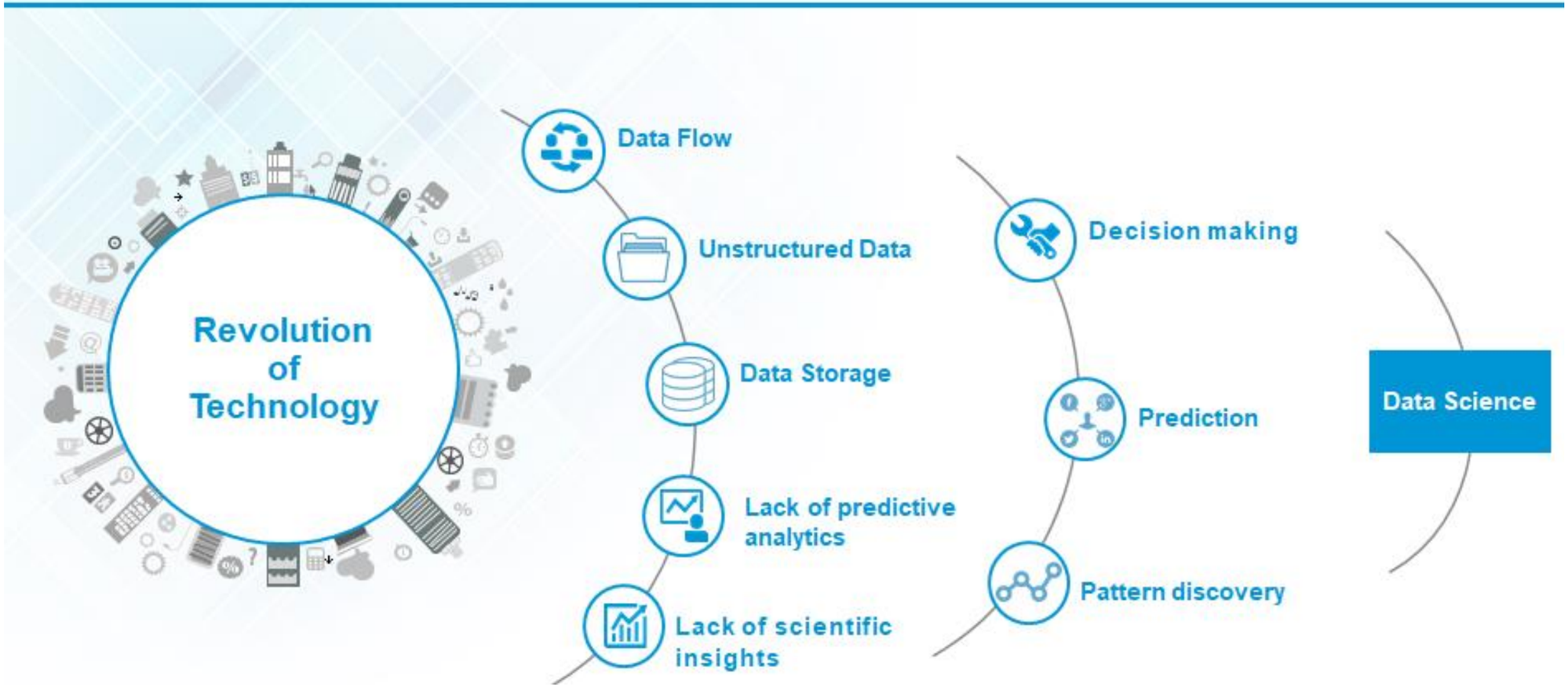
Veri Bilimi

Dr. Cahit Karakuş

Istanbul, Turkey

Need Of Data Science



Veri Bilimi

- Veri Yapıları ve Kodlama
- Bilgisayar Organizasyonu
- Veri Analitiği
- Veri Analitiği için İstatistik ve Olasılık
- Computational Methods for Data
- Uygulamalı Matematik (Lineer Denklem Sistemi, Vektörler, Matrisler, Özdeğer ve özvektçrler, ikinci dereceden diferansiyel denklemlerde kararlılık durumları)
- Sinyaller ve Sistemler
- Yapay Zeka - Makine Öğrenmesi – Derin Öğrenme

Veri Bilimi

Visualization / Reporting / Knowledge

- Dashboard (Kibana / Datameer)
- Maps (InstantAtlas, Leaflet, CartoDB…)
- Charts (GoogleCharts, Charts.js…)
- D3.js / Tableau / Flame

Analysis / Statistics / Artificial Intelligence

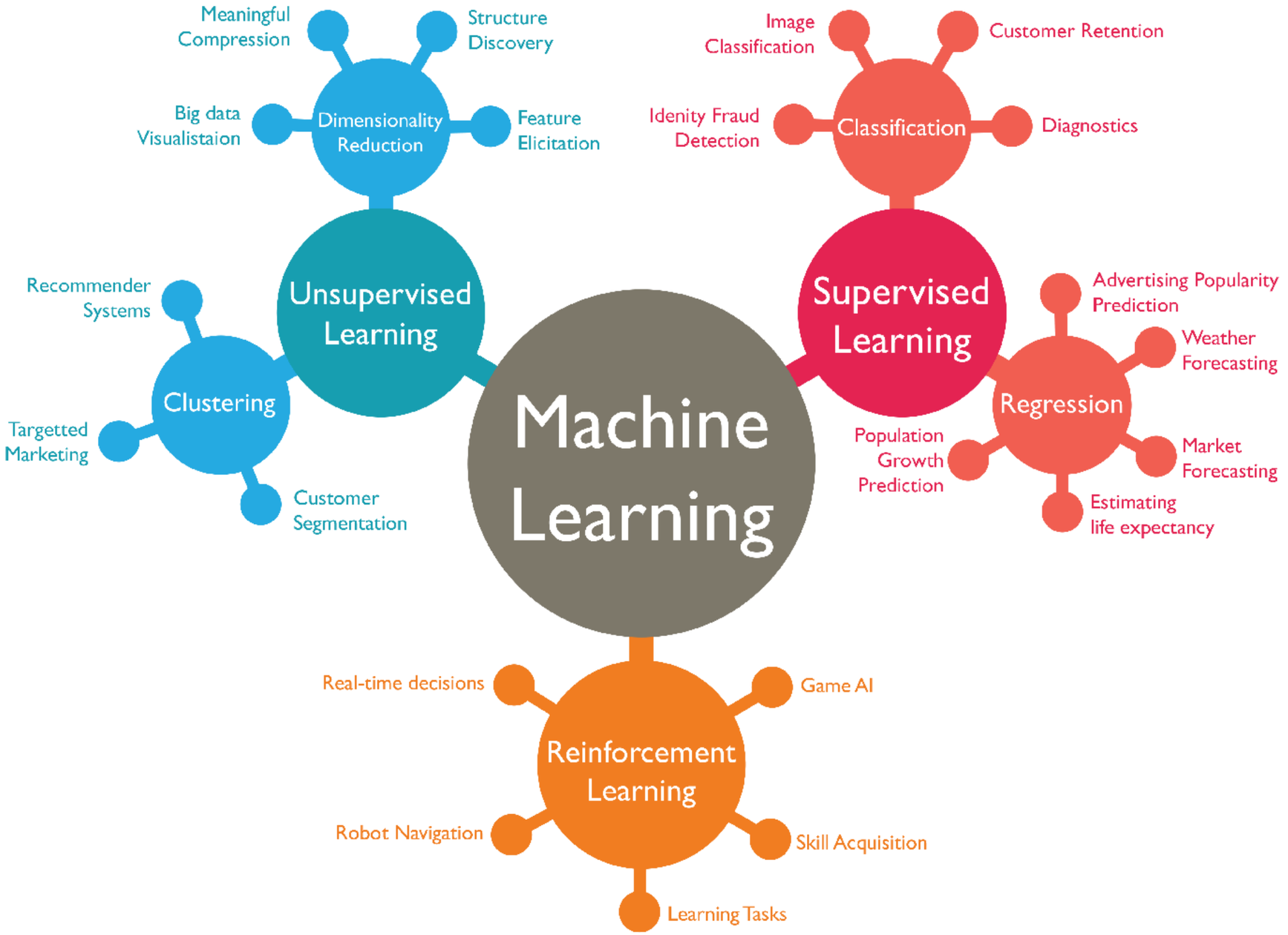
- Machine Learning
- Search / retrieval

Storage / Access / Exploitation

- File System
- Access
- Databases / Indexing (SQL / NoSQL / Both…, MongoDB, HBase, Infinispan)
- Exploit (LogStash, Flume…)

Infrastructures

- Grid Computing / HPC (High Performance Computing)
- Cloud / Virtualization


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

What is Computer Science?

- Alanın adından, bilgisayar biliminin “bilgisayarların incelenmesi” olduğu düşünülebilir.
- “Bilgisayar Bilimi, astronomi ne kadar teleskoplarla ilgili ise bundan daha fazla bilgisayarla ilgili değildir”
- CS bilgisayarların çalışması değilse, nedir?
- Bu soruyu yanıtlamak için önce bilgisayarların ne olduğunu ve ne yapabildiklerini soralım - sonra ana soruya “bilgisayar bilimi nedir?” sorusuna geri dönelim.

Elektronik Beyinler

- Bilgisayarlar 1950'lerde yaygın olarak kullanılmaya başladığında, genel olarak "elektronik beyinler" denildi.
- Onlar kocamandı, pahalıydı, gizemliydi
- yalnızca en büyük şirket, devlet veya üniversite laboratuvarlarında bulunurdu.
- Bilim kurgu filmleri bu "elektronik beyin" bakış açısını kullanıldı.

- Neden başka bir metafor yerine bir “beyin”?

- bilgisayarlar tarafından yapılan hesaplamalar, yalnızca yüksek eğitimli kişilerin yapabileceği türden şeylerdi.

- örnek: bir roketin yörüngesini hesaplanması.

Appliances

- Bugün bilgisayarlar modern toplumda temel araçlardır
- Kişisel bilgisayarlar (masaüstü ve dizüstü bilgisayarlar)
  - makaleler yaz, kişisel finansı yönet, ...
  - eğlence: oyunlar, video, ses, ...
- İş bilgisayarları:
  - günlük işlemler: bordro, faturalandırma, ...
  - müşteri hizmetleri: web siteleri, müşteri verileri, ...
- Gömülü bilgisayarlar: arabalarda, telefonlarda, binalarda kontrolör olarak kullanılan mikroçipler...
- süper bilgisayarlar:
  - bilimsel araştırmalarda ve diğer alanlarda kullanılan büyük veri hesaplamalar
  - paralel işleme: birkaç düzineden birkaç bin CPU yongasına kadar

Bilgisayarların Yapamayacakları

- “Bilgisayar nedir?” sorusuna yaklaşmanın bir yolu bir bilgisayardan yapmasını beklemediğimiz bazı şeylere bakmaktır.
- Bir sorunu sizin için çözecek bir bilgisayara güvenir miydiniz?
  - Bir çoğunu ölçmek zor olan birkaç faktör söz konusudur
  - başka birinin bu sorunu sizin için çözmesini de beklemiyorsunuz.
  - buradaki sorun bilgisayar değil, sorunun doğası
  - En iyi seçimi “hesaplamak” için her faktöre kesin bir ağırlık vermeniz gerekir.
- Cep telefonlarımız daha çok kişisel asistan gibi olsaydı iyi olurdu
  - telefonu açın, "Erica ve Katie'yi DVD izlemeye davet edin" deyin
  - telefonunuz, işe yarayan bir zaman seçmek için telefonlarıyla görüşür
  - bu, insanların yapabileceği türden bir şeydir (“insanların benim halkımı aramasını sağla...”)
- Önceki problemden farklı olarak, bu, makinelerin (şimdiye kadar) yapamayacağı, insanların yapabileceği bir şeye bir örnektir.
- Gelecekteki bir bilgisayar bu sorunu çözebilecek mi?

Bilgisayarların Yapamayacakları

- Bir bilgisayarın bir satranç oyununu kazanmasının kolay olacağını düşünebilirsiniz.
- Oyunun kuralları basittir ve bir bilgisayarın olası tüm hareketleri incelemesini sağlayacak bir program yazmak kolaydır.
- Sorun: dikkate alınması gereken çok fazla hareket var
  - tahmini 10^43 olası oyun var
  - 10^12 kart/sn kontrol eden bir süper bilgisayarın hepsine bakması için 10^21 yıl gerekir
- Yani burada yeni bir tür sınırlamamız var: pratik bir sınır
  - insanlar bu görevi yerine getirmede makinelerden daha iyi değiller
  - büyük ustalar tüm olası hamleleri dikkate almaz
- Bilgisayar bilimlerindeki ünlü bir problem, "durma problemi" olarak bilinir.
- Amaç: başka bir programın takılıp takılmadığını belirleyen bir program yazmak
  - bir makale yazdığınızı ve işaretçinin "meşgul" simgesine dönüştüğünü varsayalım
  - kelime işlemcinin çöküp çökmediğini görmek için bir "durma denetleyicisi" yazmak imkansız
- Bu sorun mantıktaki paradokslarla ilgilidir (örneğin, “bu ifade yanlıştır”)
  - yeni bir sınırlama türü: matematiksel bir engel
  - insanların sorunu çözmede makinelerden daha iyi olmadığı başka bir örnek

Özet: Hesaplamalı Limitler

- Önceki slaytlarda ilginç bir tema ortaya çıktı
- Bilgisayarların yapamadığı bazı görevler de insanlar için imkansızdır.
- Zorluk, sorunu çözmeye çalışan kişi ya da şeyde değil, sorunun doğasındadır.
- Bir problem insanlar veya bilgisayarlar tarafından çözülemeyebilir, çünkü bazı nitelikler ölçülemez (bir üniversitede yaşam kalitesi) pratik değil (satranç) imkansız (denetleyiciyi durdur)
- İnsanların çözebildiği ancak bilgisayarların çözemediği problemler genellikle “zeka” terimleriyle tanımlanır.
- doğal dil işleme, planlama, tasarım, ...bilgisayar biliminin araştırmasının aktif bir alanı: yapay zeka (AI)

Hesaplama

- Bir hesaplama, bir başlangıç noktasından istenen bir nihai sonuca götüren iyi tanımlanmış bir işlemler dizisidir.
- bu tanımın "bilgisayar" kelimesini içermediğine dikkat edin
- hesaplama, bir kişi veya bir makine tarafından gerçekleştirilebilen bir işlemdir.
- aynı hesaplama, bir dizi farklı teknolojiden herhangi biri kullanılarak gerçekleştirilebilir.

