Veri Madenciliği

Dr. Arş. Gör. Caner Erden

mail: cerden@sakarya.edu.tr

Tel: +90 264 295-7323

Görüşme Zamanları: Salı 10:00-12:00 ve Çar 14:00-16:00

Yer: M5 5220

Derse Giriş

Ders Kodu: ENM 424

Sınıf: 5102

Ders Saati: 1. Öğr Çar 10:00-13:00 2. Öğr 17:00-20:00

Google Class Kodu: jvobjs

jvobjs

Veri Madenciliği Endüstri Mühendisliği

Tanışalım

- Hangi sektör? (Mühendislik, veri bilimi, yönetim?)
- İstatistik ve Matematik (İyi, Orta, Kötü?)
- Makine öğrenmesi deneyimi? (Var, yok)
- Programlama bilgisi (Kullanıyorum, Biraz Bilgim Var, Sıfırım)

Ödevler

- 2 adet ödev
- 1 adet performans ödevi

Performans Ödevi

1. İlginç bulduğunuz bir problem üzerine veri madenciliği uygulaması
2. En fazla 3 kişilik gruplar olabilir.
3. Gruplar 19 Şubat Tarihine kadar bildirilmelidir.

Performans Ödevi Konuları

1. Regresyon Analizi
2. Kural Tabanlı Sınıflandırıcılar
3. Sınıflandırma Algoritmaları
4. Kümeleme Algoritmaları
5. Özellik İndirgeme
6. Yapay Sinir Ağları
7. Kaggle tipi bir alanda yarışmaya katılmak ve en az ortalama bir derece almak.

Kaggle Hakkında

1. Kaggle'da bir hesap açınız.
2. Kaggle'ı inceleyiniz, nasıl kullanıldığı hakkında araştırma yapınız.
3. Yarışmalara göz gezdiriniz.
4. Veri setlerini inceleyiniz.

Tarihler

Çalışma	Tarih
Ödev 1	12 Şub 20 - 19 Şub 20 (2. Hafta)
Kısa Sınav 1	26 Şub 20 (4. Hafta)
Ödev 2	11 Mar 20 - 18 Mar 20 (7. Hafta)
Ara Sınav	30 Mar 20 - 3 Nis 20 (9. Hafta)
Kısa Sınav 2	29 Nis 20
Final	11-24 May 2020 (15. Hafta)

Ödev Puanlandırma

Ödev süresini 1-2-3-4 gün geciktirmeye ceza uygulancak olup daha geç gönderilmesine izin verilmeyecek. (0.8, 0.6, 0.5, 0.3, 0)

Değerlendirme Sistemi

YARIYIL İÇİ ÇALIŞMALARI	SIRA	KATKI YÜZDESİ
AraSinav	1	40
KisaSinav	1	15
Odev	1	7
KisaSinav	2	15
Odev	2	8
Proje	1	15
Finalin Başarıya Oranı		50

Neden Python

1. Yazılması, okunması, kod geliştirmesi en kolay dil.
2. En çok kullanılan dil (Video)
3. Ücretsiz, açık kaynak.
4. Çok amaçlı, çok güçlü.

Dersin Amacı

Bilgisayar programlamayı öğretmek ve uygulamak değil programlamayı kullanarak veri madenciliği uygulaması geliştirmektir.

> Araba motorunun nasıl çalıştığını bilmeyebiliriz ancak yine de iyi bir sürücü olabiliriz.

Yazılım Kurulumu

Bu adresten Anaconda kurulumunun nasıl yapıldığını öğrenebilirsiniz.

Yükleme Adımları

1- Anaconda.com/downloads adresini ziyaret edin

2- Windows Seçin .exe yükleyicisini indirin.

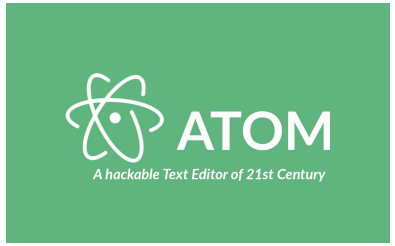
3- .Exe yükleyicisini açın ve çalıştırın.

4- Anaconda komut istemini açın ve bir Python kodu çalıştırın.

