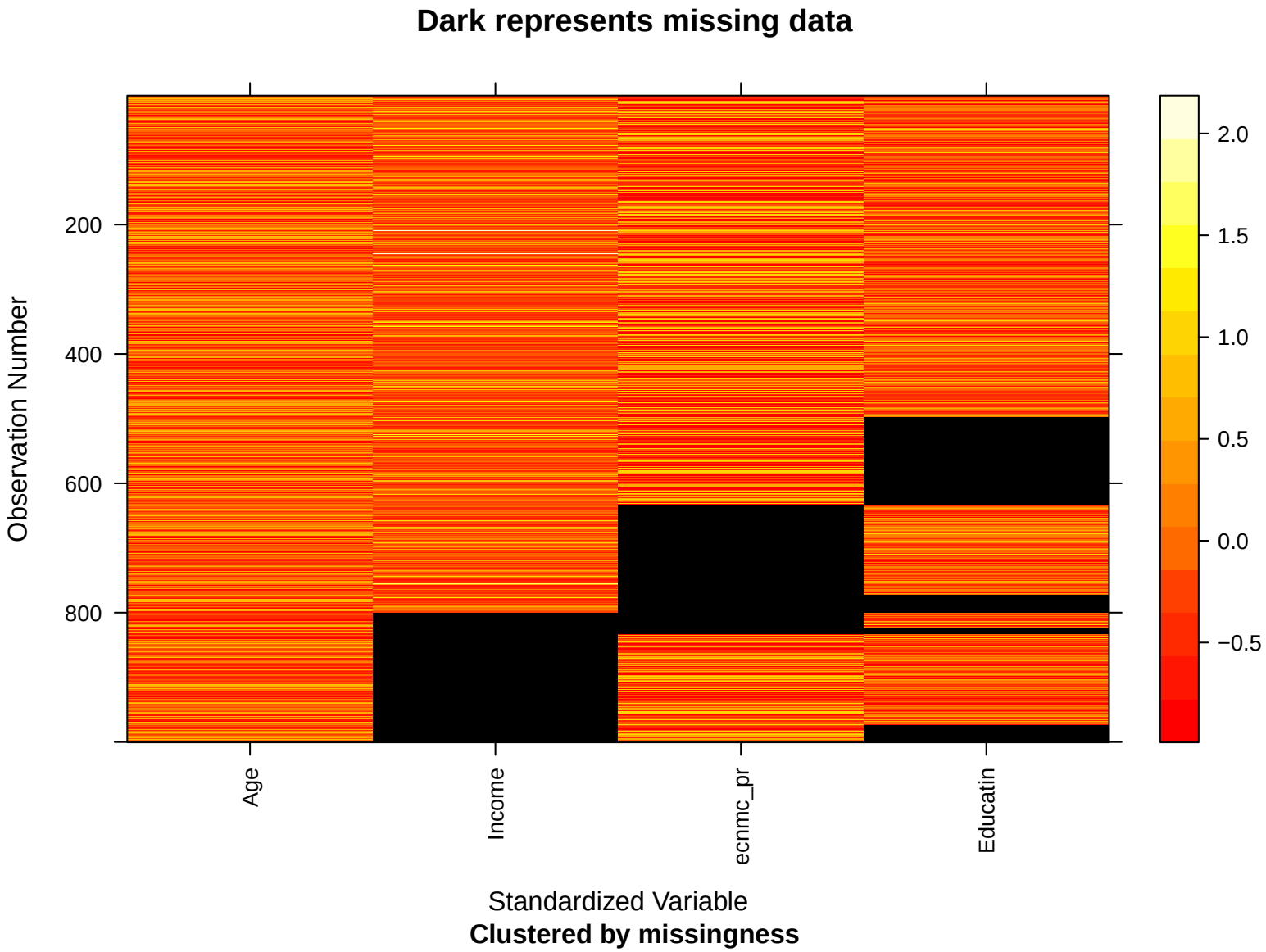
Missing Data:
Theory and Practice

Marco Morales
marco.morales@columbia.edu

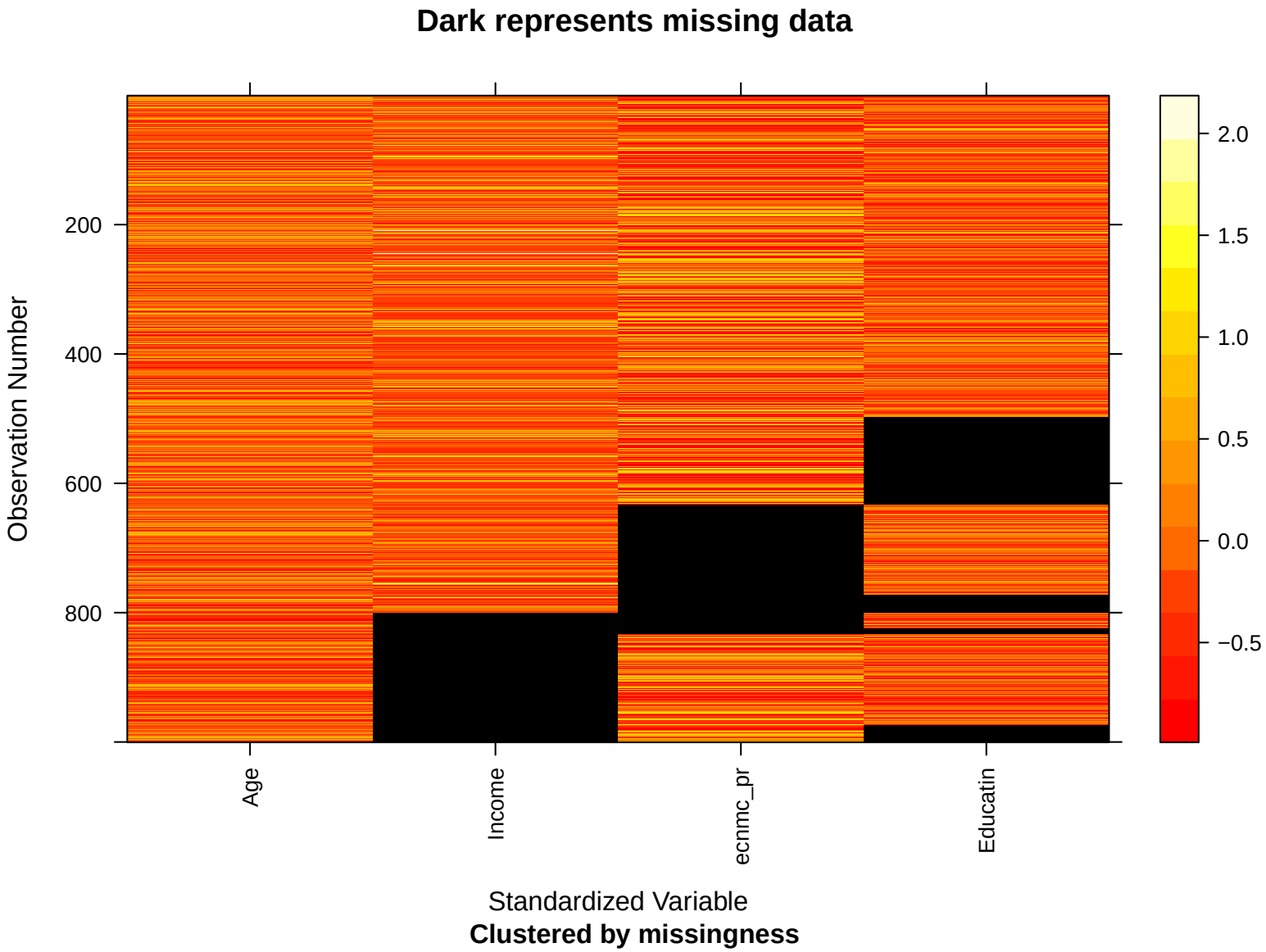
GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2019
Columbia University

what do you normally do when you find this?



what **should you do** when you find this?



the nature of the missing data problem is known

unit	age	income	economic perceptions	education
1	33	25	3	14
2	22	?	-2	12
3	50	300	0	?
4	?	220	1	20
5	18	?	-1	11
6	45	180	2	13
7	76	50	-3	16
8	29	98	?	14

Total Error approaches have captured its causes

- **item non-response:** units provide information selectively
  - e.g. not everyone wants to reveal their income
- **unit non-response:** "units" provide no information
  - e.g. war prevents collecting information
- **lost information:** information was collected but is no longer available
  - e.g. records lost in a natural disaster

