Introduction to Data Science

2025. 8

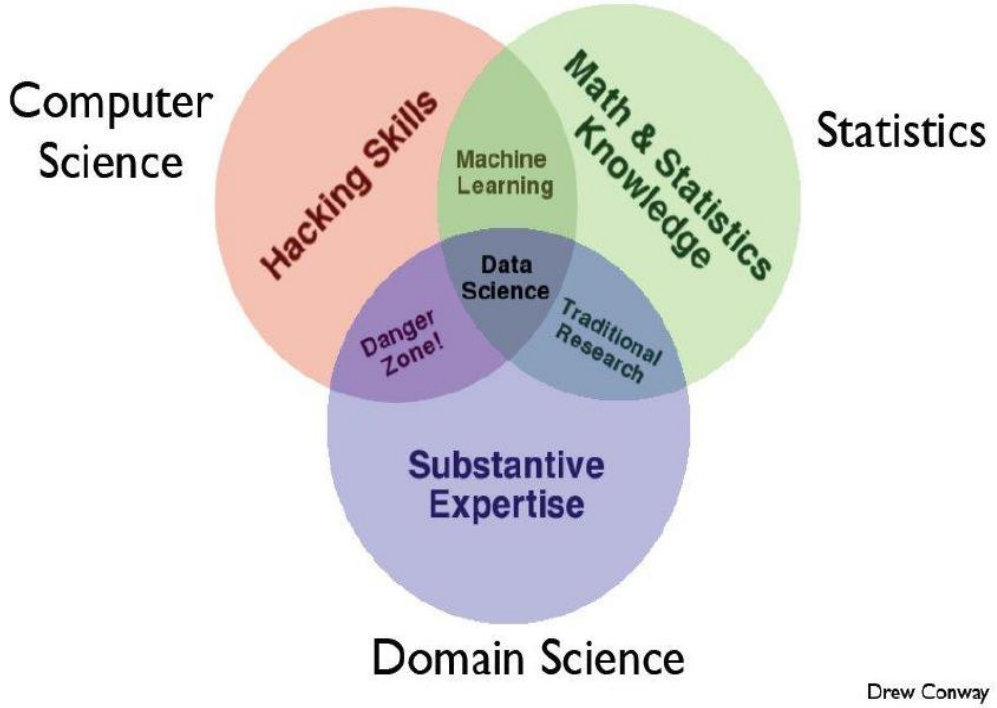
Yongjin Jeong, KwangWoon University

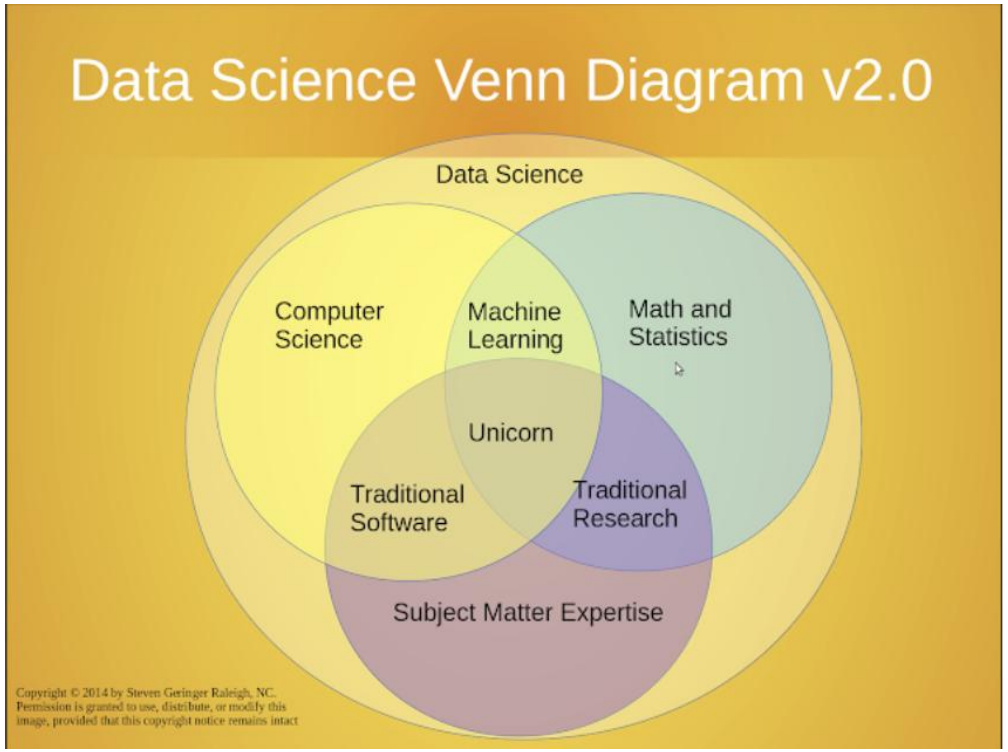
What is Data Science?

