Veri Madenciliği

2023-2024

Güz

Ders 1

Ders Tanıtımı

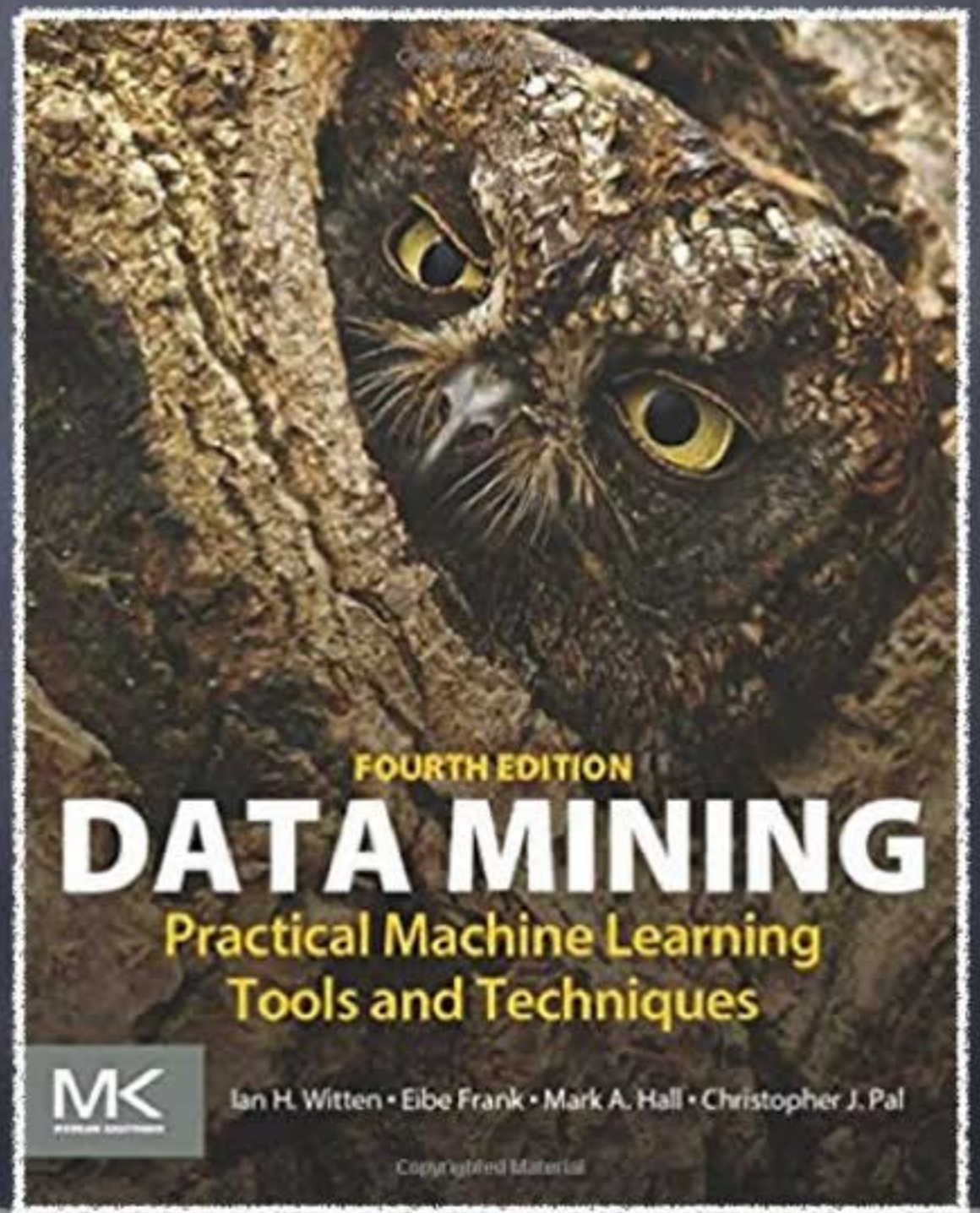
Veri Madenciliğine Giriş

Bonferroni'nin İlkesi

Doç. Dr. Ali YILMAZ

Dersin Kitabı

- Data Mining: Practical Machine Learning Tools and Techniques, 4th Ed., by Ian Witten, Eibe Frank, Mark Hall, and Christopher Pal (Morgan Kaufmann Publishers. 2017. ISBN: 978-0-12-804291-5)



Ders Tanımı

- Veri Madenciliği, verilerdeki yararlı, muhtemelen beklenmedik kalıpların keşfidir.

