Computational Analysis of Big Data

Week 2

A Data Scientist's most fundamental tools

Data visualization

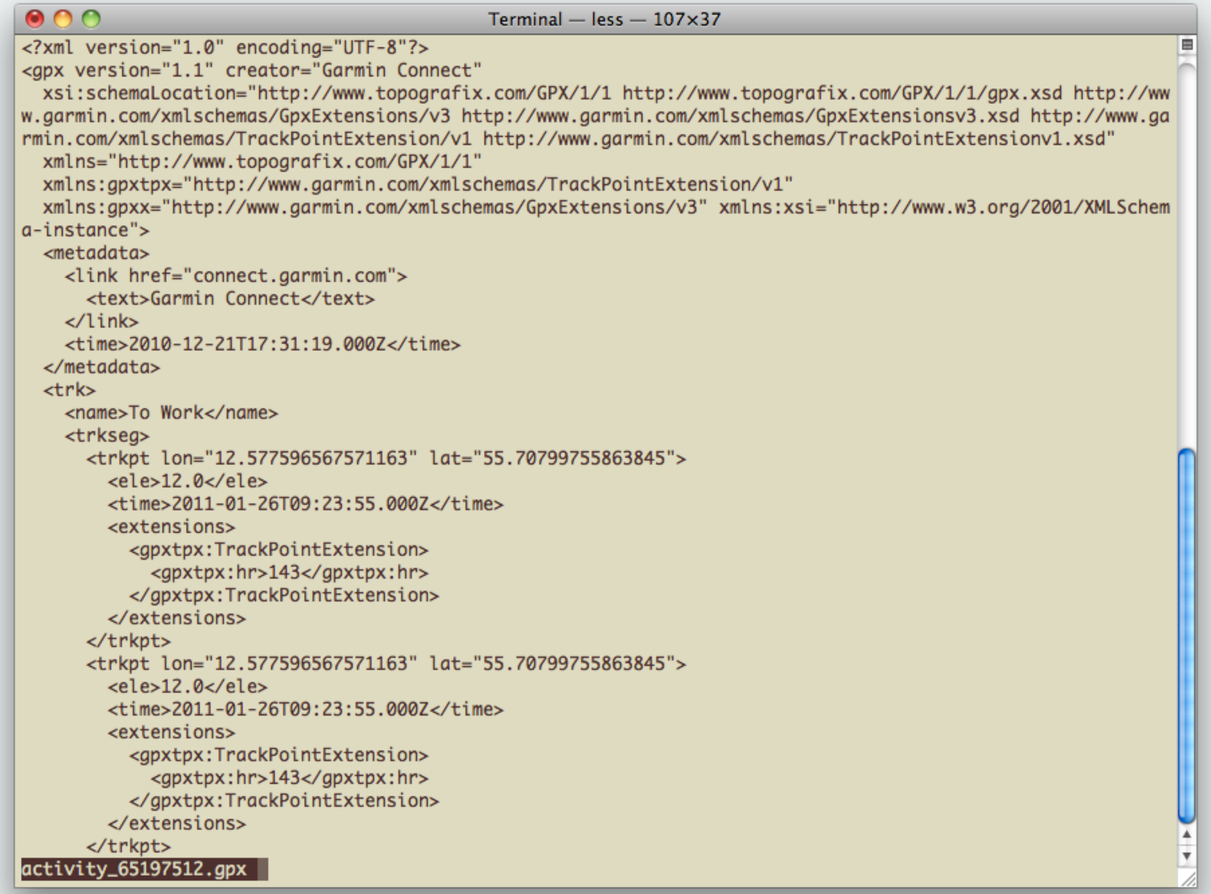
1
1000

1
1000

.

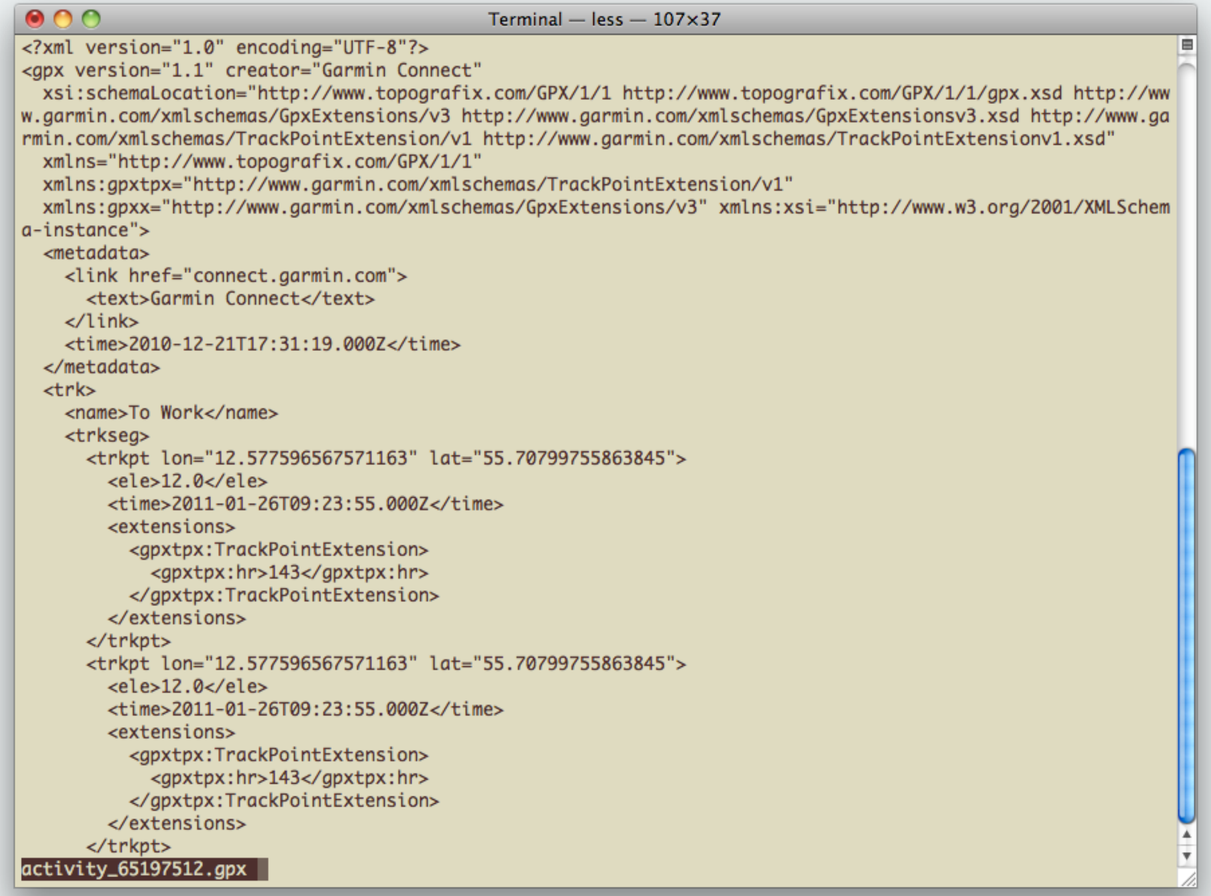
This is data

It's usually some (large) file full of text and numbers



This is location data

It's usually some (large) file full of text and numbers



And if you're lucky there is also some kind of **<markup>**

Most raw data is incomprehensible to humans

We have:

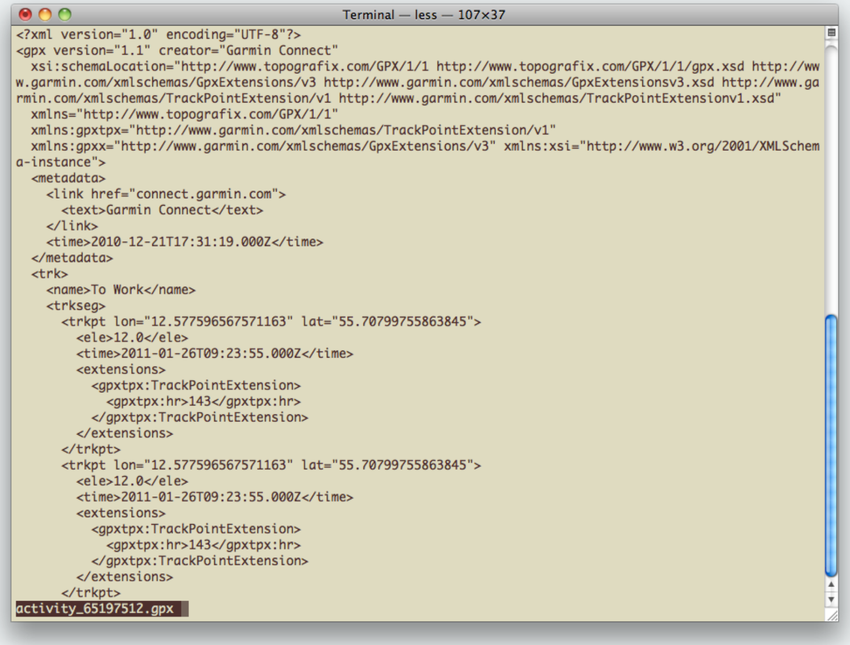
- Narrow spectrum of data that we can process and understand
- Limited memory for processing new information
- Limited attention for undertaking focussed tasks

The human eye is made for advanced pattern recognition

It can:

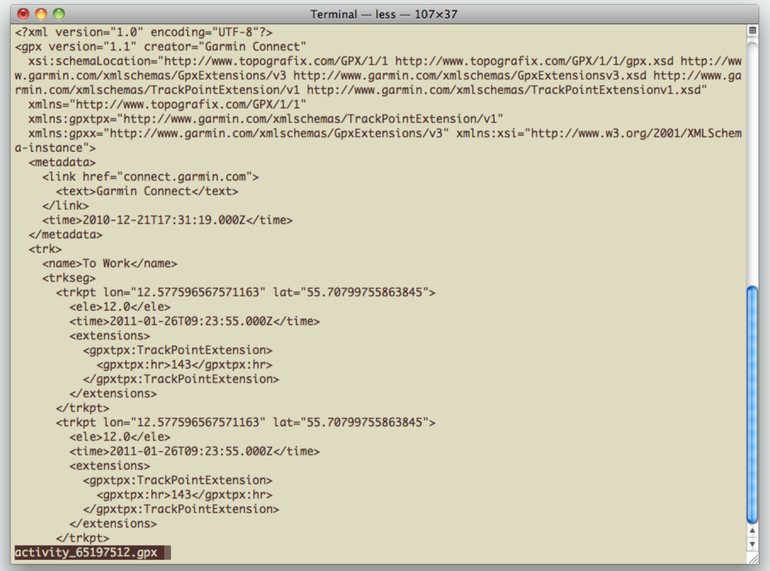
- Immediately recognize a pattern in a highly complex image
- Quickly spot things that deviate from patterns (outlier detection)
- Process streams of images and recognize patterns over time

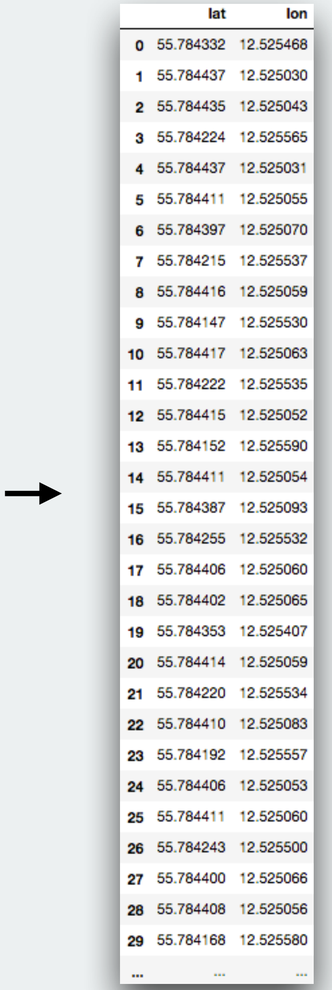
Data must be rendered in human-friendly format



?

Data must be rendered in human-friendly format





	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	12.525590
14	55.784411	12.525054
15	55.784387	12.525093
16	55.784255	12.525532
17	55.784406	12.525060
18	55.784402	12.525065
19	55.784353	12.525407
20	55.784414	12.525059
21	55.784220	12.525534
22	55.784410	12.525083
23	55.784192	12.525557
24	55.784406	12.525053
25	55.784411	12.525060
26	55.784243	12.525500
27	55.784400	12.525066
28	55.784408	12.525056
29	55.784168	12.525580
...	...	...

?

