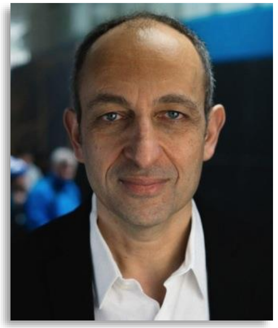
Causal Matrix Completion

Dennis Shen

Joint work with



Anish Agarwal

Munther Dahleh

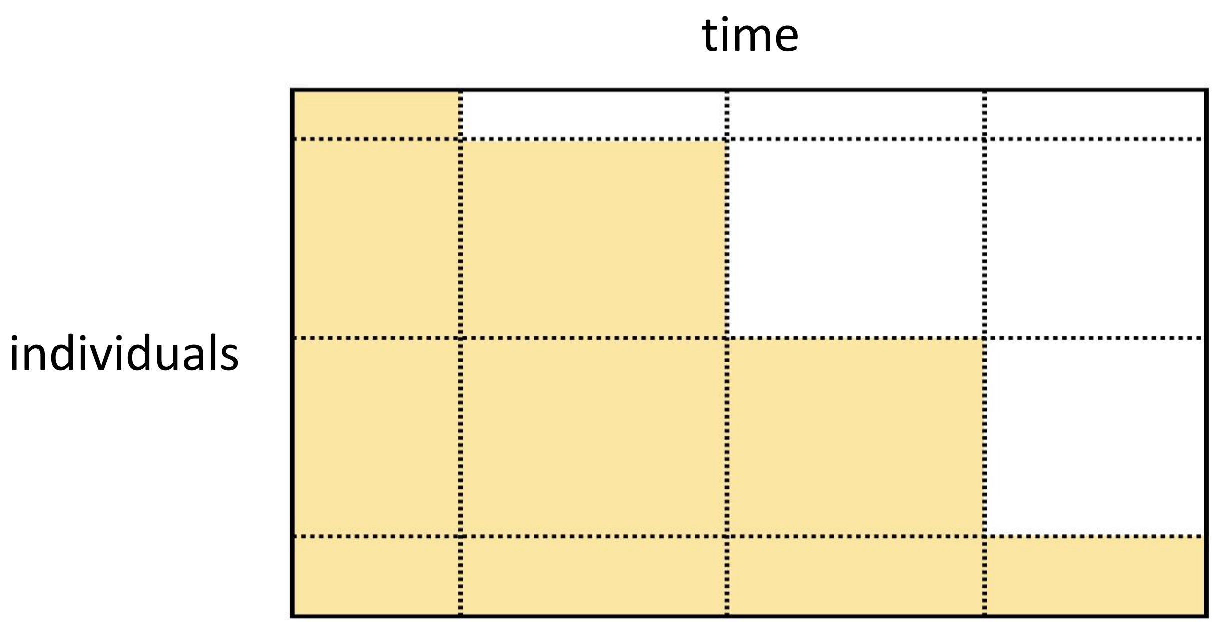
Devavrat Shah

Matrix completion

★★★★	?	?	?	★★★★★	?
★★★★	?	?	★★★★	★★★★★	?
★★★	★	?	?	?	★★★★★
★★★	★★	★★★★	?	?	?

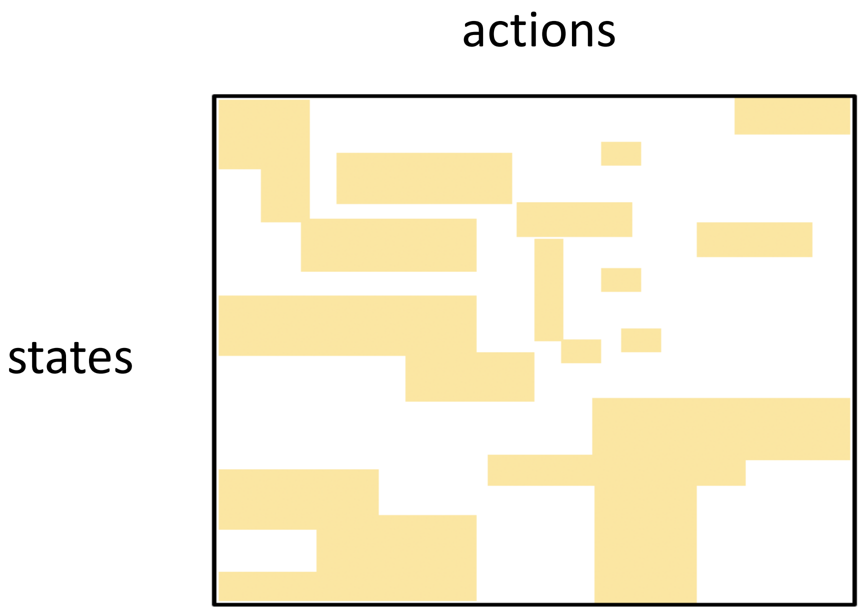
Can we recover the missing entries?

Policy evaluation



Staggered adoption
e.g., Medicare

Contextual bandits



Confounded data

e.g., (state, action) pairs with high reward will be exploited more

Matrix completion encodes wide variety of applications

Common goal
impute missing entries
& de-noise observed entries

Impediment to a unified approach
different applications induce
different sparsity patterns

Formal setup

Expected outcomes: $\boldsymbol{M} \in \mathbb{R}^{m \times n}$

Random outcomes: $Y_{ij} = M_{ij} + \varepsilon_{ij}$

Binary mask: $\boldsymbol{A} \in \{0,1\}^{m \times n}$

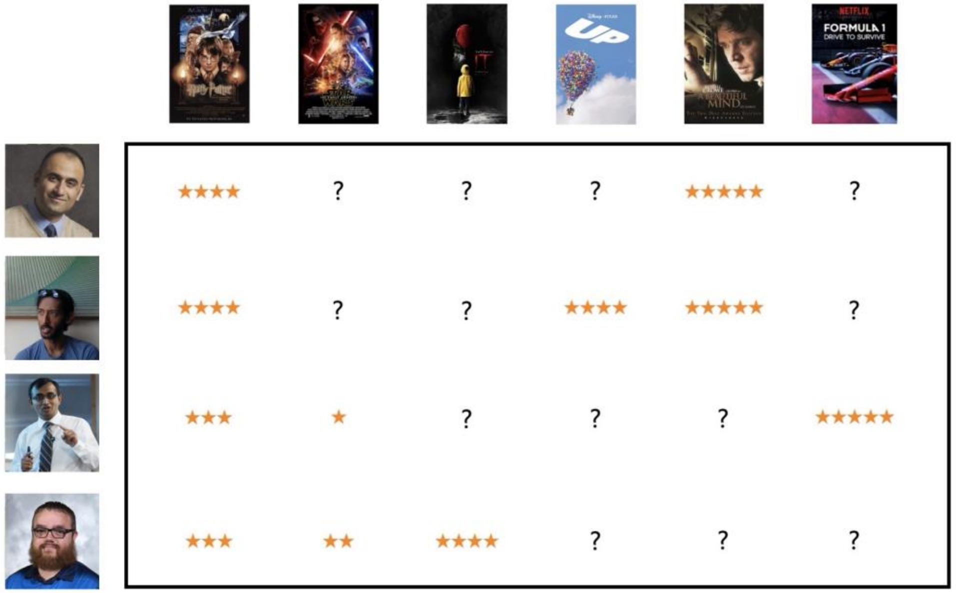
Observation: $\widetilde{Y}_{ij} = \begin{cases} Y_{ij}, & \text{if } A_{ij} = 1 \\ ?, & \text{if } A_{ij} = 0 \end{cases}$

Given $(\widetilde{\boldsymbol{Y}}, \boldsymbol{A})$, produce $\widehat{\boldsymbol{M}}$ such that $\widehat{\boldsymbol{M}} \approx \boldsymbol{M}$

error measured with respect to $\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_q : q \in \{F, 2, \infty, \dots\}$

Where does causality come into play?

Causality = missingness mechanism



Under what conditions is $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{A}$?

A brief history

Missing Completely at Random (MCAR)

↓

Missing at Random (MAR)

↓

Missing Not at Random (MNAR)

A brief history

Missing Completely at Random (MCAR)

↓

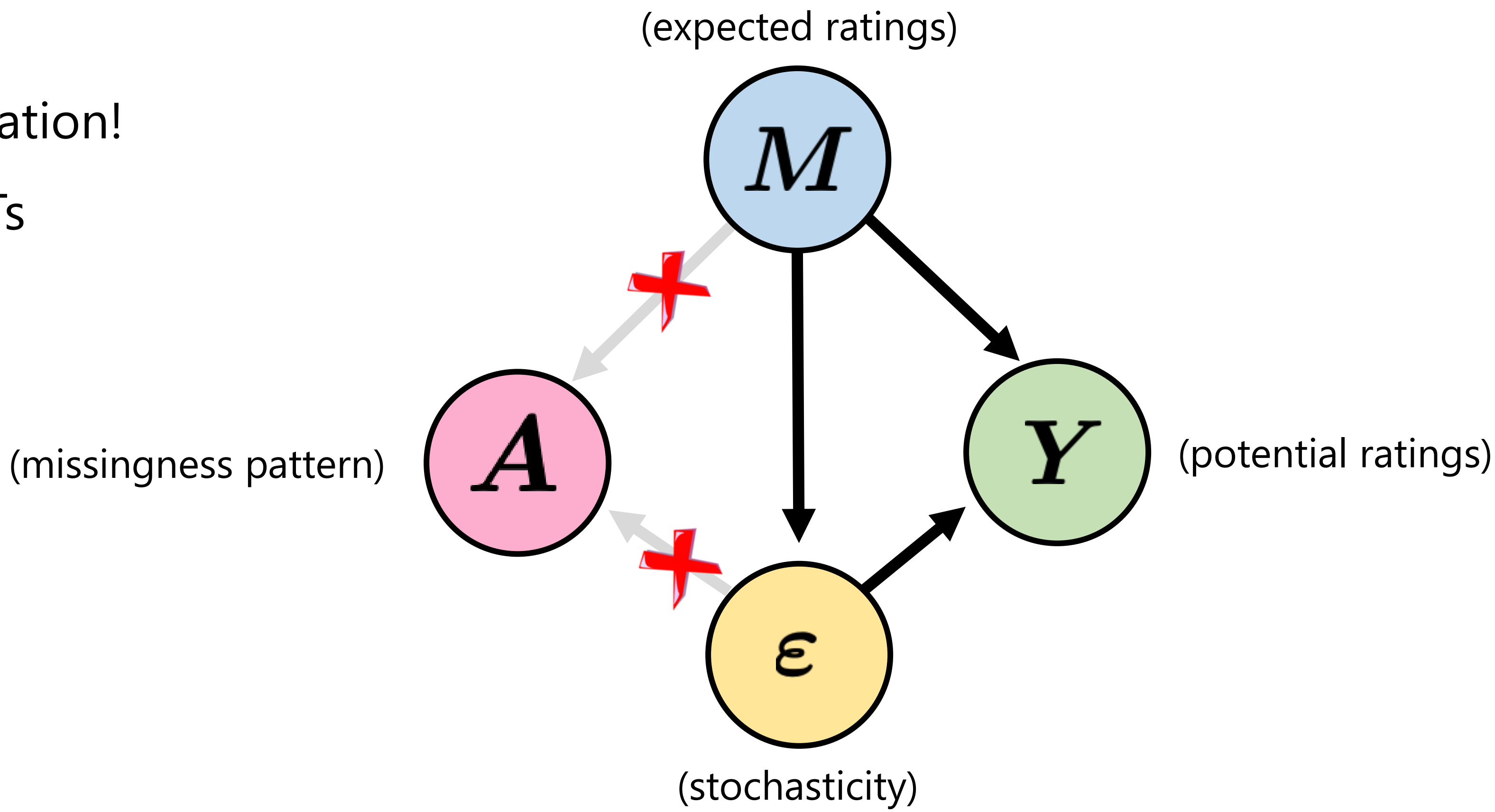
Missing at Random (MAR)

↓

Missing Not at Random (MNAR)

Missing completely at random (MCAR)

- A_{ij} i.i.d. samples from $\mathrm{Bernoulli}(\rho)$
- $\rho > 0$
- $\boldsymbol{A} \perp \boldsymbol{Y}$
  - correlation = causation!
  - e.g., A/B tests, RCTs



A brief history

Missing Completely at Random (MCAR)

