VERİ MADENCİLİĞİ
(Data Mining)
(Veri Madenciliğine Giriş)

Dr.Öğr.Üyesi Kadriye ERGÜN
kergun@balikesir.edu.tr

Ders Bilgileri

- BMM4202 Veri Madenciliği
- Ders ile ilgili duyurular
  - http://kergun.baun.edu.tr/
- Kaynaklar
  - İTÜ Veri Madenciliği Ders Notları, Şule Gündüz Öğüdücü
  - Veri Madenciliği Yöntemleri, Yalçın Özkan.
  - Veri Madenciliği: Kavram ve Algoritmaları, Gökhan Silahtaroğlu.
  - Veri Madenciliği (Kavram ve Teknikler), Aysan Şentürk.
  - RapidMiner ile Uygulamalı Veri Madenciliği, Ufuk Çelik, Eyüp Akçetin, Murat Gök.
- Başarı Notu
  - Vize (%40)
  - Final (%60)

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
  - Sınıflandırma
  - Kümeleme
  - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

VERİ MADENCİLİĞİNE GİRİŞ

İçerik

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin tarihçesi
- Veri madenciliğinin uygulama alanları
- Veri madenciliğinde temel kavramlar
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

Veri Madenciliği Giriş

- İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızlı artışı ve bilginin fazlalaşması sebebiyle öncelikle, genelde Veri Tabanlarında Bilgi Keşfi olarak adlandırılan yeni bir paradigma ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana Veri Madenciliği denilmektedir.

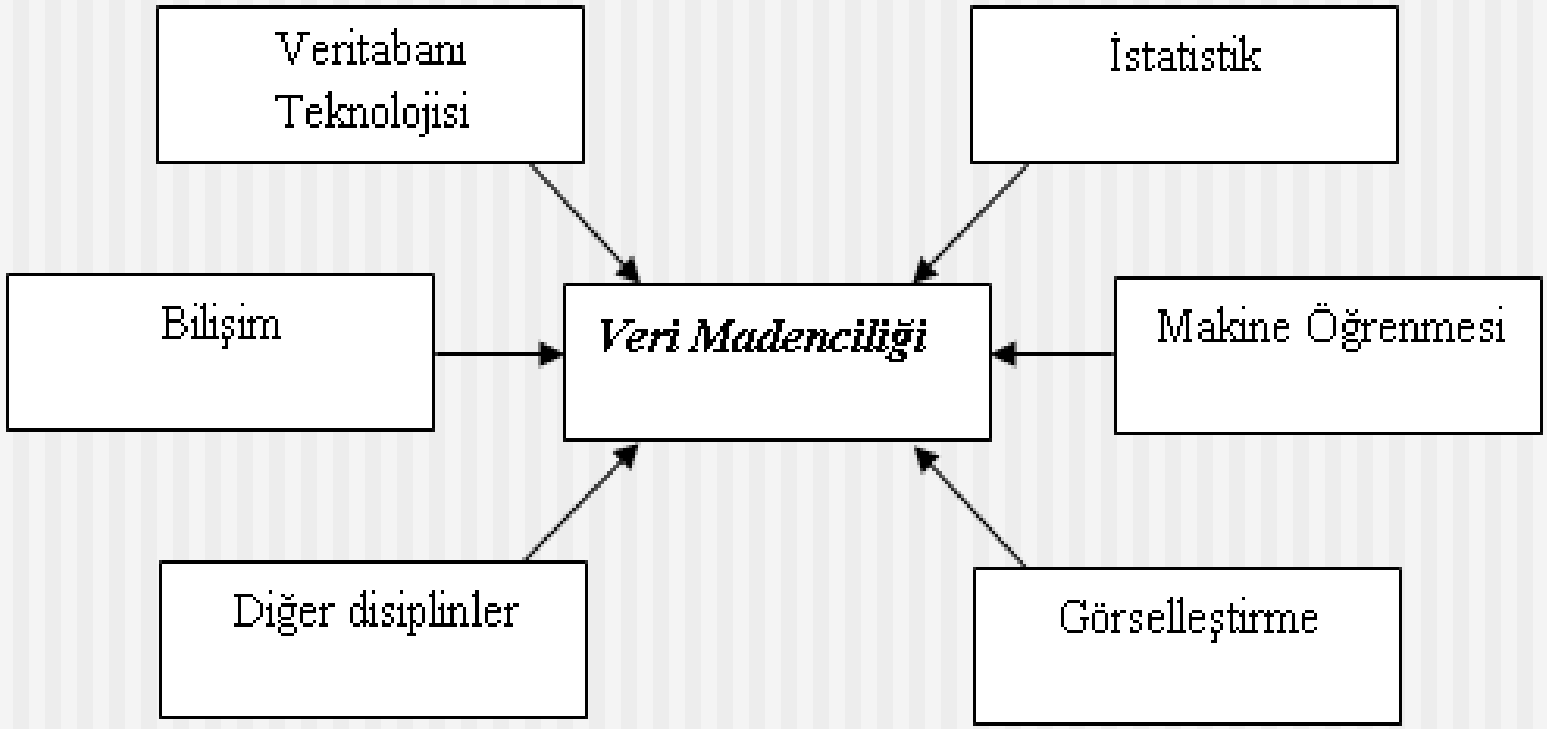
Veri Madenciliği Tanımları (1/2)

