Linear prediction with NA, Imputation versus specific methods

Alexis Ayme

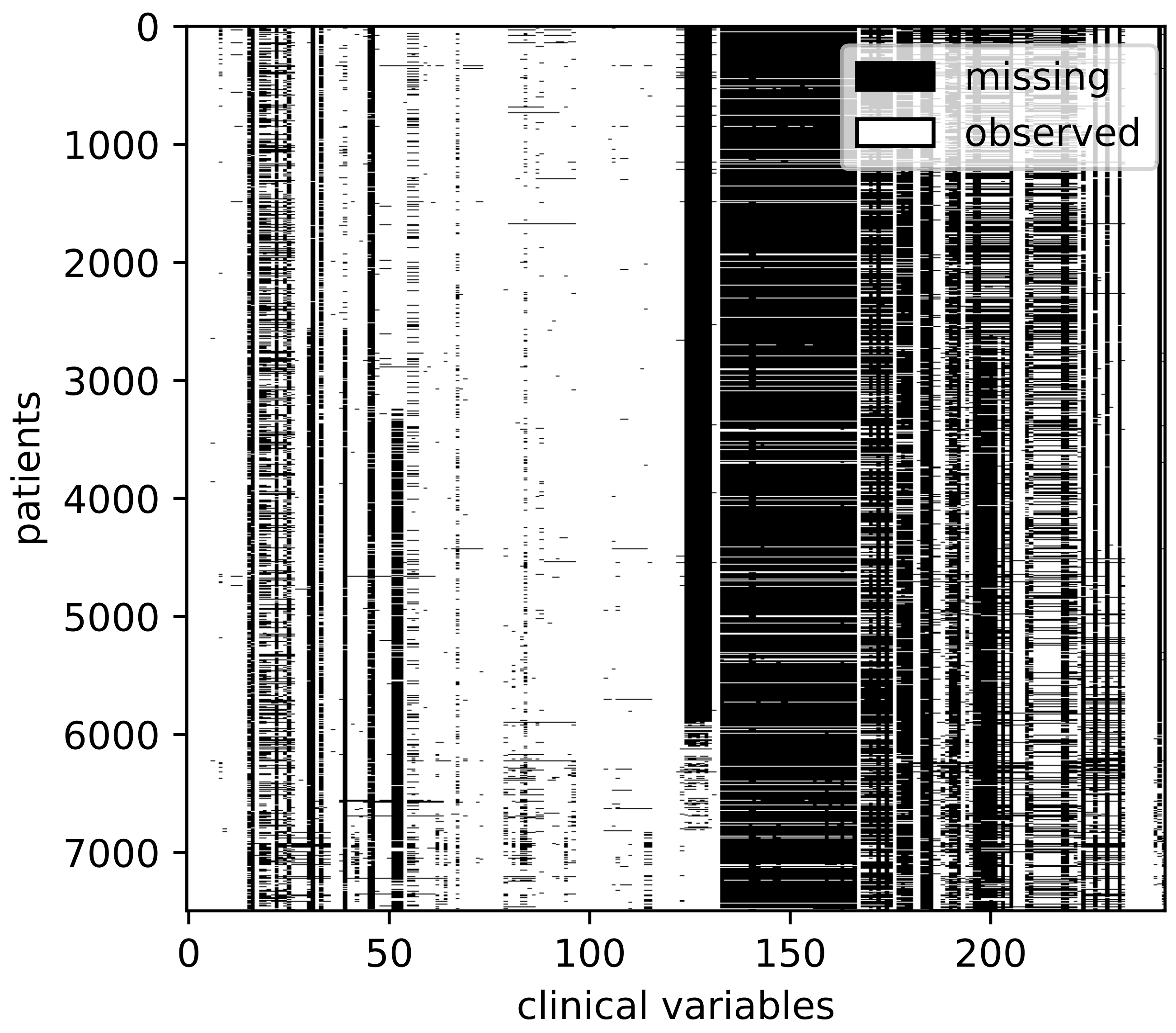
Under the supervision of:
Claire Boyer, Aymeric Dieuleveut and Erwan Scornet

Background

- **Growing mass of data => NA (not attributed)/missing values**

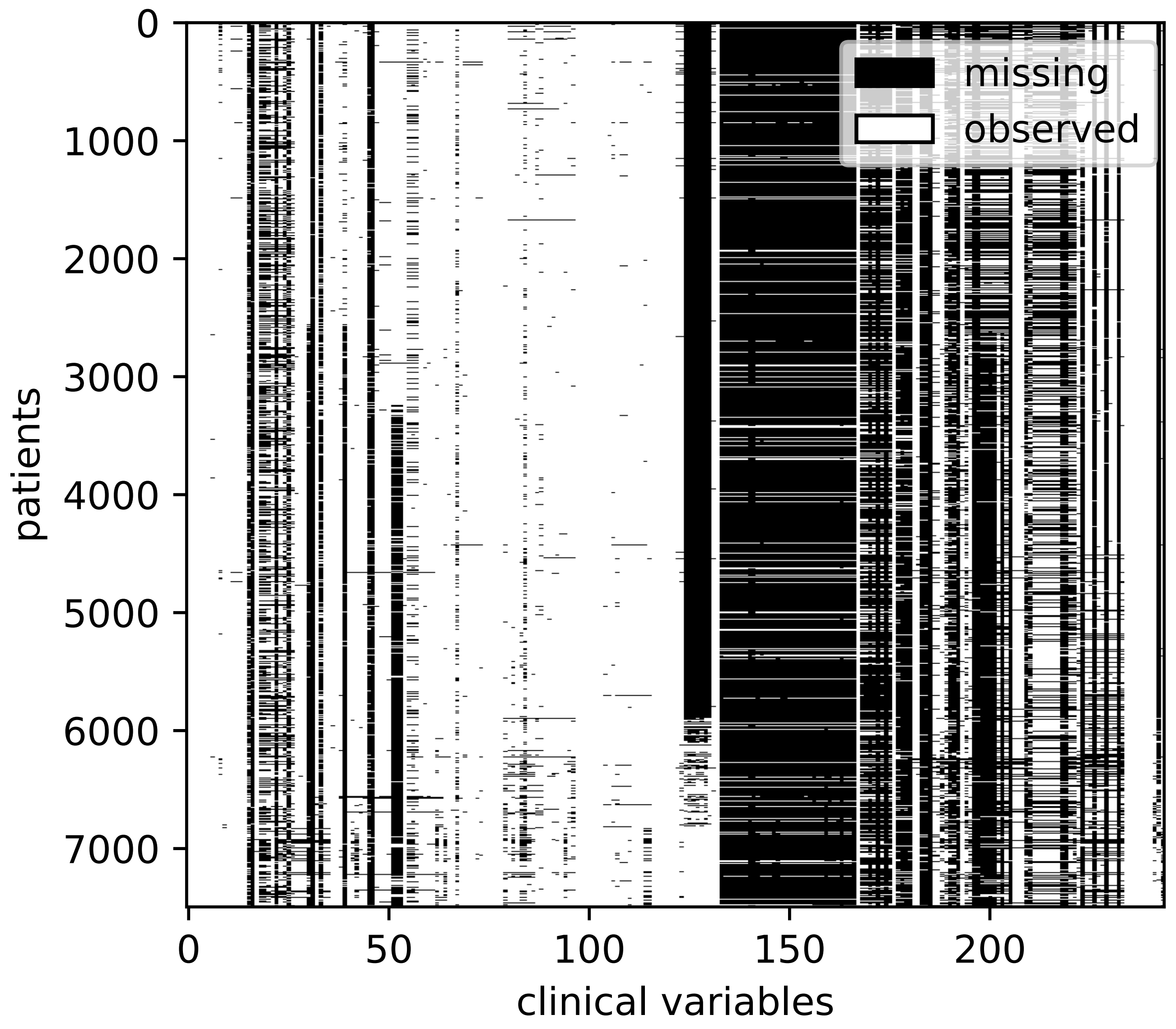
Background

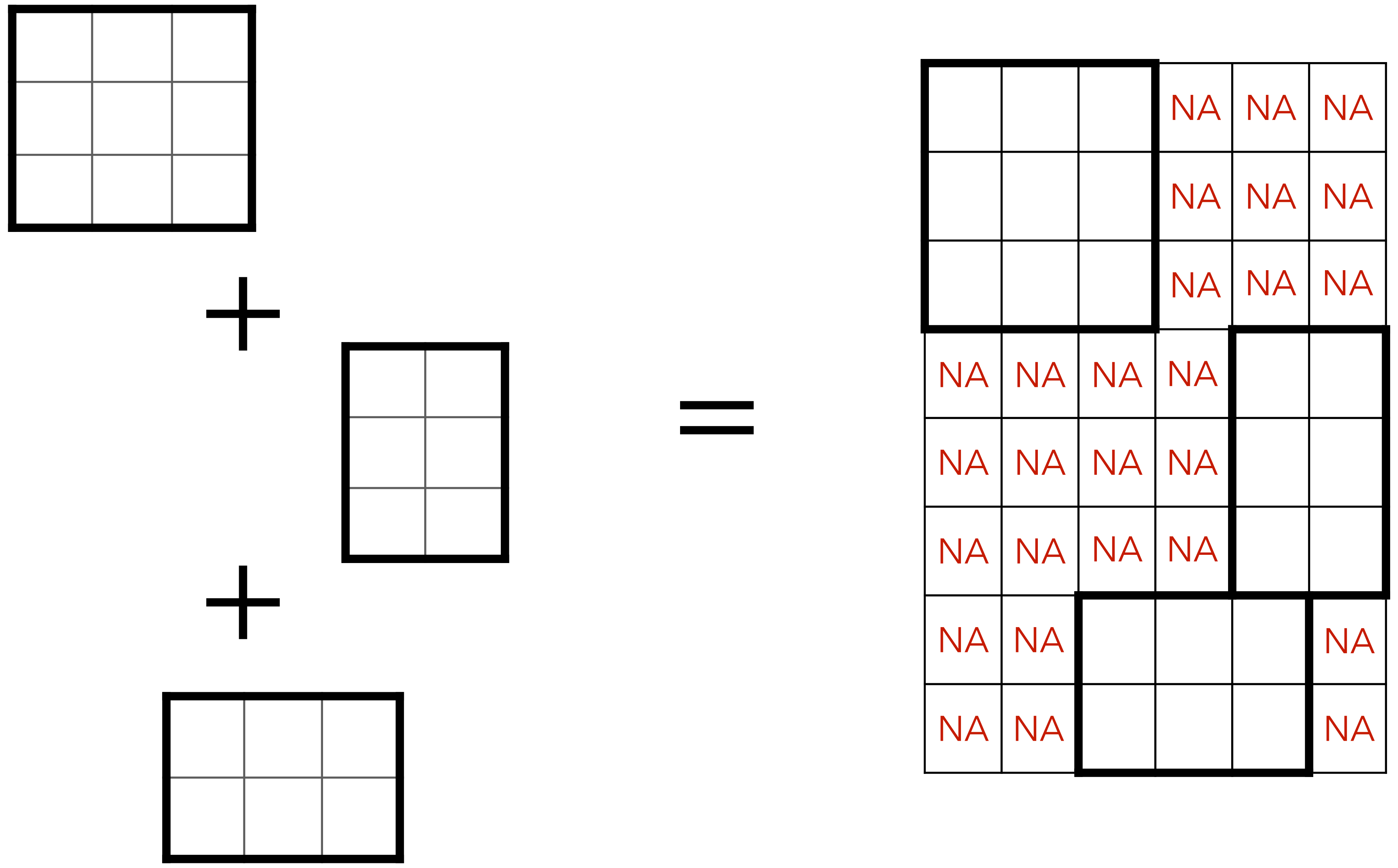
- **Growing mass of data => NA (not attributed)/missing values**
  - **Different sources:**
    1. Bugs
    2. Cost, sensitive data
    3. **Multiplication of sources** (i.e. merging)



Background

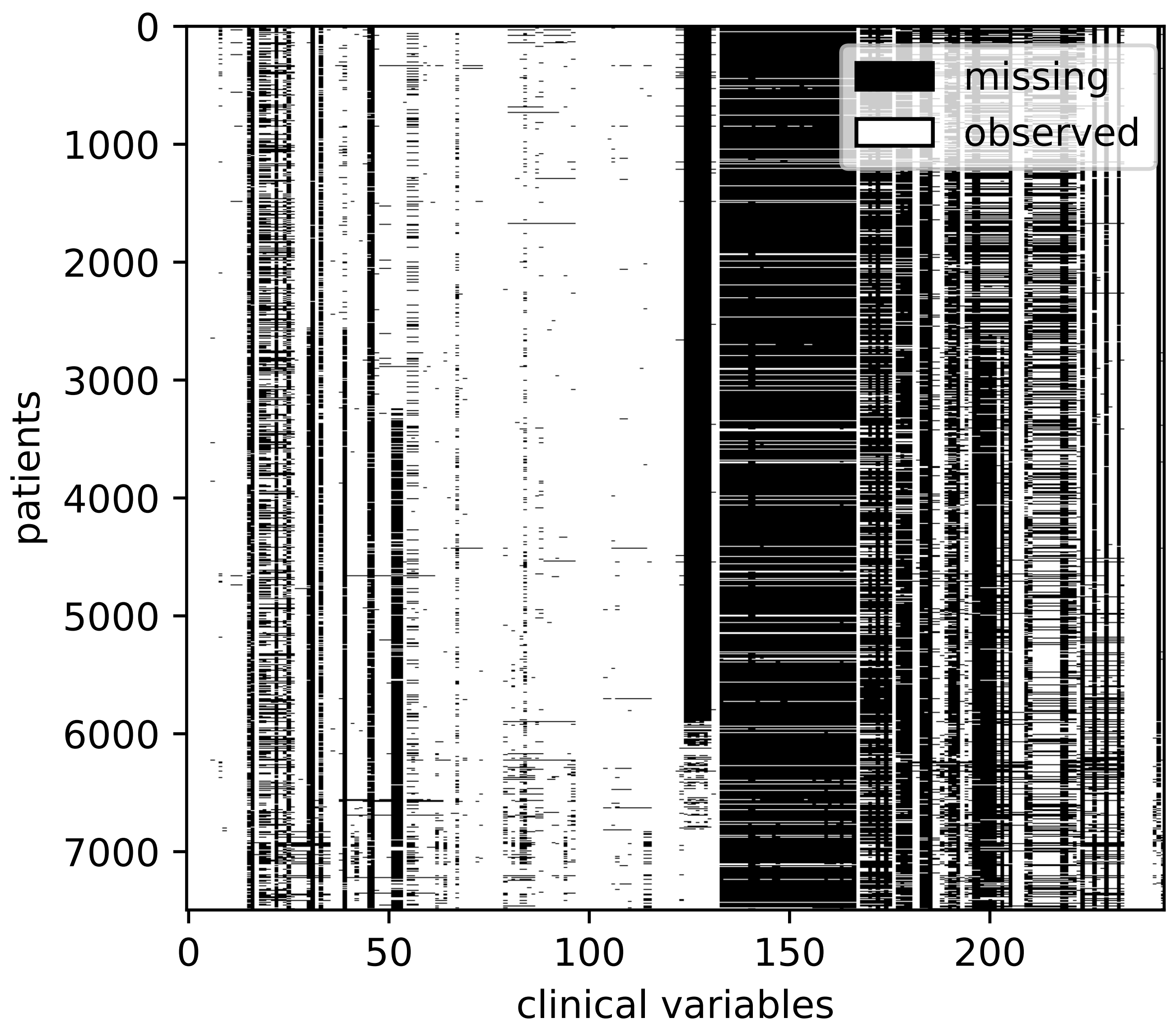
- **Growing mass of data => NA (not attributed)/missing values**
  - **Different sources:**
    1. Bugs
    2. Cost, sensitive data
    3. **Multiplication of sources** (i.e. merging)



