IRA A. FULTON SCHOOLS OF engineering

school of computing, informatics, & decision systems engineering

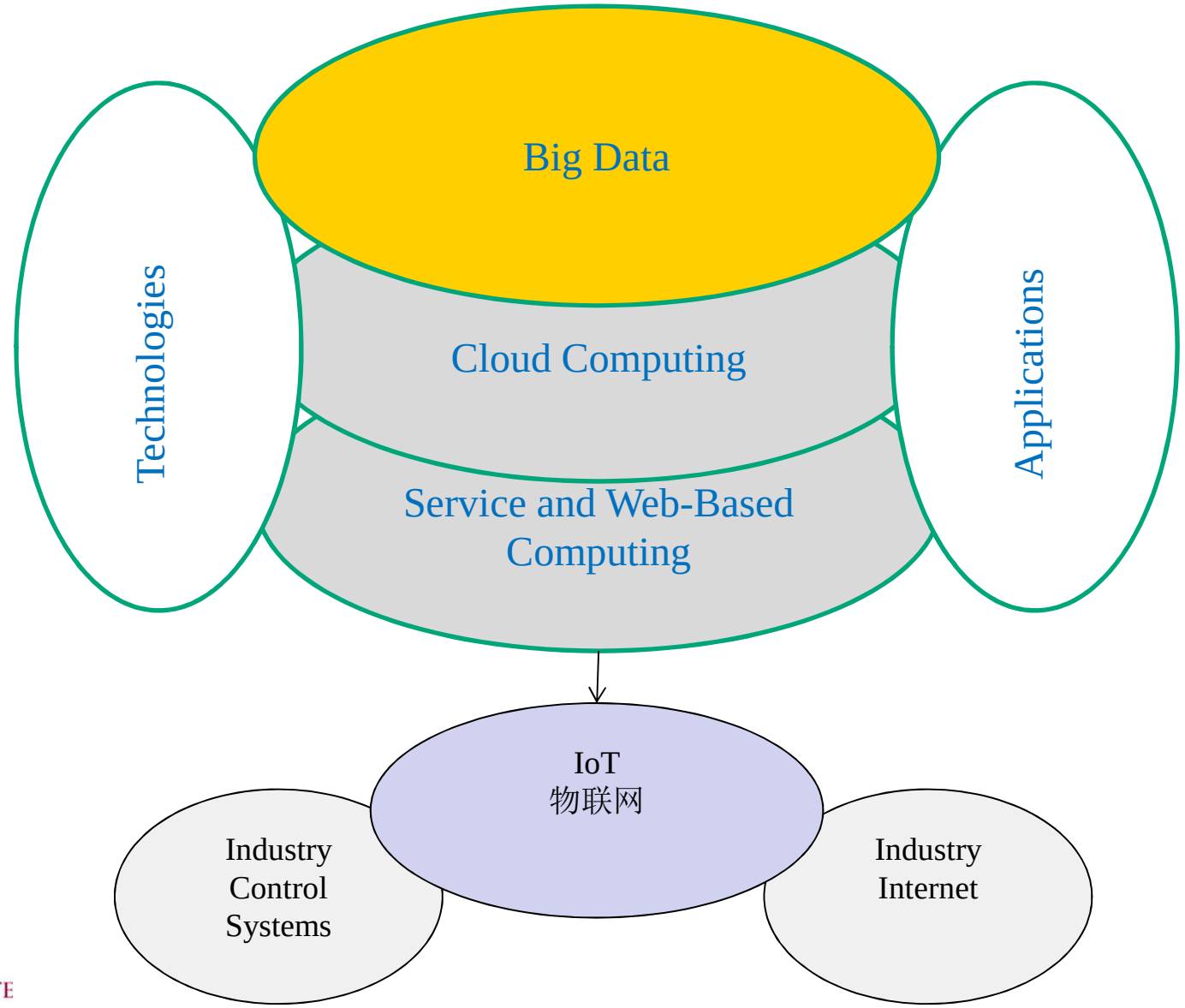
Introduction to Engineering Using Robotics Experiments

Lecture 17

Big Data

Yinong Chen

Big Data


Big Data
Technologies
Applications
Cloud Computing
Service and Web-Based Computing
IoT
物联网
Industry Control Systems
Industry Internet

Lecture Outline

- What is Big Data?
  - Characteristics
  - Challenges
- Big Data Collection:
  - Sources
  - Mechanisms
- Big Data Applications:
  - Road Traffic Enforcement
  - Network Traffic Analysis
  - MOOC and Big Data Application in Education
  - Other Applications

What is Big Data?

