Computational Analysis of Big Data

Week 2

A Data Scientist's most fundamental tools

More specifically: Visualization, linear algebra and statistics

Agenda

Review of Week 1 Exercises

Work through Visualisation Exercises
- **Review together 2.1.2**

Work through Linear Algebra Exercises
- **Stop before 2.2.4**

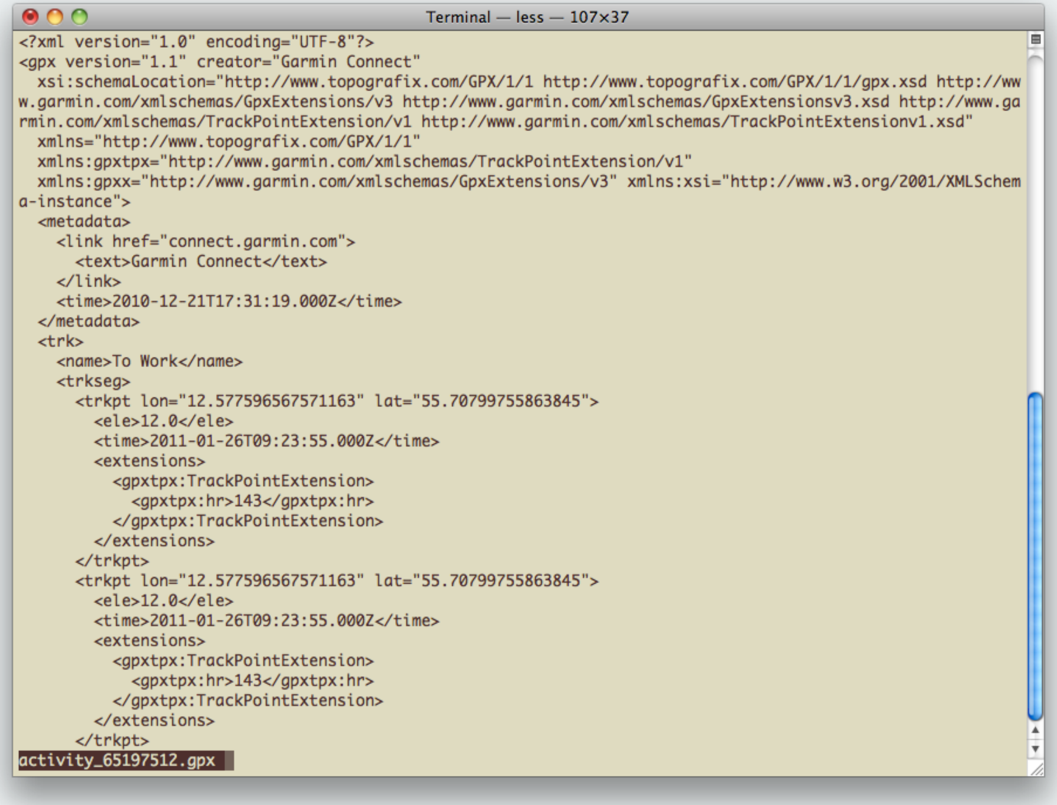
Lecture on Stats and PCA

Work on 2.2.4 and 2.2.5
- **Review together 2.2.4/5**

Work on section 2.3

This is GPS data

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



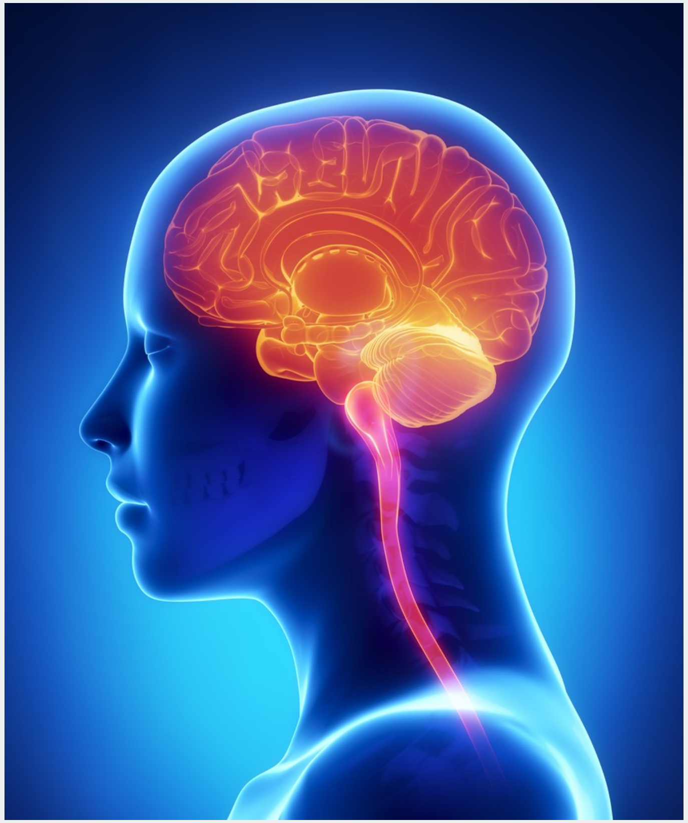
```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

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

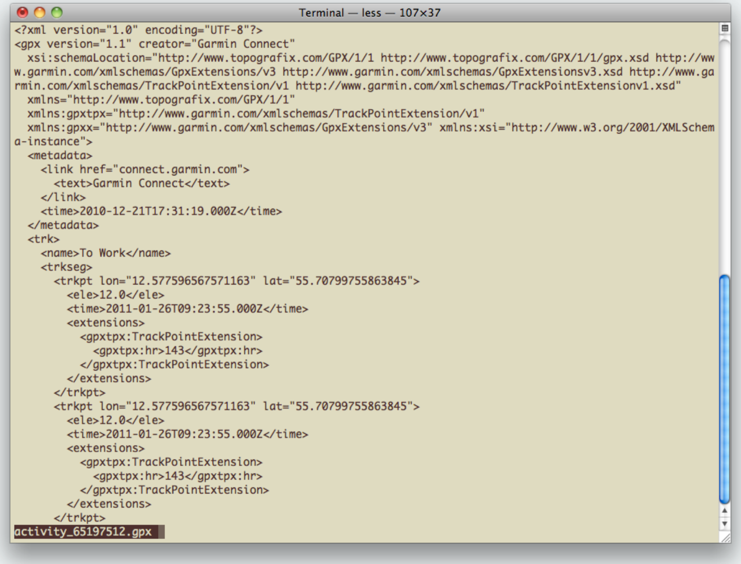


The human eye is made for advanced pattern recognition

It can:

- Immediately **recognize patterns** in highly complex images
- Notice **outliers**
- Process streams of images and recognize **patterns over time**

Data must be rendered in human-friendly format

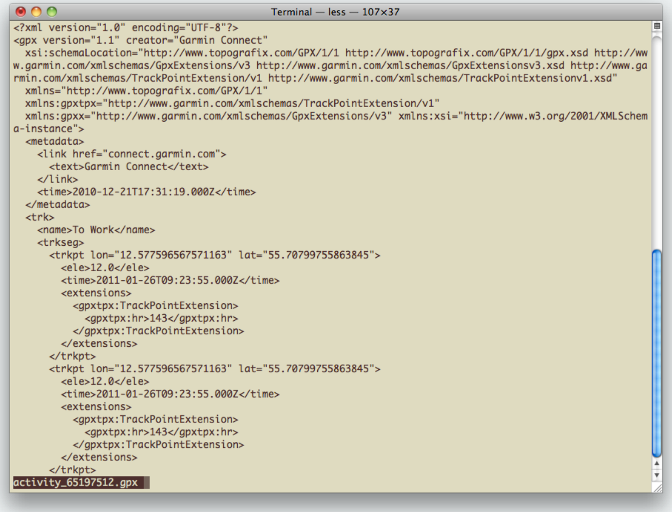


```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

→

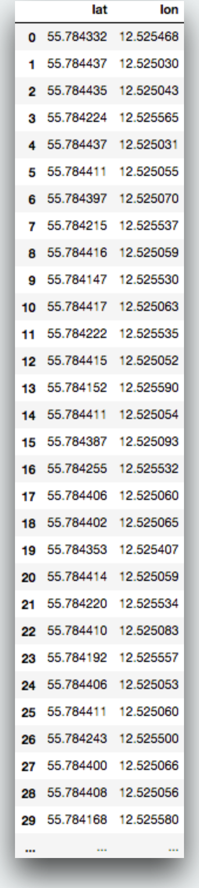
?

Data must be rendered in human-friendly format



```
<?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://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

→

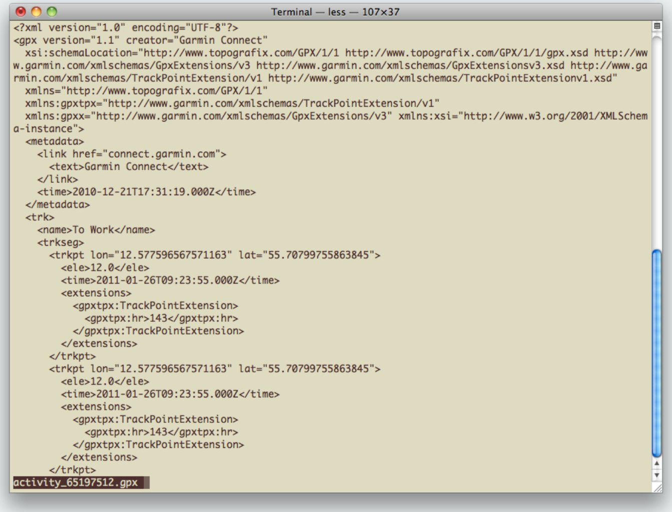


	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	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
...	...	...

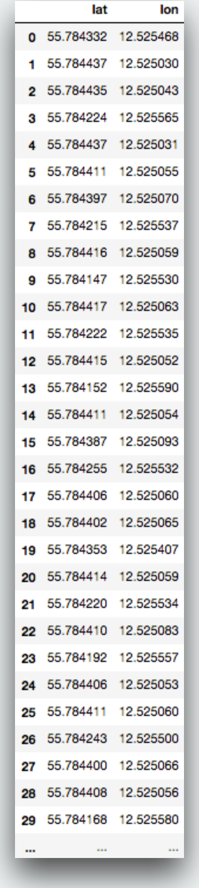
→

?

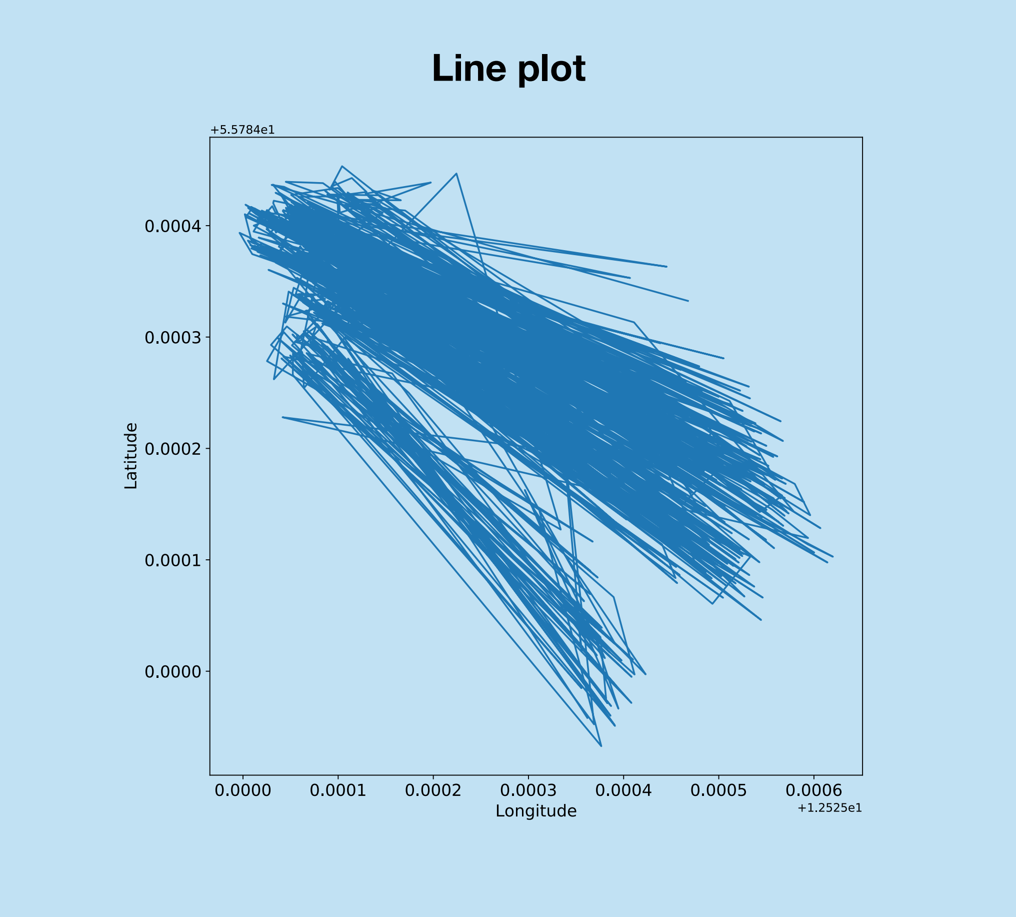
Data must be rendered in human-friendly format



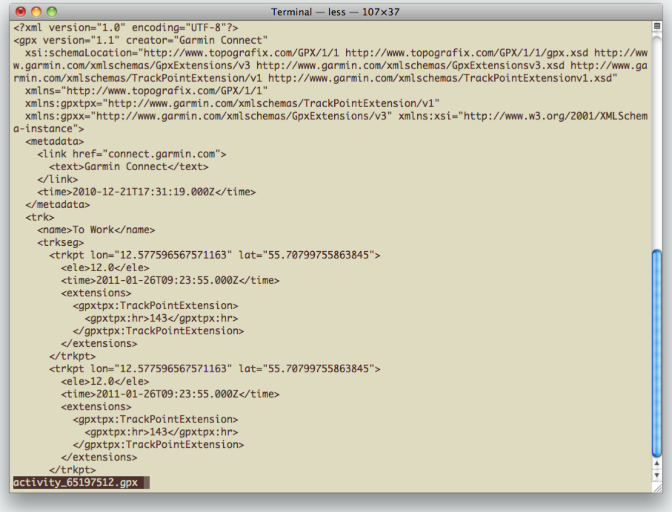
```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
...	...	...

