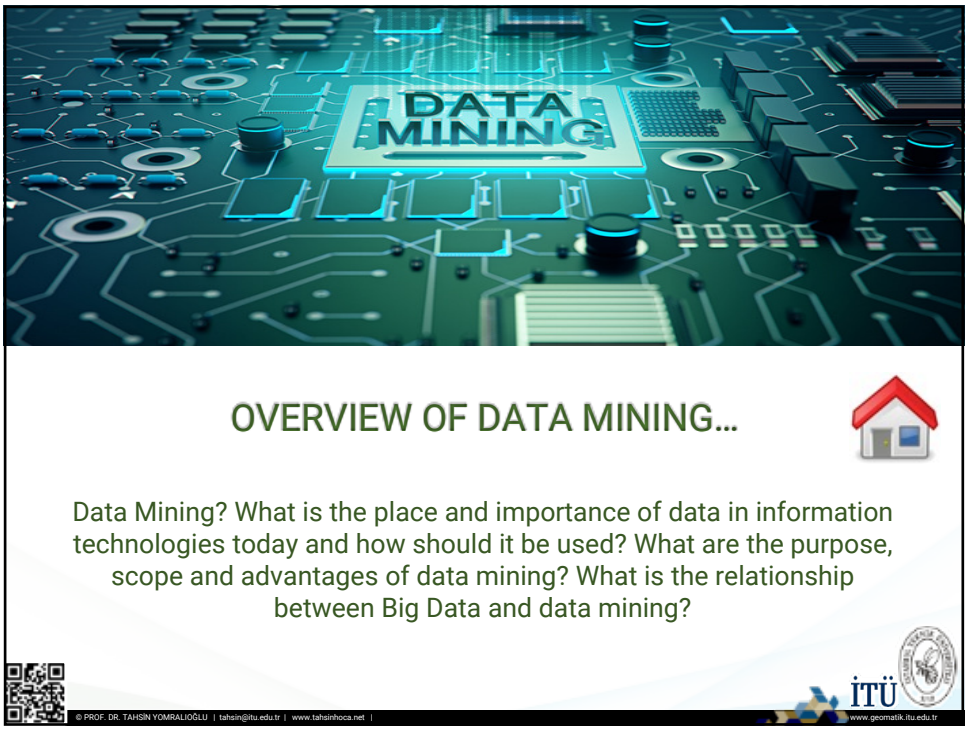

URBAN INFORMATION SYSTEMS
Prof. Dr. Tahsin YOMRALIOĞLU
İTÜ
GEOMATİK

390

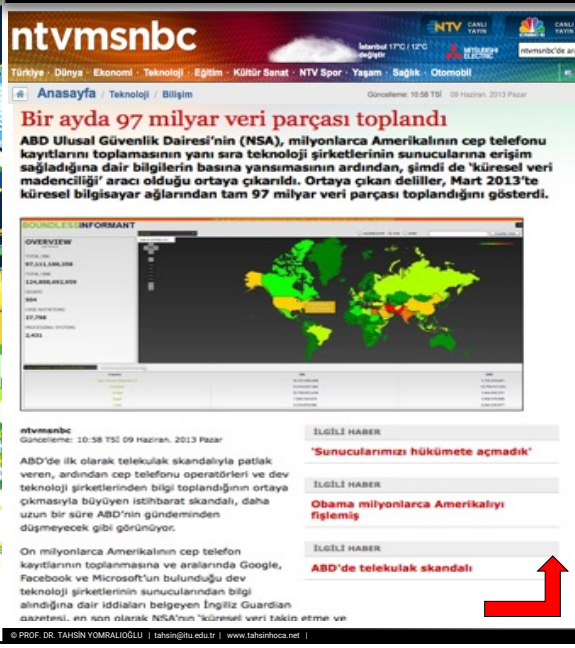

DATA MINING
OVERVIEW OF DATA MINING...
Data Mining? What is the place and importance of data in information technologies today and how should it be used? What are the purpose, scope and advantages of data mining? What is the relationship between Big Data and data mining?
İTÜ

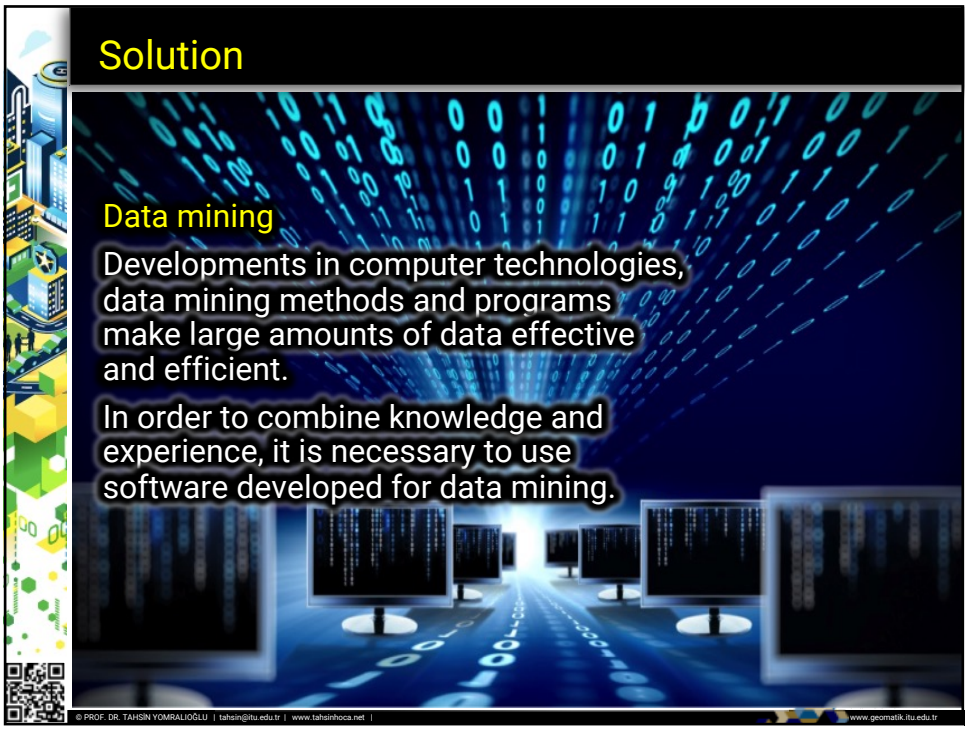
391

Reason…!

