Introduction to SIT742 - Modern Data Science

Dr Guangyan Huang
School of Information Technology
Guangyan.huang@deakin.edu.au

Tuesday, March 6th, 2018

Data Science Definition

- Data Science is the art of turning data into actions
- A data product provides actionable information without exposing decision makers to the underlying data or analytics. Examples include:
  - Movie Recommendations
  - Weather Forecasts
  - Stock Market Predictions
  - Production Process Improvements
  - Health Diagnosis
  - Flu Trend Predictions
  - Targeted Advertising

[1] Booz Allen Hamilton Inc., *The Field Guide to Data Science*, 2^{nd} Edition, 2015.

What makes Data Science Different?

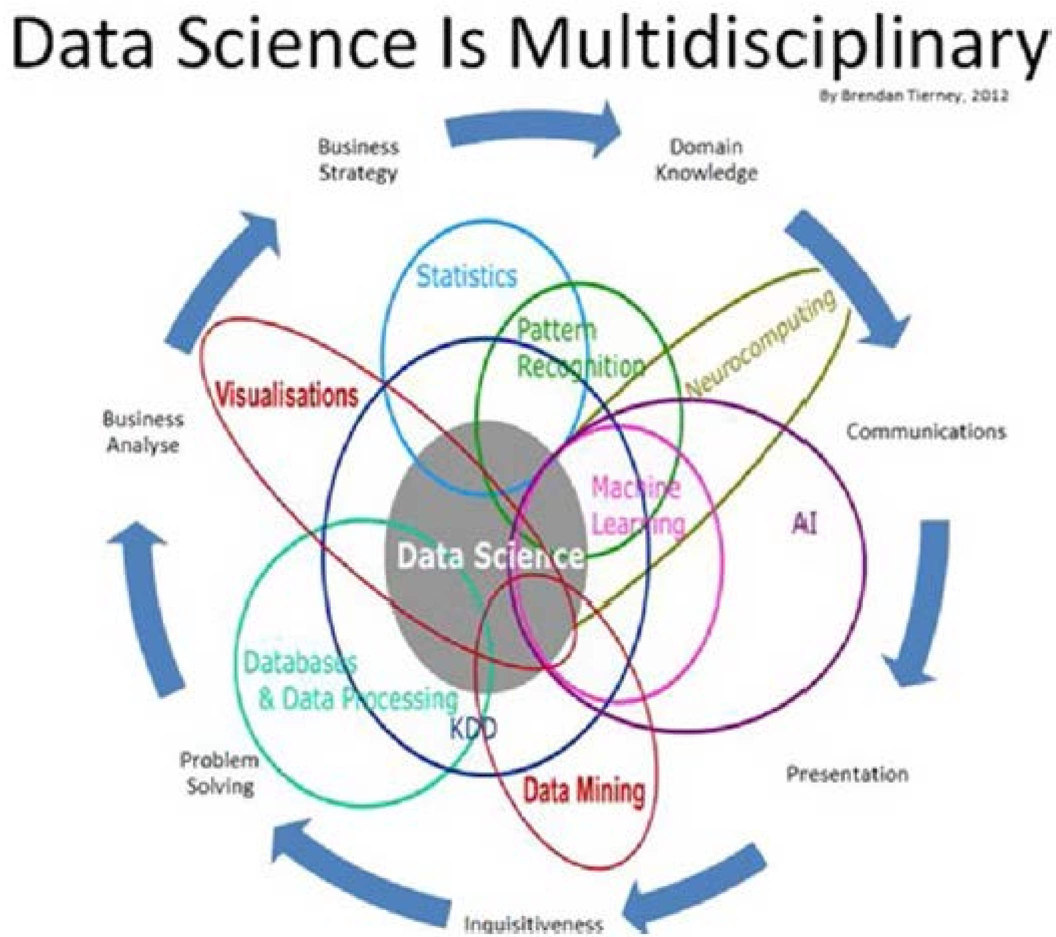
DEDUCTIVE REASONING

- Formulate hypotheses about relationships and underlying models.
- Carry out experiments with the data to test hypotheses and models.

INDUCTIVE REASONING

- Exploratory data analysis to discover or refine hypotheses.
- Discover new relationships, insights and analytic paths from the data.

Brendan Tierney's depiction of Data Science as a true multi-disciplinary field



Looking Backward and Forward

First There Was
BUSINESS INTELLIGENCE

- Deductive Reasoning
- Backward Looking
- Slice and Dice Data
- Warehoused/Siloed Data
- Analyze the Past/Guess the Future
- Creates Reports
- Analytic Output

Now We've Added
DATA SCIENCE

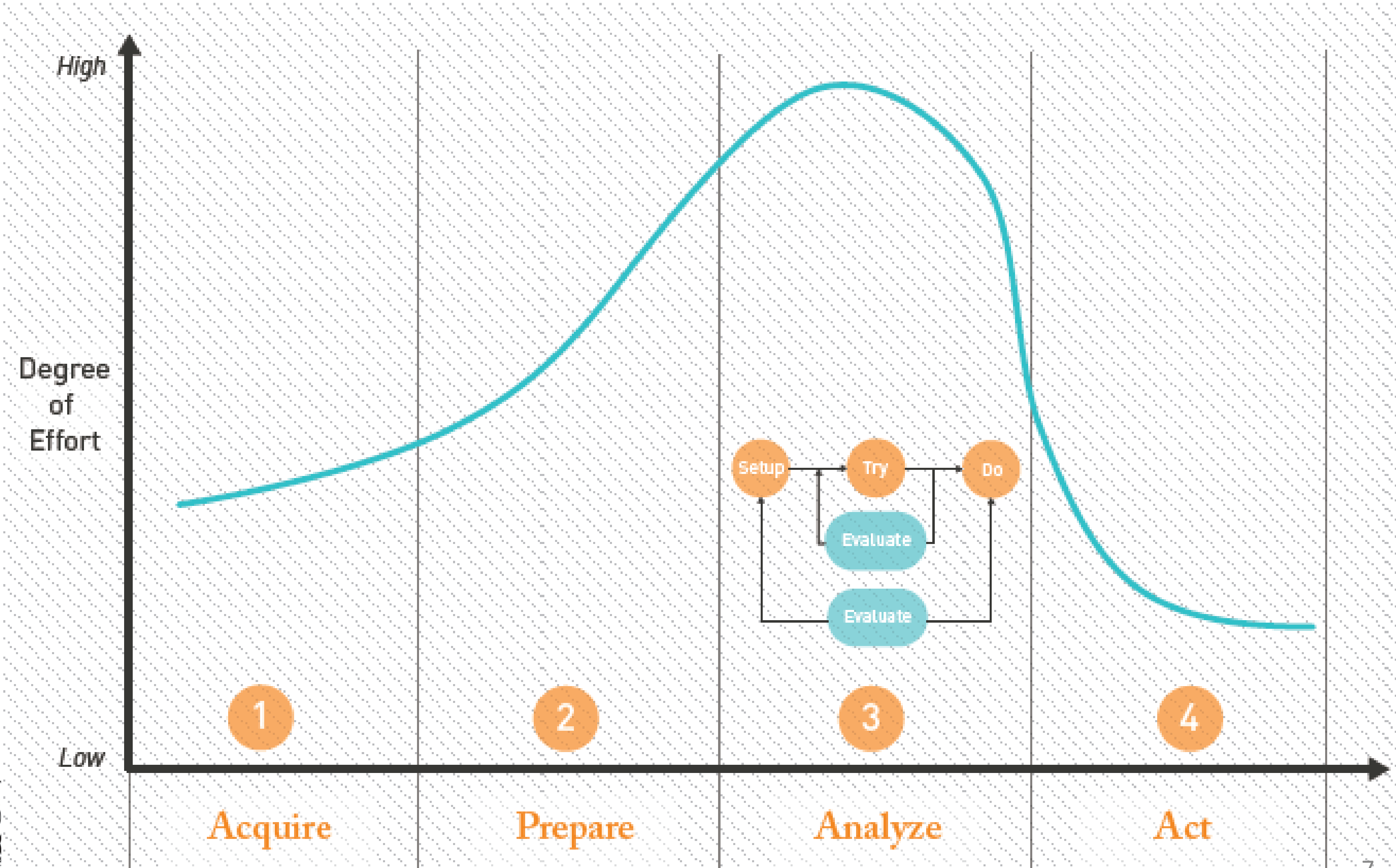
- Inductive/Deductive Reasoning
- Forward Looking
- Interact with Data
- Distributed, Real Time Data
- Predict and Advise
- Creates Data Products
- Answer Questions and Create New Ones
- Actionable Answer

What is the Impact of Data Science?

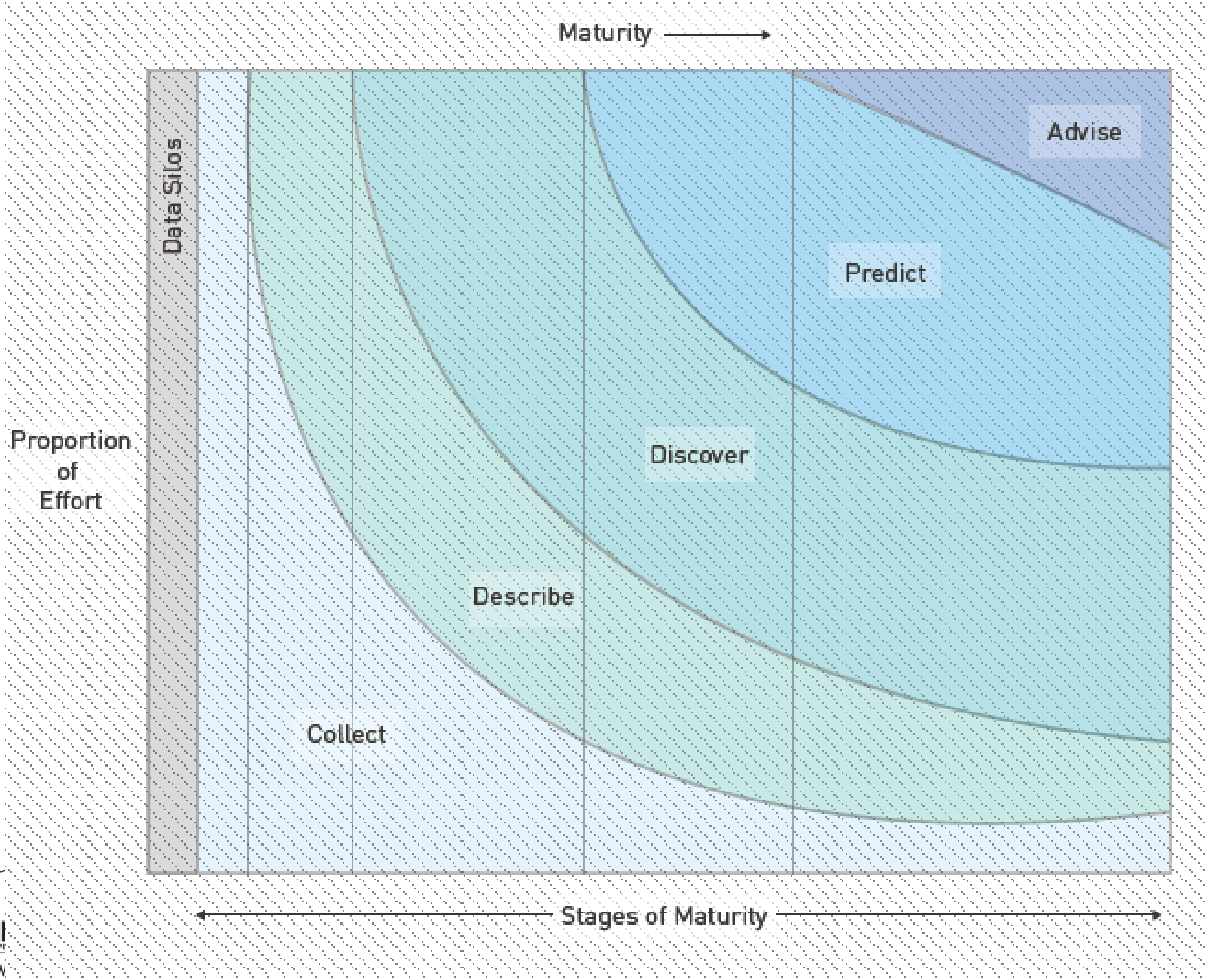
DATA SCIENCE IS NECESSARY	
17-49%	increase in productivity when organizations increase data usability by 10%
11-42%	return on assets (ROA) when organizations increase data access by 10%
241%	increase in Return on Investment (ROI) when organizations use big data to improve competitiveness
1000%	increase in ROI when deploying analytics across most of the organization, aligning daily operations with senior management's goals, and incorporating big data
5-6%	performance improvement for organizations making data-driven decisions.

DEAKIN UNIVERSITY AUSTRALIA Worldly

How does Data Science Actually Work?



The Data Science Maturity Model.



