Prediction in Excel

Through Filling Missing Values

Devavrat Shah

Professor
EECS / IDSS, MIT

Director
Statistics & Data Science, MIT

Cofounder
Celect, Inc.
Ikigailabs

Machine Learning

Term was coined by Arthur Samuel in 1959

Known for	Samuel Checkers-playing Program Alpha–beta pruning (an early implementation) Pioneer in Machine Learning [1] TeX project (with Donald Knuth)

Evolves from Statistical learning, pattern recognition

Goal to make computers “learn” from “data”

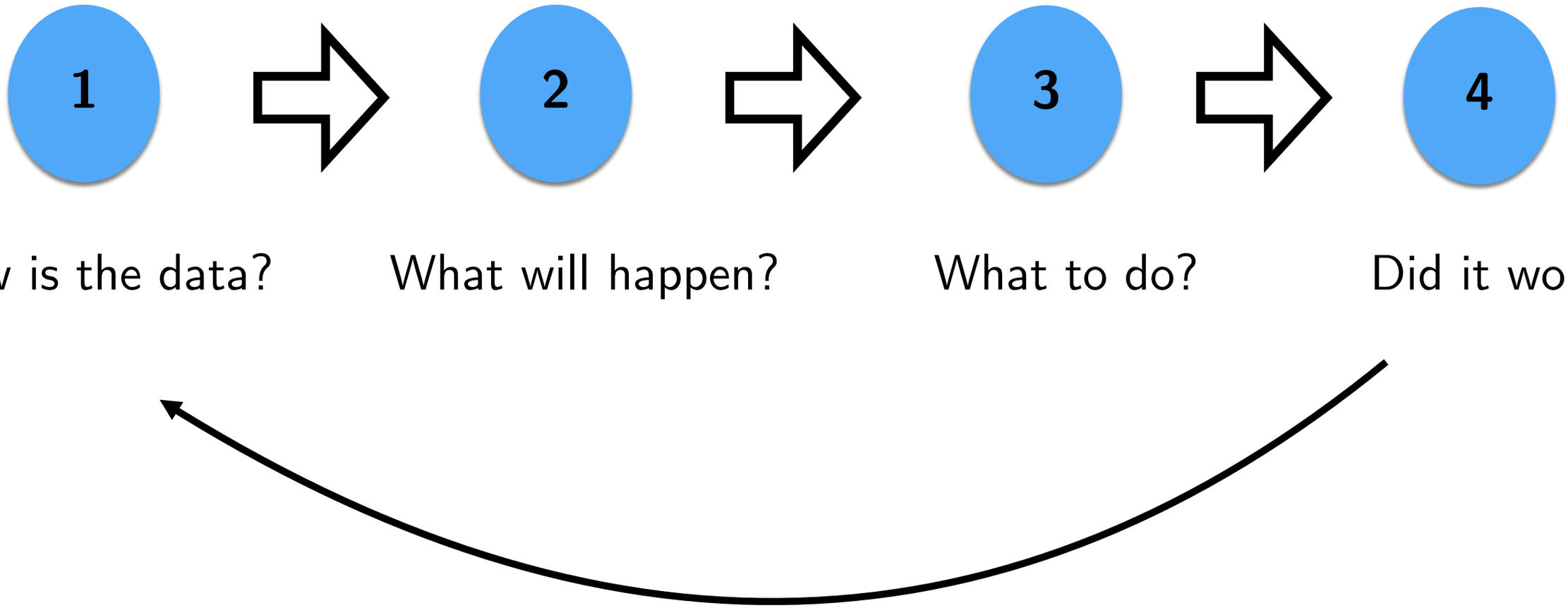
An end-user’s perspective

Obtaining insights from data and to make decisions

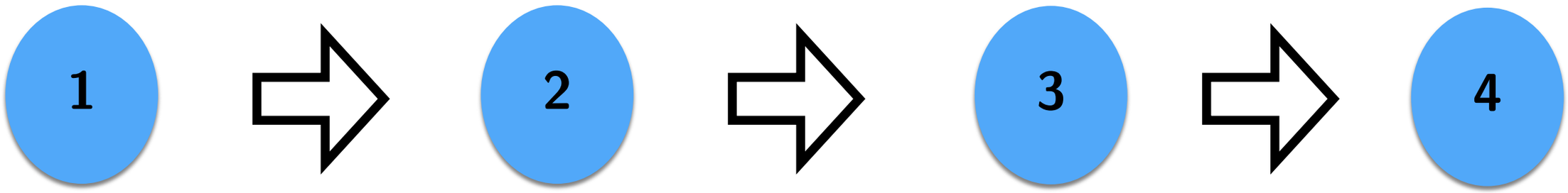
Intellectually

Methods, models, algorithms and some

Machine Learning: The building blocks



Machine Learning: The building blocks



How is the data? What will happen? What to do? Did it work?