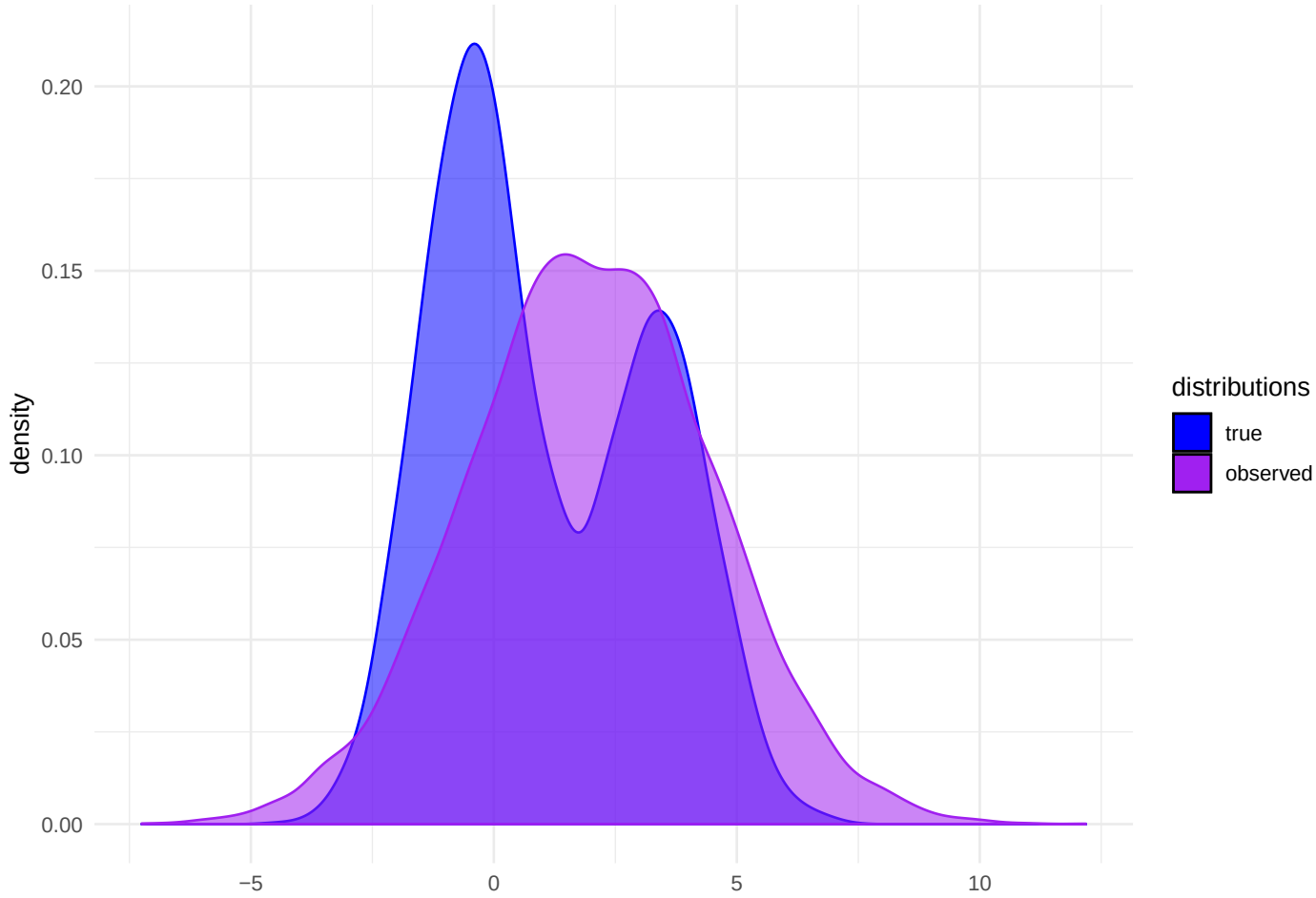
and its consequences are also well known...

though quite often ignored

- **Fundamental concerns:**
  - using only **available** information but not all **possible** information
  - observed distributions $\neq$ true distrbutions because of **missingness mechanism**
- **Consequences:**
  - **valid inferences/predictions** for the **wrong population**
  - **invalid inferences/predictions** for **unknown population segments**

and its consequences are also well known...

though quite often ignored



and its consequences are also well known...

