Büyük Veri Uygulamaları – Ders 2

Doç. Dr. Yuriy Mishchenko

Plan

- Büyük veri nedir, kaynaklar nedir, kullanım alanları
- Örnekler
- Felsefesi ve temel yaklaşım
- Diagramlar
- Yaklaşım prensibi
- Ayrıntılı örnekler

Büyük veri kavramı

Büyük veri nedir?

Diğer değişle, Büyük Veri eski, alıştığımız metotlarla işletebilmek için fazla miktarda olan veriler demektir

https://en.wikipedia.org/wiki/Pointy-haired_Boss

Büyük veri nedir?

- Zamanımızda veri çök yüksek hızla üretilip toplanmaktadır
  - İnternette ziyaretçilerin tıklamaları
  - Alişveriş ödemeleri
  - Sensör kayıtları
  - Güvenlik kameraların kayıtları
  - GPS verileri
  - Sosial media etkileşimleri
  - ...
- Bu tip verileri kaydedip işletmek ve analiz etmek gerçekten zor olmaya başlamıştır

WWW'DEN VVV'E


VOLUME
VELOCITY
Web logs
Audios
Images
Videos
Sensor Data
VARIETY

WWW'DEN VVV'E

- **V**olume (Hacim)
  - Bugünkü iş/kurum/şirketlerin verileri inanılmaz miktarda dır
- **V**ariety (Çeşitliği)
  - İş/kurum/şirketlerin verilerinin karmaşıklığı artmaktadır
  - Her gün yeni tür veriler toplanmaya başlamaktadır
- **V**elocity (Hız)
  - Verilerin toplama hızı artmaktadır
  - Baze verilerin doğasına göre anlık işlenmesi ve tepki zorunludur – bu tür verilere "veri akışları" (data streams) denir

Büyük verinin 4üncü V

- **V**eracity (Kalitesi)
  - Veri toplamanın hızı artmakla beraber, verilerin kalitesi ve güvenilirlik düşmektedir
  - Veriler sağlayan kaynaklardan yanlış, eksik, bozuk biçimdee veriler gelebilmektedir

Beklenti

- Bügünkü sistemler, şirketler ve kurumlar Terabyte'ten Petabyte'e kadar rutin olarak bilgi üretmekte dir
- Bilgi, şirket/kurumun başarısı için büyük önem taşımakta dır
- İyi kararlar verebilmek için anlamlı verilerin var olması ve dikkate alınması şarttır

Büyük Verinin Üç V

Volume

- Terabyte
- Veri kayıtları
- İşlem kayıtları
- Tablolar
- Dosyalar
- Loglar

Variety

- Yapılandırılmış
- Yapılandırılmamış
- Yarı-yapılanıdırılmış
- Karışık

Velocity

- Batch veri analizi
- Neredeyse gerçek zamanlı veri
- Gerçek zamanlı veri
- Veri akışları

ÇEŞITLIĞI

- XML dosyaları → yarı yapılandırılmış veri
- Word, PDF, TXT dosyaları → yapılandırılmamış veri
- Email metinleri → yapılandırılmamış
- Veri tabanları ve Excel tabloları → yapılandırılmış
- Sosial media mesajları → yapılandırılmamış
- İşlem logları → yarı yapılandırılmış


BÜYÜK VERILERDEN –
BÜYÜK BEKLENTILER !

KULLANıM ALANLARı

- İnternet ve e-komerse
  - Tavsiye (recommender) motorları
  - Reklam hedeflenmesi
  - Arama kalitesi artırma
  - Yasal olmayan faaliyetlerin keşfetme
- Telekom
  - Müşterilerin memnuniyeti takip etme
  - Telekom ağın performansı takip etme
  - Telekom ağları optimizasyon
  - Müşterilerin davranışları analizi
  - Ağın çalışması analizi ve arza durumlarının tahmin etme

KULLANıM ALANLARı

- Devlet
  - Siber güvenlik
  - Emniyet ve yasal araştırmaları
  - Nüfus bilgi toplama ve analizi
  - Ekonomik bilgi toplama ve analizi
  - Diğer
- Tıp ve medikal
  - Tıp kayıtların işlenmesi
  - Genetik araştırmaları
  - Hizmet iyileştirilmesi
  - İlaç/tedavi güvenliği araştırmaları

KULLANıM ALANLARı

- Banka ve finans
  - Risk modellenmesi
  - Tehdit analizi
  - Dolandırıcılık keşfetme
  - Kredi skorlar
- Perakende
  - POS kayıtların analizi
  - Müşteri memnuniyet takibi
  - İmaj araştırmaları ve analizi

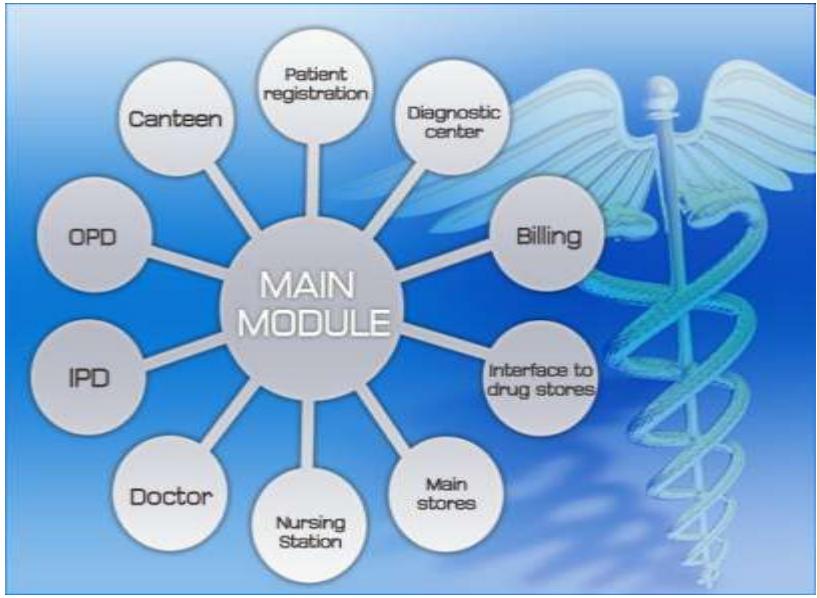
SPESIFIK ÖRNEKLERI: SPOR

SPESIFIK ÖRNEKLERI: SPOR

- Büyük sporda büyük veri yaklaşımı bilet satışı, pazarlama ve reklam stratejilerinin geliştirilmesi için kullanılır,
- Bunun için sosial media kullanılarak reklam kampanyaların verimliği ve alması gereken yönler seçiliyor
- Spor takımları “büyük veri” modelleri oyun strateji, oyuncular seçimi vb konular için de kullanılır

SPESIFIK ÖRNEKLERI: TıP

- Tıpta hastaların kayıtları analiz edilerek tedavi sonuçları ve daha iyi müdahale yapmak için yöntemler araştırılmakta
- Genel test sonuçları kullanılan teşhis tespiti yöntemleri de geliştirilmekte



SPESIFIK ÖRNEKLERI: E-KOMERS

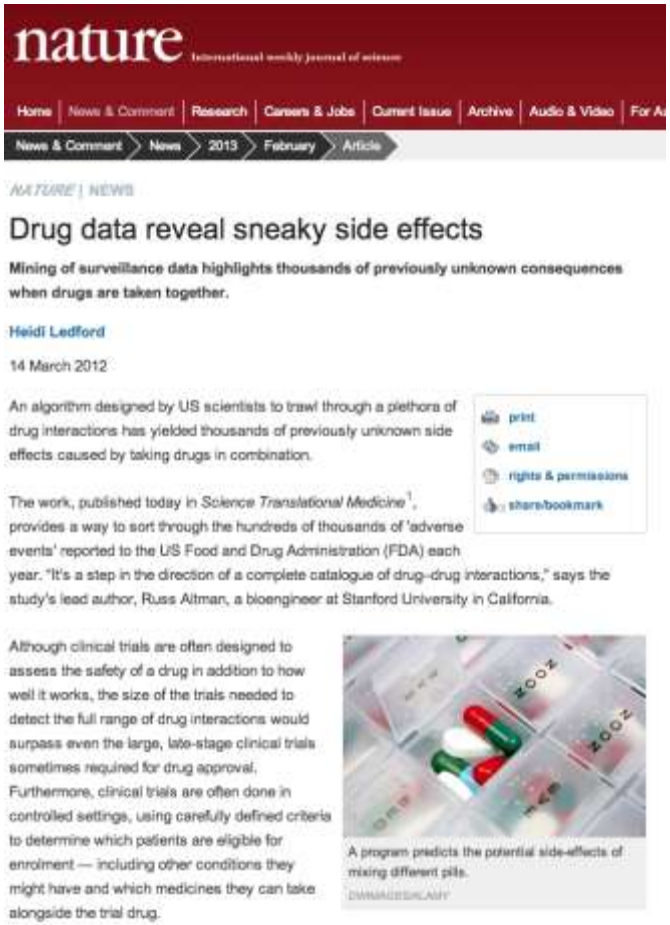
- Online satıcılar inanılmaz miktarda kullanıcı ve ürün hakkında bilgilere sahiptir
- Bu bilgi, kullanıcıların davranışları analiz etmek, hedeflenen reklam üretmek ve alişverişteyken daha faydalı ürün önerileri yapmak için kullanılmakta