- Big data is the term for a collection of data sets so big and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications [http://en.wikipedia.org/wiki/Big_data].
- Sources of big data are mainly from
  - Human through social networking and
  - Devices, particularly, IoT
- The challenges in big data processing lie not only in the volume, but also in the types of data and the velocity of new data are generated, etc.

Three Types of Data

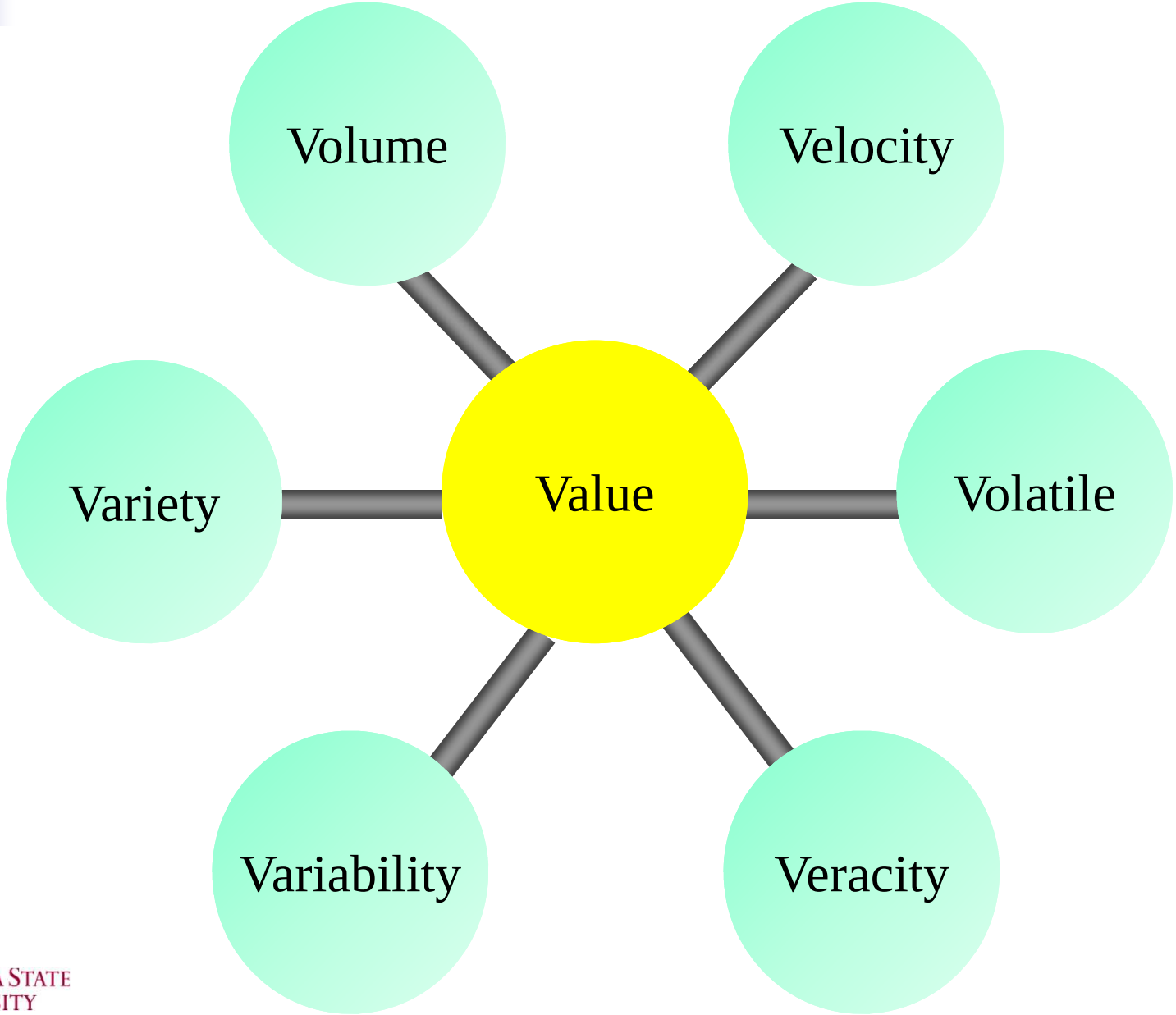
Three types of data stored in computer systems:

- Structured Data: Tables of data in traditional relational databases. SQL is the typical query language
- Semi-Structured Data: XML files stored in XML databases, as we discussed in Chapter 4 and this chapter.
- Unstructured Data: Data that are not structured or semi-structured, mainly streamed data like voice, photos, and video files.

