Data Analytics Best Practices

1.1

Agenda

Day 1

- Analytics basics
  - Analytics Processes
  - Data Requirements
  - Types of datasets in the industry
  - Data issues
  - Data Cleaning
  - Data Integration
- Workshop on arranging data elements
  - Functions
  - Data Formats
  - Date-time in R

Day 2

- Analytics best practices
  - Data Transformation
  - Exploratory Visualisation
  - Feature Engineering
  - Decision Engineering
  - Model Deployment
  - Model Maintenance
  - ROI Models
- Workshop on Analytics best practices
  - Intro to Data Cleaning
  - Data Preparation

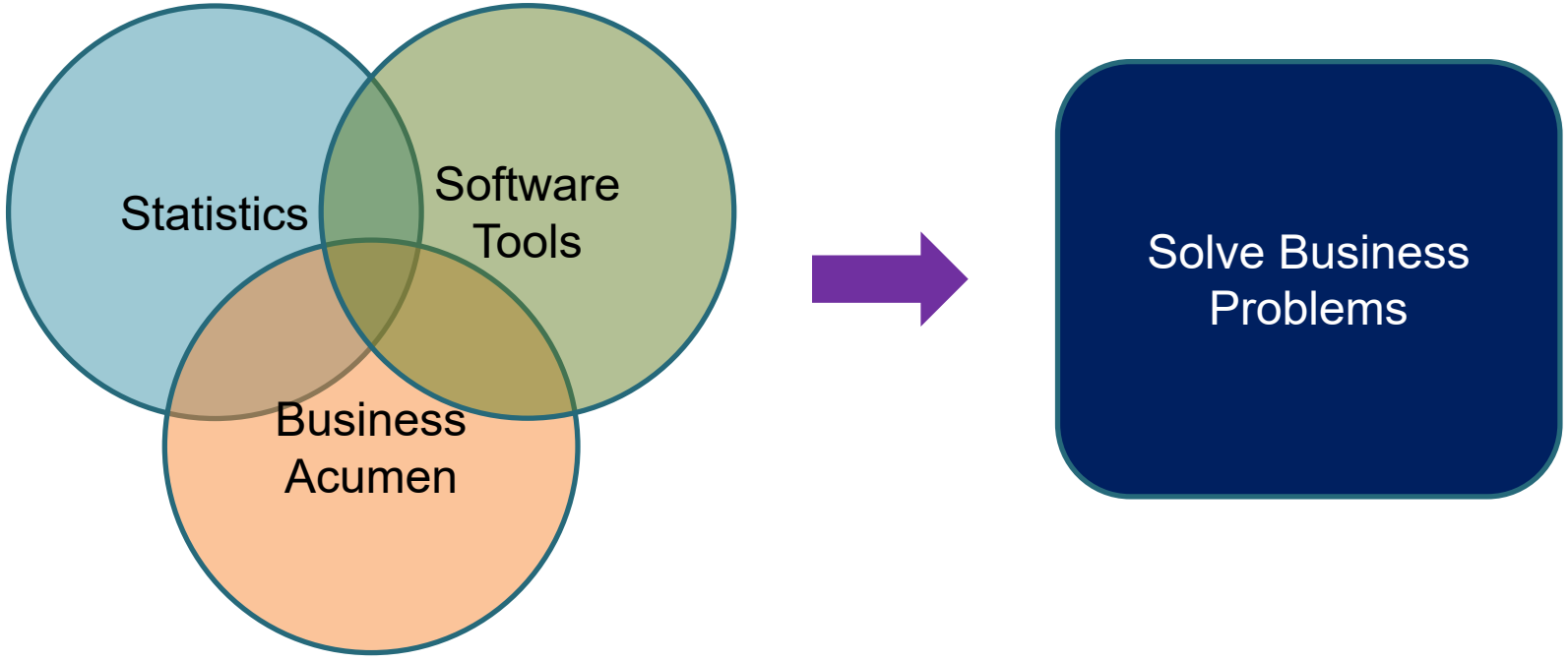
Day 3

- Workshop on Data exploration
  - Visual Data Exploration
  - Non-Visual Data Exploration
- Data Warehousing Basics
  - Data Warehousing Introduction
  - Data Modelling Essentials

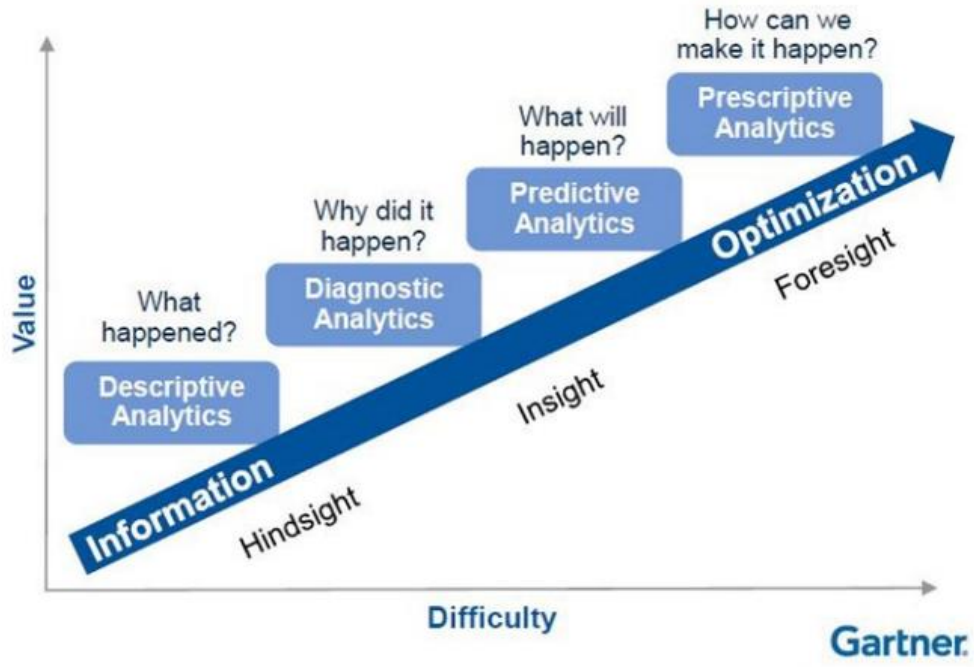
Analytics Processes

Data Analytics Best Practices

What is Analytics?



