Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining, 2nd Edition

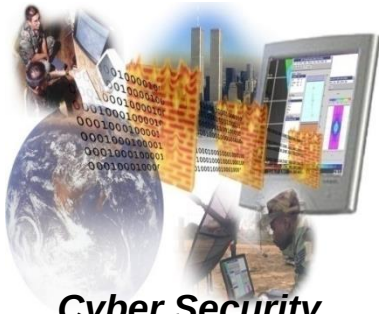
by

Tan, Steinbach, Karpatne, Kumar

Orijilal slaytların Türkçe çevirisidir.

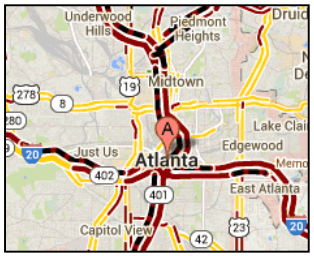
Large-scale Data is Everywhere!

- Veri oluşturma ve toplama teknolojilerindeki ilerlemeler nedeniyle hem ticari hem de bilimsel veri tabanlarında muazzam bir veri büyümesi olmuştur.
- New mantra (kutsal söz)
  - Mümkün olduğunda (**whenever**) ve mümkün olan her yerde (**wherever**) her türlü (whatever) veriyi toplayın.
- Beklentiler
  - Toplanan veriler ya toplanan amaç için ya da öngörülmeyen bir amaç için değerli olacaktır.

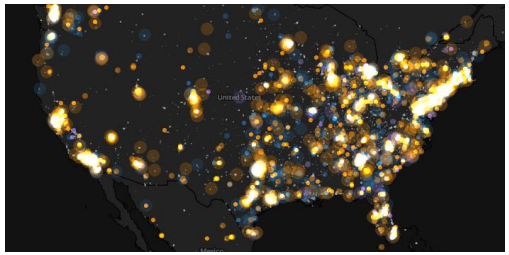


Cyber Security

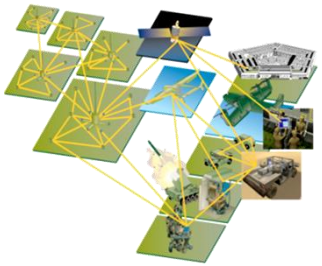
E-Commerce



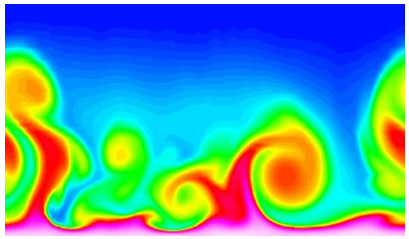
Traffic Patterns



Social Networking: Twitter



Sensor Networks



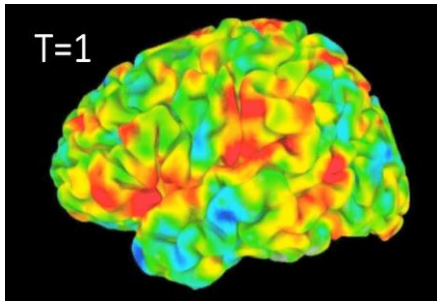
Computational Simulations

Why Data Mining? Commercial Viewpoint

- Çok sayıda veri toplanıyor ve depolanıyor
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - mağaza / marketlerde alışveriş, e-ticaret
    - Amazon.com'u her gün milyonlarca kullanıcı ziyaret ediyor
  - Bank/Credit Card transactions
- Bilgisayarlar daha ucuz ve daha güçlü hale geldi
- Rekabetçi baskı güçlü hale geldi
  - Avantaj yakalamak için daha iyi, özelleştirilmiş hizmetler sunmak (örneğin, Müşteri İlişkileri Yönetimi'nde- **C**ustomer **R**elationship **M**anagement)

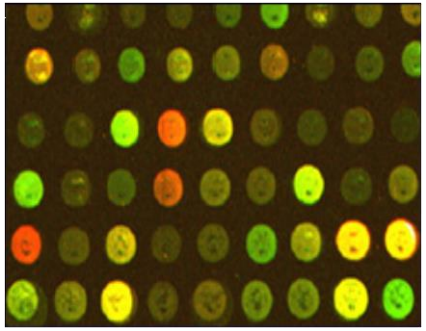
Why Data Mining? Scientific Viewpoint

- Çok yüksek hızlarda toplanan ve depolanan veriler
  - Uydudaki sensörler (remote sensors on a satellite)
    - NASA EOSDIS yılda petabyte'ların üzerinde dünyaya ilişkin bilimsel veri arşivler
  - gökyüzünü tarayan teleskoplar
    - Sky survey data
  - Yüksek-hacimli biyolojik veriler (High-throughput biological data)
  - Bilimsel simülasyonlar
    - birkaç saat içinde üretilen terabaytlarca veri
- Veri madenciliği bilim insanlarına yardımcı olur
  - büyük veri kümelerinin otomatik analizinde
  - Hipotez oluşturmada

