

# Computational Analysis of Big Data

Week 2

## A Data Scientist's most fundamental tools

More specifically: Visualization, linear algebra and statistics

Visualization ○ ○ ○ ○ ○

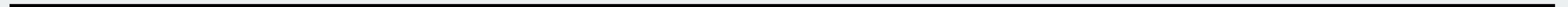
Linear algebra ○ ○

Statistics ○

1  
1000

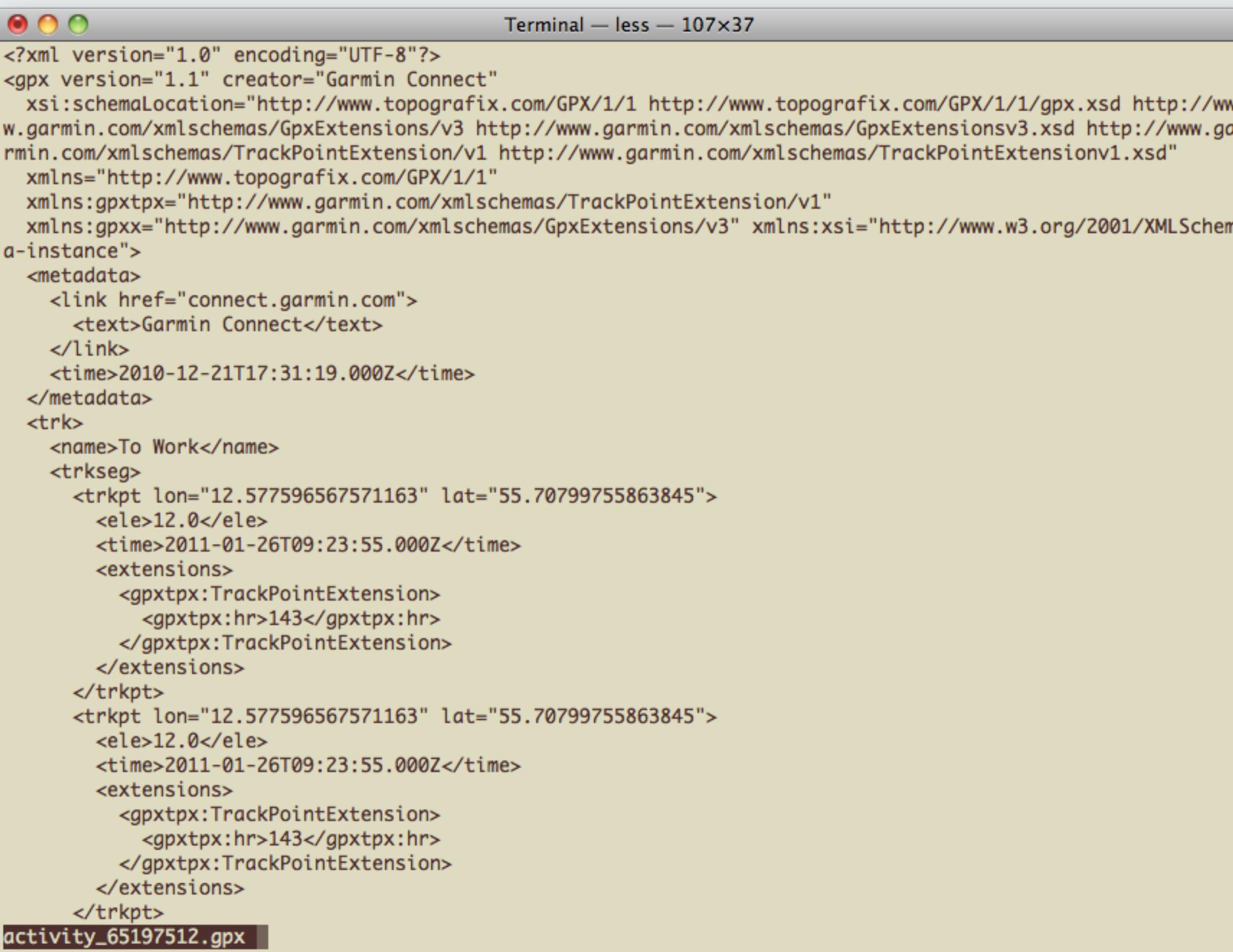
1  
1000

.



# This is data

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



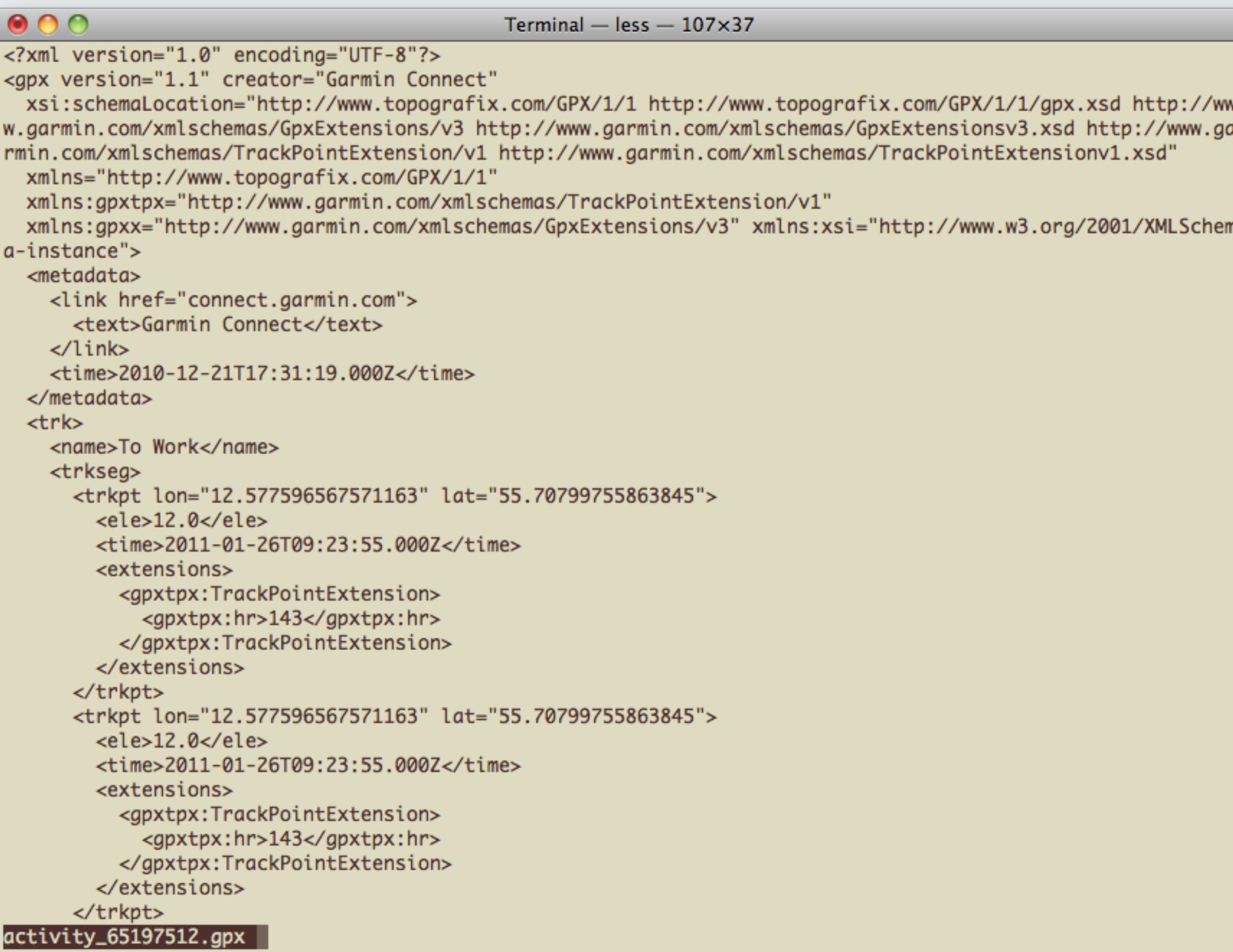
The screenshot shows a terminal window titled "Terminal — less — 107x37". The window displays the contents of a GPX (GPS Exchange Format) file named "activity\_65197512.gpx". The XML code is as follows:

```
<?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>
    </trkseg>
  </trk>
</gpx>
```

The terminal window has a scroll bar on the right side. The file name "activity\_65197512.gpx" is visible at the bottom of the terminal window.

# This is GPS data

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



A screenshot of a terminal window titled "Terminal — less — 107x37". The window displays an XML file named "activity\_65197512.gpx". The XML code describes a GPS track with two points. The first point is at coordinates (12.577596567571163, 55.70799755863845) with elevation 12.0 and time 2011-01-26T09:23:55.000Z. It includes a "gpxtpx:TrackPointExtension" element with a "gpxtpx:hr" value of 143. The second point is identical. The file ends with the closing tag of the first point.

```
<?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>
    </trkseg>
  </trk>
</gpx>
```

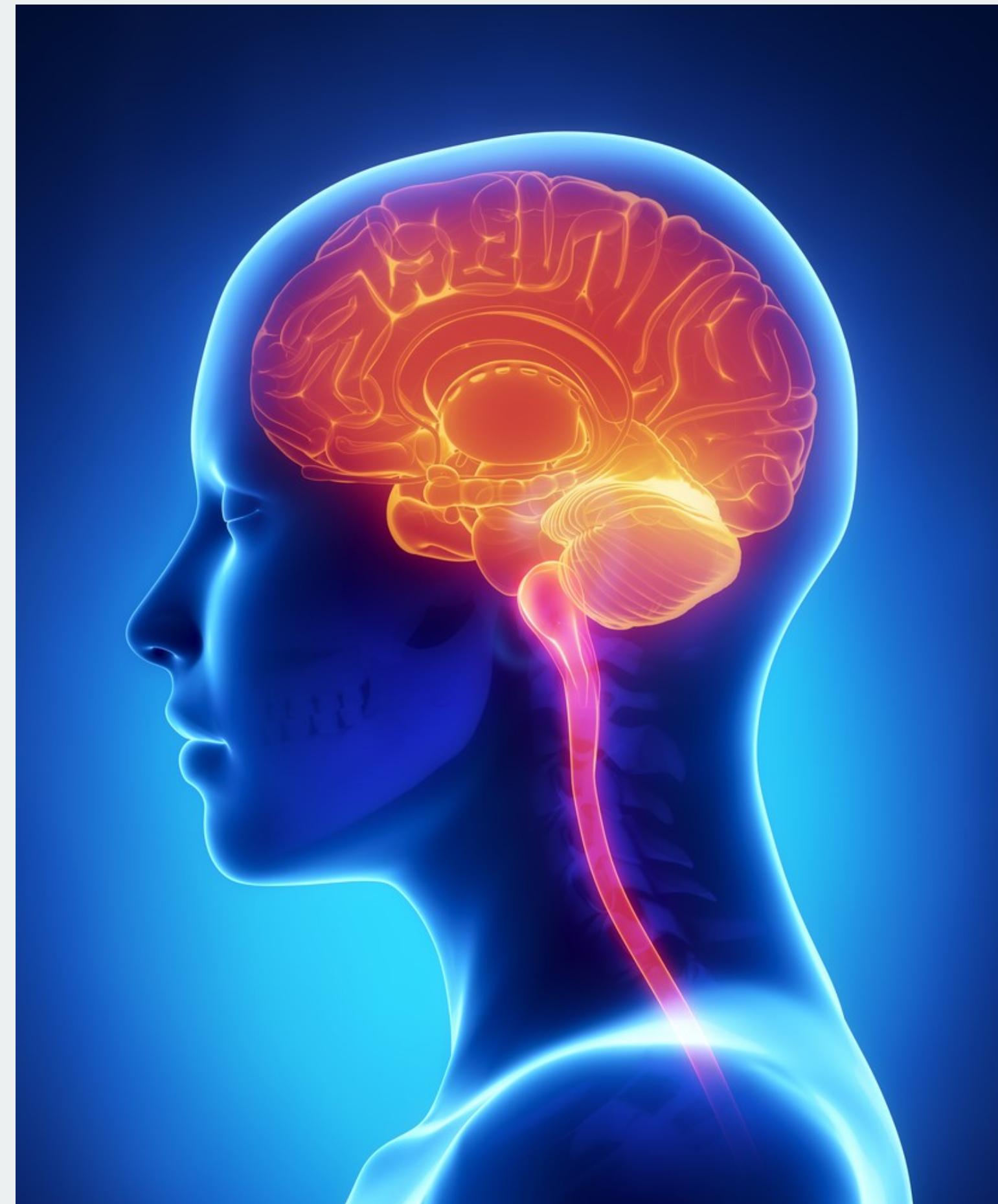
activity\_65197512.gpx

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

# Most raw data is incomprehensible to humans

## We have:

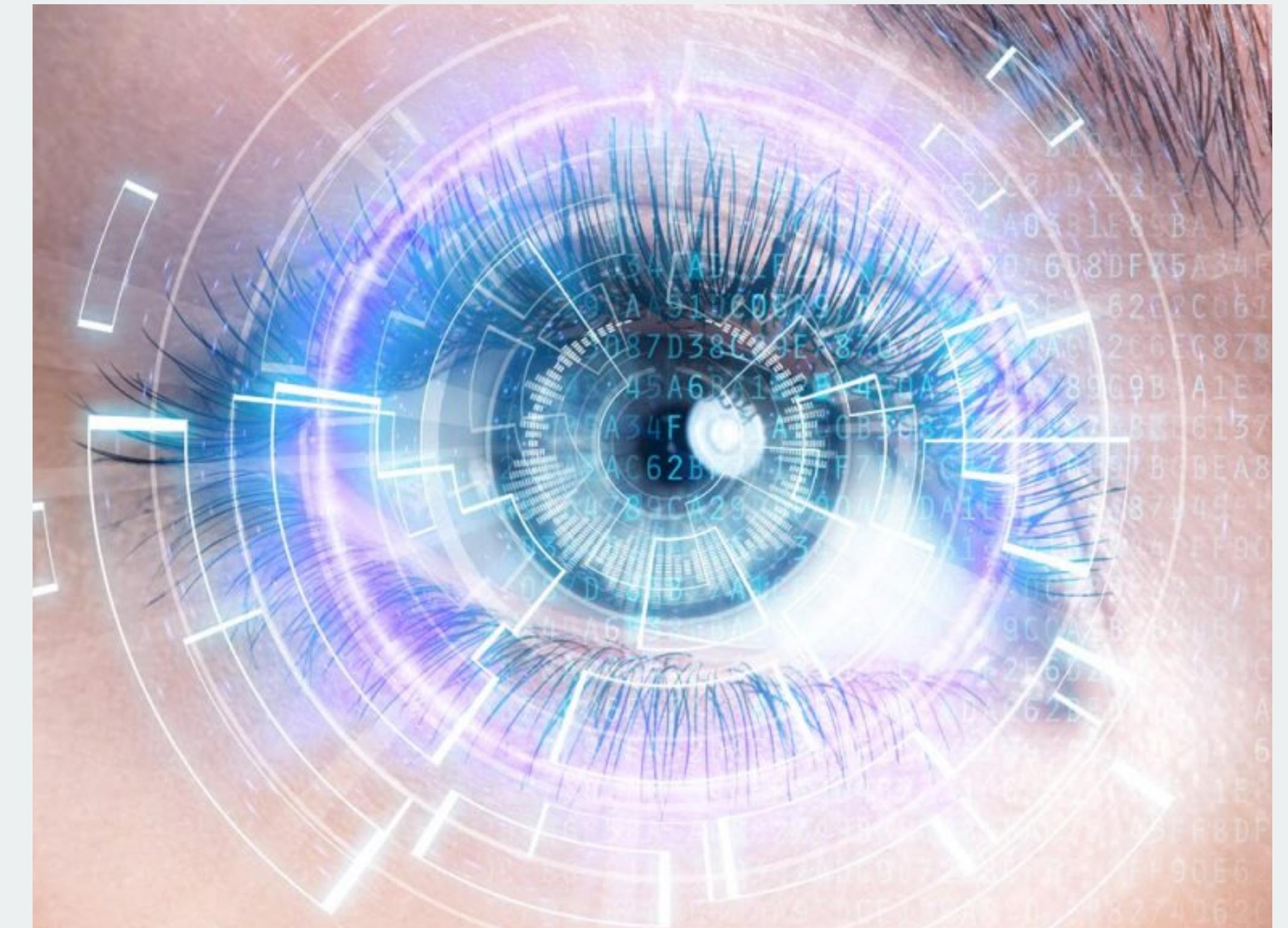
- Narrow spectrum of data that we can process and understand
- Limited memory for processing new information
- Limited attention for undertaking focussed tasks



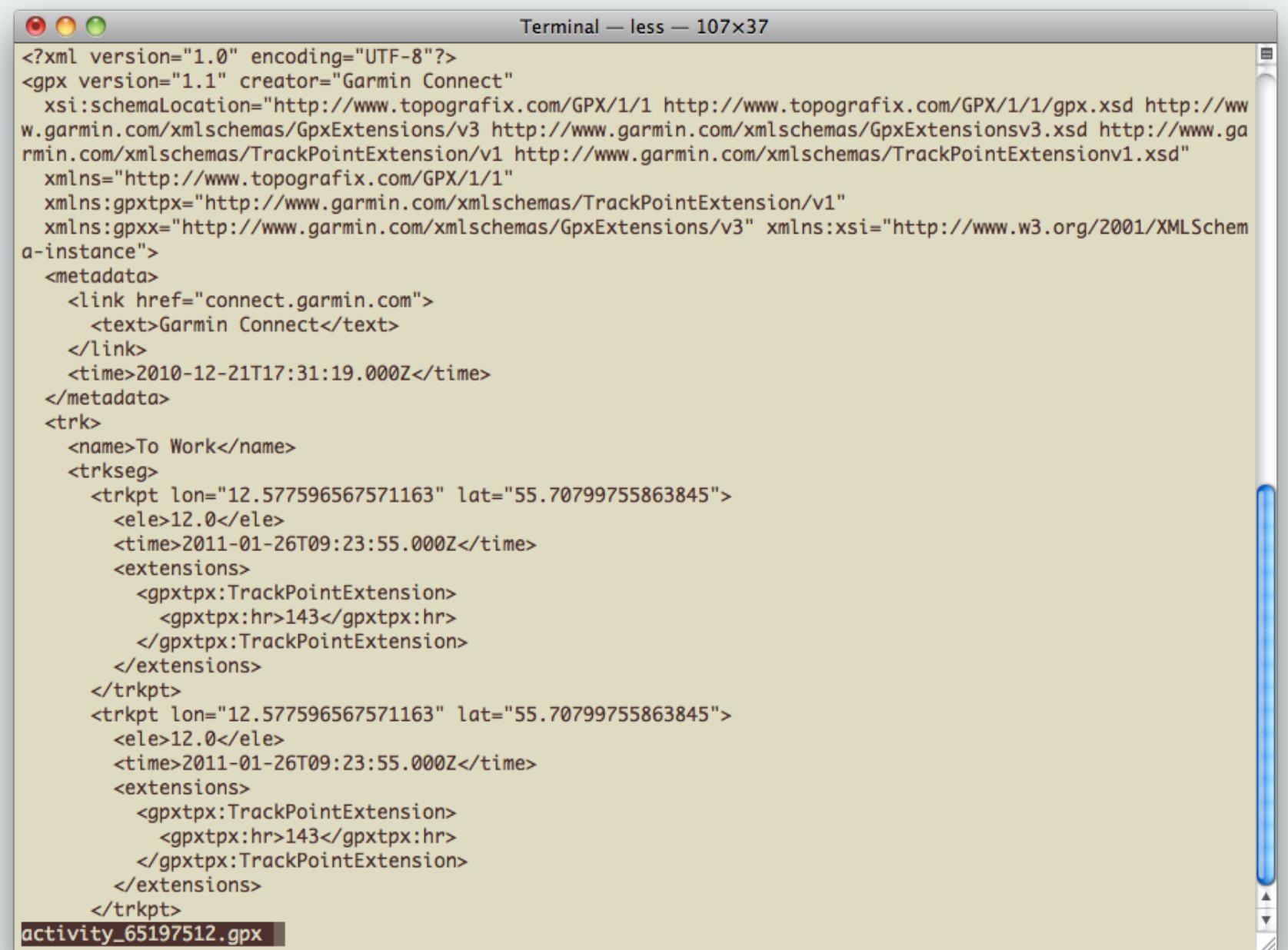
# The human eye is made for advanced pattern recognition

