Veri Bilimine Giriş:

Veri Okuryazarlığı

Prof. Dr. A. Egemen YILMAZ
Ankara Üniversitesi Mühendislik Fakültesi Elektrik-Elektronik Mühendisliği Bölümü
Ankara Üniversitesi Mühendislik Fakültesi Yazılım Mühendisliği Bölümü
Ankara Üniversitesi Fen Bilimleri Enstitüsü Disiplinler Arası Yapay Zeka Ana Bilim Dalı
Ankara Üniversitesi Akıllı Sistemler ve Teknolojiler Uygulama ve Araştırma Merkezi (ASTAM)

Veri (*Data*) Nedir?

- Latince'de "gerçek, reel" anlamına gelen "*datum*" ("*Data*", "*datum*"un çoğulu
- Sözlük anlamı olarak gerçeklik temel alınsa da, günümüzdeki kullanım şekliye her zaman somut gerçeklik göstermemekte
- Kavramsal anlamda veri: 'kayıt altına alınmış her türlü ölçüm, gözlem, olay, durum, fikir'

Veri Okuryazarlığı Nedir?

- «Verileri okuma, verilerle çalışma, verileri analiz etme ve veriler üzerinden tartışma» yeteneği
- Tüm vatandaşların sahip olması gereken bir beceri
- İnsanların, bağlı bulundukları kuruluşlara daha faydalı olacak şekilde kararlar vermesini ve eyleme geçmesini sağlayacak bir unsur
- Normal okuryazarlığın harflerden seslere dönüştürme, oradan da anlam çıkarma yeteneği olması gibi; verinin bilgiye dönüştürülmesi, oradan da makul-mantıklı kararlar alınarak eyleme geçilebilmesi yeteneği

Veri Okuryazarlığı Nedir?

- Verinin okunması, yazılması, belirli bir bağlam üzerinde tartışmaya açılması
- Verinin kaynağı da göz önünde bulundurularak verinin nasıl işlenmesi gerektiği konusunda bir anlayış oluşturulması
- Uygulanması gereken analitik yöntemler ve tekniklerin bilinmesi
- Verinin işlenmesi sonucunda ortaya çıkacak fayda ve değer; hatta olası uygulamalar hakkında bilgi sahibi olunması
- Grafiklerin nasıl okunacağının bilinmesi, buradan nasıl bulgular çıkarılacağına ve yargılara varılacağına hakim olunması
- Manipüle edilmiş verinin farkına varılması, bu gibi istismarlara karşı dikkatli ve uyanık olunması

Yeni Görevler – Yeni Kariyerler

- Veri Analisti
- İş Zekası Geliştiricisi

Büyük Veri Kavramı ile Birlikte:

- Veri Mühendisi
- Veri Bilimcisi

Yeni Görevler – Yeni Kariyerler

Veri Analisti

- Kurumunda veri sorgulayabilen ve işleyebilen, raporlar sunan, verileri özetleyen ve görselleştirebilen deneyimli veri uzmanı
- Bir sorunu çözmek için mevcut araç ve yöntemlerden nasıl yararlanılacağına dair güçlü bir anlayışa sahiptir
- Kurun genelindeki kişilerin geçici raporlar ve çizelgelerle spesifik sorguları anlamalarına yardımcı olur
- Ancak büyük verinin analiziyle uğraşmaları veya belirli problemlere yeni algoritmalar geliştirmek için matematiksel veya araştırma geçmişlerine sahip olmaları beklenmemektedir
- İstatistik, veri akışı, veri görselleştirme, keşifsel veri analizi gibi bazı temel becerilerde bilgili olmaları beklenmektedir

Kaynak: toptalent.co

Yeni Görevler – Yeni Kariyerler

İş Zekası Geliştiricisi

- Raporlama ihtiyaçlarını anlamak ve kurum için iş zekası ve raporlama çözümleri oluşturmak, gereksinimlerini toplamak, tasarlamak ve oluşturmak için iç paydaşlarla daha yakından etkileşimde bulunan veri uzmanı
- Yeni ve mevcut veri ambarlarını, ETL paketlerini, küpleri, gösterge tablolarını ve analitik raporları tasarlamak, geliştirmek ve desteklemek zorundadır
- Ek olarak hem ilişkisel hem de çok boyutlu veri tabanları ile çalışır ve farklı kaynaklardan gelen verileri entegre etmek için SQL geliştirme becerilerine sahip olmalıdır
- Tüm bu becerileri, kurum genelinde self servis ihtiyaçlarını karşılamak için kullanır
- Ancak genellikle veri analizi yapması beklenmemektedir
- ETL (*Extract-Transform-Load*), rapor oluşturma, OLAP (*On Line Analytical Processing*), web zekâsı, iş nesneleri tasarımı konularında bilgili olması beklenmektedir

Kaynak: toptalent.co

Yeni Görevler – Yeni Kariyerler

Veri Mühendisi

- Veri bilimcileri tarafından analiz edilecek "büyük veri" altyapısını hazırlayan veri uzmanı
- Çeşitli kaynaklardan veri tasarlayan, derleyen, birleştiren ve büyük veriyi yöneten yazılım mühendisi
- Ardından verilerin kolay erişilebilir olduğu ve sistemin sorunsuz çalıştığından emin olduktan sonra hedeflerinin doğrultusunda şirketlerinin büyük veri ekosistemi performansını optimize etmesini sağlamak için çalışır
- Ayrıca, büyük veri kümelerinin üstüne bazı ETL'ler yapabilir ve veri bilimcileri tarafından raporlama veya analiz için kullanılabilecek büyük veri ambarları oluşturabilir
- Bunun ötesinde tasarım ve yapıya daha fazla odaklandığı için, genellikle büyük veri için herhangi bir makine öğrenmesi veya analitiği bilmesi beklenmemektedir
- Hadoop, MapReduce, Hive, Pig, Veri akışı, NoSQL, SQL ve programlama konularında yetkin olması beklenmektedir

Kaynak: toptalent.co

Yeni Görevler – Yeni Kariyerler

Veri Bilimcisi

- Ham verileri arıtılmış içgörülere dönüştürebilir
- Kritik iş problemlerini çözmek için istatistik, makine öğrenmesi ve analitik yaklaşımlar uygular
- Öncelikli işlevi, kuruluşların büyük veri hacimlerini değerli ve eyleme geçirilebilir içgörülere dönüştürmelerine yardımcı olmaktır
- Veri analitik becerilerine ek olarak, güçlü programlama becerilerine, yeni algoritmalar tasarlama becerisine, büyük veriyi ele alma ve alan bilgisinde bazı uzmanlıklara sahip olması beklenmektedir
- Ayrıca, bulgularının sonuçlarını görselleştirme teknikleri kullanarak, veri bilimi uygulamaları oluşturarak ya da veri (işletme) sorunlarının çözümleriyle ilgili ilginç hikayeler anlatarak yorumlaması ve sunması beklenmektedir
- Farklı boyut ve şekillerde farklı veri kümeleriyle çalışma deneyimine sahip olmalı ve algoritmalarını büyük boyutlu veriler üzerinde etkin ve verimli bir şekilde çalıştırabilmelidir
- Python, R, Scala, Apache Spark, Hadoop, makine öğrenmesi, derin öğrenme ve istatistik konularında yetkin olmalıdır

Kaynak: toptalent.co

Veri Bilimi veya Veri Analitiği Nedir?

Elimizde bir büst olsun

Söz konusu büst, tarihteki ünlü bir şahsa dair olsun; ayrıca ünlü bir heykeltıraşın eseri olduğu düşünülüyor olsun

Veri Bilimi veya Veri Analitiği Nedir?

İstatistik, gerçeklere dayalı (*factual*) bir bilim dalıdır.

Elimizdeki büste dair ölçümlere (alın genişliği, göz küreleri arasındaki mesafe, vb.) dayalı olarak söz konusu şahsın genel popülasyon içerisinde hangi yüzdelik dilimde olduğunu, standart sapma içerisinde olup olmadığını, vb. söyler.

Veri Bilimi veya Veri Analitiği Nedir?

Veri Analitiği ise çıkarımcı (*inferential*) bir bilim dalıdır.

Büste dair ölçümlere (alın genişliği, göz küreleri arasındaki mesafe, vb.) dayalı olarak söz konusu şahsın boyu, kilosu, ayakkabı numarası gibi değerleri tahmin etmeye çalışır!...

Veri Bilimi veya Veri Analitiği Nedir?

Veri Analitiği ise çıkarımcı (*inferential*) bir bilim dalıdır.

Önceden yeterince veri de varsa, elimizdeki büste dair ölçümlere (alın genişliği, göz küreleri arasındaki mesafe, vb.) dayalı olarak büstü yapılmış şahsa veya büste ilişkin olarak anomali tespiti yapabilir (örneğin şahsın alın genişliği aşırı derecede büyük; veya büst yanlış yapılmış vb.)

Veri Bilimi veya Veri Analitiği Nedir?

Veri Analitiği ise çıkarımcı (*inferential*) bir bilim dalıdır.

Büstün ölçümlerine, A ve B ırklarının antropometrik verilerine dayalı olarak büstü yapılmış olan şahsın A ırkından mı, B ırkından mı olmasının daha yüksek olasılıklı olduğunu tahmin etmeye çalışabilir.

Büstün ölçümlerine ve genel toplumun antropometrik verilerine dayalı olarak büstü yapılmış olan şahsın Akromegali hastası olup olmadığını belirlemeye çalışabilir.

Veri Bilimi veya Veri Analitiği Nedir?

Veri Analitiği ise çıkarımcı (*inferential*) bir bilim dalıdır.

Büstün ölçümlerine ve büstü yaptığı iddia edilen heykeltıraşın diğer eserlerinin özelliklerine dayalı olarak büstün gerçekten de ilgili şahsın eseri olup olmadığını belirlemeye çalışabilir.

vb.

Veri Bilimi veya Veri Analitiği Nedir?

Betimleyici Veri Analitiği (*Descriptive Data Analytics*): Mevcut durumun fotoğrafını çeker; «Ne Olmuş?» sorusunun cevabını verir

Teşhis Edici Veri Analitiği (*Diagnostic Data Analytics*): Mevcut durumun fotoğrafını çeker; «Neden Olmuş?» sorusunun cevabını verir

Öngörücü Veri Analitiği (*Predictive Data Analytics*): Veri değişim eğilimlerini irdeleyerek gelecekte ne olacağını tahmin eder

Reçete Yazıcı Veri Analitiği (*Prescriptive Data Analytics*): What-If senaryoları koşturarak bunlara ilişkin sonuçlar bulur; önerilerde bulunur

Veri Bilimi veya Veri Analitiği Nedir?

Betimleyici Veri Analitiği (*Descriptive Data Analytics*): X Firması, 34 ayrı lokasyondaki şubesi ve 165 çalışanıyla 2023 yılında Y lira kâr elde etmiştir

Teşhis Edici Veri Analitiği (*Diagnostic Data Analytics*): X firmasının 2023 yılındaki karlılığının Y lira seviyesinde kalmasının nedeni, A, B ve C lokasyonlarında bulunan ve verimliliği düşük olan şubeleridir

Öngörücü Veri Analitiği (*Predictive Data Analytics*): İlk 7 aydaki verilere dayalı olarak X Firması'nın 2024 yılındaki kârının Z lira olacağı öngörülmektedir

Reçete Yazıcı Veri Analitiği (*Prescriptive Data Analytics*): X firması, A, B ve C lokasyonlarında bulunan ve verimliliği düşük olan şubelerini kapatıp D, E ve F lokasyonlarında aynı büyüklükte 3 yeni şube açarsa toplam çalışan sayısı sabit kalmak kaydıyla kârlılık oranını %7 artırabilir

Veri Bilimi veya Veri Analitiği Nedir?

