Visual Analytics & Machine Learning

Course Info

- Instructor: Matthew Berger
- Contact: matthew.berger@vanderbilt.edu
- Office Hours: TR 2:00-3:00, JH-379
- **Course Website**: https://matthewberger.github.io/teaching/vaml/spring2020/

Agenda

- Visual analytics … ?

- Course logistics

Visual Analytics

- Data analysis
    - You have some high-level goal, often cannot be formulated by a crisp, computable, objective.
    - You have some dataset of interest, either specific to your goal, or potentially a proxy.
    - You then enter an analysis loop: you compute/extract something on your data, consume this information, gain knowledge, and *iterate*, until … goal solved.
- Visual **analytics**

Example (1)

- Consider the following dataset of movies:

Title	Genre	Rating	Budget	Revenue
Toy Story	adventure	3.9	30M	373M
Magnolia	drama	3.7	37M	48M
Zodiac	drama	3.7	65M	84M
Frank	comedy	4.3	1M	2M

- Question of interest: *what are high-grossing movies?*
- What steps would you take to answer this question?

Analytic Activity

Title	Genre	Rating	Budget	Revenue
Toy Story	adventure	3.9	30M	373M
Magnolia	drama	3.7	37M	48M
Zodiac	drama	3.7	65M	84M
Frank	comedy	4.3	1M	2M

- **Sort** by revenue
- Derive the **difference** of revenue and budget
- Observe **range** of revenues
- Characterize **distribution** of revenues

Example (2)

Title	Genre	Rating	Budget	Revenue
Toy Story	adventure	3.9	30M	373M
Magnolia	drama	3.7	37M	48M
Zodiac	drama	3.7	65M	84M
Frank	comedy	4.3	1M	2M

- Question of interest: *are dramas more successful than comedies?*

- What steps would you take to answer this question?

Analytic Activity

Title	Genre	Rating	Budget	Revenue
Toy Story	adventure	3.9	30M	373M
Magnolia	drama	3.7	37M	48M
Zodiac	drama	3.7	65M	84M
Frank	comedy	4.3	1M	2M

- **Compare distributions** between **filtered** movies
- **Correlate** budget with revenue
- “Success”: eye of the beholder, dependent on analysis
- How a human performs analysis is, thus, crucial

Example (3)

Title	Genre	Rating	Budget	Revenue	User	Rating
Frank	comedy	4.3	1M	2M	1	5
Frank	comedy	4.3	1M	2M	3	3
Zodiac	drama	3.7	37M	48M	4	2
Zodiac	drama	3.7	37M	48M	8	4

- Question of interest: *are dramas more popular than comedies?*

- What steps would you take to answer this question?

Analytic Activity

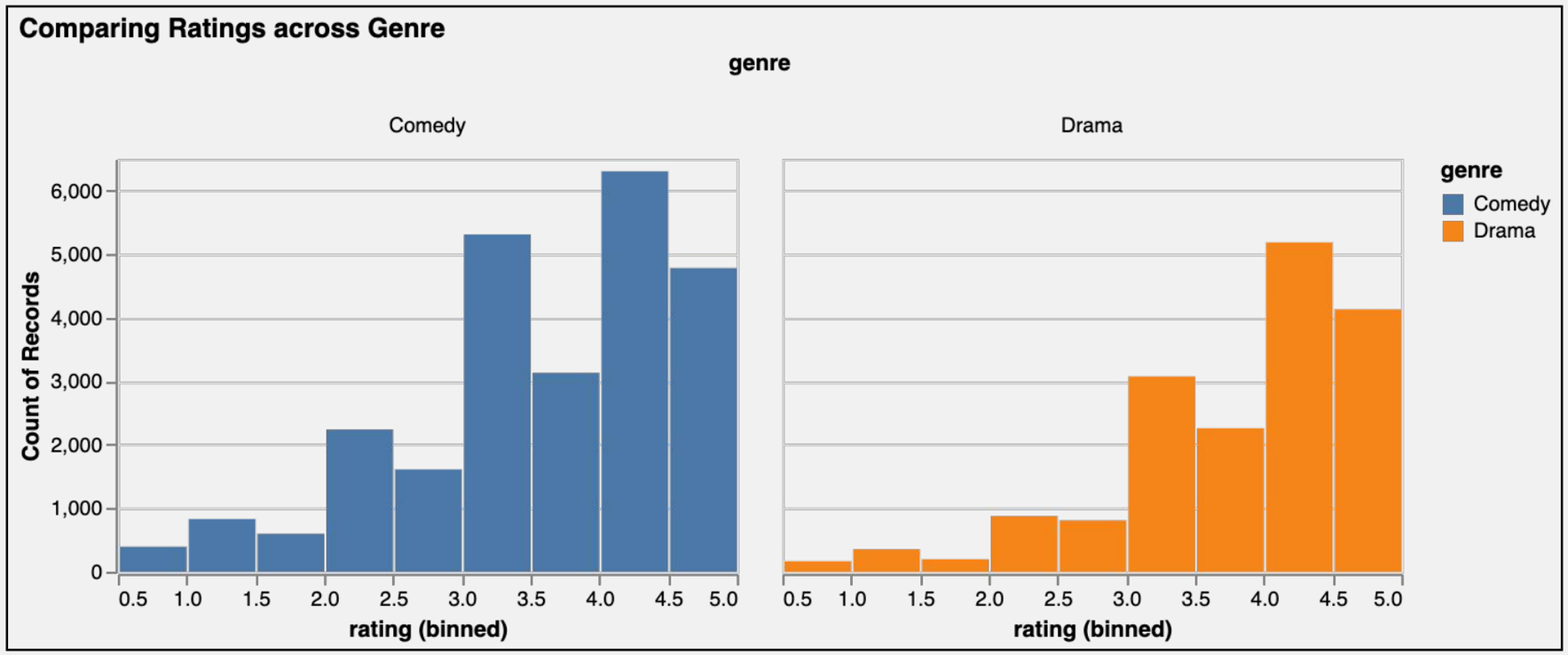
Title	Genre	Rating	Budget	Revenue	User	Rating
Frank	comedy	4.3	1M	2M	1	5
Frank	comedy	4.3	1M	2M	3	3
Zodiac	drama	3.7	37M	48M	4	2
Zodiac	drama	3.7	37M	48M	8	4