Data must be rendered in human-friendly format

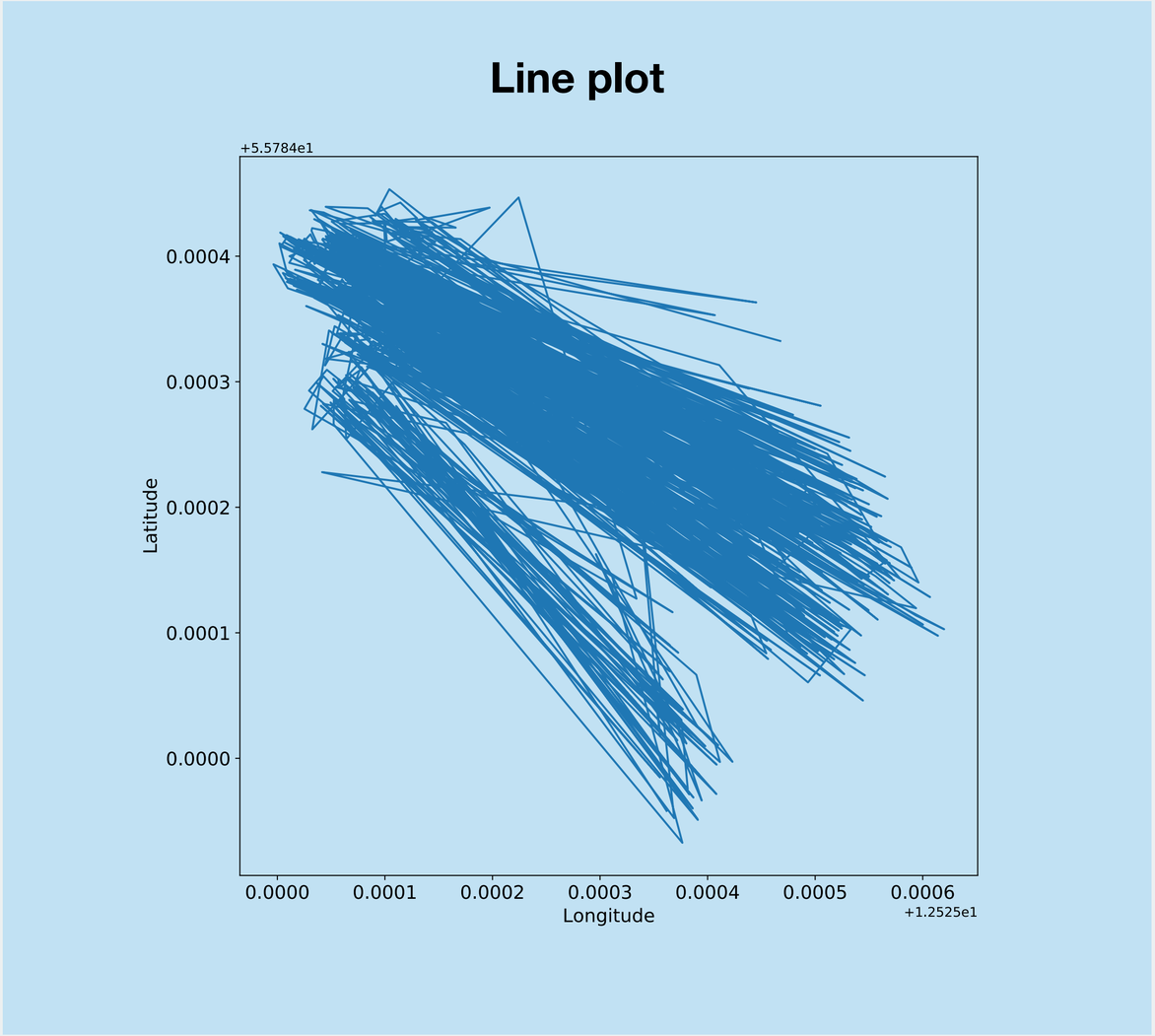
Terminal — less — 107×37

```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://www.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.garmin.com/xmlschemas/TrackPointExtension/v1 http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd"
  xmlns="http://www.topografix.com/GPX/1/1"
  xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
  xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <metadata>
    <link href="connect.garmin.com">
      <text>Garmin Connect</text>
    </link>
    <time>2010-12-21T17:31:19.000Z</time>
  </metadata>
  <trk>
    <name>To Work</name>
    <trkseg>
      <trkpt lon="12.577596567571163" lat="55.70799755863845">
        <ele>12.0</ele>
        <time>2011-01-26T09:23:55.000Z</time>
        <extensions>
          <gpxtpx:TrackPointExtension>
            <gpxtpx:hr>143</gpxtpx:hr>
          </gpxtpx:TrackPointExtension>
        </extensions>
      </trkpt>
      <trkpt lon="12.577596567571163" lat="55.70799755863845">
        <ele>12.0</ele>
        <time>2011-01-26T09:23:55.000Z</time>
        <extensions>
          <gpxtpx:TrackPointExtension>
            <gpxtpx:hr>143</gpxtpx:hr>
          </gpxtpx:TrackPointExtension>
        </extensions>
      </trkpt>
activity_65197512.gpx
```

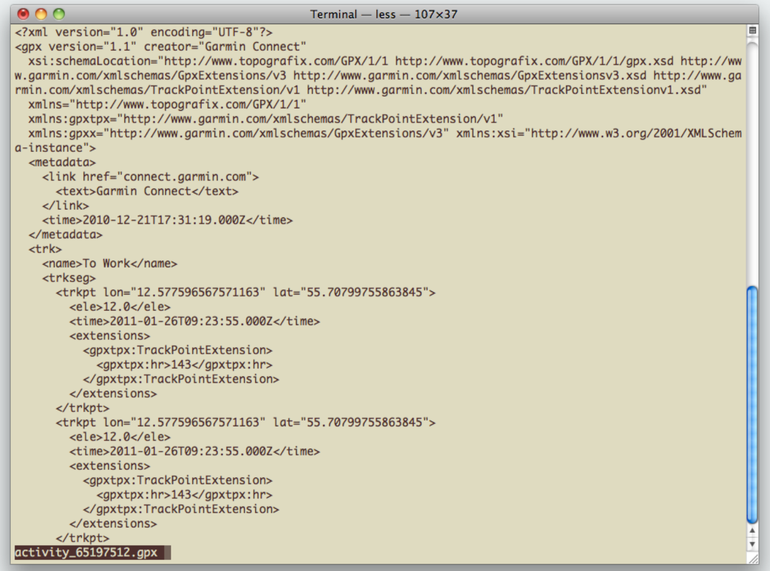
→

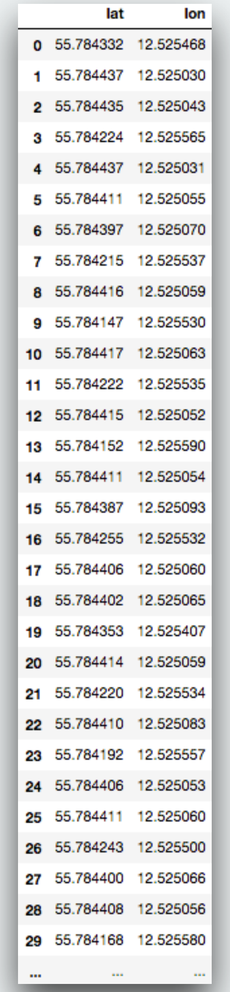
	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	12.525590
14	55.784411	12.525054
15	55.784387	12.525093
16	55.784255	12.525532
17	55.784406	12.525060
18	55.784402	12.525065
19	55.784353	12.525407
20	55.784414	12.525059
21	55.784220	12.525534
22	55.784410	12.525083
23	55.784192	12.525557
24	55.784406	12.525053
25	55.784411	12.525060
26	55.784243	12.525500
27	55.784400	12.525066
28	55.784408	12.525056
29	55.784168	12.525580
...	...	...

→

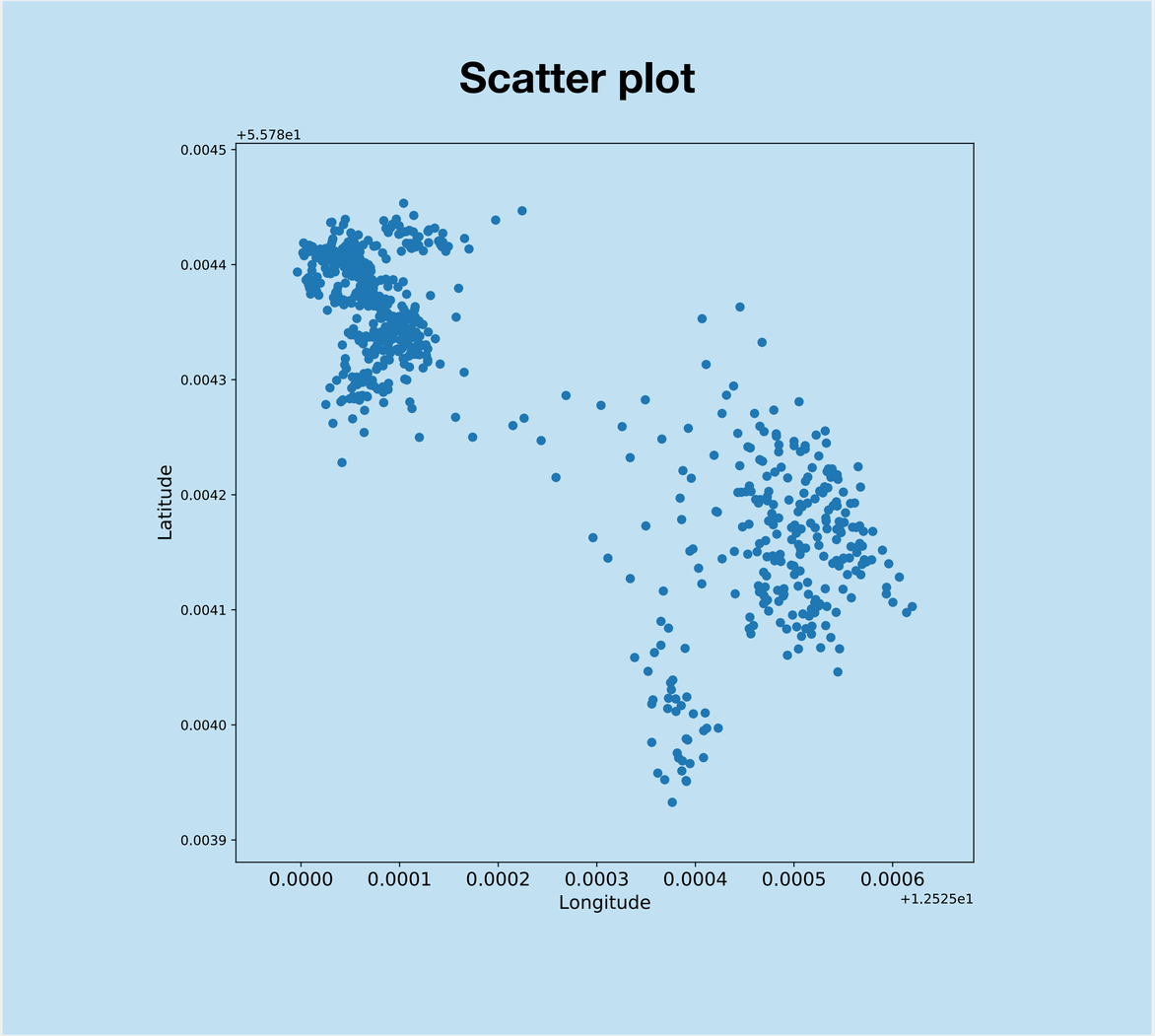


Data must be rendered in human-friendly format

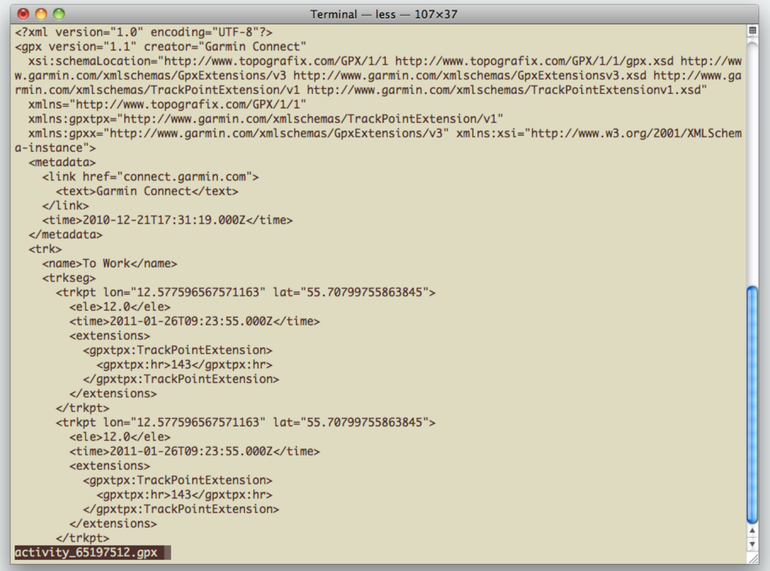


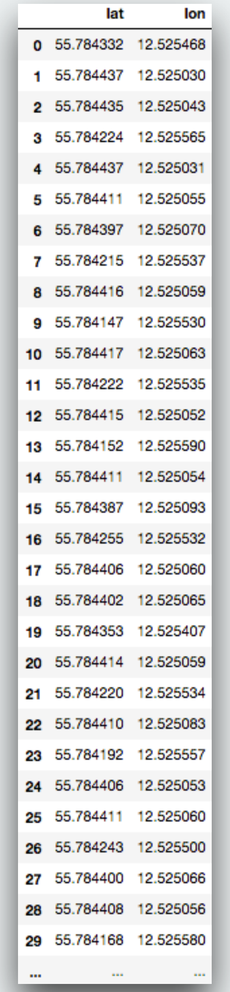


	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	12.525590
14	55.784411	12.525054
15	55.784387	12.525093
16	55.784255	12.525532
17	55.784406	12.525060
18	55.784402	12.525065
19	55.784353	12.525407
20	55.784414	12.525059
21	55.784220	12.525534
22	55.784410	12.525083
23	55.784192	12.525557
24	55.784406	12.525053
25	55.784411	12.525060
26	55.784243	12.525500
27	55.784400	12.525066
28	55.784408	12.525056
29	55.784168	12.525580
...	...	...