- **Definition (from Wikipedia)**
  - ✓ concept to unify statistics, data analysis, machine learning, domain knowledge and their related methods in order to understand and analyze actual phenomena with data
  - ✓ the application of computational and statistical techniques to address or gain insight into some problem in the real world (Pat Virtue, CMU)
  - ✓ It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge, and information science
- **Components of Data Science**
  - ✓ Software Programming -> Data mining, Database
  - ✓ Statistics/mathematical modeling -> Machine Learning, Scientific Computing
  - ✓ Domain Knowledge -> Data driven business analytics
- **Data Science** (Pat Virtue, CMU)
  - ✓ = **statistics** + **data processing** + **machine learning** + **scientific inquiry** + **visualization** + **business analytics** + **big data** + …

What is Data Science? – One definition

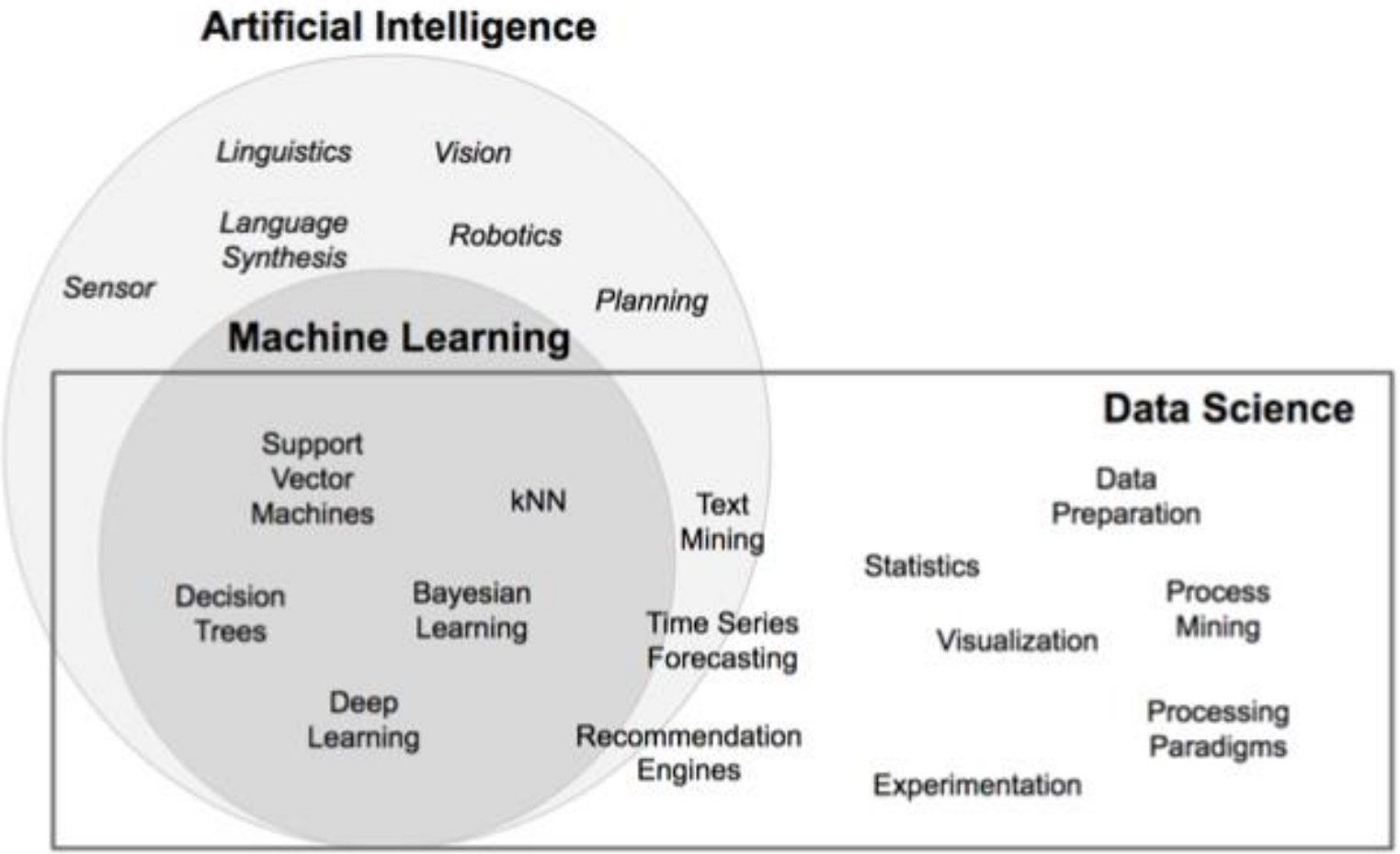
- **Venn Diagrams (Drew Conway 2010, Steven Geringer 2014)**





(*) 이 세 가지 분야를 모두 마스터한 사람을 '유니콘'이라 부를 정도로 드물기 때문에, 현대의 데이터 사이언스는 각 분야 전문가들이 협력하는 '팀 스포츠'의 성격이 강하다.

