Recap 5
Dimensionality Reduction & Clustering

ISLR 12, ESL 14, tSNE

Jilles Vreeken
Krikamol Muandet

UNIVERSITÄT DES SAARLANDES

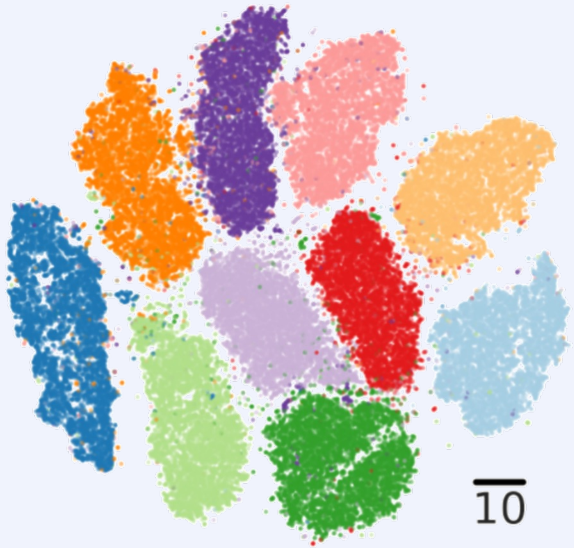
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Lecture Recap 1

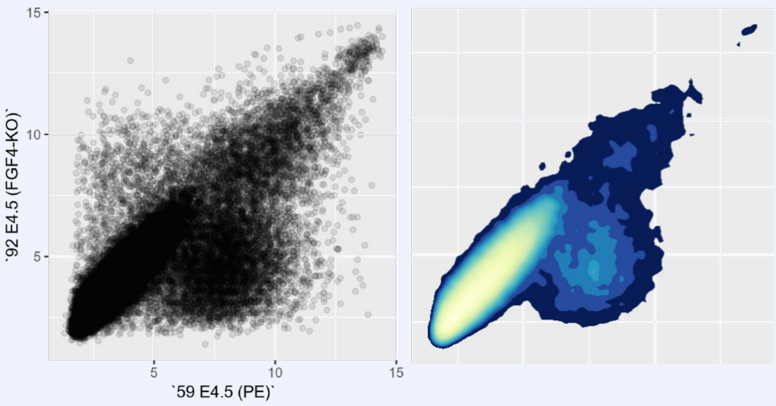
- Unsupervised Learning
  - No prediction of known label Y, instead exploration, visualization and clustering

Flavors of Unsupervised Learning

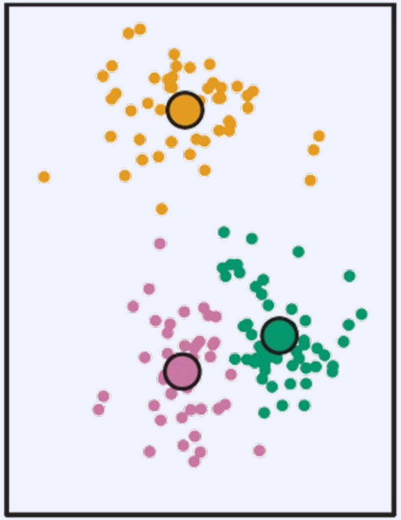
t-SNE Embedding of MNIST
(Dimensionality Reduction)



Density Estimation
& Data Visualization



K-Means
(Clustering)



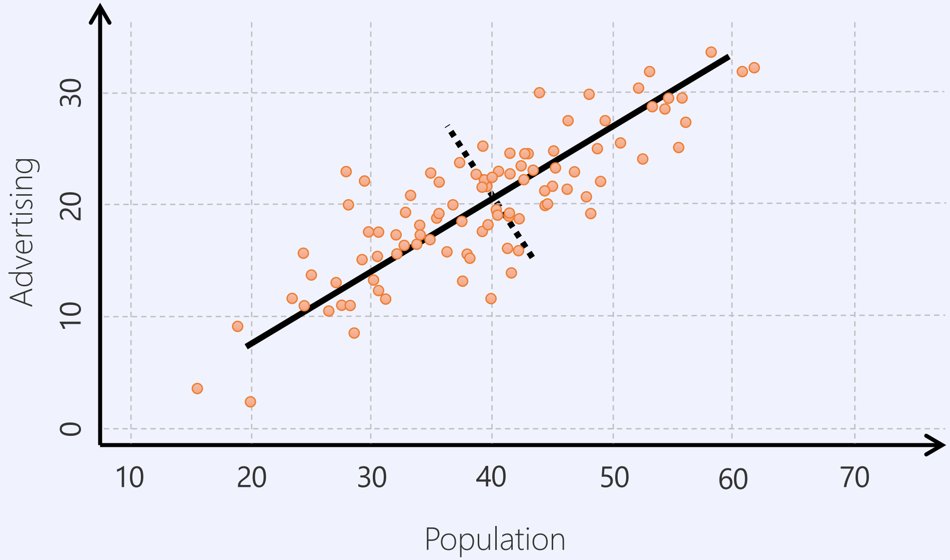
Lecture Recap 1

- Unsupervised Learning
  - No prediction of known label Y, instead exploration, visualization and clustering
- Principal Component Analysis
  - Dimensionality reduction: transform high-dimensional data into interpretable low dimensional data
  - Manifold hypothesis: data lives on low-dimensional manifold (e.g. 10 MNIST digits)
  - Find *linear* combination ϕ of features X that maximizes the *variance* of the embedded data

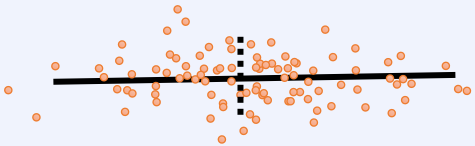
Principal Component Analysis

Example: population and ad spending for 100 different cities shown as circles

- Data are roughly linear along one direction with a small variance along a second direction
- Solid line indicates the first principal component (PC) direction, and dotted line the second PC
- Most of the variation is along the first PC



- The PCs define a new coordinate system



- Project points onto the first PC

Lecture Recap 1

- Unsupervised Learning
  - No prediction of known label Y, instead exploration, visualization and clustering
- Principal Component Analysis
  - Dimensionality reduction: transform high-dimensional data into interpretable low dimensional data
  - Manifold hypothesis: data lives on low-dimensional manifold (e.g. 10 MNIST digits)
  - Find *linear* combination ϕ of features X that maximizes the *variance* of the embedded data

- t-SNE
  - PCA drawback: only linear mapping possible. T-SNE: designed for high dimensional data
  - Idea: embed neighbouring points in high-dim space close to each other in low-dim
  - Minimize KL-Divergence between source and target distribution

Stochastic Neighbor Embedding