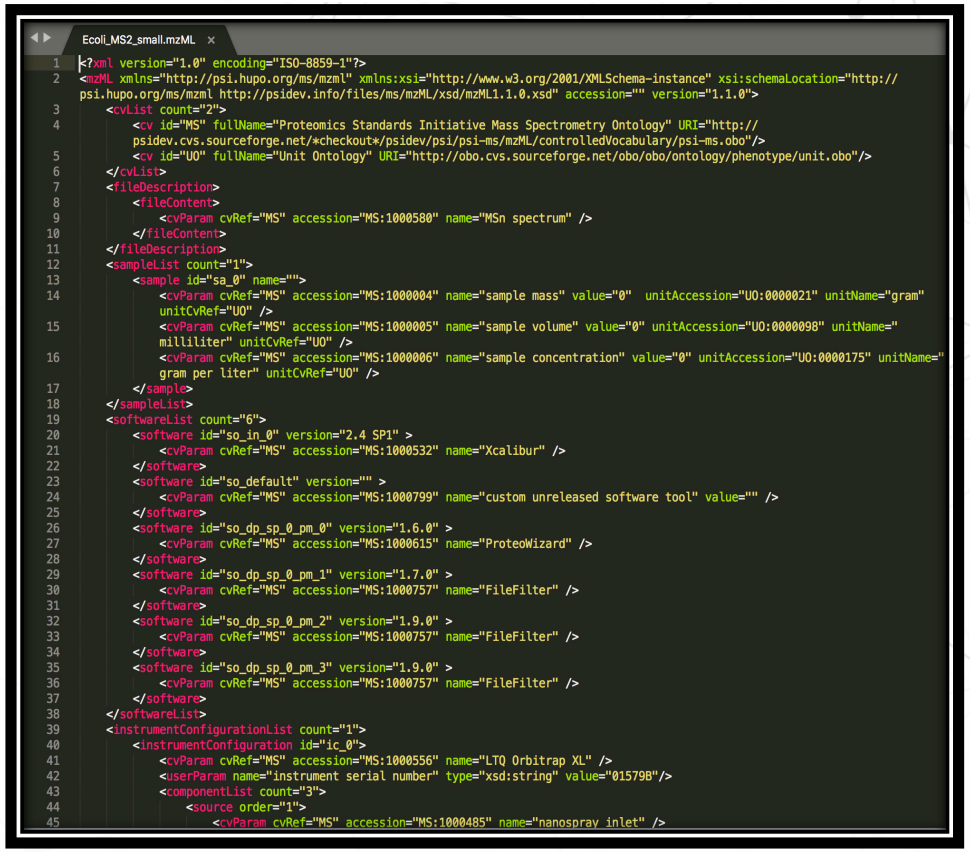
Other Database Model

Proteomics uses a lot of XML-based databases

Schematic representation key elements of the MZMLformat.



Example of actual MZML file



Source: Martens et al. mzML—a Community Standard for Mass Spectrometry Data. MCP 2011

