The Future Marriage of Big Data and Railroad Engineering

Nii Attoh-Okine, PhD, P.E., F.ASCE
Interim Academic Director, Cybersecurity Initiative
Professor
Department of Civil and Environmental Engineering
Department of Electrical and Computer Eng
University of Delaware
Newark, DE, USA

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

Railway Engineering

- 3 terabytes of data are generated in a year for bearing temperature detectors for Class I US railroad.
- Data collected from different sensors need to be integrated for predicting track failures.

Learning to Predict Train Wheel Failures

Goal: Optimize maintenance and operation of trains.

Approach: Decision Trees and Naïve Bayes.

A Simple and Efficient Parallel Approach to Large-Scale Railway Freight Data Analysis

Goal: Integrate the national-widely distributed railway freight data sets.

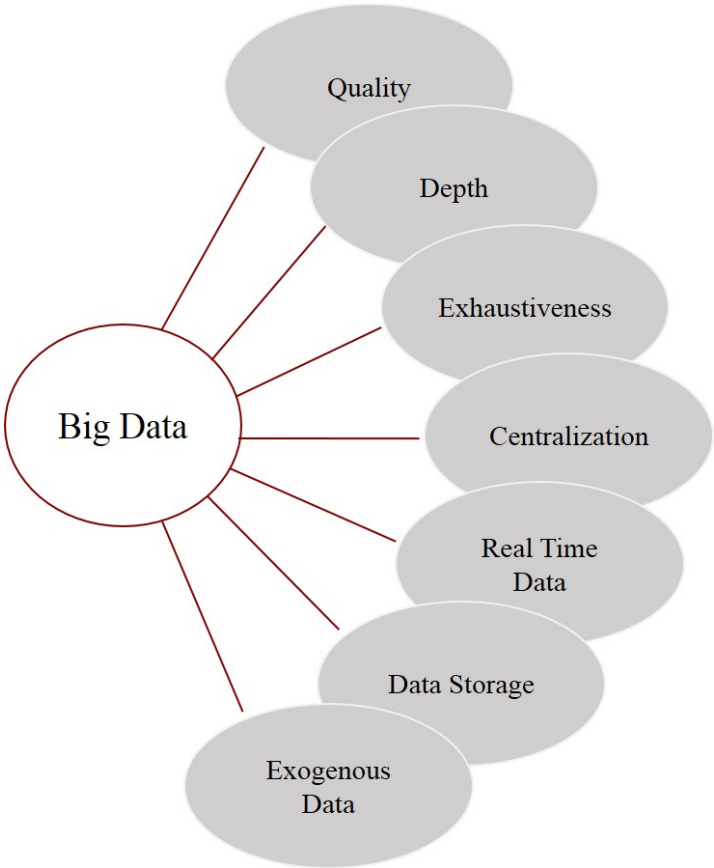
Approach: Parallel optimization techniques.

Improving rail network velocity: A machine learning approach to predictive maintenance

Data from disparate sources: Historical detector data, failure data, maintenance action data, inspection schedule data, train type data and weather data.

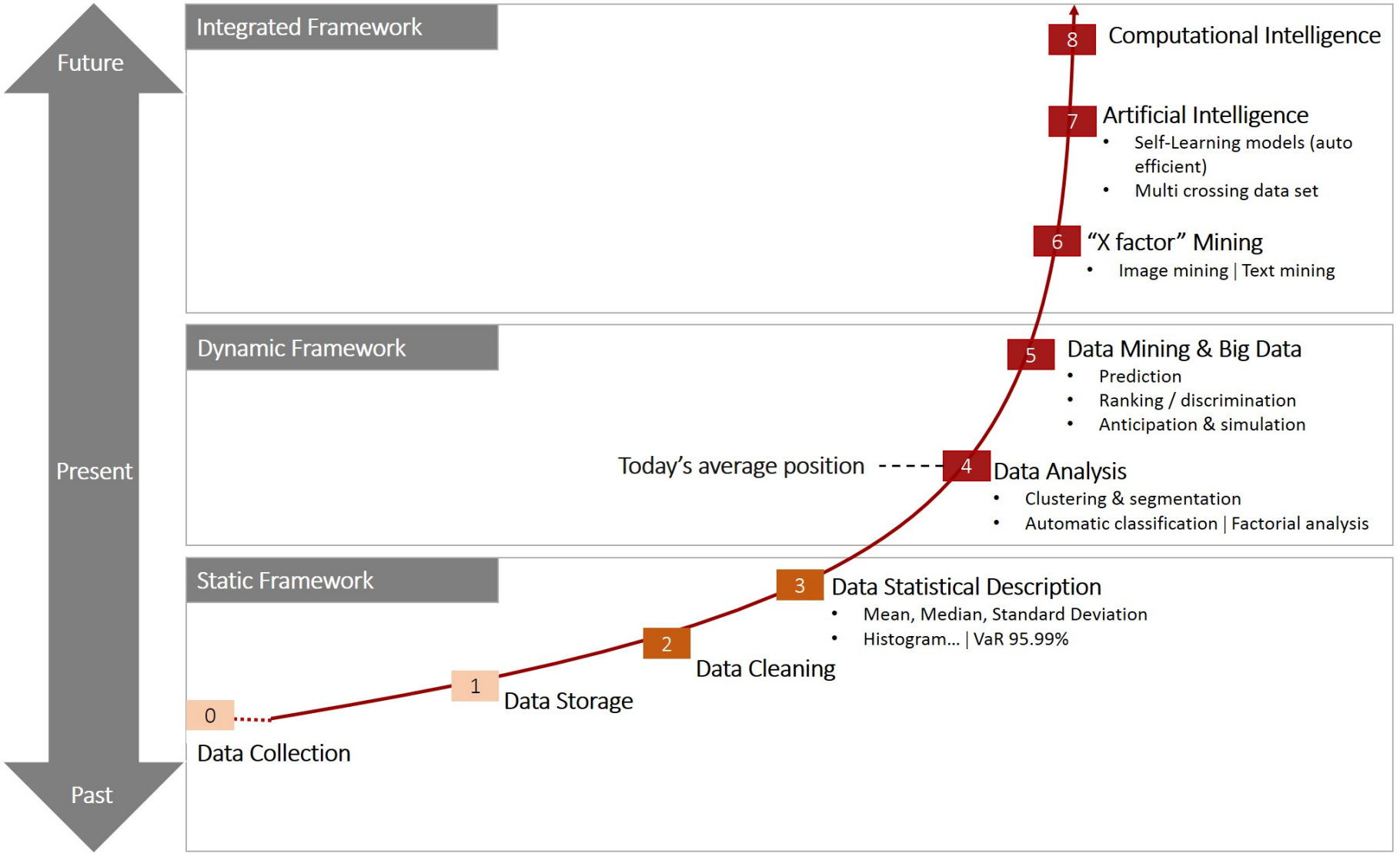
Approach: Principal Component Analysis, Support Vector Machine.

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.

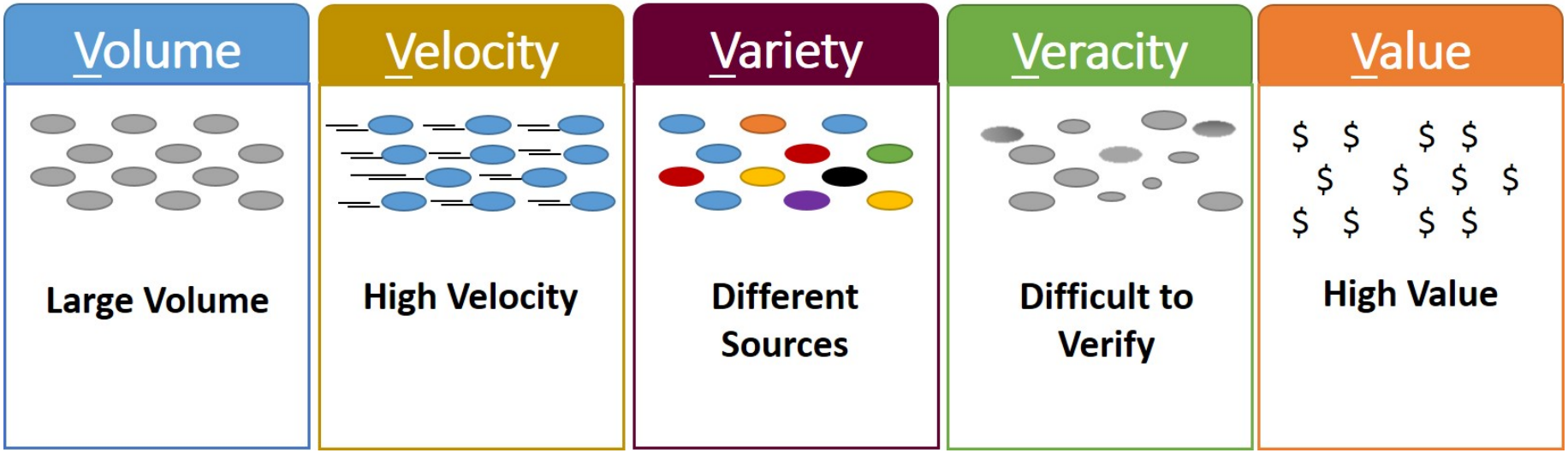
Big Data Requires Investing in New Forms of Data Processing



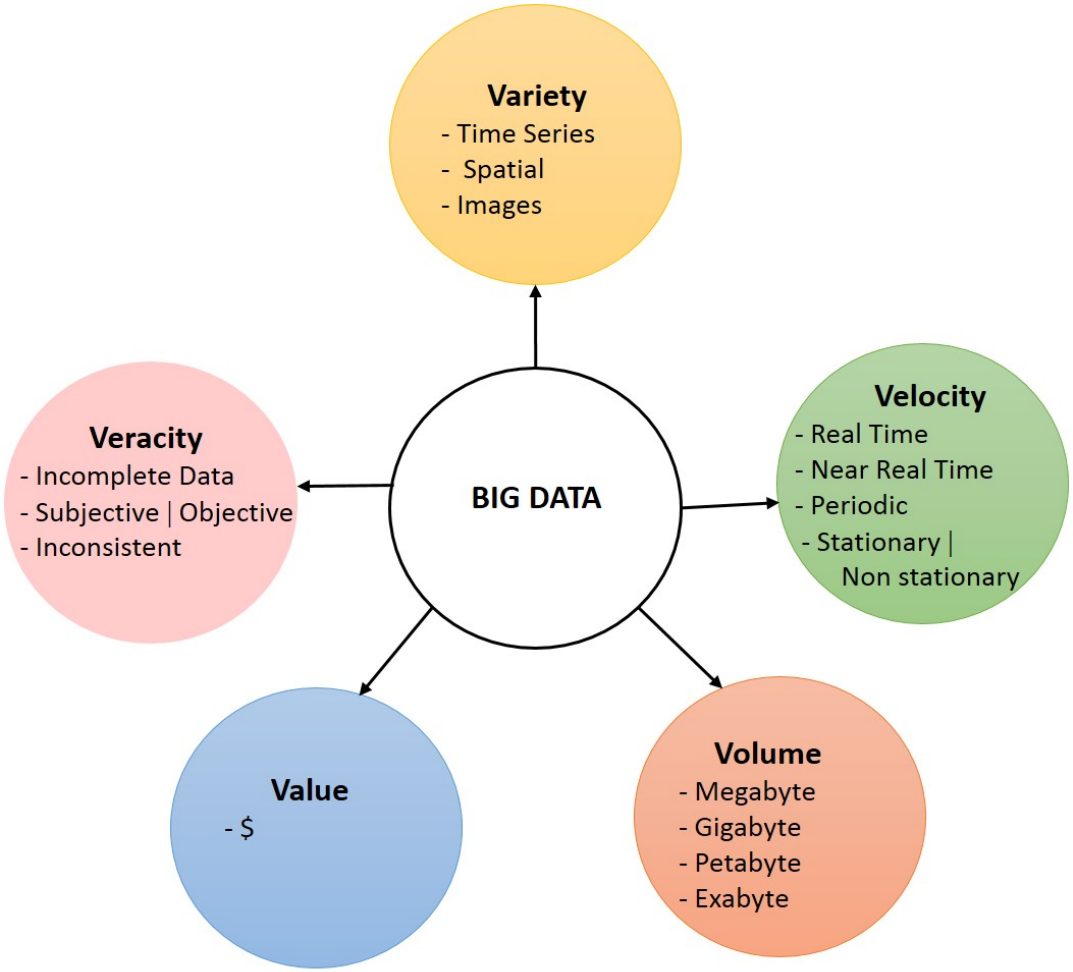
Rail Track Geometry Defects

Divisions	Length in Miles	Annual MGT	Reported Geo Defects	Unique Geo Defects	Geo Defect Per Mile	Rail Defects	Rail Defects Per Mile	Matches	Matches Per Mile
Baltimore	1042.79	24.05	18351	11462	10.992	1543	1.480	204	0.196
Atlanta	2042.1	25.68	27536	16044	7.857	2285	1.119	150	0.073
Albany	1871.05	25.02	22738	14645	7.827	1774	0.948	157	0.084
Appalachian	2221.24	17.29	49616	26383	11.878	4786	2.155	763	0.344
C&O	2051.42	19.03	61677	39658	19.332	2374	1.157	496	0.242
Chicago	1630.73	18.65	16978	10053	6.165	1464	0.898	128	0.078
Florence	3056.51	15.37	38785	22457	7.347	4019	1.315	285	0.093
Great Lakes	2326.16	32.8	18485	13202	5.675	2520	1.083	195	0.084
Jacksonville	2920.07	14.97	28308	15727	5.386	2272	0.778	178	0.061
Louisville	1406.79	17.24	31123	17435	12.393	1657	1.178	156	0.111
Nashville	1658.86	30.61	21340	15275	9.219	1746	1.054	206	0.124
Total	22227.72	21.33	334937	202341	9.103	26440	1.190	2918	0.131

Big Data and selected 5V Properties (10 Vs)



5 V's of Big Data



Variety

Structured and Semi-structured data.

- Sensor data
- Audio
- Video
- Combination of such data for efficient analysis.

Velocity

Increasing speed at which data is created, and the speed at which it can be processed, stored, and analyzed.
E.g: Track data collection, Traffic streams.

Volume

It is defined as the amount of data generated in a period of time (e.g. Gigabytes, Terabytes) from different sources, locations to be stored and processed.

Value

Mainly refers to the worth of the information being managed.

Veracity

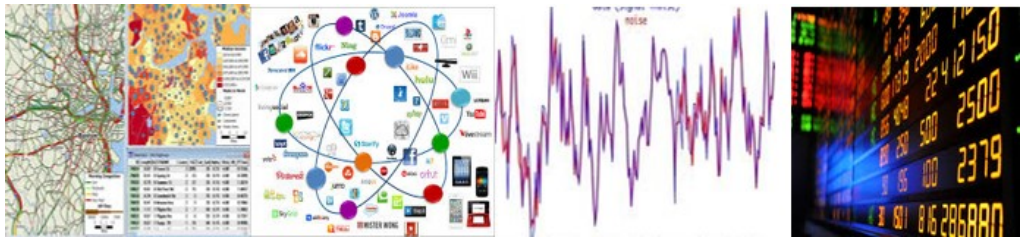
Refers to the trustworthiness of the data. It is an indication of the reliability and accuracy of information for making decisions.

Data from the Physical World

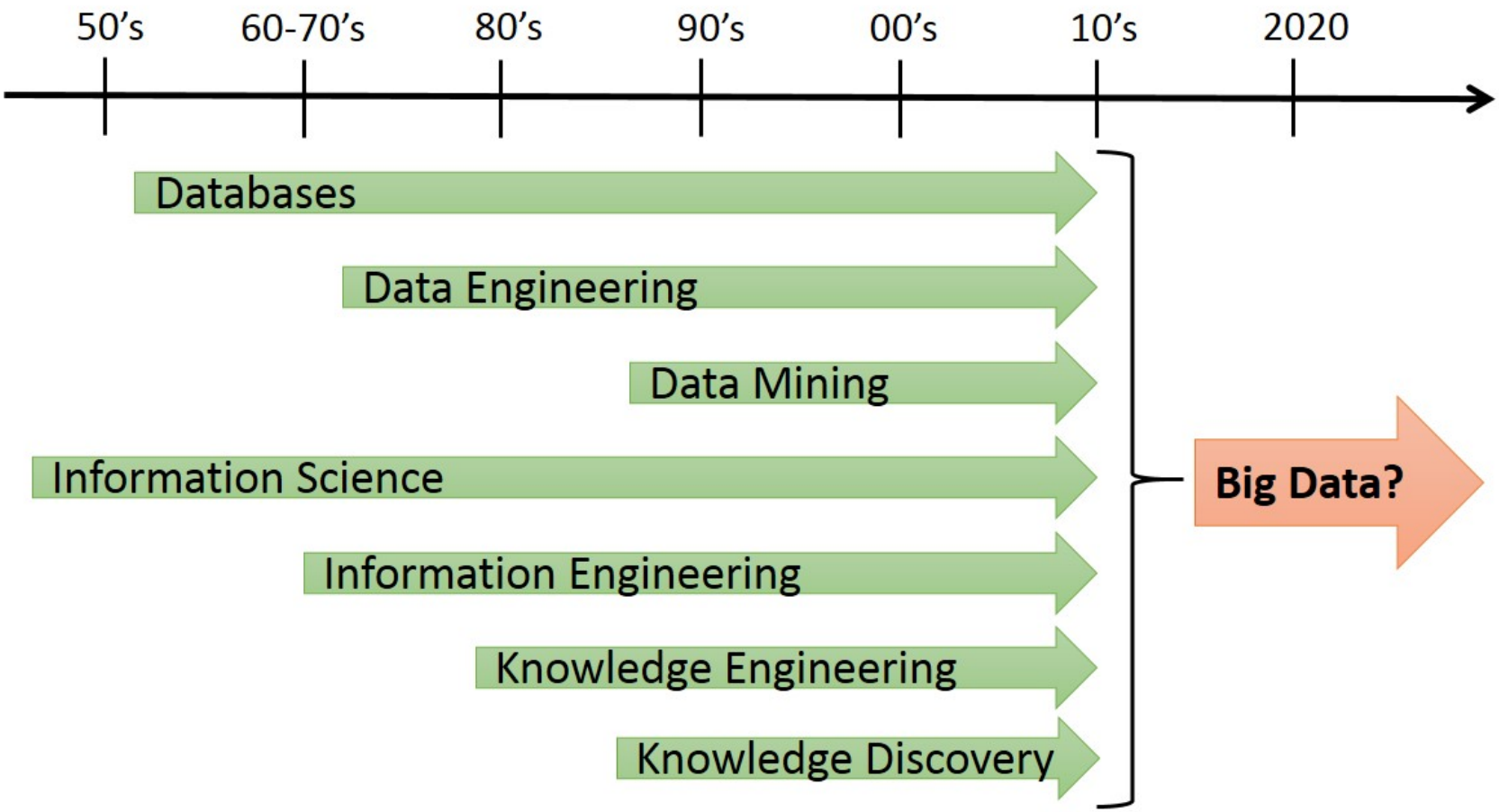
- Obtained through sensors, scientific experiments & observations.

Data from Human Activities

- Obtained through social networks, internet, health, finance, economics, transportation, etc.

