University of BRISTOL

Advanced Data Analytics: Course Introduction

Professor Ian Nabney

ian.nabney@bristol.ac.uk

bristol.ac.uk

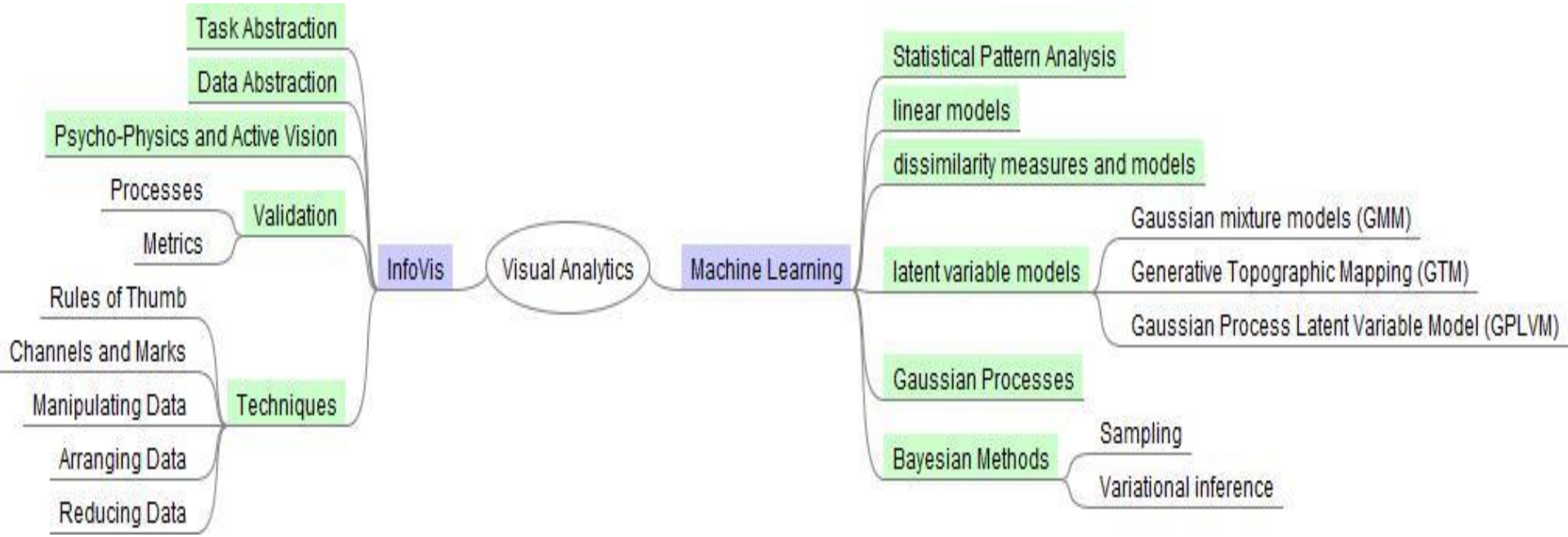
University of BRISTOL

Overview

- Structure of the unit
  - This unit extends the material taught in the co-requisite unit 'Introduction to Data Analytics' by giving students a solid grounding in contemporary advanced machine learning. The unit also covers some more advanced techniques and analytical approaches to information visualisation.
  - In visual analytics, such methods serve to as useful tools to change the data representation, e.g. through dimensionality reduction, or as a way of analysing visual data) in a framework of statistical pattern recognition. Machine learning topics covered by this unit include: principles of Statistical Pattern Recognition (probabilistic models for data, curse of dimensionality generalisation error, bias-variance dilemma); linear models (Probabilistic Principal Component Analysis; Discriminant Analysis); generalised dissimilarity mappings and neighbour embedding techniques; Gaussian Processes; latent variable models (Gaussian Mixture Models, Generative Topographic Mapping and Gaussian Process Latent Variable Model); Bayesian model regularisation and combination; feature selection; challenges of large datasets and potential solutions.
  - In text analytics, such methods serve to produce powerful analyses that traditional methods fail to deliver using learning, along with crowd-sourced data annotation.
- Data visualisation and machine learning