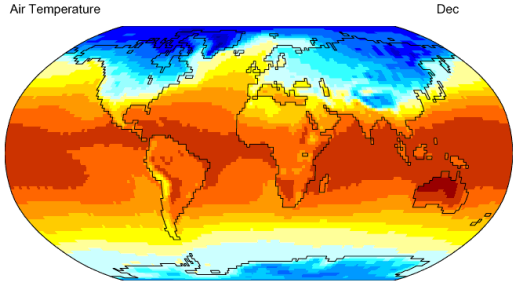


fMRI Data from Brain

Sky Survey Data



Gene Expression Data

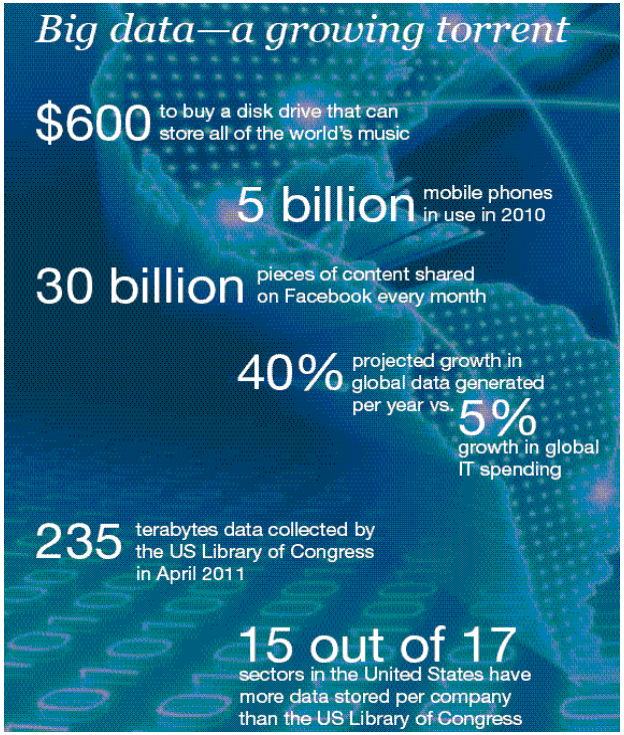


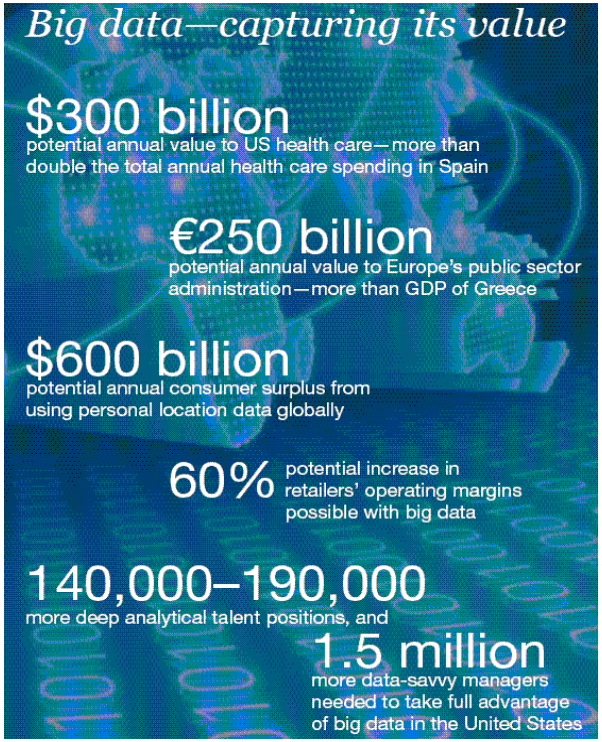
Surface Temperature of Earth

Hayatın her alanında verimliliği artırmak için harika fırsatlar

McKinsey Global Institute

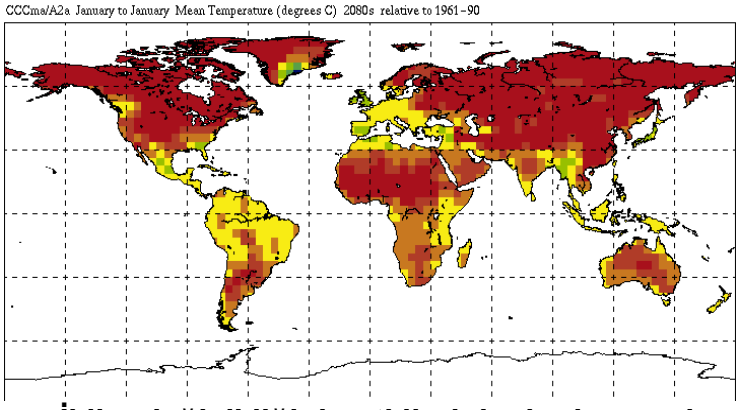
Big data: The next frontier for innovation, competition, and productivity





Toplumun Önemli Sorunlarını Çözmek için Büyük Fırsatlar

Sağlık hizmetlerini iyileştirmek ve maliyetleri düşürmek



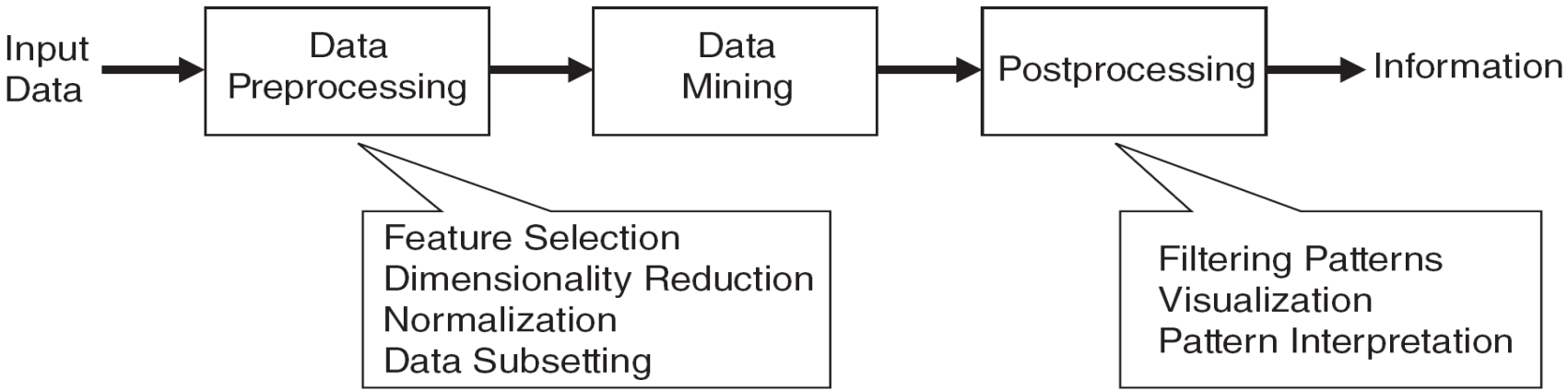
İklim değişikliğinin etkilerini tahmin etmek

Alternatif / yeşil enerji kaynakları bulmak

Tarımsal üretimi artırarak açlığı ve yoksulluğu azaltmak

What is Data Mining?