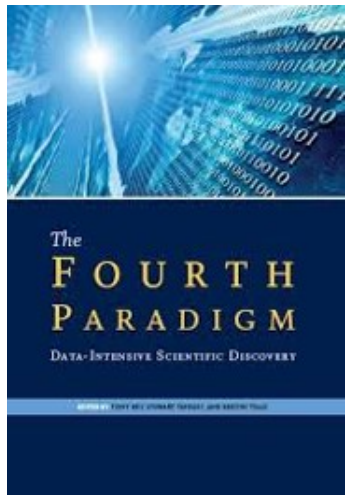


Evolution of Big Data

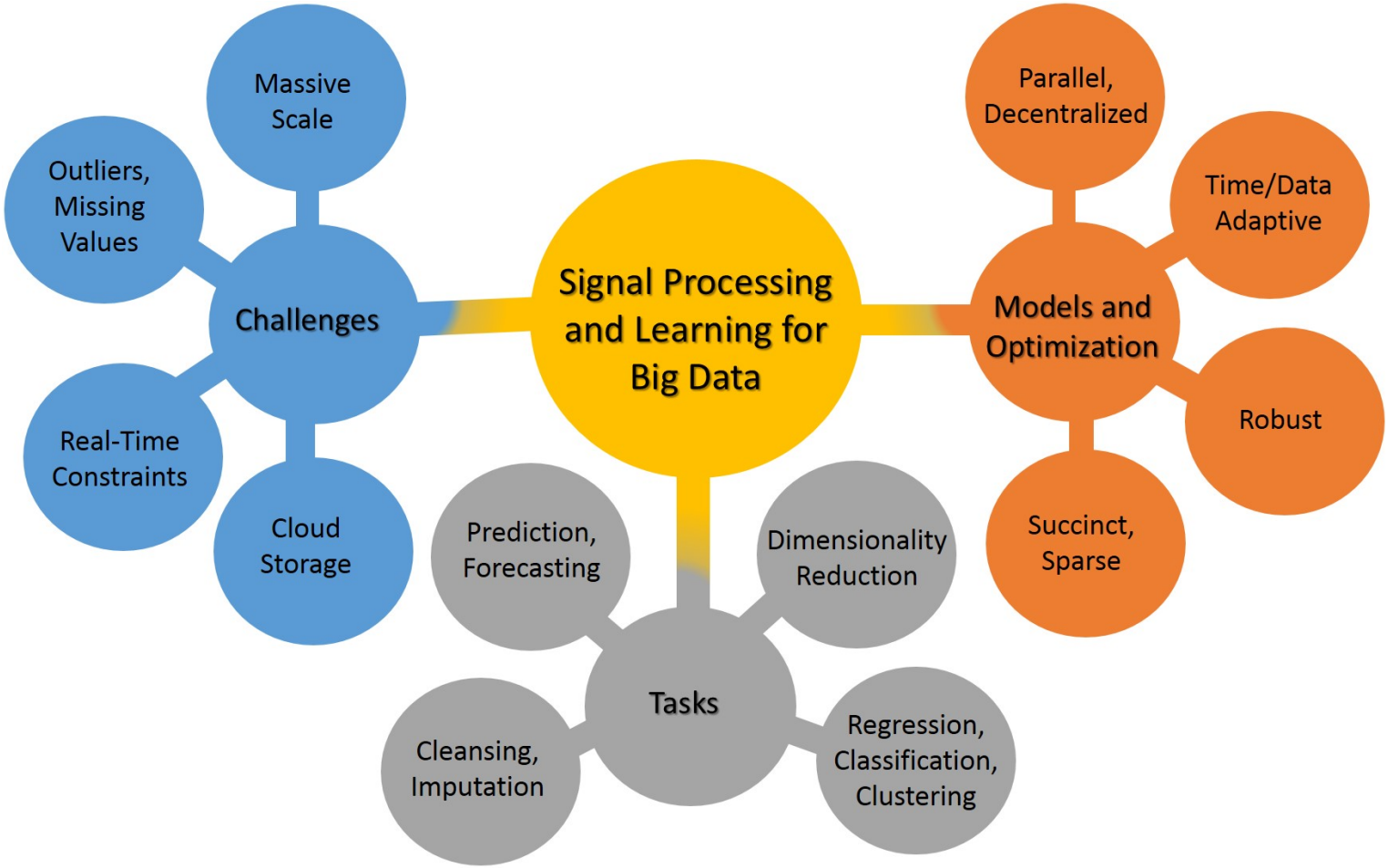


Jim Gray's Fourth Paradigm

- Thousand years ago.
  - Experimental Science - description of natural phenomena.
- Last few hundred years.
  - Theoretical Science - Newton's Laws, Maxwell's Equations ...
- Last few decades.
  - Computational Science - simulation of complex phenomena.
- Today
  - Data Intensive Science - from hypothesis-driven to data-driven.



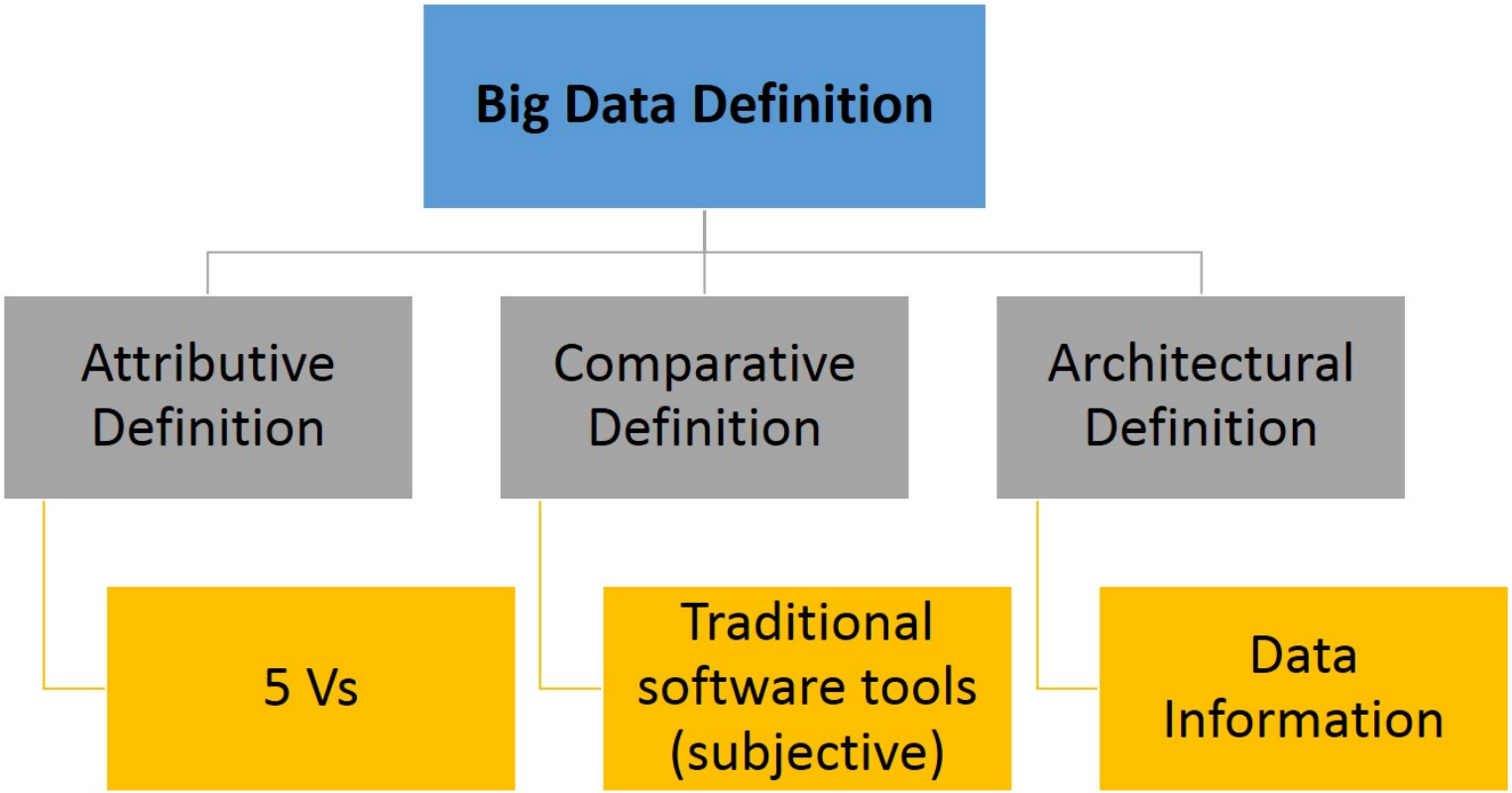
Learning for Big Data



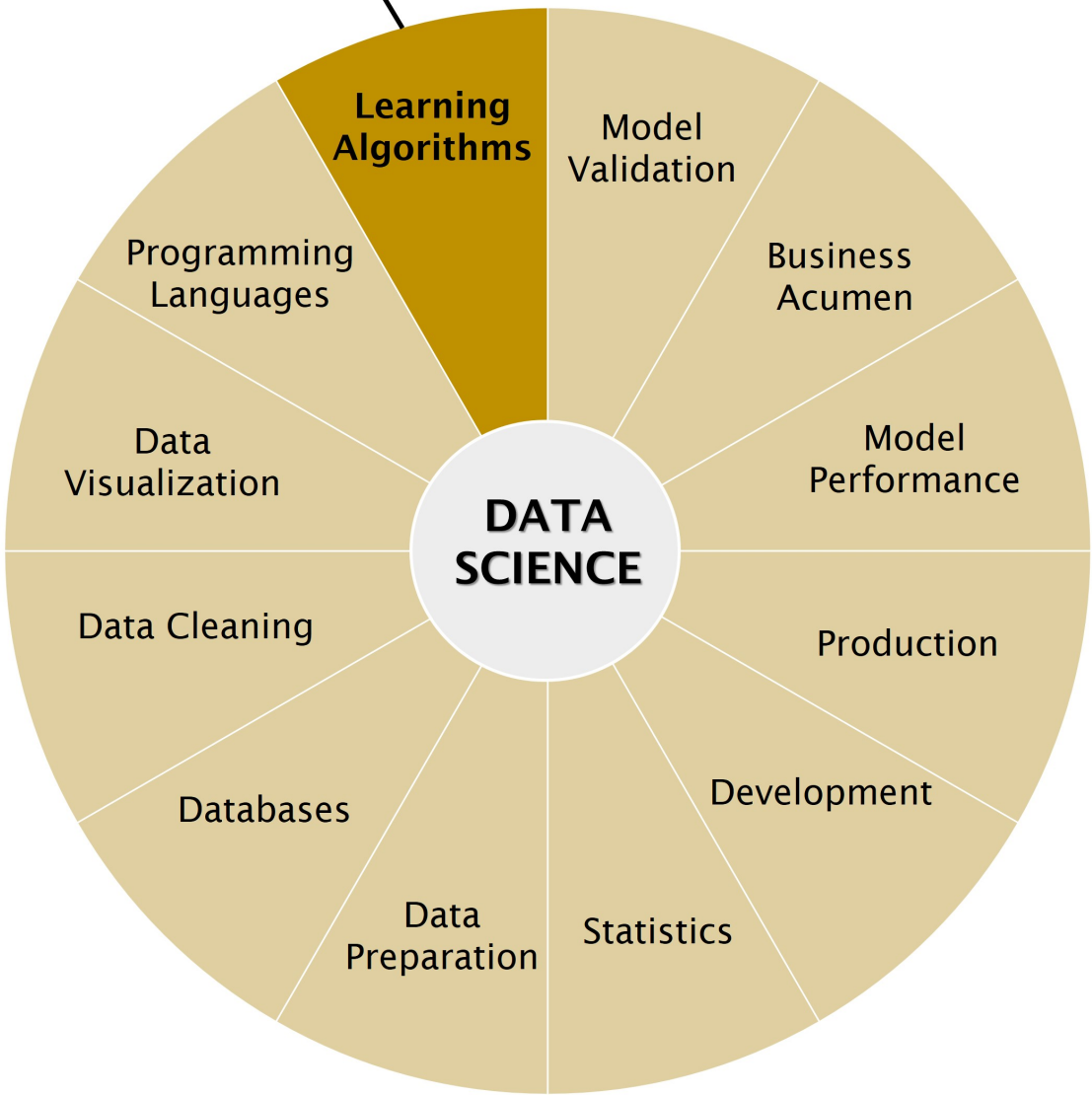
7D of Big Data Research

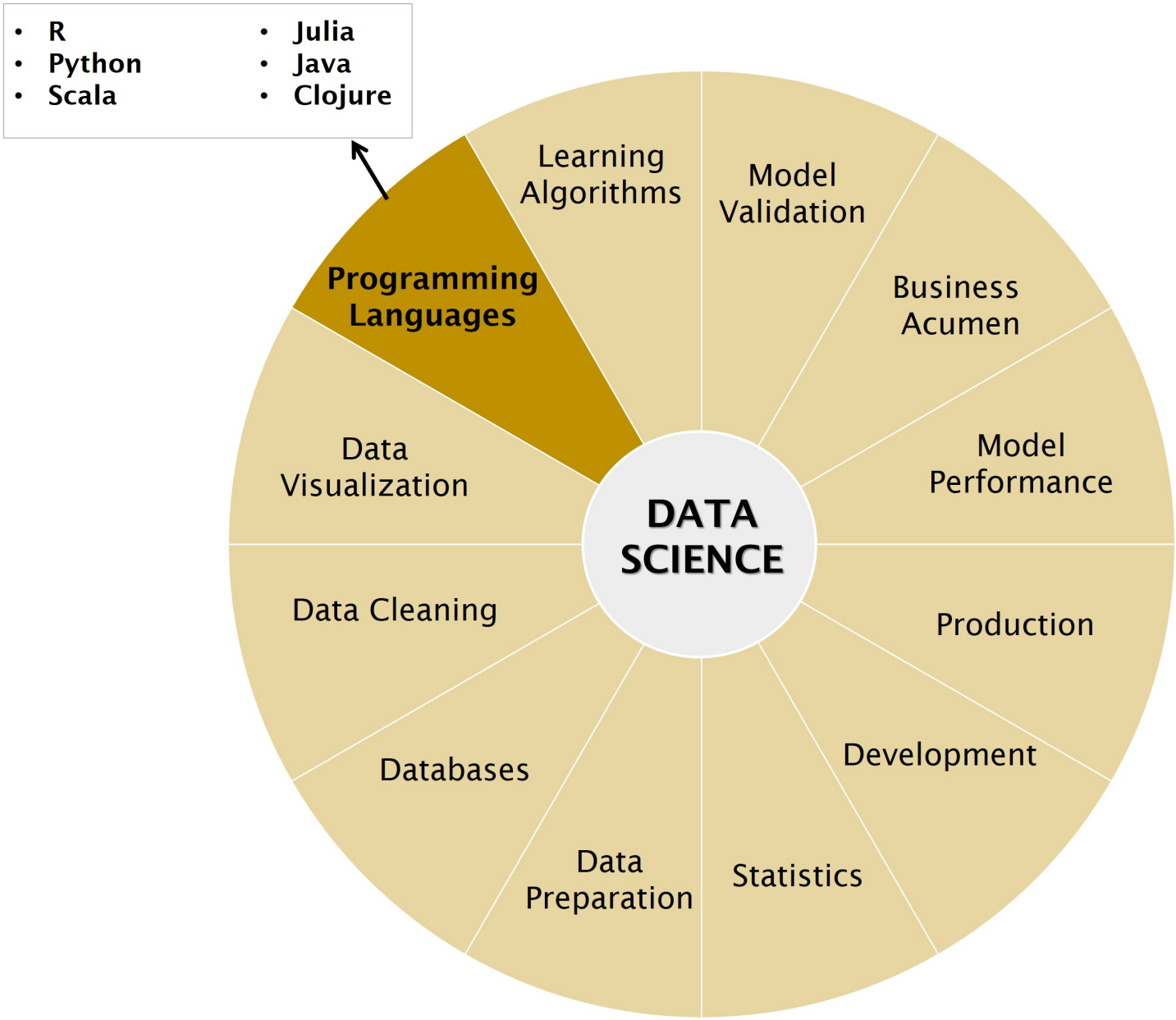
- Diversity on Data Properties.
- Diversity on Representations.
- Diversity on Applications.
- Diversity on Goals / Objectives.
- Diversity on Algorithms.
- Diversity on Theoretical Foundation.
- Diversity on Infrastructures.

Big Data is a Phenomenon


Big Data Definition
Attributive Definition
Comparative Definition
Architectural Definition
5 Vs
Traditional software tools (subjective)
Data Information

- Instance-Based Methods
- Probabilistic Classifiers
- Decision Trees
- Rule Learners
- Regression
- Artificial Neural Networks
- Kernel Methods
- Association
- Rules
- Clustering



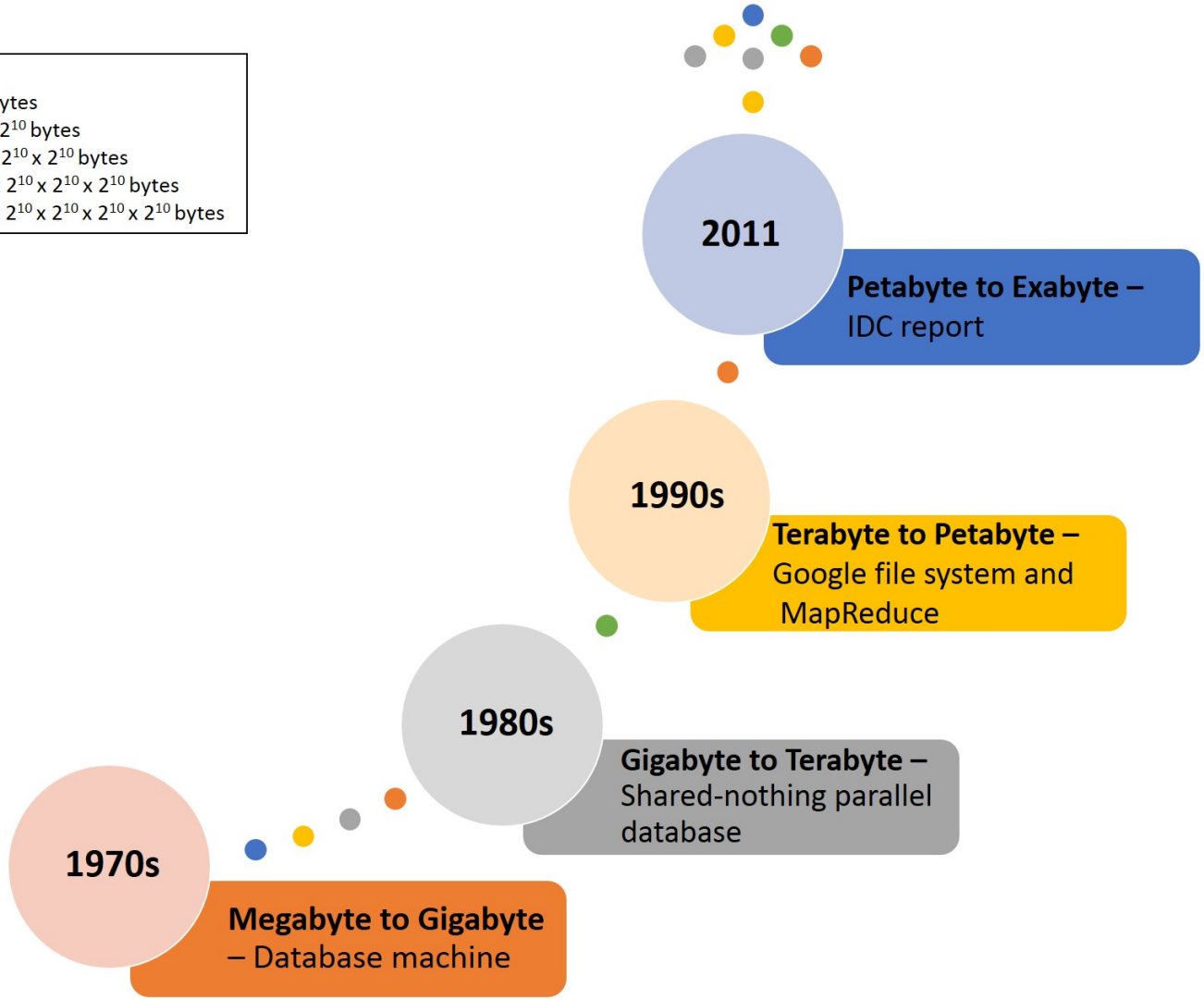

R
Python
Scala
Julia
Java
Clojure
Programming Languages
Learning Algorithms
Model Validation
Business Acumen
Model Performance
Production
Development
Statistics
Data Preparation
Databases
Data Cleaning
Data Visualization
DATA SCIENCE