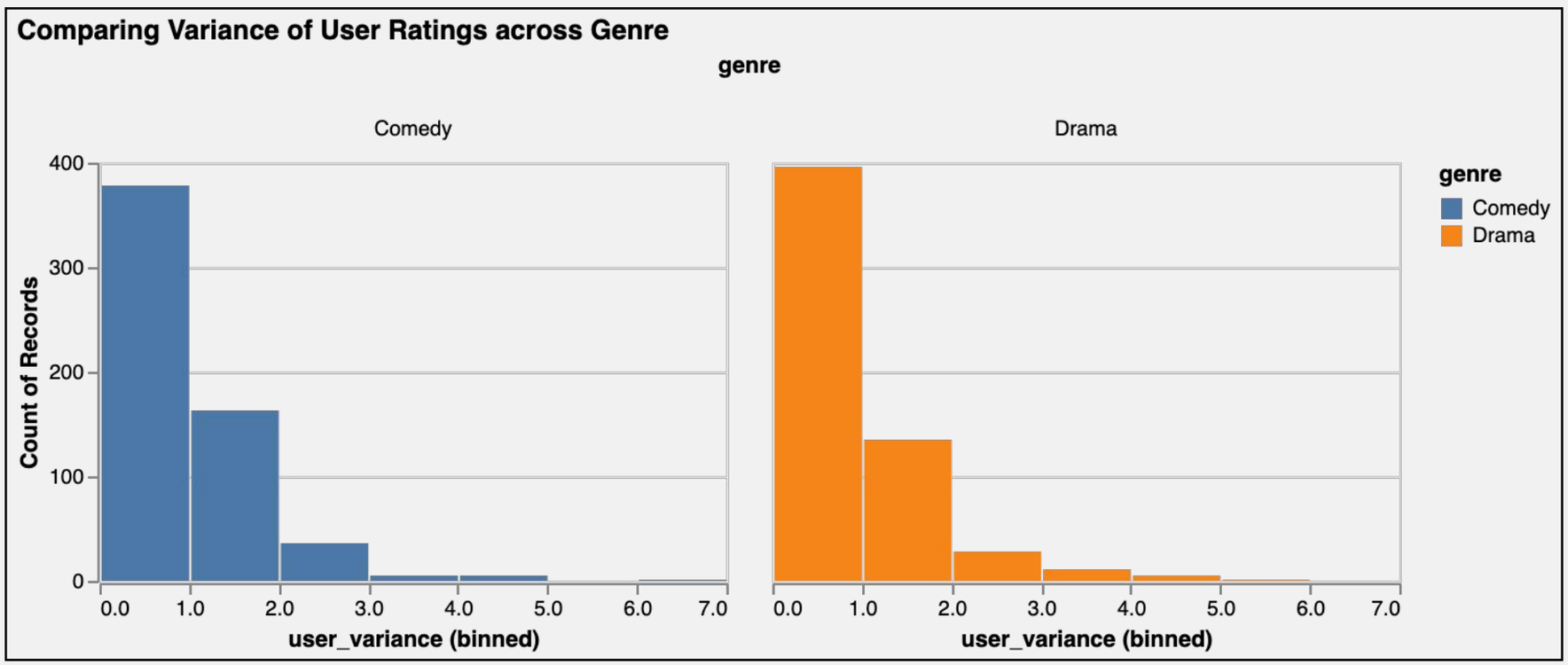
- What if certain users are spammers?
  - Inspect per-user **distributions**
  - **Model** user behavior, **compare** nominal users with **outliers**
- **Aggregate** and **compare** ratings, informed by user **models**

Back to visual analytics

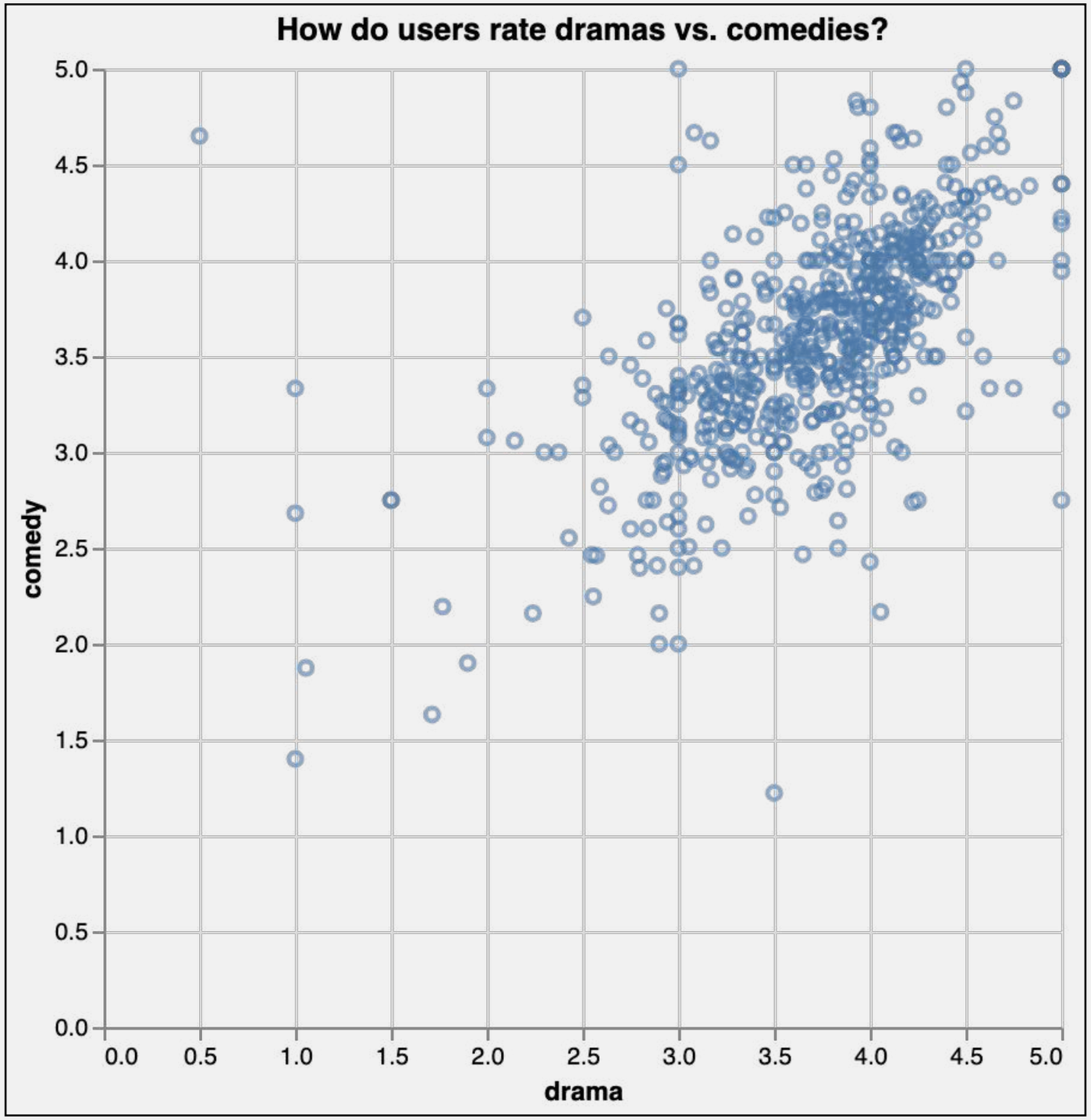
- Visualization: typically not necessary to achieve analysis goals
  - Can always stare at a table of numbers 😳
- Main purpose of visual analytics: *support* and *facilitate* humans in **analytical reasoning** through the use of **interactive visual interfaces** **[Thomas & Cook 2005]**
  - Make humans *more efficient*
  - Make humans *more effective*

