FLOATWAY SYSTEMS

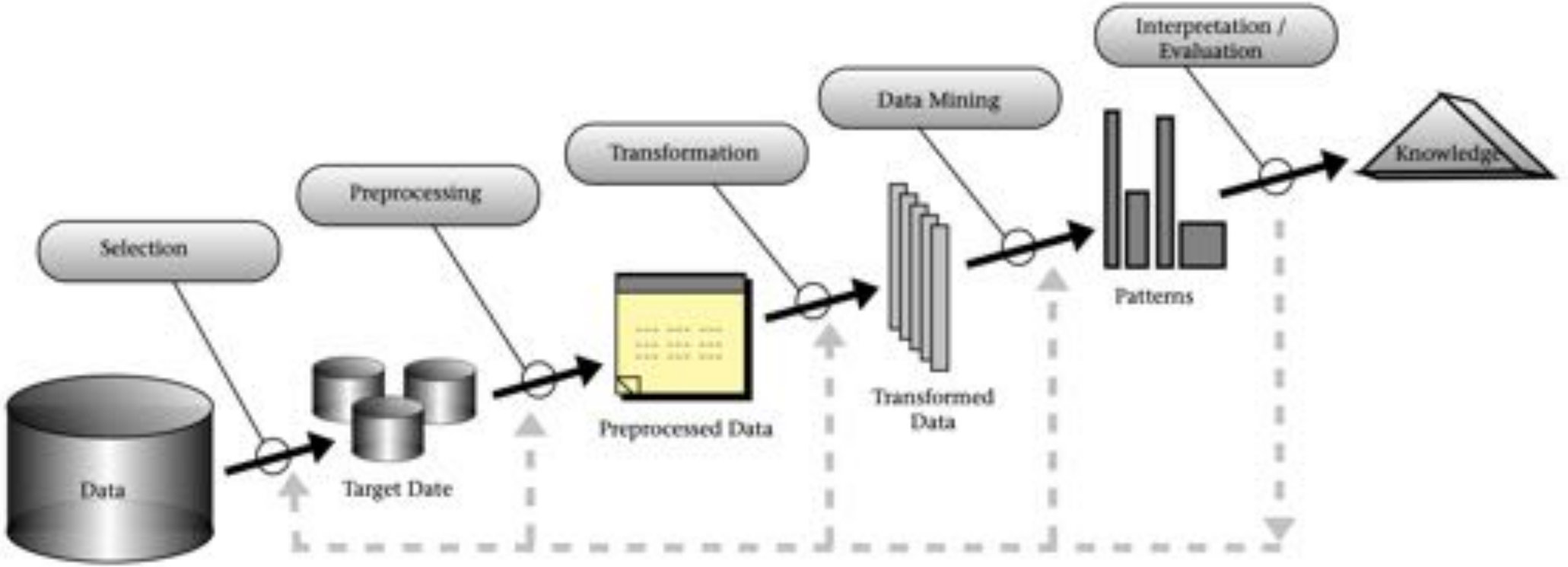
Data Science Project Framework

Data Science Project Framework

“Standard for doing data science project”

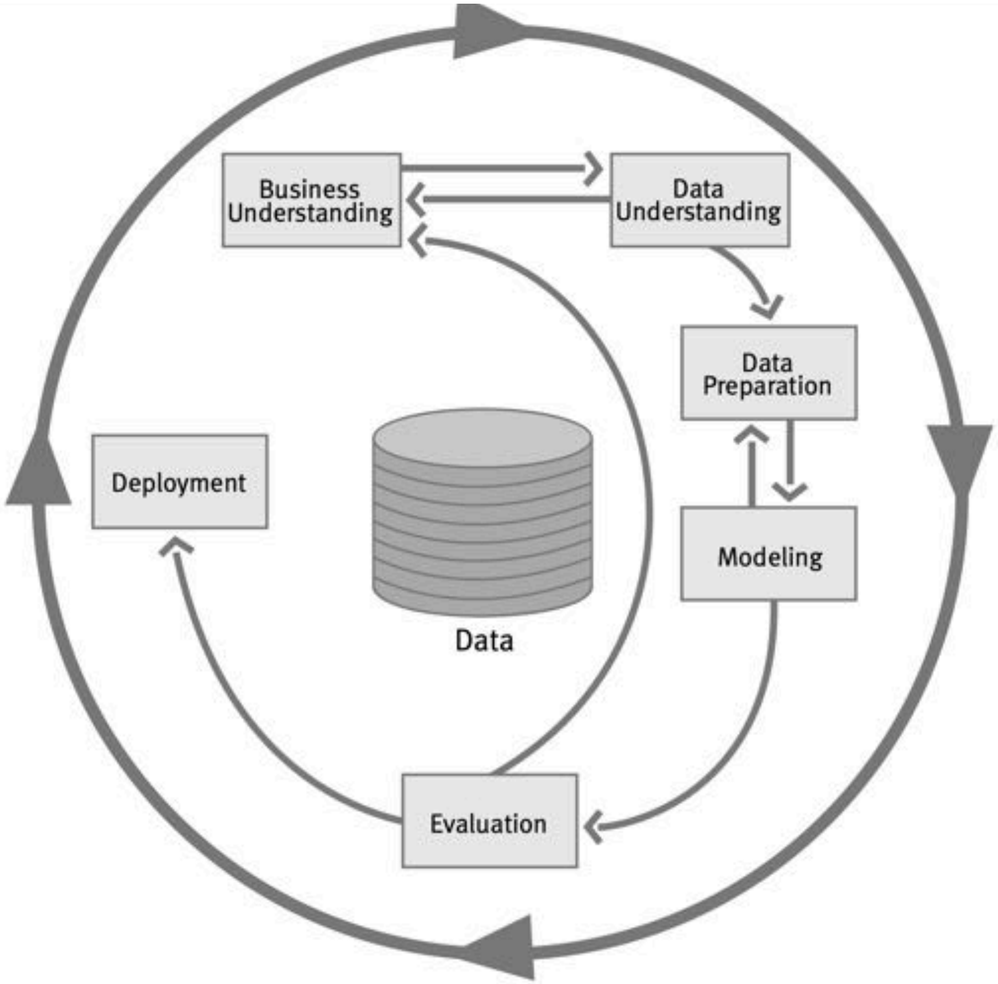
KDD **(Knowledge Data Discovery)**	**SEMMA** **(Sample, Explore, Modify, Model, and Assess)**
CRISP-DM **(Cross-industry Standard Process for Data Mining)**	**ASUM-DM** **(Analytics Solutions Unified Method for Data Mining)**

