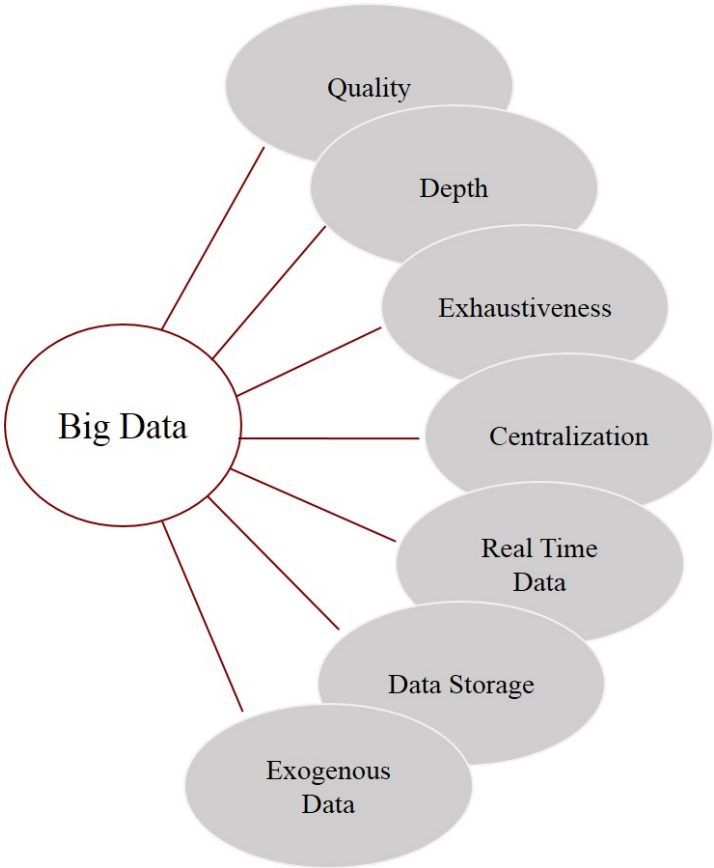
The Future Marriage of Big Data and Railroad Engineering

Nii Attoh-Okine, PhD, P.E.
Professor
Department of Civil and Environmental Engineering

