Data Analytics
A (Short) Tour

Venkatesh-Prasad Ranganath

http://about.me/rvprasad

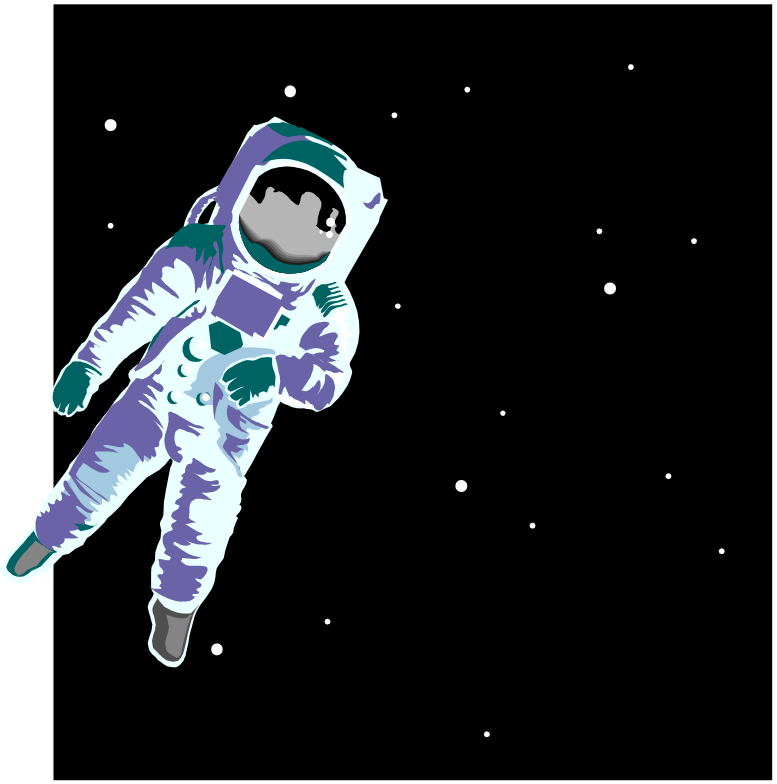
Is it Analytics or Analysis?

Analytics uses analysis to recommend actions or make decisions.

Why Data Analysis?

