

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Data Management

CT051-3-M

Topic 2 – Organizational Data Preparation

Topic and Structure of the Lesson

At the end of the lecture you should be able to

- Importance of Data Management
- List and define various sources of data
- Explain the fundamental differences between databases, data warehouses and datasets
- Explain some of the ethical dilemmas associated with data mining and outline possible solutions



Data Realities...

Living with Cancer
The changing science

Beyond Baghdad: Where The Enemy Has Its Own Surge

The Sopranos' Last Song: What Exit Will Tony Take?

TIME

SPECIAL DOUBLE ISSUE

The Global Warming Survival Guide

51 Things You Can Do to Make a Difference

2ND-QTR SIZZLE
PROFITS AT 900 COMPANIES (P. 74)

PAYING FOR COLLEGE
BEWARE OF THOSE HIGH 529 FEES (P. 96)

TERRORISM
WHAT COMPANIES STILL NEED TO DO (P. 26)

BusinessWeek

August 16, 2004

GLOBAL WARMING

Why Business Is Taking It So Seriously

BY JOHN CAREY (P. 60)

APRIL 6, 2004

TIME

SPECIAL REPORT GLOBAL WARMING

BE WORRIED. BE VERY WORRIED.

Climate change isn't some vague future problem—it's already damaging the planet at an alarming pace. Here's how it affects you, your kids and their kids as well

EARTH AT THE TIPPING POINT
HOW IT THREATENS YOUR HEALTH

INDIA CAN HELP D—OR DESTROY IT
CRUSADERS

Adapted for A NEW GENERATION
from the New York Times Bestseller

an inconvenient truth

the crisis of global warming

AL GORE

January 1997

nature

Environmental Science & Technology

Eocene global warming

Hydrothermal vents prompt methane release

Malina passes
Flows accentuate asymmetric variations

Photonic crystals
Perfected the device

Gastropod giant tortoise
Siphonous muds made rock foam

August 26, 2002

TIME

SPECIAL REPORT

HOW TO SAVE THE EARTH

The hot and wild weather is a sign of things to come. But fresh ideas and new technology can cool us down and make this a **GREEN CENTURY**

TIME

Where's the Beach?

America's Vanishing Coastline

APRIL 8, 2001

TIME

GLOBAL WARMING

Climbing temperatures. Melting glaciers. Rising seas. All over the earth we're feeling the heat. Why isn't Washington?

OCTOBER 19, 1997

TIME

The Heat Is On

How the Earth's Climate Is Changing

Why the Ozone Hole Is Growing

Dr. Bush's Rx for Health Care

TIME

VANISHING OZONE

THE DANGER MOVES CLOSER TO HOME

THE BIG DRY

GLOBAL WARMING CARTOONS - 2007-2100

NOVEMBER 28, 2005

TIME

New Orleans Blues

It's worse than you think. Three months after Katrina, the city still suffers

BY CATHY BOOTH THOMAS

Data deluge

Data is collected from sensors, sensor networks, remote sensing, observations, and more - - this calls for increased attention to data management and stewardship



Photo courtesy of
<http://www.futurlec.com>



Photo courtesy of
<http://modis.gsfc.nasa.gov/>



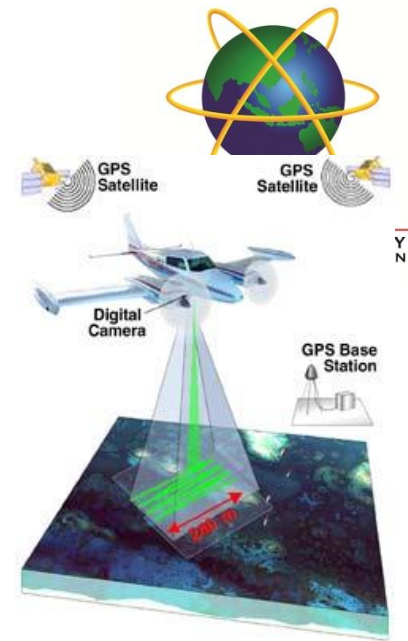
CC image by CIMMYT on Flickr



Image collected by Viv Hutchinson



Photo courtesy of www.carboafrika.net



CC image by tajai on Flickr

NY

The World of Data Around Us



A • P • U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

How Big is the Digital Universe?