5- Orange Paketini yükleyin

Diğer Çalışma Ortamları

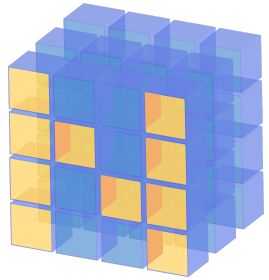
- Jupyter (Interactive Code Editor)
- Visual Studio Code (Editor)
- Pycharm (IDE)
- Atom (Editor)
- Sublime Text (Editor)
- Kaggle (Online)
- Google Colab (Online)
- Spyder (IDE)



Sublime Text

Jupyter Defterleri (Jupyter Notebooks)

Python_cheatsheet.pdf



Kullanılacak Paketler

- Python
  - Numpy
  - pandas
  - scikit-learn
  - Scipy
  - matplotlib
- Orange Aracı

Veri Madenciliği ve Python

Numpy – Bilimsel hesaplama için gerekli bir paket. Scikit-learning'in giriş verileri için kullandığı birincil veri formatı olan dizilerle çalışmak için çok yönlü bir yapı içerir.

Matplotlib – Python'da veri görselleştirme için temel paket. Bu paket basit dağılım grafiklerinden 3 boyutlu kontur grafiklerine kadar her şeyin oluşturulmasını sağlar.

Scipy – python'da istatistik için bir araç koleksiyonu. İstatistikler, regresyon analizi işlevlerini scipy modülü ile gerçekleştirebiliriz Kaynak.

scikit-learn – makine öğrenmesi için geliştirilmiştir. NumPy, SciPy ve matplotlib üzerine inşa edilen bu kütüphane, makine öğrenmesi ve sınıflandırma, regresyon, kümeleme ve boyutluluk azaltma gibi istatistiksel modelleme için birçok verimli araç içerir.

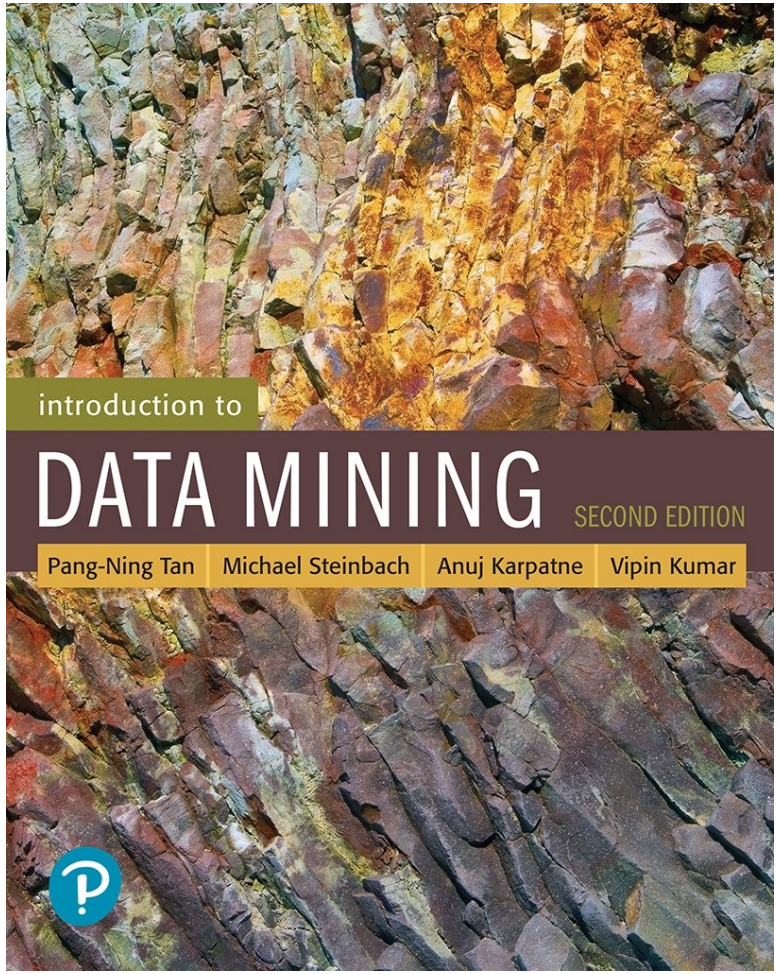
pandas - Veri toplama ve hazırlamada yaygın olarak kullanılır.