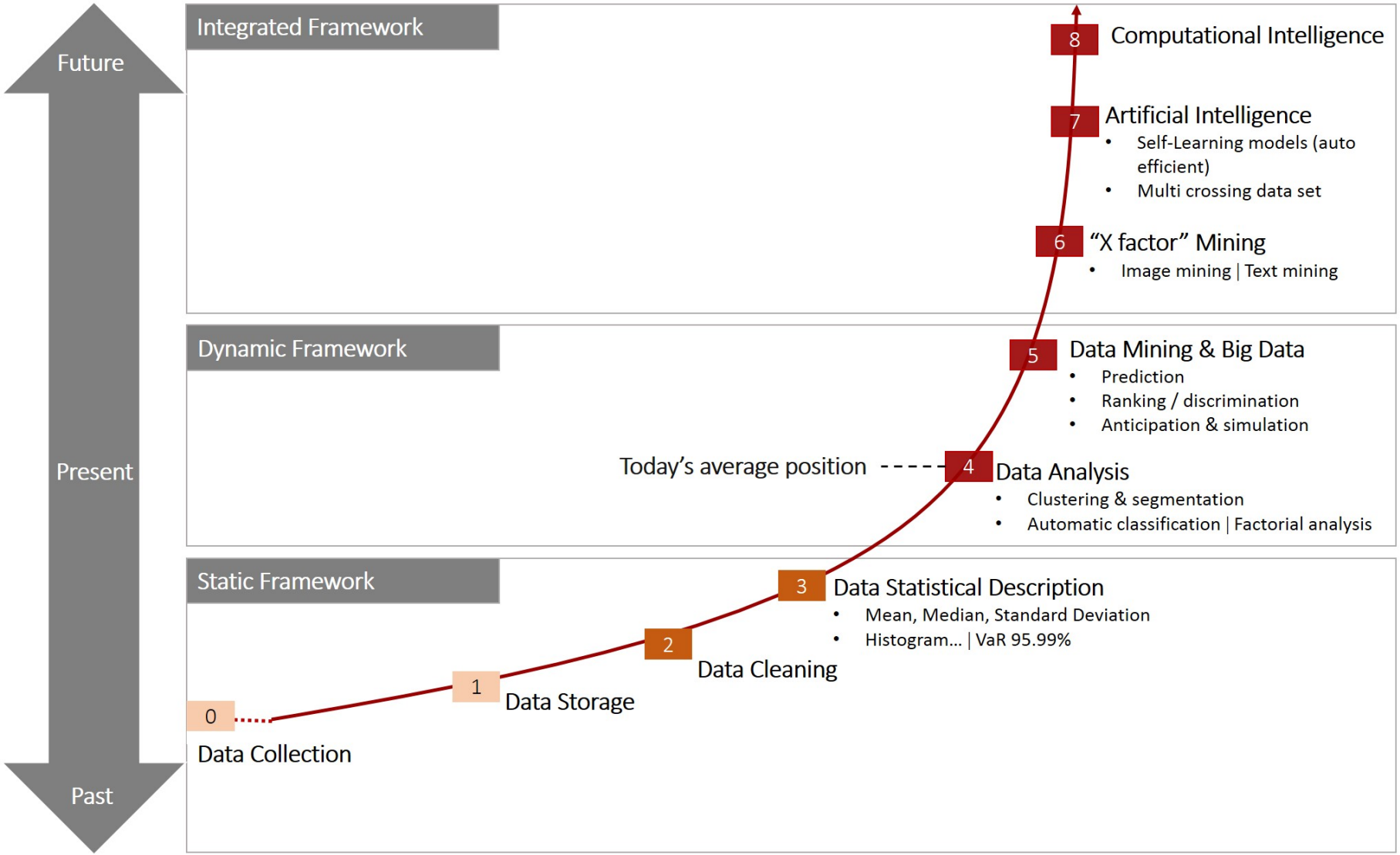
"The challenge is utilizing Big Data to improve efficiency, reliability, velocity, productivity, and safety." - *Railway Age. August, 2014.*