Challenges of Big Data System

- Type: Poly-Structured Data: including structured, semi-structured, and unstructured data. unstructured data are the essential part.
- Volume: Huge and rapidly growing.
- Processing: Adequate capacity and algorithms to extract information in adequate time, sometimes, in real time.
- Compromise: in consistency, accuracy and response time, and in long-term and short-term storage (Volatile).

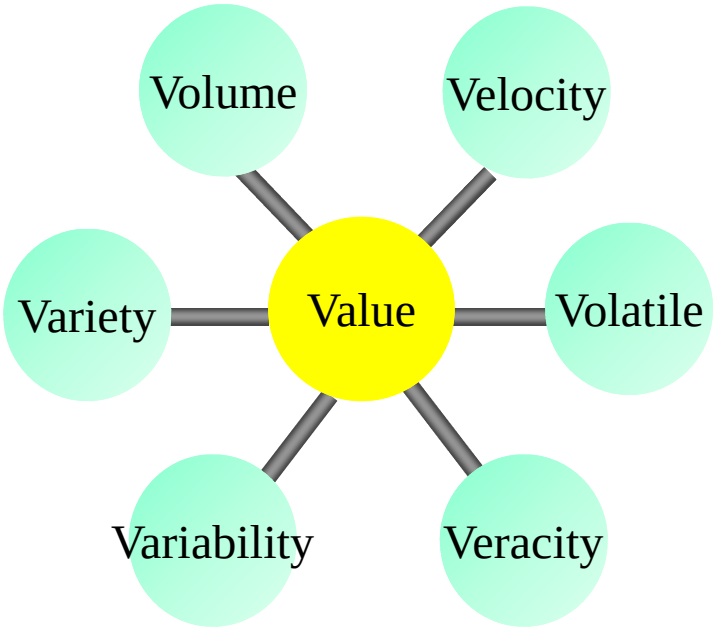
Aspects of Big Data Systems: Many Vs



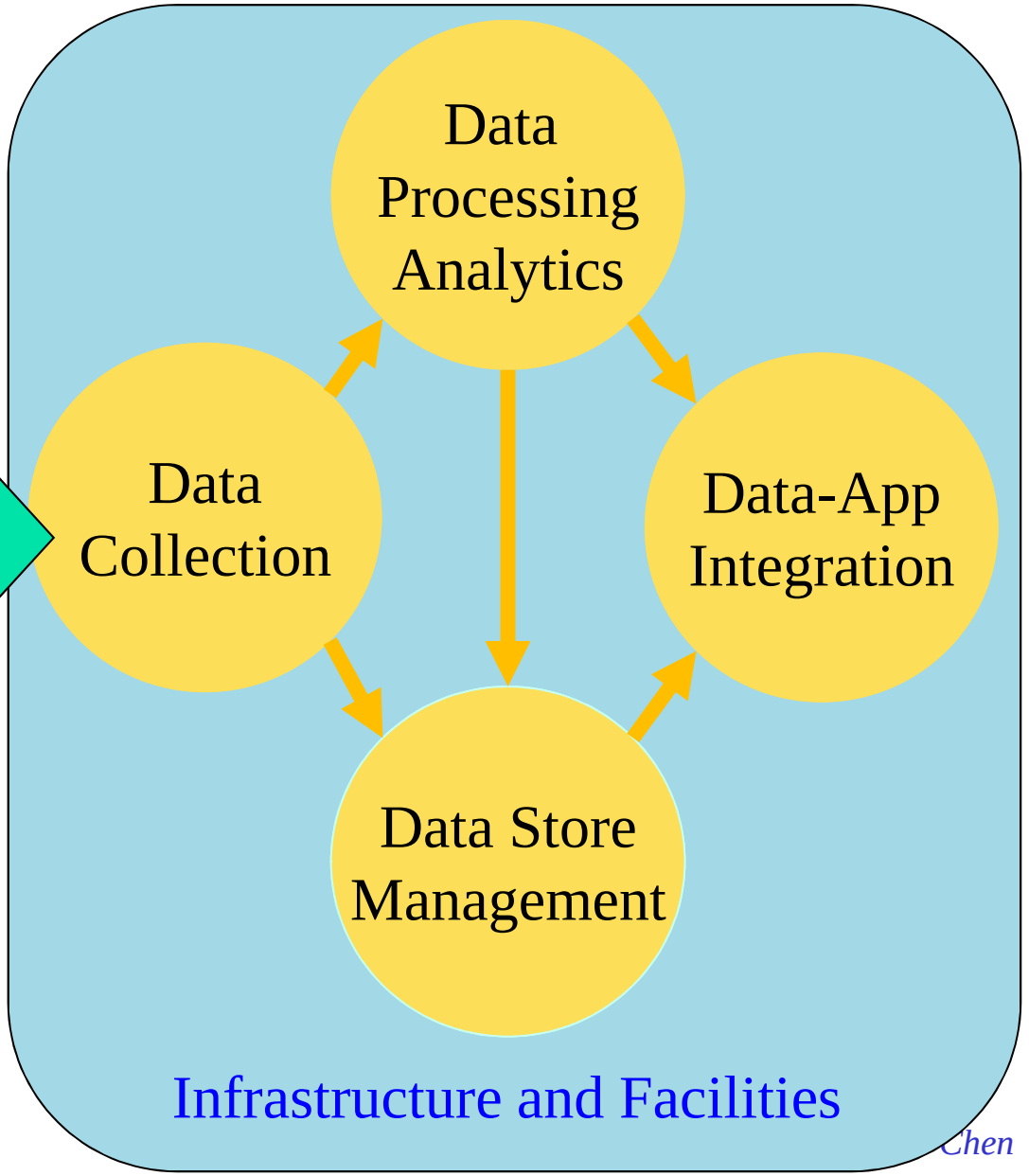
Aspects of Big Data Systems: Many Vs

- Value: It is the next big thing after Internet (Communication) and Cloud Computing (Computation). Big data is about Data.
- Volume: A moving target from petabyte (10^{15} bytes), exabyte (10^{18}), zetabyte (10^{21}), to more
- Velocity: Real-time data require real-time responses.
- Variety: Data from different sources with different semantics are integrated into different applications.
- Variability in data structures: poly-structured data
- Veracity: Accuracy issue: Noise elimination and fault tolerance
- Volatile: Not all data will be stored, and some will be permanently deleted, big data processing systems are required to selectively store and organize the data to maximize its value.

