DATA MINING

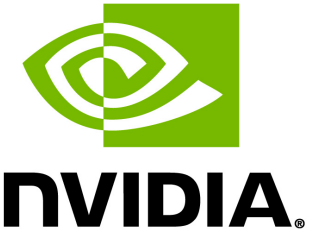
/CISC 873 - Steven Ding
/Week #1/Lecutre 1

MEME Winter of the Week:

[Current meme winner: tindur-sigurdarson - Checkout the competition]

whomai

- Dr. Steven Ding
  - AI, Machine Learning, Data Mining, and Cybersecurity
    - PhD, McGill University (2019)
    - Assistant professor, Queen's (2019–)
    - Director, L1NNA Research Lab, l1nna.com
      - AI for security, and security for AI
    - Created Kam1n0
  - The father of a child



Terminology:

Data mining?

Terminology:

Data mining?

Terminology:

Data mining?

Terminology:

Data mining

Data analytics

Knowledge discovery

Inductive modelling of systems from data

Machine learning

Data Science

Topics & Schedule

- W1
  - Introduction & Ethical/Bias/Security Issues
- W2-W5
  - Tabular/Relational/Multi-View Data Mining
- W6-W9
  - Time Series/Sequential/Transactional Data Mining
- W10-W13
  - Graph/Social Network Data Mining

CISC 251+271+371+372 with more topics covered

In the context of DM

(rapid-fire DS bootcamp)

Workload and Grading

- Be prepared to spend adequate time and effort on this course.

Assignments (Option #1 100% total)	Due Date	Individual Project (Option #2 100% total)	Due Date
Assignment #0	TBD	Project Proposal	TBD
Assignment #1	TBD	Phase 1 source code and results	TBD
Assignment #2	TBD	Phase 2 source code and results	TBD
Assignment #3	TBD	Phase 3 source code and results	TBD
Assignment #4	TBD	Final report	TBD
Assignment #5	TBD		TBD
Assignment #6	TBD		TBD
Assignment #7	TBD		TBD

Official Learning Outcome

- LO1: Develop a rigorous understanding of **fundamental concepts** in data mining.
- LO2: Describe the general **infrastructure** required for data mining workflow.
- LO3: Describe the **mechanisms and requirements** of major data mining functions.
- LO4: Design and execute a **process** for modeling complex systems that is appropriate, effective, and revealing.
- LO5: **Fine-tune** a process design to improve the performance following the general data science workflow.
- LO6: Acquire hands-on experiences with basic and advanced **tools** for data mining.

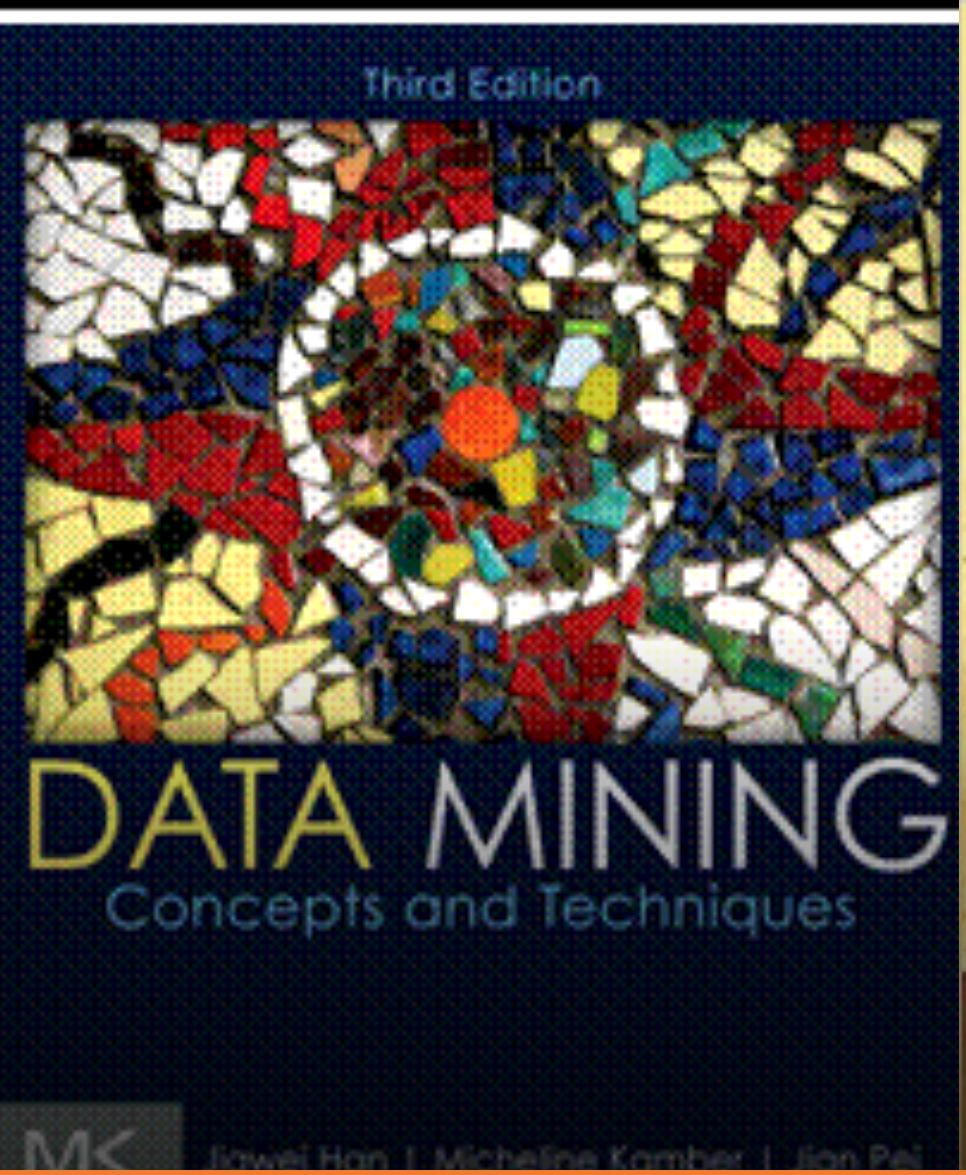
Unofficial Learning Outcome

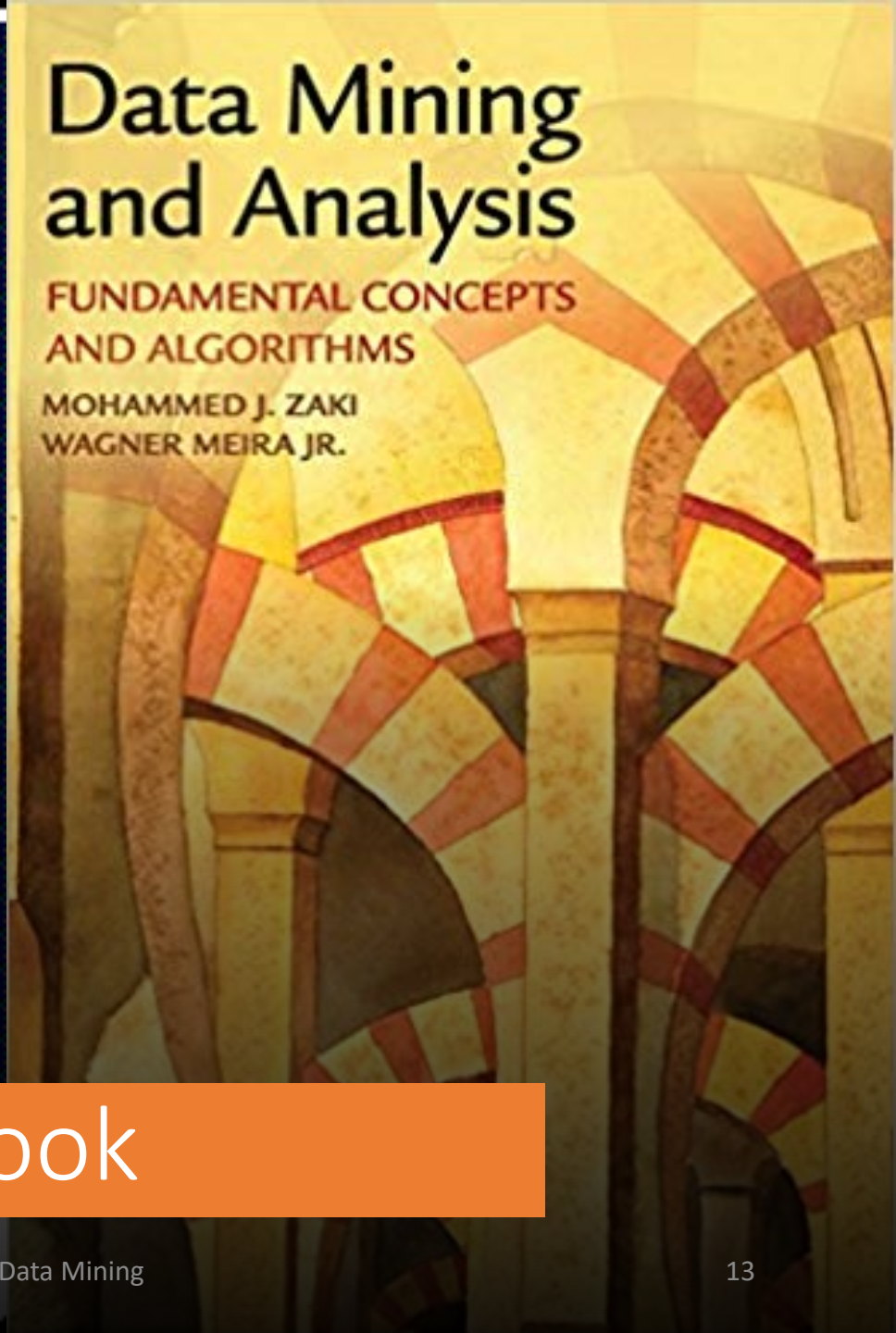
- LO1: Data => Knowledge => $ & Impact
- LO2: Have **Fun**
- LO3: Good collections of DS/ML/DM **Memes**

E-mail Policy

- When you send e-mail to me, put **"873"** in the subject area, so that it can pass the spam filter.

- **Course email list is a must-read**.

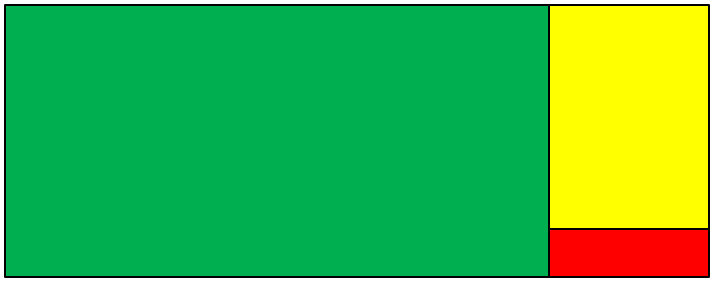




(Optional) Textbook

Academic Integrity

- *READ*:
  - https://www.queensu.ca/artsci/students-at-queens/academic-integrity
  - https://www.queensu.ca/academicintegrity/students/avoiding-plagiarismcheating
- It is **not** allowed to:
  - Falsify data or research results.
  - Copy from ANYWHERE **without** giving citation.
  - …



Feedback and Suggestions

- Your feedback and suggestions are most welcome!

- Two anonymous course evaluations:
  - Mid-course evaluation
    - Unofficial
    - Gathering feedback, so I can improve in the rest of this semester.
  - Official course evaluation
    - Official -- for administrative purpose
    - For improving the next course offering

DATA MINING

/CISC 873 - Steven Ding
/Week #1/Lecture 2
Introduction to Data Mining

Terminology:

Data mining

Data analytics

Knowledge discovery

Inductive modelling of systems from data

Machine learning

Data Science

What is Data Mining?

- Data mining (knowledge discovery from data)
  - Process of extracting **interesting** (non-trivial, implicit, previously unknown and potentially useful) **patterns** or **knowledge** from **huge** amount of data, preferably in an efficient, scalable, and practical approach.
  - Data mining: a misnomer?

- Alternative names
  - **Knowledge discovery** (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Data Science

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
  - 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/analytics 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

Watch out: Is everything “data mining”?

