O'REILLY®

Security

BUILD BETTER DEFENSES

oreillysecuritycon.com

#oreillysecurity

Foundations of Security Data Science

Data Visualization

Jay Jacobs
Charles Givre

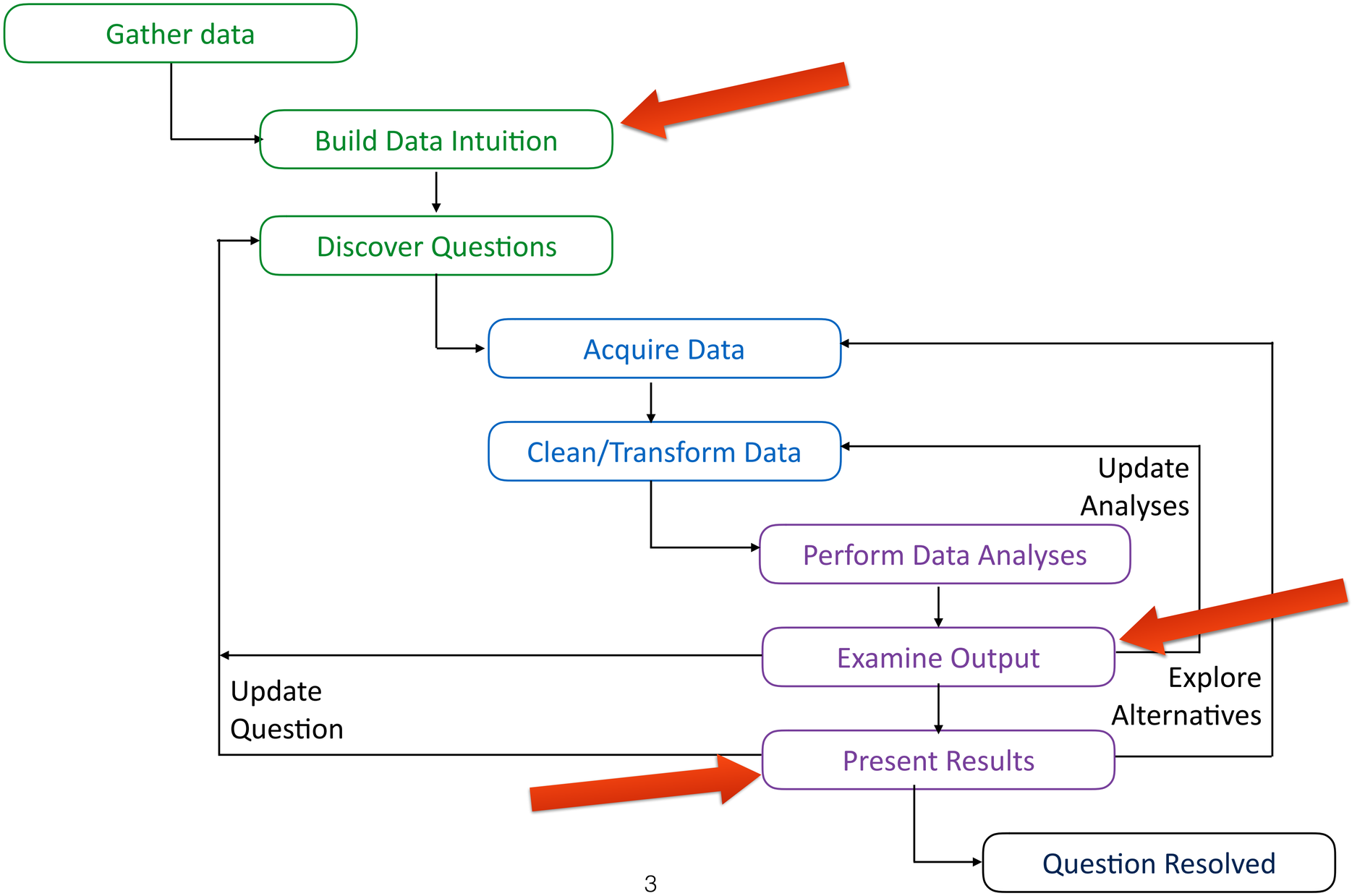
Communication

“The true promise of the information age isn’t tons of data but decisions and actions that are better because they’re based on an understanding of what’s really going on in the world.

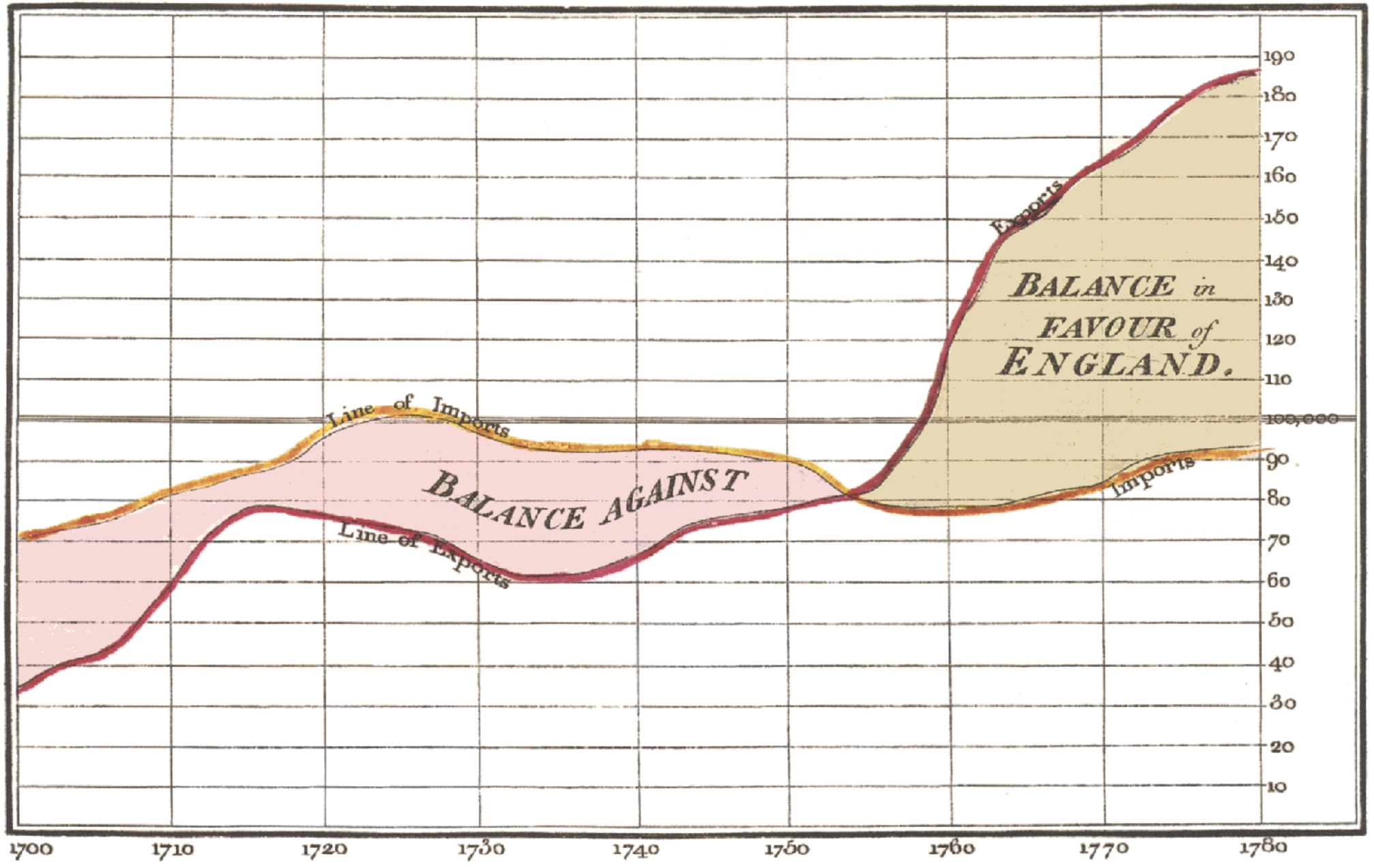
Any knowledge you gain that could be used to make better decisions will amount to nothing if you can’t communicate it to others in a way that makes sense to them.

Stephen Few
“Show me the Numbers”

Reality of Research Workflow



Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

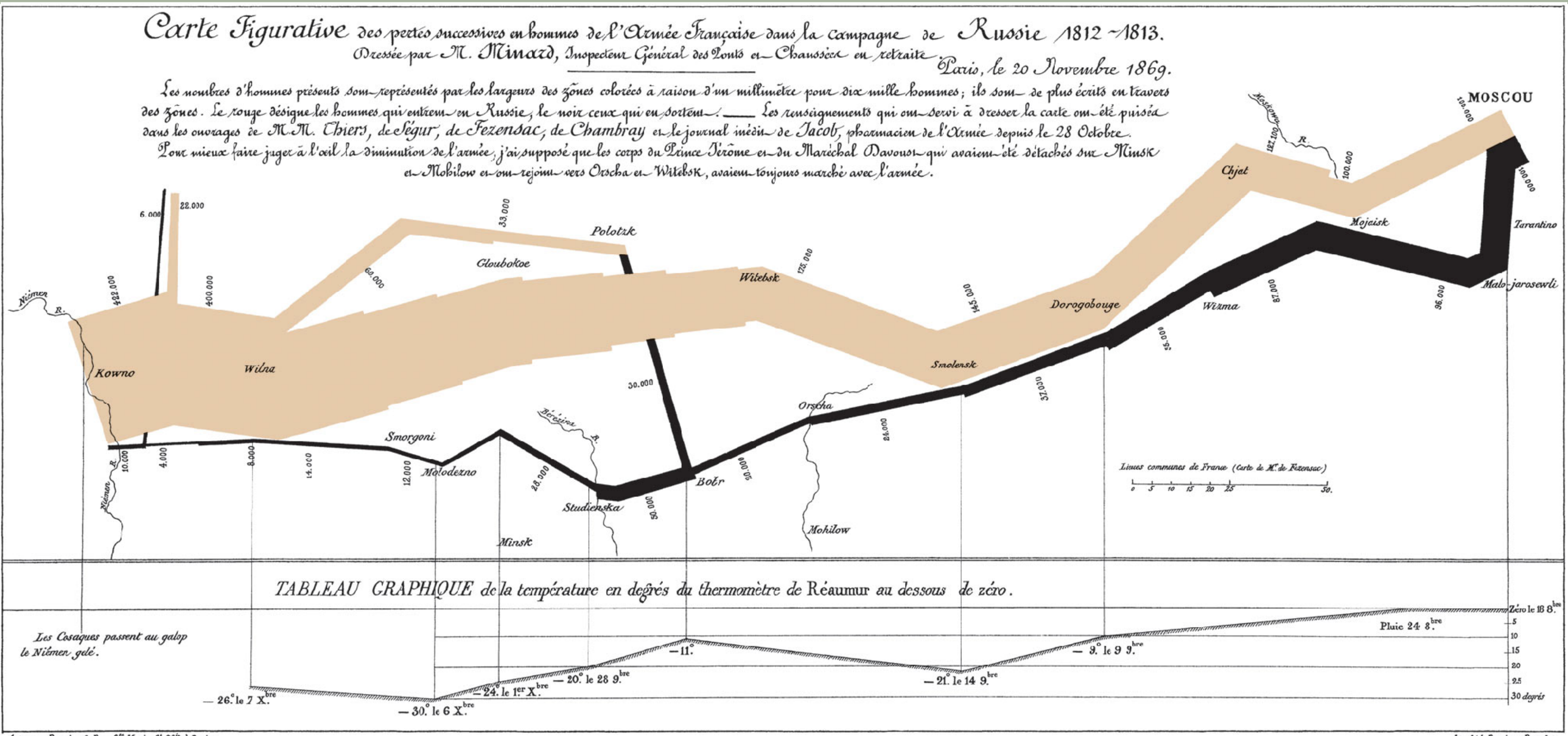


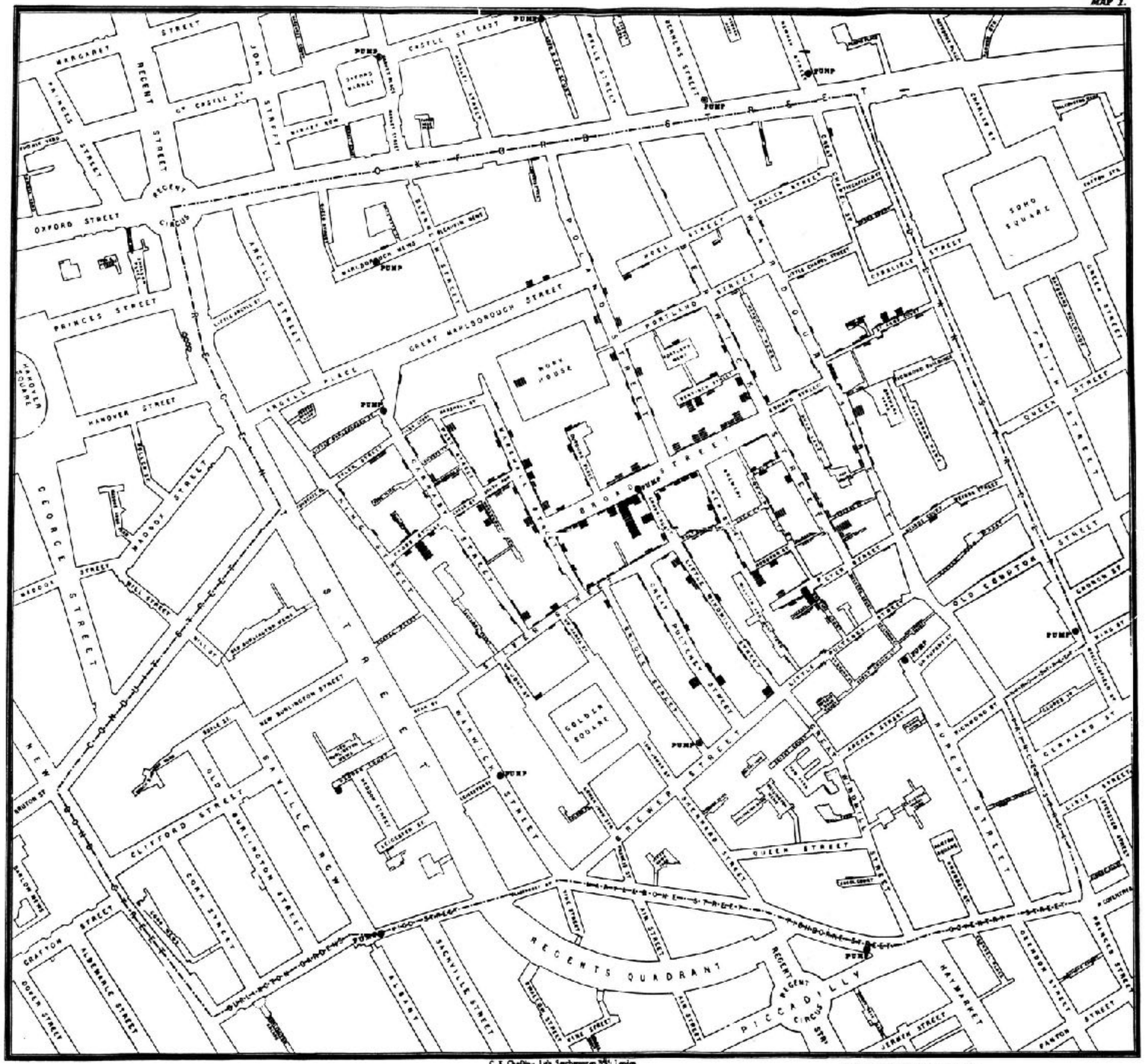
The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 1st May 1786. by Wm. Playfair

Neele sculpt 352, Strand, London.

Napoleon's March to Moscow, The War of 1812

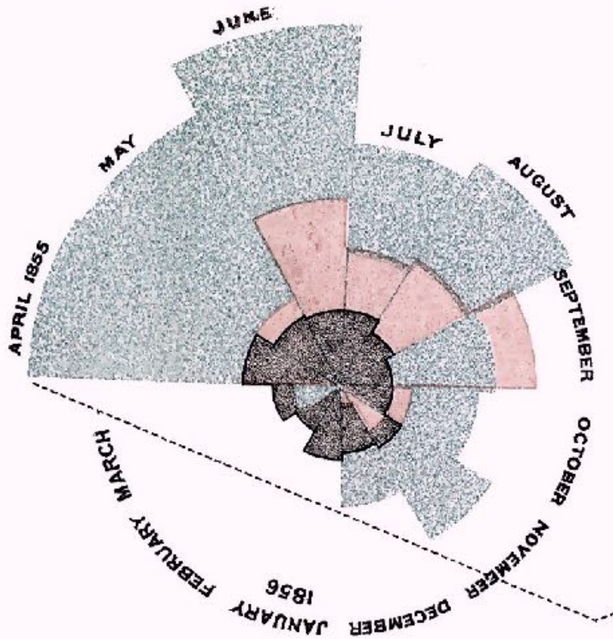



MAP 1.
OXFORD STREET
REGENT STREET
PRINCES STREET
HANOVER STREET
GREAT MARLBOROUGH STREET
BROAD STREET
PORTLAND STREET
GEORGE STREET
ST JAMES'S STREET
REGENTS QUADRANT
PICCADILLY
REGENT CIRCUS
QUEEN STREET
GOLDEN SQUARE
SOHO SQUARE
BURY HOUSE
PUMP
C.F. Cheffins, Ltd. Southampton St. London
SCALE 80 INCHES TO A MILE.

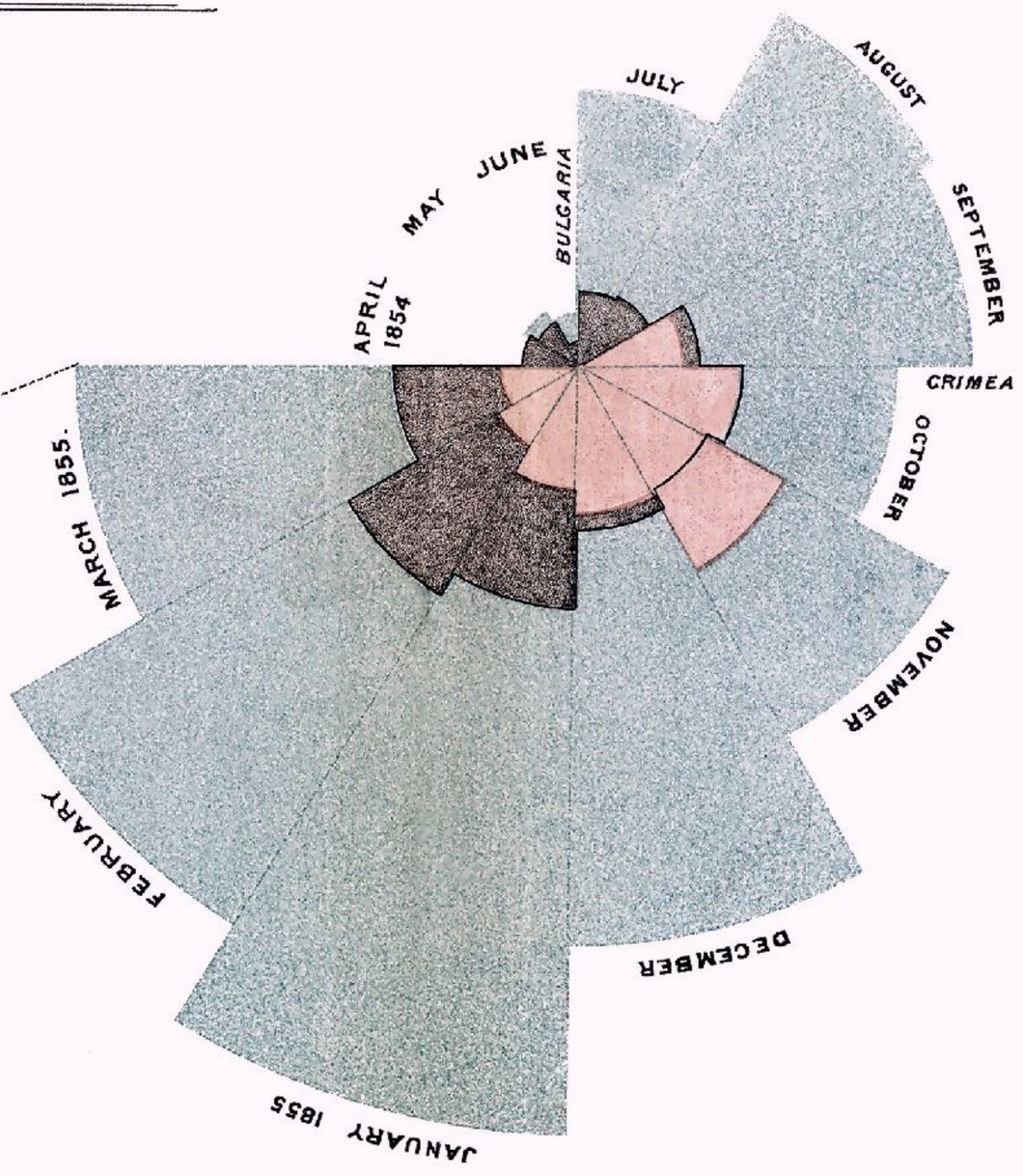
MARLBOROUGH
PUMP
GREAT MARLBOROUGH STREET
ARGYLL PLACE
LITTLE ARGYLL ST
PUMP
WORK HOUSE
PORTLAND STREET
PORTLAND MEWS
LITTLE CHAPEL STREET
CARLISLE
ST ANNS COURT
RICHMOND BUILDINGS
RICHMOND MEWS
MARSHALL ST
TYLER COURT
TYLER STREET
MARSHALL STREET
DUFOURS PLACE
WEST STREET
CARNABY
KING STREET
CROSS STREET
CROSS ST
SOUTH ROW
BROAD STREET
PUMP
BREWERY
NEW STREET
DUCK LANE
HEARDS STREET
HEARDS COURT
HUSBAND ST
PETER STREET
OLD COMPTON
SILVER STREET
BRIDLE STREET
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
WILLIAM AND MARY YARD
LITTLE PULTENEY
PUMP
UP RUPERT ST
UP JAMES ST
UP JOHN ST
OLD BURLINGTON MEWS
BURLINGTON STREET
STREET
STRE

DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 to MARCH 1856.



1.
APRIL 1854 to MARCH 1855.



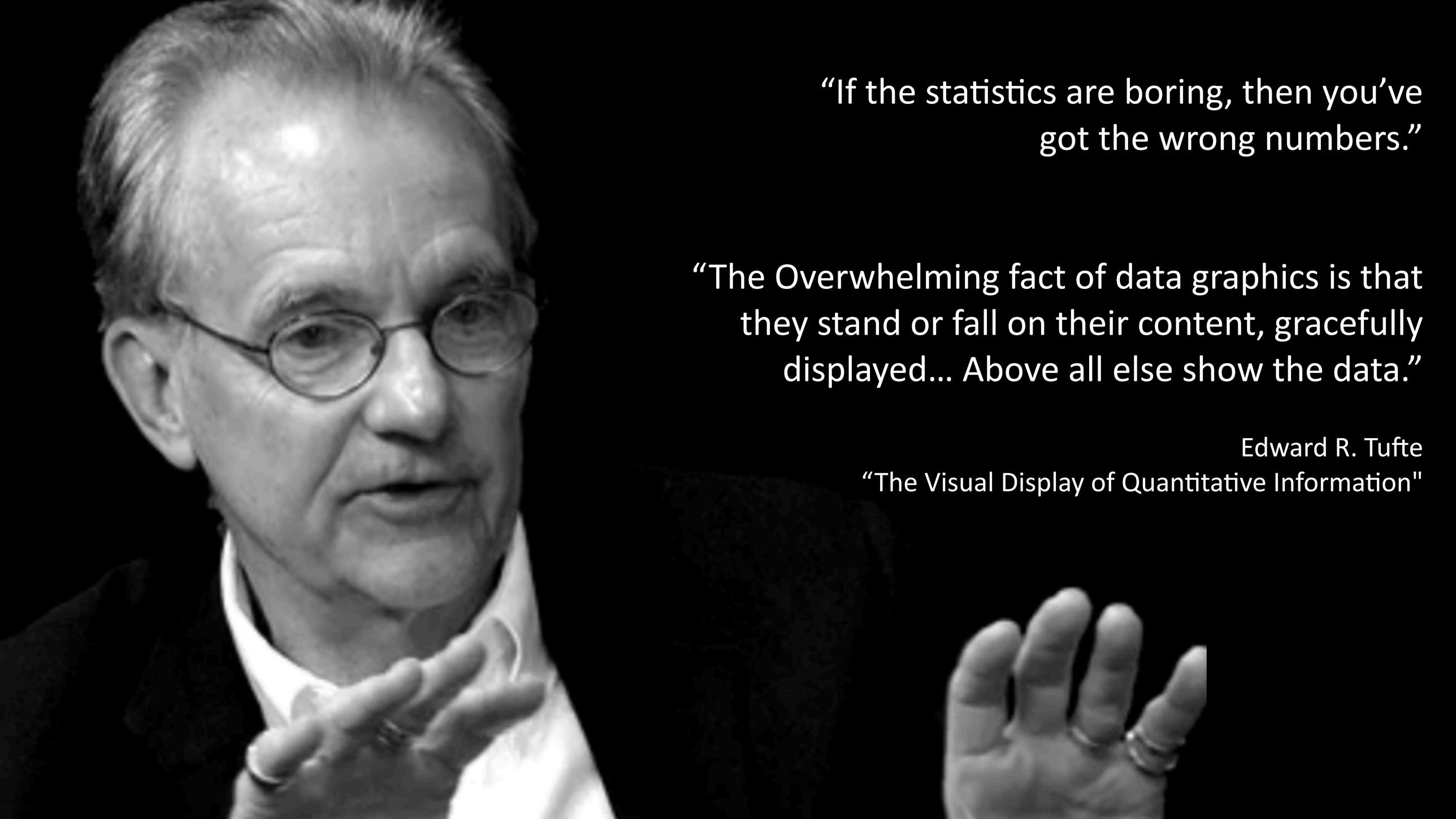
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.



“If the statistics are boring, then you’ve got the wrong numbers.”

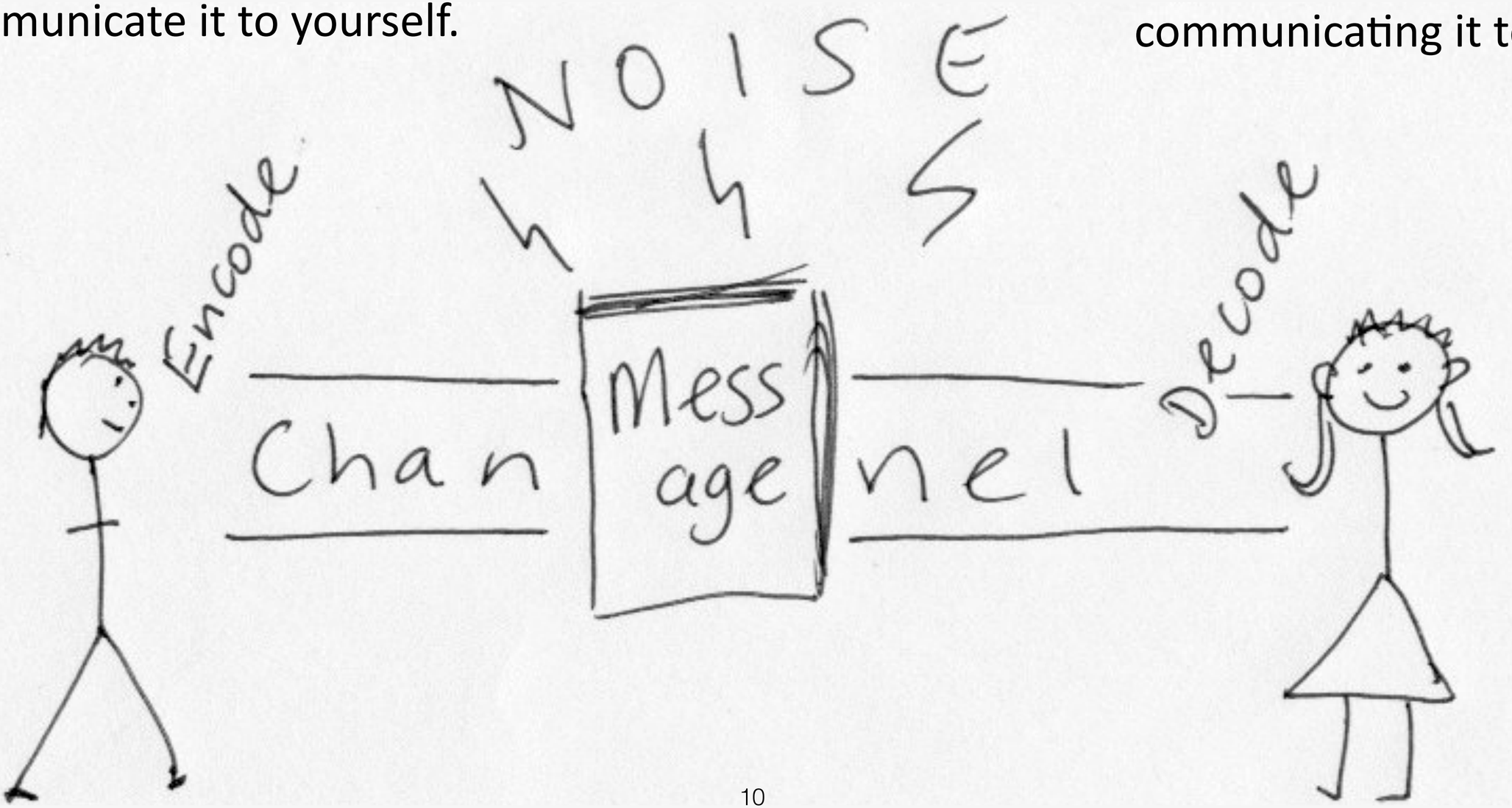
“The Overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed… Above all else show the data.”

Edward R. Tufte
“The Visual Display of Quantitative Information"

Visualization is Communication

For yourself: To see into the data and communicate it to yourself.

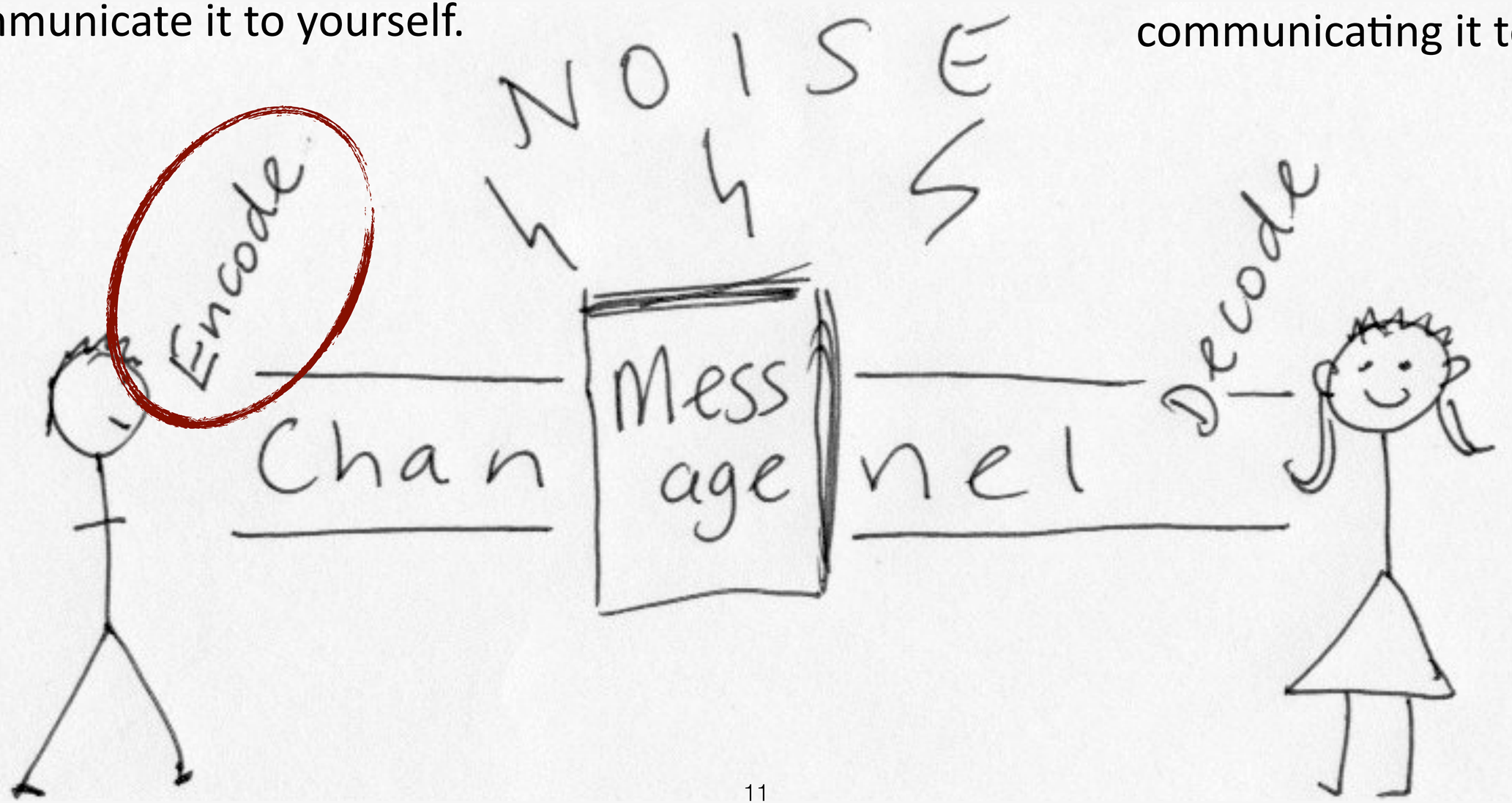
For Others: After finding a story, communicating it to others.



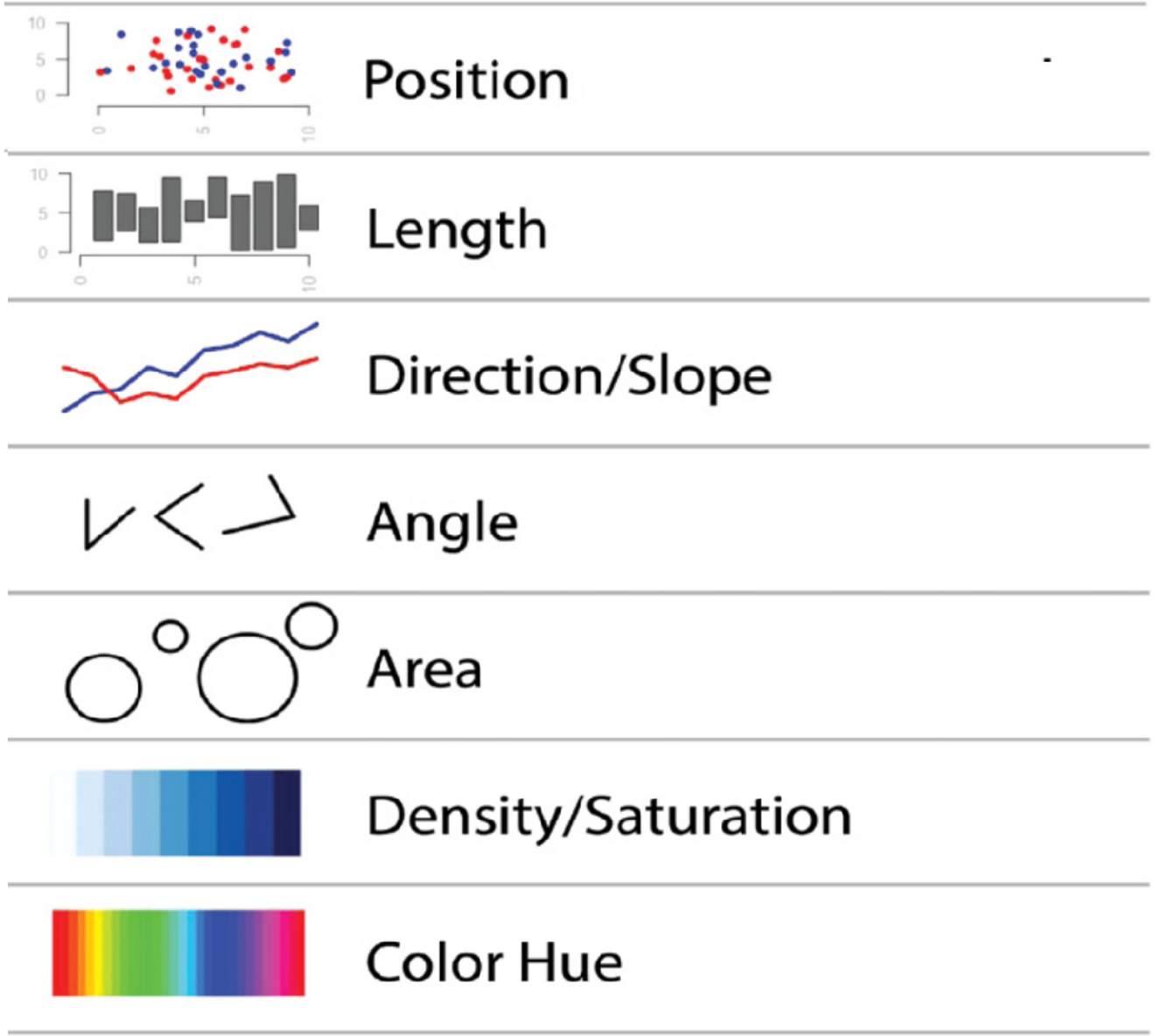
Visualization is Communication

For yourself: To see into the data and communicate it to yourself.