Common goals of Analytics



Analytics drives decisions across Industries

Banking

- Credit Decisioning
- Fraud Detection
- Anti Money Laundering

Healthcare

- Clinical Trial Analysis
- Drug Discovery
- Preventive Care

Telecom

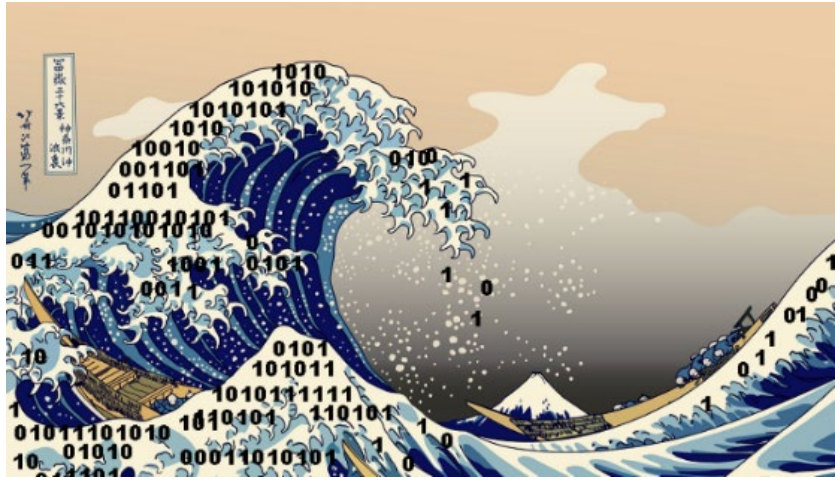
- Churn Prevention
- Product Development
- Service Quality Improvement

Manufacturing

- Process Improvement
- Supply chain optimisation
- Preventive Maintenance

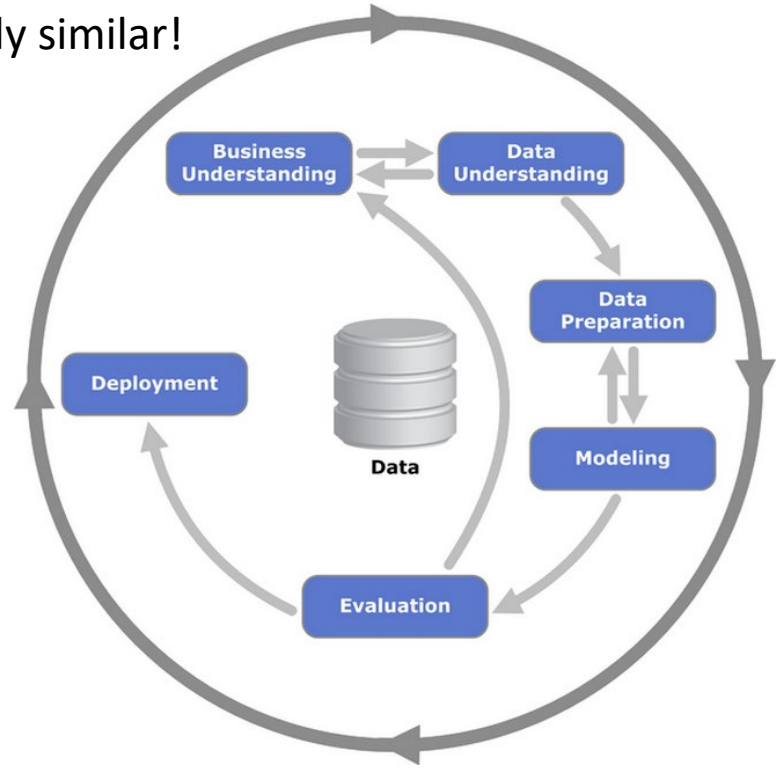
Data Type Explosion is leading to new kinds of Analytics

- Big Data Analytics
- Streaming Analytics
- Unstructured Data Analytics



CRISP DM Process

Many methodologies exist – mostly similar!



We follow (mostly) the Cross Industry Standard Process for Data Mining (CRISP-DM).
(CRISP-DM was conceived in late 1996 by collaboration between vendors and end-user orgs, including SPSS, Daimler-Benz, NCR)

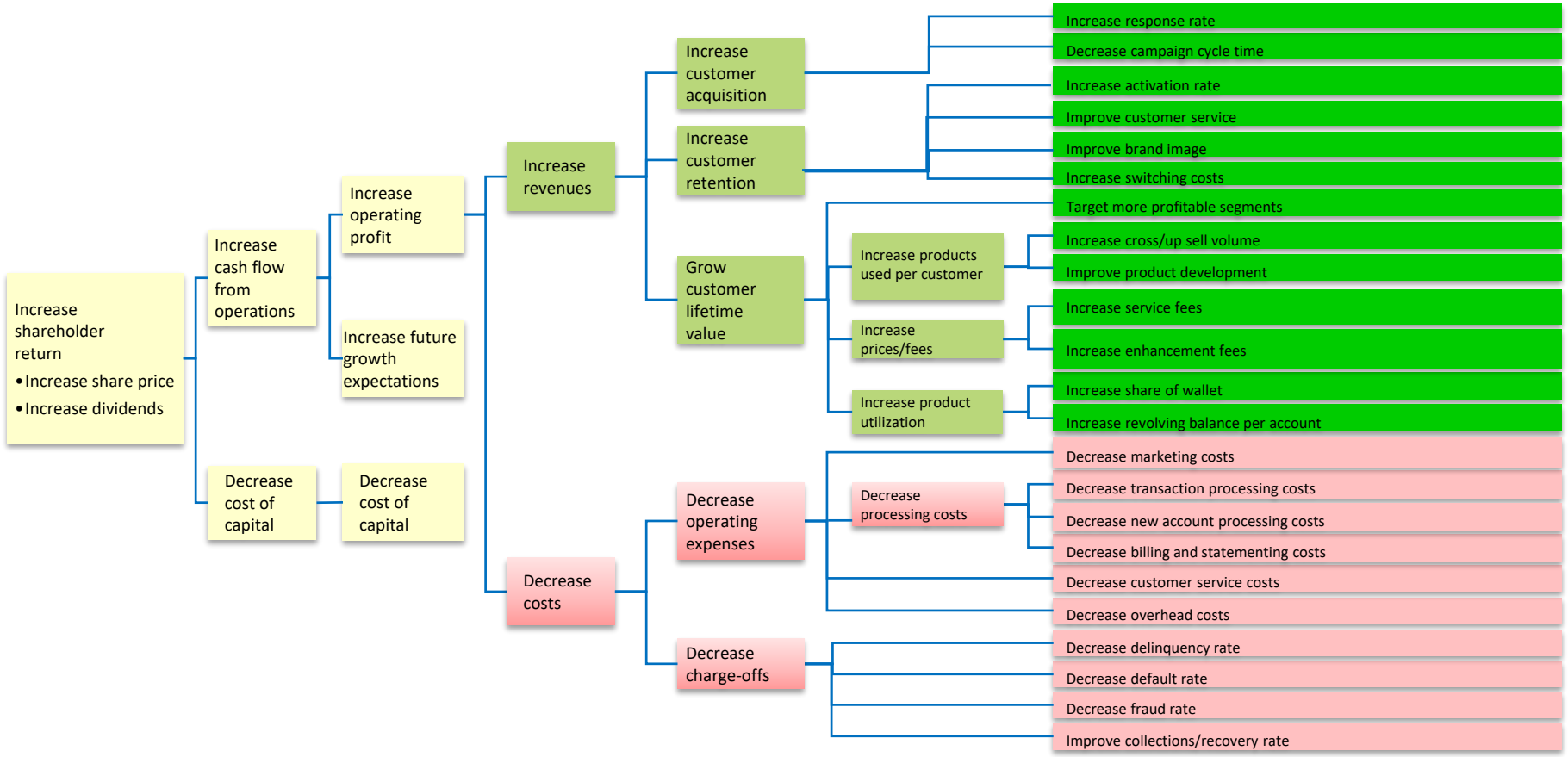
Business Understanding

- Who is the end user?
- What is the end benefit?
- How will "it" be deployed?
- What is the champion?
- What are the issues with the champion?
- Collect hypotheses
- Convert to Analytics Goals