KDD Process (Fayyad, 1996)



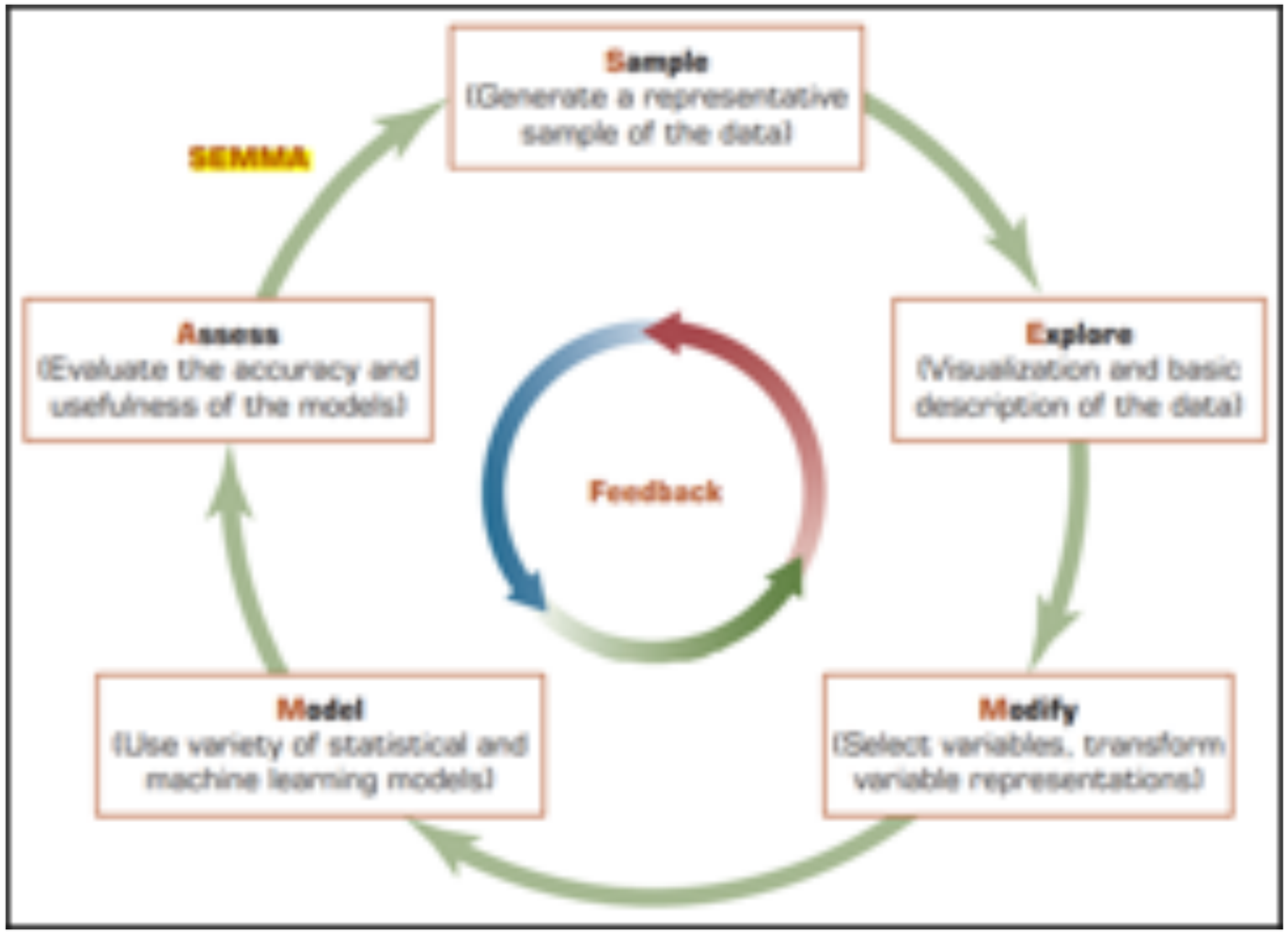
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

CRISP-DM Process (European Union, 1997)



