IF 4061
Data and Information Visualization:

Data Viz Methodology:
Learning About Your Data

Semester 2 2018/2019

Dessi Puji Lestari
(dessipuji@stei.itb.ac.id)
School of Electrical and Informatics Engineering

Acknowledgement

Most of the contents of the slides were taken from Andy Kirk. Data Visualization: A Successful Design Process. Pact Publishing. 2012, chapter 3

Content

- How to develop and refine editorial focus
- Acquiring and preparing of the data
  - ensuring it is fit for purpose, and in good shape in advance of the design stage.
- Example of how we can use visual analysis techniques to combine the task of familiarizing data and discovering key insights

Methodology

1. Purpose & parameters
2. Prepare & explore data
3. Formulate questions
4. Design concepting
5. Construct & launch

The Importance of Editorial Focus

"*Good content reasoners and presenters are rare, designers are not.*"

Edward Tufte

- Some of the most influential and esteemed visualization work comes from newspaper and magazine organizations:
  - The New York Times
  - The Guardian (UK)
  - National Geographic (US), etc.
- A key reason behind the success of the work is *the editorial focus*.

Editorial Focus

- The story we are trying to tell
- The key narrative we are looking to portray
- Questions do we wish readers to be able to answer through the visualization?

In 12 Minutes, Everything Went Wrong

How the pilots of Lion Air Flight 610 lost control.

https://www.nytimes.com/interactive/2018/12/26/world/asia/lion-air-crash-12-minutes.html

Preparing and Familiarizing with Data

- Data = raw material, the principle ingredient in creative recipe.
- Data is very important:
  - If there is no data, *or* the data is not interesting there is nothing we can do about it.
  - An incomplete, error strewn or just plain dull dataset will simply contaminate our work

Mechanisms

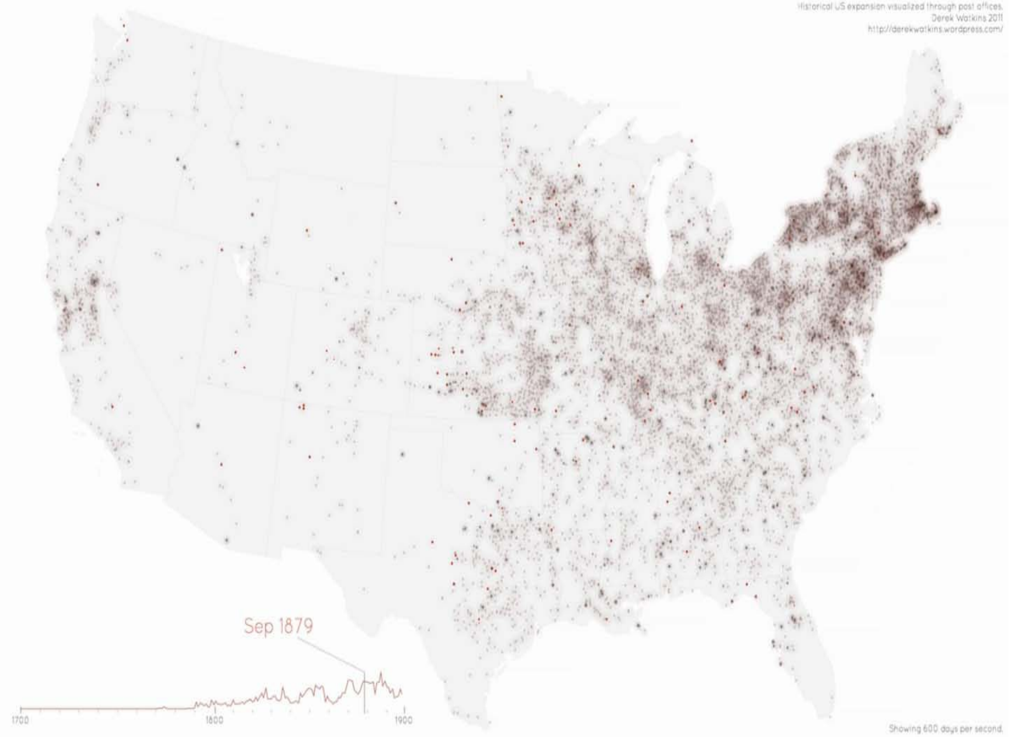
1. Acquisition
2. Examination
3. Understand Data Type
4. Transforming for quality
5. Transforming for analysis
6. Consolidating

1. Acquisition (1)

- Data origins:
  - Obtained from a colleague, client, or other third-party entity
  - A download taken from an organizational system
  - Manually gathered and recorded
  - Extracted from a web-based API
  - Scraped from a website
  - Extracted from a PDF file, etc.

