Büyük Veri İçin İstatistiksel Öğrenme (Statistical Learning for Big Data)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017." ve "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitapları kullanılarak hazırlanmıştır.

Konular

- Büyük Veri
- Akış Verisi
- Akış Verisi Kaynakları
- Büyük Veri Analitiği
- Büyük Veri Uygulamaları

Büyük Veri

- Büyük veri kendine özgü özelliklere sahip olan ve genel olarak **yüksek hacimlerdeki veriler için kullanılan bir terimdir.**
- **Dünyadaki verilerin %90'ı son 3-4 yılda oluşmuştur.**
- Büyük veri **çok farklı kaynaklardan elde edilebilir.**

- **Büyük veri analitiği yöntemleri**, farklı kaynaklardan elde edilen düzenli veya düzensiz **verileri anlamlı ve işlenebilir hale dönüştürür.**

3

Büyük Veri

- **Sosyal medya paylaşımları, blog yazıları, fotoğraf, müzik, video arşivleri, müşteri veya çalışan bilgileri, IoT verileri** ve kullanıcı hareketlerinin kaydedildiği **log dosyaları** gibi çeşitli kaynaklardan elde edilen veriler kullanılır.
- Sosyal medyadaki veri miktarı **petabyte, exabyte** veya **zettabyte** seviyelerine çıkabilmektedir.
- **Geçmişte** bilgi kirliliği olarak görülen **bu veriler gereksiz ve faydasız olarak görülmekteydi.**
- İlişkisel veritabanı sistemlerinde (Relational Database Management Systems - RDMS) oluşturulan sorgular sonucunda alınan kararlar, yanlış ve eksik bilgi nedeniyle hatalı olabiliyordu.

4

Büyük Veri

- Büyük veri terimi ilk ortaya çıktığından itibaren farklı sayıdaki özellikleriyle ifade edilmiştir.
- Bu özellikler 3V, 5V, 7V, 10V ve hatta 42V olarak ifade edilmiştir.
- **Yaygın kullanılan 10V:**
  - **Volume**
  - **Velocity**
  - **Variety**
  - **Variability**
  - **Veracity**
  - **Validity**
  - **Vulnerability**
  - **Volatility**
  - **Visualization**
  - **Value**

5

Büyük Veri

Volume

- Büyük verinin **en çok bilinen karakteristiğidir.**
- **Son birkaç yıl içerisinde** önceki **tüm zamanların yaklaşık 10 katı veri oluşturulmuştur.**
- **YouTube'a** her **bir dakikada 300 saatlik video** yüklenmektedir.
- 2016 yılında **1.1 trilyon fotoğraf** çekildiği tahmin edilmektedir.
- 2016 yıllında **cep telefonu** veri trafiği **6.2 exabyte** tahmin edilmektedir.
- Twitter kullanıcıları her **bir dakikada 277.000 tweet** atmaktadır.
- Apple kullanıcıları her **bir dakikada 48.000 uygulama** indirmektedir.
- Facebook kullanıcıları her **bir dakikada 2.460.000 içerik** paylaşmaktadır.
- Her **bir dakikada 204.000.000 e-posta mesajı** gönderilmektedir.
- Google her **bir dakikada 2.400.000 arama sorgusu** almaktadır.

6

Büyük Veri

Velocity

- Velocity **verinin üretilme**, **tüketilme**, **oluşturulma** ve **güncellenme** hızıdır.
- **Facebook** günde **600 terabyte verinin geldiğini ifade etmektedir.**
- **Google** her **saniyede 40.000 sorguya cevap** oluşturduğunu ifade etmektedir. Günde yaklaşık 3,5 milyar sorguya cevap verdiği söylenebilir.

Variety

- Büyük veride **yapılandırılmış**, **yarı yapılandırılmış** ve **çoğunlukla yapılandırılmamış veri bulunur.**
- Bunlar; **ses**, **video**, **görüntü**, **sosyal medya güncellemeleri**, **log dosyaları**, **click verileri**, **makine** ve **sensör verileri** vb. olabilir.