From Big Data Concepts to Domains of Study



Concepts



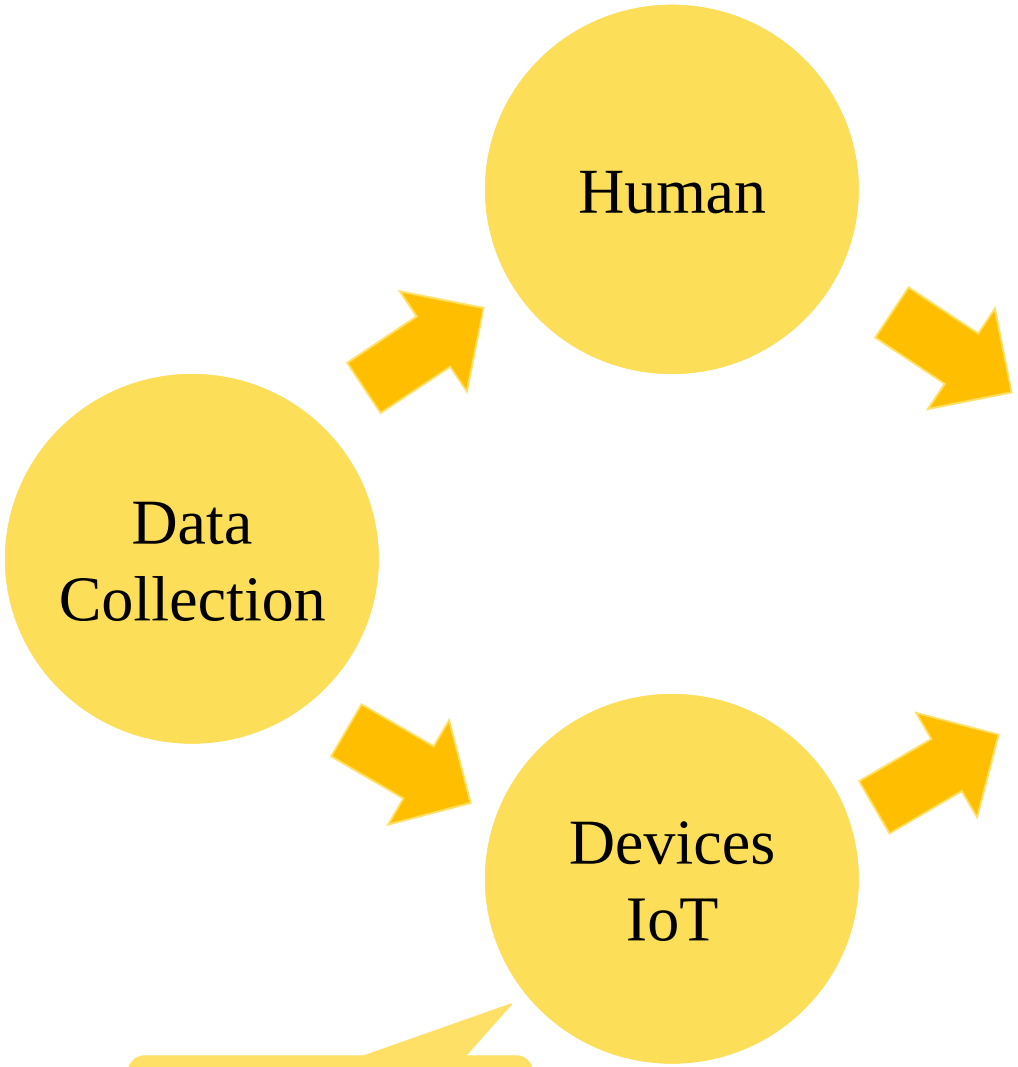
Key Domains of Study

- Data collection: Human and Devices
- Infrastructure:
  - Cloud computing and parallel computing
  - Storage and database facilities
  - Communication
- Management: Organizing data and facilities
- Processing and Analytic techniques: specifically developed for processing and analyzing big data in specific applications
- Big Data Applications

Lecture Outline

- What is Big Data?
  - Characteristics
  - Challenges
- Big Data Collection:
  - Sources
  - Mechanisms
- Big Data Applications:
  - Road Traffic Enforcement
  - Network Traffic Analysis
  - MOOC and Big Data Application in Education
  - Other Applications

Sources and Collection Mechanisms



- Social networks, Facebook, Twitter, …
- Web sites from different organizations
- Databases and data warehouses
- Publications: books, journals, …
- Archives and legacy documents
- Computer generated data
- Photos and images
- Surveillance camera steam video/audio
- Telescope and astronomical data
- Sensor networks outputs

2013 talks on IoT

Lecture Outline

- What is Big Data?
  - Characteristics
  - Challenges
- Big Data Collection:
  - Sources
  - Mechanisms
- Big Data Applications:
  - Road Traffic Enforcement
  - Banking Applications
  - Network Traffic Analysis
  - MOOC and Big Data Application in Education
  - Other Applications

Road Traffic Rule Enforcement

- Cameras are installed not any in every intersections, but also between intersections
- Video taping all road behaviors
- Illegal driving detection
  - Red light running
  - Speedy
  - Illegal U-turn
  - Double yellow lines crossing
- Ticket issuing
- Driver's license suspension

Finding a Person via Facial Recognition

- Facial recognition algorithms can extract relative positions, sizes, shapes of the eyes, nose, ears, cheekbones, and jaw and hash these numbers into a single number
- The video cameras on the road film and analyze everyone's facial characteristics.
- To find a person, enter the person's facial characteristics
- Compare them with data coming from every camera on the road.

Big Data Applications in Banking

- A bank typically service millions of customers;
- Big data capabilities provides banks the ability to understand their clients with more detail and can deliver targeted personalized offers faster.
- It enables higher offer and cross sell acceptance rates, which improve customer profitability, satisfaction and retention.
- While all banks currently perform analysis of customer data to obtain insight to improve customer offers, many are unable to properly process the big amount of data to optimize their offers.

