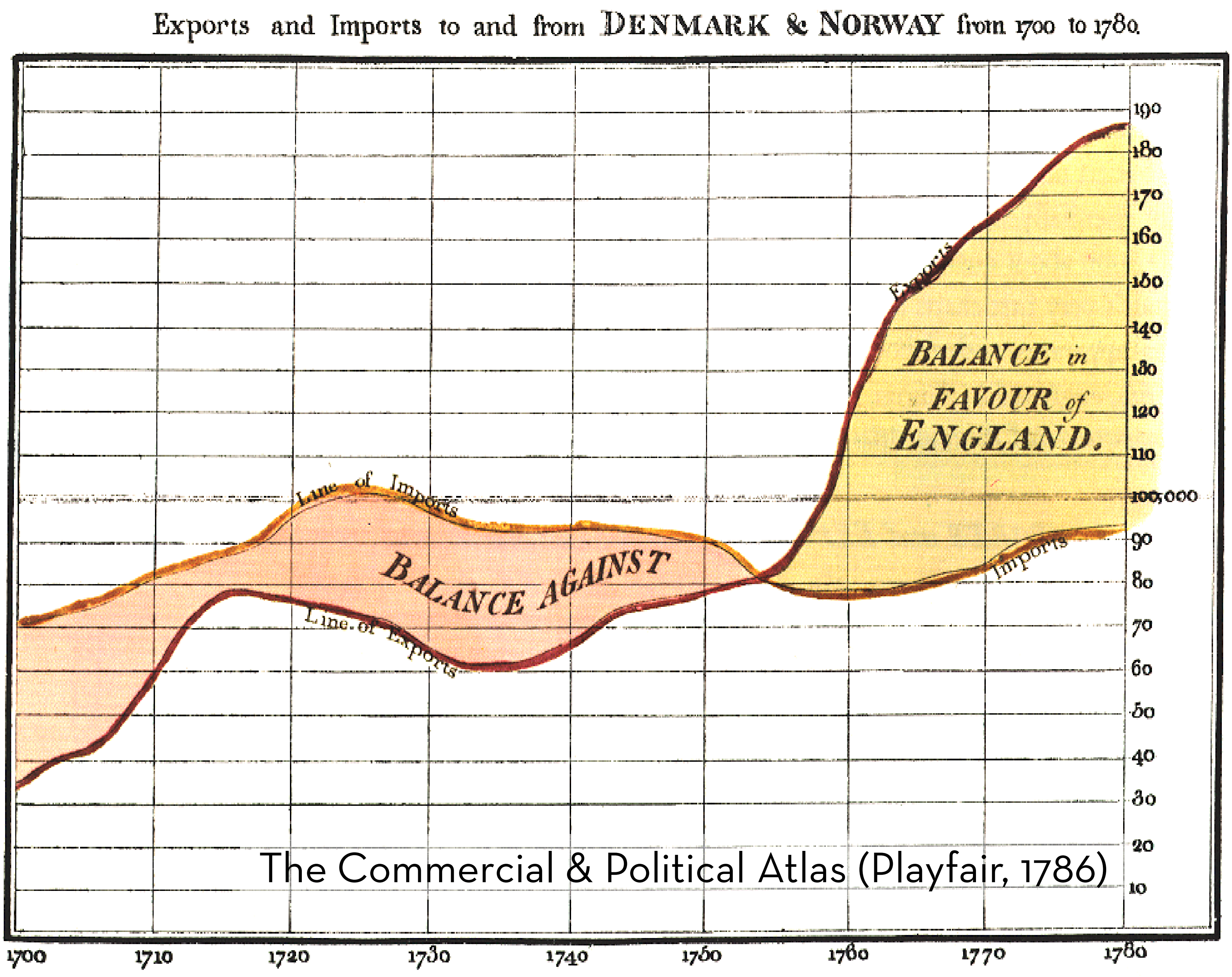
ReVision

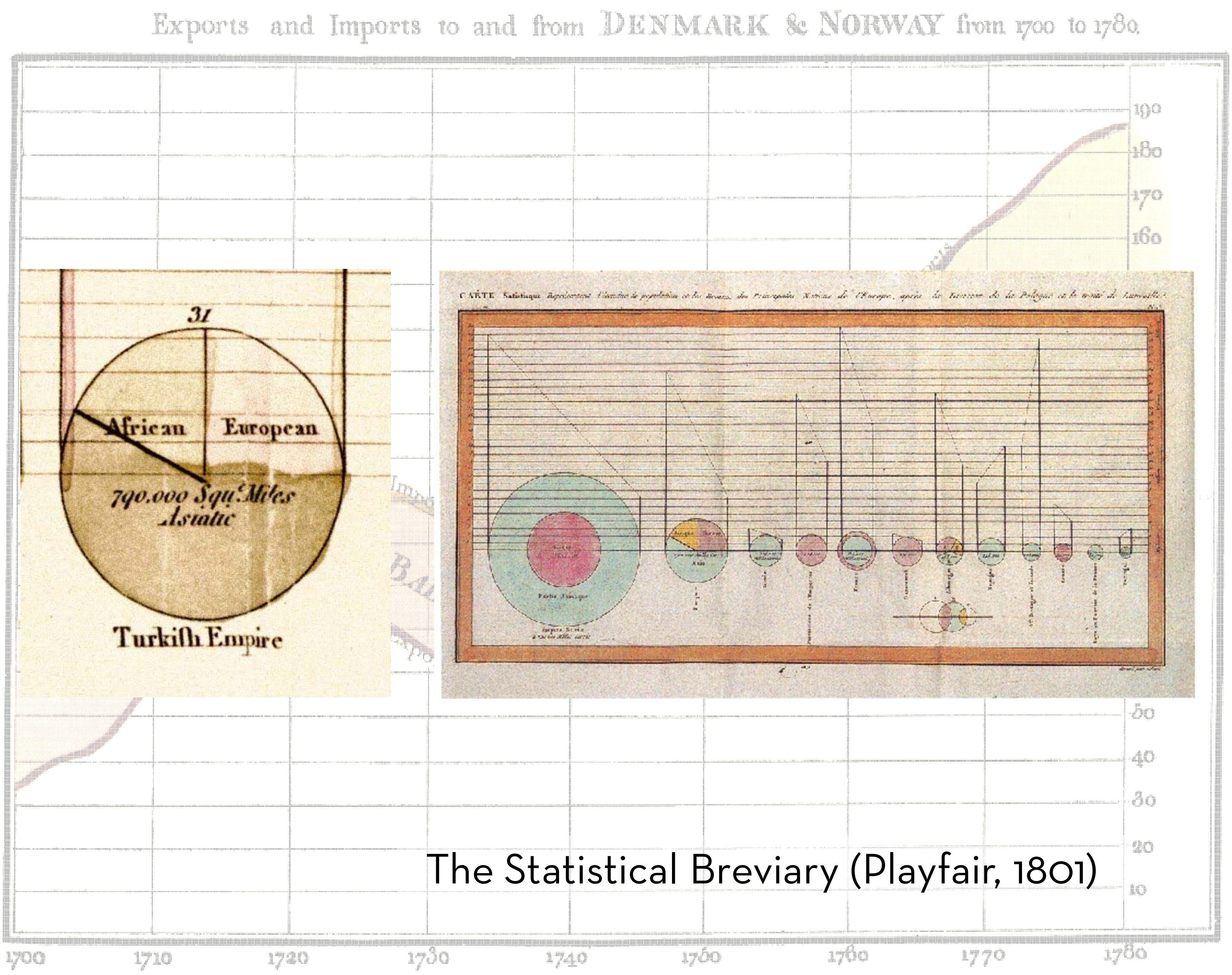
Automated Classification, Analysis and Redesign of Chart Images

Jeffrey Heer & Fei-Fei Li

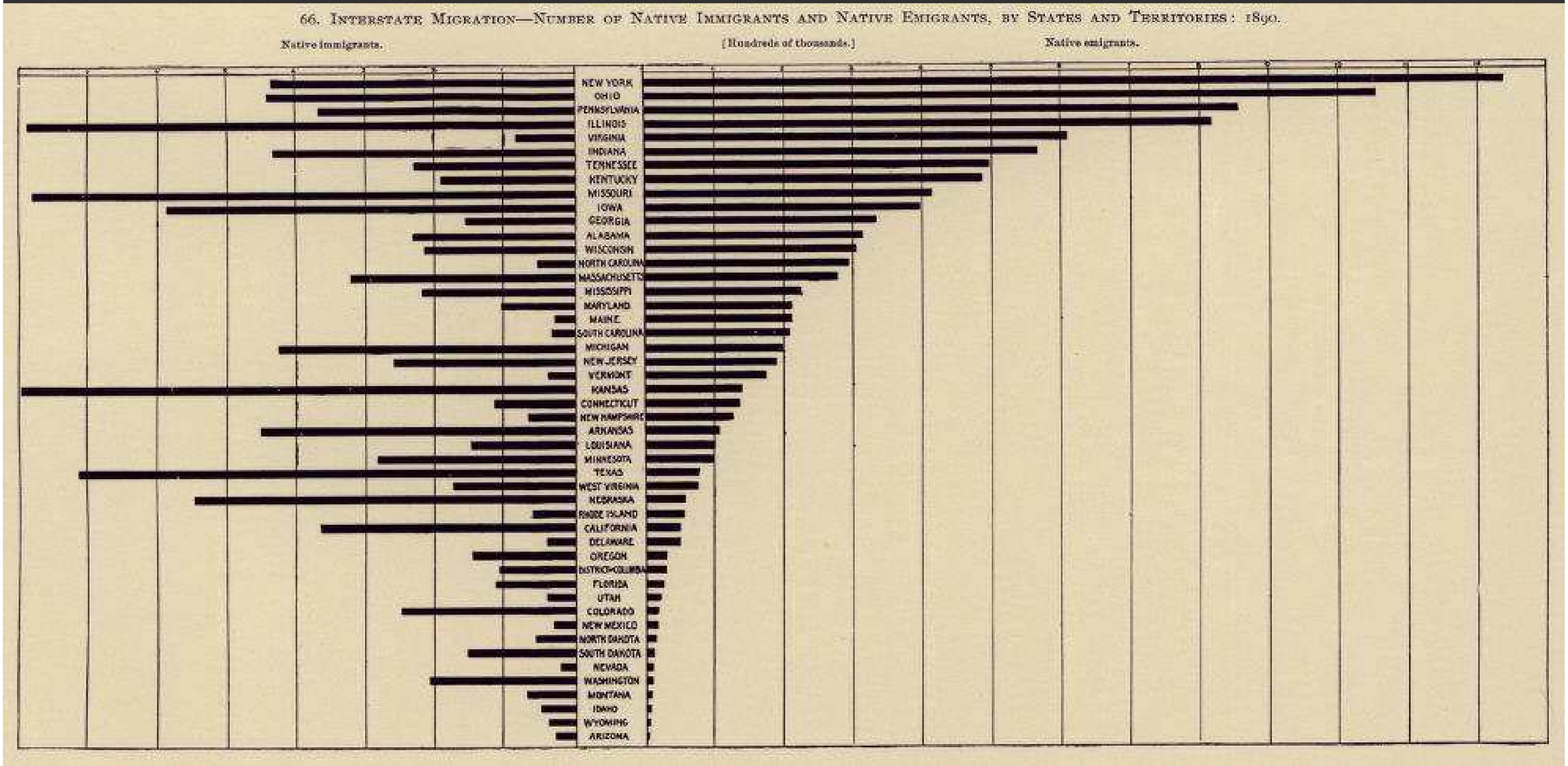
with Manolis Savva, Nick Kong, Arti Chhajta, & Maneesh Agrawala



The Commercial & Political Atlas (Playfair, 1786)

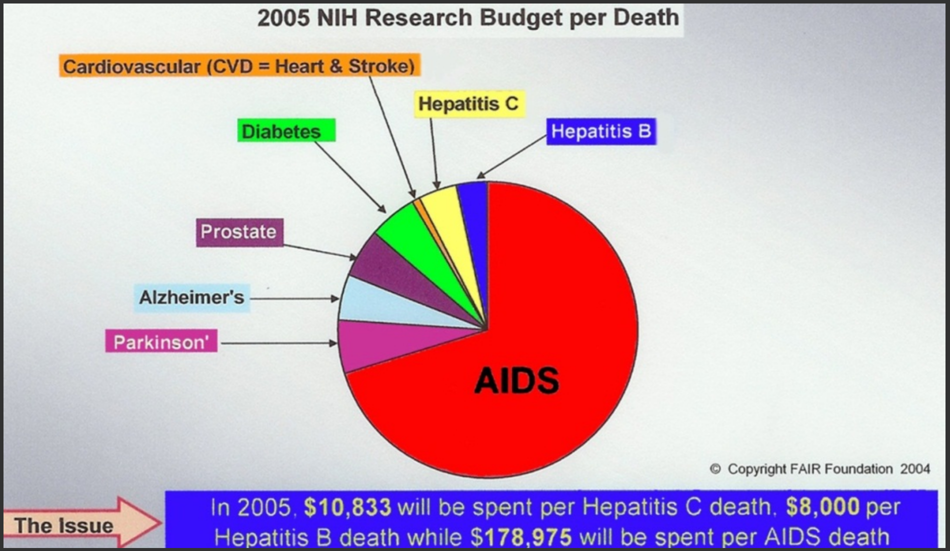


The Statistical Breviary (Playfair, 1801)



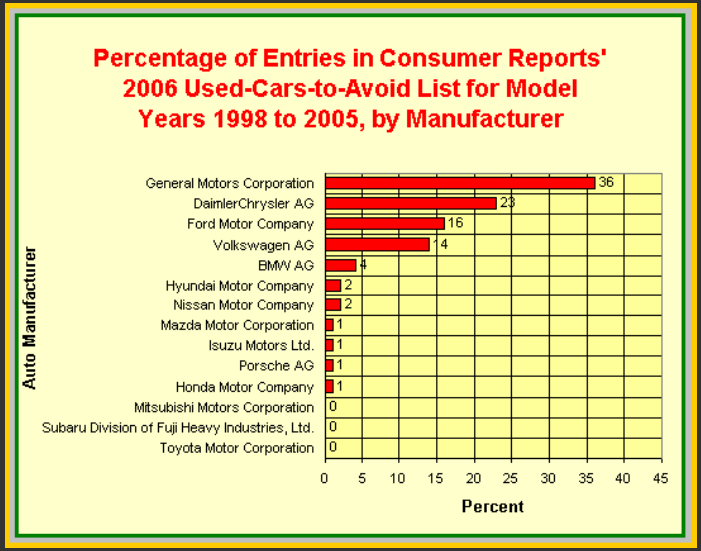
Statistical Atlas of the Eleventh U.S. Census (1890)

Poor Designs are Prevalent

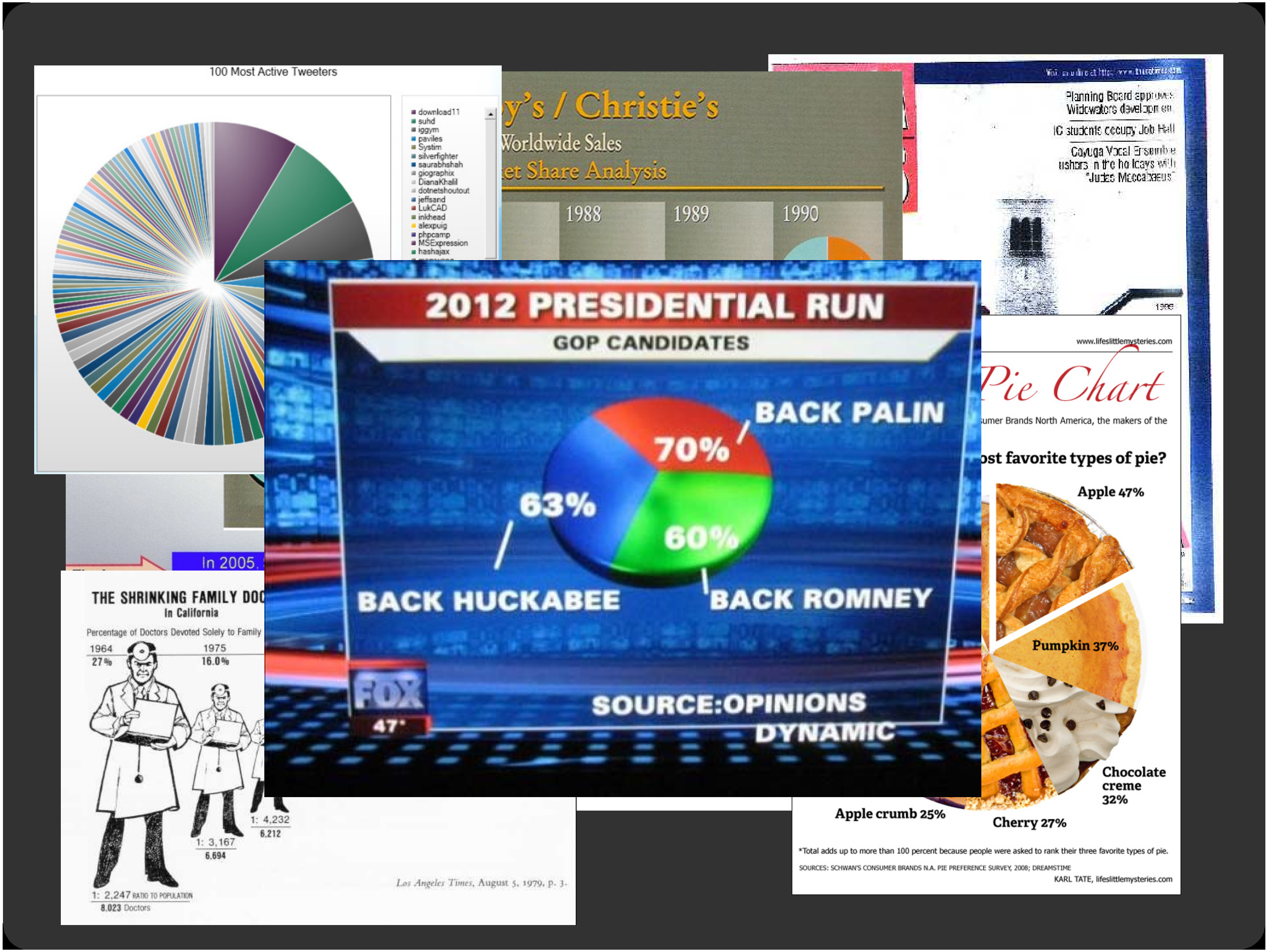


2005 NIH Research Budget per Death

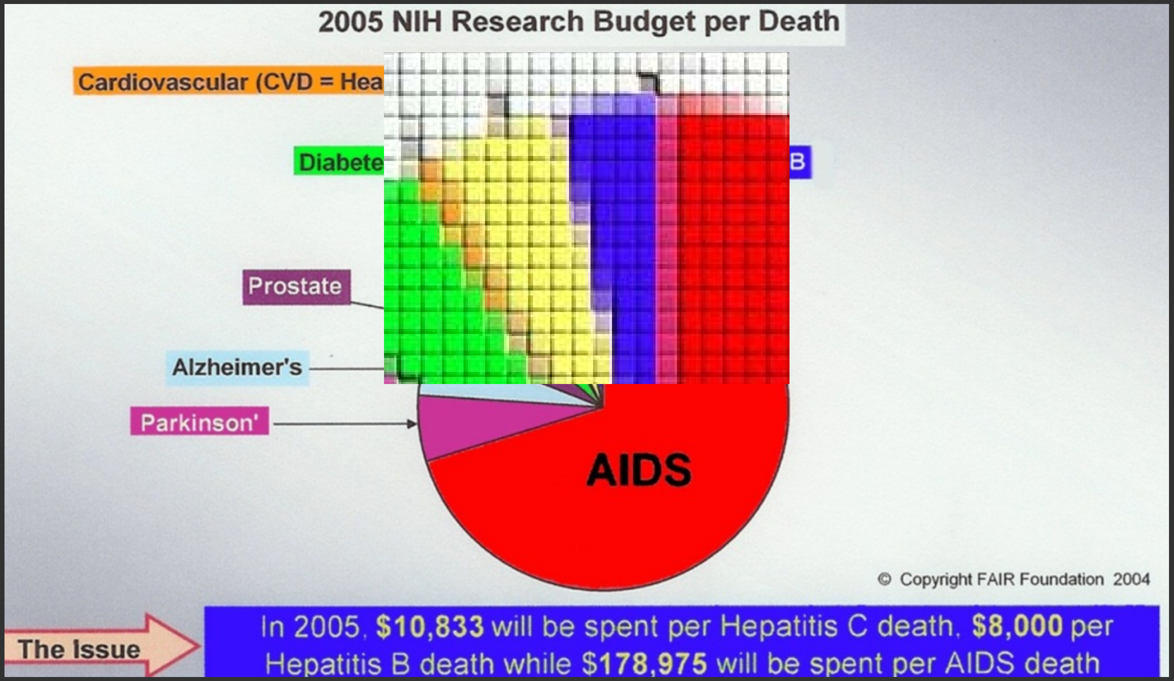
In 2005, **$10,833** will be spent per Hepatitis C death. **$8,000** per Hepatitis B death while $**178,975** will be spent per AIDS death



Percentage of Entries in Consumer Reports' 2006 Used-Cars-to-Avoid List for Model Years 1998 to 2005, by Manufacturer


100 Most Active Tweeters
download11
suhd
iggym
paviles
Systim
silverfighter
saurabhshah
giographix
DianaKhalil
dotnetshoutout
jeffsand
LukCAD
inkhead
alexpuig
phpcamp
MSExpression
hashajax
y's / Christie's
Worldwide Sales
et Share Analysis
1988
1989
1990
Planning Board approves
Widewaters development
IC students occupy Job Hall
Cayuga Vocal Ensemble
ushers in the holidays with
"Judas Maccabaeus"
In 2005,
THE SHRINKING FAMILY DOC
In California
Percentage of Doctors Devoted Solely to Family
1964
27%
1975
16.0%
1: 2,247 RATIO TO POPULATION
8,023 Doctors
1: 3,167
6,694
1: 4,232
6,212
Los Angeles Times, August 5, 1979, p. 3.
2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47°
SOURCE:OPINIONS DYNAMIC
www.lifeslittlemysteries.com
Pie Chart
umer Brands North America, the makers of the
ost favorite types of pie?
Apple 47%
Pumpkin 37%
Chocolate creme 32%
Apple crumb 25%
Cherry 27%
*Total adds up to more than 100 percent because people were asked to rank their three favorite types of pie.
SOURCES: SCHWAN'S CONSUMER BRANDS N.A. PIE PREFERENCE SURVEY, 2008; DREAMSTIME
KARL TATE, lifeslittlemysteries.com

Bitmap Images are Common



In 2005, **$10,833** will be spent per Hepatitis C death, **$8,000** per Hepatitis B death while $**178,975** will be spent per AIDS death

Hard to extract data

Hard to index and search

Hard to redesign

The Challenge

Computational Visualization Interpretation

Given a bitmap image of a chart:

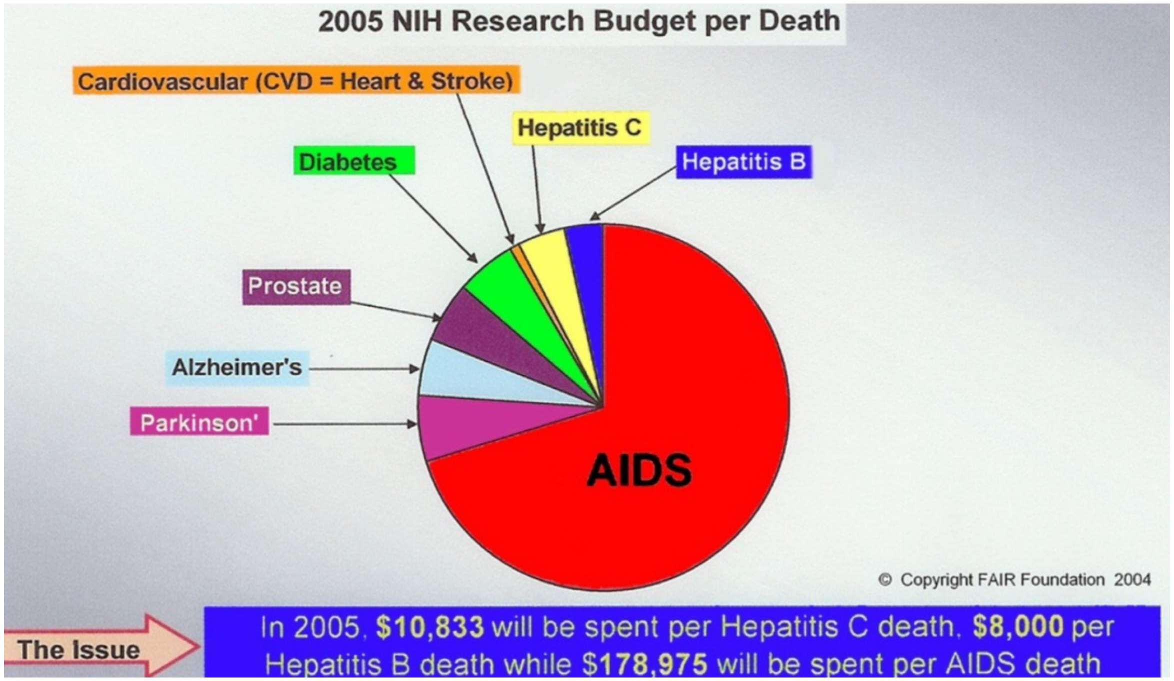
- Determine the chart type (bar, pie, *etc*)
- Infer a model of the underlying data

Applications

Retarget to new charts or visual styles

Data extraction, indexing & retrieval

Evaluate visualization designs



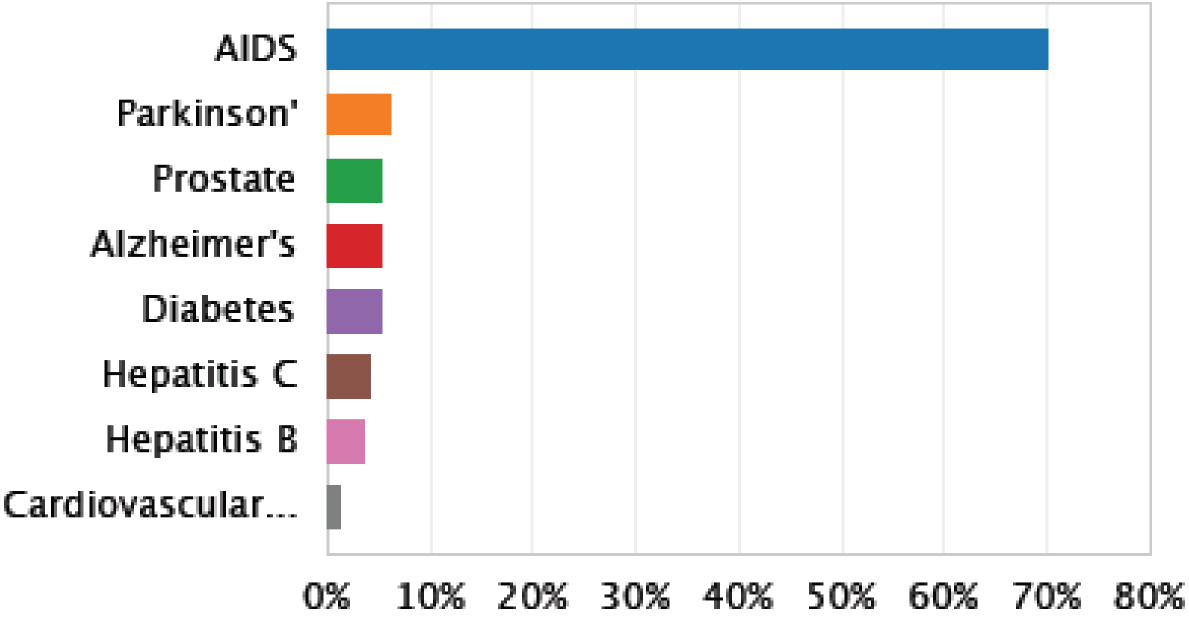
© Copyright FAIR Foundation 2004

In 2005, **$10,833** will be spent per Hepatitis C death, **$8,000** per Hepatitis B death while **$178,975** will be spent per AIDS death

Type: Pie

Data Table

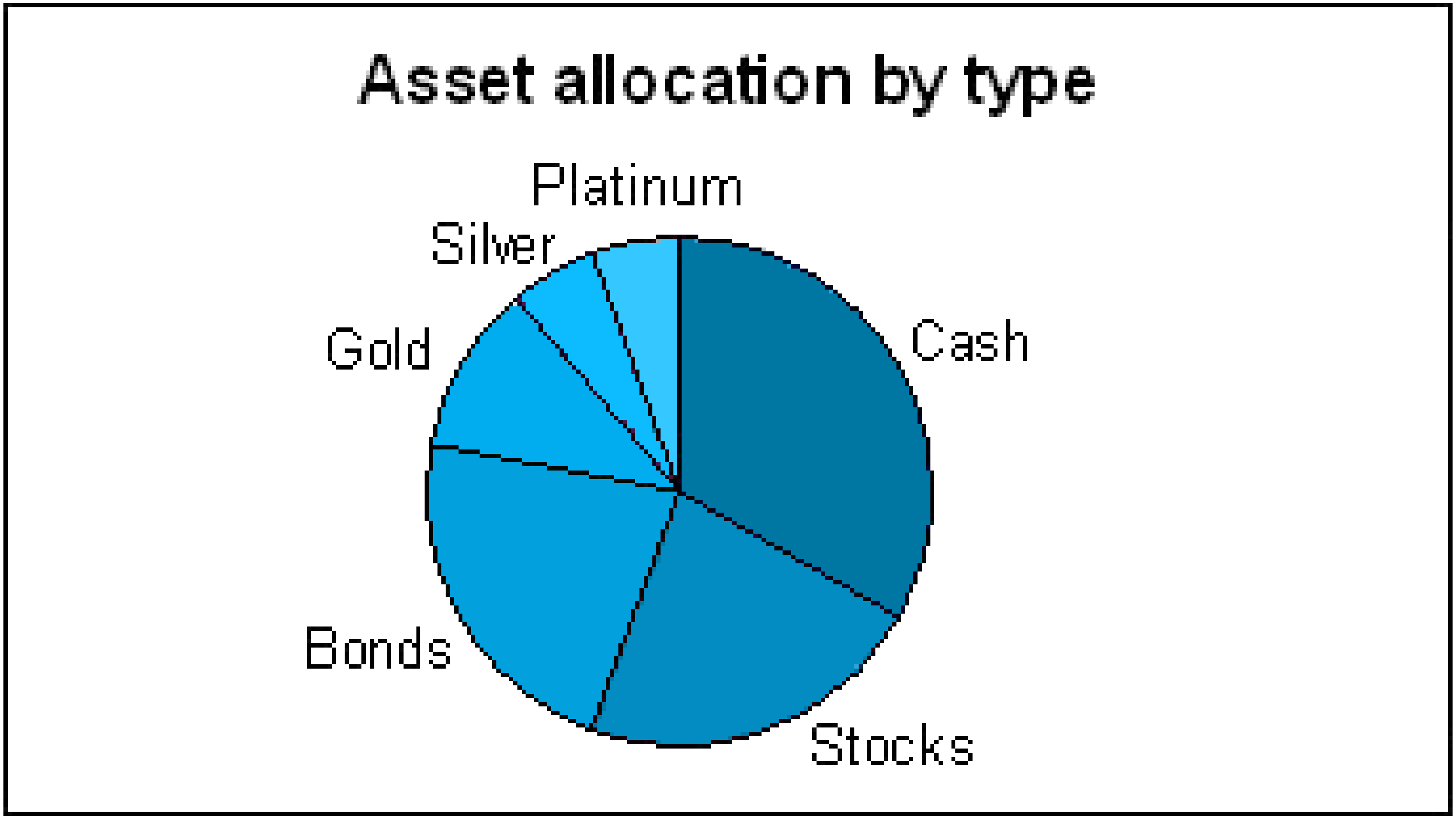
AIDS	70.0%
Parkinson'	6.0%
Prostate	5.2%
Alzheimer's	5.1%
Diabetes	5.1%
Hepatitis C	4.0%
Hepatitis B	3.5%
Cardiovasc...	1.1%