- Veri Madenciliği(Data Mining): Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır. *(Knowledge Discovery in Databases)*
- Daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır.
- Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.
- *Yapısal* veritabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir örüntülerin tanımlanması işlemidir.

Veri Madenciliği Tanımları (2/2)

- 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.
- Veri madenciliği çalışmaları, *sınıflandırma, ilişki kurma, kümeleme, regresyon, veri özetleme, değişikliklerin analizi, sapmaların tespiti* gibi belirli sayıda teknik yaklaşımları içerir.

Veri Madenciliği ile İlişkili Diğer Disiplinler



Veri Madenciliğinin Tarihçesi (1/4)

- Data FishingData Dredging: 1960
  - istatistikçiler
- Data Mining: 1990
  - veritabanı kullanıcıları, ticari
- Knowledge Discovery in Databases (KDD): 1989
  - Yapay zeka, makine öğrenmesi toplulukları
- Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction,...

Veri Madenciliğinin Tarihçesi (2/4)

- Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. O dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenilmiştir. Bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilmiştir.

Veri Madenciliğinin Tarihçesi (3/4)

- 1990'lı yıllara gelindiğinde Veri Madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atıldı. Bu camianın amacı, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktı. Bu noktadan sonra bilimadamları veri madenciliğine çeşitli yaklaşımlar getirmeye başladılar. Bu yaklaşımların kökeninde istatistik, makine öğrenmesi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar yatmaktaydı.

Veri Madenciliğinin Tarihçesi (4/4)

- İstatistik, süre gelen zaman içerisinde verilerin değerlendirilmesi ve analizleri konusunda hizmet veren bir yöntemler topluluğuydu. Bilgisayarların veri analizi için kullanılmaya başlamasıyla istatistiksel çalışmalar hız kazandı. Hatta bilgisayarın varlığı daha önce yapılması mümkün olmayan istatistiksel araştırmaları mümkün kıldı. 1990lardan sonra istatistik, veri madenciliği ile ortak bir platforma taşındı. Verinin, yığınlar içerisinden çekip çıkarılması ve analizinin yapılarak kullanıma hazırlanması sürecinde veri madenciliği ve istatistik sıkı bir çalışma birlikteliği içine girmiş bulundular.