Confirm a hypothesis

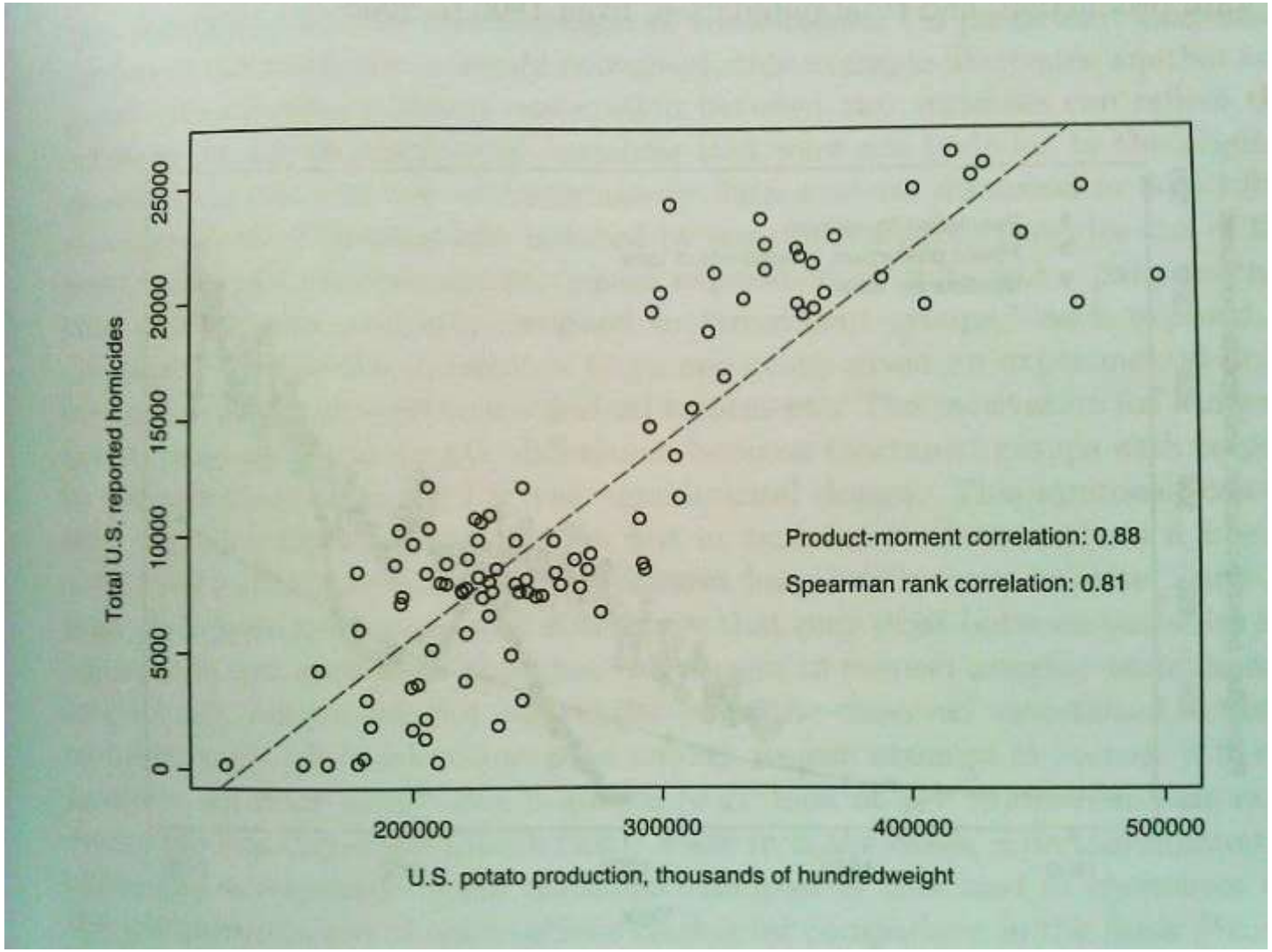
Confirmatory



Explore the data

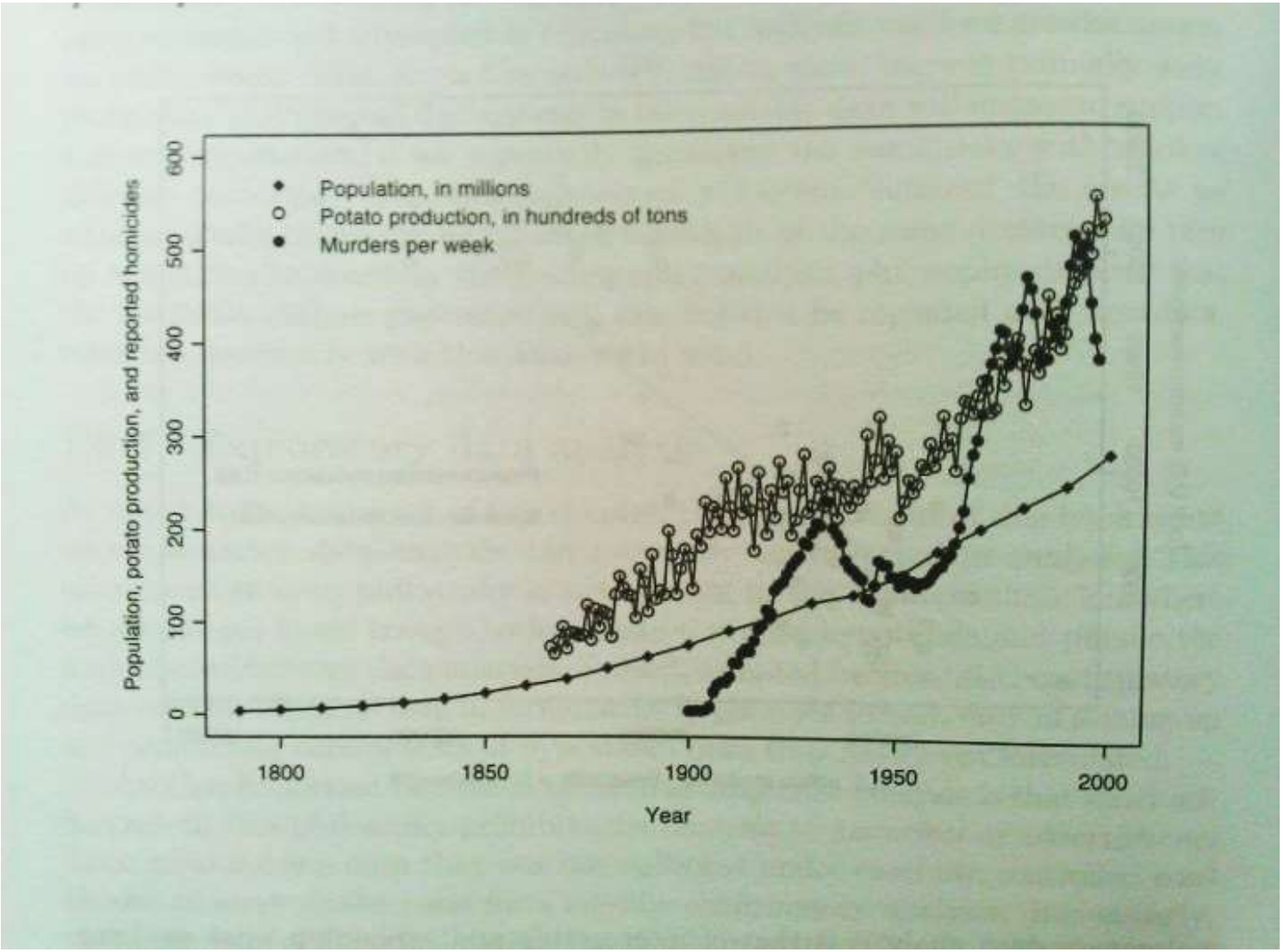
Exploratory (EDA)

Word of Caution – Case of Killer Potatoes?



This is figure 1.5 in the book “Exploring Data” by Ronald K Pearson.

Word of Caution – Case of Killer Potatoes?

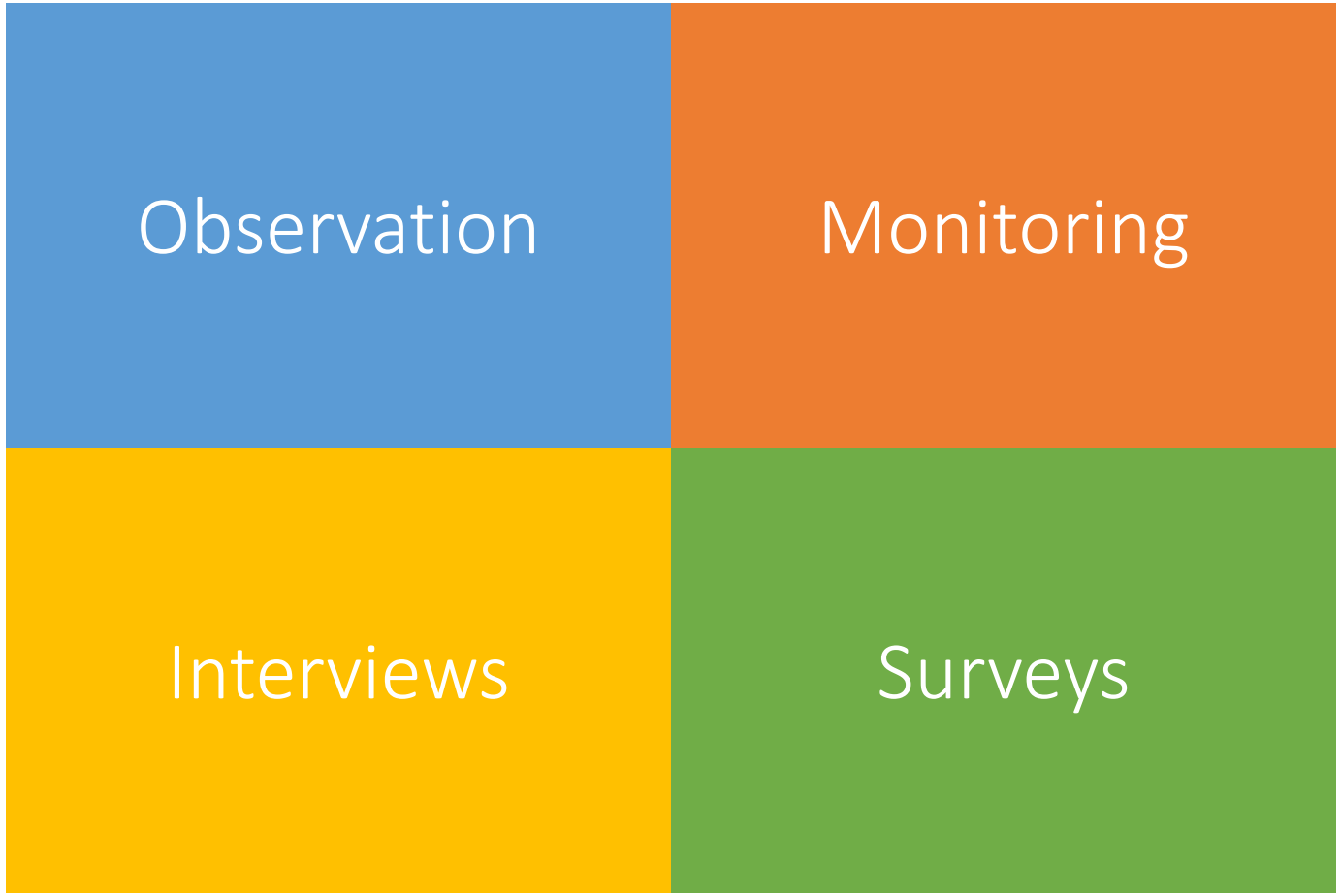


This is figure 1.6 in the book “Exploring Data” by Ronald K Pearson.

Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Data Collection – Approaches



Data Collection – Comparing Approaches

	Observation	Interviews	Surveys	Monitoring
Technique	Shadowing	Conversation	Questionnaire	Logging
Interactive	No	Yes	No	No
Simple	No	No	Yes	Yes
Automatable	No	No	Yes	Yes
Scalable	No	No	Yes	Yes
Data Size	Small	Small	Medium	Huge
Data Format	Flexible	Flexible	Rigid	Rigid
Data Type	Qualitative	Qualitative	Qualitative	Quantitative
Real Time Analysis	No	No	No	Yes
Expensive	Yes	Yes	No	No

Data Collection – Comparing Approaches

	Observation	Interviews	Surveys	Monitoring
What to capture?	Flexible	Flexible	Fixed	Fixed
How to capture?	Flexible	Flexible	Fixed	Fixed
Human Subjects	Yes	Yes	Yes	No
Transcription	Yes	Yes	Yes/No	No
SnR	High	High	High	Low
Involves NLP	Unlikely	Unlikely	Likely	Likely
Kind of Analysis	Confirmatory	Confirmatory	Confirmatory	Exploratory
Kind of Techniques	Statistical Testing	Statistical Testing	Statistical Testing	Machine Learning

Data Storage – Choices

- Flat Files
- Databases
- Streaming Data (but there is no storage)

Data Storage – Flat Files

- Simple
- Common / Universal
- Inexpensive
- Independent of specific technology
- Compression friendly
- Very few choices
  - Plain text, CSV, XML, and JSON
- Well established

- Low level data access APIs
- No support for automatic scale out / parallel access
- Unoptimized data access
  - Indices
  - Columnar storage

Data Storage – Databases

- High level data access API
- Support for automatic scale out / parallel access
- Optimized data access
  - Indices
  - Columnar storage
- Well established

- Complex
- Niche / Requires experts
  - Optimization
  - Distribution
- Expensive
- Dependent on specific technology
- DB controlled compression
- Lots of choices
  - SQL, MySQL, PostgreSQL, Maria, Raven, Couch, Redis, Neo4j, ….