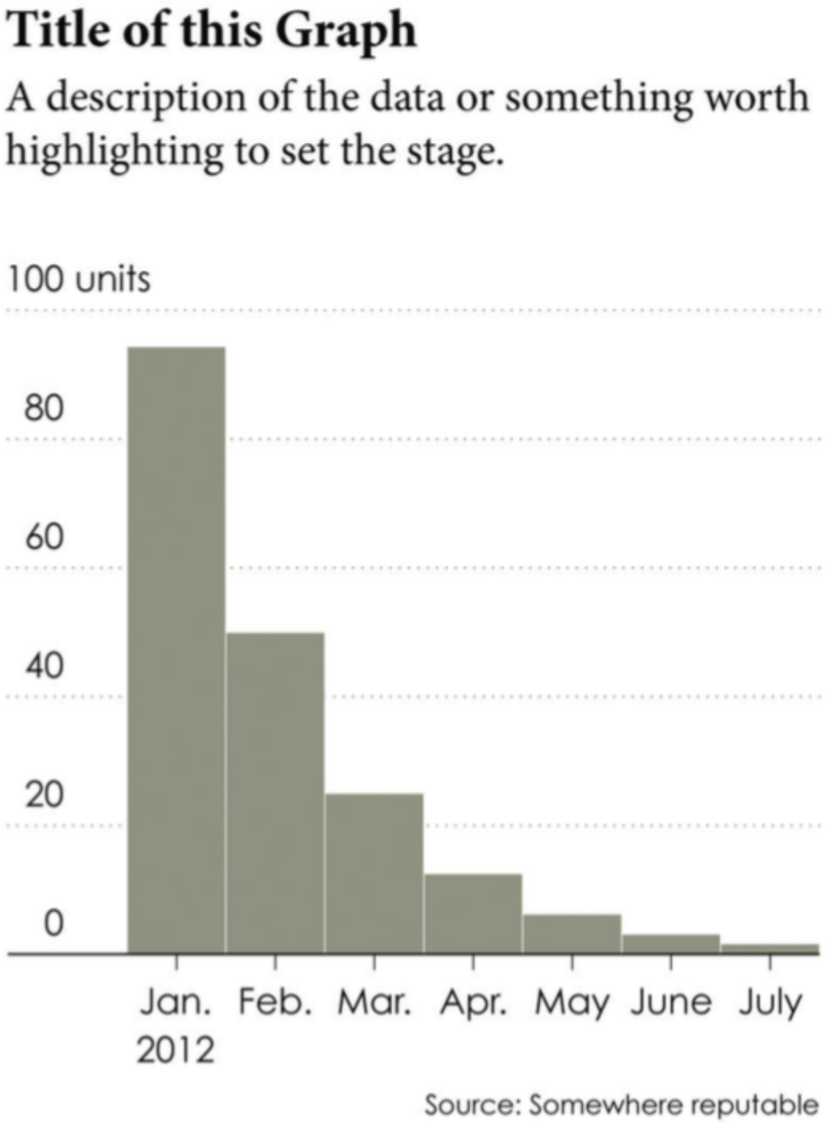
For Others: After finding a story, communicating it to others.



Visual Cues

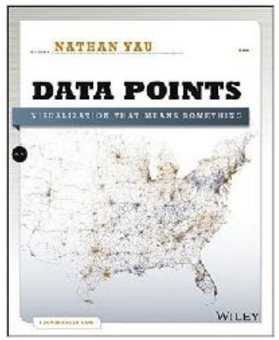

Position
Length
Direction/Slope
Angle
Area
Density/Saturation
Color Hue

Building Blocks

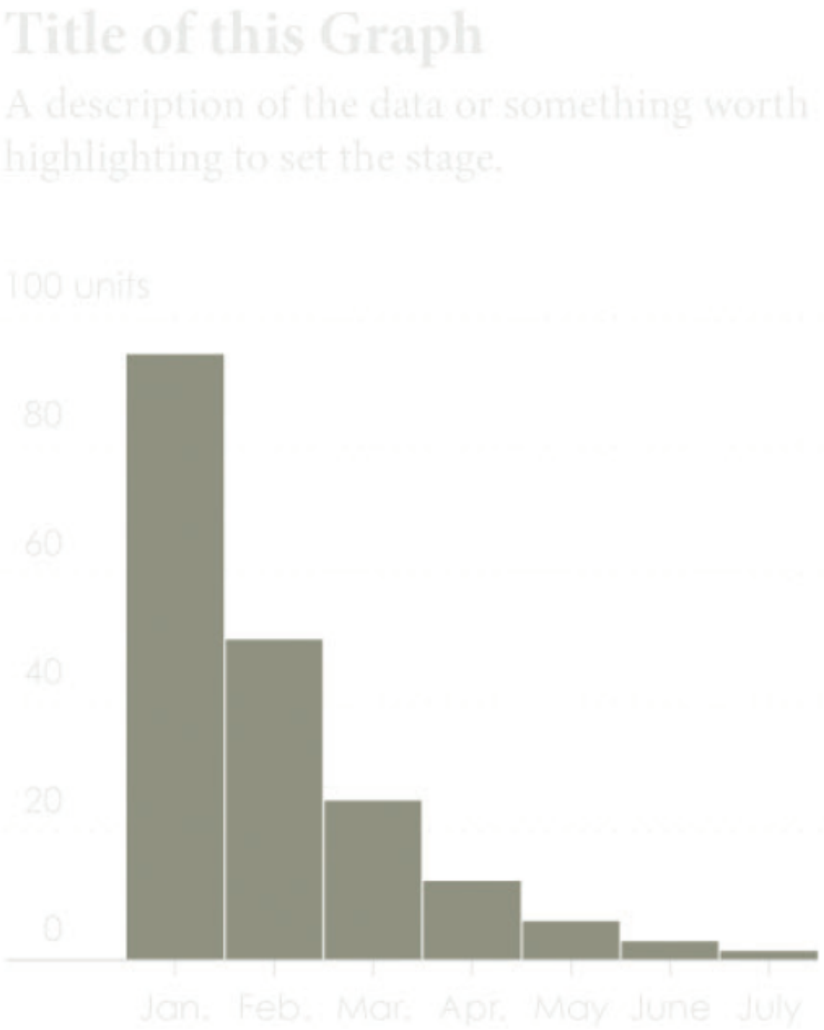


Visualizations are made up of multiple pieces working together.

From Nathan Yau

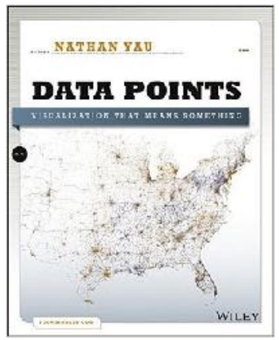


Visual Cues

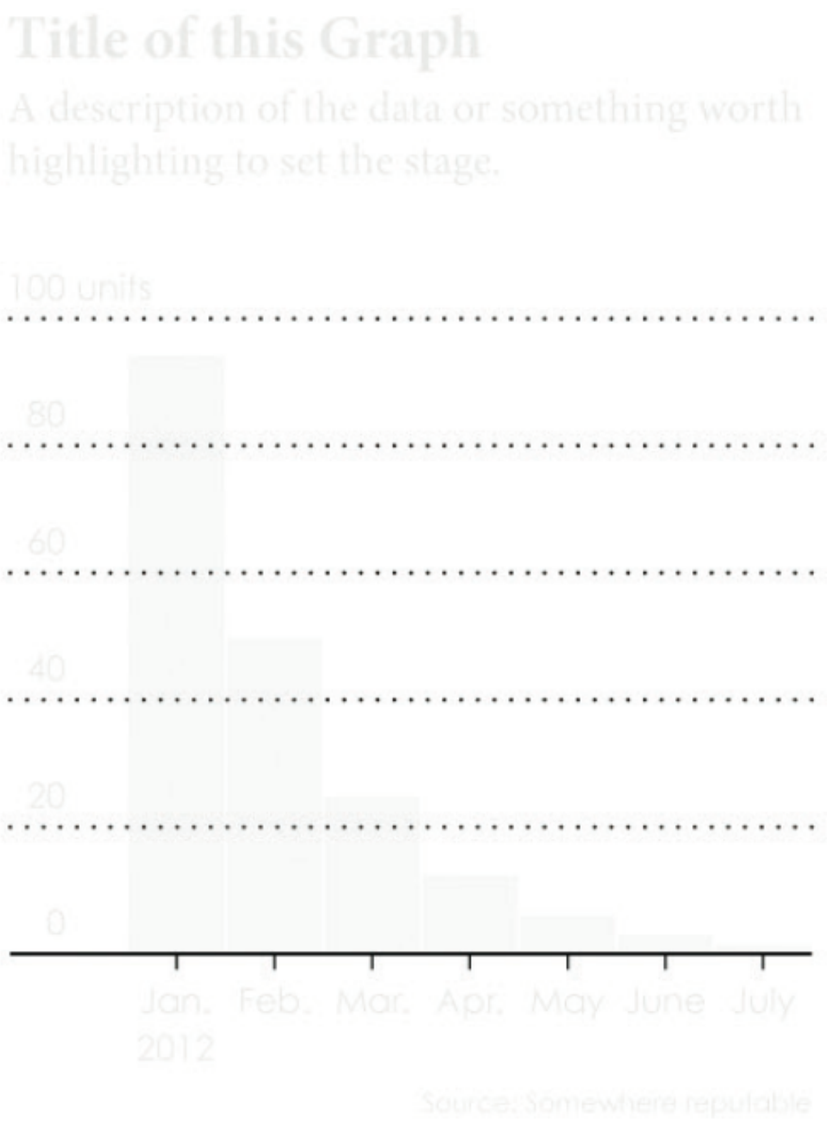


"Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals."

From Nathan Yau

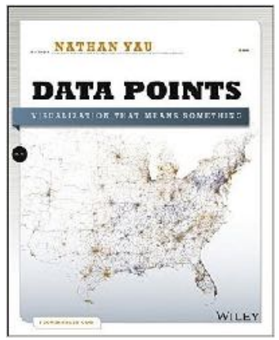


Coordinate System



Scatter plot versus pie chart, map projections use a variety of different coordinate systems.

From Nathan Yau



Map Projections

Equirectangular

Typically used for thematic mapping, but doesn't preserve area or angle

Albers

Scale and shape not preserved; angle distortion is minimal

Mercator

Preserves angles and shapes in small areas, making it good for directions

Lambert conformal conic

Better for showing smaller areas and often used for aeronautical maps.

Sinusoidal

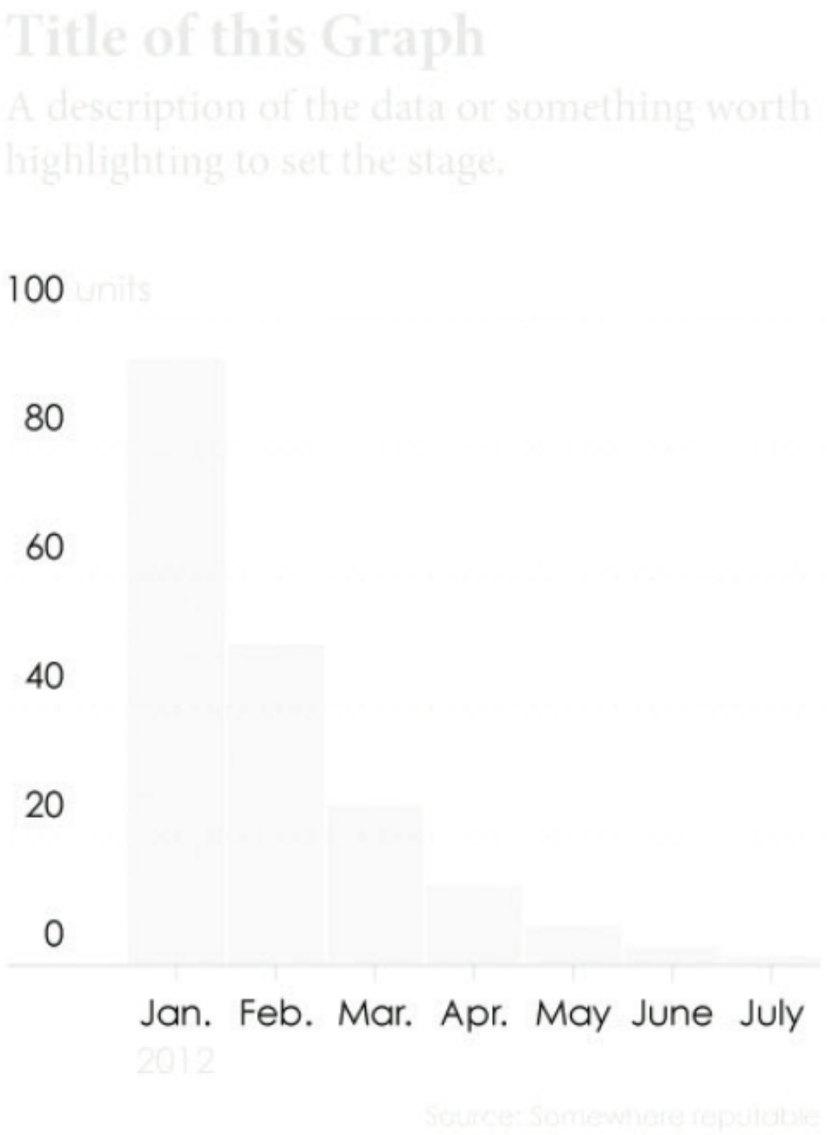
Preserves area; useful for areas near the prime meridian

Polyconic

Was often used to show US in the mid-1900s; little distortions in small areas near merdian

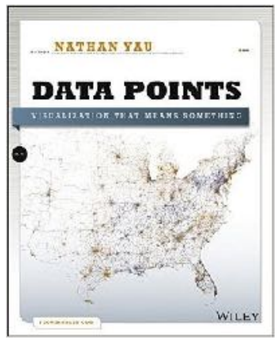
Source: Natural Earth

Scale



Linear, logarithmic, categorical, time, spatial, and so on.

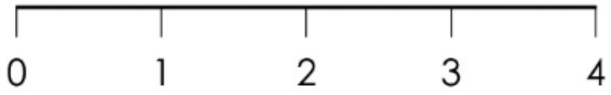
From Nathan Yau



Scales

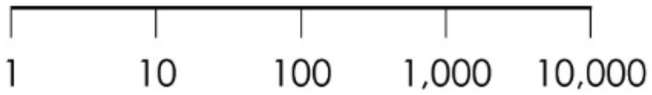
Linear

Values are evenly spaced



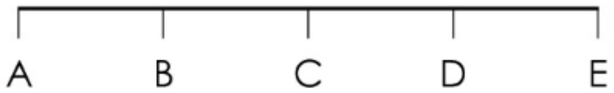
Logarithmic

Focus on percent change



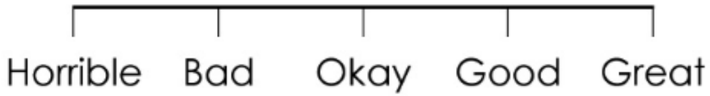
Categorical

Discrete placement in bins



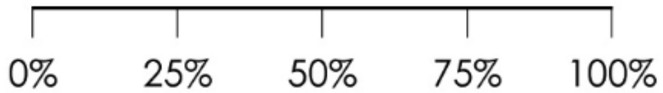
Ordinal

Categories where order matters



Percent

Representing parts of a whole

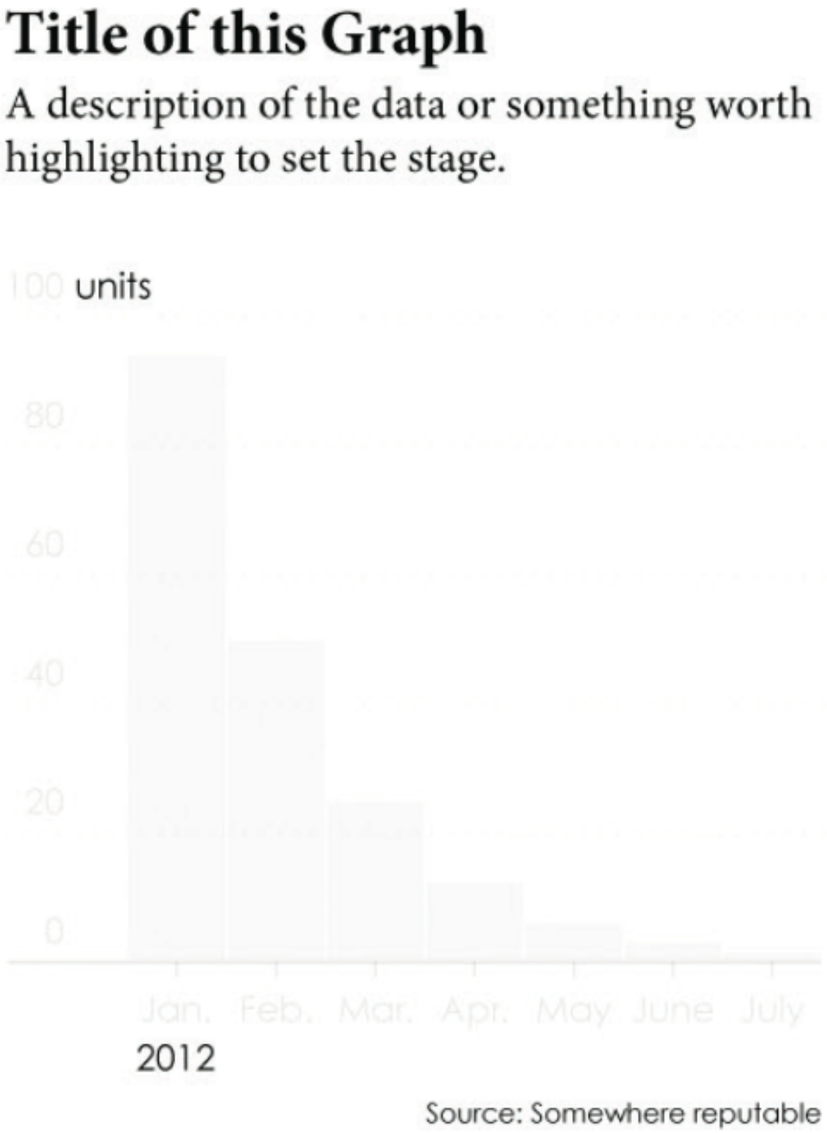


Time

Units of months, days, or hours

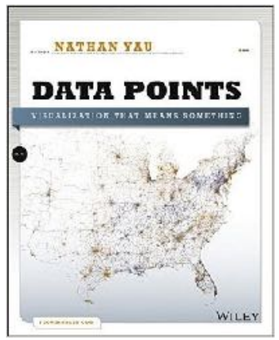
Source: Nathan Yau, Data Points

Context

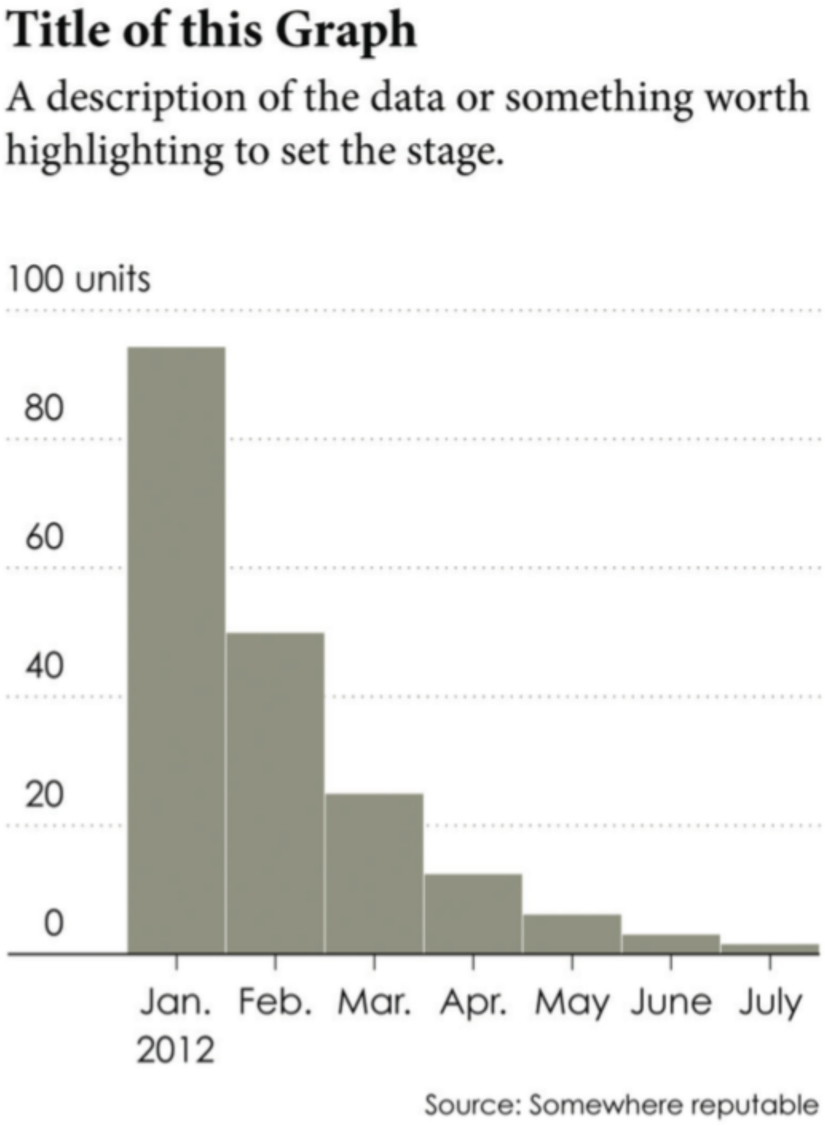


Title, subtitle, caption, clarifications of what the data represents all help the reader understand the visualization.

From Nathan Yau

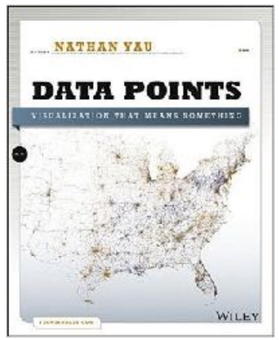


Sum of the parts



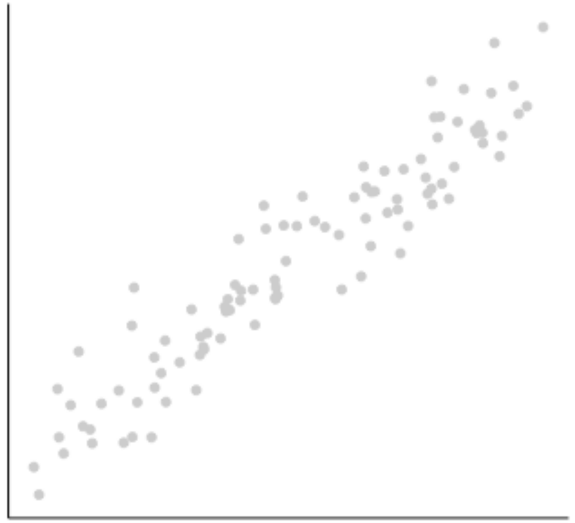
Everything comes together to (hopefully) create a coherent message for the reader.

From Nathan Yau

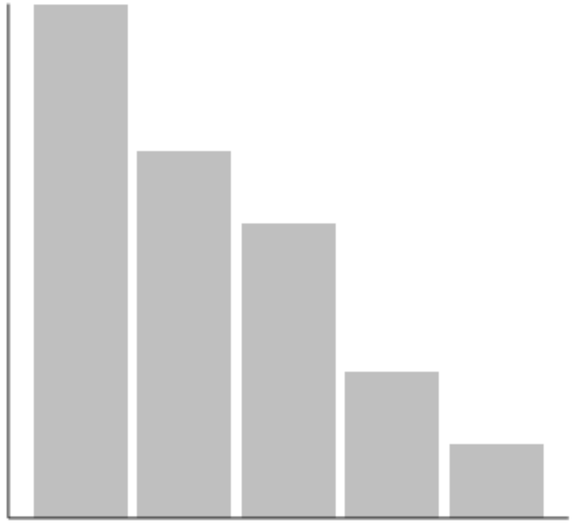


Common Visualizations

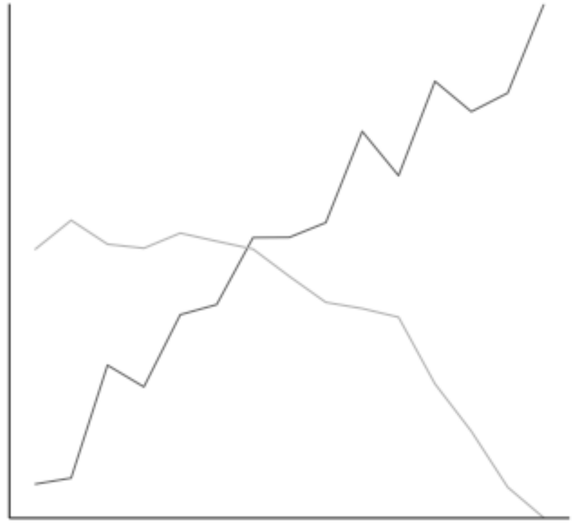
Scatter Plot



Bar plot

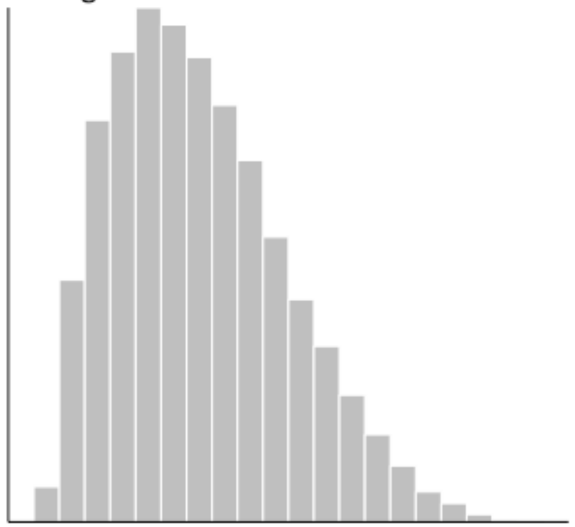


Line Plot

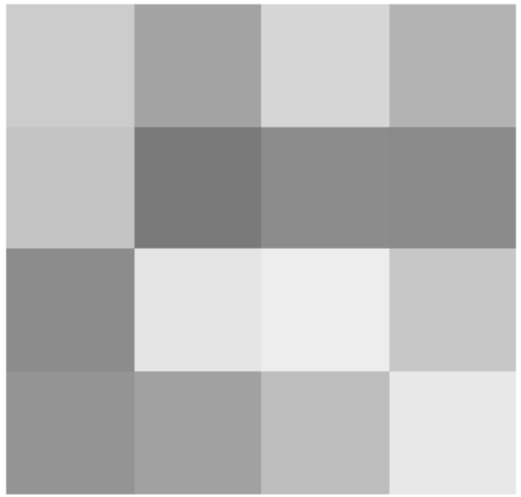


Pie Chart

Histogram



Heat Map



Visualization Building Blocks

Visual Cues

- Position, Length, Size, Shape, Color, Angle, Area

Coordinate System

- Cartesian, Polar, other projection

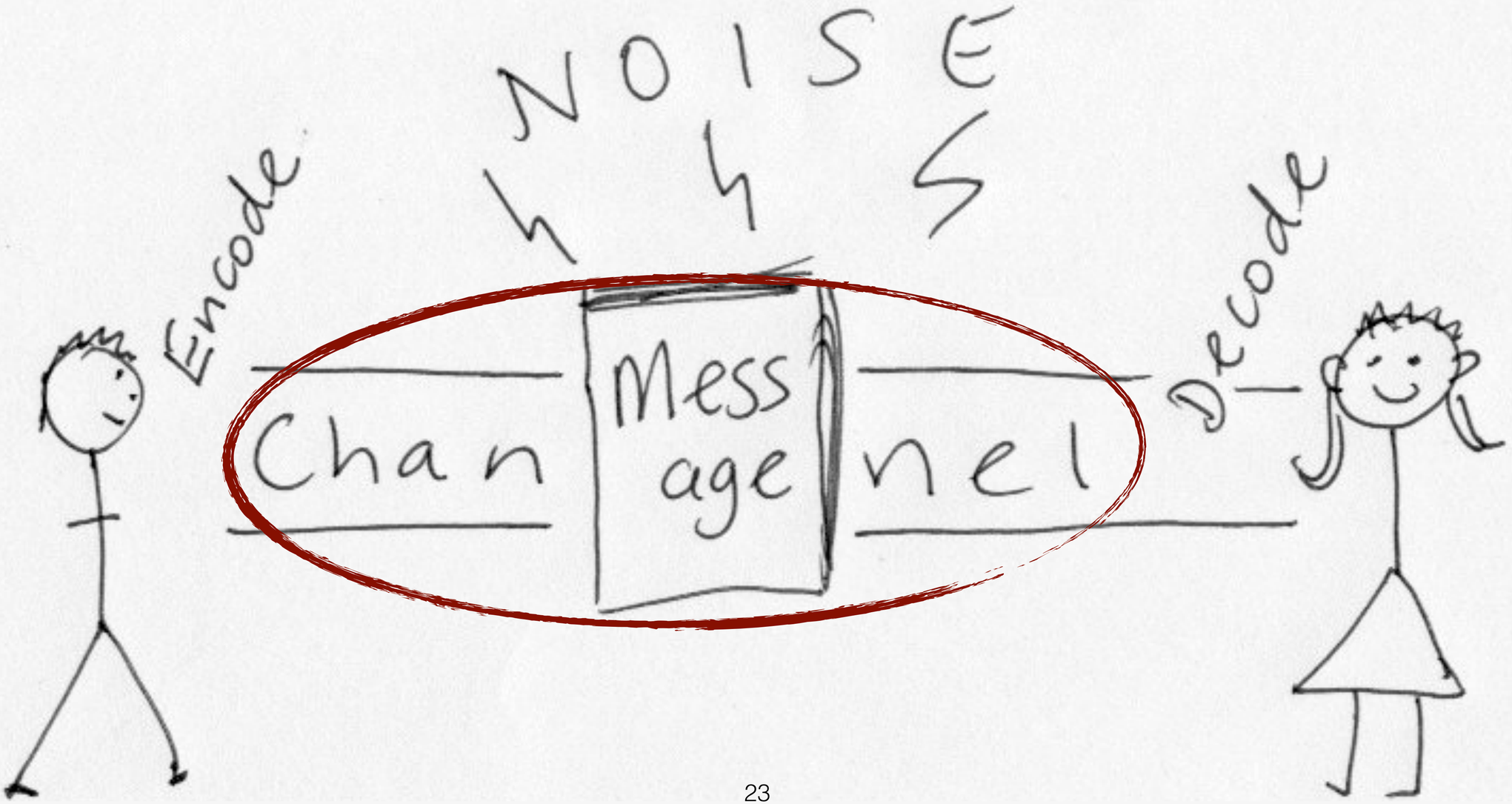
Scale

- Linear, logarithmic, categorical, time

Context

- Title, subtitle, labels, source, annotations

Be aware of the Channel



Printer Friendly?

Obama's Divided Nation

s over
ivided
in 50
riven
gath-
ct its
presi-
viding
f eco-
n re-
evi-
Mr.
close
e has
ween
and
cked

true
nney

proble
verball
that in
revert
that 9
who
Tues
42%
Sand
tie
fact
vot
Mr.
pol

ac
th
h
is
B
i

R.I.
Conn.
Del.
D.C.

Obama Romney

Source: AP

drawn attention to what hap- Obama spokestotum replied:

Color Blindness

Color Blindness

Original

Green-Blind / Deuteranopia

8% of male
0.5% female

http://www.color-blindness.com/coblis-color-blindness-simulator/

Red-Blind / Protanopia

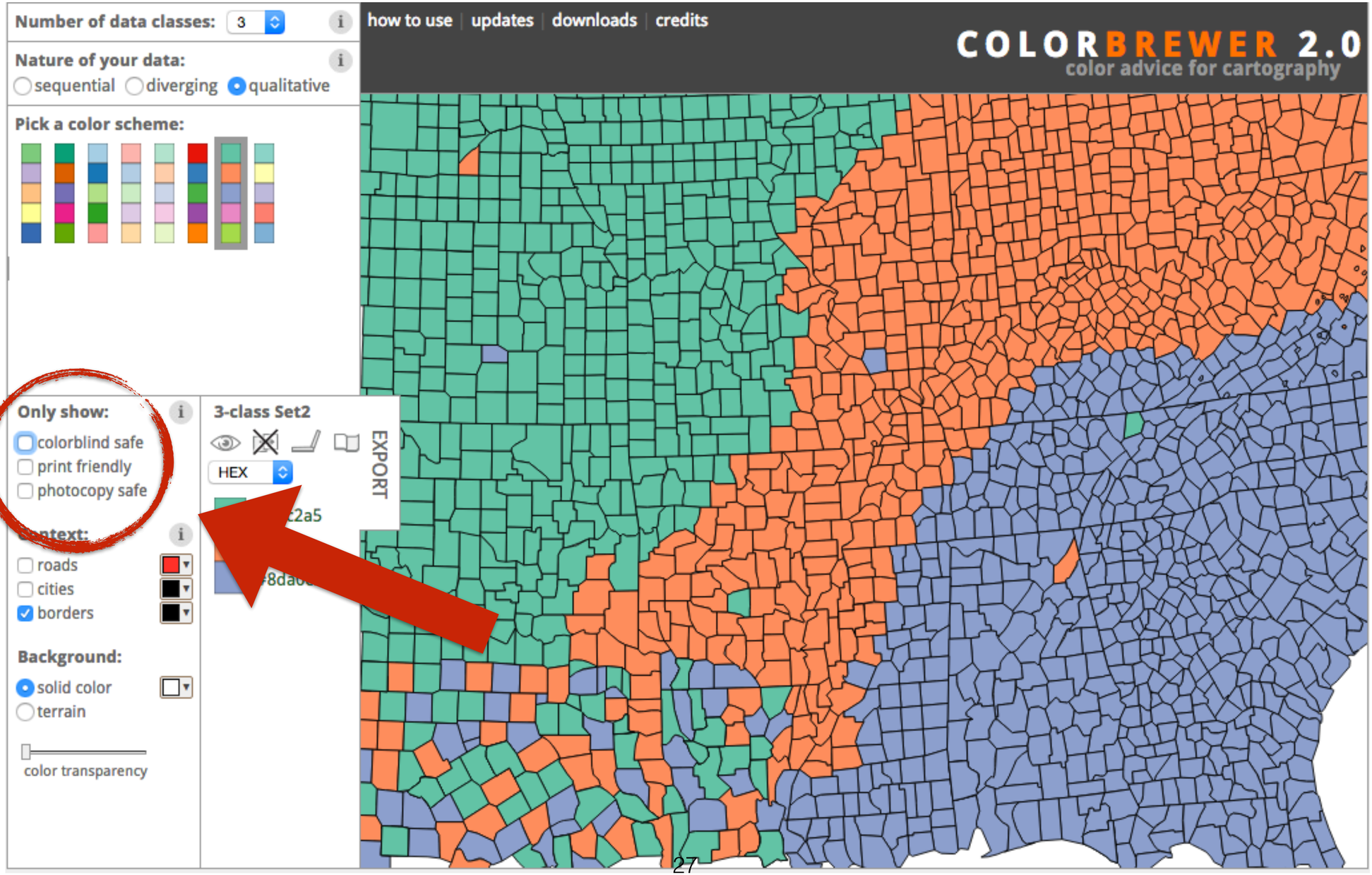
1% of males

Blue-Blind / Tritanopia

rare

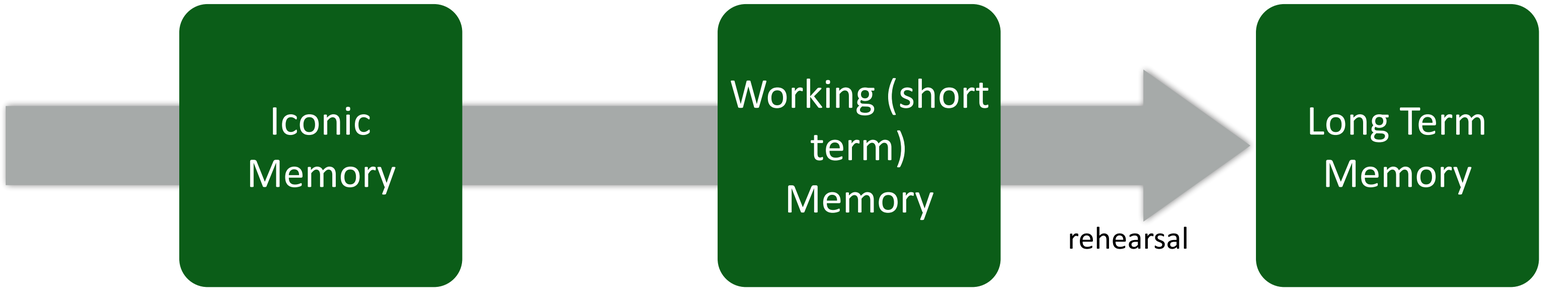
Color Brewer

http://colorbrewer2.org/



The Visual Processing System

Human Visual Processing

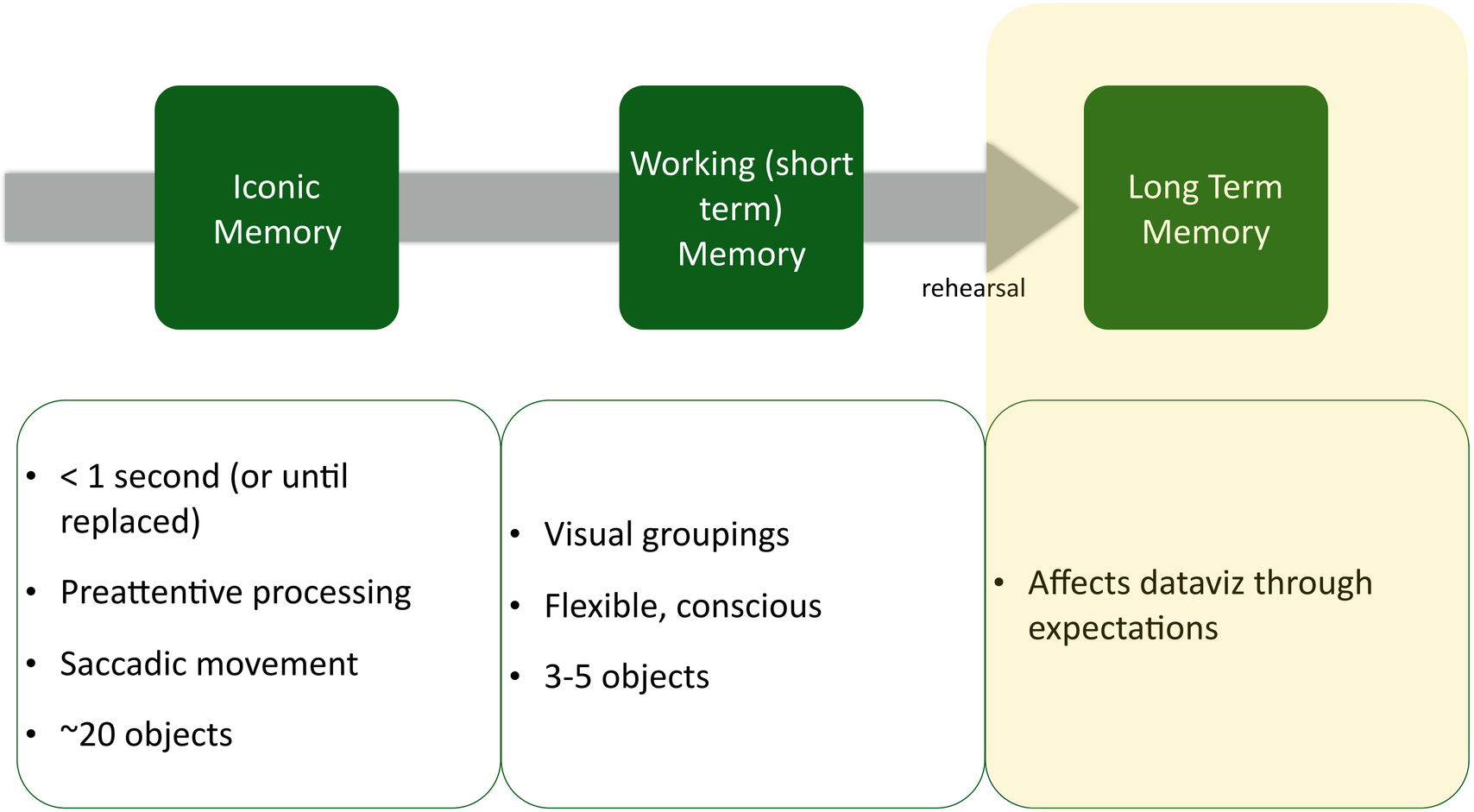


- < 1 second (or until replaced)
- Preattentive processing
- Saccadic movement
- ~20 objects