Ders Kitapları

Derste kullanılan görsel ve sunumlar Introduction to Data Mining (Second Edition)

Ana Kitap

1. Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, ve Vipin Kumar. *Introduction to Data Mining*. 2 edition. NY NY: Pearson, 2018.



Yardımcı Kitaplar

1. Han, Jiawei, Micheline Kamber, ve Jian Pei. *Data Mining: Concepts and Techniques, Third Edition*. 3 edition. Haryana, India; Burlington, MA: Morgan Kaufmann, 2011.
2. Kantardzic, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. 3 edition. Wiley-IEEE Press, 2019.
3. Leskovec, Jure, Anand Rajaraman, ve Jeffrey David Ullman. *Mining of Massive Datasets*. 2 edition. Cambridge: Cambridge University Press, 2014.
4. Sumathi, S., ve S. N. Sivanandam. *Introduction to Data Mining and Its Applications*. 2006 edition. Berlin ; New York: Springer, 2006.
5. Tan, Pang-Ning, Michael Steinbach, ve Vipin Kumar. *Introduction to Data Mining*. 1 edition. Boston: Pearson, 2005.
6. Akkucuk, Ulas. *Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları*. İstanbul: Yalın Yayıncılık, 2011.

Uygulama Kitapları

1. Kane, Frank. *Hands-On Data Science and Python Machine Learning*. Birmingham Mumbai: Packt Publishing - ebooks Account, 2017.
2. Porcu, Valentina. *Python for Data Mining Quick Syntax Reference*. 1st ed. edition. New York, NY: Apress, 2018.
3. "Python Machine Learning - Third Edition". Erişim 23 Aralık 2019. https://www.packtpub.com/data/python-machine-learning-third-edition.

Ders Asistanları

...

Derse Devamlılık

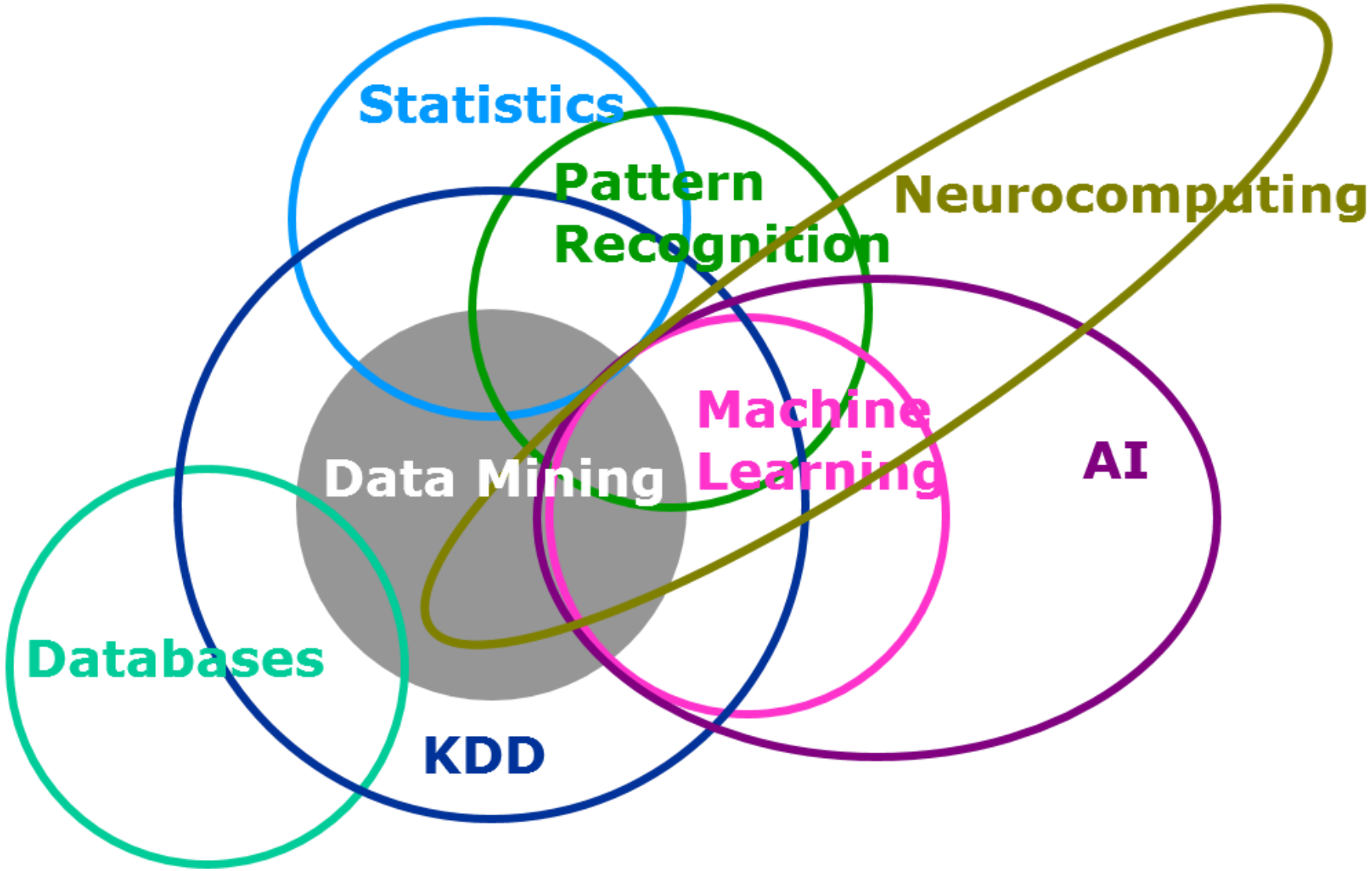
%70 oranında derse devam ve katılım gerekmektedir. Her dersin başında yoklama alınacaktır. Derse katılım not değerlendirilmesine katılmayacaktır.

