CS-GY-9223

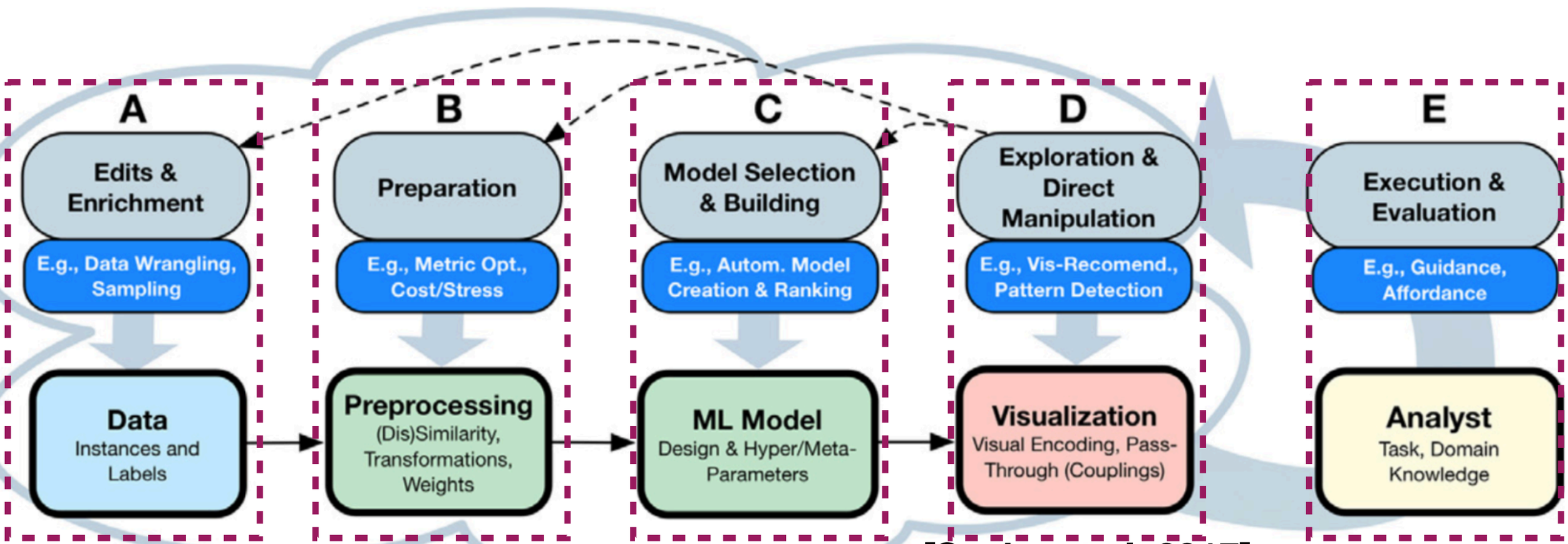
Designing Visualizations for Machine Learning

Course materials based on CS 8395-03 Visual Analytics & Machine Learning
By Matt Berger, Vanderbilt University
https://matthewberger.github.io/teaching/vaml/spring2019/

Interactive Machine Learning

- Wait, why make machine learning interactive?
- Major purpose of ML: **automation**.
- What is the role of the human in ML?
- To help answer this question, we need to consider the different pieces in which a machine learning model is prepared, constructed, and evaluated.

Conceptual Framework for VA/ML



[Sacha et al. 2017]

Let’s make this concrete

- Machine Learning can be roughly categorized into:
    - Supervised learning: we are given *source inputs* and *target outputs* (labels, annotations, etc..), the objective is to construct a model from training data that is able to *predict outputs* from novel *inputs*.
    - Unsupervised learning: we are only given *unlabeled data*. We want to **learn the data distribution**. Examples: linear subspace, nonlinear manifold, clusters, generative models.

Supervised Learning

- Classification, regression
- Different scenarios: semi-supervised, transductive, etc..
- Even more specific:
  - multi-label classification
  - multi-task learning
  - low-shot learning (zero-shot, few-shot)

Unsupervised Learning

- Clustering: vanilla clustering, hierarchical clustering, clustering with side information, etc..

- Dimensionality reduction: linear techniques, nonlinear techniques

- Generative models: autoencoders, adversarial techniques, etc…

Where should we place interactive data visualization?

- We need an understanding of:
    - **problem domain**
    - **tasks**
    - **data**
- Let us first start with data

Exercise: Data in Classification

- Training data
- Validation data
- Test data
- Loss function
- Hyperparameters
- Model parameters
- Input features
- Output labels

- Learned features (training/test)
- Training/validation predictions
- Training/validation loss
- Test predictions/loss
- Multiple models

Exercise: Tasks in Classification

- High-level:
    - How well does my model generalize?
    - How well does my model transfer (e.g. fine-tuning)?
    - Why is one model better than another?
    - What is the sensitivity of the model to: hyperparameters? model parameters? input features?
    - Why does training fail?
    - How do I interpret a (single) prediction made by the model?
    - What does the space of predictions tell me about the model?

Breaking Down Tasks

- How well does my model generalize?
  - How do we *measure* generalization?
    - Can generalization be boiled down to a single number?
    - Generalization across different categories? Across individual instances?
    - How can we enable the human to explore the space of predictions/losses/categories/instances?

