Structured Query Language (20m)

The following should require no explaination.

```
  SELECT *
    FROM Invoice
   WHERE BillingCity = 'London'
ORDER BY Total DESC
   LIMIT 2
```

One table only

When you are interested in total amount of bills from each city.

```
  SELECT BillingCity, SUM(Total) AS CityTotal
    FROM Invoice
GROUP BY BillingCity
```

Use of `CASE WHEN ... THEN ... ELSE ... END`

```
SELECT InvoiceId,
       CustomerId,
       InvoiceDate,
       Total,
       CASE WHEN Total >= 10 THEN "High"
            WHEN Total >= 5 THEN "Medium"
            ELSE "Low"
            END AS RevenueClass
  FROM Invoice
```

Use of `DATETIME`. `CURRENT_TIMESTAMP` is current time.

```
  SELECT InvoiceId, CustomerId, InvoiceDate from Q030HighRevenue
   WHERE InvoiceDate >= DATETIME("2013-01-01 00:00:00")
ORDER BY InvoiceDate DESC
```

Use of `IFNULL`. The column 'CompanyNew' contains is the column 'Company' but the missing values replaced with 'Missing Name'.

```
SELECT *,
       IFNULL(Company, "Missing Name") AS CompanyNew
  FROM Customer
```

Multiple tables

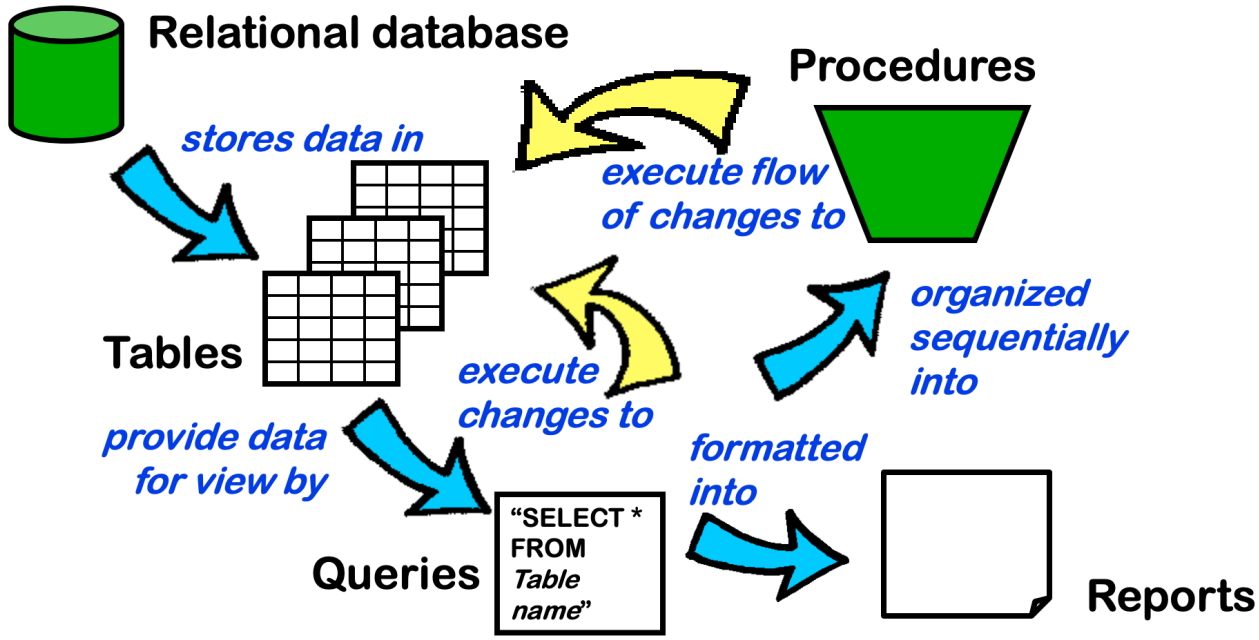
This filters all the combination that fulfils the where condition. (This may be inefficient?)

```
SELECT *
  FROM Album,
       Artist
 WHERE Album.ArtistId = Artist.ArtistId
```

`INNER JOIN` is used together with `ON` to merge tables. Entries in 'Album' with null values in `ArtistId` is excluded.

```
SELECT *
  FROM Album
       INNER JOIN
       Artist ON Album.ArtistId = Artist.ArtistId
```

There are two types of join `INNER JOIN` and `LEFT JOIN`. `RIGHT JOIN` and `FULL OUTER JOIN` not supported.



Try to understand the following queries.

Which employee serves the most number of customers?

```
  SELECT ee.FirstName,
         ee.LastName,
         COUNT(cs.CustomerId) AS CustomerNo
    FROM Employee as ee
         INNER JOIN
         Customer as cs
      ON ee.EmployeeId = cs.SupportRepId
GROUP BY ee.EmployeeId
ORDER BY CustomerNo DESC
```

Which artist has the greatest sales?

```
SELECT Artist.Name,
       SUM(InvoiceLine.UnitPrice * InvoiceLine.Quantity) AS ArtistSales
  FROM (
           (
               InvoiceLine
               INNER JOIN
               Track ON InvoiceLine.TrackId = Track.TrackId
           )
           INNER JOIN
           Album
        ON Track.AlbumId = Album.AlbumId
       )
       INNER JOIN
       Artist
    ON Album.ArtistId = Artist.ArtistId
 GROUP BY Artist.Name
 ORDER BY ArtistSales DESC;
```

Common errors

- Forgetting a comma in the middle of an array.
- Please create a new table first, if you want to modify data.