SPESIFIK ÖRNEKLERI: ONLINE

- Netflix online film seyretme hizmeti dir
- Kullanıcının daha önce seyrettiği film tarihçesi kullanılarak, kullanıcının ilgisi özel olarak çekebilecek film önerileri makine öğrenme yöntemleri yardımıyla Netflix'in seçtiği ünlü dür
- Netflix'in film deposu da 1Petabyte geçtiği bilinmekte

SPESIFIK ÖRNEKLERI: ILAÇ

- Tıpta ilaç ve tedavi güvenliği ve verimliği araştırmaları yeni değildir
- Bu araştırmalar, büyük veri kavramı temel olarak oluşturan veri modelleme ve veri analizi yaklaşımları yardımıyla gerçekleştirilmekte



nature

International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment > News > 2013 > February > Article

NATURE | NEWS

Drug data reveal sneaky side effects

Mining of surveillance data highlights thousands of previously unknown consequences when drugs are taken together.

Heidi Ledford

14 March 2012

An algorithm designed by US scientists to trawl through a plethora of drug interactions has yielded thousands of previously unknown side effects caused by taking drugs in combination.

print
email
rights & permissions
share/bookmark

The work, published today in *Science Translational Medicine*[1], provides a way to sort through the hundreds of thousands of 'adverse events' reported to the US Food and Drug Administration (FDA) each year. "It's a step in the direction of a complete catalogue of drug–drug interactions," says the study's lead author, Russ Altman, a bioengineer at Stanford University in California.

Although clinical trials are often designed to assess the safety of a drug in addition to how well it works, the size of the trials needed to detect the full range of drug interactions would surpass even the large, late-stage clinical trials sometimes required for drug approval. Furthermore, clinical trials are often done in controlled settings, using carefully defined criteria to determine which patients are eligible for enrolment — including other conditions they might have and which medicines they can take alongside the trial drug.

A program predicts the potential side-effects of mixing different pills.

SPESIFIK ÖRNEKLERI: ALIŞVERIŞ

- Target, Amerikadaki bir alişveriş zinciri, müşterilerinin alış tarihçesine bakarak hedeflenen reklam üretmeye çalışmakta
- 2012'de bu yöntemler kullanılarak aylesi bilmeyen bir genç kızın hamile olduğunu ve yaklaşık doğum tarihi tespit ettiği ile haberlere çıkmıştır


HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | Global Edition

The New York Times | International Herald Tribune

Magazine

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

How Companies Learn Your Secrets

By CHARLES DUHIGG

Published: February 16, 2012 | 870 Comments

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department

FACEBOOK

TWITTER

GOOGLE+

SPESIFIK ÖRNEKLERI: TARıM

- Tarım ve hayvan bilimi, tarımda kullanılan bitki ve hayvanların genetik bilgileri ve üreme tarihçeleri gibi veri analiz ederek, performansı artırmak için yapay seleksiyon planları modellenir ve kullanılmakta

Daha çok örnek

- Spor
  - Basketbol oyunların planlanmasında veri analiz oldukça kullanılır
  - Futbolda benzer eğilim görünmektedir
- Eğlence sektöründe
  - Bilgisayar oyunları müşteri veri analizine büyük önem vermektedir
  - Yapılacak filmlerin seçiminde benzer eğilim vardır
- Finans
  - Viza otomatik ödeme bilgilerinin analizi ile dolandırıcılık keşfetme sistemleri geliştirmiştir
- Google ve Facebook
  - Kullanıcıların bilgilerinde veri madenciliği yaparak reklam ve benzer faaliyetleri hadefliyor

Daha daha çok

- Tarım başkanlığı
  - Tarım şirket ve çiftliklerin verileri
  - Doğum, ölüm, taşınma, tedaviler, örnekler gibi verileri
- Enerji üretim
  - Elektrik enerji üretimi ve tüketim, en uygun dağıtım şekli, dinamik elektrik fiyatları, arza ihtimali, müşterin sayaç kurcalama
- Petrol ve madencilik
  - Geolojik veriler, işlem veriler, lojistik, mühendislik
- Perakendeciler
  - Müşteri modellenmesi, önceki Target örneğine bakın
  - Satışlar ve hava, sezon vb durumlarla ilişkileri, lojistik ve stok yönetimi

ORTAK DESEN …

Genel biçimde olan çeşitli veri serileri

- Alişveriş kayıtları
- Üreme tarihçesi
- Süt/et üretimleri
- Çeşitli tıp test sonuçları
- Sosiyal mediya postları
- Doktora ziyaret kayıtları
- Haber parçaları
- …

Saklı ilişki

İş kararı

ORTAK DESEN …

Nasıl

- İstatistik
- Olasılık
- Diferensiyal
- Linear cebir
- Algoritmalar
- Programlama
- …