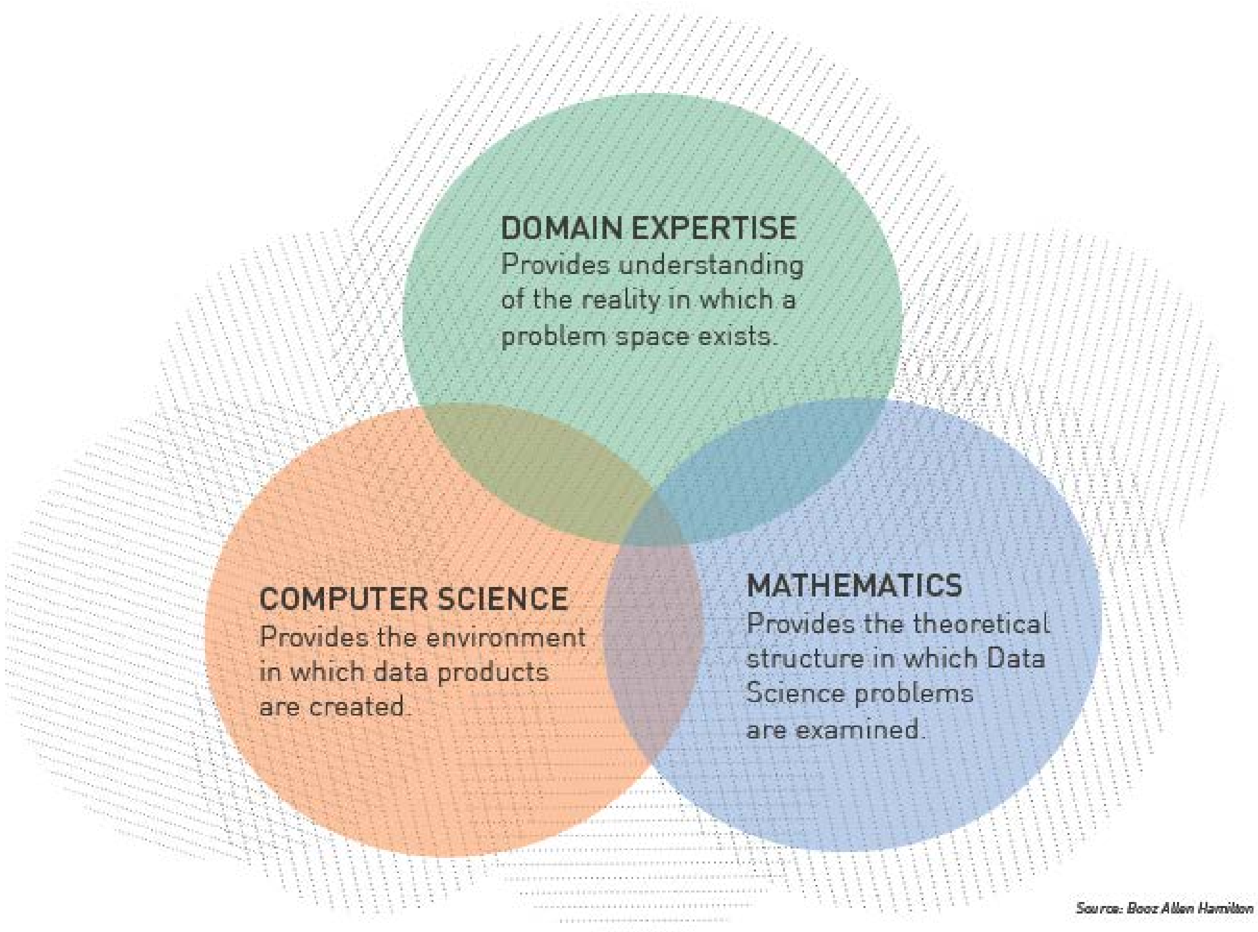
Source: Booz Allen Hamilton

The Stages of Data Science Maturity

Stage	Description	Example
Collect	Focuses on collecting internal or external datasets.	Gathering sales records and corresponding weather data.
Describe	Seeks to enhance or refine raw data as well as leverage basic analytic functions such as counts.	How are my customers distributed with respect to location, namely zip code?
Discover	Identifies hidden relationships or patterns.	Are there groups within my regular customers that purchase similarly?
Predict	Utilizes past observations to predict future observations.	Can we predict which products that certain customer groups are more likely to purchase?
Advise	Defines your possible decisions, optimizes over those decisions, and advises to use the decision that gives the best outcome.	Your advice is to target advertise to specific groups for certain products to maximize revenue.

Source: Booz Allen Hamilton

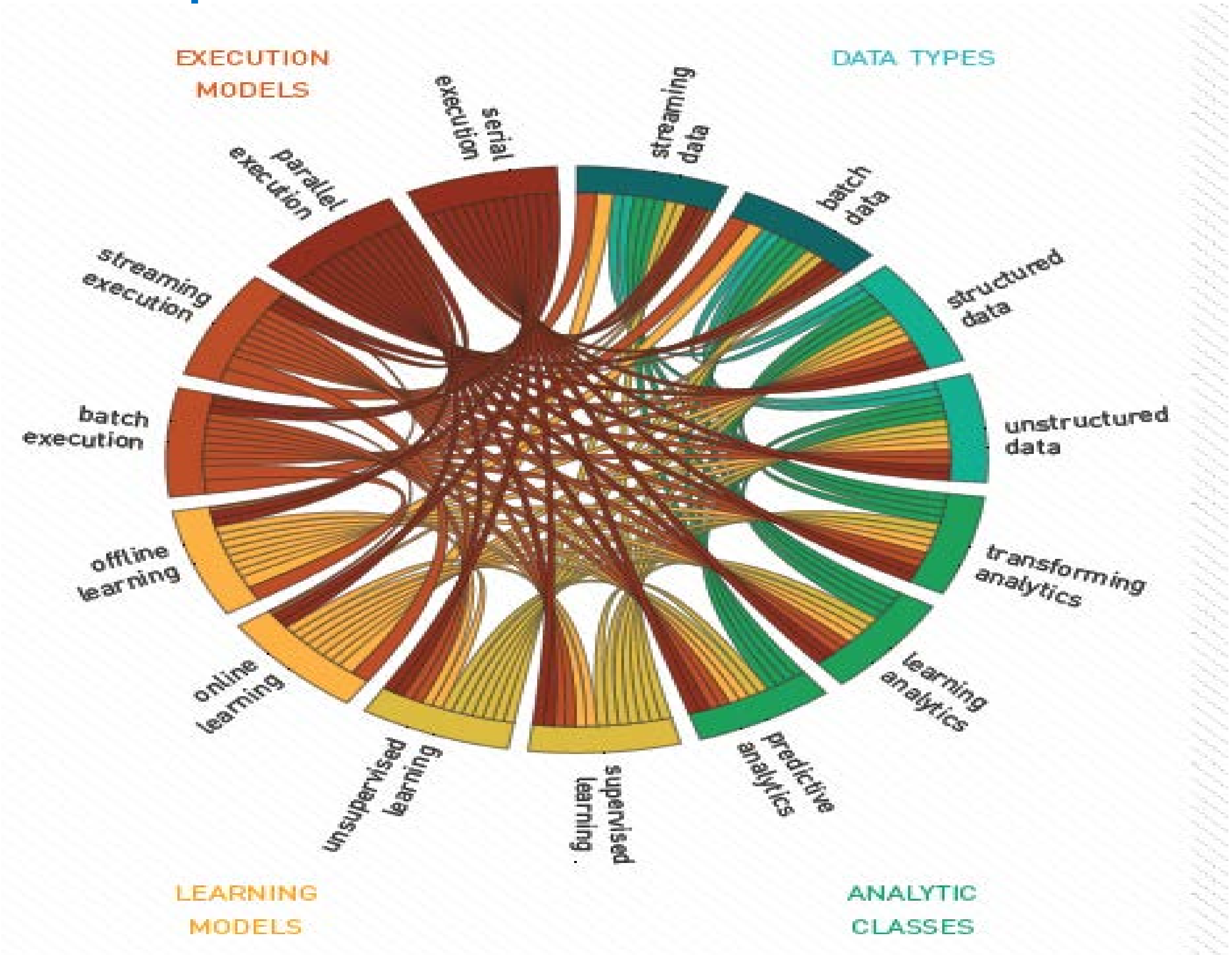
The Data Science Venn Diagram



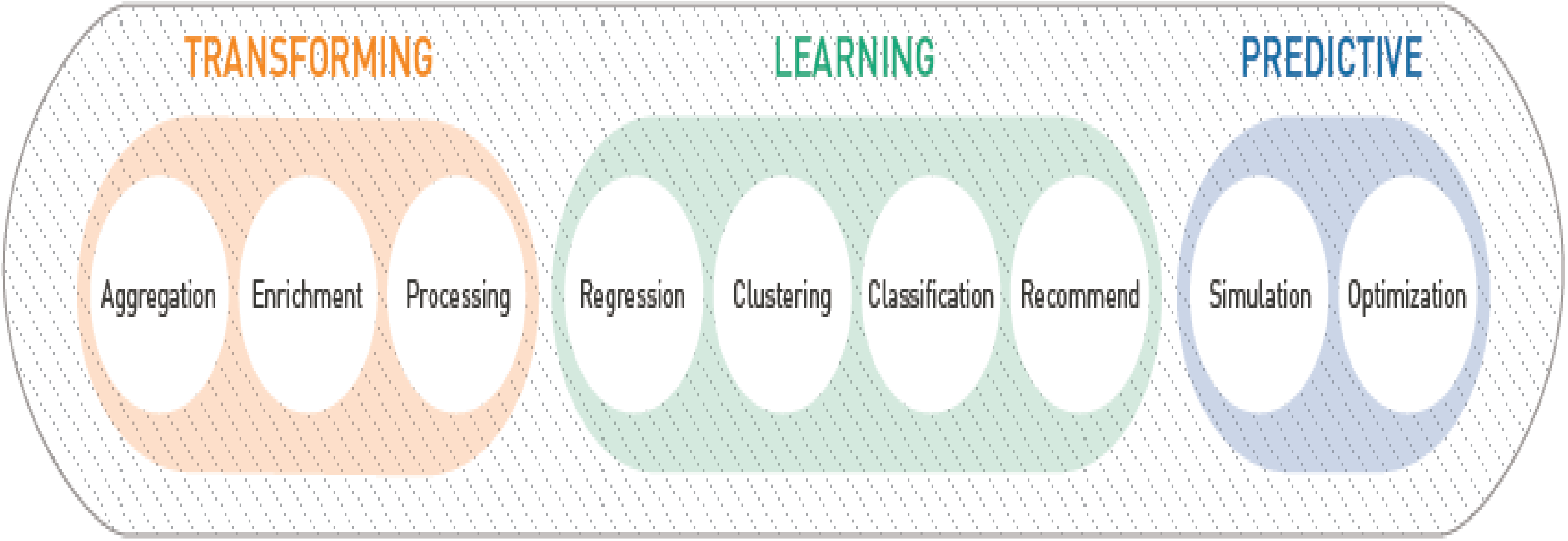
Understanding What Makes a Data Scientist

Clusters	Competencies	Description
Technical: “Knows How and What to do”	Advanced Mathematics; Computer Science; Data Mining and Integration; Database Science; Research Design; Statistical Modeling; Machine Learning; Operations Research; Programming and Scripting	the foundational technical and specialty knowledge and skills needed for successful performance in each job or role.
Data Science Consulting: “Can Do in a Client and Customer Environment”	Collaboration and Teamwork; Communications; Data Science Consulting; Ethics and Integrity	help Data Scientists easily integrate into various market or domain contexts and partner with business units to understand the environment and solve complex problems.
Cognitive: “Able to Do or Learn to Do”	Critical Thinking; Inductive and Deductive Reasoning; Problem Solving	the type of critical thinking and reasoning abilities (both inductive and deductive) a Data Scientist should have to perform their job.
Personality: “Willing or Motivated to Do”	Adaptability/Flexibility; Ambiguity Tolerance; Detail Orientation; Innovation and Creativity; Inquisitiveness; Perseverance; Resilience and Hardiness; Self-Confidence; Work Ethic	The personality traits that drive behaviors that are beneficial to Data Scientists, such as inquisitiveness, creativity, and perseverance.

Component Parts of Data Science



Classes of Analytic Techniques



Source: Booz Allen Hamilton

Transforming Analytics

- **Aggregation: Techniques to summarize the data. These** include basic statistics (e.g., mean, standard deviation), distribution fitting, and graphical plotting.
- **Enrichment: Techniques for adding additional information** to the data, such as source information or other labels.
- **Processing: Techniques that address data cleaning,** preparation, and separation. This group also includes common algorithm pre-processing activities such as transformations and feature extraction.