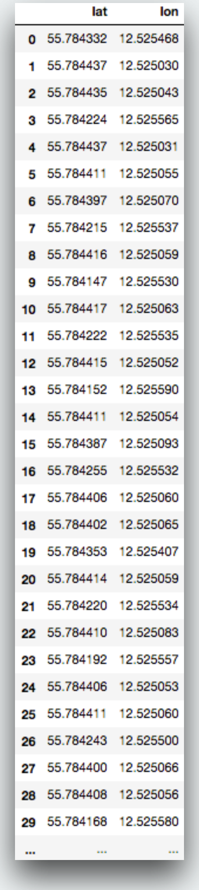


Data must be rendered in human-friendly format



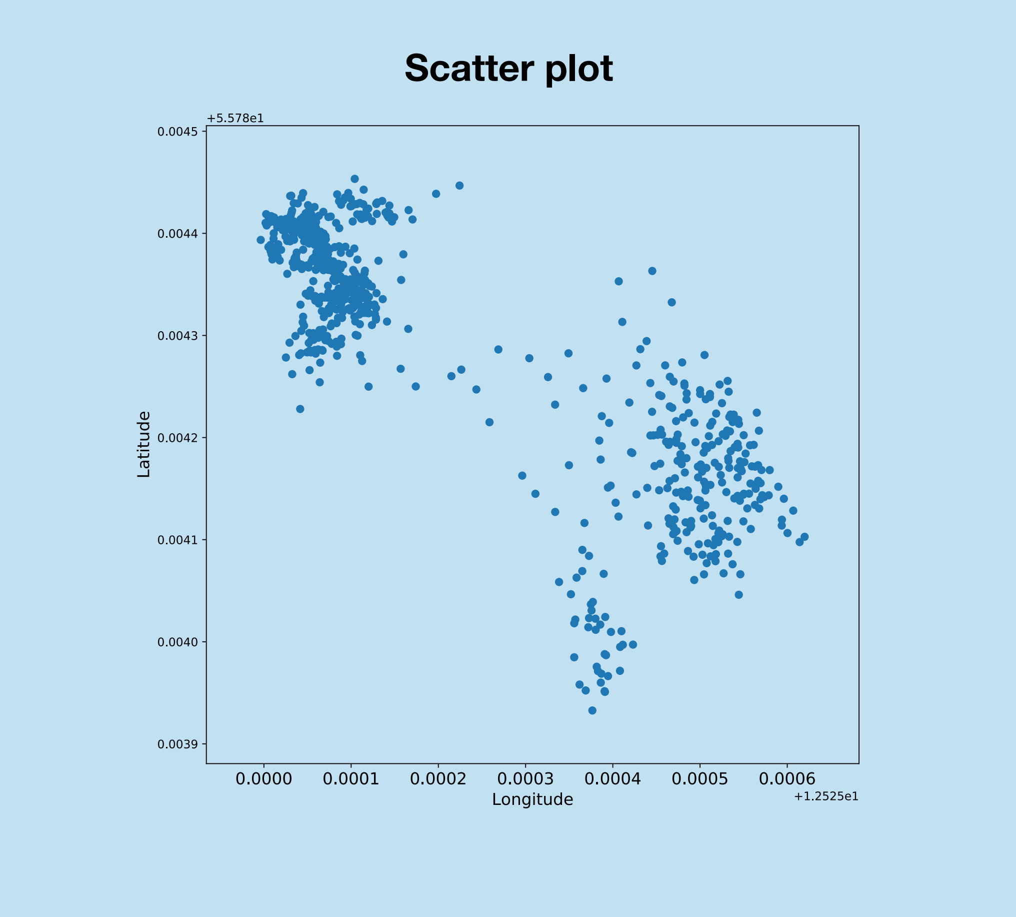
```
<?xml version="1.0" encoding="UTF-8"?>
<gpx version="1.1" creator="Garmin Connect"
  xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

→

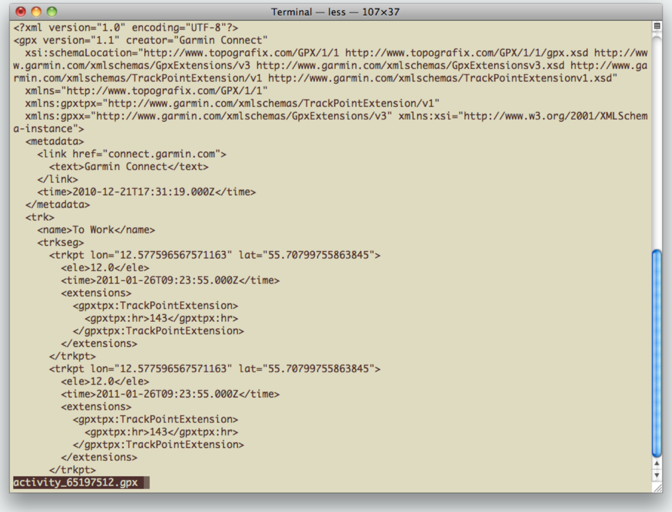


	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	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
...	...	...

→

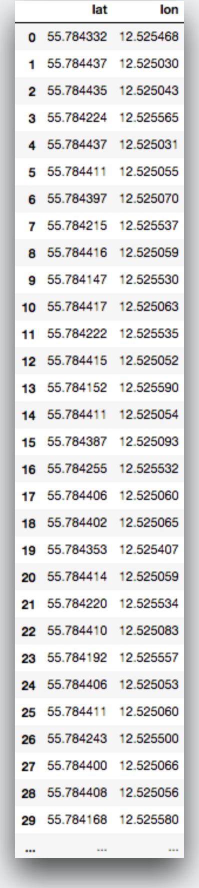


Data must be rendered in human-friendly format



```
<?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://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

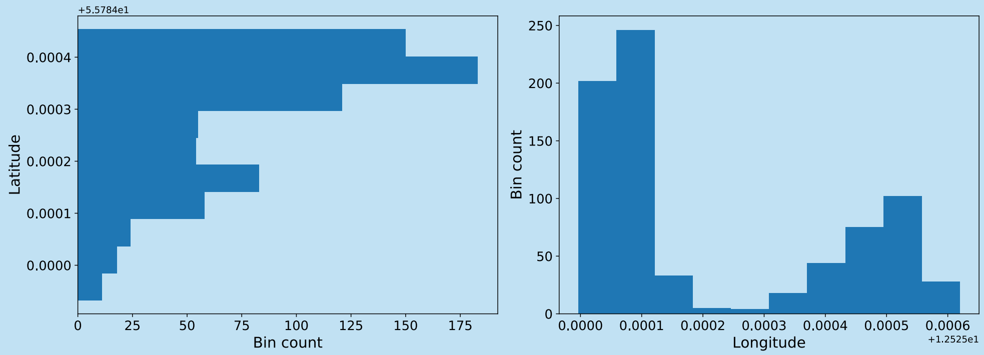
→



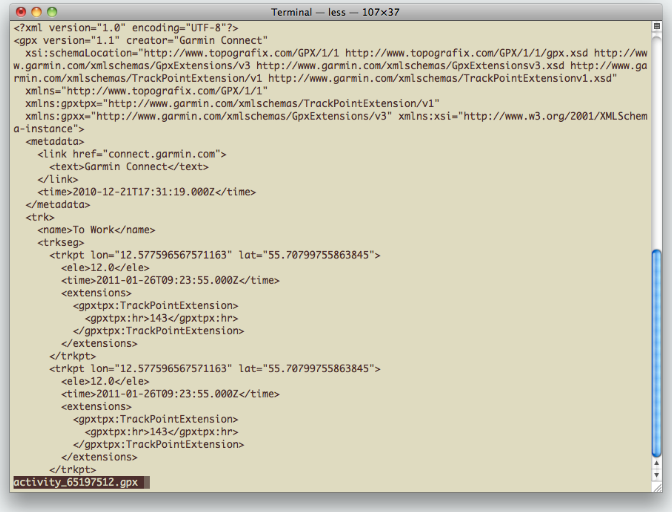
	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152	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
...	...	...

→

Histogram

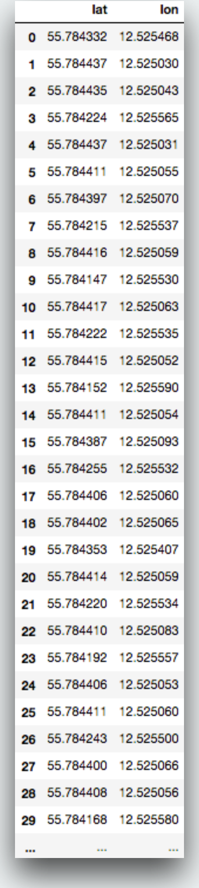


Data must be rendered in human-friendly format



```
<?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://ww
w.garmin.com/xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.ga
rmin.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/XMLSchem
a-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
```

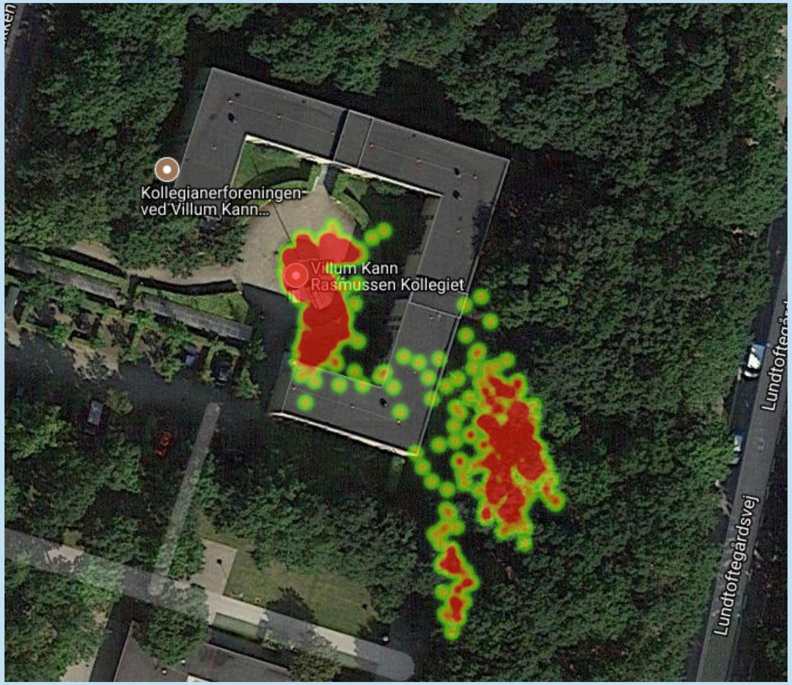
→



	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
...	...	...

