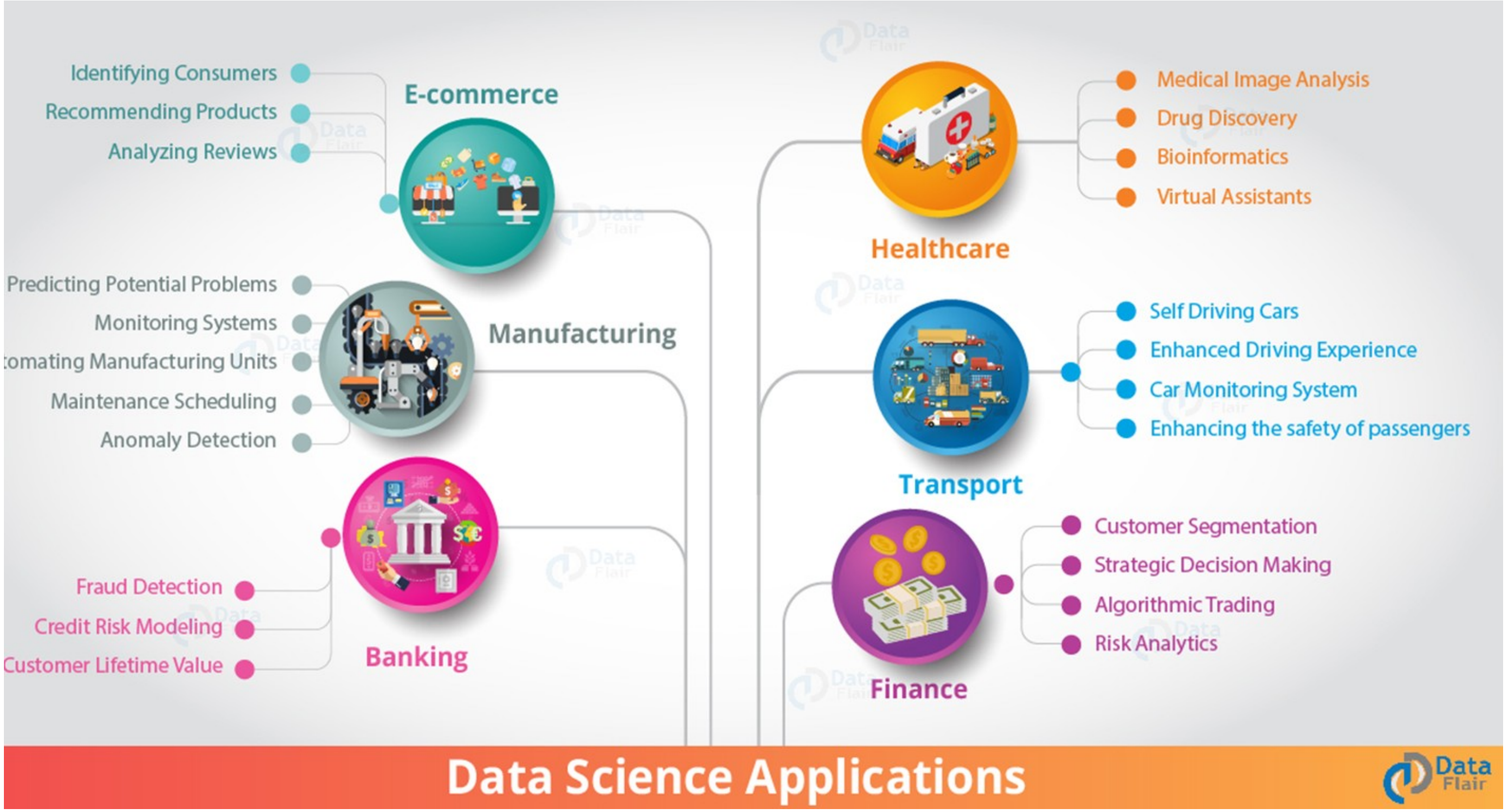
What is data science?

- Data is everywhere! Data science starts with data!



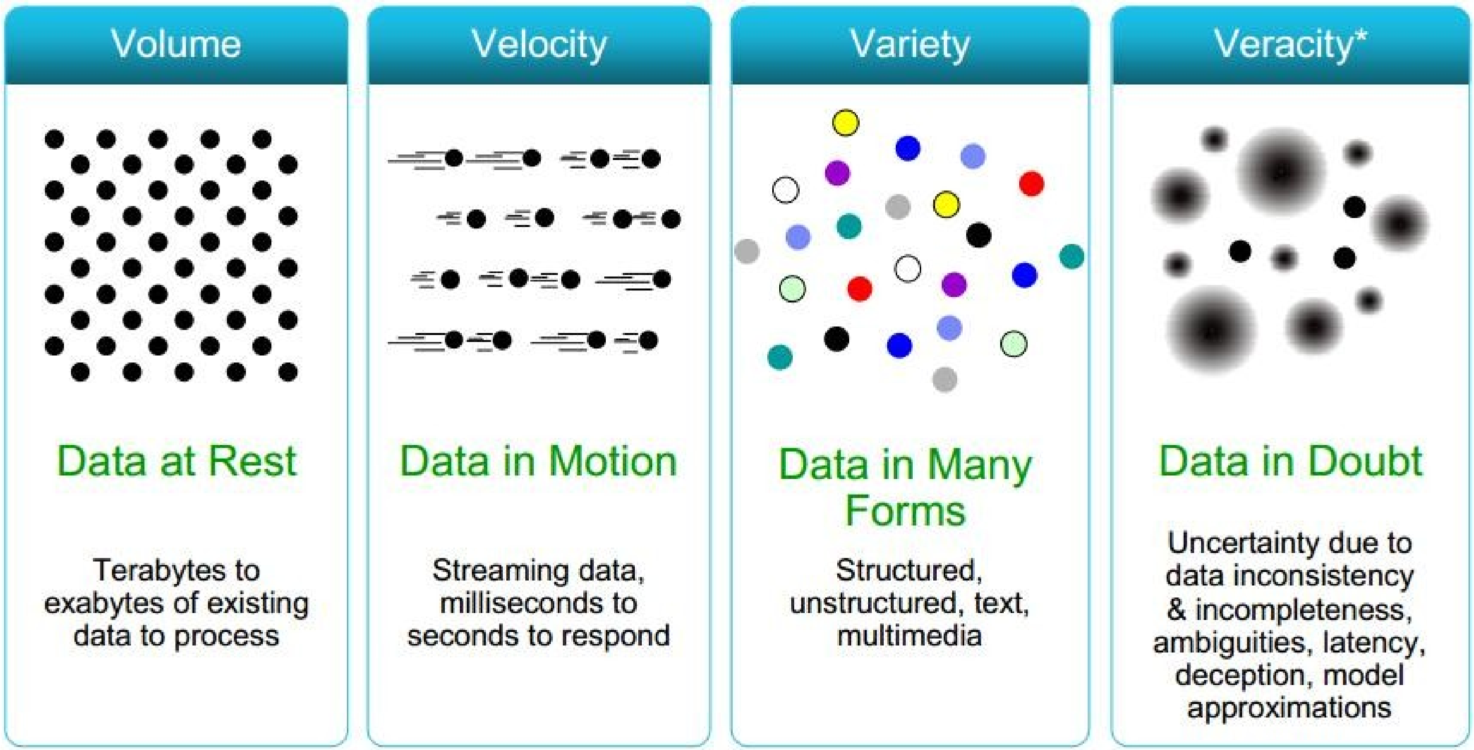
What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Big Data Era

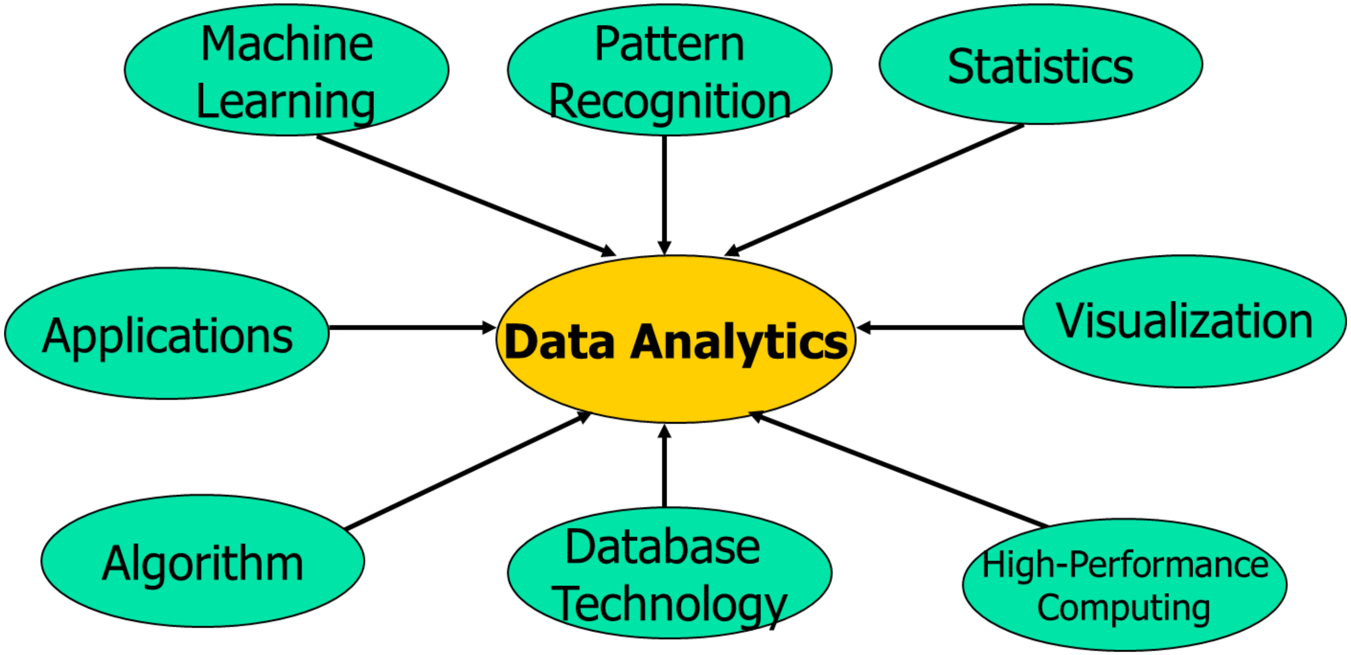
- Google: every 2 days we create as much data as we did up to 2003.
- Facebook: 500+ TB of new data every day including
  - 2.5 billion items shared
  - 2.7 billon Likes
  - 300 million photos
  - 100+ PB Hadoop cluster
- Twitter: 500 million tweets per day
- Many applications for streaming data, e.g., sensors

4V


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

What is data science?

- Data science combines the fields of computer science, mathematics, statistics, and information systems with a focus on the generation, organization, modeling, and use of data to make scientific and business decisions.



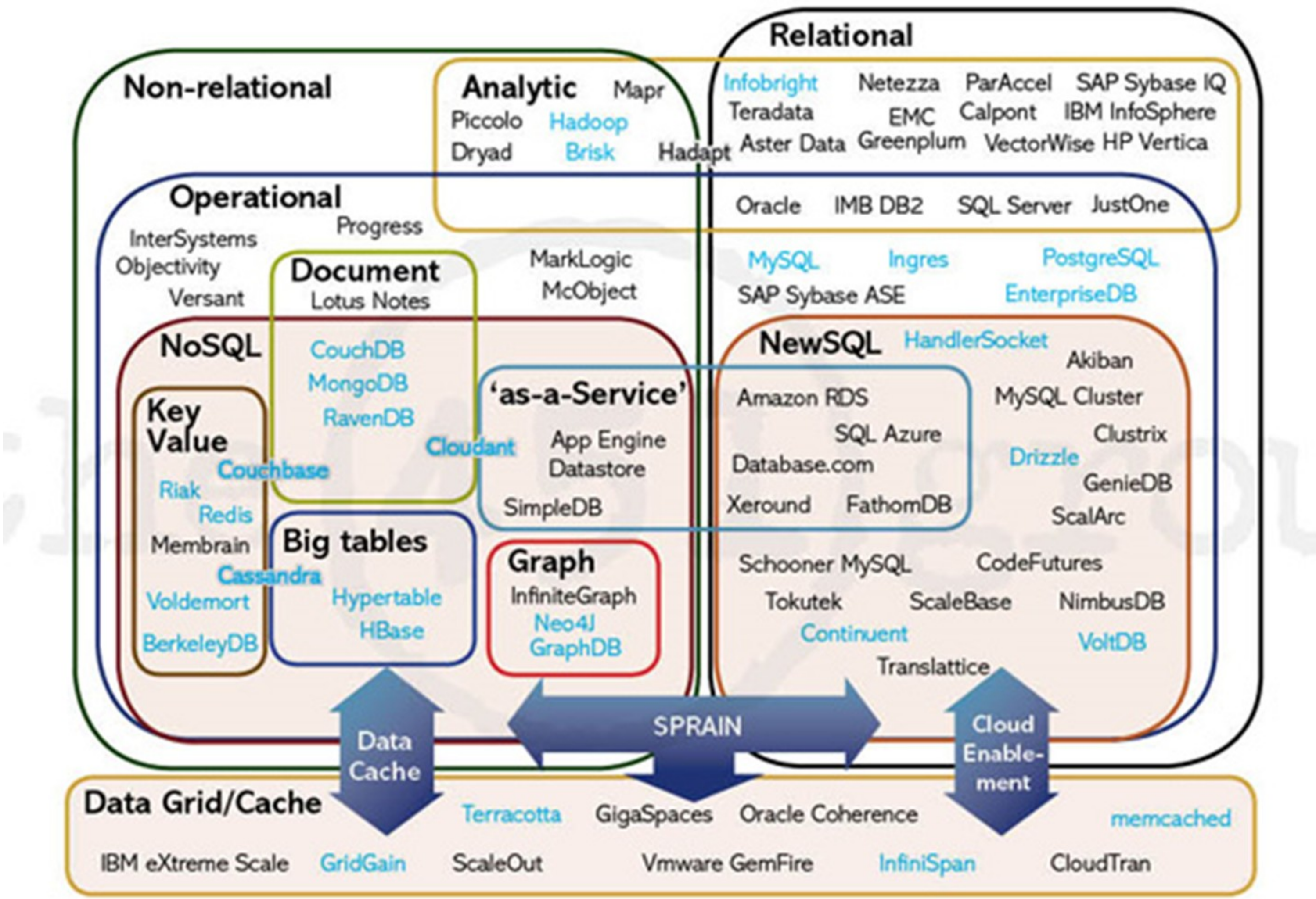
Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

