Veri Madenciliği

Doç. Dr. Ufuk ÇELİK
Bandırma Onyedi Eylül Üniversitesi
Yönetim Bilişim Sistemleri
www.ufukcelik.com.tr

Veri Madenciliği Kavramları

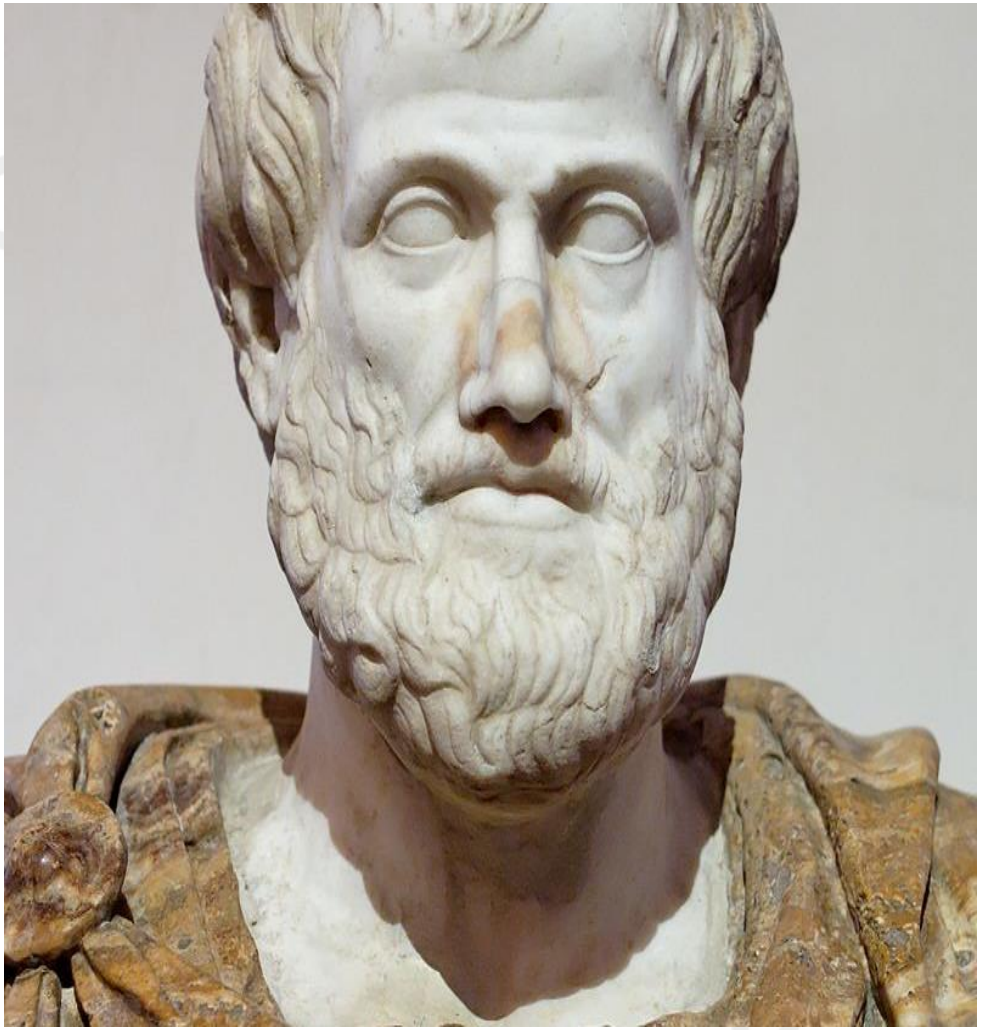
- Veri: bilişimde olgu, kavram yada komutların, iletişim, yorum ve işlem için elverişli biçimsel gösterimi (satış rakamları, döviz fiyatları, kar vs.)
- Enformasyon: en genel anlamda belirli ve görece dar kapsamlı bir konuya ilişkin derlenmiş bilgi parçası
- Bilgi: insan aklının alabileceği gerçek, olgu ve ilkelerin tümüdür
- Olgu: bir takım olayların dayandığı neden ya da bu nedenlerin yol açtığı sonuç (reklam ile satış rakamlarının artması)
- Veri Madenciliği: geniş ölçekli, potansiyel faydası bulunan veriler arasından anlamlı bilgiyi keşfetme işlemidir.
- Örüntü: veri içindeki benzer yapılar