7

Büyük Veri

Variability

- Büyük veride **tutarsızlıklara neden olan bazı farklı verilerde olabilir.**
- **Bu verilerin anomaly** veya **outlier algılama yöntemleri ile bulunup** yapılan analizlerin daha anlamlı hale getirilmesi gereklidir.

Veracity

- **Büyük veride boyut**, **çeşitlilik** ve **tutarsızlık artarken** verinin **güvenilirliği** ve **doğruluğu düşer.**
- **Veracity veri kaynaklarının güvenilirliğini ifade eder.**
- Verinin **kim tarafından oluşturulduğu**, **hangi metodoloji ile toplandığı**, **aynı türdeki kaynaklardan mı toplandığı**, **veriyi toplayanın özetleme yapıp yapmadığı**, **veri başka birisi tarafından değiştirildi mi** gibi sorulara cevap aranır.

8

Büyük Veri

Validity

- Verinin **nasıl doğrulandığı** ve **geçerliliğinin nasıl test edildiğiyle ilgilenir.**
- Verinin analiz işleminden önce doğrulanması gereklidir.

Vulnerability

- **Büyük veri yeni güvenlik konularını da beraberinde getirir.**
- Verinin hack'lenmemesi gereklidir.
- **Kaynağından elde edildiği gibi** herhangi bir bozulmaya veya güvenlik saldırısı sonucu **değişmeye uğramaması gereklidir.**

Volatility

- **Verinin ne kadar eski olduğu**, hala **güncel olup olmadığı**, **kullanılabilir olup olmadığı** ile ilgilenir.

9

Büyük Veri

Visualization

- Büyük **verinin görselleştirilmesi analizini kolaylaştırır.**
- Görselleştirmeyle ilgili hafıza gibi teknik kısıtlar halen bulunmaktadır.
- **Klasik grafik araçları ve yöntemleriyle** büyük verideki milyarlarca noktanın **görselleştirilmesi mümkün değildir.**
- Büyük veri için **kümeleme**, **ağaç haritaları**, **dairesel ağ diyagramları** gibi görselleştirme yöntemlerinin kullanılması gereklidir.

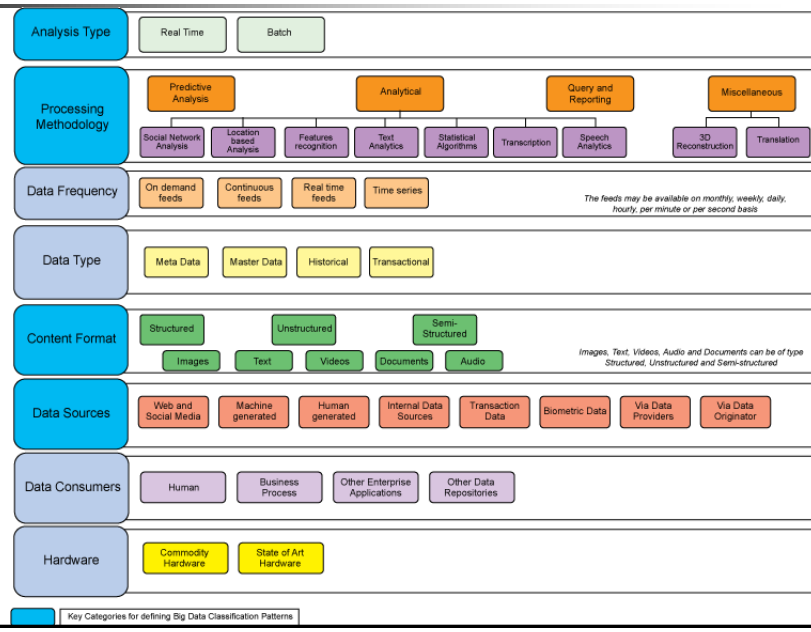
Value

- Büyük **veriden anlamlı ve değerli bilgiyi çıkarmadıkça diğer bütün karakteristikleri anlamsızdır.**
- Büyük veriden anlamlı ve değerli verinin elde edilmesi için veri madenciliği yöntemleri gibi karmaşık süreçlerin **büyük veriye özgü bir şekilde uyarlanıp kullanılması gereklidir.**