Modeling

Visualization

Probability Distribution

Un-supervised Learning

• • •

Prediction

Supervised Learning

Semi-Supervised Learning

Model Learning

• • •

Decision Systems

Markov Decision Process

Reinforcement Learning

Model Predictive Control

• • •

Causal Inference

Randomized Control Trials

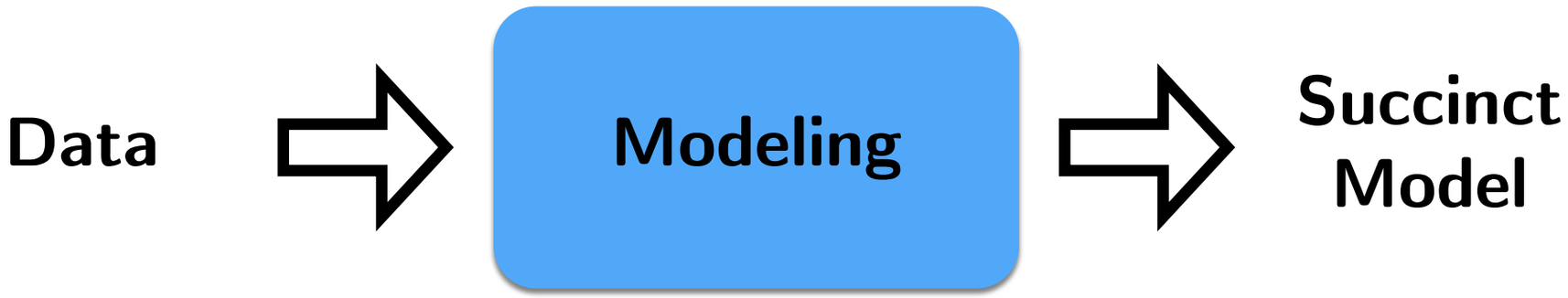