Öğrenciden Beklentimiz

Derse

- zamanında (ne geç ne erken)
- hazırlıklı (ders metaryallerini inceleyerek)
- öğrenmeye istekli

geliniz :)


Statistics
Pattern Recognition
Neurocomputing
Machine Learning
AI
Data Mining
Databases
KDD

Veri Madenciliği

Veri yeni petrol kaynağı olarak görülüyor.

Kaynak

Ücretsiz Veri Kaynakları

- UCI The UCI Machine Learning Repository (476)
- Kaggle (19,515)
- Scikit-learn
- Drivendata.org
- FiveThirtyEight Kaynak

Temel Kavramlar

Veri: araştırma sonucunda, işlem sonucunda elde ettiğimiz en küçük ham bilgidir. Veriyi ölçüm yaparak, deney yaparak, sayarak, gözlemleyerek elde edebiliriz.

Yapısal Olmayan Veri: Binlerce e-kitap verisi, milyonlarca internet sayfaları veya harddiskinizdeki resim dosyaları yapısal olmayan verilerdir.

Yapısal Veri: Nümerik, Kategorik, İkili, Metin, İlişkisel vb. veriler.

Büyük Veri: Farklı kaynaklardan elde edilen ve sürekli olarak boyutu artan veriler büyük veri adını alır. Büyük veri sürekli hızlı bir şekide artar ve birçok kaynaktan gelen verilerden meydana gelir.

Veri Bilimi: Veriler; işlenmiş veya işlenmemiş veri olabilir. Veriyi işlerken, veriyi elde ederken, veriyi anlamlı hale getirirken istatistik, makine öğrenmesi,matematik, programlama alanlarından yararlanırız. Bu alanlardan elde ettiğimiz bilgileri kullanarak yeni algoritmalar, yeni yöntemler ortaya koyarız. Bunların hepsini kapsanyan multidisipliner alan ise veri bilimidir.

Veri Analitiği

Veri analitiği elde edilen verilere farklı yöntemleri uygulayarak çeşitli sonuçlar elde eder. Uyguladığı yötemlere göre mevcut durumu ortaya koyabilir, mevcut durumun nedenini ortaya koyabilir, geleceğe yönelik tahminler de bulanabilir.Bu rada bazı istatistik ve makine öğrenmesi yöntemlerinden yararlanılabilir.