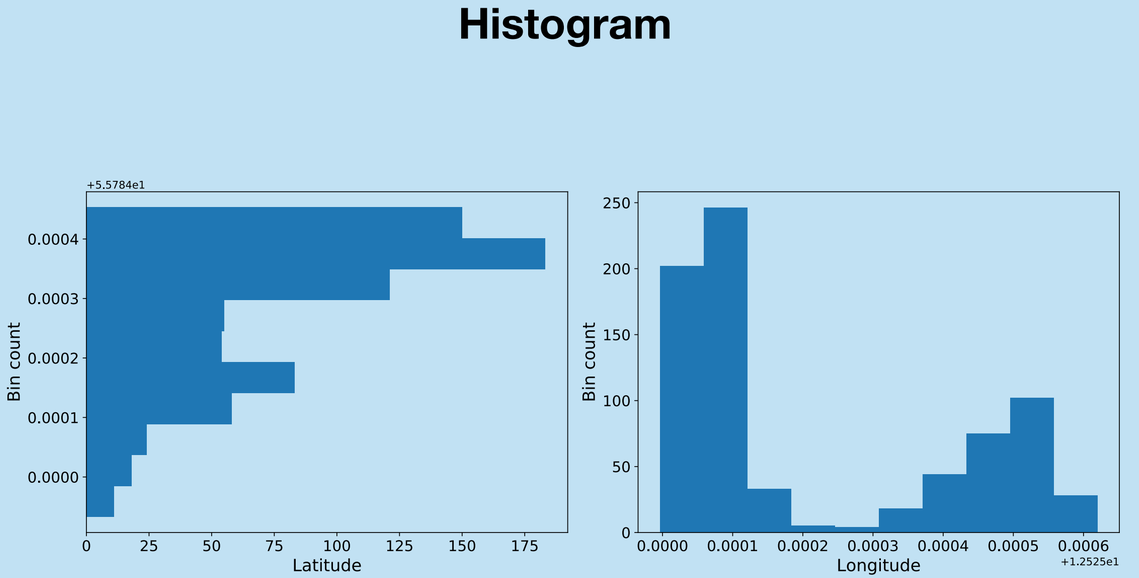


Data must be rendered in human-friendly format





	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	12.525590
14	55.784411	12.525054
15	55.784387	12.525093
16	55.784255	12.525532
17	55.784406	12.525060
18	55.784402	12.525065
19	55.784353	12.525407
20	55.784414	12.525059
21	55.784220	12.525534
22	55.784410	12.525083
23	55.784192	12.525557
24	55.784406	12.525053
25	55.784411	12.525060
26	55.784243	12.525500
27	55.784400	12.525066
28	55.784408	12.525056
29	55.784168	12.525580
...	...	...



Data must be rendered in human-friendly format

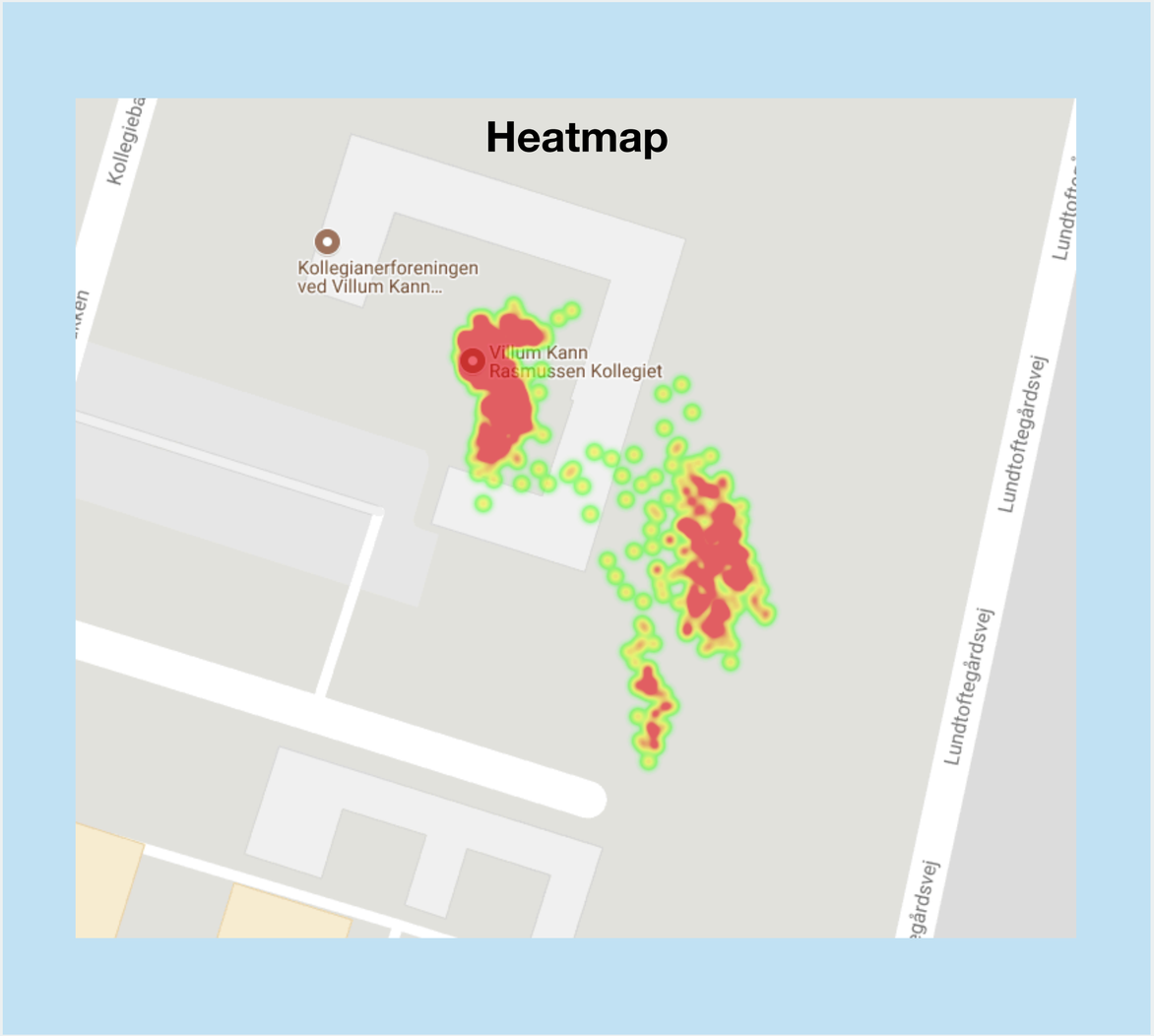
Terminal — less — 107×37

```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://www.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.garmin.com/xmlschemas/TrackPointExtension/v1 http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd"
  xmlns="http://www.topografix.com/GPX/1/1"
  xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
  xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <metadata>
    <link href="connect.garmin.com">
      <text>Garmin Connect</text>
    </link>
    <time>2010-12-21T17:31:19.000Z</time>
  </metadata>
  <trk>
    <name>To Work</name>
    <trkseg>
      <trkpt lon="12.577596567571163" lat="55.70799755863845">
        <ele>12.0</ele>
        <time>2011-01-26T09:23:55.000Z</time>
        <extensions>
          <gpxtpx:TrackPointExtension>
            <gpxtpx:hr>143</gpxtpx:hr>
          </gpxtpx:TrackPointExtension>
        </extensions>
      </trkpt>
      <trkpt lon="12.577596567571163" lat="55.70799755863845">
        <ele>12.0</ele>
        <time>2011-01-26T09:23:55.000Z</time>
        <extensions>
          <gpxtpx:TrackPointExtension>
            <gpxtpx:hr>143</gpxtpx:hr>
          </gpxtpx:TrackPointExtension>
        </extensions>
      </trkpt>
activity_65197512.gpx
```

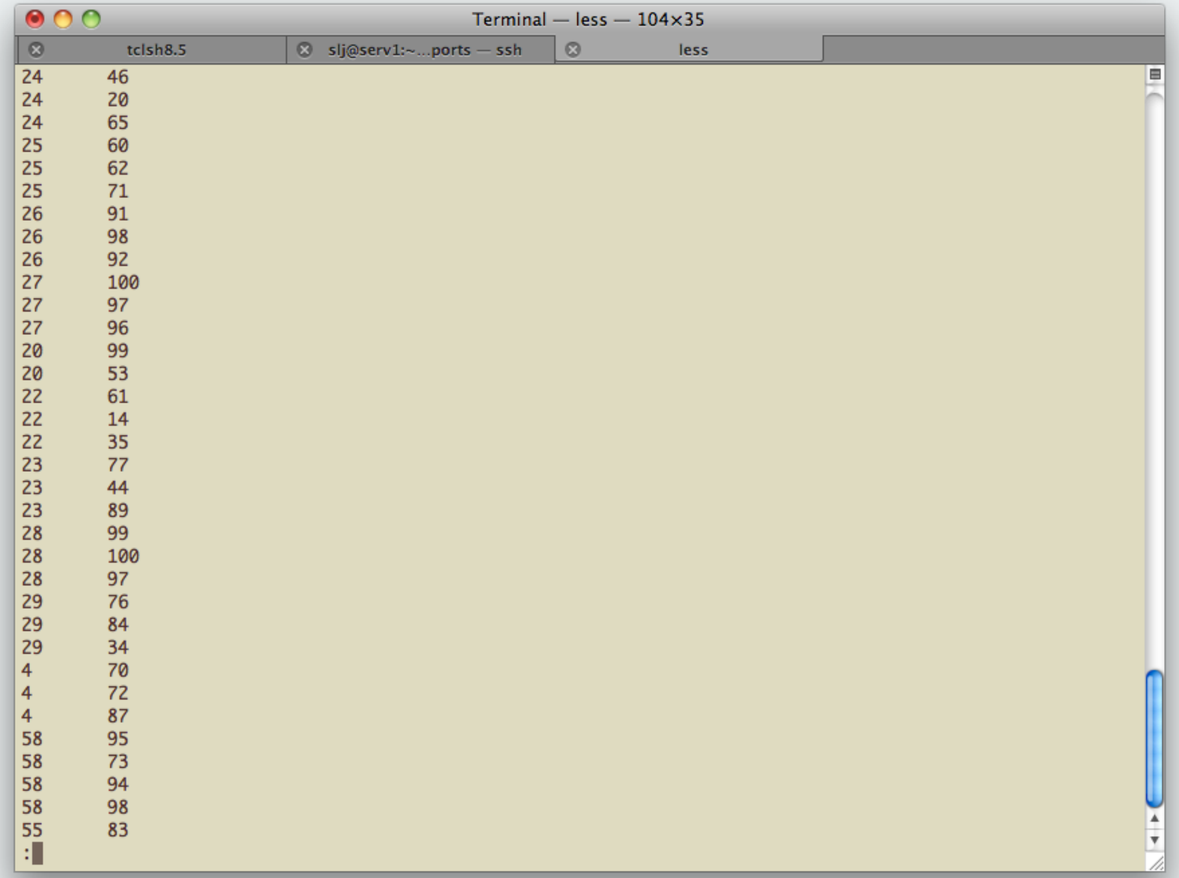
→

	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	12.525590
14	55.784411	12.525054
15	55.784387	12.525093
16	55.784255	12.525532
17	55.784406	12.525060
18	55.784402	12.525065
19	55.784353	12.525407
20	55.784414	12.525059
21	55.784220	12.525534
22	55.784410	12.525083
23	55.784192	12.525557
24	55.784406	12.525053
25	55.784411	12.525060
26	55.784243	12.525500
27	55.784400	12.525066
28	55.784408	12.525056
29	55.784168	12.525580
...	...	...