1. Acquisition (2)

- Can be a painful work
- A project to demonstrate the social expansion of the US using the story of the spread of post


Historical US expansion visualized through post offices.
Derek Watkins 2011
http://derekwatkins.wordpress.com/

- Data was scraped from the US Postal website recorded between 1700 and 1900
- Almost 1,500 records had to be discarded and final dataset contains 11,000+ post office locations

2. Examination

- To determine your *level of confidence* in the suitability of the acquired data.
- We can use available tools to quickly:
  - scan, filter, sort, and search through dataset to establish its state of quality

    (Excel, Tableau, or Google Refine)
- What to be examined:
  - **Completeness**
  - **Quality**

Completeness

- Is it all there or do you need more?
  - Does it have all the categories you were expecting?
  - Does it cover the time period you wanted?
  - Are all the fields or variables included?
  - Does it contain the expected number of records?

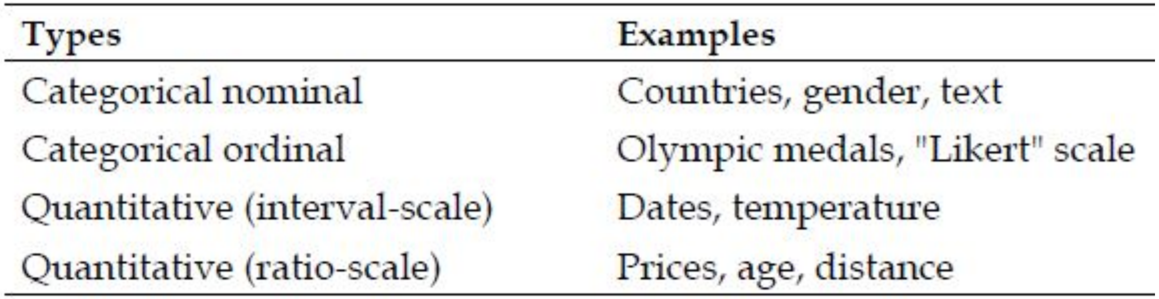
Quality

- Are there noticeable errors?
- Are there any unexplained classifications or coding?
- Any formatting issues such as unusual dates, ASCII characters?
- Are there any incomplete or missing items?
- Any duplicates?
- Does the accuracy of the data appear fine?
- Are there any unusual values or obvious outliers?

To be continued ...

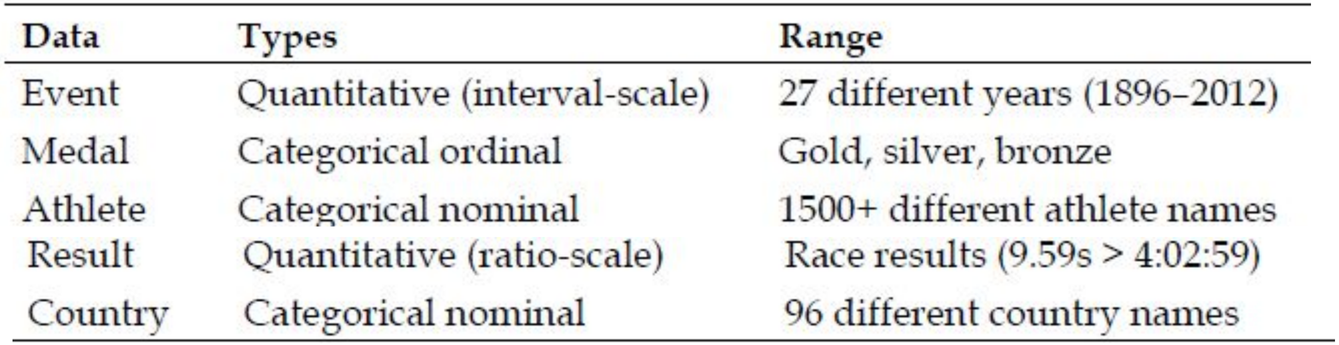
3. Understand Data Type

- To understand the fundamental structure of data in terms of the variables types



Types	Examples
Categorical nominal	Countries, gender, text
Categorical ordinal	Olympic medals, "Likert" scale
Quantitative (interval-scale)	Dates, temperature
Quantitative (ratio-scale)	Prices, age, distance

- Example:



Data	Types	Range
Event	Quantitative (interval-scale)	27 different years (1896–2012)
Medal	Categorical ordinal	Gold, silver, bronze
Athlete	Categorical nominal	1500+ different athlete names
Result	Quantitative (ratio-scale)	Race results (9.59s > 4:02:59)
Country	Categorical nominal	96 different country names

4. Transforming for Quality

- This task is about tidying and cleaning your data by resolving any of the errors we have discovered:
  - Plugging the gaps caused by missing data
  - removing duplicates
  - cleaning up erroneous values
  - handling uncommon characters, etc.