Assessing Movie Popularity

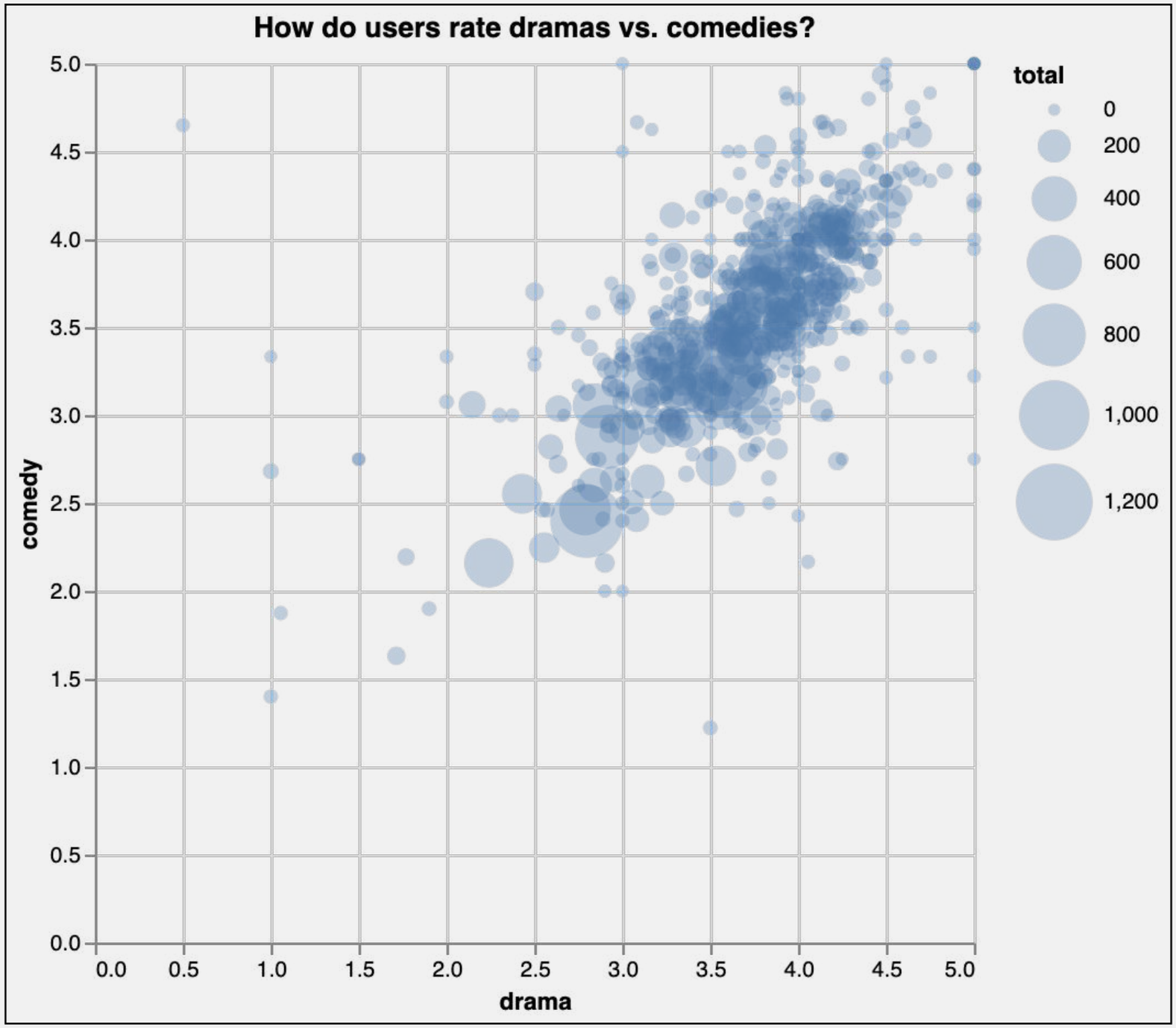




Assessing User Behavior

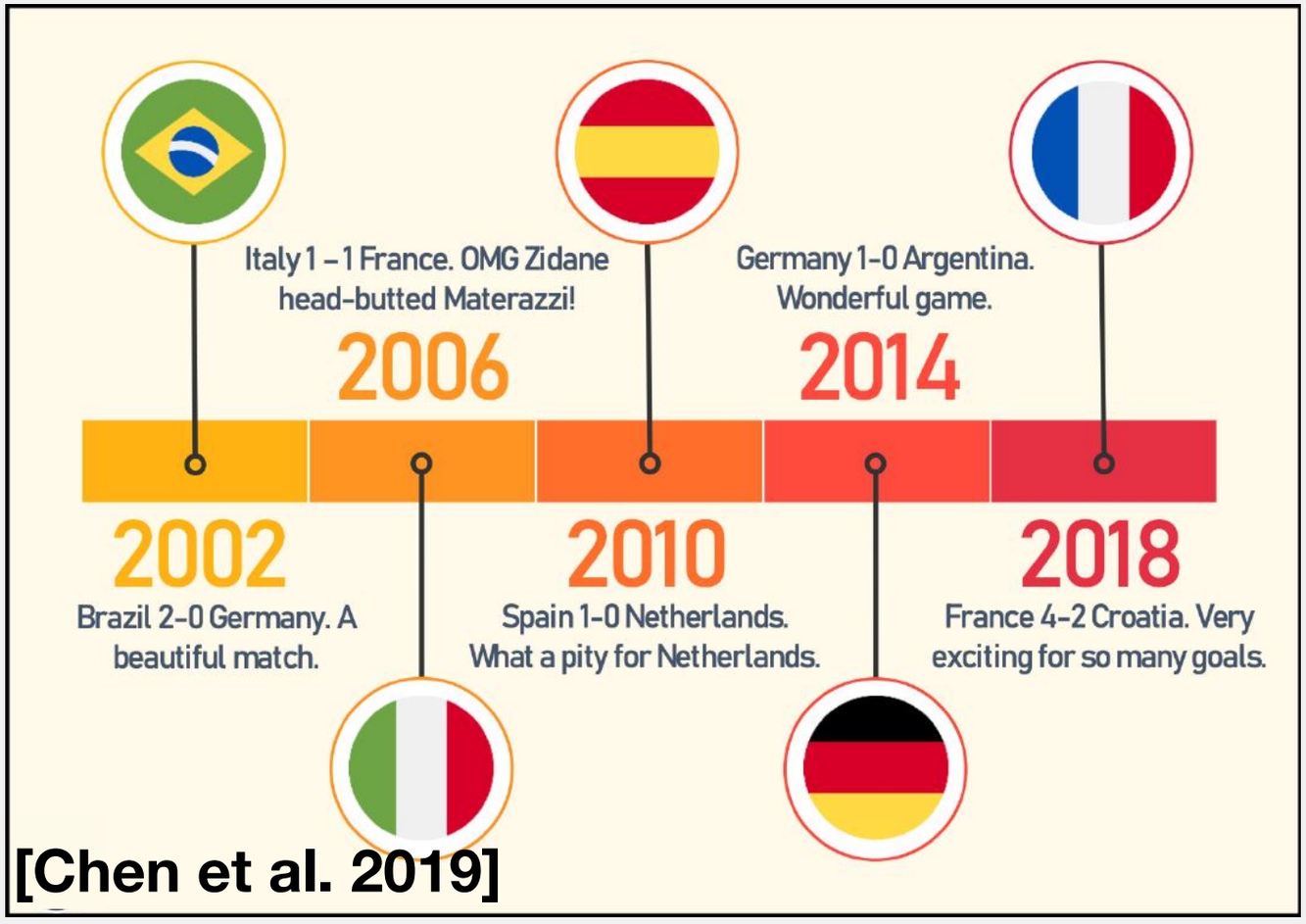


Assessing User Behavior



Visual Analytics vs … Visualization?

- Interactivity is key to visual analytics
  - We are trying to facilitate analytical reasoning, which is an **active process**; so our visualizations should be **responsive**
- In contrast: infographics



[Chen et al. 2019]

Developing effective visual analytics systems …..

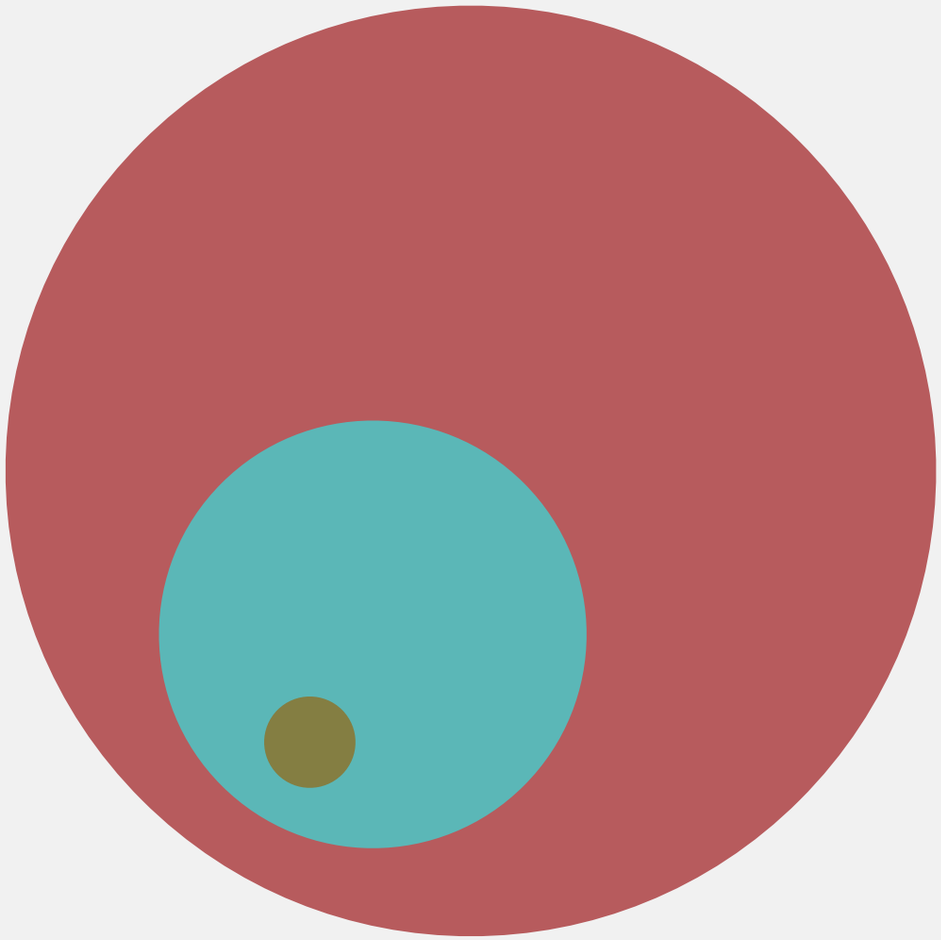
- … requires understanding good visualization practices 😀

- Necessary to approach visualization development as **design:**

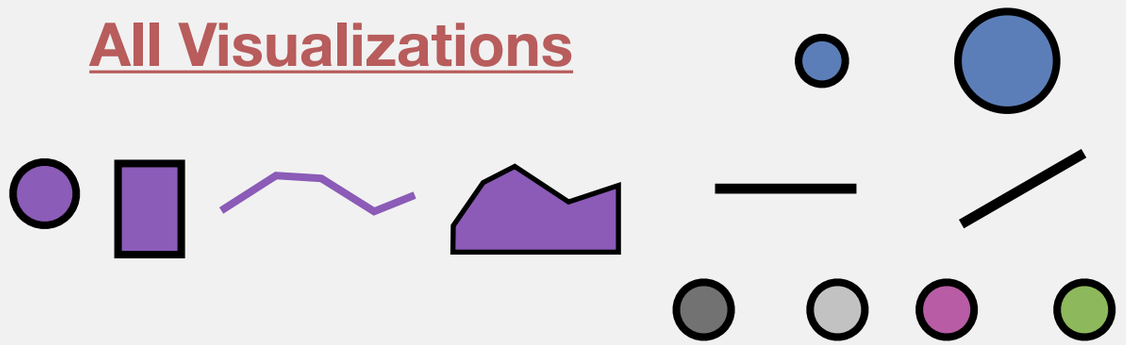
“… the creative process of searching through a vast space of possibilities to select one of many possible good choices from the backdrop of the far larger set of bad choices.”

[Sedlmair et al. 2012]

Design Space



All Visualizations



Suitable Visualizations for Problem

Good Visualizations for Problem

- So, what makes a visualization good for a given problem?
- **Visualization design** should best support **analytical activity**.

Example (4)

Title	Genre	Rating	Budget	Revenue	Review
Frank	comedy	4.3	1M	2M	“this was funny”
Frank	comedy	4.3	1M	2M	“this was bonkers”
Zodiac	drama	3.7	37M	48M	“this movie was dark”
Zodiac	drama	3.7	37M	48M	“it was dull”

- Question of interest: *how do user opinions differ between comedies and dramas?*

- What steps would you take to answer this question?