**It can:**

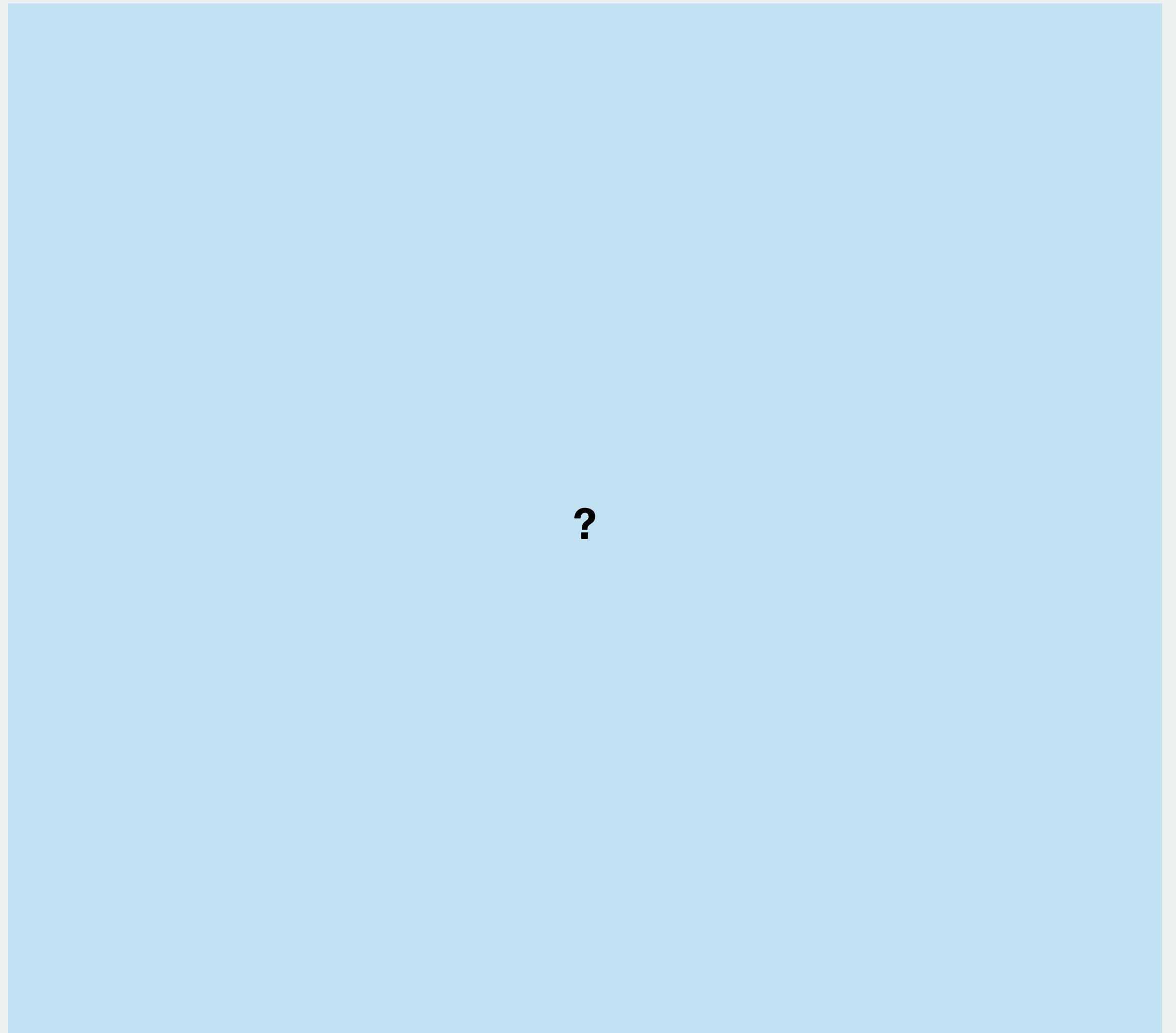
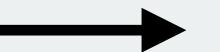
- Immediately **recognize patterns** in highly complex images
- Notice **outliers**
- Process streams of images and recognize **patterns over time**



# Data must be rendered in human-friendly format



A screenshot of a Mac OS X terminal window titled "Terminal — less — 107x37". The window displays a large block of XML code representing a GPX file. The XML includes metadata like the creator ("Garmin Connect") and a timestamp ("2010-12-21T17:31:19.000Z"). It also contains track segments with points, each with coordinates (lon, lat), elevation (ele), time, and extensions (gpxtpx:TrackPointExtension). The file is named "activity\_65197512.gpx".



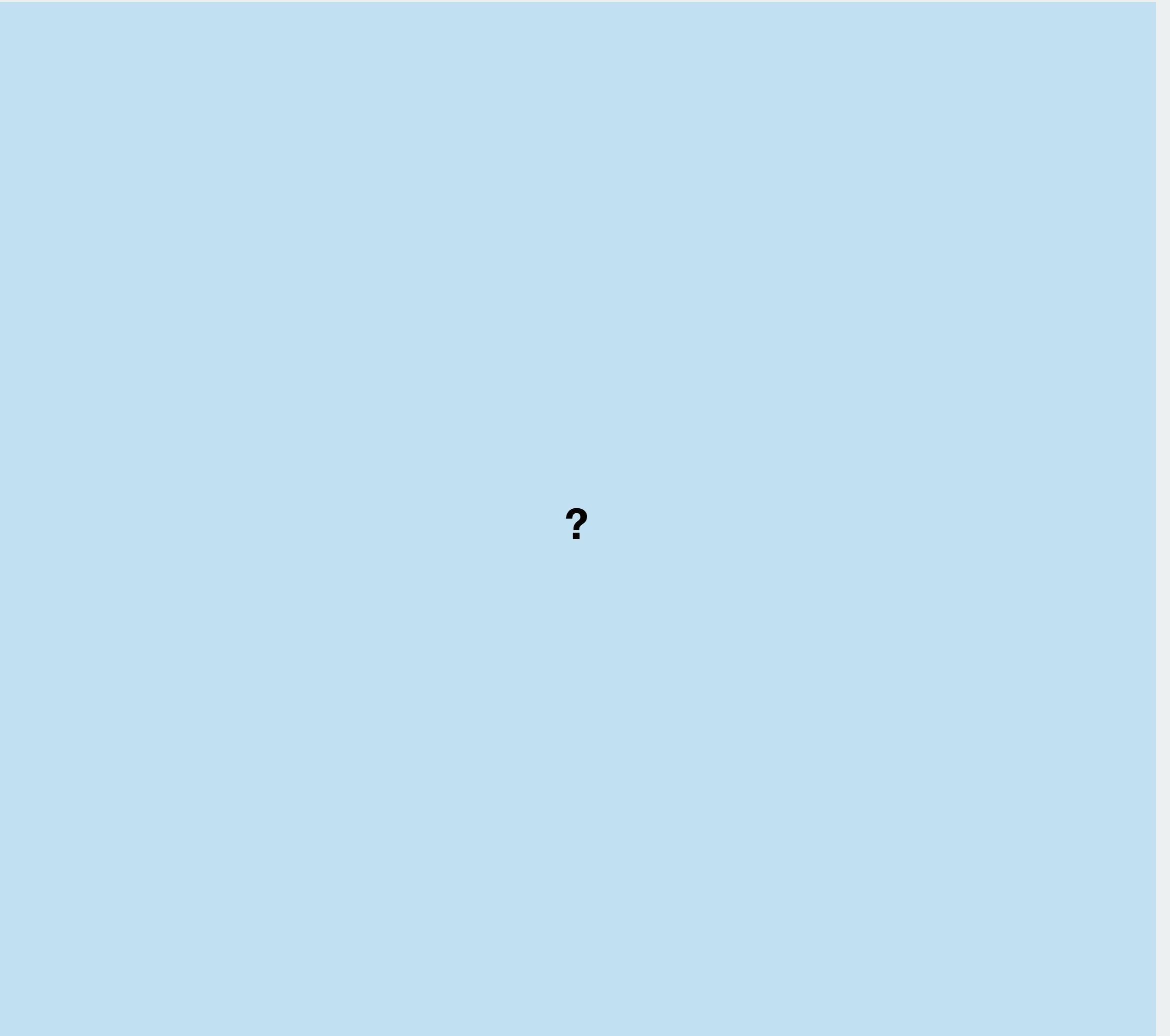
# Data must be rendered in human-friendly format

```
Terminal — less — 107x37
<?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:gpxpx="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>
    </trkseg>
  </trk>

```



	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
...		



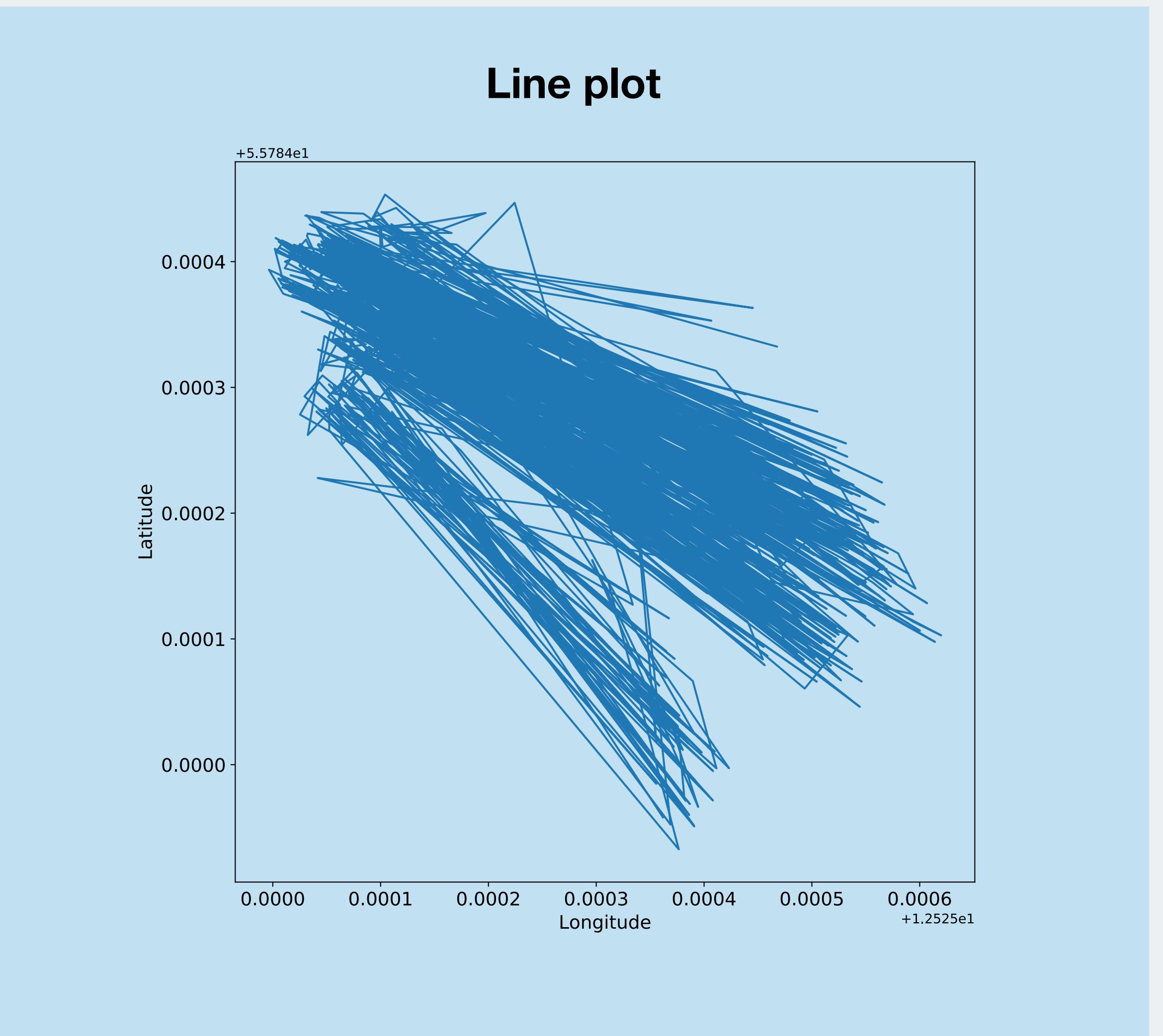
# Data must be rendered in human-friendly format

Terminal — less — 107x37

```
<?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:gpxpx="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>
    </trkseg>
  </trk>
</gpx>
```

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
...	...	...



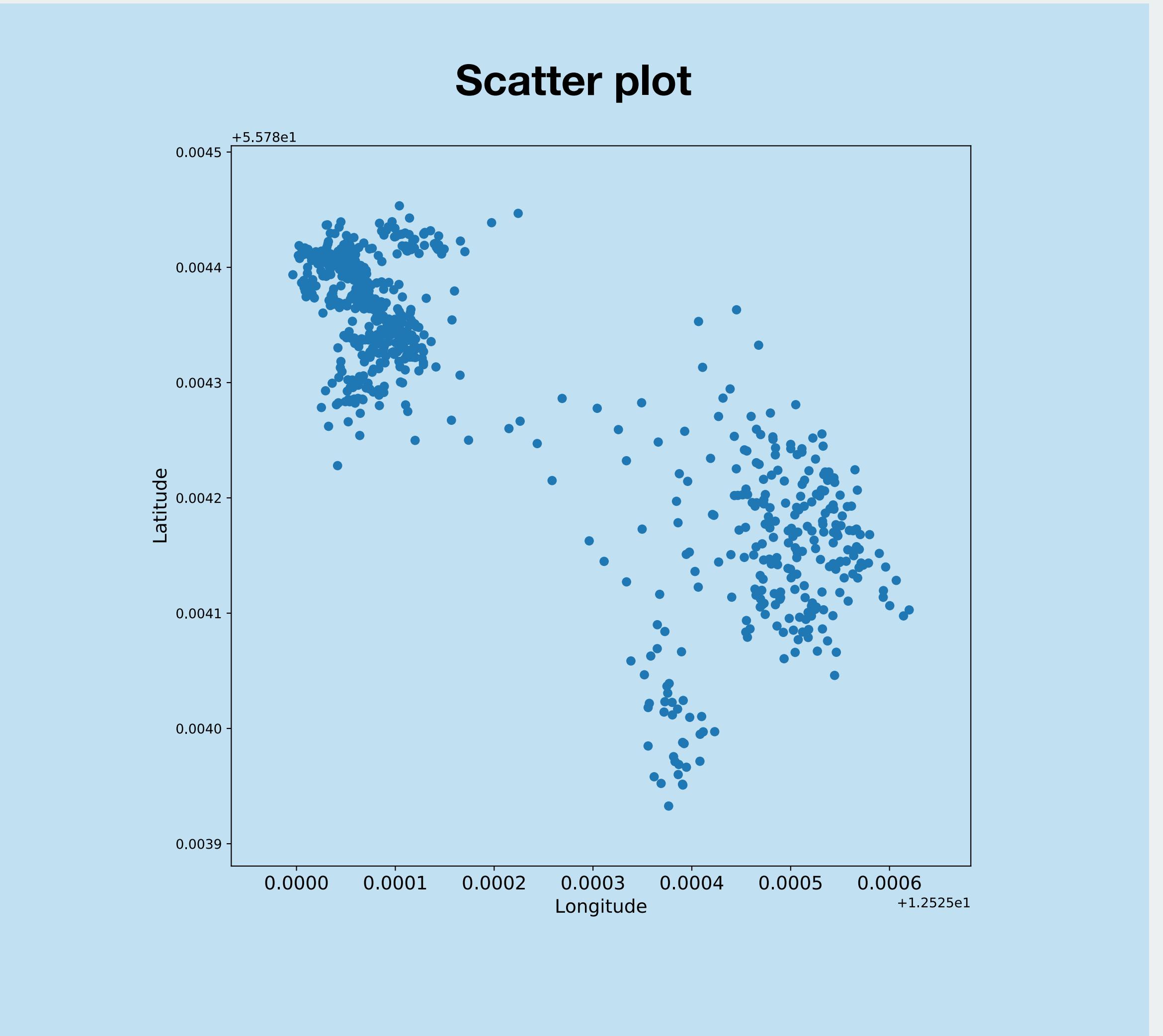
# Data must be rendered in human-friendly format

Terminal — less — 107x37

```
<?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.w3.org/2001/XMLSchema.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:gpxpx="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>
    </trkseg>
  </trk>
</gpx>
```

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
...	...	...