↓

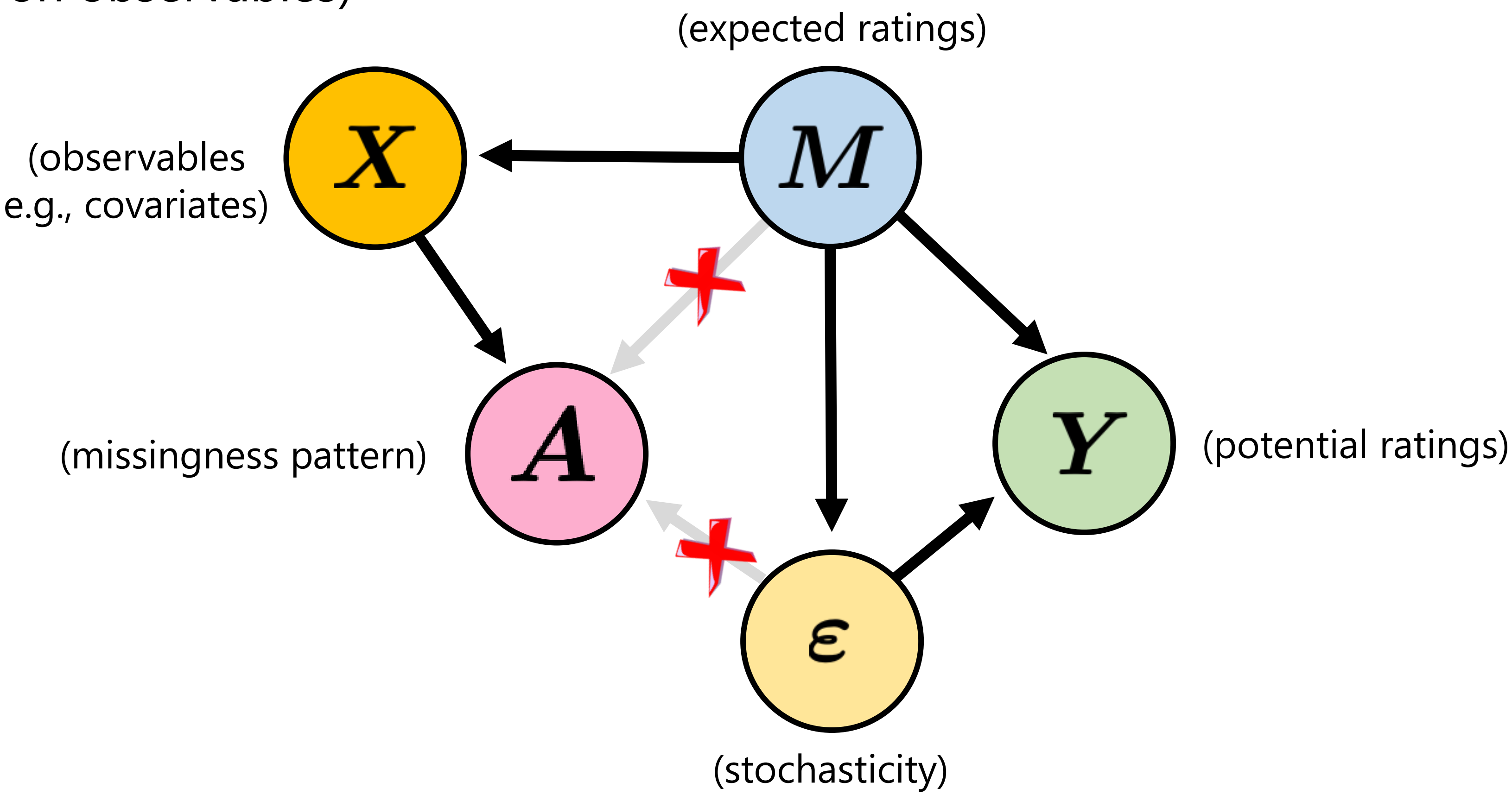
Missing at Random (MAR)

↓

Missing Not at Random (MNAR)

Missing at random (MAR)

- A_{ij} independent sample from $\mathrm{Bernoulli}(\rho_{ij})$
- $\rho_{ij} > 0$
- $\boldsymbol{A} \perp \boldsymbol{Y} \mid \boldsymbol{X}$ (selection on observables)



A brief history

Missing Completely at Random (MCAR)

↓

Missing at Random (MAR)

↓

Missing Not at Random (MNAR)

Missing not at random (MNAR)

Not MCAR nor MAR…

An overview of matrix completion algorithms

- Spectral methods
- Optimization based methods
- Nearest neighbors or collaborative filtering

Spectral methods

- Estimate $\widehat{\rho}_{ij}$ (e.g., logistic regression)

- Replace missing entries with 0

- $\boldsymbol{Y}_0 = \sum_i s_i u_i v_i^T$

- $\widehat{\boldsymbol{Q}} = \sum_{i=1}^{k} s_i u_i v_i^T$

- $\widehat{M}_{ij} = (1/\widehat{\rho}_{ij}) \cdot \widehat{Q}_{ij}$

- [Gavish-Donoho '14, Chatterjee '15, Bhattacharya-Chatterjee '21, ...]

Optimization based methods

- Estimate $\widehat{\rho}_{ij}$ (e.g., logistic regression)

- $$\widehat{\boldsymbol{M}} = \operatorname{argmin} \sum_{(i,j):A_{ij}=1} (1/\widehat{\rho}_{ij}) \cdot \operatorname{dist}(Q_{ij}, \tilde{Y}_{ij}) + \lambda \cdot \operatorname{regularize}(\boldsymbol{Q})$$

  - Example: $$\widehat{\boldsymbol{M}} = \operatorname{argmin} \sum_{(i,j):A_{ij}=1} (1/\widehat{\rho}_{ij}) \cdot (Q_{ij} - \tilde{Y}_{ij})^2 + \lambda \cdot \|\boldsymbol{Q}\|_*$$

$$\|\boldsymbol{Q}\|_* = \min_{\boldsymbol{U},\boldsymbol{V}:\boldsymbol{Q}=\boldsymbol{U}\boldsymbol{V}^T} \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$$

[Mazumder et al. '10]

- [Candes-Tao '10, Keshavan et al. '10, Mazumder et al. '10, Recht '11, Hastie et al. '15, ...]

Estimating propensities without covariates

- Assumption
  - $\boldsymbol{\rho} = [\rho_{ij}]$ is "nice" (e.g., low rank)

- Algorithm
  - Run matrix completion on $\boldsymbol{A}$ (fully observed) to yield $\widehat{\boldsymbol{\rho}} = [\widehat{\rho}_{ij}]$

- [Ma-Chen '19, Wang et al. '20, Bhattacharya-Chatterjee '21, ...]
  - Relative performance guarantees, Ma-Chen '19
  - Consistency, Bhattacharya-Chatterjee '21

Nearest neighbors aka collaborative filtering

- Find k "nearest neighbors" then average
    - Cosine similarity
    - Euclidean distance
    - Manhattan distance
    - and much much more...
- [Goldberg '92, Linden '03, Kleinberg '08, Koren '15, Lee et al. '18, '20, ...]

MCAR scorecard

- All wrt Frobenius norm except for nearest neighbors wrt l_∞

Algorithm	References	Function Class	Noise Model	Guaranteed Recovery	Observations mnp (m=n)
USVT	[Chatterjee]	Lipschitz	Arbitrary	Approx.	$n^{\frac{2r+2}{r+2}} \log^6 n$
USVT	[Chatterjee]	Low-rank	Arbitrary	Approx.	$nr \log^6 n$
Convex	[Recht]	Low-rank	No Noise	Exact	$nr \log^2(n)$
Convex	[CandesPlan]	Low-rank	Additive	Approx.	$nr \log^2(n)$
Near Nghbr	[LeeLiSoSh]	Lipschitz	Additive	Approx.	$n^{\frac{3}{2}} \text{polylog} n$
Near Nghbr	[BoChLeeSh]	Low-rank	Arbitrary	Approx.	$nr^5 \omega(1)$
Non-Convex	[KeMonOh]	Low-rank	No Noise	Exact	$nr \log n$
Non-Convex	[KeMonOh]	Low-rank	Additive	Approx.	$nr \log n$

Open question

[Ma-Chen '19]

"In terms of theoretical analysis, we have not addressed the full generality of MNAR data.

Our theory breaks down when probability of observation is exactly 0.

We still assume that each entry is revealed independent of other entries.

These are two open problems among many for robustly handling MNAR data with guarantees."

+ Few formal results for entry-wise recovery of matrix

Open question

[Ma-Chen '19]

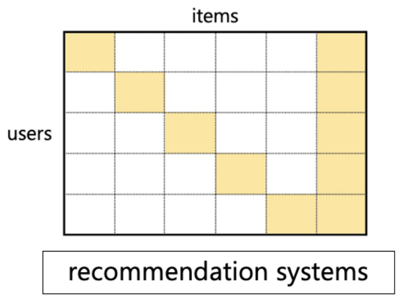
"In terms of theoretical analysis, we have
not addressed the full generality of MNAR data.

Our theory breaks down when
probability of observation is exactly 0.

We still assume that each entry is
revealed independent of other entries.

These are two open problems among many
for robustly handling MNAR data with guarantees."

+ Few formal results for entry-wise recovery of matrix



e.g., a vegetarian will **never** go to steakhouse

Open question

[Ma-Chen '19]

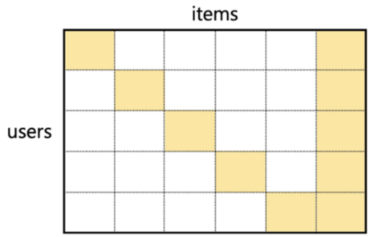
"In terms of theoretical analysis, we have not addressed the full generality of MNAR data.

Our theory breaks down when probability of observation is exactly 0.

We still assume that each entry is revealed independent of other entries.

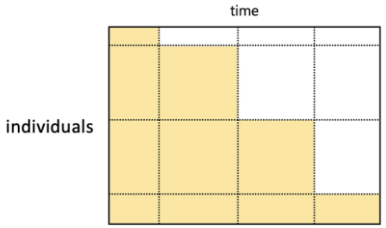
These are **two open problems among many** for robustly handling MNAR data with guarantees."

+ Few formal results for **entry-wise** recovery of matrix



recommendation systems

e.g., a vegetarian will **never** go to steakhouse



staggered adoption

e.g., if data missing at **time t**, then missing at **time t+1**

Open question

[Ma-Chen ‘19]

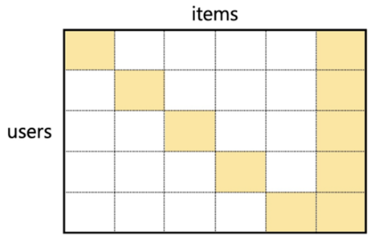
“In terms of theoretical analysis, we have not addressed the full generality of MNAR data.

Our theory breaks down when probability of observation is exactly 0.

We still assume that each entry is revealed independent of other entries.

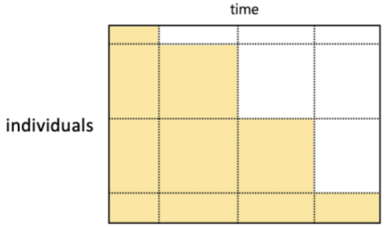
These are two open problems among many for robustly handling MNAR data with guarantees.”

+ Few formal results for entry-wise recovery of matrix



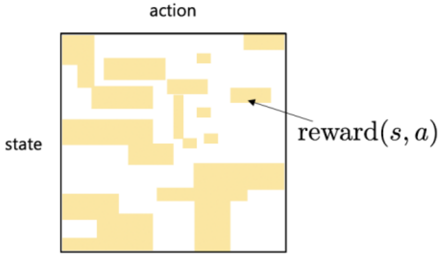
recommendation systems

e.g., a vegetarian will **never** go to steakhouse



staggered adoption

e.g., if data missing at **time t**, then missing at **time t+1**



contextual bandits

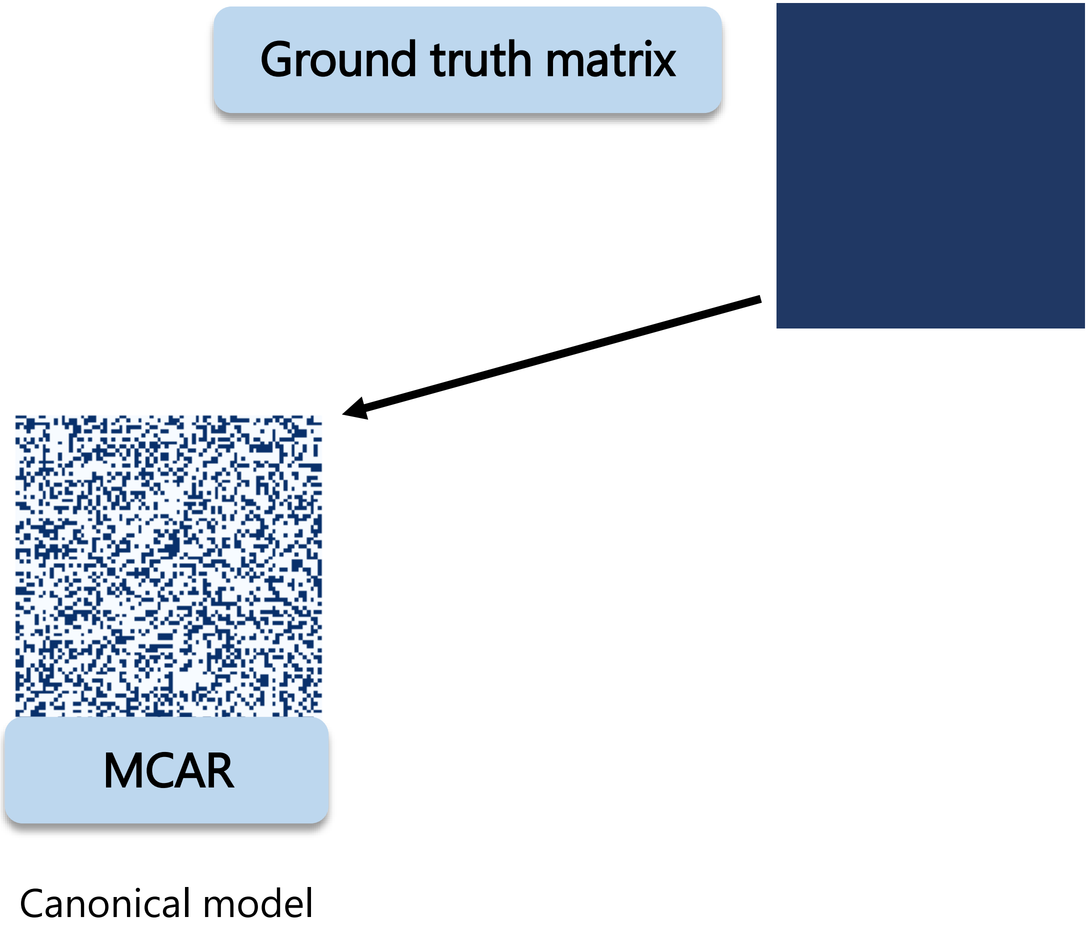
e.g., (state, action) visited **confounded with** expected reward

But…does the missingness mechanism matter?

