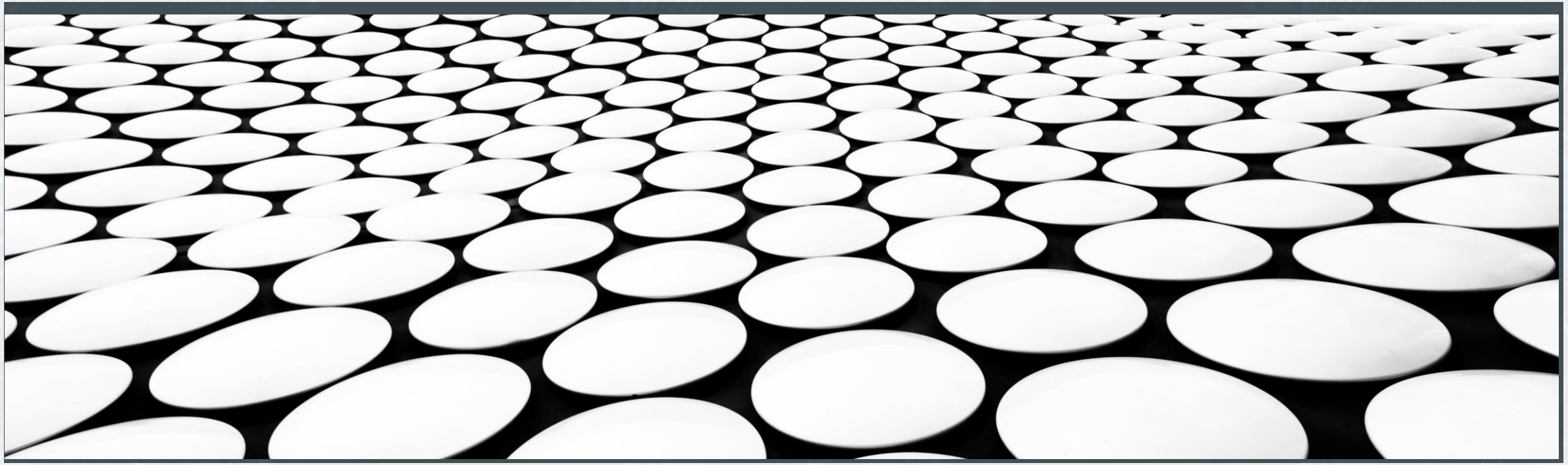
BLG2292 - VERİ MADENCİLİĞİ VE BÜYÜK VERİ (Data Mining and Big Data)

Dr. Günay TEMÜR

gunaytemur@duzce.edu.tr



VERİ MADENCİLİĞİNE GİRİŞ

GÖLYAKA MESLEK YÜKSEKOKULU / BİLGİSAYAR PROGRAMCILIĞI

DERS BİLGİLERİ

- BLG2292 - Veri Madenciliği ve Büyük Veri
- Ders ile ilgili duyurular ve Kaynaklar
  - http://www.gunaytemur.com
- Kaynaklar
  - Balıkesir Üniversitesi Veri Madenciliği Ders Notları, Dr. Öğr. Üyesi. Kadriye ERGÜN
  - Veri Madenciliği Yöntemleri (Kitap), Dr. Yalçın ÖZKAN
  - Hacettepe Üniversitesi Veri Ambari ve Veri Madenciligi Ders Notları, Prof.Dr. Pinar Duygulu Sahin
- Değerlendirme
  - Vize
  - Final
  - ???

- Devamsızlık
  - %70 Devam etmek zorundasınız ☺

İÇERİK

- Veri Madenciliğine Giriş
- Veri Madenciliği Uygulama Alanları
- Veri Ambarları ve OLAP
- Veri Madenciliği Süreci
- Veri Madenciliği Yöntemleri
  - Sınıflandırma,
  - Kümeleme,
  - Birliktelik Kuralları
- ARA SINAV
- Karar Ağaçları ile Sınıflandırma
  - Uygulama Örnekleri
- Sınıflandırma ve Regresyon Ağaçları
  - Uygulama Örnekleri
- Kümeleme Analizi
  - Uygulama Örnekleri
- Birliktelik Kuralları
  - Uygulama Örnekleri
- Metin Madenciliği ve Web Madenciliği
- Büyük Veri Kavramı ve Büyük Veri Analizi

VERİ MADENCİLİĞİNE GİRİŞ

- TANIM_1 : Büyük miktardaki veriler içerisinden önemli olanlarını bulup çıkarmaya **Veri Madenciliği** denir.
- TANIM_2 : Veri Madenciliği (Data Mining): Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır. (Knowledge Discovery in Databases).
- Tanım_3 : Kurumlardaki büyük ölçekli olarak tanımlanan ve milyonlarca veriye sahip yazılım sistemlerinden, ihtiyacı karşılayacak değerli verilerin elde edilmesi işlemine Veri Madenciliği denilmektedir.
- Tanım_4 : Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.

