**Exam notes**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data abstraction**  **Attribute(field) type**  **Categorical**   * No implicit ordering * e.g., name, id, personality, color, location, Boolean, species, model   **Ordinal**   * Implicit ordering exists * Cannot compare numerically * e.g., rank, size (S, M, L), age, weight class, drama episode, queue, 관람등급   **Quantitative**   * Implicit ordering exists * Can compare numerically * e.g., time, rating, weight, length, year, date   **Attribute(field) property**  **Cardinality**   * # Of unique values * Exist in categorical or ordinal attribute   **Range**   * [Minimum, maximum] * Exist in quantitative attributes   **Dataset type**  **Table**   * item: row of table * attribute(field): column  |  |  | | --- | --- | | Table type | Key | | scatterplot | 0 key | | flat table | 1 key | | 2 dimensional | 2 keys |  * Attribute is attached to item.   **Network**   * In network dataset, attribute is attached to node or link. * Tree: special type of network which has root, and each connected pairs have parent and child relation.   **Spatial (geometry)**   * Attribute is attached to location(position) or region. |

|  |
| --- |
| **Mark**   * Point: can encode on size, shape * Line: can encode on length, width * Interlocking area: can encode on length, width. cannot shape code.   + cannot change without change of another area   + area in pie chart, geometry in map * Containment * Connection line |

|  |
| --- |
| **Magnitude channel: how big?**   * it is used for ordinal, quantitative attribute * effectiveness: position>length>angle>area>depth>luminance>saturation>curvature>volume   **Identity channel: What type?**   * It is used for categorical attribute * effectiveness: spatial region>hue>motion>shape * Grouping using identity channel * proximity: group using the same spatial region * similarity: group using same identity channel (e.g., hue)   **Channel expressiveness: match attribute type to suitable channel**  Low expressiveness: Using length for categorical attribute  **Channel effectiveness: 채널의 순위 (더 나은 채널이 있다)**   * 순위는 accuracy, discriminability, separability, pop out 네 가지 종목을 기준으로 선정 되었다.   **Discriminability(distinguishability): How many usable steps?**   * 1개의 채널이 **구별이 되게 사용 되었는가** * It should be able to compare two lengths * Example of **low discriminability: length1=1.5, length2=1.51** * Angle has low discriminability   **Separability: Human ability to use channel affected by another one?**   * 1개의 mark가 사용하는 2개의 채널이 서로 영향을 미치는가 * 영향을 안미치는 경우(fully separable): position&hue(color) * 약간 영향을 미치는 경우: size&hue(color) * 서로 영향이 큰 경우: width&height * 영향을 무조건 미치는 경우: red&green value in color   **Accuracy**   * how precisely can we tell the difference?   **Position channel usage**   * express value * separate, order, align |

|  |
| --- |
| **Color**   * luminance (how bright): used in detecting edge * saturation (how colorful) * hue (what color)   A group of squares with different colors  Description automatically generated  **Color palette**  **Categorical color palette**   * Pros: maximize distinguishability of marks * [red, yellow, orange, purple] * Cons: limited number of discriminable bins due to relative comparison (6~12 bins including background)   **Sequential color palette (quantitative/ordinal attribute)**   * Useful when going from min to max (Contiguous) * Only change luminance or only change saturation   **Diverging color palette (quantitative/ordinal attribute)**   * 2 hues at end point and interpolate (luminance or saturation) * useful when dataset has midpoint * neutral color(w/y/g) at midpoint   A close up of a chart  Description automatically generated  **Color map**   * Univariate and bivariate colormaps are good for quantitative attribute, ordinal attribute   **Univariate color map: single scale**   * range of attributes: [0~100] (only positive)   **Bivariate color map: 2 different attributes at row and column**   * Show 2 types of attributes using only color channel * Example of selected attributes: candidate A, candidate B * range of attributes: [-100,100] (negative exist)   A diagram of a mathematical model  Description automatically generated with medium confidenceCategorical+Saturation  **Color deficiency**   * Human use luminance to detect color * solution: vary luminance, change shape of mark   **용어**   * segmentation: bin by attribute by range * continous: interpolate color continuously |