ReVision

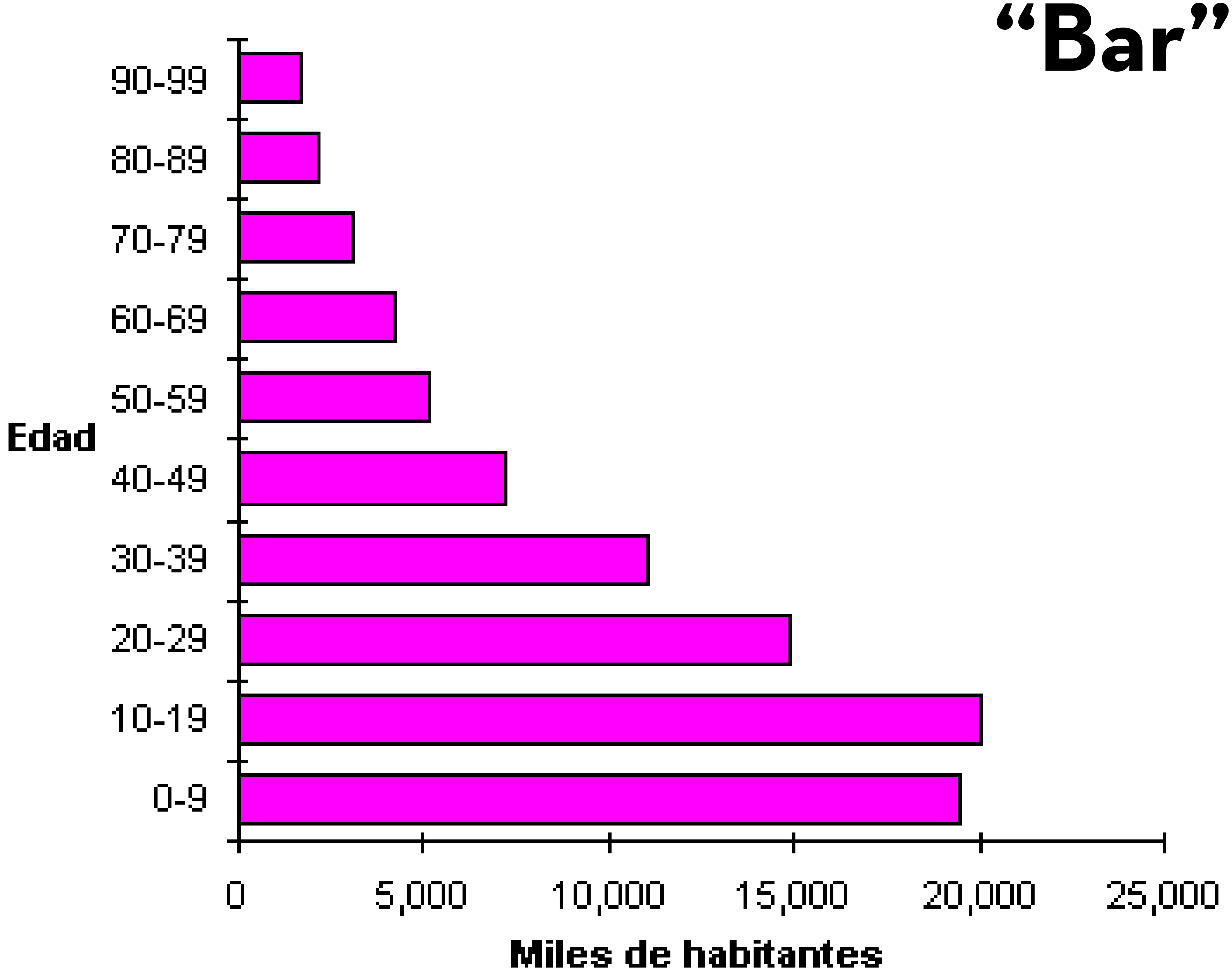
1. Classification
2. Mark & Data Extraction
3. Automated Redesign

1. **Classification**
2. Mark & Data Extraction
3. Automated Redesign


Asset allocation by type
Platinum
Silver
Gold
Cash
Bonds
Stocks

“Pie”

“Bar”


Edad
90-99
80-89
70-79
60-69
50-59
40-49
30-39
20-29
10-19
0-9
0
5,000
10,000
15,000
20,000
25,000
Miles de habitantes

Prior Work

State-of-the-Art: **Prasad et al. 2007**

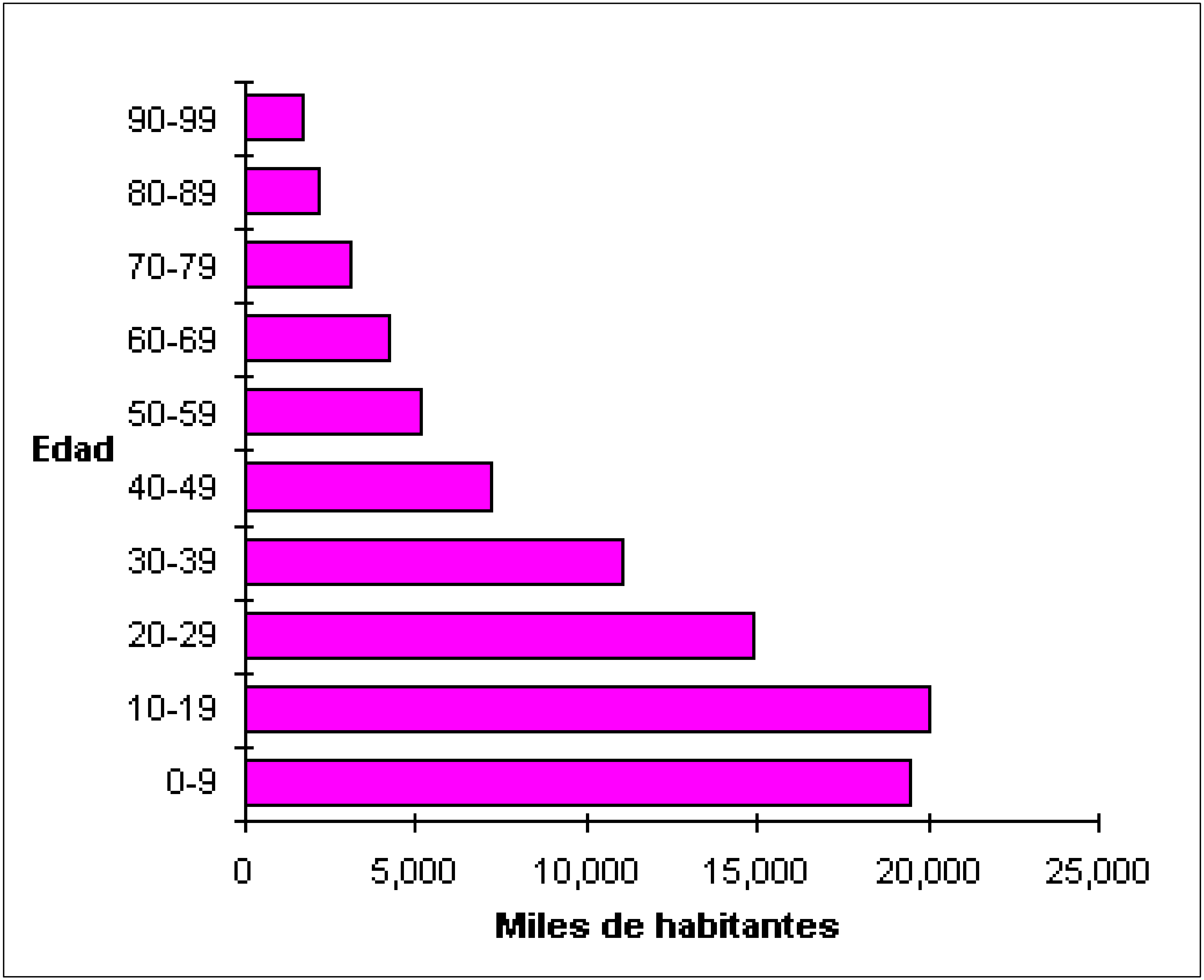
Combines many vision features: Image Segmentation, HOG, SIFT, and more

Input to a Multi-Class Support Vector Machine

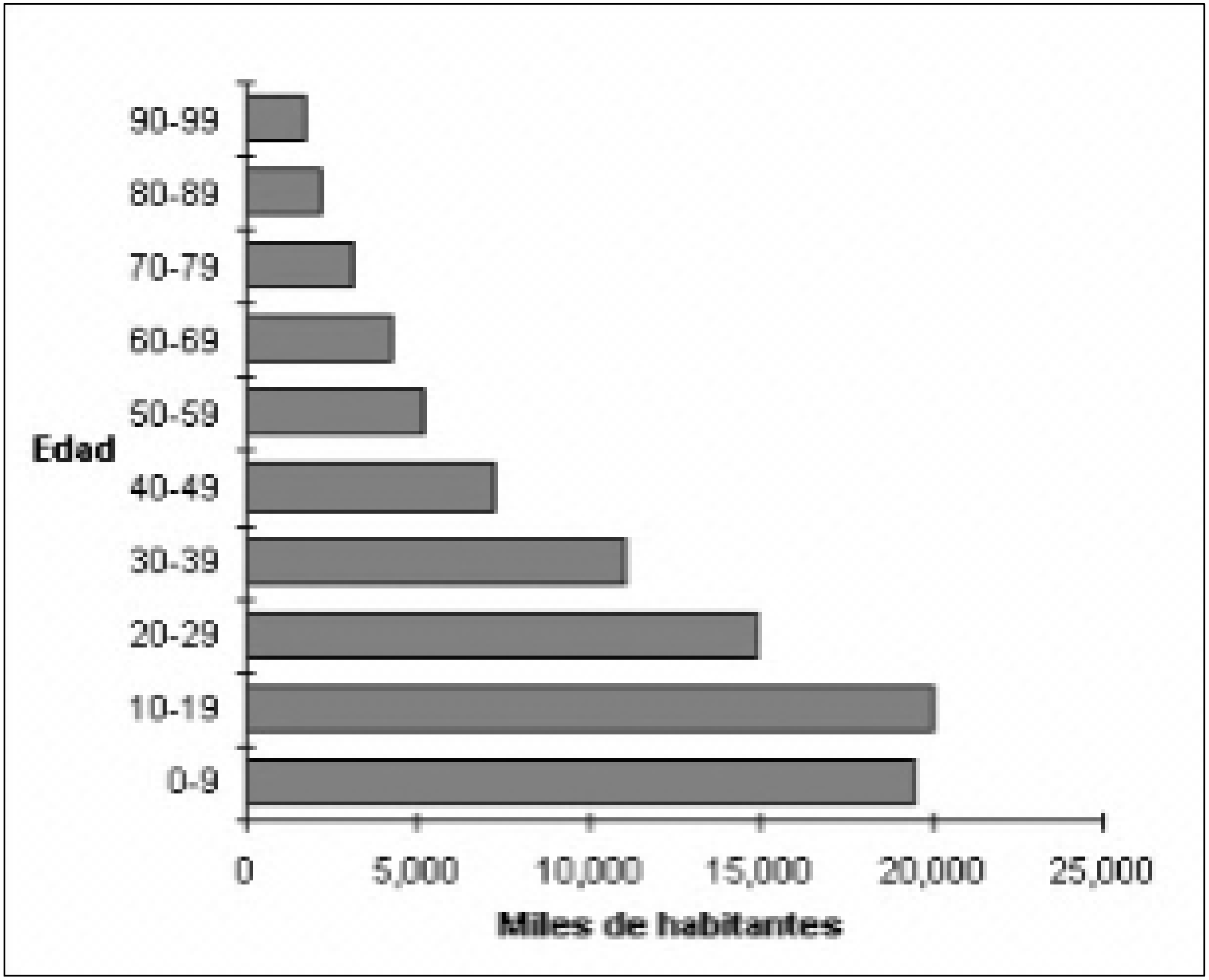
Achieves **80%** classification accuracy across 5 classes: bar, pie, line, scatter & surface charts

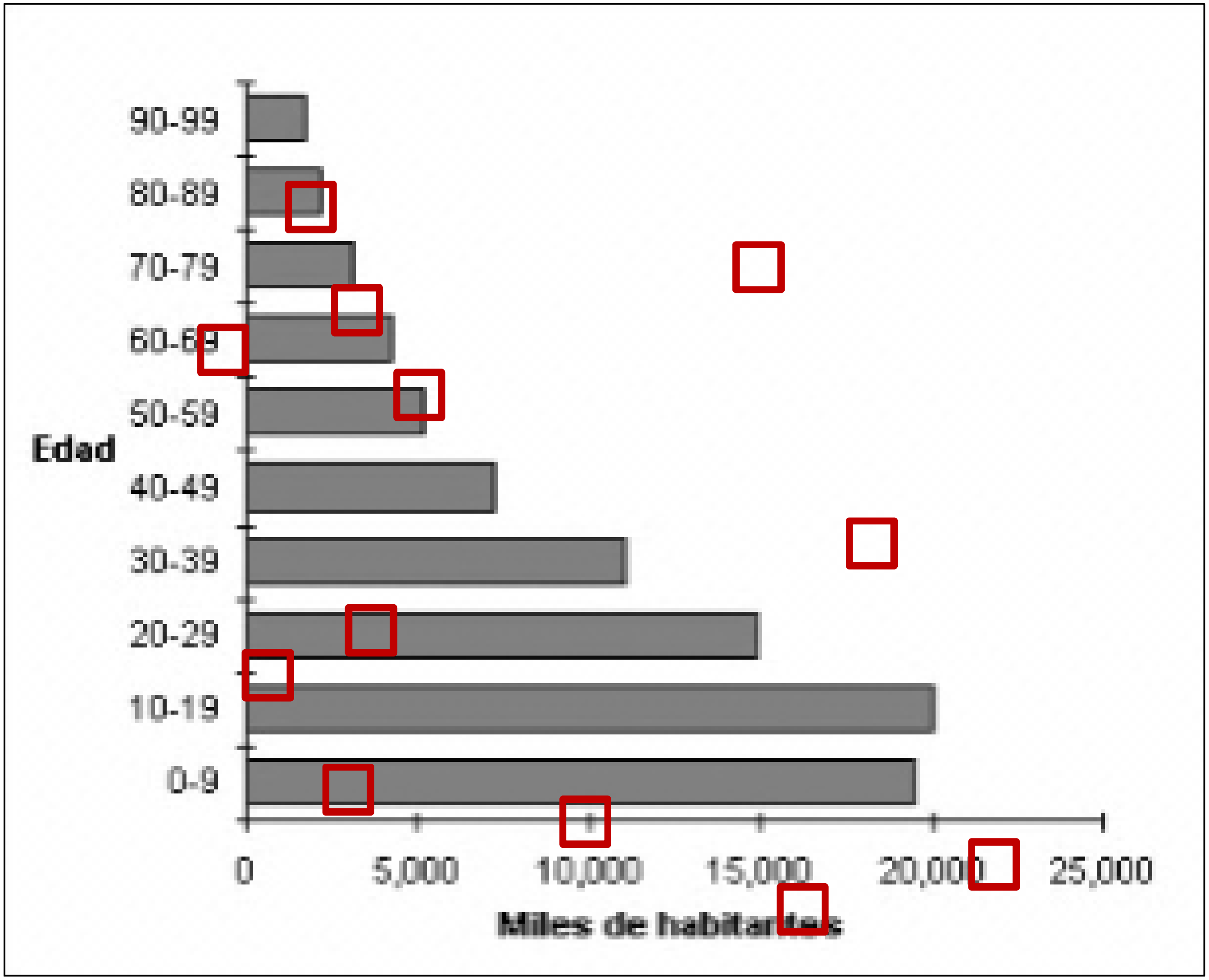
Our Approach: **Unsupervised Feature Learning**

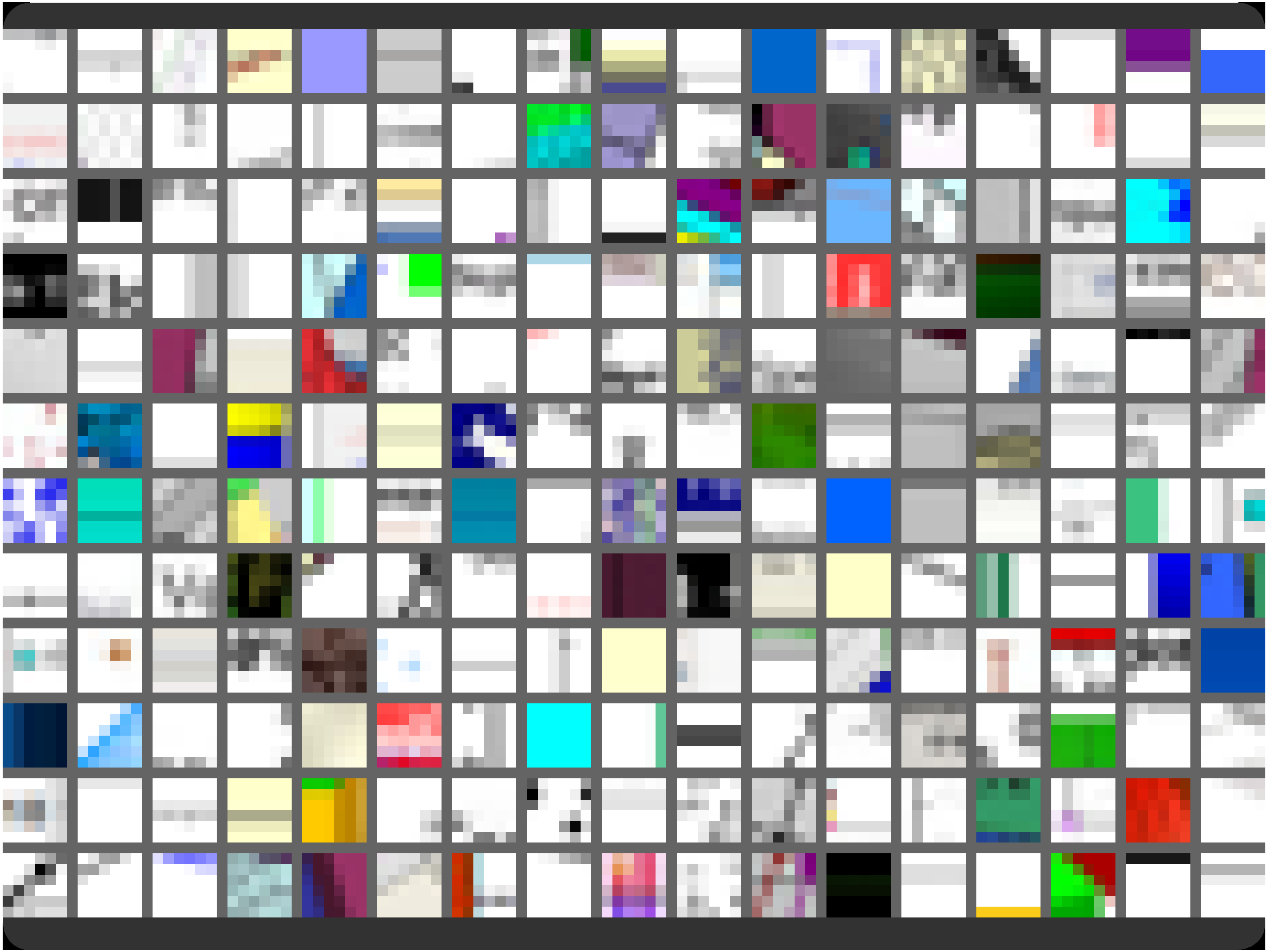
Image features at the level of *graphical marks*

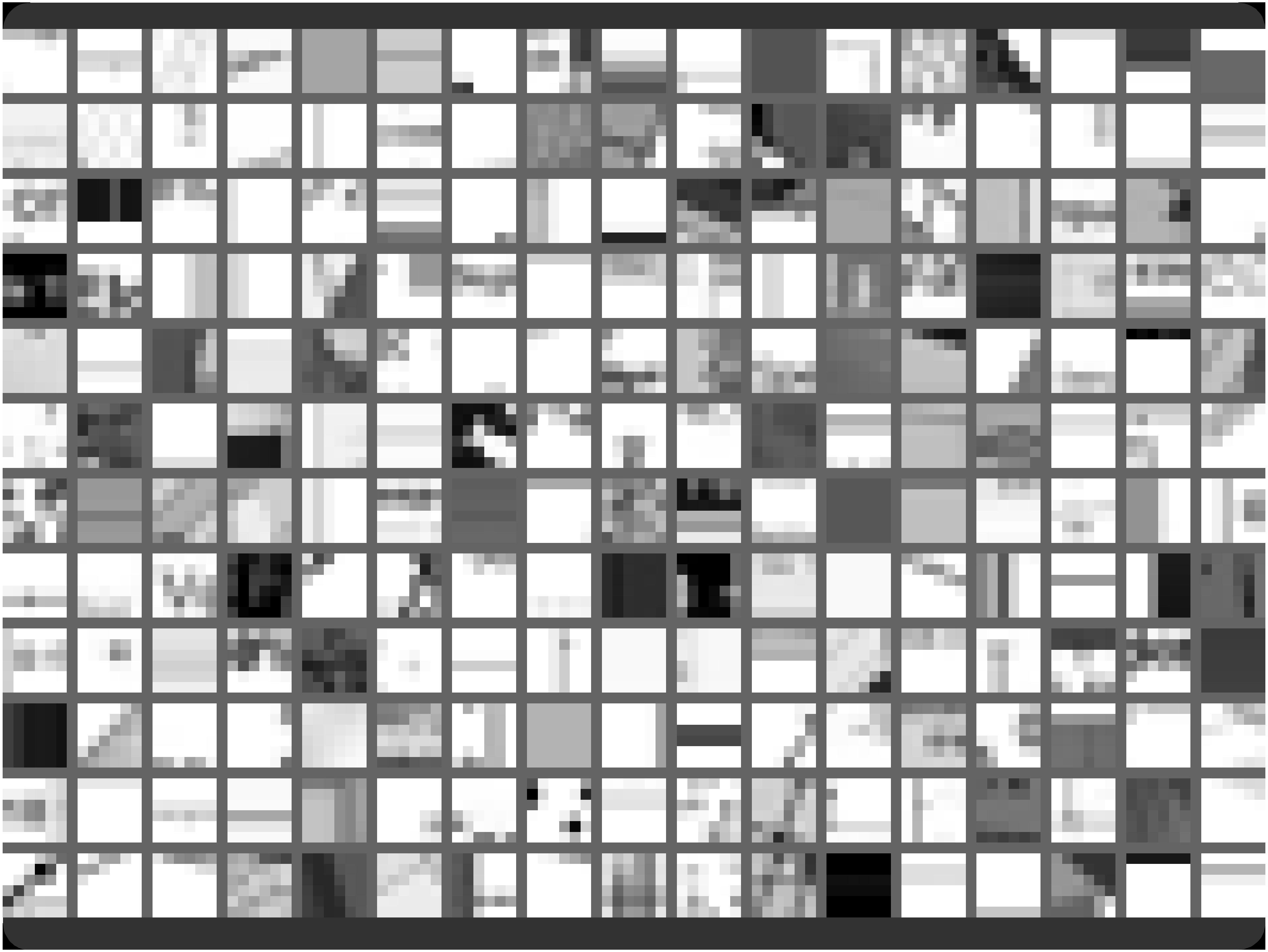


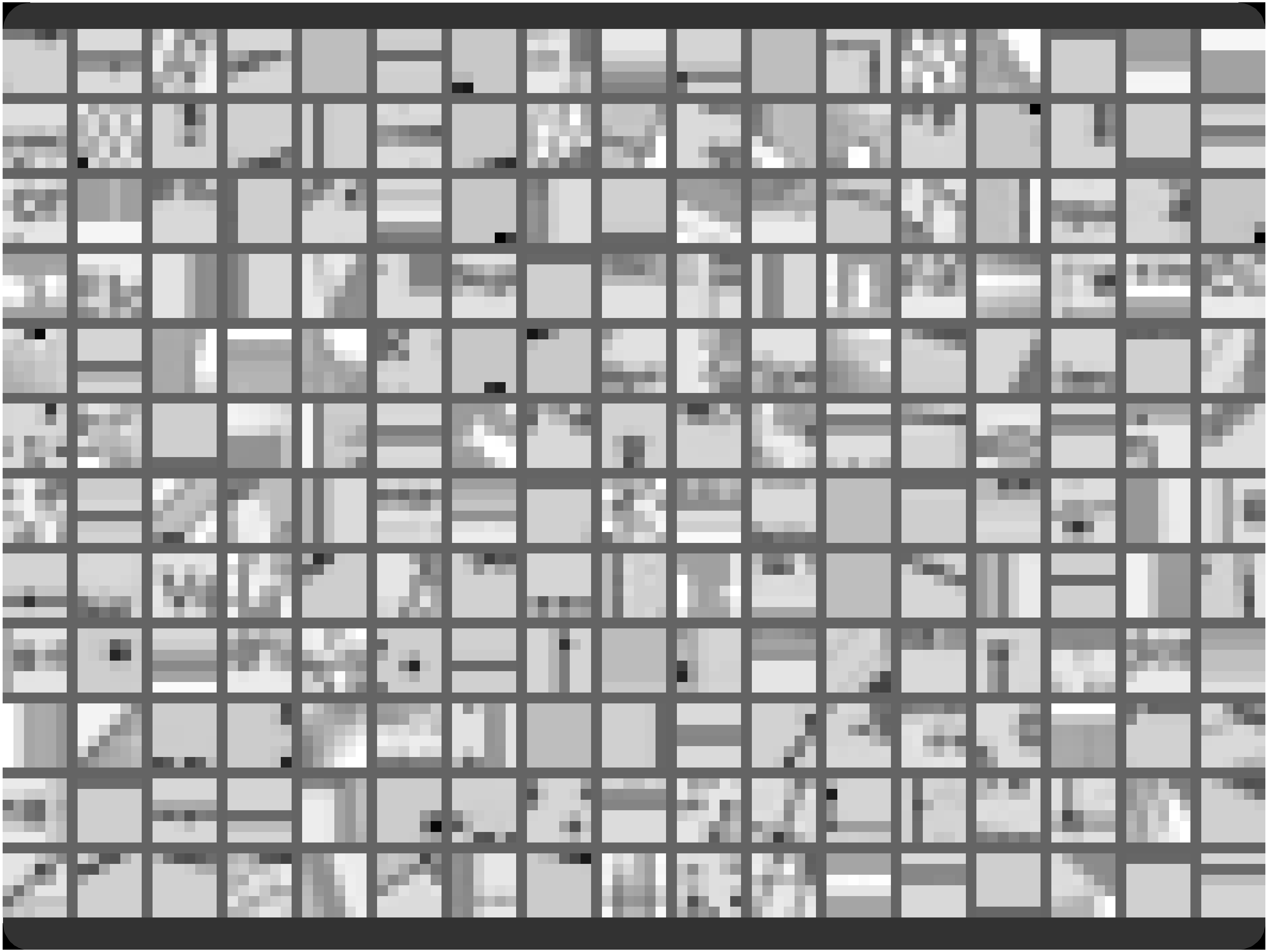
Edad
90-99
80-89
70-79
60-69
50-59
40-49
30-39
20-29
10-19
0-9
0
5,000
10,000
15,000
20,000
25,000
Miles de habitantes

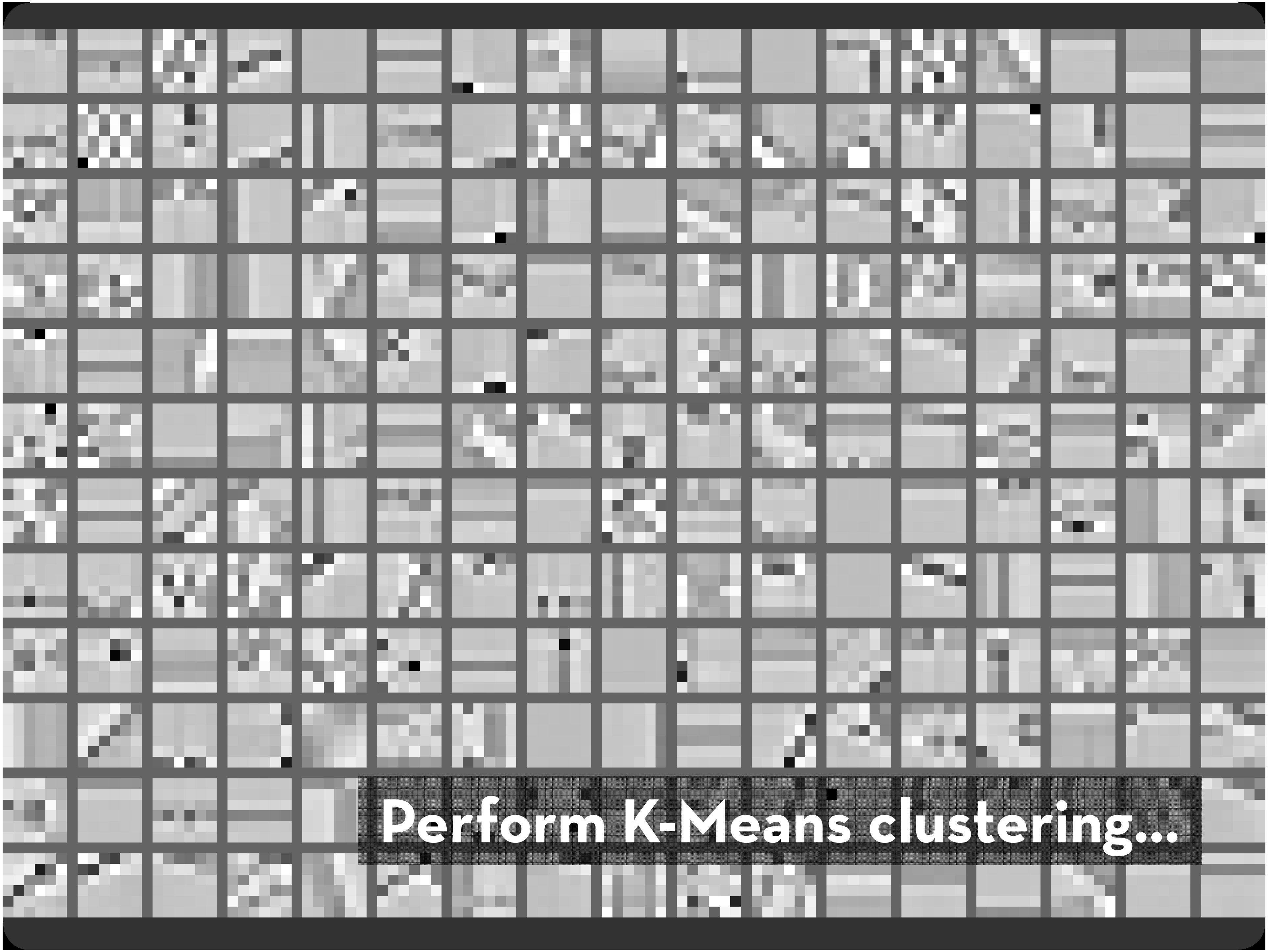

90-99
80-89
70-79
60-69
50-59
Edad
40-49
30-39
20-29
10-19
0-9
0
5,000
10,000
15,000
20,000
25,000
Miles de habitantes