VERİ MADENCİLİĞİNE GİRİŞ

- Sonuç olarak; Teknolojinin gelişmesi ve yaygınlaşması ile daha önce fiziksel olarak yapılan birçok iş ve işlem artık bilgisayardan, cep telefonundan ya da tabletten yapılıyor. Sensörlerden alınan bilgiler, güvenlik amaçlı kullanılan retina ve parmak izi verileri, meteorolojik ve jeofizik veriler, tıbbi kayıtlar, Banka ödemeleri, alışveriş, hastane randevusu alma ve daha birçok iş, telefonun birkaç tuşuna basarak gerçekleştirilebilmektedir. Bu durum dijital veri toplamanın ve saklamanın ne kadar yaygın olduğunu göstermektedir.
- Telefon ya da bilgisayar ile internet üzerinden yapılan her işlem sonucunda, veriler ilgili veri tabanlarında birikmektedirler.
- İşletmelerin sunucularında biriken bu verilerin toplanması, analiz edilmesi ve aralarından "İşe yarar" olanların ayıklanması işine "Veri madenciliği" denilmektedir.

ARTAN VERİ BOYUTU

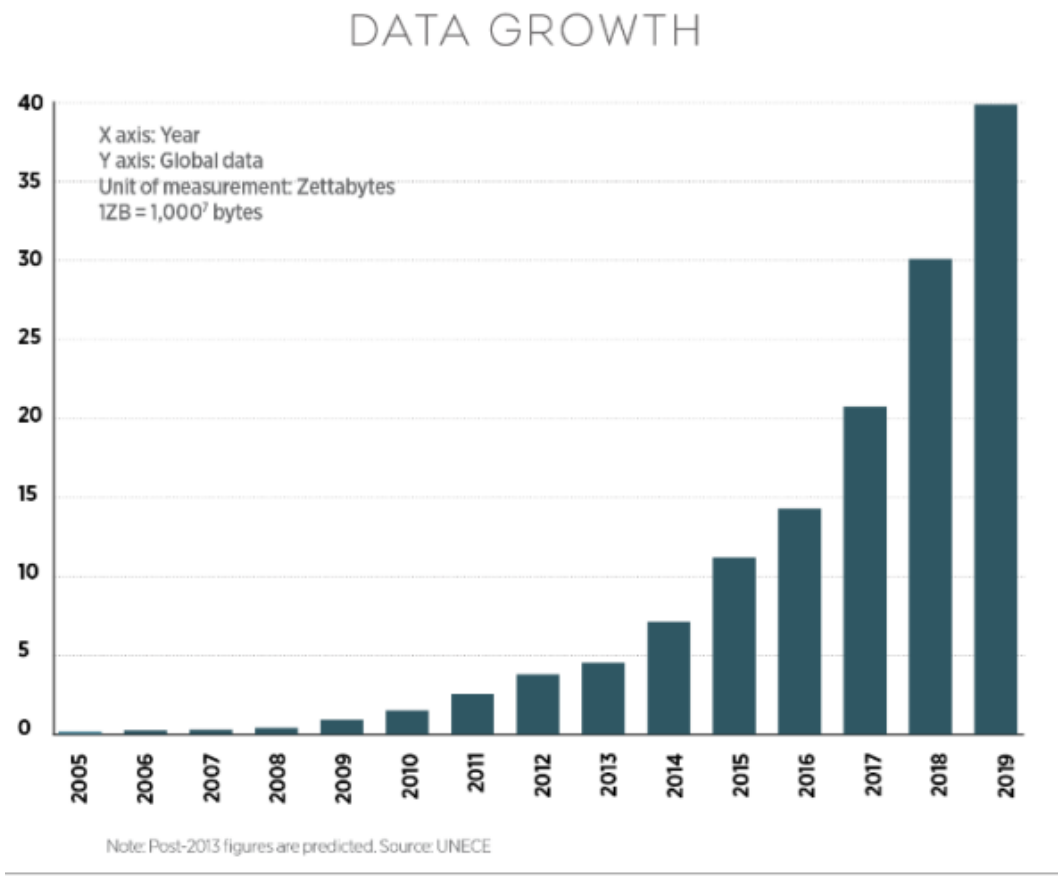
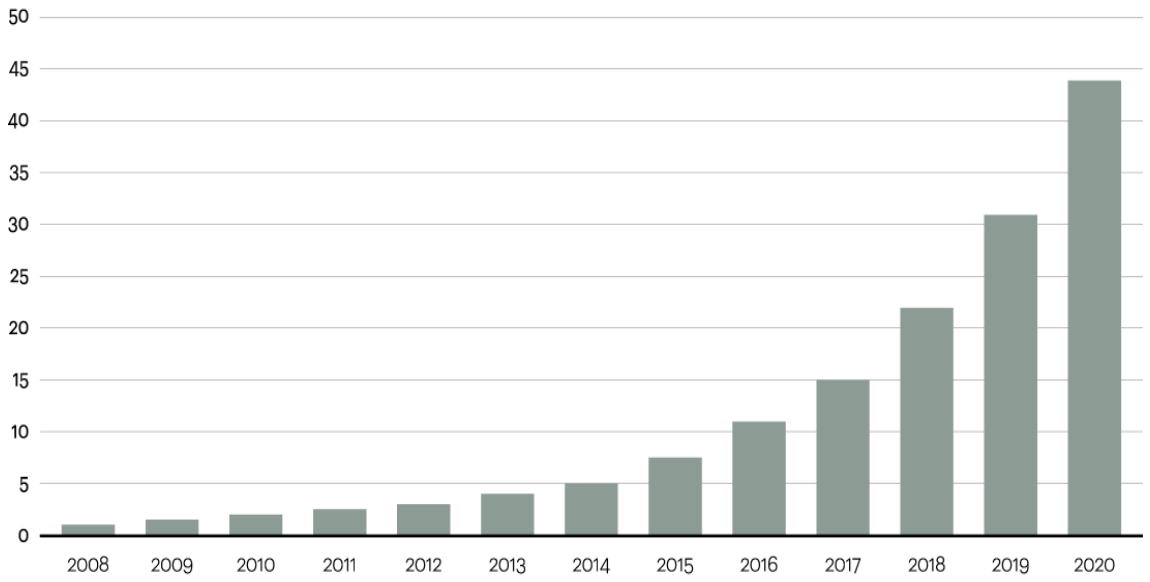