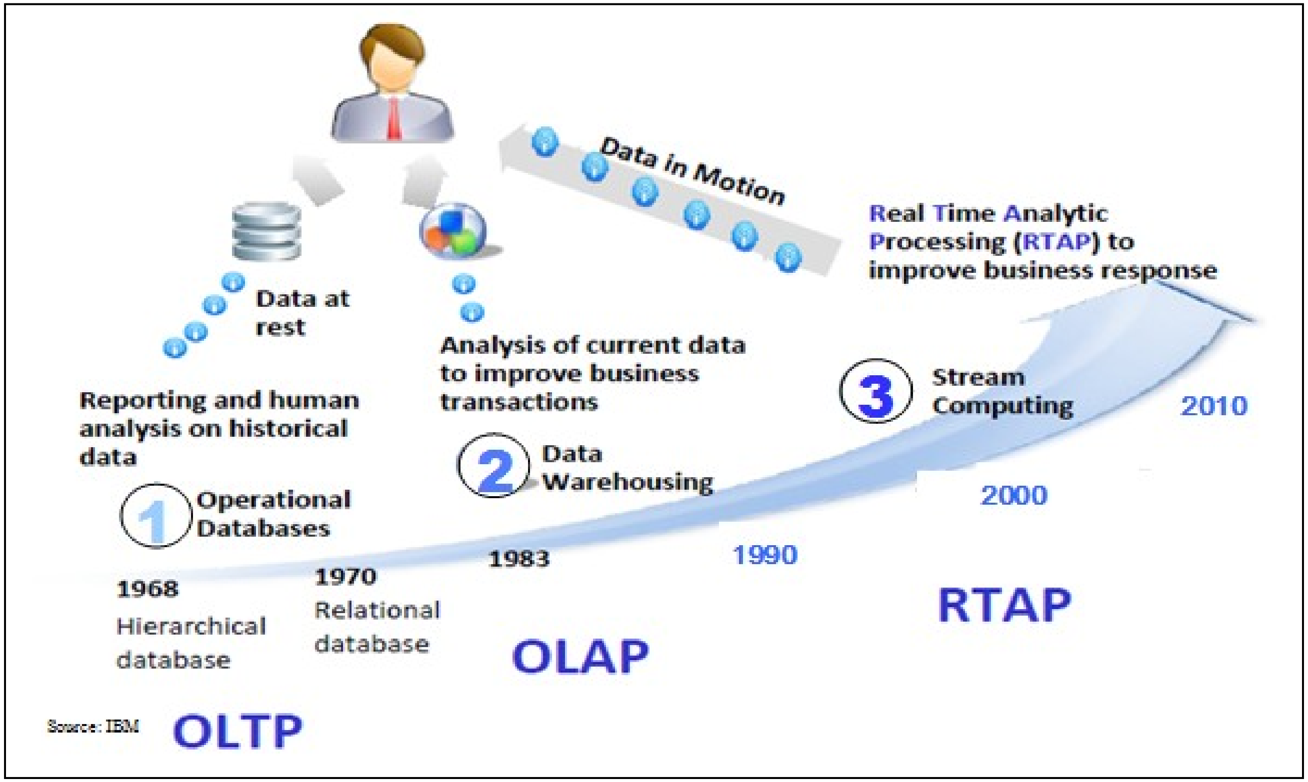
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
- Latest
  - NoSQL, NewSQL, Column DB
  - MapReduce, Spark

Database Categorization



Evolution of Business Intelligence



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Big Data Landscape

Vertical Apps
PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps
splunk>
loggly
sumologic

Ad/Media Apps
rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence
ORACLE | Hyperion
SAP BusinessObjects
RJMetrics
Microsoft | Business Intelligence
IBM COGNOS
birst
Autonomy
MicroStrategy
QlikView
bime
DOMO
Chart.io
GoodData

Analytics and Visualization
tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
sas
TIBCO
KARMASPHERE
panopticon
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service
kaggle
factual.
knoema
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
SPACE CURVE
LOQATE

Analytics Infrastructure
Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure
COUCHBASE
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service
amazon web services
Windows Azure
infochimps
Google BigQuery

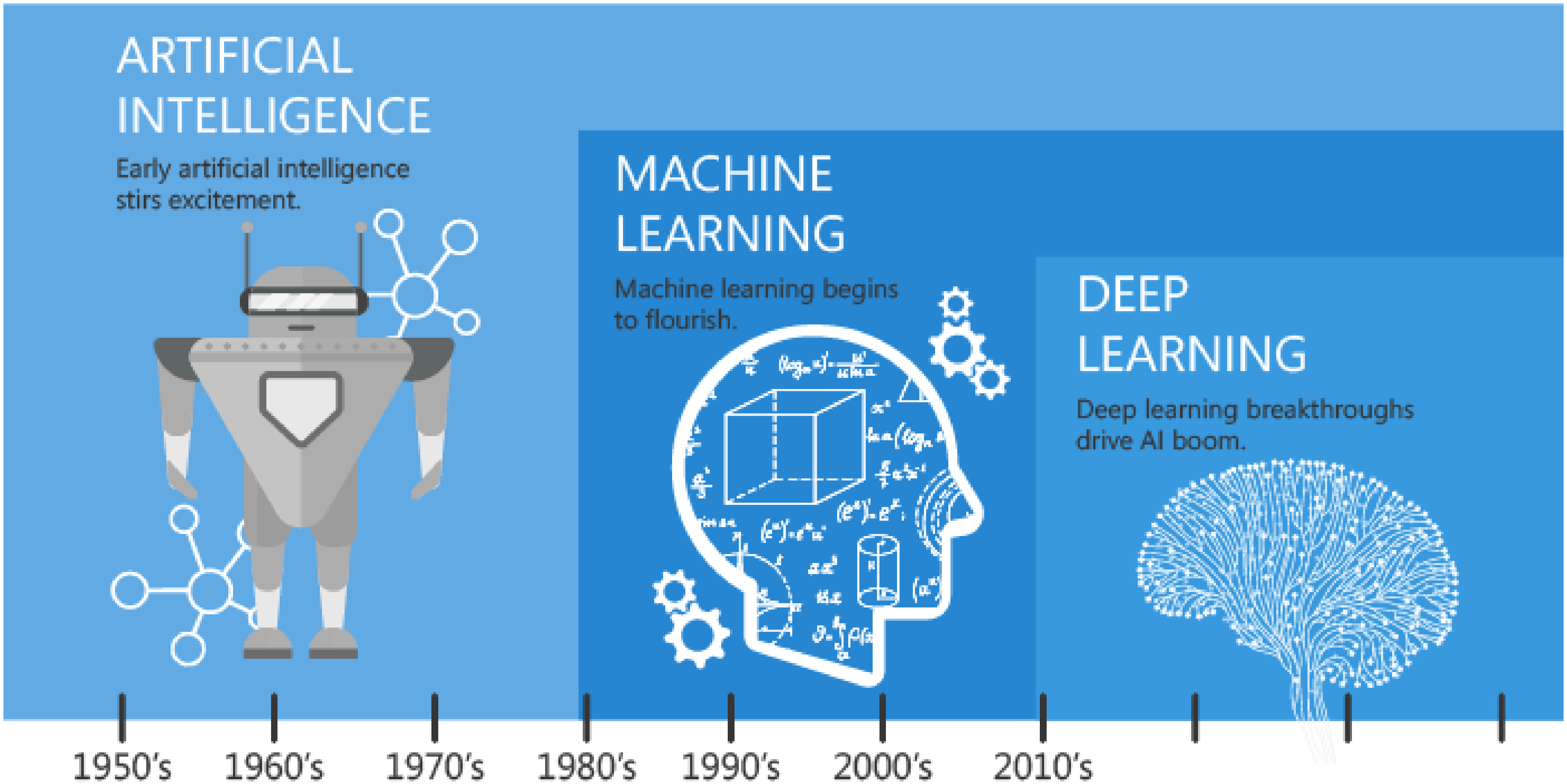
Structured Databases
ORACLE
MySQL
SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies
hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