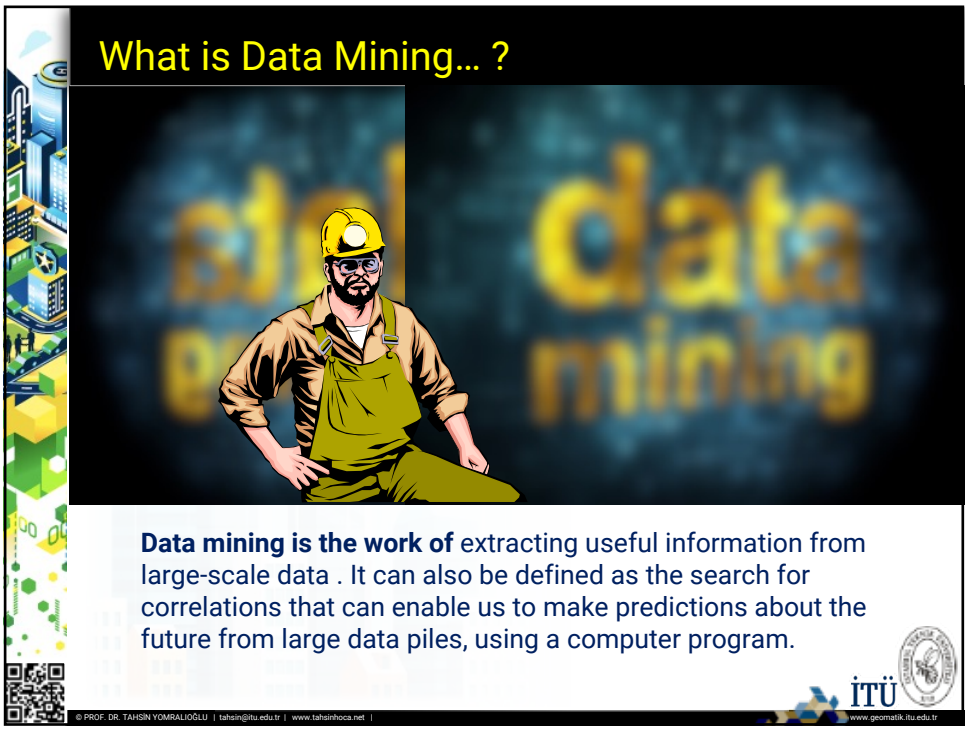
Developments in data collection tools and database technologies require the storage and analysis of large amounts of information in information wareho

Reason…!




Solution
Data mining
Developments in computer technologies, data mining methods and programs make large amounts of data effective and efficient.
In order to combine knowledge and experience, it is necessary to use software developed for data mining.
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net |
www.geomatik.itu.edu.tr

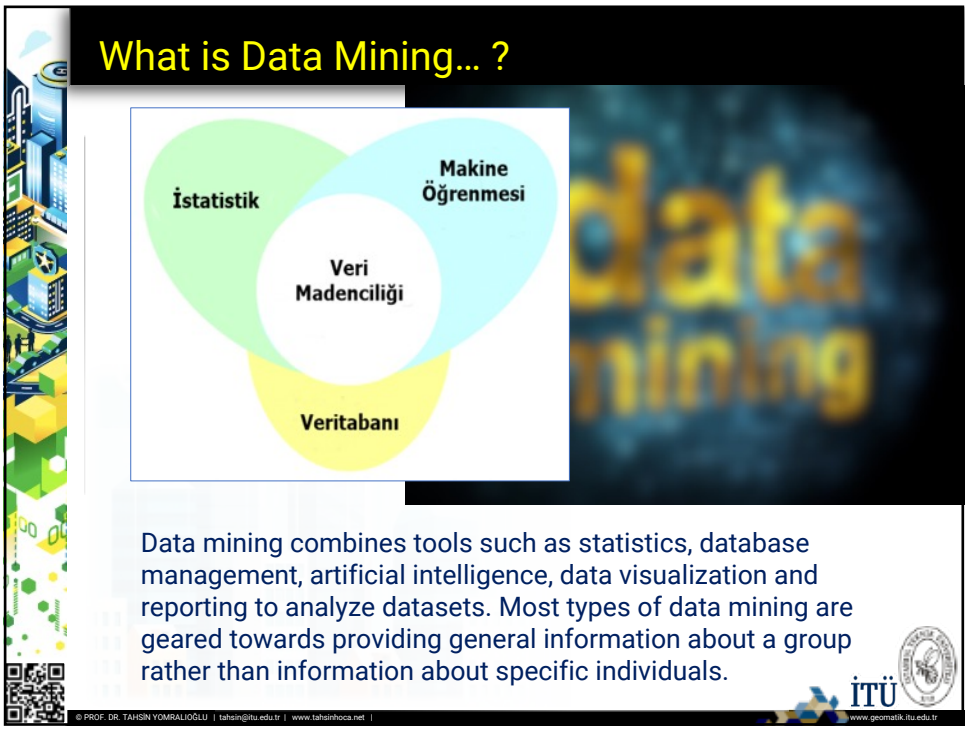
394


What is Data Mining… ?
data
mining
Data mining is the work of extracting useful information from large-scale data . It can also be defined as the search for correlations that can enable us to make predictions about the future from large data piles, using a computer program.
İTÜ
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net |
www.geomatik.itu.edu.tr