Analytic Activity

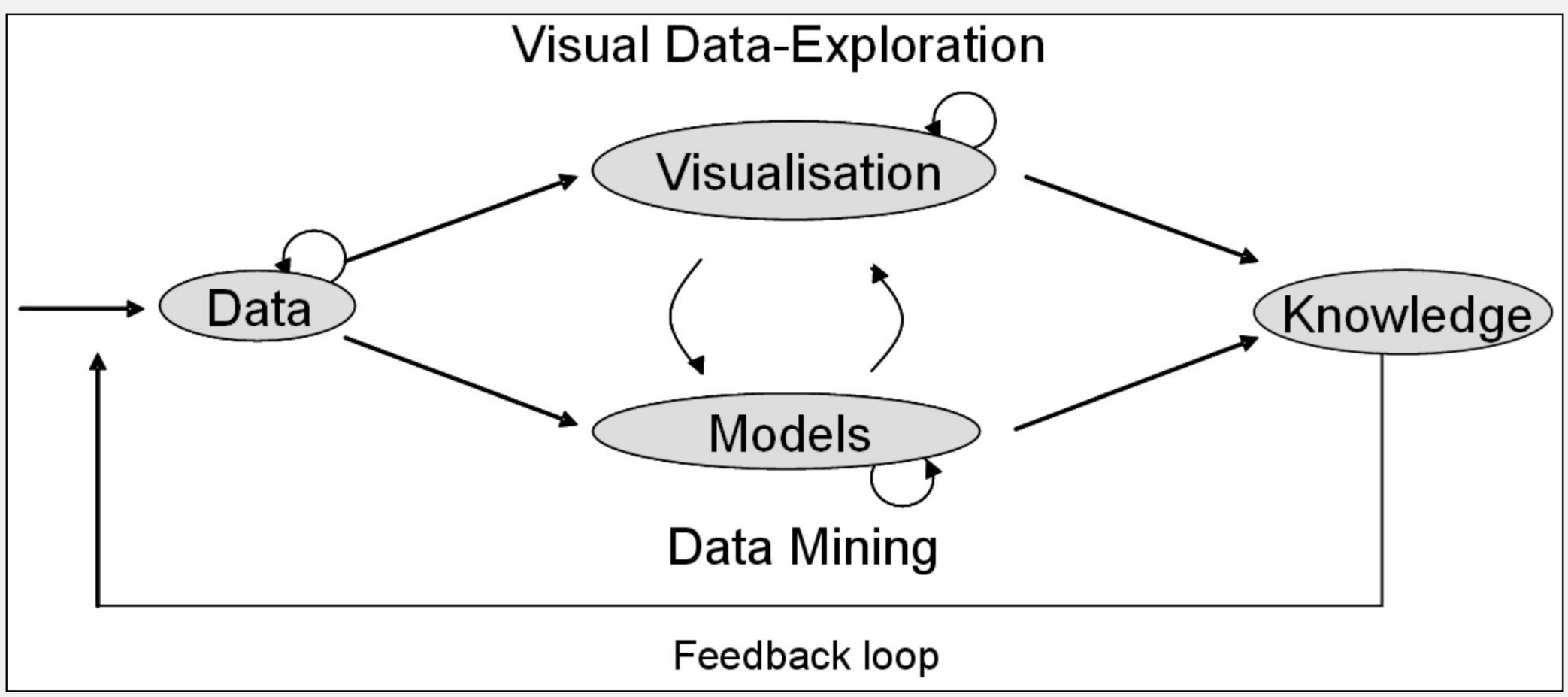
Title	Genre	Rating	Budget	Revenue	Review
Frank	comedy	4.3	1M	2M	“this was funny”
Frank	comedy	4.3	1M	2M	“this was bonkers”
Zodiac	drama	3.7	37M	48M	“this movie was dark”
Zodiac	drama	3.7	37M	48M	“it was dull”

- Data is no longer structured into clean fields
- Necessary to build models, and have the user *visually interact* with **models**
- Analytic tasks need to transfer to analyzing models

Visual Analytics: Another View

- Combine **automated analysis** with **interactive visualizations** to facilitate analytical reasoning for large, complex datasets.

[Keim et al. 2008]



Visual Analytics and Machine Learning

- Course focus: current research that brings together visual analytics with machine learning

- Emphasis on complex datasets: large number of data items, high-dimensional, e.g. images, text

- **Data**: raw, derived from model, model itself, provided by human

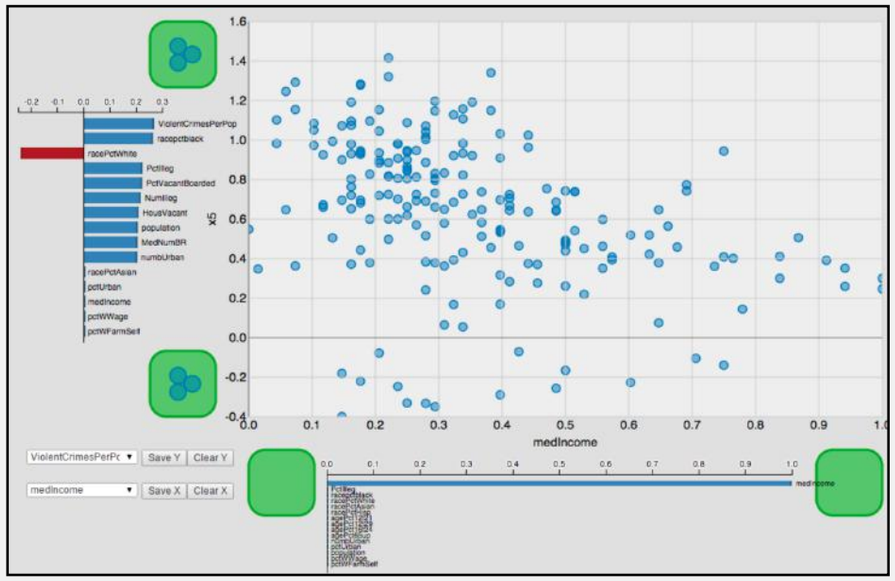
- **Visualization design**: spatial organization, visual encodings of data/model, interactions that handle user interfaces, direct manipulation, model interactions

Mixed-Initiative Visual Exploration

- Interactive visualization of **data**, and **models of data**
  - Emphasis on unsupervised learning
  - What aspects of model are useful for humans to analyze data?
  - Traditional visualization design, coupled with **model steering**: user manipulates model via visual interfaces, visualization changes in response