İş Analitiği

Veri analitiğinin bir çeşidir. Veri analitiğinde bahsettiğimiz yöntemlerden elde edilen sonuçlar eğer bir şirkete veya bir işletmeye fayda sağlıyorsa bu durumda iş analitiği denir.

İş Zekası

İŞ zekası şirketlerin geçmiş verilerini tuttuğu veri ambarlarından verileri çekerek yani mevcut verilerini kullanarak raporlanması işlemidir. Bu süreçteki tüm metodları, yöntemleri kapsar.Raporlar sonucunda şirket veya işletme belirli karar alır.Microsoft Power BI programı yaygın olarak kullanılmaktadır.

Veri Ambarları

İş zekasında bahsettiğimiz veri çekme işlemi veri ambarları sayesinde gerçekleşir. Veri ambarları hem şirketin hem de dış kaynaklardan elde verilen verilerin depolandığı yerdir.Veri ambarları verileri tarihsel olarak tutar ve farklı farklı veri depolarının bir araya gelmesiyle meydana gelir.

Veritabanı

Veri tabanı; verilerin anlamlı olarak ve birbirleri ile ilişkilerinin olduğu bir şekilde depolanmasını sağlayan depolama alanlarıdır.Veritabanlarında bilgiler hızlıca güncellenebilir ve düzenli olarak yer alır. Veritabanı için kullanılan birçok yazılım vardır. Bunlardan en yaygın olanı SQL yazılımıdır. Bu tür yazılımlar ile hem veritabanı oluşturabilir hem de mevcut veritablarından sorgulama kodları ile anlamlı verileri, bilgileri elde edebiliriz Kaynak.

Veri Madenciliği Nedir?

Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır (Alpaydın 2000)

Veritabanı uygulaması:

1. Adı Ahmet olan kredi kartı sahiplerini bul.
2. Bir ayda 12000 TL'den fazla harcama yapan kredi kartı sahiplerini bul.
3. DVD satın alan tüm müşterileri bul.

Veri madenciliği uygulaması

1. Riski az olan tüm kredi kartı başvurularını bul (sınıflandırma)
2. Harcama alışkanlığı benzer olan kredi kartı sahiplerini bul(demetleme)
3. DVD birlikte sıkça satın alınan ürünü bul (ilişkilendirme kuralları)

Neden Veri Madenciliği

1- Veri her zaman, her yerde anlık olarak toplanıyor.
2- Verilerden anlamlı bilgiler çıkarma ihtiyacı (Han, 2011)

SQL

SQL uzun haliyle Structured Query Language yani Yapılandırılmış Sorgu Dili demektir. Basit tanımıyla, veritabanı içindeki depolanan verilere ulaşmak ve onlar üzerinde işlem yapmak için kullanabileceğimiz bir dildir.

OLAP

OLTP(On Line Transactional Processing) türü sistemler, her gün çok sayıda işleme, girdi-çıktıya ve güncellemeye uygun sistemlerdir. Fakat canlı sistemlerde karmaşık sorgulama işlemleri sistemde bir takım sorunlara ve yavaşlamalara yol açabilir.

Bunun için OLAP sistemleri geliştirilmiştir. **OLTP** günlük operasyonel kullanım için uygun bir yöntem iken, OLAP arka planda uzun soluklu analizler için uygun bir yöntemdir. OLAP'a ihtiyaç duyulması için veri boyutunun yüksek ve karmaşık ilişkilerin çok olması gerekir.