5. Transforming for Analysis (1)

- Preparing and refining data to be used for analysis and presentation:
  1. Parsing (split up) any variables, such as extracting *year* from a date value
  2. Merging variables to form new ones, such as creating a whole name out of *title*, *forename*, and *surname*
  3. Converting qualitative data/free-text into coded values or keywords

5. Transforming for Analysis (2)

4. Deriving new values out of others, such as *gender* from *title* or a sentiment out of some qualitative data

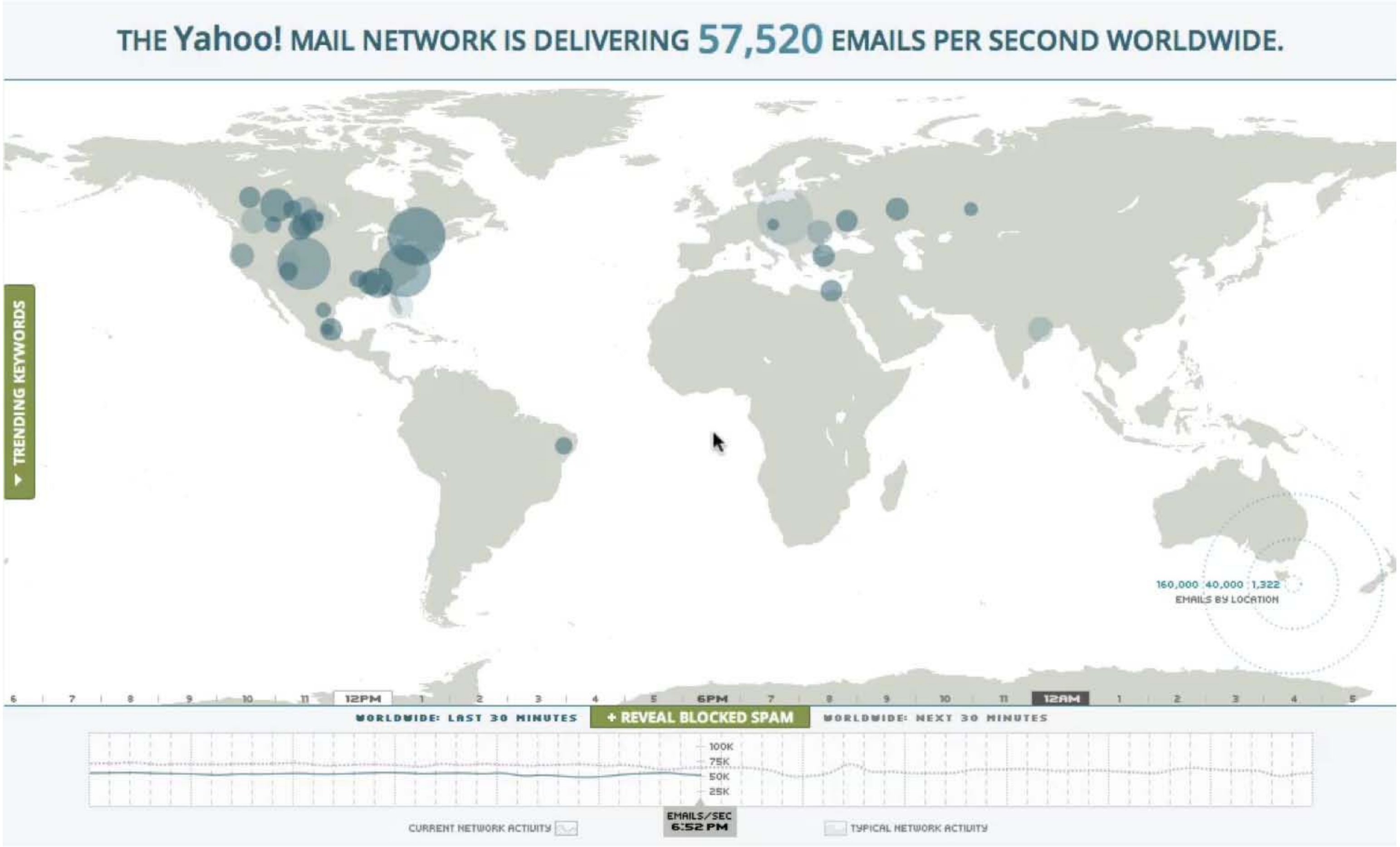
5. Creating calculations for use in analysis, such as percentage proportions

6. Removing redundant data for which you have no planned use

7. Determine what level of resolution you need to present your data.
   - may require you to aggregate or disaggregate your data to achieve get the right level of detail.

Example

- Approximately 5.6 billion e-mails (and a further 20.5 billion spam) sent every day, the sheer amount of data posed a challenge in terms of what level of detail they could reasonably show.



THE Yahoo! MAIL NETWORK IS DELIVERING 57,520 EMAILS PER SECOND WORLDWIDE.