Exercise: Data in Clustering

- Input features
- Side information
- Loss function
- Hyperparameters
- Initialization (if nonconvex)
- Cluster measures
- Validation data
- Measures during optimization

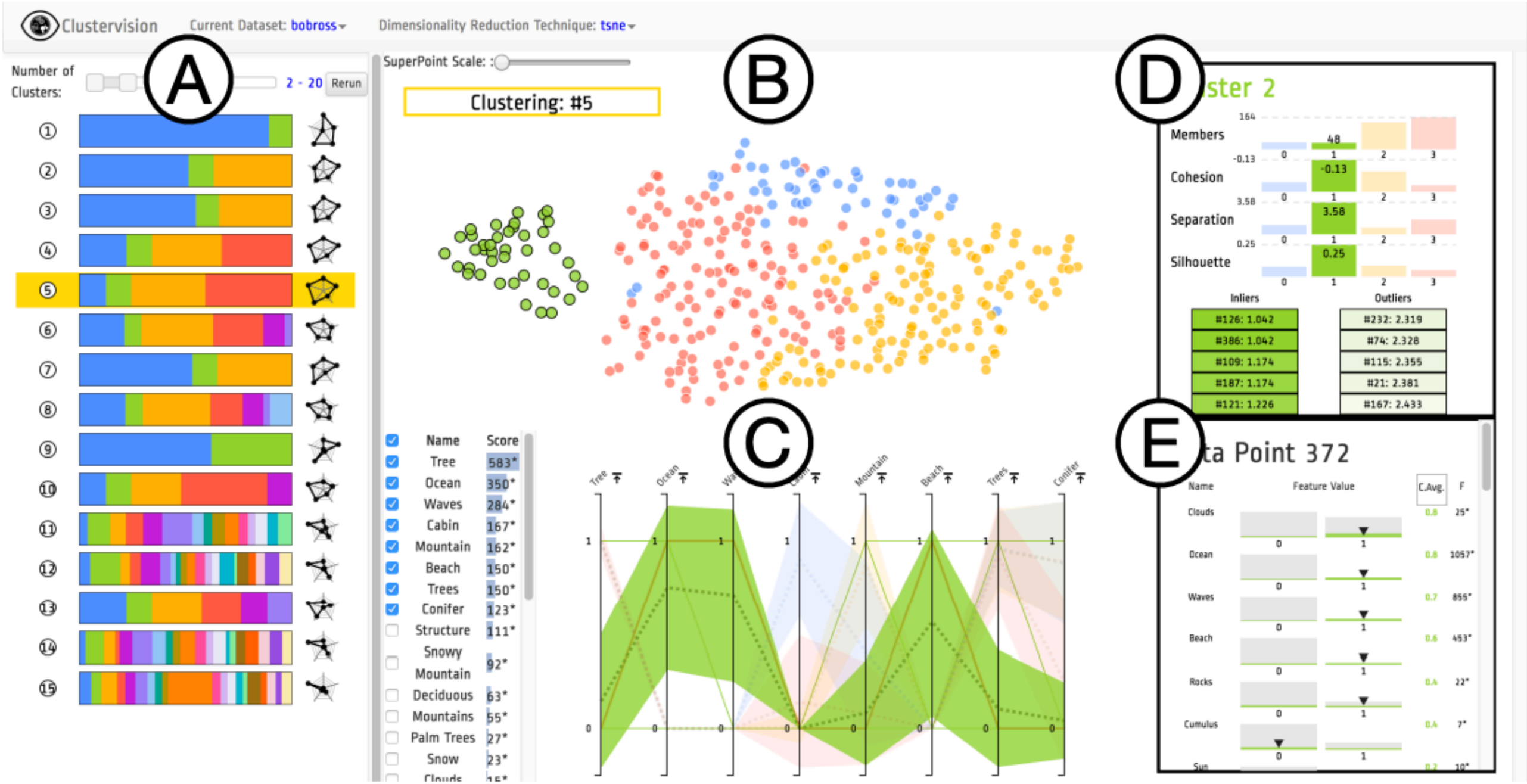
Exercise: Tasks in Clustering

- High-level:
    - How many clusters should there be?
    - What is the sensitivity of certain clusters?
    - How do I compare clustering results?
    - Why are certain data points clustered together (or not)?

Breaking Down Tasks

- Why are certain data points clustered together?
    - What is the feature distribution within a cluster?
    - How do we define a measure of similarity between data points? Between a data point and a cluster?
    - How can we score a feature within a cluster?
    - How do we measure the quality of a cluster?

Visually Understanding Clusters



[Kwon et al. 2018]

Exercise: Data in Dimensionality Reduction

- Input features
- Derived data proximities
- Output projection (typically 2D)
- Side information
- Loss function
- Hyperparameters
- Quality measures

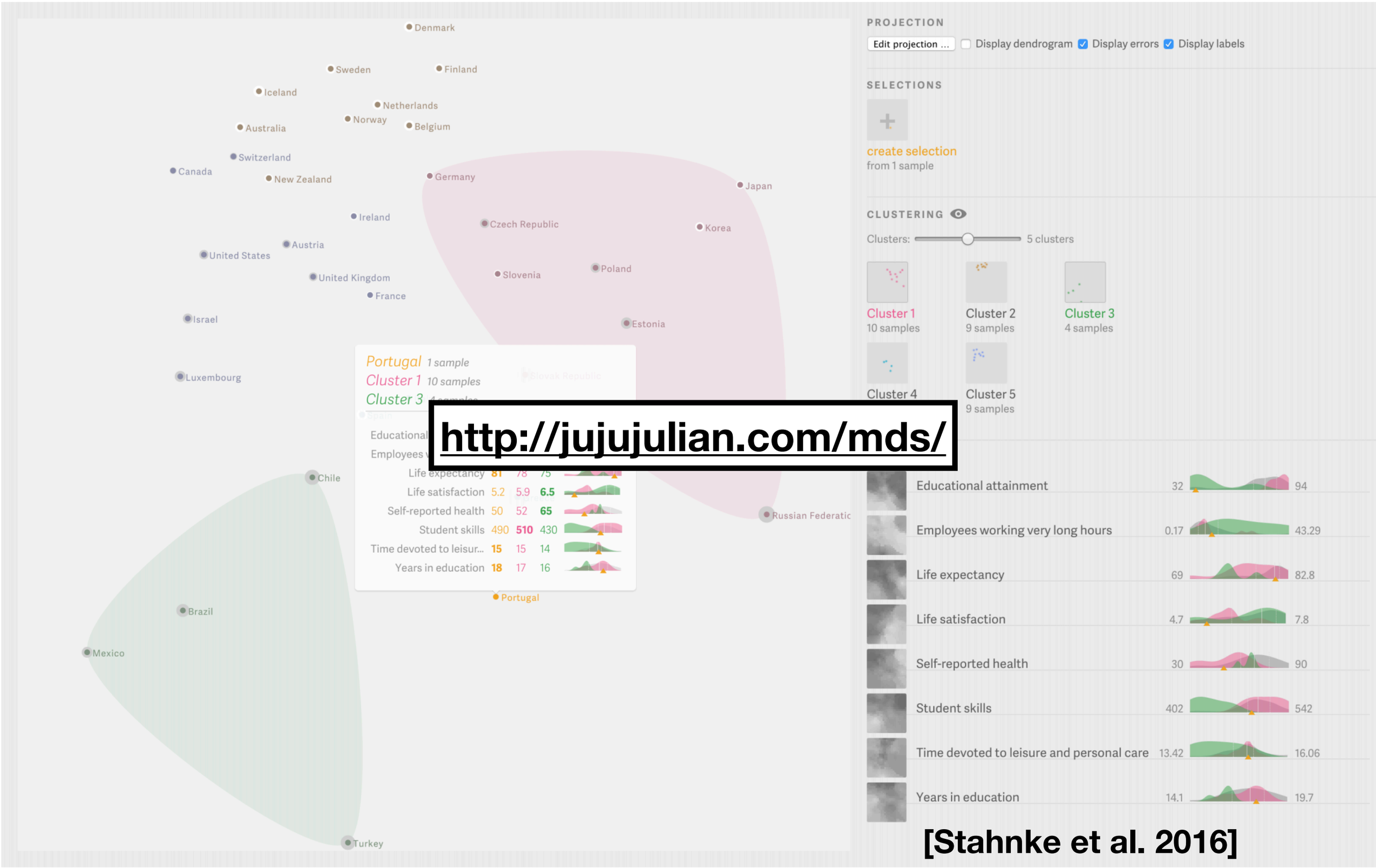
Exercise: Tasks in Dimensionality Reduction

