Foresight: Recommending Visual Insights

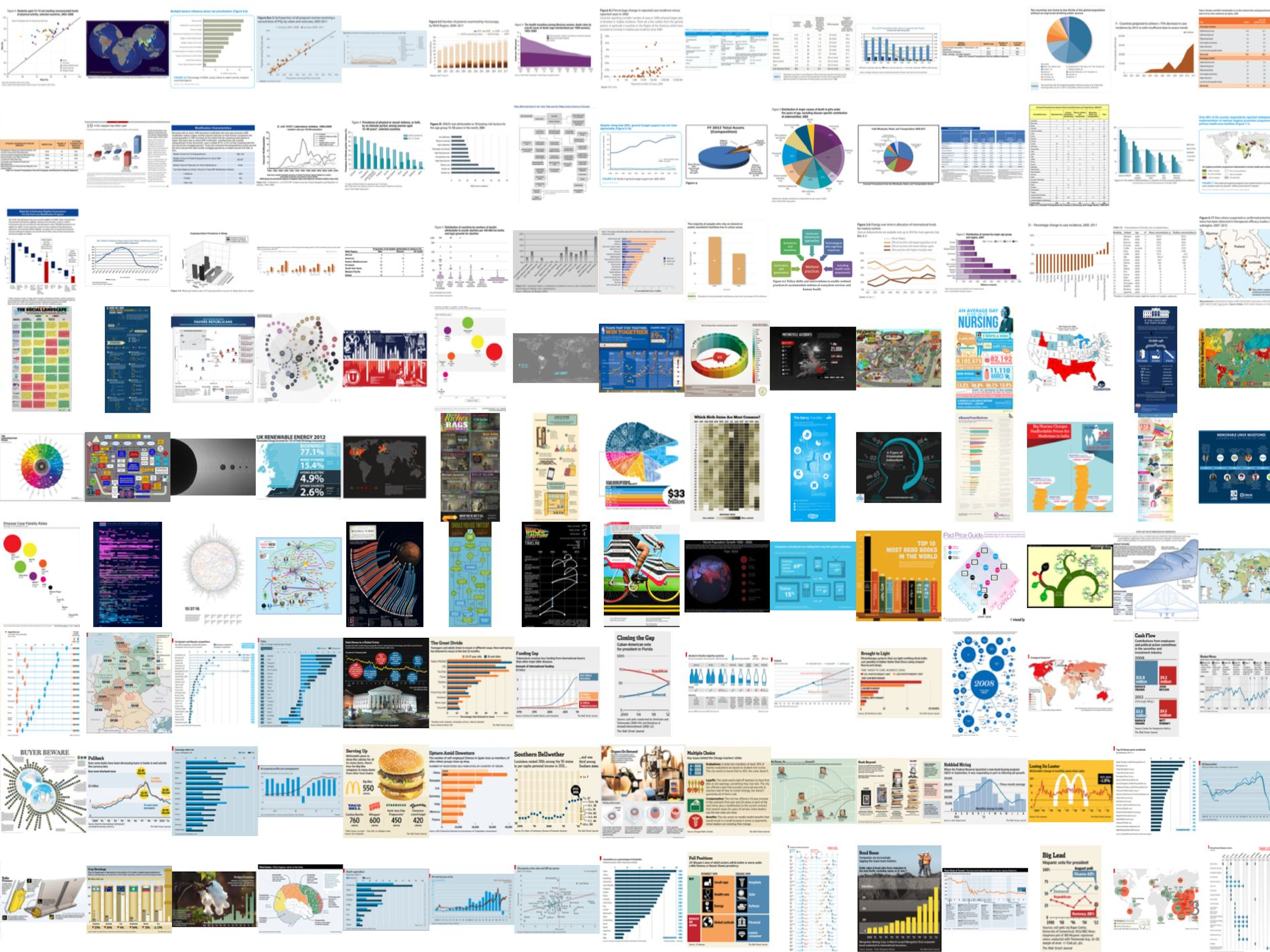
Çağatay Demiralp
Peter Haas
Srinivasan Parthasarathy
Tejaswini Pedapati

IBM Research

Foresight: Recommending Visual Insights

Çağatay Demiralp
Peter Haas
Srinivasan Parthasarathy
Tejaswini Pedapati

IBM Research



THE SOCIAL LANDSCAPE
NURSING
UK RENEWABLE ENERGY 2012
77.1%
15.4%
4.9%
2.6%
Which Birth Dates Are Most Common?
$33 billion
Disease Case Fatality Rates
TOP 10 MOST READ BOOKS IN THE WORLD
The Great Divide
Funding Gap
Closing the Gap
Brought to Light
2008
Cash Flow
BUYER BEWARE
Pullback
Serving Up
Upturn Amid Downturn
Southern Bellwether
Multiple Choice
Hobbled Hiring
Losing Its Luster
Poll Positions
Big Lead

