Introduction

Intro to Data Visualization

Gaston Sanchez

Motivation: Anscombe Dataset

Motivation

Consider some data (four pairs of variables)

```
    x1    y1 x2   y2 x3    y3 x4    y4
1   10  8.04 10 9.14 10  7.46  8  6.58
2    8  6.95  8 8.14  8  6.77  8  5.76
3   13  7.58 13 8.74 13 12.74  8  7.71
4    9  8.81  9 8.77  9  7.11  8  8.84
5   11  8.33 11 9.26 11  7.81  8  8.47
6   14  9.96 14 8.10 14  8.84  8  7.04
7    6  7.24  6 6.13  6  6.08  8  5.25
8    4  4.26  4 3.10  4  5.39 19 12.50
9   12 10.84 12 9.13 12  8.15  8  5.56
10   7  4.82  7 7.26  7  6.42  8  7.91
11   5  5.68  5 4.74  5  5.73  8  6.89
```

What things would you like
to calculate for each variable?

Motivation

```
##        x1             x2             x3             x4
##  Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8
##  1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8
##  Median : 9.0   Median : 9.0   Median : 9.0   Median : 8
##  Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9
##  3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8
##  Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19
```

```
##        y1               y2              y3              y4
##  Min.   : 4.260   Min.   :3.100   Min.   : 5.39   Min.   : 5.250
##  1st Qu.: 6.315   1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
##  Median : 7.580   Median :8.140   Median : 7.11   Median : 7.040
##  Mean   : 7.501   Mean   :7.501   Mean   : 7.50   Mean   : 7.501
##  3rd Qu.: 8.570   3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
##  Max.   :10.840   Max.   :9.260   Max.   :12.74   Max.   :12.500
```

What things would you like to calculate for each pair of variables (e.g. `x1, y1`)?

Motivation

```
cor(anscombe$x1, anscombe$y1)

## [1] 0.8164205

cor(anscombe$x2, anscombe$y2)

## [1] 0.8162365

cor(anscombe$x3, anscombe$y3)

## [1] 0.8162867

cor(anscombe$x4, anscombe$y4)

## [1] 0.8165214
```

Motivation

- Mean of $\mathtt{x}$ values $= 9$
- Mean of $\mathtt{y}$ values $= 7.5009091$
- least squares equation: $y = 3 + 0.5x$
- Sum of squared errors: 110
- Correlation coefficient: 0.8164205

Data Visualization

Using only numerical reduction methods in data analyses is far too limiting

Why Graphics?

Are you able to see any patterns, associations, relations?

```
##      x1    y1 x2   y2 x3    y3 x4    y4
## 1    10  8.04 10 9.14 10  7.46  8  6.58
## 2     8  6.95  8 8.14  8  6.77  8  5.76
## 3    13  7.58 13 8.74 13 12.74  8  7.71
## 4     9  8.81  9 8.77  9  7.11  8  8.84
## 5    11  8.33 11 9.26 11  7.81  8  8.47
## 6    14  9.96 14 8.10 14  8.84  8  7.04
## 7     6  7.24  6 6.13  6  6.08  8  5.25
## 8     4  4.26  4 3.10  4  5.39 19 12.50
## 9    12 10.84 12 9.13 12  8.15  8  5.56
## 10    7  4.82  7 7.26  7  6.42  8  7.91
## 11    5  5.68  5 4.74  5  5.73  8  6.89
```

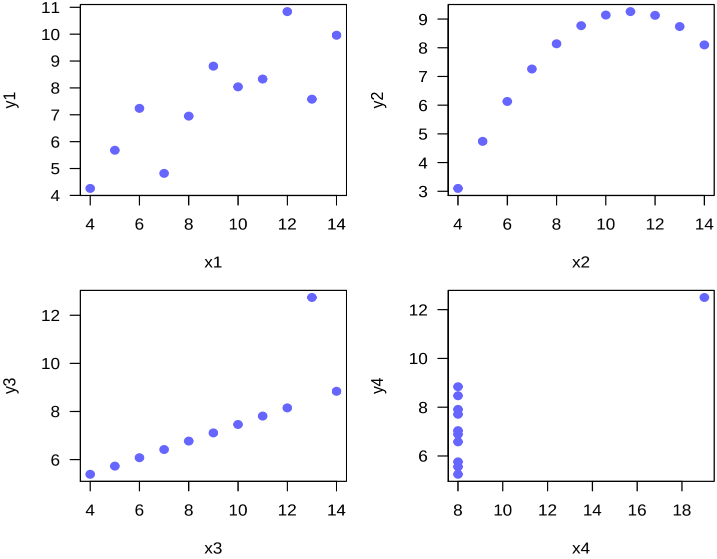
Famous dataset `"anscombe"` (four data sets)

Why Graphics?