What is Data Science?



ref: Data Science: Concepts and Practice (by V.Kotu and B.Deshpande), 2022

Data in Data Science

- **Contrast to Databases**

	Databases	Data Science
data values	"Precious"	"Cheap"
data volume	Modest	Massive
examples	Bank Records, Personal Records, Census, Medical Records	On line clicks, GPS logs, Tweets, Web surfing, building censor readings
structured	Strongly (Schema)	Weak or None (Text)
priority	Consistency, Error recovery, auditability	Speed, Availability, Query richness
realizations	SQL	No SQL Python, R, TensorFlow, Keras
	Querying the Past	Querying the Future

ref: CS194 berkeley.edu by John Canny

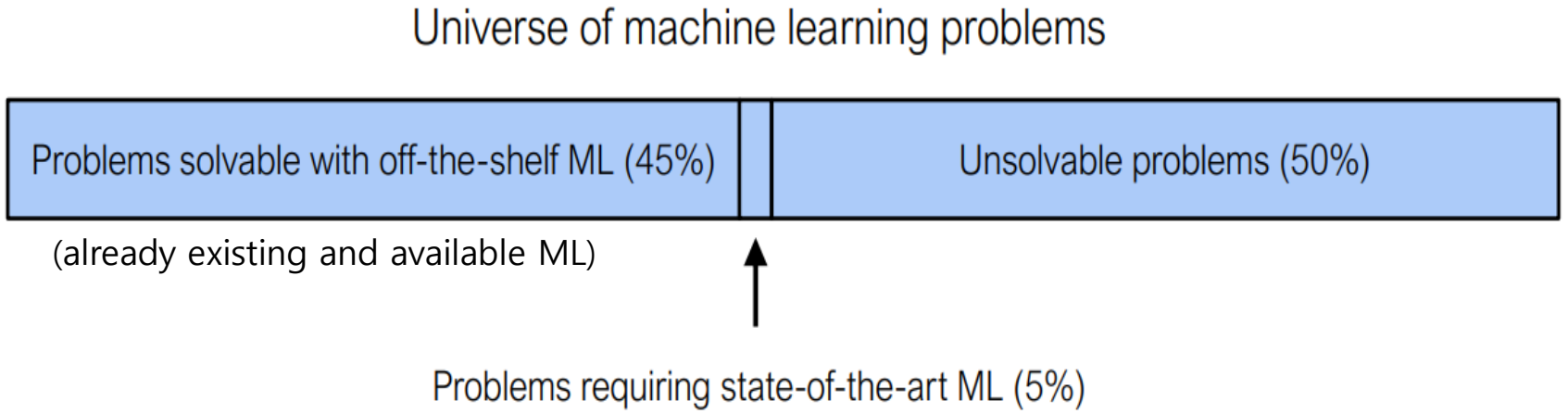
- "Garbage In, Garbage Out"
- 데이터의 질과 양이 AI의 성능을 결정
- 구글, 메타, 넷플릭스, OpenAI 등 세계 최강 기업들은 모두 '데이터 기업

Statistics vs. Data Science

	Statistics	Data science
Common	aim to extract knowledge from data - exploratory data analysis & Visualization - characterization & prediction - use sample data to make conclusions about either **future data (machine learning)** or **the population (statistics)**	
Main goal	- Inference and **explanation** - emphasis on **understanding** relationships and estimating population parameters	- **prediction** accuracy and complex datasets - focus on accurate predictions often **without detail understanding** or interpretability
Language	estimating (inferencing) data point/observation independent variable dependent variable dummy variable	learning (training) & prediction example/instance/sample feature label or target one-hot encoding
Data	small or medium sized mostly structured more manual data collection (or surveys) In general, no web scraping or data processing	huge (big data) structured or unstructured more data collection/acquisition (from web and S NS)
Processing	query (past)	predict (future)
Tools	Mathematics prefer R SAS (statistics package)	programming (prefer Python) ML libraries (sklearn, tensorflow, etc.)

Some Comments from Experts

- **Data science is not machine learning** (Pat Virtue, CMU)
  - Machine learning involves computation and statistics, but has not (traditionally) been very concerned about answering scientific questions
  - Machine learning has a heavy focus on fancy algorithms…
  - … but sometimes the best way to solve a problem is just by visualizing the data, for instance.



Some Comments from Experts

- **Data science is not big data** (Pat Virtue, CMU)
  - Sometimes, in order to truly understand and answer your question, you need massive amounts of data…
  - …But sometimes you don't
  - **Don't create more work for yourself than you need to !**

Data Science Applications and Examples

- (ref) https://builtin.com/data-science/data-science-applications-examples
- **Healthcare**
  - Google: machine learning for metastasis (identifying breast cancer)
  - CLUE, Germany: predict periods and forecast conditions for pregnancy
  - Oncora Medical: cancer care recommendations
- **Road Travel**
  - UPS: optimizing package routing (save up to $200 million)
  - Streetlight data: traffic patterns for cars, bikes, and pedestrians (use for commuter transit design)
  - Uber Eats (Uber's delivery app): optimize full delivery process
- **Sports**
  - Liverpool F.C.: recruited undervalued soccer players
  - RSPCT: basketball-coaching sensor (shooting analysis system)
  - British Olympic Rowing team: model athlete evolution and find a promising newbie rower