90-99
80-89
70-79
60-69
50-59
Edad
40-49
30-39
20-29
10-19
0-9
0
5,000
10,000
15,000
20,000
25,000
Miles de habitantes

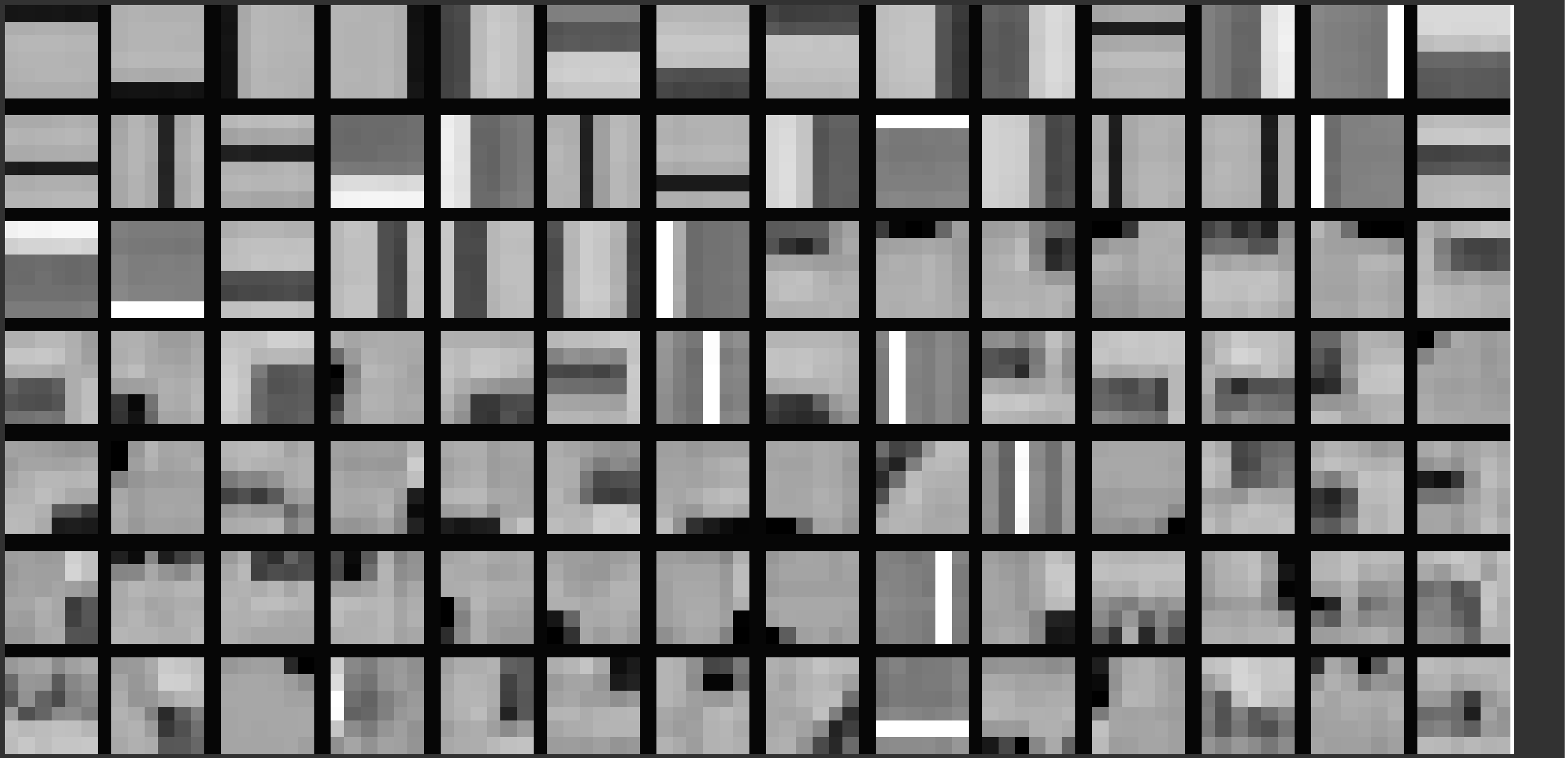




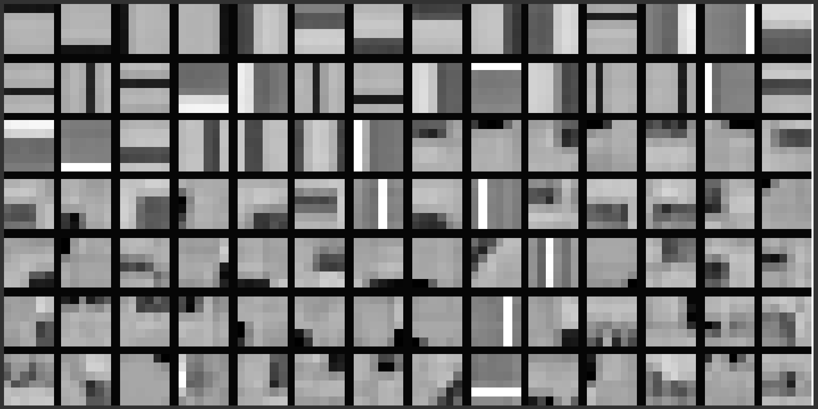




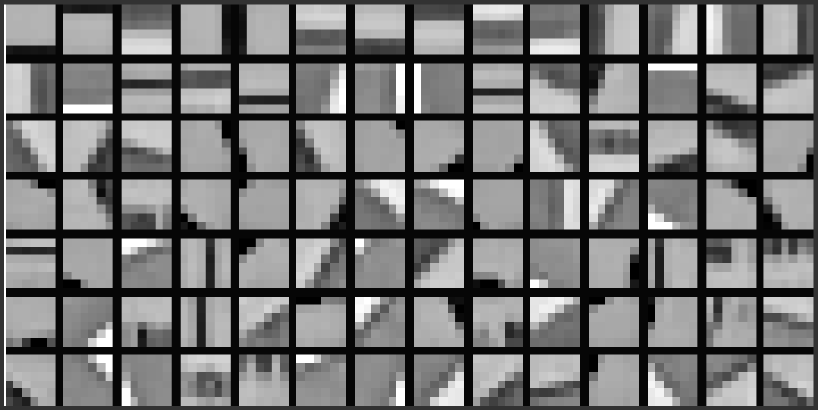
Perform K-Means clustering...



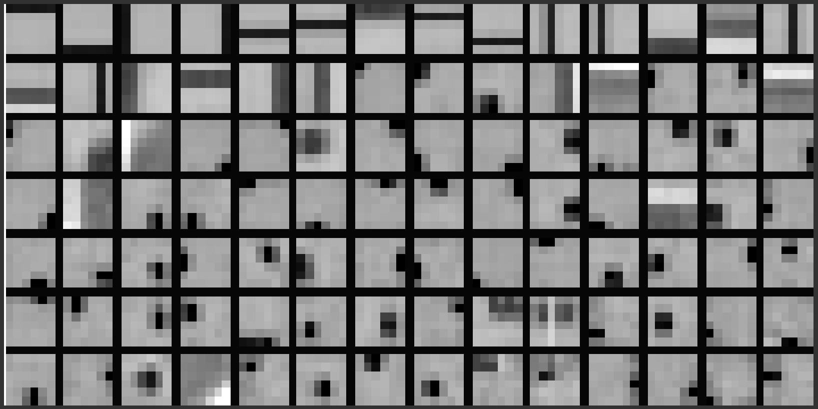
Cluster centroids are treated as a codebook.
These patches serve as our image features.



Bar Charts



Pie Charts



Scatter Plots

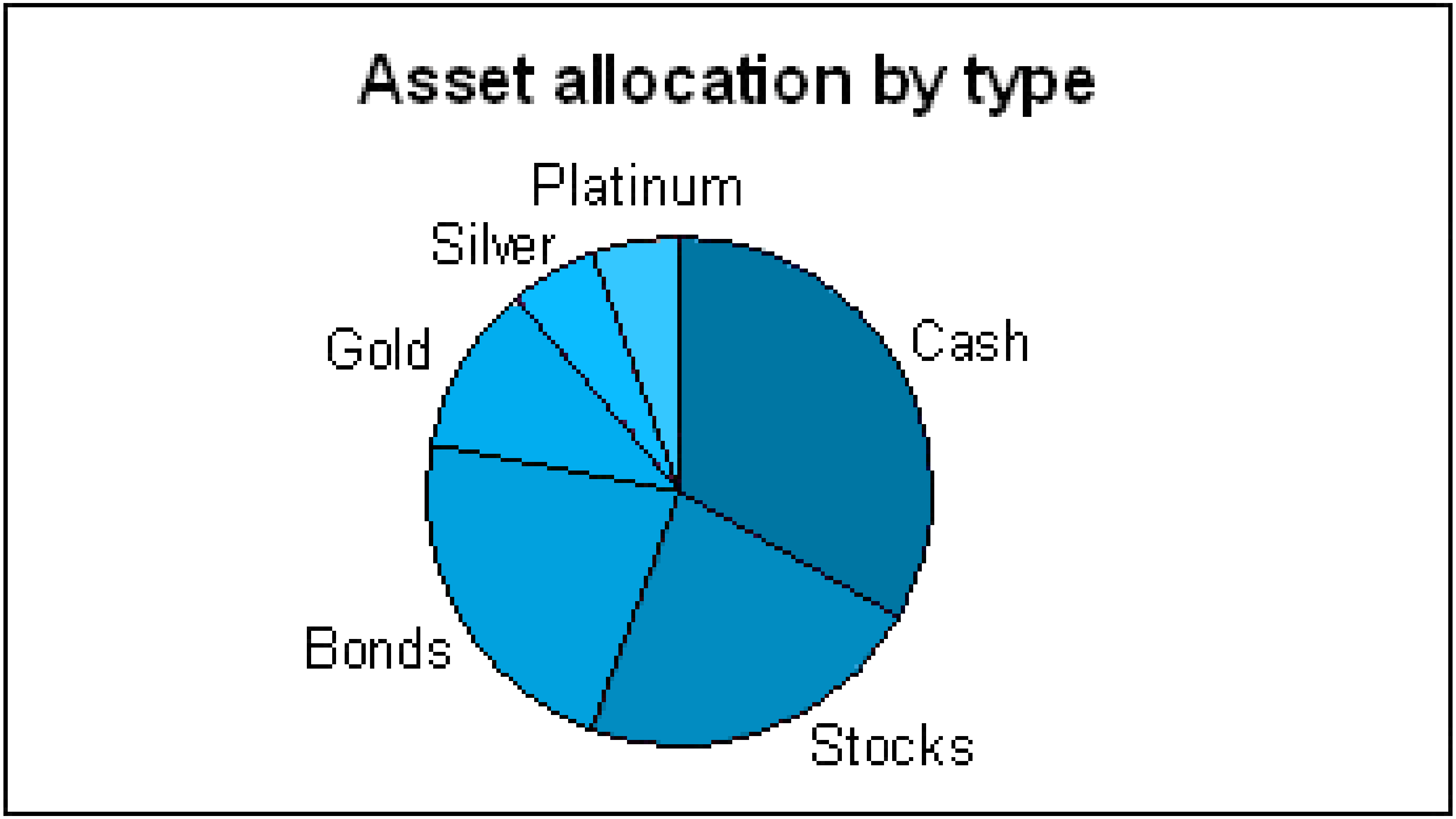
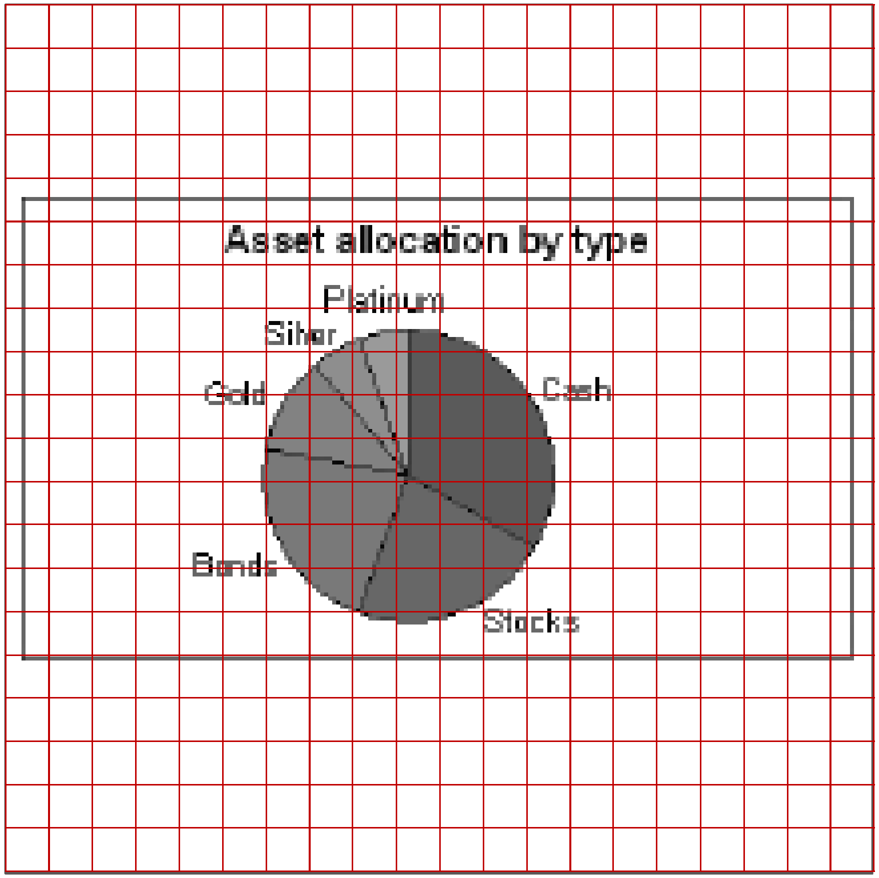

