Master of Technology in Knowledge Engineering

Identifying & Planning Data Mining Projects

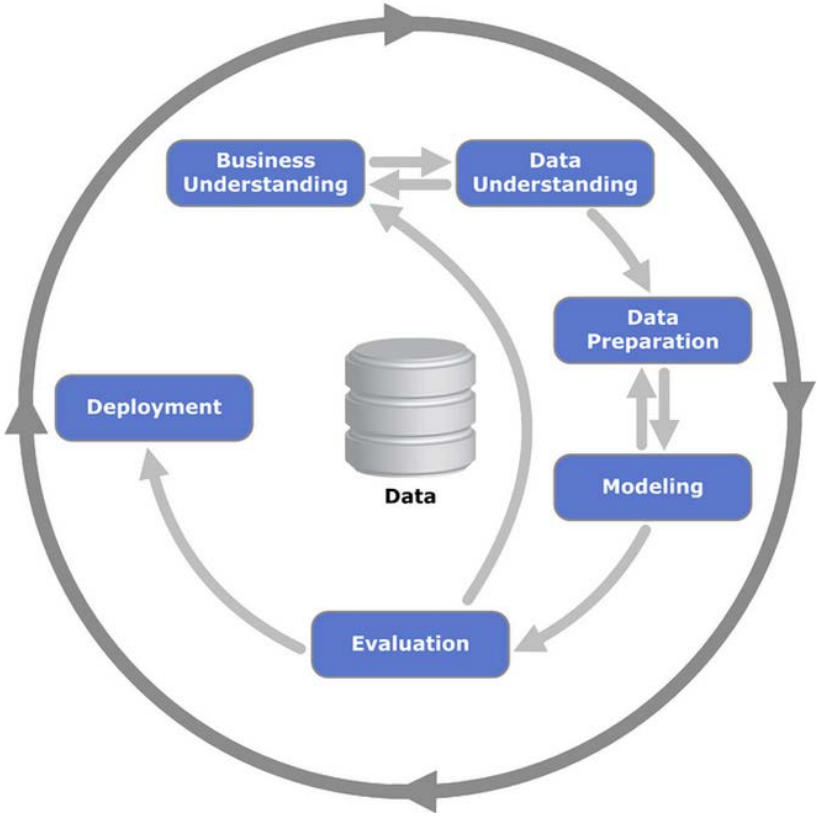
Dr. Zhu Fangming
Institute of Systems Science,
National University of Singapore.
E-mail: isszfm@nus.edu.sg

Agenda for Day1

- Data Mining – What and Why?
- Applications Overview
- Identifying and planning data mining projects
- Data Mining Planning Workshop
- Data Mining Tools Overview/Demo

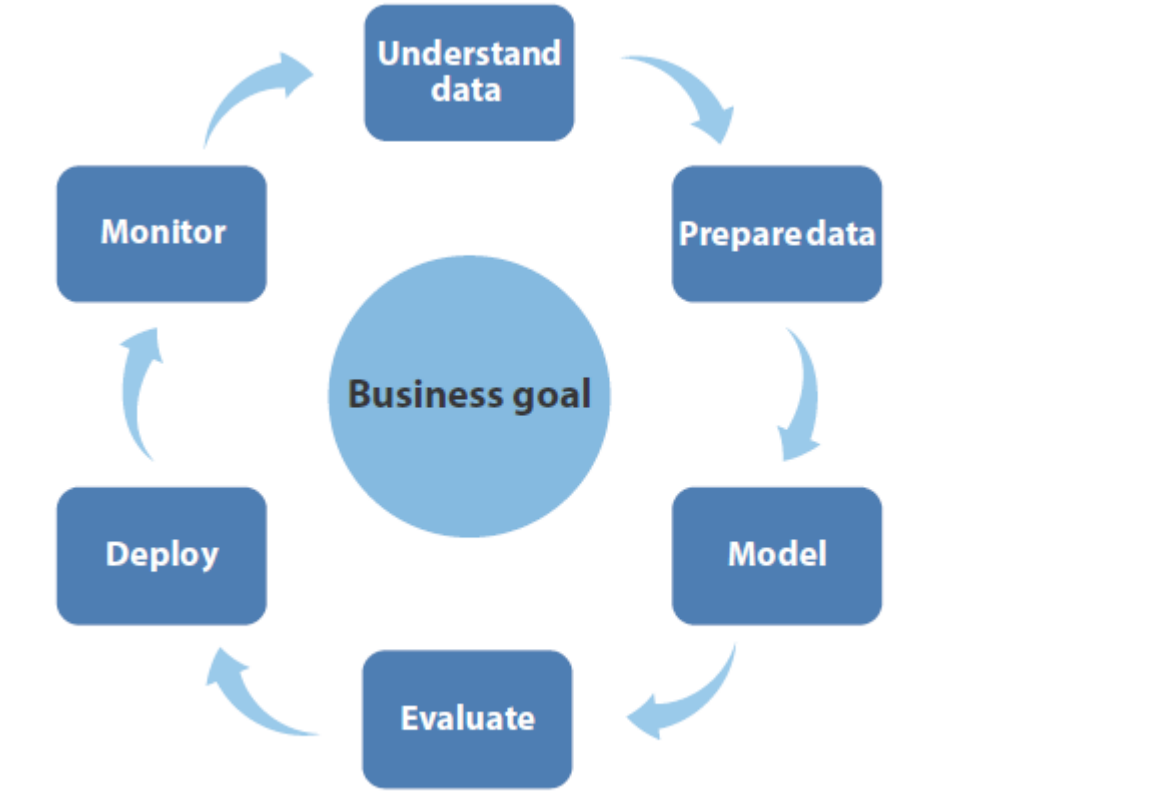
The Data Mining Process

- Cross Industry Standard Process for Data Mining
  - v1.0 published 1999 (DaimlerChrysler, NCR, SPSS), 300 orgs contributed



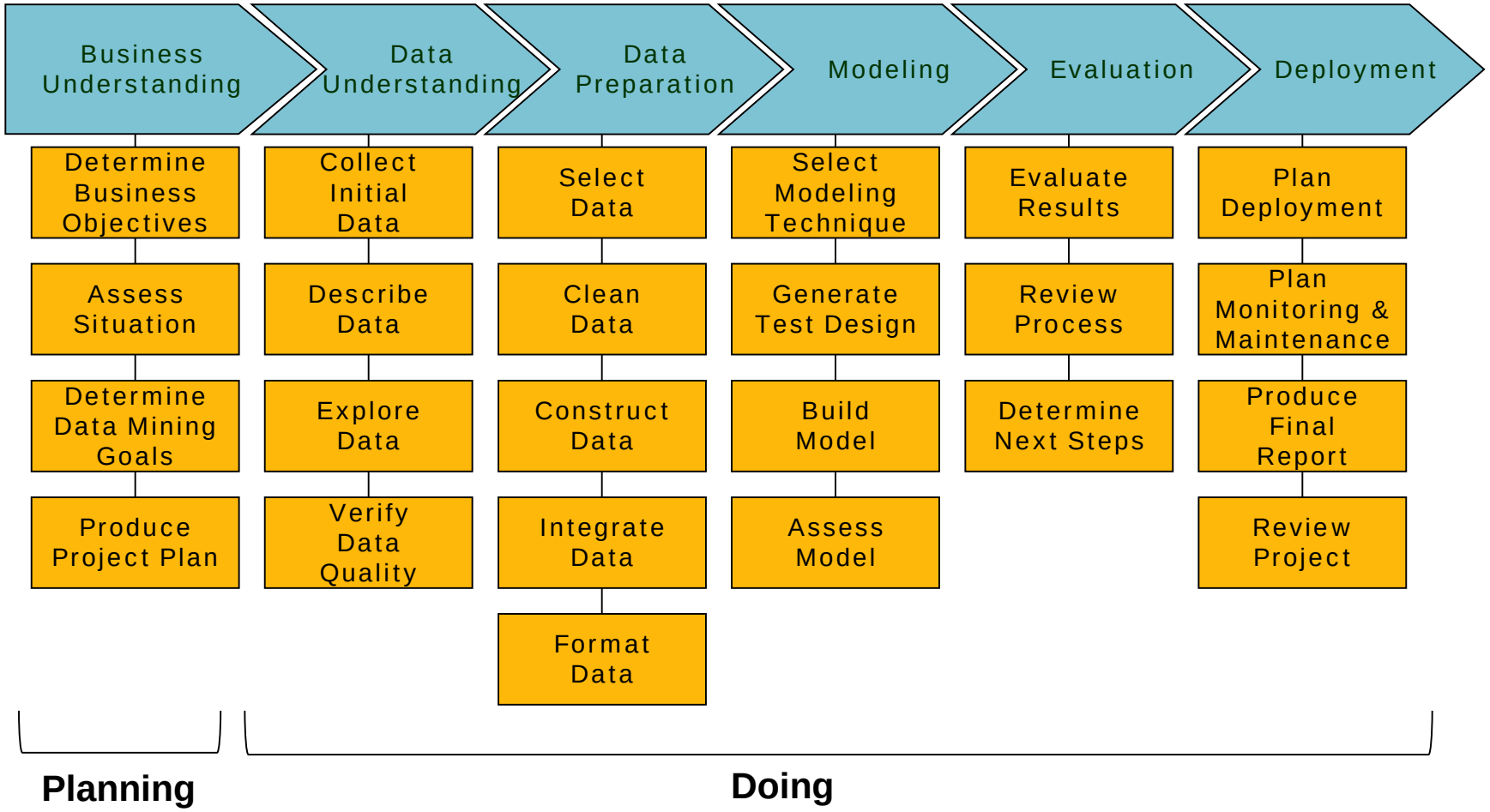
Forrester Research

- Predictive Analytics Process (2013)



Source: Forrester Research, Inc.