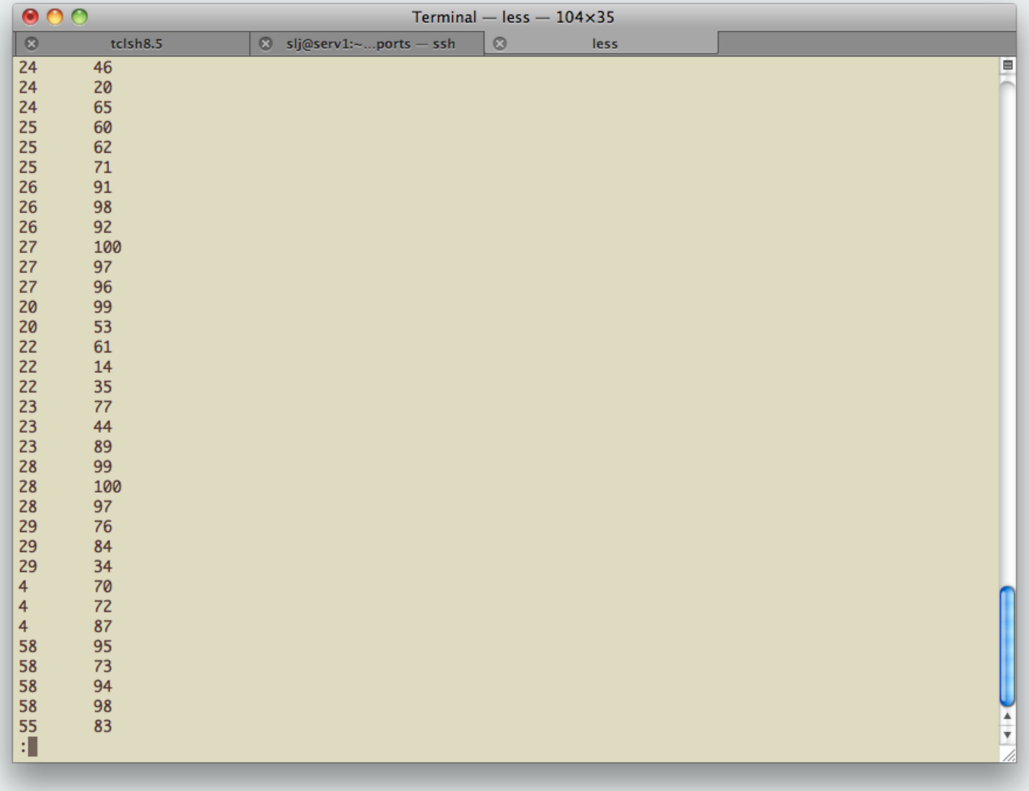
→

Heatmap



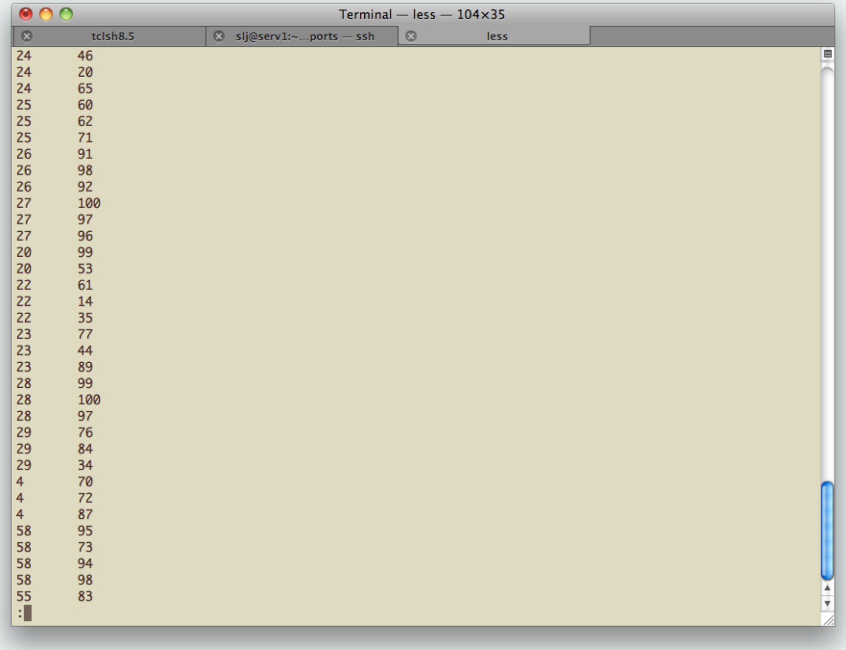
Kollegianerforeningen ved Villum Kann...
Villum Kann Rasmussen Kollegiet
Lundtoftegårdsvej
Lundtoftegå...

Relational data

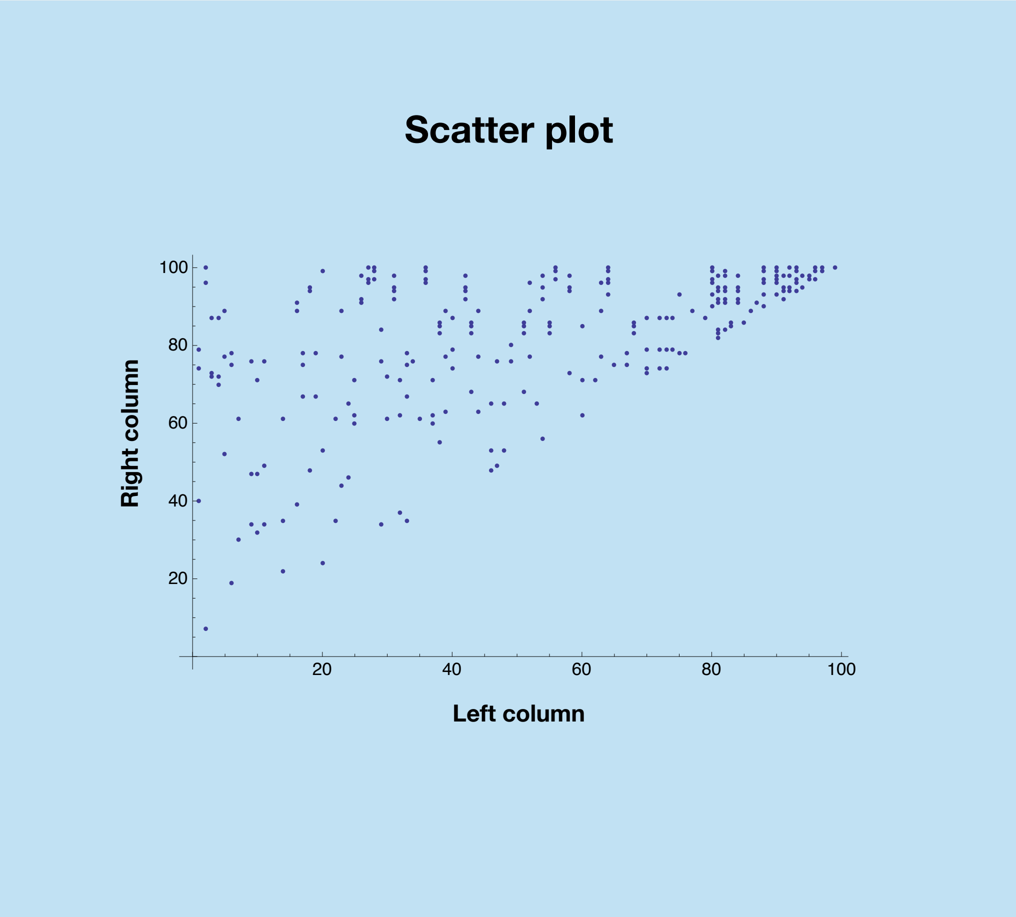


```
24      46
24      20
24      65
25      60
25      62
25      71
26      91
26      98
26      92
27      100
27      97
27      96
20      99
20      53
22      61
22      14
22      35
23      77
23      44
23      89
28      99
28      100
28      97
29      76
29      84
29      34
4       70
4       72
4       87
58      95
58      73
58      94
58      98
55      83
:
```