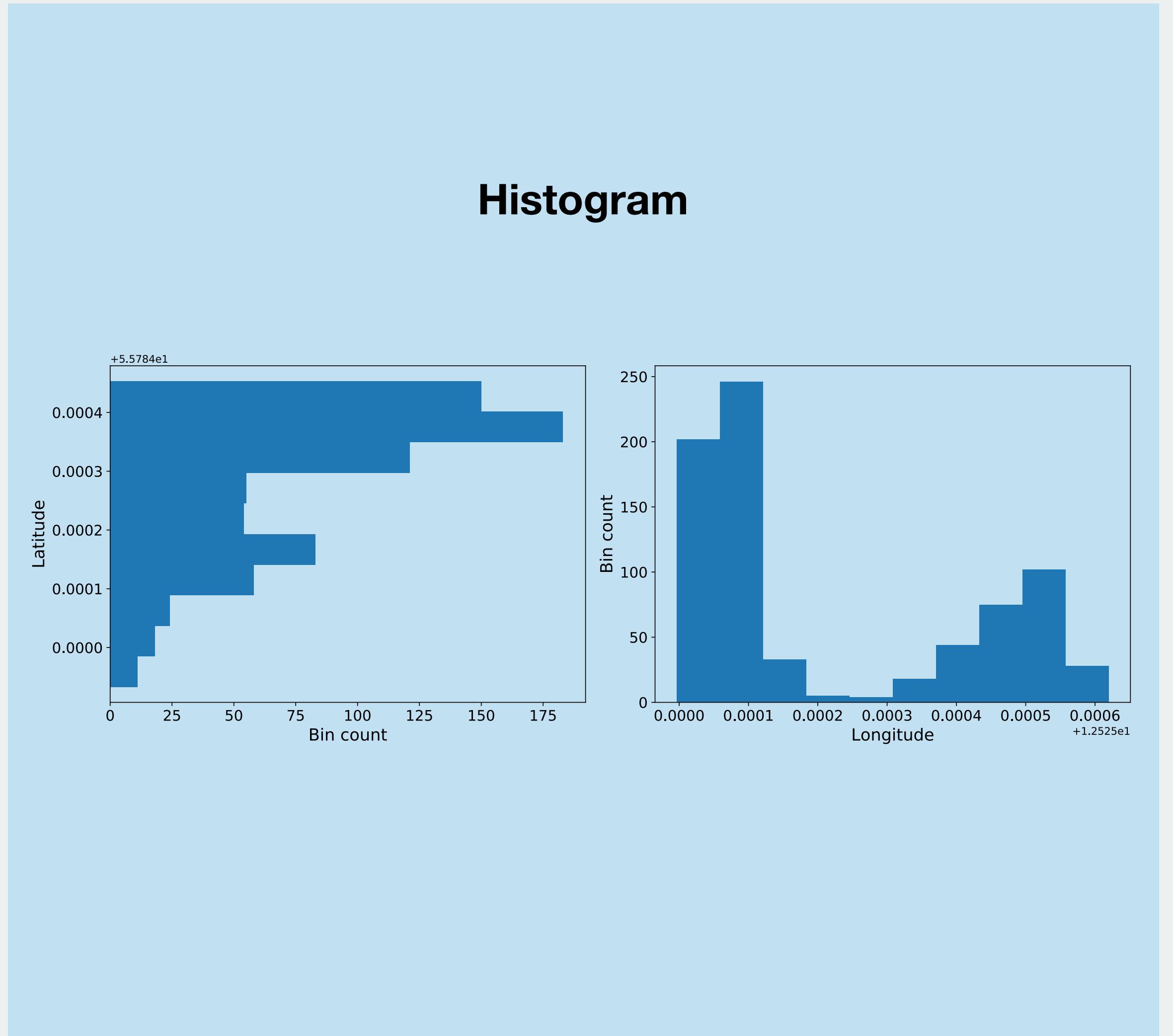
# Data must be rendered in human-friendly format

Terminal — less — 107x37

```
<?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:gpxpx="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>
    </trkseg>
  </trk>
</gpx>
```

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

Terminal — less — 107x37

```
<?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:gpxpx="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>
    </trkseg>
  </trk>
</gpx>
```

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
...	...	...



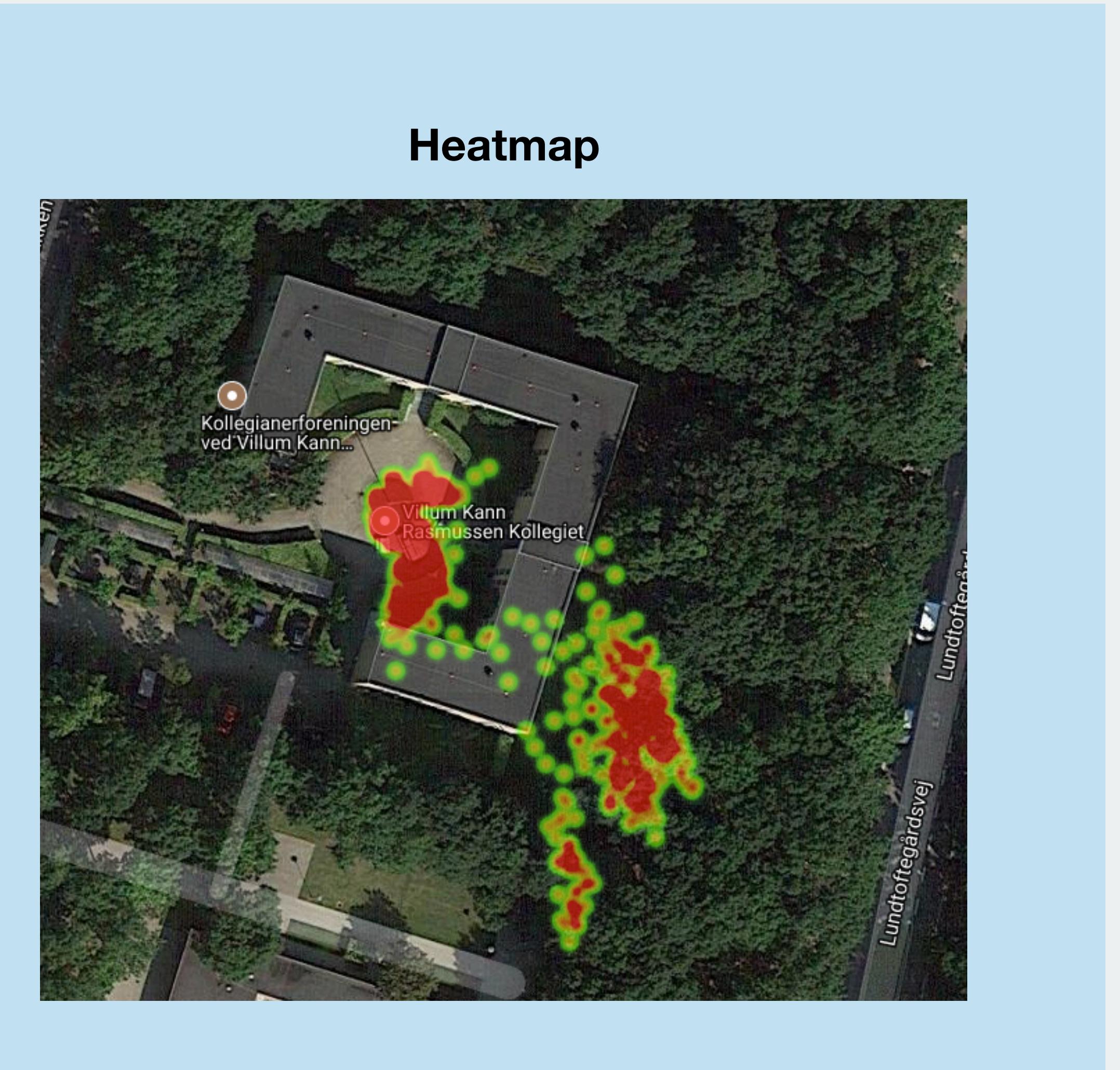
# Data must be rendered in human-friendly format

Terminal — less — 107x37

```
<?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:gpxpx="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>
    </trkseg>
  </trk>
</gpx>
```

activity\_65197512.gpx

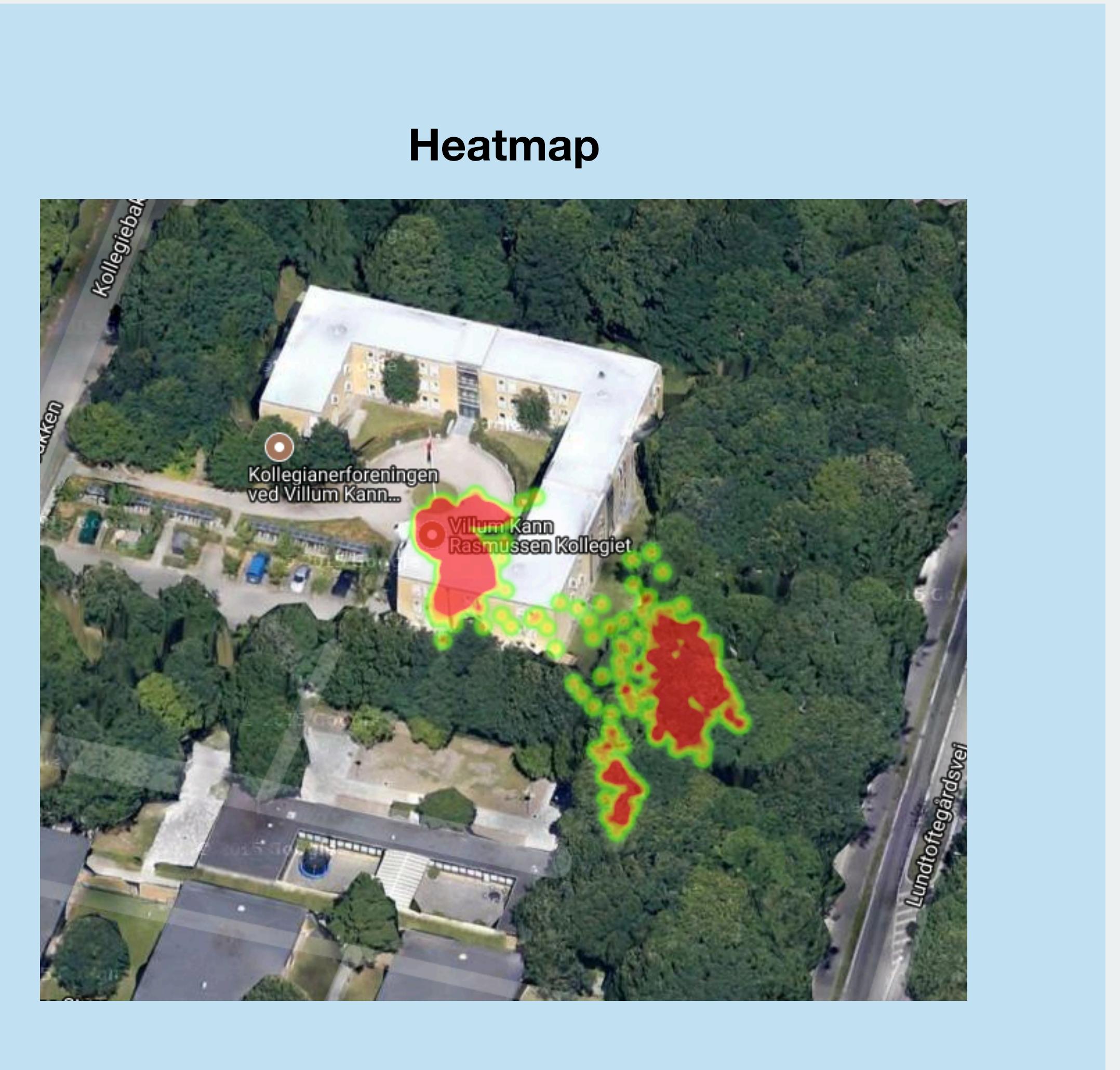
	lat	lon
0	55.784332	12.525468
1	55.784437	12.525030
2	55.784435	12.525043
3	55.784224	12.525565
4	55.784437	12.525031
5	55.784411	12.525055
6	55.784397	12.525070
7	55.784215	12.525537
8	55.784416	12.525059
9	55.784147	12.525530
10	55.784417	12.525063
11	55.784222	12.525535
12	55.784415	12.525052
13	55.784152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



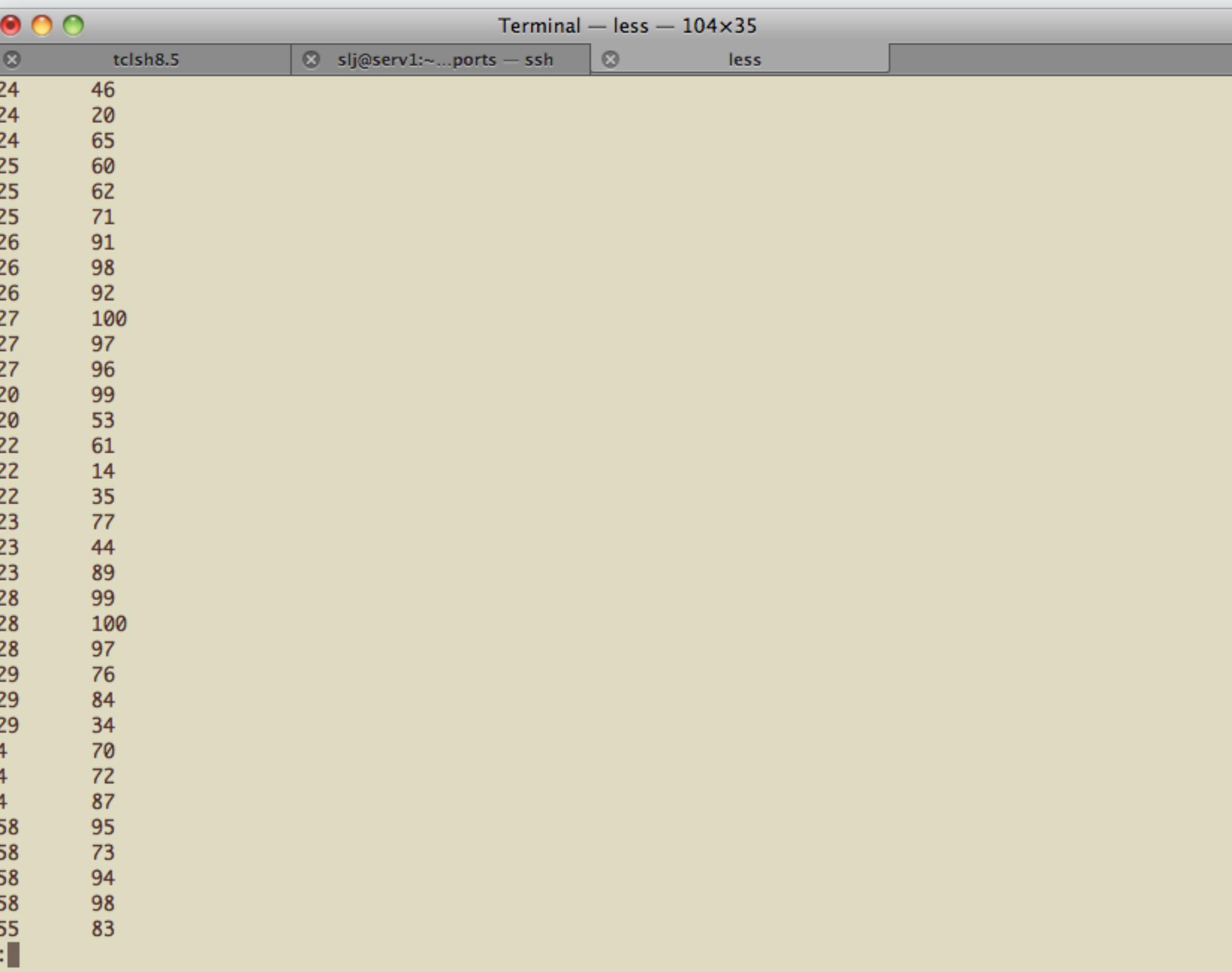
# Data must be rendered in human-friendly format

```
Terminal — less — 107x37
<?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:gpxpx="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>
          <gpxpx:TrackPointExtension>
            <gpxpx:hr>143</gpxpx:hr>
          </gpxpx:TrackPointExtension>
        </extensions>
      </trkpt>
      <trkpt lon="12.577596567571163" lat="55.70799755863845">
        <ele>12.0</ele>
        <time>2011-01-26T09:23:55.000Z</time>
        <extensions>
          <gpxpx:TrackPointExtension>
            <gpxpx:hr>143</gpxpx:hr>
          </gpxpx:TrackPointExtension>
        </extensions>
      </trkpt>
    </trkseg>
  </trk>
</gpx>
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

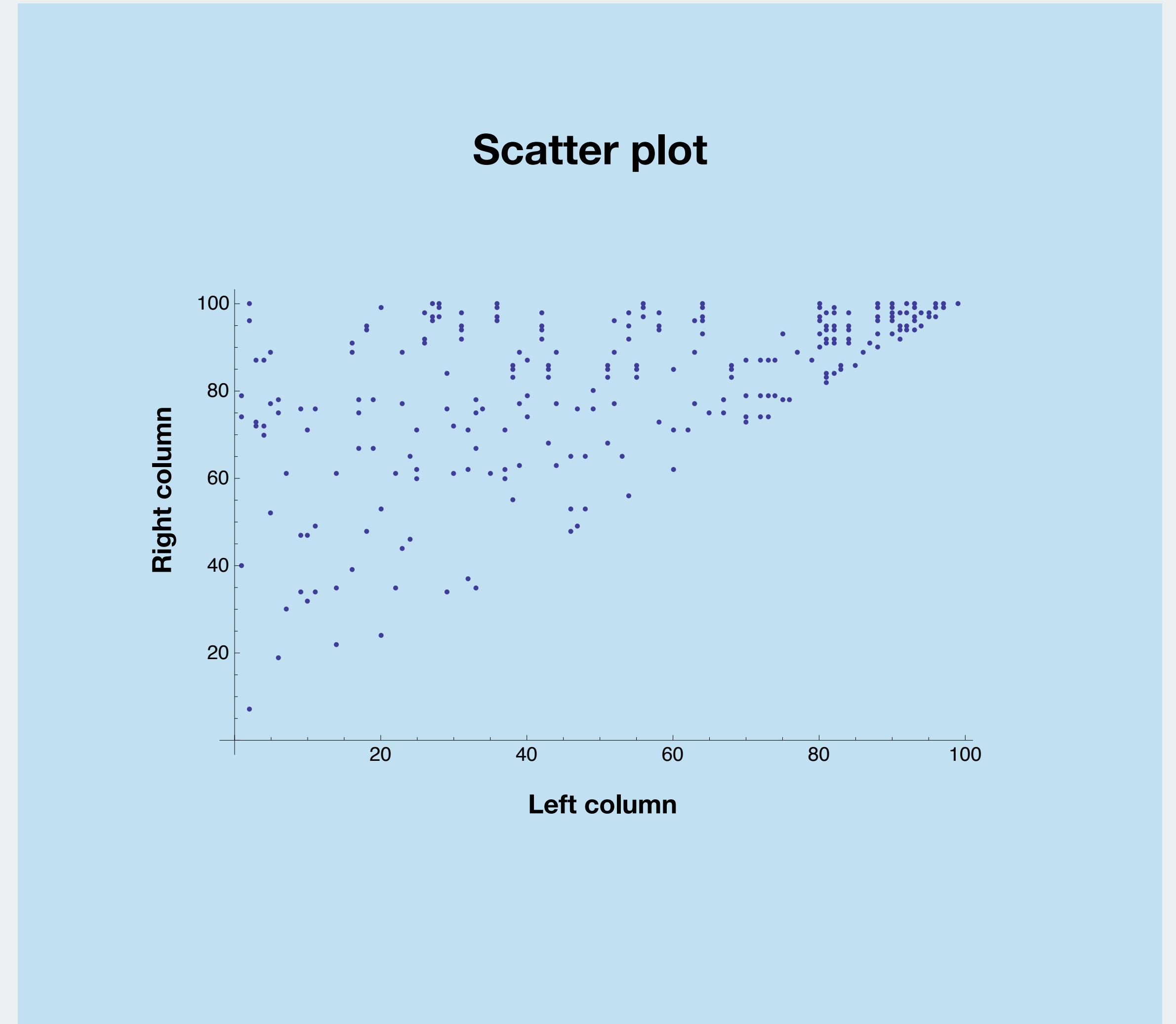


A screenshot of a Mac OS X terminal window titled "Terminal — less — 104x35". The window has three tabs: "tclsh8.5", "slj@serv1:~...ports — ssh", and "less". The "less" tab is active and displays a list of approximately 50 numerical values, likely integers, arranged vertically. The values range from 4 to 100, with some duplicates. A vertical scroll bar is visible on the right side of the terminal window.