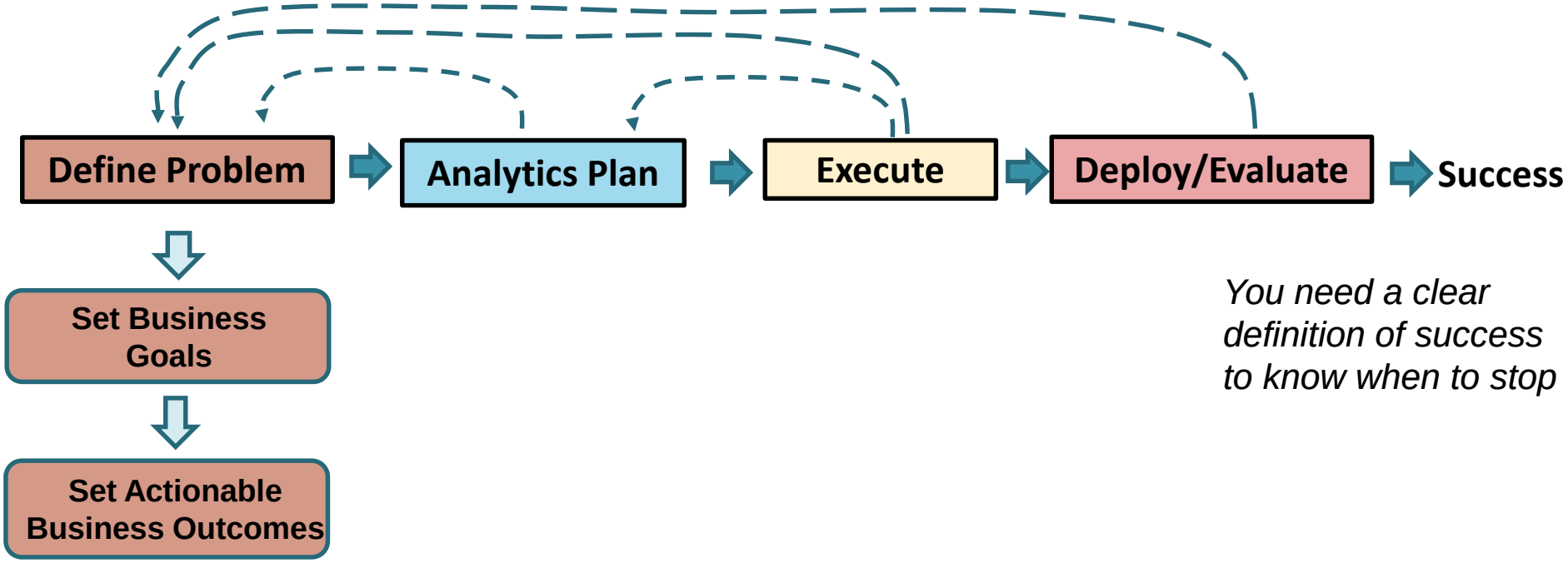
CRISP-DM: Drilling down....



Lots of detail for the doing part but less for the planning part!

Birds Eye View: The Data Mining Process

Data Mining is not a linear process. The results of each step may require you to go back to previous steps.



You need a clear definition of success to know when to stop

You need to be clear about what you want to achieve

Setting Business Goals

- Identifying the primary business objective is critical to managing a data mining project successfully, and ensuring that the project does not result in producing the right answers to the wrong questions.
- Usually a two way process between the chief data scientist(s) and the business domain experts
  - The Data Scientist usually needs ***some*** domain knowledge for this conversation to succeed

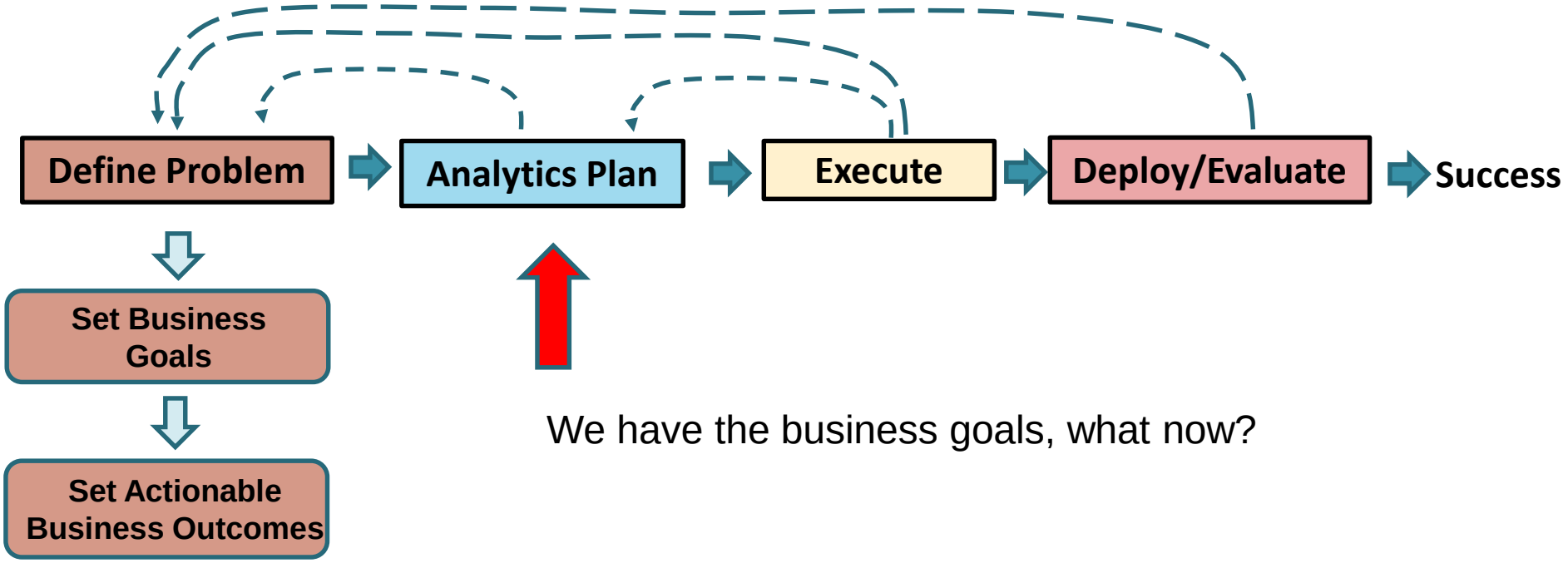
Business Goal Guidelines

1. There should be no mention of analytics methods
2. There must be an actionable outcome
3. Must be able to measure success (quantifiable metrics)