Bilgi Piramidi

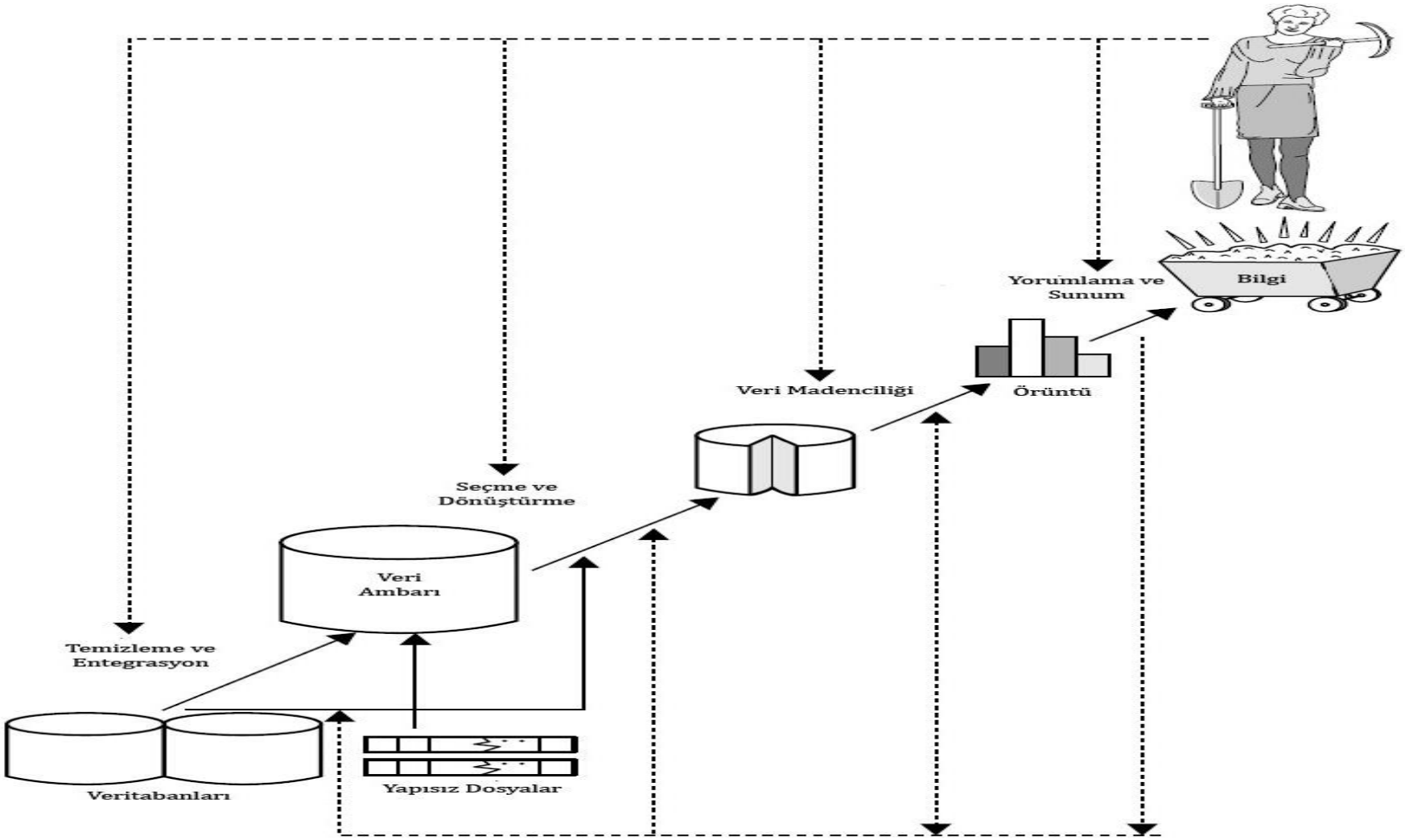
Anlama	İçerik
Keşfetme, Buluş, Muhakeme, Değerler	Bütünü ilişkilendirme, Birleştirme
Enformasyon işleme, Yeniden şekillendirme, Öğrenme, Yapma	Bütünü oluşturma
Veri işleme, Düzenleme, Raporlama	Parçaları ilişkilendirme
Semboller	Parçalara ulaşma

Aristoteles

“Kimse tesadüfle veya onun vasıtasıyla doğru ve akıllı olamaz!”