Asset allocation by type

Chart Image Classification



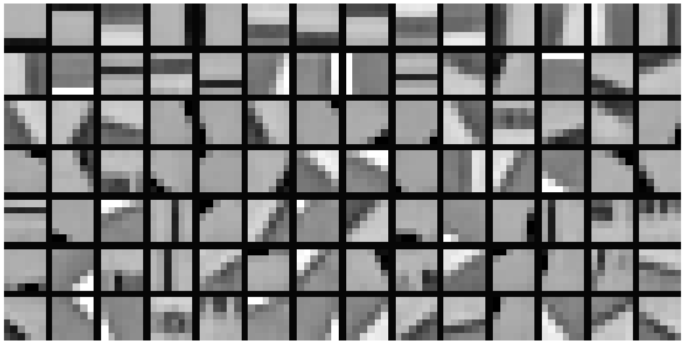


Chart Image Classification

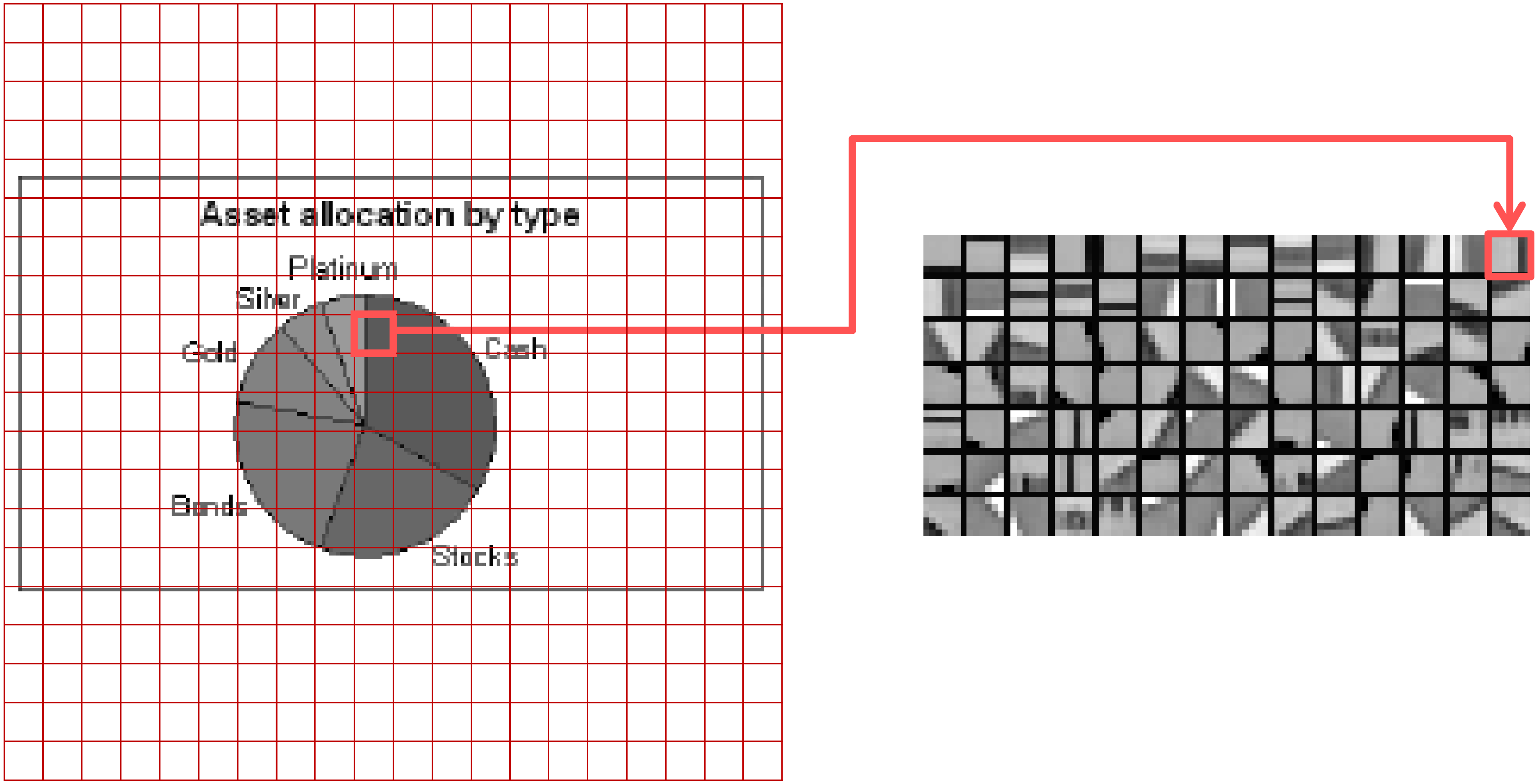


Chart Image Classification

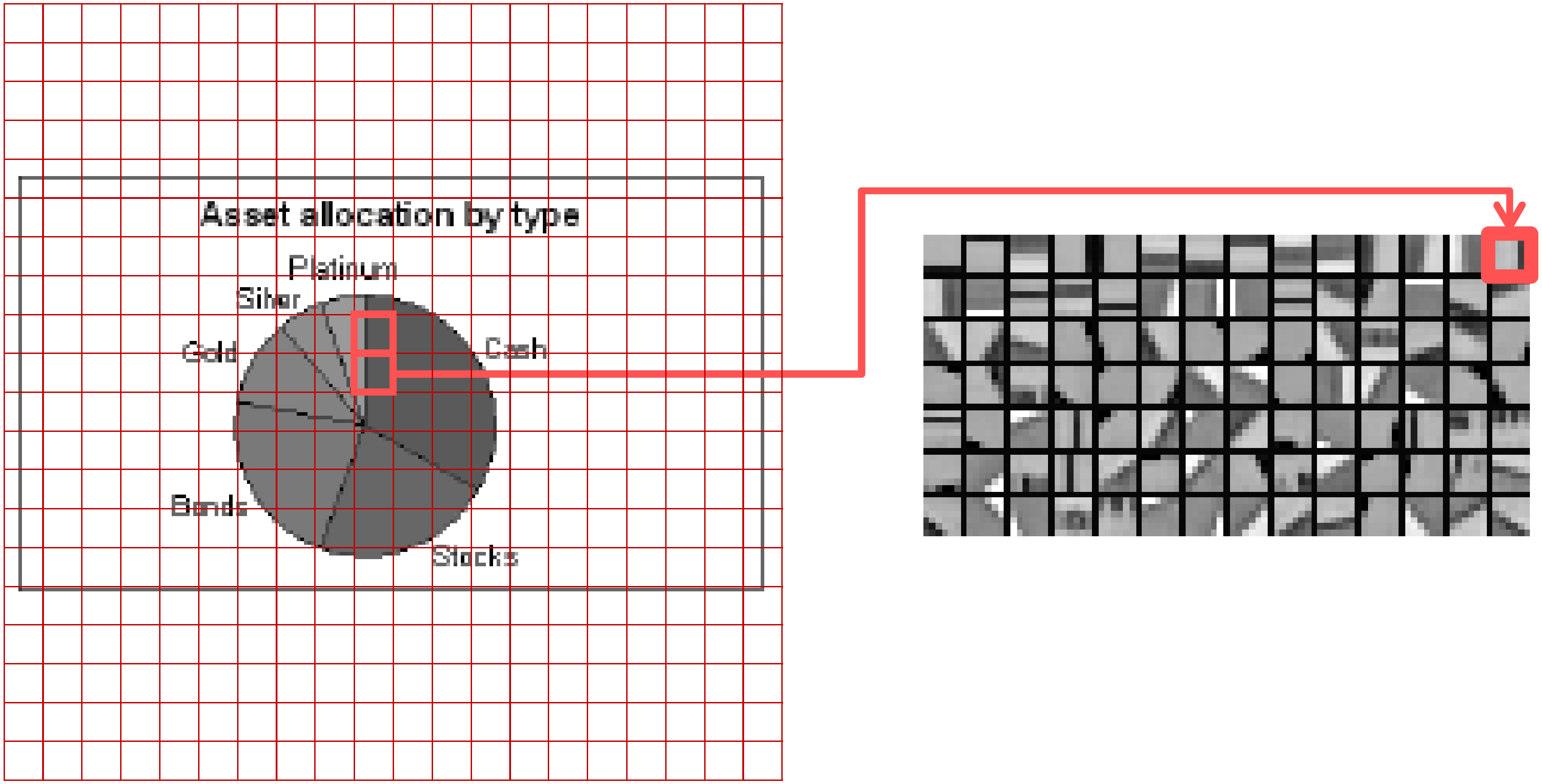


Chart Image Classification

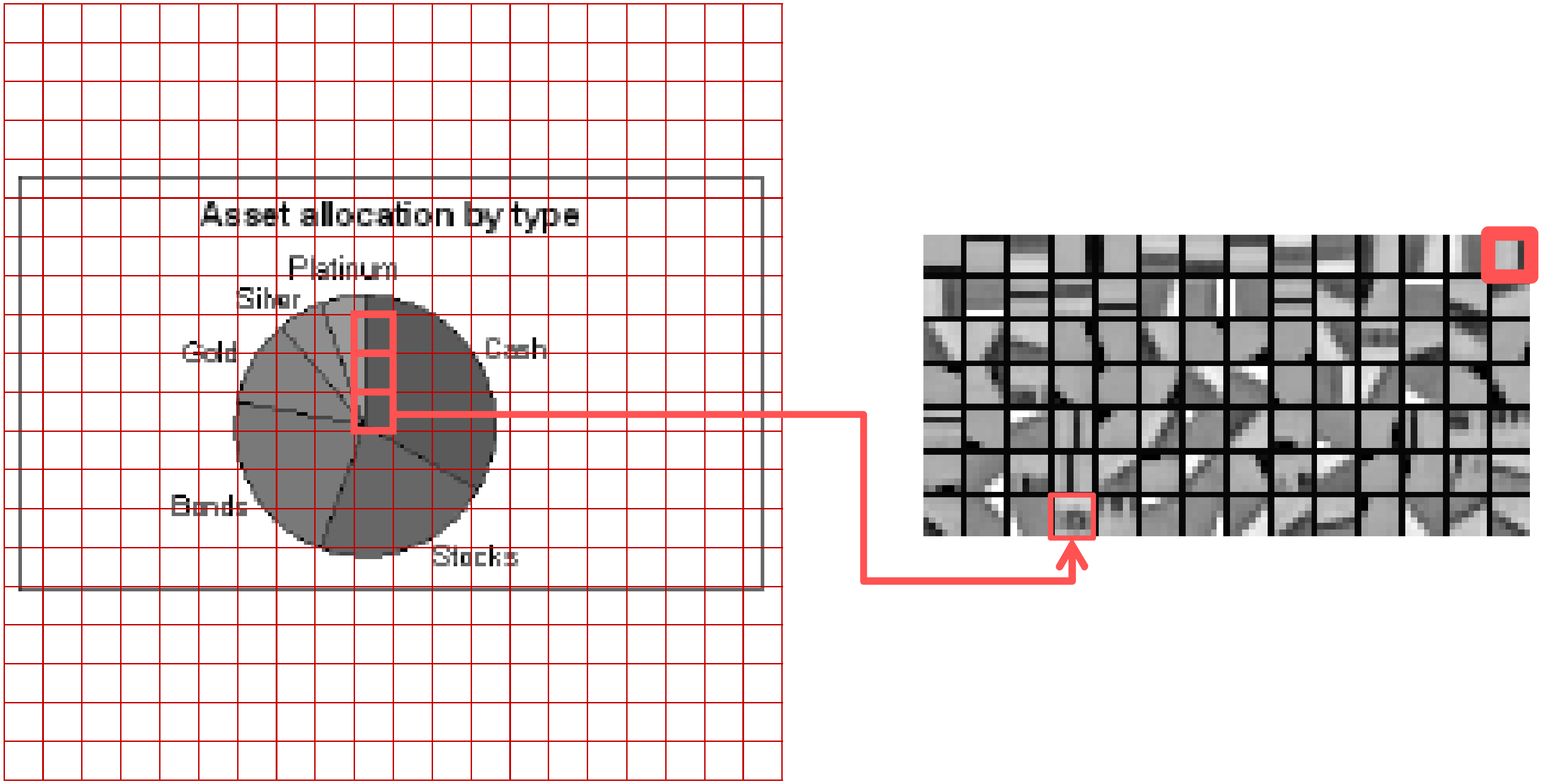


Chart Image Classification

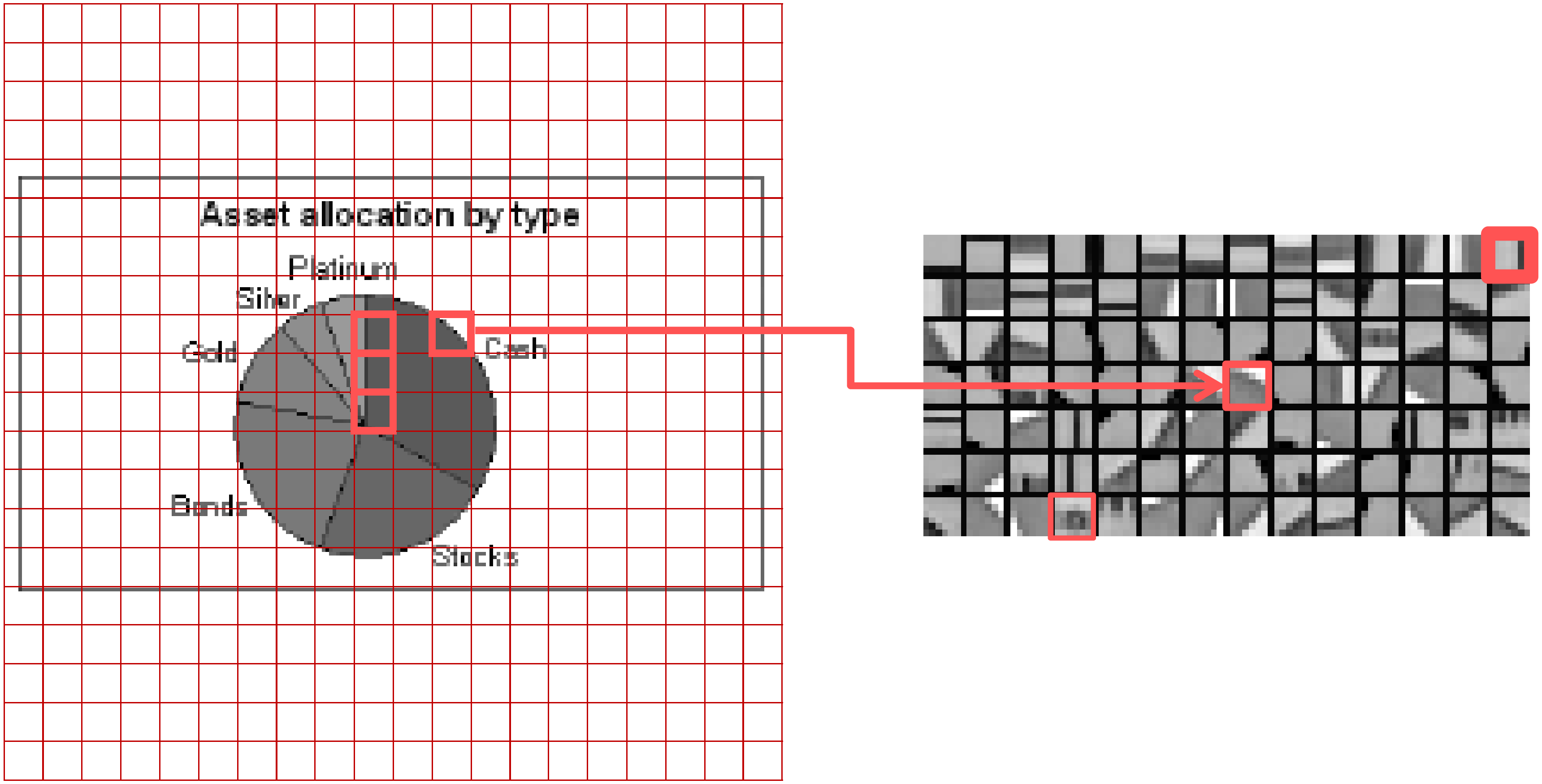


Chart Image Classification

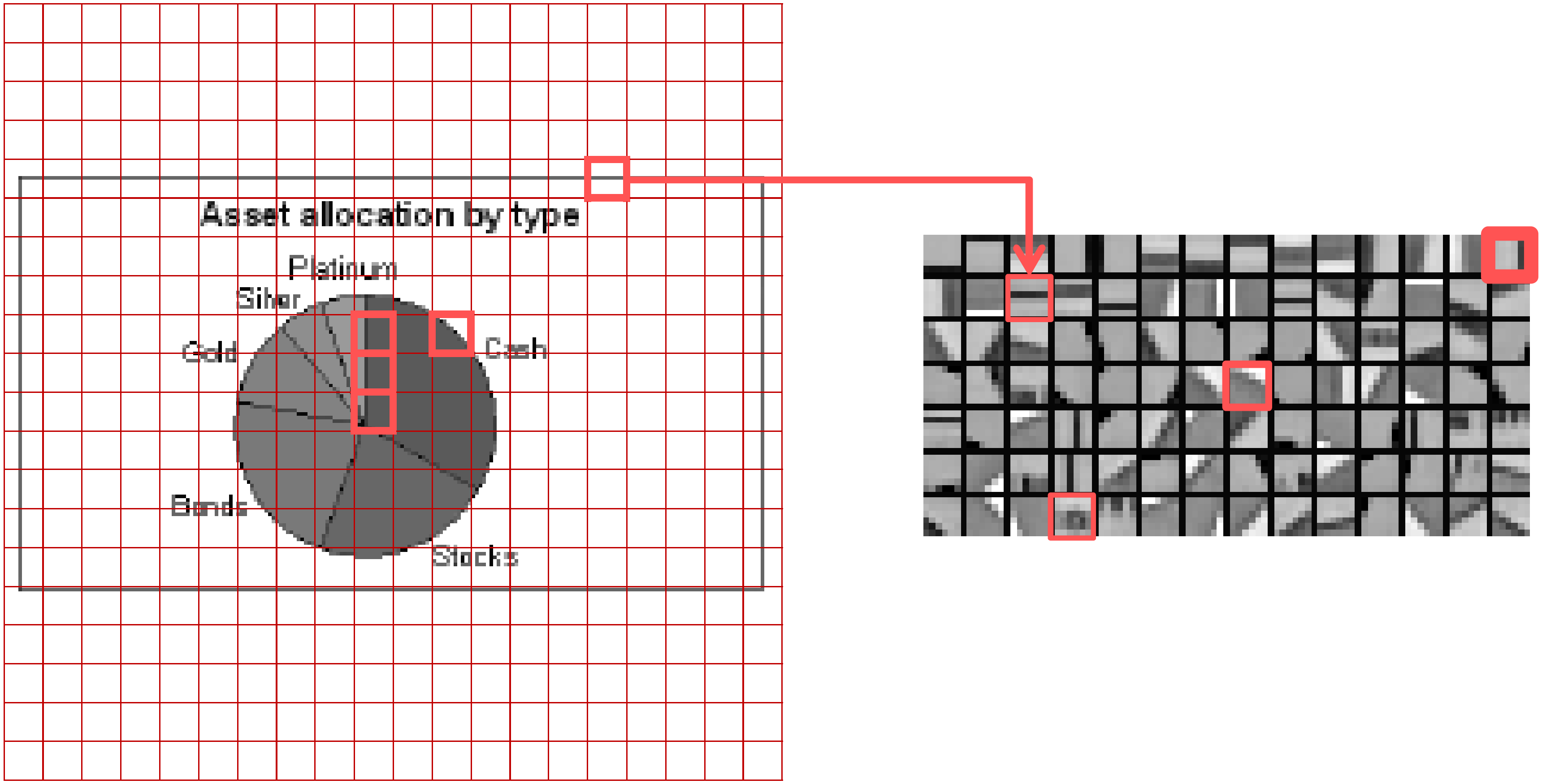


Chart Image Classification

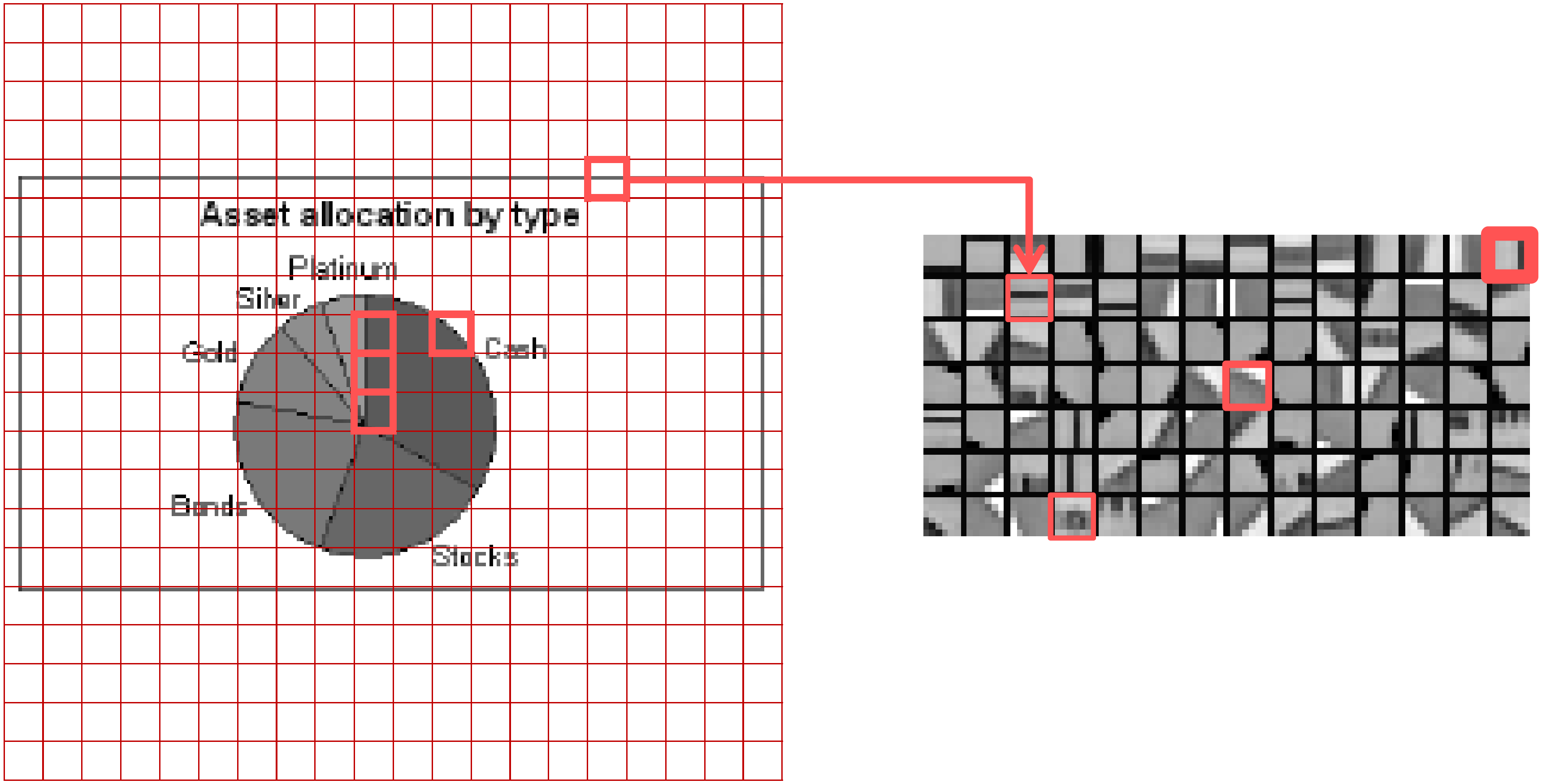


Chart Image Classification

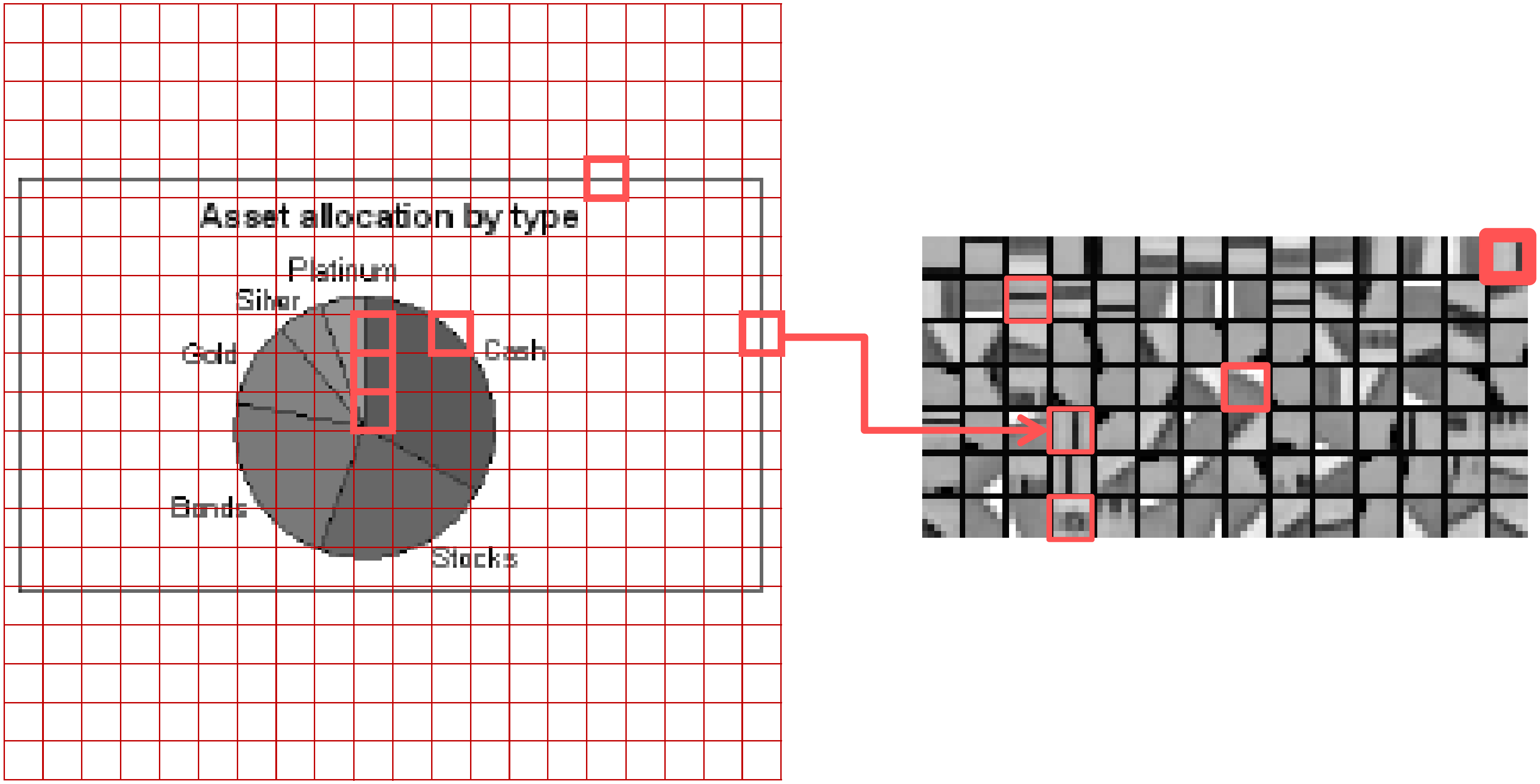
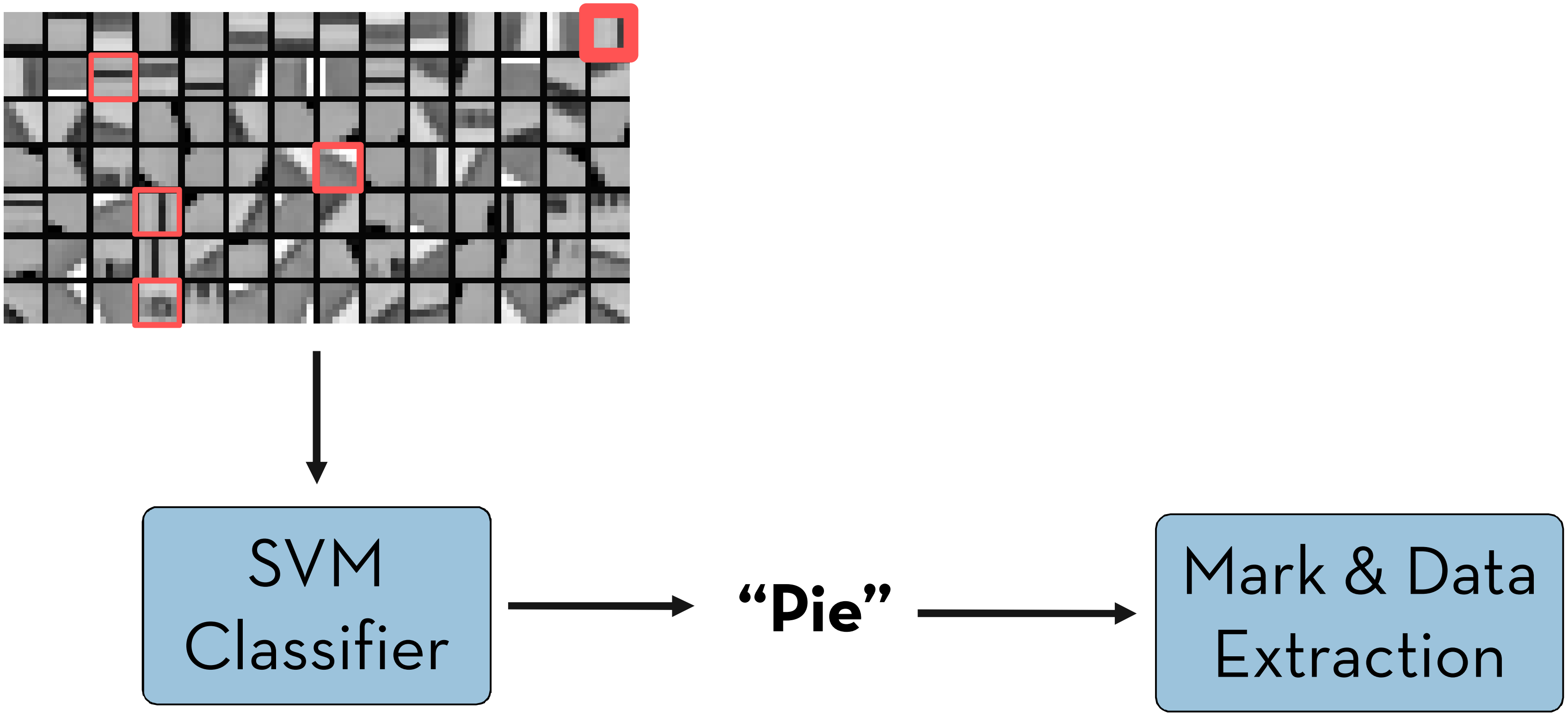


Chart Image Classification



SVM Classifier

“Pie”

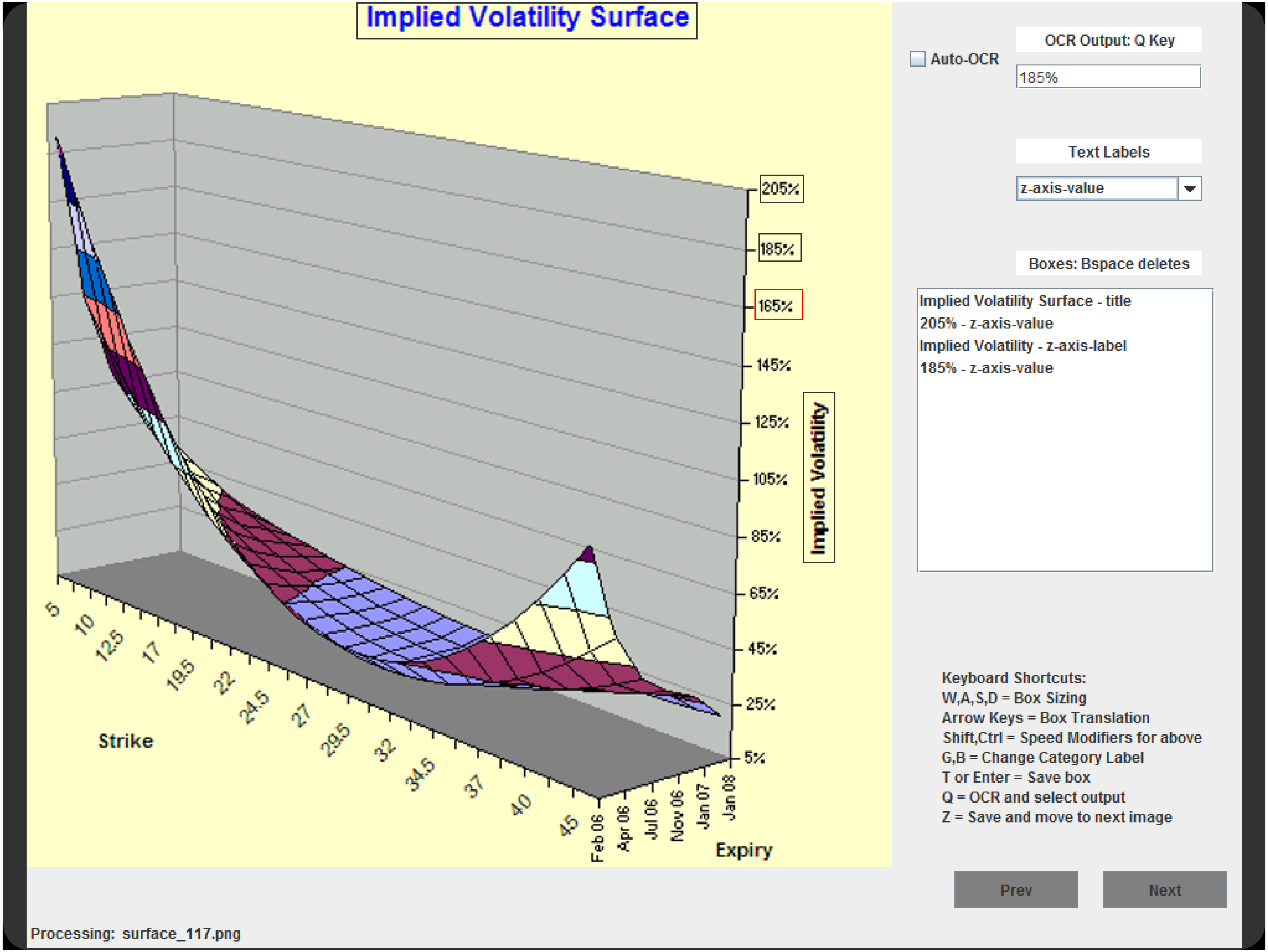
Mark & Data Extraction

Classification Results

Prasad Corpus: 667 images, 5 chart types

5-way Classifier	**Accuracy**
Prasad et al.	84%
ReVision Image Features	88%

What about text features?



□ Auto-OCR

OCR Output: Q Key

185%

Text Labels

z-axis-value

Boxes: Bspace deletes

Implied Volatility Surface - title
205% - z-axis-value
Implied Volatility - z-axis-label
185% - z-axis-value

Keyboard Shortcuts:
W,A,S,D = Box Sizing
Arrow Keys = Box Translation
Shift,Ctrl = Speed Modifiers for above
G,B = Change Category Label
T or Enter = Save box
Q = OCR and select output
Z = Save and move to next image

Processing: surface_117.png

Textual Features

Histograms of:

- Text pixels per image region (8 x 8 grid)
- Text label geometry (position, width, height)
- Angles between label pairs
- Distances between label pairs

Each histogram is normalized and then concatenated to the final feature vector.

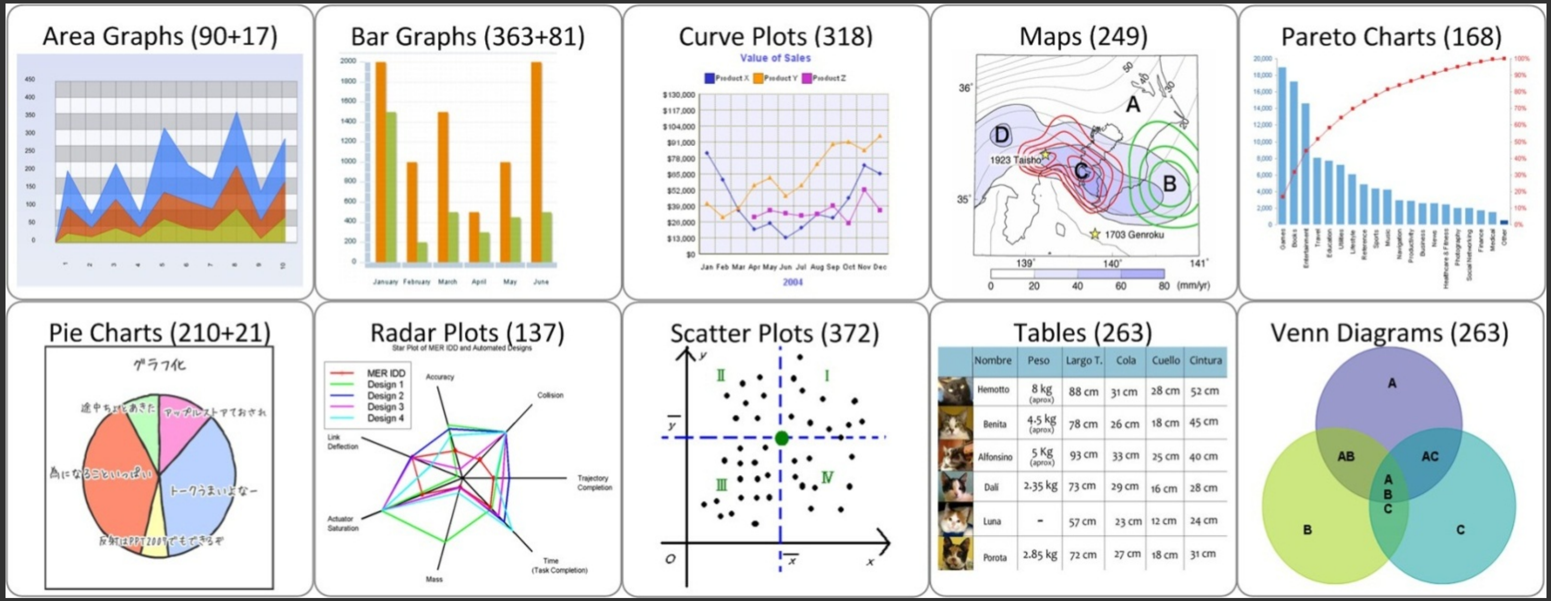
Classification Results

Prasad Corpus: 667 images, 5 chart types

5-way Classifier	Accuracy
Prasad et al.	84%
ReVision Image Features	88%
ReVision Text Features	66%
ReVision All Features	90%

Binary Classifiers	Accuracy
ReVision All Features	96%

Classification Results



Area Graphs (90+17)

Bar Graphs (363+81)

Curve Plots (318)

Maps (249)

Pareto Charts (168)

Pie Charts (210+21)

Radar Plots (137)

Scatter Plots (372)

Tables (263)

Venn Diagrams (263)

Extended Corpus: 2,601 images, 10 chart types

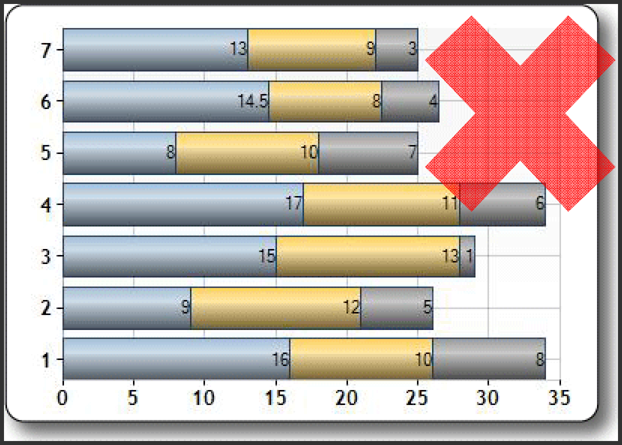
Classifier	Accuracy
ReVision Image Feat., 10-way	80%
ReVision Image Feat., Binary	96%

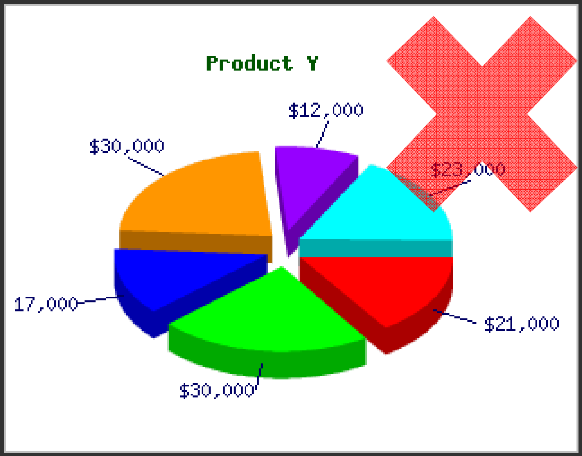
1. **Classification**
2. Mark & Data Extraction
3. Automated Redesign

1. Classification
2. **Mark & Data Extraction**
3. Automated Redesign

Initial Assumptions

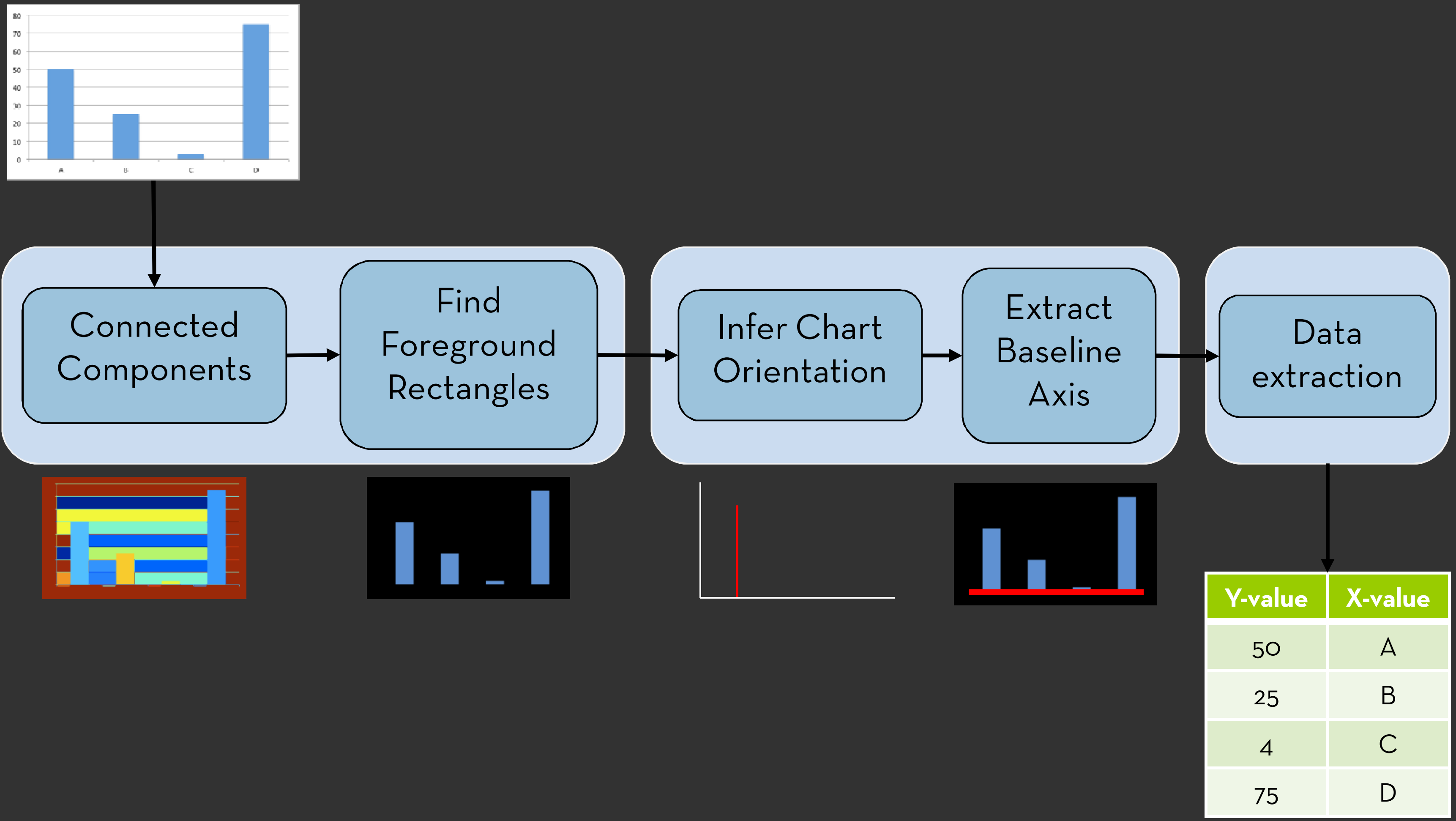
No 3D effects, stacked bars, or exploded pies





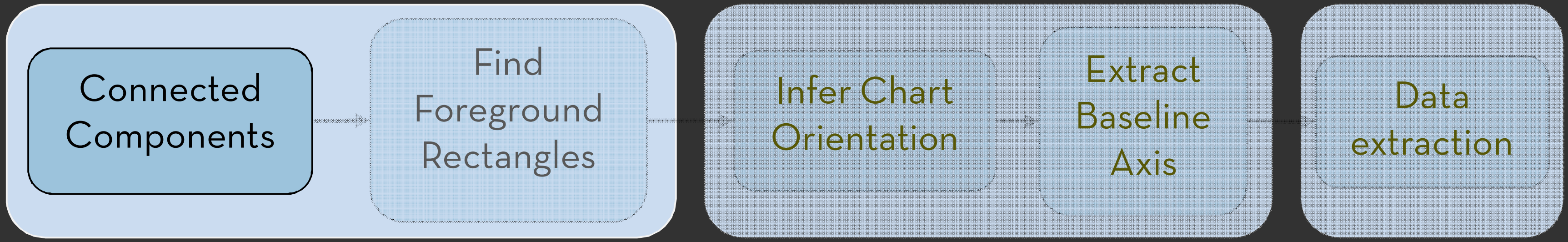
Marks are shaded with (nearly) constant colors

Bar chart extraction



Y-value	X-value
50	A
25	B
4	C
75	D

Bar charts: Bar extraction



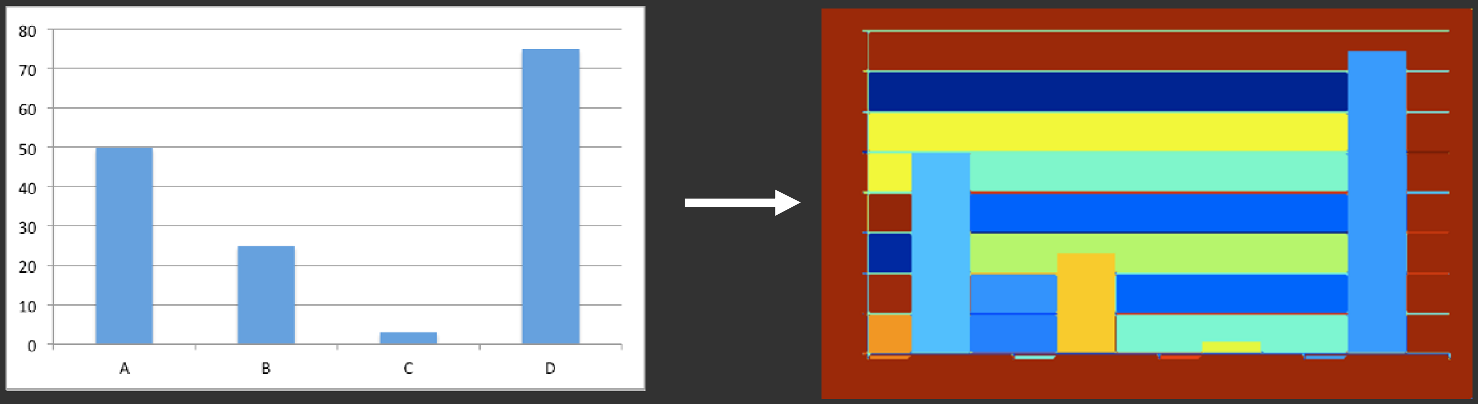
Connected Components

Find Foreground Rectangles

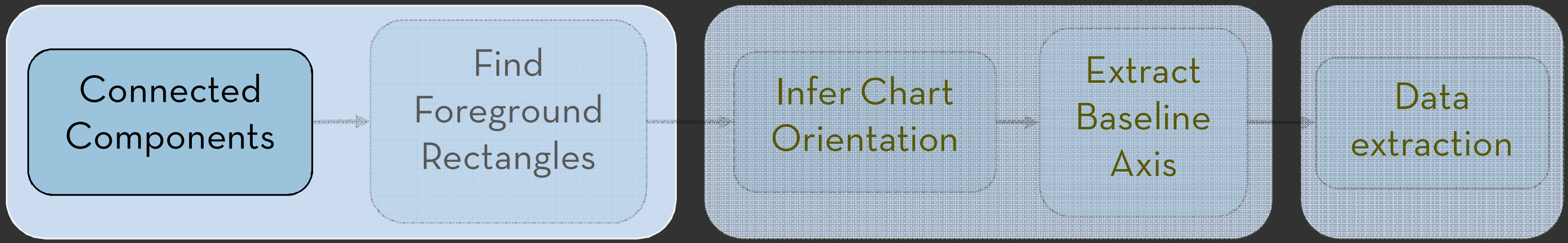
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Bar extraction



