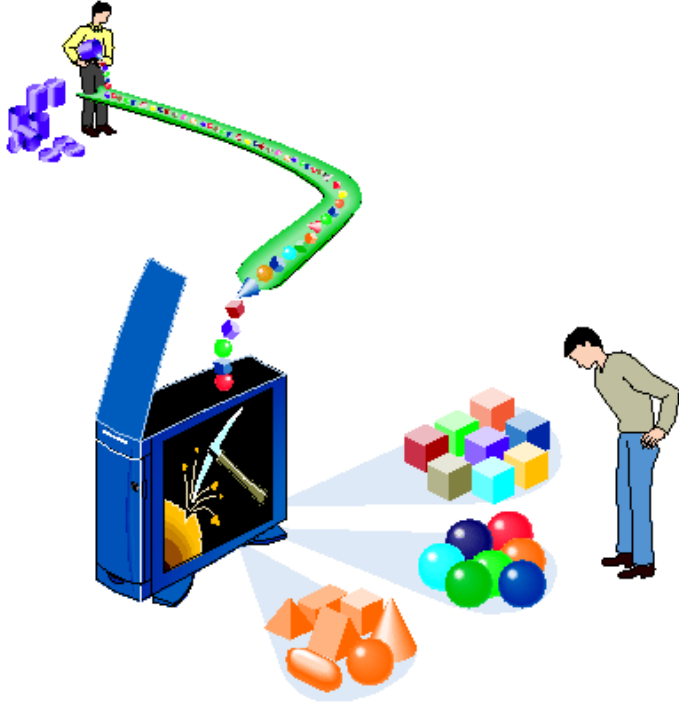




Veri Madenciliđi



Dr. Öğretim Üyesi Mustafa Gökçe Baydoğan

mustafa.baydogan@boun.edu.tr

www.mustafabaydogan.com

blog.mustafabaydogan.com

İçerik

□ Veri Madenciliği nedir?

■ Bir örnek

- Boğaziçi Üniversitesi 2014 yılı ders kayıt zamanı atılan tweetlerin incelenmesi

■ Veri madenciliğinde örnek problemler ve uygulamalar

Veri madenciliği nedir?

▣ Veri madenciliği

- Büyük miktarda veri içinden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilgi ve örüntülerin çıkarılması olarak tanımlanmaktadır.



Veri madenciliği nedir?

Bir örnek

- Boğaziçi Üniversitesi 2014 Bahar dönemi kayıt zamanı olan Şubat 10-14, 2014 tarihleri arası atılan tweetlerin incelenmesi
 - Analiz için R (<http://www.r-project.org/>) kullanıldı.
 - R paketlerden oluşur
 - **twitteR** ve **tm** paketleri kullanıldı
 - twitteR: twitterdan veri alabilmek için
 - tm: metin verisi işleme için
 - Sadece 10-20 satır R kodu



Kayıt zamanı twitter aktivitesi

□ #boun hashtagli tweetler aranır

```
tweets<- searchTwitter(`#boun`,since=`2014-02-10`, until=`2014-02-14`)  
tweet_texts<- sapply(tweets, function(x) x$getText())
```

□ Veri manipulasyonu

```
text_corpus <- Corpus(VectorSource(tweet_texts))  
text_corpus <- tm_map(text_corpus, tolower)  
text_corpus <- tm_map(text_corpus, removePunctuation)  
wordcloud(text_corpus)
```

Kayıt zamanı twitter aktivitesi



Kayıt zamanı twitter aktivitesi

- Kelime bulutu sadece görünme sayısını vermekte

- Söylenenler ne anlam ifade ediyor?