```
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
:
```

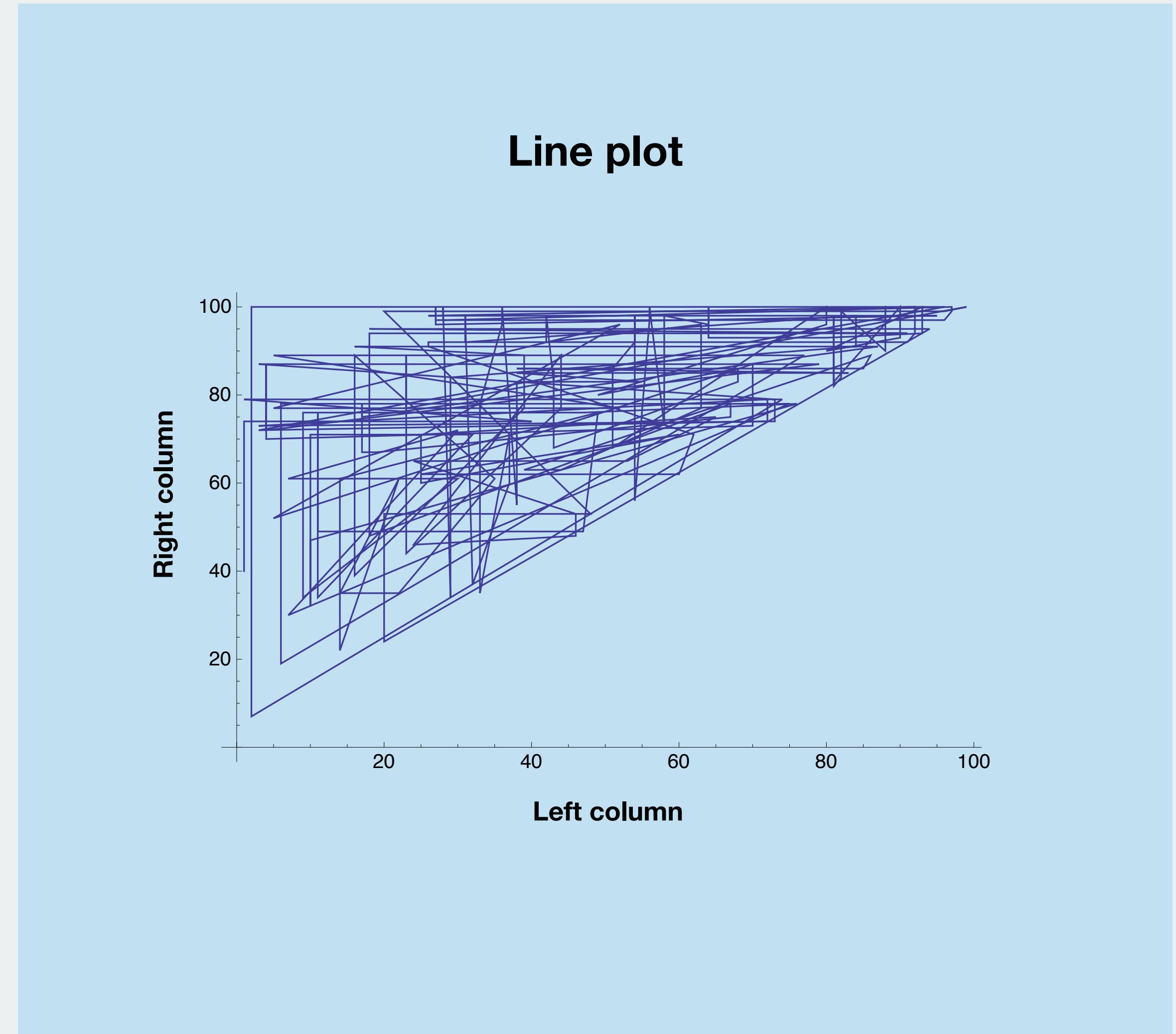
# Relational data

```
Terminal — less — 104x35
tclsh8.5      slj@serv1:~...ports — ssh      less
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

```
Terminal — less — 104x35
tclsh8.5      slj@serv1:~...ports — ssh      less
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
:|
```

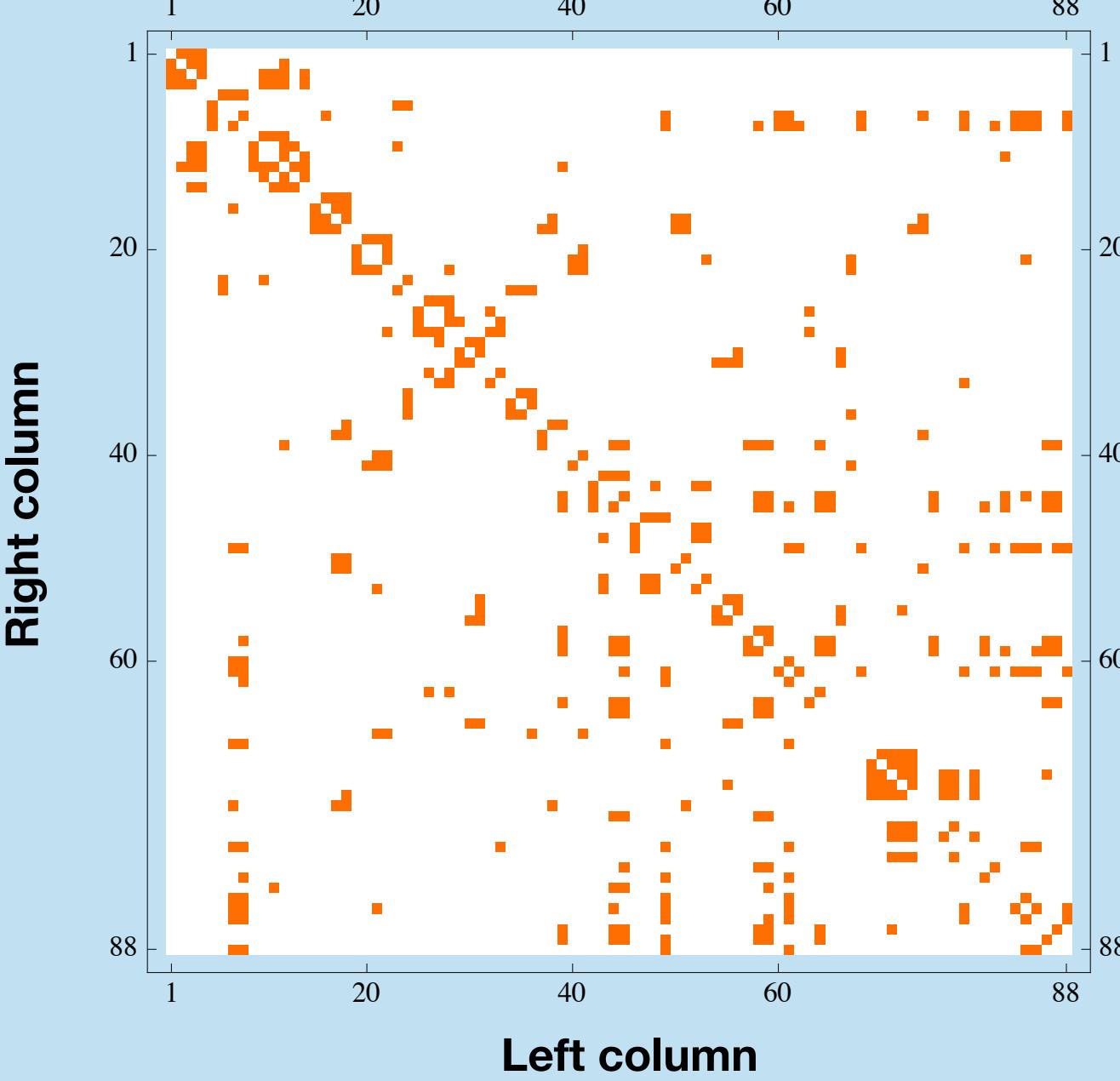


# Relational data

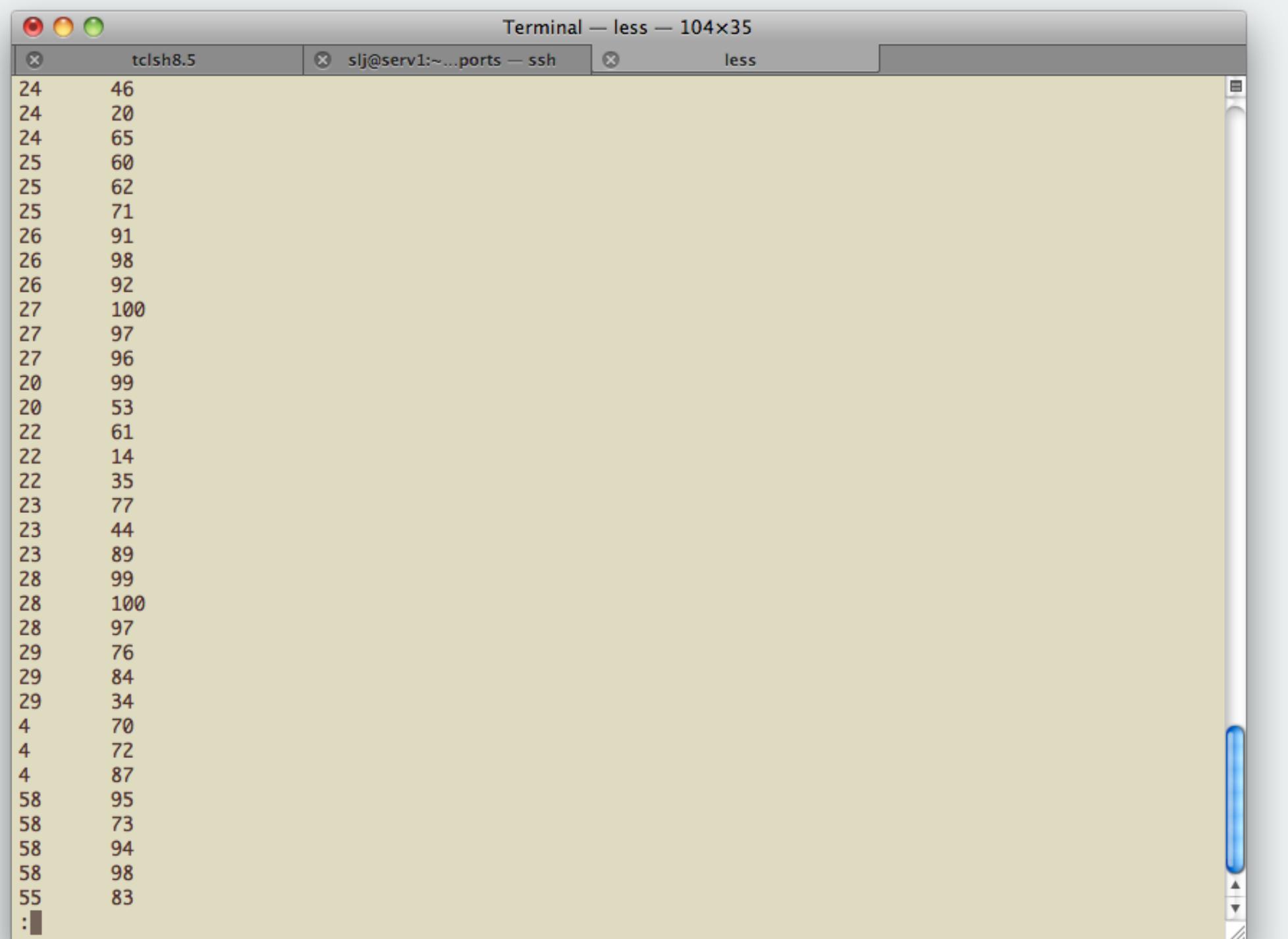
```
Terminal — less — 104x35
tclsh8.5      slj@serv1:~...ports — ssh      less
24 46
24 20
24 65
25 60
25 62
25 71
26 91
26 98
26 92
27 100
27 97
27 96
20 99
20 53
22 61
22 14
22 35
23 77
23 44
23 89
28 99
28 100
28 97
29 76
29 84
29 34
4 70
4 72
4 87
58 95
58 73
58 94
58 98
55 83
:
:
```



**Row-ordered adjacency matrix**



# Relational data

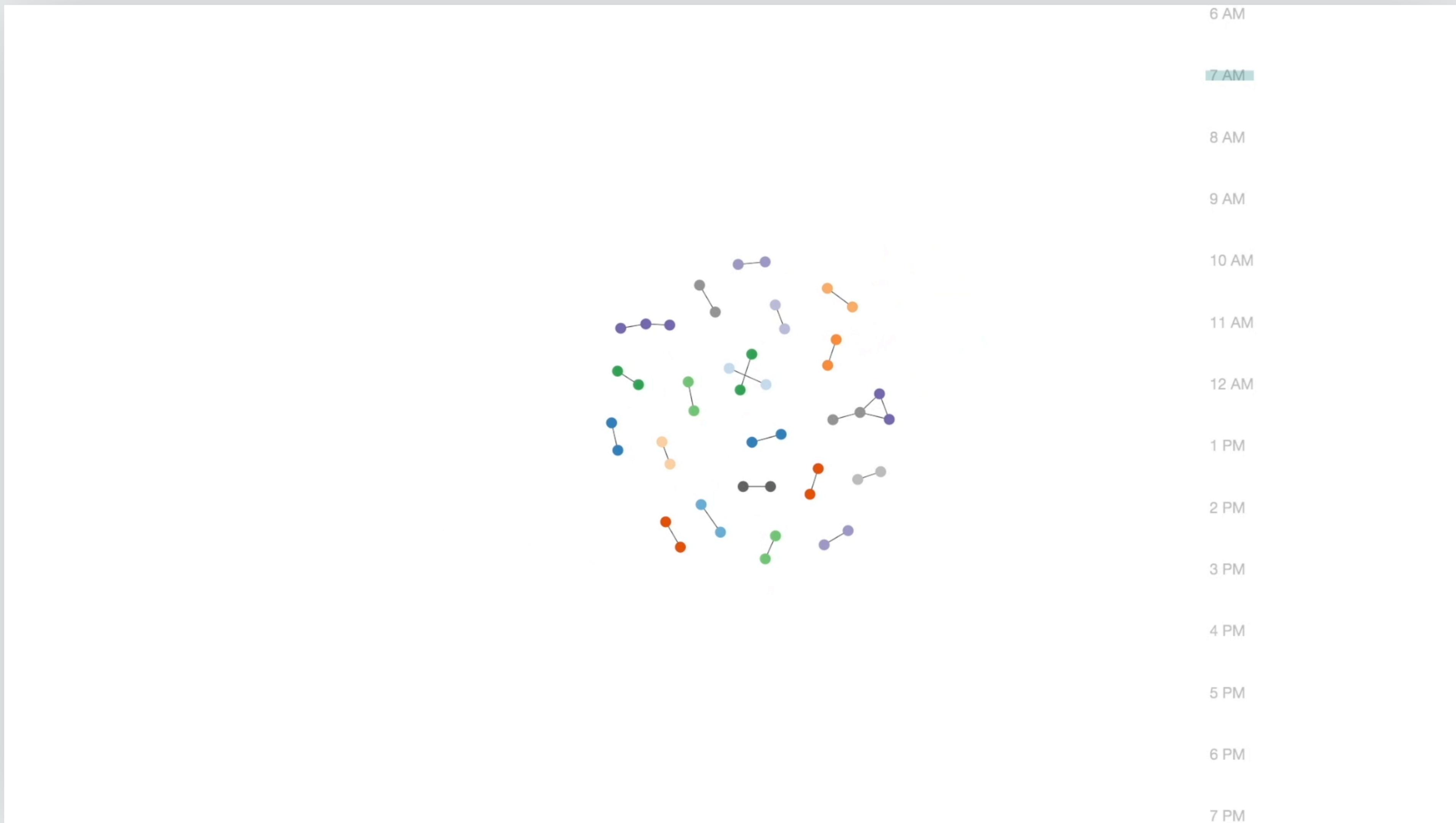


A screenshot of a Mac OS X terminal window titled "Terminal — less — 104x35". The window contains a list of numerical values, likely port numbers, displayed in two columns. An arrow points from this terminal window to the word "Network" on the right side of the slide.