Data Science Applications and Examples

- **Government**
  - Equivant: data-driven crime prediction
  - ICE (Immigrations and Customs Enforcement): facial recognition in ID databases
  - IRS: tax-fraud detection
- **E-commerce**
  - SOVRN: automated AD placement (target campaigns to custmoers)
  - Instagram: convert users' likes and comments, their usage of other apps and their web history into predictions about the products they might buy
  - Airbnb: search that highlights areas of cool neighborhoods (high density of bookings)
- **Social life**
  - Tinder (most popular dating app): find a good match for singles
  - Facebook: "people you may know" sidebar (based on friend list, photos, schools, etc.)

www.kaggle.com

- **What is Kaggle?**
  - Owned by Google, and over 3 million data scientist registered.
  - The world's largest data science and machine learning community with powerful tools and resource. (over 50,000 public datasets and 400,000 public notebooks)
  - You can find and publish data sets, and all data sets are **free**.
  - Can participate competitions to solve data science challenges.
  - Provide self-learning courses (from Python to Deep Learning)
  - Explore and run machine learning code with Kaggle Notebooks (with source codes).
  - Can discuss any data science issues with experts.
  - Try it at https://www.kaggle.com

Data in Korea

- 데이터 이용을 활성화하기 위한 데이터 3법 통과 (2020.1)
  - 개인정보 보호법, 정보통신망법, 신용정보법
  - 핵심: 가명정보를 통계작성, 연구, 공익적 기록 보존 용도로 본인 동의없이 활용 가능
- 가명정보 (pseudonym or alias) (예: https://brunch.co.kr/@jaeyunchoi/18)

	개념	예시	활용가능 범위
개인정보	특정 개인에 관한 정보, 개인을 알아볼 수 있게 하는 정보	강하늘, 1990년 2월 21일생, 남성, 2019년 12월 신용카드 사용금액 150만 원	사전적, 구체적 동의를 받은 범위 내에서만 활용 가능
가명정보	추가정보의 사용 없이는 특정 개인을 알아볼 수 없게 조치한 정보	강XX, 1990년생, 남성, 2019년 12월 신용카드 사용금액 150만 원	개인정보 범위에 포함되나, 다음 목적에 한하여 동의없이 활용 가능 ① 통계작성(상업적 목적 포함) ② 연구(상업적 연구 포함) ③ 공익적 기록보존 목적 등
익명정보	더 이상 개인을 알아볼 수 없게 (복원 불가능할 정도로) 조치한 정보	남성, 20대, 2019년 12월 신용카드 사용금액 100만 원 이상	개인정보가 아니므로 제한없이 자유롭게 활용 가능

Different Views for Data

- **Simpson’s Paradox (심슨의 역설)**
  - 각 그룹 데이터에서 개별적으로 나타나는 특징과 전체의 경향이 달라지는 현상 (같은 데이터가 분석 방법에 따라 해석이 달라질 수 있음) -> 데이터의 분석에 주의 (데이터의 특성과 구조적 차이를 잘 고려해야 함)
  - (예)

도시	A사	B사
서울	정상품 90, 불량품 10 (불량률 10%)	정상품 920, 불량품 80 (불량률 8%)
부산	정상품 980, 불량품 20 (불량률 2%)	정상품 99, 불량품 1 (불량률 1%)
전체	A사 총 불량률 30/1,100 = 3%	B사 총 불량률 81/1,100 = 8%