What is Computer Science?

- Bilgisayar bilimi, hesaplama çalışmasıdır
- Hesaplamalı olarak çözülebilecek problemleri araştırmak
- hesaplamaları tanımlamak için kullanılan programlama dilleri
- hesaplamalar yapan makineler
- hesaplamanın teorik sınırları (hesaplanabilir olan veya olmayan)
- matematik, bilim, tıp, işletme, eğitim, gazetecilik, ...
- Bilgisayarlar kilit rol oynuyor ama bilgisayar bilimi “bilgisayarlarla ilgili” değildir.

Algorithms

- Bir hesaplama sırasında gerçekleştirilen adımların sırası bir algoritma tarafından tanımlanır.
  - bir algoritma bir "reçete" olarak düşünülebilir“
  - bu adımları takip edin, sorununuzu çözeceksiniz”
- Bir algoritma, aşağıdakilerin tam bir tanımını içerir:
  - girdi seti veya başlangıç koşulları
  - çözülecek sorunun tam bir özelliği
  - çıktı seti
  - soruna geçerli çözümlerin açıklamaları
  - sonunda çıktıyı üretecek bir dizi işlem
  - adımlar basit ve kesin olmalıdır

Algoritmaların Nitelikleri

- Bir algoritmayı tam olarak neyin tanımladığını belirtmek zordur. “Basit ve kesin adımlar dizisi” ile ne demek istiyoruz?
- Algoritmalar hakkında yazan çoğu kişi, adımların olması gerektiği konusunda hemfikirdir.
  - kesinlik: herkes tarafından anlaşılabilir terimlerle yazılmalıdır, ama "kesin" ne anlama geliyor? bir adım ne kadar kesin olmalı?
  - etkili: bir adım, algoritmanın nihai hedefe ilerlemesine yardımcı olmalıdır, ama ne kadar etkili? “etkili”nin resmi bir tanımı var mı?
  - pratik: bir dizi kesin ve etkili adım pratikte faydalı olmayabilir
    - örnek (Knuth'tan): bir satranç turnuvasını kazanmak için varsayımsal bir algoritma: "her oyun için tüm olası hamleleri düşünün, en iyisini seçin"ancak bu adım en az 10^13 yıl sürecektir.“etkili” bir strateji değil.

History

- Bilinen en eski algoritmalar, örneğin Yunan matematikçiler tarafından tanımlanmıştır. İki tamsayının en büyük ortak böleni için Öklid yöntemi, 300 M.Ö.
- Modern "algoritma" kelimesi, İranlı bilgin Muhammed ibn Mūsā al-Ḵwārizmī'nin (yaklaşık 780 - yaklaşık 850) adından gelmektedir.
  - eseri Latince yayınlandığında adı Algoritma olarak yazıldı.
  - matematik ve doğa bilimleri üzerine birkaç etkili çalışmanın yazarıydı.
  - lineer denklemlerin sistematik çözümü üzerine kitabı birkaç algoritma içeriyordu.
  - Bu kitabın başlığı da cebir kelimemizin kaynağıdır.

Veri Bilimi

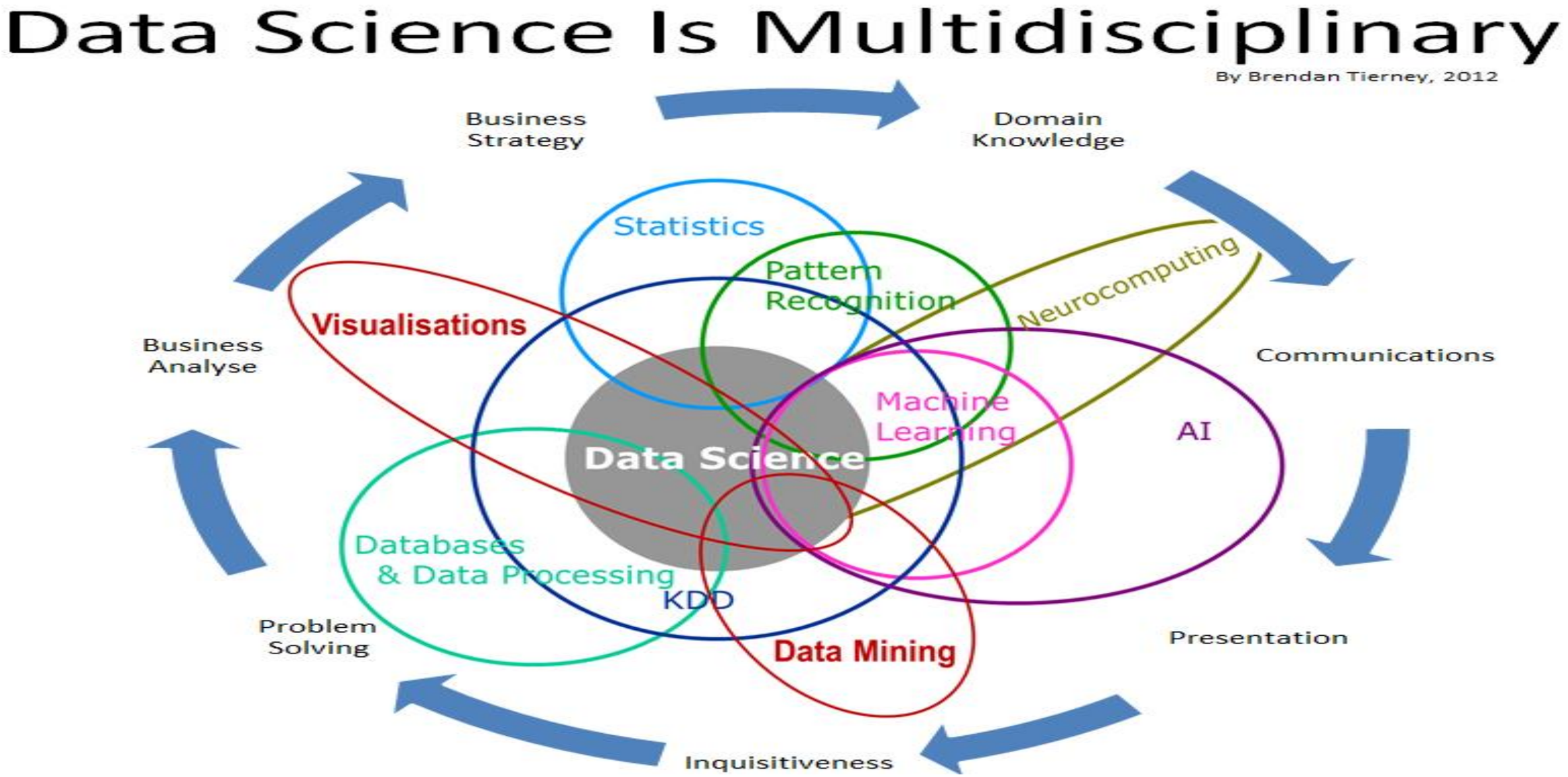
Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity. İletim ve işletim
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

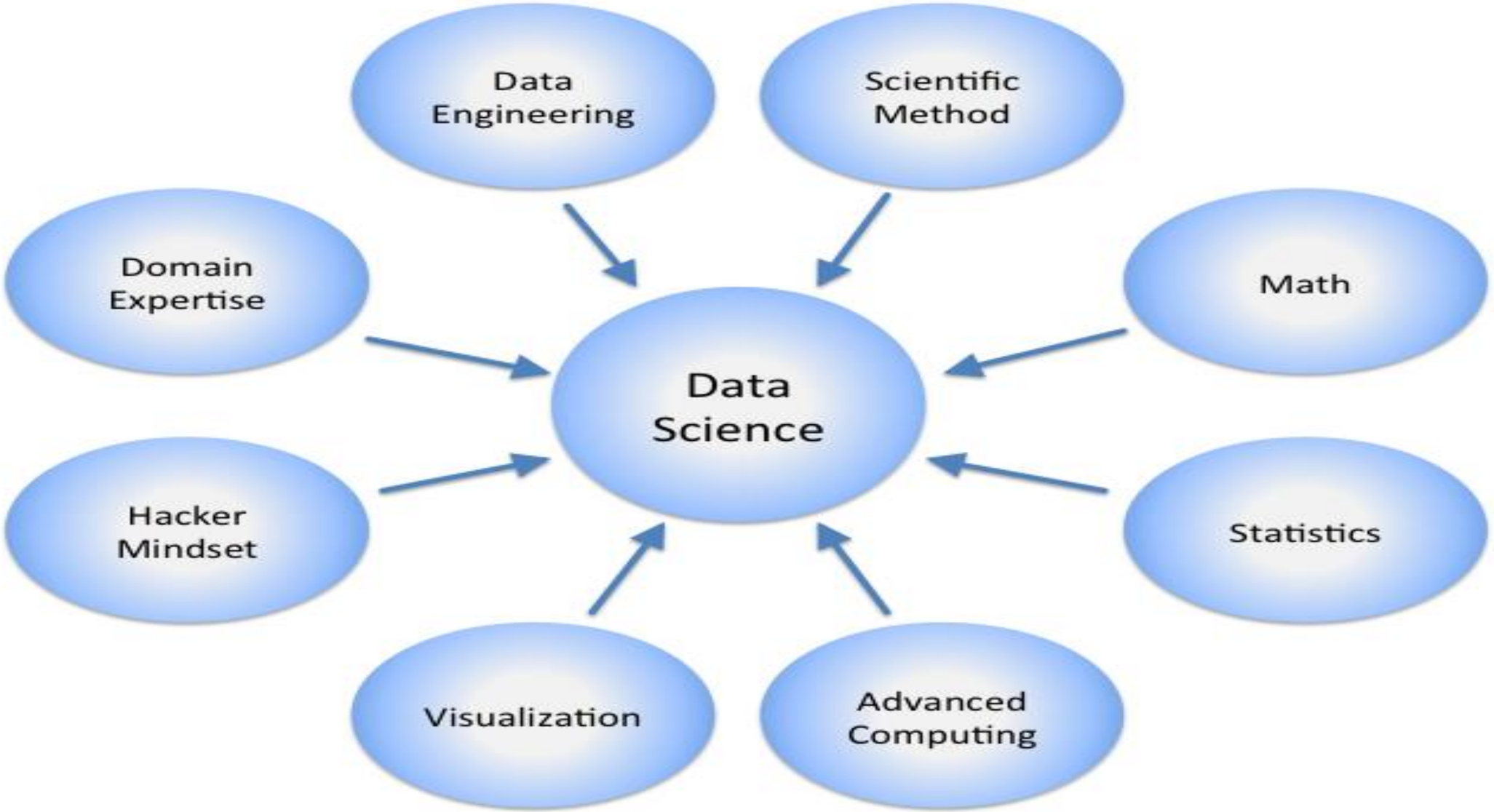
Big Data

- Büyük Veri, yönetilmesi pahalı ve değer elde edilmesi zor olan herhangi bir veridir.
- Hacşm: veri boyutu
- Hız: Artan etkileşim talebine göre veri işleme gecikmesi.
- İletim ve hizmet
- Çeşitlilik ve Karmaşıklık: kaynakların, biçimlerin, kalitenin, yapıların çeşitliliği.

Data Science



Data Science



Bilim nedir?

- Bilim, mühendislik, ekonomi, siyaset, finans ve eğitim gibi birçok sektördeki karar vericilere yardımcı olmak için birçok alandan ve disiplinden teoriler ve teknikler **büyük miktarda veriyi araştırmak ve analiz etmek** için kullanılır.
- Bilgisayar Bilimi: Desen tanıma, görselleştirme, veri ambarlama, Yüksek performanslı hesaplama, Veritabanları, Yapay Zeka
- Matematik: Matematiksel modelleme
- İstatistik: İstatistiksel ve Stokastik modelleme, Olasılık.