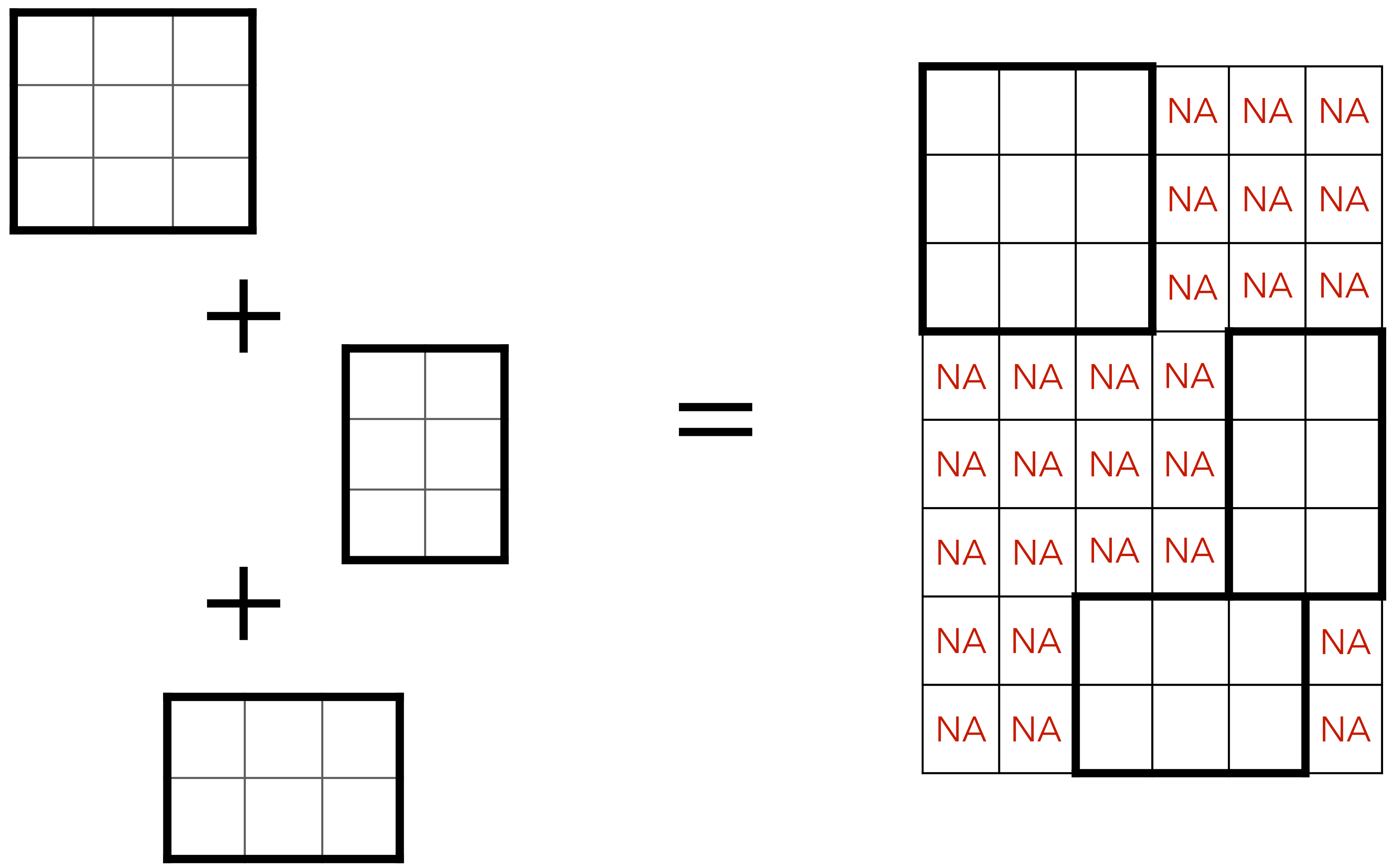


Background

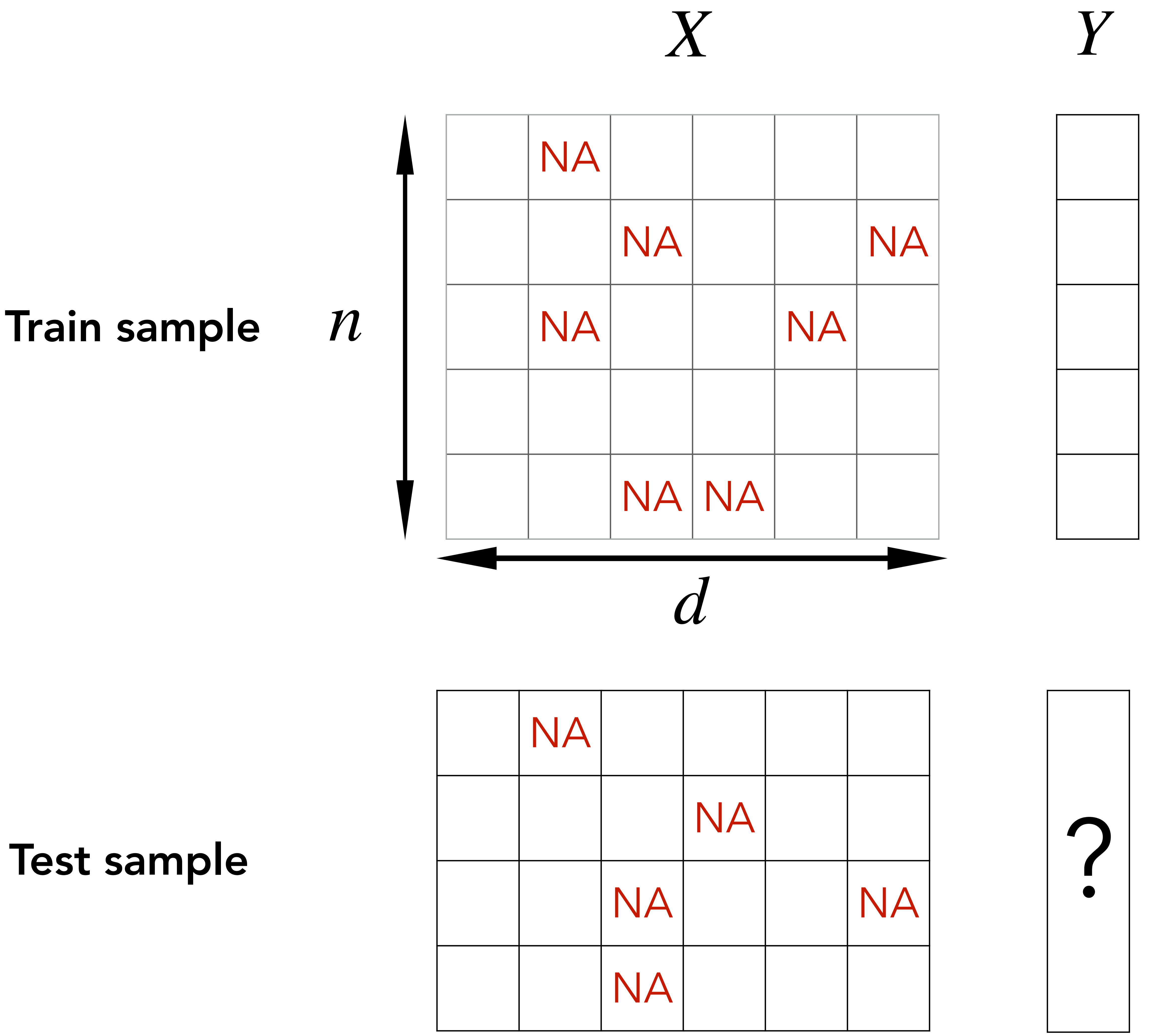
- **Growing mass of data => NA (not attributed)/missing values**
  - **Different sources:**
    1. Bugs
    2. Cost, sensitive data
    3. **Multiplication of sources** (i.e. merging)



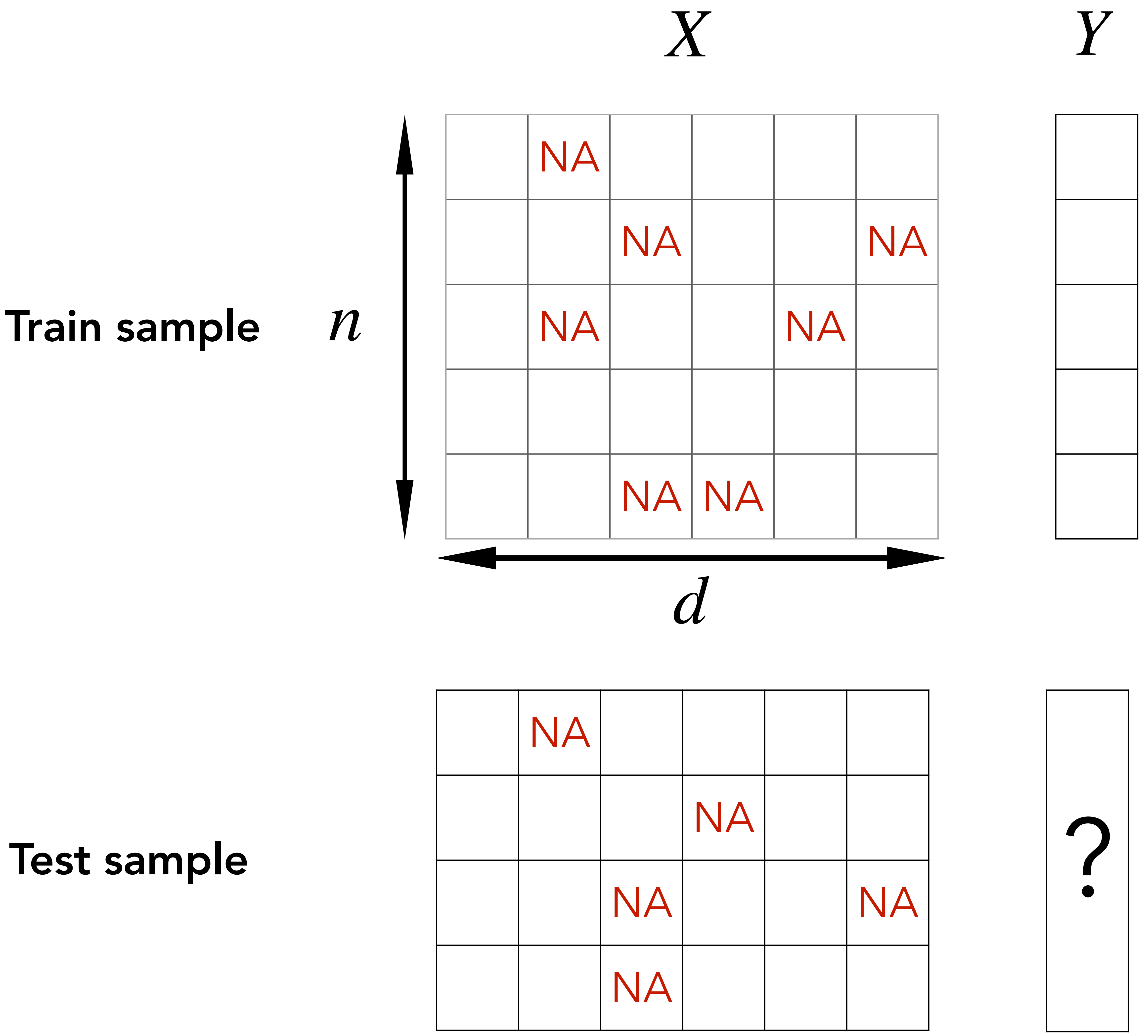
- **Growing mass of data => High-dimensional** dataset
  1. Cost
  2. **Multiplication of sources** (i.e. merging)
  3. Genotype, text



Supervised learning with missing values (NA)



Supervised learning with missing values **(NA)**



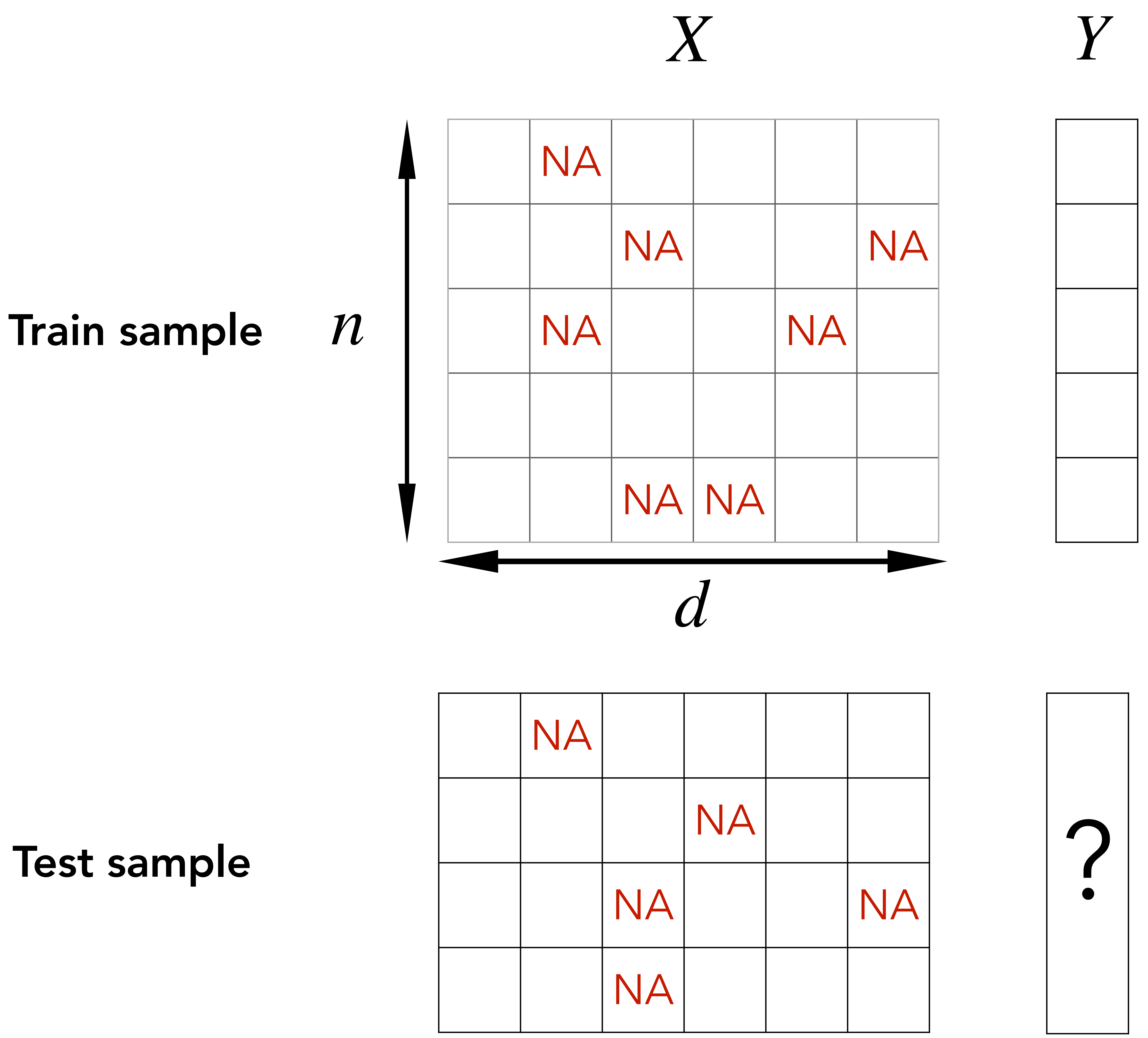
- **Missing pattern:** $M_i \in \{0,1\}^d$

$$X_i = (\,\text{NA}, 8, 0, \text{NA}, 6, 2\,)$$

$$\downarrow$$

$$M_i = (\,1,\ 0,\ 0,\ 1,\ 0,\ 0\,)$$

Supervised learning with missing values (NA)



