The RMSE Methodology (Root Mean Square Error)

Real World

$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$

Mathematical Model

$\vec{o}_1, \vec{o}_2, \ldots, \vec{o}_N$

$$RMS\ Error = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\|\vec{x}_i - \vec{o}_i\right\|_2^2}$$

The best constant prediction is the mean

- Data points: $\vec{x}_1, \ldots, \vec{x}_N \in R^d$
- What is the vector $\vec{a} \in R^d$ that minimizes
- RMSE= $\sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \|\vec{x}_i - \vec{a}\|_2^2}$

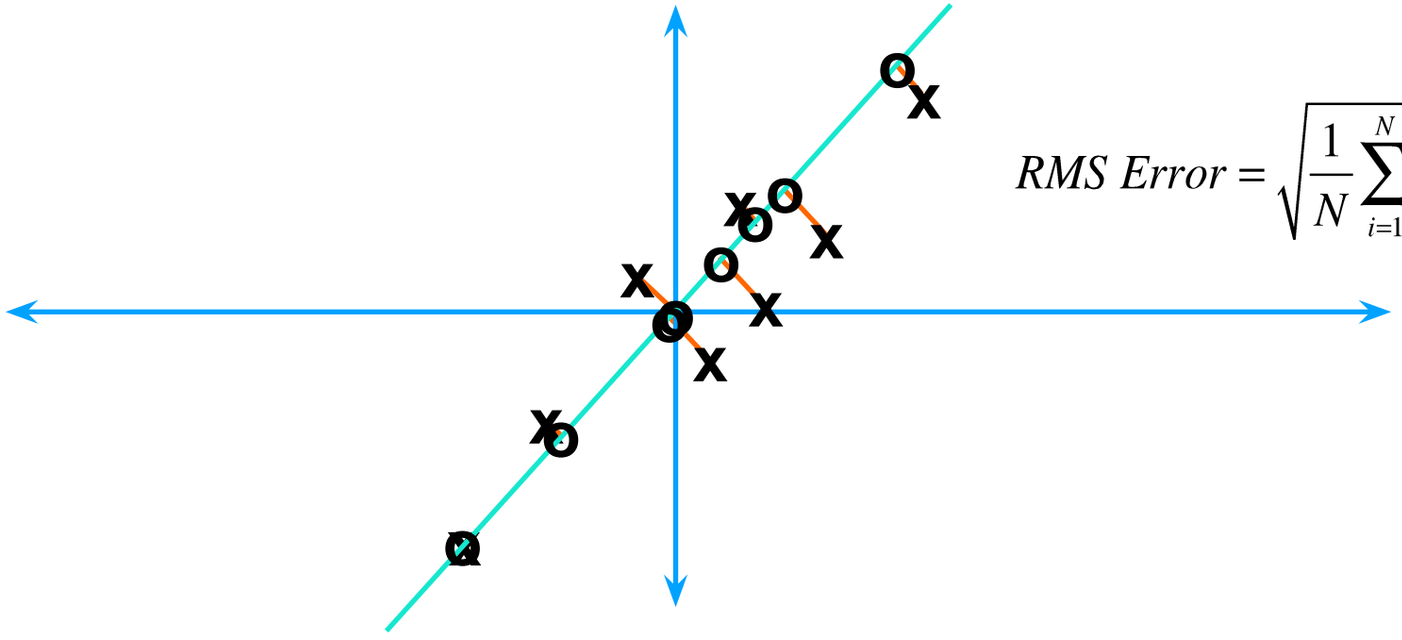
- It is the mean: $\vec{a} = \vec{\mu} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \vec{x}_i$

PCA based prediction

Data: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in R^d$

Mean vector: $\vec{\mu}$ Top k eigenvectors: $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$

Approximation of $\vec{x}_j$: $\vec{o}_j = \vec{\mu} + \sum_{i=1}^{k} \left(\vec{v}_i \cdot \vec{x}_j \right) \vec{v}_i$



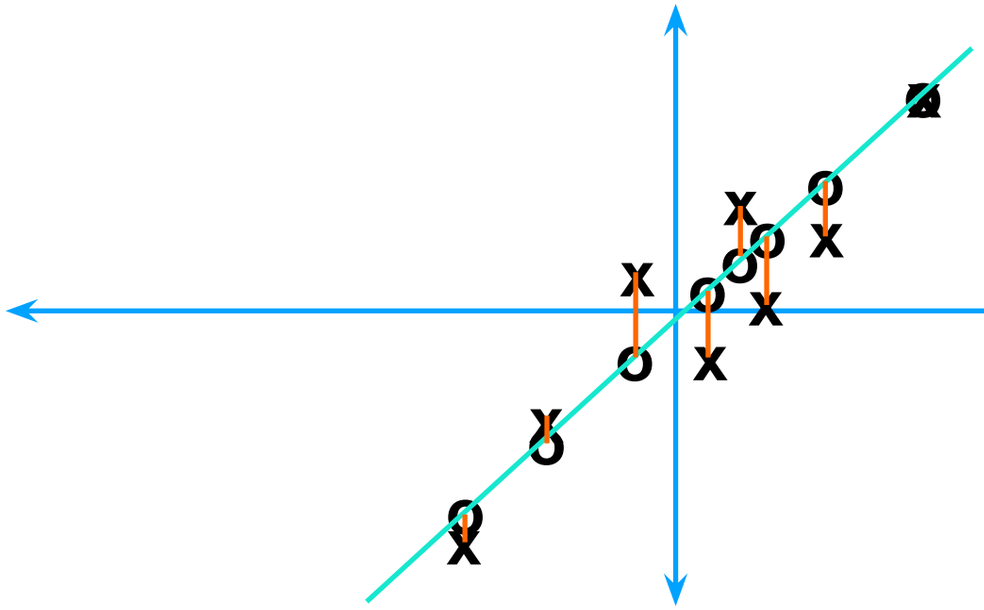
$$RMS\ Error = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left\| \vec{x}_i - \vec{o}_i \right\|_2^2}$$