Regression with R (10-20m)

Method	**Linear Regression**
Target	Number
Predicts	Number
Model	$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon_i$
Loss	Mean square error
Quality of fit	R-square Adjusted R-square AIC
Comments	Choose only the statistically significant variables This cannot predict binary objectives

Method	**Logistic Regression**
Target	Binary
Predicts	Probability
Model	$P(y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon_i)}}$
Loss	$LL(\beta)$ $= \sum_{i=1}^{n} \sum_{k=1}^{2} y_{ik} \log(P(y_{ik} = 1))$ $= \sum_{i=1}^{n} \sum_{k=1}^{2} y_{ik} \log\left(\frac{e^{\beta' x_{ik}}}{\sum_{l=1}^{k} e^{\beta' x_{il}}} \right)$
Quality of fit	$AIC = -2LL(\hat{\beta}) + 2(p + 1)$ Confusion matrix AUC-ROC
Comment	

Time series analysis (15m)

Moving average

Pretty obvious, just take the average of last n elements.

Exponential Smoothing (Holt 1) with $\alpha = 0.2$

Calculations	Forecast	Actual	Error
(based on previous row)	-	600	-
600	600	580	20
$(1-\alpha)\cdot 600 + \alpha \cdot 580$	596	620	-24
$(1-\alpha)\cdot 596 + \alpha \cdot 620$	600.8	590	10.8
$(1-\alpha)\cdot 600.8 + \alpha \cdot 590$	598.64	610	-11.36
$(1-\alpha)\cdot 598.64 + \alpha \cdot 610$	600.912	570	30.912
	F	A	
$(1-\alpha)F + \alpha X$			

Double Exponential Smoothing (Holt 2) with $\alpha = 0.2$ and $\beta = 0.1$

A	B	Forecast $\hat{x}_n$	Actual x_n
-	-	-	700
700	50 (asmp, default 0)	750	760
$(1-\alpha)\cdot 750$ $+ \alpha \cdot 760$	$(1-\beta)\cdot 50$ $+ \beta \cdot (a_n - a_{n-1})$	802.2	800
A_2	B_2	F	X
$A_3 = (1-\alpha)F$ $+ \alpha X$	$B_3 = (1-\beta)B_2$ $+ \beta(A_2 - A_3)$	$A_3 + B_3$	-

```
# Holt 1
plot(HoltWinters(AirPassengers, gamma=FALSE, beta=FALSE))
# Holt 2
plot(HoltWinters(AirPassengers, gamma=FALSE, beta=TRUE))
```

Process Analysis (10-20m)

Performance measures

Throughput

- Output **rate** of a process or a stage of a process
- Example: average number of patients served per day = 60

Flow time

- Average **time spent in system** by unit single output
- Often includes waiting time.
- Example: average time spent in clinic = 2 days

Work-in-process (WIP)

- Average number of units in systems over a time interval.
- Example: average number of patients in a clinic over time = 120

Little's Law Work-in-process = Throughput * Flow Time

Calculation of throughput

Process	Calculation
Single-stage	Throughput = Output / Time
Multi-stage (simple)	The lowest throughput among all the stages.
Stage with Parallel Activities	The overall throughput is the **minimum** throughput among all the parallel activities. (i.e. you need to **wait** for the slowest person to complete his/her part).
Stage with Multiple Paths	The overall throughput is the harmonic weighted sum. $1 \Big/ \Sigma_i^m \frac{p_i}{Th_i}$

Throughout analysis

The **bottleneck** is the process stage with the lowest throughput rate.

- Downstream operations will be starved
- Upstream operations will be blocked

Consequence of bottleneck
It slows down the whole process, limits the process capacity, leads to low utilization at other stages, leads to job waiting, requires extra inventory/buffer/stock to place waiting jobs

Parallel Coordinates (10-15m)

Standard operating procedures

- "high Y is associated with high/low X1, X2, X3"
- "low Y is associated with low/high X1, X2, X3"
- "the converse is not true: low Y have both high and low X2"
- "with the exception of the first year of the study..."

Common talking points

- answer the question "with a focus on understanding what may cause enrollments to increase"
- causation is not correlation
- consider causation in the other direction
- "the fact that many of the variables are correlated with time makes drawing conclusions difficult" - inflation, income growth
- irrelevant factors should be ignored
- "the data are consistent with many different theories"
- sometimes two variables when combined provides a strong outcome

GIS (0-5m)

Geographical decisions - point, line, area/polygon. Polygon layer should be at the bottom, so that it will not cover the line or point layer.

X is longitude, Y is latitude

Layers and Features

- A **feature** is described by a geometry (point, line segment, polygon)
- Geometries are described using latitude and longitude
- Each feature is referenced by a unique index, the ‘feature id’
- Each feature has a row in a database table, indexed by the feature id
- The fields (columns) of the table are called **attributes**
- A **layer** is a collection of features with the same geometry type (point, line, or polygon)
    - Every feature on a layer has the same set of attributes

Miscellaneous

Correlation

Pearson's Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \leq r \leq 1$$

Decision Tree

Which tree to select - one that classifies at least one category perfectly.

```
credit_data <- read.csv("wk6a-credit.csv")
plot(creditdata[,c(5,9:14)])  # scatterplot matrix
library(tree)
tree.credit = tree(Status~., data = credit_data)
plot(tree.credit); text(tree.credit, pretty=0);
```