Automated Visualization Systems

Chart Typologies

Declarative Encoding Languages

Component Model Architectures

Graphics APIs

speed

expressiveness

Excel
Google Charts
Tableau

D3
ggplot
VizQL
VizML

Processing
Prefuse

OpenGL
DirectX
Java2D
HTML Canvas

Majority of Users

Automated Visualization Systems

Chart Typologies

Declarative Encoding Languages

Component Model Architectures

Graphics APIs

speed

expressiveness

Foresight

Excel
Google Charts
Tableau

D3
ggplot
VizQL
VizML

Processing
Prefuse

OpenGL
DirectX
Java2D
HTML Canvas

Majority of Users

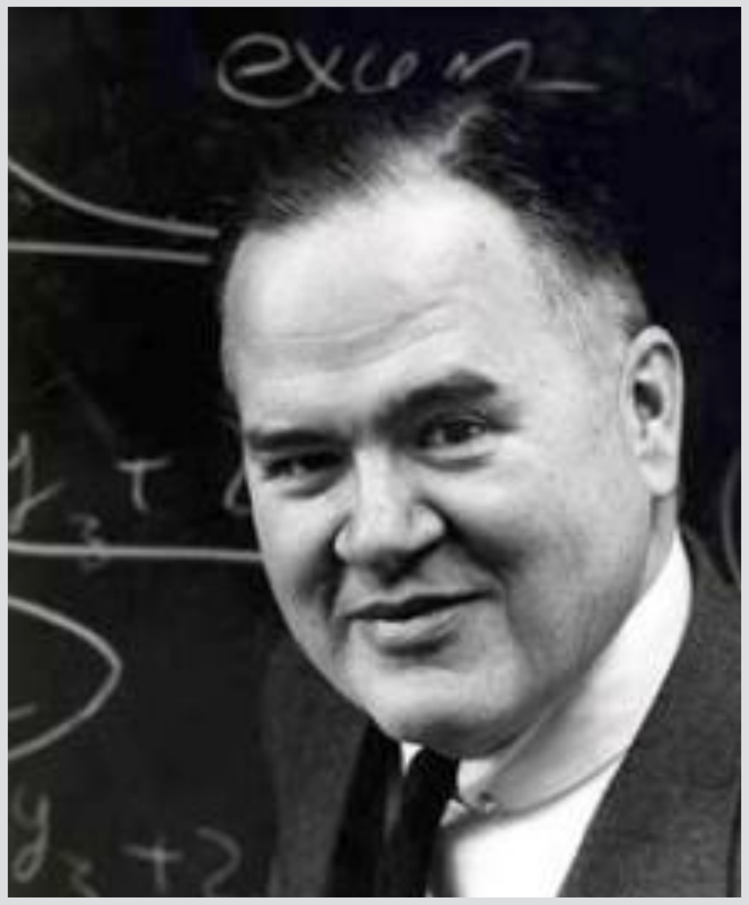
Exploratory Data Analysis (EDA)

Explore patterns and relations in data, ask questions and (re)form hypotheses

Statistics + visualizations

“Here is the data! Which questions does it want us to ask? What seems to be going on?”

Exploratory vs. confirmatory



John W. Tukey
(1915 - 2000)

EDA CHALLENGES

Data complexity

Insufficient time and skills

Cognitive limitations

Transient working memory

Tendency to fit evidence to existing expectations and schemas

[Tversky & Kahneman'75,Nickerson'98,Card et al.'05]

FORESIGHT

Structured, rapid first order EDA

Framework for exploring datasets through ranked and neighborhood based visualizations