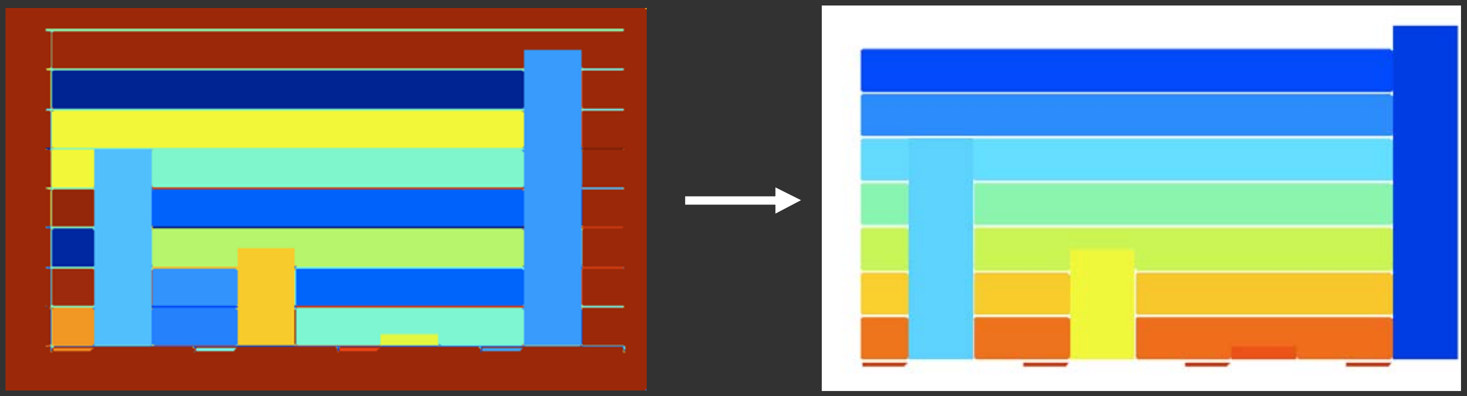
Connected Components

Find Foreground Rectangles

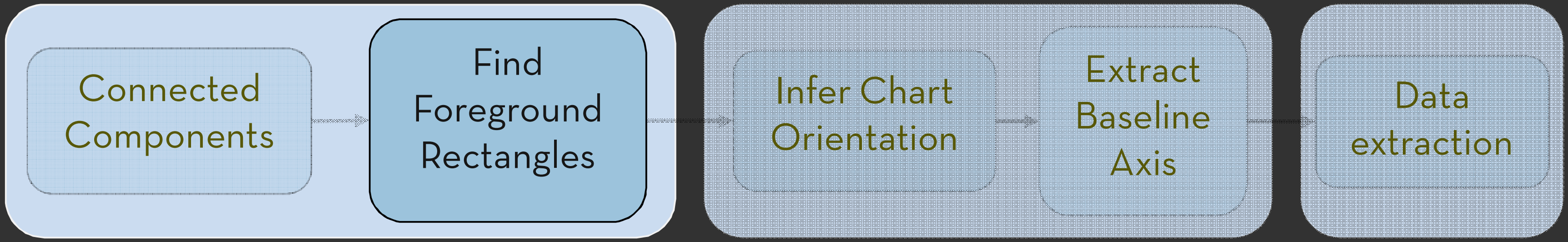
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Bar extraction



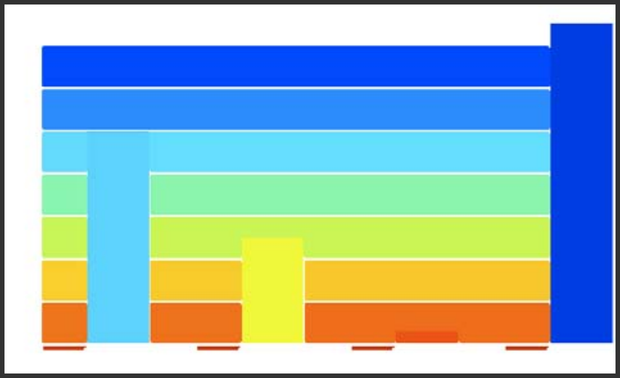
Connected Components

Find Foreground Rectangles

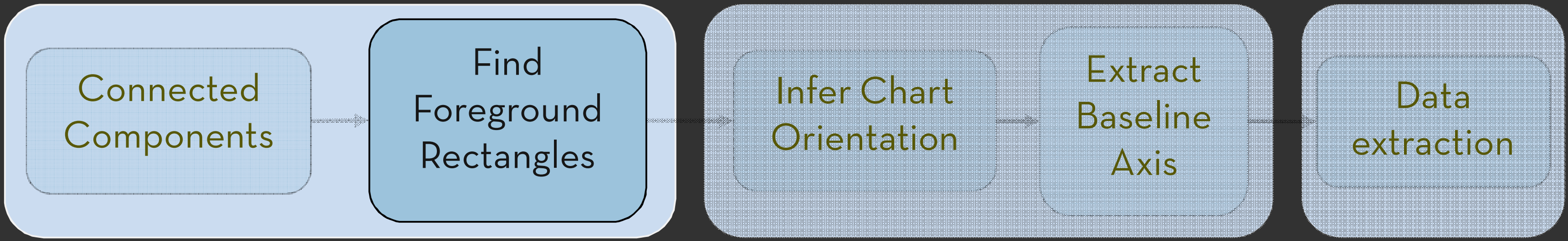
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Bar extraction



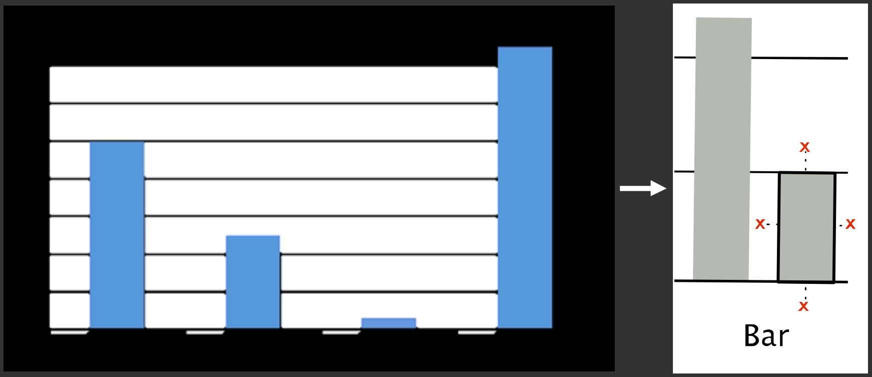
Connected Components

Find Foreground Rectangles

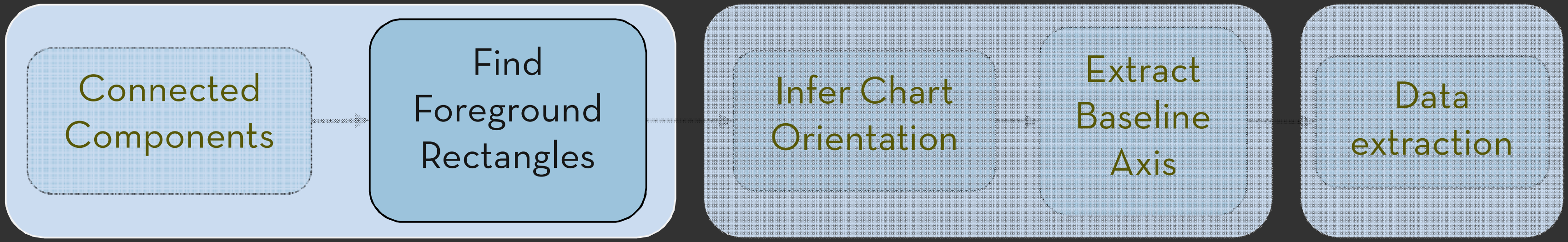
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Bar extraction

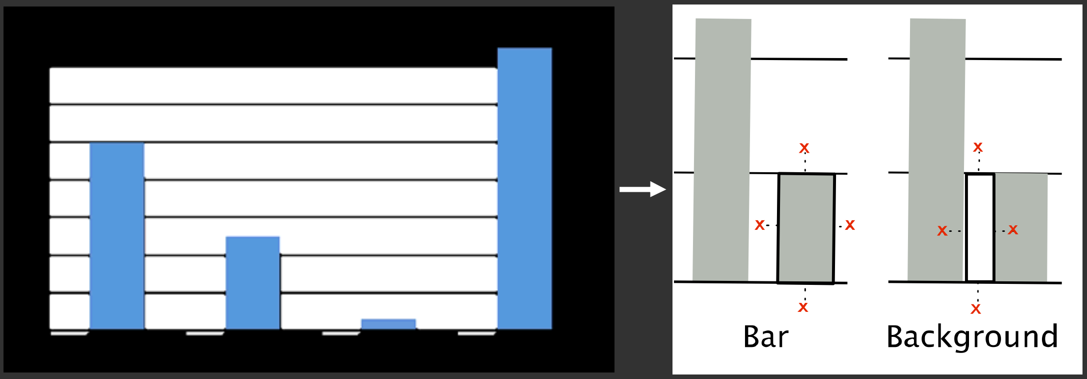

Connected Components

Find Foreground Rectangles

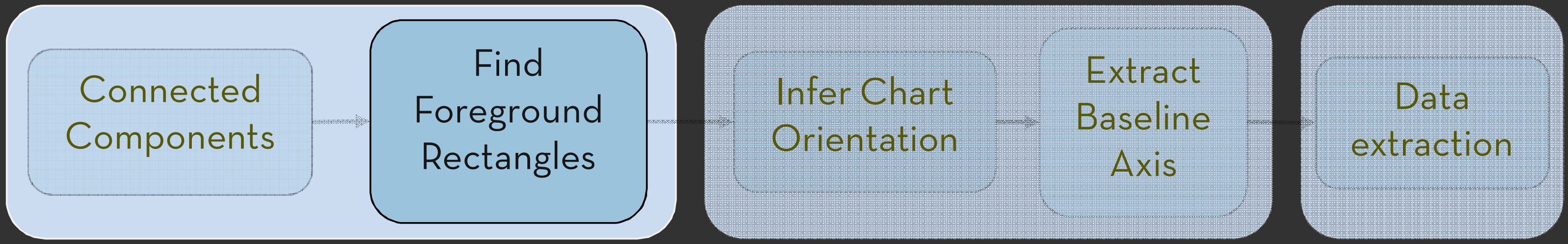
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Bar extraction



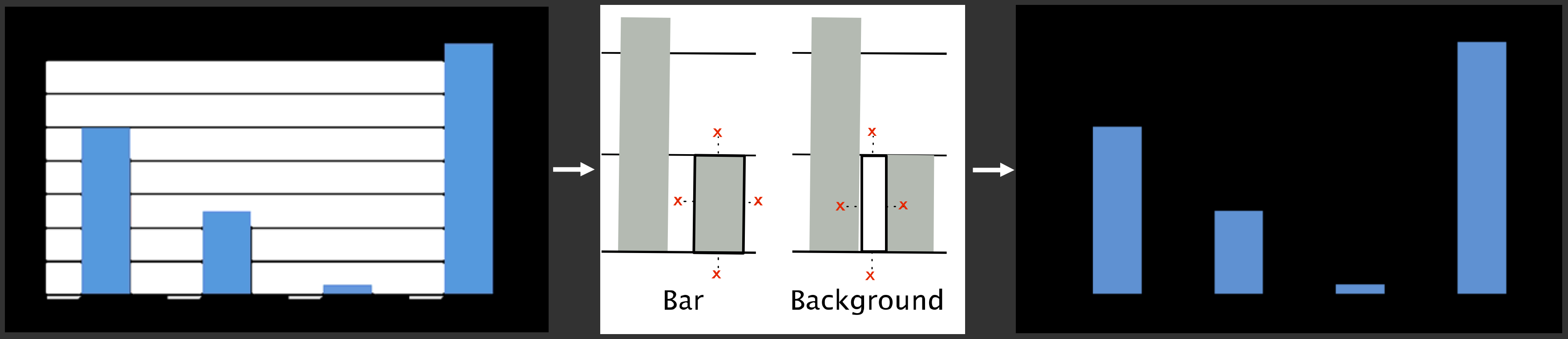
Connected Components

Find Foreground Rectangles

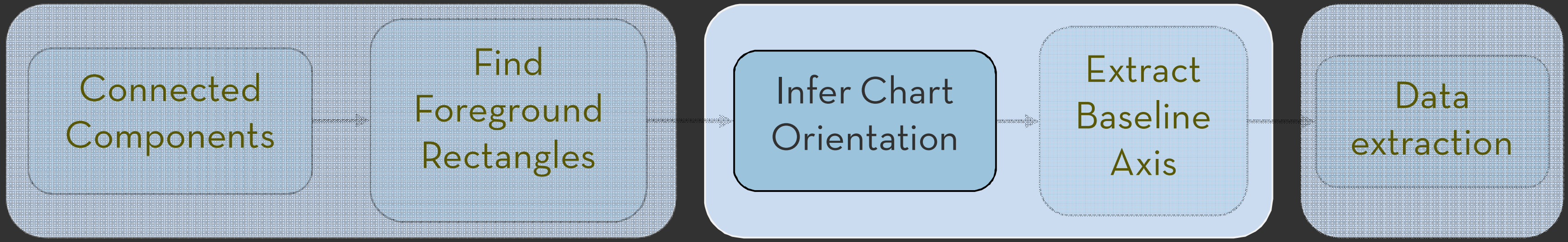
Infer Chart Orientation

Extract Baseline Axis

Data extraction



Bar charts: Axis extraction



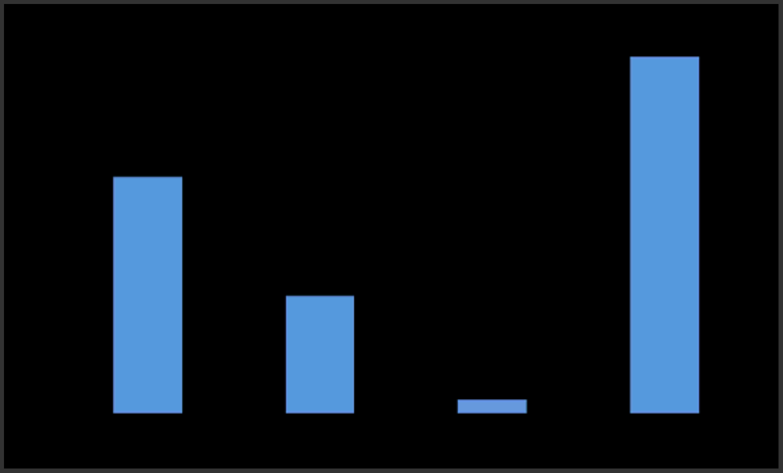
Connected Components

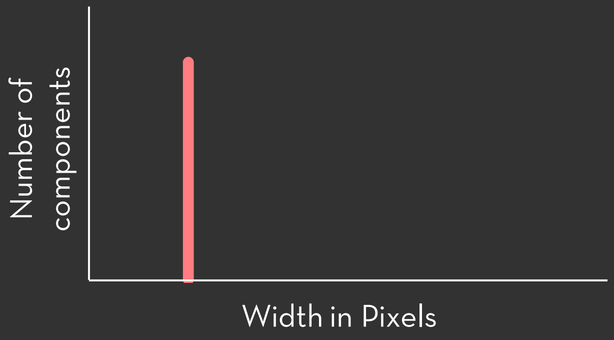
Find Foreground Rectangles

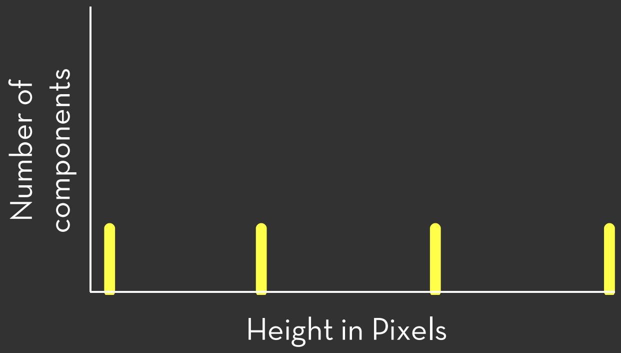
Infer Chart Orientation

Extract Baseline Axis

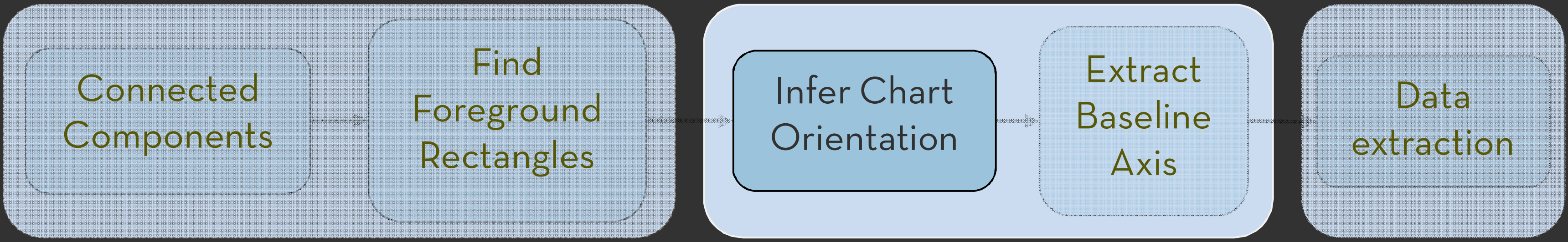
Data extraction







Bar charts: Axis extraction



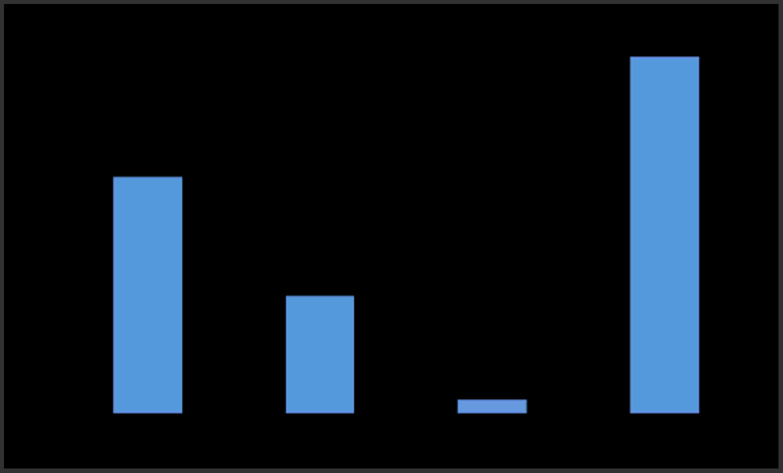
Connected Components

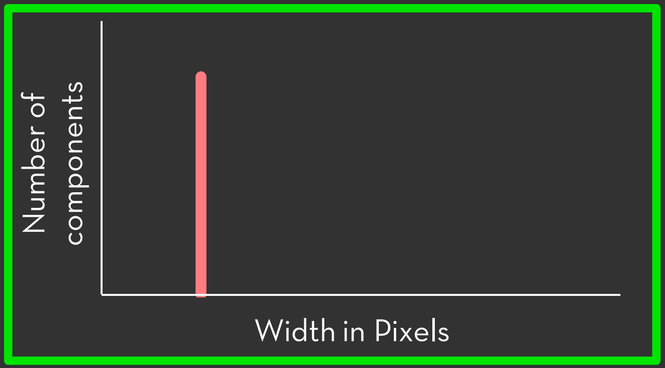
Find Foreground Rectangles

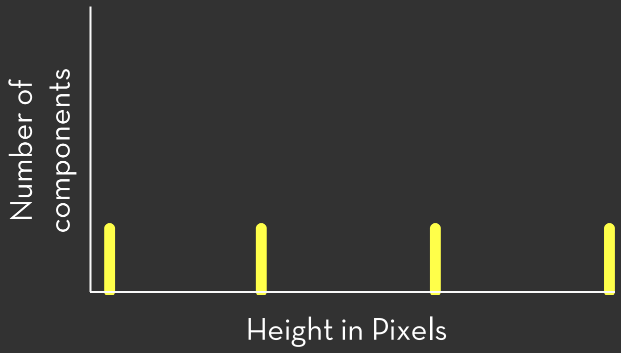
Infer Chart Orientation

Extract Baseline Axis

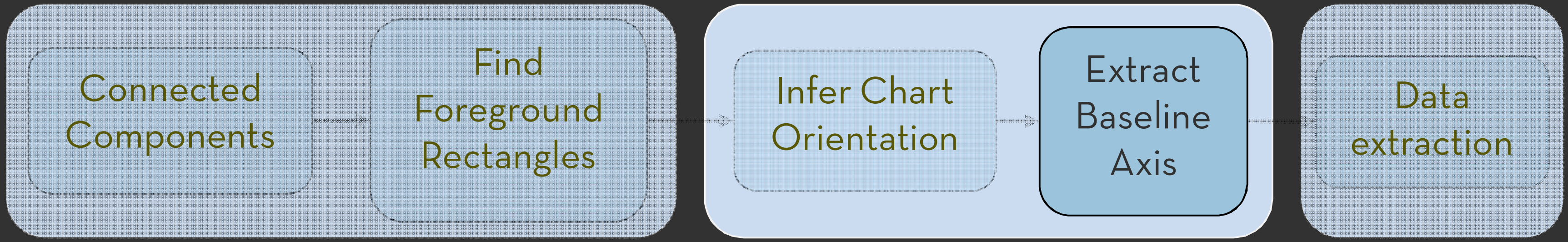
Data extraction

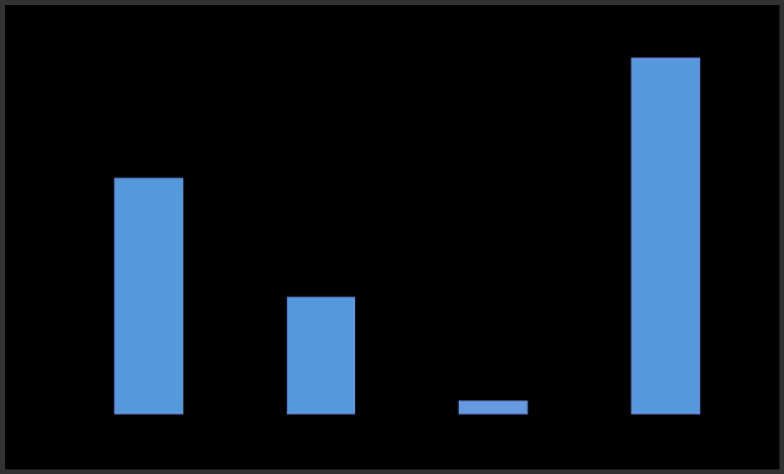


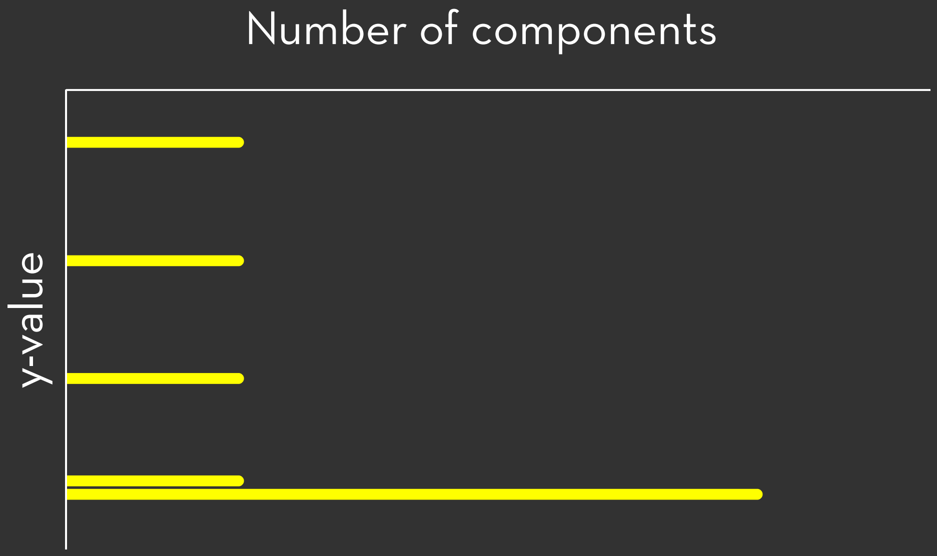




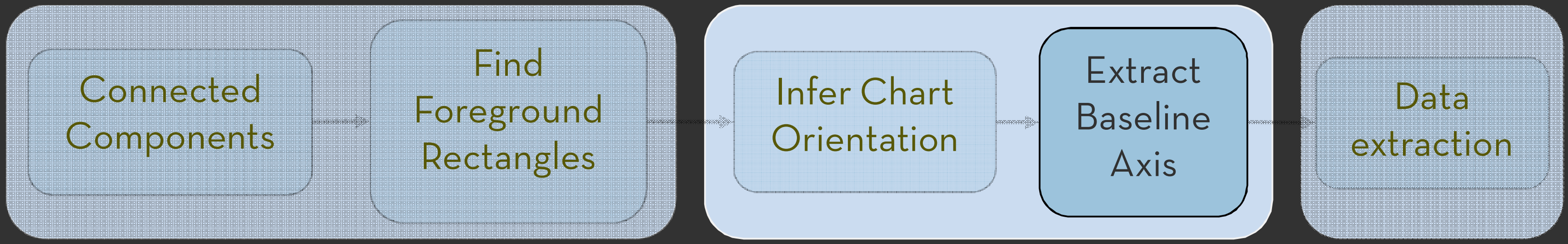
Bar charts: Axis extraction

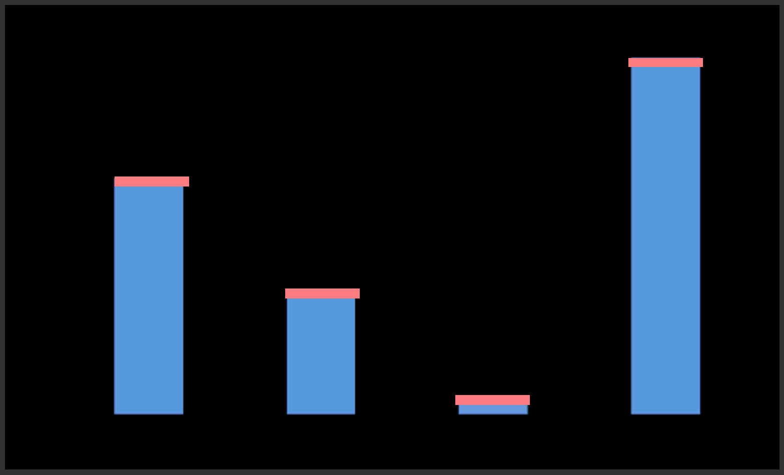


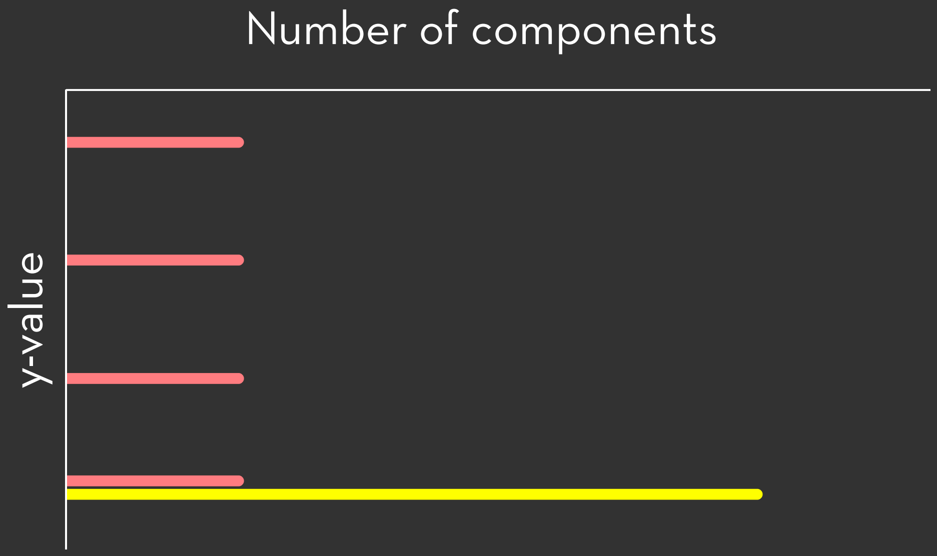




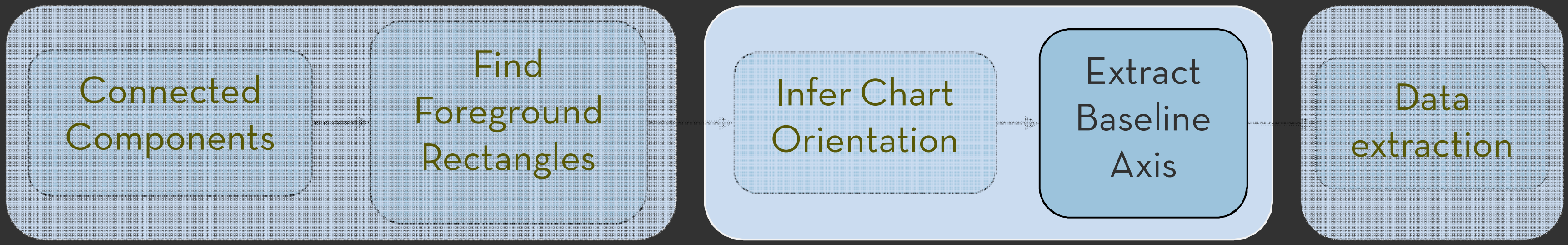
Bar charts: Axis extraction


Connected Components
Find Foreground Rectangles
Infer Chart Orientation
Extract Baseline Axis
Data extraction