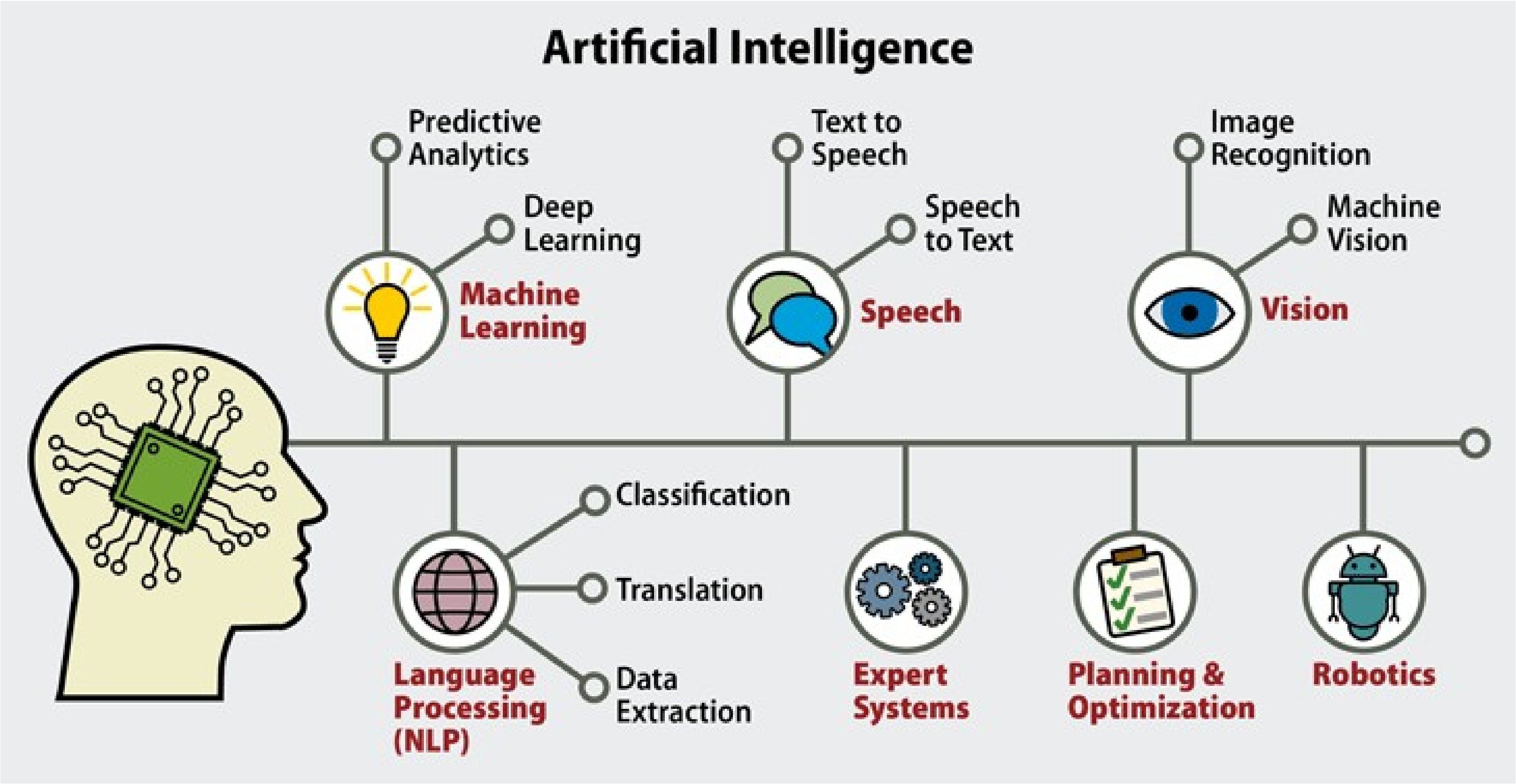
dave@vcdave.com

blogs.forbes.com/davefeinleib

Evolution of AI

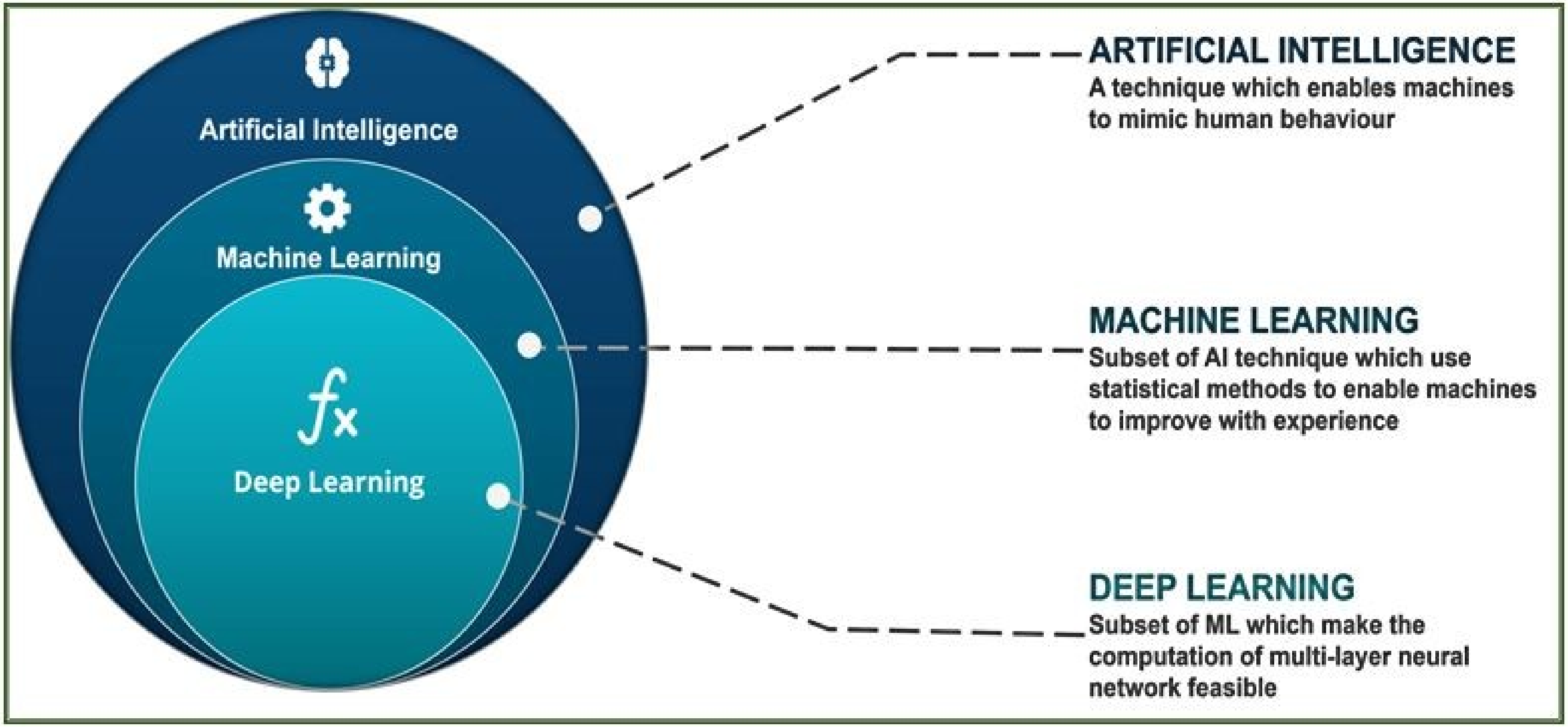


AI



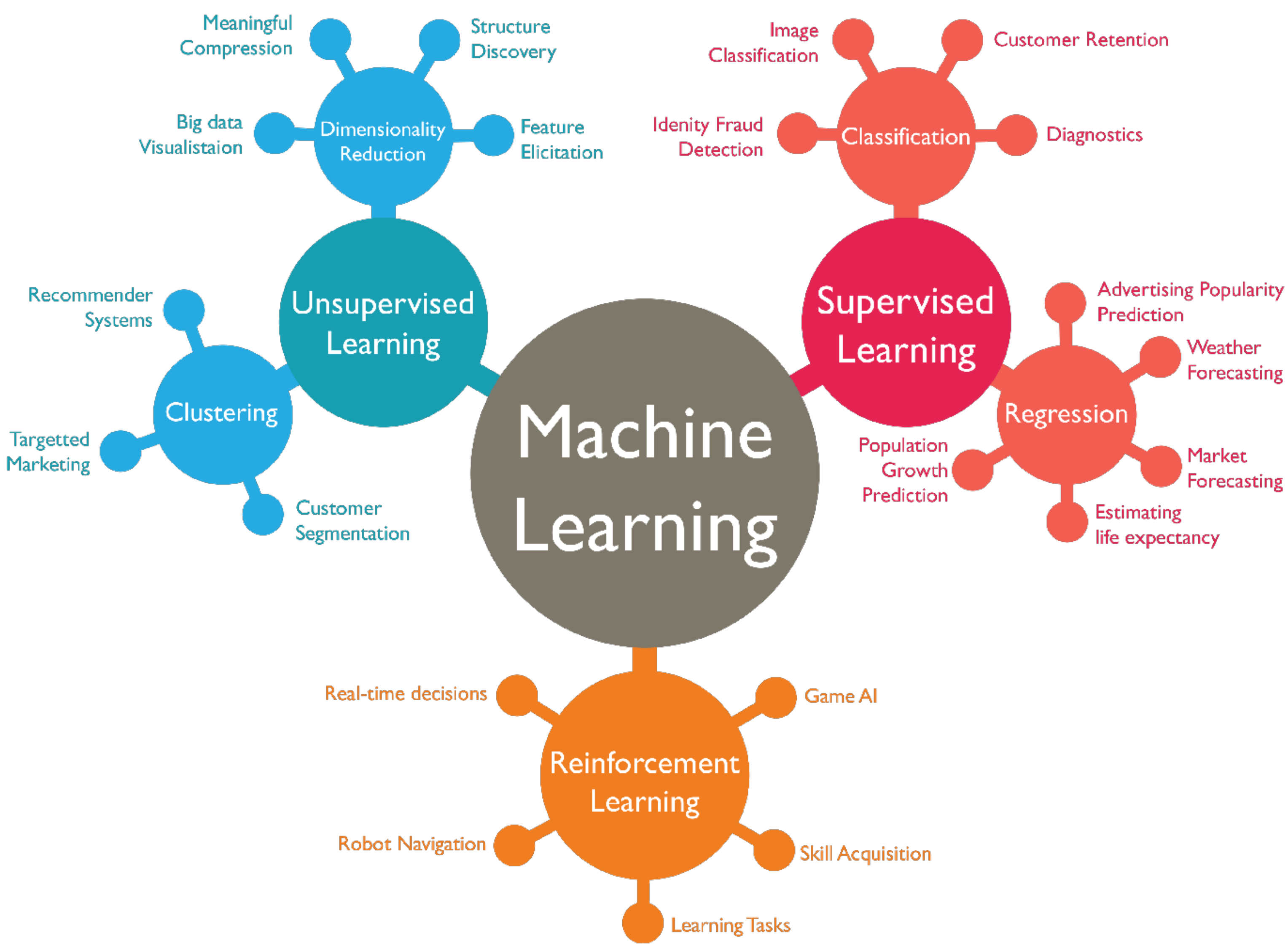
Artificial Intelligence
Predictive Analytics
Deep Learning
Machine Learning
Text to Speech
Speech to Text
Speech
Image Recognition
Machine Vision
Vision
Classification
Translation
Data Extraction
Language Processing (NLP)
Expert Systems
Planning & Optimization
Robotics

AI vs. ML vs. DL

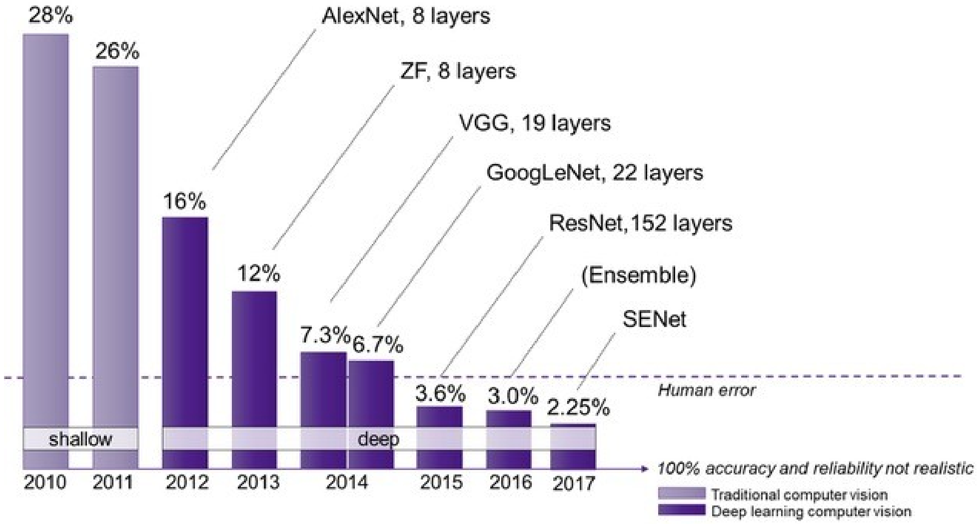


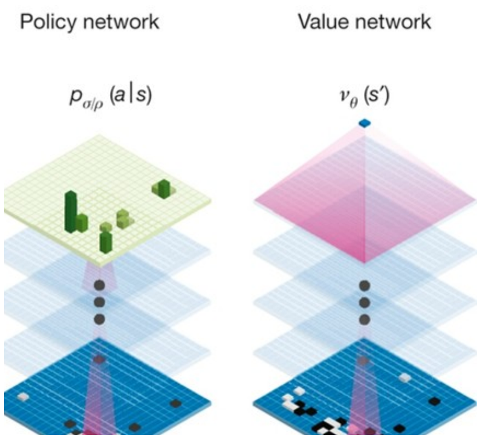
Artificial Intelligence
Machine Learning
Deep Learning
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