Big Data Revolution is in Motion in the Railway Industry

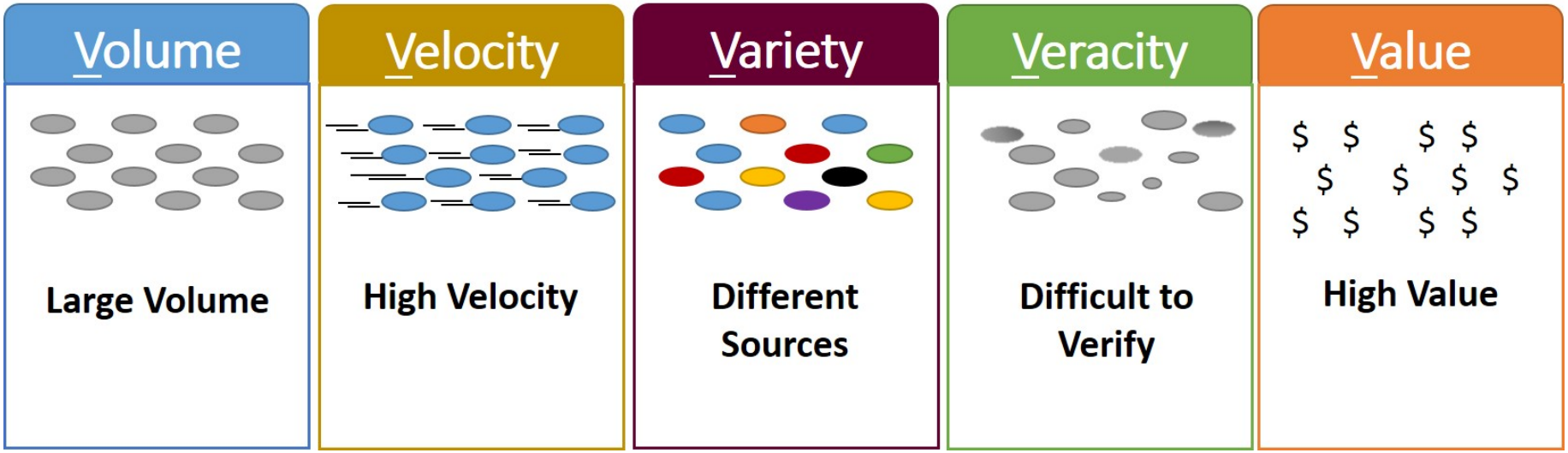


- Given recent evolution in the railway industry, remaining competitive requires an ability to process a growing flow of data.
- We strongly believe that the keys to success are twofold when it comes to designing a successful Big Data strategy:
  - A structured and robust framework.
  - A continuous upgrade of hardware and infrastructure to stick to volume of data and complexity of analyses.
- Big Data represents business opportunities for major players in the railway industry...
- ... from rethinking client relationship to crafting tomorrow's operations and risk management.