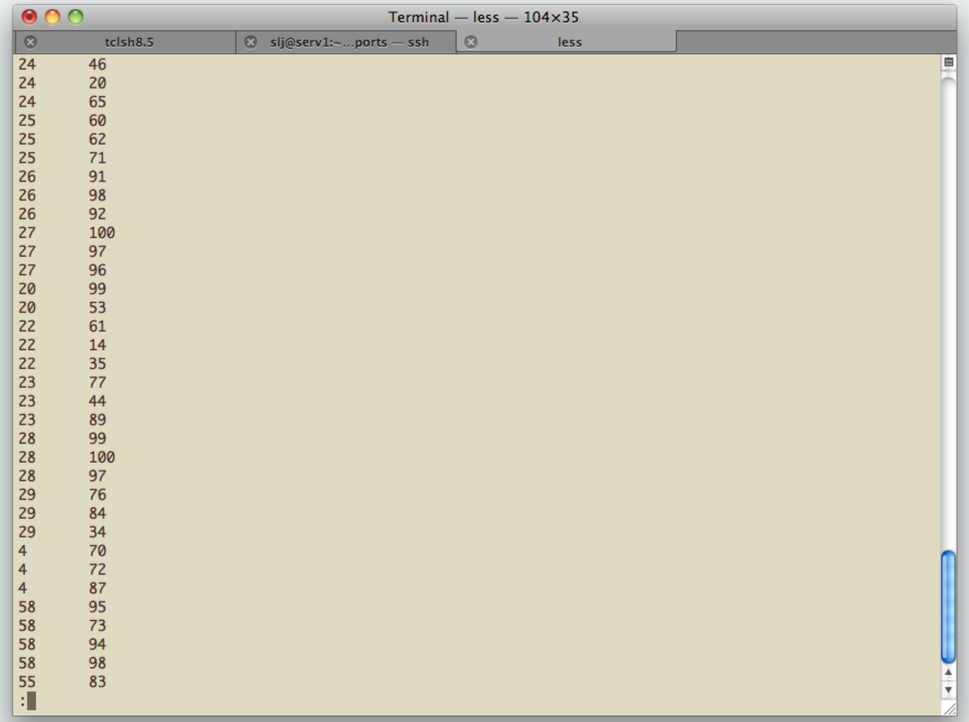
→

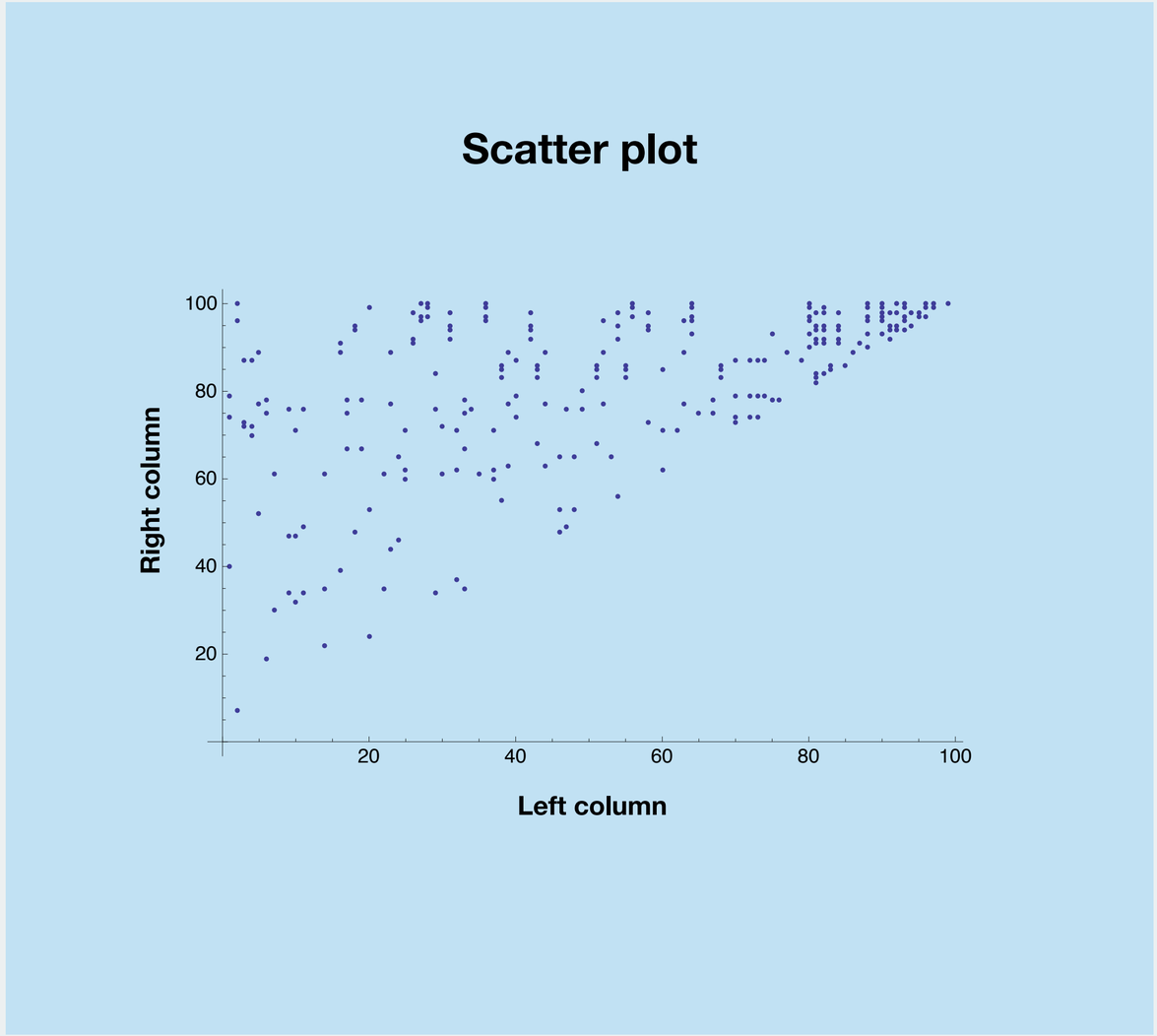


Relational data

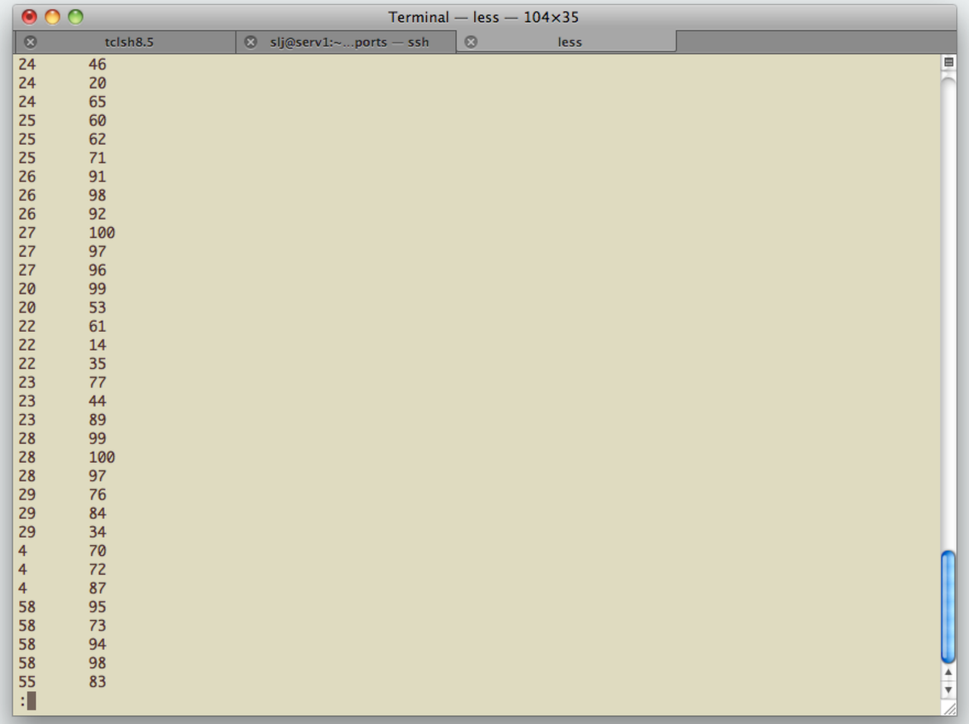


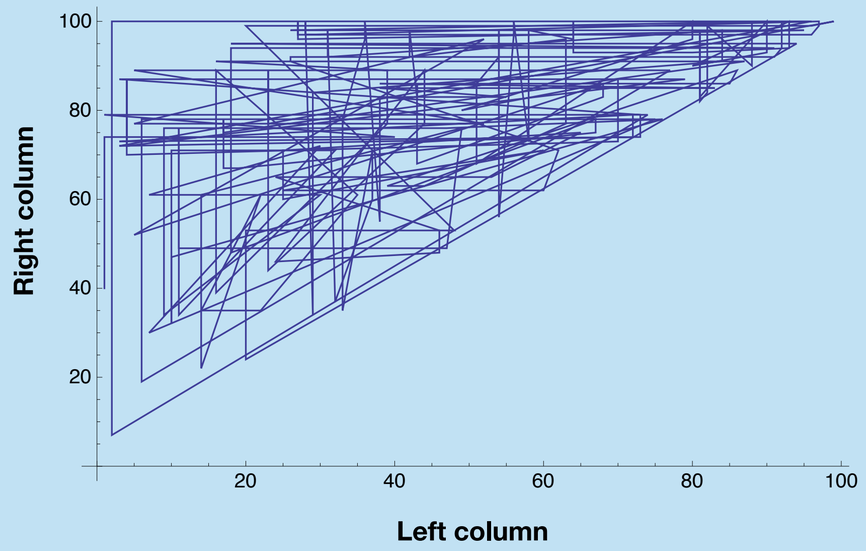
Relational data



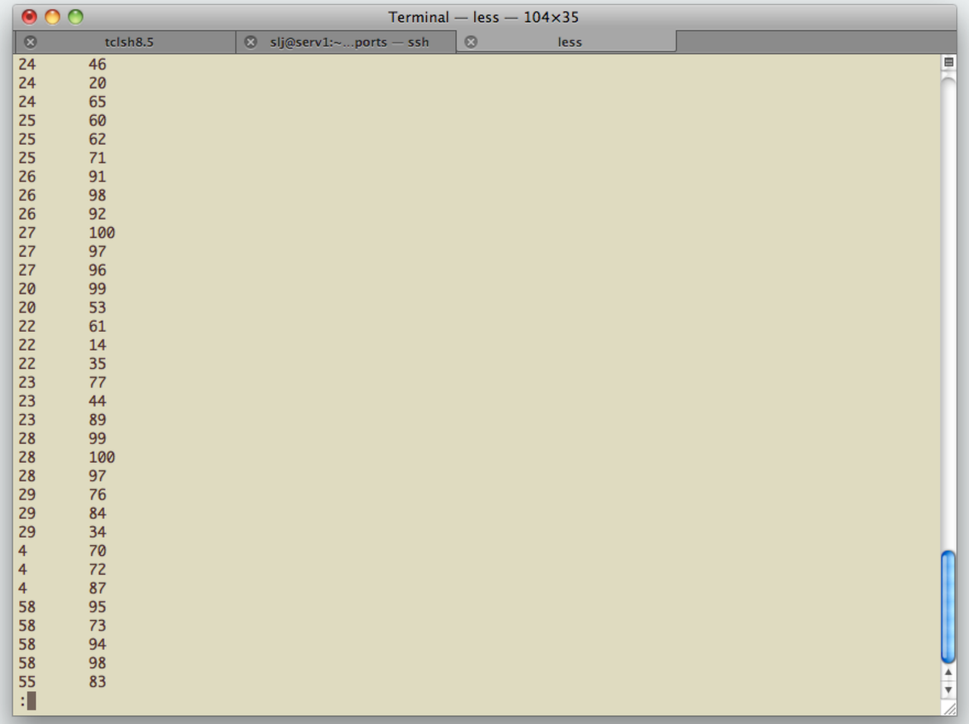


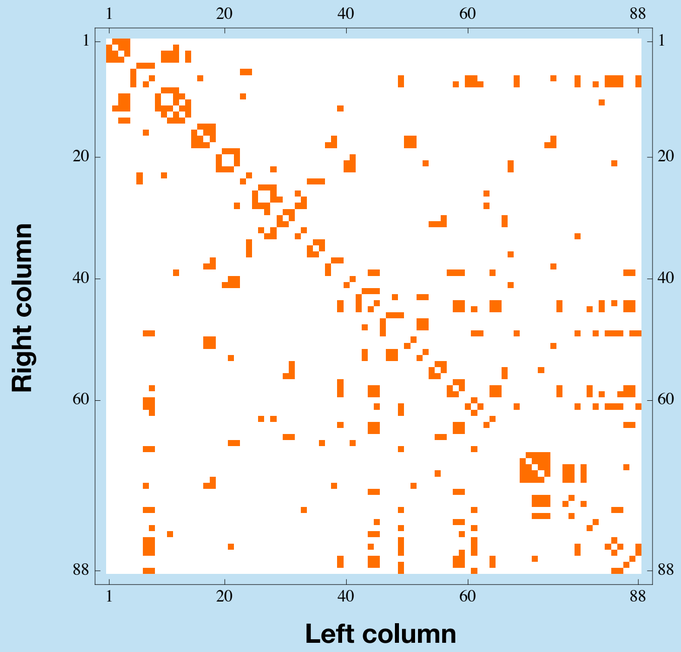
Relational data



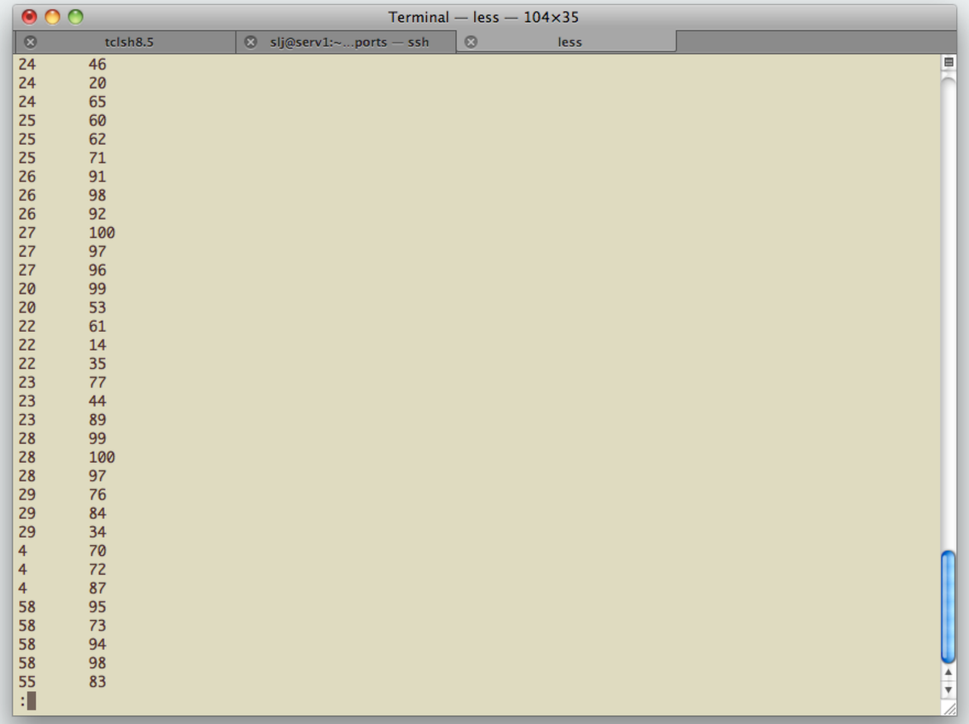


Relational data



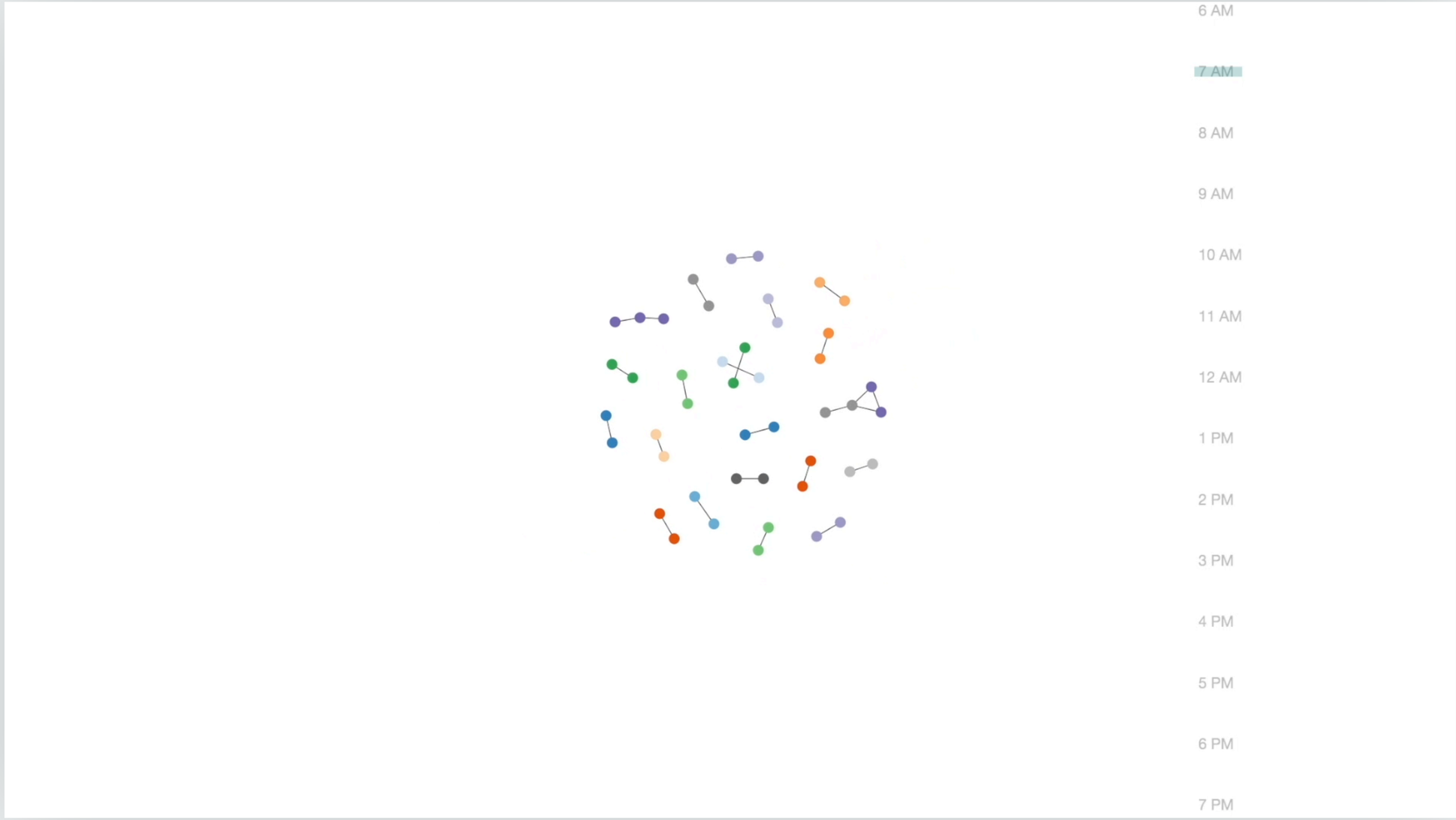


Relational data



Network

Very complex data that changes in time!



link

link

Most fancy visualizations break down to very simple things

For understanding how data is distributed

- Histograms
- Kernel density plots
- Box plots/violin plots
- Heatmaps

Most fancy visualizations break down to very simple things

For understanding how data is distributed

- Histograms
- Kernel density plots
- Box plots/violin plots
- Heatmaps

For understanding how variables in data compare and develop

- Scatter plots
- Pairs plot
- Time series plot
- Line plot
- Bar plot

Most fancy visualizations break down to very simple things

For understanding how data is distributed	• Histograms • Kernel density plots • Box plots/violin plots • Heatmaps
For understanding how variables in data compare and develop	• Scatter plots • Pairs plot • Time series plot • Line plot • Bar plot
For understanding interrelations in highly connected data	• Networks

Linear algebra

Linear algebra

A principled and scalable method for manipulating data

Linear algebra

A principled and scalable method for manipulating data

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

Linear algebra

A principled and scalable method for manipulating data

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

0D

```
In [2]: print np.random.randint(1, 100)
        Last executed 2018-01-25 11:52:52 in 5ms
        82
```

Linear algebra

A principled and scalable method for manipulating data

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

0D

```
In [2]: print np.random.randint(1, 100)
Last executed 2018-01-25 11:52:52 in 5ms
82
```

vector

1D

```
In [3]: print np.random.randint(1, 100, size=3)
Last executed 2018-01-25 11:53:37 in 5ms
[83 80 84]
```

Linear algebra

A principled and scalable method for manipulating data

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

matrix

2D

```
In [4]: print np.random.randint(1, 100, size=(3, 3))
