Automated Data Imputation: A Versatile Tool in JMP® Pro 14 for Handling Missing Values

Milo Page, PhD

JMP Research Statistician Developer

Missing Values

Imputed_Age	Imputed_BMI	Imputed_BP	Imputed_Total Cholesterol	Imputed_LDL	Imputed_HDL	Imputed_TCH	Imputed_LTG	Imputed_Glucose
59	32.1	101	157	93.2	38	4	4.8598	87
48	21.6	87	183	103.2	70	3	3.8918	69
72	30.5	93	156	93.6	41	4	4.6728	85
24	25.3	84	198	131.4	40	5	4.8903	89
50	23	101	192	125.4	52	4	4.2905	80
23	22.6	89	139	64.8	61	2	4.1897	68
36	22	90	160	99.6	50	3	3.9512	82
66	26.2	114	255	185	56	4.55	4.2485	92
60	32.1	83	179	119.4	42	4	4.4773	94
29	30	85	180	93.4	43	4	5.3845	88
22	18.6	97	114	57.6	46	2	3.9512	83

- Missing Values are a common occurrence
- Nearly all predictive models require complete data
- Data imputation replaces missing values with estimates

Outline

Introducing ADI: An Automated, Streaming Imputation Method

- The Many Uses of Streaming Imputation
- Recommended Workflow for Predictive Modeling in JMP Pro
- Automated Data Imputation
- JMP Pro Demos

Streaming Imputation

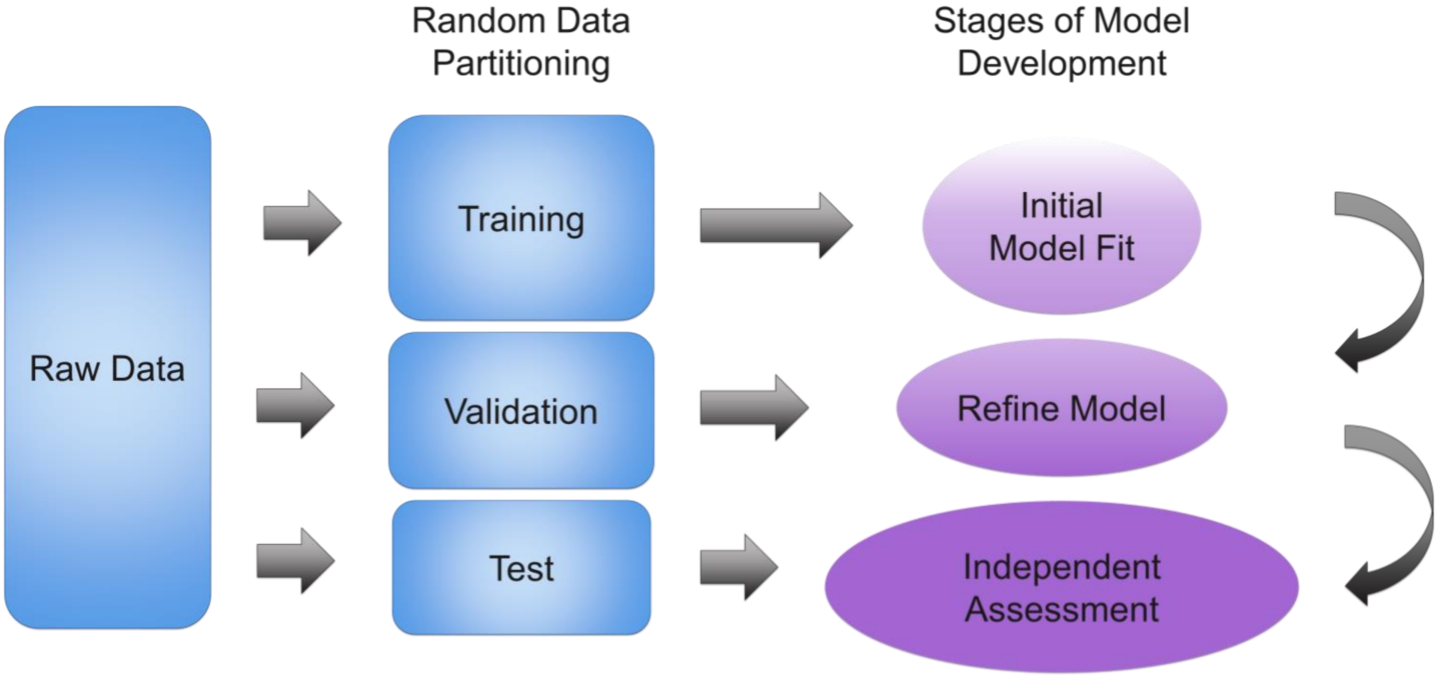
What Is It?