DEAKIN UNIVERSITY AUSTRALIA Worldly

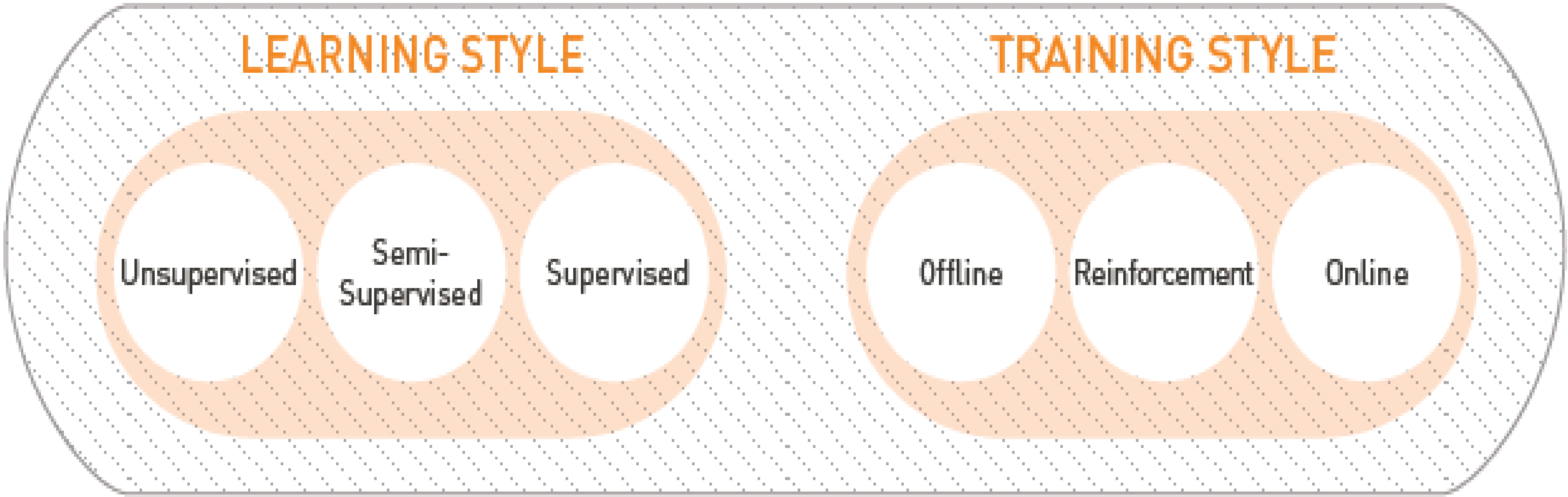
Learning Analytics

- **Regression: Techniques for estimating relationships among** variables, including understanding which variables are important in predicting future values.
- **Clustering: Techniques to segment the data into naturally** similar groups.
- **Classification: Techniques to identify data element** group membership.
- **Recommendation: Techniques to predict the rating or** preference for a new entity, based on historic preference or behavior.

Predictive Analytics

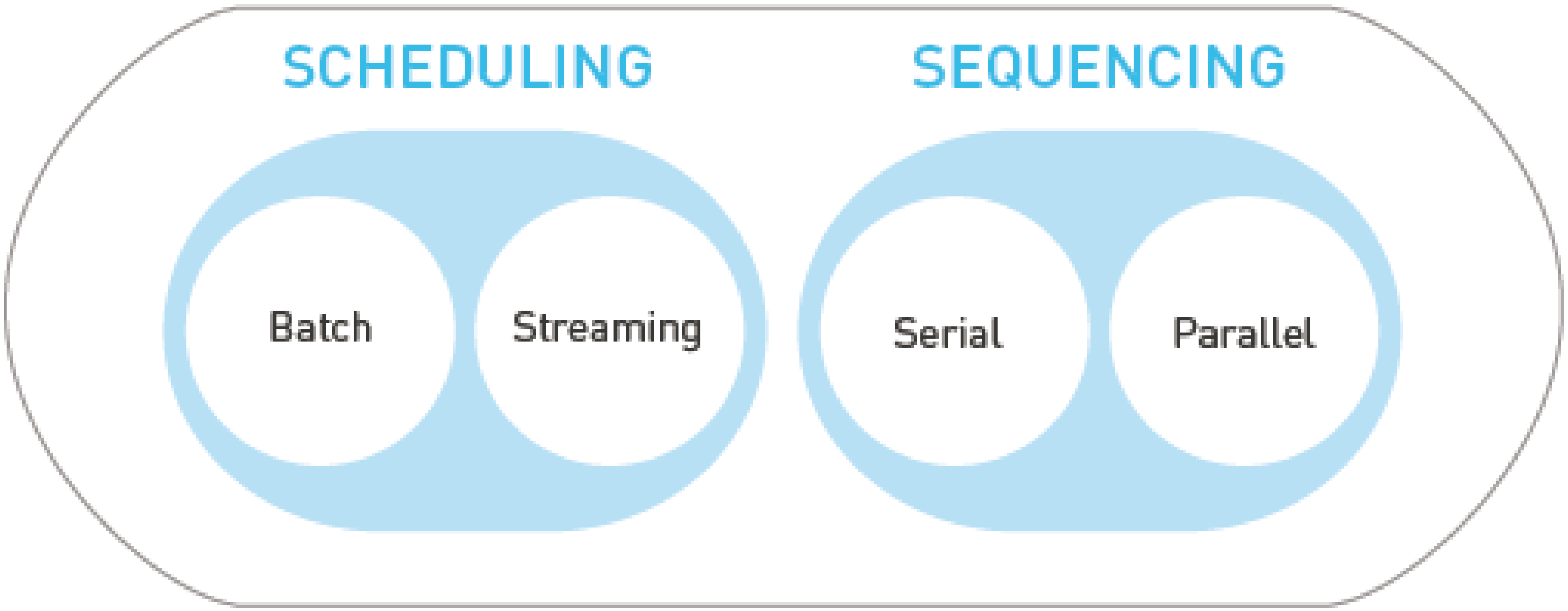
- **Simulation: Techniques to imitate the operation of a realworld** process or system. These are useful for predicting behavior under new conditions.
- **Optimization: Operations Research techniques focused on** selecting the best element from a set of available alternatives to maximize a utility function.

Analytic Learning Models



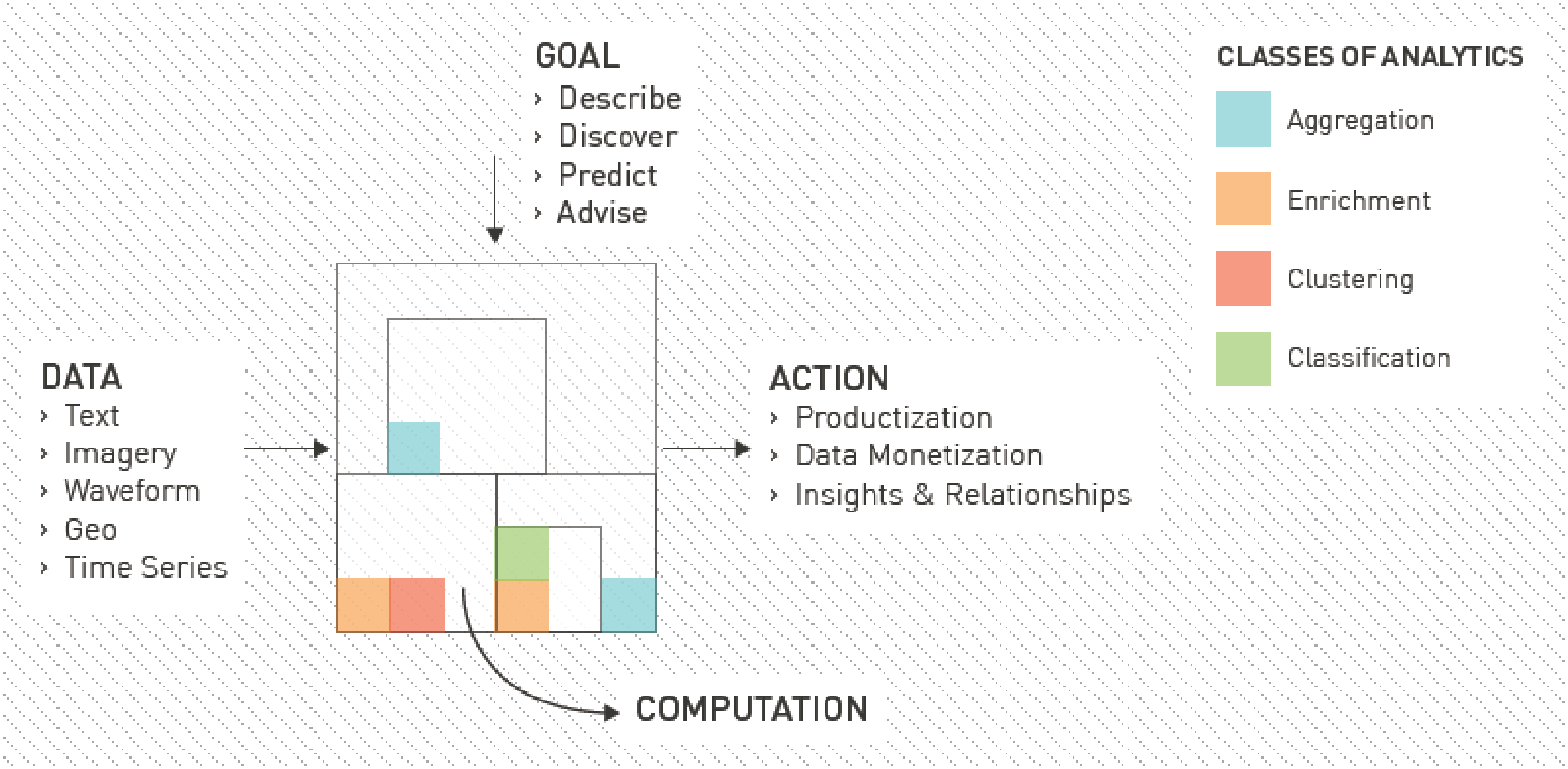
Source: Booz Allen Hamilton

Analytic Execution Models



Source: Booz Allen Hamilton

The Fractal Analytic Model



Goal

- You must first have some idea of your analytic goal and the end state of the analysis. Is it to Discover, Describe, Predict, or Advise?
- It is probably a combination of several of those. Be sure that before you start, you define the business value of the data and how you plan to use the insights to drive decisions, or risk ending up with interesting but non-actionable trivia.

DATA

- Data dictates the potential insights that analytics can provide.
  - Data Science is about finding patterns in variable data and comparing those patterns.
- If the data is not representative of the universe of events you wish to analyze,
  - you will want to collect that data through carefully planned variations in events or processes through A/B testing or design of experiments.
- Datasets are never perfect so don't wait for perfect data to get started.
  - A good Data Scientist is adept at handling messy data with missing or erroneous values. Just make sure to spend the time upfront to clean the data or risk generating garbage results.

DEAKIN UNIVERSITY AUSTRALIA Worldly

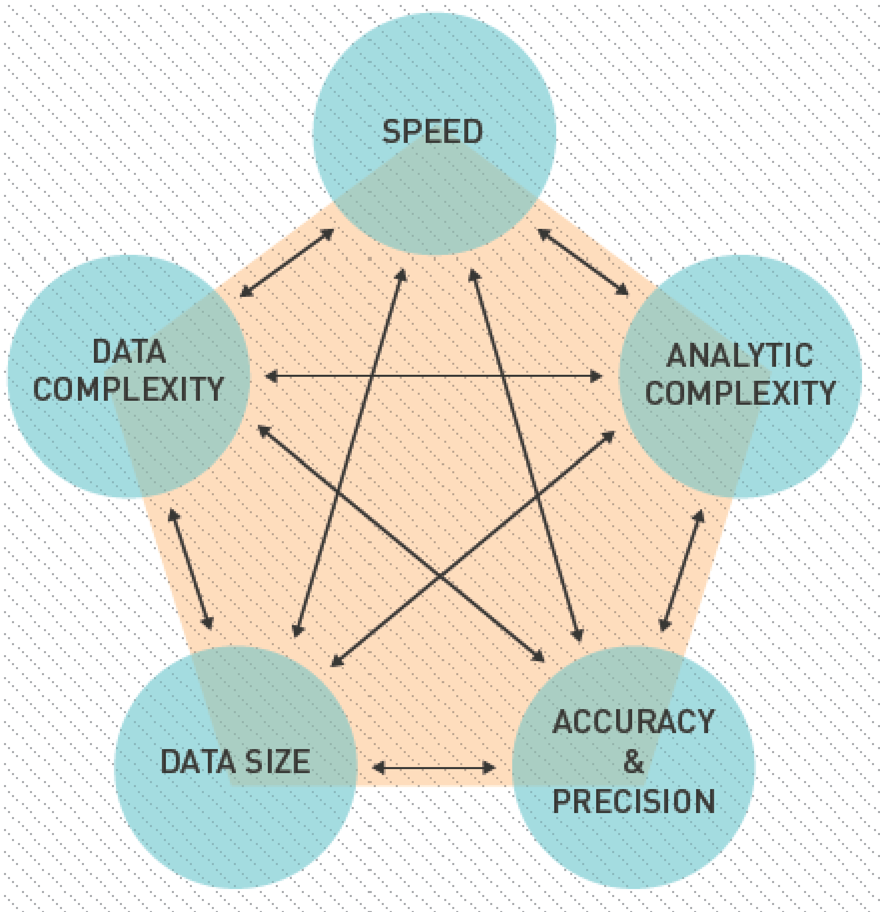
COMPUTATION

- Computation aligns the data to goals through the process of creating insights. Through divide and conquer, computation decomposes into several smaller analytic capabilities with their own goals, data, computation and resulting actions, just like a smaller piece of broccoli maintains the structure of the original stalk.
- In this way, computation itself is fractal. Capability building blocks may utilize different types of execution models such as batch computation or streaming, that individually accomplish small tasks. When properly combined together, the small tasks produce complex, actionable results.

ACTION

- How should engineers change the manufacturing process to generate higher product yield? How should an insurance company choose which policies to offer to whom and at what price?
- The output of computation should enable actions that align to the goals of the data product. Results that do not support or inspire action are nothing but interesting trivia.