Examples

Dimensionality Reduction

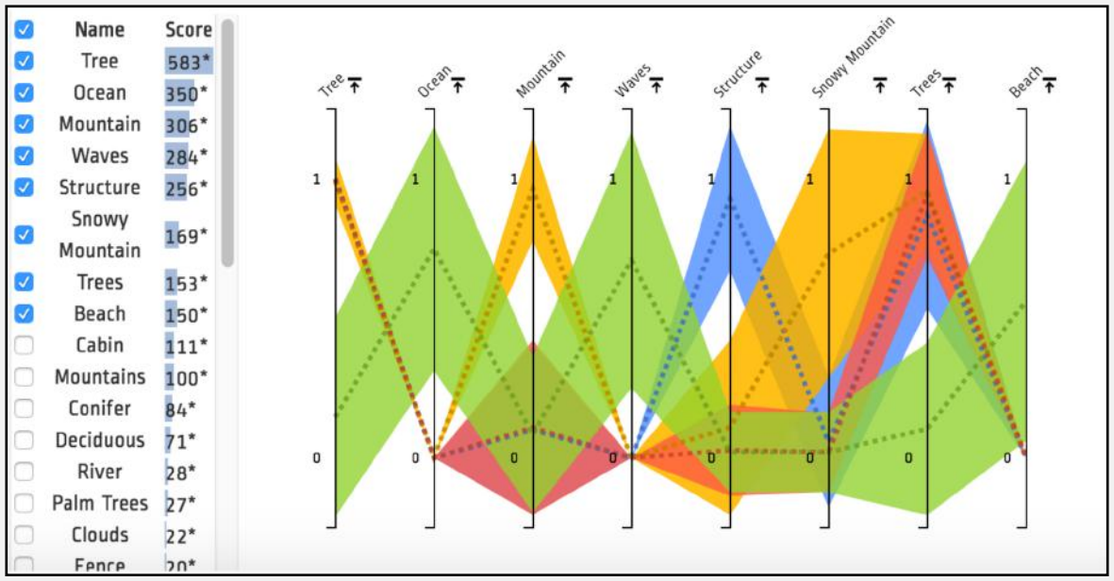


[Kim et al. 2015]

Topic Modeling

[Alexander et al. 2014]

Clustering



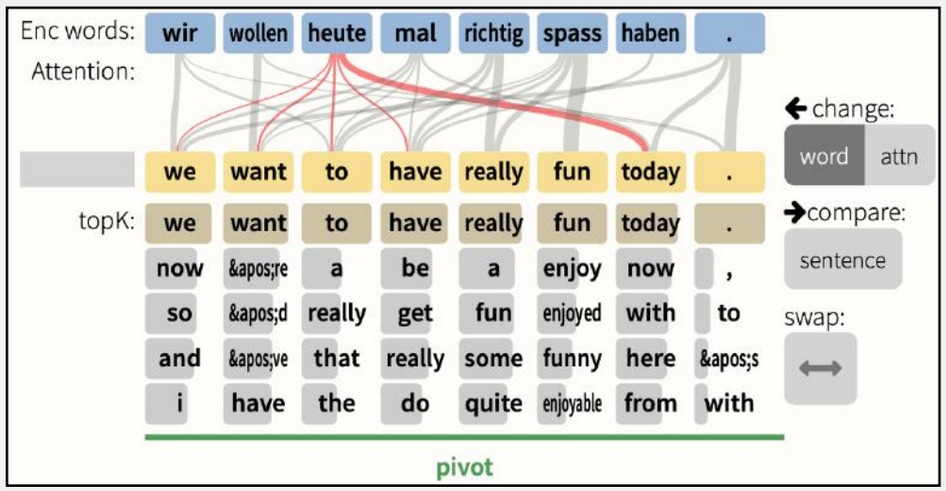
[Kwon et al. 2017]

Model Understanding

- The machine learning model is the object of study
- Visual analytics used to help understand the model
  - “Why did training fail?”
  - “What features did a model learn?”
  - “How can we characterize model generalization?”
- Must sufficiently understand a model to extract relevant information for visualization

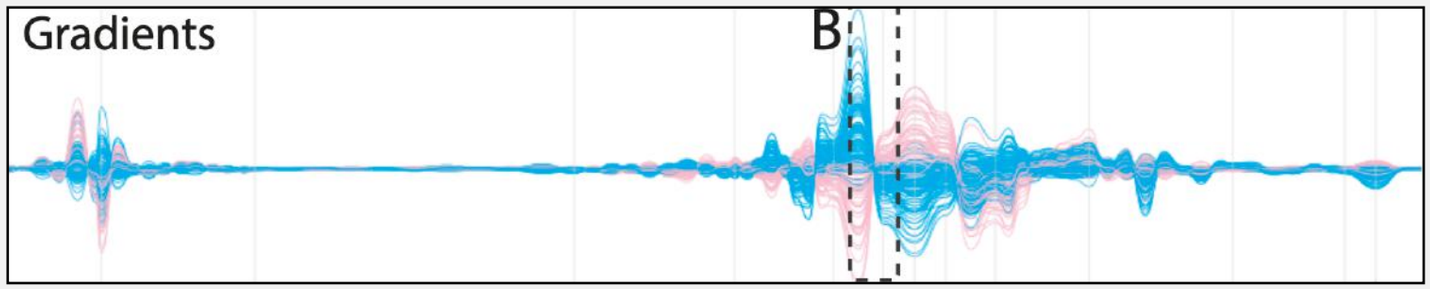
Examples

Model Predictions



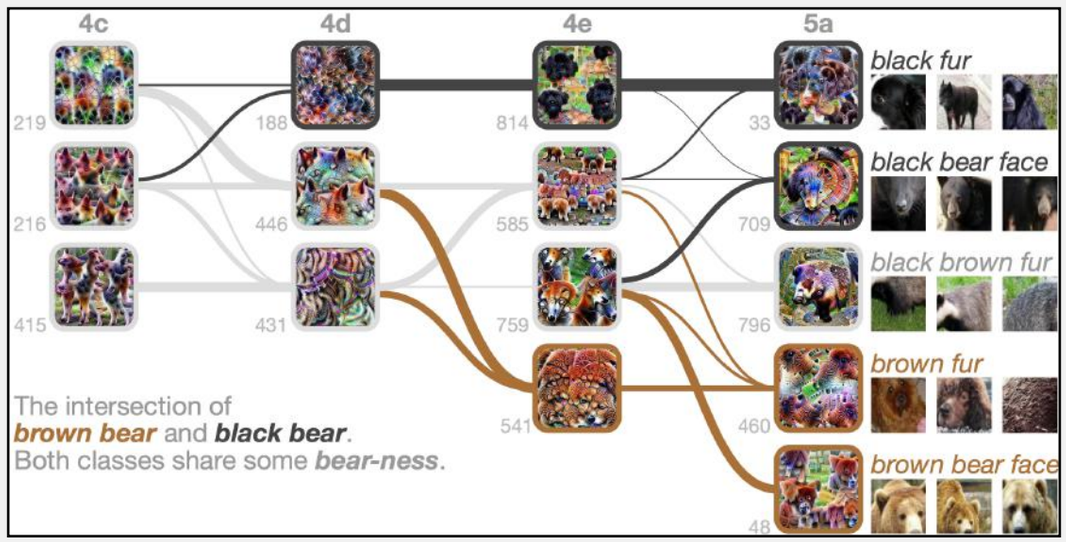
[Strobbelt et al. 2018]