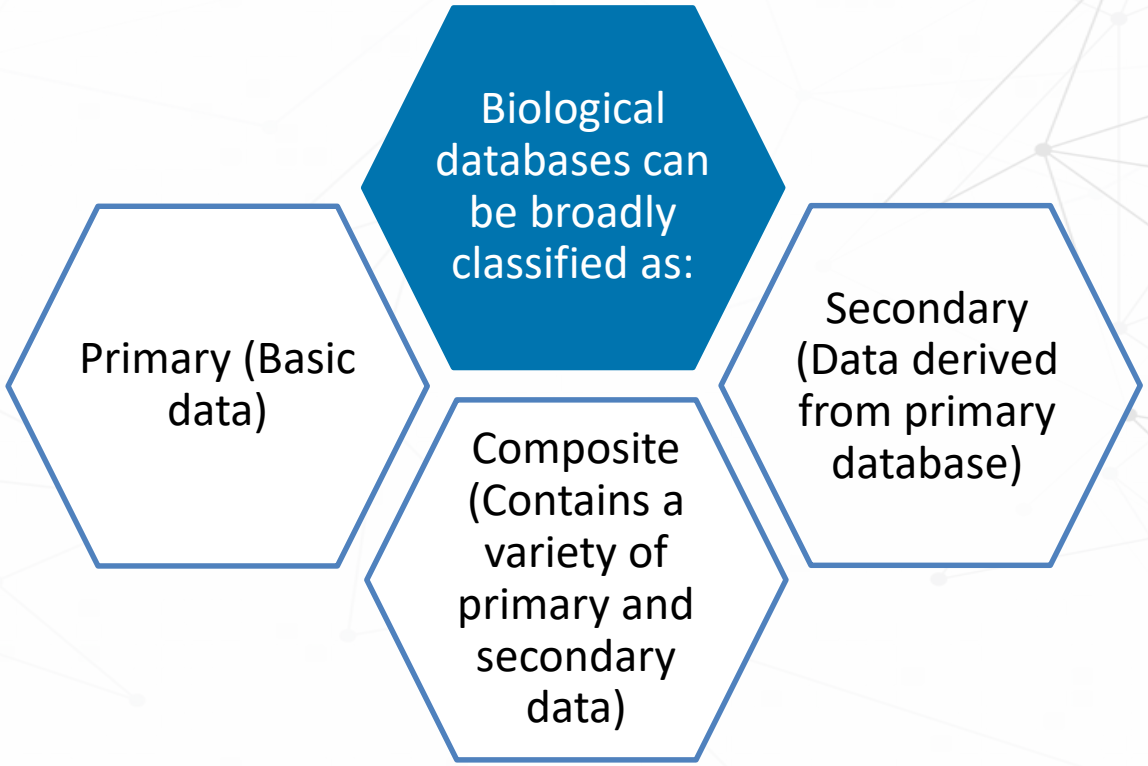
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Biological Databases

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Kinds of Biological Databases



Primary Databases

Primary databases contain information for sequence or structure only:

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures

Secondary Databases

- Secondary databases contain information derived from primary databases.
- Secondary databases store information such as conserved sequences, active site residues, and signature sequences:
  - SCOP and CATH for structural classification of proteins.
  - PROSITE for protein domains.

Composite Databases

- Composite databases contain a variety of primary databases, which eliminates the need to search each one separately.
- Each composite database has different search algorithms and data structures.
- Best known examples include NCBI (https://www.ncbi.nlm.nih.gov/) and ENSEMBL (https://www.ensembl.org/).

Kinds of Biological Databases

Biological databases can be classified by data type/ focus:

- Gene Sequence (e.g NCBI, EMBL)
- Protein Sequence (Uniprot, SwissProt)
- Genome Assembly (ENSEMBL, SGD, TAIR)
- Bibliographic (Pubmed and Web of Science)
- Disease (OMIM)
- Metabolic pathways (KEGG, WikiPathways, IPA)
- Experimental/Expression (GEO, PRIDE)

Invaluable Resource

- Every year, Nucleic Acids Research publishes an update on newly created biological databases, updates on existing ones, and also information on databases previously published in other journals.

- Access at https://academic.oup.com/nar/issue/45/D1.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Overview on Logic

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

What is logic?

Logic is a system for distinguishing between correct and incorrect **arguments**.

Arguments --- not people fighting. An argument is a chain of reasoning that leads towards a conclusion.

Logic includes a set of principles that when applied to arguments, allows us to demonstrate what is true/false.

$$\begin{array}{l} p \\ p \rightarrow q \\ \hline \therefore \quad q \end{array}$$

An example of an argument

1. All men are mortal.
2. Socrates is a man.
3. Therefore, Socrates is mortal.

In this case, the first 2 statements are premises, with the third being the conclusion.

An example of an argument

Let’s look at another example of an argument:

1. There’s a draught in the room.
2. A window is open.
3. Therefore, the draught is coming from the window.

Are the premises correct?

Is the conclusion necessarily correct?

What if the door is also open?

Premises

A premise --- what you already know, or assume to be true. Note that just because you assume it is true, does not mean it is necessarily so.

It amounts to an answer of ‘true’ or ‘false’ (A truth value).

Statements that are premises:

- All men are mortal. (true or false)
- There’s a draught in the room. (true or false)

Statements that are not premises:

- Break a leg
- What time is it?

