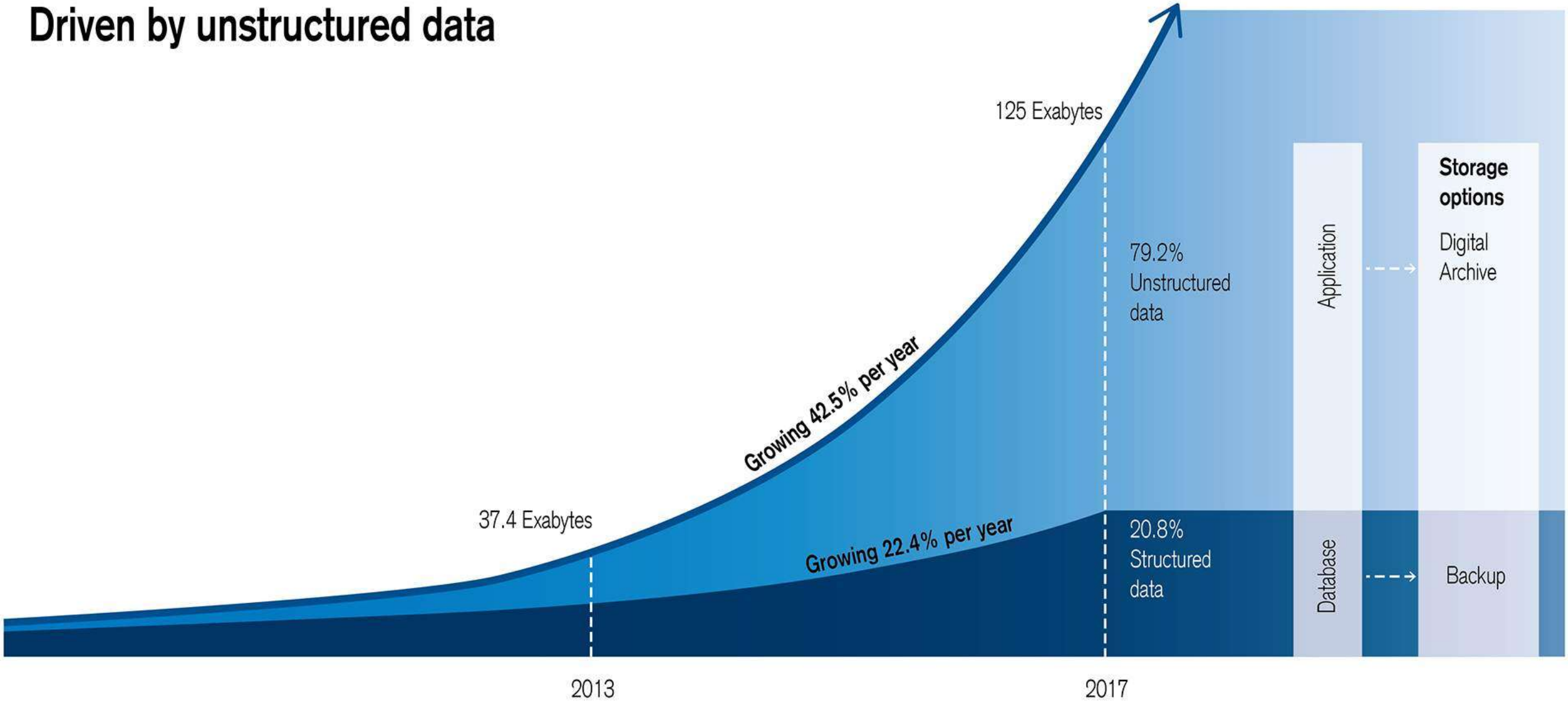
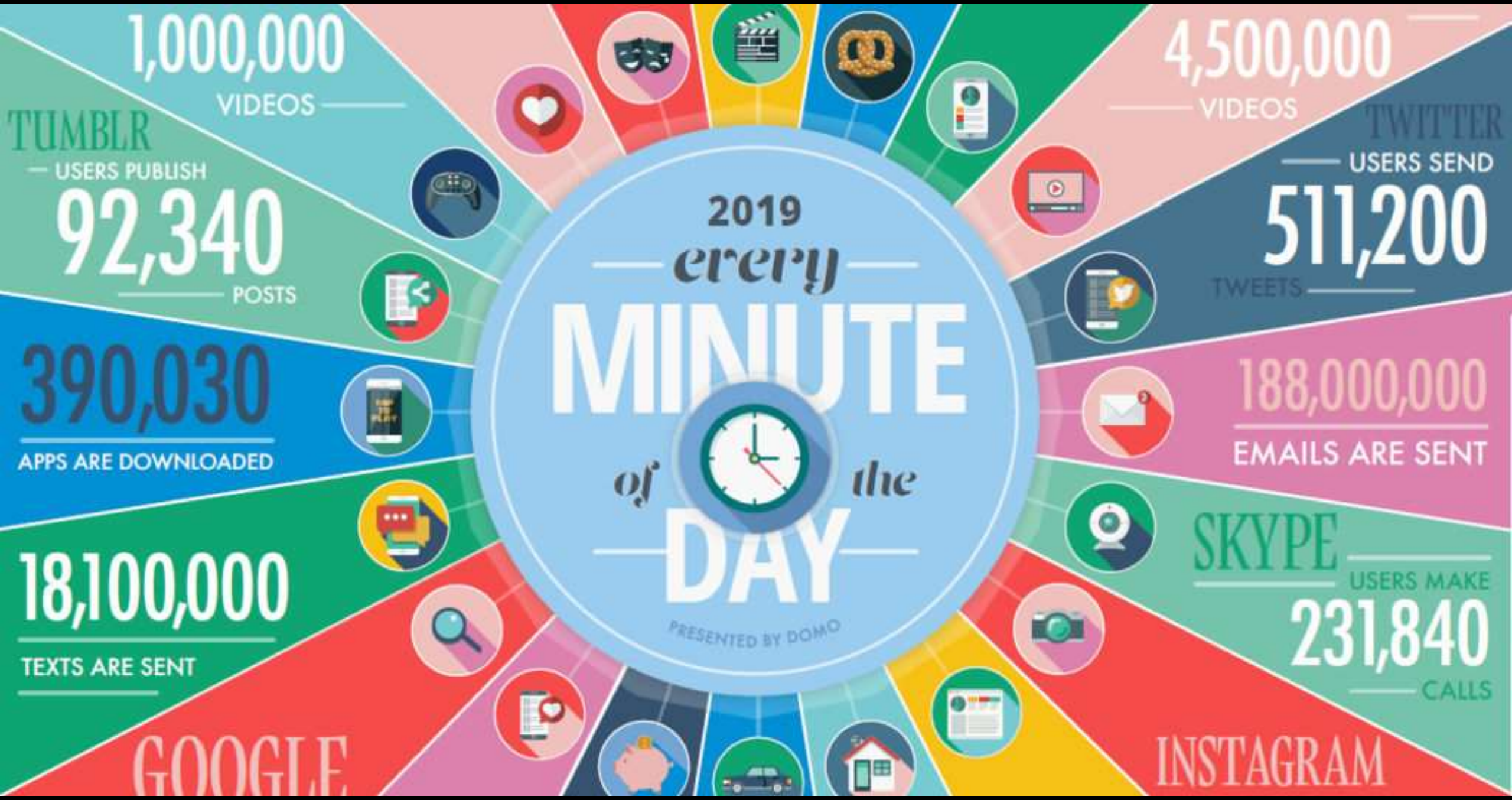
Pengenalan Sains Data

Randy Cahya Wihandika


“
The sexiest job of the 21st century.
”

Data growth
Driven by unstructured data




2019
every
MINUTE
of the
DAY
PRESENTED BY DOMO
1,000,000
VIDEOS
TUMBLR
USERS PUBLISH
92,340
POSTS
390,030
APPS ARE DOWNLOADED
18,100,000
TEXTS ARE SENT
GOOGLE
4,500,000
VIDEOS
TWITTER
USERS SEND
511,200
TWEETS
188,000,000
EMAILS ARE SENT
SKYPE
USERS MAKE
231,840
CALLS
INSTAGRAM