- The most important step
- The most neglected step

Business << >> Analytics

A business has multiple objectives

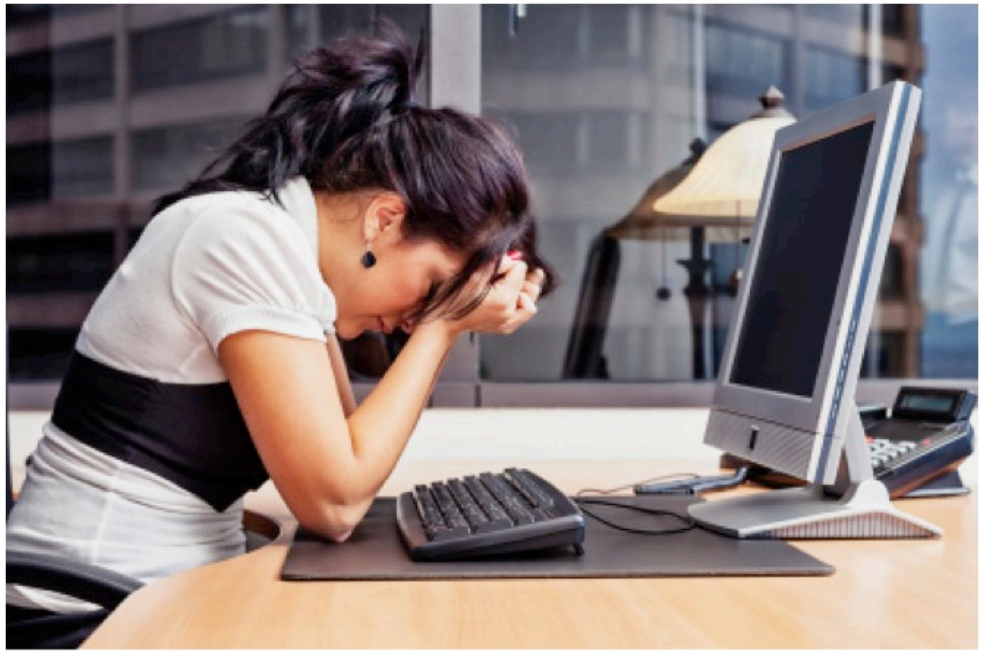


Business Objectives at a Retail Bank

Source: Accenture

Data Understanding

- What data do you need?
- What data is available?
- How easy is it to access the data?
- Where is the data?
- Will the data be available during deployment?
- What are the privacy norms to comply with?

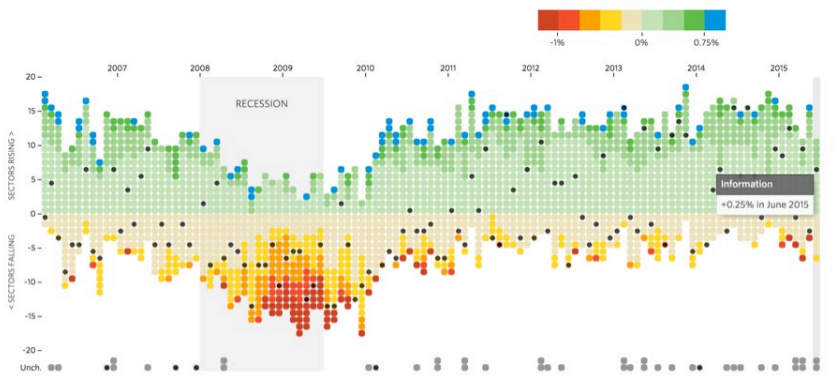


- Can you create data?
- How dirty is the data?

IT << >> Analytics

Data Preparation

- The most time consuming step
- Data Janitoring
- Data Quality
- Data Transformation
- Analyst code usually is not production quality
- What are the overall trends?
- Any special patterns?
- Does business agree?
- Does this throw new insights?

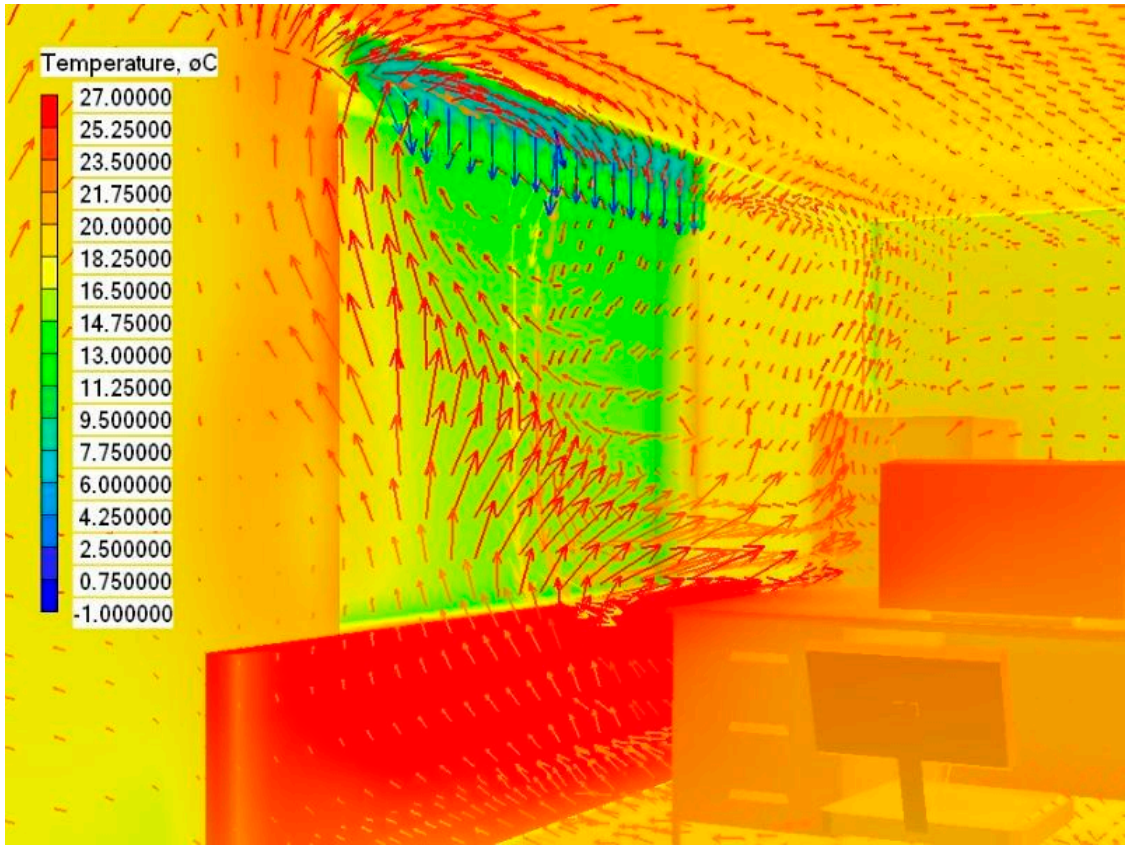


- How reusable is your code?
- Garbage in Garbage out
- Visualization is the Biggest wow factor, value addition

IT << >> Analytics

Modeling

- Is an art
- Choose appropriate technique(s)
- Do tools support technique choice?