The value of logic in scientific reasoning

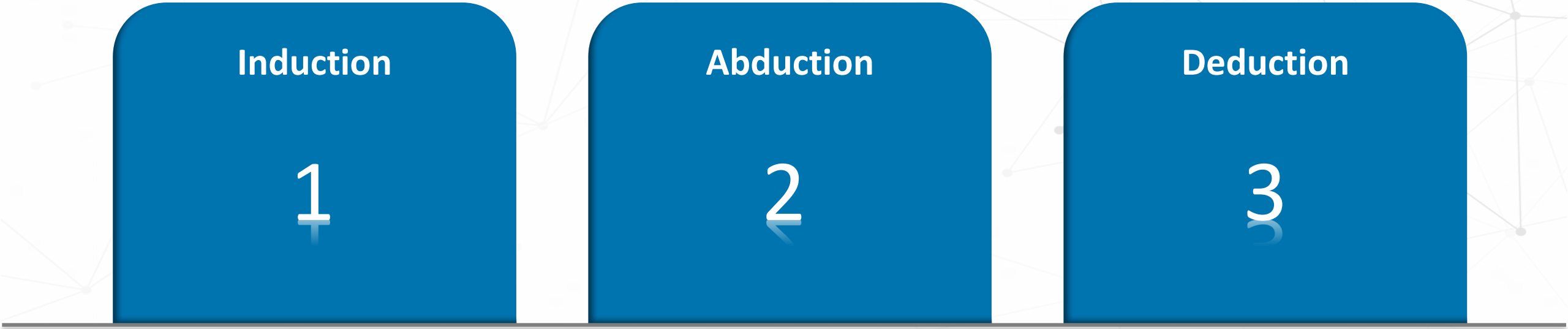
Applying logic is a way of developing and testing a hypothesis.

Using this way of thinking, applying logic assumes you already know something for sure.

And you use that knowledge to arrive at some further conclusions.

Three types of Reasoning

There are 3 major forms of logical reasoning.



Some forms of reasoning are stronger than others.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

What is Induction?

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Induction

In inductive reasoning, the conclusion need not follow exactly from the premises (i.e., the conclusion is breakable). There is always a chance that it may not. Let us consider the first example.

Induction

- Socrates is a man
- Socrates is mortal

$\Rightarrow$ All men are mortal, provided there is no counter example

In this example, we generalise to a law based on 2 attributes found on the entity, Socrates. Since Socrates is both a man and is mortal, we conclude that all men are mortal. This conclusion is true, for as long as no counter examples are observed.

Tip: Inductions often takes the form of bottom up reasoning. That is, reasoning from observations towards a law.

Induction

While it is easy to remember or think of induction as “bottom up” (creating laws from observation) while deduction is “top down” (making conclusions based on laws).

It is important to note that not all forms of inductive reasoning takes this form. Let us look at another example:

1. A bag contains 99 red balls and 1 black ball.
2. 100 people each drew one ball from the bag.
3. Sarah is one of those 100 people.
4. Therefore, Sarah probably drew a red ball.

Notice in this example, premises need not be limited to only 2 statements. On top of that, there is the aspect of probability involved as well.

In this case, Sarah likely drew a red ball with a probability of 99%. However, we cannot assure that this conclusion (that she got a red) is guaranteed.

Other Forms of Induction

- Premise 1: All **swans** we have seen are white. (True)
- Therefore all swans are white. (True/False?)
- Counter example: A black swan was observed in Africa.
- Note also that not all arguments need to have multiple premises.

- Premise 1: These beans are from this bag. (True)
- Premise 2: These beans are white. (True)
- Therefore All the beans from this bag are white. (True/False?)
- Counter example: Some green beans from this bag are discovered.

Can any signature predict breast cancer survival and its relation to inductive reasoning?

A paper by Venet et al claimed that after testing several non-breast-cancer gene signatures, any signature will do as well in predicting breast cancer survival.

- Premise 1: Social defeat signature is predictive of survival outcome. (true)
- Premise 2: Sneezing signature is predictive of survival outcome. (true)
- Conclusion: All signatures can predict survival outcome.

Two additional signatures from schizophrenia were evaluated on the breast cancer data.