- Pek çok tanımı vardır
  - Verilerden örtük (**implicit**), önceden bilinmeyen ve potansiyel olarak yararlı (önem arz eden) bilgilerin çıkarılması
  - Anlamlı örüntüleri keşfetmek için büyük miktarlarda verinin otomatik veya yarı otomatik olarak keşfi (**exploration**) ve analizi



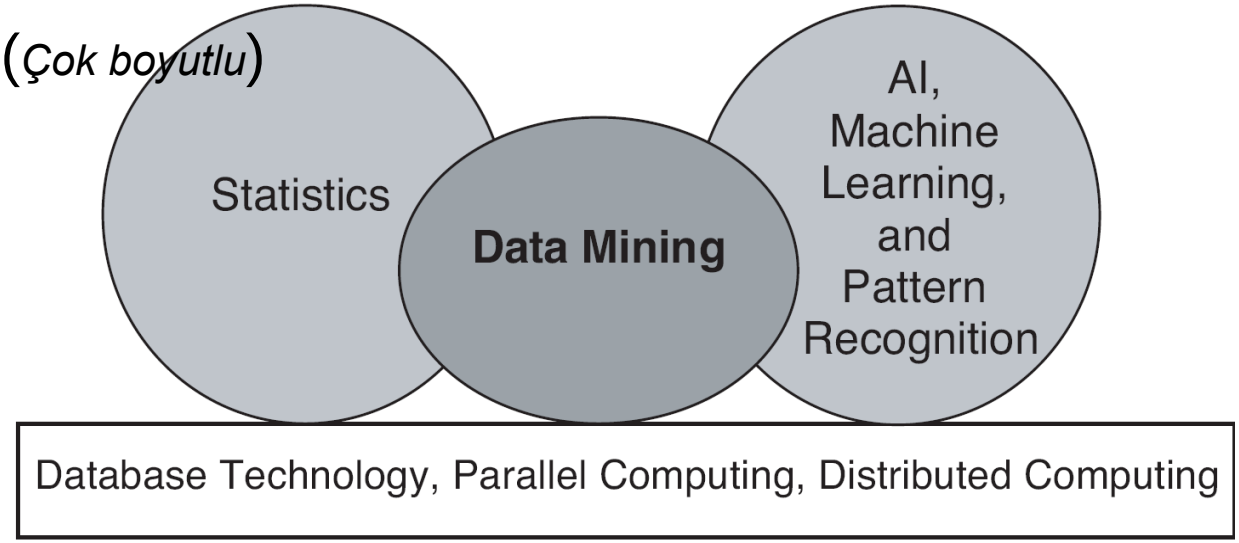
What is (not) Data Mining?

- **What is not Data Mining?**
  - Telefon rehberinde telefon numarasını aramak
  - "Amazon" hakkında bilgi için bir Web arama motorunu sorgulamak

- **What is Data Mining?**
  - Belirli isimler ABD'nin belirli bölgelerinde daha yaygındır (O'Brien, O'Rourke, O'Reilly… Boston bölgesinde)
  - Arama motoru tarafından döndürülen benzer belgeleri içeriklerine göre gruplandırın (örn. Amazon yağmur ormanları, Amazon.com)

Origins of Data Mining

- Makine öğrenmesi / yapay zeka, örüntü tanıma, istatistik ve veritabanı sistemlerinden faydalanır
- Geleneksel teknikler uygun olmayabilir, çünkü veri
  - Large-scale (*Büyük ölçekli*)
  - High dimensional (*Çok boyutlu*)
  - Heterogeneous
  - Complex
  - Distributed



- Yeni ortaya çıkan veri bilimi (**data science**) ve veri güdümlü keşif (**data-driven discovery**) alanının önemli bir bileşeni

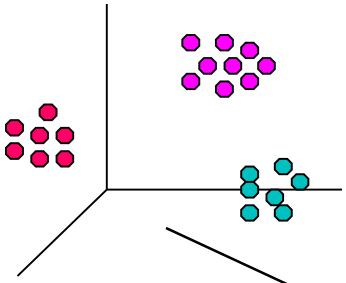
Data Mining Tasks

- Tahmin/Öngörü Yöntemleri (**Prediction Methods**)
  - Diğer değişkenlerin bilinmeyen veya gelecekteki değerlerini tahmin etmek için bazı değişkenler kullanır.

- Tanımlama/Açıklama Yöntemleri (**Description Methods**)
  - Verileri tanımlayan, insan tarafından yorumlanabilen örüntüleri bulur.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

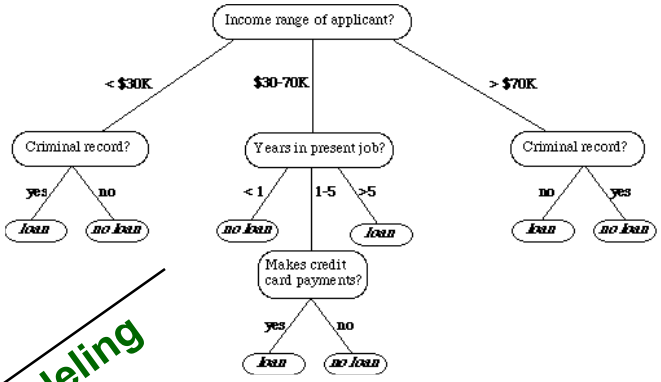
Data Mining Tasks ...