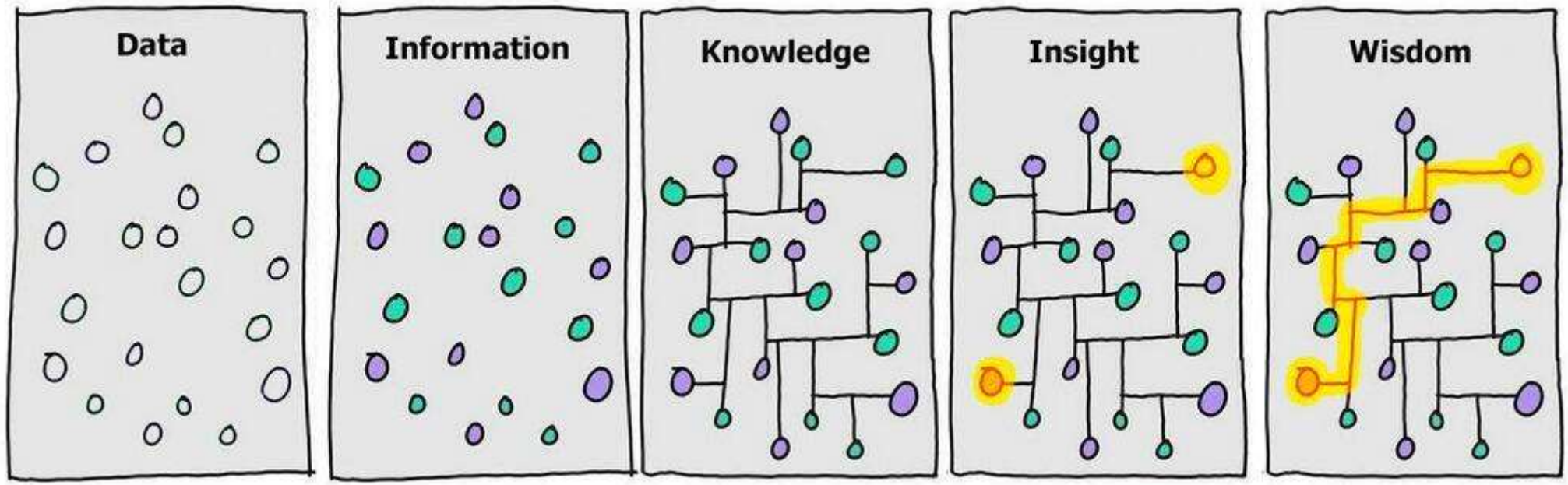

“
Data is the new oil.
”

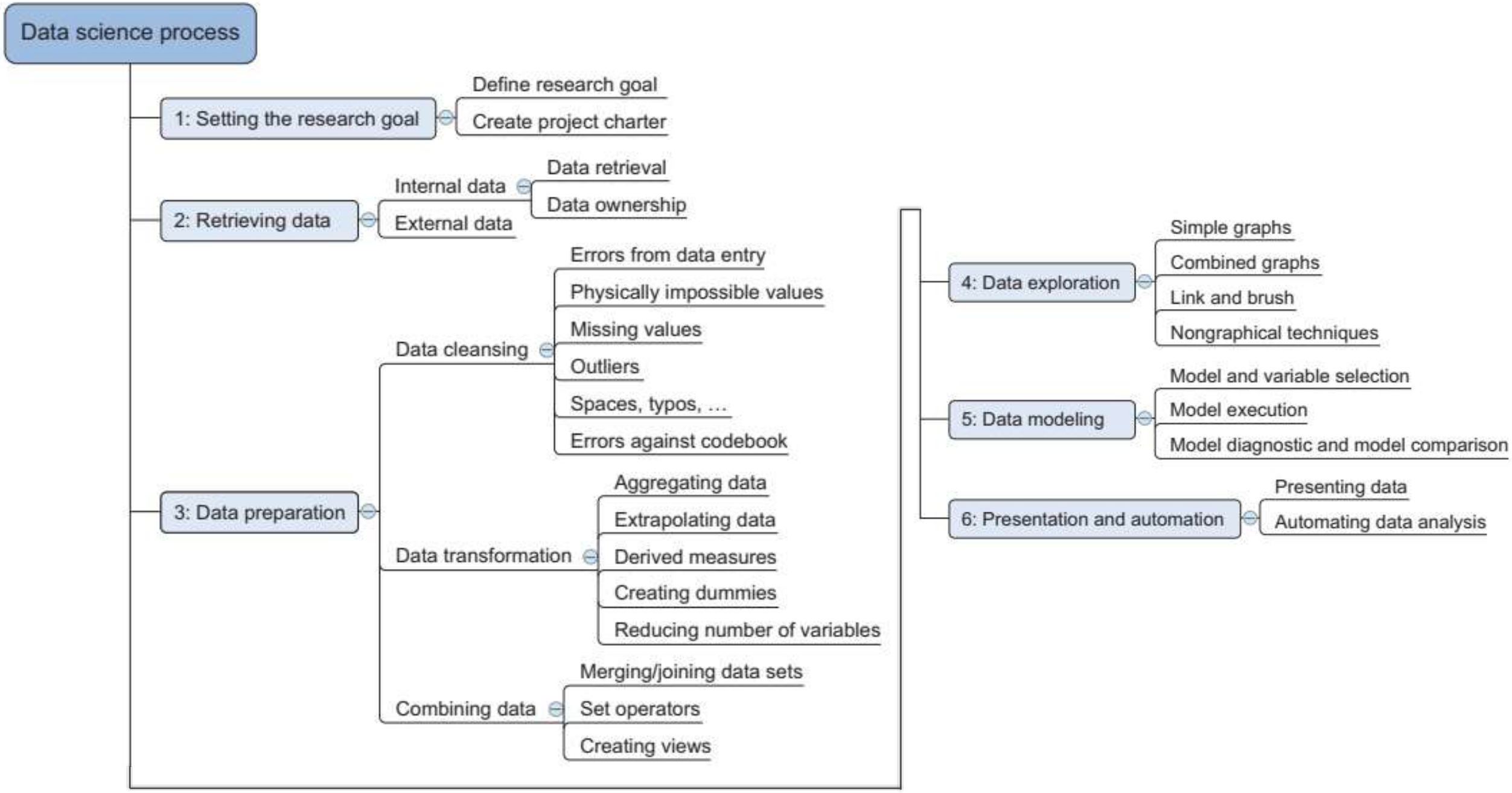
Sains Data

- Perusahaan: “Gunakan dan olah data apapun untuk meningkatkan penjualan produk.”
- Saintis data: “Diolah seperti apa? Dengan *tool* apa?”
- Perusahaan: “TERSERAH!”

Definisi

- Praktik penggunaan data untuk memahami dan menyelesaikan masalah
- Segala yang berhubungan dengan data: pengumpulan, analisis, pemodelan, dll.