- A method for imputing missing values for a new row of data without refitting the entire imputation model

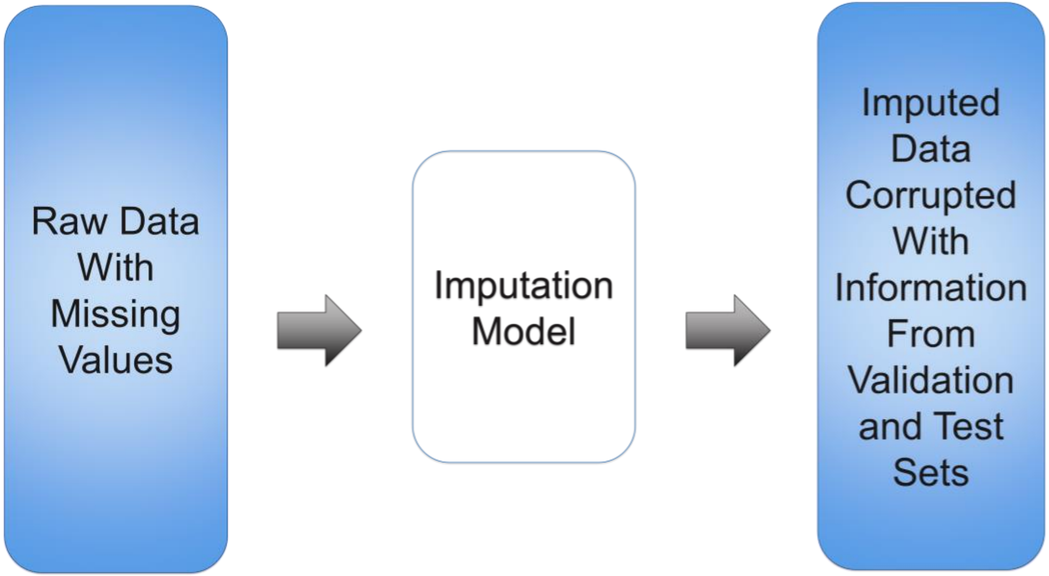
Why Use It?

- For streaming data with missing values
  - E.g., data from manufacturing process that comes available over time
- To deploy an analysis conducted on a sample of a distributed data system
  - E.g., very large customer sales data sets
- To maintain integrity of training/validation partitioning when fitting a prediction model using the imputed data
  - This applies even when the data are not streaming!
  - Most imputation methods pre-date holdout set validation techniques

