

Makine Öğrenmesinde Etik Sorunlar



YÜCEL SAYGIN
SABANCI ÜNİVERSİTESİ

Temel bazı tanımlar



Veri Sahibi

Veriyi Tutan

Veriyi İşleyen

Kişisel Veri

Hassas Veri

Kavramlar ve İlişkileri



Trust

Privacy

Kavramlar ve İlişkileri



Trust

Privacy

Discrimination

Fairness

Transparency

Explainability

Açık Veri



- Veriyi kullanıma açmak
- Makine öğrenmesi ve çeşitli modellemelerle toplum için fayda sağlamak
- Aynı zamanda mahremiyeti korumak

Anonimleştirme



- “Anonim” verilerde hukuki anlamda mahremiyet sorunu ortadan kalkar
- Nasıl yapılmalı?
 - Kişiyi direkt olarak belirleyen kısımları atmak

İsim Soyisim	TC Kimlik	Doğum Tarihi	Cinsiyet	Posta Kodu	Teşhis	Tedavi
Yücel Saygın	1511...	26.04.72	Erkek	34956	Şizofren	Elektroşok
...	...				...	...

Problem!

Doğum Tarihi	Cinsiyeti	Posta Kodu	Teşhis	Tedavi
26.04.72	Erkek	34956	Şizofren	Elektroşok

İsim Soyisim	TC Kimlik	Doğum Tarihi	Cinsiyeti	Posta Kodu
Yücel Saygın	1511....	26.04.72	Erkek	34956

Biraz daha karışık bir durum



- Ağustos 2006'da, AOL
 - 650.000 kullanıcının 3 ay boyunca yaptığı 20 milyon arama kelimelerini bir web sayfasında yayınladı.
 - Bu veriler içinde kullanıcı bilgileri yoktu ama yine de büyük bir skandala yol açtı
- AOL'e dava açıldı
- Benzer skandallar devam etti

Daha da karışabilir



facebook

twitter

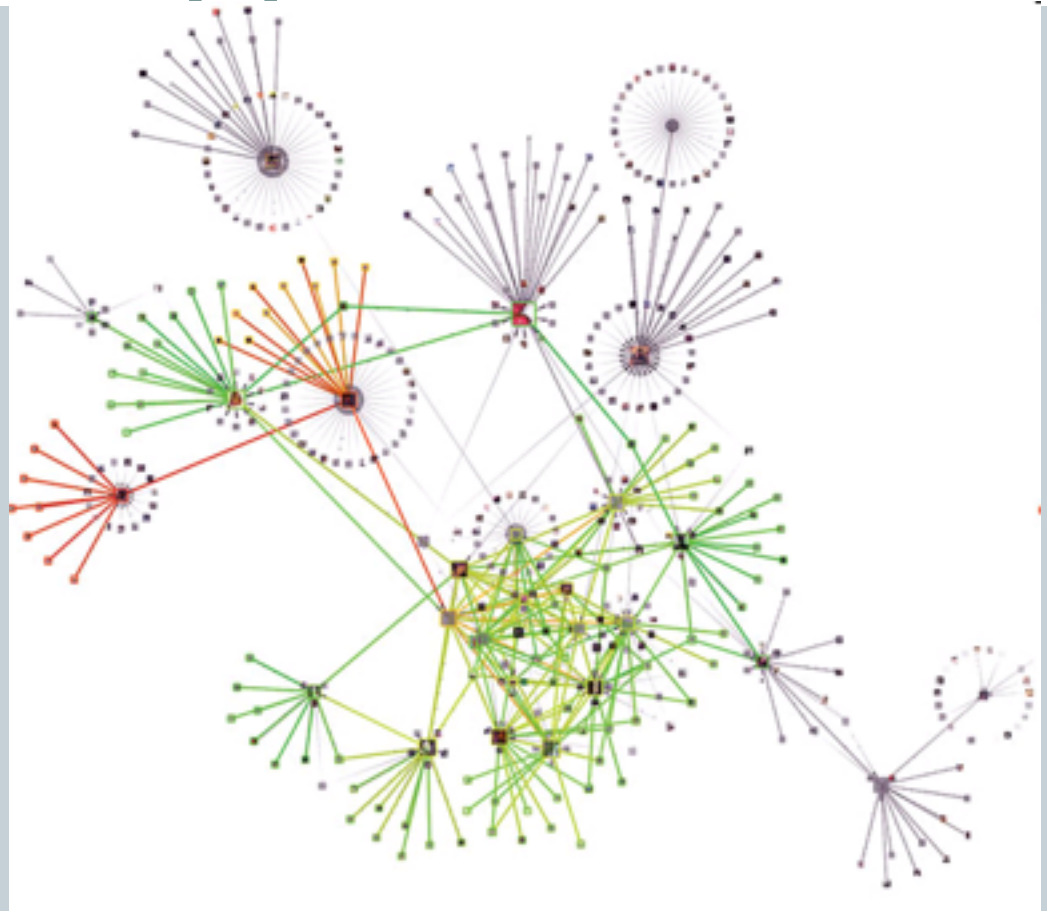
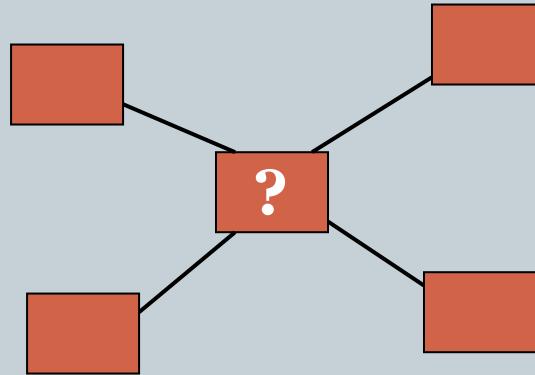