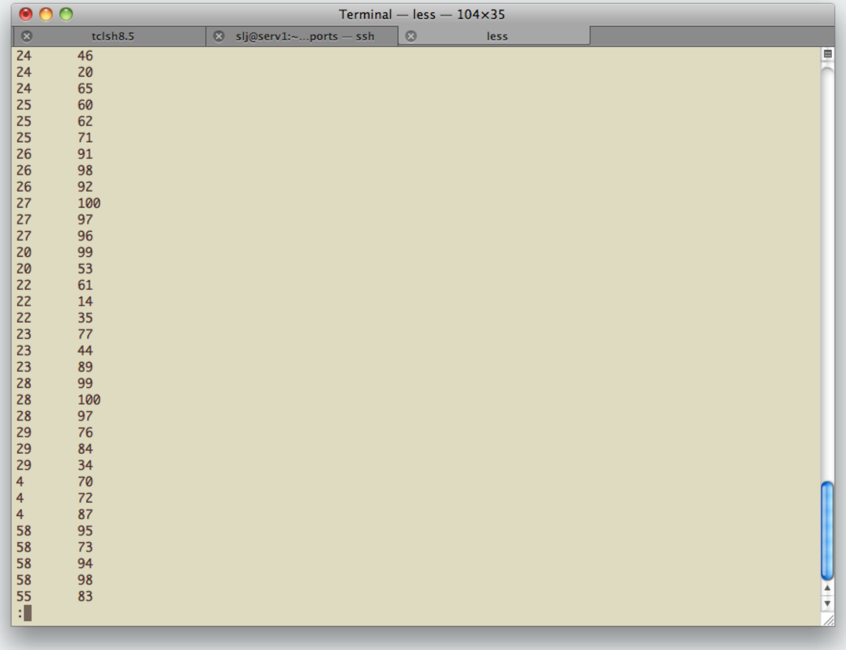
Relational data



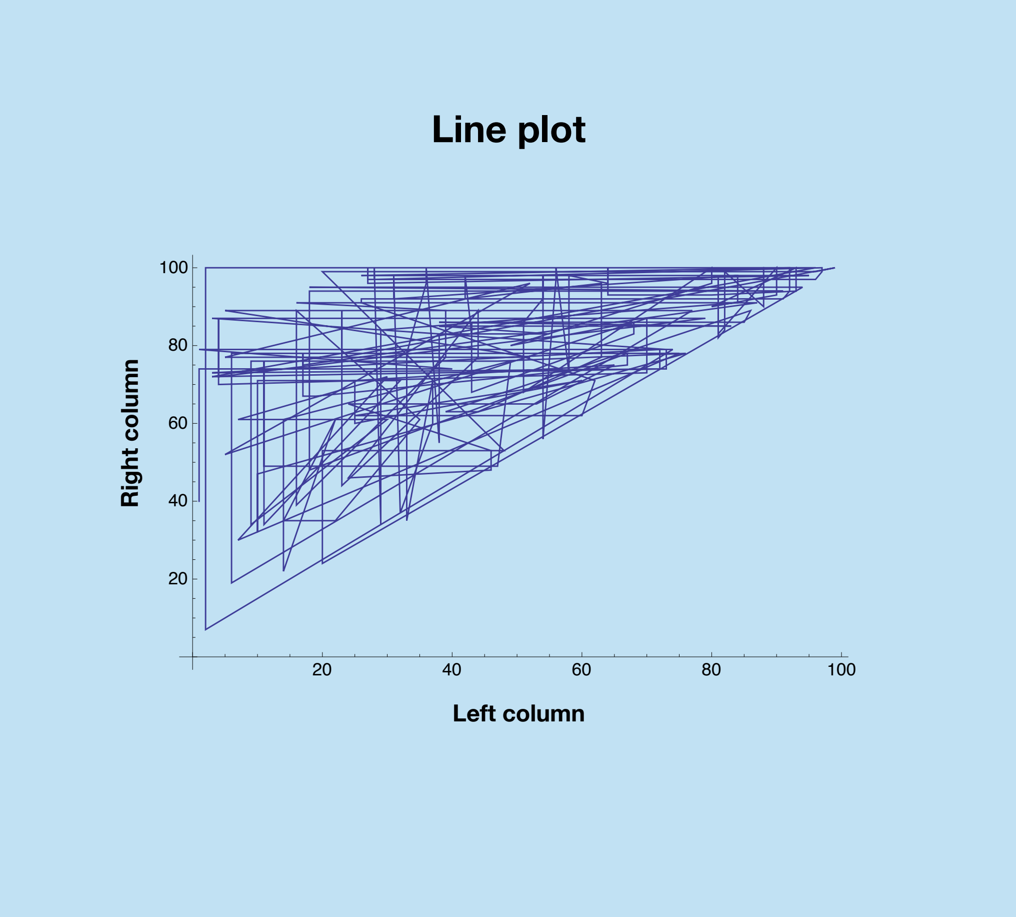
```
24      46
24      20
24      65
25      60
25      62
25      71
26      91
26      98
26      92
27      100
27      97
27      96
20      99
20      53
22      61
22      14
22      35
23      77
23      44
23      89
28      99
28      100
28      97
29      76
29      84
29      34
4       70
4       72
4       87
58      95
58      73
58      94
58      98
55      83
:
```



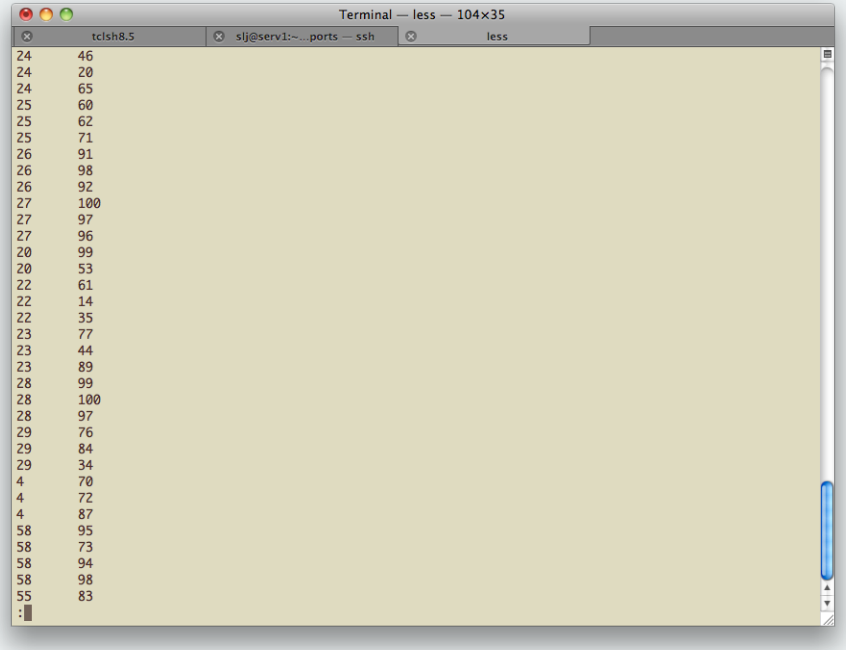
Relational data



```
24      46
24      20
24      65
25      60
25      62
25      71
26      91
26      98
26      92
27      100
27      97
27      96
20      99
20      53
22      61
22      14
22      35
23      77
23      44
23      89
28      99
28      100
28      97
29      76
29      84
29      34
4       70
4       72
4       87
58      95
58      73
58      94
58      98
55      83
:
```

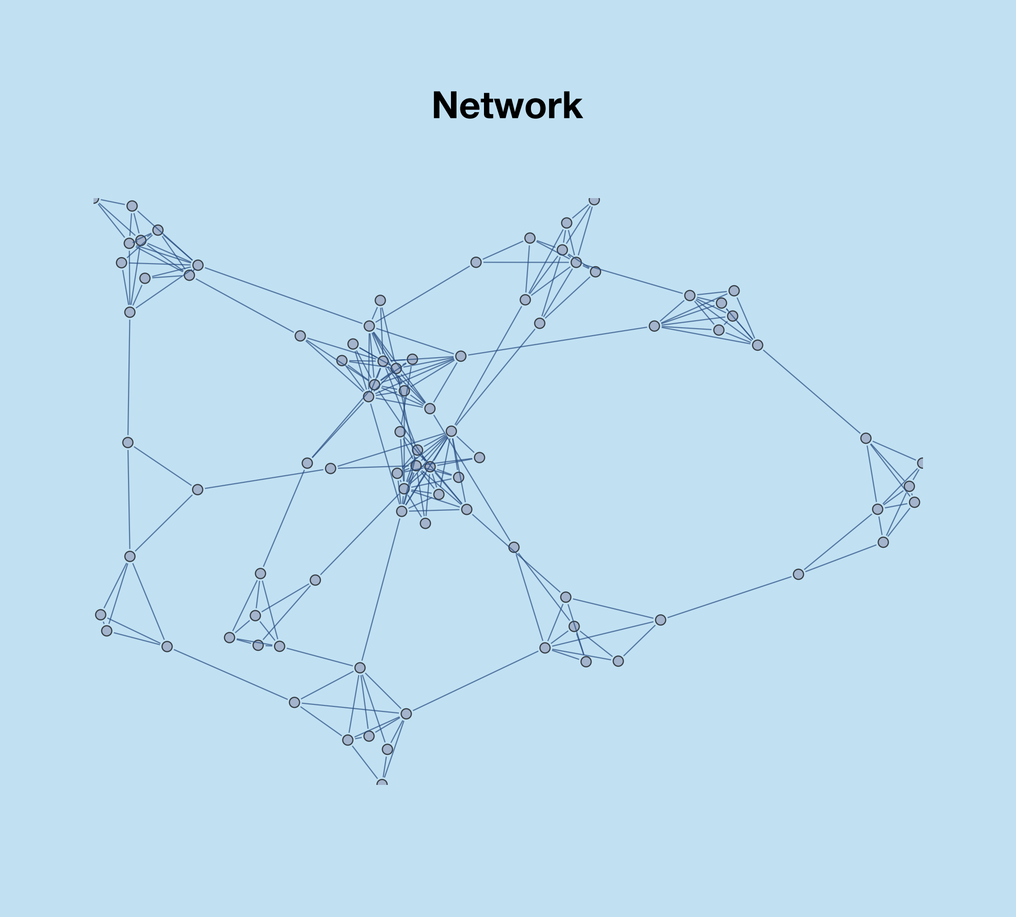


Relational data

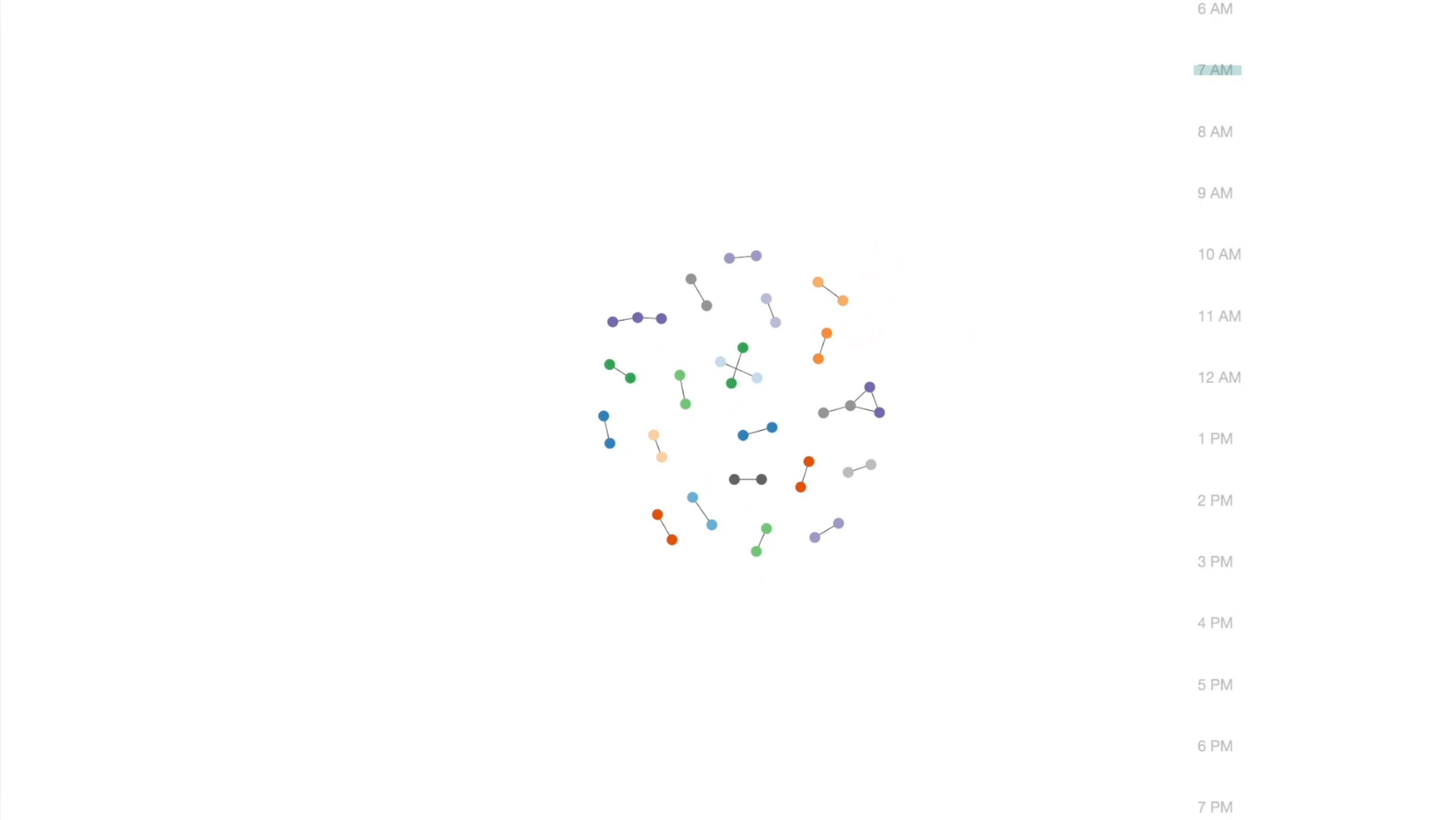


```
24      46
24      20
24      65
25      60
25      62
25      71
26      91
26      98
26      92
27      100
27      97
27      96
20      99
20      53
22      61
22      14
22      35
23      77
23      44
23      89
28      99
28      100
28      97
29      76
29      84
29      34
4       70
4       72
4       87
58      95
58      73
58      94
58      98
55      83
:
```

→



Very complex data that changes in time!