Data
Information
Knowledge
Insight
Wisdom


Data science process
1: Setting the research goal
Define research goal
Create project charter
2: Retrieving data
Internal data
Data retrieval
Data ownership
External data
3: Data preparation
Data cleansing
Errors from data entry
Physically impossible values
Missing values
Outliers
Spaces, typos, …
Errors against codebook
Data transformation
Aggregating data
Extrapolating data
Derived measures
Creating dummies
Reducing number of variables
Combining data
Merging/joining data sets
Set operators
Creating views
4: Data exploration
Simple graphs
Combined graphs
Link and brush
Nongraphical techniques
5: Data modeling
Model and variable selection
Model execution
Model diagnostic and model comparison
6: Presentation and automation
Presenting data
Automating data analysis

1. *Research Goal*

- Semua *stakeholder* harus memahami tujuan proyek *data science* yang dibuat
- *What*, *how*, dan *why*

2. *Data Retrieval*

- Mencari, mendapatkan akses, dan mengumpulkan data
- Data yang didapatkan kemungkinan masih “mentah” yang belum dapat diproses

3. *Data Preparation*

- "Membersihkan" data dari *error*
- Mengubah format data "mentah" menjadi siap untuk digunakan
- Sebagian besar waktu (sekitar 80%) digunakan pada tahap ini
- *"Garbage in equals garbage out."*

4. *Data Exploration*

- Mendapatkan pemahaman terhadap data
- Dilakukan secara manual dengan bantuan *tool*
- Mencari pola, korelasi, dan perhitungan-perhitungan statistik lainnya
- Membuat grafik data

5. *Data Modeling*

- Mengaplikasikan algoritma-algoritma *machine learning* untuk mendapatkan informasi penting
- Tidak harus menggunakan algoritma yang kompleks
- Tujuan sebenarnya adalah menghasilkan *impact* sebesar mungkin

6. *Data Presentation and Automation*

- Menyajikan hasil sehingga dapat dipahami semua *stakeholder*: manajemen, investor, UX *designer*, dll.
- Mengotomasi proses analisis, jika diperlukan

Proses Sains Data

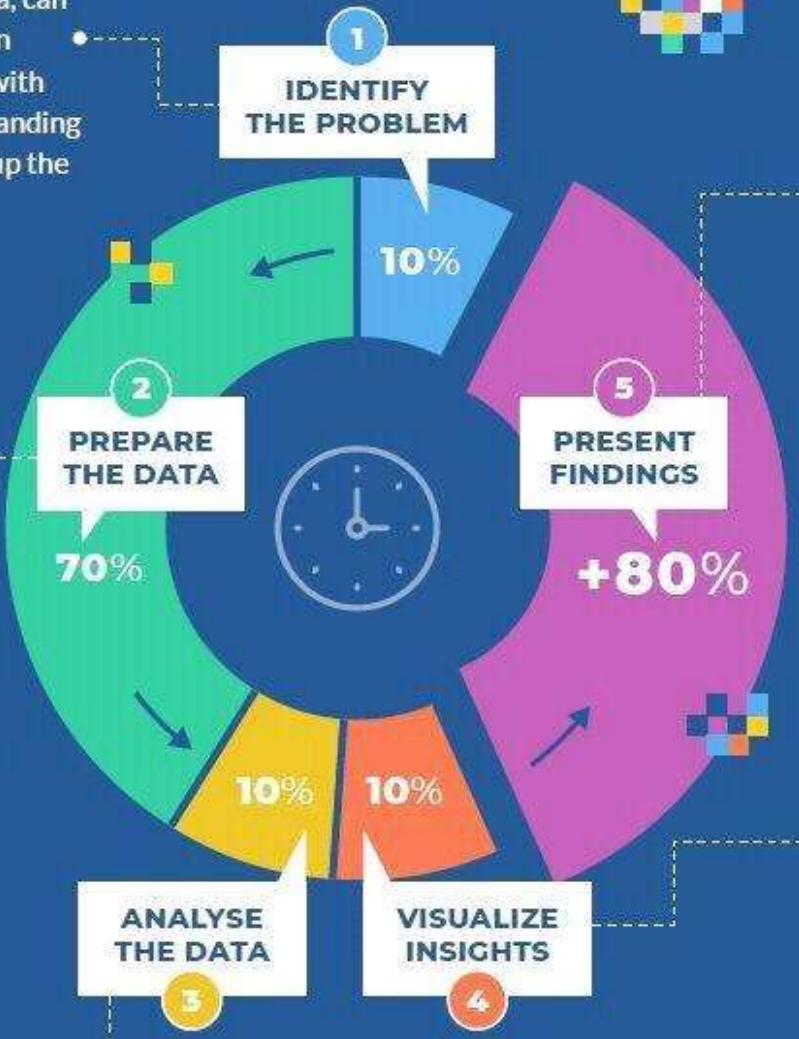
- Langkah-langkah tadi tidak bersifat linier dan tidak mengikat
- Namun penting untuk digunakan untuk menghasilkan proyek yang sukses
- Membantu mendefinisikan rencana dan produk/*deliverables* yang jelas

LIFECYCLE OF A DATA SCIENCE PROJECT

Ever heard the phrase "Here's some data, can you find some insights?" Right? Too often stakeholders approach Data Scientists with vague or even undefined goals. Understanding the end goal is very important and sets up the rest of the project for success.

By far, everybody's least favourite stage, but perhaps the most important one. Data can come from many sources, be in the wrong format, have anomalies and a myriad of other problems. A single mistake in this stage can render the rest of the analysis useless.

Creating models, performing data mining, running text analytics, setting up simulations - the list goes on! This is the most fun and exciting part and if the previous stages have been done correctly, analyzing the data and deriving insights will feel like a breeze.



We've reached 100% the project is over! Actually, not yet. Presenting findings is a whole separate "Bonus" stage. You need to not only convey the insights in your audience's language but also get buy-in from them to take action based on those insights. This is an art in its own right.

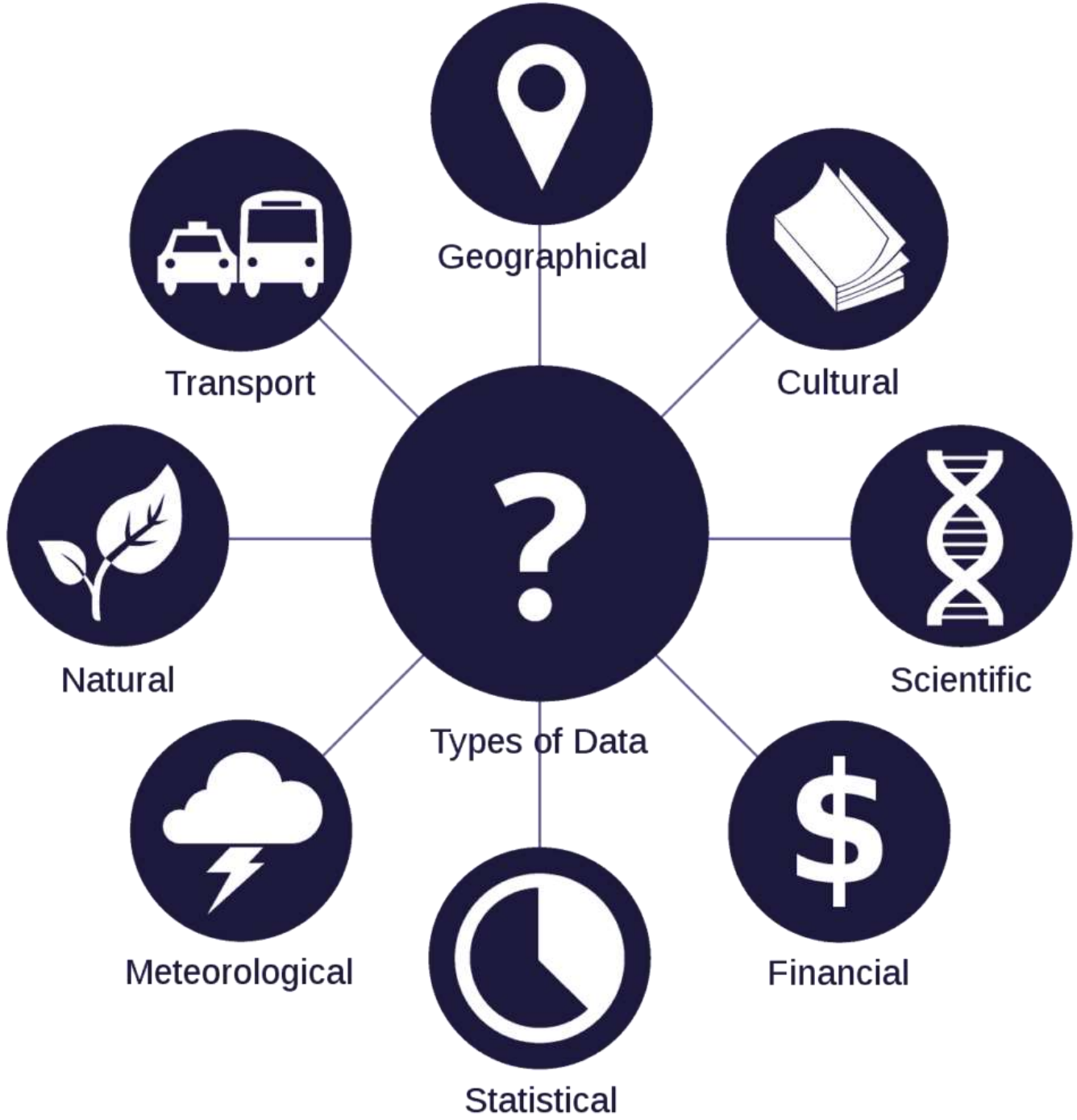
Visualizing comes hand-in-hand with analyzing. This is a very powerful technique as seeing the data in various forms and shapes can help uncover insights that are otherwise not evident. Also some projects such as BI dashboards don't require much analysis but rely heavily on visualization instead.