Model Training



[Liu et al. 2017]

Model Features



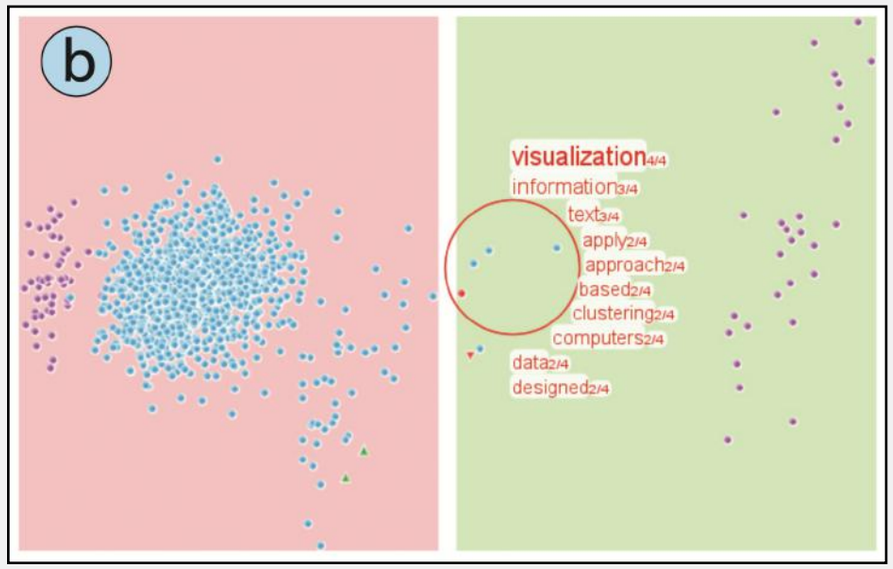
[Hohman et al. 2019]

Model Training

- More direct objective: minimize human time and effort in annotating data/model for training
- Visualization design: data, current state of model
- Involves active learning: what *should* a human annotate?
- Model steering: ensembles, neural architectures
- User interactions: how does a user annotate data? steer model?

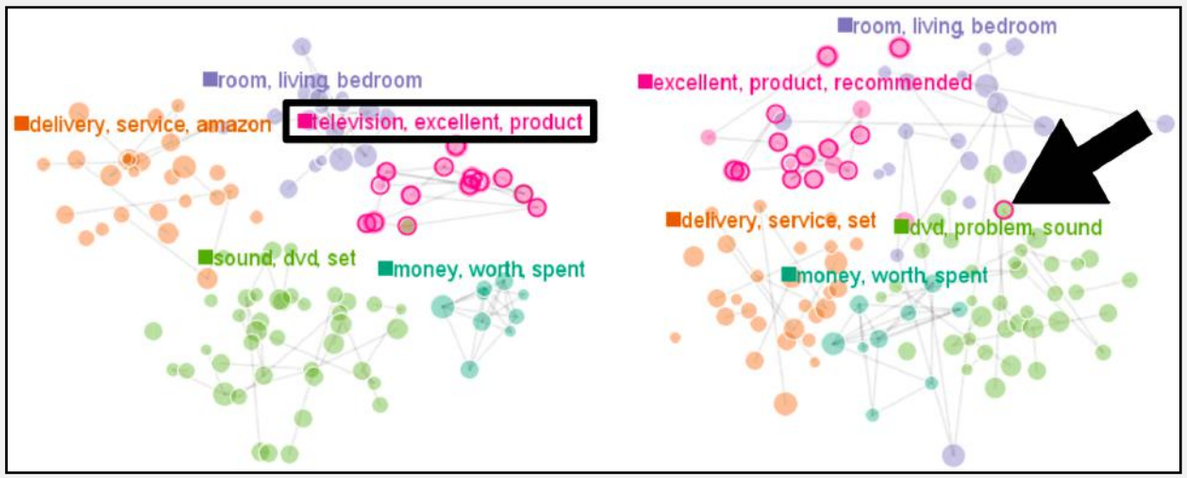
Examples

Interactive Training



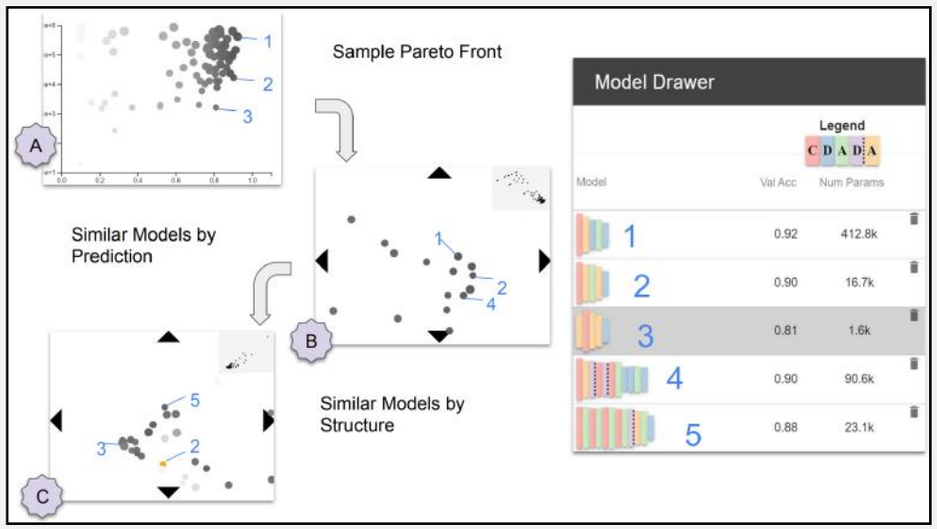
[Heimerl et al. 2012]

Steering Topic Models



[Choo et al. 2013]

Steering Model Ensembles



[Cashman et al. 2019]