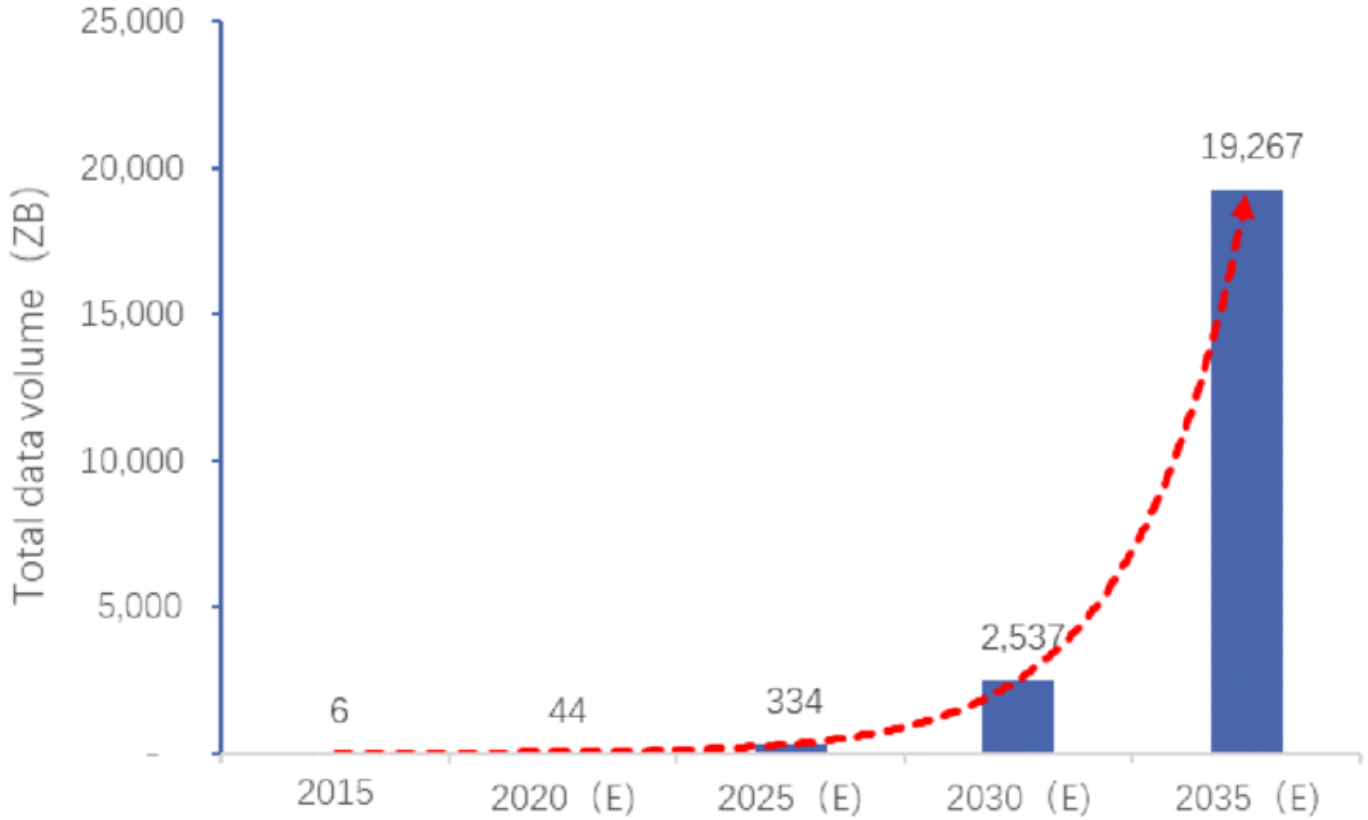
Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