Natural Experiments

Causality Indices

• • •

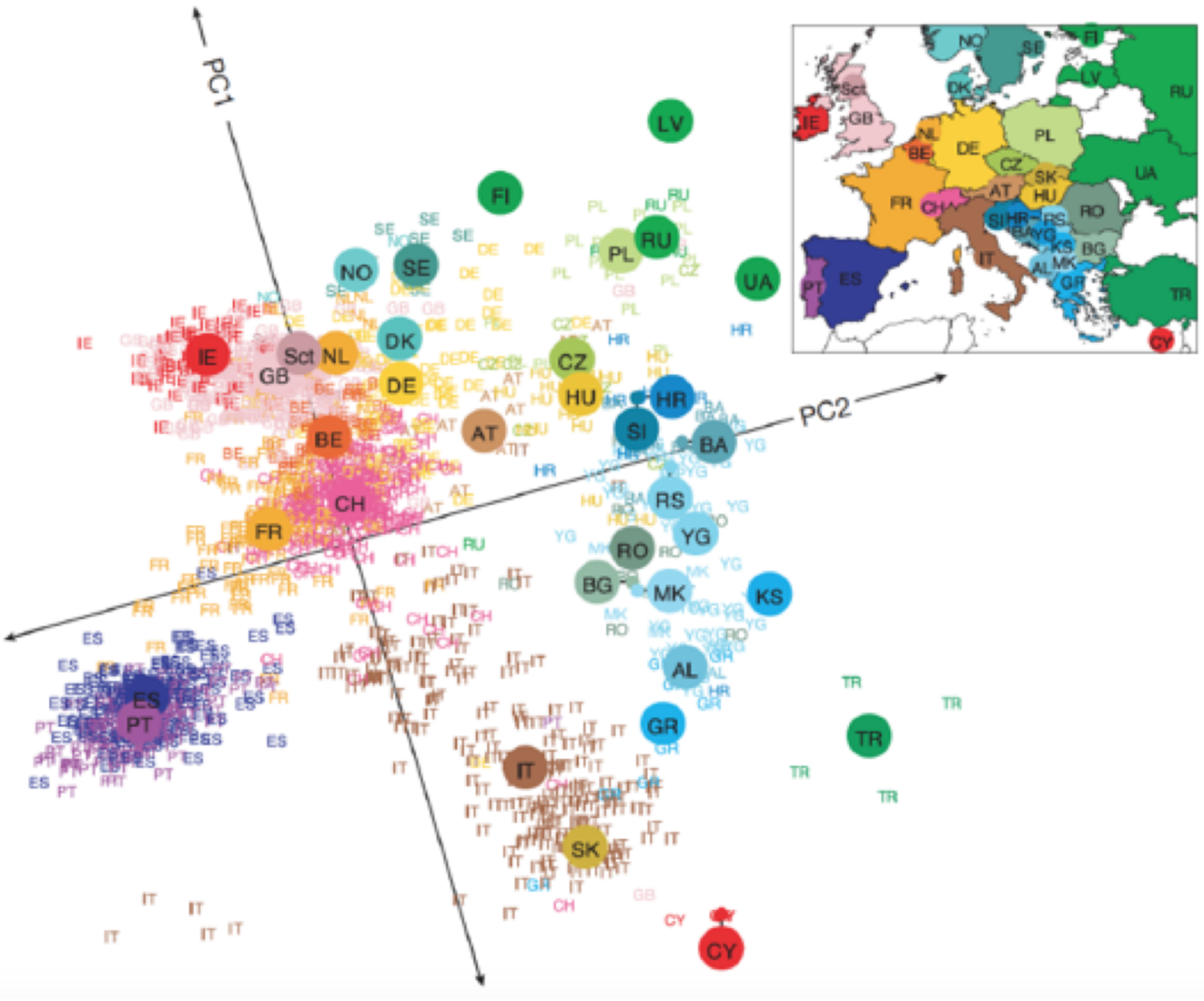
Understanding data

Understand the "structure" within the data

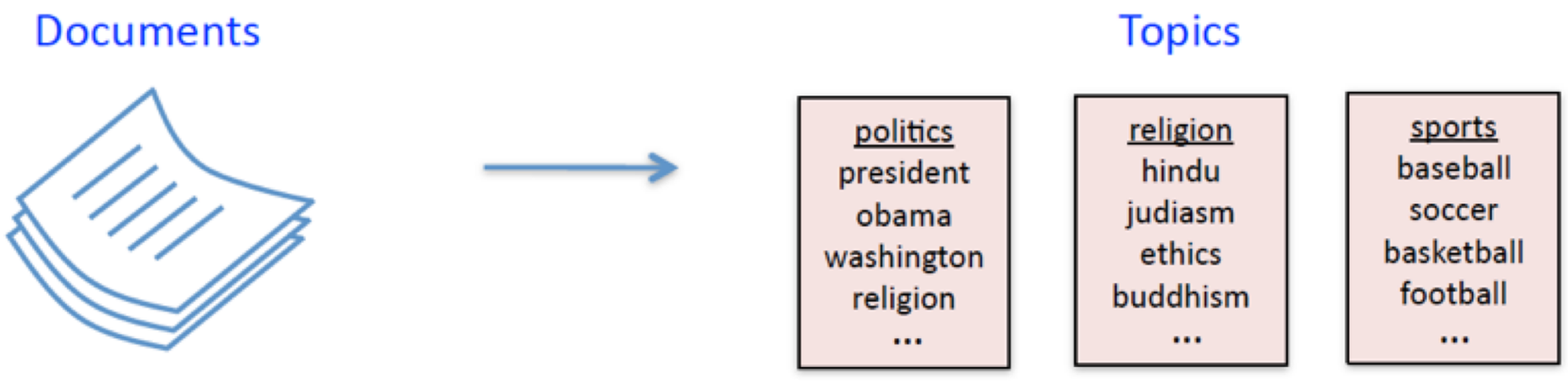


Customer Data	Clustering	Segmentation
Text Documents	Topic Model	Document Classification
	⋮	
Natural Image	Neural Network	Generating "fake-real" images