Resolution Options

- **Full resolution**: Plotting all data available as individual data marks.
- **Filtered resolution**: Exclude records based on a certain criteria.
- **Aggregate resolution**: "Roll-up" the data by, for instance, month, year, or specific category.
- **Sample resolution**: Apply certain mathematical selection rules to extract a fraction of your potential data. This is a particularly useful tactic during a design stage if you have very large amounts of data and want to quickly develop mock-ups or test out ideas.
- **Headline resolution**: Just showing the overall statistical totals.

6. Consolidating

- After the examination and preparation work there might still exist certain gaps in your subject matter.
- Additional layers of data may be required to be combined with our existing dataset
  - applied to perform additional calculations
  - or just to sit alongside this initial resource to help contextualize and enhance the scope of our communication.

Tips:

Always spend a bit of time considering if there is anything else you anticipate needing to supplement your data to help frame the subject or tell the stories you want to communicate.

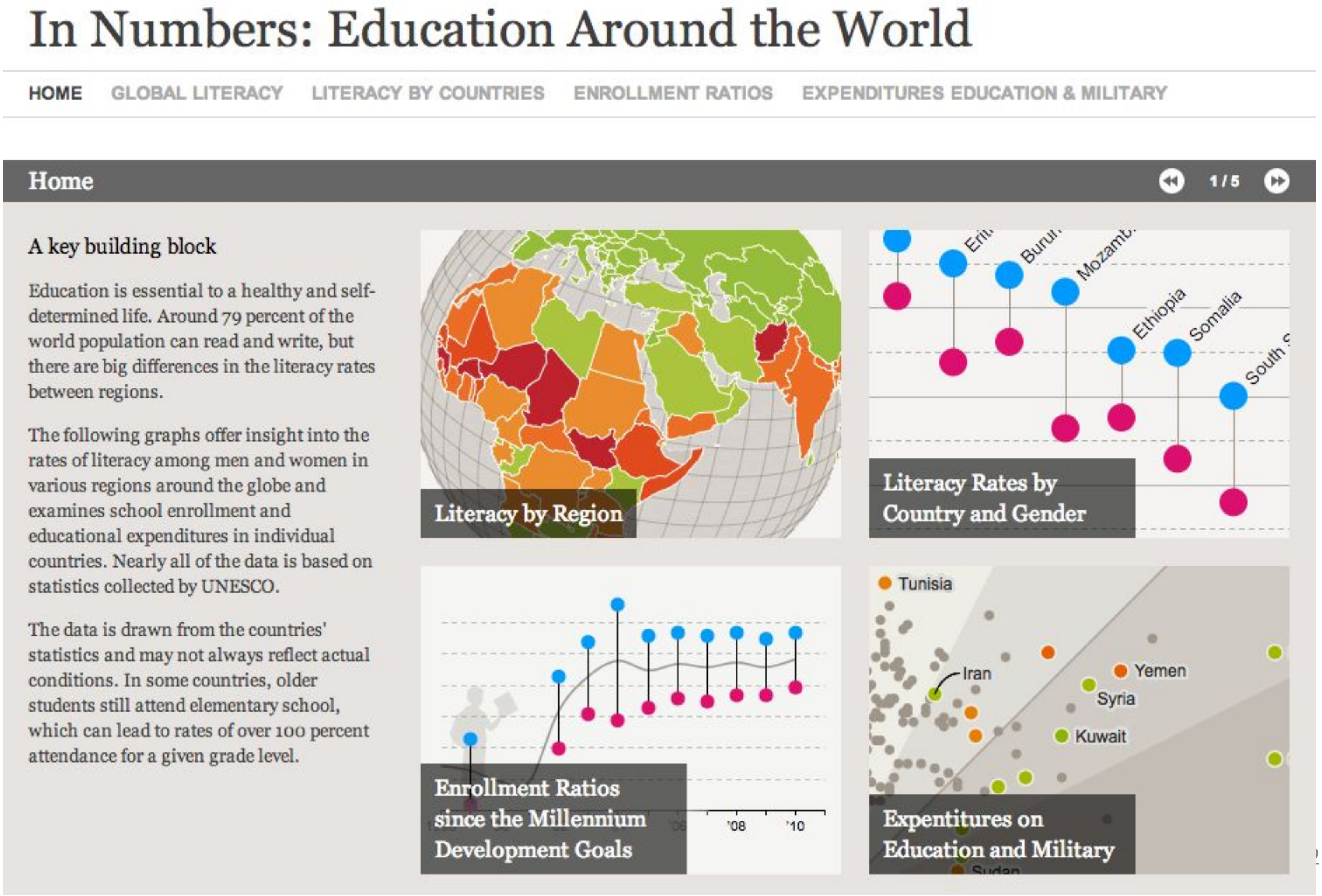
Acquiring, handling, and preparing your data is often the most time-consuming and intensive activity involved in any visualization project

Refining Editorial Focus

"Different forms do better jobs at answering different questions.“

Amanda Cox

Example



In Numbers: Education Around the World

HOME GLOBAL LITERACY LITERACY BY COUNTRIES ENROLLMENT RATIOS EXPENDITURES EDUCATION & MILITARY

Home 1 / 5

A key building block

Education is essential to a healthy and self-determined life. Around 79 percent of the world population can read and write, but there are big differences in the literacy rates between regions.

