High-Dimensional Data Visualization
Some examples

- Text data.
- Finance.
  - Time Series Data.
- Educational data
  - Spatial Data.
  - Spatio-temporal Data.
- Biological Data
- Many others...

US Industry Ratios.
Each industry is characterized by eight numerical values (market capitalization, price/earning ratio, etc.) Source: visumap.net
Curse of Dimensionality

Estimating the Gaussian density function around zero:

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>Required Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>786</td>
</tr>
<tr>
<td>7</td>
<td>10,700</td>
</tr>
<tr>
<td>10</td>
<td>842,000</td>
</tr>
</tbody>
</table>

Phenomena in high dimensions

• High dimensional samples → More noise!
• Samples in high dimensional space tend to be **equidistant**!
  
  Concentration of measure phenomenon

• The mathematical and statistical properties of high-dimensional data spaces are often poorly understood or inadequately considered.

• Data are rarely randomly distributed in high-dimensions and are highly correlated, often with spurious correlations.

• Literature:
Representative Methods

- Parallel coordinates
- Scatter plots
- Mosaicplot
- Grand tour
- Trellis displays
- Linked views
- Multivariate visualization
- Others ...
Parallel Coordinates

• Orthogonal coordinates fail after three dimensions

• Basic idea:
  – Give up orthogonality
    • Draw the coordinate system as a series of parallel axes in a plane
  – Plot a point in parallel coordinates as a polyline
    • This is a unique representation of a point.
    • The number of axes is effectively unlimited.
Parallel Coordinates

Parallel coordinate plot for 10 variables on almost 400 cars

The polyline represents the point $C = (c_1, c_2, c_3, c_4, c_5)$
Parallel Coordinates

- Points in Cartesian space map into lines in parallel coordinate space and vice versa
  - There is a fundamental projective geometry duality in this representation
Parallel Coordinates

• Points in Cartesian space map into lines in parallel coordinate space and vice versa
  – There is a fundamental projective geometry duality in this representation
Parallel Coordinates

• Points in Cartesian space map into lines in parallel coordinate space and vice versa
  – There is a fundamental projective geometry **duality** in this representation
Parallel Coordinates

• Points in Cartesian space map into lines in parallel coordinate space and vice versa
  – There is a fundamental projective geometry duality in this representation

• More dualities
  – Circles and ellipses in Cartesian space map into hyperbolas in parallel coordinate space
  – Rotations in Cartesian space map into translations in parallel coordinate space and vice versa
  – Points of inflection in Cartesian space map into cusps in parallel coordinate space.

• These dualities allow easy interpretation in parallel coordinate space
Utility of Parallel Coordinates

• Overview
  – No other statistical graphic can plot so much information (cases and variables) at a time. Thus parallel coordinate plots are an ideal tool to get a first overview of a data set.

All axes have been scaled to min-max. Several features, like a few very expensive cars, three very fuel-efficient cars, and the negative correlation between car size and gas mileage, are immediately apparent.
Utility of Parallel Coordinates

• Profiles
  – Parallel coordinate plots can be used to visualize the profile of a single case via highlighting.
  – Profiles are not only restricted to single cases but can be plotted for a whole group, to compare the profile of that group with the rest of the data.

The most fuel-efficient car (Honda Insight 2dr.) highlighted
Utility of Parallel Coordinates

- Monitor
  - When working on subsets of a data set parallel coordinate plots can help to relate features of a specific subset to the rest of the data set.
  - When looking at the result of a multidimensional scaling procedure, parallel coordinate plots can help to find the major axes, which influence the configuration of the multidimensional scaling (MDS).

MDS is a class of techniques where a set of given distances is approximated by distances in low-dimensional Euclidean space

The four cars in the lower right of the MDS are highlighted and happen to be the most expensive and most powerful cars in the data set.
Parallel Coordinate Limitation

• A common question about parallel coordinates involves the adjacency issue. Axes that are adjacent allow for easier comparison than axes that are not adjacent.
• How many parallel coordinate displays do you need so that all pairwise adjacencies are present?
Parallel Coordinate Limitation

- If the parallel coordinate axes are ordered from 1 through d, then there is an easy pairwise comparison of 1 with 2, 2 with 3 and so on.