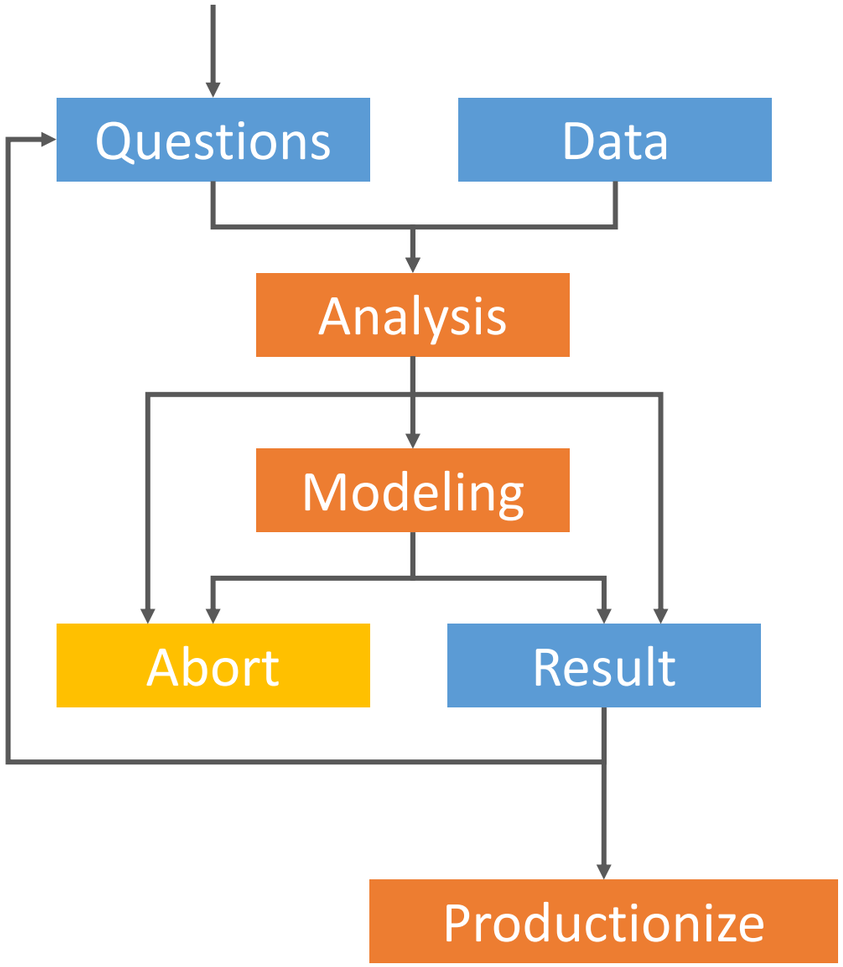
Veri bilimi üç geniş kategoriye girme eğilimindedir

- **Investigating:** Şu anda neler olup bittiğine dair temel içgörüler elde etmek için verileri toplama ve inceleme.

- **Predicting:** Verileri almak ve gelecekte ne olacağını anlamak için kullanmak.

- **Optimizing (Eniyileme):** Makine öğrenmesi ile yakın bağı vardır. birçok öğrenme probleminde, bir öğrenme seti örneğindeki işlevlerin en aza indirilmesine odaklanır.

The data science process



- Tüm analizler sorularla başlar
- Bir veri bilimcisi soruyu alır ve verileri araştırır
- Vakalar:
  - Verilerin soruyu yanıtlamak için doğru olmadığı açıktır
  - gelişmiş modelleme gerekli
  - sorunun cevabı verilerde hemen bulunabilir
- Bir sonuç bulunursa
- Üretilebilir
- Daha fazla soru ortaya çıkarabilir

The five skills needed to do data science

Technical Skills

- Statistics and Math – Veriler üzerinde kullanılan farklı teknikler: regresyonlar, kümeleme algoritmaları, zaman serisi modelleri
- Software Development – Kod nasıl yazılır, bir kod tabanı nasıl yönetilir, bir veritabanında nasıl veri depolanır

- Business Experience – Şirketlerin para israf ettiği yer, bir projenin başarılı olmasını sağlayan nedir, şirket içinden nasıl veri alınır?

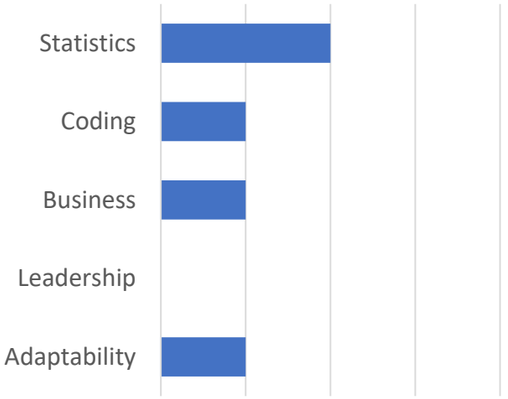
Personal Skills

- Leadership – Diğer veri bilimcilerine nasıl yardımcı olunur, onları nasıl eğitirsiniz ve iyi sonuçlar elde etmek için bir müşteriyle nasıl çalışılır?

- Adaptability – Tamamen yeni bir problemle sunulduğunda bir çözüm bulma yeteneği

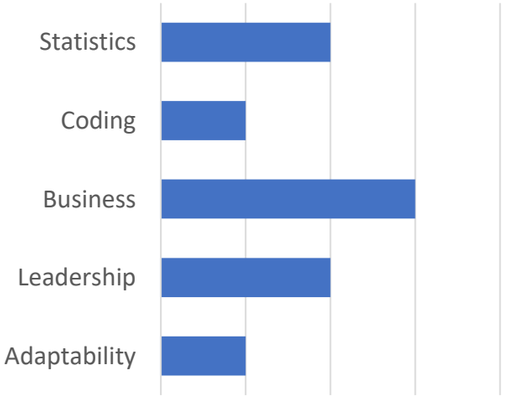
Data scientist archetypes

Junior data scientist (J) – BS ve sektörde üç yıldan az deneyime sahiptir. Tek basit istatistiksel teknikleri bilme eğilimindedir. Çok fazla rehberlik gerektirir, ancak daha az ilginç olan işi yapmaktan mutluluk duyar.



Expert junior scientist (E) – 5 yıldan uzun süredir çalışan genç bir veri bilimcisi. Basit şeyleri yaparken çok rahat olur ve iş alanları hakkında derinlemesine bilgi sahibidir. Kariyere yardımcı olmak için MS almış olabilir.

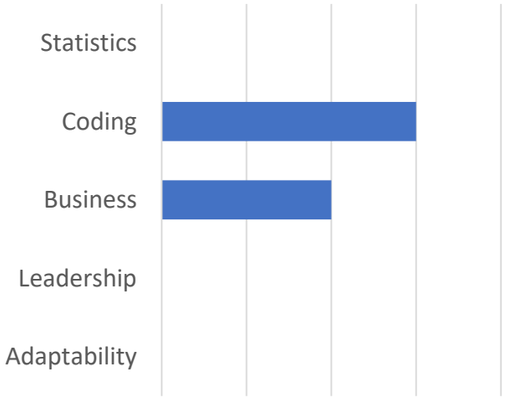
Senior data scientist (S) – İleri derece ve onunla ne yapacağını bilmek için yeterli iş tecrübesine sahip bir kişi. İşleri doğru yapmak için yeterince iyi kodlamayı anlar. Yine de iş yapmaya yeni bir veri bilimcisinden daha az isteklidir, ancak etrafta kimse yoksa işe yarayacaktır. Kıdemli ve uzman genç arasındaki en büyük fark, bağımsız olarak öğrenme yeteneğidir.



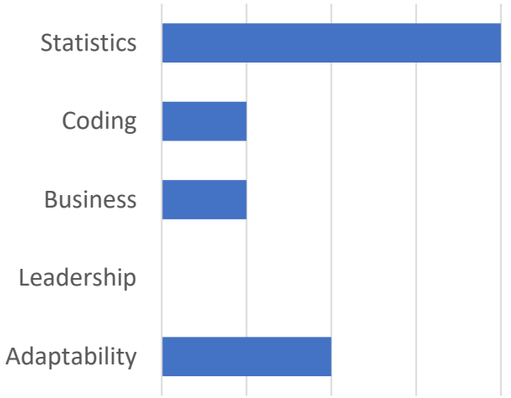
Principal data scientist (P) – Tıpkı kıdemli bir veri bilimcisi gibi, aynı zamanda bir ekibi ve bir projeyi yönetme deneyimine sahip. Bulunması zor)

Data scientist archetypes: danger

Business intelligence analyst (B) – İşletme ve ona güç veren veriler hakkında çok şey anlar. İstatistikler veya verilerle ne yapılacağı hakkında pek bir şey bilmiyor. Uygun rehberlik olmadan tehlikeli olabilir.



Academic (A) – Yüksek lisans / doktora derecesine sahip ve çok fazla iş tecrübesi yok. İlginç problemler hakkında düşünmeyi sever, ancak bir projeyi tamamlamak için sıradan işleri yapmaya zaman harcamaya daha az isteklidir (ve nasıl yapılacağını bilemeyebilir!).



Veri her yerde!

- Var olan dijital veri miktarı hızla artıyor, iki yılda bir ikiye katlanıyor ve yaşama şeklimizi değiştiriyor.
- Verilerin her zamankinden daha hızlı büyüdüğü görülmektedir, gezegendeki her insan için her saniye yaklaşık 1.7 megabayt yeni bilgi oluşturulacak, bu da en azından alanın temellerini bilmeyi son derece önemli hale getiriyor.
- Sonuçta, geleceğimiz burada yatıyor.
- Bu derste, Veri Bilimi, Büyük Veri ve Veri Analitiği arasında ne olduğuna, nerede kullanıldığına, bu alanda profesyonel olmanız için ihtiyaç duyduğunuz becerilerinize ve beklentilerinize göre ayrım yapacağız.

Büyük Veri nedir?

- Hem yapılandırılmamış hem de yapılandırılmış çok sayıda veriyi tanımlamak için kullanılan moda bir sözcüktür.
- Büyük Veri, var olan geleneksel uygulamalarla etkili bir şekilde işlenemeyen muazzam hacimli verileri ifade eder.
- Büyük Verinin işlenmesi, toplanmayan ham verilerle başlar ve çoğu zaman tek bir bilgisayarın belleğinde saklanması imkansızdır.
- Büyük Veri, depoyu günden güne doldurur.
- Büyük Veri, daha iyi kararlara ve stratejik iş hareketlerine yol açabilecek içgörüleri analiz etmek için kullanılabilen bir iş modelidir.
- Büyük veri, yüksek hacimli ve yüksek hızlı veya yüksek çeşitlilikte bilgi içeren varlıklarıdır. Gelişmiş içgörü, karar verme ve süreç otomasyonu ile bilgelik kazandıracak veri yığınıdır.
- Geçmişin öğreti birikimlerinin tutsaklığında yaşama devam etme yerine; kendisine aktarılan, bilgi ve tecrübe birikiminden, "Kendini yeniden var etme sürecini" başlatmaktır.

Büyük Veri

- Gelişen bilgi ve iletişim teknolojilerinin kapsamında kabul edilen internet teknolojileri; web sayfaları, bloglar, sosyal medya uygulamaları, algılayıcılar ve daha pek çok veri toplayan cihaz ve uygulamalar sürekli verileri toplamaktadır.
- Toplanan veriler pek çok alanda araştırmalarda kullanılabilmektedir.
- Veriyi toplama, işleme, kullanıcılara hazır hale getirme, erişime sunma, saklama, analiz etme gibi aşamalarda pek çok farklı teknikler kullanılabilmektedir.
- Verilerin günümüzde hız, çeşitlilik, kapasite (hacim) açısından büyük artış göstermesi ve bu artışa teknolojinin de destek vererek, yeni çözümler üretmesi ile birlikte "Büyük Veri" kavramı ortaya çıkmıştır.
- Günümüzde analizler yapılarak yeni bilgilerin üretilmesi ve bu bilgilerin farklı ihityaçlarda kullanılması ihtiyacı doğmuştur. Günümüzde pek çok büyük teknoloji şirketi büyük veri konusunda çok büyük yatırımlar yapmaktadır.

Veri Bilimi

- Veri bilimi, verilerden bilgi elde etmek ve öngörüde bulunmak için bilimsel yöntemleri, süreçleri, algoritmaları ve sistemleri kullanan çok disiplinli bir alandır.
- Veri bilimi büyük verilerle ilişkilidir.
- Veri bilimi, gerçek olayları verilerle anlamak ve analiz etmek için olasılık matematiğini, istatistik veri analizini, makine öğrenimini ve ilgili yöntemlerini birleştirmek için kullanılan bir kavramdır.

Veri Bilimi nedir?

- Yapılandırılmamış ve yapılandırılmış verilerle uğraşan Veri Bilimi, veri temizleme, hazırlama ve analiz ile ilgili her şeyi içeren bir alandır.

- Veri Bilimi, istatistik, matematik, programlama, problem çözme, verileri ustaca yöntemlerle yakalama, olaylara farklı bakma yeteneği geliştirme ve verileri temizleme, hazırlama ve hizalama etkinliğinin birleşimidir.

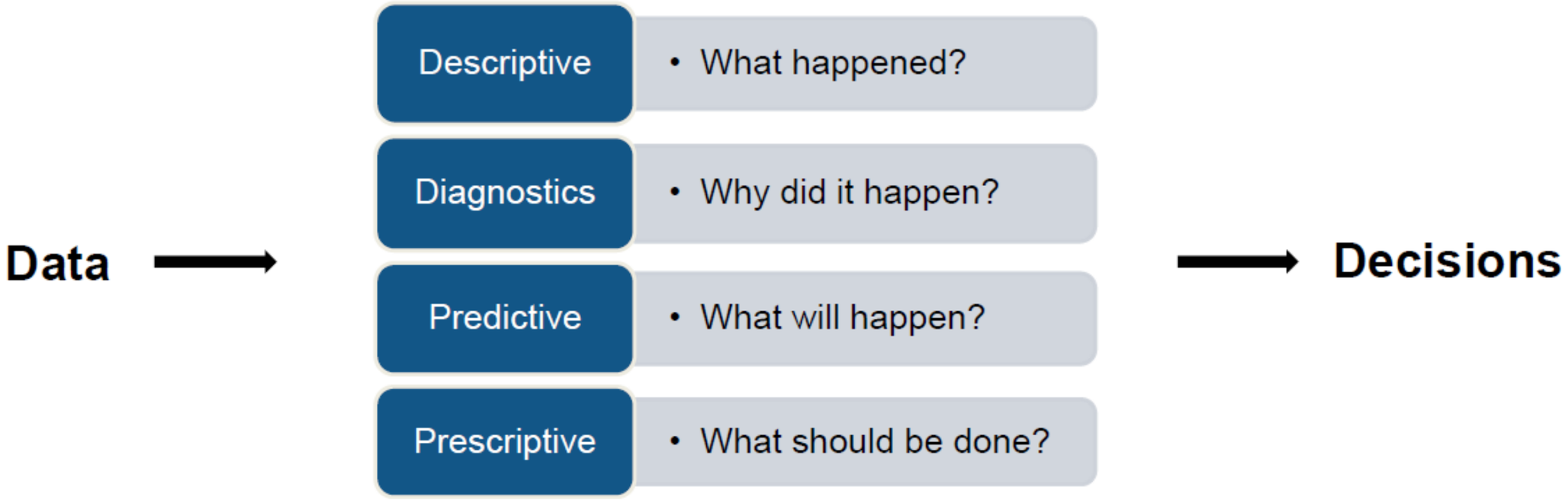
- Basit bir ifadeyle, verilerden içgörüler (derin anlamlar) ve bilgiler çıkarmaya çalışırken kullanılan teknikler şemsiyesidir.