Clustering

Data

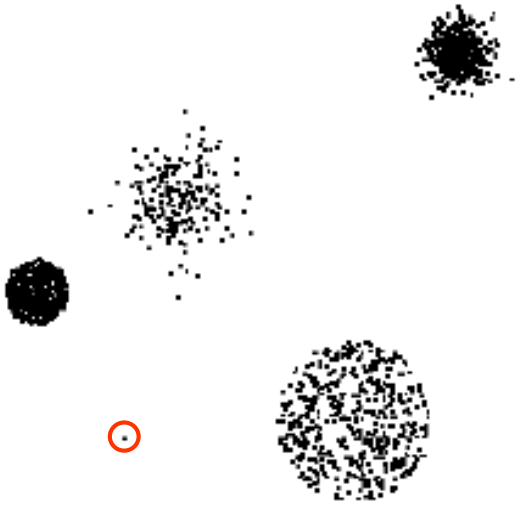
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Rules

Anomaly Detection



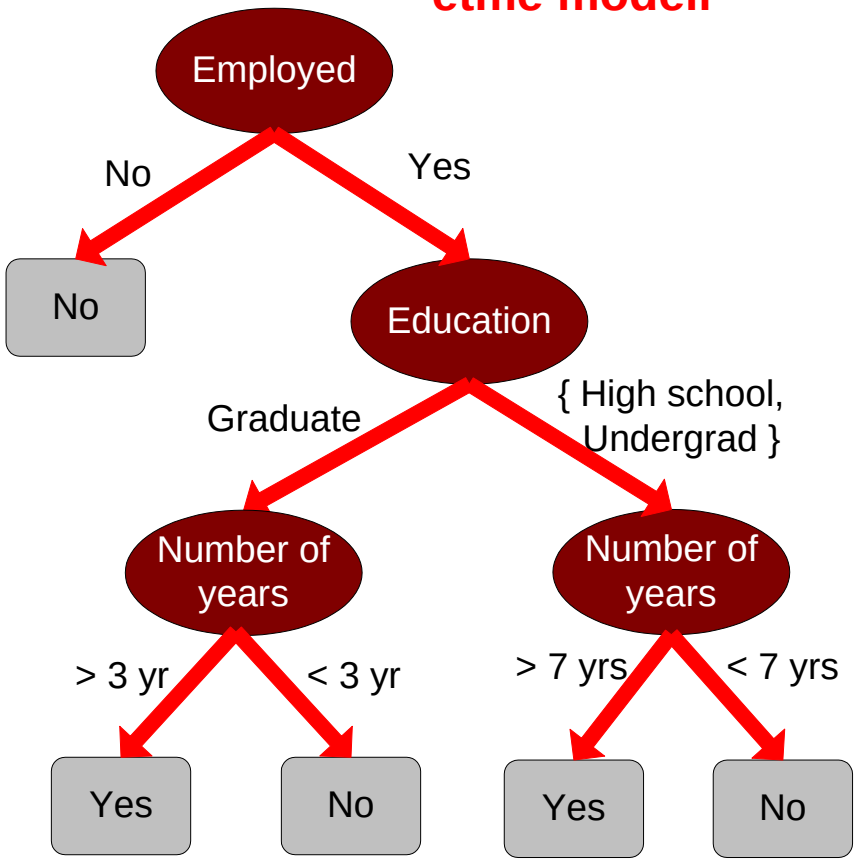
Predictive Modeling: Classification

- Sınıf özniteliği (class attribute) için diğer özniteliklerin değerlerinin bir fonksiyonu olarak bir model bulma

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Kredi liyakatini tahmin etme modeli

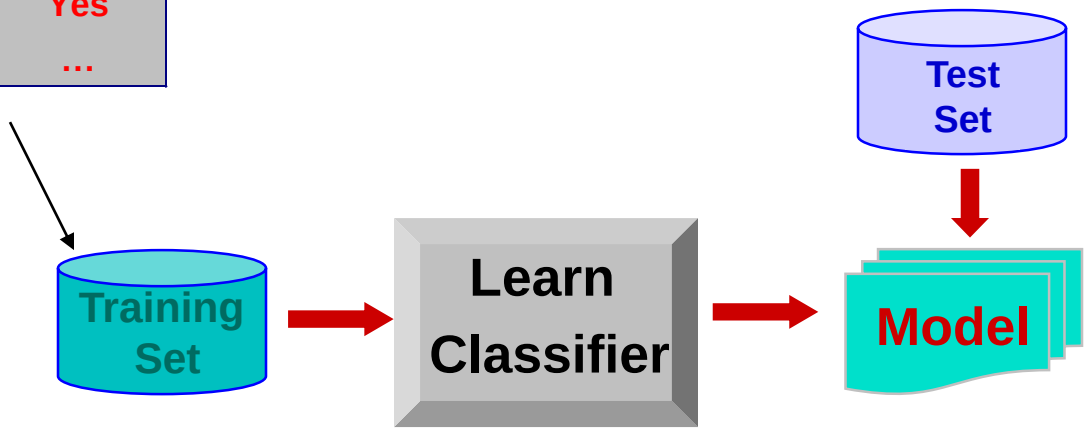


Classification Example

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

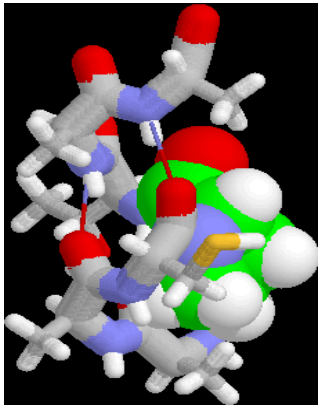


Examples of Classification Task

- Kredi kartı işlemlerini yasal veya hileli olarak sınıflandırma

- Uydu verilerini kullanarak Arazi örtülerini (su havzaları, kentsel alanlar, ormanlar, vb.) sınıflandırma

