

# Statistics for genomics

Mayo-Illinois Computational Genomics Course

June 11, 2019

Dave Zhao  
Department of Statistics  
University of Illinois at Urbana-Champaign



# Preparation

- `install.packages(c("Seurat", "glmnet", "ranger", "caret"))`
- Download sample GSM2818521 of GSE109158 from <https://urlzs.com/7UNr6>

Objective

# We will illustrate how to identify the appropriate statistical method for a genomics analysis

Cell Press  
Current Biology  
Article

## Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq

Daniel Parthey,<sup>1\*</sup> Karthik Shalhu,<sup>2</sup> Aviv Heger,<sup>2,7</sup> and Alexander F. Schaefer<sup>1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100</sup>

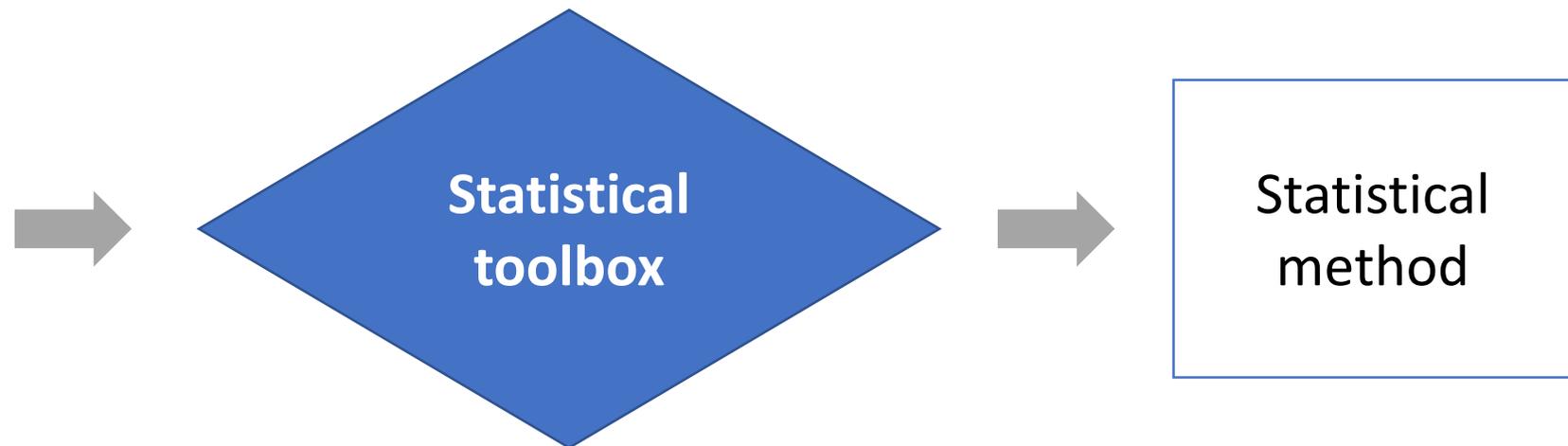
<sup>1</sup>Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA  
<sup>2</sup>Harvard Cell Observatory, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA  
<sup>3</sup>Howard Hughes Medical Institute and Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA  
<sup>4</sup>Center for Brain Science, Harvard University, 32 Oxford Street, Cambridge, MA 02138, USA  
<sup>5</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>6</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>7</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>8</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>9</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>10</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>11</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>12</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>13</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>14</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>15</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>16</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>17</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>18</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>19</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>20</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>21</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>22</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>23</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>24</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>25</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>26</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>27</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>28</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>29</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>30</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>31</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>32</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>33</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>34</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>35</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>36</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>37</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>38</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>39</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>40</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>41</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>42</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>43</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>44</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>45</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>46</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>47</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>48</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>49</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>50</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>51</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>52</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>53</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>54</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>55</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>56</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>57</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>58</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>59</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>60</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>61</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>62</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>63</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>64</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>65</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>66</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>67</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>68</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>69</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>70</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>71</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>72</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>73</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>74</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>75</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>76</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>77</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>78</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>79</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>80</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>81</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>82</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>83</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>84</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>85</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>86</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>87</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>88</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>89</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>90</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>91</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>92</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>93</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>94</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>95</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>96</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>97</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>98</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>99</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA  
<sup>100</sup>Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA

\*Lead Contact  
Correspondence: parthey@fas.harvard.edu (D.P.), schaefer@fas.harvard.edu (A.F.S.)  
DOI: 10.1016/j.cub.2018.03.014

**SUMMARY**  
The identification of cell types and marker genes is critical for dissecting neural development and function, but the size and complexity of the brain has hindered the comprehensive discovery of cell types. We combined single-cell RNA-seq (scRNA-seq) with anatomical brain registration to create a comprehensive map of the avian habenula, a conserved forebrain hub involved in pain processing and learning. Single-cell transcriptomes of ~15,000 habenular cells with 4× cellular coverage identified 18 neuronal types and dozens of marker genes. Registration of marker genes onto a reference atlas created a resource for anatomical and functional studies and enabled the mapping of active neurons onto neuronal types to tracing aversive stimuli. Strikingly, despite brain growth and functional maturation, cell types were retained between the larval and adult habenula. This study provides a gene expression atlas to dissect habenular development and function and offers a general framework for the comprehensive characterization of other brain regions.

**INTRODUCTION**  
The study of formation and function of neural circuits relies on the ability to identify specific cell types. But are defined by location, morphology, connectivity, and molecular composition. Classical histological and gene expression analyses have recently been extended through cell techniques that enable more detailed characterization of cell types based on their transcriptomes [1–3]. Such studies have provided valuable resources for cataloging cell types, but are limited in their comprehensive identification by the large number and diversity of neurons in vertebrate brains. This complexity results in the naming of new cell types over

1000 Current Biology 28, 1005–1015, April 2, 2018 © 2018 Elsevier Inc.



# Classifying statistical tools

## Data structure

## Statistical task

	No dependent variables	Continuous outcome	Censored outcomes	Etc.
Visualize				
Identify latent factors				
Cluster observations				
Select features				
Etc.				

**APPROPRIATE  
STATISTICAL  
METHODS**

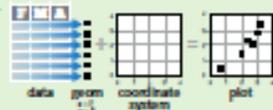
# Examples from basic statistics

## Data Visualization with ggplot2 Cheat Sheet

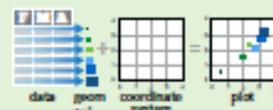


### Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Build a graph with `ggplot()` or `qplot()`

```
ggplot(data = mpg, aes(x = cty, y = hwy))
```

Begins a plot that you finish by adding layers. No defaults, but provides more control than `qplot()`.

```
ggplot(mpg, aes(hwy, cty)) +
  geom_point(aes(color = cyl)) +
  geom_smooth(method = "lm") +
  coord_cartesian() +
  scale_color_gradient() +
  theme_bw()
```

- add layers, elements with +
- layer = geom + default stat + layer specific mappings
- additional elements

Add a new layer to a plot with a `geom_*()` or `stat_*()` function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

```
qplot(x = cty, y = hwy, color = cyl, data = mpg, geom = "point")
```

Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

```
last_plot()
```

Returns the last plot

```
ggsave("plot.png", width = 5, height = 5)
```

Saves last plot as 5" x 5" file named "plot.png" in working directory. Matches file type to file extension.

**Geoms** - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

### One Variable

#### Continuous

```
a <- ggplot(mpg, aes(hwy))
```



```
a + geom_area(stat = "bin")
```

x, y, alpha, color, fill, linetype, size  
b + geom\_area(aes(y = ..density..), stat = "bin")



```
a + geom_density(kernel = "gaussian")
```

x, y, alpha, color, fill, linetype, size, weight  
b + geom\_density(aes(y = ..county..))



```
a + geom_dotplot()
```

x, y, alpha, color, fill



```
a + geom_freqpoly()
```

x, y, alpha, color, linetype, size  
b + geom\_freqpoly(aes(y = ..density..))



```
a + geom_histogram(binwidth = 5)
```

x, y, alpha, color, fill, linetype, size, weight  
b + geom\_histogram(aes(y = ..density..))

#### Discrete

```
b <- ggplot(mpg, aes(fill))
```



```
b + geom_bar()
```

x, alpha, color, fill, linetype, size, weight

### Graphical Primitives

```
map <- map_data("state")
```

```
c <- ggplot(map, aes(long, lat))
```



```
c + geom_polygon(aes(group = group))
```

x, y, alpha, color, fill, linetype, size

```
d <- ggplot(economics, aes(date, unemployment))
```



```
d + geom_path(lineend = "butt",
  linejoin = "round", linemitre = 1)
```

x, y, alpha, color, linetype, size



```
d + geom_ribbon(aes(ymin = unemployment - 900,
  ymax = unemployment + 900))
```

x, ymax, ymin, alpha, color, fill, linetype, size

```
e <- ggplot(seals, aes(x = long, y = lat))
```



```
e + geom_segment(aes(xend = long + delta_long,
  yend = lat + delta_lat))
```

x, xend, y, yend, alpha, color, linetype, size



```
e + geom_rect(aes(xmin = long, ymin = lat,
  xmax = long + delta_long,
  ymax = lat + delta_lat))
```

xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

### Two Variables

#### Continuous X, Continuous Y

```
f <- ggplot(mpg, aes(cty, hwy))
```



```
f + geom_blank()
```

(Useful for expanding limits)



```
f + geom_jitter()
```

x, y, alpha, color, fill, shape, size



```
f + geom_point()
```

x, y, alpha, color, fill, shape, size



```
f + geom_quantile()
```

x, y, alpha, color, linetype, size, weight



```
f + geom_rug(sides = "bl")
```

alpha, color, linetype, size



```
f + geom_smooth(method = "lm")
```

x, y, alpha, color, fill, linetype, size, weight



```
f + geom_text(aes(label = cty))
```

x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

#### Discrete X, Continuous Y

```
g <- ggplot(mpg, aes(class, hwy))
```



```
g + geom_bar(stat = "identity")
```

x, y, alpha, color, fill, linetype, size, weight



```
g + geom_boxplot()
```

lower, middle, upper, x, ymax, ymin, alpha, color, fill, linetype, shape, size, weight



```
g + geom_dotplot(binaxis = "y",
  stackdir = "center")
```

x, y, alpha, color, fill



```
g + geom_violin(scale = "area")
```

x, y, alpha, color, fill, linetype, size, weight

#### Discrete X, Discrete Y

```
h <- ggplot(diamonds, aes(cut, color))
```



```
h + geom_jitter()
```

x, y, alpha, color, fill, shape, size

#### Continuous Bivariate Distribution

```
i <- ggplot(movies, aes(year, rating))
```



```
i + geom_bin2d(binwidth = c(5, 0.5))
```

xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight



```
i + geom_density2d()
```

x, y, alpha, colour, linetype, size



```
i + geom_hex()
```

x, y, alpha, colour, fill size

#### Continuous Function

```
j <- ggplot(economics, aes(date, unemployment))
```



```
j + geom_area()
```

x, y, alpha, color, fill, linetype, size



```
j + geom_line()
```

x, y, alpha, color, linetype, size



```
j + geom_step(direction = "hv")
```

x, y, alpha, color, linetype, size

#### Visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4.5, se = 1:2)
k <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))
```



```
k + geom_crossbar(tatten = 2)
```

x, y, ymax, ymin, alpha, color, fill, linetype, size



```
k + geom_errorbar()
```

x, ymax, ymin, alpha, color, linetype, size, width (also `geom_errorbarh()`)



```
k + geom_linerange()
```

x, ymin, ymax, alpha, color, linetype, size



```
k + geom_pointrange()
```

x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

#### Maps

```
data <- data.frame(murder = USArrests$Murder,
  state = tolower(rownames(USArrests)))
```

```
map <- map_data("state")
```

```
l <- ggplot(data, aes(fill = murder))
```



```
l + geom_map(aes(map_id = state), map = map) +
  expand_limits(x = map$long, y = map$lat)
map_id, alpha, color, fill, linetype, size
```

### Three Variables



```
seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))
m <- ggplot(seals, aes(long, lat))
```