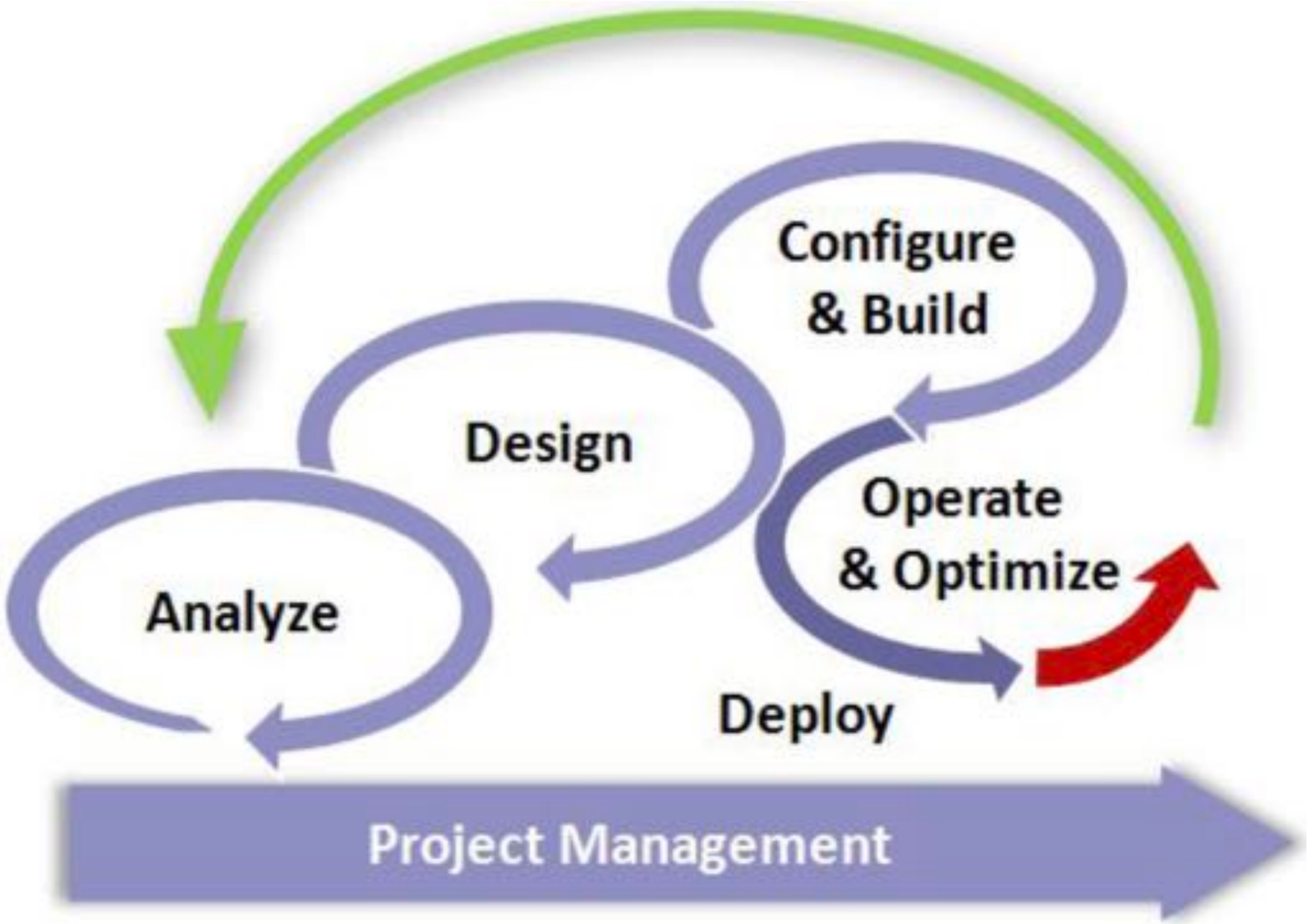
Chapman, P., et al (1999, March). The CRISP-DM user guide. In 4th CRISP-DM SIG Workshop in Brussels

SEMMA Process (SAS, 2005)



Sharda, R., Delen, D., Turban, E. (2018). Big data Intelligence, Analytics, and Data Science: A Managerial Perspective. 04. Pearson Education. New Jersey.

ASUM-DM Process (IBM, 2015)



IBM Analytics (2016). Analytics Solutions Unified Method: Implementations with Agile principles.

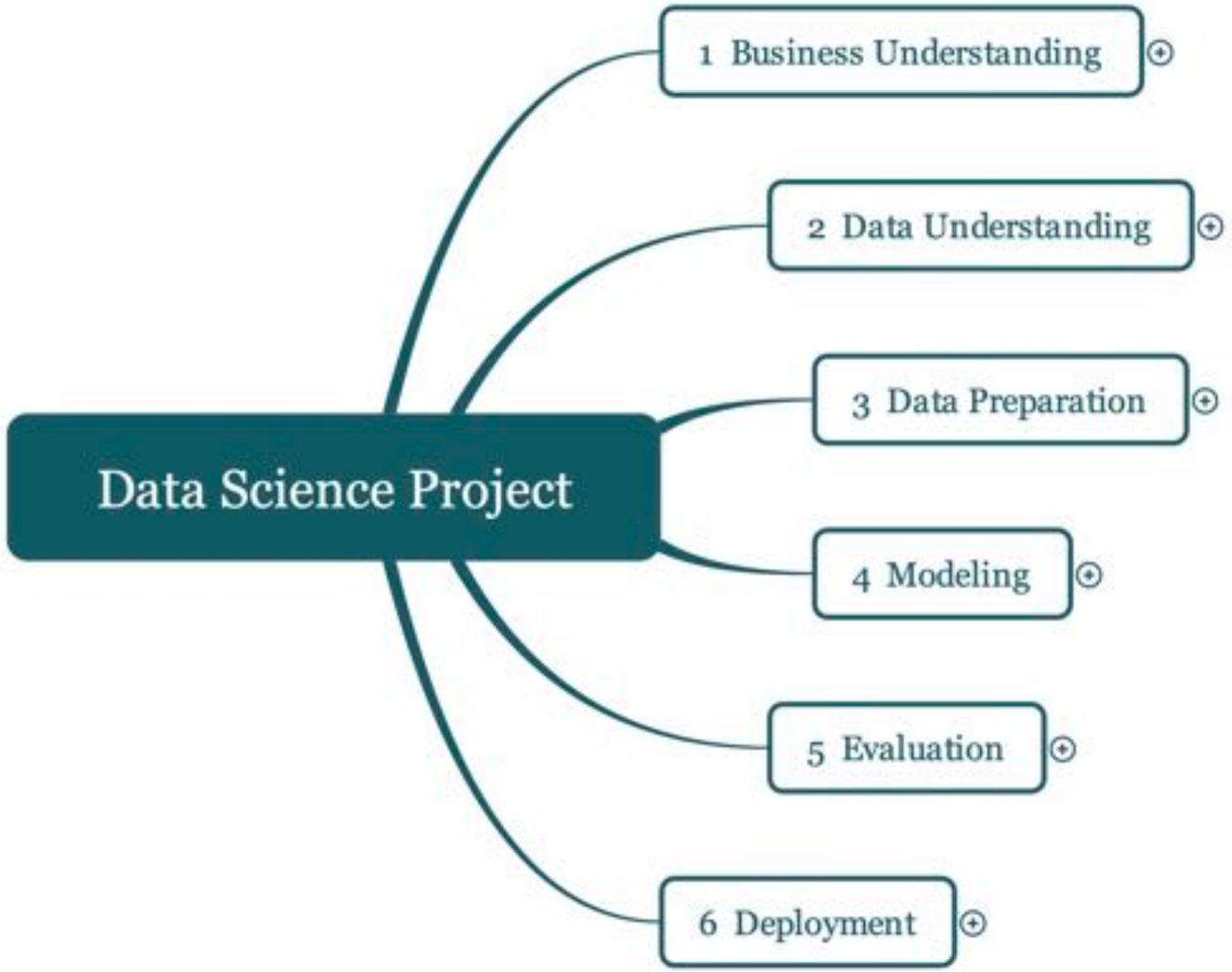
Framework Comparison

KDD	CRISP-DM	SEMMA	ASUM-DM
Pre KDD	Business Understanding	-	Analyze
Selection	Data Understanding	Sample	
Preprocessing		Explore	
Transformation	Data Preparation	Modify	Design
Data Mining	Modeling	Model	Configure and Build
Interpretation/Evaluation	Evaluation	Assessment	
Post KDD	Deployment	-	Deploy
-	-	-	Operate and Optimize