Source: Oracle, 2012

TAHMİNİ VERİ BOYUTU



VERİ BOYUTU

1	zettabyte
1.024	eksabyte
1.048.576	petabyte
1.073.741.824	terabyte
1.099.511.627.776	gigabyte
1.125.899.906.842.620	megabyte
1.152.921.504.606.850.000	kilobyte
1.180.591.620.717.410.000.000	byte
1.208.925.819.614.630.000.000.000	bit

VERİ MADENCİLİĞİ TANIMLARI

- Bu tanımlamalardan da anlaşıldığı üzere veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik tahminlerde bulunmak veri madenciliği çalışmaları sayesinde mümkün olmaktadır.
- Bunun anlamı, veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi ortaya çıkarma süreci olarak değerlendirilmesidir. Bu açıdan bakıldığında veri madenciliği işinin kurumların Karar Destek Sistemleri için önemli bir yere sahip olduğu söylenebilir.

VERİ MADENCİLİĞİ İLE İLİŞKİLİ DİĞER DİSİPLİNLER

- Veri madenciliği disiplinlerarası bir çalışmadır. İstatistik, veritabanı teknolojileri, makine öğrenmesi, yapay zeka ve görselleştirme gibi bir çok farklı disiplin bünyesinde gelişen yöntemleri kullanır.

VERİ MADENCİLİĞİNİN TARİHÇESİ

1950'ler

İlk bilgisayarlar (Sayımlar için bilgisayarlar kullanılıyor)

1960'lar

Veri kolleksiyonları Veritabanı yaratımı (Hiyerarşik ve ağ modelleri)

VERİ MADENCİLİĞİNİN TARİHÇESİ

1970'ler

İlişkisel veri modeli İlişkisel VTYS uygulamaları

1980'ler

İlişkisel VTYS yaygınlaşıyor Uygulamaya yönelik VTYS (Mekansal, Bilimsel, Mühendislik, vs.)

VERİ MADENCİLİĞİNİN TARİHÇESİ

1990'lar

Günlük işlemlerden derlenen büyük miktarda verinin nasıl değerlendirilebileceği sorgulanmaya başlıyor. Bu noktada bahsedilmesi gereken birkaç önemli olay söz konusu: 1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı

1991,

KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek 'Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop' makalenin KDD ile ilgili temel tanım ve kavramları ortaya koyması

VERİ MADENCİLİĞİNİN TARİHÇESİ

1992,

Veri Madenciliği konusunda ilk yazılımın geliştirilmesi

1995,

1. Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı'nın (KDD-95) açılış konuşması

VERİ MADENCİLİĞİNİN TARİHÇESİ

2000'ler

Veri Ambarları,

Veri Madenciliğinin yaygınlaşması.