- Haber sayfalarını finans, hava durumu, eğlence, spor vb. olarak kagetorize etme
- Siber dünyada izinsiz giriş yapmaya çalışanları belirleme
- Tümör hücrelerini iyi huylu veya kötü huylu olarak tahmin etme



- Proteinin sekonder yapılarını alfa-sarmal, beta-yaprak veya rastgele spiral olarak sınıflandırmak

Classification: Application 1

- Sahtekarlık Tespiti (Fraud Detection)
  - **Amaç:** Kredi kartı işlemlerindeki hileli vakaları tahmin etmek.
  - **Yaklaşım:**
    - Kredi kartı işlemlerini ve hesap sahibinin bilgileri öznitelik olarak kullanmak
      - müşteri ne zaman satın alır, ne satın alır, ne sıklıkta zamanında ödeme yapar, vb.
    - Geçmiş işlemler sahtekarlık veya yasal işlem olarak etiketlenir. Bu, sınıf niteliğini oluşturur.
    - İşlemlerin sınıfı için bir model eğitilir/öğrenilir.
    - Bir hesaptaki kredi kartı işlemlerini gözlemleyerek sahtekarlığı tespit etmek için bu model kullanılır.

Classification: Application 2

- Telefon operatörlerinin müşterileri için kayıp tahmini (Churn prediction)
  - **Amaç:** Bir müşterinin bir rakibe kaptırılıp kaptırılmayacağını tahmin etmek.
  - **Yaklaşım:**
    - Nitelikleri bulmak için geçmiş ve mevcut müşterilerin her biriyle ilgili işlemlerin ayrıntılı kaydını kullanılır.
      - Müşterinin ne sıklıkta aradığı, nereden aradığı, günün hangi saatinde en çok aradığı, finansal durumu, medeni durumu vb.
    - Müşteriler sadık veya sadık olmayan olarak etiketlenir.
    - Sadakat için bir model oluşturulur

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 3

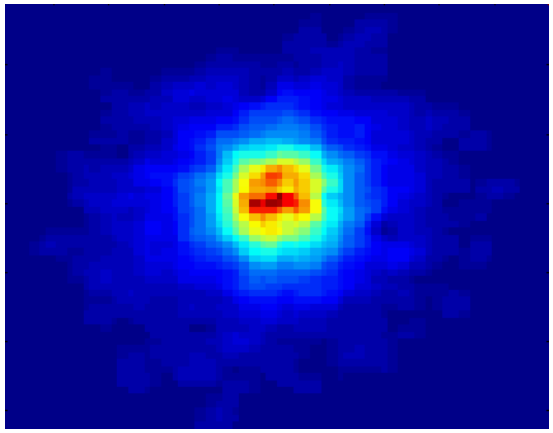
- Gök Haritası Kataloğu (Sky Survey Cataloging)
  - **Amaç:** Teleskopik inceleme görüntülerine (Palomar Gözlemevi'nden) dayalı olarak gökyüzü nesnelerinin, özellikle görsel olarak soluk olanların sınıfını (yıldız veya galaksi) tahmin etmek.
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Yaklaşım:**
    - Görüntüyü segmentlere ayırın.
    - Görüntü özniteliklerini (özellikler) ölçün - nesne başına 40 tane.
    - Sınıfı bu özelliklere göre modelleyin.
    - Başarı Hikayesi: Bulması zor olan en uzak nesnelerden biri olan (galaksi dışında) 16 yeni kırmızı yıldızsı gökcisim (red-shift quasars) bulunabildi!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Courtesy: http://aps.umn.edu

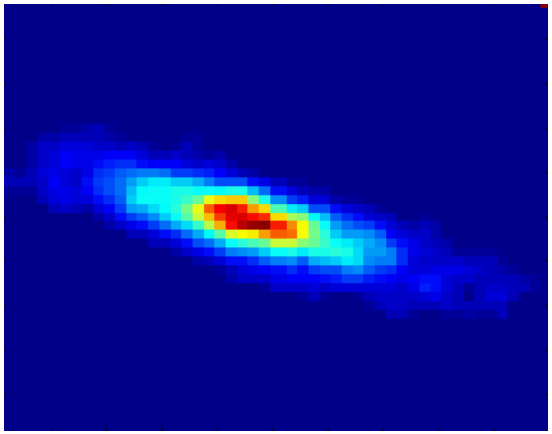
Early



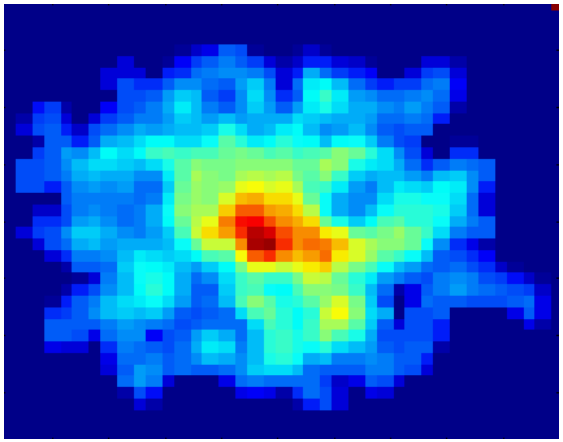
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