Image source: <http://net-savvy.com/executive/social-media-analysis/visualizing-networks-based-on-communication.html>

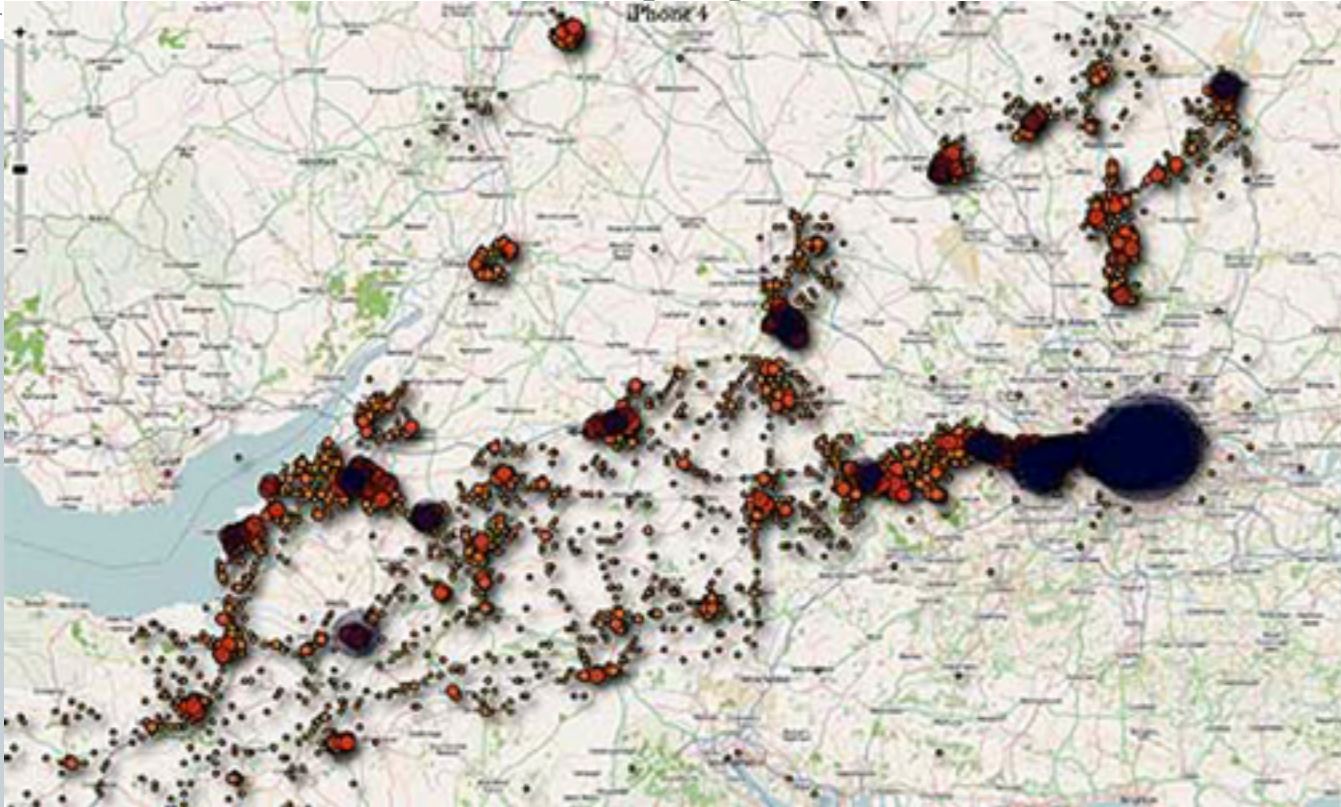


Makine Öğrenmesi ile Hassas Veri Tahmini

Bana arkadaşını söyle sana kim olduğumu söyleyeyim!



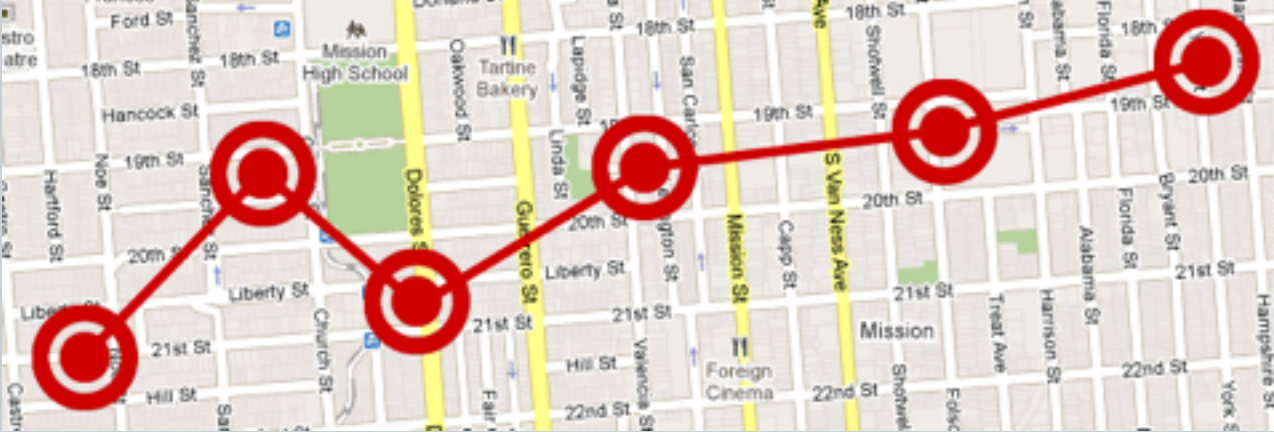
Konum Verileri



A location map from an iPhone that had been used in the southwest of England

Image source: <http://9to5mac.com/2011/04/20/ios-secretly-storing-your-location-data-say-researchers/>

Kim nerde ne yapıyor?



Konum+zaman, biraz veri temizleme ve görselleştirme

Imge Source: <http://mashable.com/2010/05/26/location-hype/>

Peki ne yapmalıyız?