Gene Mirror Geography



Extracting information from text



Generating fake-real images

ah yes, they are truly *fake* (and truth is truth)

Understanding data

Data in Retail pertains

Customers, Products, Operations

Involves Unstructured Data

Retail Questions

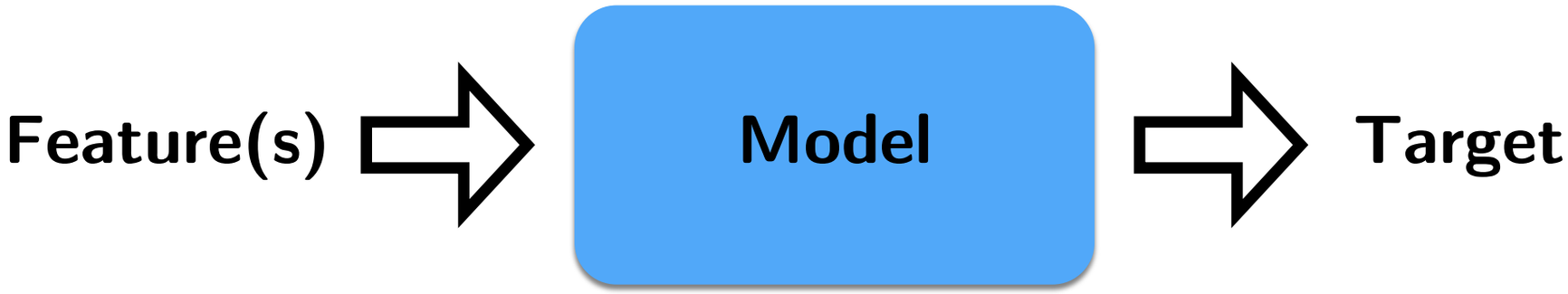
How to Segment Customers per Their Preferences

