# CS-GY-9223
# Visualization: Connections with Machine Learning

**Course materials based on CS 8395-03 Visual Analytics & Machine Learning**
**By Matt Berger, Vanderbilt University**
**https://matthewberger.github.io/teaching/vaml/spring2019/**

# Agenda

- What is Visual Analytics?
- Visual Analytics & Machine Learning
- Course Logistics

# Visual Analytics

- A combination of **analytic techniques** and **interactive data visualization** to help people **make sense of data**.

- Analytic techniques?

- Interactive Data Visualization?

- Sensemaking?

# Data Analytics

- I am given some dataset, and I want to understand something about it. What do I do?

- Well, I can stare at a table of numbers.

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| symboling | normalized_l | make | fuel_type | aspiration | num_of_doo | body_style | drive_wheels | engine_locat | wheel_base | length | width | height |
| 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | 168.8 | 64.1 | 48.8 |
| 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | 168.8 | 64.1 | 48.8 |
| 1 | ? | alfa-romero | gas | std | two | hatchback | rwd | front | 94.5 | 171.2 | 65.5 | 52.4 |
| 2 | 164 | audi | gas | std | four | sedan | fwd | front | 99.8 | 176.6 | 66.2 | 54.3 |
| 2 | 164 | audi | gas | std | four | sedan | 4wd | front | 99.4 | 176.6 | 66.4 | 54.3 |
| 2 | ? | audi | gas | std | two | sedan | fwd | front | 99.8 | 177.3 | 66.3 | 53.1 |
| 1 | 158 | audi | gas | std | four | sedan | fwd | front | 105.8 | 192.7 | 71.4 | 55.7 |
| 1 | ? | audi | gas | std | four | wagon | fwd | front | 105.8 | 192.7 | 71.4 | 55.7 |
| 1 | 158 | audi | gas | turbo | four | sedan | fwd | front | 105.8 | 192.7 | 71.4 | 55.9 |
| 0 | ? | audi | gas | turbo | two | hatchback | 4wd | front | 99.5 | 178.2 | 67.9 | 52 |
| 2 | 192 | bmw | gas | std | two | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 |
| 0 | 192 | bmw | gas | std | four | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 |
| 0 | 188 | bmw | gas | std | two | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 |
| 0 | 188 | bmw | gas | std | four | sedan | rwd | front | 101.2 | 176.8 | 64.8 | 54.3 |
| 1 | ? | bmw | gas | std | four | sedan | rwd | front | 103.5 | 189 | 66.9 | 55.7 |
| 0 | ? | bmw | gas | std | four | sedan | rwd | front | 103.5 | 189 | 66.9 | 55.7 |
| 0 | ? | bmw | gas | std | two | sedan | rwd | front | 103.5 | 193.8 | 67.9 | 53.7 |
| 0 | ? | bmw | gas | std | four | sedan | rwd | front | 110 | 197 | 70.9 | 56.3 |
| 2 | 121 | chevrolet | gas | std | two | hatchback | fwd | front | 88.4 | 141.1 | 60.3 | 53.2 |
| 1 | 98 | chevrolet | gas | std | two | hatchback | fwd | front | 94.5 | 155.9 | 63.6 | 52 |
| 0 | 81 | chevrolet | gas | std | four | sedan | fwd | front | 94.5 | 158.8 | 63.6 | 52 |
| 1 | 118 | dodge | gas | std | two | hatchback | fwd | front | 93.7 | 157.3 | 63.8 | 50.8 |

# Analytics

- Suppose I wanted to compare front-wheel drive and rear-wheel drive automobiles in terms of their width.

- For each type of “wheel drive”, we have a set of values for width. What would be some ways of analyzing this data?

- A distribution! We could show all of the values, but we could also summarize the distribution: minimum, maximum, upper/lower quartiles, median.

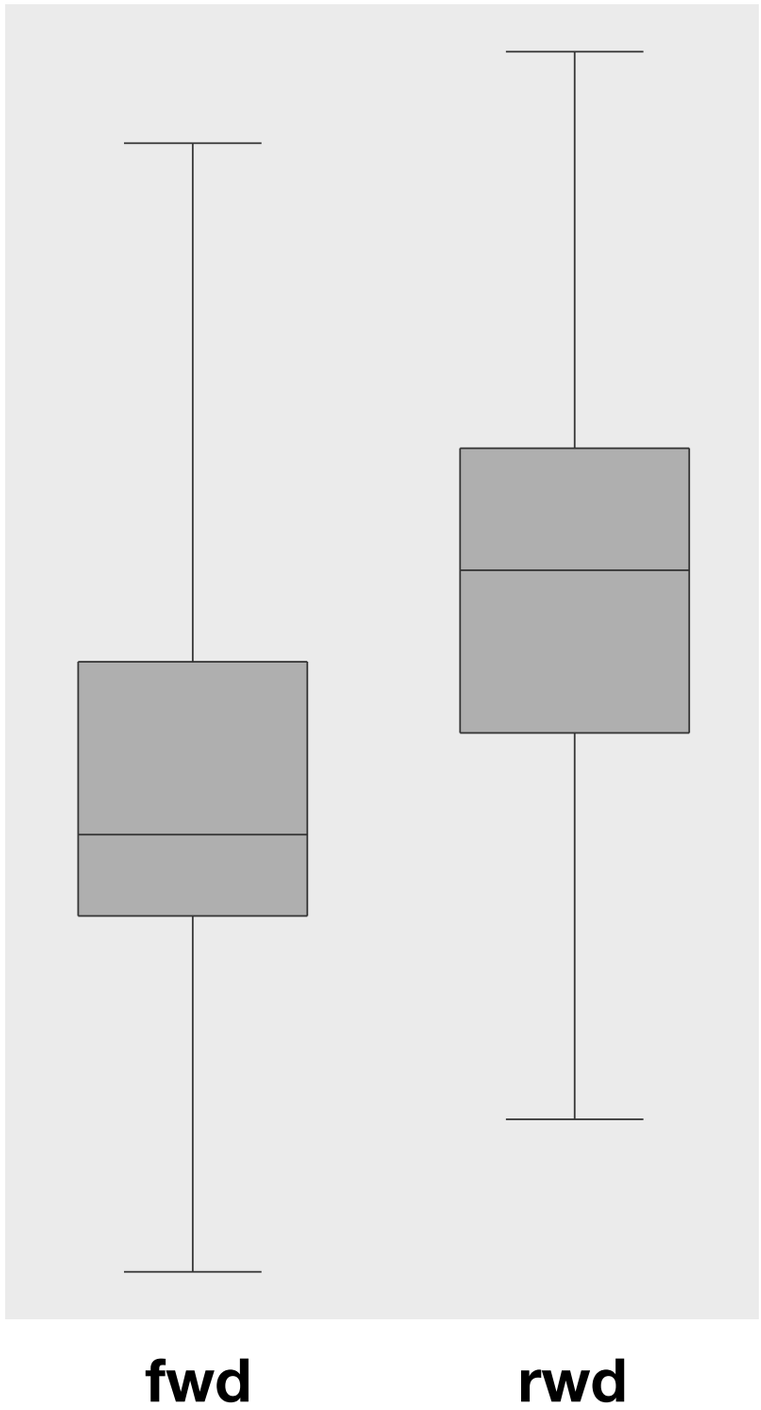
# Distribution of Cars

| | min | lq | med | uq | max |
|---|---|---|---|---|---|
| **fwd** | 60.3 | 63.8 | 64.6 | 66.3 | 71.4 |
| **rwd** | 61.8 | 65.6 | 67.2 | 68.4 | 72.3 |

**If you are staring at a bunch of numbers, plot them!**

**(e.g. data vis)**

# Boxplots



# Visual Analytics

- We visualize:
    - Data we are given
        - Models built on top of data (aforementioned trivial model: coarse statistics)
- Interactions with data *and* models to gain insight
- Interactions?

# Car Similarity

- I want to explore the similarity of individual cars (rather than isolate attributes).