Bar charts: Axis extraction



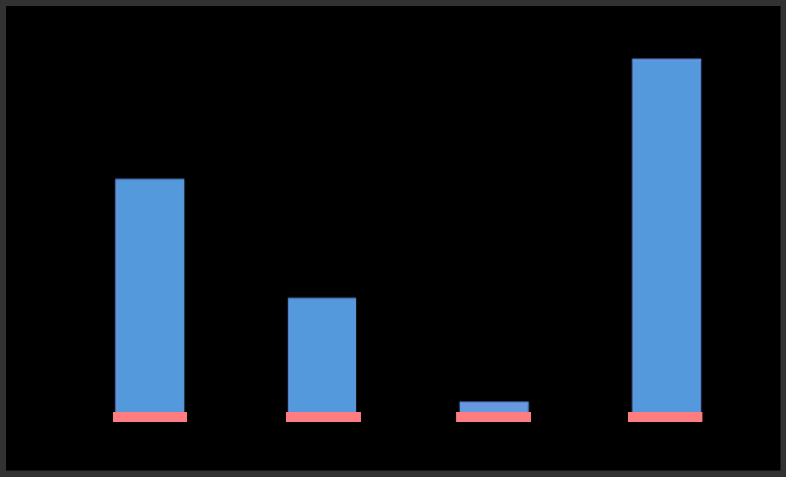
Connected Components

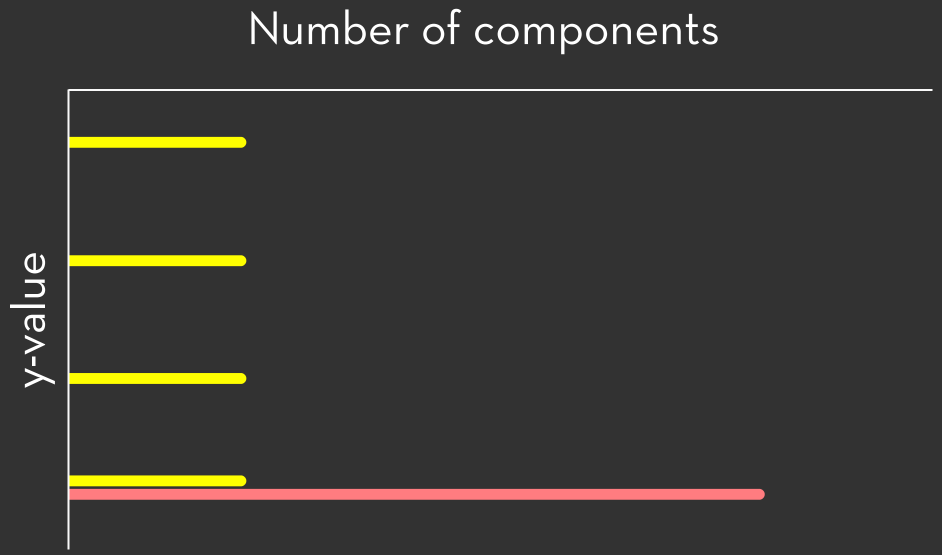
Find Foreground Rectangles

Infer Chart Orientation

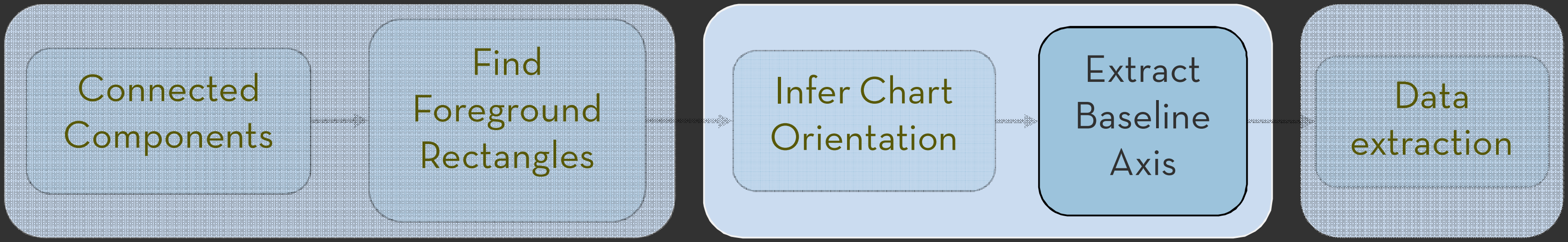
Extract Baseline Axis

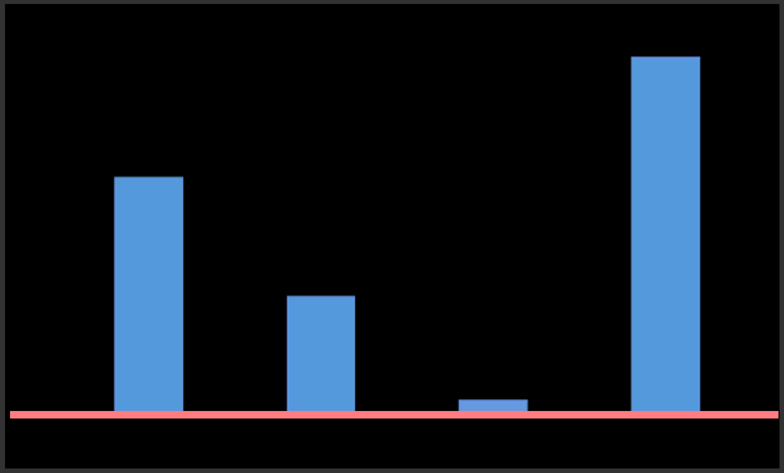
Data extraction

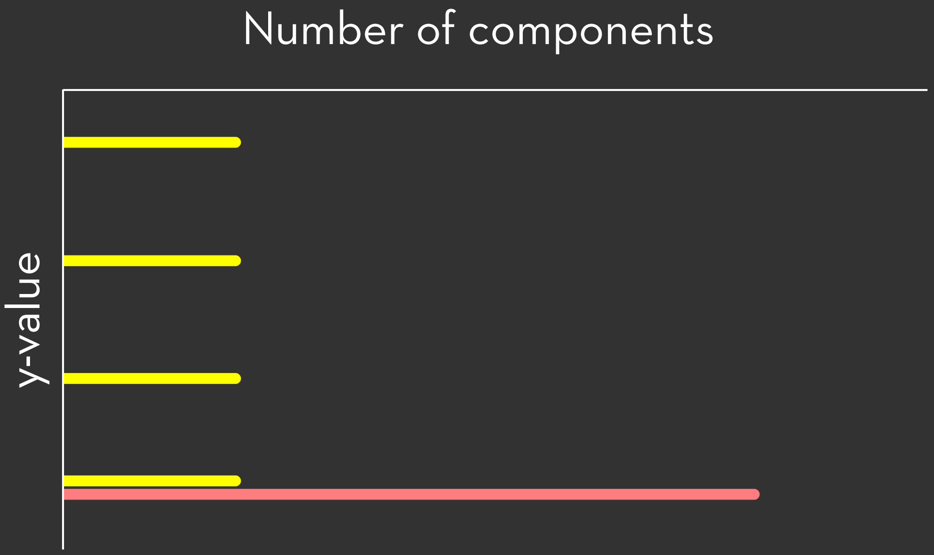




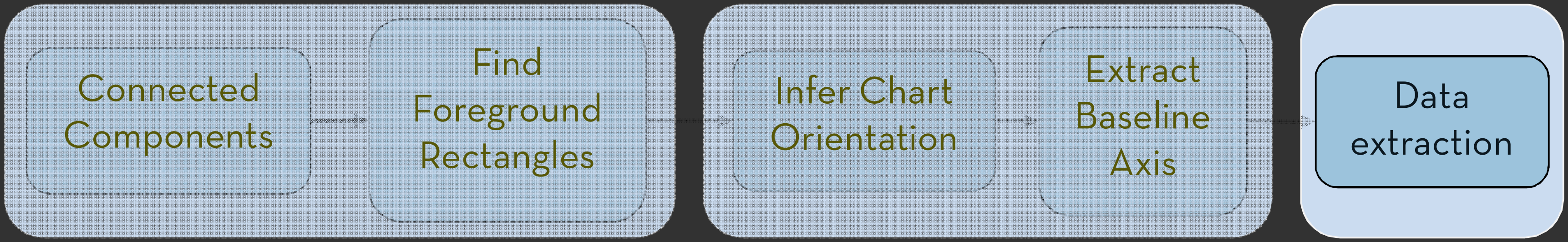
Bar charts: Axis extraction







Bar charts: Data extraction



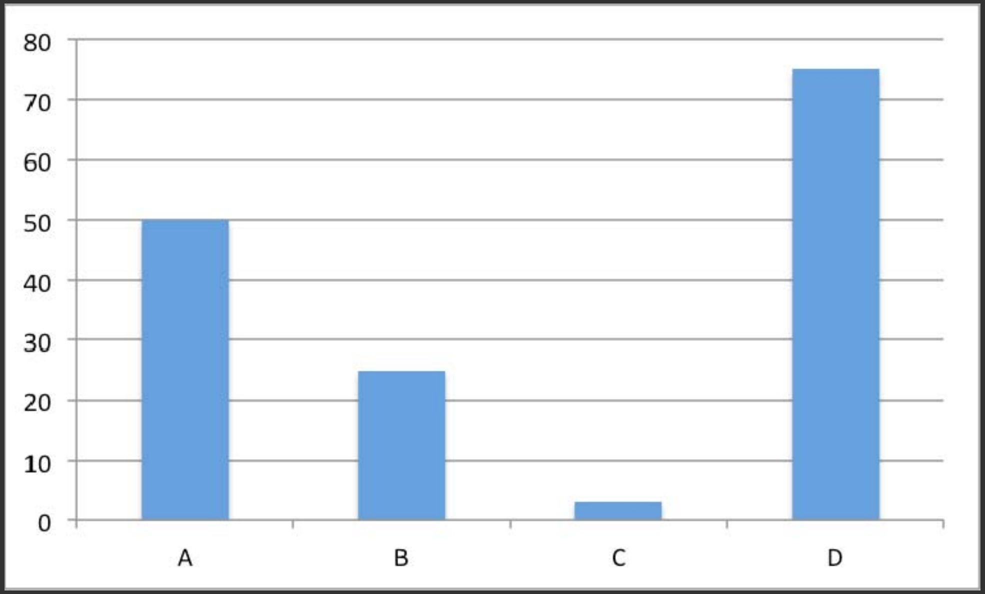
Connected Components

Find Foreground Rectangles

Infer Chart Orientation

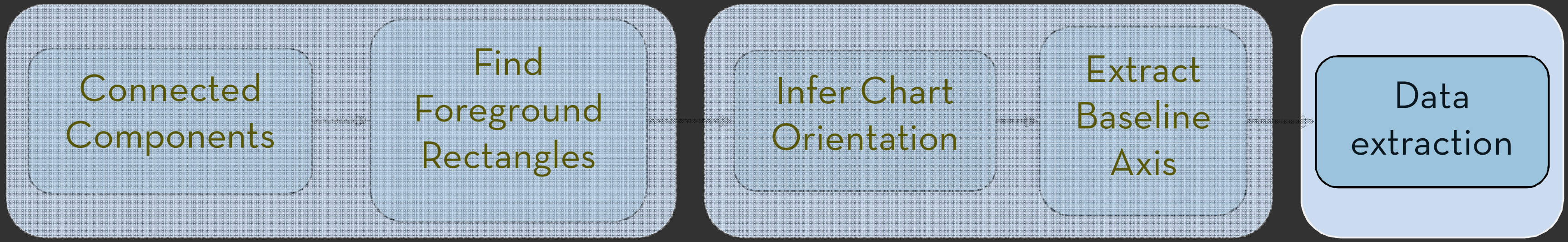
Extract Baseline Axis

Data extraction

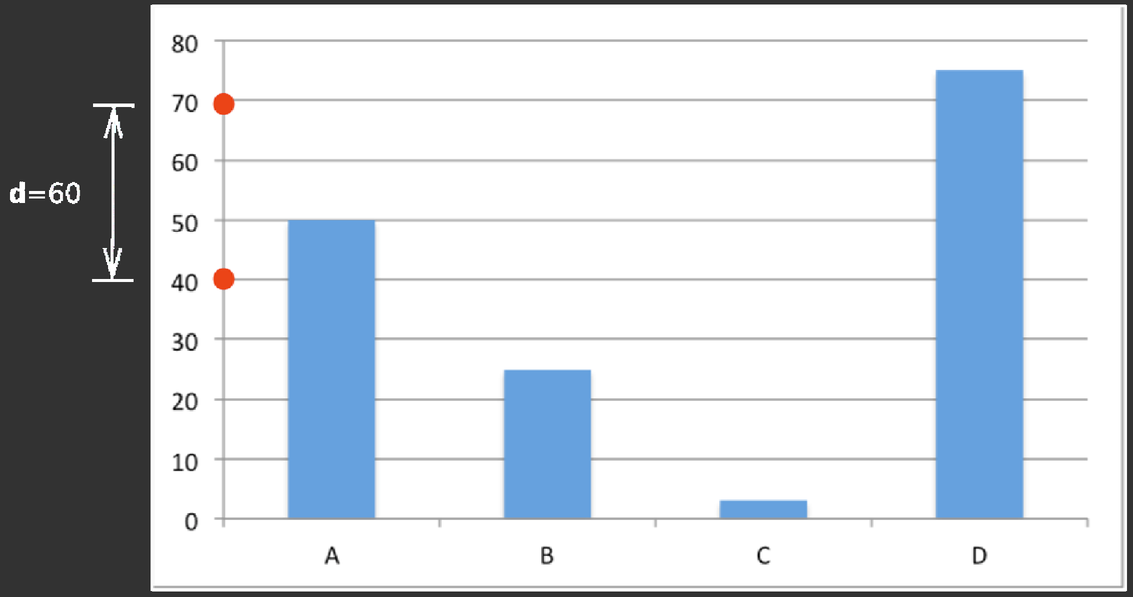


Y-value	X-value

Bar charts: Data extraction



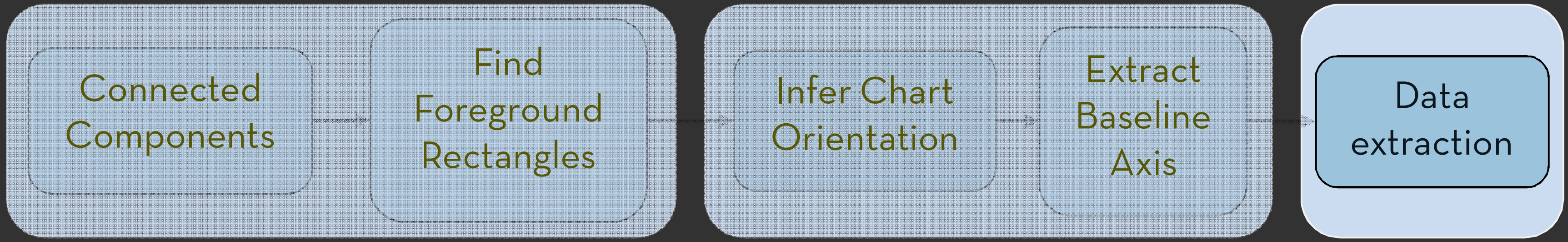
Connected Components → Find Foreground Rectangles → Infer Chart Orientation → Extract Baseline Axis → Data extraction



Y-value	X-value

Scale: 60 / (70-40) = 2 pixels/unit

Bar charts: Data extraction



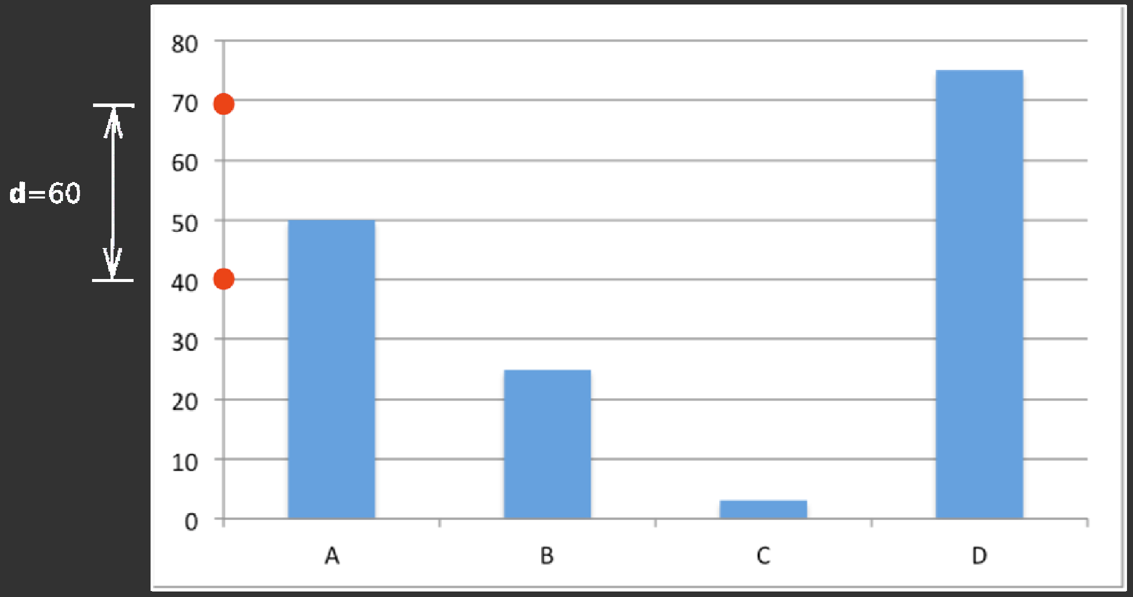
Connected Components

Find Foreground Rectangles

Infer Chart Orientation

Extract Baseline Axis

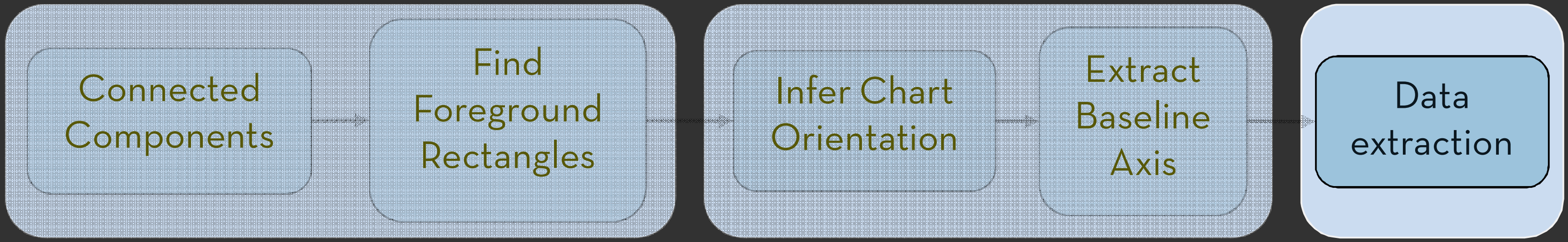
Data extraction



Y-value	X-value
50	
25	
4	
75	

Scale: 60 / (70-40) = 2 pixels/unit

Bar charts: Data extraction

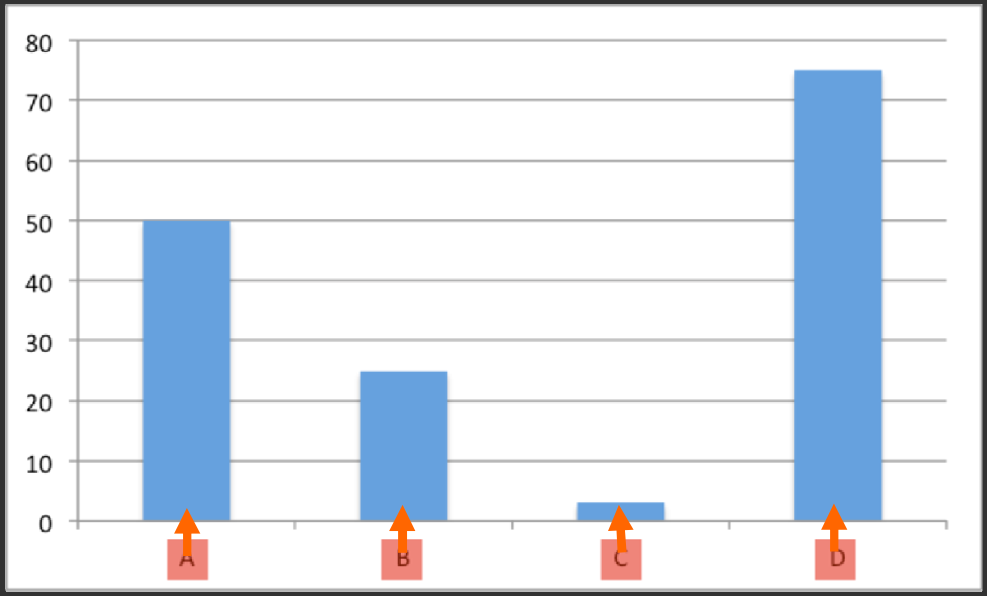

Connected Components

Find Foreground Rectangles

Infer Chart Orientation

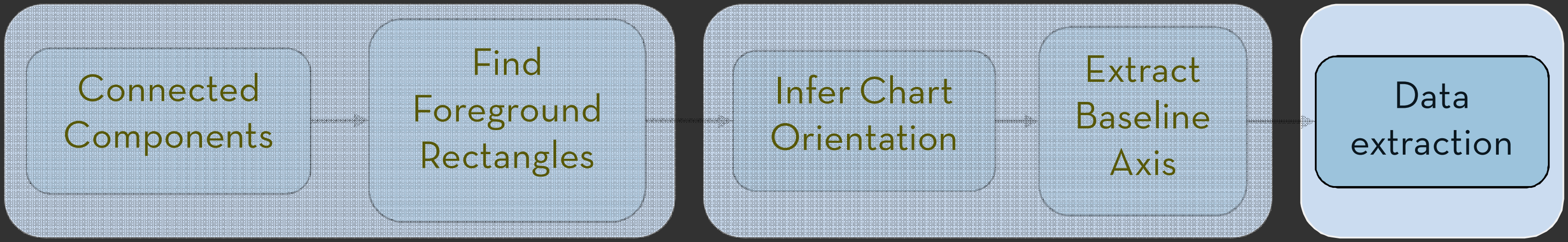
Extract Baseline Axis

Data extraction



Y-value	X-value
50	
25	
4	
75	

Bar charts: Data extraction



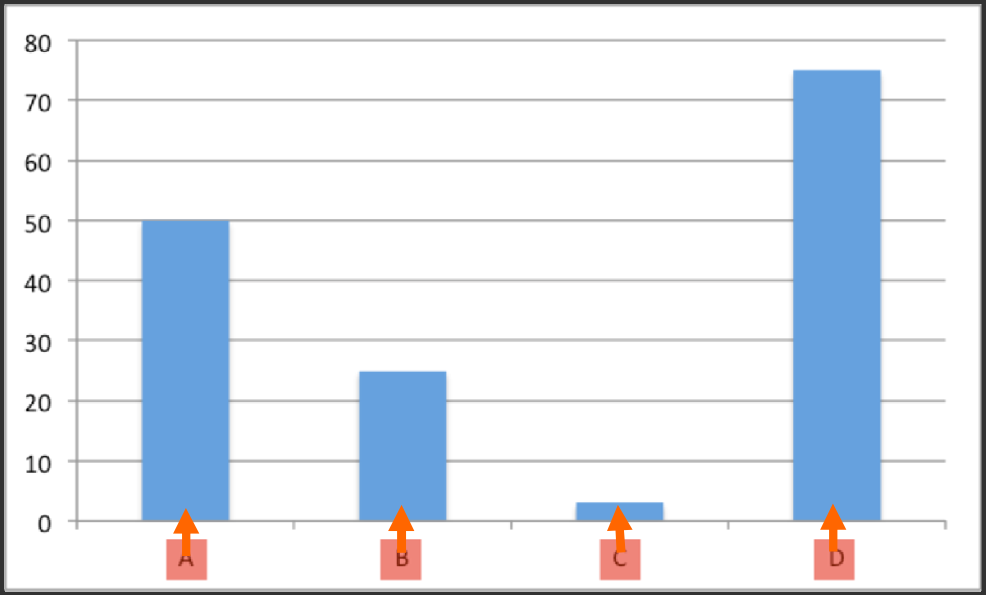
Connected Components

Find Foreground Rectangles

Infer Chart Orientation

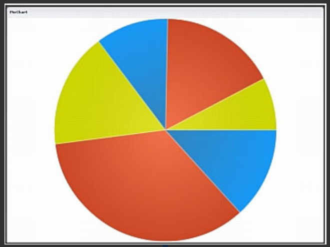
Extract Baseline Axis

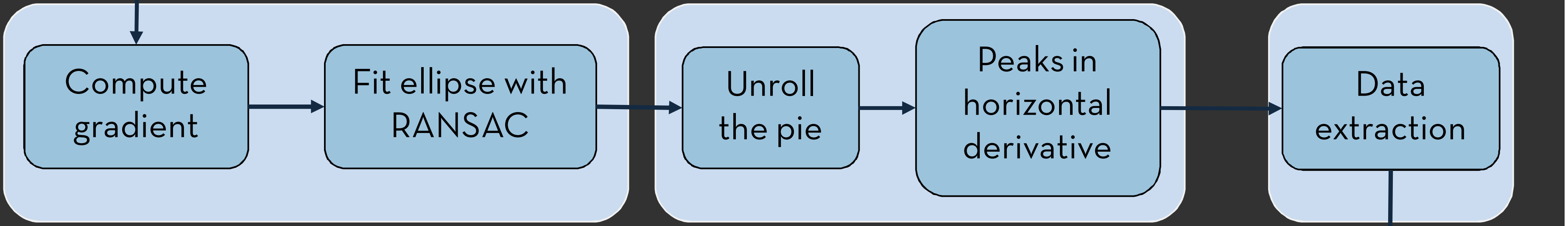
Data extraction



Y-value	X-value
50	A
25	B
4	C
75	D

Pie chart extraction





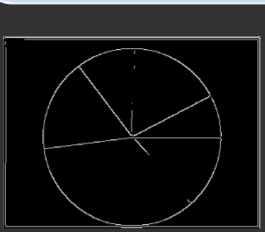
Compute gradient

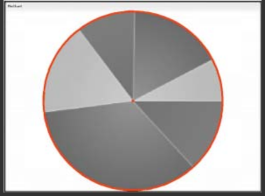
Fit ellipse with RANSAC

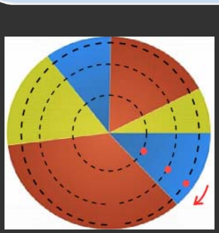
Unroll the pie

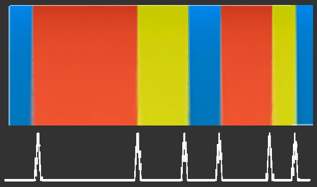
Peaks in horizontal derivative

Data extraction



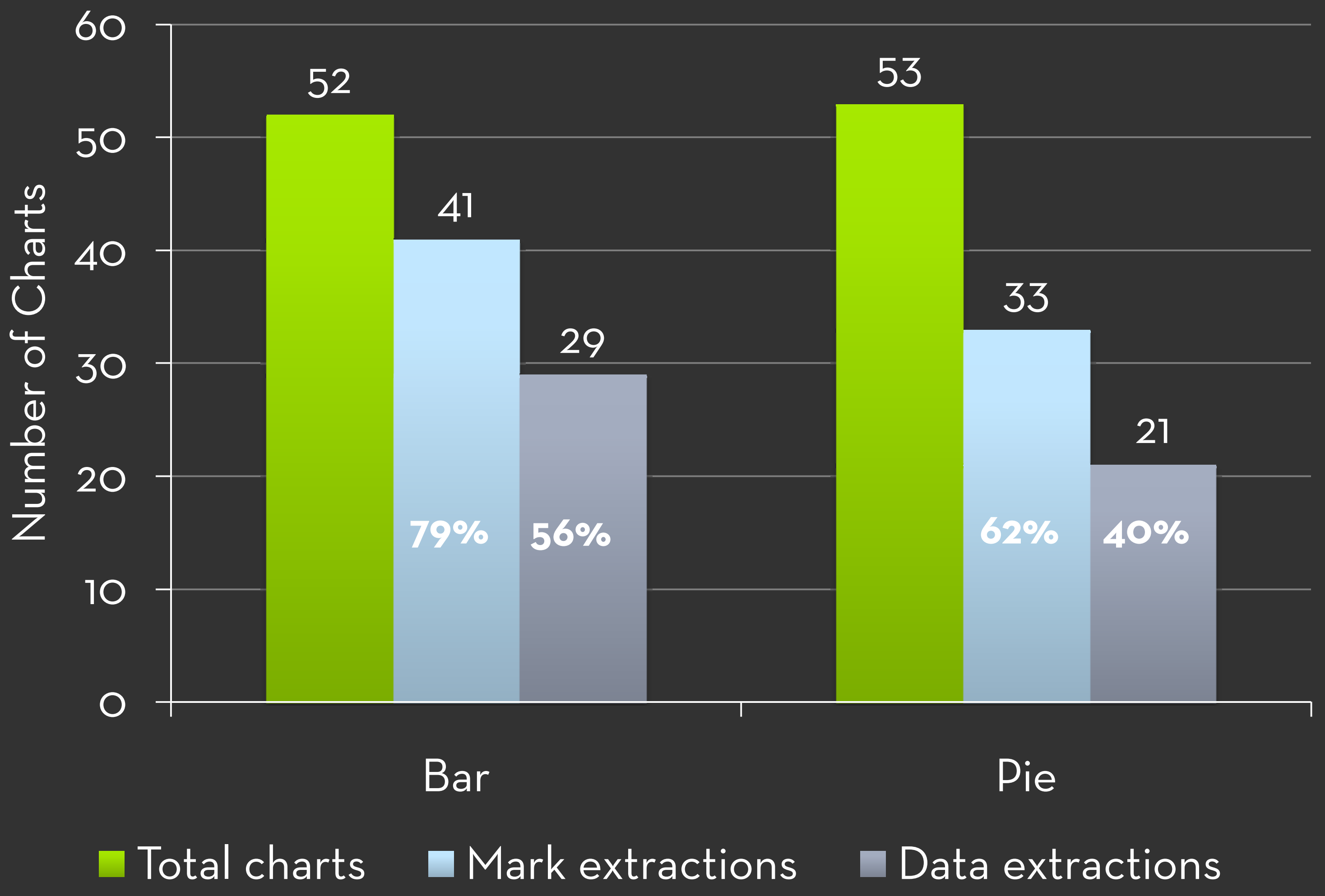






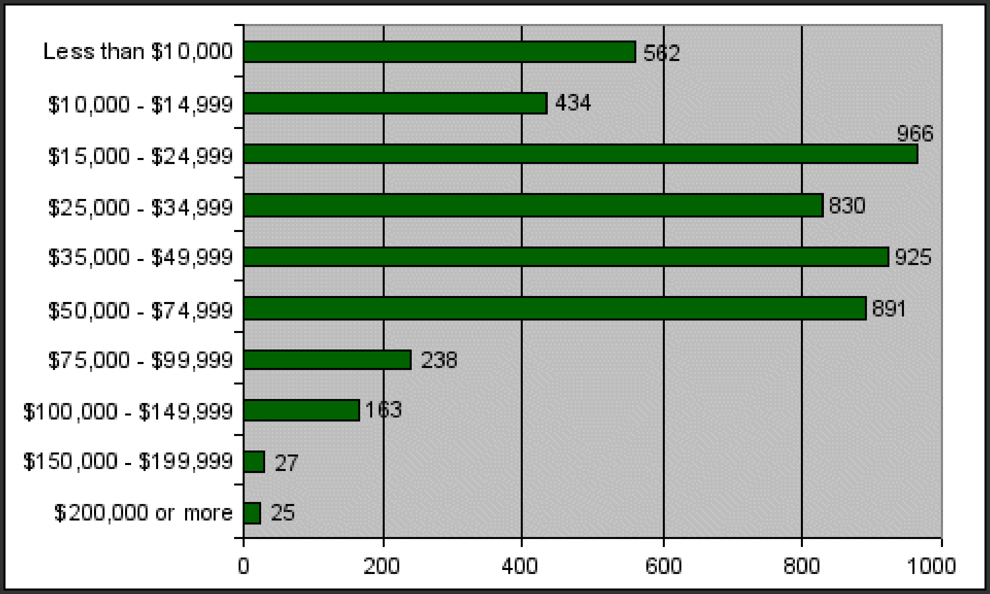
Percentage	Category
22.3	A
22.4	B
10.8	C
5.6	D
5.6	E
33.3	F