How are these two variables associated?

What does these data values look like?

```
    x1    y1
1   10  8.04
2    8  6.95
3   13  7.58
4    9  8.81
5   11  8.33
6   14  9.96
7    6  7.24
8    4  4.26
9   12 10.84
10   7  4.82
11   5  5.68
```


y1
11
10
9
8
7
6
5
4
4
6
8
10
12
14
x1
y2
9
8
7
6
5
4
3
4
6
8
10
12
14
x2
y3
12
10
8
6
4
6
8
10
12
14
x3
y4
12
10
8
6
8
10
12
14
16
18
x4

Visualization

Visualize

Visualize

- To form a mental image of
- To make visible

Visualization

Process of representing information or ideas by diagrams or graphs.
Ross Ihaka

Visualization

To convey information through visual representations

What is visualization?

Definition by OED

The action or fact of visualizing; the power or process of forming a mental picture or vision of something not actually present to the sight

What is visualization?

Definitions

- The action or process of rendering visible
- Transformation of the symbolic into the geometric McCormick et al 1987
- The use of computer-generated, possibly interactive visual representations of data to amplify cognition Card, Mackinlay, & Shneiderman 1999

What is visualization?

Visualization

Often referred to as the process of making a graphic or an image. Actually it is a cognitive process

Part of our language

- “I see what you are saying”
- “Seeing is believing”
- “A picture is worth a thousand numbers”

Vision

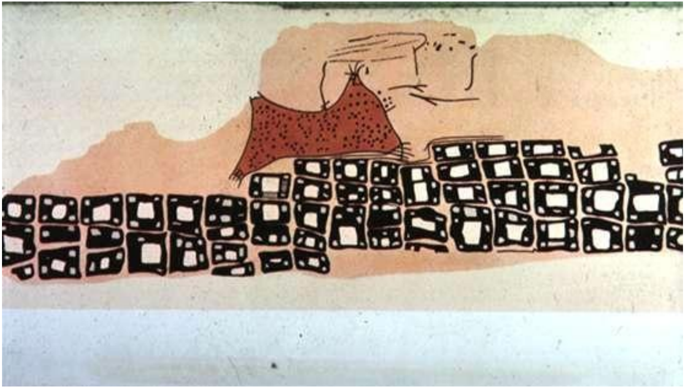
Vision, of our all senses, is the most powerful and efficient channel for receiving information from the physical world.

Why do we create visualizations?

Why do we create visualizations?

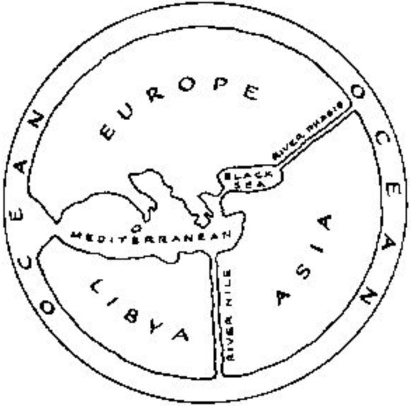
- Map
- Record
- Abstract
- Discover
- Clarify
- Interact
- Communicate
- Entertain

Maps



Konya town map, Turkey (c. 6200 BC)

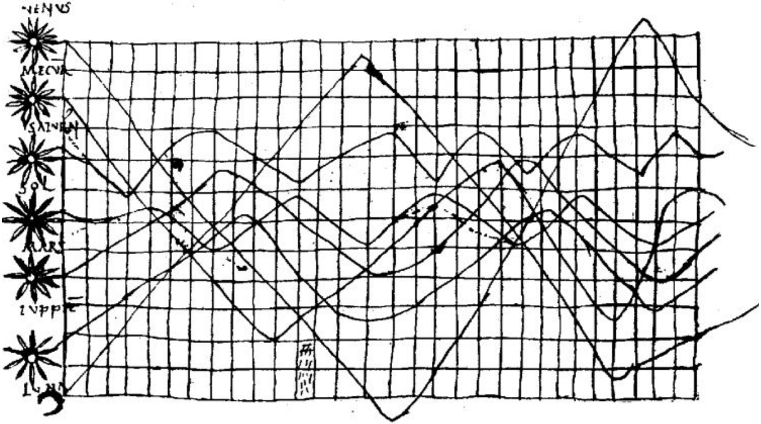
Maps



Anaximader's Map of the World

Anaximander of Miletus (c. 550 BC)