Modul Pembelajaran

Belajar dengan materi pembelajaran yang menyeluruh dalam 3 bulan

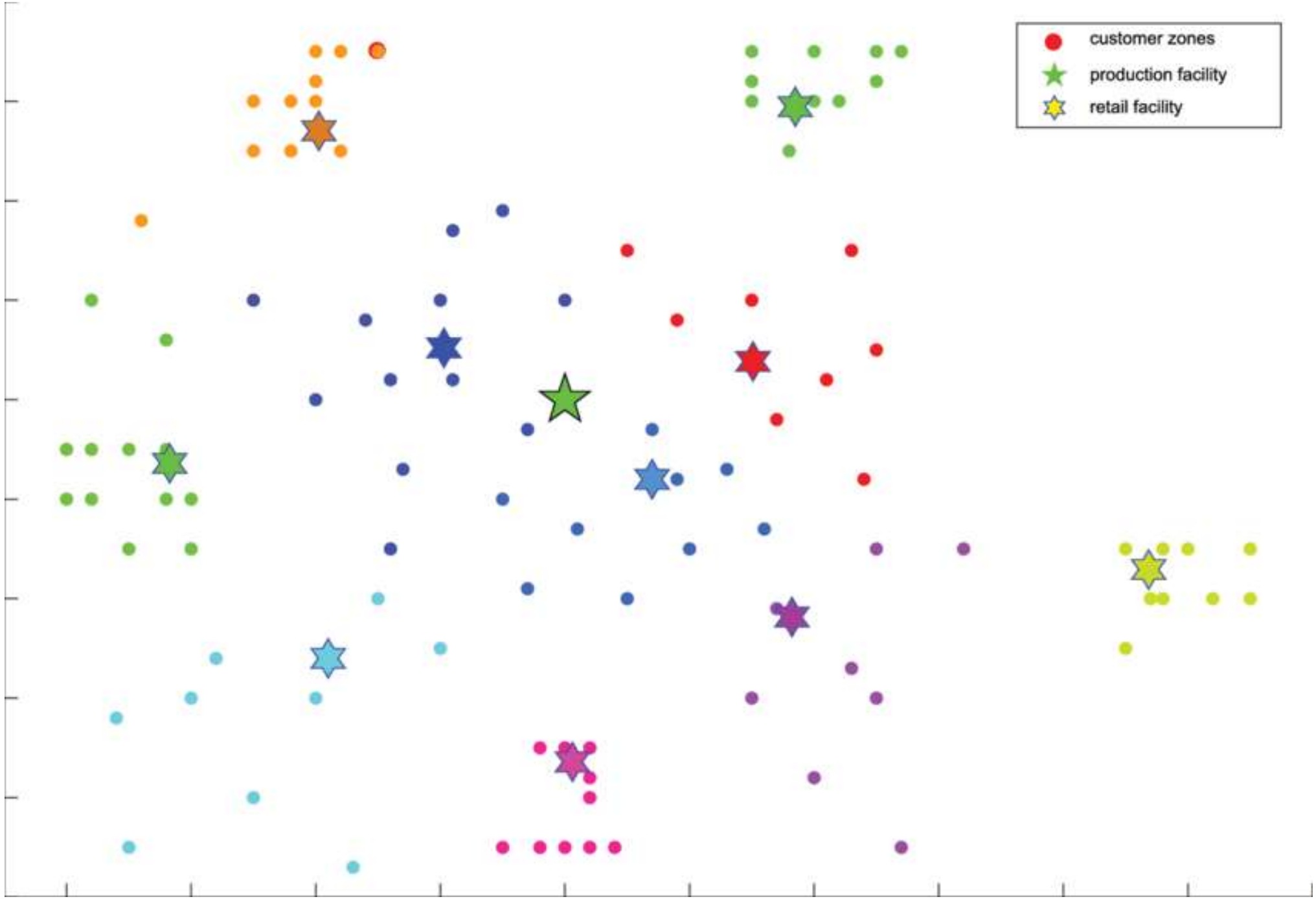
- Data Science Methodology
- Introduction to Data and Database
- SQL
- Versioning/Version Control
- Introduction to Python Programming
- Python Programming
- Introduction to Numpy
- Introduction to Basic Dataframe (Pandas)
- Dataset dan Modelling Project
- Analytical and Critical Thinking
- Dataframe
- API
- Statistics
- Data Visualization
- Data Preprocessing for Machine Learning
- Introduction to Machine Learning and Regression
- Machine Learning - Supervised & Unsupervised
- Advance Machine Learning Topics
- Business Intelligence
- Mentor Experience Sharing
- Communication and Presentation Skills
- Career Coaching and Mentoring

Sumber: https://www.digitalskola.com/skolaclass/data-science.html


Geographical
Transport
Cultural
Natural
?
Scientific
Types of Data
Meteorological
$
Financial
Statistical

Contoh: Perusahaan *Retail*

- Suatu perusahaan *retail* berencana membuka cabang toko baru
- Di mana lokasi yang terbaik?
- Saintis data dapat melihat data lokasi pembeli
- Atau dikombinasikan dengan data demografi dan gaji pembeli
- Kemudian saintis data mempresentasikan rekomendasinya


customer zones
production facility
retail facility

Contoh: *e-Commerce*

- Website menampilkan rekomendasi produk yang sifatnya *personalized* berdasarkan data histori pembelian

For Randy | Special Discount | Produk Populer WIB | Tempat Sampah | Sikat | Mata G

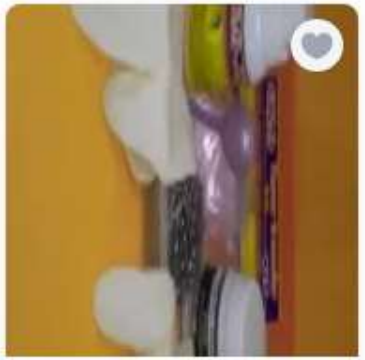
Alat pemotong kaca/pisau potong kaca
Rp35.000
Jakarta Pusat
★★★★★ (1)

Cashback
Tang Potong Dorong Pemotong Keramik Kaca
Rp52.000
Kab. Bekasi

Cashback
Alat Pemotong Kaca / Glass Cutter / Pisau
Rp70.000
Jakarta Timur
★★★★★ (21)
Grosir

Alat Pemotong Kaca Engraving Model Pulpen
Rp20.600
Tangerang Selatan
★★★★★ (21)

Solder listrik 100w 220 vac gagang kayu- mata
Rp 60.000
Kota Malang
★★★★★ (1)



GK9 kit toner transfer pcb dengan gk9 formula
Rp 75.000
Kota Malang
★★★★★ (13)

Cashback

Cashback

Cashback

Contoh: *Chicken Nuggets*

- Jika telur menetas hari ini, berapakah perkiraan bobot ayam 30 hari yang akan datang?