though quite often ignored

- most algorithms **assume no missingness** in the data
  - implementations treat missingness as a **nuissance** and "handle" it in different ways
- potential **biases**:
  - projections outside of the **support region**
  - projections based on samples that **differ systematically** from target population
  - **underestimated variances** (relevant on inferential problems)

let's start with some theory and notation

$$D = \begin{bmatrix} 1 & 33 & 25 & 3 & 14 \\ 2 & 22 & 20 & -2 & 12 \\ 3 & 50 & 300 & 0 & 16 \\ 4 & 30 & 220 & 1 & 20 \\ 5 & 18 & 10 & -1 & 11 \\ 6 & 45 & 180 & 2 & 13 \\ 7 & 76 & 50 & -3 & 16 \\ 8 & 29 & 98 & 2 & 14 \end{bmatrix} \quad M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where

$D : \{D^{obs}, D^{miss}\}$

D^{miss} = **missing** data

D^{obs} = **observed** data

$M : \{1, 0\}$= missingness indicator matrix

now we can characterize missingness mechanisms

- **Missing Completely at Random (MCAR)**: the probability of missingness is independent from the data (D)

$$P(M|D) = P(M)$$

- **Missing at Random (MAR)**: the probability of missingness only depends on observed data (D^{obs})

$$P(M|D^{obs}) = P(M|D)$$

- **Non-Ignorable (NI)**: the probability of missingness depends both on observed (D^{obs}) and unobserved (D^{miss}) data

$$P(M|D^{obs}, D^{miss}) = P(M|D)$$

not all missingness mechanisms are theoretically plausible for data imputation

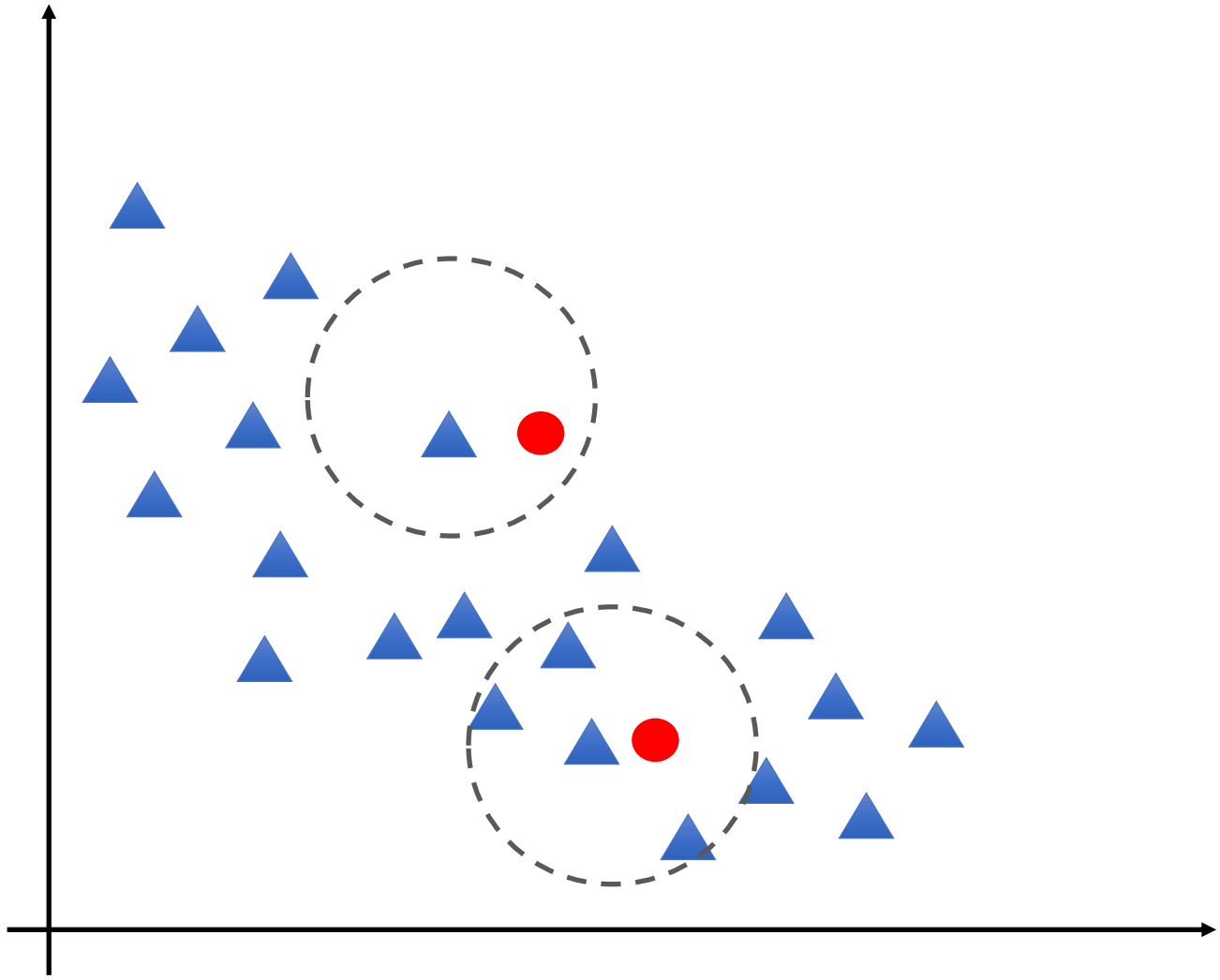
Mechanism	**Predict using**
Missing Completely at Random (MCAR)	–
Missing at Random (MAR)	D^{obs}
Non-ignorable (NI)	D^{obs} & D^{miss}