Analytics

- A model is as good as the underlying data

Model Evaluation

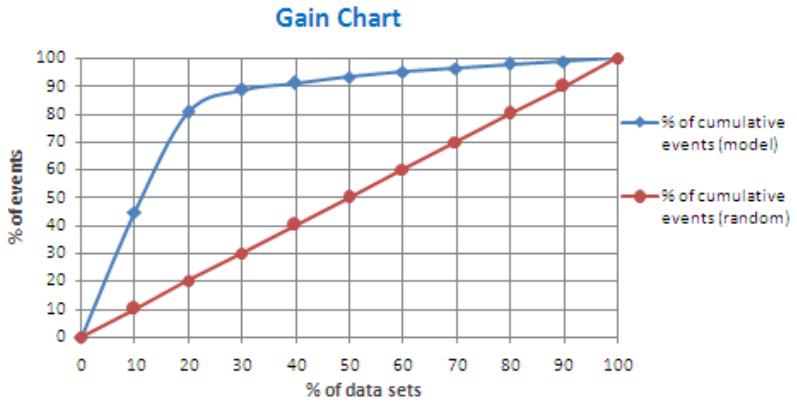
- Training vs. Testing
- Out of time vs. Out of sample
- How "good" is the model?
- Challenger defeats the champion

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Analytics

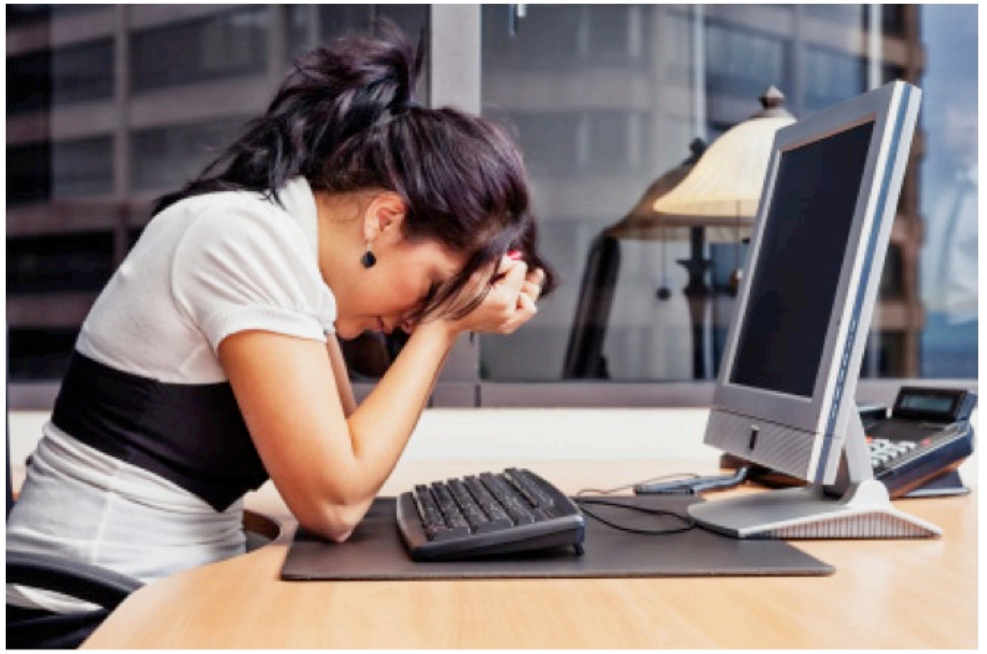
Measurement	Calculation
Accuracy	(TP+TN)/Total
Misclassification Rate/Error rate	1- accuracy



Deployment

- Scoring Engine
- Decision Engineering
- Model Health
- Recalibration

- Handed over to the deployment team
- Prone to handover issues



Analytics>>IT>>Business

Selecting an Analytics Approach

Does the problem map to a generic type?

- Are there suspected correlations, relationships ? *Exploration/Visualization*
- 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 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! *Visualization & Exploration*

Setting Analytics Goals

- ***Example***: You are the marketing VP for a bank and your primary business objective is to retain current customers who are at risk of moving to a competitor.

- Possible Approaches
  - Identify ***likely attriters*** then offer them incentives to stay.
    To do this get customer profile data and account usage data for both loyal customers and churners.
    Use this to build a churn prediction model.
  - Identify ***the issues causing customers to attrite*** – then fix these issues!
    What data is required for this?

Target Variable

- Definition of the target variable needs to take into account the implementation details
- Examples of target variables
  - Ever 30 DPD
  - Active customer who has voluntarily cancelled
  - Fraudulent Transaction as discovered

Analytics approaches could be of the following types

Business Logic Rules

- Simple business logic based algorithms that could be implemented as rules. The logic of such algorithms can usually be handcrafted

Summarisation

- Algorithms whose purpose is to create aggregations, classifications or groupings of data

Prediciton

- Algorithms whose purpose is to predict a metric from observations in the past

Event Models

- Algorithms that use events as inputs for predictions. Managing would need the ability to process and manipulate events

Analytics Approach Examples

Business Logic Rules	Summarization	Prediction	Event Models
→ Next best offer rules	→ Customer profiling	→ Next best interaction models	→ Location patterning
→ Contact Policies	→ Value based Segmentation	→ Profile inferring	→ Intent models
→ Targeted Advertising rules	→ Life stage segmentation	→ Intent models	→ Context models
→ Custom Audiences	→ Need based segmentation	→ Drop off prediction models	→ Profile inferring
→ Retargeting rules	→ RFM Analysis	→ Issue prediction models	→ Interaction stage association
→ Context Creation Rules	→ Look alike models	→ Response models	→ Timing response models
→ Process sequence rules	→ Path analysis	→ Activation models	
→ Spend triggers	→ Statistical lift testing	→ X Sell models	
→ Inactivity triggers		→ Up Sell models	
→ Interaction based rules		→ Conversion models	
→ Root cause analysis		→ Attrition models	
→ Points based triggers		→ Propensity models	
		→ Time to contact models	
		→ Affinity models	

