

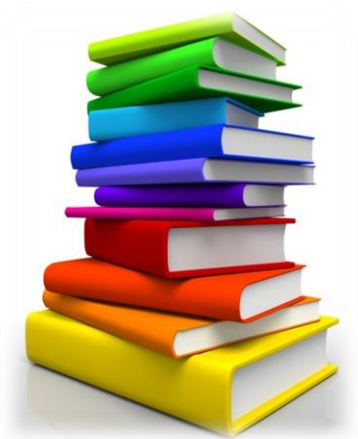
BİL 210

Veri Bilimi ve Makine Öğrenmesine Giriş

Doç. Dr. Deniz KILINÇ
İzmir Bakırçay Üniversitesi
Bilgisayar Mühendisliği

BÖLÜM - 1

Veri Dünyasına ve Kavramlara Genel Bakış



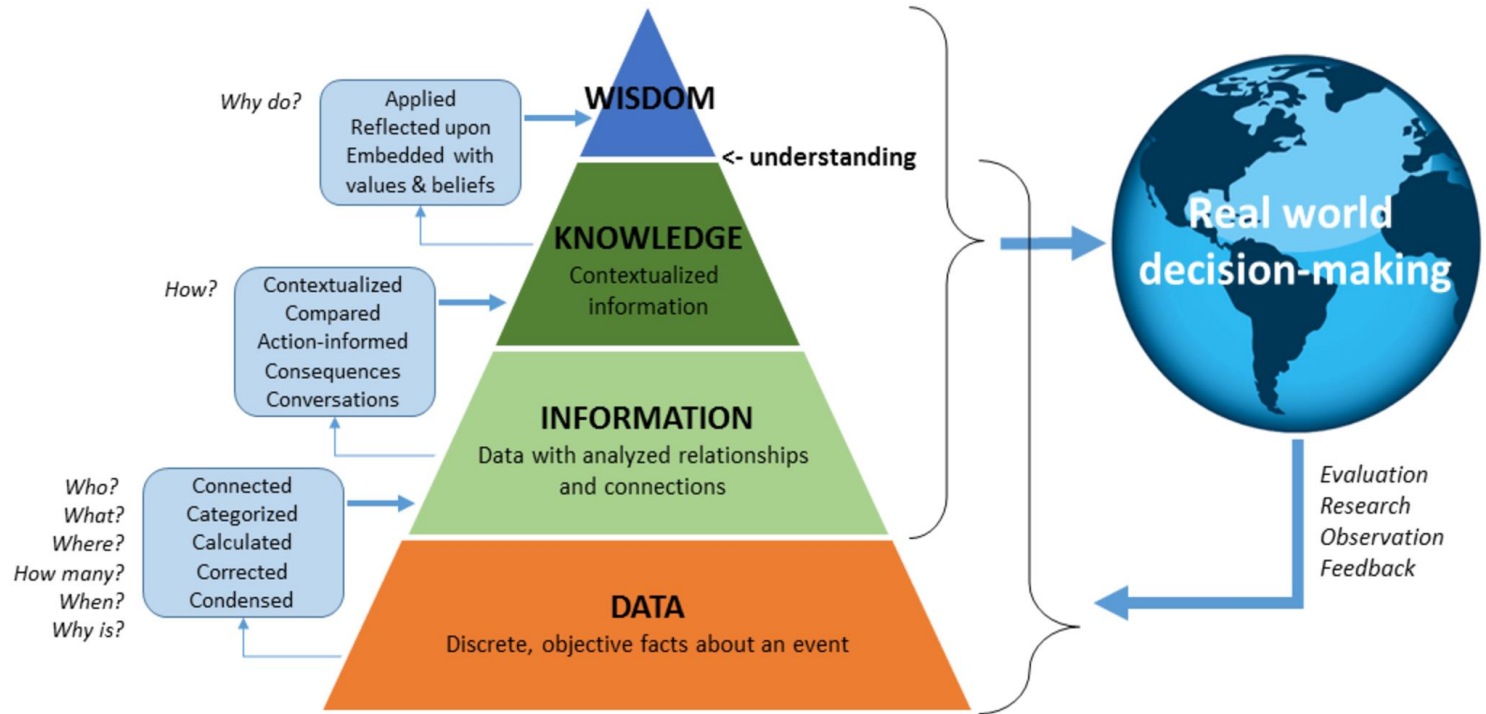
Bu bölümde;

- Veri, Enformasyon, Bilgi, Bilgelik (Data, Information, Knowledge, Wisdom)
- Veri Bilimi Nedir?
- Yapay Zeka Nedir?
- Makine Öğrenmesi Nedir?
- Makine Öğrenmesi İş Akışı
- Veri Bilimi ve Python

ile ilgili konular anlatılacaktır.

Veri, Enformasyon, Bilgi, Bilgelik

- İngilizcesi → Data, Information, Knowledge, Wisdom



Veri, Enformasyon, Bilgi, Bilgelik (devam...)

- **Veri:** Gözlem, araştırma, deney, ölçüm ve sayım gibi değişik yöntemlerle elde edilen ve herhangi başka birşeyle ilişkisi kurulmamış, işlenmemiş bilgi parçacığdır. **Kendi başlarına bir anlam ifade etmezler**. Sayılar, rakamlar, sözcükler, metinler, resimler, olaylar vb. biçiminde temsil edilen ham gerçekliklerdir.
 - Sayısal veriler: Rakamlar (0-9), harfler (a-z, A-Z)
 - Görüntüsel veriler: Grafikler, resimler, foto
 - Ses verileri: Her türlü ses, melodi ve müzik
 - Video verileri: Hareketli görüntüler.

Veri, Enformasyon, Bilgi, Bilgelik (devam...)

- **Enformasyon:** Verinin belli bir formülle düzenlenerek *anlamlı* hâle dönüştürülmesidir. Başka bir deyişle veriler; özetleme, hesaplama, sınıflandırma, gruplandırma ve analizler aracılığıyla enformasyona dönüştürülmektedir. Yani veriler, **düzenli ve anlamlı ilişkiler çerçevesinde bir araya getirildiğinde** önem kazanarak enformasyona dönüşürler.
 - Enformasyon, elde edilmiş veriler ve varsa başka diğer enformasyonların arasındaki ilişkilerin ortaya çıkarılmasıyla oluşur. **NE, NEREDE, NE ZAMAN, KİM, KAÇ TANE** gibi soruların yanıtıdır.

Veri, Enformasyon, Bilgi, Bilgelik (devam...)

- **Bilgi:** Enformasyonun rasyonel bir biçimde akıl süzgecinden geçmesi, yorumlanması ve kullanımıyla ortaya çıkar.
 - Bilgi, bir veya daha fazla **enformasyonun nasıl olduğu öğrenildiğinde** sahip olunur, enformasyonların ortaya çıkardığı **örüntünün yorumlanmasıdır**
 - **NASIL** sorusunun yanıtıdır.

Veri, Enformasyon, Bilgi, Bilgelik (devam...)

- **Bilgelik:** Elde edilen bilgilerin *nedenini*, o şekilde oluşmalarının arkasındaki ilkelerin öğrenilmesidir,
- **NEDEN sorusunun** yanıtıdır. Ayrıca diğer adımlar hep olan biten, gerçekleşmiş olgularla uğraşırken, bilgelik seviyesi bize **GELECEKTE OLABİLECEK olayları bilebilmemizi**, bir vizyon, bir gelecek öngörüsü geliştirebilmemizi, buna göre bir tasarım, bir plan yapabilmemizi olanaklı kılar.
 - Başka bir bakış açısıyla, *değişen şartlar çerçevesinde ileriye görebilme yeteneğine sahip olmaktır.*

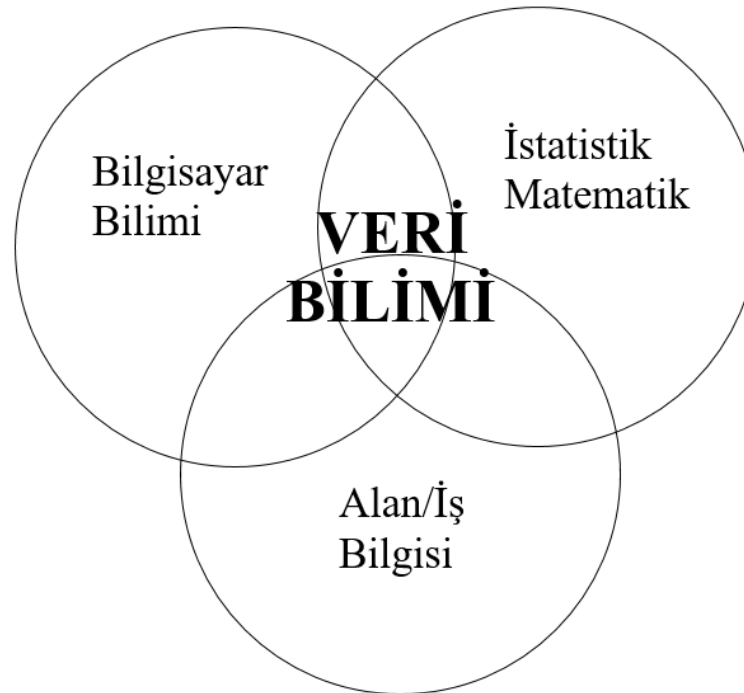
Veri, Enformasyon, Bilgi, Bilgelik (devam...)