Exploring engine supporting a faceted interface

Sketch based composition for fast approximate computation

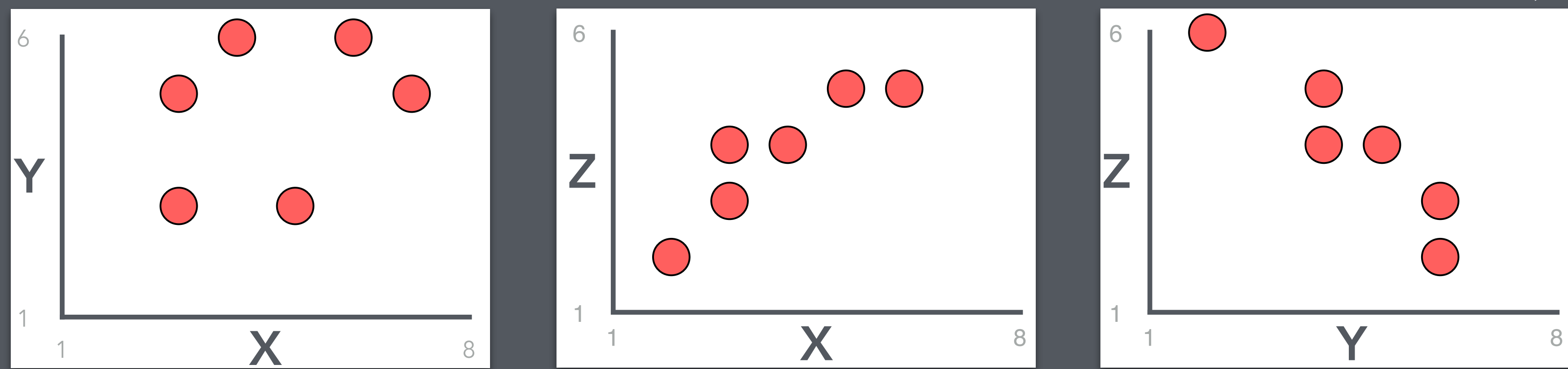
DEMO

OECD Dataset: 25 well-being indicators (columns) for 36 OECD member countries (rows)

PRIOR WORK

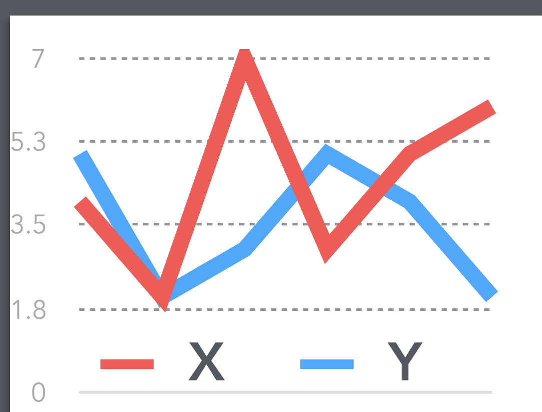

data
visual
encoding
Y
X
Z
X
Z
Y
X
Y
X
Y

measure + data



. . .