- These are **not** considered to be data mining:
- Simple **search**
  - A simple inverted index in your GLIS 617 assignment
- **Query** processing
  - Submitting a SQL query to a database to calculate the average age of the class
- (Deductive) **expert** systems
  - Runny nose + Sore throat + Cough + Low-grade fever → Cold

Why Data Mining?

- "**Necessity** is the mother of invention"
  - Data mining
  - Automated analysis of massive data sets
- The Explosive **Growth** of Data: from terabytes to petabytes to …
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data but starving for knowledge!

Why Data Mining?

- Make **use** of data assets

- There is a big **gap** from stored data to knowledge; and the transition won't occur automatically.

- Many interesting information that one wants to find **cannot be found using database queries**
  - "Find people likely to buy my products"
  - "Which movies should be recommended to each customer?"
  - "Who are likely to respond to my promotion?"

Why Data Mining?

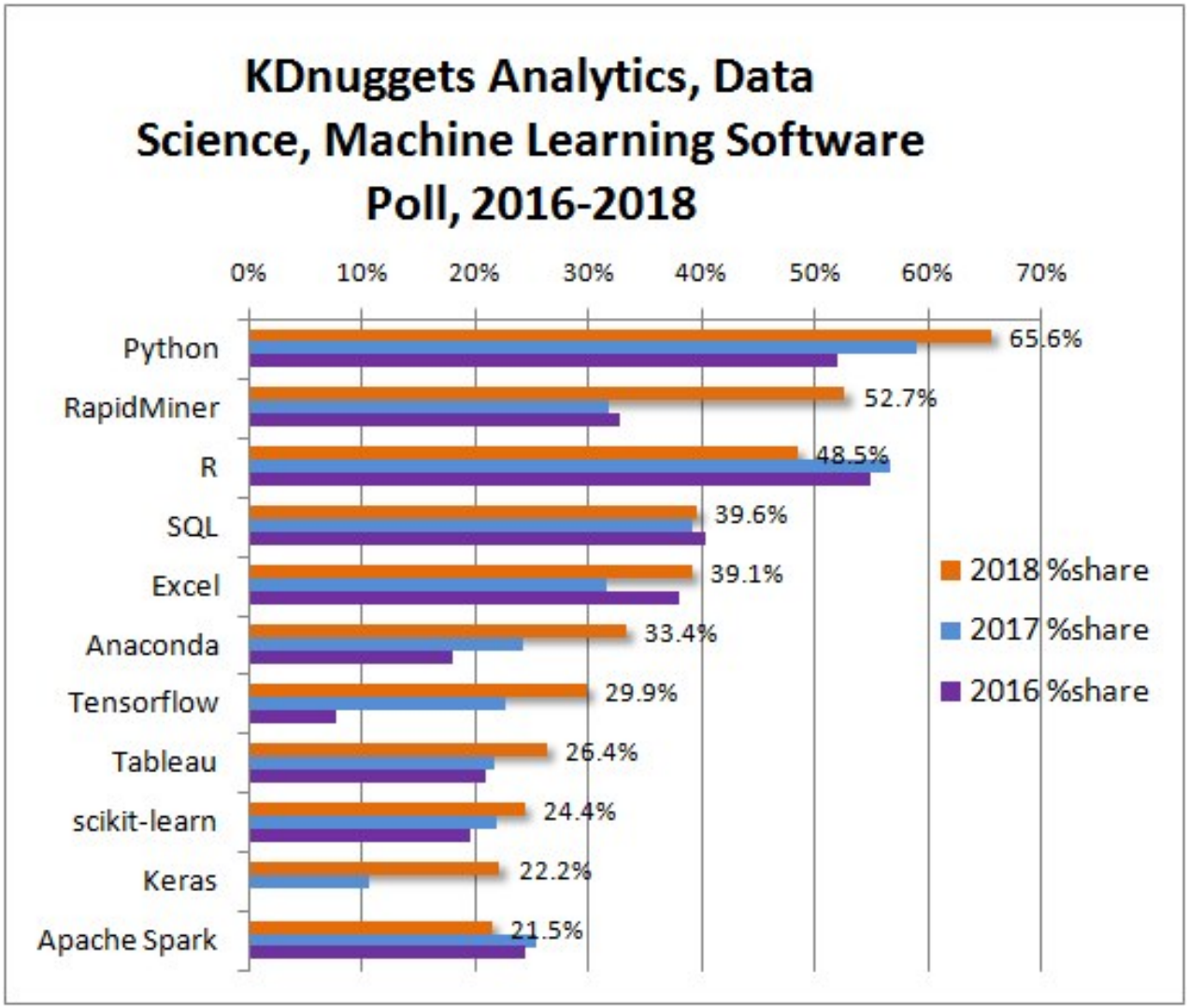
- The **data is abundant**.
- Computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every large company is doing it

Data Mining Requires Programming?

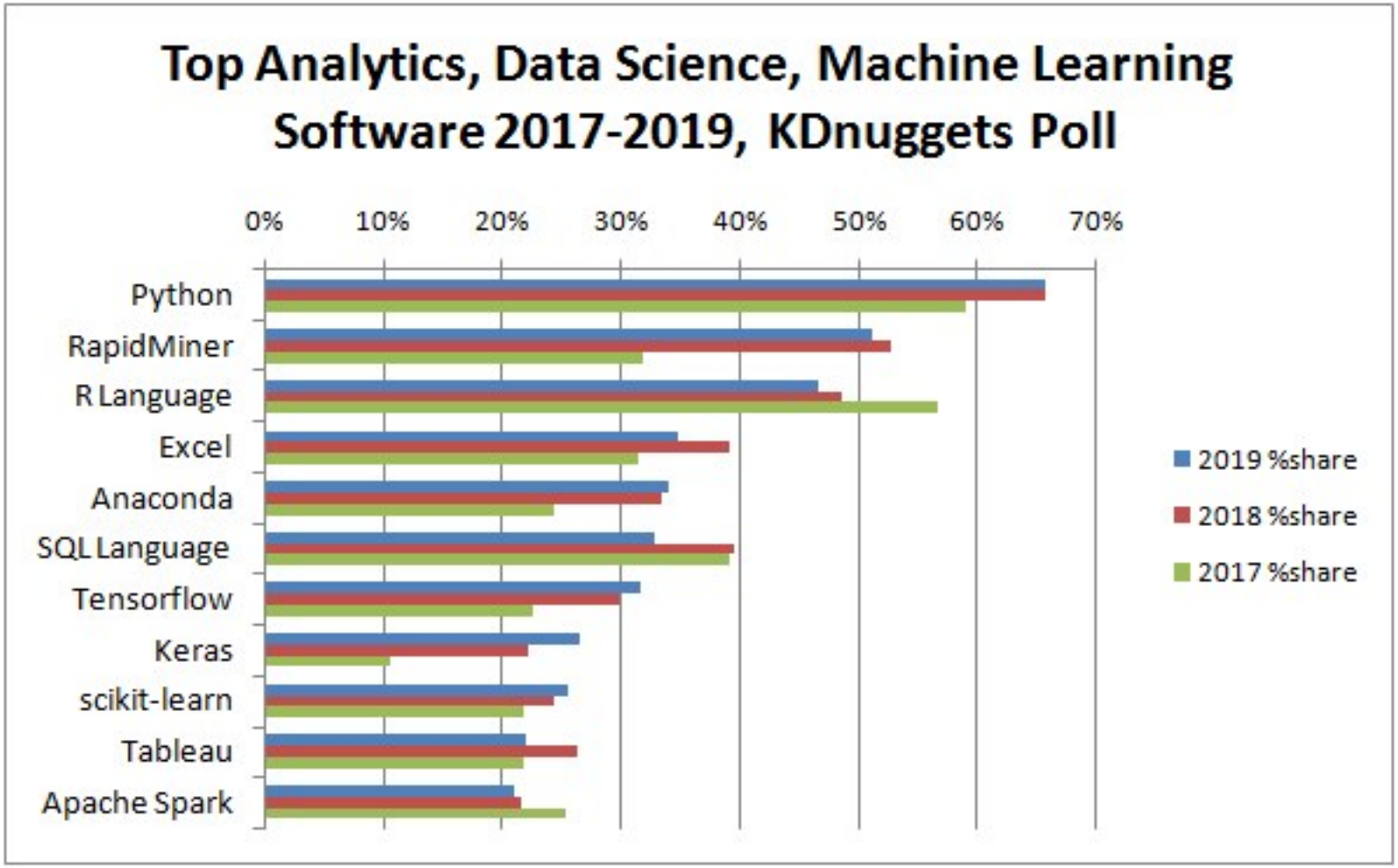
- Lots of available UI-based tools for DM.
  - Rapid Miner
  - Weka
- Programming provides better flexibility.
- Programming requires more debugging.
- Programming for new data mining models.

Data Mining Requires Programming?

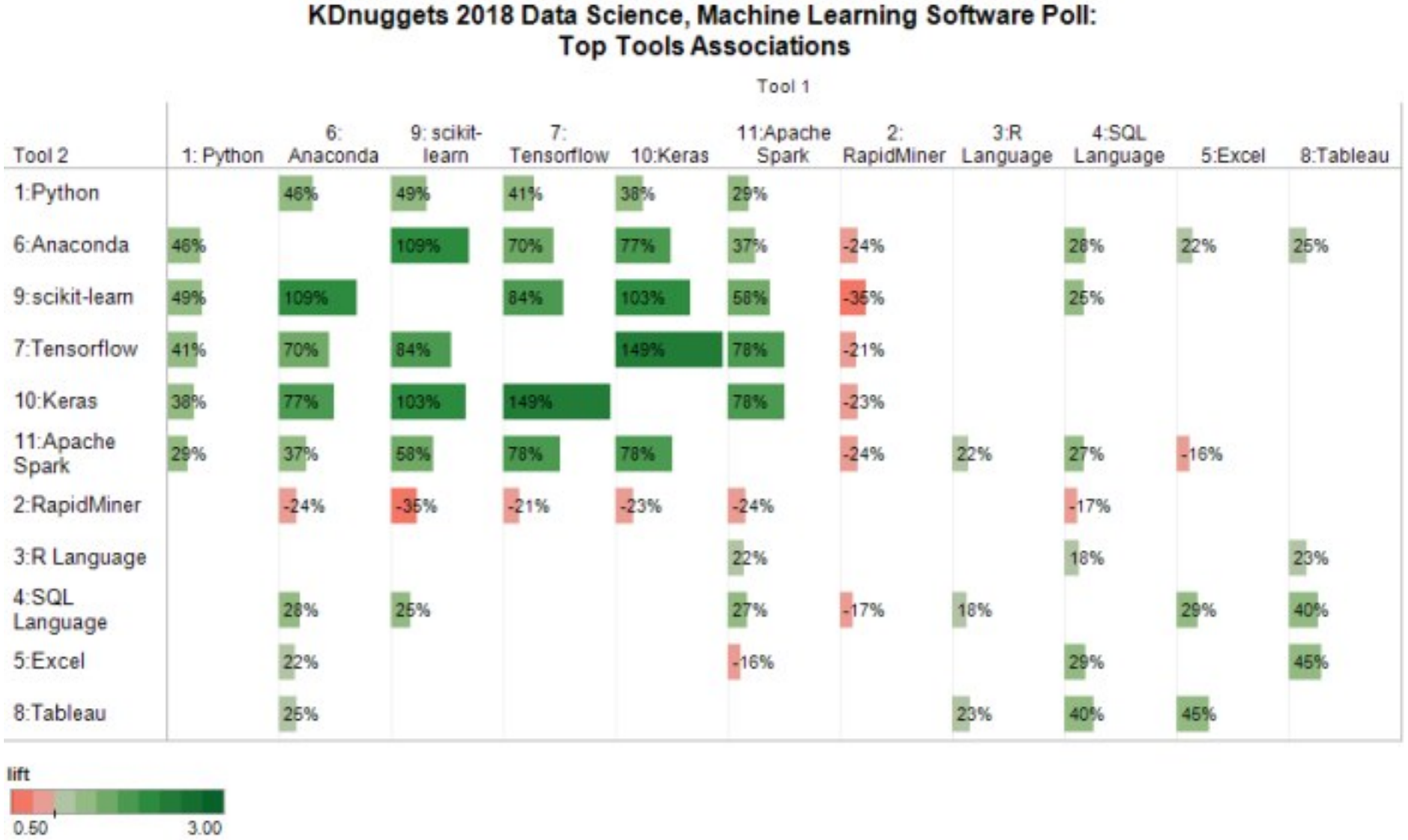
Data Mining Requires Programming?