Last executed 2018-01-25 11:54:38 in 4ms
[[99 47 77]
 [15 82  9]
 [59 55 48]]
```

scalar

0D

```
In [2]: print np.random.randint(1, 100)
Last executed 2018-01-25 11:52:52 in 5ms
82
```

vector

1D

```
In [3]: print np.random.randint(1, 100, size=3)
Last executed 2018-01-25 11:53:37 in 5ms
[83 80 84]
```

Linear algebra **A principled and scalable method for manipulating data**

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

0D

```
In [2]: print np.random.randint(1, 100)
Last executed 2018-01-25 11:52:52 in 5ms
82
```

vector

1D

```
In [3]: print np.random.randint(1, 100, size=3)
Last executed 2018-01-25 11:53:37 in 5ms
[83 80 84]
```

matrix

2D

```
In [4]: print np.random.randint(1, 100, size=(3, 3))
Last executed 2018-01-25 11:54:38 in 4ms
[[99 47 77]
 [15 82  9]
 [59 55 48]]
```

3D–tensor

3D

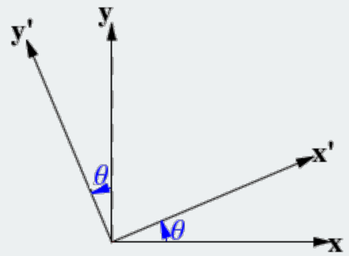
```
In [5]: print np.random.randint(1, 100, size=(3, 3, 3))
Last executed 2018-01-25 11:55:19 in 5ms
[[[45 11 73]
  [84 50 88]
  [13 22 97]]

 [[10  5 12]
  [27 23 76]
  [43 84 53]]

 [[86 58 61]
  [71 95 86]
  [92 19 68]]]
```

Linear algebra **A principled and scalable method for manipulating data**

Operations

- Products: ***dot***, *cross*
- Elementwise: *addition*, *subtraction*, *multiplication*, *division*
- Mutations: *transpose*, *inverse/pseudo-inverse*, *scaling*, *rotation*

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{bmatrix}$$



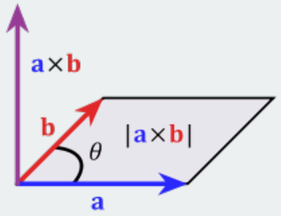
used frequently for basis transformation

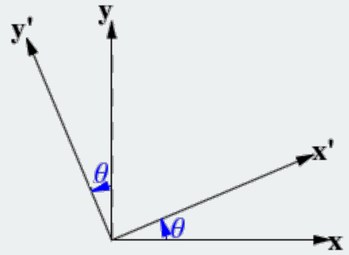
Linear algebra **A principled and scalable method for manipulating data**

Operations

- Products: *dot*, ***cross***
- Elementwise: *addition*, *subtraction*, *multiplication*, *division*
- Mutations: *transpose*, *inverse/pseudo-inverse*, *scaling*, *rotation*

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{bmatrix}$$





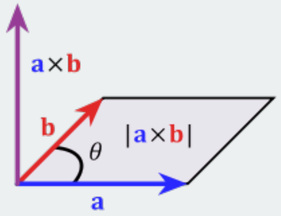
used frequently for basis transformation

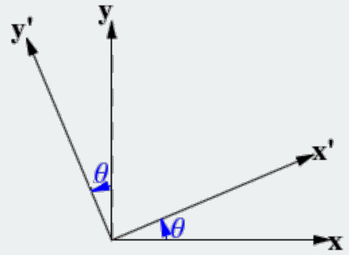
Linear algebra **A principled and scalable method for manipulating data**

Operations

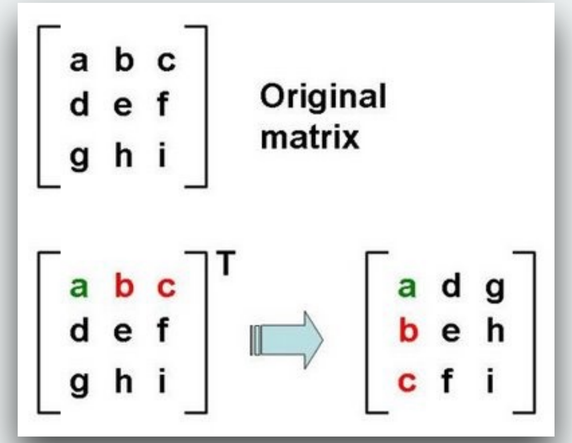
- Products: *dot*, *cross*
- Elementwise: *addition*, *subtraction*, *multiplication*, *division*
- Mutations: ***transpose***, *inverse/pseudo-inverse*, *scaling*, *rotation*