```
m + geom_raster(aes(fill = z), hjust = 0.5,
  vjust = 0.5, interpolate = FALSE)
```

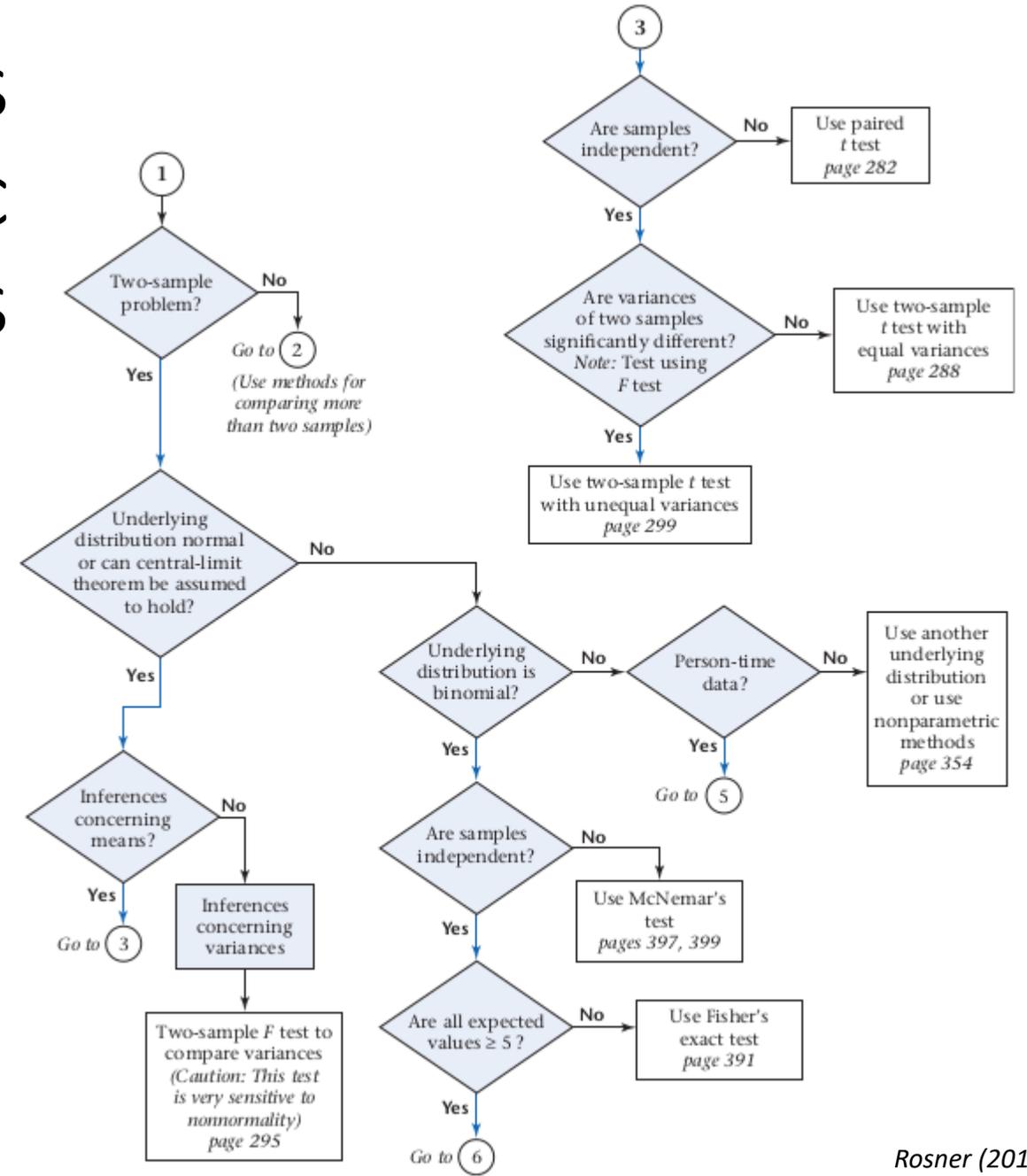
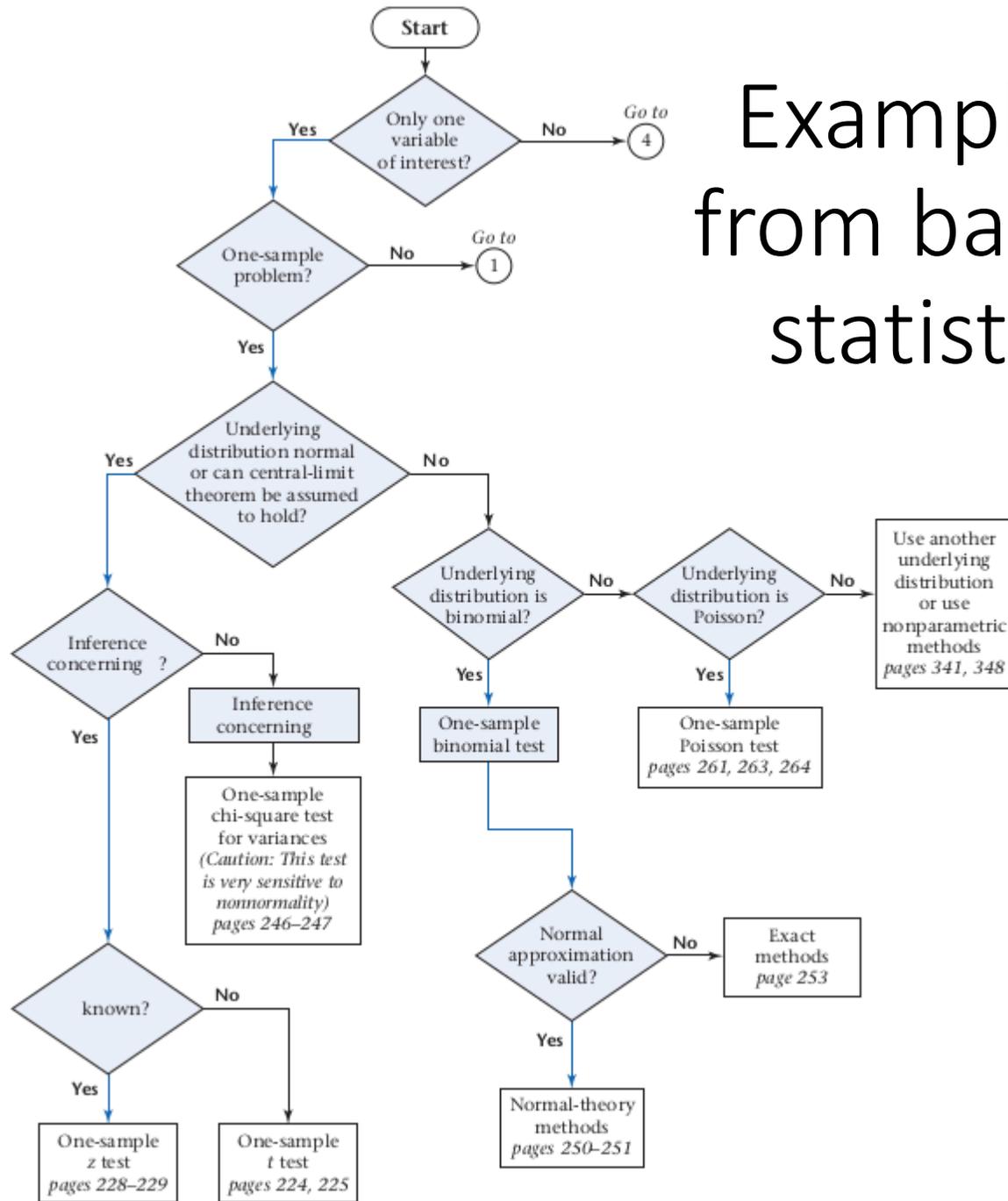
x, y, alpha, fill (fast)

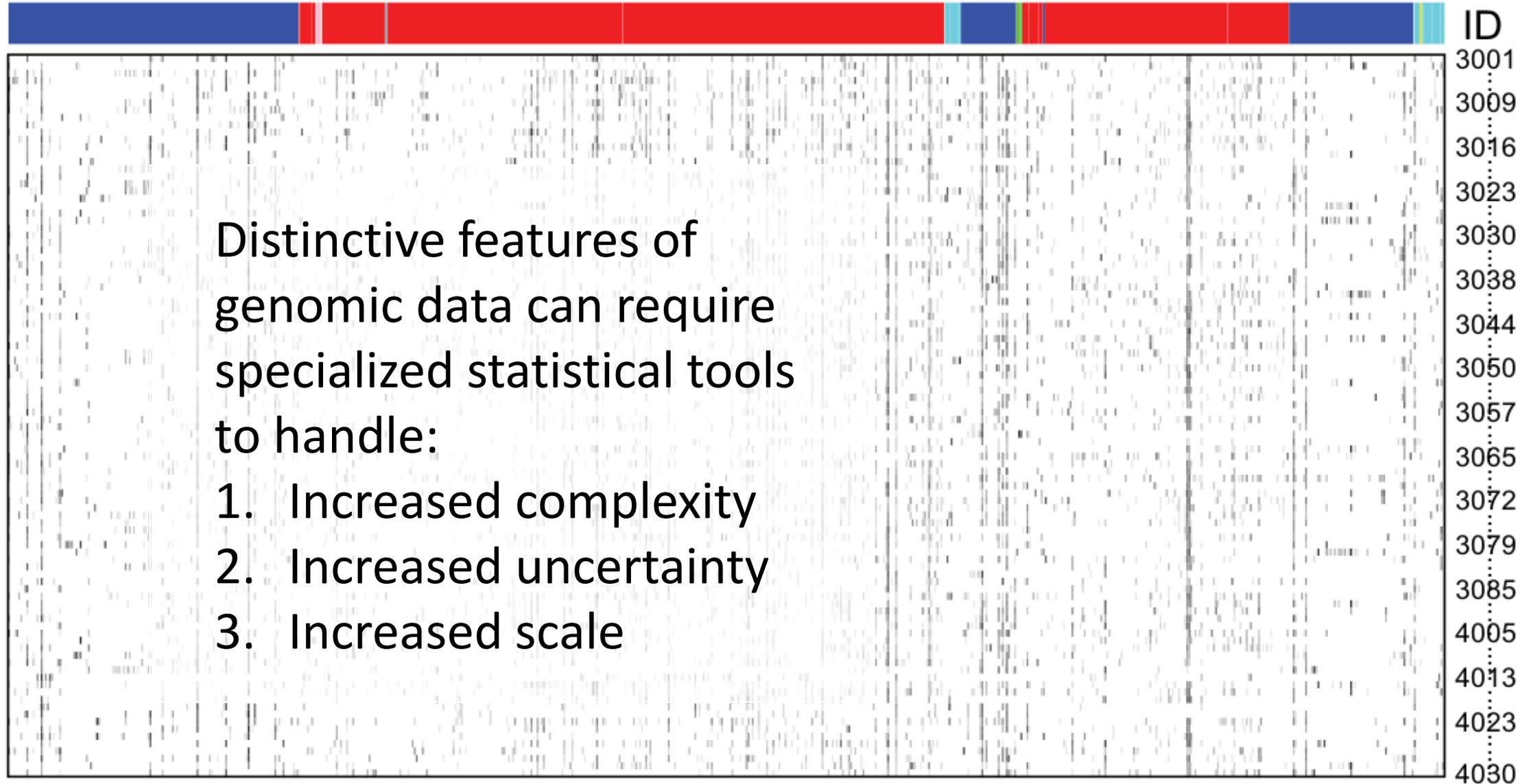