measure
+
visual
encoding



⋮

ShowMe’07

Voyager-2’17
Voyager’16

Foresight
Rank-by-Feature’04
AutoVis’10

Zenvisage’16

SeeDB’15

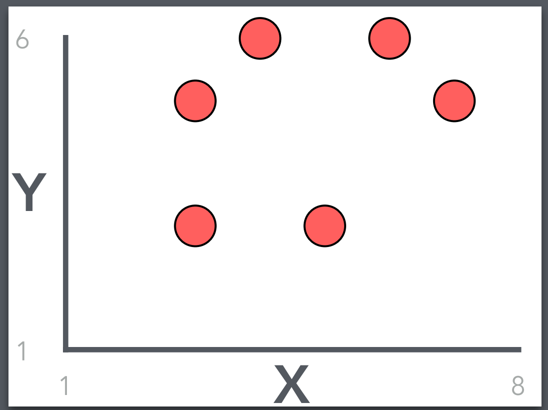
GrandTour’84
PRIM-9’79

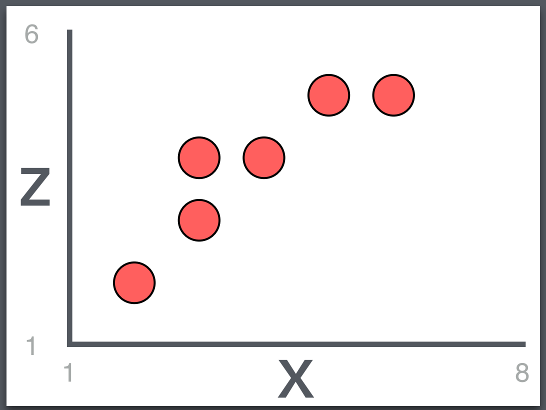
VizDeck’13

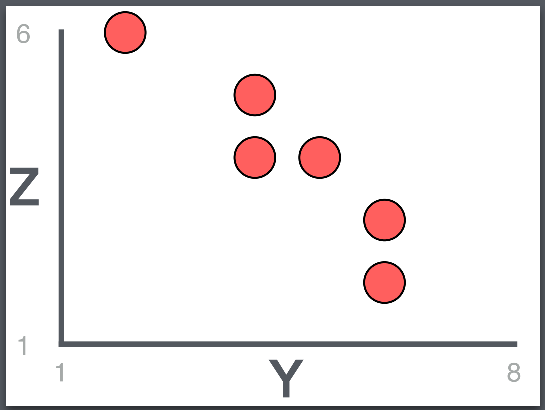
Gotz & Wen’09
Zhou & Chen'03
SAGE’94

measure + data

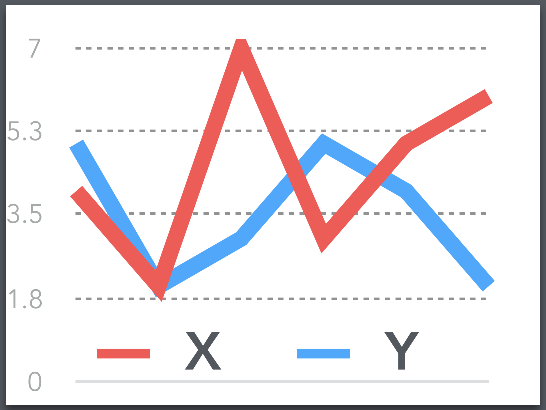
measure + visual encoding

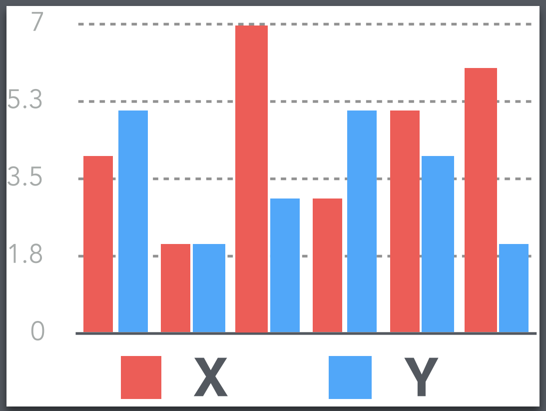






. . .