Big Data Requires Investing in New Forms of Data Processing



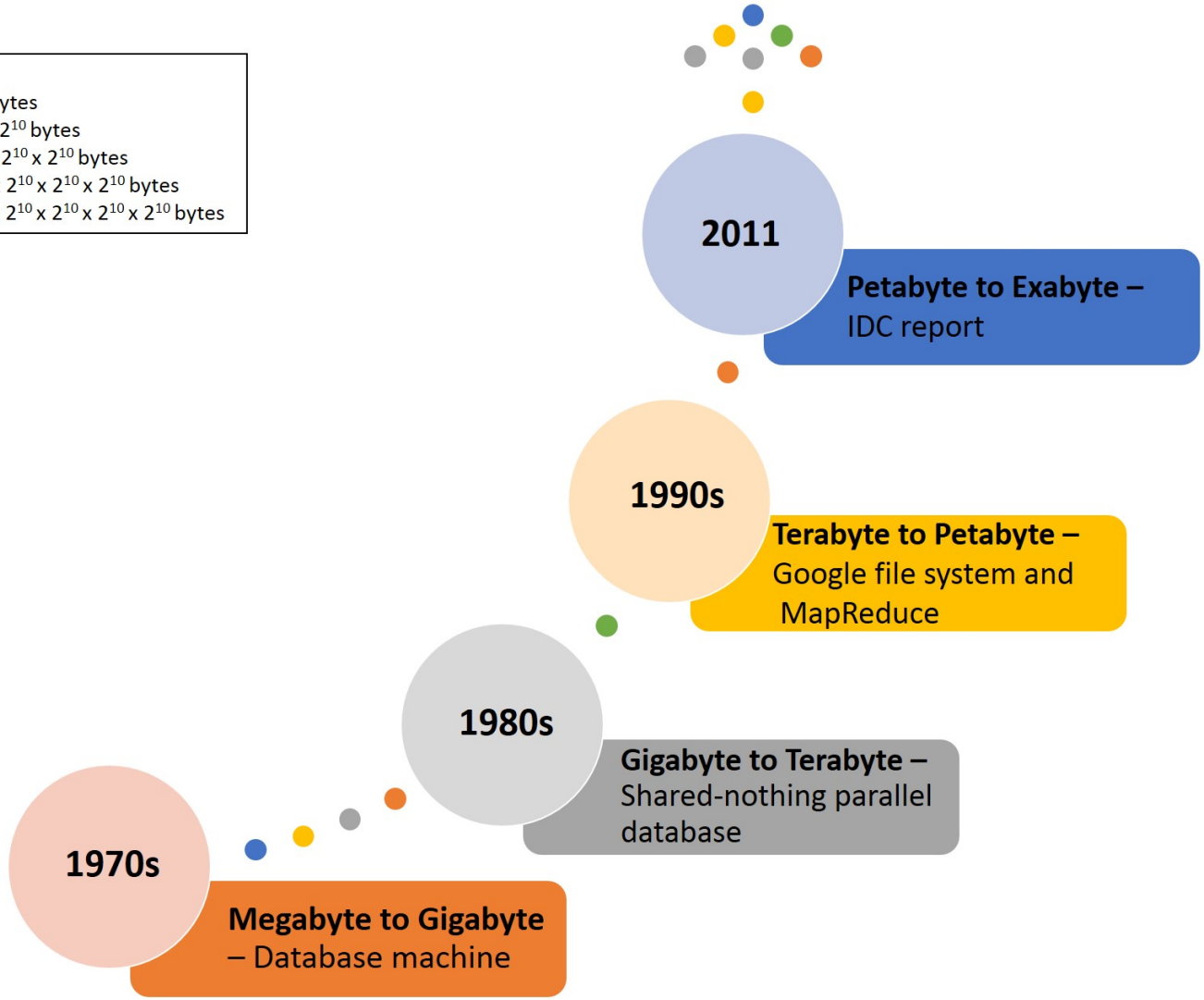
Big Data and its 5V Properties



Learning for Big Data

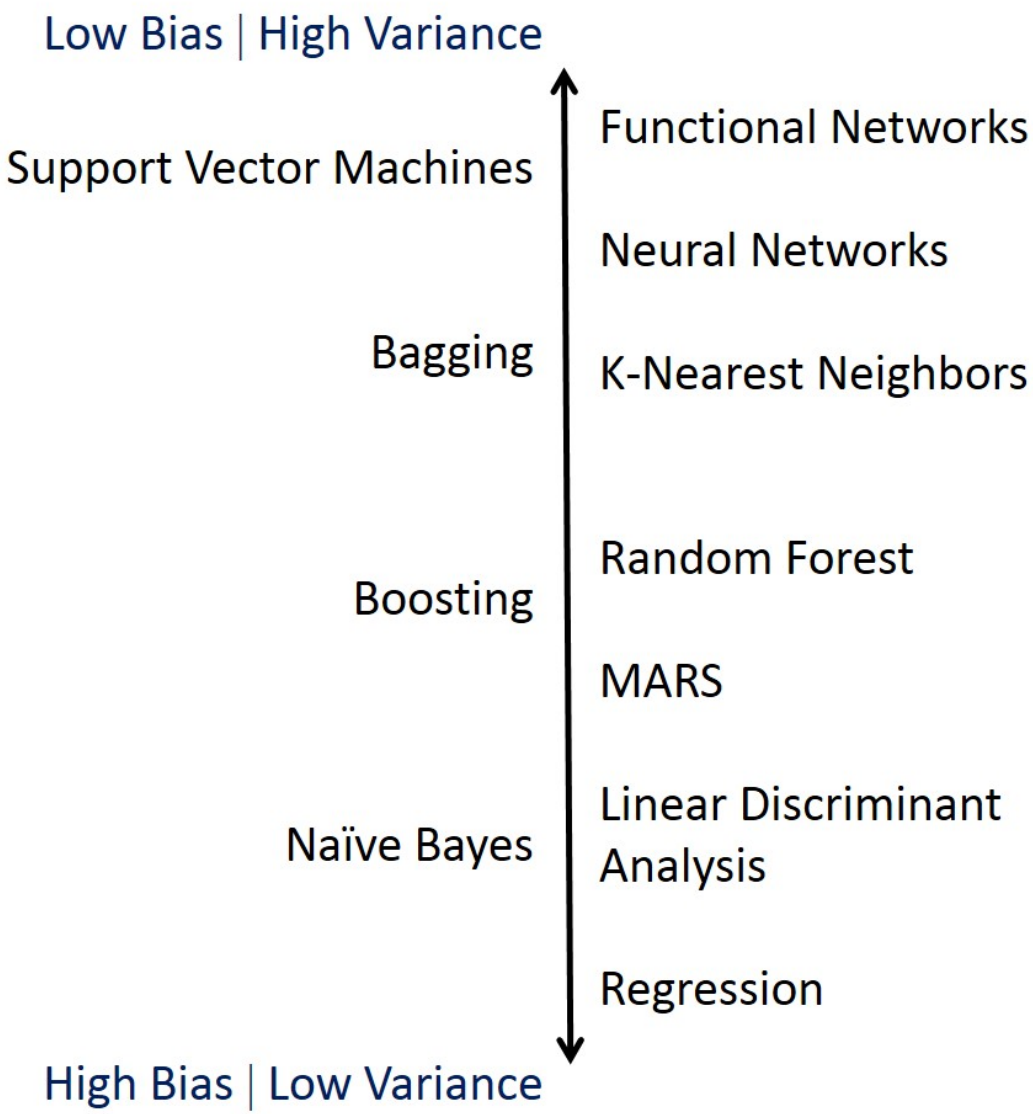
A Brief History of Big Data with Major Milestones