Veri Madenciliği Aşamaları

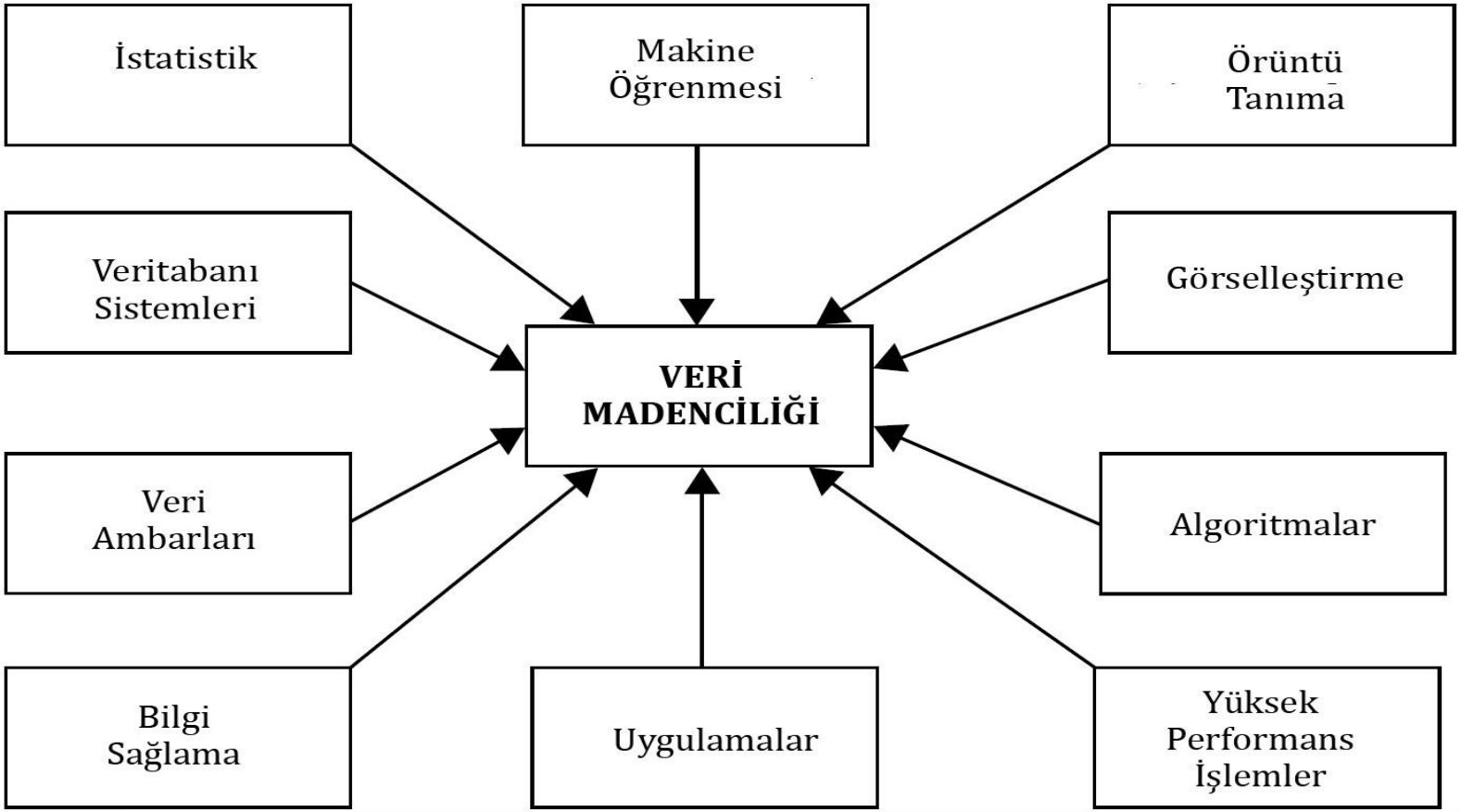


Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Third Edition, USA, The Morgan Kaufman Publishers, 2012, ISBN 978-0-12-381479-1

Veri, Madenciliği Aşamaları

1. Veri temizleme (gürültülü ve tutarsız verileri çıkarmak)
2. Veri bütünleştirme (birçok veri kaynağını birleştirebilmek)
3. Veri seçme (yapılacak olan analizle ilgili olan verileri belirlemek)
4. Veri dönüşümü (verinin veri madenciliği tekniğinden kullanılabilecek hale dönüşümünü gerçekleştirmek)
5. Veri madenciliği (veri örüntülerini yakalayabilmek için akıllı metotları uygulamak)
6. Örüntü değerlendirme (bâzı ölçümlere göre elde edilmiş bilgiyi temsil eden ilginç örüntüleri tanımlamak)
7. Bilgi sunumu (mâdenciliği yapılmış olan elde edilmiş bilginin kullanıcıya sunumunu gerçekleştirmek).