- High-level:
  - What is the distribution of the data?
  - Is the projection faithful to the original data?
  - What is the sensitivity of the algorithm (initialization, hyperparameters, etc..)?
  - What would happen were we to adjust the feature space?

Breaking Down Tasks

- Is the projection faithful to the original data?
- How do we measure the proximity between points?
- How do we compare features across nearby points?
- How do we measure distortion?

Probing Projections



http://jujujulian.com/mds/

[Stahnke et al. 2016]

Strategies for Design

- Ok, great, ***so how do I actually design visualizations!?!***

- Sketch it out on paper! Recall: our visual working memory is limited. Physically drawing helps organize our thoughts.

- Necessary pieces:

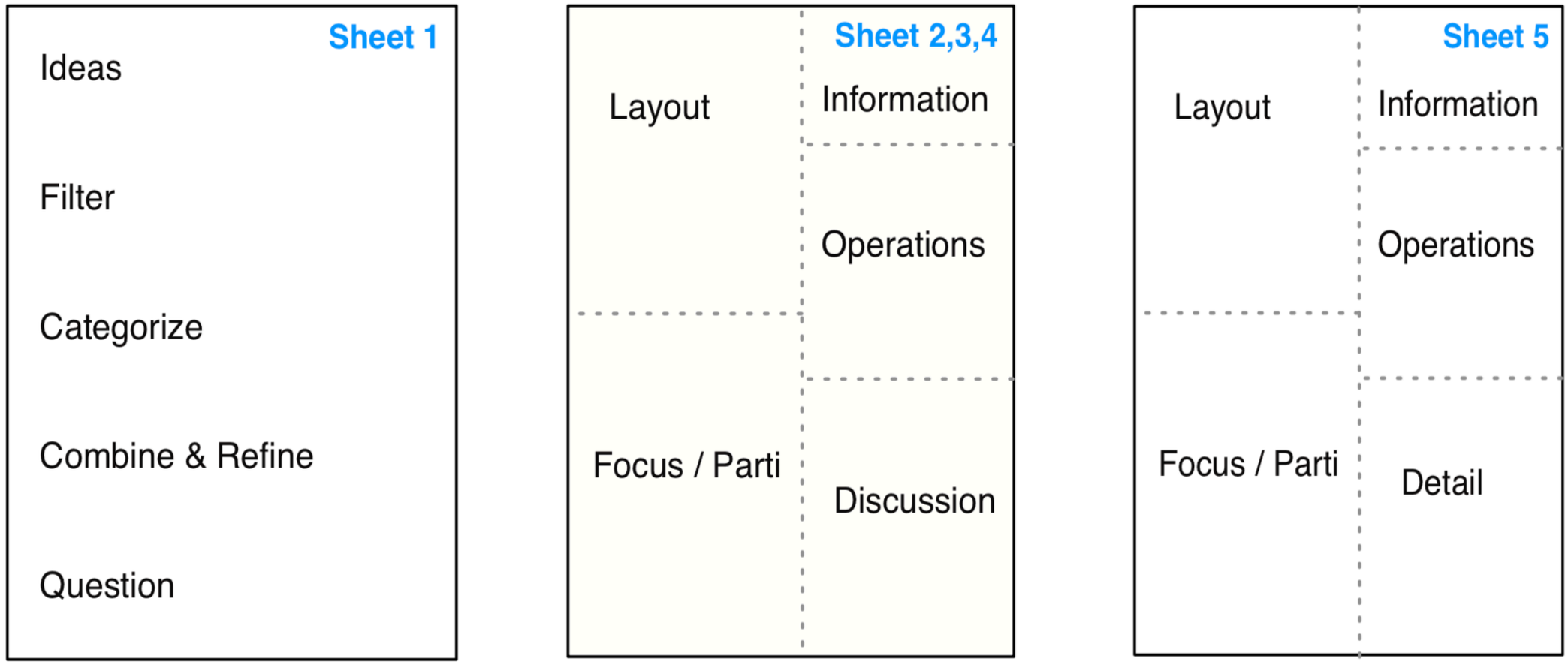
  - **Data**

  - **Task**

  - **Components**

  - **Resources**

Five Design-Sheet Methodology



[Roberts et al. 2015]

Sheet 1: Ideation

- **Ideate**: *Sketch lots of little drawings*. What are all the possible ways you could visualize the data?