Learning for Big Data

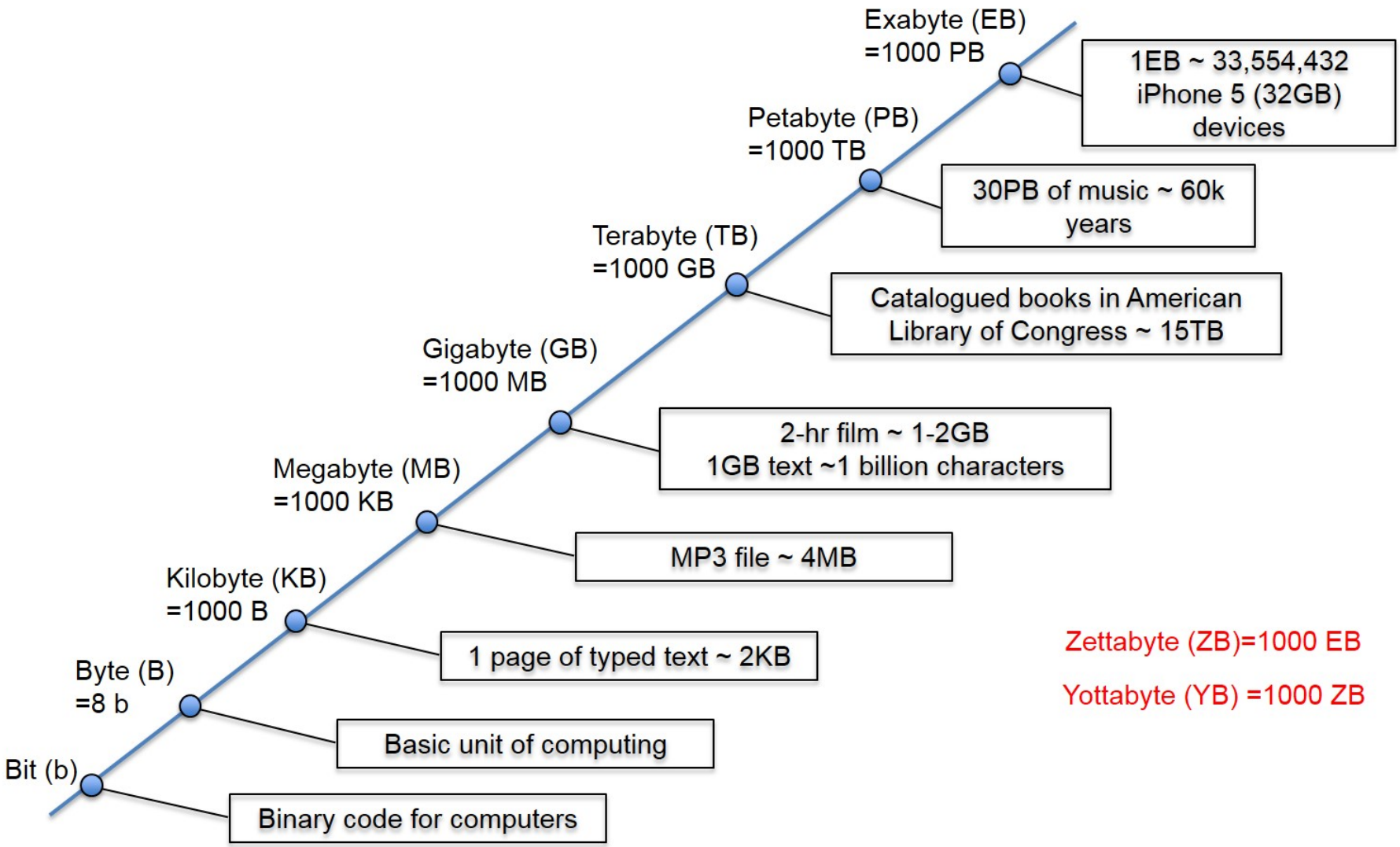
	Traditional Data	**Big Data**
Volume	GB	constantly updated (TB or PB currently)
Generated Rate	per hour, day, ...	more rapid
Structure	structured	semi-structured or un-structured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Access	interactive	batch or near real-time

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



Data Size



Zettabyte (ZB)=1000 EB

Yottabyte (YB) =1000 ZB

Streaming Processing

- Online applications, second, or even milisecond level

Batch Processing

- Data shared and analyzed

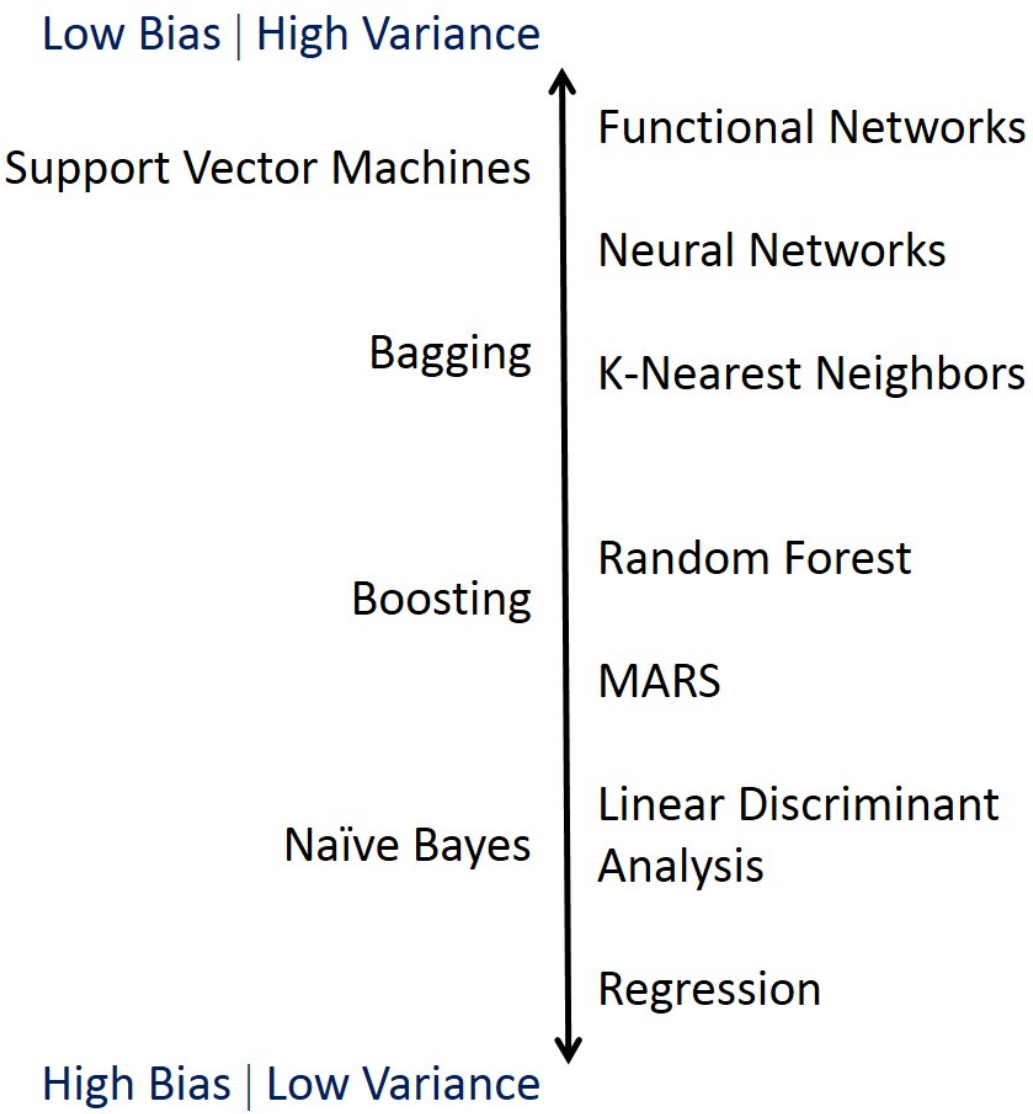
	Streaming Processing	**Batch Processing**
Input	stream of new data or updates	data chunks
Data Size	infinite or unknown in advance	known and finite
Storage	not store or store non-trivial portion in memory	store
Hardware	typical single limited amount of memory	multiple CPUs, memories
Processing	a single or few pass(es) over data	processed in multiple rounds
Time	a few seconds or even miliseconds	much longer
Applications	web mining, sensor networks, monitoring	widely adopted in almost every domain

Taxonomy of Big Data Analytics

Analysis Domains	Sources	Characteristics	Approaches
Structured data analytics	Customer transaction records Scientific experiments data	Structured records Less volume and real time	Data mining Statistical analysis
Text analytics	Logs Email Corporate documents Government Rules and Regulations Text content of webpages Citizen feedback and comments	Unstructured Rich textual Context Semantic Language dependent	Document presentation NLP Information extraction Topic model Summarization Categorization Clustering Question answering Option mining
Web analytics	Various webpages	Integration of text and hyperlink Symbolic Metadata	Web content mining Web structure mining Web usage mining
Multimedia analytics	Corporation produced multimedia User generated multimedia Surveillance Health and patient media	Image, audio, video Massive Redundancy Semantic gap	Summarization Annotation Indexing and retrieval Recommendation Event detection
Mobile analytics	Mobile apps Sensors RFID	Location based Person specific Fragmented information	Monitoring Location based mining

- 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.

Big N

N - Data sets easy to acquire.

- Sampling mechanism.
- Biases in the data.
- To reduce variance in to increase sample size.
- Big N reduce the variance.

- Low bias, high variance model.
- High bias - low variance models.
  - More stable.
  - Lack flexibility to capture many types of non-linear relationships between the predictors and outcomes.
- Less computational expensive.
- Big N can reduce variance and enable highly accurate parameter estimates.
- Does not help reducing bias.

- Low bias / high variance can still be used - if modeler has time to reduce bias.

Low Variance

- Constant function.
- Model independent of training data.
- Model depends on stable measures of data.
  - Mean.
  - Median.

High Variance

- High degree polynomial.

Reduce Variance without Increasing Bias

- Averaging reduces variance:

$$Var(\overline{\mathbf{X}}) = \frac{Var(\mathbf{X})}{N}$$

- Average models to reduce model variance.
- One problem:
  - Only one train set.
  - Where do multiple models come from?

Big P

- Double-edged sword for small - to medium - N data sets.
- Over fitting.

Rail life (MGT) $= f$(Geometry defects)

$max(0, x - t)$ or $max(0, t - x)$

where x represents the independent variables (in this case, the presence of one of more geometry defects) and t represents an extreme point or knot (which is calculated in the analysis itself).

$$MGT = 477.37 + 86.6031(BF_2) + 28.6544(BF_4) - 47.8978(BF_5)$$

$BF_2 = max(0, 6 - WARP)$
$WARP =$ number of WARP defects

$BF_4 = max(0, 18 - RAIL_CANT)$
$RAIL_CANT =$number of rail cant defects

$BF_5 = max(0, ALIGNMENT - 0)$
$ALIGNMENT =$ number of alignment defects

$Max\ MGT(no\ defects) =$ 1513 MGT

MARS Analysis in Tangent Track

$$CUM.MGT = 355.9443 + 69.37212 * pmax(0, 5 - CROSSLEVEL) - 156.7451 * pmax(0, GAGE - 1) + 213.6788 * pmax(0, 2 - RAIL.CANT) + 188.8701 * pmax(0, 2 - WARP.62)$$

Key variables:

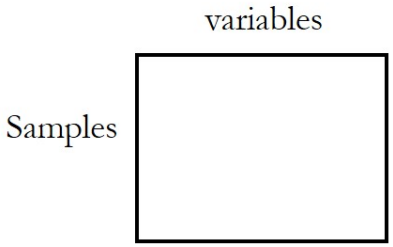
- Warp 62
- Rail Cant
- Gage
- Cross-Level

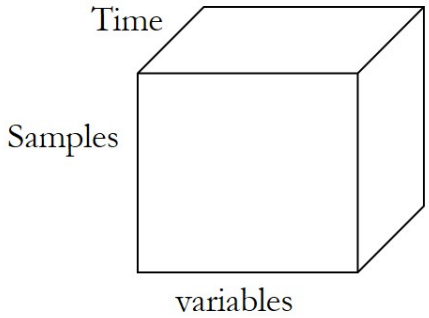
MARS Analysis in Curve Track

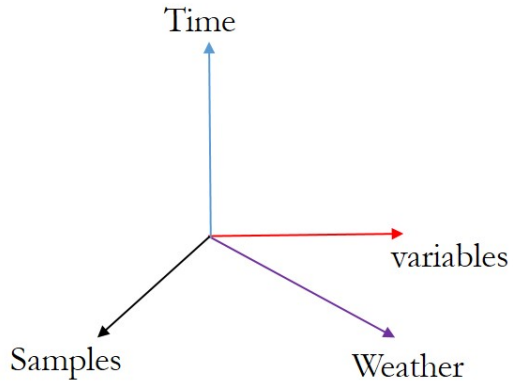
$$CUM.MGT = 258.4052 + 172.4302 * pmax(0, 2 - ALIGNMENT) + 115.7155 * pmax(0, 2 - CROSSLEVEL) + 66.55255 * pmax(0, 5 - RAIL.CANT) + 94.51797 * pmax(0, 4 - WARP.31)$$

What are Tensors?

Generalizations of scalars, vectors and matrices.





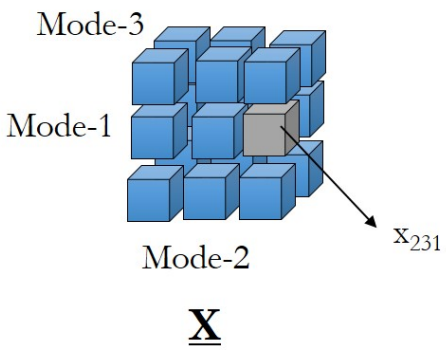


Multiway Data Analysis

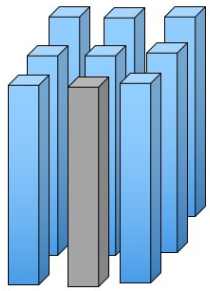
- Multiway data is everywhere.
- Capture variation in dataset while preserving multidimensional nature of data.
- Reveal hidden patterns in data.

Multiway Data Analysis

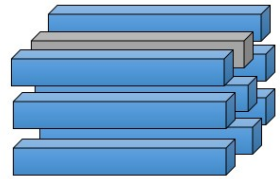
Typical 3D dataset



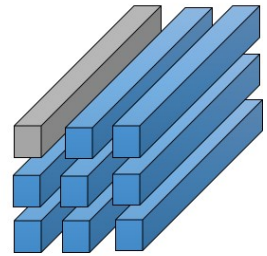
Subarrays



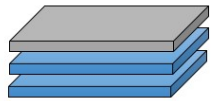
Mode-1 Fiber
$x_{:21}$



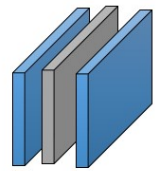
Mode-2 Fiber
$x_{1:2}$



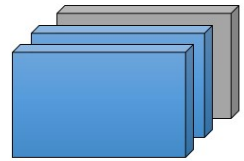
Mode-3 Fiber
$X_{11:}$



Horizontal slice
$X_{1::}$



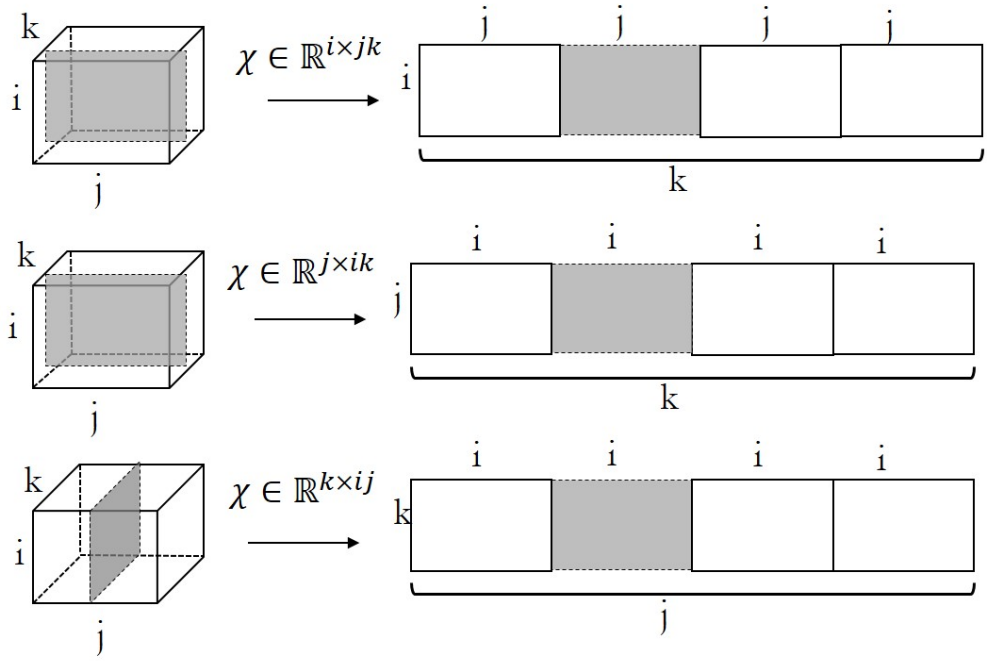
Lateral slice
$X_{:2:}$



Frontal slice
$X_{::3}$

4

Matricization (Unfolding)



5

Summation: $\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{C}}$, where $\boldsymbol{\mathcal{C}}_{ijk} = a_{ijk} + b_{ijk}$

Outer Product: $\boldsymbol{\mathcal{X}} = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c}$ s.t. $x_{ijk} = a_i b_j c_k$

Khatri-Rao Product: $\boldsymbol{P}^{I \times J} \odot \boldsymbol{Q}^{K \times J} = \boldsymbol{R}^{IK \times J}$

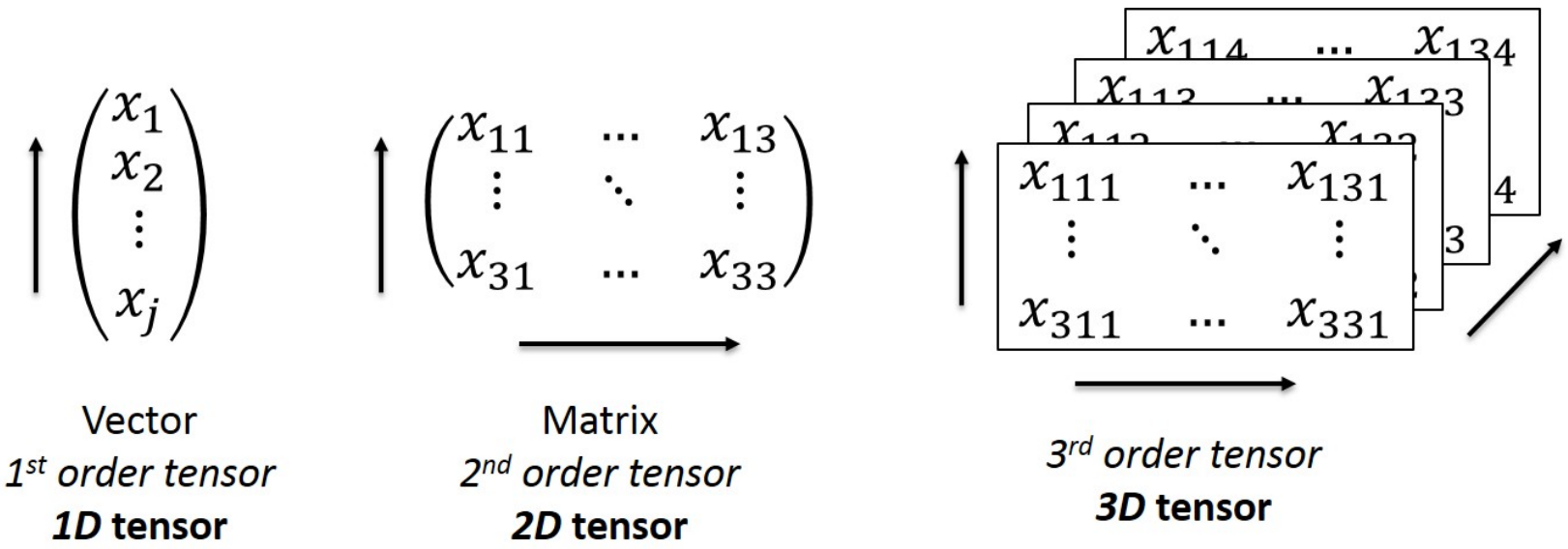
Kronecker Product: $\boldsymbol{P}^{I \times J} \otimes \boldsymbol{Q}^{K \times L} = \boldsymbol{R}^{IK \times JL}$

Hadamard Product: $\boldsymbol{P}^{I \times J} \circledast \boldsymbol{Q}^{I \times J} = \boldsymbol{R}^{I \times J}$

Conventional N-way tensor notations

Number of ways	Name	Notation
Zero	Scalar	x
One	Vector	$\boldsymbol{x} \in \mathbb{R}^{I}$
Two	Matrix	$\boldsymbol{X} \in \mathbb{R}^{I \times J}$
Three	3-way tensor	$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$
N-way	N-way tensor	$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$

Illustration of one-way, two-way and three-way tensors

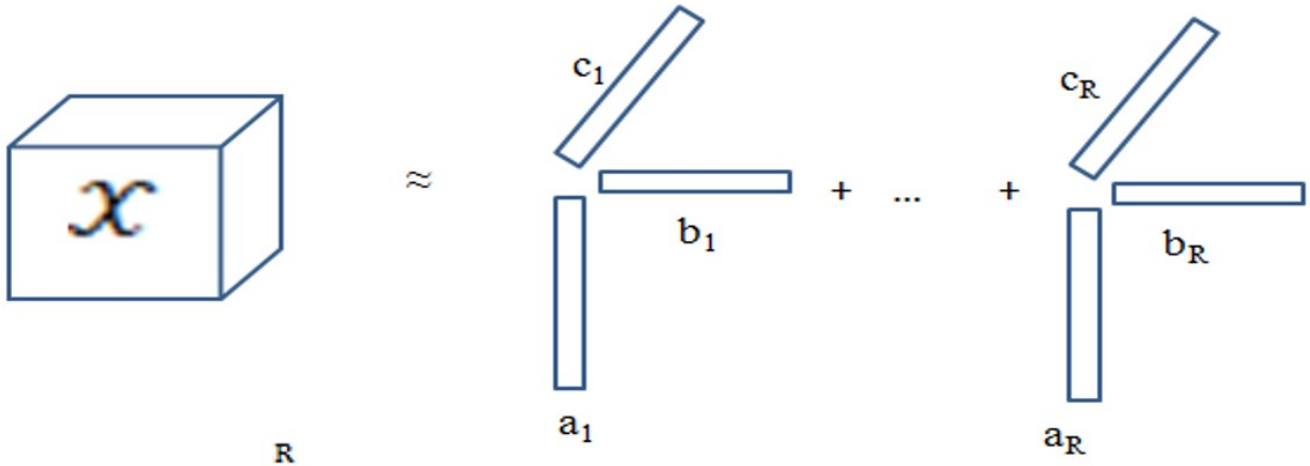


Tensor Decomposition

- Reducing data to lower-order forms for analysis (classification, prediction, clustering).
- 2 main approaches.
  - Canonical decomposition/Parallel Factors(CANDECOMP/PARAFAC; CP).
  - Tucker decomposition

The CP Decomposition

- Canonical Decomposition/ Parallel Factors



$$\mathcal{X} = \sum_{r=1}^{R} a_r \circ b_r \circ c_r$$

a_i, b_i and c_i - factor loadings

R- number of components

Rank of a tensor=min (R) required to approximate $\underline{\mathbf{X}}$

Choosing the appropriate CP model

- Alternating Least Squares Approach (ALS)

$$min_{\widehat{x}} \|x - \widehat{x}\|$$

  where $\widehat{x}$ is the PARAFAC model.
- % Variance Explained.
- Core Consistency Diagnostic (CORCONDIAG).
  - High CORCONDIAG $\longrightarrow$ stable model.
  - Low CORCONDIAG $\longrightarrow$ invalid and/or problematic model.