How to Create Ontology of Products

How to Find (dis)similar Stores

Prediction

Using Prior Data, "Predict" Unknown from Given Observation

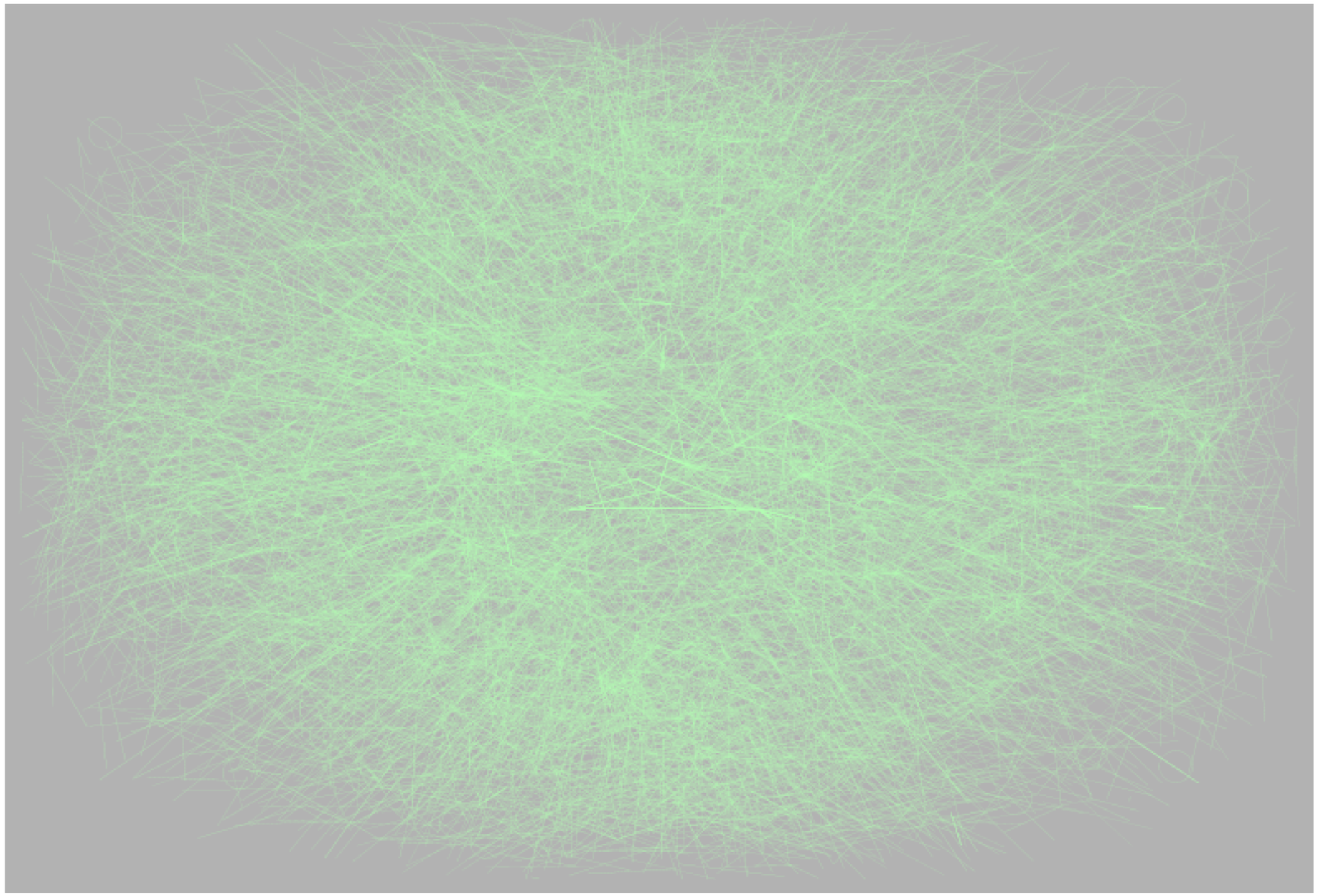


Demographics	Linear Regression	Life Expectancy
Noisy Locations	Kalman Filter	GPS
	⋮	
Image	Neural Network	Which Digit? Cat or Not?

Alzheimer's or not

https://www.nature.com/articles/nature25456

Will it Trend or Not?



Chen, Nikolov and Shah, NIPS 2012

Prediction in Retail

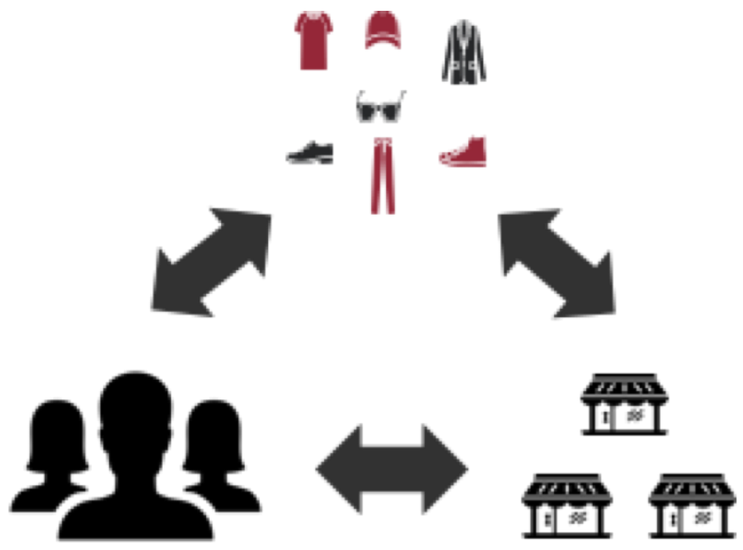
The Prediction Problem in Retail

Granular, Accurate Demand Forecasting

Extremely Hard As Data is Sparse, Unstructured

Natural Solution

Stitch data across people, product, location, ...