link

link

Linear algebra

Linear algebra

Tools for manipulating tabular data

Linear algebra

Tools for manipulating tabular data

Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

Linear algebra

Tools for manipulating tabular data

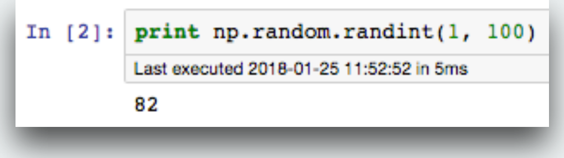
Objects

- 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

Tools for manipulating tabular data

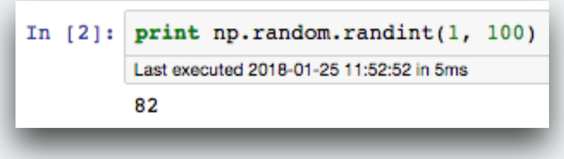
Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

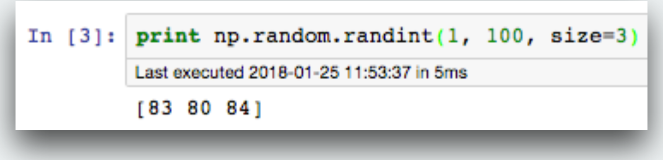
0D



```
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

Tools for manipulating tabular data

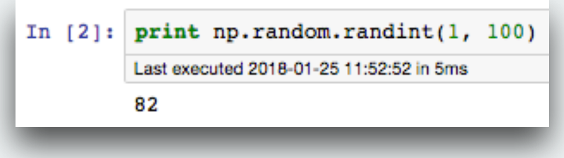
Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

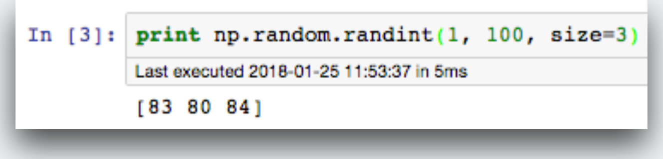
0D



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

vector

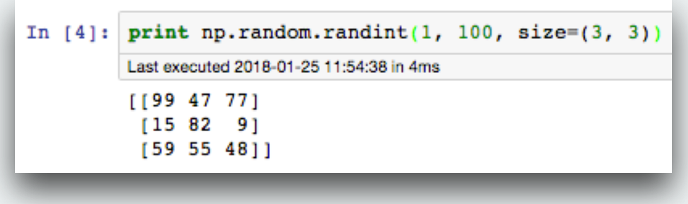
1D



```
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]]
```

Linear algebra

Tools for manipulating tabular data

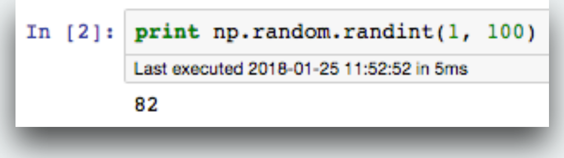
Objects

- Scalars
- Vectors
- Matrices

Everything is a Tensor!

scalar

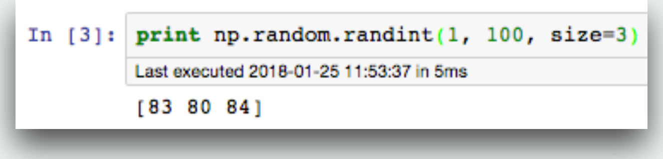
0D



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

vector

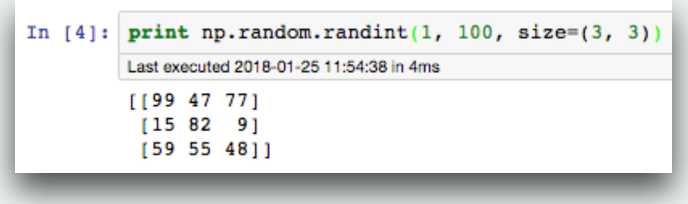
1D



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

matrix

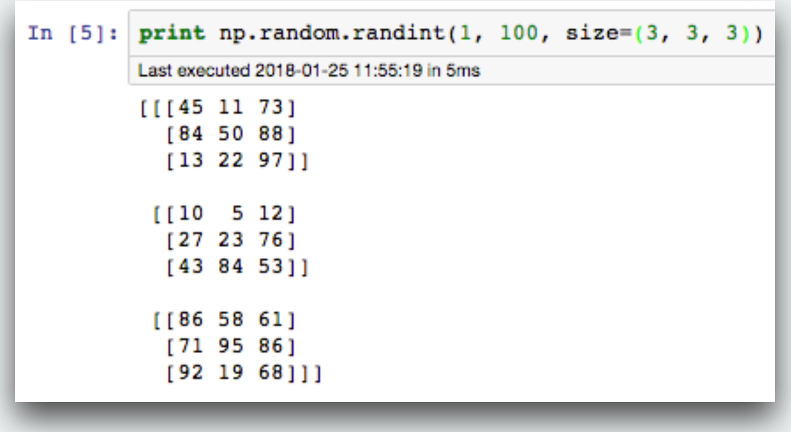
2D