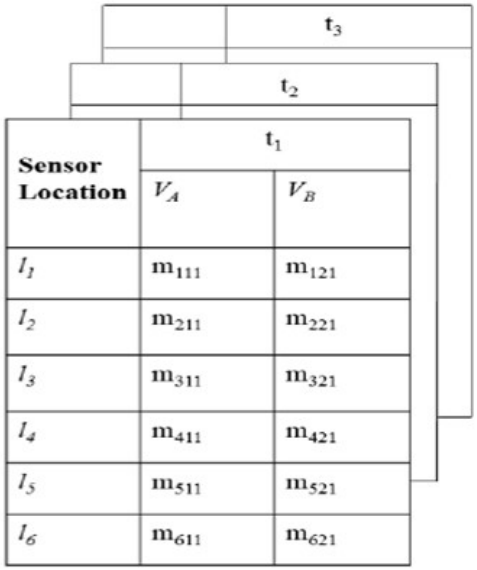
- Kartunnen (2012) noted that a common quantification measure of track geometry quality is the standard deviation of irregularities over a specified length.
- Standard deviation $= \sqrt{\frac{1}{n-1}\sum\left(\delta_i - \overline{\delta}\right)^2}$
- Where δ_i - measured irregularity, $\overline{\delta}$ - mean value, n - number of measurements.

- The standard deviations or, in some cases, average deterioration of analytical segments are represented as a deterioration matrix.

$$[D_i]_{kXm} = \begin{bmatrix} D_{i11} & D_{i12} & \cdots & D_{i1m} \\ D_{i21} & D_{i22} & \cdots & D_{i2m} \\ \vdots & \vdots & \vdots & \vdots \\ D_{ik1} & D_{ik2} & \cdots & D_{ikm} \end{bmatrix}$$

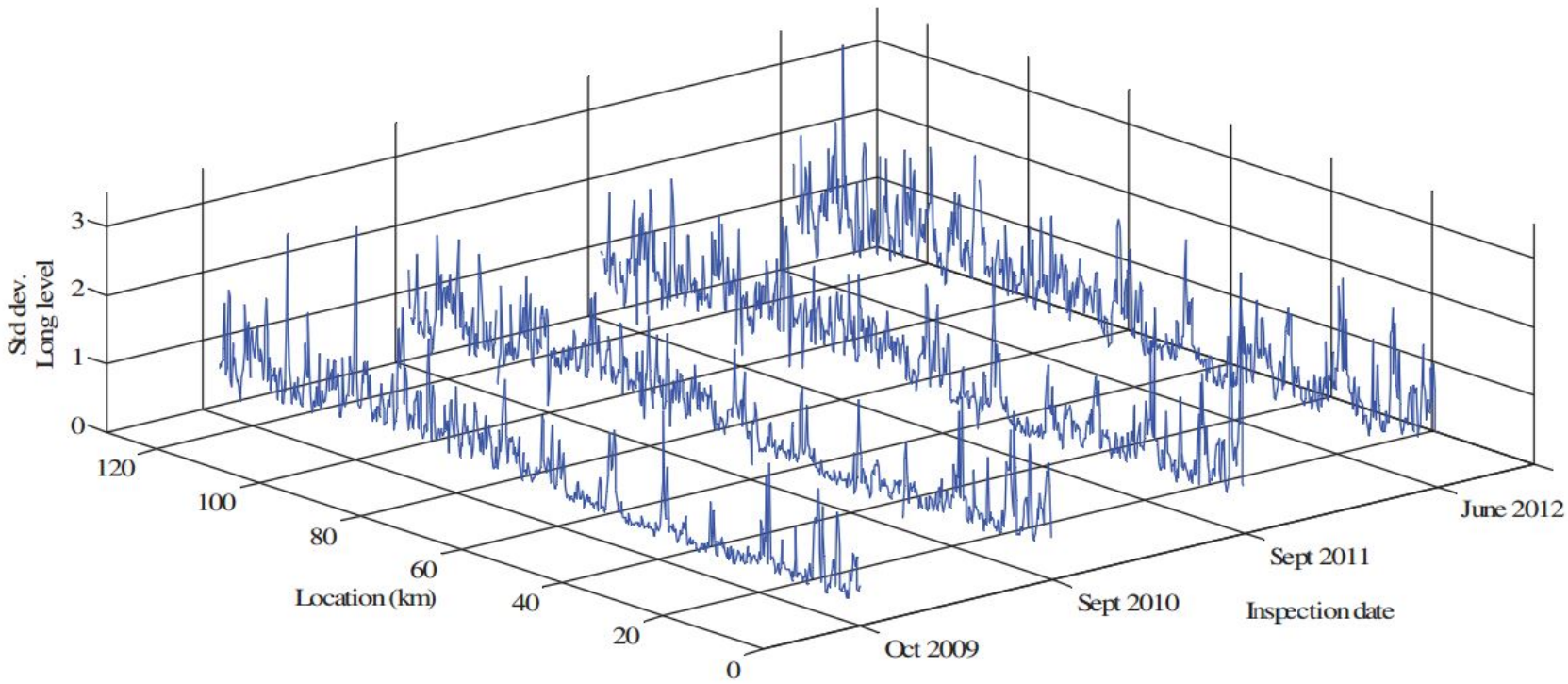
Sensor Location	t_1	
	V_A	V_B
l_1	m_{11}	m_{12}
l_2	m_{21}	m_{22}
l_3	m_{31}	m_{32}
l_4	m_{41}	m_{42}
l_5	m_{51}	m_{52}
l_6	m_{61}	m_{62}

(i)



(ii)

Standard Deviation of Longitudinal Levels



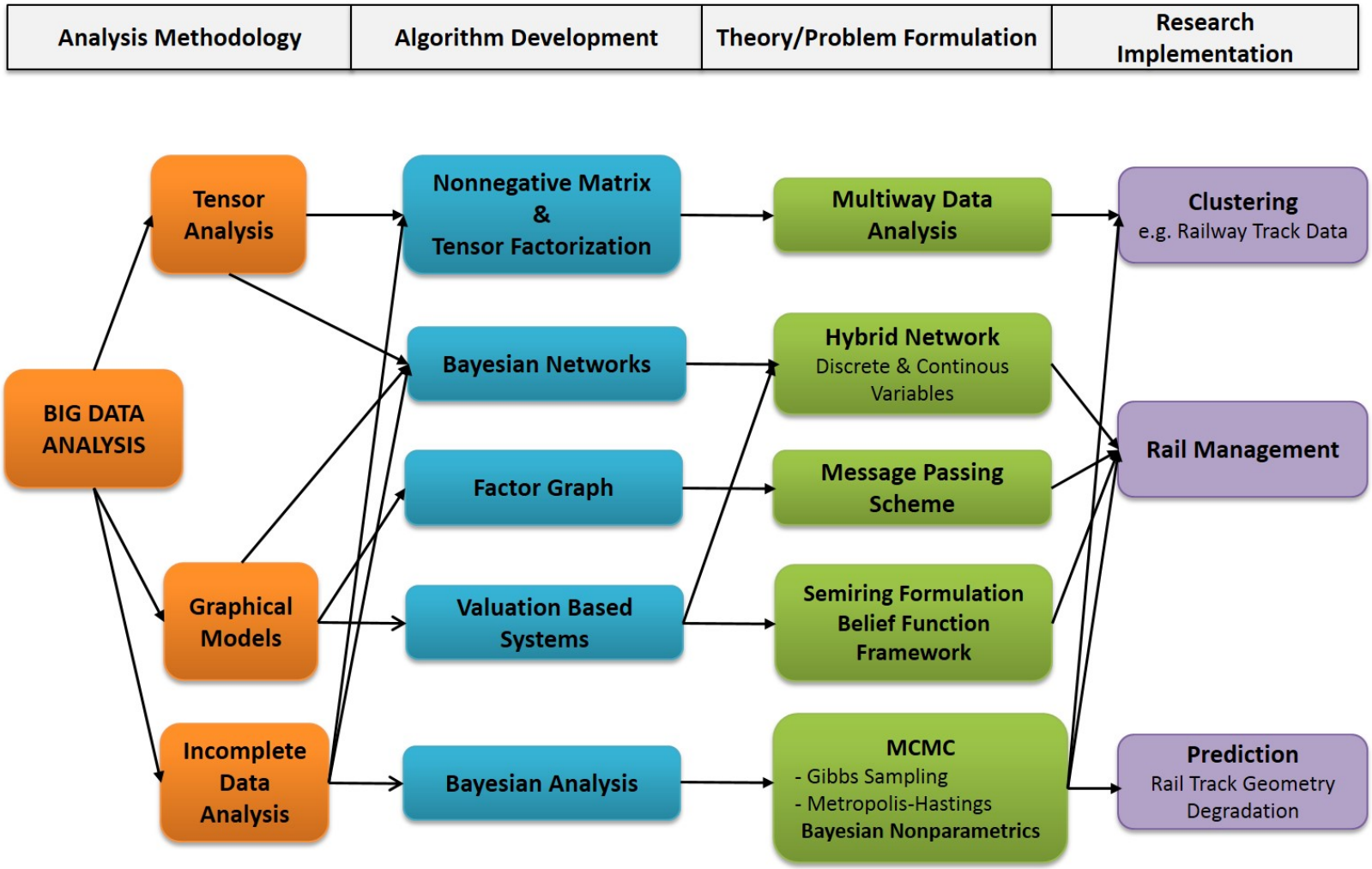
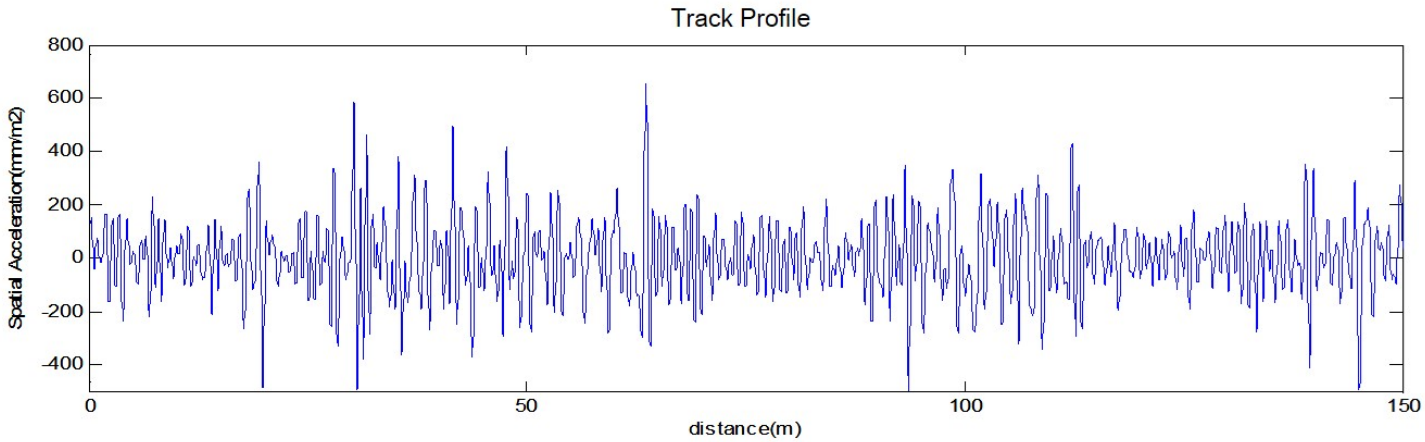
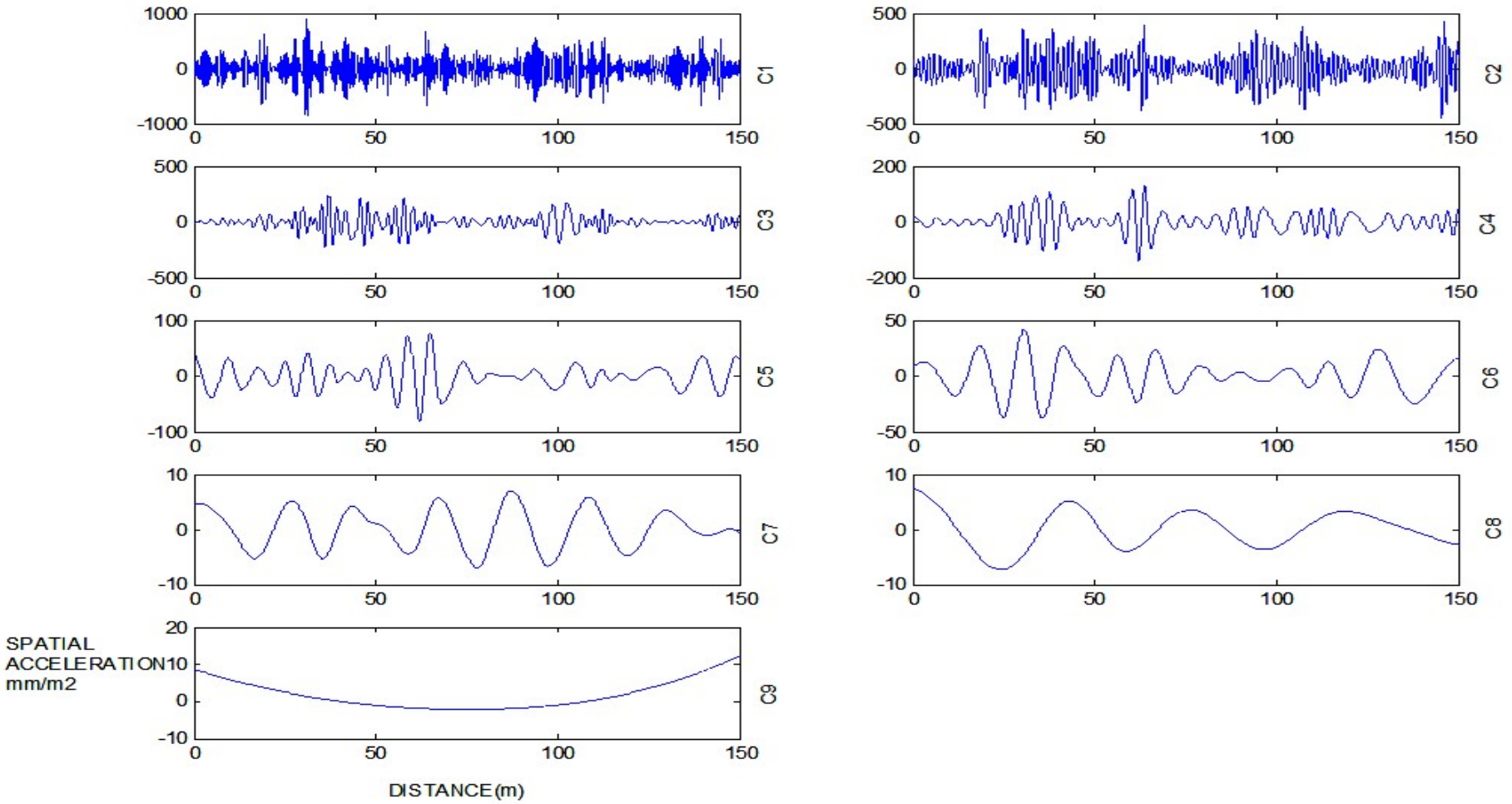


Figure 1: Example of Railway Big Data (Track Data)

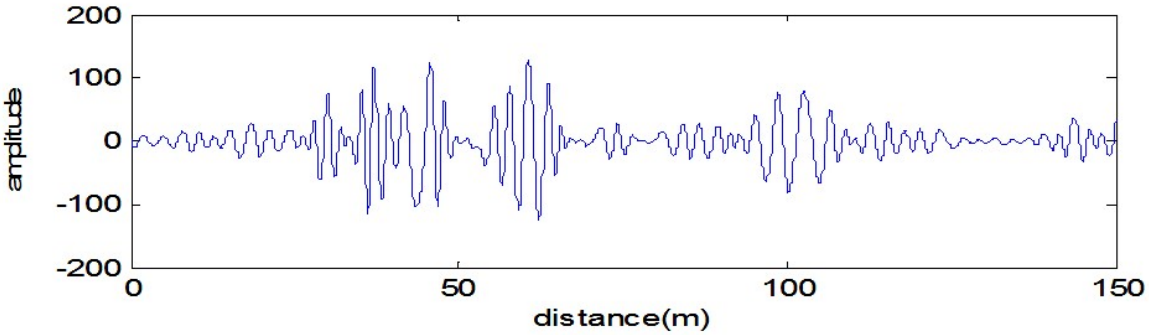
- Profile is differentiated. We desire the:
  - Deterioration trend.
  - Localized profile information.