Data Mining Requires Programming?



Data Mining Requires Programming?



Evolution of Database Technology

- 1960s:
  - Data collection, database creation, information management systems (IMS), and network database management systems (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.)

Evolution of Database Technology

- 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
- 2010s:
  - Big data: large, complex, fast growing, inconsistent
  - NoSQL: efficient key-value stores and document-oriented databases; avoid join operation; designed to scale horizontally.
- 2020s:
  - Heterogeneous Information Data

Multi-Dimensional View of Data Mining

- Data to be mined
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- Knowledge to be mined (or: Data mining functions)
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- Techniques utilized
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On 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

Data Mining Functions

1. Generalization and Summarization
2. Association and Correlation
3. Classification & Prediction
4. Clustering
5. Outlier/Anomaly Analysis
6. Time and Ordering
7. Structure and Network Analysis

(1) Generalization & Characterization

- **Generalization** - A process that abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.

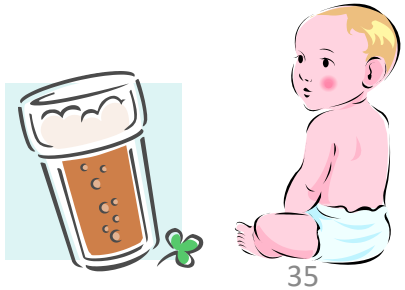
- **Characterization** - A summarization of general features of objects in a target class and produces what is called *characteristic rules*.

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics

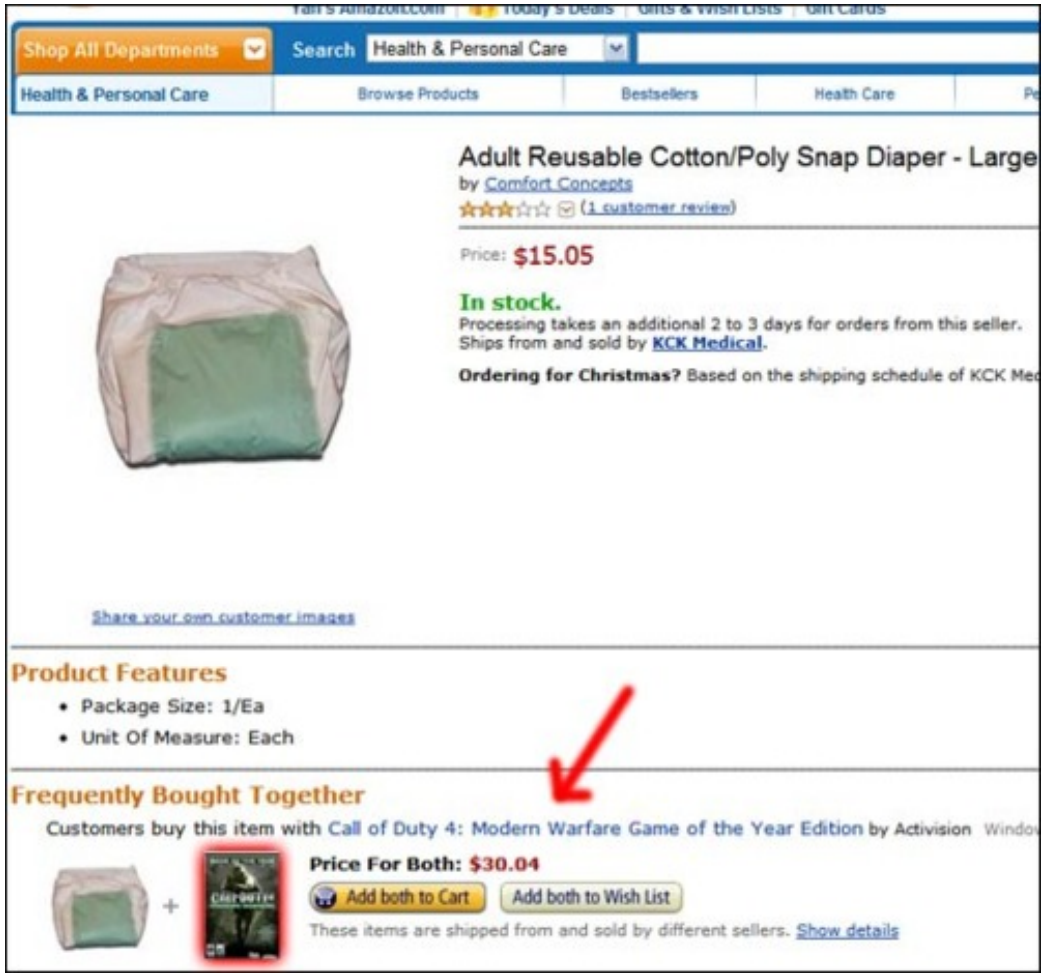
(2) Association and Correlation

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer
    - [Support=40%, Confidence=67%]
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Transaction database	
TID	**Items bought**
100	bread, butter, diaper
200	bread, butter, diaper, beer
300	bread, butter, pencil
400	orange, pencil, beer
500	diaper, beer, pencil, bread



(2) Association and Correlation



(2) Association and Correlation

- "Our recommendation of accounts to follow is based on user behavior: *users who follow account A also follow account B, so if you follow A you are likely to also want to follow B*"

https://www.vox.com/recode/2020/2/25/21152585/tiktok-recommendations-profile-look-alike

Ex. 1: Market Analysis

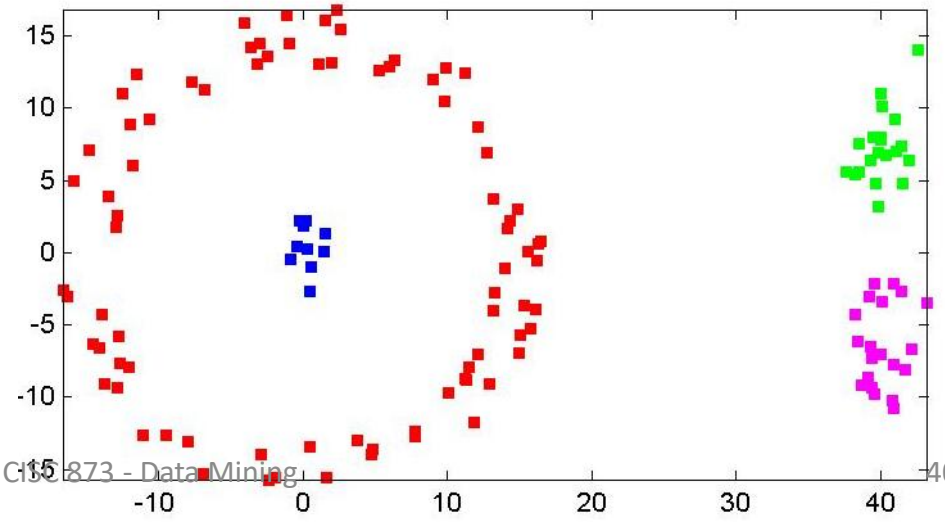
- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- *Cross-market analysis*—Find associations/co-relations between product sales, & predict based on such association
- *Customer profiling*—What types of customers buy what products (clustering or classification)
- *Customer requirement analysis*
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data *central tendency and variation*)

(3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …
- What is the difference between classification and prediction?

(4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing inter-class similarity
- Many methods and applications



(5) Outlier Analysis

- Outlier analysis
  - ***Outlier***: A data object that does not comply with the general behavior of the data
  - ***Noise*** or ***exception***? — One person's garbage could be another person's treasure
  - Methods: byproduct of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

(6) Time and Ordering

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis

- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

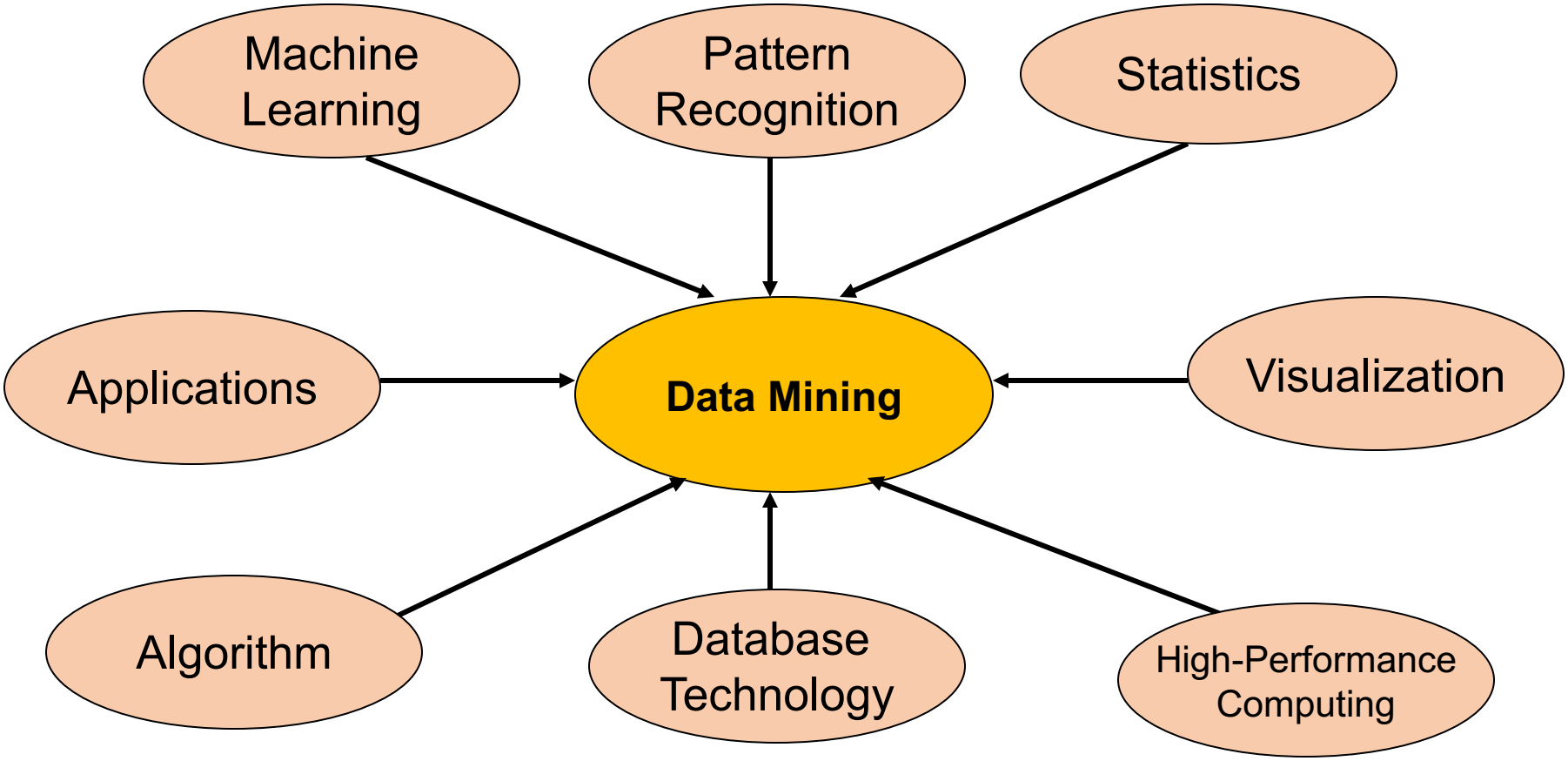
(7) Structure and Network Analysis

- Graph mining
  - Finding frequent **subgraphs** (e.g., chemical compounds), trees (XML), **substructures** (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be in multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- ***Descriptive*** vs. ***predictive***
- Coverage
- Typicality vs. novelty
- Accuracy
- Timeliness
- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient
- Evaluation of mined knowledge → directly mine only interesting knowledge?