The following graphs offer insight into the rates of literacy among men and women in various regions around the globe and examines school enrollment and educational expenditures in individual countries. Nearly all of the data is based on statistics collected by UNESCO.

The data is drawn from the countries' statistics and may not always reflect actual conditions. In some countries, older students still attend elementary school, which can lead to rates of over 100 percent attendance for a given grade level.

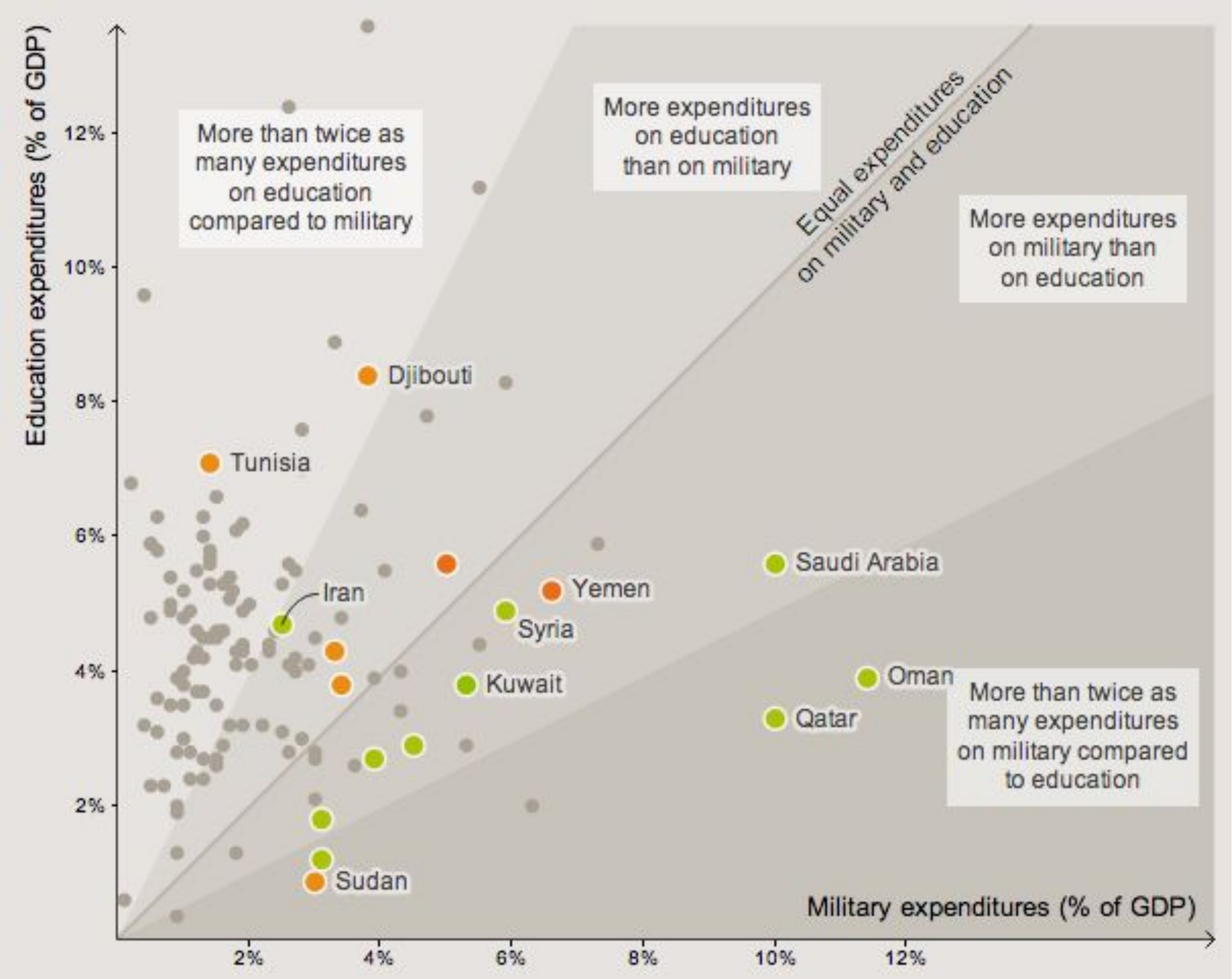
Literacy by Region

Literacy Rates by Country and Gender

Enrollment Ratios since the Millennium Development Goals

Expentitures on Education and Military

Example (cont.)