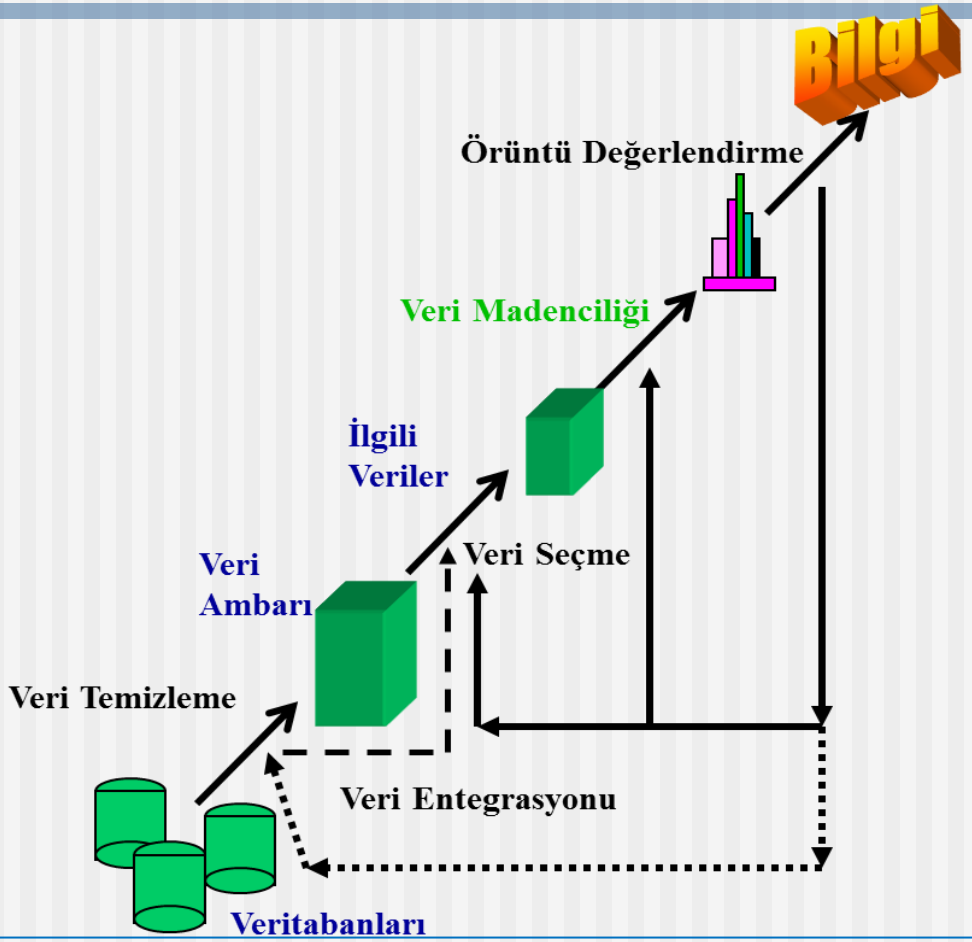
- Bunun yanısıra veri madenciliği, veritabanları ve makine öğrenimi disipliniyle birlikte yol aldı. Günümüzdeki Yapay Zeka çalışmalarının temelini oluşturan makine öğrenimi kavramı, bilgisayarların bazı işlemlerden çıkarsamalar yaparak yeni işlemler üretmesidir. Önceleri makineler, insan öğrenimine benzer bir yapıda inşa edilmeye çalışıldı. Ancak 1980lerden sonra bu konuda yaklaşım değişti ve makineler daha spesifik konularda kestirim algoritmaları üretmeye yönelik inşa edildi. Bu durum ister istemez uygulamalı istatistik ile makine öğrenim kavramlarını, veri madenciliği altında bir araya getirdi.

Bilgi Keşfi

- 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.