Signature	Size	p_val (P)
Schwarz	147	0.45
Hess	89	0.44

Neither were found to be significant for predicting breast cancer survival. So how do we evaluate this outcome?

Can any signature predict breast cancer survival and its relation to inductive reasoning

Would you conclude that any signature from any other disease can cross-predict breast cancer survivability?

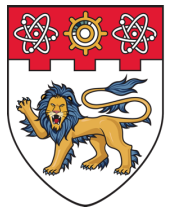
- Premise 1: Social defeat signature is predictive of survival outcome. (True)
- Premise 2: Sneezing signature is predictive of survival outcome. (True)
- Conclusion: All signatures can predict survival outcome.
- The conclusion reached is incorrect.
- Counter example: The schizophrenia signature cannot predict survival outcome.

Induction

– Socrates is a man

– Socrates is mortal

⇒All men are mortal, provided there is no counter example



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

What is Deduction?

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Deduction

- Deduction is the strongest form of reasoning --- its conclusion must follow from its premises.
- If the premises are true, then the conclusion must also be true.

Deduction:

1. All men are mortal.
2. Socrates is a man.

$\Rightarrow$ Socrates is mortal.

- 1 and 2 are true, therefore the conclusion must be true.
- Deductive reasoning has very strict standards. Not meeting any of these can result in failure.
- A tip: Deductions often takes the form of top down reasoning. That is, reasoning from a law towards an observation.

Deduction

- How can deduction fail?
- **One of the premises could be false.**

1. All dogs are brown.
2. Missie is a dog.

$\Rightarrow$ Therefore, Missie is brown.

- Premise 1 is false: not all dogs are brown.
- Even though the statements above take the form of a deductive argument, it fails because one of the premises is false (“rubbish in, rubbish out”).

Deduction

- How can deduction fail?
- **The conclusion does not follow from the premises.**

1. All tennis balls are round.
2. The earth is round.

$\Rightarrow$ Therefore, the Earth is a tennis ball.

- This argument fails because of faulty logic.
- While it is true that all tennis balls are round, but so too are many other things!

Deduction

Generalising deduction: You have seen the deductive argument in this form

1. All men are mortal.
2. Socrates is a man.
3. Therefore, Socrates is mortal.

We may express the form of this argument symbolically as follows:

- All As are Bs.
- C is an A.
- Therefore, C is B.

To be able to conclude something deductively is relatively rare. It requires having laws with absolute truth values that are not easily broken (e.g. “All men are mortal” or “The sun will rise from the east”). Compare this to “laws” like “All swans are white”.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

What is Abduction?

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Abduction

- Abduction seeks to find the simplest and most likely explanation/conclusion given a set of premises or observations.
- Like induction, it is also breakable in the sense that the conclusion can be wrong, even if the premises are true.

Abduction

- All men are mortal
- Socrates is mortal

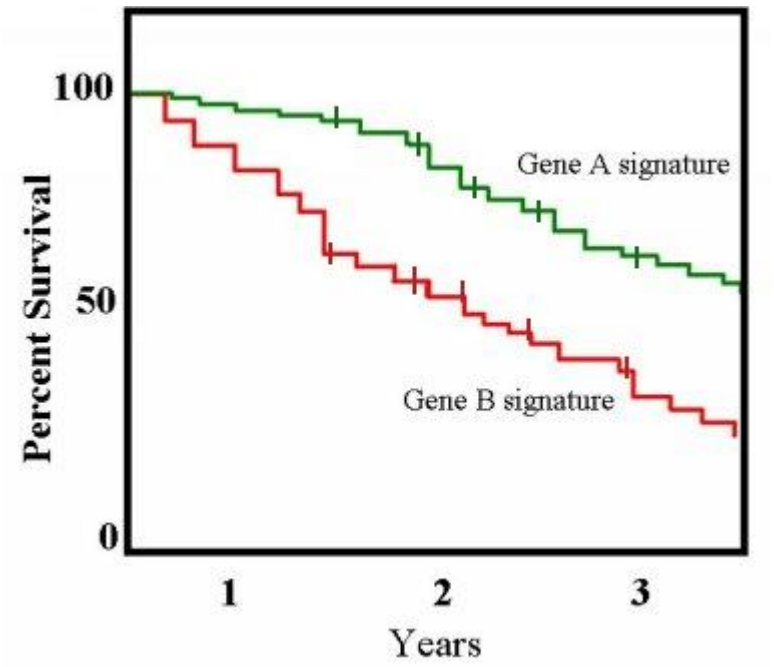
$\Rightarrow$Socrates is a man, provided there is no other explanation of Socrates' mortality

- Tip: Abduction involves generalisation via shared attributes.