FLOATWAY SYSTEMS

CRISP-DM

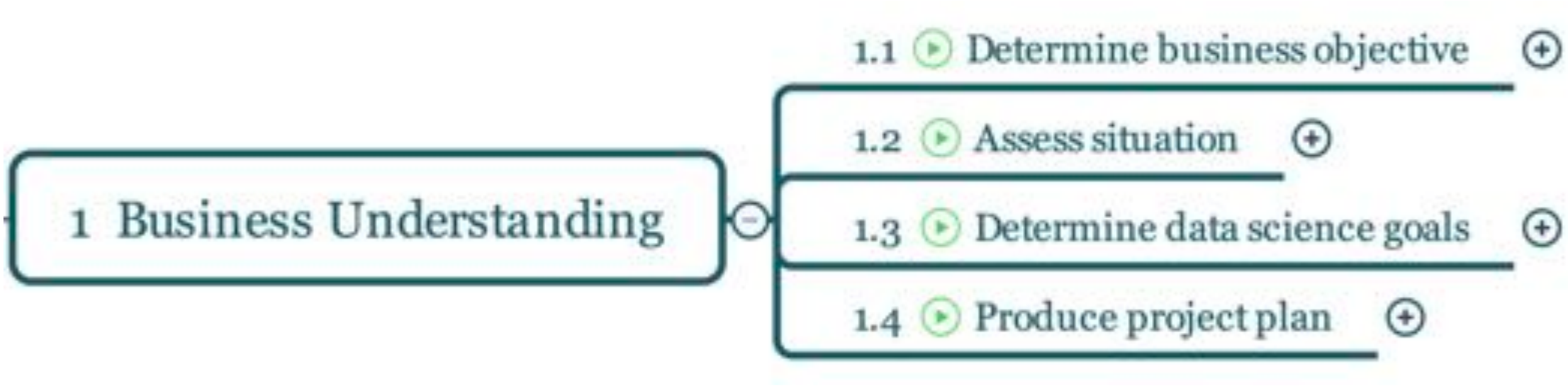
CRISP-DM Process



FLOATWAY SYSTEMS

1. Business Understanding

1. Business Understanding



1.1. Determine Business Objectives

- Memahami apa yang sebenarnya klien ingin capai dalam perspektif bisnis.
- Output:

1.1.1. Background

- Informasi yang diketahui tentang situasi bisnis dari klien. Berisi informasi bisnis, masalah, dan solusi saat ini.
- Contoh:
  - Sebuah perusahaan X bergerak di bidang …… dipimpin oleh…..dst
  - Masalah-masalah yang dihadapi yaitu: penjualan produk mengalami penurunan, banyak pelanggan yang tidak kembali, ……dst
  - Solusi yang sudah diterapkan: menambah promosi produk, tapi ini membutuhkan biaya yang tidak sedikit, dan tidak signifikan hasilnya…..dst

1.1.2. Business Objectives

- Tentukan **tujuan utama** yang ingin dicapai klien, dalam perspektif bisnis.
- Contoh:
  - Membuat pelanggan tidak beralih ke produk perusahaan lain
  - Membuat kebijakan yang sesuai keinginan rakyat
  - Menurunkan jumlah kasus korupsi di kalangan pejabat
  - Mengetahui karakteristik masyarakat pengguna media sosial
  - …..dll

1.1.3. Business Success Criteria

- Kriteria yang menjadikan business objective dikatakan berhasil atau tidak. Spesifik dan bisa diukur.
- Contoh:
  - Customer retention rate > 90%
  - Jumlah protes kebijakan di media sosial berkurang 30%
  - Jumlah kasus korupsi berkurang 50%
  - **Bidang HR** dapat memahami karakteristik masyarakat.
  - Dll

1.2. Assess Situation

- Menjelaskan tentang sumber daya, batasan, asumsi, dan faktor-faktor lain yang bisa berpengaruh.
- Output:

1.2.1. Inventory of Resources

- Daftar sumber daya yang tersedia untuk projek
- Contoh:
  - Daftar hardware yang tersedia (komputer, server, dll)
  - Sumber data dan pengetahuan (data apa saja yang dimiliki)
  - Sumber daya manusia (ekspertis yang tersedia, teknisi, dll)
  - Sumber dana
  - dll

1.2.2. Requirement, Assumptions, and Constraints

- Daftar kebutuhan, daftar asumsi, daftar batasan
- Contoh:
  - Kebutuhan: jadwal pelaksanaan, data yang dibutuhkan, sumber daya, dll
  - Asumsi: kualitas data (ketersediaan, akurasi, dll), faktor eksternal, dll
  - Batasan: dana, waktu, sumber daya, data, dll

1.2.3. Risks and Contingencies

- Daftar resiko yang mungkin akan ada dan rencana mengatasinya
- Contoh:
  - Resiko: data yang didapat sangat “kotor”, data di komputer hilang, dll
  - Rencana mengatasi: tambah proses “cleansing”, menyimpan di cloud, dll

1.2.4. Terminology

- Penjelasan tentang istilah-istilah bisnis (spesifik di klien) dan data science yang berkaitan dengan projek
- Contoh:
  - Bisnis (spesifik): churn rate adalah…., R-naught adalah….., dll
  - Data Science: MSE adalah…., regresi adalah…., recall adalah…..,dll

1.2.5. Costs and Benefits

- Perkiraan biaya-biaya yang dibutuhkan serta manfaat yang terkait.
- Contoh:
  - Pengambilan data 100 juta rupiah
    - Semakin banyak dana untuk pengambilan data -> data semakin banyak -> prediksi lebih akurat
  - Biaya sewa server 2 juta per bulan
    - Semakin mahal server (kapasitas bagus) -> proses modeling menjadi lebih cepat
  - Dll