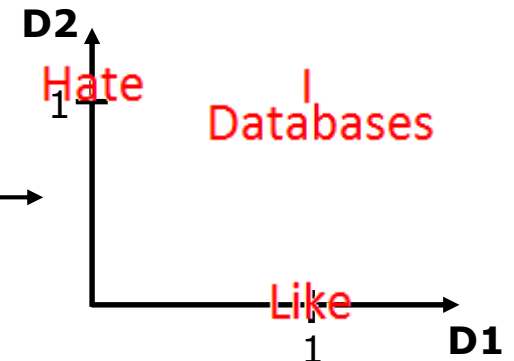
- Metni sayıya çevirme

- D1 = "I like databases"
 - D2 = "I hate databases",

	I	like	hate	databases
D1	1	1	0	1
D2	1	0	1	1



	Document 1	Document2
I	1	1
Like	1	0
Hate	0	1
Databases	1	1



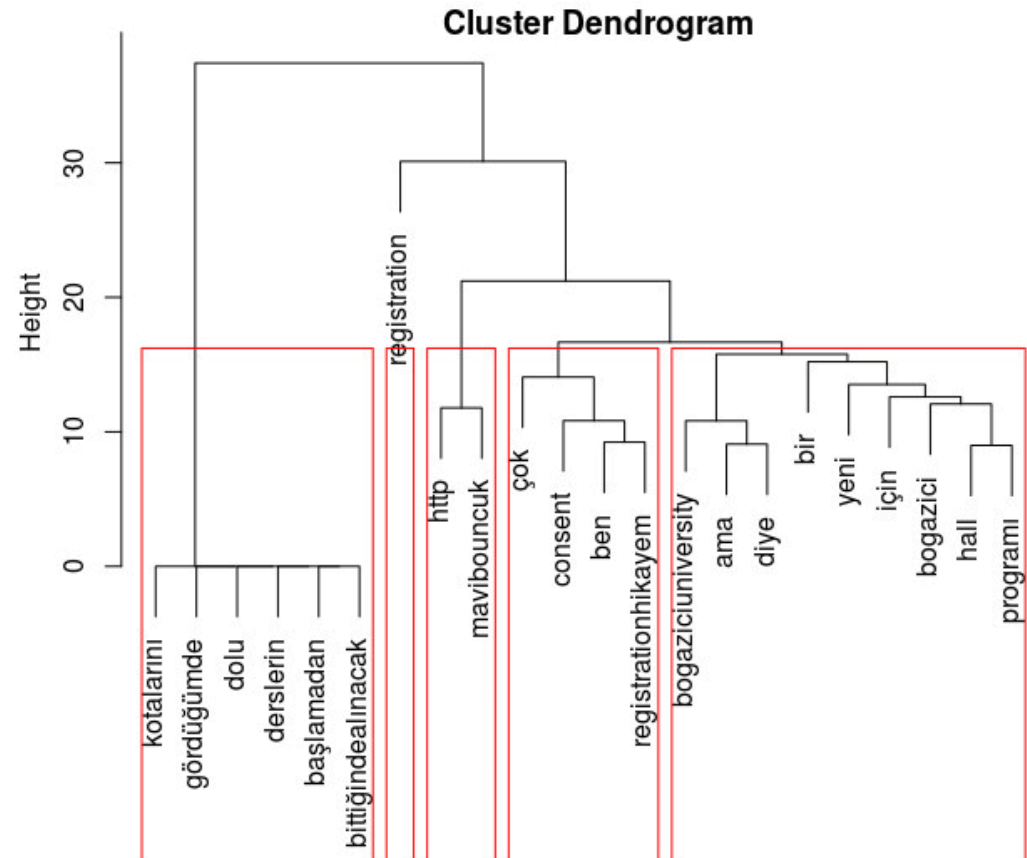
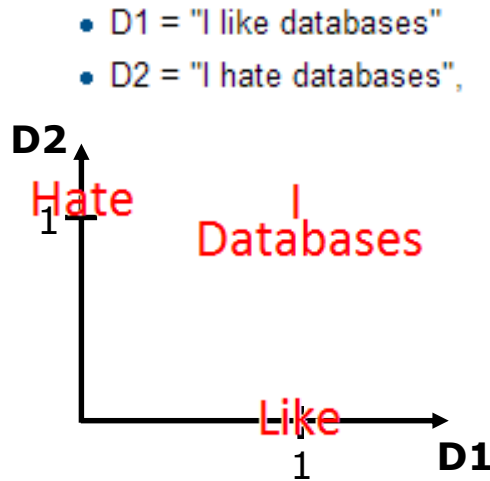
- Döküman-terim matrisi oluşturma ve az geçen kelimeleri atma

```
dtm=TermDocumentMatrix(text_corpus)
dtm=removeSparseTerms(dtm,sparse=0.95)
```

Kayıt zamanı twitter aktivitesi

□ Bir tür kümeleme yapalım*

- Burada hiyerarşik kümeleme kullanıldı

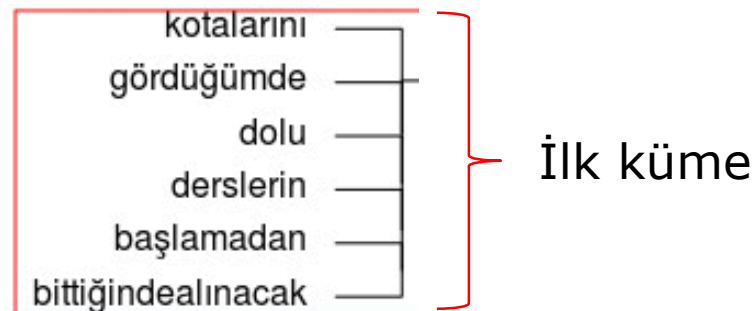


(*ilerleyen zamanda öğrenilecek)

Kayıt zamanı twitter aktivitesi

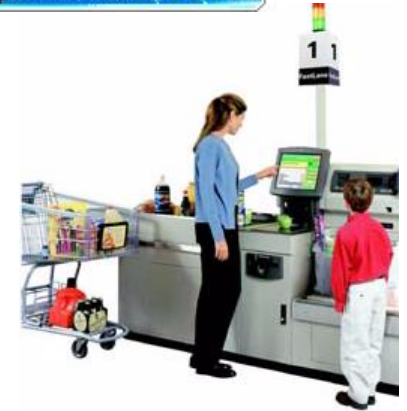
■ Özet

- Büyük bir twitter verisi içinden alakalı olabilecek kısmı seçip, bir takım veri manipülasyonu ve kümele algoritmaları kullanarak ortaya bilgi çıkarıldı.
 - Kayıt en popüler konu
 - mavibouncuk kayıt zamanı öncesi Boğaziçi Üni. öğrencilerini çekmeye çalışan bir sosyal medya platformu
 - <https://twitter.com/mavibouncuk>
 - Kayıt sistemi ile ilgili problemler var



Neden Veri Madenciliği?

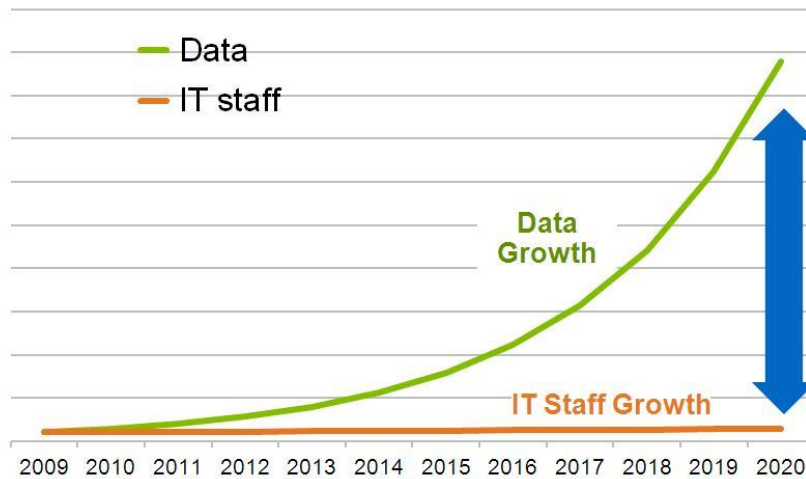
- Veri patlaması veya seli
 - Mağazalardaki satış/alış işlemleri
 - Banka ve Kredi kartı işlemleri
 - Bir çok sektördeki veri ve işlemler
 - Web verileri
- Teknolojinin ucuzlaması
- Rekabetin artması
 - Veri analizi sonucunda alınan kararların etkinliği birçok alanda ispatlanmıştır