Veri Bilimi

- Veri Biliminde projenin planlaması yapılır.
- Doğru araçlara, verilere, bilgi kaynaklarına erişim sağlanır.
- Takıma dahil olanlarla gerekli bilgi alışverişi yapılır.
- Planlamaz, iş bölümleri, performans, hangi verilere ihtiyaç olduğu, model yaklaşımları, izleyecek yol ile ilgili yapılan araştırmalar detaylı olarak paylaşılır.
- Geri bildirimler alınır. Elde edilen bütün bilgiler sisteme aktarılır.
- Bilgiler görselleştirilir.
- Veri analitiğinde kullanılacak veriler keşfedilir, analiz edilir.
- Modeller geliştirilir.
- Algoritmalar oluşturulur.
- Kodlar yazılır.
- İş süreçleri detaylı dökümanlandırılır. Süreklilik sağlanır. Çalışmaların herhangi bir aşamasında nerede, ne yapıldığının bilinmesi ve başka birinin kodu kolaylıkla anlayabilmesi için kurumsal dökümantasyon saklama önemldir.
- Hataları gidermek için test aşaması yapılmalıdır.
- Geri bildirim ve görüntüleme yapılır. Doğru çalışıp çalışmadığı geri bildirimlerle kontrol edilir.

What is Data Analytics?

Turn large volumes of complex data into actionable information



Tanım, tanı, tahmin, kural

Veriden Faydalı Bilgi Elde Etmek

- **Betimleyici Analitik:** “Ne olmuş?” sorusuna yanıt aranır. Veriyi betimlediğimizde mod, medyan, standart sapma veya görselleştirme teknikleriyle basit raporlar oluşturduğumuzda betimleyici analitik yapmış oluruz.

- **Teşhis Tanı Analitiği:** “Neden, Neden olmuş, Nasıl olmuş?” sorularının yanıtını verir. Betimledikten sonra görmüş olduğumuz durumun neden olduğunu sorgular, yani teşhis tanı analitiği yapmış oluruz.

- **Tahminsel Veri Analitiği:** “Ne olacak?” sorusuna yanıt verir. Geleceksel tahmin yapmak için kullanılır.

- **Yönergeli Analitik:** “Nasıl olmalı, Ne olmalı?” sorularına yanıt verir. Olasılıkları tahmin ettikten sonra, başarıyı arttırmak için ne yapılmalı sorularına yanıt aranır.

CRISP (1)

- Bir veri bilimi projesine nereden başlanacağı, hangi adımların izlenmesi gerektiği, projenin aşamalarının çıktıları ve proje süresince ölçülebilir adımları CRISP-DM olarak kısaltılan yöntemle yönetilebilmektedir. CRISP-DM modeli veri madenciliği için CROSS Endüstriler Arası Standart İşleme Süreci olarak tanımlanmaktadır. Cross-Industry Standard Process for Data Mining (CRISP-DM) olarak kısaltılmıştır.
- **CRISP sürecinde kullanılan aşamalar:** İş anlayışı, veriyi anlamayı, verinin hazırlanması, modelleme, başarısının değerlendirme ve kullanıma sokulması olarak sıralanır.
- Problem bulma anlayışı: Problemin çözülmek üzere veri bilimi takımına bildirilmesi. Problemin bileşenleri ve fayda - zarar katsayılarının toplamından oluşur. Katsayıların otonom öğrenmesi makine öğrenmesinin temelini oluşturur.
- $Y = x1*b1 + x2*b2 + \ldots + xp*bp$
- **Veri yığınındaki mesajı anlamak:** Veri kaynaklarından veri elde edildiğinde, verinin anlaşılıp incelenmesi gerekir. Ortalama, standart sapma, özetle istatistiksel sonuçlar çıkarılması gerekir.
- **Verinin hazırlanması:** Matematiksel modelleme yapılmadan önce veri setinin hazırlanması gerekir.

CRISP (2)

Veri setindeki yapısal bozuklukların düzenlenmesi gereken veriler:

- **Missing value:** Kaybolmuş değer anlamına gelir, eksik veridir.
- **Aykırı değer:** Verinin yapısının oldukça dışında kalan değerlere denir.
- **Gürültü:** Bozuk değer anlamına gelir.

Bu aşamada verinin bozukluğu giderilir.

CRISP (3)

- **Matematiksel modelleme, algoritmalar:** Veri içerisinde yer alan yapıların algoritmalara, fonksiyonlara öğretilmesi olayıdır. Veri setinin içindeki yapıların etkilerinin bulunması gerekir. Bu sebeple çoklu doğrusal regresyon modeli önemlidir. Veri setine çoklu doğrusal regresyon modelini verip aralarında ki ilişkiyi modellediğimizde;
- $Y = x1*b1 + x2*b2 + \ldots + xp*bp$
- fonksiyonunu bulmaya çalıştığımızı biliyoruz.
- Y: tahmin fonksiyonu
- **X değişkenleri:** veri setinin içinden öğrenecek olduğumuz değişkenlerin etki düzeylerini ve yönlerini belirtecek olan katsayılardır. Yönünü ve şiddetini etkileyecek katsayıları ortaya çıkarır.
-
- **Değerlendirme (Model başarı değerlendirme):** Kurulan modelin başarısının ve doğruluğunun değerlendirilmesi gerekir. Makine öğrenmesi modeli kurup değerlendirme işlemini ele alalım. Değerlendirme derken performans ele alınır. Modele, çıktısının ne olduğunu bildiğimiz değerler verilir ve tahmini olarak ne değer vereceği, gerçek değer ile modelin tahmin ettiği değer karşılaştırılır. Başarısı değerlendirilir. Gerçek değerler ile tahmin edilen değerler arasındaki farkları minimum seviyeye indirgemek için farklı algoritmalar denenebilir.
-
- **Kullanıma Sokma:** Nihai olarak karar verilen şekliyle, canlı sistemlere entegrasyon yapılır. Bu excel tablosu, veritabanı tablosu sekliyle ya da rapor şeklinde kullanıma sokulabilir.

Data science and AI

Yapay Zeka, akıllı makineler yapma bilimi ve mühendisliği olarak tanımlanır. AI, bir insan gibi çevrelerinden girdi alabilen ve buna dayalı eylemler gerçekleştiren akıllı sistemlerin araştırılması ve tasarımı ile ilgilenen Bilgisayar Bilimi dalıdır.

At the end of this chapter, students will have a brief introduction applications of data science in AI. They will know

- Applications of data science
- Analytics on text data
- Analytics on image data
- Overview of AI

Data Science and its role in Data analytics

Data Science

- New kinds of data
- Interdisciplinary
- Data as product
- New methods for making sense to data

Interdisciplinary

- DS, verilerden anlam çıkarma ve veri ürünleri oluşturma amacıyla matematik, istatistik, veri mühendisliği, örüntü tanıma ve öğrenme, gelişmiş bilgi işlem, görselleştirme, belirsizlik modelleme, veri ambarı ve yüksek performanslı bilgi işlemi içerir.
- Veri bilimi alanı, sosyal bilimler ve istatistik, bilgi ve bilgisayar bilimi ve tasarım alanlarının kesiştiği noktada ortaya çıkmaktadır.

New kinds of data

- Büyük veri yığınından bilgi çıkarma
- Bilgi keşfi ve veri madenciliği (KDD) olarak da bilinen, saha veri madenciliği ve tahmine dayalı analitiğin bir devamı olan, yapılandırılmış veya yapılandırılmamış büyük veri hacimlerinden bilginin çıkarılması.
- "Yapılandırılmamış veriler" e-postaları, videoları, fotoğrafları, sosyal medyayı ve kullanıcı tarafından oluşturulan diğer içerikleri içerebilir.
- İlk olarak, Veri Biliminin “veri” kısmı olan ham madde giderek daha heterojen ve yapılandırılmamış hale geliyor. İkincisi, bilgisayarlar verileri otomatik olarak yorumlayarak, onları anlamlandırma sürecinde aktif faktörler haline getirir.

Data as product

- "Veri bilimi" ile kastettiğimiz şey, yalnızca verileri kullanmak değildir.
- Bir veri uygulaması, değerini verinin kendisinden alır ve sonuç olarak daha fazla veri oluşturur.
- Bu sadece veri içeren bir uygulama değildir; bu bir veri ürünüdür.
- Veri bilimi, veri ürününün oluşturulmasını sağlar.

New methods for making sense to data

- Veri bilimi, bilginin nereden geldiğinin, neyi temsil ettiğinin ve iş ve BT stratejilerinin oluşturulmasında nasıl değerli bir kaynağa dönüştürülebileceğinin incelenmesidir.
- Özünde, veri bilimi, büyük miktarda veriyi analiz etmek ve onlardan bilgi elde etmek için otomatik yöntemler kullanmayı içerir.

Veri Bilimi manzarası

- Alanlar
- Teknikler
- Yaklaşımlar
- Nesneler

Alanlar

- Nanoteknolojiler
- Fizik
- Robotik
- Matematik
- İstatistik
- Bilgi teorisi
- Bilgi Teknolojisi
- Yapay zeka

- Nanotechnologies
- Physics
- Robotics
- Mathematics
- Statistics
- Information theory
- Information technology
- AI

Teknikler

- Sinyal işleme
- Olasılık modelleri
- Makine öğrenme
- İstatistiksel öğrenme
- Veri madenciliği
- Veri tabanı
- veri mühendisliği
- Desen tanıma
- Görselleştirme
- Tahmine dayalı analitik
- Belirsizlik modelleme
- Veri depolama
- Veri sıkıştırma
- Bilgisayar Programlama
- Yüksek Performanslı Bilgi İşlem

Yaklaşımlar

- Tahmine dayalı modellerin geliştirilebileceği verilerdeki kalıpları ortaya çıkarmak için kullanılan bir yapay zeka dalı olan makine öğreniminin gelişimi, veri biliminin büyümesini ve önemini artırdı.

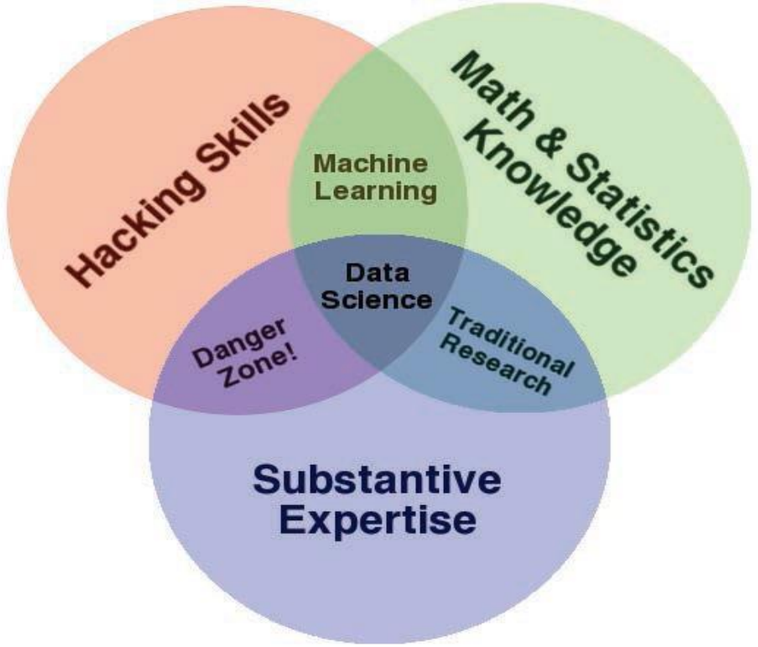
Nesneler

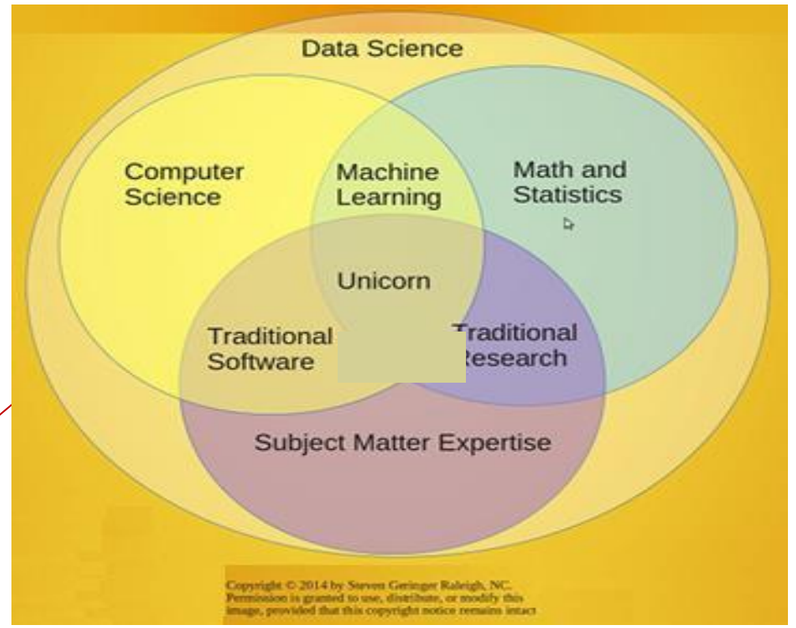
- Büyük Veriye ölçeklenen yöntemler, disiplinin genellikle bu tür verilerle sınırlı olduğu düşünülmese de, veri biliminde özellikle ilgi çekicidir.