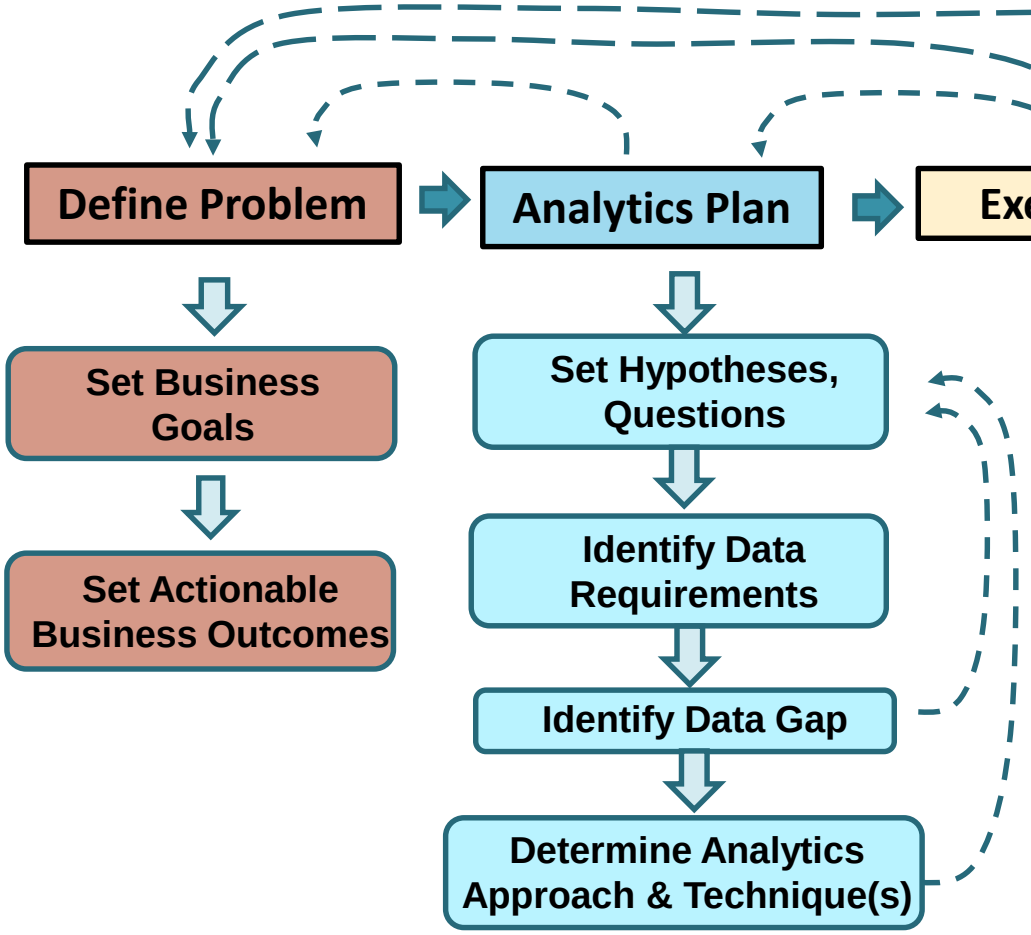
Setting Business Goals

- Possible examples are …
  - Improve the response rate for a direct marketing campaign
  - Increase the average order size
  - Determine what drives customer acquisition
  - Forecast the size of the customer base in the future
  - Retain profitable customers
  - Recommend the next, best product for existing customers
  - Choose the right message for the right groups of customers

The Data Mining Process



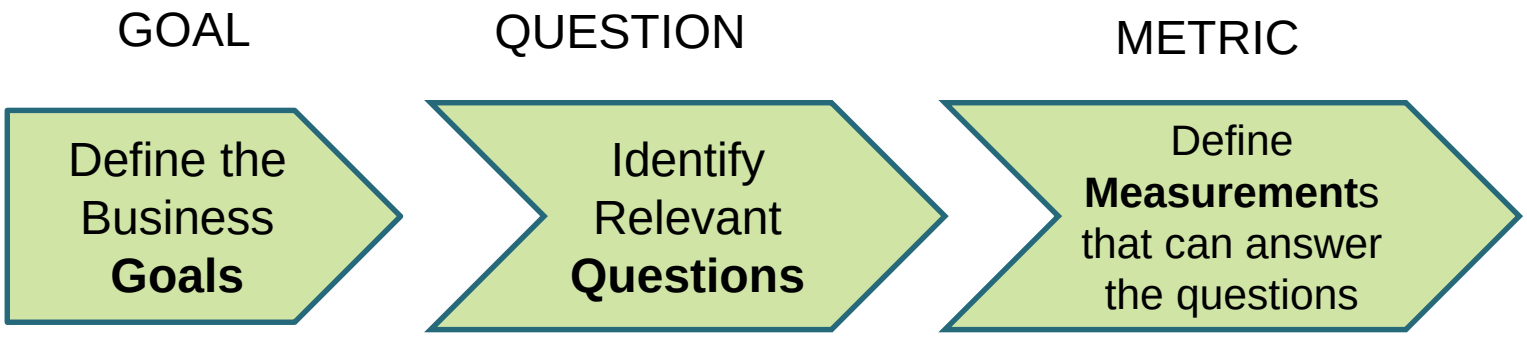
Birds Eye View: The Data Mining Process



Planning can be done in many ways, but a common approach is similar to GQM method (goal->question->metric) widespread in s/w industry for establishing metrics

The GQM Approach

- Originally developed to help an organisation identify appropriate software metrics*



- For data mining it can help determine the measurements and tests to make and the data required to make these measurements
- Try to identify all of the known business issues related to addressing your strategic objective to ensure that your data mining project is as business-focused as possible.

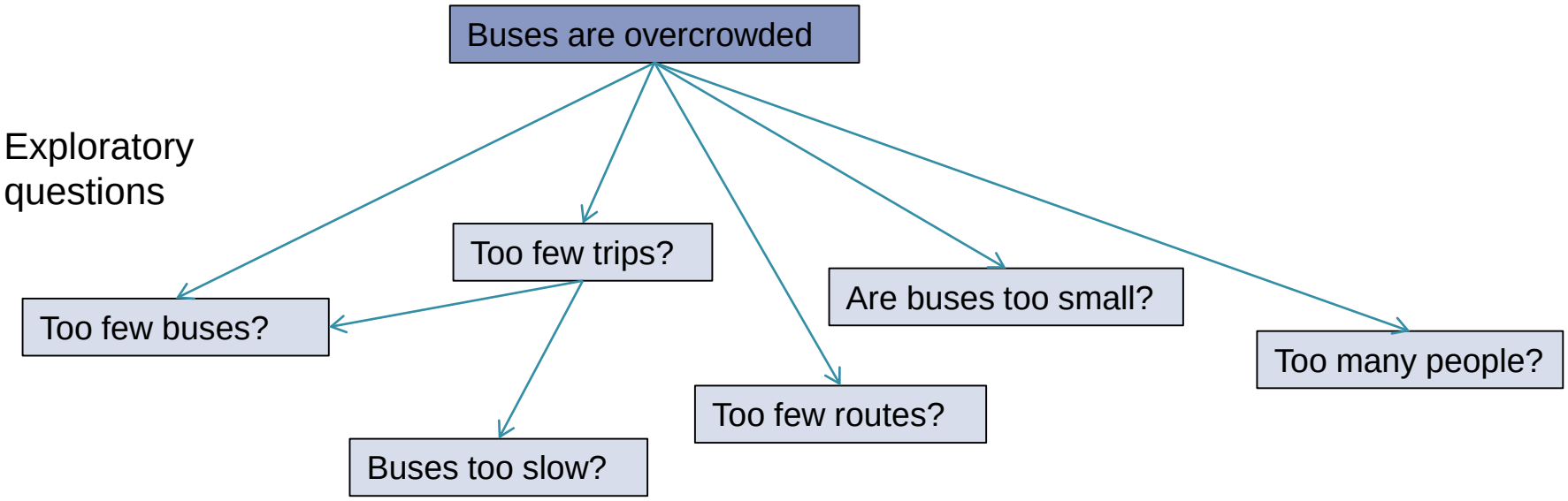
[1] Victor Basili, *Software Modeling and Measurement: The Goal/Question/Metric Paradigm*, CS-TR-2956, University of Maryland, 1992

Scenario: Public Transport Optimisation

- Public transport is currently a hot topic in Singapore, increasing population is driving the need for optimisation and innovation
  - One problem is **Bus Overcrowding** ~ what are the root causes?
  - Another problem is how to ensure **Bus Lanes** are effective. What are the characteristics of a successful bus lane?

Example: Bus Overcrowding

- **Business Problem:** Buses are overcrowded
- **Business Goal :** Recommend a strategy to reduce overcrowding
- **Success Criteria:** Reduce overcrowding by 25%
- **Initial Questions:** What are the main causes of overcrowding?
- **Measurements:** The data needed to answer these questions



Example: Bus Lane Effectiveness

- Define the Business Problem
  - Will my planned bus lane(s) be effective?
  - If not effective then why not? Can it be fixed & how?
- Define the Business Outcome
  - E.g. make a go/no-go decision on a planned new bus lane
  - Give confidence level & justification for the decision
- Set Quantitative Evaluation Criteria
  - How do we measure "being effective" ?
    - Increase in passengers on buses using the bus lane?
    - Increase in bus punctuality (less clumping)?
    - Shortened bus journey times?
    - Reduced traffic along the route? (and further afield?)
    - "Gains" should last and not drop off (too much) over time
    - All of the above (e.g. create a weighted success function)

How much increase/decrease constitutes success?

Bus Lanes: Setting Initial Questions

- How effective are existing bus lanes?
- What distinguishes effective from ineffective lanes?

Sub-Questions for Each Bus Lane	Data required to answer
Is there an increase in bus riders?	

Bus Lanes: Setting Initial Questions

- How effective are existing bus lanes?
- What distinguishes effective from ineffective lanes?