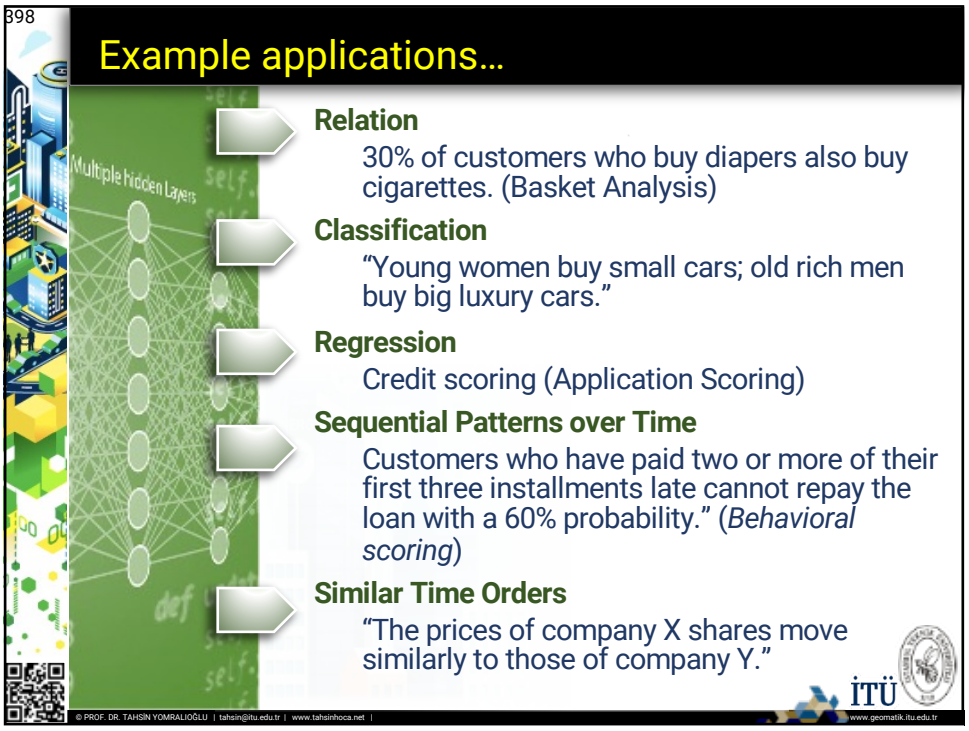
395


What is Data Mining… ?
data
mining
Data mining is the search for correlations and rules that will enable us to predict the future from large amounts of data.
Knowledge Discovery in Databases
İTÜ

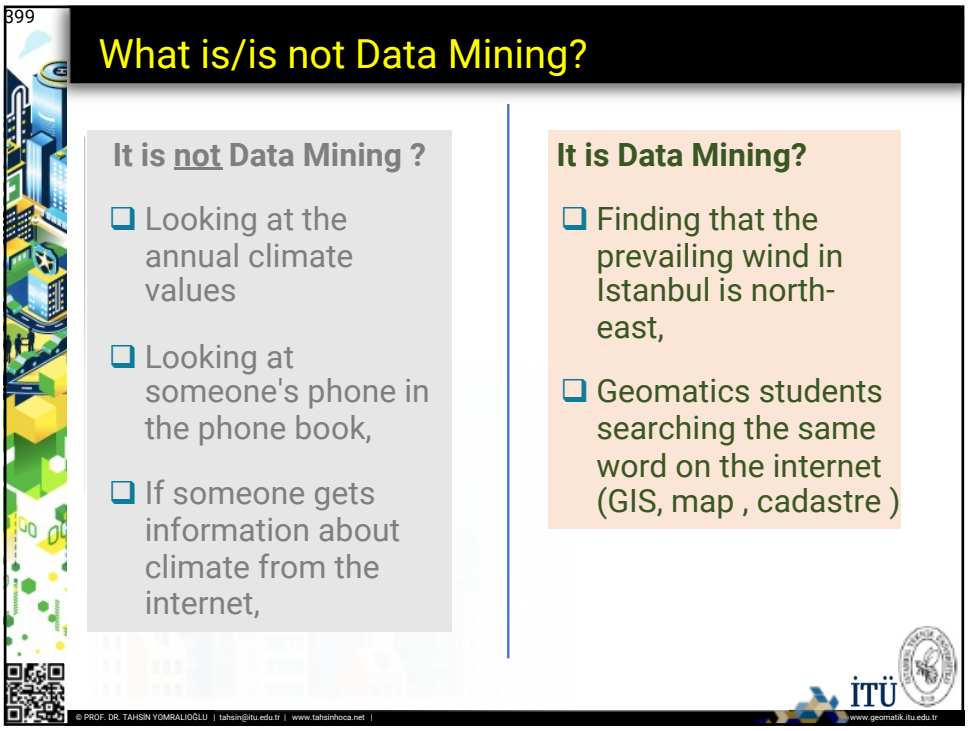
396


What is Data Mining… ?
İstatistik
Makine Öğrenmesi
Veri Madenciliği
Veritabanı
Data mining combines tools such as statistics, database management, artificial intelligence, data visualization and reporting to analyze datasets. Most types of data mining are geared towards providing general information about a group rather than information about specific individuals.
İTÜ