```
m + geom_contour(aes(z = z))
```

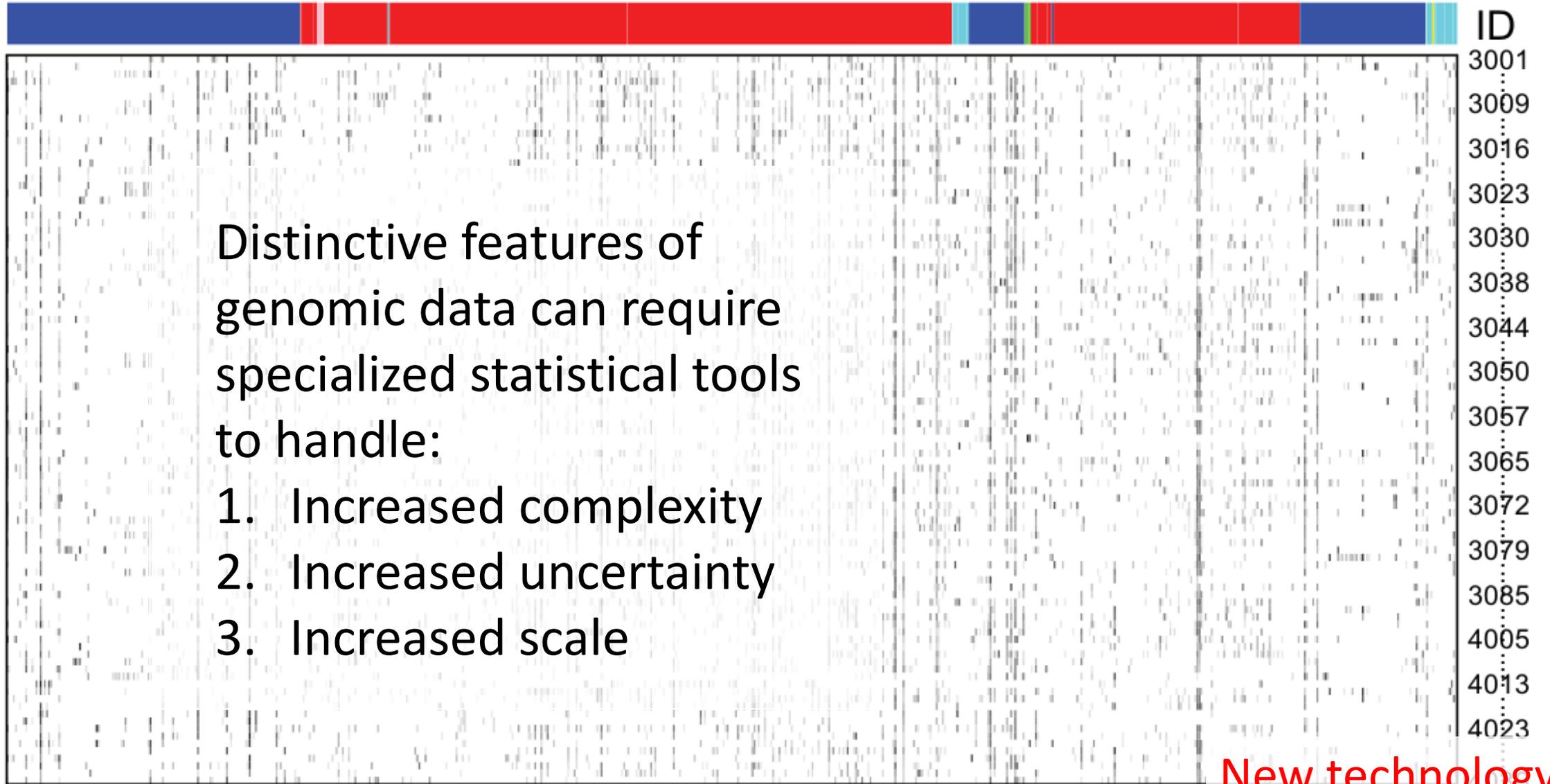
x, y, z, alpha, colour, linetype, size, weight

# Examples from basic statistics





Actinobacteria
  Bacteroidetes
  Firmicutes
  Fusobacteria
  Proteobacteria
  TM7

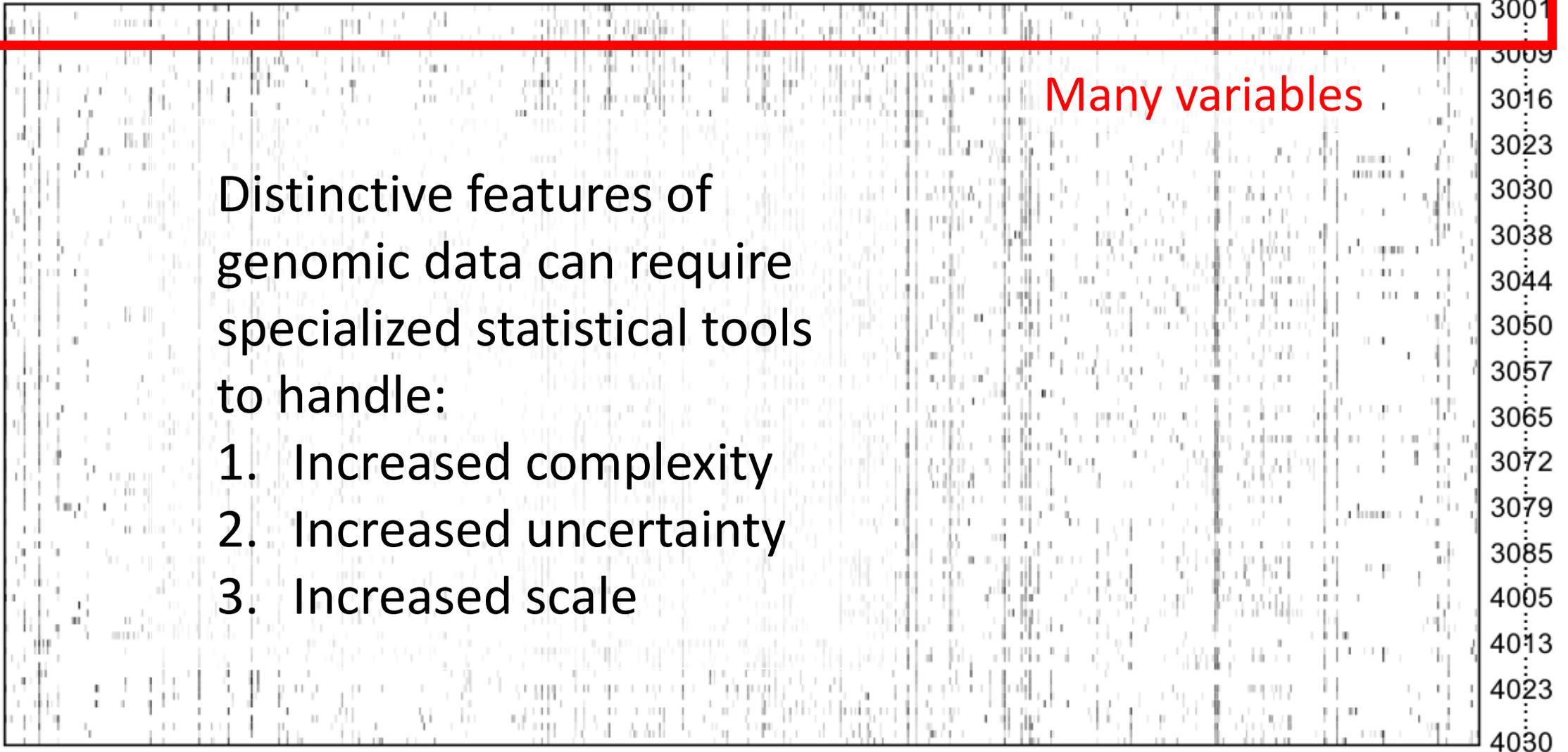


Distinctive features of genomic data can require specialized statistical tools to handle:

1. Increased complexity
2. Increased uncertainty
3. Increased scale

New technology

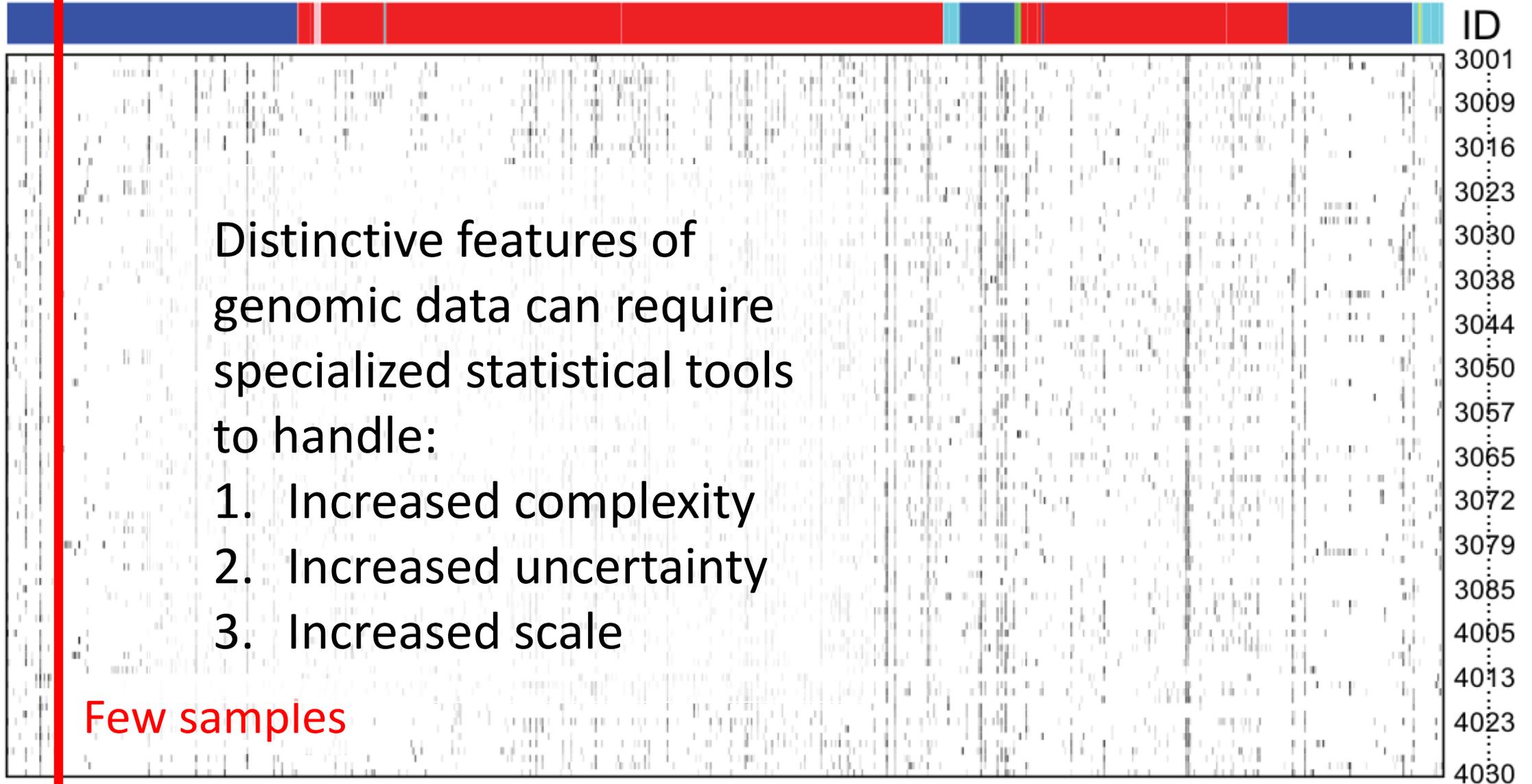




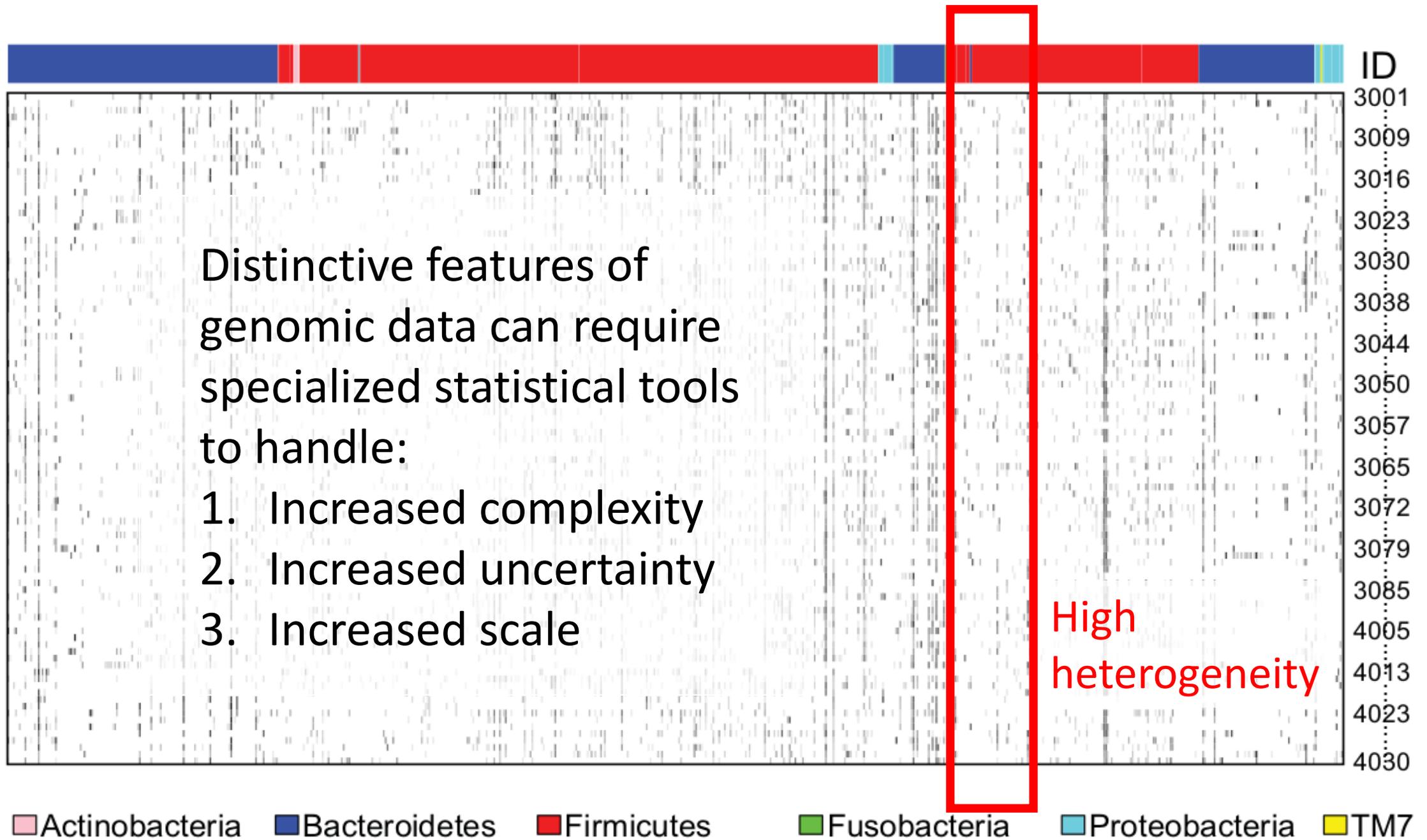
Distinctive features of genomic data can require specialized statistical tools to handle:

1. Increased complexity
2. Increased uncertainty
3. Increased scale

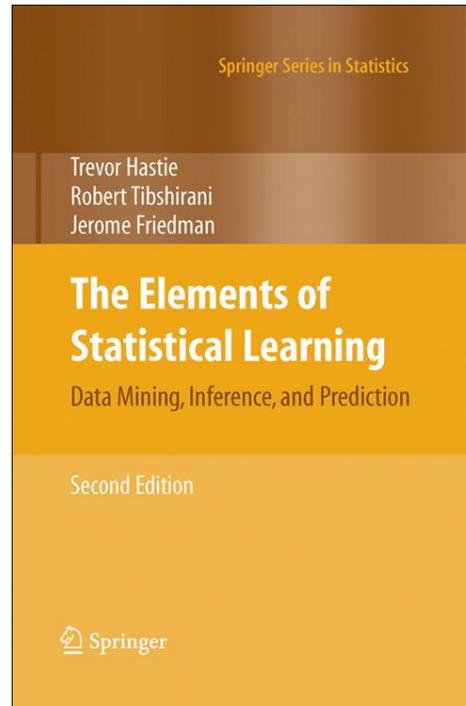
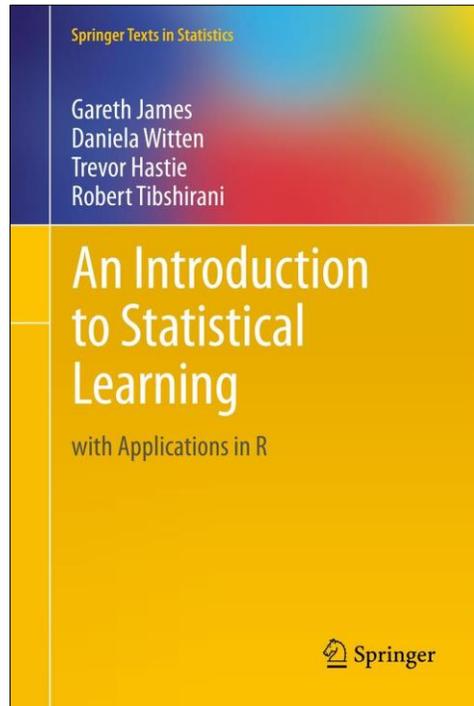




Actinobacteria
  Bacteroidetes
  Firmicutes
  Fusobacteria
  Proteobacteria
  TM7



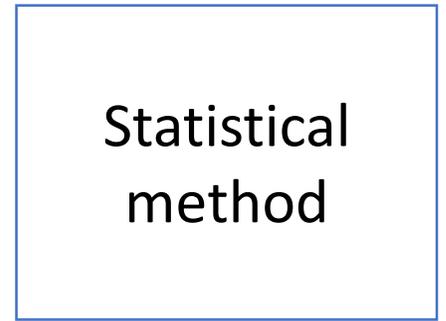
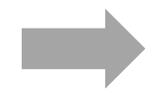
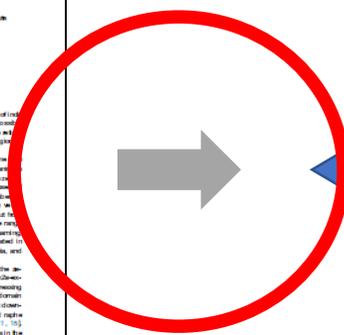
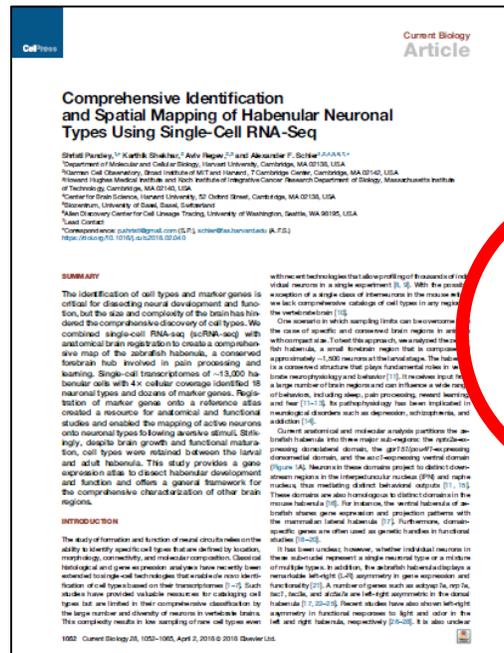
# We will discuss some statistical tools that are frequently used in genomics



We will also discuss a general framework for organizing new tools.



# Choosing the right tool for a given biological question requires creativity and experience



There are other frameworks that are organized by biological question rather than statistical tool

The image shows three browser windows illustrating course content organized by biological question rather than statistical tool.

- Left Window:** [www.biostat.jhsph.edu/~iruczins/teach](http://www.biostat.jhsph.edu/~iruczins/teach). The page lists various topics, each with a PDF icon:
  - Introduction to statistical genomics
  - Summarizing and presenting genomic data
  - Statistical modeling I : means and two-group comparisons
  - Multiple hypothesis testing
  - Differential expression
  - Pathway and gene set analyses
  - Experimental design
  - Dimension reduction
  - Batch effects
  - Statistical modeling II : linear models in genomics
  - Statistical modeling III : pre-processing genomic data
- Middle Window:** [https://www.jmp.com/en\\_us/academic/jmpg-com](https://www.jmp.com/en_us/academic/jmpg-com). The page is titled "Step-by-Step Guides" and lists several guides:
  - Step-by-Step Guide for Genetics QK Modeling #1: K
  - Step-by-Step Guide for Genetics QK Modeling #2: F
  - Step-by-Step Guide for Genetics QK Modeling #3: A
  - Step-by-Step Guide for Genetics QK Modeling #4: U
  - Step-by-Step Guide for Genetics QK Modeling #5: K
  - Step-by-Step Guide for Expression #1: Importing an
  - Step-by-Step Guide for Expression #2: Predictive M
  - Model Comparison for Model Selection
  - Step-by-Step Guide for Expression #3: Learning Cu
  - Samples Needed
  - Step-by-Step Guide for Expression #4: Subset Data
  - Step-by-Step Guide for Expression #5: Final Model
  - Step-by-Step Guide for Importing Expression Data i
  - Step-by-Step Guide for Importing Genetics Data into
  - Additional Step-by-Step Guides
- Right Window:** <https://onlinecourses.science.psu.edu/statprogram/stat555>. The page is titled "Course Topics" and lists the topics to be covered:

The topics that will be covered in this course will likely include:

  1. Introduction to R and RStudio
  2. Introduction to cell biology
  3. Introduction to measurement technologies: microarrays, sequencing, SNPs and ChIP
  4. Basic statistics
  5. Gene Expression Microarrays: experimental designs, preprocessing and normalization, differential expression.
  6. RNA-seq: experimental designs, preprocessing and normalization, differential expression, splice variants
  7. SNPs
  8. ChIPs
  9. Replication and pooling
  10. Gene Set enrichment analysis
  11. Clustering samples and genes
  12. Classifying samples using statistical machine learning

Example analysis using R



# To illustrate these tools, we will analyze single-cell RNA-seq data in R

CellPress

Current Biology  
Article

## Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq

Shristi Pandey,<sup>1,\*</sup> Karthik Shekhar,<sup>2</sup> Aviv Regev,<sup>2,3</sup> and Alexander F. Schier<sup>1,2,4,5,6,7,\*</sup>

<sup>1</sup>Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA  
<sup>2</sup>Klarman Cell Observatory, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA  
<sup>3</sup>Howard Hughes Medical Institute and Koch Institute of Integrative Cancer Research Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA  
<sup>4</sup>Center for Brain Science, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA  
<sup>5</sup>Biozentrum, University of Basel, Basel, Switzerland  
<sup>6</sup>Allen Discovery Center for Cell Lineage Tracing, University of Washington, Seattle, WA 98195, USA  
<sup>7</sup>Lead Contact

\*Correspondence: [p.shristi@gmail.com](mailto:p.shristi@gmail.com) (S.P.), [schier@fas.harvard.edu](mailto:schier@fas.harvard.edu) (A.F.S.)  
<https://doi.org/10.1016/j.cub.2018.02.040>

# Anatomy of a basic R command

```
pandey =  
read.table("GSM2818521_larva_counts_matrix.txt")
```

- Case-sensitive
- **Function()**: performs pre-programmed calculations given **inputs and options**
- **Variable**: stores values and outputs of function, name cannot contain whitespace and cannot start with a special character

# Basic preprocessing of single-cell RNA-seq data using Seurat

```
library(Seurat)

s_obj = CreateSeuratObject(counts = pandey,
min.cells = 3, min.features = 200)

s_obj = NormalizeData(s_obj)

s_obj = FindVariableFeatures(s_obj)

