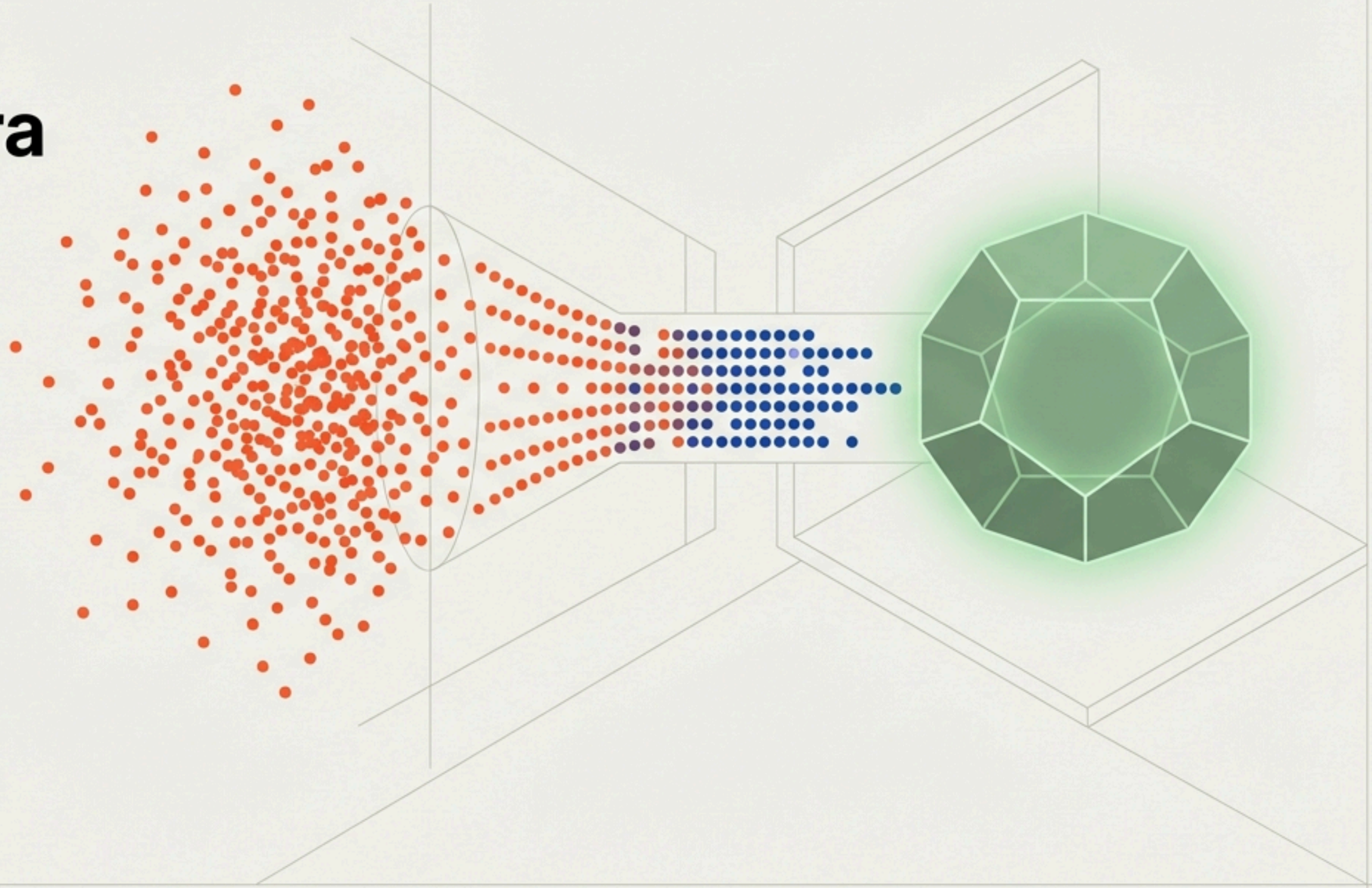
Veri Bilimi: Ham Veriden Akıllı Kararlara

Temeller, Süreçler ve Geleceğin Teknolojileri



Bu doküman, veri biliminin temel kavramlarını, yaşam döngüsünü ve sektörel uygulamalarını kapsayan kapsamlı bir rehber niteliğindedir.

NotebookLM

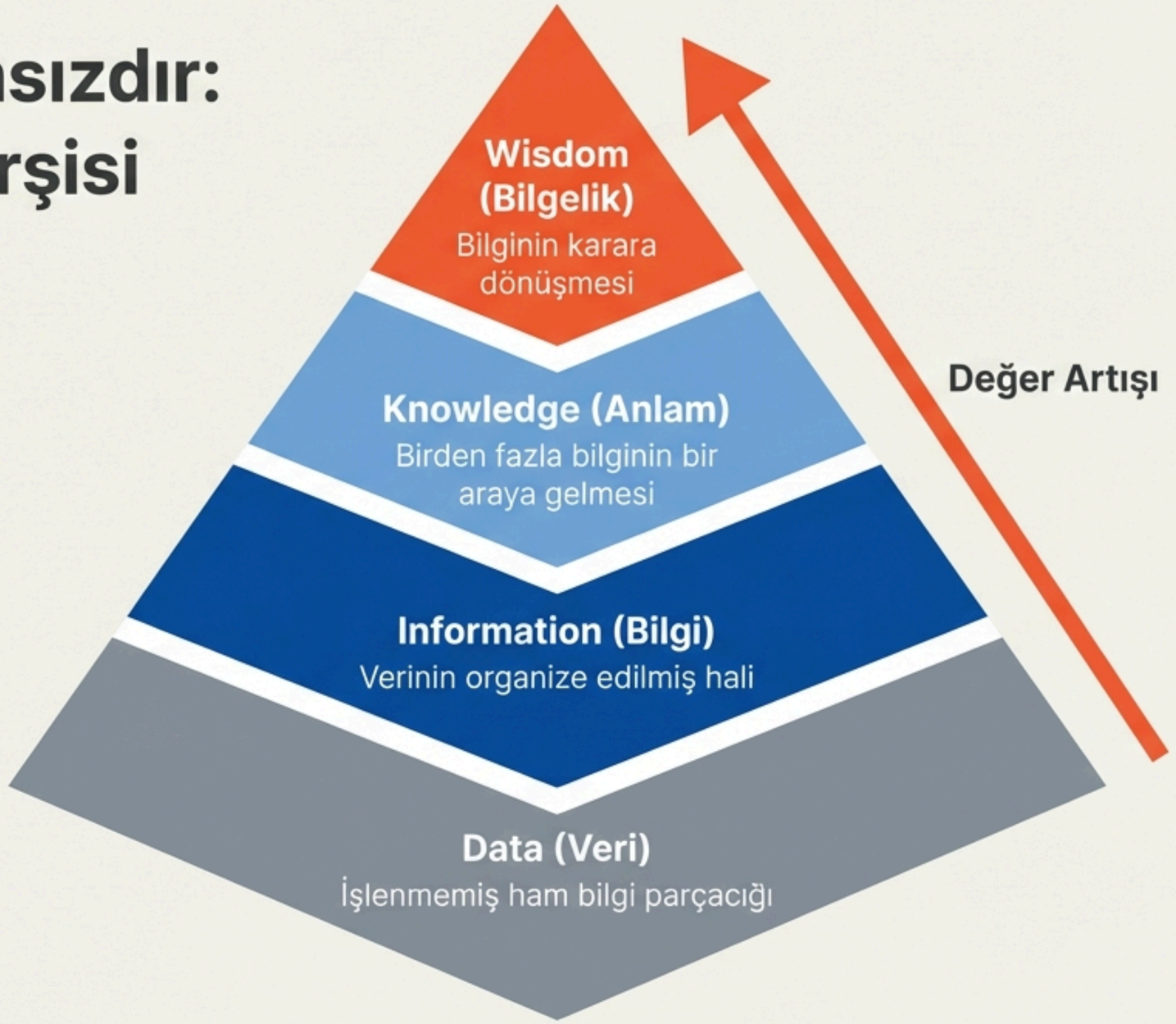
The Economist
MAY 6TH–12TH 2017
Theresa May v Brussels
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
The world's most valuable resource
amazon
Google
Data and the new rules of competition

NotebookLM

Veri Tek Başına Anlamsızdır: Değer Yaratma Hiyerarşisi

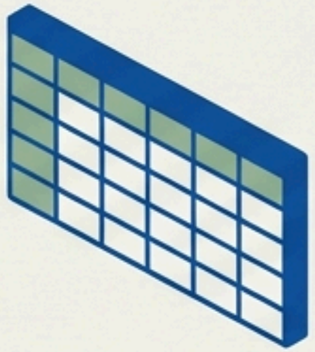
Veri Bilimi Nedir?

Verinin toplanması, işlenmesi, analiz edilmesi ve anlamlı bilgilere dönüştürülmesi sürecidir. İstatistik, matematik ve bilgisayar bilimlerinin birleşimidir.



NotebookLM

Hammaddemizi Tanıyalım: Veri Çeşitleri



Yapılandırılmış Veri (Structured)

Belirli bir formatta organize edilmiştir.

- Örnekler: Excel dosyaları, SQL tabloları.

Yarı Yapılandırılmış Veri (Semi-Structured)

Tam organize değil ama etiketler içerir.

- Örnekler: JSON, XML formatları.

Yapılandırılmamış Veri (Unstructured)

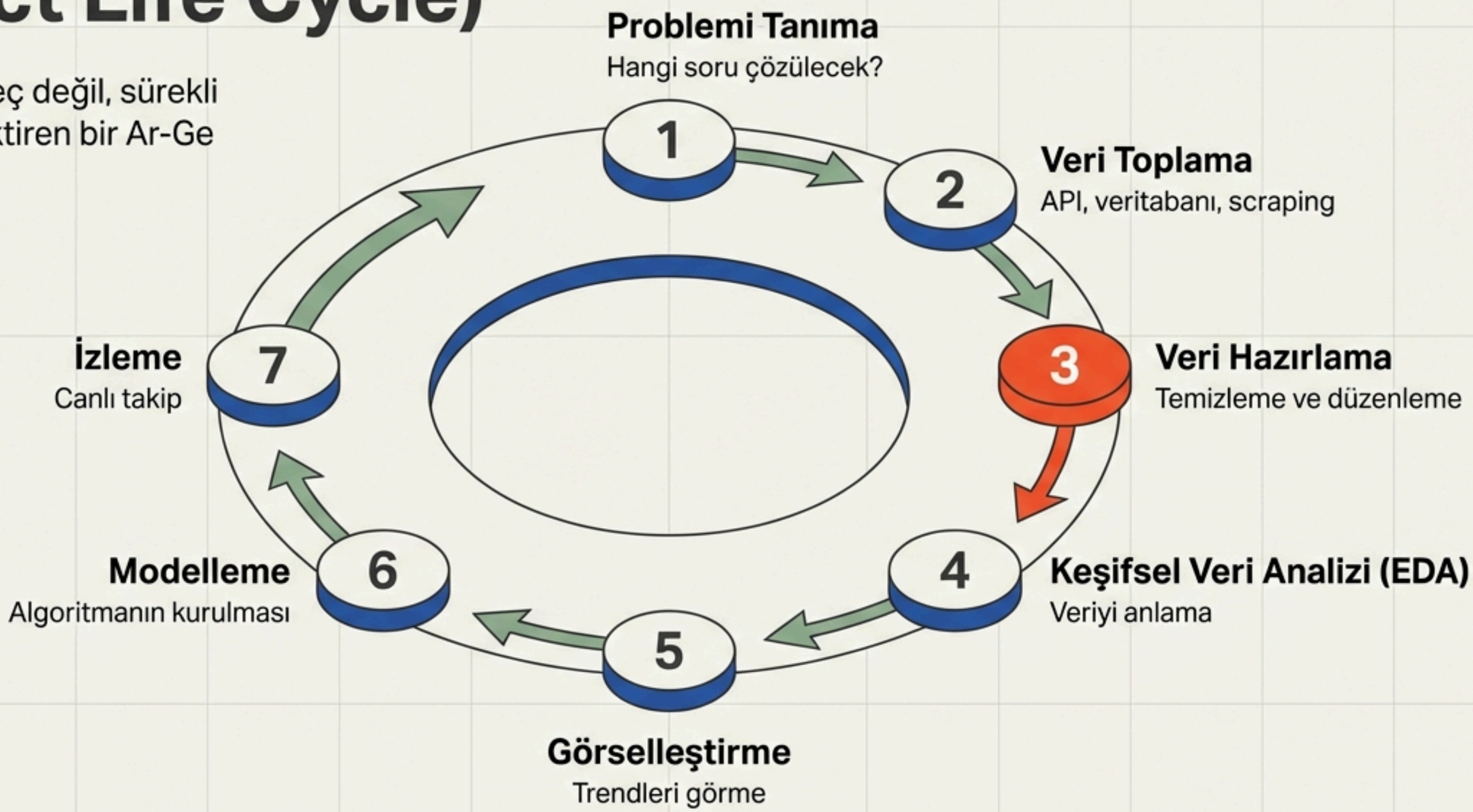
Belirli bir formatı yoktur, işlemesi en zor olandır.