- **Örnek:** Bir öğrenci veri tabanındaki ad, soy ad, doğum yeri, öğrencilerin her birinin
 1. Algoritma ve Programlama dersinden aldığı notlar, öğrencilerin bölüme giriş puanları, mezun oldukları lise türleri vb. gerçeklikler veri;
 2. Öğrenci listesi, notlar, harf notları, sınıf ortalamasından oluşan liste enformasyon;
 3. Öğrencilerin mezun olduğu lise ya da üniversiteye giriş puanına göre başarı durumlarının analiz edilerek bir örüntü çıkarılması bilgi;
 4. Bu örüntüleri çıkarabilme yetisi, programlama bilgisi ve analiz ve yorumlama yeteneği ise bilgelik olarak düşünülebilir.

Veri Bilimi Nedir?

- **Veri bilimi**, i) bilimsel problem tanımlama ve çözme yöntemlerinin, ii) matematiğin, iii) istatistiğin, iv) yazılım geliştirmenin ve v) teknolojinin birleşiminden oluşan *çok disiplinli* bir bilim dalıdır.
- **Veri bilimi**, karmaşık problemleri çözmek için hem yapılandırılmış hem de yapılandırılmamış veriyi (data), işe yarar/anlamlı/değerli bilgiye (knowledge) dönüştürmeye yarar (Wiki, 2018).
- Donanım altyapılarının ve teknolojinin gelişimi sayesinde daha önceleri işlenemeyecek büyüklükte olan veriler işlenebilir hale gelmiş ve yoğun kaynak gerektiren algoritmaların bu veriler ile çalışabilmesi sağlanmıştır.
- Data Scientist: The Sexiest Job of the 21st Century (HBR)

Veri Bilimi Nedir? (devam...)



Yapay Zeka Nedir?

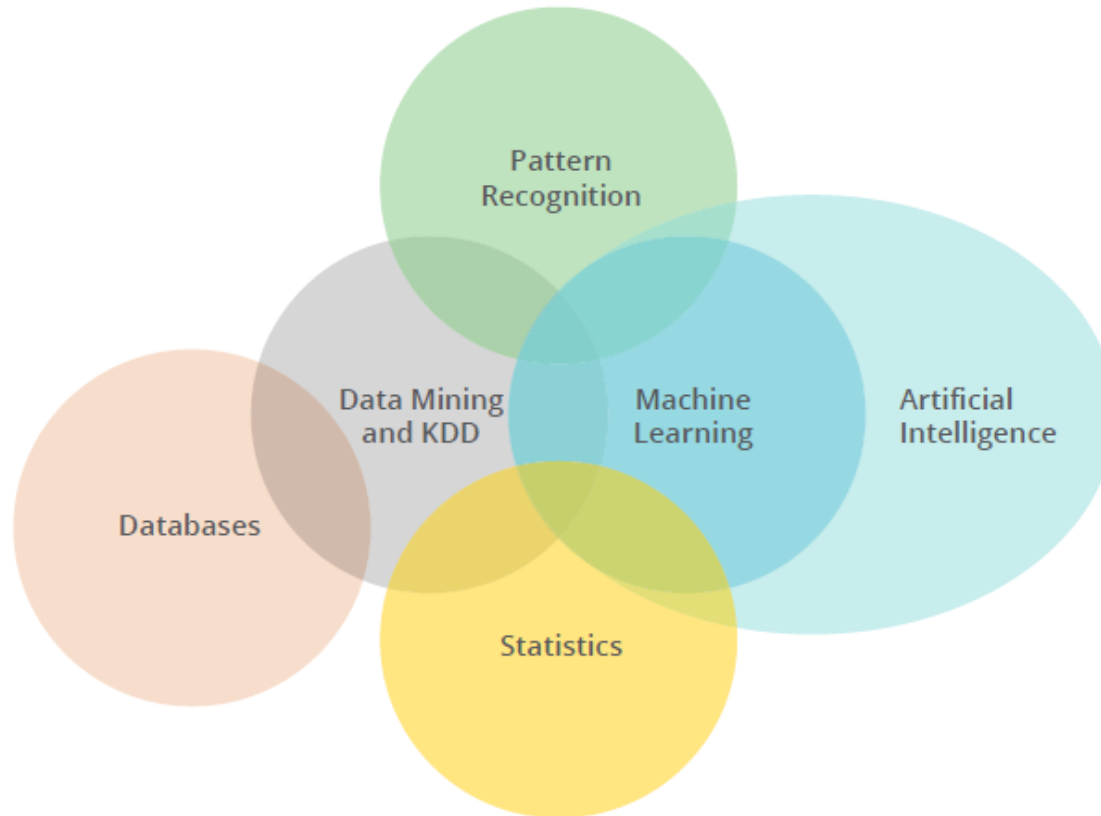
- **Zekâ;** İnsanın, karşılaştığı bir olay ve durumu algılayabilme ve buna karşı çözüm üretme yeteneği olarak tanımlanmıştır. Zekâ, zaman içinde, eğitim, öğrenme ve çevre etkenleri ile gelişmektedir.
- **Yapay Zekâ:** İnsan zekasına özgü olan, algılama, öğrenme, çoğul kavramları bağlama, düşünme, fikir yürütme, sorun çözme, iletişim kurma, çıkarım yapma ve karar verme gibi yüksek bilişsel fonksiyonları veya otonom davranışları sergilemesi beklenen zeki etmenler/programlar yaratmak üzere yapılan tüm çalışmalarının toplandığı geniş bir alandır.
- **Günümüzde** yapay zekâ araştırmaları; makine öğrenmesinin bir alt dalı olan, derin yapay sinir ağları ya da diğer adıyla derin öğrenmeye odaklanmış durumdadır.

Yapay Zeka Nedir? (devam...)

- **Yapay Zekâ Avantajları**

- Yapay zekâ paylaşılabılır: Zekâ, insanda eğitim, öğrenme ve çevre etkenleri ile gelişmektedir. Bu birikimin bir başka insana aktarılması, usta çırak ilişkisi ile olabilmektedir ve aktarım uzun süre almaktadır.
- Yapay zekâ daha kolay elde edilebilir: Bir bilgisayarın zekâ düzeyinin yükseltilmesi, bir insanın zekâ düzeyinin yükseltilmesine göre daha kolaydır; kolay olduğu için maliyeti de düşüktür.
- Yapay zekâ tutarlıdır: Aynı olay karşısında verilecek tepki insandan insana değişik olacaktır. Hatta benzer iki olayda, aynı insan farklı davranabilmektedir.

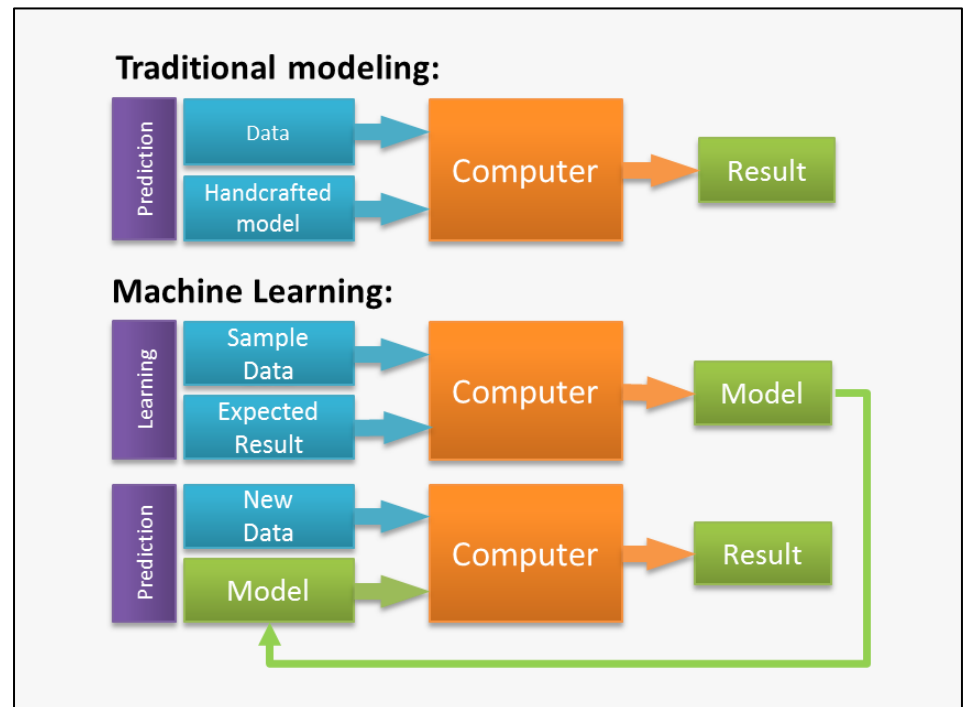
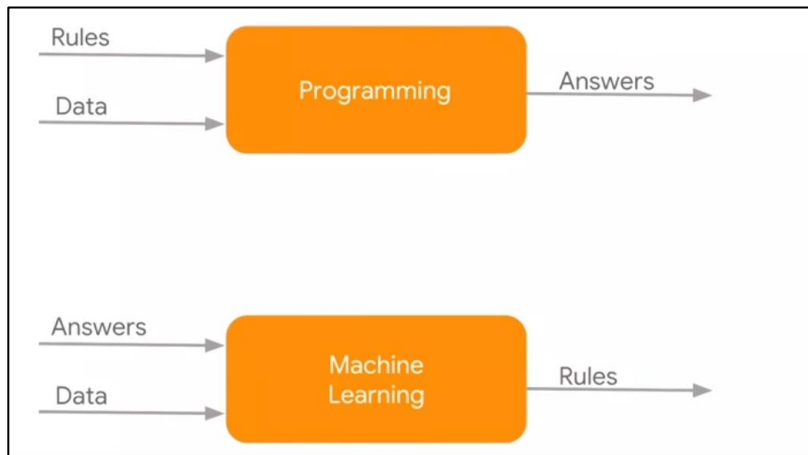
Yapay Zeka Nedir? (devam...)