Data Storage – Streaming

- Well, there is not storage ☺
- Novel
- Many streaming data sources

- Breaks traditional data analysis algorithms
  - No access to the entire data set
- Too many unknowns
  - Expertise
  - Cost
  - Best practices
  - Accuracy
  - Benefits
  - Deficiencies
  - Ease of use

Data Storage – Algorithms and Necessity

- Flat Files
- Databases
- Streaming Data

- Offline
- Online
- Streaming
- Real-time

- Do we need fast?
- How fast is quick enough?
- How often do we need fast?
- Is it worth the cost?
- Is it worth the loss of accuracy?

Data Representation – Structured

- Easy to process
- One time schema setup cost
- Common schema types
  - CSV, XML, JSON, …
  - You can cook up your schema
- Eases data exploration & analysis
- Off-the-shelf techniques to handle data
- Requires very little expertise
- Ideal with automatic data collection
- Ideal for storing quantitative data

- Rigid
  - Changing schema can be hard
- Upfront cost to define the schema

Data Representation – Unstructured

- Flexible
- Off-the-shelf techniques to preprocess data but requires expertise
- Ideal for manual data collection

- Requires lots of preprocessing
- Complicates data exploration and analysis
- Requires domain expertise
- Extracting data semantics is hard
- Requires schema recovery *

Data Access – Security

- Who has access to what parts of the data?
- What is the access control policy?
- How do we enforce these policies?
  - What techniques do we employ to enforce these policies?
- How do we ensure the policies have been enforced?

Data Access – Privacy

- Who has access to what parts of the data?
- Who has access to what aspects of the data?
- How do you ensure the privacy of the source?
- What are the access control and anonymization policies?
- How do we enforce these policies?
  - What techniques do we employ to enforce these policies?
- How do we ensure the policies have been enforced?
- How strong is the anonymization policy?
  - Is it possible to recover the anonymized information? If so, how hard?

Data Scale

- Nominal
  - Male, Female
  - Equality operation
- Ordinal
  - Very satisfied, satisfied, dissatisfied, and very dissatisfied
  - Inequalities operations
- Interval
  - Temperature, dates
  - Addition and subtraction operations
- Ratio
  - Mass, length, duration
  - Multiplication and division operations

Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Data Cleansing

Let’s get our hands dirty!!

The data set is from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

Data Cleansing – Common Issues

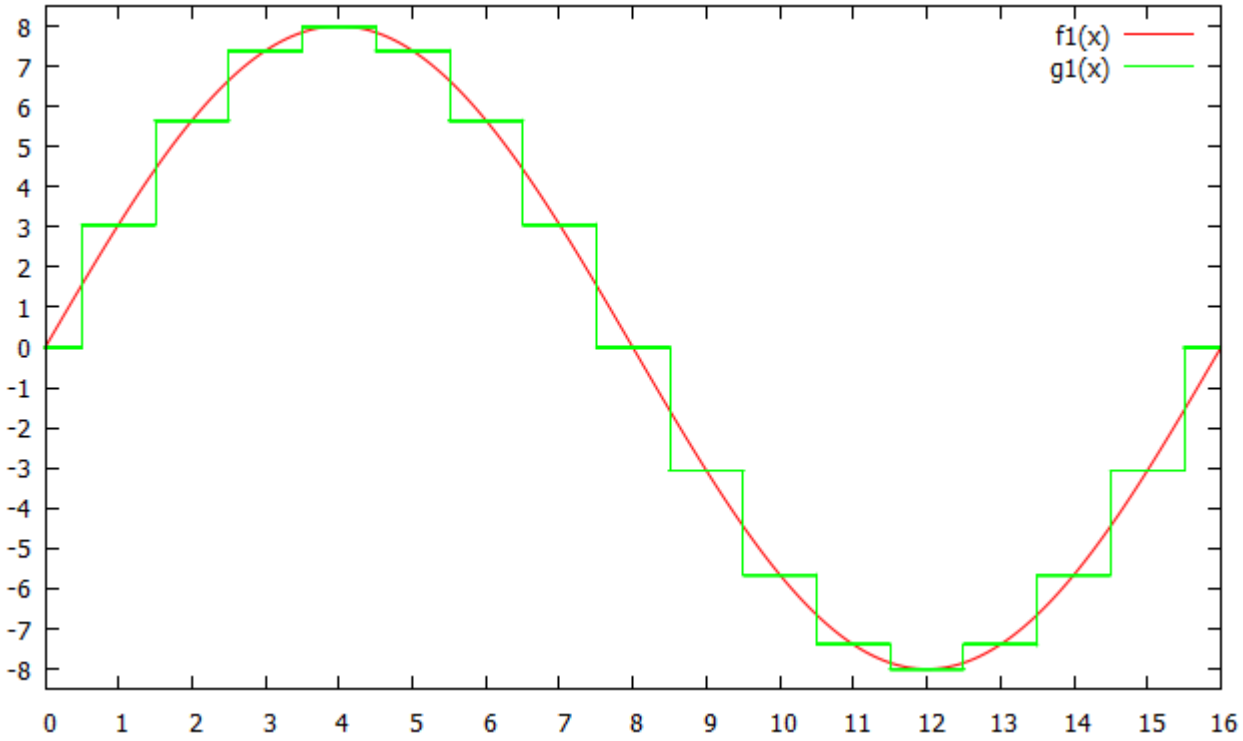
- Missing values
- Extra values
- Incorrect format
- Encoding
- File corruption
- Incorrect units
- Too much data
- Outliers
- Inliers

Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Data Transformation (Feature Engineering)

- Analyze specific aspects of the data
- Coarsening data
  - Discretization
  - Changing Scale
  - Normalization



Data Transformation (Feature Engineering)

- Analyze specific aspects of the data
- Coarsening data
  - Discretization
  - Changing Scale
  - Normalization

BMI	BMI Categories
< 18.5	Underweight
18.5 – 24.9	Normal Weight
25 – 29.9	Overweight
> 30	Obesity

Data Transformation (Feature Engineering)

- Analyze specific aspects of the data
- Coarsening data
  - Discretization
  - Changing Scale
  - Normalization

Actual Weight	Normalized
78	0.285
88	0.322
62	0.227
45	0.164

Data Transformation (Feature Engineering)

- Analyze relations between features of the data
- Synthesize new features
  - Relating existing features
  - Combining existing features

Data Transformation

Let’s get out hands dirty!!

Data Transformation (Feature Engineering)

Keep in mind the following:

- Scales
  - What the permitted operations?
- Data Collection
  - What is the trade-offs in data collection?
- Parsimony
  - Can we get away with simple scales?