Balancing the Five Analytic Dimensions



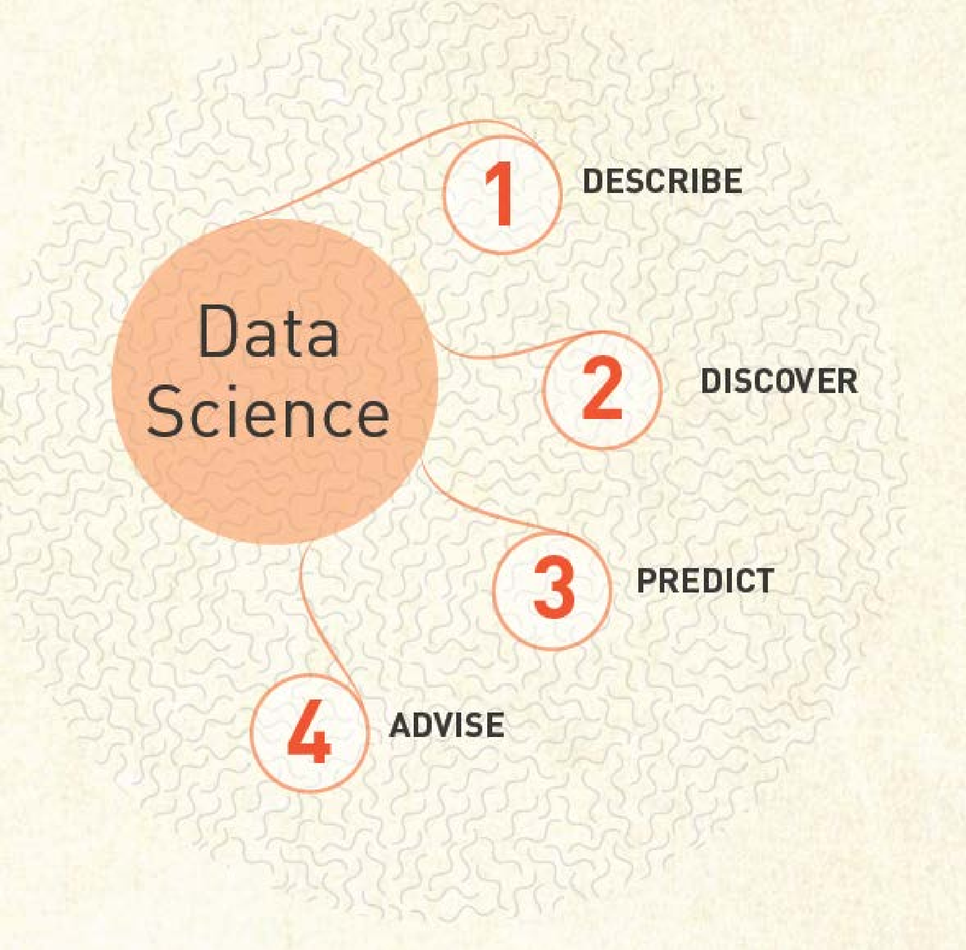
SPEED: The speed at which an analytic outcome must be produced (e.g., near real-time, hourly, daily) or the time it takes to develop and implement the analytic solution

ANALYTIC COMPLEXITY: Algorithmic complexity (e.g., complexity class and execution resources)

ACCURACY & PRECISION: The ability to produce exact versus approximate solutions as well as the ability to provide a measure of confidence.

DATA COMPLEXITY: The data type, formal complexity measures including measures of overlap and linear separability, number of dimensions /columns, and linkages between datasets

Guide to Analytic Selection

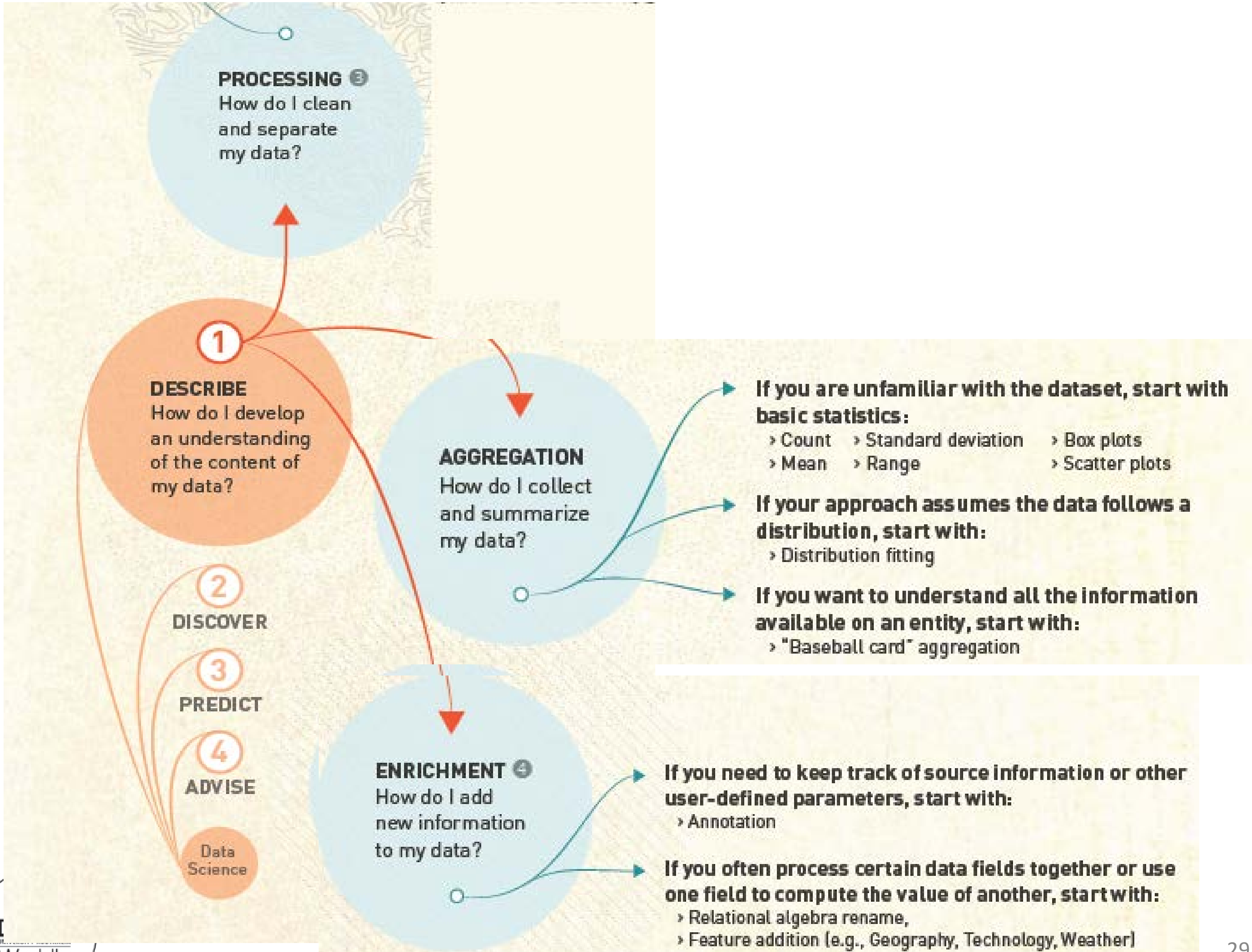


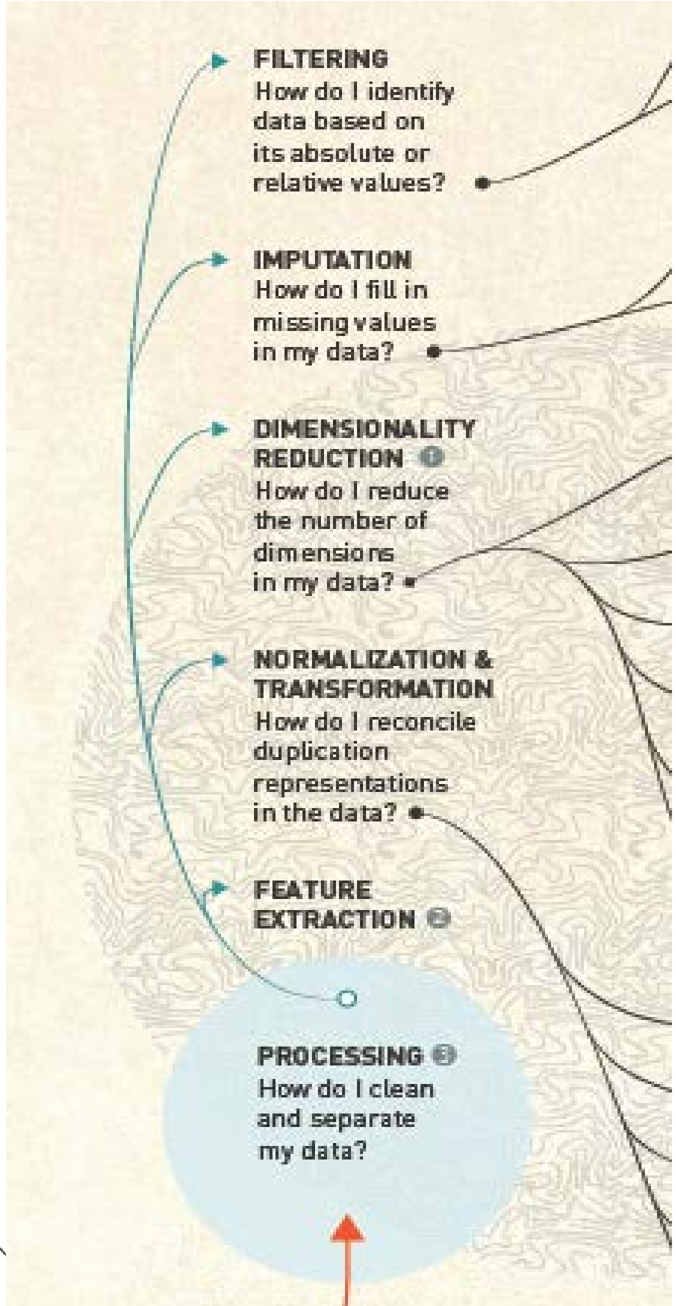
- **There are several situations where dimensionality reduction may be needed**:
  - Models fail to converge,
  - Models produce results equivalent to random chance,
  - You lack the computational power to perform operations across the feature space,
  - You do not know which aspects of the data are the most important.

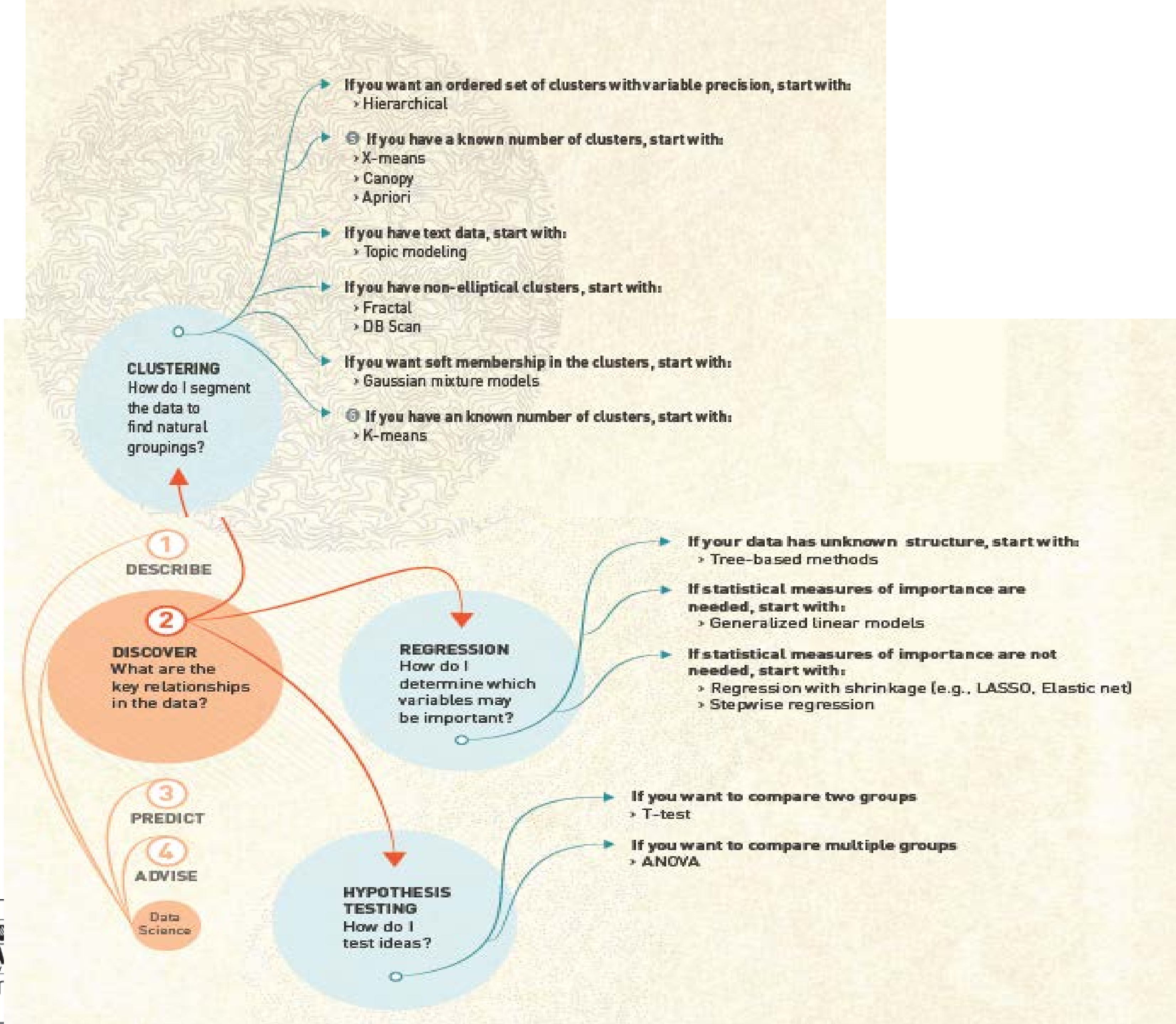
- **Feature Extraction is a broad topic and is highly dependent upon the domain area.**
  - This topic could be the subject of an entire book. As a result, a detailed exploration has been omitted from this diagram. See the *Featuring Engineering and Feature Selection sections in the Life in the Trenches chapter for additional information.*