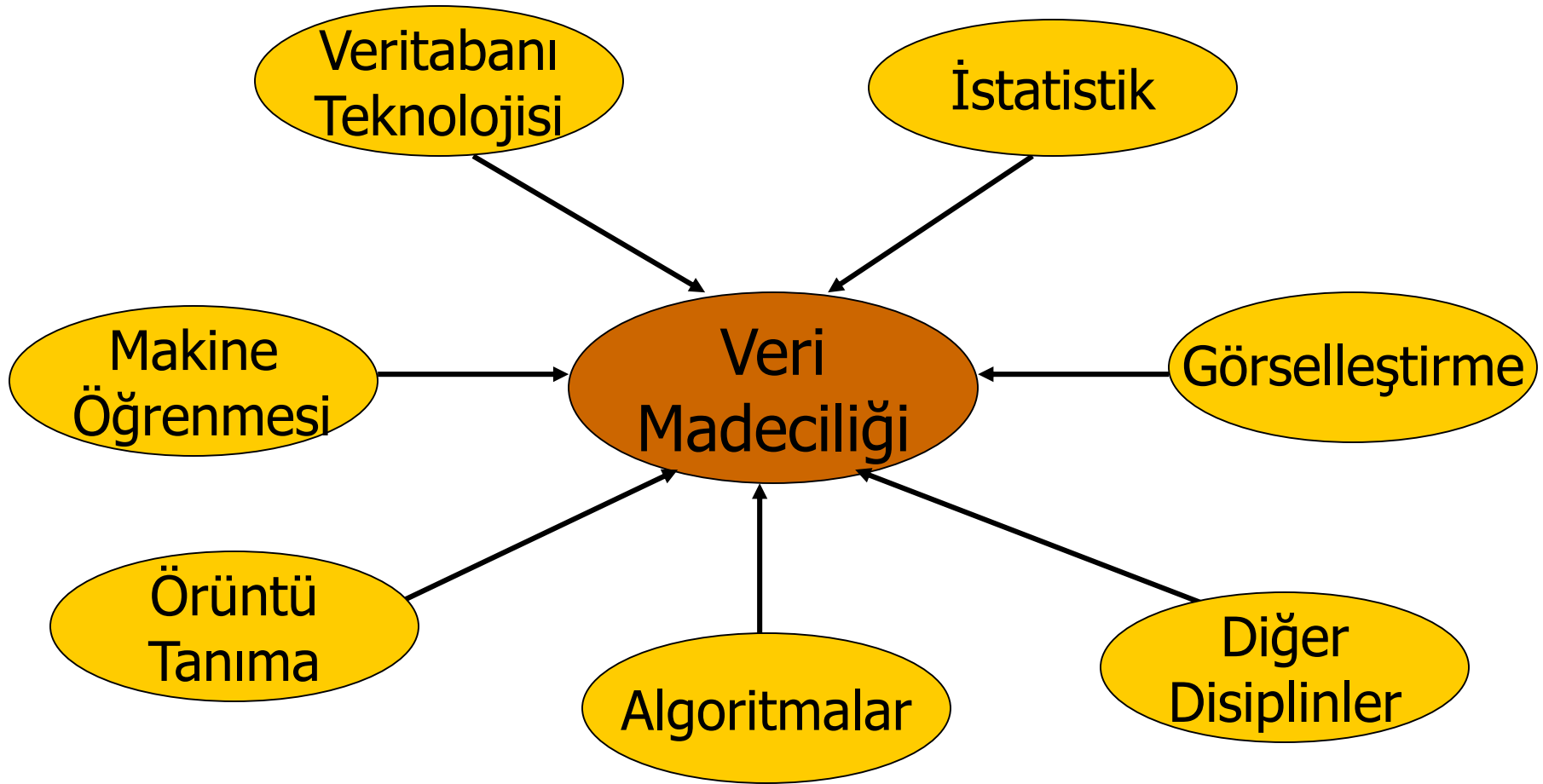
Neden Veri Madenciliği?

- Büyük verilerde klasik yaklaşımlarla ortaya çıkarması zor olan 'gizli' bilgiler vardır.
- Anlamlı bilginin bulunması haftalarca sürebilir.
- Eldeki verinin büyük bir kısmı hala tamamen analiz edilmemiştir.