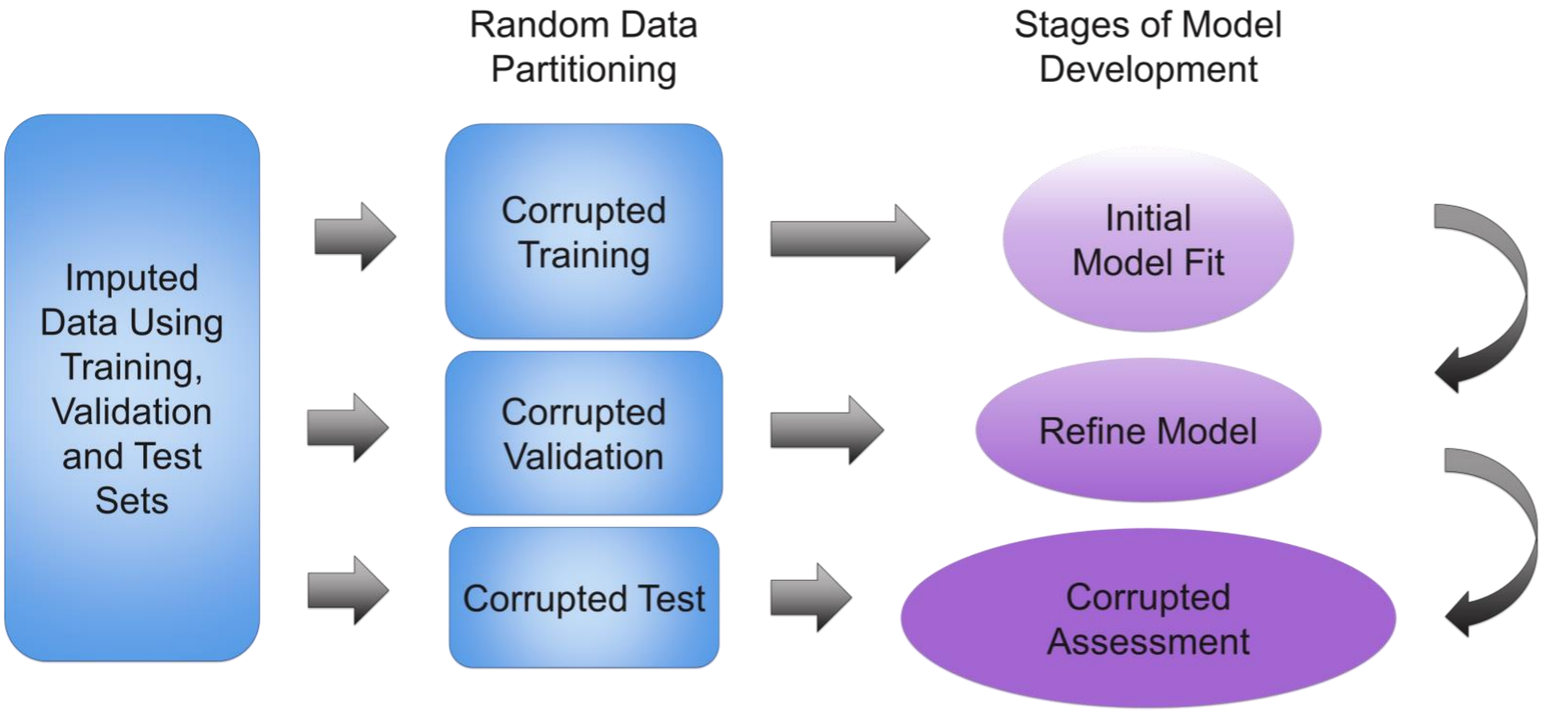
Holdout Validation Sets (Without Missing Values)