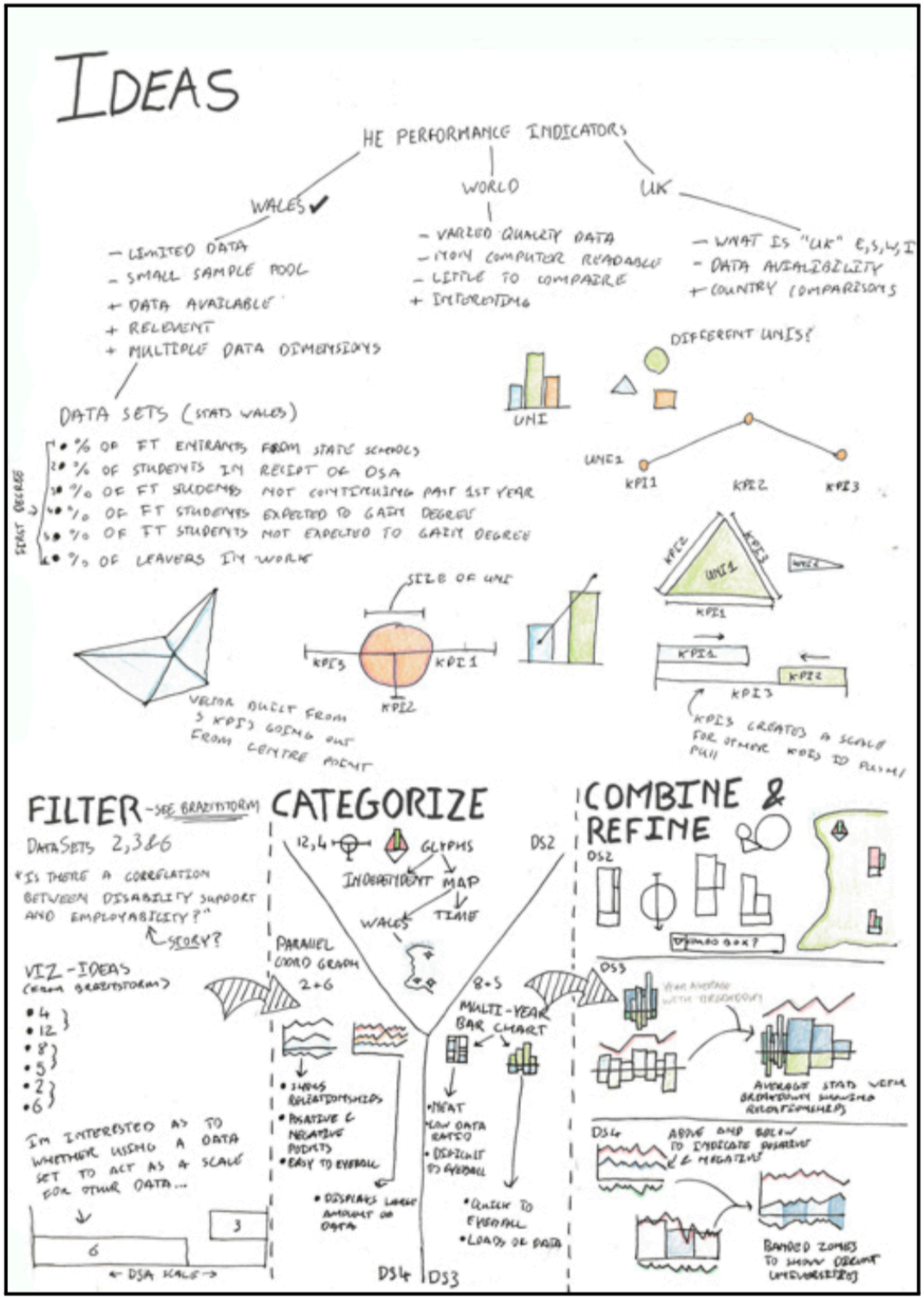
- **Filter**: *Remove irrelevant, or duplicated ideas*. Consider the task abstraction as a guide…

- **Categorize**: *Determine similarities/differences in drawings*. Group ideas, also recognize whats missing.

- **Combine & Refine**: *Organize mini-ideas into bigger solutions*. Consider how they relate to one another - **multiple views**.

- **Question**: *Consult nested design validation*.