Analytics Approach Techniques

Business Logic Rules	Summarization	Prediction	Event Models
→ If then else rules	○ Directed Segmentation	○ Regression	→ Event trigger rules
→ Trigger based rules	○ Cluster Analysis	○ Forecasting	→ Complex event processing rules
	○ Topic Modeling	○ Decision trees	→ Hoeffding Decision Trees
	○ Association Mining	○ Neural Networks	→ Streaming Apriori
		○ Support Vector Machine	
		○ Naïve Bayes	
		○ Memory-based reasoning	
		○ Ensemble approaches	

Data Requirements

Data Analytics Best Practices

Identifying Data Requirements

- Questions to Answer:
  - What data is available?
  - What must the data contain?
  - What would be useful? (whether available or not) – **innovate!**
  - What is the right level of granularity?
  - What volume of data is needed?
  - How much history is required?
- What data is required for comparison?
  - What is currently being done?
    - E.g. what is the existing churn rate, response rate, failure rate?
  - Obtain a control group ~ data describing the status quo
    - E.g. what happened to patients who did not receive the treatment?
    - E.g. what did customers buy who did not see the ad?

?

Data dictionary

- Explains what the data contains
- Is a critical requirement for data understanding

Field	Description	Values
pclass	Passenger Class	1,2,3 for class
survived	Whether the passenger survived or not	1:survived, 0:perished
Name	Name of the passenger	Text String
Sex	Gender of the passenger	male, female
Age	Age of the passenger	Number
Sibsp	Siblings/spouses on board	Number
Fare	Fare paid	??
Cabin	Cabin number	Text String
Embarked	Port of embarkation	??
Home.dest	Home Destination	Text String

Data Masking is required for data like

- Customer/Patient Personally Identifiable information
  - Names
  - Phone numbers
  - Email id
  - Credit Card number
- Employee Details
  - Employee id
  - Salary
  - Family Data
  - HR status (termination, personnel issues)
- Company Secrets
  - Pricing
  - Confidential Details
  - Contract Details
  - Financials

Masking Process

- Step 1: Confidential/Sensitive fields identified
- Step 2: Fields not required for analytics dropped
- Step 3: Unique key for all data sets identified (if not available, created)
- Step 4: Unique key masked and converted to an encrypted key
- Step 5: Data sent to Analytics team with encrypted key
- Analysis
- Step 6: Analytics output sent back to masking team with encrypted key
- Step 7: Masking team decrypts the key to map back to original data

Exclusion and Inclusion Criteria

- It is important to pick the right data points. Some of the common exclusion and inclusion criteria are
  - Start and End time for analysis
  - Separating a special group of customers example staff credit cards
  - Customer joining date (recently joined customers might be excluded)
- Ensure that the exclusion criteria applies to all the raw data sets

- Exercise: Discuss the exclusion and inclusion criteria for studying seasonality based trends in credit card data

Types of Datasets in the Industry

Data Analytics Best Practices

Analytics Data Stack

Category	Definition	Example
Structured Data	Data which can be structured into a relation databases or data tables	Customer Date of Birth
Semi structured Data	Data that does not conform to data models associated with relational databases or other forms of data tables, but contains tags or markers to separate semantic elements and enforce hierarchies of fields within data. (Wikipedia)	Location traces
Meta Data	Data about data. There are two types: structural, about design and specification of structures; and descriptive, about individual instances of data content. (Wikipedia)	Data type of variable
Unstructured Data	Information without a pre-defined data model. Unstructured information is typically text- heavy, but may contain data such as dates, numbers, and facts. (Wikipedia)	Email from Customer

Structured Data Examples

- Customer profiles
- Prospect profiles
- Social profiles
- Customer preferences
- Agent profiles
- Offer reaction data
- Redemption data
- Loyalty points data
- Fraud data

Semi Structured Data Examples

- Click streams
- Location streams
- Customer PFM tagging
- Process instance data
- Handoff instance data
- Interaction streams
- FAQ knowledgebase
- Intent data
- Customer Service interaction data
- ATM logs
- Transaction data

Meta Data Examples

- Click stream metadata
- Merchant category mapping
- Map data
- Open graph standards
- Context definitions
- Public data like calendars
- Process map repository
- Product data
- Messaging data

Unstructured Data Examples

- Customer documents
- Customer statements
- Chat transcripts
- Voice records
- Social media posts
- Incoming Email

Types of Datasets in the Industry

- Demographics
- Behavioural Data
- Product Information
- Derived Data
- Third Party Data
- Publicly Available Data

Demographics

- Data about customers and their profiles
- Fields include
  - Gender
  - Age
  - Marital Status
  - Income
- Demographics is slowly changing data
- Multiple sources of data could be available, challenges include deduplication and updating
- Challenge is data preparation where a lot of derived variables can be created from Behavioural data
- The predictive power of demographic data is usually minimal and hence more data should always be sought

Behavioural Data

- Data about customer interactions with respect to a company or a product
- Example Fields
  - Spend on Credit Card
  - Amount paid on Bill
  - Customer Service Interactions
  - Credit Default Information
- Behavioural data is temporal in nature, so a time stamp with the data is critical to capture
- The predictive power of behavioural data is higher hence is used extensively in modelling

Product Information

- Data about the product being studied
- Example Fields
  - Credit Limit
  - Affiliation (Visa/Master/Amex)
  - Type of Card (Classic/Gold/Platinum)
  - Colours available (iPhone colours)
- Product Information is used to enhance the value of Behavioural or Demographic data
- Product Information is rarely used on its own in Analysis

Derived Data

- The previous three examples are of raw data
- Derived data consists of new variables created using raw data and combinations
- Example Fields
  - Percentage Revolved
  - Time since last payment
  - Time since last complaint
  - Number of Products held
- Derived data used well could have the highest predictive power amongst data types

Third Party Data

- Data that could be bought from vendors like DMPs or Bureaus to enhance the analysis
- Example Fields
  - Credit Scores
  - Blacklist Entries
  - IAB profiles
  - Facebook custom data
- Security mechanisms like hashing protect the confidentiality of data from either party
- A good practice is to test the value of such data with a sample before purchasing the entire dataset

Examples of External Third Party Data

- Data as a service
- APIs
- Credit Bureau data
- Negative lists
- DNC registry data
- National identification services
- Facebook data
- MCC Categories

Publicly Available Data

- There is a wealth of data available in the public domain
- Examples Include
  - Linked Open Data
  - Weather Information
  - Industry Standards
  - Published Reports
- Often analysis can be greatly enhanced with the use of public data sources
- A good understanding of the domain is required to extract and make use of such data

Data Issues

Data Analytics Best Practices

Data sources

- Transaction Processing Systems
  - This is the "rawest" form of data, right at the source
  - An application expert would be required to extract the right data
  - Data would have to be cleaned for quality issues
- Data warehouses
  - Mature organizations have one or more data warehouses that data can be extracted from
  - This data is usually in a cleaner format
  - Data latency should be taken into account when using such data
- Third Party APIs
  - Data could be available from third party sources via APIs
  - This data will have to be integrated with existing data for proper analysis
  - Challenges include throttling limits and accuracy