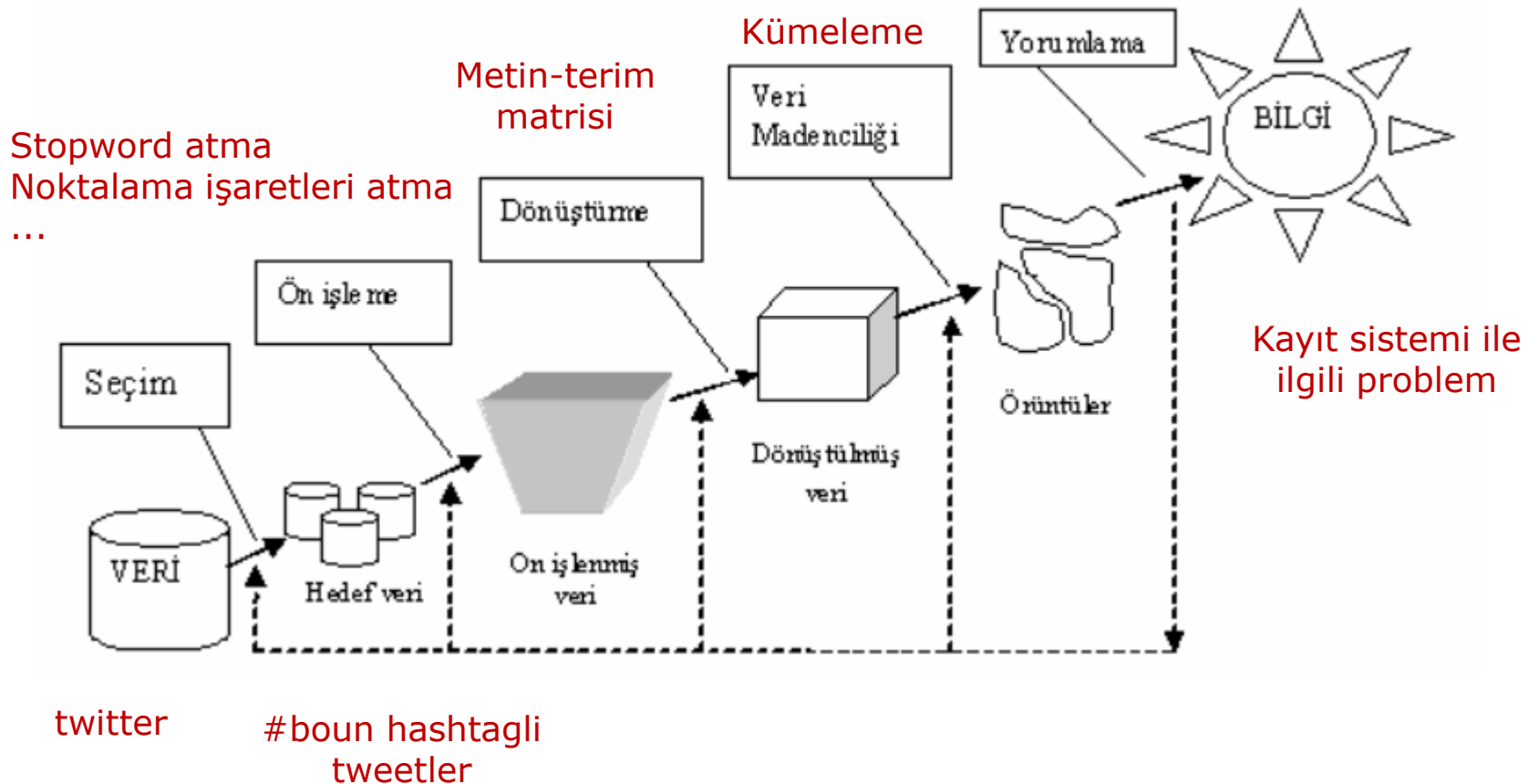
The Data Management Gap



Veri madenciliđi sihir deđildir



Adımlar



Adımlar

1. Amaç tanımlama:

- Ürünler arasında *bağıntı* ?
- Yeni pazar *segmentleri* veya potansiyel müşteriler?
- *Zaman içindeki* satın alma örüntüleri veya ürün satım eğrileri?
- Müşterileri *gruplamak, sınıflandırmak* ?

2. Veri hazırlama

- Veriyi birleştir, seç ve önüle
(Eğer veri ambarı varsa zaten yapılmıştır)
- Var olan verinin dışında, amaç için kullanılabilecek ek bilgi var mı?

Adımlar

2. Veri hazırlama – devam

(En önemli adımlardan biridir)

- **Veri seçimi**: Önemli değişkenlerin saptanması
- **Veri temizleme**: Hata, tutarsızlık, tekrar ve eksik verilerin ayıklanması/düzeltilmesi
- **Veri fırçalama**: Gruplama, dönüşümler
- **Görsel inceleme**: Veri dağılımı, yapısı, istisnalar, değişkenler arasında bağıntılar
- **Değişken analizi**: Gruplama

Adımlar

3. Yöntem seçme

- Amaç sınıfının tanımlanması

Gruplama (Clustering/Segmentation), Regresyon Analizi (Regression), Sınıflandırma (Classification), Bağıntı kurma (Association), Zaman içinde örüntü bulma/tahmin yapma (Pattern detection/Prediction in time)

- Çözüm sınıfının tanımlanması

Açıklama (Karar ağaçları,kurallar) vs Kara kutu (sinir ağı)

- Model değerlendirme, geçerleme ve karşılaştırma

k-kat çapraz geçerleme, istatistiksel testler

- Modellerin birleştirilmesi

Adımlar

4. Yorumlama

- Sonuçlar (açıklamalar/tahminler) doğru mu, dikkate değer mi?
- Uzmanla danışma

Veri madenciliği yöntemleri

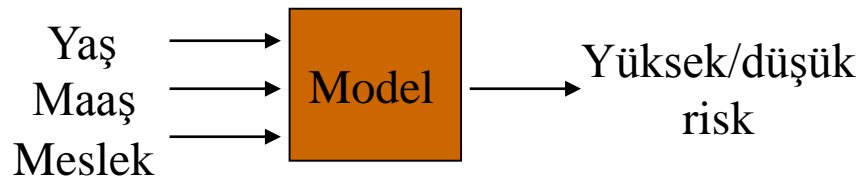
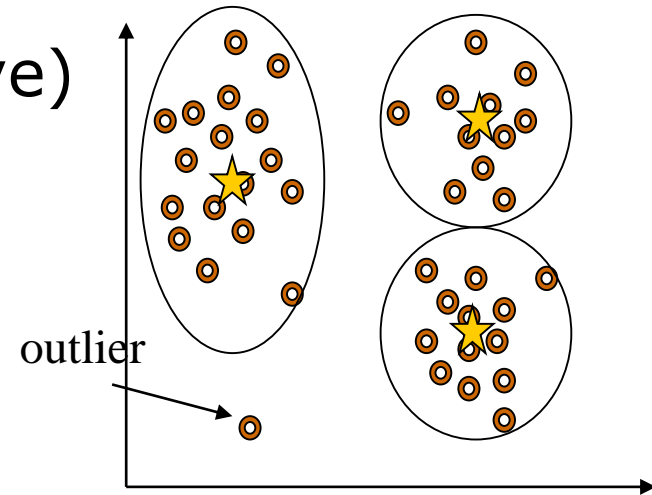
Genel olarak veri madenciliği yöntemleri iki sınıfa ayrılabilir:

■ Tanımlayıcı Yöntemler (Descriptive)