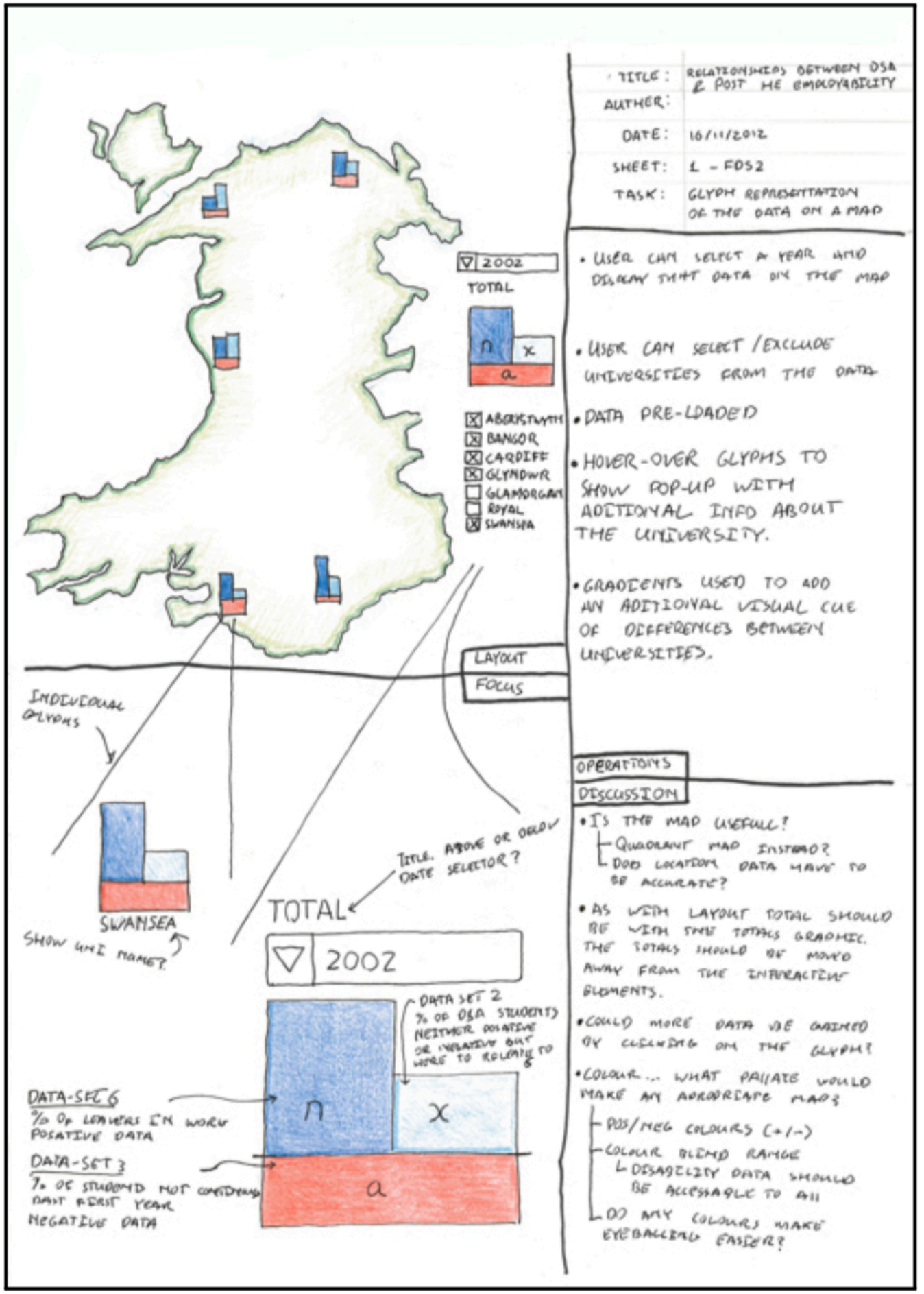
Sheet 1: Example



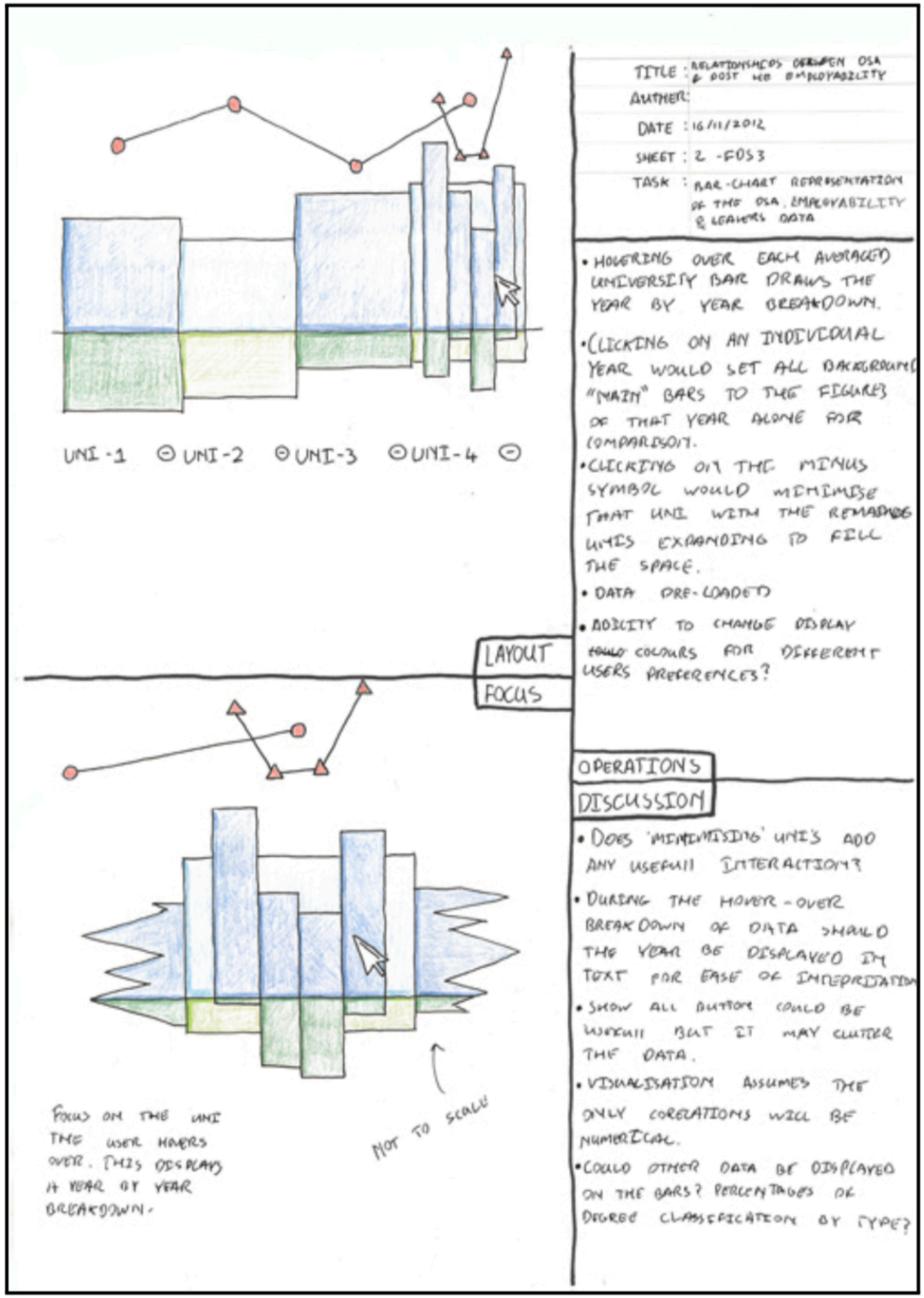
Sheets 2-4: Design Alternatives

- Three diverse design sketches based on Sheet 1.
- **Information**: meta-data like title, task, sheet number, etc…
- **Operations**: action-result pairs. A sketch of your interactions.
- **Focus**: the core concept. Flow diagram of user interaction. Particular piece of your layout.
- **Discussion**: critique the design. Suitability, implementation feasibility, scalability, clarity.