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

Bilgi Keşfi


Bilgi
Örüntü Değerlendirme
Veri Madenciliği
İlgili
Veriler
Veri Seçme
Veri
Ambarı
Veri Temizleme
Veri Entegrasyonu
Veritabanları

Bilgi Keşfinin Aşamaları

- 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 : Verinin veri madenciliği yöntemine göre hale dönüşümünü gerçekleştirmek
- Veri Madenciliği : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin uygulanması
- Örüntü Değerlendirme: Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumunu

Veri Madenciliği Uygulama Alanları

- 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

Veri Madenciliği Uygulama Alanları

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	• Terörle Mücadele • Kanun Yaptırımı • Vergi Kaçakçılarının Profilinin Çıkarılması

Uygulamalar

- 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?

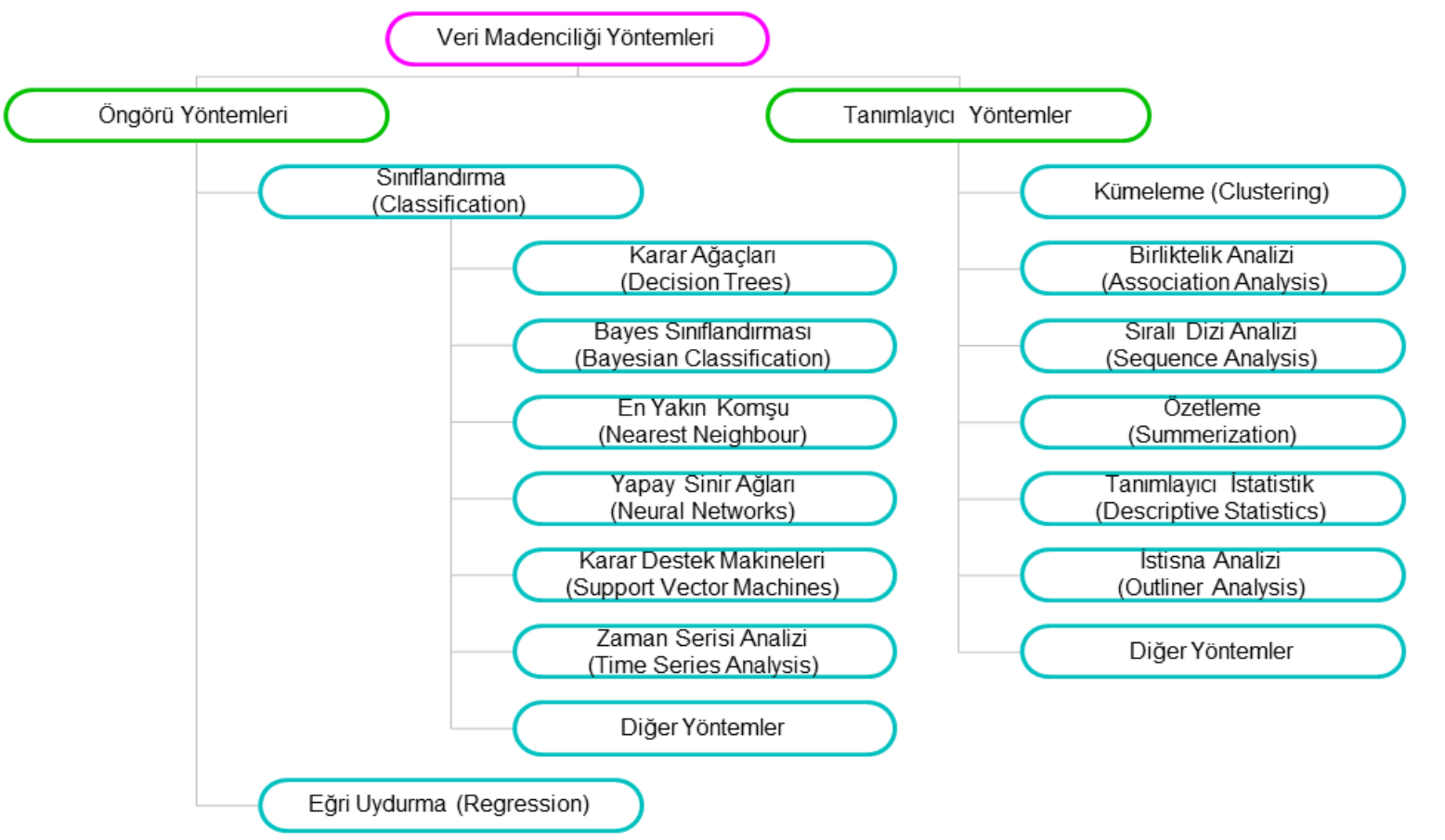
Veri Kaynakları

- 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

Veri Madenciliği Algoritmaları

- 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

Veri Madenciliği Yöntemleri



Veri Madenciliği İşlevleri (1/2)

- 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

Veri Madenciliği İşlevleri (2/2)

- 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.

Veri Madenciliğinde Temel Kavramlar

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

Veri (Data) (1/2)

- Veri kelimesi Latince'de "gerçek, reel" anlamına gelen "datum" kelimesine denk gelmektedir. "Data" olarak kullanılan kelime ise çoğul "datum" manasına gelmektedir. Her ne kadar kelime anlamı olarak gerçeklik temel alınsa da her veri her daim somut gerçeklik göstermez. Kavramsal anlamda veri, kayıt altına alınmış her türlü olay, durum, fikirdir. Bu anlamıyla değerlendirildiğinde çevremizdeki her nesne bir veri olarak algılanabilir.

Veri (Data) (2/2)

- 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.

Bilgi (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.

Bilgelik (Wisdom)

- 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.
- 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 bilginin teferruatlı ve hassas kullanımını gerektirir.
- Bilgelik karar alma ve kararın uygulanması sırasında tecrübe edilir.

