

# How Big Data Techniques are Changing Pavement and Railway Track Engineering Condition Assessment Protocols

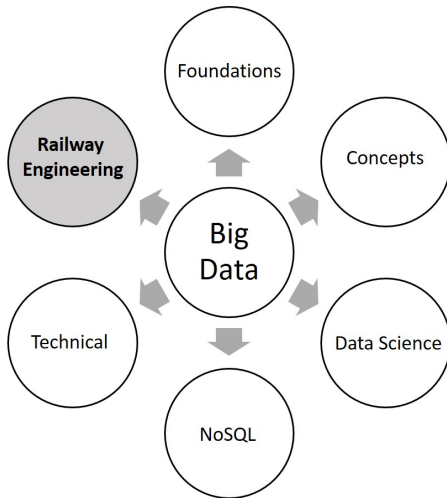
**Nii Attoh-Okine, Ph.D., P.E.,F.ASCE**

Professor

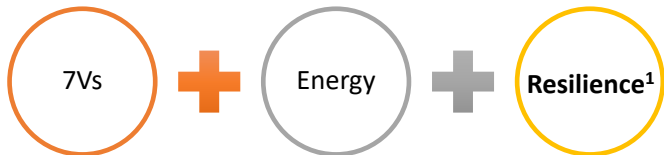
Department of Civil and Environmental Engineering

University of Delaware

Newark, DE, USA



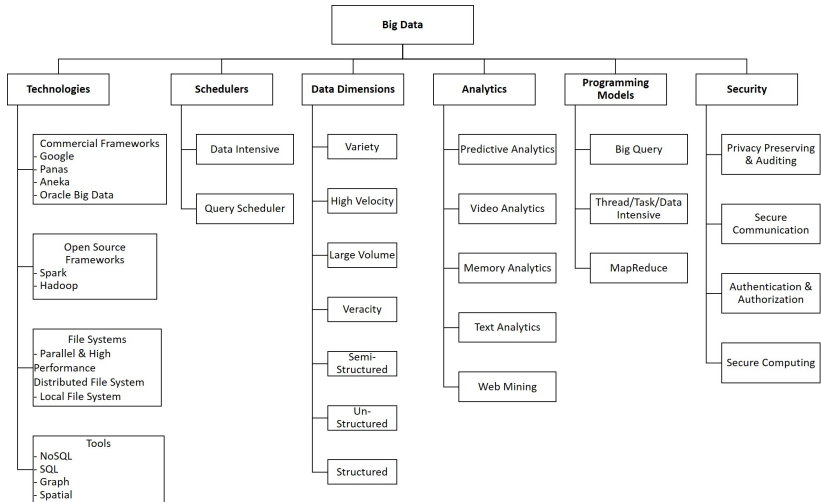
# Issues with small data sets



---

<sup>1</sup>N. Attoh-Okine, Resilience Engineering: Models and Analysis. Cambridge University Press, 2016.

# Taxonomy of Big Data



# Classification of Track Vertical Defects Upon their Wavelengths (Salvador et al.,2016)

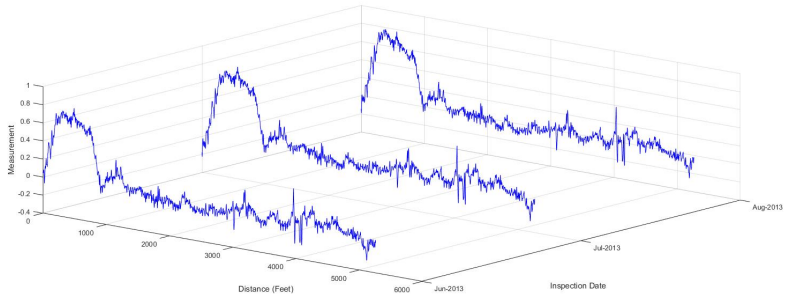
Type of defect	Classification	Wavelength range (m)	Examples of defects
Rail corrugation and isolated rail defects	Very short	0.03–0.06	Rail joints, very short wavelength rail corrugation, small squats
	Short	0.06–0.25	Short wavelength rail corrugation, medium size squats
	Medium	0.25–0.60	Medium wavelength rail corrugation, large squats, turnout frogs
	Long	0.60–2	Long wavelength rail corrugation, ballast fouling
Loss of track vertical alignment	Short	2-25	Changes on track vertical stiffness
	Medium	25-70	Medium wavelength vertical misalignment
	Long	70-120	Long wavelength vertical misalignment