- **Missing pattern:** $M_i \in \{0,1\}^d$

$$X_i = (\text{NA}, 8, 0, \text{NA}, 6, 2)$$

$$\downarrow$$

$$M_i = (1, 0, 0, 1, 0, 0)$$

- **Input:** $Z = \left(X_{\text{obs}}, M\right)$
- **Output:** $Y \in \mathbb{R}$

Goal: Predict on **test sample** minimizing

$$R_{\text{missing}}(f) = \mathbb{E}_{Z,Y}\left[\left(Y - f(Z)\right)^2\right]$$

Supervised learning vs inference

- **Linear model** for complete inputs

$$Y_i = \beta^\top X_i + \epsilon_i$$

with $\mathbb{E}[\epsilon_i^2] = \sigma^2$ and:

  - if model is well specified: $\mathbb{E}[\epsilon_i | X_i] = 0$
  - else: $\mathbb{E}[\epsilon_i X_i] = 0$

Supervised learning vs inference

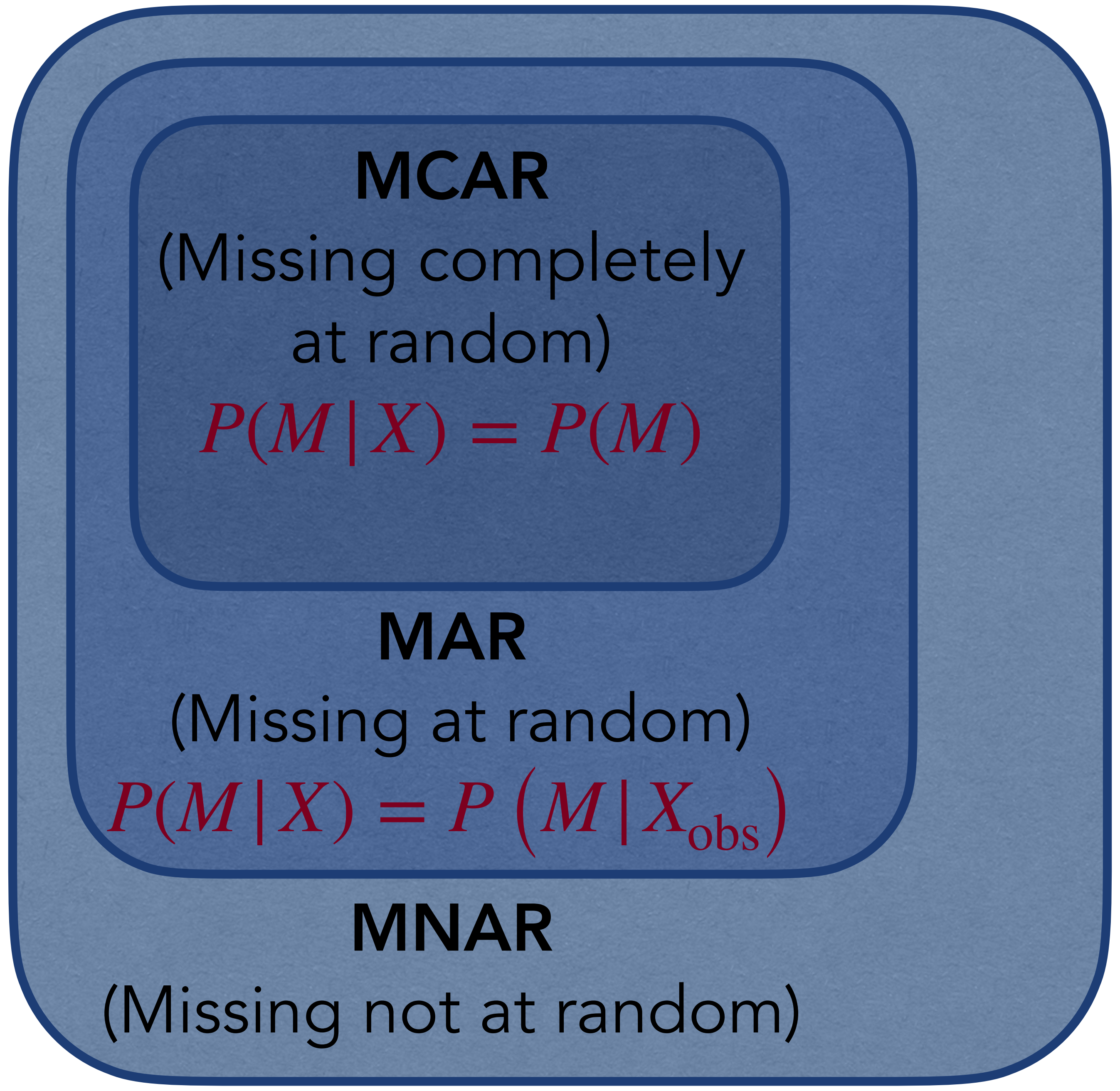
- **Linear model** for complete inputs

$$Y_i = \beta^\top X_i + \epsilon_i$$

with $\mathbb{E}[\epsilon_i^2] = \sigma^2$ and:

  - if model is well specified: $\mathbb{E}[\epsilon_i | X_i] = 0$
  - else: $\mathbb{E}[\epsilon_i X_i] = 0$

- **Missing data mechanism**



Supervised learning vs inference

- **Linear model** for complete inputs

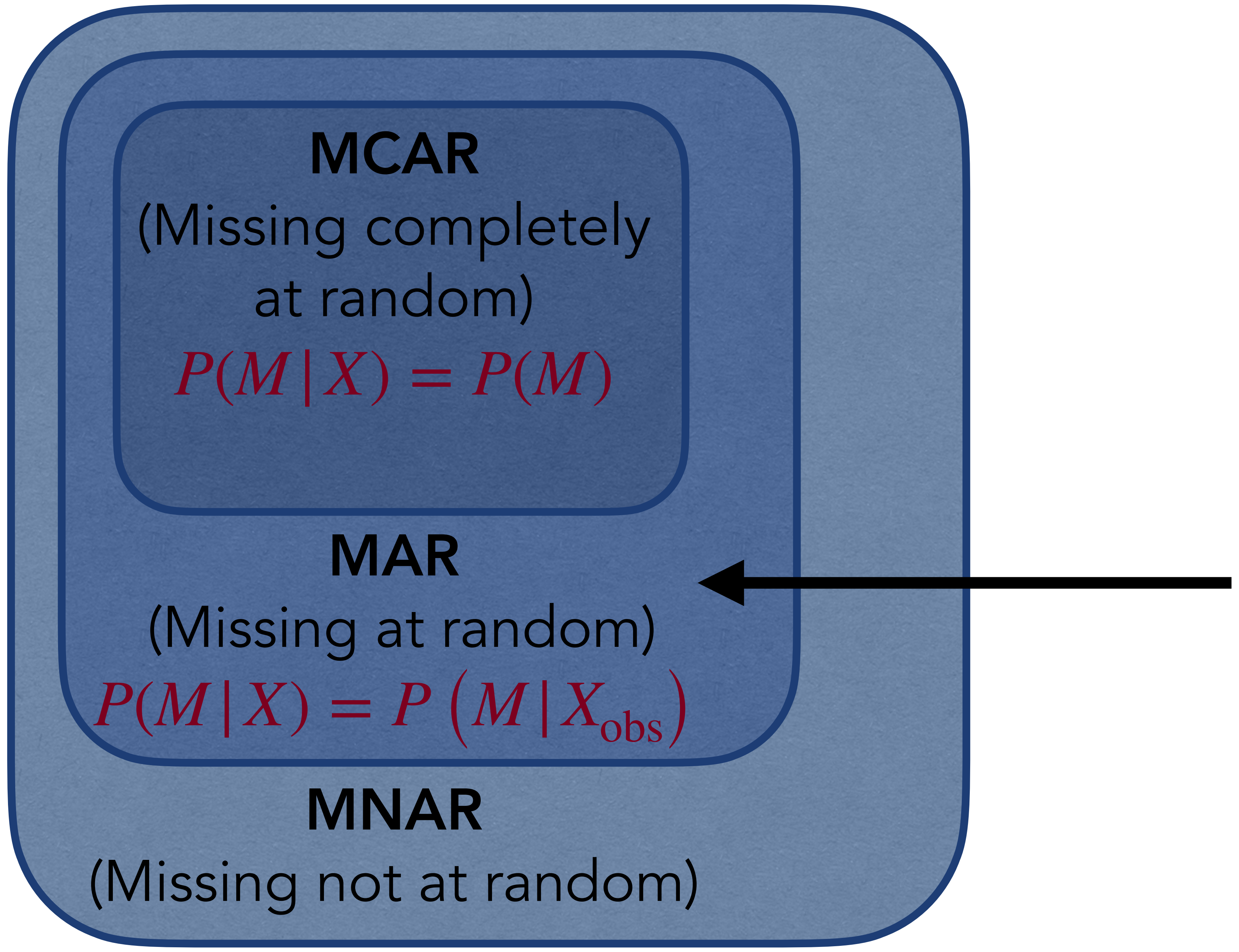
$$Y_i = \beta^\top X_i + \epsilon_i$$

with $\mathbb{E}[\epsilon_i^2] = \sigma^2$ and:

  - if model is well specified: $\mathbb{E}[\epsilon_i | X_i] = 0$
  - else: $\mathbb{E}[\epsilon_i X_i] = 0$

- **Inference**: estimate the model parameter β

- **Missing data mechanism**



Rubin 76, Little 92, Jones 96; Robins et al 94

Supervised learning vs inference

- **Linear model** for complete inputs

$$Y_i = \beta^\top X_i + \epsilon_i$$

with $\mathbb{E}[\epsilon_i^2] = \sigma^2$ and:

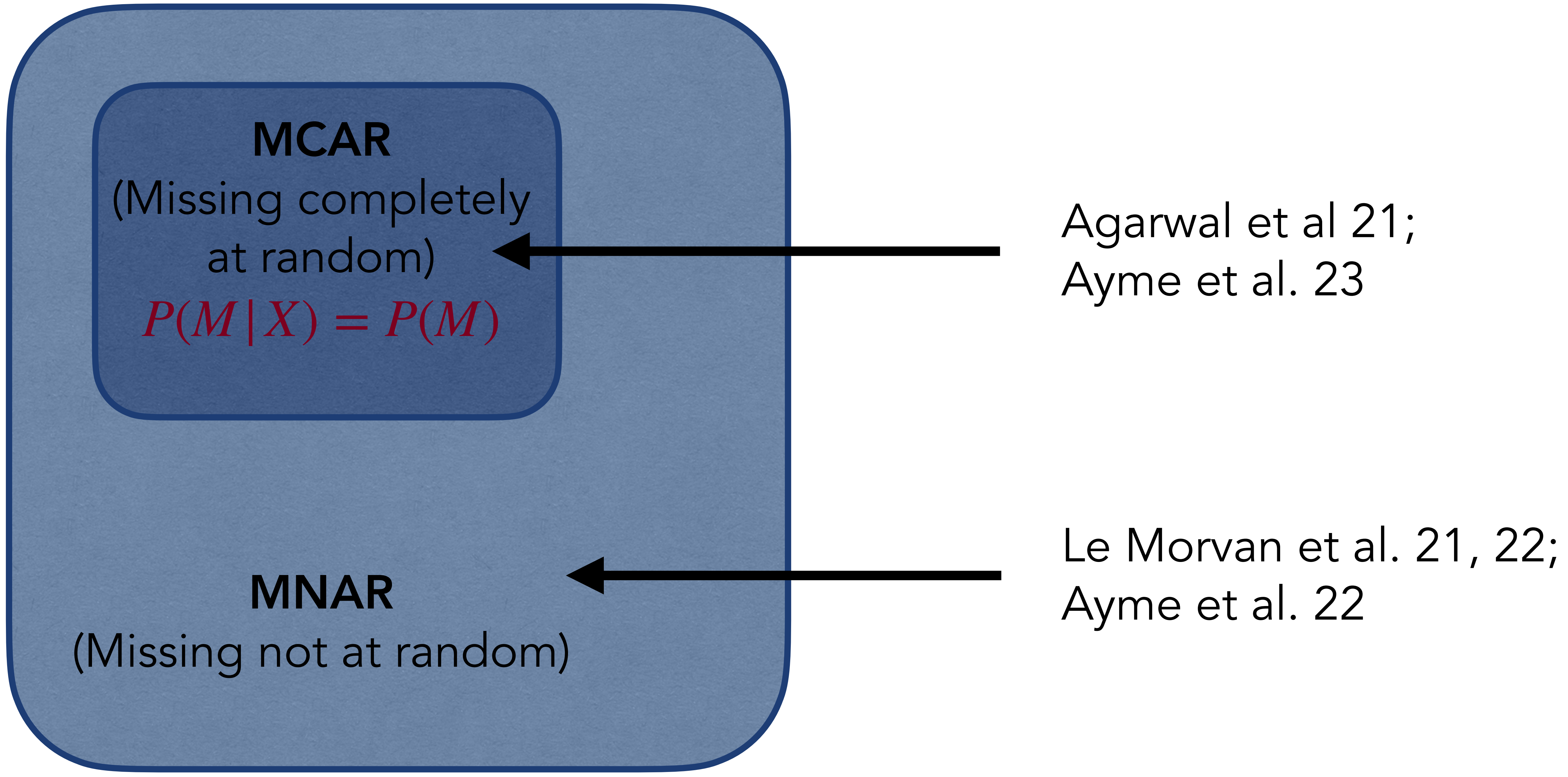
  - if model is well specified: $\mathbb{E}[\epsilon_i | X_i] = 0$
  - else: $\mathbb{E}[\epsilon_i X_i] = 0$

- **Inference**: estimate the model parameter β
- **Prediction**: predict Y on a new observation X