1.3. Determine Data Science Goals

- Penjelasan tujuan projek data science dalam perspektif teknis

1.3.1. Data Science Goals

- Tujuan yang bersifat teknis dan spesifik menjelaskan masalah yang ingin dipecahkan.
- Tipe masalah: deskripsi, eksplorasi, segmentasi, klasifikasi, regresi, atau asosiasi
- Contoh:
  - Klasifikasi produk yang akan dipilih pelanggan
  - Prediksi berapa banyak pelanggan yang akan membeli lagi
  - Dll

1.3.1. Data Science Goals (2)

Deskripsi
Ringkasan karakteristik suatu data

Klasifikasi
Memprediksi label/kelas suatu data

Eksplorasi
Mengungkap *insight* dalam suatu data

Regresi
Memprediksi nilai kontinyu dari data

Segmentasi
Pemisahan data ke dalam grup-grup

Asosiasi
Mengungkap keterkaitan antar data, grup, atau variabel

1.3.2. Data Science Success Criteria

- Kriteria keluaran yang dianggap sukses dalam istilah teknis
- Contoh:
  - Akurasi model prediksi > 95%
  - Indeks Silhouette > 0.8
  - <subjective assessment>
  - Dll

1.4. Produce Project Plan

- Penjelasan tentang rencana dalam melaksanakan projek
- Output:

1.4.1. Project Plan

- Daftar langkah-langkah dalam projek serta kebutuhan sumber daya untuk setiap langkah.
- Bisa dibuat menggunakan gantt chart.

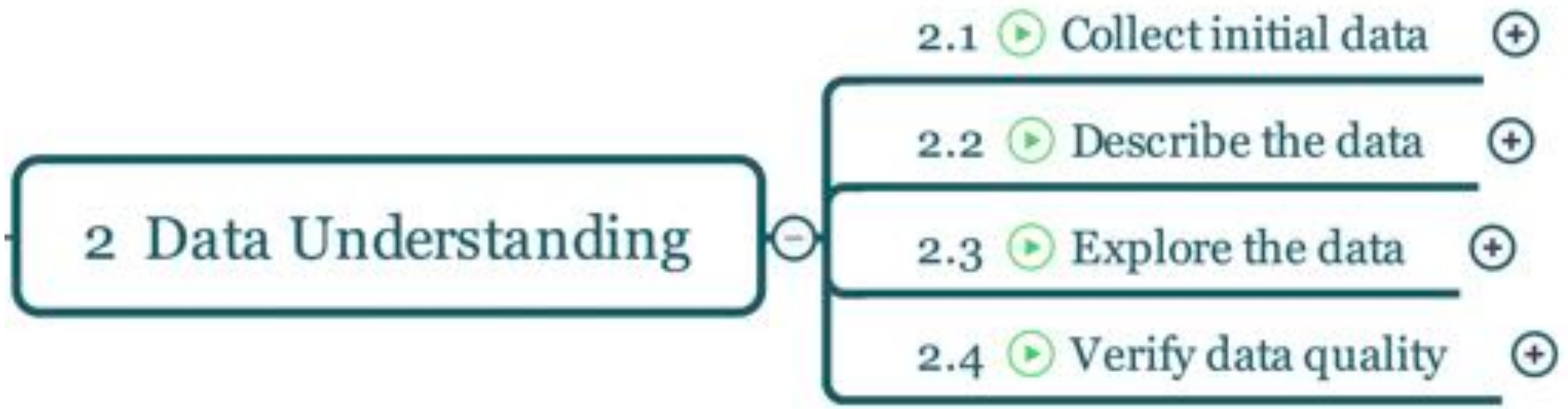
1.4.2. Initial Assessment of Tools and Techniques

- Memilih alat dan metode yang potensial digunakan pada setiap fase dalam proyek.
- Bisa disertakan plus minus masing-masing.
- Contoh:
  - Manajemen proyek: Ms Project, Ganttproject, dll
  - Eksplorasi: Tableau, dll
  - Model: Python, R, Knime, dll
  - Report: Latex, Ms Word, dll
  - Teknik prediksi: XGBoost, CNN, dll
  - Dll

FLOATWAY SYSTEMS

2. Data Understanding

2. Data Understanding



2.1. Collect Initial Data

- Mencoba mengambil data dari sumber data yang sudah dituliskan sebelumnya.

2.1.1. Initial Data Collection Report

- Menjelaskan data-data yang digunakan dalam proyek. Termasuk bagaimana cara mendapatkan/mengaksesnya secara teknis.
- Contoh:
  - Data pelanggan dapat diakses dari tabel pelanggan yang ada di database X dengan akses ……..
  - Data komentar warganet diakses menggunakan API Twitter dengan metode pengambilan ……..
  - Dll

2.2. Describe Data

- Memeriksa gambaran “kasar” dari suatu data
- Jika diperlukan, bisa ubah asumsi setelah memeriksa data ini

2.2.1. Data Description Report

- Penjelasan umum tentang data meliputi format data, kuantitas, tipe kolom, dan sebagainya.
- Bisa disajikan dalam tabel
- Contoh:
  - Ada 5 tabel, tiap tabel ada 1000 baris dan 16 kolom
  - Kolom 1 adalah ….., merepresentasikan…….
  - Statistika dasar untuk tiap tabel
  - Dll

Hands-on with Python

2.3. Explore Data

- Mengeksplorasi data, meliputi: visualisasi dasar, verifikasi hipotesis, dll
- Proses ini sering disebut sebagai Exploratory Data Analysis (EDA)
- Mungkin terkait langsung dengan tujuan teknis data science tertentu.