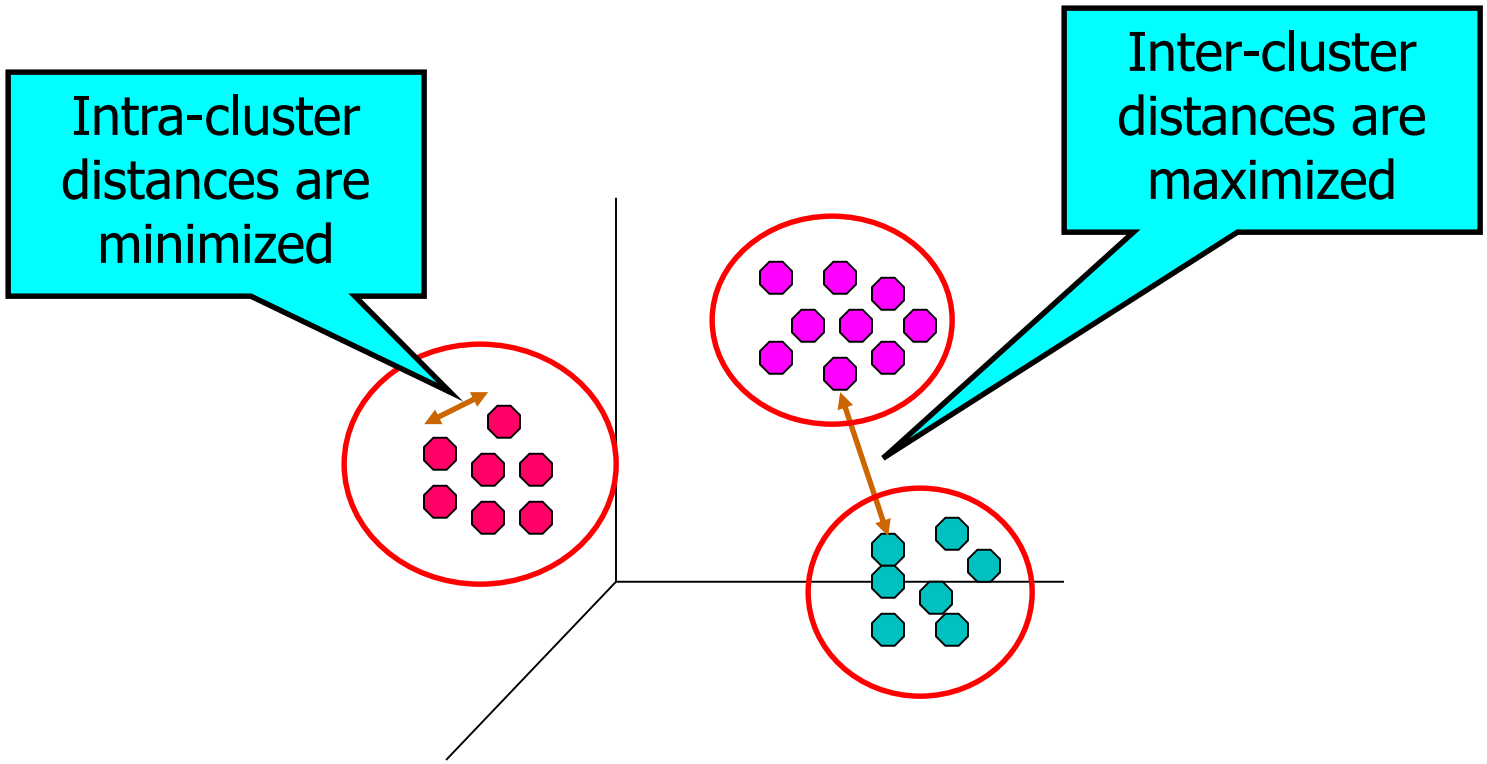
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Regression

- Doğrusal veya doğrusal olmayan bir bağımlılık modeli varsayarak, belirli bir sürekli değerli değişkenin değerini diğer değişkenlerin değerlerine göre tahmin etmek
- İstatistik ve sinir ağı alanlarında üzerinde yoğun bir şekilde çalışılmıştır.
- Örnekler:
  - Reklam harcamalarına dayalı olarak yeni ürünün satış miktarlarını tahmin etme.
  - Sıcaklık, nem, hava basıncı vb. nin bir fonksiyonu olarak rüzgar hızlarını tahmin etme
  - Borsa endekslerinin zaman serisi tahmini.

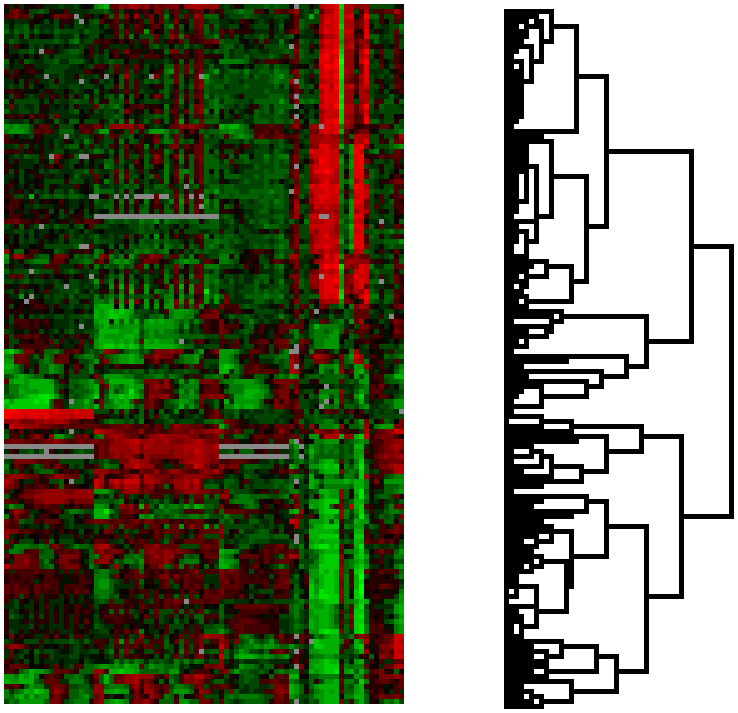
Clustering (Kümeleme)

- Bir gruptaki nesnelerin birbirine benzeyeceği (veya ilişkilendirileceği) ve diğer gruplardaki nesnelerden farklı (veya ilgisiz) olduğu nesne gruplarını bulma