Find Editorial Focus by Reasoning

- Unless you've already had the editorial focus specifically outlined to you, an effective approach to refine it can be drawn from the practice of logical reasoning, such as:
  - **Deductive** reasoning
  - **Inductive** reasoning

Deductive Reasoning

- Confirming or finding evidence to support specific ideas:
  1. A certain predetermined sense of what stories might be interesting, relevant, and potentially available within your data.
  2. You are pursuing a curiosity by interrogating your dataset in order to substantiate your ideas of what may be the key story dimensions.

Inductive Reasoning

- It works the opposite way
- Open-ended and exploratory.
  - Use analytical and visualization techniques to try and unearth potentially interesting discoveries, forming different and evolving combinations of data questions.
  - may end up with nothing, we may find plenty
- Fundamentally, this is about using visual analysis to find stories.

Using Visual Analysis to Find Stories

"Visualization gives you answers to questions you didn't know you had.“

Ben Schneiderman

- This activity can also be described as data sketching or preproduction visualization.
- We are using visualization techniques to:
  - become more intimate with our raw material
  - to start to form an understanding of what we might portray to others
  - And how we might accomplish that.

Characteristics to be Observed

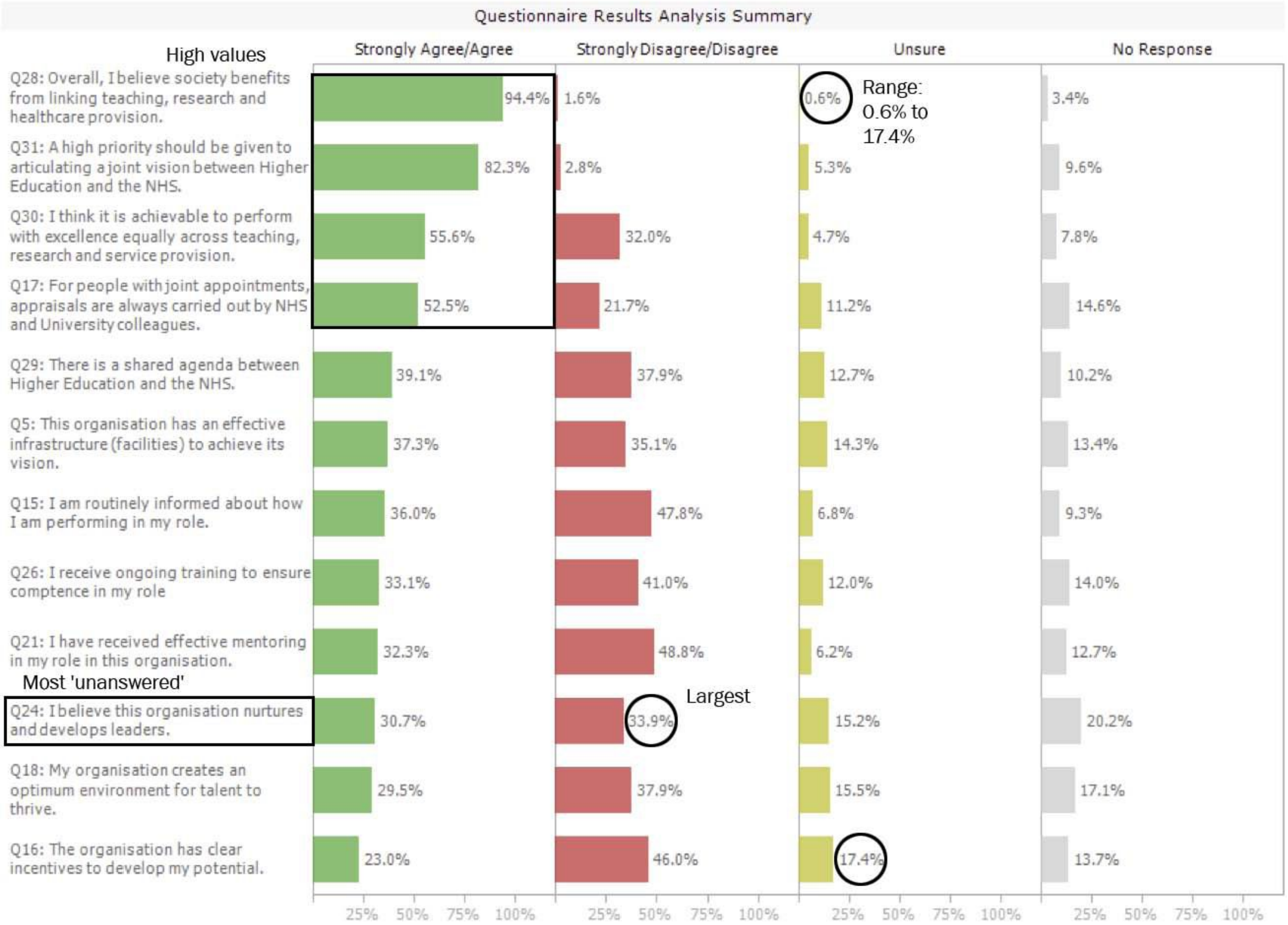
- Comparisons and proportions
- Trends and Patterns
- Relationships and Connections

Comparisons and Proportions

- **Range and distribution**: Discovering the range of values and the shape of their distribution within each variable and across combinations of variables
- **Ranking**: Learning about the order of data in terms of general magnitude, identifying the big, medium, and small values.
- **Context**: Judging values against the context of averages, standard deviations, targets, and forecasts.

