

Veri Madenciliđi

Doç. Dr. Mustafa ULAŞ
mustafaulas@firat.edu.tr



INDEX

- Bilimin Evrimi
- Veri Tabanı Teknolojilerinin Gelişimi
- Veri Madenciliğinin Tarihçesi
- Veri Madenciliği Kullanım Alanları
- Veri Madenciliğine Giriş

Bilimlerin Evrimi

- 1600'den önce ampirik bilim
- 1600-1950'ler, teorik bilim
 - Her disiplin teorik bir bileşen geliştirmiştir. Teorik modeller genellikle deneyleri motive eder ve anlayışımızı genelleştirir.
- 1950'ler-1990'lar, hesaplamalı bilim
 - Son 50 yılda çoğu disiplin üçüncü bir hesaplamalı dal geliştirdi (örneğin ampirik, teorik ve hesaplamalı ekoloji veya fizik veya dilbilim).
 - Hesaplamalı Bilim geleneksel olarak simülasyon anlamına geliyordu. Karmaşık matematiksel modeller için kapalı form çözümler bulma konusundaki beceriksizliğimizden kaynaklandı.
- 1990-şimdi, veri bilimi
 - Yeni bilimsel araçlardan ve simülasyonlardan elde edilen veri seli
 - Petabaytlarca veriyi çevrimiçi olarak ekonomik bir şekilde depolama ve yönetme yeteneği
 - Tüm bu arşivleri evrensel olarak erişilebilir kılan İnternet ve Bilgi İşlem Grid'i
 - Bilimsel bilgi, yönetim, satın alma, organizasyon, sorgulama ve görselleştirme görevleri veri hacimleriyle neredeyse doğrusal olarak ölçeklenir. Veri madenciliği yeni ve büyük bir zorluktur!
- Jim Gray ve Alex Szalay, Dünya Çapında Teleskop: Çevrimiçi Bilim için Bir Arketip, Comm. ACM, 45(11): 50-54, Kasım 2002

Veri Tabanı Teknolojilerinin Gelişimi

- 1960'lar:
 - Veri toplama, veri tabanı oluşturma, IMS ve ağ DBMS
- 1970'ler:
 - İlişkisel veri modeli, ilişkisel DBMS uygulaması
- 1980'ler:
 - RDBMS, gelişmiş veri modelleri (genişletilmiş-ilşkisel, OO, tümdengelimli vb.)
 - Uygulama odaklı DBMS (mekansal, bilimsel, mühendislik vb.)
- 1989 'da
 - Knowledge Discovery in Databases (KDD):
 - Yapay zeka, makine öğrenmesi toplulukları
- 1990'lar:
 - Veri madenciliği, veri ambarı, multimedya veri tabanları ve Web veri tabanları
- 2000'ler
 - Akış veri yönetimi ve madenciliği
 - Veri madenciliği ve uygulamaları
 - Web teknolojisi (XML, veri entegrasyonu) ve küresel bilgi sistemleri

Veri Madenciliğinin Tarihçesi

- Günümüzde bilgi çok hızlı bir şekilde çoğalmaktadır ve bu bilgilerin içerisinde işe yarayacak ve yaramayacak olanları ayırt etmek zorlaşmaktadır. Bu bilgi yığınları arasından, işe yarayacak olanlarını elde etme işi için kullanılan yöntem ise veri madenciliğidir.

Veri Madenciliğinin Tarihçesi

- 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

- 1990'larda içindeki veri miktarı katlanarak artan veri tabanlarından, faydalı bilgilerin nasıl bulunabileceği düşünülmeye başlanmıştır.
 - 1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin KDD ile ilgili temel tanım ve kavramları ortaya koyması ile süreç daha da hızlanmıştır.
 - 1992 yılında veri madenciliği için ilk yazılım gerçekleştirilmiştir.
 - Geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktaydı.
 - Bu noktadan sonra bilim adamları 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

- 2000'li yıllarda veri madenciliği sürekli gelişmiş ve hemen hemen tüm alanlara uygulanmaya başlanmıştır. Alınan sonuçların faydaları görüldükçe bu alana yönelik çalışmalar ve uygulamalar sürekli artmaktadır.
- 2010'lu yıllar Büyük Veri kavramının gittikçe yaygınlaştığı zaman dilimi olarak tarihte yerini almıştır. Buna en büyük etkiyi sosyal medya sitelerinin yaygınlaşması ve mobil yaşam sağlamıştır.
- 2020 artık; Nesnelerin İnterneti, Bulut Bilişim, Derin Öğrenme gibi disiplinlerle birlikte artan veri çeşitliliğine paralel olarak veriyi işleme algoritmaları da çeşitlenmiştir.