Sub-Questions	Data required to answer (suggestions)
Is it a physical property of the bus lane?	• length (kms)

Identifying Data Requirements

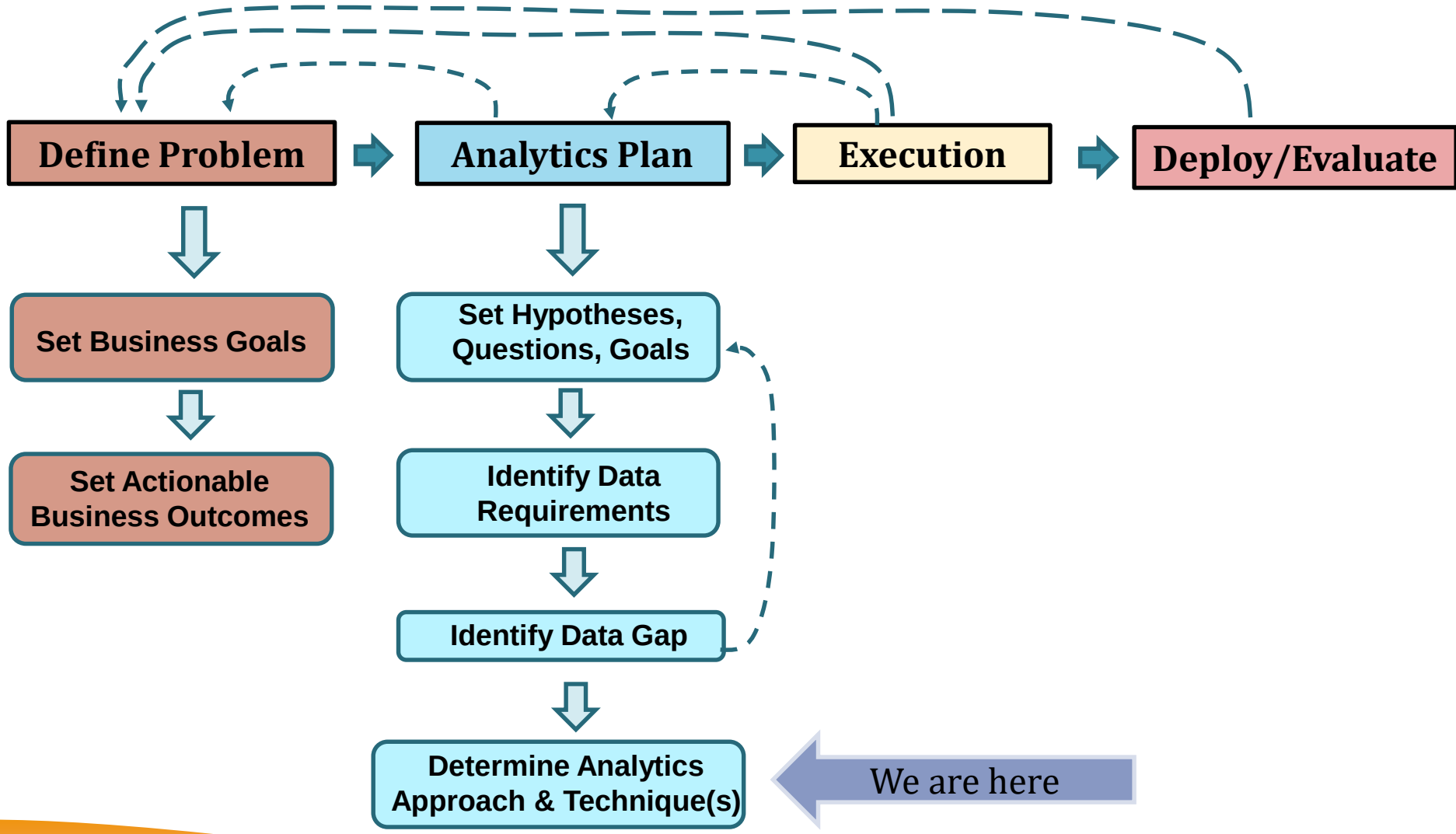
- General questions
  - What data is available?
  - What must the data contain?
  - What would be useful? (whether available or not)
  - What is the right level of granularity?
  - How much data is needed?
  - How much history is required?
    how far back in time should the data go?

- What is the base rate for comparison?
  - What would the above measures have been if the bus lane was not implemented?
  - Can we use the numbers before the bus lane introduction?
  - What was the traffic growth trend? – any increase may have happened anyway

Identify any Data Gap

- Consolidate all of your data requirements
- Determine what (if any) essential data is missing
- How to bridge the gap?
  - Put in place mechanisms to start collecting the missing data (delay the analytics)
  - Get the data from elsewhere (e.g. 3rd party, the web)
  - Innovate to get missing data or data you think may be useful
    (e.g. run in-bus surveys via smartphone apps)

Recap: The Data Mining Process



Identifying an Analytics Approach

- The best approach depends on the ***Business Goals*** and the ***Data Available***
- ***Another Example:*** Company X wishes to increase sales to existing customers. How can we use analytics?

(1) Examine **past purchase data**, identify big spenders and target them with promotions.

(2) Examine customer profiles (demographics, interests etc.) – identify the big spenders, then target low spenders who ***"look like"*** the big spenders

(3) Examine the records of **past marketing campaigns,** combine this with **customer profile data** and **past purchase data** to build a ***response model*** to predict which customers will respond best to new campaigns

Selecting an Analytics Approach

Factor Analysis

Data Visualisation

Times-Series Analysis

Clustering

Bagging

Market Basket Analysis

K-Means

CHAID

Bayesian networks

Outlier detection

Random Forests

Nearest Neighbour Methods

Prediction Models

Segmentation

Association Finding

Collaborative Filtering

Sequence Mining

C5.0

Pattern Matching

Ensemble Methods

Neural Networks

Network Analysis

Support Vector Machines

RFM Analysis

Cox regression

Link analysis

Logistic regression

Supervised Learning

Text Mining

Multi-Dimensional Scaling

Unsupervised learning

Semi-supervised learning

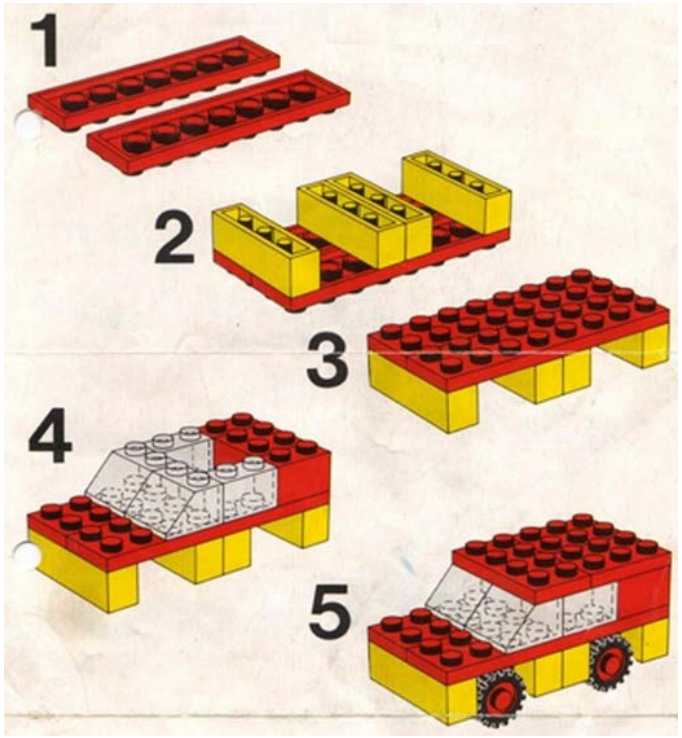
Boosting

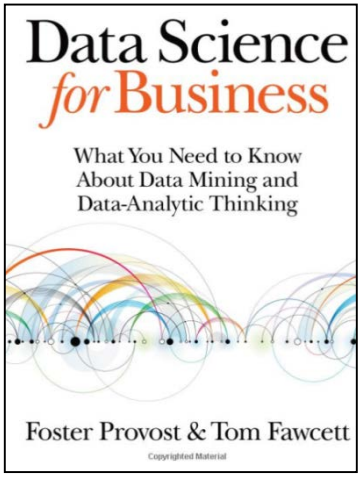
Principle Component Analysis

Selecting an Analytics Approach

"A critical skill in data science is the ability to decompose a data analytics problem into pieces such that each piece matches a known task for which tools are available"

"Recognising familiar problems and their solutions avoids wasting time and resources reinventing the wheel"