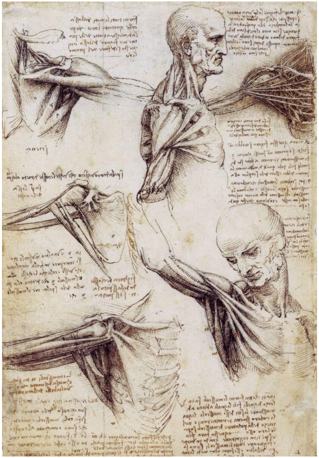
Maps



Planetary Movements (source: wikimedia)

Record

Leonardo Da Vinci (ca. 1500)

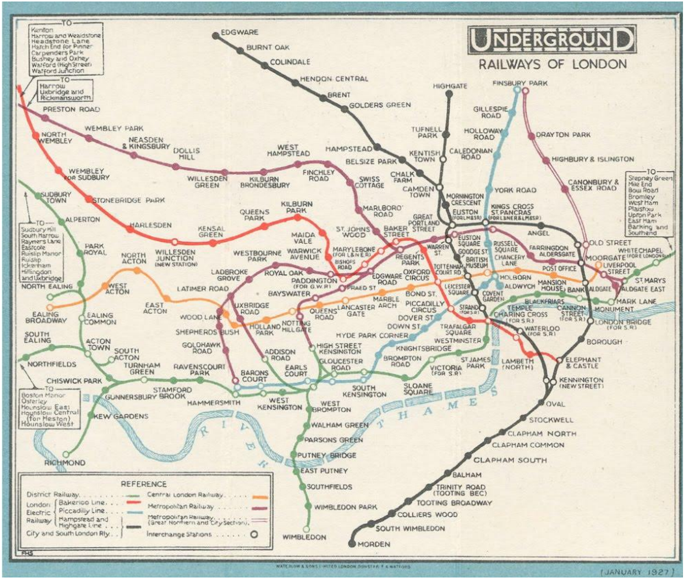


Leonardo Da Vinci (ca. 1500)

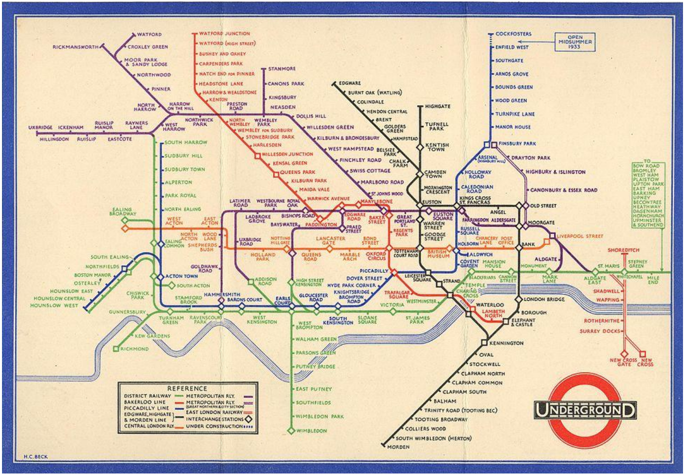
Record

William Curtis (1746-1799)

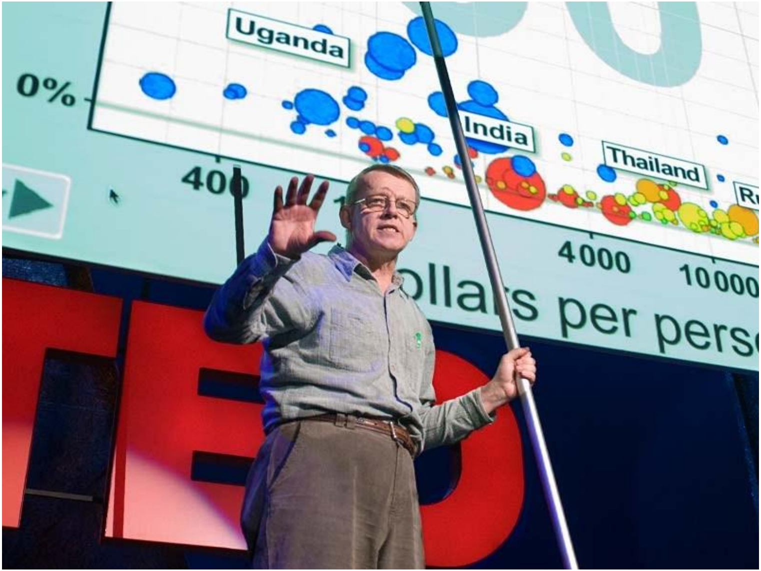
Clarify: Stingemore’s London Underground (1927)



Clarify: Harry Beck's London Underground (1933)



