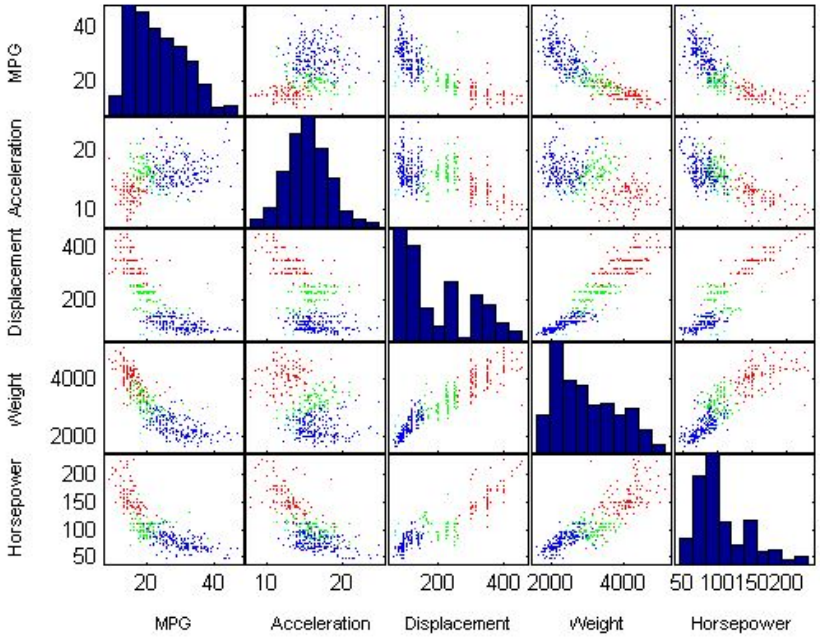
Convex Hulls, Voronoi Diagrams, & k-Means Clustering

Today

- Reading: "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Let's do Computational Geometry
  - Convex Hull
  - Voronoi Diagram
  - k-Means Clustering
- HW4: Data Collection
- Today's Worksheet: Brainstorming Data