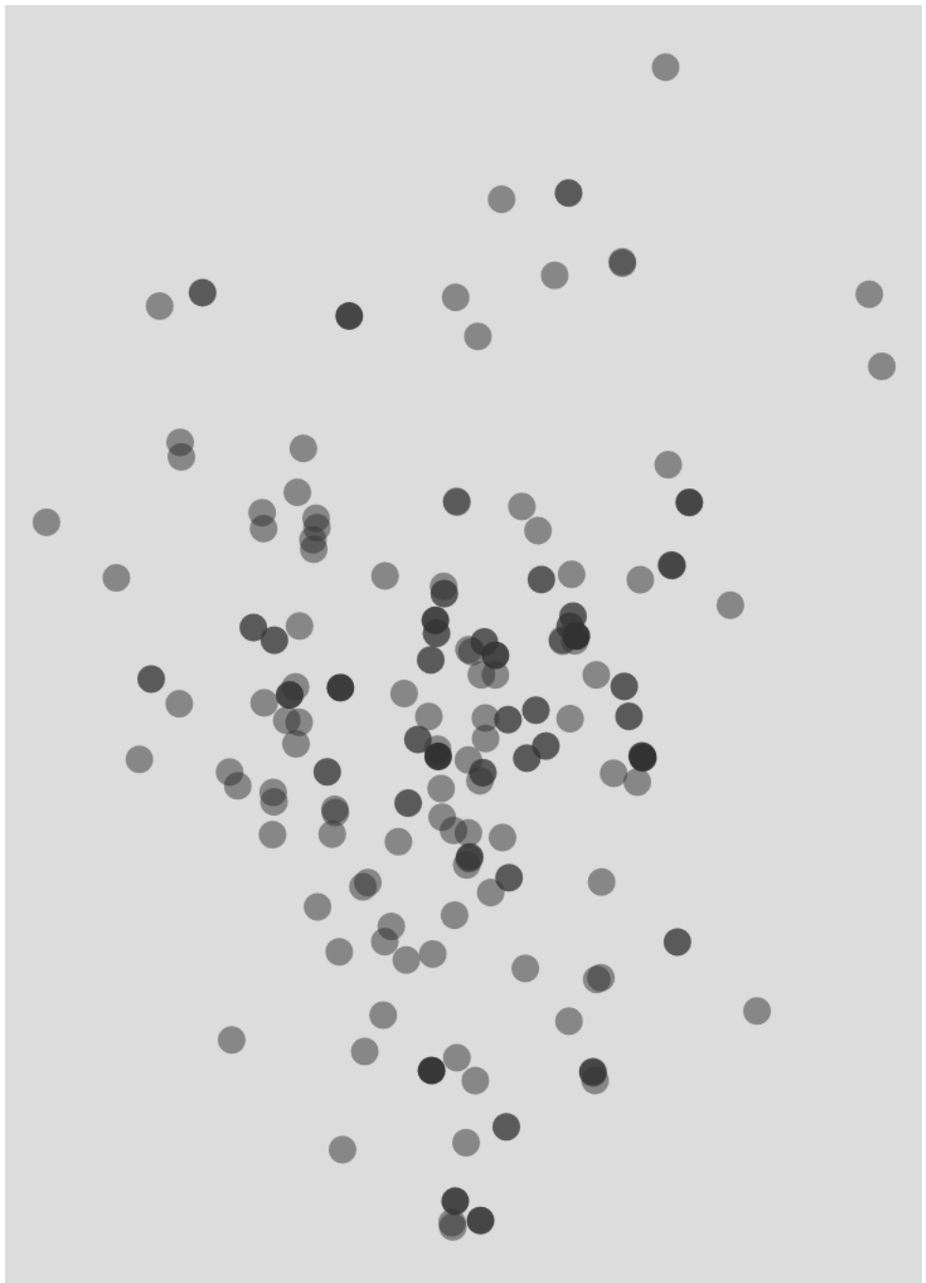
- I will take *wheel size*, *length*, *width*, *height*, *MPG (in city)*, *MPG (on highway)* as my data fields.

- Gives a 6-dimensional space, (moderately) high-dimensional data.

- How do we visualize high-dimensional data?

- Dimensionality reduction! Let's look at PCA.

# PCA



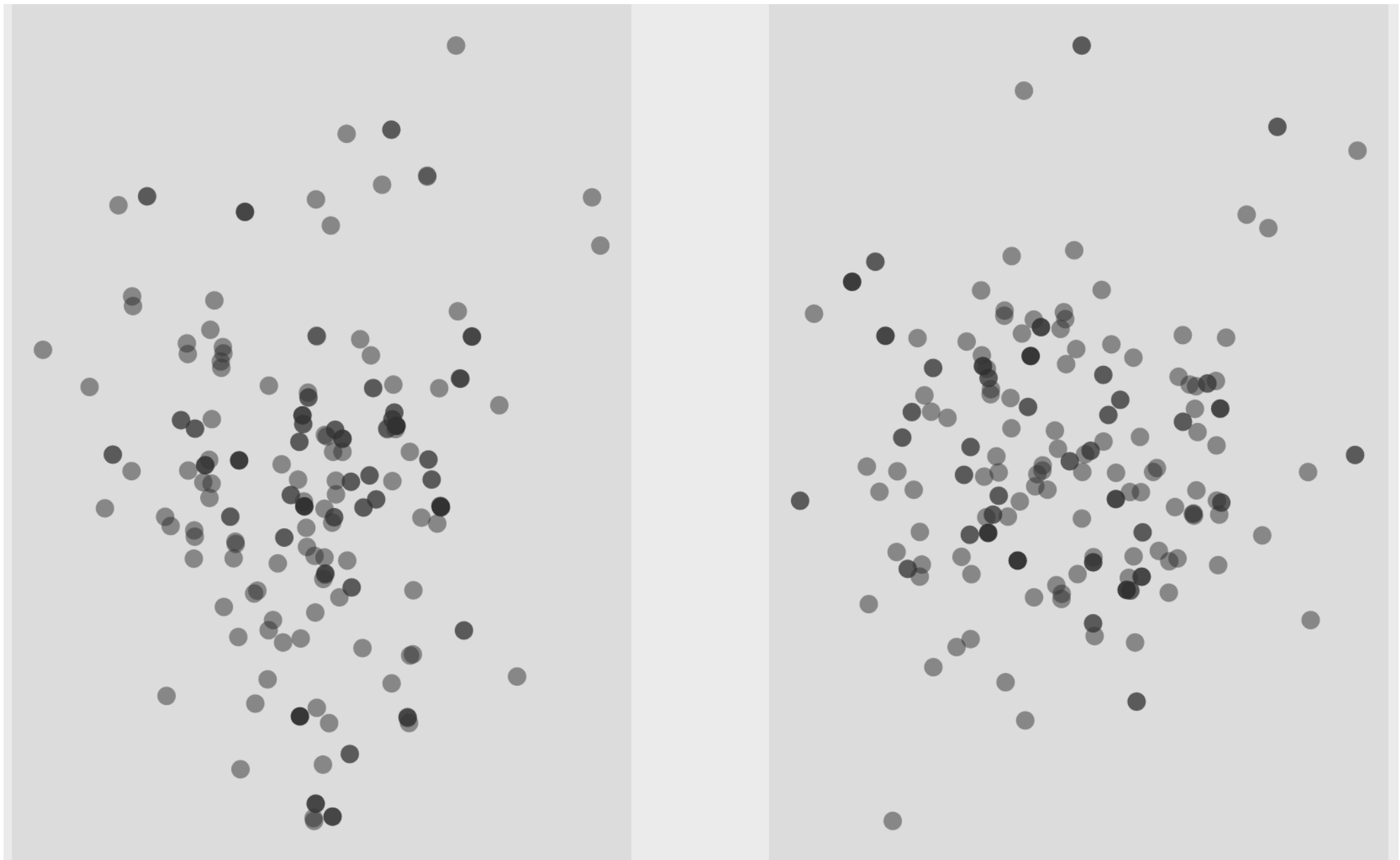
- Points that are close in 2D should (ideally!) be “close” in high-dimensional space.

- Any potential problems you can see here in using this plot for analysis?