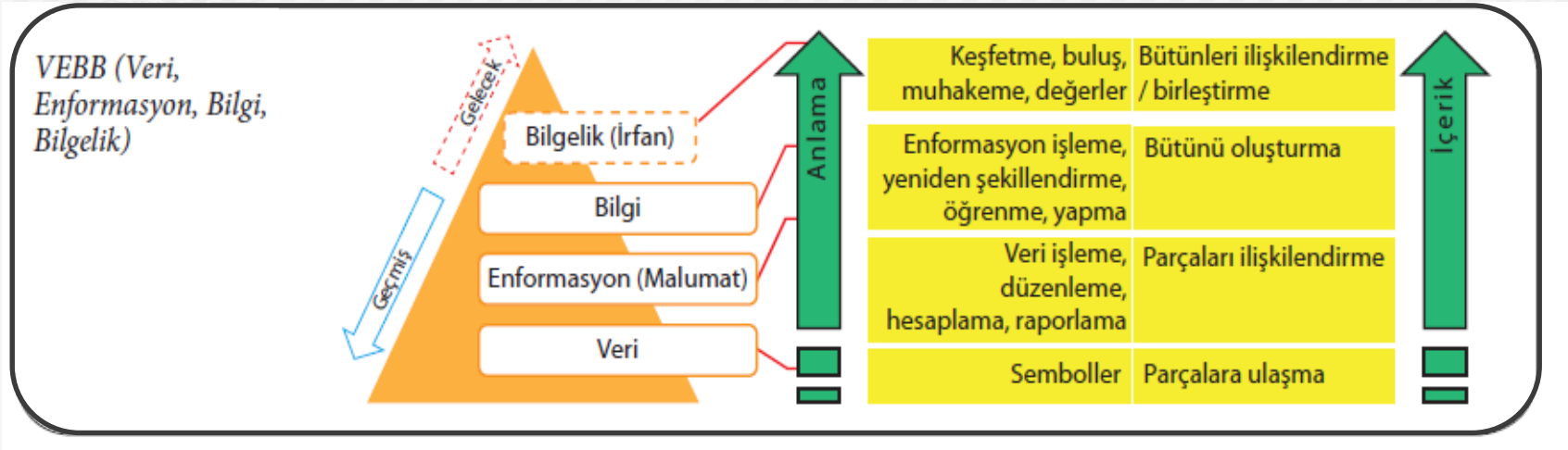
Bilgi Piramidi

Bilgi piramidi hiyerarşisi incelenecek olursa bilgiye ulaşmanın kolay olmadığı görülür. Yeni teknolojiler enformasyona ulaşmayı daha kolay hale getirmektedir buna karşın, doğru ve güvenilir, yeterli enformasyona ulaşmak zordur. Eğer ulaşılan enformasyon hatalı ya da eksik ise doğal olarak elde edilecek bilgi ve uygulama sonuçları da sağlıklı olmayacaktır.

Bilgi Piramidi

- Bilgeliğe ulaşabilmek için geçilmesi gereken yollar bilgi piramidinin aşamalarına benzemektedir. Veriden bilgeliğe kadar olan yükselme sırasında, gözlemlerden iletişime varan boyutlarda değişiklik gerekmekte ve bilge olana sağlanacak değerin buradan çıkacağı varsayılmaktadır. Bilgelik için gereken şartlara bakıldığında ise, hem bağlam hem de anlayış açısından, gerçekleştirilmesi gereken bir bakış açısının ortaya çıktığı görülmektedir. Bilgelik, deneyimlerin düşünme becerilerine dahil edilmesi ile oluşmaktadır

Veri, Enformasyon, Bilgi, Bilgelik Piramidi (Bilgi Piramidinin Geliştirilmiş Hali)



Kaynak:*Temel Bilgi Teknolojileri-I , AÖF Yayını*

Veri Ambarı

- **Veritabanı:** birbirleriyle ilişkili bilgilerin depolandığı alanlardır.
- **Veri Ambarı**: ilişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işlemsel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar.
- **Data Mart:** veri ambarlarının alt kümeleridir. Veri ambarları bir iş probleminin tamamına yönelik bir bakış sağlarken, data mart'lar sadece belli bir kısma bakış sağlarlar. Veri pazarları ile veriye hızlı erişim sağlayabiliriz. İkinci olarak, verinin gruplanmamış yapıda olması ve farklı iş birimlerinin farklı verileri görmesidir. Bu da bize gereksiz bir iş yükü ve güvenlik sorununa neden olmaktadır. İşte tam bu noktada, veri pazarları konuya, bölümlere uygun, veri ambarının küçük bir kopyası halinde çözüm sunmaktadır.

Veri Ambarı

- Amaca yönelik
- Birleştirilmiş
- Zaman değişkenli
- Değişken değil

Veri Ambarları: Amaca Yönelik

- Müşteri, ürün, satış gibi belli konular için düzenlenebilir.
- Verinin incelenmesi ve modellenmesi için oluşturulur.
- Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar.

Veri Ambarları: Birleştirilmiş

- Veri kaynaklarının birleştirilmesiyle oluşturulur.
  - Canlı veri tabanları, dosyalar.
- Veri temizleme ve birleştirme teknikleri kullanılır.
  - Değişik veri kaynakları arasındaki tutarlılık sağlanır.