- Visual groupings
- Flexible, conscious
- 3-5 objects

- Affects dataviz through expectations

Human Visual Processing



The New York Times

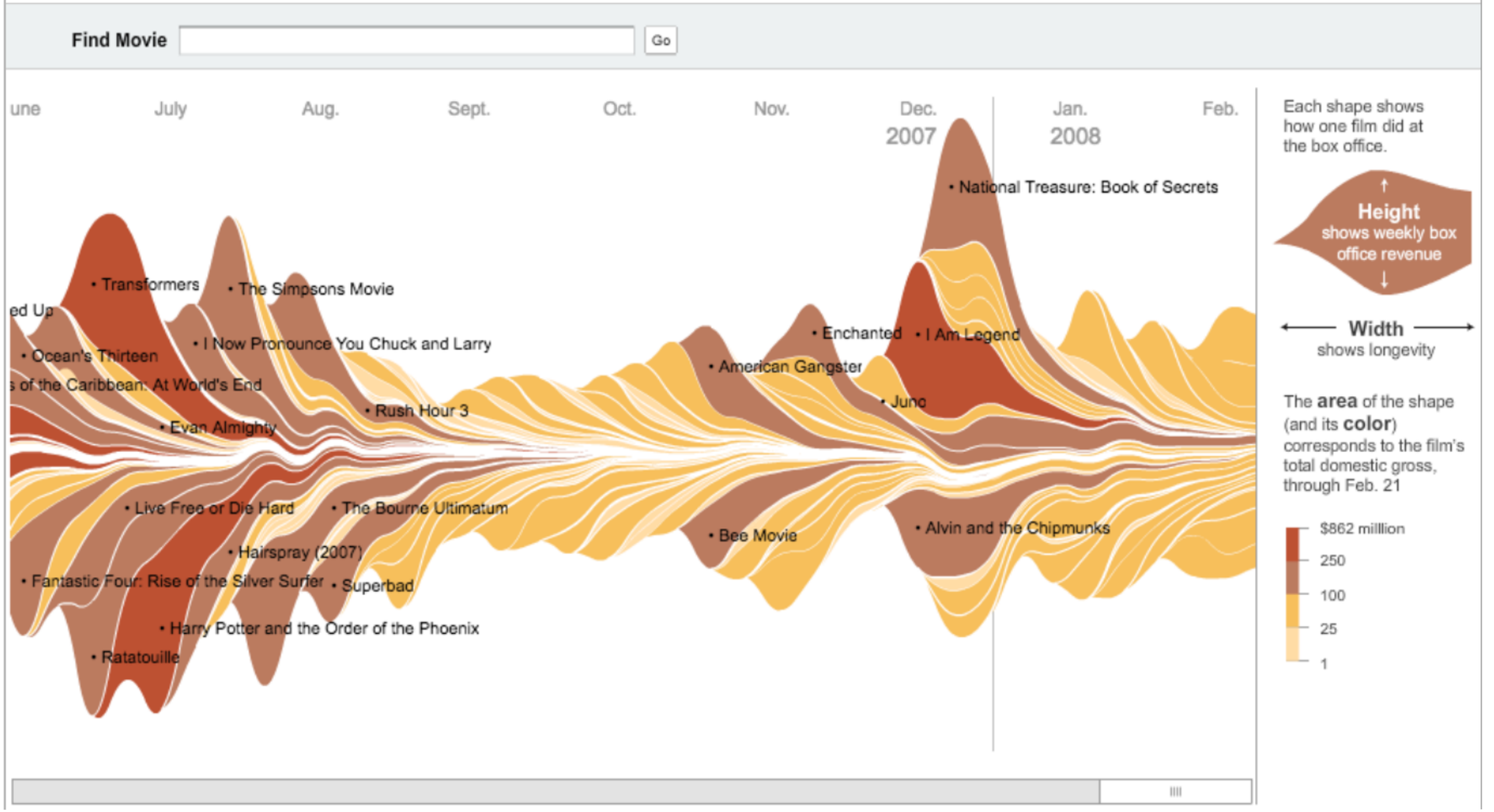
Movies

February 23, 2008

SIGN IN TO E-MAIL OR SAVE THIS | FEEDBACK

The Ebb and Flow of Movies: Box Office Receipts 1986 — 2008

Summer blockbusters and holiday hits make up the bulk of box office revenue each year, while contenders for the Oscars tend to attract smaller audiences that build over time. Here's a look at how movies have fared at the box office, after adjusting for inflation.

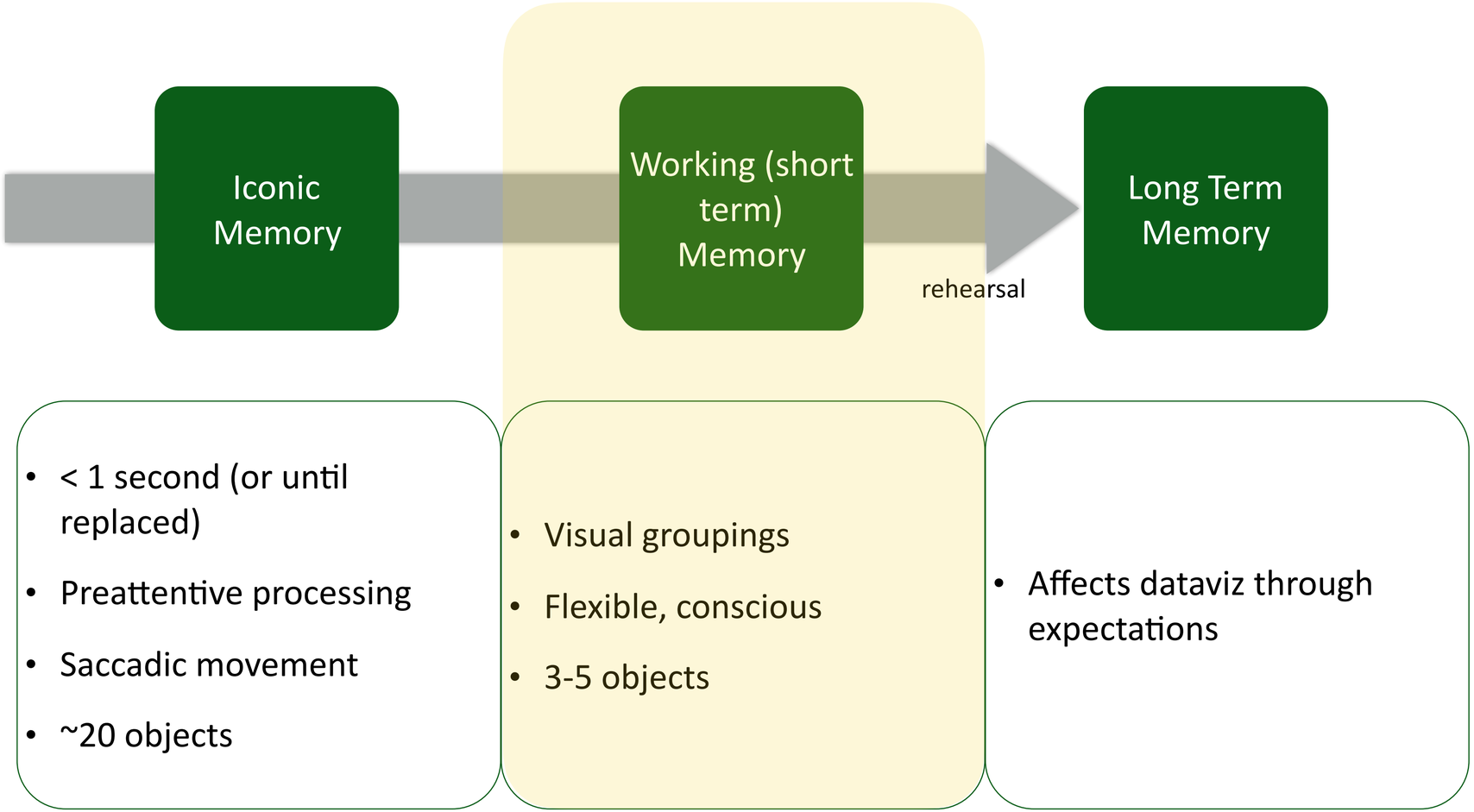


Sources: Baseline StudioSystems; Box Office Mojo

Mathew Bloch, Lee Byron, Shan Carter and Amanda Cox

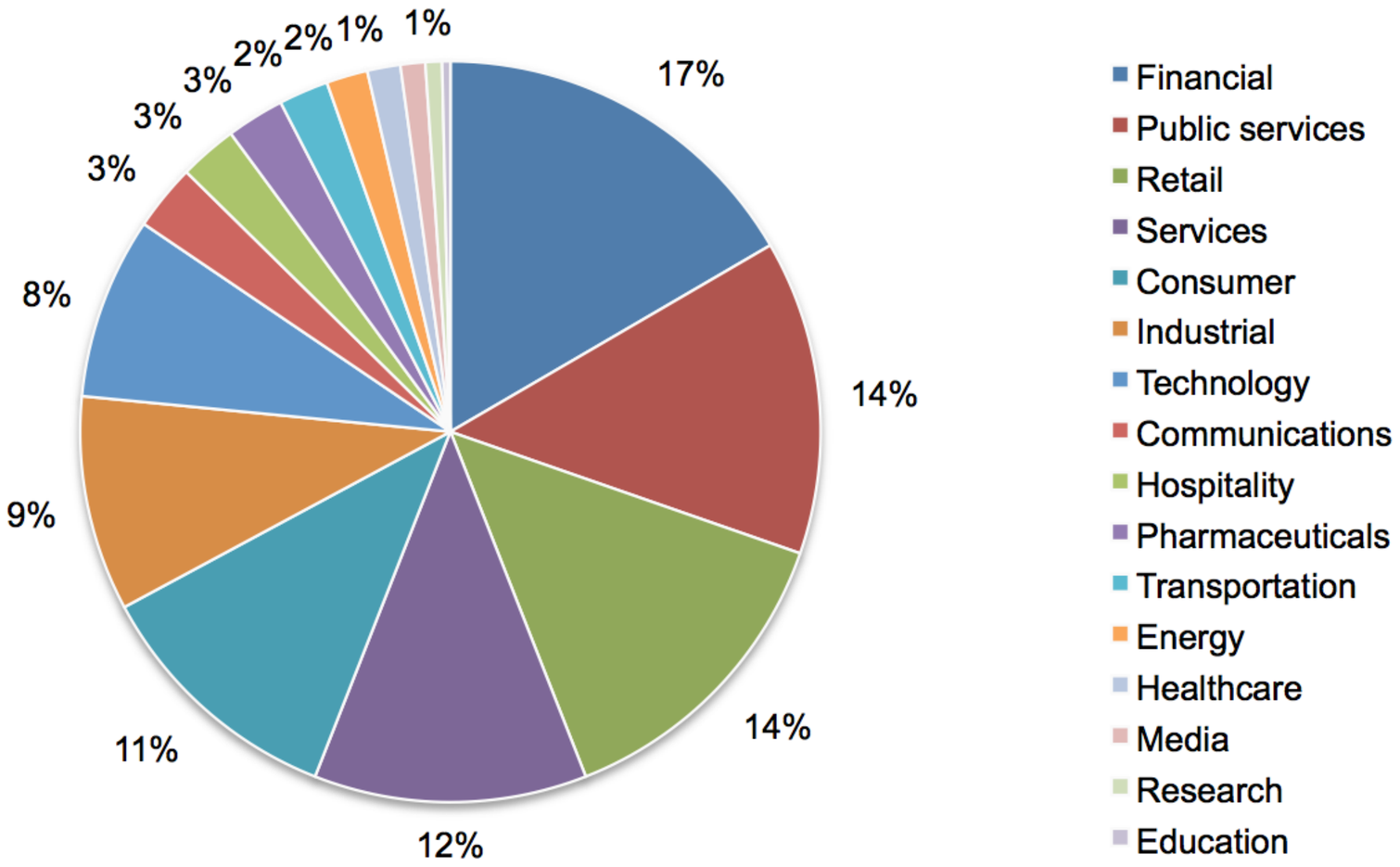
http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html?_r=0

Human Visual Processing



Overworking Working Memory

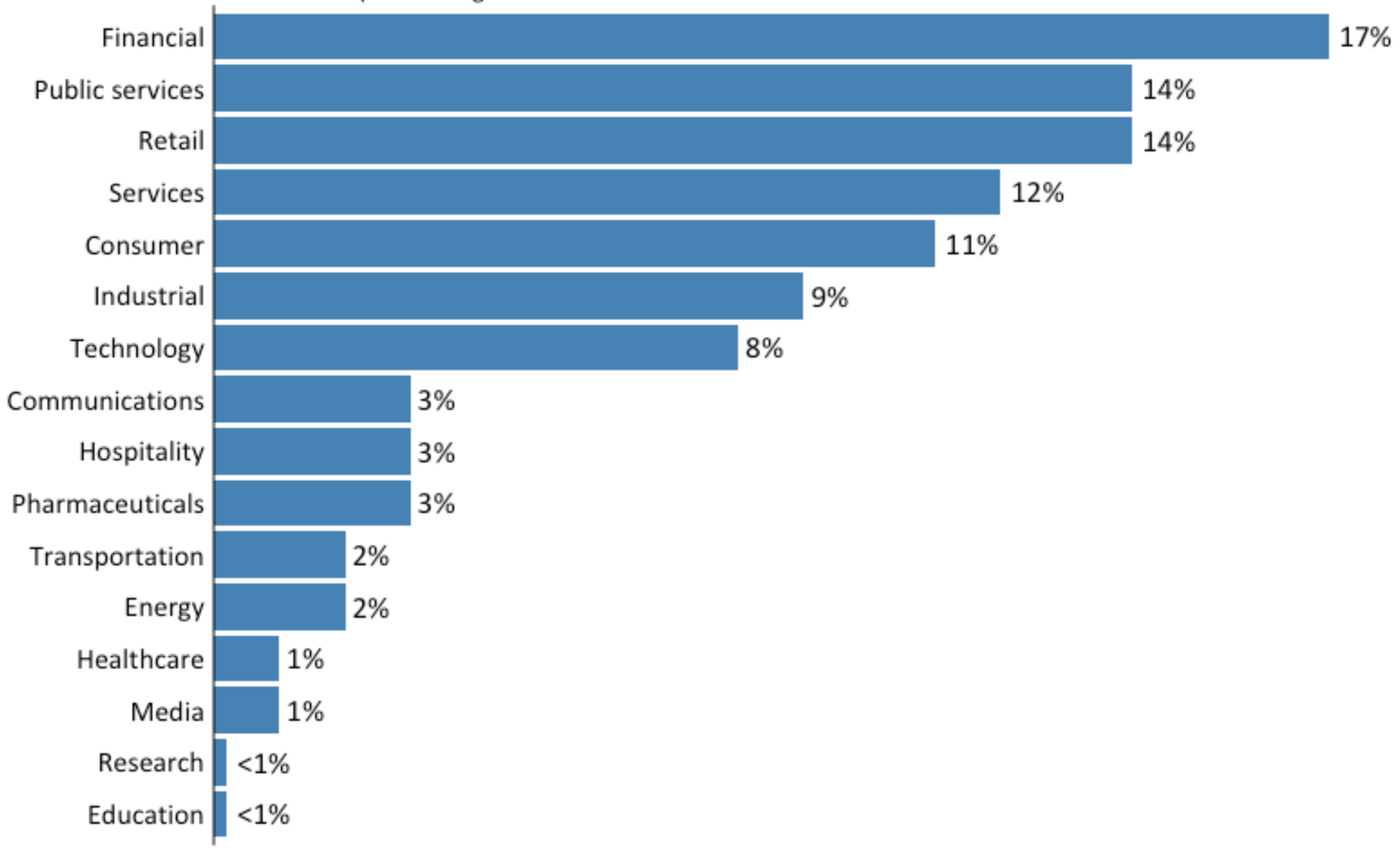
Figure 20. Distribution of the benchmark sample by industry segment
Consolidated (n = 277 organizations)



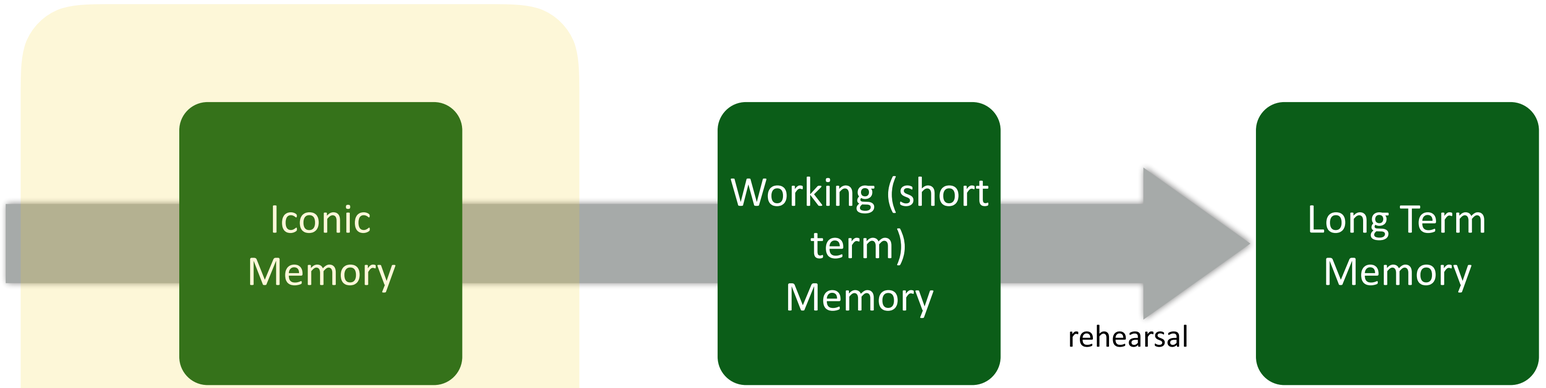
https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf

Working Memory: Direct Labels

Figure 20. Distribution of the benchmark sample by industry segment
Consolidated (n = 277 organizations



Human Visual Processing



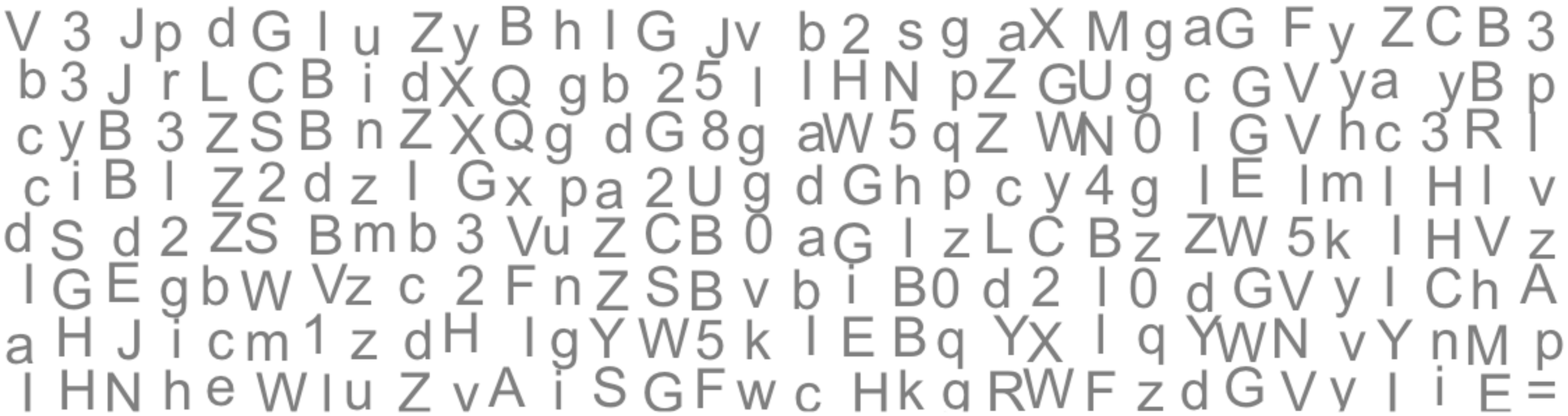
- < 1 second (or until replaced)
- Preattentive processing
- Saccadic movement
- ~20 objects

- Visual groupings
- Flexible, conscious
- 3-5 objects

- Affects dataviz through expectations

Pre-attentive Processing

Count the number of times “X” appears:



Pre-attentive Processing

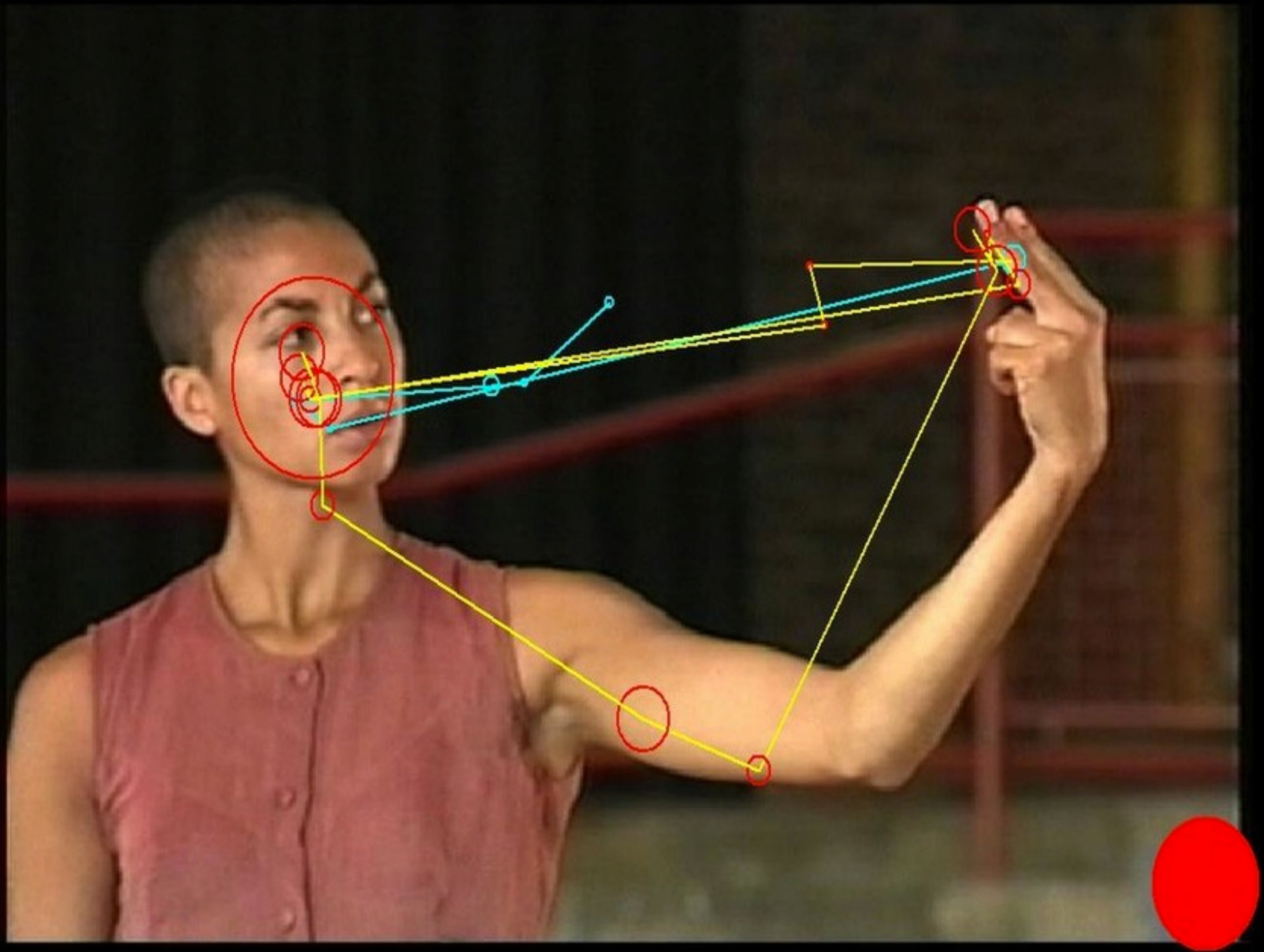
Count the number of times “X” appears:

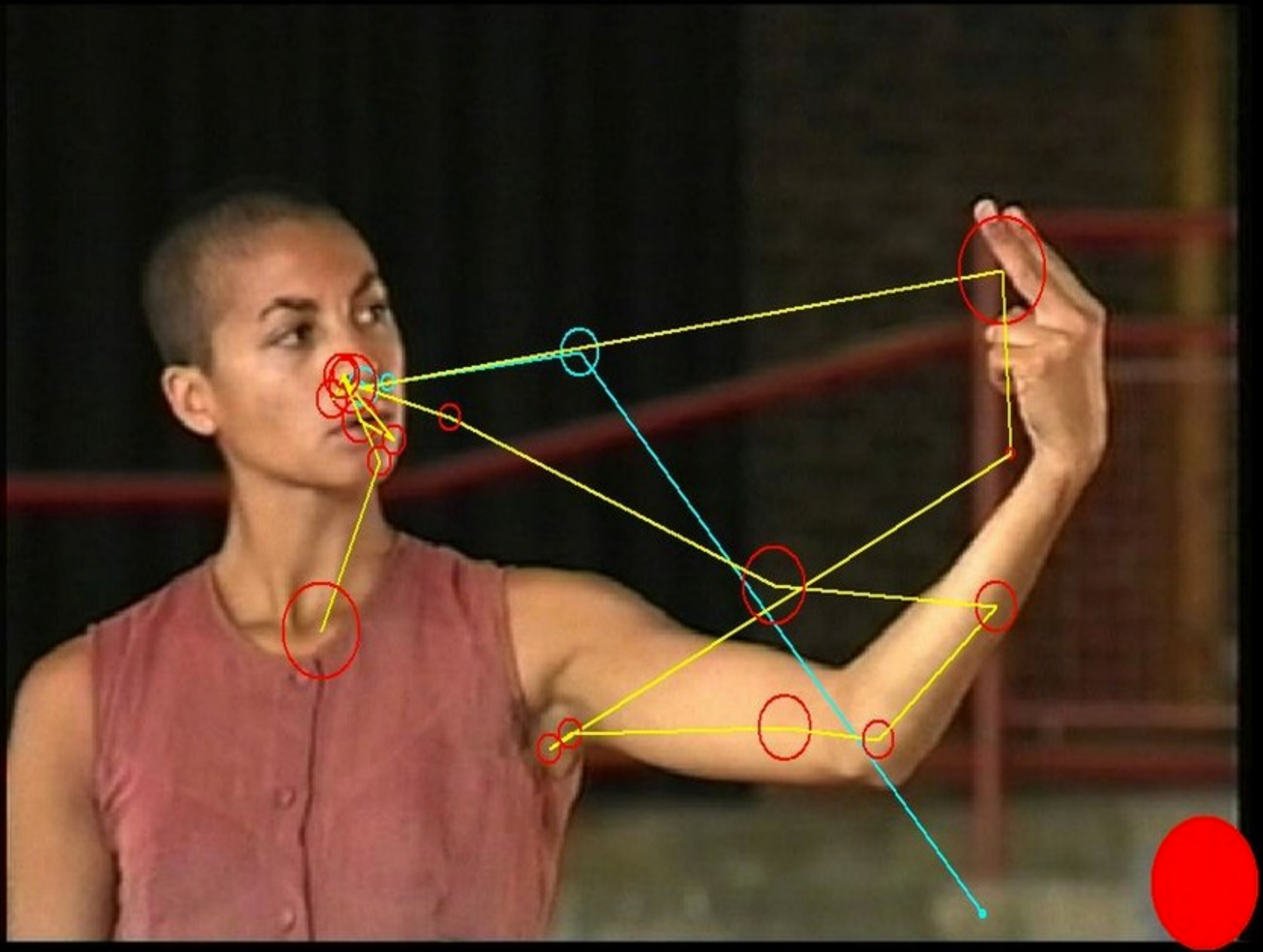
Saccadic Eye Movement

Saccadic Eye Movement

- Pre-attentively Driven
- Ballistic
  - Visual Suppression
  - Cannot Change after start

Saccadic Eye Movement





Mobile Ops Dashboard

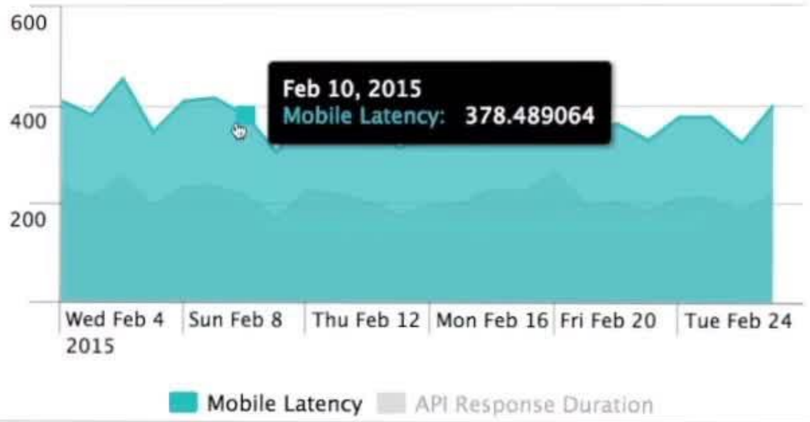
Edit | More Info

End to End Time

368
milliseconds

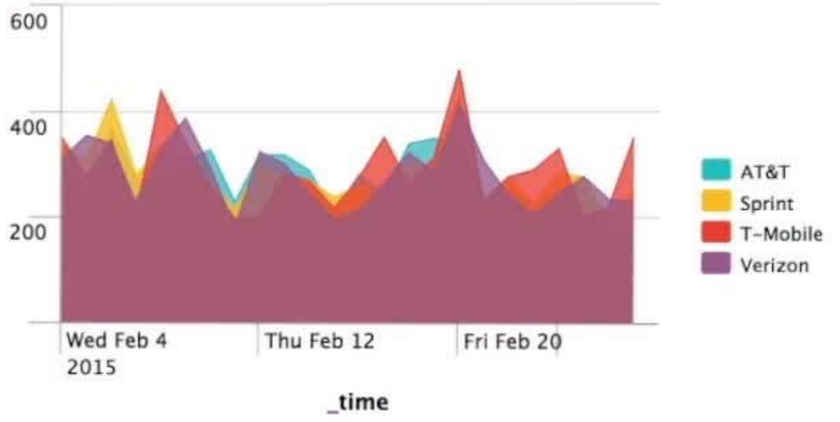
Mobile & Server Latency

12m ago



Mobile Latency by Carrier

12m ago



Mobile App Crashes

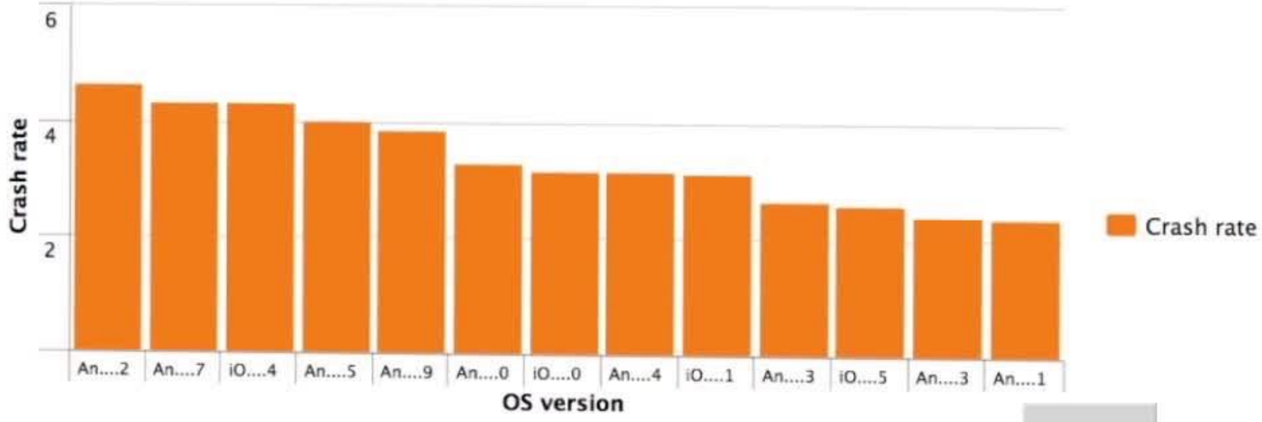
12m ago

	platform	message	where	Trend	Occurrences
1	iOS	SIGSEV	[CrashController Handled1]		110
2	Android	java.lang.NullPointerException	AsyncThread.java:55		95
3	iOS	SIGILL	[CrashController Handled1]		91
4	Android	java.lang.IllegalStateException: Could not execute method of the activity	AsyncThread.java:55		79
5	iOS	SIGILL	[UIKITController Handled5]		73

« prev 1 2 3 next »

Crash Rate by OS Version

12m ago



Revenue at Risk due to Latency

12m ago

Revenue at Risk due to HTTP Errors

12m ago

Mobile Ops Dashboard

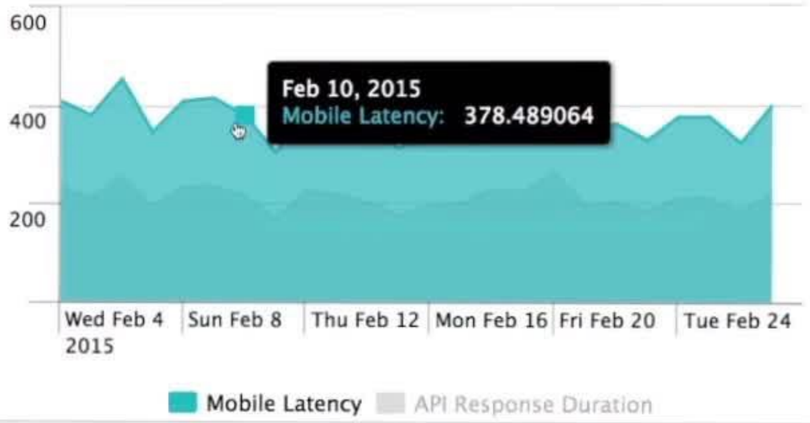
Edit | More Info

End to End Time

368
milliseconds

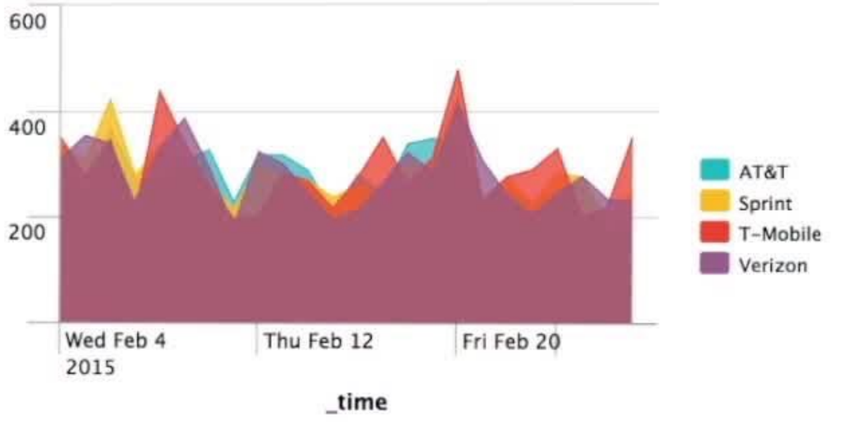
Mobile & Server Latency

12m ago



Mobile Latency by Carrier

12m ago



Mobile App Crashes

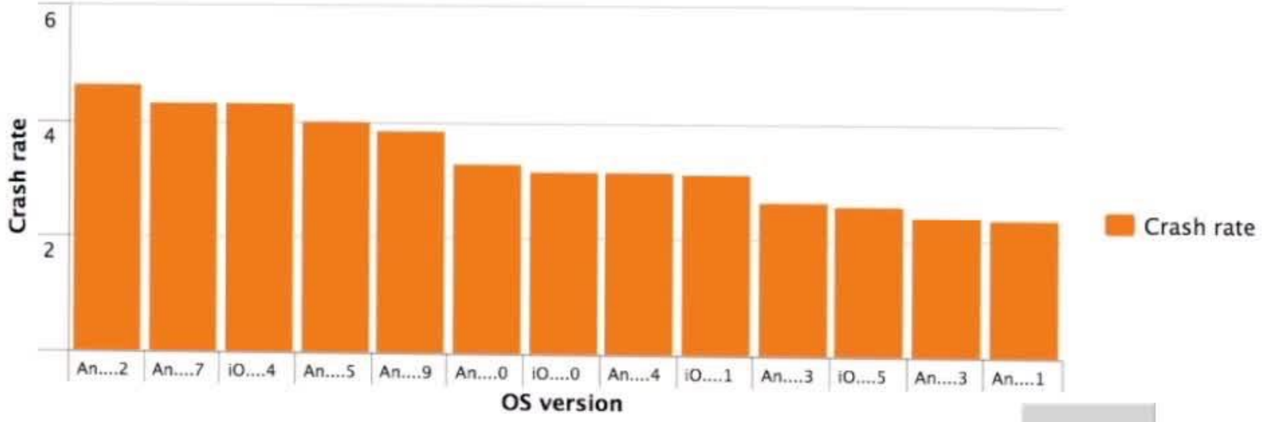
12m ago

	platform	message	where	Trend	Occurrences
1	iOS	SIGSEV	[CrashController Handled1]		110
2	Android	java.lang.NullPointerException	AsyncThread.java:55		95
3	iOS	SIGILL	[CrashController Handled1]		91
4	Android	java.lang.IllegalStateException: Could not execute method of the activity	AsyncThread.java:55		79
5	iOS	SIGILL	[UIKITController Handled5]		73