10

Büyük Veri

Büyük veri sınıflandırma

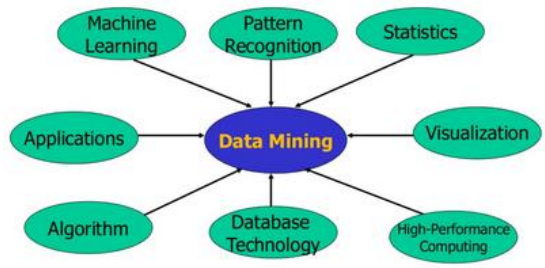


11

Büyük Veri

Büyük verinin kullanım alanları

- Büyük veri, **veri madenciliği, makine öğrenmesi, örüntü tanıma, istatistik, görselleştirme, yüksek performanslı hesaplama, veritabanı teknolojisi, algoritma, uygulamalar** gibi disiplinlerde kullanılmaktadır.
- Veri madenciliği yöntemleri birçok alanda başarılı bir şekilde kullanılmaktadır.



12

Büyük Veri

İşletme

- Özellikle büyük ölçekli işletmeler **müşteri analizi**, **müşteriye özelleştirilmiş tavsiye**, **reklam veya öneri oluşturma**, **ürün dağıtımı** ve **lojistik optimizasyonu** gibi çok sayıdaki alanda büyük veri analiz yöntemlerini kullanmaktadır.

Perakende Satış

- **Personel gelir optimizasyonu**, **müşteri davranış analizi**, **müşteri ilişkileri analizi**, **ürün çeşitliliği**, **kampanya yönetimi** ve **fiyat optimizasyonu** gibi uygulamalarda yaygın bir şekilde büyük veri analiz yöntemleri kullanılmaktadır.

13

Büyük Veri

Kamu

- Verilere **kolay ve güvenli erişebilirliği sağlama**, **gizlilik** ve **şeffaflık** oluşturma, **uygun ürün** ve **hizmetlerin sunumu**, **risk** ve **sahtekarlığı azaltmaya yönelik** alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Teknoloji

- **Gerçek zamanlı analiz** ve **işlem (menü) özelleştirme**, **işlem süresini azaltma**, **riskleri azaltma** konusunda otomatik sistemler ile karar verme gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

14

Büyük Veri

Eğitim

- **Eğitimde öğrenci analizi**, **ders planlaması** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Kişisel Konum Verileri

- **Konum tabanlı reklam**, **akıllı yönlendirme**, **acil müdahale** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

Sağlık

- **Hastalık tespiti**, **hasta izlenmesi**, **kişisel DNA analizi** gibi alanlarda büyük veri analiz yöntemleri kullanılmaktadır.

15

Konular

- Büyük Veri
- Akış Verisi
- Akış Verisi Kaynakları
- Büyük Veri Analitiği
- Büyük Veri Uygulamaları