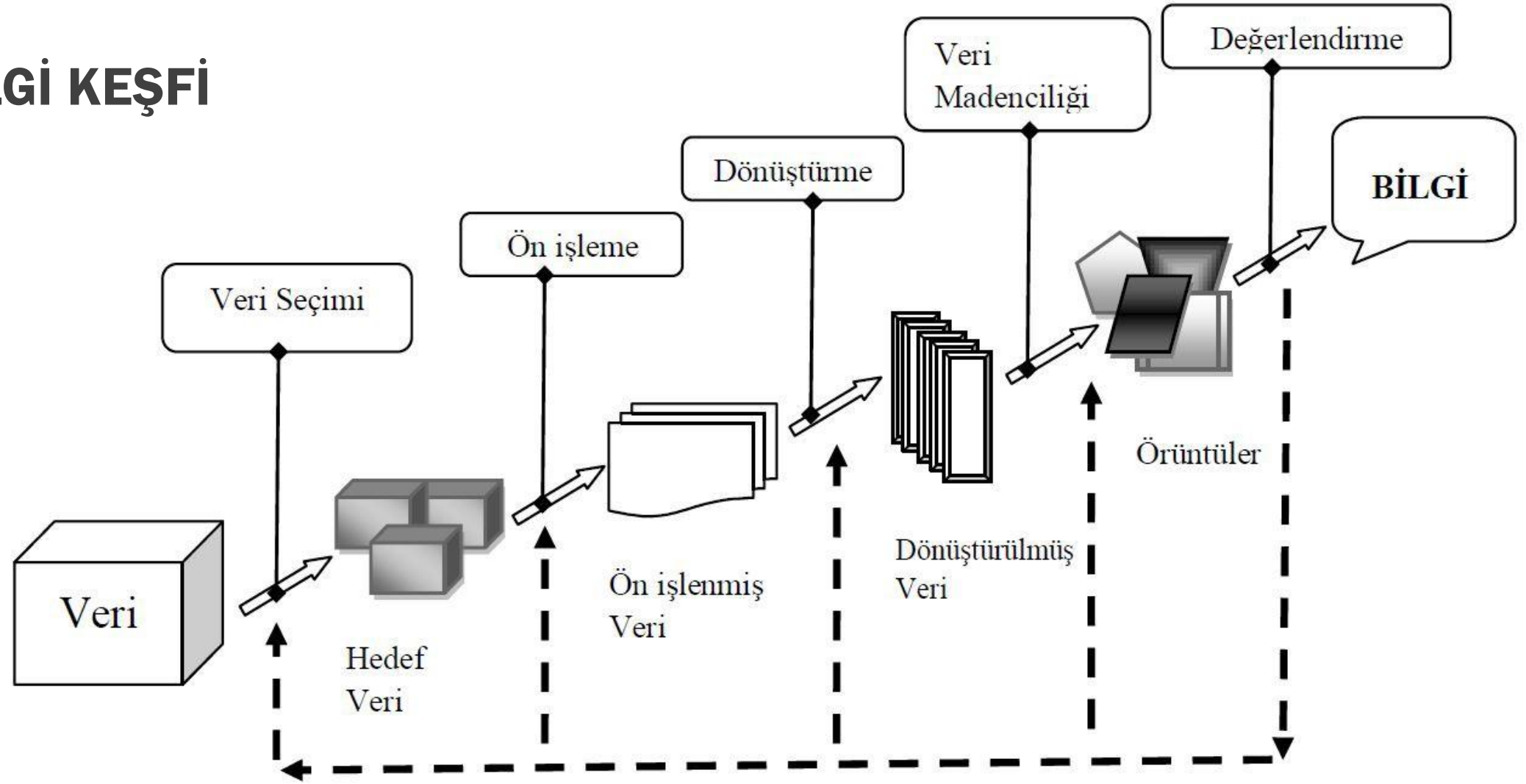
BİLGİ KEŞFİ

- Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir.
- Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.
- Veri madenciliği teknikleri veriyi belli bir modele uydurur.
  - veri içindeki örüntüleri bulur
  - örüntü: veri içindeki herhangi bir yapı
- Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.

BİLGİ KEŞFİ

- Büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak
- Bulunan bilgi
  - gizli,
  - önemli,
  - önceden bilinmeyen,
  - yararlı olmalı.

