Week 1: Intro, Tasks and Data, Marks and Channels

Tamara Munzner
Department of Computer Science
University of British Columbia

JRNL 520H, Special Topics in Contemporary Journalism: Data Visualization
Week 1: 13 September 2016

http://www.cs.ubc.ca/~tmm/courses/journ16
Who’s who

• Instructor: Tamara Munzner
  – UBC Computer Science

• Instructor: Caitlin Havlak
  – Discourse Media
Class time

• 6 weeks, Sep 13 - Oct 18
  – once/week, 3 hr session 9:30am-12:30pm

• standard week
  – foundations lecture/discussion: 80 min
  – break: 15 min
  – demos: 45 min
  – lab: 30 min

• office hrs: 1-3pm most weeks
Structure

• participation, 10%
  – attend lectures and demos, discuss
    • tell us in advance if you’ll miss class (and why)
    • tell when us recover if you were ill

• homework, 90%
  – gradual transition from structured to open-ended
  – 60%: 5 assignments
    • best 4 out of 5 marks used, so 15% each
    • start in lab time, finish over the subsequent week
    • due just before next class session (9am)
      – some solo, some in groups of 2
  – 30%: final assignment
    • find your own interesting data and design your own visualization for it
Further reading

• optional textbook for following up on visualization foundations lectures
    – library has multiple ebook copies
    – to buy yourself, see course page

• optional textbook for more about Tableau software
    • [http://dataremixed.com/books/cdwt/](http://dataremixed.com/books/cdwt/)

• optional papers/books
  – links and references posted on course page
  – if DL links, use library EZproxy from off campus
Finding us

• office hours in Sing Tao bldg
  – 1-3pm Tuesdays: Tamara and/or Caitlin
  – by appointment: Tamara in ICICS/CS bldg Room X661

• email other times
  – tmm@cs.ubc.ca, caitlin@discoursemedia.org

• course page is font of all information
  – don’t forget to refresh, frequent updates
  – http://www.cs.ubc.ca/~tmm/courses/journ16
Topics

• Week 1
  – Intro
  – Tasks and Data
  – Marks and Channels

• Week 2
  – Arrange Data Tables

• Week 3
  – Color
  – Arrange Spatial Data

• Week 4
  – Manipulate, Facet, Reduce

• Week 5
  – Wrangle
  – Stories
  – Rules of Thumb

• Week 6
  – Networks
  – Regression Lines
  – Vis in Newsrooms
Introduction: 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 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>7.5</td>
</tr>
<tr>
<td>y variance</td>
<td>3.75</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>0.816</td>
</tr>
</tbody>
</table>
Why focus on tasks and effectiveness?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more 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
Why analyze?

- imposes structure on huge design space
  - scaffold to help you think systematically about choices
  - analyzing existing as stepping stone to designing new
  - most possibilities ineffective for particular task/data combination

**What?**

- **Tree**

**Why?**

- **Actions**
  - Present
  - Locate
  - Identify
  - Encode
  - Navigate
  - Select
  - Filter
  - Aggregate

- **Targets**
  - Path between two nodes
  - *SpaceTree* 
  - *TreeJuxtaposer* 

**How?**

- **SpaceTree**


- **TreeJuxtaposer**

Analysis framework: Four levels, three questions

- **domain** situation
  - who are the target users?

- **abstraction**
  - translate from specifics of domain to vocabulary of vis

- **what** is shown? **data abstraction**
  - often don’t just draw what you’re given: transform to new form

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

- **idiom**

- **how** is it shown?
  - visual encoding idiom: how to draw
  - interaction idiom: how to manipulate

- **algorithm**
  - efficient computation

---


Why is validation difficult?

- different ways to get it wrong at each level

- **Domain situation**
  - You misunderstood their needs

- **Data/task abstraction**
  - You’re showing them the wrong thing

- **Visual encoding/interaction idiom**
  - The way you show it doesn’t work

- **Algorithm**
  - Your code is too slow
Why is validation difficult?

- solution: use methods from different fields at each level

- **Domain situation**
  - Observe target users using existing tools

- **Data/task abstraction**
  - **Visual encoding/interaction idiom**
    - Justify design with respect to alternatives
  - **Algorithm**
    - Measure system time/memory
    - Analyze computational complexity
  - Analyze results qualitatively
  - Measure human time with lab experiment (*lab study*)
  - Observe target users after deployment (*field study*)
  - Measure adoption

- **Problem-driven work**
- **Technique-driven work**

**What?**

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

- **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>
</tbody>
</table>

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

- **Dataset Types**
  - Tables
  - Networks
  - Fields (Continuous)

- **Ordering Direction**
  - Sequential
  - Diverging
  - Cyclic

- **Geometry (Spatial)**

- **Dataset Availability**
  - Static
  - Dynamic
Three major datatypes

Dataset Types

→ Tables

→ Multidimensional Table

→ Networks

→ Trees

→ Spatial

→ Fields (Continuous)

→ Geometry (Spatial)

• visualization vs computer graphics
  – geometry is design decision
Dataset and data types

Data and Dataset Types

- Tables
  - Items
  - Attributes

- Networks & Trees
  - Items (nodes)
  - Links
  - Attributes

- Fields
  - Grids
  - Positions
  - Attributes

- Geometry
  - Items
  - Positions

- Clusters, Sets, Lists
  - Items

Data Types

- Items
- Attributes
- Links
- Positions
- Grids

Dataset Availability

- Static
- Dynamic
Attribute types

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

- **Ordering Direction**
  - Sequential
  - Diverging
  - Cyclic
• \{action, target\} pairs
  – discover distribution
  – compare trends
  – locate outliers
  – browse topology
Actions: Analyze