- Büyük veri yığınlarındaki örüntüleri keşfetme, yorumlama ve görselleştirme tekniklerini öğreneceğiz.

Ders Tanımı

- Konular şunları içerir:
  - Sınıflandırma
  - Kural tabanlı öğrenme
  - Karar ağaçları
  - Birliktelik kuralları
  - Veri görüntüleme
- Teorik Temel: (Derinlemesine değinmeyeceğiz)
  - Yapay zeka
  - Makine öğrenme
  - İstatistiksel analiz, vesaire.

Proje, kısaca

Çalışma grubunuzla birlikte çalışın:

- Çözmek için bir veri madenciliği problemi seçin.
- Bir veri kümesi bulun, dönüştürün ve temizleyin.
- Bir model ve birkaç algoritma seçin.
- Testleri ve sonuçları açıklayın.
- Bir rapor yazın ve bir sunum oluşturun.

Neden Veri Madenciliği?

Muazzam Veri Büyümesi

Terabaytlardan petabaytlara

- Veri toplama ve veri kullanılabilirliği
  - Otomatik veri toplama araçları, veritabanı 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 hala açız!

"Gereklilik buluşun anasıdır"—Veri madenciliği—Büyük veri kümelerinin otomatik analizi

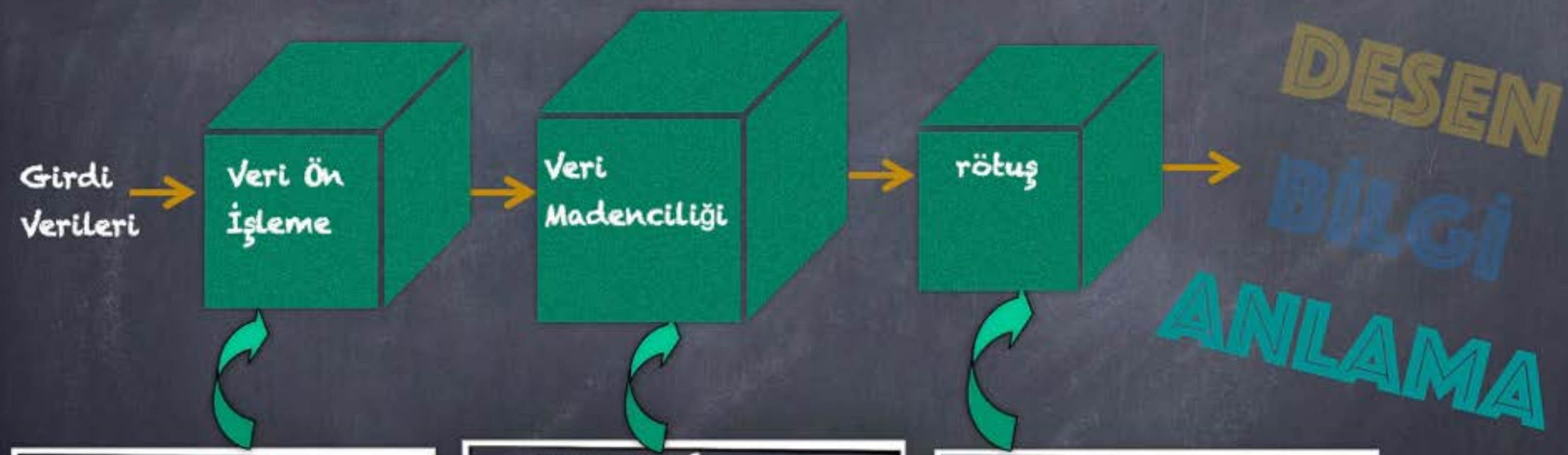
Veri Madenciliği Nedir?

- Verilerdeki yararlı, muhtemelen beklenmedik kalıpların keşfi.
  - Büyük miktarda veriden ilginç (önemsiz, örtük, önceden bilinmeyen ve potansiyel olarak yararlı) kalıpların veya bilgilerin çıkarılması
  - Yardımcı konular:
    - Veri temizleme: sahte verilerin tespiti.
      - örneğin, yaş = 150.
    - Görselleştirme: megabaytlarca çıktı dosyasından daha iyi bir şey.

Veri Madenciliği Nedir?

- Alternatif İsimler
  - Veritabanlarında (KDD) bilgi keşfi (madencilik),
  - bilgi çıkarma,
  - veri/kalıp analizi,
  - veri arkeolojisi,
  - veri tarama,
  - bilgi toplama,
  - iş zekası vb.