- MAR would imply that missingness is **ignorable**
  - observed data can be used to recover unobserved data
- MAR is an **assumption** (not directly verifiable)
  - ... but supported by some sort of theory about how missingness was generated

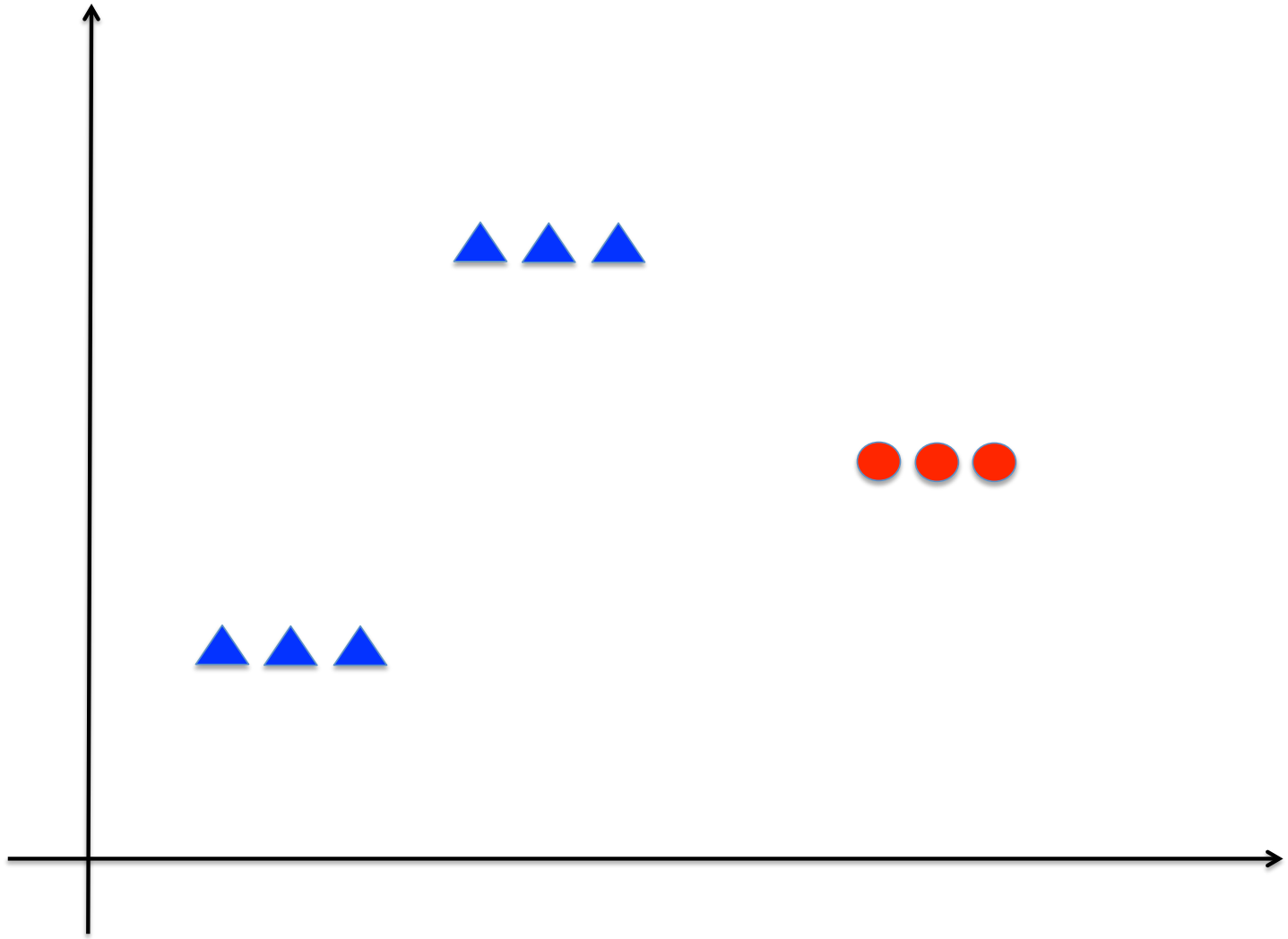
with appropriate assumptions, we can rely on a few methods to impute missing data

- **hot/cold deck imputation**: missing data is provided by a "nearest neighbor" donor unit
- **mean imputation**: missing data is provided by the mean of observed data
- **regression-based imputation**: missing data is generated by a regression model, conditional on observed data
- **multiple imputation**: model-based imputation that produces $m > 1$ vales for each missing value, conditional on observed data

hot/cold deck Imputation of missing data

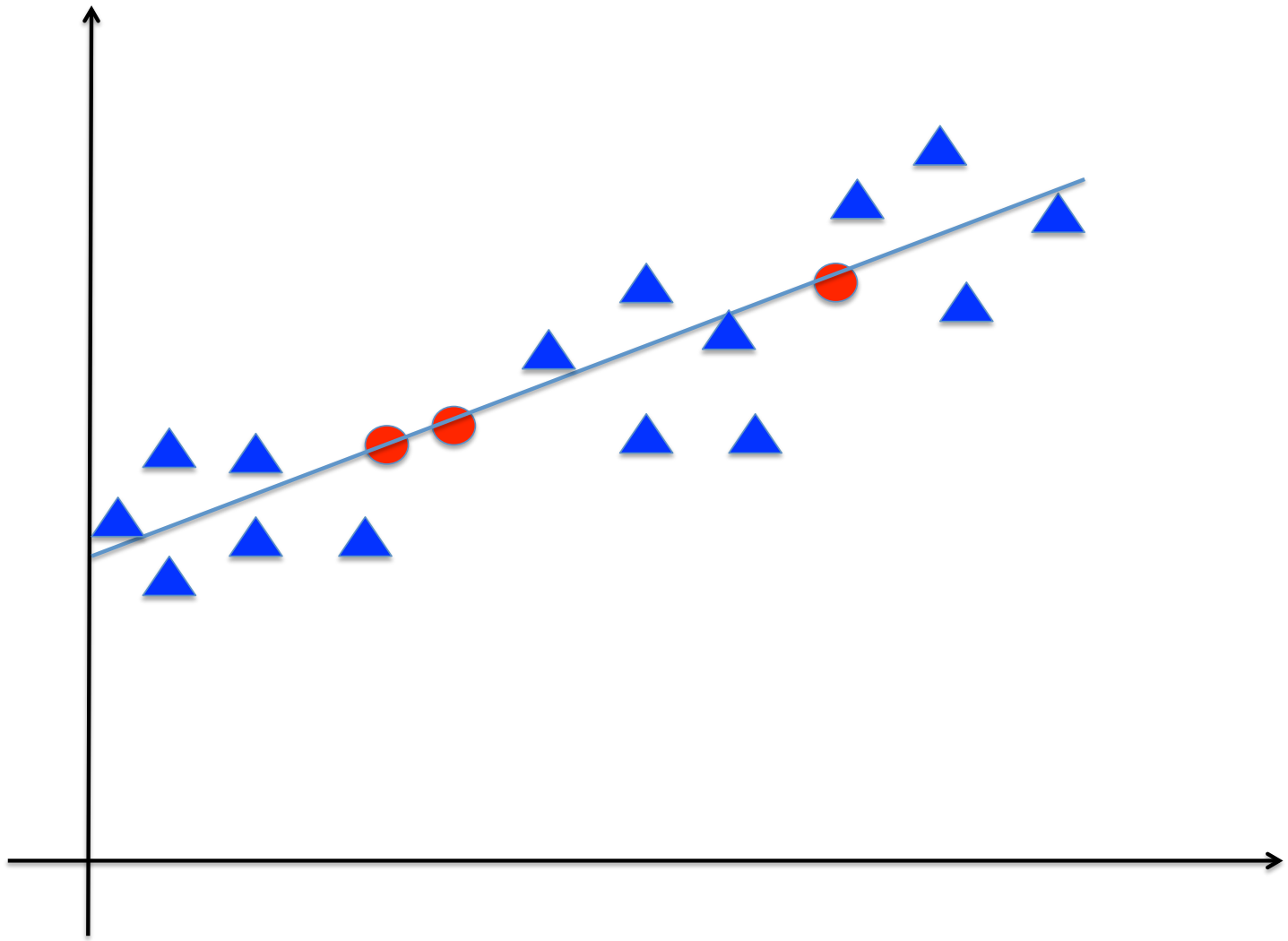


mean imputation of missing data



King, Honaker, Joseph & Scheve (1999)

regression-based imputation of missing data



King, Honaker, Joseph & Scheve (1999)

why do we distinguish between single-value and multiple imputation?

- **single-value imputations** may have important shortcomings:
  - potential bias in point estimates
  - understate uncertainty surrounding imputed values (underestimate variances)
- **multiple imputation** overcomes some of these shortcomings
  - assigns $m > 1$ plausible values from a conditional distribution
  - generates variance estimates that converge to the true variance
- particularly important when trying to generate **valid inferences**

in theory, Multiple Imputation involves solving 3 tasks - Rubin (1987)

(i) **modeling:** specify a (hypothetical) joint distribution for all data

$$P(D, M) \tag{1}$$

(ii) **imputation:** derive a posterior predictive distribution of missing data (D^{miss}) given observed data (D^{obs})

$$P(D^{miss}|D^{obs}, M) \tag{2}$$

(iii) **estimation:** compute the posterior distribution of parameters in (2) to produce random draws and enable imputations of missing data

in practice, Multiple Imputation works in 3 steps - Rubin (1977)

1) **impute** $m > 1$ values for each missing data
   - employ an algorithm to impute missing data m times
   - existing data remain unchanged
   - a stochastic value is assigned for missing values
2) **analyze** each one of the $m > 1$ data bases
   - use each of m data bases *as if* it had full information
   - perform analyses on each data base: compute descriptive statistics, regression, etc
3) **combine** $m > 1$ estimates to compute point estimates and variances of quantities of interest (q)

Multiple Imputation - quantities of interest (q)

- **Point estimates** of quantities of interest

$$\tilde{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

where

$\tilde{q}$ = point estimate of quantities of interest
q_j = quantity of interest for imputation j
m = number of imputations

Multiple Imputation - quantities of interest (q)

- **Variance of the point estimate** of the quantity of interest
- sum of *within* and *between* imputation variance

$$
\begin{aligned}
SE(q)^2 &= \bar{w} + b \\
&= \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + \left(1 + \frac{1}{m}\right)\frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{m-1}
\end{aligned}
$$

where

$\bar{w}$ = *within* imputation variance
b = *between* imputation variance
$\tilde{q}$ = point estimate of the quantity of interest
q_j = quantity of interest on imputation j
m = number of imputations

Multiple Imputation - quantities of interest (q)

- q is distributed t with degrees of freedom defined by

$$d.f. = (m-1)\left[1 + \frac{1}{m+1}\frac{\bar{w}}{b}\right]^2$$

where

$\bar{w} =$ *within* imputation variance
$b =$ *between* imputation variance
$m =$ number of imputations

some advantages of multiple imputation

- Does not assume that we have recovered the **true value of each missing observation**
- Accurately reflects **imputation uncertainty**
  - imputation with useful information have low variances
  - includes *between* imputation variance to avoid underestimating the general variance
- Preserves information on **data distributions**
  - uses all available information to characterize the distribution that governs data generation
  - each imputed value comes from this distribution

(1) multiple imputation through **joint distributions**

- **main assumption:** full data — $D : \{D^{obs}, D^{miss}\}$ — has a multivariate distribution, e.g. multivariate normal:

$$D \sim MVN(\theta), \quad \theta = (\mu, \Sigma)$$

- **substantive problem:** estimating θ from the observed data (D^{obs})
- **imputation** boils down to predictions from a regression that uses observed data (D^{obs}) and random draws of parameters — β, ϵ — derived from θ

$$\tilde{D}^{miss}_{i,j} = D^{obs}_{i,-j}\tilde{\beta} + \tilde{\epsilon}_i$$

- solves a k-dimensional problem
- implemented in the `Amelia` package

(2) multiple imputation through **chained equations**

- **different approach**: specify a **sensible imputation model** (2), skipping the joint distribution specification
- **variable-by-variable imputations** from a **conditional model for each variable** (D_j)

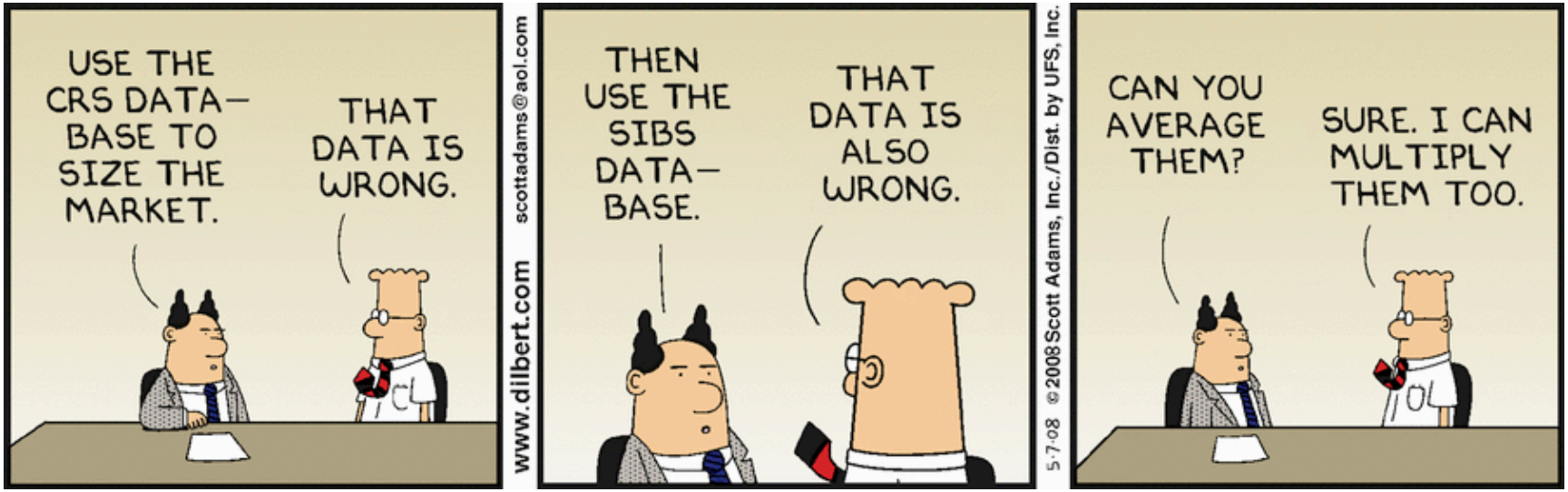
$$P(D_j^{miss}|D_{-j}^{obs}, M), \quad j = 1, ..., k$$

- breaks a k-dimensional problem into k one-dimensional problems
- allows **different models** and **different predictors** for each variable

(2) multiple imputation through **chained equations**

- **algorithm:**
  - (i) impute simple values for all missing data
  - (ii) model a variable (x_i) by a regression conditional an all other variables (x_{-i})
  - (iii) use the fitted regresion to impute missing values for that variable (x_i)
  - (iv) cycle through all variables using the same procedure and update imputations iteratively
  - (v) iterative imputation will converge to a distribution
- implemented in the `mi` and `mice` packages

a comment on missing data in big data contexts



a comment on missing data in big data contexts

- **belief:** missingness becomes less relevant as size tends towards **big data**
  - belief that **asymptotics** kick in and solve everything
  - belief that **large samples** are, by definition, unbiased
  - belief that **implementations of common algorithms** handle missing data natively (and thus appropriately)
- **problem:** missingness may be generating a **biased sample** of observed data... regardless of size
- **consequences:**
  - training and testing sets $\neq$ general population
  - changes in missingness parameters "change" data when true distributions remain unchanged
- **food for thought:** big data as big n v big data as big p

can ML algorithms recover original conditional distributions?

- **question:** how good are ML algorithms at recovering the original conditional distribution of a target variable $P(y|x)$ if using data from a biased sample (Zadrozny 2004)?
  - **local learners:** output depends asymptotically on $P(y|x)$
    - logistic regression, hard margin SVM
  - **global learners:** output depends asymptotically on $P(y|x)$ and $P(x)$
    - Bayesian classifiers, decision trees, soft margin SVM
- **results:** *iff* **MAR** is assumed, **local learners are not affected by data missingness**, but global learners are
- **note:** these results refer to $P(y|x)$ not to D

Missing Data:
Theory and Practice

Marco Morales
marco.morales@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2019
Columbia University