# Comparison Between Fourier, Wavelet and Hilbert-Huang Transform (HHT)

	<b>Fourier</b>	<b>Wavelet</b>	<b>HHT<sup>2</sup></b>
Basis	A priori	A priori	Adaptive
Frequency	Convolution: global, uncertainty	Convolution: re- gional, uncertainty	Differentiation: local, certainty
Presentation	Energy-frequency	Energy time- frequency	Energy time- frequency
Nonlinear	Not easily defined	Not easily defined	Not easily defined
Nonstationary	No	Yes	Yes
Feature Extraction	No	Yes	Yes
Theoretical Base	Theory complete	Theory complete	Empirical

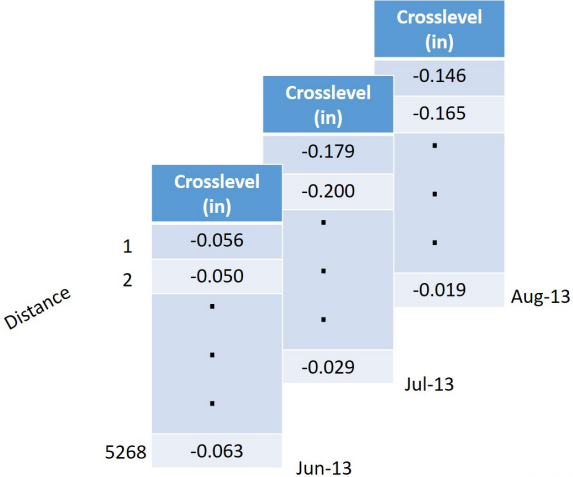
<sup>2</sup>N. E. Huang and N. O. Attoh-Okine, The Hilbert-Huang Transform in Engineering. CRC Press, 2005.

# Tensor – Crosslevel





# Tensor – Crosslevel



5268 points x 1 measure x 3 months

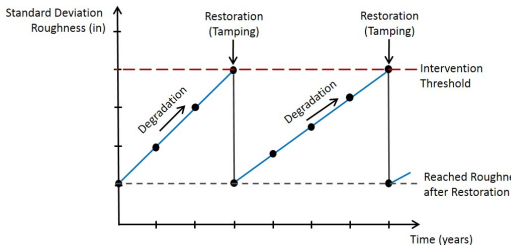
# Case Studies

# Track Geometry Degradation Model: Linear Representation

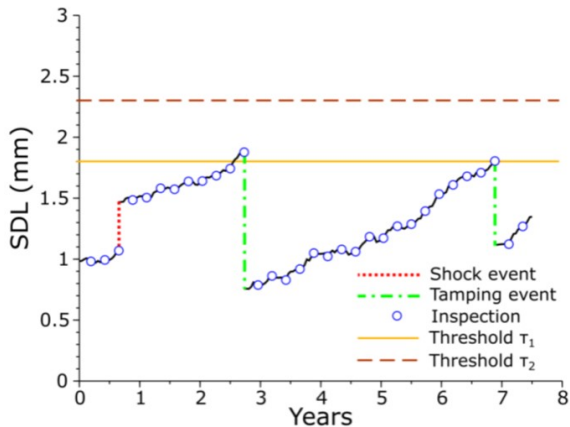
$$\sigma_s = \theta_1 + \theta_2 t + \varepsilon$$

Where:

- $\sigma_s$  : Standard deviation of surface (in)
- $\theta_1$  : Intercept (in)
- $\theta_2$  : Degradation rate (in/month)
- $t$  : Time (months)
- $\varepsilon$  : White noise  $\sim N(0, s)$



# Track Geometry Degradation-Shock Model Representation<sup>3</sup>



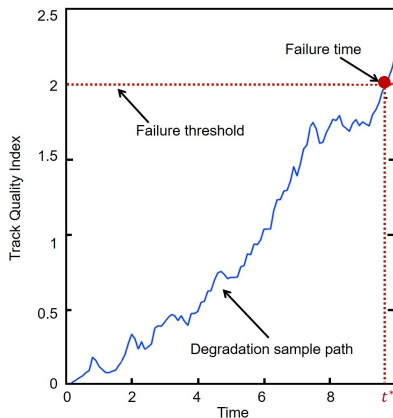
<sup>3</sup>Soleimanmeigouni et al.,2016