Contoh: *Customer Churn*

- Suatu perusahaan TV berlangganan tentu ingin mempertahankan pelanggannya
- Berapakah persentase kemungkinan pelanggan berpindah ke produk kompetitor?
- Ini dikenal dengan *customer churn*

Contoh: *Fraud Detection*

- Mencegah terjadinya transaksi ilegal
- Digunakan pada banyak industri, seperti perbankan, asuransi, dan pinjaman *online*
- Kendala yang sering terjadi: *class imbalance*

Contoh: *Targeted Ads*

- Menampilkan iklan yang berbeda-beda pada setiap orang

https://www.facebook.com

Search Facebook

Kate Home 2

Facebook Marketing Partners

Sponsored ·

Like Page

Great things happen with the right partner. Find yours now and supercharge your marketing.

Facebook Marketing Partners

Move fast with confidence with the help of best-in-class partners. From managing complexity to adding scale to reaching new audiences, Marketing Partners can help you get more from your marketing.

WWW.FACEBOOKMARKETINGPARTNERS.COM

Learn More

Like · Comment · Share · 123 9 6

SUGGESTED GROUPS See All

London Startup Events
Kenan Mujezinovic joined
+ Join

SPONSORED Create Ad

€22.8 & Envío gratis
es.sheinside.com
Forme el vestido €22.8 & Envío gratis! Adicional 30% Off En 1r pedido,Compras Tu Amor!

Chat (39)

RESEARCH ARTICLE

Private traits and attributes are predictable from digital records of human behavior

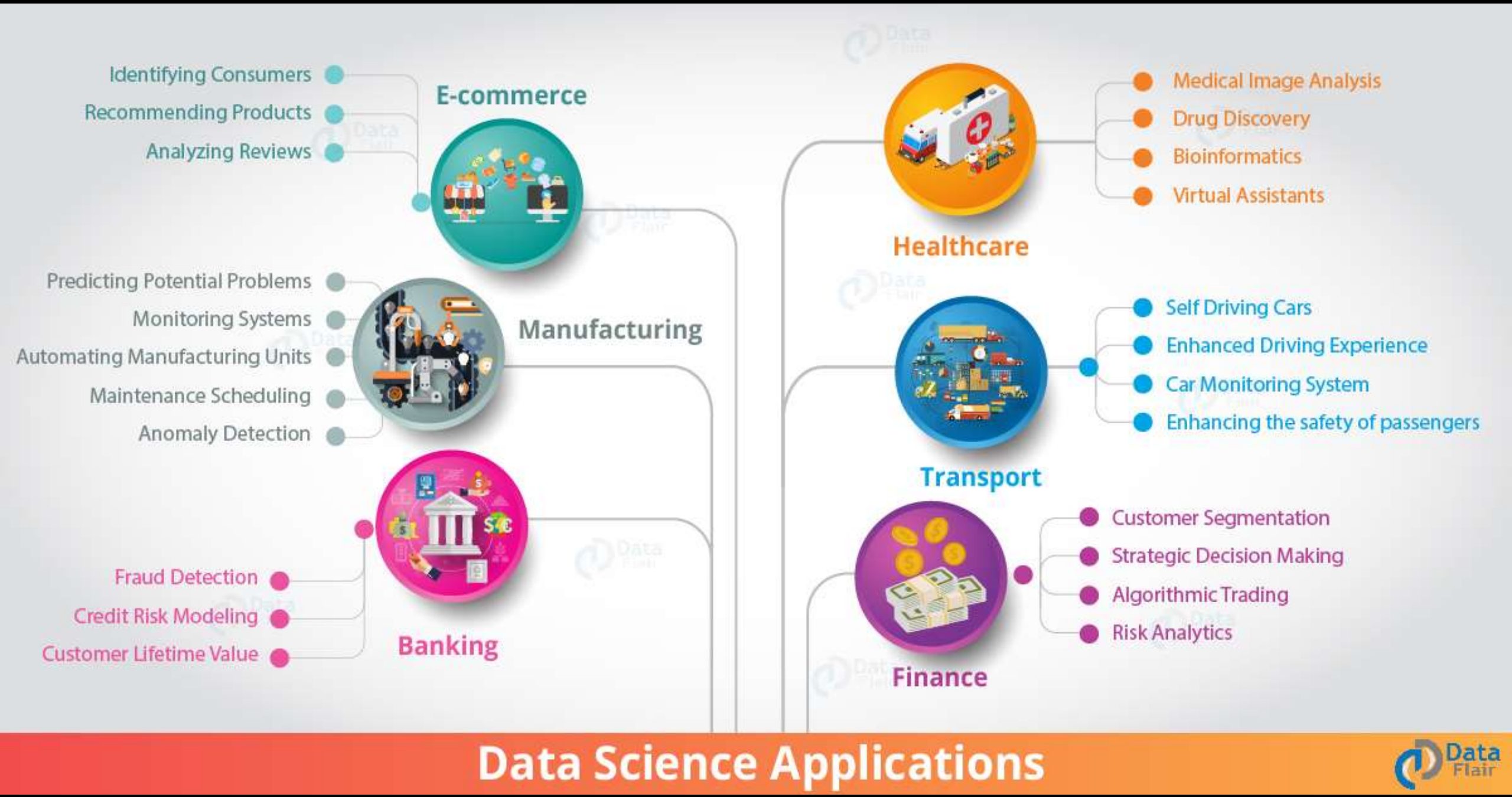
Michal Kosinski, David Stillwell, and Thore Graepel