Selecting an Analytics Approach

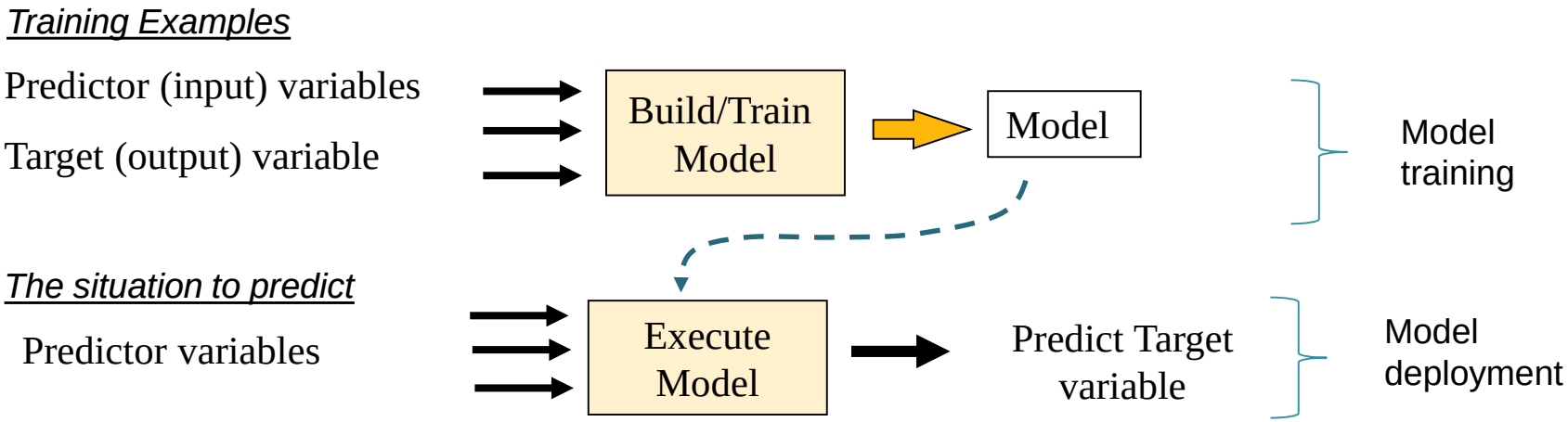
Does the problem map to a generic problem type?

- Are there suspected correlations, relationships ? *Exploration/Visualisation*
- Is there something that could be useful to predict? *Predictive Modelling*
- Do you hope to find things that happen (close) together? *Association Finding*
- Do you want to compare the current situation with past situations? *Look-alike*
- Do you hope/expect to find groupings/clusters? *Statistical Clustering*
- Are there exceptional cases that need investigation? *Outlier detection*
- None of the above – just find me some insights!

Problem Type → **Analytics Approach** → **Analytics Techniques**

Is it a Prediction Problem?

- Many problems can map to a predictive modeling approach
  - Who will buy X? How many $ will they spend?
  - Which insurance claim is a likely fraud?
  - Which patients will respond best to drug Y?



You need enough data to train and test the model!

(Many techniques exist: Decision tree learning, Regression, Neural Nets, SVM, …)

Is it a Prediction Problem?

- Can we apply predictive modeling to the bus lane investigation?
- Possible if we have a large enough data set containing sufficient instances of successful and failed bus lanes, e.g.

length (kms)	lights density	service density	destination type	route coverage%	success
2.50	6.42	2.16	Shopping	64.7	T
9.73	9.49	2.61	Leisure	76.4	T
7.33	10.61	3.73	Leisure	101.9	T
4.70	6.33	3.35	Residential	35.6	F
3.65	6.92	1.39	Leisure	36.1	F
7.31	7.12	3.34	Offices	17.3	T
5.44	4.81	1.59	Residential	83.3	T
3.34	8.36	2.85	Residential	56.6	F
5.27	5.29	4.62	Shopping	57.7	T
5.20	7.35	2.19	Shopping	23.6	T
4.66	9.81	0.95	Residential	25.5	F
5.84	5.81	2.30	Residential	5.2	F
5.44	5.91	4.14	Offices	63.5	T
5.68	7.28	3.40	Shopping	27.3	T

Predictor variables

Target variable

Is it a Prediction Problem?

Will my planned bus lane be effective?

- We can answer this question with a variety of approaches:
  - Predict success or failure (with probability)
  - Predict degree of success, e.g.
    - **% gains in riders**
    - **% reduction in bus clumping**
    - **% increase/decrease in car traffic along route**

Could also create a composite success measure as a weighted sum

- We can also build a prediction model as a means to identify what factors are important for bus lane success and in what circumstances they apply

Predicting Bus Lane Effectiveness

- Possible Training Set….

To select the best set of predictor variables we start by including all conceivable influencers and then use feature selection techniques to distill down (eliminate the irrelevant variables) to generate the modeling data set

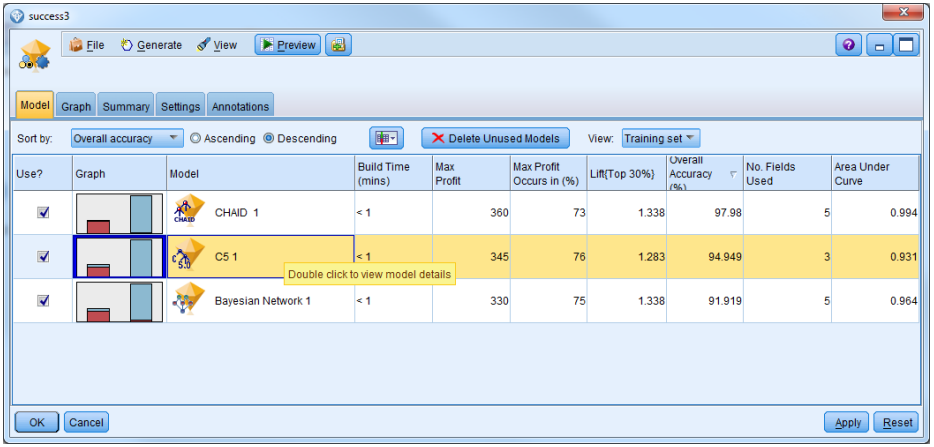
kms	tlightdensity	bus route density	dest.type	avg route coverage%	success
2.497	6.420	2.161	Shopping	64.741	T
9.730	9.494	2.611	Leisure	76.380	T
7.325	10.614	3.727	Leisure	101.871	T
4.696	6.329	3.353	Residential	35.642	F
3.652	6.916	1.388	Leisure	36.077	F
7.308	7.123	3.336	Offices	17.320	T
5.437	4.814	1.594	Residential	83.285	T
3.337	8.359	2.848	Residential	56.559	F
5.266	5.292	4.621	Shopping	57.698	T
5.202	7.347	2.187	Shopping	23.604	T
4.662	9.815	0.950	Residential	25.458	F
5.835	5.814	2.299	Residential	5.211	F
5.436	5.914	4.141	Offices	63.533	T
5.675	7.279	3.399	Shopping	27.349	T
3.634	6.300	2.462	Residential	76.438	T