Using the IDC / EMC Study of the Topic

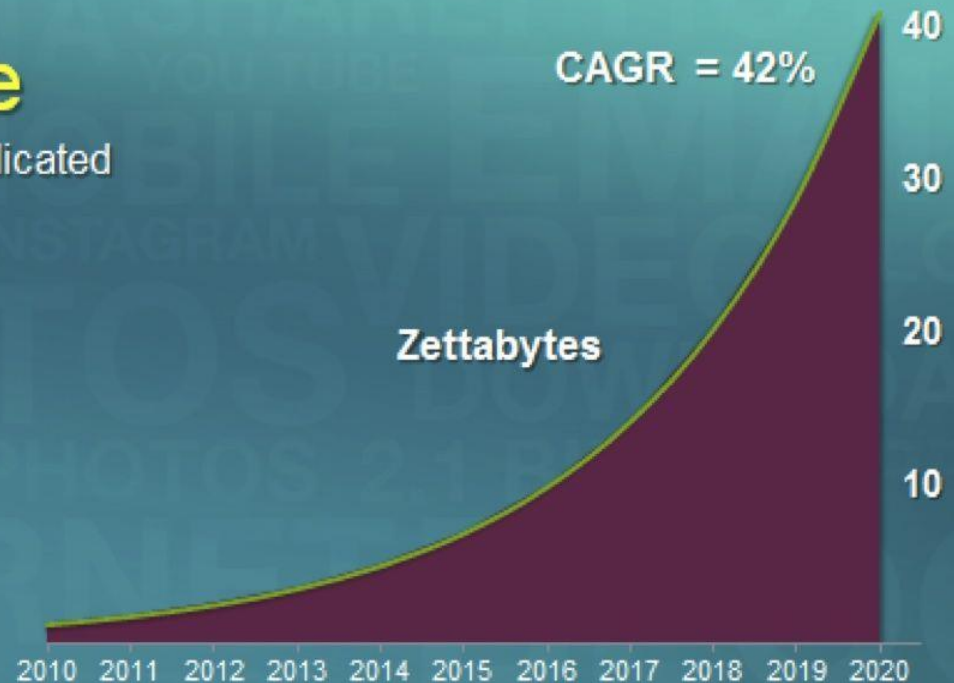
The Digital Universe

The measure of all digital data created, replicated
and consumed in a single year

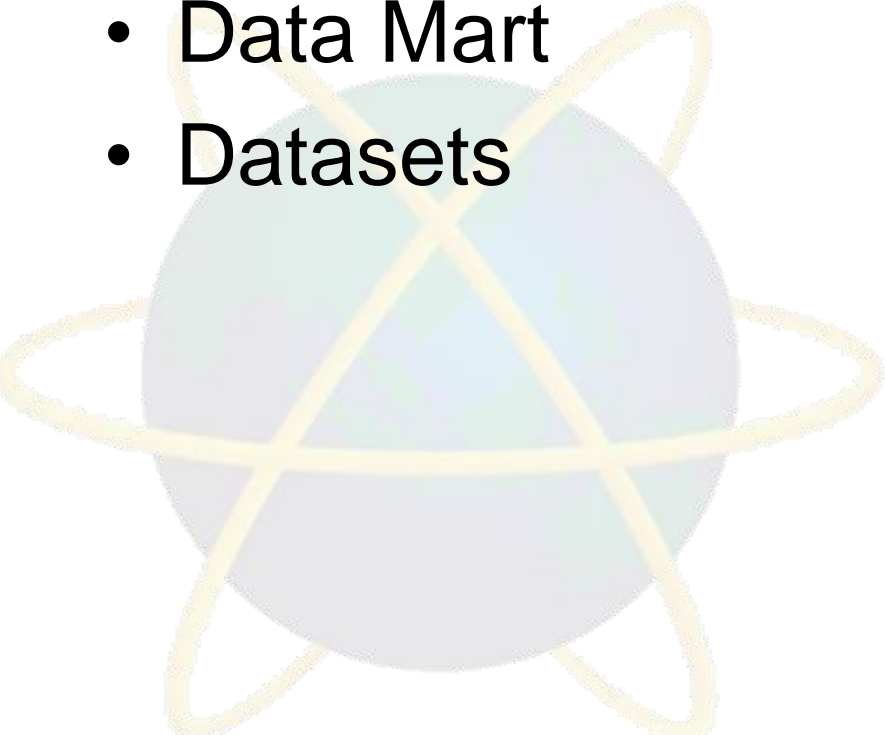
=

40 Zettabytes

in 2020



Data Storage

- Database
 - Data Warehouse
 - Data Mart
 - Datasets
- 

Data Storage: Databases

	A	B	C	D
1	3989.408	3989.408	140.4029	2654.278
2	140.4029	4125.044	4125.044	1335.467
3	2654.278	1335.467	2789.76	2789.76
4	5777.168	1788.068	5912.553	3123.153
5	2050.529	6039.689	1915.155	4704.363
6	1435.265	2554.287	1571.295	1219.56
7	4006.104	7994.156	3872.258	6659.535
8	671.2763	3318.277	807.9208	1983.314
9	2622.699	1367.091	2758.56	43.64889
10	8364.031	12353.06	8229.223	11018.06

Data arranged in columns and rows.