Example

Using methods like a bar chart will enable comparison across values and categories to pick out the type of physical qualities just listed

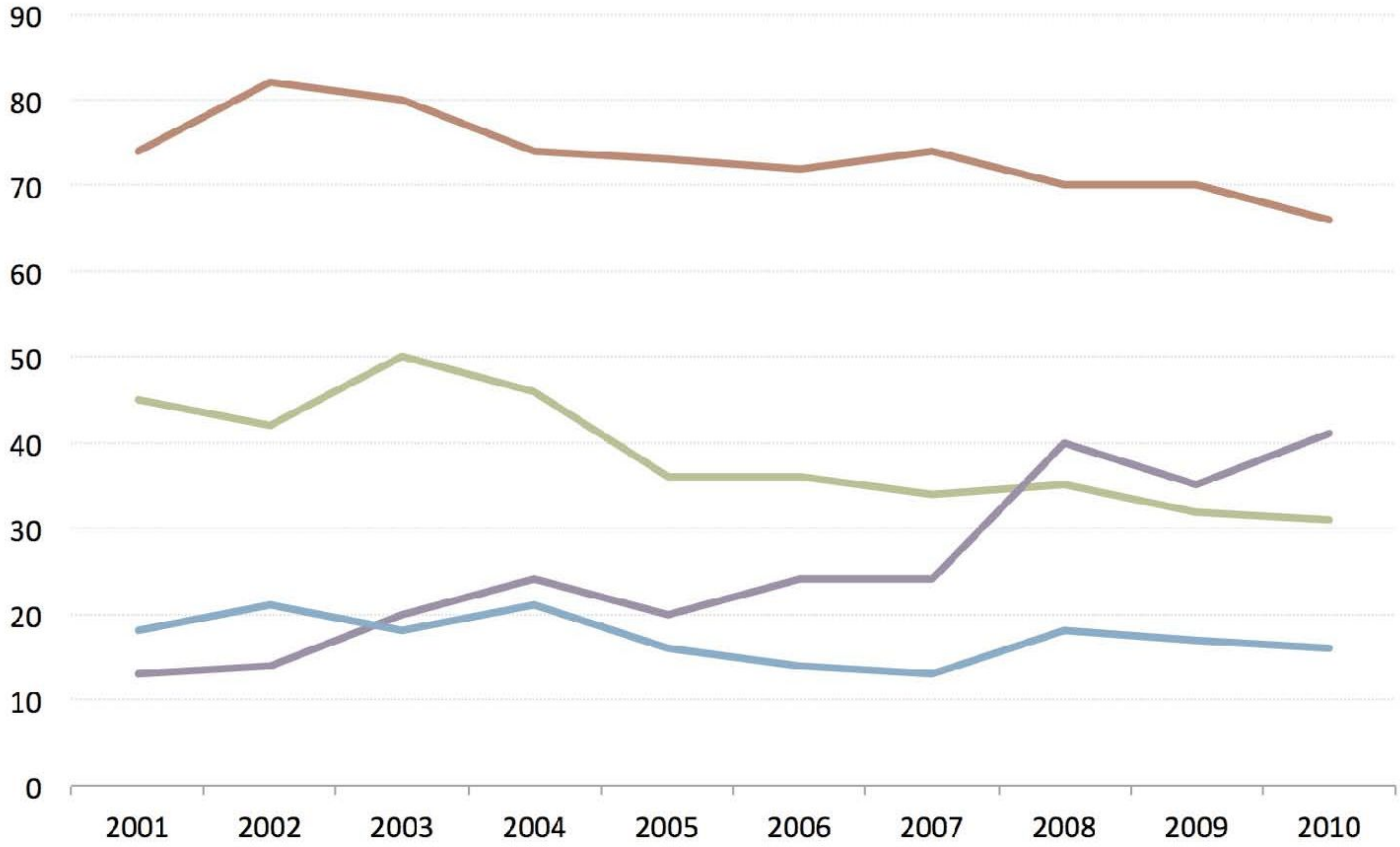


Trends and Patterns

- **Direction**: Are values changing in an upward, downward, or flat motion?
- **Rate of change**: How steep or flat do pattern changes occur? Do we see a consistent, linear pattern, or is it much more exponential in shape?
- **Fluctuation**: Do we see evidence of consistent patterns or is there significant fluctuation? Maybe there is a certain rhythm, such as seasonality, or perhaps patterns are more random
- **Significance**: Can we determine if the patterns we see are meaningful signals or simply represent the noise within the data?
- **Intersections**: Do we observe any important intersections or overlaps between variables, crossover points that indicate a significant change in relationship?