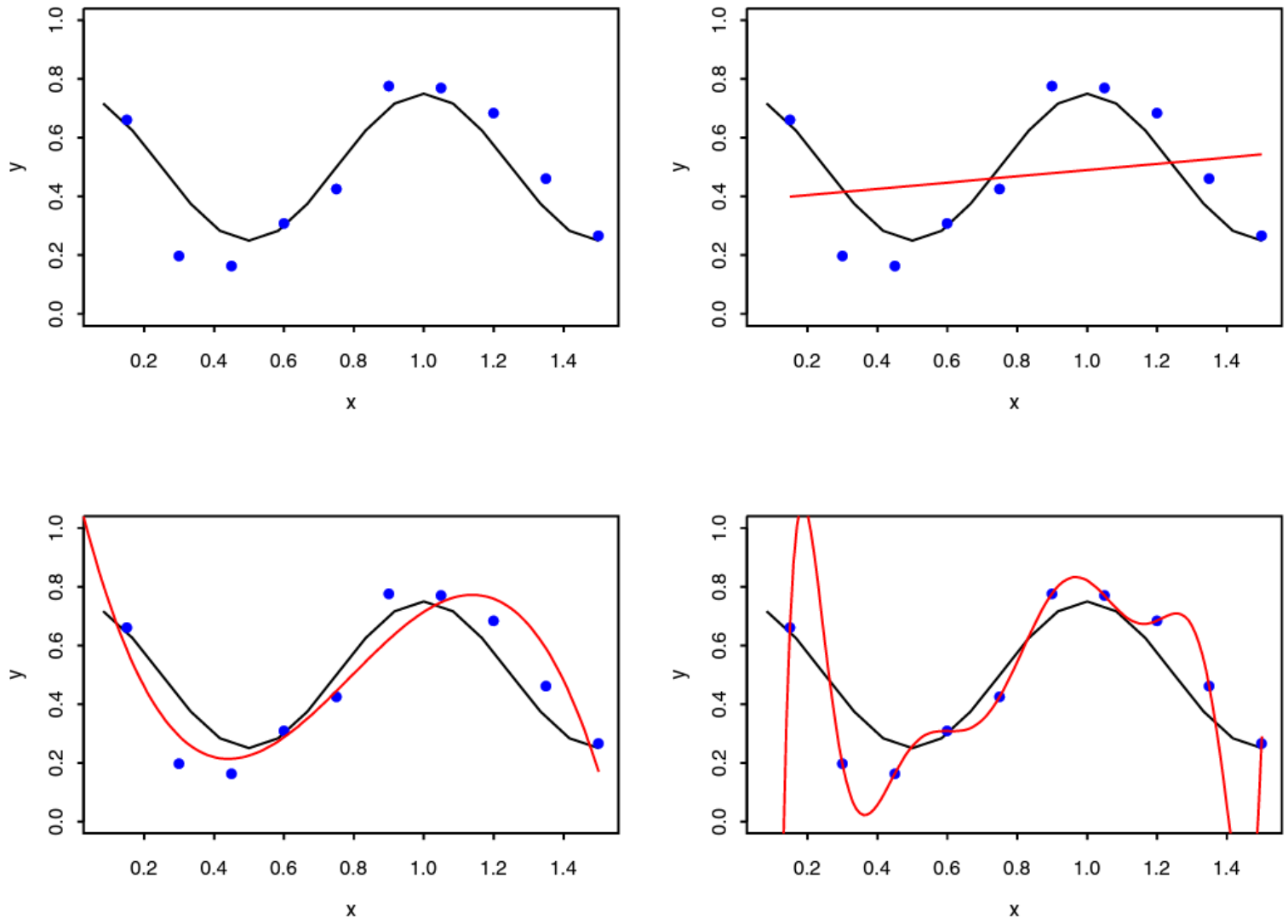
Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Data Analysis

- Features
  - Attributes of each datum
- Labels
  - Expert’s input about datum
- Data sets
  - Training
  - Validation
  - Test
- Work flow
  - Model building (training)
  - Model tuning and selection (validation)
  - Error reporting (test)

Data Analysis – Models



The figure is from the book “Modern Multivariate Statistical Techniques” by Alan Julian Izenman.

Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Result Validation – Approaches

- Expert Inputs
- Cross Validation
  - K-fold cross validation
  - 5x2 cross validation
  - Bootstrapping

Result Validation – Basic Terms

Consider a 2-class classification problem.

	Classification			
Actuals		X	Y	
	X	True X (tx)	False Y (fy)	p = tx + fy
	Y	False X (fx)	True Y (ty)	n = fx + ty
		p’ = tx + fx		N = p + n

Result Validation – Basic Terms

Now, consider X as positive evidence and Y as negative evidence.

	Classification			
Actuals		X	Y	
	X	True Positive (tp)	False Negative (fn)	p = tp + fn
	Y	False Positive (fp)	True Negative (tn)	n = fp + tn
		p’ = tp + fp		N = p + n

Result Validation – Measures

error = (fp + fn) / N

accuracy = (tp + tn) / N

sensitivity = tp / p = tp-rate

specificity = tn / n = 1 – fp-rate

tp-rate = tp / p

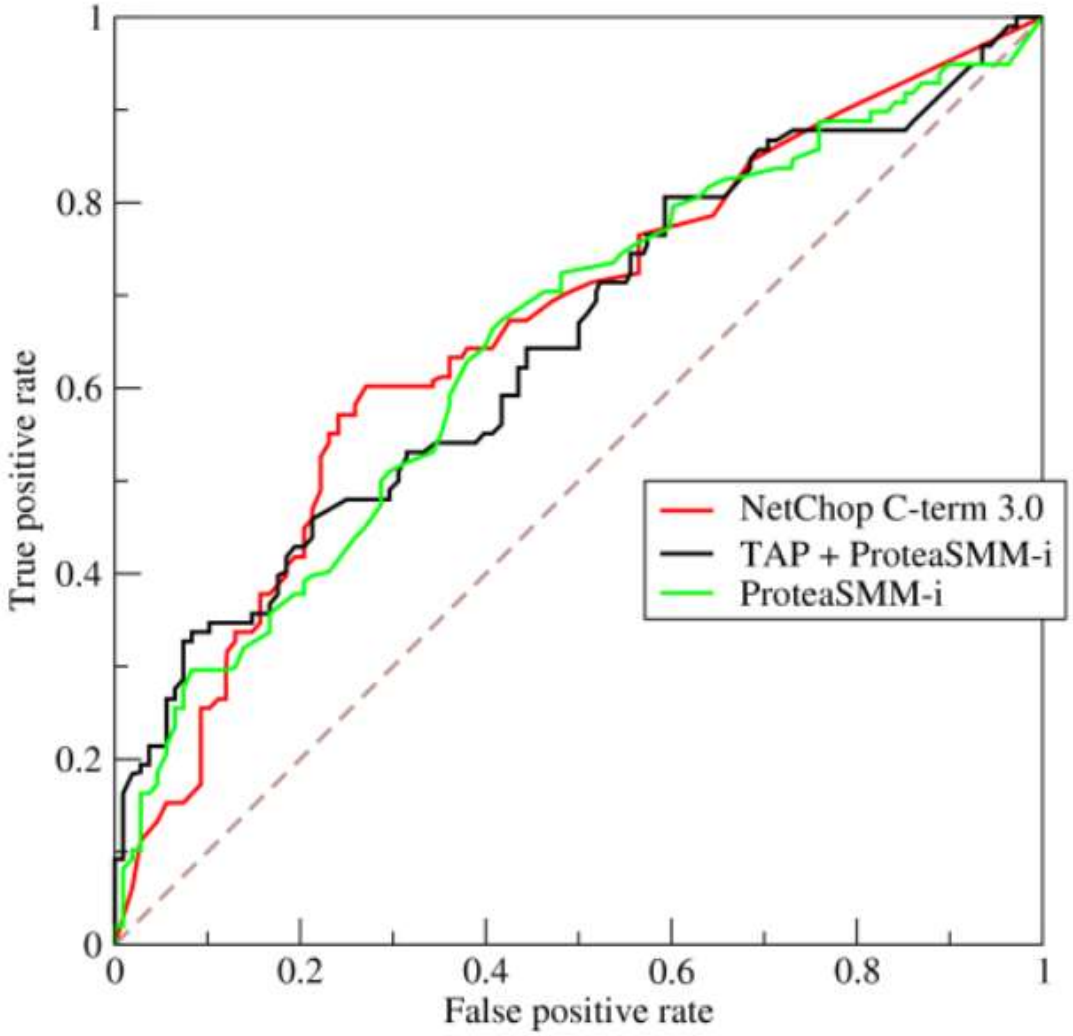
fp-rate = fp / n

precision = tp / p'

recall = tp / p = tp-rate

	Classification			
Actuals		X	Y	
	X	True Positive (tp)	False Negative (fn)	p = tp + fn
	Y	False Positive (fp)	True Negative (tn)	n = fp + tn
		p' = tp + fp		N = p + n

Result Validation – ROC (Receiver Operating Characteristics)



The figure is from the Wikipedia page about “Receiver Operating Characteristics”.