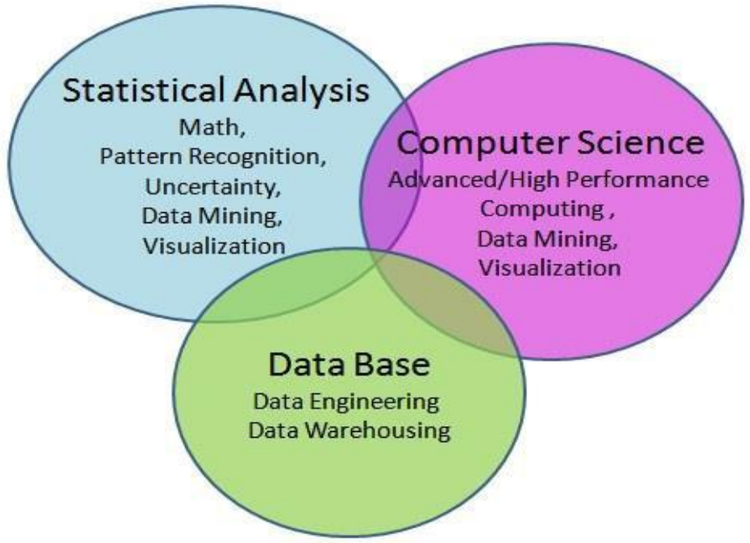
Veri Bilimcisi kimdir?

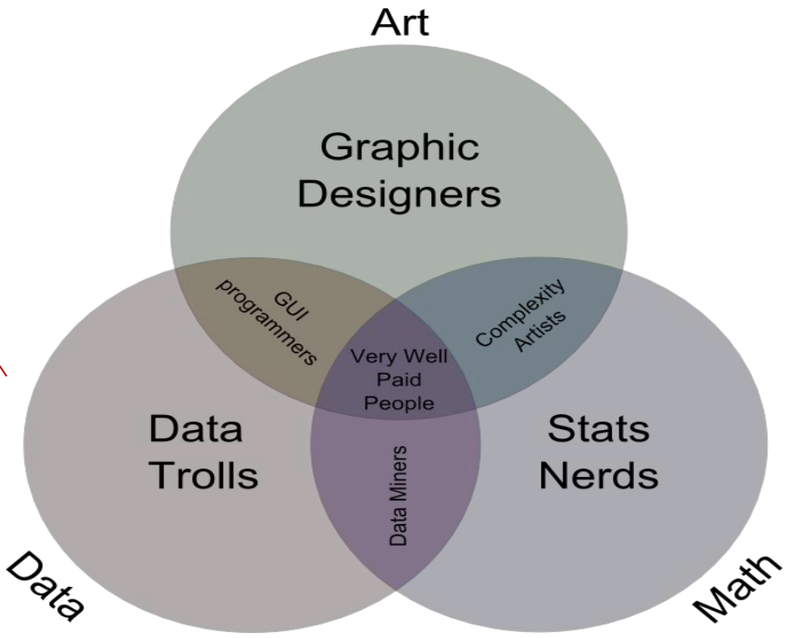
- Görevler: Gelişmiş analitik becerilere ek olarak, bu kişi aynı zamanda büyük, çeşitli veri kümelerini entegre etme ve hazırlama, özel veritabanı ve bilgi işlem ortamları tasarlama ve sonuçları iletme konusunda da yetkindir.
- Misyon, Amaç: Bir veri bilimcisi, iş problemlerini modellemeye ve verileri anlama ve hazırlamaya yardımcı olacak özel sektör bilgisine sahip olabilir veya olmayabilir.
- Öz geçmiş: Veri bilimcisi, iş zekası (BI) analistleri ve istatistikçilerinden farklı, ancak benzerleri olan yeni bir rol olarak ortaya çıktı.
- Yetenekler: Veriden değer yaratmak, bir dizi yetenek gerektirir: veri entegrasyonu ve hazırlığından, özel bilgi işlem/veritabanı ortamlarının mimarisine, veri madenciliği ve akıllı algoritmalara kadar
- Sorumluluk: İstatistiksel, algoritmik, madencilik ve görselleştirme tekniklerini kullanarak karmaşık iş problemlerini modellemekten, iş içgörülerini keşfetmekten ve fırsatları belirlemekten sorumlu kişi.
- Farkındalıklar: Veri bilimcileri, özellikle "büyük veri"den kavramlar üretmede çok değerli olabilir; ancak teknik ve iş becerilerinin benzersiz kombinasyonu, artan talepleriyle birlikte onları bulmayı veya geliştirmeyi zorlaştırıyor.

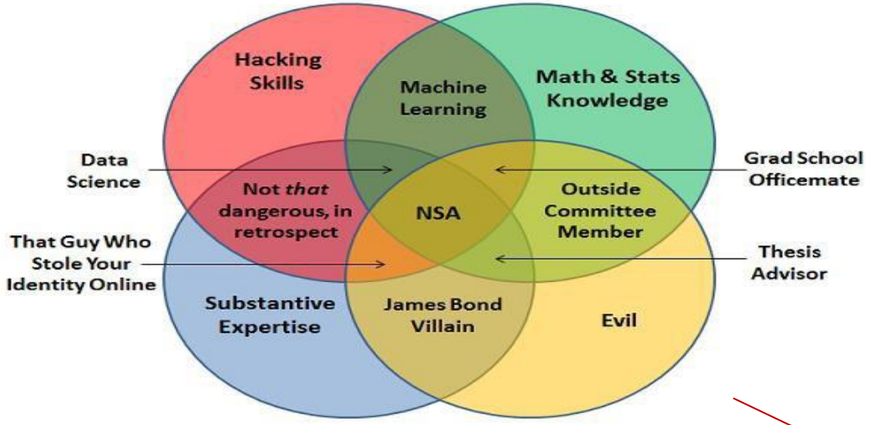
Venn diagram

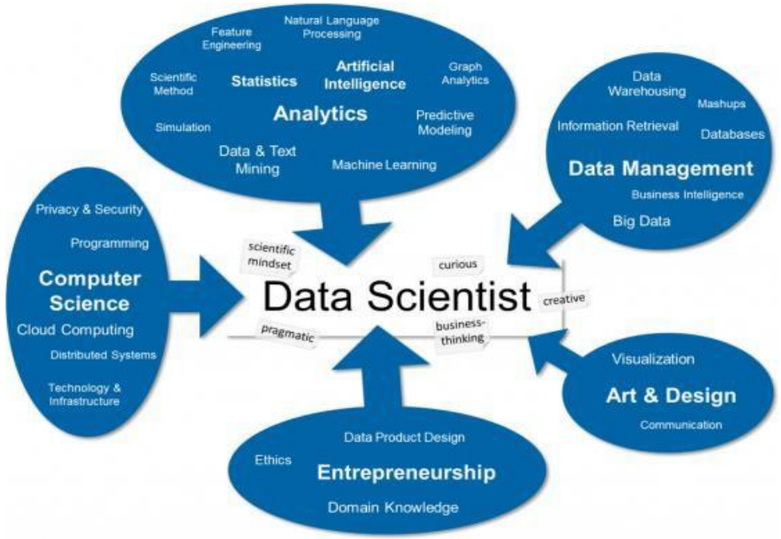


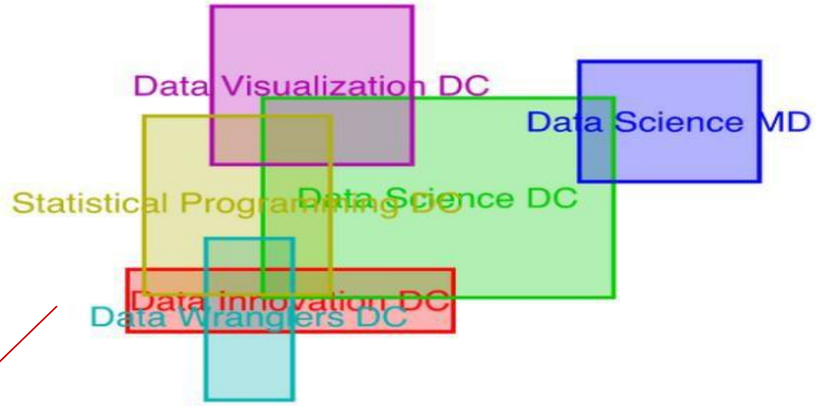


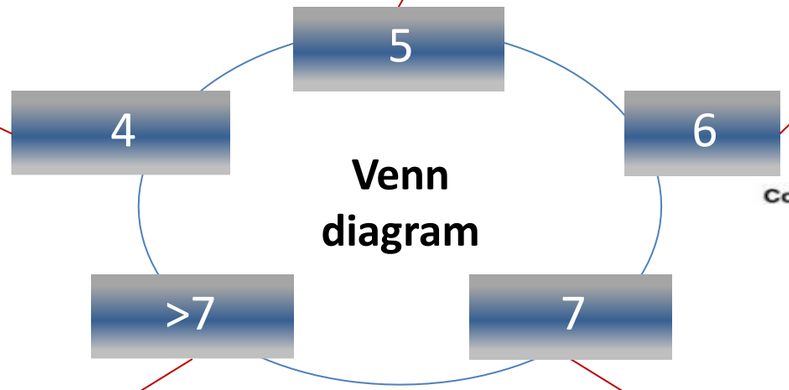


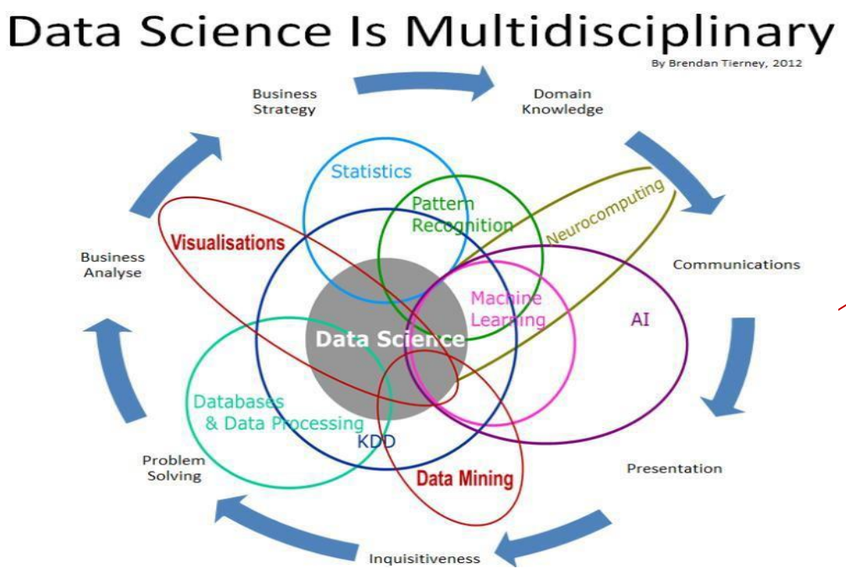


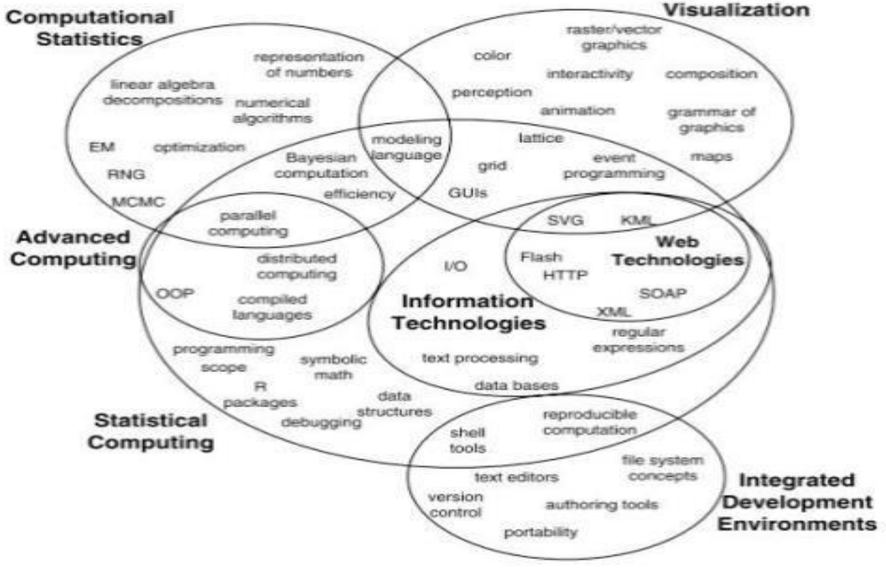

Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
Grad School Officemate
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