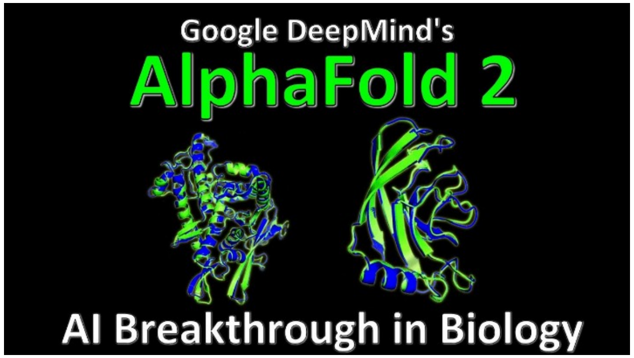
Machine Learning



Deep Learning

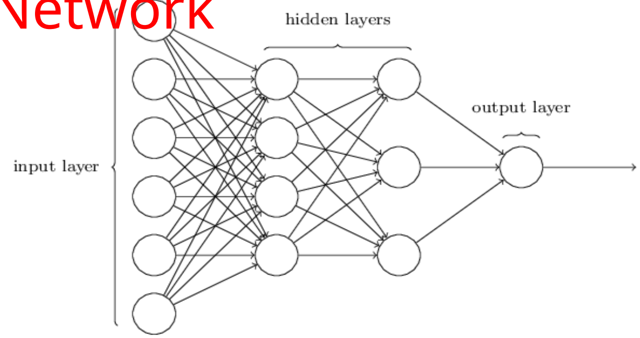




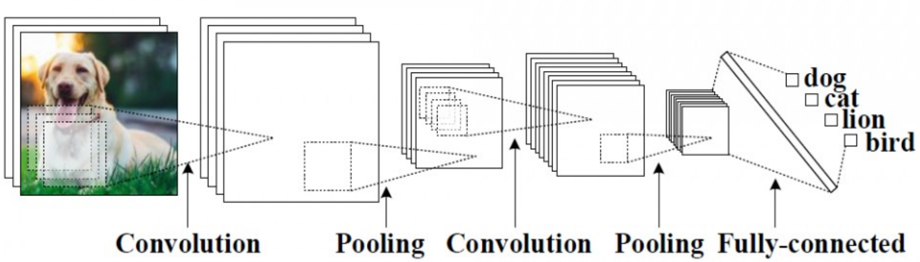


Basic Deep Learning Structures

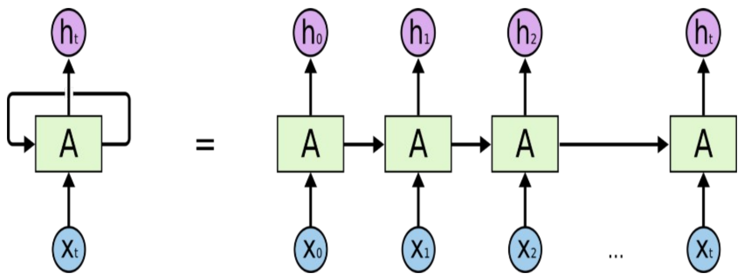
Feedforward Neural Network



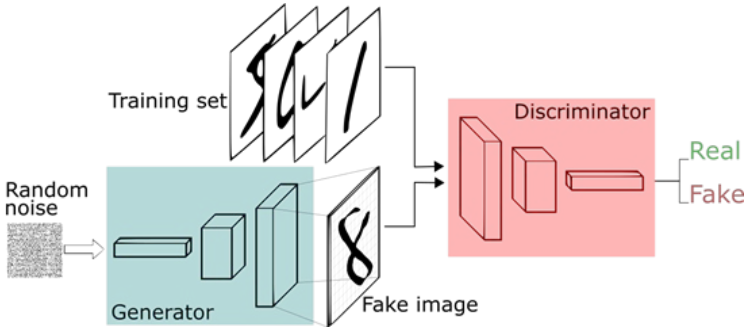
Convolutional Neural Network



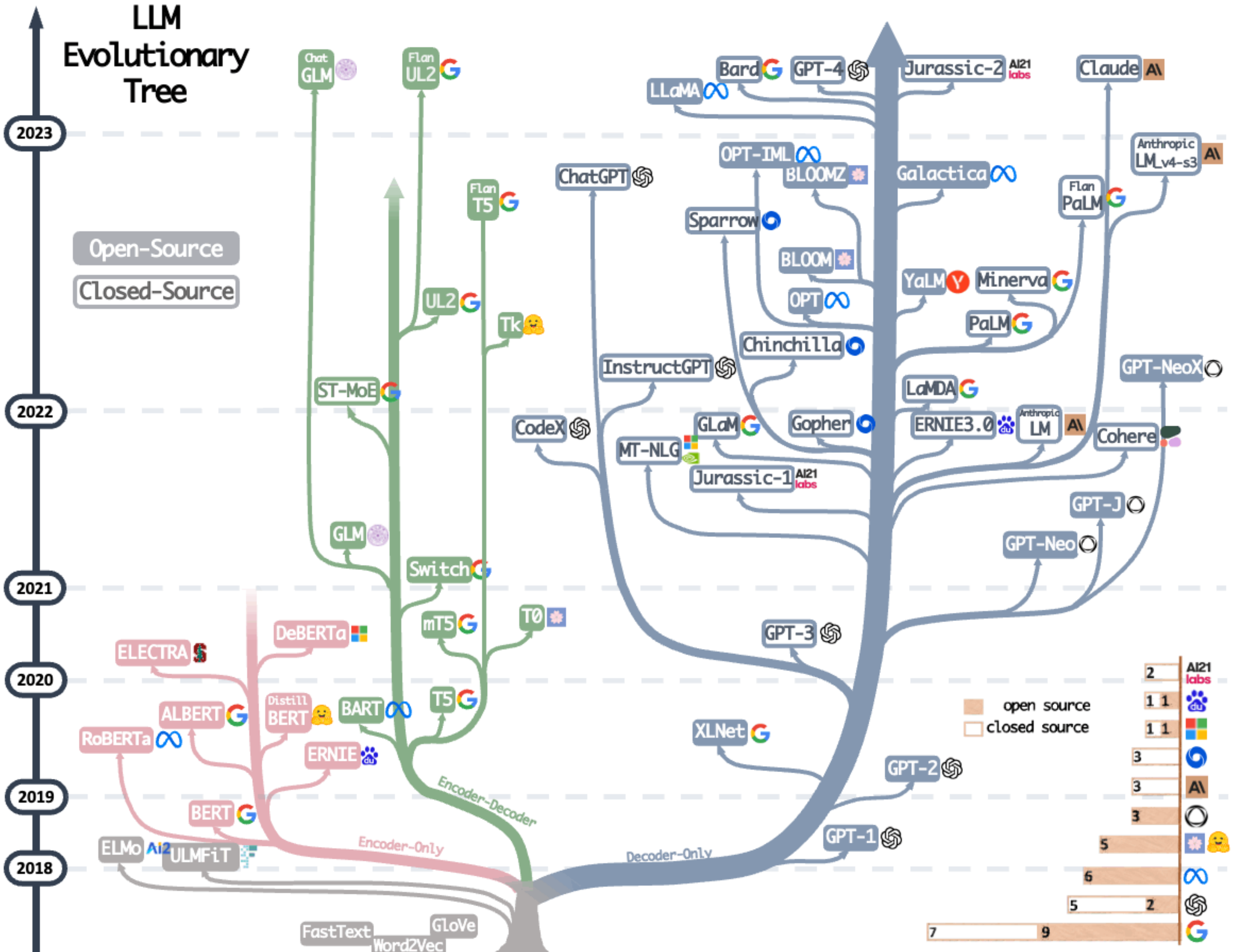
Recurrent Neural Network



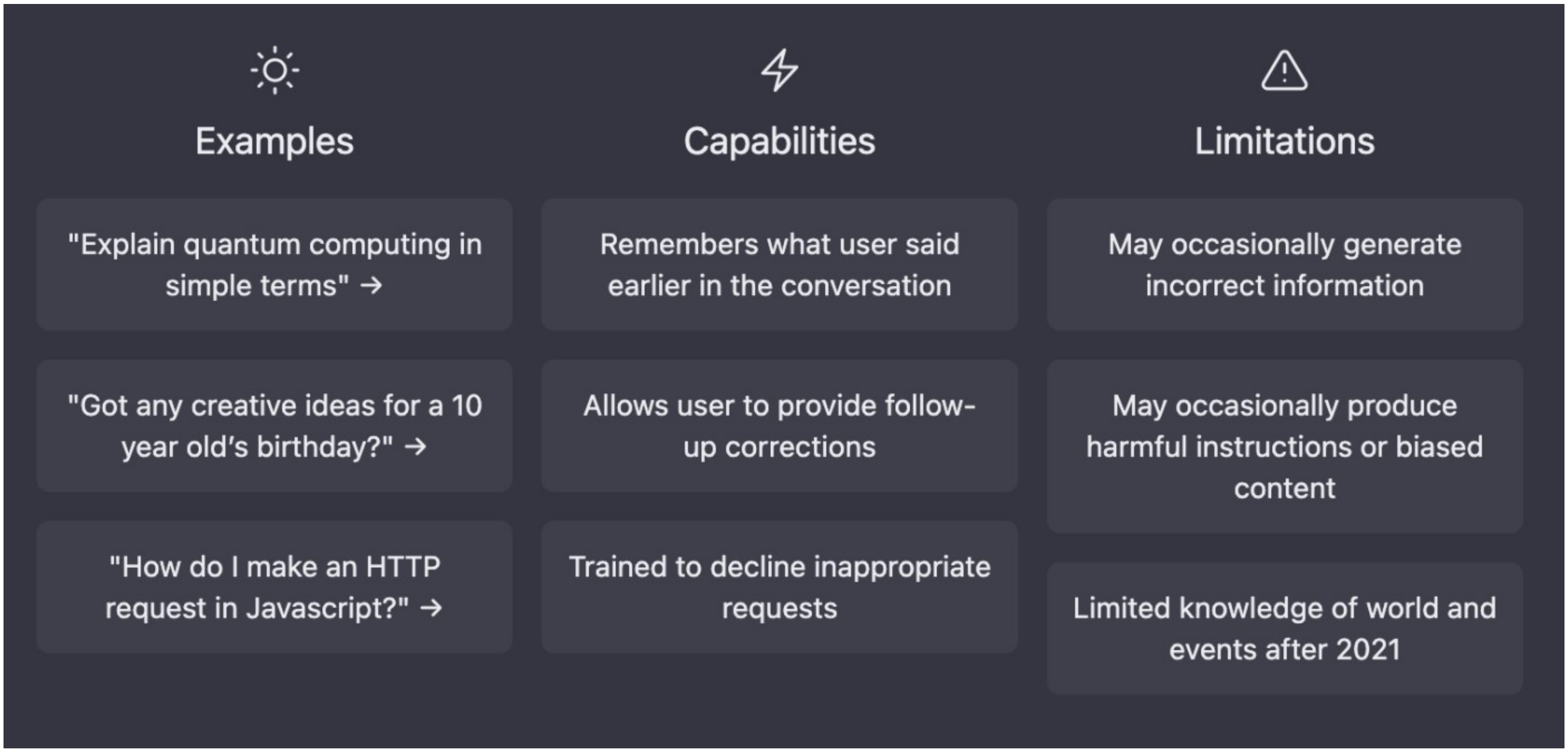
Generative Adversarial Network



Large Language Models

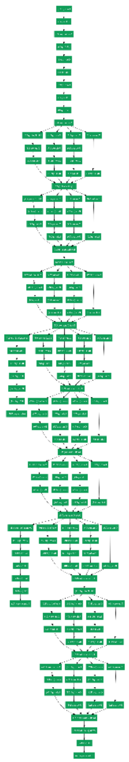


ChatGPT

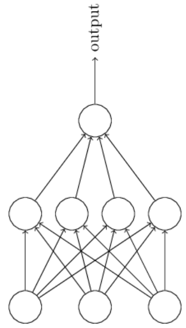

Examples
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
Limitations
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021

Why does DL work so well?

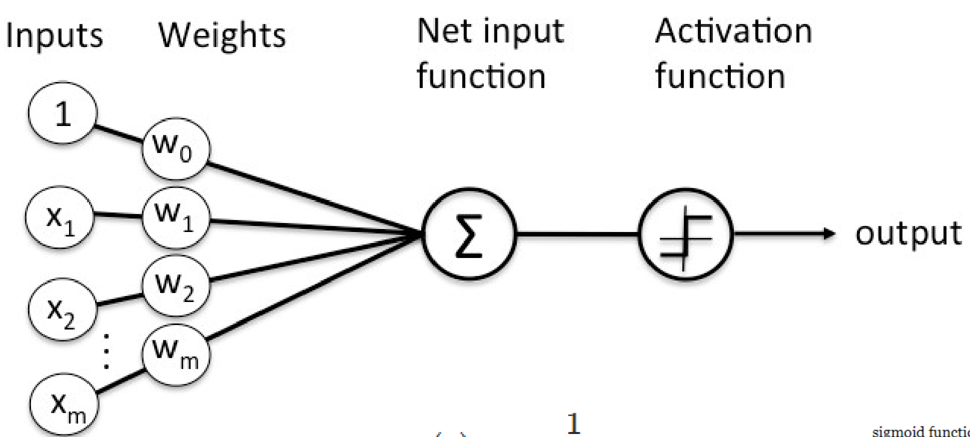
- lots of data (Big Data)
- Very flexible models
- GPGPU (powerful machines)
- Advanced algorithms for optimization, activation, regularization
- Huge research society (vision, speech, NLP, bioimaging, etc.)