« prev 1 2 3 next »

Crash Rate by OS Version

12m ago



Revenue at Risk due to Latency

12m ago

Revenue at Risk due to HTTP Errors

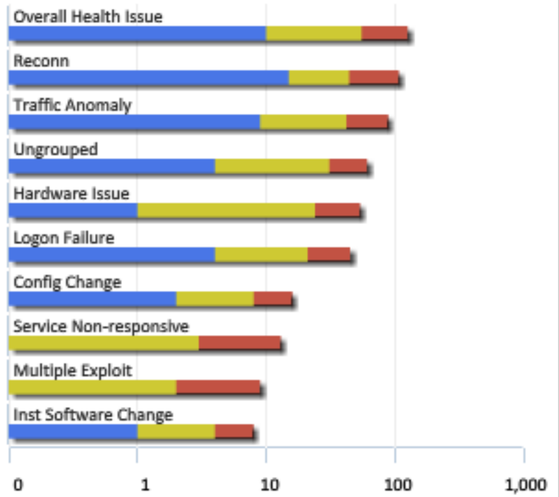
12m ago


accelops
Dashboard
Analytics
Incidents
CMDB
Admin
Benjamin ...
Exchange
Auth Servi...
DHCP/DN...
Dashboard > Device Dashboard > Event Status
Summary Dashboards
Biz Service Summary
Device Summary
All Devices
Servers
Network Devices
My Devices
VMWare Summary
Hardware Summary
Storage Summary
Application Summary
Incident Dashboard
Availability Dashboard
Performance Dashboard
BizSvc
Device
Security Dashboard
BizSvc
Device
Biz Svc Dashboard
Device Dashboard
Network
Server
Application
Event Status
Storage
My Dashboard
Ungrouped
Incidents: Top Incidents Categories Ranked By Count (last 30 min...
Overall Health Issue
Reconn
Traffic Anomaly
Ungrouped
Hardware Issue
Logon Failure
Config Change
Service Non-responsive
Multiple Exploit
Inst Software Change
0
1
10
100
1,000
Total Event Rate (Per Sec) (last 30 mins,4,8 hours updated@19:33)
0
100
200
300
400
500
600
Events: Top Dest IP By Count (last 8 hours updated@19:33)
172.16.10.36
312,206
172.16.10.16
299,195
172.16.10.15
296,025
172.16.10.18
246,624
Top Ports, Events By Severity, Count (last 1 hour updated@19:43)
10,000
1,000
100
10
1
0
18:45
18:55
19:05
19:15
19:25
19:35
TCP(6),HTTPS(443),WEB-MISC PCT Client_Hello ove
UDP(17),DOMAIN(53),Denied IP traffic by policy,Seve
UDP(17),59866,Windows Filtering blocked a connectio
UDP(17),57968,Windows Filtering blocked a connectio
UDP(17),49359,Windows Filtering blocked a connectio
Settings
Title
Top Rpt IP By Event Rate
Description
Condition
Please see Report Template section for details.
Display
Combo View
Time
Last 1 hour
# of results
10
OK
Cancel
Events: Top Network Connections, Events By Severity, Count (las...
2,090
28
49
172.16.20.12
192.168.24.1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
224.0.0.252,1
Top Events By Severity, Count (last 1 hour updated@19:43)
Event Name
COUNT
Trend
A user account was locked 2
Scanning detected and suc 5
Host Vulnerability score ex 1
Memory hardware health: 20
Power supply hardware he 19
WEB-MISC PCT Client_Hell 2
Windows Filtering blocked 10,674
Denied IP traffic by policy 2,105
Hardware health status 19
Power supply hardware he 19
Incidents: Top Incidents Ranked By Severity, Count (last 1 hour u...
Event Name
COUNT
Trend
Account Lockout: Server 2
Important application stop 1
User_Added_to_Domain_ 1
Biz Srvc Performance Heal 3
Difference in Running and 2
Available reports for dashboard
Copyright © 2008-2010 AccelOps, Inc. All rights reserved.
Powered by AccelOps

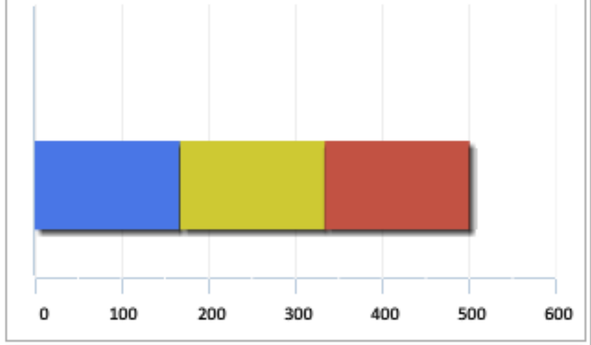
Dashboard > Device Dashboard > Event Status

- Summary Dashboards
  - Biz Service Summary
  - Device Summary
    - All Devices
    - Servers
    - Network Devices
    - My Devices
  - VMWare Summary
  - Hardware Summary
  - Storage Summary
  - Application Summary
- Incident Dashboard
- Availability Dashboard
- Performance Dashboard
  - BizSvc
  - Device
- Security Dashboard
  - BizSvc
  - Device
- Biz Svc Dashboard
- Device Dashboard
  - Network
  - Server
  - Application
  - Event Status
  - Storage
- My Dashboard
- Ungrouped

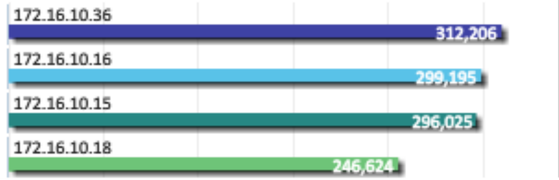
Incidents: Top Incidents Categories Ranked By Count (last 30 min...



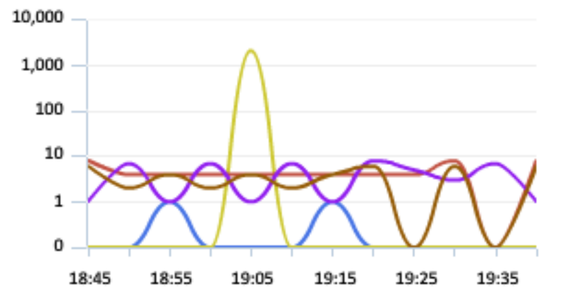
Total Event Rate (Per Sec) (last 30 mins,4,8 hours updated@19:33)



Events: Top Dest IP By Count (last 8 hours updated@19:33)



Top Ports, Events By Severity, Count (last 1 hour updated@19:43)



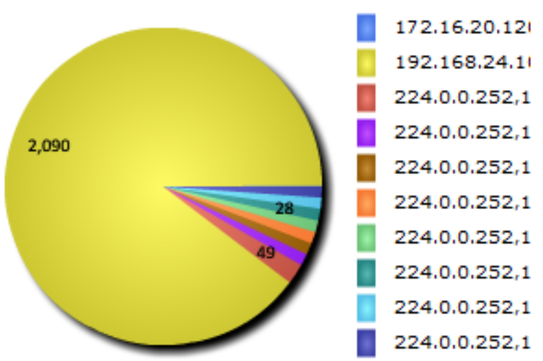
TCP(6),HTTPS(443),WEB-MISC PCT Client_Hello ove
UDP(17),DOMAIN(53),Denied IP traffic by policy,Seve
UDP(17),59866,Windows Filtering blocked a connectio
UDP(17),57968,Windows Filtering blocked a connectio
UDP(17),49359,Windows Filtering blocked a connectio

Settings

Title	Top Rpt IP By Event Rate
Description	
Condition	Please see Report Template section for details.
Display	Combo View
Time	Last 1 hour
# of results	10

OK | Cancel

Events: Top Network Connections, Events By Severity, Count (las...



Top Events By Severity, Count (last 1 hour updated@19:43)

Event Name	COUNT	Trend
A user account was locked	2	
Scanning detected and suc	5	
Host Vulnerability score ex	1	
Memory hardware health:	20	
Power supply hardware he	19	
WEB-MISC PCT Client_Hell	2	
Windows Filtering blocked	10,674	
Denied IP traffic by policy	2,105	
Hardware health status	19	
Power supply hardware he	19	

Incidents: Top Incidents Ranked By Severity, Count (last 1 hour u...

Event Name	COUNT	Trend
Account Lockout: Server	2	
Important application stop	1	
User_Added_to_Domain_	1	
Biz Srvc Performance Heal	3	
Difference in Running and	2	

Available reports for dashboard

Copyright © 2008-2010 AccelOps, Inc. All rights reserved.

RSA Security Operations Management

Preferences | Reports | Help | Logout | Search: Incident Response

Security Operations Management | Enterprise Management

Navigation Menu

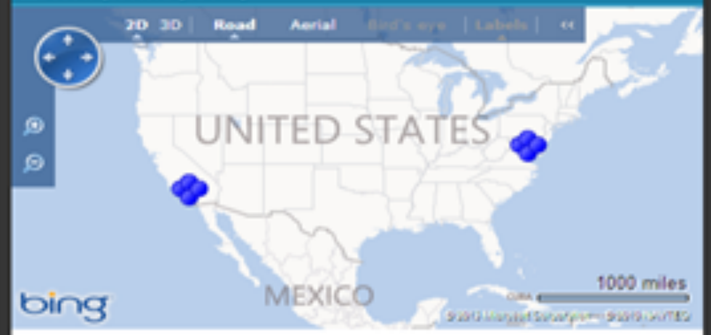
- **Incident Response**
  - Security Alerts
  - Security Incidents
  - Incident Investigations
  - Forensic Analysis
  - Incident Response Tasks
  - Incident Journal
- **Data Breach Response**
  - Data Breaches
  - Breach Tasks
  - Breach Risk Assessment
  - Notifications & Call Trees
  - Notification History
- **SOC Program Management**
  - Shift Handover
  - SOC Procedure Library
  - Security Controls
  - SOC Policies
  - Contacts
  - Teams
  - Team Members
  - Competencies
  - Training & CPE Tracking
- **Issue Management**
  - Findings
  - Remediation Plans
  - Exception Requests
  - Policy Change Requests

Dashboard: SOC Manager

Welcome, Jim SOC Manager

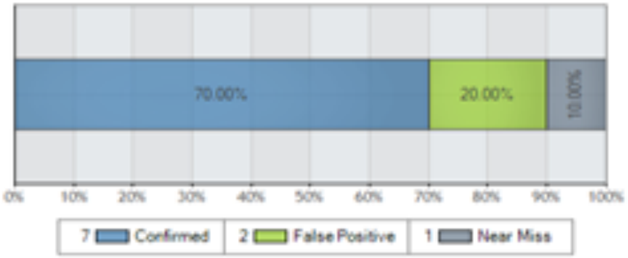
Security Alerts by External Destination

Security Alerts by Facility



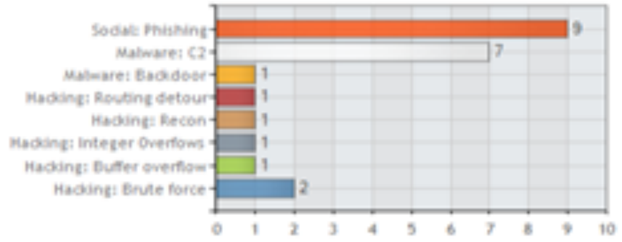
Incident Metrics

Incidents by Confirmation



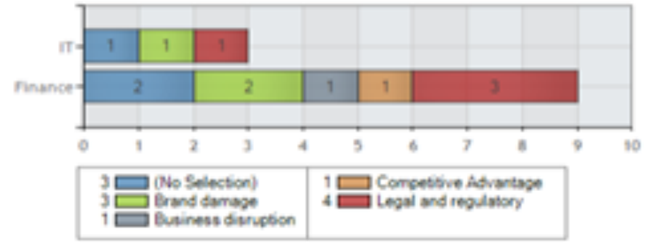
Threat Metrics

Incidents by Threat Categories



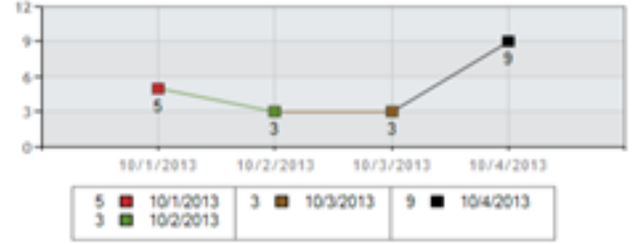
Impact Metrics

Incidents by Business Impact

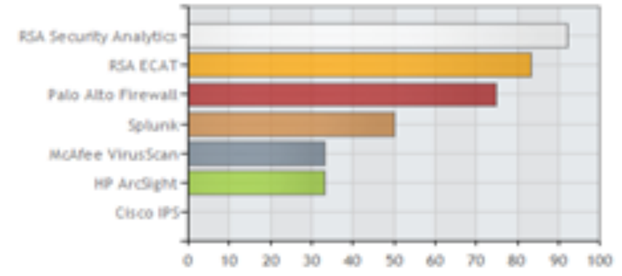


Trend for All Queues

Incident Trends - Daily



Security Tools Efficacy



Data Breaches

ID	Date of Breach	Breach Name	Breach Type	Breach Status
DB-205533	10/3/2013 10:30 AM	PII Disclosed	Unauthorized Disclosure Hacking or Malware	Active

Page 1 of 1 (1 records)

Declared Incidents

ID	Incident Summary	Priority	Threat Category	Threat Valid	Incident Status
INC-6	Beaconing activity grouped by destination IP:202.111.175.31	P-1	Malware C2	Suspected	Escalated

Page 1 of 1 (1 records)

RSA Security Operations Management

Preferences | Reports | Help | Logout | Search: Incident Response

Security Operations Management | Enterprise Management

Navigation Menu

Incident Response
- Security Alerts
- Security Incidents
- Incident Investigations
- Forensic Analysis
- Incident Response Tasks
- Incident Journal

Data Breach Response
- Data Breaches
- Breach Tasks
- Breach Risk Assessment
- Notifications & Call Trees
- Notification History

SOC Program Management
- Shift Handover
- SOC Procedure Library
- Security Controls
- SOC Policies
- Contacts
- Teams
- Team Members
- Competencies
- Training & CPE Tracking

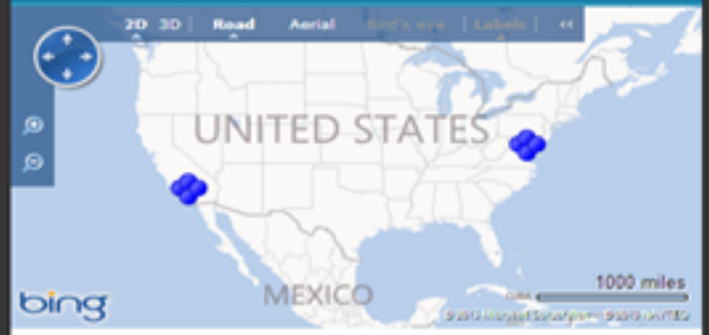
Issue Management
- Findings
- Remediation Plans
- Exception Requests
- Policy Change Requests

Dashboard: SOC Manager

Welcome, Jim SOC Manager

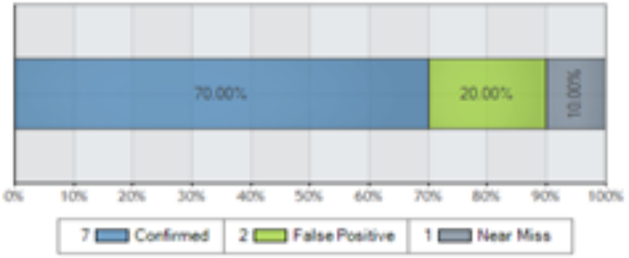
Security Alerts by External Destination

Security Alerts by Facility



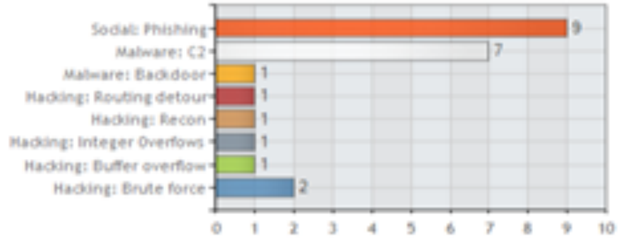
Incident Metrics

Incidents by Confirmation



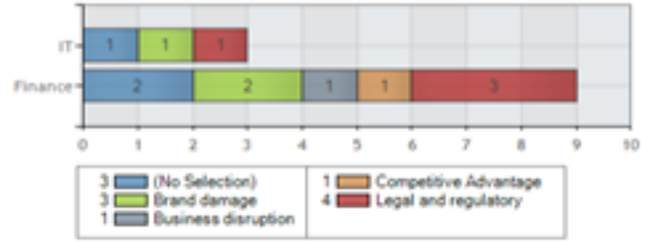
Threat Metrics

Incidents by Threat Categories



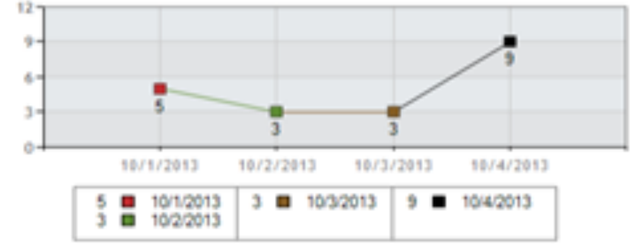
Impact Metrics

Incidents by Business Impact

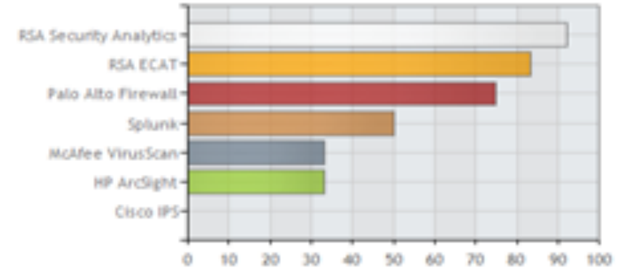


Trend for All Queues

Incident Trends - Daily



Security Tools Efficacy



Data Breaches

ID	Date of Breach	Breach Name	Breach Type	Breach Status
DB-205533	10/3/2013 10:30 AM	PII Disclosed	Unauthorized Disclosure Hacking or Malware	Active

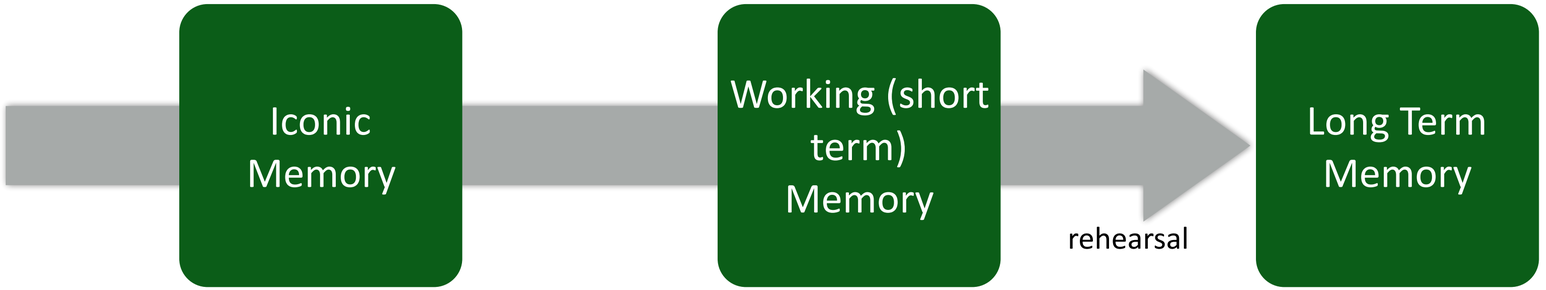
Page 1 of 1 (1 records)

Declared Incidents

ID	Incident Summary	Priority	Threat Category	Threat Valid	Incident Status
INC-6	Beaconing activity grouped by destination IP:202.111.175.31	P-1	Malware C2	Suspected	Escalated

Page 1 of 1 (1 records)

Human Visual Processing



- < 1 second (or until replaced)
- Preattentive processing
- Saccadic movement
- ~20 objects

- Visual groupings
- Flexible, conscious
- 3-5 objects

- Affects dataviz through expectations

Gestalt Laws

Law of Similarity:

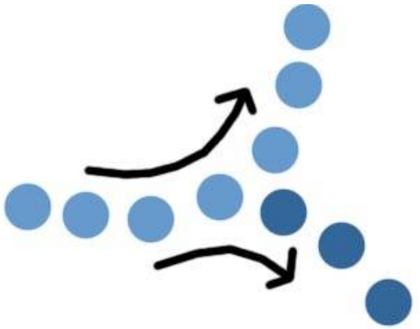
Items that are similar tend to be grouped together.

In the image above, most people see vertical columns of circles and squares.

Law of Pragnanz:

Reality is organized or reduced to the simplest form possible.

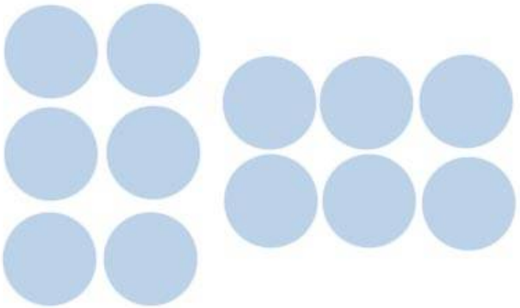
For example, we see the image above as a series of circles rather than as many much more complicated shapes.



Law of Continuity:

Lines are seen as following the smoothest path.

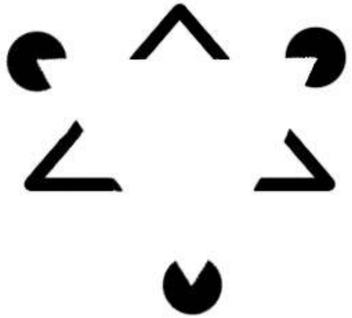
In the image above, the top branch is seen as continuing the first segment of the line. This allows us to see things as flowing smoothly without breaking lines up into multiple parts.



Law of Proximity:

Objects near each other tend to be grouped together.

The circles on the left appear to be grouped in vertical columns, while those on the right appear to be grouped in horizontal rows.



Law of Closure:

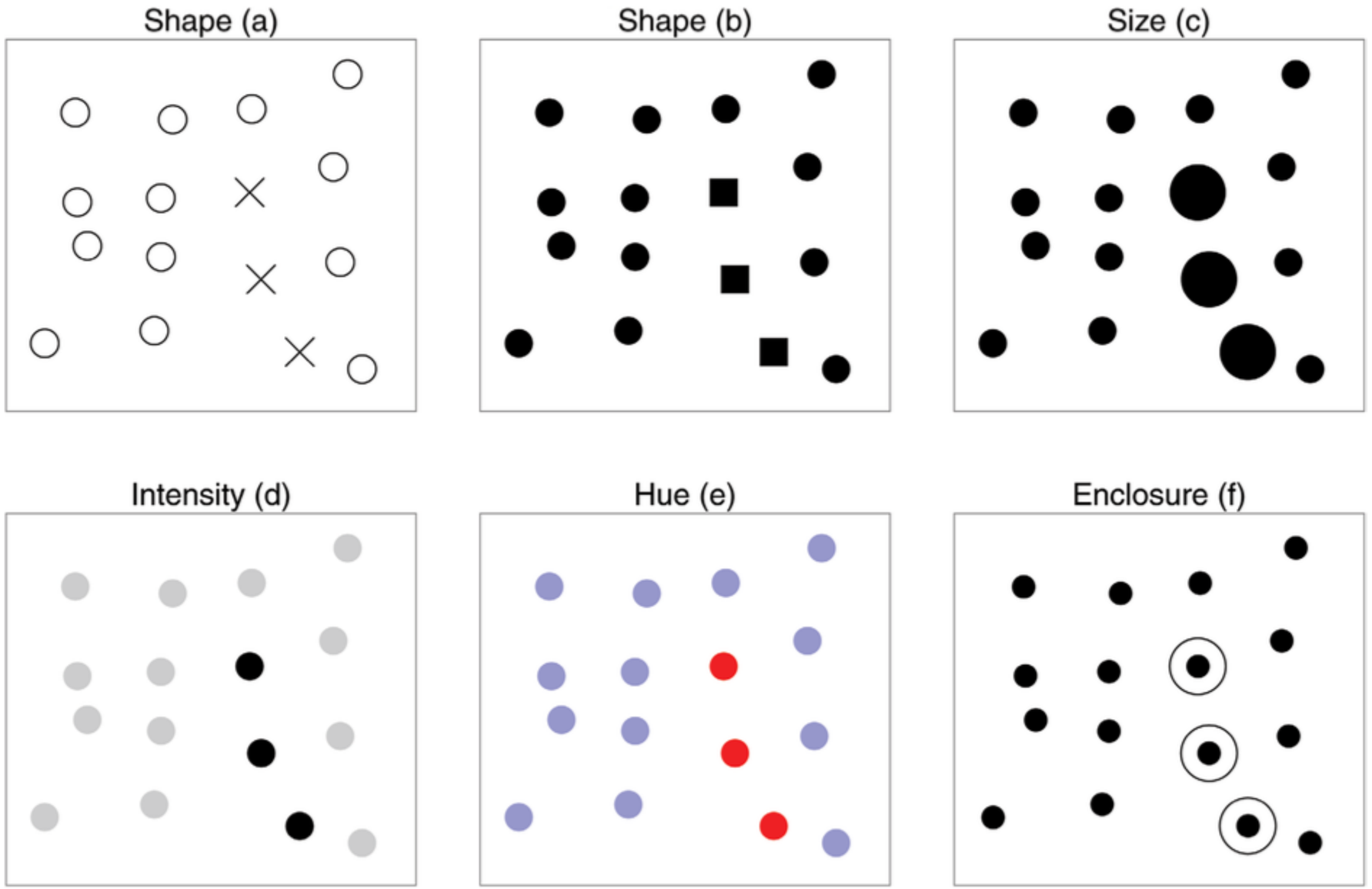
Objects grouped together are seen as a whole.

We tend to ignore gaps and complete contour lines. In the image above, there are no triangles or circles, but our minds fill in the missing information to create familiar shapes and images.

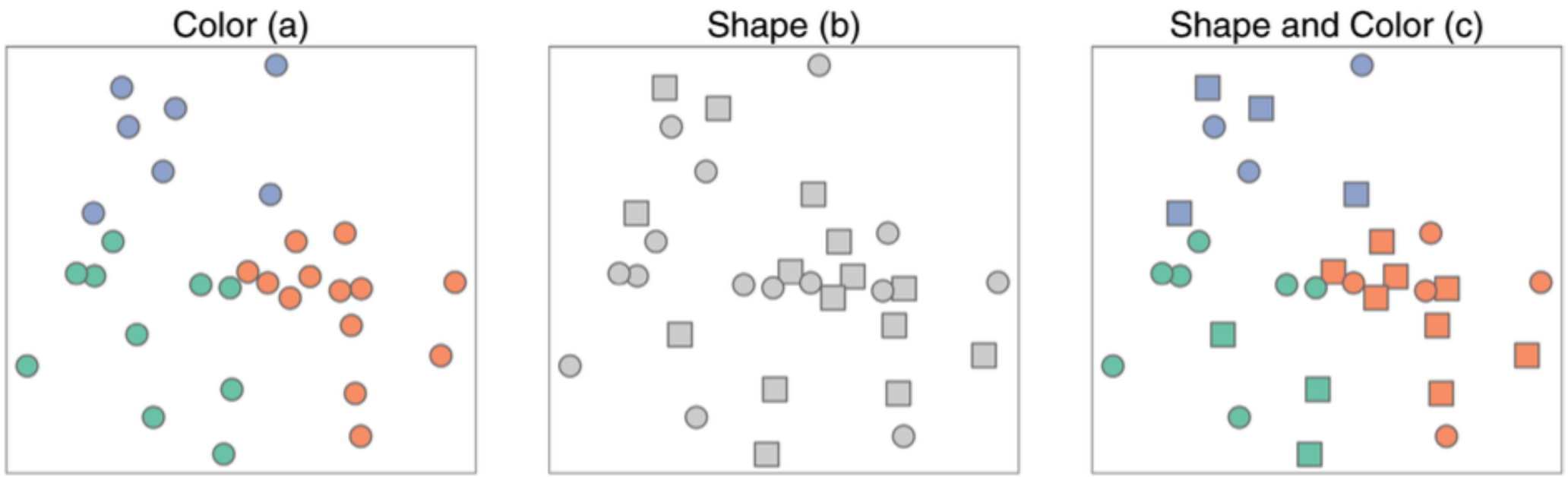
http://psychology.about.com/od/sensationandperception/ss/gestaltlaws.htm

Gestalt Logos

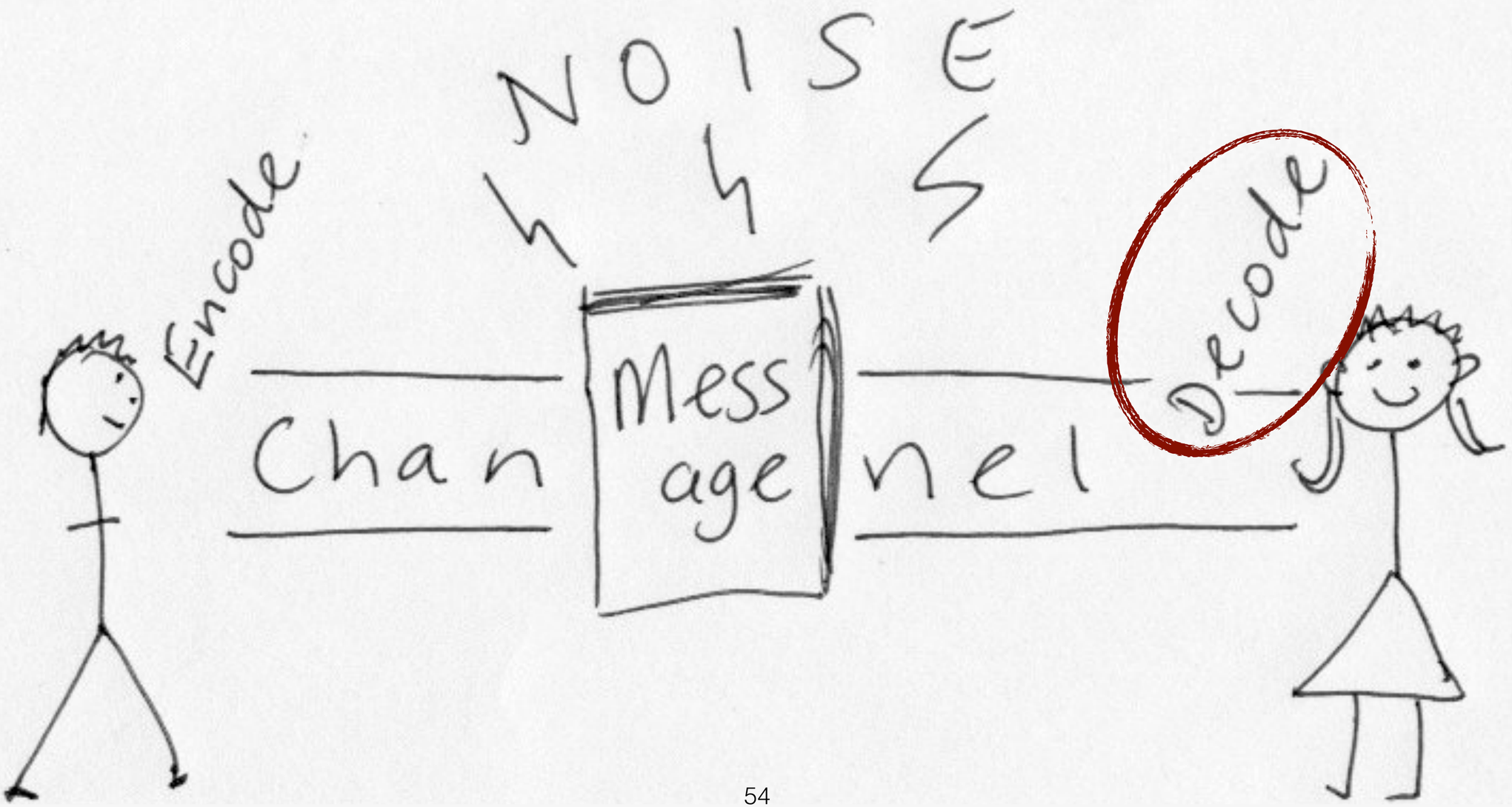
Groupings



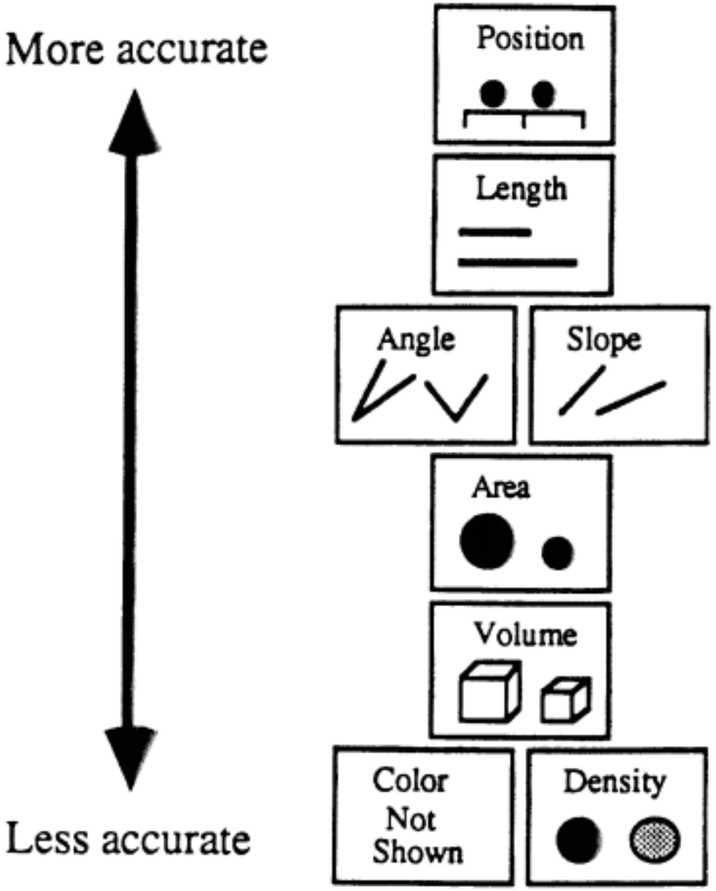
Too many encodings…



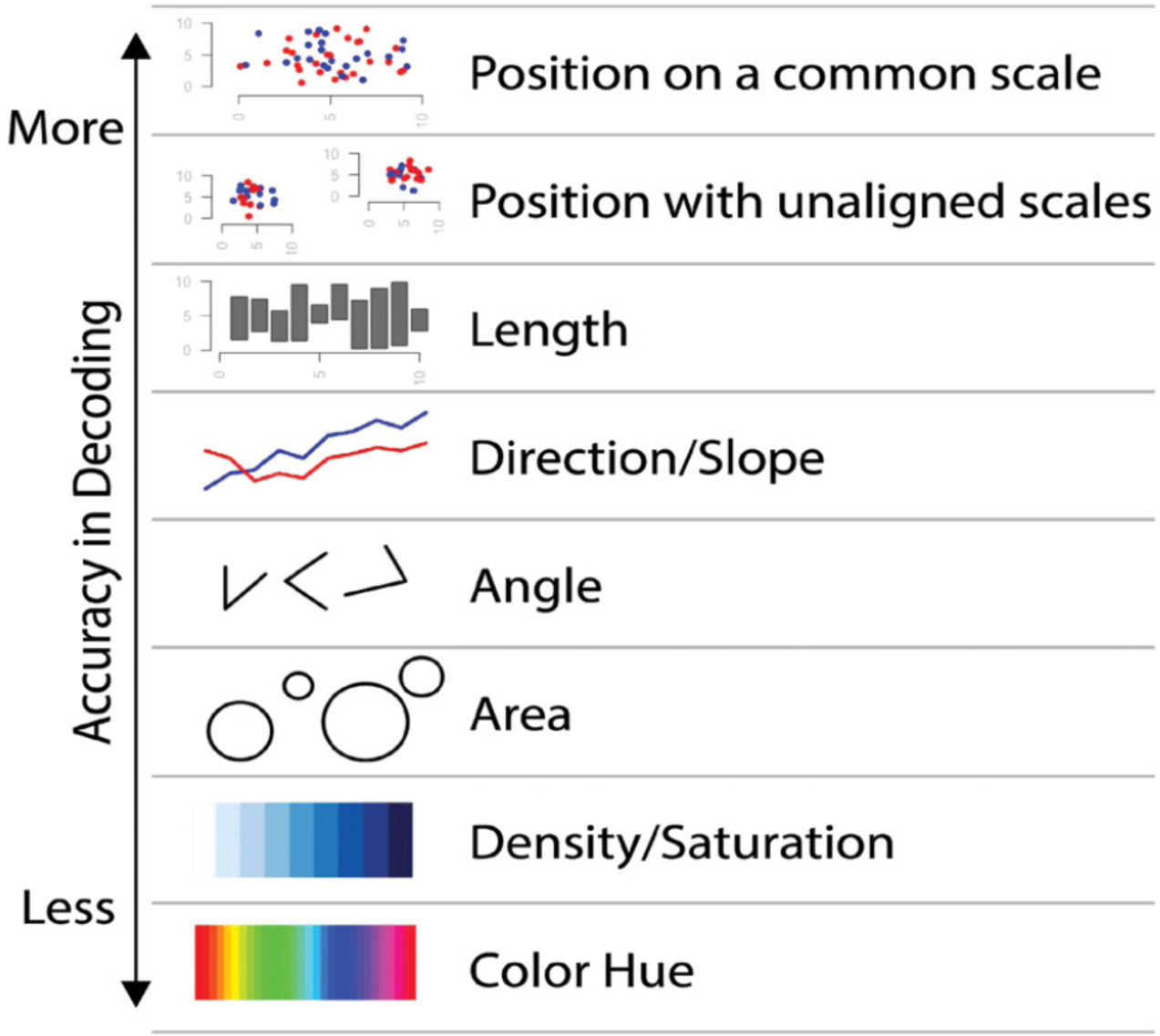
Targeting the Audience



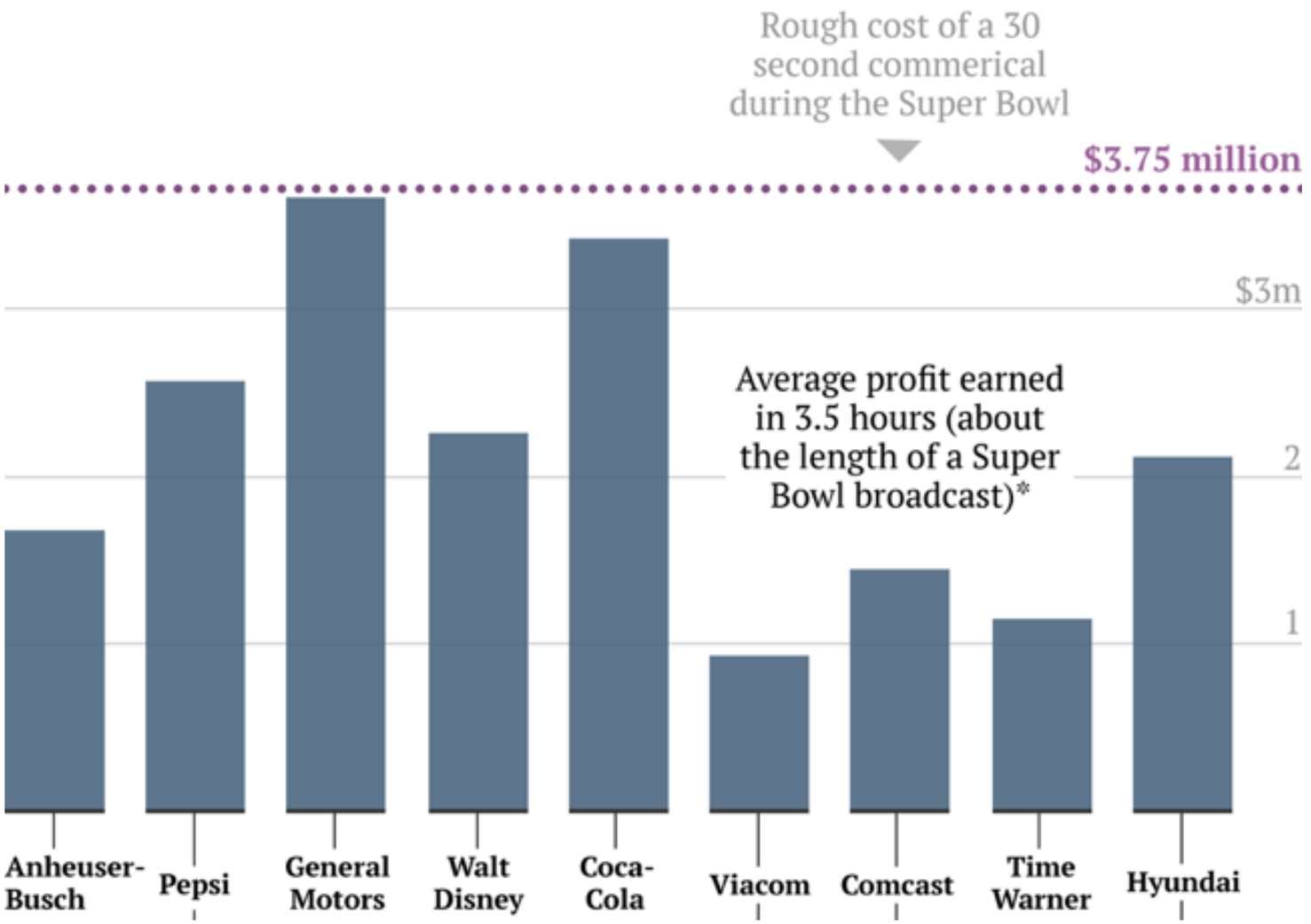
Ease of Decoding



"Automating the Design of Graphical Presentations of Relational Information", J. D. Mackinlay