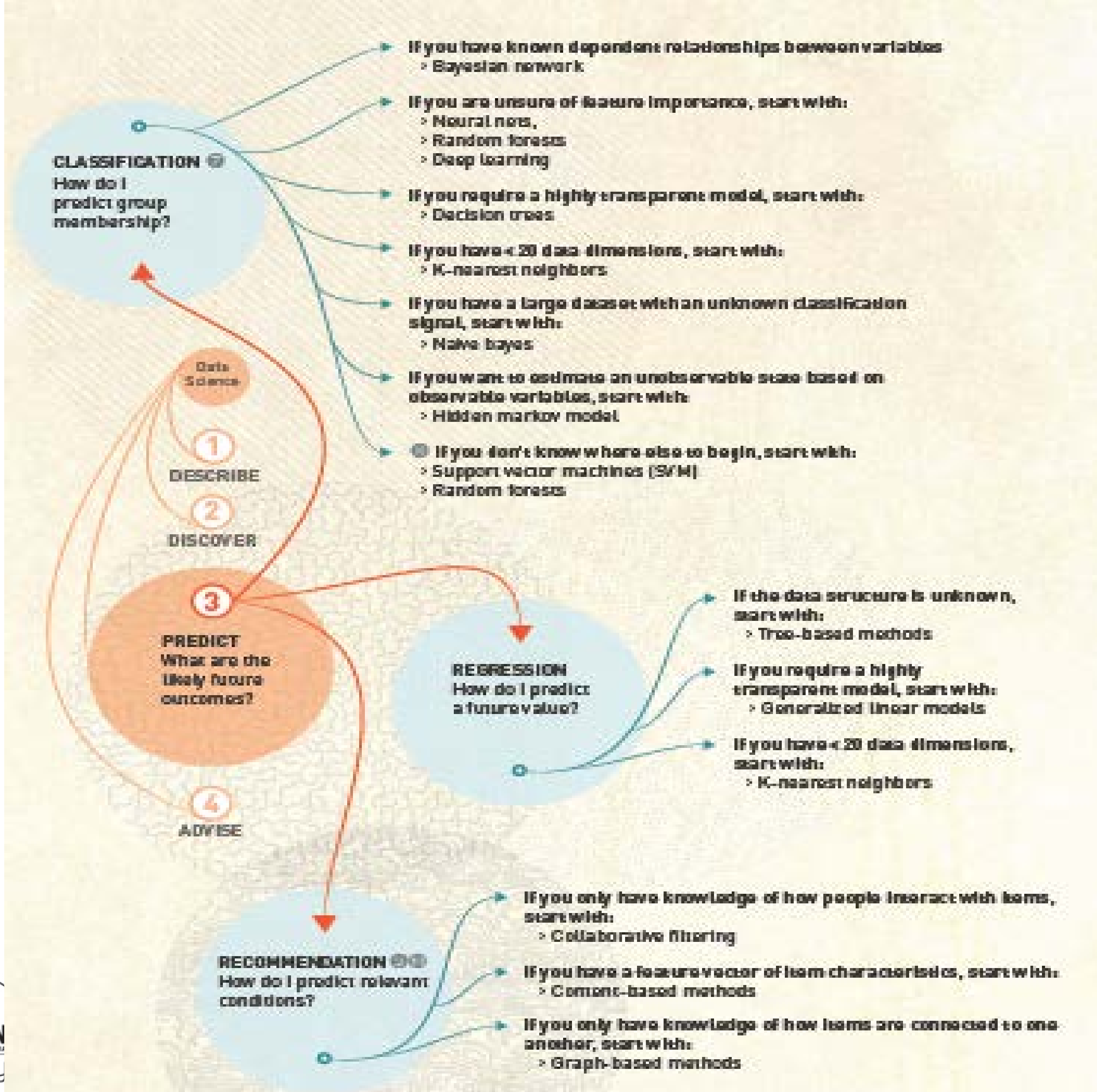
DEAKIN UNIVERSITY AUSTRALIA Worldly

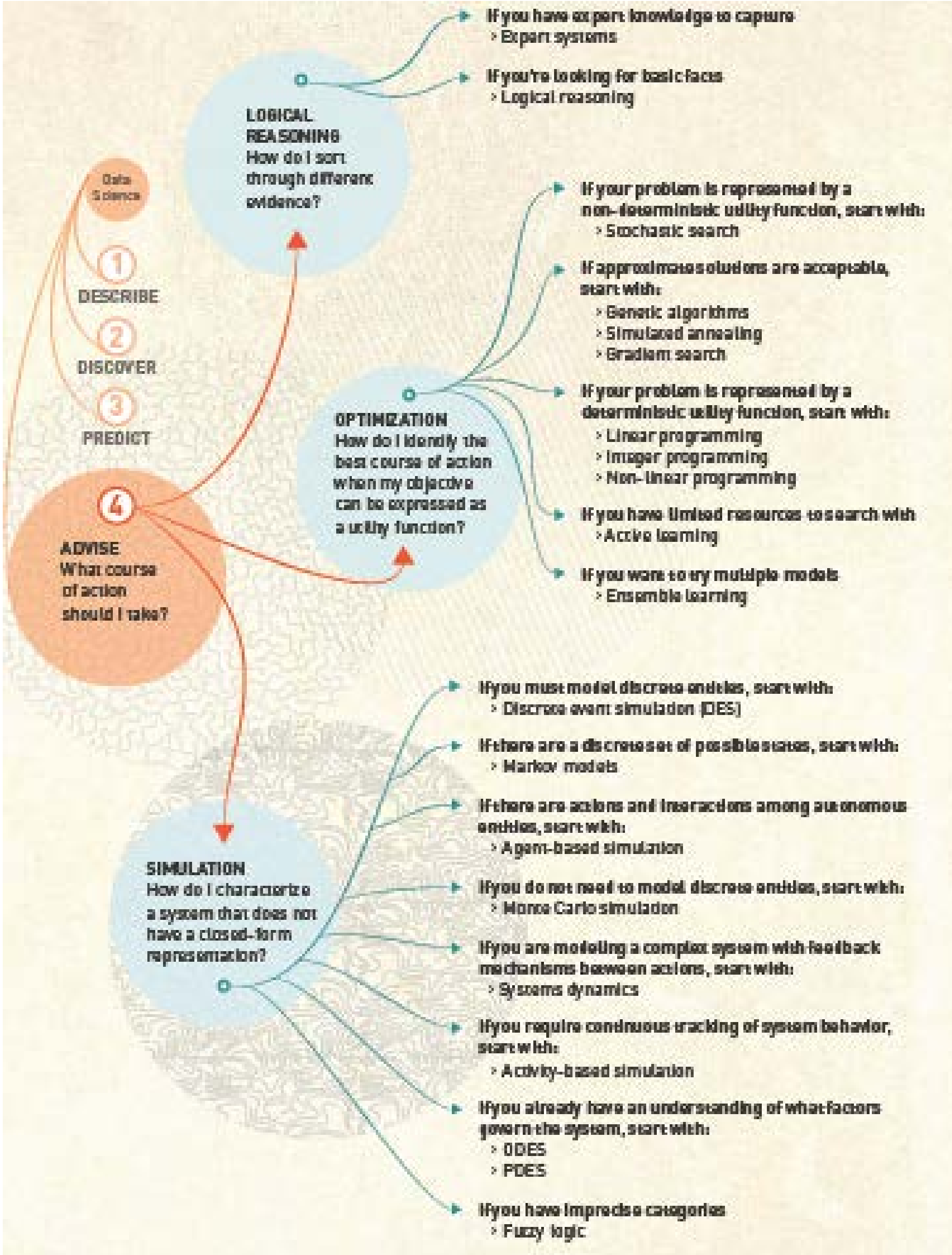
- **Always check data labels for correctness**. This is particularly true for time stamps, which may have reverted to system default values.
- **Smart enrichment can greatly speed-up computational time**. It can also be a huge differentiator between the accuracy of different end-to-end analytic solutions.


PROCESSING 3
How do I clean and separate my data?
1
DESCRIBE
How do I develop an understanding of the content of my data?
2
DISCOVER
3
PREDICT
4
ADVISE
Data Science
AGGREGATION
How do I collect and summarize my data?
If you are unfamiliar with the dataset, start with basic statistics:
› Count › Standard deviation › Box plots
› Mean › Range › Scatter plots
If your approach assumes the data follows a distribution, start with:
› Distribution fitting
If you want to understand all the information available on an entity, start with:
› "Baseball card" aggregation
ENRICHMENT 4
How do I add new information to my data?
If you need to keep track of source information or other user-defined parameters, start with:
› Annotation
If you often process certain data fields together or use one field to compute the value of another, start with:
› Relational algebra rename,
› Feature addition (e.g., Geography, Technology, Weather)


FILTERING
How do I identify data based on its absolute or relative values?
IMPUTATION
How do I fill in missing values in my data?
DIMENSIONALITY REDUCTION
How do I reduce the number of dimensions in my data?
NORMALIZATION & TRANSFORMATION
How do I reconcile duplication representations in the data?
FEATURE EXTRACTION
PROCESSING
How do I clean and separate my data?

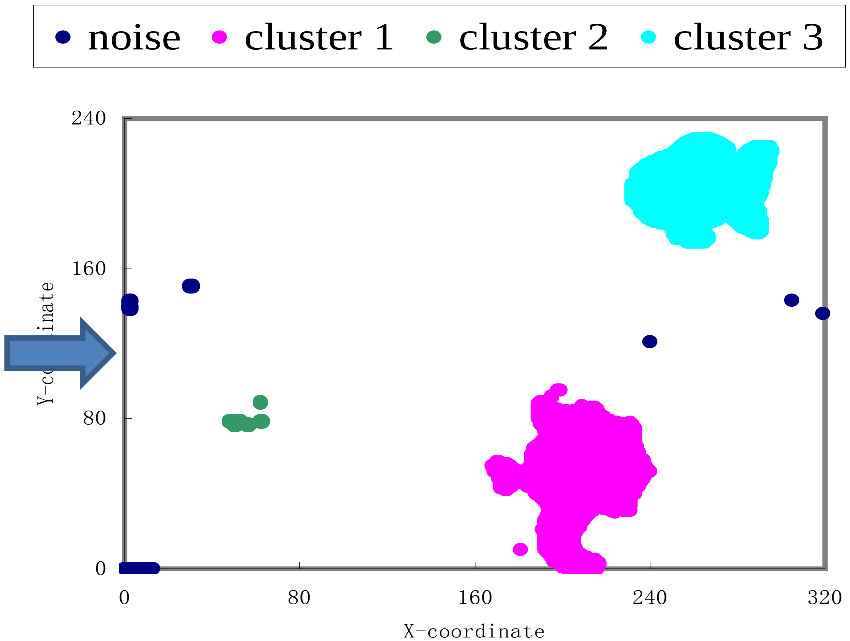

If you want an ordered set of clusters with variable precision, start with:
› Hierarchical
If you have a known number of clusters, start with:
› X-means
› Canopy
› Apriori
If you have text data, start with:
› Topic modeling
If you have non-elliptical clusters, start with:
› Fractal
› DB Scan
If you want soft membership in the clusters, start with:
› Gaussian mixture models
If you have an known number of clusters, start with:
› K-means
CLUSTERING
How do I segment the data to find natural groupings?
1
DESCRIBE
2
DISCOVER
What are the key relationships in the data?
3
PREDICT
4
ADVISE
Data Science
REGRESSION
How do I determine which variables may be important?
If your data has unknown structure, start with:
› Tree-based methods
If statistical measures of importance are needed, start with:
› Generalized linear models
If statistical measures of importance are not needed, start with:
› Regression with shrinkage (e.g., LASSO, Elastic net)
› Stepwise regression
HYPOTHESIS TESTING
How do I test ideas?
If you want to compare two groups
› T-test
If you want to compare multiple groups
› ANOVA


If you have known dependent relationships between variables
> Bayesian network
If you are unsure of feature importance, start with:
> Neural nets,
> Random forests
> Deep learning
CLASSIFICATION
How do I predict group membership?
If you require a highly transparent model, start with:
> Decision trees
If you have < 20 data dimensions, start with:
> K-nearest neighbors
If you have a large dataset with an unknown classification signal, start with:
> Naive bayes
If you want to estimate an unobservable state based on observable variables, start with:
> Hidden markov model
If you don't know where else to begin, start with:
> Support vector machines (SVM)
> Random forests
Data Science
1
DESCRIBE
2
DISCOVER
3
PREDICT
What are the likely future outcomes?
4
ADVISE
REGRESSION
How do I predict a future value?
If the data structure is unknown, start with:
> Tree-based methods
If you require a highly transparent model, start with:
> Generalized linear models
If you have < 20 data dimensions, start with:
> K-nearest neighbors
RECOMMENDATION
How do I predict relevant conditions?
If you only have knowledge of how people interact with items, start with:
> Collaborative filtering
If you have a feature vector of item characteristics, start with:
> Content-based methods
If you only have knowledge of how items are connected to one another, start with:
> Graph-based methods