Betimleyici Veri Analitiği (*Descriptive Data Analytics*): Firmamızda 30 yaş altı personelin firmada ortalama çalışma süresi 3 yıl, 30 yaş üstü personelin firmada ortalama çalışma süresi ise 10.4 yıldır. Firmamızın personel yaş ortalaması 36.9'dur.

Öngörücü Veri Analitiği (*Predictive Data Analytics*): Mevcut verilere dayalı olarak firmamızın 5 yıl sonra personel yaş ortalaması 8.2 yıl artarak 45.1 olacaktır.

Reçete Yazıcı Veri Analitiği (*Prescriptive Data Analytics*): 30 yaş altı personelin sirkülasyonunu azaltmak için … çalışma usullerinin getirilmesi, … uygun olacaktır.

Veri Bilimi veya Veri Analitiği Nedir?

Kaynak: https://www.usatoday.com/story/news/health/2020/10/23/current-covid-strategies-could-cause-more-than-500-k-deaths-feb-28/3729661001/

Betimleyici Veri Analitiği (*Descriptive Data Analytics*): 15 Ekim 2020 tarihi itibarı ile ABD'de Covid19'a dayalı ölümler 250,000'i aşmış durumdadır

Öngörücü Veri Analitiği (*Predictive Data Analytics*): Şubat 2021 sonu itibarı ile ABD'de Covid19'a dayalı ölümler 500,000'i aşmış olacaktır

Reçete Yazıcı Veri Analitiği (*Prescriptive Data Analytics*): Nüfusun %85'inin maske vb tedbirlere uyması durumunda bu ölümlerin 96,000 tanesi; %95'inin uyması durumunda ise 131,000 tanesi engellenebilir

Veri Bilimi veya Veri Analitiği Nedir?

- Mevcut artış eğilimi devam ettiği takdirde kuraklığın 2030 yılında Asya ülkelerinin %....'sinde büyük sıkıntı oluşturacağı öngörülmektedir. (Öngörücü)

- Yeşil gözün çekinik olmasından ötürü 20XX yılında yeşil gözlü insanların tüm nüfusa oranının milyonda birin altına düşeceği öngörülmektedir. (Öngörücü)

- Tüm Dünya genelinde nüfusun %...'sini kadınlar, %...'sini ise erkekler oluşturmaktadır. (Betimleyici)

- Emniyet kemeri kullanımının %....'e çıkması halinde trafik kazalarında can kayıplarının %... oranında azalacağı öngörülmektedir. (Reçete Yazıcı)

Veri Bilimi veya Veri Analitiği Nedir?

- Mevcut azalma eğilimi devam ettiği takdirde 2035 yılı itibarı ile … hastalığının tamamen ortadan kalkacağı Dünya Sağlık Örgütü tarafından öngörülmektedir. (Öngörücü)

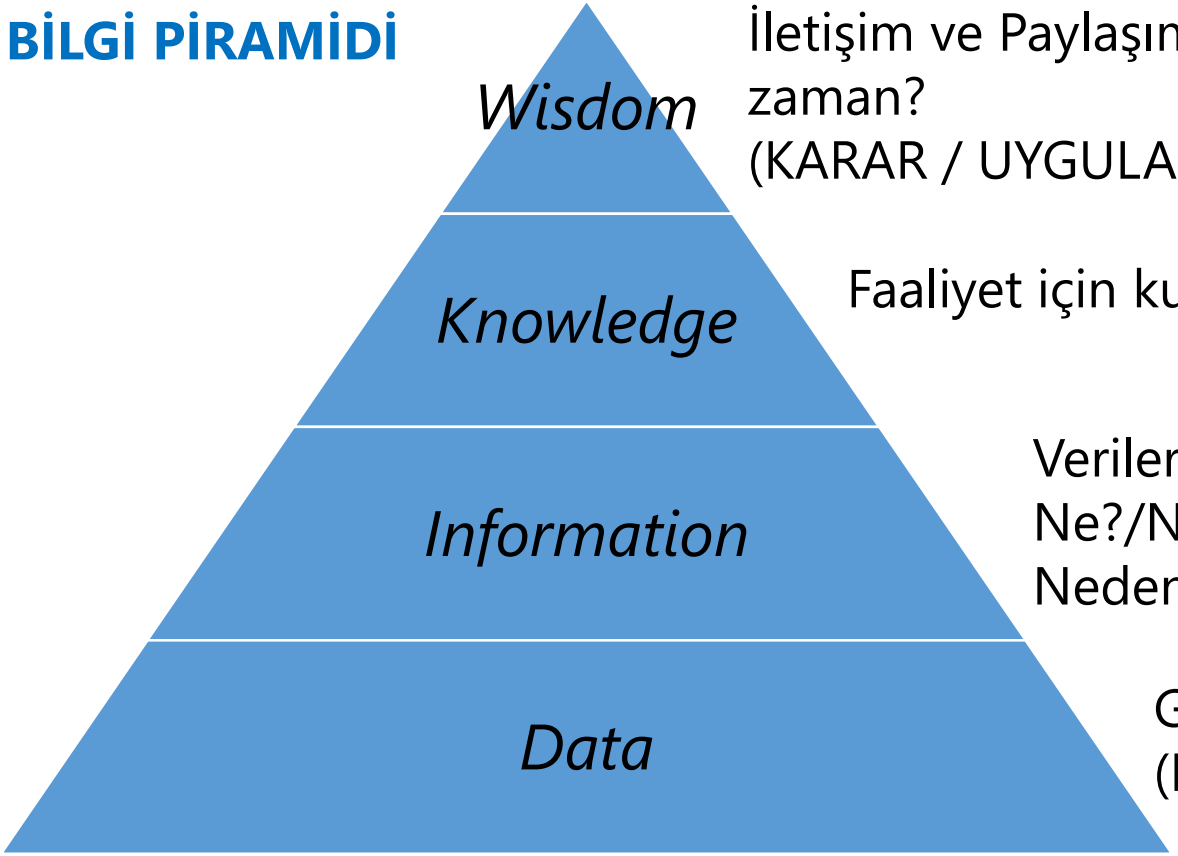
- Mevcut artış eğilimi devam ettiği takdirde 2040 yılı itibarı ile çocuk işçi sayısının tüm Dünya'da … sayısını aşacağı Uluslararası Çalışma Örgütü tarafından öngörülmektedir. (Öngörücü)

- Otoyollarda hız sınırının 130km/s'den 120km/s'e düşürülmesi halinde, trafik kazalarında %… oranında azalma olacağı Emniyet Genel Müdürlüğü tarafından öngörülmektedir. (Reçete Yazıcı)

- Mevcut artış eğilimi devam ettiği takdirde 2050 yılında Dünya nüfusunun … milyarı aşması Birleşmiş Milletler tarafından öngörülmektedir. (Öngörücü)

Bilgi Piramidi

BİLGİ PİRAMİDİ



İletişim ve Paylaşım, Farkındalık: Neden/Ne zaman?
(KARAR / UYGULAMA)

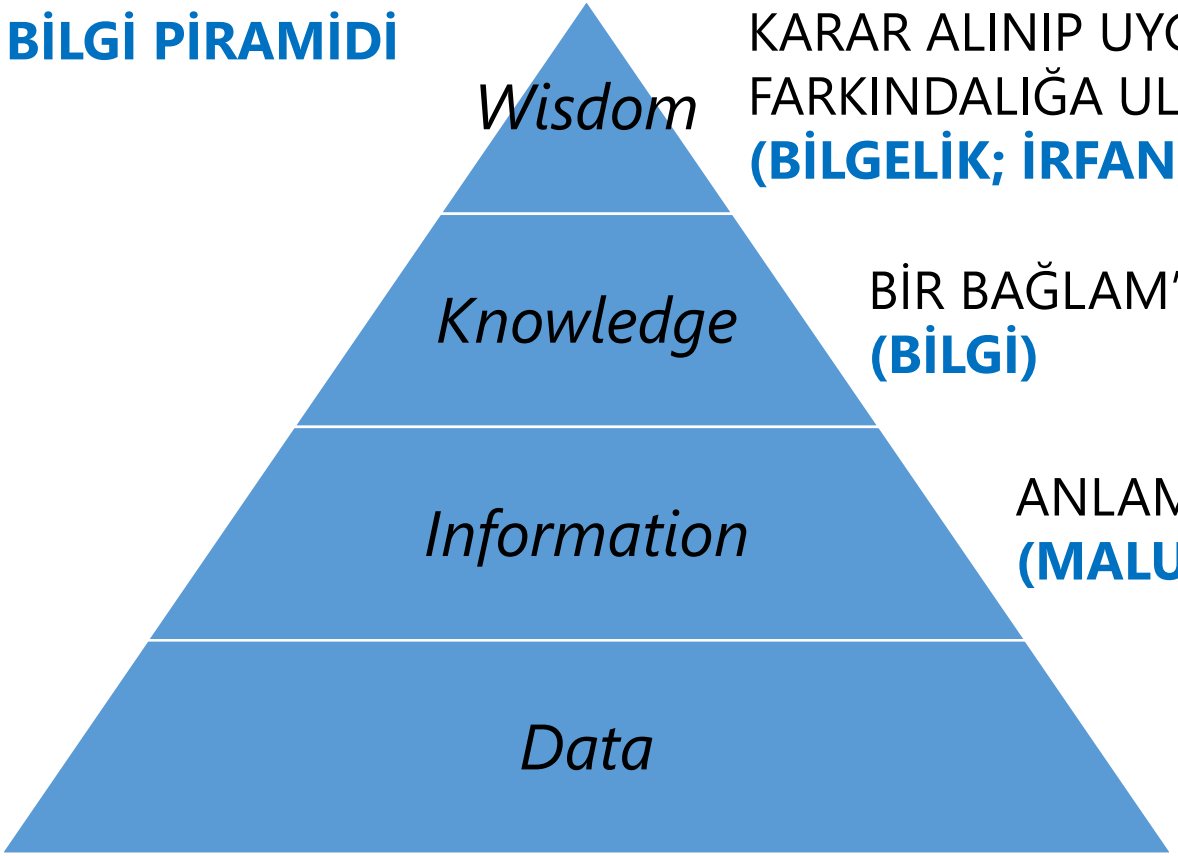
Faaliyet için kullanımı: Nasıl? (BAĞLAM)

Veriler arasındaki ilişkiler:
Ne?/Nerede?/Ne zaman?/
Neden?/Kim? (ANLAM)

Gözlemler: Ölçümler
(HAM)

Bilgi Piramidi

BİLGİ PİRAMİDİ



KARAR ALINIP UYGULAMAYA GEÇİLECEK FARKINDALIĞA ULAŞILMIŞ
(BİLGELİK; İRFAN; 'HİKMET')