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

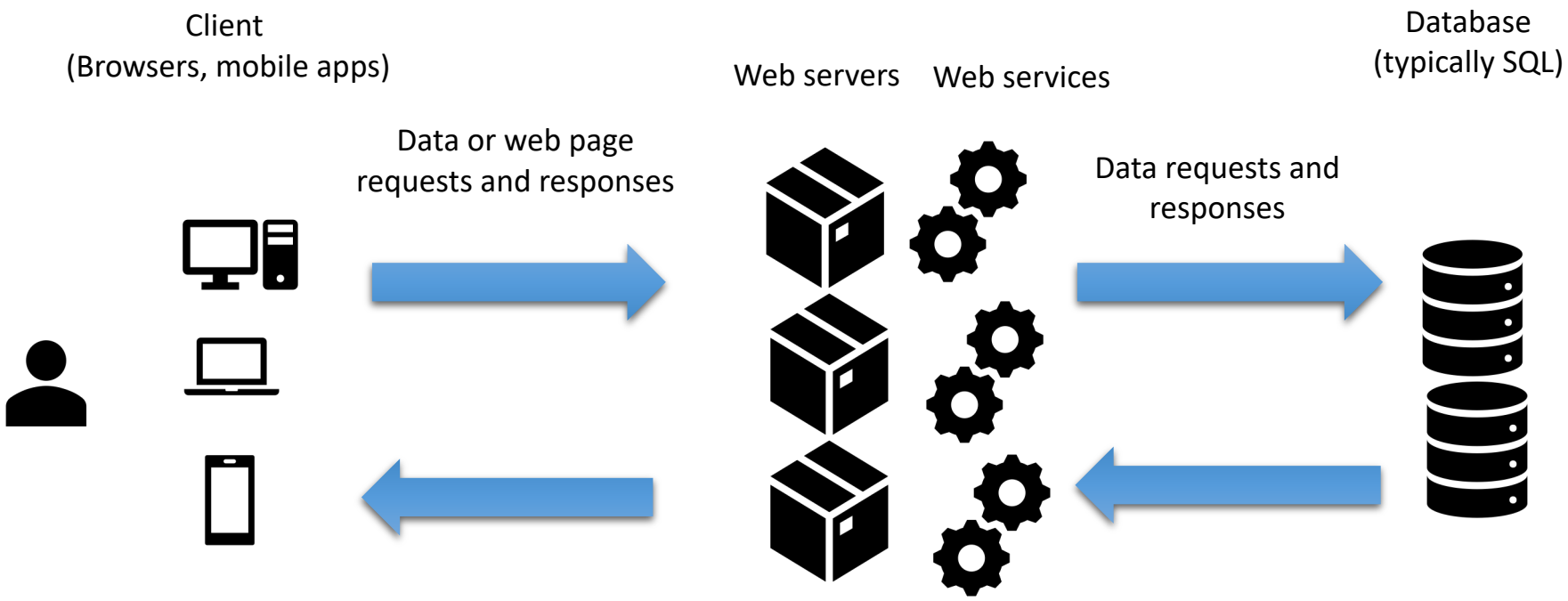
Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

DATA MINING

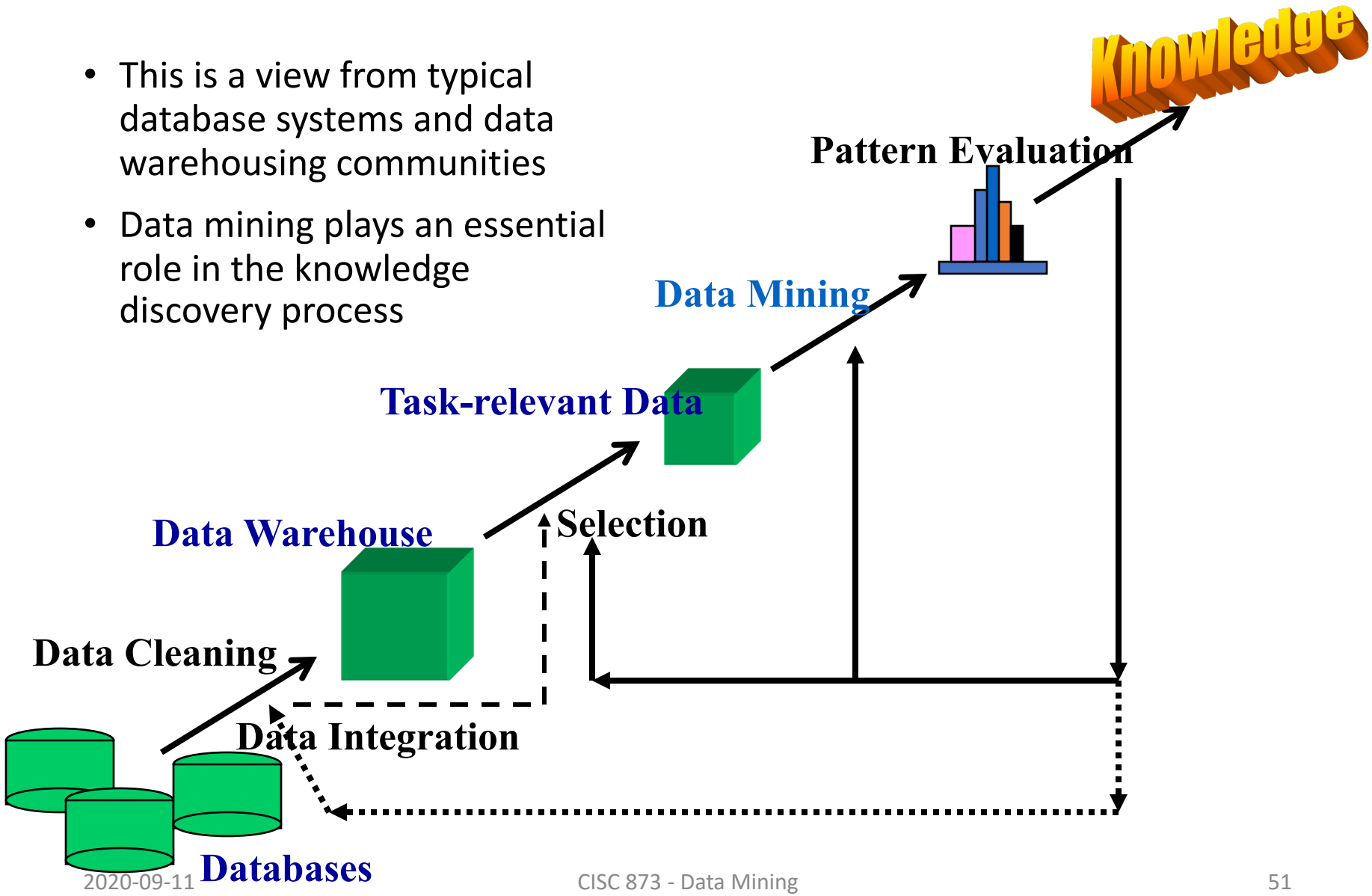
/CISC 873 - Steven Ding
/Week #1/Lecture 3
Data Mining Process & Infrastructure

3-Tier Architecture



Knowledge Discovery (KDD) Process

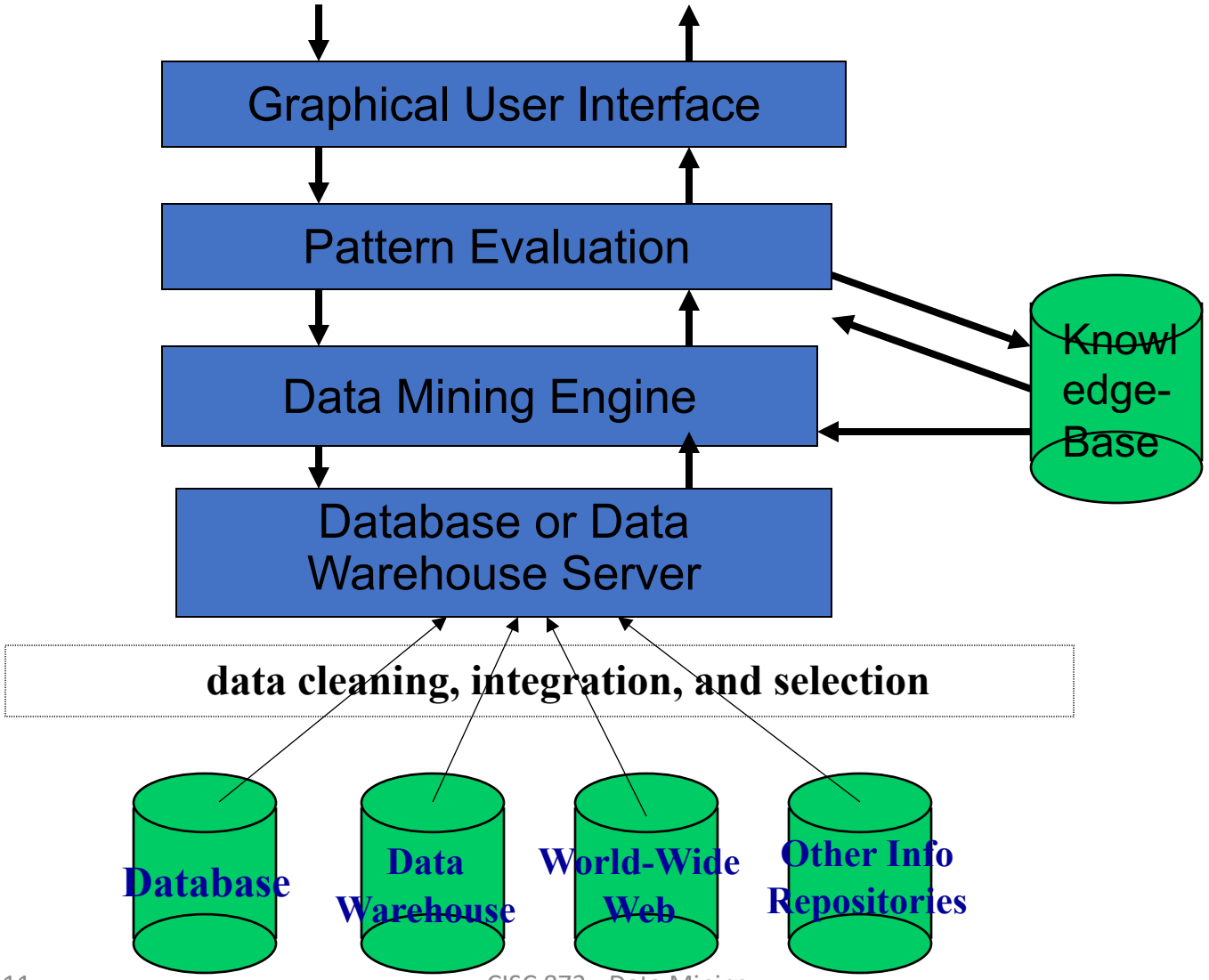
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Example: Transactions in supermarket

- Large-scale data mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

Architecture: Typical Data Mining System



What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - **A decision support database** that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of **constructing** and **using** data warehouses

Data Warehouse—Subject-Oriented

- Organized **around major subjects**, such as customer, product, sales
- Focusing on the **modeling** and **analysis** of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: **provide information from a historical perspective** (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains **an element of time**
  - But the key of operational data may or may not contain "time element"

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- *Operational update of data does not occur in the data warehouse environment*
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - initial loading of data and access of data