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
:	■

## Network

# Very complex data that changes in time!



# 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 **relate 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 **relate and**  
**develop**

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

For understanding  
**complex structure** in  
interconnected data

- Networks

# 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!**

**0D**

*scalar*

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

**1D**

*vector*

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

**2D**

*matrix*

```
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!**

**0D**

*scalar*

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

**1D**

*vector*

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

**2D**

*matrix*

```
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**

*3D-tensor*

```
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]]]
```

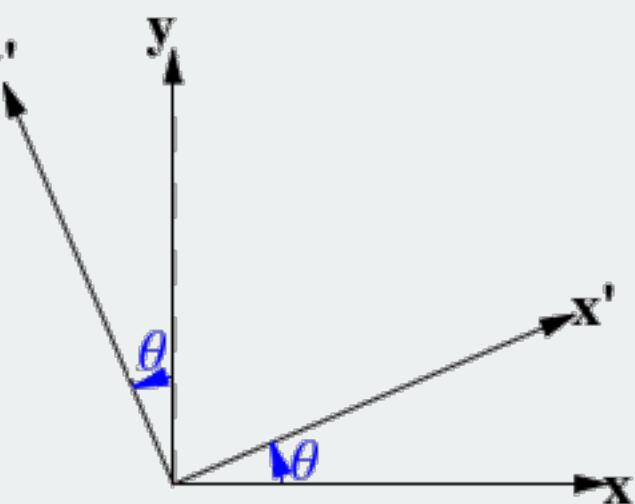
# Linear algebra

# 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

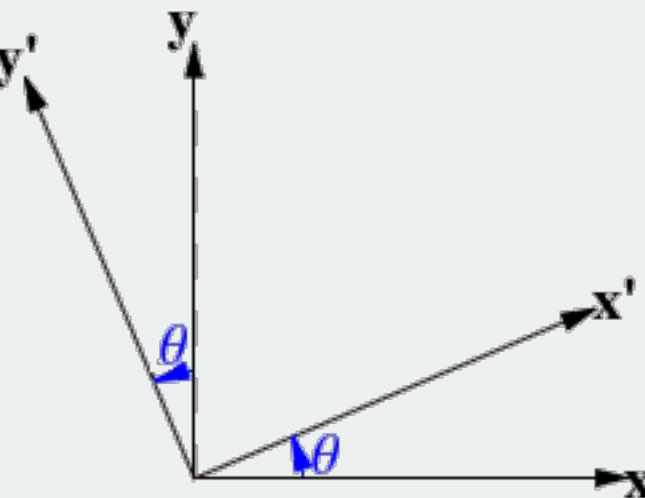
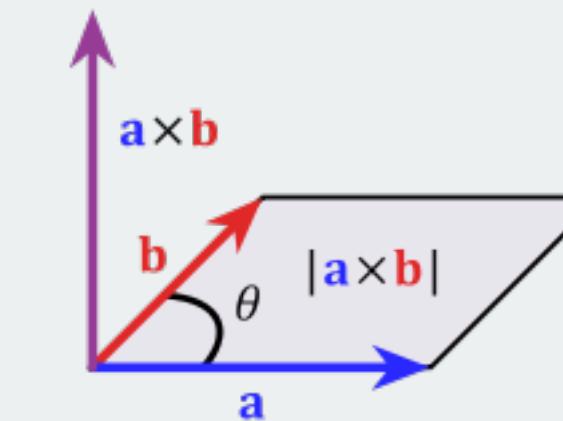
# Linear algebra

# 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

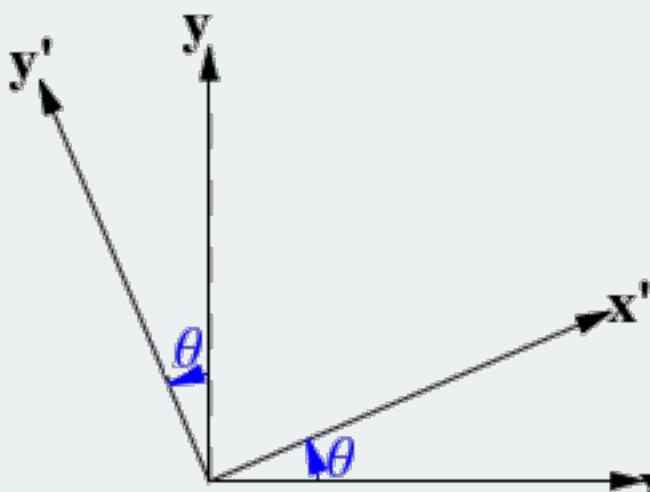
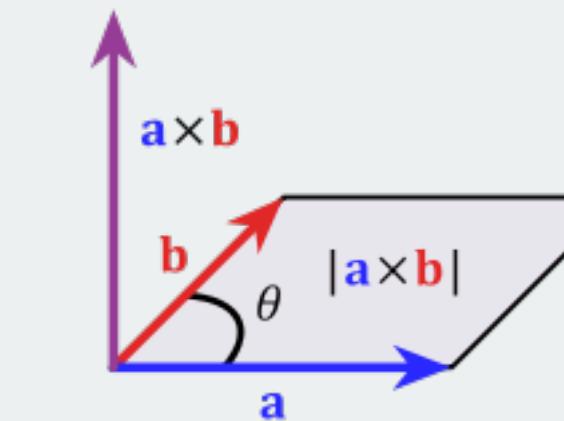
# Linear algebra