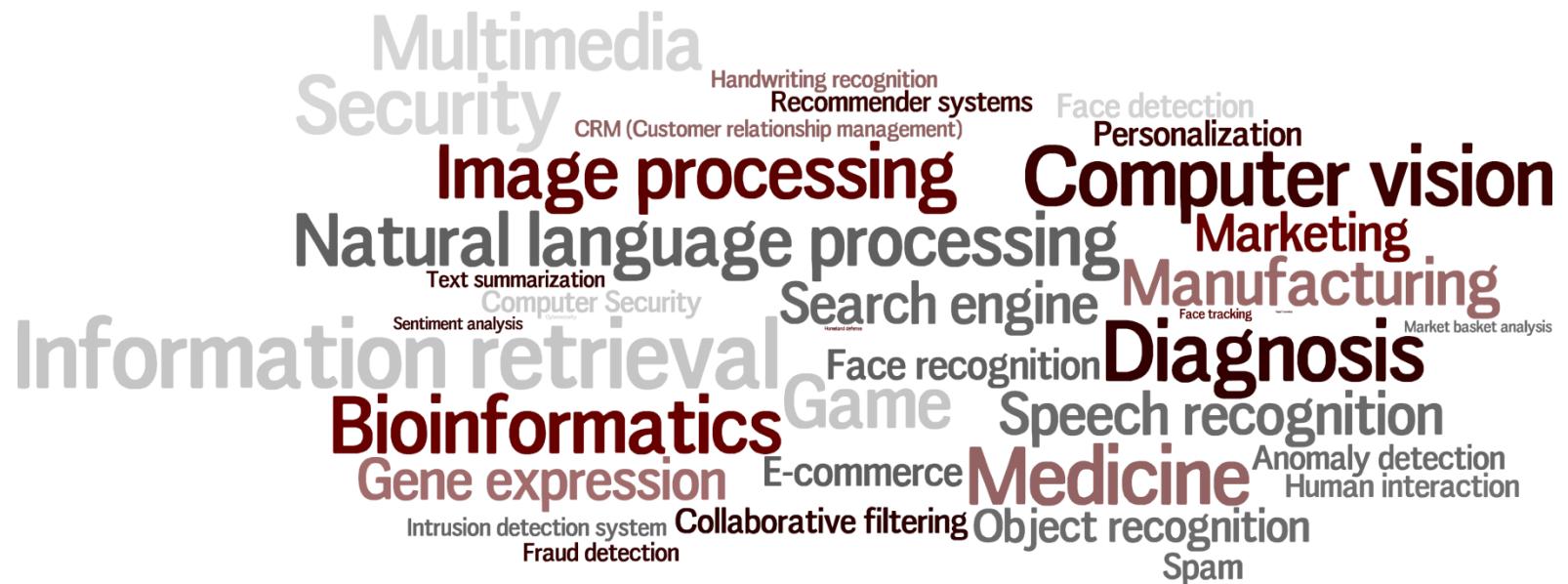
Makine Öğrenmesi Nedir?

- 1959 yılında bilgisayar biliminin **yapay zekada sayısal öğrenme ve model tanıma** çalışmalarından geliştirilmiş bir alt dalıdır.
- Makine öğrenmesi (Machine Learning) ayrı bir alan olarak **1990 yıllarında** yeniden gelişmeye başladı.
- **Amaç;** geçmişteki verileri (bir kısmını) kullanarak bir modeli oluşturmak/eğitmek ve yine bu modeli kullanarak gelecek ile ilgili tahminlerde bulunmaktır. Bu süreçte de en uygun model belirli ölçütleri değerlendirerek (Accuracy, ROC AUC, F-measure) bulunmaya çalışır. Sürekli denemeler yapılır.

Makine Öğrenmesi vs. Programlama



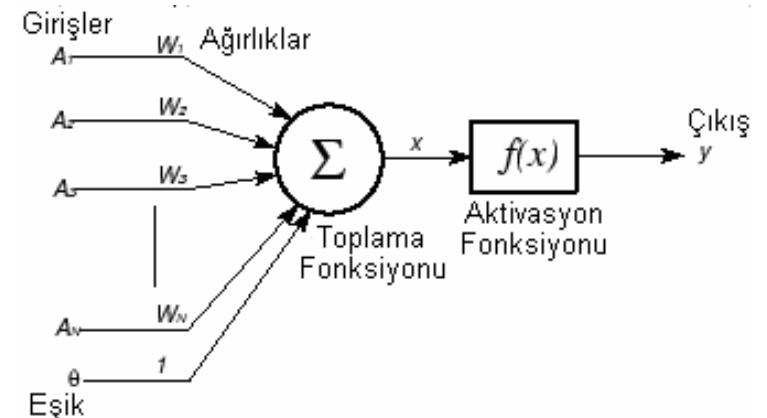
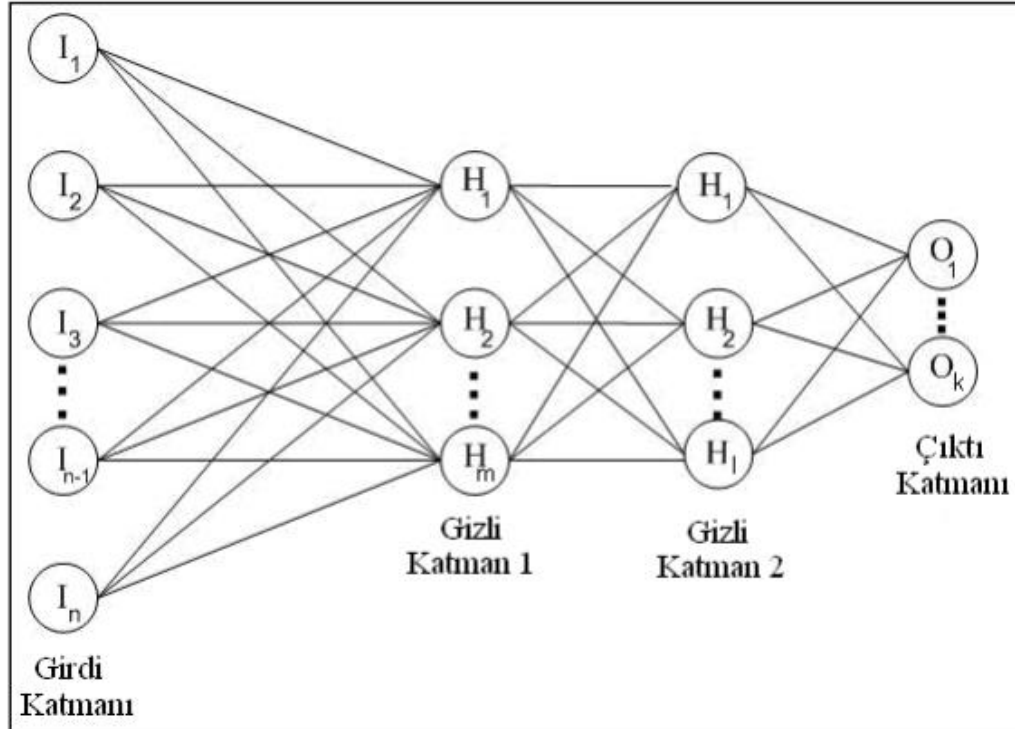
Makine Öğrenmesi Uygulama Alanları



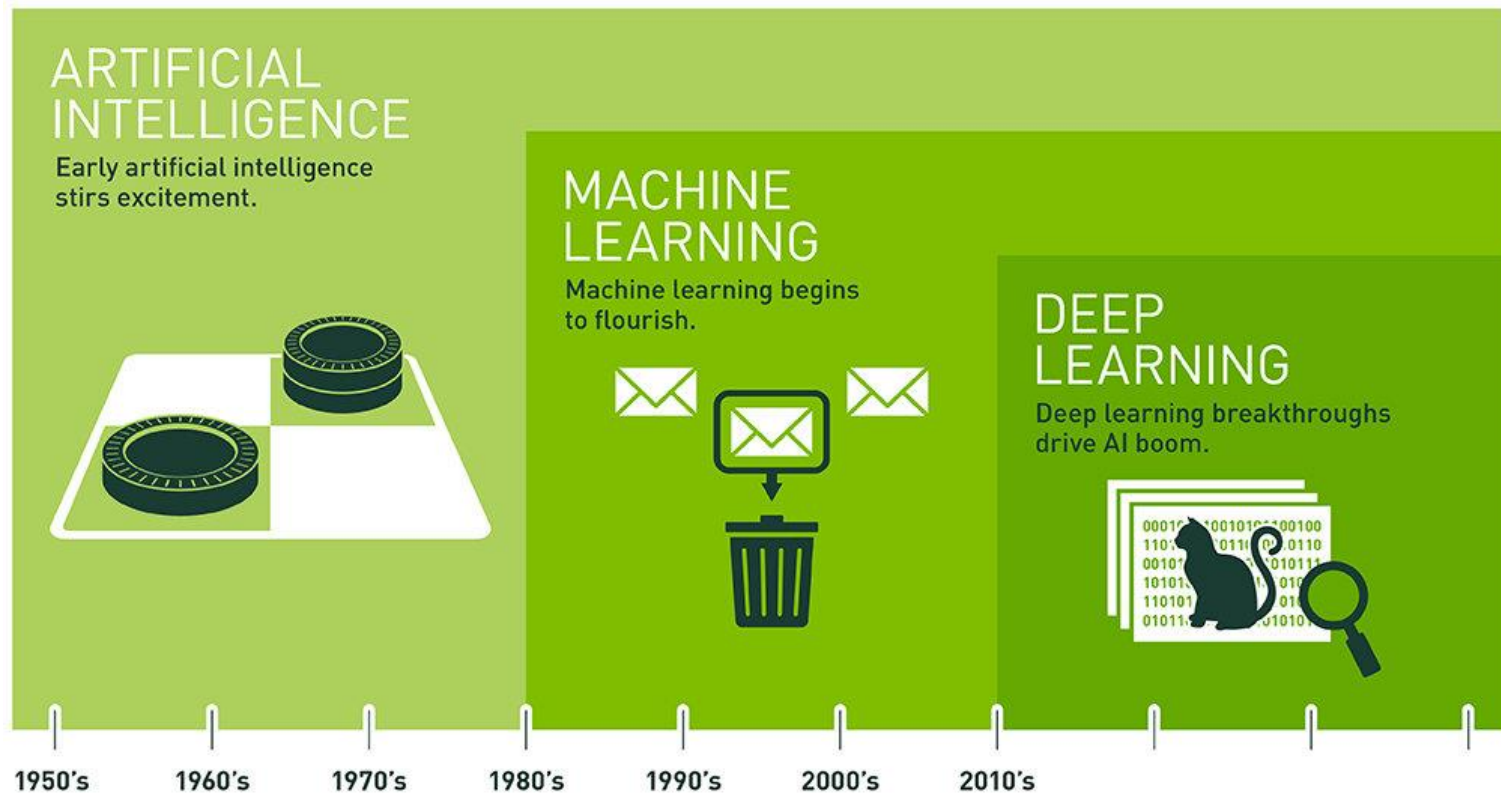
Derin Öğrenme Nedir?