# Predictive Model in Track Geometry Degradation: Stochastic Process<sup>4</sup>

$$W(t) = \omega_0 + \mu t + \sigma B(t).$$

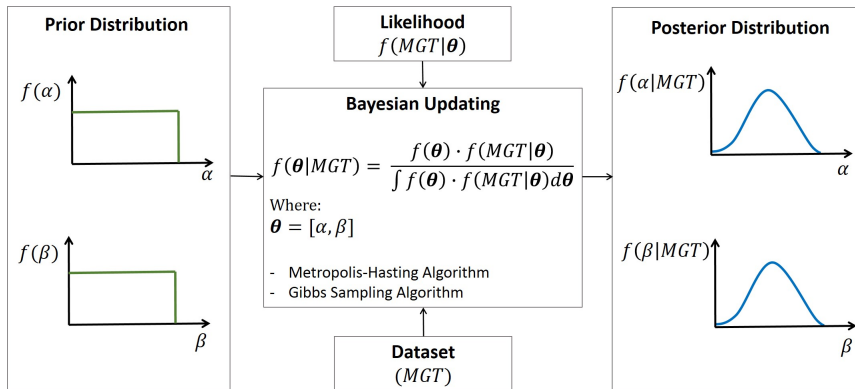
Where:

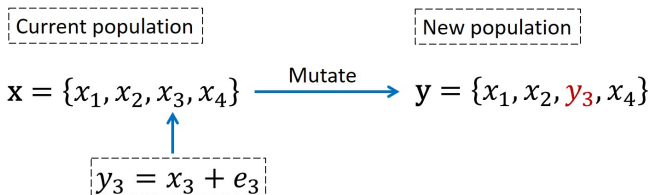
- $W(t)$ : Degradation at time  $t$
- $\omega_0$ : Initial degradation
- $\mu$ : Deterioration rate (drift parameter)
- $\sigma$ : Diffusion coefficient
- $B(t)$ : Standard Brownian motion



<sup>4</sup>Silvia A. Galván-Núñez, University of Delaware

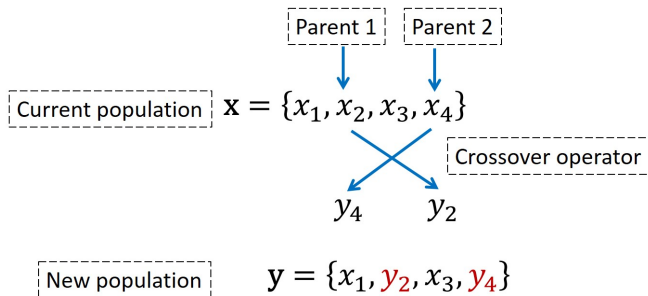
# Parameter Estimation: Markov Chain Monte Carlo





The new population  $\mathbf{y}$  is accepted with probability  $\min(1, r_m)$  according to the Metropolis-Hastings rule:

$$r_m = \frac{f(\mathbf{y}) T(\mathbf{x} | \mathbf{y})}{f(\mathbf{x}) T(\mathbf{y} | \mathbf{x})} = \exp \left\{ (H(\mathbf{y}_k)) - \frac{H(\mathbf{x}_k)}{t_k} \right\} \frac{T(\mathbf{x} | \mathbf{y})}{T(\mathbf{y} | \mathbf{x})}$$

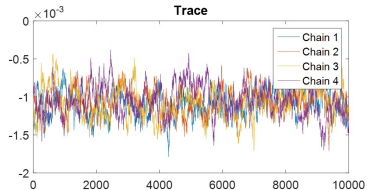


The new population  $\mathbf{y}$  is accepted with probability  $\min(1, r_c)$  according to the Metropolis-Hastings rule:

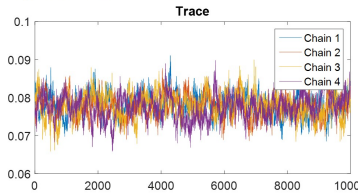
$$r_c = \frac{f(\mathbf{y}) T(\mathbf{x} | \mathbf{y})}{f(\mathbf{x}) T(\mathbf{y} | \mathbf{x})} = \exp \left\{ -\frac{H(\mathbf{y}_i) - H(\mathbf{x}_i)}{t_i} - \frac{H(\mathbf{y}_j) - H(\mathbf{x}_j)}{t_j} \right\} \frac{T(\mathbf{x} | \mathbf{y})}{T(\mathbf{y} | \mathbf{x})}$$