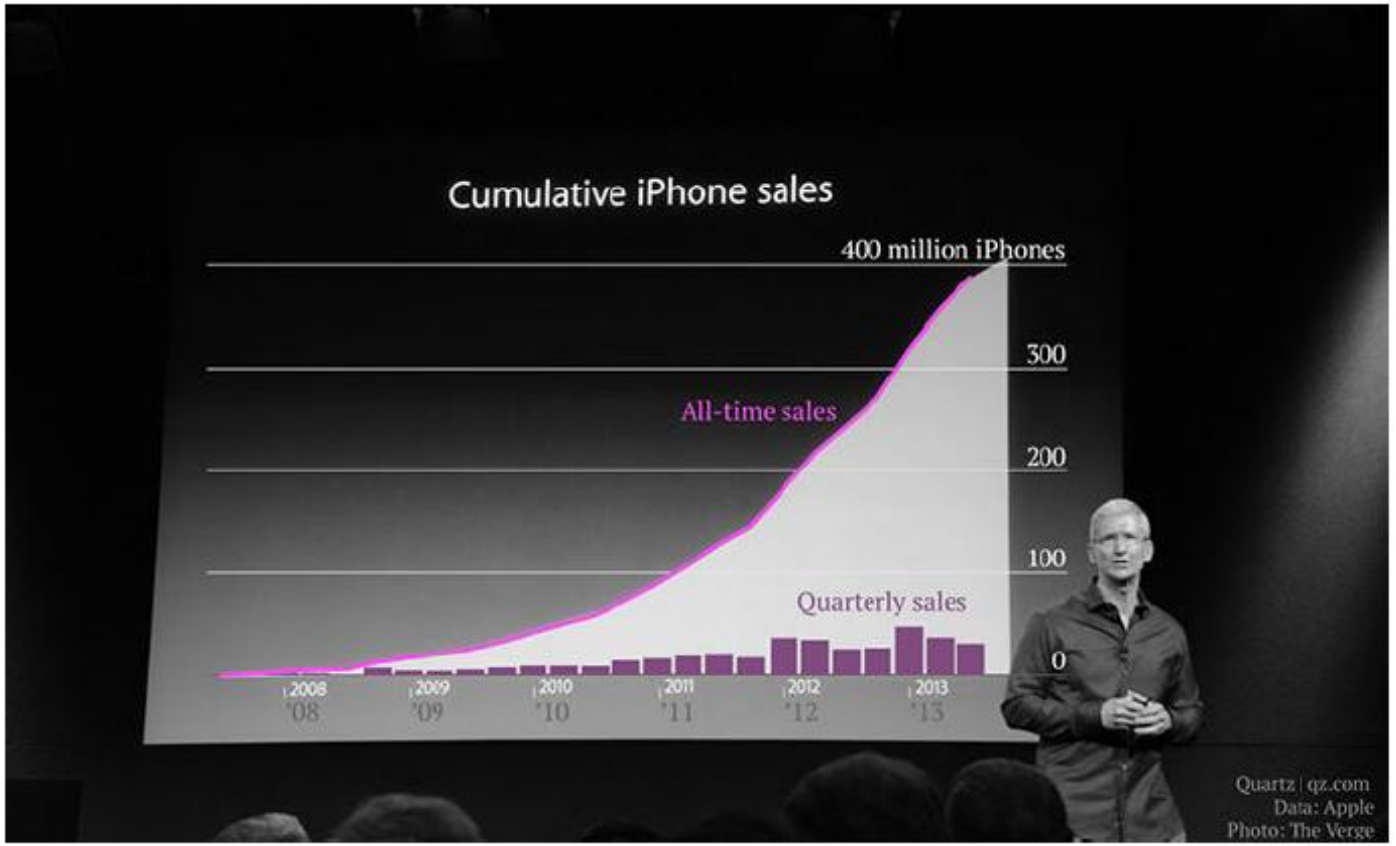
$$\left(\frac{a1}{A1} > \frac{b1}{B1}\right) \;\&\; \left(\frac{a2}{A2} > \frac{b2}{B2}\right) \overset{?}{\rightarrow\rightarrow} \frac{a1+a2}{A1+A2} > \frac{b1+b2}{B1+B2}$$

Cumulative data vs. Incremental Data

- Apple iPhone sales have been exploding, right?

Cumulative data vs. Incremental Data

- Cumulative distributions present a misleading view of growth rate



- 누적 데이터는 모든 이전 값이 포함되어 있기 때문에 최근의 변화나 경향을 숨길 수 있다. 예를 들어, 초기에 좋은 성과를 냈던 데이터가 이후 성과의 하락을 감출 수 있다.
- 증분 데이터는 각 기간별 성과를 보여주므로 최근의 변동성을 더 잘 보여줄 수 있다. 하지만 짧은 기간의 증분 데이터만 본다면 전체적인 장기 추세를 놓치기 쉽다.
- 즉, 데이터 분석 시 맥락을 잘못 파악하면 오해가 발생할 수 있다. 따라서, 데이터의 특성과 구조적 차이를 잘 고려해야 한다.

Public datasets for data scientist

- **Dataset finders**
  - Google dataset search: contains over 25 million datasets (https://toolbox.google.com/datasetsearch)
  - Kaggle (www.kaggle.com)
  - UCI machine learning repository (https://archive.ics.uci.edu/ml)
  - VisualData (https://www.visualdata.io/discovery) : computer vision dataset
  - CMU Libraries (https://guides.library.cmu.edu/machine-learning/datasets)
  - NLP Database by Quantum Stat. (https://datasets.quantumstat.com): natural language processing, sentiment analysis dataset
  - US Government and official dataset (https://www.data.gov)
  - Korea government's public dataset (https://www.data.go.kr)
  - Eurostat: open data from the EU statistical office

Public datasets for data scientist

- **Machine learning dataset**
  - Mall customers dataset : people visiting the mall (gender, id, age, income, spending score, etc.)
  - IRIS dataset: simple and beginner-friendly (flower petal and sepal width)
  - MNIST dataset: 60,000 training images and 10,000 testing images
  - Cifar-10: 60,000 32x32 color images with 10 classes
  - Boston housing dataset: collected by US Census Service
  - Fake news detection: 7,796 rows with 4 columns (news, title, news text, result)
  - Wine quality dataset: different chemical information about the wine
  - Titanic dataset: 891 training and 418 test passengers (name, age, sex, no of siblings abroad, etc.)
  - Credit card fraud detection dataset: recognize credit card transactions

Public datasets for data scientist

- **Computer vision dataset**
  - xView: one of the most massive publicly available image dataset (images from complex scenes annotated using bounding boxes) (xviewdataset.org)
  - ImageNet: largest image dataset for computer vision, used in ILSVRC(ImageNet Large Scale Visual Recognition Challenge), more than 1.2 million images with 1,000 classes (hierarchically organized)
  - Kinetics-700: (https://deepmind.com/research/open-source/kinetics) a collection of large-scale, high-quality datasets of URL links of up to 650,000 video clips (human -object interactions)
  - Google open images dataset: more practical than ImageNet, ~9 million images annotated with labels over 6,000 categories
  - Cityscapes dataset: video sequences taken from 50 city streets
  - IMDB-Wiki dataset: dataset for face images with labeled gender and age
  - Color detection dataset
  - Stanford Dogs dataset