- Dimensionality reduction is imperfect! **Visual analytics is all about confronting these imperfections!**

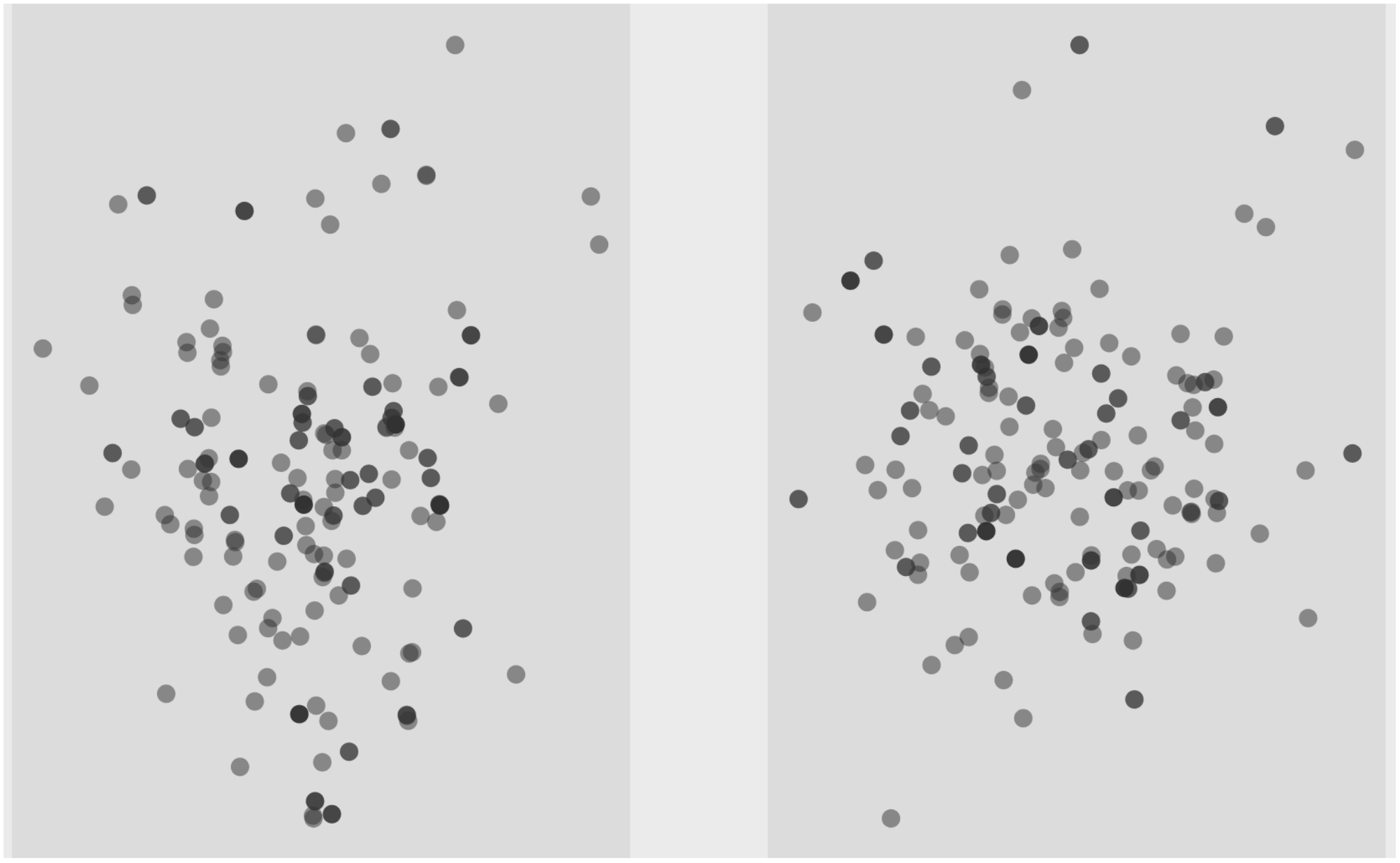
# Exploring PCA a bit

- Why restrict ourselves to the top 2 principal components?