- Örnekler: Fotoğraf, ses, video, metin dosyaları.

NotebookLM

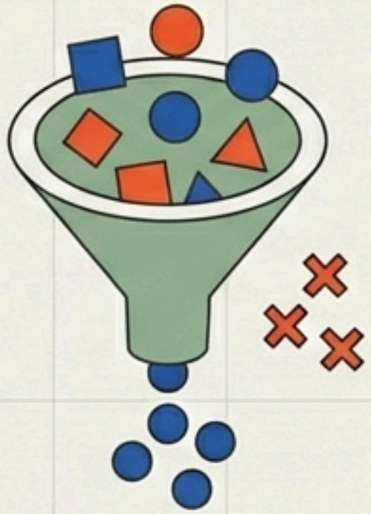
Veri Bilimi Yaşam Döngüsü (Project Life Cycle)

Bu lineer bir süreç değil, sürekli iyileştirme gerektiren bir Ar-Ge döngüsüdür.



NotebookLM

Mutfağın Arkası: Hazırlık ve Keşif

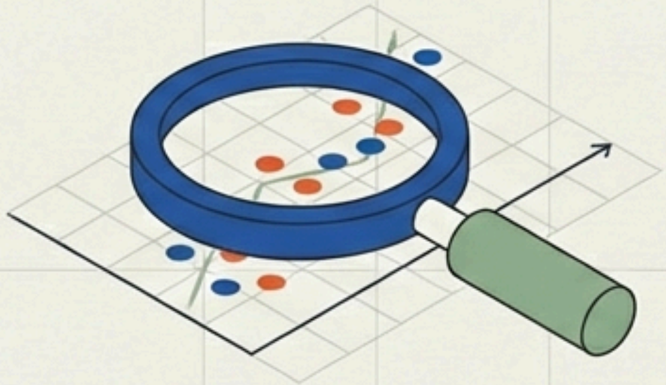


01. Veri Hazırlama ve Temizleme

Neden Önemli? Verideki eksikler, hatalar ve uyumsuzluklar giderilmelidir.

Temizleme: Tekrarlı veriler silinir, aykırı değerler düzeltilir (Örn: Yaş verisinin '200' girilmesi).

Manipülasyon: Veri, bilgiyi bozmadan algoritmaya uygun hale getirilir (Normalizasyon).



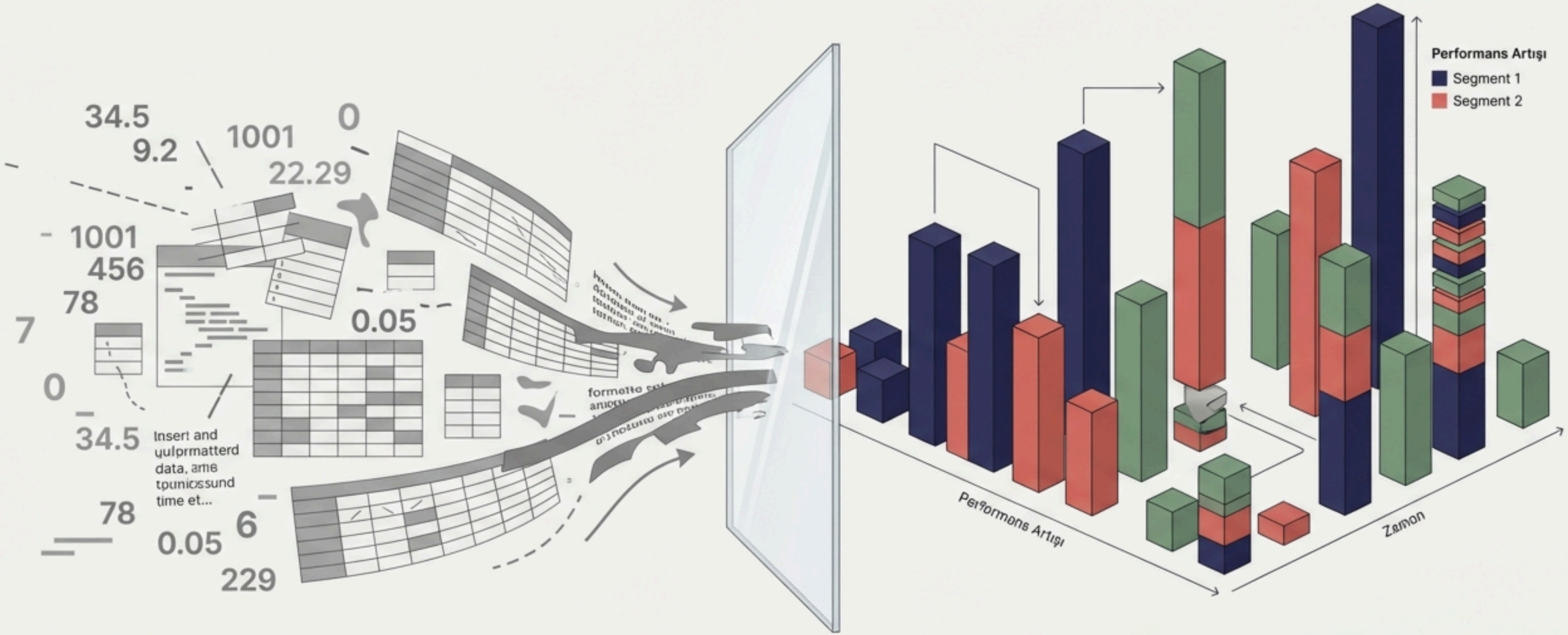
02. Keşifsel Veri Analizi (EDA)

Verinin genel özellikleri ve dağılımı incelenir. Veriyi anlamadan modele geçmek hatalı sonuç verir. Algoritma seçimi bu aşamada yapılır.

NotebookLM

Veri Görselleştirme Sanatı

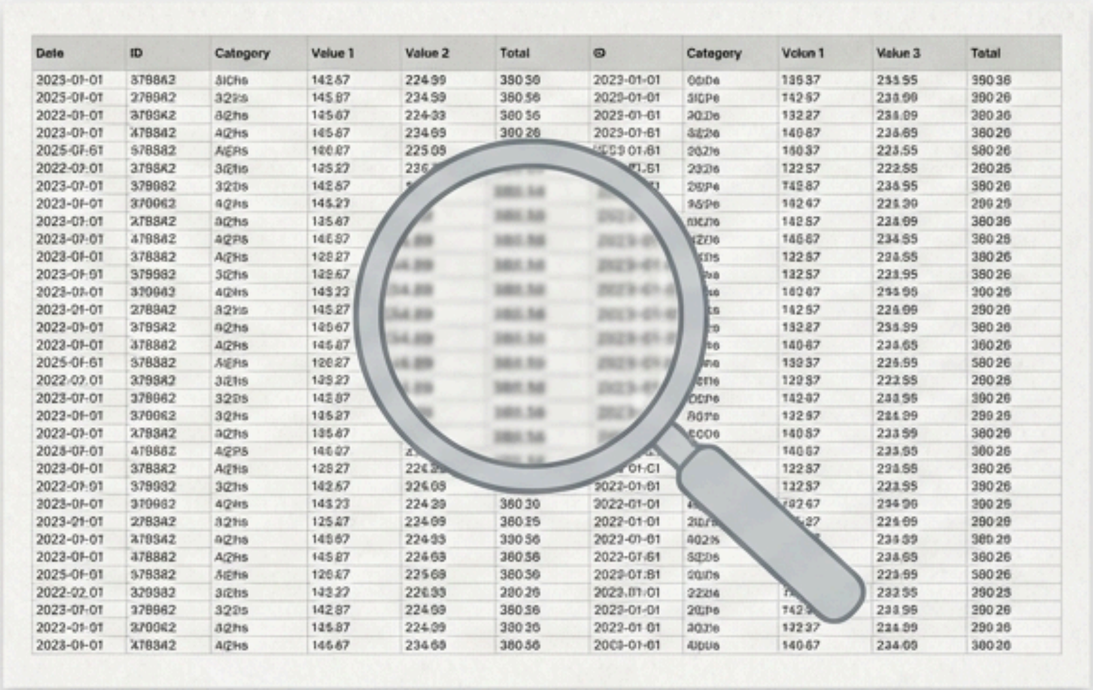
Ham Veriden Anlamlı İçgörülere



NotebookLM

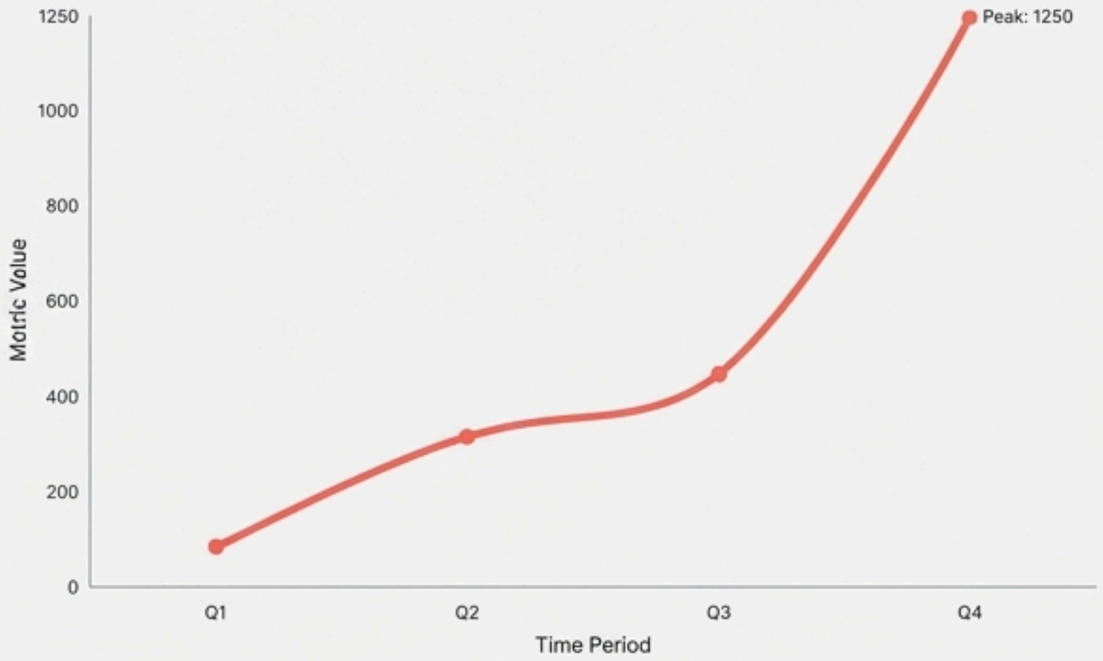
Neden Görselleştiriyoruz?

Gürültü



Sadece bir tabloya baktığımızda—bu bir Excel dosyası veya veritabanı çıktısı olabilir—trendi, uyumu veya aykırılığı çıplak gözle görmek mümkün olmayabilir.

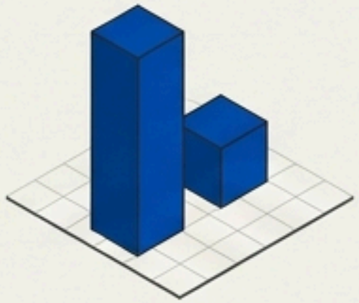
Sinyal



Görselleştirme tekniklerini, ayrıntıları kaçırmamak ve büyük resmi netleştirmek için kullanırız.