- Anonimleştirme farklı veri türleri ve güçlü saldırganlar düşünüldüğünde zor ya da imkansız.
- Aslında ne yaparsak yapalım herhangi bir veri paylaşımında mahremiyet ihlali olabilir.
- O zaman veriyi paylaşmak yerine sadece istatistikleri paylaşalım!

İstatistik paylaşımı ve sorunları



- **İstatistiksel Veri Tabanları**

- *Sorgu I: “HIV taşıyan hastaların sayısı”*
- *Sorgu II: “HIV taşıyan ve İstanbulda yaşayan hastaların sayısı”*
- *Sorgu III: “HIV taşıyan, İstanbulda yaşayan ve 20-22 yaşları arasında olan hastaların sayısı”*

İstatistik paylaşımı ve sorunları



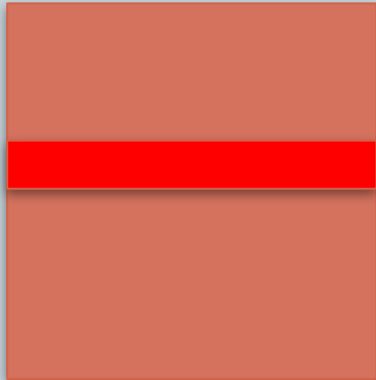
- İstatistiksel Veri Tabanlarında sorgulara karşı mahremiyeti korumak zor
 - *Auditing*
 - *Sorgu kısıtlama*
 - *Ard arda gelen sorguların kesişimine bakarak kısıtlama*
 -

Çözüm: Yeni bir mahremiyet standardı

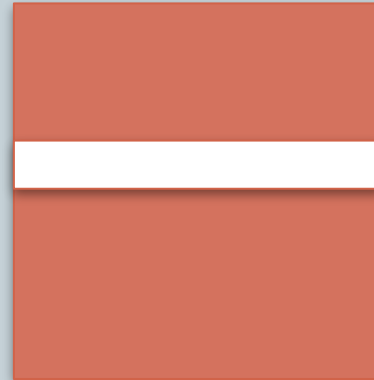


- **Diferansiyel Mahremiyet:**

- *Paylaşılan istatistik veri içindeki tek bir kaydın değişiminden etkilenmemeli*



D

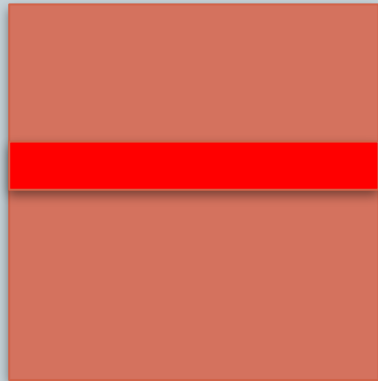


D'

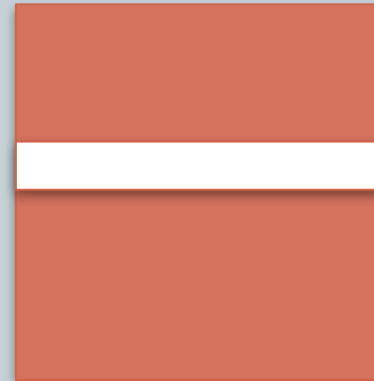
Diferansiyel mahremiyetin avantajları



- Verinin sızma riskini minimize ediyor
- İnsanlara verisini paylaşması için bir motivasyon oluşturuyor



D



D'

Formal Tanım



- Neighboring datasets:

Two datasets D and D' , such that D' can be obtained by changing a single tuple in D

- Differential privacy (ϵ -DP) :

A randomized algorithm A satisfies ϵ -differential privacy, iff for any two neighboring datasets D and D' and for any output S of A ,

$$\Pr[A(D) = S] \leq \exp(\epsilon) \cdot \Pr[A(D') = S]$$

- ϵ is a threshold set by the data owner
 - Lower ϵ -> stricter privacy requirement

Hassasiyet (Sensitivity)