- Deep Learning (Derin Öğrenme) aslında yepyeni bir kavram değildir.
- İnsan beyninin çalışma şeklinden ilham alınarak geliştirilen yapay sinir ağları (YSA) uzun süredir üzerinde çalışılan bir konu.
- Deep Learning bu yapay sinir ağlarının geliştirilmesiyle ortaya çıktı. Önceki yıllarda geliştirilmiş olan yapay sinir ağlarının belirgin özelliği bir veya iki katmandan oluşmasıydı.
- Derin Öğrenme modellerinde ise onlarca hatta yüzlerce katman olabilir, GPU üzerinde paralel çalışma gerçekleştirilebilir.

Derin Öğrenme Nedir? (devam...)



Genel Zaman Çizgisi Gösterimi (AI, ML, DL)

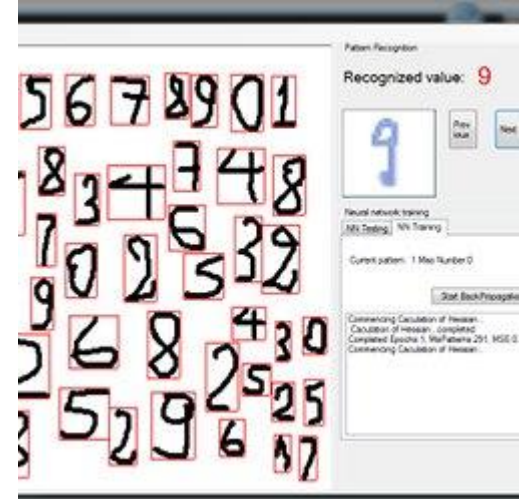


Günlük Hayattan Örnekler

El Yazısı / Kitap Yazısı Tanımı (HCR / OCR)

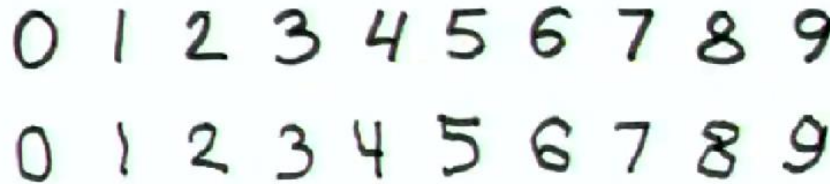
- İşlem: Şekillerin hangi harf olduğunu tahmin etme.

To markon manson - its TAKEN ME ALONG TIME TO GET
there from where I ~~could~~ Could touch M, manson
now I gat a Card to play - you may look in to my
none pruffit A.T.W.A + Give manson what you
think hes gat coming for Air, TREES, WATER,
+ you. or I will pay manson what you
think manson gat coming - the music has
make manson in to a ~~drax~~ Diver + I'm SURE
you would want some of what I gat from
what I gat. ITS a FAH OUT BALANCE + Behind
good + bad right wrong - what you Dont Dais
what I will do - what you did a song along
I let it roll + said how you saved me a lot
if steeps - I dont need it but I need a want
couple - COUP, Ghost dancers SLAY together + you
just in my Grane S UNSTROKER CORONA - CORONA - CORONA - YOU
seen me from under with it all standig on me + HOT
O DUMP TRUCKS - being the same as C.M.F 000007



Günlük Hayattan Örnekler

El Yazısı Tanıma (HCR)



0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9

Features: 

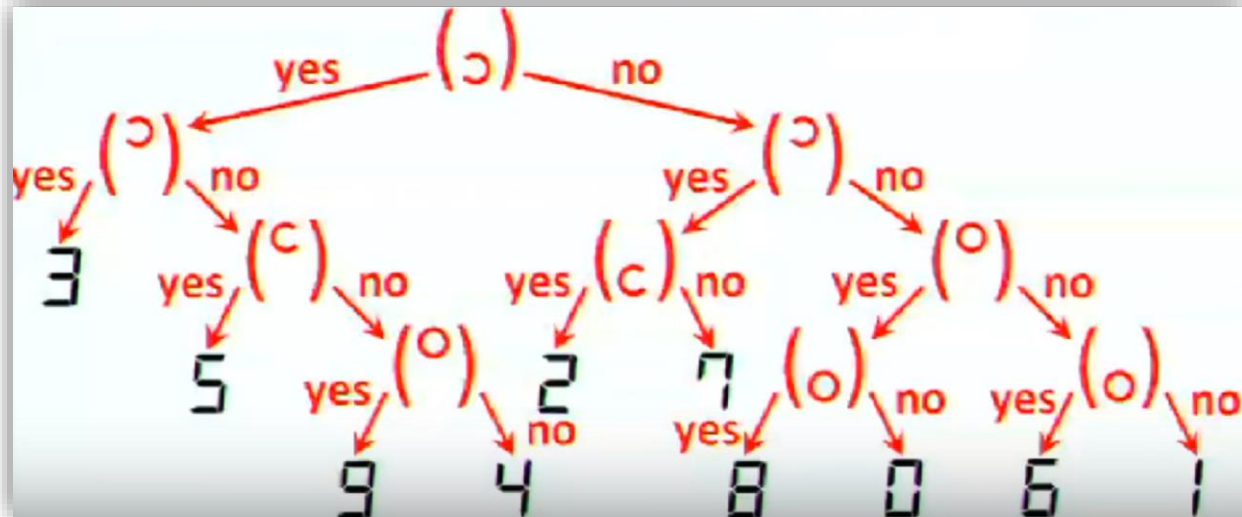
El Yazısı Tanıma (HCR)



Günlük Hayattan Örnekler

El Yazısı Tanıma (HCR)

0 1 2 3 4 5 6 7 8 9
(\emptyset) (\emptyset) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$) (\emptyset) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$) ($\begin{smallmatrix} \circ \\ \emptyset \end{smallmatrix}$) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$) ($\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$)



Günlük Hayattan Örnekler (devam...)

Kredi Taleplerini Değerlendirme

- **Senaryo:** Kredi veren bir kuruluş, iki önemli konuda karar vermelidir:
 - Birincisi, kredi başvurusunda bulunan kişiye kredi verip vermeyeceğine karar vermelidir,
 - İkincisi, mevcut kredi çekenlerin kredi limitini yükseltip yükseltmeyeceğine karar vermelidir.



Günlük Hayattan Örnekler (devam...)

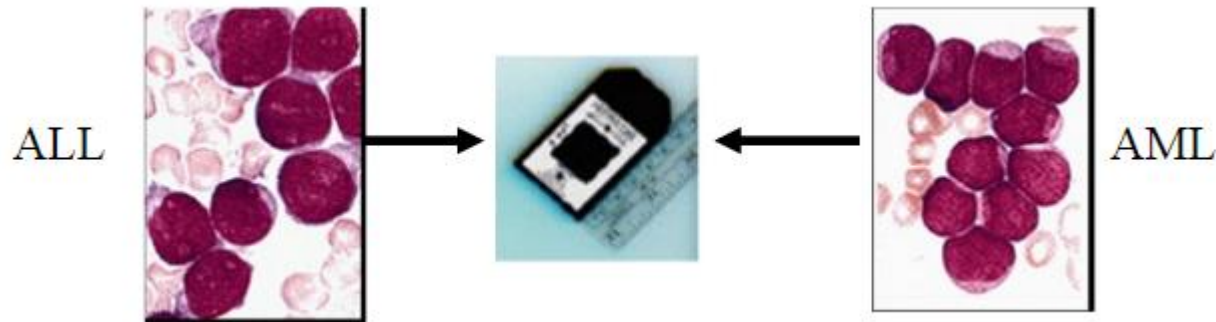
E-ticaret

- **Görev:** Müşteriye alması muhtemel kitaplar önerilir.
- Nasıl?
 - Kitapları
 - konularına
 - yazarlarına
 - birlikte satılış şekillerine
 - göre kümelemek.

Günlük Hayattan Örnekler (devam...)