- However, the pairwise comparison of 1 with 3, 2 with 5 and so on was not easily done because these axes were not adjacent.

- One simple mathematical question then is what is the minimal number of permutations of the axes in order to guarantee all possible pairwise adjacencies.

- Although there are d! (factorial of d) permutations, many of these are duplicate adjacencies. Actually far fewer permutations are required.
Parallel Coordinate Theory

- Graph representation for the axes
  - A graph is drawn with vertices representing coordinate axes, labeled clockwise 1 to d. Edges represent adjacencies, so that vertex one connected to vertex two by an edge means axis one is placed adjacent to axis two.

- Based on the graph,
  - To construct a minimal set of permutations that completes the graph is equivalent to finding a minimal set of orderings of the axes so that every possible adjacency is present.
Parallel Coordinate Limitation

1 2 6 3 5 4
2 3 1 4 6 5
3 4 2 5 1 6
Hamiltonian Paths

• It is about how to represent all the spanning trees of the complete graph that the union of the edges in the obtained spanning trees is the original set of the edges and the intersection of the edges in the obtained spanning trees is empty or close to empty.

• A technique called Hamiltonian paths and Hamiltonian cycles of the graph can be used.
Hamiltonian paths and Hamiltonian cycle

A Hamiltonian path or traceable path is a path that visits each vertex exactly once. A graph that contains a Hamiltonian path is called a traceable graph. A graph is Hamiltonian-connected if for every pair of vertices there is a Hamiltonian path between the two vertices.

A Hamiltonian cycle, Hamiltonian circuit, vertex tour or graph cycle is a cycle that visits each vertex exactly once (except for the vertex that is both the start and end, which is visited twice). A graph that contains a Hamiltonian cycle is called a Hamiltonian graph.

Examples
- a complete graph with more than two vertices is Hamiltonian.
- every cycle graph is Hamiltonian.
- every tournament has an odd number of Hamiltonian paths (Rédei 1934).
- every platonic solid, considered as a graph, is Hamiltonian.
- Every maximal planar graph with no separating triangles is Hamiltonian.
A Simple Way to Find Hamiltonian Paths

- A basic zig-zag pattern can be used in the construction. This creates an ordering which in the example is 1 2 7 3 6 4 5.

- For \( d \text{ even} \) this general sequence can be written as
  \[ 1, 2, d, 3, d - 1, 4, d - 2, \ldots \frac{d + 2}{2} \]
  and for \( d \text{ odd} \) as
  \[ 1, 2, d, 3, d - 1, 4, d - 2, \ldots \frac{d + 3}{2} \]

- An even simpler formulation is:
  \[ d_{k+1} = (d_k + (-1)^{k+1} k) \mod d, \quad k = 1, 2, \ldots, d - 1 \]
  With \( d_1 = 1 \). Here \( 0 \mod d = d \mod d = d \).
Parallel Coordinates Limitation II

Hard to scale to large scale data sets!

400 data points

3848 data points
Parallel Coordinates Limitation II

Certain clustering technique is needed + enhanced visual representation for the visualization of the obtained clusters!

Original

Density-based

3848 data points
Parallel Coordinates Limitation II

Other improvements

arc-edge  [Zhou et al. EuroVis08]

splatting  [Zhou et al. EuroVis09]

Continuous representation  [Lehmann and Theisel TVCG11]
Parallel Coordinates Limitation II

Other improvements

Force-direct  [Walker et al. 2013]

PCP + Angular histogram  [Geng et al. Vis2011]

Parallel Coordinates Art  [Heinrich and Weiskopf Vis2013]
Additional Readings

• Alfred Inselberg. "Parallel Coordinates: Interactive visualization for high dimensions", Trends in Interactive Visualization, Advanced information and knowledge processing 2009, pp 49-78.


• Hong Zhou, Weiwei Cui, Huamin Qu, Yingcai Wu, Xiaoru Yuan, Wei Zhou, “Splatting the lines in parallel coordinates”, EuroVis09.

• Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, Baoquan Chen. “Visual clustering in parallel coordinates”. EuroVis08.