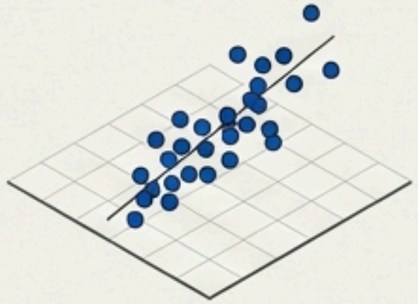
NotebookLM

Veriyi Görselleştirmek: Hikayeyi Görmek

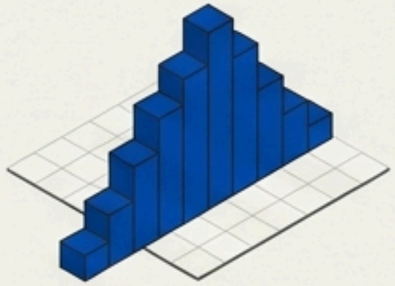


Bar Chart (Çubuk): Kategorik karşılaştırma (Kadın/Erkek sayısı).

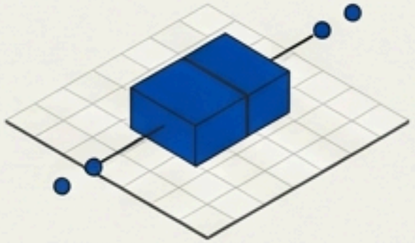
Line Chart (Çizgi): Zaman içindeki trendler (Borsa, sıcaklık).



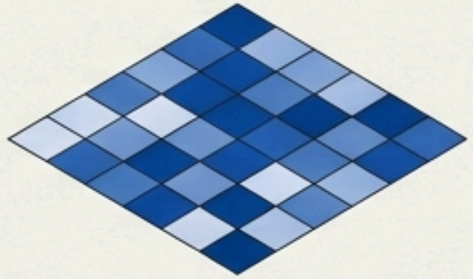
Scatter Plot (Dağılım): İki değişken arası ilişki/korelasyon.



Histogram: Veri dağılımı ve yoğunluğu (Yaş aralıkları).



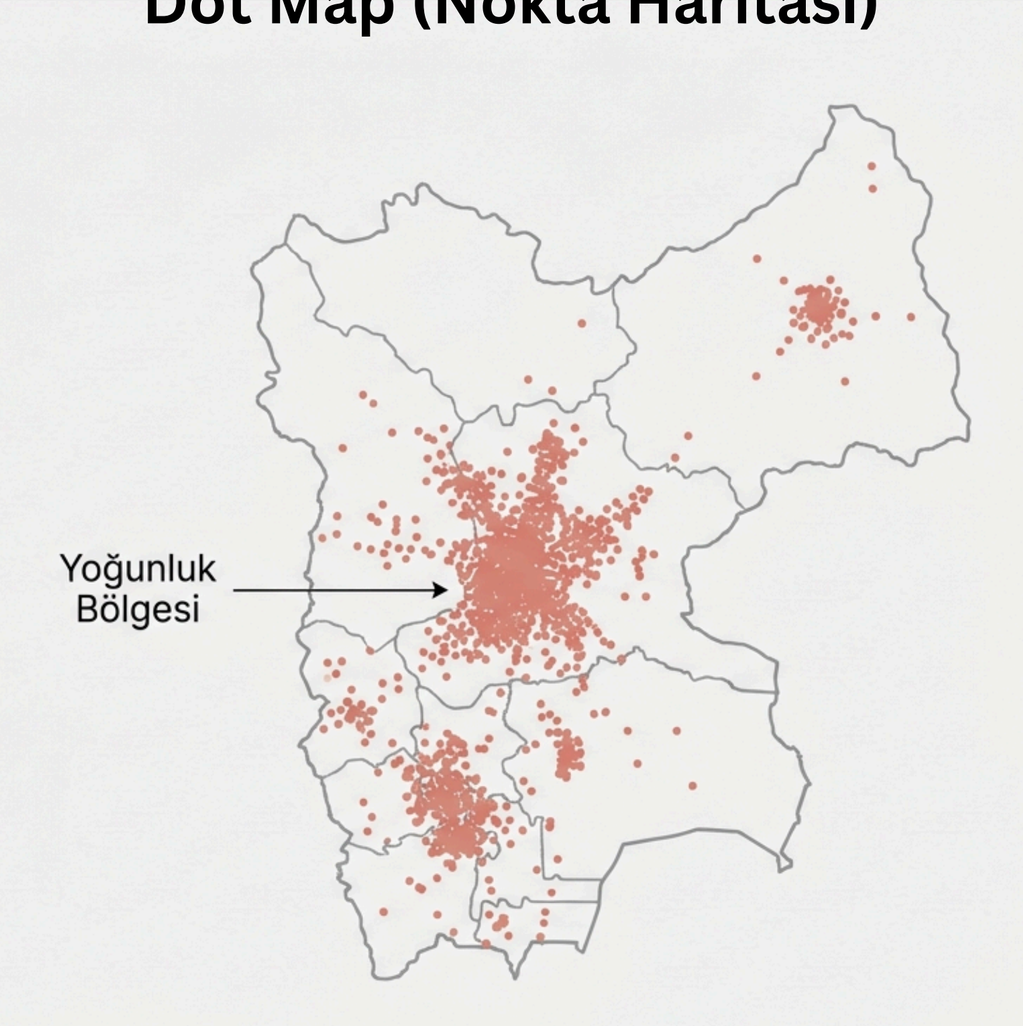
Box Plot (Kutu): Aykırı değerler (outliers) ve çeyrekler.



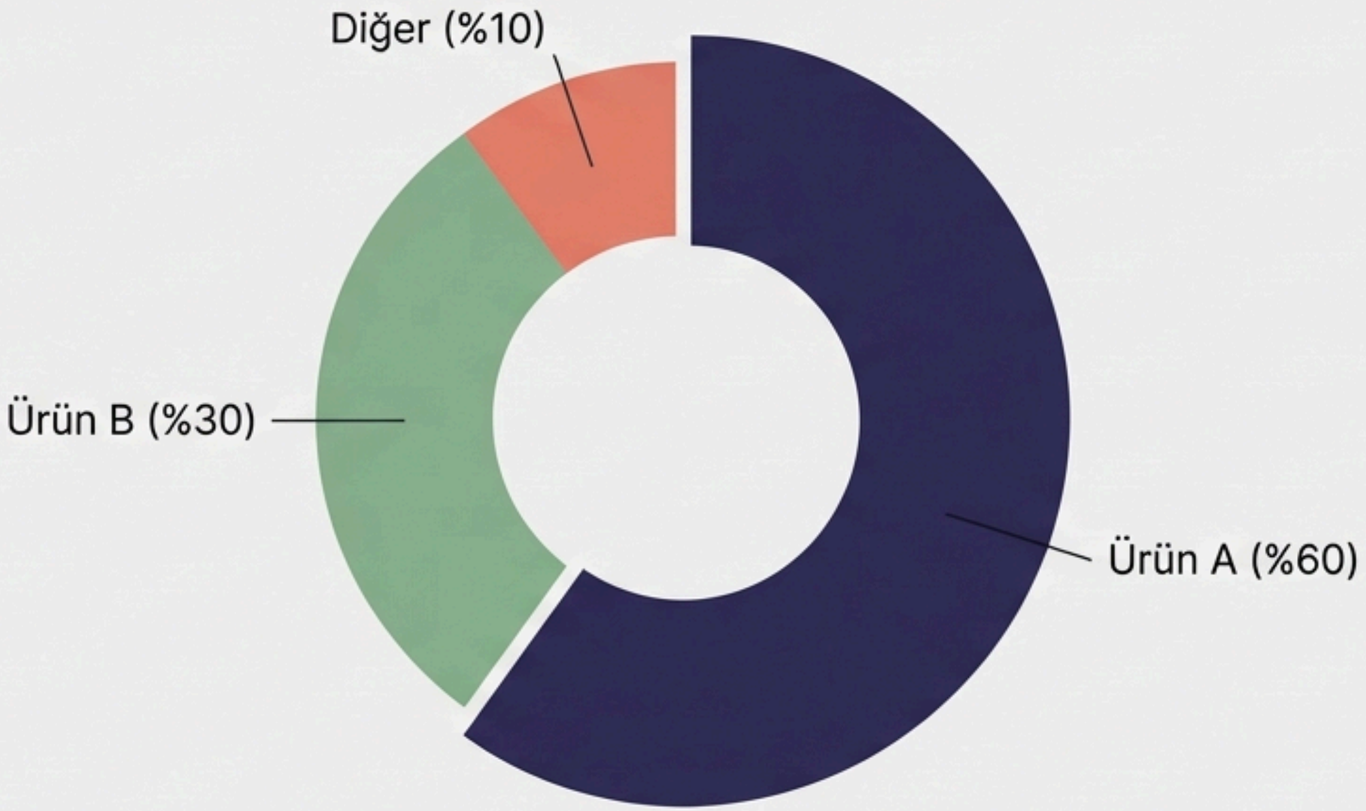
Heatmap (Isı Haritası): Değer yoğunluğu ve korelasyon matrisi.

NotebookLM

Dot Map (Nokta Haritası)



Pie Chart (Pasta Grafik)

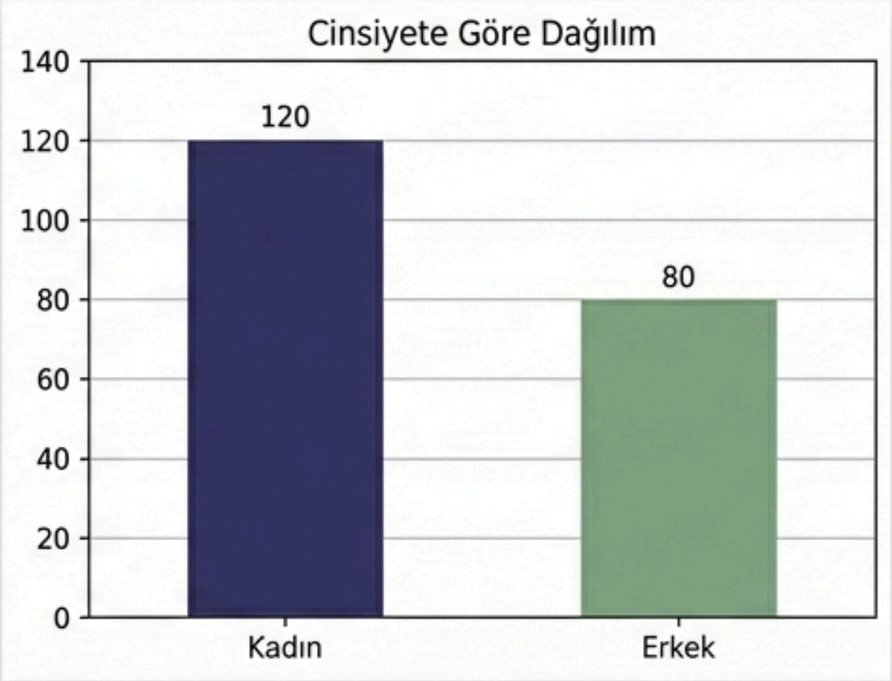


NotebookLM

Mutfağın İçinden: Koddan Görsele

Veri setinden anlamlı bir görsel çıkarmak, doğru kodlama yapısı ile başlar.