Veri Madenciliğinin Tarihçesi

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük, statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veri tabanları, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı? Boston'a aşağı inebilecek"	OLAP, Çok Boyutlu Veri tabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir? Neden?"	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veri tabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft, vs.	Geleceğe dönük, proaktif enformasyon dağıtımı

Veri Madenciliğinin Kullanıldığı Alanlar

- Büyük hacimde veri bulunan her yerde veri madenciliğini kullanmak mümkündür. Günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır.
- Örneğin, eğitim, biyoloji, finans, borsa, genetik, sağlık, sigorta, endüstri, istihbarat vb. birçok dalda başarılı uygulamaları görülmektedir.
- Son 30 yıldır tüm bu belirtilen sektörlerde başarılı Veri Madenciliği uygulamalarının yapıldığı bilinmektedir.
- Veri madenciliğini kullanan alanlar ve kullandıkları uygulamalar şöyle belirtilebilir:

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,
- Müşteri değerlendirmesi,
- Satış tahmini,
- Müşteri dağılımı,
- Çeşitli pazarlama kampanyaları,
- Pazarlama stratejilerinin oluşturulması,
- Çapraz satış analizleri

Bankacılık, Sigortacılık ve Borsa

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi,
- Müşteri dağılımı,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,
- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Hisse senedi fiyat tahmini,
- Hileli hesapların ve dolandırıcılıkların ortaya çıkarılması

Perakendecilik

- Satış noktası veri analizleri,
- Alışveriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin optimizasyonu.

Eğitim

- Eğitsel Veri Madenciliği
- Geleneksel Sınıf Ortamında Veri Madenciliği
- Uzaktan Eğitimde Veri Madenciliği
- Üniversiteye Giriş ve Öğrenci Kabulü İle İlgili Çalışmalar
- Öğrenci Profilinin Tahmin Edilmesi İle İlgili Yapılan Çalışmalar
- Öğrenci Performansının Belirlenmesi İle İlgili Yapılan Çalışmalar
- Öğretmen Performansının Tahmin Edilmesi
- Müfredat Geliştirme İle İlgili Yapılan Çalışmalar
- Öğrencinin Dersten Geçmesini Tahmin Etme
- Öğrencinin Ders Seçimini Tahmin Etme
- Öğrencilerin Üniversiteyi Kazanma Tahmin Etme

Telekomünikasyon

- Kalite ve iyileştirme analizleri,
- Hatların yoğunluk tahminleri,
- Web sitesi ziyaretçilerinin profil analizi.

Bilim ve Mühendislik

- Laboratuvar veya bilgisayar ortamında sistemlerin simülasyonu ve analizi sürecinde üretilen yüksek miktarda bilimsel verilerin anlamlandırılması

Sağlık, Biyoloji ve Genetik

- Gen araştırmaları,
- Yeni virüs türlerinin keşfi ve sınıflandırılması,
- Hastalıkların özelliklerinin belirlenerek teşhislerin kolaylaştırılması,
- Birlikte kullanılan ilaçların yan etkilerinin araştırılması,
- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tedavi sürecinin belirlenmesi,
- Tıbbi teşhis.

Kriminoloji

- Suç eğilimlerinin tanımlanması,
- İstihbarat birimleri.