BİLGİ KEŞFİ


Veri
Veri Seçimi
Hedef Veri
Ön işleme
Ön işlenmiş Veri
Dönüştürme
Dönüştürülmüş Veri
Veri Madenciliği
Örüntüler
Değerlendirme
BİLGİ

BİLGİ KEŞFİNİN AŞAMALARI

- Veri Temizleme : Gürültülü ve tutarsız verileri çıkarmak
- Veri Bütünleştirme : Birçok data kaynağını birleştirebilmek
- Veri Seçme : Yapılacak olan analiz ile ilgili olan verileri belirlemek
- Veri Dönüşümü gerçekleştirmek : Verinin veri madenciliği yöntemine göre hale dönüşümünü
- Veri Madenciliği uygulanması : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin
- Örüntü Değerlendirme : Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumu

VERİ TABANLARINDA BİLGİ KEŞFİ

- Veri Tabanlarında Bilgi Keşfi, veriden faydalı bilginin keşfedilmesi sürecinin tamamına atıfta bulunmakta ve veri madenciliği bu sürecin bir adımına karşılık gelmektedir.

VERİ MADENCİLİĞİ CEVAPLAYABİLECEĞİ SORULAR NELERDİR?

- Bilinmeyen bilgiyi ortaya çıkarmak için çalışır
- Bu sene potansiyel en iyi 10 müşterimiz kimler olacaktır sorusuna cevap arar
- Gelecek 5 sene şirketimiz hangi alanda büyümeli/büyüyebilir
- Bu hastanın hastalığı ne olabilir
- Kanseri erken teşhis edebilir miyim

VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

- Finans Sektörü
- Haberleşme Sektörü
- Sağlık Sektörü
- Devlet Uygulamaları

Büyük hacimde veri olan her yerde veri madenciliği kullanılabilir.

VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

Veritabanı analizi ve karar verme desteği

- Pazar araştırması
  - Hedef Pazar, müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi
- Risk analizi
  - Kalite kontrolü, rekabet analizi, öngörü
- Sahtekarlıkların saptanması
- Diğer Uygulamalar
  - Belgeler arası benzerlik (haber kümeleri, e-posta)
  - Sorgulama sonuçları

VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

Bilim	İş Hayatı	Web	Devlet
• Astronomi • Biyoinformatik • İlaç keşfi	• Reklam • CRM (Müşteri İlişkileri Yönetimi) ve Müşteri Modelleme • E-ticaret • Yatırım değerlendirme ve karşılaştırma • Sağlık • Üretim • Spor/eğlence • Telekom (telefon ve iletişim) • Hedef pazarlama	• Metin Madenciliği (haber grubu, email, dokümanlar) • Web analizi • Arama motorları	• Kanun Yaptırımı • Vergi Kaçakçılarının Profilinin Çıkarılması

VERİ MADENCİLİĞİNİN UYGULAMA ALANLARI

- Hangi promosyonu ne zaman uygulamalıyım?
- Hangi müşteri aldığı krediyi geri ödemeyebilir?
- Bir müşteriye ne kadar kredi verilebilir?
- Sahtekarlık olabilecek davranışlar hangileridir?
- Hangi müşteriler yakın zamanda kaybedilebilir?
- Hangi müşterilere promosyon yapmalıyım?
- Hangi yatırım araçlarına yatırım yapmalıyım?

