EuroVis 2016
Machine Learning Methods
in Visualization for Big Data
6 June 2016, Groningen, the Netherlands

Machine Learning Methods in Visualisation for Big Data

Daniel Archambault[1]
Ian Nabney[2]
Jaakko Peltonen[3]

1Swansea University

2Aston University

3Aalto University and University of Tampere

Outline

- Why machine learning and visualisation?
- How do (can) we work together?
- Preliminary Introduction to ML
- Agenda for the Tutorial

Entering the Big Data Age

- Machine learning and visualisation methods have the same goal: finding interesting things in data
  - machine learning – emphasis on algorithms
  - Visualisation – emphasis on interfaces/interaction
- Machine learning has the advantage of scalability in terms of the data sets it can handle
- Visualisation has the advantage of interactive exploration

Data Models

Machine learning is the computer-based generation of models from data

A model is a parameterised function from input attributes to a target prediction

Parameters in the model express the hidden connection between inputs and predictions.

Parameters are learned from data by changing them to optimise a cost function that expressed the model quality. Optimisation may be an iterative process and there may not be a unique global solution.

‘Non-parametric’ models usually work by combining local models for individual data points: issues with scaling.

Uncertainty

“Doubt is not a pleasant condition, but certainty is absurd”
-- Voltaire

Real data is noisy.

We are forced to deal with uncertainty, yet we need to be quantitative.

The optimal formalism for inference in the presence of uncertainty is probability theory.

We assume the presence of an underlying regularity to make predictions. Bayesian inference allows us to reason probabilistically about the model as well as the data.

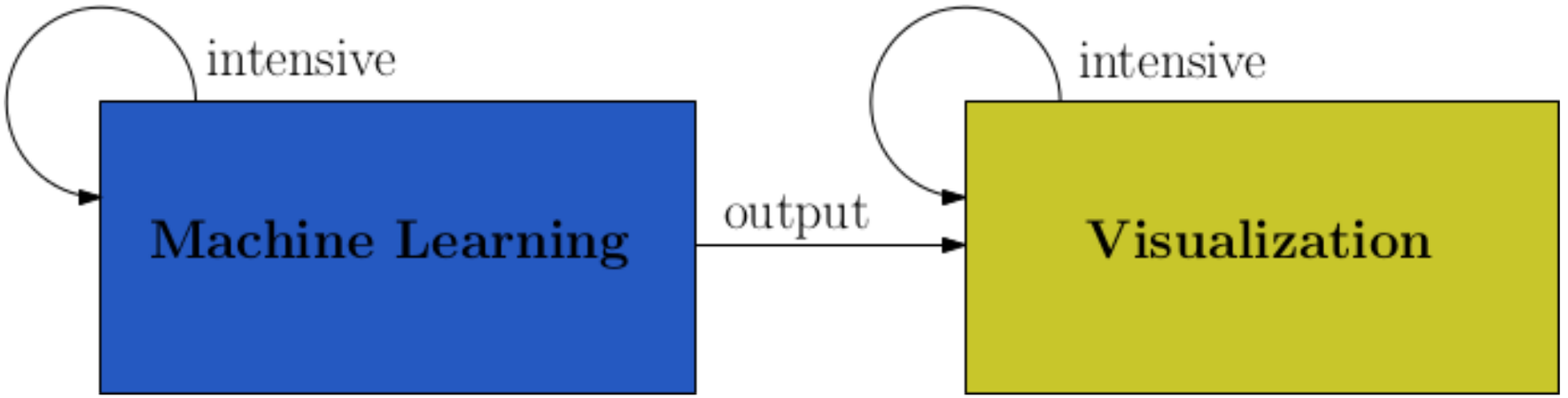
Why ML and Visualisation?

- Idea: Allow interactive visualisation methods to scale to larger data sets
  - Find a way to leverage the advantages of each approach

- How can we do this well?
  - still an open research question
  - however there are some solutions

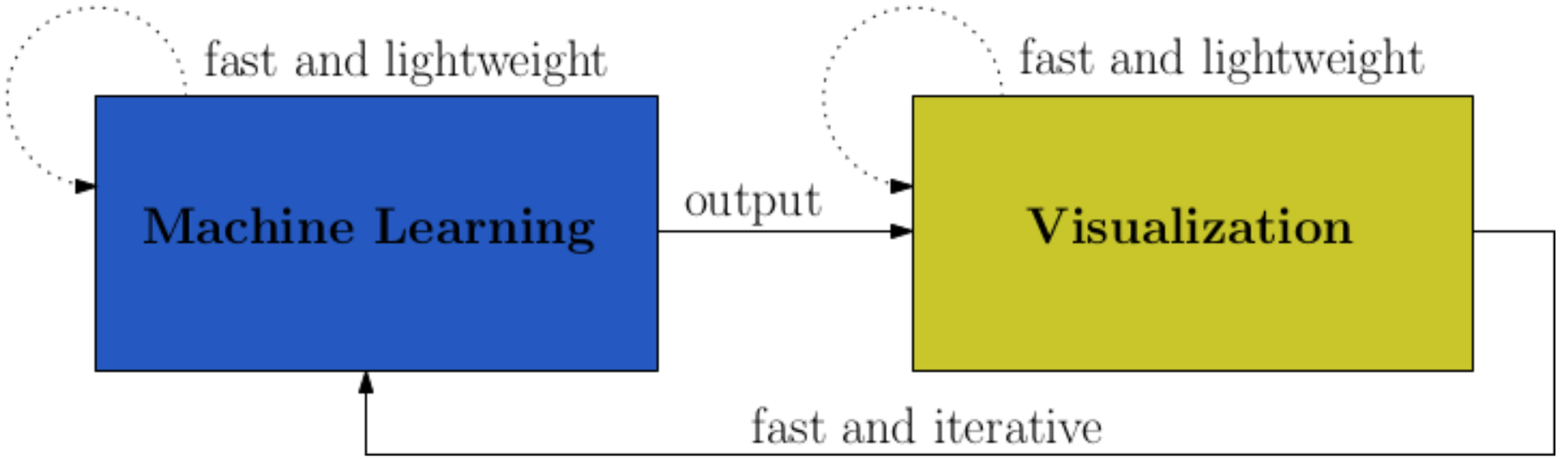
Visualisation as Output

- An easy way is to use machine learning as a preprocessing step
- In this way, sumarize first and visualise second
- Issue: adjusting machine learning results



Steerable Visualisation

- Visualisation and machine learning are integrated
- Quick approximate results give overview of data
- Once satisfied, run heavyweight process
- More fruitful partnership, fewer systems



Overview of the Day

[14:20-15:20] Dimensionality Reduction
[15:20-15:40] Software Activity
[15:40-16:10] Coffee Break
[16:10-16:40] Clustering
[16:40-17:20] Multivariate Graphs and Graph Mining
[17:20-17:45] BYOD: Bring Your Own Data

Dimensionality Reduction

- Generative models and non-generative models
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS) and its variants
- Methods that preserve similarities (neighbourhood relationships)

Software Activity

- An opportunity to work with these models and dimensionality reduction techniques is provided.
- The session will consist of demonstrations and activities

Clustering

- Generative models: mixture models and links to dimensionality reduction
- Bayesian methods for generative models
- Hierarchical models

Multivariate Graphs and Graph Mining

- Community finding approaches
- Evaluation of Ouput using Normalised Mutual Information (NMI)
- Multivariate graphs and graph mining
- Discussion on how to best integrate ML and Visualisation

Bring Your Own Data (BYOD)

- Opportunity for participants to work with tutorial leaders to apply some of the learned techniques to their own data