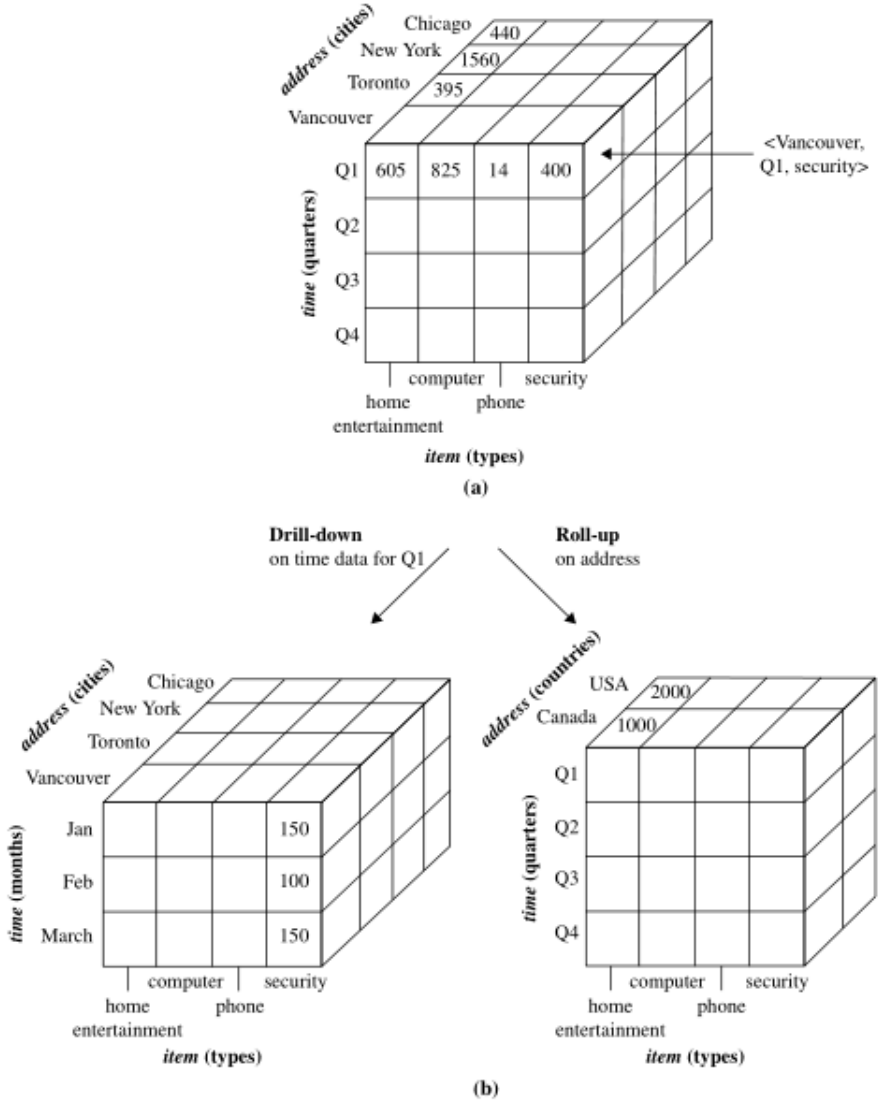
OLAP verileri ile alttaki tarzdaki verilere cevap bulabiliriz:

– Geçen yılın aynı döneminde, belirlenen satış temsilcileri bu yıl yüzde kaç fazla satış yaptı?

– Bu güne kadar toplanan müşteri verisi geçen 2 yıla oranla hangi durumda?

– Her bir ürün grubunun müşteri cinsiyetine ve aylara göre dağılımı nasıl? Kaynak