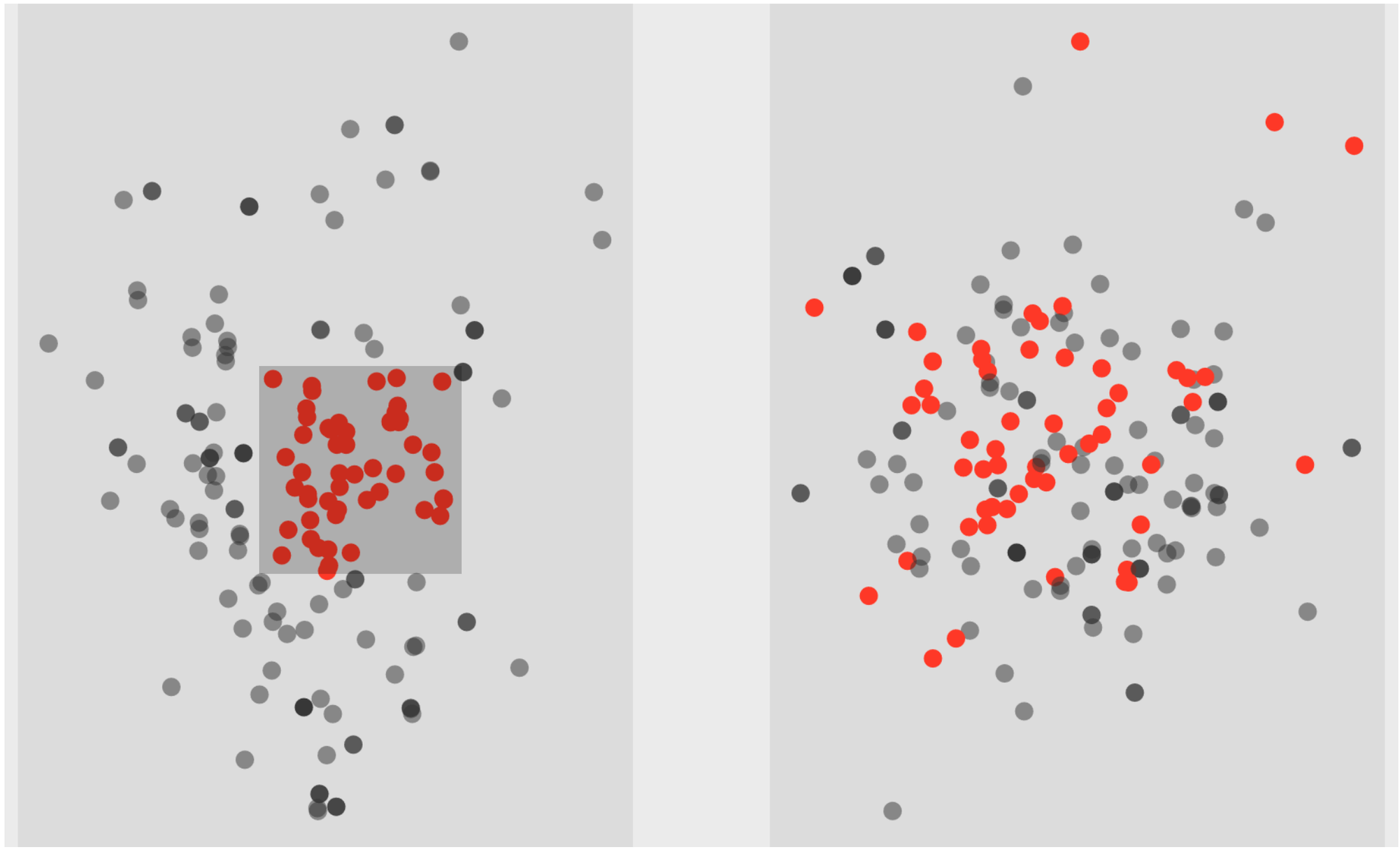
# Interaction with PCA

- Interaction is an essential component to visual analytics



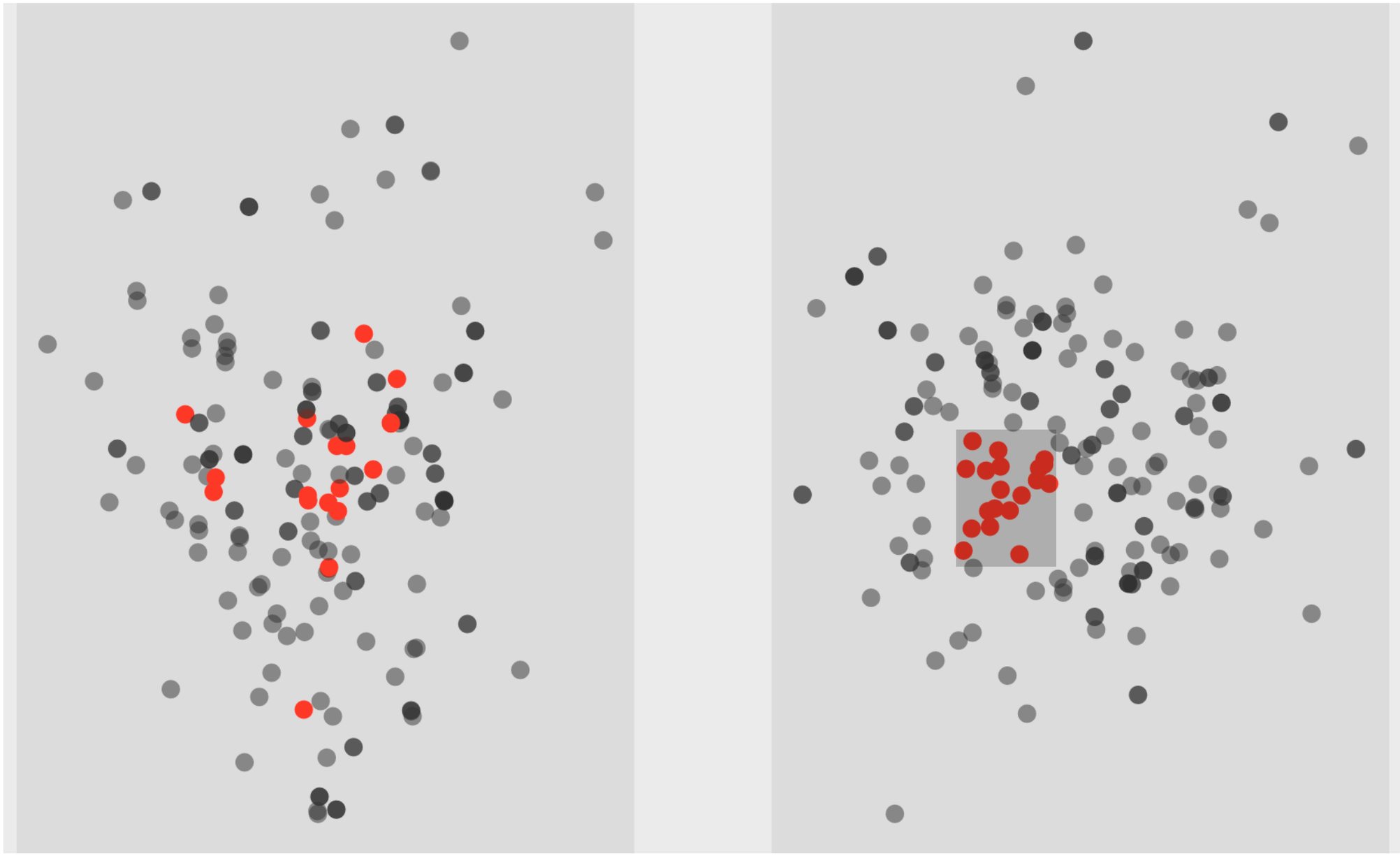
# Interaction with PCA

- Interaction is an essential component to visual analytics



# Interaction with PCA

- Interaction is an essential component to visual analytics



# Understanding Model Parameters

- Let's move away from a simple model like PCA, and instead consider tSNE (t-distributed Stochastic Neighbor Embedding)

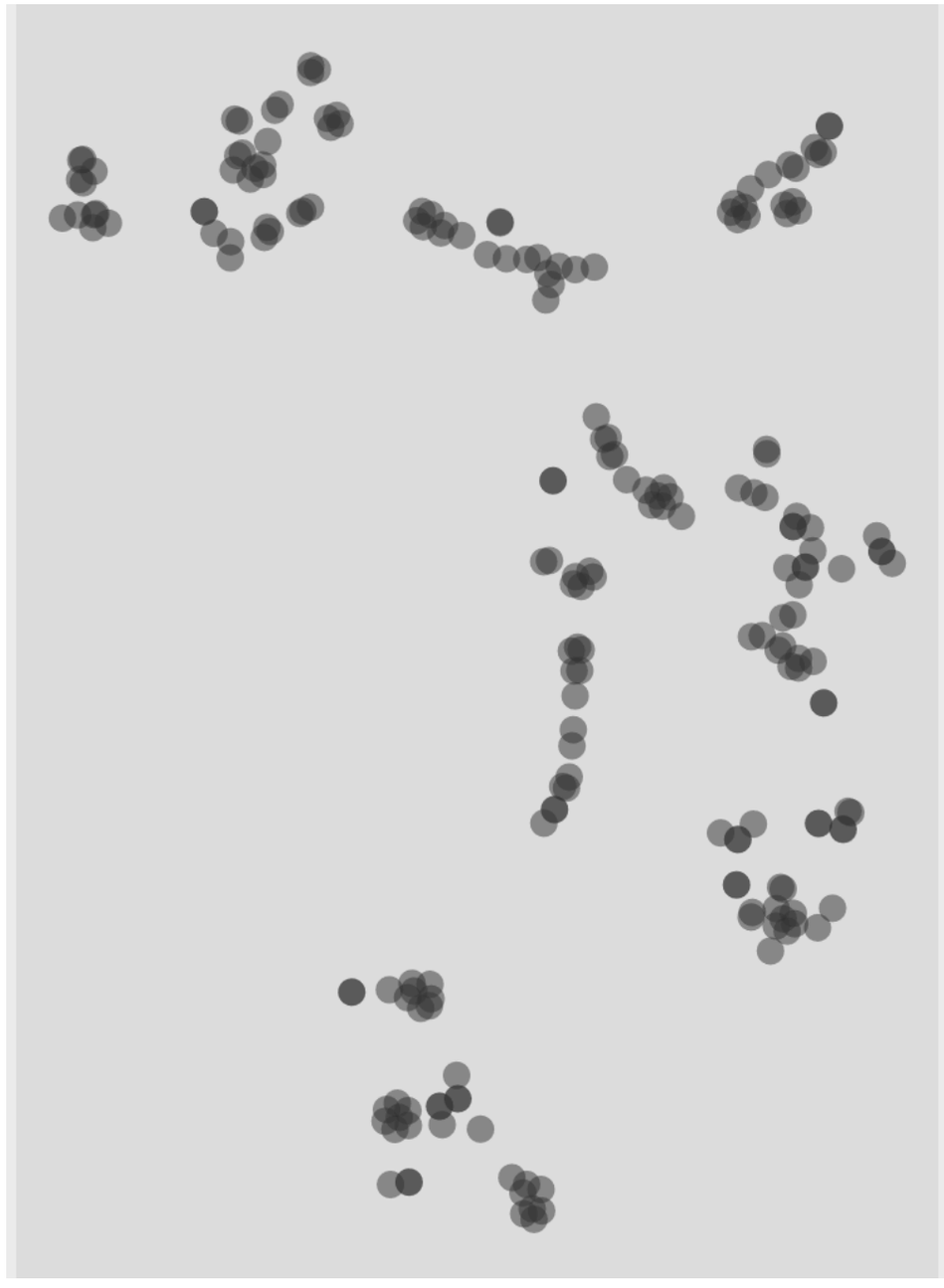
- One key parameter to tSNE: perplexity

- Controls neighborhood influence: larger perplexity, larger neighborhoods

- What should the perplexity be?