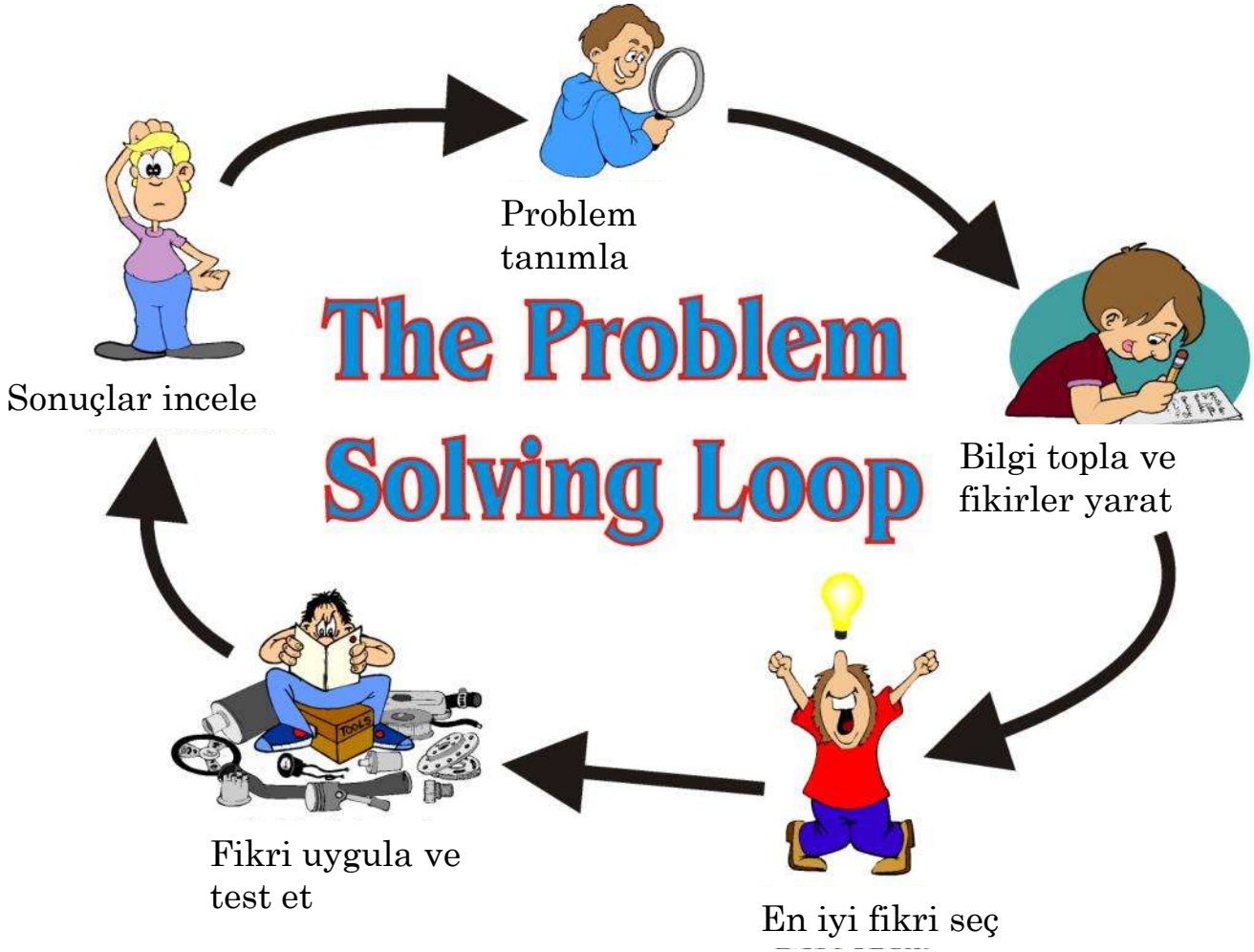
28

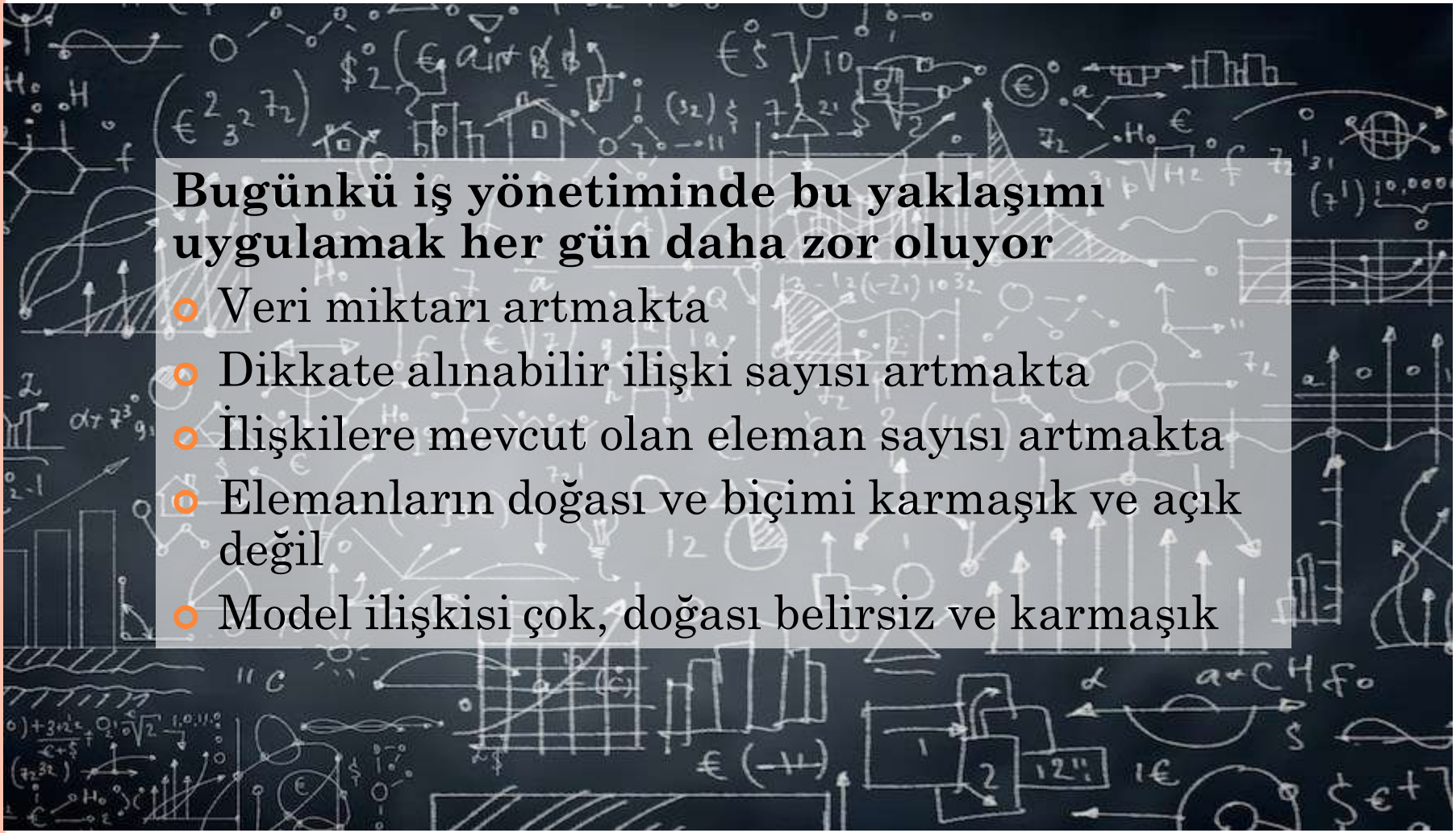
Büyük veri felsefesi

İKI PROBLEM ÇÖZME YAKLAŞıMı

- Alıştığımız problem çözme
  - Problemin mantığını anlamak
  - Mevcut kavramlar arasındaki ilişkiler kesinleştirmek
  - Mevcut olan faktörlerin muhtemel etkileri belirtmek
  - Belirli müdahale olduğunda problemde değişiklikleri tahmin etmek

ESKI PROBLEM ÇÖZME YAKLAŞıMı (NORMAL)

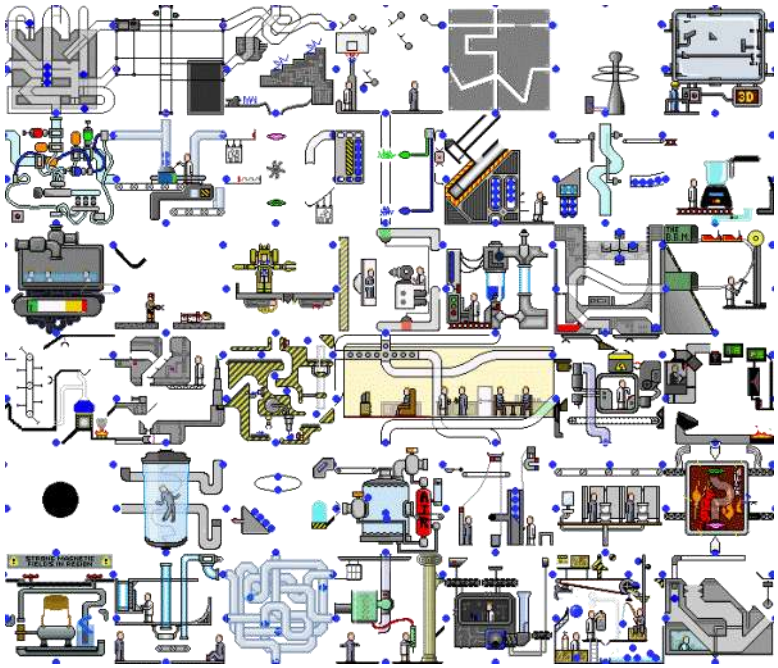




Bugünkü iş yönetiminde bu yaklaşımı uygulamak her gün daha zor oluyor

- Veri miktarı artmakta
- Dikkate alınabilir ilişki sayısı artmakta
- İlişkilere mevcut olan eleman sayısı artmakta
- Elemanların doğası ve biçimi karmaşık ve açık değil
- Model ilişkisi çok, doğası belirsiz ve karmaşık

Problem çözme



Alternatif (modern) yaklaşım

- Yeni problem çözmeye yaklaşım
  - Belirli problemi için genel bir parametreli model oluştur
  - Bu modelin parametreleri mevcut verilerden tahmin et
  - Müdahalenin sonucunu bu modelden tahmin et

ALTERNATIF PROBLEM ÇÖZME YAKLAŞıMı (YENI)

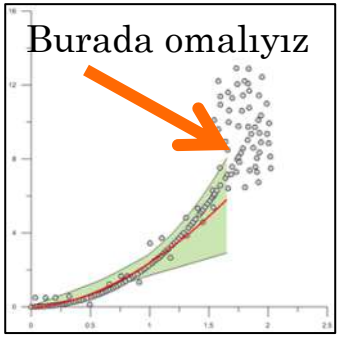
Gerçek problem ve dahil olacak değişkenler tanımla

a, c, X, Y, ξ, μ, …

Bir genel parametreli model seç

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots$$

Parametreleri tahmin et ve model kullan

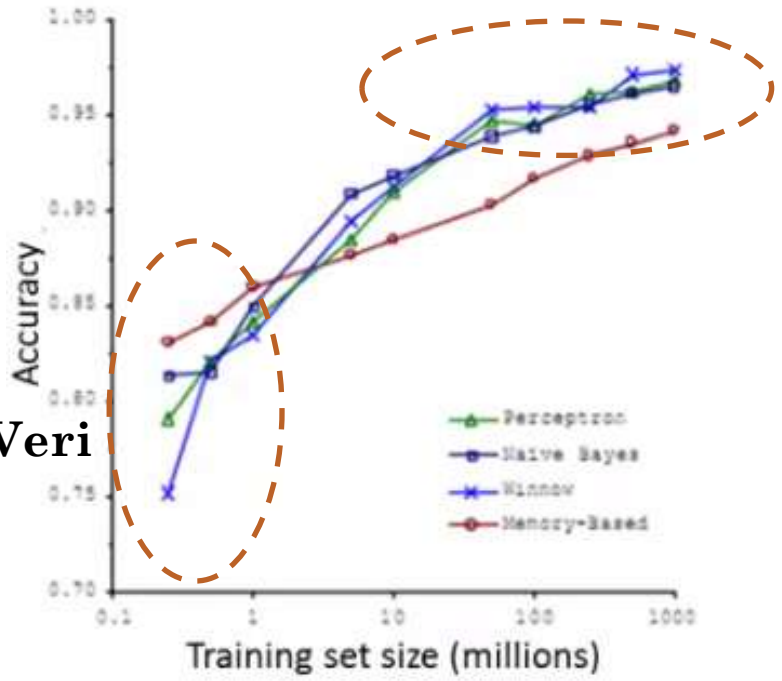


BÜYÜK VERI "TEOREMI"