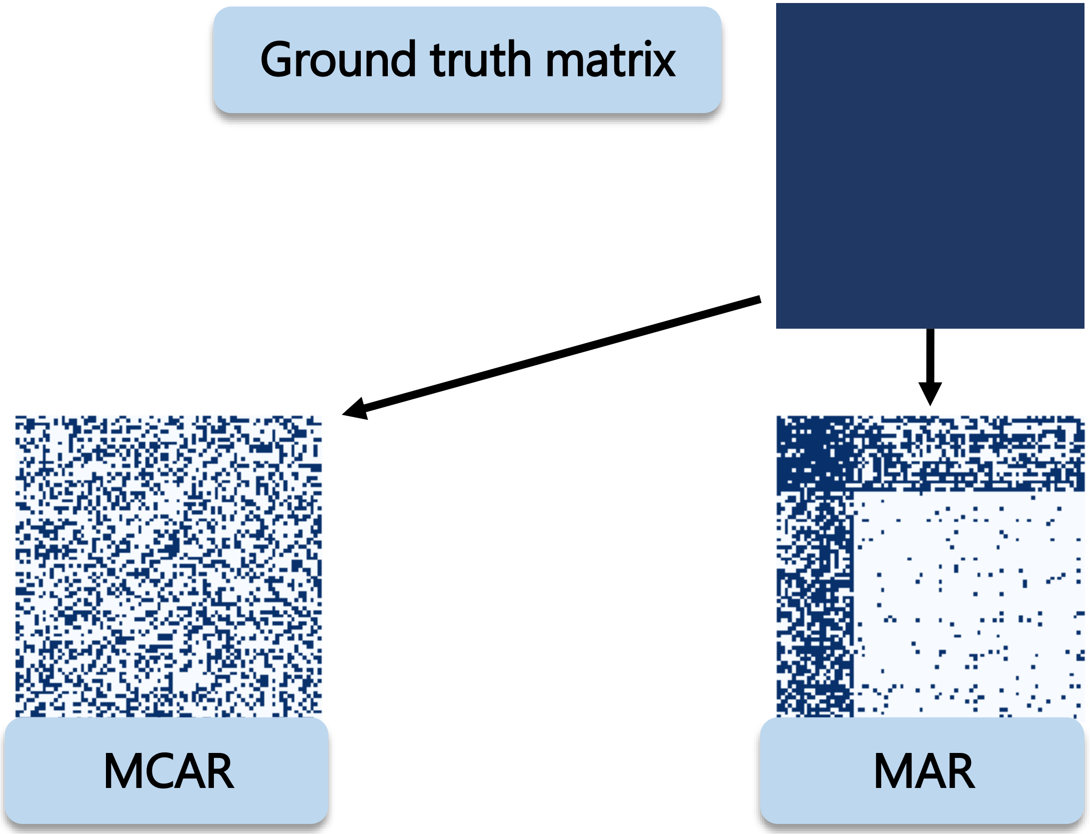
A toy illustration

Ground truth matrix

A toy illustration



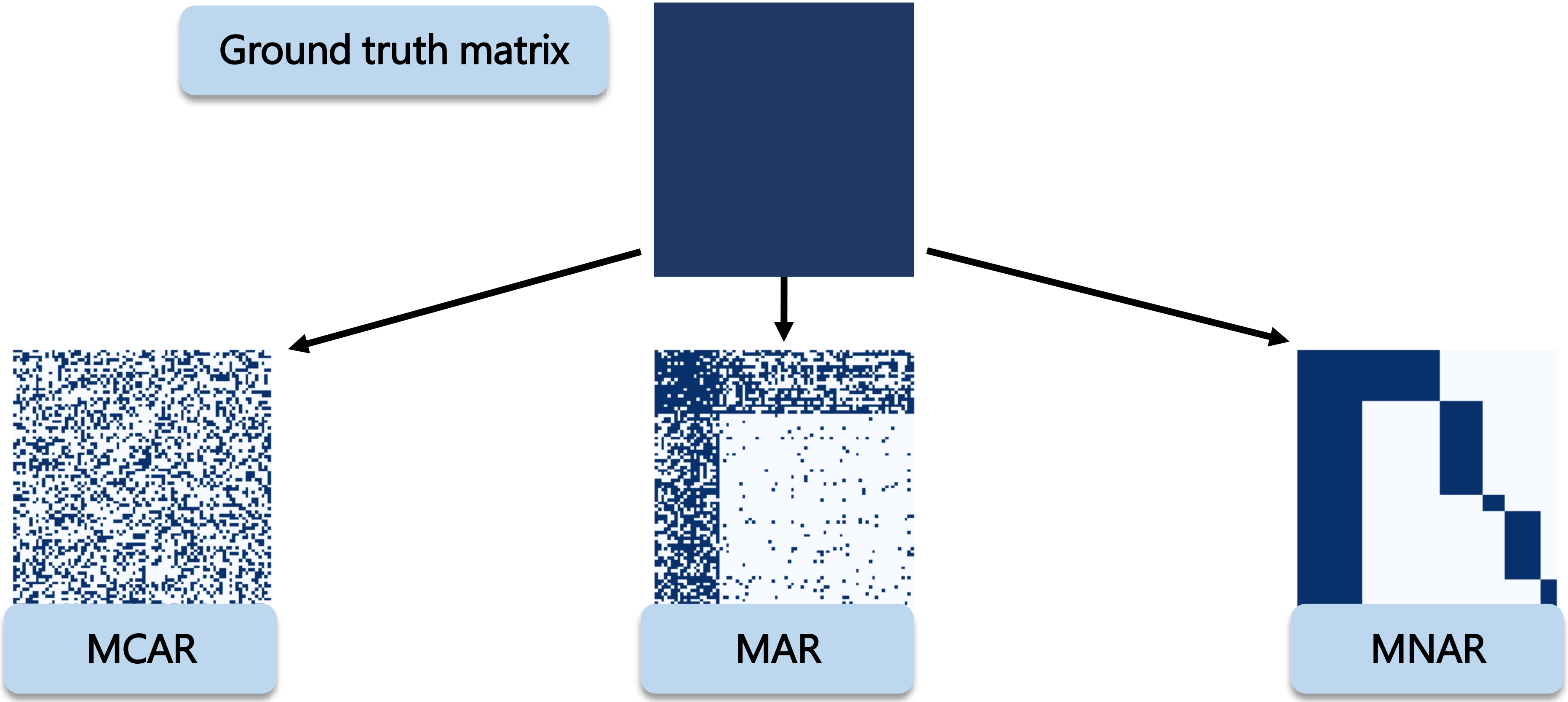
A toy illustration



Canonical model

"State-of-the-art" model

A toy illustration



Canonical model

"State-of-the-art" model

$$\min_{ij} \mathbb{P}(A_{ij} = 1) = 0$$

$$A_{ij} \not\perp A_{k\ell} \qquad \boldsymbol{M} \not\perp \boldsymbol{A}$$

Open problem in [Ma-Chen'19]

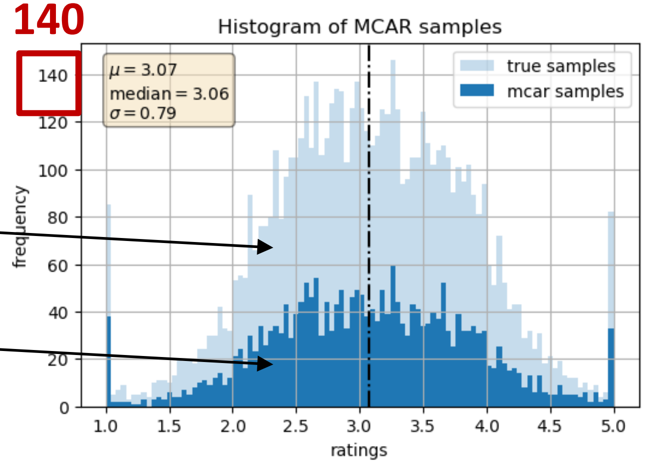
MCAR

Goal:
recover **true** from **observed**

True distribution

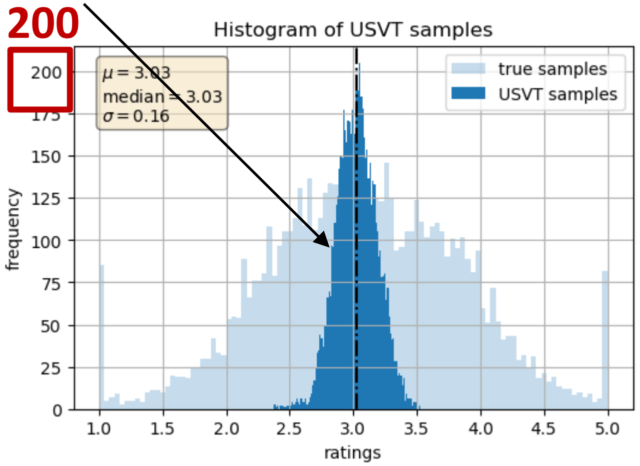
Observed MCAR samples

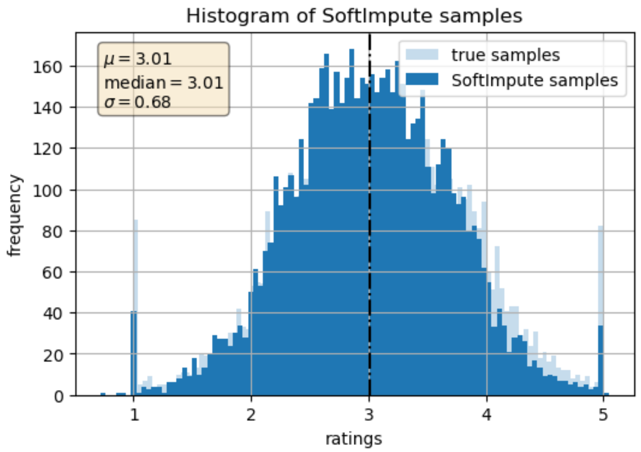
140

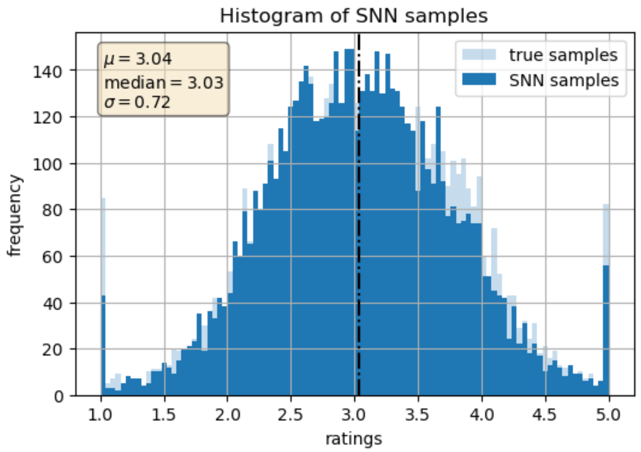


Recovered distribution

200







Universal singular value thresholding [Chatterjee '15]

spectral method

Softimpute [Hastie et al. '14]