Data Science
1 DESCRIBE
2 DISCOVER
3 PREDICT
4 ADVISE What course of action should I take?
LOGICAL REASONING How do I sort through different evidence?
If you have expert knowledge to capture
› Expert systems
If you're looking for basic facts
› Logical reasoning
OPTIMIZATION How do I identify the best course of action when my objective can be expressed as a utility function?
If your problem is represented by a non-deterministic utility function, start with:
› Stochastic search
If approximate solutions are acceptable, start with:
› Genetic algorithms
› Simulated annealing
› Gradient search
If your problem is represented by a deterministic utility function, start with:
› Linear programming
› Integer programming
› Non-linear programming
If you have limited resources to search with
› Active learning
If you want to try multiple models
› Ensemble learning
SIMULATION How do I characterize a system that does not have a closed-form representation?
If you must model discrete entities, start with:
› Discrete event simulation (DES)
If there are a discrete set of possible states, start with:
› Markov models
If there are actions and interactions among autonomous entities, start with:
› Agent-based simulation
If you do not need to model discrete entities, start with:
› Monte Carlo simulation
If you are modeling a complex system with feedback mechanisms between actions, start with:
› Systems dynamics
If you require continuous tracking of system behavior, start with:
› Activity-based simulation
If you already have an understanding of what factors govern the system, start with:
› ODES
› PDES
If you have imprecise categories
› Fuzzy logic

Clustering: Discovery of Common Patterns

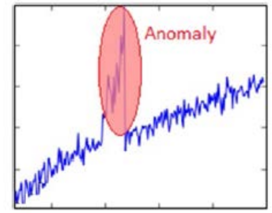
Clustering (e.g., DBSCAN algorithm)

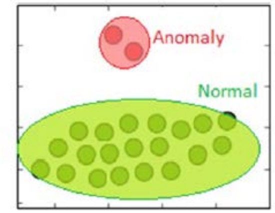


G. Huang, Jing He and Zhiming Ding, *International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

DEAKIN UNIVERSITY AUSTRALIA Worldly

Unsupervised Learning: Anomaly/Outlier Detection

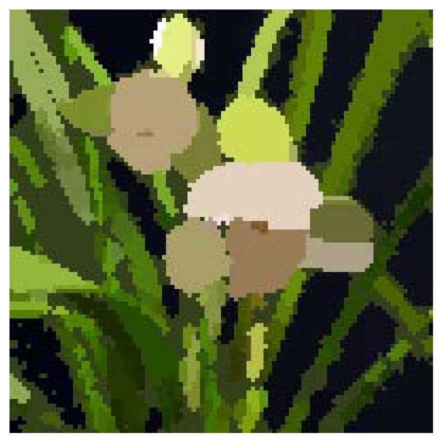




- Motivation:
  - Automatically generate traditional style painting from photos taken by camera
  - By product: Noises of DBSCAN for Pencil Sketch

byproduct

(a) Photo



(b) Watercolor

(c) Pencil Sketch

G. Huang, et al, *ICDM Workshop'2007*

Recommendation: Singular Value Decomposition (SVD)

Original rating matrix	Titanic	WALL·E	X-Men	Avatar
Tom	5	3	?	?
John	2	4	5	?

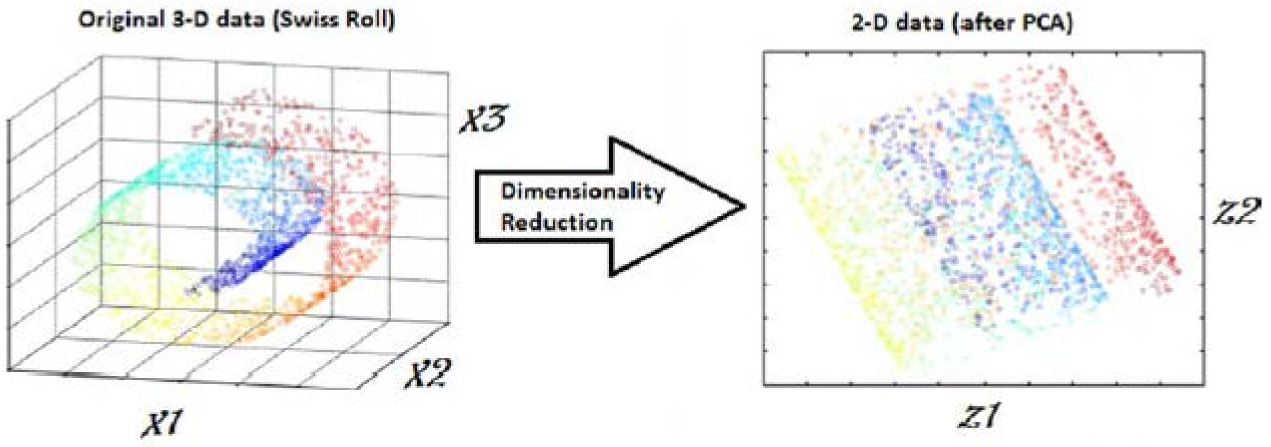
User preference vector	Romantic	Sci-Fi
Tom	1	0.1
John	0.2	1

Prediction Using Incremental ApproSVD algorithm

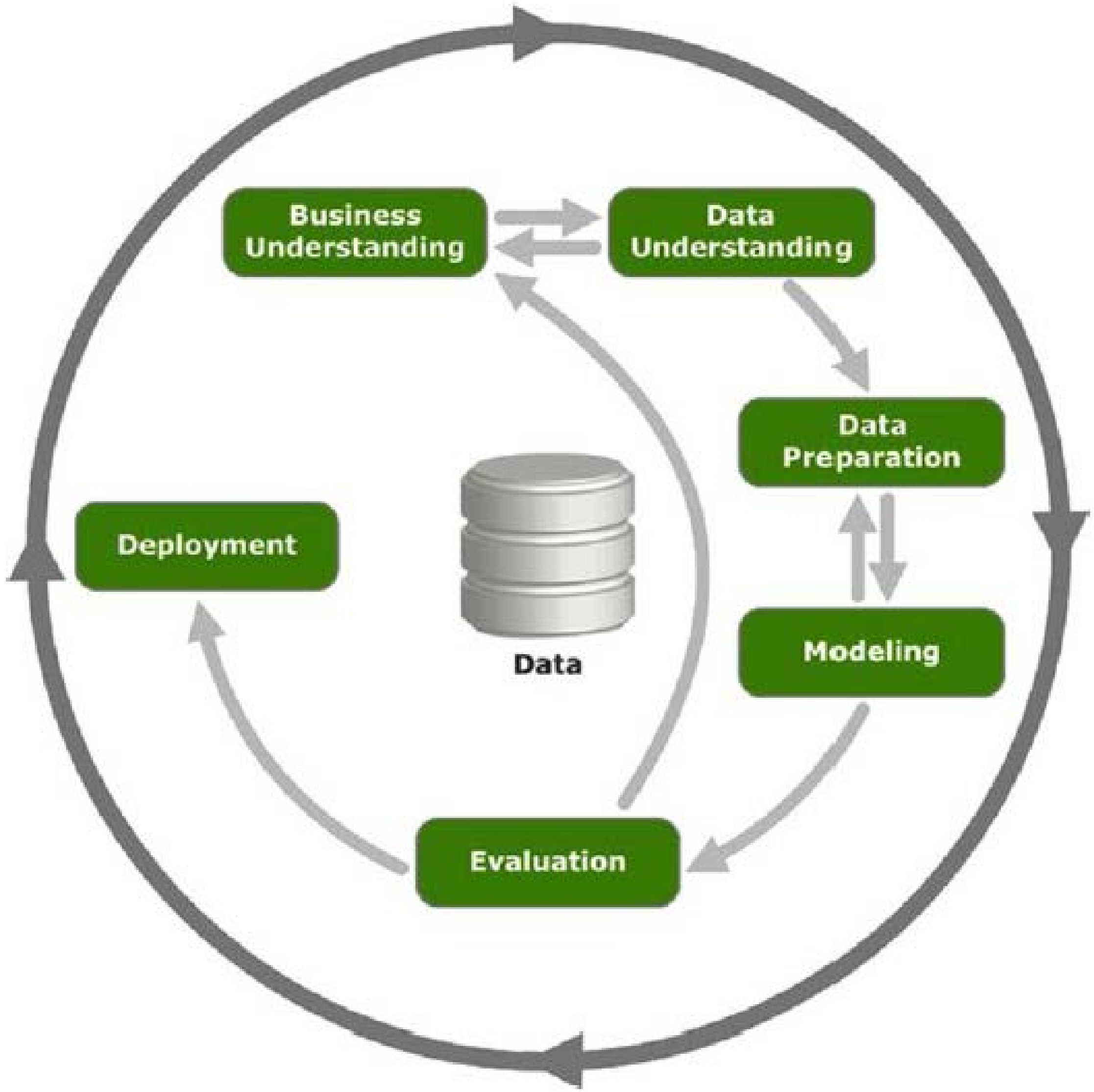
Predicted rating matrix	Titanic	WALL·E	Avatar
Tom	5	3.3	2.4
John	1	3.6	4.4

X. Zhou, J. He, G. Huang, Y. Zhang, SVD-Based Incremental Approaches for Recommender Systems, *Journal of Computer and System Sciences*, 81(4): 717-733, 2015.