Gen Mikro-dizilimleri

- **Görev:** 100 kişinin (hasta/sağlam) elimizde gen dizilimleri var. Bu dizilimleri analiz ederek hasta olup olmadığı bilinmeyen birisinin hasta olup olmadığını ya da hastalığının türünü öğrenebilir miyiz?
- En iyi tedaviyi önerebilir miyiz?
- Nasıl? Elimizde hangi bilgiler olmalı ?



Günlük Hayattan Örnekler (devam...)

Bu kişi kim, girişı izni var mı?



Günlük Hayattan Örnekler (devam...)

Kalabalık içerisinde aradığımız kişiyi bulabilir miyiz?



Günlük Hayattan Örnekler (devam...)

Bu metnin konusu nedir? Bu e-posta spam mi?

- Anti spam yazılımları nasıl çalışır ? Spam'ciler nasıl çalışıyor ?
- Yeni nesil spam e-postalar: Mesaj resimde, metinde ise anti spamlardan kaçmak için gereken kelimeler var.
- Makine öğrenmesi metotlarını hem spamciler hem anti spamciler kullanıyor.

From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - \$100
Medicine (any kind) - \$50
Also low cost M0rgages
available.

Spam

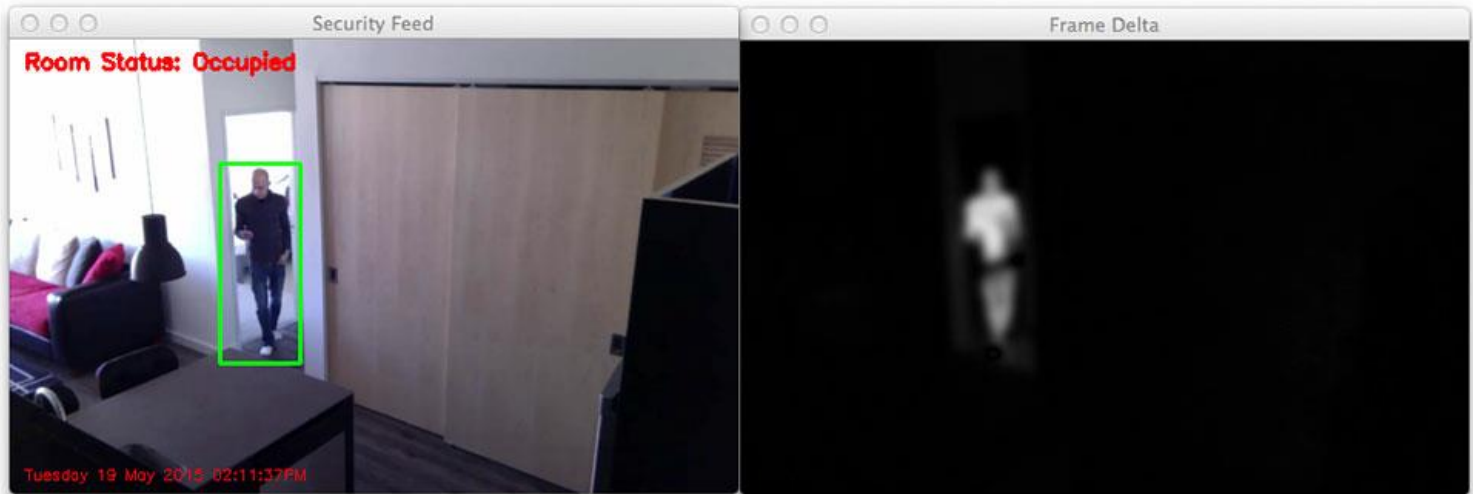
From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf

Non-spam

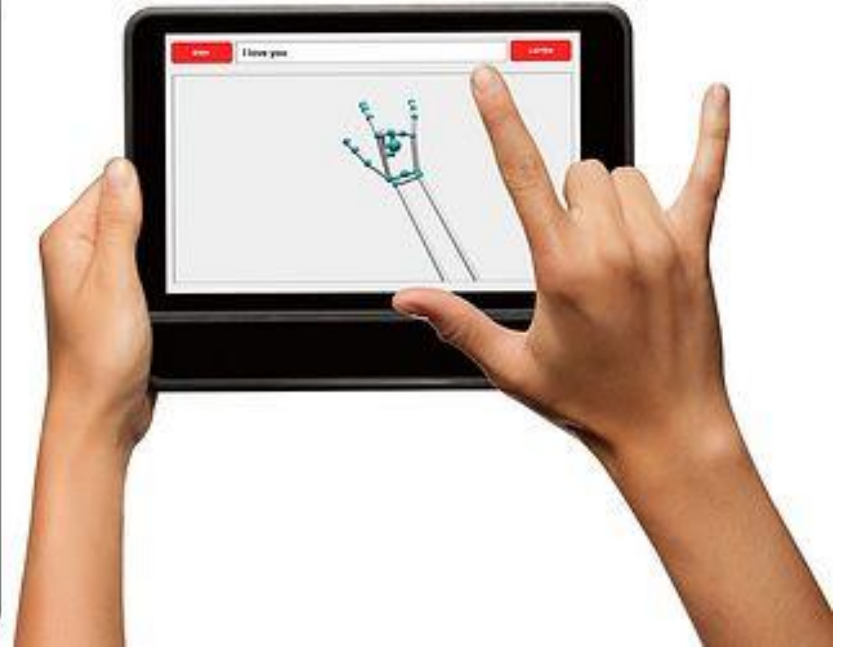
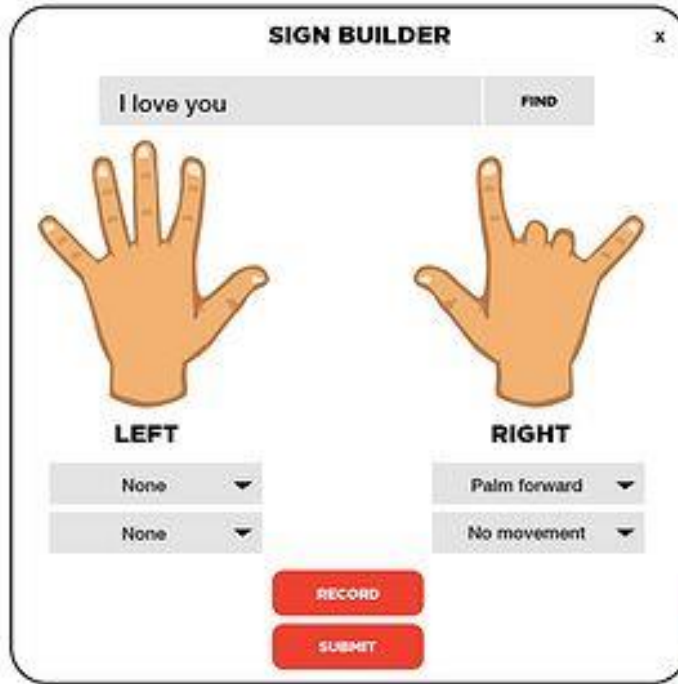
Günlük Hayattan Örnekler (devam...)

Olağan dışı bir durum var mı? Güvenlik kamerası kayıtları



Günlük Hayattan Örnekler (devam...)

Videodaki işaret dilini anlayabilir miyiz?



Makine Öğrenmesi Alan Terminolojisi

- **Gözlemler / Örnekler (Observations/Samples):** Öğrenmek ya da değerlendirmek için kullanılan her bir **veri parçası**. **Örn:** her bir e-posta bir gözlemdir.
- **Öznitelikler (Features):** Bir gözlemi temsil eden verilerdir. **Örn:** e-posta'nın uzunluğu, tarihi, bazı kelimelerin varlığı.
- **Etiketler (Labels):** Gözlemlerdeki **kategoriler**. **Örn:** spam, spam-değil.
- **Eğitim Verisi (Training Data):** **Algoritmanın öğrenmesi için** sunulan gözlemler dizisi. Algoritma bu veriye bakarak çıkarımlarda bulunur, kafasında **model kurar**. **Örn:** çok sayıda spam/spam-değil diye etiketlenmiş e-posta gözlemi.
- **Test Verisi (Test Data):** Algoritmanın kafasında şekillendirdiği modelin ne kadar gerçeğe yakın olduğunu test etmek için kullanılan veri seti. Eğitim esnasında saklanır, eğitim bittikten sonra **etiketsiz olarak algoritmaya verilerek** algoritmanın (vermediğimiz etiketler hakkında) tahminlerde bulunması beklenir. **Örn:** spam olup olmadığı bilinen (ama gizlenen), eğitim verisindekilerden farklı çok sayıda e-posta gözlemi

Veri Türleri

1. Nümerik Veriler: Sayısal-Nümerik-Nicel veriler de denmektedir. Boy, Yaş gibi süreklilik arz eden değerler Nümerik verilerdir. “Daha fazla” ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:

a) Sürekli Nümerik Veriler: Yaş, Sıcaklık

b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı

2. Nominal Veriler: Kategorik bir veri çeşidir. “Daha fazla” ifadesi ile kullanılmazlar. İkiye ayrılır:

a) Dikotom Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı

b) İkiden Çok Kategorili: Medeni Durum-Renk-Irk-Şehir, İsim, Forma Numarası

Veri Türleri (devam...)