16

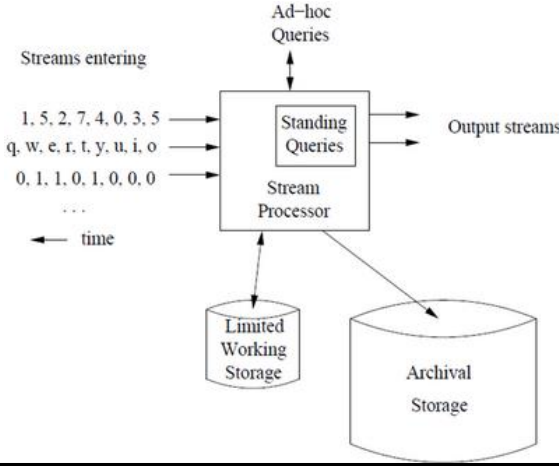
Akış Verisi

- **Akış verisi geldiği anda işlem yapılmazsa** (depolama, data process vs.) **kalıcı şekilde kaybedilebilir.**
- Veriyi işleme hızından daha hızlı veri gelmesi durumunda da kaybedilebilir.
- **Akış verisinde işlem yapan algoritmalar akış verisini özetler.**
- Akış madenciliği algoritmaları, **faydalı örnekleri seçer** ve **istenmeyen örnekleri filtreler.**
- Başka bir özetleme yaklaşımında ise, **sabit boyutlu bir pencere içerisindeki elemanlarla** (belirli bir süre için geçmiş veri) **özetleme yapılmaktadır.**
- Akış verisinin özetlenmesiyle birlikte **daha küçük alanda saklanması da sağlanmış olur.**

17

Akış Verisi

- Akış işlemcisi bir tür veri yönetim sistemi olarak görülebilir.
- **Sisteme çok sayıda farklı stream'den veri gelebilir.**
- **Veri türleri**, **veri oranları** ve **veri gelme aralıklarının dağılımları** farklı olabilir.



18

Akış Verisi

- Stream'lerden gelen **veriler büyük bir depolama biriminde (archival storage) saklanabilir.**
- Bu depolama birimindeki veri üzerinde **uzun zaman alan işlemlerin ardından sorgulama yapılabilir.**
- **Working storage** depolama birimi ise **akış verisinin özetini** veya **bir parçasını saklar.**
- Working storage birimi, işlem hızı gereksinimine göre **disk** veya **ana hafıza olabilir.**
- Working storage birimi **sınırlı kapasiteye sahiptir** ve **akış verisinin tamamını saklayamaz.**

19

Konular

- Büyük Veri
- Akış Verisi
- Akış Verisi Kaynakları
- Büyük Veri Analitiği
- Büyük Veri Uygulamaları

20

Akış Verisi Kaynakları

Sensor data

- **Bir okyanus yüzeyindeki ısı sensörü** her saat ölçtüğü ısı değerini reel sayı olarak bir istasyona göndersin.
  - Bu durumda **veri oranı çok düşük** olduğundan günümüz teknolojisinde tüm veri ana hafızada tutulabilir.
- GPS birimindeki sensör **yüzeydeki yükseklik değişimini** ölçüp bir istasyona göndersin.
  - Bu durumda **veri oranı yüksektir** ve ancak ana hafızada veya ayrı bir diskte tutulabilir.
- Bir **okyanusun tüm davranışını ölçmek istersek**, milyonlarca sensör kullanılır ve günlük birkaç terabyte veri alınabilir.

21

Akış Verisi Kaynakları

Image data

- Uydulardan sürekli dünyaya ilişkin **görüntüler alınıp yeryüzündeki istasyonlara gönderilir.**
- Bu görüntü verilerinin boyutları **günlük birkaç terabyte düzeyinde olabilir.**
- Şehirlerdeki **güvenlik kameraları uyduya göre düşük çözünürlüktedir**, ancak her birisi akış verisi oluşturur.
- Londra'da **6 milyon kamera** olduğu belirtilmektedir ve her birisi akış verisi oluşturur.

22

Akış Verisi Kaynakları

İnternet ve Web trafiği

- İnternet **anahtarlama düğümleri** (router) IP paketlerinden oluşan **stream'leri alır ve çıkış portlarına yönlendirme yapar.**
- Anahtarlama elemanlarının görevi verileri sorgulamak veya tutmak değildir.
- Ancak, **günümüzde anahtarlama elemanlarının kapasitesinin artırılmasına** yönelik eğilim (DOS ataklarının algılanması, tıkanıklık denetimi yapılması) vardır.
- **Web siteleri** her gün **milyonlarca sorgu almaktadır** (Google her gün yüzlerce milyon arama sorgusu almaktadır, Yahoo milyarlarca click almaktadır.).
- Bu tür **verilerden faydalı bilgiler elde edilebilir** (sorgulardaki ani yükselme, click sayısındaki ani yükselme veya düşme).