kilobyte (KB)	$= 2^{10}$ bytes
megabyte (MB)	$= 2^{10} \times 2^{10}$ bytes
gigabyte (GB)	$= 2^{10} \times 2^{10} \times 2^{10}$ bytes
terabyte (TB)	$= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ bytes
petabyte (PB)	$= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ bytes
exabyte (EB)	$= 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$ bytes



- As data sets gets larger, complexity of false findings grow exponentially.
- Serious statistical skill is required to avoid being misled.

- Large N, Small P.
- Small N, Large P.
    - Use of statistical significance is inappropriate.

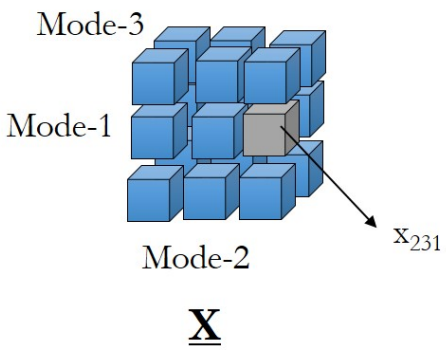

Low Bias | High Variance
Support Vector Machines
Bagging
Boosting
Naïve Bayes
Functional Networks
Neural Networks
K-Nearest Neighbors
Random Forest
MARS
Linear Discriminant Analysis
Regression
High Bias | Low Variance

Bias and Variance in Statistics

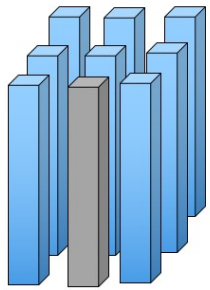
- Bias is error that cannot be corrected by repeated experiments.
- Bias-variance decomposition states that expected squared error is equal to the bias plus the random error.
- You can reduce the variance but not the bias.
- True value of the parameter is a constant.
- Experimental estimate is a probabilistic variable.
- Bias is the systematic or average difference between these two variables and variance is the probabilistic component.