Scatterplot Matrix

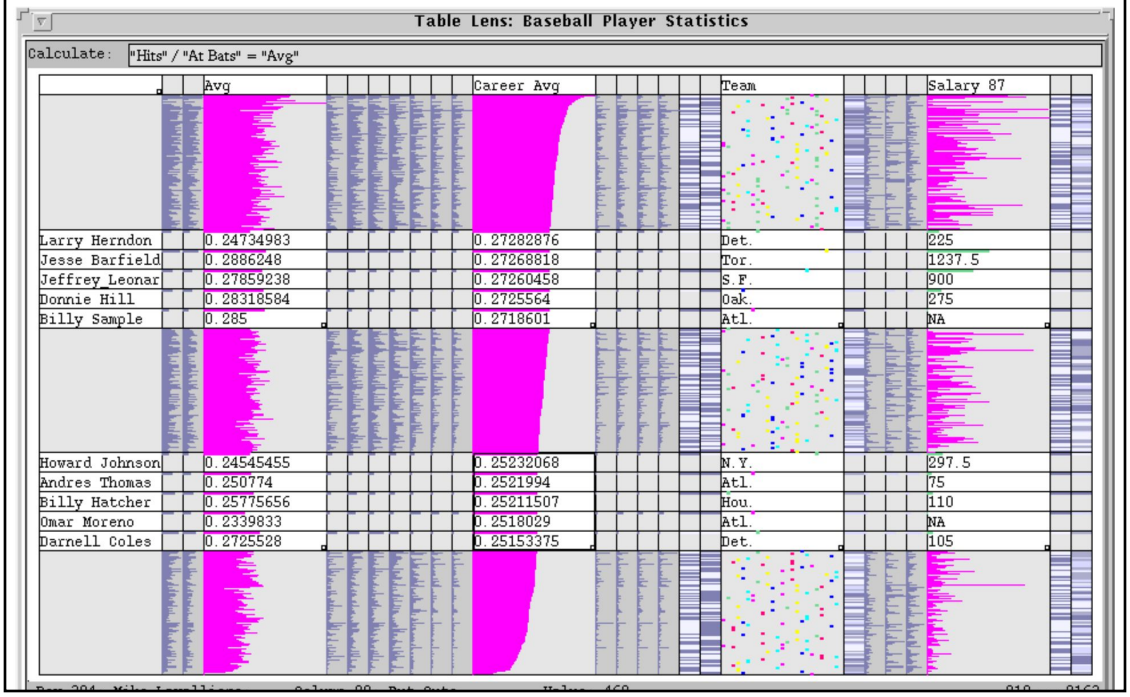


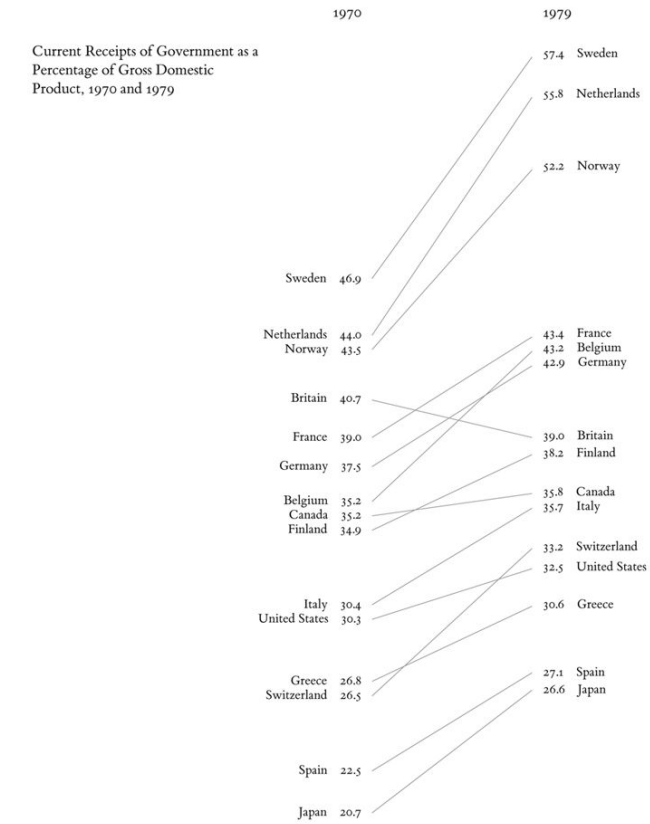
Color coded by # of cylinders

http://www.mathworks.com/help/stats/examples/visualizing-multivariate-data.html

"The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information" Ramana Rao, SIGCHI Conference on Human Factors in Computing Systems 1994

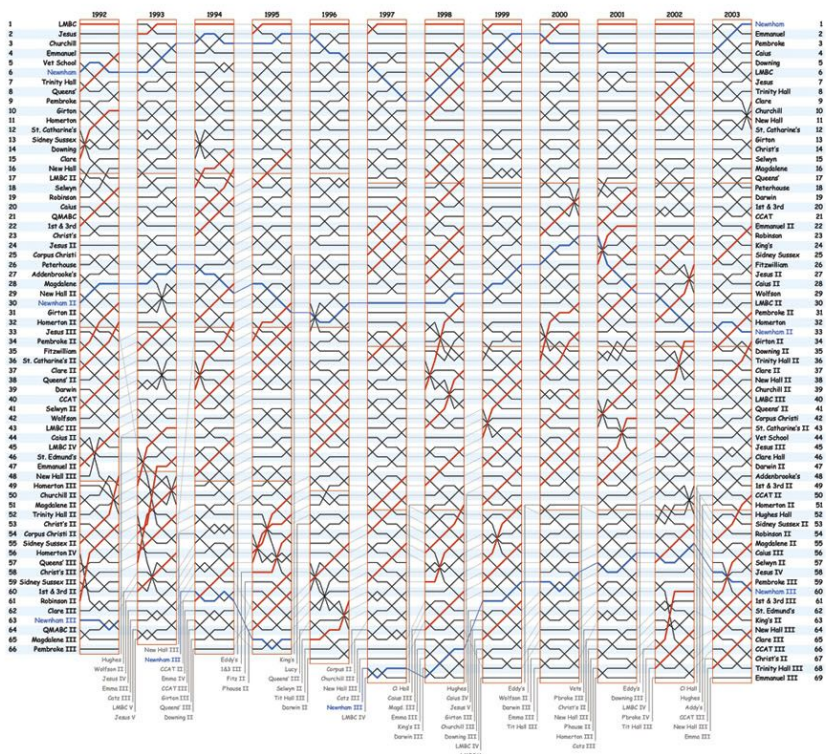
Focus+context





"Slope graph"

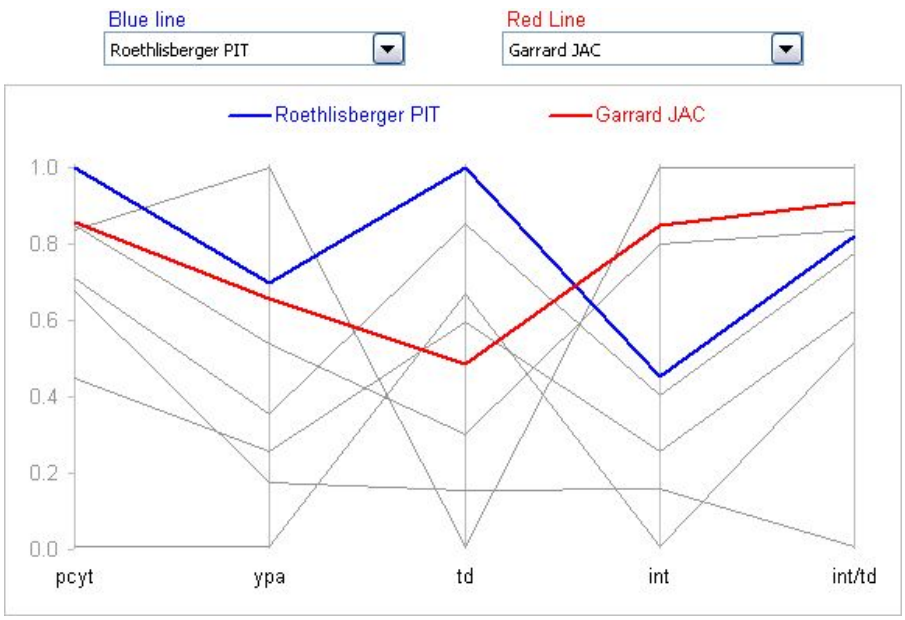
from:
Edward Tufte
The Visual Display of Quantitative Information 1983



"Bump chart"

Time Granger, University of Cambridge

Parallel Coordinates



https://peltiertech.com/Excel/Charts/ParallelCoord.html

RankExplorer: Visualization of Ranking Changes in Large Time Series Data

Conglei Shi, Weiwei Cui, Shixia Liu, *Member, IEEE*, Panpan Xu, Wei Chen, and Huamin Qu, *Member, IEEE*