Result Validation – Class Confusion Matrix

	A	B
A	True A (ta)	False A (fa)
B	False A (fa)	True B (tb)

2 class problem

4 class problem

	A	B	C	D
A	True A (ta)	False B (fb)	False C (fc)	False D (fd)
B	False A (fa)	True B (tb)	False C (fc)	False D (fd)
C	False A (fa)	False B (fb)	True C (tc)	False D (fd)
D	False A (fa)	False B (fb)	False C (fc)	True D (td)

Result Validation – Bias and Variance

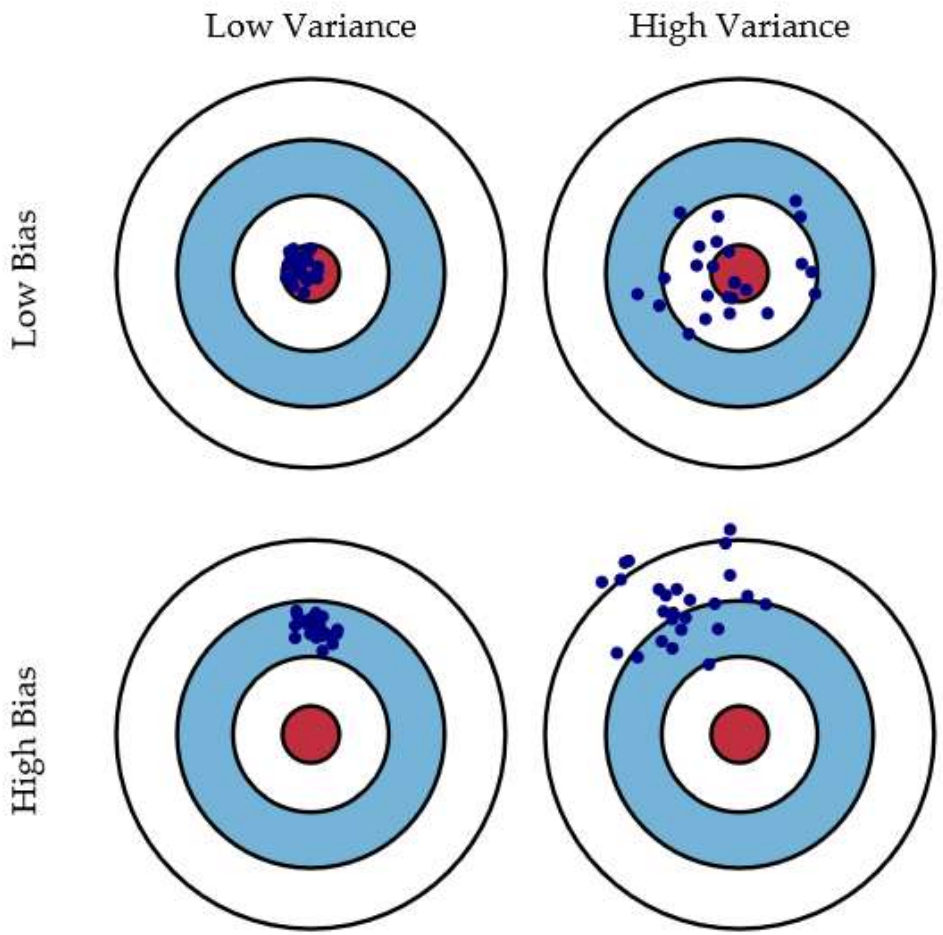
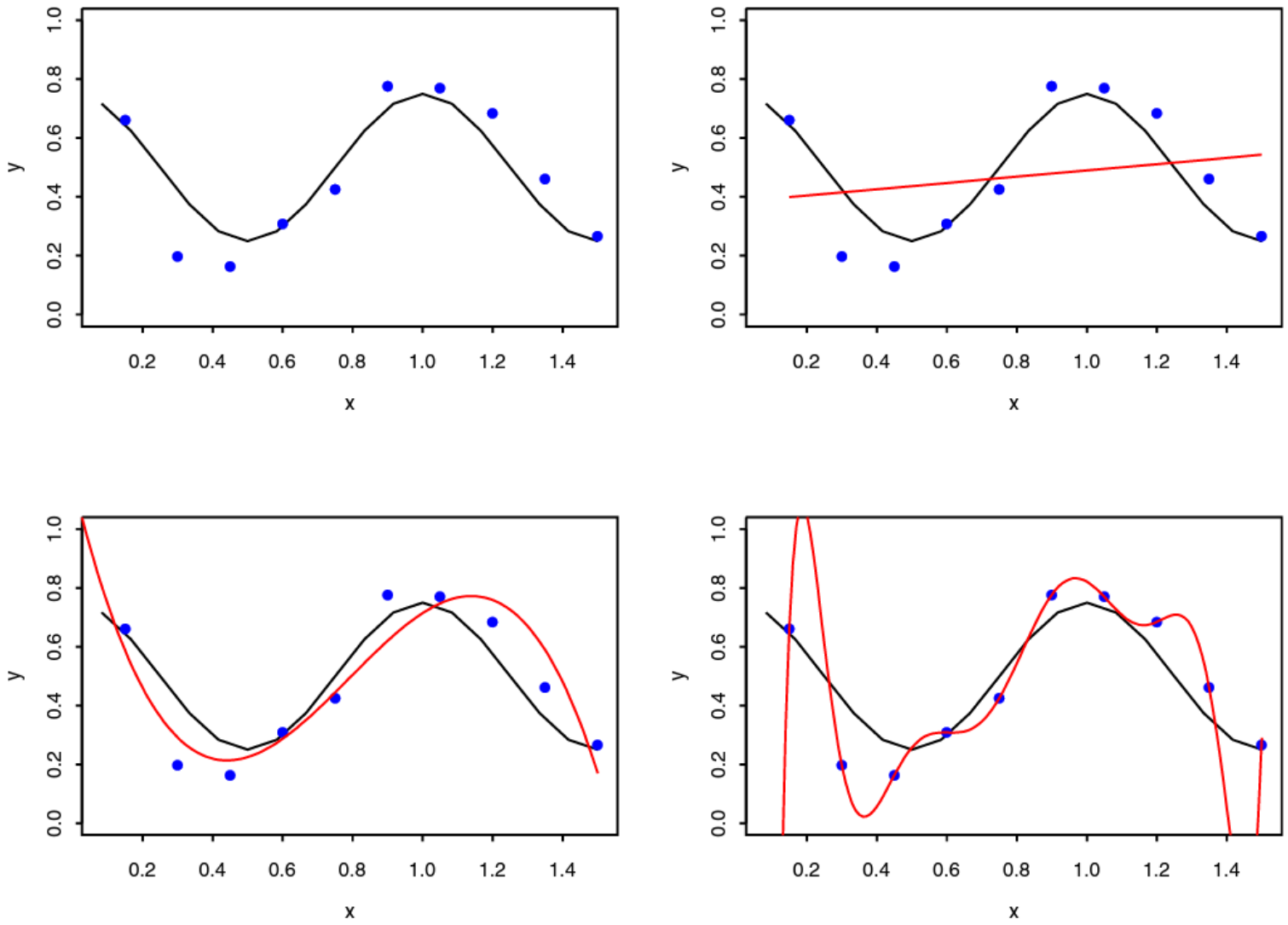


Fig. 1 Graphical illustration of bias and variance.

The figure is from http://scott.fortmann-roe.com/docs/BiasVariance.html.

Result Validation – Underfitting & Overfitting



The figure is from the book “Modern Multivariate Statistical Techniques” by Alan Julian Izenman.

Result Validation

Let’s get out hands dirty!!