s_obj = ScaleData(s_obj)
```

# Statistical methods for genomics

# Where statistics appears in a standard genomic analysis workflow

1. Experimental design ← Not covered today
  2. Quality control
  3. Preprocessing
  4. Normalization and batch correction
  5. Analysis ← The focus of today's discussion
  6. Biological interpretation
- Uses statistics but is highly dependent on technology

# Classifying statistical tools

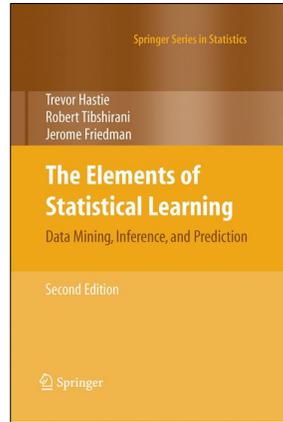
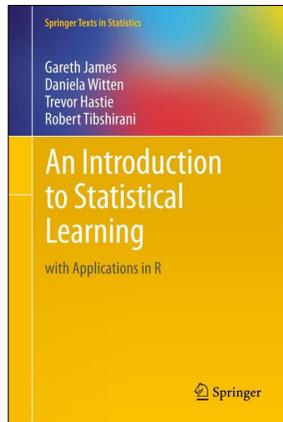
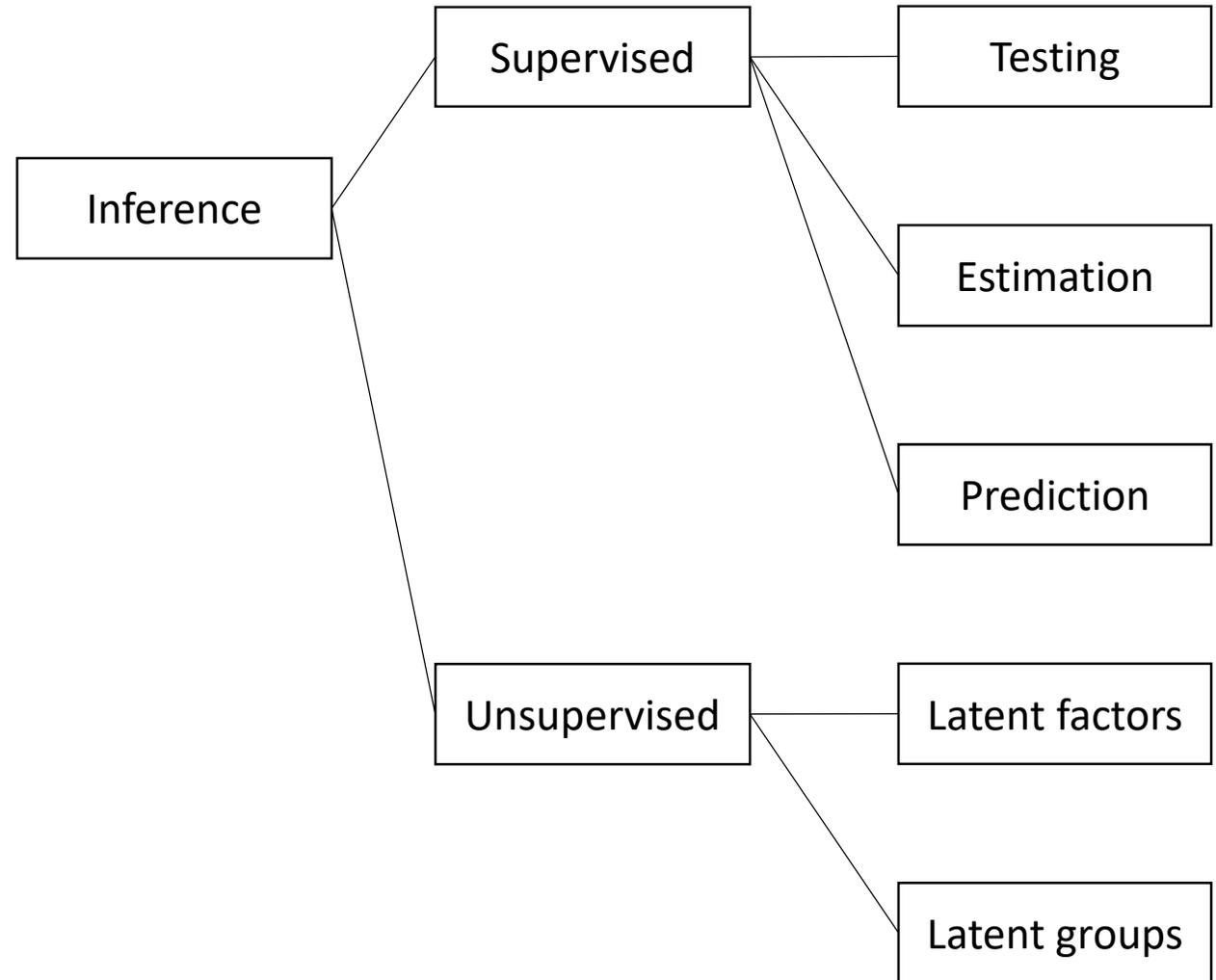
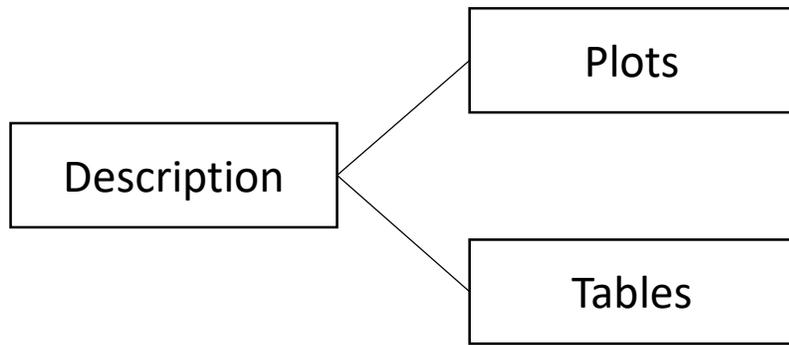
## Data structure

## Statistical task

	No dependent variables	Continuous outcome	Censored outcomes	Etc.
Visualize				
Identify latent factors				
Cluster observations				
Select features				
Etc.				

**APPROPRIATE  
STATISTICAL  
METHODS**

# Classifying statistical tasks



# Classifying data structures

- Can vary widely, and classification is difficult
- Important factors in genomics:
  1. Data type
  2. Number of samples relative to number of variables

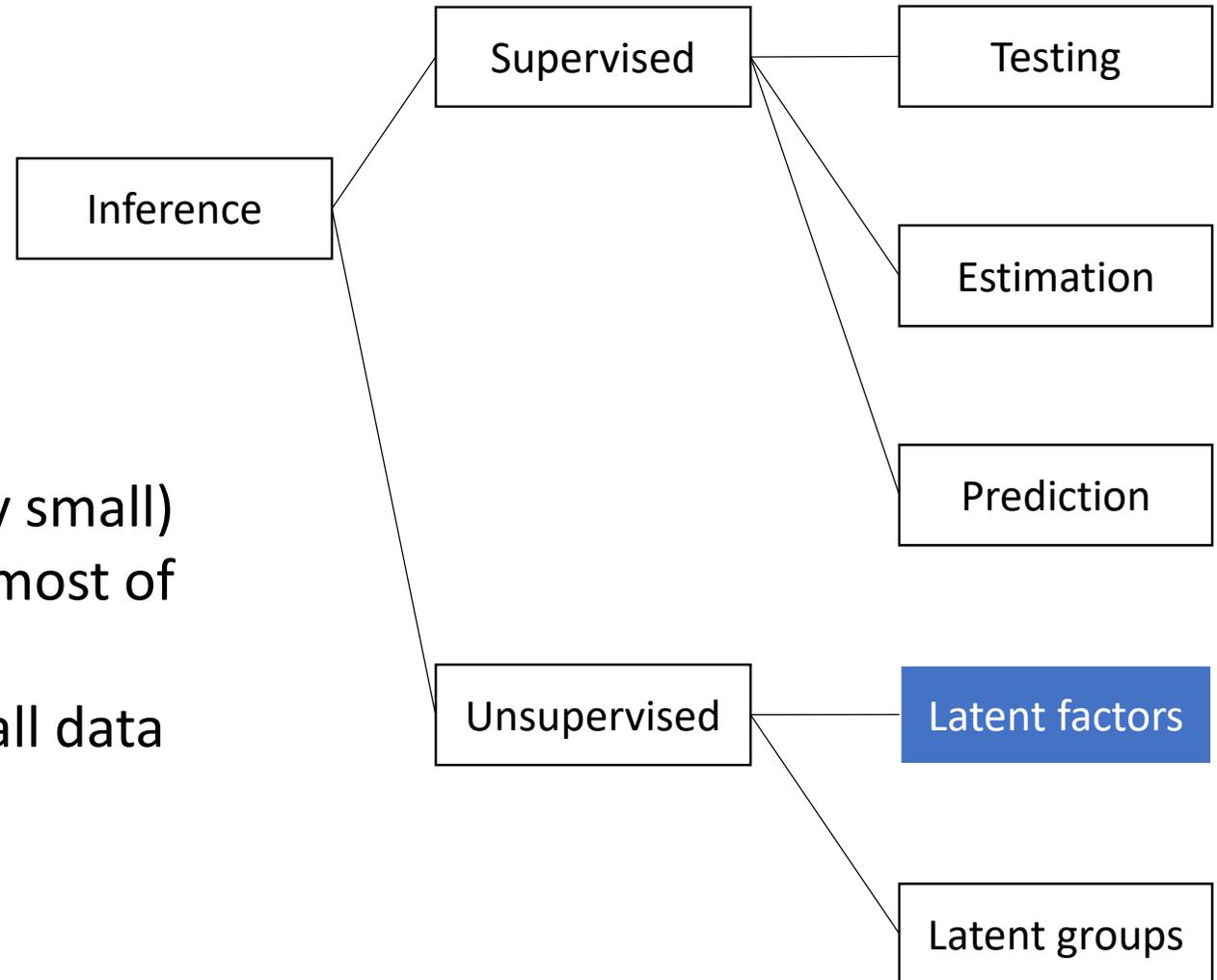
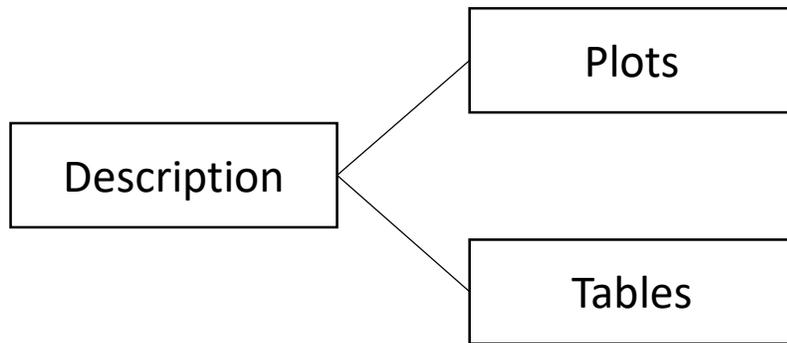
# PCA

```
> dim(pandey)
[1] 24105  4365
> pandey[1:3, 1:2]
      larvalR2_AAACCTGAGACAGAGA.1 larvalR2_AAACCTGAGACTTTCG.1
SYN3                             0                             0
PTPRO                             1                             0
EPS8                              0                             0
```

## Research question:

Can the gene expression information be summarized in fewer features?

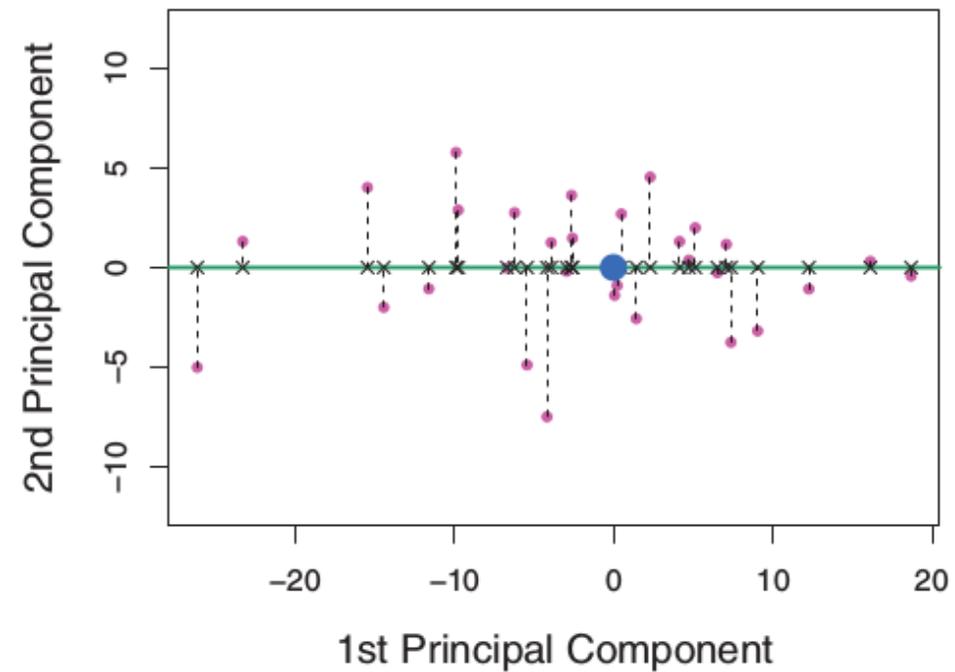
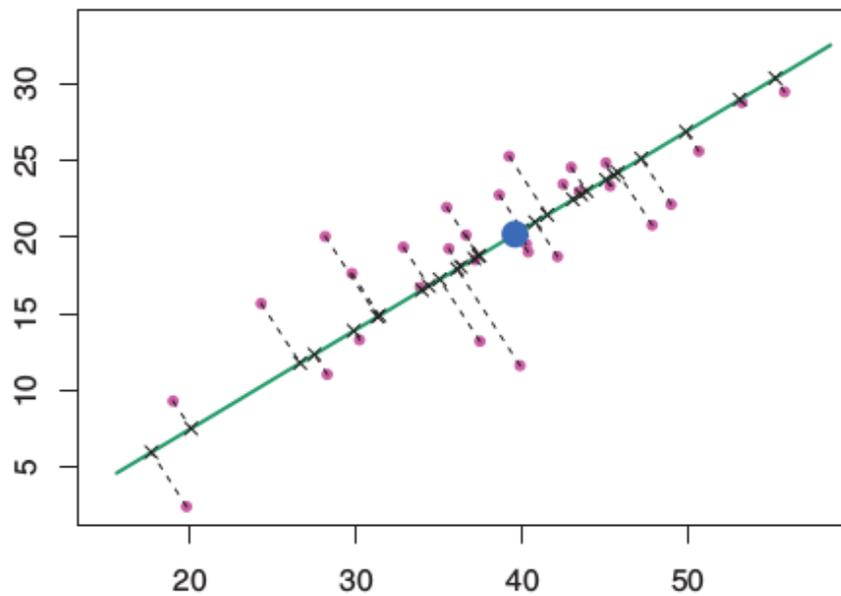
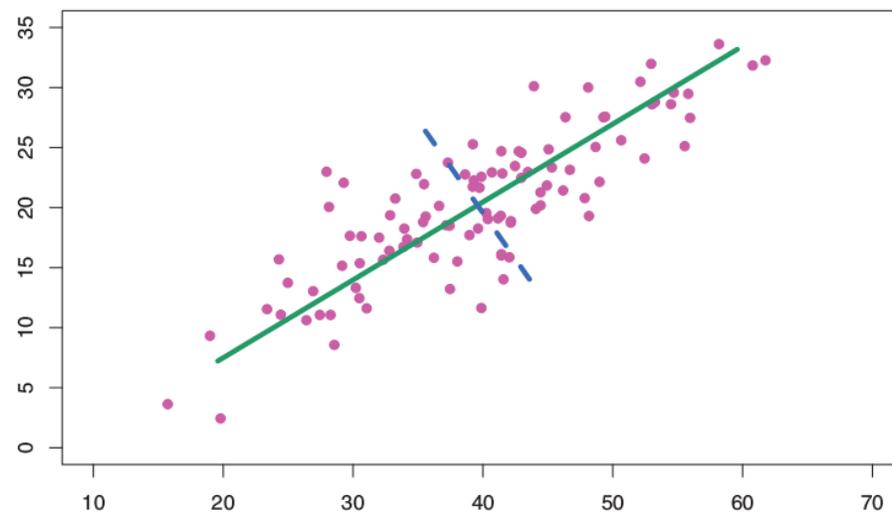
# PCA



**Statistical task:** calculate a (usually small) set of latent factors that captures most of the information in the dataset

**Data structure:** can be applied to all data structures

# PCA



# PCA using Seurat