Week	Asynchronous Content	Lectorial/Discussion	Lab Class
1	Topographic dimensionality reduction. Dissimilarity measures (stress etc.) and models for dimensionality reduction (Neuroscale, SNE).	Error analysis	Application of dissimilarity-measure based models for data projection
2	Statistical pattern recognition. Principles of statistical modelling, likelihood, curve fitting, model selection. Linear models for dimensionality reduction (PCA)	Information theory	Application of linear data projection
3	Latent variable models. Density estimation: kernel estimation, Gaussian mixture models, probabilistic PCA, Generative Topographic Mapping	Evaluation of dimensionality reduction methods	GTM and variants for data projection
4	Advanced Information Visualisation 1: Task Abstraction, Visual Queries, Validation	Visual queries: example analysis	Tableau: data sources and calculations
5	Advanced Information Visualisation 2: Structuring Space, Multiple Views	Visual thinking and design critiques	Tableau: data joins and advanced calculations
-	READING WEEK		
6	Advanced Information Visualisation 3: Colour perception and advanced mapping marks	Visual thinking conclusions	Tableau: advanced interaction
7	Bayesian methods in machine learning: principles, sampling, basic variational methods (and application to Gaussian mixture models)	Optimisation algorithms	Variational Bayesian GMM for robust density modelling
8	Gaussian Processes and Gaussian Process Latent Variable Models	Summary and overview of machine learning	Coursework
9	Advanced Text Analytics 1: Skipgram word embeddings; neural text classifiers, RNNs, LSTMs	Neural networks for text analytics	Jupyter notebook: neural text classifier
-	SPRING VACATION		
10	Advanced Text Analytics 2: Self-attention, transformers, transfer learning; question answering	Large language models and transfer learning	Jupyter notebook: pretrained language models
11	Advanced Text Analytics 3: continuing advanced text analytics 1 and 2; data annotation	Data annotation	Completing lab report based on Jupyter notebooks

Course Mind Map: Visual Analytics



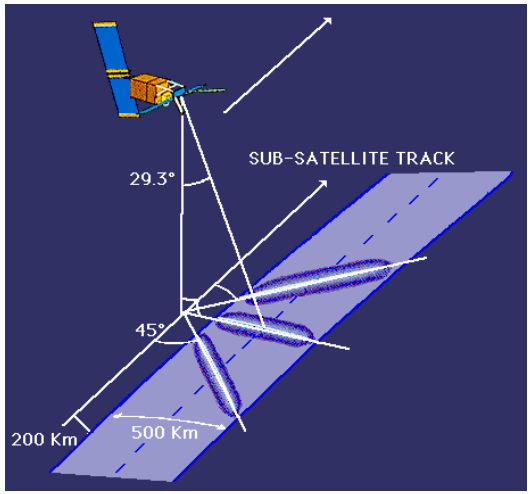
University of BRISTOL

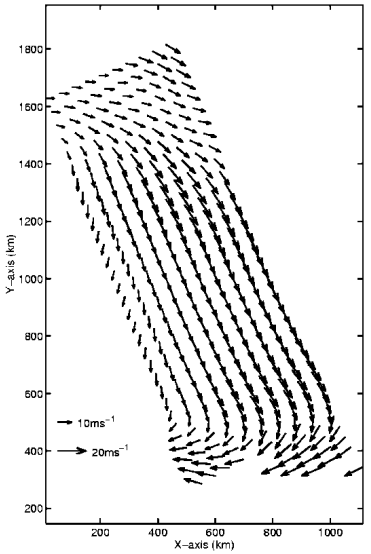
Data projection

- The goal is to project data to a lower-dimensional space (usually 2d or 3d) while preserving as much information or structure as possible.
- Once the projection is done standard information visualisation methods can be used to support user interaction. These may need to be modified for Big Data.
- The quantity and complexity of many datasets means that simple visualisation methods, such as Principal Component Analysis, are not very effective.

Hidden Knowledge

- Understanding the vast quantities of data that surround us is a real challenge
- We can understand more of it with help. Machine learning is the computer-based generation of models from data.
- Parameters in the model express the hidden connection between inputs and predictions.





University of BRISTOL

Uncertainty

Doubt is not a pleasant condition, but certainty is absurd.