⋮

ShowMe’07

Voyager-2’17
Voyager’16

Foresight *statistical*
Rank-by-Feature’04
AutoVis’10

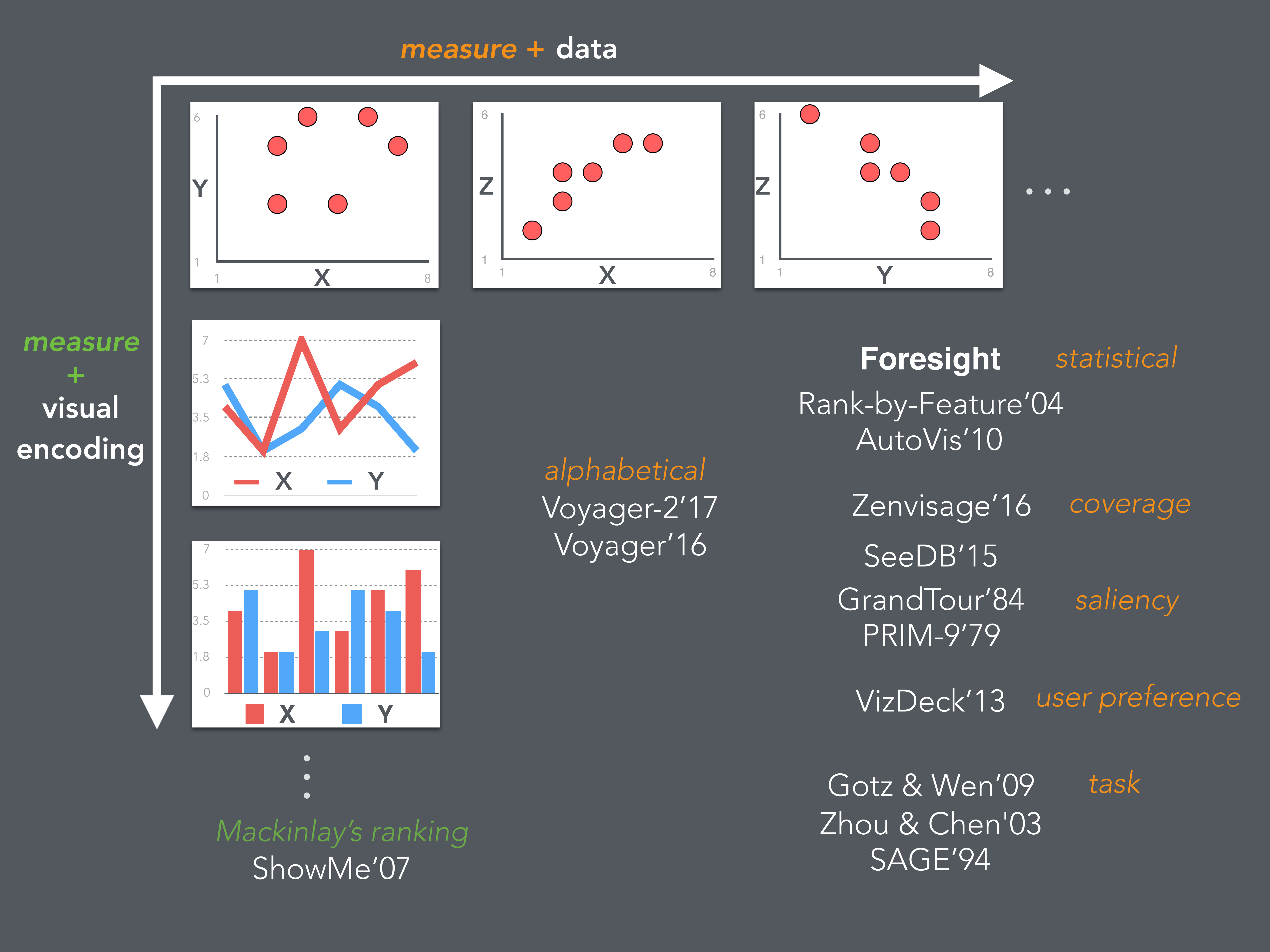
Zenvisage’16

SeeDB’15

GrandTour’84
PRIM-9’79

VizDeck’13

Gotz & Wen’09
Zhou & Chen'03
SAGE’94



measure + data
measure
+
visual
encoding
Y
X
Z
X
Z
Y
X
Y
X
Y
Mackinlay's ranking
ShowMe'07
alphabetical
Voyager-2'17
Voyager'16
Foresight
statistical
Rank-by-Feature'04
AutoVis'10
Zenvisage'16
coverage
SeeDB'15
GrandTour'84
saliency
PRIM-9'79
VizDeck'13
user preference
Gotz & Wen'09
task
Zhou & Chen'03
SAGE'94

DESIGN

INTERVIEW STUDY

Participants:

10 data scientists (2 female + 8 male)

- IBM Research
- Diverse domains, e.g., healthcare, marketing , finance, etc.
- MS & PhDs
- Predictive modeling

INTERVIEW STUDY

Sought answers for:

How do analysts start exploratory data analysis?

What tools do analysts generally work with?

What visualizations and statistics do analysts frequently use?

How do analysts decide on what is “interesting” in data?

What strategies do analysts use with large data?

What are productivity challenges in general and for specific tools?

INTERVIEW STUDY

Procedure & analysis:

- Face to face, open ended
- Walk through a recent experience
- Three note takers & audio recorded
- Lasted ~30 mins
- Merged & grouped through iterative coding

INTERVIEW STUDY

Results:

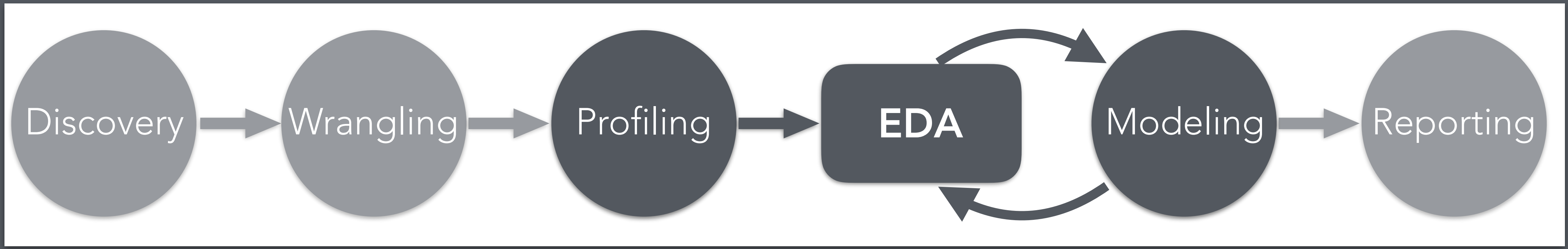
1) EDA in Data Analysis Process
2) Junior versus Senior Analysts
3) Stratified Greedy Navigation
4) Handling Big Data
5) Tools
6) Challenges

INTERVIEW RESULTS

EDA in Data Analysis Process

Analysts spent most of their time on EDA, after data is readied for analysis

First order understanding dominated EDA



INTERVIEW RESULTS

Junior versus Senior Analysts

Senior analysts (5+ years experience) spent more time on domain understanding and EDA than junior analysts

Junior analysts transitioned to modeling faster, relied more on ML based techniques

Senior analysts relied on basic statistical techniques but put more emphasis on domain specific—causal/semantic—relations