VS



Neural Network



cost function

$$C(w,b) \equiv \frac{1}{2n} \sum_x \| y(x) - a \|^2.$$

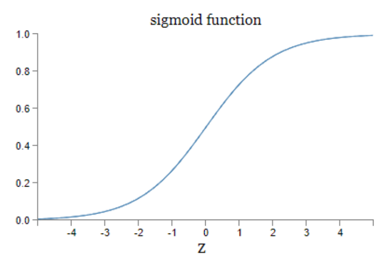
gradient descent

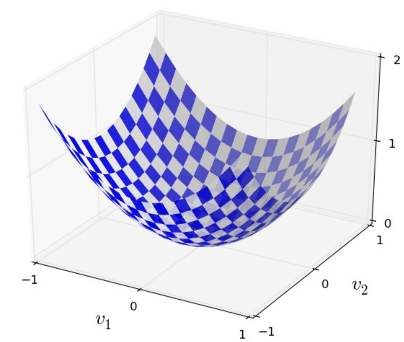
$$w_k \rightarrow w'_k = w_k - \eta \frac{\partial C}{\partial w_k}$$

$$b_l \rightarrow b'_l = b_l - \eta \frac{\partial C}{\partial b_l}.$$

$$\sigma(z) \equiv \frac{1}{1+e^{-z}}.$$

$$\frac{1}{1+\exp(-\sum_j w_j x_j - b)}.$$



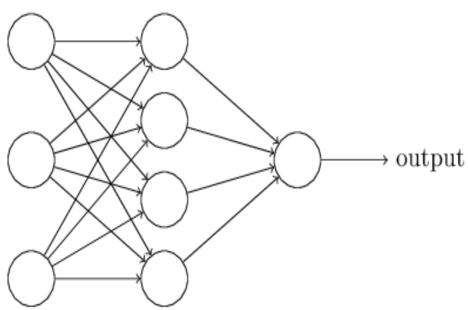


http://neuralnetworksanddeeplearning.com/chap1.html

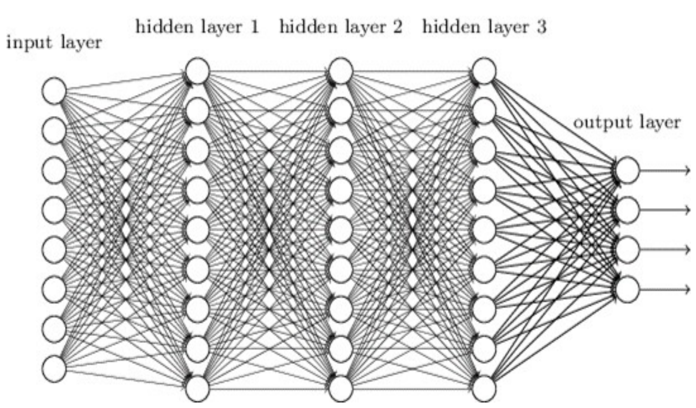
Deep Neural Network

- Machine learning algorithms based on multiple levels of representation/abstraction
  - Automatically learning good features or representations
  - Not simply using human-designed representations or input features

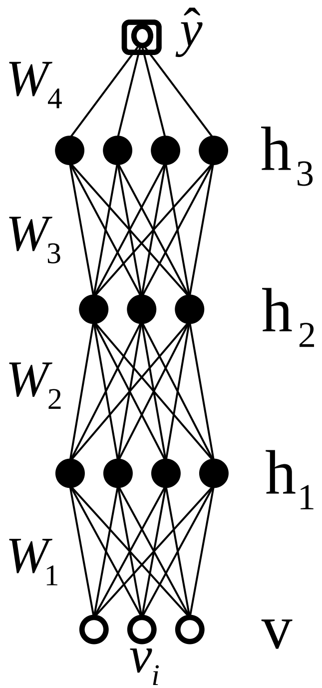
Neural Network

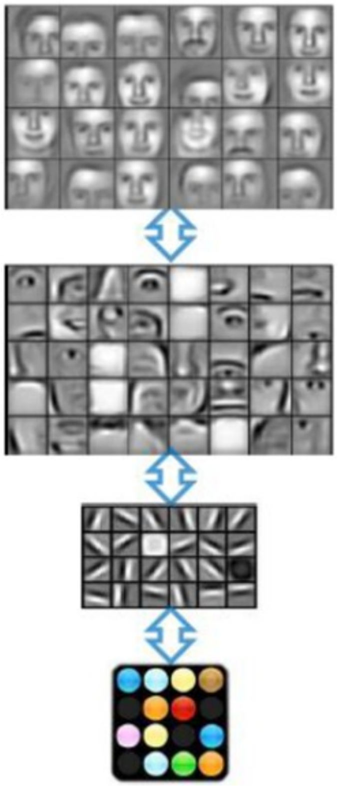


Deep neural network



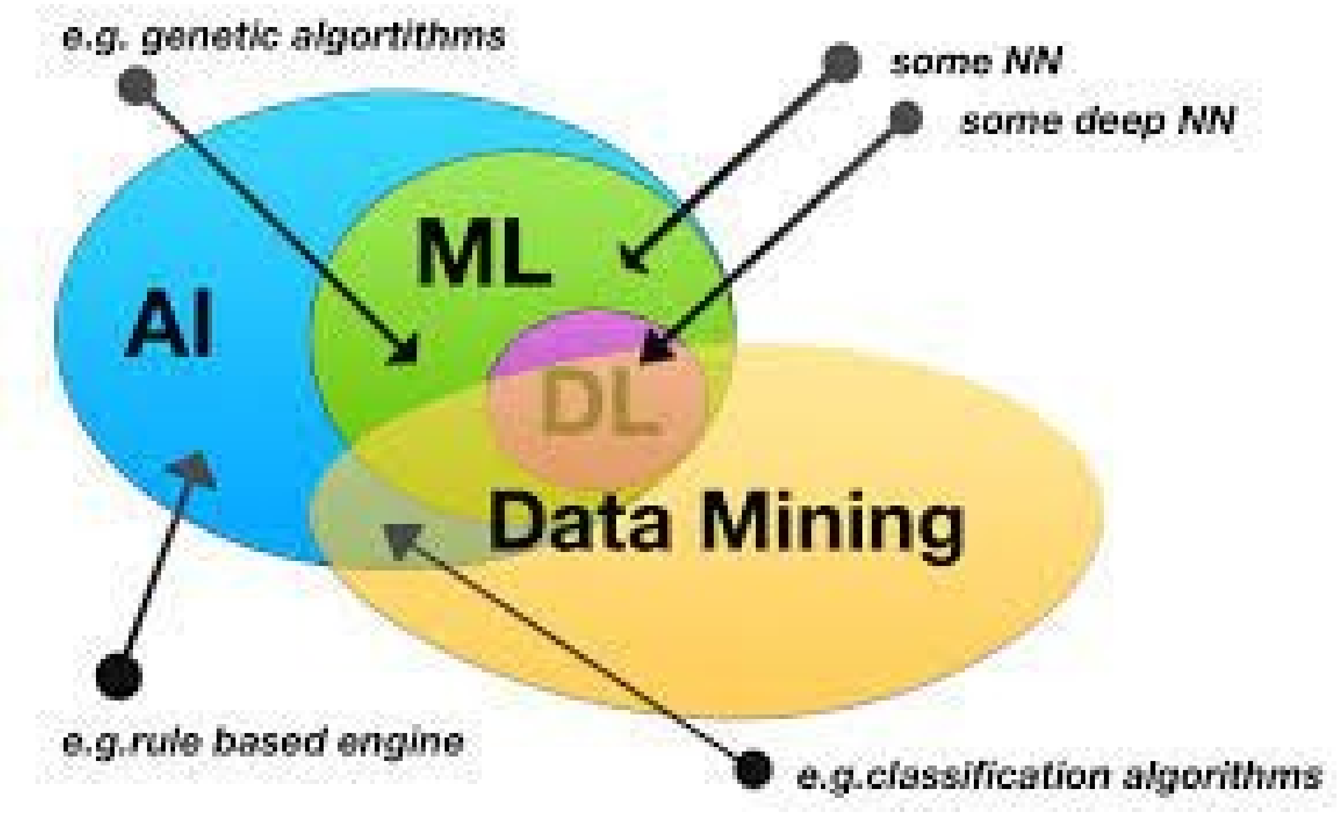
Learning of Representations

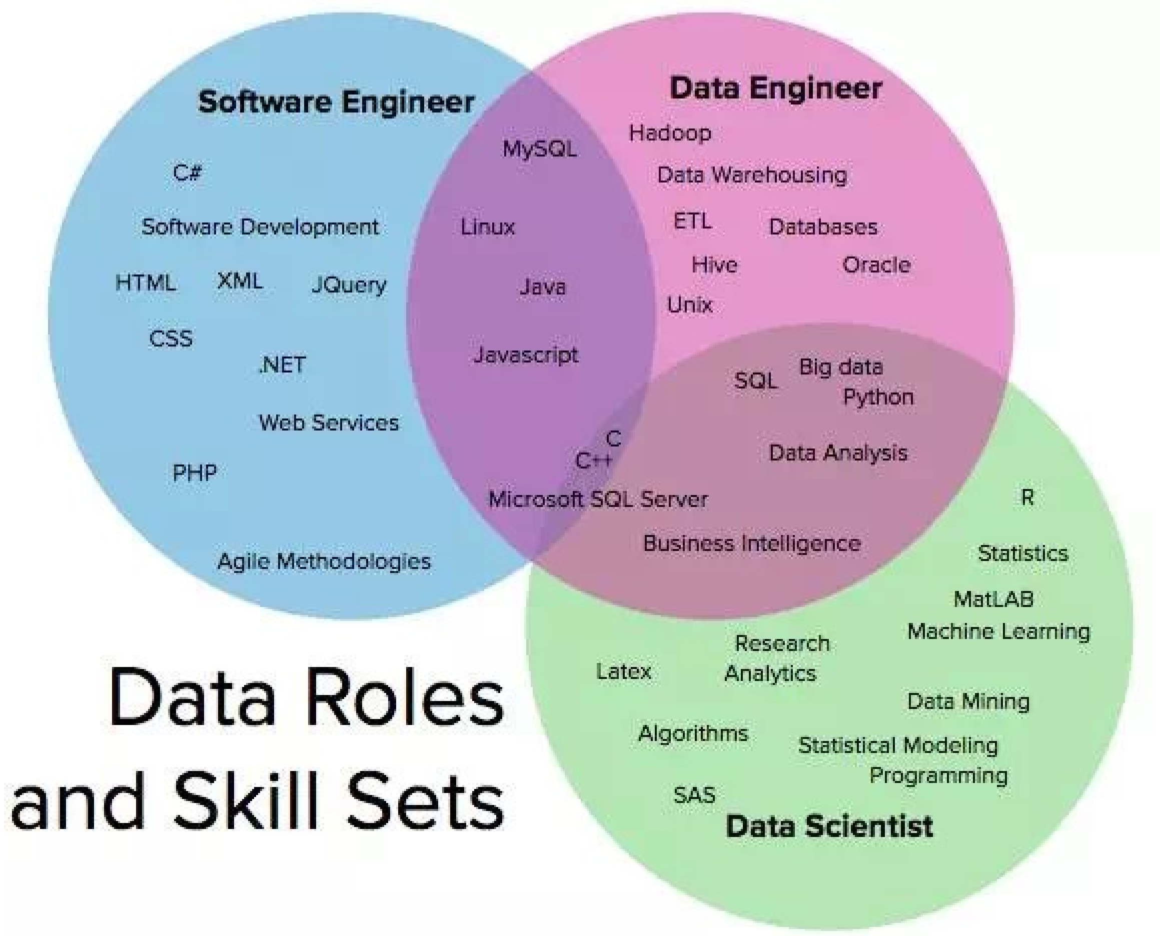




[Andrew Ng]

Data Mining vs. ML




Software Engineer
C#
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
MySQL
Linux
Java
Javascript
C
C++
Microsoft SQL Server
Data Engineer
Hadoop
Data Warehousing
ETL
Databases
Hive
Oracle
Unix
SQL
Big data
Python
Data Analysis
Business Intelligence
R
Statistics
MatLAB
Machine Learning
Research
Analytics
Latex
Data Mining
Algorithms
Statistical Modeling
Programming
SAS
Data Scientist
Data Roles
and Skill Sets

What do data scientists do?

- A data scientist is a person that has expert knowledge for turning observations into decisions.

- A data scientist devotes time to collecting data and answering questions of interest based on analyzing data.
  - Data scientists think about the physical processes and man-made systems that generate data and how to extract and organize the data in order to get answers.
  - Data scientists make the connection between observation and decision making by applying analytics to the data.
  - Data scientists observe and describe what happened, predict what might happen, and prescribe solutions for what to do.

Data Analyst Skills

Xintao Wu, Ph.D.

Professor and Charles D. Morgan/Acxiom Endowed Graduate Research Chair

Areas of Expertise

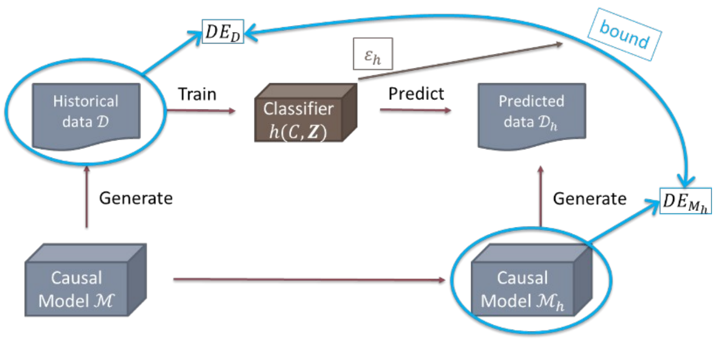
- Data Mining, Privacy and Security
- Fraud Detection
- Fair Machine Learning
- Causal Modeling and Inference

Contact Information

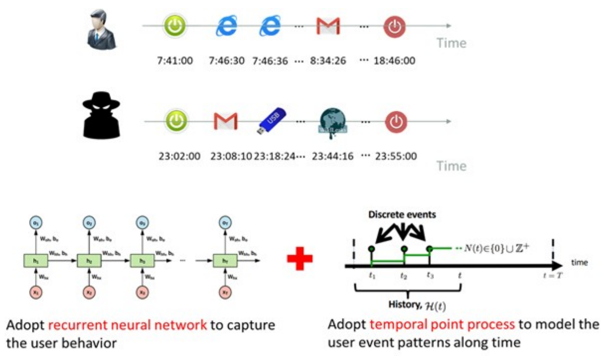
Dr. Xintao Wu
Office: JB Hunt 516
Phone: (479) 575-6519
xintaowu@uark.edu
http://csce.uark.edu/~xintaowu