Estimation of β is not sufficient

$$X = (\text{NA}, 8, 0, \text{NA}, 6, 2)$$

- **Missing data mechanism**



Introduction: Handle missing values

- **Handle missing values with:**

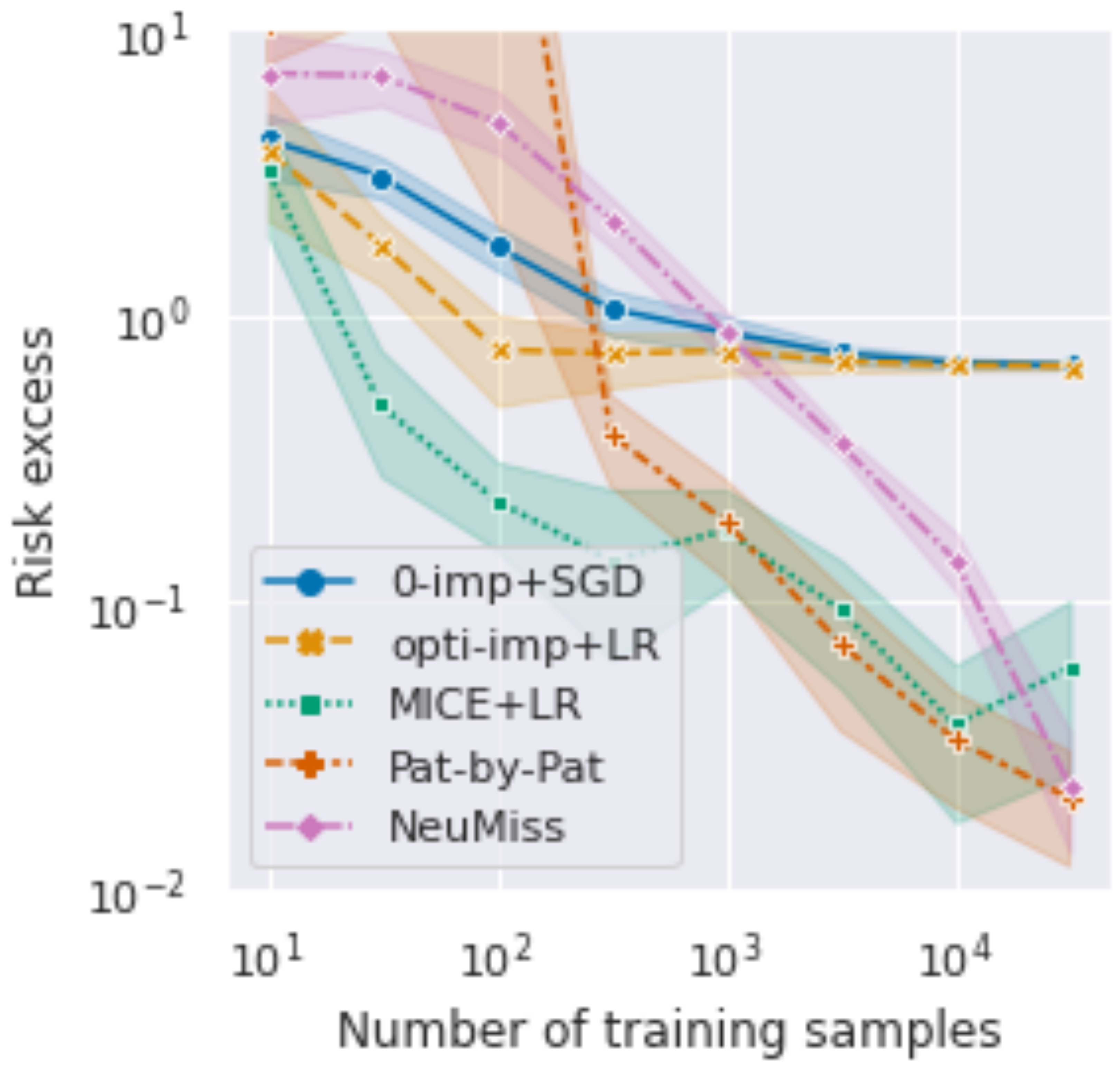
1. Impute-then-regress procedure (e.g. imputation by 0)
2. Specific method (e.g. pattern-by-pattern)

Introduction: Handle missing values

- **Handle missing values with:**

1. Impute-then-regress procedure (e.g. imputation by 0)
2. Specific method (e.g. pattern-by-pattern)

- **Low dimension $n \to +\infty$**

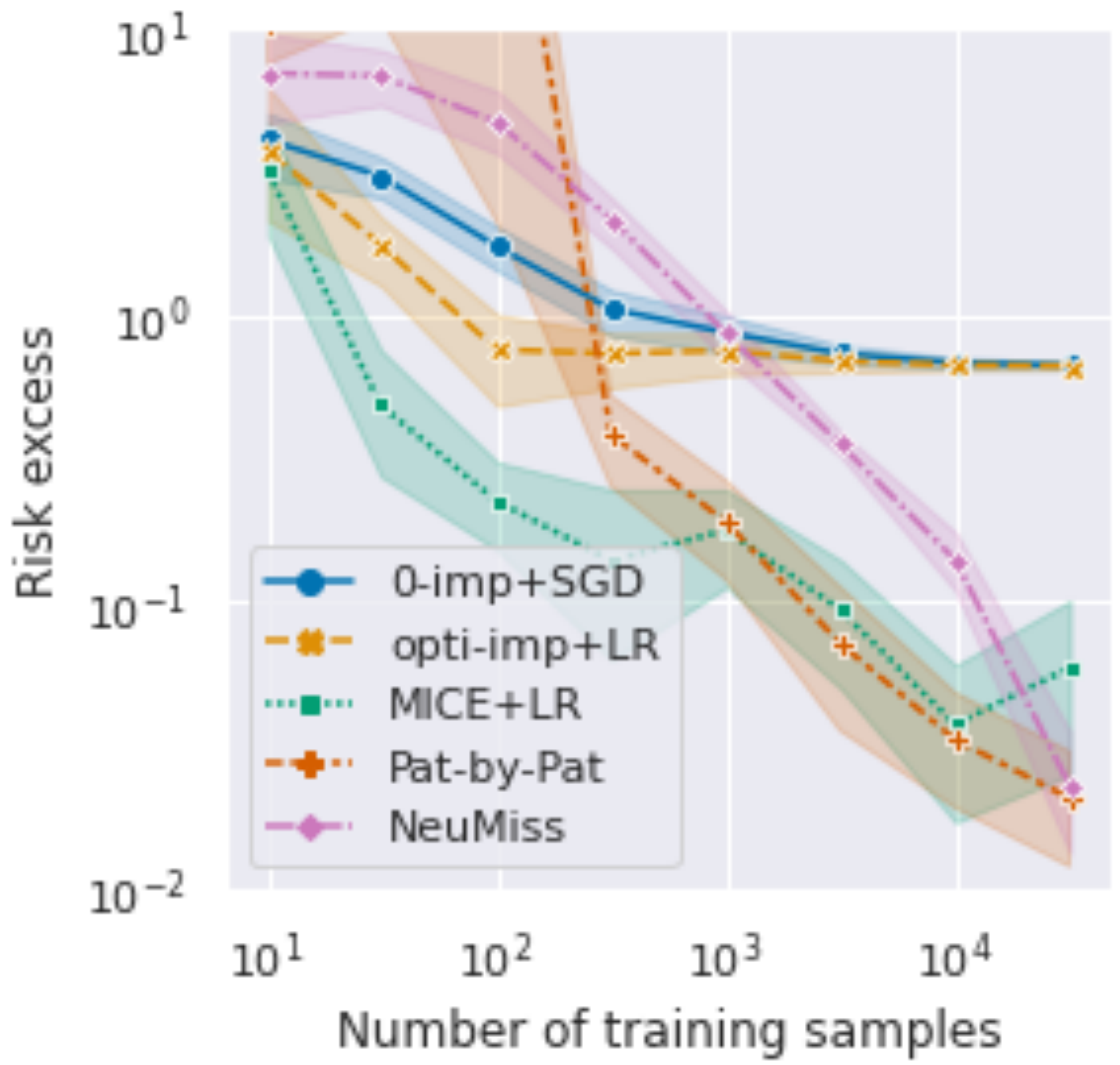


Introduction: Handle missing values

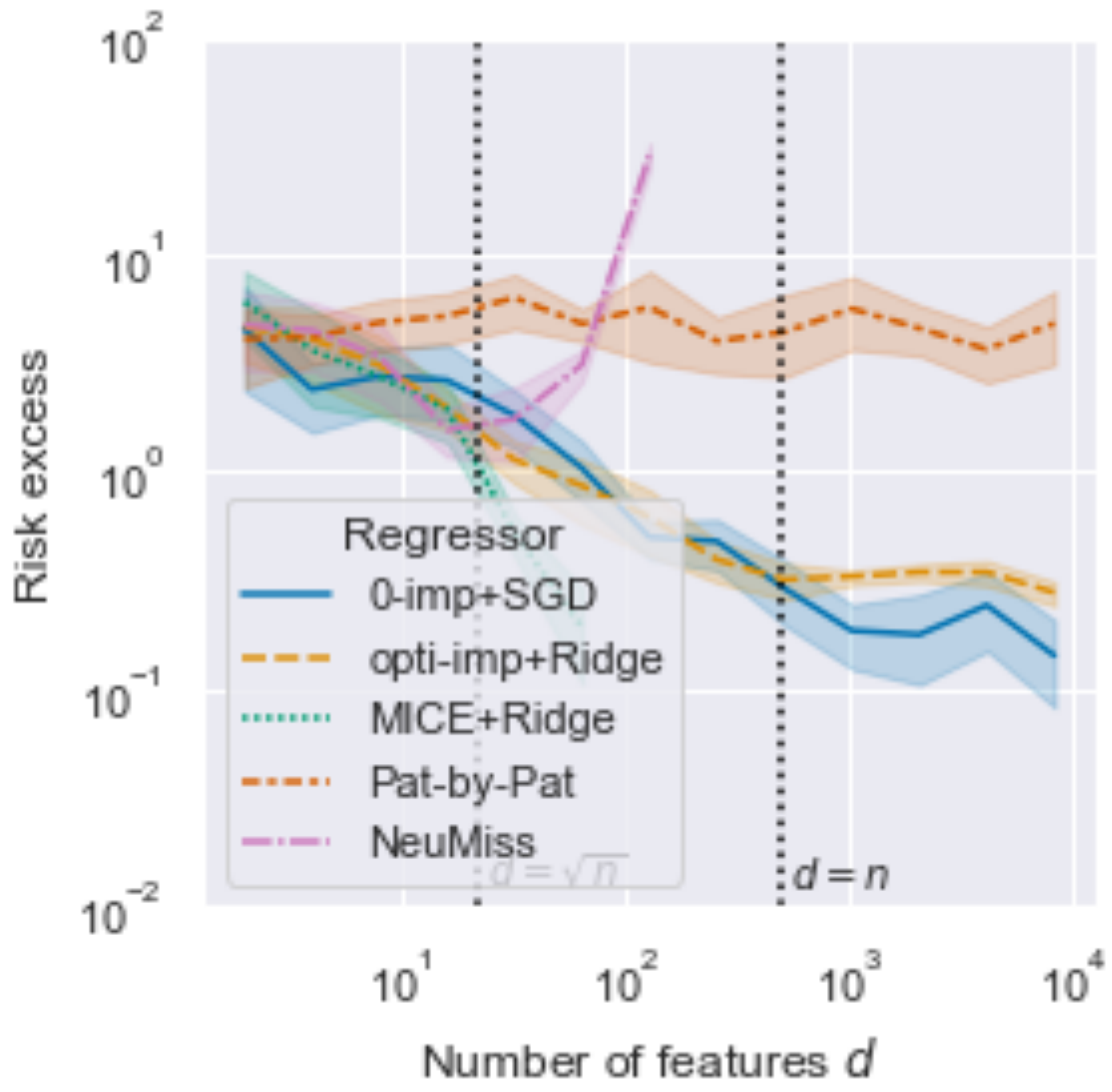
- **Handle missing values with:**

1. Impute-then-regress procedure (e.g. imputation by 0)
2. Specific method (e.g. pattern-by-pattern)

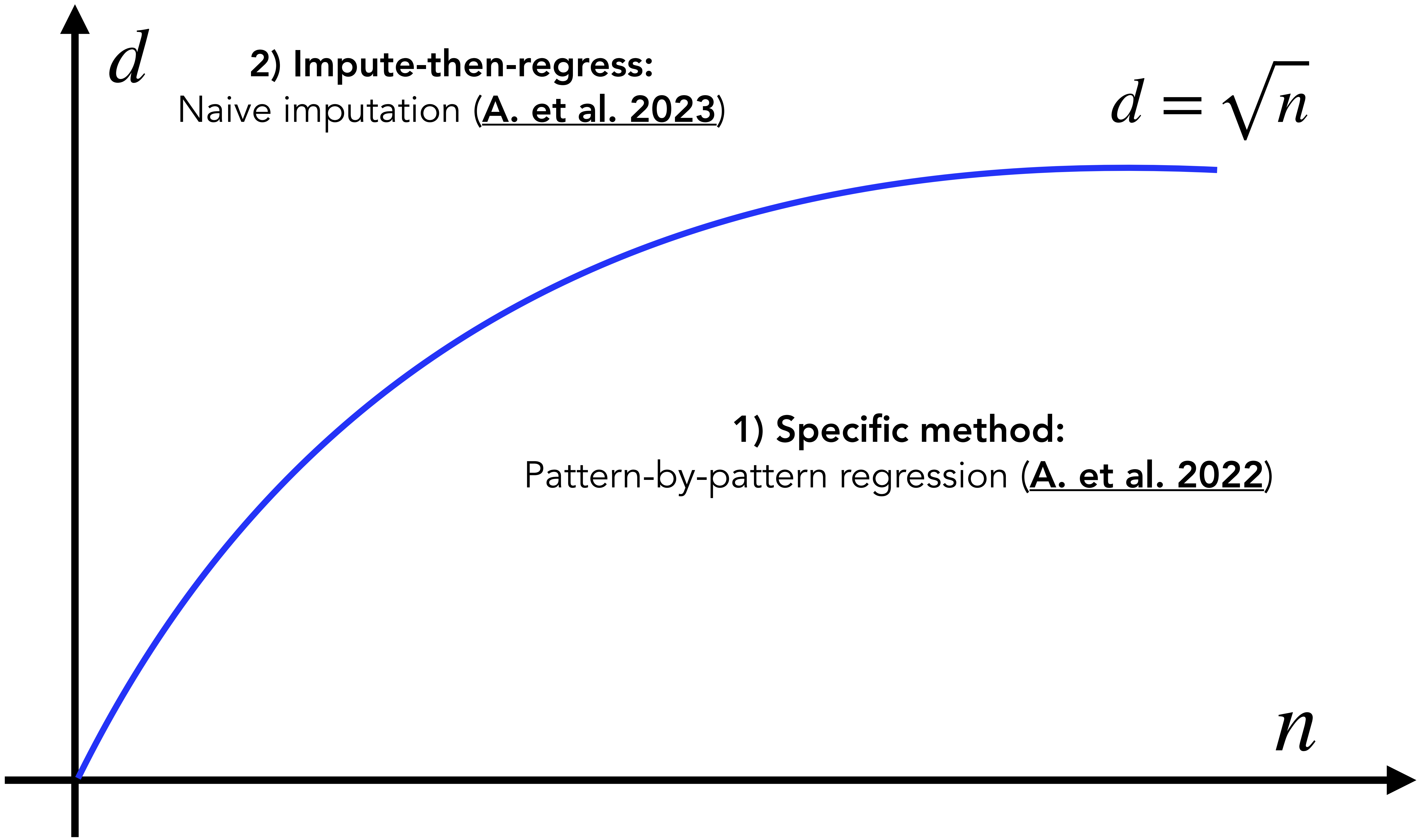
- **Low dimension $n \to +\infty$**



- **High dimension $d \to +\infty$**



In this talk



d
2) Impute-then-regress:
Naive imputation (**A. et al. 2023**)
$d = \sqrt{n}$
1) Specific method:
Pattern-by-pattern regression (**A. et al. 2022**)
n

1) Specific method: Pattern-by-Pattern regression

- **Bayes predictor** decomposition

$$f^{\star}(Z) = \sum_{m \in \{0,1\}^d} f^{\star}_m(X_{\mathrm{obs(m)}}) \mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^{\star}_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{\mathrm{obs(m)}}) \mathbf{1}_{M=m}$$