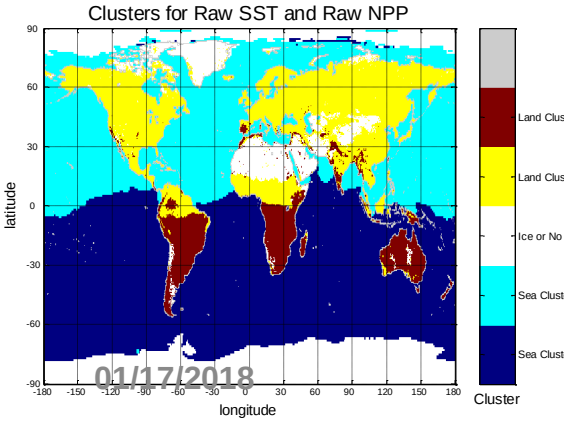


Applications of Cluster Analysis

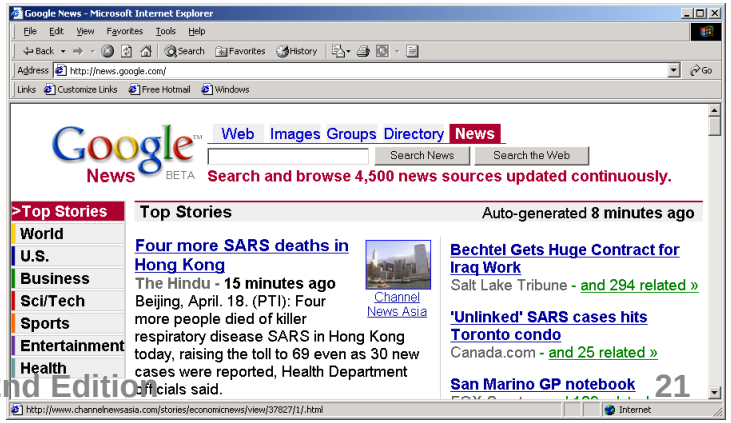
- **Anlama (Understanding)**
  - Hedeflenen pazarlar için özel profil oluşturma
  - «Browsing» için ilgili belgeleri gruplama
  - Benzer işlevselliğe sahip genleri ve proteinleri gruplama
  - Benzer fiyat dalgalanmalarına sahip hisse senetlerini gruplama
- **Özetleme(Summarization)**
  - Büyük veri kümelerinin boyutunu küçültme



Courtesy: Michael Eisen



K-means yönteminin, Deniz Yüzeyi Sıcaklığı (SST) ve Net Birincil Üretimi (NPP) Kuzey ve Güney Yarımküre'yi yansıtan kümelere ayırmak için kullanılması.



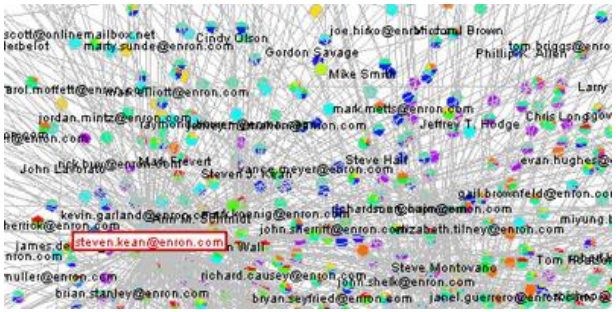
Clustering: Application 1

- Pazar Bölümlemesi (Market Segmentation):
  - **Amaç:** herhangi bir alt kümenin farklı bir pazarlama karmasıyla ulaşılacak bir pazar hedefi olarak seçilebileceği bir pazarın farklı müşteri alt kümelerine bölünmesi.
  - **Yaklaşım:**
    - Coğrafi ve yaşam tarzı ile ilgili bilgilere dayanarak müşterilerin farklı özelliklerini toplayın.
    - Benzer müşteri kümelerini bulun.
    - Farklı kümelerdekilerle aynı kümedeki müşterilerin satın alma örüntülerini gözlemleyerek kümeleme kalitesini ölçün.

Clustering: Application 2

- Document Clustering:
  - **Amaç:** İçinde geçen önemli terimlere dayalı olarak birbirine benzeyen belge gruplarını bulmak
  - **Yaklaşım :** Her bir belgede sık görülen terimleri tanımlayıp farklı terimlerin frekanslarına dayalı bir benzerlik ölçüsü oluşturun ve bunları kümeleme için kullanın.

Enron email dataset



Association Rule Discovery: Definition (Birliktelik Kuralı Keşfi)

- Her biri belirli bir koleksiyondan birkaç öğe içeren bir kayıt kümesi verildiğinde
  - Diğer öğelerin olma durumlarına dayalı olarak bir öğenin olmasını tahmin edecek bağımlılık kuralları üretmek