# Tools for manipulating tabular data

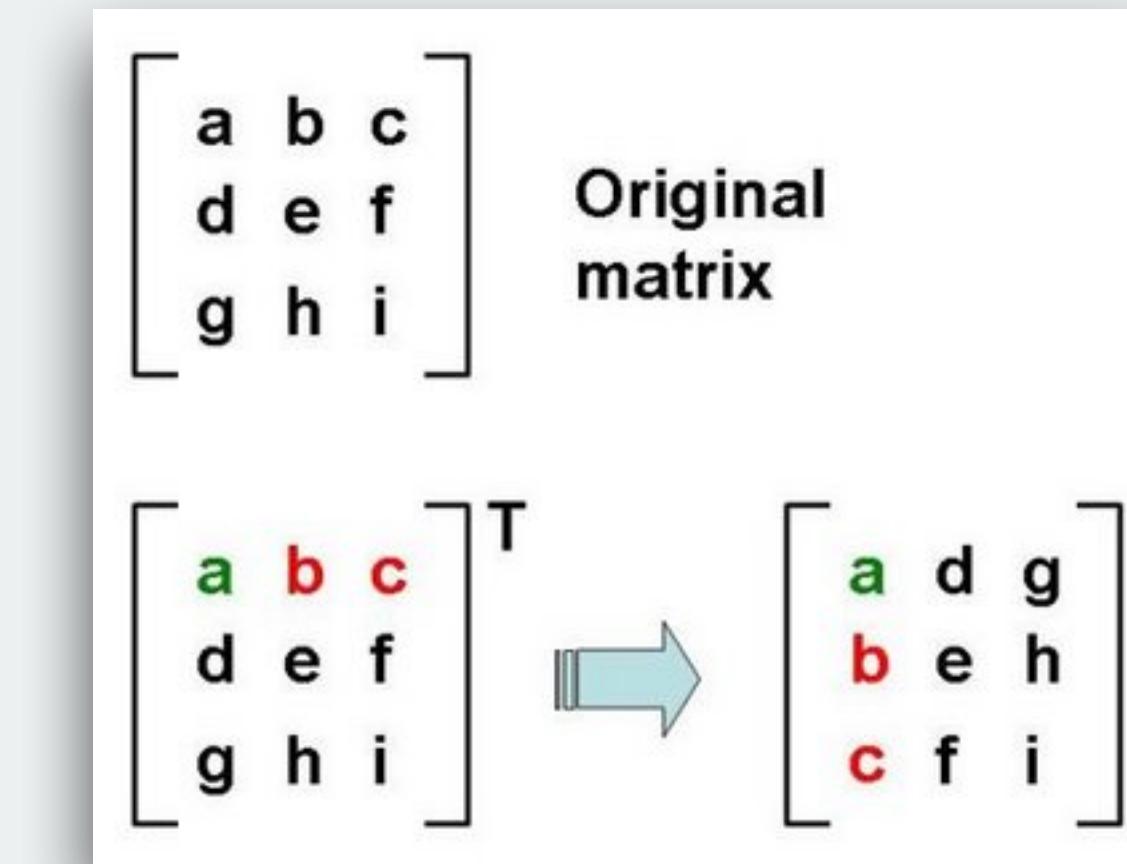
## Operations

- 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**

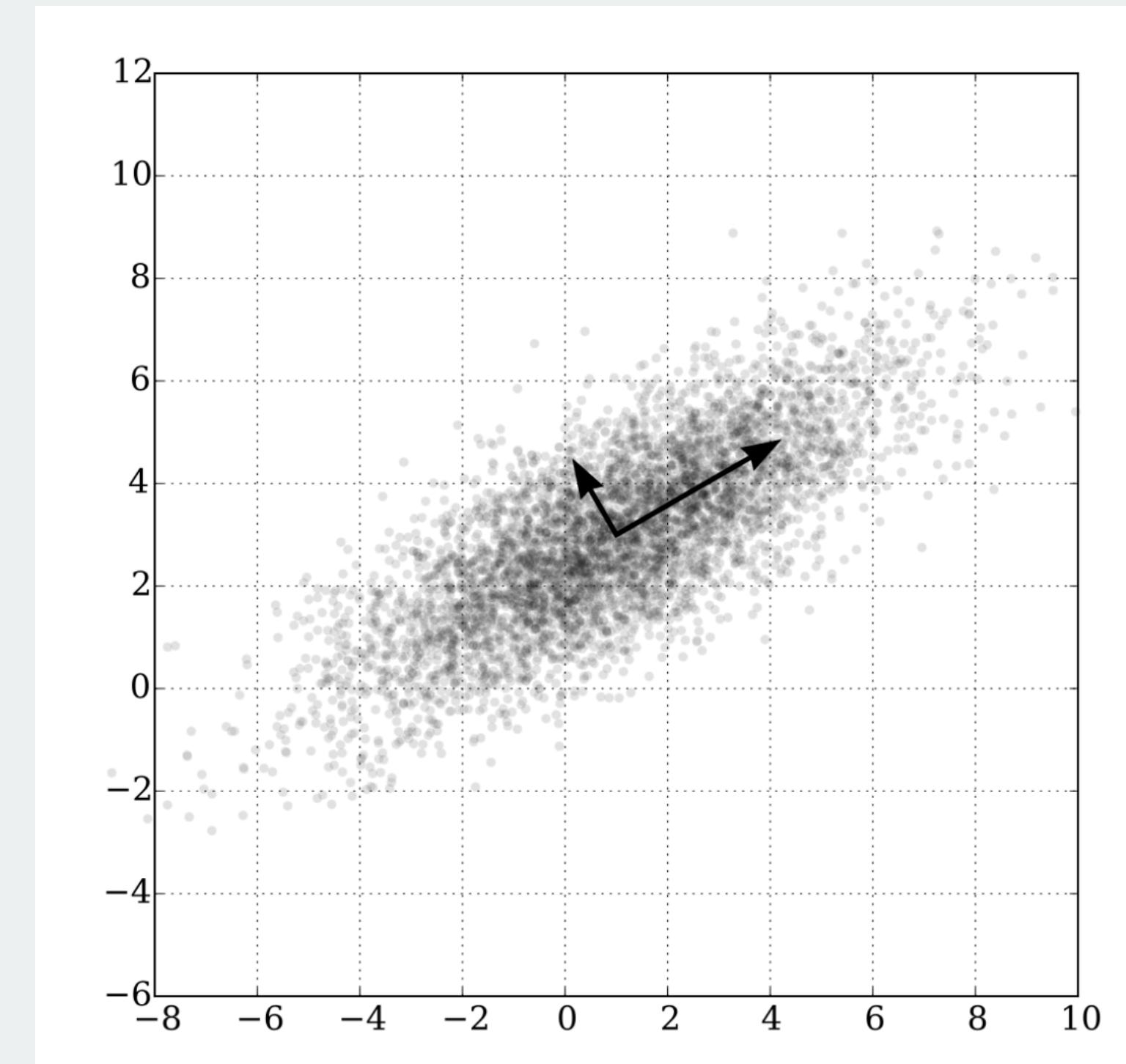


# Linear algebra

# Tools for manipulating tabular data

## Tools

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

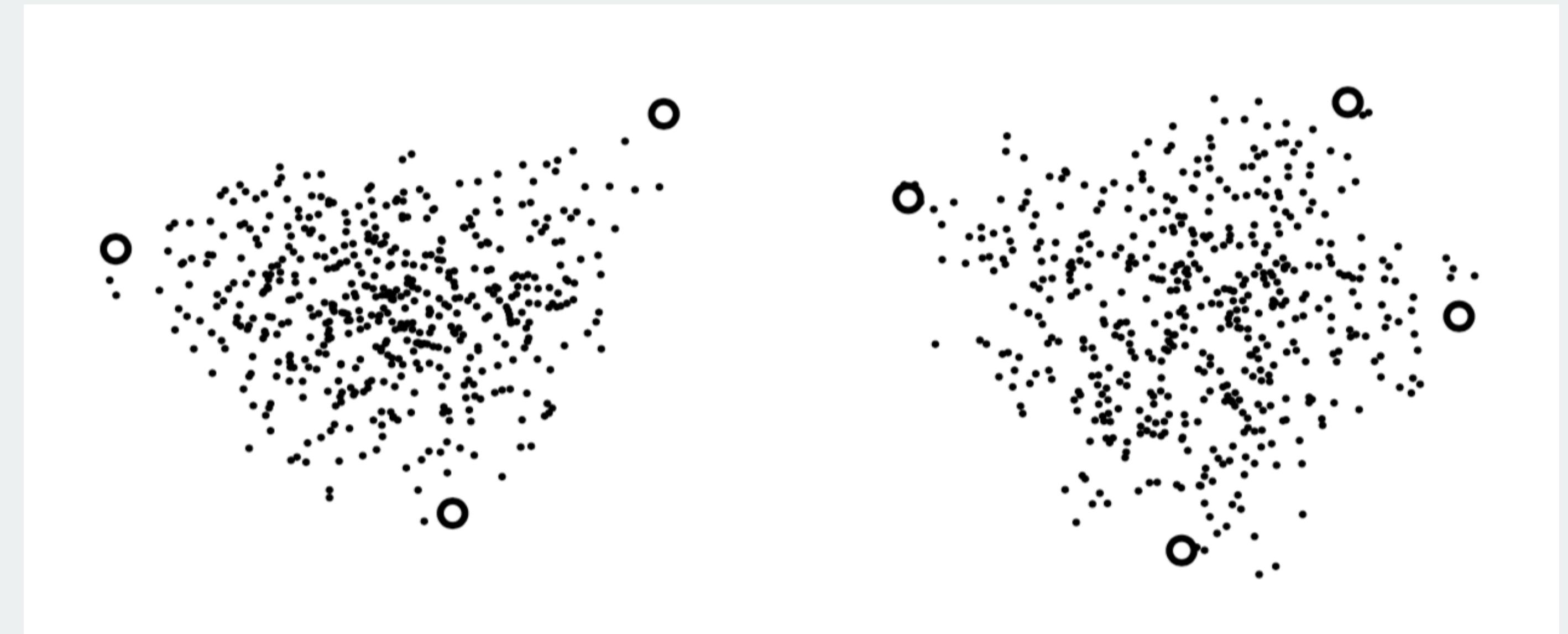


# Linear algebra

# Tools for manipulating tabular data

## Tools

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

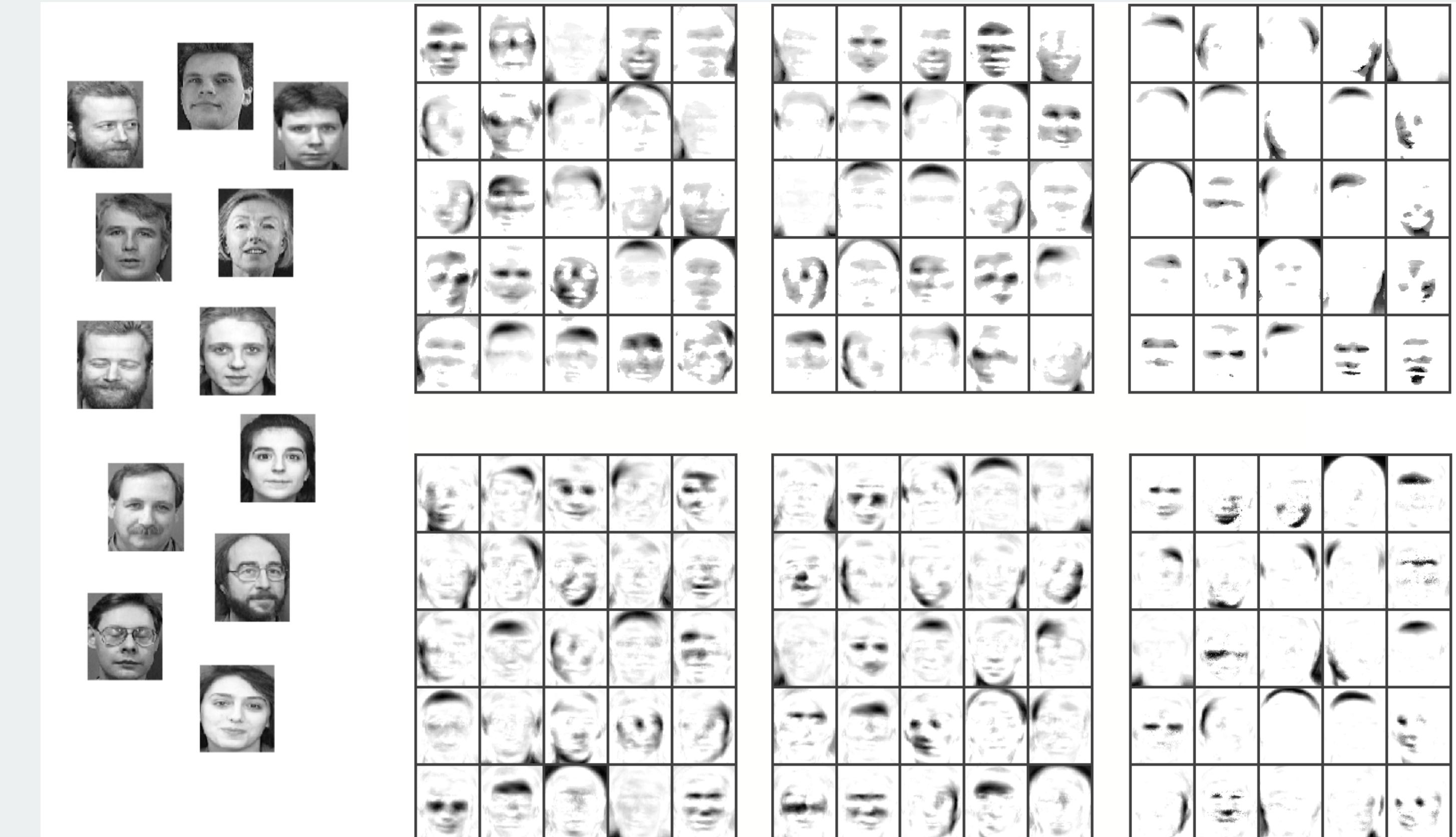


# Linear algebra

# Tools for manipulating tabular data

## Tools

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



# Statistics

# Statistics

## Framework for describing data

# Statistics

## Framework for describing data

### Vocabulary

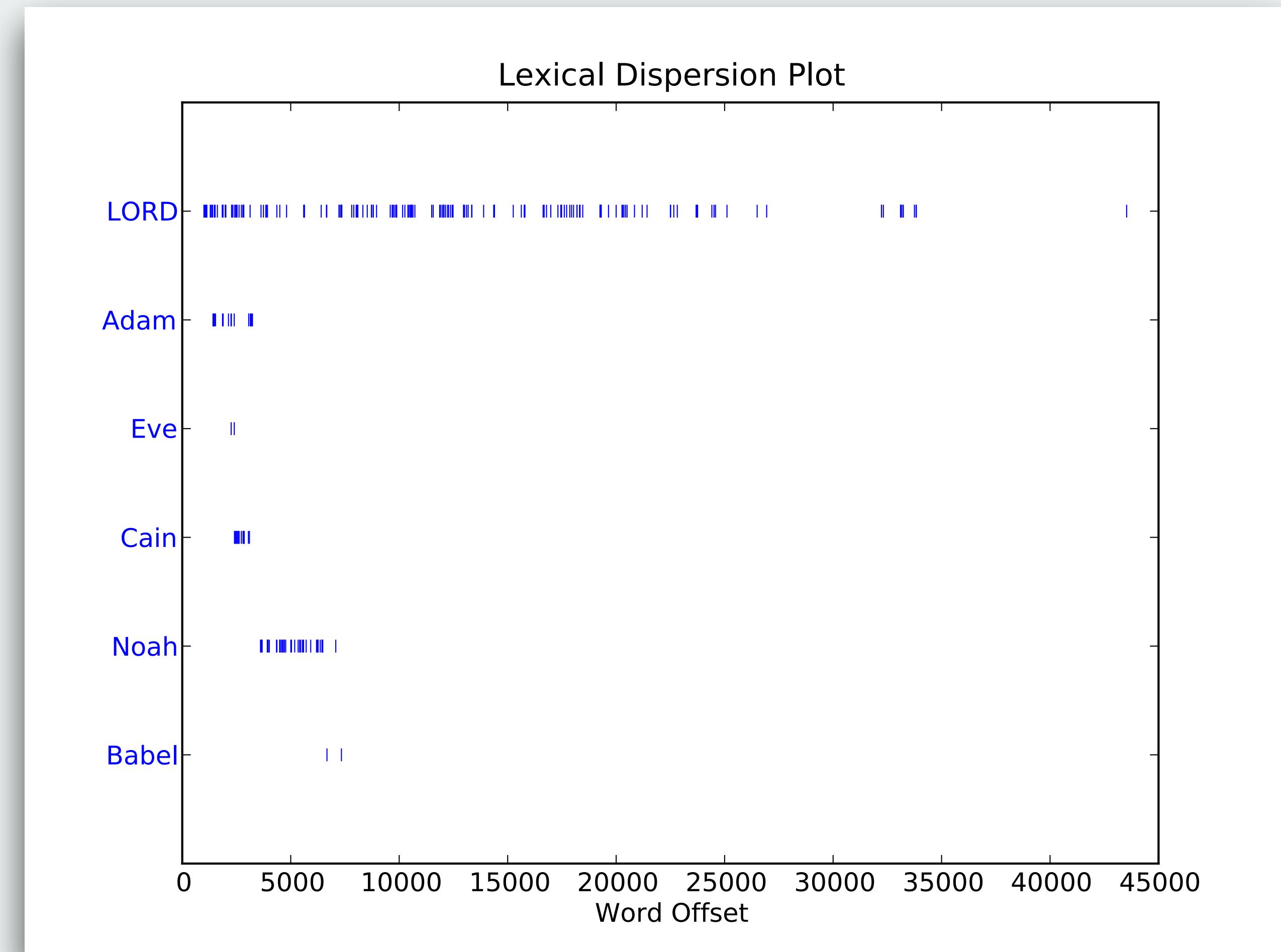
- Mean, median
- Variance, standard deviation, range
- Correlation, covariance