Regression based Prediction

Data: $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ \ldots \ , (\vec{x}_N, y_N) \in R^d$

Input: $\vec{x} \in R^d$ Output: $y \in R$

Approximation of y given $\vec{x}$: $\hat{y} = a_0 + \sum_{i=1}^{d} a_i x_i$



$$RMS\ Error = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}$$

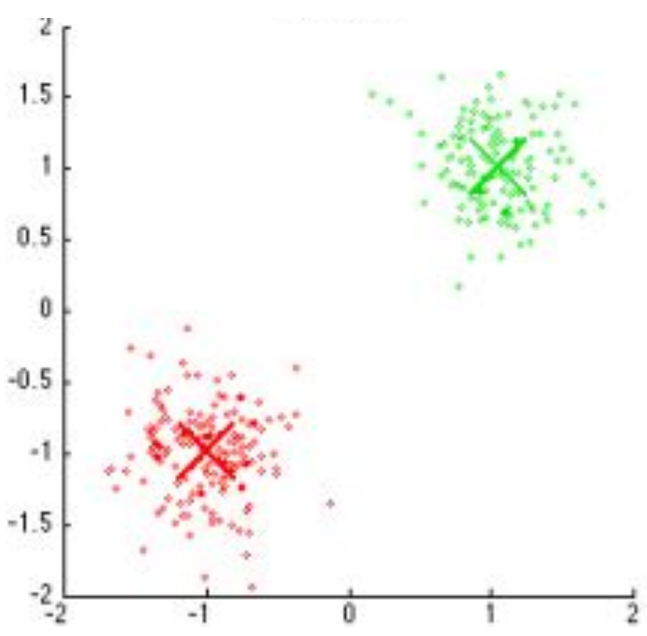
K-means clustering

Data: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in R^d$

Model: k representatives: $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_k \in R^d$

Approximation of $\vec{x}_j$: $\vec{o}_j = \text{argmin}_{\vec{r}_i} \left\| \vec{x}_j - \vec{r}_i \right\|_2^2$

= the representative closest to $\vec{x}_j$



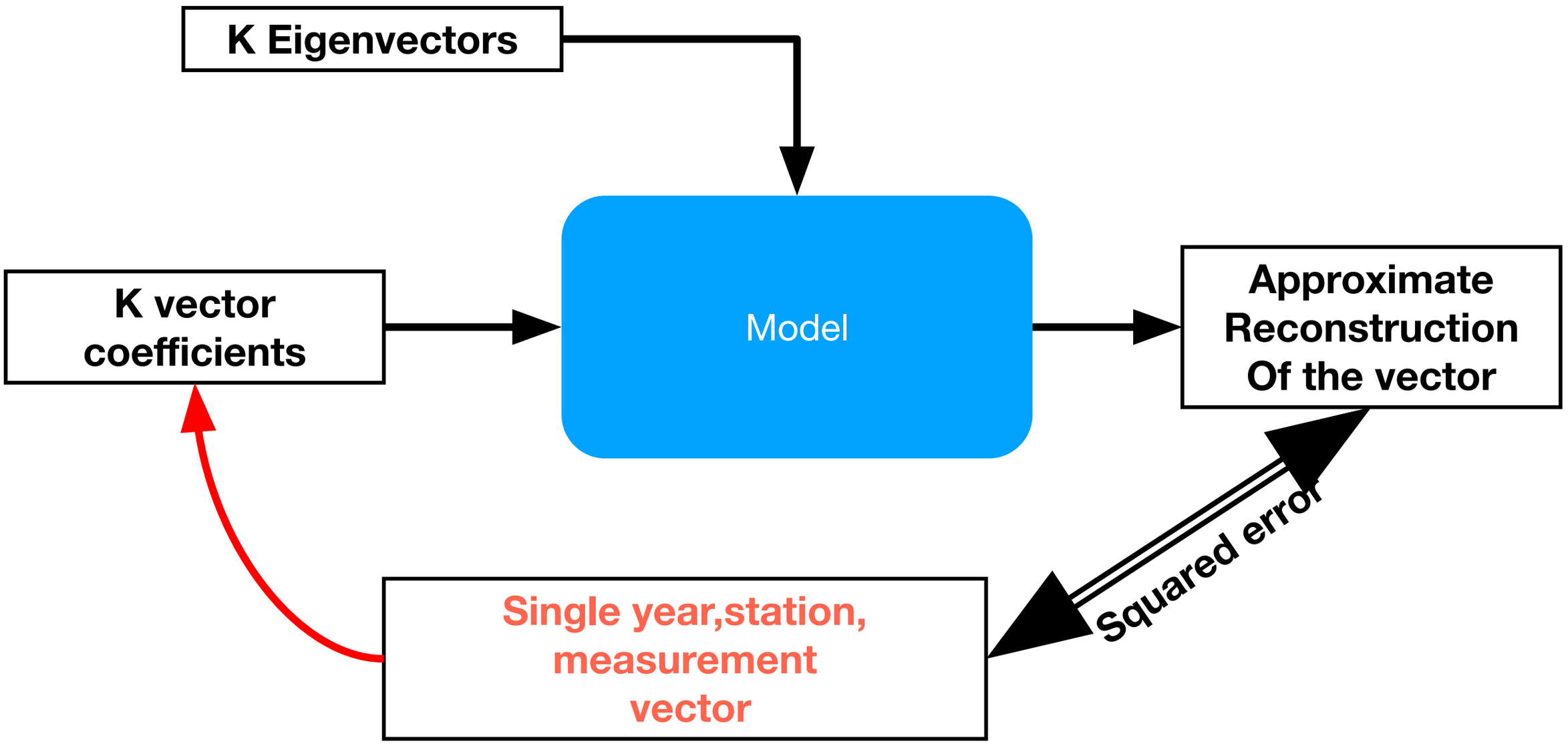
$$RMS\ Error = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left\| \vec{x}_i - \vec{o}_i \right\|_2^2}$$