Local **Least-Square** regression on
$\left\{(X_{i,obs}, Y_i), M_i = m\right\}$

1) Specific method: Pattern-by-Pattern regression

- **Bayes predictor** decomposition

$$f^\star(Z) = \sum_{m \in \{0,1\}^d} f^\star_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

↑

Local **Bayes prediction** for the missing pattern $(M = m)$

> **Proposition: (Le Morvan et al. 2020)**
> Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

↑

Local **Least-Square** regression on

$\{(X_{i,obs}, Y_i), M_i = m\}$

- Definition: excess risk

$$\mathscr{E}(\hat{f}) = \mathbb{E}\left[\left(\hat{f}(Z) - f^\star(Z)\right)^2\right]$$

- Definition: missing pattern **complexity**

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m \in \{0,1\}^d} p_m \wedge \frac{d}{n}$$

1) Specific method: Pattern-by-Pattern regression

- **Bayes predictor** decomposition

$$f^\star(Z) = \sum_{m\in\{0,1\}^d} f^\star_m(X_{obs(m)})\mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat f(Z) = \sum_{m\in\{0,1\}^d} \hat f_m(X_{obs(m)})\mathbf{1}_{M=m}$$

Local **Least-Square** regression on
$\{(X_{i,obs}, Y_i), M_i = m\}$

- Definition: excess risk

$$\mathscr{E}\left(\hat f\right) = \mathbb{E}\left[\left(\hat f(Z) - f^\star(Z)\right)^2\right]$$

- Definition: missing pattern **complexity**

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m\in\{0,1\}^d} p_m \wedge \frac{d}{n}$$

Theorem:
Under Lipschitz and Sub-Gaussian assumptions

$$\mathscr{E}(\hat f) \leq A\log(n)\mathfrak{C}_p\left(\frac{d}{n}\right) + \text{Approx}$$

1) Specific method: Pattern-by-Pattern regression

- **Bayes predictor** decomposition

$$f^\star(Z) = \sum_{m\in\{0,1\}^d} f^\star_m(X_{obs(m)})\mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat{f}(Z) = \sum_{m\in\{0,1\}^d} \hat{f}_m(X_{obs(m)})\mathbf{1}_{M=m}$$

Local **Least-Square** regression on
$\{(X_{i,obs}, Y_i), M_i = m\}$

- Definition: excess risk

$$\mathscr{E}\left(\hat{f}\right) = \mathbb{E}\left[\left(\hat{f}(Z) - f^\star(Z)\right)^2\right]$$

- Definition: missing pattern **complexity**

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m\in\{0,1\}^d} p_m \wedge \frac{d}{n}$$

Theorem:
Under Lipschitz and Sub-Gaussian assumptions

$$\mathscr{E}(\hat{f}) \leq A\log(n)\mathfrak{C}_p\left(\frac{d}{n}\right) + \text{Approx}$$

Examples:

1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d\frac{d}{n}$
2. **Bernoulli** distribution: $M_j \sim \mathscr{B}(1-\rho)$ and $1-\rho \leq \frac{d}{n}$

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq \frac{d^2}{n}$$

1) Specific method: Pattern-by-Pattern regression

- **Minimax risk**

Worst case on a class of problem $\mathscr{P}_p$

$$\mathscr{E}_{\text{mini}}\left(p\right) = \inf_{\tilde{f}} \sup_{\mathbb{P}\in\mathscr{P}_p} \mathbb{E}_{\mathbb{P}}\left[\left(\tilde{f}(Z) - f^{\star}(Z)\right)^2\right]$$

Best algorithm

where $\mathscr{P}_p$ represents a class of data distributions

- for which the missing pattern distribution is p
- under Lipschitz and Sub-Gaussian assumptions

1) Specific method: Pattern-by-Pattern regression

- **Minimax risk**

Worst case on a class of problem $\mathscr{P}_p$

$$\mathscr{E}_{\text{mini}}(p) = \inf_{\tilde{f}} \sup_{\mathbb{P} \in \mathscr{P}_p} \mathbb{E}_{\mathbb{P}} \left[\left(\tilde{f}(Z) - f^\star(Z) \right)^2 \right]$$

Best algorithm

Theorem:

$$\sigma^2 \mathfrak{C}_p \left(\frac{1}{n} \right) \lesssim \mathscr{E}_{\text{mini}}(p) \underbrace{\leq \mathscr{E}(\hat{f}) \leq A \log(n) \mathfrak{C}_p \left(\frac{d}{n} \right)}_{\text{previous thm}}$$

- Lower bound still holds when $\mathscr{P}_p$ includes **MAR** missing values

where $\mathscr{P}_p$ represents a class of data distributions

- for which the missing pattern distribution is p
- under Lipschitz and Sub-Gaussian assumptions

1) Specific method: Pattern-by-Pattern regression

- **Minimax risk**

Worst case on a class of problem $\mathscr{P}_p$

$$\mathscr{E}_{\text{mini}}(p) = \inf_{\tilde{f}} \sup_{\mathbb{P}\in\mathscr{P}_p} \mathbb{E}_{\mathbb{P}}\left[\left(\tilde{f}(Z) - f^\star(Z)\right)^2\right]$$

Best algorithm

where $\mathscr{P}_p$ represents a class of data distributions

- for which the missing pattern distribution is p
- under Lipschitz and Sub-Gaussian assumptions

Theorem:

$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \mathscr{E}_{\text{mini}}(p) \underbrace{\leq \mathscr{E}(\hat{f}) \leq A\log(n)\mathfrak{C}_p\left(\frac{d}{n}\right)}_{\text{previous thm}}$$

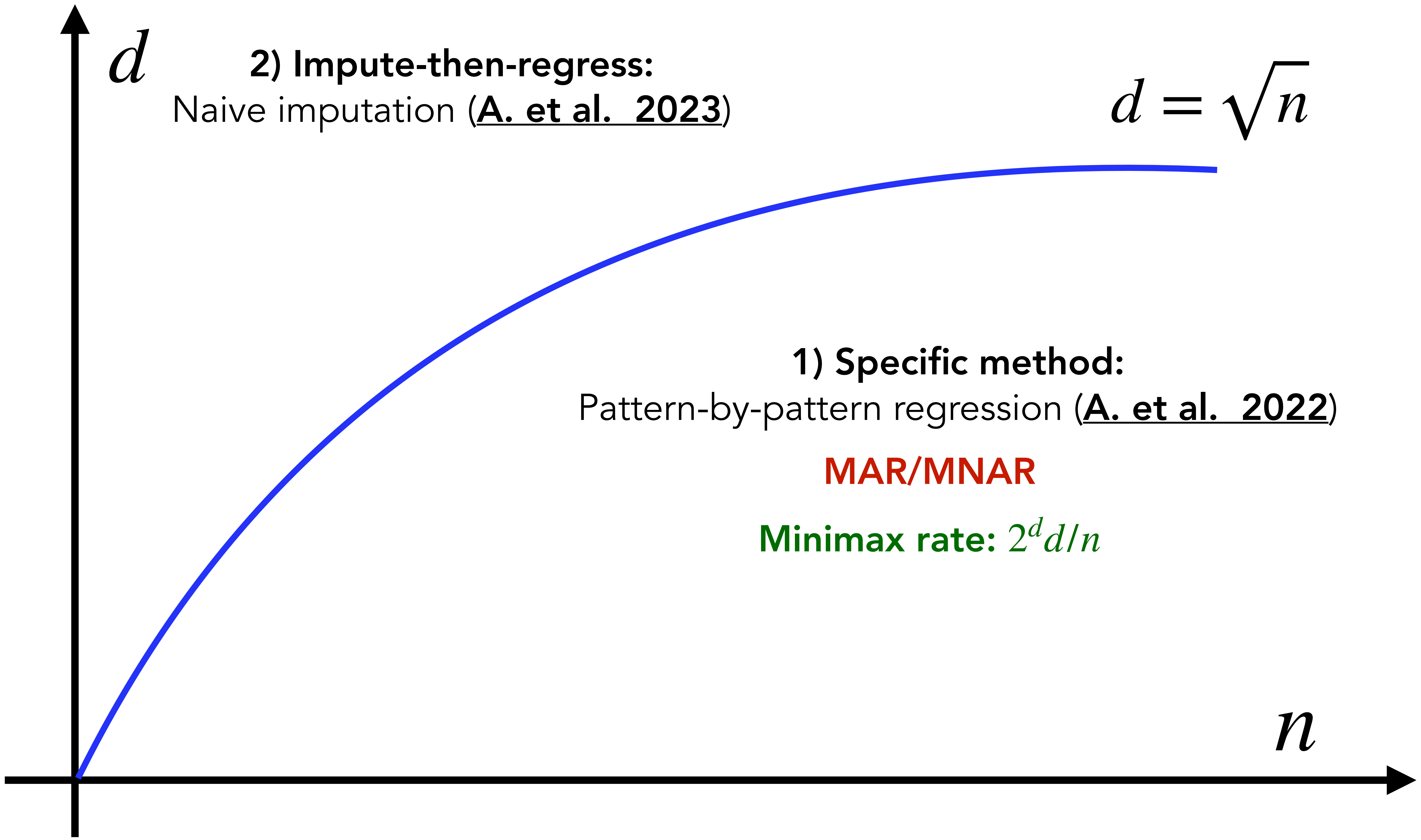
- Lower bound still holds when $\mathscr{P}_p$ includes **MAR** missing values

Examples

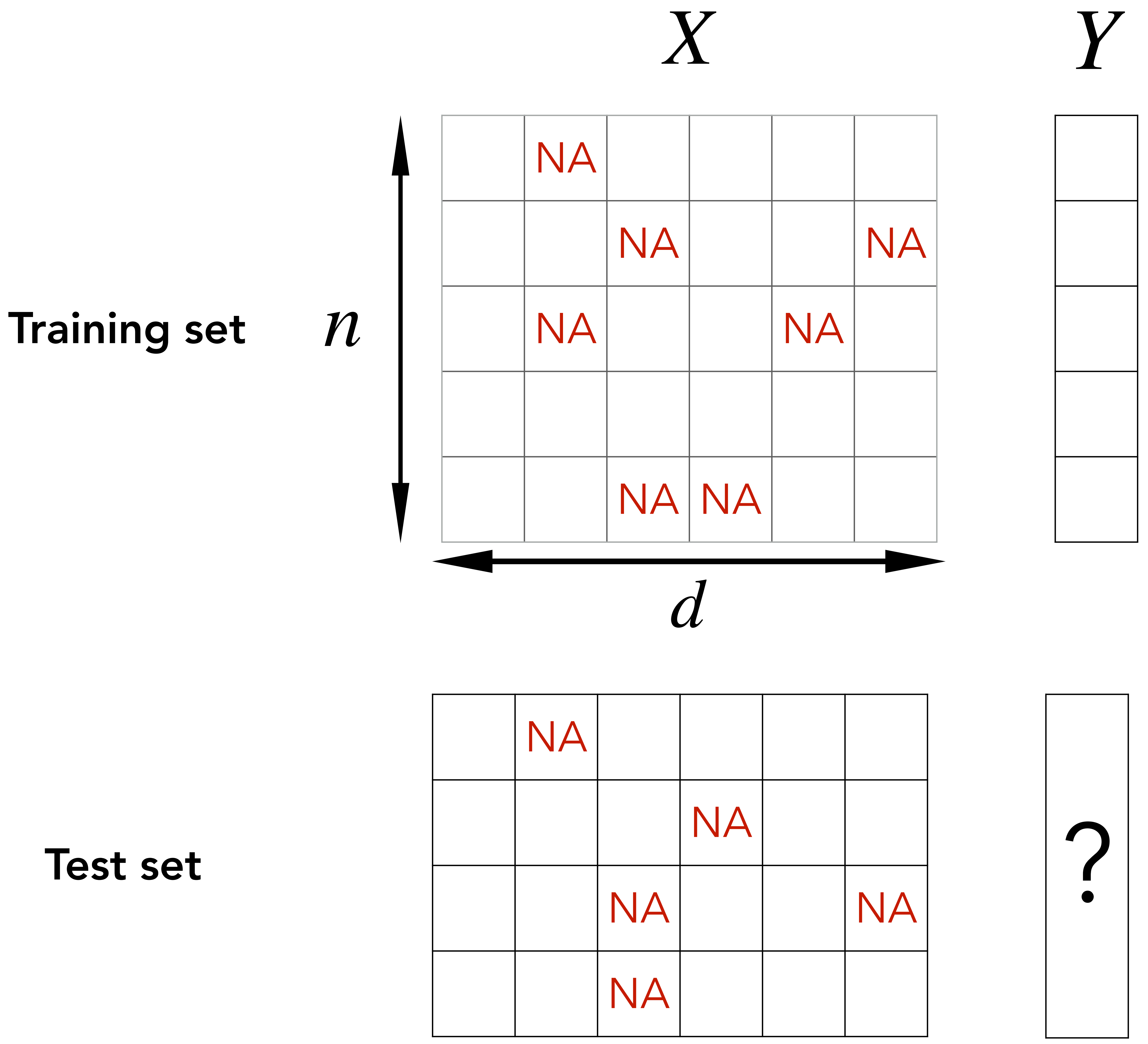
1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{2^d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d\frac{d}{n}$
2. **Bernoulli** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = \frac{d^2}{n}$

- **without stronger assumption, best rate can be exponential!**

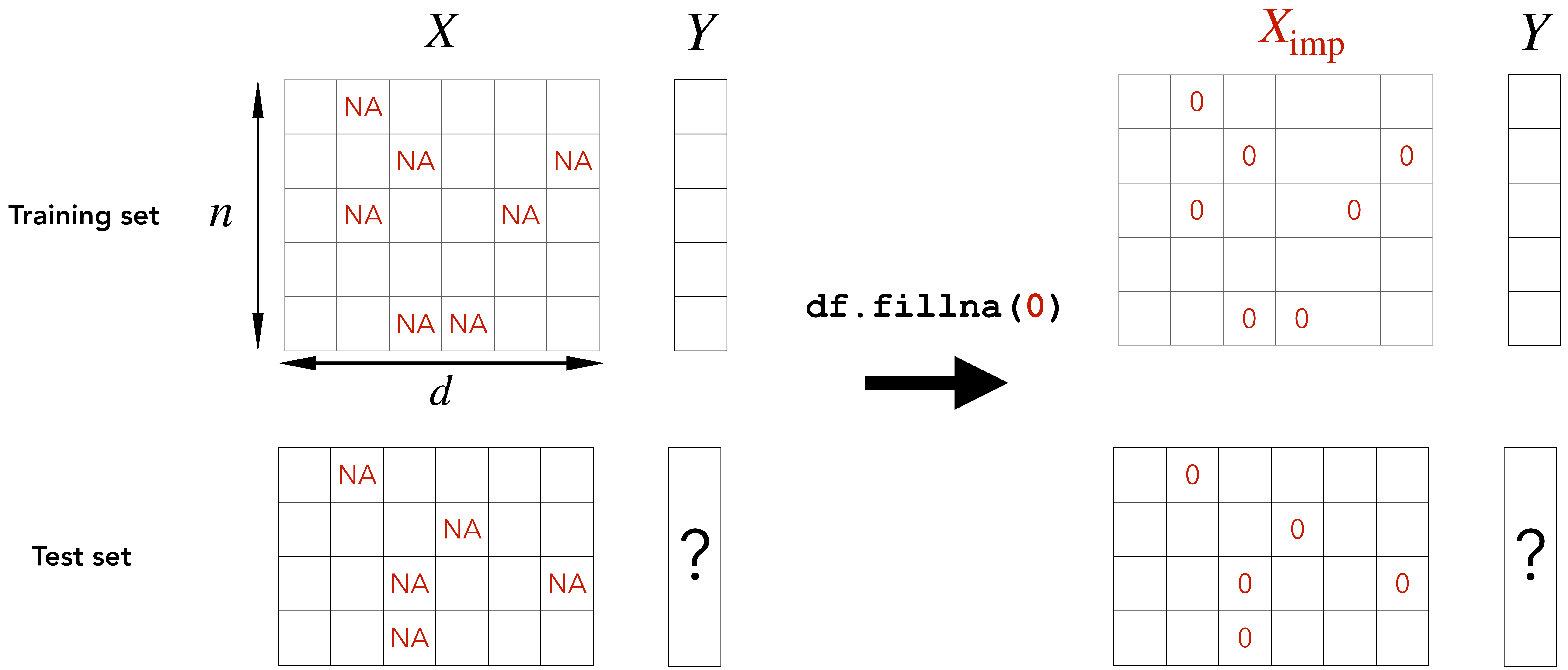
1) Specific method:



2) Imputation by 0



2) Imputation by 0



2) Imputation by 0: Framework

- Linear prediction risk on **imputed data:**

$$R_{\text{imp}}(\theta) = \mathbb{E}\left[\left(Y - \theta^\top X_{\text{imp}}\right)^2\right]$$

2) Imputation by 0: Framework

- Linear prediction risk on **imputed data:**

$$R_{\text{imp}}(\theta) = \mathbb{E}\left[\left(Y - \theta^\top X_{\text{imp}}\right)^2\right]$$

- **Imputation Bias:**

Risk of the optimal **linear** predictor on **complete** data

$$B_{\text{imp}} = R^\star_{\text{imp}} - R^\star$$

Risk of the optimal **linear** predictor on **0-imputed** data

2) Imputation by 0: Framework

- Linear prediction risk on **imputed data:**

$$R_{\text{imp}}(\theta) = \mathbb{E}\left[\left(Y - \theta^\top X_{\text{imp}}\right)^2\right]$$

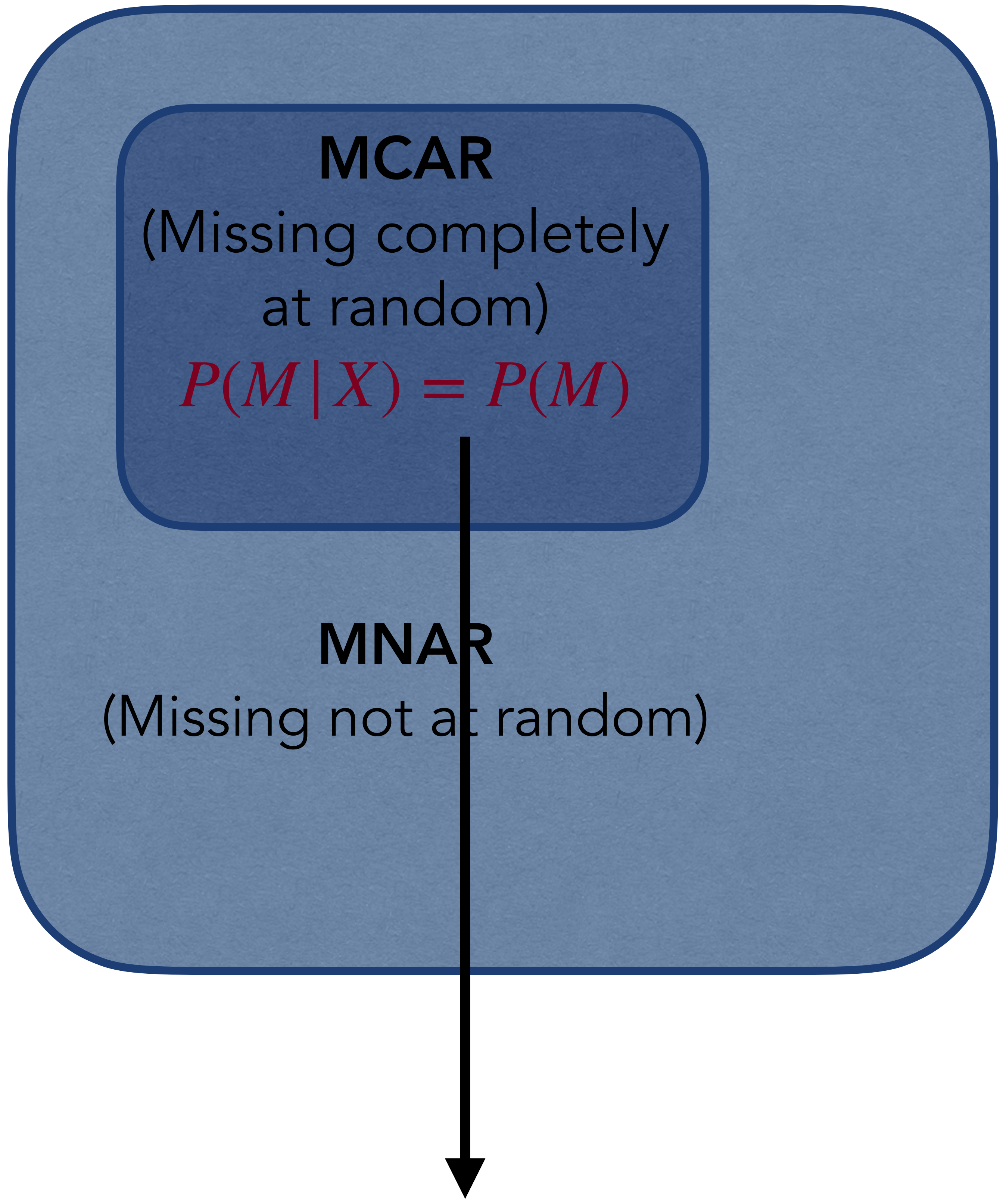
- **Imputation Bias:**

Risk of the optimal **linear** predictor on **complete** data

$$B_{\text{imp}} = R^\star_{\text{imp}} - R^\star$$

Risk of the optimal **linear** predictor on **0-imputed** data

- **Missing values**



Bernoulli Model: Missing values i.i.d

$M_1, \ldots, M_d \sim \mathscr{B}(1-\rho)$

2) Imputation by 0: Toy example

- **Complete Model:**

$Y = X_1$.

$X = (X_1, X_1, \ldots, X_1)$

$\theta^\star = (1,0,...,0)^\top$

$R^\star = 0$

- **With imputed missing values:** $M_1, .., M_d \sim \mathscr{B}(1/2)$

2) Imputation by 0: Toy example

- **Complete Model:**

$Y = X_1$.

$X = (X_1, X_1, \ldots, X_1)$

$\theta^\star = (1,0,\ldots,0)^\top$

$R^\star = 0$

- **With imputed missing values:** $M_1, \ldots, M_d \sim \mathscr{B}(1/2)$

$\theta_1 = (1,0,\ldots,0)^\top$

↓

$\theta_1^\top X_{\text{imp}} = X_1 M_1$

↓

$R(\theta_1) = \frac{1}{2}\mathbb{E}[X_1^2]$

2) Imputation by 0: Toy example

- **Complete Model:**

$$Y = X_1 \, .$$

$$X = (X_1, X_1, \ldots, X_1)$$

$$\theta^\star = (1,0,\ldots,0)^\top$$

$$R^\star = 0$$

- **With imputed missing values:** $M_1, \ldots, M_d \sim \mathscr{B}(1/2)$

$$\theta_1 = (1,0,\ldots,0)^\top \quad \longrightarrow \quad \theta_1^\top X_{\text{imp}} = X_1 M_1 \quad \longrightarrow \quad R(\theta_1) = \frac{1}{2}\mathbb{E}[X_1^2]$$

$$\theta_2 = 2(1/d, 1/d, \ldots, 1/d)^\top \quad \longrightarrow \quad \theta_2^\top X_{\text{imp}} = \frac{2X_1}{d}\sum_j M_J \quad \longrightarrow \quad R(\theta_2) = \frac{1}{d}\mathbb{E}[X_1^2]$$

$$B_{\text{imp}} = R^\star - R_0^\star \leq \frac{1}{d}\mathbb{E}[X_1^2]$$

2) Imputation by 0 = implicit ridge?

- **Ridge penalization**

$$R_\lambda(\theta) = R(\theta) + \lambda \|\theta\|_2^2$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|_2^2$$

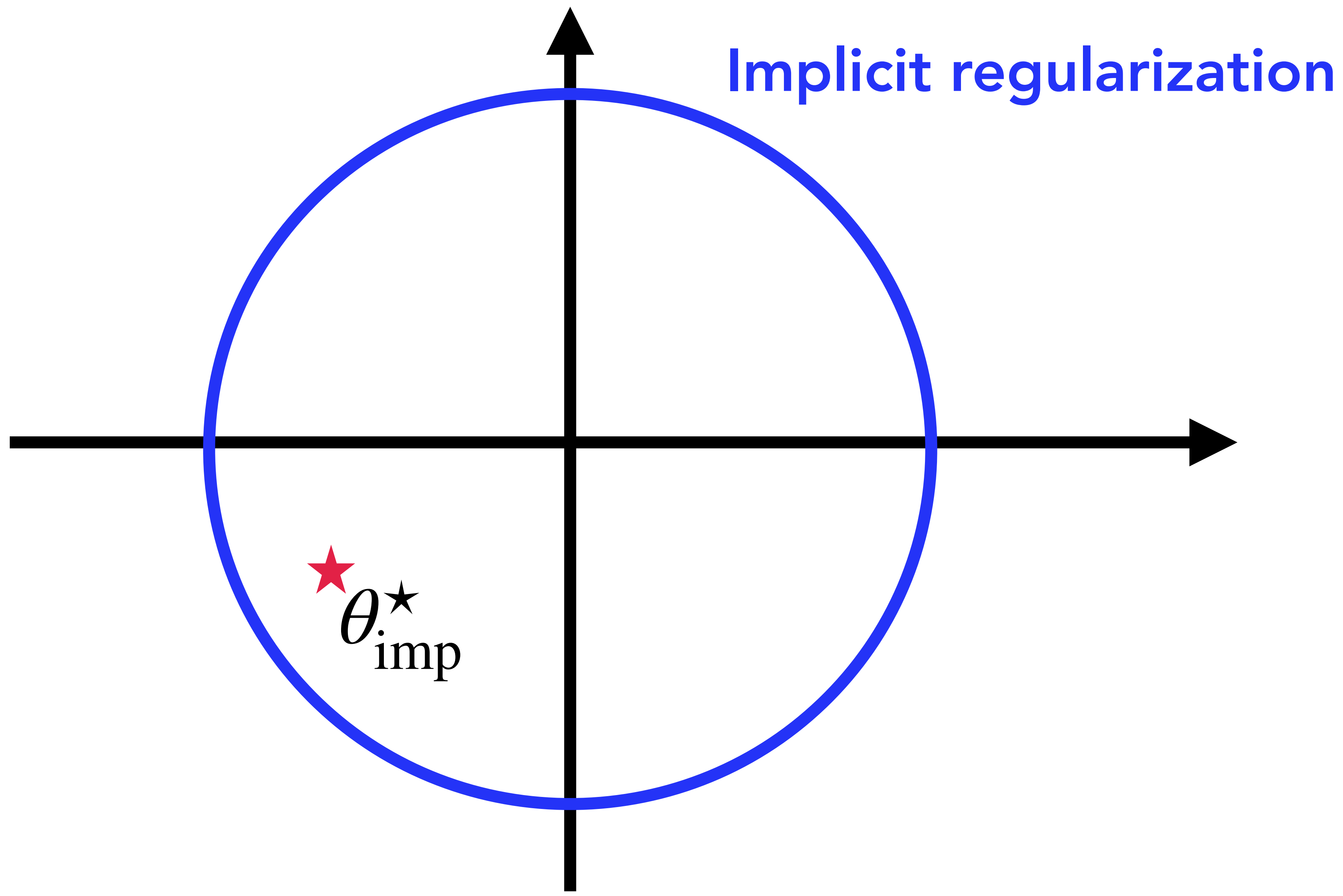
2) Imputation by 0 = implicit ridge?

- **Ridge penalization**

$$R_\lambda(\theta) = R(\theta) + \lambda\|\theta\|_2^2$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|_2^2$$



1. **Imputation induce a ridge penalization** (Optimal predictor has a small norm)

2) Imputation by 0 = implicit ridge?

- **Ridge penalization**

$$R_\lambda(\theta) = R(\theta) + \lambda \|\theta\|_2^2$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|_2^2$$

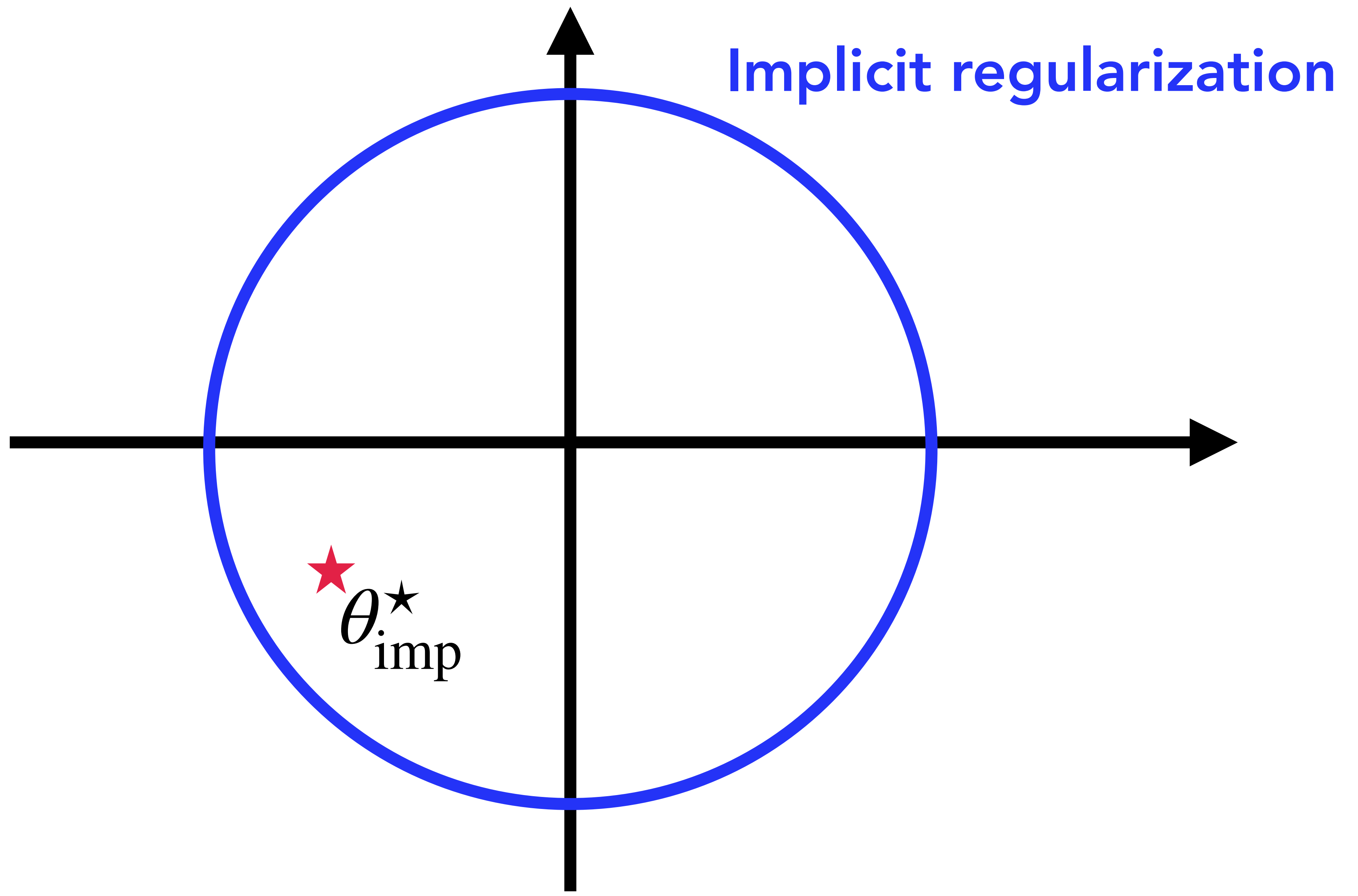
- **Ridge bias**

$$B_{\text{ridge},\lambda} = \inf_\theta \{R(\theta) - R(\theta_\star) + \lambda \|\theta\|_2^2\}$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$B_{\text{imp}} = B_{\text{ridge},\lambda_{\text{imp}}}$$

where $\lambda_{\text{imp}} = \dfrac{\rho}{1-\rho}$



1. **Imputation induce a ridge penalization** (Optimal predictor has a small norm)

2) Imputation by 0 = implicit ridge?