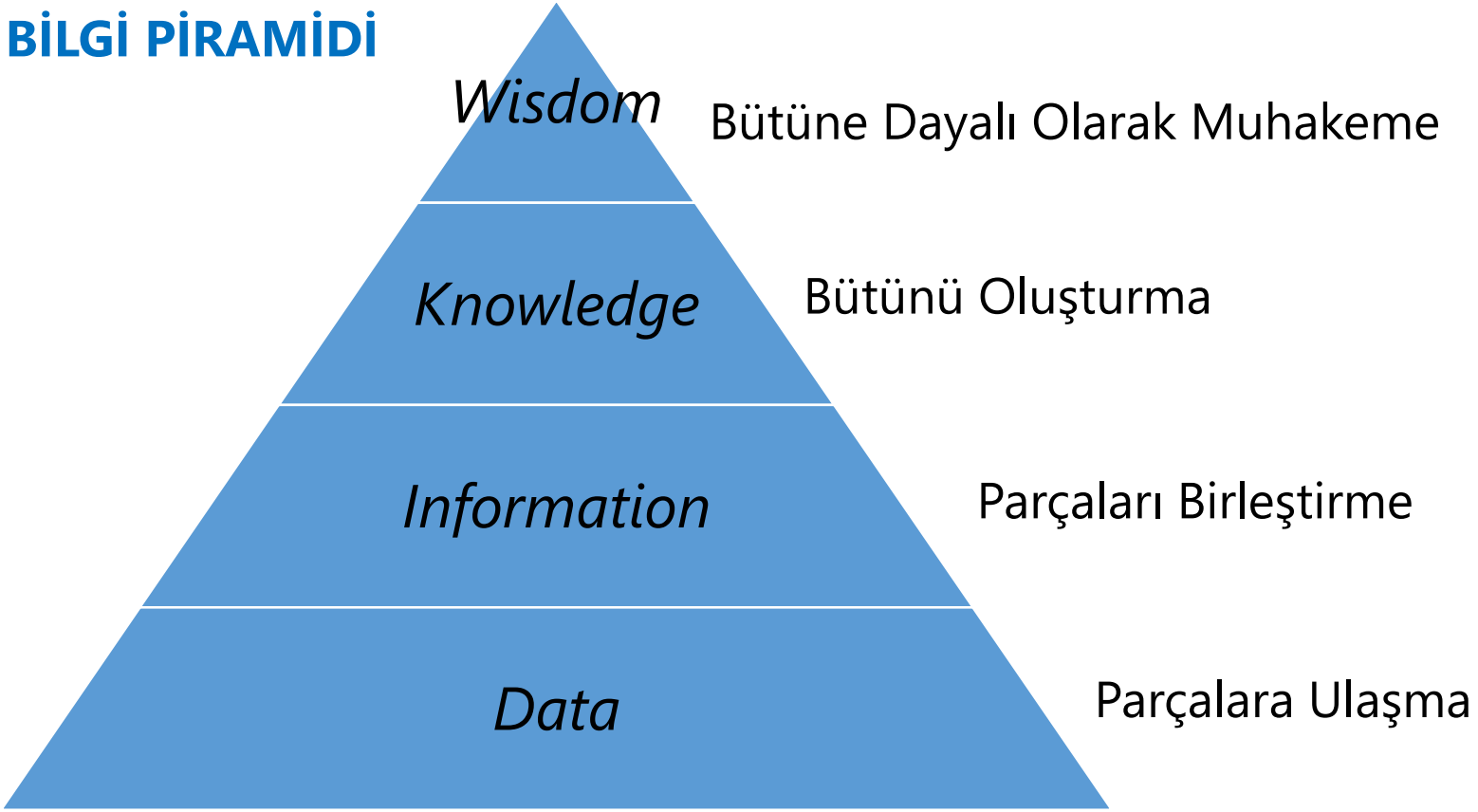
BİR BAĞLAM'A OTURTULMUŞ
(BİLGİ)

ANLAM KATILMIŞ
(MALUMAT?)

HAM
(VERİ)

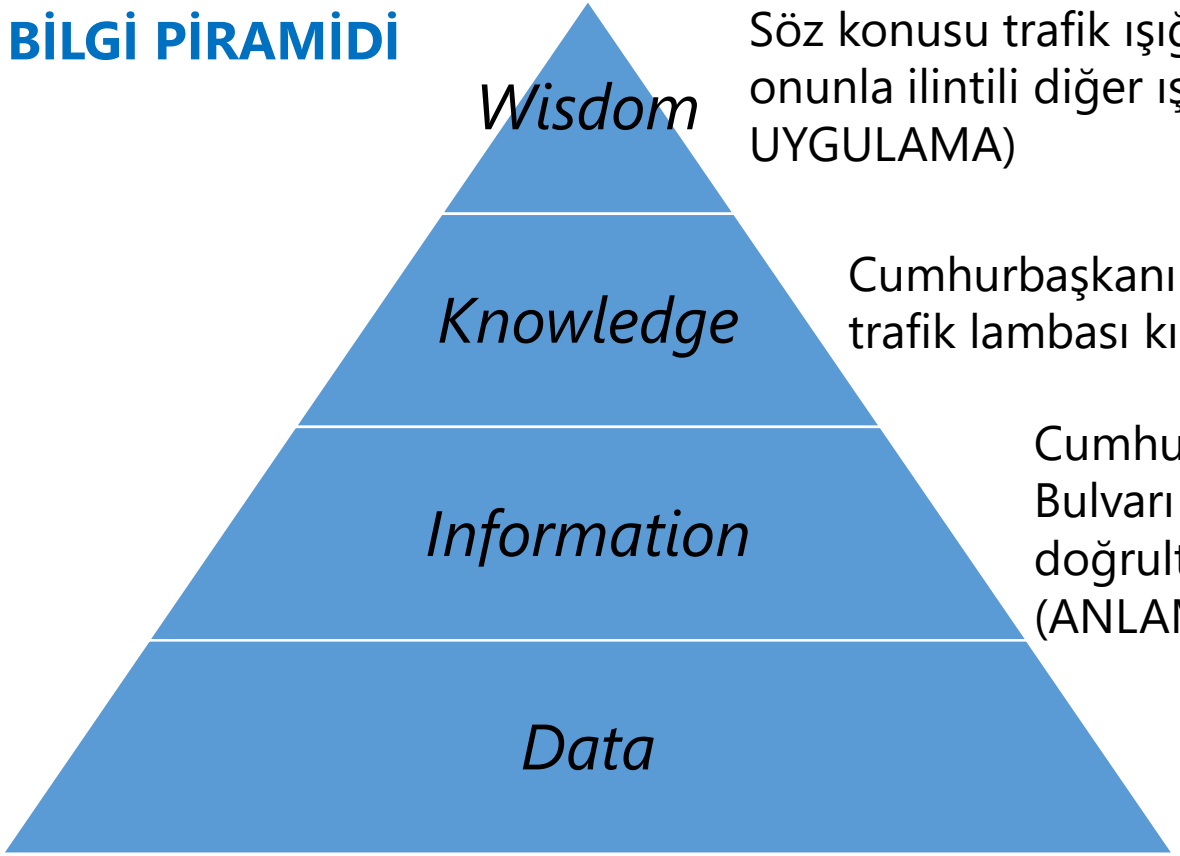
Bilgi Piramidi

BİLGİ PİRAMİDİ



Bilgi Piramidi

BİLGİ PİRAMİDİ



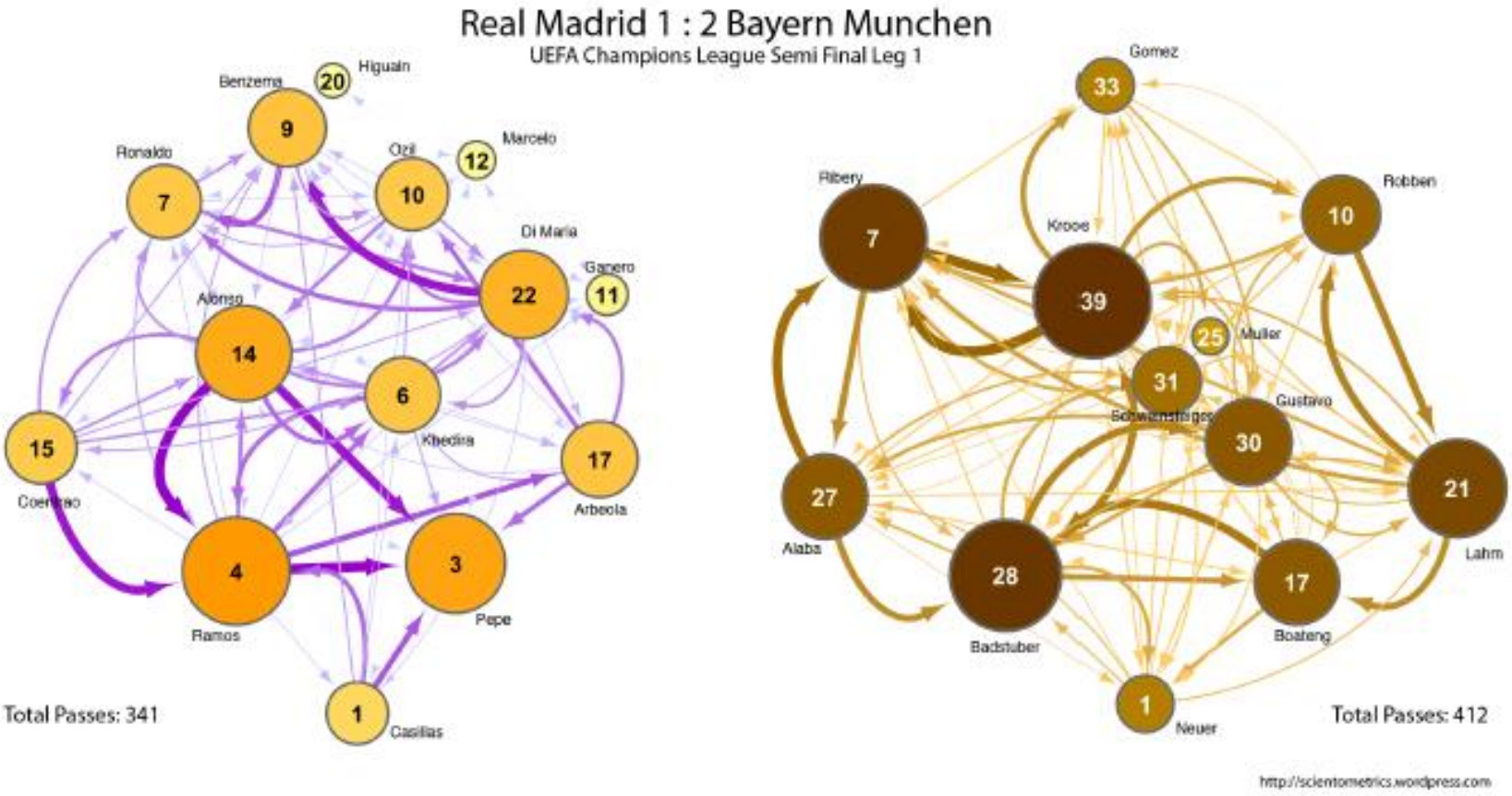
Söz konusu trafik ışığının durumunu derhal yeşile; onunla ilintili diğer ışıkları da kırmıza çevir (KARAR / UYGULAMA)

Cumhurbaşkanı Konvoyu'nun gittiği yöndeki trafik lambası kırmızı (BAĞLAM)

Cumhuriyet Caddesi ve Atatürk Bulvarı kesişimindeki güney-kuzey doğrultusundaki trafik lambası kırmızı (ANLAM)