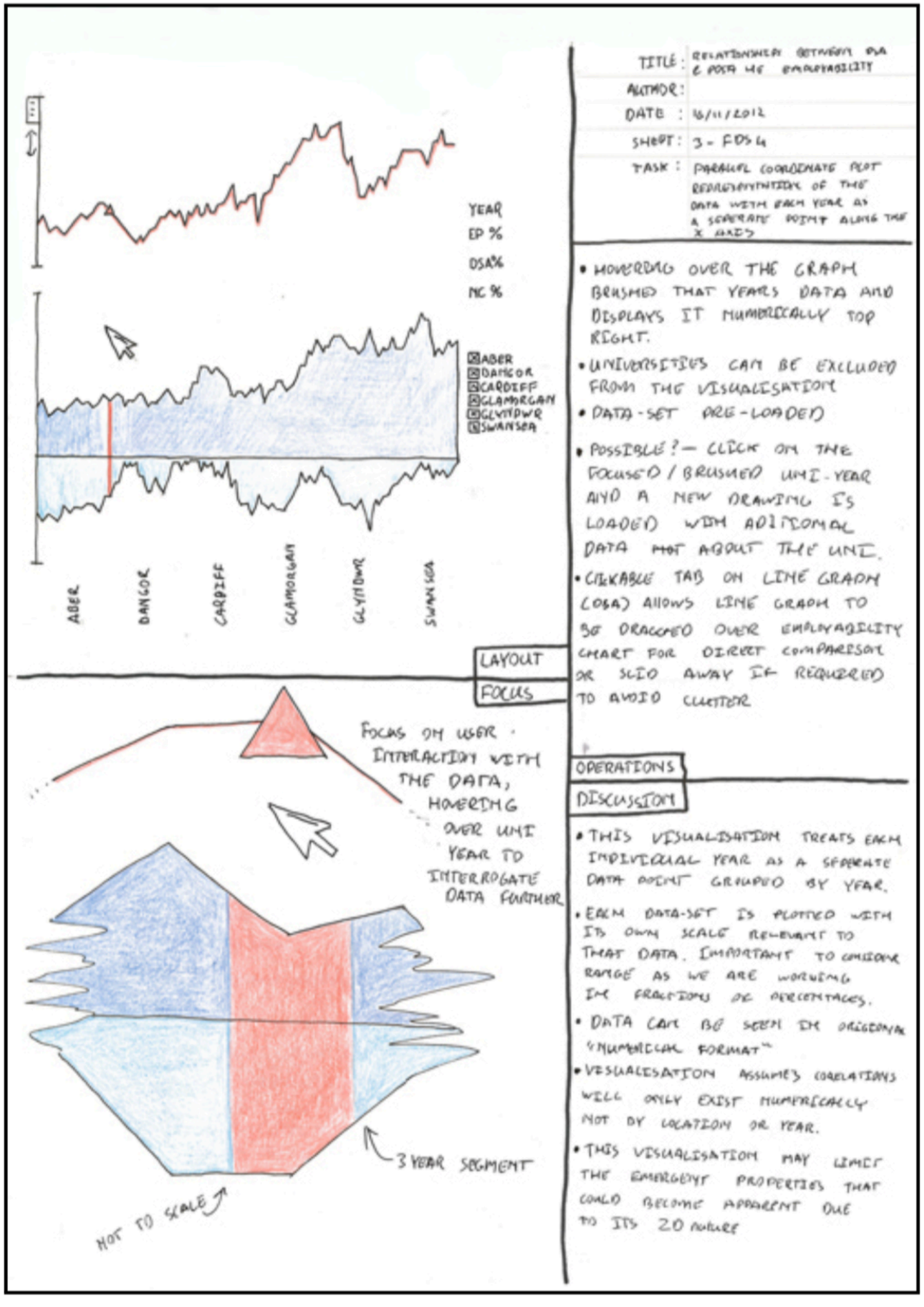
Sheet 2: Example



Sheet 3: Example



Sheet 4: Example

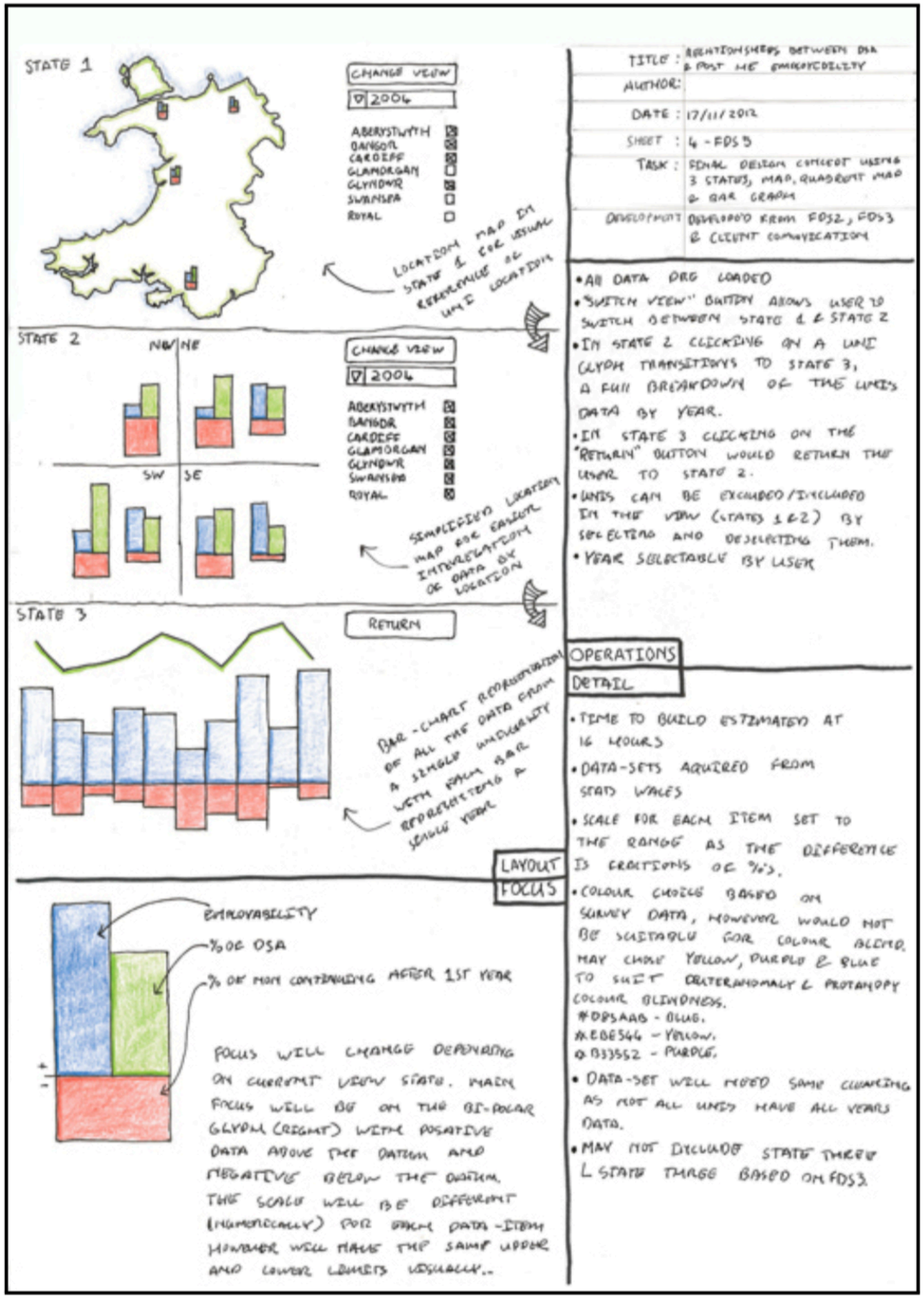


Sheet 5: Realization

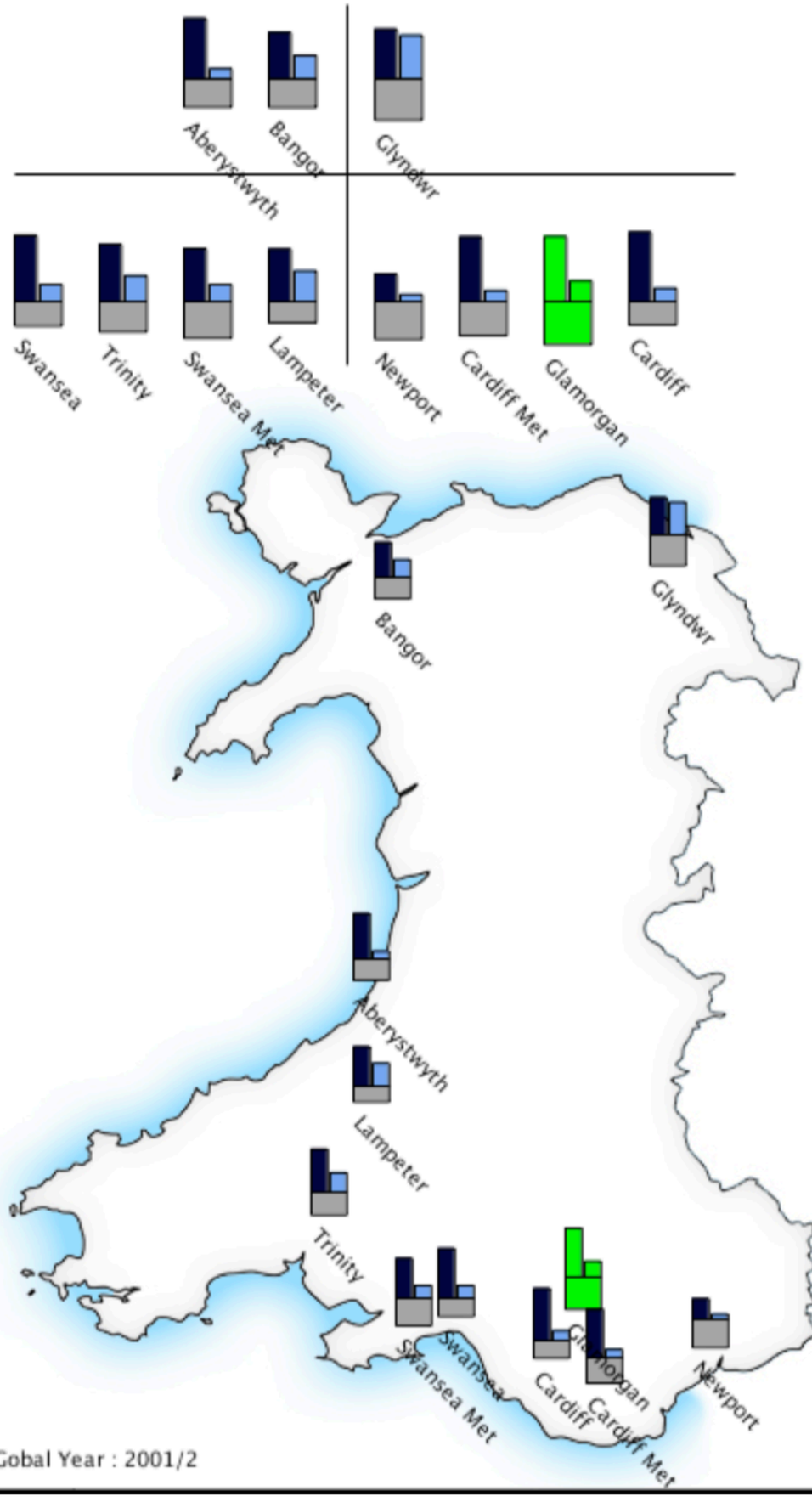
- *The* design.
- Main decision: from critiques of sheets 2-4, pick the best one, or some amalgamation.

- Same components as previous sheet, but also include **detail**: how are you going to actually implement the design?