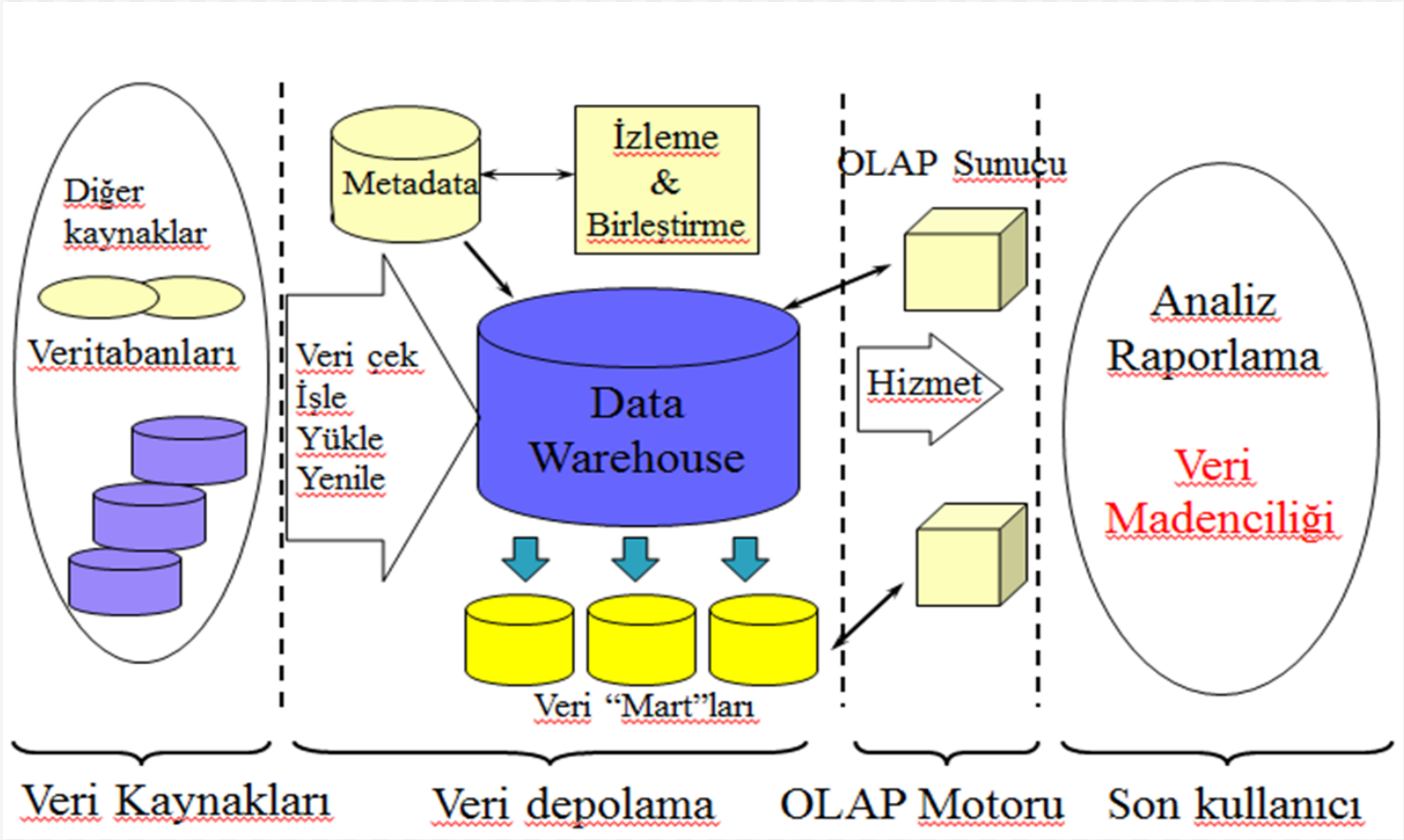
Veri Ambarları: Zaman Değişkenli

- Zaman değişkeni canlı veri tabanlarına göre daha uzundur.
  - Canlı veri tabanları: Güncel veriler bulunur (en çok geçmiş 1 yıl)
  - Veri ambarları: Geçmiş hakkında bilgi verir (geçmiş 5-10 yıl)

Veri Ambarları: Değişken Değil

- Canlı veritabanlarından alınmış verinin fiziksel olarak başka bir ortamda saklanması.
- Canlı veritabanlarındaki değişimin veri ambarlarını etkilememesi.