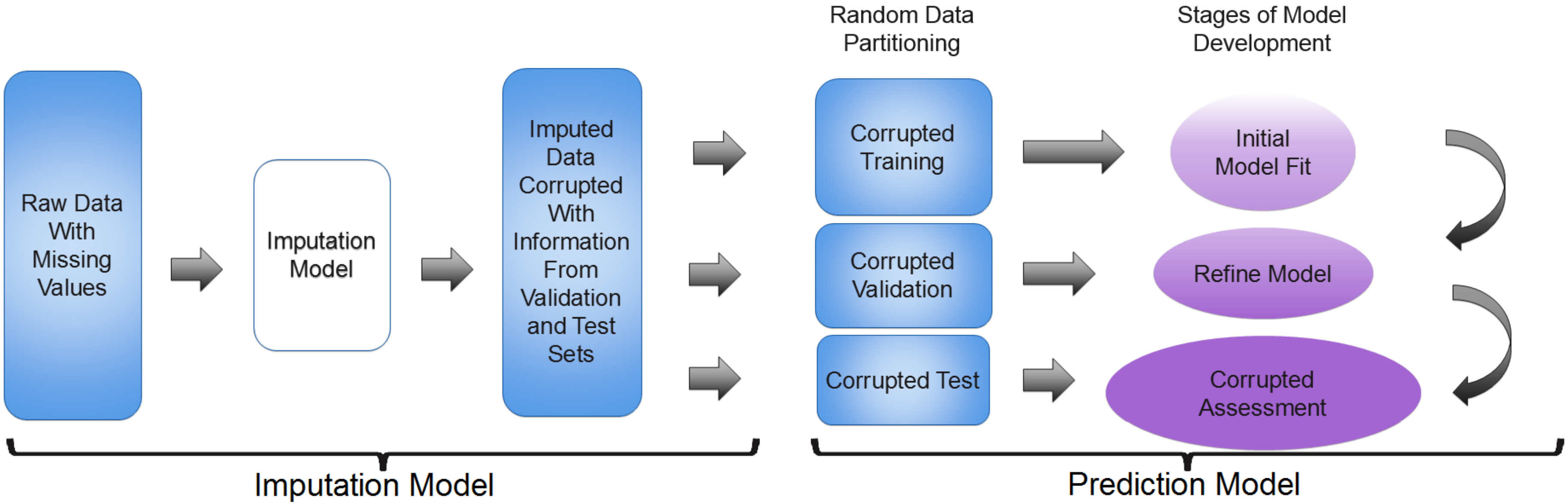
Classical Imputation Approach



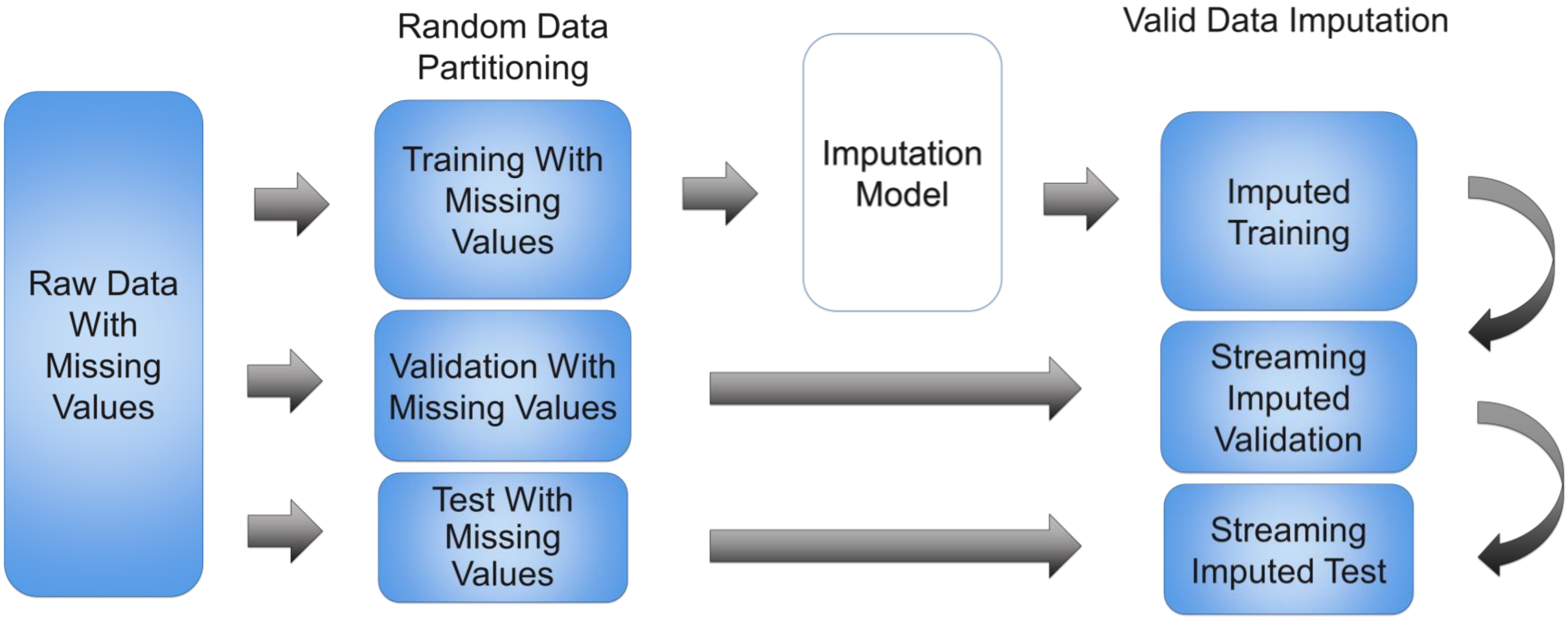
Prediction Model With Classical Imputation