Sheet 5: Example



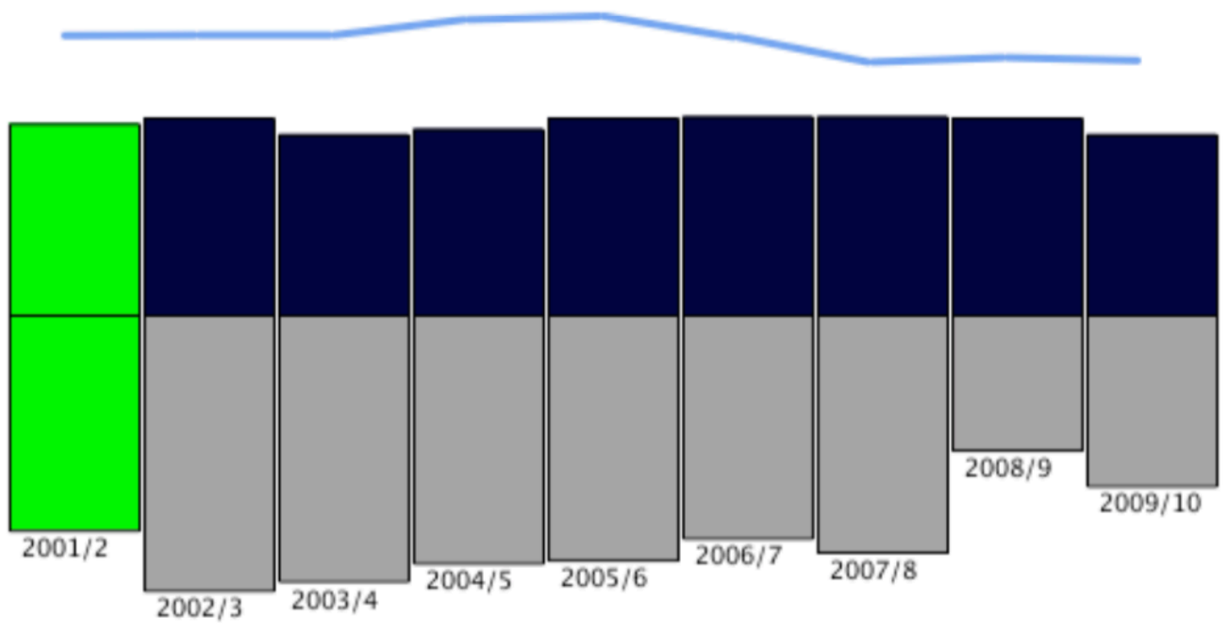
Quadrant Map of Wales, Corners represent NE, NW, SW and SE



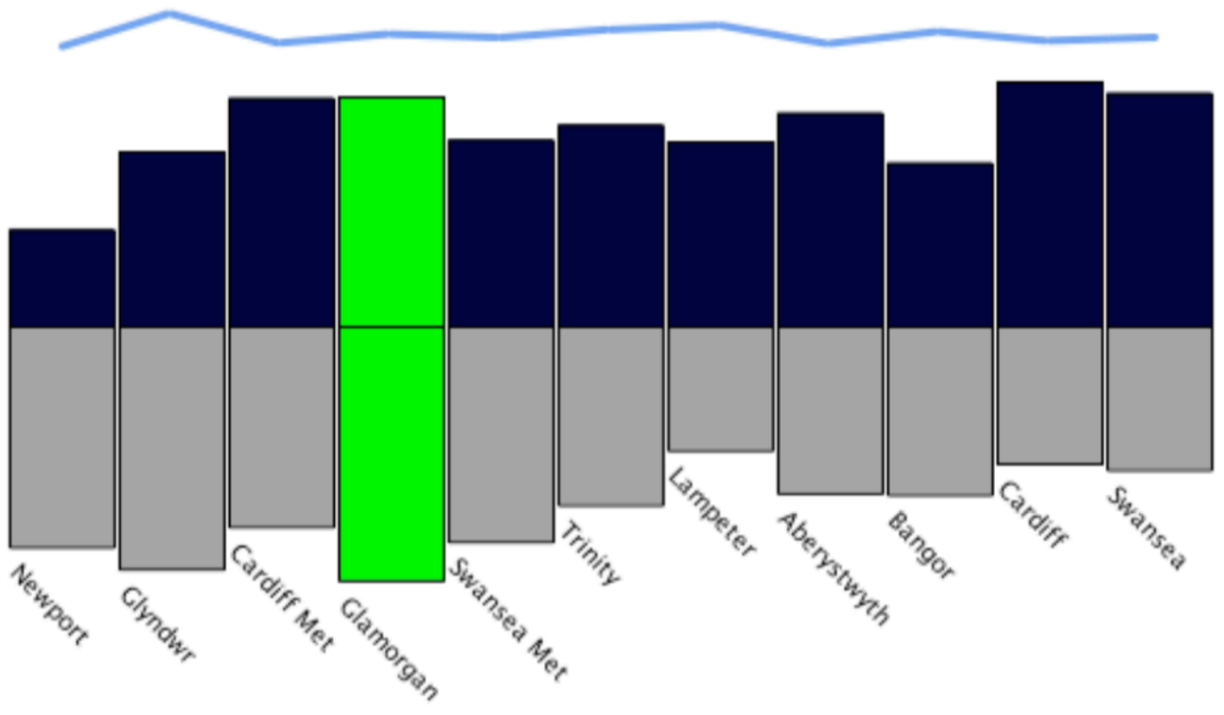
Gobal Year : 2001/2

% Disabled Students Allowance Applications
% Students in work after university
% Of students not continuing past their first year

University : Glamorgan



Year : 2001/2

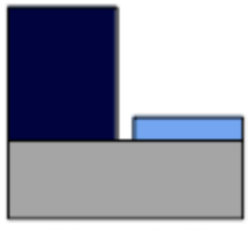


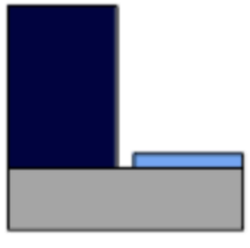
View Options

OVERVIEW
LOCATION MAP
QUADRANT
2001/02
AVERAGE (ALL YEARS)
2001/02
2002/03
2003/04
CHANGE THEME


All Welsh HEIs

All English HEIs


All Scottish HEIs


All N I HEIs