optimization

Synthetic nearest neighbor

Our approach

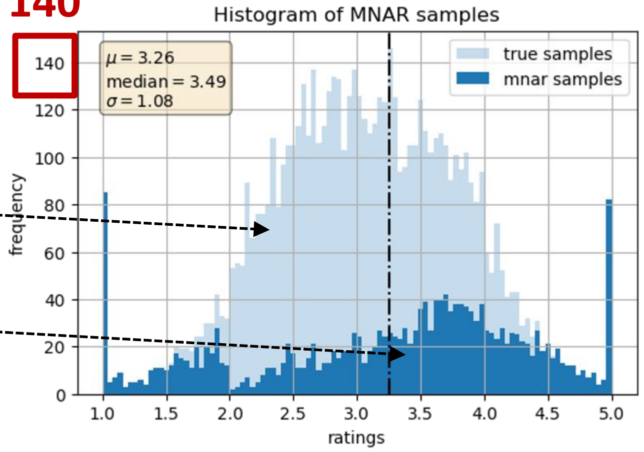
MAR

Goal:
recover **true** from **observed**

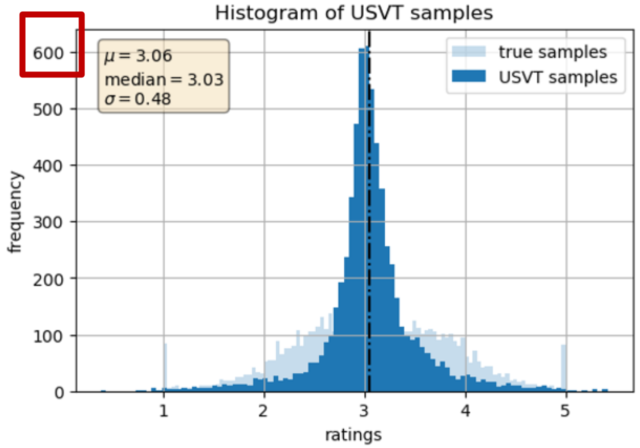
True distribution

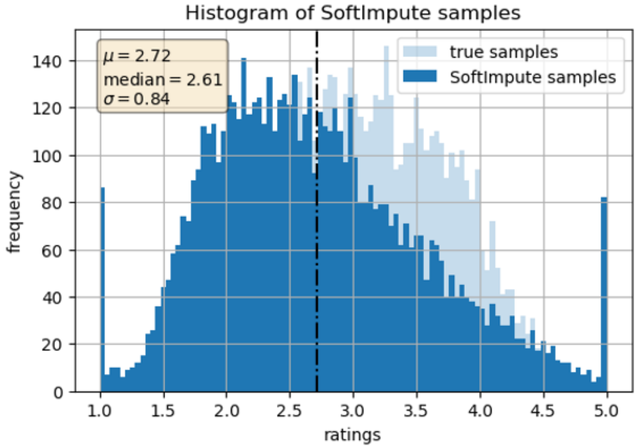
Observed L-MNAR samples

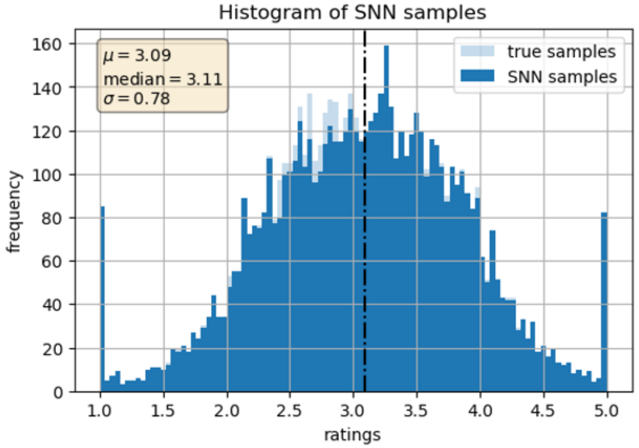
140



600







Universal singular value thresholding [Chatterjee '15]

Softimpute [Hastie et al. '14]

Synthetic nearest neighbor

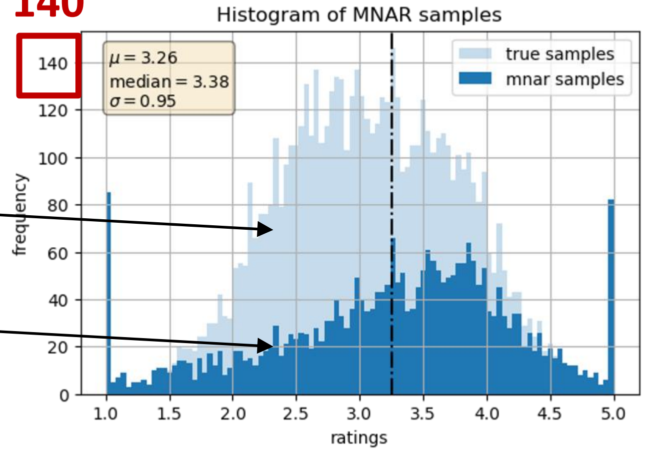
MNAR

Goal:
recover **true** from **observed**

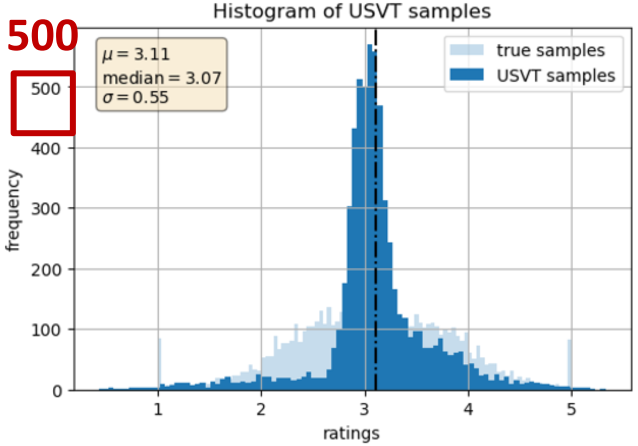
True distribution

Observed MNAR samples

140

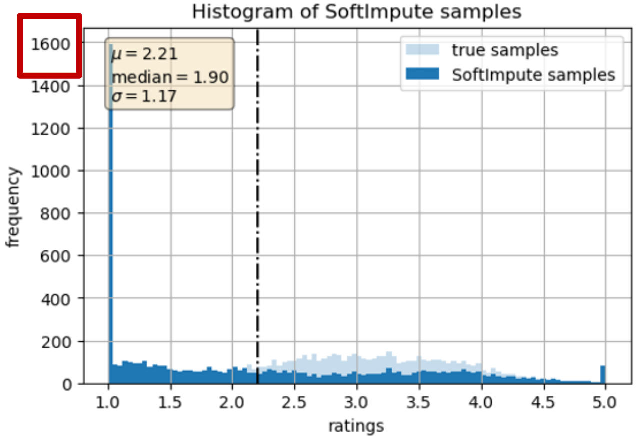


500

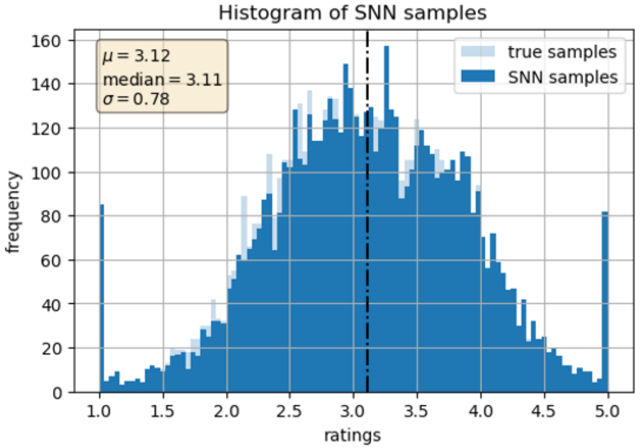


Universal singular value thresholding [Chatterjee '15]

1600!



Softimpute [Hastie et al. '14]



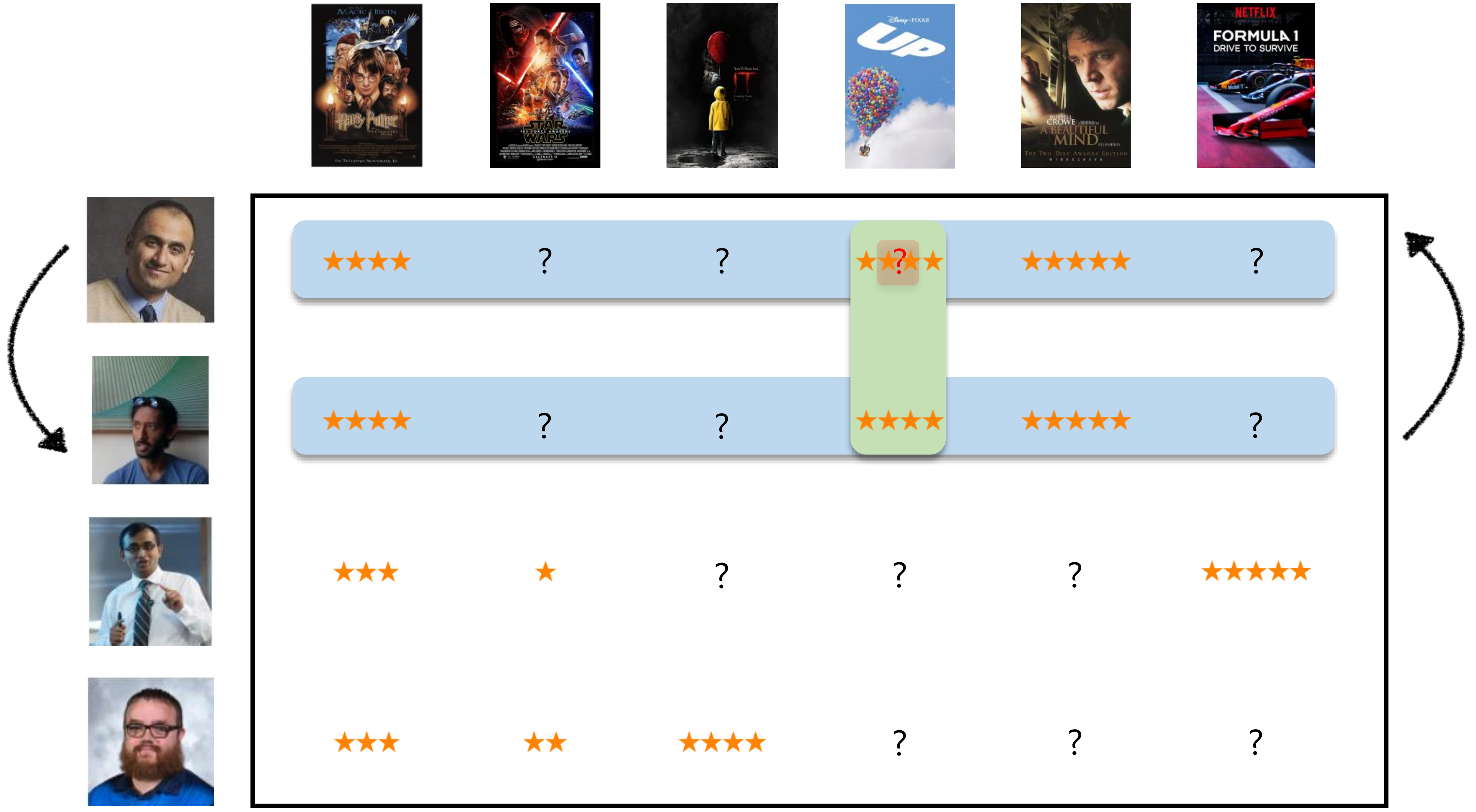
Synthetic nearest neighbor

MNAR data is abundant

- Recommendation systems
  - Movies, products, news articles…
- Clinical trials
  - ~35% dropout rate
- U.S. census
  - ~40% data missing

Synthetic nearest neighbors (SNN)

Nearest neighbors (NN)

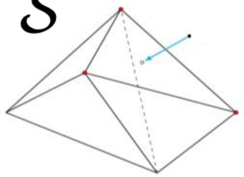


Synthetic controls (SC)

"What if California never passed Prop 99?"