INTERVIEW RESULTS

Stratified Greedy Navigation

Simpler, univariate to more complex, multivariate

Hierarchical both in statistical computation and data relations

Rarely considered trivariate relations

Greedy strategy deciding on what to focus

May cause premature fixation

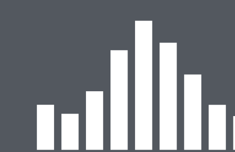
DESIGN CRITERIA

1. *Structure data variation around statistical descriptors*
2. *Use descriptor strength to drive the promotion of data variation*
3. *Give user control over the definition of descriptor strength*
4. *Use the best visualizations for communicating statistical descriptors*
5. *Facilitate stratified work flow to minimize the cost of exploration*
6. *Enable access to raw data on demand*

DESCRIPTORS

Dispersion: Quartile coefficient of dispersion; visualized with histogram

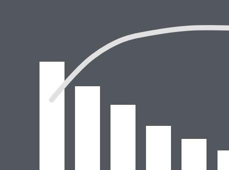
Skew: Standardized skewness coefficient; visualized with histogram



Heavy tails: Kurtosis; visualized with histogram

Outliers: Number of points outside the inlier range of Tukey box-and-whisker plot; visualized using box-and-whisker plot

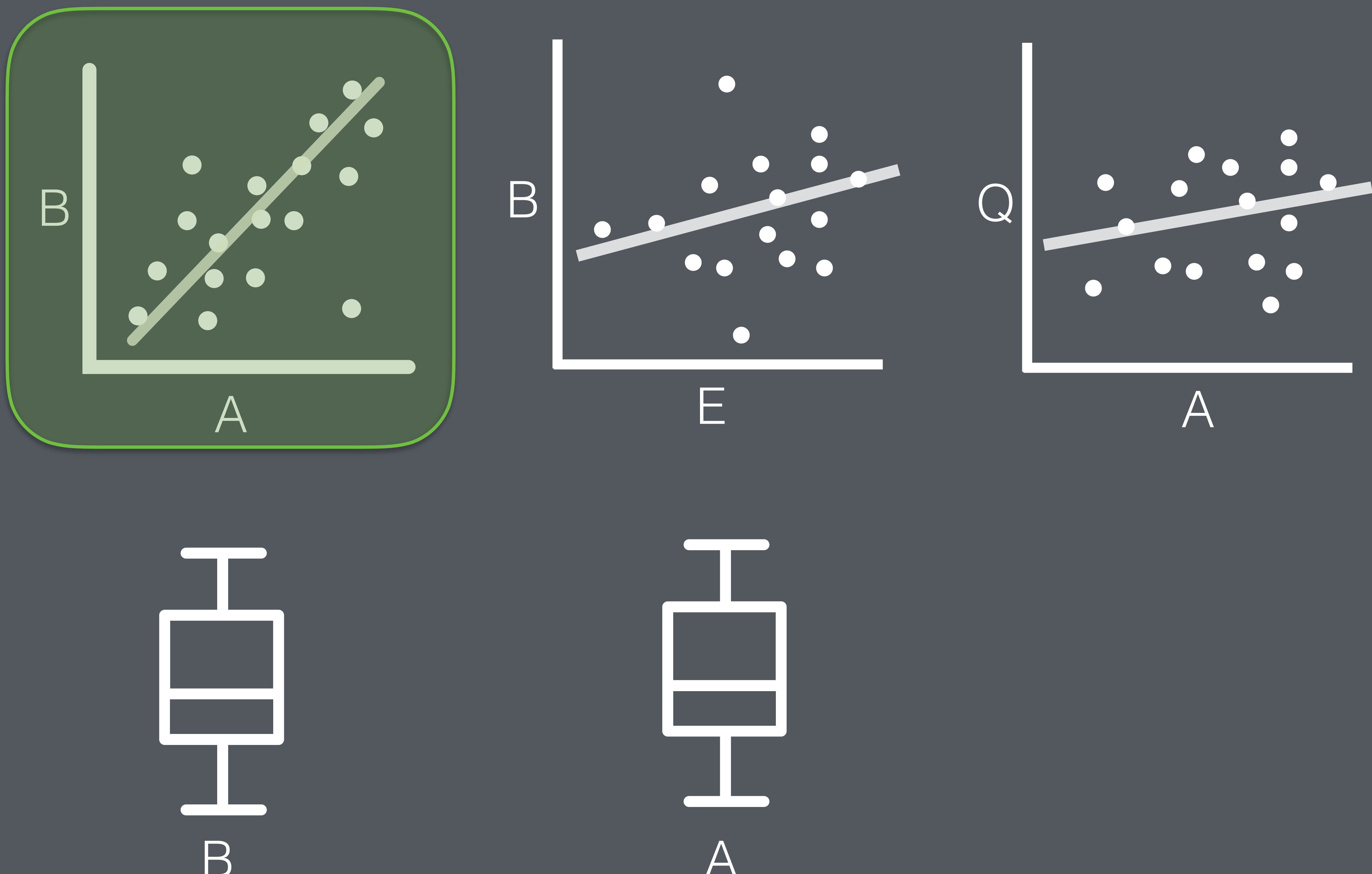
Heterogeneous frequencies: Normalized Shannon Entropy; visualized with Pareto chart