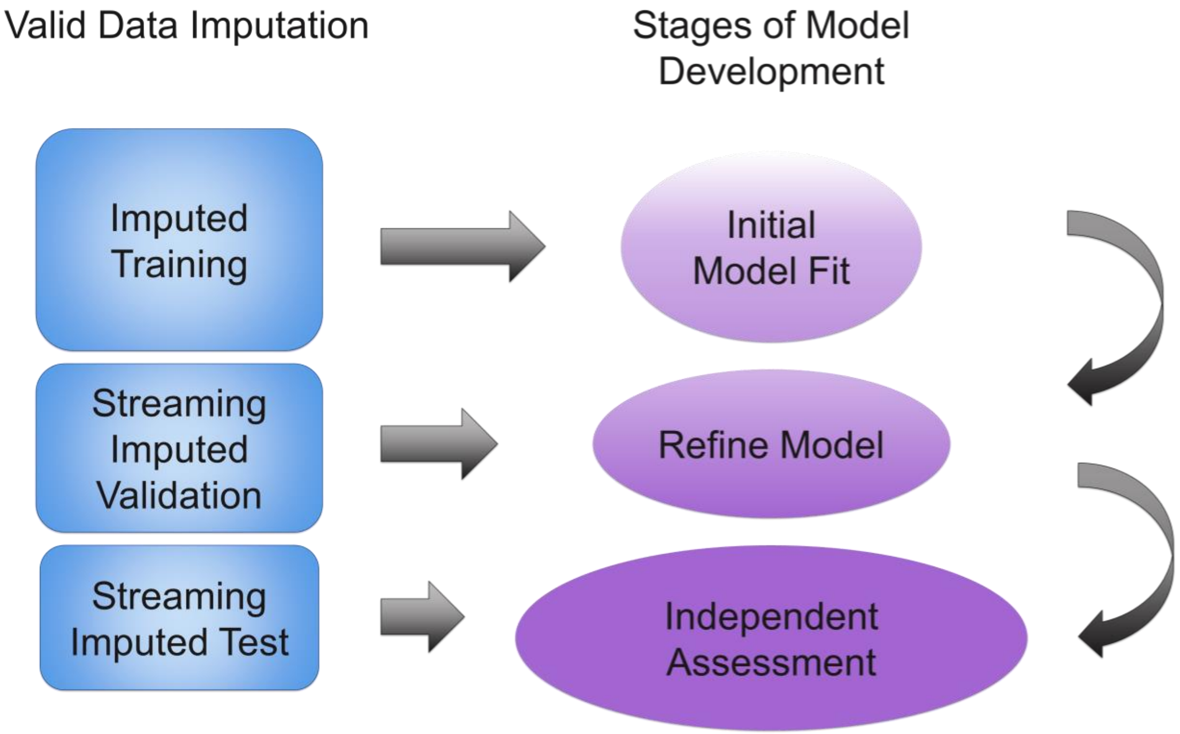
Classical Imputation And Prediction Model Pairing