PNAS April 9, 2013 110 (15) 5802-5805; https://doi.org/10.1073/pnas.1218772110

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

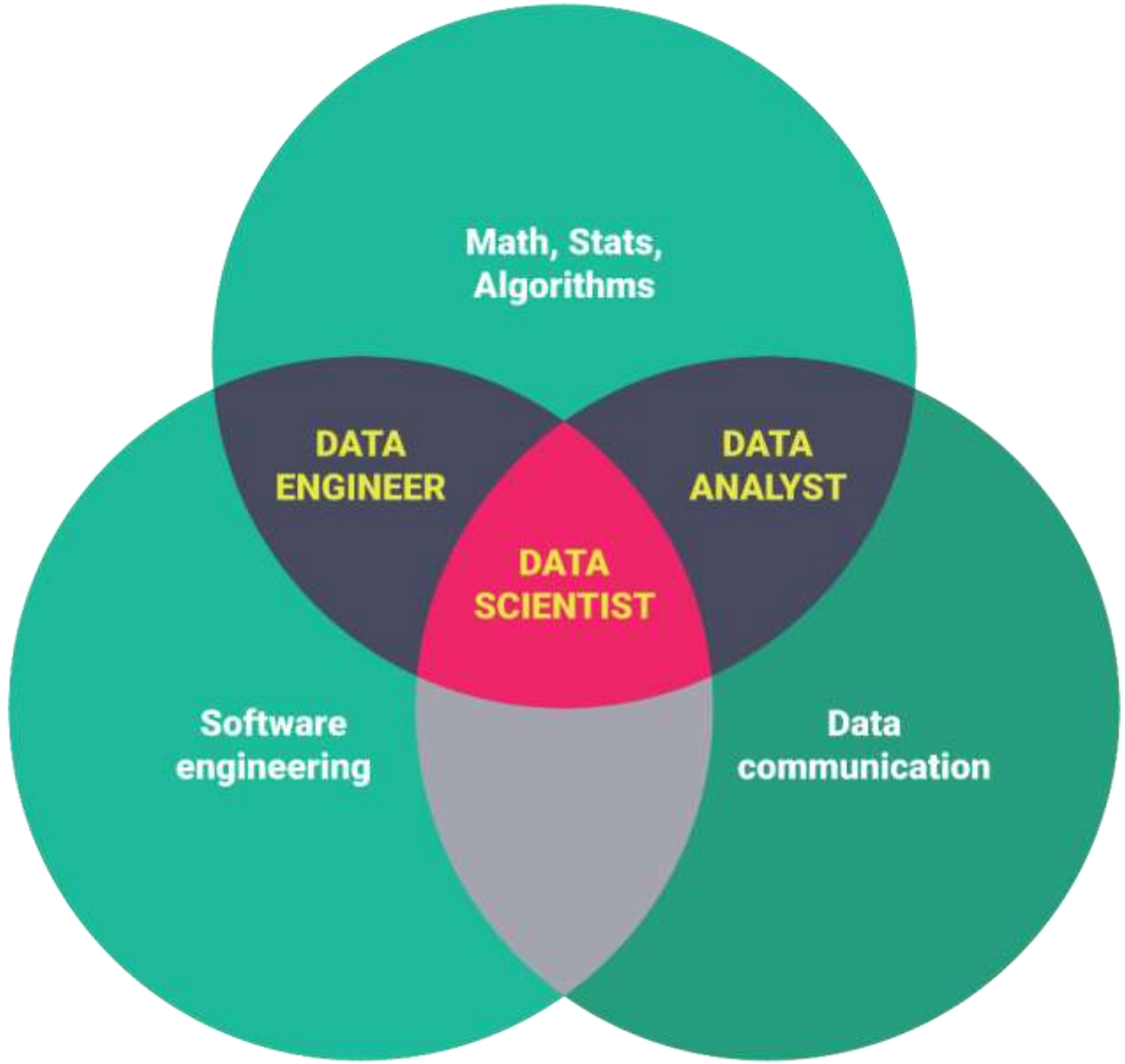
Abstract

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization and privacy.

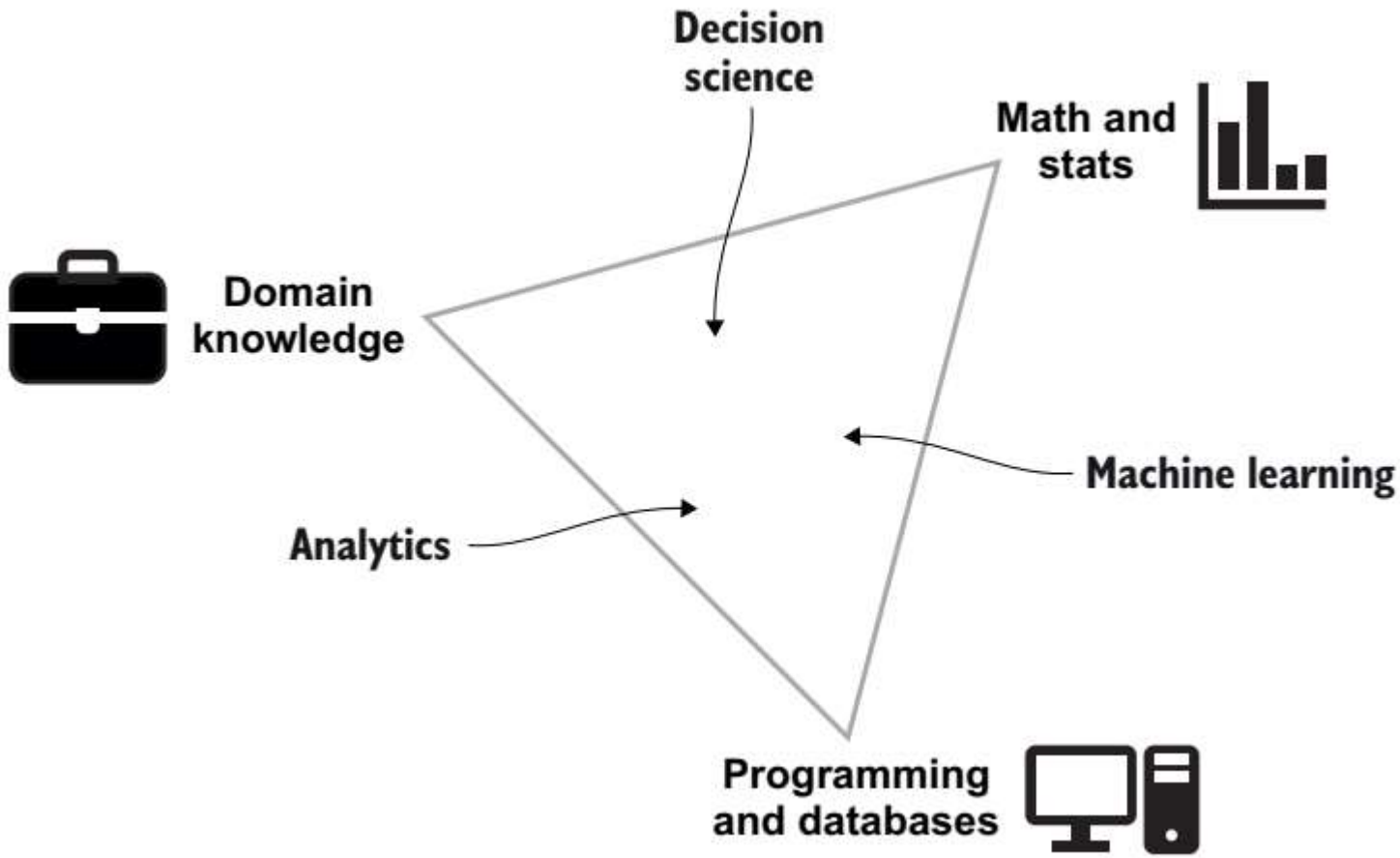

Identifying Consumers
Recommending Products
Analyzing Reviews
E-commerce
Predicting Potential Problems
Monitoring Systems
Automating Manufacturing Units
Maintenance Scheduling
Anomaly Detection
Manufacturing
Fraud Detection
Credit Risk Modeling
Customer Lifetime Value
Banking
Medical Image Analysis
Drug Discovery
Bioinformatics
Virtual Assistants
Healthcare
Self Driving Cars
Enhanced Driving Experience
Car Monitoring System
Enhancing the safety of passengers
Transport
Customer Segmentation
Strategic Decision Making
Algorithmic Trading
Risk Analytics
Finance
Data Science Applications
Data Flair

Skills

- Bisa berasal dari beragam latar belakang ilmu
- *Structured query language* (SQL)
- *Big data*
- Pemrograman: Python atau R
- Matematika dan statistika
- *Machine learning*
- *Domain knowledge*
- *Version control*: Git


Math, Stats,
Algorithms
DATA
ENGINEER
DATA
ANALYST
DATA
SCIENTIST
Software
engineering
Data
communication