Sample Projects

Fairness Aware Machine Learning

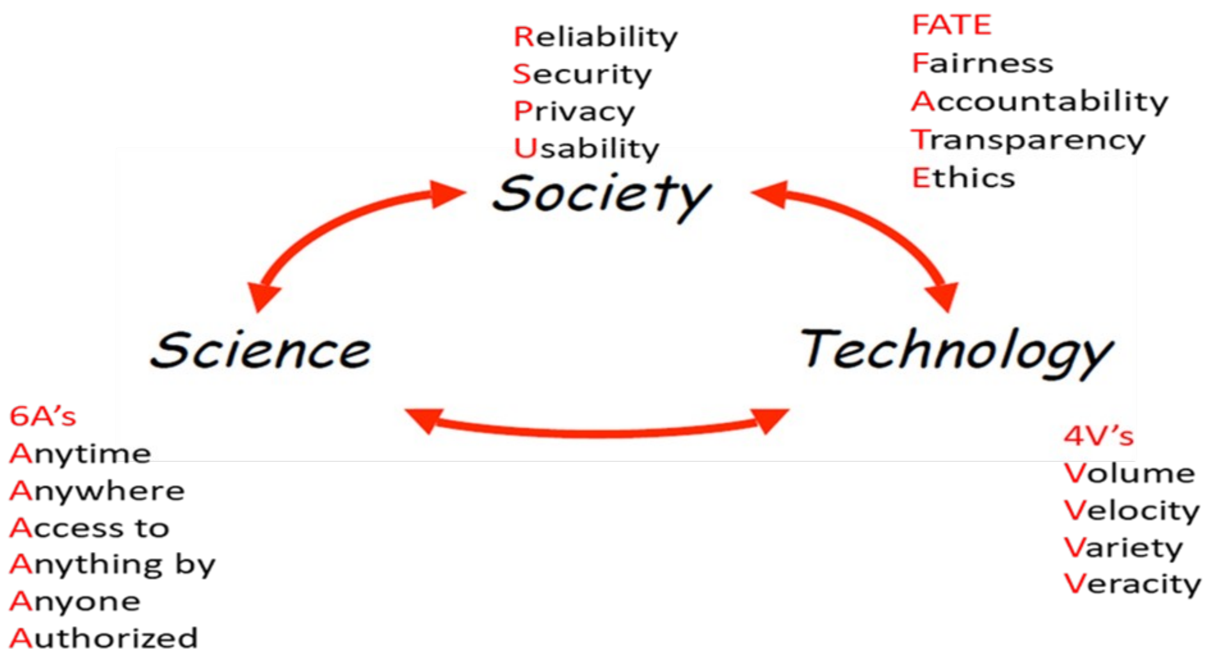


oSafari – Online Social Network Fraud and Attack Research and Identification

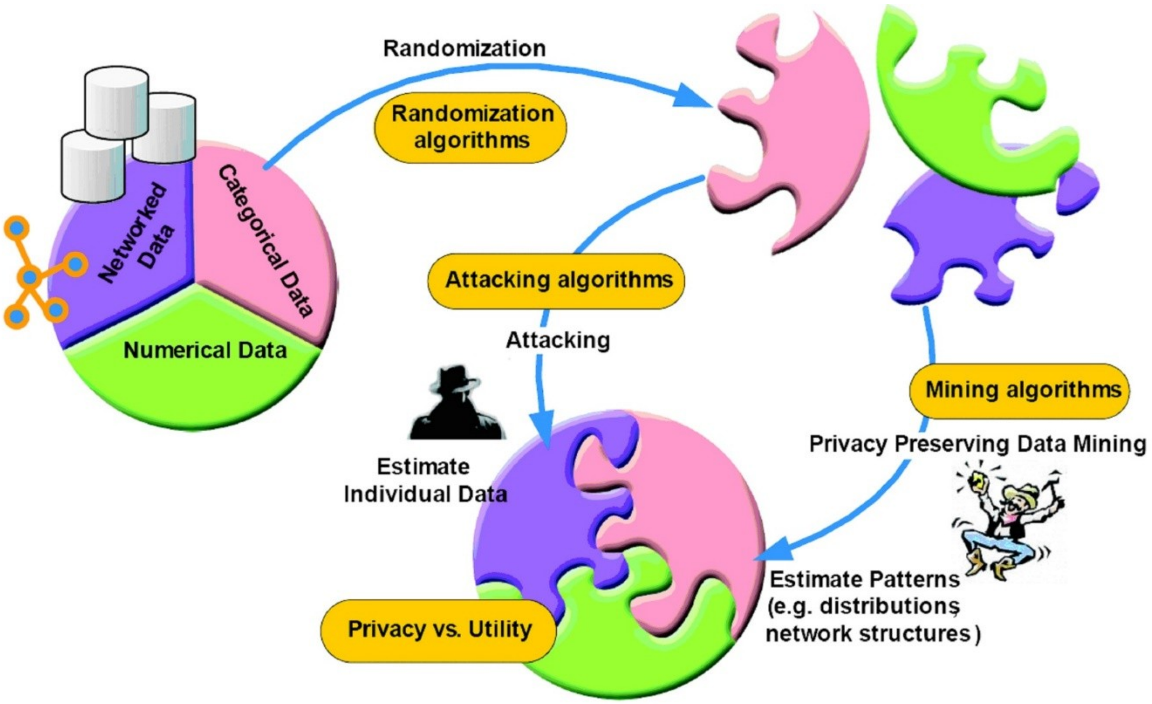


Social Awareness and Intelligent Learning

Develop cutting-edge techniques to provide privacy preservation, fairness, safety, and robustness to a variety of data analytics and learning algorithms



Privacy Preserving Data Publication

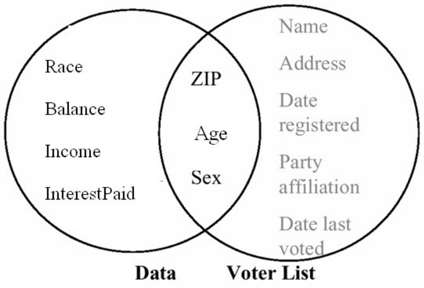


No Guarantee for Privacy Protection

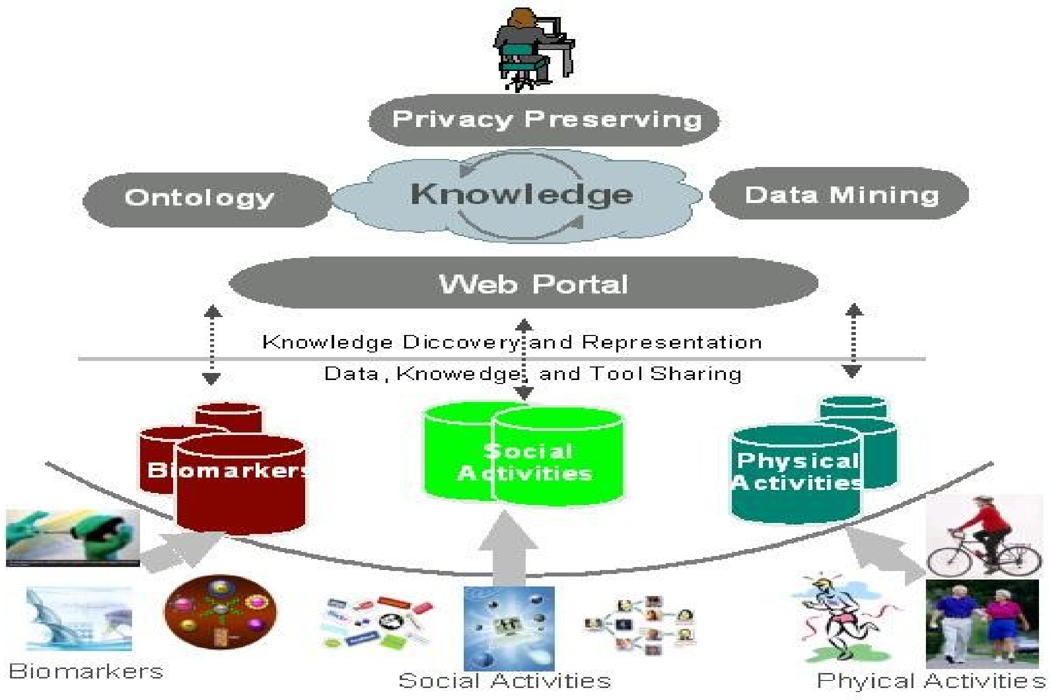
ssn	name	zip	race	…	age	Sex	income	…	disease
		28223	Asian	…	20	M	85k	…	Cancer
		28223	Asian	…	30	F	70k	…	Flu
		28262	Black	…	20	M	120k	…	Heart
		28261	White	…	26	M	23k	…	Cancer
		.	.	…	.	.	.	…	.
		28223	Asian	…	20	M	110k	…	Flu

69% unique on zip and birth date
87% with zip, birth date and gender

Generalization (k-anonymity, l-diversity, t-closeness) Randomization



Differential Privacy Preserving Data Mining/Collection



Differential Privacy

Data owner

name	sex	age	disease	salary
Ada	F	18	cancer	25k
Bob	M	25	heart	110k
Cathy	F	20	cancer	70k
Dell	M	65	flu	65k
Ed	M	60	cancer	300k
Fred	M	24	flu	20k
George	M	22	cancer	45k
Harry	M	40	flu	95k
Irene	F	45	heart	70k

Data miner

Query f

Query result + noise

Cannot be used to derive whether any individual is included in the database

Differential Guarantee

name	disease
Ada	cancer
Bob	heart
Cathy	cancer
Dell	flu
Ed	cancer
Fred	flu

K

f   count(#cancer)

f(x) + noise

3 + noise

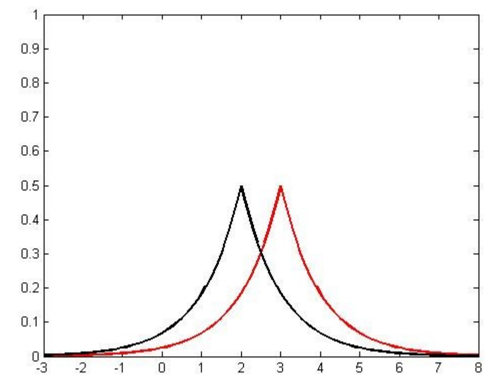
name	disease
Ada	cancer
Bob	heart
~~Cathy~~	~~cancer~~
Dell	flu
Ed	cancer
Fred	flu

K

f   count(#cancer)

f(x') + noise

2 + noise



achieving Opt-Out

Differential Privacy Preserving ML

Mechanisms to Achieve Differential Privacy

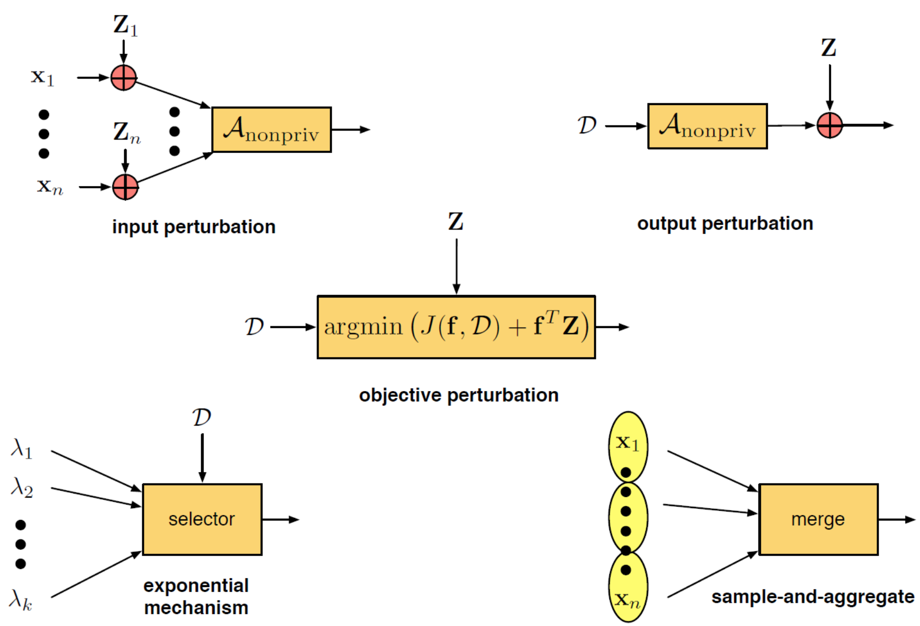


Image credit: Chaudhuri & Sarwate

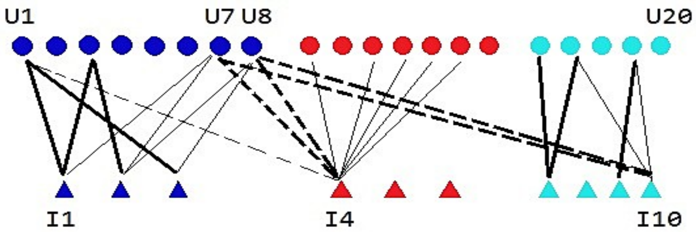
Applications of Differential Privacy

- ✓ Data Collection
- ✓ Data Streams
- ✓ Logistic Regression
- ✓ Stochastic Gradient Descents
- ✓ Recommendation
- ✓ Spectral Graph Analysis
- ✓ Causal Graph Discovery
- ✓ Embedding
- ✓ Deep Learning