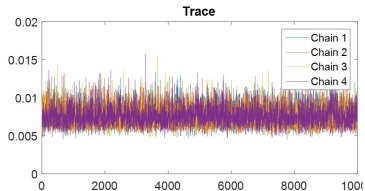
**a) Intercept**



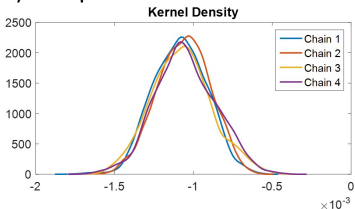
**b) Deterioration Rate**



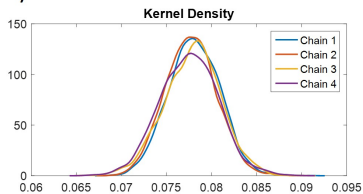
**c) White Noise**



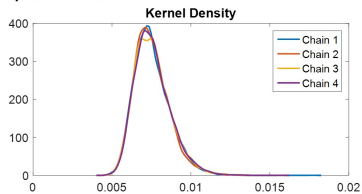
a) Intercept



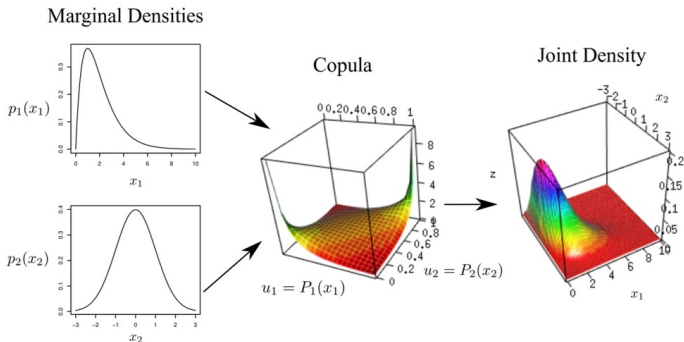
b) Deterioration Rate



c) White Noise



- Copulas are functions that combine or link multiple distribution functions to their univariate marginal distribution functions.



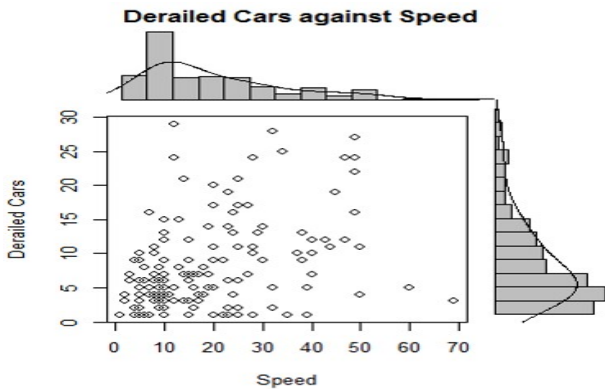
- Copulas are multivariate distribution functions with uniform margins on the unit interval  $[0,1]$

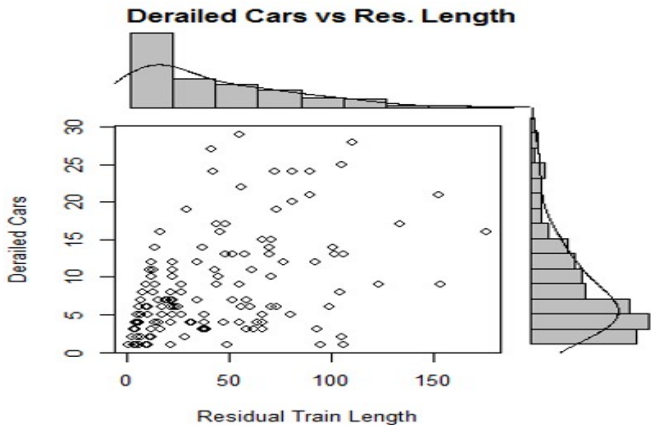
- Copulas are suitable for modelling tail dependence and skewness.
- Copulas allow for the separate modelling of marginals and dependence structure.
- Since copulas are invariant under monotone transformations, concordance measures such as Kendall's Tau and Spearman's Rho are more suitable for evaluating dependence.
- Copulas are important because of Sklar's Theorem.