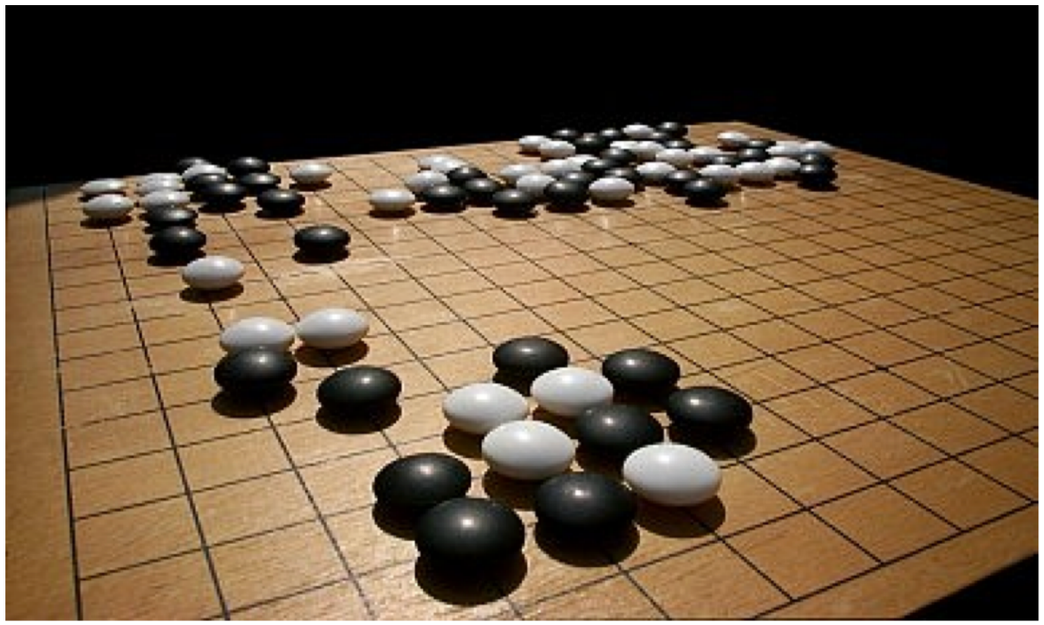
Decision Systems

Using History on State, Decisions and Outcomes
learn "Policy" to take Action Given current State

Market, Portfolio | Model Predictive Control | Buy / Sell / Hold

Board of Go | Neural Network | Next Move

Automated Go Player



AlphaGo Zero is a version of DeepMind's Go software AlphaGo. AlphaGo's team published an article in the journal *Nature* on 19 October 2017, introducing AlphaGo Zero, a version created without using data from human games, and stronger than any previous version.[1] By playing games against itself, AlphaGo Zero surpassed the strength of AlphaGo Lee in three days by winning 100 games to 0, reached the level of AlphaGo Master in 21 days, and exceeded all the old versions in 40 days.[2]

Decision Systems in Retail

The Decision Problem in Retail

Given Demand Forecast, What To Make / Buy / Charge / Sell

Challenging as Possible Options / Actions are Massive

and Opportunities to "Explore" are too Little

Retail Challenge / Opportunity

Data-driven decisions for Make / Buy / Charge / Sell

(getting Model Predictive Control to work will be a Success)

Causal Inference

Extracting causal connection between
Conditions of the occurrence of an *Effect*

Conditions ⇨ **Effect**

Drug Efficacy	Randomized Control
Study of Brain	Granger Causality
⋮	⋮
London Cholera Outbreak	Natural Experiments

Experiment design

Does Beer cause Baldness?