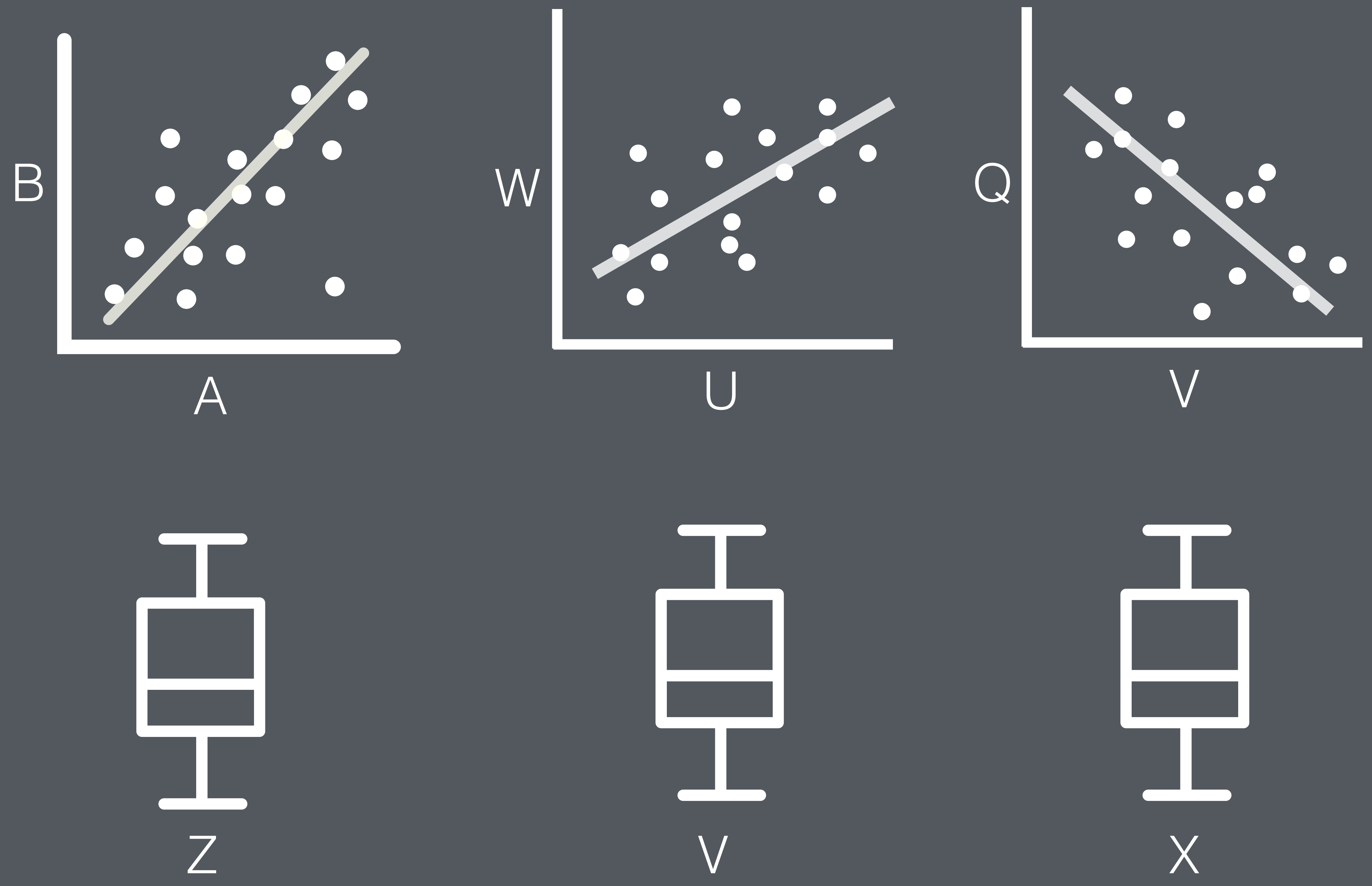
Linear relationship: Absolute value of the Person correlation coefficient; visualized with a scatter plot with a best line fit overlaid