Data Warehouse vs. Operational DBMS

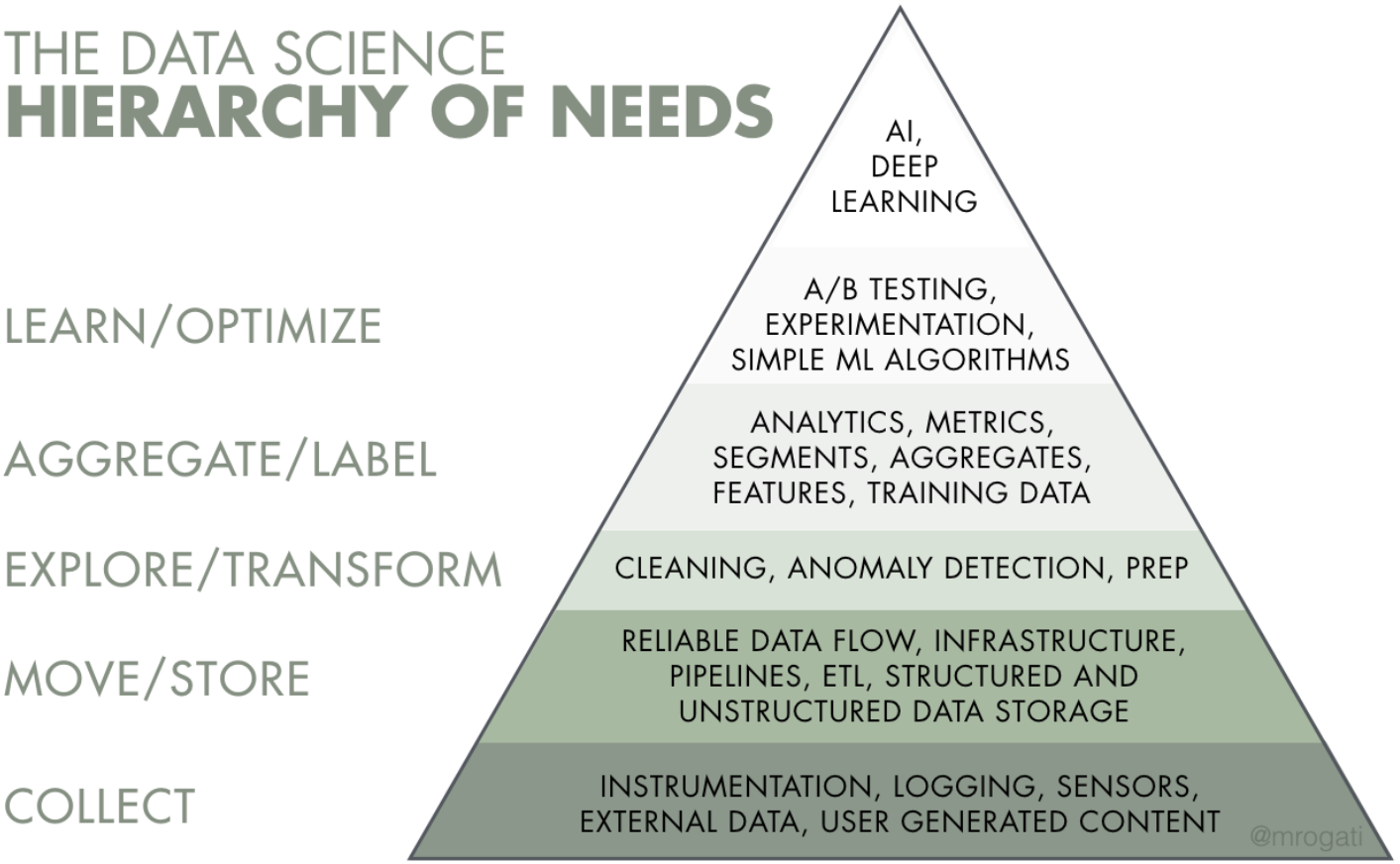
- OLTP (on-line transaction processing) ← Operational DBMS
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing) ← Data Warehouse
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: read and write vs. read-only but complex queries

OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support (DS) requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform analysis directly on relational databases



https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

DATA MINING

/CISC 873 - Steven Ding
/Week #1/Lecture 4
Security, Privacy, and Ethical Issues

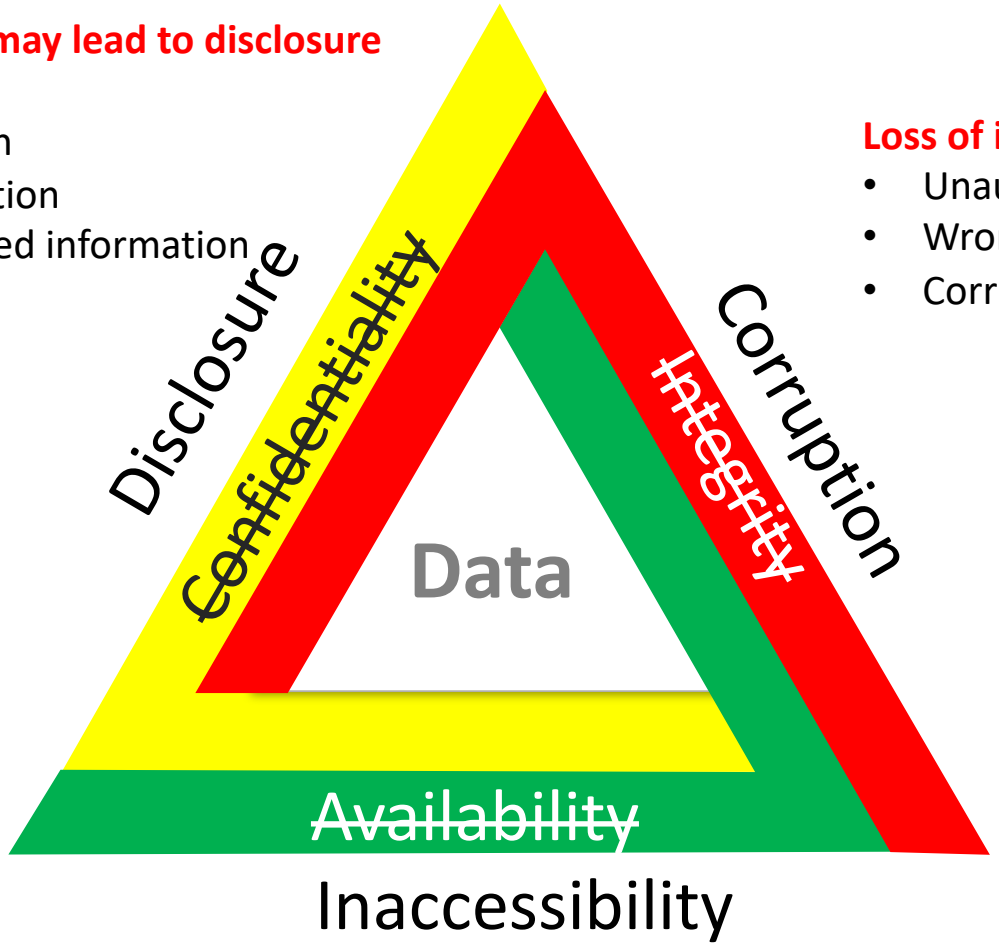
- **Marriott data breach**
- 500M victims
- $123 million fine

- **British Airways website breach**
- 380,000 victims
- $230 million fine

CIA Threats

Loss of confidentiality may lead to disclosure of:

- Personal information
- Proprietary information
- Government classified information



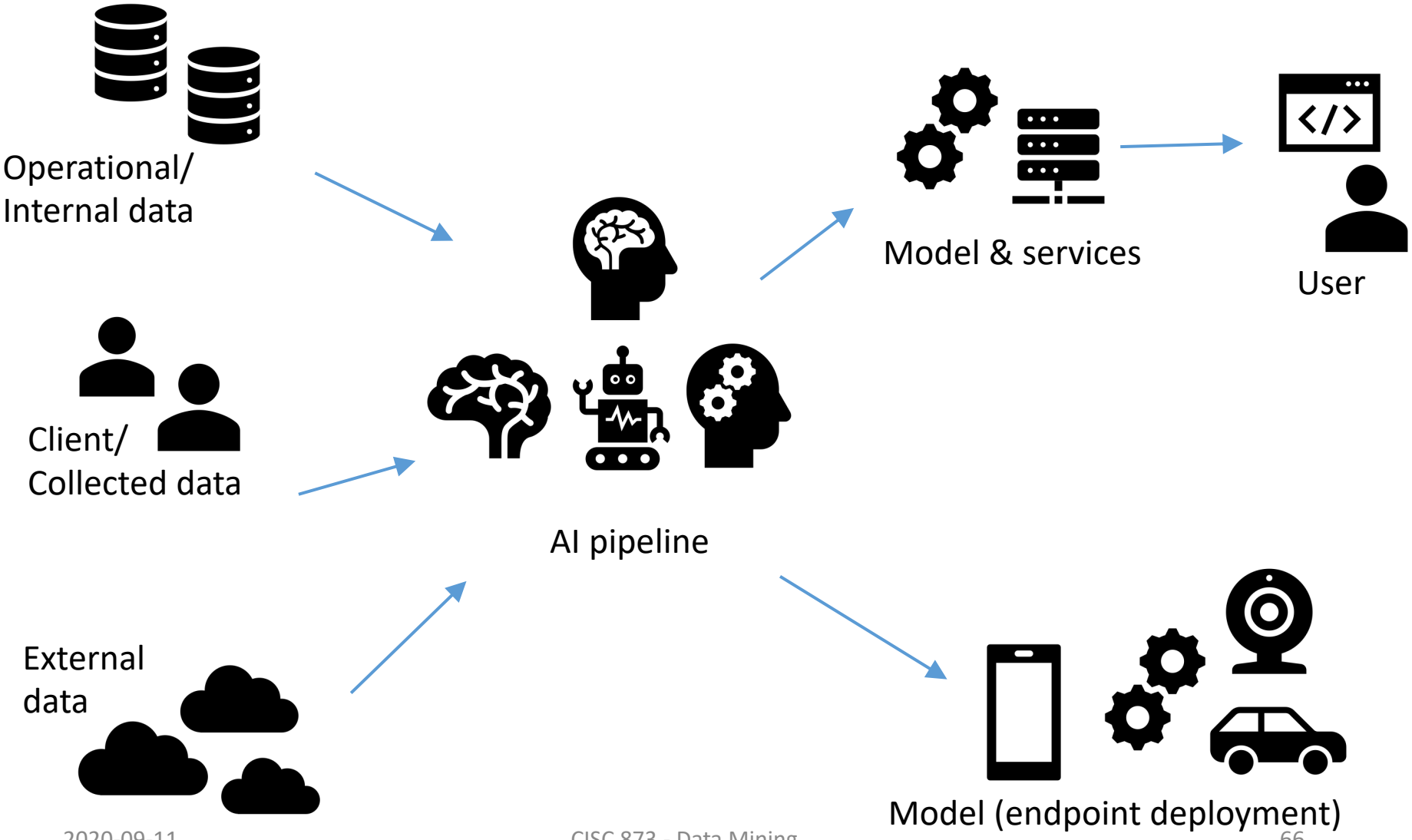
Loss of integrity may lead to:

- Unauthorized transactions
- Wrong execution of software
- Corruption of data

Loss of availability may lead to:

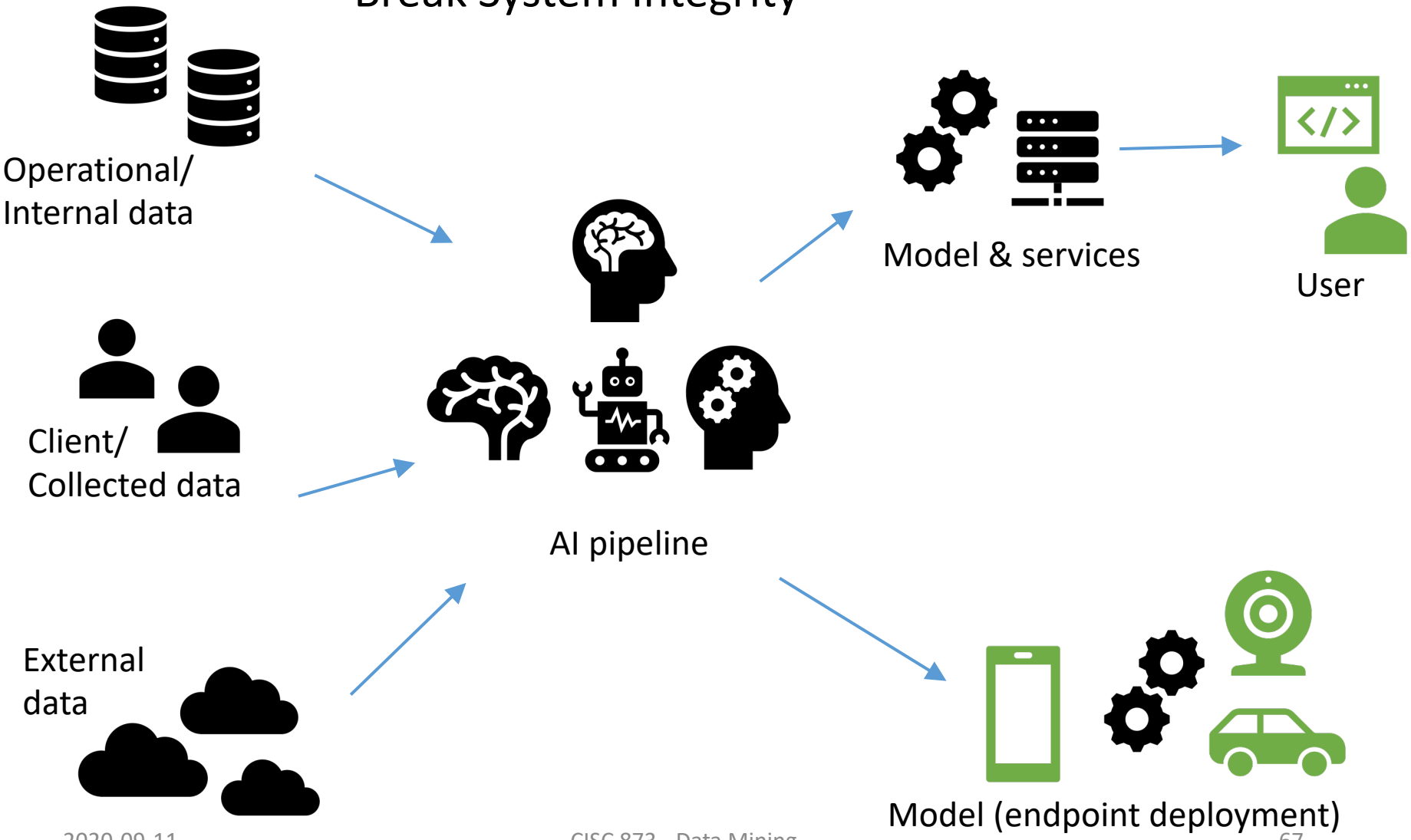
- Denial of Service
- Loss of Data

Data Flow


Operational/
Internal data
Client/
Collected data
External
data
AI pipeline
Model & services
User
Model (endpoint deployment)

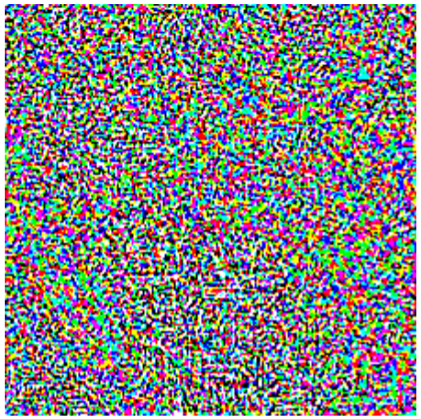
Adversarial Samples

- Break System Integrity



Adversarial Samples

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

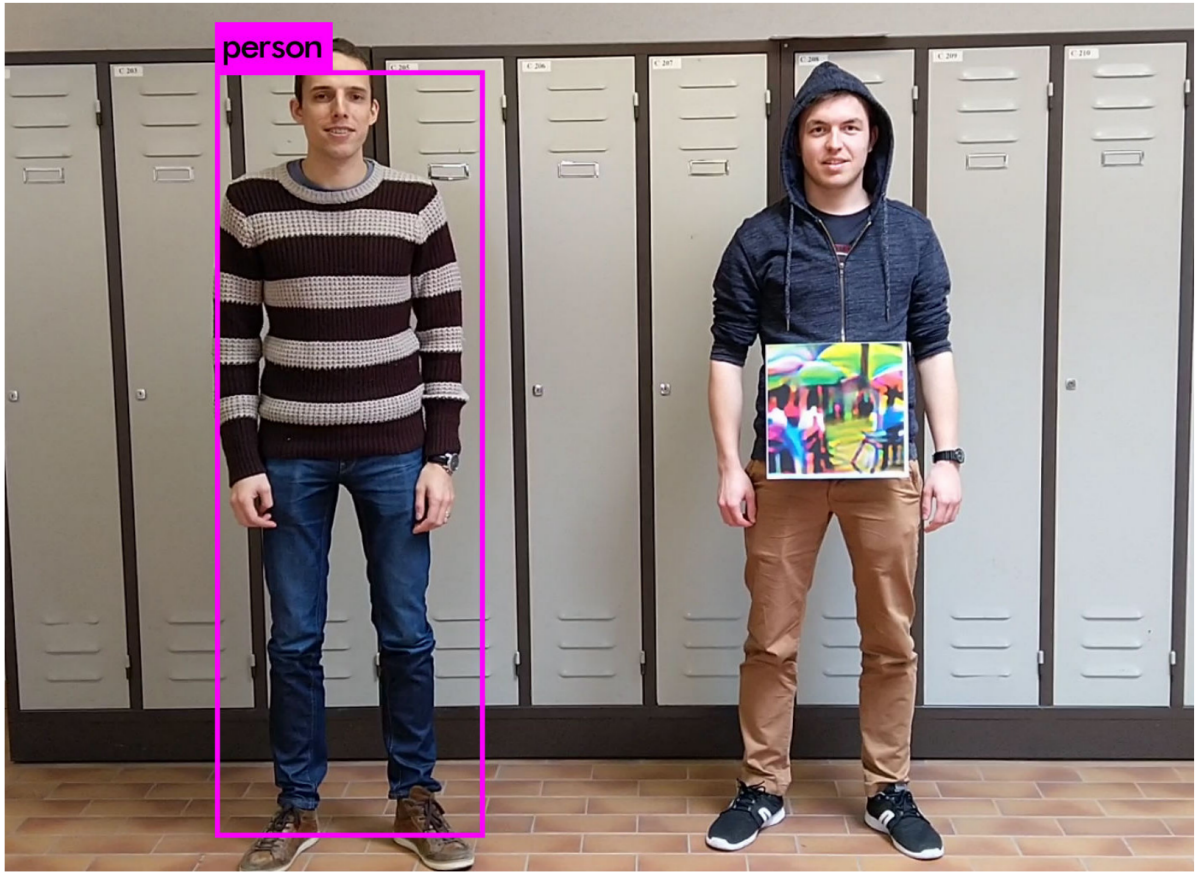
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

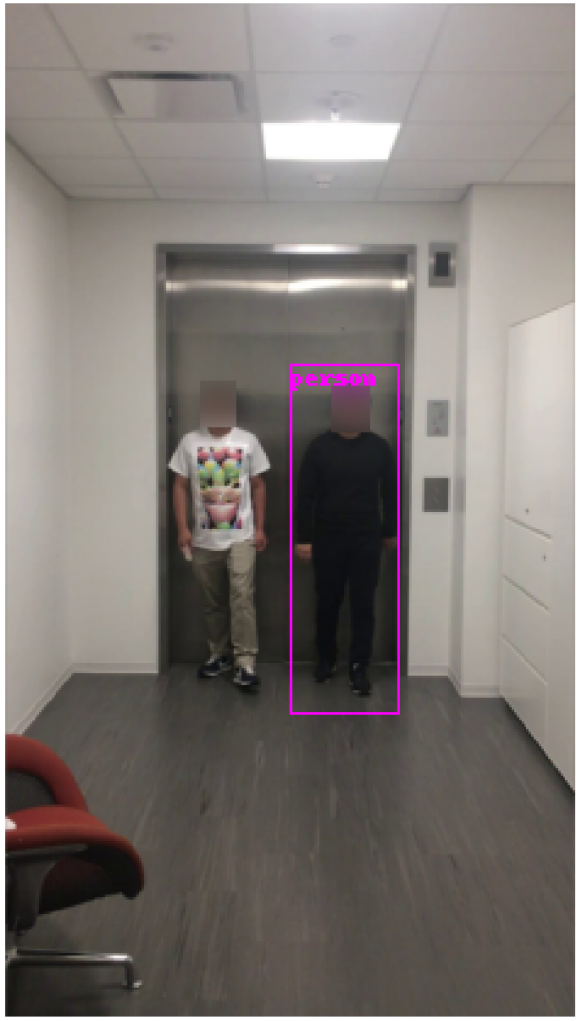
Source: https://arxiv.org/pdf/1412.6572.pdf