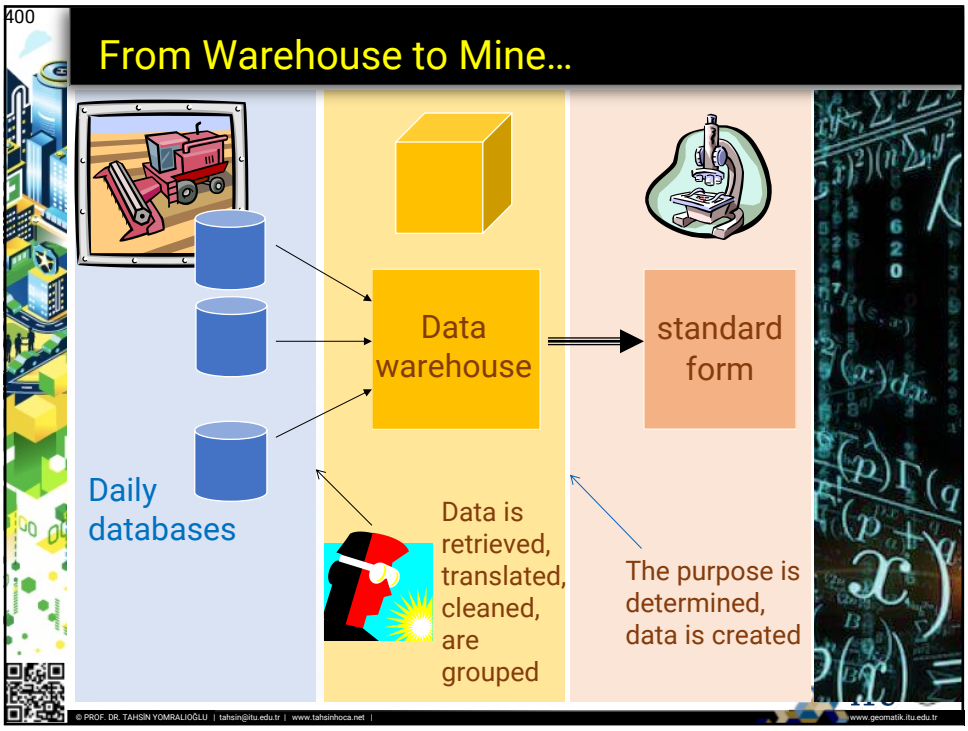
397


398
Example applications…
Multiple hidden Layers
Relation
30% of customers who buy diapers also buy cigarettes. (Basket Analysis)
Classification
"Young women buy small cars; old rich men buy big luxury cars."
Regression
Credit scoring (Application Scoring)
Sequential Patterns over Time
Customers who have paid two or more of their first three installments late cannot repay the loan with a 60% probability." (*Behavioral scoring*)
Similar Time Orders
"The prices of company X shares move similarly to those of company Y."
İTÜ

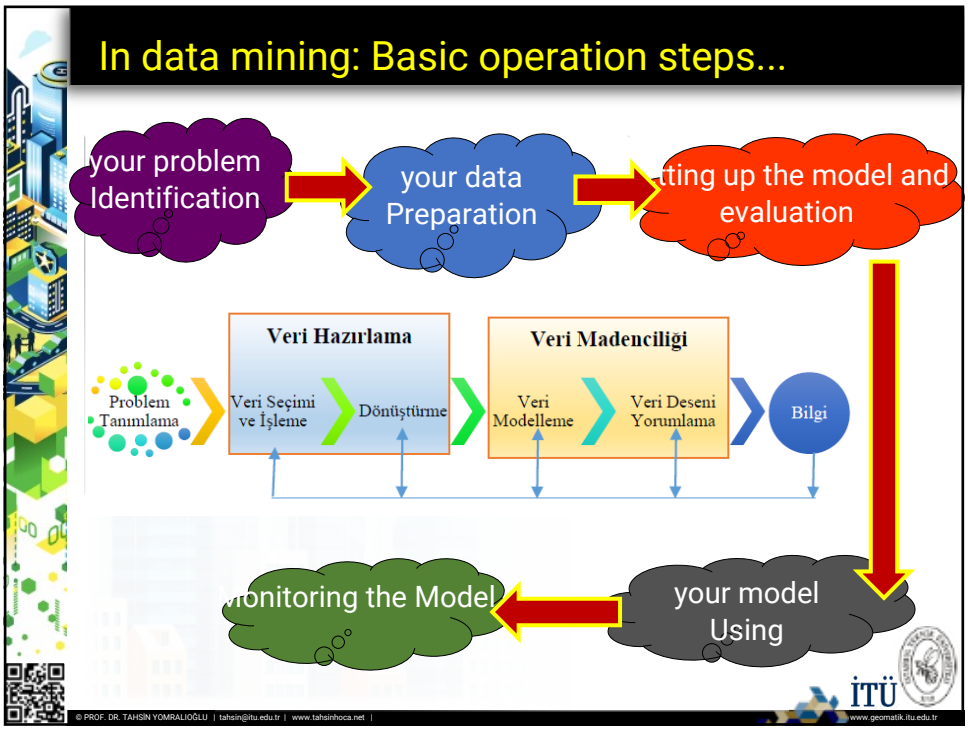
398


399
What is/is not Data Mining?
It is not Data Mining ?
Looking at the annual climate values
Looking at someone's phone in the phone book,
If someone gets information about climate from the internet,
It is Data Mining?
Finding that the prevailing wind in Istanbul is north-east,
Geomatics students searching the same word on the internet (GIS, map , cadastre)
İTÜ

399


400
From Warehouse to Mine…
Data warehouse
standard form
Daily databases
Data is retrieved, translated, cleaned, are grouped
The purpose is determined, data is created

400


In data mining: Basic operation steps...
your problem Identification
your data Preparation
ting up the model and evaluation
Veri Hazırlama
Veri Madenciliği
Problem Tanımlama
Veri Seçimi ve İşleme
Dönüştürme
Veri Modelleme
Veri Deseni Yorumlama
Bilgi
your model Using
onitoring the Model
İTÜ

401

402

Steps: 1. Defining Purpose

1. Relationship between products?
2. New market segments or potential customers?
3. Buying patterns or product selling curves over time?
4. Grouping, classifying customers?