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

• produce
  – annotate, record
  – derive
    • crucial design choice
Derive

• don’t just draw what you’re given!
  – decide what the right thing to show is
  – create it with a series of transformations from the original dataset
  – draw that

• one of the four major strategies for handling complexity

Original Data

\[
\text{trade balance} = \text{exports} - \text{imports}
\]

Derived Data

\[
\text{trade balance} = \text{exports} - \text{imports}
\]
Actions: Search, query

- what does user know? (Search)
  - target, location

- how much of the data matters?
  - one, some, all

- independent choices for each of these three levels
  - analyze, search, query
  - mix and match

<table>
<thead>
<tr>
<th>Location known</th>
<th>Target known</th>
<th>Target unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lookup</td>
<td></td>
<td>Browse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location unknown</th>
<th>Target known</th>
<th>Target unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate</td>
<td></td>
<td>Explore</td>
</tr>
</tbody>
</table>
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

- **What?**
  - In Tree
  - Out Quantitative attribute on nodes
- **Why?**
  - Derive

Task 2

- **What?**
  - In Tree
  - In Quantitative attribute on nodes
- **Why?**
  - Summarize
- **How?**
  - Reduce
  - Filter

- **What?**
  - In Tree
  - Out Quantitative attribute on nodes
- **Why?**
  - Topology
- **How?**
  - Reduce
  - Filter

Out
Filtered Tree
Removed unimportant parts
Why: Targets

- **All Data**
  - Trends
  - Outliers
  - Features

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

- **Network Data**
  - Topology
    - Paths

- **Spatial Data**
  - Shape
How?

**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

- **Reduce**
  - Filter
  - Aggregate
  - Embed

---

What?

Why?

How?
Encoding visually

• analyze idiom structure
### Definitions: Marks and channels

- **marks**
  - geometric primitives

    - **Points**
    - **Lines**
    - **Areas**

- **channels**
  - control appearance of marks

    - **Position**
      - Horizontal
      - Vertical
      - Both

    - **Color**

    - **Shape**

    - **Size**
      - Length
      - Area
      - Volume
Encoding visually with marks and channels

• 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

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)
Spatial region
Color hue
Motion
Shape
Channels: 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 principle**
  - encode most important attributes with highest ranked channels

- **expressiveness principle**
  - match channel and data characteristics
Accuracy: Fundamental Theory

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

Cleveland & McGill's Results

Crowdsourced Results

Positions

Angles

Circular areas

Rectangular areas (aligned or in a treemap)

Log Error

Log Error


Discriminability: How many usable steps?

• must be sufficient for number of attribute levels to show
  – linewidth: few bins

[mappa.mundi.net/maps/maps_014/telegeography.html]
Separability vs. Integrality

Position + Hue (Color)

2 groups each

Size + Hue (Color)

2 groups each

Width + Height

3 groups total: integral area

Red + Green

4 groups total: integral hue

Fully separable

Some interference

Some/significant interference

Major interference
• 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
Popout

- 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

- proximity
  - same spatial region
- similarity
  - same values as other categorical channels
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

![Diagram of length and position along aligned and unaligned scales]

Relative luminance judgements

- perception of luminance is contextual based on contrast with surroundings

[Image: A diagram showing the effect of shadow on perceived luminance.]

http://persci.mit.edu/gallery/checkershadow
Relative color judgements

• color constancy across broad range of illumination conditions

http://www.purveslab.net/seeforyourself/
Further reading

  – Chap 1, What’s Vis, and Why Do It?
  – Chap 2, What: Data Abstraction
  – Chap 3, Why: Task Abstraction
  – Chap 4, Analysis: Four Levels for Validation
  – Chap 5, Marks and Channels

• Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. Jeffrey Heer and Michael Bostock. Proc. CHI 2010

• Perception in Vision web page with demos, Christopher Healey.

Next

• Break (15 min)

• Demos (45 min)
  – Caitlin will walk through Tableau demos
  – you follow along step by step on your own laptop
  – Tamara will rove the room to help out folks who get stuck

• Lab (30 min)
  – you’ll get started on Tableau assignment
Demo 1: Basic Visual Encoding & Dashboarding

• Tableau Lessons
  – Dimensions (categorical) and Measures (quantitative)
  – drag and drop to create visual encodings
  – combining multiple charts side by side into dashboards

• Big Ideas
  – see different patterns with different visual encodings
Demo 2: Vancouver Election Results

• Tableau Lessons
  – sorting along axis
  – disaggregate into multiple charts

• Big Ideas
  – absolute numbers can sometimes mislead
  – check hunches with relative percentages!
Demo 3: Vancouver Crime

• Tableau Lessons
  – multiple pills on a shelf, pill ordering
  – show filters
  – undo
  – duplicate & rename tabs

• Big Ideas
  – underlying causes can be tricky to understand
Demo 4: Back to the Future

• Tableau Lessons
  – simple analytics: totals
  – more disaggregation practice
  – Show Me

• Big Ideas
  – beyond simple bars
  – challenges of missing data
Assignment

- **Music Sales**
  - work through workbook on your own
  - submit finished version (in workbook .twbx format)
- **Vancouver Crime**
  - analyze further on your own
  - write up brief news story (submit in PDF format)
    - < 500 words
    - up to 2 screenshots from Tableau
  - write up reflections (submit in PDF format)
    - discuss dead ends
    - include Tableau screenshots
- submit before next class (9am Tue Sep 20)
  - email tmm@cs.ubc.ca and caitlin@discoursemedia.org with subject JOURN Week 1