```
s_obj = RunPCA(s_obj)
```

# Graph clustering

## Research question:

How many cell types exist in the larval zebrafish habenula?

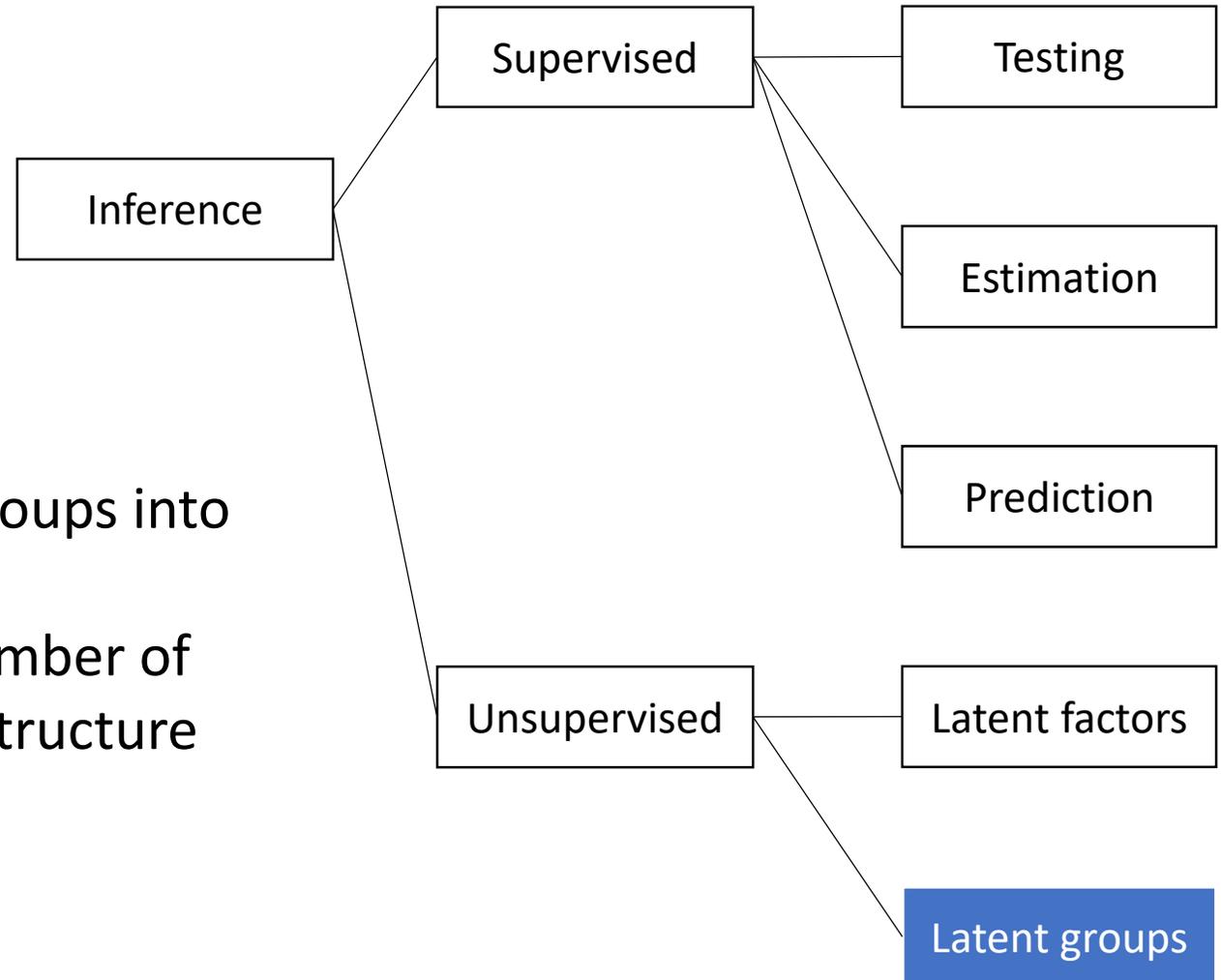
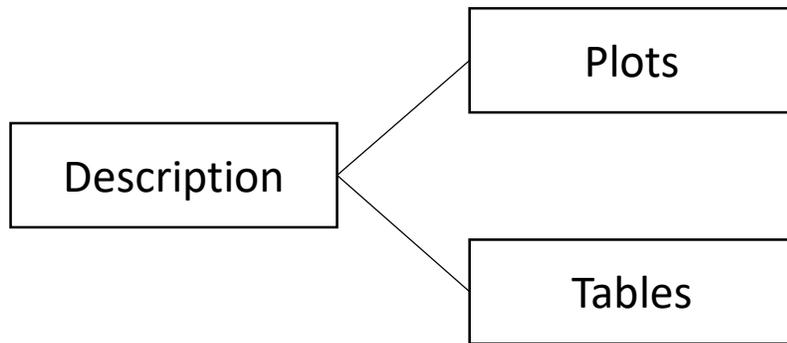
Cell Systems  
**Voices**

CellPress

### What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism?

What Is an Adult Cell Type, Really?	Defining Cell Type Space	Cellular Demographics, Recorded
		
<b>Hans Clevers</b> Hubrecht Institute	<b>Susanne Rafelski</b> Allen Institute for Cell Science	<b>Michael Elowitz</b> Caltech
The human body is home to hundreds of cell types. Some are rather unobtrusive; others,	Canonical cell types, e.g., muscle and nerve, were originally defined by the functions of	It seems to me that we are at the beginning of a paradigm shift on the issue of cell type.

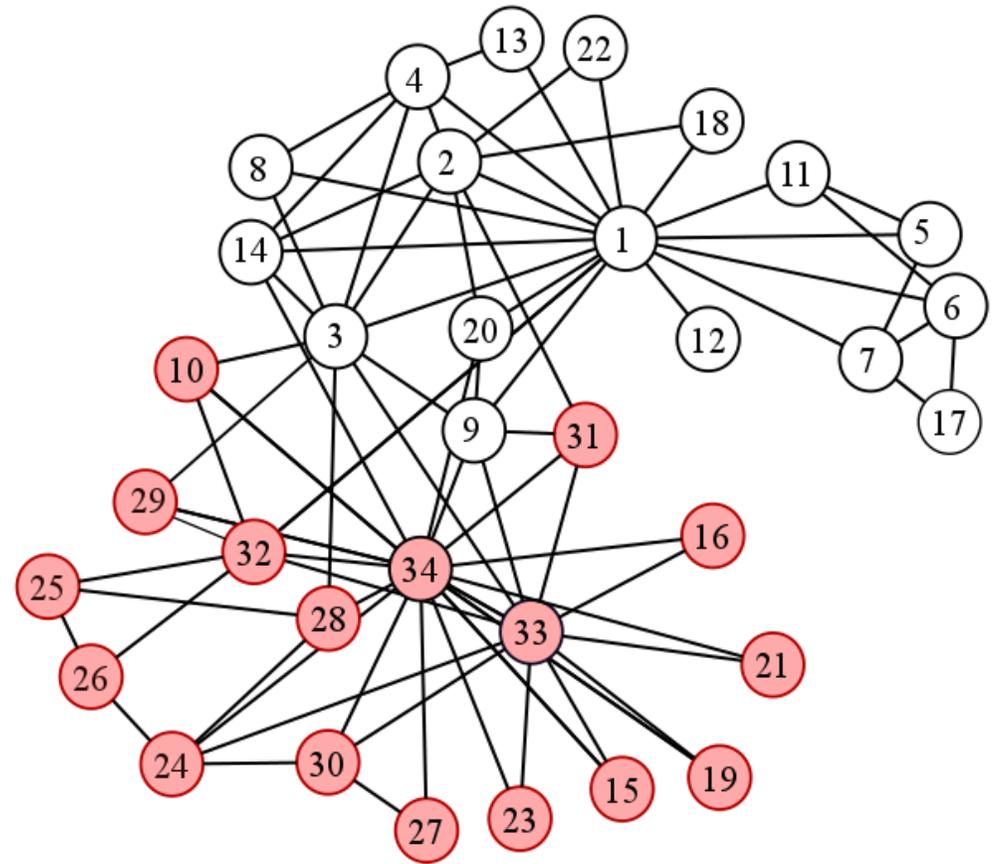
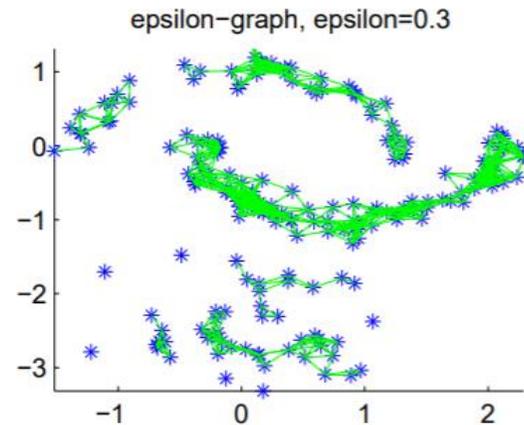
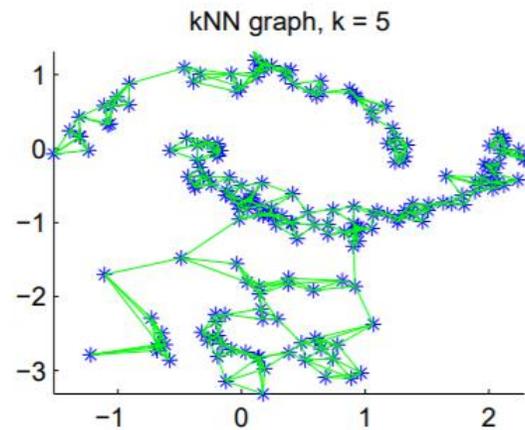
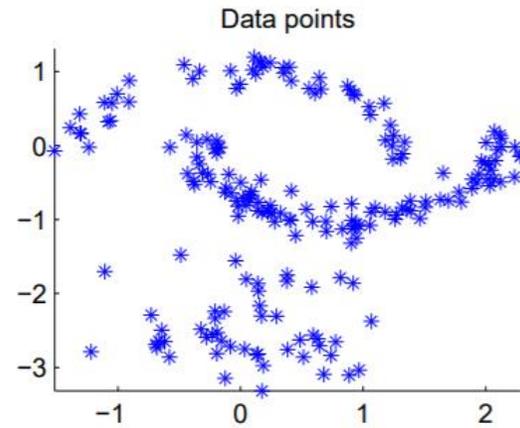
# Graph clustering



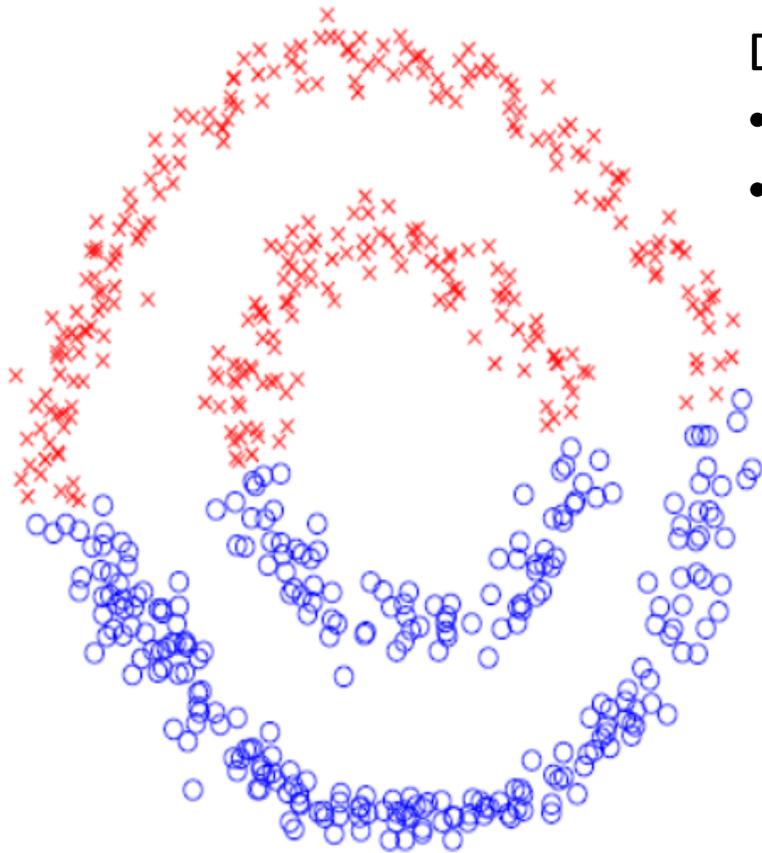
**Statistical task:** construct latent groups into which the observations fall

**Data structure:** relatively small number of features and complicated cluster structure

# Graph clustering



# Comparison to other clustering methods



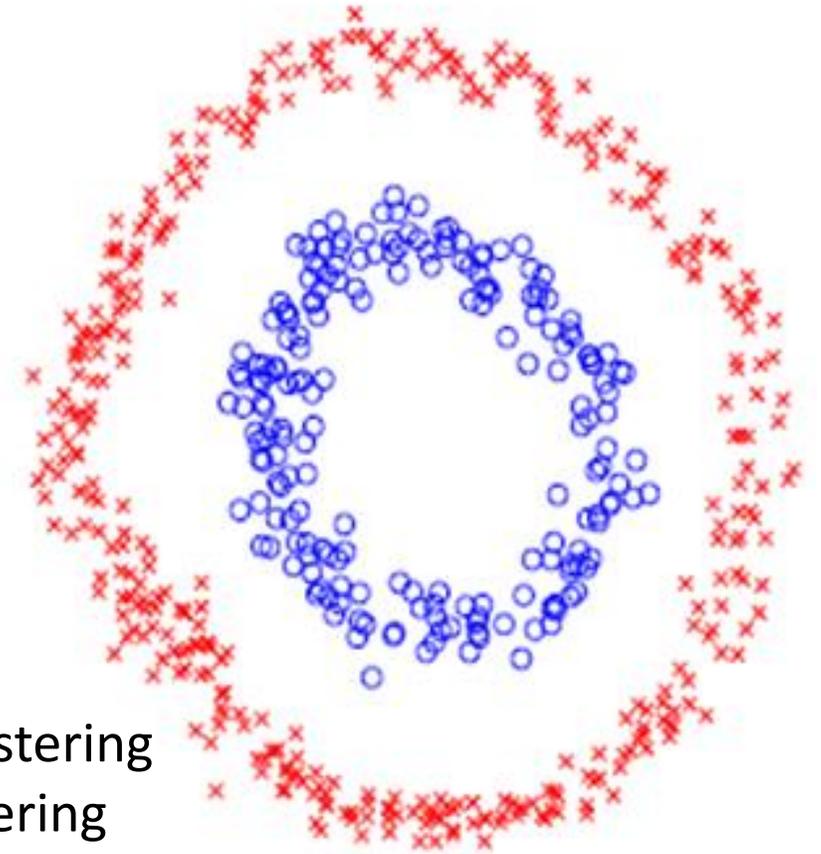
Distance

- Hierarchical clustering
- K-means clustering

vs.

Relationship

- Spectral clustering
- Graph clustering



# Graph clustering using Seurat

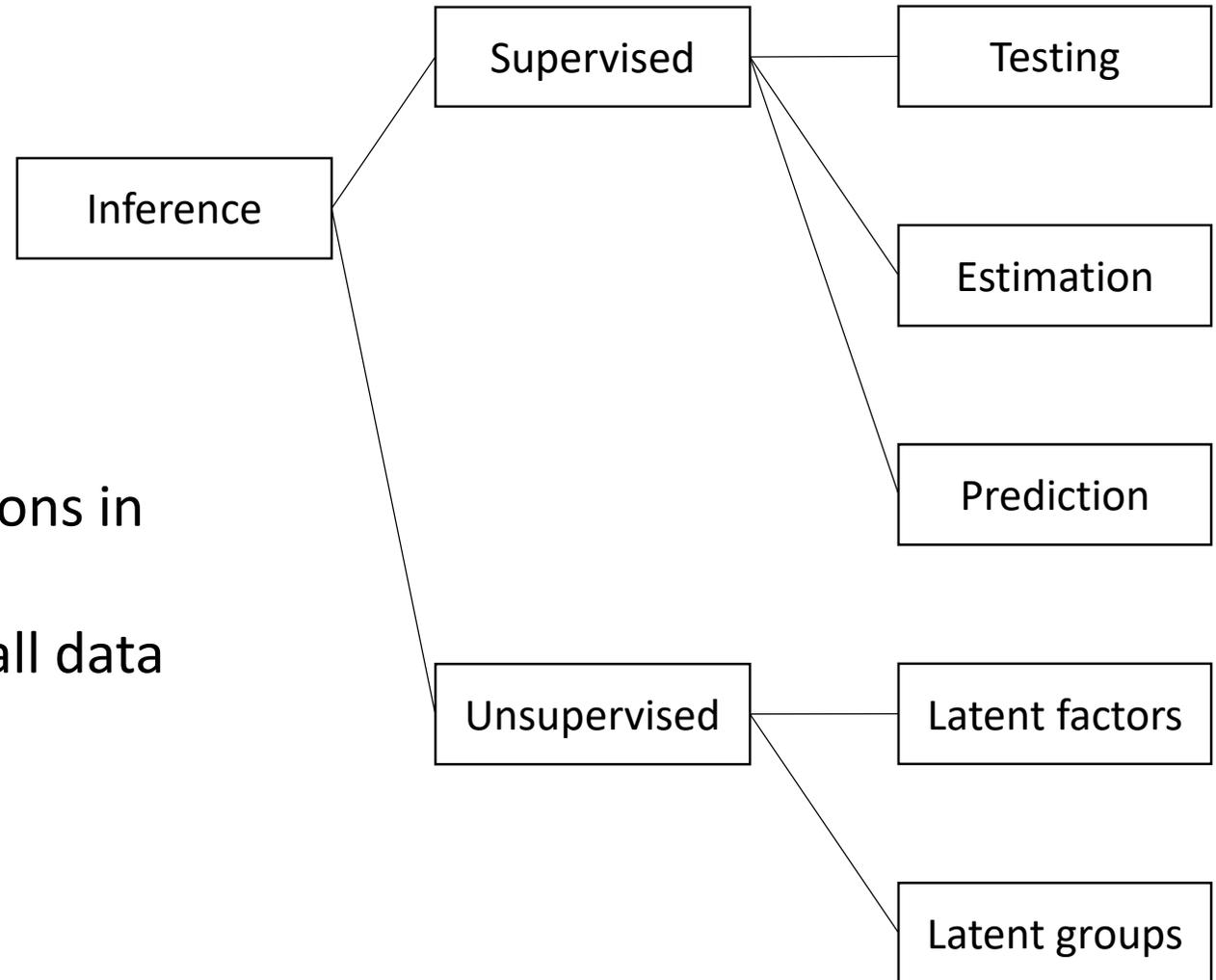
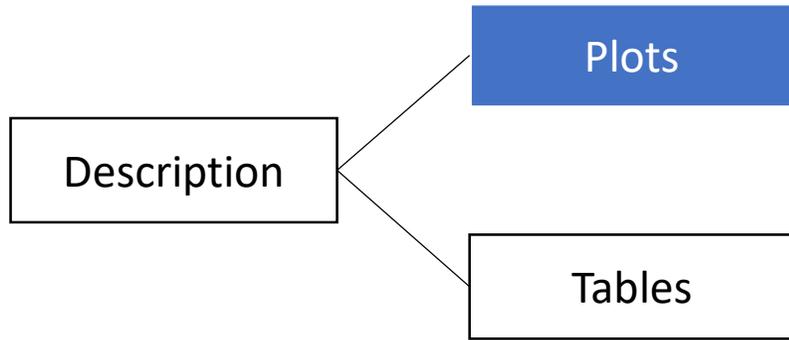
```
s_obj = FindNeighbors(s_obj, dims = 1:10)  
s_obj = FindClusters(s_obj, dims = 1:10,  
resolution = 0.1)
```

# t-SNE plot

## **Research question:**

How to visualize the different cell types?

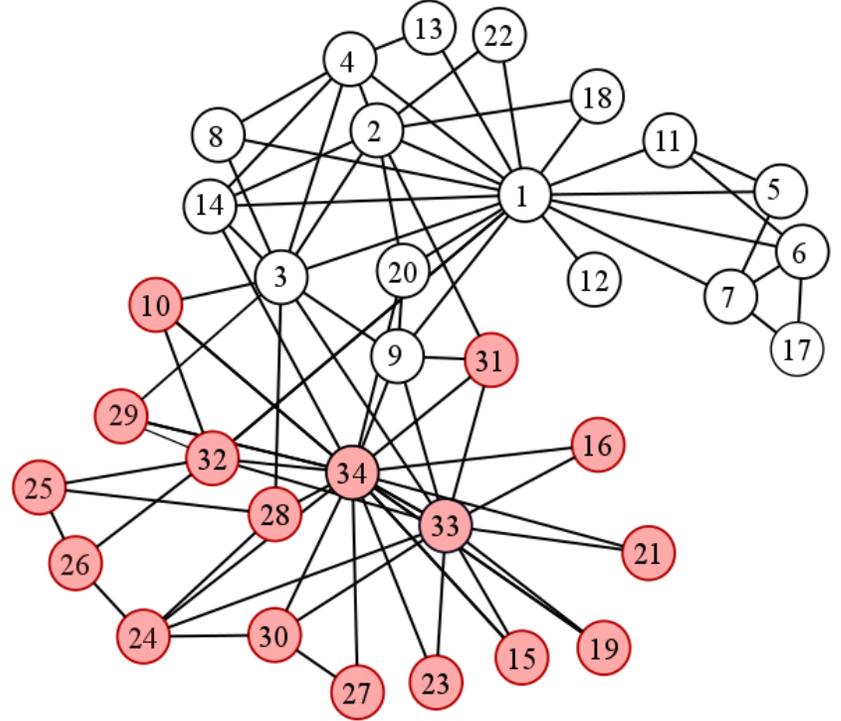
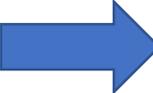
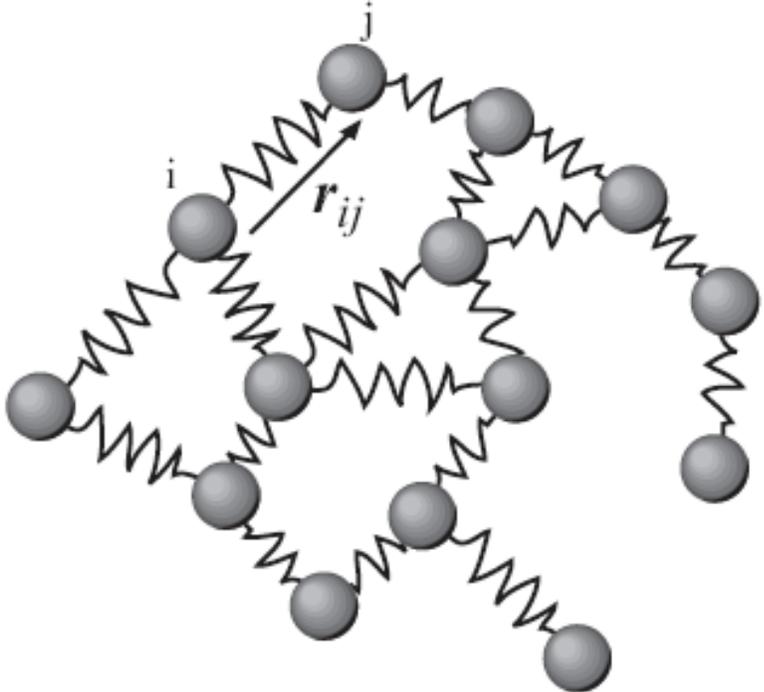
# t-SNE plot



**Statistical task:** visualize observations in low dimensions

**Data structure:** can be applied to all data structures

# t-SNE plot



# t-SNE plot

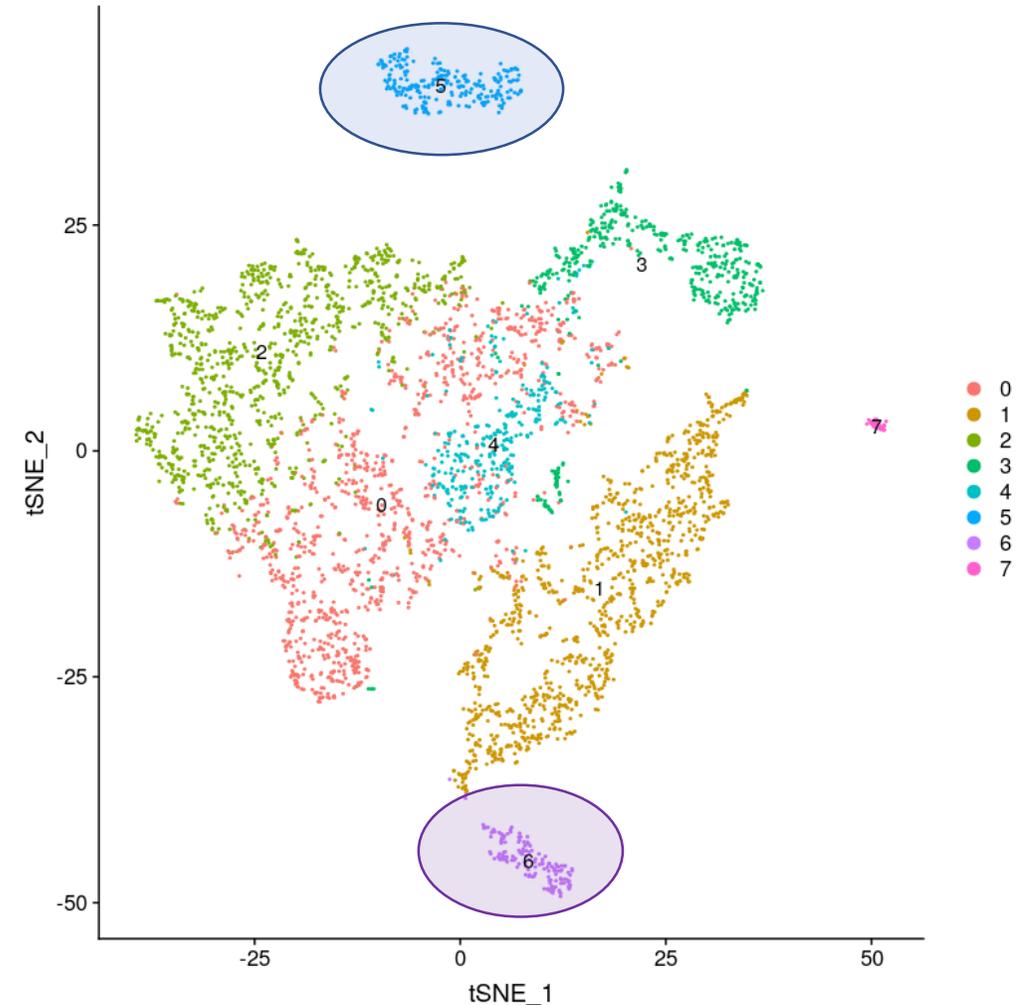
```
s_obj = RunTSNE(s_obj)
```