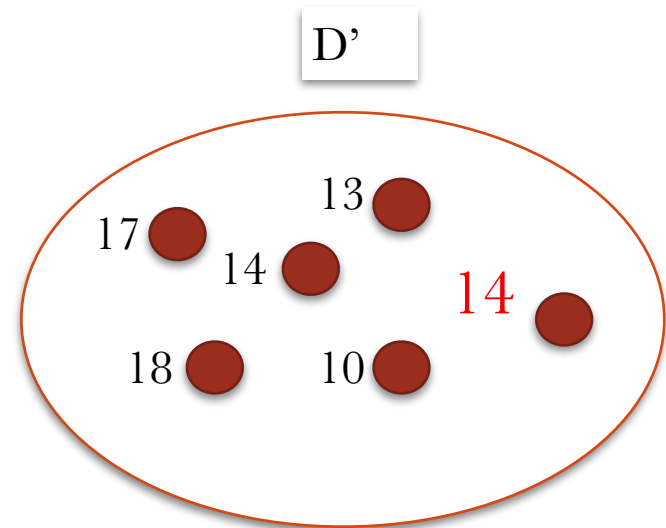
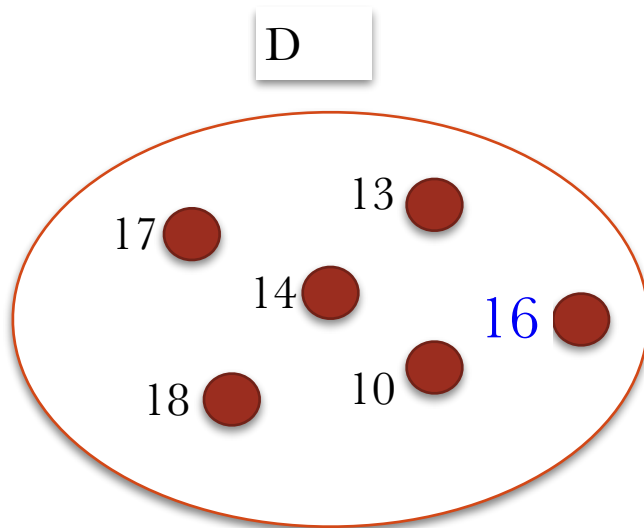


- Sensitivity of a function/algorithm

$$\Delta f := \max_{D, D'} ||f(D) - f(D')||$$

Örnek

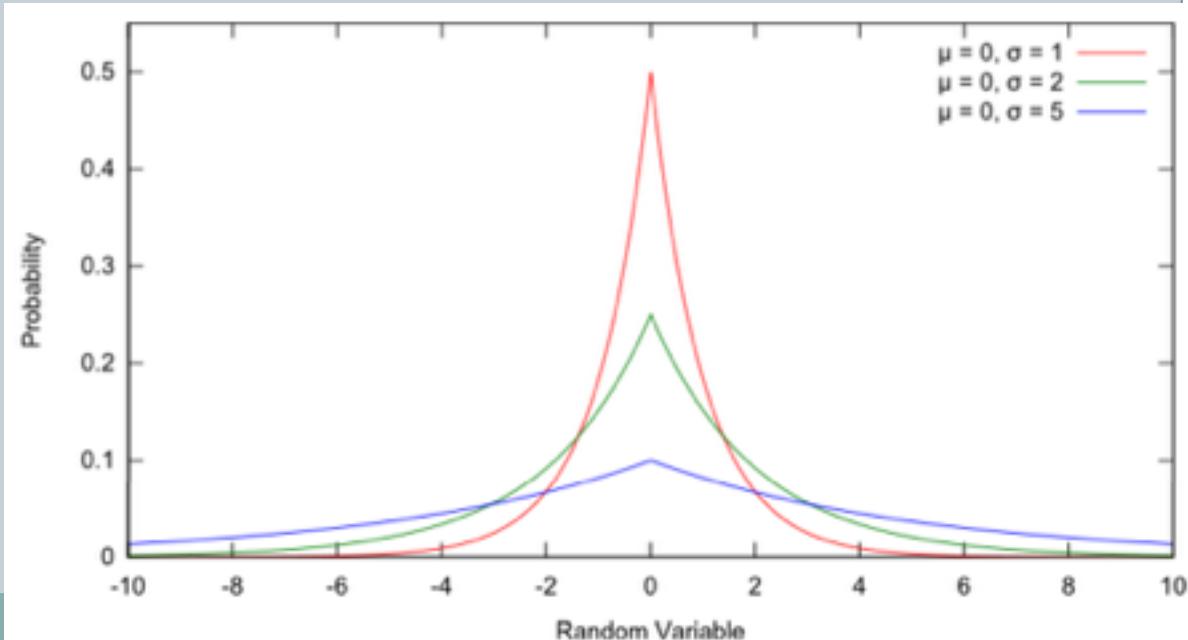
- Q: select count(*) from Students where age > 15
- Sensitivity of Q?