- Visual analytics: let the user explore this parameter space in understanding data

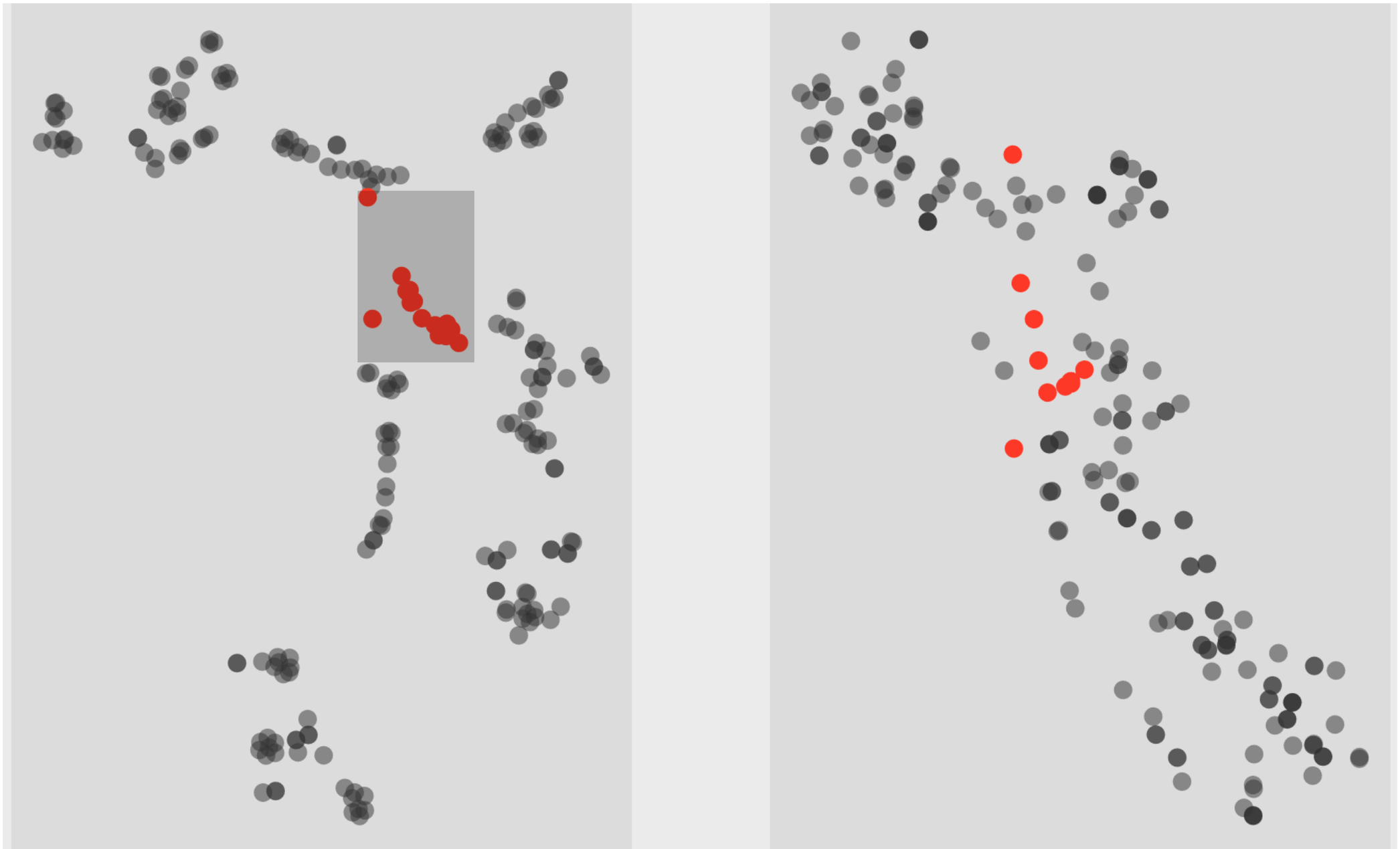
# Interaction with tSNE



perplexity: 10

perplexity: 50

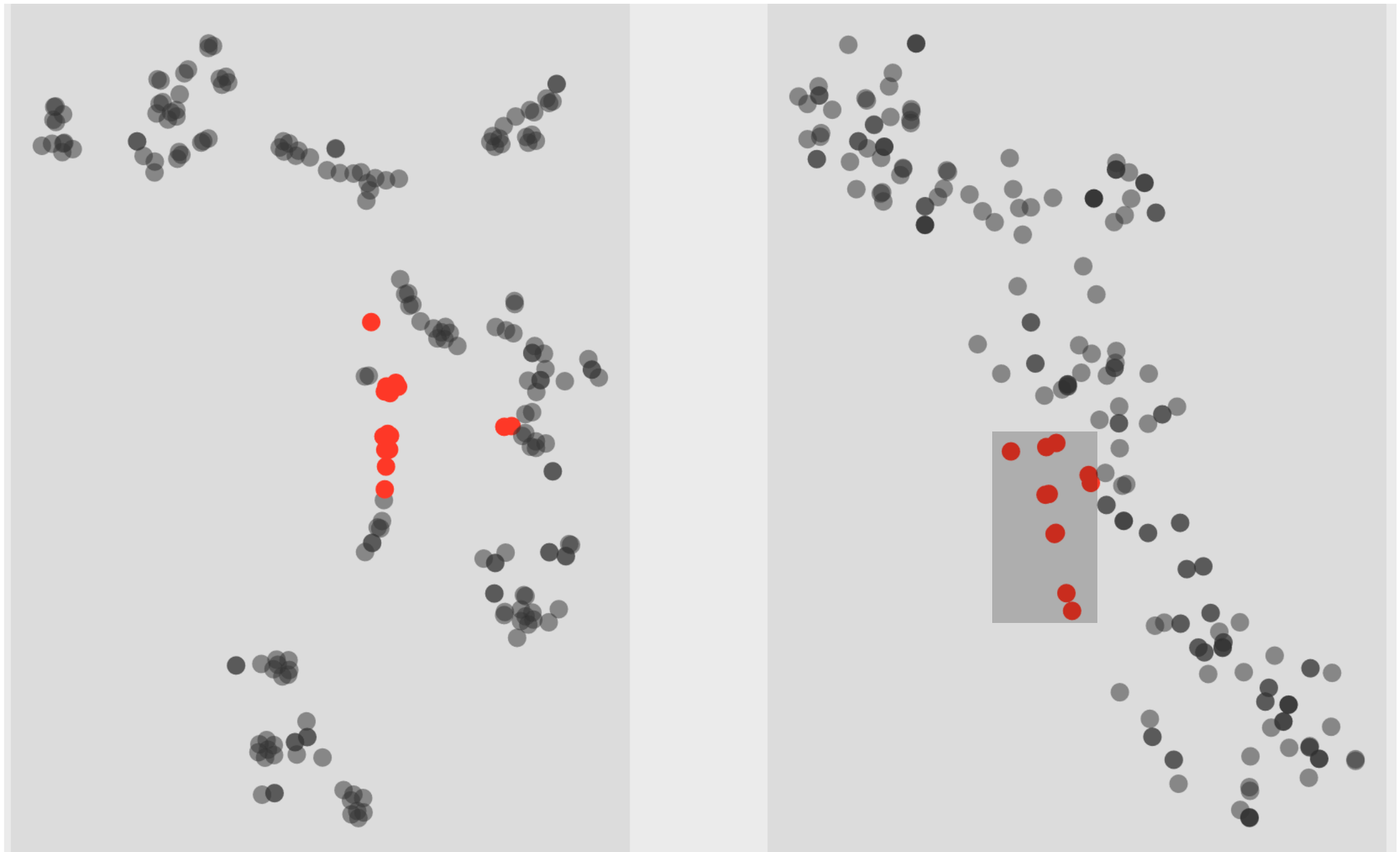
# Interaction with tSNE



perplexity: 10