```
DimPlot(s_obj, reduction = "tsne", label = TRUE)
```

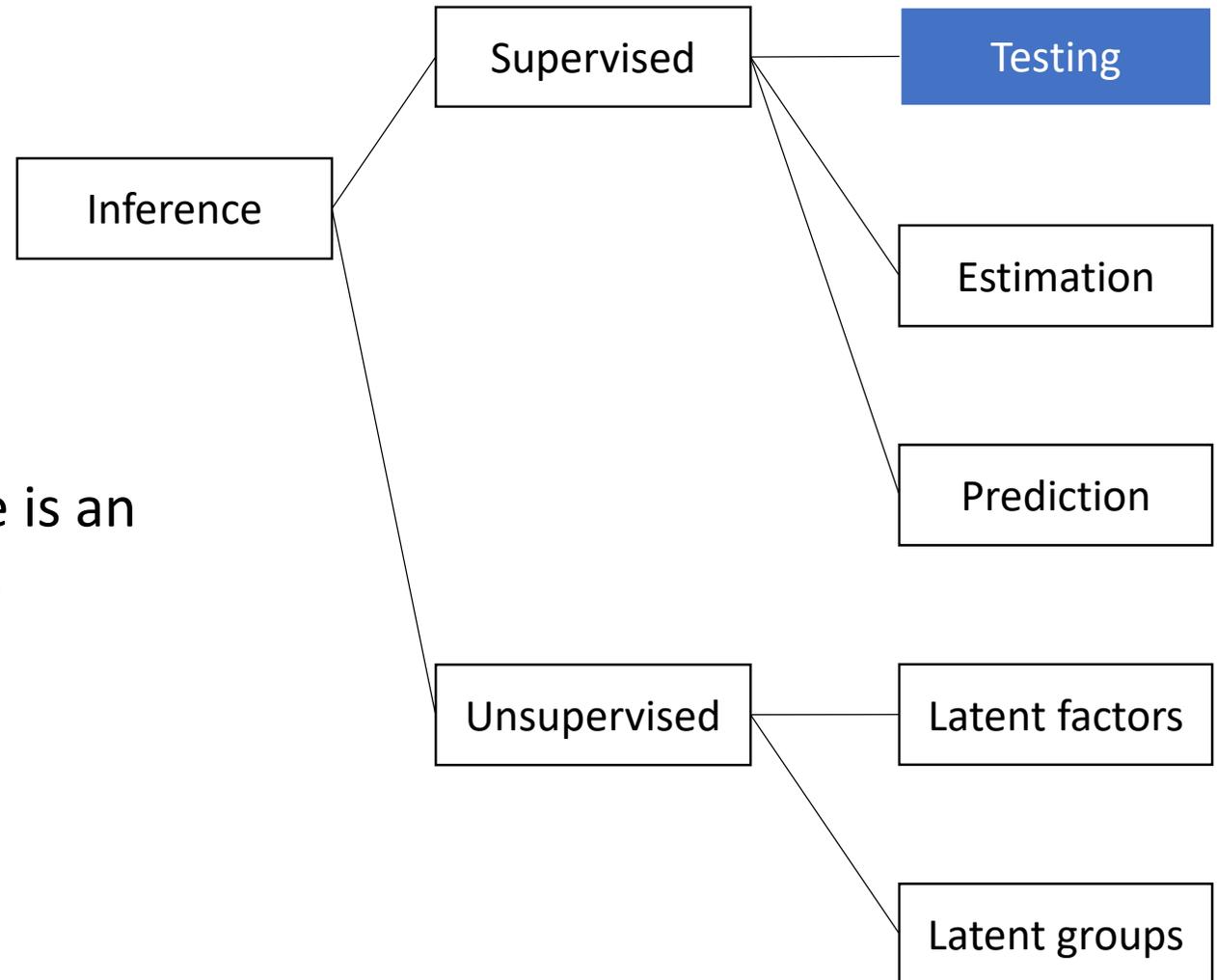
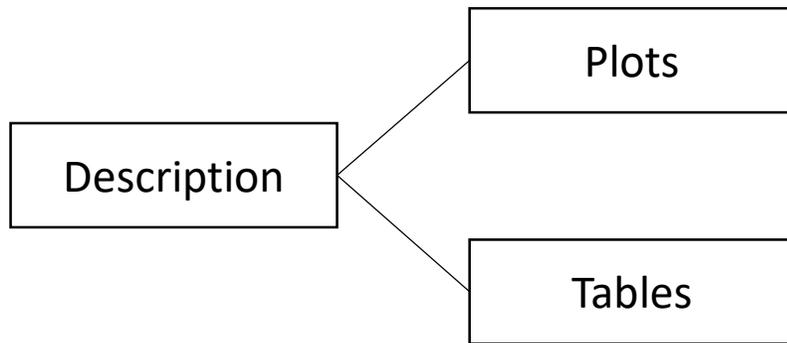
# Wilcoxon test

## Research question:

Does the expression of the gene PDYN differ between clusters 5 and 6?



# Wilcoxon test



**Statistical task:** test whether there is an association between two variables

**Data structure:** one variable is dichotomous

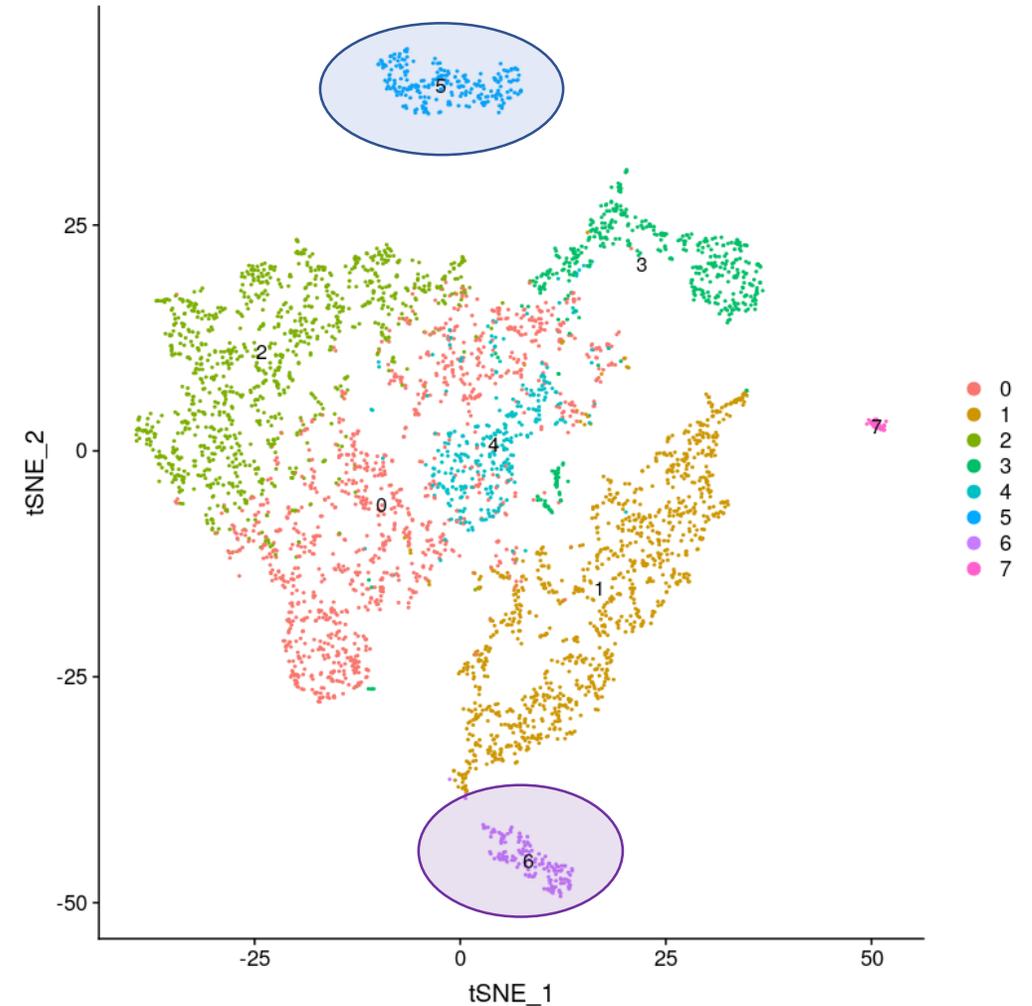
# Wilcoxon test using Seurat