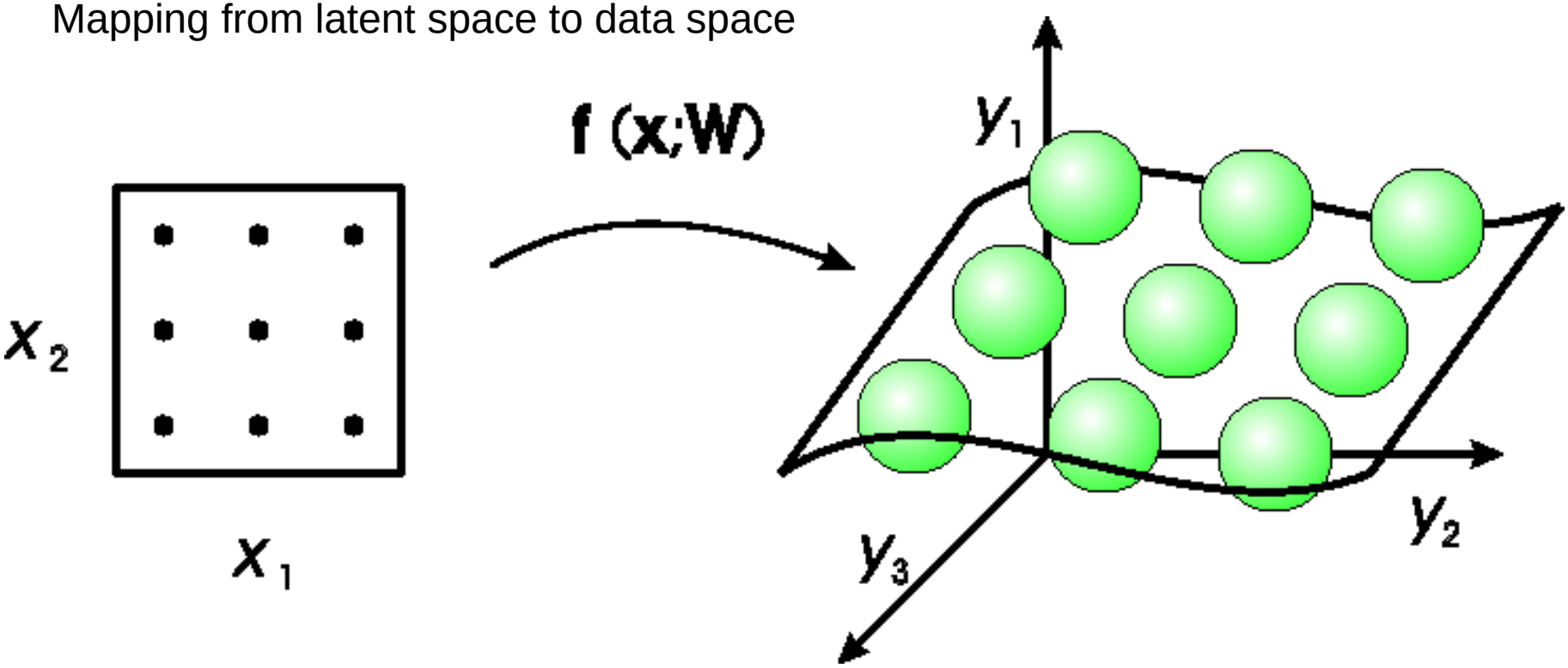
Voltaire

- 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.