perplexity: 50

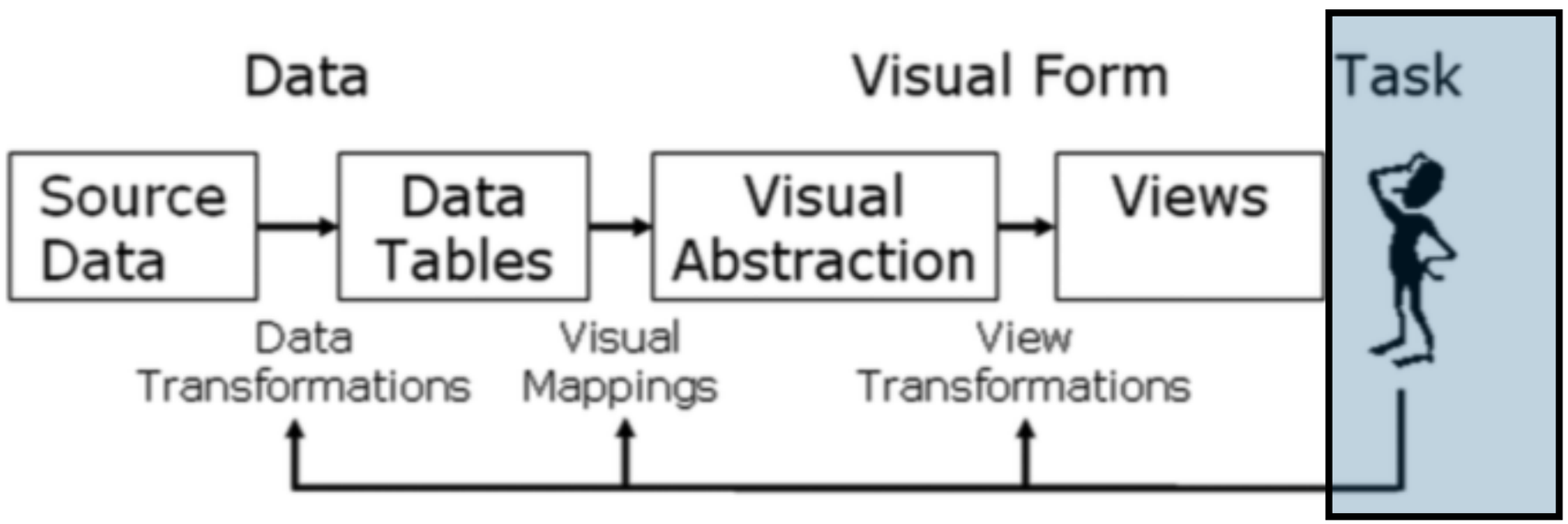
# Interaction with tSNE



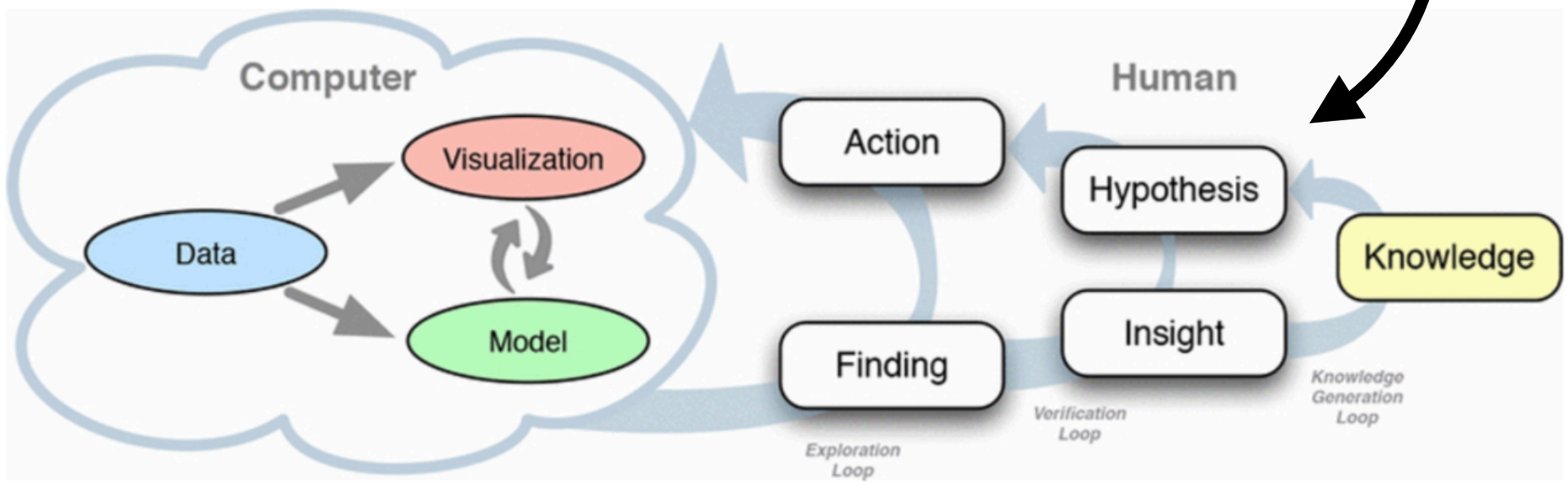
**perplexity: 10**

**perplexity: 50**

# Data Visualization for Insight

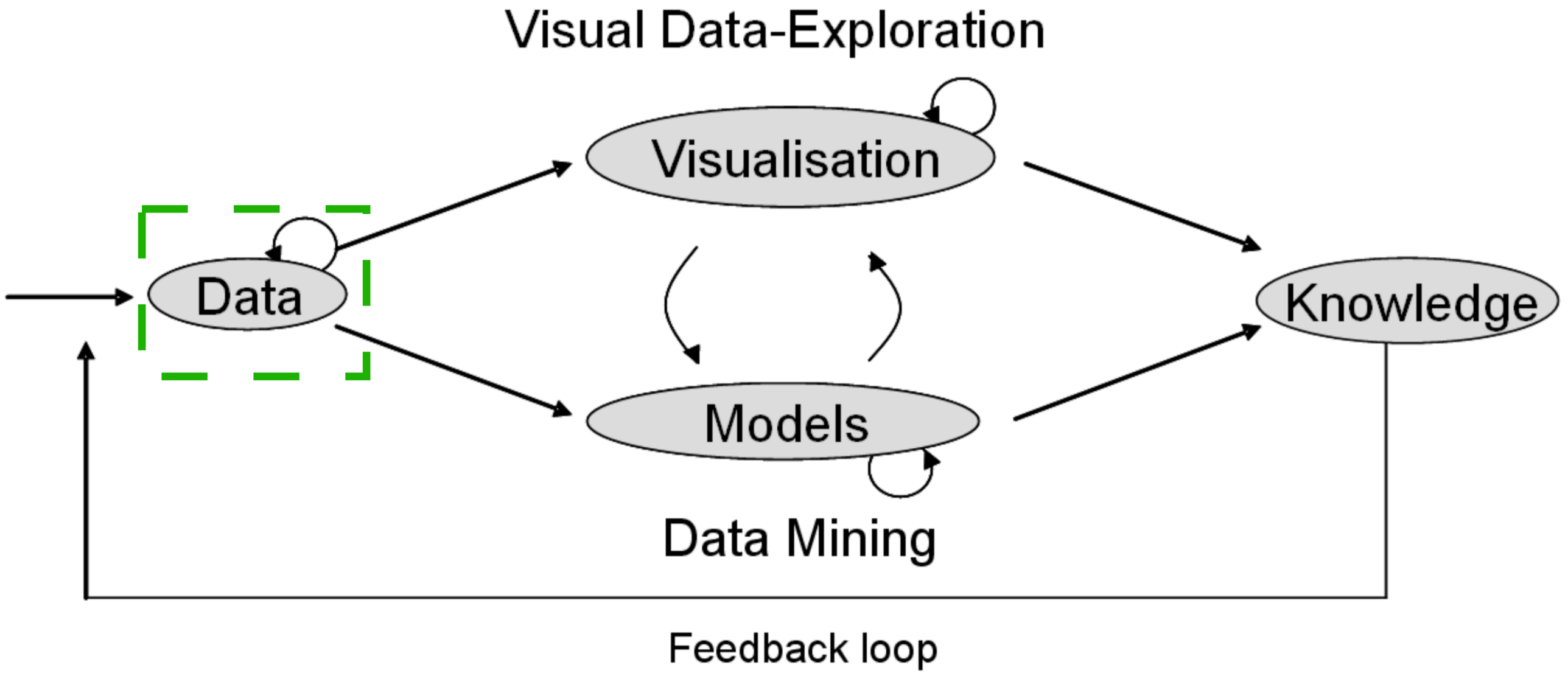



[Heer 2006]




[Sacha et al. 2014]