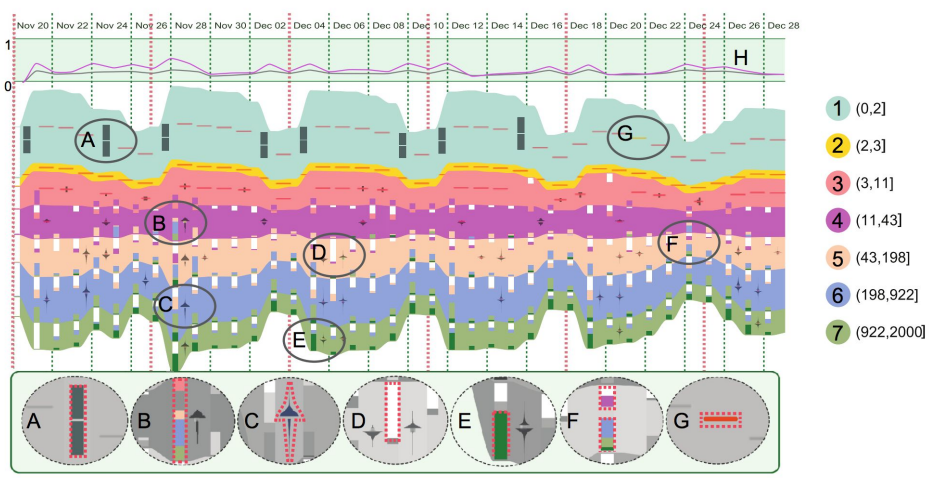


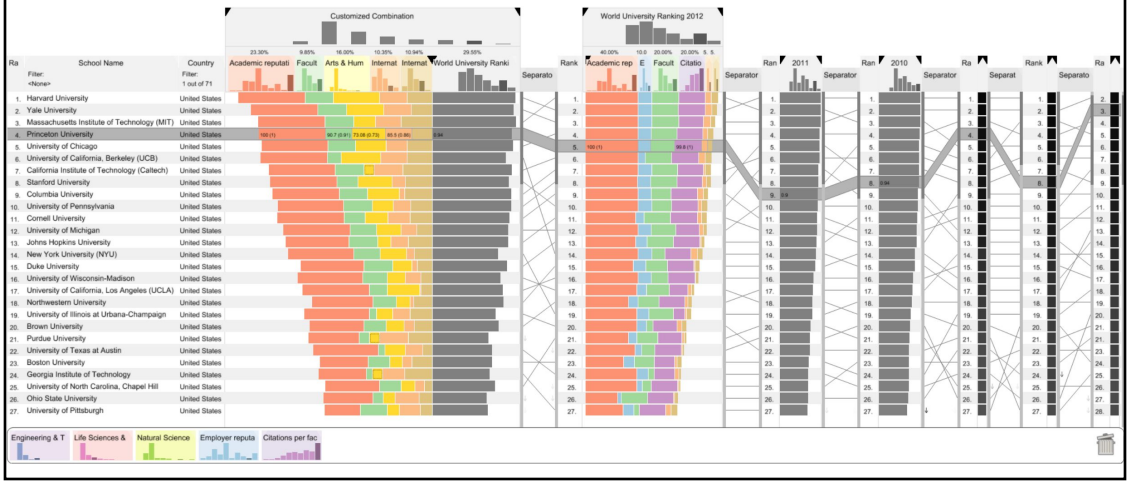
Fig. 1. RankExplorer visualization of the top 2000 Bing search queries from Nov. 20 to Dec. 29 in 2011. All queries are divided into seven categories. The width of each layer at a time point encodes the total query count at that time. The color bar and glyphs encode the content changes in each ranking category. From the color bar, we can observe: 1) the change between layers (the bar segments with the colors of other layers in B and F); 2) new queries coming in (the white segment in D); 3) recurring queries (the dark green segment in E). From the changing glyphs, we can see: 1) a non-change pattern (only red line in G); 2) a swap pattern (the two equal-height segments in A represent that the two queries swap their rankings); 3) a shift pattern (the increasing part is significantly larger than the decreasing part in C). From the trend curve (H), we can see the degree of ranking change over time.

Tableau

http://www.tableausoftware.com/

Reading for Today

- "LineUp: Visual Analysis of Multi-Attribute Rankings", Gratzl, Lex, Gehlenborg, Pfister and Streit, IEEE INFOVIS 2013



- Visualizing overall ranking is simple, interpretation of ranking is not simple
- Alternative rankings exists
- Customize weights of multiple heterogeneous attributes
- Customize function for a particular attribute: linear, log
  is this easy for the user to do? Give a larger set of typical options rather than type javascript?
- Compare different publications rankings, compare rankings over time
- Cost-benefit analysis for improving rankings
- Rankings should be personalized, we each have our own priorities
- Take "rankings" to a new level
- Previously thought bar charts were boring, not creative