Slide: Data Visualization

2.3.1. Data Exploration Report

- Berupa temuan awal atau hipotesis awal serta dampaknya dalam proyek keseluruhan.
- Hasil verifikasi hipotesis awal juga dapat disampaikan.
- Contoh:
  - Temuan tentang adanya tren dari penjualan produk…..
  - Temuan adanya anomali pada data penderita C-19…..
  - Hipotesis awal tentang…..
  - dll

Hands-on with Python

2.4. Verify Data Quality

- Memeriksa kualitas data: apakah datanya lengkap (untuk semua kasus)?, apakah ada data eror? Seberapa banyak erornya? Apakah ada data kosong? Dll

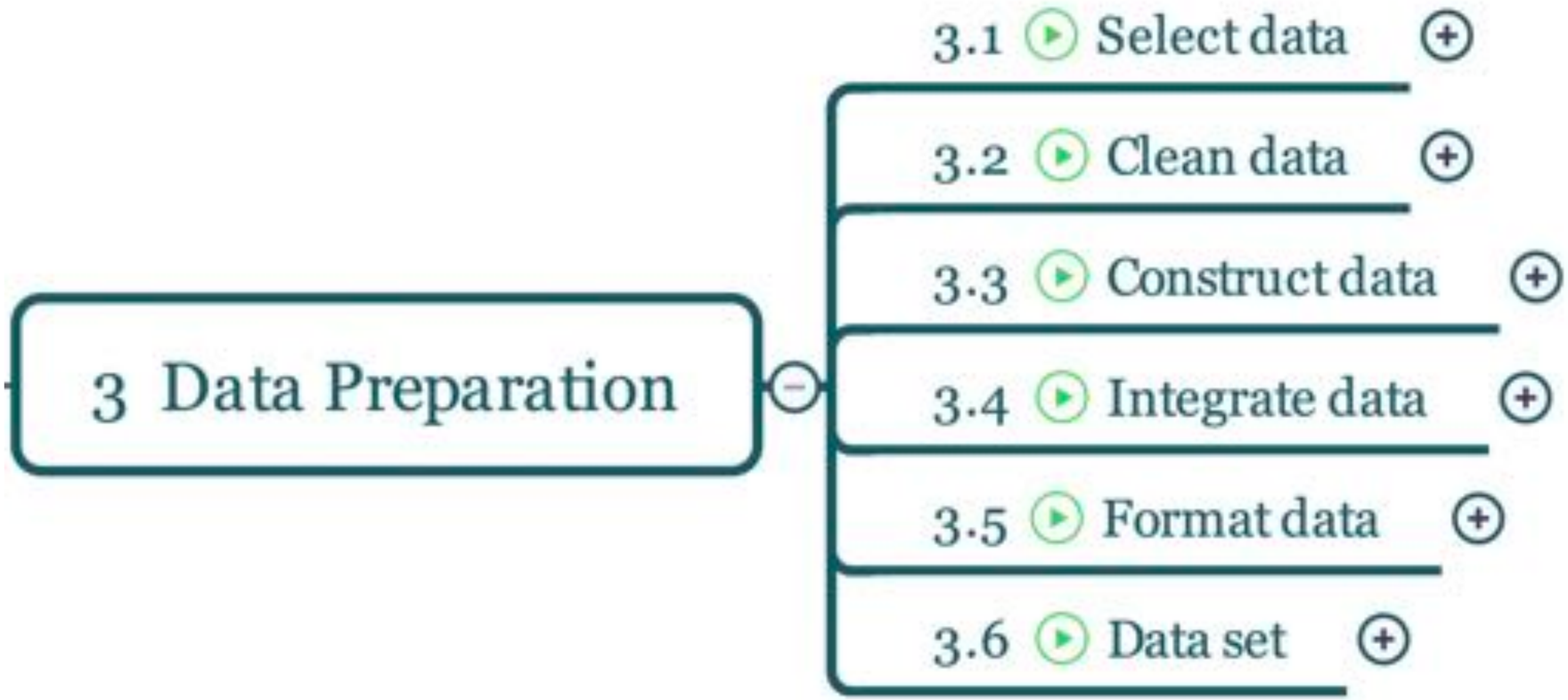
2.4.1. Data Quality Report

- Daftar hasil pengamatan kualitas data
- Jika ada masalah terkait kualitas, berikan juga solusi yang mungkin
- Contoh:
  - Penulisan jenis kelamin tidak sama. Solusi: proses standardisasi
  - Ada 50 data kosong pada kolom. Solusi: imputasi
  - Kolom profesi semua berisi mahasiswa. Solusi: hapus kolom/tambah data
  - Dll

FLOATWAY SYSTEMS

3. Data Preparation

3. Data Preparation



3.1. Select Data

- Memilih subset data, dapat berupa kolom atau tabel yang sesuai dengan tujuan data science
- Proses yang digunakan: feature selection dan sampling

3.1.1. Rationale for Inclusion/Exclusion

- Daftar alasan mengapa memilih atau membuang data yang bersangkutan
- Dapat memanfaatkan uji statistika
- Contoh:
  - Kolom pendapatan dipilih karena berkorelasi tinggi dengan pengeluaran berdasarkan uji korelasi
  - Membagi data menjadi data training dan testing menggunakan random sampling
  - Dll

Hands-on with Python

3.2. Clean Data

- Meningkatkan kualitas data hingga mencapai tingkat yang dibutuhkan untuk melakukan analisis tertentu

3.2.1. Data Cleaning Report

- Menjelaskan keputusan serta langkah-langkah dalam mengatasi masalah kualitas data yang ada di data quality report (2.4.1)
- Contoh:
  - Proses mengatasi data kosong
  - Mengatasi typo
  - Dll

Hands-on with Python

3.3. Construct Data

- Membangun data dengan menambah kolom baru atau menambah baris baru
- Menambah kolom (derived attributes) biasa disebut feature engineering
- Menambah baris (generated records) biasa disebut oversampling