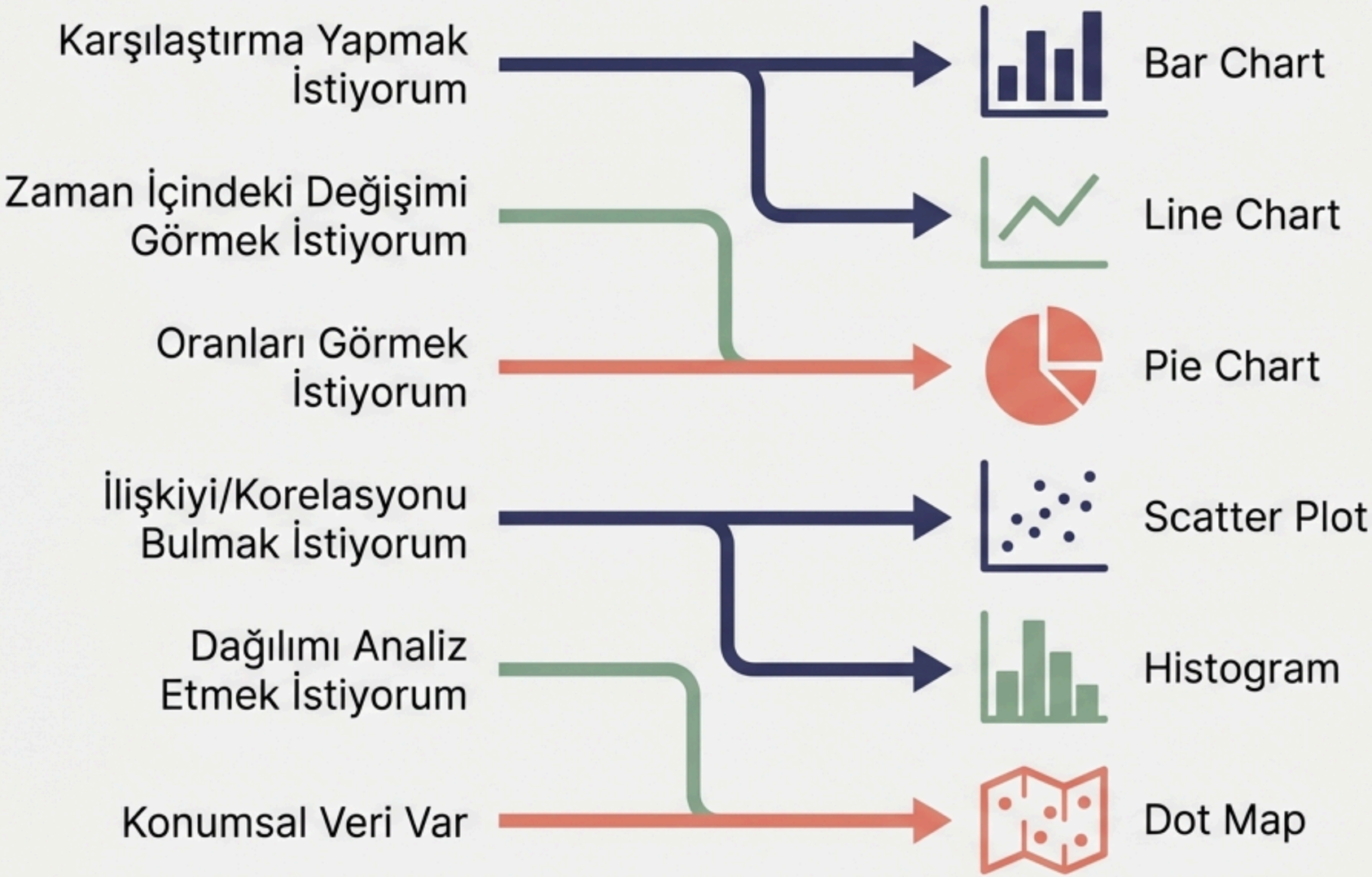
```
# 1. Veriyi Grupla
cinsiyet_sayilari = df.groupby('cinsiyet').size()

# 2. Görselleştir (Matplotlib)
import matplotlib.pyplot as plt
cinsiyet_sayilari.plot(kind='bar')
plt.title('Cinsiyete Göre Dağılım')
plt.show()
```



Kod, '120 Kadın, 80 Erkek' gibi ham sayısal sonuçları alır ve görsel bir nesneye dönüştürür.

NotebookLM

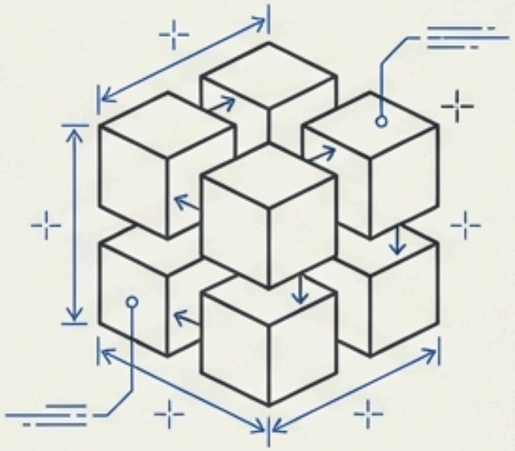
Hangi Grafiği Seçmeliyim?



NotebookLM

Başarılı Bir Model İki Temel Aşamadan Oluşur

01 Modelleme (Modeling)



Tahmin veya sınıflandırma modellerinin kurulması, sürecin sadece başlangıcıdır. Bu aşama tasarım ve test odaklıdır.

02 İzleme ve Bakım (Deployment & Monitoring)

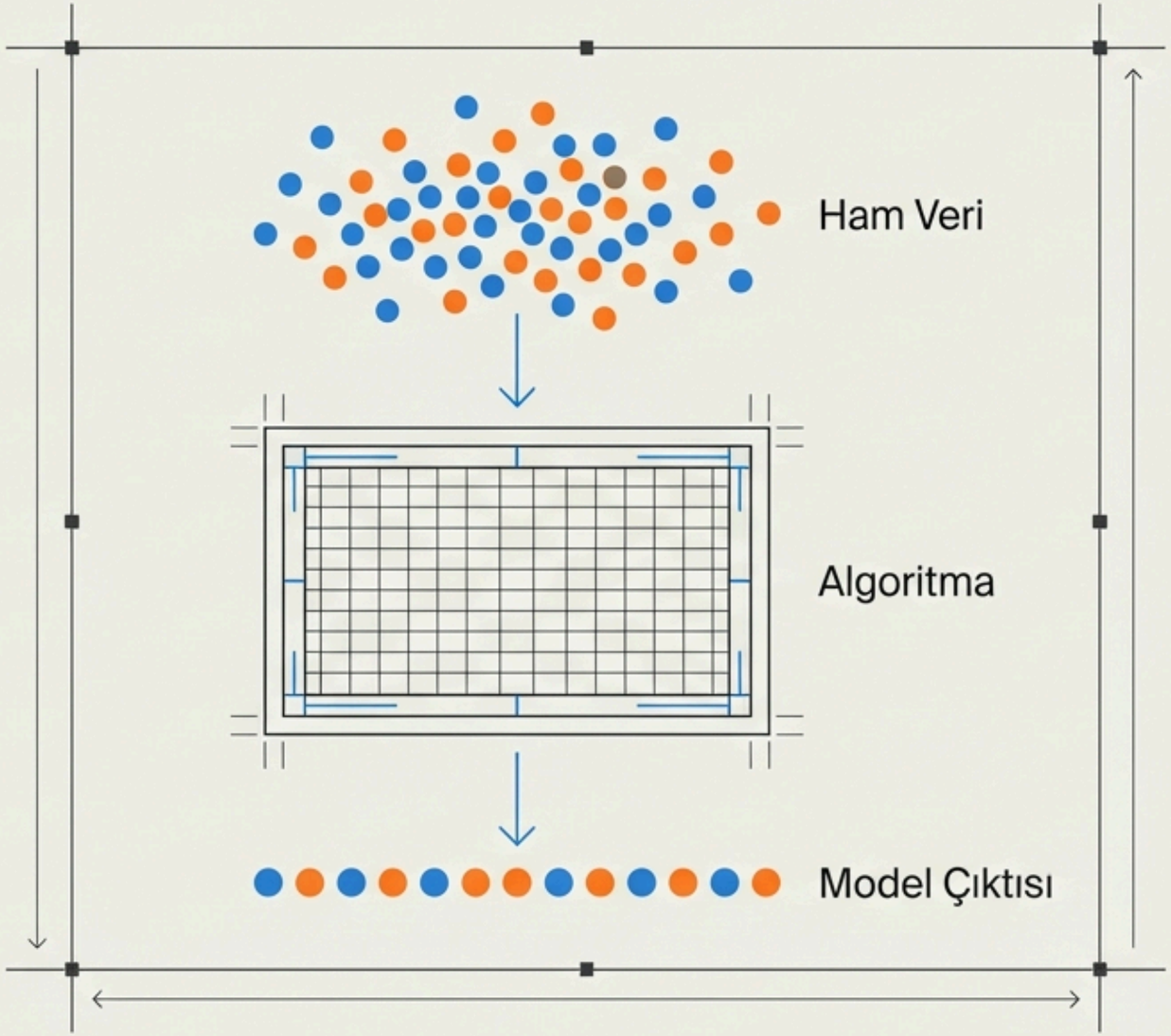
Modelin başarısı, canlıya alındıktan sonra değişen veri koşullarına ne kadar adapte olabildiği ile ölçülür.

NotebookLM

Faz 1: Modelleme ve Kapsam

Temel Aktivite:
Tahmin (Prediction) veya sınıflandırma (Classification) modelleri kurulur.

Bağlam:
İş problemini çözmek için doğru algoritma ve mimari seçimi yapılır.



NotebookLM

Temel Kriter: Genellenebilirlik

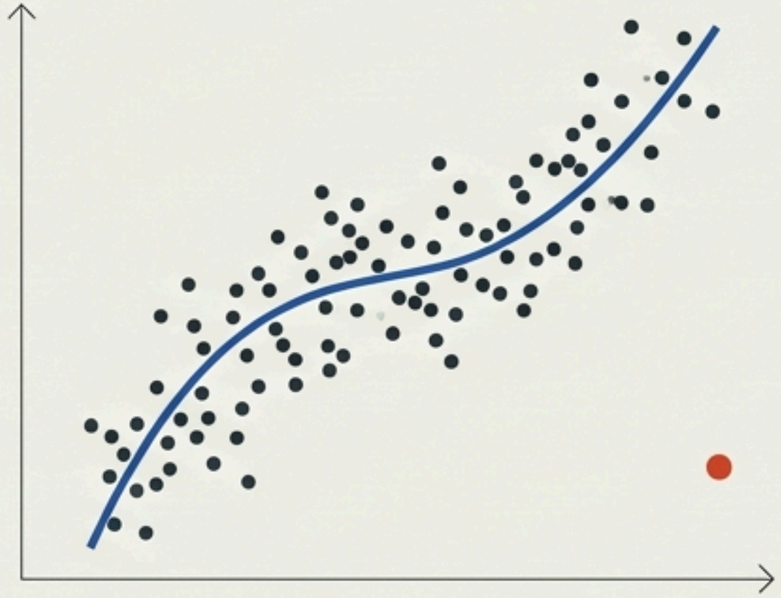
“Modelin ‘genellenebilir’ olması (farklı verilerle de çalışabilmesi) esastır.”

Ezberleme (Overfitting) Riski

Sadece eğitim verisini öğrenen bir model, gerçek dünyada başarısız olur.

Genel Geçerlilik Hedefi

Tasarım önceliği, modelin daha önce hiç görmediği veri setlerinde de mantıklı sonuçlar üretmesidir.



NotebookLM

Başarı Ölçümü ve Test Verileri

Modelin başarısı test verileriyle ölçülür. Eğitim sırasında kullanılmayan ayrıştırılmış veri setleri, modelin gerçek dünya performansının tek objektif göstergesidir.

NotebookLM

- **MAE – Mean Absolute Error**

Ortalama mutlak hata

- **MSE – Mean Squared Error**

Hataların karesinin ortalaması

- **RMSE – Root Mean Squared Error**

MSE’nin karekökü

- **R^2 – R-kare (Belirlilik Katsayısı)**

Model, verideki değişimin
ne kadarını açıklıyor?

- **Accuracy**

Doğru tahmin oranı

- **Precision**

Pozitif dediklerinin kaçı gerçekten pozitif?

- **Recall (Sensitivity)**

Gerçek pozitiflerin kaçını yakaladın?