Learning for Vis

- Machine learning used as a **tool** for the visualization designer

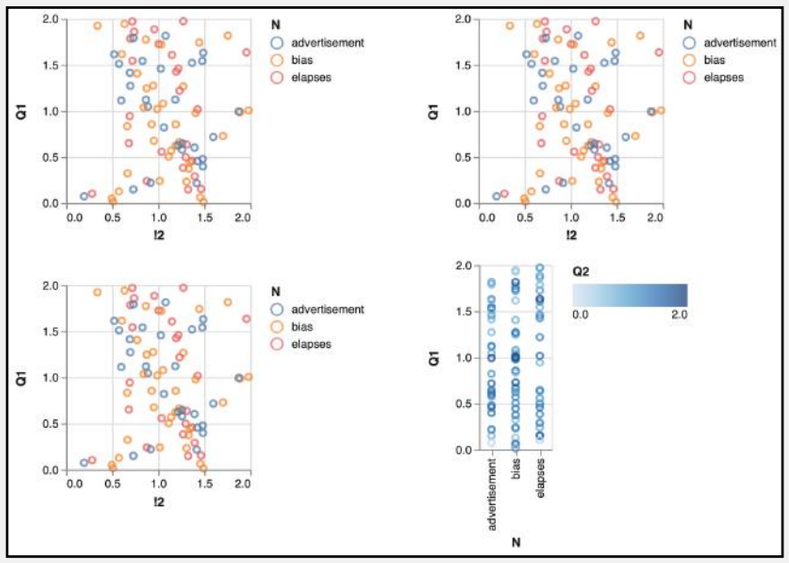
- Formulate visualization design as problems that we may optimize

- Nascent field, so we will place less emphasis on this topic

    - Challenges compared to traditional machine learning formulations in terms of datasets, benefits, etc…

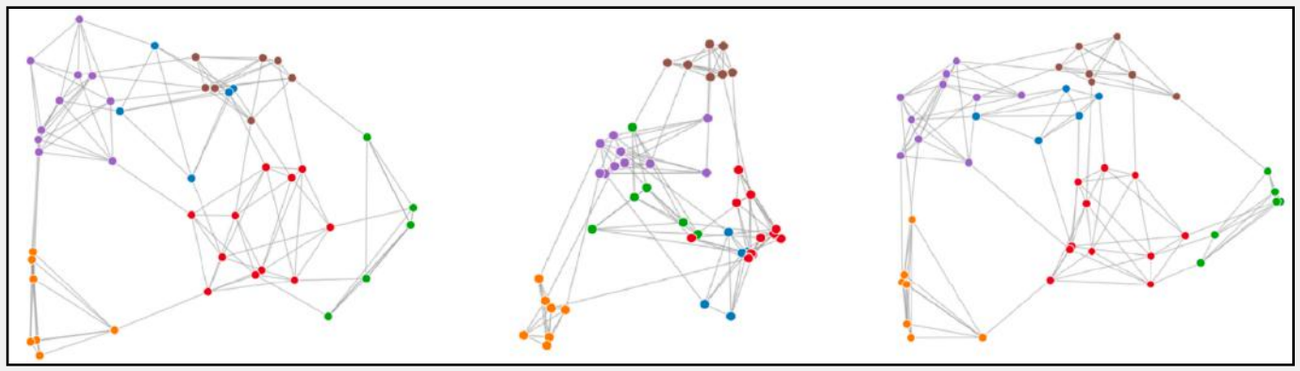
Examples

Visualization Recommendation



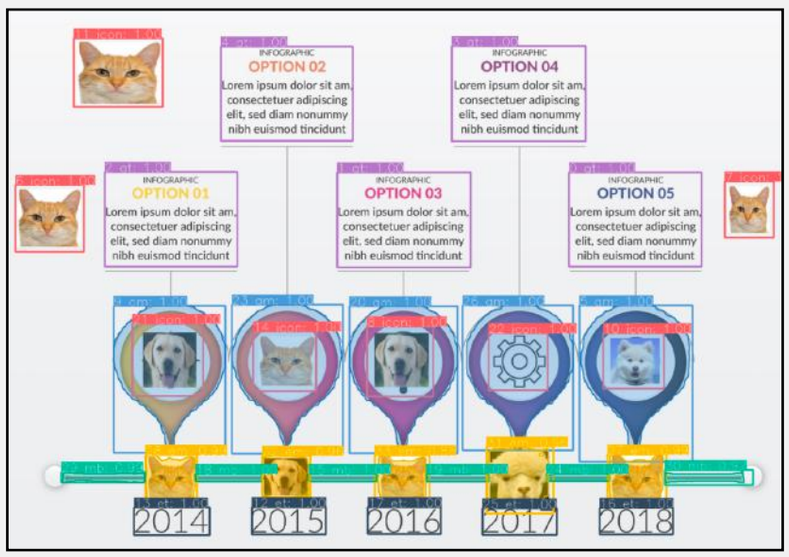
[Moritz et al. 2018]

Learning Graph Layouts



[Yang et al. 2019]

Infographic Generation



[Chen et al. 2019]

Lecture Format

- We will, predominantly, cover research papers throughout the semester.

- Emphasis on: analytical tasks (high/low-level), data, visualization design, how the design addresses the stated problems.

- Papers posted on course website prior to lecture: please read papers in advance of lecture.

- Notable exception to this format: 1st two weeks of class …

Programming

- We will use Javascript for development. Specifically, the main library we will use is D3. We will also use a bit of Vega-Lite.

- D3 requires an understanding of the Document Object Model (DOM), Scalable Vector Graphics (SVG), we will cover the basics as part of the course.

- Environment for development: Observable notebooks

    - Geared towards developing visualizations that are responsive.

https://observablehq.com/d/85073fc618044677

Programming Assignments

- First part of the semester: three programming assignments, each worth 10% of grade.
- Intention of assignments:
  - Visualization design is hard.
  - The best way to improve on design is practice.
  - Assignments will provide you with practice.

Paper Presentation

- Each student will be expected to present a research paper at some point in the semester, worth 10% of grade.

- You will choose a paper from the “Papers” portion of the course webpage.

  - (may choose a paper not listed, but I will need to approve it)

- Presentation: describe the problem being addressed, the approach taken, a design critique, and alternative visualization designs, e.g. how else *could* the paper have addressed the problem?

Project

- Second half of the semester: form a team of 2 to work on a project (50% of grade)

- Prepare a project proposal describing problem, data, tasks, etc.. and present this proposal to the class.

- A prototype of the project will be due about halfway through the project.

- End of the semester: final presentation, and project submission.

- Should use Observable notebook to handle most aspects of project - if this will not work for you, then please let me know.

Prerequisites (1)

- Background in machine learning
  - Unsupervised learning (clustering, dimensionality reduction)
  - Supervised learning (classification, regression)
  - Basic understanding of deep learning
  - Basics of optimization (e.g. gradient descent)
- We will cover machine learning material as necessary for visual analytics

Prerequisites (2)

- Linear algebra
    - Comfortable with matrix notation.
    - Familiarity with basics: matrix inversion, eigendecomposition, singular value decomposition.
    - Practice using these in the context of machine learning techniques.

Prerequisites (3)

- Visualization: minimal background expected. We will cover basics as part of the course.

- Good to have familiarity with some visualization tool(s): matplotlib, ggplot2, Tableau, etc..

- Javascript will be the main programming language used for assignments/projects.

  - Not necessary to have Javascript knowledge, if you are skilled in Python, transitioning to Javascript is not too bad.

  - Please see resources section on course website.

Visual Analytics and Machine Learning

- Any questions on the structure of the course?