ID	NAME	DATE OF BIRTH	GENDER	CREDIT RATING	COUNTRY	SALARY
0034	Brian	22/05/78	male	aa	ireland	67,000
0175	Mary	04/06/45	female	c	france	65,000
0456	Sinead	29/02/82	female	b	ireland	112,000
0687	Paul	11/11/67	male	a	usa	34,000
0982	Donald	01/12/75	male	b	australia	88,000
1103	Agnes	17/09/76	female	aa	sweden	154,000

Diagram labels and arrows:

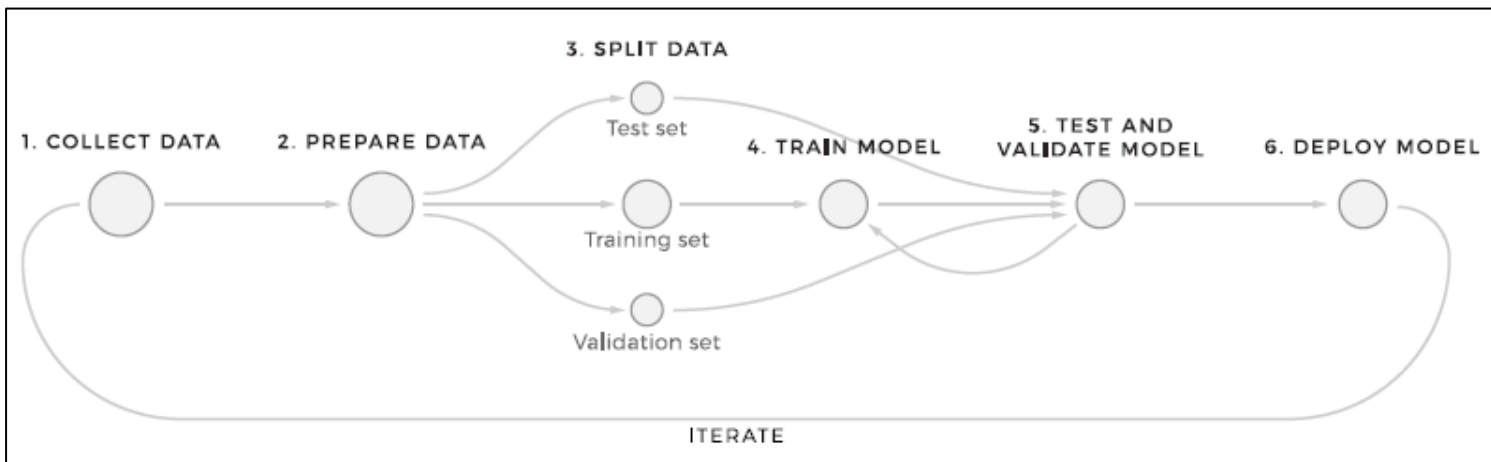
- Ordinal** (two arrows): one to **ID**, one to **CREDIT RATING**
- Ordinal** (one arrow): to **GENDER**
- Categorical** (one arrow): to **COUNTRY**
- Textual** (one arrow): to **NAME**
- Interval** (one arrow): to **DATE OF BIRTH**
- Binary** (one arrow): to **GENDER**
- Numeric** (one arrow): to **SALARY**

Makine Öğrenmesi İş Akışı

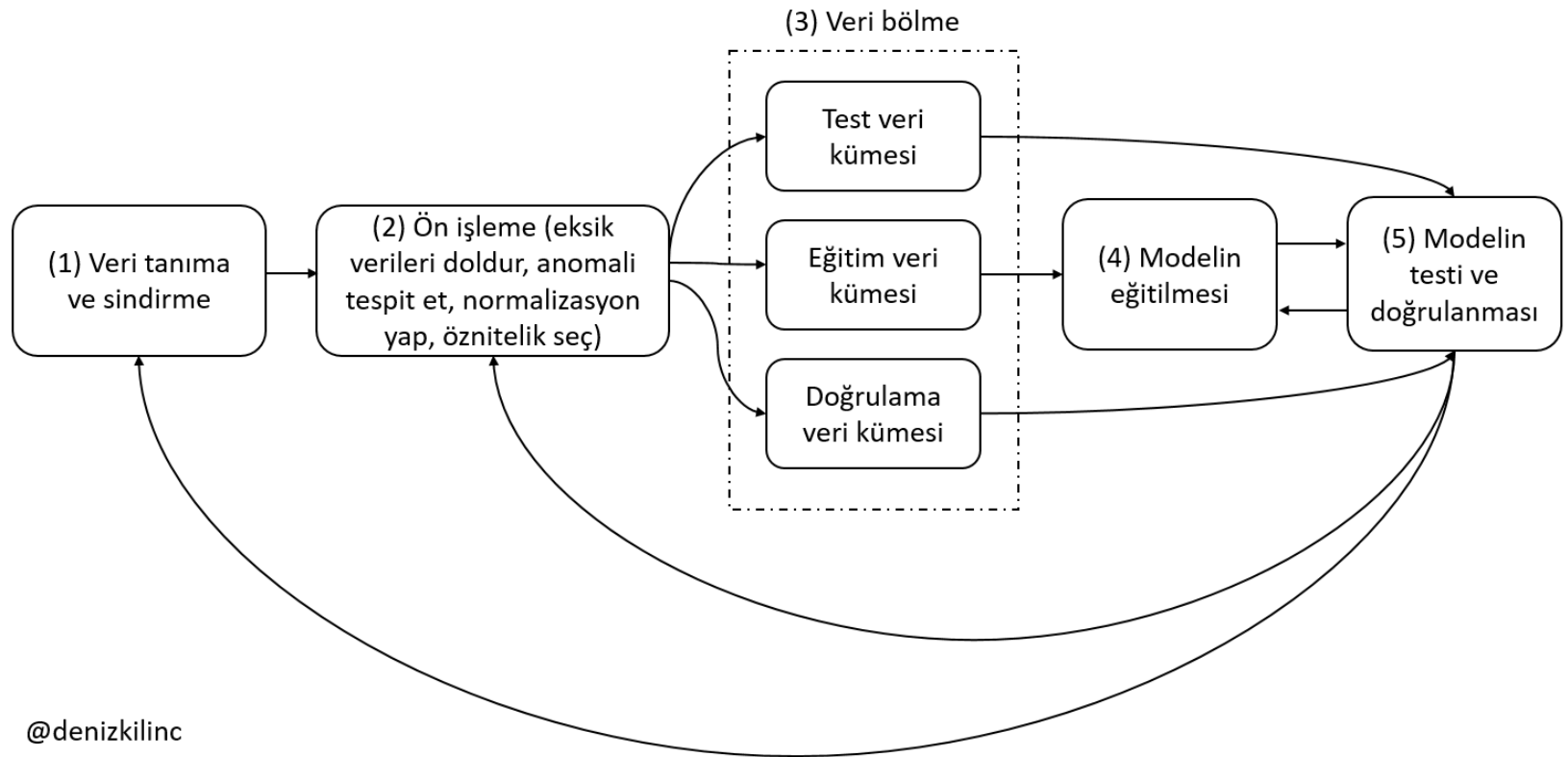
- 1) Veri Topla (Collect data):** Farklı veri kaynakları kullanılarak işe yarayacağını düşündüğün olabildiğince veriyi topla.
- 2) Veri Hazırla (Prepare data):** Veri üzerinde ön işleme çalışmalarını gerçekleştir. Eksik verileri tamamla, anomalileri bul ve düzelt. Normalizasyon yap. Öznitelik seçimi gerçekleştir.
- 3) Veriyi böl ve modeli eğit (Split data and Train model):** Veriyi uygun şekilde test ve eğitim veri kümeleri olarak ayır, uygun bir öğrenme algoritması seç (Naive Bayes, Karar Ağaçları, SVM vb.) ve eğitim işlemini gerçekleştir.

Makine Öğrenmesi İş Akışı (devam...)

- 4) Modeli test et ve doğrula (Test and validation):** Modelin ürettiği sonuçların gerçekten ne kadar doğru tahmin yaptığını anla.
- 5) Modeli deploy et:** Test ettiğin ve doğruladığın modeli, bir analitik çözüm programı olarak son kullanıcılara aç.
- 6) Tekrarla (Iterate):** Zaman içerisinde topladığın yeni verilerle modeli sürekli iyileştir.



Makine Öğrenmesi İş Akışı (devam...)



@denizkilinc

Makine Öğrenmesi Türleri

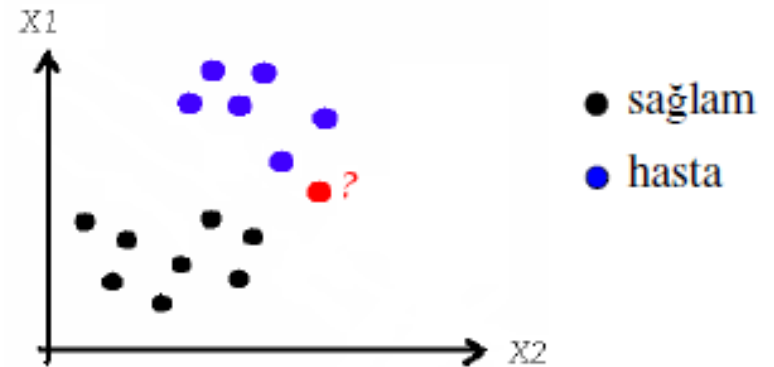
1. Denetimli Öğrenme (Supervised Learning)

- Etiketlenmiş gözlemlerden öğrenme sürecidir.
- Etiketler, algoritmaya gözlemleri nasıl etiketlemesi gerektiğini öğretir.
- Örneğin içinde "para kazan" ifadesi geçiyorsa *spam demelisin* gibi yol göstermelerde bulunur.

a) Sınıflandırma (Classification):

Geçmiş gözlemlerin hangi sınıftan olduğu biliniyorsa, yeni gelen verinin hangi sınıfa dahil olacağını bulmaya çalışır.