Requirements analysis

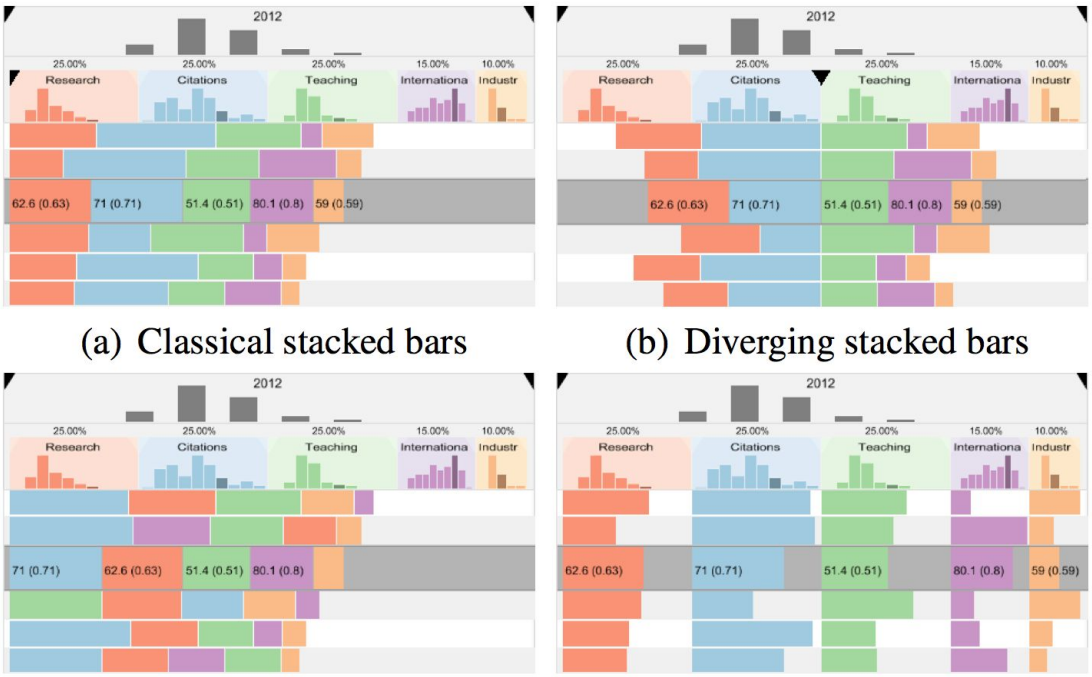
- List of 10 detailed requirements, including:
  - Nested model for visualization design and validation
  - Domain problem characterization
  - Data must be normalized, allow filters of data, handle missing data (unfortunately discussion/solution rather vague in paper)
  - But keeping track of 10 criteria was a little hard to follow
  - Some of the 10 requirements seemed redundant
- Did good job developing & describing the problem & requirements
- Formalizes general (intuitive, agreed upon) criteria
- Interaction is key to effective user interpretation of data

Features

- Column types
  - Rank columns
  - Textual attribute column – labels, nominal attributes
  - Categorical attribute column
  - Numerical attribute column – encoded as bars!
  - Combined attribute column
- Histogram above the attributes – confusing, not sufficiently explained
- Interactivity
  - Interactive bar graph plotter with customizable filters & parameters
  - Immediate feedback when weights are changed
  - Animated transitions, color highlight – strength/length of time applied, based on amount of change
  - "Snapshot" for comparing rankings before/after weight edits
  - Collapsing/compressing columns
    - Use fewer pixels to display same data (can be done with bar graphs, not words/raw numbers)
    - "Fisheye" view allowing a gradual falloff in level-of-detail/screen space
    - Clean up clutter
    - Replace connection lines to invisible targets with arrows

Features II

- Memo pad: place to put attributes not currently in use.
- Consistent, customize-able mapping between attribute and color
  - too many colors might become too busy/confusing
- Parallel mapping editor
- Orthogonal mapping editor
- Allows user to infer the importance of attributes to ranking
- Is this too many features in one industrial strength application? Would it be better to have a collection of individual tools/plugins?
- Too many controls, not enough guidance for novices on how to use them
  - Tool (and/or presentation of tool) is needlessly complex?



(a) Classical stacked bars (b) Diverging stacked bars

(c) Ordered stacked bars (d) All-aligned bars

Evaluation

- Heavy focus on User Study in paper
- Compare to existing tools: Confirmed that experts in Excel could complete tasks that novices would be asked to perform with LineUp – However, did not do formal user study comparison between Excel & Lineup because expertise is such a huge component
- Check for colorblindness – why?
- 12 separate tasks
- NASA TLX workload assessment questionnaire, task completion times
- Justification for their experiments was early in paper & detailed, synced well with study & results, study measured the appropriate things
- Well designed study, but only a tiny # of participants (8)
- Users in study were unrealistic? Target audience is “expert amateurs”?

Results

- Authors surprised that filter task took longer than most other tasks
- Users thought it was visually pleasing, potentially helpful, easy to understand, and (compared to Excel) would save time, help gather more insights
- Drag & drop is easier to create complex rankings than typing formulas
- Some negative comments about rankings change from user who did not [know to] use the snapshot feature
- Does this scale to bigger datasets? Is it visually clear? Is it reasonable to navigate?

Future work

- Features to help optimize rankings, what needs to be changed to achieve a certain ranking?
  - Why is this an important feature? (ask any university administrator…)
- Statistical techniques to deal with large number of attributes
- Large differences in rankings become steep slopes which are hard to interpret
- Need to test actual users in real-world analyses and observe adoption rates
  - Web-based implementation
  - Apply tool for original purpose: ranking of genes, clusters, pathways for analysis of genomic data

Writing Quality

- Paper was well written
- Justification for their experiments was early in paper & detailed, synced well with study & results, study measured the appropriate things
- Who is the audience?
  - Clients designing rankings
  - Or people who need to look at rankings
  - Domain experts?
  - Novices?
- Lots of figures! Great. But some details in some figures were confusing/hard to follow
- Would have been nice to include algorithm in paper
- Multiple real-world examples good -- Nutrition example helped explain the data mapping editor
- Scrolling up to the top of the paper every time they mention Figure 1 or a specific # requirement from the list of 10 was annoying
- Good use of bold text
- Some parts were repetitive

Definitions / Misc.