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{bmatrix}$$





used frequently for basis transformation



Statistics

Statistics

A set of tools and jargon for describing data

Statistics

A set of tools and jargon for describing data

Vocabulary

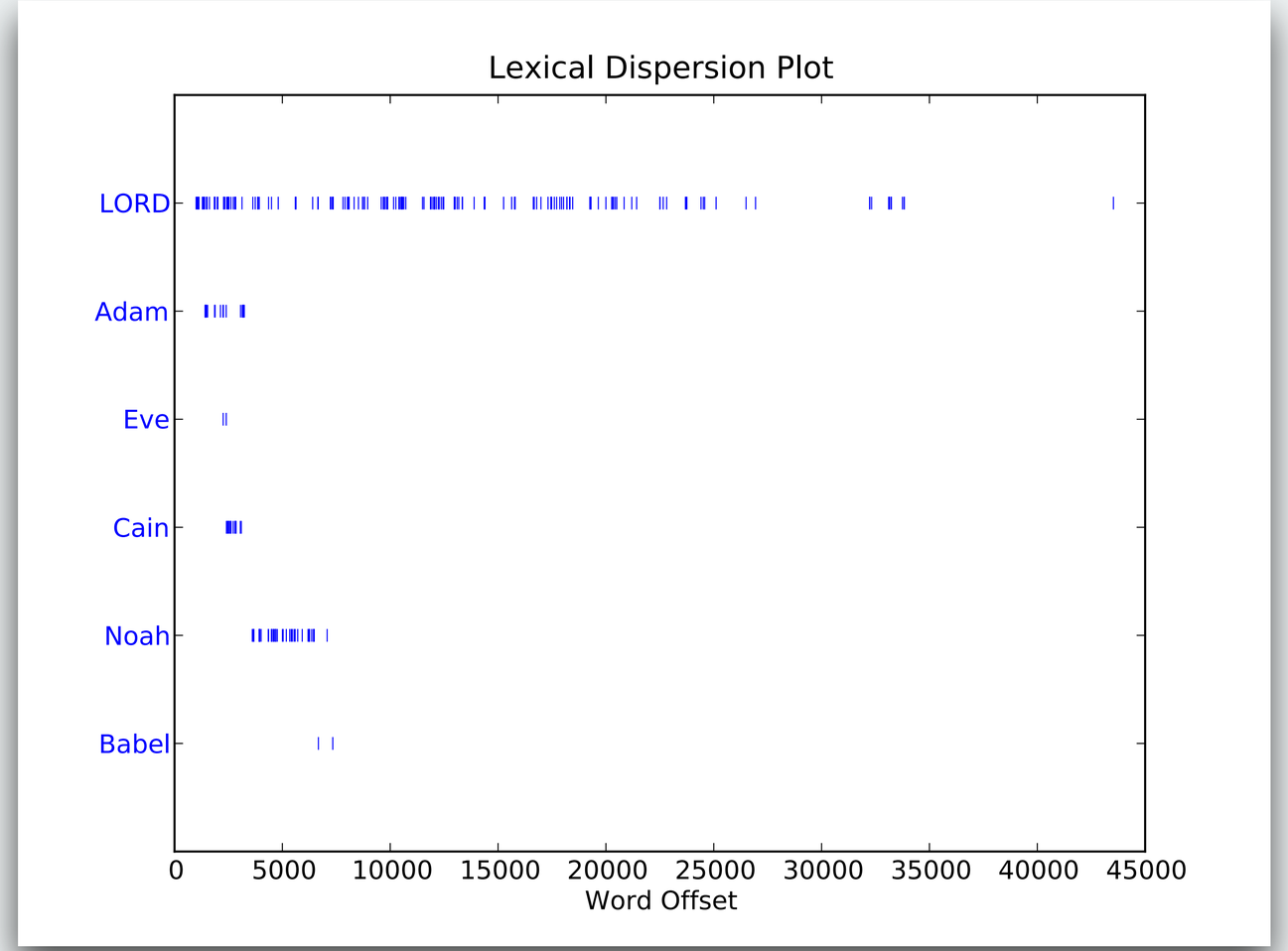
- Mean, median
- Variance, standard deviation, range
- Correlation, covariance

Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- Variance, standard deviation, range
- Correlation, covariance



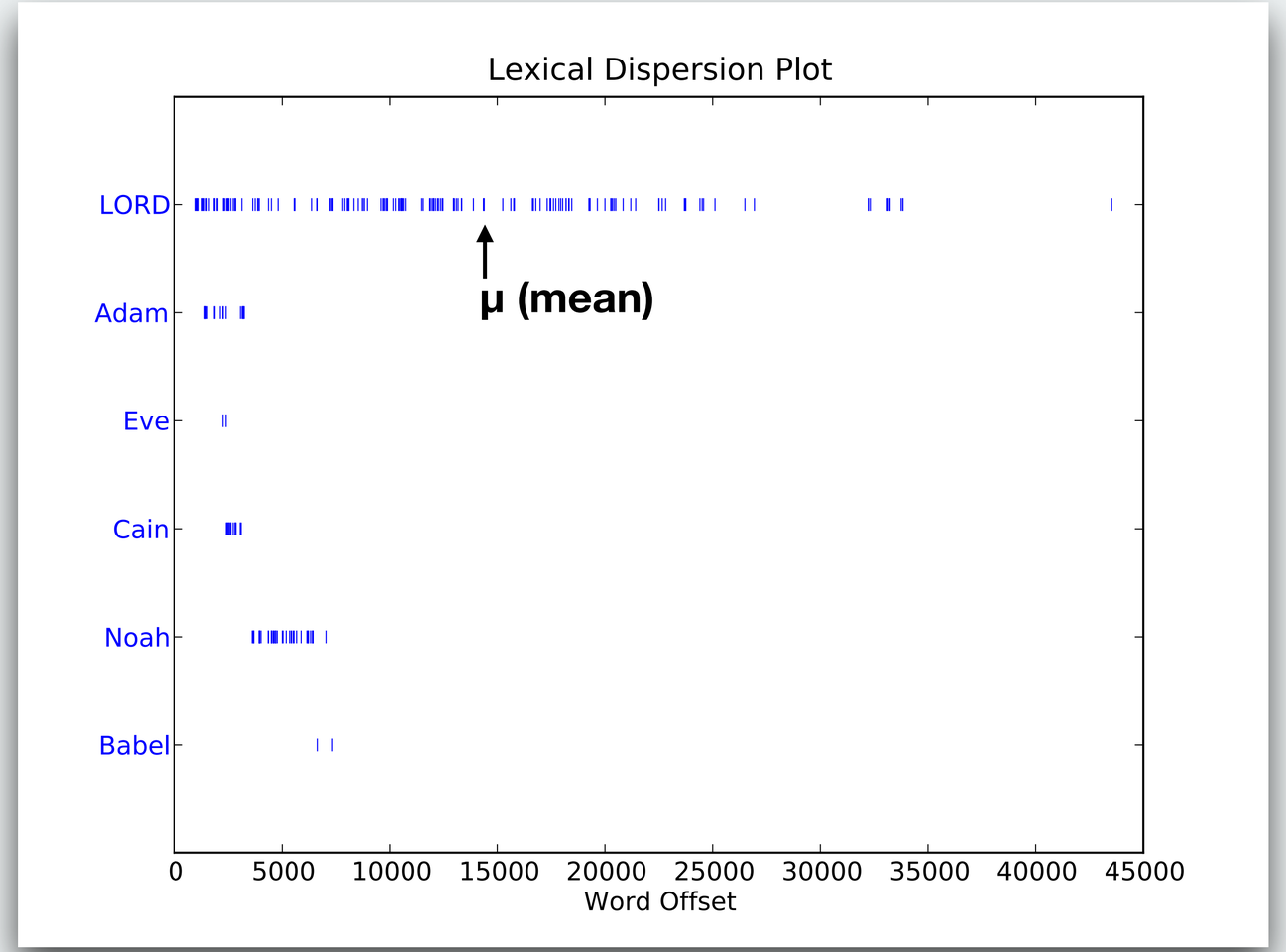
Statistics

A set of tools and jargon for describing data

Vocabulary

- **Mean**, median
- Variance, standard deviation, range
- Correlation, covariance

$$\mu = \frac{\text{Sum of values}}{\text{Number of values}}$$



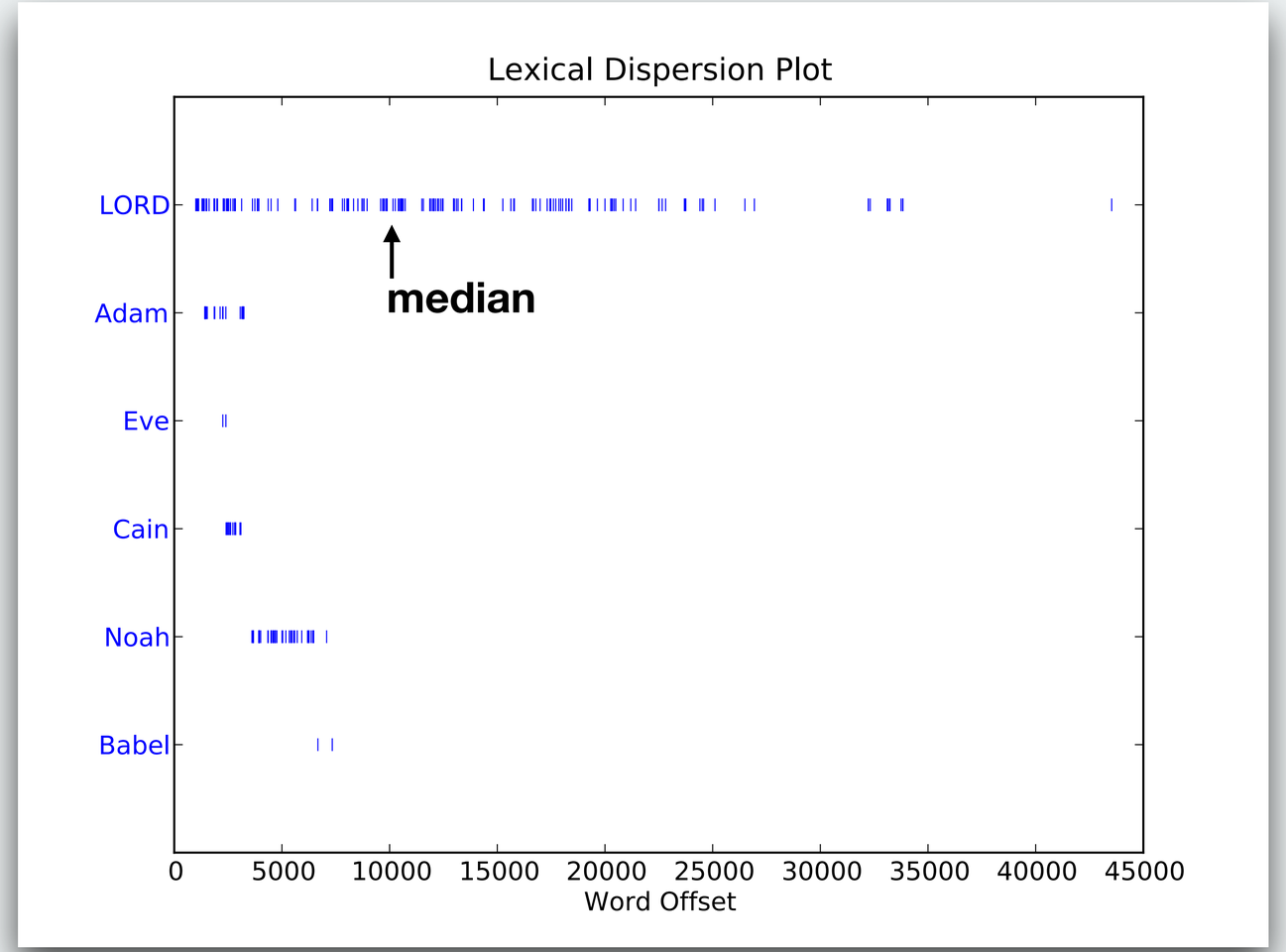
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, **median**
- Variance, standard deviation, range
- Correlation, covariance