Data access challenges

- Freshness
  - How often does the ETL process need to be run?
  - Different data sources might need different frequencies
- Master data management
  - Single source of meta data
  - New entries for newly created variables during the analytics process
  - Analytics team needs to be aware of newly created data
- Processing load
  - Ensuring the ETL process does not slow down the day to day operations
  - Ensuring that the new process can handle the required load
  - Ensuring that the prepared data is sized properly

Data Integrity Concerns

- Incomplete Datasets
- Loss in transmission
- Missing Fields
- Missing data dictionary entries
- Data dictionary conformance

- Good Practice requires you to create a Data Integrity Report
  - List out the tables, fields and number of rows per field received
  - Ensure that you have all fields explained in the data dictionary
  - Check that the values of the fields correspond to the data dictionary

Data Audit

- A useful starting point is the Data Audit. Audit goals include:
  - Is the data adequate?
  - Is it what you expect?
  - Does it look sensible?
  - What are the data quality issues? (What cleaning is required?)
- Data Exploration is more concerned with analysis and discovery (can also be done on the prepared data)
  - Find answers to questions asked
  - Make recommendations
  - Find Insights
  - Data visualization is a key tool

Data Quality Concerns

- Wrongly formatted Data
- Missing Data
- Duplicated Data
- Extreme Outliers in Data

Data Cleaning

Data Analytics Best Practices

Data Cleaning

- Data may not be perfectly collected, or collected with the right purpose.
- Many reasons exist for data to be dirty
  - Data entry errors
  - Misplaced decimal points
  - Inherent error in counting or measuring devices
  - External factors, etc.
- Data exploration can discover anomalous patterns, leading to the questioning of data quality
  - E.g. categories with very low frequency counts ➔ mistyping?
  - Name and addresses recorded in multiple ways in data integrated from multiple sources (can be up to 20~30 variations)
  - Missing data

Data Cleaning Tasks

- Data cleaning tasks
  - Handle missing values
  - Handle noisy / erroneous data
  - Handle outliers
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Values

- Common feature of any dataset
- Various reasons:
  - Information not available
  - Lost data / accidentally deleted
  - Purposefully left out with a reason
- Missing does not always imply an empty/blank value. There may be a value entered in the data that signifies missing
  - E.g. *"9999", "1 Jan 1900", "*", "?", "#", "$"*, etc
- The presence of missing values in data can make problems for the modeling tools.

Handling Missing Values

- Ignore attributes that have majority of values missing?
- Ignore data records with missing values?
  - Throwing away data ~ but this is bad if you do not have much data!
  - Especially poor when the percentage of missing values per attribute varies considerably – one attribute (which may not even be important) with few values could cause the whole data to be discarded!

Gender	Children	Salary	Bought PEP
M	-	29,000	Y
M	-	65,000	Y
F	2	26,500	Y
M	-	47,000	Y
F	-	15,000	N
-	1	23,000	N
F	-	36,000	N

What should we do here?

Handling Missing Values

- Data Imputation - fill in the missing values automatically
  - Guiding Principle: Avoid adding bias and distortion to the data
  - Understand why the data is missing can help guide the imputation
  - Often a missing value means zero or the default value. E.g. for '*rainfall*' variable, a missing value may mean no rain on that day ➔ 0
    - Common Options
      - A global **constant** : e.g., "unknown" or 0 (zero)

        Easy, but modeling algorithms may mistakingly treat "unknown" as a concept
      - The **attribute mean** (or median, mode)

        Simple and quick though not always satisfactory
      - The **attribute mean** for all samples belonging to the **same class**

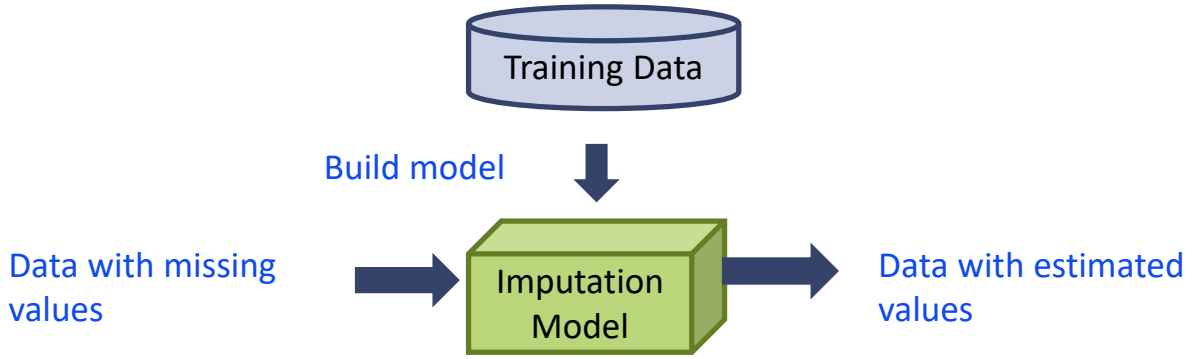
        Often a better estimate than attribute mean

Gender	Children	Salary	Bought PEP
M	1	29,000	Y
M	0	65,000	Y
F	2	-	Y
M	0	47,000	Y
F	-	15,000	N
-	1	23,000	N
F	1	36,000	N

What should we do here?

Data Imputation

- Train a prediction model (e.g. regression model, decision tree) to predict the most probable value
  - Use variables containing values to estimate the variable with missing values
  - Can produce good estimates.
  - Need training data and additional modeling



Missing data - plan

- We don't anticipate missing data
- Before data collection begins, determine how missing data will be recorded and entered
- Possibilities:
  - Not applicable N/A
  - Not available N.A.
  - Unknown
  - True missing (attribute recording missed!)

Missing data

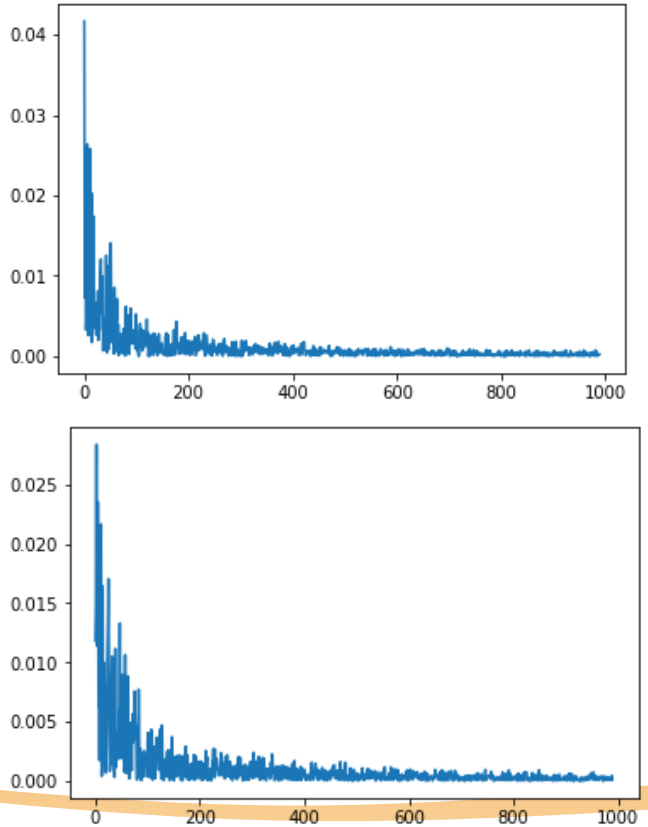
- Determine how much data is missing?
- "Missingness"
  - Missing completely at random
  - Missing at random
  - Not missing at random
- Analyze for patterns of missingness
  - Systematic
  - Random