|  |
| --- |
| **Multiview**  **Multiform**   * 각 view들이 **같은 input dataset을 사용**한다 + **encoding이 다르다** * 각 view에서 **다른 column들의 집합을 사용**하는 것을 허용한다 * 같은 dataset을 사용하되, **각 view에서 다른 item(row)의 집합을 사용**하는 것을 허용한다 (일부 overlap허용)   **Small multiples**   * 각 view들이 **같은 input dataset을 사용**한다. + **encoding이 같다** * 같은 dataset을 사용하되, **각 view에서 다른 item(row)의 집합을 사용**하는 것을 허용한다 (일부 overlap허용) * **지도 등이 나열되어 있는 경우**   **Overview+detail+same form**   * 각 **detail view들**이 original dataset의 **일부 item(row)**를 사용한다+**encoding이 같다** * 돋보기 view에 대한 영역 선택자가 있어야 한다. * **돋보기 view가 존재한다**   **Overview+detail+multiform**   * 각 **detail view들**이 original dataset의 **일부 item(row)를 사용한다**+**encoding이 다르다** * 돋보기 view에 대한 영역 선택자가 있어야 한다. * **돋보기 view가 존재한다**   **Fully redundant (copies of same visualization)**  **No linkage (not related across views)**  **A screenshot of a data analysis  Description automatically generated** |
| **Combination**  **Small multiples+multiform**   * small multiple인 view들이 있고, 또 다른 view가 있다. * 같은 item에서 다른 column을 사용하는 것을 허용한다(dendrogram)   A diagram of a diagram of a diagram  Description automatically generated with medium confidence  **Small multiples+overview detail**   * small multiple인 상태에서 작은 view에 대한 부연 설명인 view가 있는 경우   A chart of weather forecast  Description automatically generated |

|  |
| --- |
| **Tradeoff: Maps**  **Choropleth**   * Usage: show only 1 quantitative attribute of each region. understand distribution of quantitative attribute across spatial regions * It needs to use sequential color palette. * Pros: easy to read (no learning curve) * Cons: effect that large region appear to be more important when the goal is to encode attribute value   **Symbol map**   * Usage: show aggregated data with mark or glyph * Pros: solve the problem that larger region appearing to be more important because we use marks or glyphs in symbol map * Cons: Symbol can overlap (occlusion)   **Cartogram**   * Type: continuous cartogram, grid cartogram * Pros: Capture the viewer's interest using differences in size * Cons: Viewer need to know original dataset in memory to compare distorted mark on the cartogram   **Dot density map**   * Usage: each circle represents constant number of items. Every circle is same size and shape. Circle mark only allow color channel. * Pros: solve the problem that larger region appearing to be more important * Cons: Region with more population appear to be more important when we want to represent attribute in each mark |

|  |
| --- |
| **Tradeoff: Orientation**  **Rectilinear layout**   * Pros: Familiar to user, easy to accurately encode horizontal vertical position * Cons: Only 2 axis is good. 3~ axes reduce effectiveness due to depth   **Parallel layout**   * Pros: can visualize many attributes simultaneous * Cons: Unfamiliar to user (need training time)   **Radial layout**   * Pros: Can bin items in radial section * Cons: Human cannot read angle accurately |