23

Konular

- Büyük Veri
- Akış Verisi
- Akış Verisi Kaynakları
- Büyük Veri Analitiği
- Büyük Veri Uygulamaları

24

Büyük Veri Analitiği

- Büyük veri analitiği, büyük ve çeşitli veri setleri üzerinde işlem yaparak **gizli örüntüleri çıkarma**, **bilinmeyen ilişkileri keşfetme sürecidir.**
- Kullanılan yöntemlerle **elde edilen bilgi**; **firmalara**, **kurumlara** veya **ticari girişimlere** yönelik **önemli bilgiler sağlamaktadır.**
- Büyük veri analitiği uygulamaları **veri bilimcilere modelleri tahmin etme**, **istatistikçilere** ve diğer analiz alanında çalışan profesyonellere **büyüyen verileri kolay analiz etme** yeteneği kazandırır.
- **Büyük veri analitiği** klasik yöntemlerle yönetilmesi çok zor olan **çok büyük**, **yapılandırılmamış** ve çok hızlı değişen veriyle uğraşır ve **anlamlı örüntüler elde eder.**
- Büyük veri analitiği yöntemleri **veriyi saklamak**, **veriyi elde etmek** ve **analiz etmek** için gelişmiş **teknolojiyi kullanır.**

25

Konular

- Büyük Veri
- Akış Verisi
- Akış Verisi Kaynakları
- Büyük Veri Analitiği
- Büyük Veri Uygulamaları

26

Büyük Veri Uygulamaları

- Ekonomik ve **ticari faaliyetlerden kamu yönetimine**, **ulusal güvenlikten bilimsel araştırmalara** kadar birçok alanda **büyük veriden yararlanılmaktadır.**
- Büyük veri uygulamalarının en önemli amacı, **tüketici deneyimlerinin iyileştirilmesi**, **maliyetlerin düşürülmesi**, daha iyi **pazarlama stratejilerinin oluşturulması** ve **mevcut süreçlerin etkinliğinin artırılmasıdır.**
- Günümüzde **güvenlik saldırıları** ve **veri gizliliği** konularında da büyük veri kullanılmaya başlanmıştır.
- Büyük verinin başlıca uygulama alanları arasında **bankacılık**, **iletişim**, **medya** ve **eğlence sektörü**, **sağlık hizmetleri**, **eğitim**, **üretim**, **devlet hizmetleri**, **sigortacılık**, **perakendecilik** ve **ticaret**, **ulaşım**, **enerji sektörü** ve **ölçüm verisinin analiz edilmesi** yer almaktadır.

27

Büyük Veri Uygulamaları

- Statista tarafında yapılan araştırmalara göre, **2016 yılı itibarıyla** büyük veri ve analitiğinin dünya genelindeki pazar payında, **üretim %20,03** ile en çok gelir sağlayan uygulama alanı olmuştur.
- Bankacılığı, **%13,10 ile bankacılık**, **%7,60 ile devlet hizmetleri** ve **%7,40 ile de profesyonel hizmetler** takip etmiştir.
- 2016 yılında büyük verinin tüm uygulama alanlarındaki **toplam pazar değeri ise 130,10 milyar dolar** seviyesine ulaşmıştır.
- Diğer bir araştırma kuruluşu IDC ise 2016'da elde edilen bu toplam gelir değerinin, yıllık %11,7'lik büyüme oranı ile **2020 yılında 203 milyar dolardan daha fazla** seviyelere ulaşmıştır.