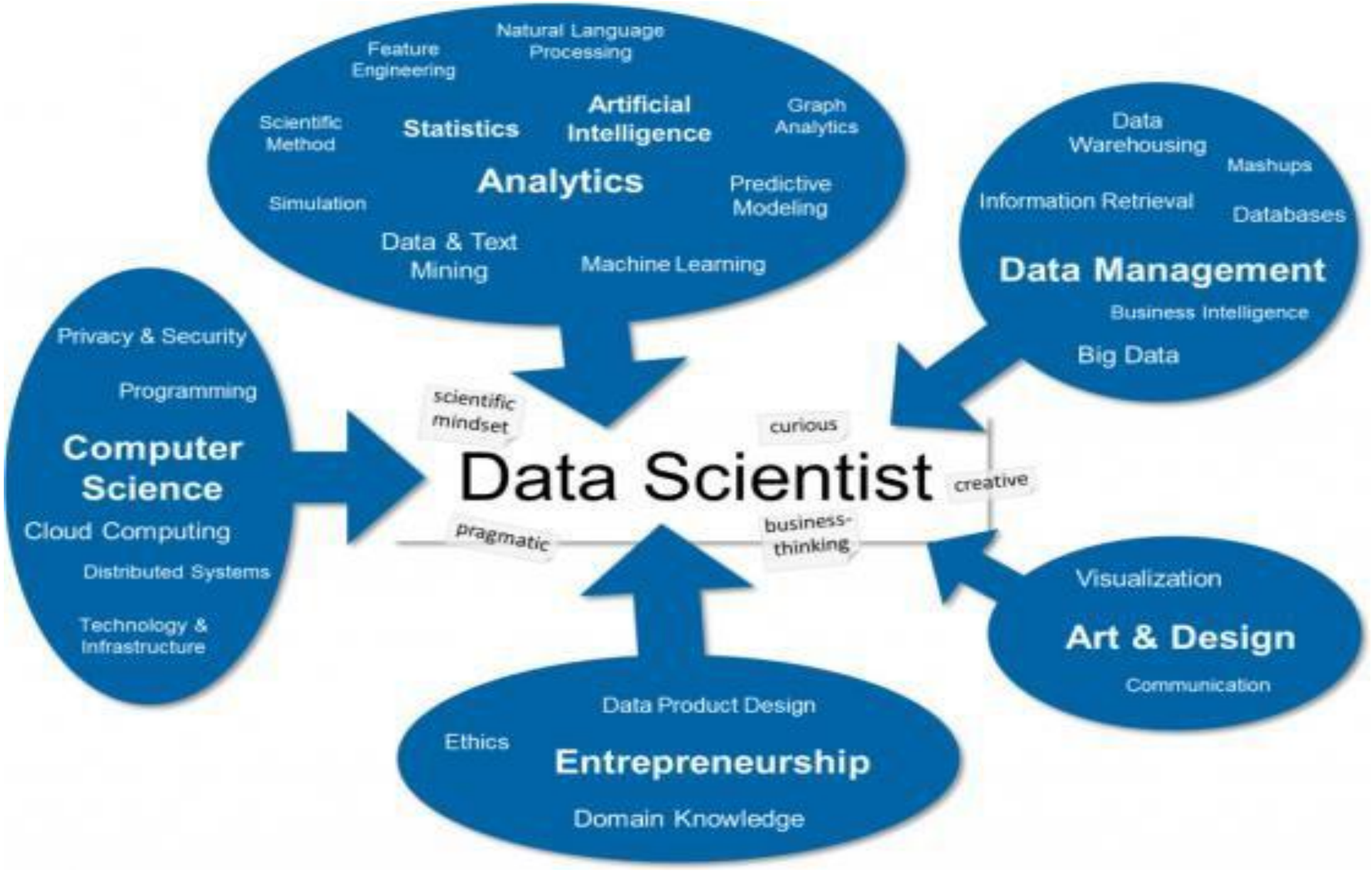

Feature Engineering
Natural Language Processing
Scientific Method
Statistics
Artificial Intelligence
Graph Analytics
Analytics
Predictive Modeling
Simulation
Data & Text Mining
Machine Learning
Data Warehousing
Mashups
Information Retrieval
Databases
Data Management
Business Intelligence
Big Data
Privacy & Security
Programming
Computer Science
Cloud Computing
Distributed Systems
Technology & Infrastructure
scientific mindset
curious
Data Scientist
creative
Pragmatic
business-thinking
Visualization
Art & Design
Communication
Data Product Design
Ethics
Entrepreneurship
Domain Knowledge


Data Visualization DC
Data Science MD
Statistical Programming DC
Data Science DC
Data Innovation DC
Data Wranglers DC

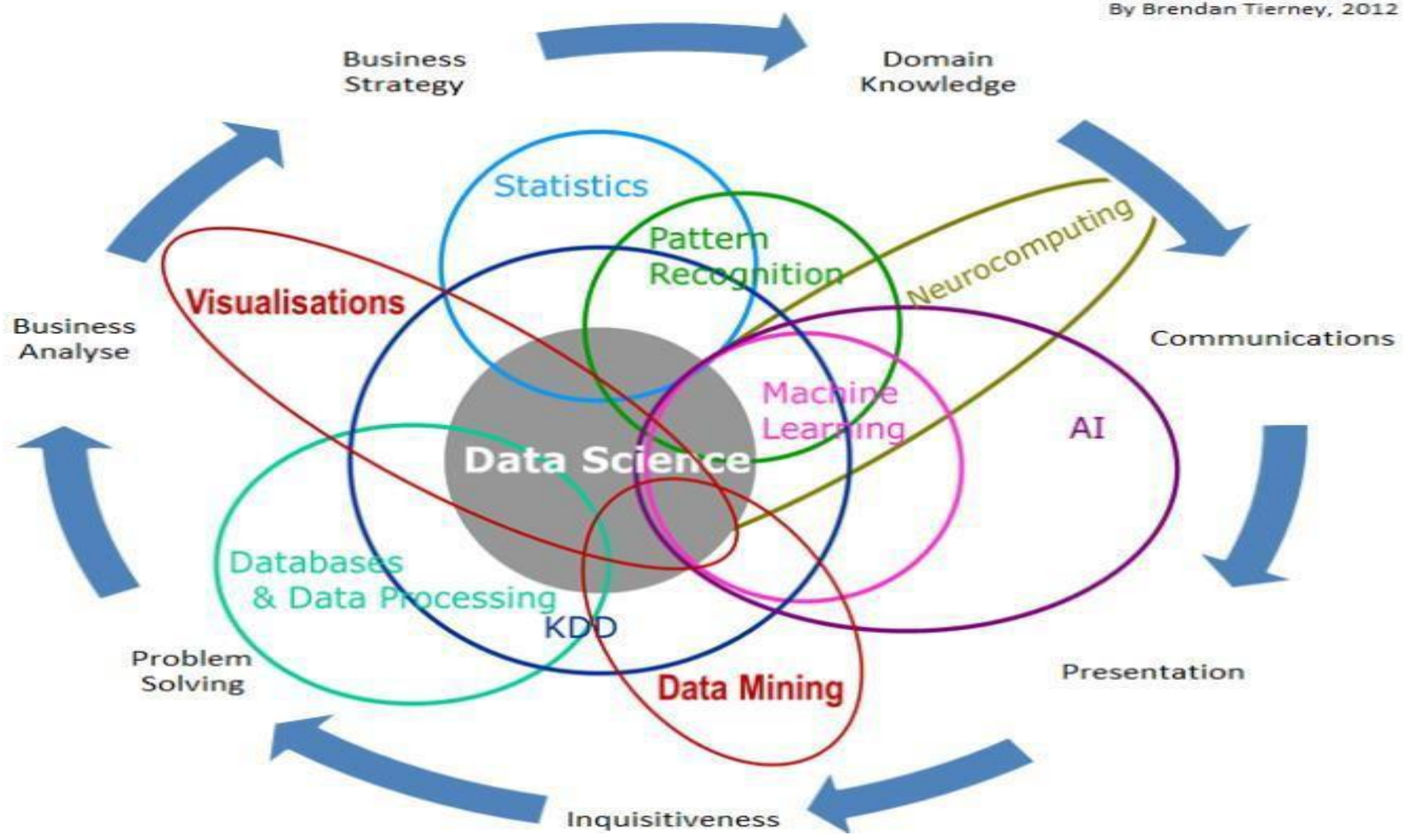

5
4
6
Venn diagram
>7
7

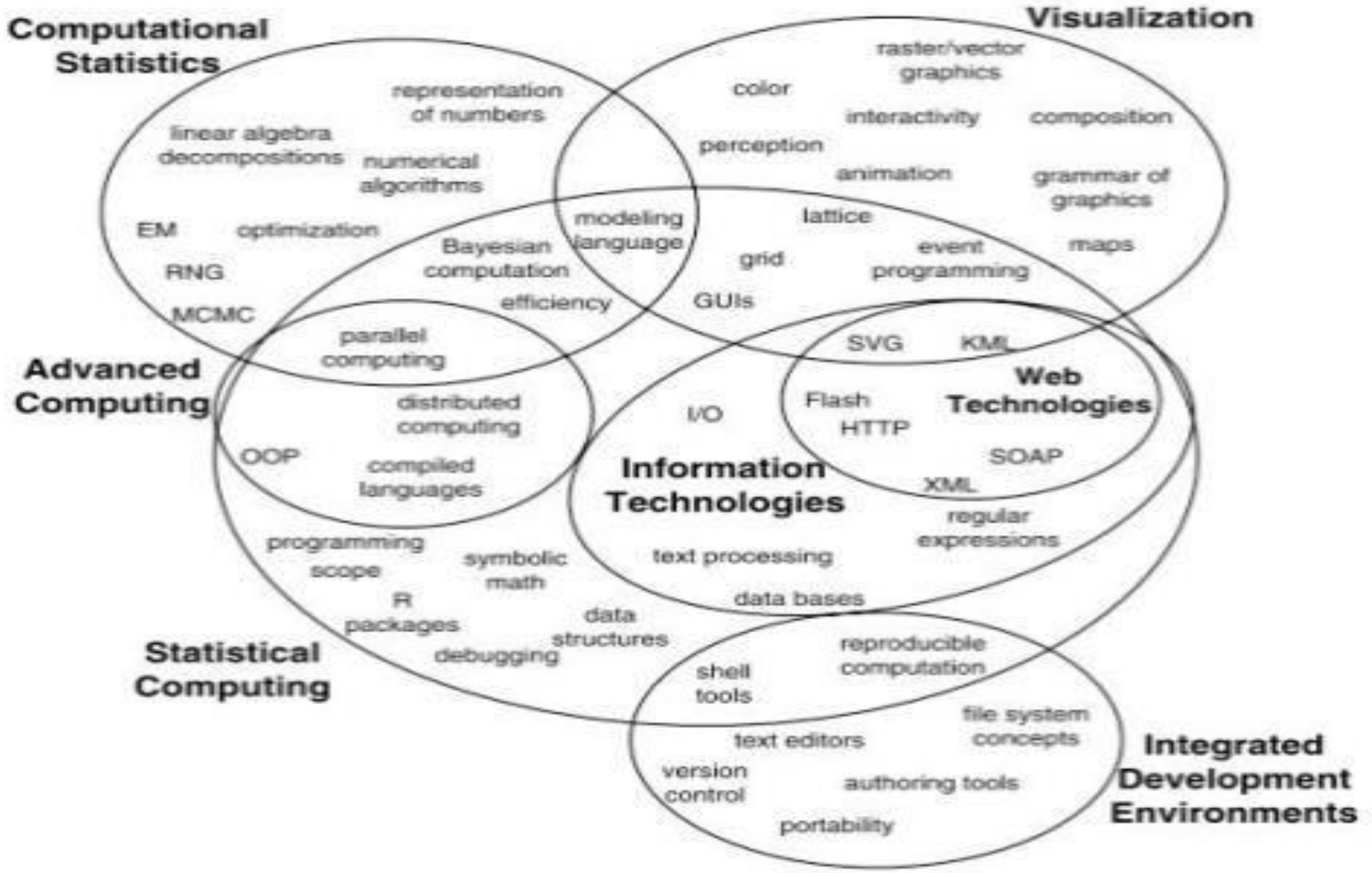

Data Science Is Multidisciplinary
By Brendan Tierney, 2012
Business Strategy
Domain Knowledge
Business Analyse
Communications
Problem Solving
Presentation
Inquisitiveness
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Data Mining