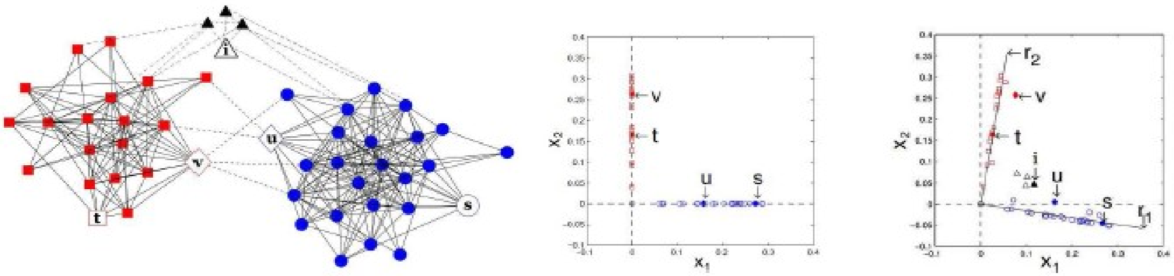
Heterogeneous Gaussian mechanism: preserving differential privacy in deep learning with provable robustness. IJCAI'19
DPNE: Differentially Private Network Embedding. PAKDD'18
Preserving differential privacy in convolutional deep belief networks. ML'17
Adaptive Laplace mechanism: differential privacy preservation in deep learning, ICDM'17
Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. AAAI'16

Spectral Graph Analysis for Fraud Detection

Reviews
Ratings
Ranks



- Bot-committed
- Money-motivated



(a) Topology snapshot

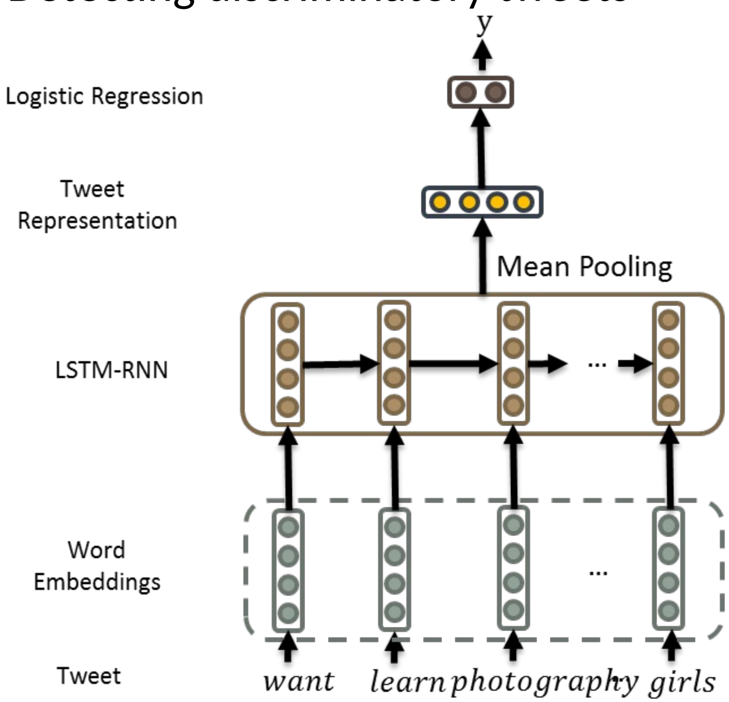
(b) w/o inter-comm edges and no attack

(c) with inter-comm edges and an attack

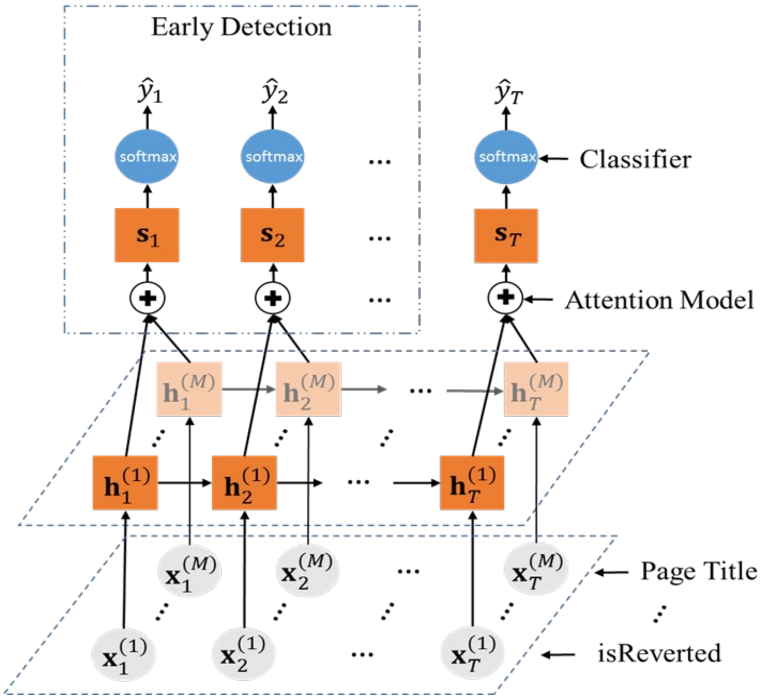
Fig. 2. Detecting random link attack in 2-D spectral space.

Deep Learning for Fraud Detection

Detecting discriminatory tweets

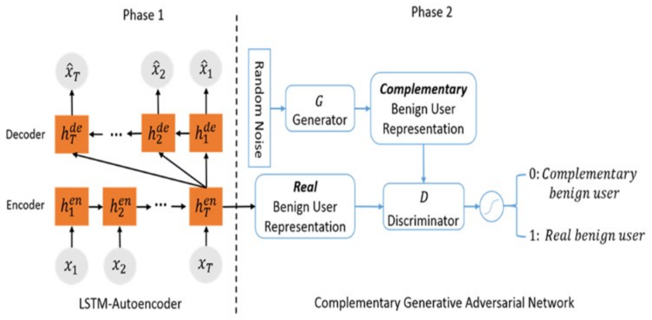


Vandal detection from Wikipedia



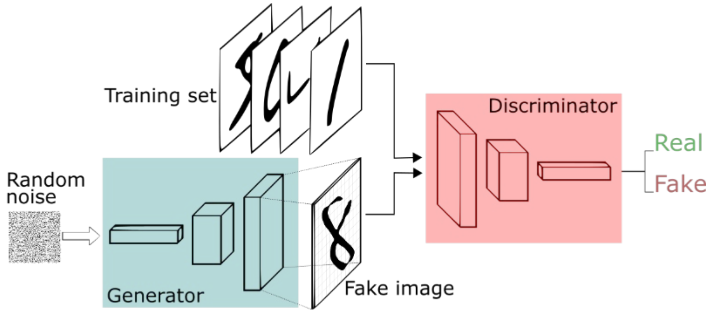
Fraud Detection

- Core Techniques
  - Spectral graph analysis and graph emebedding
  - Multi-LSTM
  - One-class generative adversarial networks
  - Neural temporal point processes

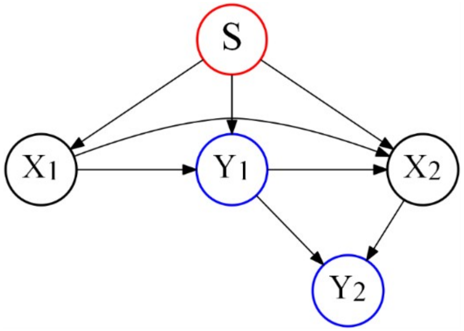


- Recent Publications
  - Insider threat detection via hierarchical neural temporal point processes (NeurIPS-WTPP'19)
  - SAFE: A neural survival analysis model for fraud early detection (AAAI'19)
  - One-class adversarial nets for fraud detection (AAAI'19)
  - Dynamic anomaly detection using vector auoregressive model (PAKDD'19)
  - Spectrum-based deep neural networks for fraud detection (CIKM'17)
  - Wikipedia vandal early detection: from user behavior to user embedding (ECML-PKDD'17)

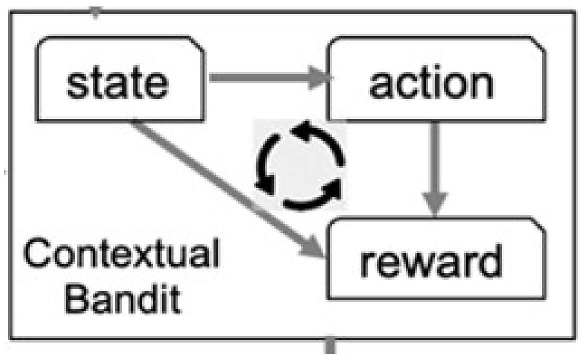
Fair and Robust Learning



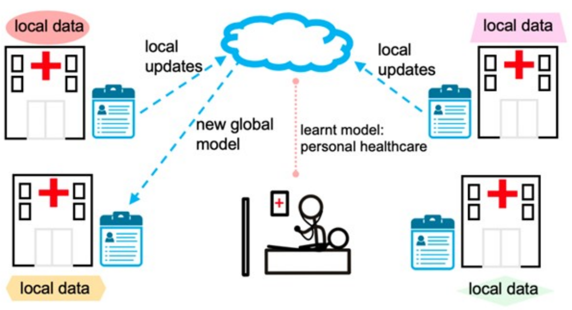
Generative adversarial networks



Causal learning



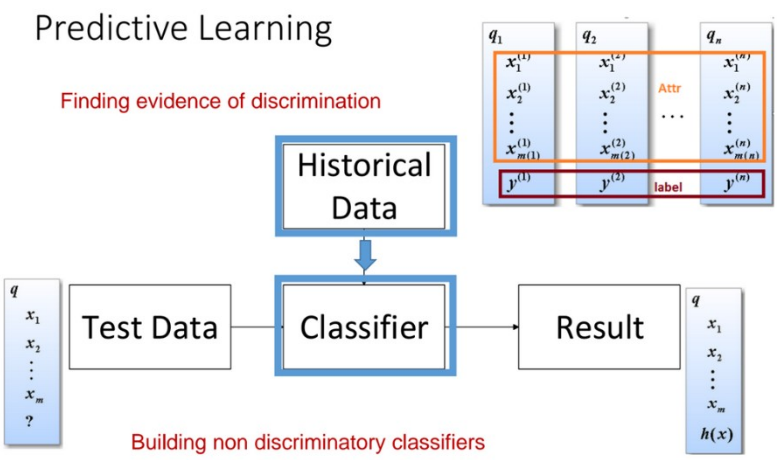
Online recommendation



Federated learning

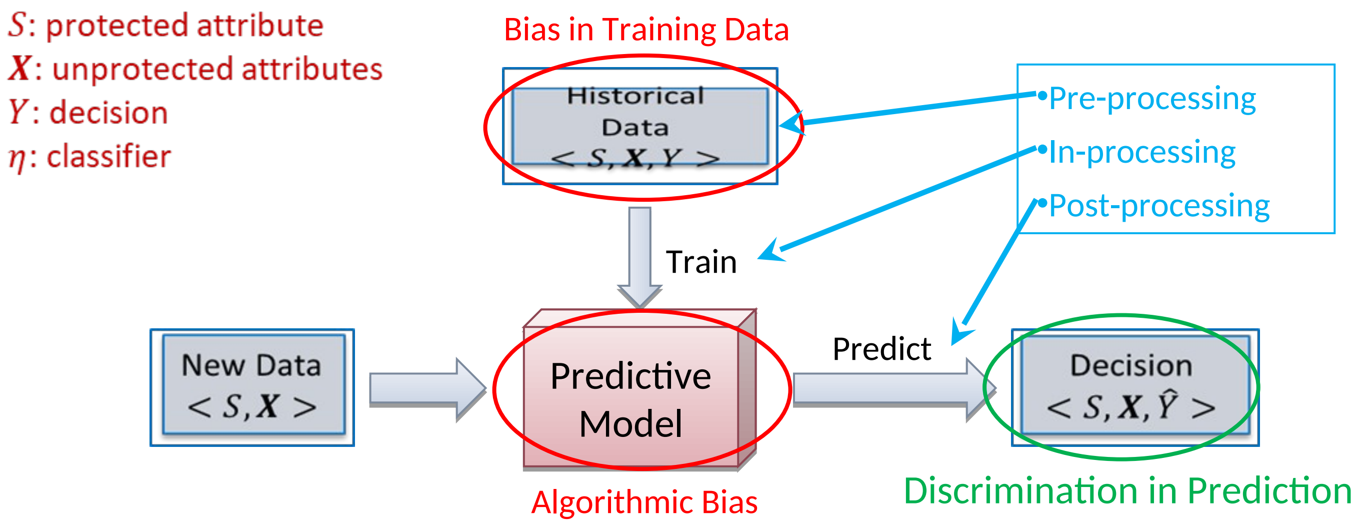
Fairness Aware Learning

Support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks.