- **Ridge penalization**

$$R_\lambda(\theta) = R(\theta) + \lambda \|\theta\|_2^2$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|_2^2$$

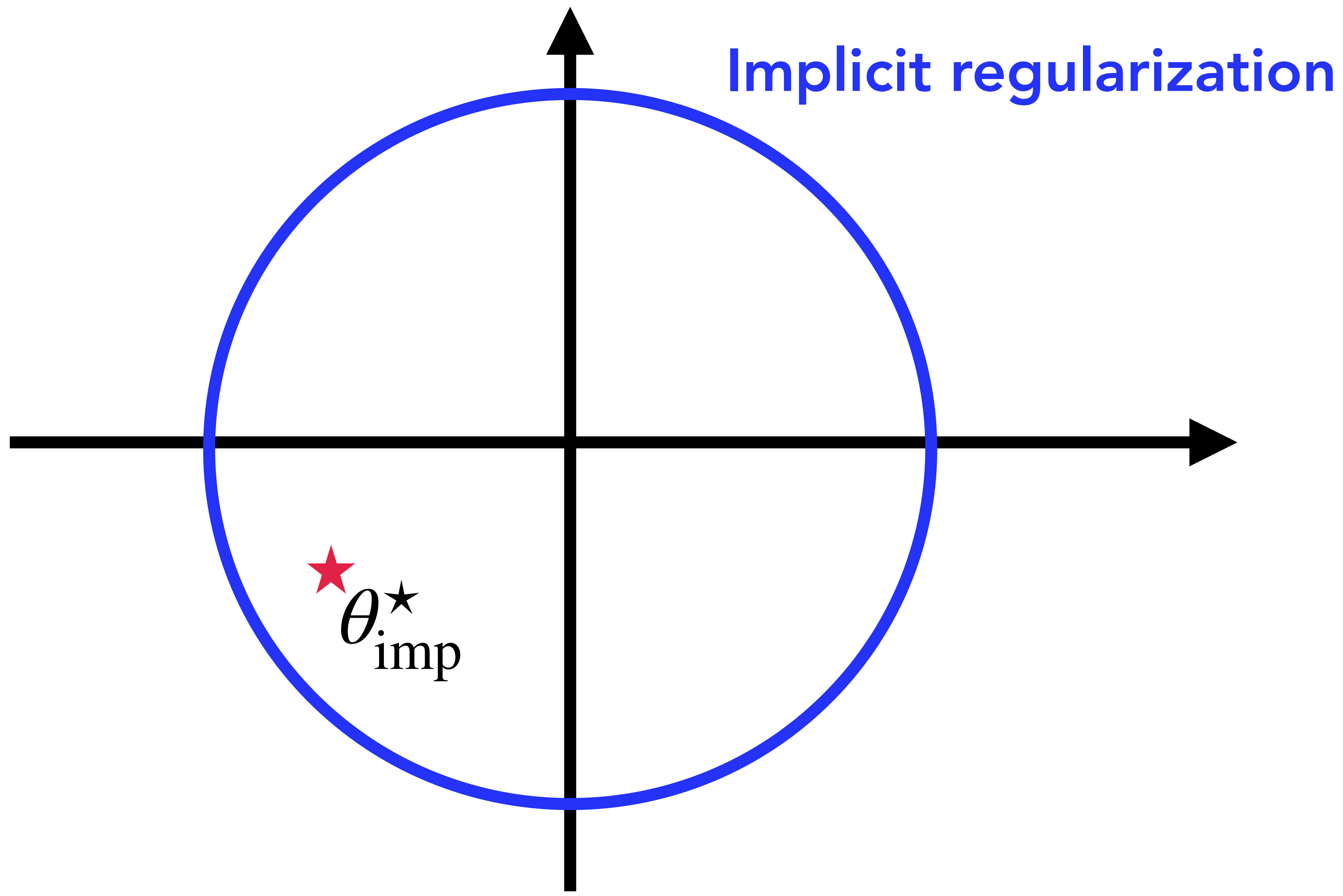
- **Ridge bias**

$$B_{\text{ridge},\lambda} = \inf_\theta \{R(\theta) - R(\theta_\star) + \lambda \|\theta\|_2^2\}$$

Theorem: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$B_{\text{imp}} = B_{\text{ridge},\lambda_{\text{imp}}}$$

where $\lambda_{\text{imp}} = \dfrac{\rho}{1-\rho}$



1. **Imputation induce a ridge penalization** (Optimal predictor has a small norm)
2. **Imputation by 0 seem to be at the same price of ridge penalization**
3. Penalization parameter λ_{imp} depends only on $1-\rho$ the proportion of missing values.
4. Available for all MCAR setting with another λ_{imp}

2) Imputation by 0: Illustration on low rank data

- **Low rank data (or spiked):** $\mathrm{rank}(\Sigma) \approx r$

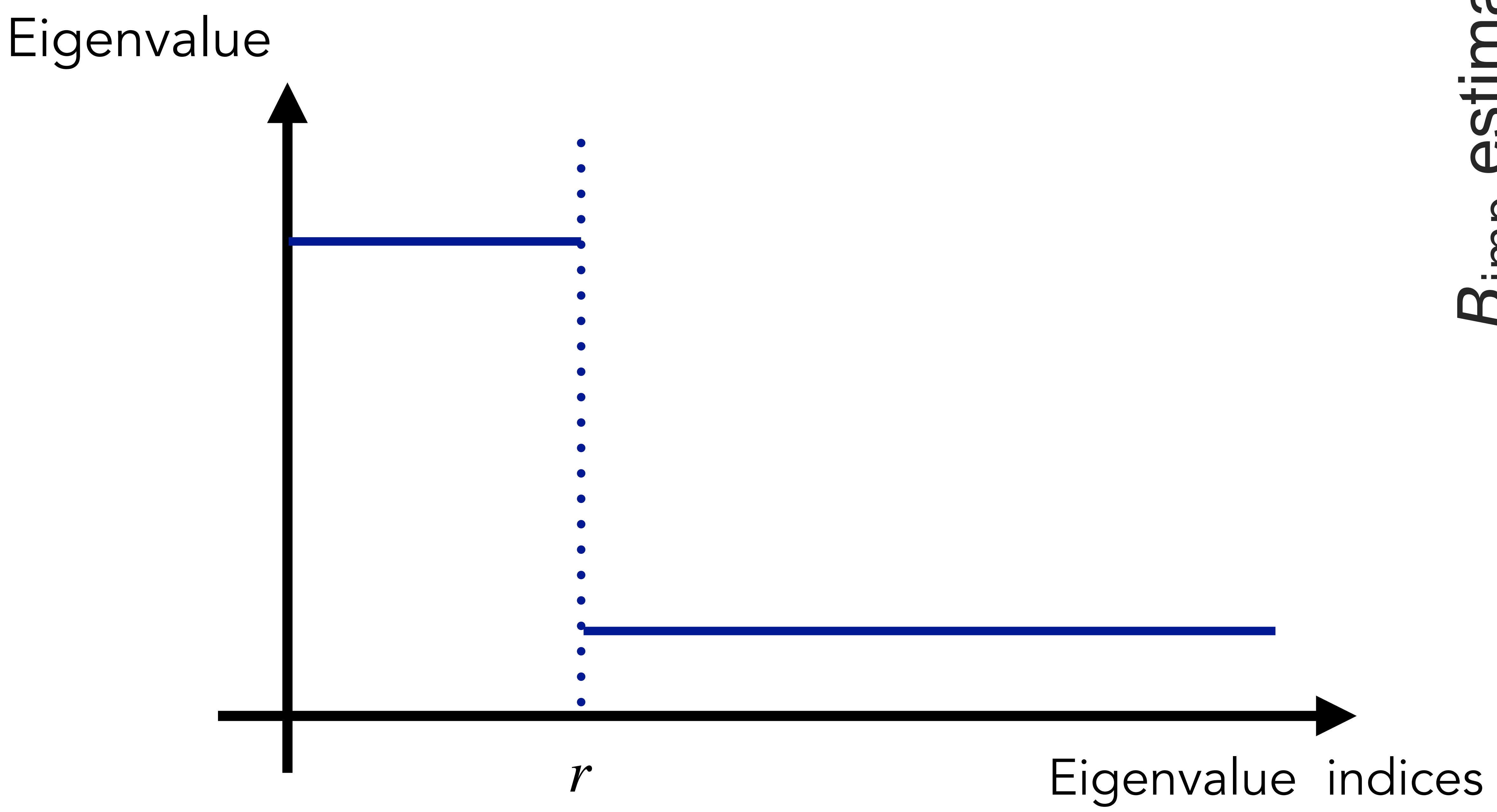
$$B_{\text{imp}} \lesssim \frac{r}{d}\mathbb{E}Y^2$$

Spectrum of covariance matrix

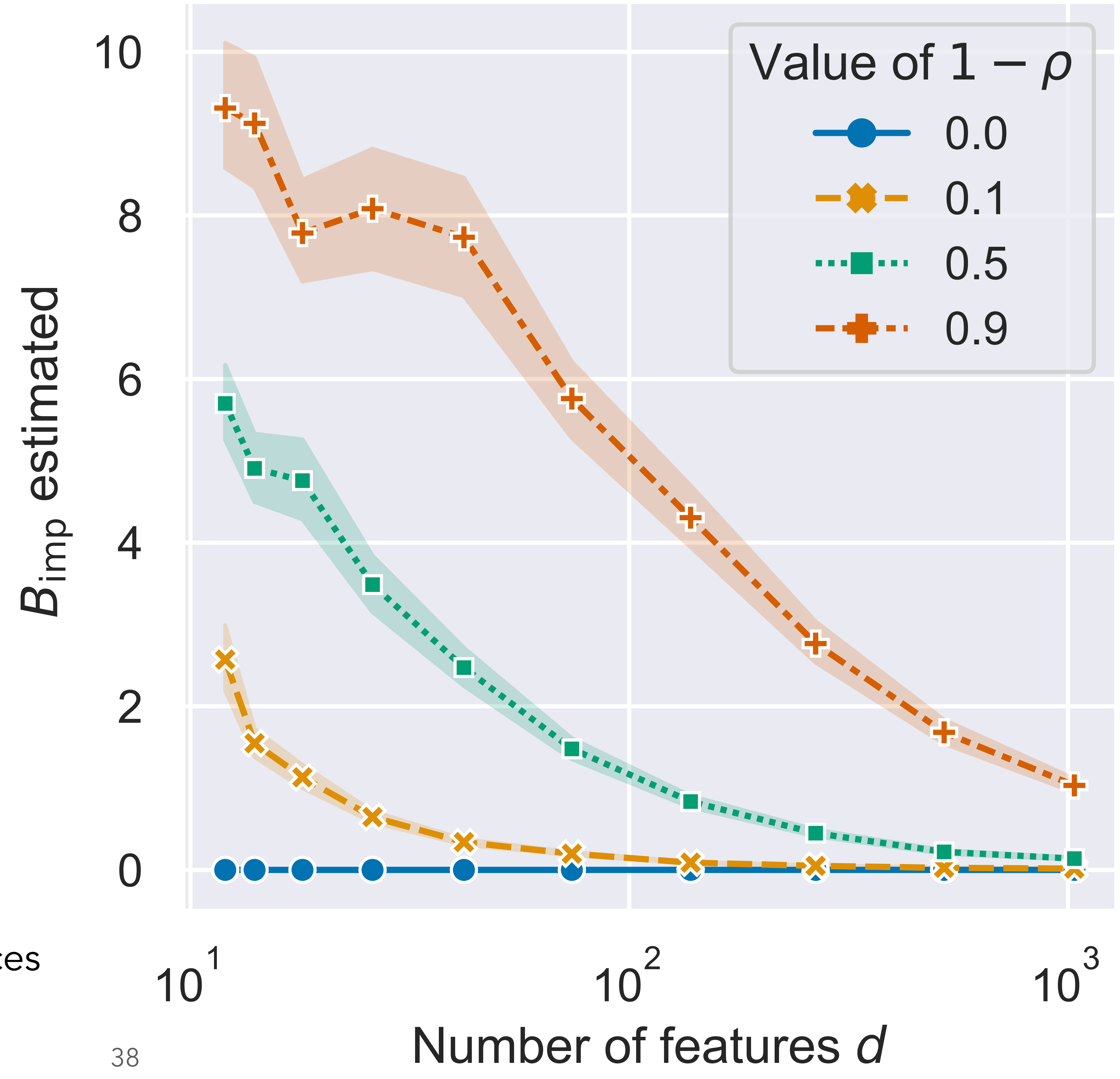
2) Imputation by 0: Illustration on low rank data

- **Low rank data (or spiked):** $\text{rank}(\Sigma) \approx r$

$$B_{\text{imp}} \lesssim \frac{r}{d}\mathbb{E}Y^2$$



Spectrum of covariance matrix



2) Imputation by 0: Learn imputed data with SGD

- **SGD recursion:** with **constant** learning rate $\gamma = \dfrac{1}{d\sqrt{n}}$

$$\begin{cases} \theta_0 = 0 \\ \theta_{\text{imp},t} = \left[I - \gamma X_{\text{imp},t} X_{\text{imp},t}^\top\right] \theta_{\text{imp},t-1} + \gamma Y_t X_{\text{imp},t} \end{cases}$$

- Polyak Ruppert average:

$$\bar{\theta}_{\text{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\text{imp},t}$$

2) Imputation by 0: Learn imputed data with SGD

- **SGD recursion:** with **constant** learning rate $\gamma = \dfrac{1}{d\sqrt{n}}$