402

403

Steps: 2. Data Preparation

- Merge, select and preprocess data.

(If there is a data warehouse it has already been done)

- Apart from the existing data, is there any additional information that can be used for the purpose?

1. **Data selection:** Identifying important variables
2. **Data cleaning:** Extracting/correcting errors, inconsistencies, duplication and missing data
3. **Data scrubbing:** Grouping, transformations
4. **Visual inspection:** Data distribution, structure, exceptions, correlations between variables
5. **Variable analysis:** Grouping

403

404

Steps: 3. Choosing a Technique

1. **Defining the purpose class**
   Grouping (Clustering/ Segmentation), Association, Classification, Pattern finding/prediction over time (Pattern detection / Prediction in time)
2. **Defining my solution class**
   Description (Decision trees, rules) vs. Black box (neural network)
3. **Model evaluation, validation and comparison**
   k-fold cross validation, statistical tests
4. **Combining models**

404

405

Steps: 4. Interpretation

- **results** (explanations/predictions) accurate or notable?
- **to the specialist advice**

405

406

Results: The Importance of Data, Expertise and Patience…

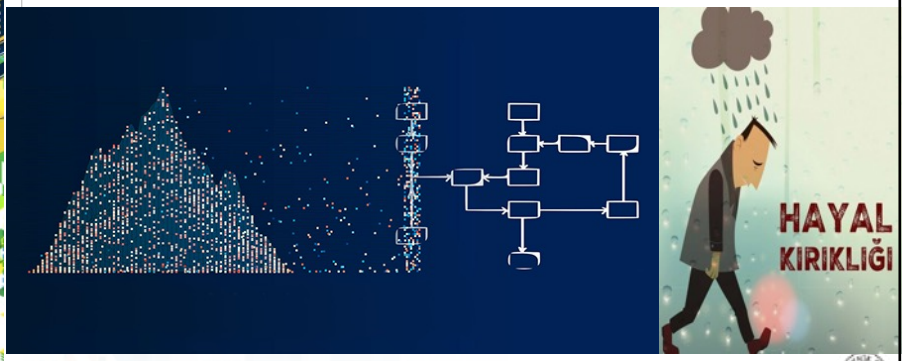
- **The goal** is to extract valuable information from large volumes of raw data…
- A large amount of **reliable data is a prerequisite** . The quality of the solution primarily depends on the quality of the data.
- Data mining **is not magic** ; We can't turn it under the stone.
- Data mining is the **joint work of the experts in the application field and the computer.**
- Any practical and useful information (symmetries, constraints, etc.) should be given to the system to aid learning.
- Consistency of results **needs to be audited by experts** .