IBM Big Data Platform Fiserv for Banking

http://www-01.ibm.com/software/data/bigdata/industry-banking.html

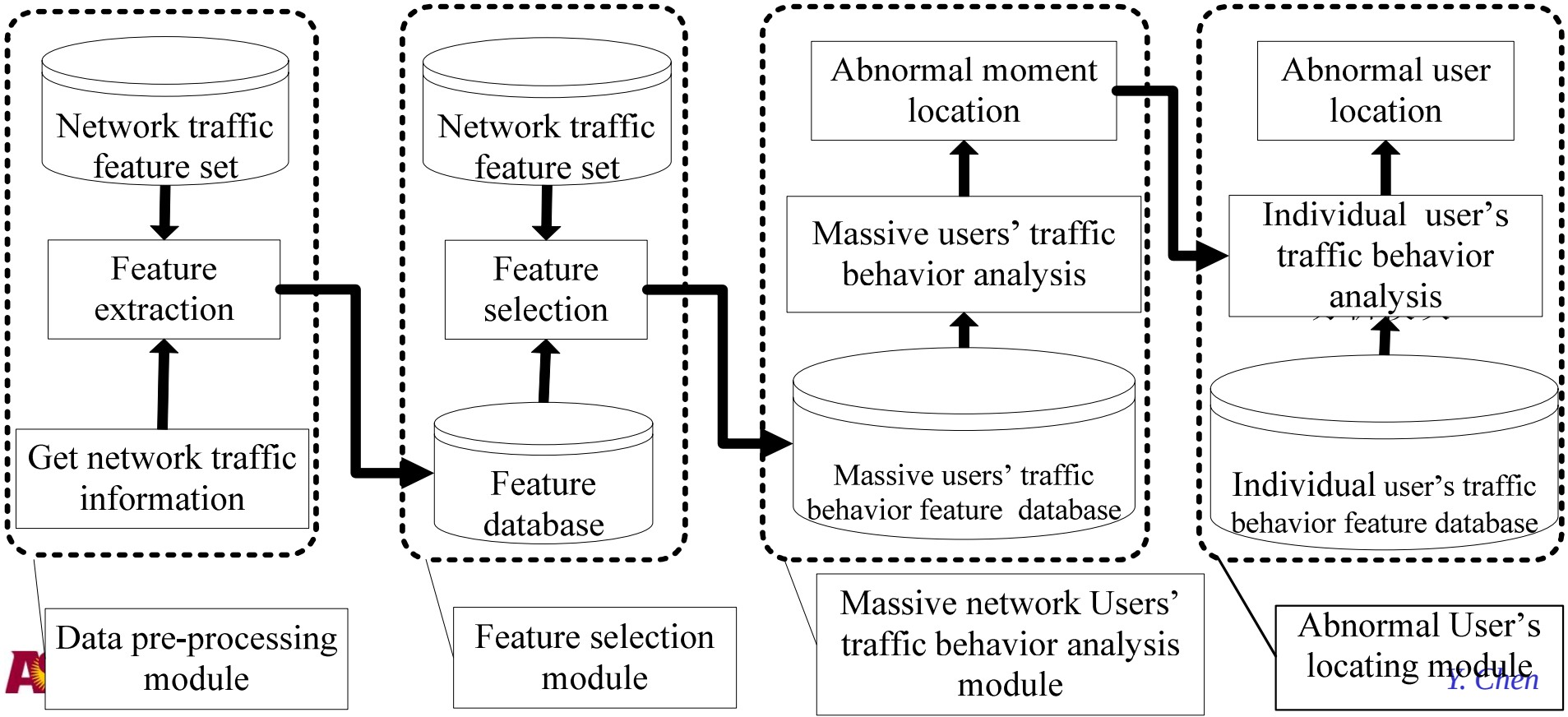
Fiserv enables banks to deliver improved actionable insight during customer interactions with the contact center. Used by more than 16,000 financial institutions. Benefits include:

- More accurately predict what happens next, understand customers, and increase customer adoption;
- Improve processing performance with increased capabilities;
- Higher revenues through improved cross sell
- Cost efficiency through higher productivity
- Higher rate of customer satisfaction/retention
- Lower campaign and infrastructure costs

Network Traffic Monitoring Example

(Text Chapter 11 for more detail)

- Traffic Analysis for Intrusion Detection & Prevention
- Define features to characterize the traffic behaviors
- Obtained features from data fields in the IP packets.
- Save normal & abnormal patterns and compare with traffic flow.