• Julian Heinrich and Daniel Weiskopf. "Parallel-Coordinates Art" . IEEE VIS 2013 Arts Program.
Software for PCP

While there is a large amount of papers about parallel coordinates, there are only few notable software publicly available to convert databases into parallel coordinates graphics. Notable software are ELKI, GGobi, Macrofocus High-D, Mondrian, and ROOT. Libraries include Protovis.js, D3.js provide basic examples, while more complex examples are also available. D3.Parcoords.js (a D3-based library) and Macrofocus High-D API (a Java library) specifically dedicated to | -coords graphic creation have also been published. The Python data structure and analysis library Pandas implements parallel coordinates plotting, using the plotting library matplotlib. The R package GGally, among others, implements parallel coordinates plotting.

https://en.wikipedia.org/wiki/Parallel_coordinates

http://www.xdat.org/

http://davis.wpi.edu/xmdv/

http://www.oicweave.org/

http://www.kdnuggets.com/software/visualization.html
Dimensionality reduction

Assume that the data of interest lie on an embedded non-linear manifold within the higher-dimensional space. If the manifold is of low enough dimension, the data can be visualized in the low-dimensional space.

Linear Decomposition Methods

- Principle component analysis (PCA)
- Singular value decomposition (SVD)

Non-linear dimensionality reduction

- Multidimensional scaling
- manifold learning
- many more!!

Visual Analytics
Numerous data
Data deluge!
... peta ($2^{40}$), exa ($2^{50}$), ...
Time-dependent
Often unstructured
Usually with uncertainty
Interrelated

Useful knowledge!

transformational science and engineering
Numerous data
Data deluge!
... peta ($2^{40}$), exa ($2^{50}$), ...

Time-dependent
Often unstructured
Usually with uncertainty
Interrelated

Useful knowledge!
transformational science
and engineering

What do we want to know?
- openconnectomeproject.org
- 450x350x50 cubic micron volume
- 4x4x45 cubic nanometers resolution
- about over 1000 images, each with 112,500 x 87,500 pixels.
Numerous data
data deluge!
... peta ($2^{40}$), exa ($2^{50}$), ...

Time-dependent
Often unstructured
Usually with uncertainty
Interrelated

Useful knowledge!

transformational science
and engineering

**Visual analytics**—the science of combining interactive visual interfaces with automatic algorithms to support analytical reasoning and build synergies between humans and computers.
Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces. Today, data is produced at an incredible rate and the ability to collect and store the data is increasing at a faster rate than the ability to analyze it. Over the last decades, a large number of automatic data analysis methods have been developed. However, the complex nature of many problems makes it indispensable to include human intelligence at an early stage in the data analysis process. Visual Analytics methods allow decision makers to combine their human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today’s computers to gain insight into complex problems. Using advanced visual interfaces, humans may directly interact with the data analysis capabilities of today’s computer, allowing them to make well-informed decisions in complex situations.

vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf
Visual Analytics

• Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces

• People use visual analytics tools and techniques to
  – Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
  – Detect the expected and discover the unexpected
  – Provide timely, defensible, and understandable assessments
  – Communicate assessment effectively for action.

• Not really an “area” per se
  – More of an “umbrella” notion
  – Combines multiple areas or disciplines
  – Ultimately about using data to improve our knowledge and help make decisions

• Main Components:
• **Combine strengths** of both human and electronic data processing
  – Gives a *semi-automated analytical process*
  – Use strengths from each
  – Below from Keim, 2008

Keim et al, chapter in Information Visualization: Human-Centered Issues and Perspectives, 2008

---

**Diagram:**

- **Machine**
  - Statistical Analysis
  - Data Mining
  - Data Management
  - Compression & Filtering
  - Semantics-based approaches

- **Human**
  - Human-centered computing
  - Information Design
  - Visual Intelligence
  - Decision Making Theory

---

"The best of both sides"
Domain Roots

- Dept. of Homeland Security supported founding VA research
- Area has thus been connected with security, intelligence, law enforcement
- Should be domain-independent, however, as other areas need VA too
  - Business, science, biology, legal, etc.

Source: Google images
Acknowledgment

• Thank the following people for the materials
  – Prof. Min Chen
  – Prof. Niklas Elmqvist
Structure of your final project report
(4-8 pages using IEEE VGTC style)
https://www.computer.org/web/tvcg/author

Your report should contain the following content:
• Abstract
• Introduction
• Related work (adapt your literature review for this)
• Implementation
• Results
• Future directions
• References