```
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



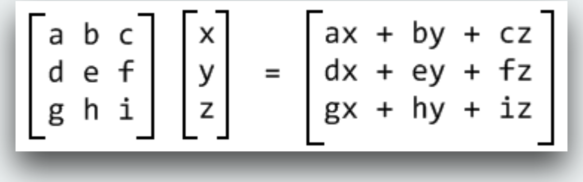
```
In [5]: print np.random.randint(1, 100, size=(3, 3, 3))
        Last executed 2018-01-25 11:55:19 in 5ms
        [[[45 11 73]
          [84 50 88]
          [13 22 97]]

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

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

Linear algebra

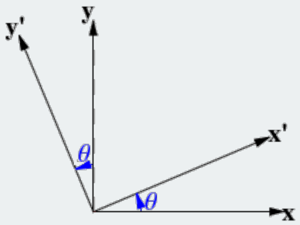
Tools for manipulating tabular data

Operations

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



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



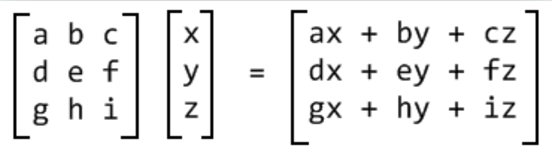
used frequently for basis transformation

Linear algebra

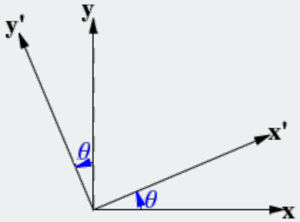
Tools for manipulating tabular data

Operations

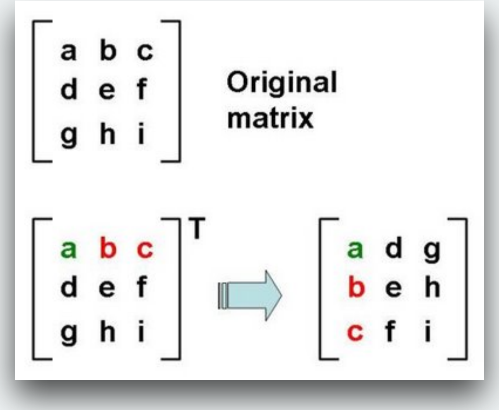
- Products: *dot*, *cross*
- Elementwise: *addition*, *subtraction*, *multiplication*, *division*
- Mutations: ***transpose***, *inverse/pseudo-inverse*, *scaling*, *rotation*



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



used frequently for basis transformation



$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

Original matrix

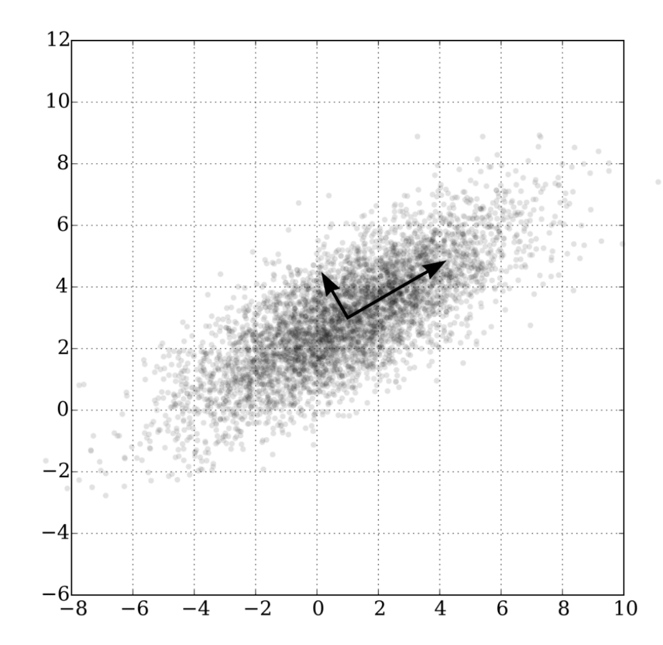
$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^T \Rightarrow \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}$$

Linear algebra

Tools for manipulating tabular data

Tools

- **Principal Component Analysis (PCA)**
- Archetypal Analysis
- Non-negative matrix factorization
- … many more



Statistics + PCA

Statistics