```
markers = FindMarkers(s_obj, ident.1 = 5,  
ident.2 = 6)  
markers["PDYN",]
```

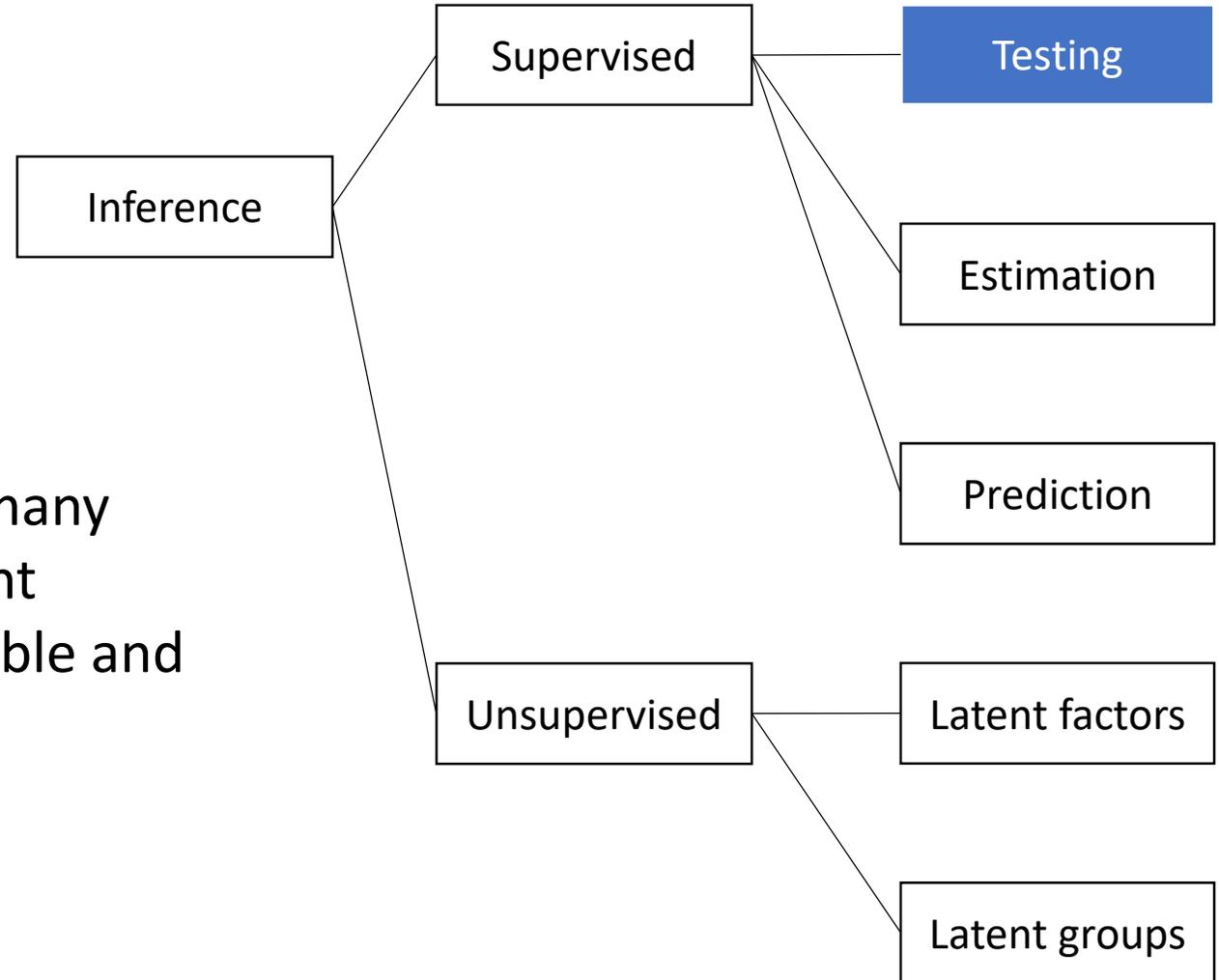
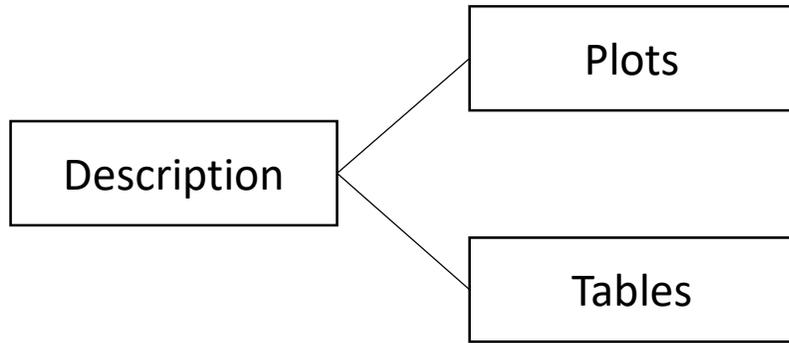
# FDR control

## Research question:

Which genes differ between clusters 5 and 6?



# FDR control



**Statistical task:** identify which of many hypothesis tests are truly significant

**Data structure:** p-values are available and statistically independent

# FDR control

- FDR = False Discovery Rate = expected value of

$$\frac{\text{\# false discoveries}}{\text{total \# of discoveries}}$$

- Statistical methods reject the largest number of hypothesis tests while maintaining  $\text{FDR} \leq \alpha$ , for some preset  $\alpha$

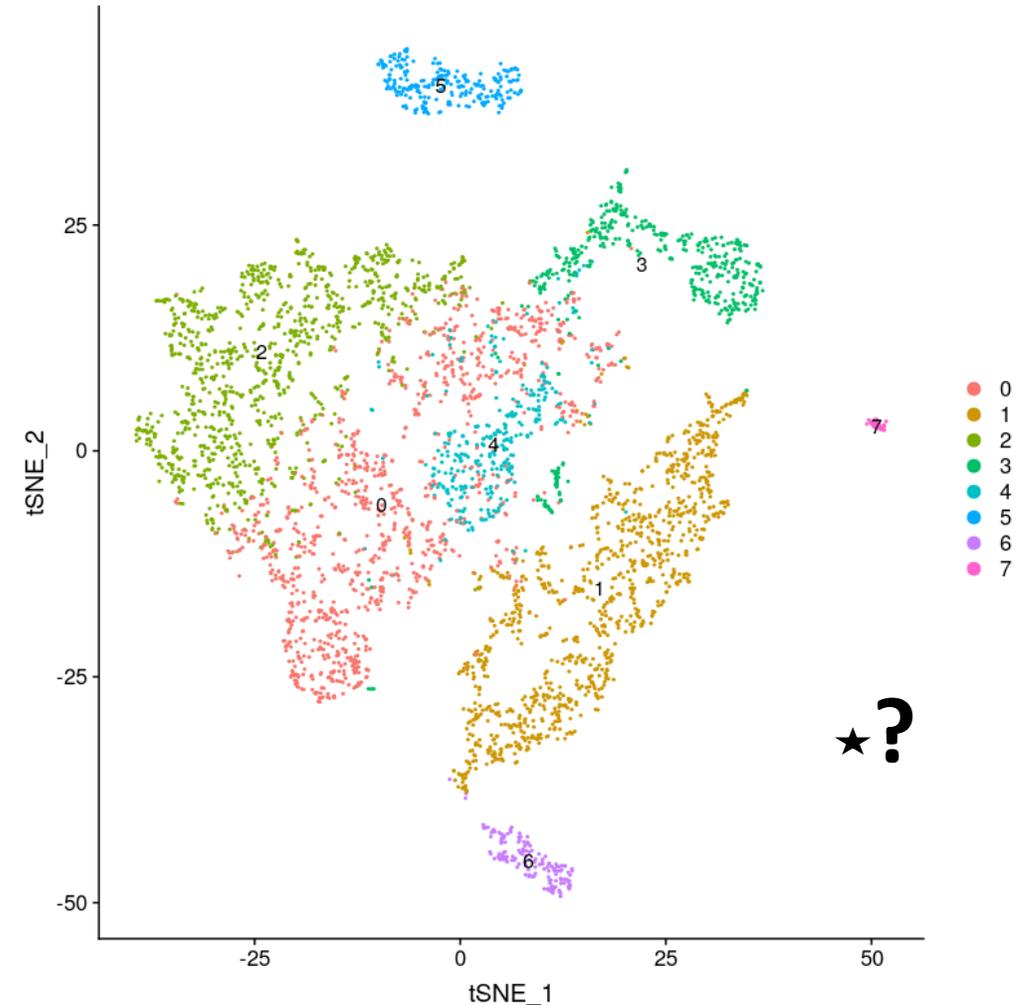
# FDR control using Seurat

```
markers = FindMarkers(s_obj, ident.1 = 5,  
ident.2 = 6)  
head(markers)  
sum(markers$p_val_adj <= 0.05)
```

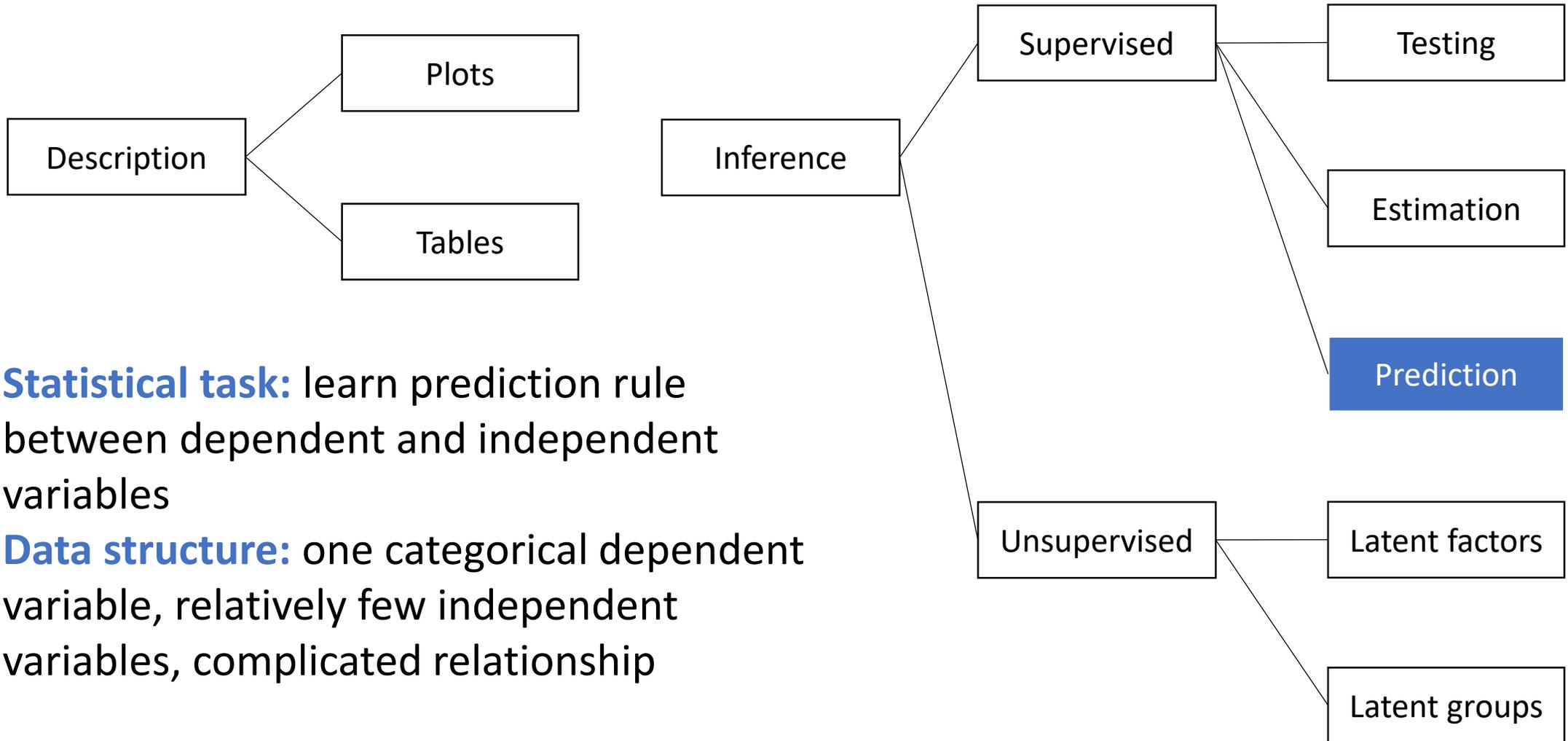
# Random forest classification

## Research question:

Given the principal components of the RNA-seq expression values of all genes from a new cell, how can we determine the cell's type?



# Random forest classification

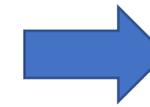
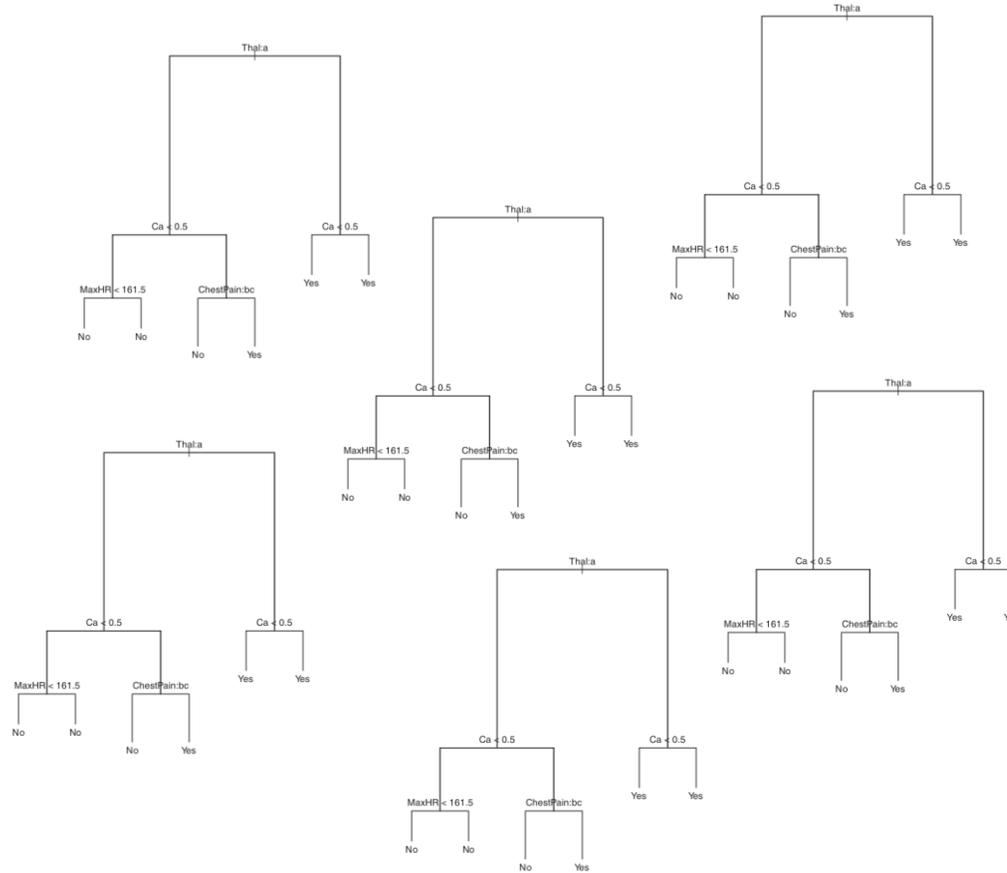
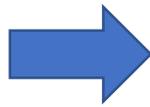


**Statistical task:** learn prediction rule between dependent and independent variables

**Data structure:** one categorical dependent variable, relatively few independent variables, complicated relationship

# Random forest classification

Independent variables



Prediction

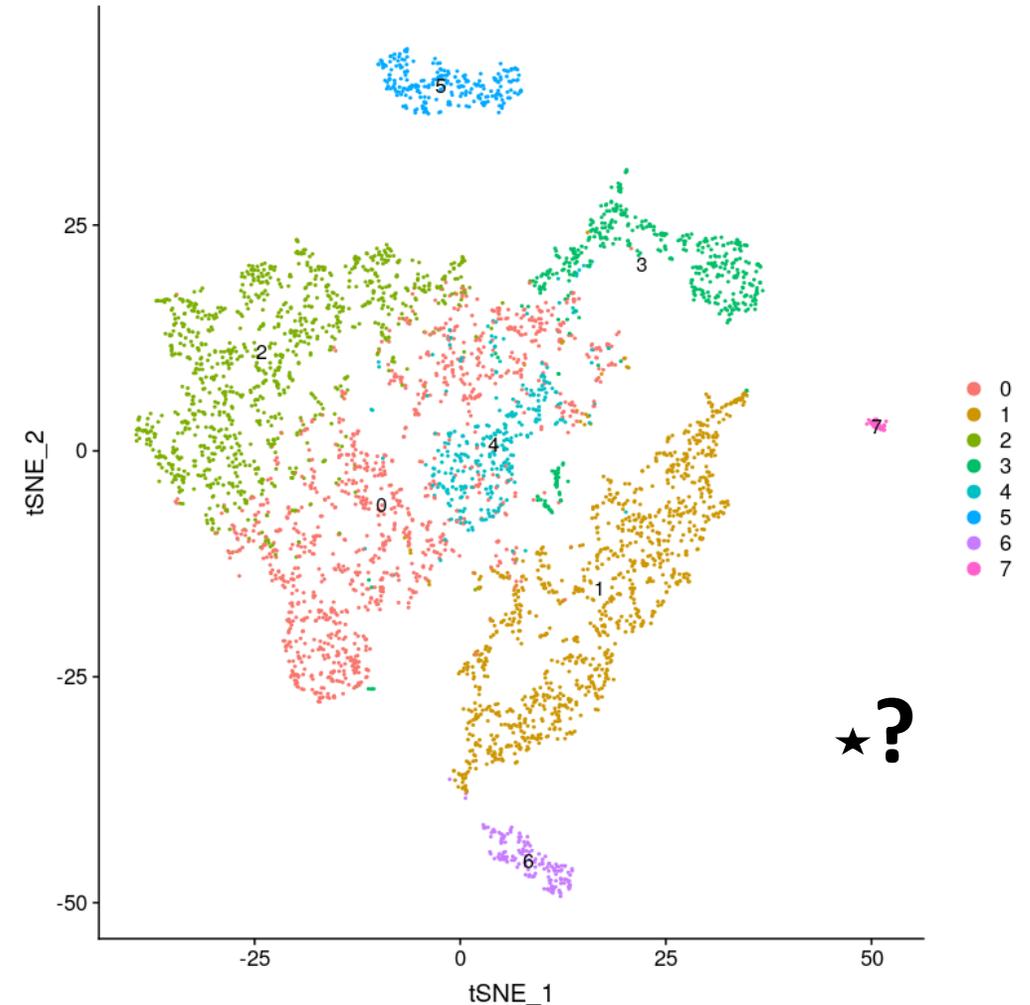
# Random forest classification using caret and ranger