- **F1-Score**

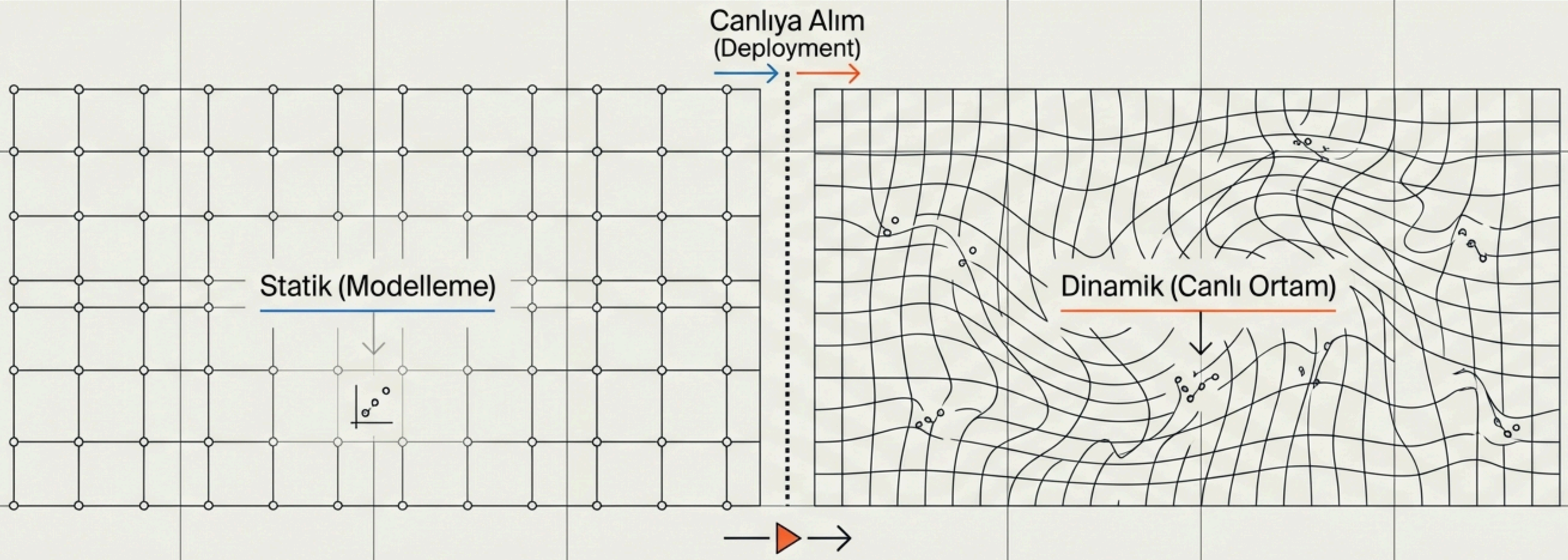
Precision ve Recall’un dengesi

	Tahmin Pozitif	**Tahmin Negatif**
Gerçek Pozitif	TP	FN
Gerçek Negatif	FP	TN

NotebookLM

Faz 2: Canlıya Alım ve Değişen Dinamikler

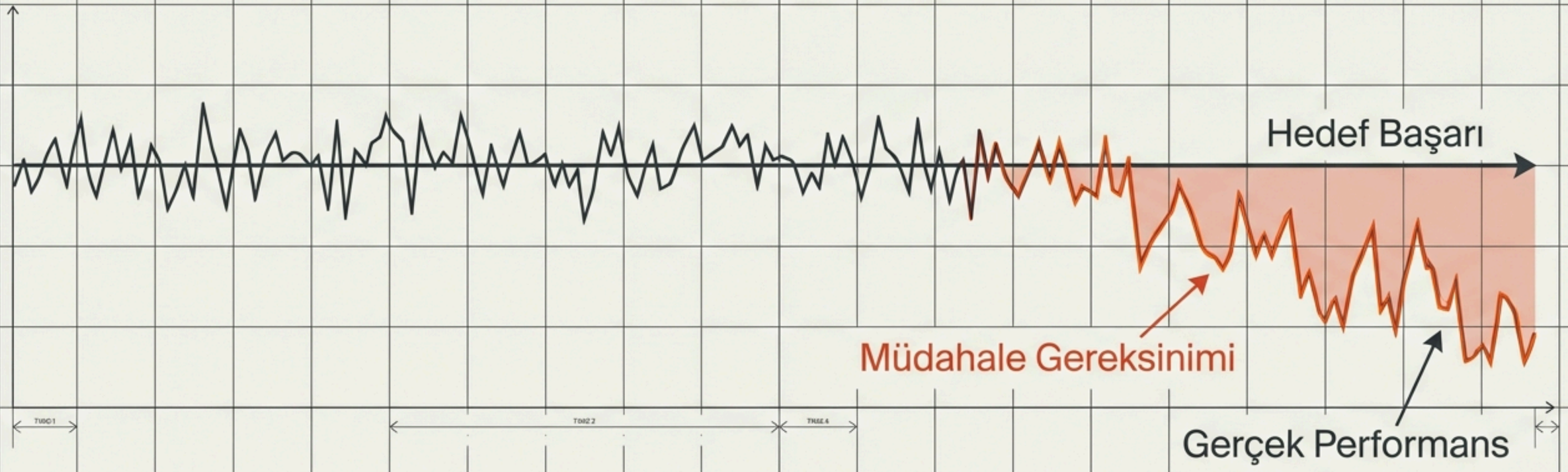
Model canlıya alındıktan sonra veri değişebilir. Laboratuvar ortamındaki sabit varsayımlar, canlı ortamın dinamik yapısıyla karşılaşır.



NotebookLM

İzleme (Monitoring) Zorunluluğu

Modelin başarısının zamanla düşüp düşmediği sürekli takip edilmelidir. Canlıya alım ‘bitti’ demek değildir; sürekli bir gözlem döngüsünün başlangıcıdır.



NotebookLM

Model Sağlığı İçin Kritik Sorular

Performans

Model başarısı zaman içinde düşüyor mu?

Kararlılık

Hala aynı başarıda mı?

Veri Kalitesi

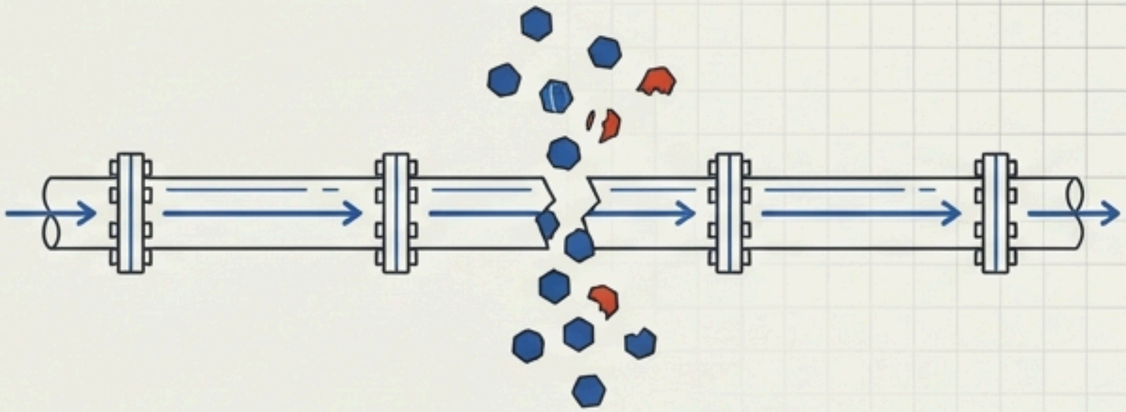
Veride bir hata var mı, veri değişti mi?

NotebookLM

Kök Neden Analizi: Hata ve Değişim

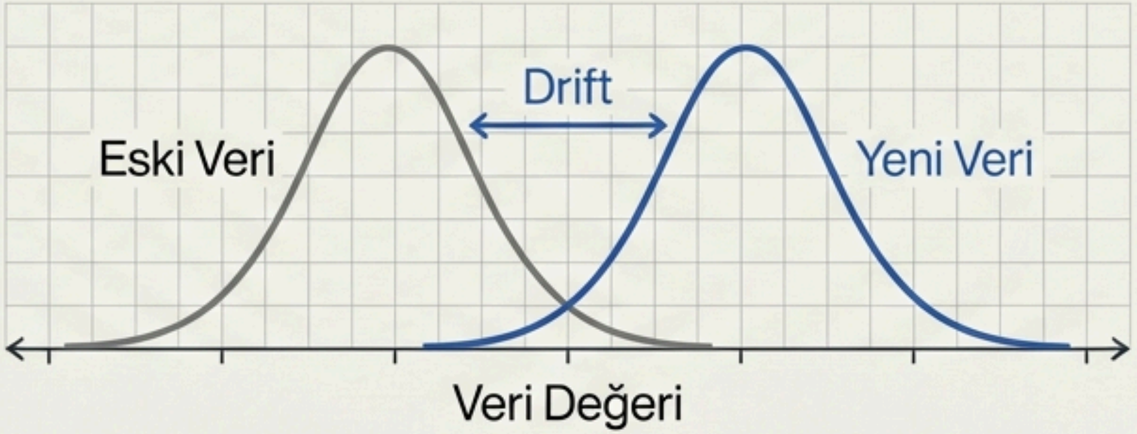
Senaryo A: Veri Hataları

Soru: Veride bir hata var mı? Teknik boru hatlarında (pipelines) kopukluk veya veri bozulması kontrol edilir.



Senaryo B: Veri Değişimi (Drift)

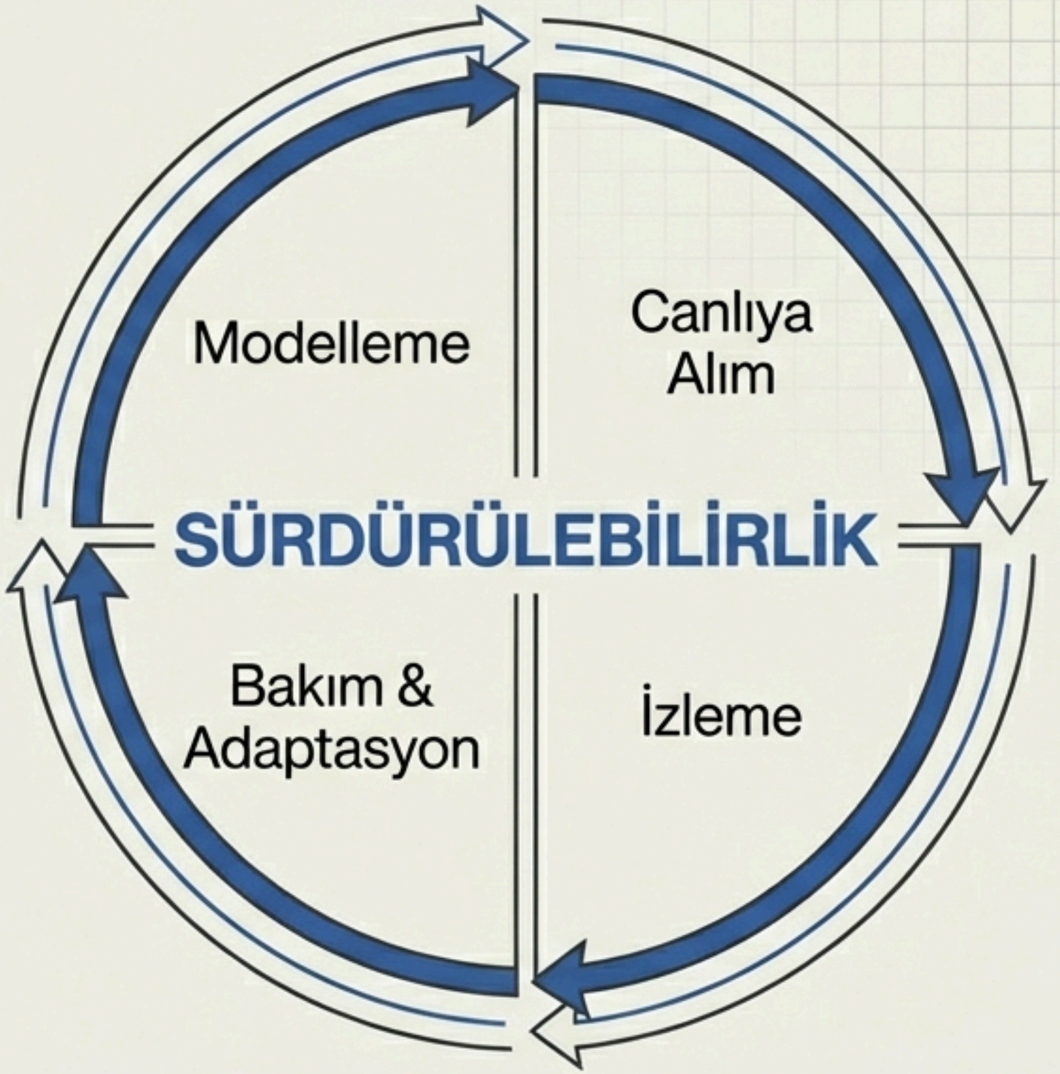
Soru: Veri değişti mi? Dünyadaki temel kalıpların değişmesi sonucu modelin öğrendiği kuralların geçersiz kalmasıdır.