Percentage of variance explained

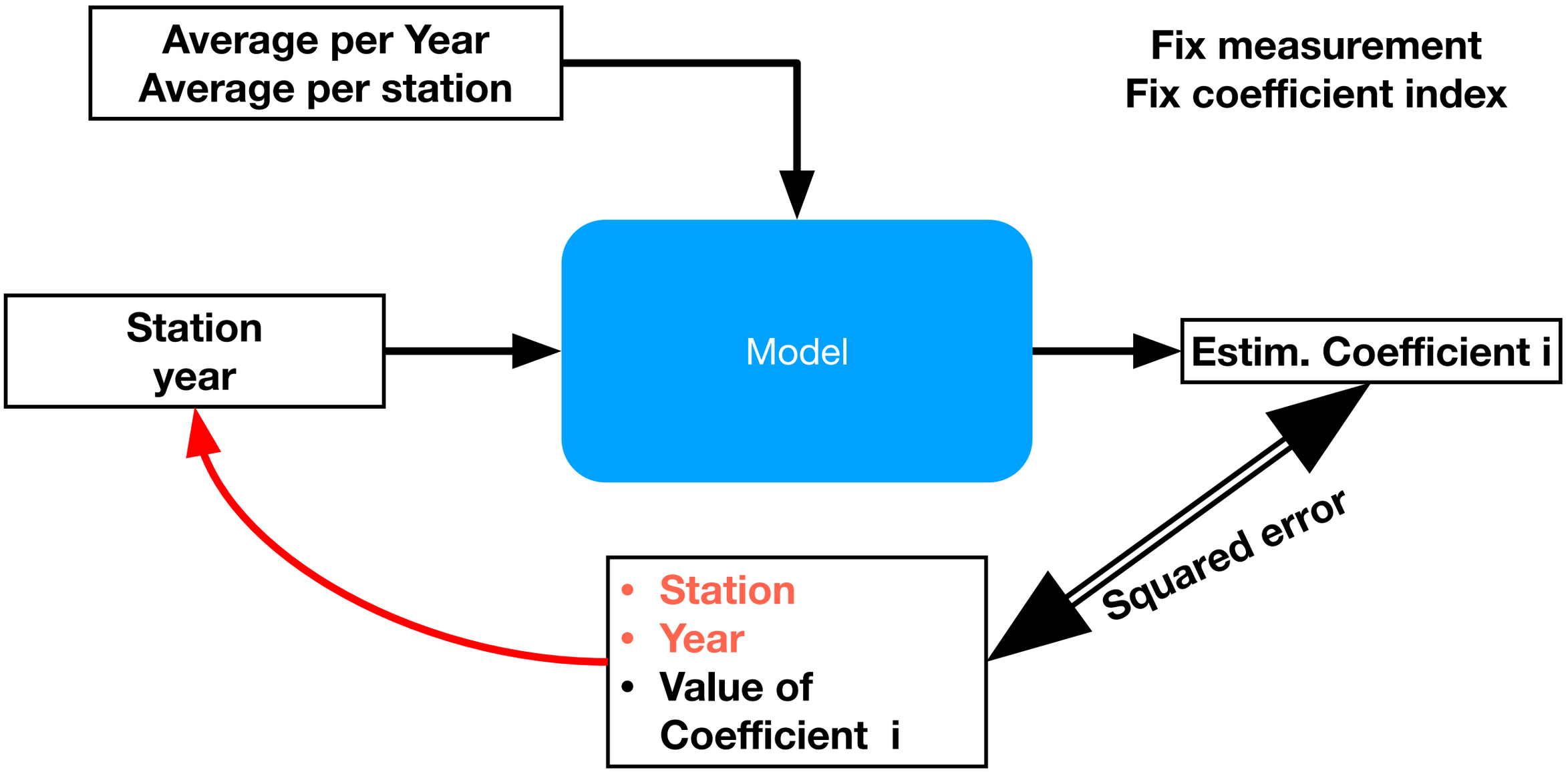
- RMSE changes with the units of measure (cm, mm, km)
- To measure improvements in a unit-free way we divide by a base-line model.