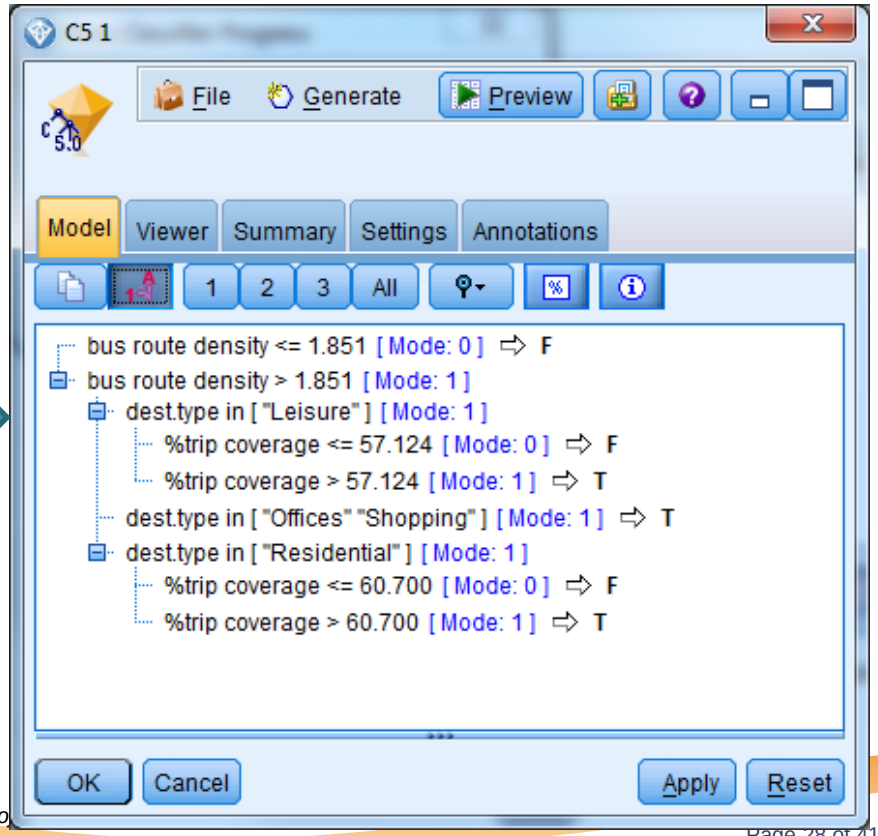
Predictor variables

Target variable

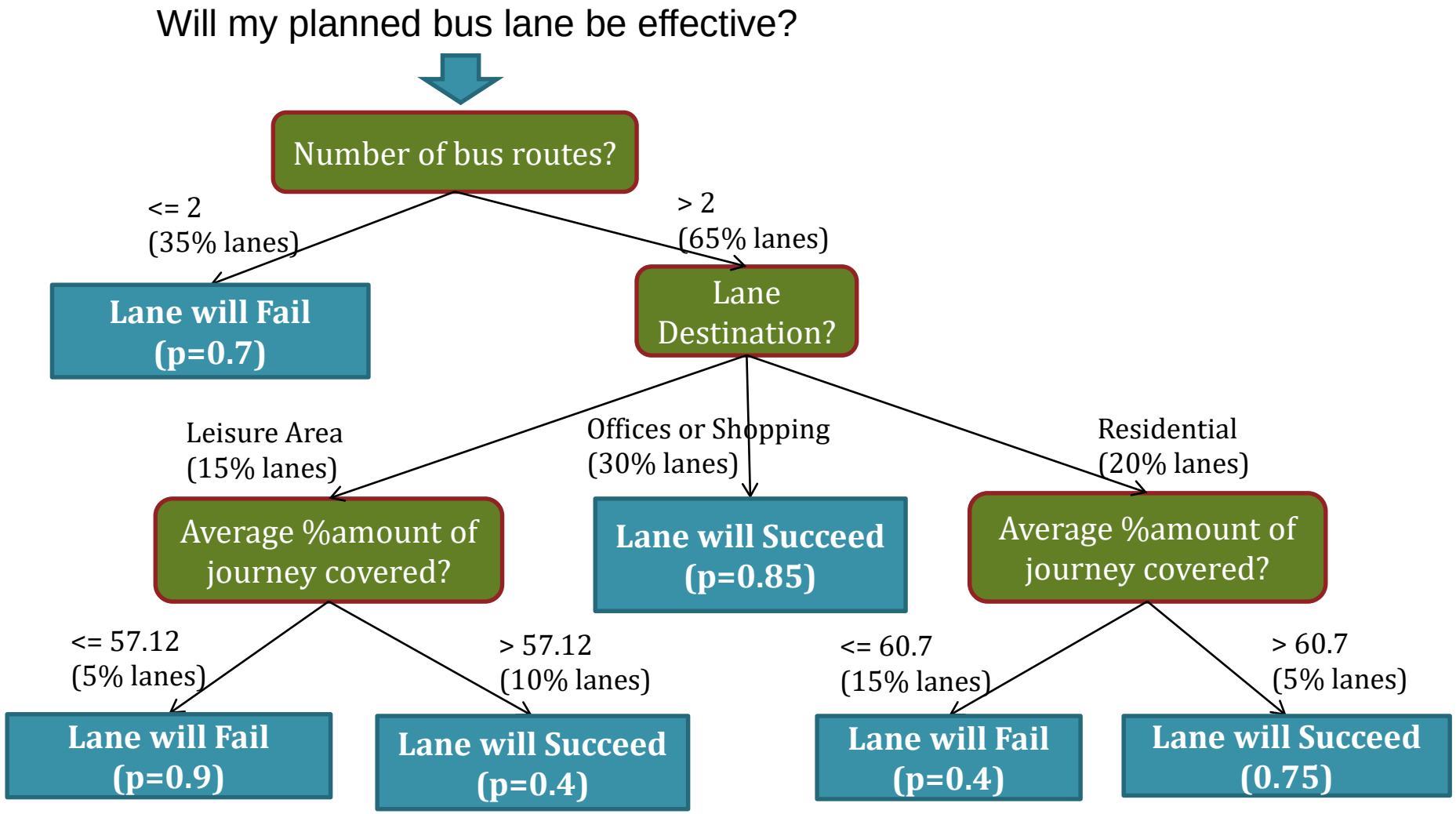
Building a Decision Tree Model

Field	Measurement	Values	Missing	Check	Role
kms	Continuous	[1.256417784335472,10.585644448748...		None	Input
tlightdensity	Continuous	[1.1428003452635096,11.85567456212...		None	Input
bus route density	Continuous	[0.7680001945797534,5.362147606796...		None	Input
dest.type	Nominal	Leisure,Offices,Residential,Shopping		None	Input
avg route coverage%	Continuous	<Read>		None	Input
success	Flag			None	Target





Predictive Models can give Insights



Find things that happen (close) together

Market basket analysis (association) methods are used to detect things that happen together

E.g. *"72% of customers who bought baby diapers also bought beer on Thursday nights"*

=> position diaper & beer together and have paired discounts!

Can be applied outside of the market basket!

Some imaginary (plausible) findings:

42% of people arriving at Changi from Bali also…..

36% of people who exited at Orchard MRT after 6pm also….

Comparing with past situations & known solns

Look-Alike modeling

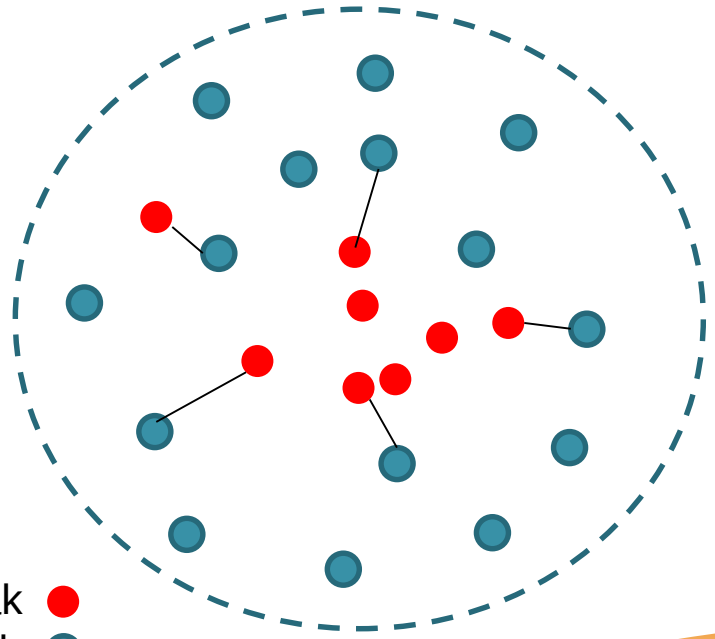
- Who does a person most look like? (what did they buy?)
- What other items look-like those I already own?
- What past case is most like the current patient?

Common algorithms

- Collaborative Filtering, Nearest Neighbour

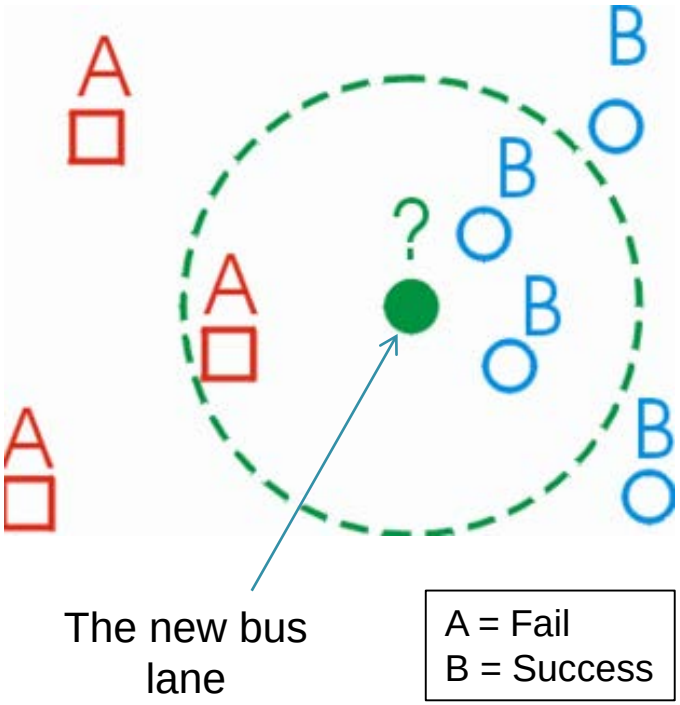
E.g. Who should we incentivize to travel off-peak?