- Likert scale: on a survey, response from a set of choices on a scale, commonly: strongly disagree, disagree, neither agree nor disagree, agree, strong agree
- Algorithm design, encoding/interaction design
  - "A nested model for visualization design and validation", Tamara Munzer, TVCG 2009
  - http://www.cs.ubc.ca/labs/imager/tr/2009/NestedModel/NestedModel.pdf

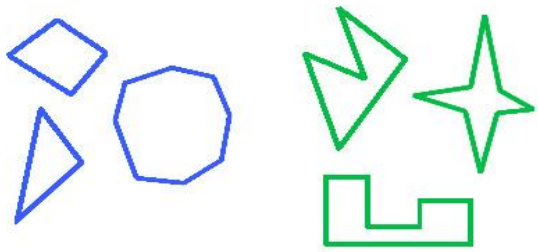
Today

- Reading: "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Let's do Computational Geometry
  - Convex Hull
  - Voronoi Diagram
  - k-Means Clustering
- HW4: Data Collection
- Today's Worksheet: Brainstorming Data

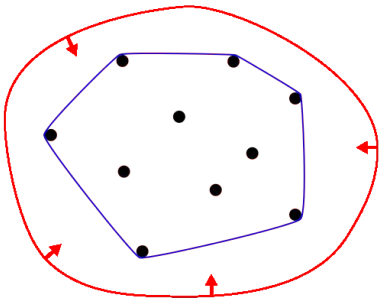
Convex vs. Concave, Convex Hull

Convex: Shape has no inward corners or curving faces.

Concave: Has inward corner(s) or inward curving face(s).



http://img.sparknotes.com/figures/B/b333d91dce2882b2db48b8ad670cd15a/convexconcave.gif

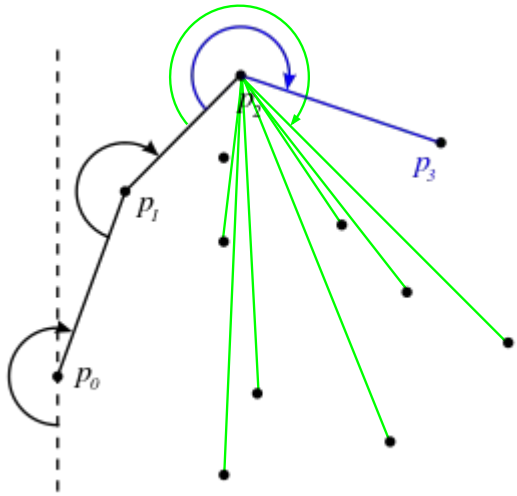


http://en.wikipedia.org/wiki/File:ConvexHull.svg

Convex Hull: The smallest convex shape that contains all of the input points / elements. 2D, 3D, higher dimension.

"Gift Wrapping"

- Find left most point
- "Walk around" the point set in the clockwise direction

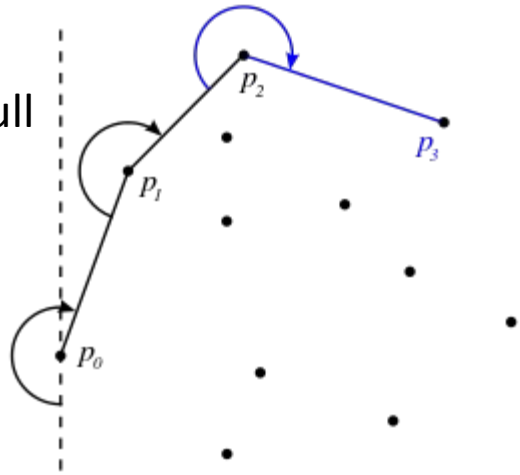


- Find the next point p_3 on the hull.
  - smallest outer angle between lines p_1p_2 & p_2p_3

Gift Wrapping Analysis?

- For n input points, and output convex hull contains h points?

  O(nh)

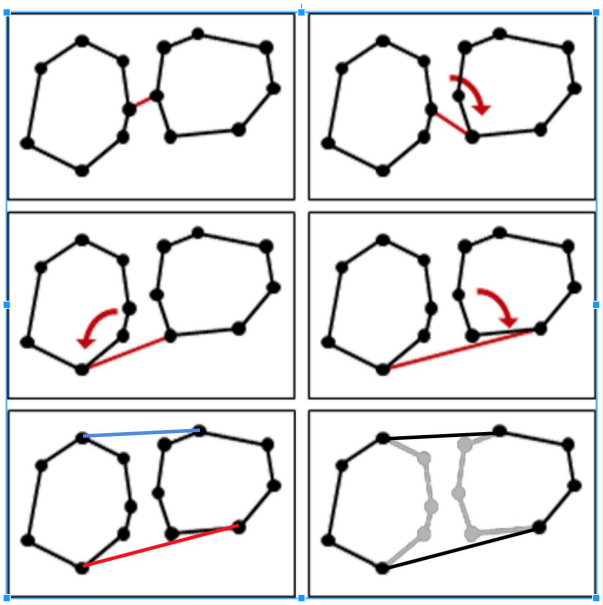


- Worst case?