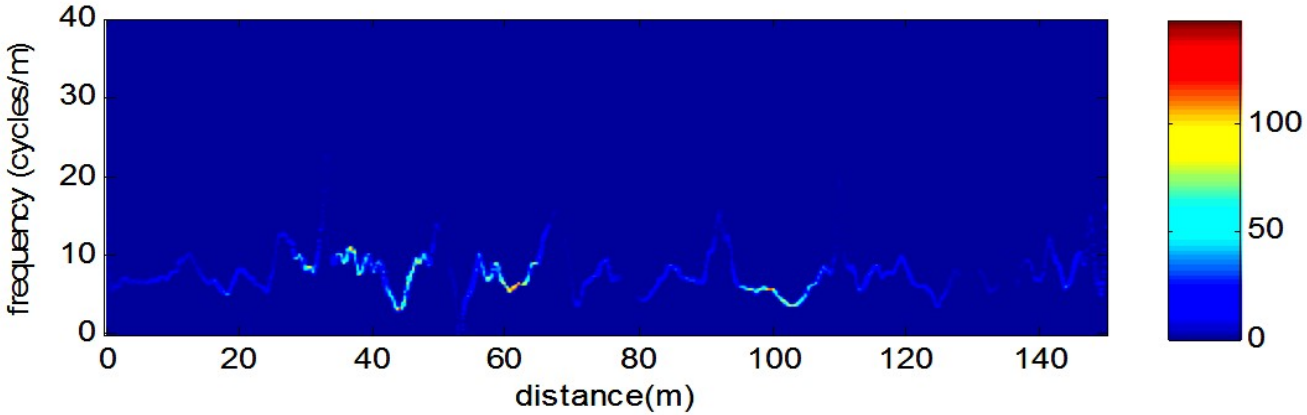
Decomposed IMFs



Energy-distance-frequency Information

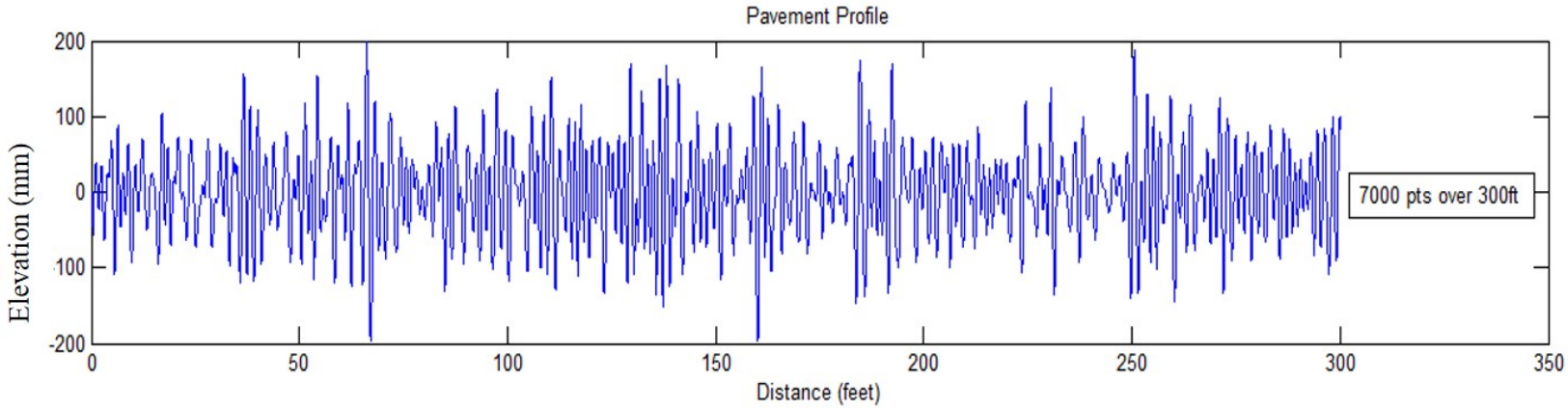


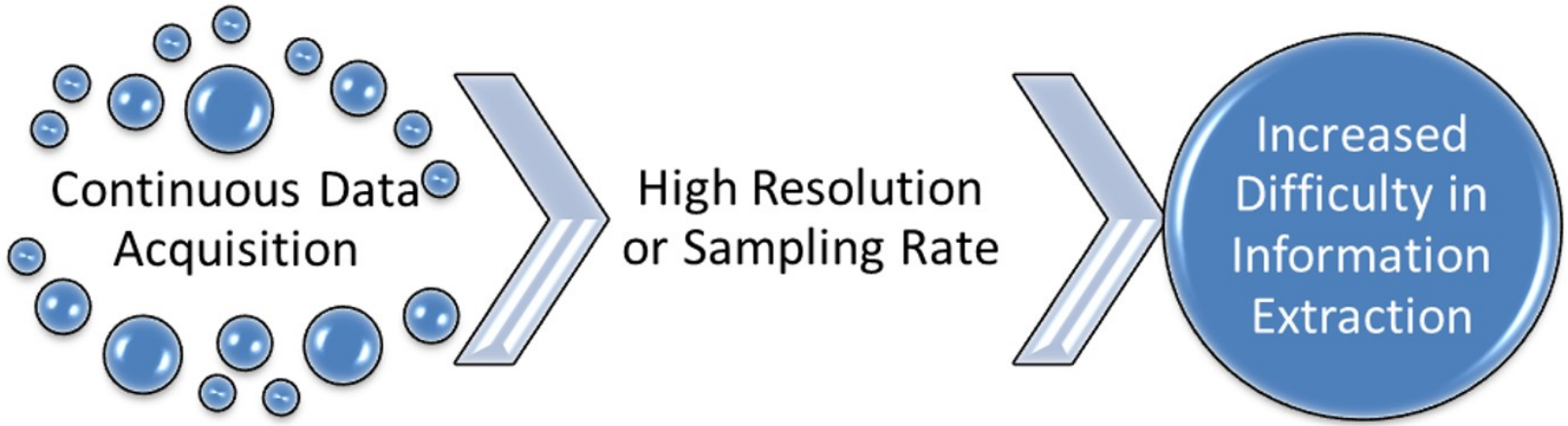
Critical Locations



Hilbert Spectrum of IMFs

Motivation


Pavement Profile
Elevation (mm)
200
100
0
-100
-200
0
50
100
150
200
250
300
350
Distance (feet)
7000 pts over 300ft


Continuous Data Acquisition
High Resolution or Sampling Rate
Increased Difficulty in Information Extraction

- Input is presented as a sequence of items.
- Can be examined in only a few passes.
- Limited memory.
- Data can arrive at enormous rates.

Same as Online Algorithms

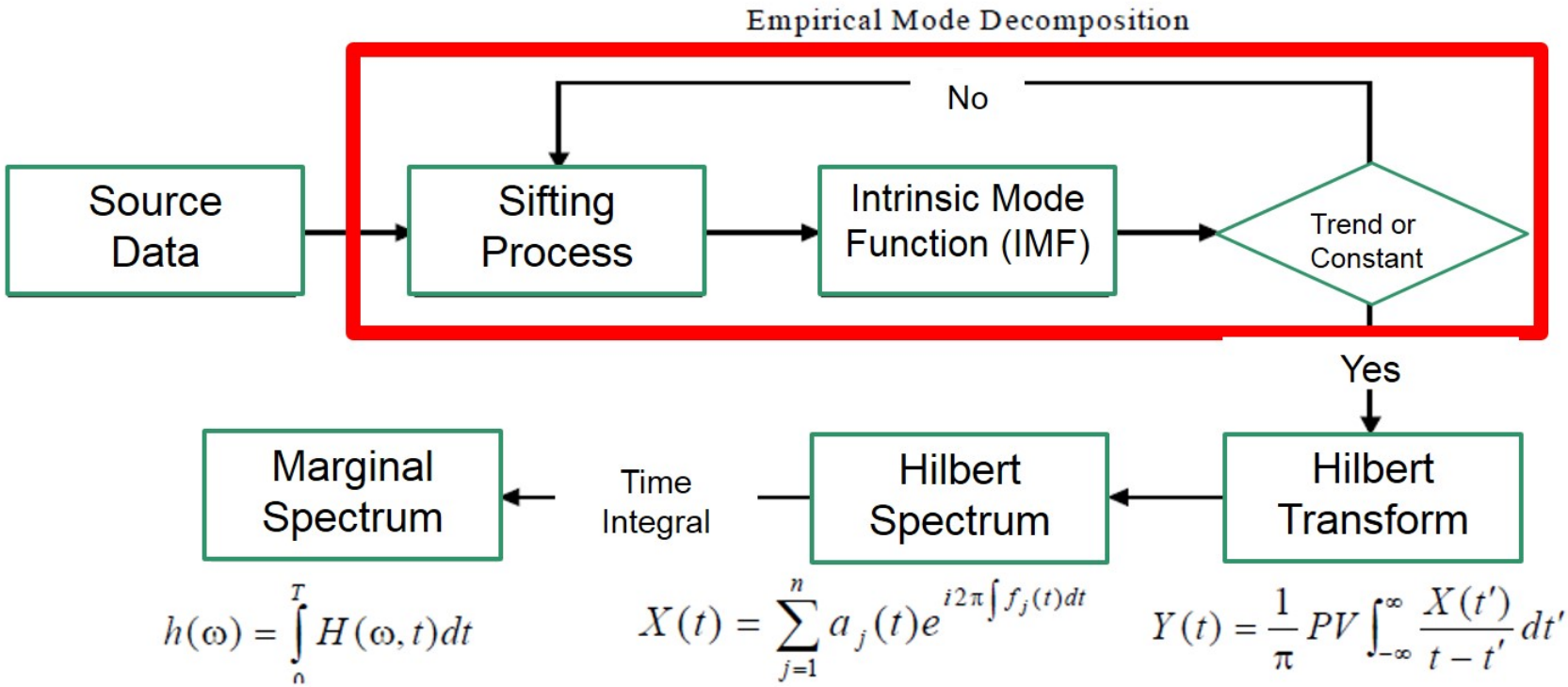
Similarities

- Decisions to be made before all data are available.
- Limited memory.

Differences

- Streaming algorithms can defer action until a group.
- Outline algorithms take action as soon as each point arrives.

Flow Chart

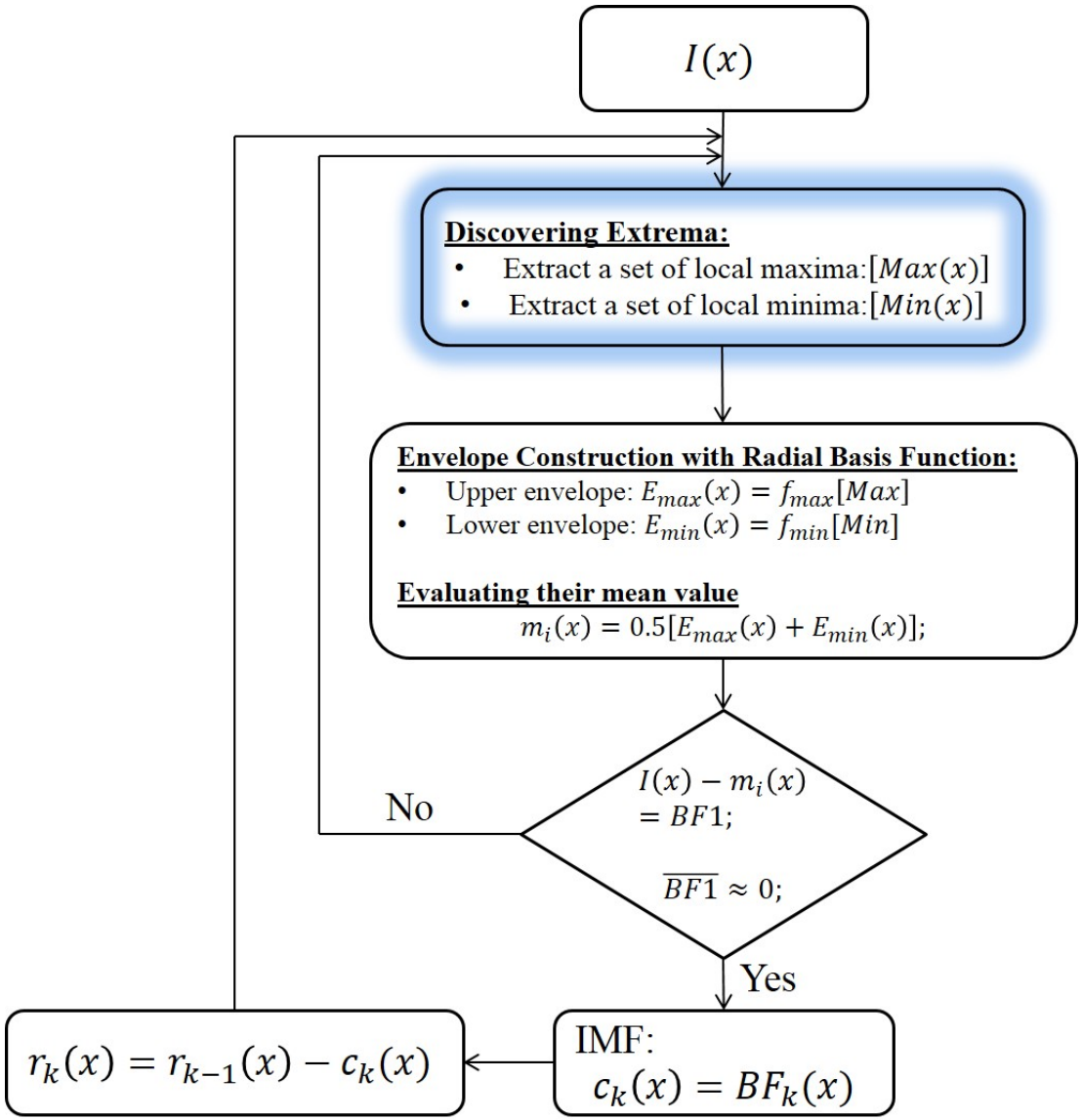

Empirical Mode Decomposition
No
Source Data
Sifting Process
Intrinsic Mode Function (IMF)
Trend or Constant
Yes
Marginal Spectrum
Time Integral
Hilbert Spectrum
Hilbert Transform
$h(\omega) = \int_0^T H(\omega, t)dt$
$X(t) = \sum_{j=1}^{n} a_j(t) e^{i2\pi \int f_j(t)dt}$
$Y(t) = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{X(t')}{t - t'} dt'$

Empirical Mode Decompositon (EMD)

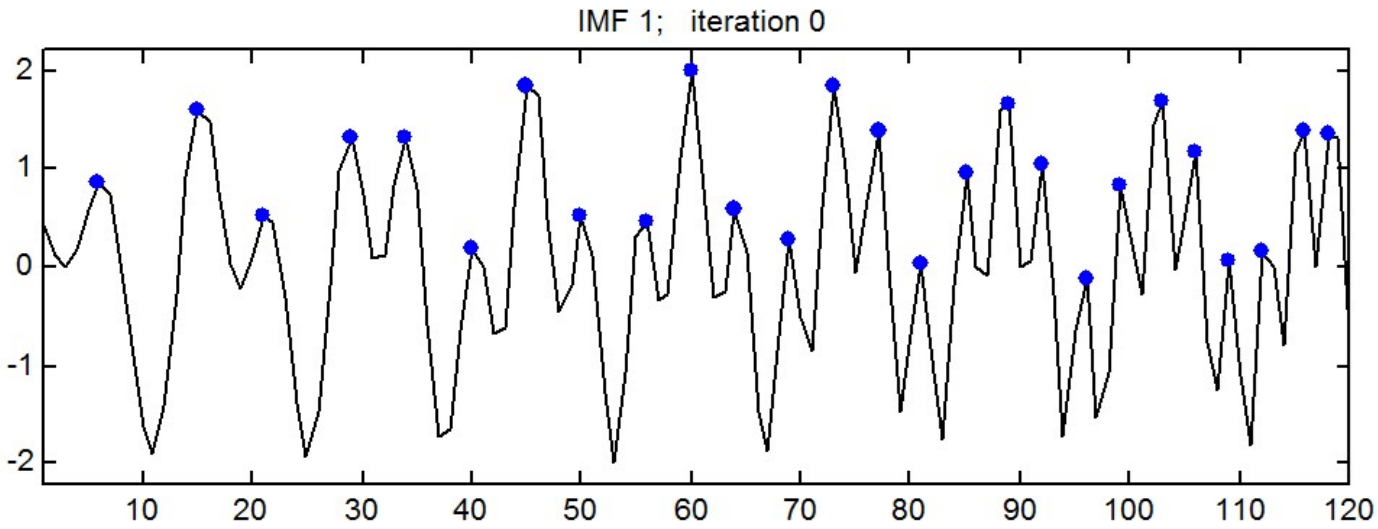
	IMF1	IMF2	IMF3	...	IMFn
	$X(t)$	$X(t) - c_1 = r_1$	$r_1 - c_2 = r_2$	$\cdots$	$r_{n-2} - c_{n-1} = r_{n-1}$
0	$X(t) - m_1 = h_1$	$r_1 - m_2 = h_2$	$r_2 - m_3 = h_3$	$\cdots$	$r_{n-1} - m_n = h_n$
1	$h_1 - m_{11} = h_{11}$	$h_2 - m_{21} = h_{21}$	$h_3 - m_{31} = h_{31}$	$\cdots$	$h_n - m_{n1} = h_{n1}$
2	$h_{11} - m_{12} = h_{12}$	$h_{21} - m_{22} = h_{22}$	$h_{31} - m_{32} = h_{32}$	$\cdots$	$h_{n1} - m_{n2} = h_{n2}$
3	$h_{12} - m_{13} = h_{13}$	$h_{22} - m_{23} = h_{23}$	$h_{32} - m_{33} = h_{33}$	$\cdots$	$h_{n2} - m_{n3} = h_{n3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$h_{1(k-1)} - m_{1k} = h_{1k}$	$h_{2(k-1)} - m_{2k} = h_{2k}$	$h_{3(k-1)} - m_{3k} = h_{3k}$	$\cdots$	$h_{n(k-1)} - m_{nk} = h_{nk}$
IMF	$h_{1k} = c_1$	$h_{2k} = c_2$	$h_{3k} = c_3$	$\cdots$	$h_{nk} = c_n$