Skills

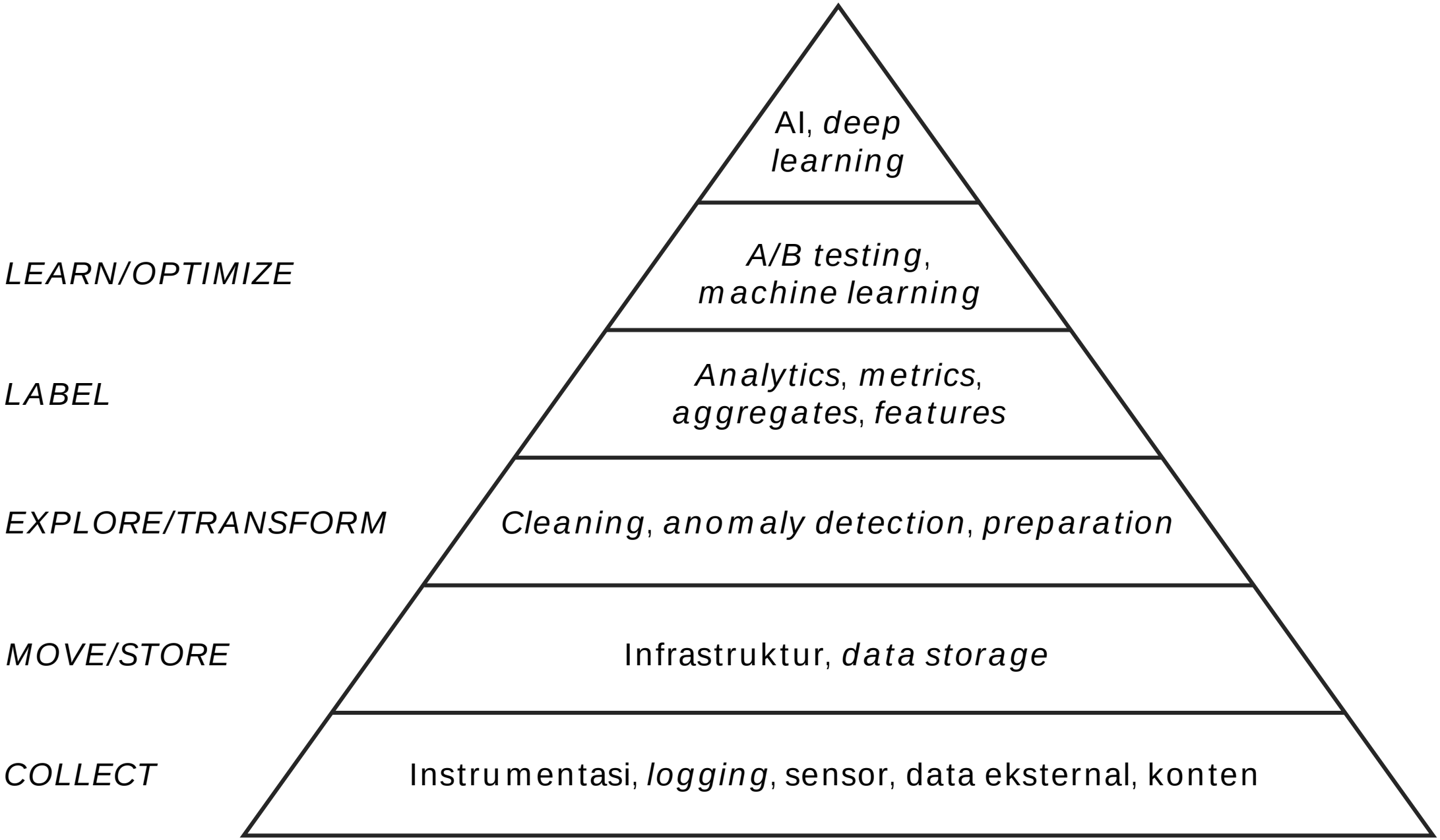

Decision science
Math and stats
Domain knowledge
Machine learning
Analytics
Programming and databases

Soft Skills

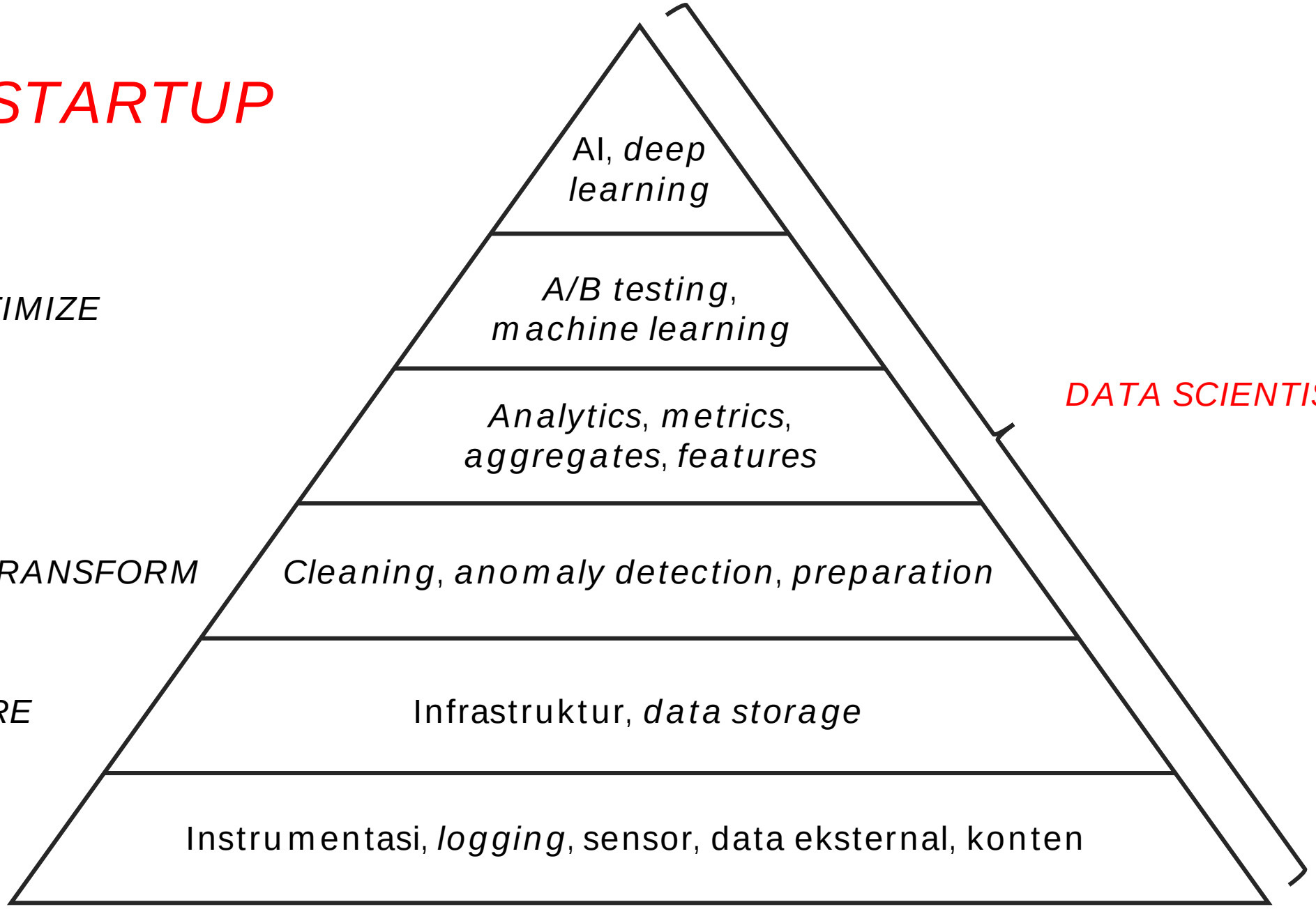
- *Curiosity*
- Kreativitas
- Kesabaran
- Ketekunan
- Komunikasi

Roles

- *Data scientist* tidak bekerja sendiri
- Ada *software engineer*, *data engineer*, dan *research scientist*


LEARN/OPTIMIZE
LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, deep learning
A/B testing, machine learning
Analytics, metrics, aggregates, features
Cleaning, anomaly detection, preparation
Infrastruktur, data storage
Instrumentasi, logging, sensor, data eksternal, konten