# Statistics

## Vocabulary

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

# Framework for describing data



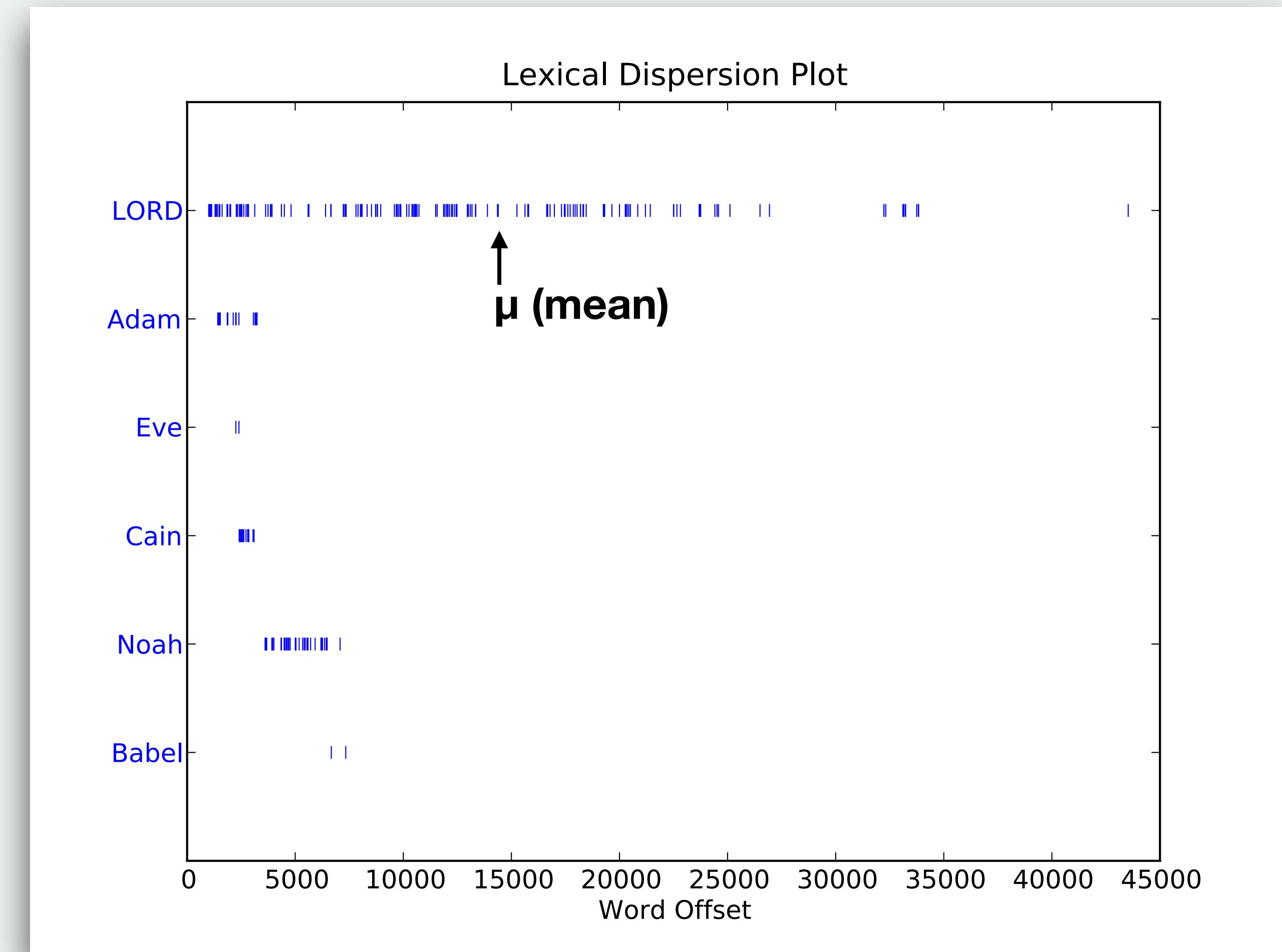
# Statistics

## Vocabulary

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

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

## Framework for describing data



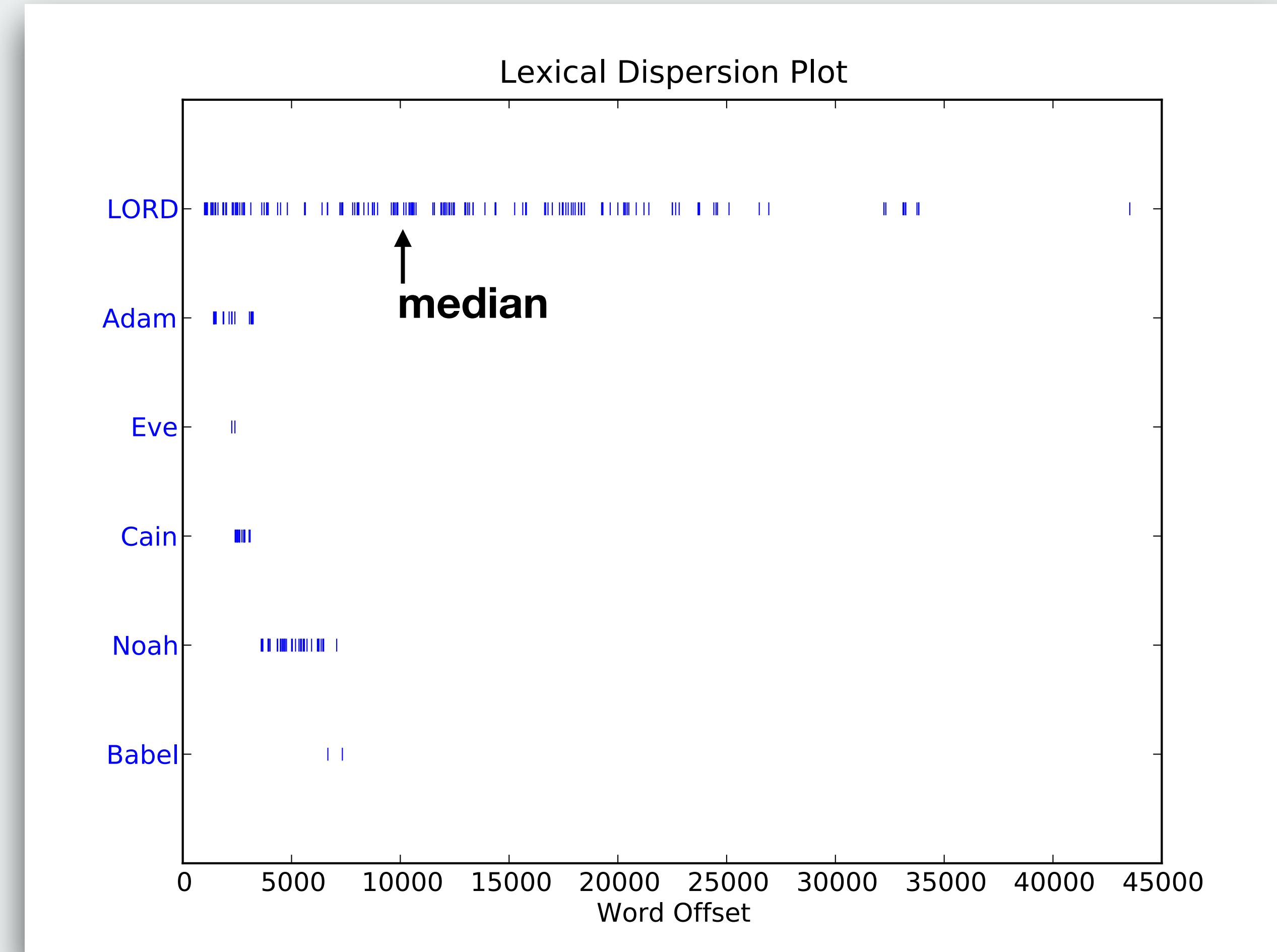
# Statistics

## Framework for describing data

### Vocabulary

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

**median** = Middle number  
in ordered list



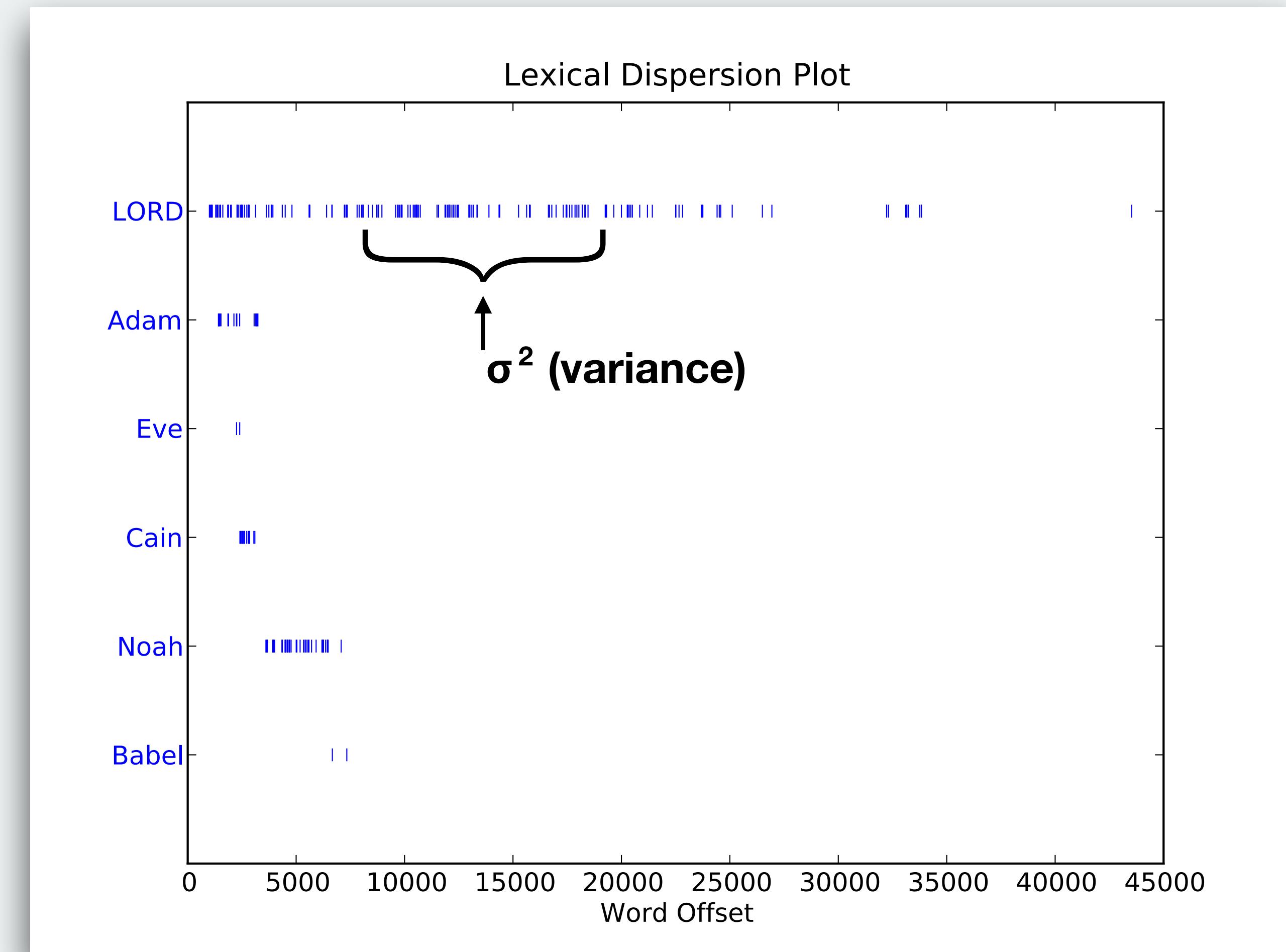
# Statistics

## Vocabulary

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

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

## Framework for describing data



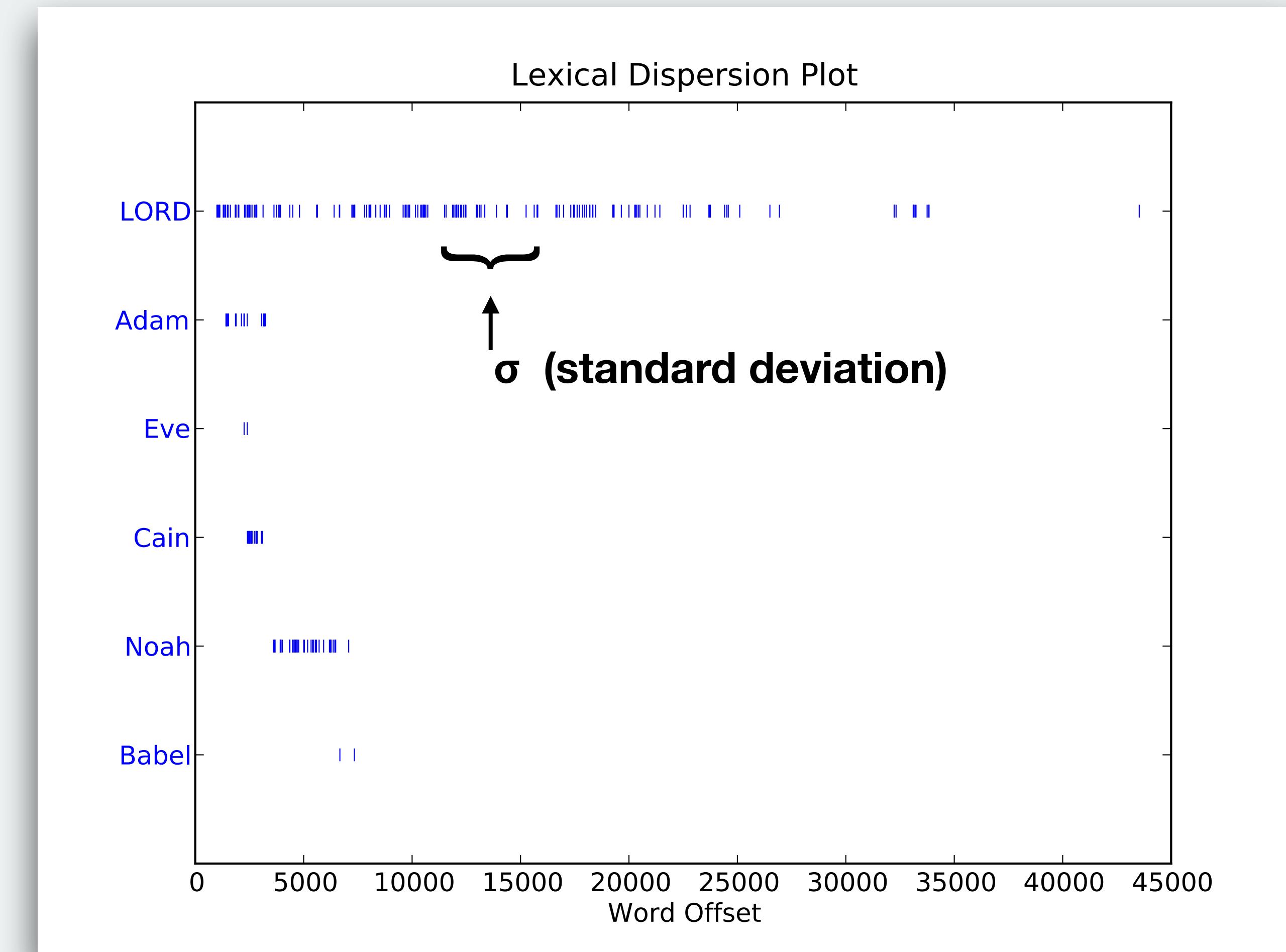
# Statistics

## Vocabulary

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

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

## Framework for describing data



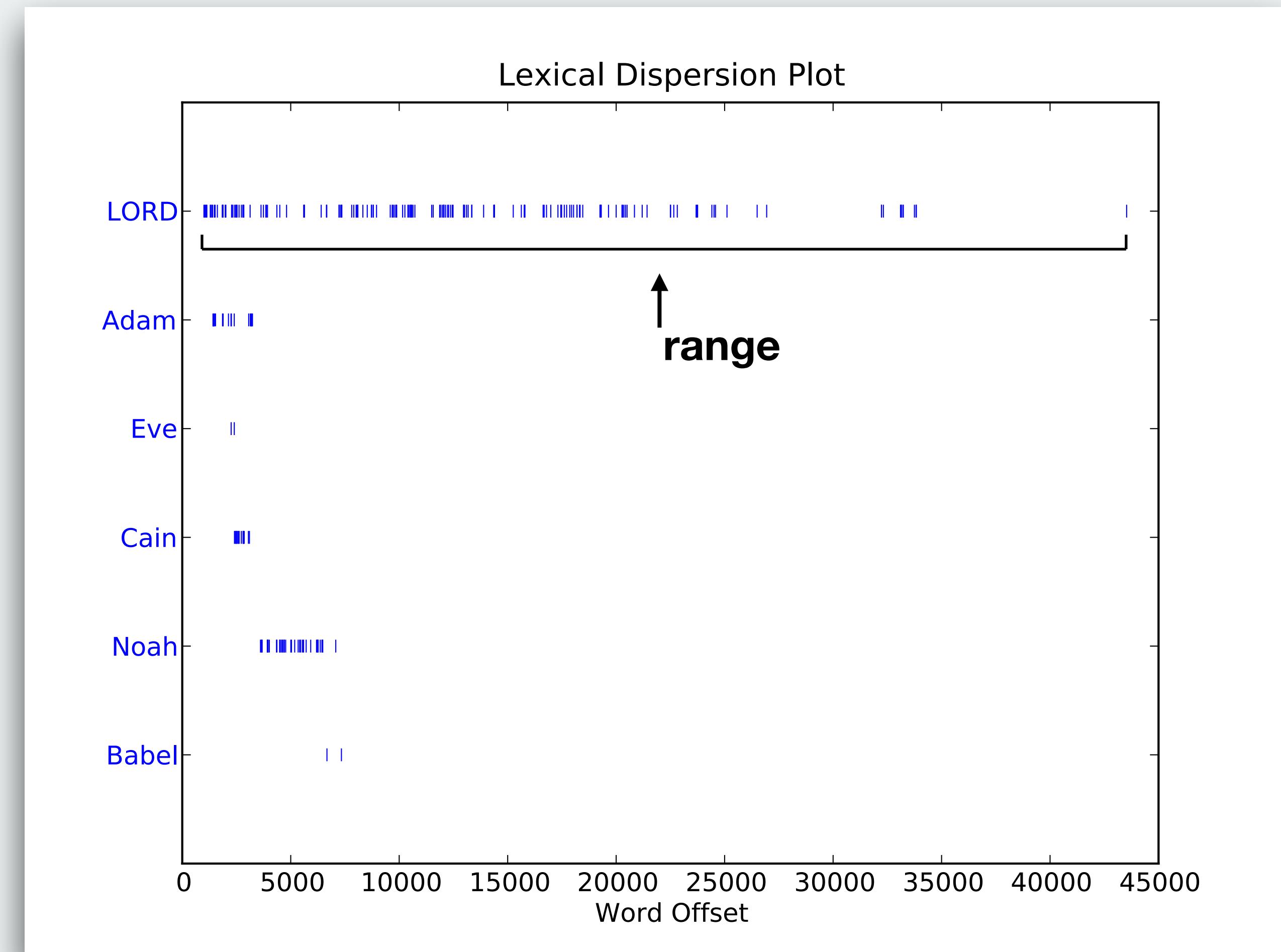
# Statistics

## Vocabulary

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

$$\text{range} = \max(\text{value}) - \min(\text{value})$$

## Framework for describing data



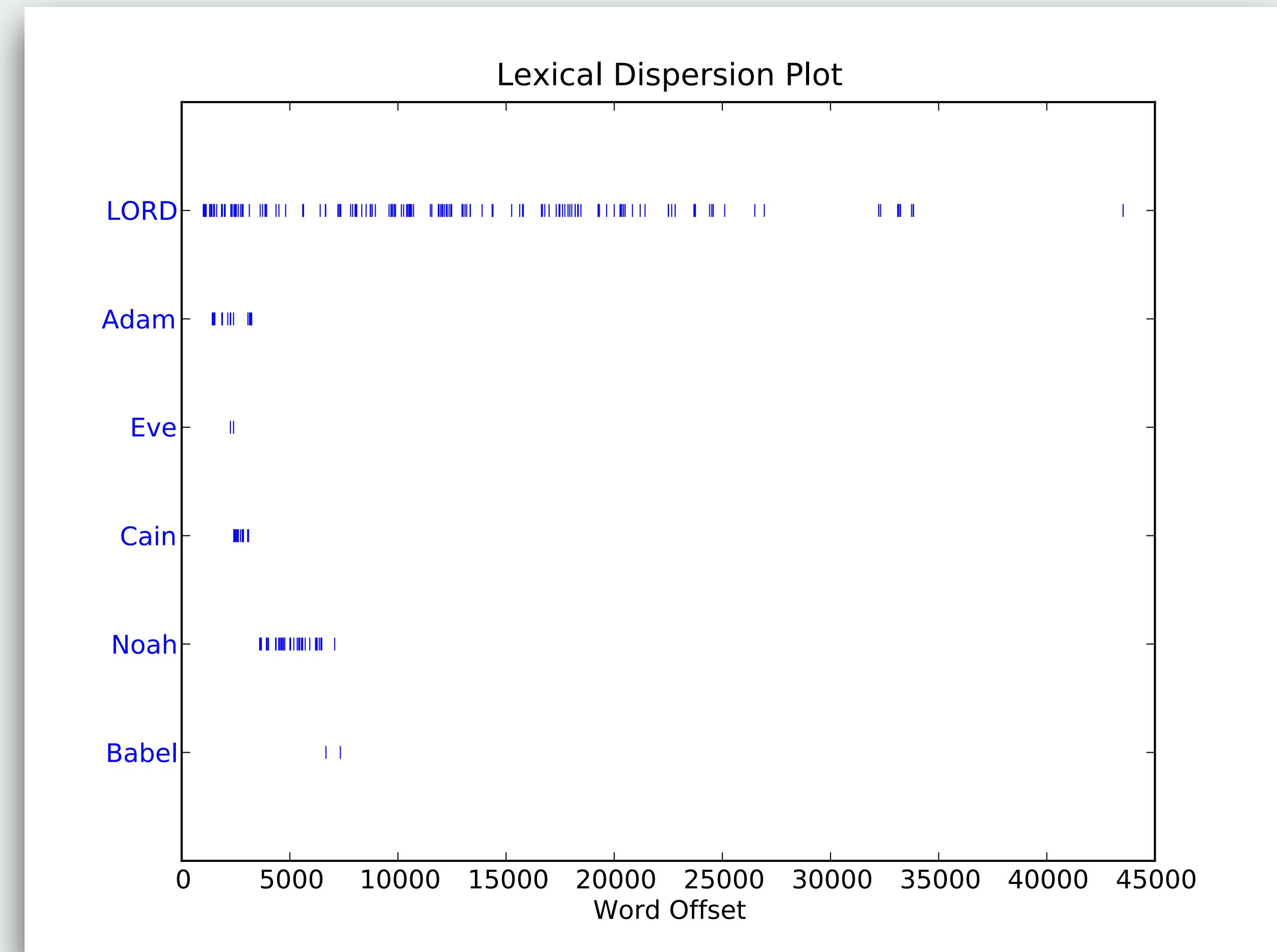
# Statistics

## Vocabulary

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