Sifting Process

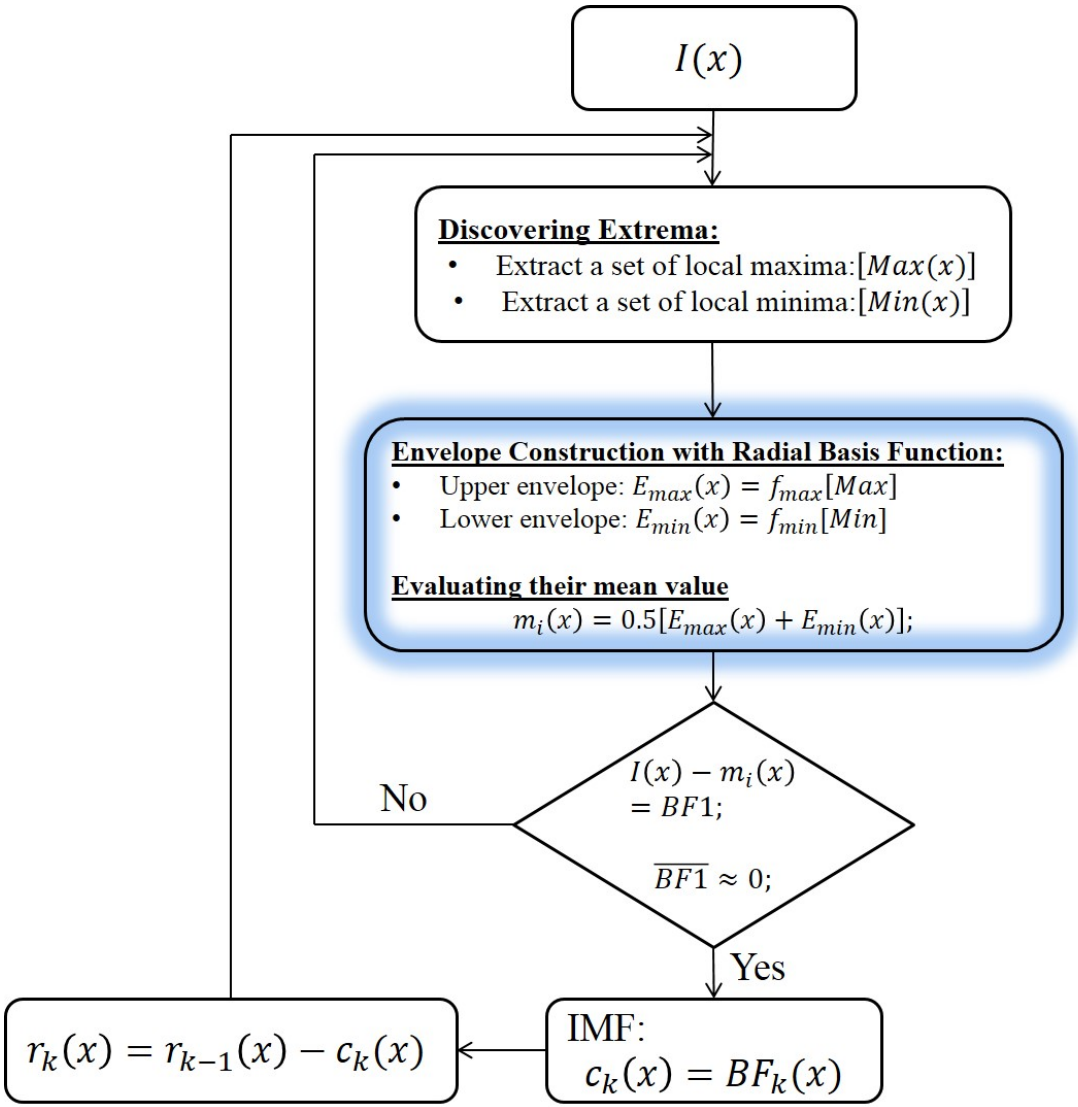
Extraction Extrema



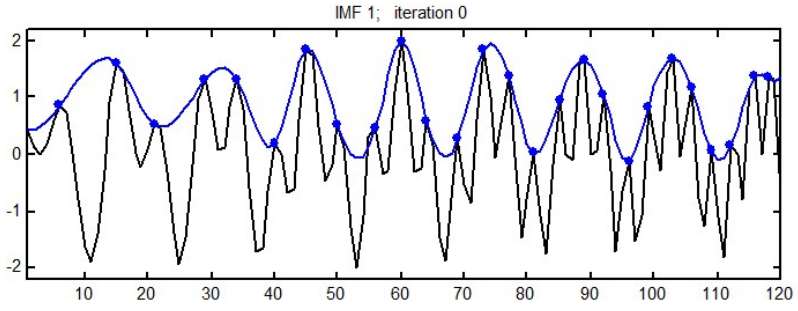
maxima: $x_{i-1} < x_i > x_{i+1}$

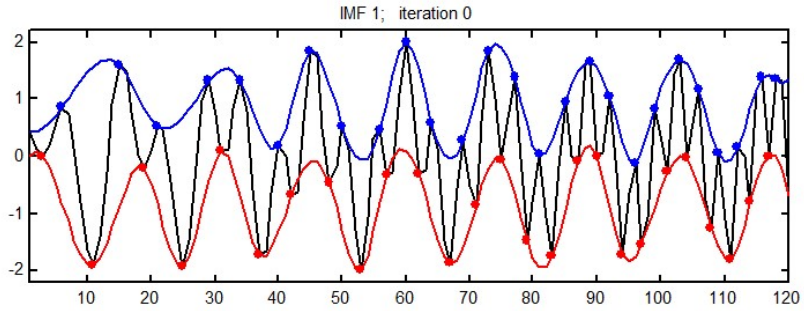


Extraction of Extrema

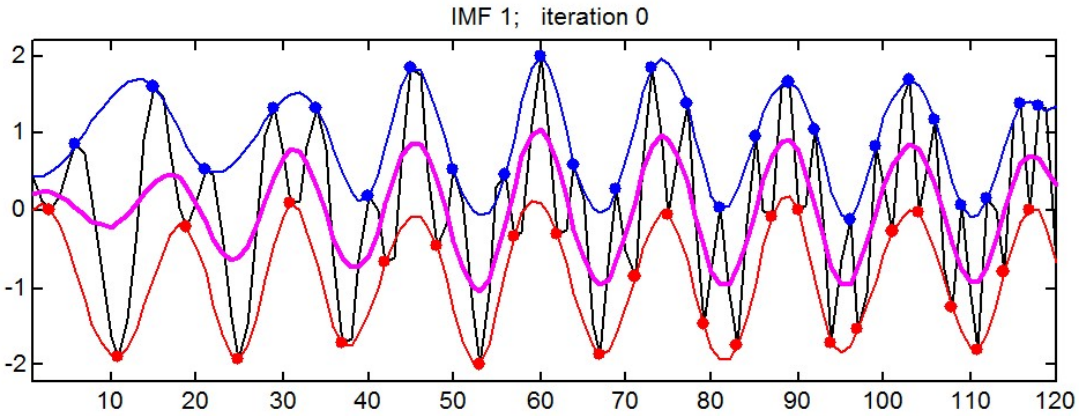


Interpolate to find Local Mean

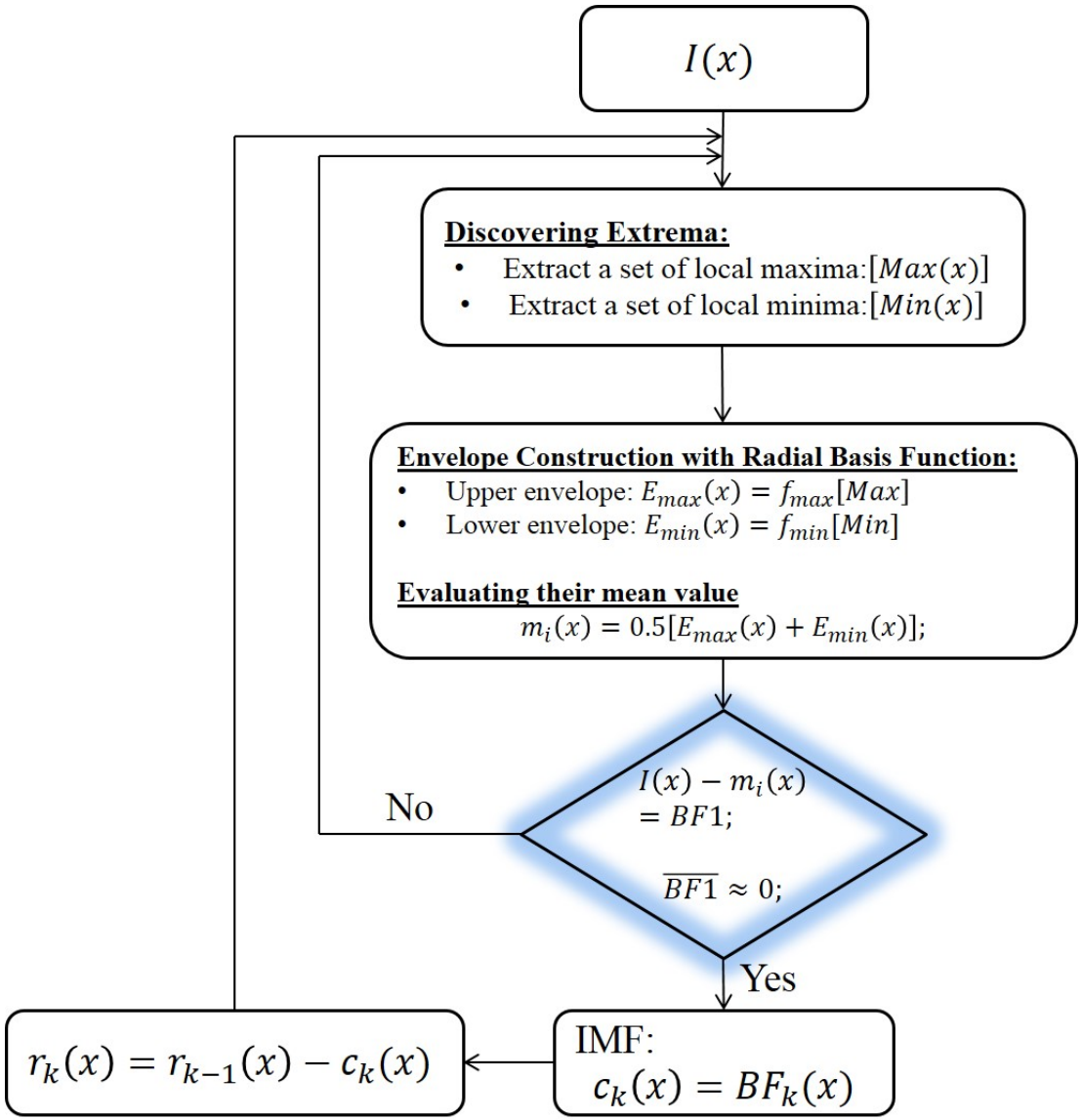




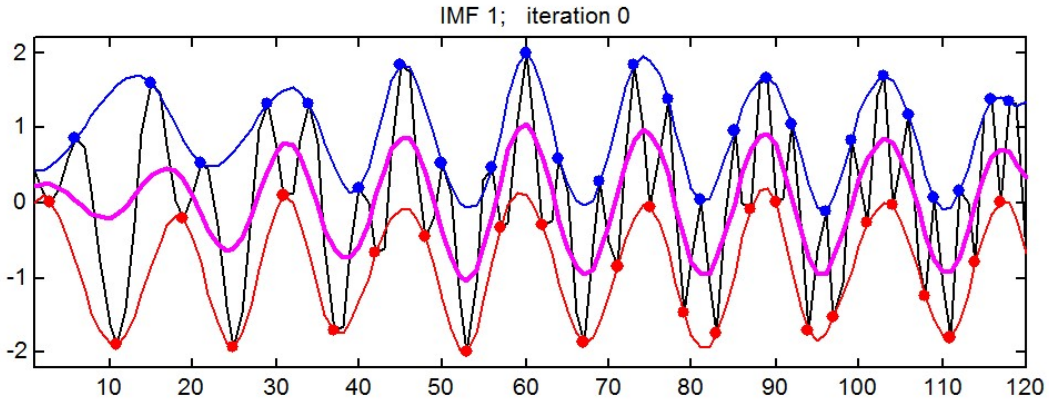
$$min: x_{i-1} > x_i < x_{i+1}$$

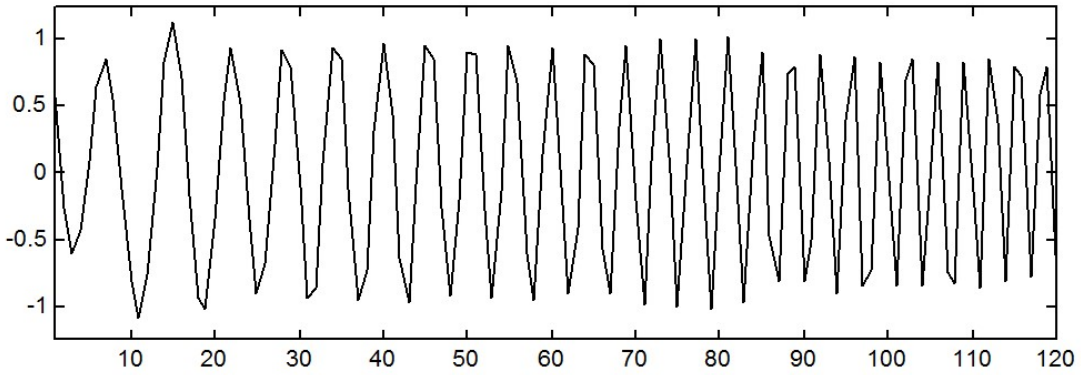


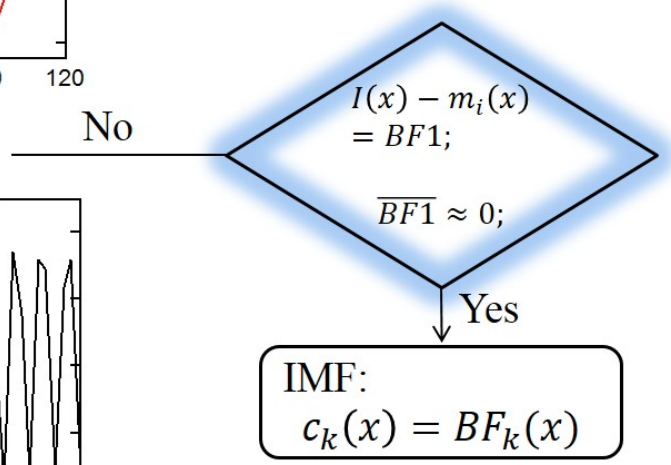
Interpolate to find Local Mean



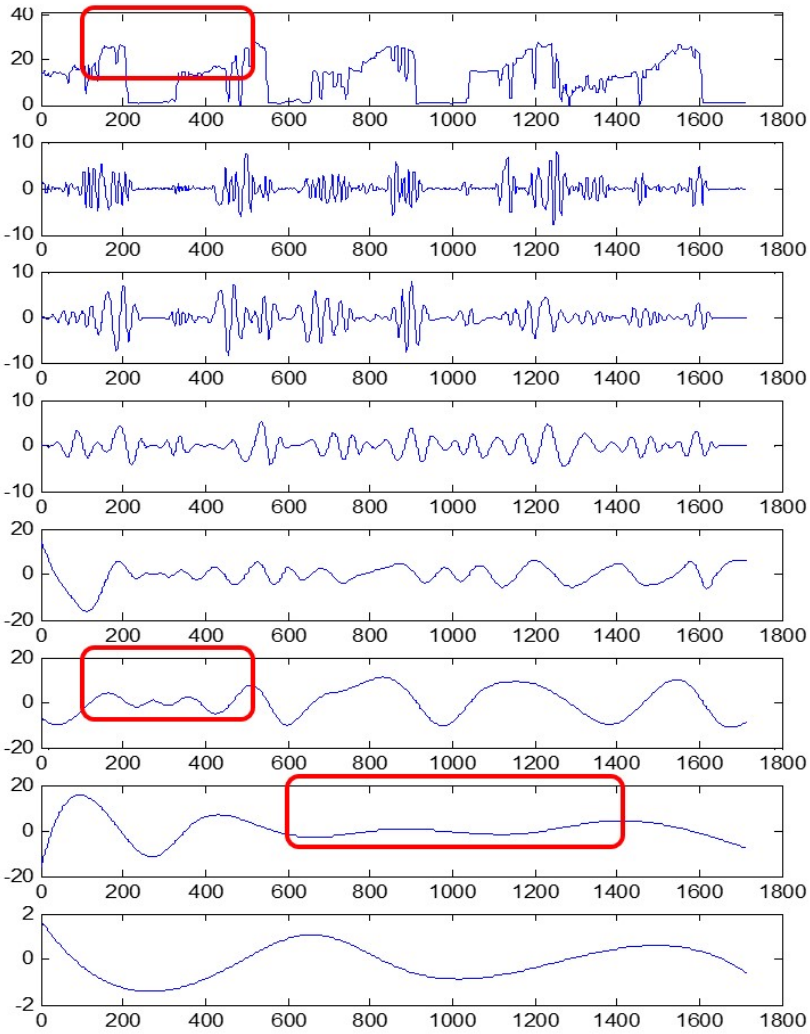
Remove Local Mean







Signal Reconstruction

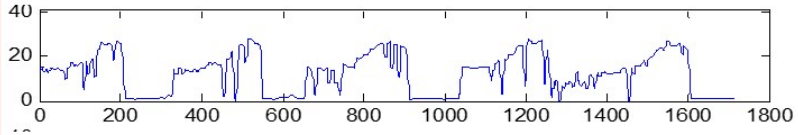


- IMFs should be physically meaningful
- Mode-Mixing and frequency leakage

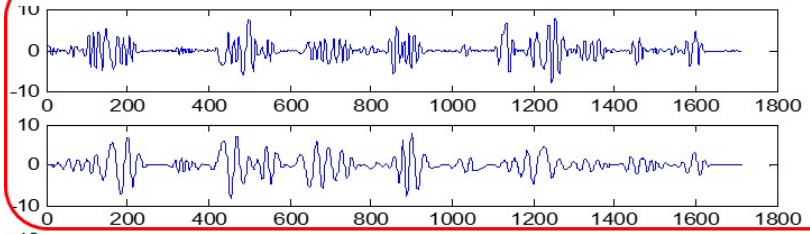
Goal

- Reconstruct an informative and physically meaningful signal

Modeling Intrinsic Mode Function



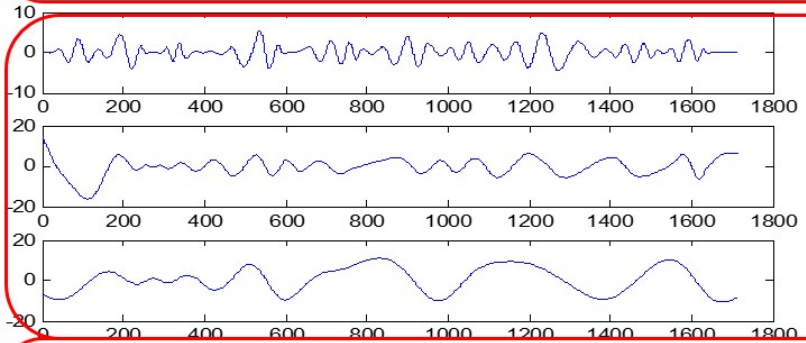
$$x_{IMFx,t} = 0.9341x_{IMFx,t-1} - 0.0656x_{IMFx,t-2} + \varepsilon_t$$



GARCH Model (Non-linear)

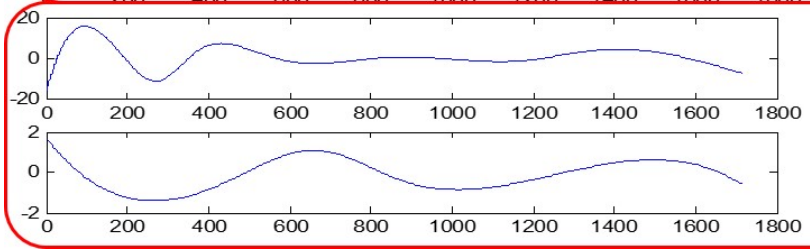
$$X_t = C + \varepsilon_t ; \varepsilon_t = z_t \sigma_t$$

$$\sigma^2_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon^2_{t-i} + \sum_{j=1}^{q} \beta_j \sigma^2_{t-j}$$



ARIMA, AR and MA models

$$x_{IMFx,t} = 0.0321x_{IMFx,t-1} - 0.1456x_{IMFx,t-2} + \varepsilon_t$$

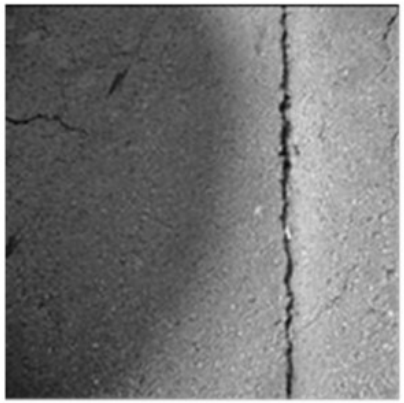


AR and MA models

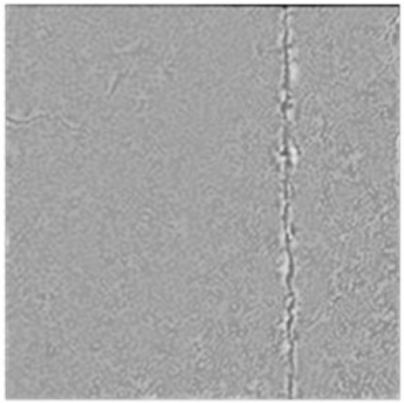
$$x_{IMFx,t} = 0.1321x_{IMFx,t-1} - 0.0456x_{IMFx,t-2} + \varepsilon_t$$