Laplace Mekanizması



- given Δf and ϵ
- Let $\text{Lap}(\sigma)$ denote a random variable sampled from the Laplace distribution with mean 0 and scale parameter σ
- Laplace mechanism for ϵ -DP by adding noise $\text{Lap}(\sigma)$ where $\sigma \geq \Delta f / \epsilon$



Sıralı Kompozisyon



- A_1 ve A_2 ϵ_1 ve ϵ_2 -DP algoritmalar olsun.
 - Sıralı Kompozisyon: $A_1(D)$ ve $A_2(D)$ algoritmalarını sonuçlarını yayınlamak $(\epsilon_1 + \epsilon_2)$ -DP sağlar.
- Dolayısıyla ϵ parameteresi mahremiyet bütçesi gibi düşünülebilir
- Herhangi bir makine öğrenmesi işi farklı alt işler bölünerek ϵ bütçesi bu işlere paylaştırılabilir.

Diferansiyel Mahremiyetin Eleştirisi



- Gürültü düşük ihtimal olsa da çok fazla olabiliyor
- Sonuçların doğruluğunun önemli olduğu kritik uygulamalar için problem
- Mahremiyet ve Yarar arasındaki denge

Yeni Trendler



- Dağıtık ortamlarda veri toplama/öğrenme
 - Local differential privacy
 - Federated Learning



Mahremiyetin Ötesinde Diğer Etik Sorunlar

Makine öğrenmesi ve ayrımcılık



- Ayrımcılık: **Din, Irk, Cinsiyet, Politik tercihler...**
- Makine öğrenmesi ile oluşturulan sınıflandırma modellerinin ayrımcılık yapmaması gerekir