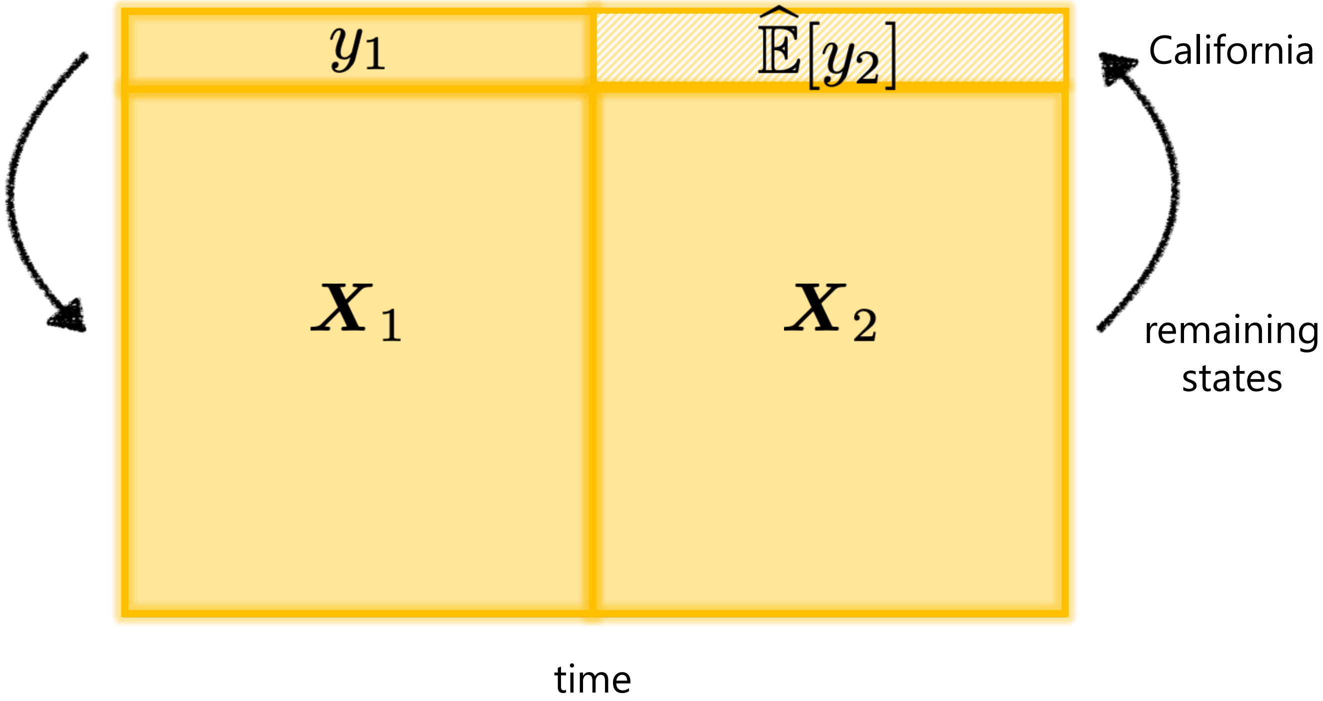
Model learning:

$$\widehat{\beta} = \underset{\beta \in \mathcal{S}}{\operatorname{argmin}} \|y_1 - \boldsymbol{X}_1 \beta\|_2^2$$



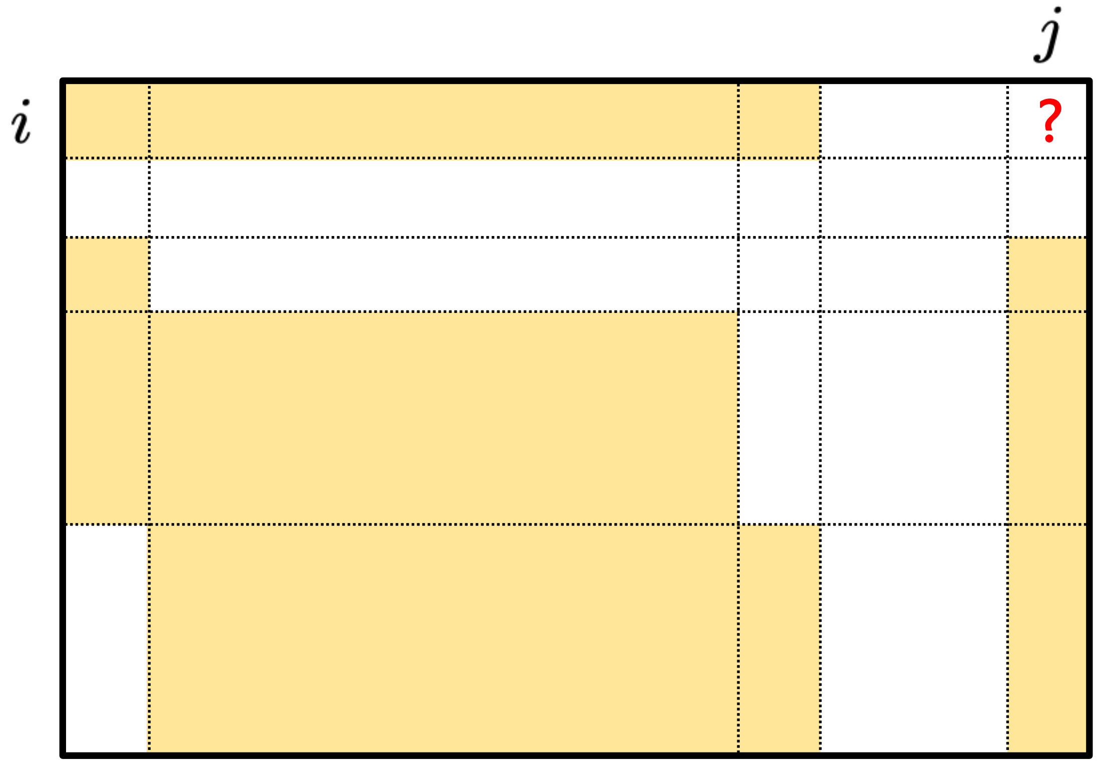
Counterfactual prediction:

$$\widehat{\mathbb{E}}[y_2] = \boldsymbol{X}_2 \widehat{\beta}$$

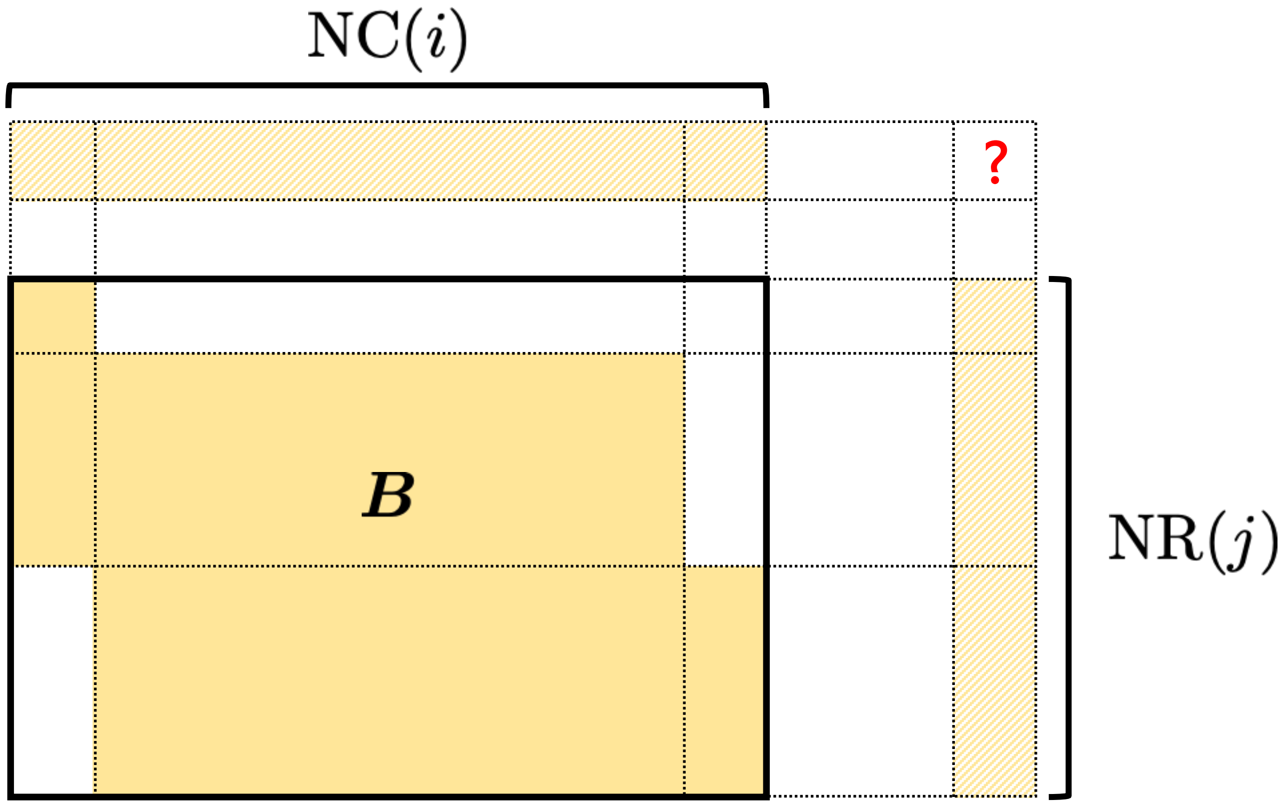


[Abadie '03, '10, '15, ...]

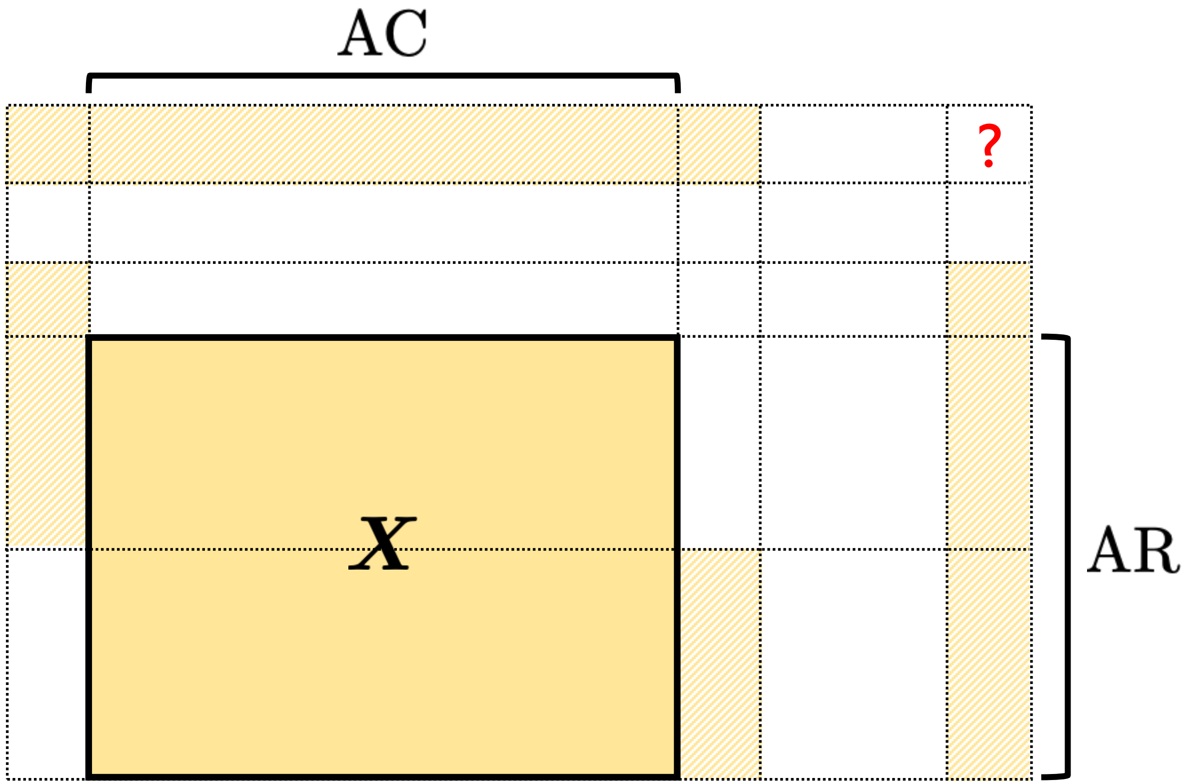
Synthetic nearest neighbors (SNN): NN meets SC



Step 1: Academic "Sudoku"



Step 1: Academic "Sudoku"