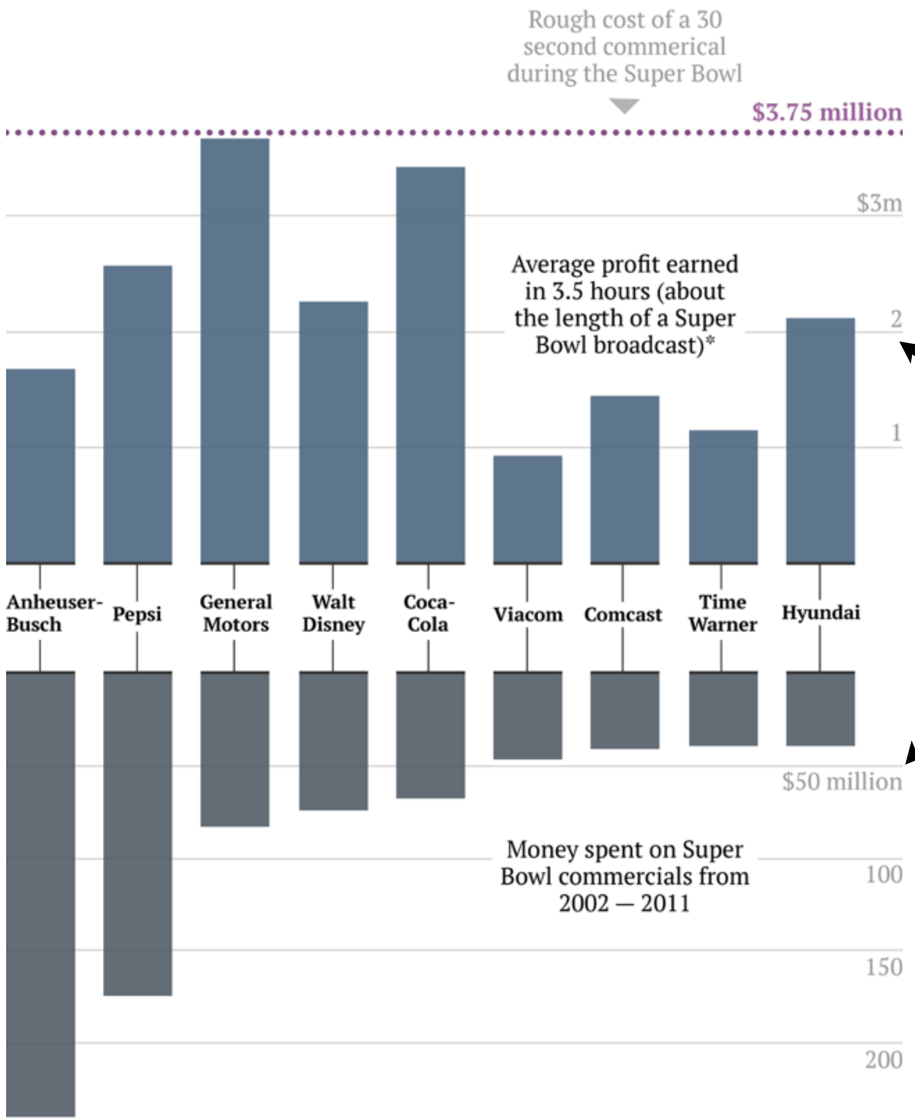


Position on a Common Scale



Scales are aligned

Position



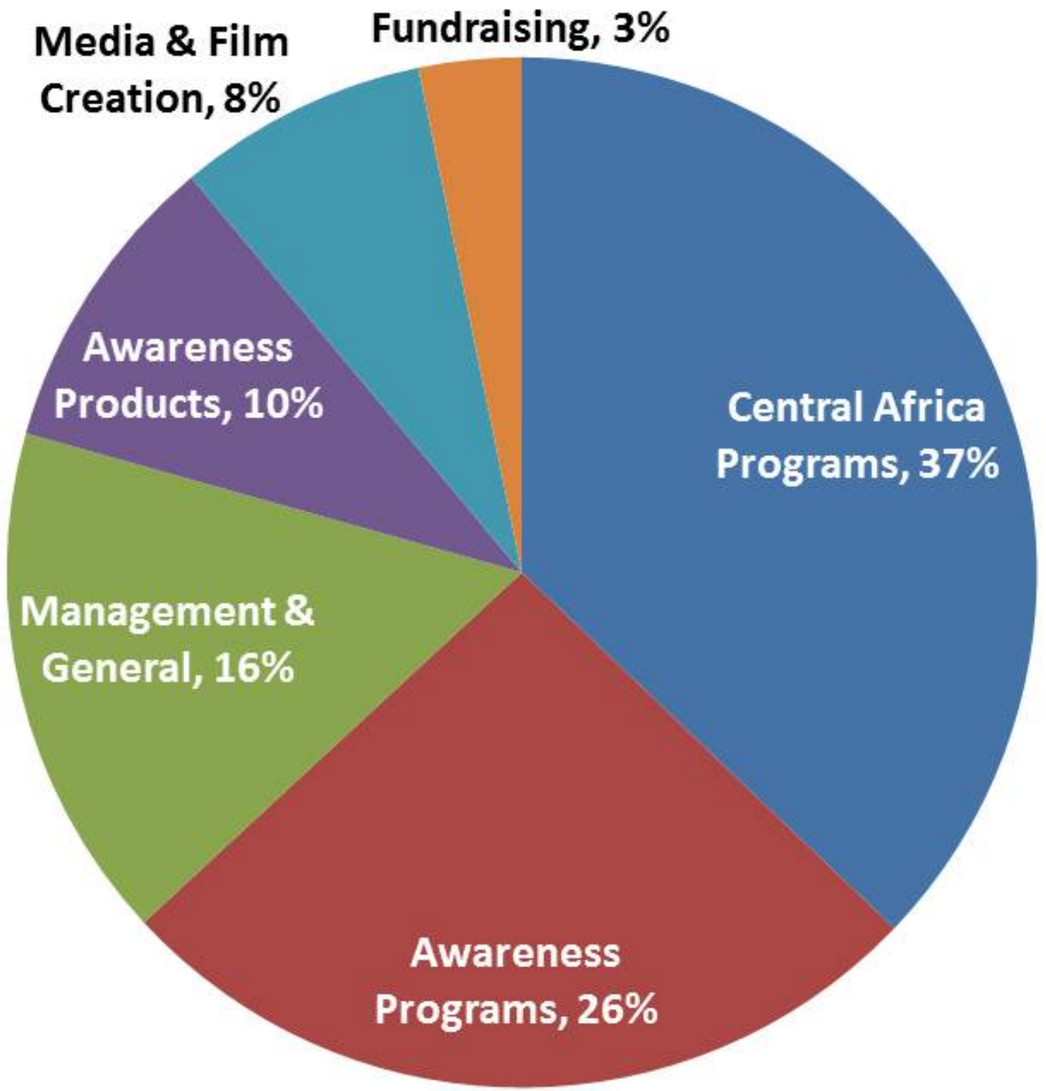
Note: this number is determined by taking the latest fiscal year earnings for each company, dividing by the number of hours in a year, then multiplying by 3.5.

Ritchie King | Quartz

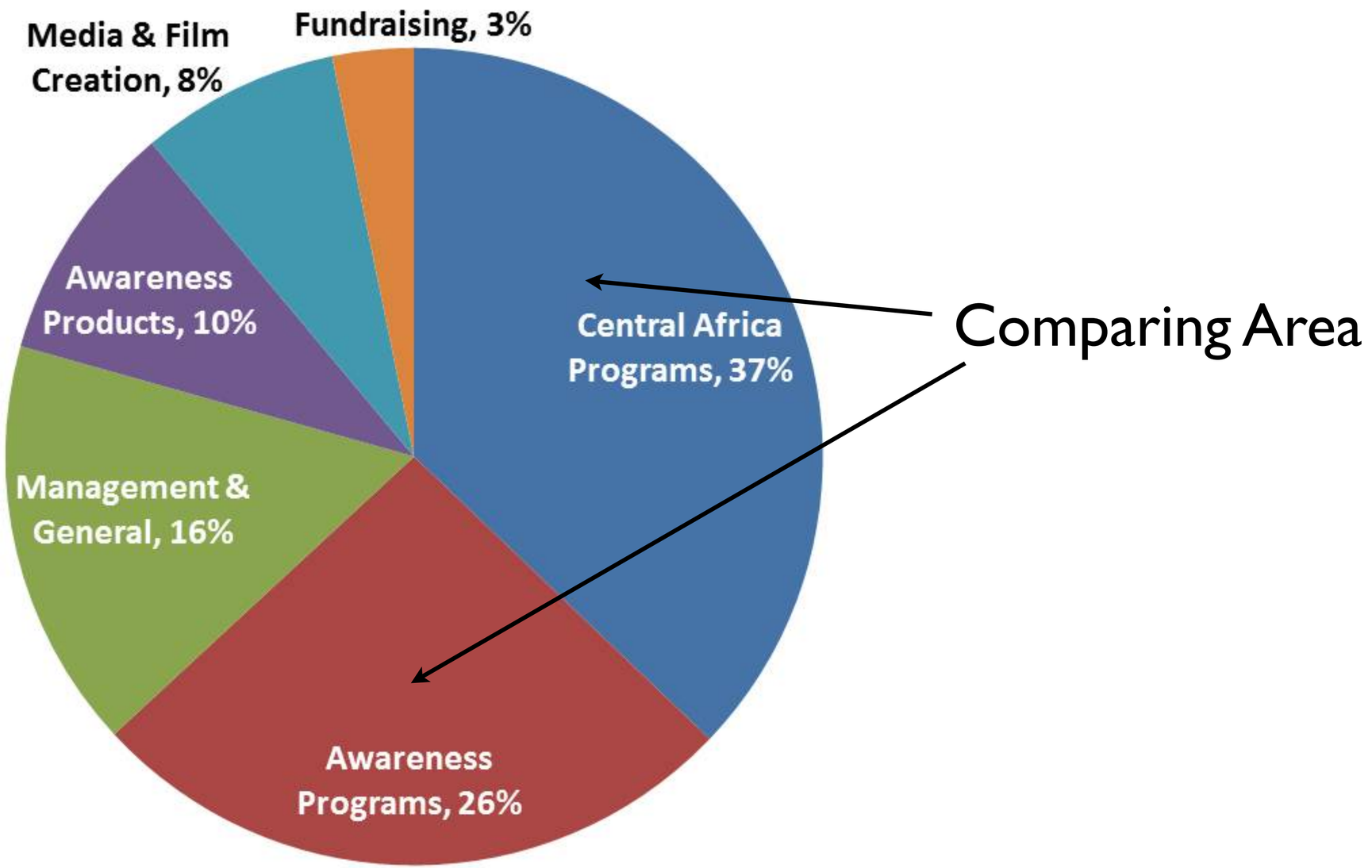
Data: Compiled by Factset, Kantar Media, New York Times

Scales are not aligned

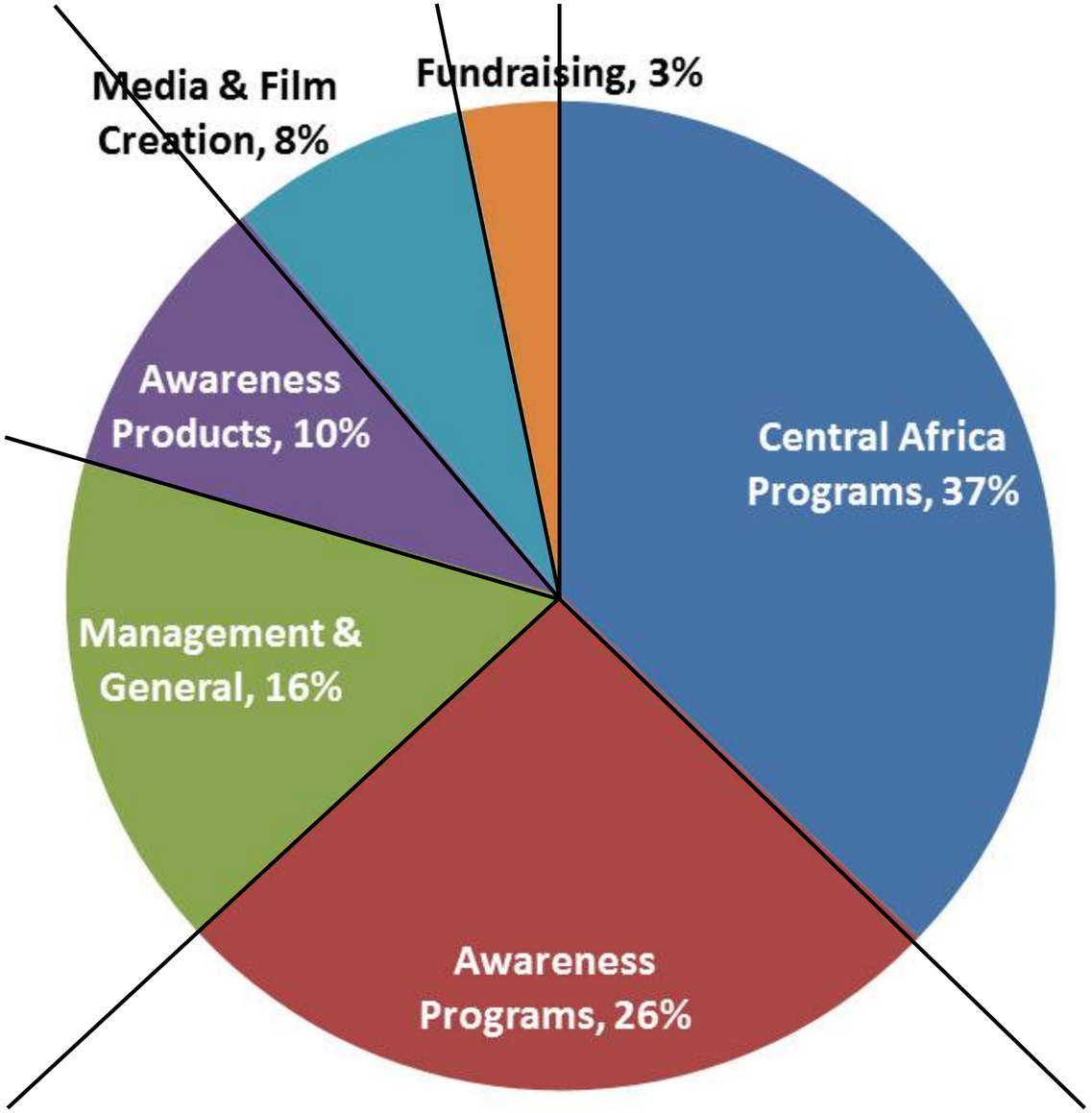
Interpreting a Pie Chart



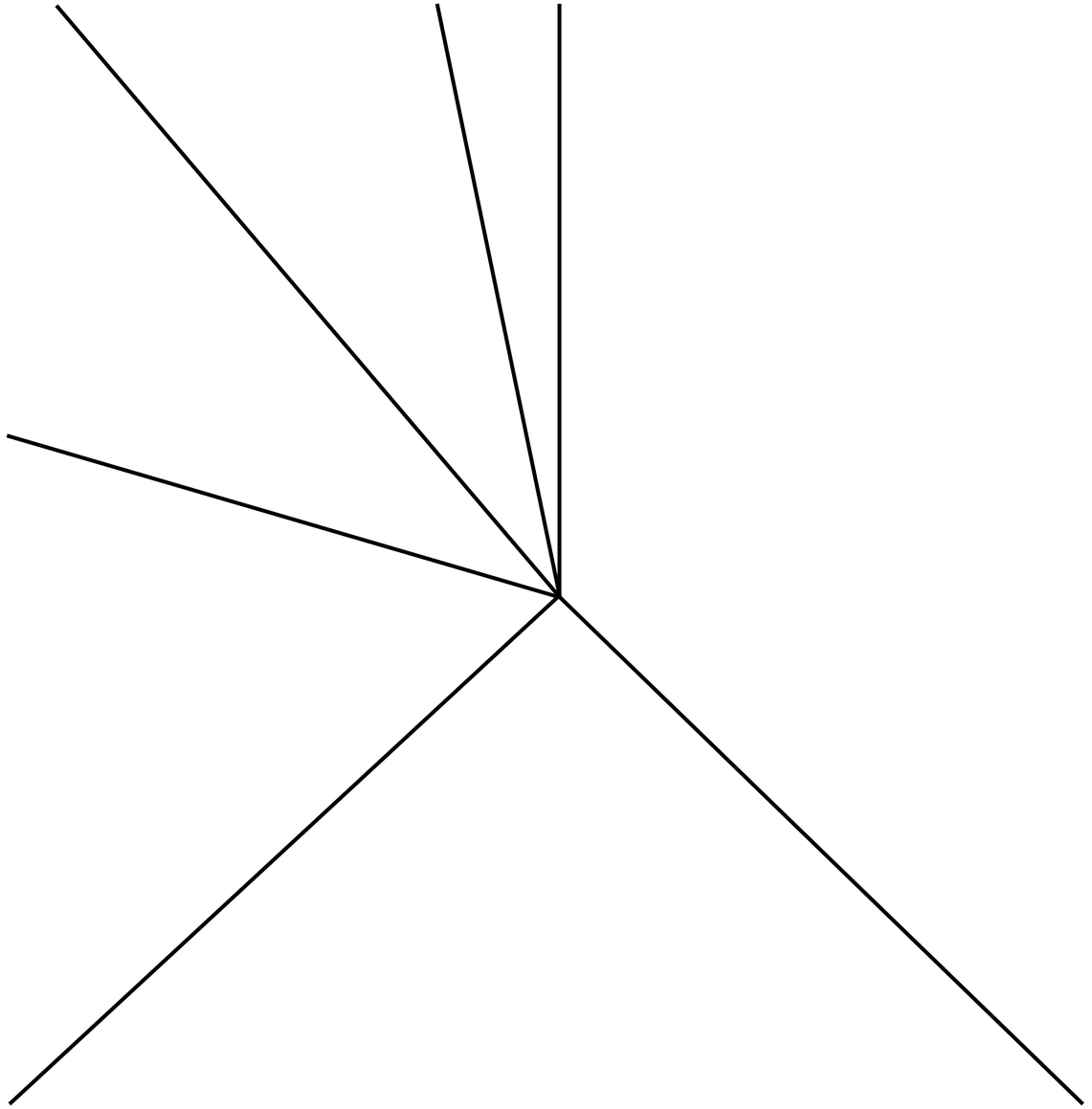
Interpreting a Pie Chart



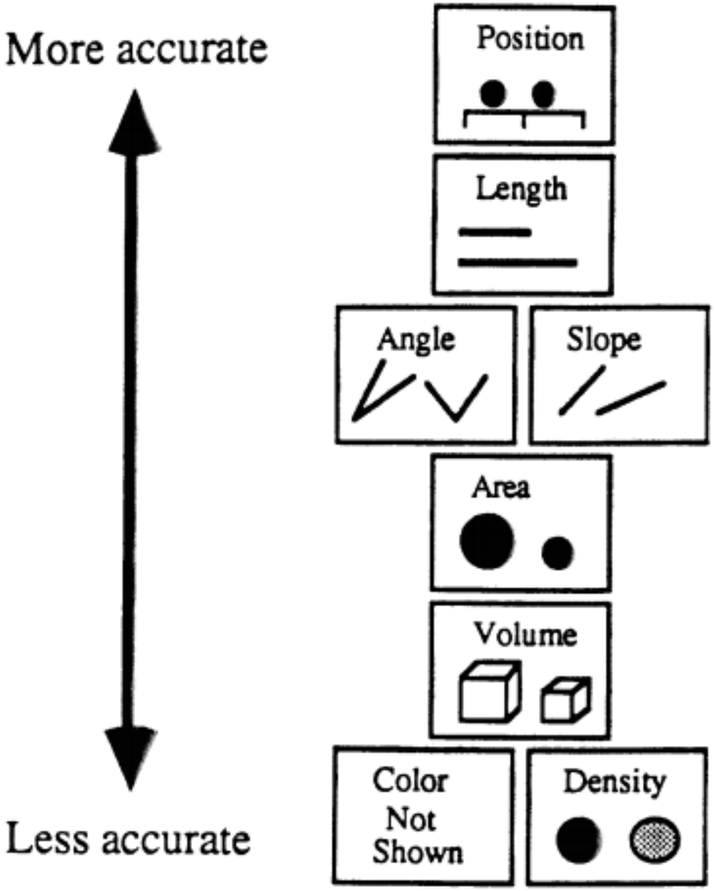
Interpreting a Pie Chart



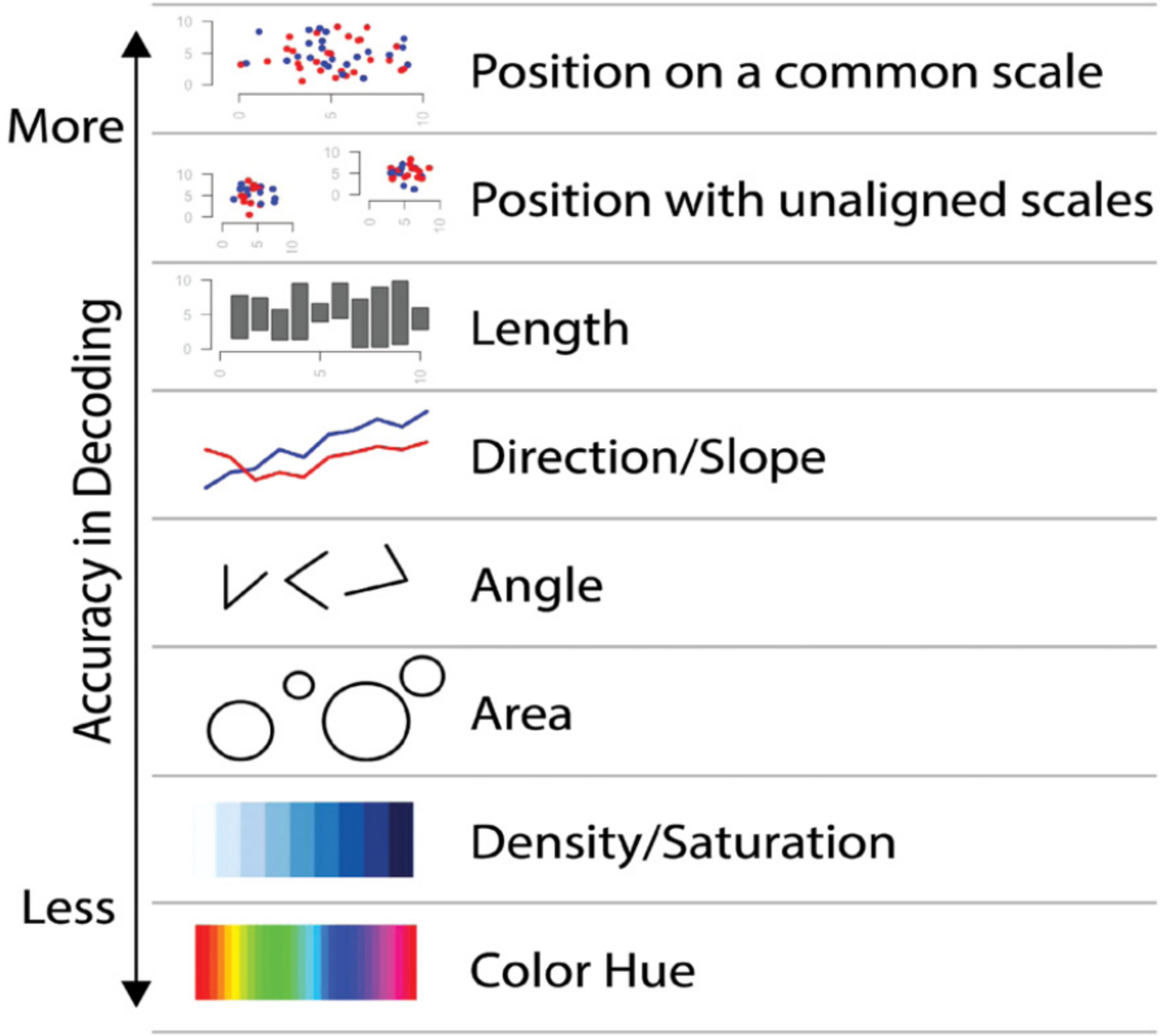
Comparing Angles



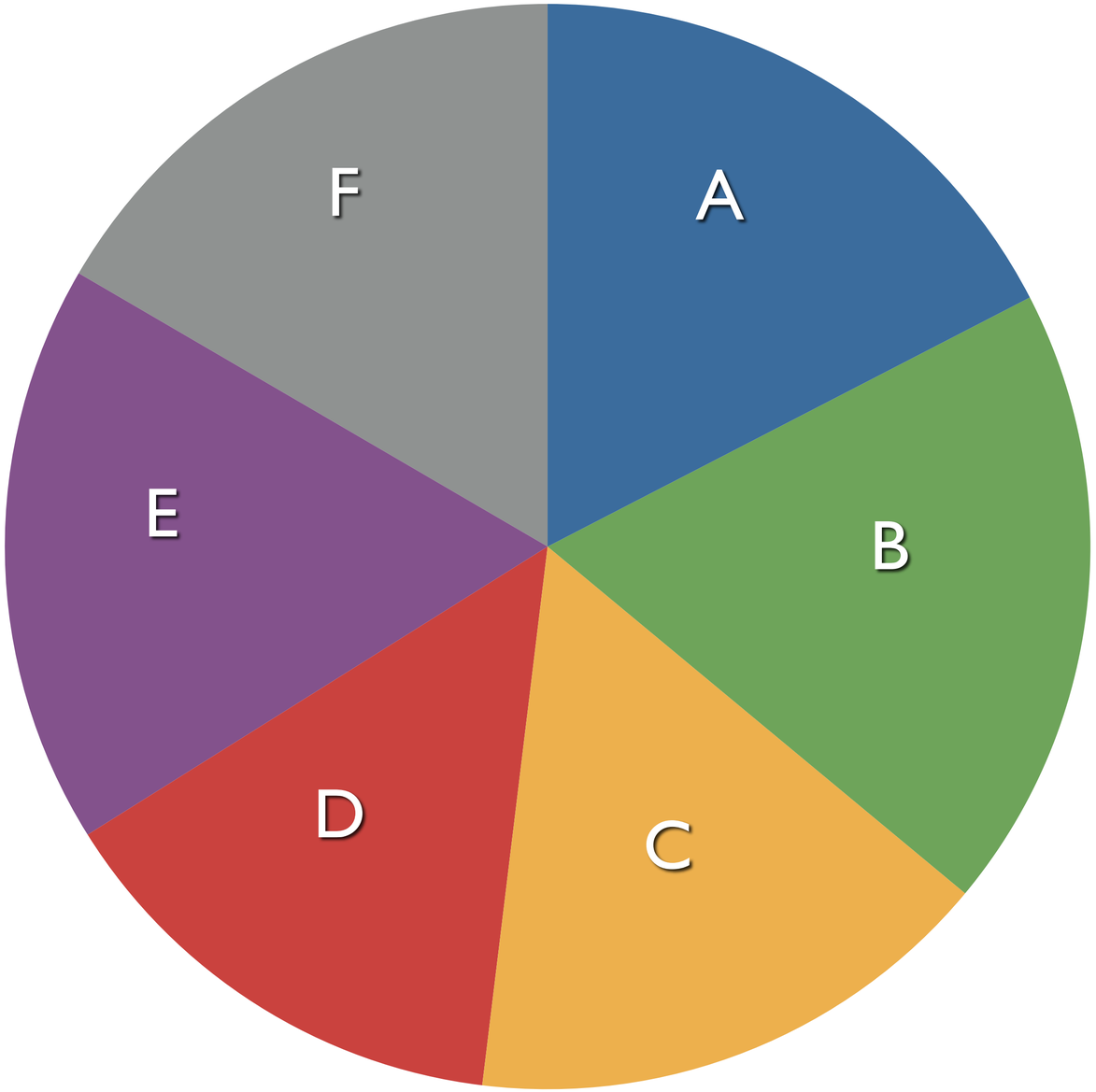
Ease of Decoding



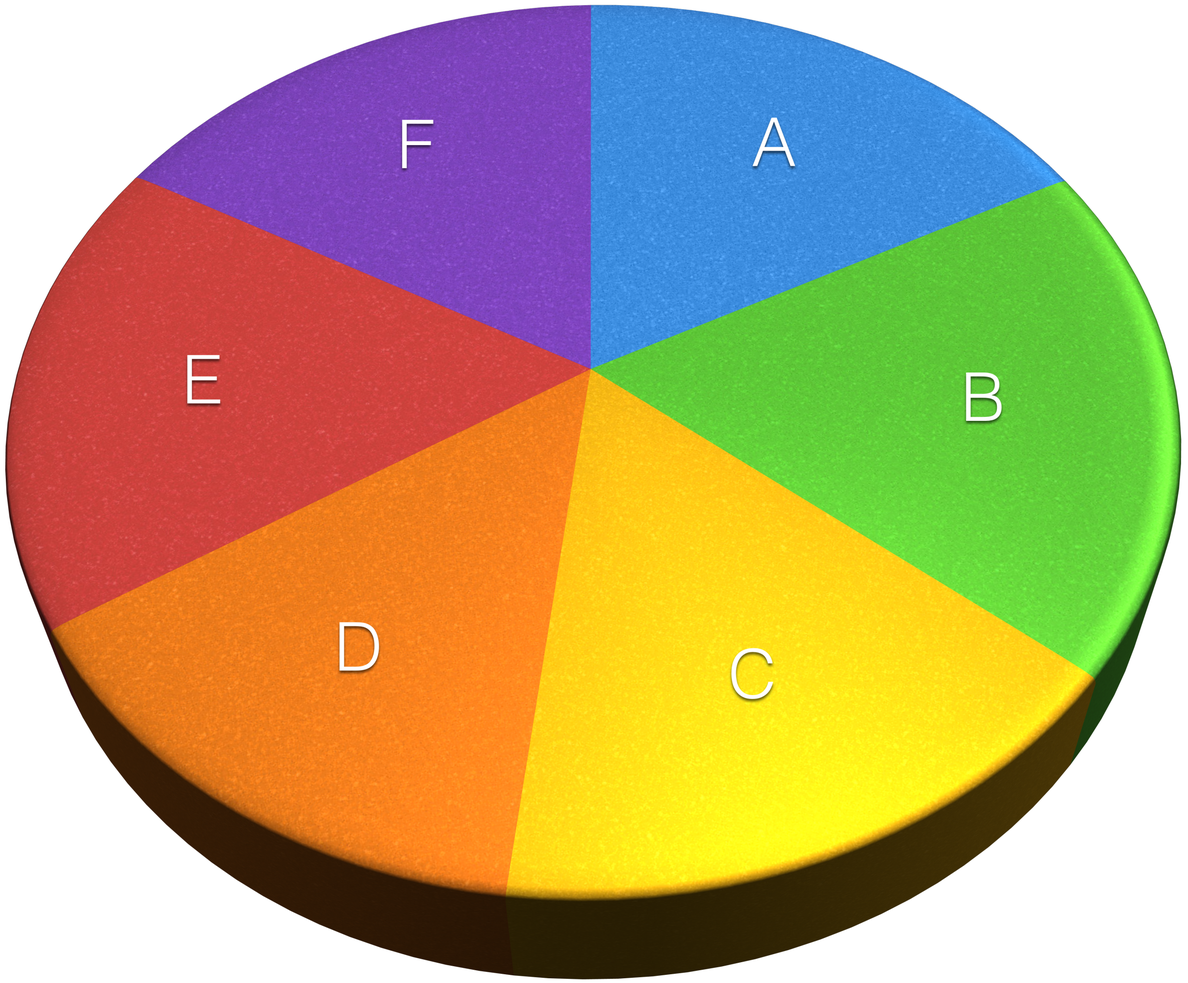
"Automating the Design of Graphical Presentations of Relational Information", J. D. Mackinlay



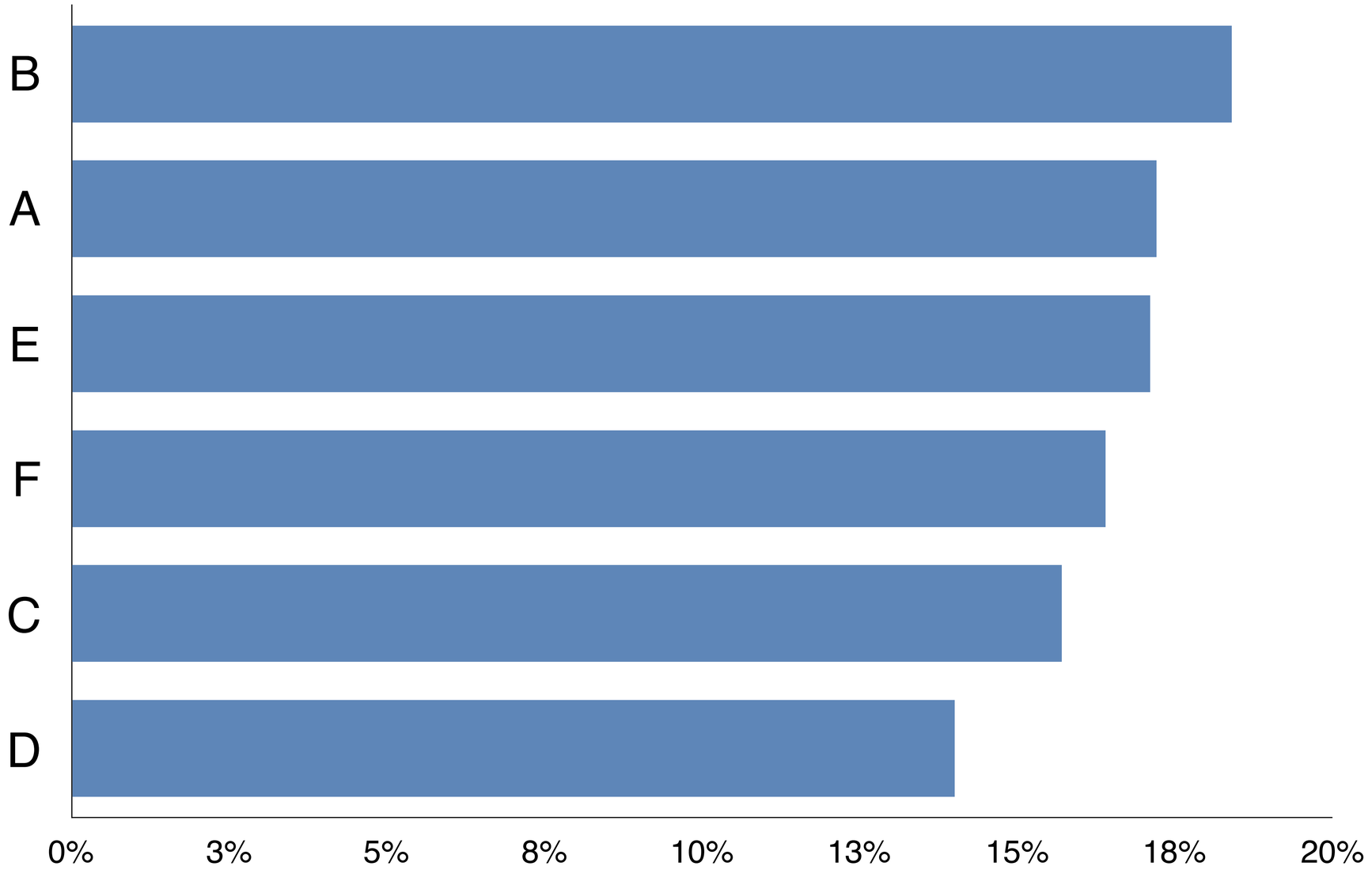
Which one is the Largest? Smallest?