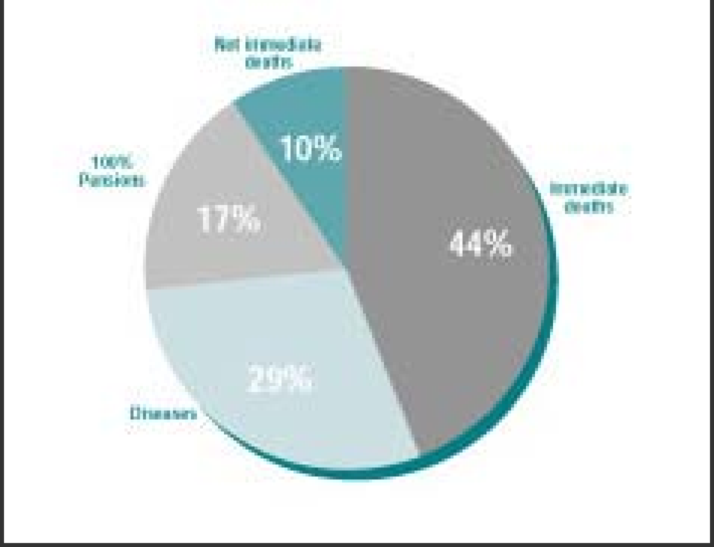
Extraction Results



Average Data Extraction Error

Error = | true_value – est_value | / max_value





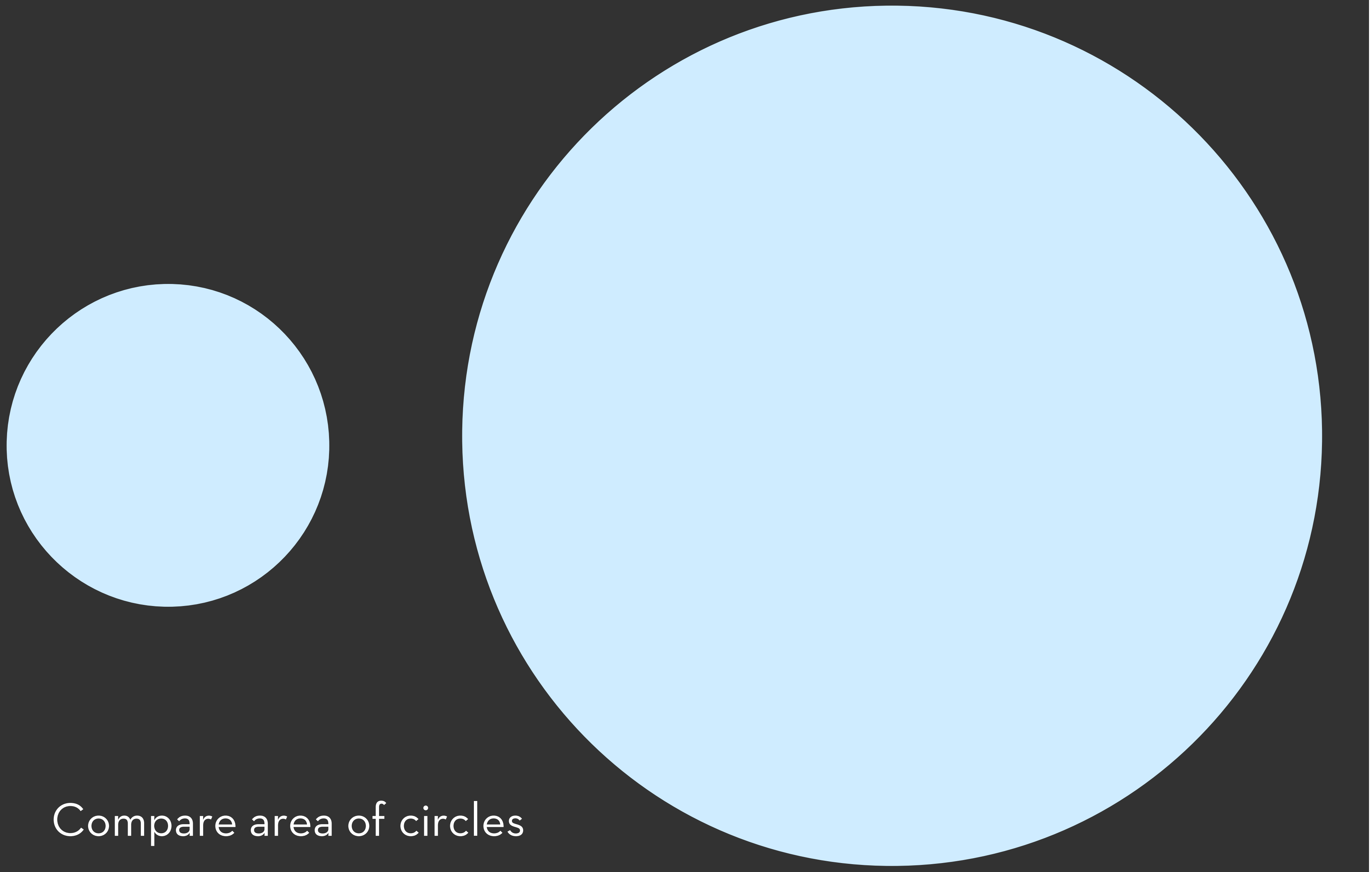
Bar Charts: 0.0093

Pie Charts: 0.0034

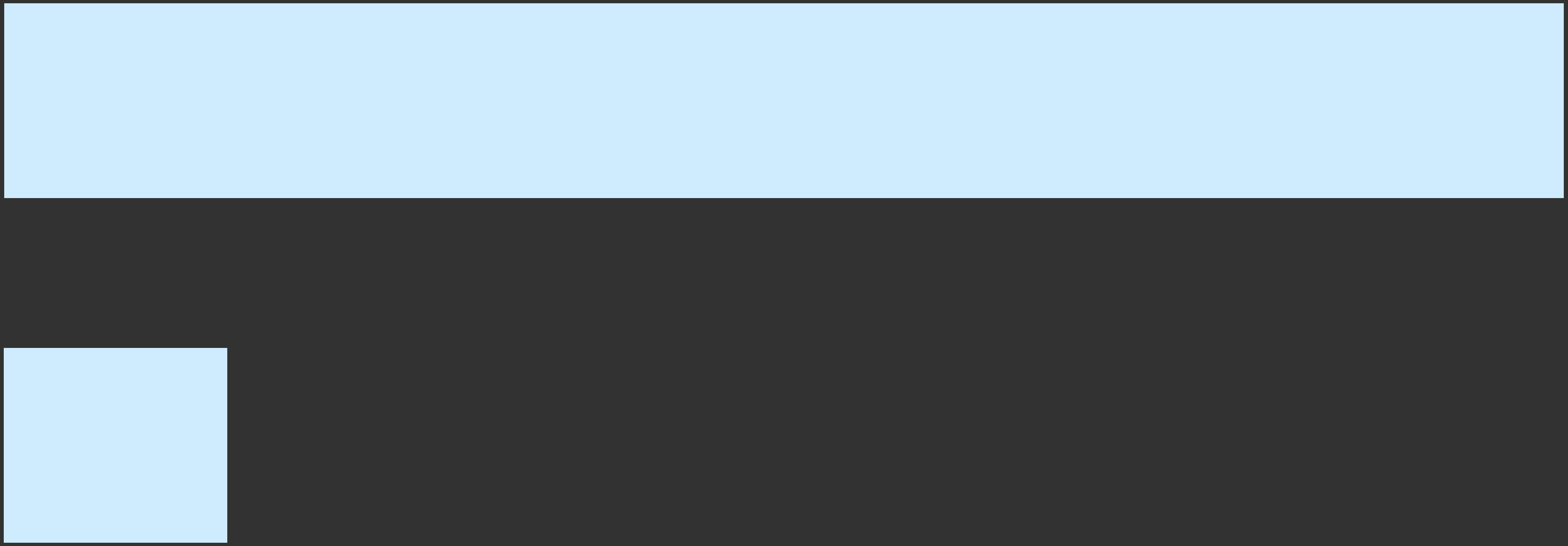
Average chart size: 452 x 342 pixels

1. Classification
2. **Mark & Data Extraction**
3. Automated Redesign

1. Classification
2. Mark & Data Extraction
3. **Automated Redesign**

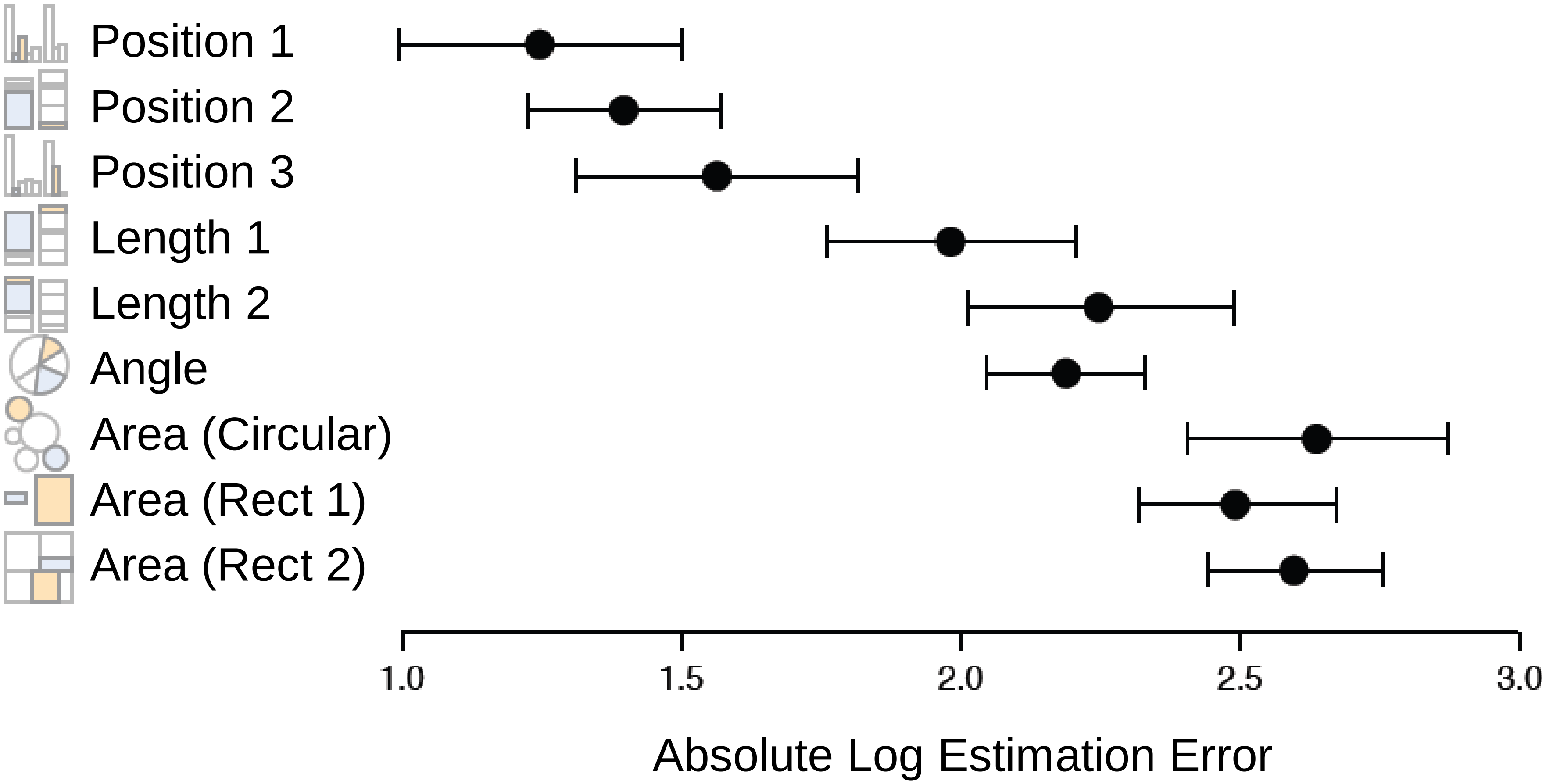


Compare area of circles

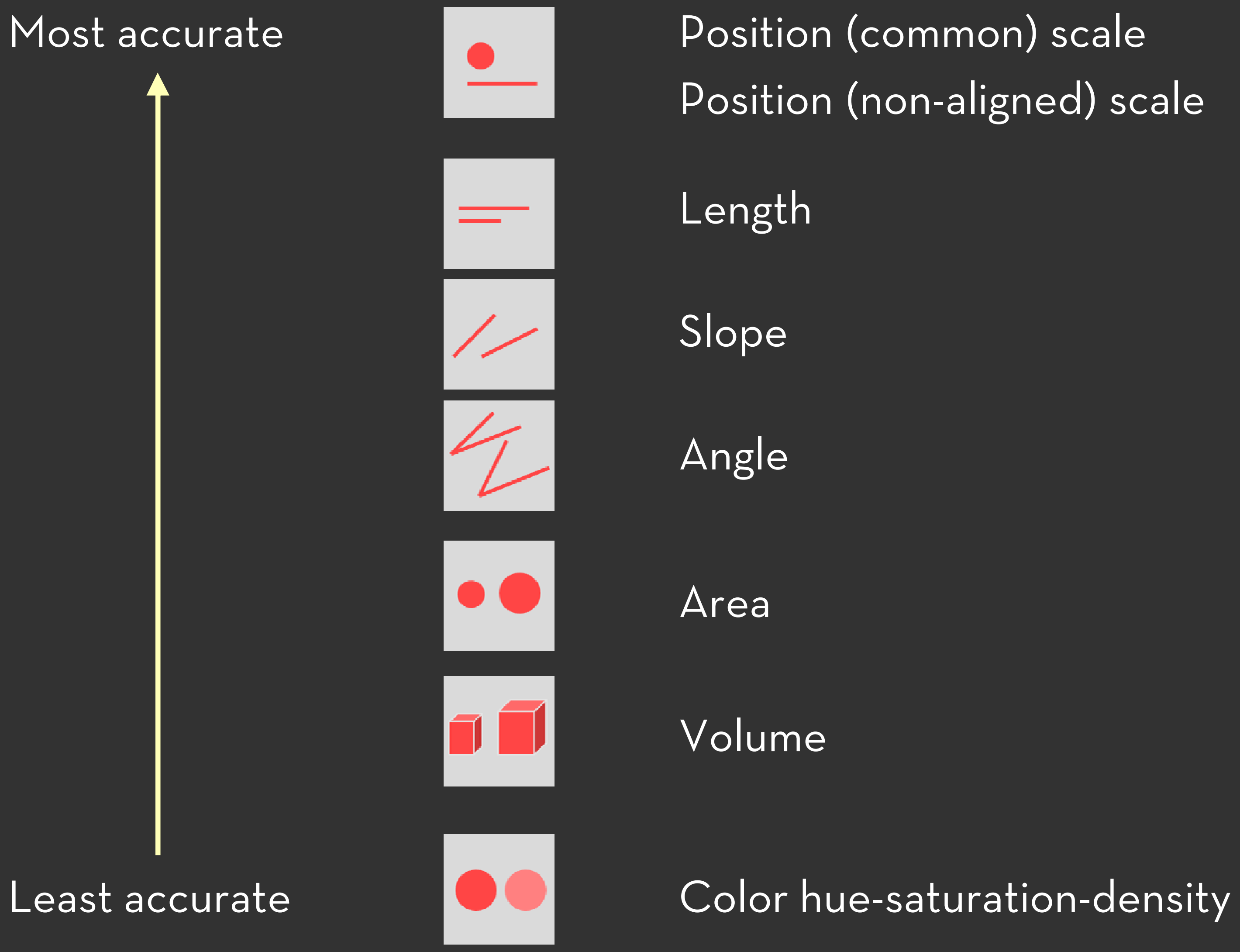


Compare length of bars

Graphical Perception Experiments



Heer & Bostock, 2010



Most accurate

Position (common) scale

Position (non-aligned) scale

Length

Slope

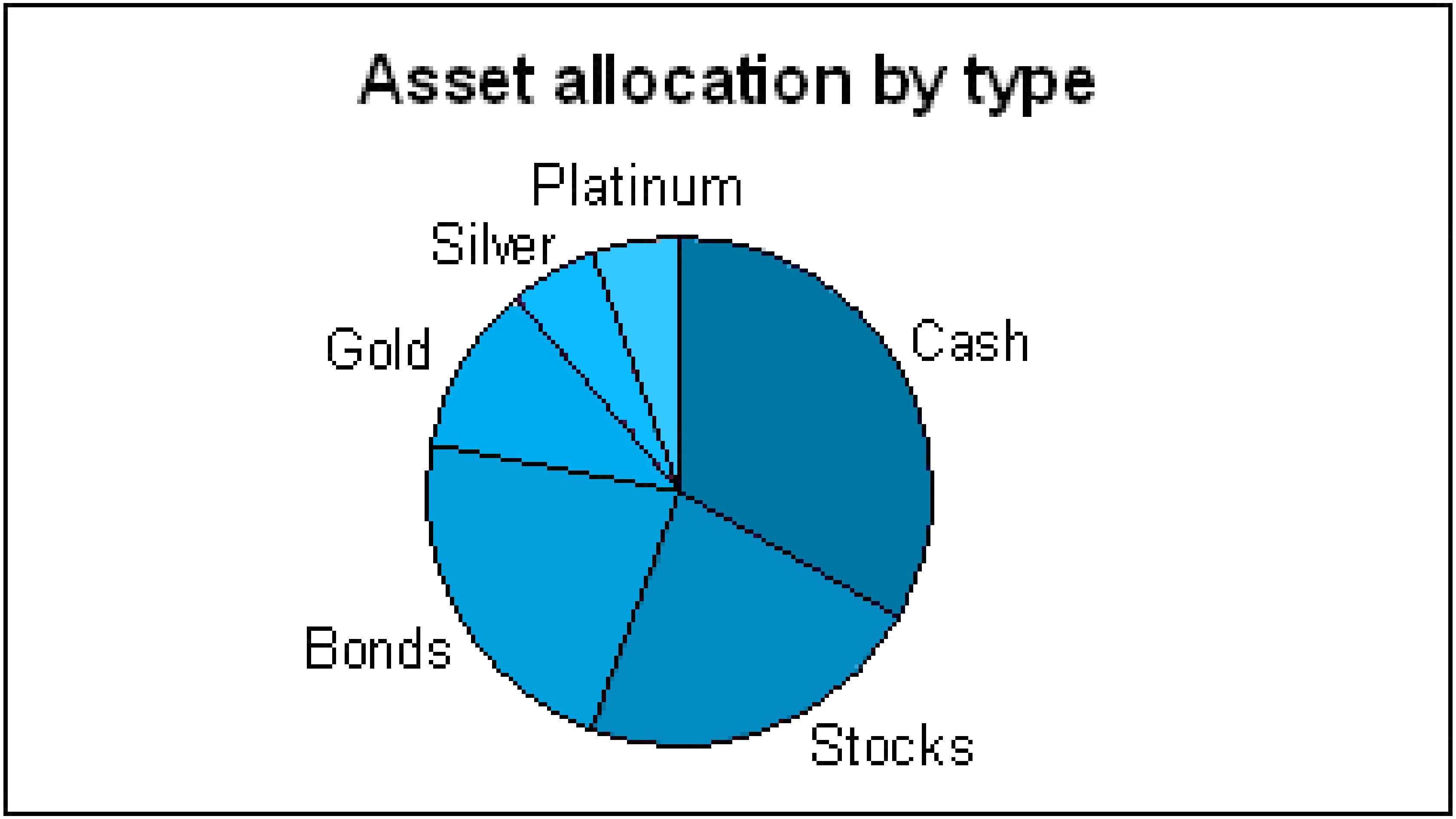
Angle

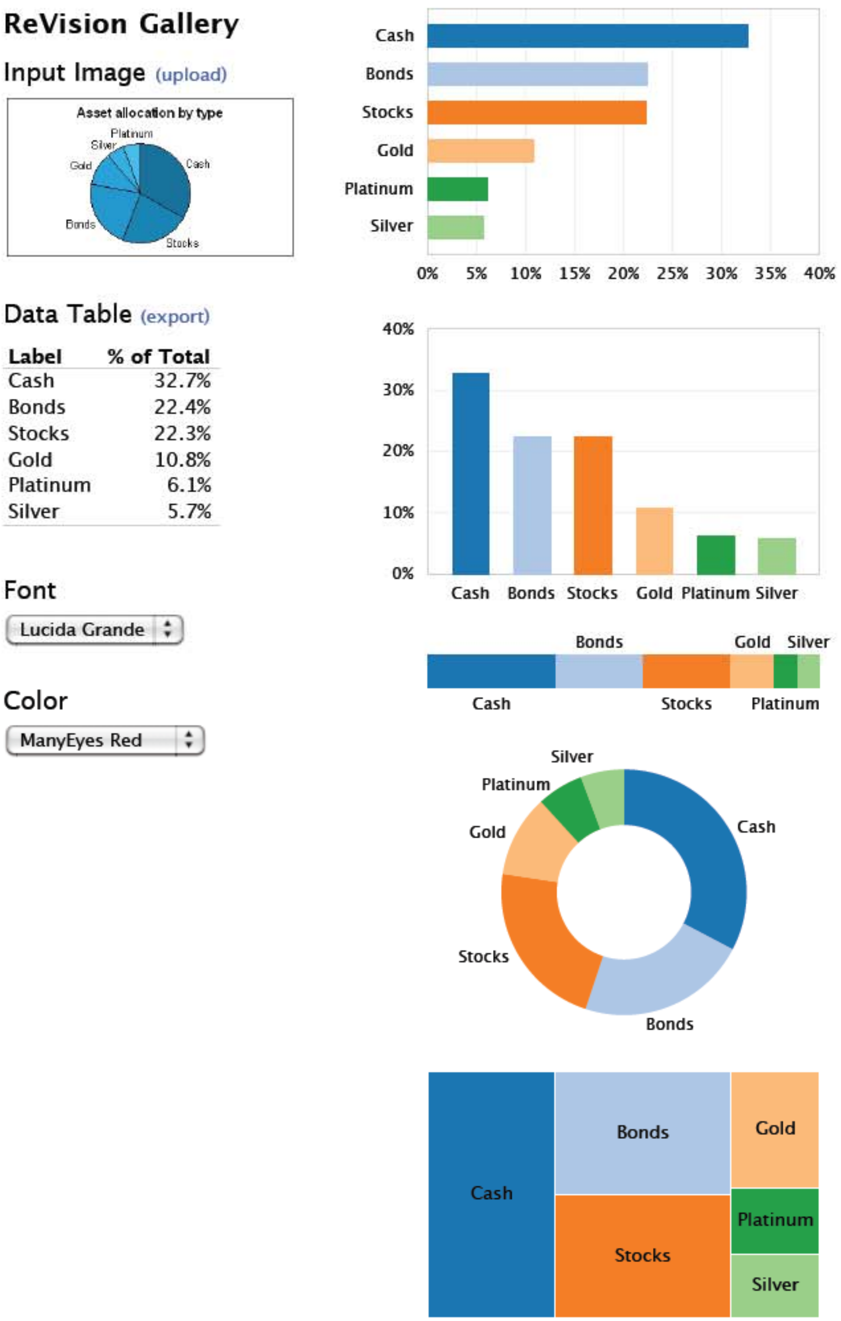
Area

Volume

Color hue-saturation-density

Least accurate


Asset allocation by type
Platinum
Silver
Cash
Gold
Bonds
Stocks



ReVision Gallery

Input Image (upload)

Data Table (export)

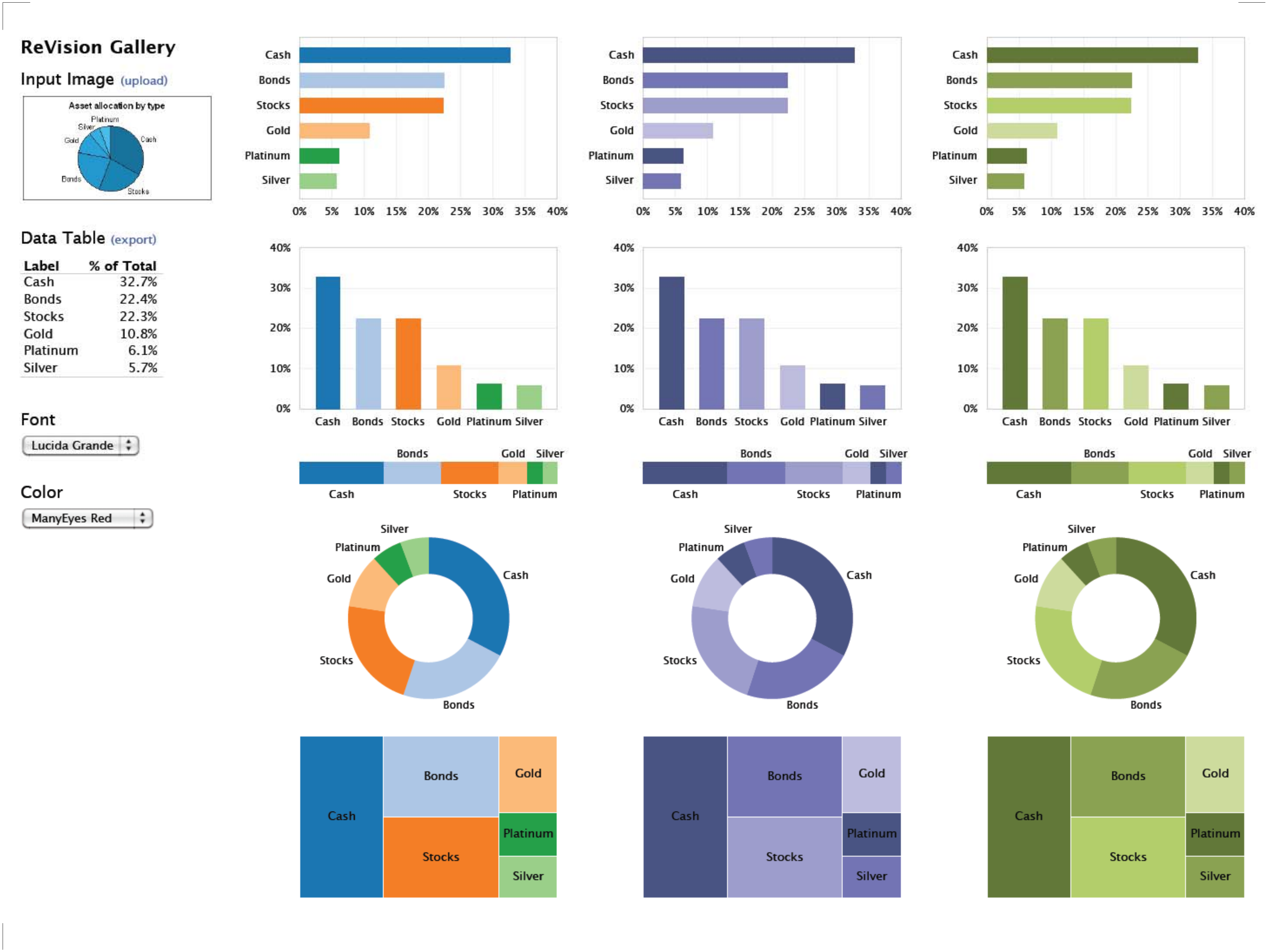
Label	% of Total
Cash	32.7%
Bonds	22.4%
Stocks	22.3%
Gold	10.8%
Platinum	6.1%
Silver	5.7%

Font

Lucida Grande

Color

ManyEyes Red



ReVision Gallery

Input Image (upload)

Data Table (export)

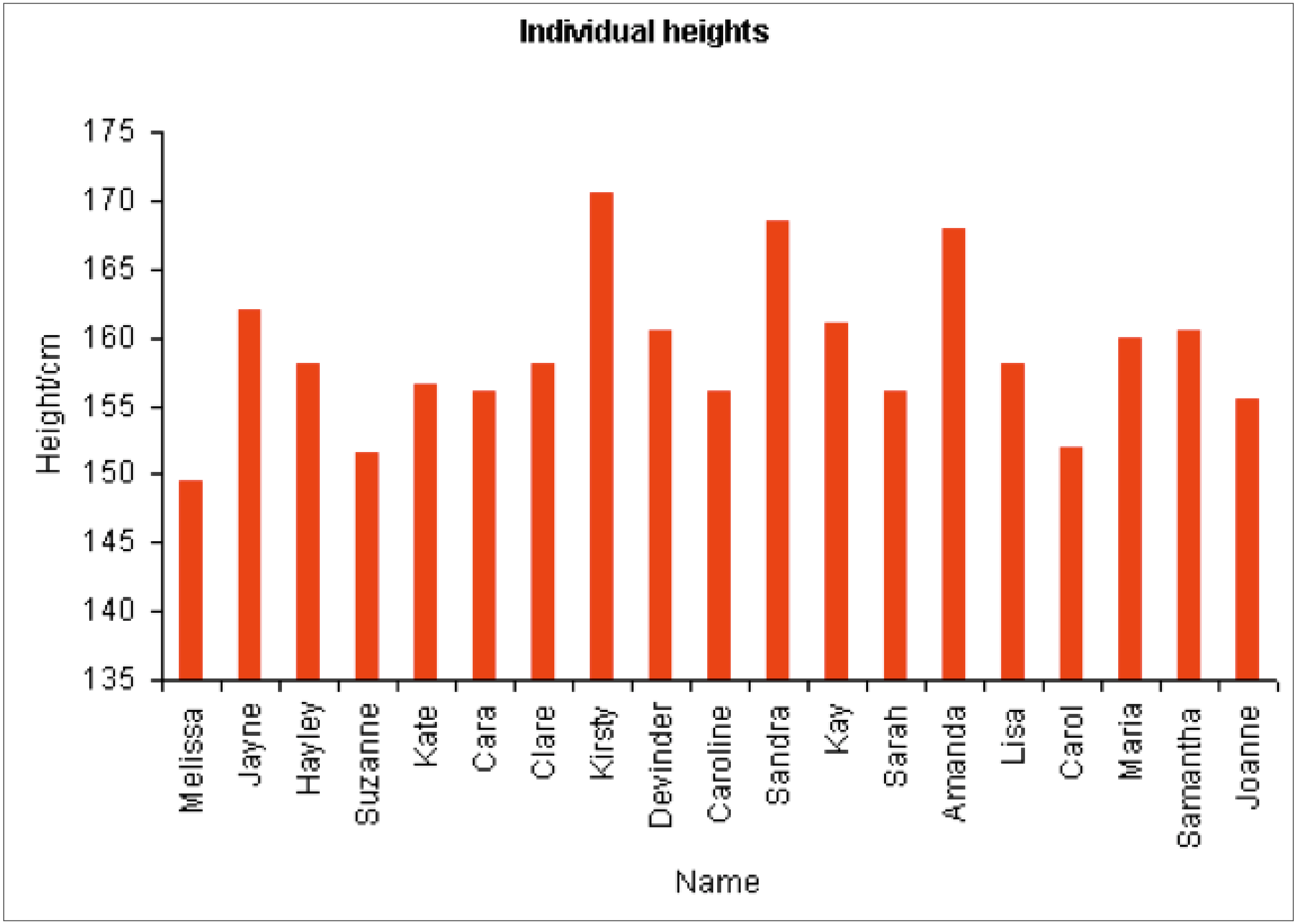
Label	% of Total
Cash	32.7%
Bonds	22.4%
Stocks	22.3%
Gold	10.8%
Platinum	6.1%
Silver	5.7%