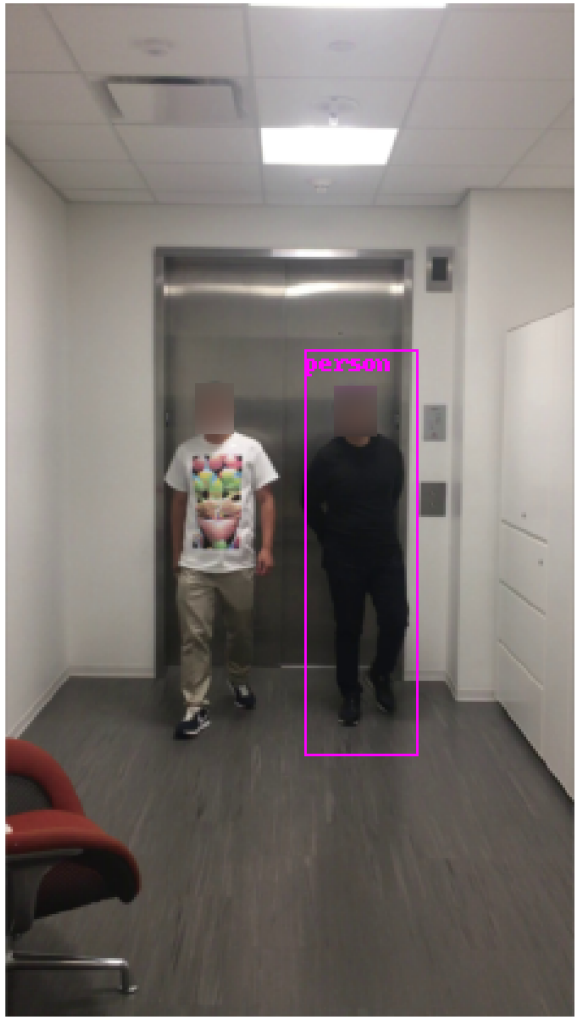
Adversarial Samples

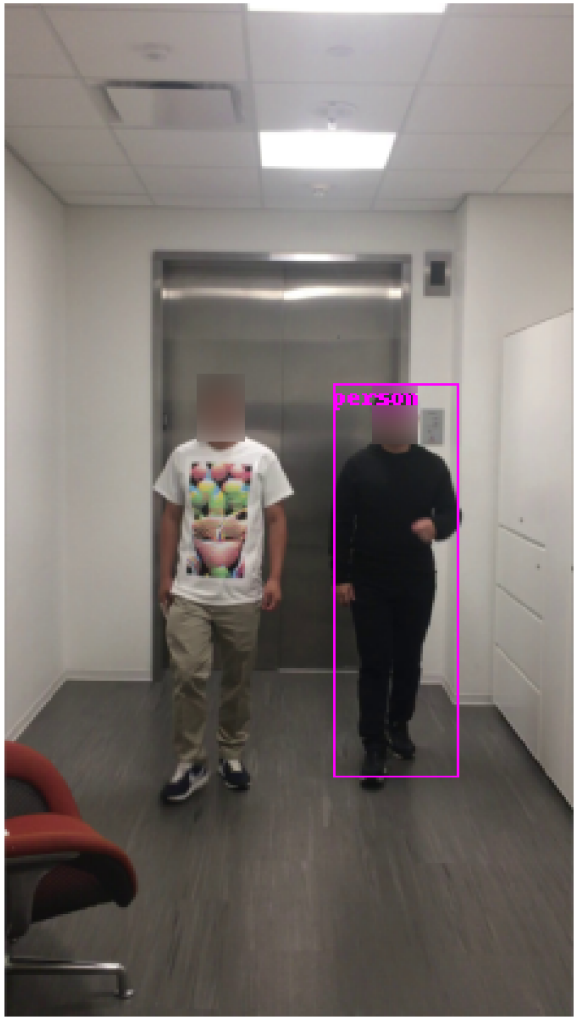


Source: https://arxiv.org/pdf/1904.08653.pdf

Adversarial T-Shirt



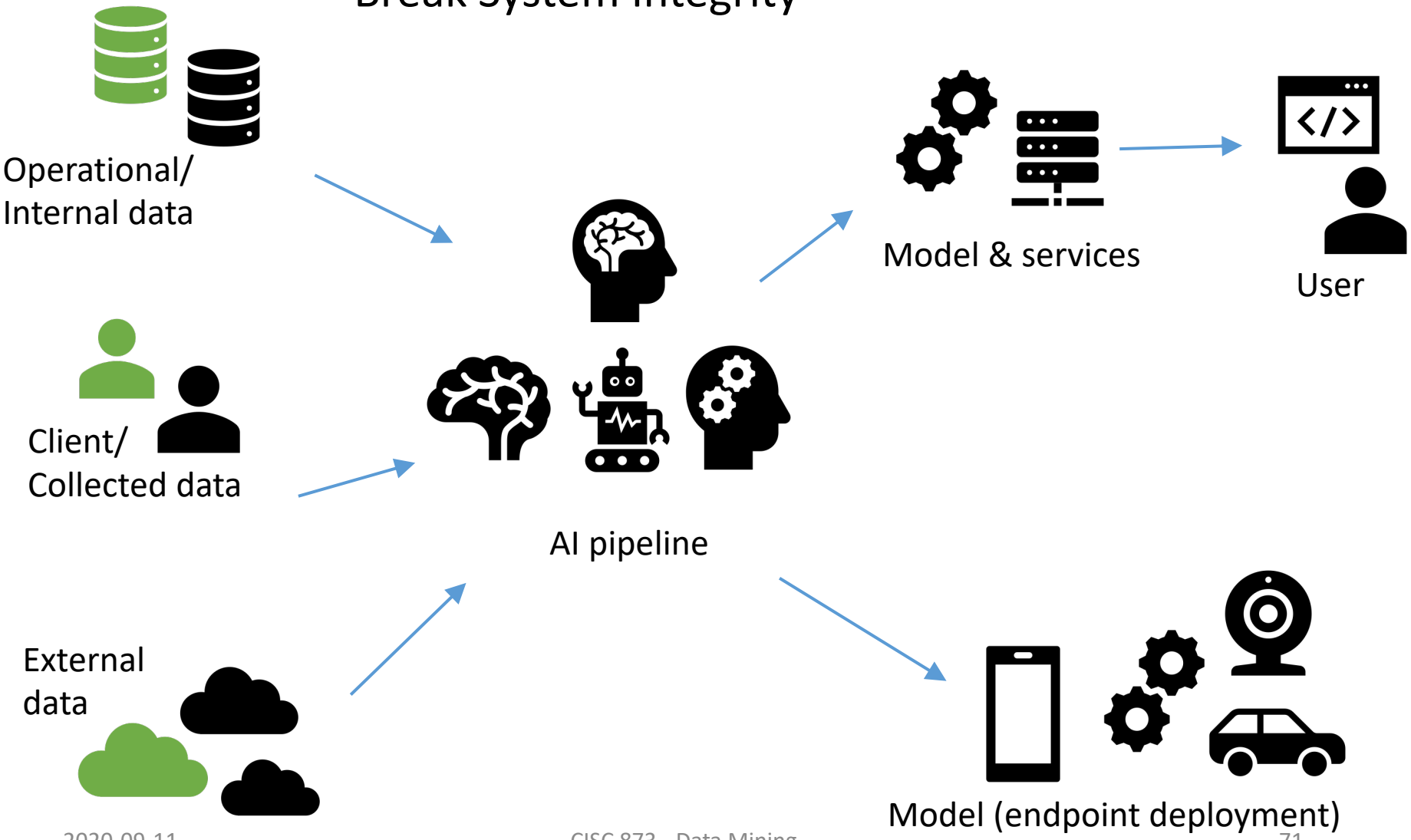




Source: https://arxiv.org/abs/1910.11099

Backdoor Attack

- Break System Integrity



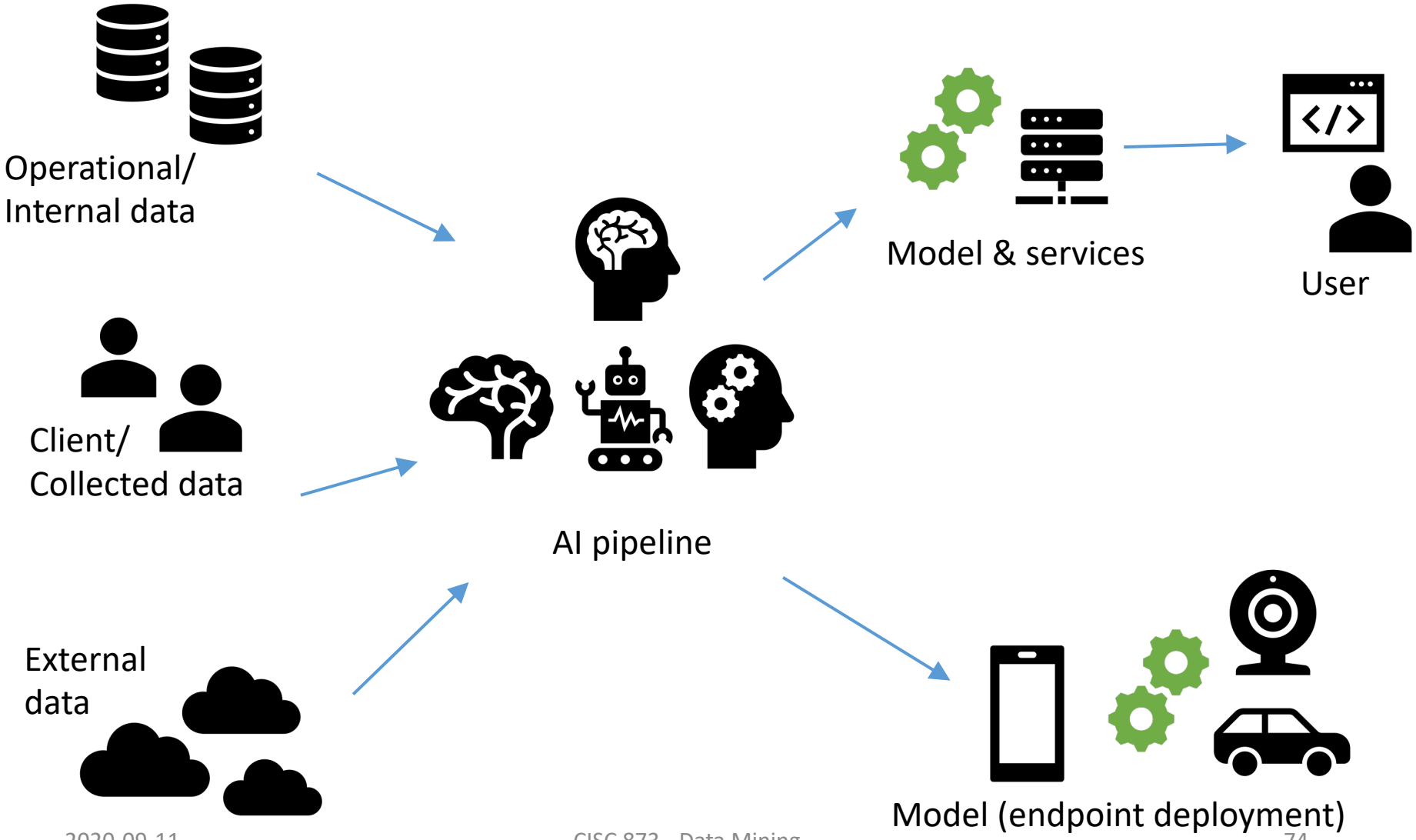
Backdoor Attack

Source: https://arxiv.org/pdf/1708.06733.pdf

Backdoor Attack

Source: https://arxiv.org/pdf/1708.06733.pdf

Information Leak

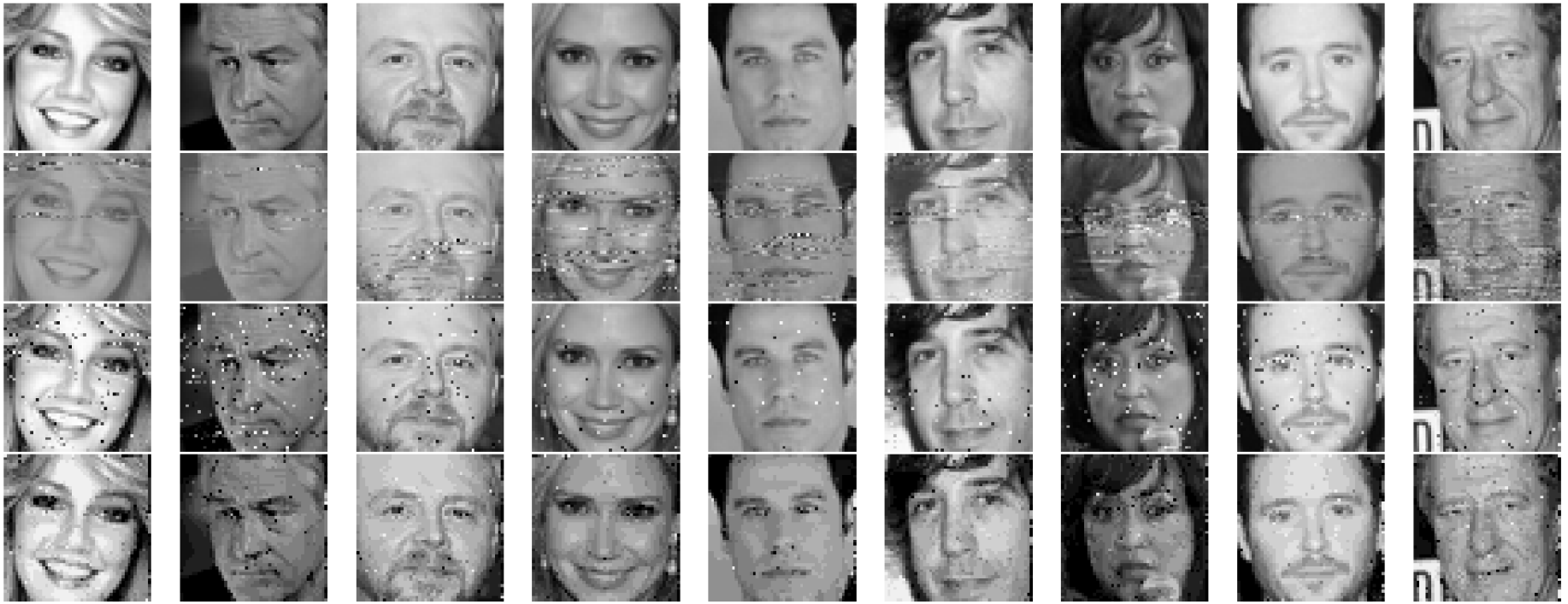


Information Leak

- Model is solely evaluated based on testing performance metric before release/deployment
  - Like Shingai mentioned: objective function

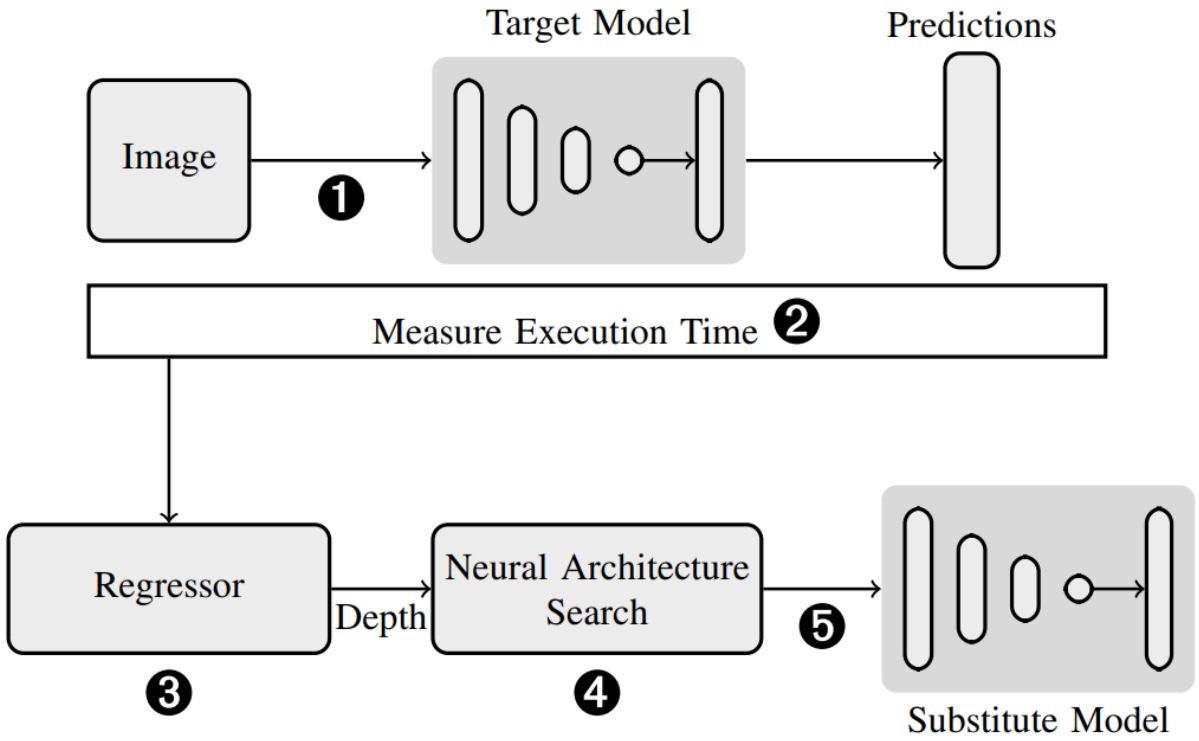
- But what else did the model captures in the data? (can be recovered by attackers?)