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

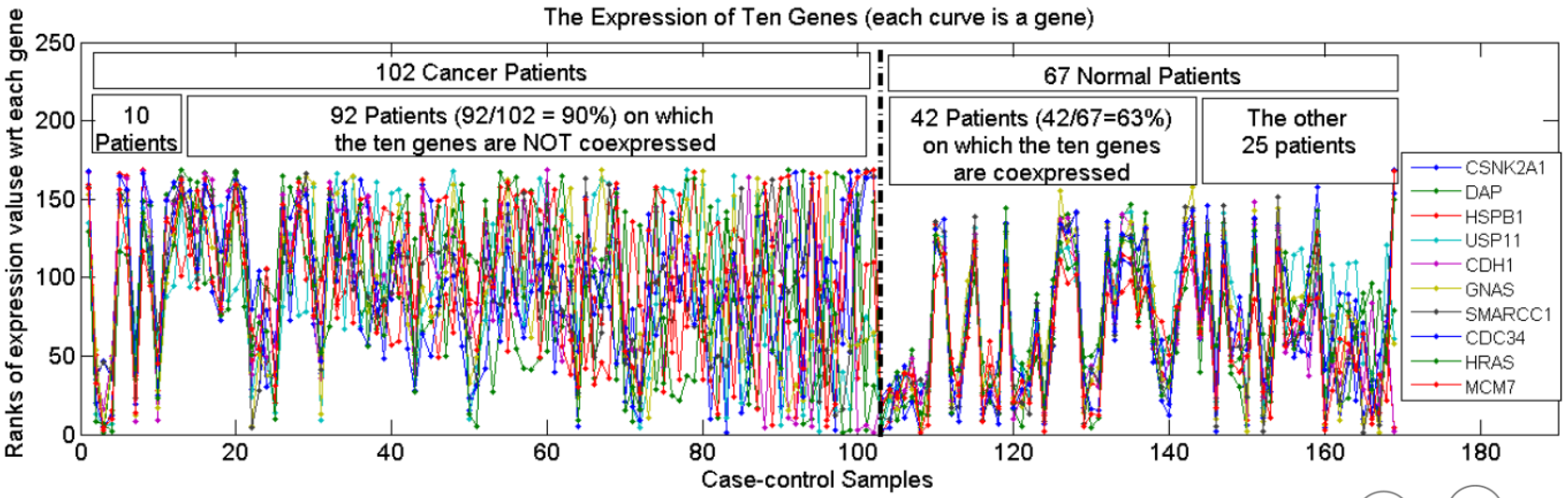
Association Analysis: Applications (Birliktelik analizi)

- Market sepeti analizi
  - Kurallar; satış promosyonu, raf yönetimi ve envanter yönetimi için kullanılır
- Telekomünikasyon alarm teşhisi
  - Kurallar, aynı zaman aralığında sık sık meydana gelen alarmların birleşimini bulmak için kullanılır
- Medical Informatics
  - Kurallar, hasta semptomları ve bazı hastalıklarla ilişkili test sonuçlarının kombinasyonunu bulmak için kullanılır

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

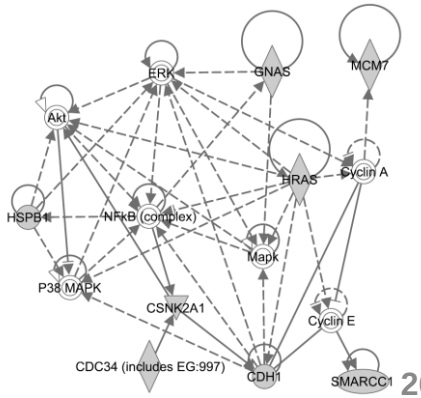
Three lung cancer datasets [Bhattacharjee et a 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway

which is well-known to be related to lung cancer

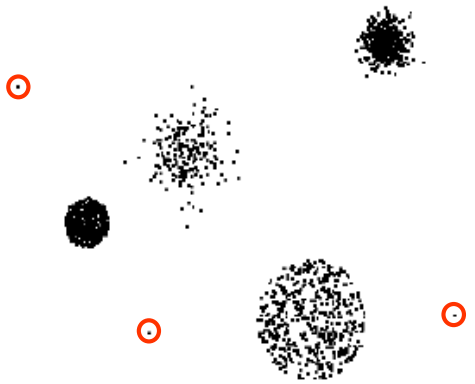
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

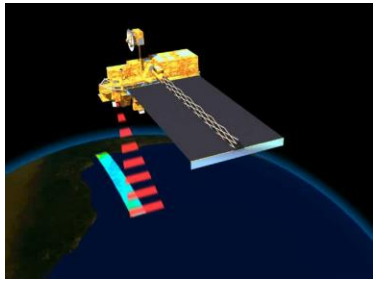


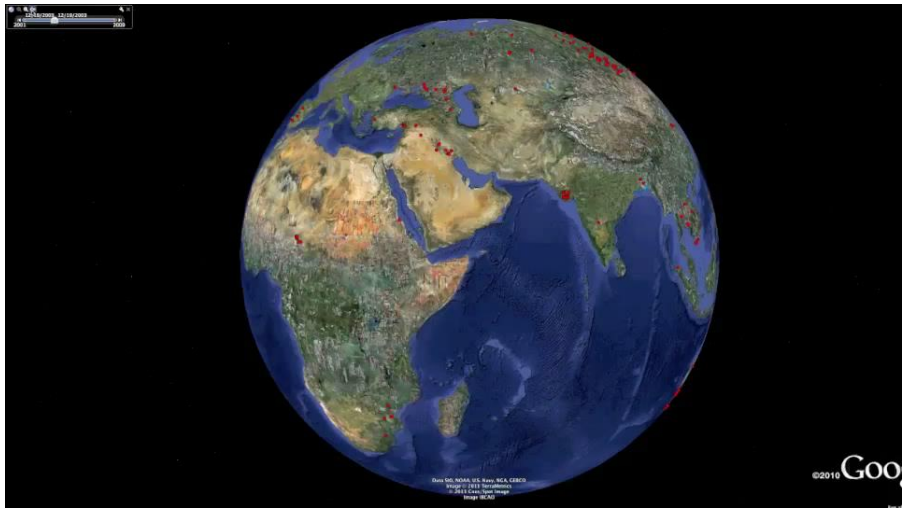
[Fang et al PSB 2010]

Deviation/Anomaly/Change Detection (Sapma/Analomali/Değişim tespiti)

- Normal davranıştan önemli derecedeki sapmaları tespit etmek
- Applications:
  - Kredi Kartı Sahtekarlık Tespiti
  - İzinsiz Ağ (Network) Giriş Tespiti
  - İzleme ve gözetim için kullanılan sensör ağlarından gelen anormal davranışı belirlemek
  - Küresel orman örtüsündeki değişiklikleri tespit etmek







Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis