# Independent Study of *A Bayesian Approach to Graphical Record Linkage and De-duplication*

Presented by Melody Jiang

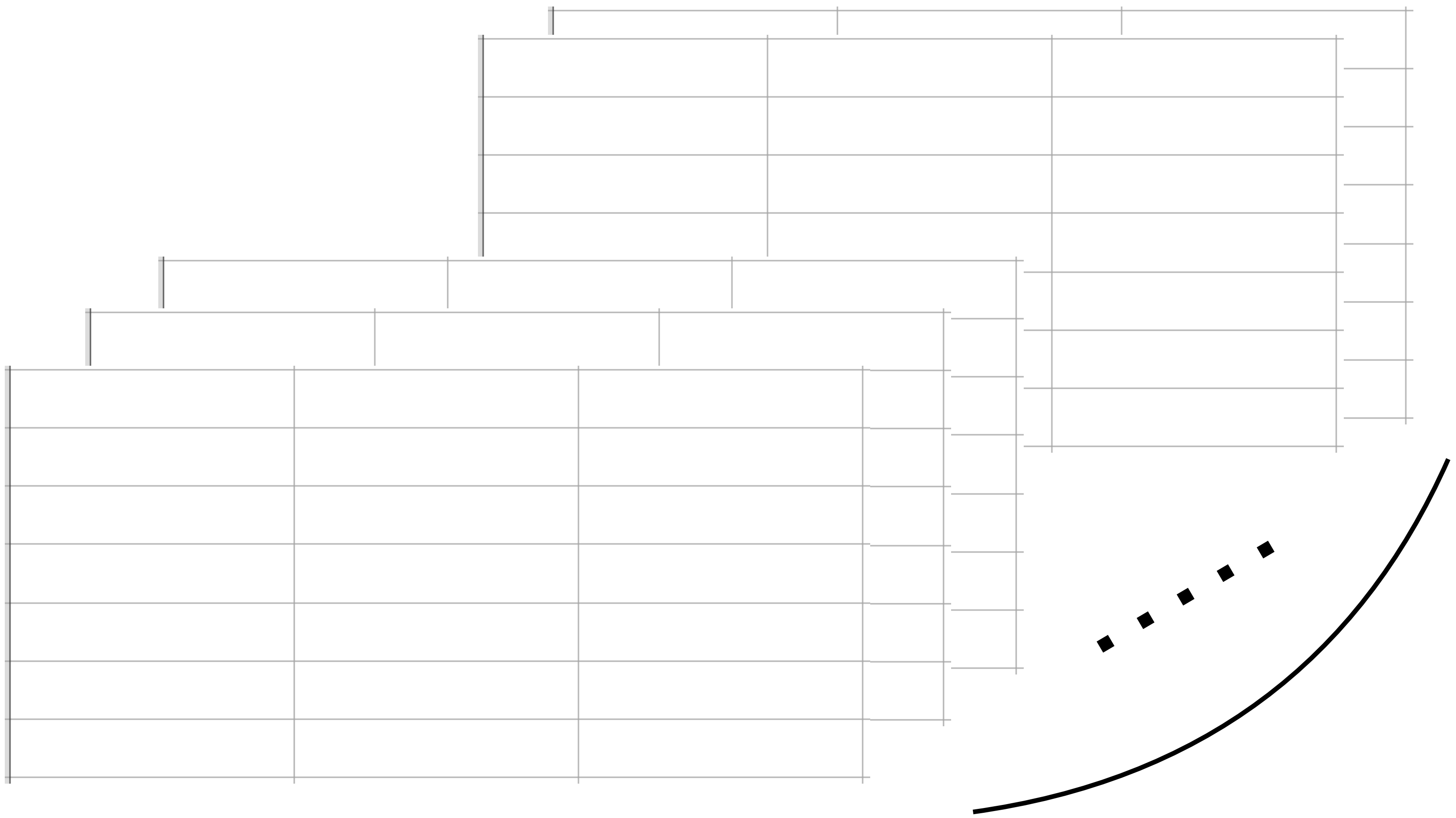
Feb 28, 2019

# Motivation

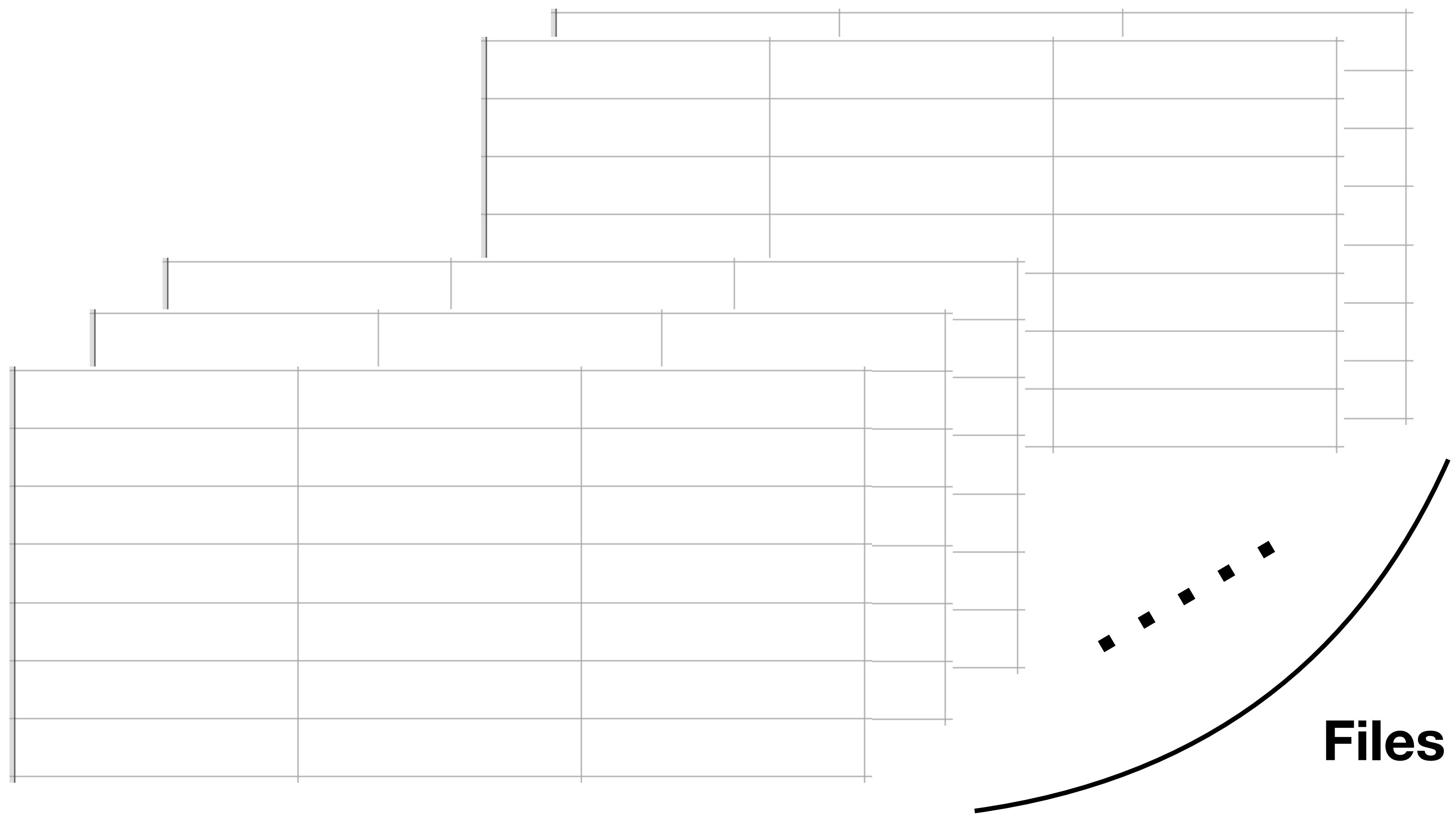
- Link data about a individual coming from different sources to the same individual
- Exciting societal applications!

# Specifically: Different Files

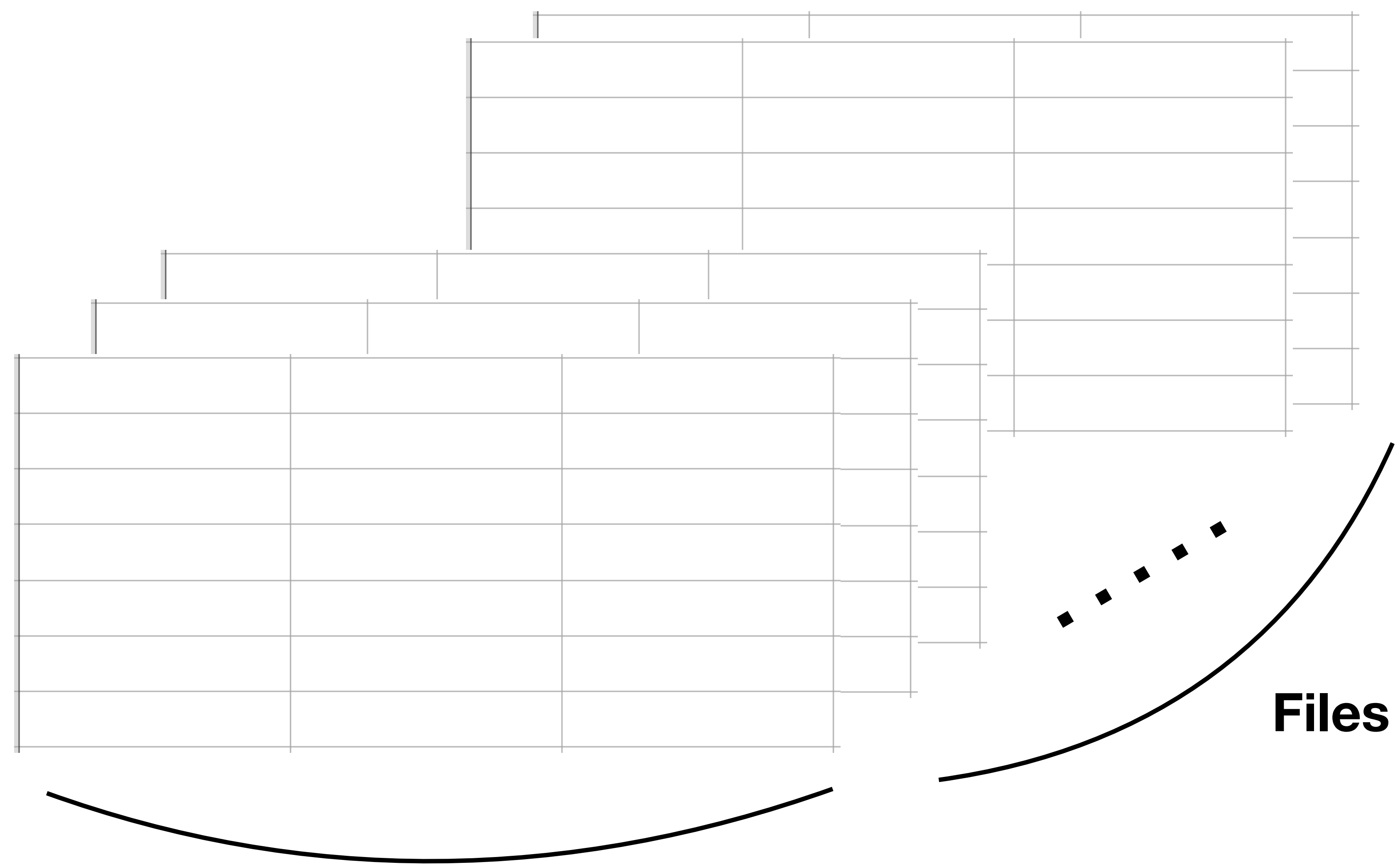
# Specifically: Different Files



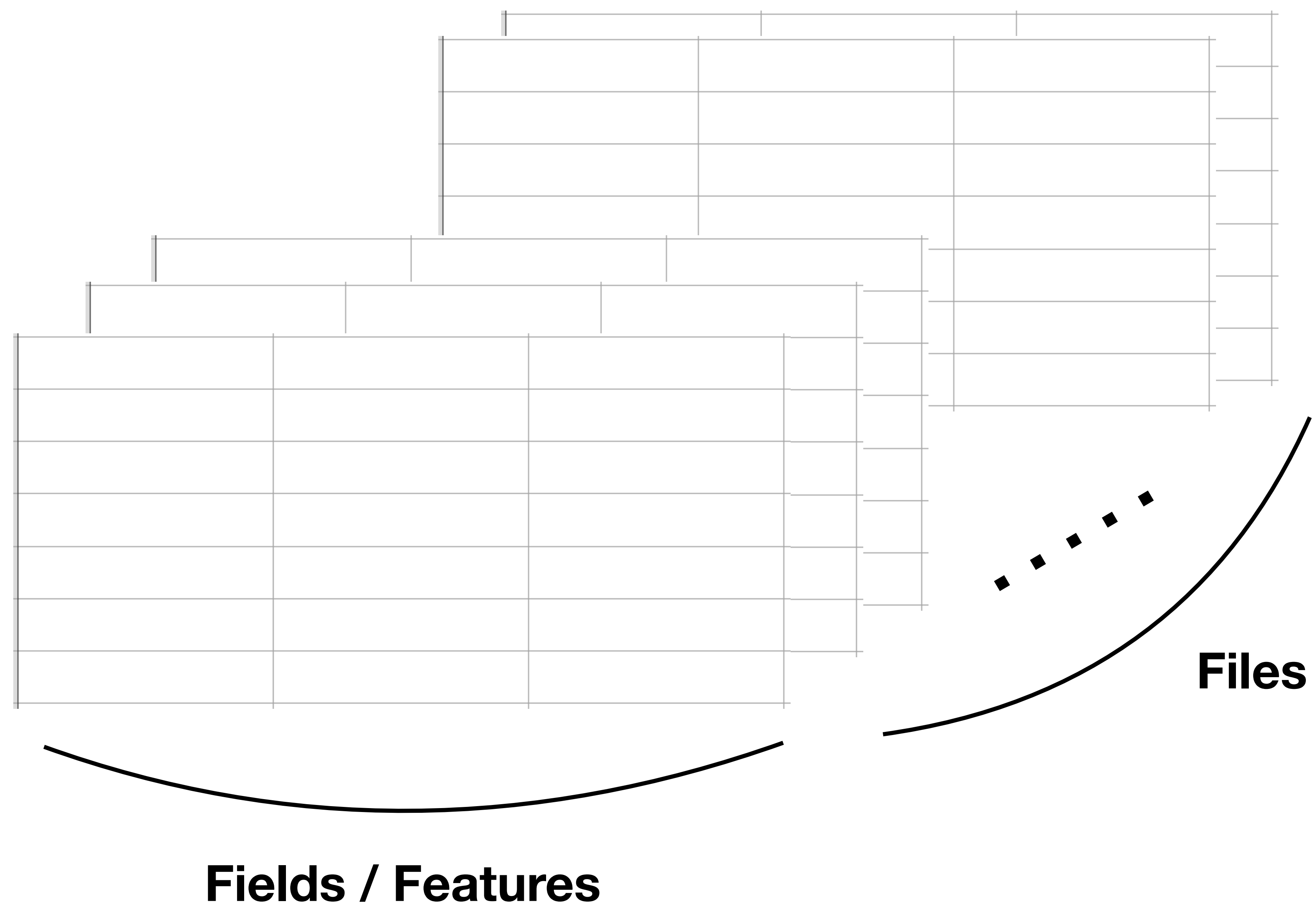
# Specifically: Different Files



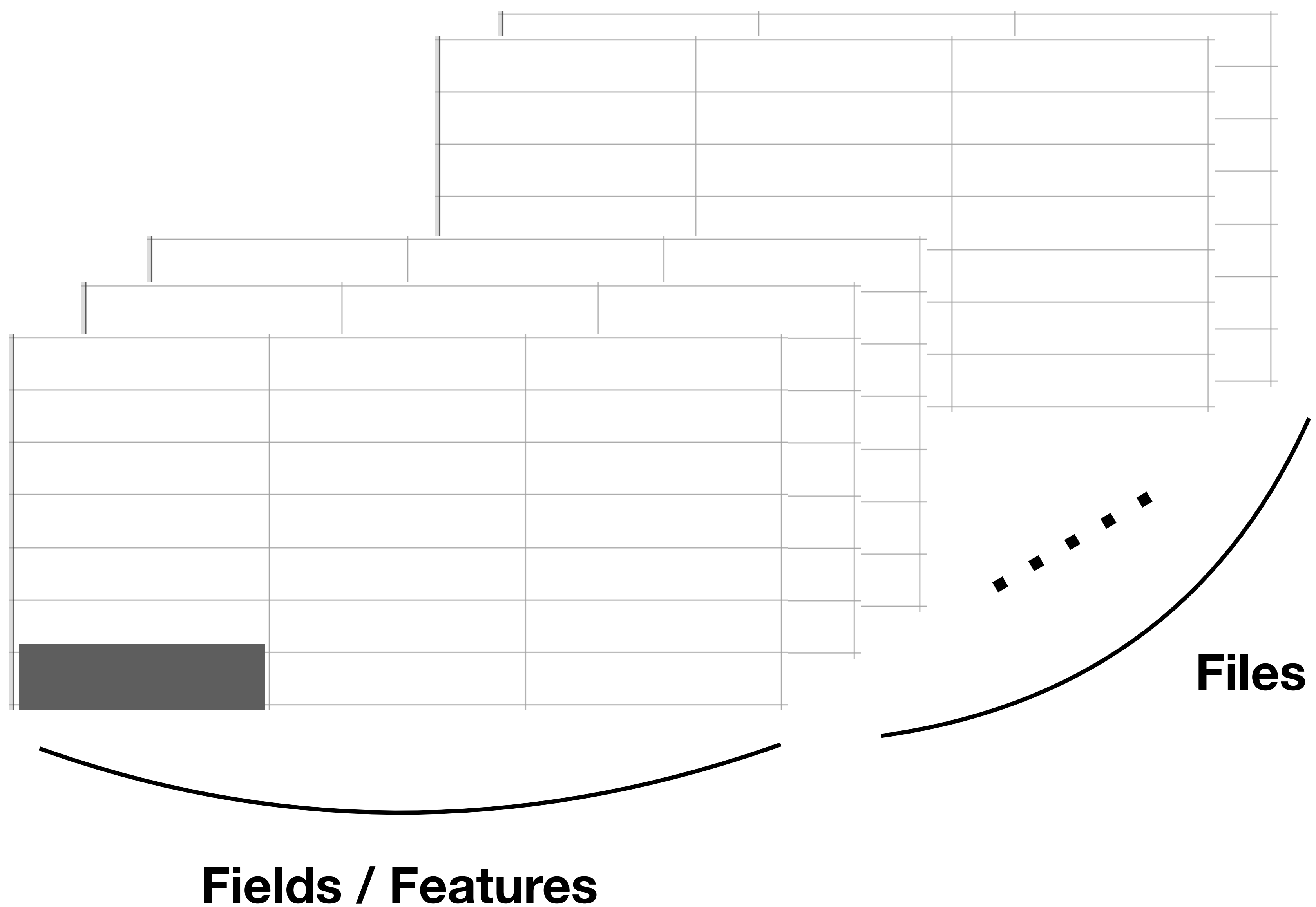
# Specifically: Different Files



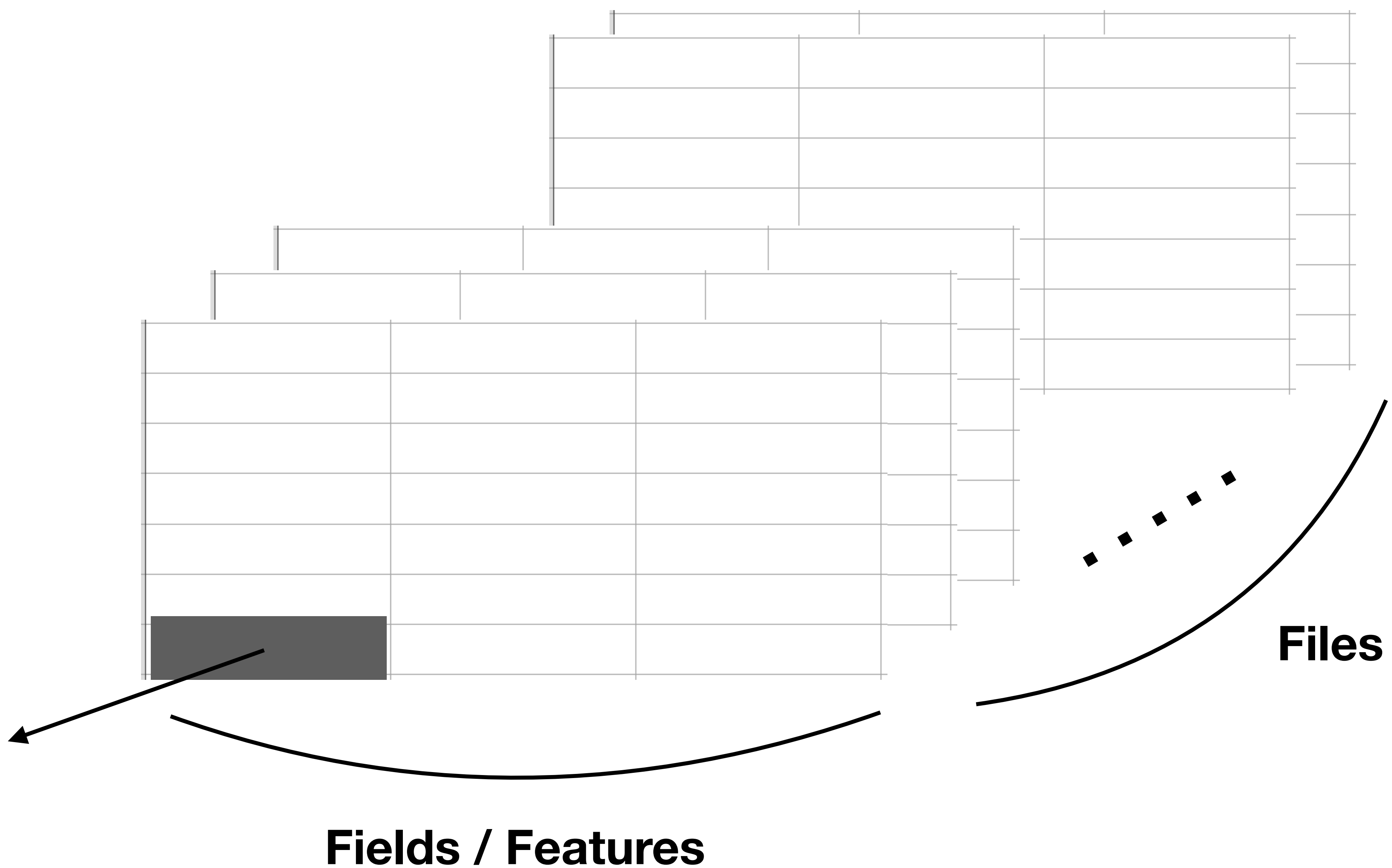
# Specifically: Different Files



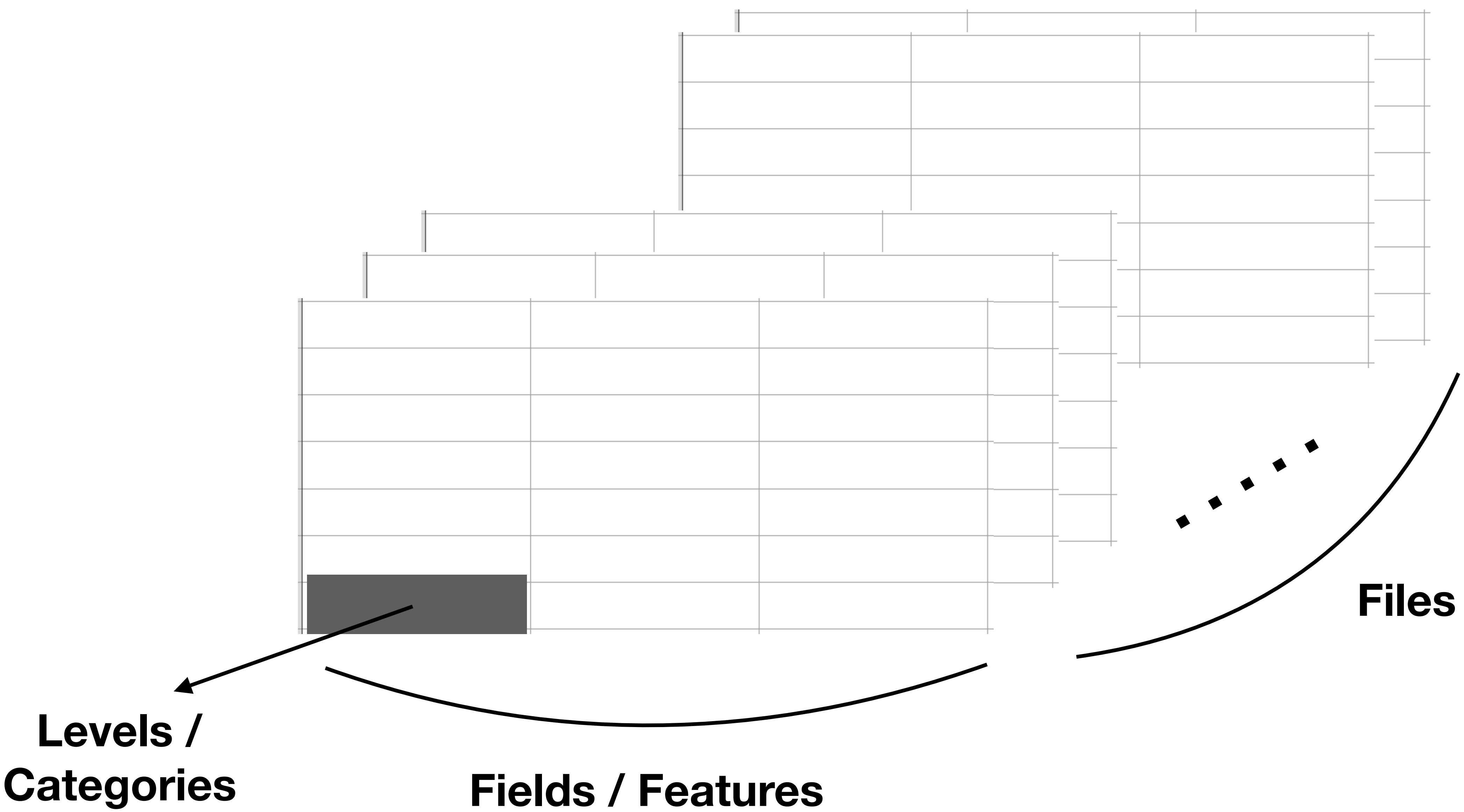
# Specifically: Different Files



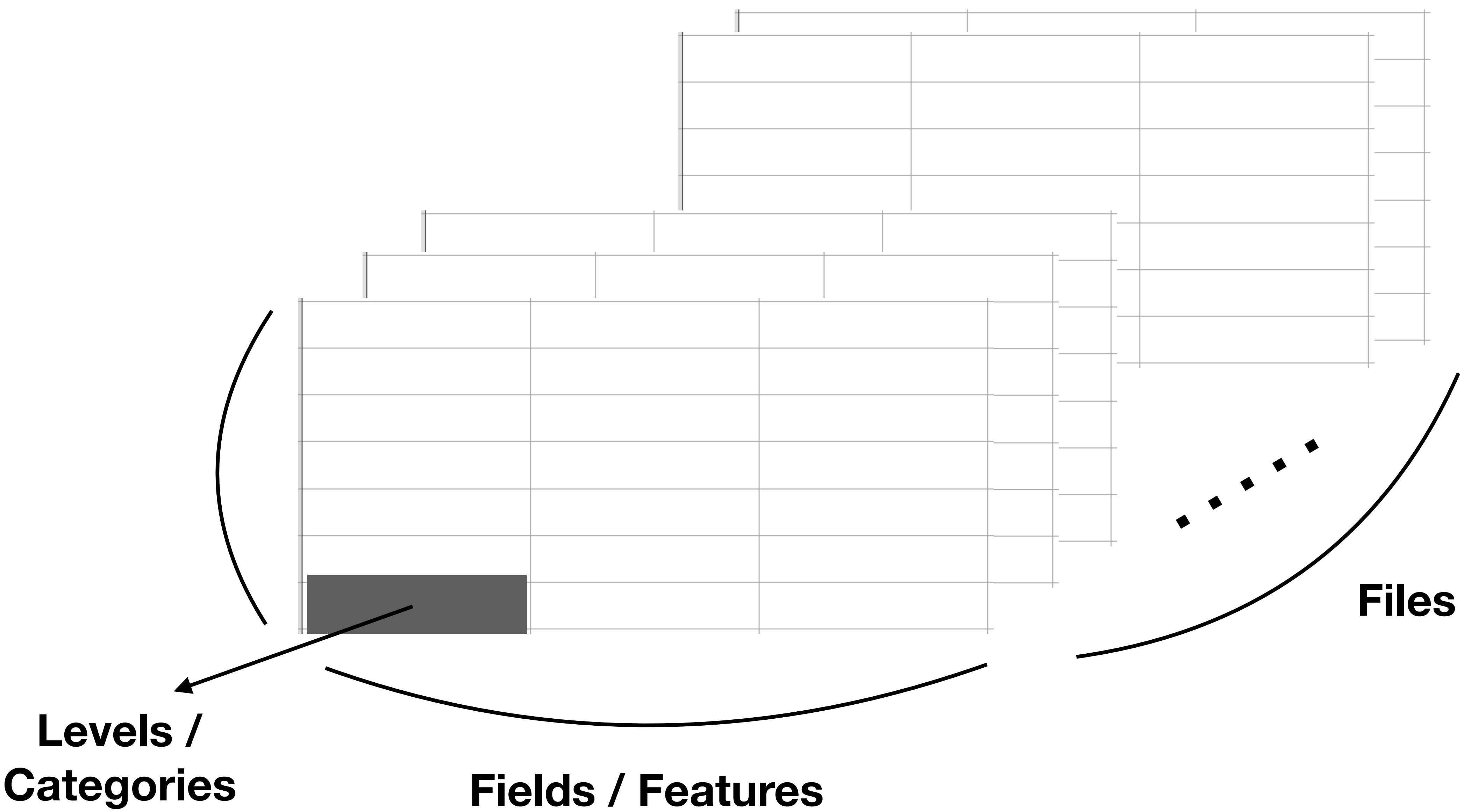
# Specifically: Different Files



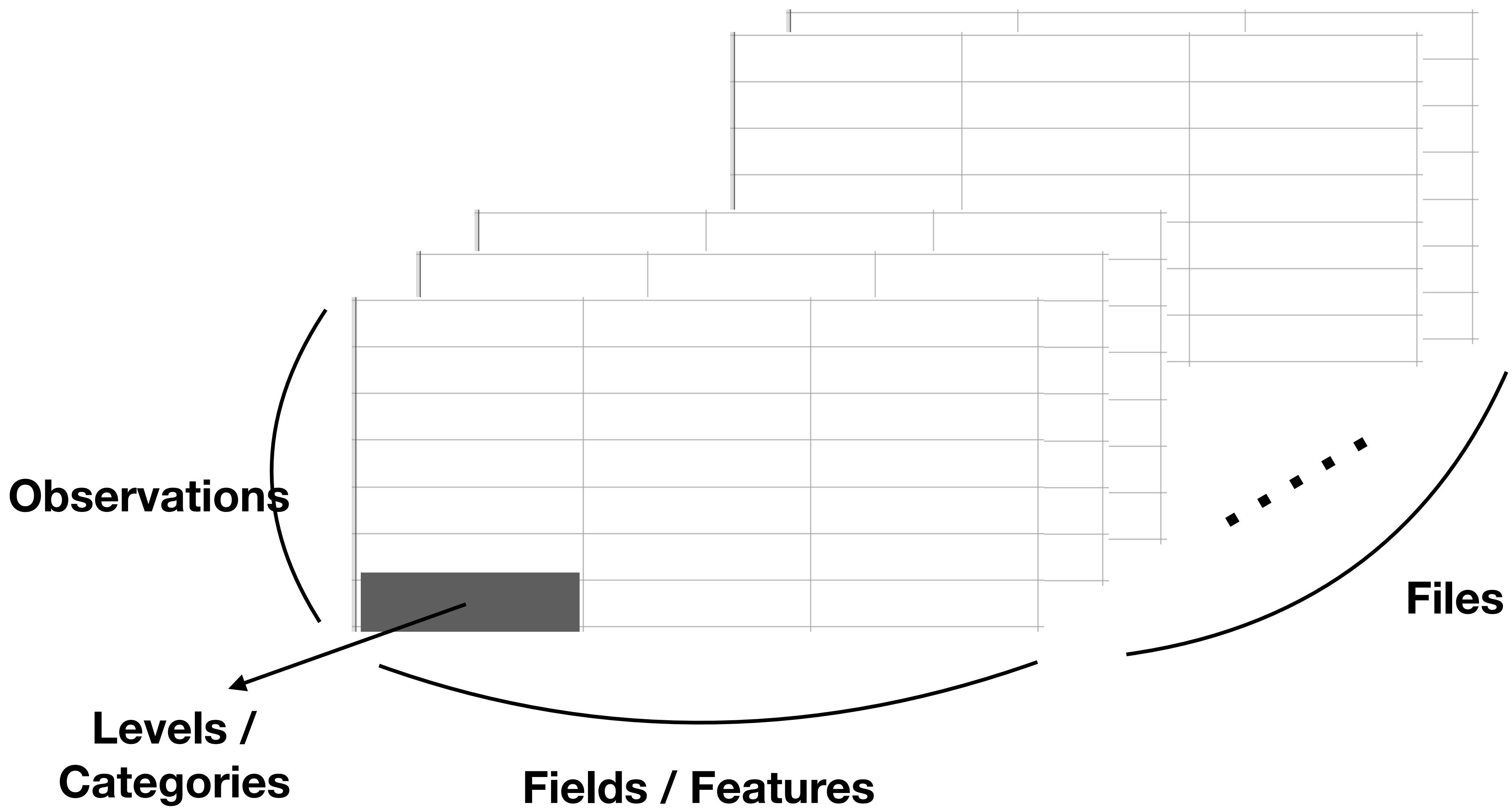
# Specifically: Different Files



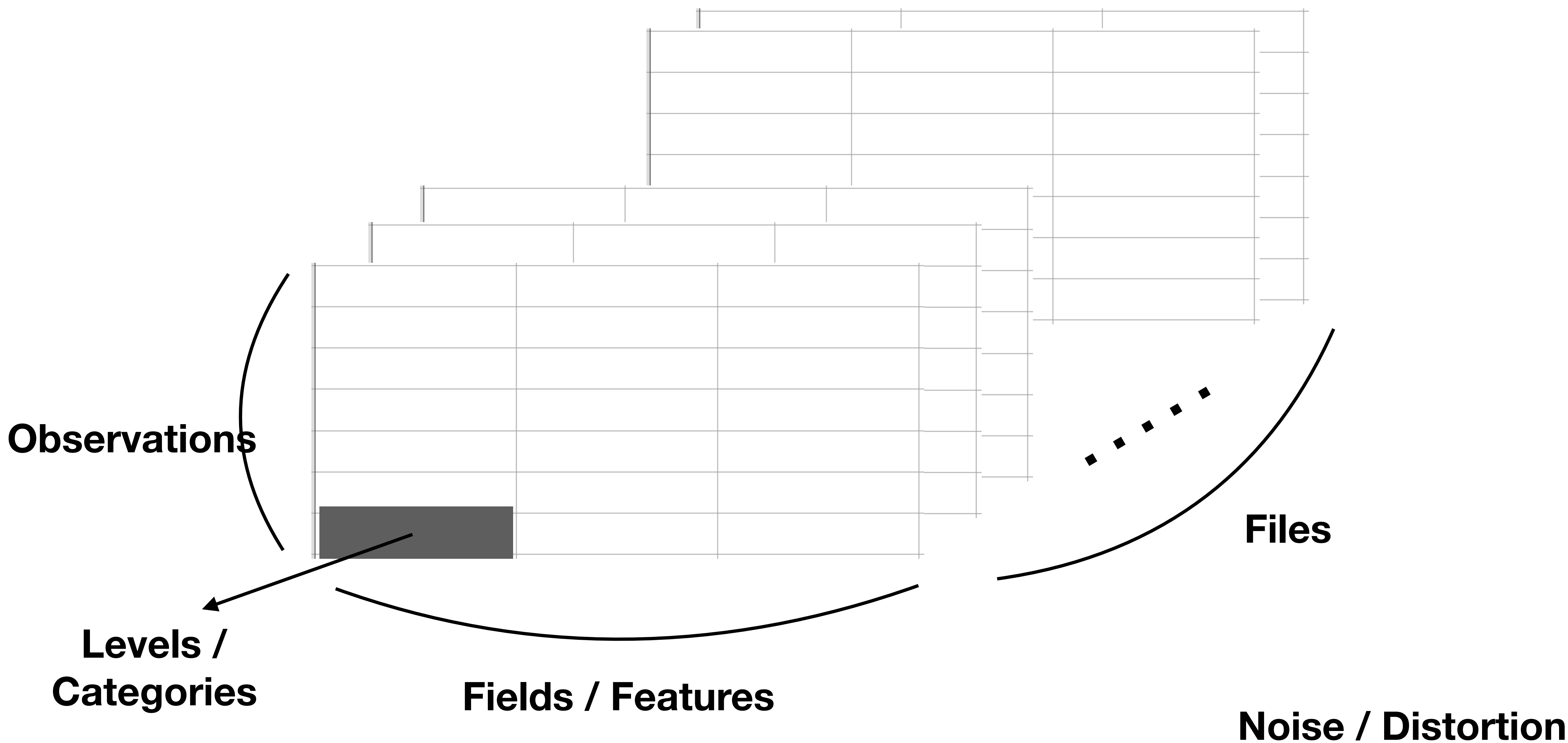
# Specifically: Different Files



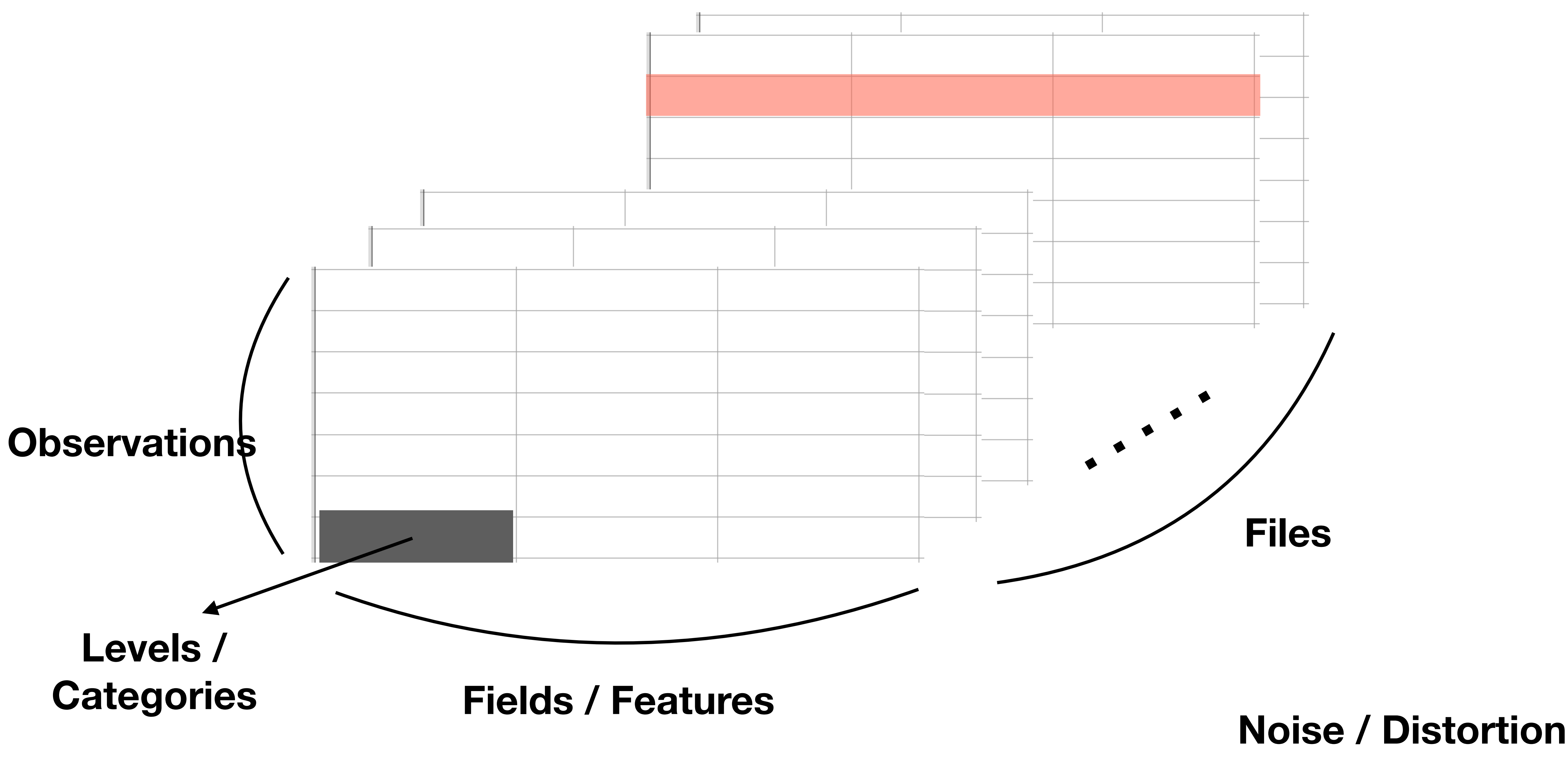
# Specifically: Different Files



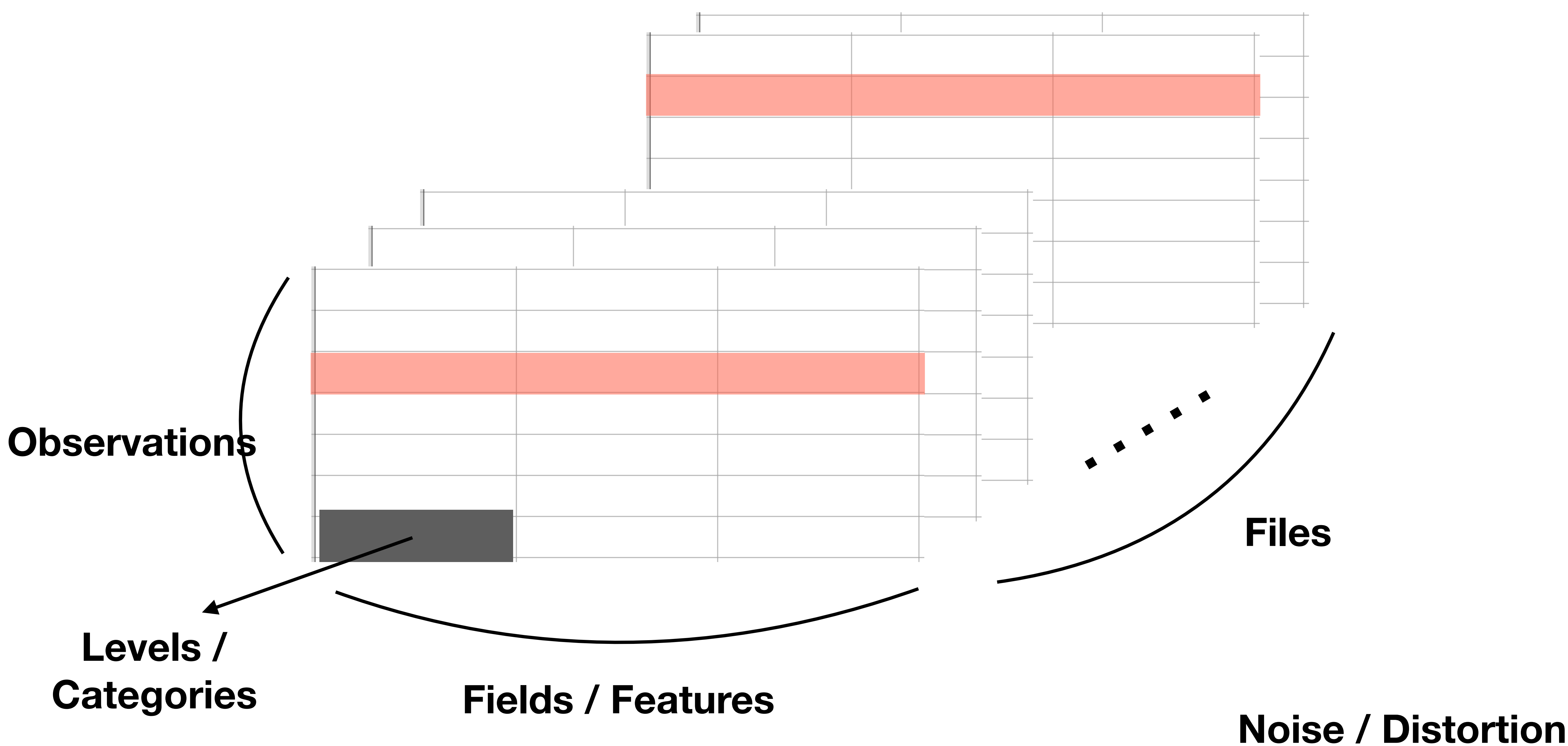
# Specifically: Different Files



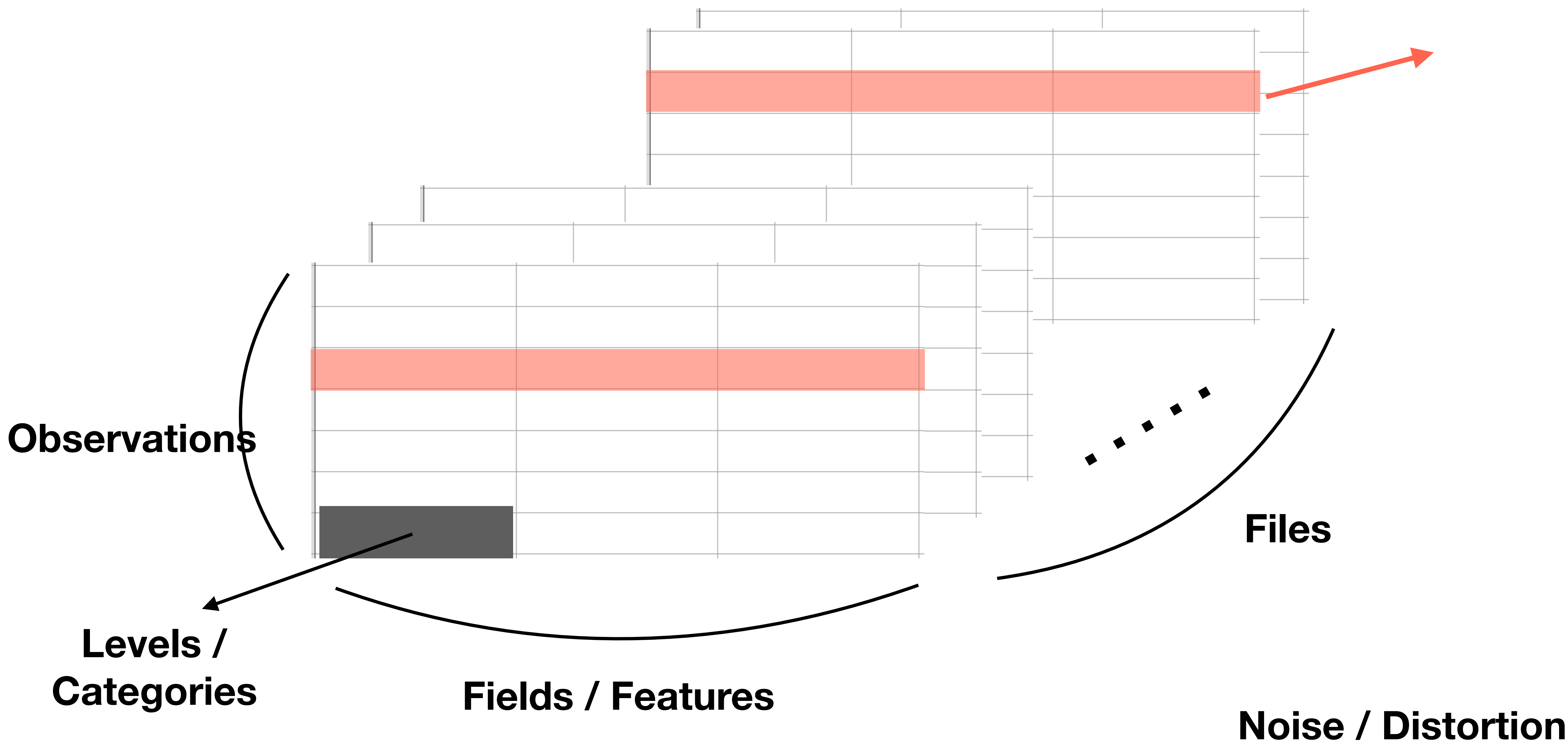
# Specifically: Different Files



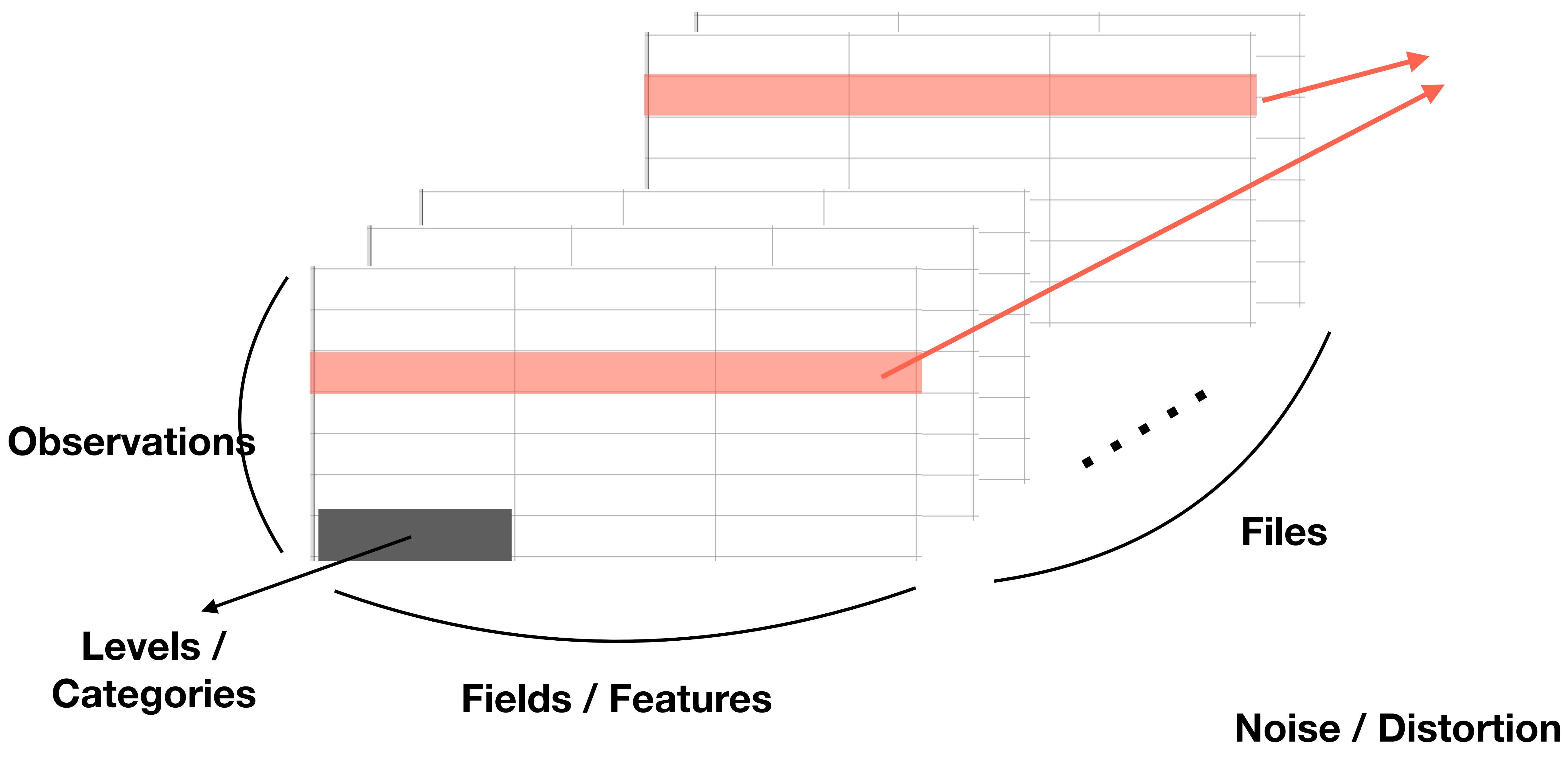
# Specifically: Different Files



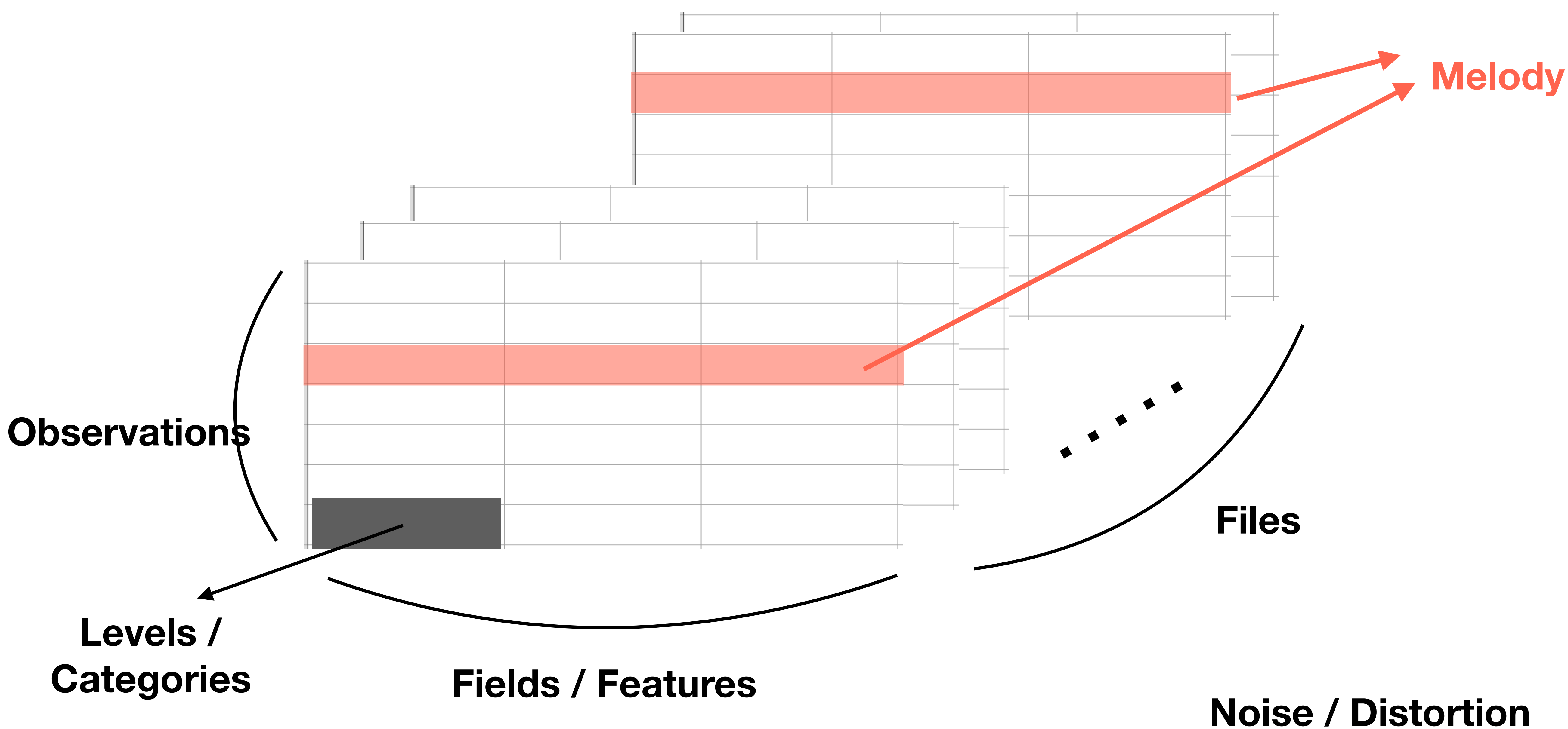
# Specifically: Different Files



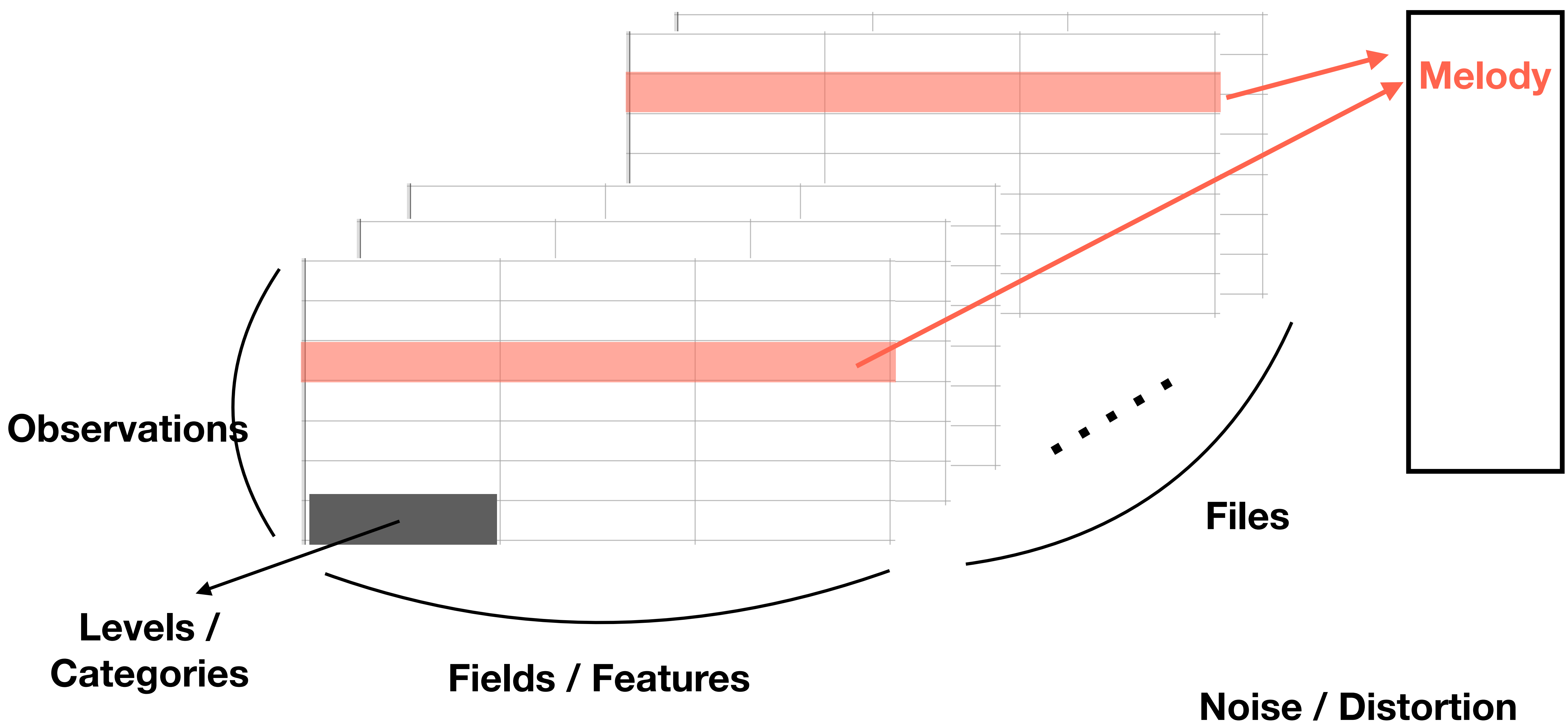
# Specifically: Different Files



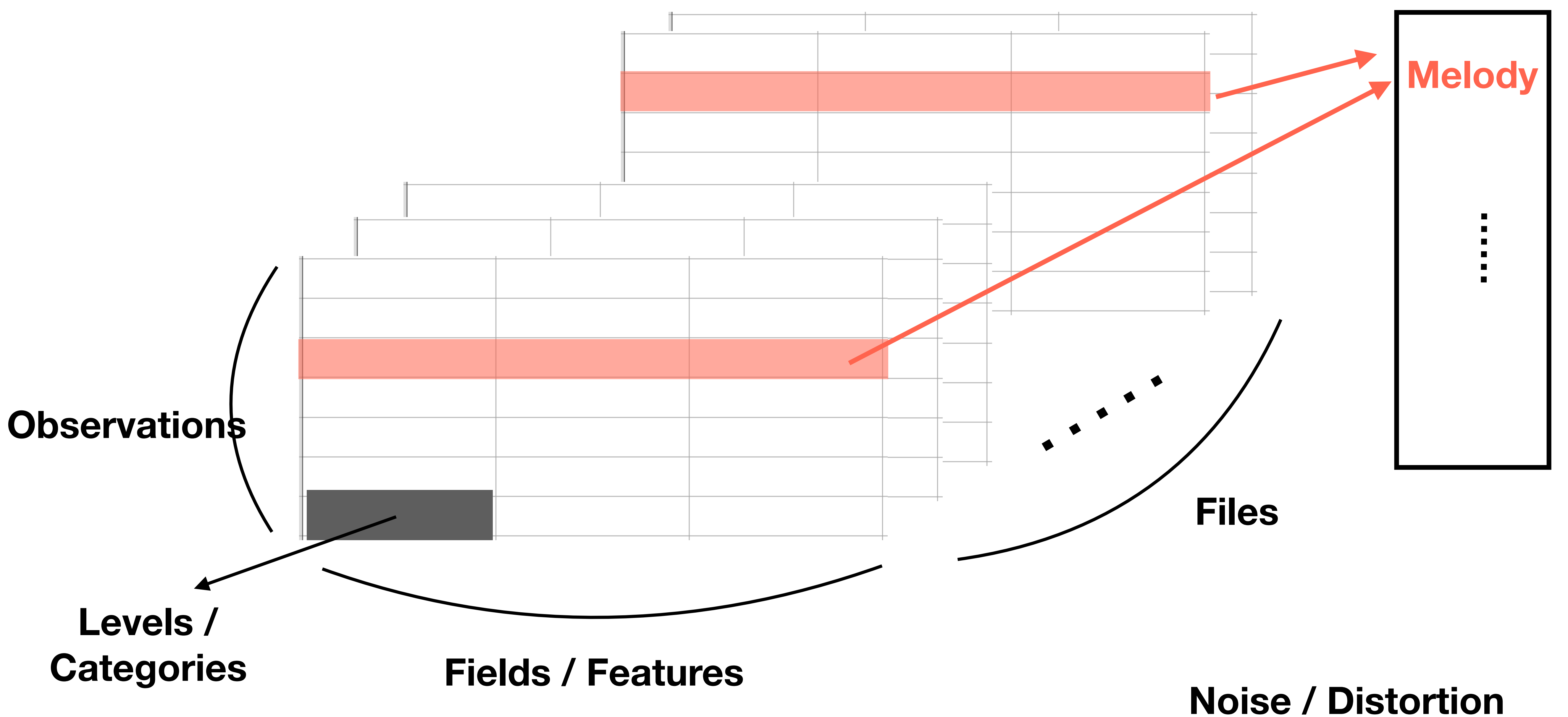
# Specifically: Different Files



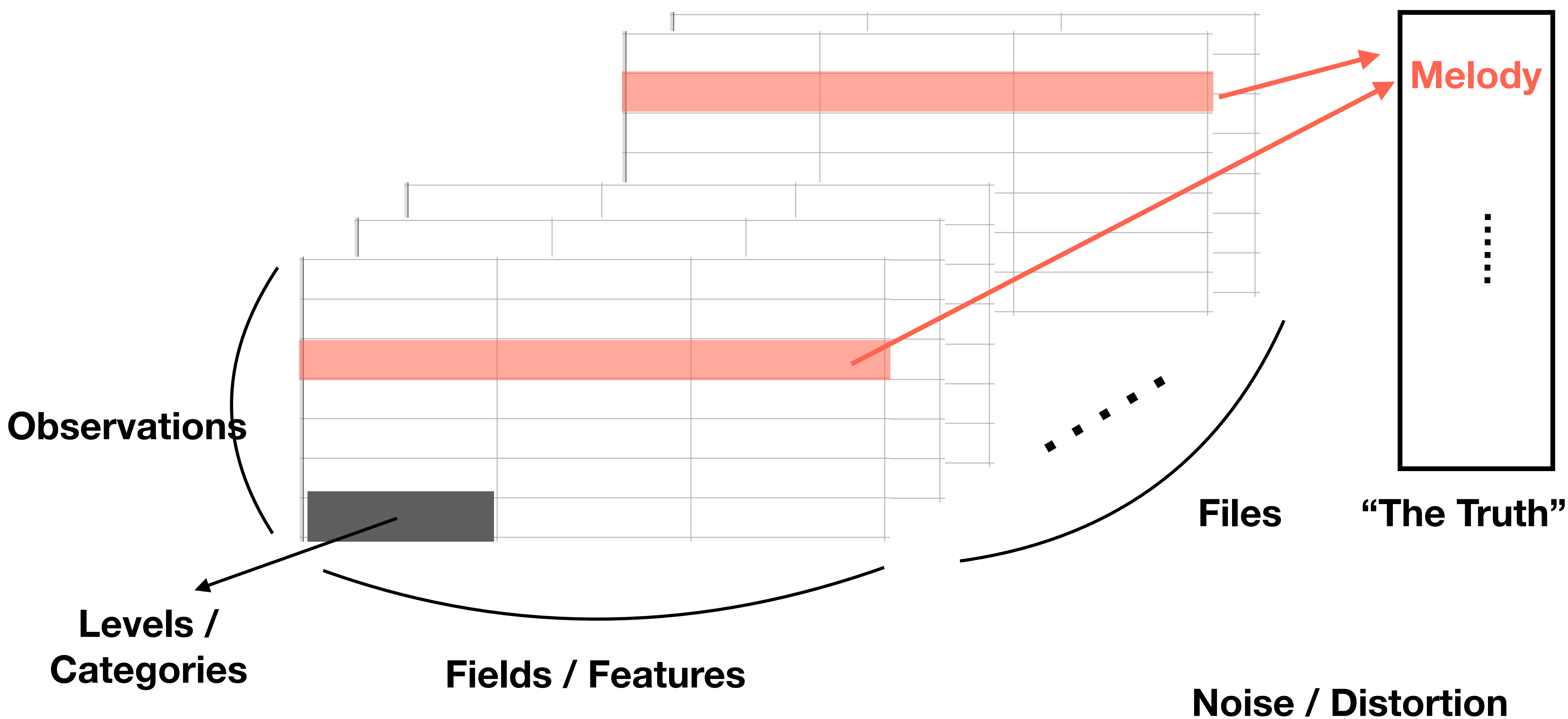
# Specifically: Different Files



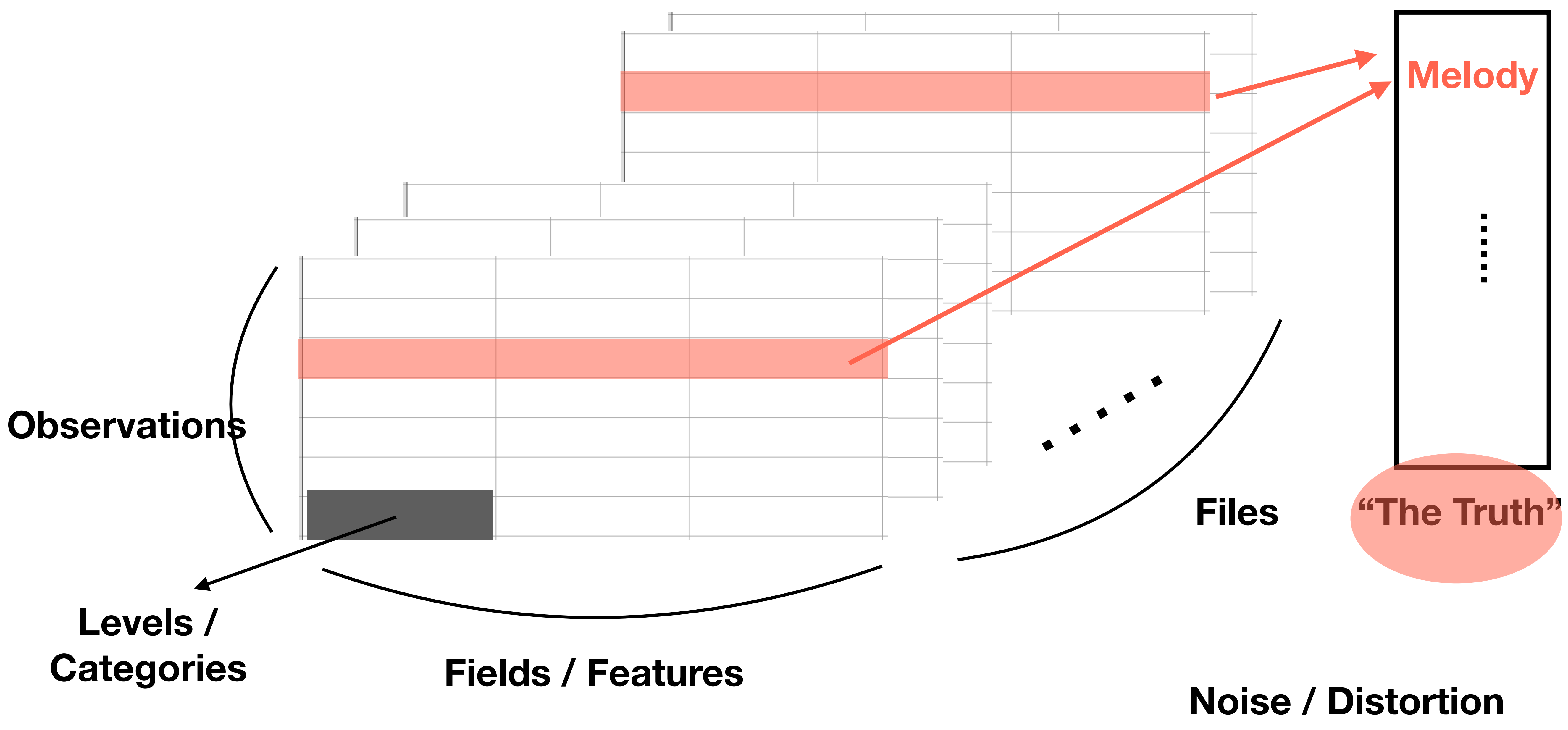
# Specifically: Different Files



# Specifically: Different Files



# Specifically: Different Files



# In English: Independent Fields Model

| | | |
|---|---|---|

# In English: Independent Fields Model

- What is given which individual it actually is, true field values, noise, and probability of each field value?

# In English: Independent Fields Model

- What is given which individual it actually is, true field values, noise, and probability of each field value?
- How is noise distributed?

# In English: Independent Fields Model

- What is given which individual it actually is, true field values, noise, and probability of each field value?

- How is noise distributed?

- How are true field values distributed?

# In English: Independent Fields Model

- What is given which individual it actually is, true field values, noise, and probability of each field value?
- How is noise distributed?
- How are true field values distributed?
- How is "the truth" distributed?

# In English: Independent Fields Model (cont.)

# In English: Independent Fields Model (cont.)

- p("the truth", true field values, noise, probability of each field value, a parameter associated with noise | data)

# In English: Independent Fields Model (cont.)

- p("the truth", true field values, noise, probability of each field value, a parameter associated with noise | data)
- Full conditionals

# In English: Independent Fields Model (cont.)

- p("the truth", true field values, noise, probability of each field value, a parameter associated with noise | data)
- Full conditionals
- So that… Split and MErge REcord linkage and De-duplication (SMERED) Algorithm —> "the Truth"

# In English: Independent Fields Model (cont.)

- p("the truth", true field values, noise, probability of each field value, a parameter associated with noise | data)
- Full conditionals
- So that… Split and MErge REcord linkage and De-duplication (SMERED) Algorithm —> "the Truth"
    - Me: simple Gibbs sampler for now

# Future Directions

- More elaborate models: “missing fields, data fusion, complicated string fields, population heterogeneity, dependence across fields, across time, or across individuals”
- Computational speed-ups: “online learning, variational inference, approximate Bayesian computation”
- Topic models?

# Future Directions

- Topic models?
    - Saw Bayesian

A Latent Dirichlet Model for Unsupervised Entity Resolution


Indrajit Bhattacharya    Lise Getoor
Department of Computer Science
University of Maryland, College Park, MD 20742


A Latent Dirichlet Allocation Model for Entity Resolution


Indrajit Bhattacharya
University of Maryland
College Park, MD, USA
indrajit@cs.umd.edu

Lise Getoor
University of Maryland
College Park, MD, USA
getoor@cs.umd.edu


1 Aug, 2005

Document clustering as a record linkage problem



DOI: 10.1145/3209280.3229109
Conference: the ACM Symposium

# Appendix: Independent Fields Model

## Notations

- There are $k$ files or lists
- There are $p$ fields in each file
- Field $l$ has $M_l$ levels

- $\mathbf{x}_{ij}$ := data for the $j^{th}$ record in file $i$, where $i = 1, ..., k$, $j = 1, ..., n_i$, and $n_i$ is the number of records in file $i$
- $\mathbf{y}_{j'}$ := latent vector of true field values for the $j'^{th}$ individual in the population, where $j' = 1, ..., N$, and $N$ being the total number of *observed* individuals from the population
- $\mathbf{\Lambda} = \{\lambda_{ij}; i = 1, ..., k; j = 1, ..., n_i\}$, where $\lambda_{ij} \in \{1, 2, ..., N_{max}\}$, indicating which latent individual the $j^{th}$ record in file $i$ refers to
- $z_{ij}$ := 1 or 0 according to whether or not field $l$ in $\mathbf{x}_{ij}$ is distorted

- $I$ denotes indicator functions
- $\delta_a$ := distribution of a point mass at $a$
- $\boldsymbol{\theta}_l$ := multinomial probabilities. Length of $\boldsymbol{\theta}_l = M_l$

- $i = 1, ..., k$ is the numbering of files
- $j = 1, ..., n_i$ is the numbering of records
- $j' = 1, ..., N$ see $\mathbf{y}_{j'}$ for definition
- $l = 1, ..., p$ is the numbering of features
- $m = 1, ..., M_l$ is the numbering of categories / levels of a feature

# Appendix: Independent Fields Model

$$
\begin{aligned}
\boldsymbol{x}_{ij\ell} \mid \lambda_{ij}, \boldsymbol{y}_{\lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell &\overset{\text{ind}}{\sim} \begin{cases} \delta_{\boldsymbol{y}_{\lambda_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\ \text{MN}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases} \\
z_{ij\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_\ell) \\
\boldsymbol{y}_{j'\ell} \mid \boldsymbol{\theta}_{j\ell} &\overset{\text{ind}}{\sim} \text{MN}(1, \boldsymbol{\theta}_\ell) \\
\boldsymbol{\theta}_\ell &\overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_\ell) \\
\beta_\ell &\overset{\text{ind}}{\sim} \text{Beta}(a_\ell, b_\ell) \\
\pi(\boldsymbol{\Lambda}) &\propto 1,
\end{aligned}
$$

$$
\begin{aligned}
&\pi(\boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{x}) \\
&\propto \prod_{i,j,\ell,m} \left[ (1 - z_{ij\ell}) \delta_{y_{\lambda_{ij}\ell}}(x_{ij\ell}) \;+\; z_{ij\ell} \theta_{\ell m}^{I(x_{ij\ell}=m)} \right] \\
&\quad \times \prod_{\ell,m} \theta_{\ell m}^{\mu_{\ell m} + \sum_{j'=1}^{N} I(y_{j'l}=m)} \\
&\quad \times \prod_{\ell} \beta_\ell^{a_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij\ell}} \\
&\times (1 - \beta_\ell)^{b_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - z_{ij\ell})}.
\end{aligned}
$$

# Appendix: Full Conditionals

$$\beta_\ell \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{x}$$
$$\sim \text{Beta}\left(a_\ell + \sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij\ell},\, b_\ell + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(1 - z_{ij\ell})\right)$$

$$\theta_{\ell m} \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{x}$$
$$\sim \text{Dirichlet}\left(\mu_{\ell m} + \sum_{j'=1}^{N} y_{j'\ell} + \sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij\ell}\, x_{ij\ell} + 1\right)$$

$$\boldsymbol{y}_{j'l} \mid \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}$$
$$\sim \begin{cases} \delta_{x_{ij\ell}} & \text{if there exist } i,\ j \in R_{ij'} \text{ such that } z_{ij\ell} = 0, \\ \text{Multinomial}(1, \boldsymbol{\theta}_l) & \text{otherwise.} \end{cases}$$

$$z_{ij\ell} \mid \boldsymbol{\Lambda}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij\ell}), \text{ where}$$
$$p_{ij\ell} = \begin{cases} 1 & \text{if } x_{ij\ell} \neq y_{\lambda_{ij}\ell} \\ \dfrac{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}}}{\beta_\ell \prod_{m=1}^{M_\ell} \theta_{\ell m}^{x_{ij\ell}} + (1 - \beta_\ell)} & \text{if } x_{ij\ell} = y_{\lambda_{ij}\ell}, \end{cases} \text{ for all } \ell.$$

$$P(\lambda_{i1} = c_1, \ldots, \lambda_{in_i} = c_{n_i} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x})$$
$$\overset{\text{ind}}{\propto} \begin{cases} 0 & \text{if there exist } j, \ell \text{ such that} \\ & \quad z_{ij\ell} = 0 \text{ and } x_{ij\ell} \neq y_{c_j\ell}, \\ 1 & \text{otherwise.} \end{cases}$$