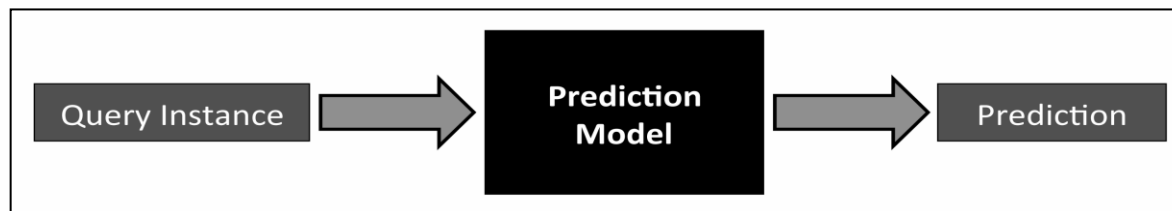
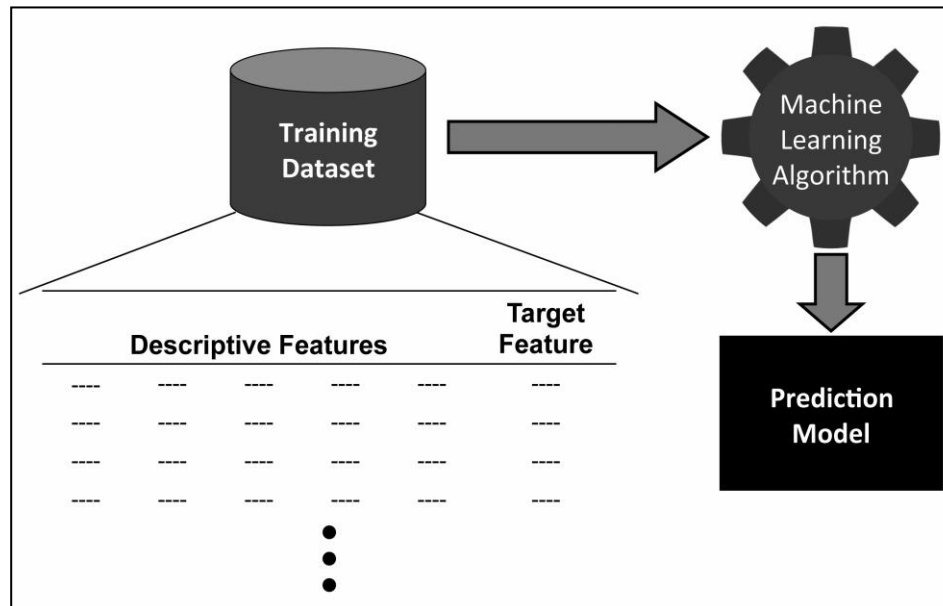
Örn: spam/spam değil. Sınıflar ayrıktır (sayı değildir) ve birbirlerine yakın/uzak olmaları gibi bir durum söz konusu değildir.



Kırmızı hangi sınıfa dahildir ?

Makine Öğrenmesi Türleri (devam...)

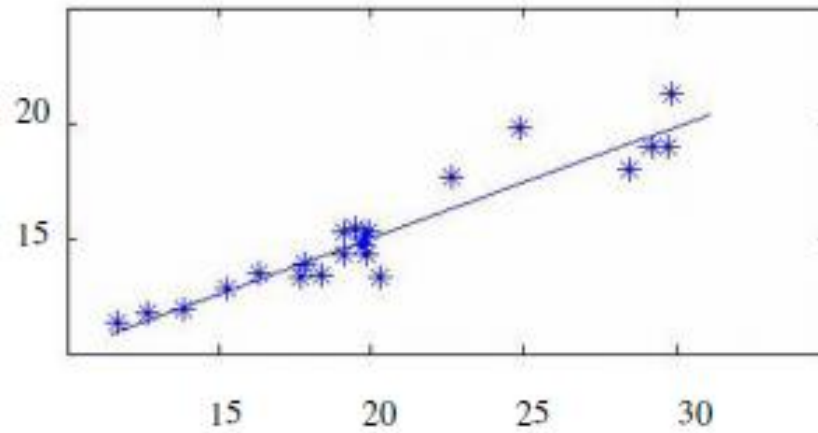
a) Sınıflandırma (Classification) (devam...)



Makine Öğrenmesi Türleri (devam...)

b) Regresyon (Regression): Geçmiş bilgilere ait sınıflar yerine sürekli bilginin yer aldığı problemlerdir. Eğri uydurma olarak da geçer. x eksenini **hava sıcaklığını**, y eksenini de **deniz suyu sıcaklığını** göstermektedir.

- Bizden istenen **hava sıcaklığına bağlı olarak deniz suyu sıcaklığının tahmin edilmesidir**. Giriş ile çıkış arasındaki fonksiyonun eğrisi bulunur.



$$y = ax + b$$

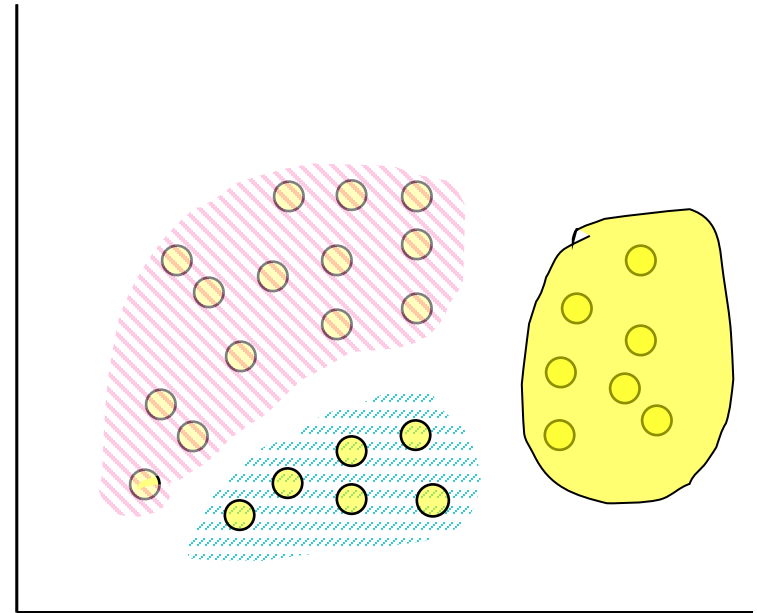
Makine Öğrenmesi Türleri (devam...)

2. Denetimsiz Öğrenme (Unsupervised Learning)

- **Etiketsiz** gözlemlerden öğrenme sürecidir.
- Algoritmanın kendi kendine keşifler yapması, **gizli** örüntüleri *keşfetmesi* beklenir.

a) Kümeleme (Clustering):

Geçmişteki verilerin sınıfları/etiketleri verilmediği/bilinmediği durumlarda verilerin birbirlerine **yakın** benzerliklerinin yer aldığı kümelerin bulunmasıdır.



Öğrenme Öncesi Veri Toplama

- Öncelikle **ham veri** (raw data) buluyoruz.
 - Metin, görüntü, genetik verisi, sayısal ölçümler, sosyal ağlar, kullanıcı puanlamaları...
- **Örn:** Regresyon tipi denetimli öğrenme:
- Elimizde hastaların **sağlık verileri** olsun.
- **Her hasta için:**
 - yaş,
 - cinsiyet,
 - sigara içiyor mu?,
 - günde kaç sigara içiyor?,
 - şeker hastalığı var mı?,
 - daha önce kalp krizi geçirdi mi?,
 - kolesterolü kaç?, nabızı kaç?, şeker seviyesi kaç?

Öğrenme Öncesi Veri Toplama (devam...)

Ad	AnneAdı	DahaOnceKriz	SigaraSayisi	Kolestrol	...	KrizGeçirdiMi
Mehmet	Leyla	1	12	100	..	1
Murat	Ayşe	0	6	170	..	1
Fahri	Tuba	0	3	30	..	0

- Burada her gözlem için girilmiş farklı öznitelik değerleri görüyoruz.
- Amacımız bu özniteliklerden yararlanarak, hastanın **önümüzdeki bir zaman diliminde kriz geçirip geçiremeyeceğini** tahmin etmek.
- Peki bu özniteliklerin *hepsi* gerçekten **değerli mi**?

Öznitelik Çıkarımı ve Seçimi

- Gözlemlerimizi en iyi temsil edecek öznitelikleri seçmek için ham veriden **Öznitelik Çıkarımı (Feature Extraction)** ve çıkarılan öznitelikler üzerinde **Öznitelik Seçimi (Feature Selection)** yapılmalıdır.
- Bu süreçte alan bilgisi (domain knowledge) uygulamak ve bu öznitelikleri *tanımlayıp* **hesaplamak** gerekli.
- Örneğin kalp hastalığı tespitinde "AnneAdı" özelliğinin *işe yaramayacağını* öngörebiliriz.
- En çok şüphelendiğimiz *sigara sayısı*, *daha önce kalp krizi geçirdi mi?* gibi öznitelikleri hesaba katmakta fayda var. Gereksiz kısımları temizlemeliyiz.

Öznitelik Çıkarımı ve Seçimi (devam...)