- Examples:
  - for PCA model we use $\frac{RMSE(k)}{RMSE(0{=}mean\ only)}$
  - For K-means we use $\frac{RMSE(k)}{RMSE(1{=}mean\ only)}$

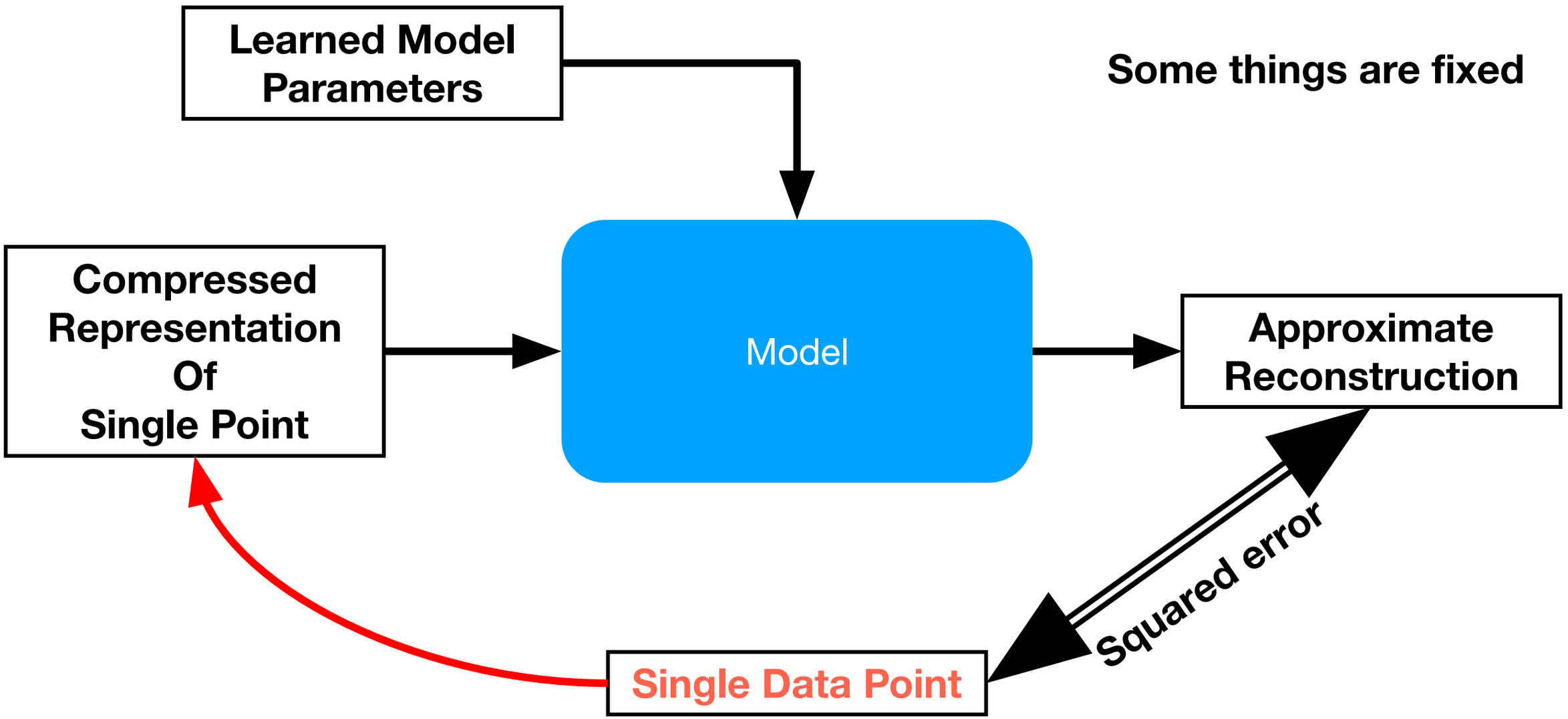
PCA block diagram



Spatial/temporal block diagram

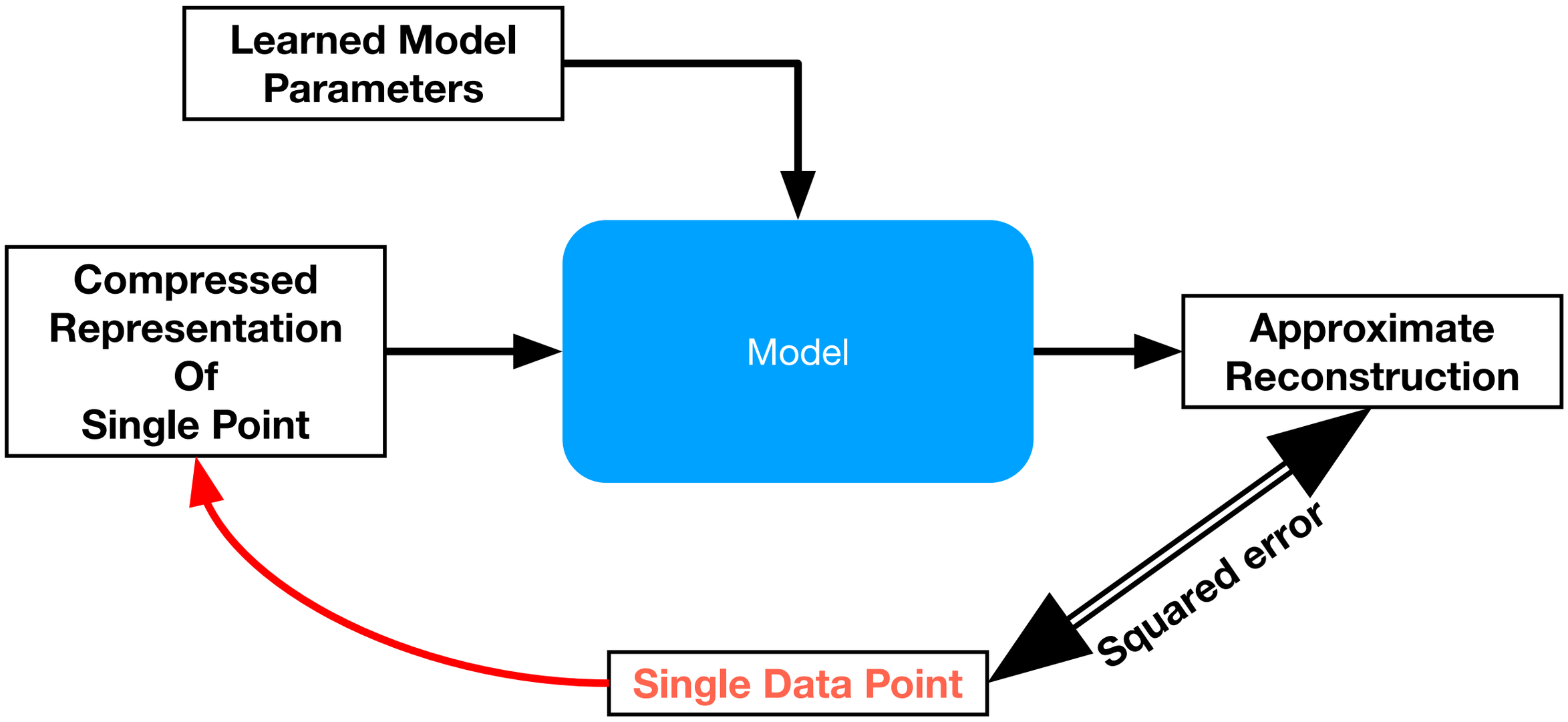


Model-based approximation block diagram



Extensive and Intensive properties

- Extensive: size scales with size of data
- Intensive: size does not scale with size of data

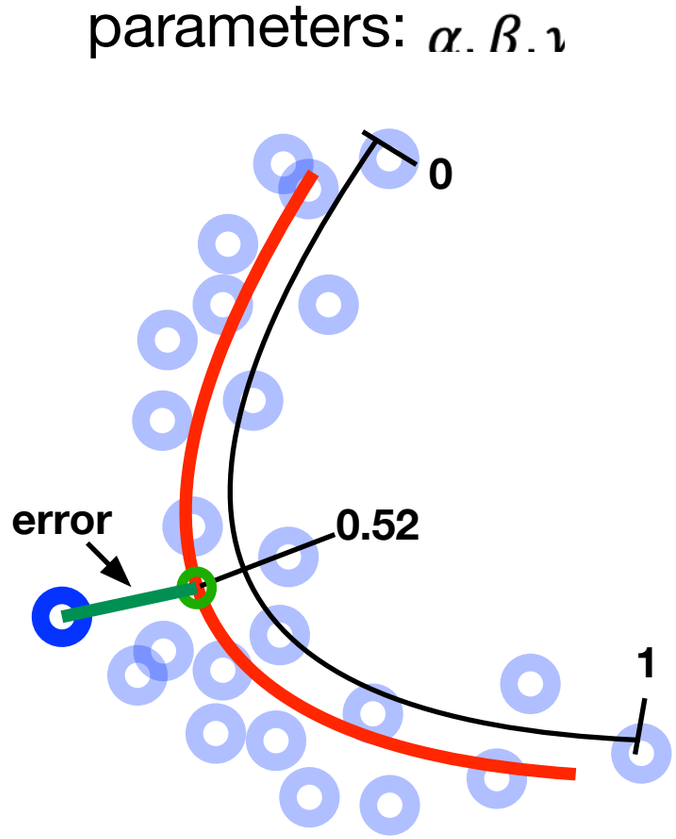


What is a good model?

- $\vec{x}_t \rightarrow \vec{r}_t$ is the dimensionality reducing mapping.
- $\vec{x}_t, \vec{r}_t, \vec{o}_t$ are all extensive
- A good model is one which reduces the dimension with only a small increase in the RMSE.
- If data is large then a larger model is justified.

Model-based approximation

- Given d-dimensional data points. (here d=2)
- Fit a simple model with few parameters. (Here Curve)
- Each data point is approximated by the closest point on the curve.
- The point on the curve is represented by a few numbers (dimensionality reduction)
- The distance between the point and the approximation is the error.
- We want the average root-mean-square-error (RMSE) or Percentage of



Summary

- Models are small approximate representation of the data distribution.
- Each data point $\vec{x}_t$ is mapped to a smaller representation $\vec{r}_t$
- The model maps the representation $\vec{r}_t$ to a reconstruction $\vec{o}_t$
- The RMSE is: $\sqrt{\frac{1}{N}\sum_{t=1}^{N}\|\vec{x}_t - \vec{o}_t\|_2^2}$
- The model and model parameter are intensive
- $\vec{x}_t, \vec{r}_t, \vec{o}_t$ are all extensive
- $\vec{x}_t \rightarrow \vec{r}_t$ is the dimensionality reducing mapping.