252.144.122.355 IP Numaralı Trafik Lambası: Kırmızı (HAM)

Bilgi Piramidi – Bir Örnek

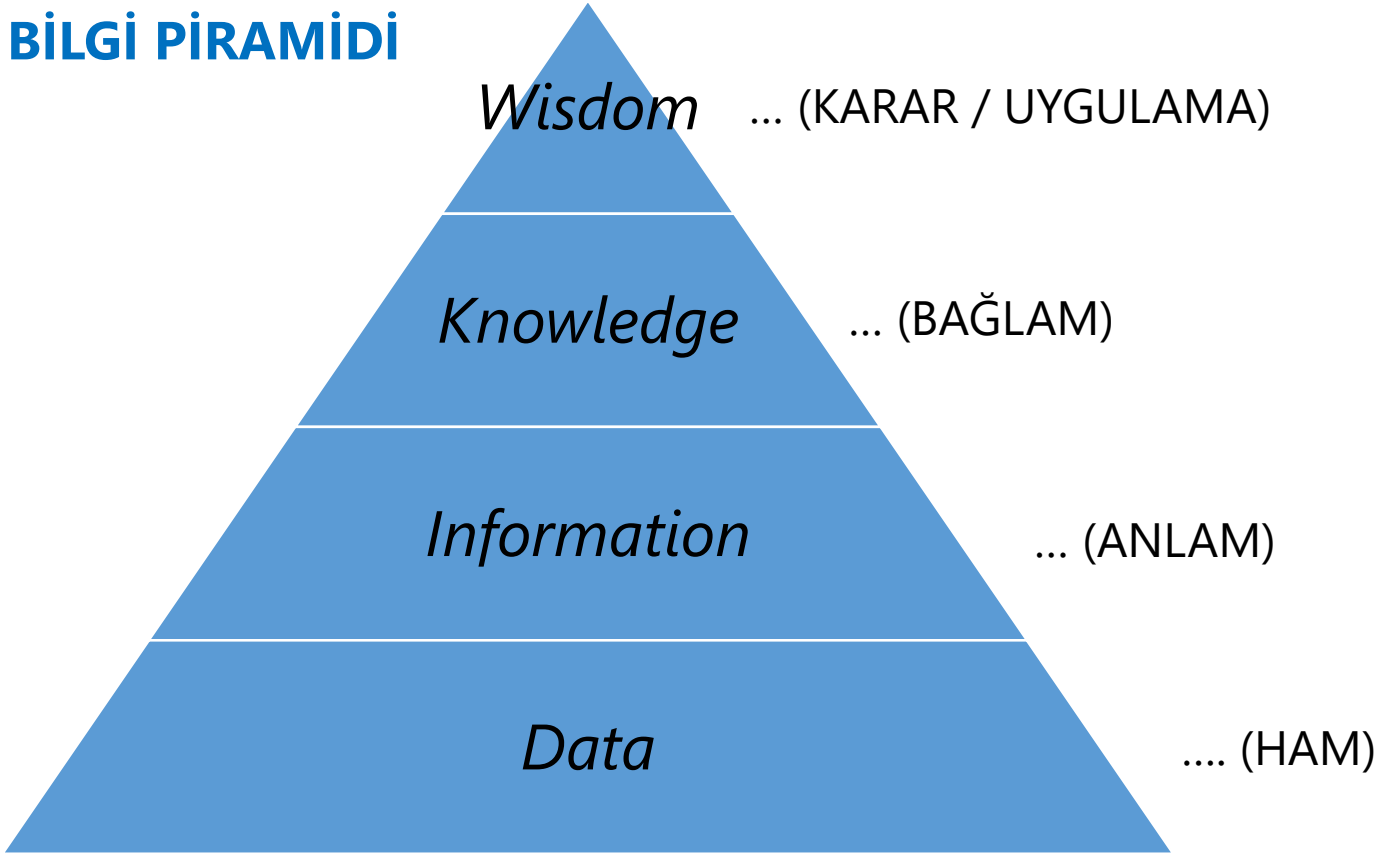


Real Madrid 1 : 2 Bayern Munchen Passing Distribution

https://scientometrics.wordpress.com/tag/social-network-analysis/

Bilgi Piramidi

BİLGİ PİRAMİDİ



Veride Nitelik (*Attribute*)

- Google E-Tablolar, Excel veya başka bir tabloda her bir sütuna karşılık gelen 'sıfat'

İsim	Cinsiyet	Doğum Yılı	Eğitim Düzeyi	Ağırlık (kg)	Boy (cm)	Aylık Gelir (TL Aralık)
…	E	1976	İlkokul	78.4	181	[2300, 4500]
…	K	1984	Yüksek Lisans	88.5	169	[2300, 4500]
…	K	1988	Lise	56.0	161	[4500, 7000]
…	E	1966	Lisans	78.8	174	[4500, 7000]
…	K	1961	Lisans	54.5	165	[7000, 10500]
…	E	1989	Ön Lisans	89.7	167	[4500, 7000]

- Nesne: Google E-Tablolar, Excel veya başka bir tabloda her bir satırda betimlenen varlık

Veride Nitelik (*Attribute*)

- Nitelik (*attribute*): özellik
- Öznitelik (*feature*): karakteristik, ön plana çıkan, ayırt edici özellik
- Zaman zaman nitelik, öznitelik ve boyut (*dimension*) kavramları birbirlerinin yerine kullanılmakta
- Nitelikler ve bunlara ait değerler bir nesneyi oluşturmakta / tanımlamakta

Veri Türleri

Niteliksel (Kalifatif: *Qualitative*) ya da Kategorik (*Categorical*) Veriler

- Sırasız (Saç Rengi, Göz Rengi, Kan Grubu, vb.)
- Sıralı (Eğitim Durumu, Akademik Unvan, vb.)

Niceliksel (Kantifatif: *Quantitative*) Veriler

- Sürekli (Yaş, Sıcaklık, vb.)
- Ayrık ya da Aralıklı (Çocuk Sayısı, Kaza Sayısı, Ceza Sayısı, vb.)

Veri (Ölçüm) Seviyeleri

Nominal Veriler (Nom: Name)

- İkilik (*Binary* veya *Boolean*) (Var/Yok, Kadın/Erkek, Hasta/Sağlıklı, vb.)
- İkiden Çok Kategorili (Medeni Durum, Irk, Şehir, İsim, vb.)

Ordinal Veriler (Ord: Order)

- Sıralı Kategorik Veriler (Eğitim Düzeyi, Rütbe, Akademik Unvan, Gelişmişlik Derecesi, Sosyoekonomik Ölçek Değeri, vb.)

Aralık (*Interval*) Tipi Veriler

- Sıralı sayısal veriler (Santigrat veya Fahrenheit Cinsinden Sıcaklık, Zaman, vb.)

Oransal/Rasyo (*Ratio*) Veriler

- Belirli bir referansa göre oranlı sayısal veriler (Mutlak Sıcaklık, Bağıl Nem, Desibel Cinsinden Sinyal Şiddeti, Richter Ölçeğine Göre Deprem Şiddeti, vb.)

Veri (Ölçüm) Seviyeleri

İşlenmemiş Değerler Kümesi

İsimlendirme Yoluyla Ayrıştırma

Nominal Veriler (Nom: Name)

Değerler Arasında Sıralama Bağıntısı

Ordinal Veriler (Ord: Order)

Değerleri Alt ve Üst Sınırlar Arasında Sıkıştırma; Değerler Arasında Mesafe Bağıntısı

Aralık (*Interval*) Tipi Veriler

Değer Kümesinde Mutlak Sıfır: «0»

Oransal/Rasyo (*Ratio*) Veriler

Veri (Ölçüm) Seviyeleri

Veri Tipi	Matematiksel İşlemler	Betimleyici İstatistikler	Çıkarımsal İstatistikler
Nominal	Sayma	Mod	Parametrik Olmayan, Chi Kare
Ordinal	Sıralama	Mod, Medyan, Aralık	Parametrik Olmayan, Chi Kare, Mann-Whiteney, Kruskall-Wallis, ANOVA
Aralık	Toplama, Çıkarma	Mod, Medyan, Ortalama, Aralık, Varyans	Parametrik Olmayan, Parametrik, t-testi, ANOVA
Oransal/ Rasyo	Toplama, Çıkarma, Çarpma, Bölme	Mod, Medyan, Ortalama, Aralık, Varyans	Parametrik Olmayan, Parametrik, t-testi, ANOVA

Verilere İlişkin Olası Birtakım Sıkıntılar

Hatalı / Kirli Veriler

- Kaydedilmemiş / Girilmemiş Veriler
  - meslek = ‘ ’
- Hatalı Girilmiş Veriler
  - maaş = ‘–10’
- Tutarsız Nitelik İsimleri
  - Bir yerde ‘ad’, başka yerde ‘isim’
  - Bir yerde ‘soyadı’, başka yerde ‘soyismi’
  - Bir yerde ‘malzeme referans no’, başka yerde ‘malzeme kayıt no’
- Tutarsız Nitelik Değerleri
  - Bir yerde ‘yaş=30-40 arası’, başka yerde ‘doğum tarihi = 01.02.1970’

Verilere İlişkin Olası Birtakım Sıkıntılar

Hatalı / Kirli Veriler

- Eksik veri kayıtlarının nedenleri
  - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
  - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
  - İnsan, yazılım ya da donanım problemleri
- Gürültülü (hatalı) veri kayıtlarının nedenleri
  - Hatalı veri toplama gereçleri
  - İnsan, yazılım ya da donanım problemleri
  - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
  - Verinin farklı veri kaynaklarında tutulması
  - İşlevsel bağımlılık kurallarına uyulmaması

Veriyi Tanıma ve Tanımlama

- Verinin merkezi eğilim özellikleri
- Verinin dağılım özellikleri
- Verinin sayısal nitelikleri ve sıralanabilir değerleri

- Merkezi eğilim, ortanca, en büyük, en küçük, varyans, sıklık derecesi (frekans), aykırılık, dağılım fonksiyonu, yoğunluk fonksiyonu, …

- Bol olasılık, istatistik ve stokastik bilgisi!...

Eksik Veriyi Tamamlama

- Eksik nitelik değerleri olan kayıtların kullanılmaması / atılması
- Eksik nitelik değerlerinin manuel olarak doldurulması
- Eksik nitelik değerleri için global bir değişken kullanılması (Null, Bilinmiyor, vb.)
- Eksik nitelik değerlerinin, o niteliğin ortalama değeri ile doldurulması
- Eksik nitelik değerlerinin, aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldurulması
- Eksik nitelik değerlerinin, olasılığı en fazla olan nitelik değeriyle doldurulması

Veri Düzeltme

Gürültülü Veri Nasıl Düzeltilir?

Bölütleme (*Segmentation*)

- Verinin sıralanması, eşit genişlik veya eşit derinlik ile bölünmesi

Kümeleme / Demetleme / Öbekleme (*Clustering*)

- Aykırılıkların belirlenmesi

Eğri Uydurma (*Curve Fitting*)

- Verinin bir fonksiyona uydurularak gürültünün düzeltilmesi

Kaynak: K. Ergün, 'Veri Madenciliği Ders Notları', Balıkesir Üniversitesi.

Veri Düzeltme

Gürültülü Veri Nasıl Düzeltilir?

Bölütleme Örneği

Veri: 8, 4, 21, 15, 21, 25, 24, 34, 28

Sıralı Veri: 4, 8, 15, 21, 21, 24, 25, 28, 34

Eşit Genişlik Yaklaşımı: Bölme sayısının belirlenmesi ve verinin eşit aralıklarla bölünmesi

Eşit Derinlik Yaklaşımı: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.