Algorithms: BEMD

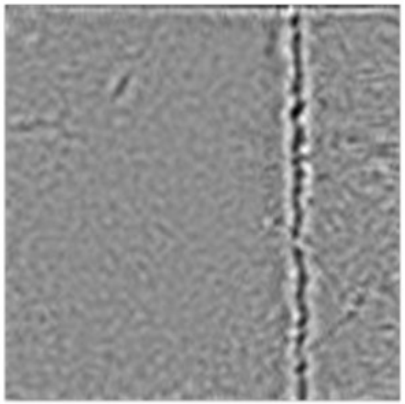
Original Image



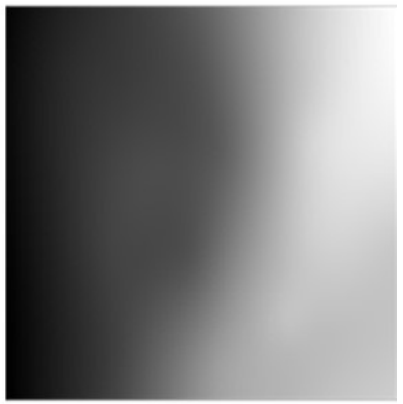
Mode 1



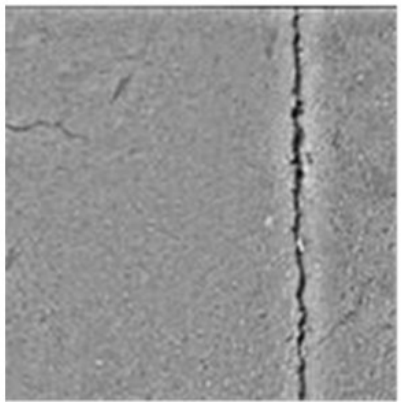
Mode 2



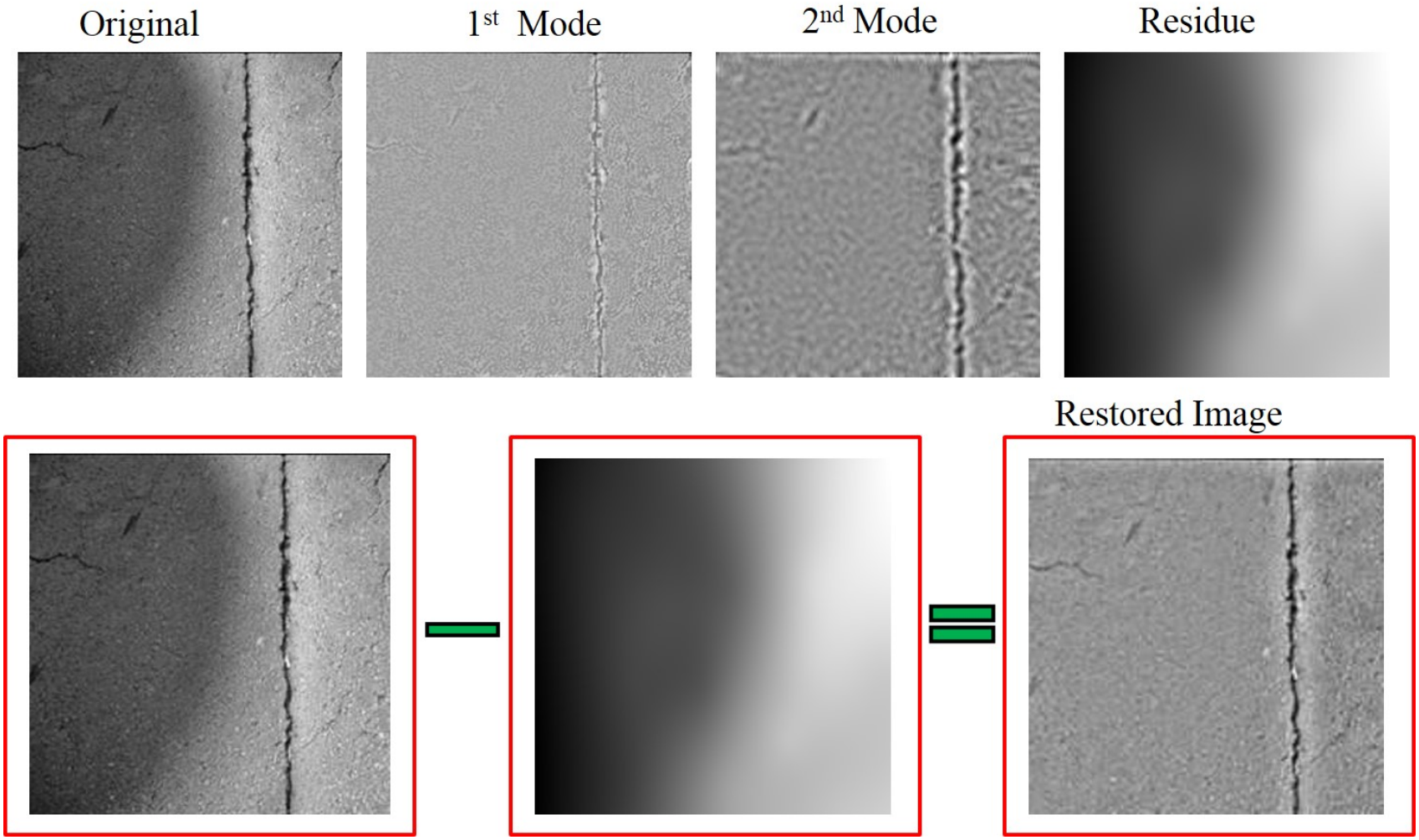
Trend



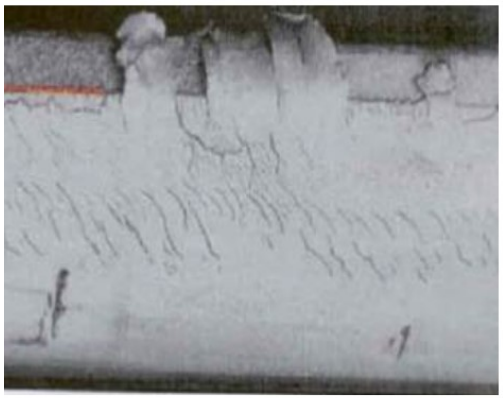
Restored Image



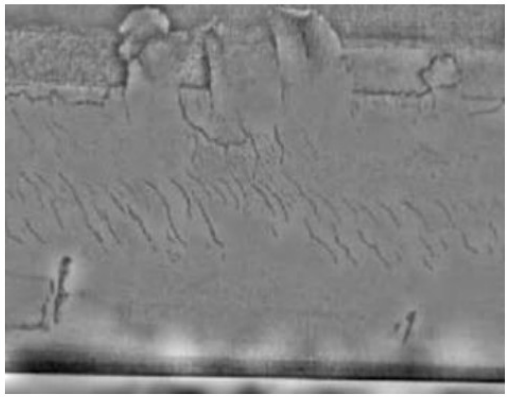
2D-Multiscale Analysis



Original

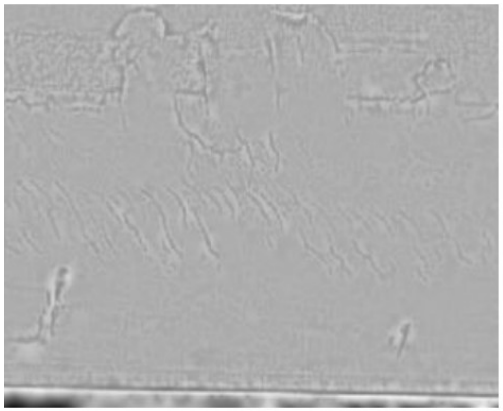


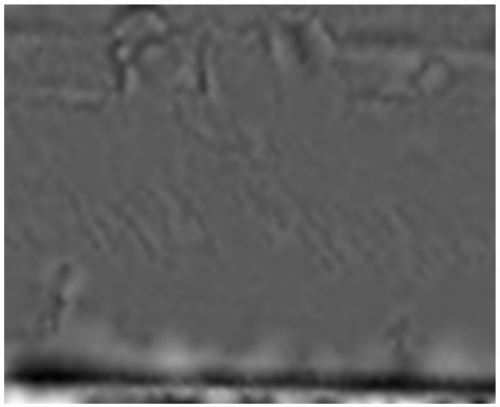
Reconstruction



Edge Detection

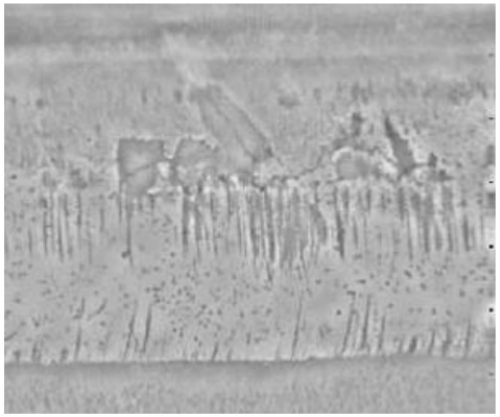
Decomposition Results





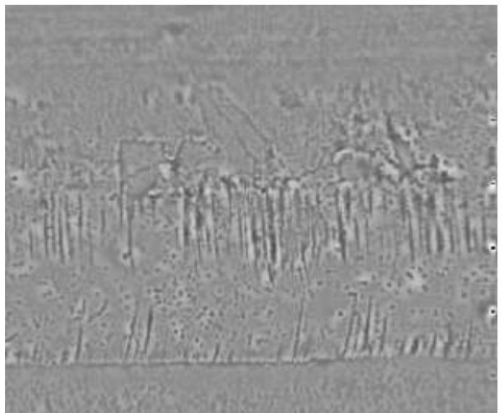
Original

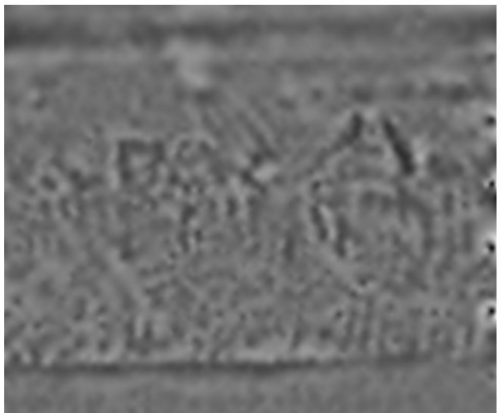
Reconstruction



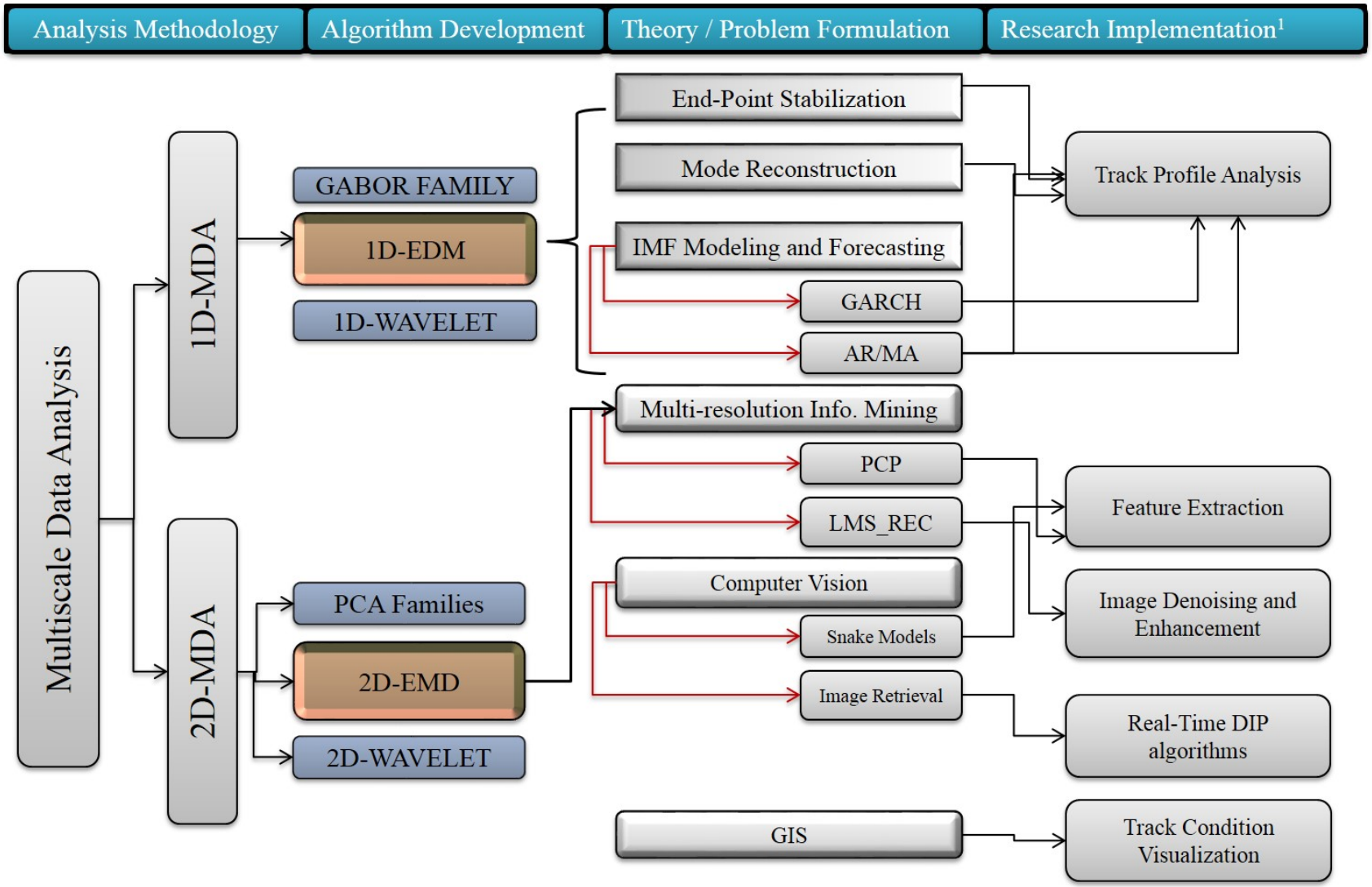
Edge Detection

Decomposition





Multiscale Data Analysis



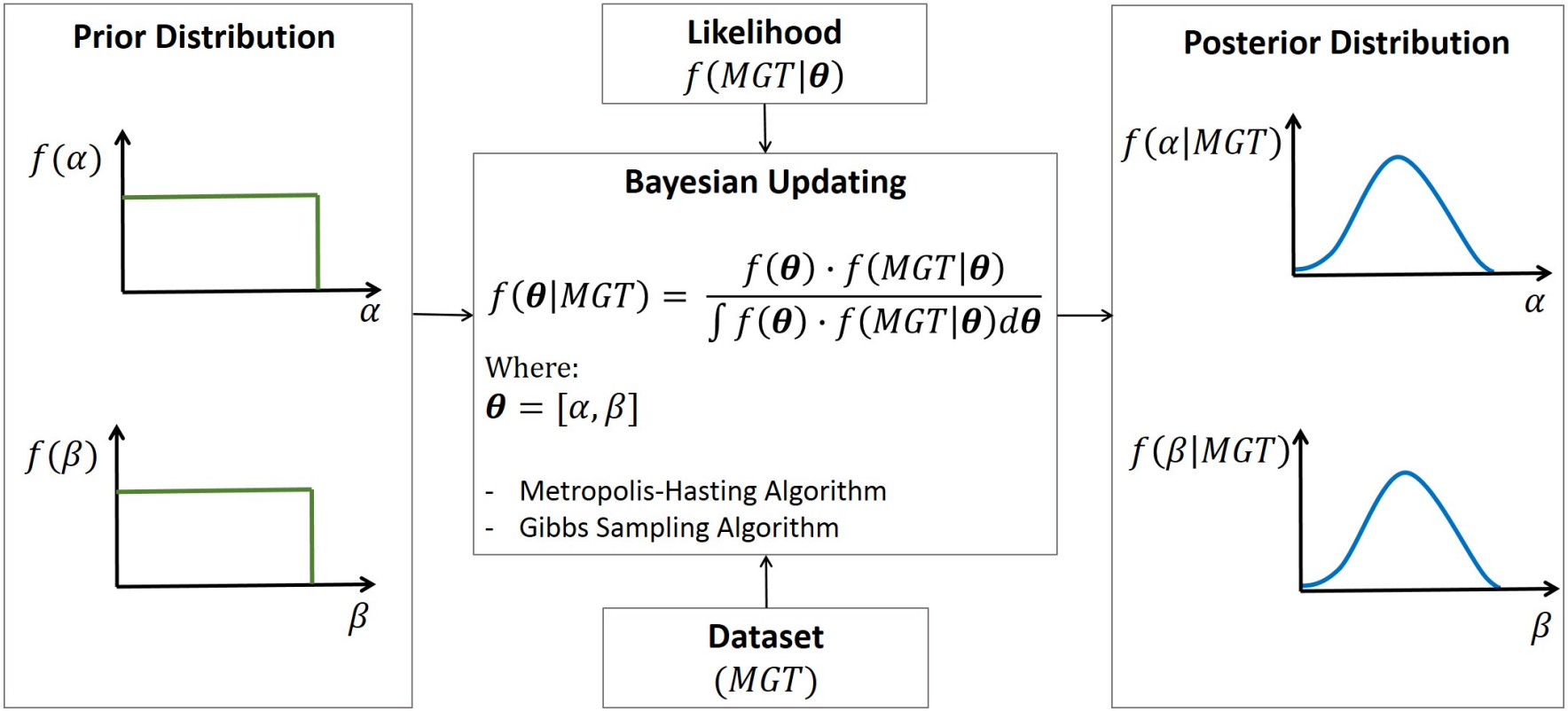


Figure 2: General Bayesian Model

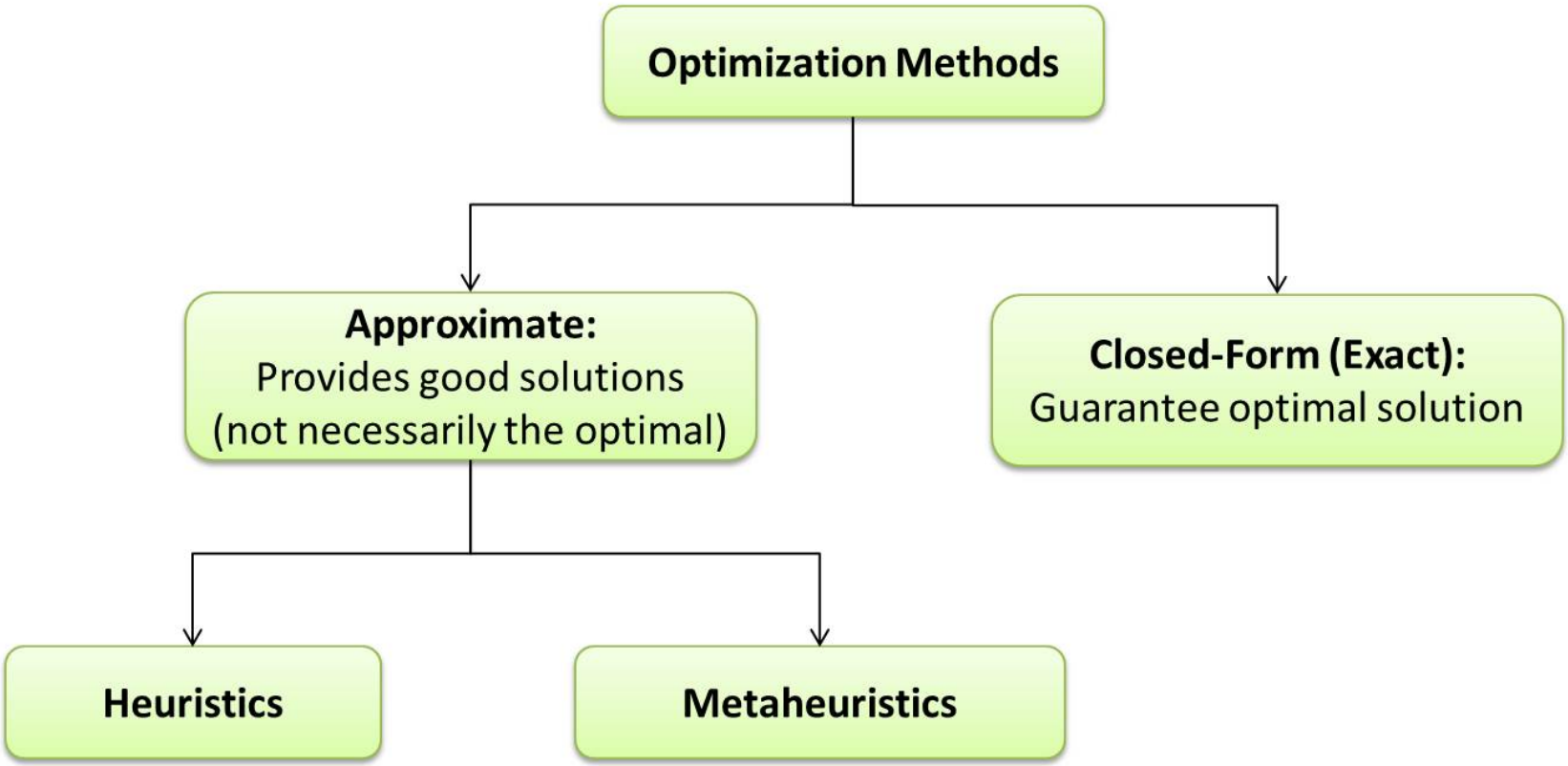


Figure 3: Classification of Optimization Techniques

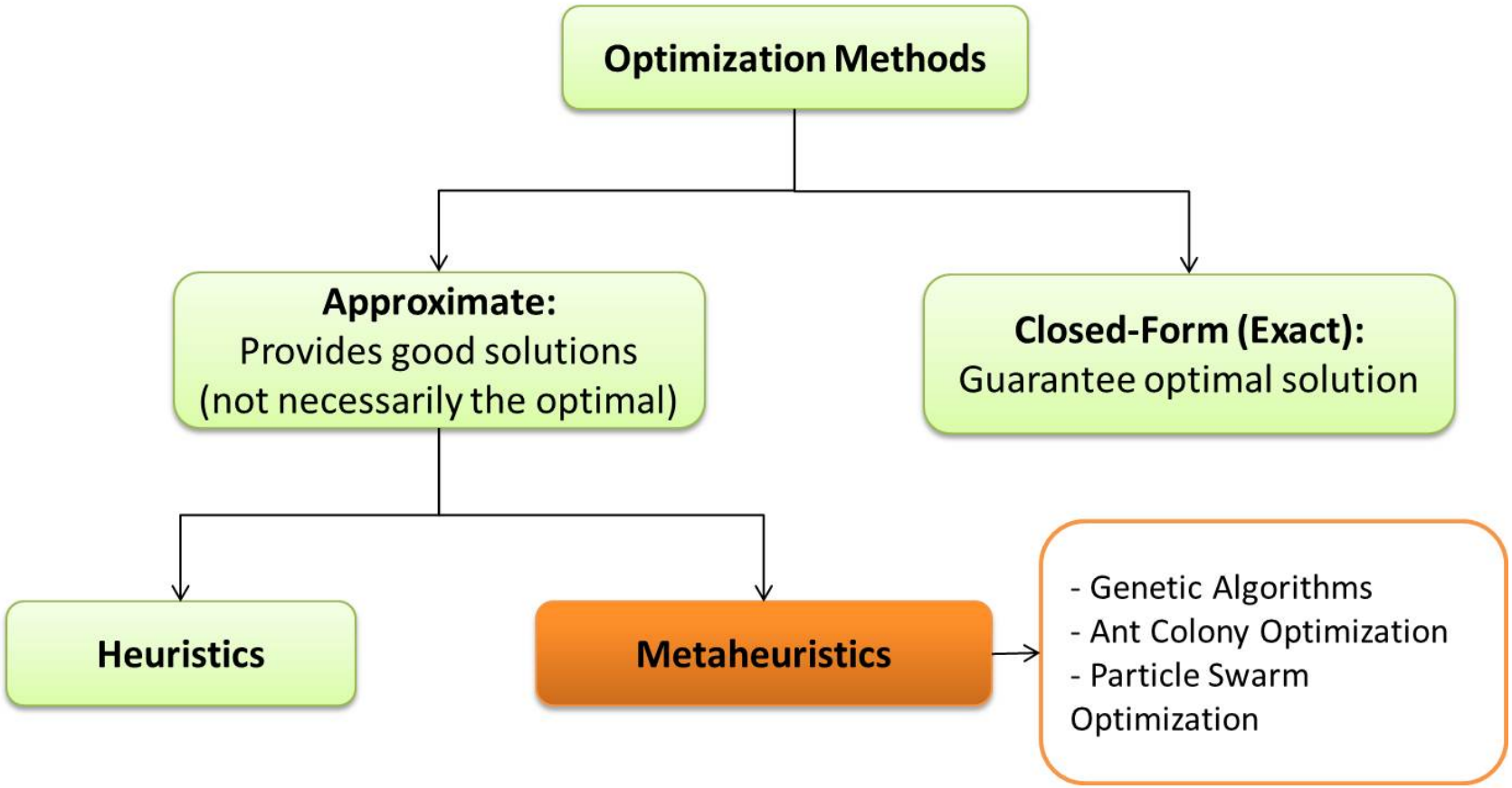


Figure 3: Classification of Optimization Techniques

Advantages

- Adequate for solving NP-Hard problems.
- Mainly used for:
  - Feature extraction.
  - Dimension reduction.
- Potential to address multi-objective problems.