KDD Süreci: Makine Öğrenimi ve İstatistiklerden Tipik Bir Görünüm



Veri Madenciliğinin Kullanımları

1. Veri özelliklerini genelleştirin, özetleyin ve karşılaştırın, örneğin; kuru ve ıslak bölge

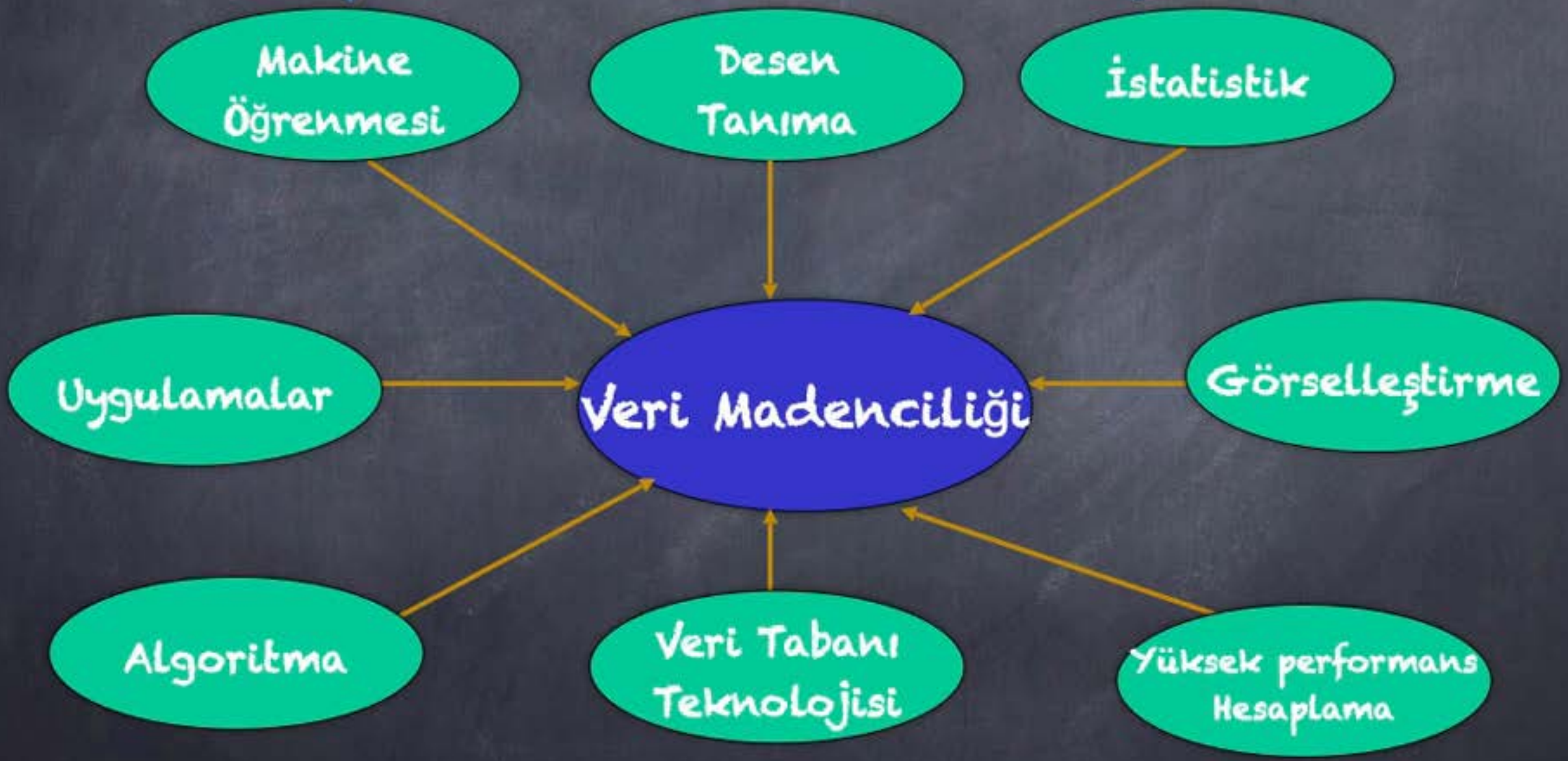
2. İlişkilendirme ve Korelasyon Analizi

Örneğin; Teknosa'da ne tür ürünler birlikte satın alınır?