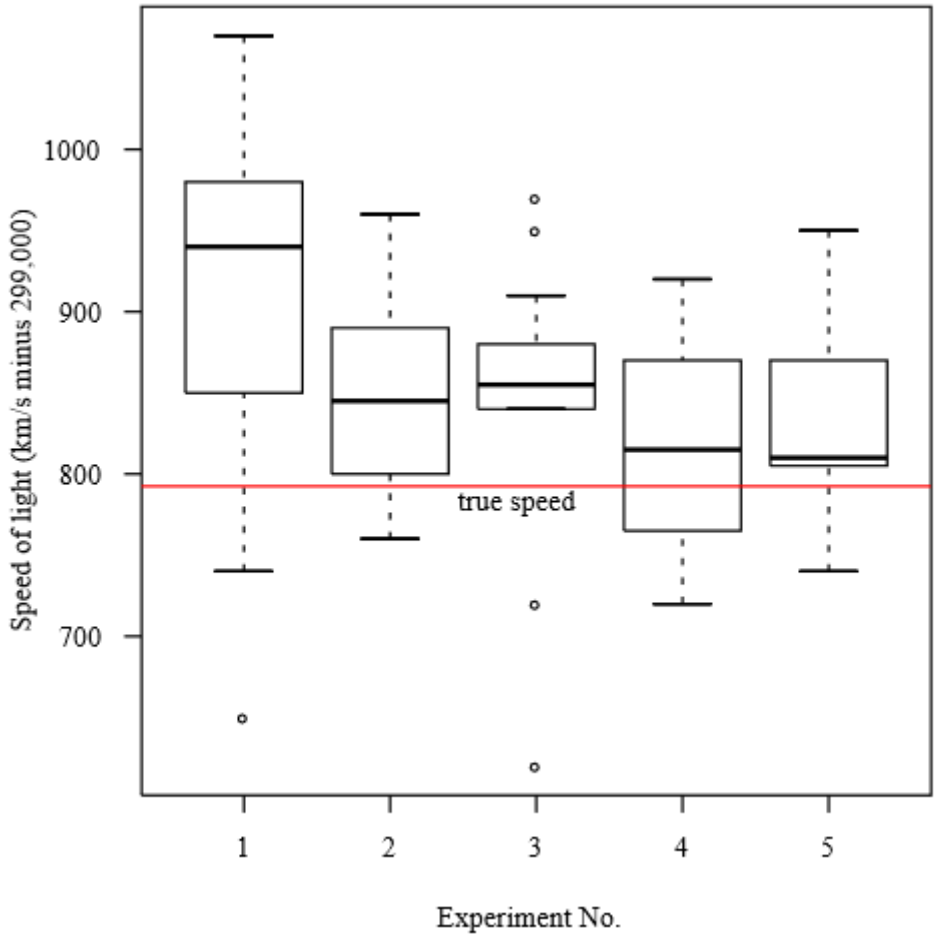
Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Result Presentation (Visual Validation)

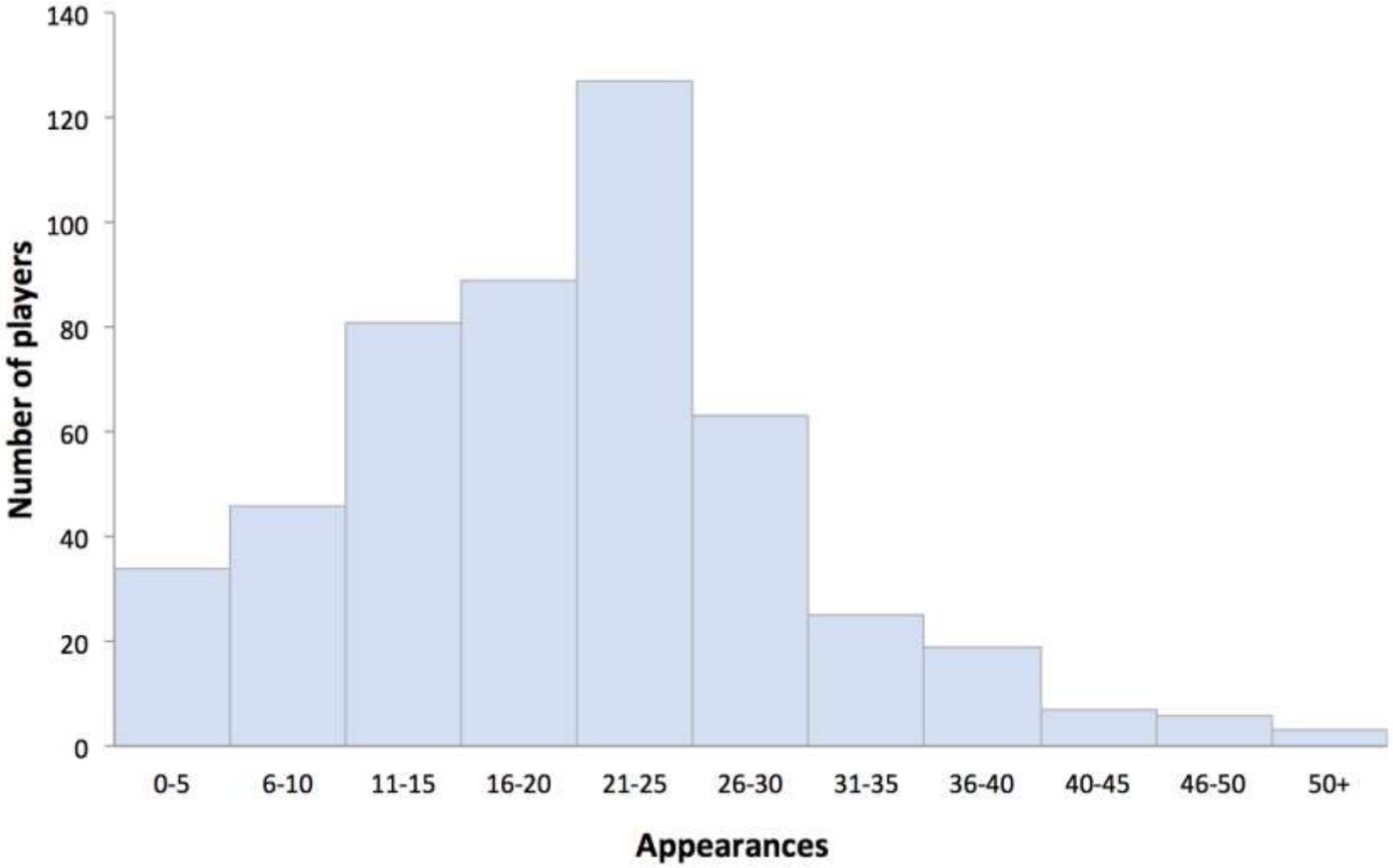
- Numbers
  - Central tendencies – mode, median, and mean
  - Dispersion – range, standard deviation
  - Five number summary
    - min, 1st quartile, median (mean), 3rd quartile, max
  - Margin of error
  - (Confidence) Interval
- Tables
- Charts