Abduction

- Using similar reasoning, let's say that the fire truck is red.
- We observe that an apple is also red.
- Therefore, by abduction, the fire truck is an apple.
- Obviously, there are other explanations for the fire truck's redness. It was painted to be red.
- Use abductive reasoning with care. It is valid, only if there are **no other explanations** that can explain the observation.

Survival Curve and Abductive Reasoning



By Deanne Taylor (made and original to submitter) - The original description page was here. All following user names refer to en.wikipedia., Public Domain, https://commons.wikimedia.org/w/index.php?curid=644003

- Premise 1: Multi-gene signature for breast cancer is predictive of survival outcome. (true)
- Premise 2: Multigene for social defeat is predictive of survival outcome. (true)
- Therefore, social defeat signatures are breast cancer signatures. (false)
- The conclusion reached is **incorrect. This is because other explanations exists**. And that has to do with the presence of **confounders** and widespread **genome instability effects** observed in cancer.

Abduction

- All men are mortal.
- Socrates is mortal.

$\Rightarrow$Socrates is a man, provided there is no other explanation of Socrates' mortality

Three Types of Reasoning

While deduction is the strongest form of reasoning, induction and abduction can also be powerful when the additional criteria is fulfilled.

1

Induction

Socrates is a man.

Socrates is mortal.

All men are mortal (provided there is no counter example).

2

Abduction

All men are mortal.

Socrates is mortal.

Socrates is a man (provided there is no other explanation of Socrates' mortality).

3

Deduction

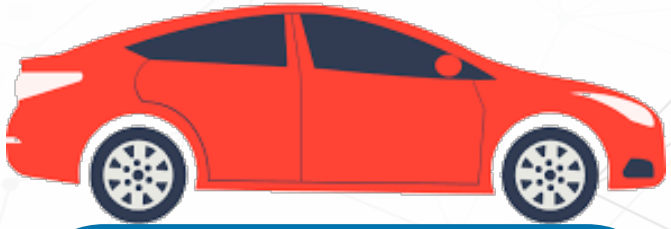
All men are mortal.

Socrates is a man.

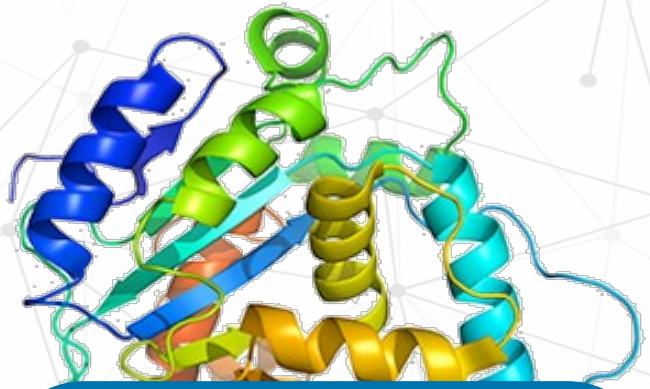
Socrates is mortal.

Which of the following are examples of each reasoning type?

Gene A performs function X; Gene B is sequentially similar to Gene A. Therefore, Gene B also performs function X.



An apple is red, a car is red, so therefore a car is an apple.

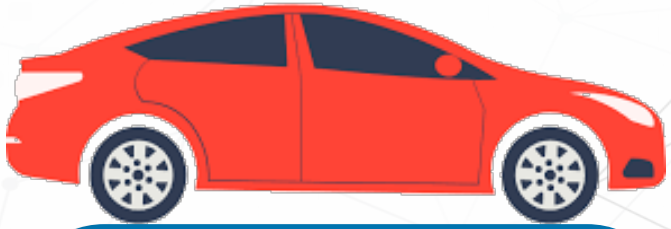


A class of proteins, C, performs function X. Protein Z is a member of C, so Protein Z must therefore perform function X.