- *Find the weekly EZ-link trip profiles of people who travel **during** the morning peak*
- *Compare with the profiles of people who are known to already travel **before** the morning peak*
- *Offer the **closest ones** off-peak travel incentives*



K-Nearest-Neighbour & Bus Lane Prediction

- Compare proposed bus lane with the existing lanes – compute their distances
- Look at the K nearest lanes to predict an outcome for the new bus lane (use majority vote). E.g. is K = 3 then…



Closest to the new lane

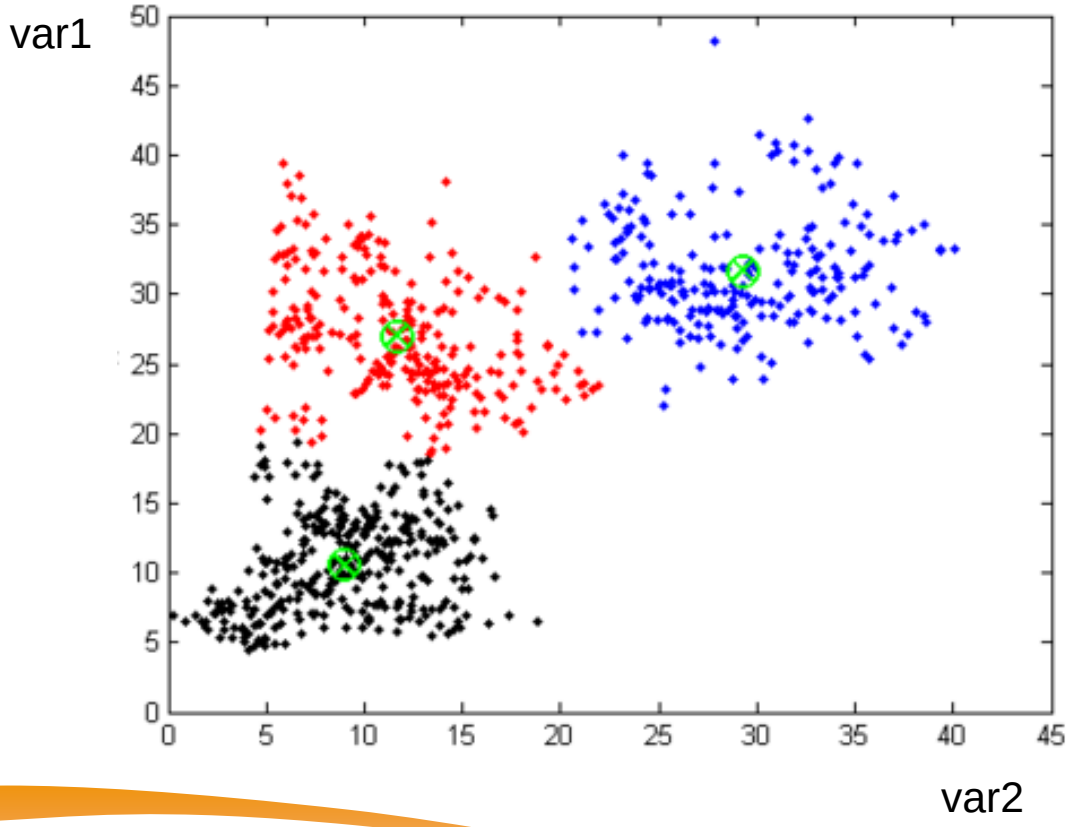
length (kms)	lights density	service density	destination type	route coverage%	success	Distance
2.50	6.42	2.16	Shopping	64.7	T	0.97
9.73	9.49	2.61	Leisure	76.4	T	0.16
7.33	10.61	3.73	Leisure	101.9	T	0.87
4.70	6.33	3.35	Residential	35.6	F	0.32
3.65	6.92	1.39	Leisure	36.1	F	0.56
7.31	7.12	3.34	Offices	17.3	T	0.87
5.44	4.81	1.59	Residential	83.3	T	0.84
3.34	8.36	2.85	Residential	56.6	F	0.12
5.27	5.29	4.62	Shopping	57.7	T	0.65
5.20	7.35	2.19	Shopping	23.6	T	0.14
4.66	9.81	0.95	Residential	25.5	F	0.82
5.84	5.81	2.30	Residential	5.2	F	0.29
5.44	5.91	4.14	Offices	63.5	T	0.84
5.68	7.28	3.40	Shopping	27.3	T	0.93

This is also called memory-based reasoning – why?

Is this method good or bad for huge data sets?

Finding Natural Groupings

- **Cluster Analysis**
  - Finding natural groupings using statistical algorithms or data visualization
  - E.g. find natural clusters of customers, products, transactions …..

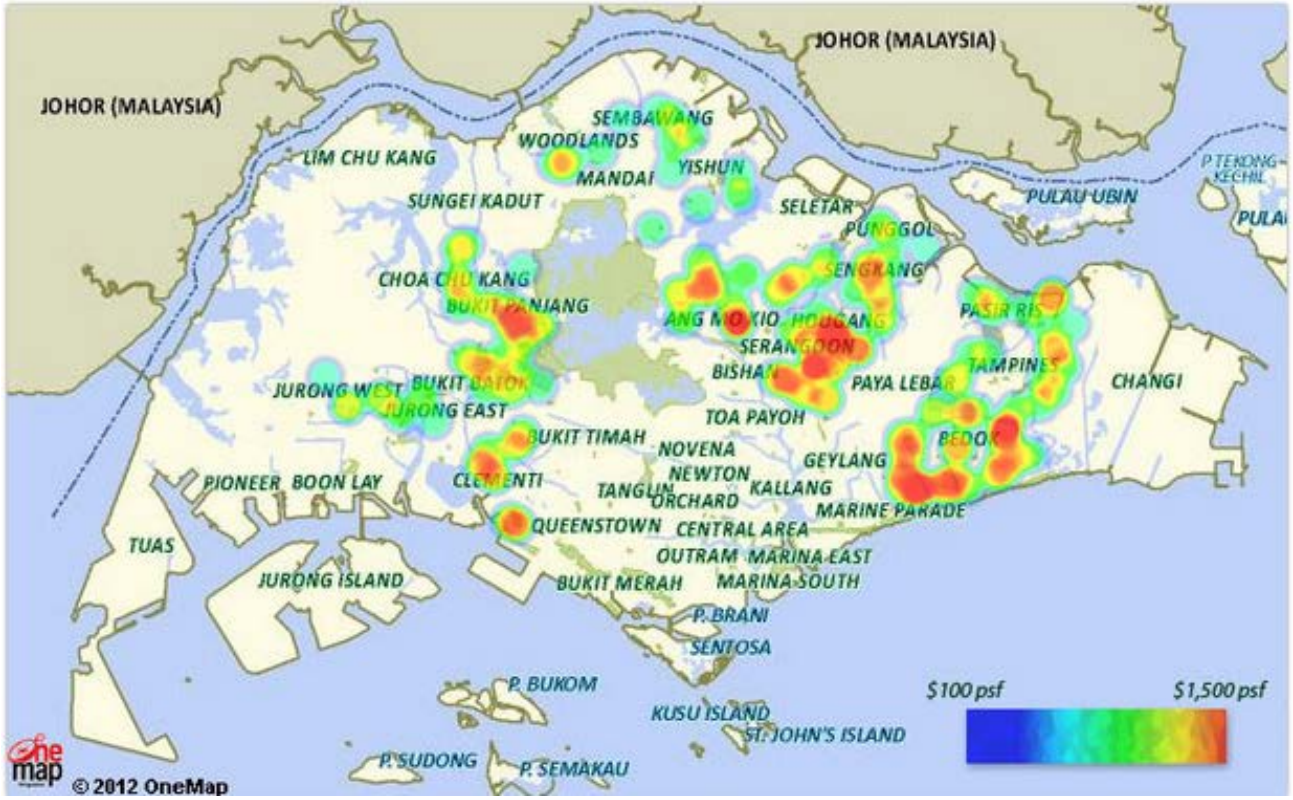


The discovered clusters can be used to:

- ***Gain insights*** *~ e.g. what common traits underlie a cluster of customers*
- ***Make predictions*** *~ e.g. classify a new case according to the majority class of the nearest cluster (faster version of kNN)*

Data Visualisation

- Leverage the ability of the human eye to see visual patterns
- Lines, bars, pies, histograms, heat maps, link analysis + many many more!