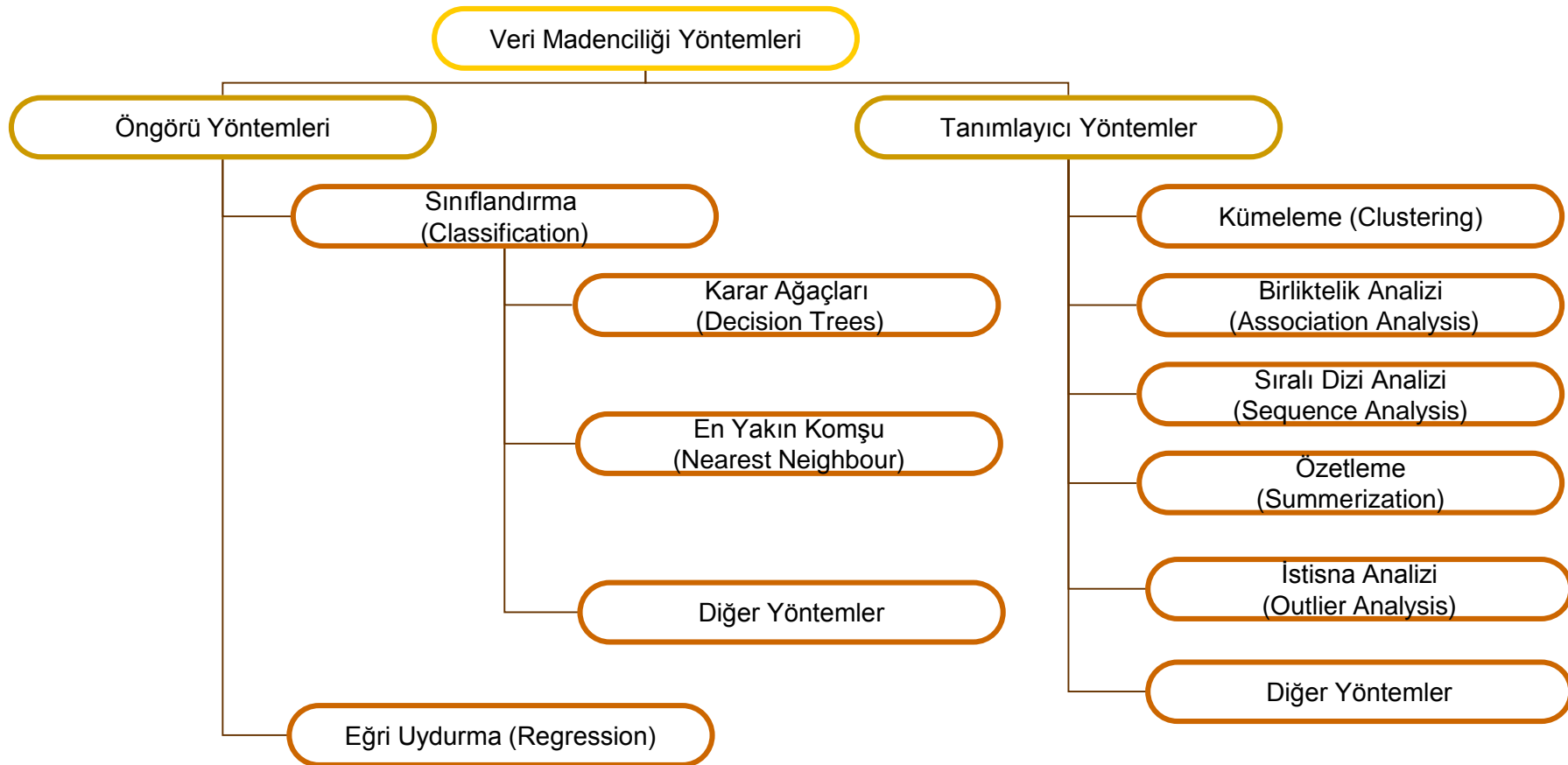
- Veriyi tanımlayan yorumlanabilir örüntülerin bulunması

■ Öngörü Yöntemleri (Predictive)

- Öngörü amacı ile var olan verilerden yorum çıkarılması



Veri madenciliği yöntemleri



Veri

▣ Veri, çok boyutlu değişkenler tablosudur

Ad	Gelir	Birikim	Medeni hali	...	Default
Ali	25,000 \$	50,000 \$	Evli	...	Hayır
Veli	18,000 \$	10,000 \$	Evli		Evet

Örnek (instance)
Kayıt (record)
Nesne (object)

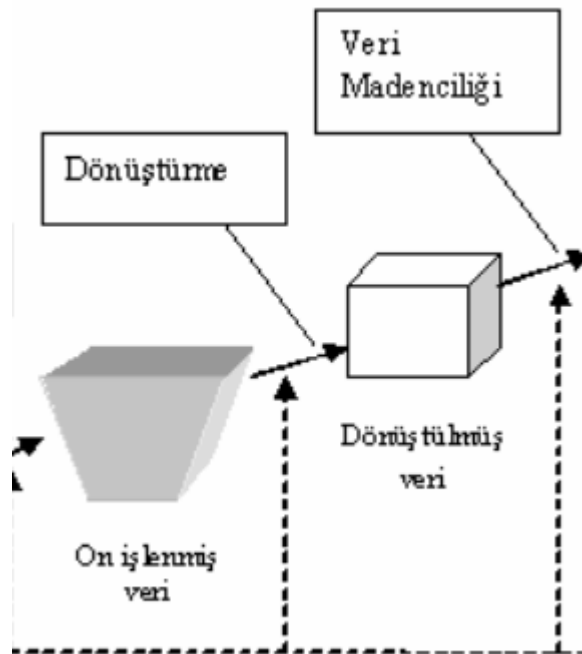
Değişken (variable)
Öznitelik (feature)

Veri

- Verinin nasıl ifade edildiği uygulamaya bağlı olarak değişir ve çok önemlidir.

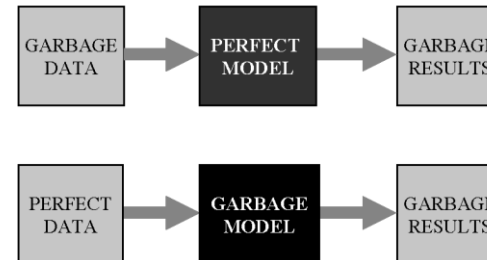
- D1 = "I like databases"
- D2 = "I hate databases",

	I	like	hate	databases
D1	1	1	0	1
D2	1	0	1	1



Bu aşamaya **öznitelik çıkarımı/gösterimi** (feature extraction/representation) de denir.