Her bölmenin, ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilmesi

Bölme Derinliği = 3

1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

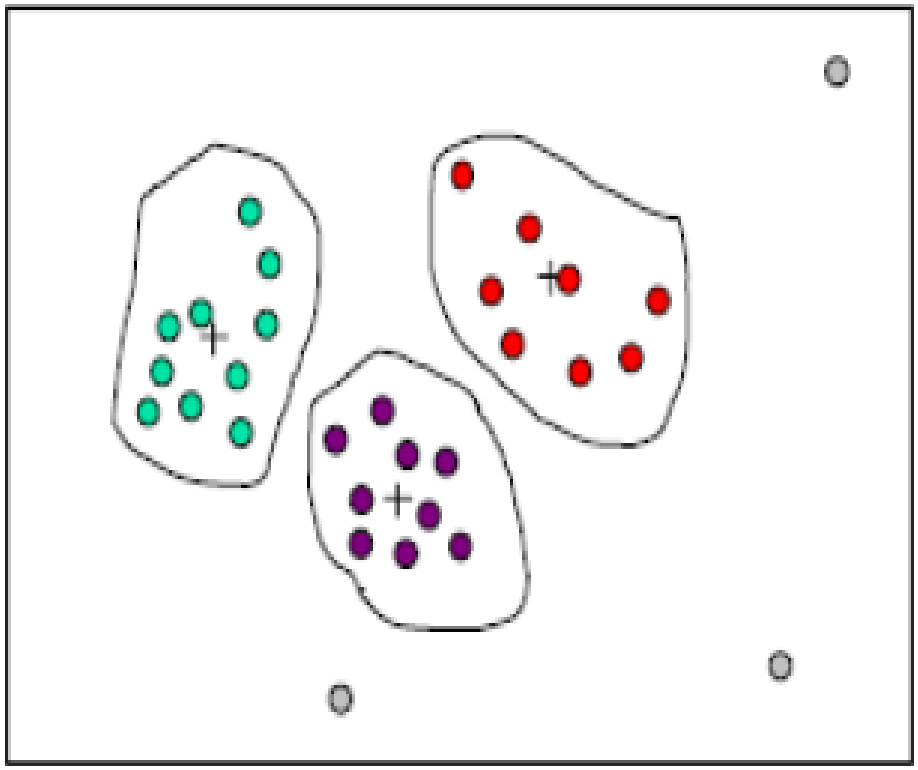
Ortalamayla düzeltme:
1. Bölme: 9, 9, 9
2. Bölme: 22, 22, 22
3. Bölme: 29, 29, 29

Alt-üst sınırla düzeltme:
1. Bölme: 4, 4, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 25, 34

Kaynak: K. Ergün, 'Veri Madenciliği Ders Notları', Balıkesir Üniversitesi.

Veri Düzeltme

Kümeleme / Demetleme / Öbekleme (*Clustering*)



Benzer verilerin aynı kümede/öbekte olacak şekilde gruplanması

Bu kümelerin/öbeklerin dışında kalan verilerin aykırılık olarak belirlenmesi ve silinmesi

Kaynak: K. Ergün, 'Veri Madenciliği Ders Notları', Balıkesir Üniversitesi.

Veri Düzeltme

Normalizasyon (*Normalization*)

Verinin, uygun ve belirli bir aralık arasında kalacak şekilde dönüştürülmesi

- min-max normalizasyon

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalizasyon

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j}$$ j: Max($|v'|$)<1 olacak şekildeki en küçük tam sayı

Kaynak: K. Ergün, 'Veri Madenciliği Ders Notları', Balıkesir Üniversitesi.

Veri Düzeltme

Nitelik Oluşturma

Mevcut nitelik/özniteliklerden, daha anlamlı nitelik/öznitelik oluşturulması

Örneğin: ‘en’ ve ‘boy’ niteliklerinden ‘alan’ niteliğinin oluşturulması

Kaynak: K. Ergün, ‘Veri Madenciliği Ders Notları’, Balıkesir Üniversitesi.

Veri Biliminin İlk Adımı

Veri Okuryazarlığı

- P21 Skills altındaki en önemli ve gerekli yeteneklerden biri
- Her türlü eğitim düzeyi, her yaş ve her meslek grubu için elzem
- Veri-Malumat-Bilgi arasındaki farklar hakkında bilinç
- Veri türleri hakkında temel düzey farkındalık
- 'Temel' düzey istatistik bilgisi

Veri Biliminin İlk Adımı

Biri 'İstatistik' mi dedi?

Ne kadarını bilmemiz gerek?

Endişeye Mahal Yok ☺

Sadece Temel Düzeyde:

- Merkezi Eğilim Ölçütleri
- Dağılım Ölçütleri

yeterli

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR? BU KADAR ÇOK KAFA KARIŞTIRICI TANIMA GEREK VAR MI?

A ve B isimli iki öğrencimiz olsun

A'nın genel başarı ortalaması 100 üzerinden 79

B'nin genel başarı ortalaması 100 üzerinden 53

'Hangi öğrenci daha başarılı?' sorusuna cevap vermek kolay

Ancak …

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

20'şer öğrencilik iki adet sınıfımız olsun. Aynı ders ve aynı sınavdan öğrencilerin aldığı notlar şu şekilde olsun:

SINIF 1 Notlar (Küçükten Büyüğe Sıralı):
25, 27, 31, 34, 37, 41, 41, 43, 56, 59, 59, 60, 61, 64, 66, 66, 70, 71, 75, 81

SINIF 2 Notlar (Küçükten Büyüğe Sıralı):
22, 22, 23, 34, 38, 40, 41, 43, 51, 56, 59, 61, 61, 64, 65, 66, 72, 72, 79, 88

Soru ve Sorun: Hangi sınıf daha başarılı? Hangi ölçüt(ler)e göre?

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Merkezi Eğilim Ölçütleri:
Bir dağılımın nerede yoğunlaştığına (merkezinin ne civarda olduğuna) cevap bulmaya yönelik ölçütler

Ortalama veya Beklenen Değer (*Mean, Expected Value,* μ):
Dağılımdaki (kümedeki veya sayı dizisindeki) tüm değerlerin aritmetik ortalaması

Kümenin (sayı dizisinin) elemanı olmak zorunda değil

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Merkezi Eğilim Ölçütleri:
Bir dağılımın nerede yoğunlaştığına (merkezinin ne civarda olduğuna) cevap bulmaya yönelik ölçütler

Tepe Değeri veya Mod (*Mode*):
Dağılımda (kümede veya sayı dizisinde) en çok görülen değer(ler)

Kümenin (sayı dizisinin) elemanı

Tek bir değer olmak zorunda değil; birden fazla değer olabilir

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Merkezi Eğilim Ölçütleri:
Bir dağılımın nerede yoğunlaştığına (merkezinin ne civarda olduğuna) cevap bulmaya yönelik ölçütler

Ortanca veya Medyan (*Median*):
Dağılım (küme veya sayı dizisi) küçükten büyüğe (veya büyükten küçüğe) sıralandığında ortada kalan değer

Kümenin (sayı dizisinin) toplam eleman sayısı tekse; tam ortada kalan değer olacağı için kümenin (sayı dizisinin) elemanı

Kümenin (sayı dizisinin) toplam eleman sayısı çiftse; tam ortada kalan bir değer olmayacağı için kümenin (sayı dizisinin) elemanı olmayabilir

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Bir Örnek

- Elimizde bir sayı dizisi bulunsun

(örn. Öğrencilerin aldığı notlar ☺):

13, 18, 13, 14, 13, 16, 14, 21, 13

- Bunları küçükten büyüğe sıralı dizersek:

13, 13, 13, 13, 14, 14, 16, 18, 21

Ortalama veya Beklenen Değer

(*Mean, Expected Value*) = 15 ← Kümenin (sayı dizisinin) elemanı bile değil

Mod (*Mode*) = 13

Medyan veya Ortanca (*Median*) = 14

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Bir Örnek Daha

- Elimizde bir sayı dizisi bulunsun

(örn. Öğrencilerin aldığı notlar ☺):

13, 18, 13, 15, 13, 16, 14, 21, 13, 17

- Bunları küçükten büyüğe sıralı dizersek:

13, 13, 13, 13, 14, 15, 16, 17, 18, 21

Ortalama veya Beklenen Değer (*Mean, Expected Value*) = 15.3 ← Kümenin (sayı dizisinin) elemanı bile değil

Mod (*Mode*) = 13

Medyan veya Ortanca (*Median*) = 14.5 ← Kümenin (sayı dizisinin) elemanı bile değil

Sayısal Verilere İlişkin Birtakım Ölçütler

İSTATİSTİK NEDEN BU KADAR ZOR?

Bir Örnek Daha (Hatalı Veri Girişi)

- Bir önceki örnekte bir öğrencinin notu yanlış girilmiş olsun (örn. Öğrencilerin aldığı notlar ☺):

13, 18, 13, 15, 13, 16, 14, 21, 13, 178

- Bunları küçükten büyüğe sıralı dizersek:

13, 13, 13, 13, 14, 15, 16, 18, 21, 178

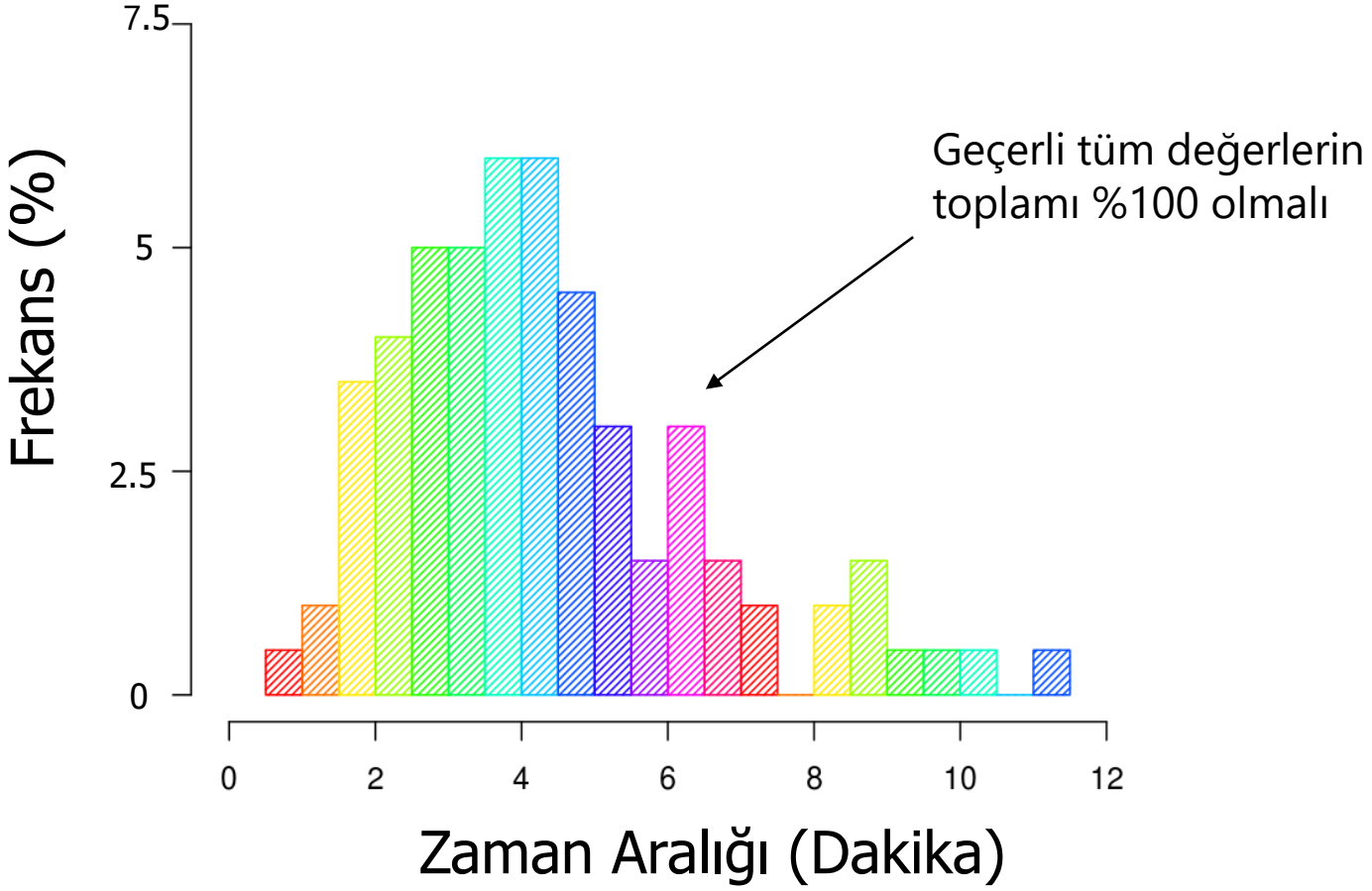
Ortalama veya Beklenen Değer
(*Mean, Expected Value*) = 31.4 ← 15.3'tü; anormal bir artış gösterdi