Missing data – imputation

- Simple method:
  - Exclude cases pairwise
  - Exclude rows
- Advanced method:
  - Imputation: we decide based on present data how to handle missing data
  - Prediction
  - Similarity
  - Mean and median

Missing data – Mean and median

- Replace the missing values with mean of the column
- It distorts the mean and median of the data
- Mean changes affect algorithms which work on **means to optimise**
  - **PCA**
  - **k-means clustering**
  - **etc.**
- **Good for few missing values**



Missing data – Interpolation

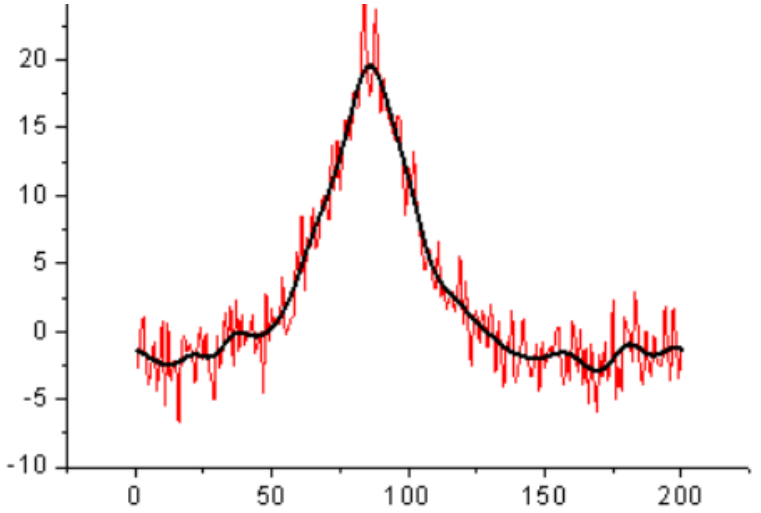
- Replace the missing values by interpolating nearby values of the column
- **Linear**
- **Time**
- **Nearest**
- **Derivatives**

Noisy / Erroneous Data

- **Noise:** random error or variance in a measured variable
- Incorrect attribute values may have been entered due to
  - Measurement error: faulty (or inaccurate) data collection instruments
  - Data entry problems
  - Data transmission problems
  - Inconsistency in naming convention

Noise handling Methods

- Binning
  - Sort and bin data, use bin means, medians etc
- Curve/Line Fitting
  - Fitting the data into regression functions
- Ensemble methods
  - Averaging the results from multiple models



Outliers

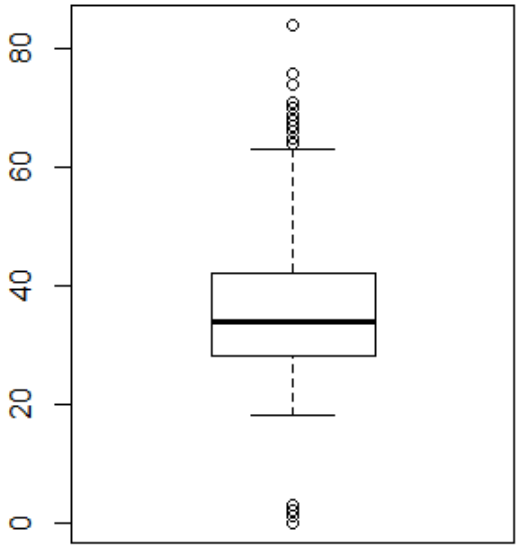
- Observations that "*deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism*". (Hawkins, 1980)
- Appearing at the maximum or minimum end of a variable, skewing or distorting the distribution
  - E.g. extreme weather conditions on a particular day, a very wealthy person financially very different from the rest of the population, etc.

Handling Outliers

- Outliers may be errors ***or*** they may be valid data!
  - Can be rare, unusual, infrequent events we are interested in.
  - They should be identified for further investigation.
  - E.g. frauds in income tax, insurance, banking, etc.
- Otherwise, outliers usually should be removed to avoid adversely affecting the modeling result (though some algorithms, like random forests and support vector machines can be robust to outliers)

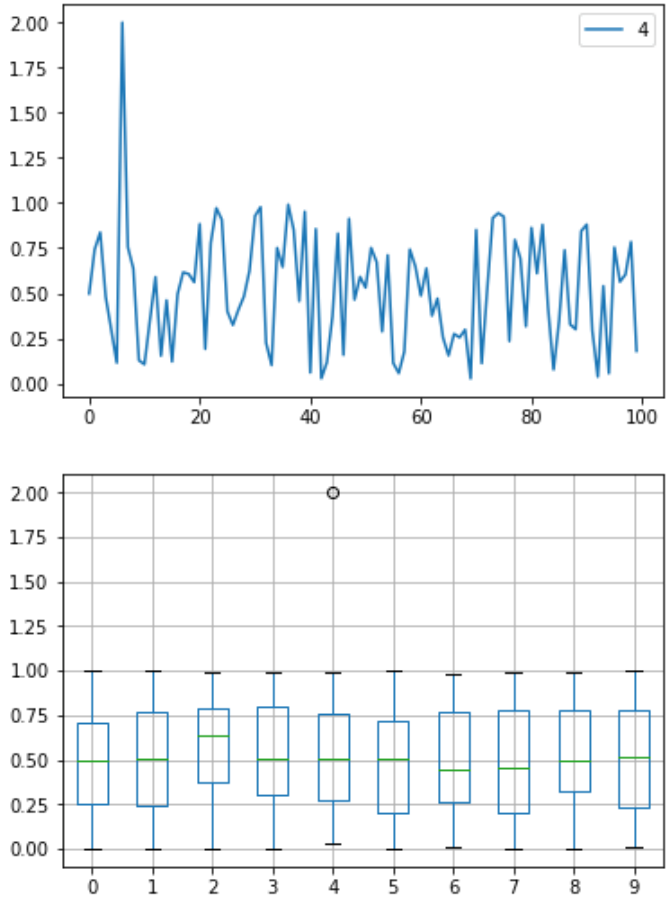
Identifying Outliers

- Statistical tests for variance
- Clustering
- Human inspection
- Others…



Outliers

- Disturb the mean of the data distribution
- Easy to identify!
- Handle?
  - Treat as missing value
  - Remove data
  - Interpolate