MODEL CALCULATIONS
"Garbage In-garbage Out" Paradigm



Sınıflandırma

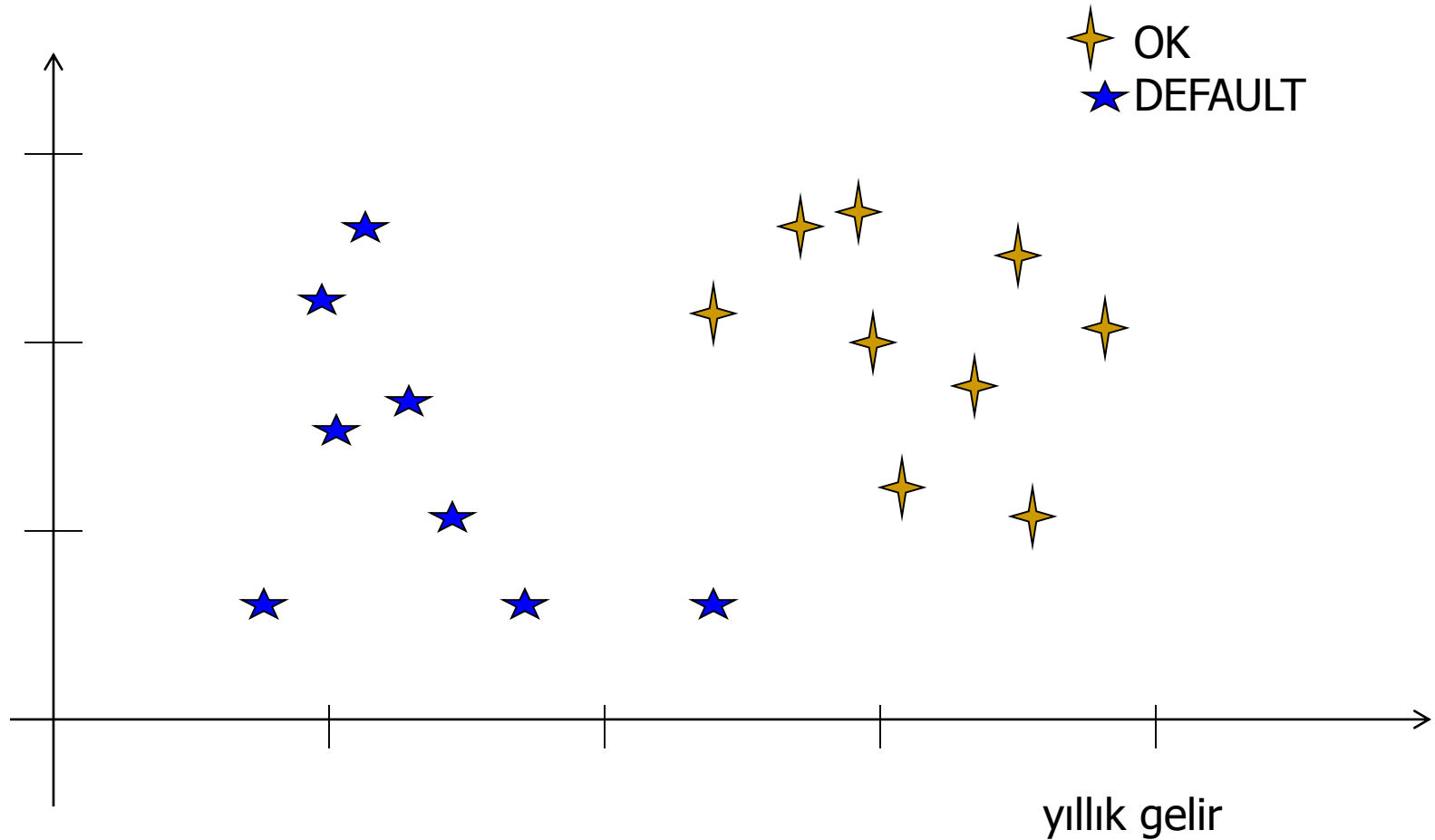
Sınıflamanın temel kuralları:

- ❑ Öğrenme **eğitici**lidir (supervised).
- ❑ **Veri setinde** bulunan **her örneğin** bir dizi **özniteliği** vardır ve bu niteliklerden biri de **sınıf** bilgisidir.
- ❑ Hangi sınıfa ait olduğu bilinen nesneler (**öğrenme kümesi**- training set) ile bir model oluşturulur
- ❑ Oluşturulan model öğrenme kümesinde yer almayan nesneler (**deneme kümesi**- test set) ile denenerek başarısı ölçülür.