NotebookLM

Sonuç: İzleme ve Bakım Döngüsü

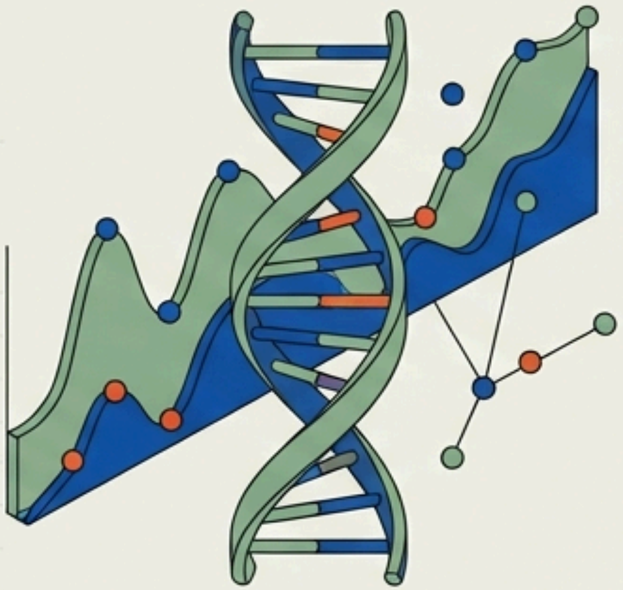
Modelleme değer yaratır, izleme bu değeri korur. “Sürekli takip” prensibi, veri değiştiğinde veya hata oluştuğunda ekibin müdahale etmesini sağlar.



NotebookLM

Sektörel Uygulamalar: İnsan ve Ticaret

Sağlık Sektörü



- Kanser gibi hastalıkların erken teşhisi.
- Radyolojik görüntülerin analizi.
- E-Nabız verileriyle hasta takibi ve analizler.

Perakende ve E-Ticaret

- **Churn Analizi:** Müşteri terk tahmini (Bizi bırakacak mı?).
- **Sepet Analizi:** Birliktelik kuralları (Örn: Ekmek alan süt de alıyor).
- **Tavsiye Sistemleri:** 'Bunu inceleyen buna da baktı.'
- **Sentiment Analizi:** Müşteri yorumlarının duygu analizi.

NotebookLM

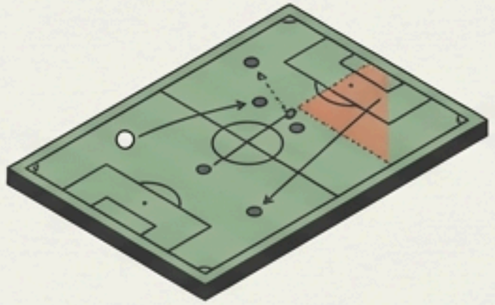
Sektörel Uygulamalar: Sistemler ve Şehirler

Finans

Fraud Detection: Kredi kartı dolandırıcılığı tespiti.
Risk Analizi: Kredi skorlama.

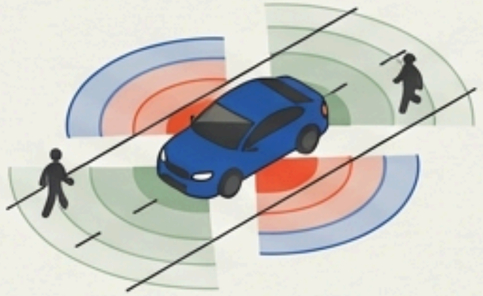
Ulaşım

Kargo rota optimizasyonu.
Araç paylaşım eşleşmeleri.



Spor

Bilgisayarlı görü ile ofsayt tespiti.
Gerçek zamanlı top/oyuncu takibi.

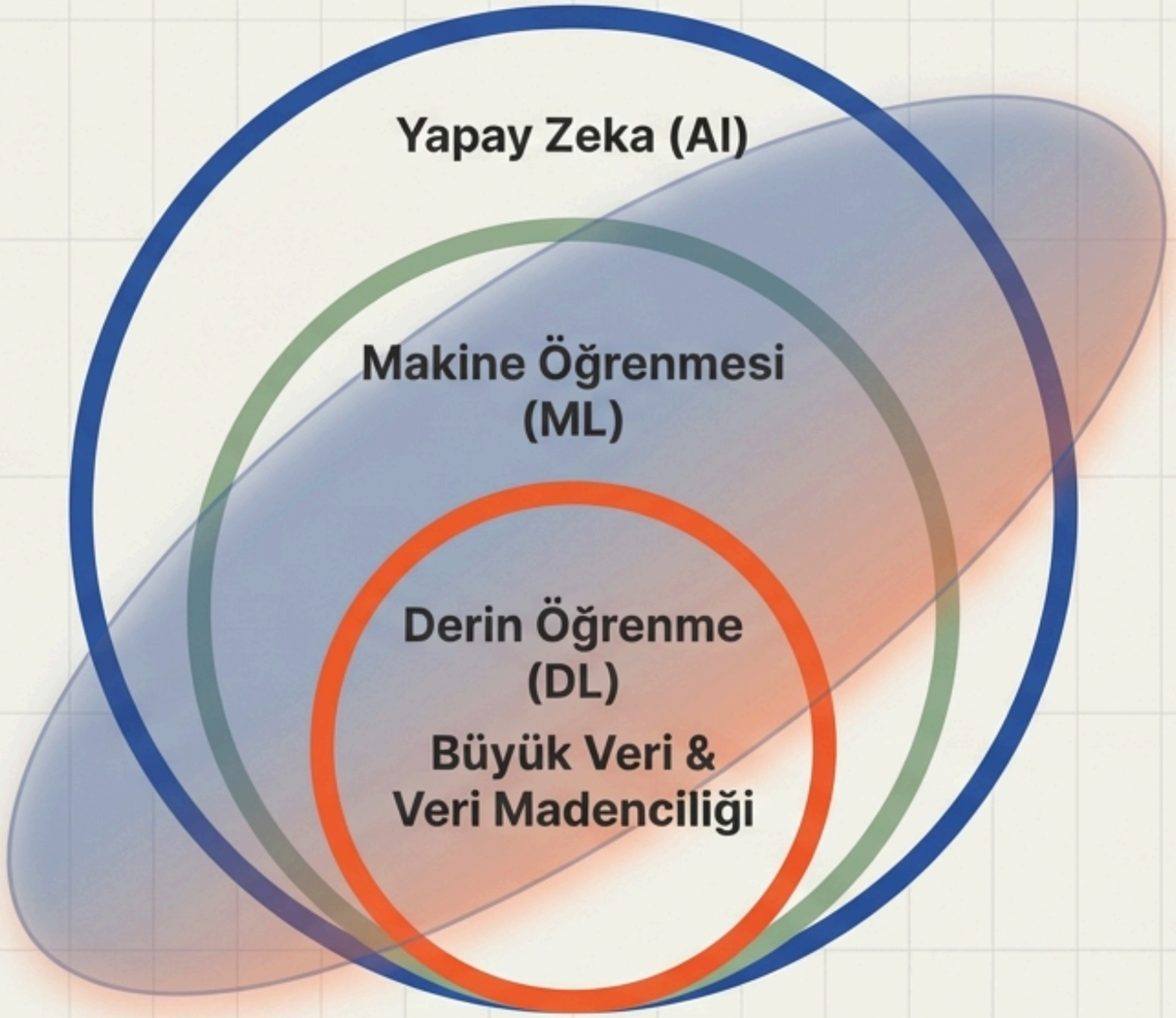


Otonom Araçlar

Algılama: Yayayı görme.
Tahmin: Araç hareketlerini öngörme.
Karar: Fren/Gaz aksiyonu.

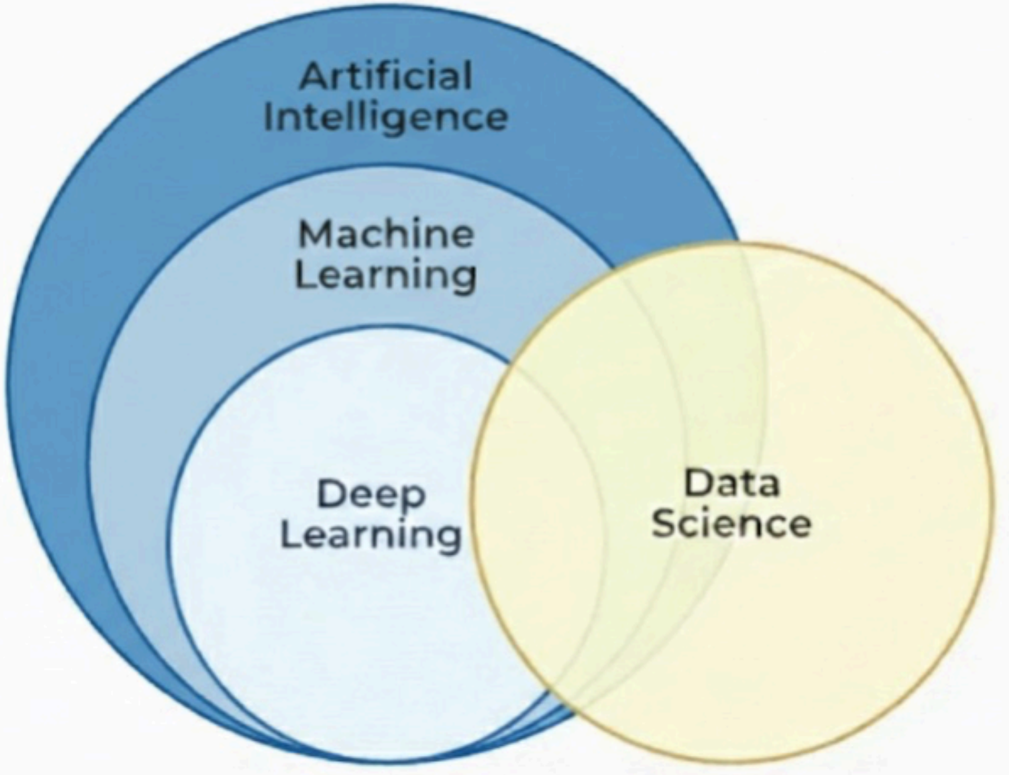
NotebookLM

Kavram Karmaşasını Çözmek: Büyük Resim



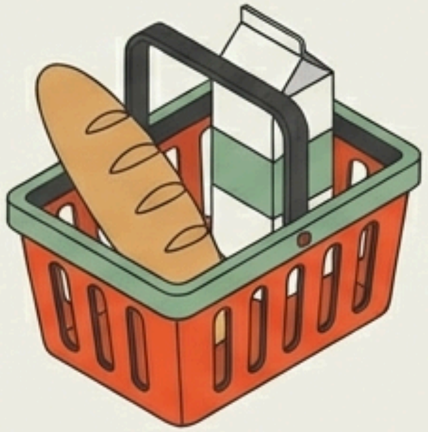
Veri Madenciliği en içteki örüntü keşfidir; Büyük Veri ise bu sistemlerin beslendiği devasa kaynaktır.

NotebookLM


Artificial Intelligence
Machine Learning
Deep Learning
Data Science

Veri Madenciliği: Örüntüleri Keşfetmek

Büvük veri setleri içinden anlamlı, gizli ve işe yarar örüntüleri (pattern) çıkarma sürecidir.