Deduplication

→Problem

→Customers acquired at different stages or for different products might not be recognized as the same customer in the bank

→Need

→Apart from wastage of communication effort, the right picture of a customer is not built leading to erroneous marketing

→Challenges

→Name matching (when there is no unique identifier); attribute finalization (when two sources disagree) and data augmentation

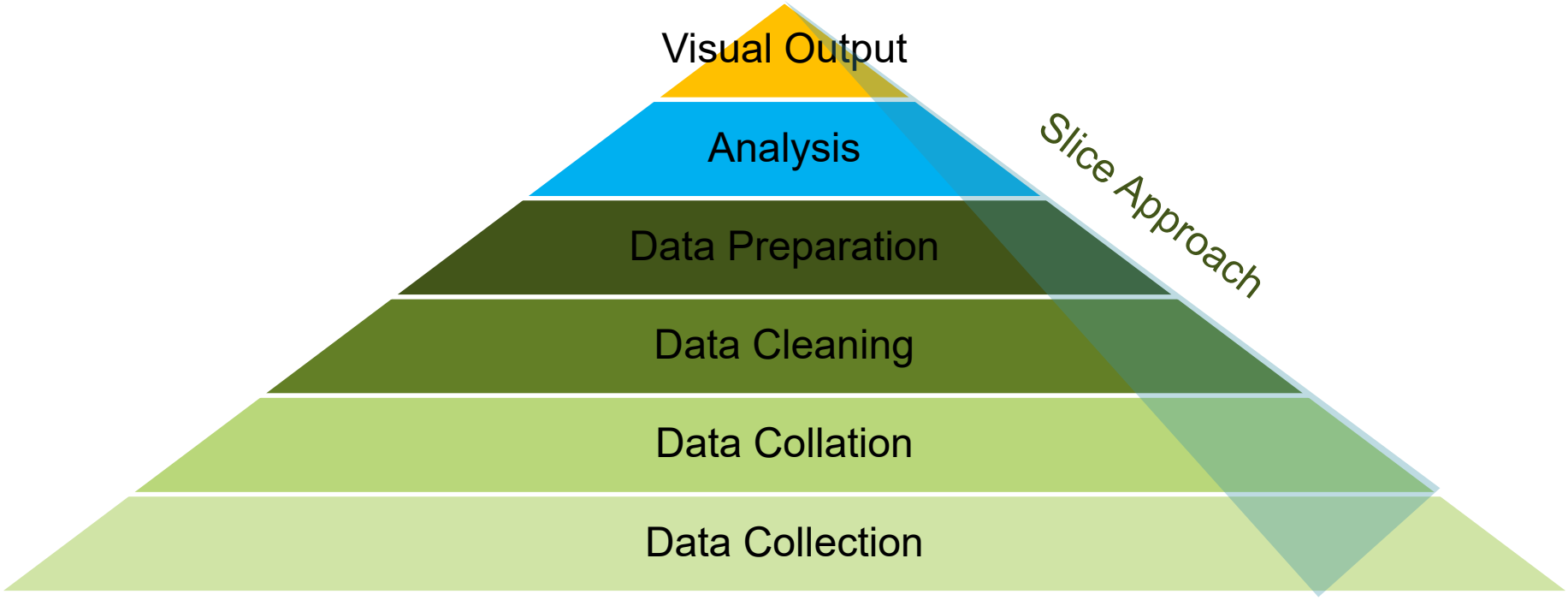
Incomplete Data Dictionary

- Sometimes Data Dictionaries are incomplete and that poses issues
- Clarifying your understanding of the data dictionary is important

- Example data dictionary with incomplete information

Field	Description	Values
pclass	Passenger Class	1,2,3 for class
survived	Whether the passenger survived or not	1:survived, 0:perished
Name	Name of the passenger	Text String
Sex	Gender of the passenger	male, female
Age	Age of the passenger	Number
Sibsp	Siblings/spouses on board	Number
Parch	??	??
Ticket	??	??
Fare	Fare paid	??
Cabin	Cabin number	Text String
Embarked	Port of embarkation	??
Boat	??	??
Body	??	??
Home.dest	Home Destination	Text String

The data value chain



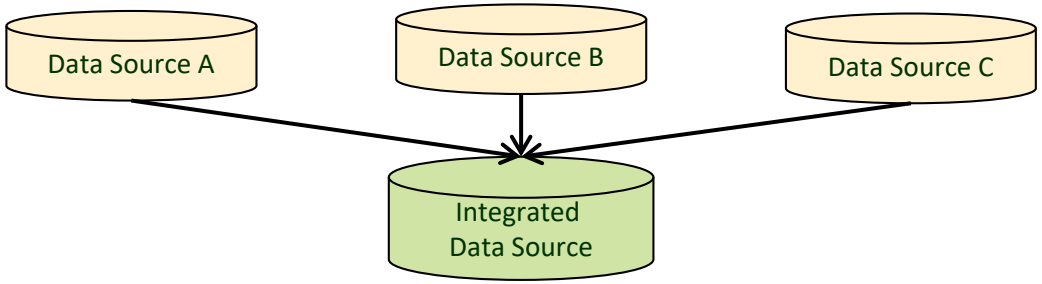
Quick wins: Slice Approach
Long term must: Horizontal scalable architecture

Data Integration

Data Analytics Best Practices

Data Integration

- Combining data from different sources into a coherent store



- Duplication & Redundancy
  - Same attribute may have different names in different databases (e.g. tenure, length of service)
  - One attribute may be derived from another in a different database (e.g. monthly and annual revenue)
  - Same user may be identified differently in different databases (John Smith vs Smith, J.)
- Inconsistency & Data Value Conflicts
  - Same attribute may occur in different databases but with different values for the same entity *(Ben is married in database1 and not in database2)*

Single View of Customer

- Benefits
  - » Have a consistent conversation with the customer across channels and products
  - » Manage X sell campaigns
- Issues
  - » The same customer would have data about her across multiple products
  - » Some of the data elements across products might not agree with each other
    - » eg: marital status depending on latest update
  - » Identifying that it is the same customer might not be trivial

Single row per customer

- A common intermediate goal of data preparation is to create a single row per customer
  - The benefit of having a single row per customer table is that it can easily be used to create models that can distinguish between customers
  - Variations include single row per transaction or single row per context where the effort is towards distinguishing between entities other than customers
  - The logical data model would help towards creating this single row per customer
  - All standardised variables in this table should over a period of time be automated into a data mart
  - Note that all data elements in this table are not static. Examples include averages across time that need a time element to be specified
- Sometimes the data could be rolled up a single row per "scorable" unit

Granularity

- An important issue to consider while rolling up data is granularity
- Choosing the right granularity should be done by looking at the implementation scenario
- Exercise:
  - An intelligent recommender system collects POS data at a retail outlet every 15 minutes
  - The company wants to enhance the recommender system by building a predictive model that can be used to plan appropriate support staff
  - What is the frequency at which the data should be rolled up?