  $O(n^2)$ *most/all input points are on the convex hull*

Divide & Conquer *(like merge sort)*

- Sort the x axis
- Split into 2 ~equal groups
- Recurse...
- Merge the 2 convex hulls
  - Find rightmost point in left hull, and leftmost point in right hull.
  - Walk down to find lower tangent
  - & up for upper tangent
  - Discard points in between upper & lower tangents
- Analysis?



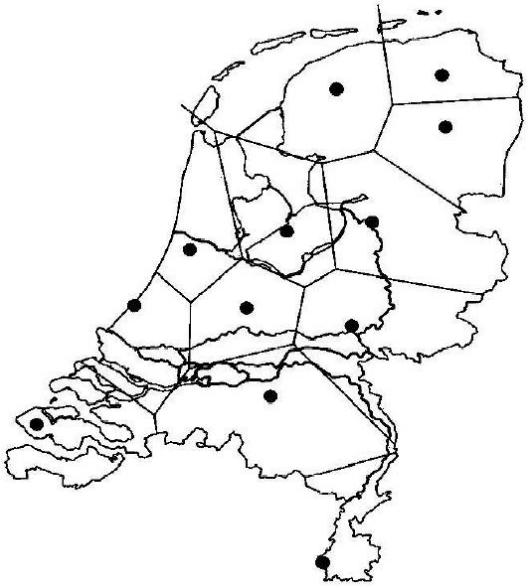
O(n log n)

Today

- Reading: "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Let's do Computational Geometry
  - Convex Hull
  - Voronoi Diagram
  - k-Means Clustering
- HW4: Data Collection
- Today's Worksheet: Brainstorming Data

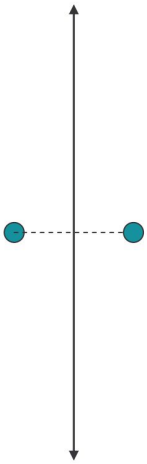
Voronoi Diagram

- Intuition: Partition space into regions that all agree on which is the closest point from a set of input points.
- Examples
  - Re-district the Netherlands into provinces so that everyone reports to the closest capital
  - Multiple ambulance helicopters distributed across the county, which should we dispatch to help?

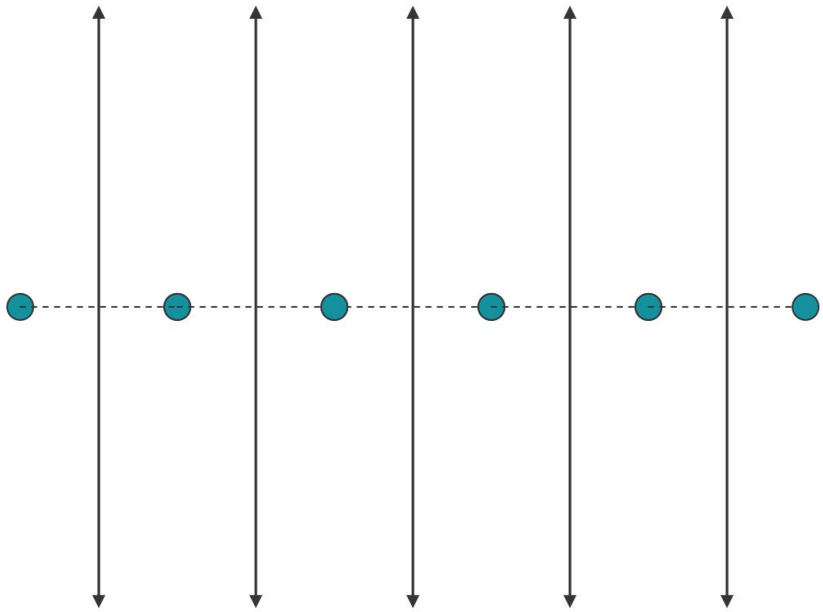


http://ccc.inaoep.mx/~rodrigo/robotica/Trigui.pdf

- If we have just 2 input points, we divide space into 2 regions.
- Region/Cell edges are the perpendicular bisectors of nearby points.



Multiple input points… but all on a line

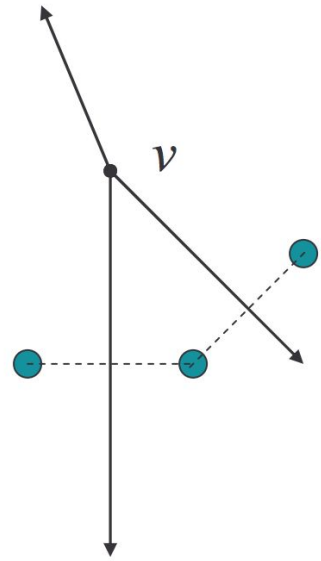


Assume "General Position"

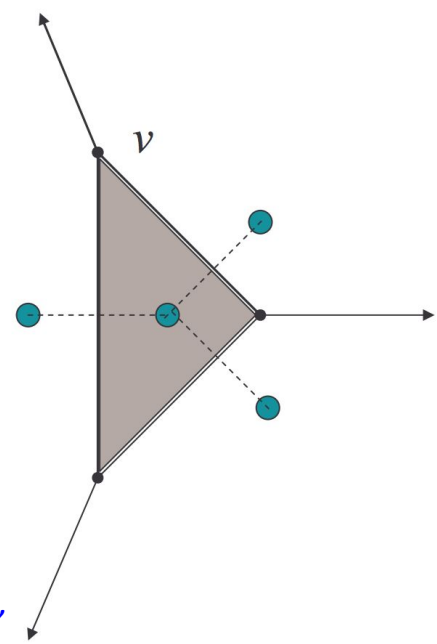
- What does this mean?
- We'll focus on the general case, where we don't have *coincidental alignments…*
  - No 2 points are exactly at the same position
  - No 3 points are exactly co-linear
  - No 4 points lie exactly on a circle
  - Etc.
- We don't have to worry about anything ever being a tie, there will only be one right answer.
- How to ensure general position in practice?