Dimensionality Reduction



Data Mining Lifecycle

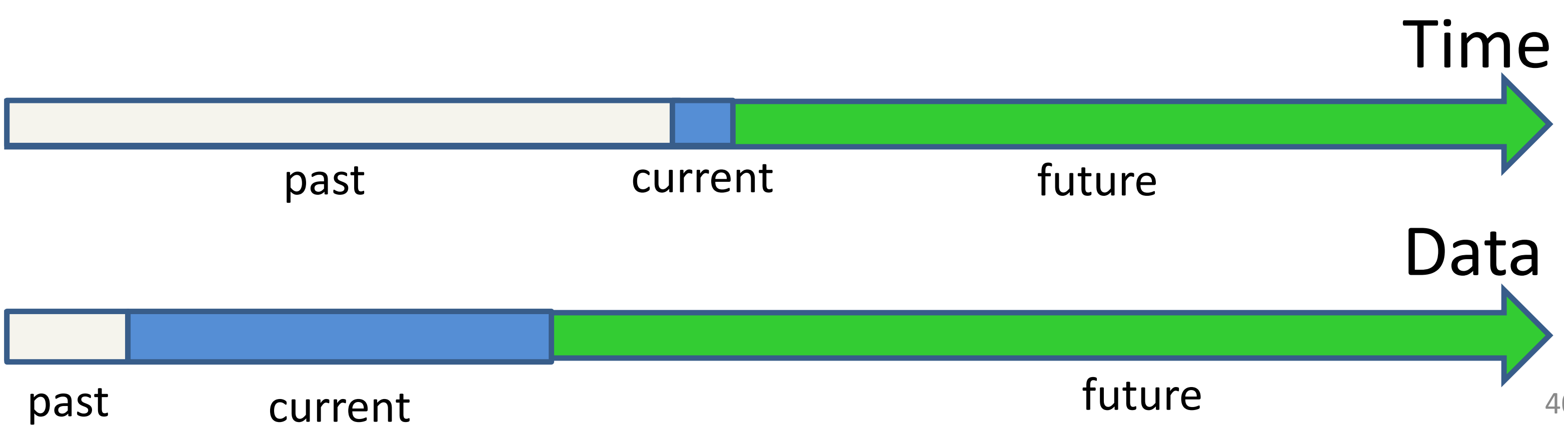


Big Stream Data

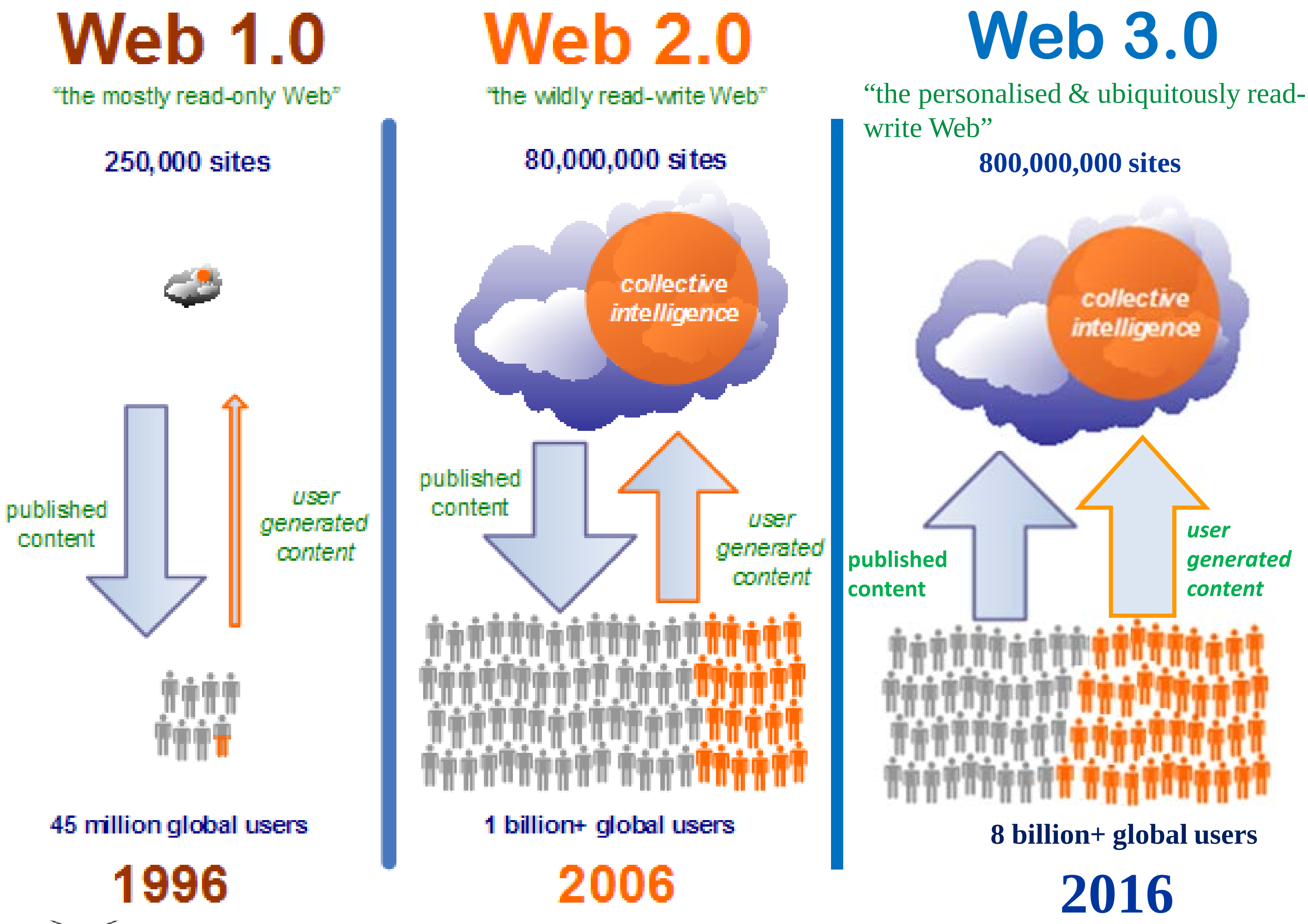
- Exponential Growing of Data
- The Development of Web
- Individual Engagement Powered by Mobile and Social Technologies
- “Data is becoming the world’s new natural resource”

Exponential Growing of Data

- Data generated in a way exceeding human limits to use them (S. H. Muggleton, *Nature*, 2006).
  - 90% of the data in the world was created in the past few years alone; ‘‘each of today’s cloud datacenters contains more computing and storage capacity than the entire Internet did just a few years ago’’ (D. A. Reed, D. B. Gannon and J. R. Larus, *Computer*, 2012).
  - The amount of data is doubling every year in every science domain (A. Szalay and J. Gray, Nature, 2006).



The Development of Web



Individual Engagement Powered by Mobile and Social Technologies

Sensor (stream) data

Text (stream) data

Sensor Generated (microphone, camera, medical sensors …)

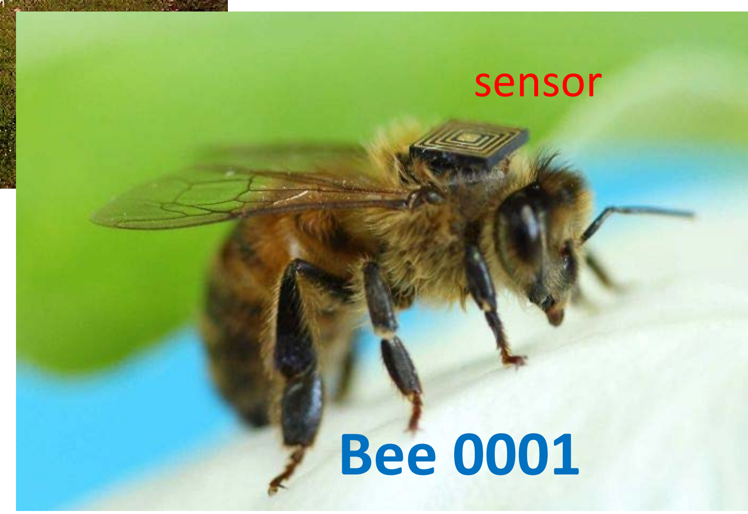
User Generated

"Data is becoming the world's new natural resource" (- IBM 2013 Annual report)

- Stream data: everything at every moment is becoming the history, recorded as a data object with a timestamp
  - All things, being recorded, are naturally buried in stream data.
- Value of stream data, which are continuously recorded phenomena , for example:
  - Natural phenomena (e.g., the change of a river's water quality)
  - Human-related phenomena
    - Individual health status (e.g., ECG curves)
    - Individual behaviors (e.g., GPS trajectories) (bee behaviors)
    - Individual thoughts (e.g. streams of text in blog, micro-blog, short messages)
    - Social events (e.g., recorded in daily news)

DEAKIN UNIVERSITY AUSTRALIA Worldly

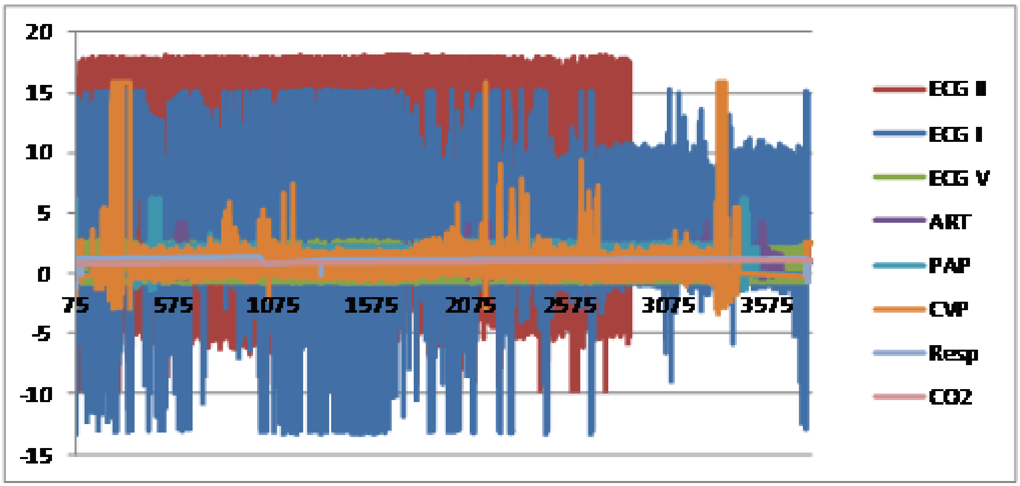
Modeling of Bee Behavior Using Sensor Data (collaborated with CSIRO)



DEAKIN
UNIVERSITY AUSTRALIA
Worldly

Big Data Application 1 – Time Series Data

- Medical Sensor Data Streams (1 patient, 1 hour 3 minutes, 8 physiological parameters)

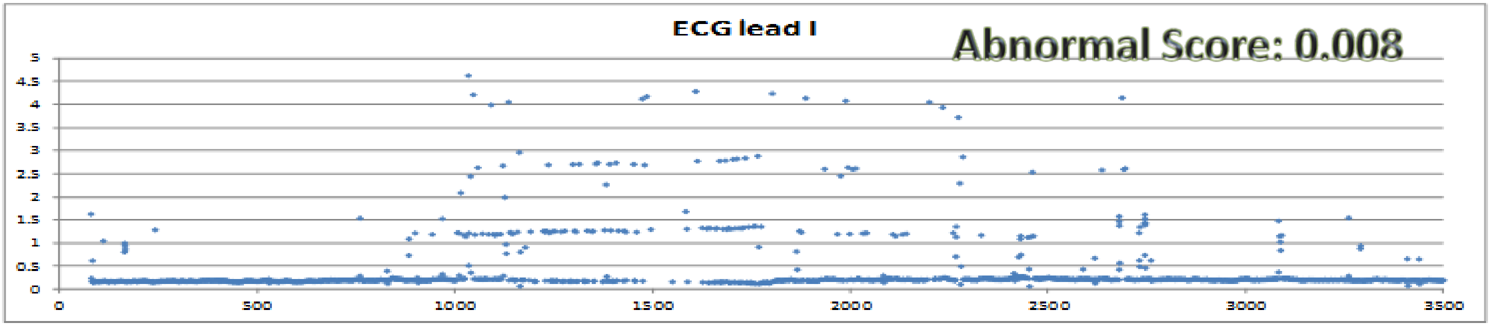


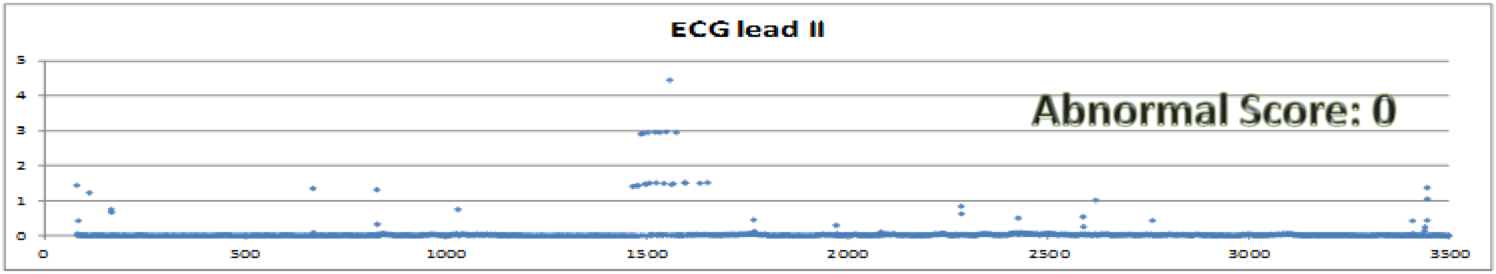
Normalized values by dividing means.

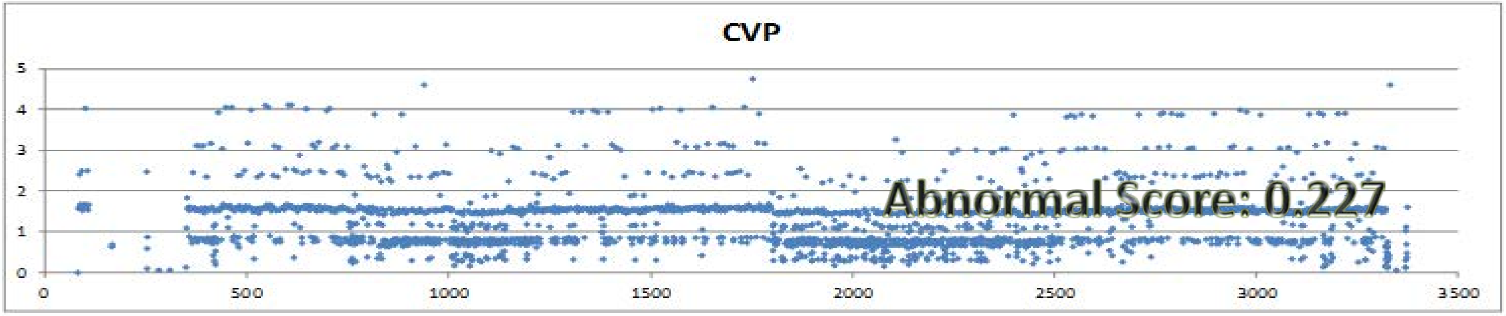
•Excel 2007 supports 1024^2=1,048,576 rows and 32,000 points for one curve
•one patient data stream, sampled once every 3ms, 1,300,000 points (83MB).

•three ECG leads (ECGs I, II and V),
•arterial pressure (ART),
•pulmonary arterial pressure (PAP),
•central venous pressure (CVP),
•respiratory impedance (Resp) and
•air way CO2 waveform (CO2).