Result Presentation – Box-and-Whisker Plots



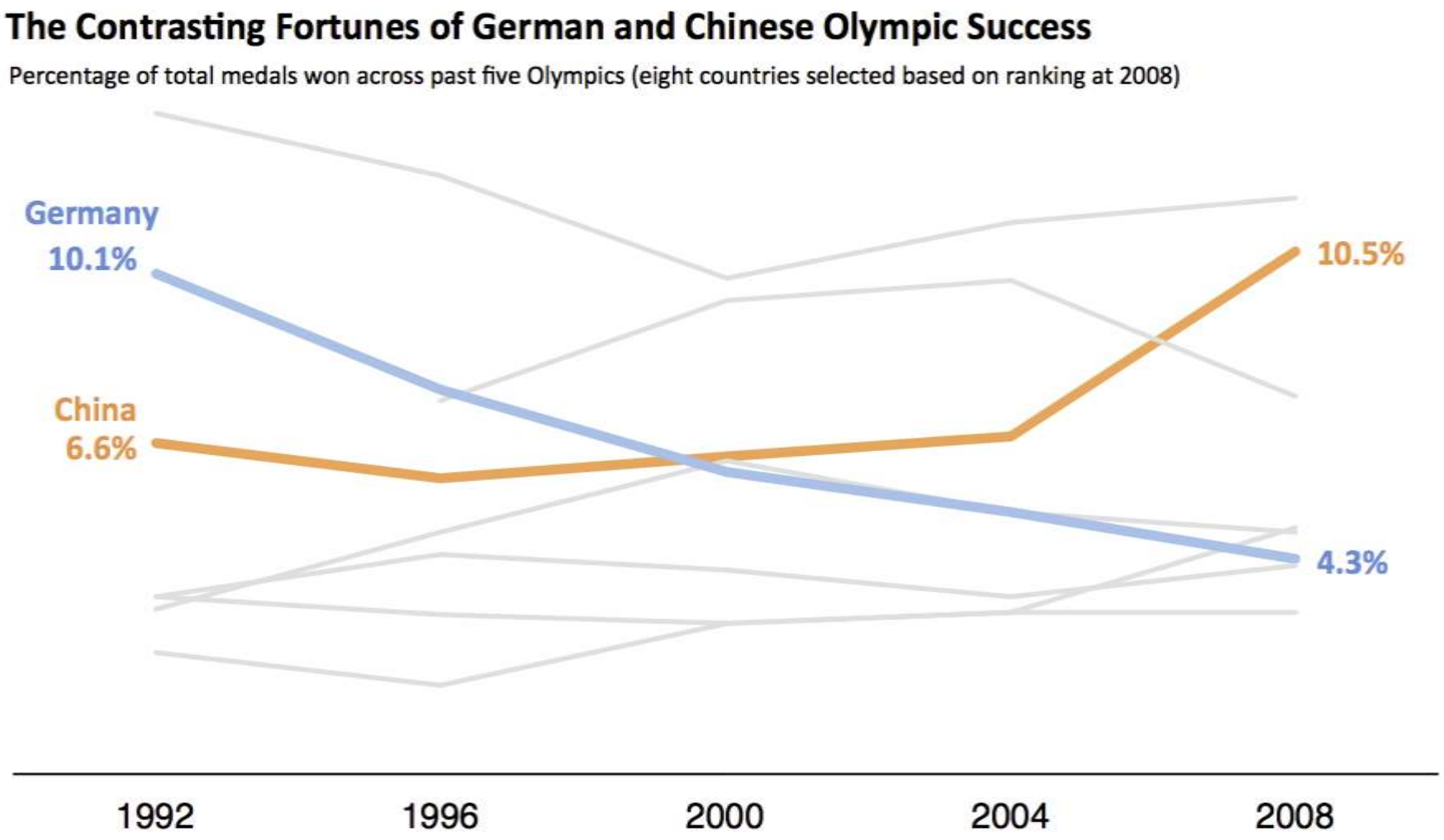
The figure is from the Wikipedia page about “Box Plot”.

Result Presentation – Histograms



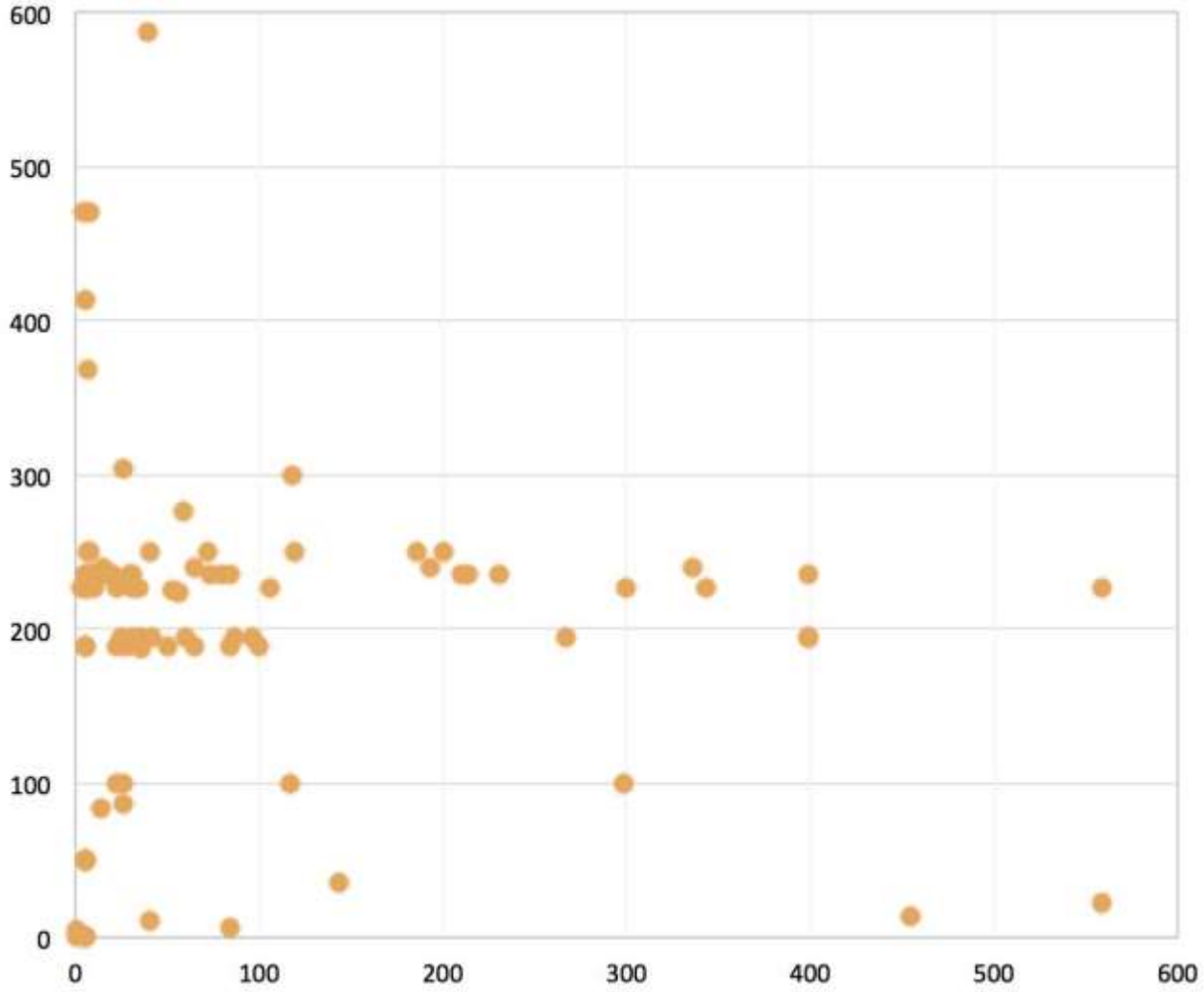
The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Result Presentation – Line Graphs