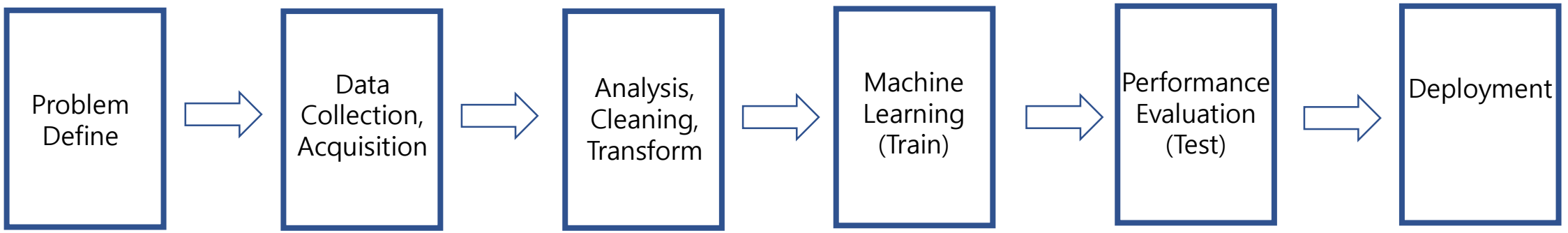
Public datasets for data scientist

- **Self-driving (autonomous driving) datasets**
  - Waymo open dataset (https://waymo.com/open/)
  - Berkeley DeepDrive BDD100k: for self-driving over 2,000 hours in NY and California
  - Bosch small traffic light dataset: traffic light for deep learning
  - WPI datasets: traffics lights, pedestrians, and lane detection
  - LISA: traffics signs, vehicle detection, and trajectory patterns
  - Comma.ai: car's speed, acceleration, steering angle, and GPS coordinates
  - Cityscape datasets: street scenes
- **Clinical datasets**
  - MaskedFace-Net: masked faces
  - COVID-19 datasets: from over 45,000 scholarly articles about COVID-19
  - MIMIC-III: from MIOT Lab for computational physiology from ~40,000 critical care patients

Public datasets for data scientist

- **Sentiment analysis datasets**
  - Lexicoder Sentiment Dictionary: over 3,000 negative words and over 2,000 positive sentiment words
  - IMDB reviews: over 50,000 movie reviews from Kaggle
  - Stanford sentiment Treebank
  - Twitter US airline sentiment: twitter data on US airlines, classified as positive, negative, and neutral
- **Natural Language Processing (NLP) datasets**
  - The Big Bad NLP database: various natural language processing tasks
  - HotspotQA datasets: for explainable question answering systems
  - Amazon reviews
  - Rotten Tomatoes Reviews: 480,000 critic reviews (fresh or rotten)
  - SMS spam collection in English: 5,574 SMS spam messages
  - UCI spambase dataset: 4,601 emails with 57 meta-information about emails

Data Science Work Flow



- Domain knowledge
- Business strategy

- CSV/Excel
- JSON
- HTML/XML
- SNS
- String(structured)
- Text(unstructured)
- Image, Voice
- Language
- Multi-modal