Veri Madenciliği Kullanım Alanları

- Veri madenciliğinin kullanıldığı yukarıda belirtilen alanlarda araştırmacılara, firma yöneticilerine ve idarecilerine sağladığı yararlar şu şekilde sıralanabilir:
 - Tıp, verilerin en çok tutulduğu alanların başında bulunmaktadır. Özellikle son yıllarda genetiğin inanılmaz hızda ilerlemesi sonucunda oluşan gen haritaları ile hastalıklar sınıflandırılmaya başlanmıştır. Hangi genlere sahip bireylerin hangi hastalıklara yakalanma olasılığı olduğuna dair çalışmaları gerçekleştirmek artık mümkündür.
 - Gen bilimi ile ortaklaşa çalışma sayesinde hangi bireylerin suç işlemeye eğilimli olduklarına dair tahminleme çalışmaları yapılarak bu olayların daha ortaya çıkmadan önlenmesinden, kullanıcıların yazım karakterlerinden yola çıkarak birçok olasılığın hesaplanmasına kadar pek çok kriminoloji bilgisi elde edilebilir.
 - Simülasyon ortamlarındaki verileri kullanarak mühendislik alanında, üretim veya problem çözme konularında öngörüler oluşturulabilir ve çözümler üretilebilir.

Veri Madenciliği Kullanım Alanları

- Veri madenciliği algoritmalarıyla yapılan değerlendirmeler, çok uzun zamanlar alabilecek deney ve test olaylarına gerek kalmadan, güvenilir ve daha kısa zaman alan sonuçlar elde etme imkanı sağlayabilir.
- Bankacılık faaliyetlerinde, küçük işletmelere yönelik olarak makine ve donanım satışı yapan dağıtıcı firmalarla ortak hareket ederek oluşturulacak satış paketleri ile pazarlama stratejileri geliştirilebilir.
- Mevcut müşterilerin yöneticiler tarafından daha iyi tanınmasını sağlayabilir.
- Özellikle finans sektöründe mevcut müşteriler bölümlere ayrılıp, kredi risk davranış modelleri oluşturularak, yeni başvuruda bulunan müşterilere karşı riskin minimize edilmesi sağlanabilir.
- Mevcut müşterilerin ödeme performansları incelenip kötü ödeme performansı gösteren müşterilerin ortak özellikleri belirlenerek, benzer özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulabilir.

- Ülkemizde en çok kullanıldığı alanlar bankacılık, sigortacılık ve borsa olarak sıralanabilir. Kaynaklar incelendiğinde ise veri madenciliğinin en çok kullanıldığı alanlar olarak tıp, biyoloji ve genetik olarak görülmektedir.

Veri Madenciliđi Giriř

Doç. Dr. Mustafa ULAř
mustafaulas@firat.edu.tr

INDEX

- Neden Veri Madenciliği?
- Veri madenciliği sistemlerinin sınıflandırılması
- En popüler 10 veri madenciliği algoritması
- Veri madenciliğinde temel sorunlar
- Veri madenciliği işlevselliği

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

- 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 veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır.
- Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.
- *Yapısal* veri tabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir örüntülerin tanımlanması işlemidir.