median = Middle number in ordered list



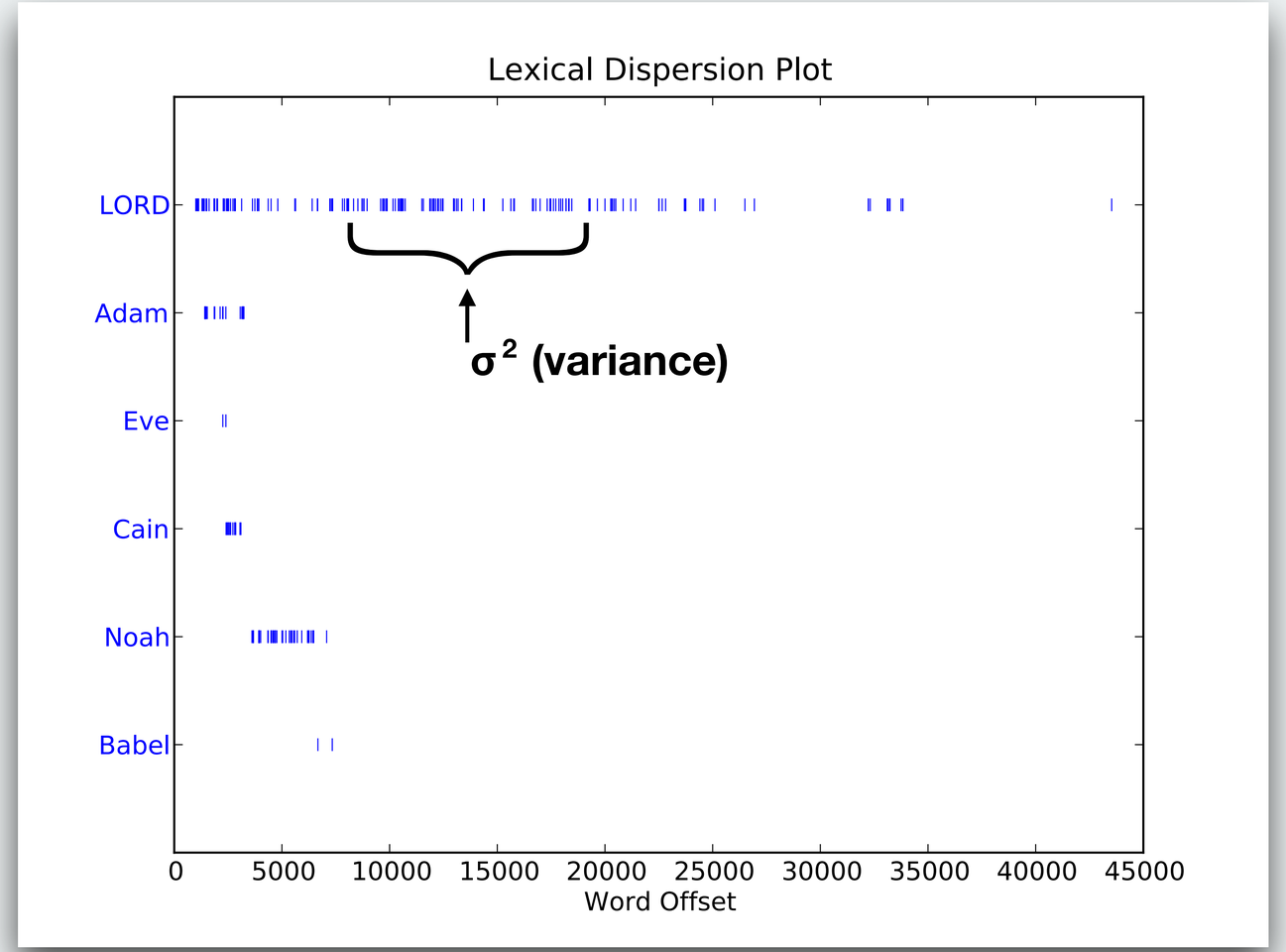
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- **Variance**, standard deviation, range
- Correlation, covariance

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2$$



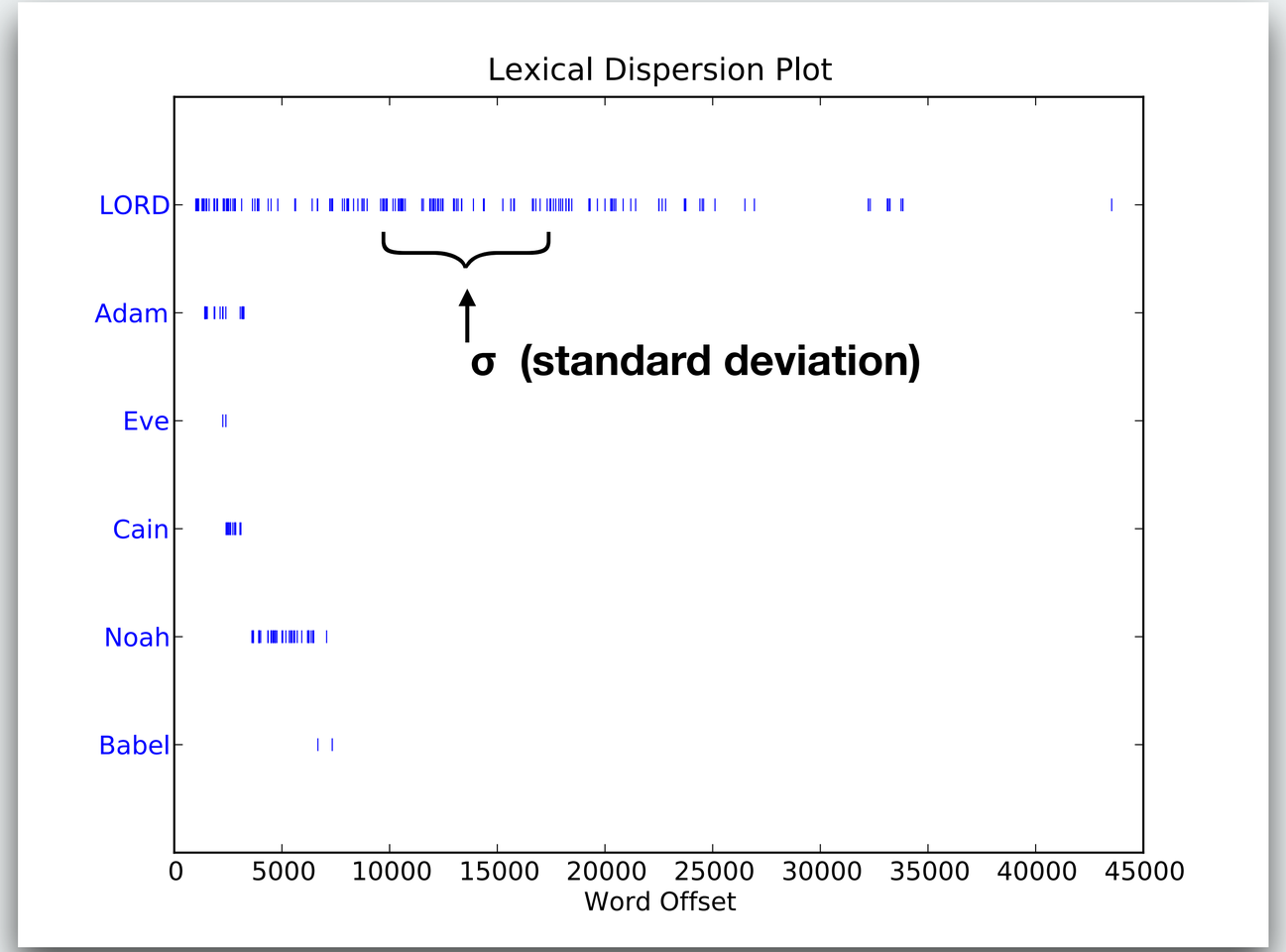
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- Variance, **standard deviation**, range
- Correlation, covariance

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$



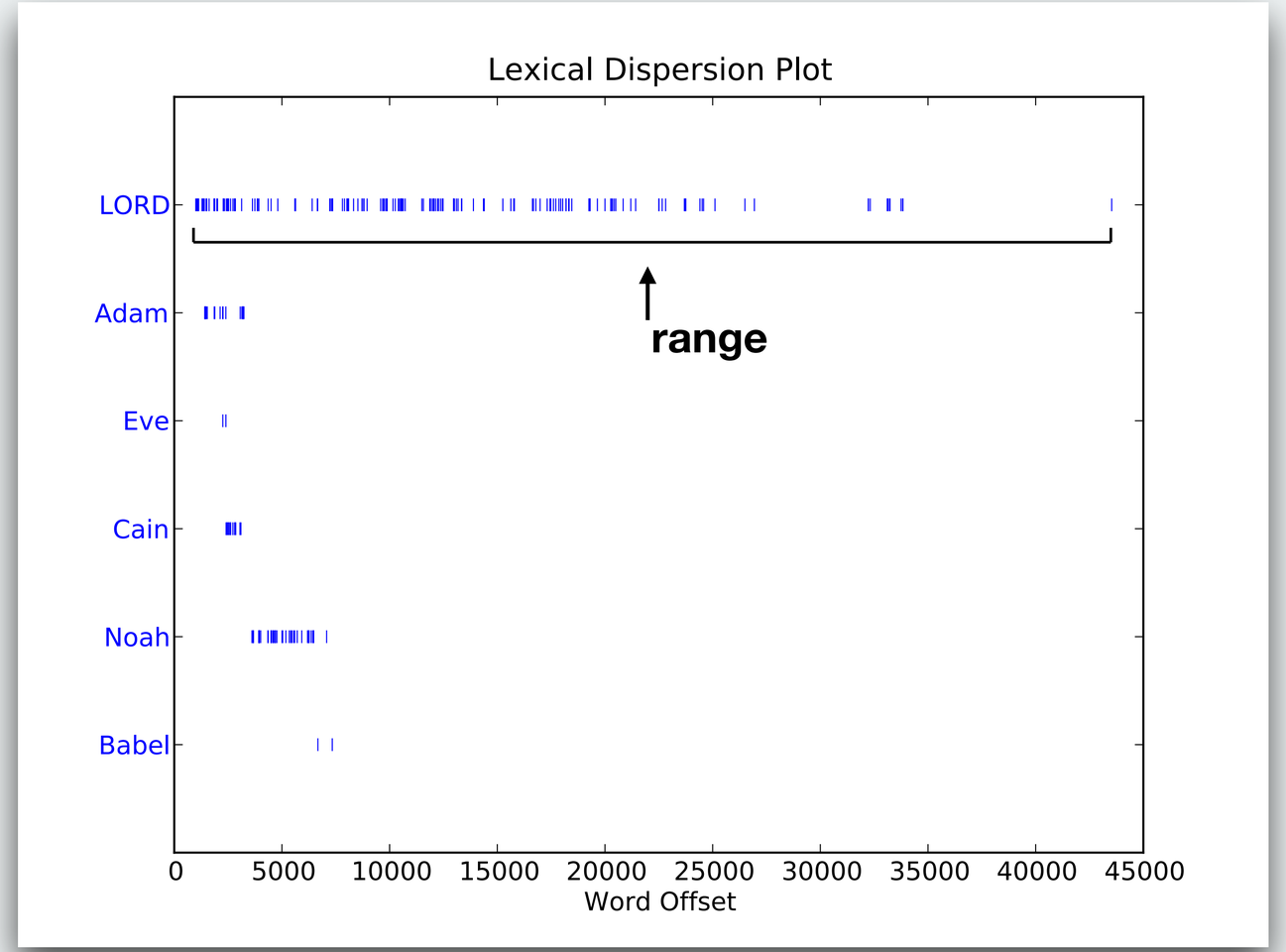
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- Variance, standard deviation, **range**
- Correlation, covariance

range = max(value) - min(value)



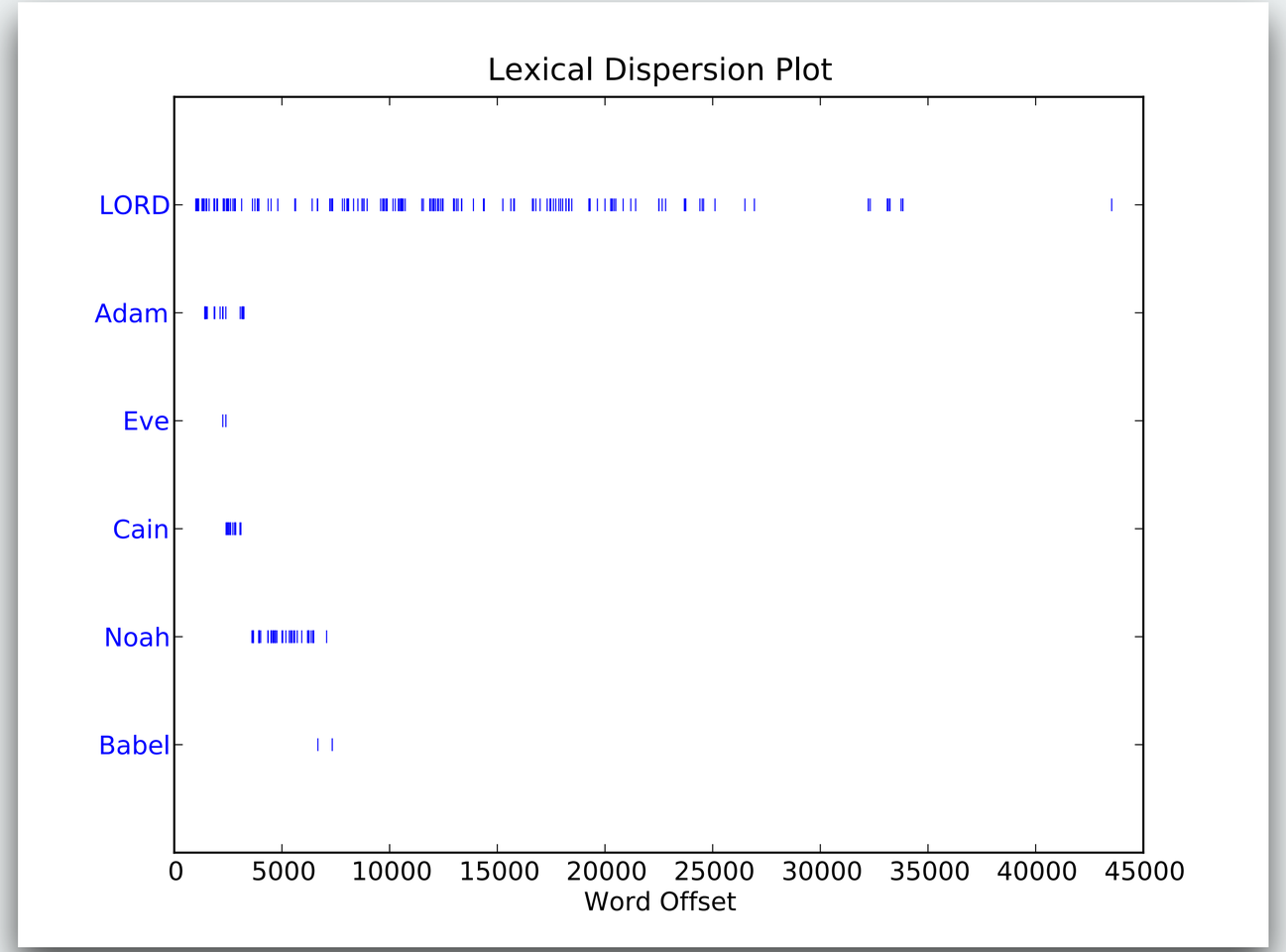
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- Variance, standard deviation, range
- Correlation, **covariance**