Multiway Data Analysis

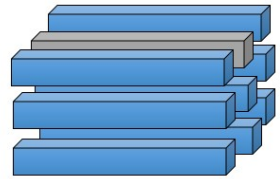
Typical 3D dataset



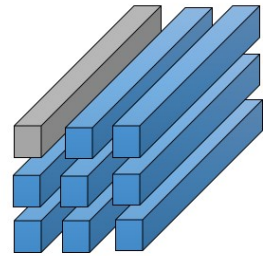
Subarrays



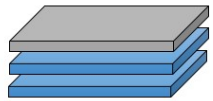
Mode-1 Fiber
$x_{:21}$



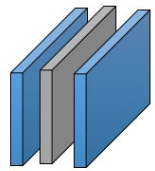
Mode-2 Fiber
$x_{1:2}$



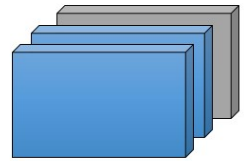
Mode-3 Fiber
$X_{11:}$



Horizontal slice
$X_{1::}$



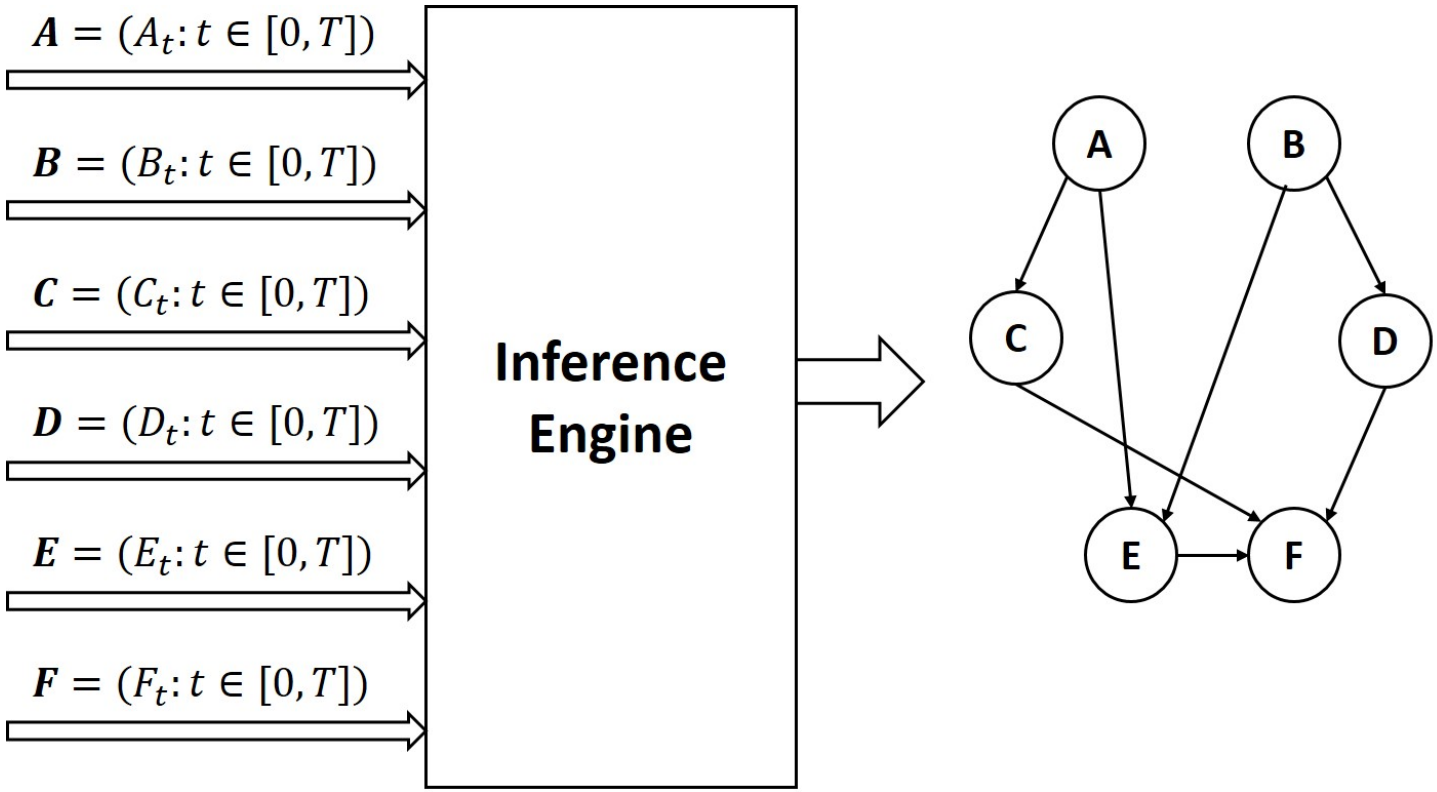