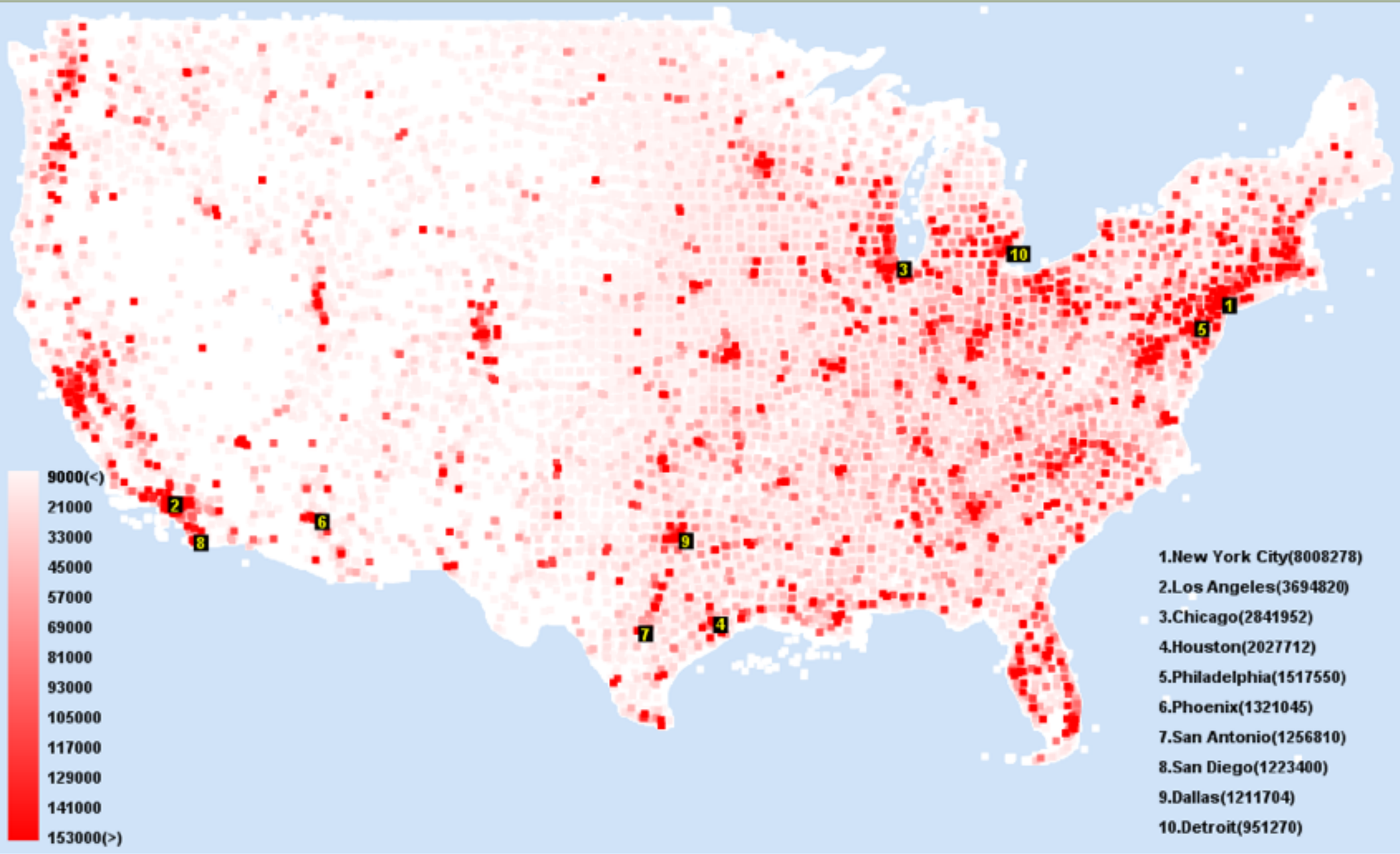
Which one is the Largest? Smallest?



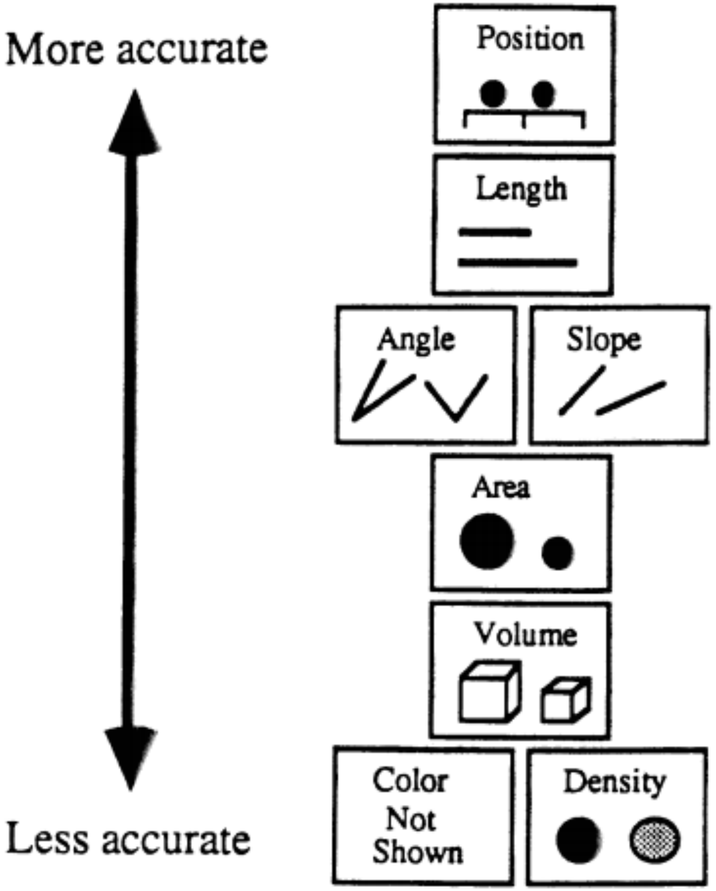
Which one is the Largest? Smallest?



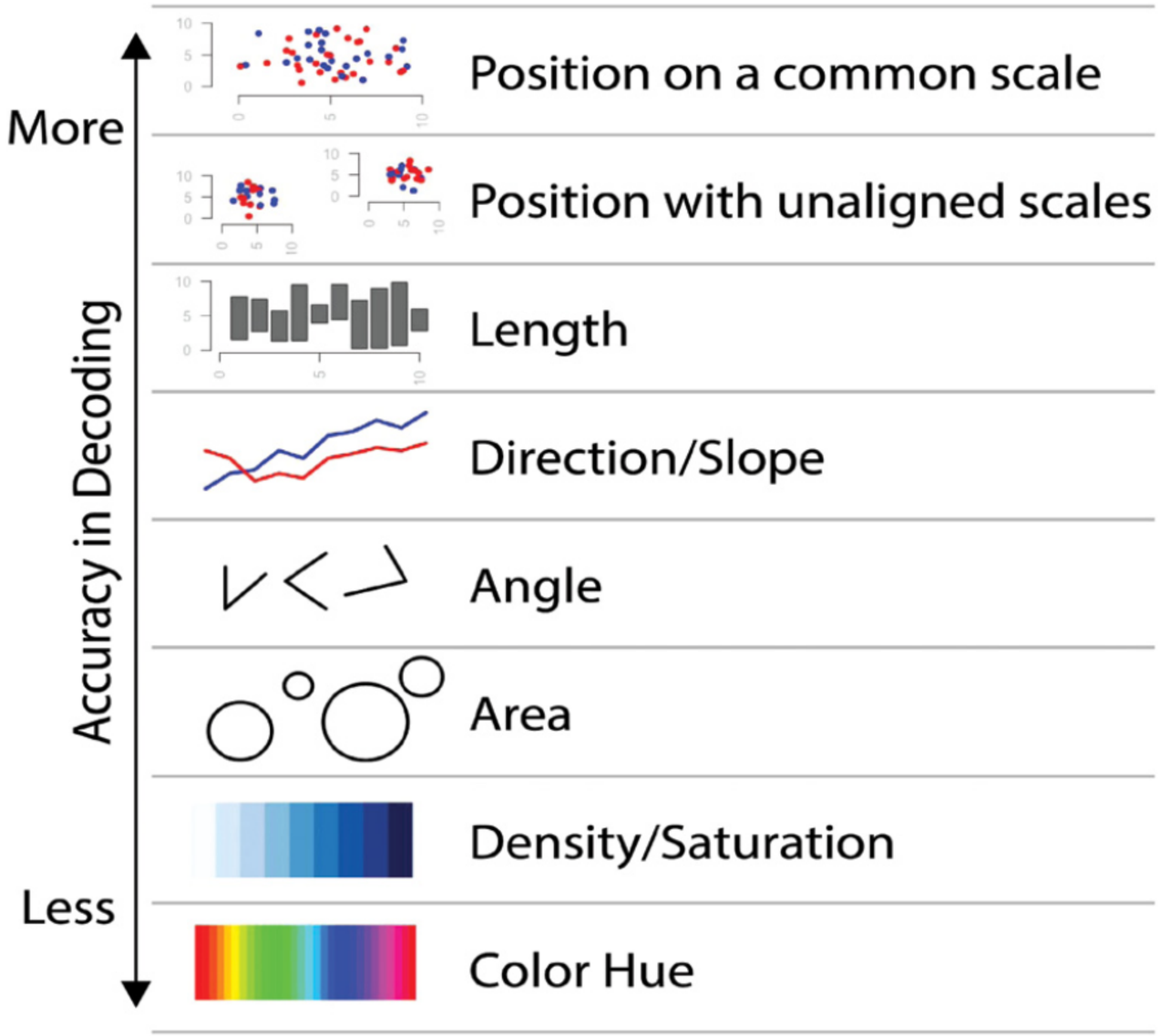
Saturation Communicating Quantity



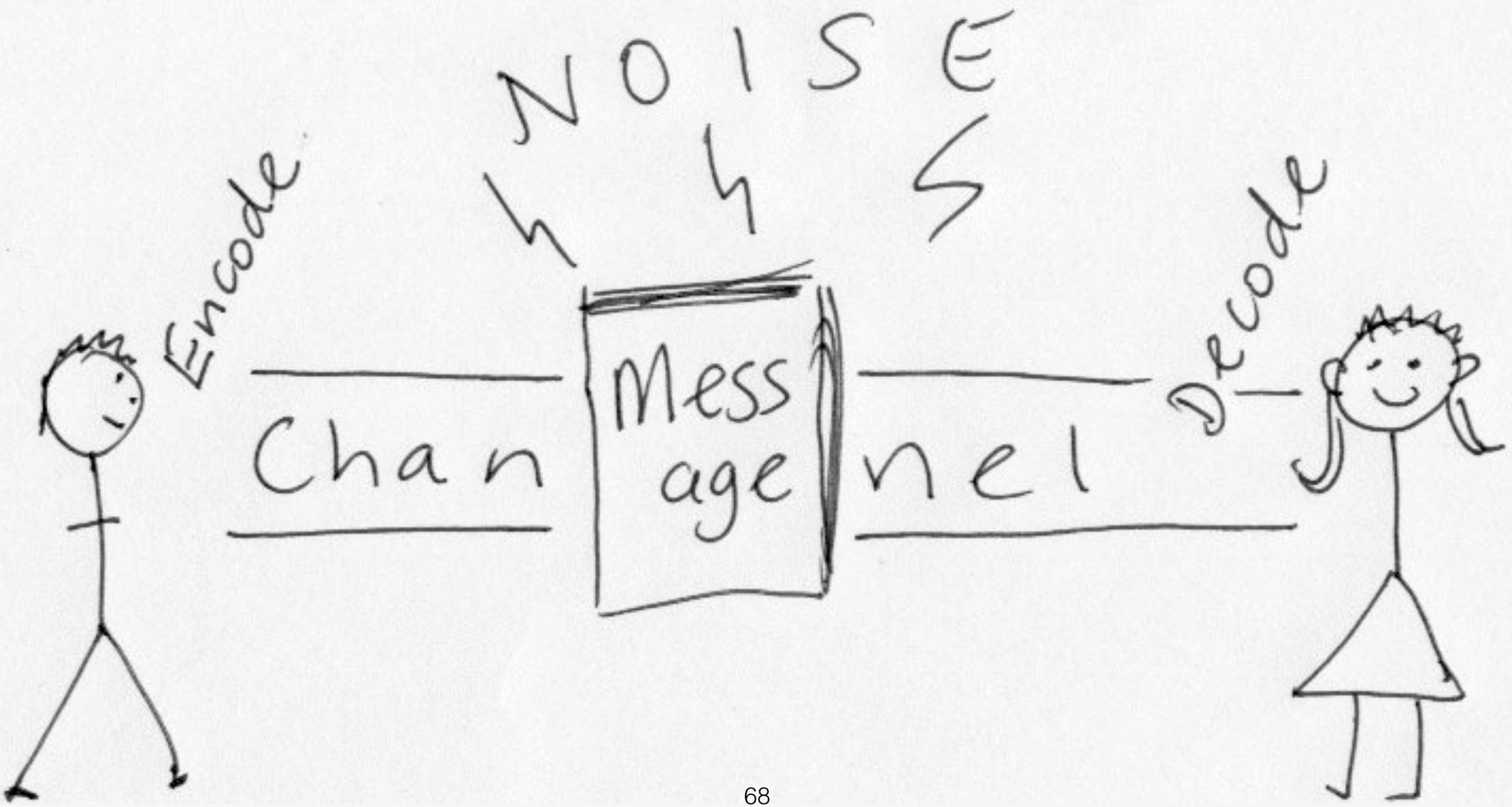
Ease of Decoding



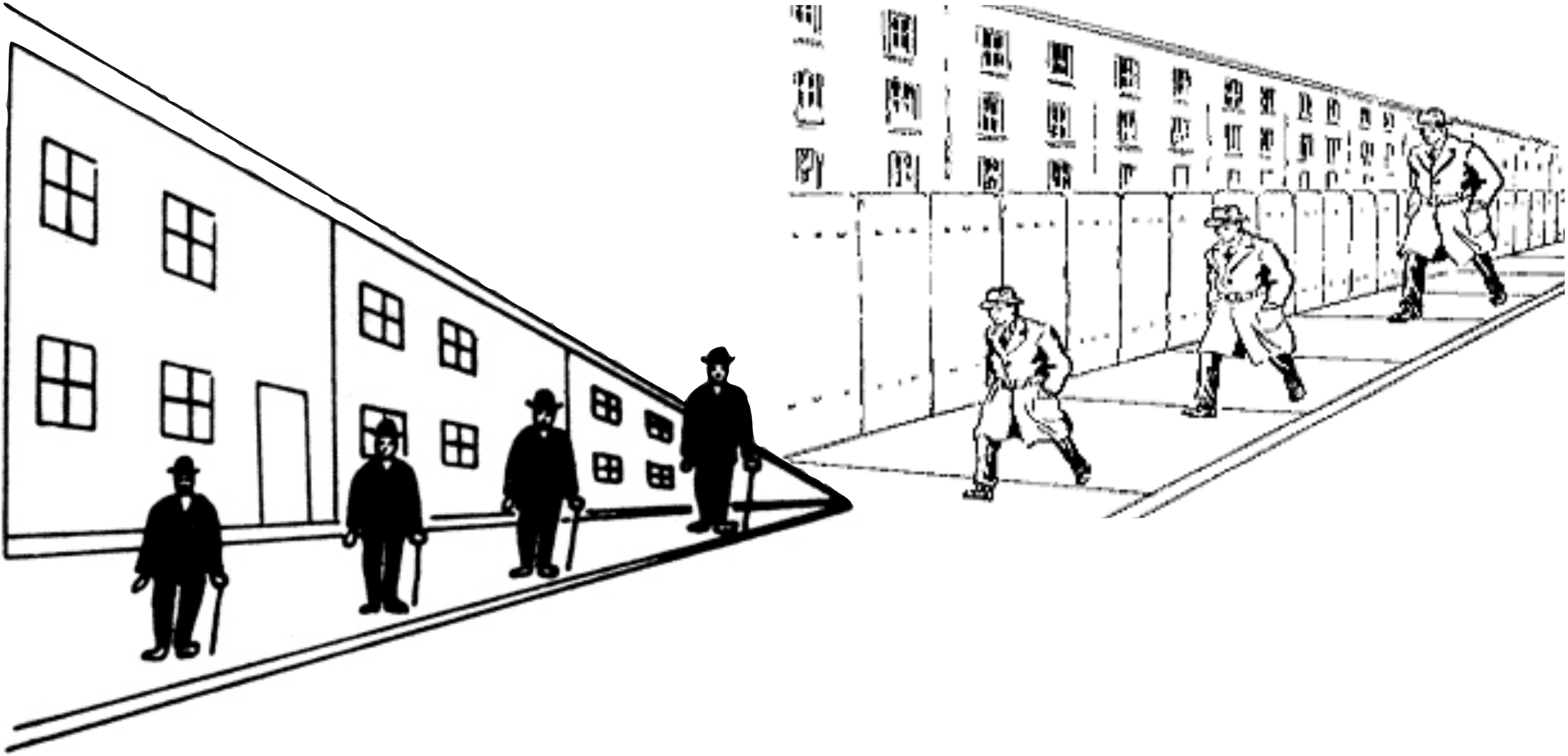
"Automating the Design of Graphical Presentations of Relational Information", J. D. Mackinlay



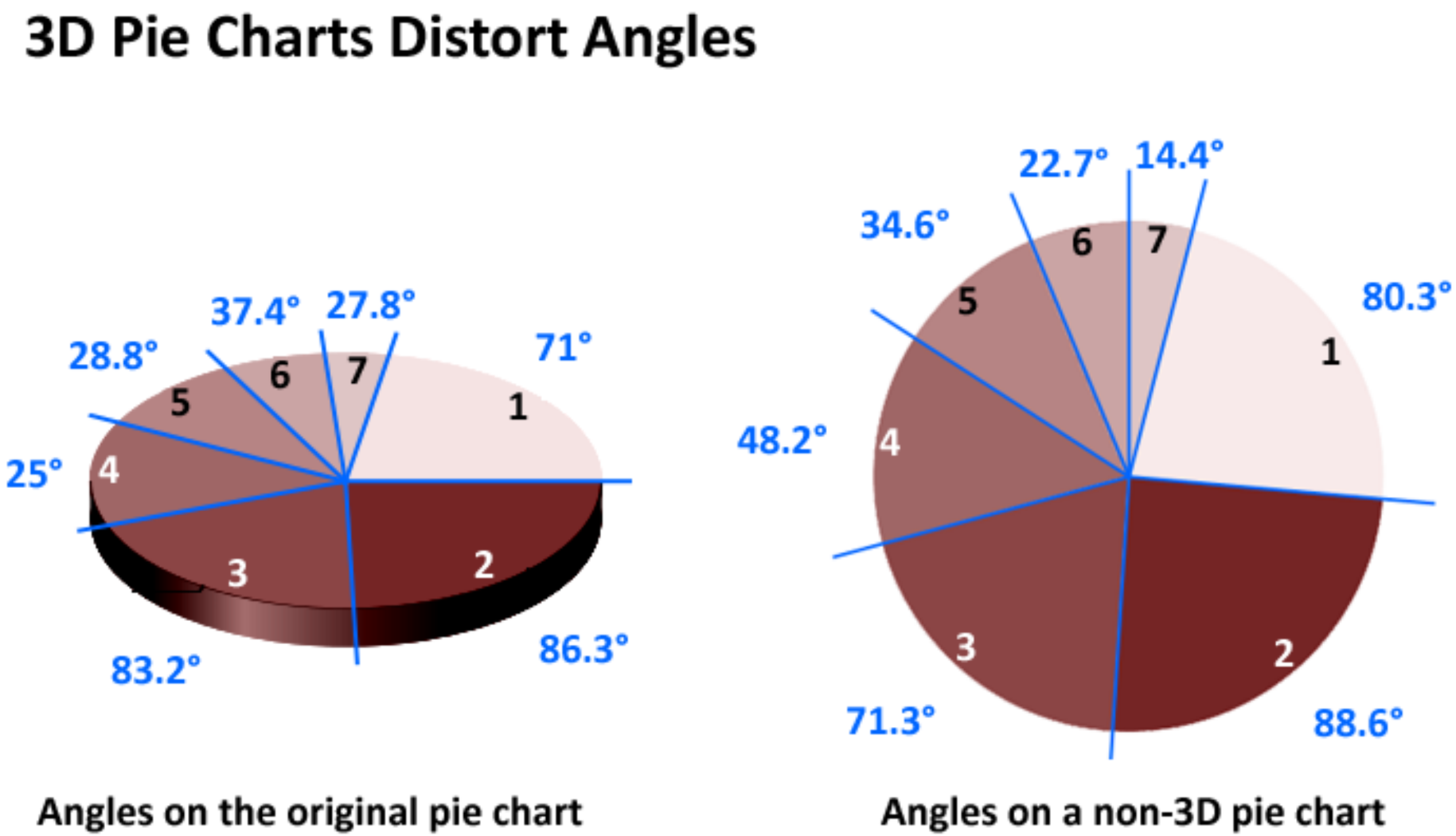
Everything Else



Avoid 3D Perspectives

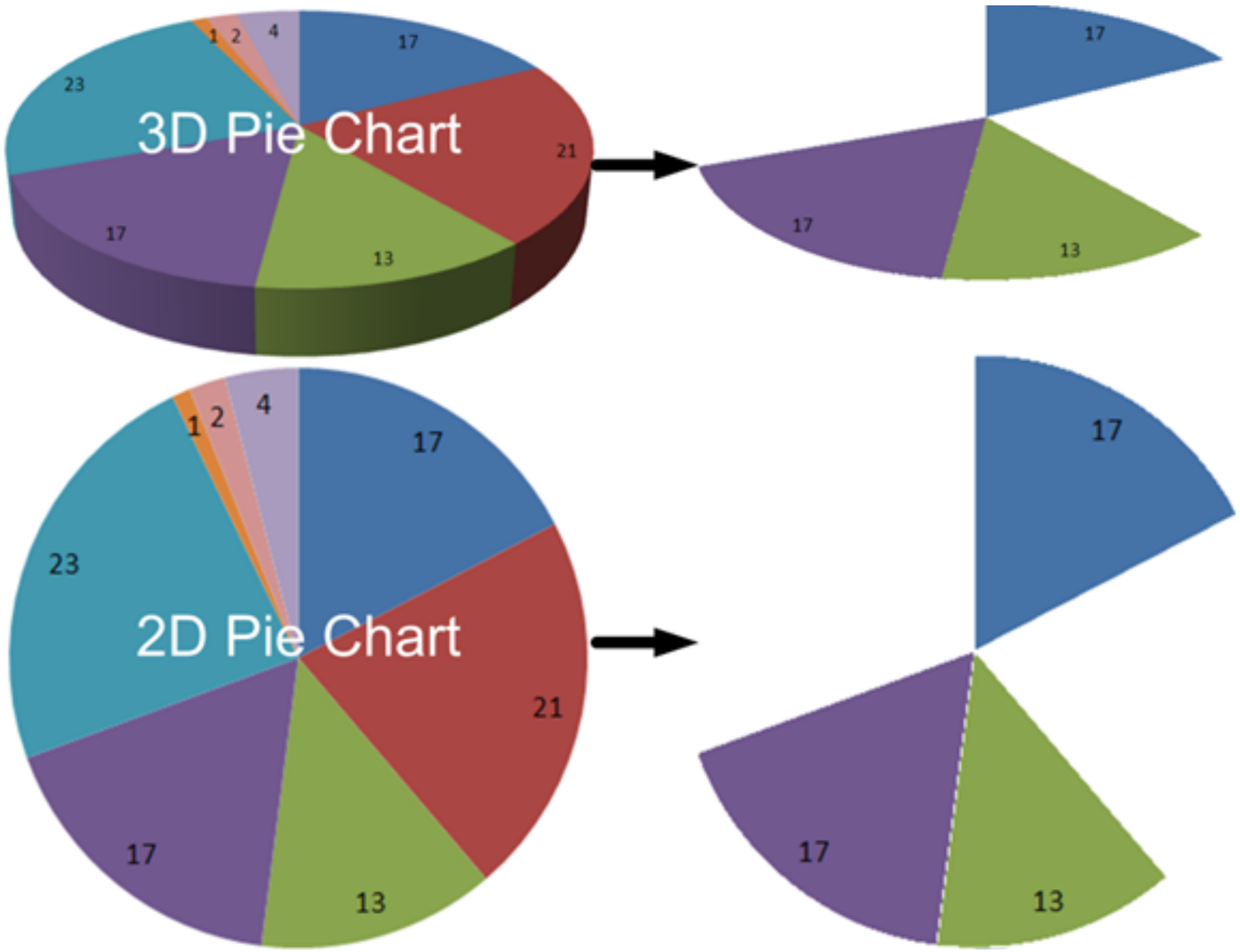


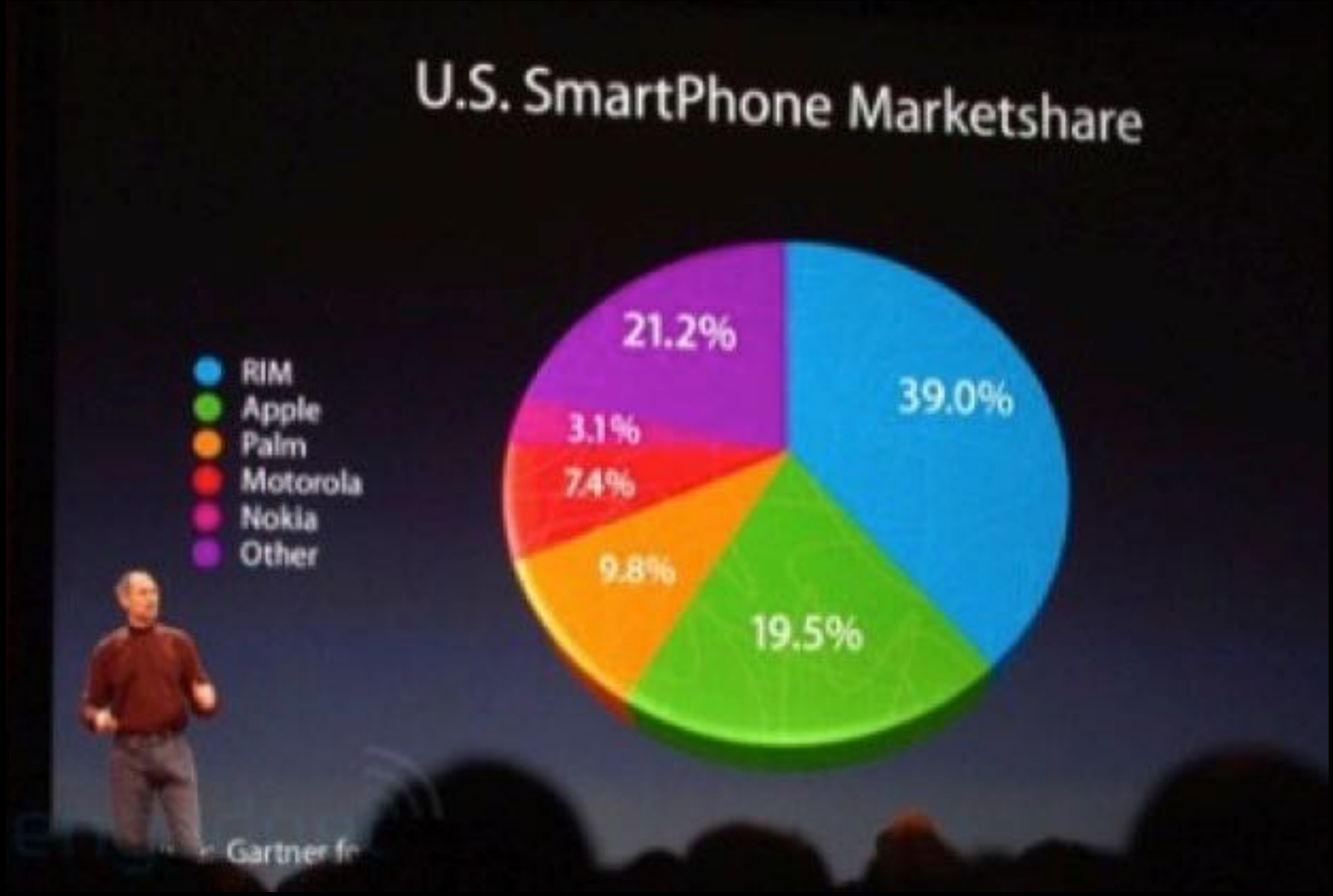
Pie charts



http://themendozaline.org/post/113828947026/3d-pie-charts-are-lie-charts

3D Pie Charts Distort Area




U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%
Gartner

3D Perspectives

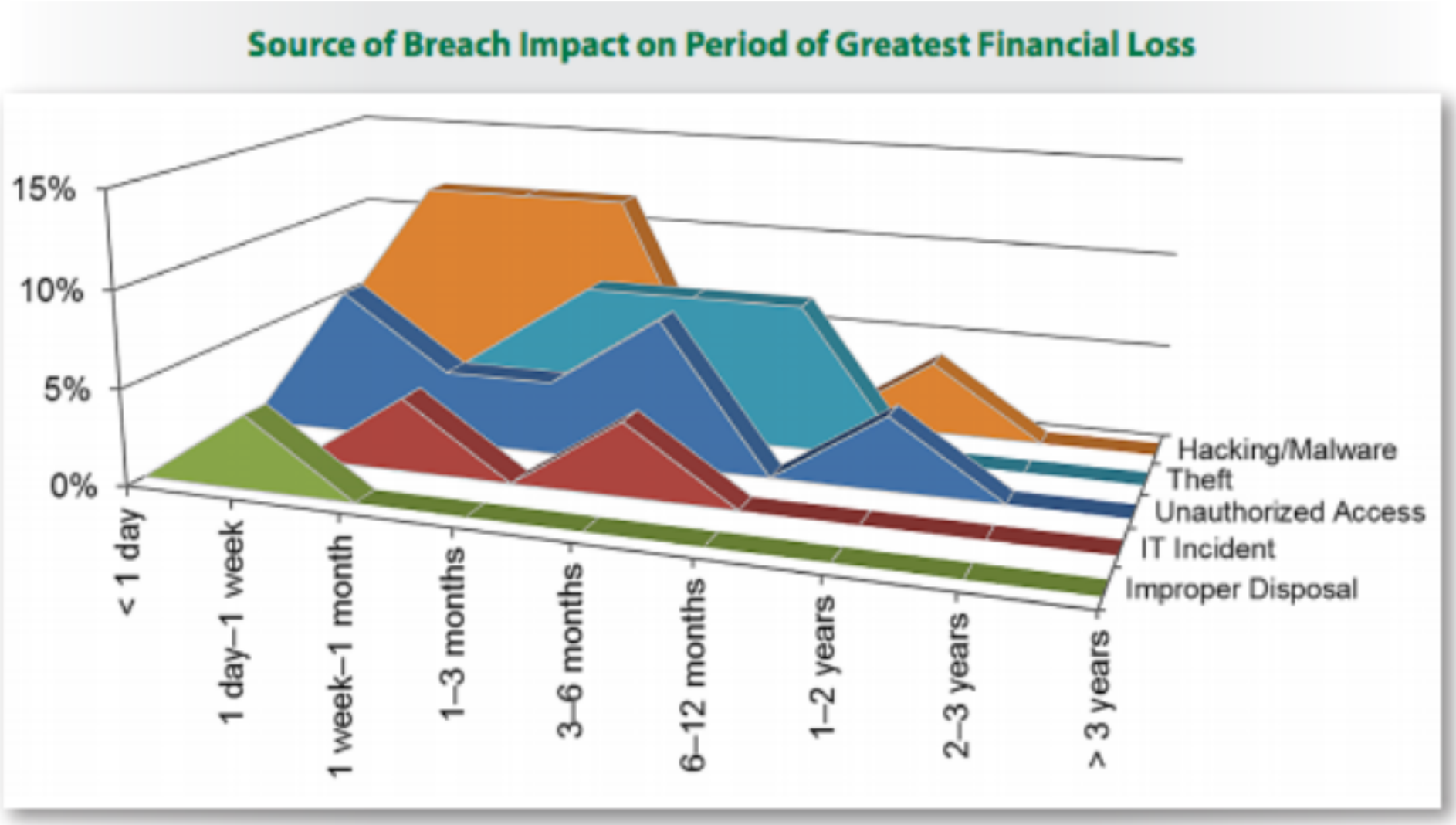
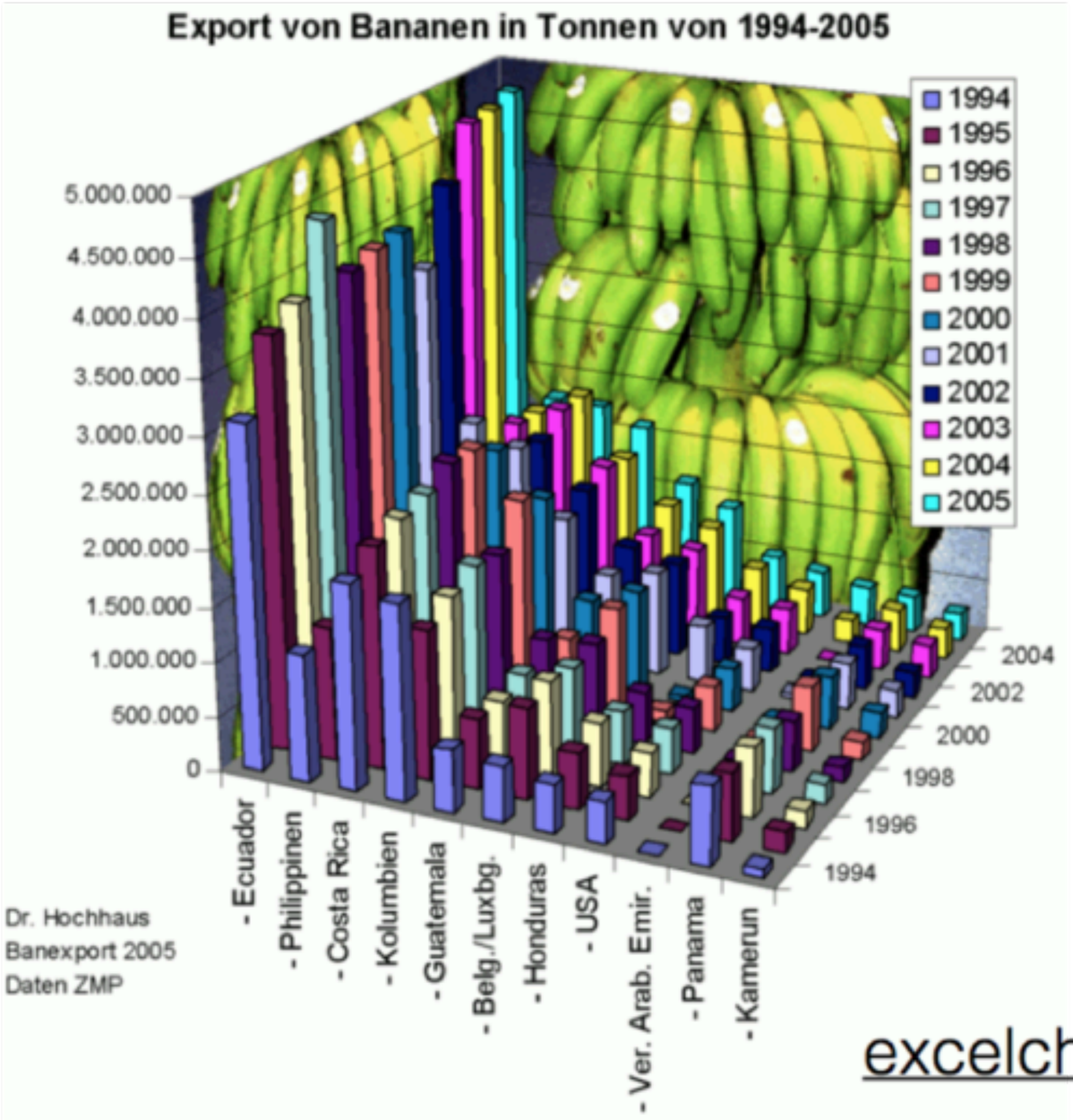