- Veri miktarı büyük olduğunda, basit modellerin performansı ve çok karmaşık model performansı arasında büyük fark yok

Büyük Veri

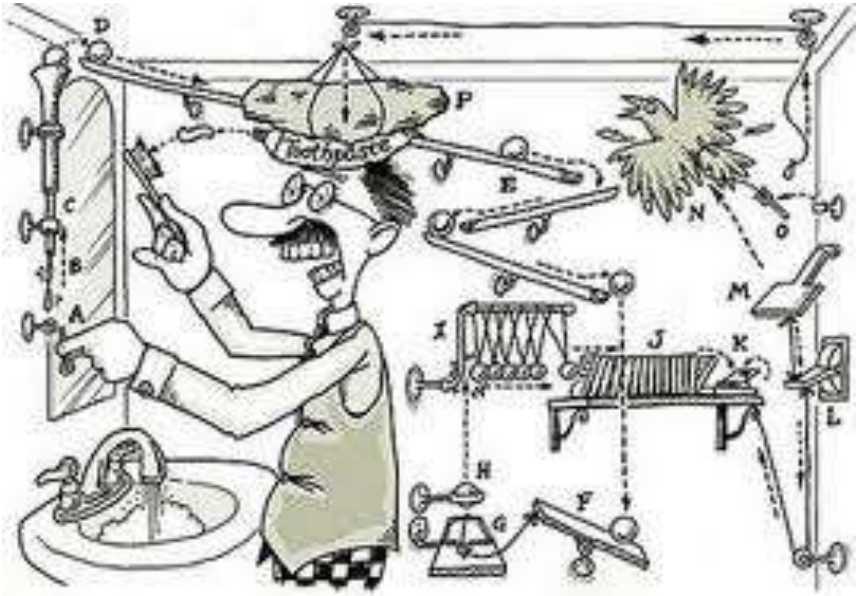
Küçük Veri

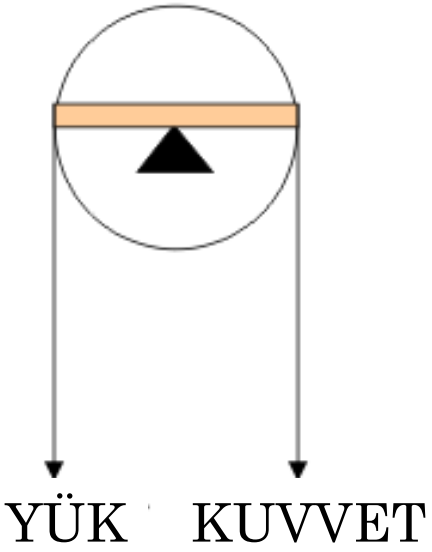


Banko and Brill [2001] Scaling to Very Very Large Corpora for Natural Language Disambiguation

BÜYÜK VERI "TEOREMI"

Büyük veri rejiminde genel basit modeller, probleme özel tasarımlanmış ve çok karmaşık olan modeller'den genelde daha başarılıdır



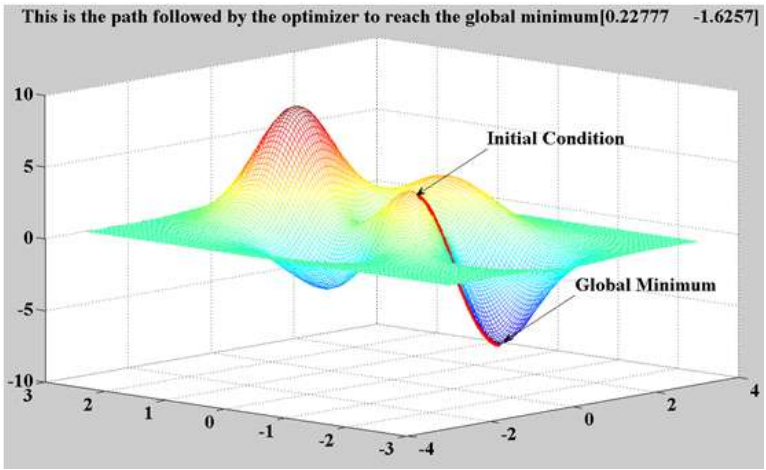


Büyük Veri teoreminden uygulamalarına

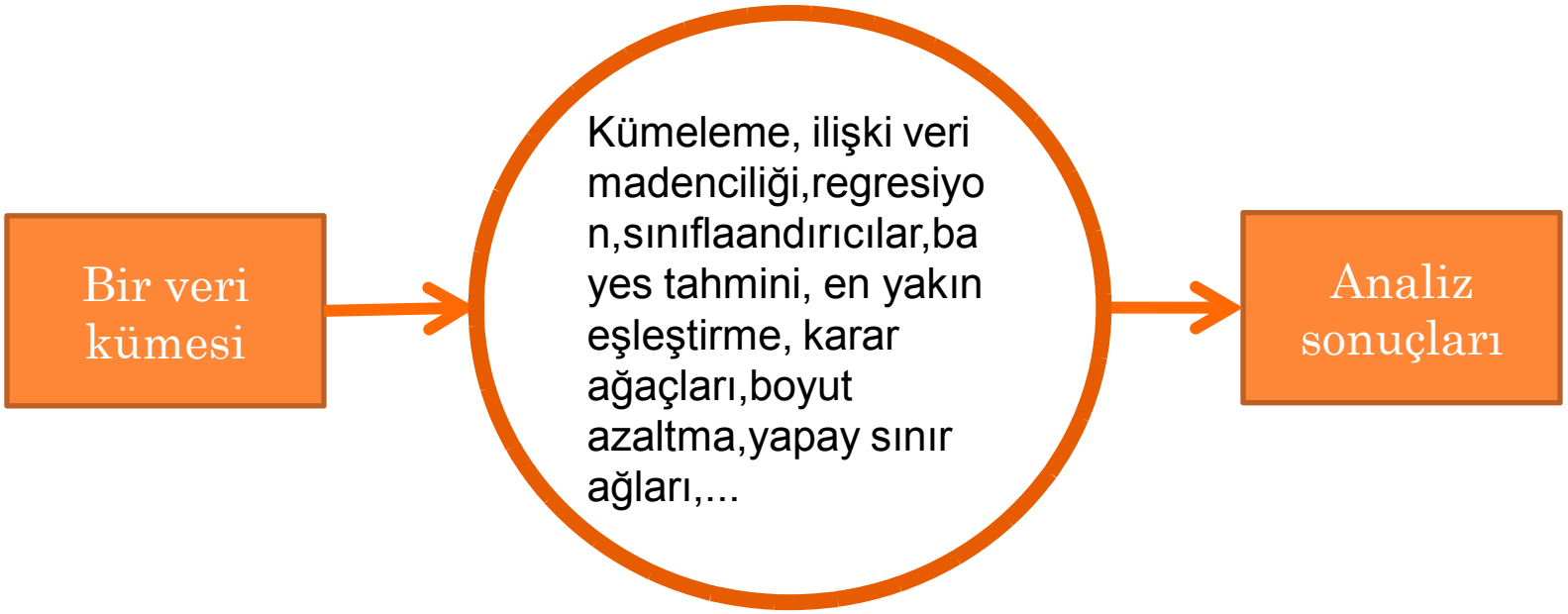
- Teknolojinin gelişimiyle mevcut olmaya başlayan büyük miktarda olan iş yönetimi ile alakalı verilerden faydalanmak için ...
- İş yönetimi ile ilgili soruların modellenmesi için genel makine öğrenme aletleri kullanılır ve ...
- Bu modellerin sonuçları iş yönetim kararları vermek için kullanılır