Example

Using a line chart is a perfectly suitable method to observe patterns and trends

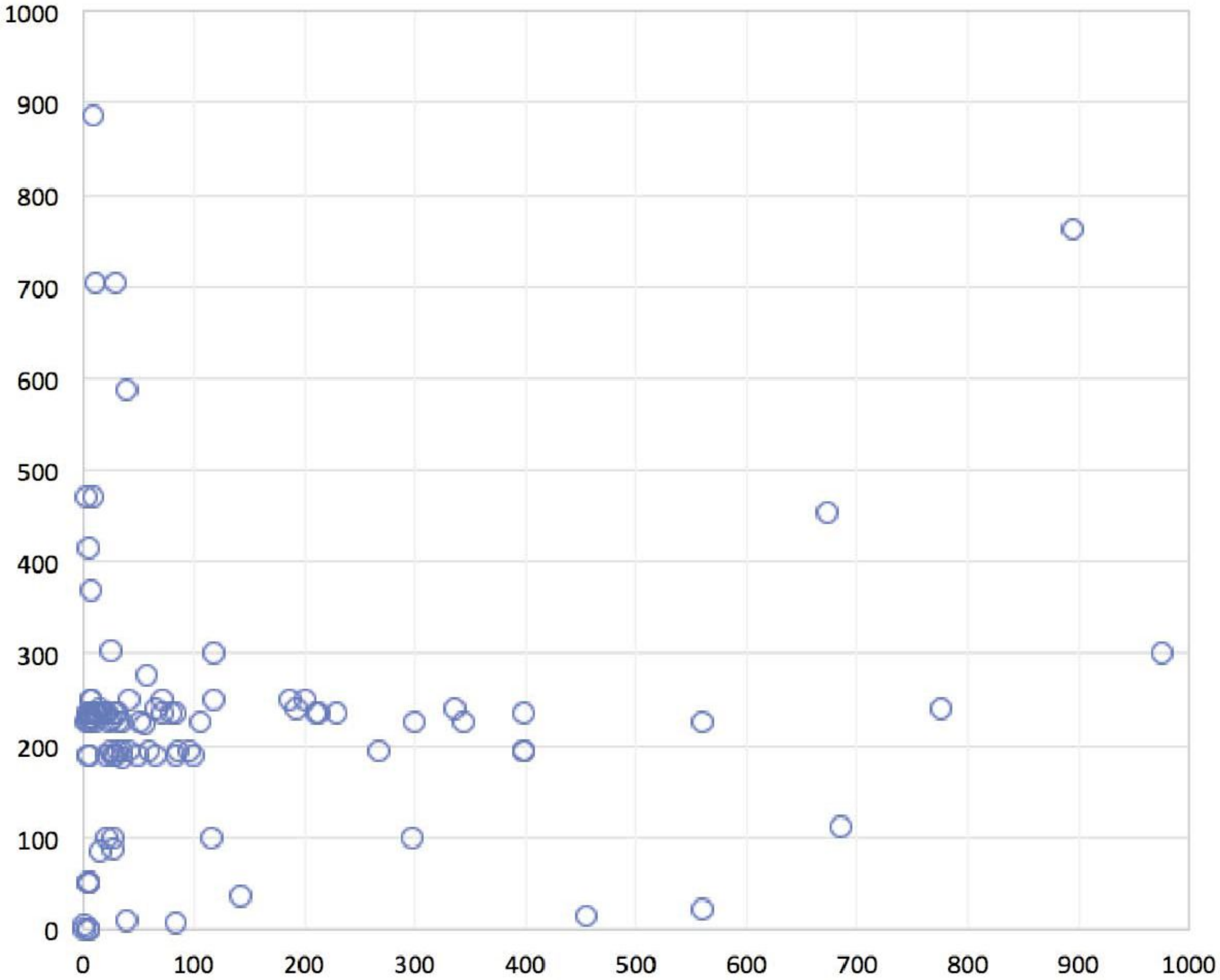


Relationships and Connections

- **Exceptions**: Can we identify any significant values that sit outside of the norm, such as outliers?
- **Correlations**: Is there evidence of strong or weak correlations between variable combinations?
- **Associations**: Can we identify any important connections between different combinations of variables or values?
- **Clusters and gaps**: are there gaps in values and data points?
- **Hierarchical relationships**: Determining the composition, distribution, and relevance of the data's categories and subcategories.

Example

Using a scatter plot will enable visibility of these types of relationships



Next Class

- Quiz