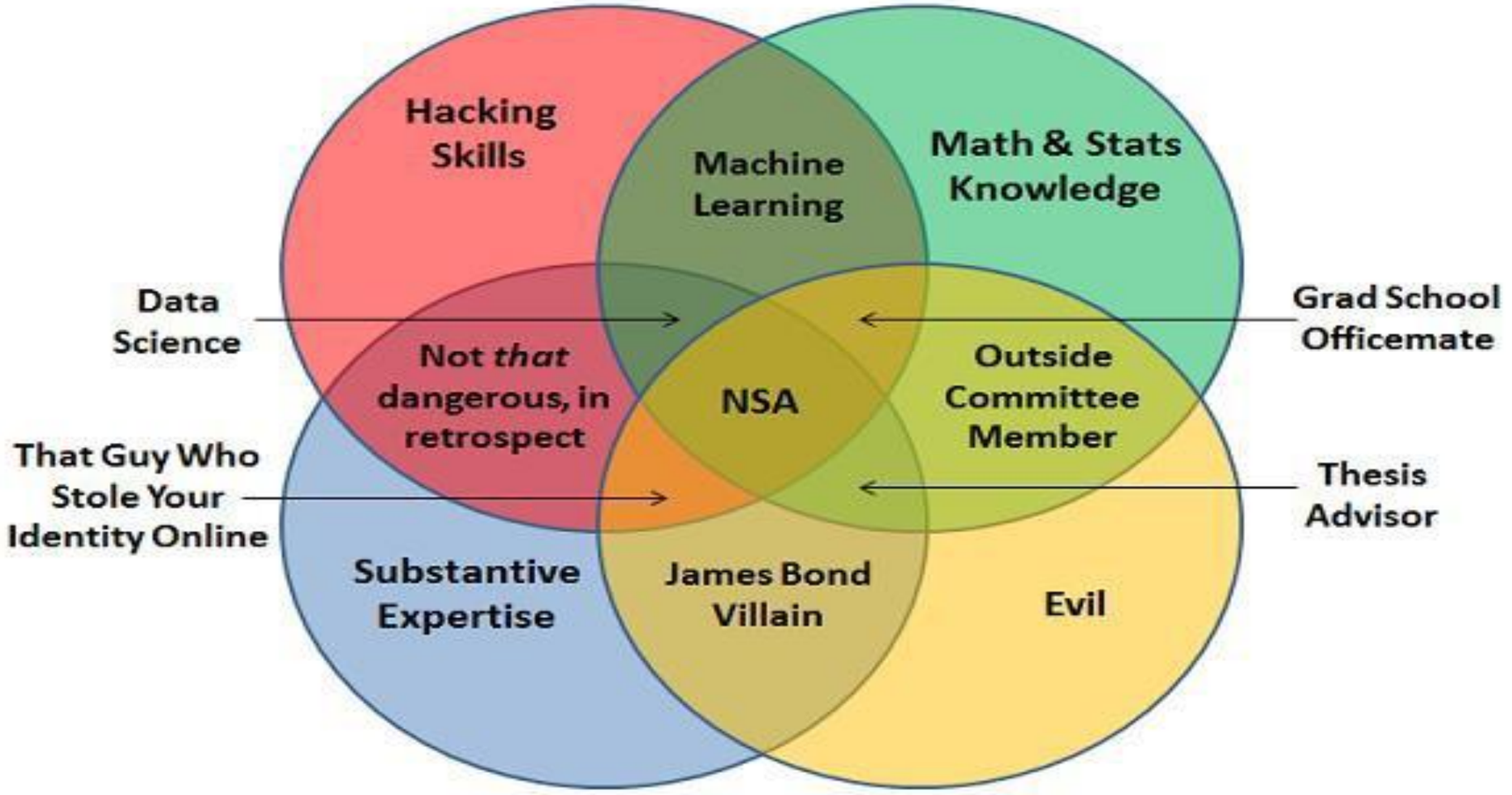

Computational Statistics
Visualization
Advanced Computing
Statistical Computing
Information Technologies
Web Technologies
Integrated Development Environments
representation of numbers
linear algebra decompositions
numerical algorithms
color
raster/vector graphics
interactivity
composition
perception
animation
grammar of graphics
EM
optimization
RNG
MCMC
Bayesian computation
modeling language
lattice
grid
event programming
maps
efficiency
GUIs
parallel computing
distributed computing
OOP
compiled languages
SVG
KML
Flash
HTTP
I/O
SOAP
XML
programming scope
symbolic math
R packages
debugging
data structures
text processing
data bases
regular expressions
reproducible computation
shell tools
text editors
file system concepts
version control
authoring tools
portability


Natural Language Processing
Feature Engineering
Scientific Method
Statistics
Artificial Intelligence
Graph Analytics
Analytics
Simulation
Predictive Modeling
Data & Text Mining
Machine Learning
Data Warehousing
Mashups
Information Retrieval
Databases
Data Management
Business Intelligence
Big Data
Privacy & Security
Programming
Computer Science
Cloud Computing
Distributed Systems
Technology & Infrastructure
scientific mindset
curious
Data Scientist
creative
pragmatic
business-thinking
Visualization
Art & Design
Communication
Data Product Design
Ethics
Entrepreneurship
Domain Knowledge

Data Science Is Multidisciplinary




Computational Statistics
representation of numbers
linear algebra decompositions
numerical algorithms
EM
optimization
RNG
MCMC
Visualization
raster/vector graphics
color
interactivity
composition
perception
animation
grammar of graphics
lattice
maps
event programming
grid
GUIs
modeling language
Bayesian computation
efficiency
Advanced Computing
parallel computing
distributed computing
OOP
compiled languages
Web Technologies
SVG
KML
Flash
HTTP
SOAP
XML
I/O
Information Technologies
regular expressions
text processing
data bases
programming
scope
symbolic math
R packages
data structures
debugging
Statistical Computing
reproducible computation
shell tools
file system concepts
text editors
version control
authoring tools
portability
Integrated Development Environments


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

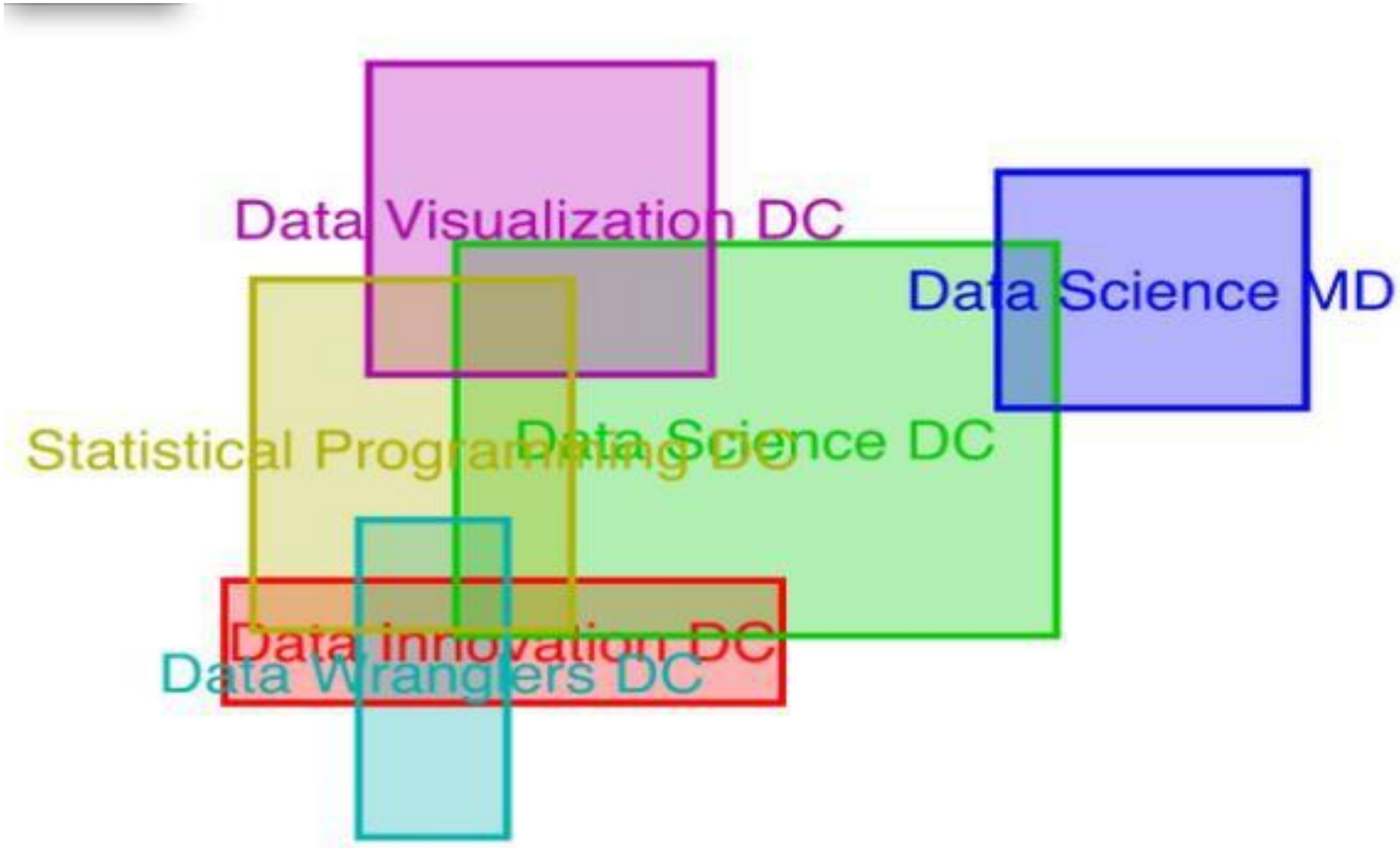
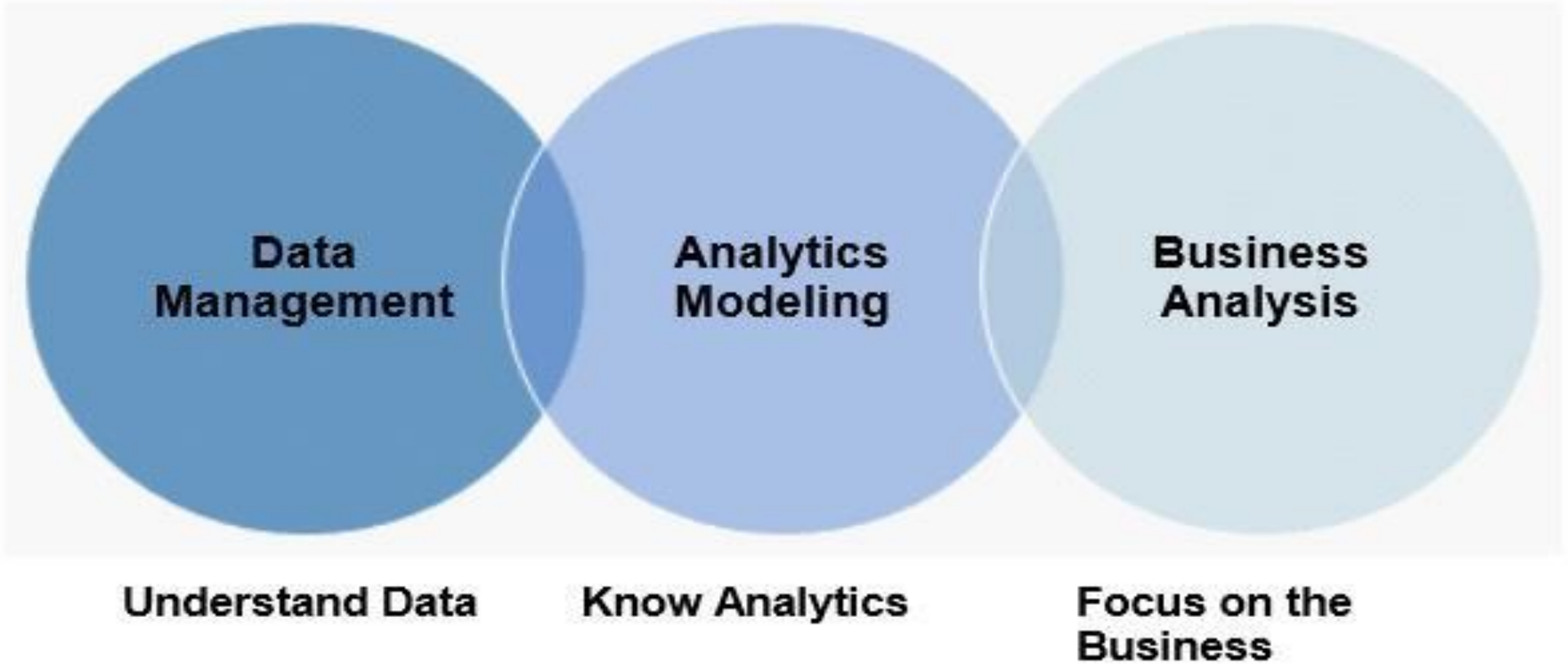

Data Visualization DC
Data Science MD
Statistical Programming DC
Data Science DC
Data Innovation DC
Data Wranglers DC

Figure 3. Core Data Scientist Skills



Source: Gartner (March 2012)

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Is Data Science a maturity science?

Types of domain dealt by an intellectual enterprises:

(a) topics (facts, data, problems, phenomena, observations, and the like)

(b) methods (techniques, approaches, and so on)

(c) theories (hypotheses, explanations, and so forth)

Feature of a new discipline:

(a) To represent an autonomous field (*unique topics*)

(b) To provide an innovative approach to both traditional and new philosophical topics (*original methodologies*);

(c) To stand beside other disciplines, offering the systematic treatment of its own conceptual foundations (*new theories*).

If a discipline attempts to innovate in more than one of these domains simultaneously is premature, as detaches itself too abruptly from the normal and continuous thread of evolution of its general field (Stent 1972).

As everyone's concern is nobody's business → crossroad of
- technical matters
- theoretical issues
- applied problems
- conceptual analyses

→ to be anyone's own area of specialisation → **Transdisciplinary** (like cybernetics or semiotics) or **interdisciplinary** (like biochemistry or cognitive science)?

Short History of Data Science

(Loosely based on Gil Press version)

1962

J. W. Tukey

The Future of Data Analysis

"I have come to feel that my central interest is in *data analysis*... Data analysis, and the parts of statistics which adhere to it, must...


THE ANNALS of MATHEMATICAL STATISTICS

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

The Future of Data Analysis ..... JOHN W. TUKEY 1

Infinitely Divisible Distributions: Recent Results and Applications ..... MAREK FISZ 68

Limit Theorems for Randomly Selected Partial Sums ..... PATRICK BILLINGSLEY 85

A Law of Large Numbers for the Maximum in a Stationary Gaussian Sequence ..... SIMEON M. BERMAN 93

A "Renewal" Limit Theorem for General Stochastic Processes ..... V. E. BENEŠ 98

Mixtures on Markov Processes ..... DAVID A. FREEDMAN 114

Some Aspects of the Emigration-Immigration Process ..... HAROLD RUBEN 119

A Transient Discrete Time Queue with Finite Storage ..... JOHN R. KINNEY 130

A General Model for the Reliability Analysis of Systems Under Various Preventive Maintenance Policies ..... B. J. FLEHINGER 137

Approximations to the Moments of the Sample Median ..... M. M. SIDDIQUI 157

Recurrence Relations Between the PDF's of Order Statistics, and some Applications ..... K. S. SRIKANTAN 169

Enumeration of Linear Graphs for Mappings of Finite Sets ..... JOHN RIORDAN 178

Asymptotic Power of Certain Test Criteria (Based on First and Second Differences) for Serial Correlation Between Successive Observations ..... A. R. KAMAT AND Y. S. SATHE 186

Ante-Dependence Analysis of an Ordered Set of Variables ..... K. R. GABRIEL 201

A Classification Problem in which Information about Alternative Distributions is Based on Samples ..... BERT E. ELLISON 213

(Continued on back cover)

Vol. 33, No. 1 — March, 1962

take on the characteristics of science rather than those of mathematics... data analysis is intrinsically an empirical science"

1974

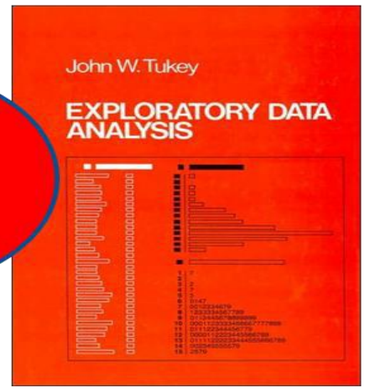
Peter Naur

Concise Survey of Computer Methods

"*[Data is]* a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."