28

Büyük Veri Uygulamaları

Bankacılıkta büyük veri uygulamaları

- Bankacılık alanında büyük veri analitiği ile geçmiş veri kümelerinden düne göre daha fazla kazanç elde edilmektedir.
- Geçmiş veri, nakit hareketlerinin, öngörülebilen felaketlerin, soygunların ve müşteri davranışlarının anlaşılmasında yardımcı olmaktadır.
- Büyük veri kullanımıyla bankalar; **para hareketlerinin detaylarını görebilmekte**, **felaketleri ve hırsızlık olaylarını önceden öngörüp önleyebilmekte**, **tüketici davranışlarını daha iyi anlayabilmekte** ve **analiz edebilmektedir.**

29

Büyük Veri Uygulamaları

İletişim, medya ve eğlence sektörlerinde büyük veri uygulamaları

- Haberleşme ve sosyalleşme aracı olan sosyal medya, her geçen gün insan hayatındaki önemini artırmaktadır.
- Akıllı telefonların kullanımının artması ve yüksek hızlı mobil ağların genişlemesi, kişiler tarafından üretilen verinin **anlık olarak Web sayfalarına yüklenmesi** kültürünü ortaya çıkarmaktadır.
- Büyük verinin en çok kullanım alanları arasında, **sosyal medya üzerinden müşteri memnuniyetinin ölçülmesi** yer almaktadır.
- **Müşterilerin ürün ve hizmetler hakkındaki düşüncelerini yakından takip** edebilmek için organizasyonlar müşteri geri bildirimlerine değer vermelidir.
- Tüketicilerin bir ürün hakkındaki düşüncelerini yansıtan Web sayfası üzerindeki beğen butonlarından elde edilen veri, Twitter üzerinden paylaşılan yorumlar örnek teşkil etmektedir.

30

Büyük Veri Uygulamaları

Sağlık hizmetinde büyük veri uygulamaları

- Sağlık hizmetleri alanında üretilen **verinin miktarı gün geçtikçe artmaya devam etmektedir.**
- Hastalıklarla mücadele eden bireylerin sağlık kayıtları büyük veriyi oluşturan önemli kaynaklar arasında yer almaktadır.
- Büyük veri, **belirli hastalıkların örüntü ve eğilimlerinin gözden geçirilmesini sağlamakta** ve **erken teşhis** fırsatını sunmaktadır.

Eğitimde büyük veri uygulamaları

- Büyük veri, birçok eğitim organizasyonu tarafından öğrencilerin sistemlere ne zaman giriş yaptıkları, **gezindikleri Web sayfaları**, sayfalarda **ne kadar süre harcadıkları** ve belirli bir zaman içindeki faaliyetleri gibi olayların genel örüntüsünün ortaya çıkartılmasında büyük veriden yararlanılabilmektedir.

31

Büyük Veri Uygulamaları

Üretimde büyük veri uygulamaları

- Üretim ve kaynak temini alanlarında karar verme süreçlerini desteklemek ve bu bağlamda rekabet avantajı elde etmek için, **büyük verinin coğrafi, grafiksel, metinsel ve zamansal unsurlarından bilgi çıkaran tahmin modellerinden yararlanılmaktadır.**
- Ayrıca, **akıllı üretim süreci** ve ürün yaşam döngüsü yönetimi gibi gelişmekte olan uygulamalar, büyük veriyle birlikte gerçek yaşamda kullanılmaya başlamıştır.
- **Akıllı üretim sistemlerinde aktif önleyici bakım**, büyük veri analitiği yoluyla uygulanabilmektedir.
- Üretim alanındaki büyük verinin desteğiyle üretim **cihazlarının sağlık durumunu değerlendirmek** ve **arızalarını önceden tespit etmek** için **cihaz alarmları**, **cihaz olay kayıtları** ve **cihaz durum bildirimleri** gibi gerçek zamanlı birçok cihaz verisi analiz edilebilmektedir.