# Visual Analytics Workflow



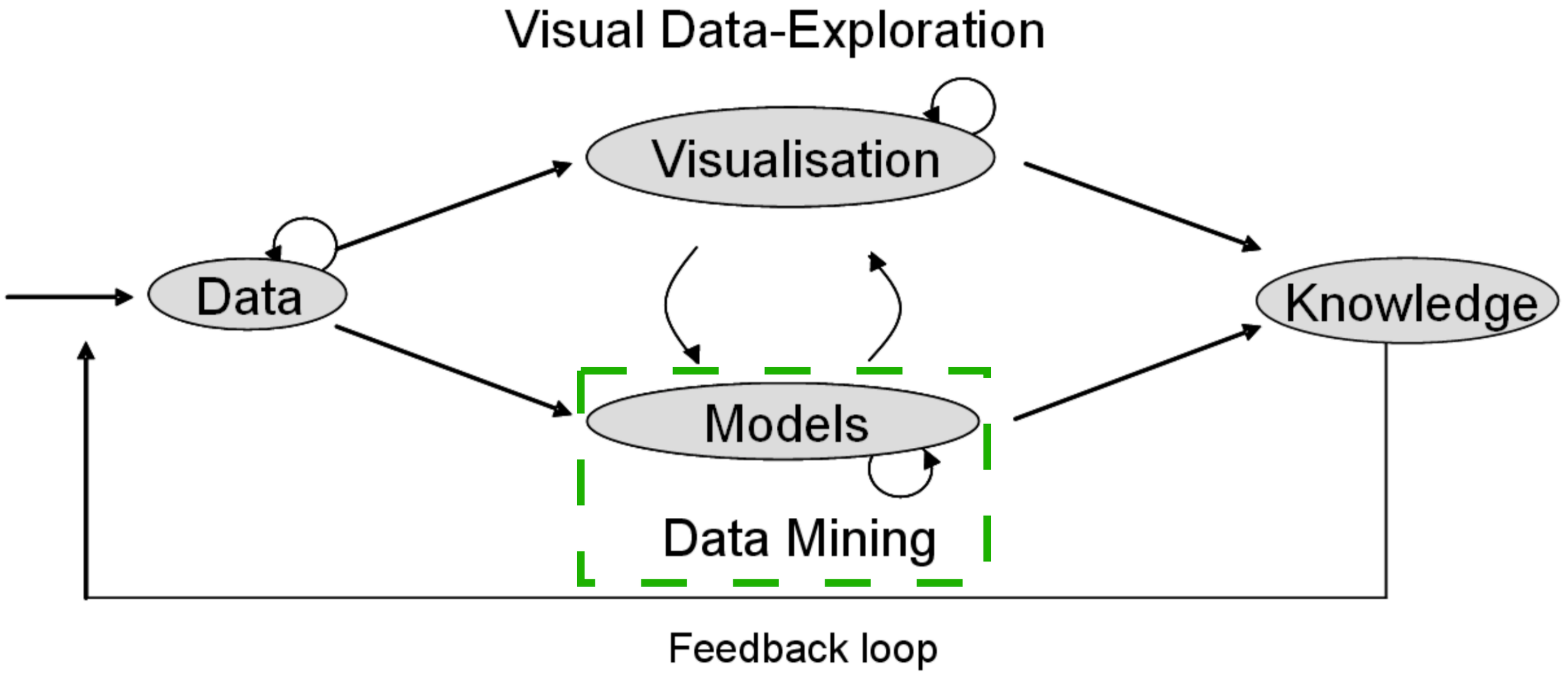

**[Keim et al. 2008]**

# Data

- Data could be (virtually) anything!
    - Images, video, text, networks, trees, etc..

- At this step: often need to perform data cleaning, filtering, aggregation, transformation, etc..

# Visual Analytics Workflow

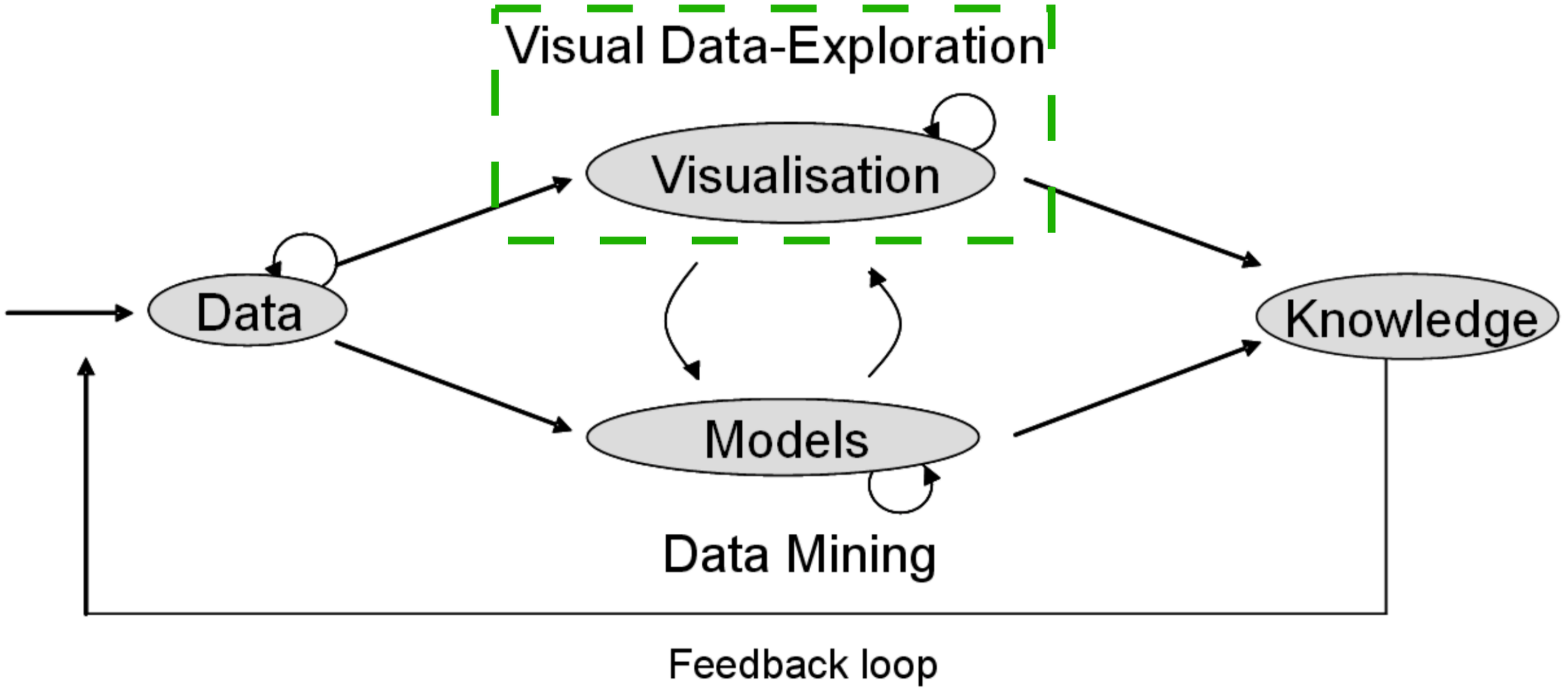



**[Keim et al. 2008]**

# Models

- Model is something built on top of data that can be used for automated data analysis
- Models that we will consider? **<u>Machine learning models</u>**
  - Classification, regression, clustering, generative models, etc…
  - SVMs, decision trees, neural networks, dimensionality reduction, topic models, language models, etc…