Sınıflandırma

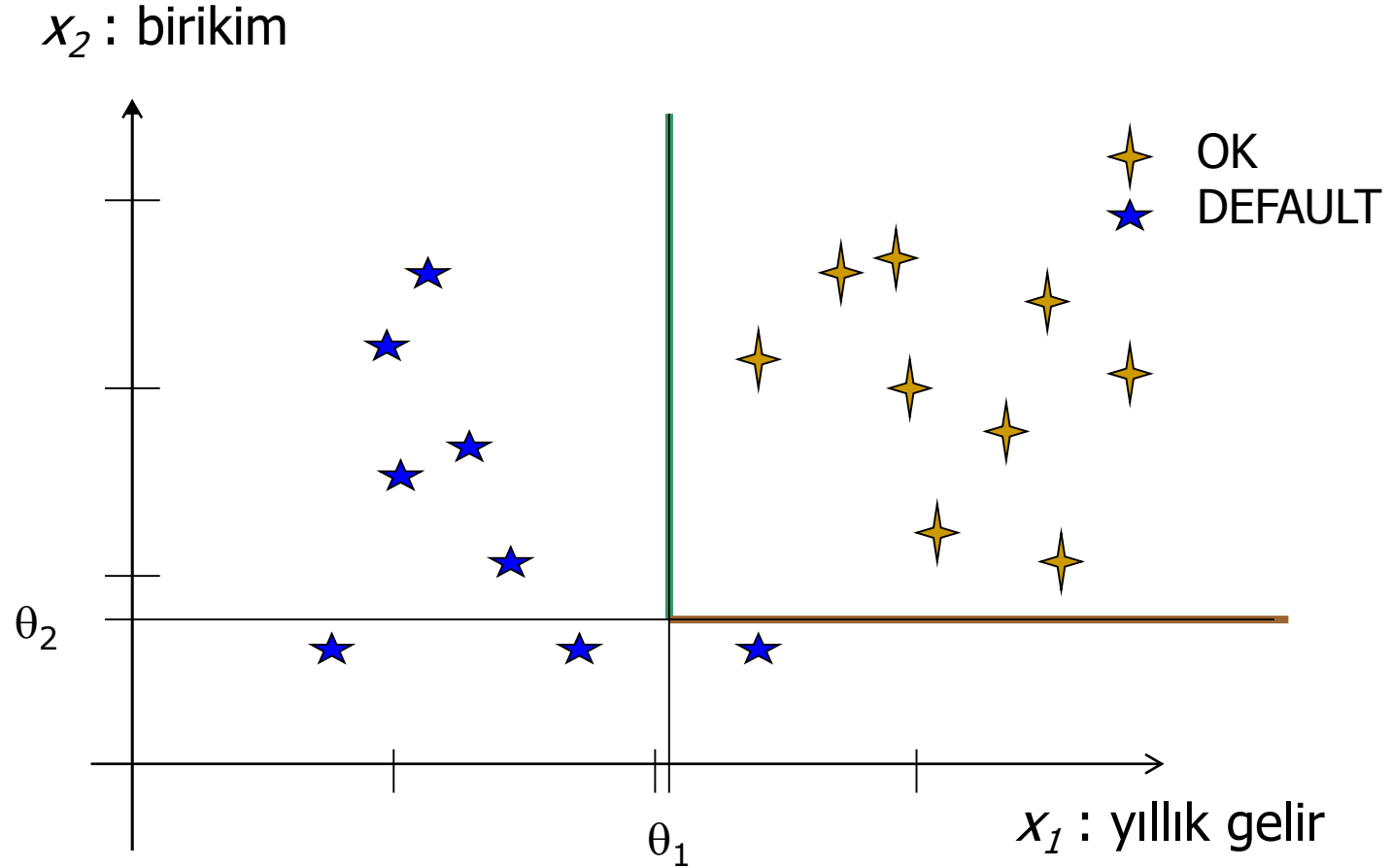
Örnek

birikim



Sınıflandırma

Örnek çözüm

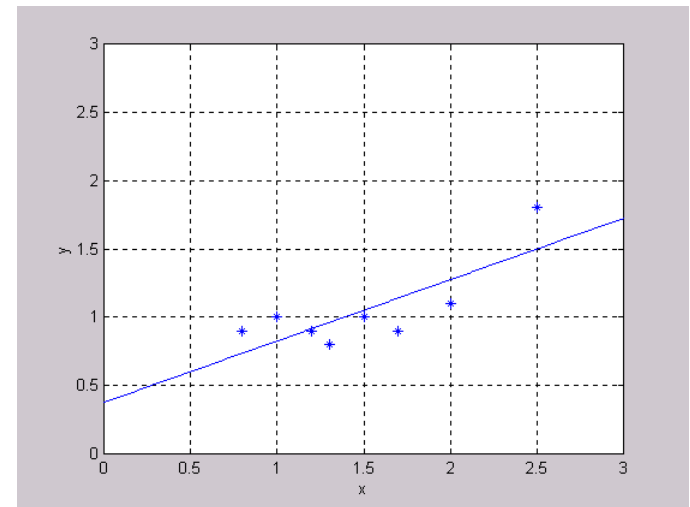
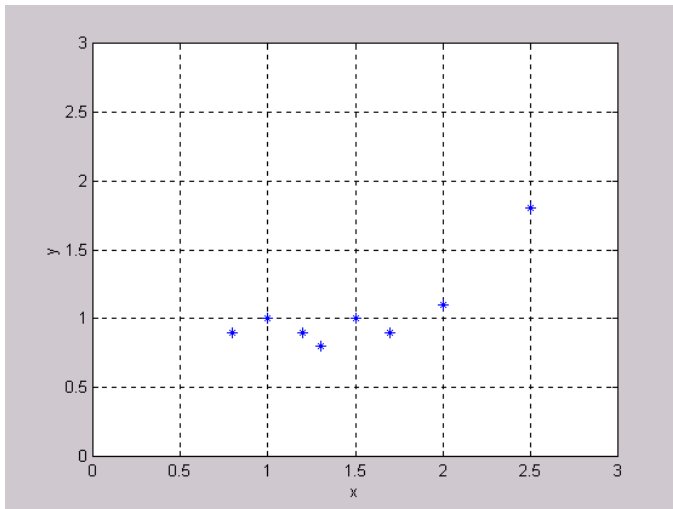


KURAL: EĞER yıllık gelir $> \theta_1$ VE birikim $> \theta_2$
İSE OK DEĞİLSE DEFAULT

Regresyon

(Eğri Uydurma, Fonksiyon Yakınsama)

- Sürekli değişkenlerin öngörüsü regresyon (eğri uydurma) olarak adlandırılan bir istatistiksel yöntemle tespit edilebilir.
- Regresyon analizinin amacı değişik girdi değişkenlerini çıktı değişkeni ile ilişkilendirecek en iyi modelin çıkarılmasıdır.