Add a small amount of random noise to every point.
Small enough to not change the result.
But large enough so things don't align.

- Where the perpendicular bisectors intersect, we create corners for our Voronoi cells



- A Voronoi cell for an input point is bounded by the perpendicular bisectors with its closest neighboring input points
- Each region will be convex!

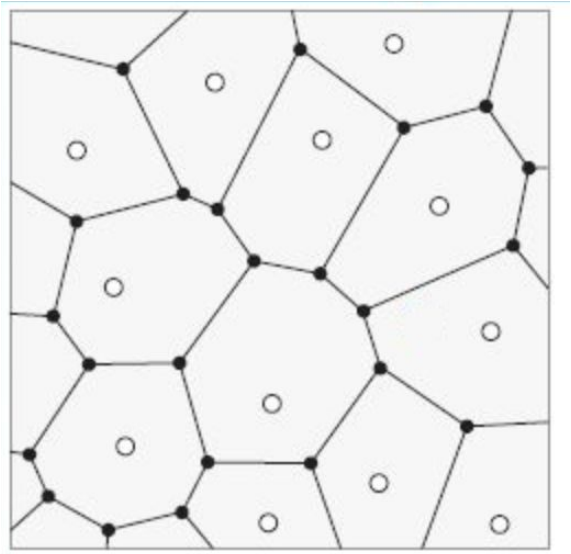


- *A Voronoi cell can have any number of edges.*
- *If the input points are evenly spaced, most cells will have approx 6 sides, with approx equal length sides and approx equal 120° angles.*

Nearest Neighbor Query

- Once we have built a Voronoi Diagram
- We can find the nearest neighbor in the input point set
- a.k.a. locate which Voronoi Cell contains the query point
- *In O(log n) time*

  … *algorithm* details omitted



Incremental Algorithm for Voronoi construction

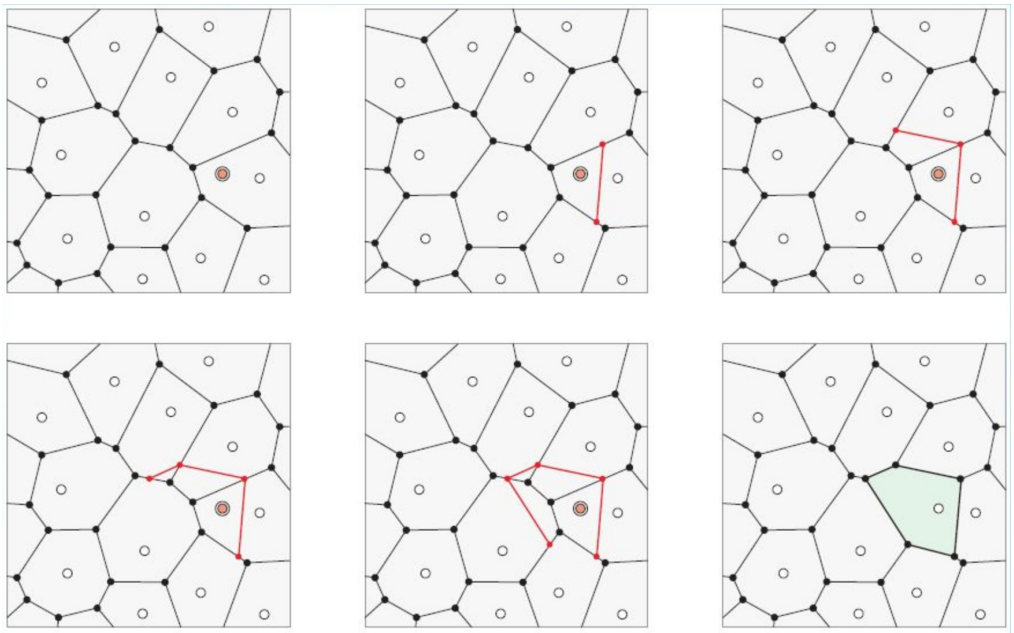
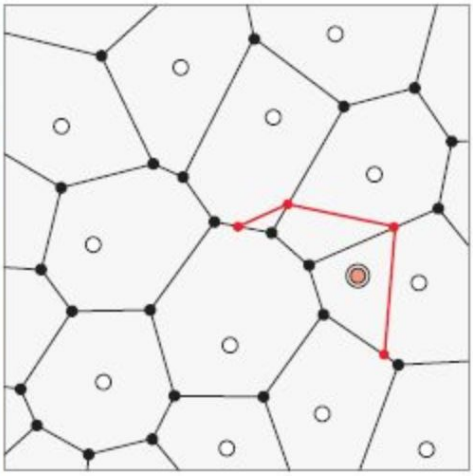


Figure 4.5: The incremental algorithm for creating Voronoi diagrams.

