Visualization Analysis & Design
Full-Day Tutorial
Session 1

Tamara Munzner
Department of Computer Science
University of British Columbia

Sanger Institute / European Bioinformatics Institute
June 2014, Cambridge UK

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14
Outline

• Visualization Analysis Framework
  Session 1 9:30-10:45am
  – Introduction: Definitions
  – Analysis: What, Why, How
  – Marks and Channels

• Idiom Design Choices, Part 2
  Session 3 1:15pm-2:45pm
  – Manipulate: Change, Select, Navigate
  – Facet: Juxtapose, Partition, Superimpose
  – Reduce: Filter, Aggregate, Embed

• Idiom Design Choices
  Session 2 11:00am-12:15pm
  – Arrange Tables
  – Arrange Spatial Data
  – Arrange Networks and Trees
  – Map Color

• Guidelines and Examples
  Session 4 3-4:30pm
  – Rules of Thumb
  – Validation
  – BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14
Outline

• Visualization Analysis Framework
  Session 1 9:30-10:45am
  – Introduction: Definitions
  – Analysis: What, Why, How
  – Marks and Channels

• Idiom Design Choices, Part 2
  Session 3 1:15pm-2:45pm
  – Manipulate: Change, Select, Navigate
  – Facet: Juxtapose, Partition, Superimpose
  – Reduce: Filter, Aggregate, Embed

• Idiom Design Choices
  Session 2 11:00am-12:15pm
  – Arrange Tables
  – Arrange Spatial Data
  – Arrange Networks and Trees
  – Map Color

• Guidelines and Examples
  Session 4 3-4:30pm
  – Rules of Thumb
  – Validation
  – BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14
Defining visualization (vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Why?...
Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

• don’t need vis when fully automatic solution exists and is trusted
• many analysis problems ill-specified
  – don’t know exactly what questions to ask in advance
• possibilities
  – long-term use for end users (e.g. exploratory analysis of scientific data)
  – presentation of known results
  – stepping stone to better understanding of requirements before developing models
  – help developers of automatic solution refine/debug, determine parameters
  – help end users of automatic solutions verify, build trust
Why use an external representation?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- external representation: replace cognition with perception

Why have a computer in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- beyond human patience: scale to large datasets, support interactivity
  – consider: what aspects of hand-drawn diagrams are important?

Why depend on vision?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- human visual system is high-bandwidth channel to brain
  - overview possible due to background processing
    - subjective experience of seeing everything simultaneously
    - significant processing occurs in parallel and pre-attentively
- sound: lower bandwidth and different semantics
  - overview not supported
    - subjective experience of sequential stream
- touch/haptics: impoverished record/replay capacity
  - only very low-bandwidth communication thus far
- taste, smell: no viable record/replay devices
Why show the data in detail?

• summaries lose information
  – confirm expected and find unexpected patterns
  – assess validity of statistical model

**Anscombe’s Quartet**

<table>
<thead>
<tr>
<th>Identical statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x mean</td>
<td>9</td>
</tr>
<tr>
<td>x variance</td>
<td>10</td>
</tr>
<tr>
<td>y mean</td>
<td>8</td>
</tr>
<tr>
<td>y variance</td>
<td>4</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>1</td>
</tr>
</tbody>
</table>
Idiom design space

The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations.

• **idiom**: distinct approach to creating or manipulating visual representation

  – how to draw it: **visual encoding** idiom
    • many possibilities for how to create

  – how to manipulate it: **interaction** idiom
    • even more possibilities
      – make single idiom dynamic
      – link multiple idioms together through interaction

Why focus on tasks and effectiveness?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks and effectively.

• tasks serve as constraint on design (as does data)
  – idioms do not serve all tasks equally!
  – challenge: recast tasks from domain-specific vocabulary to abstract forms

• most possibilities ineffective
  – validation is necessary, but tricky
  – increases chance of finding good solutions if you understand full space of possibilities

• what counts as effective?
  – novel: enable entirely new kinds of analysis
  – faster: speed up existing workflows
Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.