Which of the following are examples of each reasoning type?

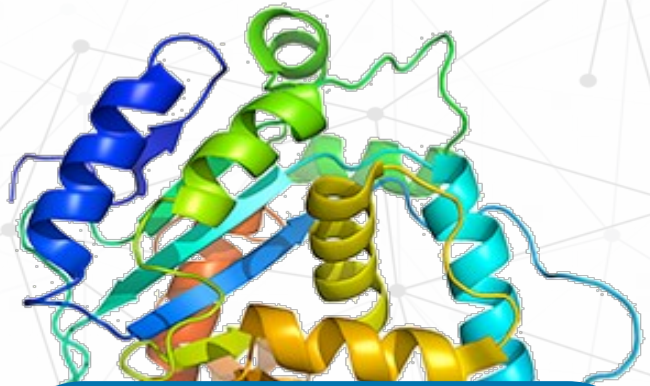
Induction

Gene A performs function X; Gene B is sequentially similar to Gene A. Therefore, Gene B also performs function X.



Abduction

An apple is red, a car is red, so therefore a car is an apple.



Deduction

A class of proteins, C, performs function X. Protein Z is a member of C, so Protein Z must therefore perform function X.

Practice

- Deduction:
  - All the beans from this bag are white.
  - These beans are from this bag:
  - Therefore These beans are white.

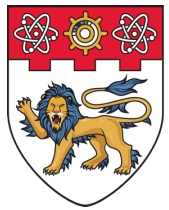
Is the conclusion always true when the premises are true?

- Induction:
  - These beans are from this bag.
  - These beans are white.
  - Therefore All the beans from this bag are white.

Is the conclusion always true when the premises are true?

- Abduction:
  - All the beans from this bag are white.
  - These beans are white.
  - Therefore These beans are from this bag.

Is the conclusion always true when the premises are true?



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

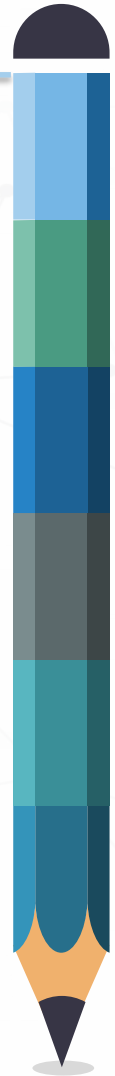
BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. A database is an organised collection of data, stored and accessed electronically. Database designers typically organise the data to model aspects of reality in a way that supports practical usage.

2. Database normalisation is the process of restructuring a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.



3. The Relational Model (RM) for database management is an approach to managing data using a tabular structure, with entries recorded as rows, and defined uniquely by a primary key.

4. A flat file database is a database that stores data in a plain text file. Each line of the text file holds one field.

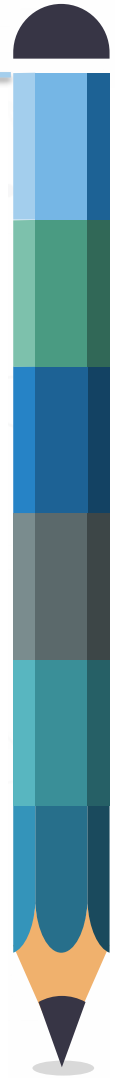
5. An XML database stores and represents data as a series of nested tags with appropriate opening and closing statements.

Key Takeaways from this Topic

6. With the advent of high-performance computational platforms, biological databases have become more important than ever in providing the infrastructure needed for biological research, from data preparation to data extraction.

7. Logic is a system for distinguishing between correct and incorrect arguments. It includes a set of principles that when applied to arguments, allows us to demonstrate what is true/false.

8. There are 3 major forms of logical reasoning, induction, abduction and deduction.



9. In inductive reasoning, the conclusion need not follow exactly from the premises (i.e., the conclusion is breakable). There is always a chance that it may not.

10. Deduction is the strongest form of reasoning --- its conclusion must follow from its premises. If the premises are true, then the conclusion must also be true.

11. Abduction seeks to find the simplest and most likely explanation/conclusion given a set of premises or observations. Like induction, it is also breakable in the sense that the conclusion can be wrong, even if the premises are true.