- 161 observations were used for the copula analysis.
- Explanatory variables (Speed, Residual Length, Track Quality)
- Response variable (Number of Derailed Cars)

---

<sup>5</sup>Emmanuel Martey, University of Delaware





# Some Bivariate Copula Families

Copula	Properties
Normal/Gaussian (N)	Tail symmetric, no tail dependence
Student t-copula (t)	Tail-symmetric, tail dependence
Clayton (C)	Tail-asymmetric, Suitable for modelling lower tail dependence
Gumbel (G)	Tail-asymmetric, Suitable for modelling upper tail dependence
Joe (J)	Tail-asymmetric, Suitable for modelling lower tail dependence
Frank (F)	Tail-symmetric, no tail dependence, Tends to work well when tail dependence is very weak.
Clayton-Gumbel (BB1)	Tail-asymmetric, Suitable for different non-zero upper and lower tail dependence
Joe- Clayton (BB7)	Tail-asymmetric, Suitable for different non-zero upper and lower tail dependence
Rotations of Archimedean copulas	Suitable for modelling various forms of negative dependence



# Topological Data Analysis

- **Data-Driven** Approach.
- Studying complex high dimensional data **without any assumptions or feature selections**.
- Shape has meaning; **extracting shapes (patterns)** of data.
- Qualitative and quantitative **summaries** of the data are provided.
- Especially, TDA using **persistent homology** provides **threshold-free** analysis.

- Apply topology to develop tools for studying qualitative features of data.
- Recover topological and geometric information from sampled data.
- Topology is the branch of mathematics which concerns itself with the study of shapes and its properties.

Properties of geometric objects that are invariant under “continuous deformations”.

- Bending
- Twisting
- Stretching
- But not tearing

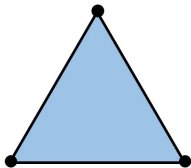
- Number of connected components.
- Number of cycles.
- Number of voids.



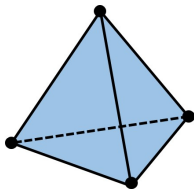
0-simplex



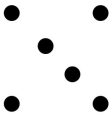
1-simplex



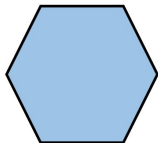
2-simplex



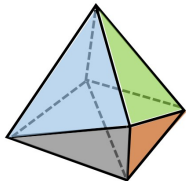
3-simplex



0-cycle

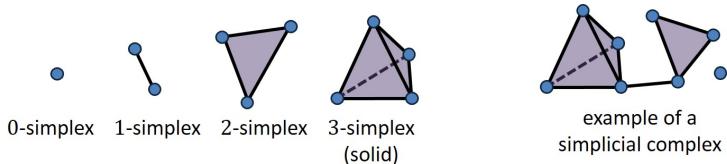


1-cycle

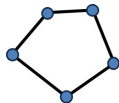


2-cycle

A **simplicial complex** is built from points, edges, triangular faces, etc.



**Homology** counts components, holes, voids, etc.



hole



void  
(contains faces but empty interior)

Homology of a simplicial complex is computable via linear algebra.

<sup>6</sup>Wright and Lesnick (2014)

- 0 – simplex point  $\Delta^0$
- 1 – simplex line segment  $\Delta^1$
- 2 – simplex triangle  $\Delta^2$
- 3 – simplex tetrahedron  $\Delta^3$



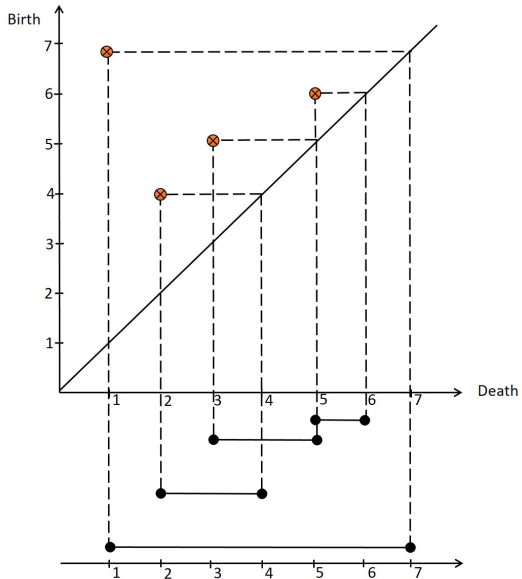
## Simplicial Complex

A simplicial complex is a finite collection of set of simplices.