Variations on a theme

- different model geometric: lines, curves, points
- different measures of error: distance from projection, distance from closest point fixing x, non-euclidean distances…

Different approaches to statistical modelling

- **Different Models for Different Goals**
  - **RMSE approach**: Model maps each example to a simplified approximation.
    - Examples PCA, Regression, K-means
    - Performance measured using average reconstruction error (RMSE)
    - **Other Goal**: capture parameters of true distribution.
  - **Generative approach**: model is a generator of examples.
    - Examples: Mixture of Gaussians, HMMs
    - Generated distribution approximates true distribution.
    - Performance measured using likelihood.
    - **Main Goal**: capture parameters of true distribution.
  - **Discriminative approach**: model predicts output given input.
    - Examples: Perceptron, Decision trees, Neural Networks
    - Generated predictor approximated input -> output relationship.
    - Performance measured using average loss (RMSE, # of mistakes, …)
    - **No Other Goal**

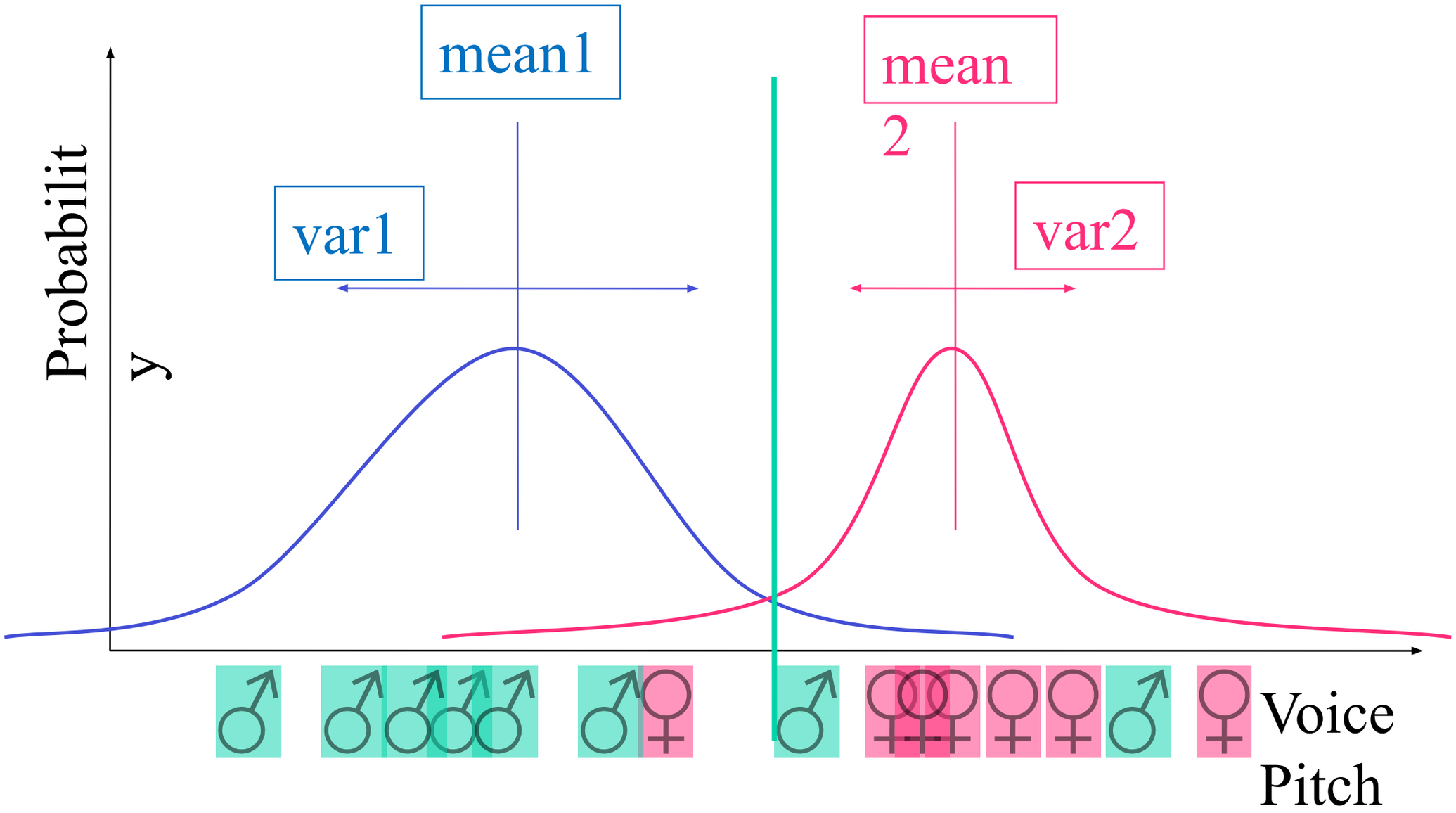
Example of generative vs predictive modeling

- Computer receives telephone call
- Measures Pitch of voice
- Decides gender of caller