406

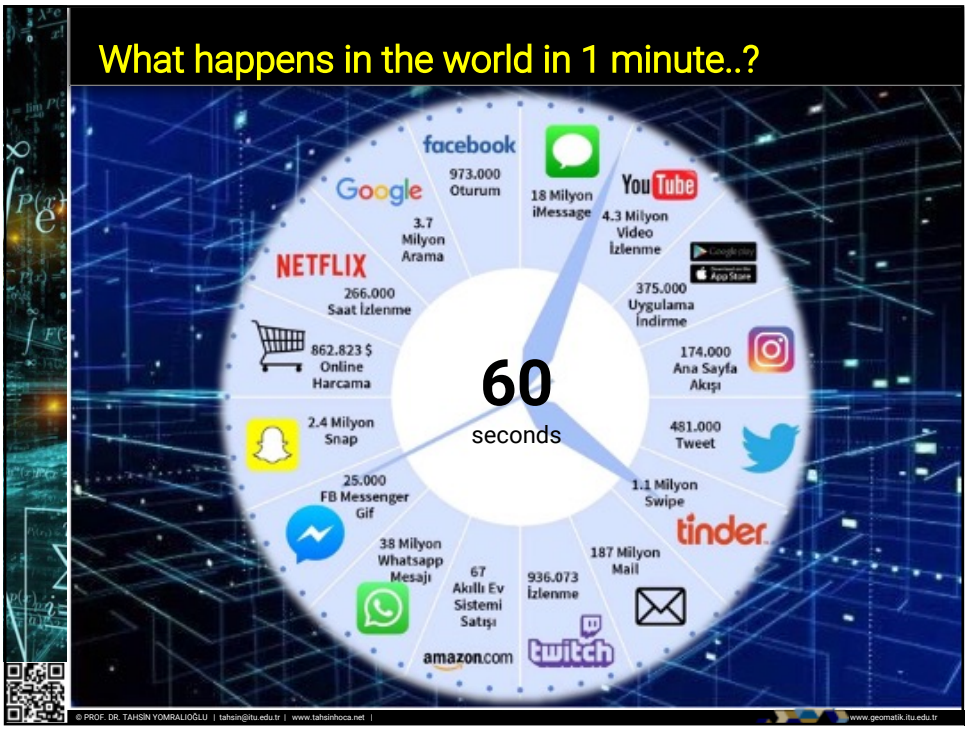
407

Results: The Importance of Data, Expertise and Patience…

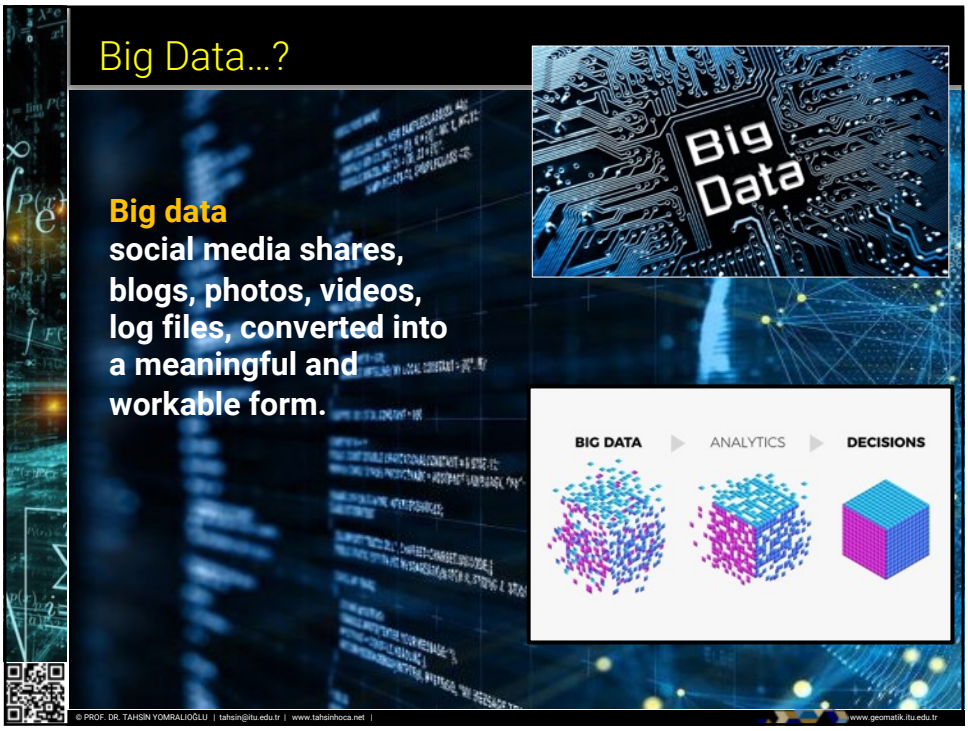
- Data mining is not a one-step study; is **repetitive** . It takes many tries until the system is set up.
- Data mining can be a long work. **Great expectations lead to great disappointments.**



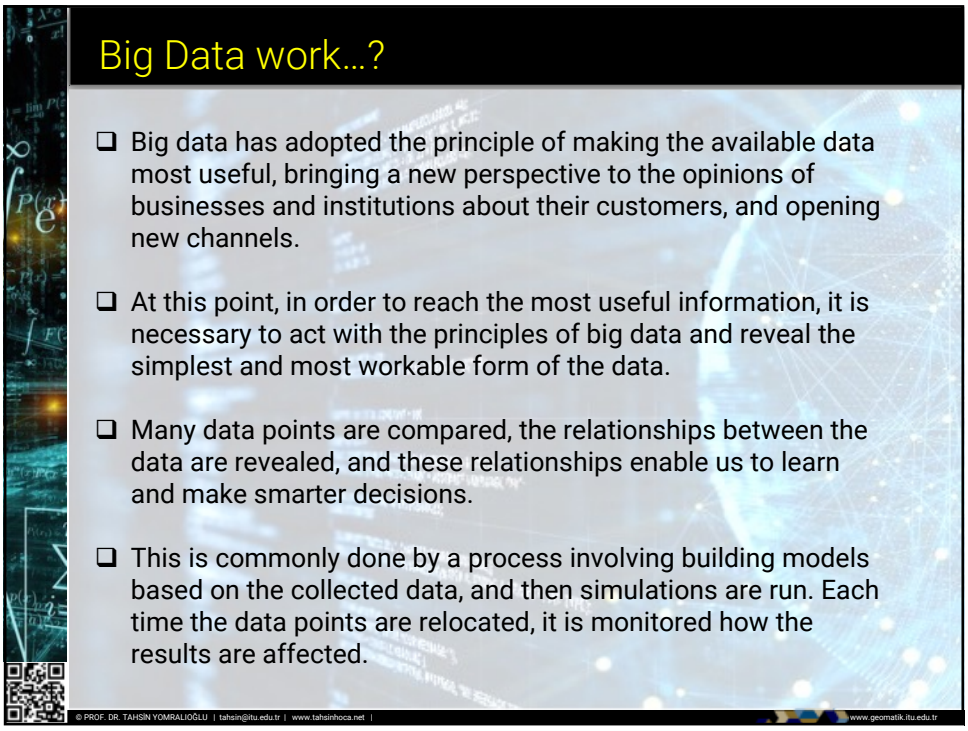
407


What happens in the world in 1 minute..?
60
seconds
facebook
973.000 Oturum
18 Milyon iMessage
YouTube
4.3 Milyon Video İzlenme
Google
3.7 Milyon Arama
NETFLIX
266.000 Saat İzlenme
375.000 Uygulama İndirme
862.823 $ Online Harcama
174.000 Ana Sayfa Akışı
2.4 Milyon Snap
481.000 Tweet
25.000 FB Messenger Gif
1.1 Milyon Swipe
tinder
38 Milyon Whatsapp Mesajı
187 Milyon Mail
67 Akıllı Ev Sistemi Satışı
936.073 İzlenme
amazon.com
twitch
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net | www.geomatik.itu.edu.tr