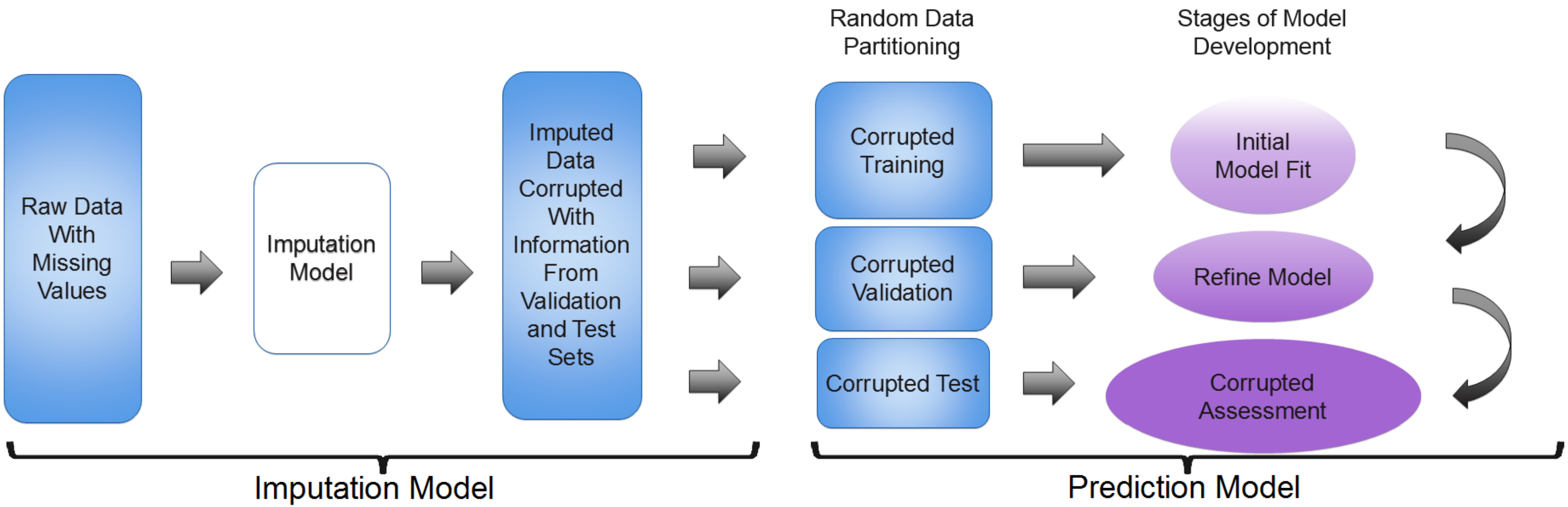
Streaming Imputation Approach



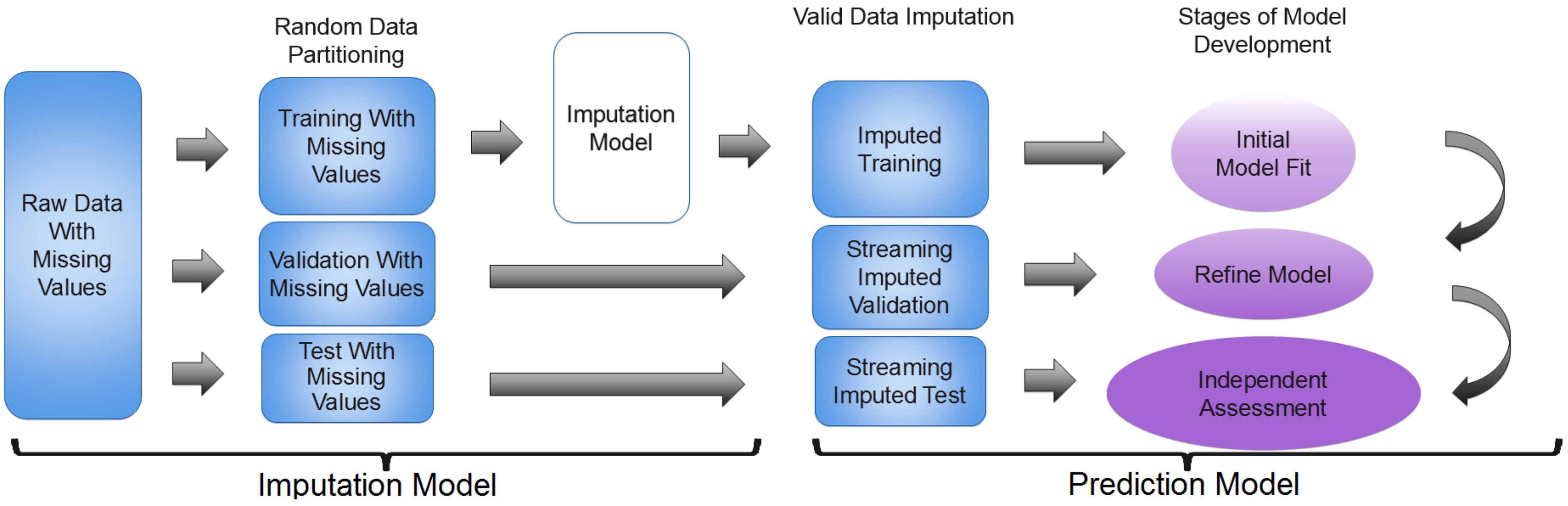
Prediction Model With Streaming Imputation