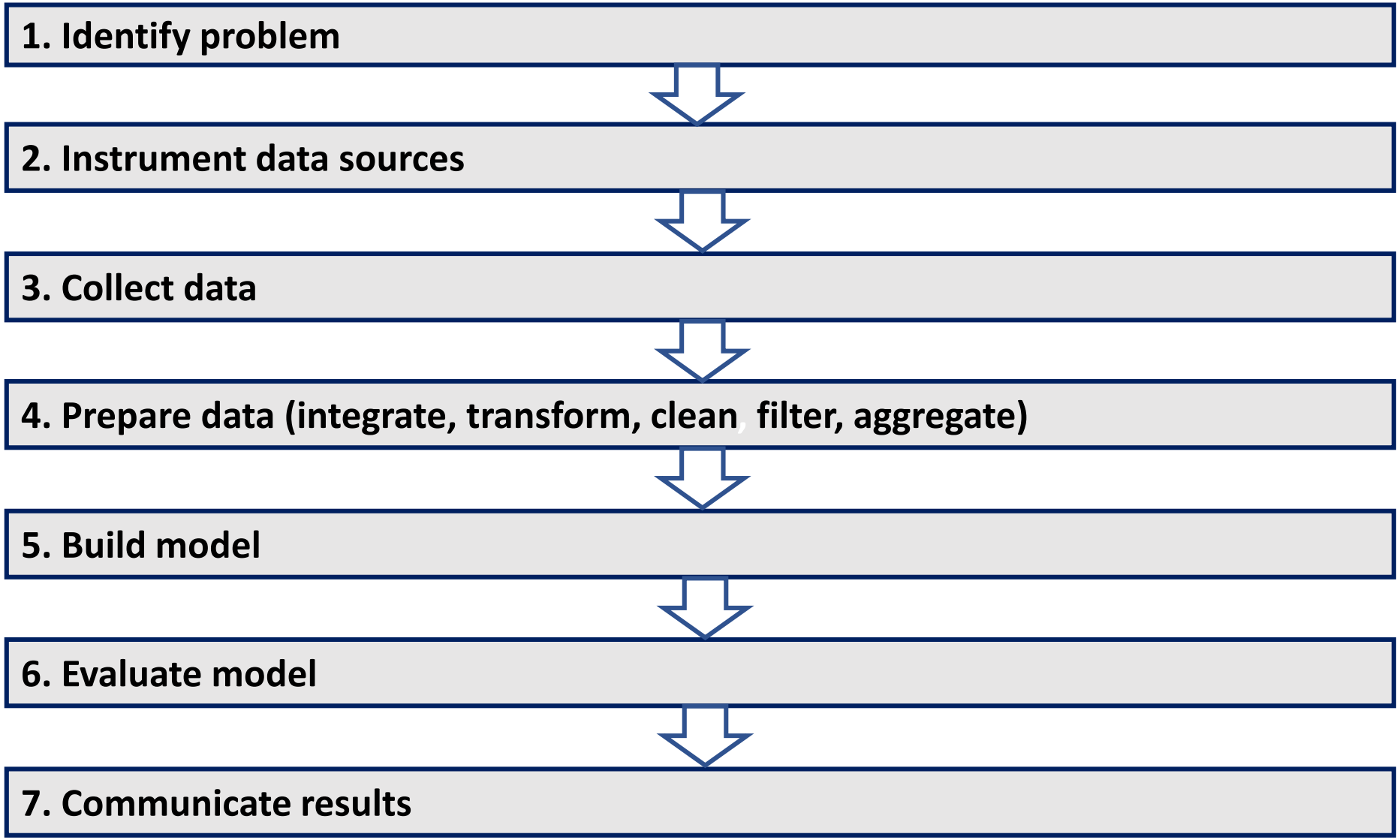
- Visualization
- Missing values
- Invalid values
- Outliers
- Categorical encoding
- Scaling
- Transform
- Feature engineering

- Supervised
- Unsupervised
- Loss (or Error)
- Bias and Variance
- Overfitting
- Regularization
- MLP/CNN/RNN
- Generative model
- Reinforcement learning
- Transformer

- R-square
- Accuracy
- Precision/recall
- F-1 score
- ROC/AUC
- mAP
- IoU

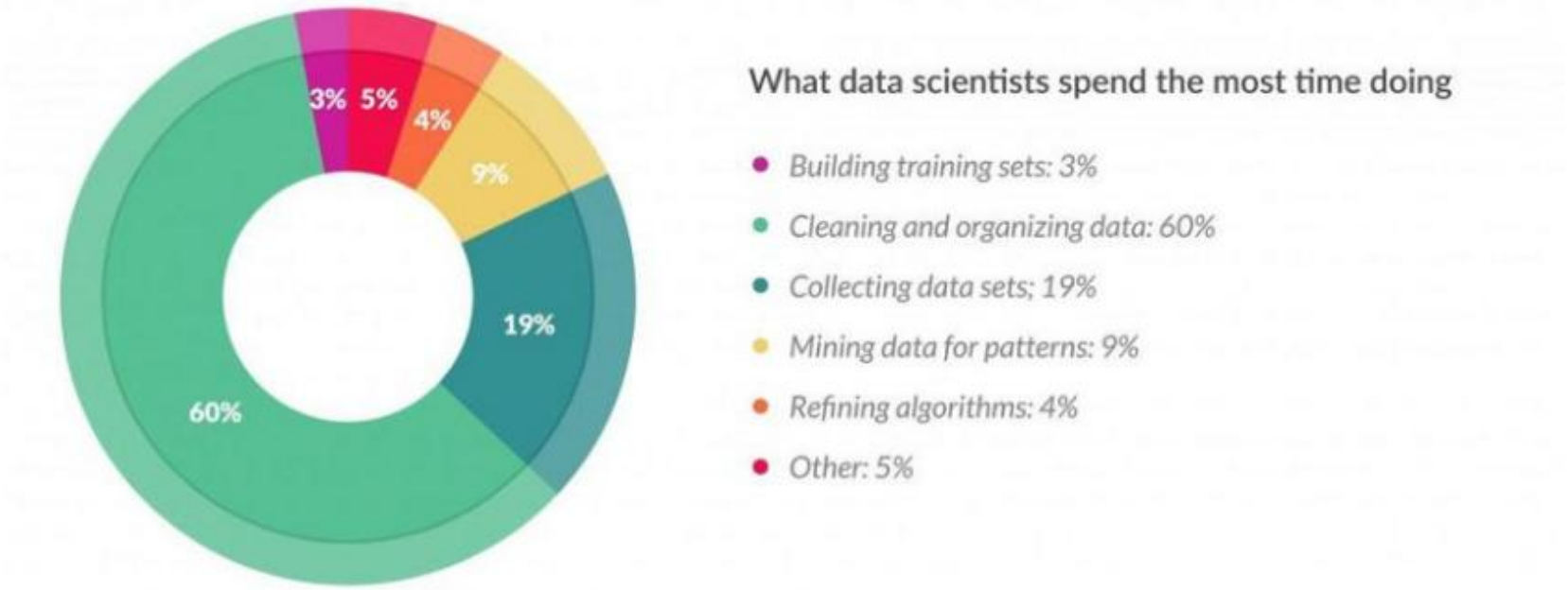
- Server
- Mobile

Jeff Hammerbacher's Model



Data Science Work Flow

- **According to a survey in Forbes,**
  - Data scientist spend 80+ % of their time on **data preparation**.