Communicate: Hans Rosling



Main functions of visualizations

- **Record**: store information
  - photographs, blueprints, sketches, diagrams
- **Analyze**: support reasoning about information
  - process and calculate
  - reason about data
  - feedback and interaction
- **Communication**: convery information to others
  - share and persuade
  - collaborate and revise
  - emphasize important aspects of data

based on J. Heer

Data Visualization

Cereals Data Set

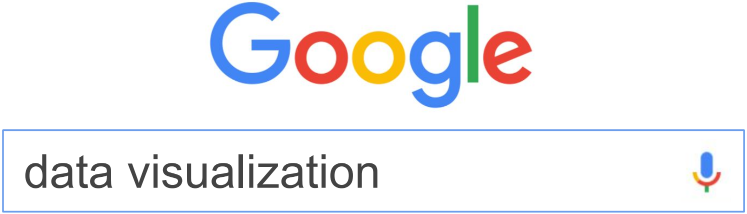
```
                 Cups Calories Carbs Fat Fiber Potassium Protein Sodium Sugars
CapnCrunch       0.75      120  12.0   2   0.0        35       1    220     12
CocoaPuffs       1.00      110  12.0   1   0.0        55       1    180     13
Trix             1.00      110  13.0   1   0.0        25       1    140     12
AppleJacks       1.00      110  11.0   0   1.0        30       2    125     14
CornChex         1.00      110  22.0   0   0.0        25       2    280      3
CornFlakes       1.00      100  21.0   0   1.0        35       2    290      2
Nut&Honey        0.67      120  15.0   1   0.0        40       2    190      9
Smacks           0.75      110   9.0   1   1.0        40       2     70     15
MultiGrain       1.00      100  15.0   1   2.0        90       2    220      6
CracklinOat      0.50      110  10.0   3   4.0       160       3    140      7
GrapeNuts        0.25      110  17.0   0   3.0        90       3    179      3
HoneyNutCheerios 0.75      110  11.5   1   1.5        90       3    250     10
NutriGrain       0.67      140  21.0   2   3.0       130       3    220      7
Product19        1.00      100  20.0   0   1.0        45       3    320      3
TotalRaisinBran  1.00      140  15.0   1   4.0       230       3    190     14
WheatChex        0.67      100  17.0   1   3.0       115       3    230      3
Oatmeal          0.50      130  13.5   2   1.5       120       3    170     10
Life             0.67      100  12.0   2   2.0        95       4    150      6
Maypo            1.00      100  16.0   1   0.0        95       4      0      3
QuakerOats       0.50      100  14.0   1   2.0       110       4    135      6
Muesli           1.00      150  16.0   3   3.0       170       4    150     11
Cheerios         1.25      110  17.0   2   2.0       105       6    290      1
SpecialK         1.00      110  16.0   0   1.0        55       6    230      3
```

Some questions

- Which cereal has the most/lest potassium?
- Is there a relationship between potassium and fiber? If so, are there any outliers?
- Which is the “healthiest” cereal?

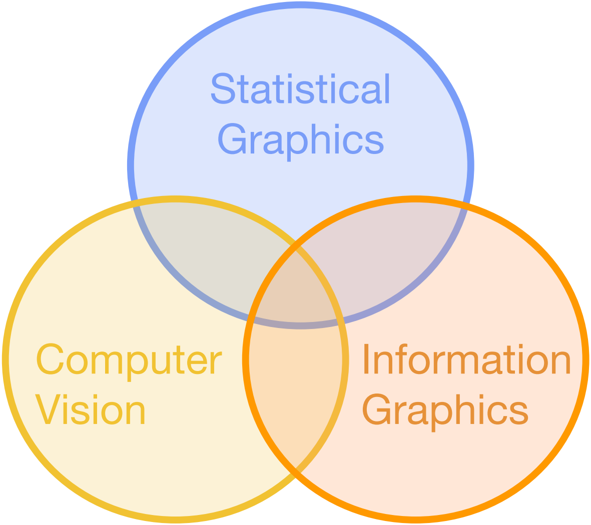
Data Visualization

A key component of computing with data consists of **Data Visualization**



data visualization

Data Visualization



Data Visualization

> *"Data visualization is an umbrella term to cover all types of visual representations that support the exploration, examination, and communication of data."*

Stephen Few

Why data visualizations?

- see overall patterns and detailed behavior
- reveal patterns
- identify trends
- identify exceptions and outliers
- summarize information

Data Visualization

Data Visualization

- Statistical Graphics?
- Computer Graphics?
- Computer Vision?
- Infographics?
- Data Art?

Data Visualization

We'll focus on statistical graphics and visual displays of data in science and technology