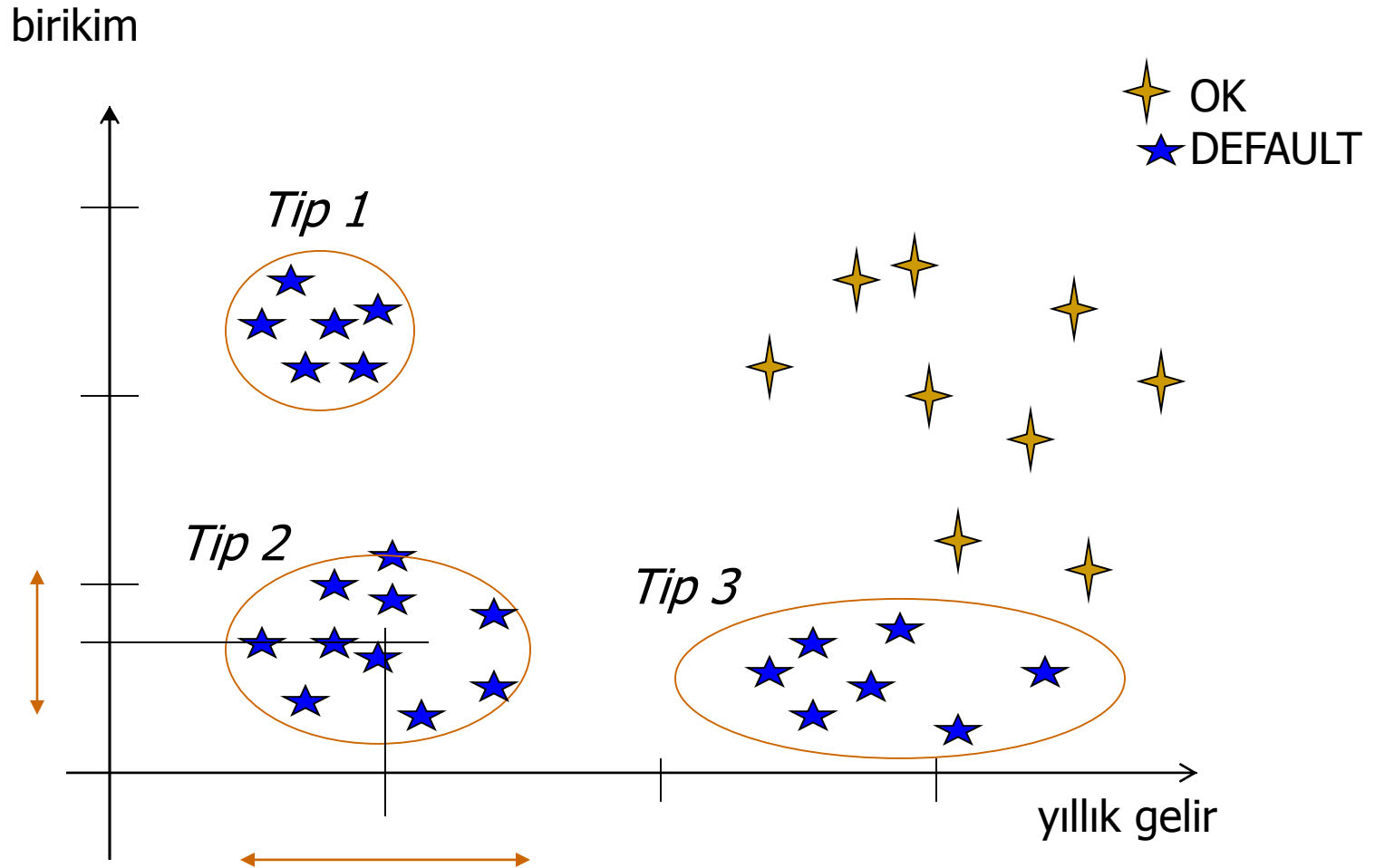


Kümeleme

- Kümeleme bir **eğitici**siz öğrenme ile gerçekleştirilir (unsupervised)
- Küme: Birbirine **benzeyen** nesnelerden oluşan gruptur.
 - Aynı kümedeki örnekler birbirine **daha çok benzer**
 - Farklı kümedeki örnekler birbirine **daha az benzer**
- Benzerlik ölçütü?

Kümeleme

Örnek



Birliktelik analizi

- Birliktelik analizi büyük veri kümeleri arasında birliktelik ilişkilerini bulur.
 - Belirli bir veri kümesinde yüksek sıklıkta birlikte görülen öznitelik değerlerine ait ilişkisel kuralların keşfidir.
- Sonuçlar birliktelik kuralları ($A \rightarrow B$) olarak sunulur.
- Birliktelik kurallarının kullanıldığı en yaygın örnek market sepeti uygulamasıdır.
 - Market sepet analizi, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını belirlemeye çalışır.

Marketlerde birliktelik kuralı keşfi

□ Örnek

<i>TID</i>	<i>Ürünler</i>
1	Ekmek, Kola, Süt
2	Bira, Ekmek
3	Bira, Kola, Çocuk Bezi, Süt
4	Bira, Ekmek, Çocuk Bezi, süt
5	Kola, Çocuk Bezi, Süt

Bulunan kurallar:

{Süt} --> {Kola}

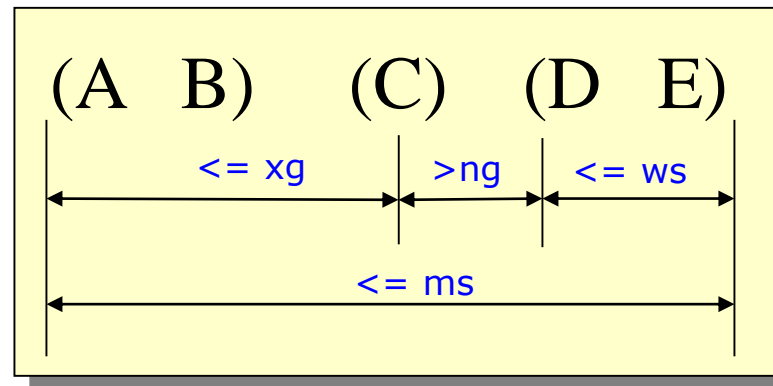
{Çocuk Bezi, Süt} --> {Bira}

Sıralı örüntü madenciliği

- Bir nesne kümesinde her nesnenin kendine ait bir zaman çizelgesi olduğu durumda (örnek: t zamanında, A olayı gerçekleşti), farklı olaylar arası güçlü sıralı birliktelik kuralları çıkarmaktır.

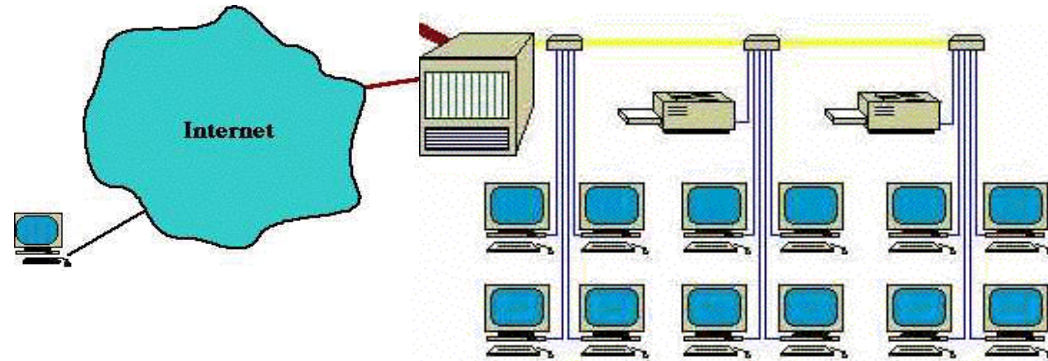
(A B) (C) (D E)

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



İstisna Analizi

- Normal davranışlardan ve eğilimlerden çok farklı sapmaları belirlemede kullanılır.
- Uygulamalar:
 - Kredi Kartı Yolsuzluğu Tesbiti
 - Ağ Saldırı (Intrusion) Tesbiti



Veri Madenciliğinde Yaşanan Zorluklar

- ❑ Veri Boyutu ve Ölçeklenebilirlik
- ❑ Karmaşık ve Heterojen Veri
- ❑ Veri Kalitesi
- ❑ Verinin Sahipleri ve Dağıtılması
- ❑ Gizlilik Koruması
- ❑ Sürekli Güncellenen Veri (Streaming Data)