Sains Data

Adila Alfa Krisnadhi*, Siti Aminah, Aruni Yasmin Azizah,
Dina Chahyati, Fariz Darari

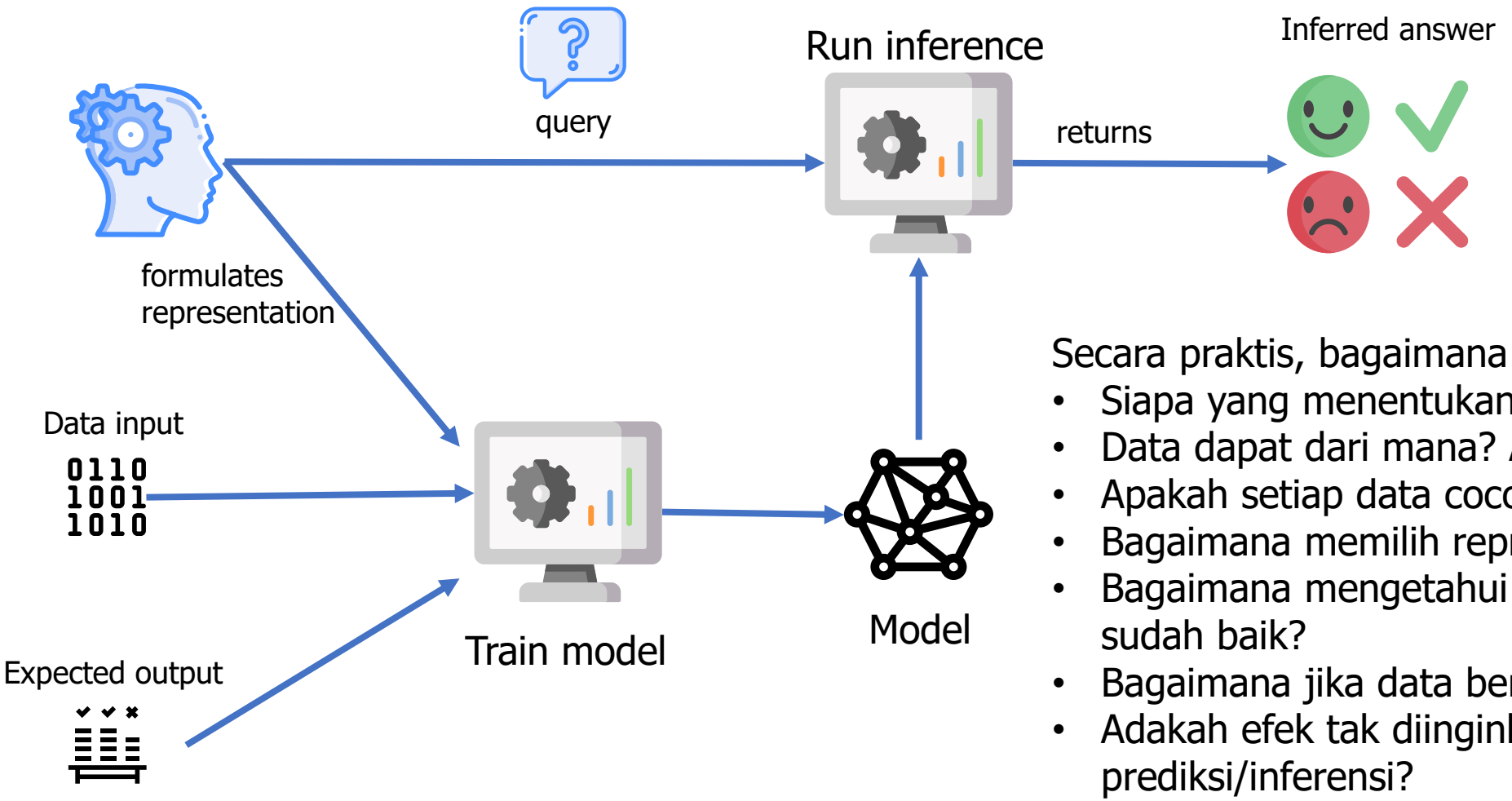
CSGE603130 - Kecerdasan Artifisial dan Sains Data Dasar

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

Kecerdasan artifisial masa kini

- … sangat dipengaruhi ketersediaan data yang massif: **Big Data**
  - **Volume**: data terakumulasi sangat besar (terabyte, petabyte, exabyte, zettabyte, yottabyte, dst.)
  - **Velocity**: data terhasilkan sangat cepat (per detik, per menit, per jam, dst)
  - **Variety**: ragam data semakin bermacam-macam: terstruktur, semi-terstrukur, tidak terstruktur (teks, suara, citra, video, dst)
  - **Veracity**: kesesuaian data dengan fakta semakin sulit ditakar
  - **Value**: kemampuan menghasilkan *value* dari data (profit, manfaat sosial, kepuasan pelanggan, dst.)
  - **Variability**: variasi penggunaan data dalam pelbagai aplikasi
- Dalam 1 menit Internet:
  - 1.4 juta Facebook scrolling, 500 jam video Youtube diunggah, 197 juta email terkirim, 69 juta pesan WA dan FB messenger terkirim, 200 ribu orang mengirim tweet, dll.

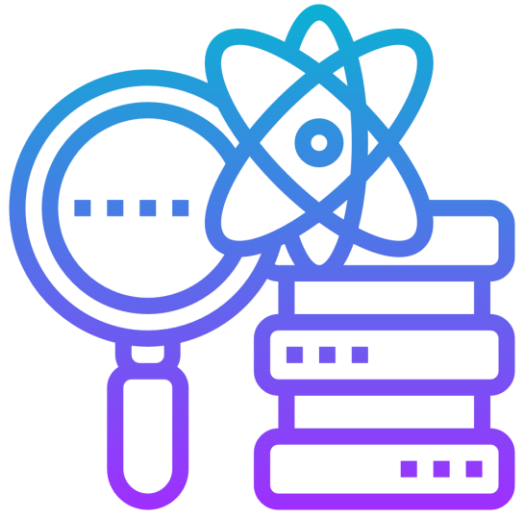
Learning agent dapat membantu, tapi ….



Secara praktis, bagaimana ini direalisasikan?

- Siapa yang menentukan kueri?
- Data dapat dari mana? Apakah cukup tersedia?
- Apakah setiap data cocok untuk dipakai?
- Bagaimana memilih representasi yang sesuai?
- Bagaimana mengetahui bahwa hasil inferensi sudah baik?
- Bagaimana jika data berubah?
- Adakah efek tak diinginkan dari prediksi/inferensi?
- …

Perlu metode yang tertib dan sistematis …

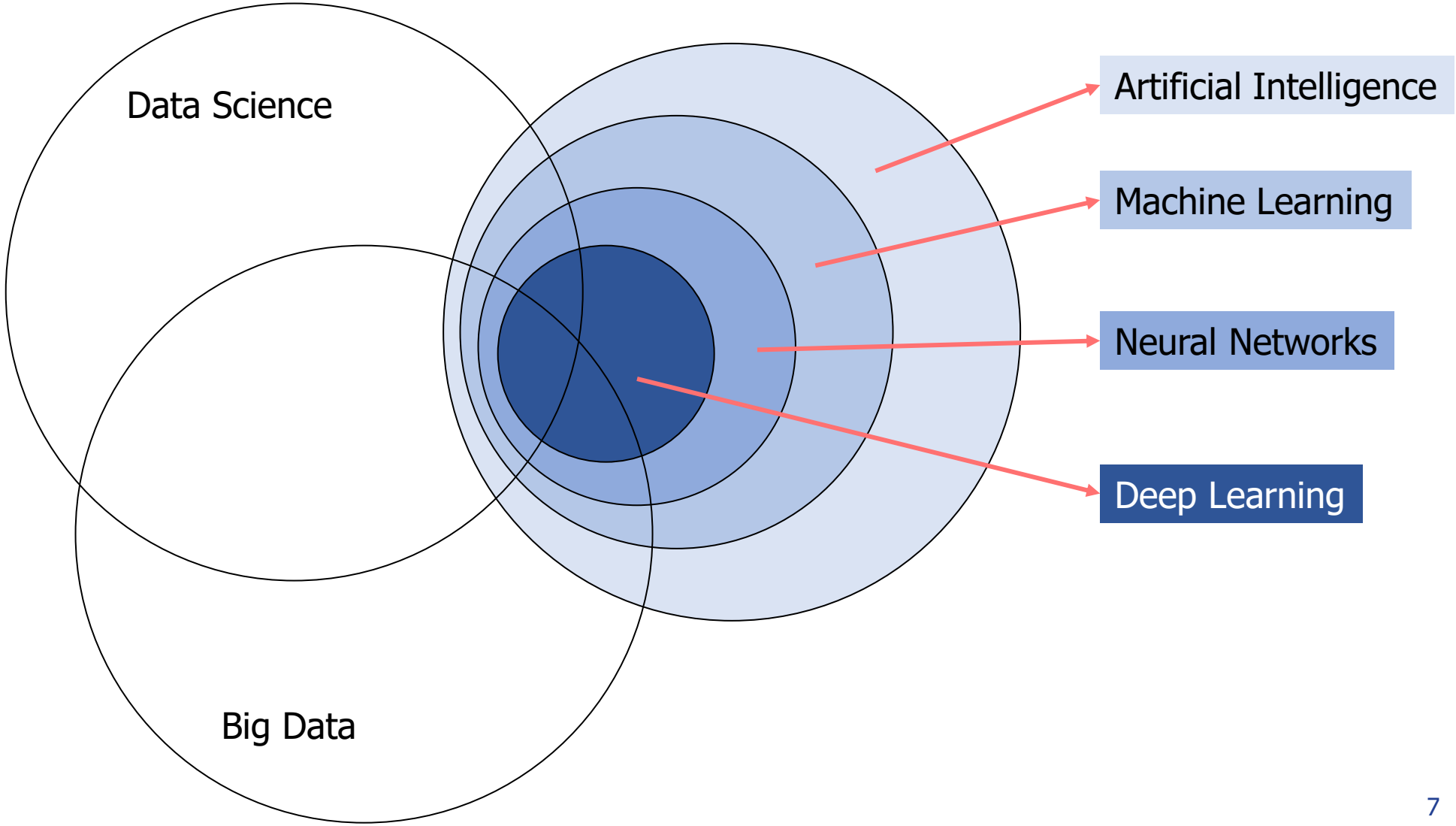


SAINS DATA

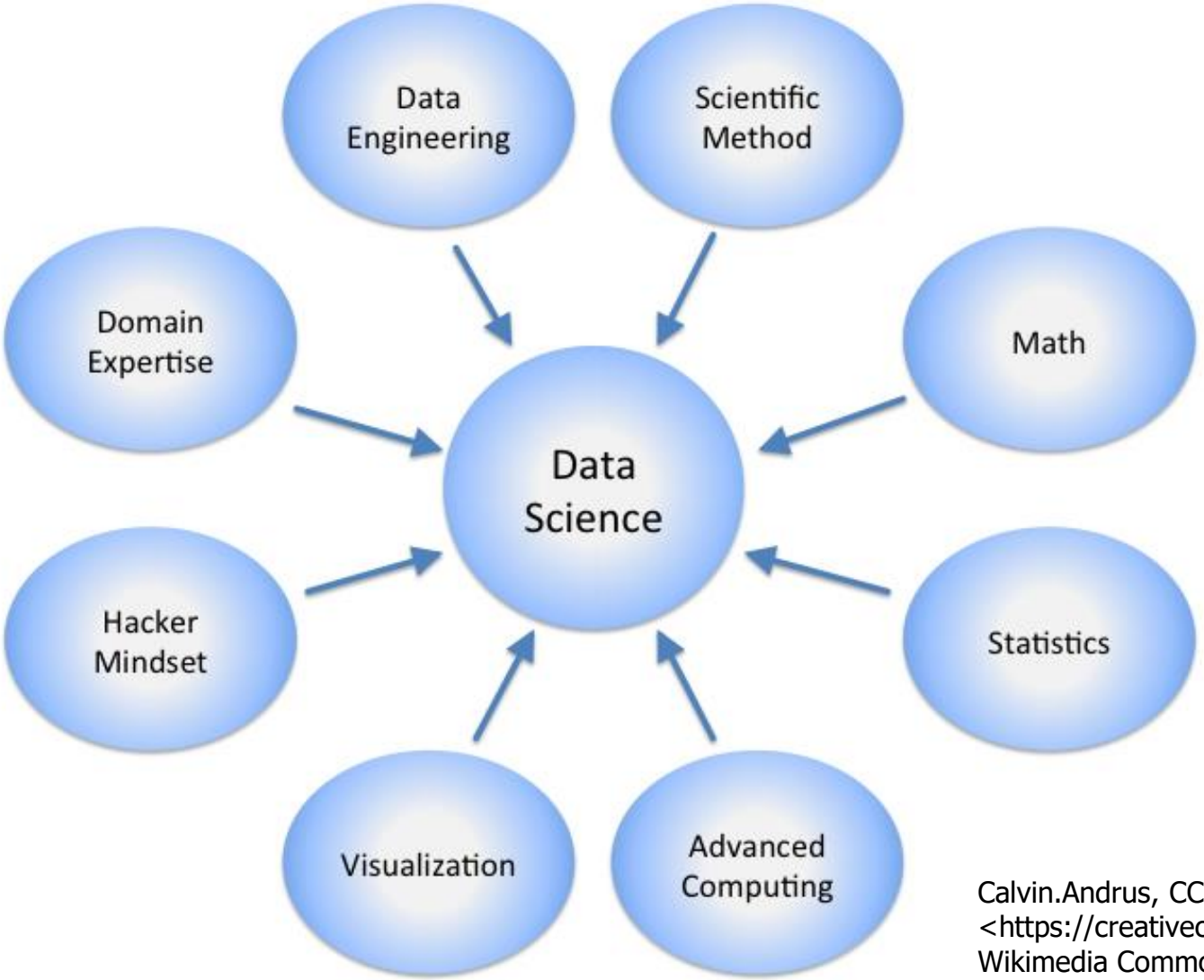
Sains Data

- Sains data: ekstraksi *insight* dari data yang *messy* (untuk pengambilan keputusan masa depan atau pemahaman masa lampau)
- Ilmuwan data = orang yang melakukan sains data, melakukan
  - **descriptive data analytics**: menjelaskan keadaan bisnis saat ini melalui data historis/lampau.
  - **diagnostic data analytics**: menjelaskan mengapa suatu masalah muncul dengan melihat data historis
  - **predictive data analytics**: memprediksi hasil di masa depan berdasarkan data historis
  - **prescriptive data analytics**: merumuskan rekomendasi upaya terbaik di masa depan berdasarkan hasil analitika prediktif dan pengetahuan lain.

KA vs. Big Data vs. Data Science



Area dasar sains data



Peran pelaku sains data

Belum ada standar; setiap industri (bahkan perusahaan) bisa bervariasi

Data engineer	Data scientist	Business stakeholder
Role: • Collect, manage, analyze and visualize data • Build, deploy data infrastructure and architecture **Core skills**: • Data ingest, ETL tools • Database systems • Distributed computing systems (Hadoop, Spark, etc.) • Data APIs • Unstructured data & data modeling • Data warehousing	**Role**: • Ascribes value to raw data via original interpretation & modeling • Use various machine learning methods to interrogate data **Core skills**: • Python, R • Distributed computing • Machine learning & deep learning • Feature engineering & predictive modeling • Statistics & math • Storytelling & visualization	**Role**: • Return on investment (ROI), Net Present Value (NPV), domain expertise • Value chains • Financial analysis **Core skills**: • Business intelligence • Statistics and math • Data stewardship & management • Storytelling & visualization • Business communication

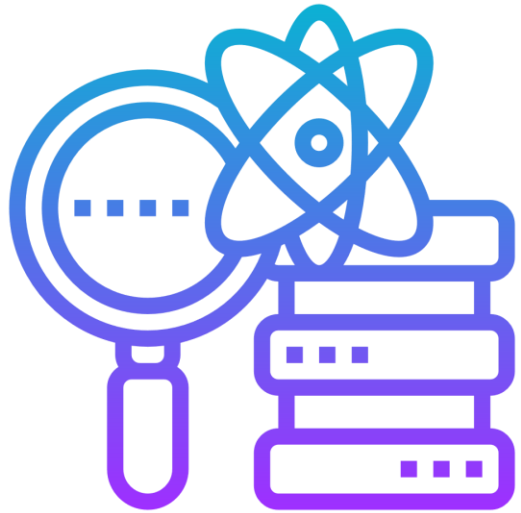
Bagaimana sistem KA dikembangkan dan dipakai?

- Tahap pengembangan/pelatihan

- Tahap penggunaan

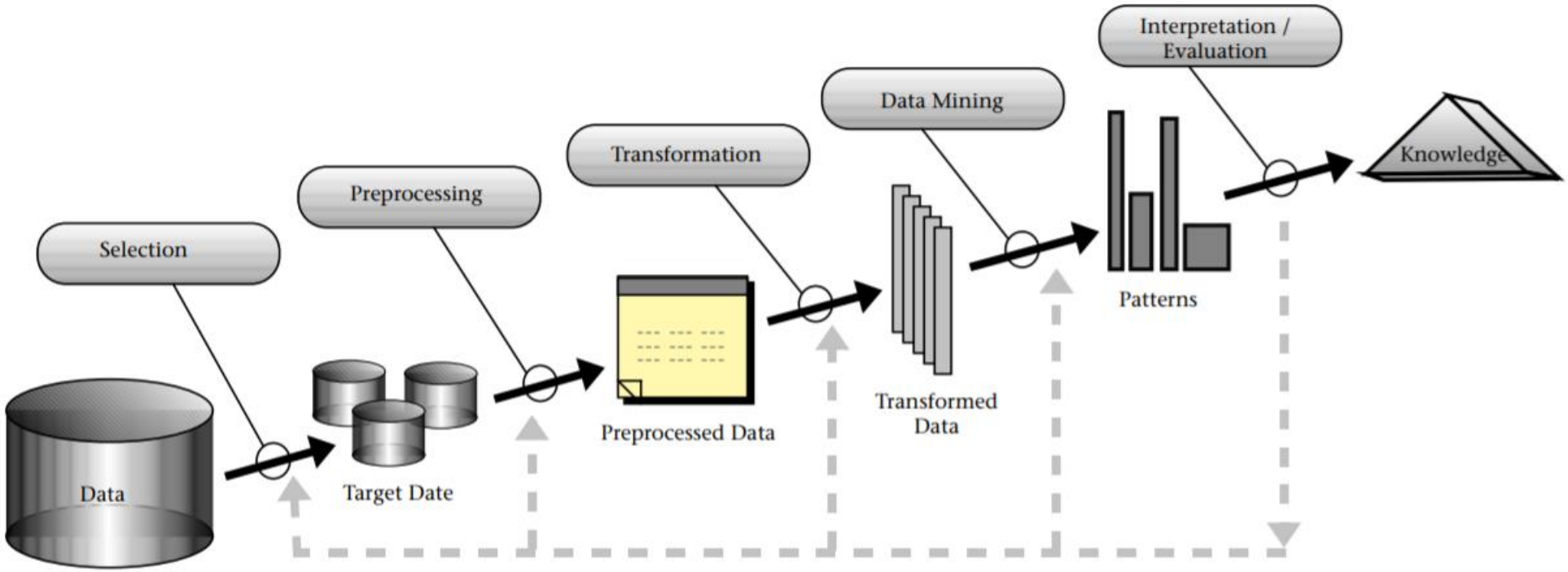
Kegagalan proyek sains data

- Menurut Gartner (2018),
  - Hanya 15-20% proyek sains data yang terselesaikan.
  - Hanya 8% dari yang terselesaikan menghasilkan value bagi organisasi.
- Penyebab kegagalan:
  - Problem tidak jelas atau salah diformulasikan; *over-promise* pada solusi problem
  - Data: tidak cukup (jumlah); tidak tepat (variabel), kualitas tidak memadai, makna/semantik data tidak jelas, bias implisit (saat sampling) tidak diperhitungkan.
  - Model: terlalu kompleks, metrik kinerja tidak tepat.
  - Algoritma: terlalu rumit sehingga tidak dapat dipahami secara teknis, tidak tepat
  - SDM: *one-man show*, dukungan *stakeholder* tidak cukup.



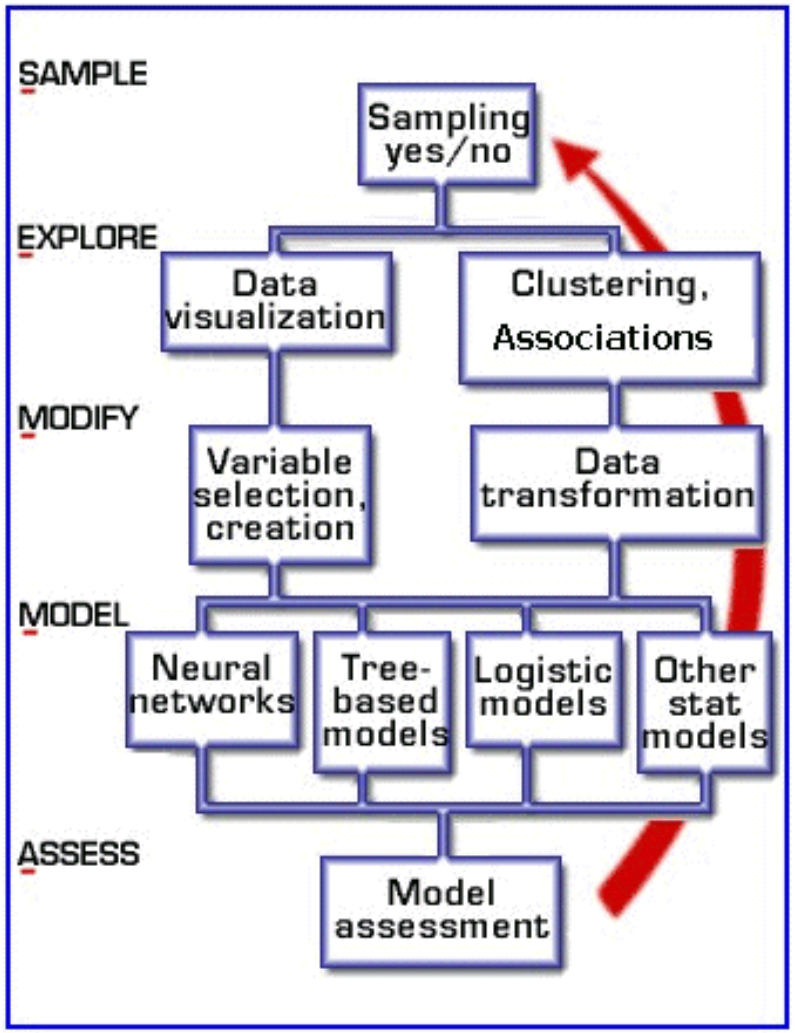
METODOLOGI SAINS DATA

KDD process: Fokus teknis



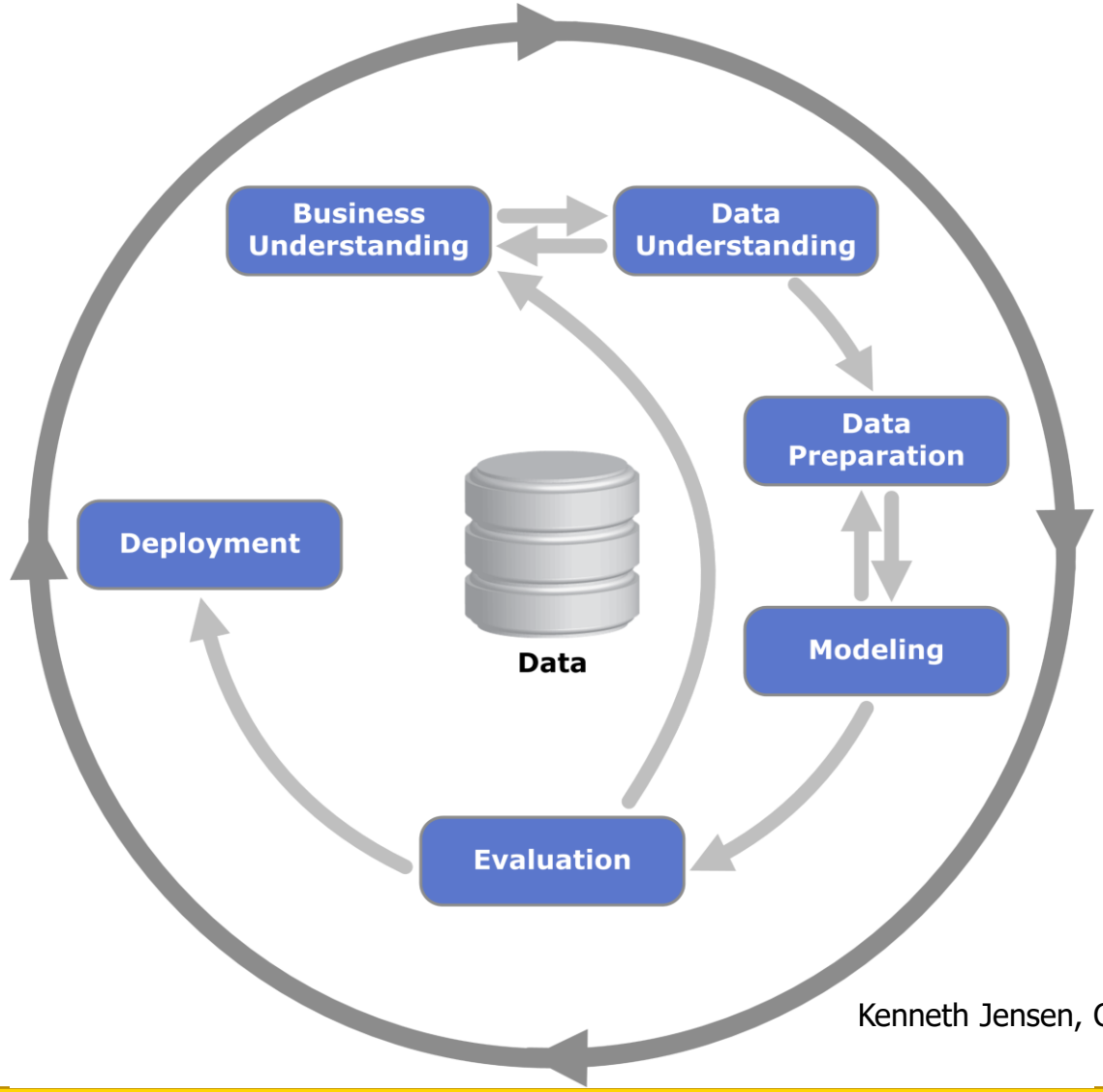
Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine 17(3): Fall 1996, 37-54
https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

SEMMA dari SAS: Fokus teknis



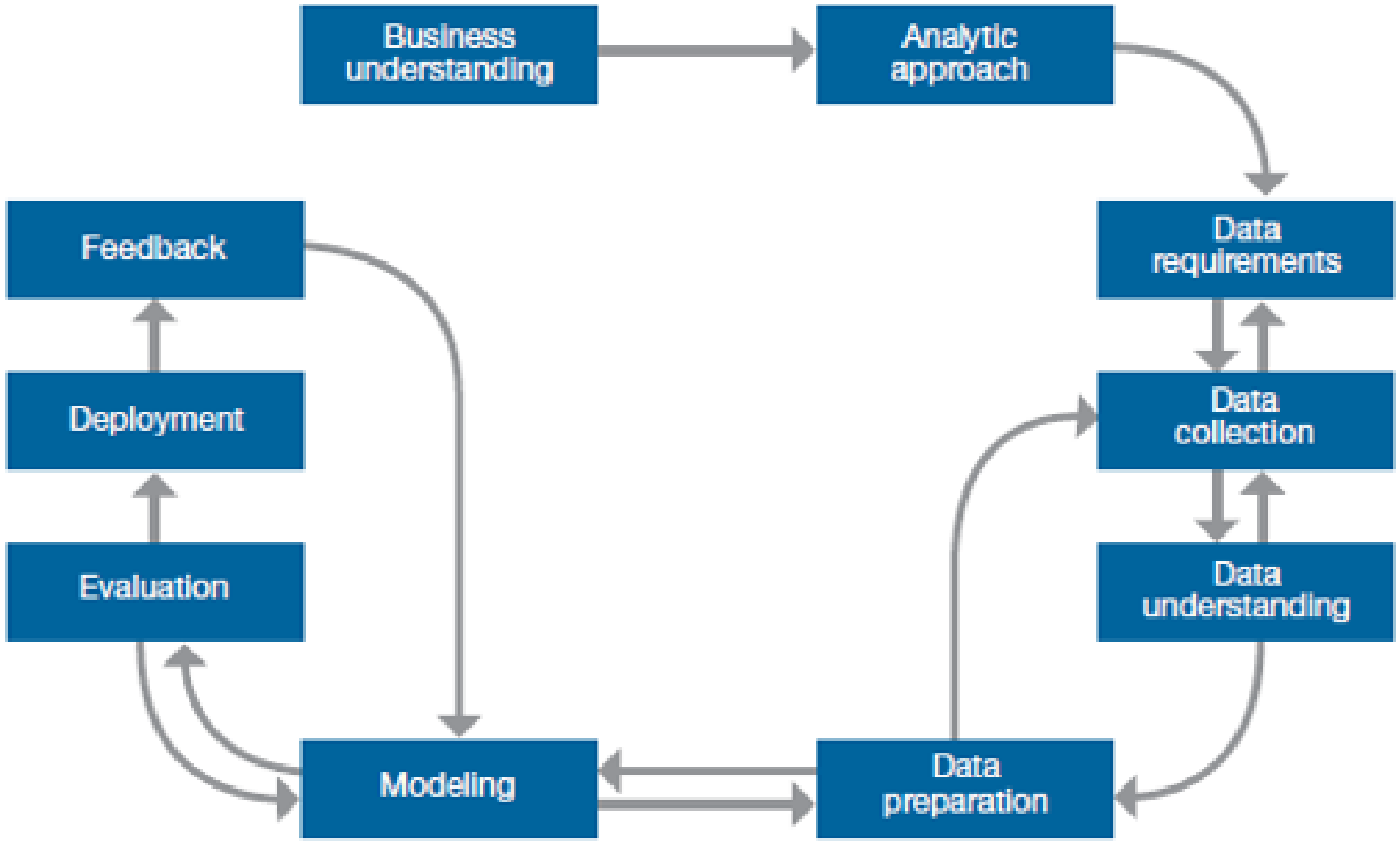
https://documentation.sas.com/doc/en/emref/15.2/n061bzurmej4j3n1jnj8bbjjm1a2.htm?homeOnFail

CRISP-DM: Cross-industry standard process for data mining (Bisnis & Teknis)



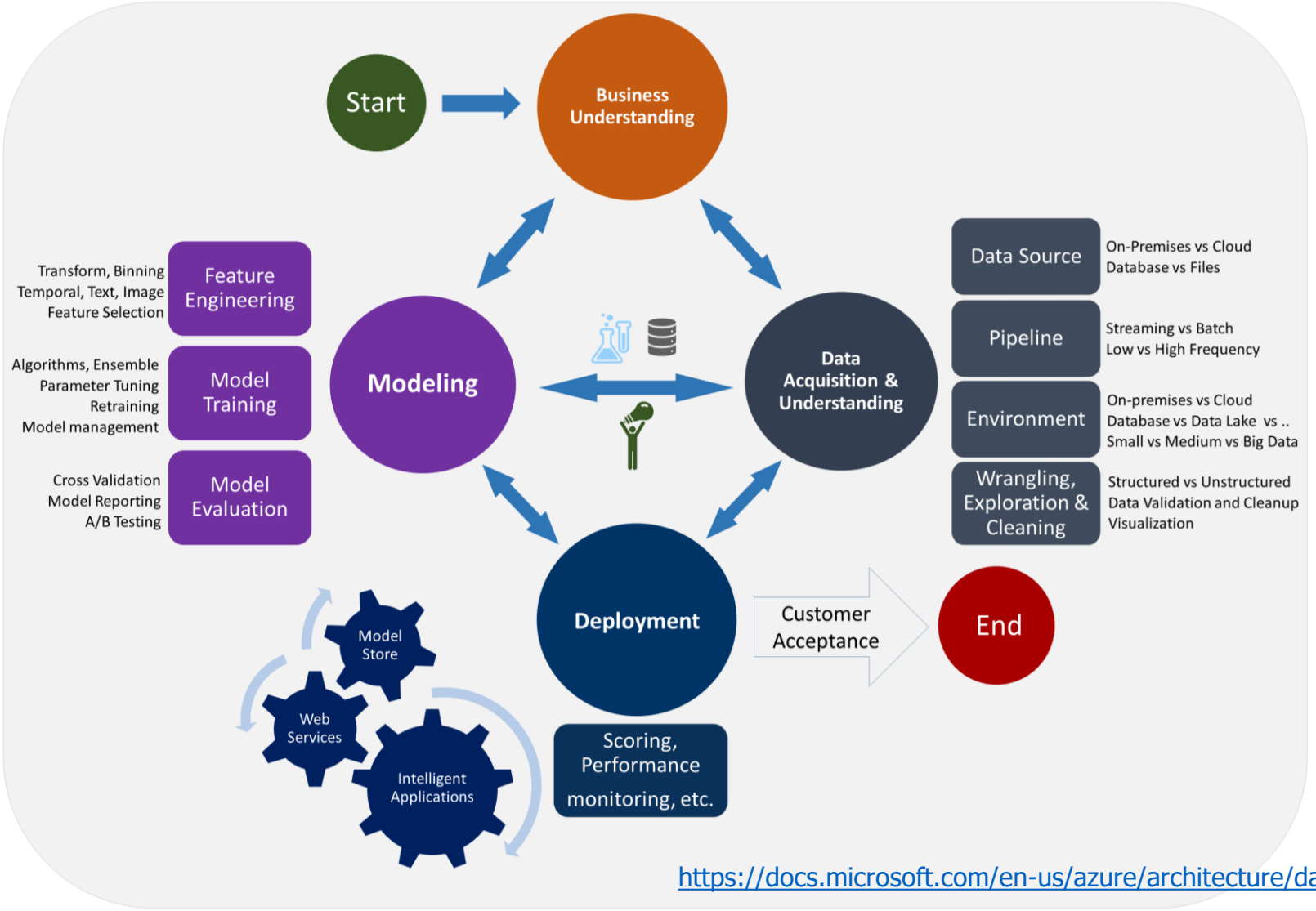
Kenneth Jensen, CC BY-SA 3.0, via Wikimedia Commons

IBM Data Science Methodology: Bisnis & Teknis



https://developer.ibm.com/articles/introduction-watson-studio/

Microsoft Team Data Science Process: Bisnis & Teknis



https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

Di Indonesia?

Standar Kompetensi Kerja Nasional Indonesia: Kepmenaker No 299 tahun 2020

MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA

KEPUTUSAN MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA
NOMOR 299 TAHUN 2020
TENTANG
PENETAPAN STANDAR KOMPETENSI KERJA NASIONAL INDONESIA
KATEGORI INFORMASI DAN KOMUNIKASI GOLONGAN POKOK AKTIVITAS
PEMROGRAMAN, KONSULTASI KOMPUTER DAN KEGIATAN YANG
BERHUBUNGAN DENGAN ITU (YBDI) BIDANG KEAHLIAN *ARTIFICIAL INTELLIGENCE* SUBBIDANG *DATA SCIENCE*

TUJUAN UTAMA	FUNGSI KUNCI	FUNGSI UTAMA	FUNGSI DASAR
Menemukan pengetahuan, *insight* atau pola yang bermanfaat dari data untuk berbagai keperluan (orang mengambil keputusan atau sistem memproses lebih lanjut)	Menganalisis Kebutuhan (Requirements) Organisasi	*Business Understanding*	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
		Data Understanding	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
	Mengembangkan model	*Data Preparation*	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
		Modeling	12. Membangun skenario pengujian 13. Membangun model
		Model Evaluation	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
	Menggunakan model yang dihasilkan	*Deployment*	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
		Evaluation	20. Melakukan review proyek 21. Membuat laporan akhir proyek

Credits

- Stuart Russell & Peter Norvig, "Artificial Intelligence: A Modern Approach", 4th ed., 2020, Section 19.9.
- Windy Gambetta, "Metodologi Data Science", Salindia Modul Pelatihan Thematic Academy, Digital Talent Scholarship, Kemenkominfo 2021.
- Joel Grus, "Data Science from Scratch: First Principles with Python", 2nd ed., O'Reilly 2019.
- Gambar dan tangkapan layar hanya untuk kebutuhan penjelasan
  - Hak cipta tetap ada pada pemilik aslinya.