Büyük Veri için makine öğrenme aletleri

- Kümeleme
- Regresiyon
- İlişkisel kural madenciliği
- Sınıflandırma
- Bayes tahmini
- En yakın eşleştirme
- Karar ağaçları
- Boyut azaltma
- Yapay sınır ağları



Büyük Veri Felsefesi

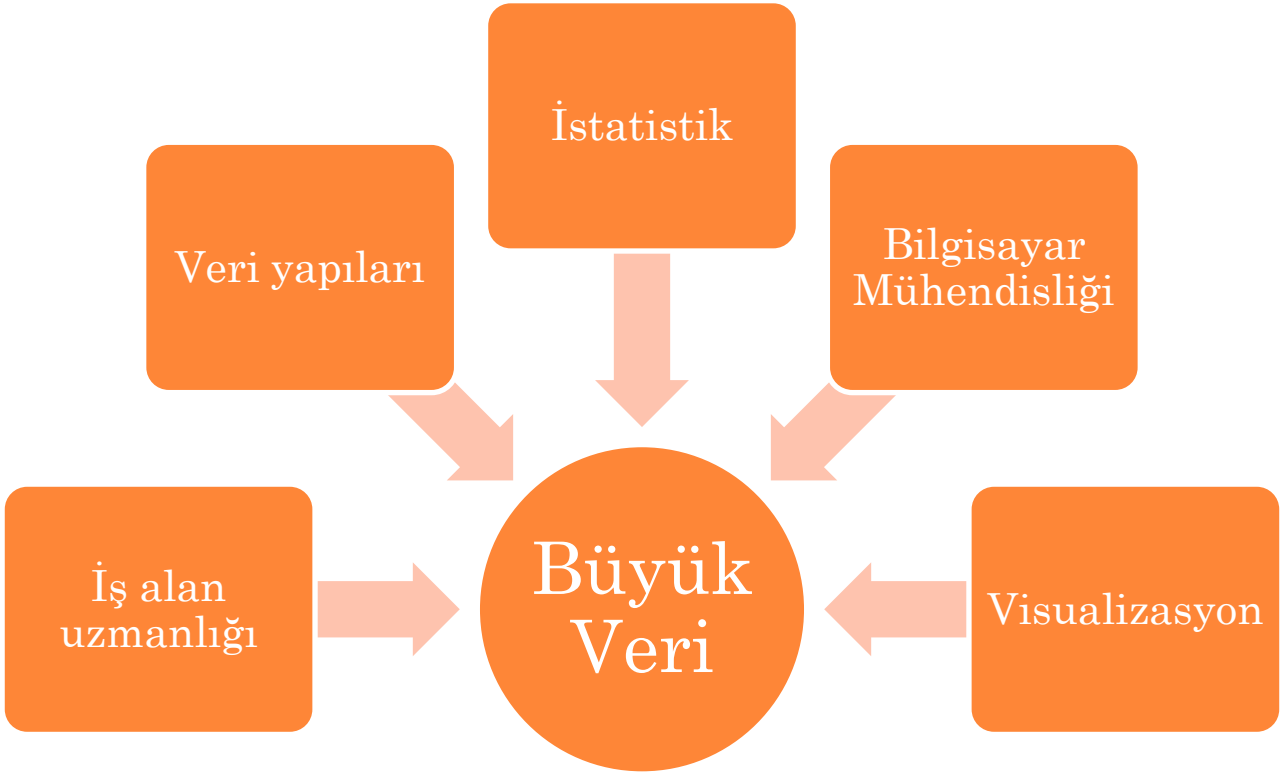


Büyük veri içerisi

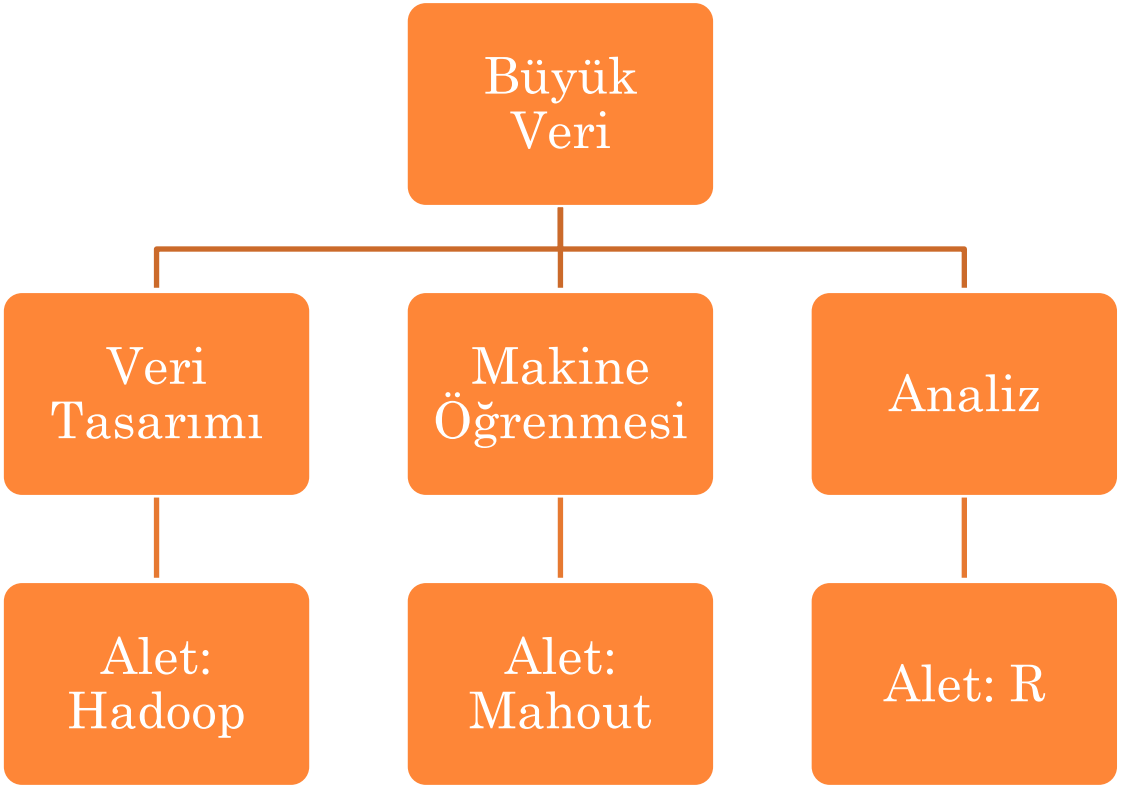
BÜYÜK VERI'NIN SORUNLARı

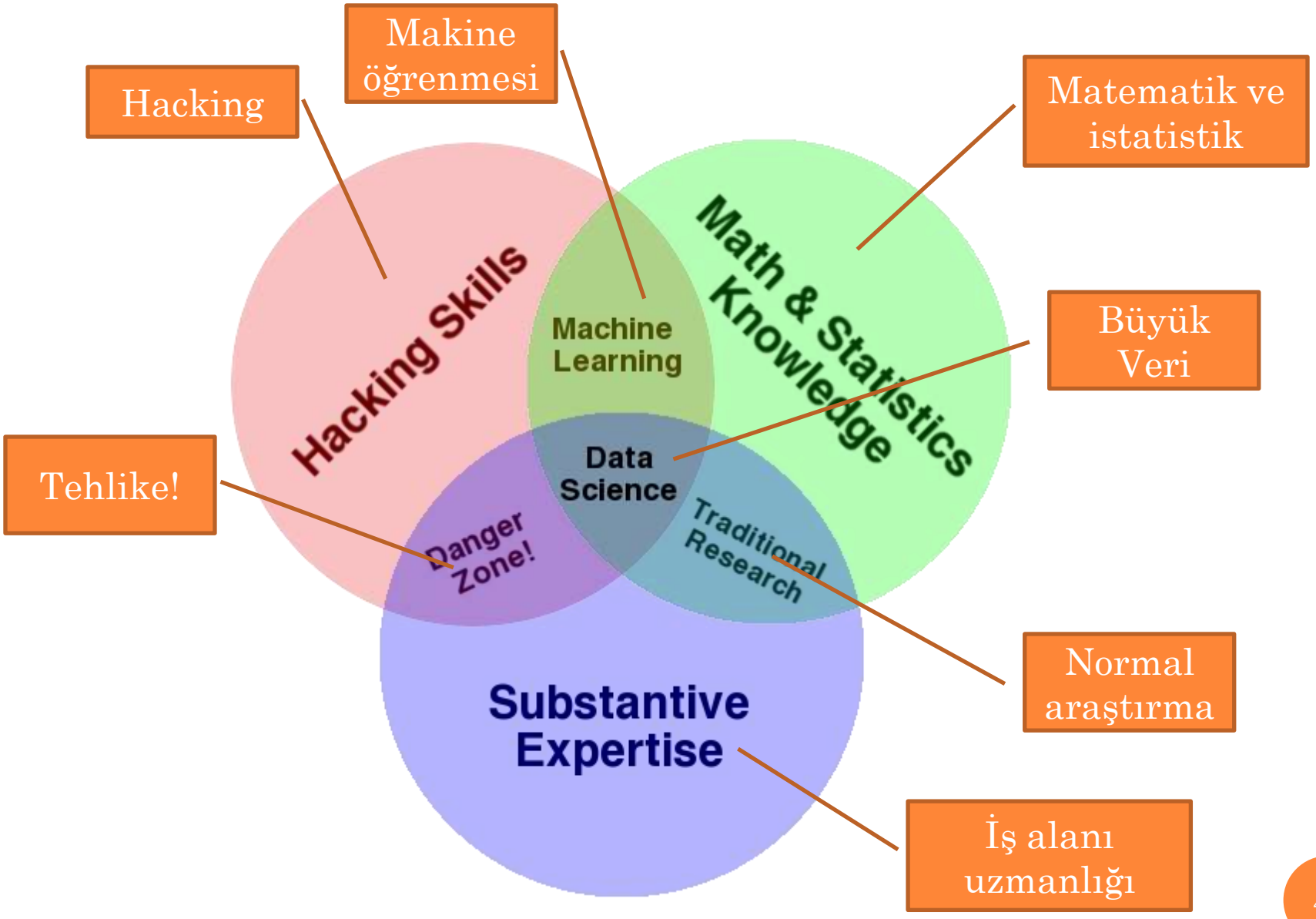
- Veri toplama
  - hangi veriler kullanılacak?
- Veri depolama
  - TerraByte/PettaByte veri nerede kaydedilecek?
- Veri transferleri
  - TB/PB veri nasıl transfer edilecek?
- Veri sorgulanması
  - TB/PB veri setleri nasıl sorgulanacak?
- Veri analizi ve sonuç çıkartma
  - Karmaşık ilişkiler nasıl ortaya çıkartılacak?
- Sonuç bilgilendirme
  - Karmaşık sonuçlar nasıl incelenip bilgilendirecek?