Mod (*Mode*) = 13 ← Değişmedi!...

Medyan veya Ortanca (*Median*) = 14.5 ← Hiç değişmedi!...

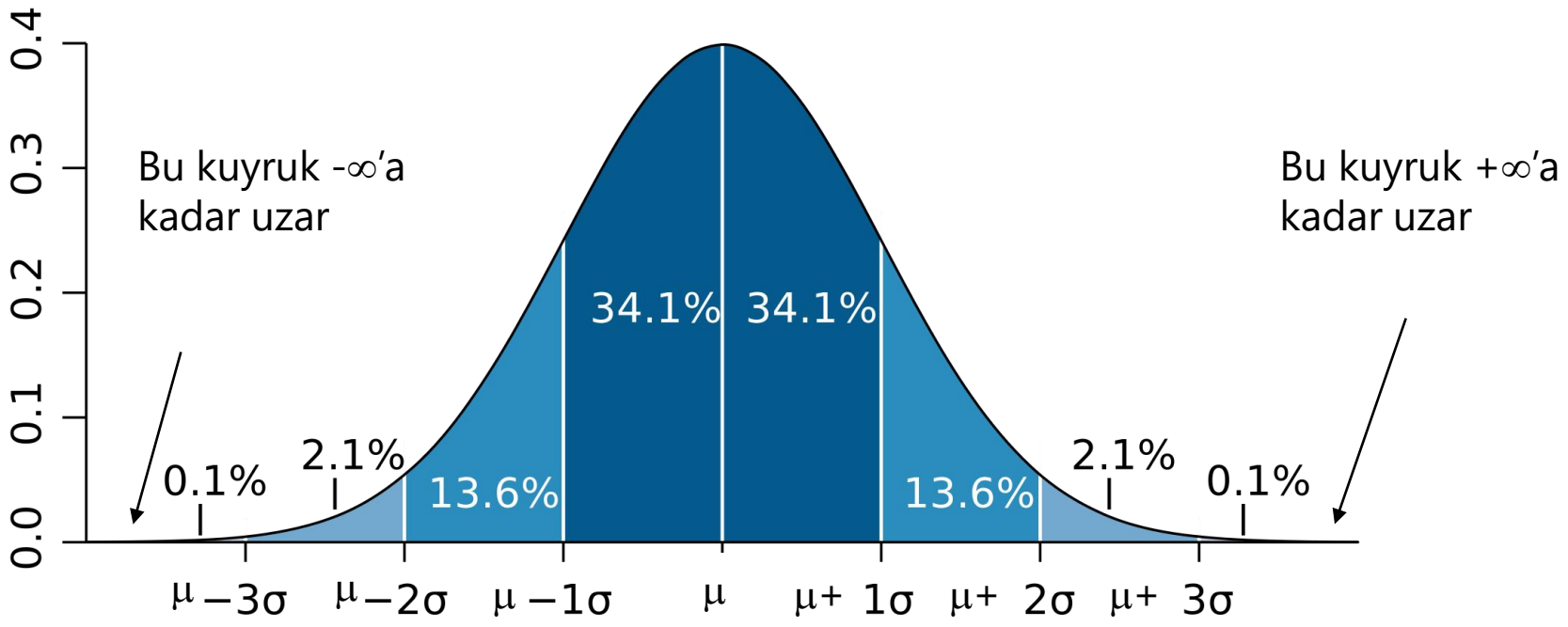
Histogram

Bir terminale giren araçların Histogram'ı



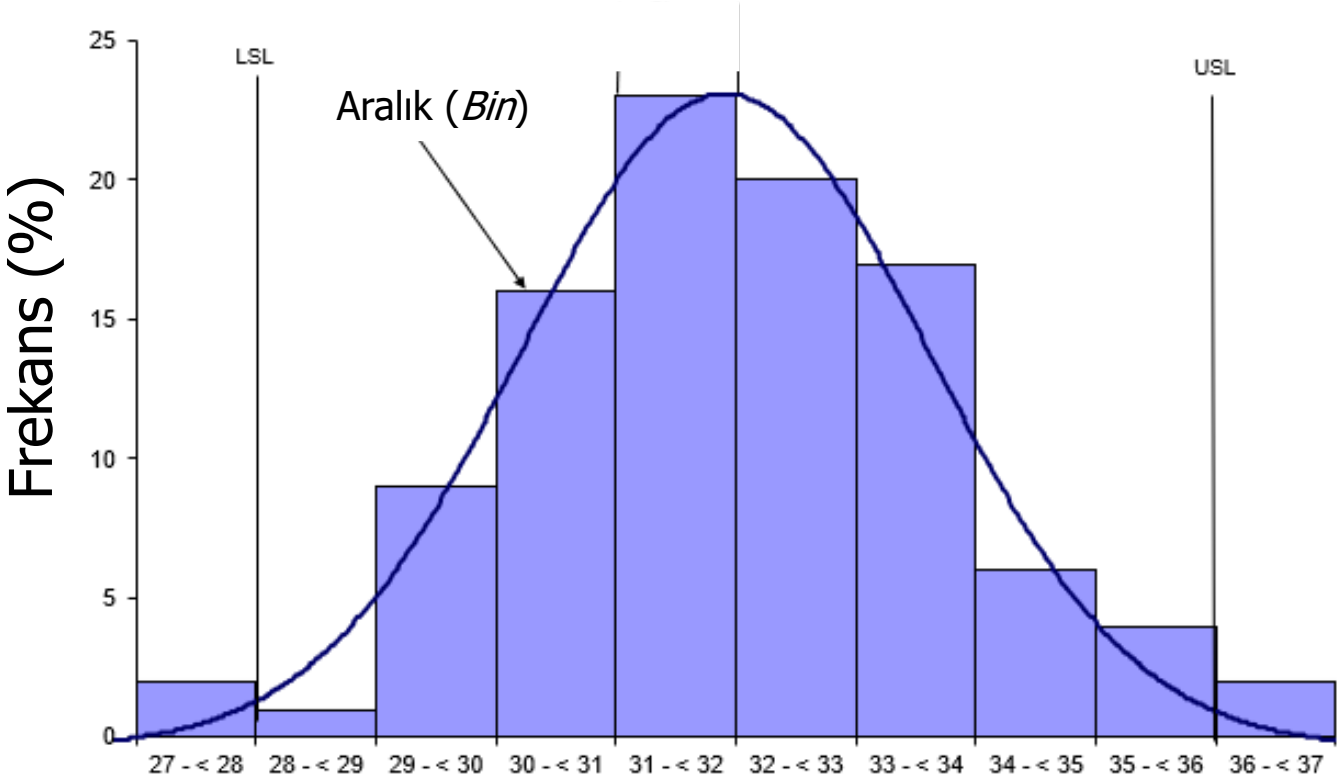
Gauss Dağılımı (Normal Dağılım)

- Meşhur Çan Eğrisi
- Düzgün Dağılım (*Uniform Distribution*) ile karıştırılmamalı



Histogram – Gauss Dağılımı İlişkisi

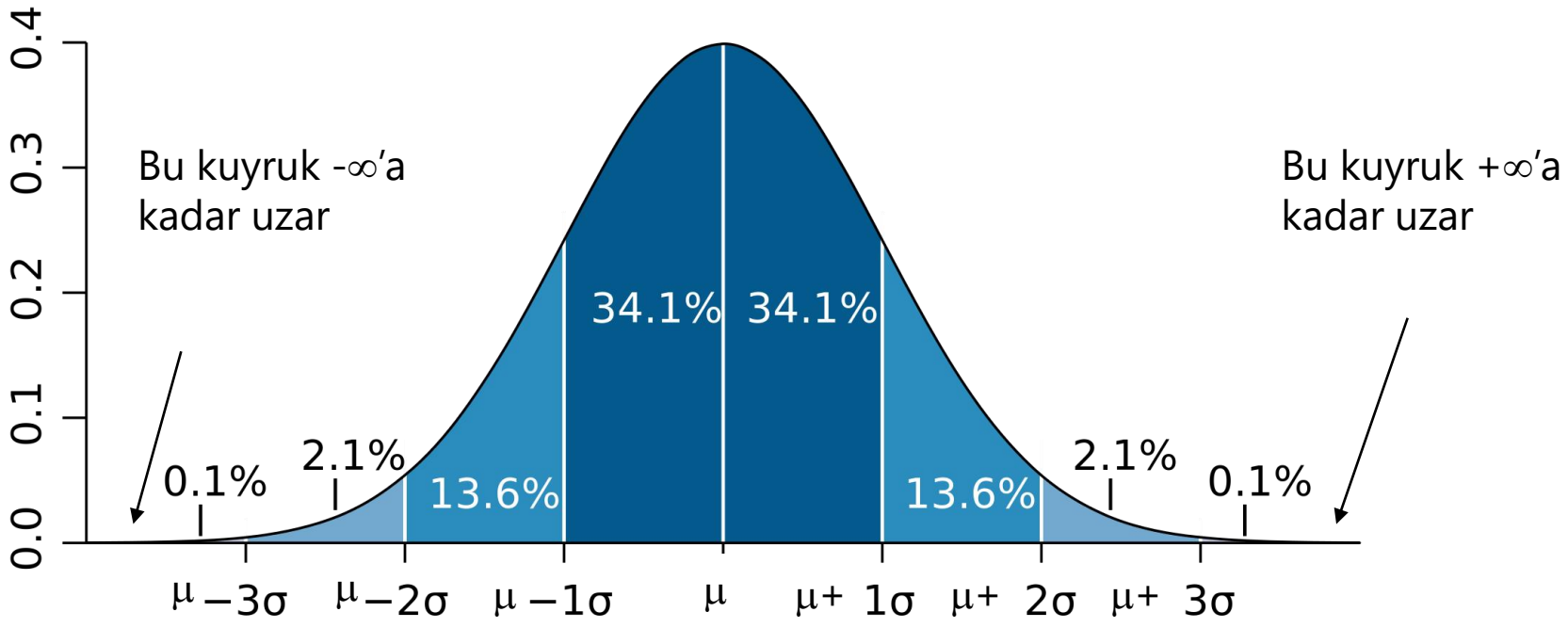
- Çok sayıda örnek varsa; aralık (*bin*) değerleri de çok küçülürse Histogram, Gauss Dağılımına dönüşmeye başlar!...