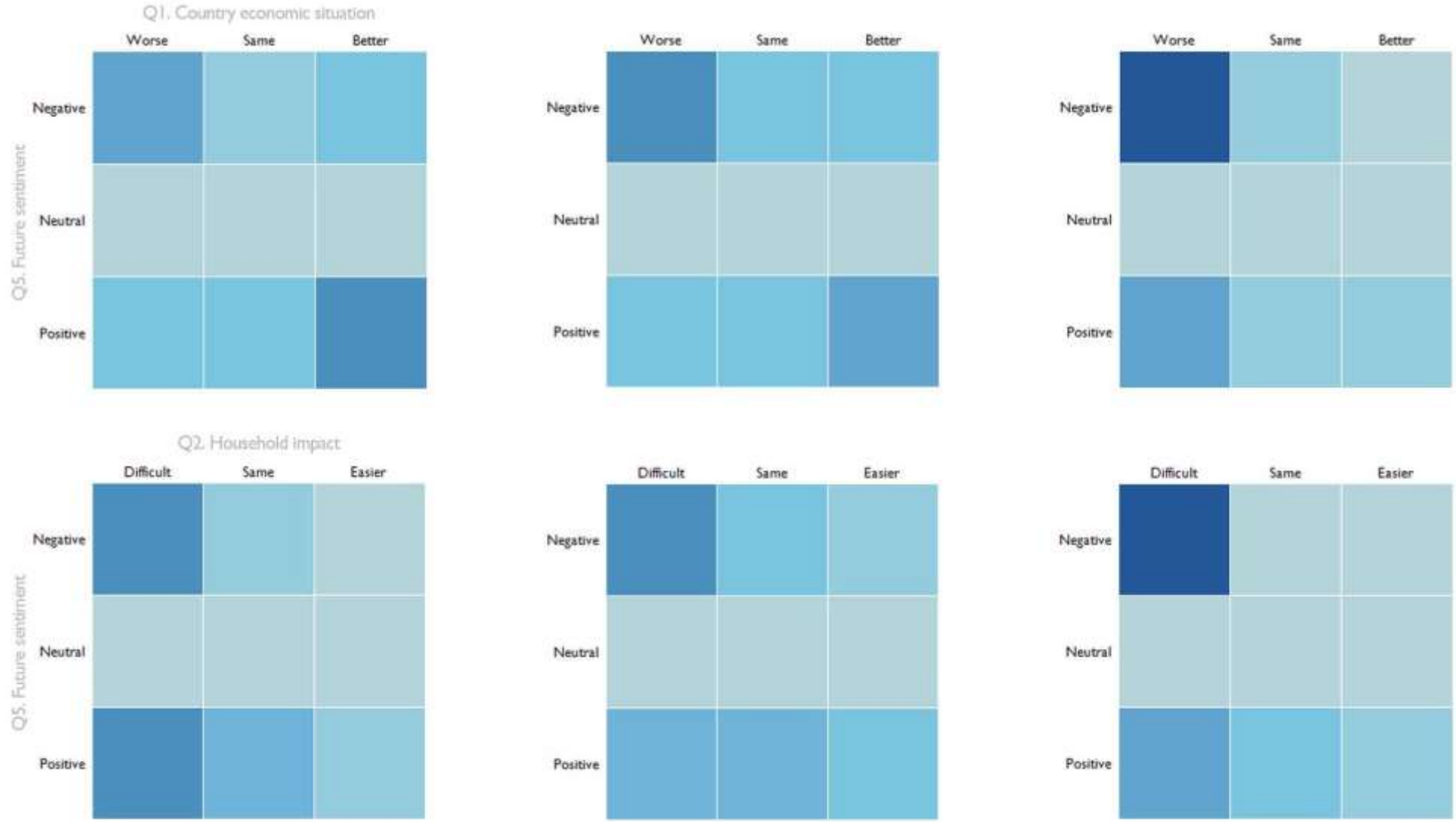
The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Result Presentation – Scatter Plots



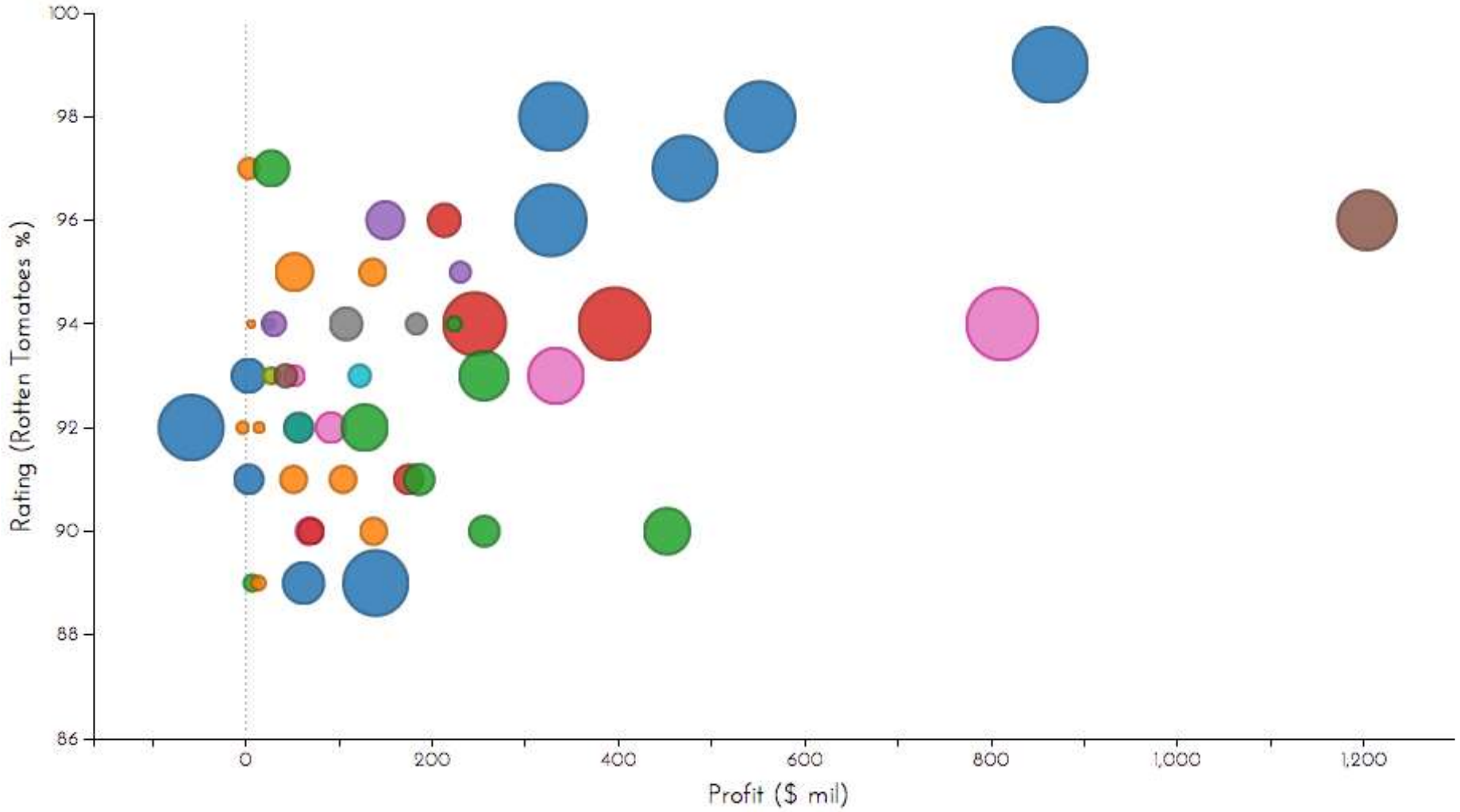
The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Result Presentation – Heatmaps



The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Result Presentation – Bubble Chart



The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Result Presentation – Word Cloud

The figure is from the book “Data Visualization: A Successful Design Process” by Andy Kirk.

Typical Data Analytics Work Flow

1. Identify Issue
2. Data Collection, Storage, Representation, and Access
3. Data Cleansing
4. Data Transformation
5. Data Analysis (Processing)
6. Result Validation
7. Result Presentation (Visual Validation)
8. Recommend Action / Make Decision

Now, are you thinking..

- What about identifying the issue/question?
  - Know where you are going before you start
- What about recommending action / making the decision?
  - Information and knowledge aren't the same
- Are data X tasks really that important and hard?
  - Garbage in, Garbage out
- Aren't data analysis techniques the most important?
  - Smart data (structures) and dumb code works better than the other way around

Some Personal Observations

- Domain Knowledge is crucial
  - Optimizing analysis
  - Improving relevance of results
- Always prefer
  - Simple solutions over complex solutions
  - Fast solutions over slow solutions
  - Most correct solutions over fully correct solutions (even no solution!!)
- If you hear “I want it now”, then say “Really? Please Explain”
- Visualization helps only if the results are relevant
- Limited data sets can only get you so far
- Security and privacy are more important than you think

Got any stories from the trenches?