BÜYÜK VERI BILEŞENLERI



BÜYÜK VERI AŞAMALARı





http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

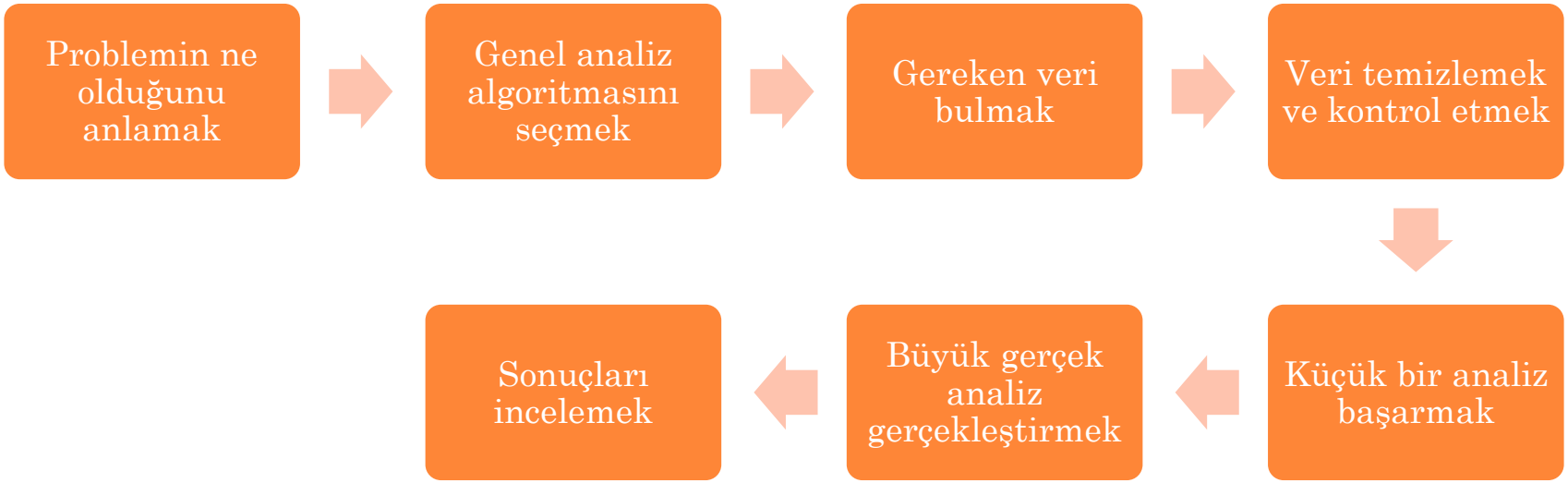
BÜYÜK VERI'NIN IKI TARAFı

- Veri analizi/makine öğrenmesi/istatistik
  - Verilerden kavrama yaratabilmek
- Büyük ölçekli bilgi işlem/bilgisayar programlama
  - Verilerin büyük hacimleri işletebilmek

- Son yıllarda Google Prediction API, Microsoft Azura ML, Amazon ML, BigML gibi Big Veri bulut çözümleri bu ikisi işi oldukça kolaylaştırmıştı

Detaylı örnekler

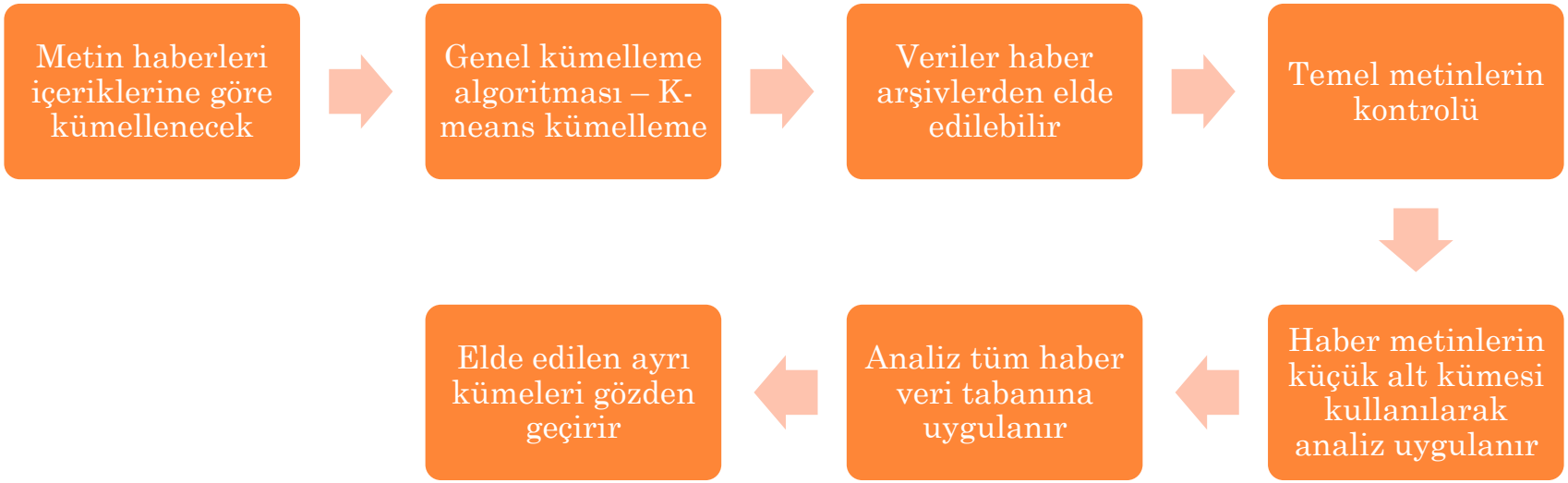
BÜYÜK VERI UYGULAMALARı



Media kullanim örneği

- Problem tanımı:
  - Bir büyük media şirketi, 1980 den itibaren belirli konuda popüler haberleri incelemek istiyor.
  - Analizi kolaylaştırmak için, tüm haberleri benzer kümelerine ayırdırmak istiyor