|  |
| --- |
| **Tradeoff:4 ways to handle complexity**  **Deriving new attribute from existing attributes**   * f (export, import)->trade balance * Pros: Doesn't need to compute in mind   **Changing view overtime (Interactive view)**   * Use widget and control (button, checkbox, dropdown) * Cons: use screen space   **Reduce**  **Filter (throw things away) == query**   * Pros: make computation cheaper * Cons: hard to reason about filtered out items with existing items   **Aggregate (group)**   * We can aggregate into cluster/histogram * Pros: sum up into simple signal * Cons: difficult to pick the right signal for group   **Facet(divide) into multiple views**  **Juxtapose (put two chart side by side)**   * difficulty: Visual encoding choice, linking two views, number of data shared in 2 views   **Partition one attribute into multiple views**   * split into regions by attribute * Order of views matter   **Superimpose (overlay charts in one view)**   * Cons: overlapping marks if put too many attributes |

|  |
| --- |
| **Tradeoff: node-link diagram vs adjacency matrix**  **Node link (good for small network)**   * It can be used for both network and tree * Pros: Easy to understand topology (path, structure) * Cons: Overlap when scaled   **Adjacency matrix (good for large network)**   * It can be used for both network and tree * Pros: Focus on edges, Scalable * Cons: Bad at path tracing   **Enclosure (Containment)**   * It can only be used for tree * Pros: Hierarchical structure * Cons: Doesn't scale well |

|  |
| --- |
| **Task abstraction {Action, Target}**  **Action**   * discover(find) * lookup: know both target and location in the idiom (dictionary) * locate: know target but do not know location (node in network) * browse: know where to look(location), but do not know what you are looking for(target) (finding book in bookstore) * explore: don't know location and target (완전 새로운 것을 찾을 때) * identify: find one item * compare: comparison between multiple items * summarize: summary on the whole dataset   **Target**  **By looking at whole dataset**   * trend (pattern): increase, decrease * outlier * feature: any particular structure of interest in DSL   **By looking at attribute**   * 1 attribute: distribution, extrema * many attributes: dependency, correlation, similarity between 2 attributes   **etc.**   * path of network * shape (tumor) |

|  |
| --- |
| **Limitation==Weakness** |
| **Occlusion**   * occlusion: marks or label overlap one another making it hard to read value (마크가 겹쳐 있는 경우) * Views are too small that text label occludes line mark   **Clutter**   * clutter: too many marks and label packed closely together leading to distraction (많은 마크가 있는경우) * Excessive line marks exist next to each other within a limited space in bar chart led to clutter and reduced readability (clutter) * Line mark does not add meaningful information (clutter) |
| **Information density**   * Information density: There is a significant amount of empty and wasted space within the view.   **Layout density**  point mark가 픽셀 1개 만큼 작은 경우. |
| **Label**   * No legend or label: it is impossible to determine the actual values * Line mark is not labeled making it hard to specify item * There are no numbers or labels in the pie chart, so it is difficult to accurately read or interpret data * The height on the y-axis does not specify which quantitative attribute it is but it only displays category labels |
| **Effectiveness**   * 채널 type을 보고 **더 나은 effectiveness를 가진 채널을 언급**한다. * Angle channel encoding makes it hard to accurately estimate, compare, and interpret unlike easily perceivable length or position differences (angle channel을 사용하는 경우) * It is more difficult to accurately judge and compare sizes based on area than to do so using length (area channel을 사용하는 경우) * Pie charts are more effective visualizing part-to-whole relationships compared to using simple circle marks * **Distinguishability**: color of point marks is not distinguishable (confusing color or using too many color bins) |
| **Radial layout vs rectilinear layout**   * Radial layout is less accurately perceived than a rectilinear layout (position on x,y axes) because human readability to read angle channel is lower precision than length channel * In polar coordinate, area in the same sector should have the similarity * 3D is bad: In rectilinear layouts, using 2 axes is optimal, as additional axes beyond 2 decrease comprehension due to the low efficiency of the depth channel * In parallel coordinate it is unfamiliar to individuals, so it takes training time to interpret idiom. |
| **Normalization에 대한 내용**   * Normalizing two circle marks to the larger of the two makes it difficult to accurately comprehend their proportional relationship |
| **View**   * Understanding per capita data is difficult because it requires mentally dividing the bars on the left view by those on the right view * Direction change: When lines go back and forth a lot in a picture, it gets harder to compare them, making it difficult understand patterns or trends * Views in small multiples use different scale axis, so there's no consistent reference point for absolute value comparison (small multiple인 경우) |
| **Inconsistency**   * The irregular spacing in spiral due to varying radius makes it difficult to find trends * Width of bar representing category is not fixed, so if width is small, it is hard to read label |
| **Color**   * Change in luminance or saturation is good for ordinal or quantitative data while different hues are good for representing categorical data * Having too many color bins for categories more than 6 bins makes it difficult to distinguish one another |
| **Task**   * Task does not match purpose of idiom |

**Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Idiom | Input attribute | | | Task | Scalability |
| Idiom | Categorical | Quantitative | Ordinal | {action, target} |  |
| Stacked bar chart | 2 | 1 |  | {Part-to-whole}, {lookup, value}, {find, trend} | stacked glyph: ~12 |
| Bar chart | 1 | 1 |  | {compare/lookup, value} | Key attribute: 12s~100s |
| Heatmap | 2 | 1 |  | {find, clusters/outliers} | item: 1M, categorical attribute level: 100s, quantitative attribute level: 3~11 |
| Dot/Line chart |  | 2 |  | {find, trend} | key categorical attribute: 100s |
| Pie chart | 1 | 1 |  | {part-whole relationship} | category: 12 |
| Scatterplot |  | 2 |  | {find, trend/outlier/distribution/correlation}, {locate, cluster} | item: 100s |
| Histogram |  | 1 |  | {find, distribution} |  |
| SPLOM |  | many | many | {find, correlation/trend, outlier} | attribute: 12 |
| parallel coordinate |  | many | many | {find, trend/outlier/extrema/correlation} | vertical line: 12s, item: 100s |
| Boxplot |  | many |  | {find, outlier/extrema/average}, {identify, skew} | attribute: 12s |
| Streamgraph | 1 (movie) | 1 (count) | 1 (time) | {discover, trend/distribution/extrema} | time key: 100s, movie key: 12s~100s |

**Geometry**

|  |  |  |  |
| --- | --- | --- | --- |
| Idiom | Input attribute | Task {action, target} | Scalability |
| Choropleth map | Geometry, 1 quantitative attribute per region | {understand, distribution} | size of input dataset |
| Symbol map | Geometry, 1 attribute per region | everything | size of input dataset |
| Cartogram | Geometry, 1 attribute per region | everything | size of input dataset |
| Dot density map | Geometry, 1 attribute per region | everything | size of input dataset |

**Modifiable Areal Unit Problem (개리멘더링으로 조작)**

* 지도 view에서 boundary의 크기를 바꾸면 region의 encoding된 weight 이 바뀐다.
* zone effect: boundary의 모양에 따라 region의 weight이 바뀌는 현상
* scale effect: boundary의 크기에 따라 weight이 바뀌는 현상

**Tree: choose based on task**

|  |  |  |  |
| --- | --- | --- | --- |
| Idiom | Input attribute | Task {action, target} | Scalability |
| Tree map | tree data structure | {query, attribute at leaf node} | Leaf node: 1M |
| Force directed placement (node-link) | Network | {explore, topology}, {locate, path} | Link<4\*Node |
| Adjacency matrix | Network | {find, edge topology} | Node: 1K, Edge 1M |

|  |
| --- |
| **Design idiom process**   1. **Analyze Dataset**  * Dataset type: table, network, geometry * type of attribute  1. **Abstract task** 2. **Consideration phase**  * Cardinality == scalability of idiom * Channel: effectiveness   Never use radial bar chart  시험은 흑백이므로 luminance만 바뀌는 sequential color palette만 사용한다  **위젯 사용**   * filter: selection (button, radio button), drag, legend, hover, dropdown, checkbox 사용 * query: search box, query language 사용 |