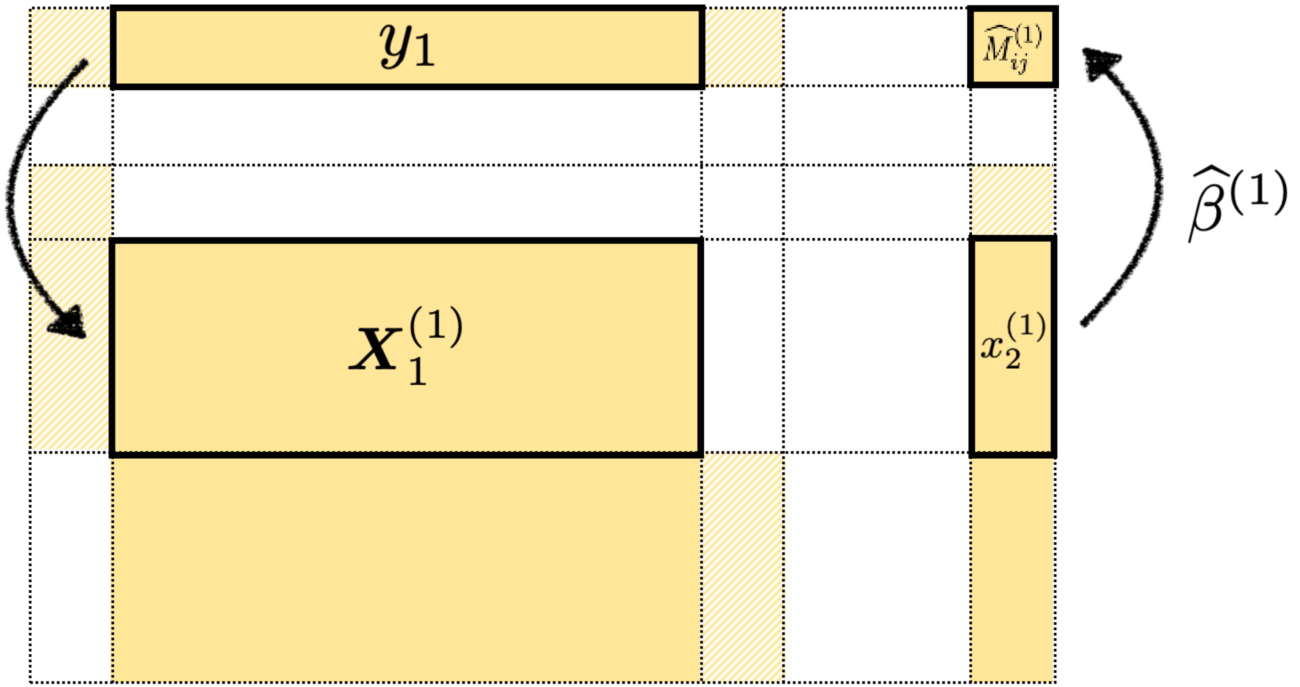


Max biclique

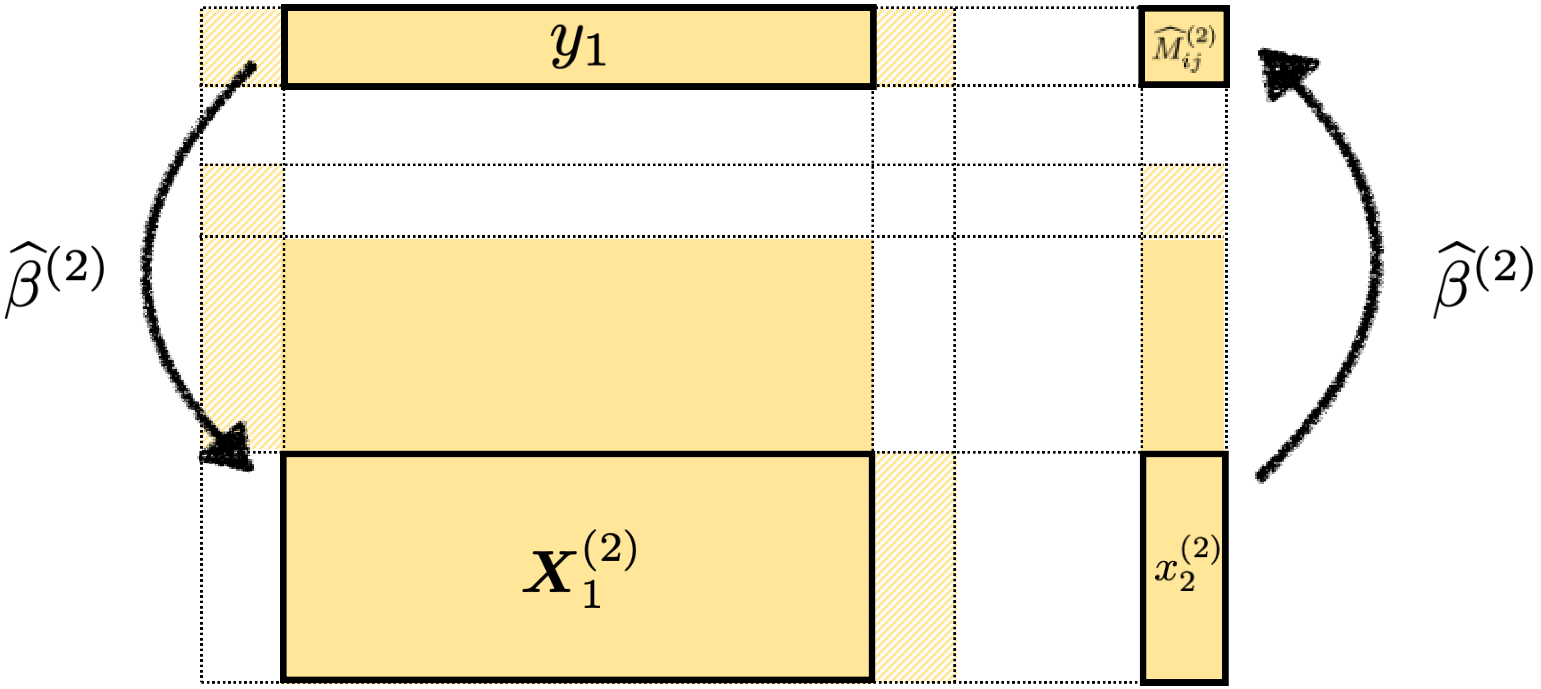
Algorithms: [Alexe '03; Zhang '14; Lyu '20; Lu '20]

Step 2: Create synthetic neighborhoods

$$\widehat{\beta}^{(1)} = \mathrm{PCR}(\boldsymbol{X}_1^{(1)}, y_1)$$



Step 3: Average

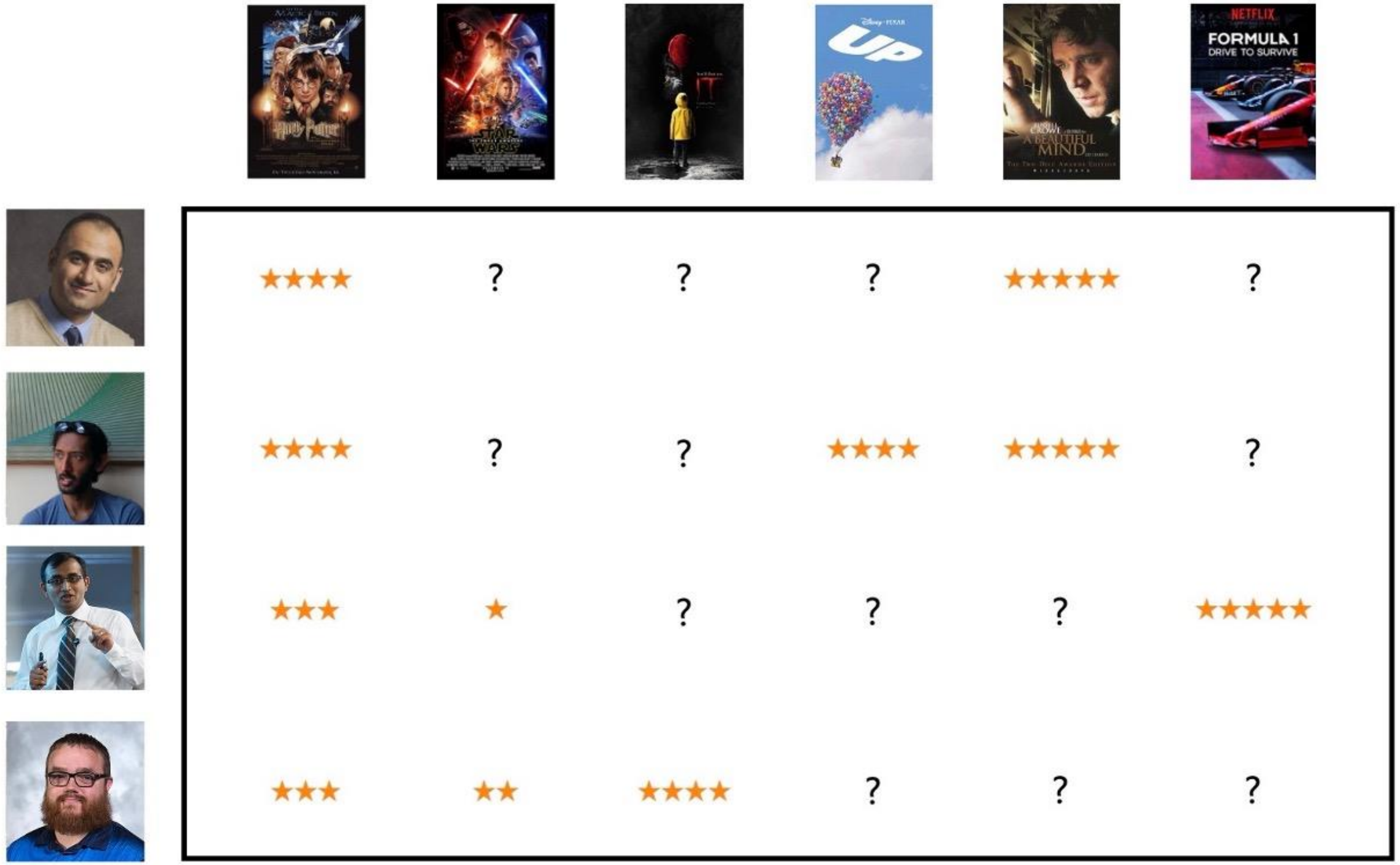


$$\widehat{M}_{ij} = \frac{1}{k}\sum_{\ell=1}^{k}\widehat{M}_{ij}^{(\ell)}$$

When does SNN work?

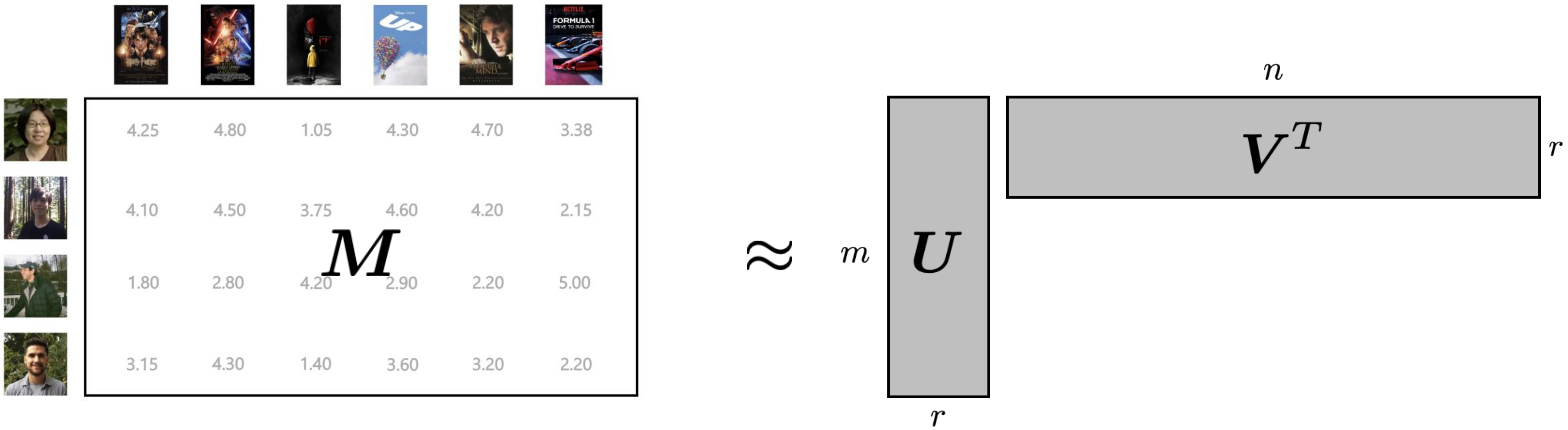
- Why linear model?
- What class of missingness models?