Veri Ambarı Mimarisi

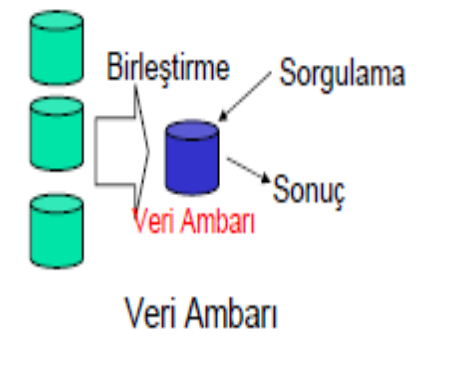


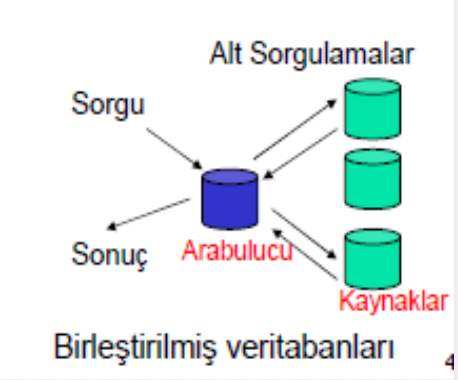
Veri Kaynakları

- İki yaklaşım:
  - sorgulamalı
  - veri ambarı

Veri Ambarı & Birleşmiş Veritabanları

- Veritabanlarının birleştirilmesi:
  - Farklı veritabanları arasında bir arabulucu katman
  - Sorgulamalı
    - Bir sorgulamayı her veritabanı için alt sorgulamalara ayır
    - Sonucu birleştir
- Veri ambarı:
  - Veri daha sonra kullanılmak üzere birleştirilip veri ambarında saklanıyor.





Veri Madenciliği & OLAP

- OLAP (On-Line Analytical Processing)
  - Veri ambarlarının işlevi
  - Veriyi inceleme ve karar verme
  - OLTP (On-Line Transaction Processing) saatler sürebilen işlemler
- OLAP avantajları
  - Daha geniş kapsamlı sonuçlar
  - Daha kısa süreli işlem
- OLAP dezavantajları
  - Kullanıcı neyi nasıl soracağını bilmesi gerekiyor
  - Genelde veriden istatistiksel inceleme yapmak için kullanılır.
- OLAP NE sorusuna cevap verir, veri madenciliği NEDEN sorusuna cevap verir.

Veri Madenciliğinde Sorunlar (1/3)

- Gizlilik ve sosyal haklar
  - Kişilere ait verilerin toplanarak, kişilerden habersiz ve izinsiz olarak kullanılması
  - Veri madenciliği yöntemleri ile bulunan sonuçların izinsiz olarak açıklanması (/paylaşılması)
  - Gizlilik ve veri madenciliği politikalarının düzenlenmesi
- Kullanıcı Arabirimi
  - Görüntüleme
    - Sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi
    - Bilginin sunulması
  - Etkileşim
    - Veri madenciliği ile elde edilen bilginin kullanılması
    - Veri madenciliği yöntemine müdahale etmek
    - Veri madenciliği yönteminin sonucuna müdahale etmek
- Veri madenciliği yöntemi
- Başarım ve ölçeklenebilirlik

Veri Madenciliğinde Sorunlar (2/3)

- Veri madenciliği yöntemi
  - Farklı tipte veriler üzerinde çalışabilme
  - Farklı seviyelerde kullanıcı ile etkileşim halinde olabilme
  - Uygulama ortamı bilgisini kullanabilme
  - Veri madenciliği ile elde edilen sonucu anlaşılır şekilde sunabilme
  - Gürültülü ve eksik veri ile çalışabilme (ve iyi sonuç verebilme)
  - Değişen veya eklenen verileri kolayca kullanabilme
  - Örüntü değerlendirme: önemli örüntüleri bulma

Veri Madenciliğinde Sorunlar (3/3)

- Başarım ve ölçeklenebilirlik
  - Kullanabilirlik ve ölçeklenebilirlik
    - Zaman karmaşıklığı ve yer karmaşıklığı kabul edilebilir
    - Örnekleme yapabilme
  - Paralel ve dağıtık yöntemler
    - Artımlı veri madenciliği
    - Parçala ve çöz
- Veri kaynağı