STARTUP



LEARN/OPTIMIZE

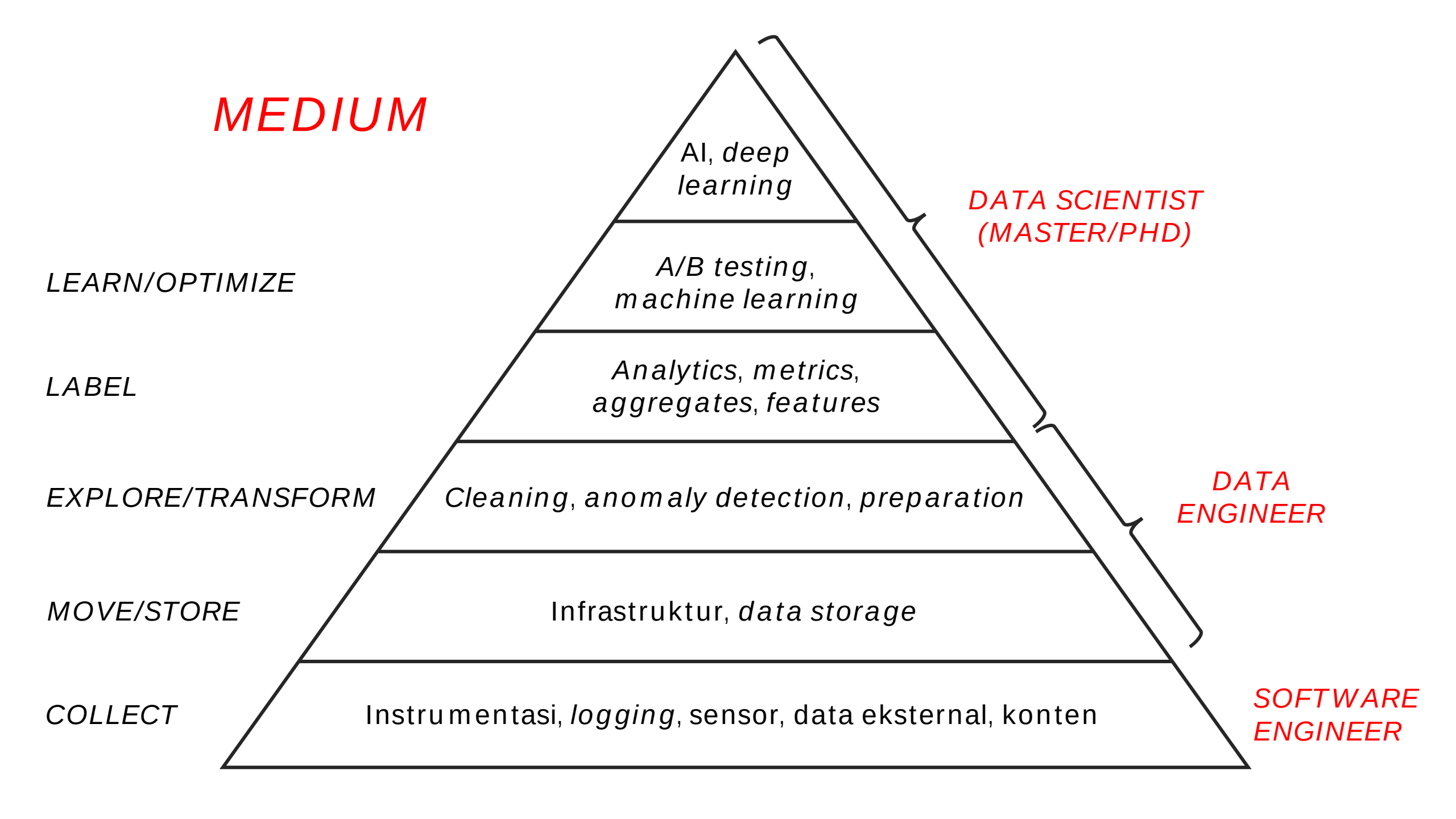
LABEL

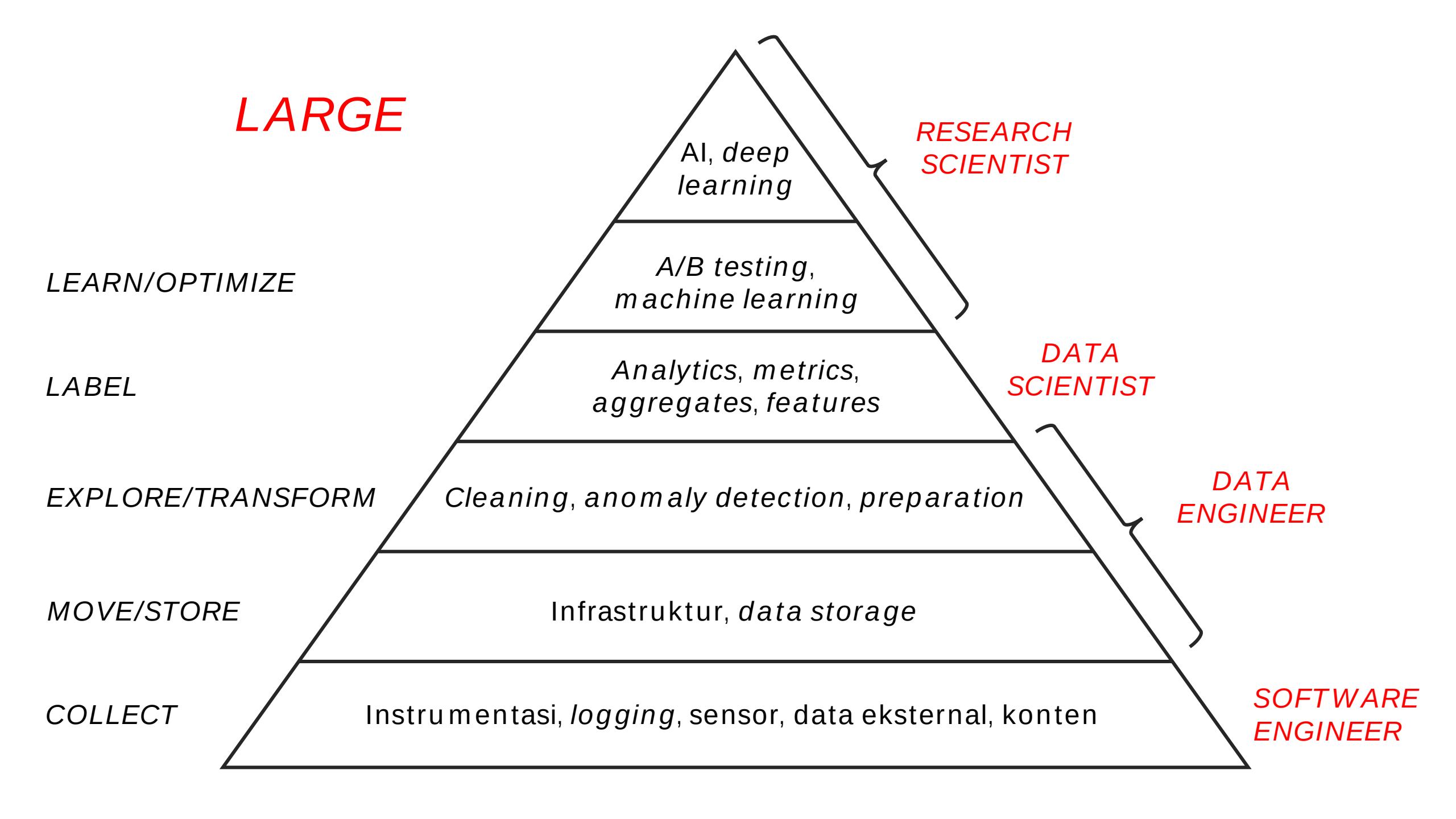
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

DATA SCIENTIST


MEDIUM
LEARN/OPTIMIZE
LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, deep learning
A/B testing, machine learning
Analytics, metrics, aggregates, features
Cleaning, anomaly detection, preparation
Infrastruktur, data storage
Instrumentasi, logging, sensor, data eksternal, konten
DATA SCIENTIST (MASTER/PHD)
DATA ENGINEER
SOFTWARE ENGINEER


LARGE
AI, deep learning
A/B testing, machine learning
Analytics, metrics, aggregates, features
Cleaning, anomaly detection, preparation
Infrastruktur, data storage
Instrumentasi, logging, sensor, data eksternal, konten
LEARN/OPTIMIZE
LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
RESEARCH SCIENTIST
DATA SCIENTIST
DATA ENGINEER
SOFTWARE ENGINEER

Algoritma

- *Natural Language Processing*
- *Classification*
- *Clustering*
- *Ensemble methods*
- *Deep learning*
- dll.


See you!