OLAP Küpleri

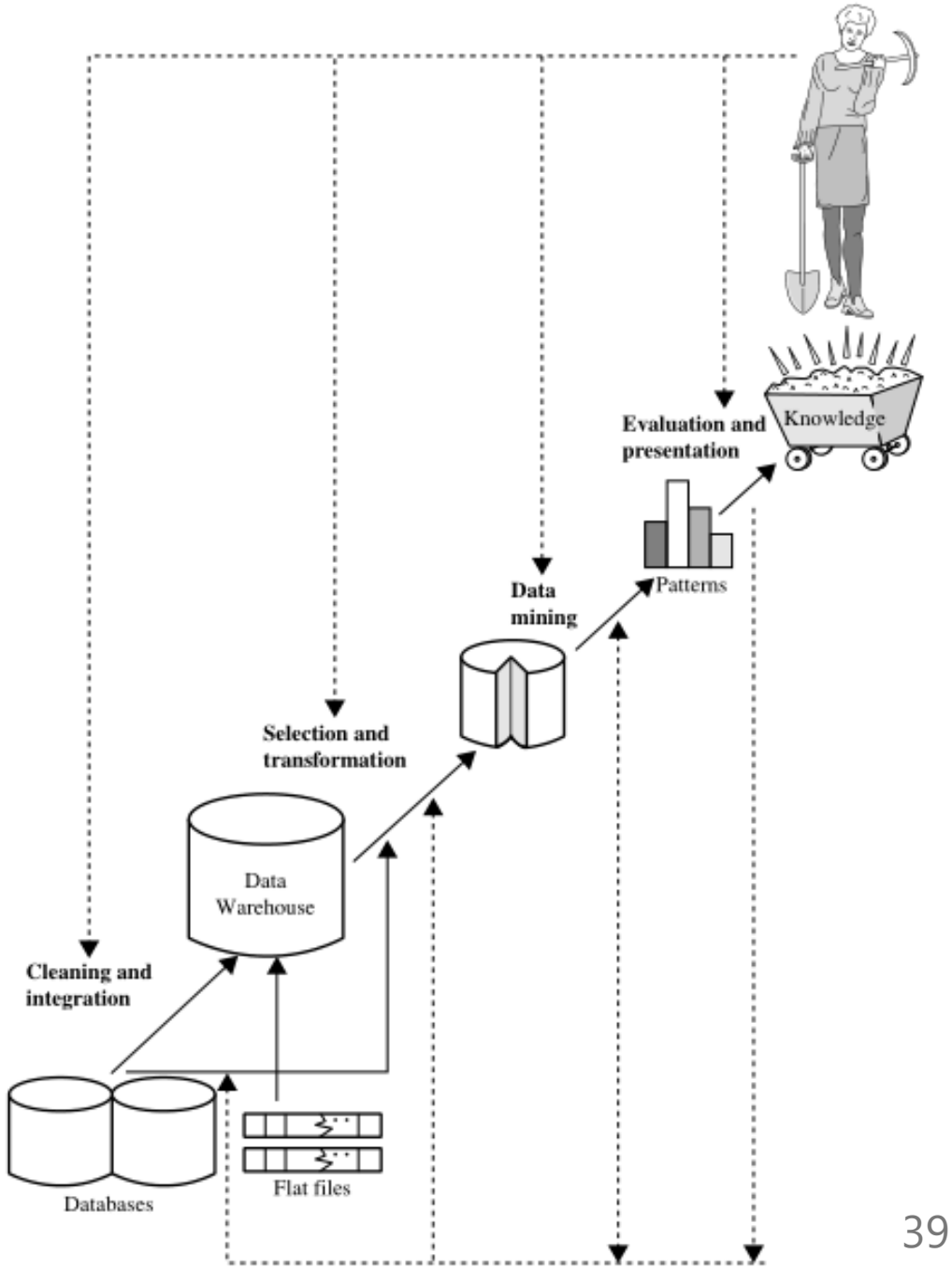


Veri Madenciliği Görevleri

1. Sınıflandırma
2. Regresyon
3. Kümeleme
4. Özetleme
5. Veri Analizi
6. Birliktelik Analizi

Adımlar

Veri tabanlarından bilgi keşfetme süreci



Veri Madenciliği Uygulamaları

Birliktelik

- “Çocuk bezi alan müşterilerin 30%’u çamaşır sodası da alır.” (Basket Analysis)

Sınıflandırma

“Genç kadınlar küçük araba satın alır; yaşlı, zengin erkekler ise büyük, lüks araba satın alır.”

Regresyon

Kredi skorlama (Application Scoring)

Zaman içinde Sıralı Örüntüler

“İlk üç taksidinden iki veya daha fazlasını geç ödemiş olan müşteriler %60 olasılıkla krediyi geriye ödeyemiyor.” (Behavioral scoring, Churning)

Benzer Zaman Sıraları

“X şirketinin hisselerinin fiyatları Y şirketinin fiyatlarıyla benzer hareket ediyor.”

İstisnalar (Fark Saptanması)

“Normalden farklı davranış gösteren müşterilerim var mı?”

Fraud detection

Döküman Madenciliği (Web Madenciliği)

“Bu arşivde (veya internet üzerinde) bu dökümana benzer hangi dökümanlar var?” (Albayrak, 2017)