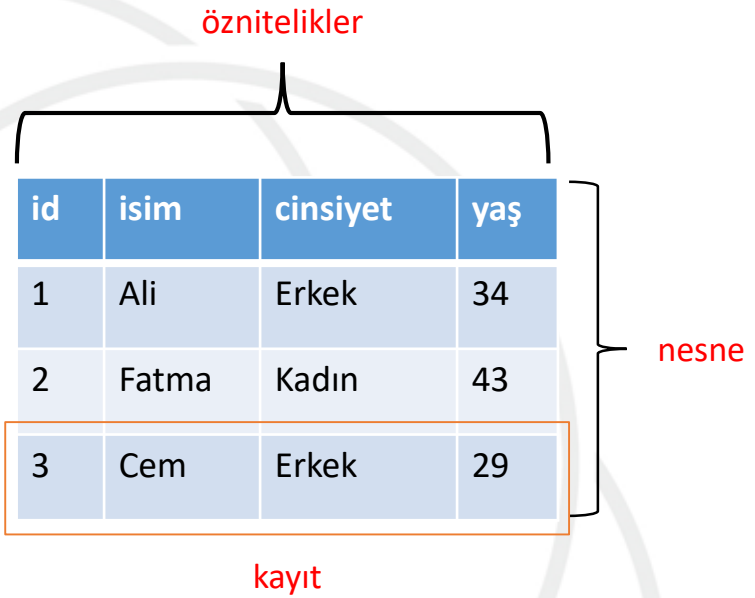
Veri Madenciliği İle Alakalı Teknolojiler



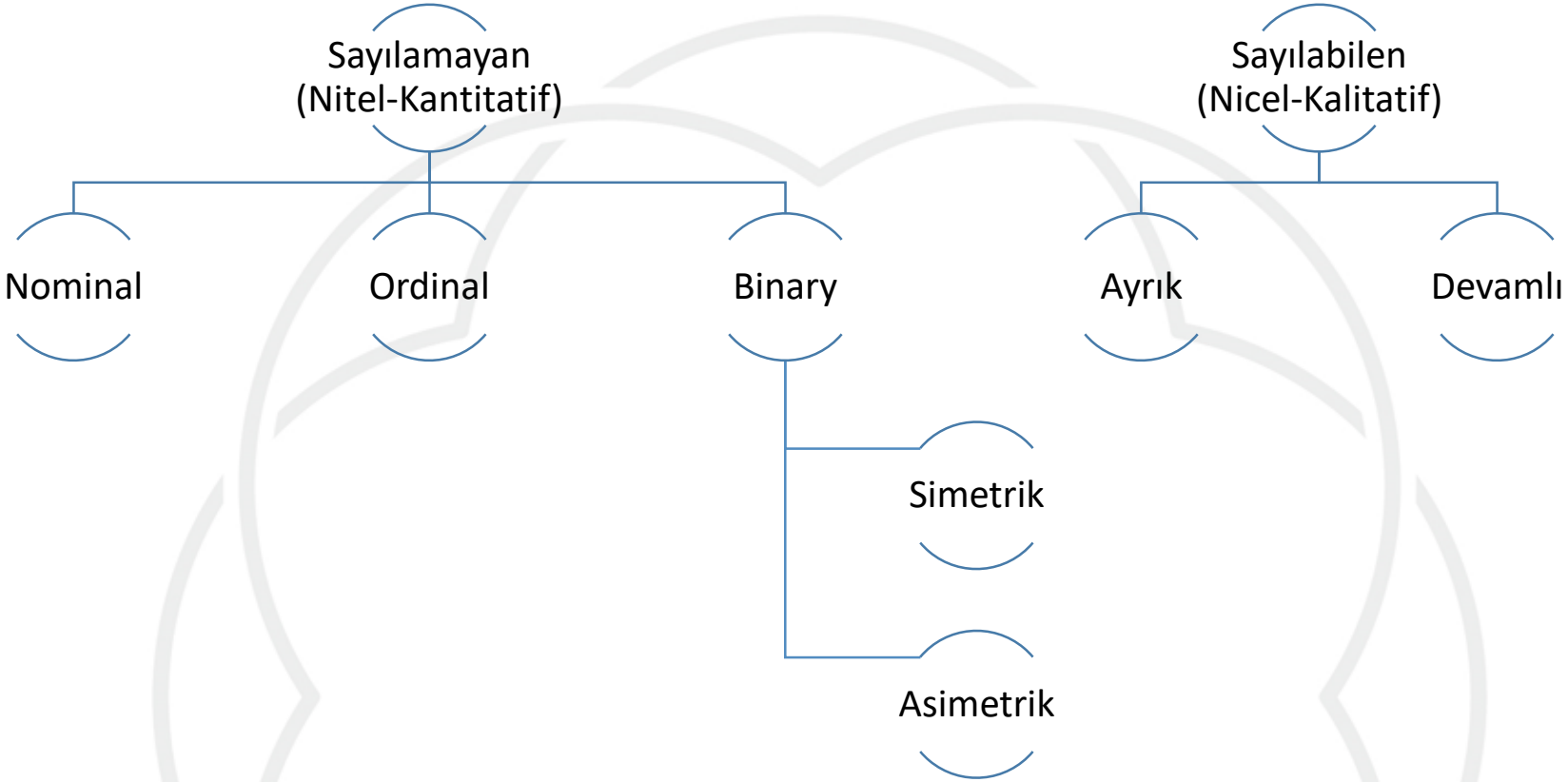
Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Third Edition, USA, The Morgan Kaufman Publishers, 2012, ISBN 978-0-12-381479-1