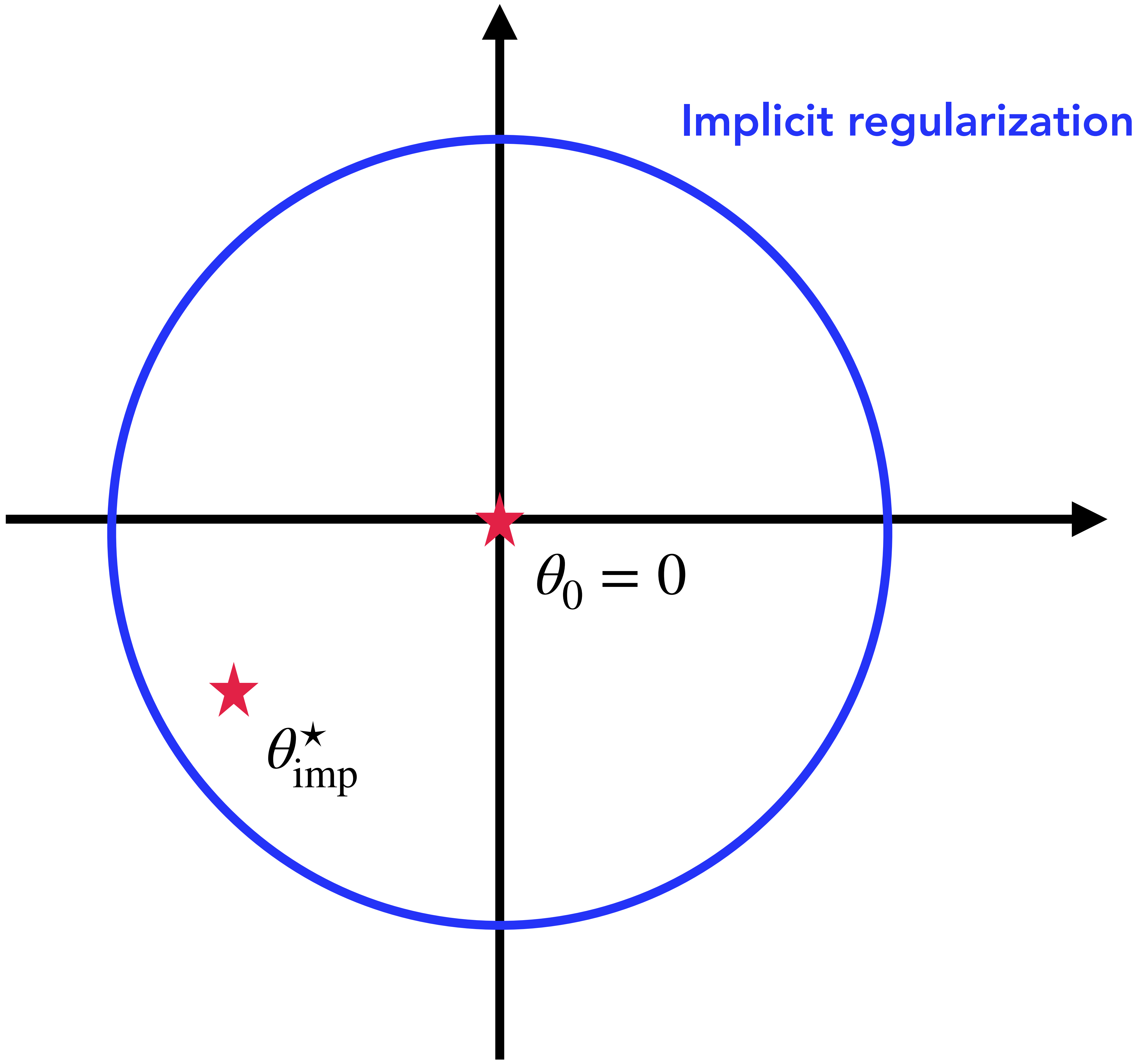
$$\begin{cases} \theta_0 = 0 \\ \theta_{\mathrm{imp},t} = \left[I - \gamma X_{\mathrm{imp},t} X_{\mathrm{imp},t}^\top\right] \theta_{\mathrm{imp},t-1} + \gamma Y_t X_{\mathrm{imp},t} \end{cases}$$

- Polyak Ruppert average:

$$\bar{\theta}_{\mathrm{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\mathrm{imp},t}$$

Theorem: Under classical SGD assumptions,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp,n}}\right)\right] - R^\star \leq B_{\mathrm{imp}} + \frac{d}{\sqrt{n}} \|\theta^\star_{\mathrm{imp}}\|_2^2 + \frac{\sigma^2}{\sqrt{n}}$$



2) Imputation by 0: Learn imputed data with SGD

- **SGD recursion:**

$$\begin{cases} \theta_0 = 0 \\ \theta_{\mathrm{imp},t} = \left[I - \gamma X_{\mathrm{imp},t} X_{\mathrm{imp},t}^\top\right] \theta_{\mathrm{imp},t-1} + \gamma Y_t X_{\mathrm{imp},t} \end{cases}$$

- Polyak Ruppert average:

$$\bar{\theta}_{\mathrm{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\mathrm{imp},t}$$

Theorem: Under classical SGD assumptions,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp,n}}\right)\right] - R^\star \leq B_{\mathrm{imp}} + \frac{d}{\sqrt{n}} \|\theta^\star_{\mathrm{imp}}\|_2^2 + \frac{\sigma^2}{\sqrt{n}}$$

- **Illustration on low rank:**

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp,n}}\right)\right] - R^\star \leq \left(\frac{1}{\rho\sqrt{n}} + \frac{1-\rho}{d}\right) \frac{r}{\rho} \mathbb{E} Y^2 + \frac{\sigma^2}{\sqrt{n}}$$

2) Imputation by 0: Learn imputed data with SGD

- **SGD recursion:**

$$\begin{cases} \theta_0 = 0 \\ \theta_{\text{imp},t} = \left[I - \gamma X_{\text{imp},t} X_{\text{imp},t}^\top \right] \theta_{\text{imp},t-1} + \gamma Y_t X_{\text{imp},t} \end{cases}$$

- Polyak Ruppert average:

$$\bar{\theta}_{\text{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\text{imp},t}$$

Theorem: Under classical SGD assumptions,

$$\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp,n}} \right) \right] - R^\star \le B_{\text{imp}} + \frac{d}{\sqrt{n}} \|\theta^\star_{\text{imp}}\|_2^2 + \frac{\sigma^2}{\sqrt{n}}$$

- **Illustration on low rank:**

$$\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp,n}} \right) \right] - R^\star \le \left(\frac{1}{\rho\sqrt{n}} + \frac{1-\rho}{d} \right) \frac{r}{\rho} \mathbb{E} Y^2 + \frac{\sigma^2}{\sqrt{n}}$$

1. We leverage on implicit regularization.
2. Streaming online (one passe)
3. Trade-off between imputation bias and initial condition.
4. **Imputation bias vanishes for $d \gg \sqrt{n}$**

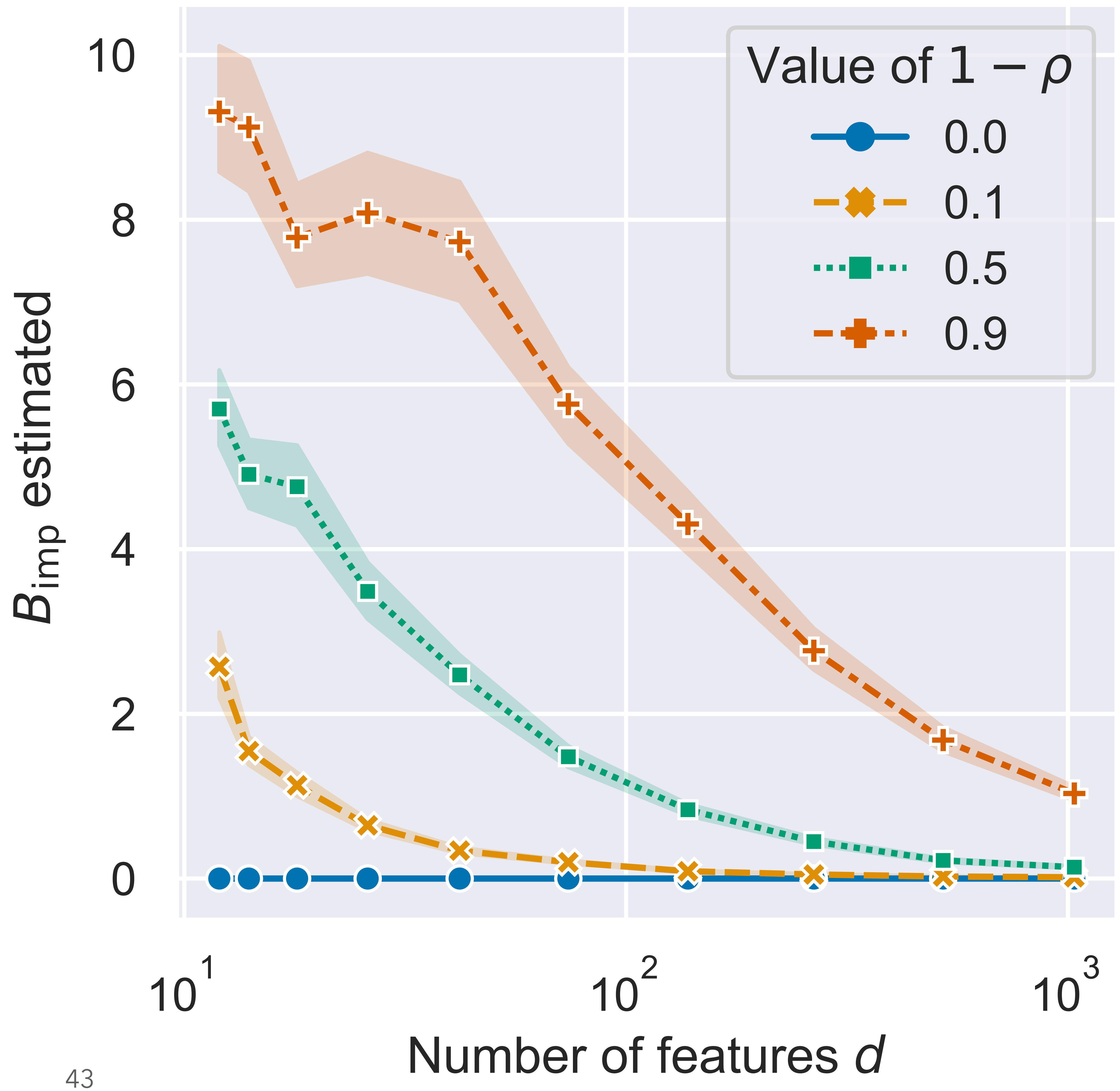
2) Imputation by 0: Conclusion

1. In practice: In high-dimension imputation (even naive) out performs specific methods designed to handle missing values.

2. Imputation by 0 induces a **Ridge penalization**.

3. Imputation bias **vanishes** with dimension. As a consequence missing values are not an issue in high dimension (correlated setting).

4. The regime $d \gg \sqrt{n}$ leads to **slow rates** of consistency.

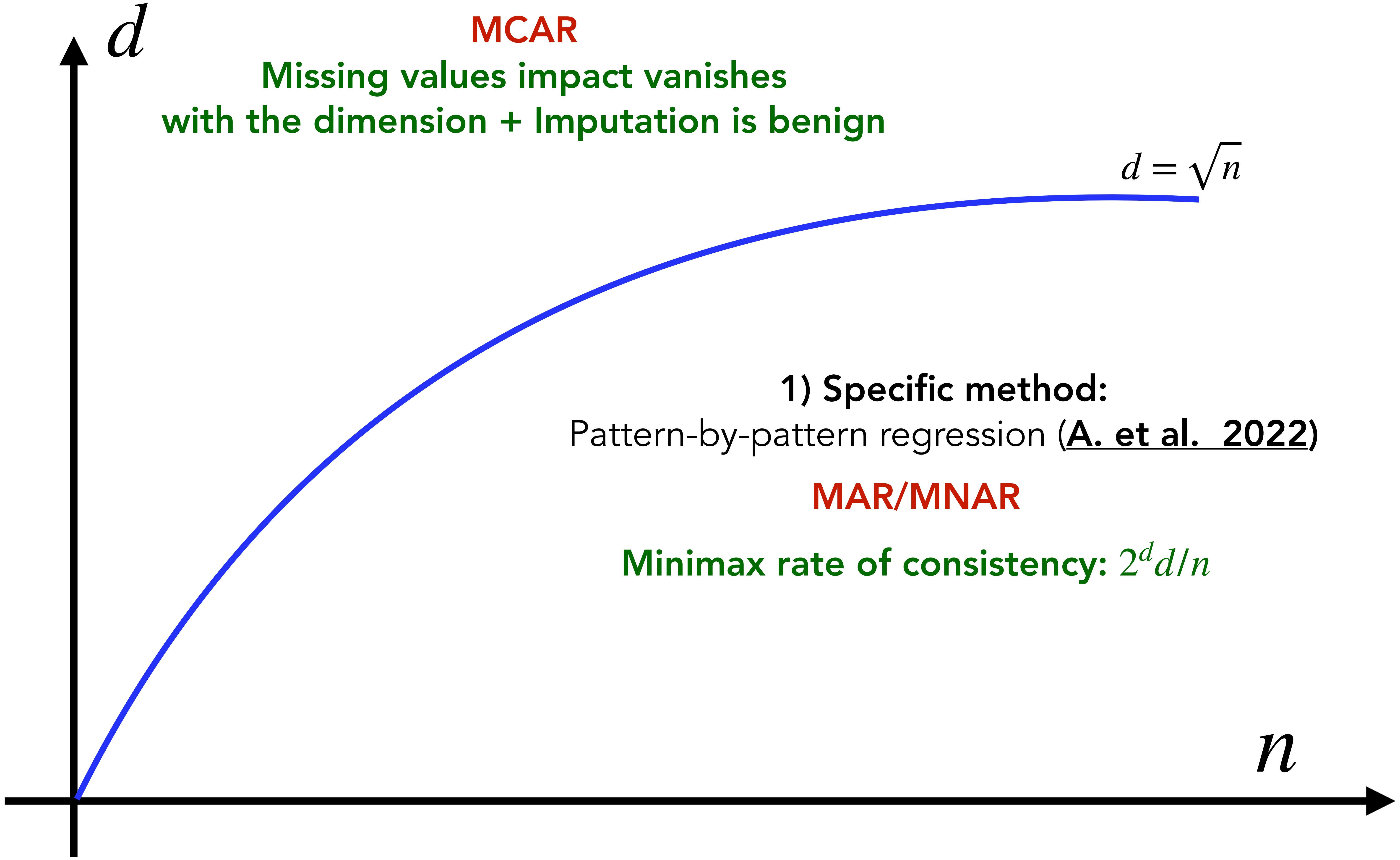


Conclusion

2) Impute then regress:
Naive imputation (**A. et al. 2023**)

MCAR

Missing values impact vanishes with the dimension + Imputation is benign



1) Specific method:
Pattern-by-pattern regression (**A. et al. 2022**)

MAR/MNAR

Minimax rate of consistency: $2^d d/n$

Corresponding Author: alexis.ayme@sorbonne-universite.fr

References

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. Naive imputation implicitly regularizes high-dimensional linear models (link)

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. Near- optimal rate of consistency for linear models (link)

Agarwal, A., Shah, D., Shen, D., and Song, D. On robustness of principal component regression

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varo- quaux, G. Linear predictor on linearly-generated data with missing values: non consistency and solutions.

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. NeuMiss networks: differentiable programming for supervised learning with missing values.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. What's a good imputation to predict with missing values?

Rubin, D. B. Inference and missing data.