- **computational limits**
  - processing time
  - system memory

- **human limits**
  - human attention and memory

- **display limits**
  - pixels are precious resource, the most constrained resource
  - **information density**: ratio of space used to encode info vs unused whitespace
    - tradeoff between clutter and wasting space, find sweet spot between dense and sparse
Further reading

  – Chap 1: What’s Vis, and Why Do It?
Outline

• Visualization Analysis Framework
  Session 1 9:30-10:45am
  – Introduction: Definitions
  – Analysis: What, Why, How
  – Marks and Channels

• Idiom Design Choices, Part 2
  Session 3 1:15pm-2:45pm
  – Manipulate: Change, Select, Navigate
  – Facet: Juxtapose, Partition, Superimpose
  – Reduce: Filter, Aggregate, Embed

• Idiom Design Choices
  Session 2 11:00am-12:15pm
  – Arrange Tables
  – Arrange Spatial Data
  – Arrange Networks and Trees
  – Map Color

• Guidelines and Examples
  Session 4 3-4:30pm
  – Rules of Thumb
  – Validation
  – BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14
Analysis: What, why, and how

- **what** is shown?
  - **data** abstraction

- **why** is the user looking at it?
  - **task** abstraction

- **how** is it shown?
  - **idiom**: visual encoding and interaction

- abstract vocabulary avoids domain-specific terms
  - translation process iterative, tricky

- what-why-how analysis framework as scaffold to think systematically about design space
## What?

- **Data Types**
  - Items
  - Attributes
  - Links
  - Positions
  - Grids

- **Data and Dataset Types**
  - **Tables**
    - Items
    - Attributes
  - **Networks & Trees**
    - Items (nodes)
    - Grids
    - Positions
    - Attributes
  - **Geometry**
    - Items
    - Positions
  - **Clusters, Sets, Lists**
    - Items

## Attributes

- **Attribute Types**
  - Categorical
  - Ordered
    - Ordinal
    - Quantitative

## Dataset Types

- **Tables**
  - **Multidimensional Table**
  - Attributes (columns)
  - Items (rows)
  - Cell containing value

- **Networks**
  - **Trees**
  - Link
  - Node (item)
  - Grid of positions
  - Attributes (columns)
  - Value in cell

- **Geometry**
  - **Spatial**
  - Position

## Ordering Direction

- **Sequential**
- **Diverging**
- **Cyclic**

## Dataset Availability

- **Static**
- **Dynamic**
Dataset types

Dataset Types

→ Tables

Attributes (columns)

Items (rows)

Cell containing value

→ Networks

Link

Node (item)

→ Fields (Continuous)

Grid of positions

Cell

Attributes (columns)

Value in cell

→ Geometry (Spatial)

Position

→ Multidimensional Table

Key 1

Key 2

Value in cell

→ Trees
# Dataset and data types

## Data and Dataset Types

<table>
<thead>
<tr>
<th>Tables</th>
<th>Networks &amp; Trees</th>
<th>Fields</th>
<th>Geometry</th>
<th>Clusters, Sets, Lists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Items (nodes)</td>
<td>Grids</td>
<td>Items</td>
<td>Items</td>
</tr>
<tr>
<td>Attributes</td>
<td>Links</td>
<td>Positions</td>
<td>Attributes</td>
<td>Positions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Data Types

- Items
- Attributes
- Links
- Positions
- Grids

## Dataset Availability

- Static
- Dynamic
Attribute types

- **Attribute Types**
  - Categorical
    - Symbols: +, ●, □, △
  - Ordered
    - Symbols: ▲, ▼
  - Ordinal
    - Symbols: ▲, ▼
  - Quantitative
    - Symbols: |

- **Ordering Direction**
  - Sequential
    - Symbols: →
  - Diverging
    - Symbols: ↔
  - Cyclic
    - Symbols: ⌘
• {action, target} pairs
  – discover distribution
  – compare trends
  – locate outliers
  – browse topology
High-level actions: Analyze

• consume
  – discover vs present
    • classic split
    • aka explore vs explain
  – enjoy
    • newcomer
    • aka casual, social

• produce
  – annotate, record
  – derive
    • crucial design choice
Actions: Mid-level search, low-level query

• what does user know?
  – target, location

• how much of the data matters?
  – one, some, all
Why: Targets

- **ALL DATA**
  - Trends
  - Outliers
  - Features

- **ATTRIBUTES**
  - One
    - Distribution
    - Extremes
  - Many
    - Dependency
    - Correlation
    - Similarity

- **NETWORK DATA**
  - Topology
    - Paths

- **SPATIAL DATA**
  - Shape
**Encode**

- **Arrange**
  - Express
  - Separate

- **Order**
  - Align

- **Use**

**Map**
from *categorical* and *ordered* attributes

- **Color**
  - Hue
  - Saturation
  - Luminance

- **Size, Angle, Curvature, ...**

- **Shape**

- **Motion**
  *Direction, Rate, Frequency, ...*

**Manipulate**

- **Change**

- **Select**

- **Navigate**

**Facet**

- **Juxtapose**

- **Partition**

- **Superimpose**

**Reduce**

- **Filter**

- **Aggregate**

- **Embed**

---

**How?**

**What?**

**Why?**

**How?**
Analysis example: Compare idioms

**SpaceTree**

- Actions: Present, Locate, Identify
- Targets: Path between two nodes

**TreeJuxtaposer**

- Actions: Encode, Navigate, Select, Filter, Aggregate
- Targets: Path between two nodes