cov(X, Y) = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_y)$



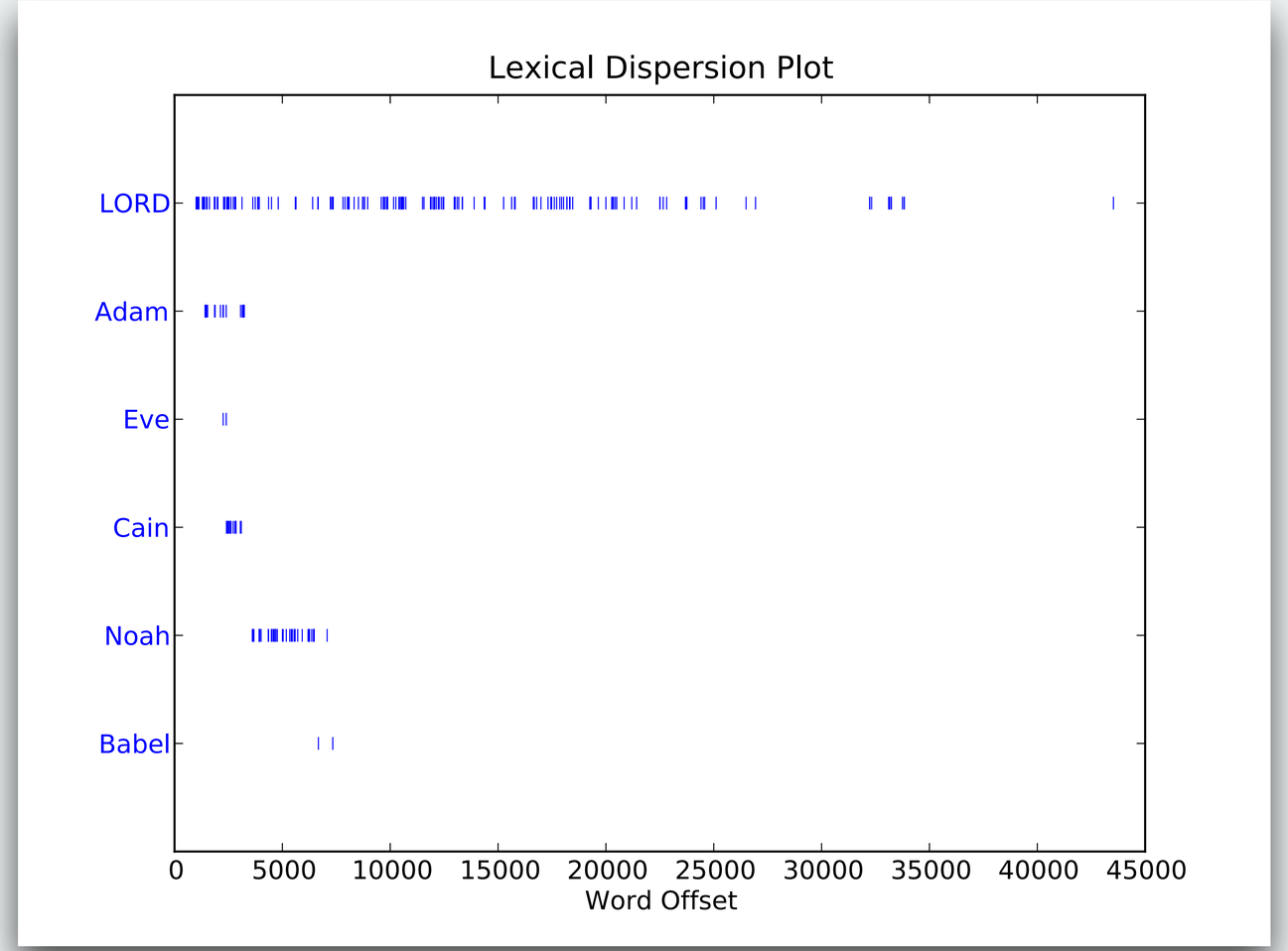
Statistics

A set of tools and jargon for describing data

Vocabulary

- Mean, median
- Variance, standard deviation, range
- Correlation, **covariance**

$$\textbf{cor(X, Y)} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$



Probability theory

Probability theory

Formalized framework for dealing with randomness

Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- Distribution and process
- Random and stochastic
- Probability functions
  - Probability *mass* function (pmf)
  - Probability *density* function (pdf)
  - *Cumulative density* function (cdf)

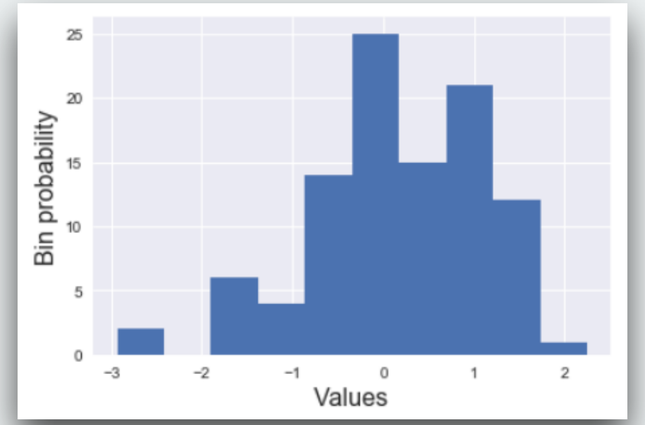
Probability theory

Formalized framework for dealing with randomness

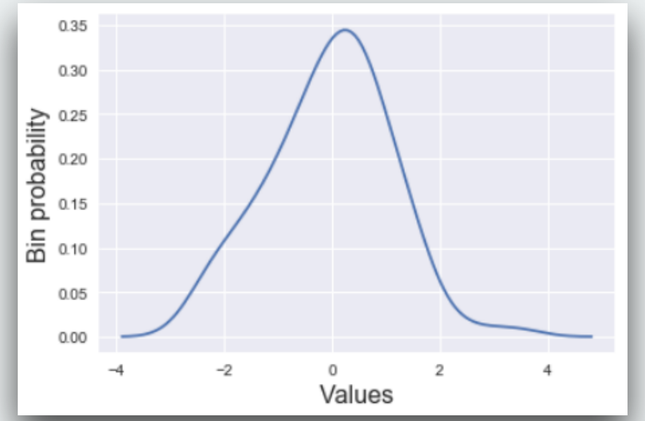
Important concepts

- **Discrete vs. continuous**
- Distribution and process
- Random and stochastic
- Probability functions
  - Probability *mass* function (pmf)
  - Probability *density* function (pdf)
  - *Cumulative density* function (cdf)

Discrete



Continuous



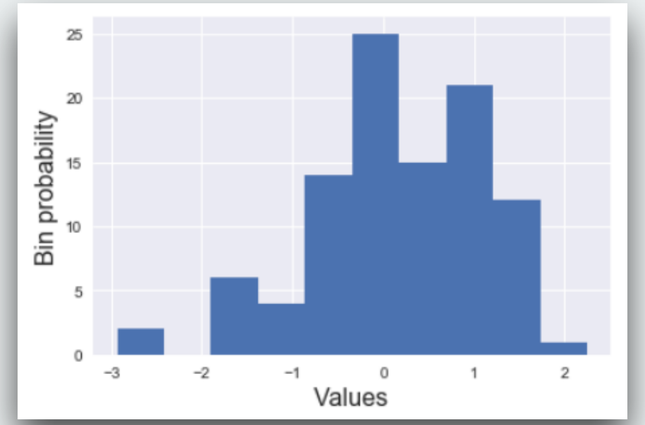
Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- **Distribution and process**
- Random and stochastic
- Probability functions
  - Probability *mass* function (pmf)
  - Probability *density* function (pdf)
  - *Cumulative density* function (cdf)

Distribution



Process

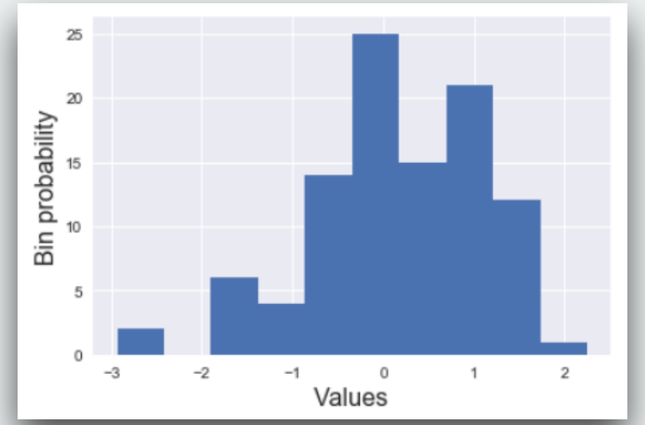
Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- Distribution and process
- **Random and stochastic**
- Probability functions
  - Probability *mass* function (pmf)
  - Probability *density* function (pdf)
  - *Cumulative density* function (cdf)

A **variable** is *random*



A **process** is *stochastic*

Otherwise the two words mean the same

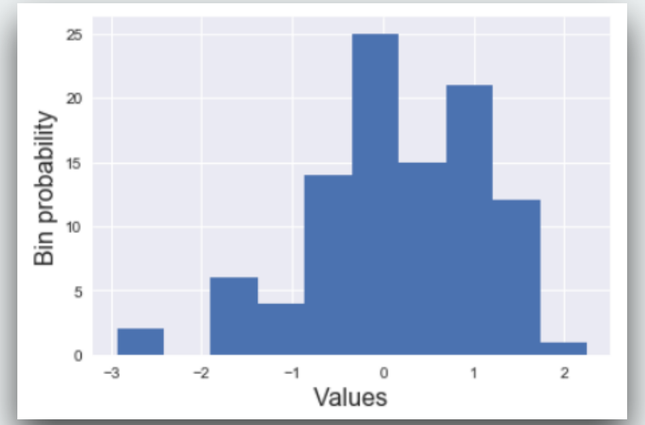
Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- Distribution and process
- Random and stochastic
- Probability functions
  - **Probability *mass* function (pmf)**
  - Probability *density* function (pdf)
  - *Cumulative density* function (cdf)

pmf



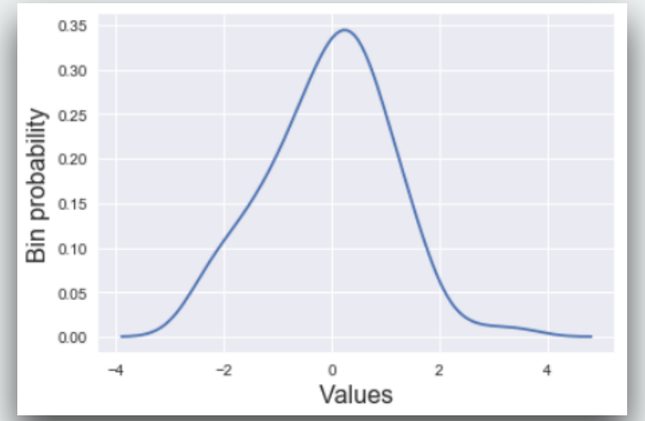
Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- Distribution and process
- Random and stochastic
- Probability functions
  - Probability *mass* function (pmf)
  - **Probability *density* function (pdf)**
  - *Cumulative density* function (cdf)

pdf



Probability theory

Formalized framework for dealing with randomness

Important concepts

- Discrete vs. continuous
- Distribution and process
- Random and stochastic
- Probability functions
  - Probability *mass* function (pmf)
  - Probability *density* function (pdf)
  - ***Cumulative density* function (cdf)**

cdf