MEDIA KULLANIM ÖRNEĞI



ALIŞVERIŞ KULLANıM ÖRNEĞI

- Problem tanımı
  - Bir magaza müşterilerinin davranışı ve satın alınları anlamak istiyor. Bu bilgi müşterilerine daha iyi hizmet sağlamak için kullanılacak.
  - Analiz magazaya, müşterilerin bir ürünle beraber başka ürünlerin satın aldığını ortaya çıkabilir; bu halde söz konusu ürüne uygulanacak kampanya ilişkili ürünlerin satışı artırıp magazanın geliri etkileyebilir

ALIŞVERIŞ KULLANıM ÖRNEĞI

Tıp kullanım örneği

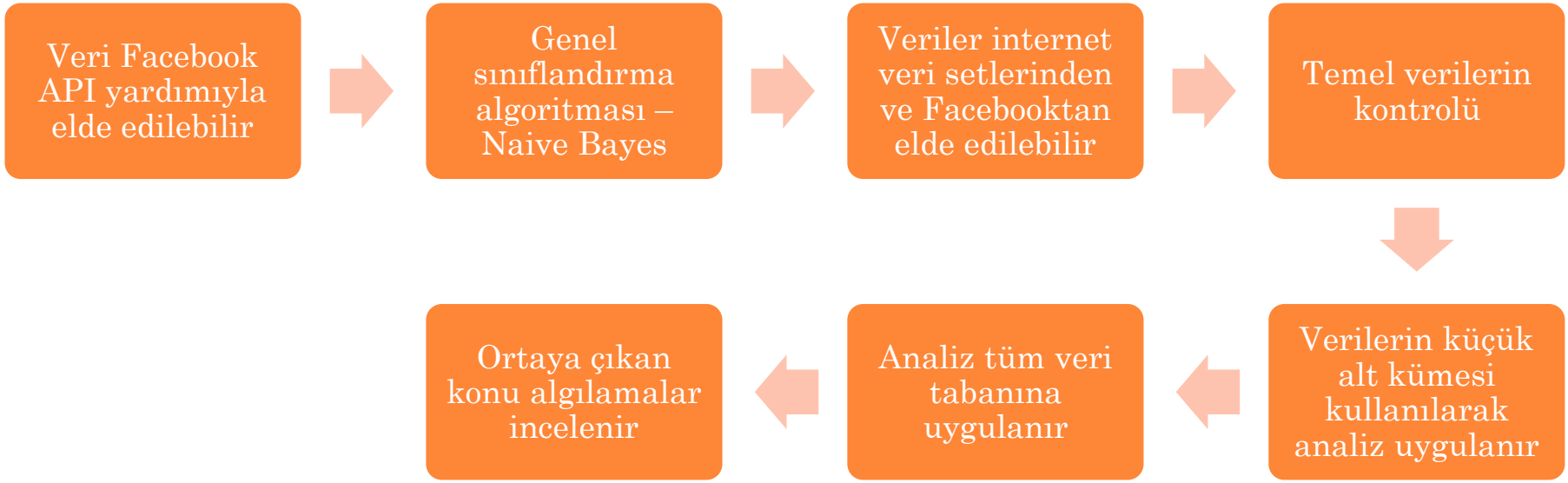
- Problem tanımı
  - Bir hastane hastaların demografik özellikleri ile beraber çeşitli test sonuçları ve hastalığın başlangıcı arasında muhtemelen ilişkiyi ortaya çıkartmak istiyor
  - Bunları kullanılarak tahmini müdahale stratejileri tasarımlamak istiyor

Tıp kullanım örneği

SOSIYAL MEDIA KULLANıM ÖRNEĞI

- Problem tanımı
  - Sosiyal media araştırma şirketi Facebook'taki "sıcak" konuları analiz ederek onları aşağıdaki kategorilerine sınıflandırmak istiyor:
    - Giyisi (ayakabı, kıyafet, saat, takı, ...)
    - Sanat (Kitaplar, film, DVD, VCD, müsik)
    - Photokameralar
    - Olaylar (seyahat, konser, film,...)
    - Sağlık (güzellik, spa,...)
    - Ev (mutfak, mobilya, bahçe,...)
    - Teknoloji (bilgisayar, laptop, tablet, smartphone,...)