```
library(caret)
pcs = Embeddings(s_obj, reduction = "pca")[-(1:2), 1:10]
class = as.factor(Idsents(s_obj))[-(1:2)]
dataset = data.frame(class, pcs)
rf_fit = train(class ~ .,
               data = dataset,
               method = "ranger",
               trControl = trainControl(method = "cv", number
= 3))
new_pcs = Embeddings(s_obj, reduction = "pca")[1:2, 1:10]
predict(rf_fit, new_pcs)
Idsents(s_obj)[1:2]
```

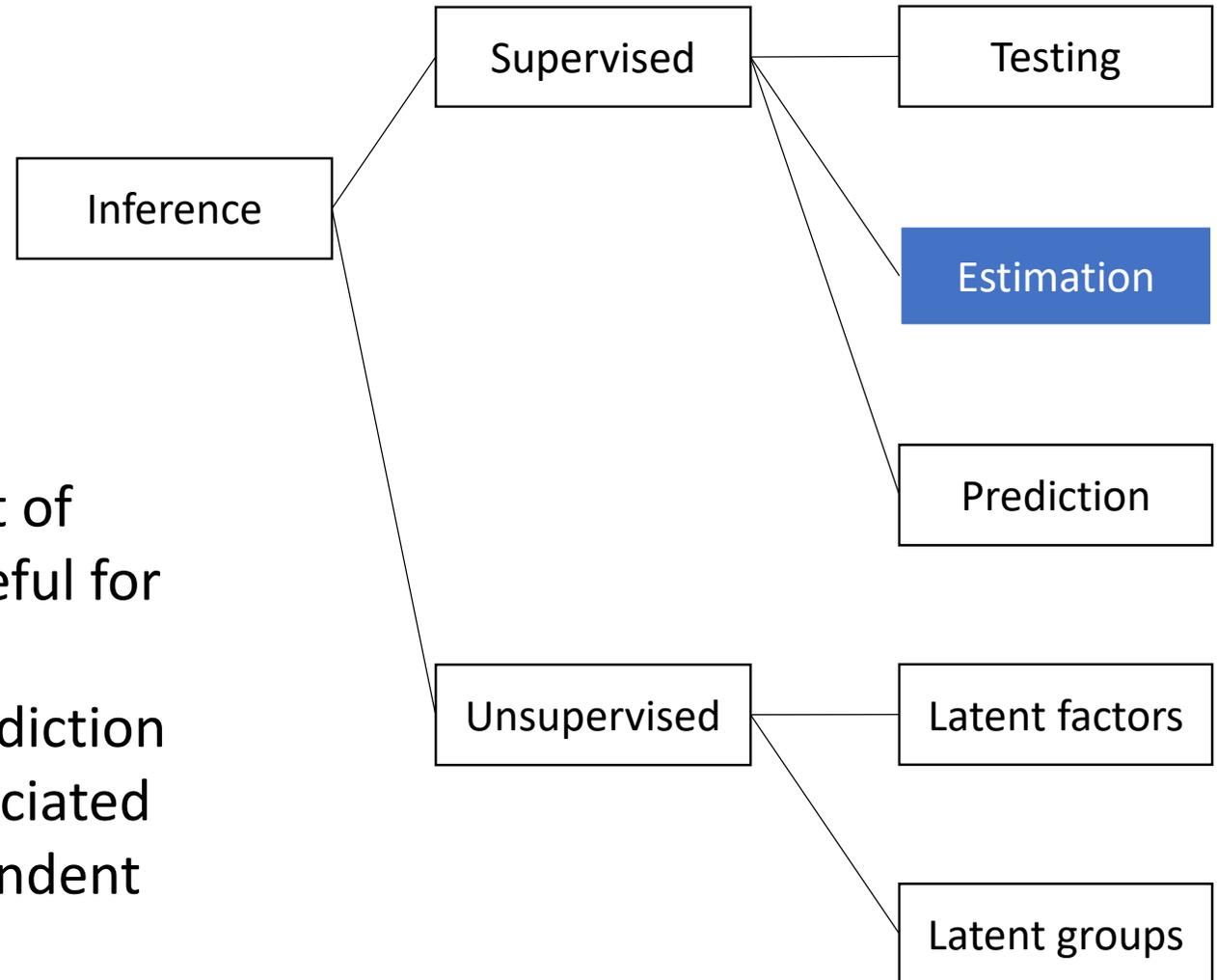
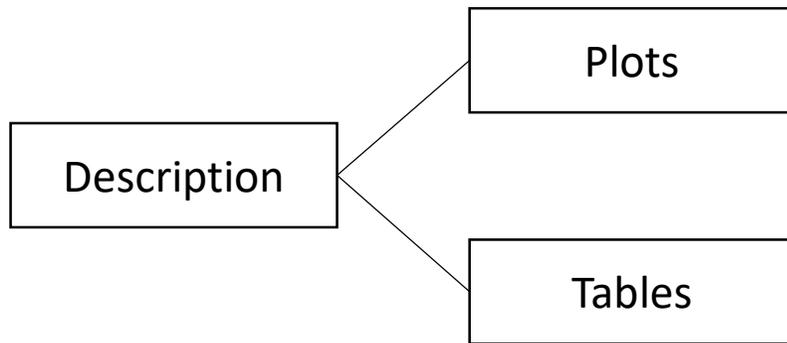
# Lasso

## Research question:

If we can only measure the expression of 10 genes in a new cell, which should we measure in order to most accurately predict the cell's type?



# Lasso



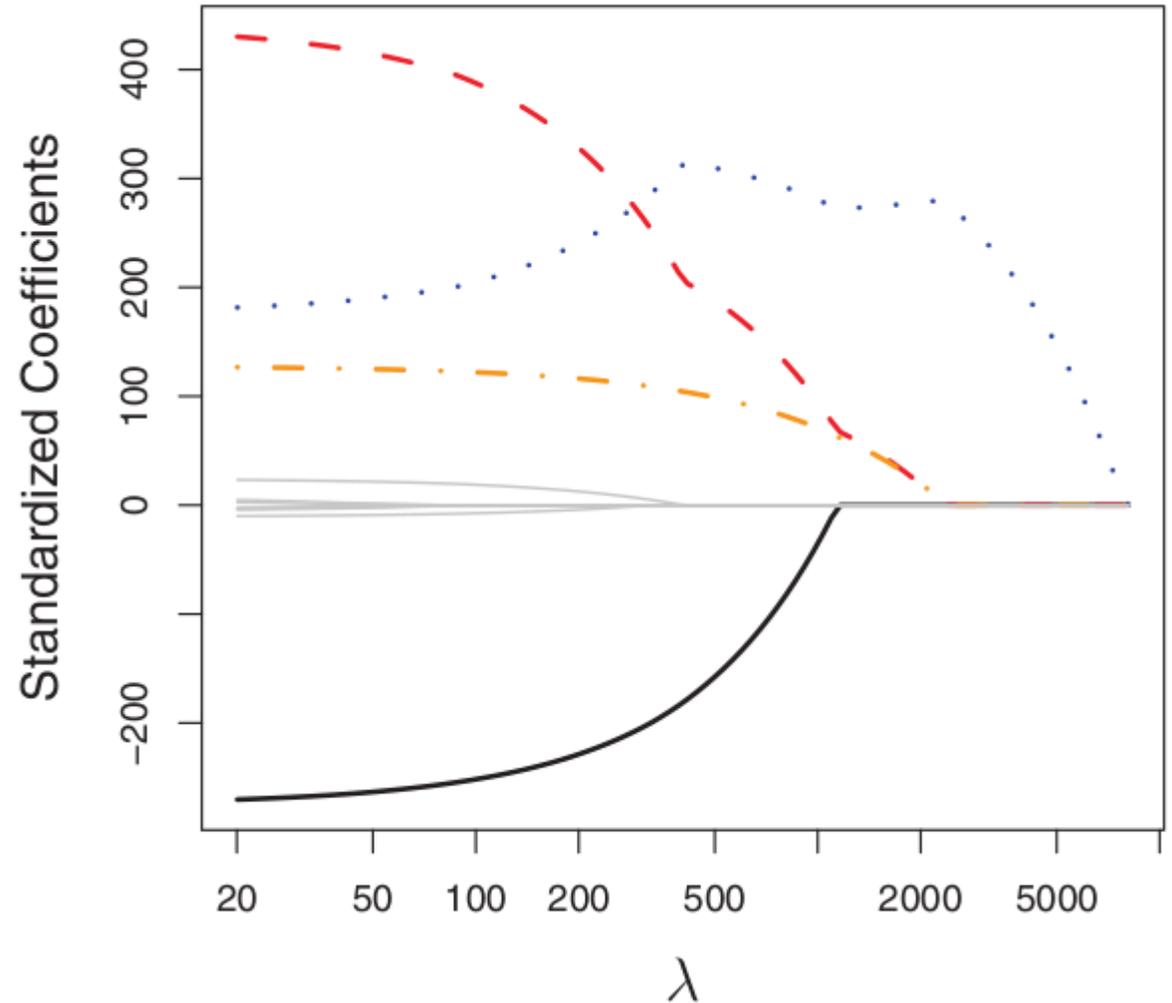
**Statistical task:** identify a small set of independent variables that are useful for predicting dependent variables

**Data structure:** simple (linear) prediction rule, dependent variables are associated with only a few (unknown) independent variables

# Lasso

Estimates coefficients in the regression model

$$Y = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p$$

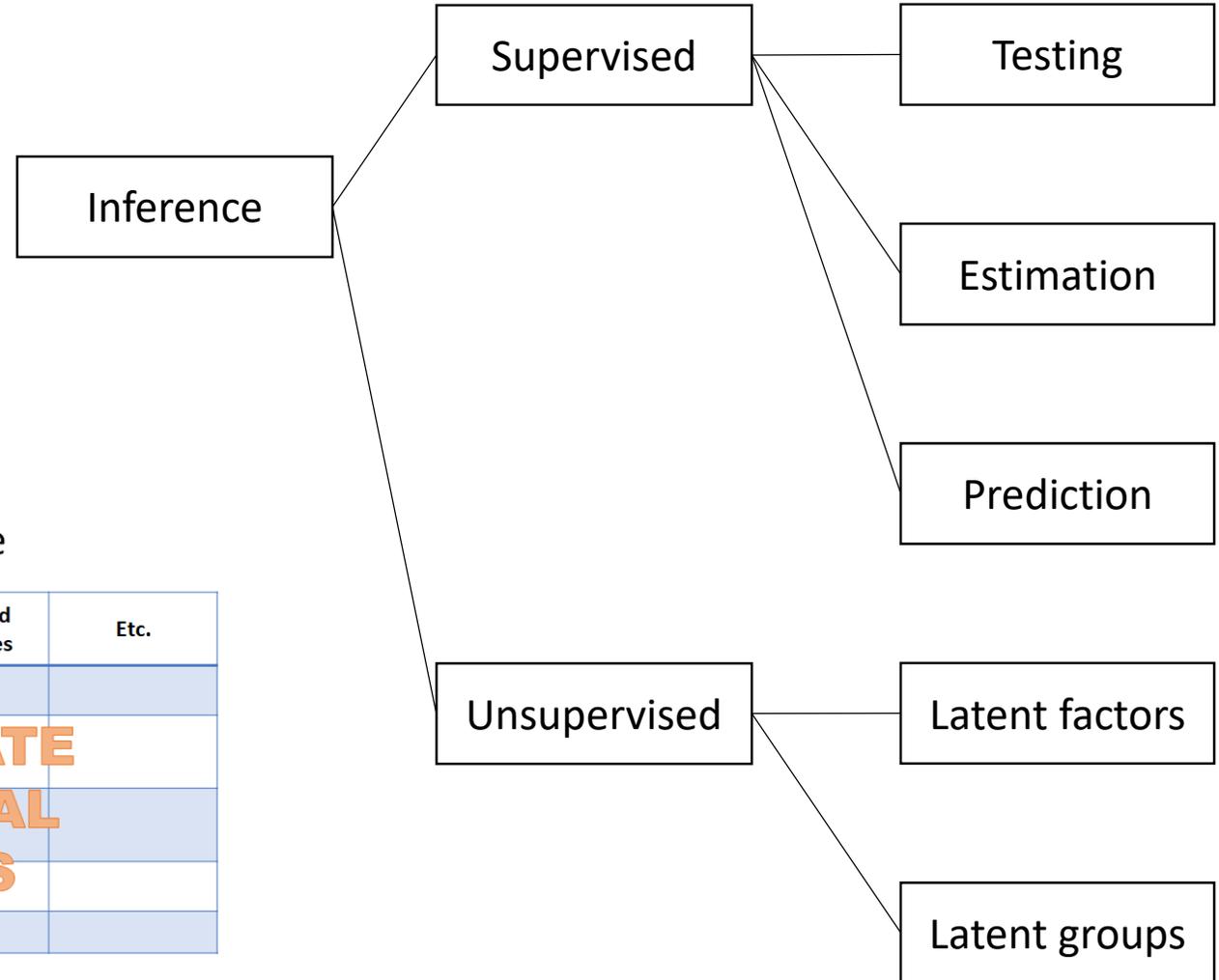
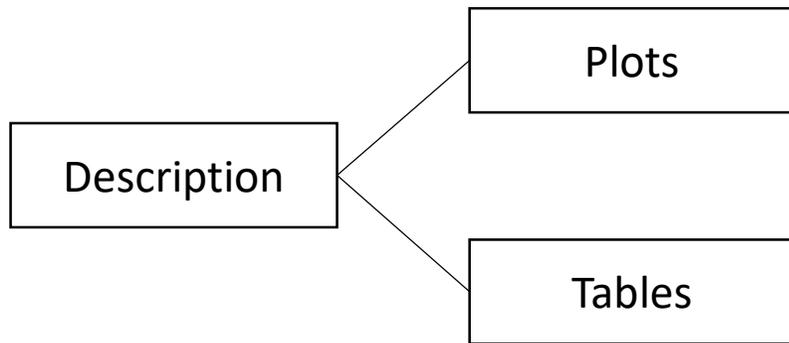


# Lasso using glmnet

```
library(glmnet)
counts = t(GetAssayData(s_obj, slot = "counts"))[-(1:2),]
pcs = Embeddings(s_obj, reduction = "pca")[-(1:2), 1:10]
lasso = cv.glmnet(counts, pcs, family = "mgaussian", nfolds =
3)
lambda = min(lasso$lambda[lasso$nzero <= 10])
coefs = coef(lasso, s = lambda)
rownames(coefs$PC_1)[which(coefs$PC_1 != 0)]
new_counts = t(GetAssayData(s_obj, slot = "counts"))[1:2,]
new_pcs = predict(lasso, newx = new_counts, s = lambda)[,, 1]
predict(rf_fit, new_pcs)
Idents(s_obj)[1:2]
```

Conclusions

# Statistical toolbox



Data structure

Statistical task

	No dependent variables	Continuous outcome	Censored outcomes	Etc.
Visualize				
Identify latent factors	<b>APPROPRIATE STATISTICAL METHODS</b>			
Cluster observations				
Select features				
Etc.				

# Statistical tools

1. PCA
2. Graph clustering
3. t-SNE plot
4. Wilcoxon test
5. FDR control
6. Random forest classification
7. Lasso

# Can be applied to single-cell RNA-seq and beyond

CellPress

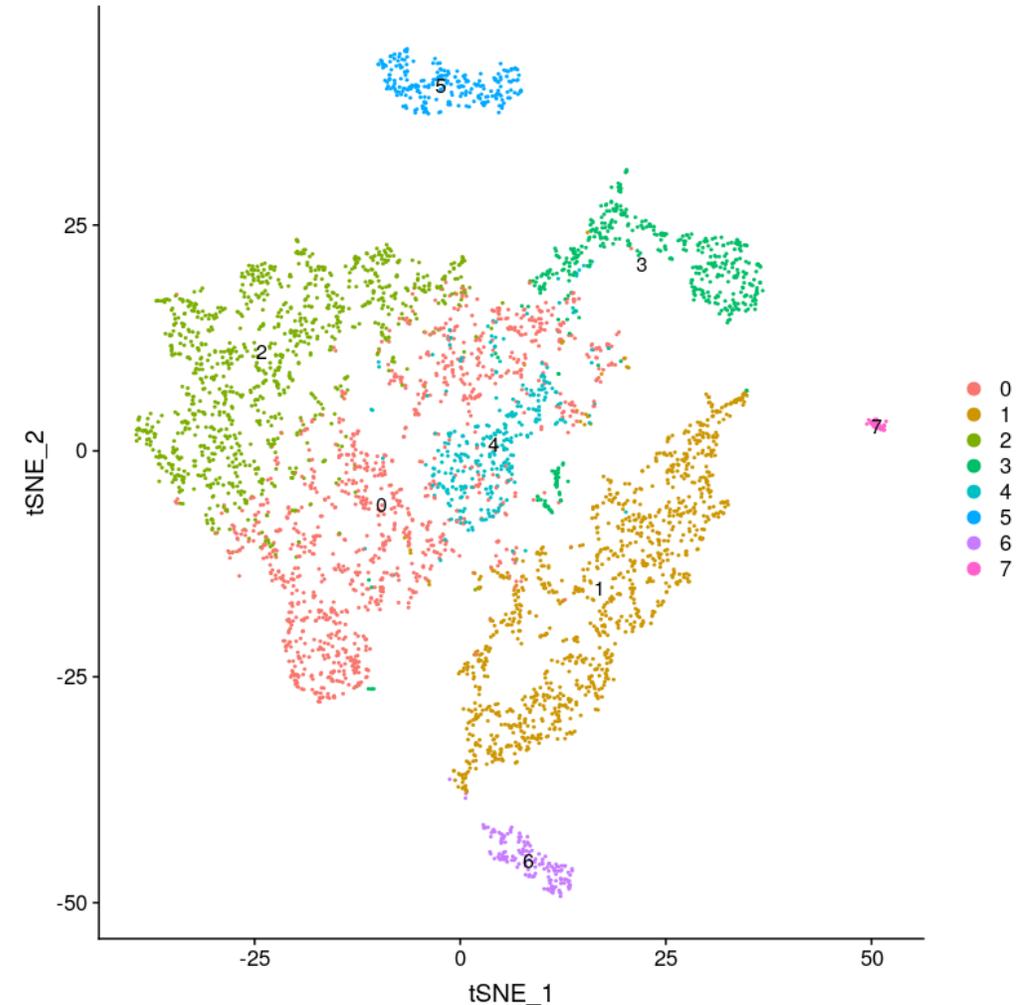
Current Biology  
Article

## Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq

Shristi Pandey,<sup>1,\*</sup> Karthik Shekhar,<sup>2</sup> Aviv Regev,<sup>2,3</sup> and Alexander F. Schier<sup>1,2,4,5,6,7,\*</sup>

<sup>1</sup>Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA  
<sup>2</sup>Klarman Cell Observatory, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA  
<sup>3</sup>Howard Hughes Medical Institute and Koch Institute of Integrative Cancer Research Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA  
<sup>4</sup>Center for Brain Science, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA  
<sup>5</sup>Biozentrum, University of Basel, Basel, Switzerland  
<sup>6</sup>Allen Discovery Center for Cell Lineage Tracing, University of Washington, Seattle, WA 98195, USA  
<sup>7</sup>Lead Contact

\*Correspondence: [p.shristi@gmail.com](mailto:p.shristi@gmail.com) (S.P.), [schier@fas.harvard.edu](mailto:schier@fas.harvard.edu) (A.F.S.)  
<https://doi.org/10.1016/j.cub.2018.02.040>



# To learn more

- Take systematic courses in basic statistics, statistical learning, and R/python
- Study recently published papers in your field of interest that use your technology of interest
- Consult tutorials, workshops, lab mates, and Google

**Thank you**