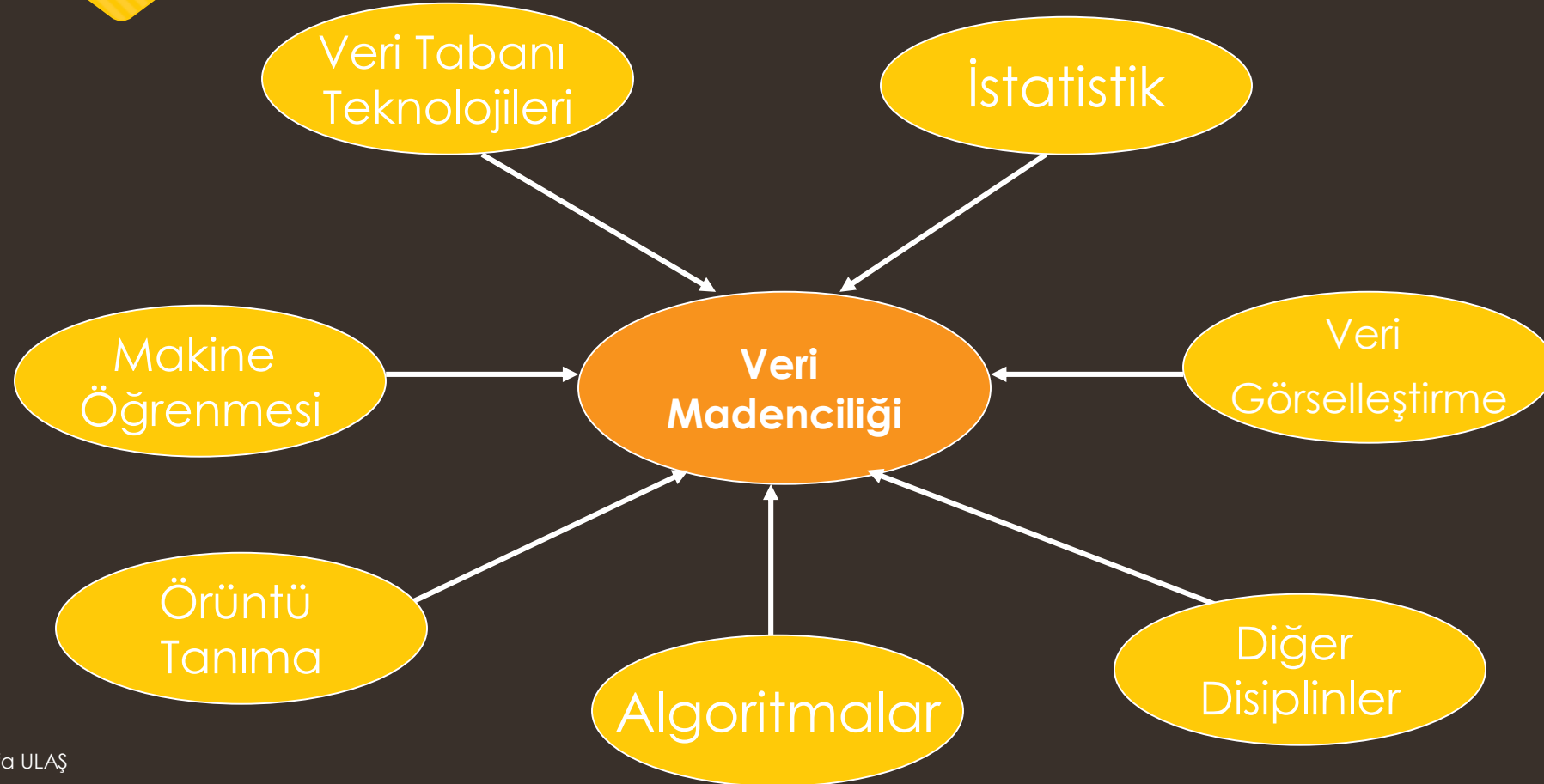
Veri Madenciliği Tanımları

- 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.

Neden Veri Madenciliği?