SOSIYAL MEDIA KULLANıM ÖRNEĞI

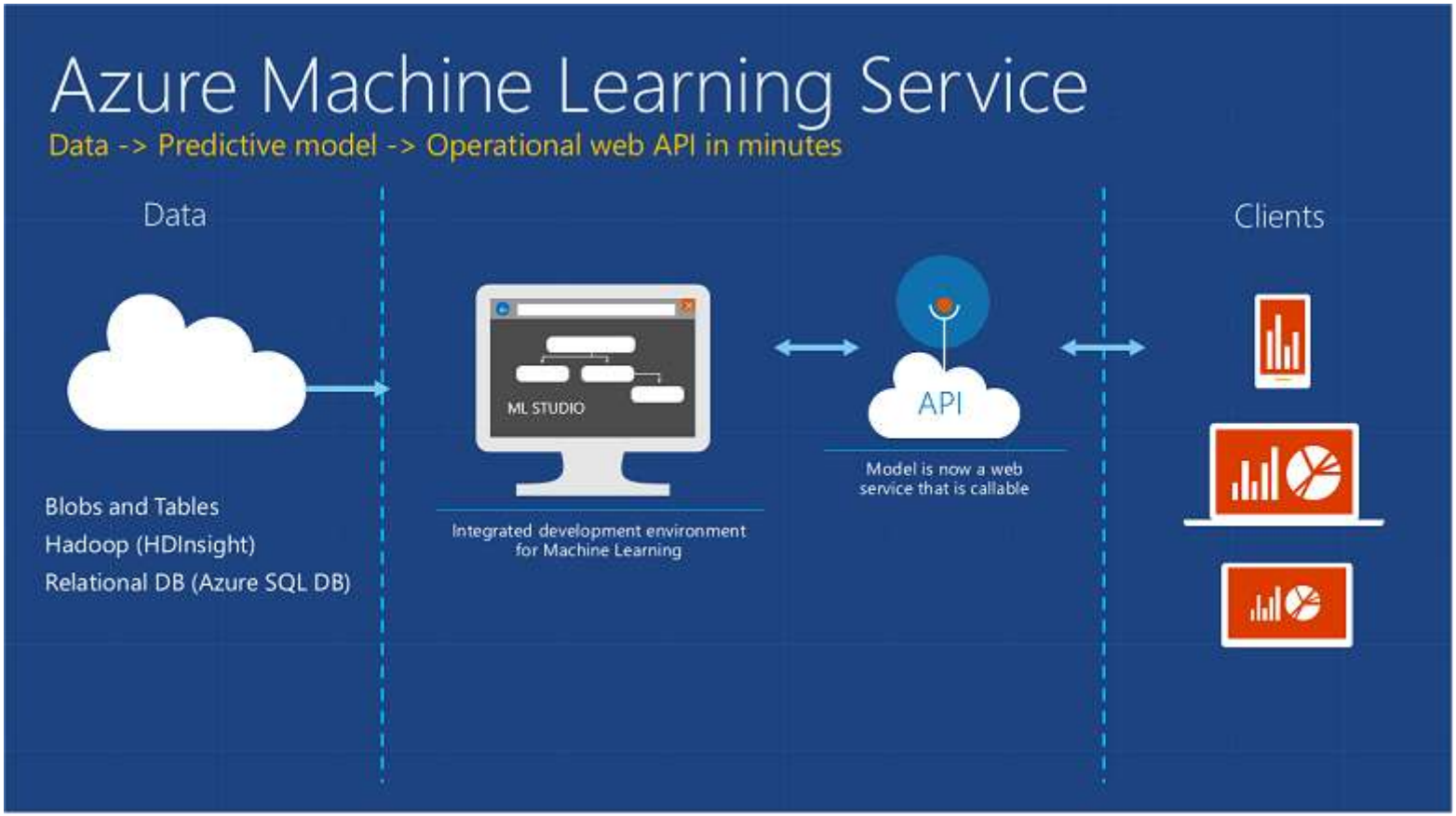


Aletler

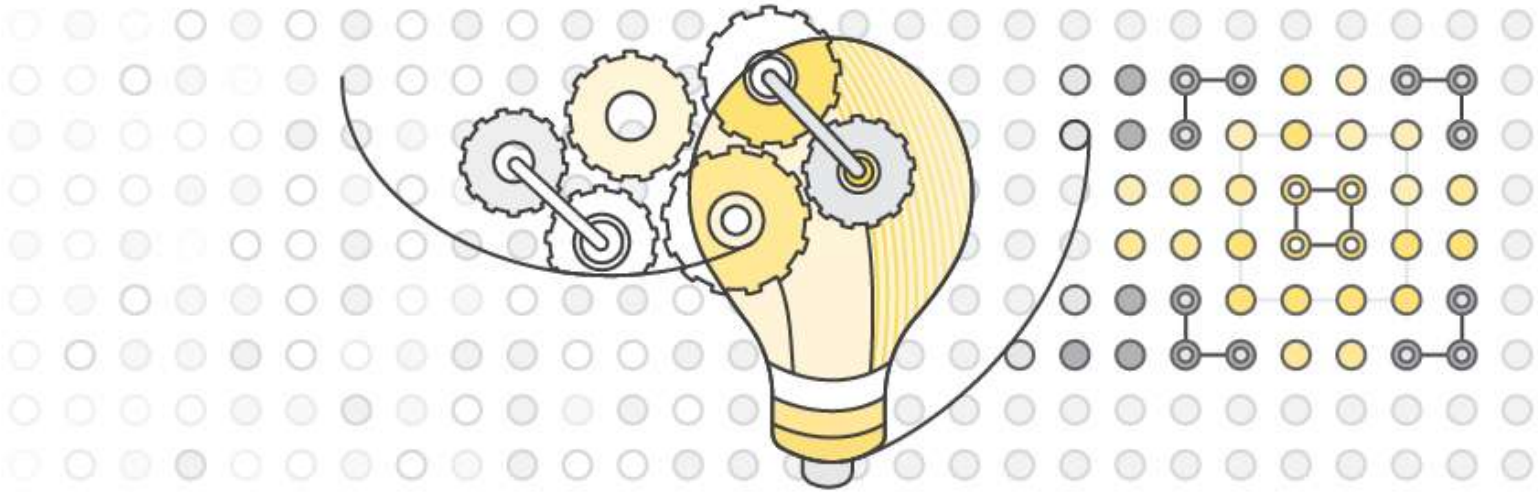
Google Prediction API

Demo

Azura ML


Azure Machine Learning Service
Data -> Predictive model -> Operational web API in minutes
Data
Blobs and Tables
Hadoop (HDInsight)
Relational DB (Azure SQL DB)
ML STUDIO
Integrated development environment for Machine Learning
API
Model is now a web service that is callable
Clients

Amazon ML



Real Statistics Excel

- “İleri istatistik MS Excel’inize…”

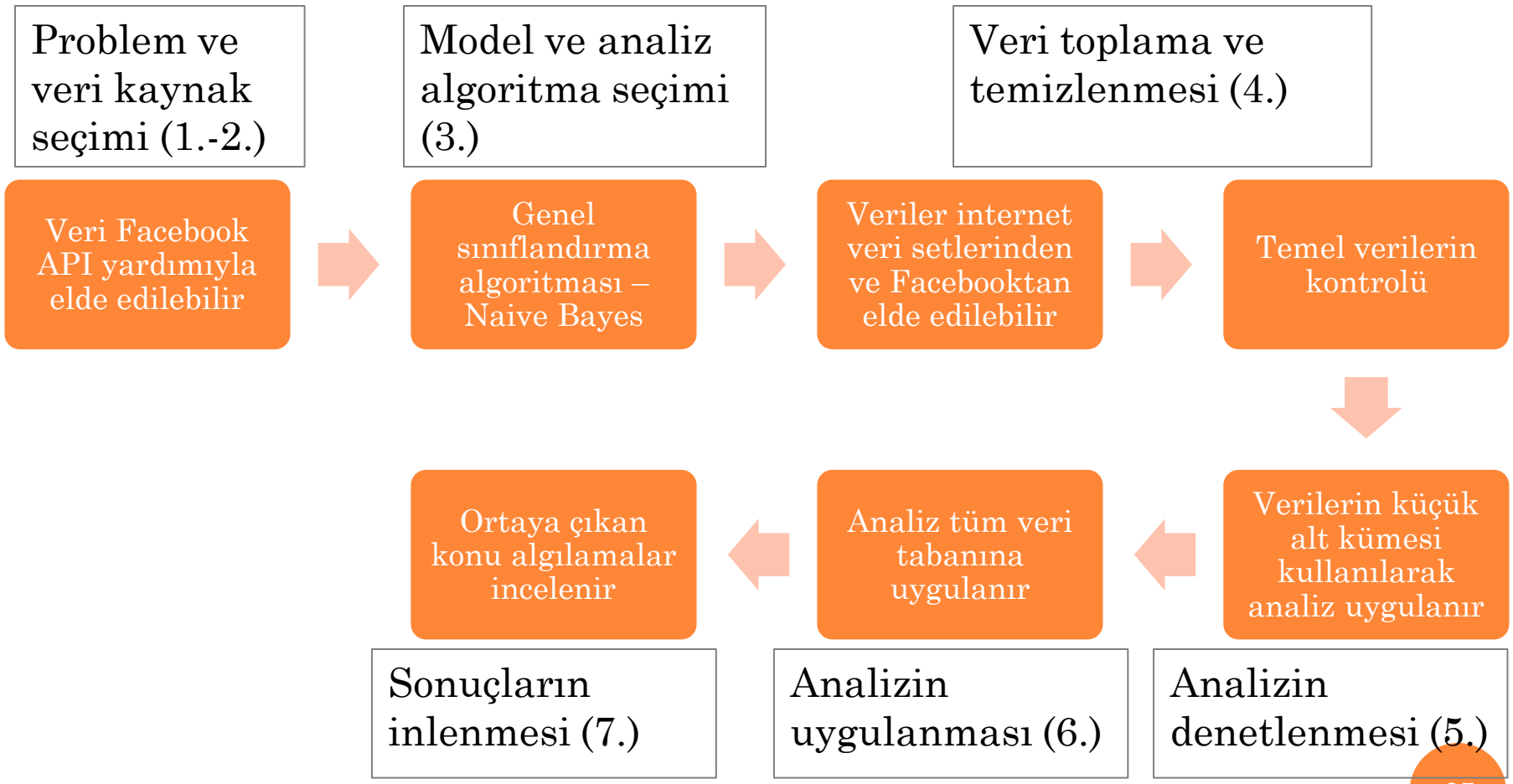
Daha …

- Open Source kaynaklar: Hadoop, Apache Mahout, MapReduce, TensorFlow Google, H2O on Hadoop, Spark MLib, Weka-Java, SHOGUN, WSO2 Data Analytics, DARWIN, scikit, FAIR Facebook, vb.
- Daha ileri ve düşük seviyeli ARGE için

BÜYÜK VERI UYGULAMA TARIFI

1. Mevcut iş sorunları ve veri kaynakları inceleyip özetleyin
2. Uygun bir sorun modelleme problemi olarak tanımlayın ve mevcut veri kaynaklarından ilgili olabilecek kaynakları belirtin
3. Sorunu ve veri bağlayacak bir genel makine öğrenme modeli ve analiz algoritmasını seçin
4. Gereken veri toplayın ve kalite açısından inceleyin, gerekirse --- veri temizleyin
5. İstediğiniz analiz işleminin adımları seçip, küçük bir denetleme veri kümesi üzerinde uygulayın
6. Analizinizi tüm verilerinize uygulayın
7. Analiz sonuçları uygun görsel şekilde gösterin ve orijinal soruna dair sonuçlar çıkartın

SOSIYAL MEDIA ÖRNEĞI



BÜYÜK VERI UYGULAMA TARIFI

1. Mevcut iş sorunları ve veri kaynakları inceleyip özetleyin
2. Uygun bir sorun modelleme problemi olarak tanımlayın ve mevcut veri kaynaklarından ilgili olabilecek kaynakları belirtin
3. Sorunu ve veri bağlayacak bir genel makine öğrenme modeli ve analiz algoritmasını seçin
4. Gereken veri toplayın ve kalite açısından inceleyin, gerekirse --- veri temizleyin
5. İstediğiniz analiz işleminin adımları seçip, küçük bir denetleme veri kümesi üzerinde uygulayın
6. Analizinizi tüm verilerinize uygulayın
7. Analiz sonuçları uygun görsel şekilde gösterin ve orijinal soruna dair sonuçlar çıkartın

ÖDEV/DÖNEM PROJE ÖN-ÇALIŞMASI

- **Öbür derse kadar lütfen, iş alanınızdan bir problemi seçin**
  - Mesleki alanınızdan
  - Kolay ulaşılabilir veri (elektronik halde)
  - İlginç bir soru
- **10 dakika geçmeyen, derste sunulmak üzere bir tanıtım sunuşu hazırlayın**

ÖDEV/DÖNEM PROJE ÖN-ÇALIŞMASI

- **Dönem projesi**
  - Problemin tanımı
  - Modelin tanımı
  - Veri elde edilmesi
  - Bulut araçları yardımıyla analiz edilmesi
  - Sonuçların incelenmesi