Figure 8. Source of Breach Effect on Period of Greatest Financial Loss

https://www.sans.org/reading-room/whitepapers/analyst/cleaning-breach-post-breach-impact-cost-compendium-36517

Just ouch…



excelcharts.com

O'REILLY®
Security
BUILD BETTER DEFENSES

oreillysecuritycon.com
#oreillysecurity

Visualizations in Code

Setting up for visualization

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
pd.options.display.mpl_style = 'default'
```

Python

R

```
library(tidyverse)
```

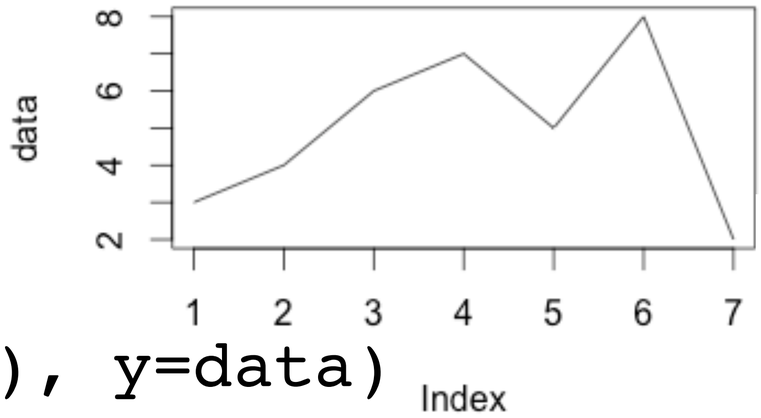
- **ggplot2**, for data visualisation.
- **dplyr**, for data manipulation.
- **tidyr**, for data tidying.
- **readr**, for data import.
- **purrr**, for functional programming.
- **tibble**, a modern re-imagining of data frames.

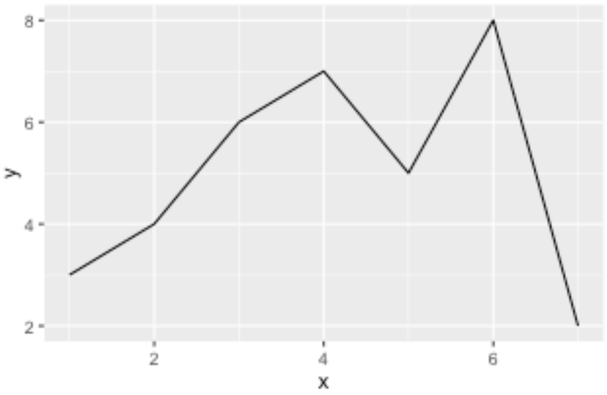
Line Plots

```
data = pd.Series([3,4,6,7,5,8,2 ])
graph = data.plot()
```

Python

```
data <- c(3,4,6,7,5,8,2)
plot(data, type="l")

dataf <- data.frame(x=seq(7), y=data)
ggplot(dataf, aes(x, y)) +
geom_line()
```

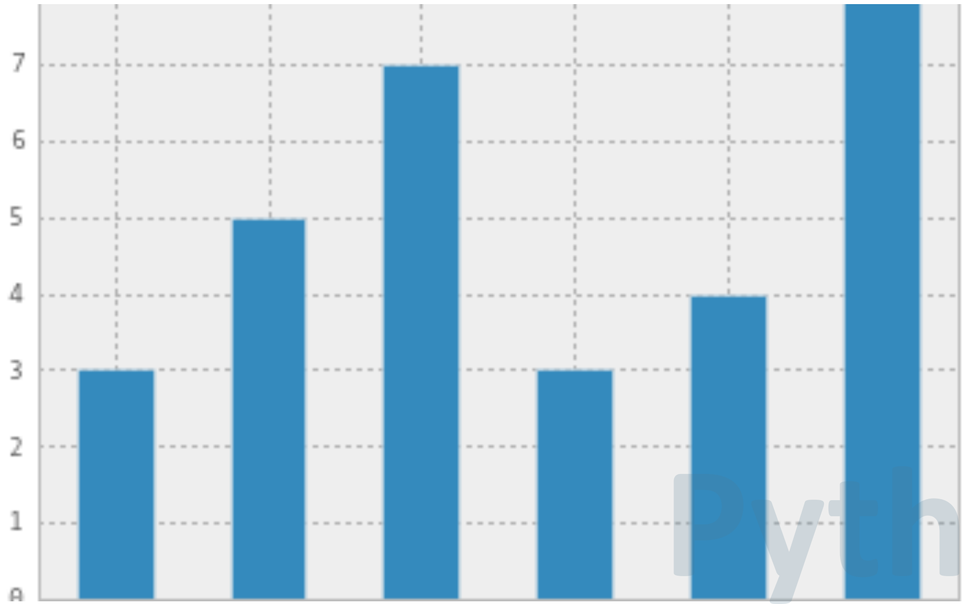




R

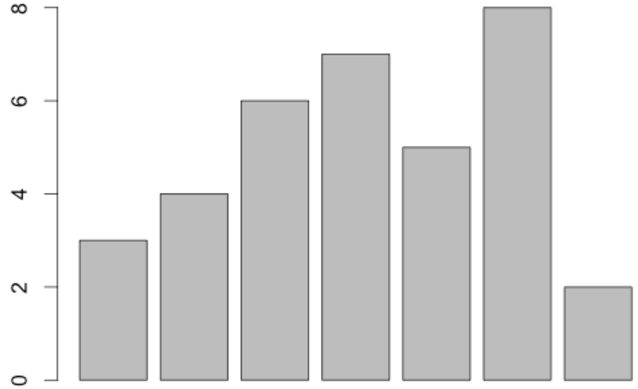
Bar Plots

```
barchart = data.plot( kind="bar" )
```



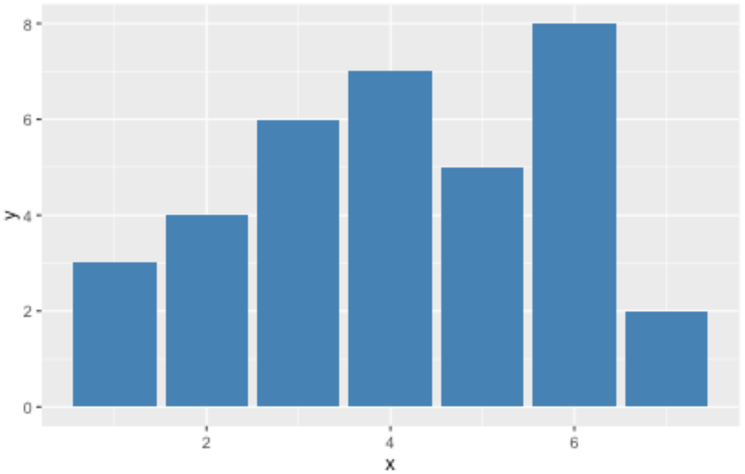
Python

```
barplot(data)
```



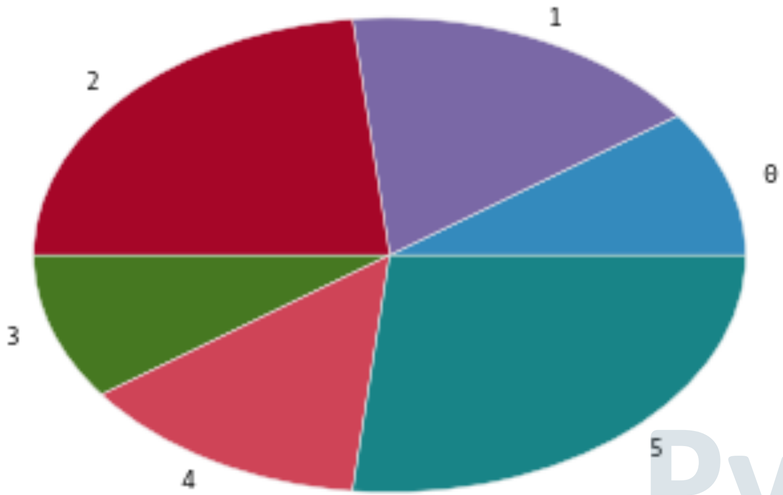
R

```
ggplot(dataf, aes(x, y)) +
geom_bar(stat="identity", fill="steelblue")
```



Pie Charts

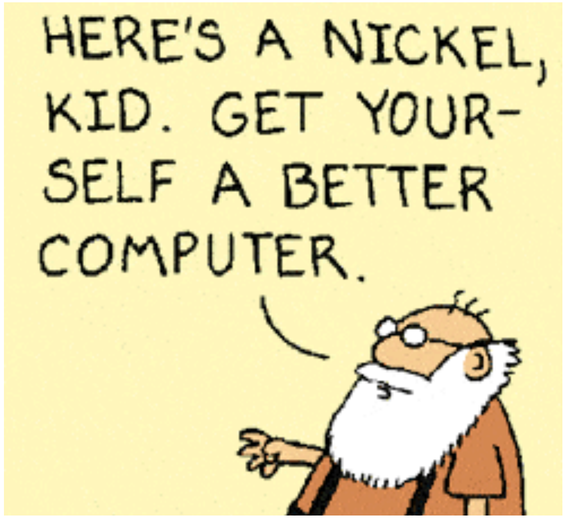
```
piechart = data.plot( kind="pie" )
```



Python

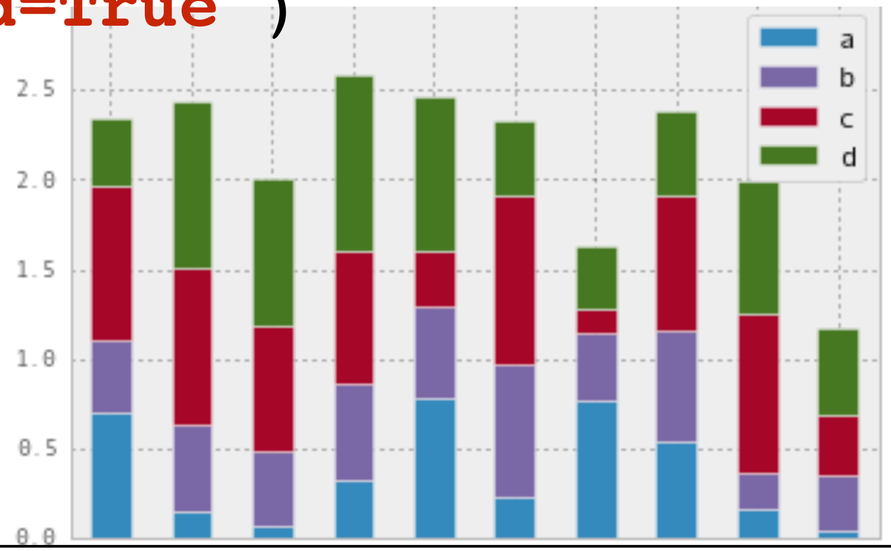
R

. . .



Stacked Bar Charts

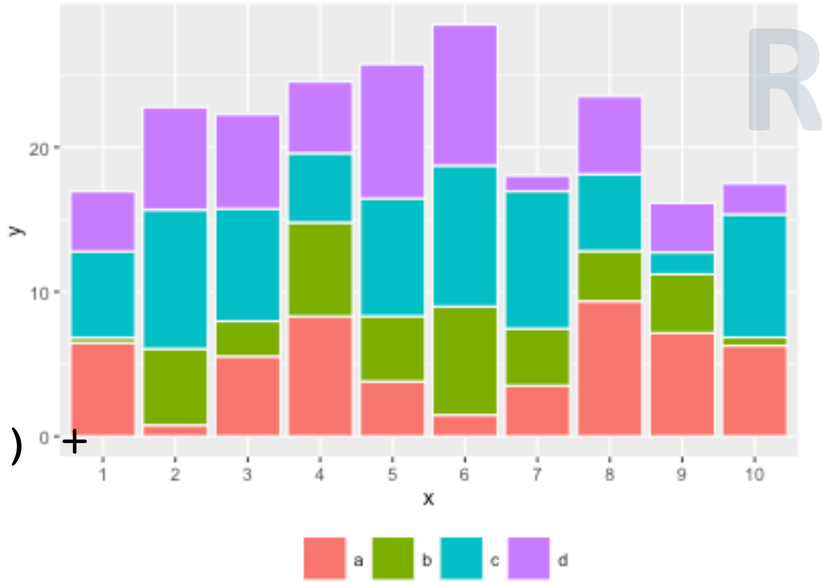
```
df2.plot( kind='bar', stacked=True )
```



Python

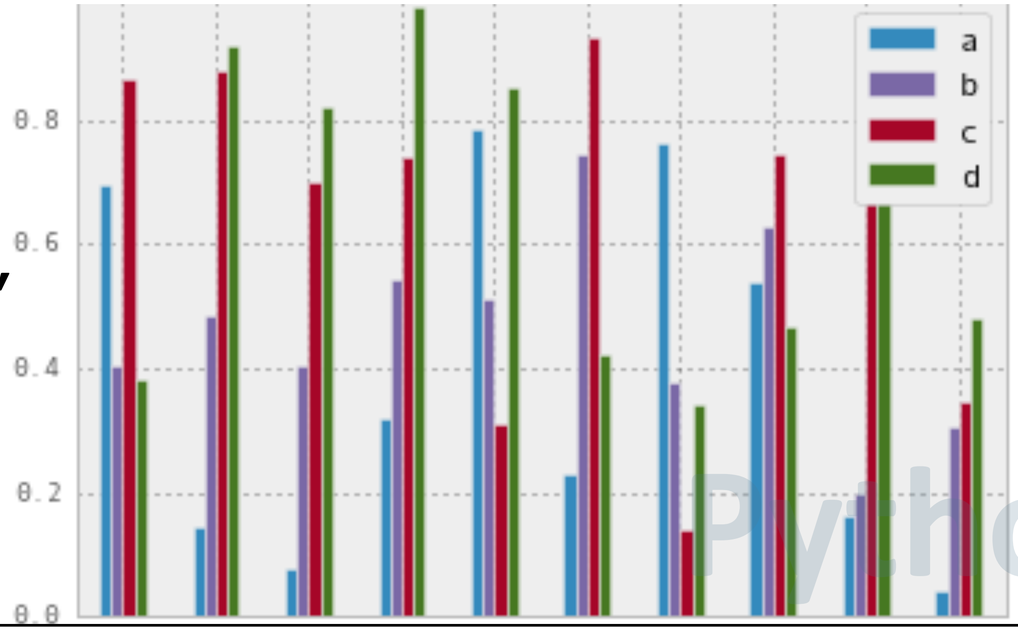
R

```
ggplot(df2, aes(x,y, fill=cat)) +
  geom_bar(stat="identity", position="stack", color="white") +
  theme(legend.position="bottom",
        legend.title = element_blank())
```



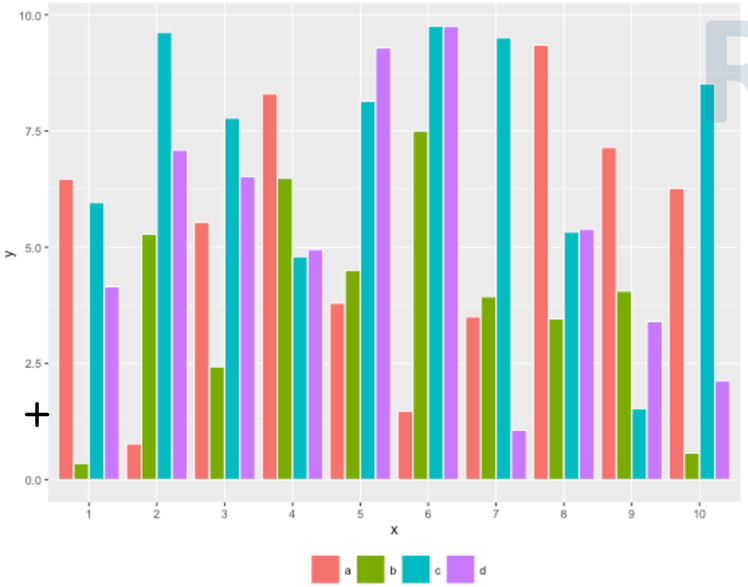
Grouped Bar Charts

```
df2 = pd.DataFrame(np.random.rand(10, 4),
columns=['a', 'b', 'c', 'd'])
df2.plot( kind='bar' )
```



Python

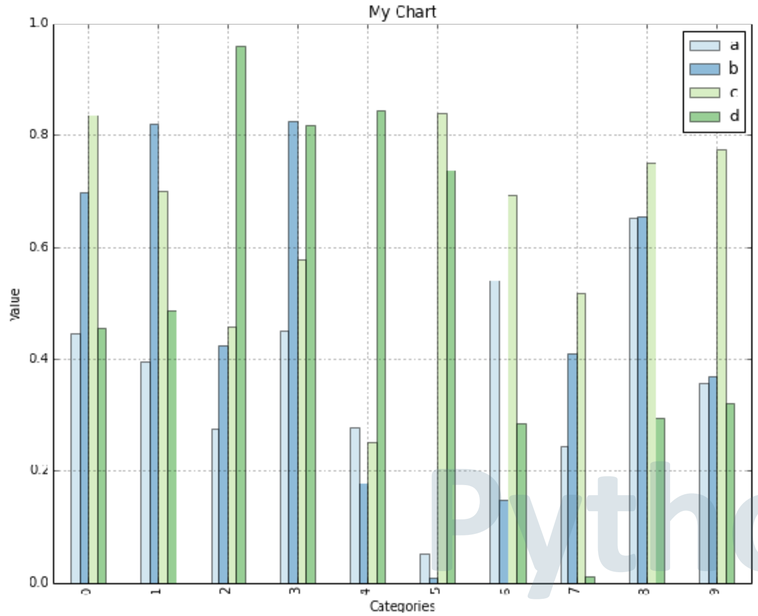
```
df2 <- tibble(x=factor(rep(seq(10), each=4)),
              y=runif(40,0,10),
              cat=rep(letters[1:4], 10))


ggplot(df2, aes(x,y, fill=cat)) +
  geom_bar(stat="identity", position="dodge", color="white") +
  theme(legend.position="bottom",
        legend.title = element_blank())
```



R

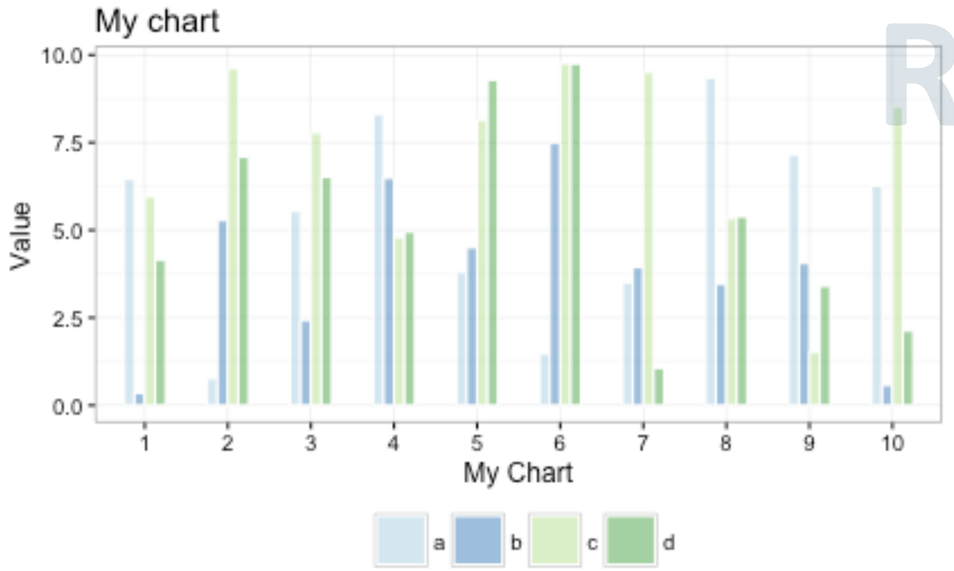
Grouped Bar Chart (custom fill)

```
df2.plot( kind='bar',
          color=('#a6cee3','#1f78b4','#b2df8a','#33a02c' ),
           alpha=0.5,
             width=0.5,
             figsize=(10, 8))
plt.title( "My Chart" )
plt.xlabel( "Categories" )
plt.ylabel( "Value" )
```



Python

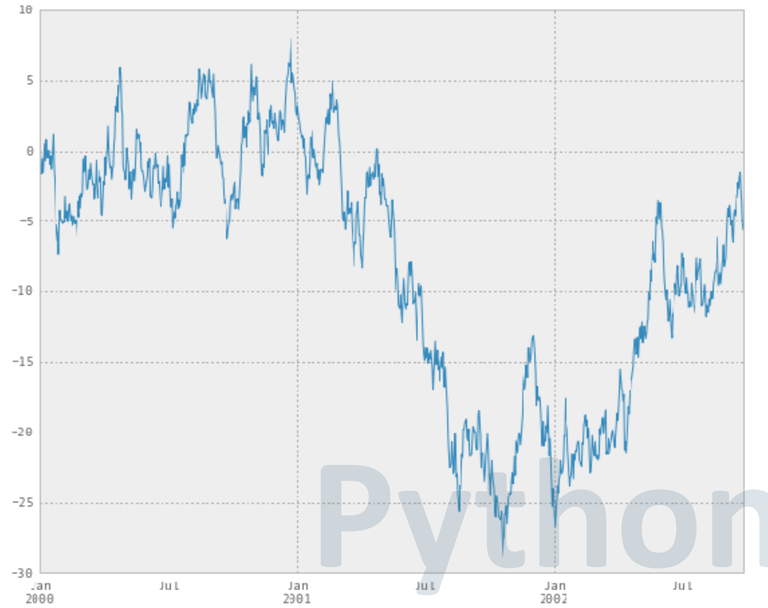
```
ggplot(df2, aes(x, y, fill=cat)) +
  geom_bar(stat="identity", position="dodge",
           color="white", alpha=1/2, width=0.5) +
  scale_color_manual(values=c('#a6cee3','#1f78b4',
                              '#b2df8a','#33a02c')) +
  ggtitle("My chart") +
  labs(x="My Chart", y="Value") +
  theme_bw() +
  theme(legend.position="bottom",
        legend.title = element_blank())
```



R

Time series (line plot)

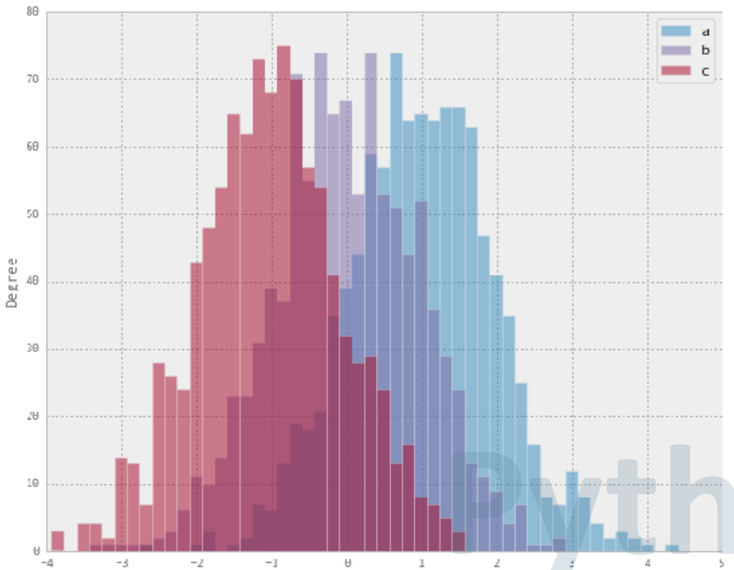
```
ts = pd.Series(np.random.randn( 1000),
     index=pd.date_range('1/1/2000',
     periods=1000))
ts = ts.cumsum()
timeseriesChart = ts.plot( figsize=(10, 8) )
```



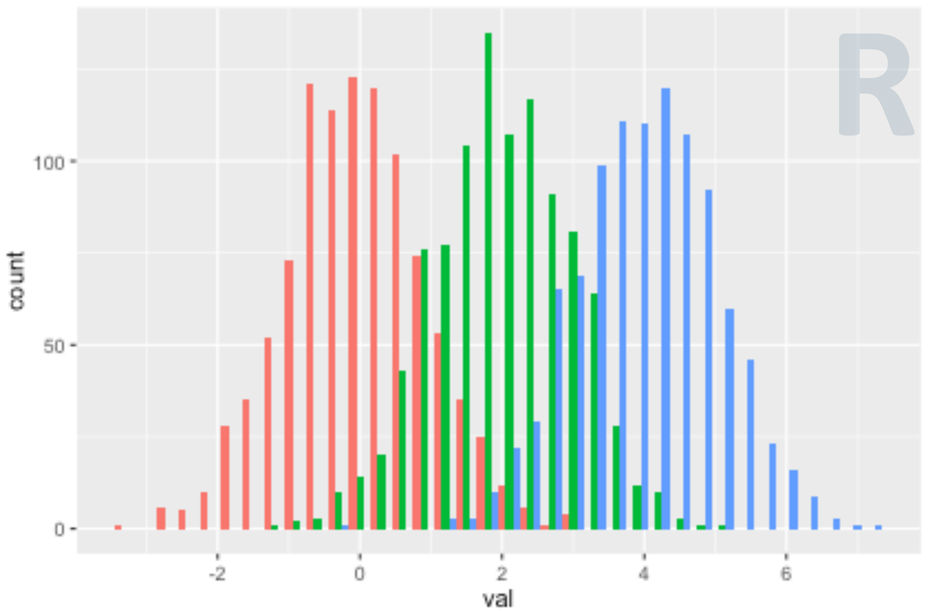

```
toplot <- data.frame(ts=cumsum(runif(1000, -1, 1)),
                     dt=as.Date("2010-01-01") + seq(1000))
ggplot(toplot, aes(x=dt, y=ts)) +
  geom_line(color="steelblue") +
  theme_bw()
```

Histograms

```
df4.plot(kind='hist',
         alpha=0.5,
         bins=50 )
```

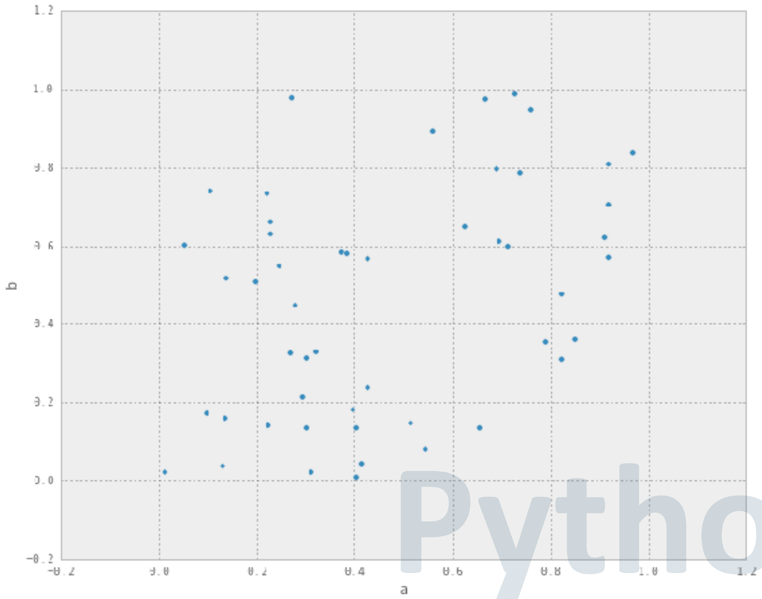



```
ggplot(df4, aes(val, fill=cat)) +
  geom_histogram(position="dodge",
                 binwidth = 0.3)
```



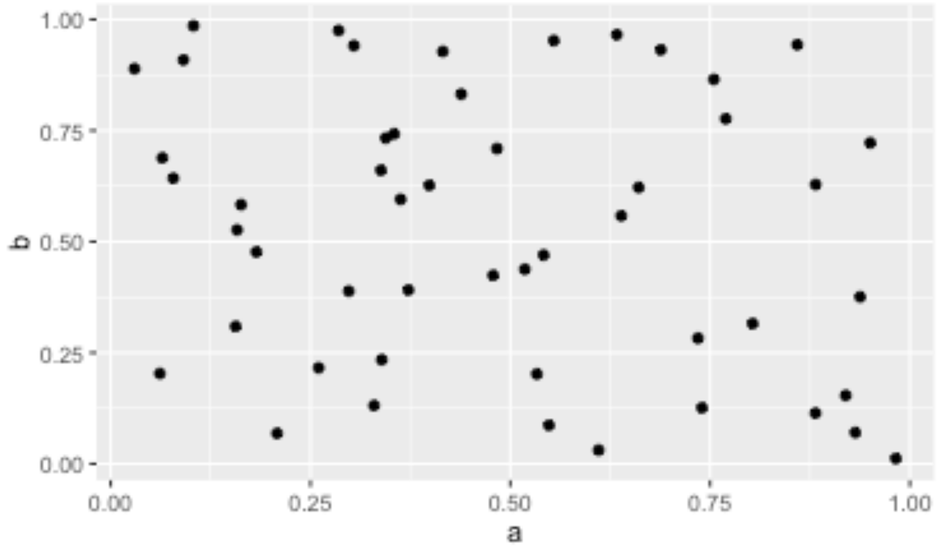
Scatter Plot

```
df5 = pd.DataFrame(np.random.rand(50, 4),
columns=['a', 'b', 'c', 'd'] )
df5.plot(kind='scatter', x='a', y='b' )
```



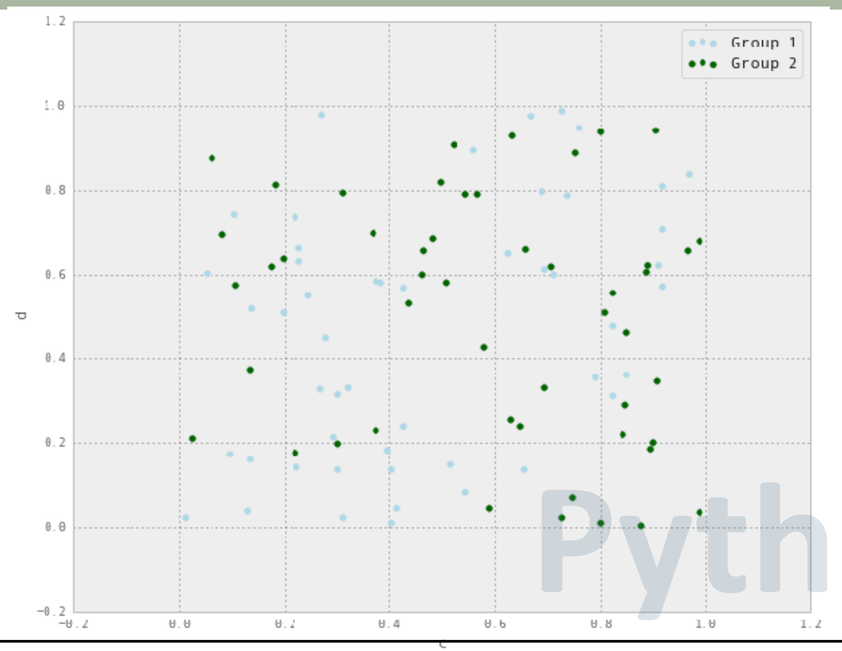
Python

```
df5 <- data.frame(a=runif(50), b=runif(50),
                  c=runif(50), d=runif(50))
ggplot(df5, aes(x=a, y=b)) + geom_point()
```



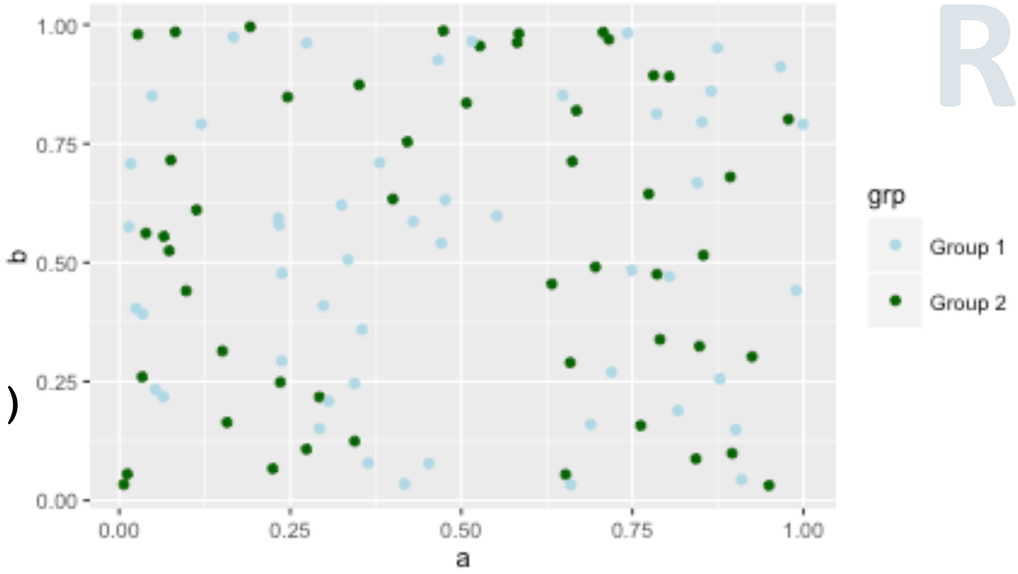
Scatter plot with groups