Lateral slice
$X_{:2:}$



Frontal slice
$X_{::3}$

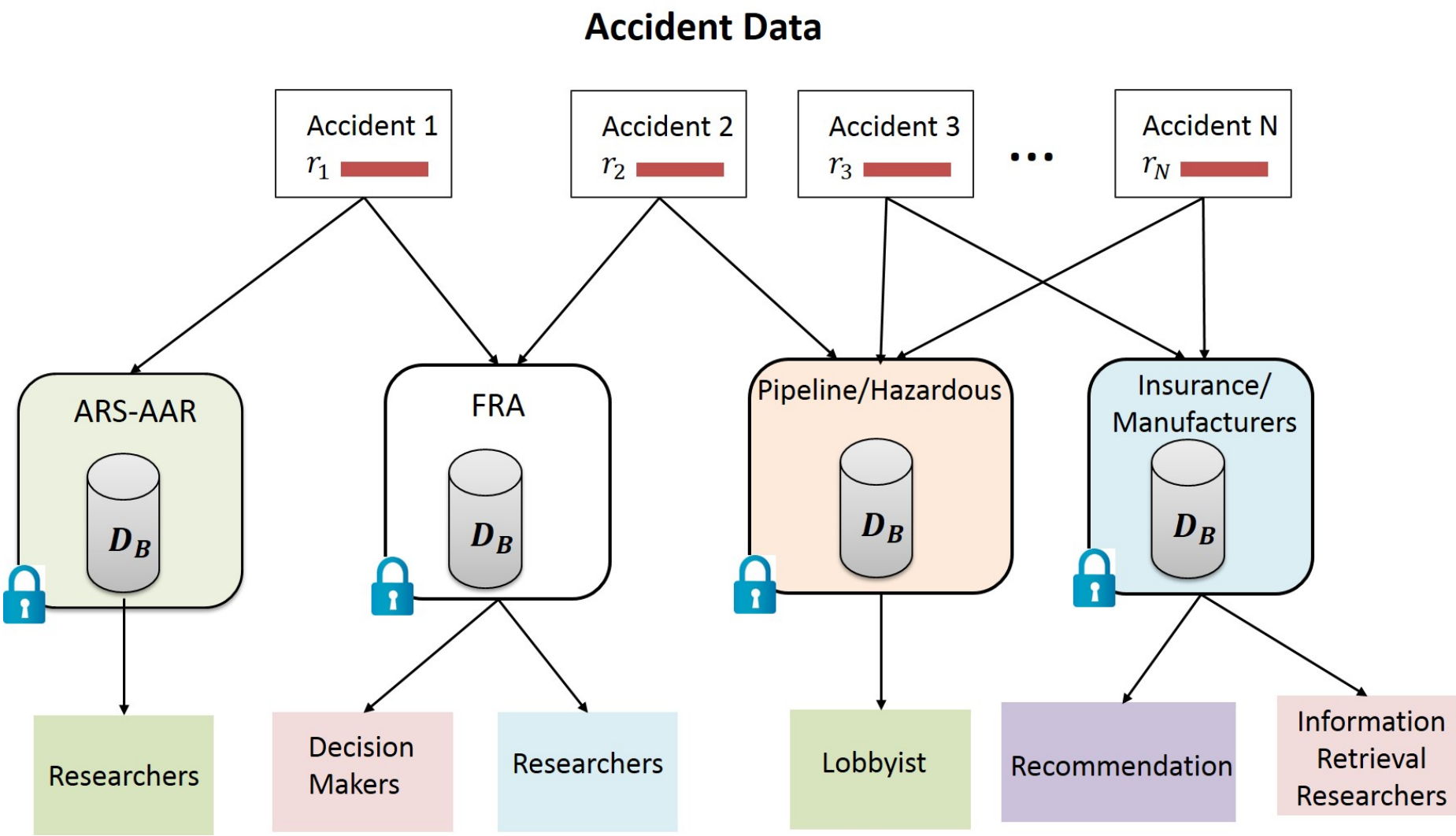
4

Beyond Correlation: Causation



Idea: Map a set of K time series to a directed graph with K nodes where an edge is placed from a to b if the past of a has an impact on the future of b

Massive Amount of Data (Tank Safety)



Traditional analytical techniques are inadequate in analyzing and drawing conclusions