32

Büyük Veri Uygulamaları

Devlet hizmetlerinde büyük veri uygulamaları

- Kamu kurum ve kuruluşları, büyük veriyi toplayan, araştıran ve analiz eden yeni araçlar ile yapısal olmayan veriden faydalı bilgi elde edebilmektedir.
- Devlet hizmetlerinde, **her gün petabaytlar seviyesinde veri üretilmektedir.**
- Büyük verinin **gerçek zamanlı analizi**; **eğitim kalitesinin artırılması**, **işsizlik oranının azaltılması**, **trafikle ilgili canlı akış verisi** temel alınarak **trafik yoğunluğunun kontrol edilmesi** ve **mobil ambulans hizmetlerinin iyileştirilmesi** gibi birçok alanda yardımcı olmaktadır.

33

Büyük Veri Uygulamaları

Sigortacılıkta büyük veri uygulamaları

- Sigortacılık alanında büyük verinin kullanılmasıyla daha iyi fiyat ayarlaması yapılarak ve daha iyi müşteri ilişkileri kurularak, **sigorta organizasyonlarının kârlılığı ve performansı artırılabilmektedir.**

Perakendecilik ve ticarette büyük veri uygulamaları

- Perakendecilikte büyük veri akışı **beş boyutta** görselleştirilebilmektedir: **Müşteriler, ürünler, zaman, yer** ve **kanallar**.
- Perakendecilikte büyük veri kullanımının sağladığı başlıca faydalar arasında; **stokların doğru bir şekilde gösterilmesi**, **zamanında analiz edilmesi**, **alışveriş örüntülerinden elde edilebilecek bilgilerin kullanılmasıyla personel istihdamının optimizasyonu** ve **müşteri ilişkilerinde devamlılığın sağlanması** yer almaktadır.

34

Büyük Veri Uygulamaları

Ulaşımda büyük veri uygulamaları

- **Trafiği kontrol etmek**, **en iyi ulaşım rotasını planlamak**, **akıllı ulaşım sistemleri geliştirmek**, **trafik koşullarını tahmin ederek oluşabilecek tıkanıklıkları yönetmek** için büyük veriden yararlanabilmektedir.
- Özel sektörde ise büyük veri sayesinde gönderilerin **nakliye hareketlerinin optimizasyonu** sağlanarak gelirlerde artış ve rekabetçi avantaj elde edilebilmektedir.
- Bireysel olarak **yakıt ve zamandan tasarruf sağlamak** amacıyla **uygun ulaşım rotasının planlanmasında büyük veri kullanılabilir.**
- **GPS** alıcı-vericileri, **CCTV** sistemleri, **dedektörler**, **cep telefonları** ve diğer taşınabilir cihazlar ile toplanan **yol durumu**, **araç ve sürücü davranışları** büyük veriyi oluşturmaktadır.
- Bu verinin kullanımıyla geliştirilen hızlı ve dinamik modellemeler, akıllı ulaşım sistemleri için daha iyi simülasyon ortamları sağlayabilmektedir.

35

Büyük Veri Uygulamaları

Enerji sektöründe büyük veri uygulamaları

- Büyük veri analiz yöntemleri **kaynak ve işgücü yönetiminin sağlanması**, **problemlerin önceden tespit edilmesi** amacıyla kullanılmaktadır.
- Enerji sektöründe **sensörlerin, bulut bilişim teknolojilerinin, kablosuz ve ağ iletişiminin** kullanılmasıyla birlikte büyük miktarda veri elde edilmektedir.

Kendi kendine ölçümde büyük veri uygulamaları

- **Kişisel aktivite ve davranışlarını ölçümleyen** bireyler tarafından üretilen **veri**, kendi kendine ölçüm verisi (self-quantification data) olarak adlandırılmaktadır.
- **Kişilerin hareketlerini**, **egzersizlerini** izleyen ve buradan elde edilen veriyi akıllı telefon uygulamasına aktararak **verinin analiz edilmesini sağlayan** bileklikler kendi kendine ölçüm verisi üretmektedir.

36