Dağılımlara İlişkin Ölçütler

Dağılım Ölçütleri:
Bir dağılımın merkezin etrafında ne kadar yoğunlaştığına (veya merkezin yakınında/uzağında ne oranda değer aldığına) cevap bulmaya yönelik ölçütler

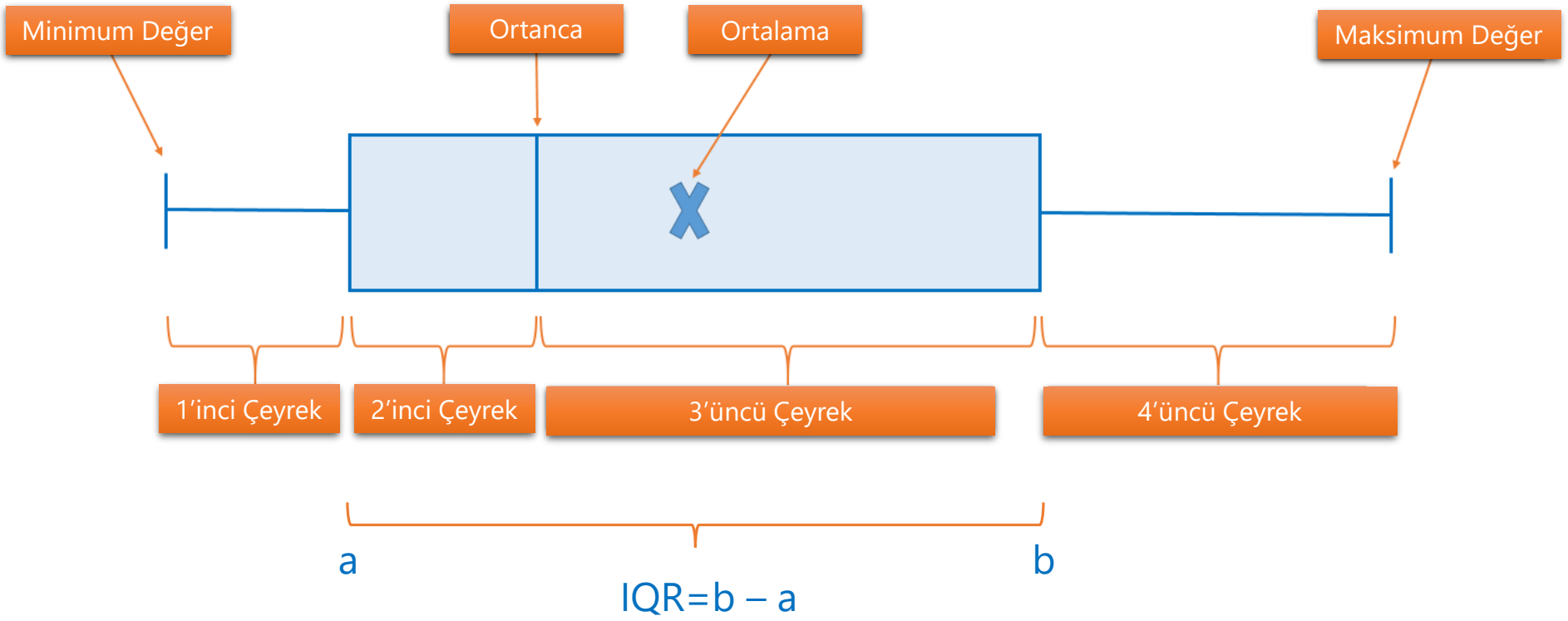
Standart Sapma (*Standard Deviation*, σ) ve Varyans (*Variance*, σ^2) :



Sayısal Verilere İlişkin Birtakım Ölçütler

Dağılım Ölçütleri:

Dörttebirlik veya Çeyreklik; IQR gibi ölçütlerin 'Kutu' veya Kutu ve Bıyık' Grafiği (*Box Plot; Box and Whisker Plot*) ile gösterimi (daha sade ve modern)



Dağılımlara İlişkin Ölçütler

Pratikte tam bir Gauss Eğrisi elde edilemez; 2,800,000 kişinin girdiği üniversite sınavında bile!... Niye?

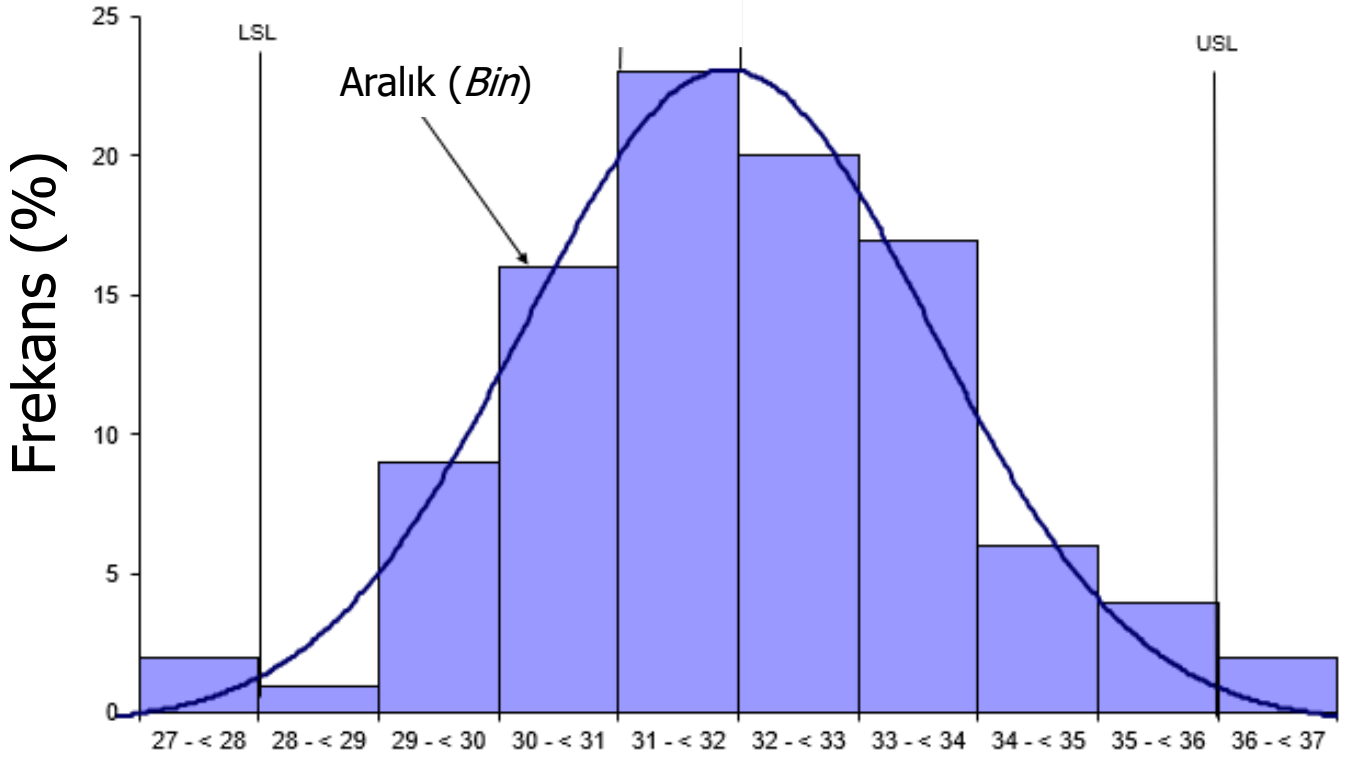
- Öğrencilerin aldığı puanlar (ya da yaptığı netler) $+\infty$ veya $-\infty$'a gitmez; sınırlıdır.
- Histogram çizilirken aralık (*bin*) değeri ne kadar küçük bir sayı da olsa tam 0'a çekilemez
- 2,800,000 sayısı ne kadar büyük olsa da sonludur!...

Yani:

- Histogramı çizilen değerlerin (puan veya net) $+\infty$ veya $-\infty$'a gitmemesi
- Histogram çizimindeki aralık (*bin*) değerinin ne kadar küçültülse de tam 0'a çekilememesi
- Örnek sayısının ne kadar büyük olsa da sonlu olması

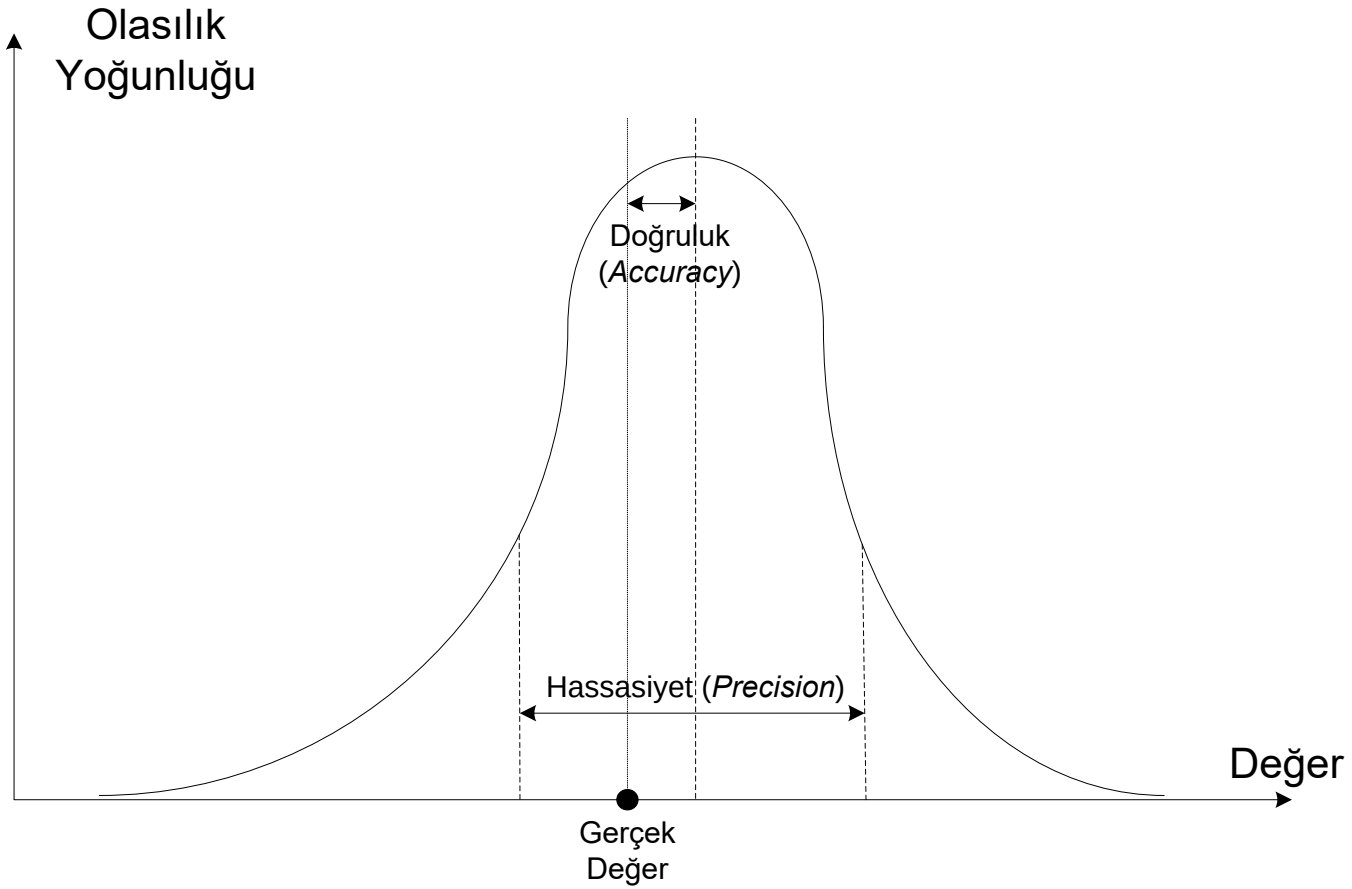
Dağılımlara İlişkin Ölçütler

Dolayısıyla bir histogramın ideal Gauss'a ne kadar benzediğine dair ölçütlere de ihtiyaç var: Çarpıklık (*Skewness*) ve Basıklık (*Kurtosis*)