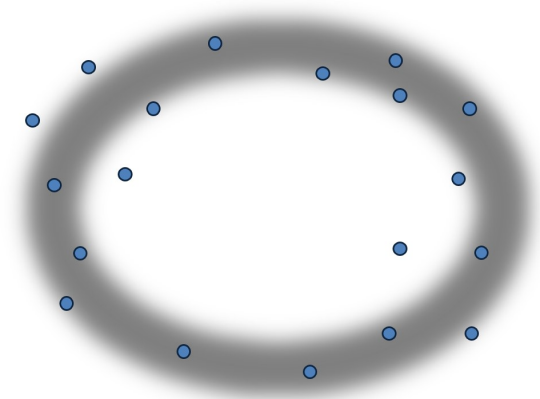
## Simplicial Homology

- Simplex, simplicial complex.
- Chain group, cycle group, boundary group.
- Homology group, homology class, Betti number.

# Persistence Diagram



**Example:** What is the shape of the data?<sup>7</sup>

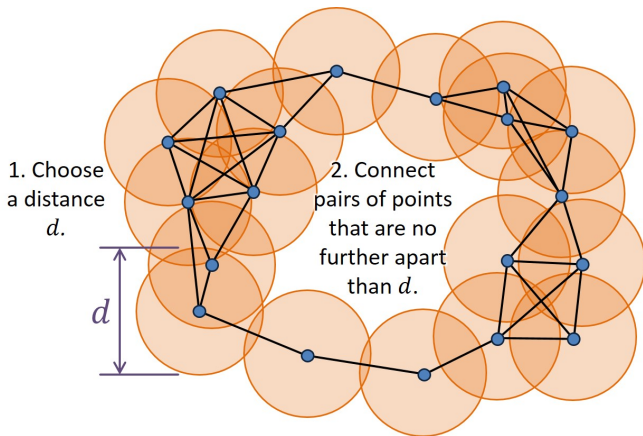


**Problem:** Discrete points have trivial topology.

---

<sup>7</sup>Wright and Lesnick (2014)

**Idea:** Connect nearby points<sup>8</sup>.

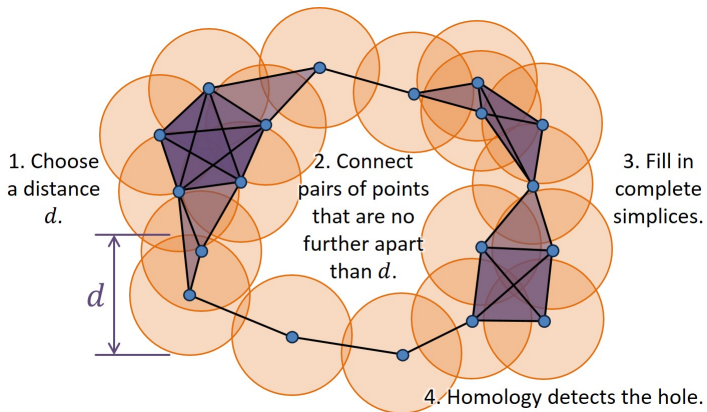


**Problem:** A graph captures connectivity, but ignores higher-order features, such as holes.

---

<sup>8</sup>Wright and Lesnick (2014)

**Idea:** Connect nearby points, build a simplicial complex<sup>9</sup>.



**Problem:** How do we choose distance  $d$ ?

---

<sup>9</sup>Wright and Lesnick (2014)



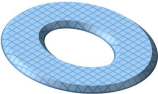
Homology is a mathematical formalism used to define and identify basic topological features called **holes**.

- 0-dimensional holes related to the gaps between connected components.
- 1-dimensional – can be viewed as tunnels (like a hole in a donut).
- 2-dimensional – holes cavities (inside balloon).
- Homology class – means individual holes.

The topological features detected by simplicial homology correspond to  $n$ -dimensional holes. The number of holes is known as **Betti Number**.