Tuples or Records or Rows

Fields, Variables or Attributes



Data Storage: Data Warehouse

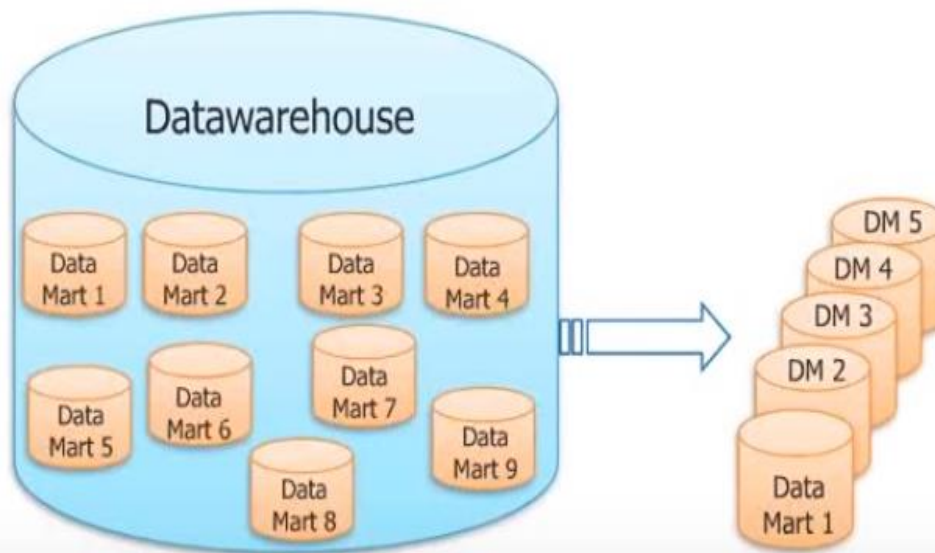


A . P . U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION



Data Storage: Data Mart

- Data mart is a smaller version of the Datawarehouse
- Data marts deal with a single subject
- Data marts are focused on one area. Hence they draw data from a limited number of sources
- Time taken to build the data marts is very low compared to the time taken to build a Datawarehouse



Data Storage: Dataset

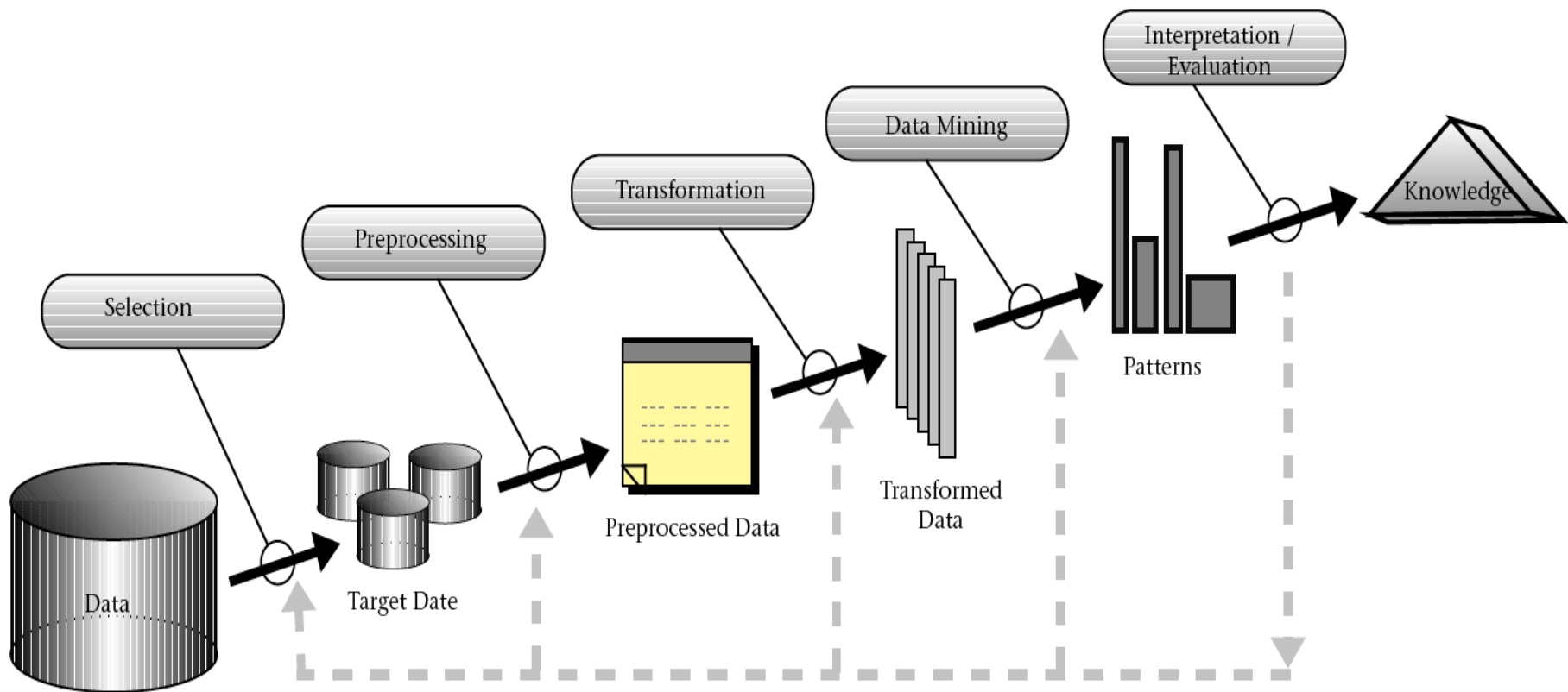
<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)