Fair Machine Learning

Classification, recommendation, ranking, resource allocation, federated learning, …



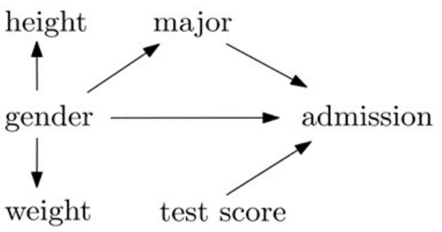
Demographic Parity: $P(\eta(\boldsymbol{X}) = 1 | S = 1) = P(\eta(\boldsymbol{X}) = 1 | S = 0)$

Equal Opportunity: $P(\eta(\boldsymbol{X}) = 1 | Y = 1, S = 1) = P(\eta(\boldsymbol{X}) = 1 | Y = 1, S = 0)$

Anti-discrimination Learning: From Association to Causation. KDD tutorial 2018

Causal Fairness

- Core Techniques
  - Causal model and causal graph
  - Intervention and *do*-operator
  - Path-specific effect
  - Counterfactual analysis

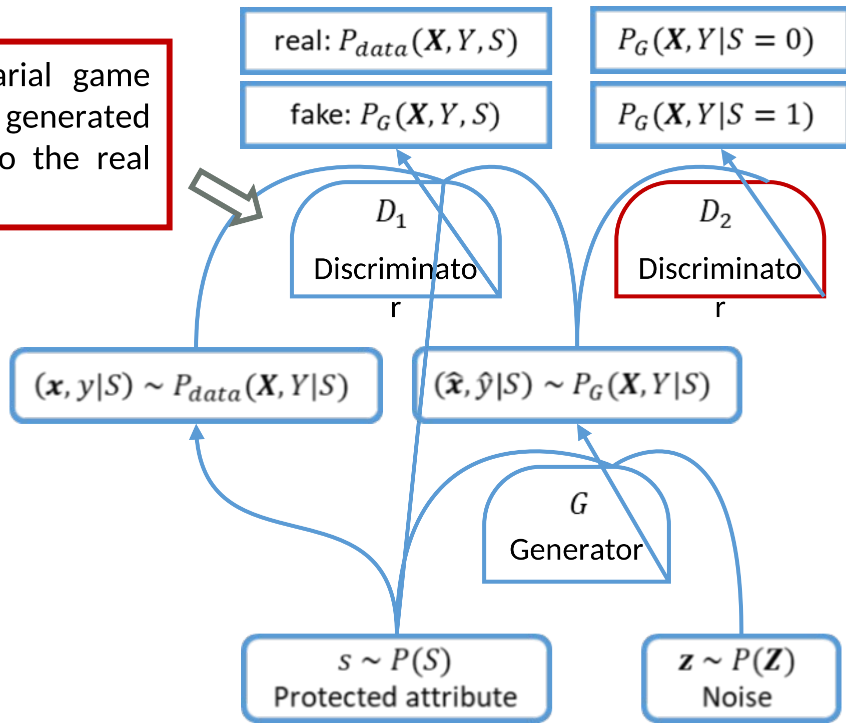


- Challenges
  - Identifiability
  - Interference
  - Representation with different types of data
- Recent Publications
  - Fair multiple decision making through soft intervention (NeurIPS'20)
  - Path-specific counterfactual fairness (NeurIPS'19)
  - Counterfactual fairness: unidentification, bound, and algorithm (IJCAI'19)
  - On convexity and bounds of fairness-aware classification (WWW'19)
  - On discrimination discovery and removal in ranked data using causal graph (KDD'18)
  - Achieving non-discrimination in prediction (IJCAI'18)
  - Achieving non-discrimination in data release (KDD'17)

FairGAN: Fairness-aware Generative Adversarial Networks

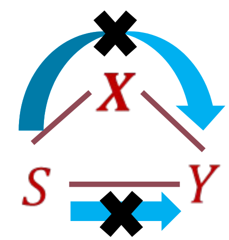
FairGAN

One adversarial game ensures the generated data close to the real data.



The other adversarial game ensures data fairness by preventing the disparate treatment and disparate impact.

$$P_G(\mathbf{X}, Y|S=0) = P_G(\mathbf{X}, Y|S=1)$$

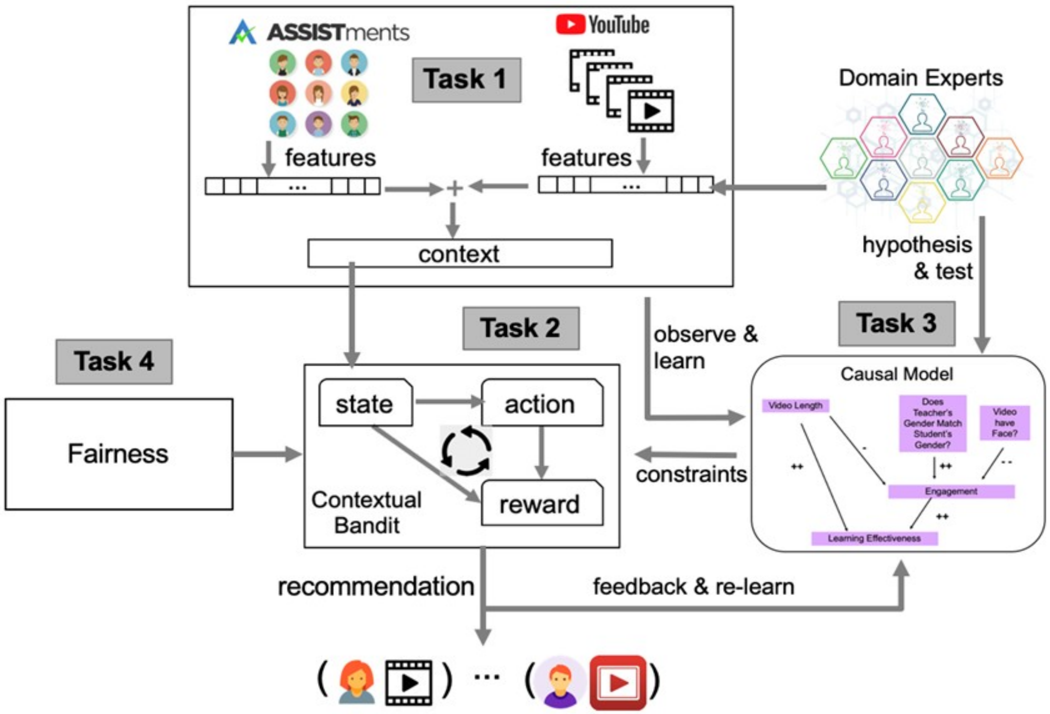


Fairness in Data

- Statistical Fairness
- ε-fairness
- Causal fairness
- ...

FairGAN: Fairness-aware Generative Adversarial Networks. BigData 2018.
FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets. BigData 2019.
Achieving Causal Fairness through Generative Adversarial Networks. IJCAI 2019.

Fair Recommendation



Challenges

- Online learning with cold-start problem
- Better regret bound via causal modeling
- Group vs. individual vs. counterfactual fairness

Achieving Counterfactual Fairness for Causal Bandit, https://arxiv.org/abs/2109.10458

Transferable Contextual Bandits with Prior Observations. PAKDD'21

Achieving User-Side Fairness in Contextual Bandits. https://arxiv.org/abs/2010.12102

Fair Federated Learning

Challenges

- Global vs. Local Fairness
- Horizontal vs. Vertical partition
- Convergence due to non-IID
- Computational vs. Communication Cost

Fairness-aware Agnostic Federated Learning. SDM’21

Robust Machine Learning

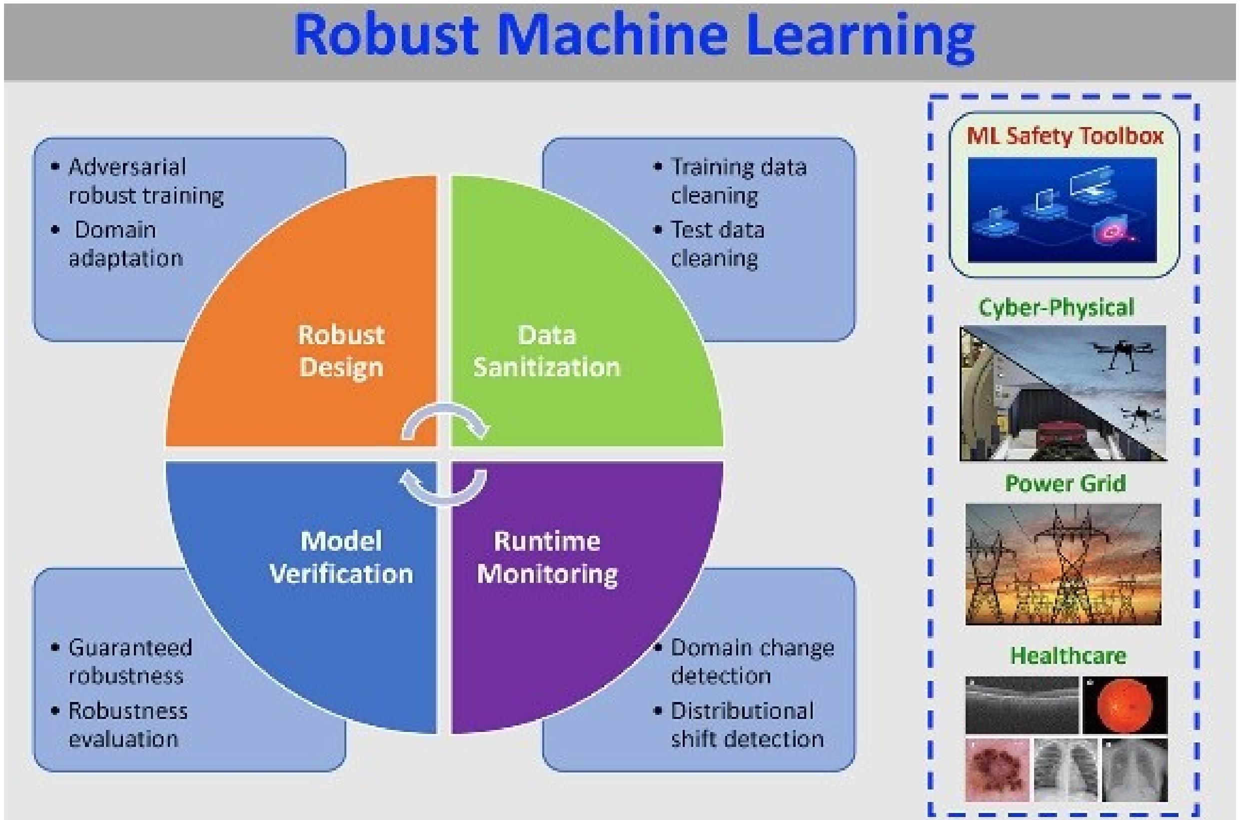


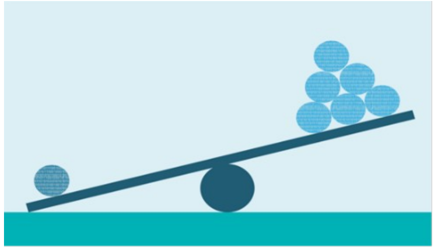
Image credit to Jeremy Thomas

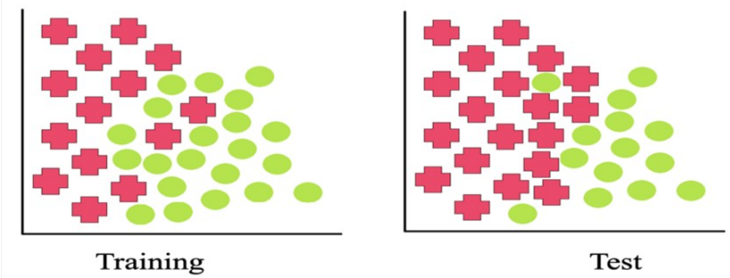
Challenges

- Robustness certification
- Robust machine learning based on causal representation

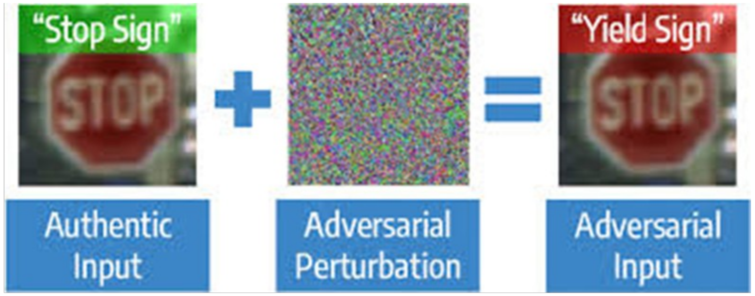
Achieving Robustness in ML

- Under Distribution Shift





- Under Adversarial Attack



Fair and Robust Classification Under Sample Selection Bias, CIKM'21
Fair Regression Under Sample Selection Bias, BigData'22
Poisoning Attacks on Fair Machine Learning, DASFAA'22
Defending Evasion Attacks via Adaptive Training, BigData'22