Framework for describing data

Vocabulary

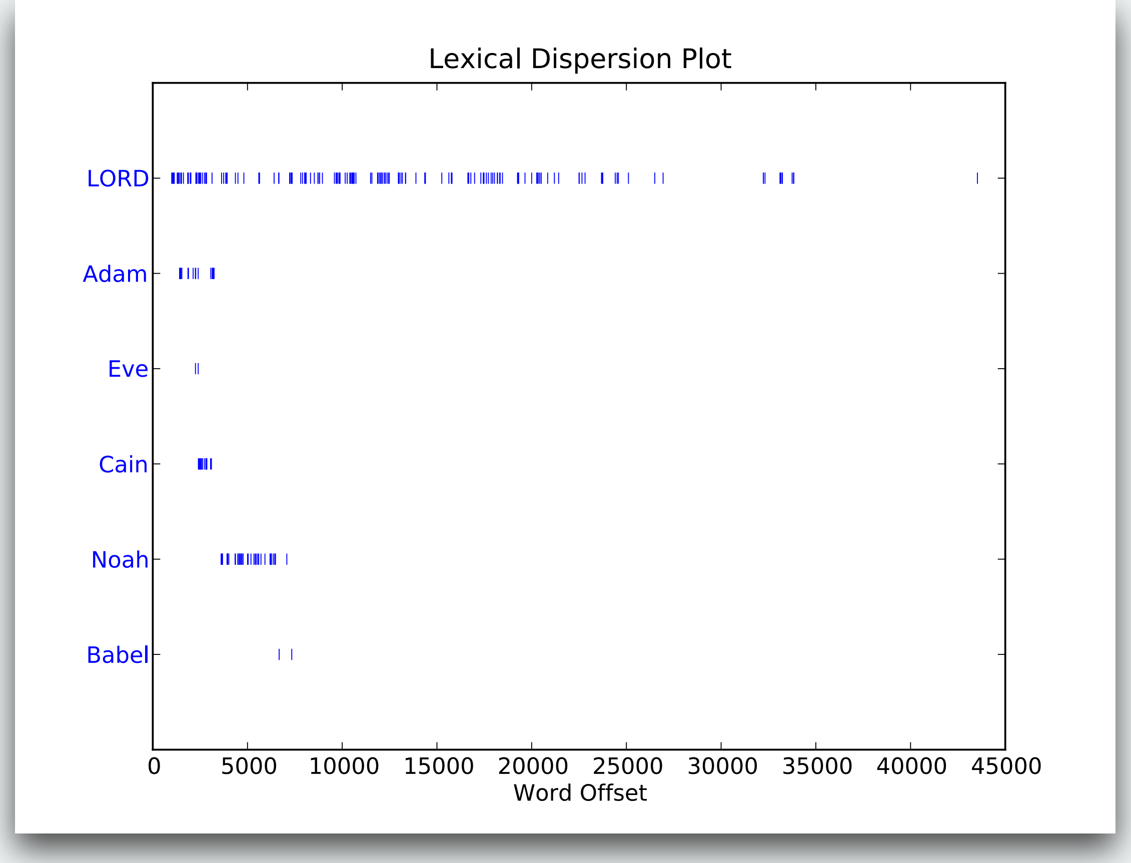
- Mean, median
- Variance, standard deviation, range
- Correlation, covariance

Statistics

Framework for describing data

Vocabulary

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



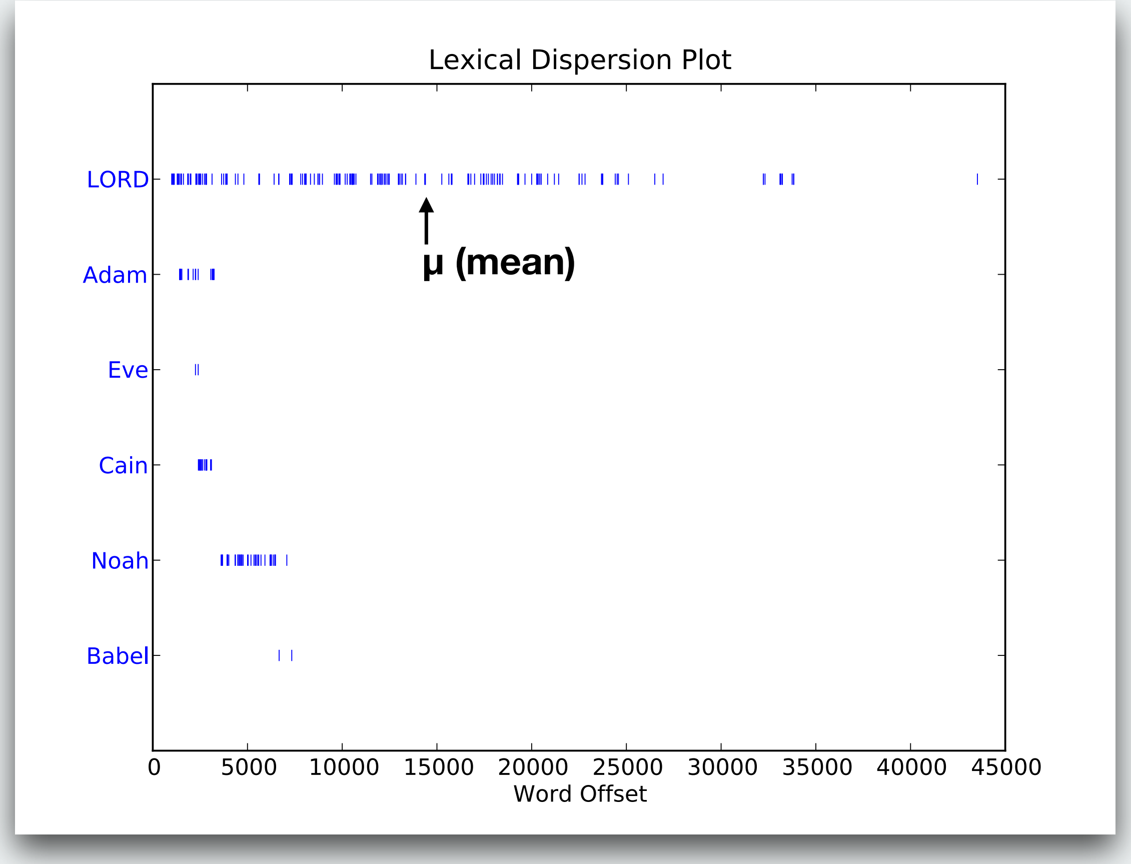
Statistics

Framework for describing data

Vocabulary

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

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



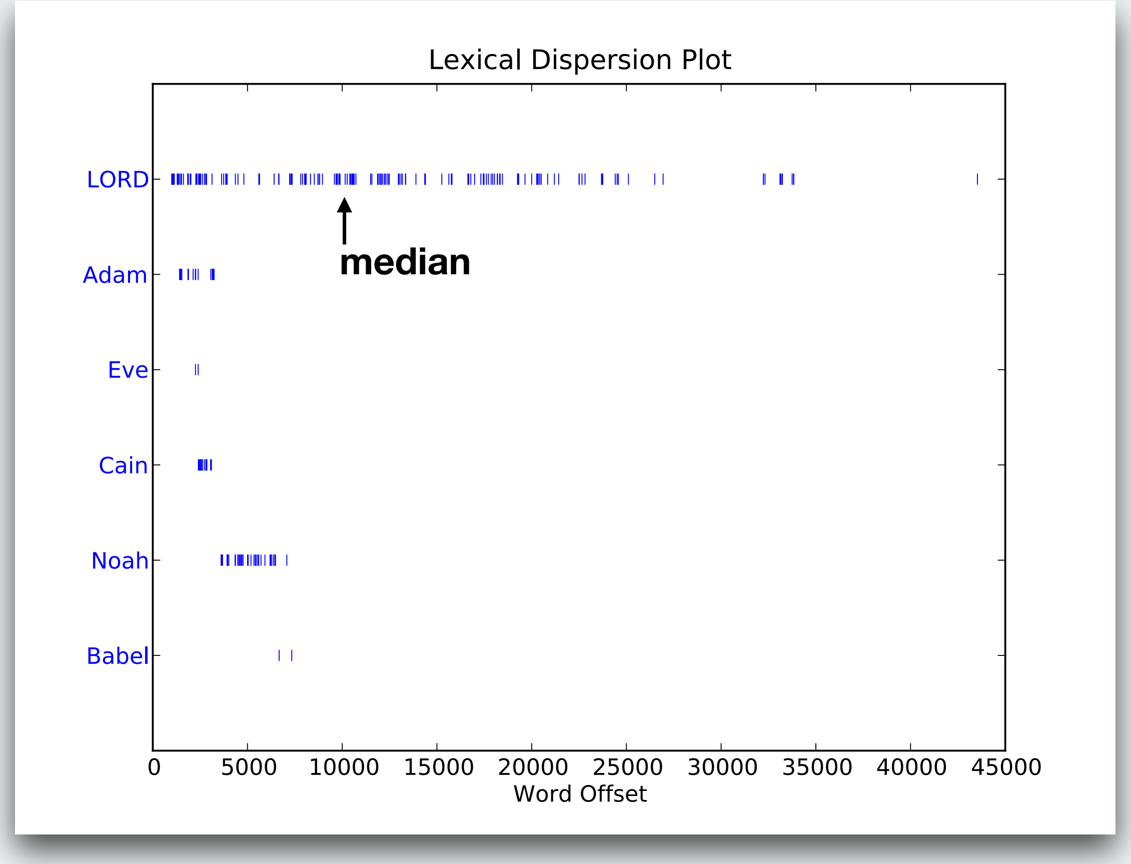
Statistics

Framework for describing data

Vocabulary

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

median = Middle number in ordered list



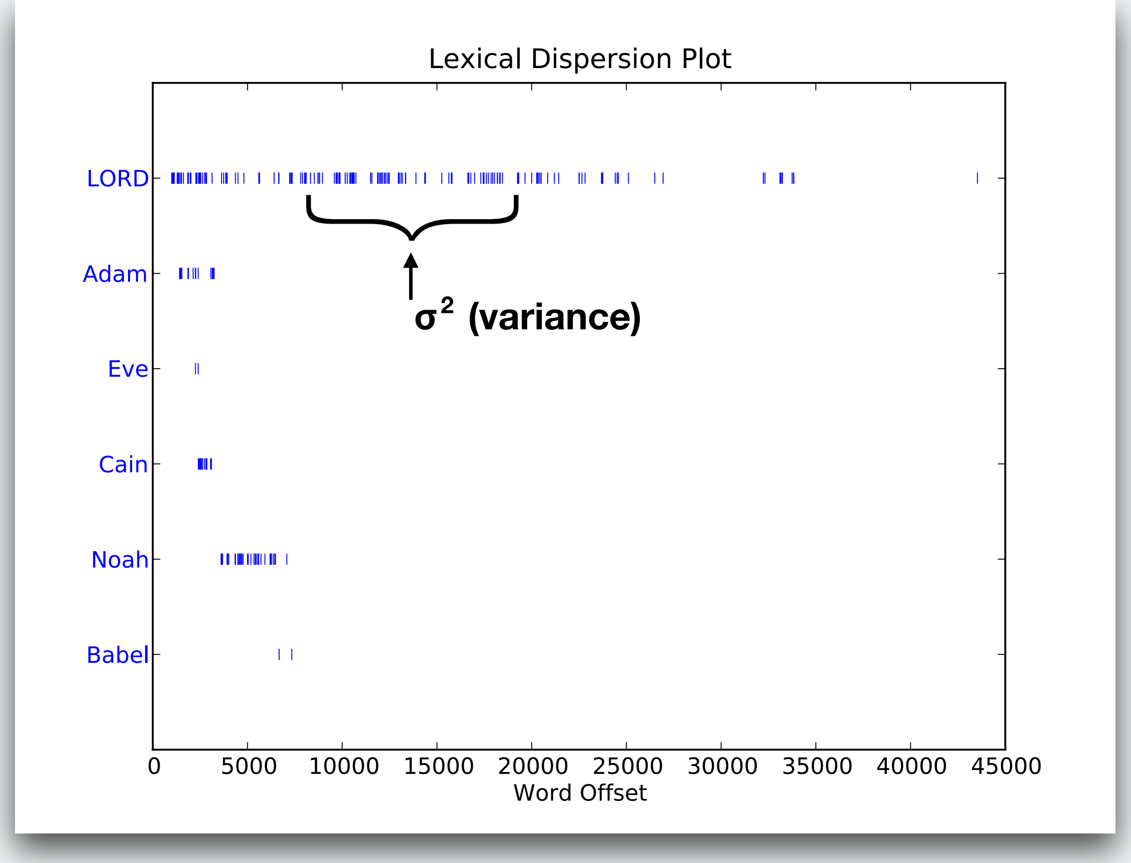
Statistics

Framework for describing data

Vocabulary

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

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



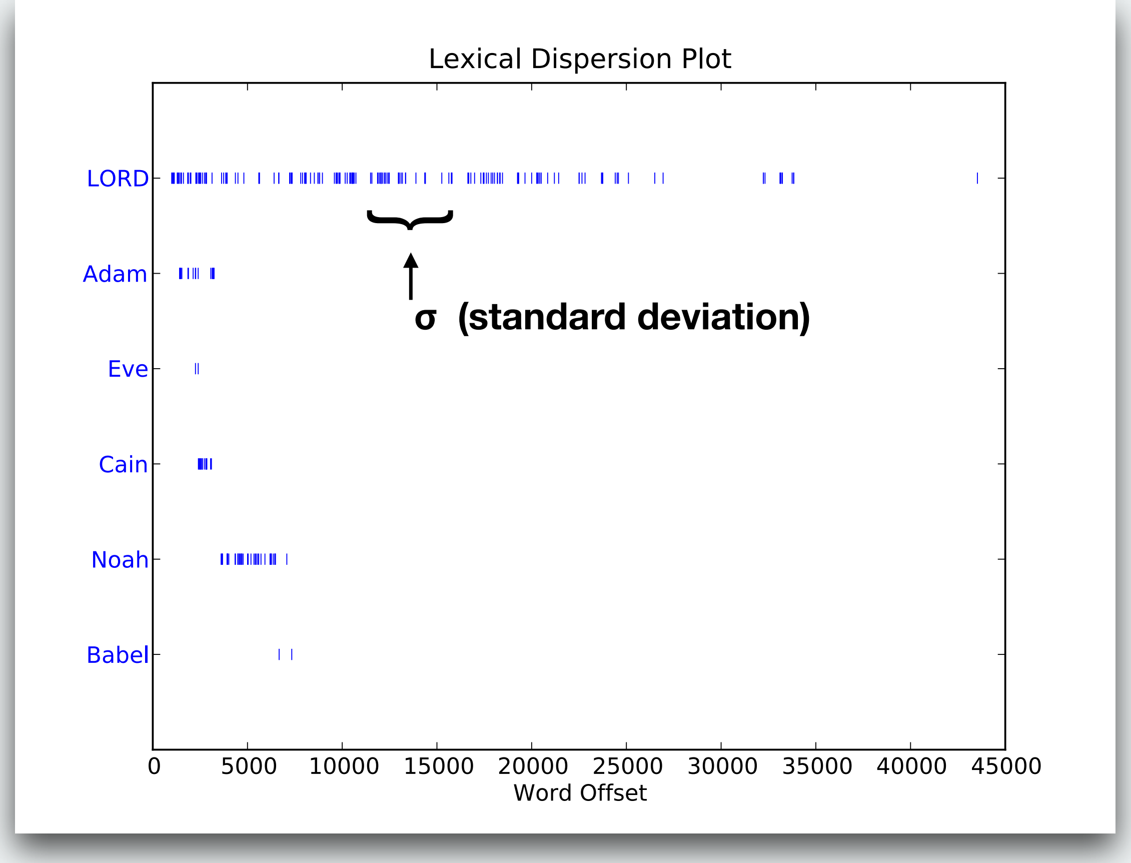
Statistics

Framework for describing data

Vocabulary

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

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



σ (standard deviation)

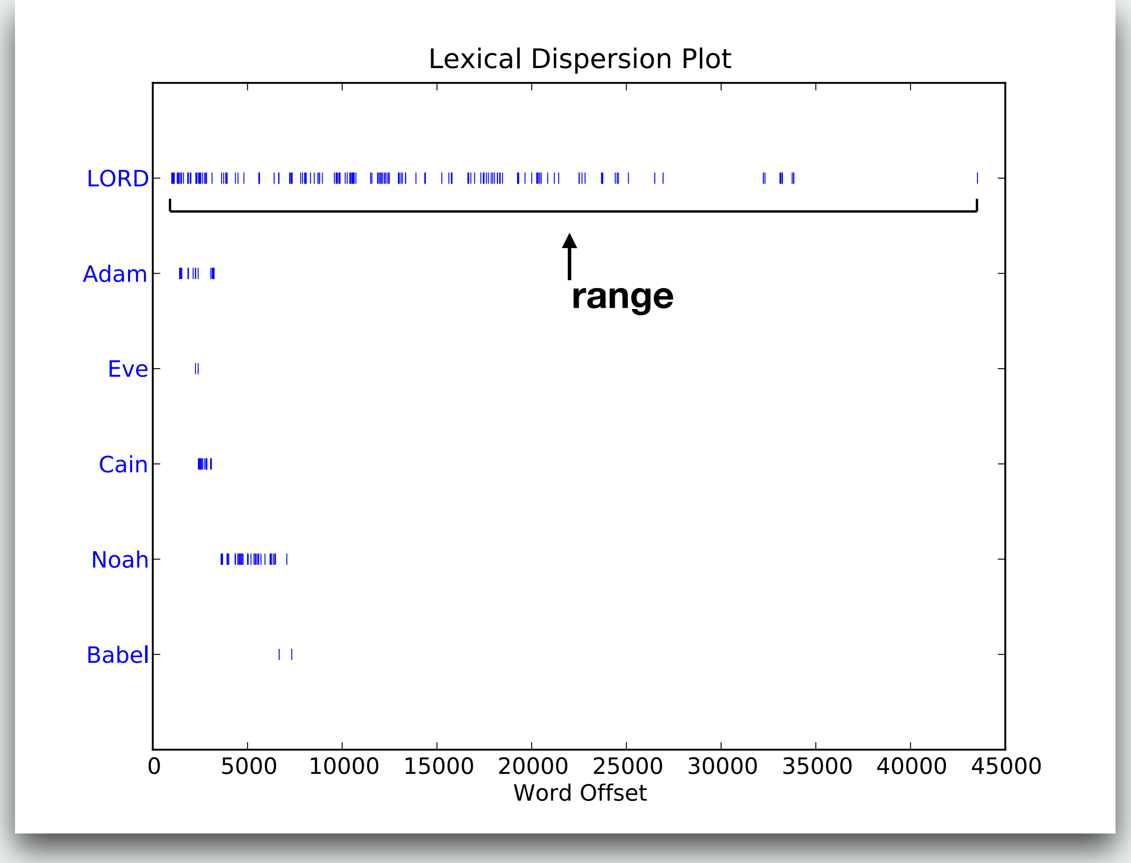
Statistics

Framework for describing data

Vocabulary

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

range = max(value) - min(value)



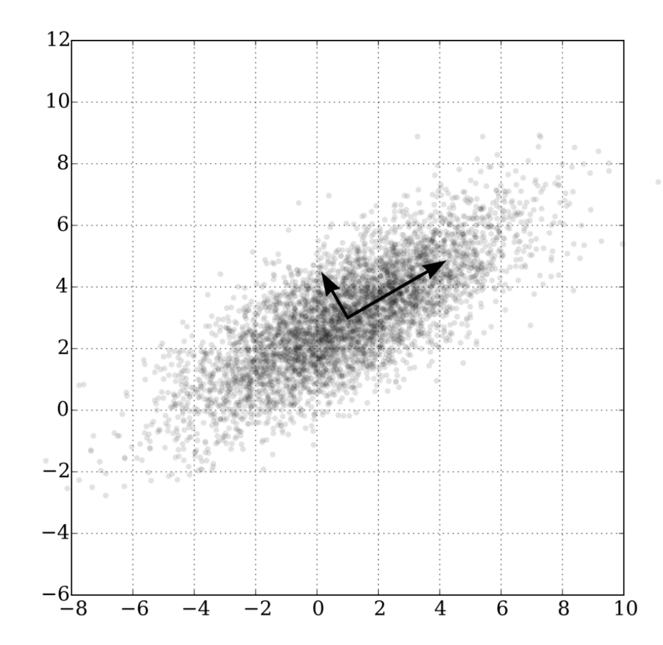
Statistics

Framework for describing data

Vocabulary

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

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



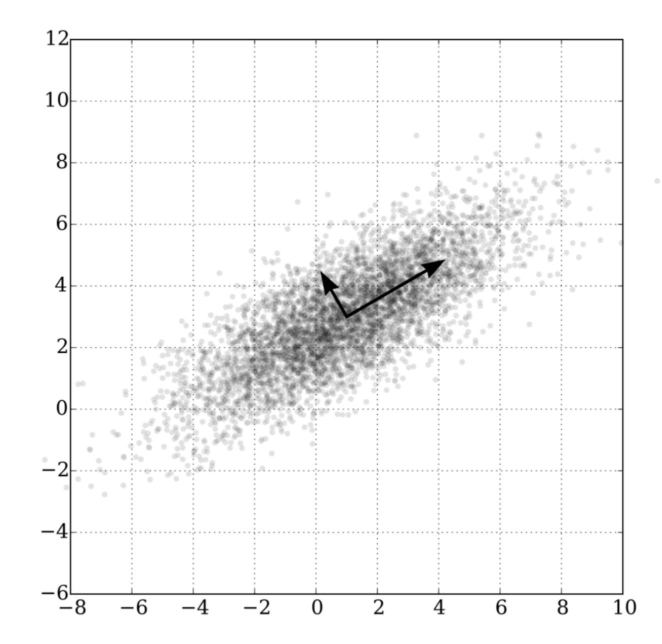
Statistics