Classical Imputation And Prediction Model Pairing

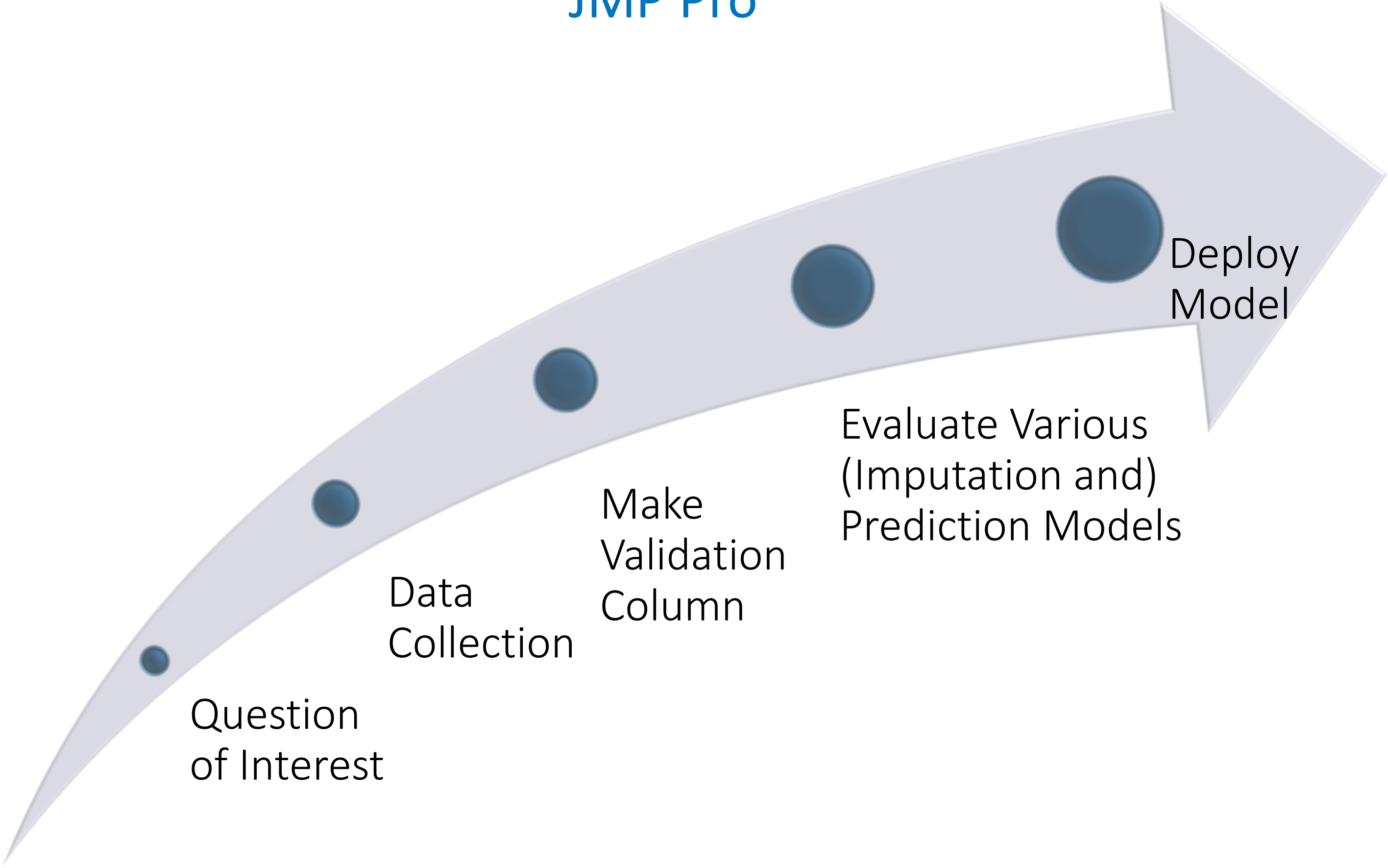


Streaming Imputation And Prediction Model Pairing



Recommended Workflow for Predictive Modeling

JMP Pro



Automated Data Imputation

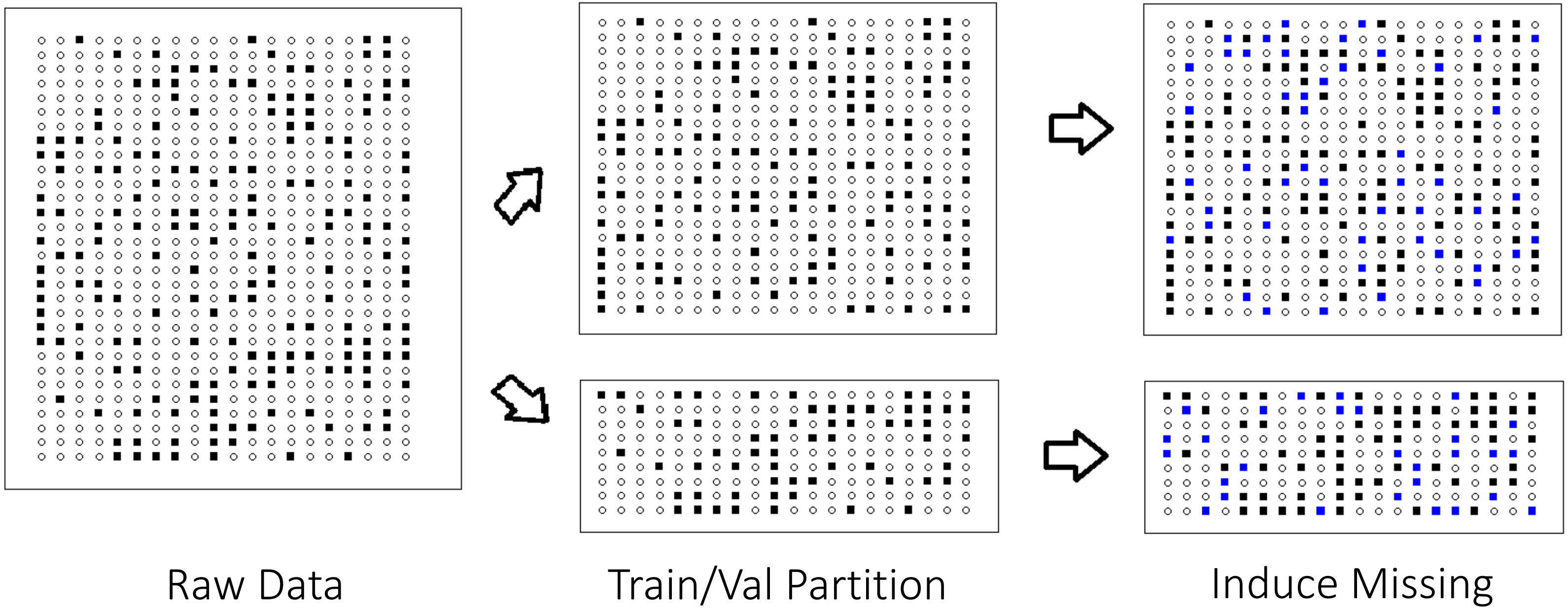
Methodology Overview

- ADI is an extension to Matrix Completion methods
- Matrix completion:
  - Gained popularity with the Netflix challenge
  - Relies on a low rank assumption: $X = UV' + \epsilon$
  - Handles high degree of sparsity
  - Useful when covariates are not independent
  - Solves:

$$\begin{aligned}&\text{minimize} \quad \operatorname{rank}(\hat{X}) \\ &\text{subject to} \sum_{\Omega}(X_{ij} - \hat{X}_{ij})^2 = \|P_{\Omega}(X - \hat{X})\|_F^2 \leq \delta\end{aligned}$$

Automated Data Imputation

Automated Dimension Selection by Induced Missing Values



Automated Data Imputation

Streaming Extension

- From training partition, calculate $\hat{X} = UV'$
- For new rows, use latent column structure, V, to estimate missing values
- Need to estimate the new row of U:

$$\hat{\mathbf{u}}_i = \underset{\mathbf{u}=[u_1,\ldots,u_k]'}{\operatorname{argmin}} ||P_{\mathbf{\Omega}_i}(X_{i.}) - (P_{\mathbf{\Omega}_i}(V))\mathbf{u}||_F$$

- This is estimated using the observed elements within each row
- More details:

Page, M. & Gotwalt, C. & Wilson, A. G. (2018). Automated Data Imputation: Extending Low Rank Matrix Imputation Techniques For Statistical Prediction Modeling. https://repository.lib.ncsu.edu/handle/1840.20/35520

Highlights of ADI

- Automatically fits to data
  - Capable of fitting complicated (high dimensional) low rank structure
  - If no structure found, resorts to column mean
- Handles high degrees of sparsity
- Unlike other imputation methods, the more columns the better
  - Assuming they provide some information on the other covariates
- Integrates seamlessly with prediction model using validation column
- Formula columns dynamically handle imputation for new rows of data

Demo

- Using JMP Pro 14.2