Knowledge Discovery in Database (KDD Process)



What is (NOT) Data Mining?



A . P . U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

□ What is not Data Mining?

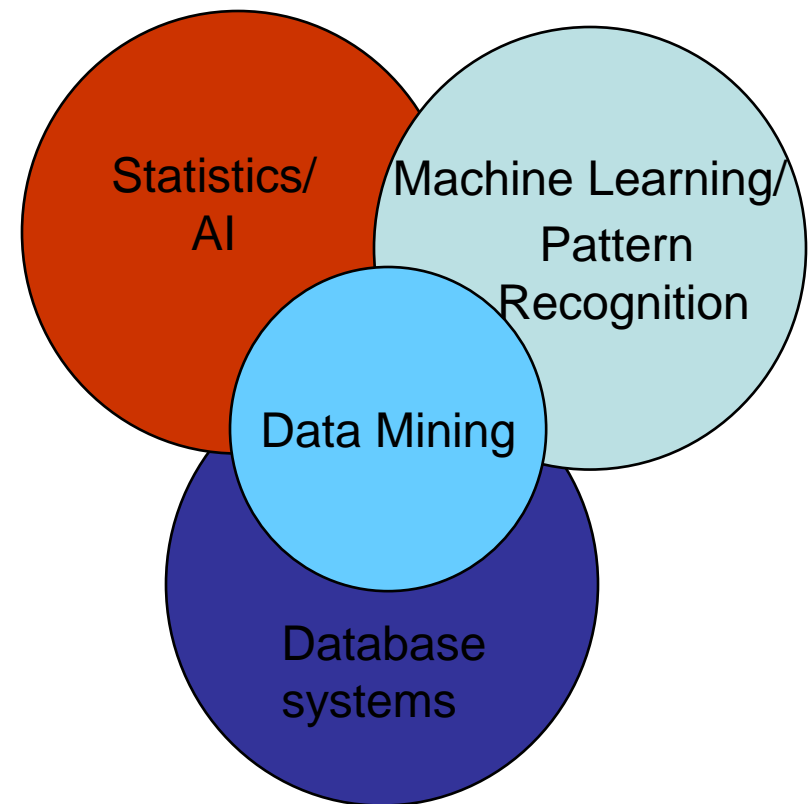
- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

□ What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Data Mining Methods – Task Categorization

- Supervised vs. Unsupervised
- Unsupervised Methods (also called **Descriptive**):
Try to find meaningful patterns in the data.
 - **Clustering**: group similar data into clusters
 - ❖ Market Segmentation, Document Clustering
 - **Association Rule Discovery**: find human interpretable patterns (associations)
 - ❖ Product Recommendations, Store Shelf Management
 - **Sequential Pattern Discovery**: describe the sequential dependencies among different events
 - ❖ Buying Patterns, Gene Sequencing
 - **Unsupervised anomaly detection** to detect anomalies in unlabeled data under the assumption that the majority of the instances are normal
 - ❖ Fraud Detection, Network Intrusion Detection