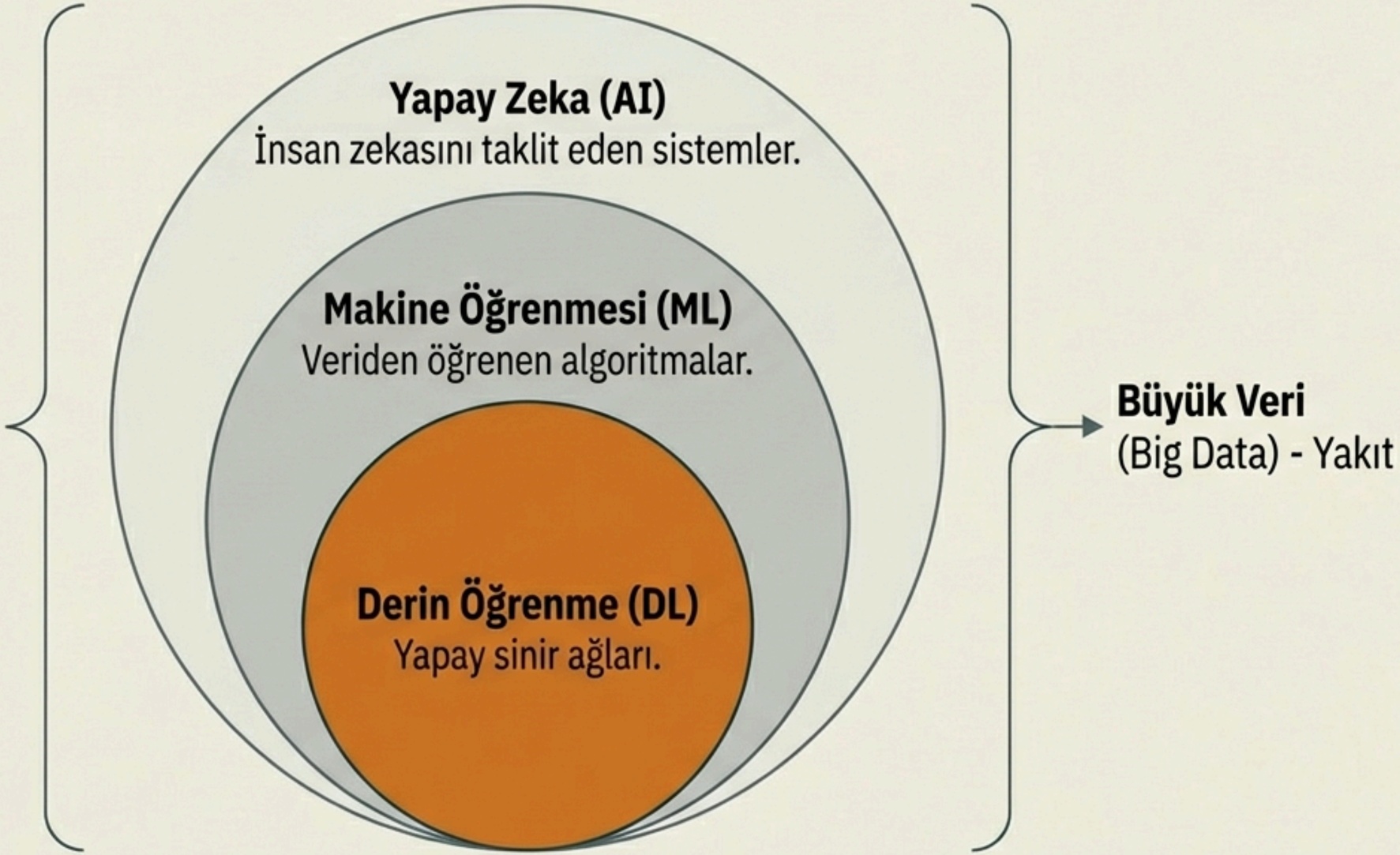
Bulgu: Ekmek alanların %80'i Süt de alıyor.

Aksiyon: Raf düzeni ve kampanya.

Veri Analizi	Veri Madenciliği
Ne oldu?	Ne olabilir?
Geçmiş odaklı	Gelecek/Örüntü odaklı
Raporlama	Modelleme

NotebookLM

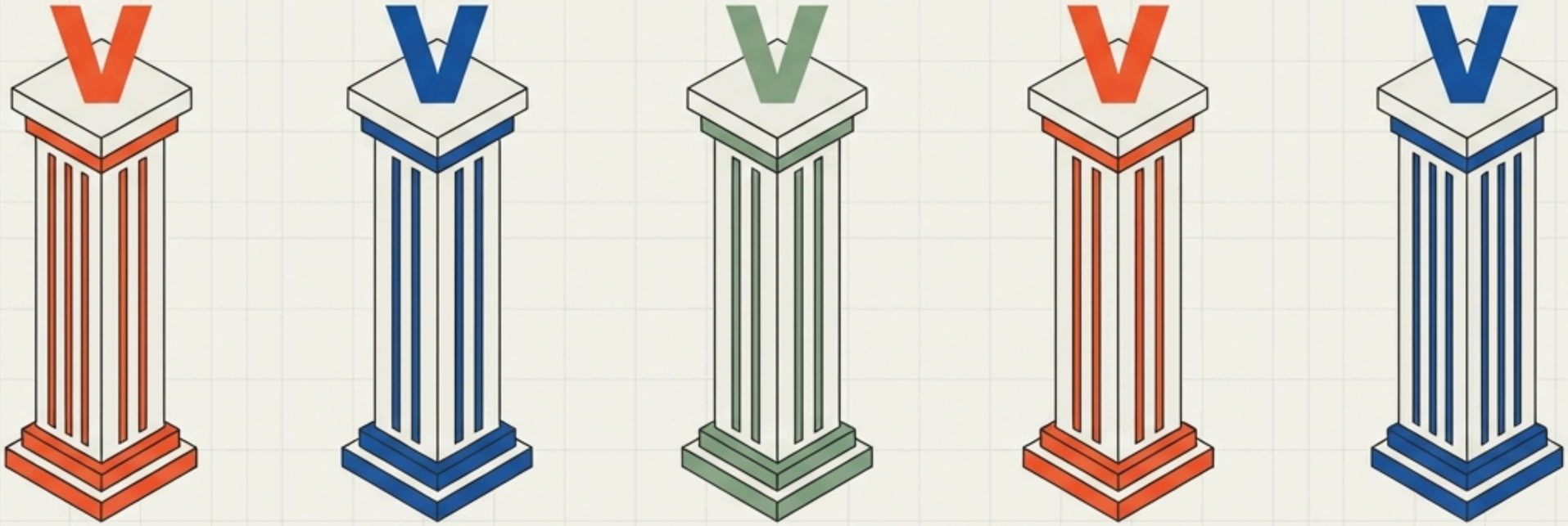
Yapay Zeka Dünyasının Haritası



NotebookLM

Büyük Veri (Big Data) ve 5V Kuralı

Klasik veritabanı sistemlerinin (SQL) işleyemeyeceği kadar büyük, hızlı ve karmaşık veri kümeleridir.



Volume (Hacim)

TB, PB seviyesinde veri.

Velocity (Hız)

Verinin çok hızlı akması (streaming).

Variety (Çeşitlilik)

Yapılandırılmış ve yapılandırılmamış karışık.

Veracity (Doğruluk)

Güvenilirlik ve temizlik.

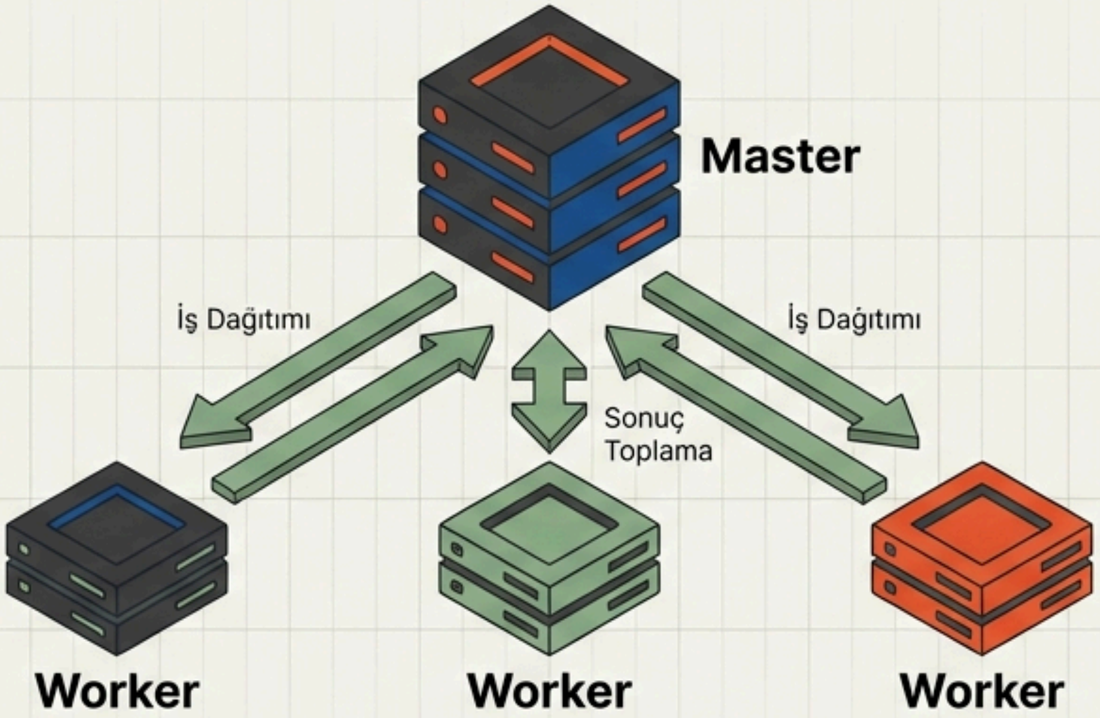
Value (Değer)

İş değeri üretme amacı.

NotebookLM

Büyük Veri Teknolojileri: Dağıtık Sistemler

Tek bilgisayarın gücü yetmediğinde iş parçalara bölünür, paralel makinelerde işlenir.



Depolama

Hadoop (HDFS), Cloud Storage (S3)

İşleme

Apache Spark (Hızlı, bellek içi), Hadoop MapReduce

Akış (Streaming)

Apache Kafka (Canlı veri)

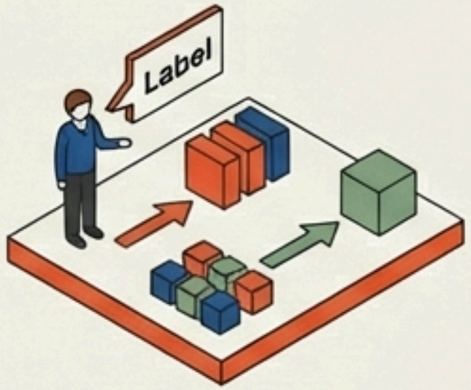
NoSQL

MongoDB, Cassandra

NotebookLM

Makine Öğrenmesi (Machine Learning) Türleri

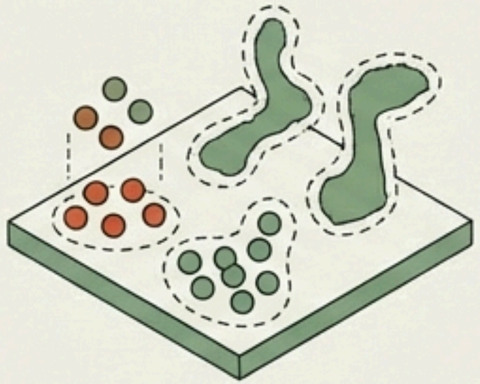
Veriden öğrenerek tahmin yapan veya karar veren algoritmalar.



Denetimli Öğrenme (Supervised)

Etiketli veri kullanılır (Label var).