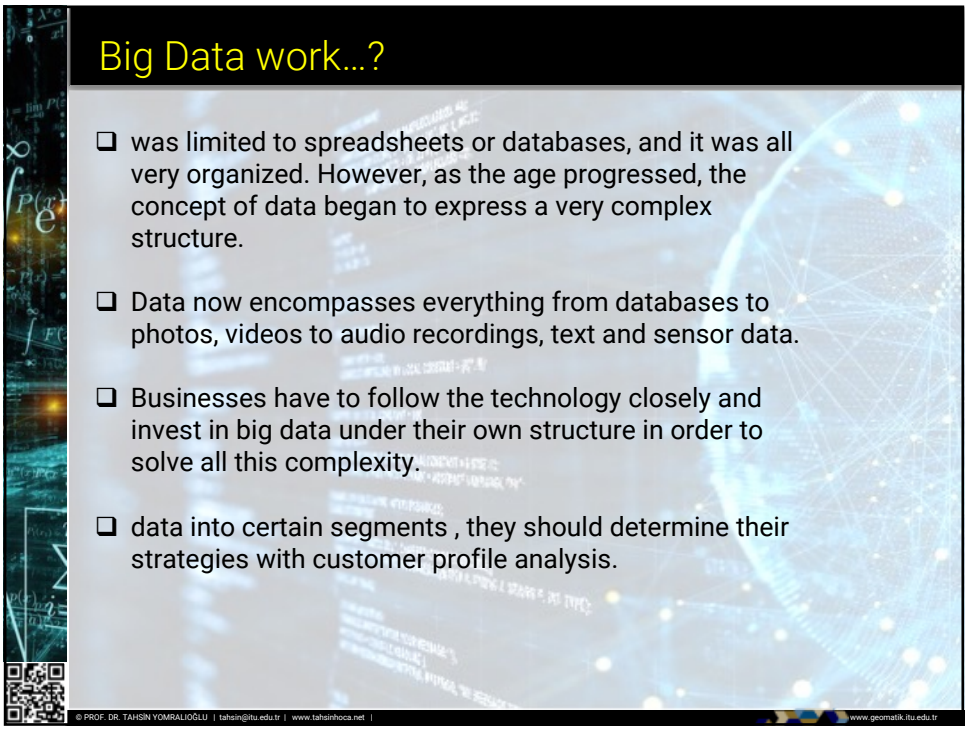
408


Big Data…?
Big data
social media shares, blogs, photos, videos, log files, converted into a meaningful and workable form.
Big Data
BIG DATA
ANALYTICS
DECISIONS
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net | www.geomatik.itu.edu.tr

409


Big Data work…?
❑ Big data has adopted the principle of making the available data most useful, bringing a new perspective to the opinions of businesses and institutions about their customers, and opening new channels.
❑ At this point, in order to reach the most useful information, it is necessary to act with the principles of big data and reveal the simplest and most workable form of the data.
❑ Many data points are compared, the relationships between the data are revealed, and these relationships enable us to learn and make smarter decisions.
❑ This is commonly done by a process involving building models based on the collected data, and then simulations are run. Each time the data points are relocated, it is monitored how the results are affected.
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net |
www.geomatik.itu.edu.tr

410


Big Data work…?
❑ was limited to spreadsheets or databases, and it was all very organized. However, as the age progressed, the concept of data began to express a very complex structure.
❑ Data now encompasses everything from databases to photos, videos to audio recordings, text and sensor data.
❑ Businesses have to follow the technology closely and invest in big data under their own structure in order to solve all this complexity.
❑ data into certain segments , they should determine their strategies with customer profile analysis.
© PROF. DR. TAHSİN YOMRALIOĞLU | tahsin@itu.edu.tr | www.tahsinhoca.net |
www.geomatik.itu.edu.tr

411

Reasons for the Increase of Big Data….

- In the 1980s , when the product was much more important, the main purpose of companies was to manufacture a particular product and make it available to the customer.
- In these years, it was seen that ERP (Enterprise Resource Planning) systems were at the forefront. Some of the main purposes in the development of ERP systems; was to gather the customer, distribution center, suppliers and production on one platform. When saturation was reached in this once very popular system, people started asking the question "**who is the right customer for me?"**
- It can be said that the birth of CRM systems started with this question. Customer Relations Management is "To present the right product or service to the right customer, at the right price, in the right channel, at the right place and at the right time."
- In other words, product turnover has started according to the customer, not the customer according to the product. This methodology, which has been a rising value for the last 10 years, is gradually increasing its importance.

412

Reasons for the Increase of Big Data….

- The main reason for the rapid increase in the amount of data can be summarized as the growing volume of social network interactions, the increase in location sensitive devices, and the increase in the number of "smart sensors" that capture and transmit information about the physical world . Of course, video and media sources can also be added to these.
- In the last few years, the importance of real-time data has increased with the developments in mobile technology and social media, and besides the volume of data, its diversity and data growth rate have been affected by these developments.
- The increase in the usage rate of smartphones, in addition to providing 24/7 access to the Internet , online messaging applications such as Whatsapp and social media applications such as Facebook, Twitter and Instagram have increased the sharing of instant messages, photos and videos, and location -based data generation has become possible thanks to GPS technology. .

413

Data Components in Big Data

There are five main components in the formation of the big data platform. These; variety, velocity, volume, verification and value. Since it is explained as 5v in general, English equivalents can be given.