In general, we cannot infer missing entries…



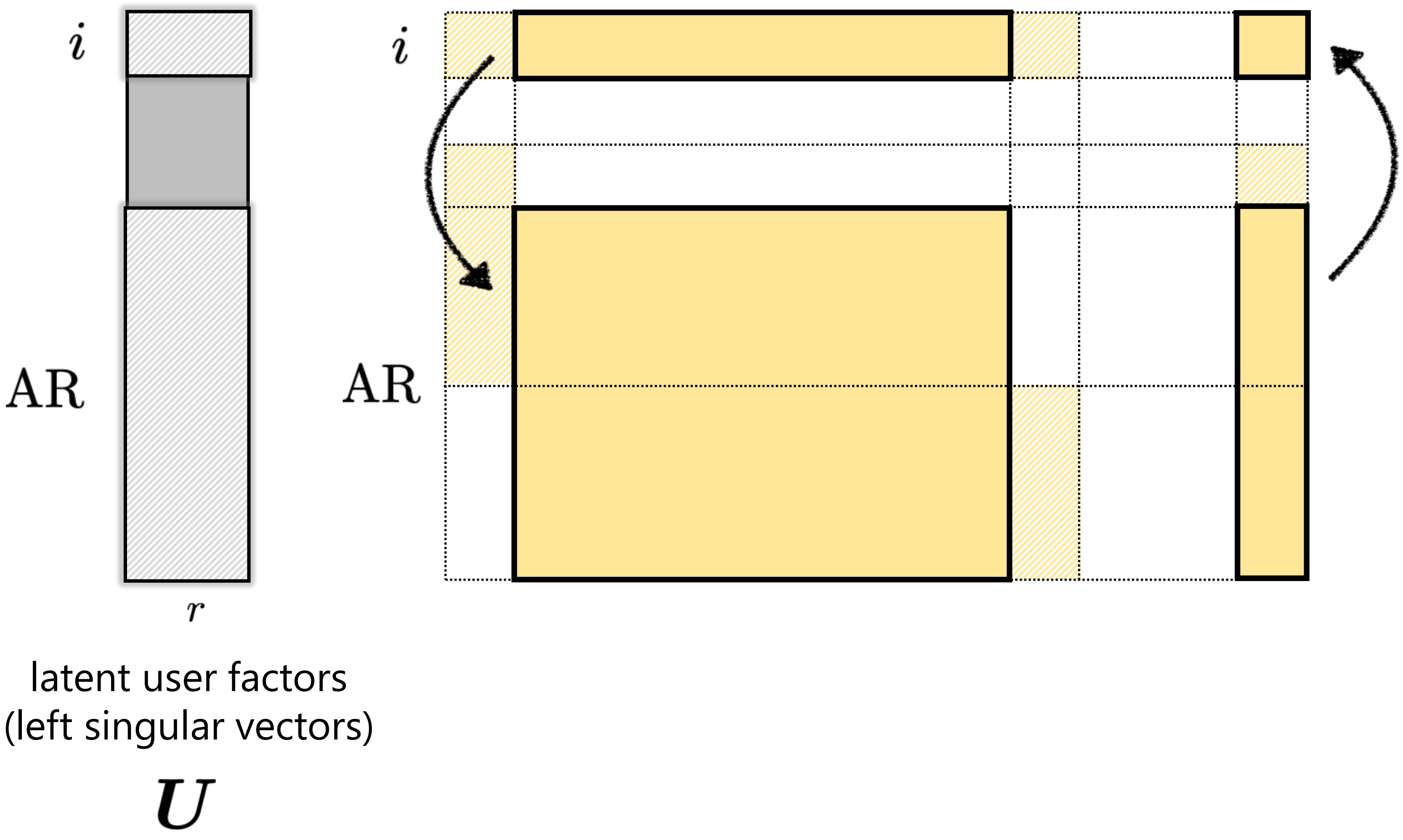
Undetermined system: more unknowns than observations

…unless underlying ratings have additional structure

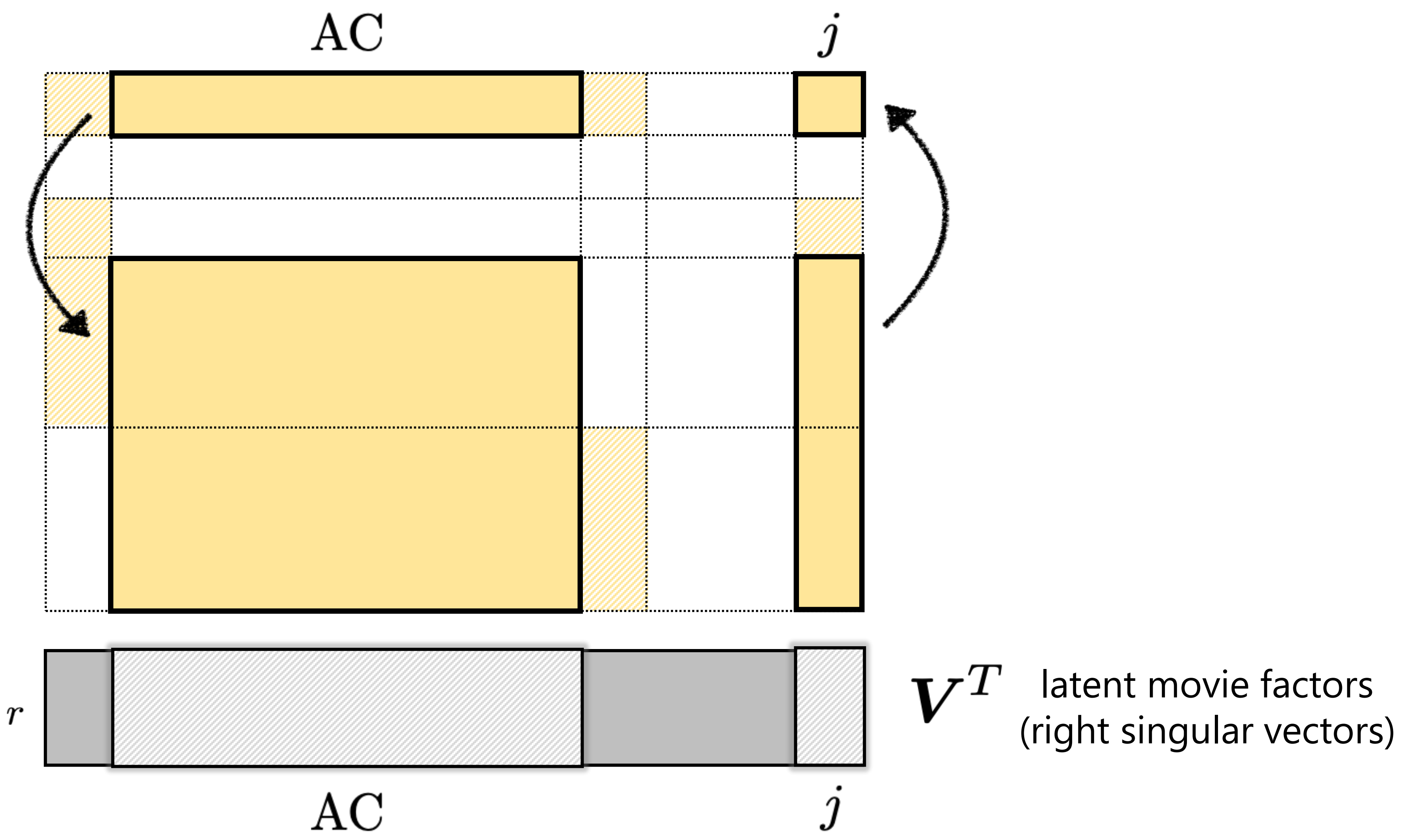


a **few** factors explain **most** of the data → **low-rank** approximation $(r \ll \min\{m, n\})$

Low rank implies linear model (across users)

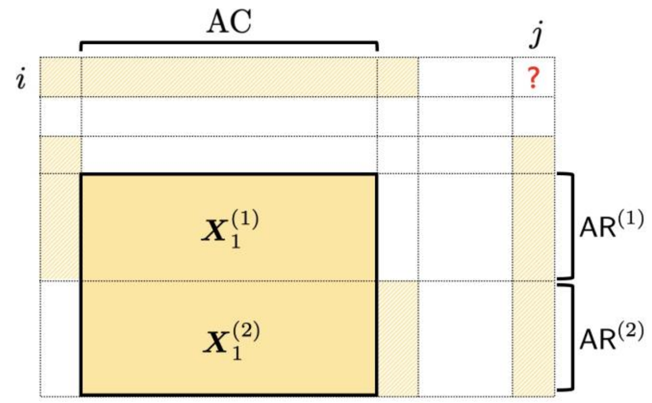


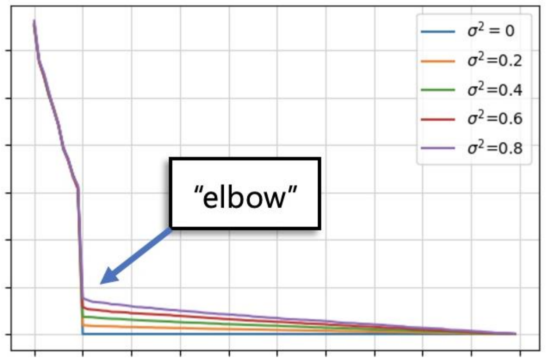
Low rank implies subspace inclusion (across movies)



Well-balanced spectra

Separation of signal & noise





$$s_{\min}(\mathbb{E}[\boldsymbol{X}_1^{(\ell)}]) \gg s_{\max}(\boldsymbol{\varepsilon}_1^{(\ell)}) \text{ for all } \ell = 1, \ldots, k$$

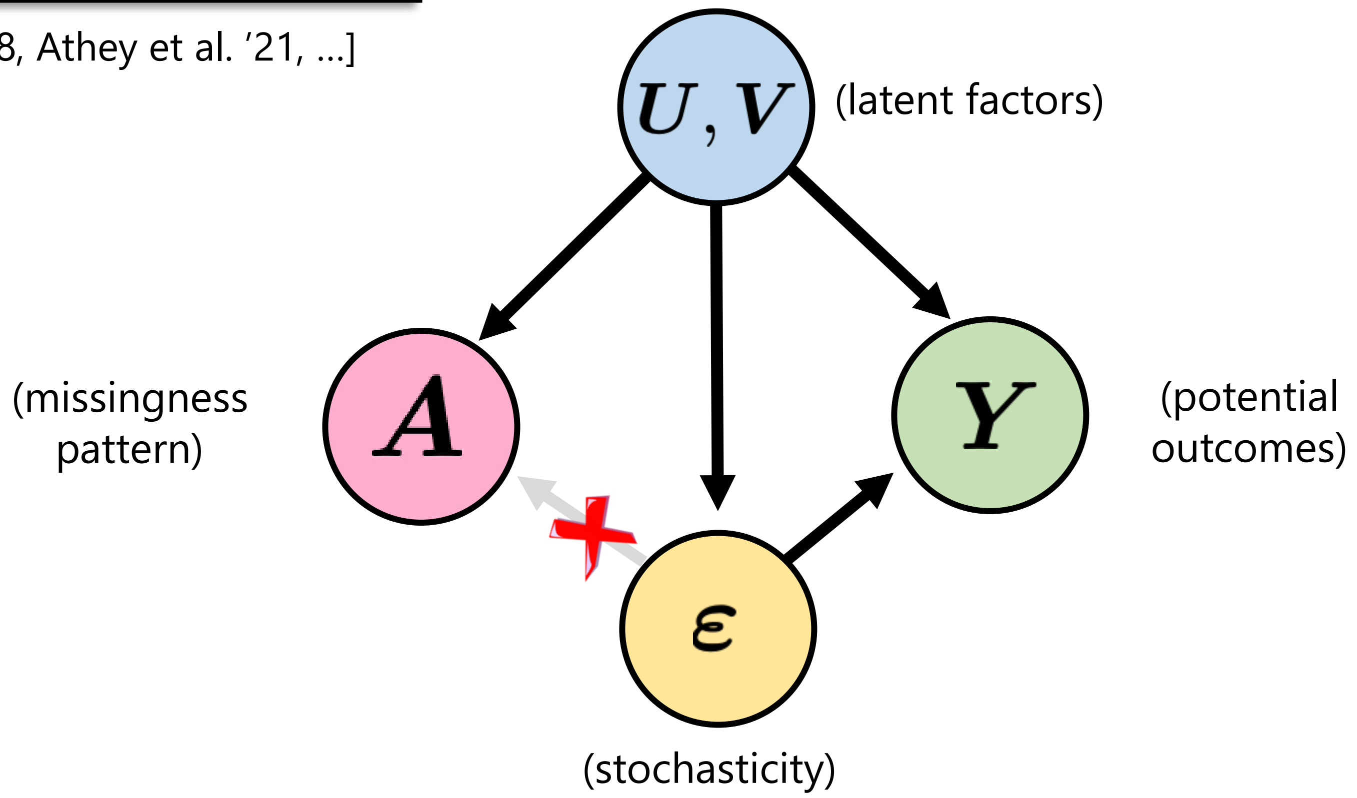
[Chamberlain '83; Fan '18; Bai '19; Cai '21]

What type of missingness is allowed?