3. Sınıflandırma

Örneğin; Ülkeleri iklime göre sınıflandırın veya arabaları gaz kilometresine göre sınıflandırın

Veri Madenciliği: Çoklu Disiplinlerin Birleşmesi



Neden? Muazzam miktarda veri; Yüksek veri karmaşıklığı; Yeni ve gelişmiş uygulamalar

Veri Madenciliği: Hibrit Alanlar

- **Veritabanları:** Yapılandırılmış Sorgu Dili, İlişkisel Veri, Ölçeklenebilirlik
- **Algoritmalar:** Tasarım, Karmaşıklık, Veri Yapıları
- **Makine Öğrenimi:** Karar ağaçları, Sinir Ağları, Genetik Algoritmalar, vb.
- **İstatistikler:** Bayes Teoremi, Regresyon, Zaman Serileri Analizi

Sosyal Etkiler

- **Gizlilik:** Sizi hedefleyen reklamlar artık "sayılarda kaybolmayı" (Tavsiye edebilir miyiz...)

- **Profil oluşturma:** Kredi kartı riski, Terör riski vb.

Cevapların Anlamlılığı

- Büyük bir veri madenciliği riski, anlamsız kalıpları "keşfedeceksiniz".

- İstatistikçiler buna Bonferroni ilkesi diyor: (kabaca) ilginç desenler için veri miktarınızın destekleyeceğinden daha fazla yere bakarsanız, anlamsız sonuçlar bulmanız kaçınılmazdır.

Bonferroni Prensibi Örnekleri

1. TBF'ye (Toplam Bilgi Farkındalığı) büyük bir itiraz, o kadar çok belirsiz bağlantı arıyordu ki, sahte olan ve dolayısıyla masumların mahremiyetini ihlal eden şeyler bulacağından emindi.

2. Ren Paradoksu: Bilimsel araştırmanın nasıl yapılmayacağının harika bir örneği.

Stanford Üniversitesi'nden Profesör Jeff Ullman "TIA" Öyküsü Örneği

- Bazı kötü niyetli grupların ara sıra otellerde kötülük yapmayı planlamak için toplandıklarına inandığımızı varsayalım.

- Aynı gün aynı otelde en az iki kez kalmış (ilgisiz) kişileri bulmak istiyoruz.

Ayrıntılar

- 10^9 kişi izleniyor.*
- 1000 gün.
- Her kişi zamanın %1'inde bir otelde kalıyor (1000'de 10 gün).
- Oteller 100 kişiliktir (yani 10^5 otel).
- Herkes rastgele davranırsa (yani, kötü niyetli kişiler yoksa) veri madenciliği şüpheli bir şey tespit edecek mi?

* Bu, Çin'deki (veya bu konuda Hindistan'daki) insan sayısından daha azdır.

Ullman Hesaplamaları - (1)

- Verilen p ve q kişilerinin d gününde aynı otelde olma olasılığı:

- Verilen d_1 ve d_2 günlerinde p ve q'nun aynı otelde olma olasılığı:

- gün çiftleri:

Ullman Hesaplamaları - (2)

- p ve q'nun bazen iki gün kadar aynı otelde olma olasılığı:

- insan çiftleri:

- Beklenen "şüpheli" insan çifti sayısı:

Sonuç

- Diyelim ki aynı otelde kesinlikle iki kez kalan 10 çift kötülük var.

- Analistler, 10 gerçek vakayı bulmak için çeyrek milyon adayı elemek zorunda.

  - Olmayacak.

  - Fakat bu düzeni/tasarımı nasıl iyileştirebiliriz?

Kıssadan Hisse

- Bir mülk ararken (örneğin, "iki kişinin aynı otelde iki kez kalmış mı"), mülkün o kadar çok olasılığa izin vermediğinden emin olun ki, rastgele veriler kesinlikle "ilgi çekici" gerçekler üretecektir.

Ren Paradoksu - (1)

- Joseph Rhine 1950'lerde bazı insanların Ekstra Duyusal Algıya (EDA) sahip olduğunu öne süren bir parapsikologdu.
- Deneklerden kırmızı veya mavi olmak üzere 10 gizli kart tahmin etmelerinin istendiği bir deney (benzeri bir şey) tasarladı.
- Neredeyse 1000'de 1'inin EDA'ya sahip olduğunu keşfetti - 10'unu da doğru yapabildiler!

Ren Paradoksu - (2)

- Bu insanlara EDA'ları olduğunu söyledi ve onları aynı tipte başka bir test için çağırdı.
- Ne yazık ki, neredeyse hepsinin EDA'larını kaybettiğini keşfetti.

PEKİ SONUÇ NE OLDU?

Kıssadan hisse:

- Bonferroni İlkesini anlamak, sonuçların benzer şekilde yanlış yorumlanmasından kaçınmanıza yardımcı olabilir.