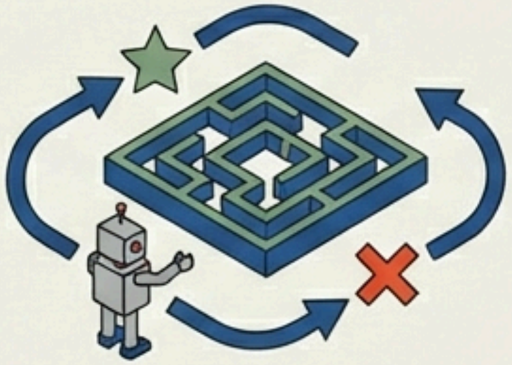
- Regresyon (Sayısal)
- Sınıflandırma (Kategorik)



Denetimsiz Öğrenme (Unsupervised)

Etiket yoktur, veri kendi içinde gruplanır.

- Kümeleme (K-Means)
- Boyut İndirgeme



Pekiştirmeli Öğrenme (Reinforcement)

Deneme-yanılma ve ödül/ceza sistemi.

- Otonom araçlar
- Oyun botları

NotebookLM

Algoritmalar Kütüphanesi

Denetimli Öğrenme (Supervised)

Regresyon (Sayısal Çıktı)

- Linear Regression
- Decision Tree Regressor
- Random Forest
- SVR

Sınıflandırma (Kategorik Çıktı)

- Logistic Regression
- KNN
- Naive Bayes
- SVM
- Decision Tree Classifier

Denetimsiz Öğrenme (Unsupervised)

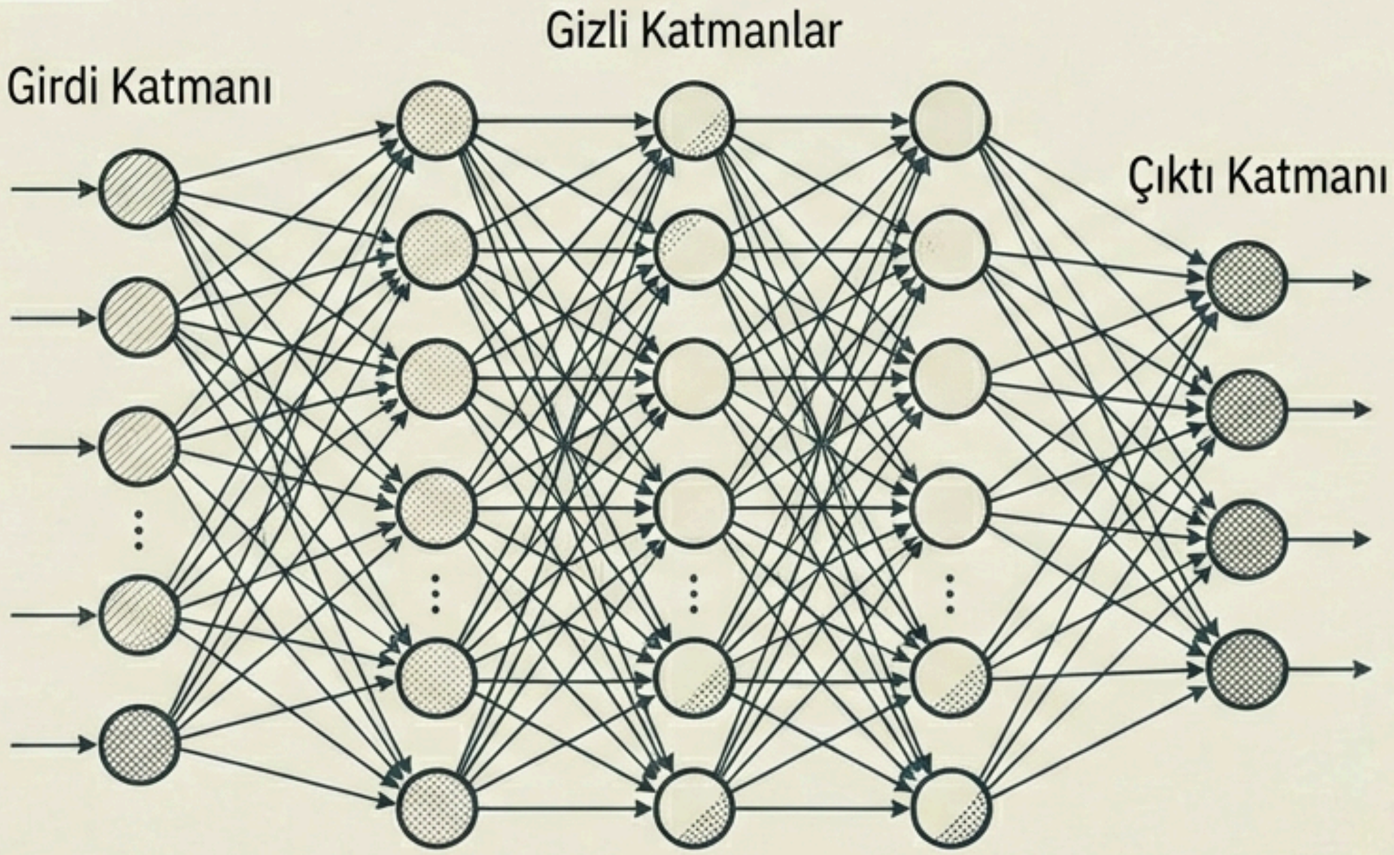
Kümeleme

- K-Means
- DBSCAN
- Hierarchical Clustering

Modelin başarısı ‘Train-Test’ veri ayrımı ile ölçülür.

NotebookLM

Derin Öğrenme: İnsan Beyninin Dijital İkizi



Ana Mimariler

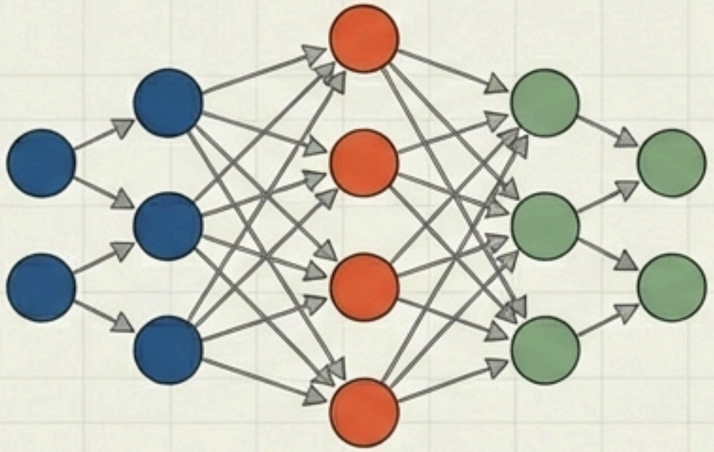
- ANN (Yapay Sinir Ağları): Temel yapı.
- CNN (Evrişimli Sinir Ağları): Görüntü işleme.
- RNN & LSTM: Metin ve zaman serileri.
- Transformers: Modern GenAI temeli.

NotebookLM

Derin Öğrenme ve Üretken Yapay Zeka (GenAI)

Derin Öğrenme (Deep Learning)

İnsan beyni simülasyonudur (Yapay Sinir Ağları).

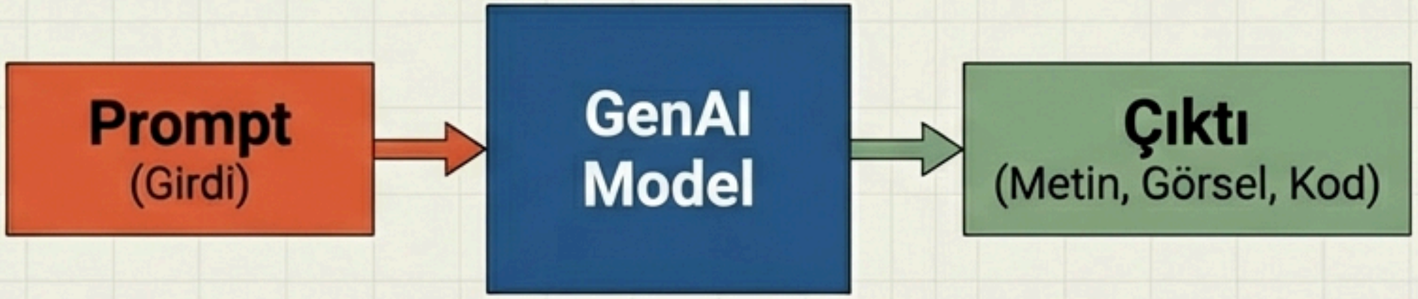


CNN (Görüntü)

RNN (Zaman Serisi)

Üretken Yapay Zeka (Generative AI)

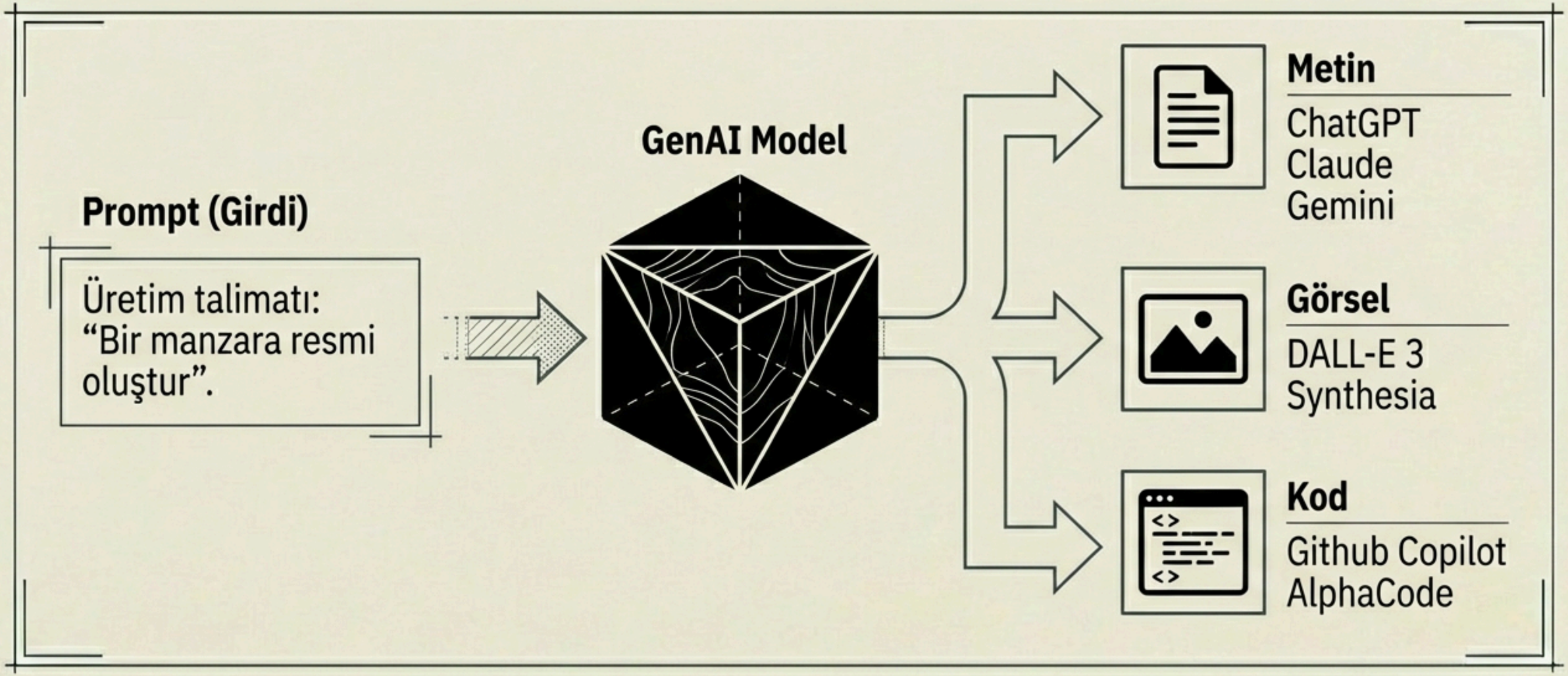
Sadece analiz etmez; yeni ve özgün içerik üretir.



NotebookLM

Üretken Yapay Zeka (Generative AI)

Analiz etmez, yaratır.



NotebookLM

Araç Çantası ve Sonuç

Veri Bilimi, geçmişi anlamak, şimdiyi yönetmek ve geleceği tahmin etmek için en güçlü araçtır.

NotebookLM