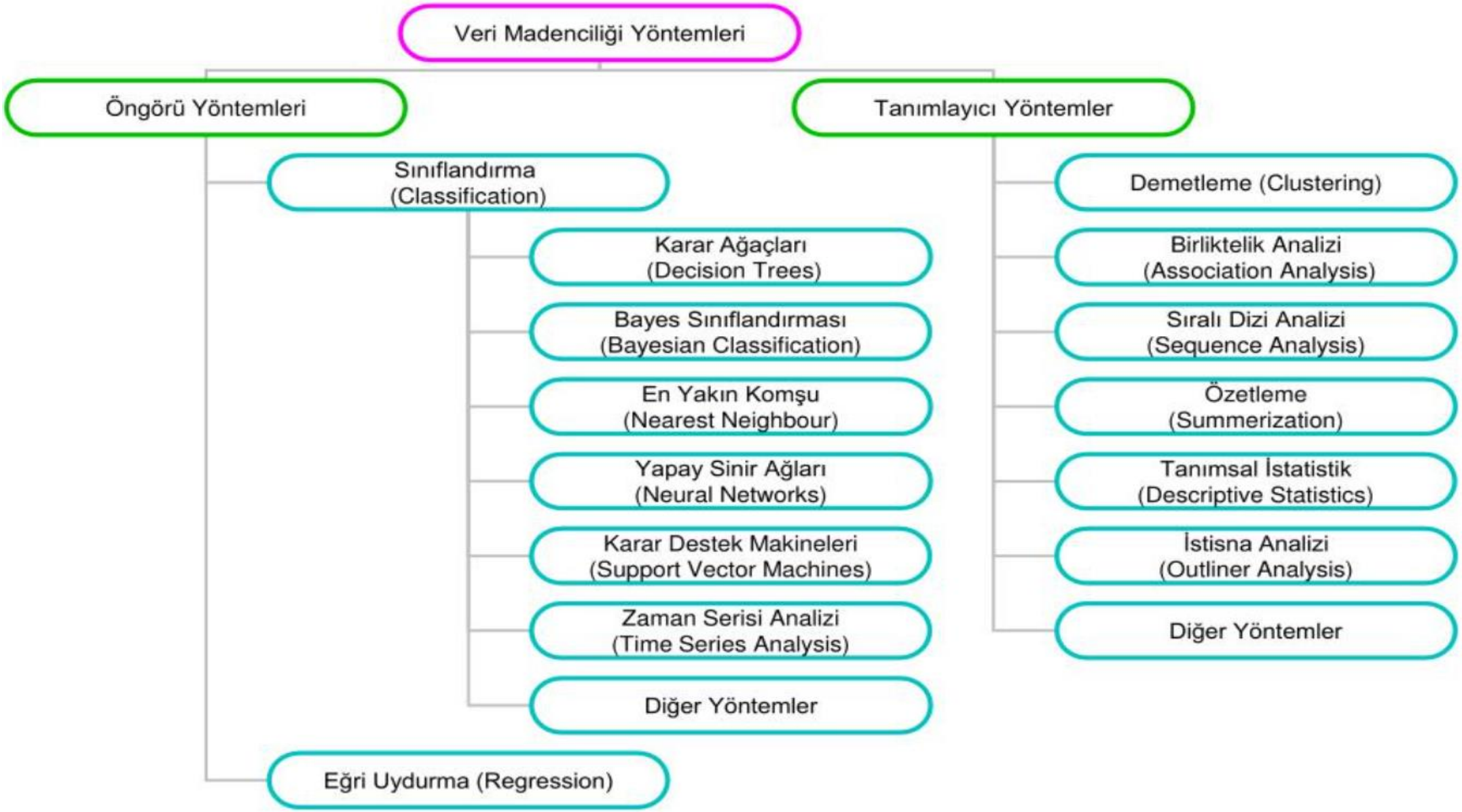
VERİ KAYNAKLARI

- Veri dosyaları
- Veritabanı kaynaklı veri kümeleri
  - ilişkisel veritabanları, veri ambarları
- Gelişmiş veri kümeleri
  - duraksız veri (data stream), algılayıcı verileri (sensor data)
  - zaman serileri, sıralı diziler (biyolojik veriler)
  - çizgeler, sosyal ağ (social networks) verileri
  - konumsal veriler (spatial data)
  - çoğul ortam veritabanları (multimedia databases)
  - nesneye dayalı veritabanları
  - www

VERİ MADENCİLİĞİ ALGORİTMALARI

- amaç: veriyi belli bir modele uydurmak
  - tanımlayıcı
    - En iyi müşterilerim kimler?
    - Hangi ürünler birlikte satılıyor?
    - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
  - kestirime dayalı
    - Kredi başvurularını risk gruplarına ayırma
    - Şirketle çalışmayı bırakacak müşterileri öngörme
    - Borsa tahmini
- seçim: veriye uyan en iyi modeli seçmek için kullanılan kriter
- arama: veri üzerinde arama yapmak için kullanılan teknik

VERİ MADENCİLİĞİ YÖNTEMLERİ



VERİ MADENCİLİĞİ İŞLEVLERİ

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
  - Danışmanlı (Gözetimli) öğrenme
  - Örüntü tanıma
  - Kestirim
- Eğri uydurma (Regression): Veriyi gerçel değerli bir fonksiyona dönüştürür.
- Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- İstisna Analizi (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