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

## Framework for describing data



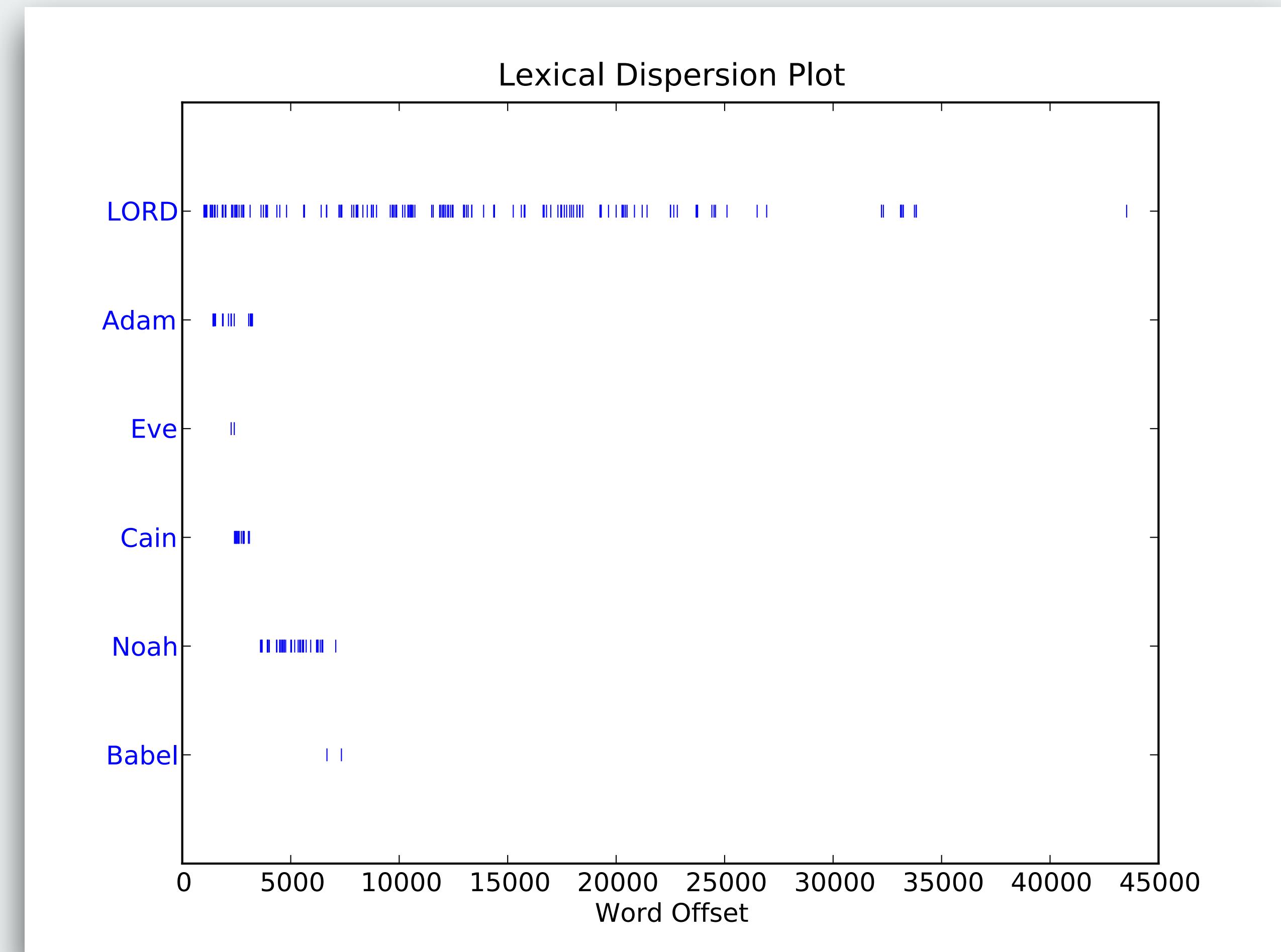
# Statistics

## Vocabulary

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

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

## Framework for describing data

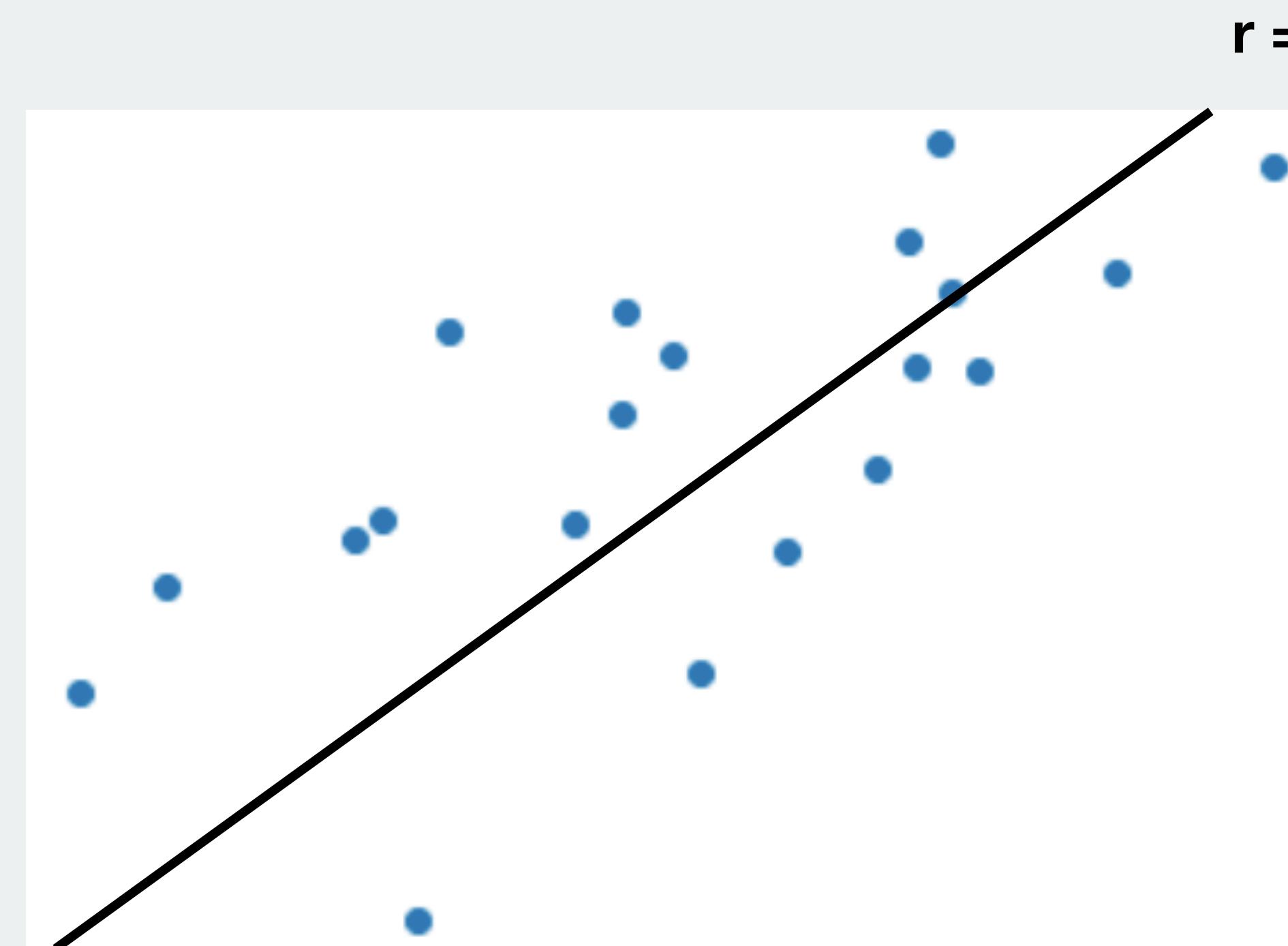


# Statistics

## Framework for describing data

### Tools

- Hypothesis testing

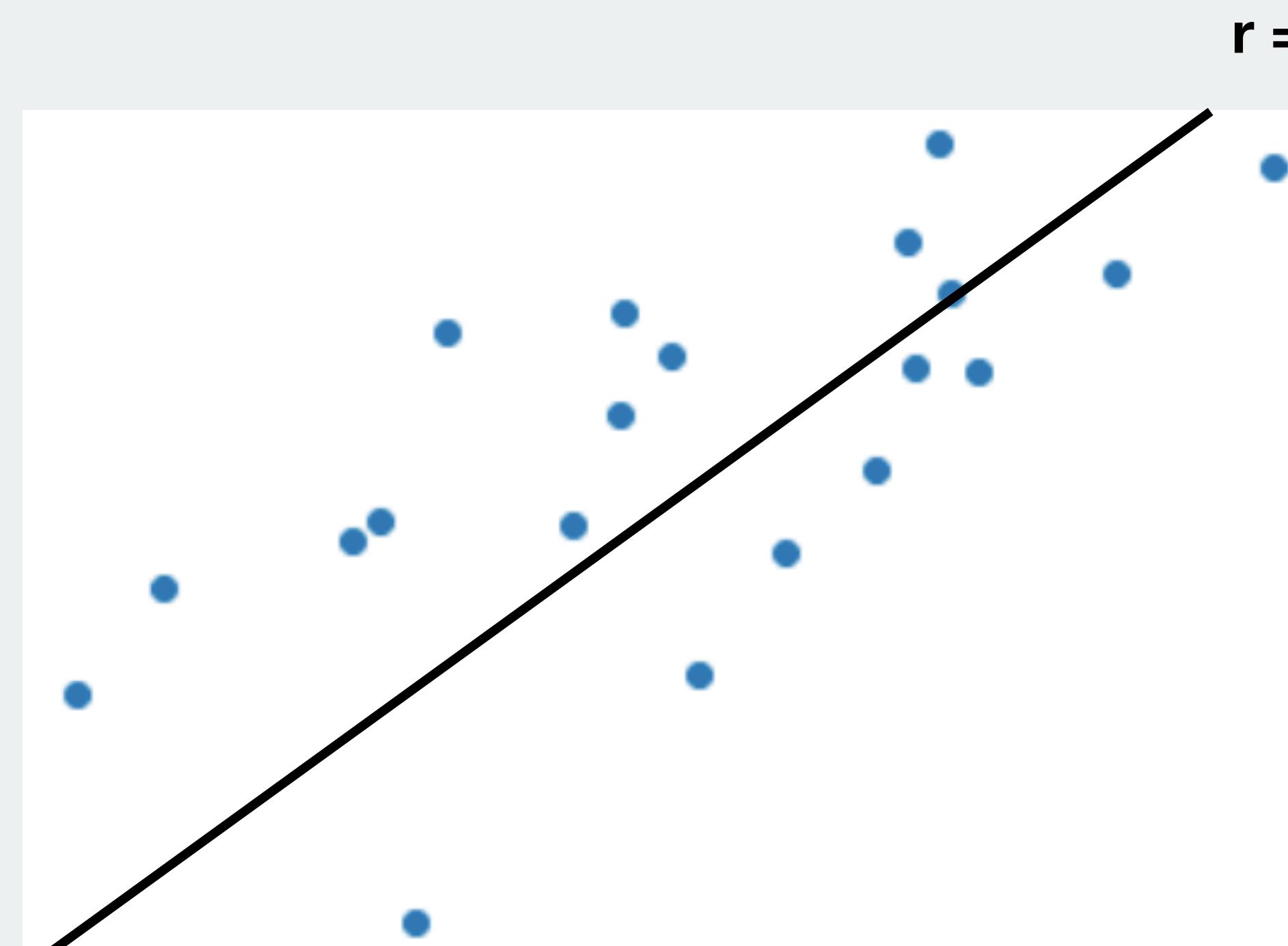
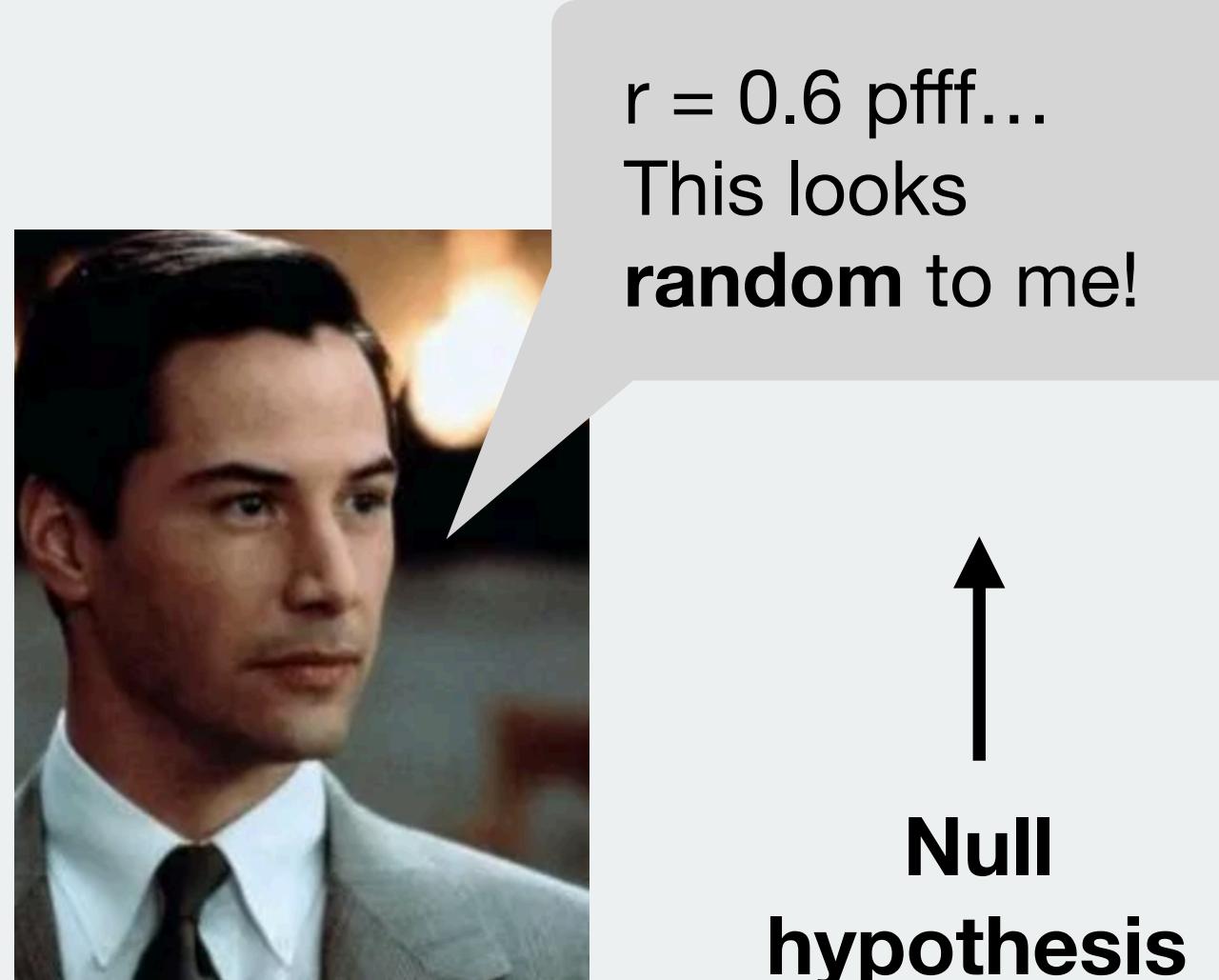


# Statistics

# Framework for describing data

## Tools

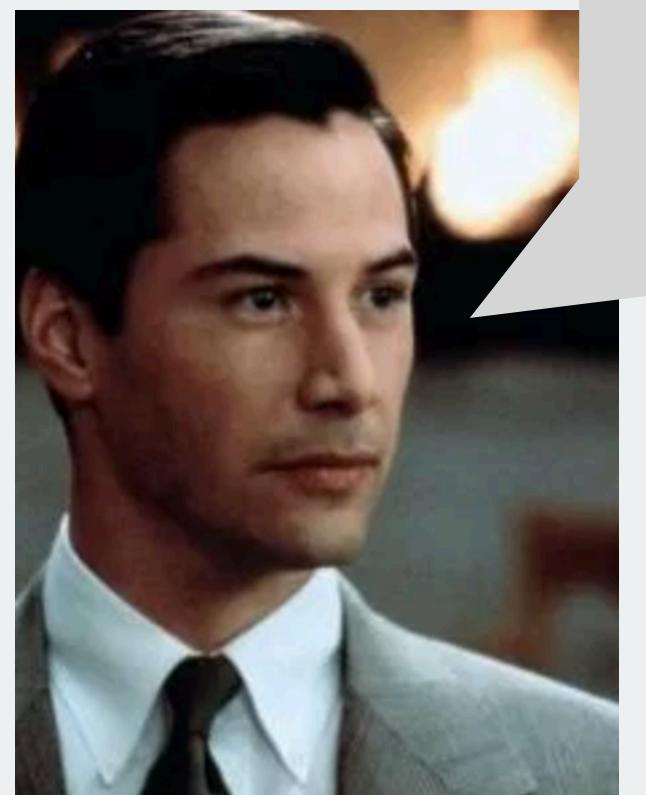
- Hypothesis testing



# Statistics

## Tools

- Hypothesis testing

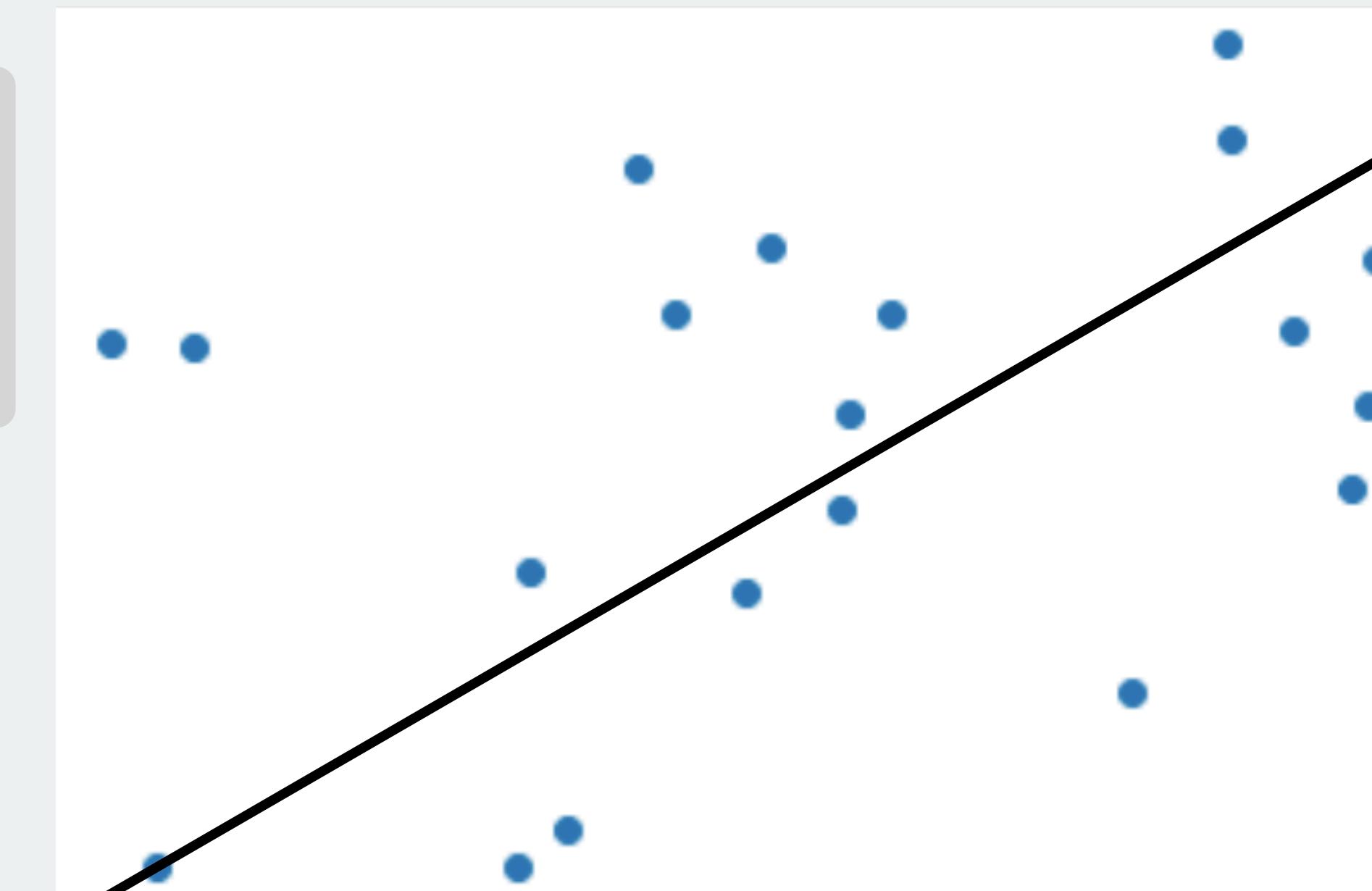


$r = 0.2$ ? But this  
is just one  
example. Still  
could be random

Null  
model

Randomization 1

$r = 0.2$

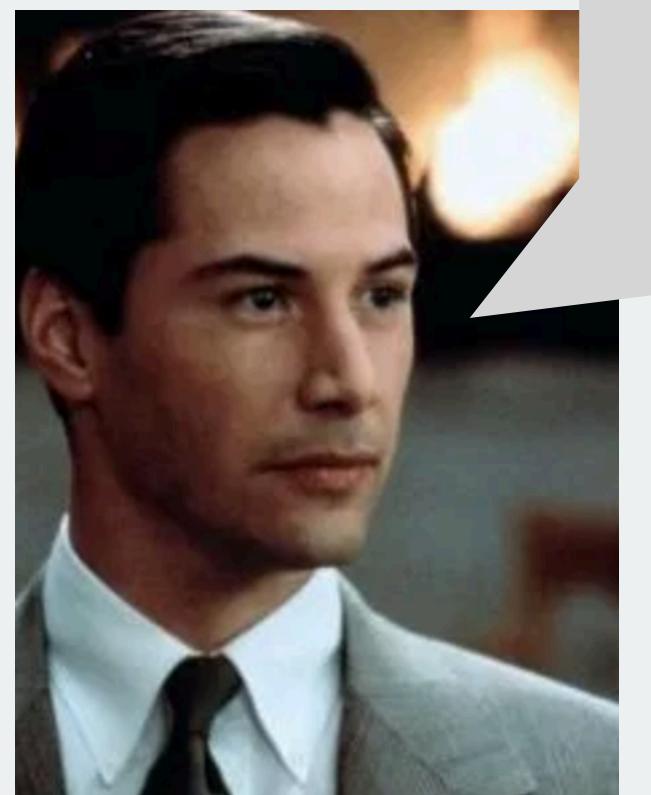


Null  
hypothesis

# Statistics

## Tools

- Hypothesis testing



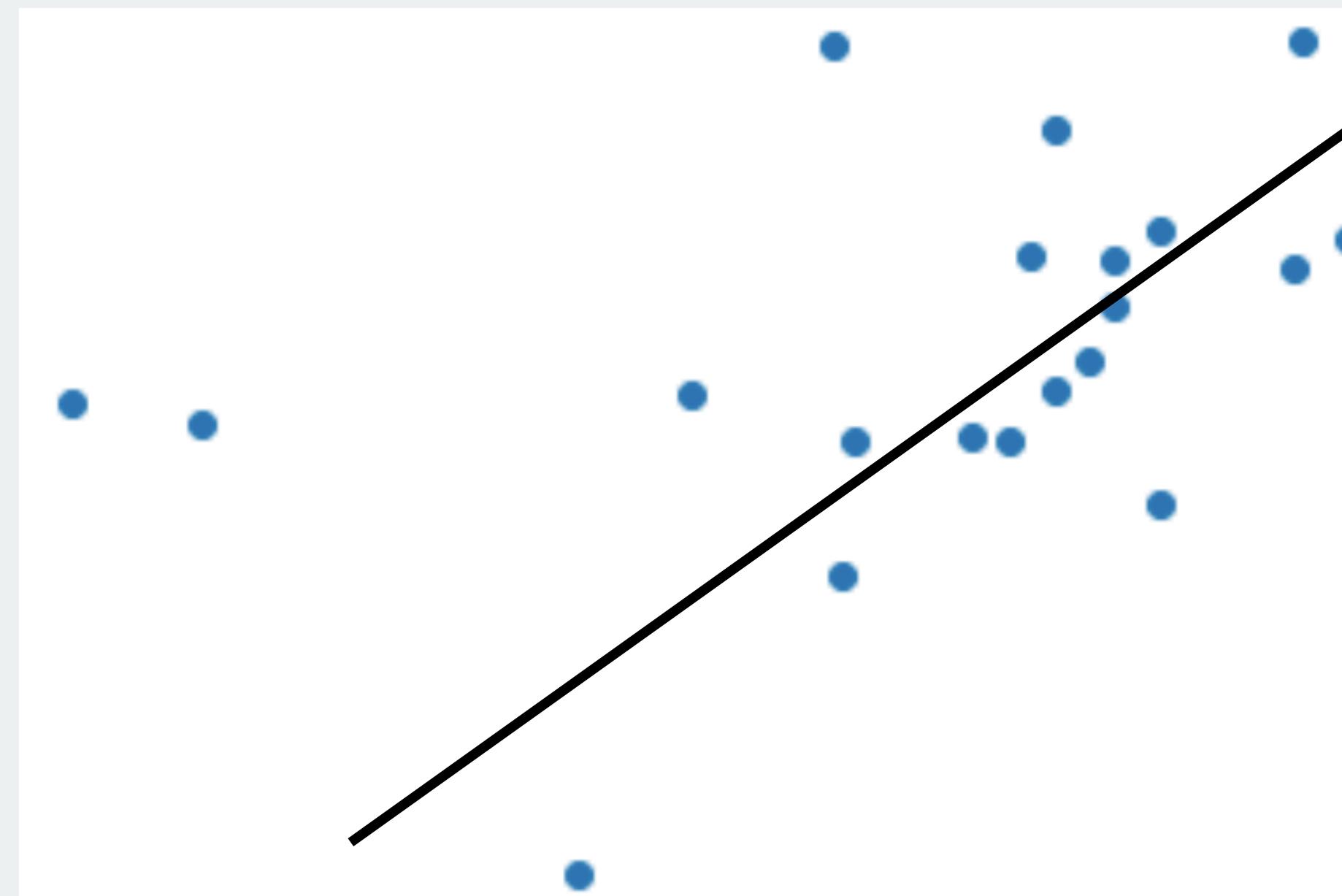
$r = 0.25$ ? Pure chance, I still think your  $r=0.6$  is random

Null hypothesis

Null model

Randomization 2

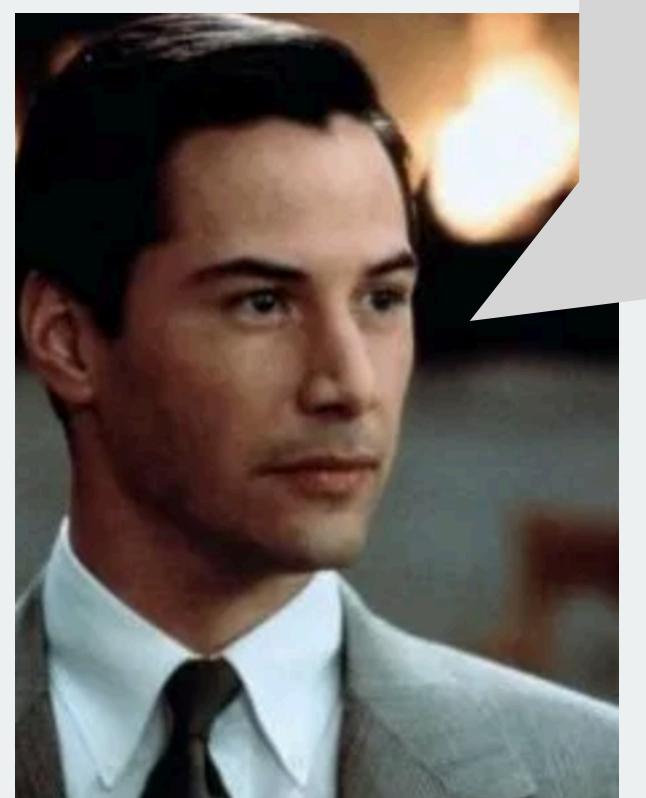
$r = 0.25$



# Statistics

## Tools

- Hypothesis testing



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

REJECTED  
Null hypothesis

Null  
model

Randomization 1000

$r = 0.3$



# Framework for describing data