Veri Kavramları

- Nesne: öznitelikleri barındıran kayıtlar (insanlar)
- Öznitelik: Bir nesneye ait temel bir özellik (isim, cinsiyet, yaş)
- Kayıt: Bir duruma ait öznitelikler topluluğu (tek bir insan)



Öznitelik Türleri



Sayılamayan Öznitelik Türleri

- Nominal: sayısal büyüklük ifade etmeyen kategorik veri tipi
  - renk, ırk, kan grupları gibi…
- Ordinal: belli bir ölçüğe göre sıralanan veriler
  - notlar (AA, BA, BB, BC vs.) beğeni (memnun, kararsız, memnun değil)
- Binary: sadece iki değer içeren nominal veri
  - Simetrik: ikisi de aynı değerde önemli olan
    - Cinsiyet
  - Asimetrik: ikisi de farklı derecede önemli olan
    - Hasta kanser veya değil, Dersten kaldı veya geçti

Sayılabilen Öznitelik Türleri

- Nümerik: tamsayı veya gerçek sayılardan oluşan verilerdir…
  - Aralıklı Ölçeklendirilmiş: sıcaklık veya tarih gibi belli aralıklara sahip değerler
  - Oranlı Ölçeklendirilmiş: genişlik, yükseklik, enlem, boylam gibi değerler
- Ayrık: sonlu sayıda belli bir aralık cinsinden sayısal veya kategorik veriler
  - meslekler (öğretmen, şef vs.), posta kodları (34800, 10200) gibi…
- Devamlı: süreklilik arz eden nümerik veriler
  - en (125 cm, 46 metre), yaş gibi…

Öznitelik Türleri

- Sürekli Öznitelikler: sonlu sayıda veya sayılabilen sonsuz sayıda gerçek değerler içeren veriler
  - Saç rengi, sigara kullanma durumu, tıbbi test sonuçları vs.
  - en, boy, yükseklik, sıcaklık
- Ayrık veya Süreksiz Öznitelikler: sadece iki değer içeren nominal veri
  - hasta veya sağlıklı, sigara içen veya içmeyen
- Ordinal: belli bir ölçüğe göre sıralanan veriler
  - notlar (AA, BA, BB, BC vs.) beğenme (memnun, kararsız, memnun değil)
- Sayısal Aralık: belli bir aralık cinsinden nümerik veriler
  - tarih, derece
- Sayısal Oran: degerlerin, oransal olarak karşılaştırılabildiği veriler
  - boyut, sıcaklık, deneyim yılı

Veri Kümelerinin Tipleri

- Kayıt (Çizgisel)
  - – Veri matrisi
  - – Döküman verisi
  - – İşlem (Transaction) verisi
- Grafik
  - – World Wide Web
  - – Moleküler yapılar
- Sıralı
  - – Uzaysal veri
  - – Geçici veri
  - – Ardışık veri
  - – Genetik dizi verisi

Veri Matrisi

- Eğer veri nesneleri sayısal özniteliklerin sabit bir kümesine sahipse o zaman veri nesneleri her bir boyutun ayrık bir özniteligi sunduğu çok boyutlu uzayda noktalar olarak düşünülebilir.
- Böylesi veri setleri m adet satır ve n adet sütunun bulundugu (m x n) boyutlu matris ile sunulabilir.
- Her bir nesne için n sütun ve bir satır bulunur.