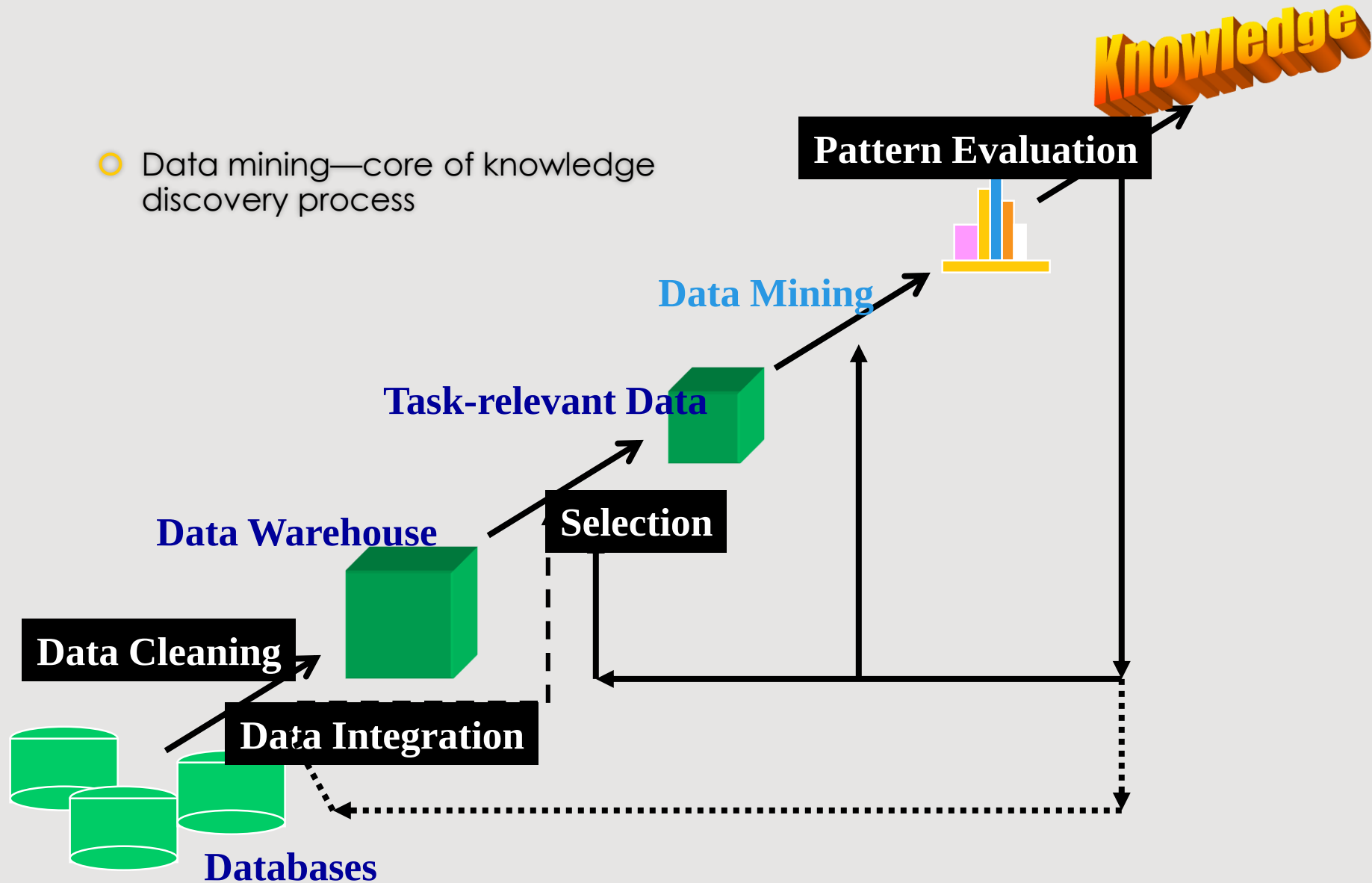
- Verilerin Aşırı Büyümesi: terabaytlardan petabaytlara
 - Veri toplama ve veri kullanılabilirliği
 - Otomatik veri toplama araçları, veri tabanı sistemleri, Web, bilgisayarlı toplum
 - Bol miktarda verinin ana kaynakları
 - İş: Web, e-ticaret, işlemler, hisse senetleri,...
 - Bilim: Uzaktan algılama, biyoinformatik, bilimsel simülasyon,...
 - Toplum ve herkes: haberler, dijital kameralar, YouTube
 - Veri içinde boğuluyoruz ama bilgiye aç kalıyoruz!
 - “İhtiyaç buluşun anasıdır”—Veri madenciliği—Çok büyük veri kümelerinin otomatik analizidir.

Veri Madenciliği ile İlişkili Diğer Disiplinler



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 verilen addır.
- 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 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 İşlevleri

- 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

- 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 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
 - Text databaseler

Veri Madenciliğinde Temel Sorunlar

- Madencilik metodolojisi
 - Biyo, akış, Web gibi farklı veri türlerinden farklı türde bilgilerin çıkarılması
 - Performans: verimlilik, etkililik ve ölçeklenebilirlik
 - Desen değerlendirmesi: ilginçlik sorunu
 - Arka plan bilgisinin dahil edilmesi
 - Gürültüyü ve eksik verileri işleme
 - Paralel, dağıtılmış ve artımlı madencilik yöntemleri
 - Keşfedilen bilginin mevcut bilgiyle entegrasyonu: bilgi füzyonu

Veri Madenciliğinde Temel Sorunlar

- Kullanıcı etkileşimi
 - Veri madenciliği sorgu dilleri ve geçici madencilik
 - Veri madenciliği sonuçlarının ifadesi ve görselleştirilmesi
 - Çoklu soyutlama seviyelerinde etkileşimli bilgi madenciliği
- Uygulamalar ve sosyal etkiler
 - Etki alanına özgü veri madenciliği ve görünmez veri madenciliği
 - Veri güvenliğinin, bütünlüğünün ve gizliliğinin korunması

Veri Madenciliği Yöntemleri