University of BRISTOL

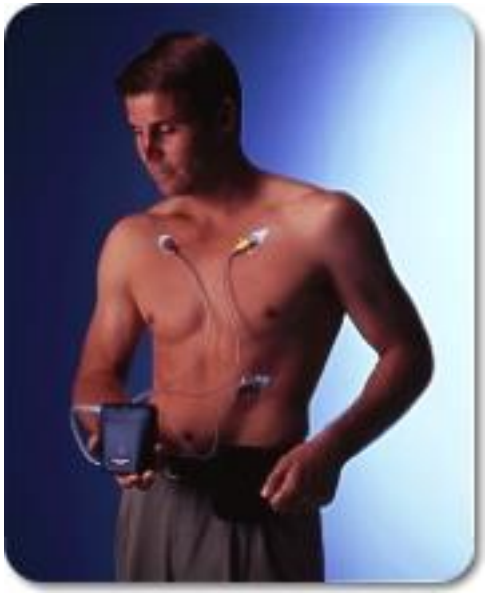
Generative Topographic Mapping

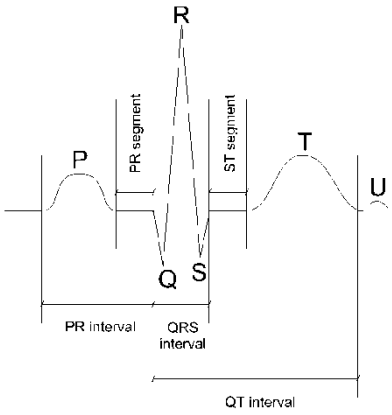
Mapping from latent space to data space

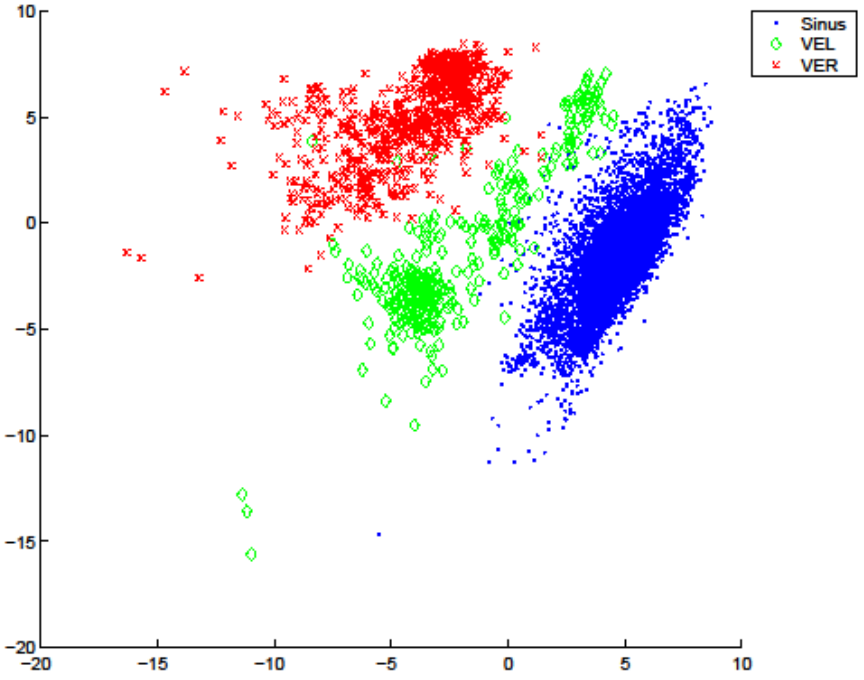


A thick rubber sheet studded with tennis balls. GTM defines p(y|x;W); use Bayes' theorem to compute p(x|y*;W) for a given point y* in data space.

University of BRISTOL



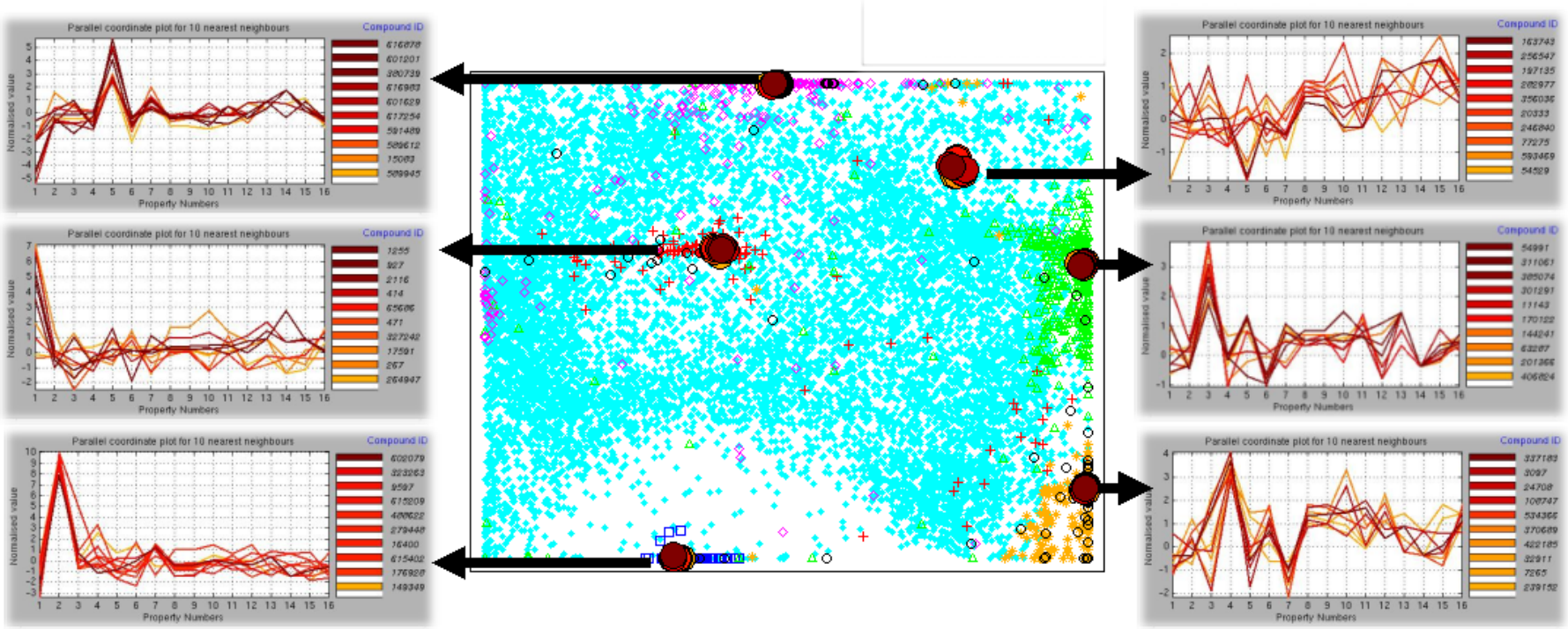




What can we learn from this?

University of BRISTOL

Interactive Visualisation Tool



University of BRISTOL

Conclusions

- We need to understand the vast quantities of data that surround us; visualisation and machine learning can help us in that task.
- Models can be used to uncover the hidden meanings of data.
- A probabilistic approach to modelling data provides many benefits.
- It is a multivariate, multi-skilled, collaborative effort.