X yükleme	Y yükleme	Uzaklık	Kalınlık	Kesim
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.25	2.8	1.1

Döküman Verisi

- Her bir doküman bir terim vektörü haline gelir,
  - Her bir terim, vektörün bir bileşenidir (öznitelik),
  - Her bir bileşenin değeri doküman içerisinde ilgili terimin kaç kez tekrarlandığı ile ilgilidir.

Kelimeler	skor	takım	kaptan	kaleci	hakem	saha	antrenör
tweet 1	3	6	4	0	2	5	1
tweet 2	2	2	4	5	0	3	8
tweet 3	6	8	2	4	6	0	4

Bir futbol maçı hakkında atılan tweetlerdeki geçen kelimelerin sayıları

İşlem Verisi

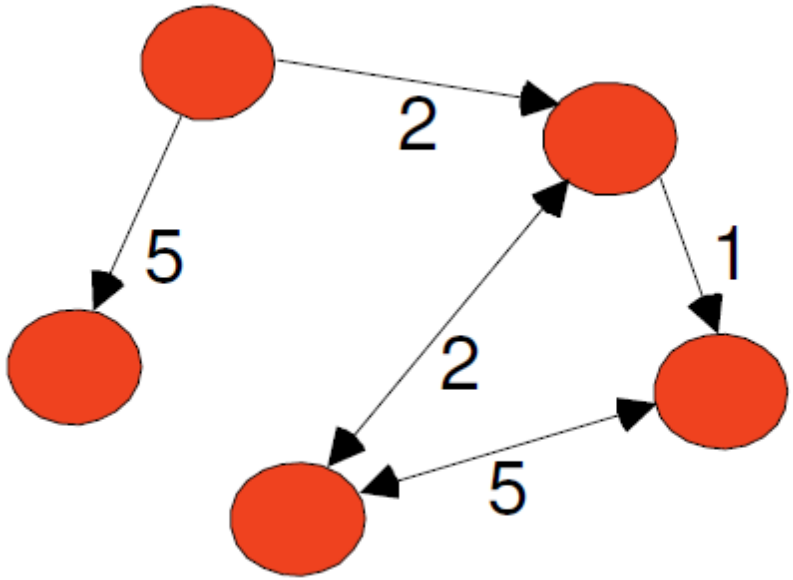
- Kayıt verisinin özel bir tipidir,
  - Her bir işlem (transaction) elemanların bir kümesini içermektedir.
  - Örneğin, bir dükkan düşünün. Burada, ödemesi yapılan ürünlerin bir kümesi bir işlem kaydını verir.

TID	ürünler
1	Ekmek, kola, süt
2	Bira, ekmek
3	Bira, kola, bebek bezi
4	Bebek bezi, süt, mama

Grafik Verisi

- Jenerik grafikler veya HTML linkleri gibi

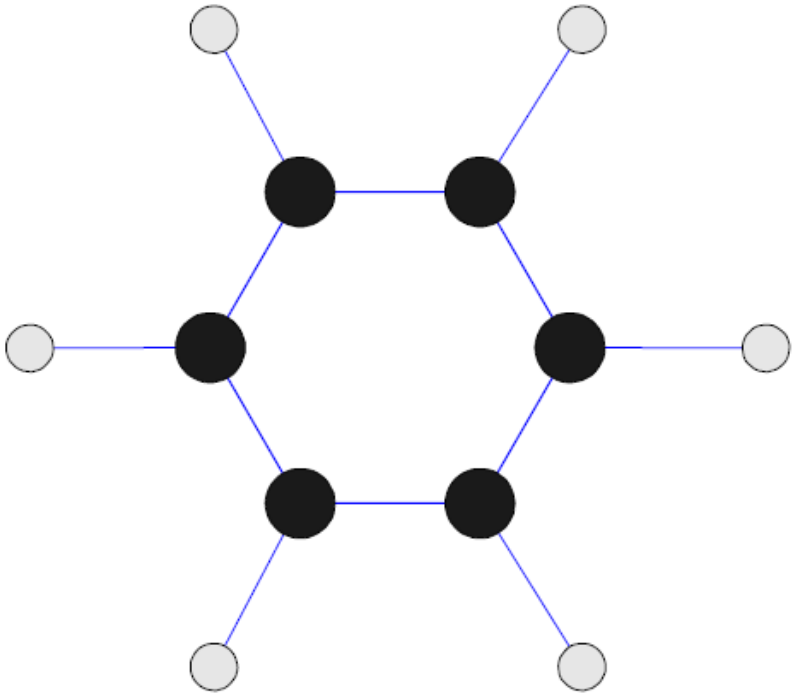
```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of
Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear
System Solvers
```