Framework for describing data

Vocabulary

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

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



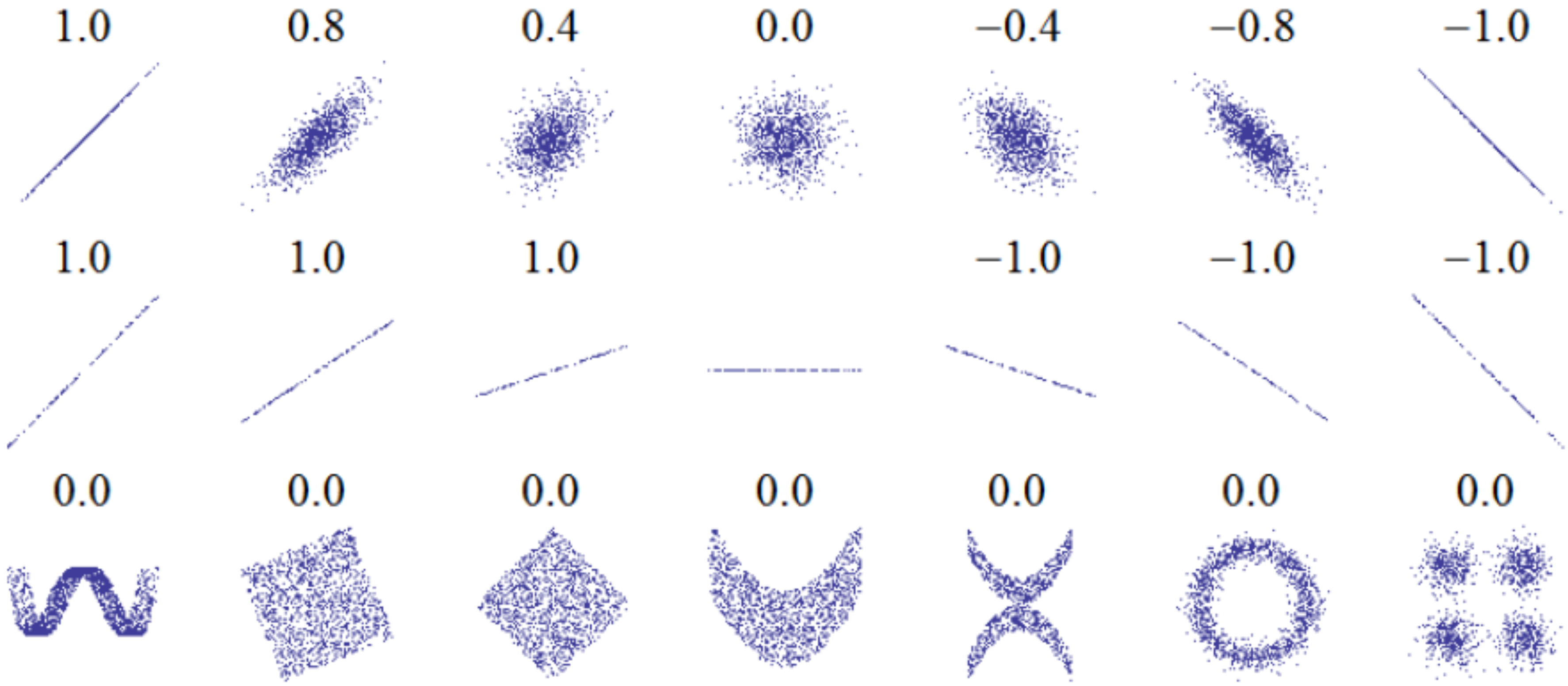


Figure 9.1: Examples of datasets with a range of correlations.

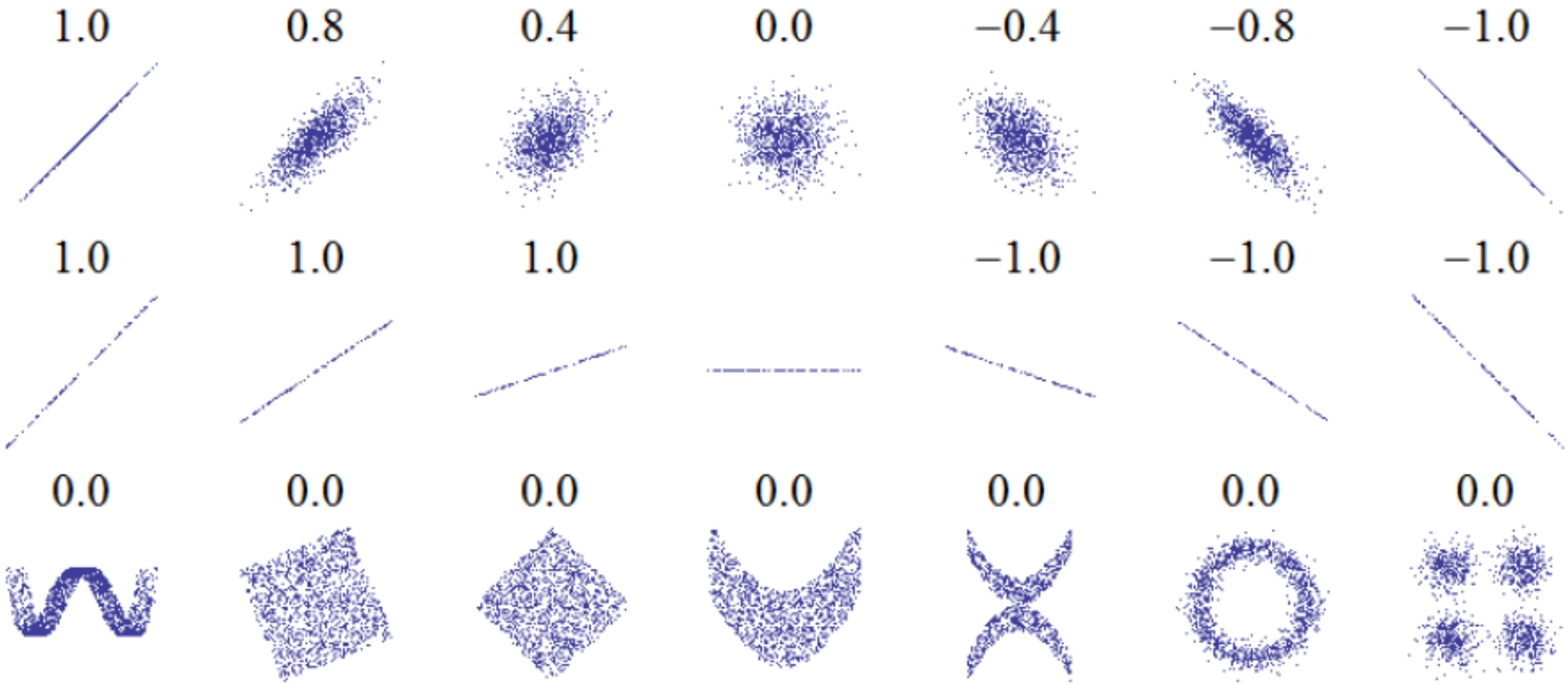


Figure 9.1: Examples of datasets with a range of correlations.

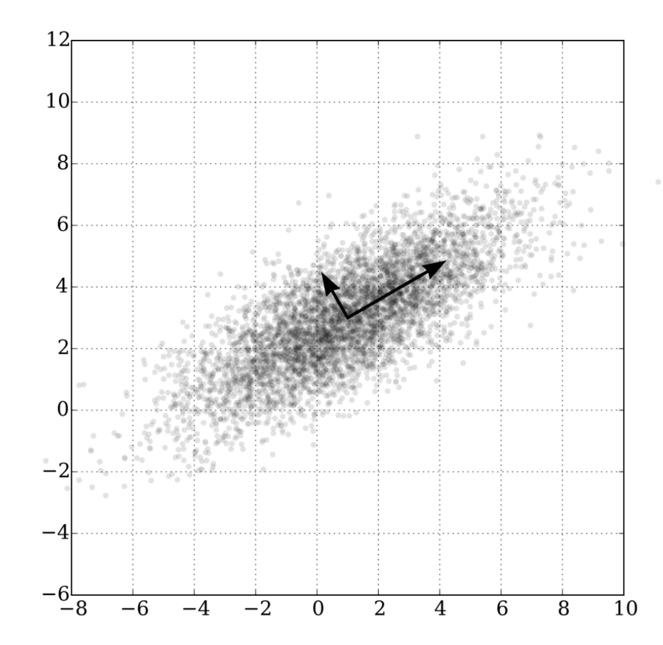
Attn: correlation only measures ***linear*** relationships!

Linear algebra

Tools for manipulating tabular data

Tools

- **Principal Component Analysis (PCA)**
  - **https://www.youtube.com/watch?v=g-Hb26agBFg**
- Archetypal Analysis
- Non-negative matrix factorization
- … many more



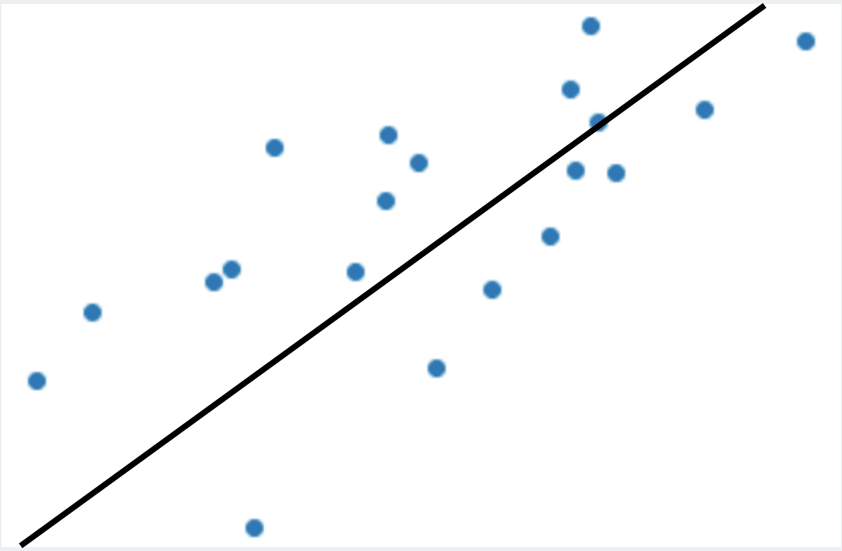
Hypothesis Testing

Statistics

Hypothesis Testing

- r = correlation coefficient

r = 0.6



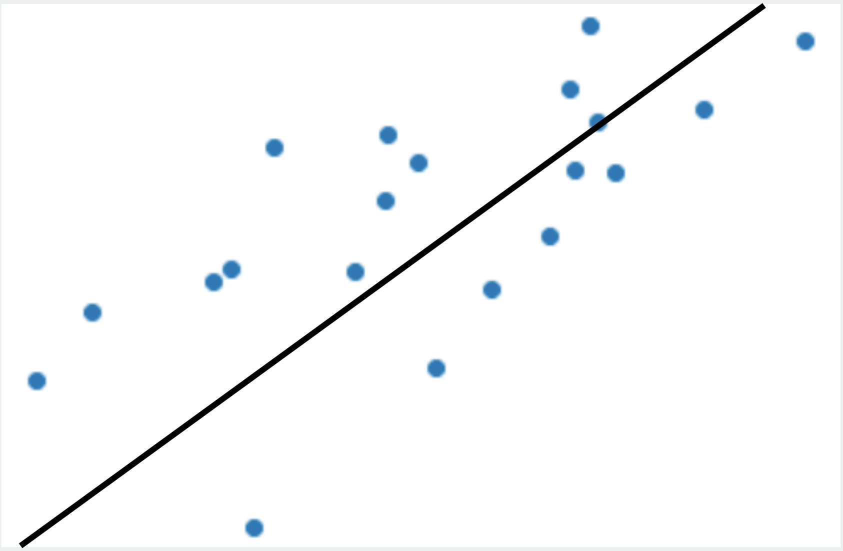
Statistics

Hypothesis Testing

r = 0.6 pfff…
This looks
random to me!

Null hypothesis

r = 0.6



Statistics

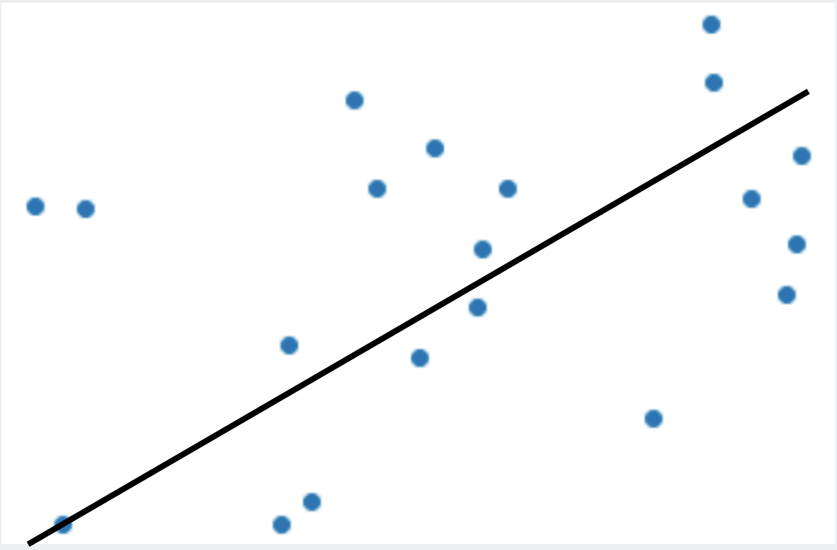
Hypothesis Testing

Null model

Randomization 1

r = 0.2? But this is just one example. Still could be random

Null hypothesis



r = 0.2

Statistics

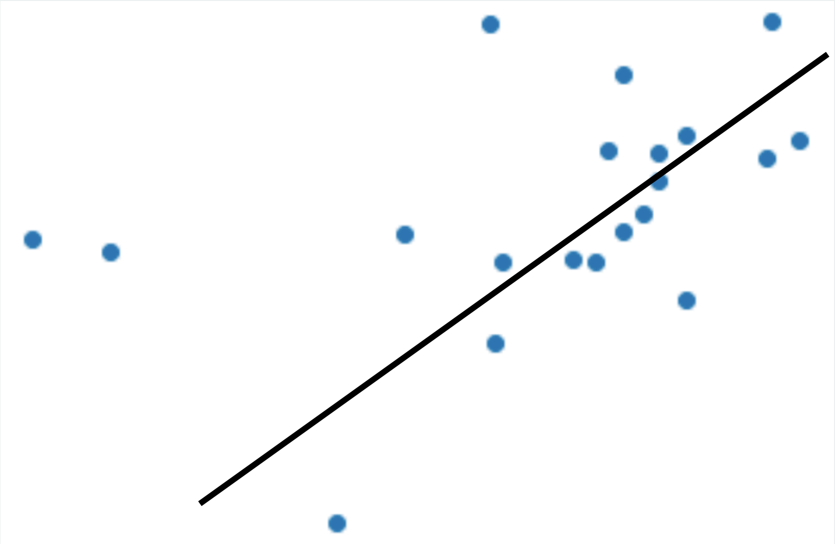
Hypothesis Testing

Null model

Randomization 2

r = 0.25? Pure chance, I still think your r=0.6 is random

Null hypothesis



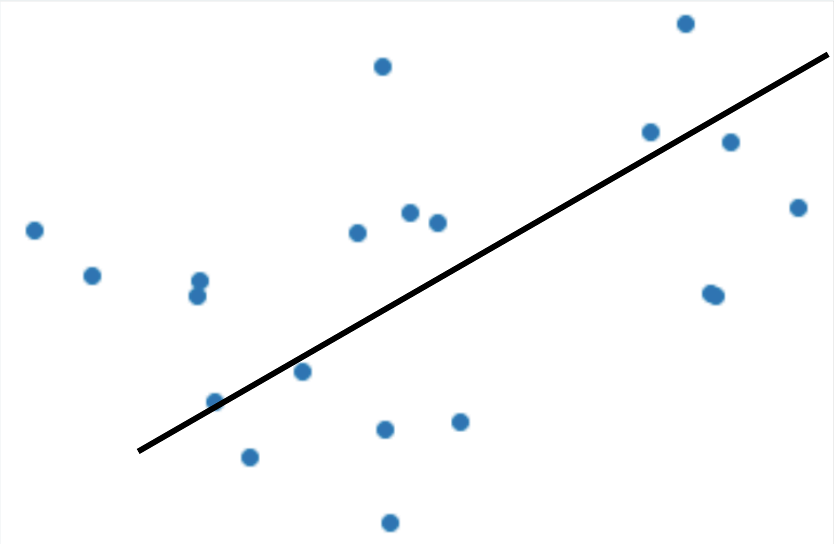
r = 0.25

Statistics

Hypothesis Testing

Null model

Randomization 1000



r = 0.3

Ok ok I give up, random almost always gives worse correlation

Null Hypothesis

REJECTED