	Bald	Not bald
Drinks beer	49%	2%
no beer	1%	48%

Randomized Control is Essential

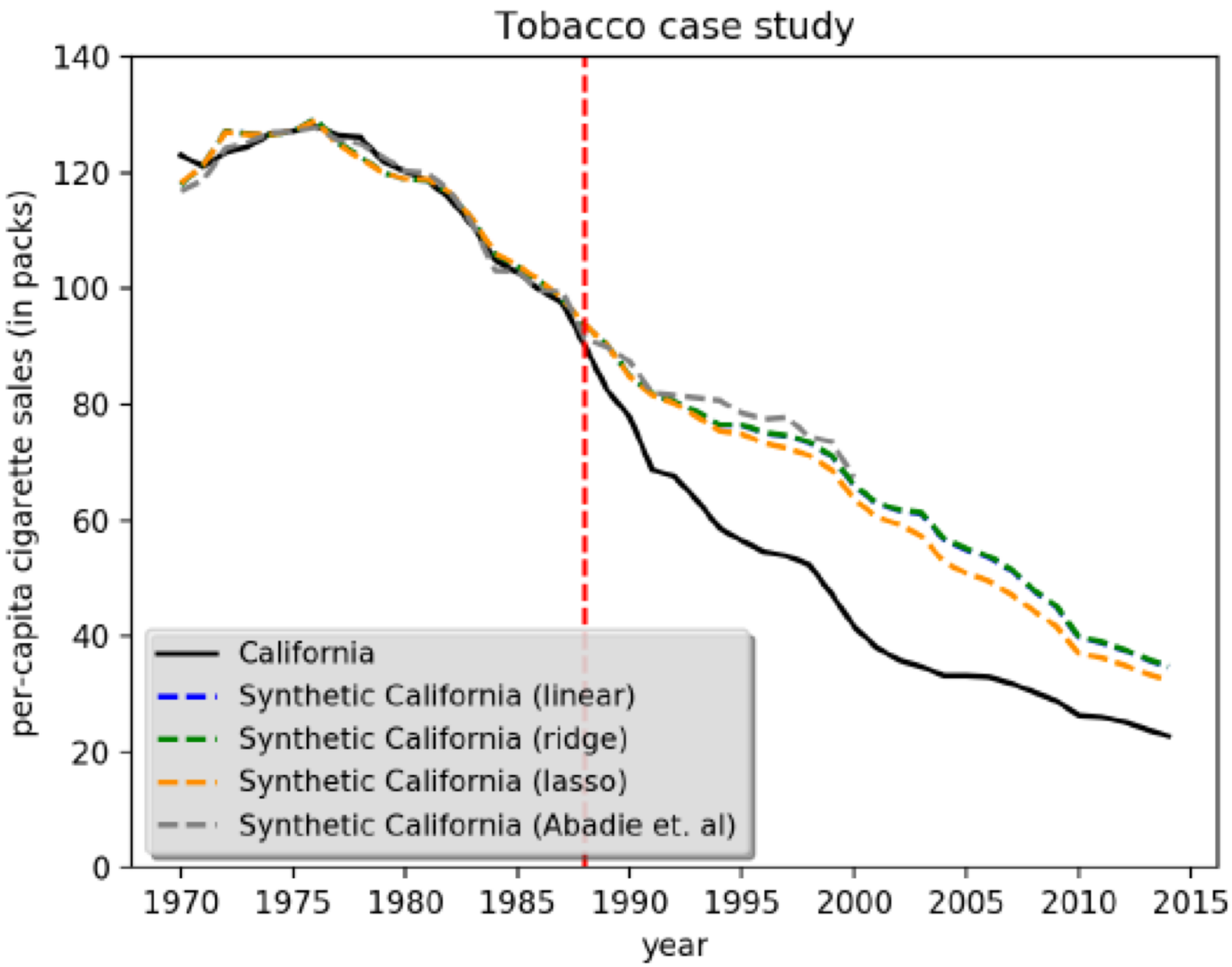
Overcome Confounding Factors

Gold standard for

Drug discovery, e-commerce (A/B testing), ...

When experiment design is not feasible

Impact of CA Prop 99



Robust Synthetic Control (Amjad, Shah, Shen 2018)

Causal Inference in Retail

Causal Inference in Retail

What "policy" lead to better performance?

Challenging as Randomized Control is Not Feasible

(other than few online scenarios)

Retail Challenge / Opportunity

Accurate attribution of performance to actions

(it's like CRISPR for Genetics)

Prexcel:
Machine Learning in Excel

Benchmark Performance

Problem	Benchmark Dataset	Prexcel	Best in the State-of-Art
Classification	MNIST	0.8%	0.21%
Regression	Ames Housing	0.107	0.117
Time Series	Electricity Demand	2.08%	2%
Matrix/Tensor Estimation	MovieLens 20M	0.771	0.765

In Summary

Tremendous Progress in Machine Learning

A Lot Is Done, Lot More Needs to be Done

But Most Importantly

It Needs to be Brought in Hands of Everyone