The number of  $k^{th}$  order

	 <b>Point</b>	 <b>Circle</b>	 <b>Torus</b>
$\beta_0$	1	1	1
$\beta_1$	0	1	2
$\beta_2$	0	0	1

- $0^{th}$  order holes: clusters.
- $1^{st}$  order holes: holes.
- $2^{nd}$  order holes: voids.

Wiley Series in Operations Research  
and Management Science

# BIG DATA AND DIFFERENTIAL PRIVACY

## ANALYSIS STRATEGIES FOR RAILWAY TRACK ENGINEERING



Nii Attoh-Okine

WILEY

Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering, Nii Attoh-Okine – John Wiley & Sons, Incorporated (2017).

---

---

### Book Chapters

---

1. Introduction
  2. Data Analysis - Basic Overview
  3. Machine Learning - Basic Overview
  4. Basic Foundations of Big Data
  5. Hilbert-Huang Transform, Profile, Signal, and Image Analysis
  6. Tensors – Big Data in Multidimensional Settings
  7. Copula Models
  8. Topological Data Analysis
  9. Bayesian Analysis
  10. Basic Bayesian Nonparametrics
  11. Basic Metaheuristics
  12. Differential Privacy
-

## A comprehensive introduction to the theory and practice of contemporary data science analysis for railway track engineering

Featuring a practical introduction to state-of-the-art data analysis for railway track engineering, *Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering* addresses common issues with the implementation of big data applications while exploring the limitations, advantages, and disadvantages of more conventional methods. In addition, the book provides a unifying approach to analyzing large volumes of data in railway track engineering using an array of proven methods and software technologies.

Dr. Attoh-Okine considers some of today's most notable applications and implementations and highlights when a particular method or algorithm is most appropriate. Throughout, the book presents numerous real-world examples to illustrate the latest railway engineering big data applications of predictive analytics, such as the Union Pacific Railroad's use of big data to reduce train derailments, increase the velocity of shipments, and reduce emissions.

In addition to providing an overview of the latest software tools used to analyze the large amount of data obtained by railways, *Big Data and Differential Privacy: Analysis Strategies for Railway Track Engineering*:

- Features a unified framework for handling large volumes of data in railway track engineering using predictive analytics, machine learning, and data mining
- Explores issues of big data and differential privacy and discusses the various advantages and disadvantages of more conventional data analysis techniques
- Implements big data applications while addressing common issues in railway track maintenance
- Explores the advantages and pitfalls of data analysis software such as R and Spark, as well as the Apache™ Hadoop® data collection database and its popular implementation MapReduce

*Big Data and Differential Privacy* is a valuable resource for researchers and professionals in transportation science, railway track engineering, design engineering, operations research, and railway planning and management. The book is also appropriate for graduate courses on data analysis and data mining, transportation science, operations research, and infrastructure management.

# Topic Modeling (Latent Dirichlet Allocation)

Collection of documents – identify underlying “topics” that organize collection.

What is it for?

Infer the latent structure behind the collection of **documents**.

Good for:

Document classification and retrieval.

# Key Assumptions behind the LDA Topic Model

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding generative process
  - Each document is assumed to be generated by this (simple) process
- A topic is a distribution over a fixed vocabulary
  - These topics are assumed to be generated first, before the documents
- Only the number of topics is specified in advance

Images are translated to words (visual words)

- Rail images
- Fasteners (worn, missing, polluted, ...)



## Digital Data: Challenges and Opportunities

- a) What is the most efficient way to store rail track information?
- b) How secure is the stored data against intervention by unauthorized agencies and individuals?
- c) How stable is the data storage platform?
- d) Ease of reproducibility
- e) Read and write capabilities

## Example

Transportation Technology Center Data

- Text  $\rightarrow$  Transform into binary (0, 1)
- Nucleotides are represented as

A C G and T

---

---

**Mapping Table**

---

<b>Binary - nts.</b>	<b>Binary - nts.</b>	<b>Binary - nts.</b>	<b>Binary - nts.</b>
0000 - AA	0100 - AC	1000 - AG	1100 - AT
0001 - CA	0101 - CC	1001 - CG	1101 - CT
0010 - GA	0110 - GC	1010 - GG	1110 - GT
0011 - TA	0111 - TC	1011 - TG	1111 - TT

---

Thank You.