Challenges

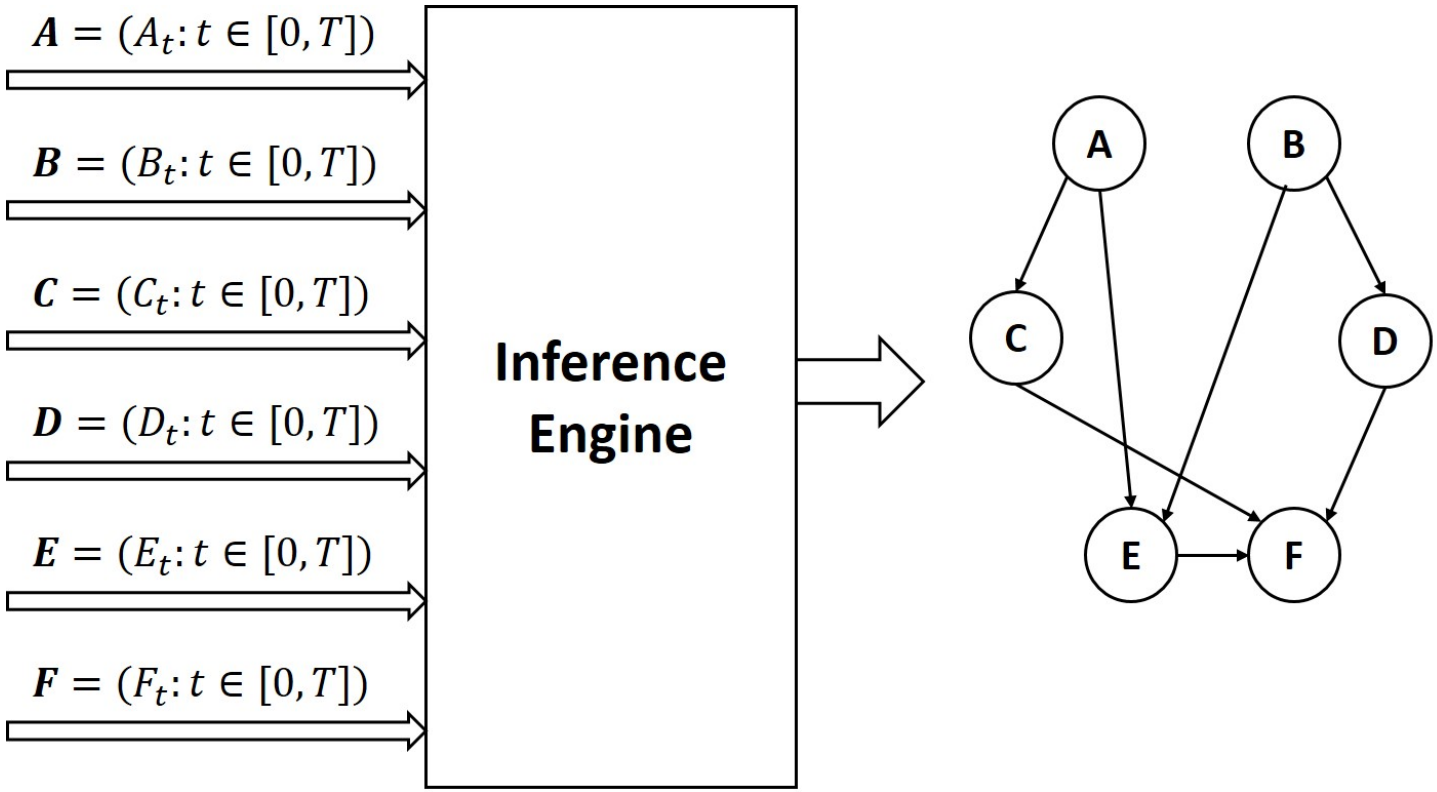
- The fitness function or the processed data is noisy.
- The fitness function suffers from approximation errors.

Puzzles in Multivariate Deterioration Models

- Do correlation change with the degree of Geometry/Rail Defects.
- To what extent.
- At what stages.
- Is the correlation uniform across the entire data set?

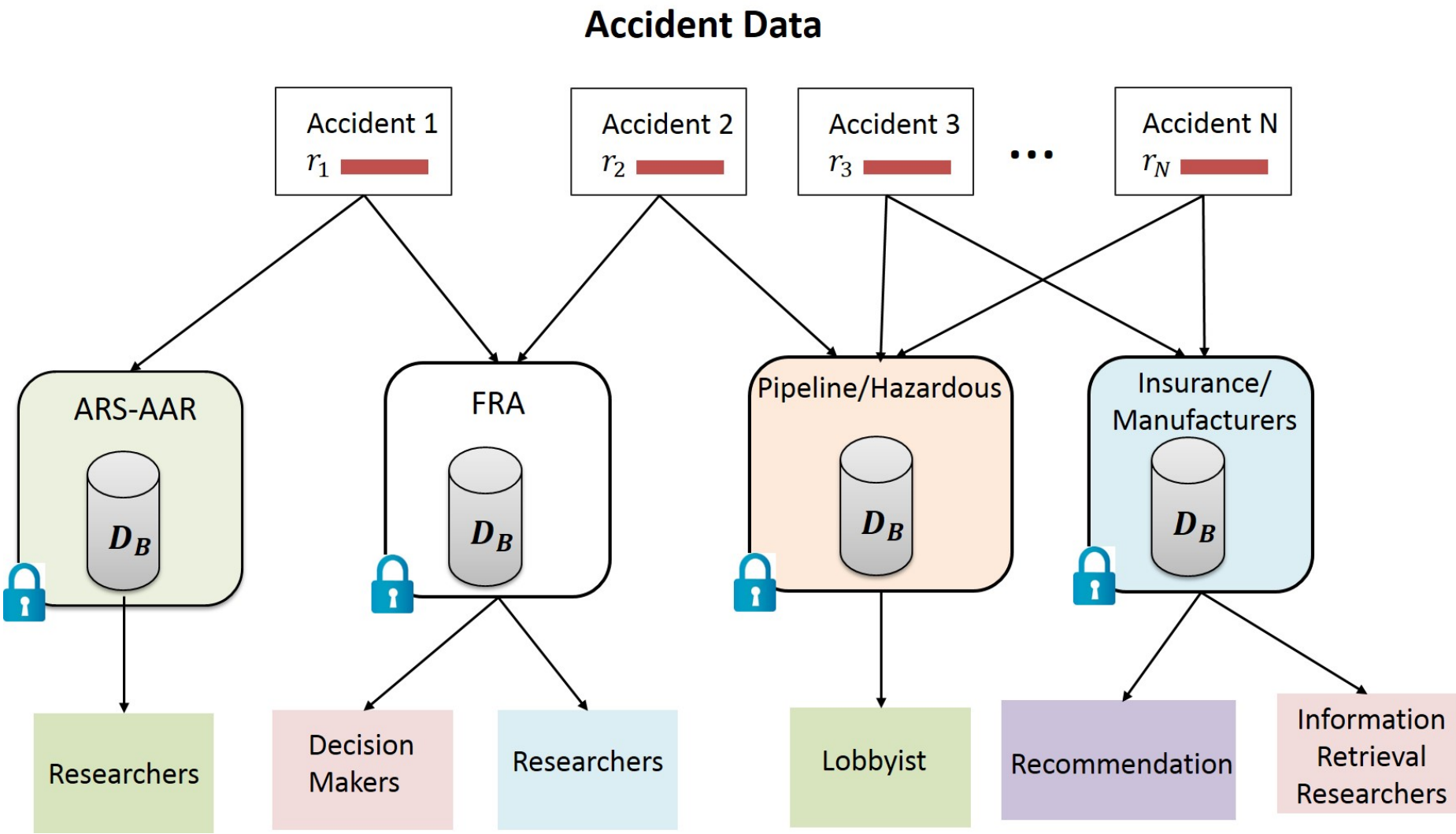
- Linear Correlation.
- Tail Dependence.
- Rank Correlation.

Beyond Correlation: Causation

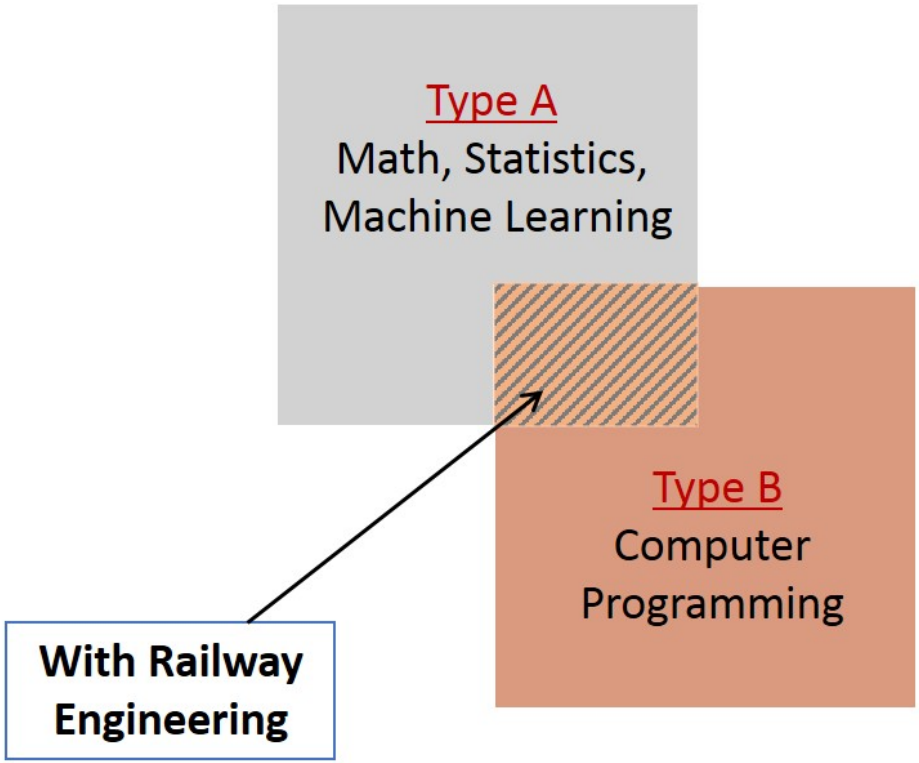


Idea: Map a set of K time series to a directed graph with K nodes where an edge is placed from a to b if the past of a has an impact on the future of b

Massive Amount of Data (Tank Safety)

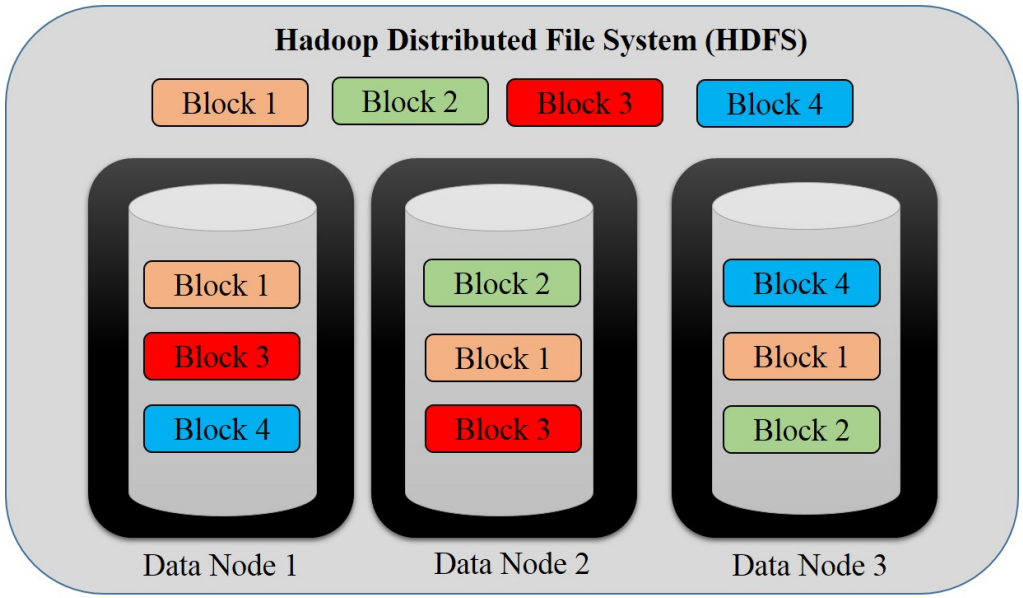


Traditional analytical techniques are inadequate in analyzing and drawing conclusions

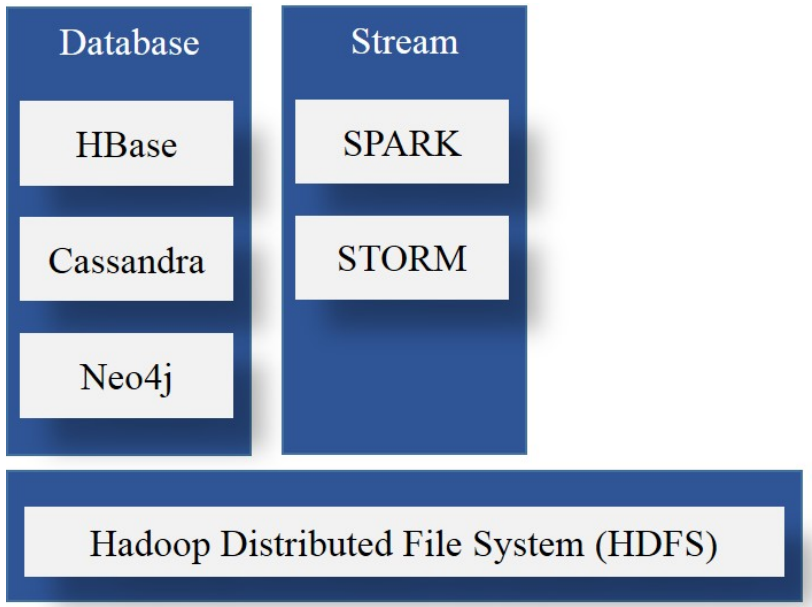


Data Storage - HDFS

- Stores large number of large files across multiple machines.
- Files smaller than 500MB may not be store efficiently
- Data is divided into blocks and replicated over data nodes.



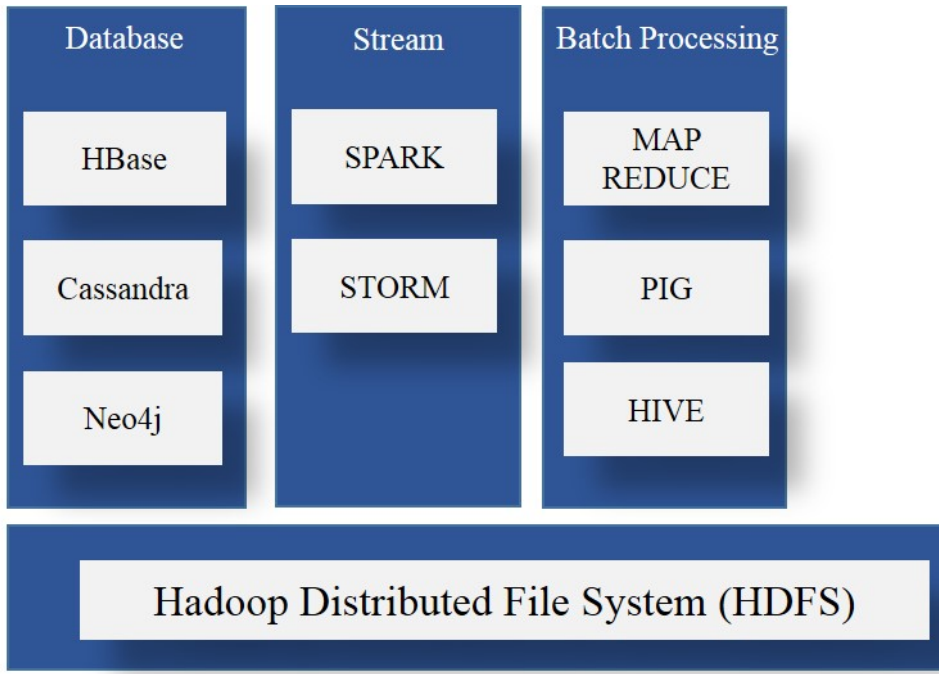
- Tables:
  - HBase - stores millions of columns and billions of row information.
  - MongoDB, Neo4j
- Images:
  - Cassandra – developed by facebook.
- Storing Streaming Data – twitter feed, facebook, linkedIn cannot be stored.
  - Rely on real-time, in-memory processing.



Data Cleaning and Processing

- Real World: data is messy, unclean and error prone.
- How do you handle:
  - Erroneous / Conflicting Data. - Standardize / normalize Data. - Missing Data. - Merge tables.
- Data cleaning is time consuming (40 to 60%) but very important.

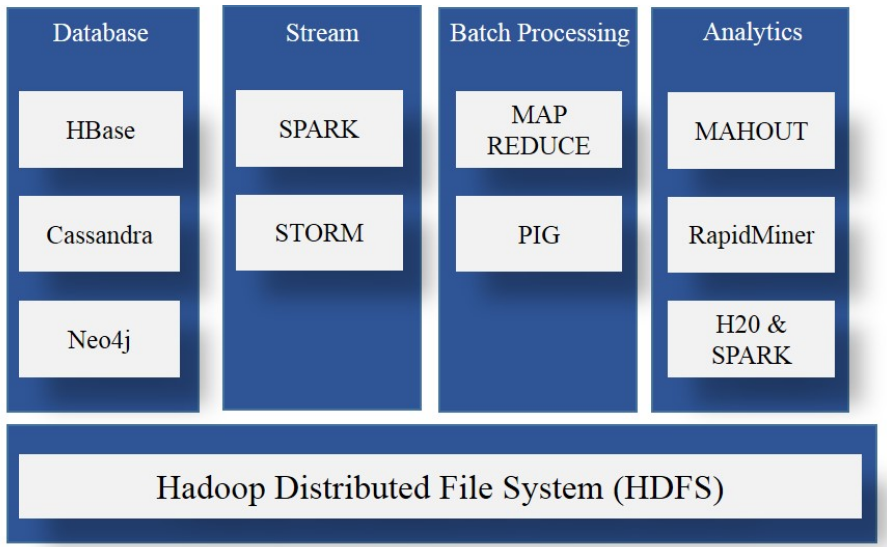
- Take advantage of programming models for processing large data datasets.
- PIG & HIVE – High level scripting language for creating MapReduce programs.



Learn from High Quality Data

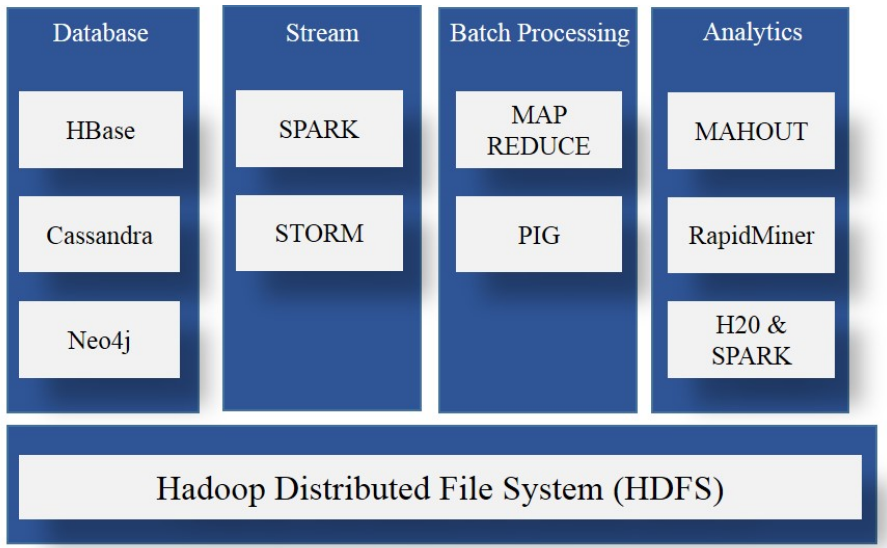
- Build good models.
- Fast Scalable Machine Learning – MAHOUT, H20 & SPARK
- Text Analytics- RapidMiner

Quote: The future is all about doing data analytics - it's not about the software. (Massimo Fubini, founder, ContactLab).



Big Data Interactive Visualization

- Visualizing gigabites and terabytes of data in a browser or phone?
- Data should be summarized.
- Provides the platform for exploring multivariate datasets



Thank You.