Model Differentiating Normal and Abnormal Behaviors

- Definition: The relative deviation distance (D_i) of feature f_i is defined

$$D_i = \frac{|f_i - E(f_i)|}{\sigma_i}$$

- Rule: Feature selection rule between normal traffic (N) and abnormal (A) traffic behavior:

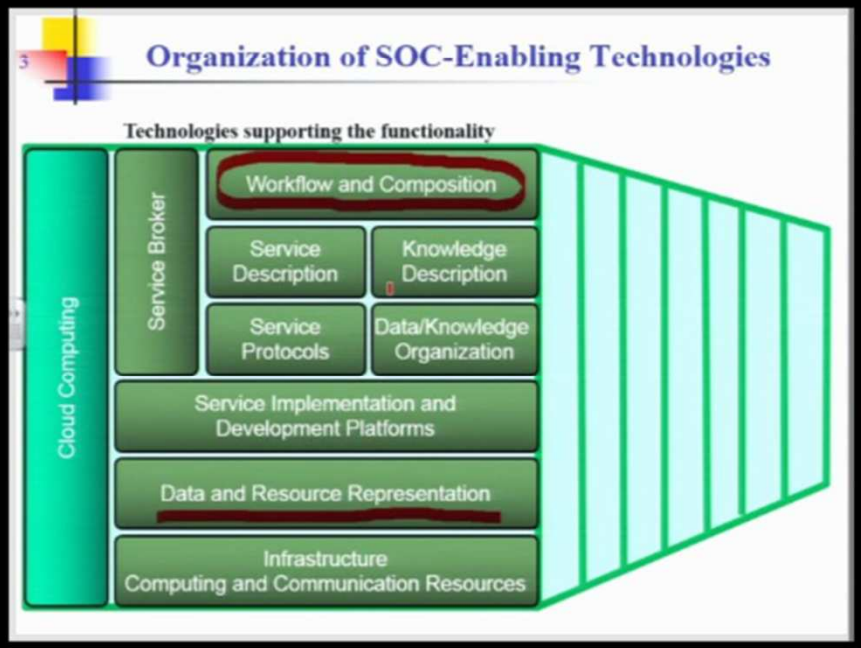
$$\frac{\frac{(D_i)_A}{(D_i)_N}}{\sum_{i=1}^{110} \frac{(D_i)_A}{(D_i)_N}} \geq threshold$$

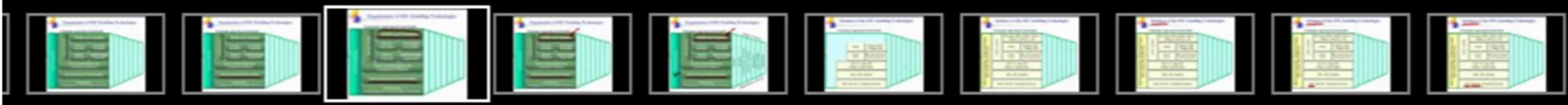
MOOC

- MOOC (Massively Open Online Course) is reshaping the education system.
- Video: http://mslgoee.asu.edu/Mediasite/Play/aae59213d64e4005aeb8d2e6d6290a561d?catalog=bac218b3-1903-4a98-9b83-93856a9b74f0
- Course ontology: Store the concepts required by the course
- Learning ontology and big data analysis: Collect statistically significant numbers of user interactions, semantic analysis of learner postings, which can be used for creating feedback to learners:
  - Discussion board is no longer adequate if there are thousands of posting in each thread. We need recommendation.
  - We need an ontology system for knowledge organizing and processing

Online Lecture

ASU Ira A. Fulton Schools of Engineering
ARIZONA STATE UNIVERSITY





Many Other Applications

- Healthcare: Link all patients' data, doctors' decisions, and outcomes, require an ontology.
- National Security: The U.S. Utah Data Center is a big data system for Comprehensive National Cybersecurity Initiative. The mission is classified.
- Tax: IRS collects all data from all organizations and individuals to detect any tax evasion.
- Credit Scores: The US companies collect all the finance-related activities of every person with an SSN.
- Retailers: Not only online retailers like Amazon and eBay, but also traditional retailers like Walmart and Target, have million transactions/hour to process.
- Real Estate: Collect GPS signals to help home buyers to determine their drive times to work throughout at different times.