Kimyasal Veri

- Benzen Mokülü

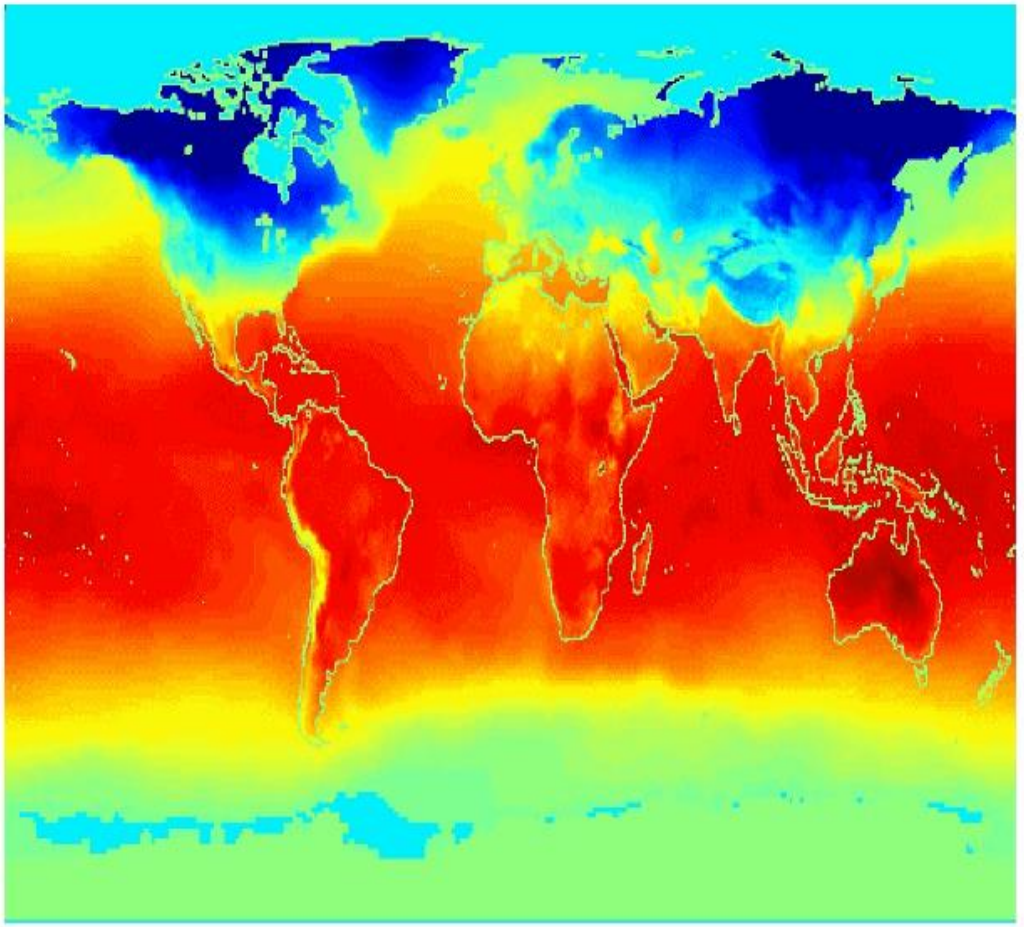
- C_6H_6



Sıralı Veri

- Uzaysal-geçici veri

Karaların ve Okyanusların Ocak ayı ortalama sıcaklıkları



Sıralı Veri

- Gen dizisi verisi

DNA genleri gibi biyoistatistik veriler

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Veri Kalitesi

- Önce çıkan bazı faktörler:
  - Veri kalitesi problemleri hangi çeşitlerdedir?
  - Veri ile ilgili problemleri nasıl tespit edebiliriz?
  - Bu problemlerle ilgili olarak ne yapabiliriz?
- Veri kalitesi ile ilgili problemler:
  - Gürültü ve taşmalar
  - Kayıp değerler
  - Veri tekrarı