```
ax = df5.plot(kind='scatter', x='a', y='b',
              color='LightBlue',
              label='Group 1',
              figsize=(10, 8) )

df5.plot(kind='scatter', x='c', y='d',
             color='DarkGreen',
            label='Group 2',
            ax=ax)
```



Python

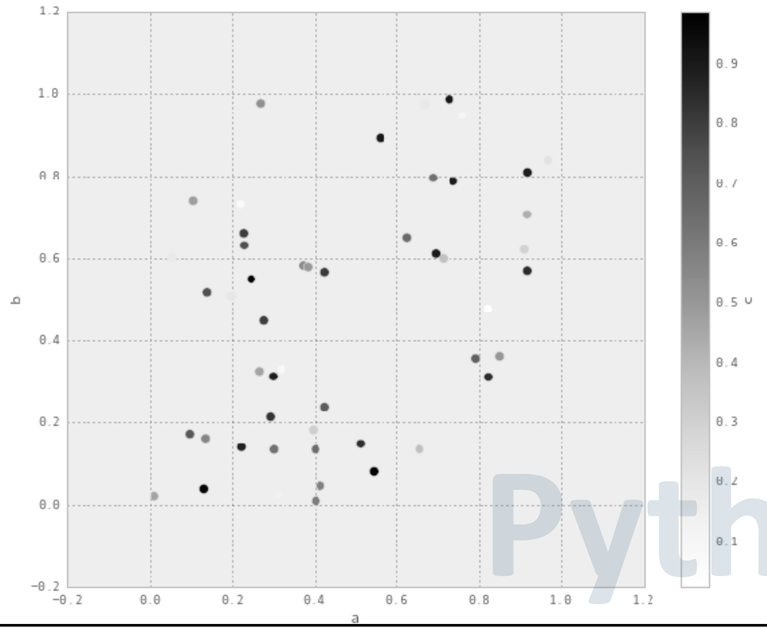
```
df5 <- data.frame(grp=paste("Group", c(1,2)),
                  a=runif(100), b=runif(100))
ggplot(data=df5, aes(x=a, y=b, color=grp)) +
  geom_point() +
  scale_color_manual(values=c("Group 1"="lightblue",
                              "Group 2"="darkgreen"))
```



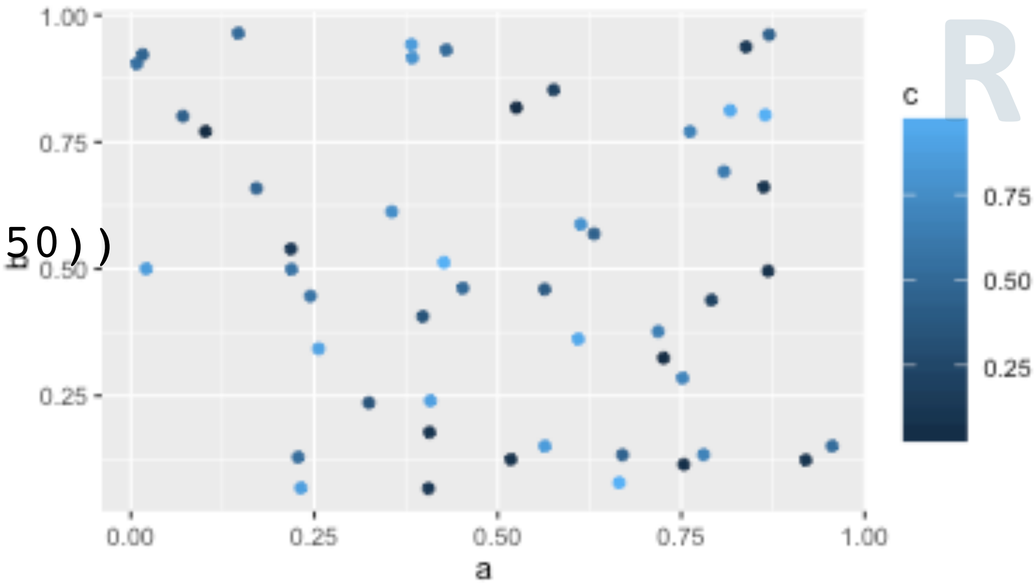
R

Scatter plots with color density representing data

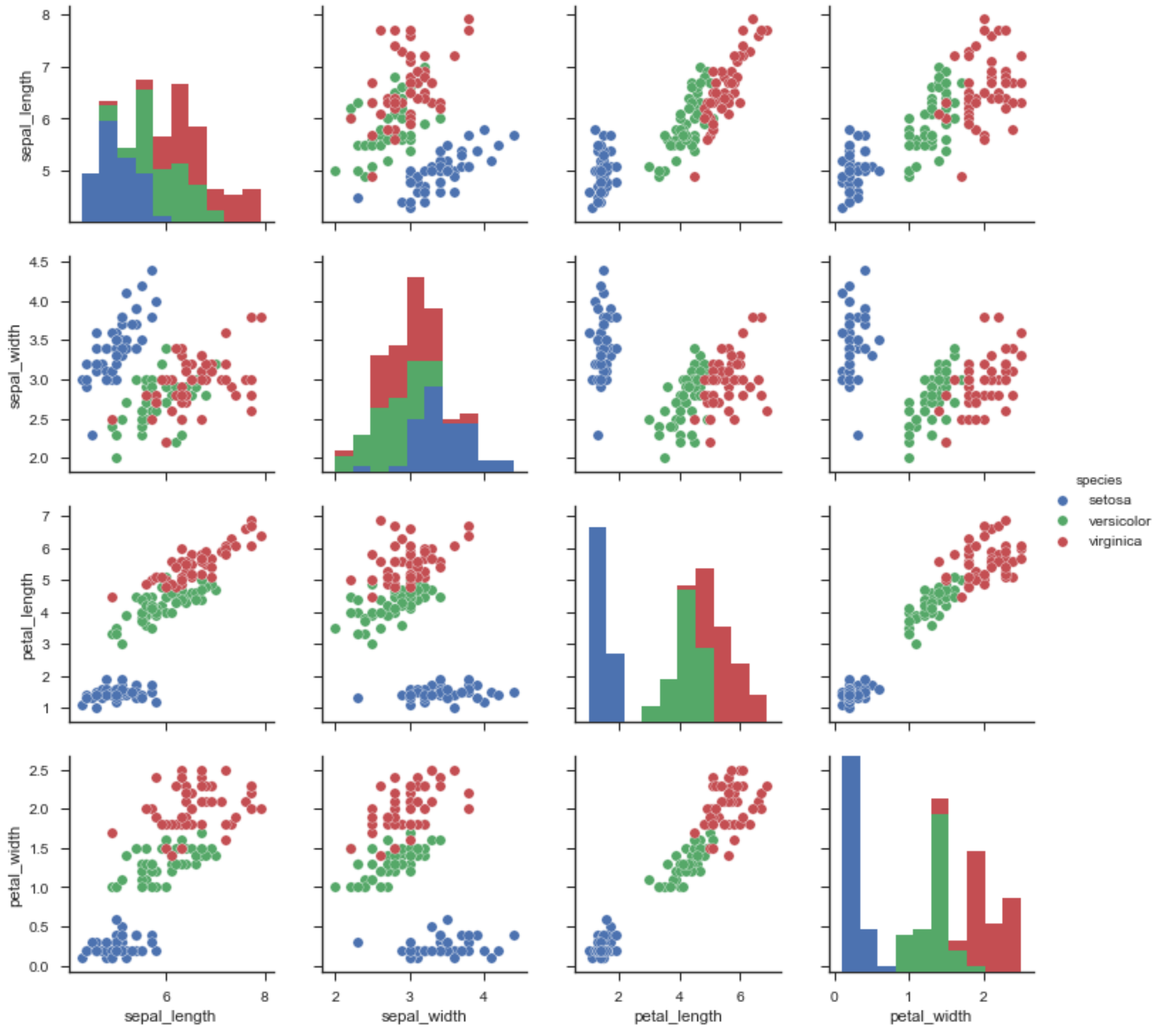
```
df5.plot(kind='scatter', x='a', y='b',
c='c', s=50 )
```




```
df5 <- data.frame(a=runif(50), b=runif(50), c=runif(50))
ggplot(df5, aes(a,b,color=c)) + geom_point()
```



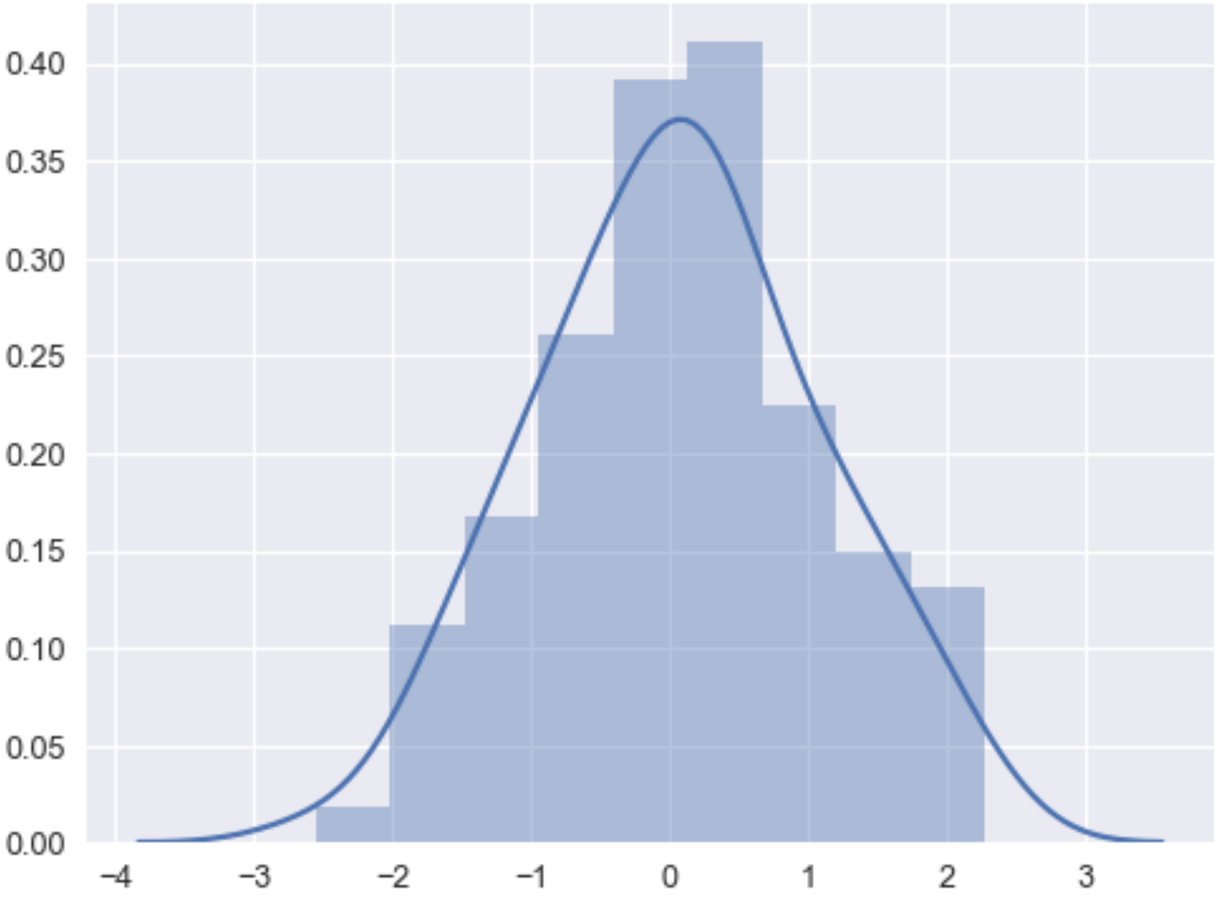
More Advanced Visualizations




```
import seaborn as sns
sns.set(style="ticks")

df = sns.load_dataset("iris")
sns.pairplot(df, hue="species")
```

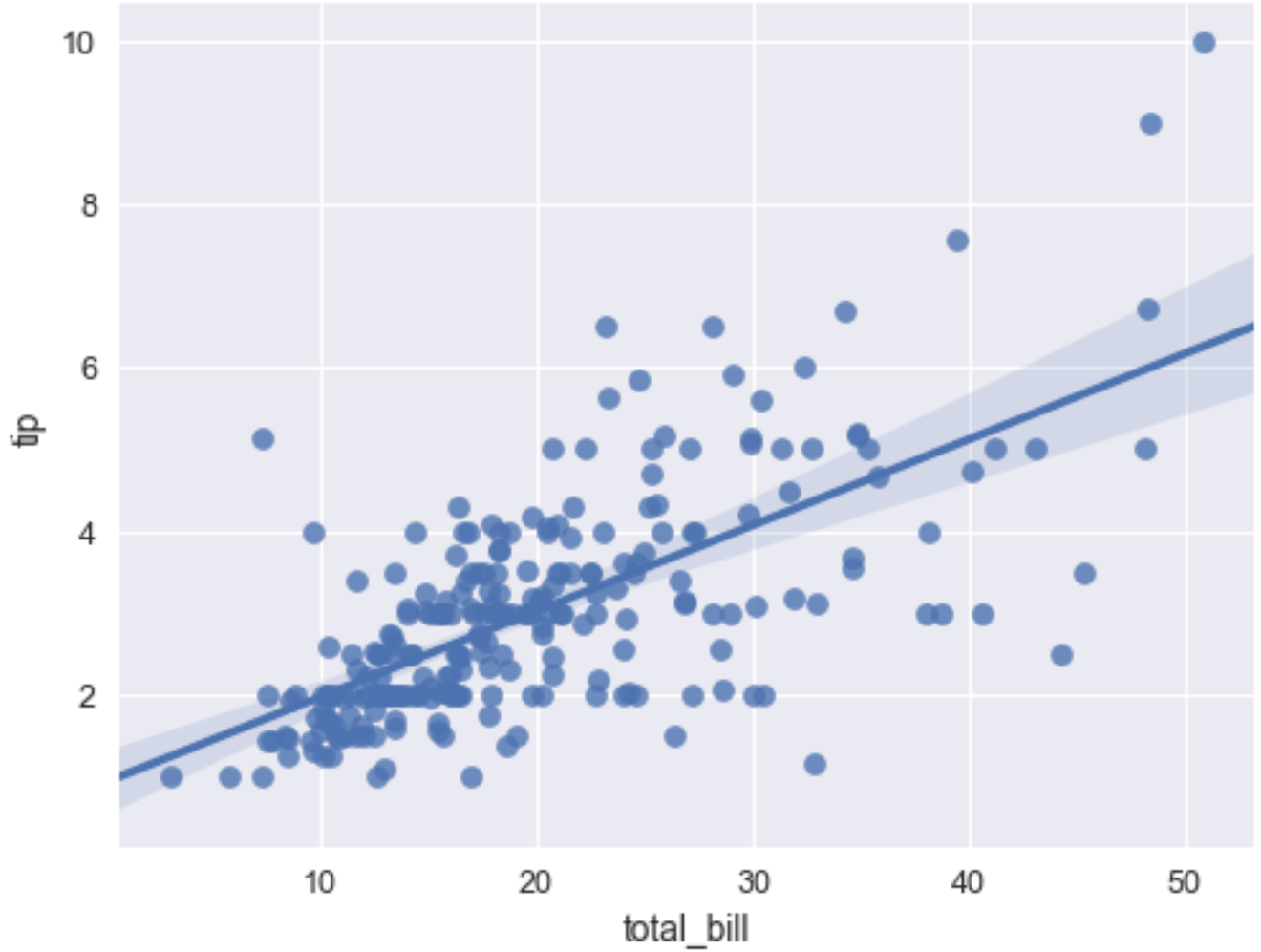
http://seaborn.pydata.org/examples/scatterplot_matrix.html

More Advanced Visualizations



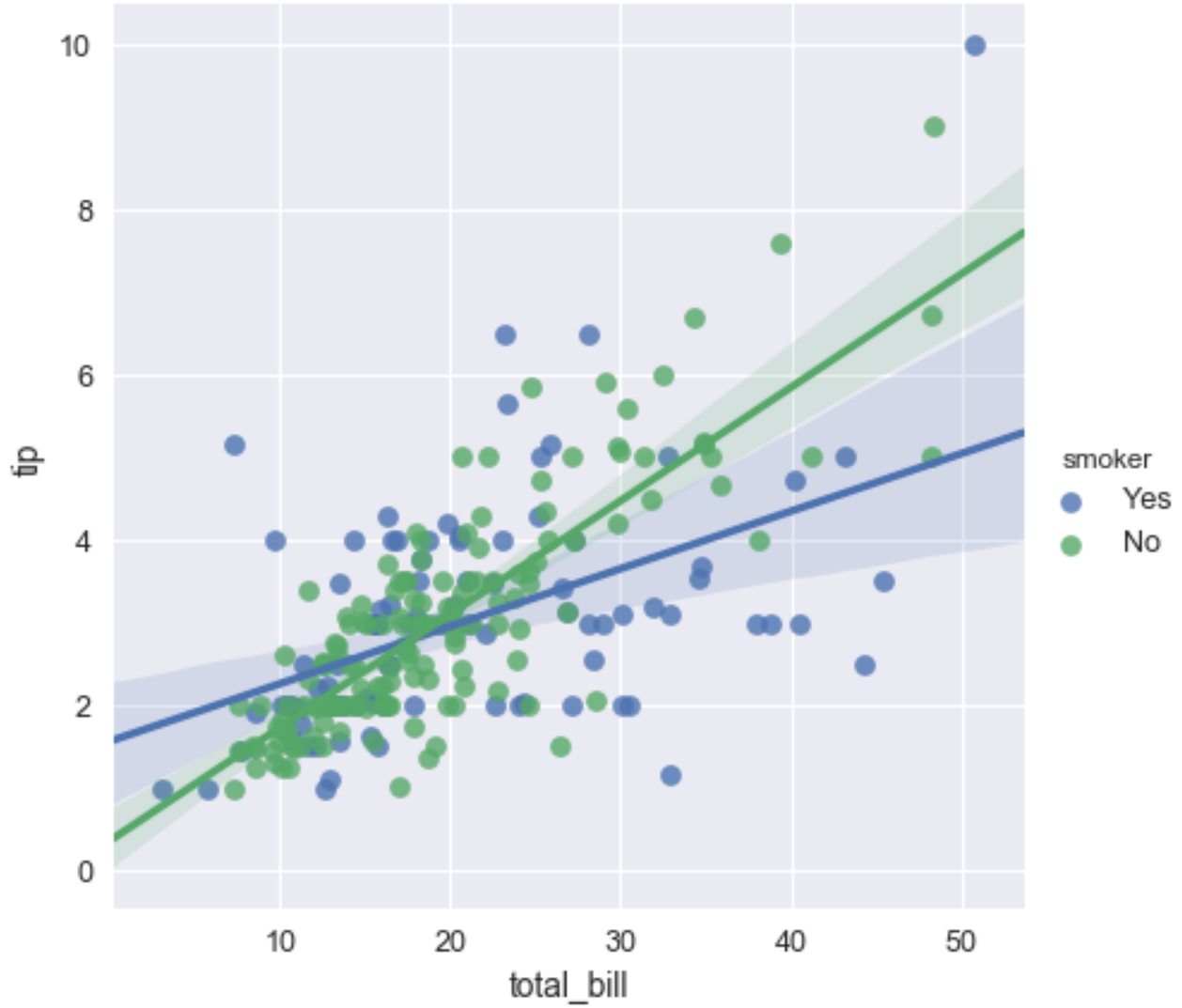

```
ax = sns.distplot(<data>)
```

More Advanced Visualizations



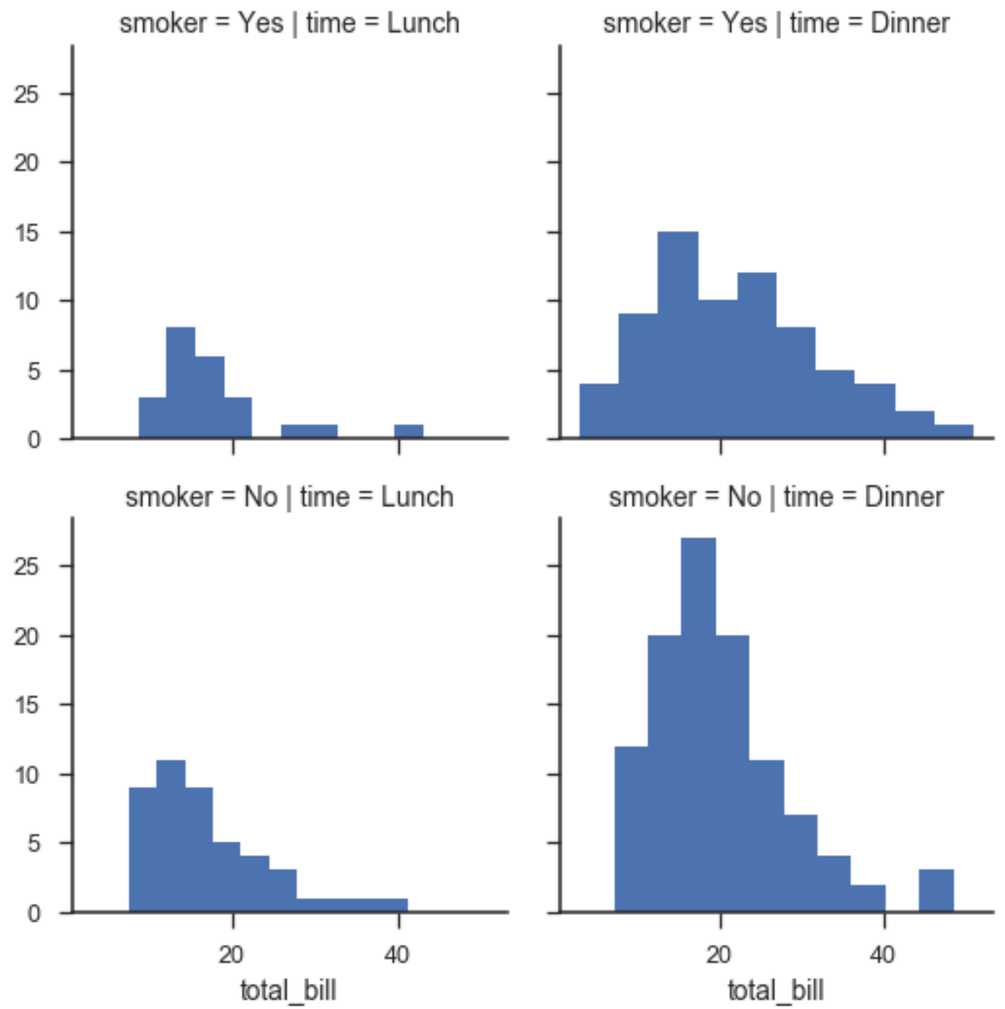

```
ax = sns.regplot(x="total_bill", y="tip", data=tips)
```

More Advanced Visualizations



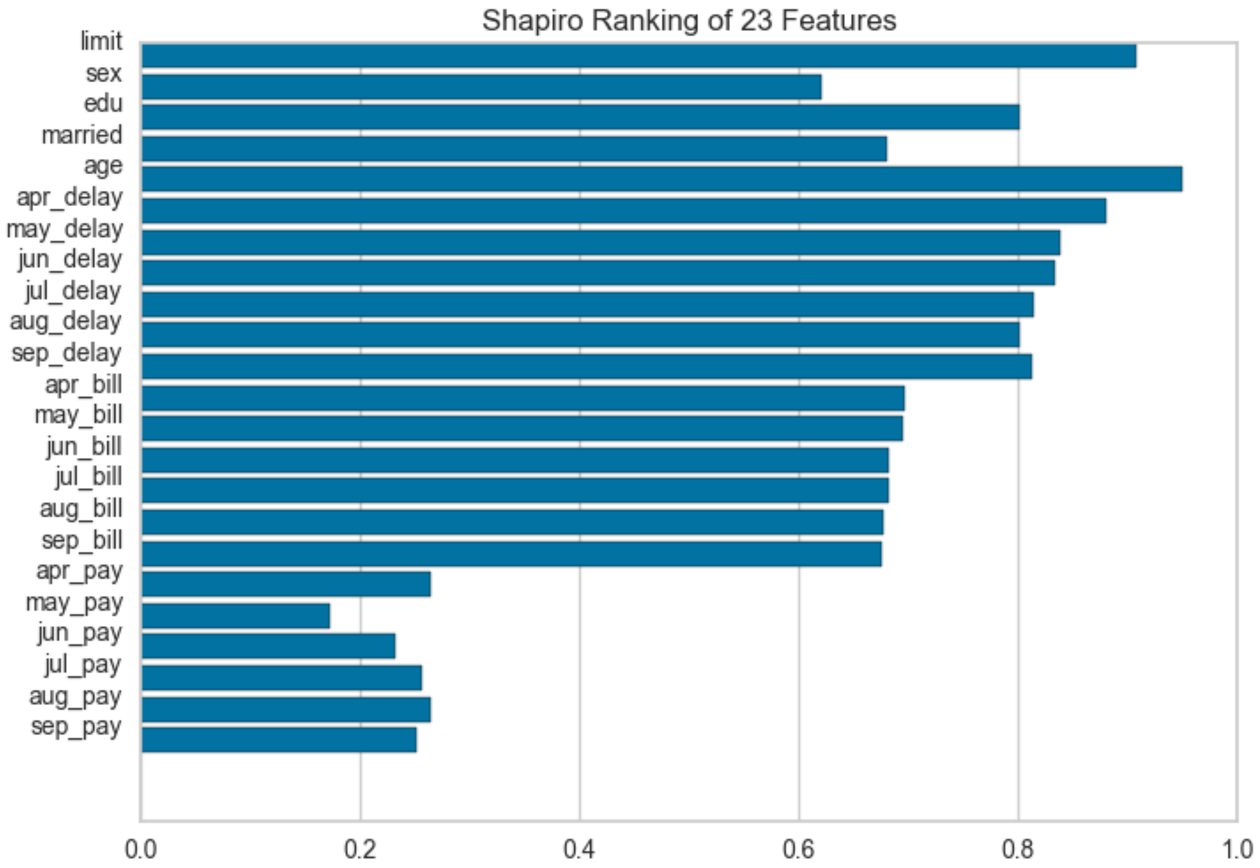

```
g = sns.lmplot(x="total_bill", y="tip", hue="smoker", data=tips)
```

More Advanced Visualizations



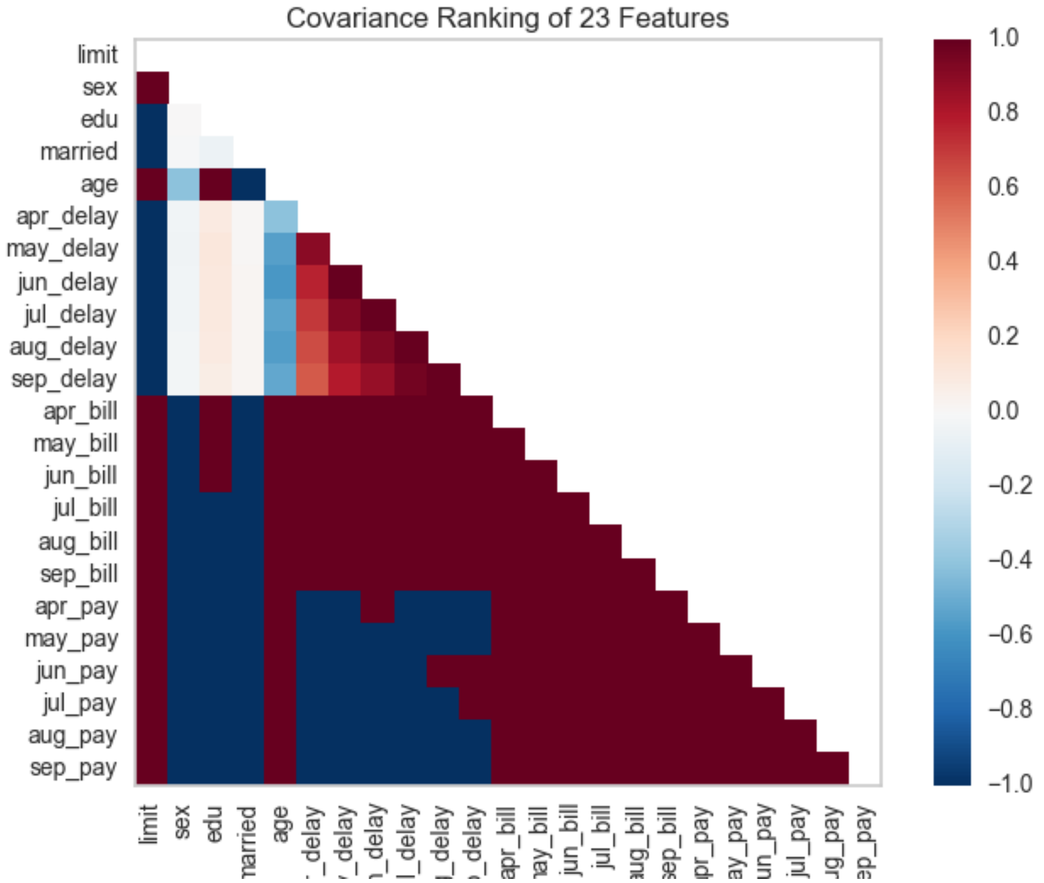

```
g = sns.FacetGrid(tips, col="time",  row="smoker")
        g = g.map(plt.hist, "total_bill")
```

More Advanced Visualizations



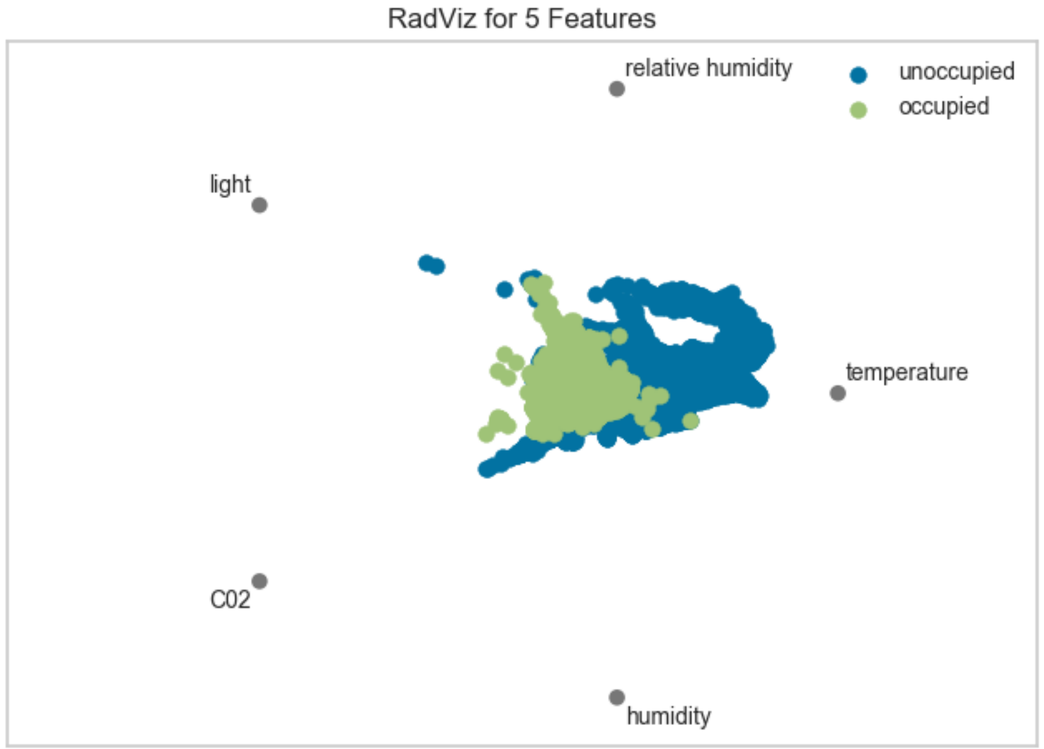

```
visualizer = Rank1D(features=features, algorithm='shapiro')

visualizer.fit(X, y)
visualizer.transform(X)
visualizer.poof()
```

More Advanced Visualizations




```
visualizer = Rank2D(features=features, algorithm='covariance')

visualizer.fit(X, y)
visualizer.transform(X)
visualizer.poof()
```

More Advanced Visualizations




```
from yellowbrick.features import RadViz
visualizer = RadViz(classes=classes, features=features)
visualizer.fit(X, y)
visualizer.transform(X)
visualizer.poof()
```

{LAB} time

Please Complete “Data Visualization Worksheet”

Answers - Step 1

Python

```
data = pd.read_csv('../../Data/dailybots.csv')
```

R

```
bots <- read_csv('../../Data/dailybots.csv')
```

Answers - Step 2 & 3

Python

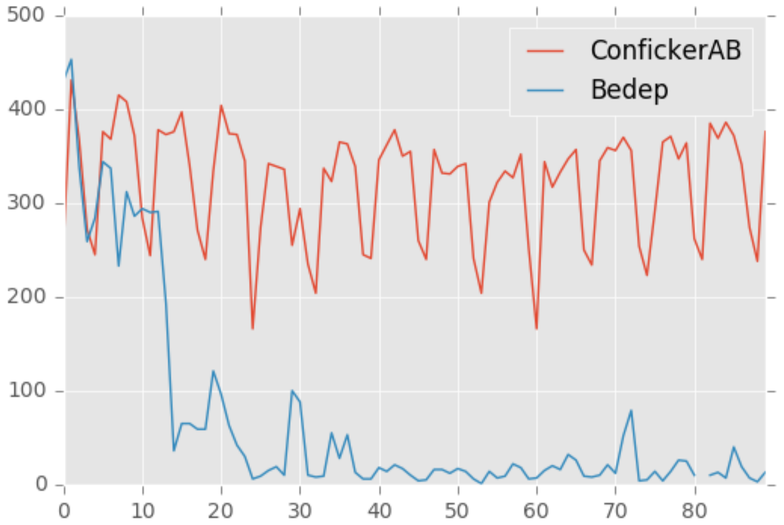
```
filteredData = data[data['industry'] == "Government/Politics"]

filteredData2 = filteredData[filteredData['botfam']== 'ConfickerAB' ][['date','hosts']]
filteredData2.columns = ['date', 'ConfickerAB']
filteredData2.date = pd.to_datetime( filteredData2.date )
filteredData3 = filteredData[filteredData['botfam']== 'Bedep' ][['date','hosts']]
filteredData3.columns = ['date', 'Bedep']
filteredData3.date = pd.to_datetime( filteredData3.date )
finalData = pd.merge(filteredData2, filteredData3, on='date', how='left')
```

R

```
toplot <- filter(bots, industry == "Government/Politics",
         botfam %in% c('ConfickerAB', 'Bedep'))
```

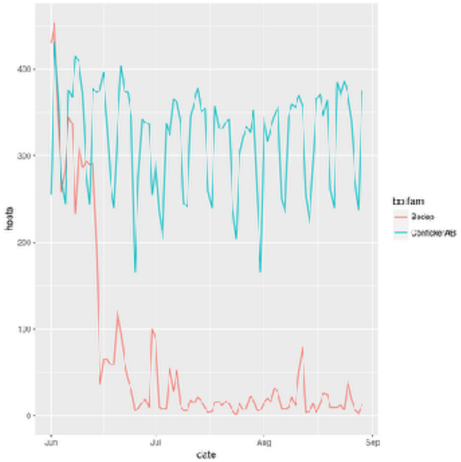
Answers - Step 4

```
finalData.plot(kind='line' )
```



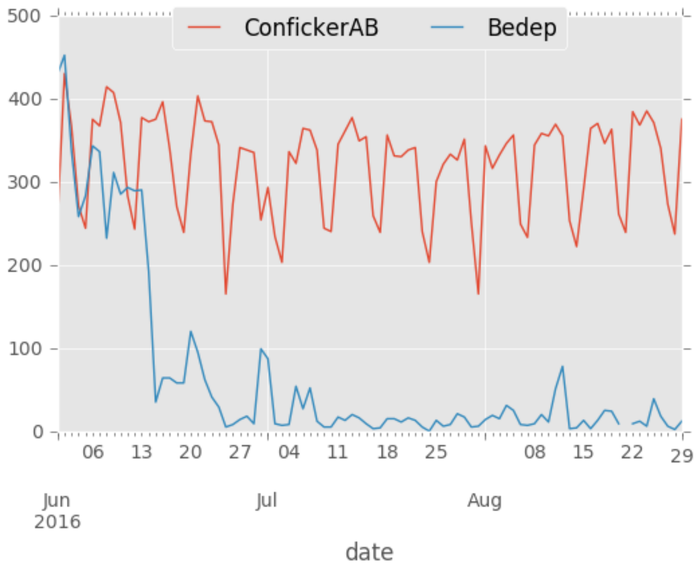
Python

R




```
ggplot(toplot, aes(date, hosts, color=botfam)) + geom_line()
```

Answers - Step 5

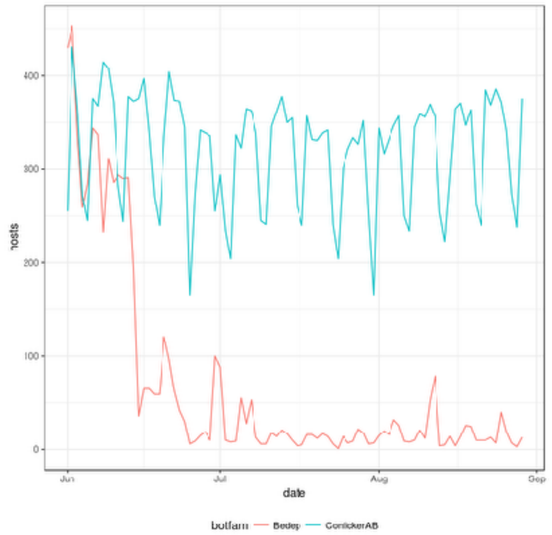



```
nicePlot = finalData.plot( kind="line")
nicePlot.legend(loc='upper center', bbox_to_anchor=(0.5, 1.05),
          ncol=3, fancybox=True, shadow=False)
```

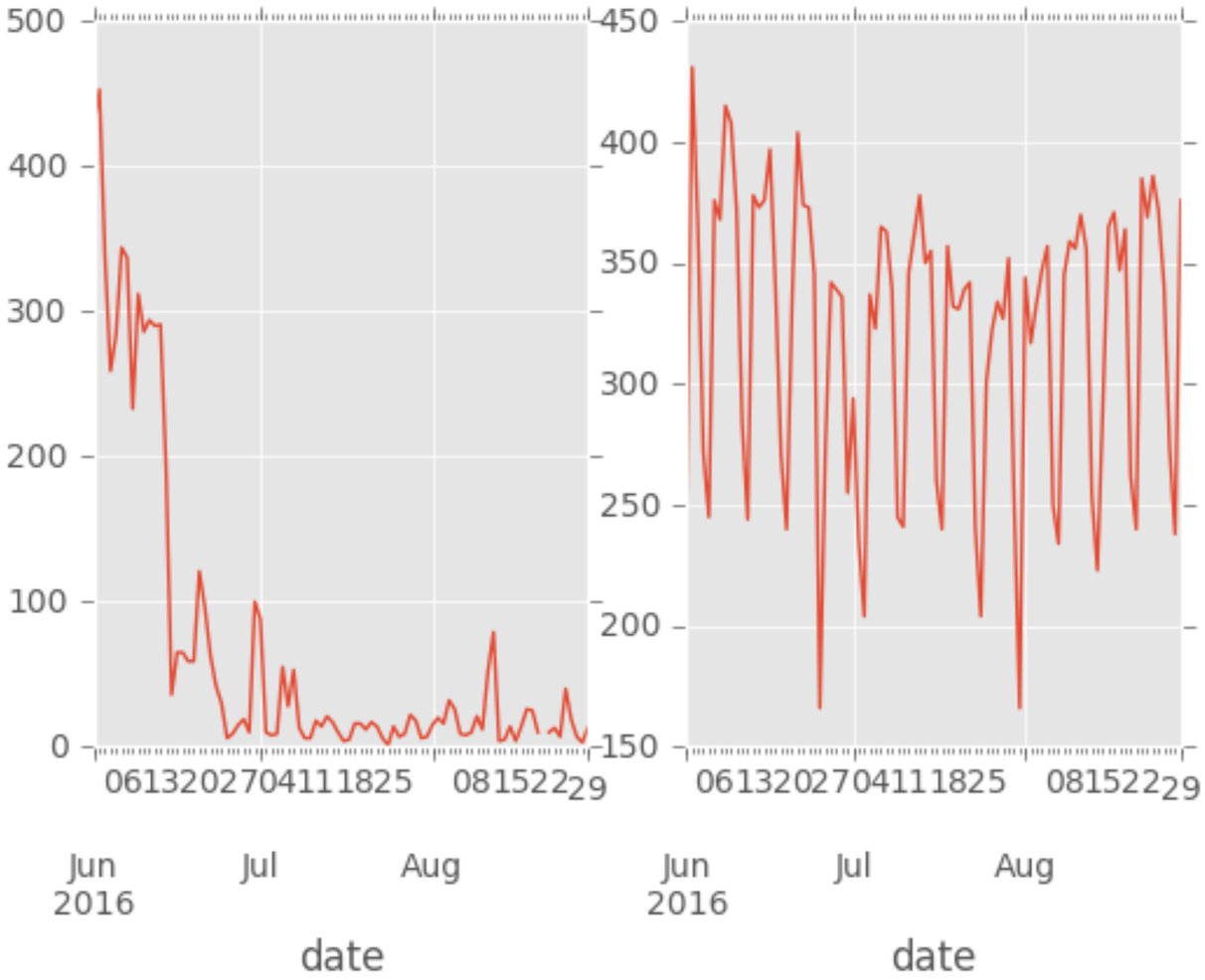
Python

R

```
ggplot(toplot, aes(date, hosts, color=botfam)) + geom_line() +
  theme_bw() +
  theme(legend.position="bottom")
```



Answers - Step 5




```
fig, axes = plt.subplots(nrows=1, ncols=2)

finalData['ConfickerAB'].plot()
finalData['Bedep'].plot(ax=axes[0])
```

Python

R