1) **Variety (Çeşitlilik):** 80 percent of the data produced is unstructured and every newly produced technology can produce data in different formats. All kinds of "Data Types" have to be dealt with from phones, tablets, integrated circuits. Also, if you think that this data can be in different languages, Non -Unicode, they need to be integrated and converted to each other.
2) **Velocity (Hız):** Big Data is produced is very high and increasing. Data that reproduces faster results in an increase in the number and variety of transactions that need that data at the same rate.
3) **Volume (Veri Hacmi):** According to IDC statistics, the amount of data to be reached in 2020 will be 44 times that of 2009. It is necessary to think about the capacities and "large systems" currently in use, which we call "large", and imagine how they will cope with data 44 times larger! It is necessary to design how the institution's data archiving, processing, integration, storage, etc. technologies will cope with such a large data volume. In the 2010s, total IT expenditures in the world increased by 5% per year, but the amount of data produced increased by 40%.

414

Data Components in Big Data

4) **Verification (Doğrulama):** Another component of this information density is that the data is "safe" during its flow. During the flow, it needs to be monitored with the right layer, at the security level it should be, visible or hidden by the right people.
5) **Value (Değer):** The most important component is value creation. Big Data , which is described with all the above efforts , should create an added value for the institution after your data production and processing layers. It needs to have an instant effect on your decision-making processes, and it should be at your fingertips to make the right decision.

   Exp: see the distribution of diseases, drugs, doctors in details such as region, province, district, etc. The Air Force should be able to see the instant location and status of all its vehicles in its aviator inventory, and track their retrospective maintenance histories. A bank should be able to monitor not only the demographic information of the person to whom it will lend, but also their eating and vacation habits, and if necessary, be able to see what they are doing on social networks.

415

Big Data Applications...

- *According to research, companies using big data; They made 50% more profits, were 41% effective in their market efforts, decreased their advertising expenditures by 37%, and were more successful in their social media use, with rates as high as 37%.*

416

Big Data Applications...

Education; In the individualization of learning processes, big data is processed by using learning analytics, and learning processes can be designed according to the learning needs, behaviors and emerging patterns of the learners.

Hospitals; In order to provide effective, individualized, personalized medical services for their patients, they store individual data in their own digital media.

Governments; They have to work with very large-scale data about processing and storing information and services for their citizens. For example, in accordance with RTÜK's decisions, television channels in our country have to keep their broadcasts for the last year. The information to be stored is of the type we define as "Big Data".

Internet accelerate the production of data imposes the task of re-presenting the growing information, especially to service providers, by blending it and transforming it into a meaningful form.

417

Big Data Applications…

Banks , through the information they collect and store about their customers , provide a rich, fast and convenient 7/7 system that recognizes the user, knows what he or she is entering the Internet branch for that day, and optimizes the homepage and menu accordingly, reminds customers, and offers customizable interfaces. It has grown into 24 branches.

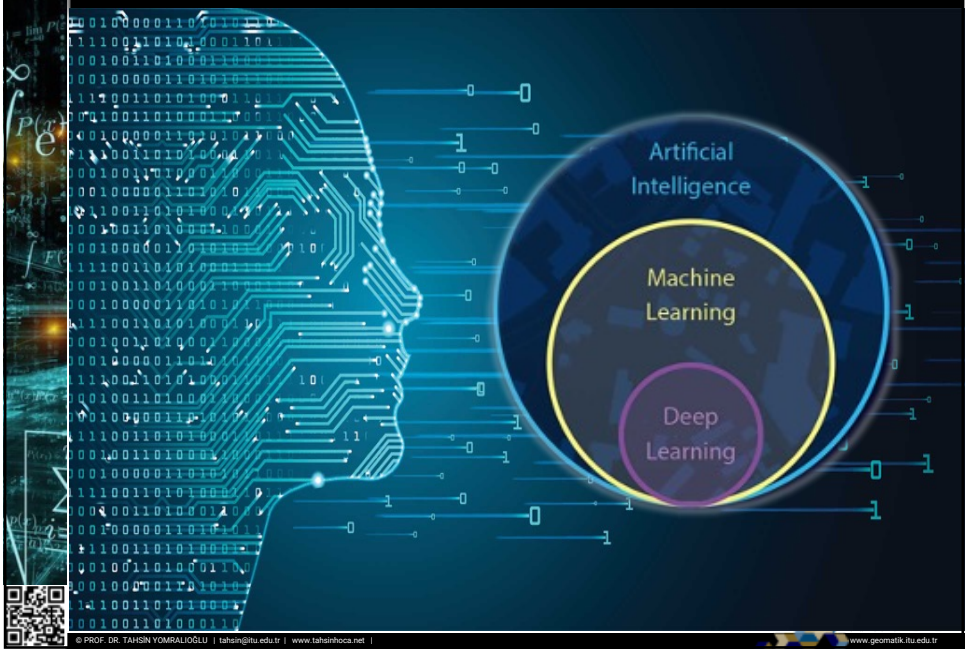
Using smart grids and meters, **energy companies have to store and process data about their subscribers' individual usage.**

Pharmaceutical industry; e.g. large genomics created for "cancer research" databases must be open to researchers' constant access.

Audio and video information produced by Satellite/Map Arrays (GPS), Smart Mobile Phones (GSM), new generation cameras that can take very high resolution photos; It pushes the limits of storage environments and reduces their efficiency. In addition to those produced by Internet-based software and applications that can work on all kinds of mobile devices, the necessity of storing information produced by users in social media environments such as Facebook and Twitter has pushed IT-related entrepreneurs into the Big Data field.

418

Future trends…?



419


URBAN INFORMATION SYSTEMS
Thanks...
Prof. Dr. Tahsin YOMRALIOĞLU
İTÜ
GEOMATİK

420