Supervised

Unsupervised

Clustering

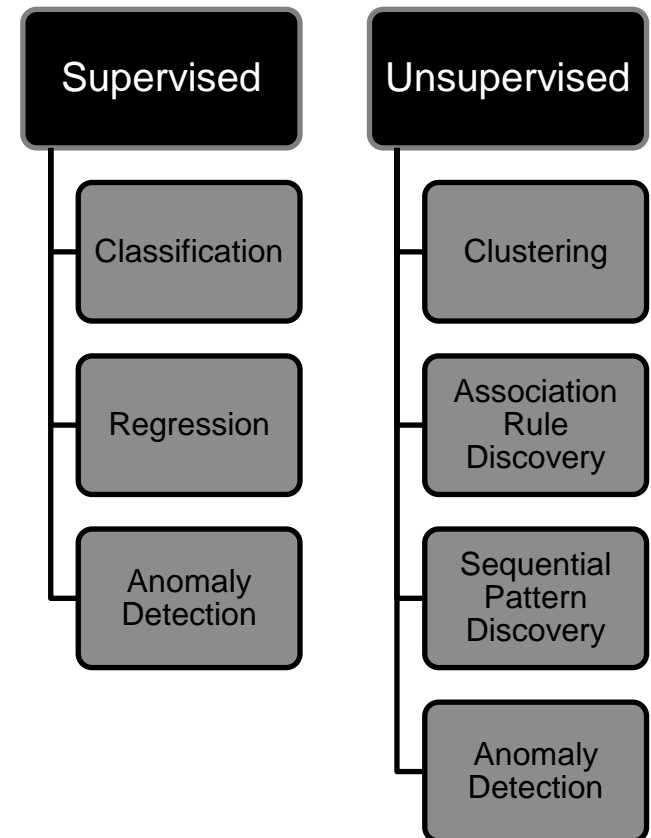
Association
Rule
Discovery

Sequential
Pattern
Discovery

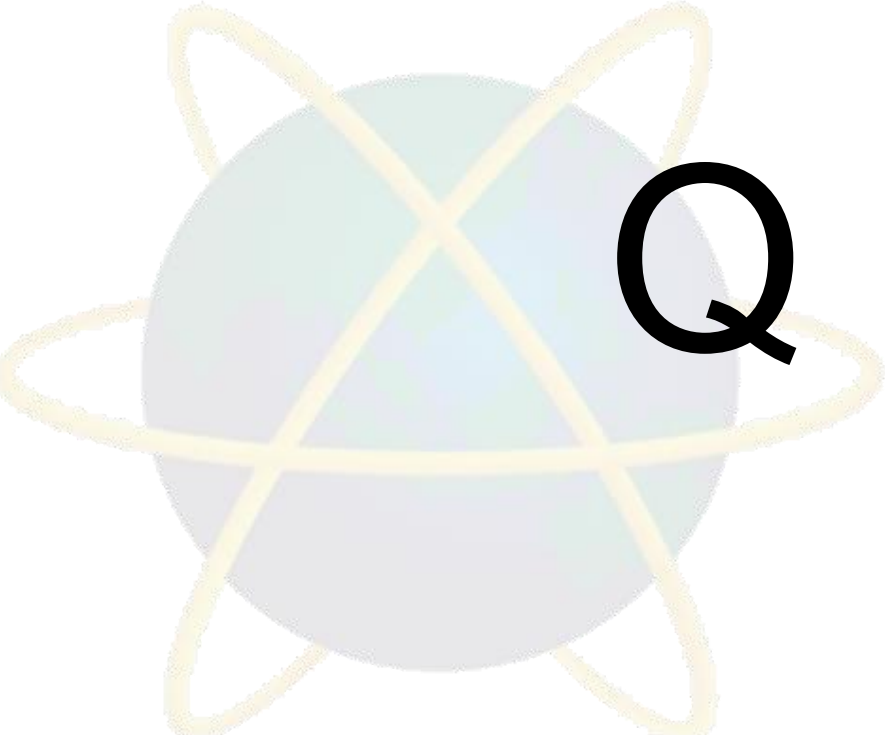
Anomaly
Detection

Data Mining Methods – Task Categorization

- Supervised vs. Unsupervised
 - Supervised Methods (also called **Predictive**):
Predict an unknown value(s) of a variable(s) from the values of some attributes
 - **Classification**: predict the type/class of new cases
 - ❖ Spam Filtering, Handwriting Character Recognition, Patient Diagnosis
 - **Regression**: predict a numerical value of new cases
 - ❖ Blood Pressure, Sales Amounts
 - **Supervised Anomaly Detection**: identify items, events or observations deviating from expected patterns using data labeled as "normal" and "abnormal" (involves training a classifier)
 - It is common to combine different methods such as clustering and classification (Hybrid methods)



Question & Answer Session



Q & A