Makine öğrenmesi ve ayırmcılık



- Örnek: banka kredi verme kararı

Makine öğrenmesi ve ayrımcılık



- Şeffaflık (Transparency)
- Derin öğrenme modellerinde şeffaflık problemi

Google Arama Skandalları

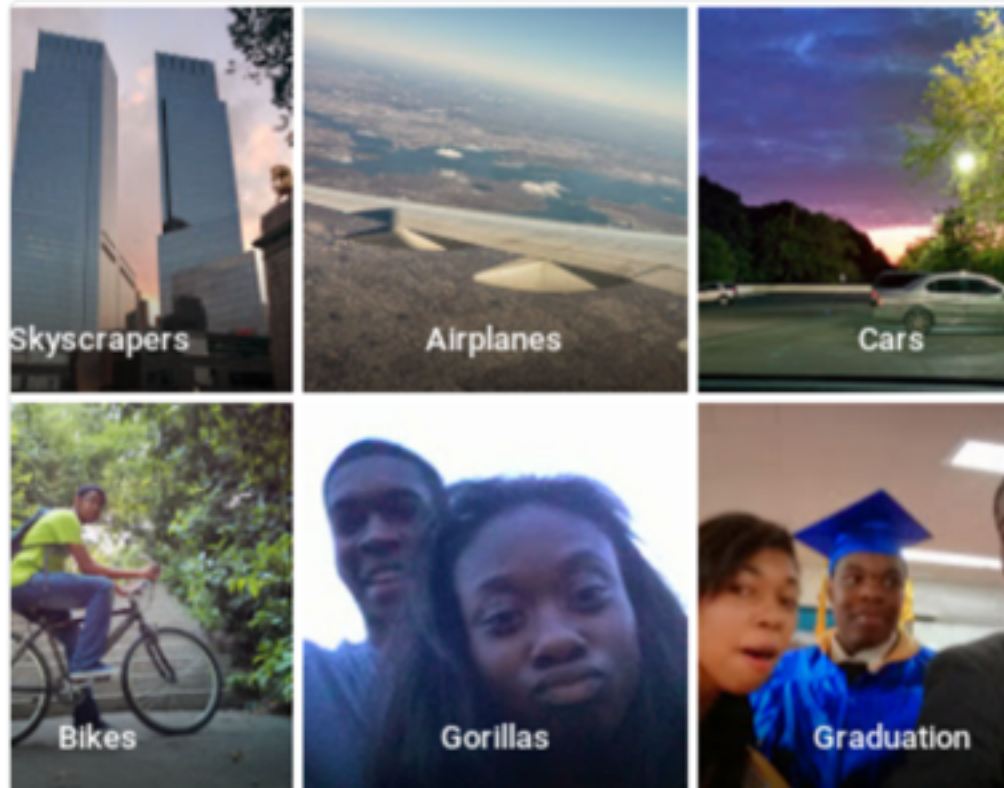


- <http://edition.cnn.com/2009/TECH/11/25/google.michelle.obama.controversy-2/>
- "Michelle Obama" Google arama sonucu

Google İmge Sınıflandırma Skandalı



- <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>



Ayrımcılığın kaynağı



- Makine öğrenmesi algoritmaları mı problemli?
- Yoksa insanların yanlı görüşleri ile oluşturulan eğitim verisi mi problemli?

Arama Motorları ve Yanlı Sonuçlar (Bias)



Geleneksel Medya



- Az sayıda kontrolü mümkün olan yayınlar
- Editörlerin/Yapımcıların süzgecinden geçmiş içerik
- Geleneksel medya meslek eğitimi ve etik değerler



Arama Motorlarının Çalışma Prensipleri



- Kullanıcı arama motoruna bir sorgu gönderir
- Arama motoru, verilen sorgu için en ilgili web sayfalarını (dokümanları) döndürür
- Bu en ilgili dokümanlar arama motorunun sıralama algoritması tarafından sıralanır

Sonuçlara Ne Zaman Yanlı Diyebiliriz?



- Eğer arama motorundan dönen sonuçlar belli web sayfalarına (dokümanlara) öncelik veriyorsa
- Ya da gönderilen sorgu ile ilgili birçok farklı perspektif varken üst sıralarda yalnızca (veya daha sıklıkla) belli perspektiflere yer veriliyorsa

Örnek



- Sorgu: Brexit
- Web’de destekleyen (pozitif) ve desteklemeyen (negatif) dokümanlar var
- Eğer kullanıcı bu sorguyu gönderdiğinde, üst sıralarda hep bu ambargoyu destekleyen (bu konu ile ilgili pozitif) dokümanları görüyorsa
- O zaman arama sonuçları Brexit sorgu bağlamında taraflı diyebiliriz.

Döner bu sıralı arama sonuçlarının yanı olmasının sebebi



- Sıralama algoritması
 - Algoritma dizaynı, ve
 - Algoritmayı eğitmek için kullanılan tıklanma verisi
- Sorgu ile ilgili web'de bulunan dokümanlar (girdi veri seti)

Kişiselleştirme



Yankı Odaları (Echo Chambers)



Kavramlara genel bir bakış



Trust

Privacy

Discrimination

Fairness

Transparency



TEŞEKKÜRLER