- Peki elimizde onlarca/yüzlerce öznitelik olsaydı ne yapardık?
- Bu noktada devreye bilgisayarları ve algoritmaları sokmamız gerekmektedir.
- İki tip öznitelik seçim yaklaşımı bulunmaktadır.
 - Öznitelikleri tek tek değerlendirmek (**Filter**)
 - Öznitelik alt kümeleri oluşturup, sınıflandırıcılar kullanıp performanslarını ölçüp, bu alt kümeleri en iyi sonucu elde etmek için değiştirerek denemek (**Wrapper**)
- **Information Gain (Bilgi Kazancı)**: Filter kategorisinde önemli bir öznitelik seçim yaklaşımıdır.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Model Oluşturma

- öznitelikler seçim işlemi tamamlandıktan sonra bir **öğrenme algoritması** bu matris üzerinde çalışır.
- Sonucunda bir takım kurallar ortaya çıkar.
- **Örneğin** basit bir algoritma az sayıda veriye bakarak şöyle bir model ortaya koyabilir:

Kriz olasılığı = günlük sigara sayısı x 0.05 + kolesterol seviyesi x 0.004

- Gözlem sayısı arttıkça model karmaşıklaşacaktır.

Model oluşturma sırada ne var?

Değerlendirme ve Test

- Modelimizi test etmek için aşağıdaki gibi bir test verisine sahibiz.

Ad	AnneAdı	DahaOnceKriz	SigaraSayisi	Kolestrol	...	KrizGecirdiMi
Mert	Beyza	1	10	80	..	1
Davut	Fatma	1	15	40	..	1
Veysel	Kadriye	0	5	15	..	0

- Elimizde etiketi gizlenmiş hasta gözlemlerini (test verisini) modele uygulayarak her birisi için bir **olasılık** çıkarıyoruz.
- Ardından **olasılık 0.5'den büyükse** kriz geçireceğini iddia ediyoruz.
- Modelin ürettiği bu tahminleri **elimizdeki gerçek kriz bilgileriyle** karşılaştırıp ne kadar başarılı bir tahmin yaptığını değerlendiriyoruz.

Değerlendirme ve Test (devam...)

Ad	AnneAdı	DahaOnceKriz	SigaraSayisi	Kolestrol	...	KrizGecirdiMi	HesaplananOlasılık
Mert	Beyza	1	10	80	..	1	0.82
Davut	Fatma	1	15	40	..	1	0.91
Veysel	Kadriye	0	5	15	..	0	0.85

- Eşik olarak 0.5 kullandığımızda bu test verisindeki **herkesin** *kalp krizi geçirmesini* bekliyoruz ama **Veysel** *geçirmemiş*.
- Eğer eşiği 0.9 deseydik *Veysel'in kriz geçirmeyeceğini* **bilecektik** ama **gerçekte kriz geçiren Mert'in kriz geçireceğini** iddia edemeycektik.

Değerlendirme ve Test (devam...)

- Değerlendirmede confusion matrisi ve doğruluk (*Accuracy*, ACC) kullanılır:

		Tahmin Edilen Sınıf	
		Sınıf=1	Sınıf=0
Gerçek Sınıf	Sınıf=1	TP (True Pozitif)	FN (False Negatif)
	Sınıf=0	FP (False Pozitif)	TN (True Negatif)

$$\text{Doğruluk} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- Sensitivity (hassaslık)**, gerçekten kriz geçiren kimselerden yüzde kaçını "kriz geçirecek" diyerek bildik?
- Specificity (belirginlik)**, kriz geçirmeyen kimselerden yüzde kaçını "kriz geçirmeyecek" diyerek bildik?

Değerlendirme ve Test (devam...)

- 0.5 olarak kullandığımız karar katsayısını **0.95 yaparak** *"iddiamdan ancak çok eminsem kriz geçirecek"* diyebiliriz, bu da **hassaslığı düşürür**. Çünkü kimse kriz geçirmeyecek demiş oluruz.
- Katsayıyı 0.3 yaparsak ise birçok kişiye **potansiyel hasta muamelesi** yapmış oluruz, durum tersine döner.
- Modelin karmaşıklığına göre tahmin sonuçları değişecektir.
- Eğer model çok basitse (az sayıda özelliğe bakarak karar veriyorsa) çok daha hızlı çalışacaktır.
- Eğer **iyi genelleme yapabiliyorsa** farklı veri setlerinde **iyi tahminler** yapabilir.

İyileştirme

- Eğer modelin yeterince başarılı olmadığını düşünüyorsak nerelerde hata yaptığını inceleyip modelimize hangi özellikleri vermemiz gerektiğini düşünmeliyiz.
 - Bazı **özellikleri çıkarmalı**,
 - Bazı **özellikleri eklemeli**,
 - Model parametrelerini **değiştir**,
 - Yeniden **bir model oluşturup** tekrar değerlendirmeliyiz.
- Bu süreç tatmin olana kadar devam edebilir.
- Ama **sonu yoktur**.

Sonuç

- Makine Öğrenmesi çok faydalı olabilmektedir ancak *sihirli bir değnek de değildir*,
 - Farklı algoritmalar,
 - Farklı öznitelikler,
 - Farklı veriler ile
 - Çok sayıda deneme yanılma girişimiyle model oluşturmak sabır gerektirir.

İstatistik ve Matematik

- Mean (Ortalama)
- Median (Medyan)
- Std. Deviation (Standart Sapma)
- Variance (Varyans)
- Covariance (Kovaryans)
- Correlation (Korelasyon)
- Matrices operations (Matris operasyonları)
- Türev
- Integral
- Polinomlar
- ...

Makine Öğrenmesi Araç ve API'ler

- Microsoft Azure Machine Learning Services
- Amazon Machine Learning Services (AWS)
- Google Machine Learning and Prediction Services
- Google Cloud Vision API
- IBM Watson Analytics
- Weka
- RapidMiner
- KNIME
- Python ve ilgili paketler (Numpy, Pandas, Scikit-learn...)
- R ve ilgili paketler

Veri Bilimi ve Python

- Python günümüzde veri bilimi denilince akla ilk gelen programlama dillerinden birisidir.
- Oldukça yaygın kullanılmasının en temel sebepleri; **kolay öğrenilebilen**, **kolay okunan** ve **birçok hazır veri bilimi kütüphanesi bulunduran** bir dil olmasıdır.
- Popülerliği ve kullanımı arttıkça, var olan kütüphaneleri daha fazla geliştirmekte ve ayrıca yeni kütüphaneler de eklenmektedir.
- Veri bilimi çalışmaları için en temel ve en yaygın kullanılan; Numpy, Pandas, Scikit-Learn, Keras gibi birkaç kütüphaneyi ve işlevlerini ayrıntılı olarak açıklayalım.

Numpy

- İsmi Sayısal Python (Numerical Python) kelimelerinin kısaltmalarından oluşan Numpy kütüphanesi, hızlı matematiksel işlemler yapabileceğimiz diziler (array) sunar.
- Python'un kendi veri yapısında bulunan liste veri yapısına nazaran oldukça hızlı çalışan Numpy kütüphanesiyle rastgele sayı üretebilir, matris çarpımlarından doğrusal cebir işlemlerine ve Fourier dönüşümlerine kadar birçok matematiksel işlemi gerçekleştirebilirsiniz.

Pandas

- Pandas kullanımı kolay, yüksek performanslı bir veri yapılandırma ve veri analizi kütüphanesidir.
- Bu kütüphane ile excel, json, metin (csv) ve veritabanı gibi birçok farklı kaynaktan veri okunabilir ve bu kaynaklara veri yazılabilir.
- Tek boyutlu olarak Serie isimli, 2 boyutlu olarak da DataFrame isimli tablo yapısını içinde barındırır.
- Pandas tabloları içerisinde birçok tipte değişken tutulabilirler (sayısal, kategorik, tarih vb.).
- Veri dönüştürme, filtreleme gibi önemli veri ön işleme aşamaları bu kütüphane ile kolayca gerçekleştirilebilir.

Scikit-Learn

- Oldukça yaygın kullanıma sahip olan bu kütüphane, birçok makine öğrenmesi algoritmasının (Naive Bayes, Decision Tree, K-NN, Multi-Layer Perceptron, SVM, Ensemble vb.) gerçekleştirimini içinde barındırmaktadır.
- Bu algoritmalara ek olarak, kütüphanede;
 - Boyutsal küçültme (dimensionality reduction),
 - Veri ön işleme ve model seçme yöntemleri de yer almaktadır.

Keras

- Keras, diğer derin öğrenme kütüphanelerine nazaran daha kolay gerçekleştirimi yapılabilen, derin öğrenme çalışmalarında tercih edebileceğimiz bir kütüphanedir. Keras arka planda Python'da bulunan *Tensorflow* ve *Theano* isimli kütüphanelerden birini tercihe bağlı olarak kullanır.
- Keras'ta ise birçok derin öğrenme algoritması hazır olarak yer almaktadır. Tek yapmamız gereken, verimizi derin öğrenme ağının yapısına uygun hale getirmek ve hiper parametre seçimlerini yapmaktır. *Tensorflow* ve *Theano* arka planı sayesinde Keras'ta eğittiğimiz modeli, CPU üzerinde çalıştırabileceğimiz gibi aynı zamanda GPU üzerinde Cuda kütüphanesini kullanarak paralel bir şekilde de çalıştırabiliriz.



Leading Open Data Science Platform Powered by Python



Leading Package and Environment Manager

OPEN DATA SCIENCE



DATA



COMPUTATION



Yararlanılan Kaynaklar

- Kitaplar
 - Introduction to Machine Learning (Ethem Alpaydın)
 - Veri Bilimi Uygulama Senaryoları (Deniz Kılınç, Nezahat Başeğmez)
- Udemy Kursları
 - <https://www.udemy.com/user/datai-team/>
- Web Siteleri
 - <http://machinelearningmastery.com/>
 - <http://veribilimi.co>
 - <http://medium.com/deep-learning-turkiye/>

İYİ ÇALIŞMALAR...

Doç. Dr. Deniz KILINÇ

drdenizkilinc@gmail.com

deniz.kilinc@bakircay.edu.tr