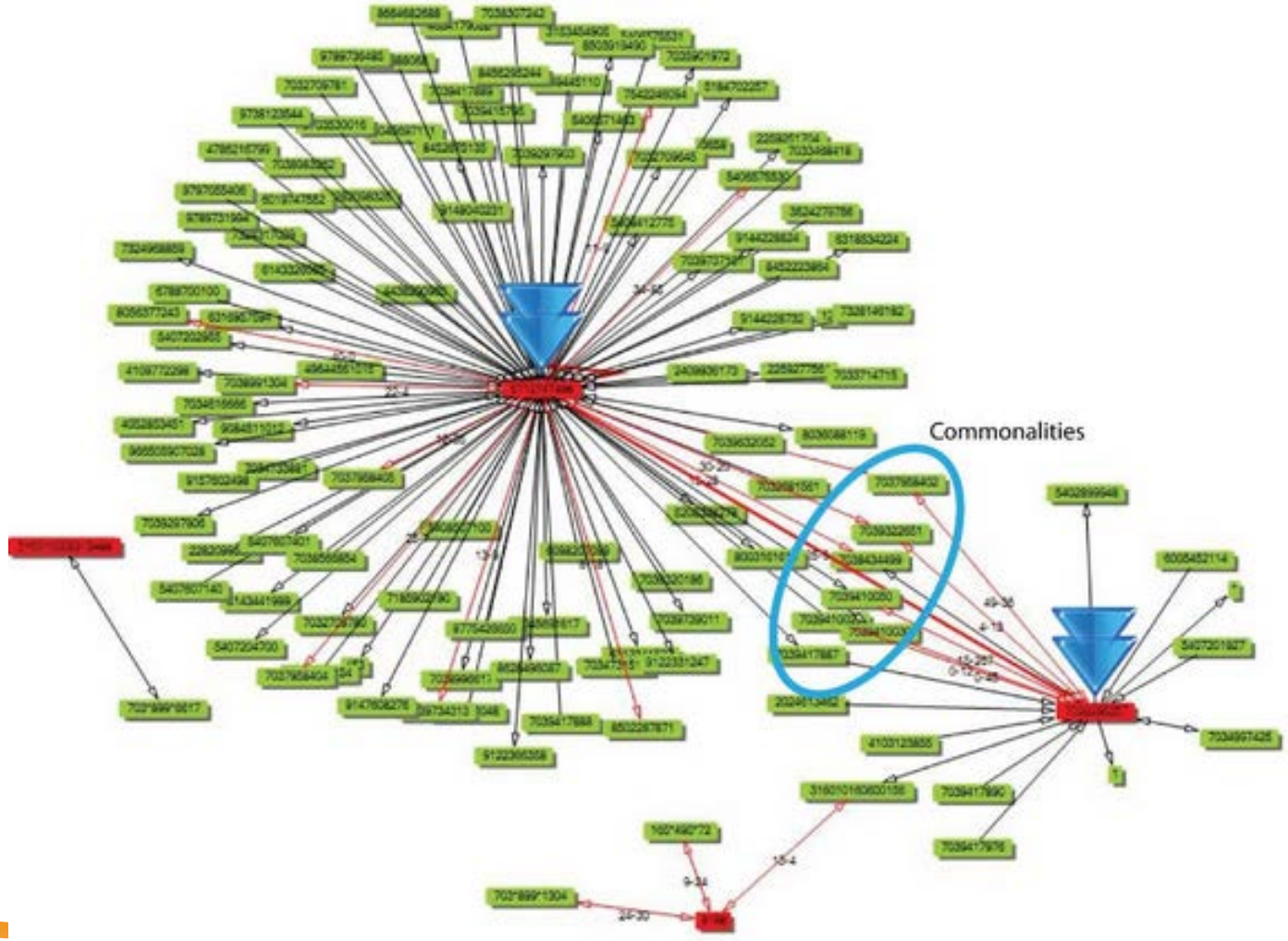


Data Visualisation can be used to:

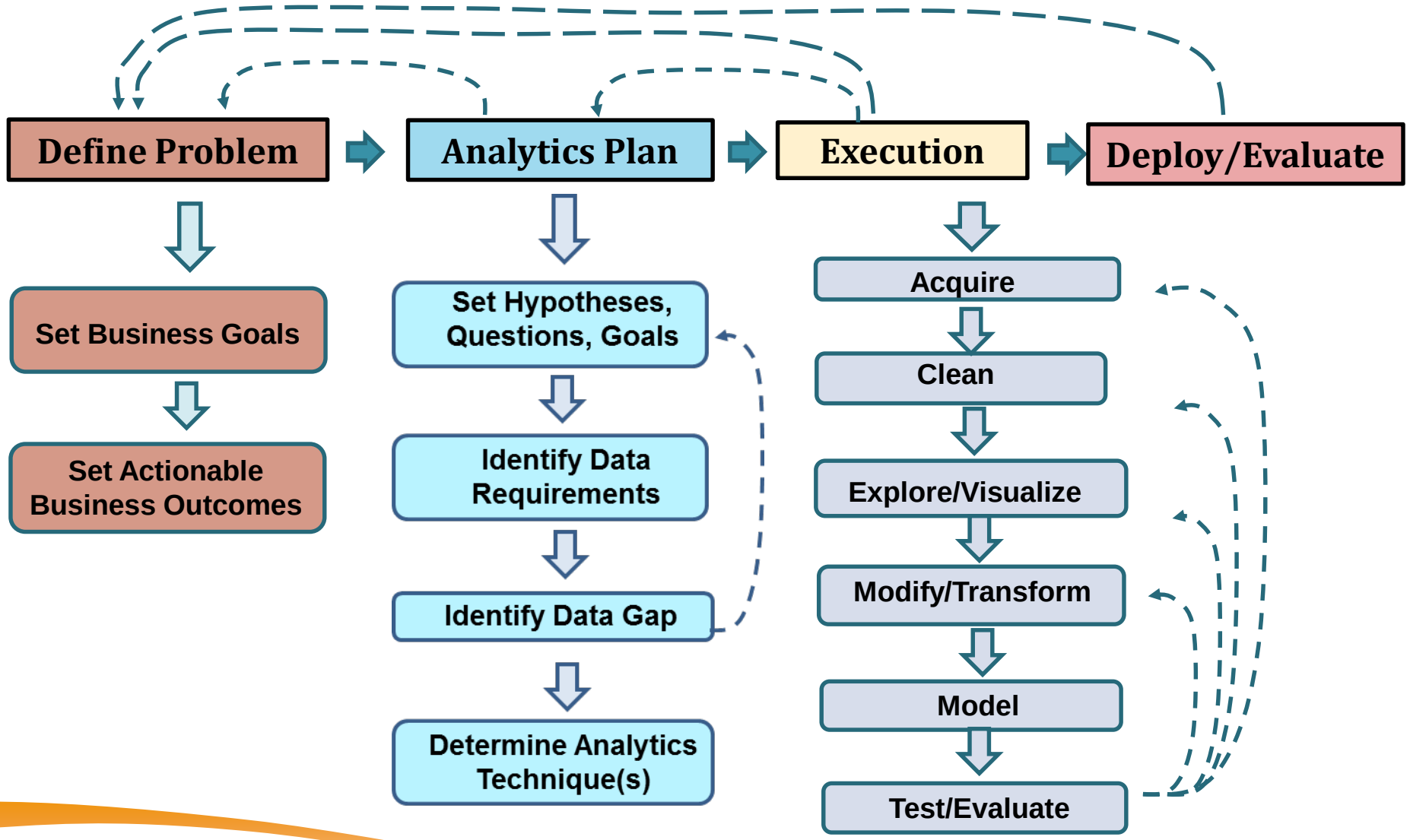
- ***Tell a Story*** ~ *display data on a story-board, business dashboard etc..*
- ***Knowledge Discovery***~ *find unexpected patterns and relationships*

Data Visualisation: Link Analysis

- E.g. tracking a churned cellphone user's calls to find their new number

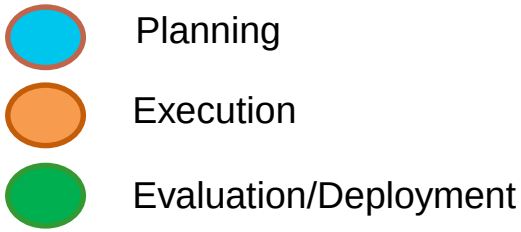


The Data Mining Process



Roles & Responsibilities (Guidelines)

Project Stage (Iterations may be required)	Champion / Domain Expert	Head Data Scientist	Data Scientists	Data Owners	IT Staff
Set Business/Project Goals	●	●		●	○
Analytics Goals & Plan	●	●		●	○
Data Collection, Integration		●	●	●	○
Data Preparation		●	●		
Exploration & Modeling		●	●		
Interpretation & Validation	●	●	●		
Results Deployment	●	●			○



Project Risk & Constraints

- What resources are available?
  - Human (business experts, data experts, tech. support etc..)
  - Data (files, live data warehouse, operational DB etc.)
  - Software (data mining tools etc.) & hardware
- What assumptions and constraints?
  - Is data availability assumed?
  - Are there legal constraints?
  - What other restrictions exist on data access
  - Are there constraints on budget, timescales etc.
- What are the risks? Contingency plan?
  - Business risk (e.g. competitor get better results first)
  - Organizational risks (e.g. department runs out of money)
  - Technical & data risks (e.g. poor quality or inadequate data)
  - Are there Risk control measures (e.g. a contingency plan)

Project Success Factors

- Get Project Sponsorship
  - Single most important determinant of success or failure
  - Must be senior and from business side and not IT side
  - They will want to see a clear objective that is aligned to the business strategy
- Set Clear Scope and expectations
  - Scope must be clear and well communicated to all
  - Start with narrow focus, then expand once the value of data mining has been shown
  - Timelines and expectations must be realistic
  - Report early results
- Good Project Team
  - Good composition with the correct skills
  - Involve the users
  - Regular progress reporting

DM Planning Workshop

- A telco company is concerned about the number of customers it is losing to competitors. They want to retain as many customers as possible.
- How might we use data analytics/data mining to help the company?

  how to reduce the churn rate?

"The service is lousy. I will terminate the contract next month" (customer)

"In recent months, we have lost many customers. We need to do something ASAP" (company CEO)

Team-Based Workshop

- Break into teams to brainstorm how we might use data analytics/data mining to deal with this problem…

Tasks
Determine starting questions & hypotheses
What data is required/could be relevant? Consider internal, partner, external data (paid & public) Identify mechanisms for acquisition & data quality issues How to bridge any data gap?
Brainstorm possible analytics approaches

Short team presentations will be held at the end
(I'll randomly pick a few teams)