Hands-on with Python

Slide: Feat. Eng. and Gen. Records

3.4. Integrate Data

- Menggabungkan data dari berbagai tabel atau dari sumber lain

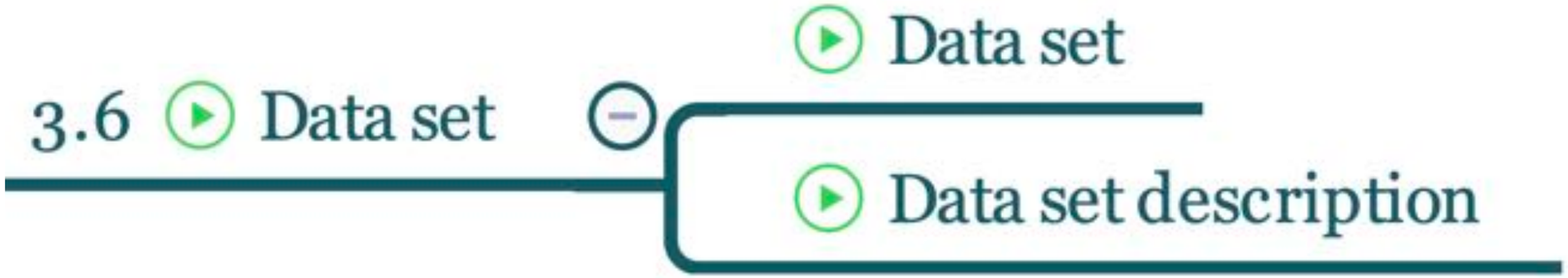
Hands-on with Python

3.5. Format Data

- Pengubahan format data dengan tidak mengubah makna namun bisa berguna untuk pembuatan model
- Contoh: mengubah urutan kolom, urutan baris, dan sebagainya

3.6. Data Set

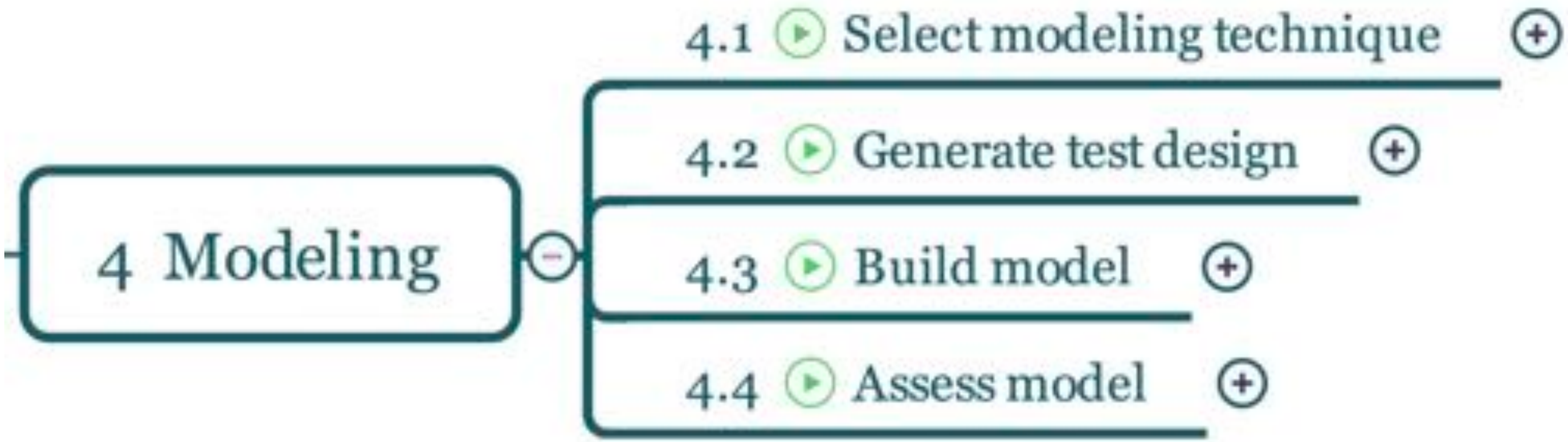
- Output akhir dari proses data understanding
- Data set = data siap untuk dibuat model
- Data set description = informasi metadata tentang dataset