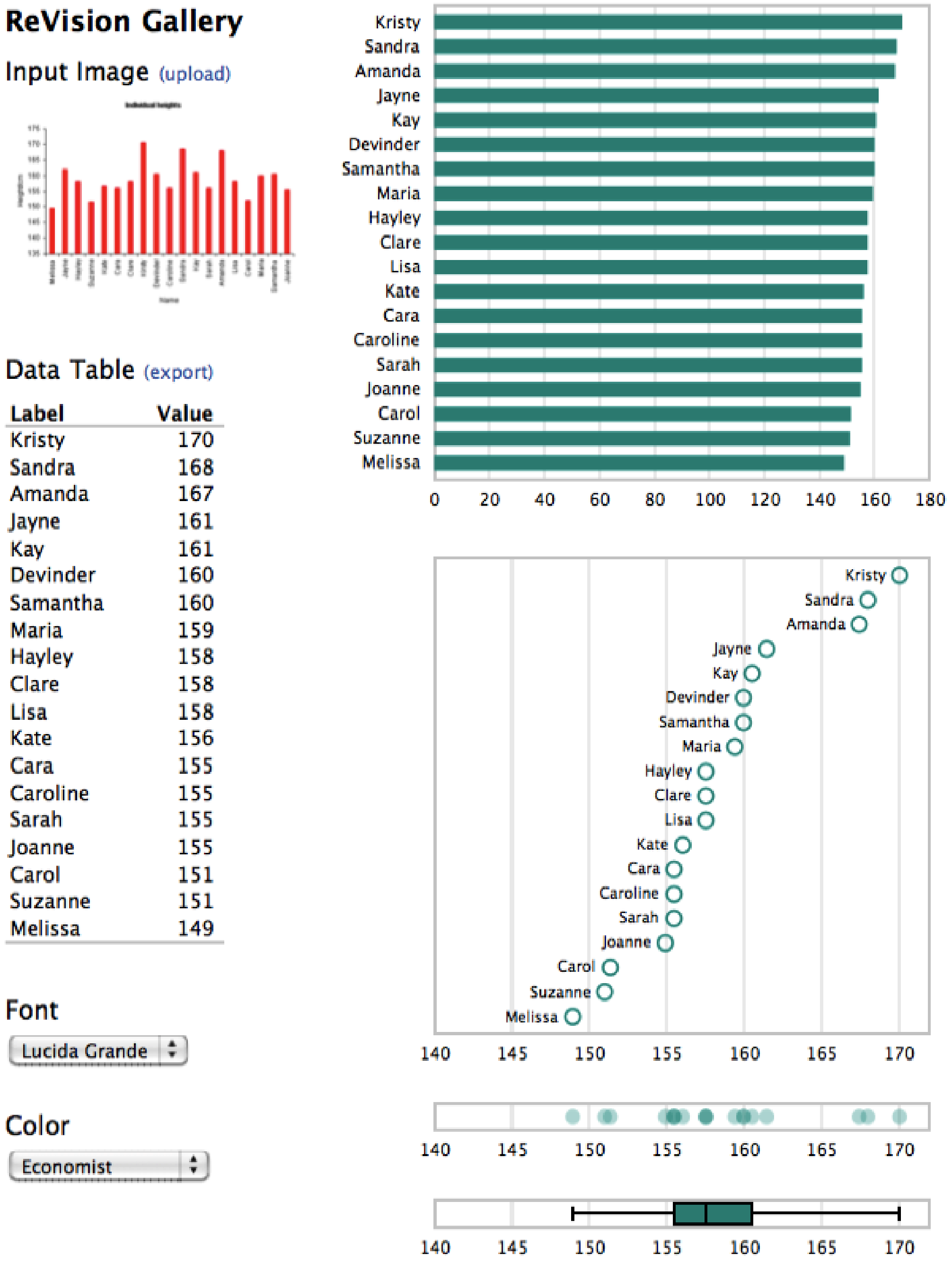
Font

Lucida Grande

Color

ManyEyes Red


Individual heights
Height/cm
175
170
165
160
155
150
145
140
135
Melissa
Jayne
Hayley
Suzanne
Kate
Cara
Clare
Kirsty
Devinder
Caroline
Sandra
Kay
Sarah
Amanda
Lisa
Carol
Maria
Samantha
Joanne
Name



ReVision Gallery

Input Image (upload)

Data Table (export)

Label	Value
Kristy	170
Sandra	168
Amanda	167
Jayne	161
Kay	161
Devinder	160
Samantha	160
Maria	159
Hayley	158
Clare	158
Lisa	158
Kate	156
Cara	155
Caroline	155
Sarah	155
Joanne	155
Carol	151
Suzanne	151
Melissa	149

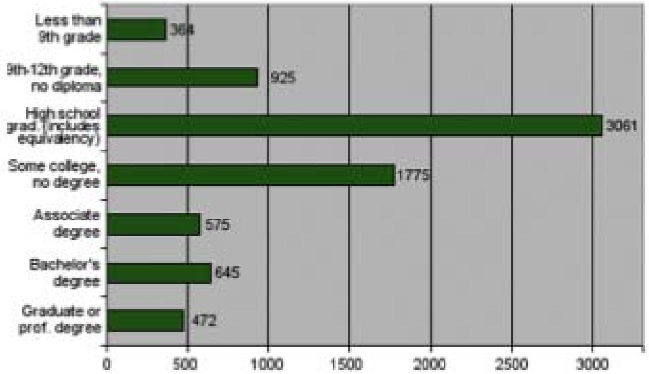
Font

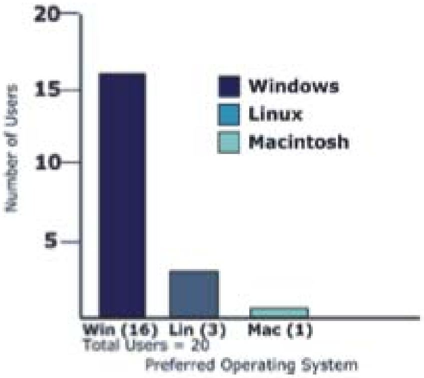
Lucida Grande

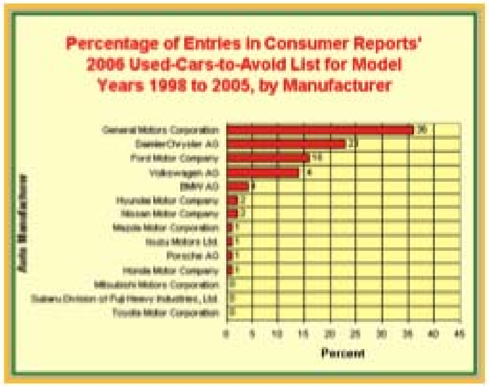
Color

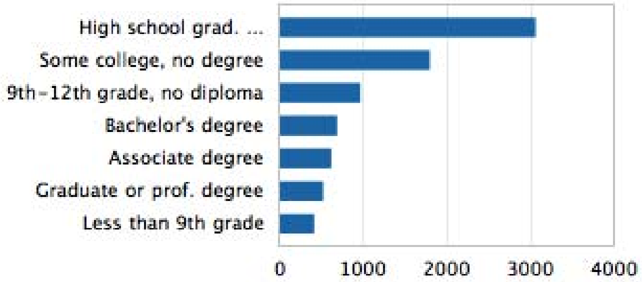
Economist

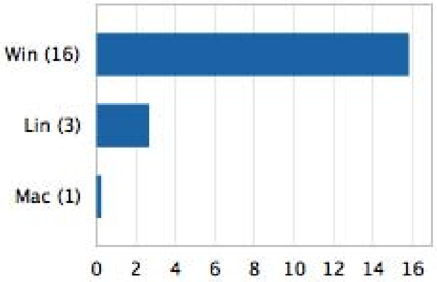
Bar Chart Redesigns

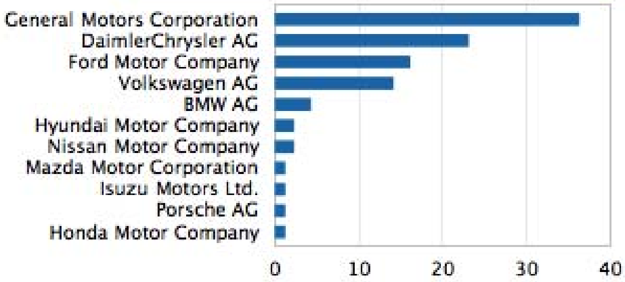




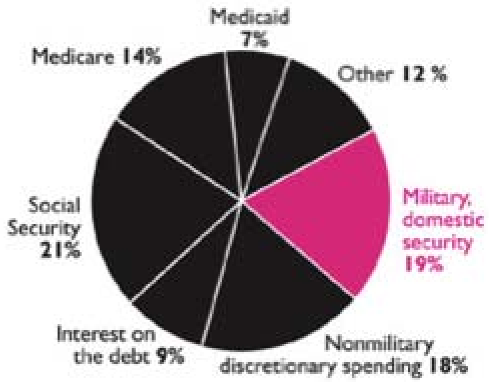


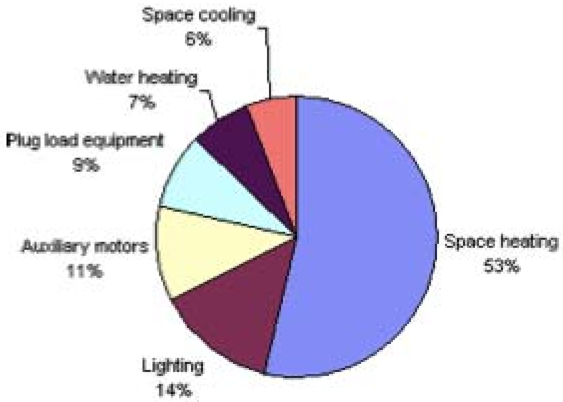


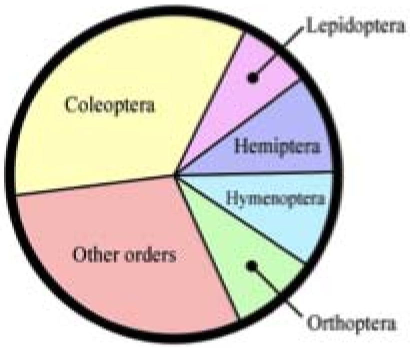


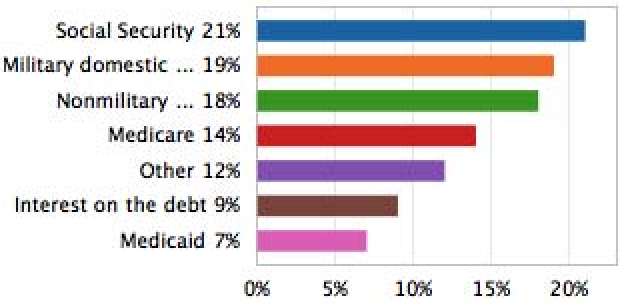


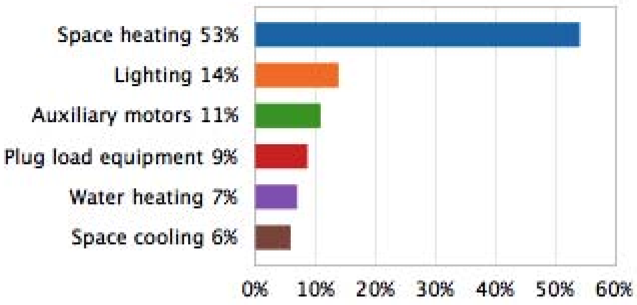
Pie Chart Redesigns

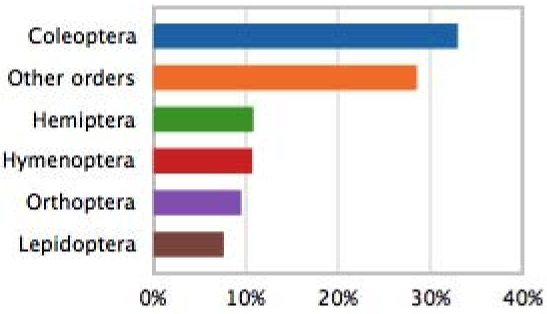






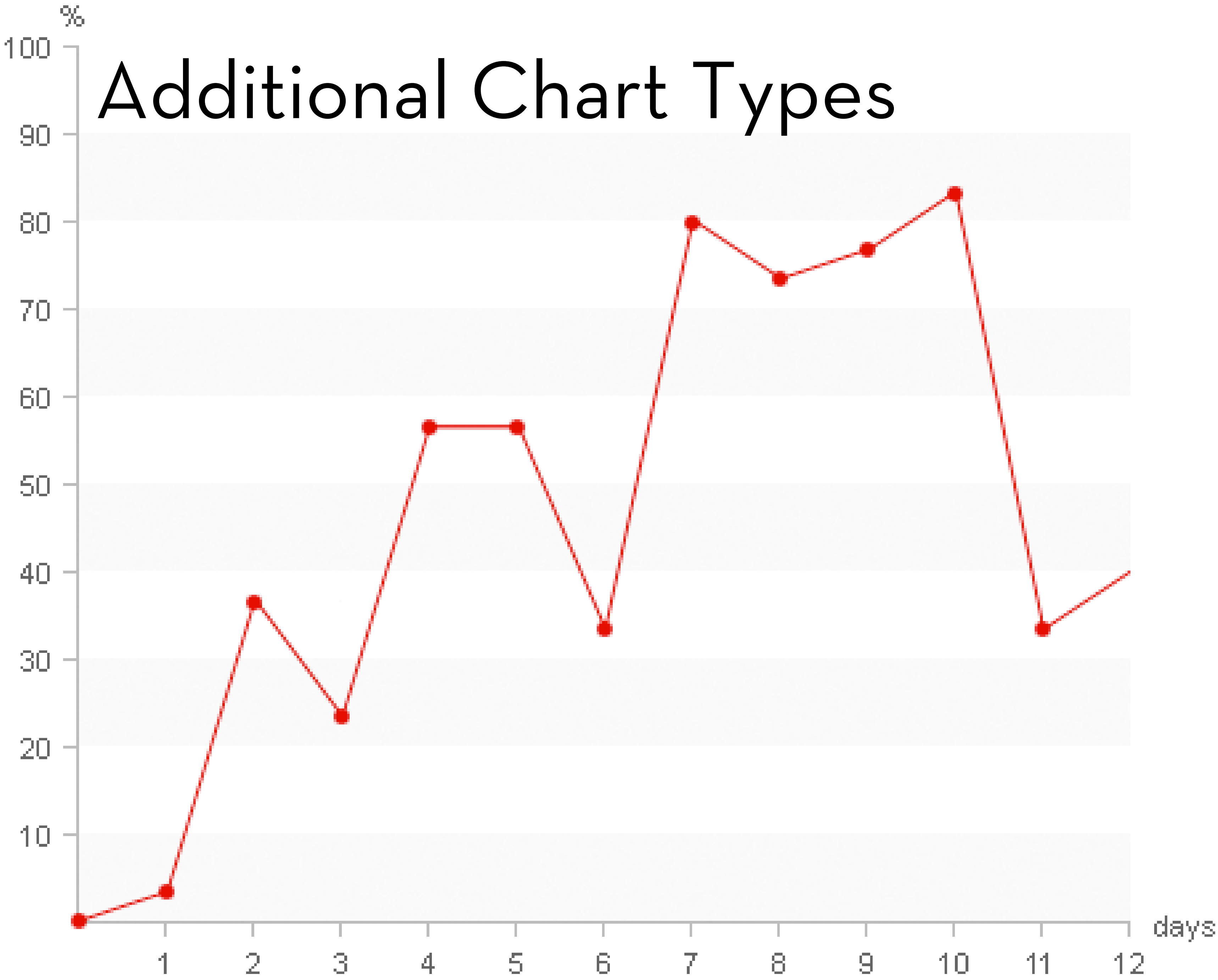




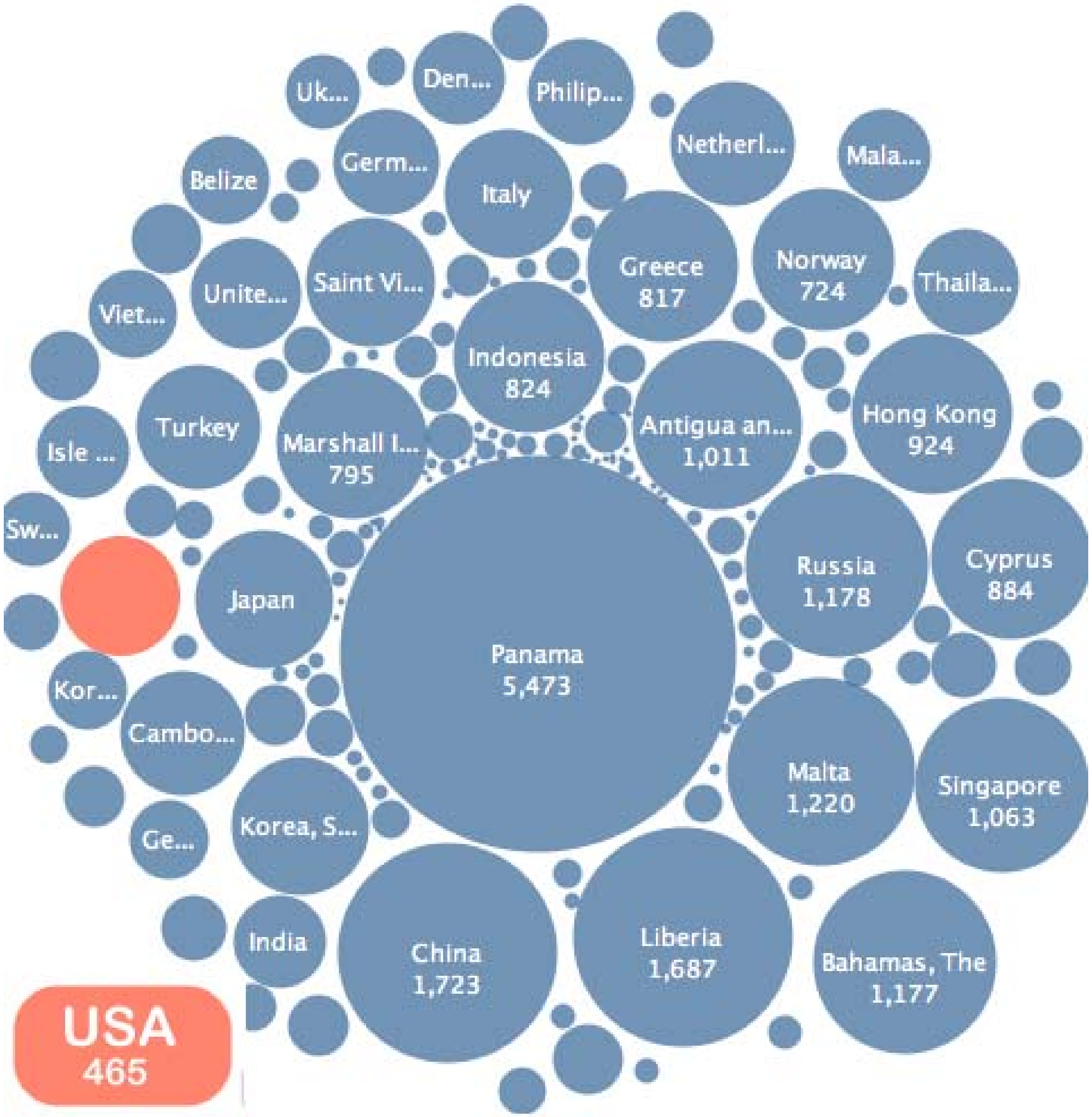


Future Work

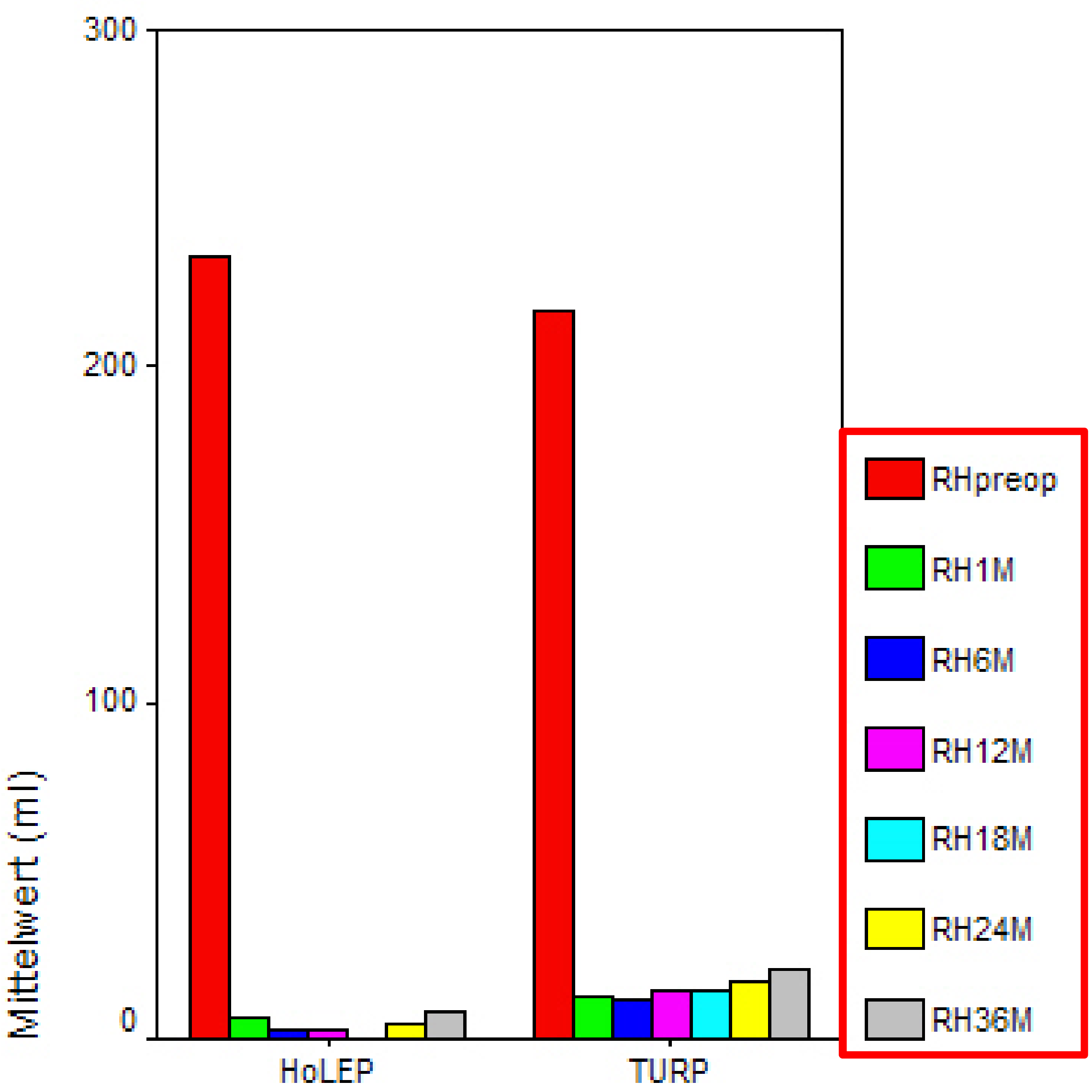
Additional Chart Types



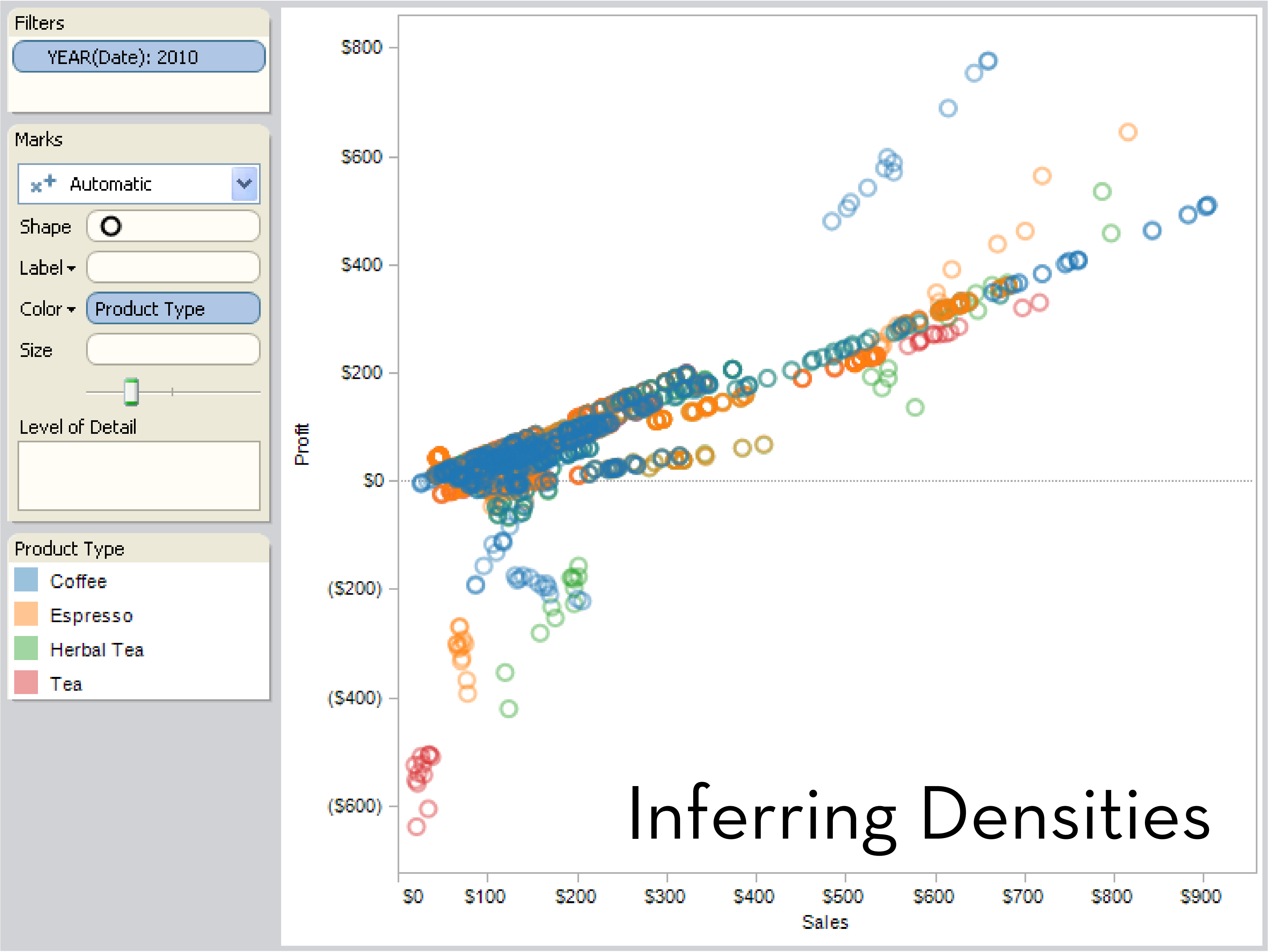
Additional Chart Types



Extracting Legends



Inferring Densities



Filters

YEAR(Date): 2010

Marks

Automatic

Shape ○

Label

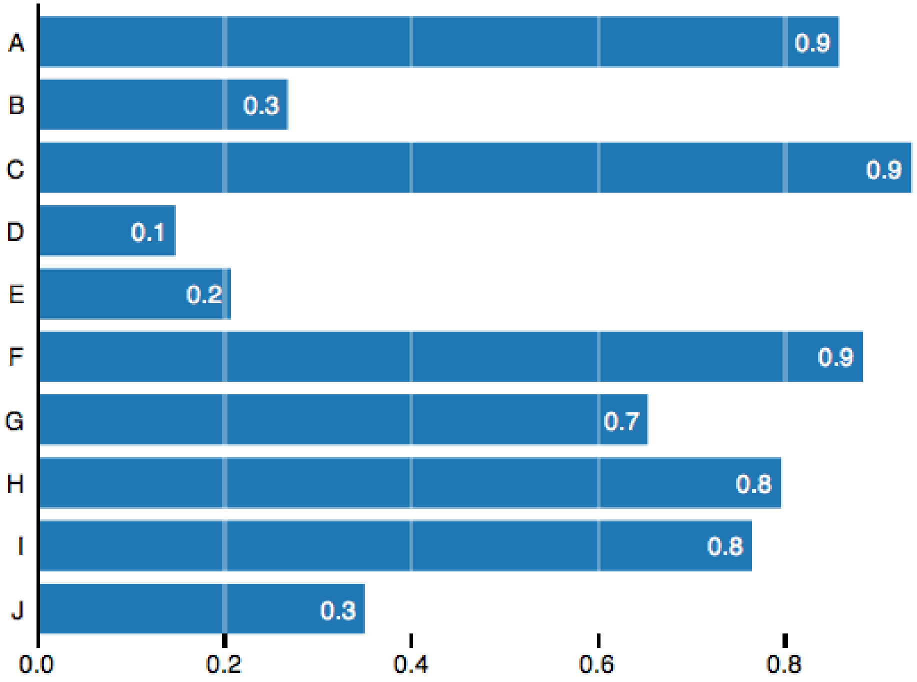
Color Product Type

Size

Level of Detail

Product Type

- 🟦 Coffee
- 🟧 Espresso
- 🟩 Herbal Tea
- 🟥 Tea



Generative Model

Infer chart specifications, not just backing data.

```
/* Sizing and scales. */
var w = 400,
  h = 250,
  x = pv.linear(0, 1.1)
     .range(0, w),
  y = pv.ordinal(pv.range(10))
     .splitBanded(0, h, 4/5);

/* The root panel. */
var vis = new pv.Panel()
   .width(w)
   .height(h)
   .bottom(20)
   .left(20)
   .right(10)
   .top(5);

/* The bars. */
var bar = vis.add(pv.Bar)
   .data(data)
   .top(function() y(this.index))
   .height(y.range().band)
   .left(0)
   .width(x);
```

ReVision

Automated Classification, Analysis and Redesign of Chart Images

Jeffrey Heer & Fei-Fei Li

with Manolis Savva, Nick Kong, Arti Chhajta, & Maneesh Agrawala

Thanks to Stanford CS SEED grant program!