Dağılımlara İlişkin Ölçütler

Ayrıca bir sürecin/işlemin başarımını nesnel olarak değerlendirmeye yönelik ölçütlere de ihtiyaç var: Doğruluk (*Accuracy*) ve Keskinlik (*Precision*)



Dağılımlara İlişkin Ölçütler

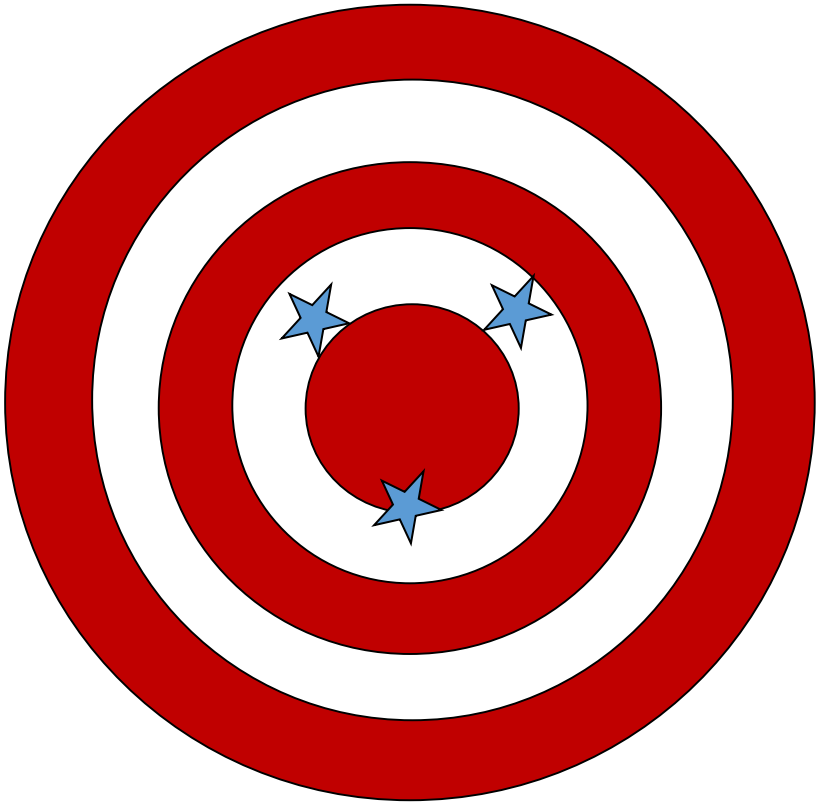
Doğruluk (*Accuracy*) ve Keskinlik (*Precision*) Kavramlarına İlişkin 4 Okçu Örneği

Accurate ve
Precise okçu

Dağılımlara İlişkin Ölçütler

Doğruluk (*Accuracy*) ve Keskinlik (*Precision*) Kavramlarına İlişkin 4 Okçu Örneği

Accurate ama
Imprecise okçu



Dağılımlara İlişkin Ölçütler

Doğruluk (*Accuracy*) ve Keskinlik (*Precision*) Kavramlarına İlişkin 4 Okçu Örneği

Inaccurate ama
Precise okçu

Dağılımlara İlişkin Ölçütler

Doğruluk (*Accuracy*) ve Keskinlik (*Precision*) Kavramlarına İlişkin 4 Okçu Örneği

Inaccurate ve
Imprecise okçu

Genel Tanımlar

Değişken:
Bir deney / gözlem esnasında değişim gösteren her nitelik/nicelik

Değişken Tipleri:
- Bağımsız
- Bağımlı
- Kontrol Altında Tutulan

Genel Tanımlar

Bağımsız Değişken:
Bilim insanı veya araştırmacı tarafından değiştirilen nitelik/nicelik

- Genellikle bir anda (bir deney veya bir gözlem esnasında) sadece bir adet
- Neden? '*ceteris paribus*'

Bağımlı Değişken:
Bilim insanı veya araştırmacı tarafından bağımsız değişken değiştirilirken; olası değişimleri gözlemlenen nitelik(ler)/nicelik(ler)

Genel Tanımlar

Kontrol Altında Tutulan Değişken:
Tüm deneyler/gözlemler esnasında; bilim insanı veya araştırmacı tarafından sabit tutulmaya çalışılan nitelik(ler)/nicelik(ler)

Örnek:

Bir süs bitkisinin, amiyane tabirle 'güneşi sevip sevmediğini' belirlemeye yönelik bir araştırma yapıyor olalım.

- Bağımsız değişkenimiz?
- Bağımlı değişkenlerimiz?
- Kontrol altında tutulan değişkenlerimiz?

Genel Tanımlar

Popülasyon (*Population*)
Hedeflenen ve üzerinde çalışılan ana kitle

Örneklem (*Sample*)
Popülasyondan çekilen alt küme

Genel Tanımlar

Tümevarım (*Induction*); Tümevarımsal Yaklaşım (*Inductive Approach*) ve Tümevarımsal Gerekçelendirme (*Inductive Reasoning*):

- Bir alt küme üzerinde gerçekleştirilen gözleme ilişkin bulguların, tüm küme için genellenmesi
  - Örneklemde Q oranında bireyin A özelliği bulunmaktadır.
  - Dolayısıyla:
    - Tüm popülasyonda Q oranında bireyin A özelliği bulunmaktadır.

Genel Tanımlar

Tümevarım (*Induction*); Tümevarımsal Yaklaşım (*Inductive Approach*) ve Tümevarımsal Gerekçelendirme (*Inductive Reasoning*):

İstatistiksel Tasım (*Statistical Syllogism*)

- İstatistiksel bilgilerin tek bir bireye dair çıkarımda kullanılması
  - Örnek: Fen Lisesi mezunlarının %90'ı üniversite sınavında ilk seferde bir bölüme yerleşiyor. Ahmet de Fen Lisesi son sınıfta. Demek ki Ahmet bu sene bir yere yerleşecek.

Genel Tanımlar

Tümdengelim (*Deduction*); Tümdengelimsel Yaklaşım (*Deductive Approach*) ve Tümdengelimsel Gerekçelendirme (*Deductive Reasoning*):

- Genel popülasyonun özelliklerine dair gözlemlere ilişkin bulgulardan yola çıkılarak bu özelliklerin bir alt kümeye ve/veya bir elemana atfedilmesi

  Örnek:
  - İfade 1: Tüm insanlar ölümlüdür.
  - İfade 2: Socrates bir insandır.
  - Çıkarım: Dolayısıyla, Socrates de ölümlüdür.

Genel Tanımlar

Tümdengelim (*Deduction*); Tümdengelimsel Yaklaşım (*Deductive Approach*) ve Tümdengelimsel Gerekçelendirme (*Deductive Reasoning*):

- *Modus ponens* (Öncülün doğrulanması): Birinci seviye kural çıkarımı

  Örnek:
  - İfade 1: Bir açı 90° ile 180° arasında ise geniş açıdır.
  - İfade 2: A açısı 120° 'dir.
  - Çıkarım: Dolayısıyla, A açısı geniş açıdır.

  Lise mantık derslerinden ufak bir hatırlatma:

  $P \Rightarrow Q$

Genel Tanımlar

Tümdengelim (*Deduction*); Tümdengelimsel Yaklaşım (*Deductive Approach*) ve Tümdengelimsel Gerekçelendirme (*Deductive Reasoning*):

- *Modus tollens* (Karşıdoğru yasası): Olmayana ergi
  Örnek:
  - Yağmur yağıyorsa havada bulutlar olmalıdır.
  - Şu anda havada bulut bulunmamaktadır.
  - Dolayısıyla şu anda yağmur yağıyor olamaz!...

  Lise mantık derslerinden ufak bir hatırlatma:
  $P \Rightarrow Q$
  $(\neg Q) \Rightarrow (\neg P)$

Genel Tanımlar

Karşılaştırmalar

Tümdengelim (*Deduction*)	Tümevarım (*Induction*)
Yukarıdan aşağıya mantık (*Top-down logic*)	Aşağıdan yukarıya mantık (*Bottom-up logic*)
Yapılan çıkarımlar kesinlikle doğru	Yapılan çıkarımlar olasılıksal olarak doğru

Peki Ya Şu Meşhur '*p*' Değeri

İstatistiksel Anlamlılık veya Önem (*Statistical Significance*):

- Tümevarımsal Yaklaşım'da Dış Geçerliğin ne derecede sağlandığına dair bir ölçüt: *p* değeri
- *p* değeri ne kadar küçükse, o kadar iyi

- Bir parada 20 yazı/tura atışı sonucunda 14 defa tura gelmiş!... Para için hileli iddiasında bulunabilir miyiz?
- Peki 15 defa gelseydi?

- Ortalama 70g fındık içerdiği iddia edilen bir gofretten 40 adet satın alıp hesap yaptığımızda; fındık ortalamasını 68.4g bulduk
- Firmayı sahtekarlıkla suçlayabilir miyiz?

Bilimsel Çalışmalarda İstatistiksel Değerlendirme

Diyelim ki çalışmamız esnasında kurduğumuz hipotezin yanlış mı; doğru mu olduğunu değerlendireceğiz.

- Örnek 1: Kapıyı çalmadan girmek kötü niyet belirtisi olduğundan; kapıyı çalmadan giren kişi hırsızdır.
- Örnek 2: Ailesinin aylık geliri 2000TL'den az olan, ebeveyninden en az birini kaybetmiş bir ergen suça meyilli olur.
- Örnek 3: Hastada X, Y, Z vital bulgularının olması/görülmesi, H sendromuna işaret eder.

Özetle: Tespit (*Detection*), Anomali Çıkarımı, İkili Sınıflandırma (*Binary Classification*) yaptığımız tüm çalışmalarımızda

"Karar Verici" Girdi-Çıktı İlişkileri

Durumsallık (*Contingency*) Matrisi

	Olay Var	**Olay Yok**
Algı Var	Doğru Pozitif (DP) *detection* veya *hit*	Yanlış Pozitif (YP) *false alarm* veya *Type I error*
Algı Yok	Yanlış Negatif (YN) *missed detection*, *miss* veya *Type II error*	Doğru Negatif (DN) *correct rejection*