# Visual Analytics Workflow




**[Keim et al. 2008]**

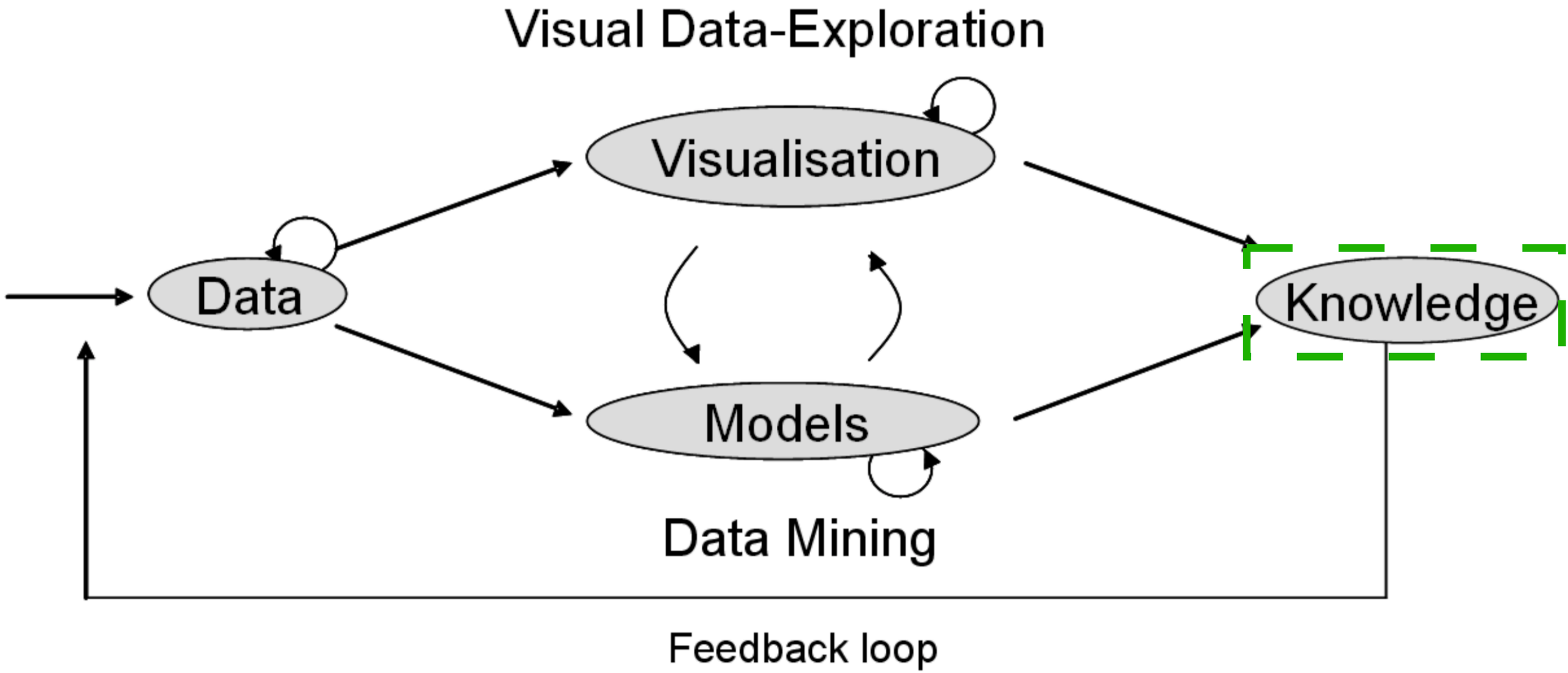
# Visualization

- Visualize **data** and/or **model**.

- Are models alone sufficient for solving the problem under consideration? (often) **no!**
    - Can we use models alongside humans?
    - Suppose we didn’t; can we trust a classifier?
    - How do we go about building models in the first place?

- Visualization plays a key role in tackling these questions.

# Visual Analytics Workflow




**[Keim et al. 2008]**

# Knowledge

- For what reason are we building models and performing visualization?

- Visual analytics often used when the question we want to answer is not crisp: typically cannot be formulated as a machine learning problem

- Loops back into visualization and models:

  - We understand one piece a little better, and then adjust our visualization to understand something else

  - We tune parameters of our model based on our better understanding

# Knowledge for VAML

- What do we hope to gain from combining visualization with machine learning? (e.g. this course?)
- Mixed-Initiative Visual Exploration
- Visual Analytics for Model Understanding
- Visual Analytics for Model Training
- Learning for Visualization

# Mixed-Initiative Visual Exploration

- **Goal**: obtain insight from data

- Blend of automated analyses provided by a machine learning technique, alongside interactive data visualization

- **Challenges**:

  - How do we visually encode both data and model?

  - What should be left to the model? Exposed as user interactions?

- Typical scenarios: dimensionality reduction, clustering, topic modeling, etc.. - *help the user understand data*

# Visual Analytics for Model Understanding

- **Goal**: obtain insight on model

- Interpretability, explainability:
  - Training process, parameters of a model, features learned by a model, outputs produced by a model

- Classification, regression, generative models, etc..

- Typical scenarios: convolutional networks, recurrent networks, generative models, neural language models - *help the user understand why models behave as they do*

# Visual Analytics for Model Training

- **Goal**: efficiently and intuitively build models
- Typically for supervised learning: classification
- Improve how humans annotate data used in training
- Incorporate human directly in to the model-building process
- Typically, a blend of *active learning* with *visual inspection / interaction* with model

# Learning for Visualization

- **Goal**: use machine learning to improve the process of visualization itself

- Recommending visualizations, (semi-)automating the creation of visualizations, constructing learning models for visualization techniques

# Course Format

- Lecture-based
- No textbook for the course
- Course material based on research papers
- (LINK)
- Assessment: class participation (10%), project (90%)

# Class Participation

- Expectations
  - You have read the papers listed on the schedule prior to lecture (LINK)
  - During lecture, you should provide critiques on the covered papers - visual design? interactions? do they successfully address intended problems?
  - Some lectures you will find more relevant than others
  - Critiques during all lectures will *help you in your project*

# Project

- Will span the entirety of the course
- Three main components:
  - Project Proposal **(Abstract & Introduction)**
  - Baseline **(Related Work)**
  - Full Project **(Technical Approach Details, Results)**
- Treat the project as a research paper
- (LINK)

# Project Proposal

- What do you want to do for your project?
- Proposal Document
    - basic info, description, background, data, baseline, ***schedule***
- Proposal Presentation
    - 5-minute talk outlining the goals of your project
- We will provide feedback on your proposal, we will agree to a refined/expanded scope, and this will serve as the basis for assessment

# Project Types

- Mirrors the structure of the course:
  - Mixed-Initiative Visual Exploration
  - Visual Analytics for Model Understanding
  - Visual Analytics for Model Training
  - Learning for Visualization

# Project Topics

- Choose a topic that is most interesting to you! It will make it much easier to invest the necessary effort.
    - Certain domain of interest? Dataset? ML model? ML training procedure?
- Unable to decide on a topic? We can meet with you to discuss  problems from a variety of domains.
    - (first come first serve, email me if interested)

# Baseline

- You will be expected to implement an existing approach that will either:
    - Serve as a piece of your project
    - A competing method, one you intend to compare against
- Baseline will help focus your project: deal with data wrangling/cleaning, Visualization/ML libraries, etc..
- Pre-baseline: use existing visualization tools to visualize your data! (see website for further references)

# Project Updates

- Throughout the semester you will be expected to give updates to the class on the progression of your project, and baseline.

- Intended to keep you *on track*.

- Updates comprise part of project assessment: you will not be graded on whether you are keeping up with your proposed schedule; you will be graded on detailing where things went wrong, unexpected challenges, etc..

# Final Project

- Project presentation: how does your project improve over prior work? What is your technical approach? What are the features of your visualization? Strengths/weaknesses? Insights gained? Should be all-encompassing.

# Project Summary

- Start thinking about projects …
  - **Now!** 😀
- Do not put off the project proposals: carefully think about what you want to do, and how you are going to achieve it.
- Proposal does not need to be perfect: We will provide feedback
- We will cover basics of VA & ML, to give you some basis; but you should also be proactive and comb through papers on the website.

# Prerequisites (1)

- (one of the following two)
- Working knowledge of machine learning
    - Experience in data cleaning, training models, evaluation, and ideally some experience in optimizing models from scratch (e.g. not *just* using TensorFlow, PyTorch, etc..)
- Working knowledge of data visualization
    - Some experience with visualization tools, and ideally, building your own!

# Prerequisites (2)

- You should have a solid linear algebra background
- Basics of optimization is a plus
    - Linear systems, eigenvalue problems, matrix factorization, gradient descent, stochastic gradient descent (and the many variants)
- For visualization: you will need to be able to think in terms of **space, shape, colors**, and how data gets mapped to these visual encodings.

# Prerequisites (3)

- Programming languages? Libraries?

- Up to you! You will need to be **self-motivated** for this course.

- The ideal languages/libraries likely to vary project-to-project (e.g. Python, Javascript, C++, Processing, etc)…

- (*Treat the baseline as an opportunity to learn libraries that you will need for your project!*)

- Additionally: I have listed a set of resources on the website (LINK)

# Course Logistics

- Please see course website for additional information on project, lectures, late submission policy, academic honesty, etc…

- Questions?