FLOATWAY SYSTEMS

4. Modeling

4. Modeling

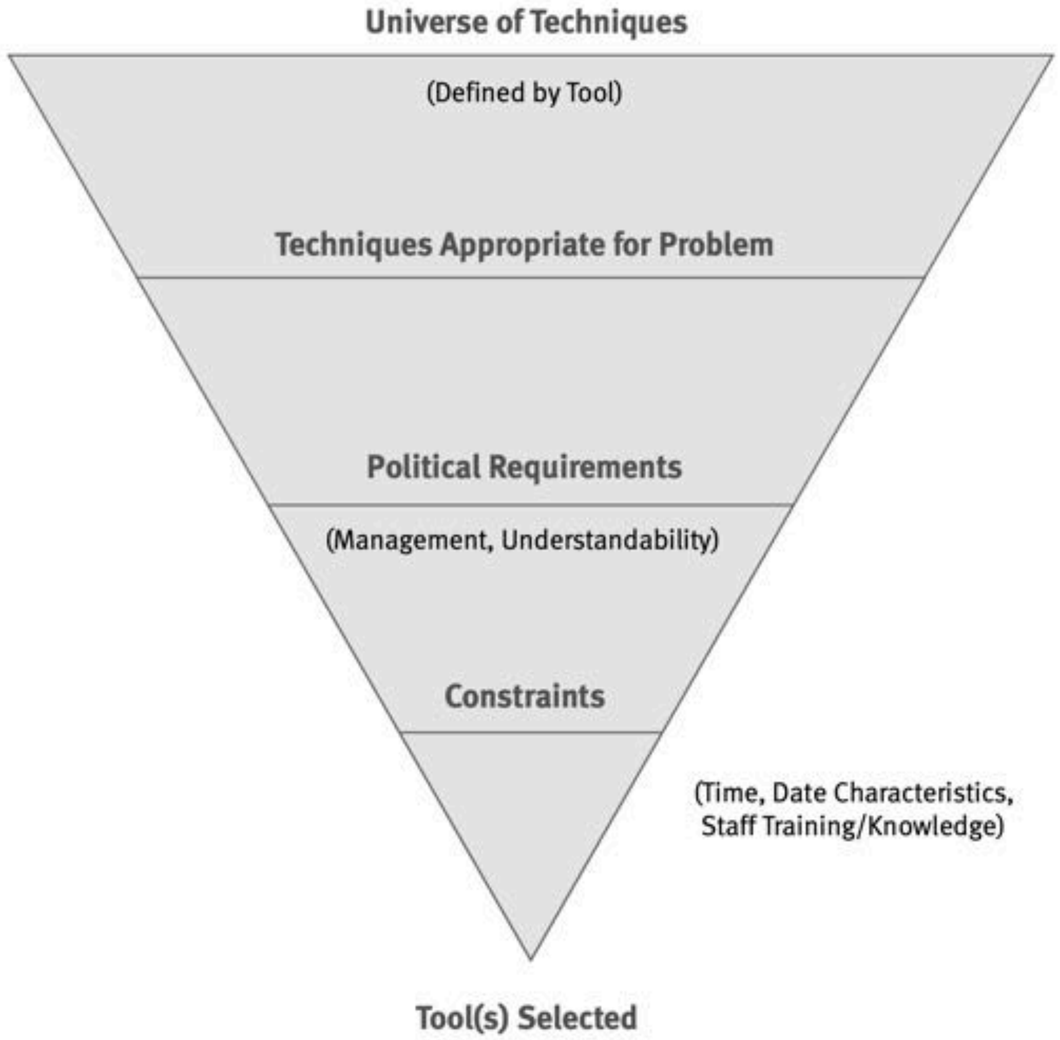

4 Modeling
4.1 Select modeling technique
4.2 Generate test design
4.3 Build model
4.4 Assess model

4.1. Select Modeling Technique

- Memilih teknik pemodelan yang akan digunakan

Hands-on with Python

4.1.1. Modeling Technique



4.1.2. Modeling Assumptions

- Banyak model yang mengharuskan suatu asumsi terhadap data.
- Contoh:
  - Linear regression membutuhkan asumsi linearitas, dll
  - Random forest tidak membutuhkan asumsi
  - SVM membutuhkan asumsi bahwa datanya independen dan tersebar merata
  - dll

4.2. Generate Test Design

- Merencanakan skema pengujian model
- Contoh:
  - Membagi dataset ke dalam training, validation, dan testing dengan proporsi……Kemudian melakukan pembuatan model di training, diaplikasikan di validation, dan diuji di testing.
  - Dsb

4.3. Build Model

- Menjalankan proses pembuatan model
- Output:
  - Parameter awal yang digunakan di model
  - Model itu sendiri
  - Deskripsi model. Bisa berisi parameter/hyperparameter yang digunakan, dan informasi lain terkait model akhir

4.4. Assess Model

- Mengevaluasi hasil model dikaitkan dengan kriteria sukses dari tujuan data science

4.4.1. Model Assessment

- Regression
  - MAE (Mean Absolute Error), MSE (Mean Square Error), RMSE (Root Mean Square Error)
- Classification
  - Accuracy, Precision, Recall, F1-Score, Sensitivity, Specivicity, TPR, FPR, ROC AUC, dll
- Clastering
  - WCSS, Silhouette Index, Rand Index, Calinski-Harabasz Index, Davies-Bouldin Index, dll

4.4.2. Revised Parameter Settings

- Berdasarkan hasil assessment, maka bisa dilakukan proses pengubahan parameter yang ada di dalam model untuk mendapat model terbaik.

- Alur proses bisa berulang dari membuat model hingga assessment

FLOATWAY SYSTEMS

5. Evaluation

5. Evaluation

- Mengevaluasi keseluruhan projek, dikaitkan dengan business objective
- Result projek = Model + Findings (temuan)

5.1. Evaluate Results

- Apakah hasil dari data science sudah sesuai dengan business objective?
- Tuliskan rekomendasi untuk projek selanjutnya
- Pilih model yang hasilnya sesuai dengan business criteria

5.2. Review Process

- Meninjau ulang proses data science di dalam projek
- Dapat dikatakan sebagai proses Quality Assurance

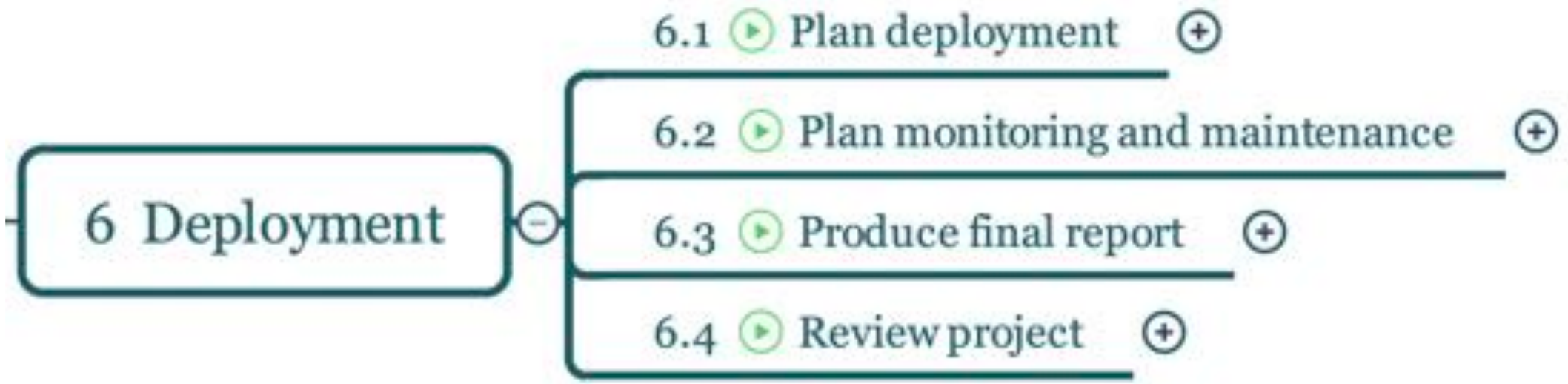
5.3. Determine Next Steps

- Membuat daftar aksi selanjutnya beserta alasannya
- Menentukan langkah mana yang diambil beserta alasannya

FLOATWAY SYSTEMS

6. Deployment

6. Deployment



6.1. Plan Deployment

- Membuat perencanaan pengaplikasian hasil data science ke dalam proses bisnis

6.2. Plan Monitoring and Maintenance

- Membuat perencanaan monitoring dan perawatan hasil data science yang sudah di-deploy ke sistem bisnis

6.3. Produce Final Report

- Membuat laporan dan presentasi akhir

6.4. Review Project

- Membuat review keseluruhan projek, bagian mana yang bisa ditingkatkan, beserta rekomendasi pengembangan selanjutnya