VERİ MADENCİLİĞİ İŞLEVLERİ

- Kümeleme (Clustering): Benzer verileri aynı grupta toplama
  - Danışmansız (Gözetimsiz) öğrenme
- Özetleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
  - Genelleştirme (Generalization)
  - Nitelendirme (Characterization)
- İlişkilendirme kuralları (Association Rules)
  - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

VERİ MADENCİLİĞİNDE TEMEL KAVRAMLAR

- Veri (Data)
- Enformasyon(Information)
- Bilgi (Knowledge)
- Bilgelik (Wisdom)

VERİ (DATA)

- Veri, oldukça esnek bir yapıdadır. Temel olarak varlığı bilinen, işlenmemiş, ham haldeki kayıtlar olarak adlandırılırlar. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. Ancak bu durum her zaman geçerli değildir. İşlenerek farklı bir boyut kazanan bir veri, daha sonra bu haliyle kullanılmak üzere kayıt altına alındığında, farklı bir amaç için veri halini koruyacaktır. Bu konuyu daha iyi açıklayabilmek için enformasyon kavramını incelemek gerekmektedir.

- a. Bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öğe.
- b. Bir sanat eserine veya bir edebî esere temel olan ana ilkeler: "Bir romanın verileri."
- c . Bilgi, data.
- d. Matematik: Bir problemde bilinen, belirtilmiş anlatımlardan bilinmeyeni bulmaya yarayan şey.
- e. Bilişim: Olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimli gösterimi.

ENFORMASYON (INFORMATION)

- Enformasyon, veri kavramının tanımından yola çıkıldığında, adreslemedeki ikinci safhadır. Yani verilerin ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış, işlenmiş halidir. Bu haliyle enformasyon, potansiyel olarak içinde bilgi barından bir veri halindedir.
- Belli bir alanda ve belli bir toplumda bilgi ve haberlerin yayılmasına olanak sağlayan araçların tümüne verilen isimdir.
- Enformasyon, genel olarak insanın dış dünyayla ilişkisinde, belirsizlik düzeyini azaltan her tür uyaran şeklinde tanımlanabilir. Daha özel olarak ise, formatlanmış ve yapılandırılmış veriler bütünü olarak tanımlanabilir.
- Yaygın anlamda enformasyon terimi, "haber" (ing. news, alm. nachrichf) veya mesaj terimiyle eşanlamlıdır.
- Veriler enformasyona dönüştürülerek kullanışlı hale getirilirler. Bu yönüyle enformasyon anlam katılmış verilerdir.

BİLGİ (KNOWLEDGE)

- Bilgi, bu süreçteki üçüncü aşamadır. Enformasyonun, bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir. Dolayısıyla bireyin algılama yeteneği, yaratıcılık, deneyim gibi kişisel nitelikleri de bu süreci doğrudan etkilemektedir.
- «İnsan aklının erebileceği olgu, gerçek ve ilkeler bütünü, malumat» olarak sözlüğümüzde tanımlanan bilgi, bilişim dilinde kurallardan yararlanarak kişinin veriye yönelttiği anlam demektir.
- Felsefi olarak ise insanların maddi ve toplumsal anlıksal etkinliğinin ürünü olarak tanımlanmaktadır.
- Enformasyonun daha yüksek biçimi olarak bilginin tüm modelleri altında yatan, bilginin ham maddelerinden onlara anlam eklenerek ortaya çıkarılması gerektiği düşüncesidir.
- Bilgiden, farklı enformasyon parçacıkları arasındaki ilişkiler anlaşılmalıdır. Örneğin bir kişiyi sadece bir T.C kimlik numarasının temsil edebileceği bilgisine sahip olunmalıdır.

BİLGELİK (WİSDOM)

- Bilgelik ulaşılmaya çalışılan noktadır ve bu kavramların zirvesinde yer alır. Bilgilerin kişi tarafından toplanıp bir sentez haline getirilmesiyle ortaya çıkan bir olgudur. Yetenek, tecrübe gibi kişisel nitelikler birer bilgelik elemanıdır.
- Bilgelik bilginin teferruatlı ve hassas kullanımını gerektirir.
- Neyin bilindiğinin (bilgi) ve en iyinin ne olduğunun (sosyal ve etnik faktörler) dikkate alınarak en uygun davranışın sergilenmesi demektir. Belirli bir alanı veya alanları anlamak için daha geniş ve genelleştirilmiş kuralları ve şemaları temsil etmesiyle bilgiden ayrılır.
- Bilgelik karar alma ve kararın uygulanması sırasında tecrübe edilir.

BİLGİ PİRAMİDİ