Human Voice

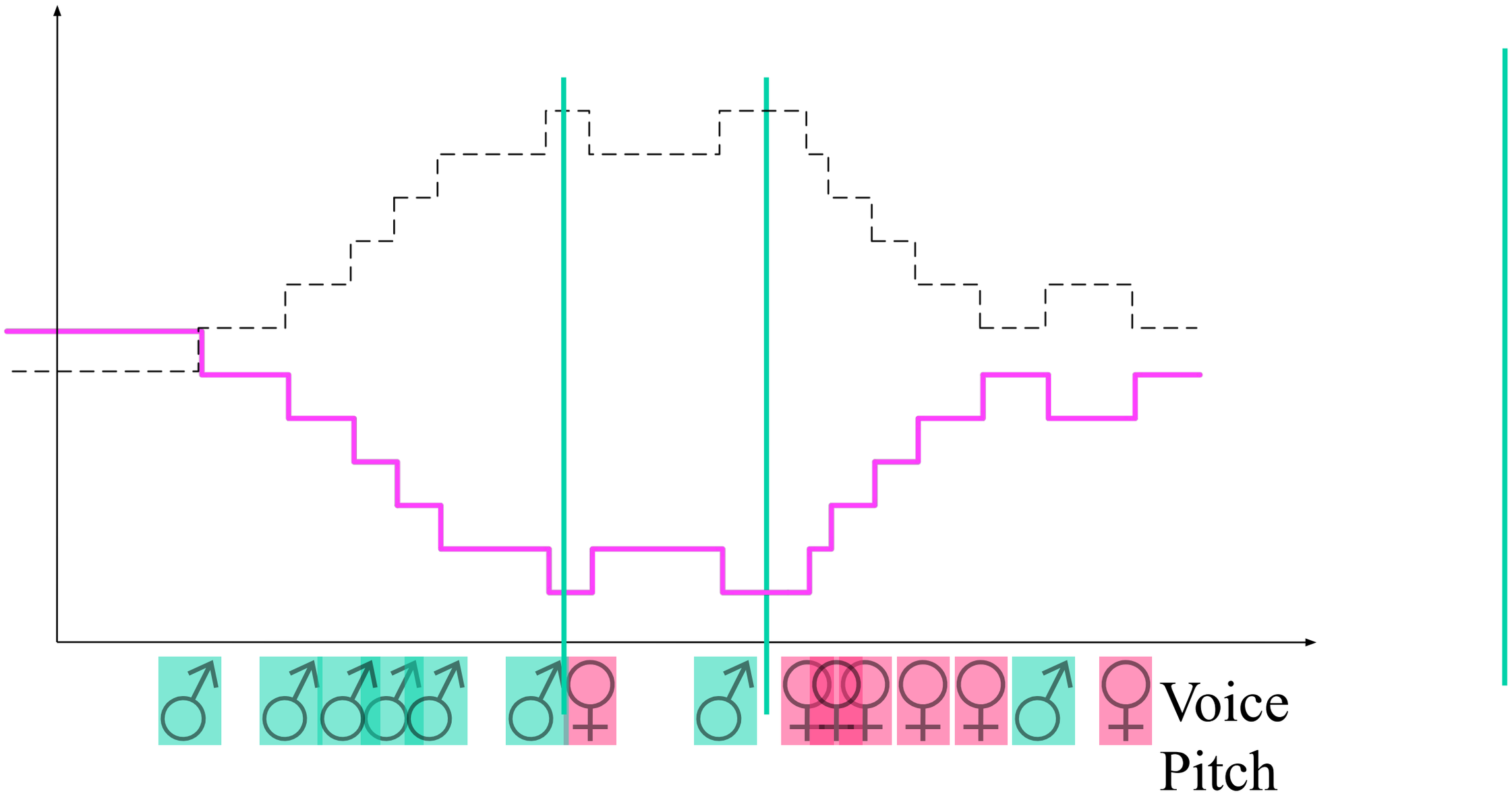
Mal
e

Generative modeling

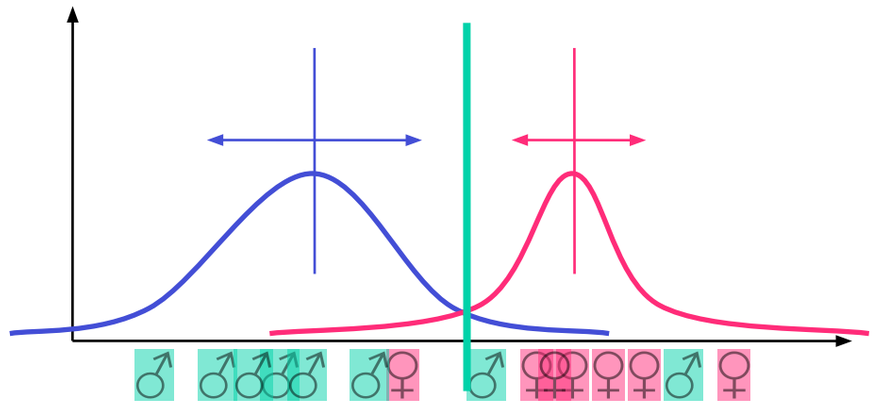


Discriminative approach

[Vapnik 85]