Chained sequences

• output of one is input to next
  – express dependencies
  – separate means from ends
Analysis example: Derive one attribute

- Strahler number
  - centrality metric for trees/networks
  - derived quantitative attribute
  - draw top 5K of 500K for good skeleton


Task 1

- In Tree
- Out Quantitative attribute on nodes

What?
- In Tree
- Out Quantitative attribute on nodes

Why?
- Derive

Task 2

- In Tree
- Out Quantitative attribute on nodes
- In Tree
- In Quantitative attribute on nodes
- Out Filtered Tree
- Removed unimportant parts

What?
- In Tree
- In Quantitative attribute on nodes
- Out Filtered Tree

Why?
- Summarize
- Topology

How?
- Reduce
- Filter
Further reading

  – Chap 2: What: Data Abstraction
  – Chap 3: Why: Task Abstraction


Outline

• Visualization Analysis Framework
  Session 1 9:30-10:45am
  – Introduction: Definitions
  – Analysis: What, Why, How
  – Marks and Channels

• Idiom Design Choices, Part 2
  Session 3 1:15pm-2:45pm
  – Manipulate: Change, Select, Navigate
  – Facet: Juxtapose, Partition, Superimpose
  – Reduce: Filter, Aggregate, Embed

• Idiom Design Choices
  Session 2 11:00am-12:15pm
  – Arrange Tables
  – Arrange Spatial Data
  – Arrange Networks and Trees
  – Map Color

• Guidelines and Examples
  Session 4 3-4:30pm
  – Rules of Thumb
  – Validation
  – BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14
Visual encoding

• analyze idiom structure
Definitions: Marks and channels

• marks
  – geometric primitives

• channels
  – control appearance of marks
  – can redundantly code with multiple channels

• interactions
  – point marks only convey position; no area constraints
    • can be size and shape coded
  – line marks convey position and length
    • can only be size coded in 1D (width)
  – area marks fully constrained
    • cannot be size or shape coded
Visual encoding

• analyze idiom structure
  – as combination of marks and channels

1: vertical position
mark: line

2: vertical position
horizontal position
mark: point

3: vertical position
horizontal position
color hue
mark: point

4: vertical position
horizontal position
color hue
size (area)
mark: point
Channels: Expressiveness types and effectiveness rankings

**Magnitude Channels: Ordered Attributes**
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**
- Spatial region
- Color hue
- Motion
- Shape
Effectiveness and expressiveness principles

• **effectiveness principle**
  – encode most important attributes with highest ranked channels

• **expressiveness principle**
  – match channel and data characteristics


• rankings: where do they come from?
  – accuracy
  – discriminability
  – separability
  – popout
Accuracy: Fundamental Theory

Steven’s Psychophysical Power Law: $S = I^N$
Accuracy: Vis experiments

Discriminability: How many usable steps?

- linewidth: only a few
Separability vs. Integrality

<table>
<thead>
<tr>
<th>Position + Hue (Color)</th>
<th>Size + Hue (Color)</th>
<th>Width + Height</th>
<th>Red + Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully separable</td>
<td>Some interference</td>
<td>Some/significant interference</td>
<td>Major interference</td>
</tr>
<tr>
<td>2 groups each</td>
<td>2 groups each</td>
<td>3 groups total: integral area</td>
<td>4 groups total: integral hue</td>
</tr>
</tbody>
</table>
Popout

• find the red dot
  – how long does it take?

• parallel processing on many individual channels
  – speed independent of distractor count
  – speed depends on channel and amount of difference from distractors

• serial search for (almost all) combinations
  – speed depends on number of distractors
• many channels: tilt, size, shape, proximity, shadow direction, ...
• but not all! parallel line pairs do not pop out from tilted pairs
Grouping

- containment
- connection

Marks as Links

- Containment
- Connection

Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape
Relative vs. absolute judgements

• perceptual system mostly operates with relative judgements, not absolute
  – that’s why accuracy increases with common frame(scale) and alignment
  – Weber’s Law: ratio of increment to background is constant
    • filled rectangles differ in length by 1:9, difficult judgement
    • white rectangles differ in length by 1:2, easy judgement

length

position along unaligned common scale

position along aligned scale

Further reading

  – Chap 5: Marks and Channels