Gürültü

- Orijinal değerlerin bozulması
  - Eksik değerler:
    - Meslek = ‘ ‘
  - Tekrarlayan veri:
    - Aynı kişinin birden fazla e-posta adresi
  - Hatalı veri girişi:
    - Maaş = -3400
  - Tutarsız veri:
    - Yaş 35 ama doğum tarihine göre 45
    - Bir kaynakta öznitelik değeri ‘isim’ diğerinde ‘ad’
    - Bir kayıtta oylama değerleri ‘1,2,3’ diğerinde ‘A,B,C’

ARAŞTIRMA

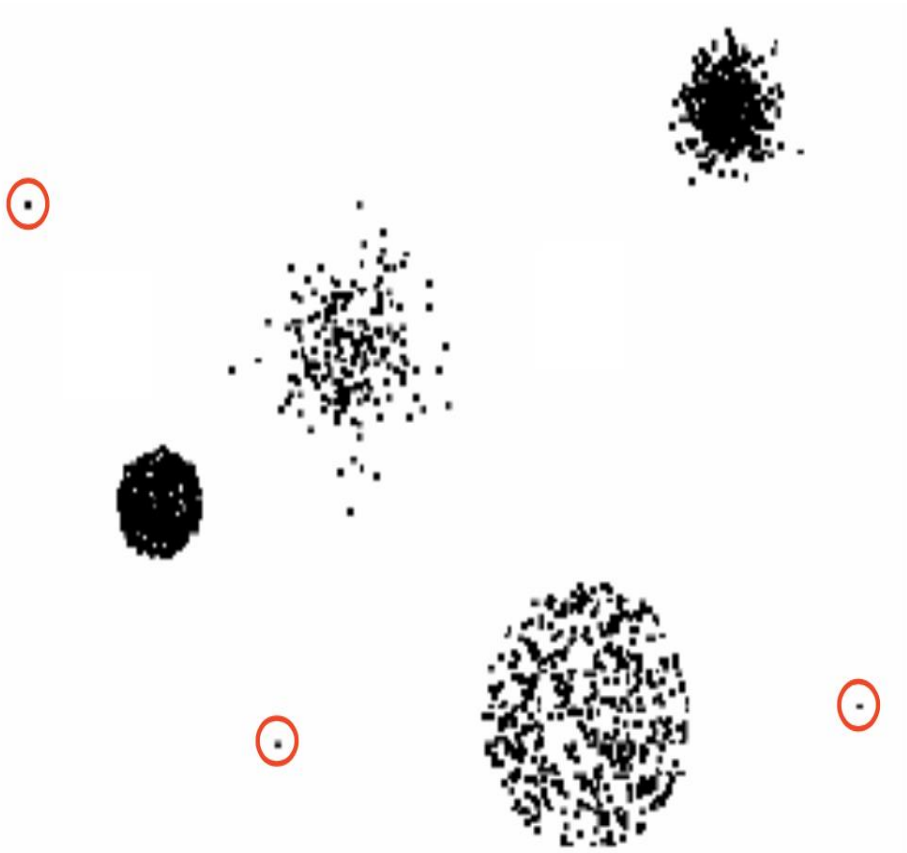
- Gürültülü verileri nasıl düzeltiriz?

Taşmalar

- Taşma (outlier), veri kümesinde, diğer nesnelerden ciddi şekilde farklı olan veri nesnelerinin gösterdiği karakteristiktir.

Taşma

- Taşma (outlier), veri kümesinde, diğer nesnelerden ciddi şekilde farklı olan veri nesnelerinin gösterdiği karakteristiktir.



Veri Ön İşleme

- Veri temizleme
  - *Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme*
- Veri birleştirme
  - *Farklı veri kaynağındaki verileri birleştirme*
- Veri dönüşümü
  - *Normalizasyon ve biriktirme*
- Veri azaltma
  - *Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma*

Veriyi Tanımlayan Özellikler

- Simetrik veya asimetrik veriyi anlamak
  - Merkezi eğilim (central tendency): ortalamalar, Ortanca (medyan), Mod
  - varyasyon, yayılma, dağılım
- Verinin dağılım özellikleri
  - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
  - verinin dağılımı
  - kutu grafiği çizimi ve sıklık derecesi incelemesi