NEIGHBORHOOD

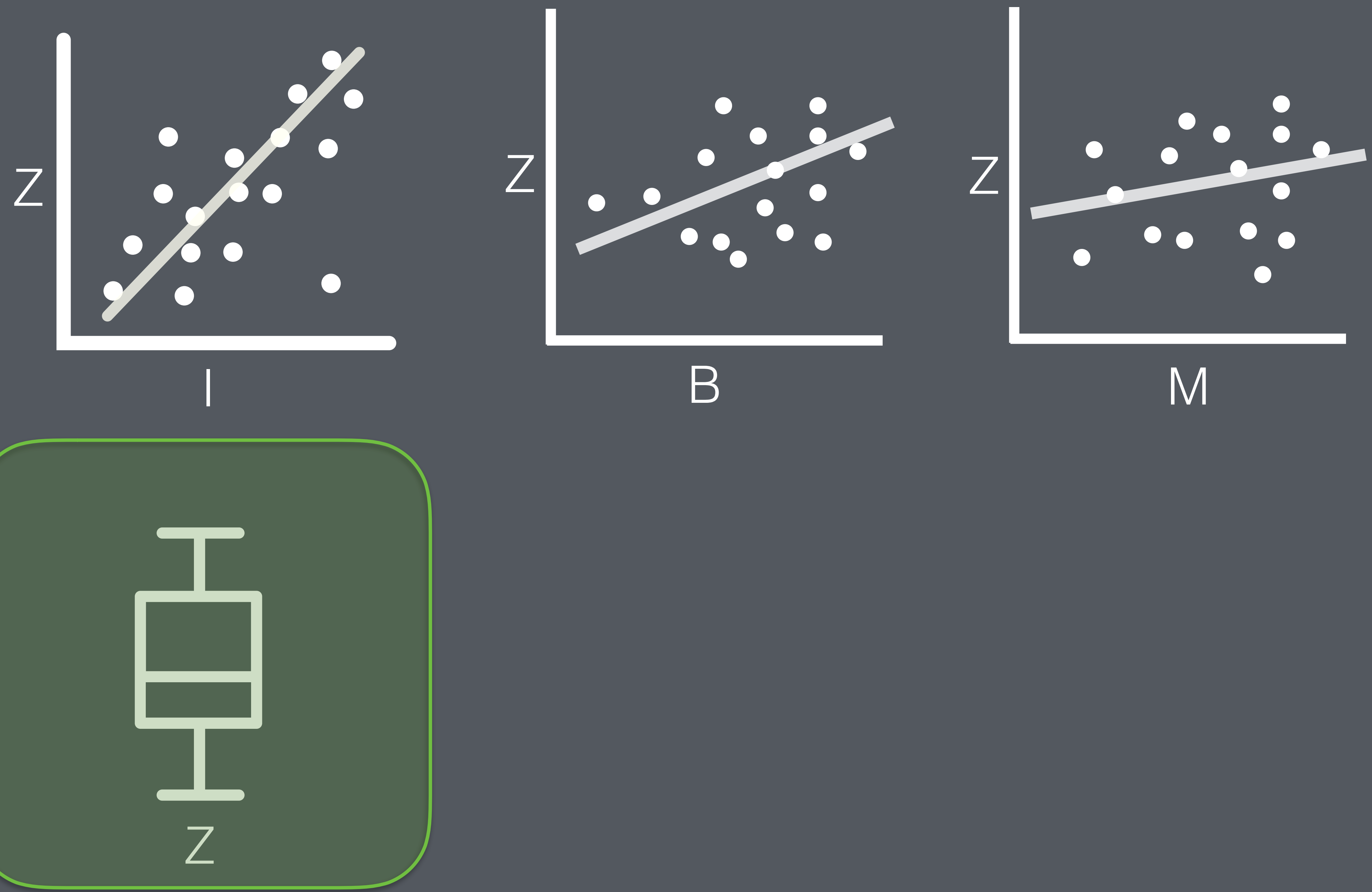
NEIGHBORHOOD



NEIGHBORHOOD



NEIGHBORHOOD



SCALABILTY VIA SKETCHING

SKETCHES

Compressed synopses for fast approximate computations

Provide desirable guarantees on approximation errors

Hyperplane sketch for correlation

CONCLUSION

"What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herb A. Simon
(1916 - 2001)

FORESIGHT

Framework for exploring datasets through ranked and neighborhood based visualizations

Exploring engine supporting a faceted interface

Sketch based composition for fast approximate computation

Interview study providing insights into the EDA practices, informing EDA tool design at large

ON GOING

Human-subjects study 

New descriptors 

Foresight: Recommending Visual Insights

Çağatay Demiralp @serravis
Peter Haas
Srinivasan Parthasarathy
Tejaswini Pedapati

IBM Research

INSIGHT

Strong manifestation of a statistical property of the data, e.g., high correlation between two attributes, high skewness or concentration about the mean of a single attribute, a strong clustering of values, etc.