SQL

```
SELECT [DISTINCT]
            {* | expr [[AS] c_alias]
            {,expr [[AS] c_alias] ...}}
        FROM tableref {, tableref}
        [[INNER | LEFT ] JOIN table_name ON qualification_list]
        [WHERE search_condition]
        [GROUP BY colname {,colname...}]
        [HAVING search_condition]
        [ORDER BY column_list]
        [LIMIT number]
        [OFFSET number of rows];
```

Syntax	Description
`SELECT column_expression_list`	List is comma-separated. Column expressions may include aggregation functions (`MAX`, `SUM`, `COUNT`, `AVG`, etc). `AS` renames columns. `DISTINCT` selects only unique rows.

`WHERE a IN cons_list`
`WHERE a IS NOT val`
`WHERE a LIKE 'p'`
	• Select rows for which the value in column `a` is among the values in a `cons_list`.
	• Selects rows for which the value in column `a` is not equal to `val` (of any data type).
	• Matches each entry in the column `a` to the text pattern `p`. The wildcard `%` matches at least zero characters. _ matches exactly one character.

`ORDER BY RANDOM() LIMIT n`	Draw a simple random sample of `n` rows.
`ORDER BY a, b DESC`	Order by column `a` (ascending by default) , then `b` (descending).
`CASE WHEN pred THEN cons ELSE alt END`	Evaluates to `cons` if `pred` is true and `alt` otherwise. Multiple `WHEN`/`THEN` pairs can be included, and `ELSE` is optional.
`LIMIT number`	Keep only the first `number` rows in the return result.
`OFFSET number`	Skip the first `number` rows in the return result.

Principal Component Analysis (PCA)

The i-th Principal Component of the matrix X is defined as the i-th column of $U\Sigma$ defined by Singular Value Decomposition (SVD).

$X = U\Sigma V^T$ is the SVD of X if U and V^T are matrices with orthonormal columns and Σ is a diagonal matrix. The diagonal entries of Σ, $[s_1, \ldots, s_r, 0, \ldots, 0]$, are known as singular values of X, where $s_i > s_j$ for $i < j$ and $r = \text{rank}(X)$.

Define the design matrix $X \in \mathbb{R}^{n \times p}$. Define the total variance of X as the sum of individual variances of the p features. The amount of variance captured by the i-th principal component is equivalent to s_i^2/n, where n is the number of datapoints.

Syntax	Description
`np.linalg.svd(X, full_matrices = True)`	SVD of `X` with shape `(M, N)` that returns `u`, `s`, `vt`, where `s` is a 1D array of `X`'s singular values. If `full_matrices=True`, `u` and `vt` have shapes `(M, M)` and `(N, N)` respectively; otherwise shapes are `(M, K)` and `(K, N)`, respectively, where `K = min(M, N)`.

Decision Trees

Suppose you have a **decision tree** classifier for k classes. For each node, define the probability for class $C \in \{1, \ldots, k\}$ as $p_C = d_C/d$, where d_C is the number of datapoints in class C (of the total d in the node). The entropy of the node (in bits) is defined as $S = -\sum_C p_C \log_2 p_C$, and the weighted entropy of the split is the average entropy of child nodes weighted by the number of datapoints in each.
Decision tree generation algorithm: all of the data starts in the root node. Repeat until every node is either pure or unsplittable.

- Pick the best feature x and best split value β to maximize the change in weighted entropy.
- Split data into two nodes, one where $x < \beta$, and one where $x \geq \beta$

A node that only has samples from one class is called a "pure" node. A node that has overlapping datapoints from different classes and thus cannot be split is called "unsplittable."

A **random forest** is a collection of many decision trees fit to variations of the same training data (e.g. bootstrapped samples, also called bagging). It is an ensemble method.