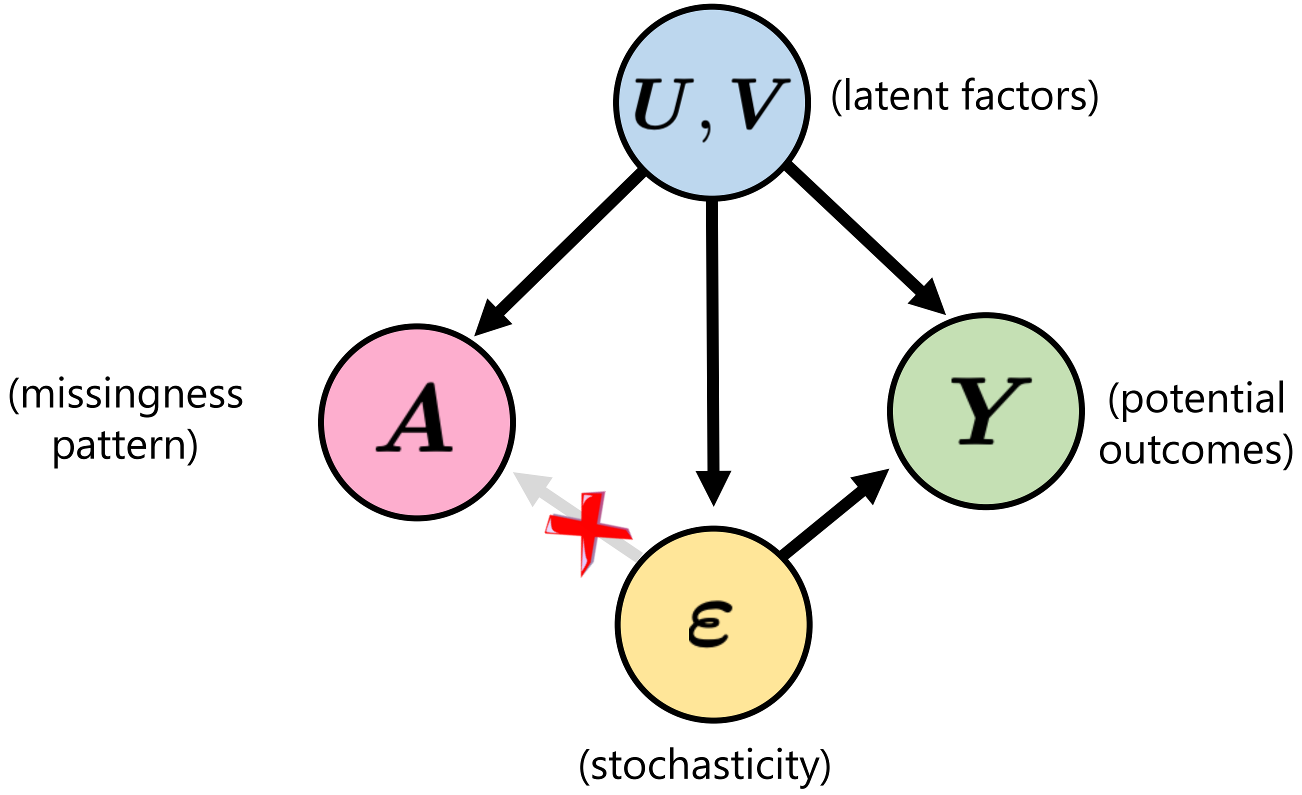
$$Y_{ij} = \langle u_i, v_j \rangle + \varepsilon_{ij}$$

$$+ \quad \boxed{A \perp\!\!\!\perp \varepsilon \mid U, V} \quad \Longrightarrow \quad \boxed{A \perp\!\!\!\perp Y \mid U, V}$$

[Kallus '18, Athey et al. '21, ...]



What this model allows for



$$\min_{ij} \mathbb{P}(A_{ij} = 1) = 0$$

$$A_{ij} \not\perp\!\!\!\perp A_{k\ell}$$

$$\boldsymbol{M} \not\perp\!\!\!\perp \boldsymbol{A}$$

Formal guarantees

Theorems (informal)

- Consistency: $\widehat{M}_{ij} - M_{ij} = o(1)$
- Asymptotic normality: $\widehat{M}_{ij} - M_{ij} \sim \mathcal{N}(0, \sigma^2)$

Entry-wise guarantees

Implications for experimental design

A motivating example from drug design

In-Vitro Studies

Clinical Trials

Inefficient

slow and rigid → 10+ years

Costly

average $1.3 billion per drug development, ethical concerns

High Failure Rate

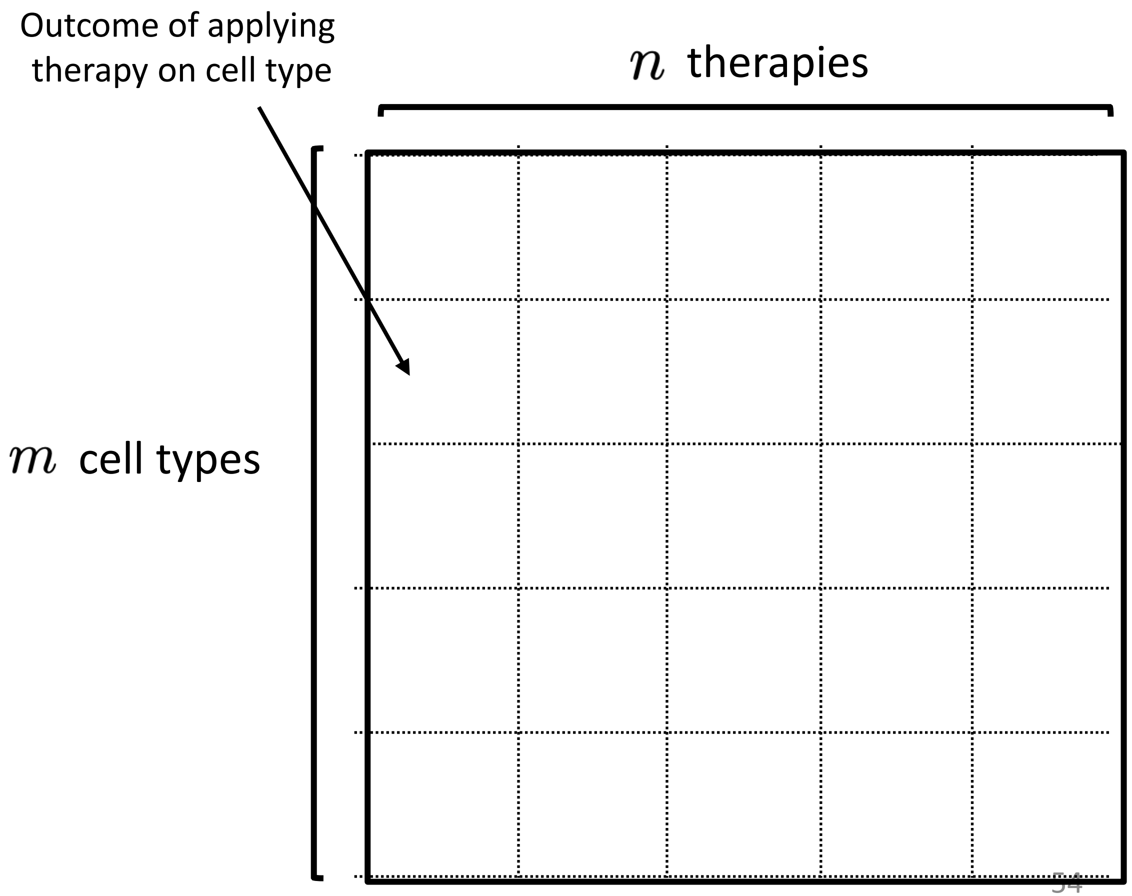
Phase II – 67% failure, Phase III – 75% failure

Can we identify the most promising therapies with a limited experimental budget?

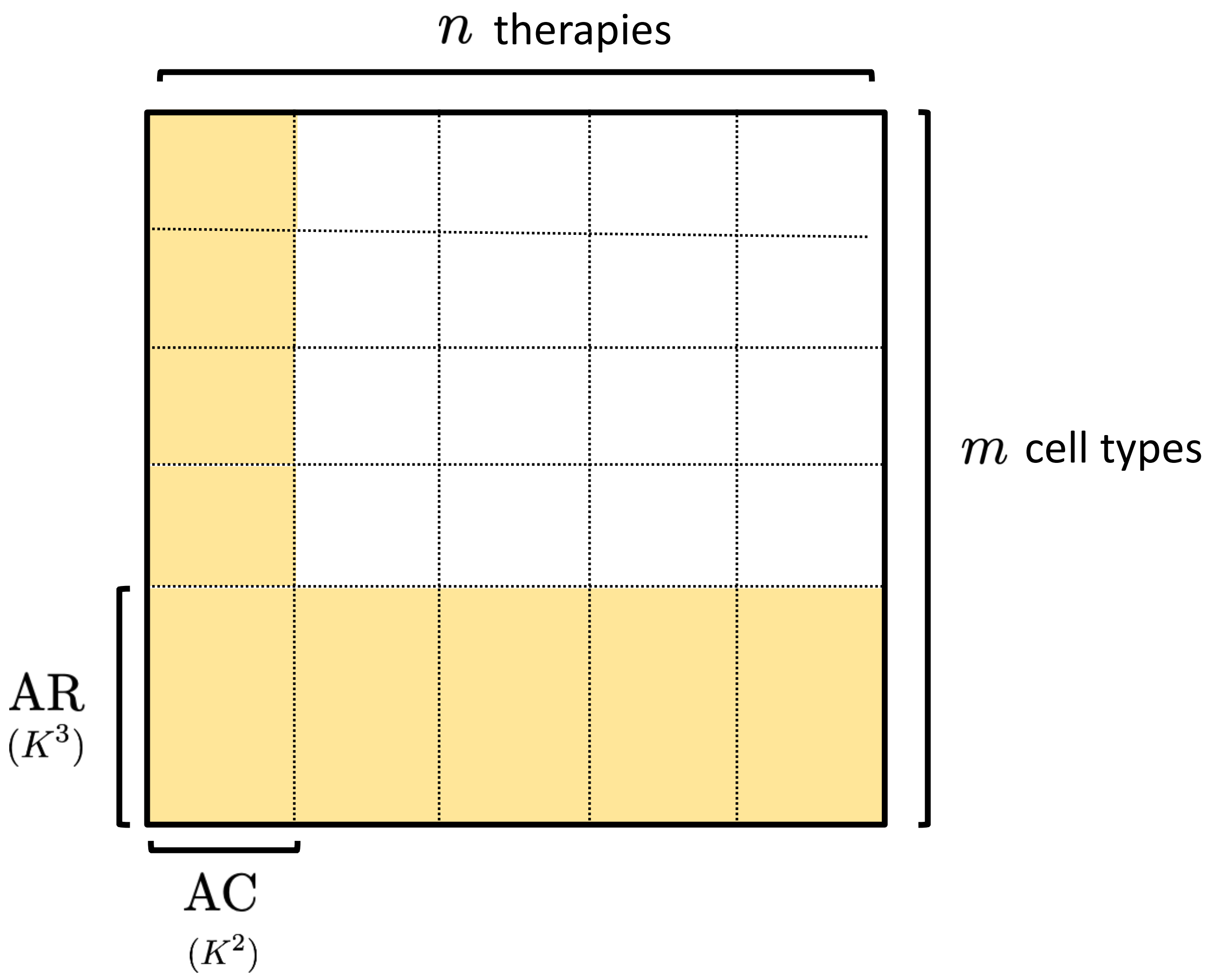
A matrix completion perspective on experimental design

Synthetic drug design: m cell types and n therapies

- **Goal**
  - Recover outcomes of all m x n experiments
- **Constraint**
  - Budget of O(m + n) experiments



"Optimal" observation pattern



Implication of causal matrix completion result

Recover all $m \times n$ experiments
with entry-wise error of $\sim 1/\sqrt{k}$
from $(m \times k^2) + (n \times k^3)$ experiments

Future question:

Optimal (adaptive) experimental design?

Towards heteroskedastic variance estimation

Towards heteroskedastic variance estimation

Confidence intervals require estimates of $\sigma_{ij}^2 = \mathbb{E}[\varepsilon_{ij}^2]$

In general, cannot recover under arbitrary heteroskedastic setting…

Assume σ_{ij}^2 are low-rank too!

Observe: $\mathbb{E}[Y_{ij}^2] = \mathbb{E}[(M_{ij} + \varepsilon_{ij})^2] = M_{ij}^2 + \sigma_{ij}^2$

Define: $\boldsymbol{\Sigma} = [\sigma_{ij}^2]$

$\boldsymbol{S} = [\mathbb{E}[Y_{ij}^2]]$

$\Longrightarrow$ $\mathrm{rank}(\boldsymbol{S}) \leq (\mathrm{rank}(\boldsymbol{M}))^2 + \mathrm{rank}(\boldsymbol{\Sigma})$

Heteroskedastic variance estimation algorithm

Estimate $\mathbb{E}[Y_{ij}] = M_{ij}$

$$\widehat{M}_{ij} = \mathrm{SNN}(Y_{ij})$$

Estimate $\mathbb{E}[Y_{ij}^2] = M_{ij}^2 + \sigma_{ij}^2$

$$\widehat{M_{ij}^2 + \sigma_{ij}^2} = \mathrm{SNN}(Y_{ij}^2)$$

Estimate σ_{ij}^2

$$\widehat{\sigma}_{ij}^2 := (\widehat{M_{ij}^2 + \sigma_{ij}^2}) - (\widehat{M}_{ij})^2$$

Statistical guarantees

Suppose

$$M_{ij} - \widehat{M}_{ij} = \mathcal{O}_p(\delta_1)$$

$$(M_{ij}^2 - \sigma_{ij}^2) - (\widehat{M_{ij}^2 - \sigma_{ij}^2}) = \mathcal{O}_p(\delta_2)$$

Then,

$$\sigma_{ij}^2 - \widehat{\sigma}_{ij}^2 = \mathcal{O}_p(\max\{\delta_1, \delta_1^2\} + \delta_2)$$

An important connection

"causal inference is a missing data problem"

vis-à-vis

"matrix completion is a missing data problem"

Causal Inference	Matrix Completion
causal estimand	error metric (norm)
confounded data	missing not at random data
observational & experimental studies	sparsity patterns
estimating potential outcomes	imputing missing entries

THANK YOU

dshen24@berkeley.edu

This talk: https://arxiv.org/abs/2109.15154

Code: https://github.com/deshen24/syntheticNN