Information Leak
- Reconstruct original training data



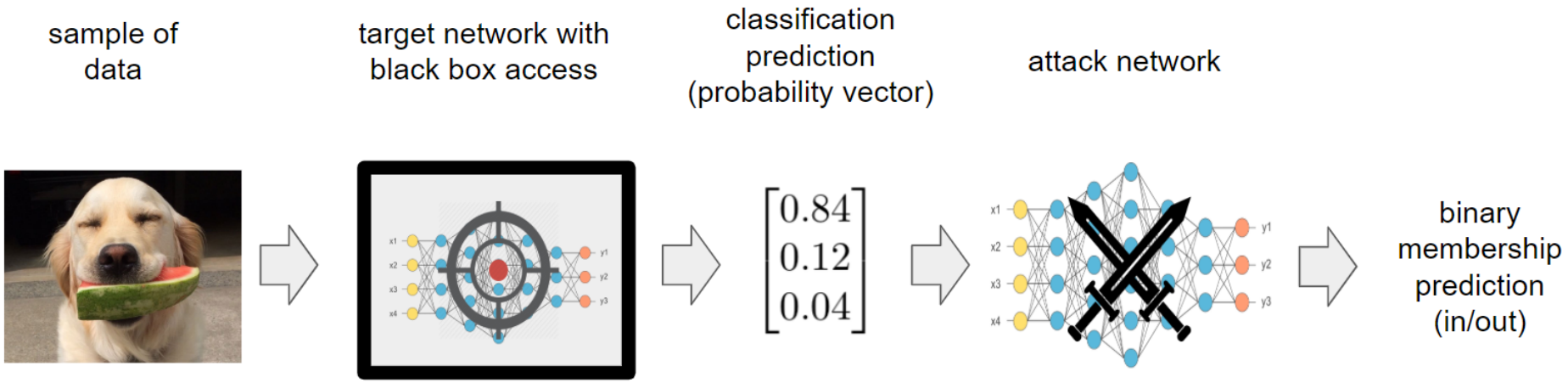
Source: https://www.cs.cornell.edu/~shmat/shmat_ccs17.pdf

Information Leak
- Stealing Neural Networks



Source: https://arxiv.org/pdf/1812.11720.pdf

Information Leak
- Membership Inference – Privacy Breach



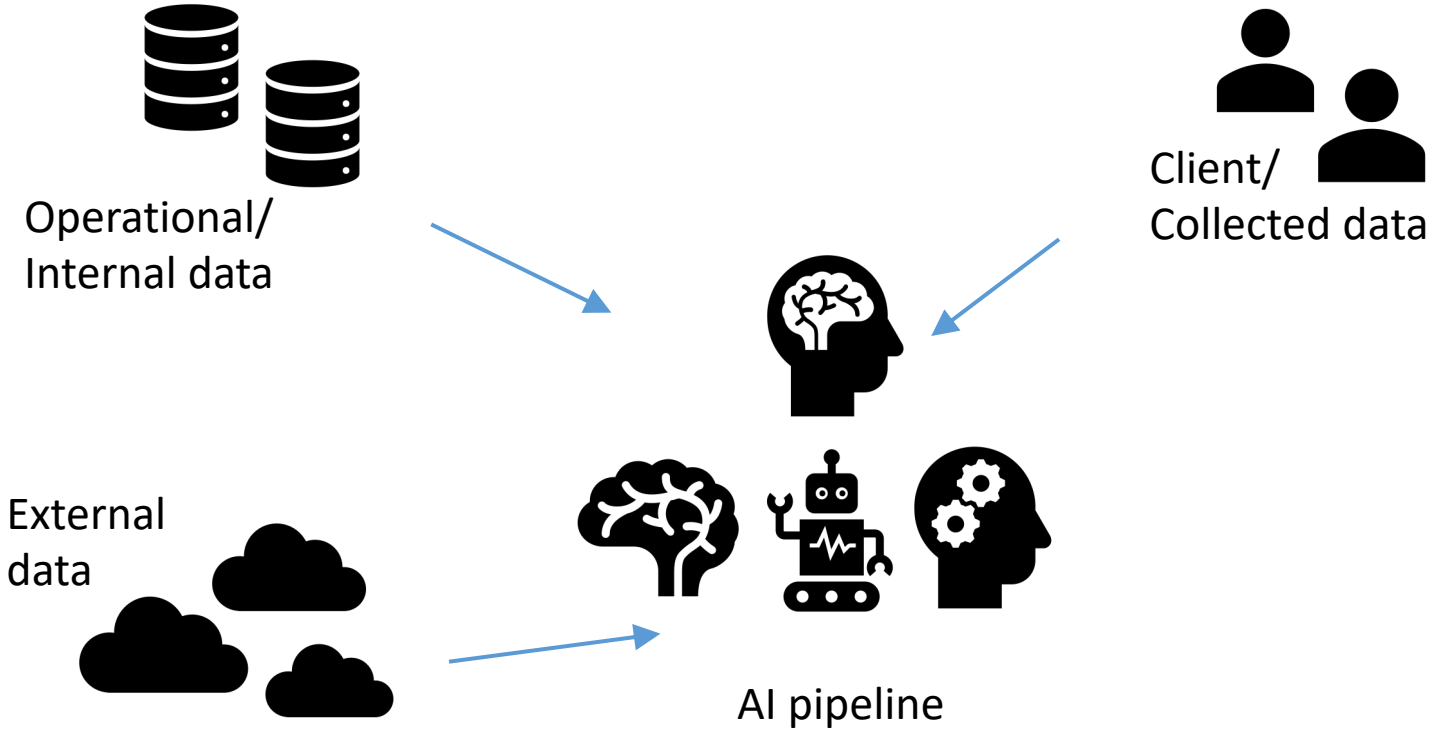
… results for the Texas hospital discharge dataset (over 70% accuracy) indicate that membership inference can present a risk to health-care datasets if these datasets are used to train machine learning models and access to the resulting models is open to the public.

AI under Information Security

- Increased risk of data breach and fine
- Increased uncertainty
- Difficult to evaluate change in AI
- Difficult to verify against compliance
- Responsibility, accountability, liability
- Should be part of the risk management framework

Ethical Issues

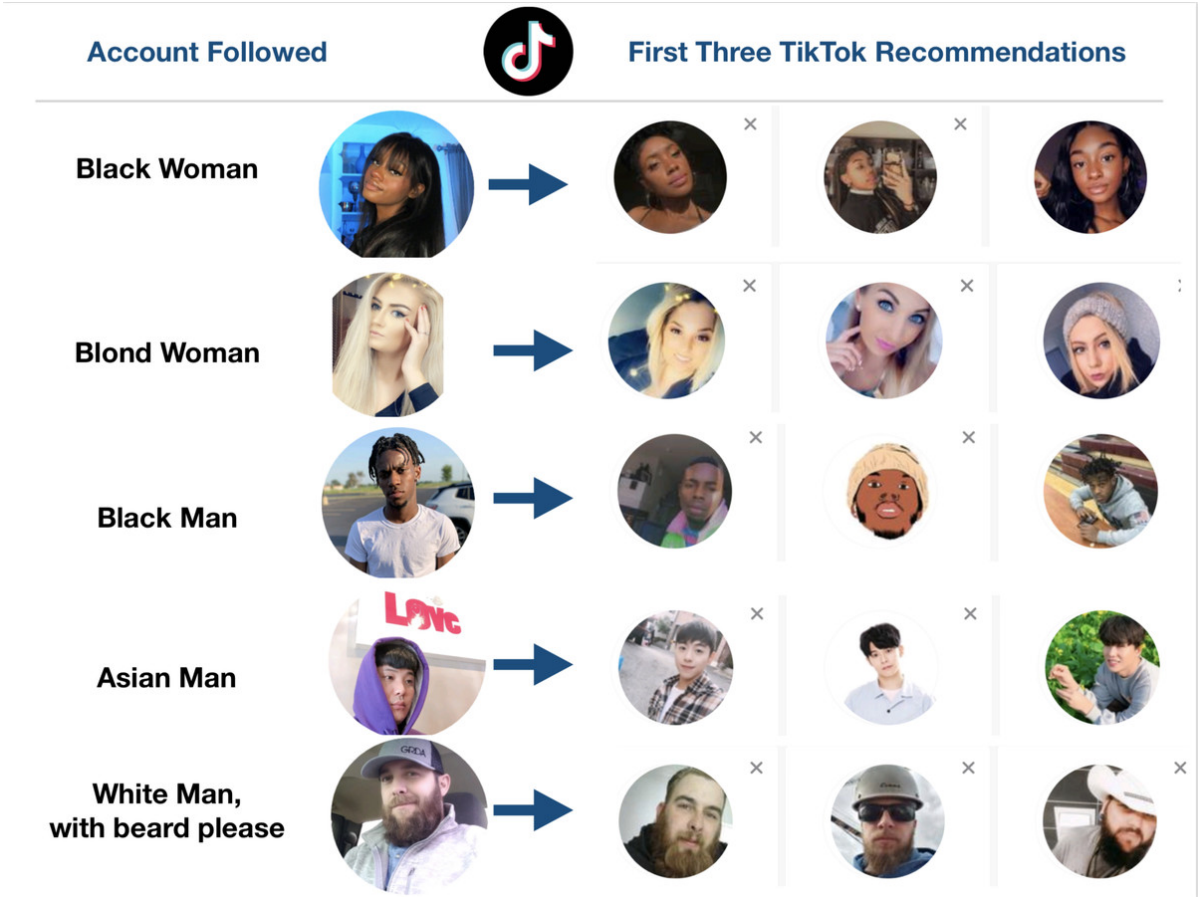
- AI systems designers choose the features, metrics, and analytics structures of the models that enable data mining. Thus, **data-driven** technologies, such as Artificial Intelligence, can potentially replicate the **preconceptions and biases** of their designer.



https://interestingengineering.com/ethics-of-ai-benefits-and-risks-of-artificial-intelligence-systems

Filter Bubbles.

- Recommended follows tended to physically resemble the initial account followed, though they were not always the exact same accounts that appeared in Faddoul's results.



https://www.vox.com/recode/2020/2/25/21152585/tiktok-recommendations-profile-look-alike

Purpose? Ethical use?

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Purpose? Ethical use?

- Google will stop building custom AI for Big Oil and Gas

https://electrek.co/2020/05/22/google-stop-building-custom-artificial-intelligence-big-oil-and-gas/

NEXT WEEK