J. W. Tukey

...arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis "can—and should—proceed side by side."



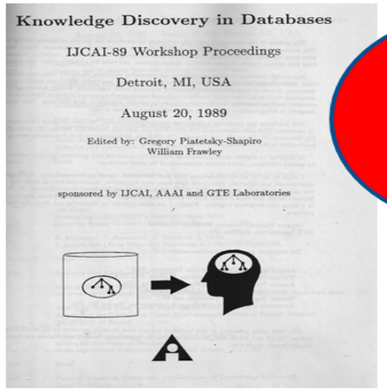
1977

I S I

1° Section of The International Association for Statistical Computing (IASC)

"It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

G.Piatetsky-Shapiro


Knowledge Discovery in Databases

IJCAI-89 Workshop Proceedings

Detroit, MI, USA

August 20, 1989

Edited by: Gregory Piatetsky-Shapiro
William Frawley

sponsored by IJCAI, AAAI and GTE Laboratories

1989

First Knowledge Discovery in Databases (KDD) workshop

BusinessWeek

Cover story on "Database Marketing"

1994

"...Many companies were too overwhelmed by the sheer quantity of data to do anything useful with the information... Still, many companies believe they have no choice but to brave the database-marketing frontier."

1996

IFCS


IFCS-96

March 27-30,1996
International Conference
Center Kobe, Japan

Data Science, Classification and Related Methods

U. Fayyad *et al.*

From Data Mining to Knowledge Discovery in Databases

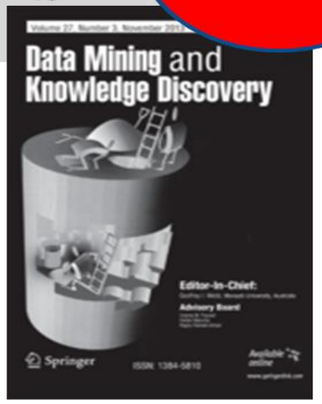
Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

For the first time, the term "**data science**" is included in the title of a conference

The journal "Data Mining and Knowledge Discovery" is launched

1997



C. F. J. Wu

From Statisticians to Data Scientist

…calls for statistics to be renamed data science and statisticians to be renamed data scientists

J. Zahavi

Born of Big Data?

"Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Scalability is a huge issue in data mining."

1999

…a plan "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'"

W. S. Cleveland

2001

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

L. Breiman

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.

"...management of data and databases in Science and Technology. The scope of the Journal includes descriptions of data systems, their publication on the internet, applications and legal issues."

2002

DATA SCIENCE Journal — CODATA

2003

Journal of Data Science

"By "Data Science", we mean almost everything that has something to do with data"

T. H. Davenport

"the emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making... "

Data Scientists: "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."

2005

NSF

Dataology

上海市数据科学重点实验室(复旦大学)

Shanghai Key Laboratory of Data Science, Fudan University

2007

The main research areas include fundamentals of data science, exploration of datanature, and data technologies and applications. Researchers are from Computer Science, Economics, Mathematics, Management, Journalism, Psychology, Chemistry, Philosophy, and so on.

As an open platform for data science research, Area 96 has invited a number of scholars to conduct joint scientific research and short term visiting.

2008

Skills, Role & Career Structure of Data Scientists & Curators

Data Scientists: "people who work where the research is carried out–or in close collaboration with the creators of the data"

2009

"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…".

H. Varian

The sexy job

"The nation needs to identify and promote the emergence of new disciplines and specialists expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data".

Interagency Working Group on Digital Data

K. D. Borne

Data Science & Astrophysic

"Training the next generation in the fine art of deriving intelligent understanding from data is needed for the success of sciences, communities, projects, agencies, businesses, and economies. ".

2009

M. Driscoll

Sexy skills

"with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity.".

N. Yau

New Fields for DS

" [a] new field that combines the skills and talents from often disjoint areas of expertise… [computer science; mathematics, statistics, and data mining; graphic design; infovis and human-computer interaction]"

2009

T. Sadkowsky

First DS group on LinkedIn

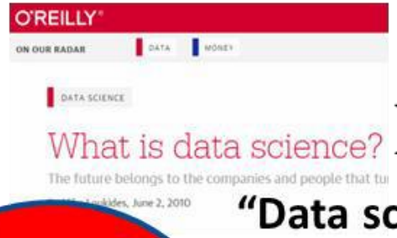
The 3 step OPD Data Science Process

K. Cukier

Data, Data Everywhere

"… a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data"

M. Loukides



2010

"Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution"

H. Mason

Data Science Taxonomy

2010

"In chronological order: Obtain, Scrub, Explore, Model, and iNterpret"

Data Science Venn Diagram

D. Conway

"a combination of computer hacking, data analysis, and problem solving "

All in one name

D. Smith

M. J. Graham

Data Science Epistemology

"Rules to follow. how data can be symbolized and communicated and what the relationships to physical space and time"

2011

"There is no widely accepted boundary for what's inside. and when I look around I see people with shared characteristics who don't fit into traditional categories. "

Counterpoints to four common data science criticisms

P. Warden

H. Harris

Carreer & eclectism

"Data Science is defined by its practitioners, that it's a career path rather than a category of activities"

D.J. Patil

Ultimate definition

"Those who use both data and science to create something new."

2012

Data Scientist: The Sexiest Job of the 21st Century

T. H. Davenport

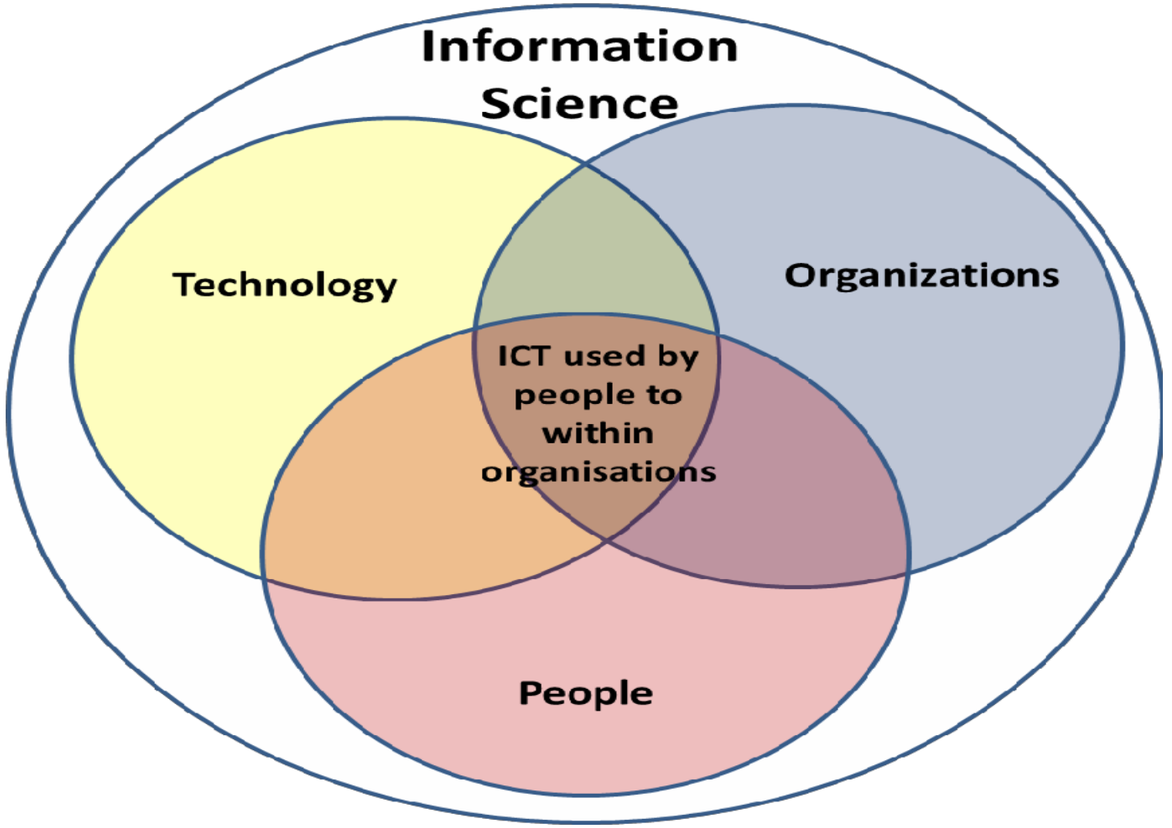
Steps to a Metaphisics of Data Science

- How does the Data Science in the context of the Knowledge Organization?
- What are its relations with other fields of scientific knowledge?
- Can DS be explained as part of the philosophy of science?

	Data	Information	Knowledge
Scientific context		Information Science	Knowledge Science
Philosophical context	Philosophy of Data	Philosophy of Information	Philosophy of Knowledge (Epistemology, Gnoseology)

Beyond Data Science?

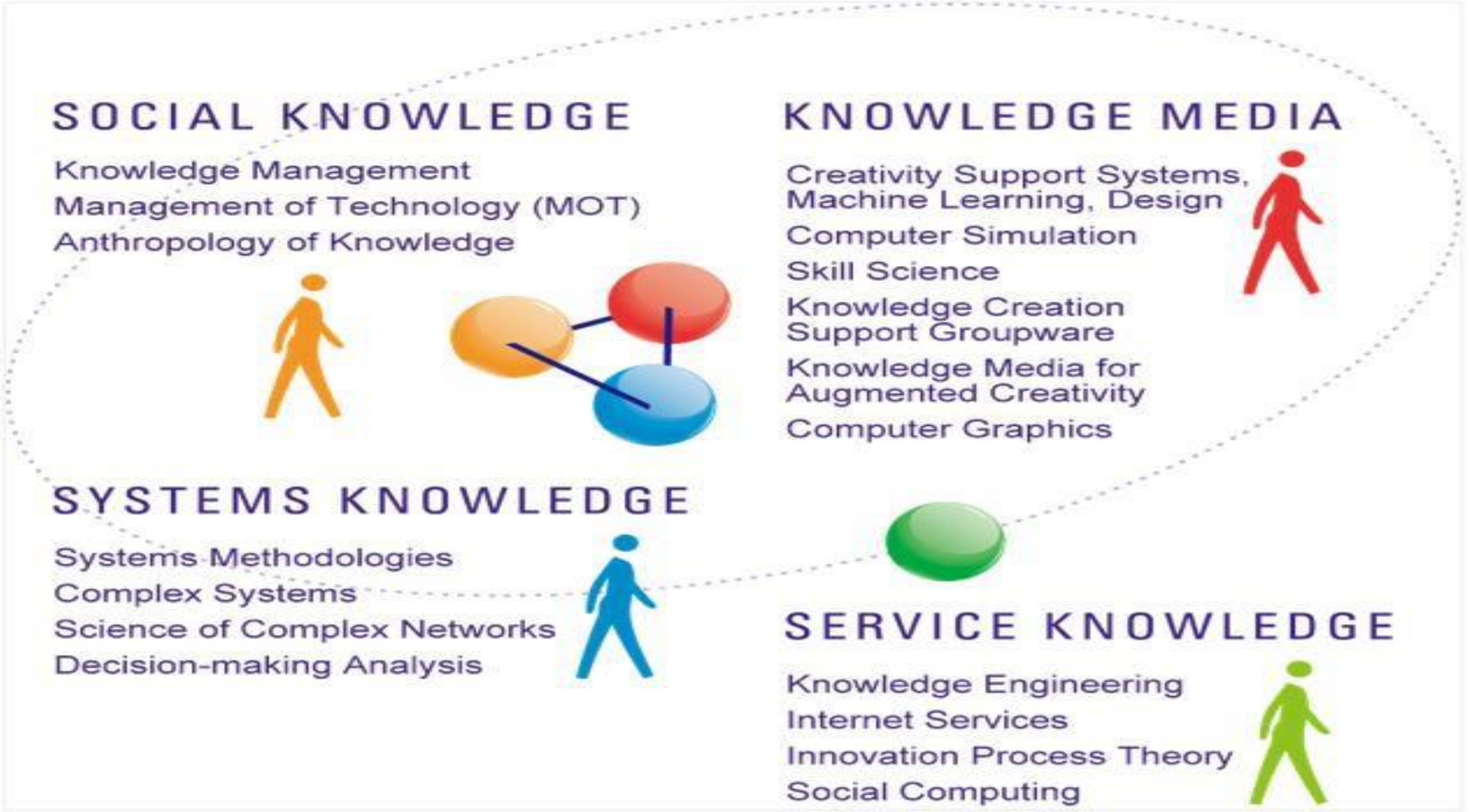
Information Science is the study of **information** and how it is used by people within **organisations**



Information Science sits at the intersection of technology, people, and organizations.
It is a distinct discipline and has a focus on Information and Communication Technologies (ICT) used by people to manage information within organisations.

Beyond Data Science?

The School of Knowledge Science consists of four major content areas.



Usage Notes

- A lot of slides are adopted from the presentations and documents published on internet by experts who know the subject very well.
- I would like to thank who prepared slides and documents.
- Also, these slides are made publicly available on the web for anyone to use
- If you choose to use them, I ask that you alert me of any mistakes which were made and allow me the option of incorporating such changes (with an acknowledgment) in my set of slides.

Sincerely,

Dr. Cahit Karakuş

cahitkarakus@gmail.com