Well Behaved Data



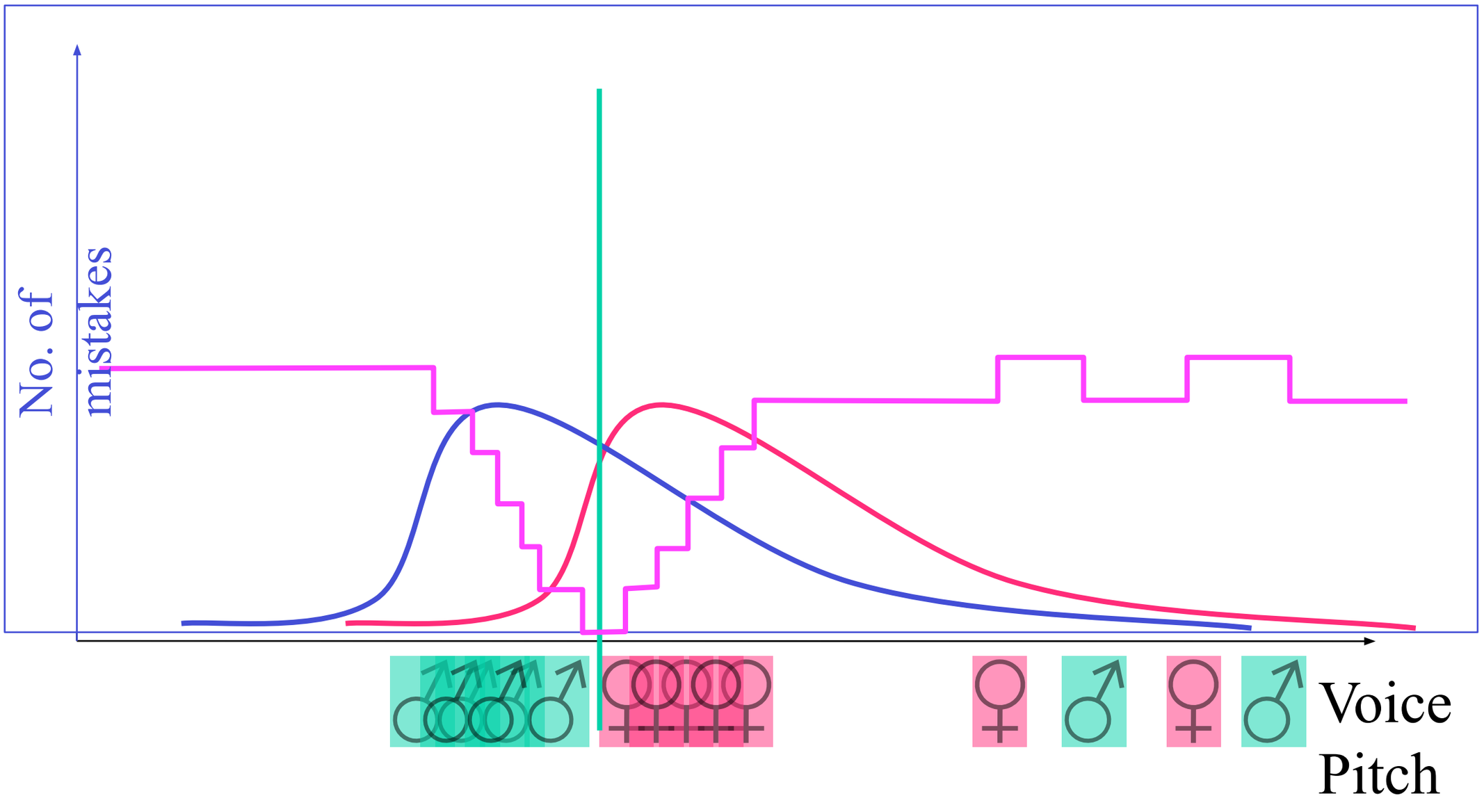
- Data for male/female generated according to normal distributions.

- Estimated Mean and variance converge quickly to true values.

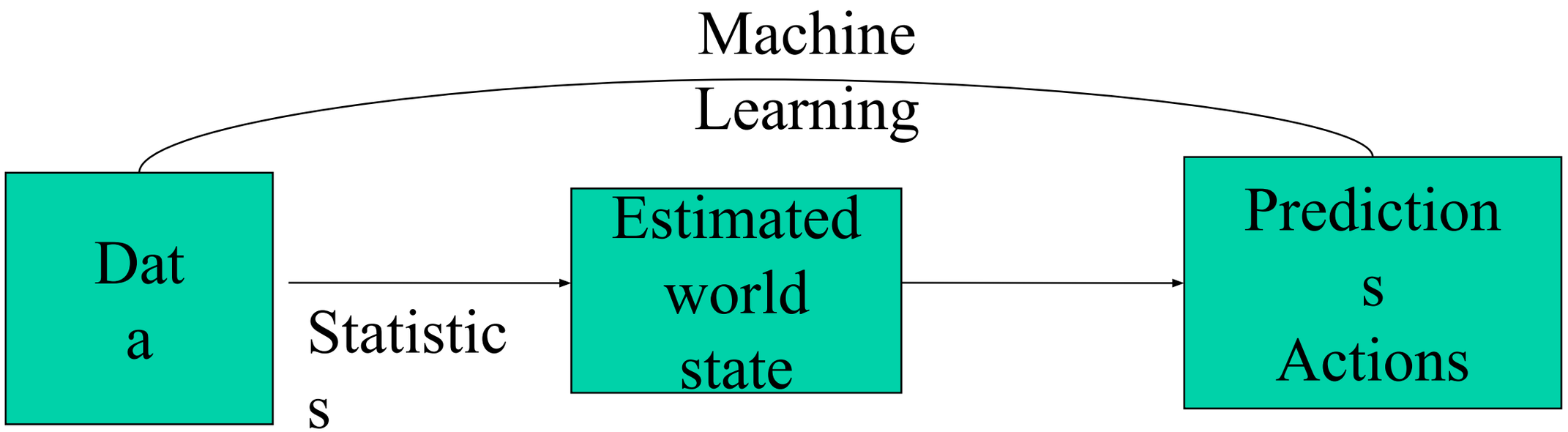
- Generative method converges (much) faster than discriminative method.

- Only viable method when number of training examples is small (<10,000).

Ill-behaved data



Traditional Statistics vs. Machine Learning



Deep Learning

- In the discriminative family.
- Claim: because we have many labeled training examples, we can use as many parameters as we want and never overfit.
- Approach: Use models with many parameters, and train until training error 0.
- Famous success cases: Alpha-Go, machine translation.
- Problems:
  - Restricted labels: can we make use of unlabeled data (auto-encoders vs PCA)
  - Collecting a representative training set can be

Back To Kmeans