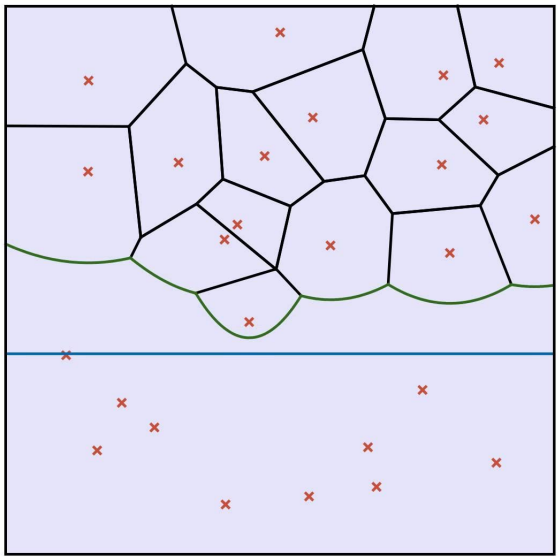
Analysis of Incremental Algorithm?

- Find the cell containing new point
- Calculate perpendicular bisector with the nearest neighbor and points in the neighboring cells



- Average case to insert 1 point?
  O(log n)
- Worst Case to insert 1 point?
  O(n)
- Average case to build complete diagram?
  O(n log n)

Sweep line algorithm for Voronoi construction

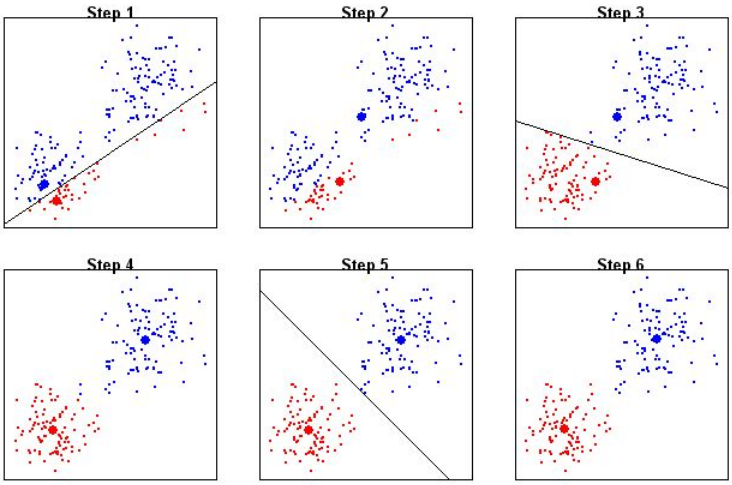


- Sort the points by y-axis
- Move a "sweep line" from top to bottom,
- Keep track of the "beach line": all points *equidistant* from the sweep line & one of the points on the top. (A curve made from parabolas)
- Above the beach line the diagram has been finalized.
- *Running time: O(n log n)*

Today

- Reading: “LineUp: Visual Analysis of Multi-Attribute Rankings”
- Let’s do Computational Geometry
  - Convex Hull
  - Voronoi Diagram
  - k-Means Clustering
- HW4: Data Collection
- Today’s Worksheet: Brainstorming Data

K-Means Clustering

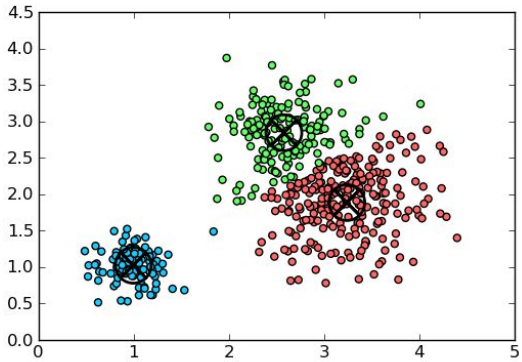


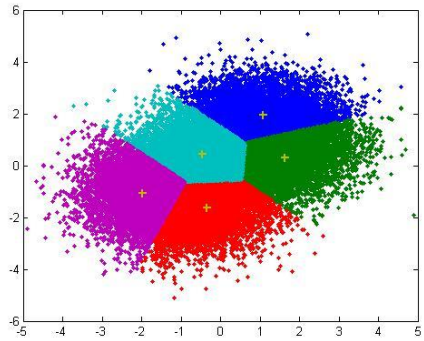
For a set of 2D/3D/*n*D points:

- Choose k, how many clusters
- Select k points from your data at random as initial “team” representatives
- Every other point determines which team representative it is closest to and joins that team
- Average the positions of all team members, this is the team’s new representative
- Repeat 3-5 times until change < threshold

K-means Clustering

- Works quite well, when the data can be meaningfully classified (and we know how many clusters to use).
- With dense data, output is visually similar to Voronoi diagram (k-Means chooses the data points that define the cells)





Today

- Reading: "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Let's do Computational Geometry
  - Convex Hull
  - Voronoi Diagram
  - k-Means Clustering
- HW4: Data Collection
- Today's Worksheet: Brainstorming Data

Homework Assignment 4: Data Collection & Data Preparation

- Identify an interesting (to you) and interestingly large, possibly high-dimensional data source that is currently "untapped"
  - Shouldn't just be a "download a file" dataset
  - Strongly encouraged to be related to your research!
- Use your programming skills to:
  - Obtain the data (what should you collect?)
  - Parse/Organize/Simplify/Post-Process the data
- Visualize the data using a tool that's new to you… Excel, LineUp, Tableau, Google Analytics, or Plotly
  - Try Parallel Coordinates? Convex Hull? Voronoi Diagram? k-Means Clustering?
- *Teams of 2 encouraged! [but optional] (… perhaps someone you didn't know before this class?)*

Reading for Tuesday

Post a comment/question on LMS by Tuesday @10am

"Angular Brushing of Extended Parallel Coordinates", Hauser, Ledermann, and Doleisch, InfoVis 2002