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

AI의 진화: 더 똑똑하게, 더 가깝게

- "2025년 현재, AI는 인간처럼 보고 듣고, 스스로 행동하며, 우리 삶 속으로 더 가까이 다가오고 있다."
- **멀티모달 AI (Multimodal AI):**
  - 텍스트(언어)만 이해하던 과거를 넘어, 이미지(시각)와 사운드(청각)까지 동시에 이해하고 소통하는 AI 등장.
  - 예시: 이미지를 보여주며 " 이 사진 분위기에 맞는 음악 추천해줘 " 와 같은 복합적인 요청 처리.
- **AI 에이전트 (AI Agent):**
  - 단순한 답변을 넘어, 목표를 주면 스스로 계획을 세우고 여러 도구를 사용해 과업을 완수하는 자율형 AI 비서.
  - 예시: " 이번 주말 부산 여행 계획 짜고, 가장 저렴한 KTX 예약해줘 " 와 같은 복잡한 임무를 스스로 수행.
- **온디바이스 AI (On-Device AI):**
  - 거대한 서버(클라우드)가 아닌, 스마트폰, 노트북, 자동차 등 개인 기기 위에서 직접 동작하는 AI.
  - 장점: 빠른 응답 속도, 인터넷 연결 불필요, 강력한 개인정보보호.

AI를 움직이는 힘 (1): 하드웨어 혁신

- AI의 놀라운 진화는 눈에 보이지 않는 반도체 기술의 혁명이 있기에 가능하다.
- **두뇌 세포의 진화 (트랜지스터):**
  - AI 칩의 성능을 결정하는 기본 소자인 트랜지스터가 더 작고 효율적인 3차원 구조(FinFET → GAA)로 발전하며 AI의 연산 능력을 극대화. (채널과 게이트가 만나는 유효 면적을 넓혀서 전류에 대한 제어력을 높이고자 함, 누설전류 감소 노력)
- **AI 전용 두뇌의 등장 (특화 반도체):**
  - GPU (그래픽 처리장치): 수천 개의 코어로 데이터를 동시에 처리하는 '병렬 처리' 방식으로 AI 학습 속도를 혁신.
  - NPU (신경망 처리장치): AI 연산만을 위해 만들어진 초저전력, 고효율의 AI 전용 칩.
- **데이터 공급 고속도로 (차세대 메모리):**
  - HBM (고대역폭 메모리): AI 반도체에 데이터를 막힘없이 공급하기 위해 메모리 칩을 수직으로 쌓아 올린 데이터 전용 고속도로.

AI를 움직이는 힘 (2): 알고리즘 혁신

- 최고의 하드웨어도 그것을 100% 활용하는 소프트웨어 기술이 없다면 무용지물에 불과하다.
- **거대 모델을 위한 협력 기술:**
  - 분산 훈련 (Distributed Training): 혼자서는 감당 못 할 초거대 AI 모델을 수천 개의 AI 반도체에 나누어 함께 학습시키는 기술.
  - ChatGPT와 같은 거대 언어 모델(LLM) 탄생의 일등 공신.
- **일상 속 AI를 위한 다이어트 기술:**
  - 모델 경량화 (Model Lightweighting): 거대 모델의 성능은 최대한 유지하면서 용량을 획기적으로 줄여, 스마트폰에 탑재(온디바이스 AI)할 수 있도록 만드는 AI 다이어트 기술.
  - 주요 기술: 양자화(Quantization), 가지치기(Pruning), 지식 증류(Knowledge Distillation) 등

    - Prunning: 잘 훈련된 대규모 AI 모델에서 성능에 거의 영향을 주지 않는 불필요한 부분(연결, 뉴런 등)을 식별하고 제거하는 기술. 거대한 신경망 모델 안에는 예측에 거의 기여하지 않는 중복되거나 중요도가 낮은 연결(가중치가 0에 가까운)이 많이 존재한다. 이러한 부분을 제거하면 모델의 크기(저장 공간)와 계산량(추론 속도)을 줄일 수 있다.
    - Knowledge Distillation: 크고 성능이 좋은 교사 모델(Teacher Model의 지식을, 작고 가벼운 학생 모델(Student Model)에게 전달하여 훈련시키는 방식으로 교사 모델의 성능 손실 (거의) 없이 작은 모델로 이전한다. (전이 + 모델 압축)

Data Science Tools

- **Python**
  - 데이터 과학 프로젝트의 전 과정을 포괄하는 사용이 쉬운 툴
  - 풍부한 라이브러리 생태계 (NumPy, Pandas, Scikit-learn...)
  - 강력한 커뮤니티와 높은 확장성
- **Platform**
  - Cloud 환경: Google Colab, Kaggle
  - Local 환경: VS Code, PyCharm, JupyterLab/Notebook