Data Product – Data Deviation Distribution for Disease Diagnosis

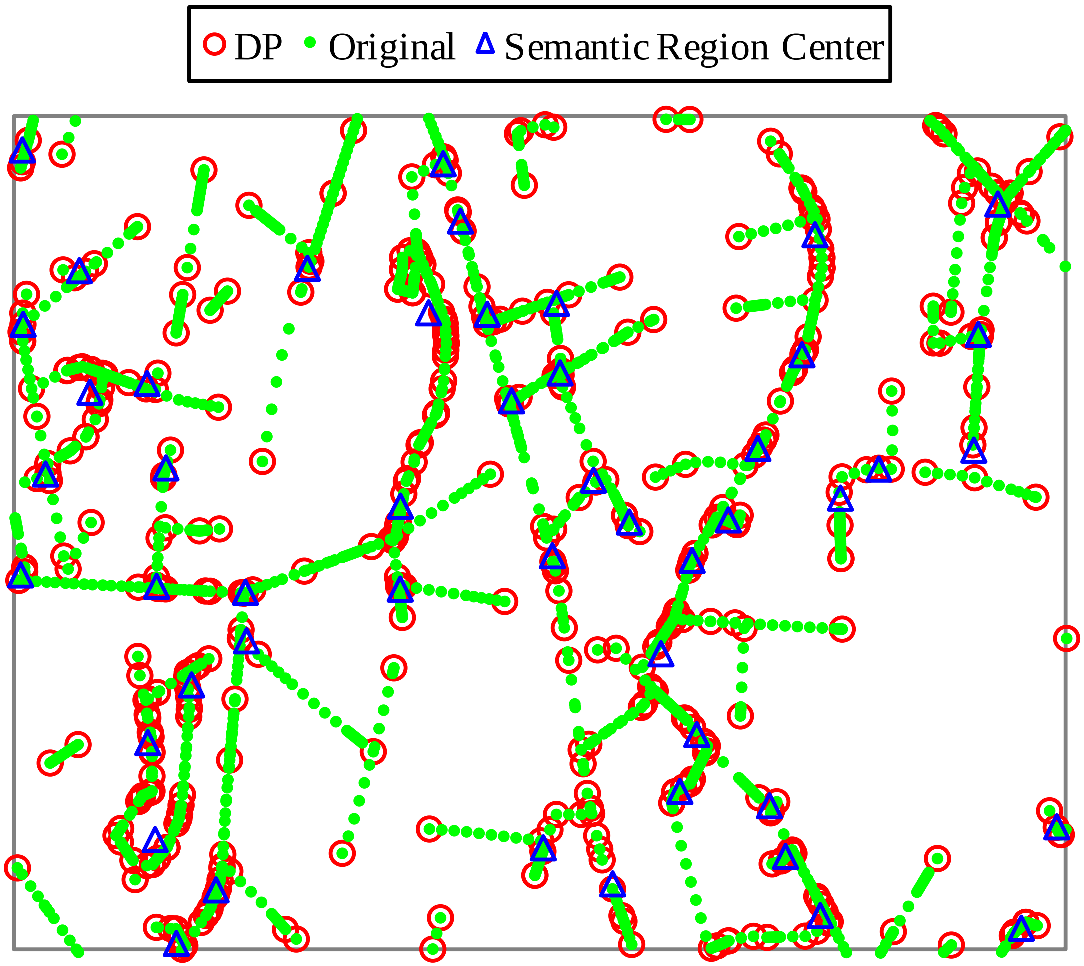






(G. Huang, et al, *World Wide Web Journal*, 2014)

Big Data Application 2 – Trajectories Represented by Semantic Regions

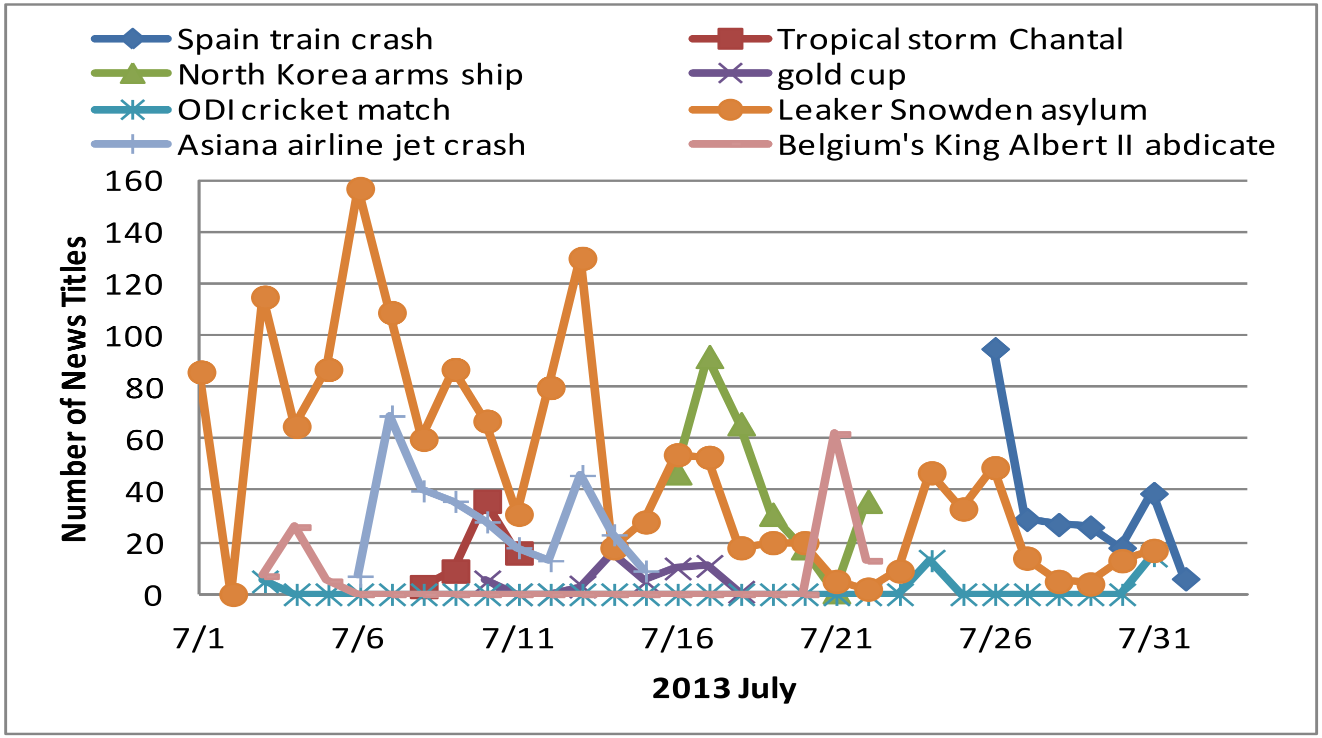


(G. Huang, Y. Zhang, J. He, Z. Ding, PAKDD 2011)

Big Data Application 3 – Short Text Analysis for Event Detection/Tracking

- Text Data (from http://www.infopig.com)
  - A corpus of over 100,000 pieces of news titles for one month (1/7/2013-31/7/2013) related to 157 countries from http://www.infopig.com. That is, averagely 3,225 pieces of news titles in a day and around 21 pieces of daily news titles in each country.
- Querying the trends/evolutions of world news events

An Example: Top 8 Events' Evolutions in July, 2013



(G. Huang, J. He, Y. Zhang, W. Zhou et al, *World Wide Web Journal*, 2015)

Three Levels of Knowledge in This Unit

Level 1: General process of using big data, including

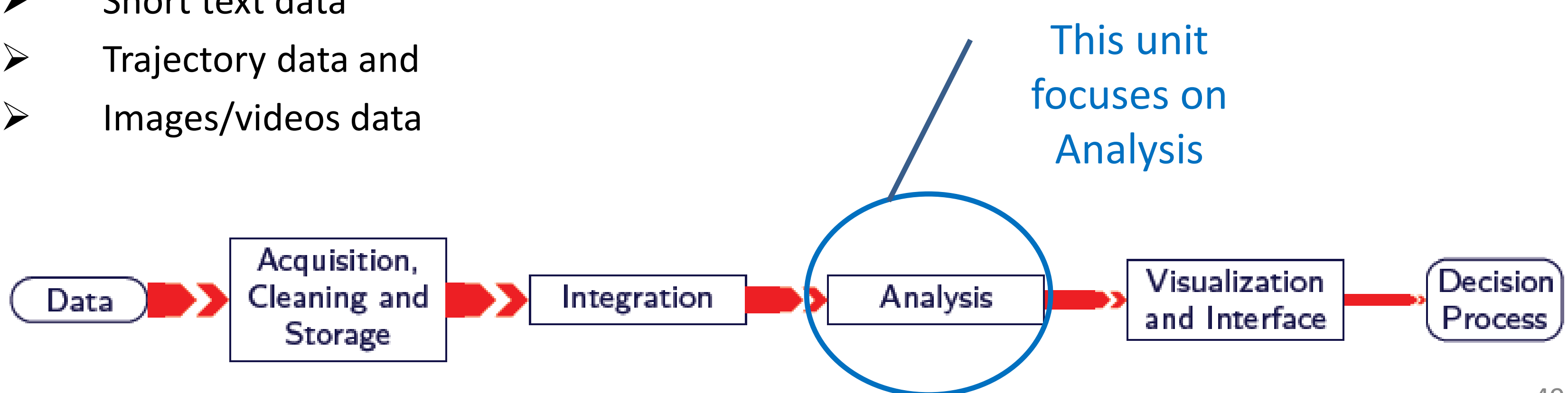
- Data acquisition
- Data cleaning
- Data analysis and
- Data applications.

Level 2: Basic data analytics techniques

- Common pattern discovery
- Outlier detection and
- Recommendation systems

Level 3: Data analytics applications

- Time series data
- Short text data
- Trajectory data and
- Images/videos data



Lecturing Content

- Introduction to Modern Data Science (Week 1)
- Mode of Teaching
  - Classes + Practices
- Basic Data Processing Techniques
  - Data Acquisition and Integration (Week 2)
  - Data Cleaning and Preparation (Week 3)
- Data Analytics Algorithms
  - Data Analytics: Common Pattern Discovery (Week 4)
  - Data Analytics: Outlier Detection (Week 5)
  - Data Analytics: Recommendation (Week 6)
- Big Data Applications
  - Time Series Data Analytics (Week 7)
  - Short Text Data Analytics (Week 8)
  - Trajectory Data Analytics (Week 9)
  - Image and Video Analytics (Week 10)
- Review of the Unit (Week 11)

Practical

- Practical 1: Python Basic (Week 1)
- Data Processing by Python
  - Practical 2: Data Acquisition by Python (Week 2)
  - Practical 3: Data Cleaning and Preparation by Python (Week 3)
  - Practical 4: Data Integration by Python (Week 4)
  - Practical 5: Plotting and Visualization (Week 5)
- Data Analytic Algorithms by Python
  - Practical 6: Demo Display (Week 6)
  - Practical 7: K-means Clustering (Week 7)
  - Practical 8: Principal Component Analysis (Week 8)
  - Practical 9: Support Vector Machines (Week 9)
  - Practical 10: Time Series Basic (Week 10)
  - Practical 11: Time Series Applications (Week 11)

Class Schedule

Class Room: LT8 (Y2.43)

TIME Tuesday (16:00-17:50)	CONTENT	CLASS TYPE
Week 1 (06 March, 2018)	Lecture 1: Introduction to Modern Data Science	
Week 2 (13 March, 2018)	Lecture 2: Data Acquisition and Integration	Skills/Talks
Week 3 (20 March, 2018)	Lecture 3: Data Cleaning and Preparation	Skills/Talks
Week 4 (27 March, 2018)	Lecture 4: Data Analytics - Common Pattern Discovery	Skills/Talks
Intra-trimester break*		
Week 5 (10 April, 2018)	Lecture 5: Data Analytics – Outlier Detection	Skills/Talks
Week 6 (17 April, 2018)	Lecture 6: Data Analytics – Recommendation	Skills/Talks
Week 7 (24 April, 2018)	Lecture 7: Big Data Applications – Time Series Data Analytics	Skills/Talks
Week 8 (01 May, 2018)	Lecture 8: Big Data Applications – Short Text Data Analytics	Skills/Talks
Week 9 (08 May, 2018)	Lecture 9: Big Data Applications – Trajectory Data Analytics	Skills/Talks
Week 10 (15 May, 2018)	Lecture 10: Big Data Applications – Image and Video Data Analytics	Skills/Talks
Week 11 (22 May, 2018)	Lecture 11: Unit Review	

Practical Schedule

	CONTENT	Lab Location and Time
Weeks 1-2	Practicals 1-2: Python Basic	Group 1 (T1.01): Mon (9:00-10:50AM) Group 2 (T1.05): Mon (11:00-12:50PM) Group 3 (B3.16): Mon (14:00-15:50PM) Group 4 (B3.16): Mon (16:00-17:50PM) Group 5 (B3.16): Thu (12:00-13:50PM) Group 6 (B4.01): Thu (14:00-15:50PM) Group 7 (B3.16) Thu (16:00-17:50PM)
Week 3	Practical 3: Data Acquisition by Python	
Week 4	Practical 4: Data Cleaning and Preparation by Python	
Week 5	Practical 5: Data Integration by Python	
Intra-trimester break*		
Week 6	Practical 6: Plotting and Visualization	
Week 7	Practical 7: K-means Clustering	
Week 8	Practical 8: Principal Component Analysis	
Week 9	Practical 9: Support Vector Machines	
Week 10	Practical 10: Time Series Basic	
Week 11	Practical 11: Time Series Applications	

*Friday 30 March – Sunday 8 April 2018

Evaluation/Marking

	Marks	Deadline	Evaluation summary	Guidelines documents at CloudDeakin
Assignment 1	15%	14 April, 2018	General data processing and using big data (about lectures in Weeks 1-3)	SIT742 Assignments Guideline
Assignment 2	20%	19 May, 2018	Basic data analytic skills and four example data types (about lectures in Weeks 4-9)	
Practical Assignment 1	10%	Open in Weeks 4-6	Quiz 1: General data processing by Python (about practicals in Weeks 1-3)	SIT742 Quiz/Practical Assignments Guideline
Practical Assignment 2	10%	Open in Weeks 7-11	Quiz 2: Three basic data analytic algorithms by Python (about practicals in Weeks 7-11)	
Examination	45%	Examination Period (Two hours, closed book)	Examination about the lectures and practical s in Weeks 1-10	SIT742 Examination Guideline

Contacts

Unit Chair: Dr Guangyan Huang

Location: Building T, T2.12
Phone: +61-3-9244 6282
Email: guangyan.huang@deakin.edu.au
Homepage: http://www.deakin.edu.au/about-deakin/people/guangyan-huang

Practical Tutors:

Mr Borui Cai (bcai@deakin.edu.au)
Mr Shuiqiao Yang (shuiqiao.yang@deakin.edu.au)