Panel Discussion – Data Analytics and Computing Challenges

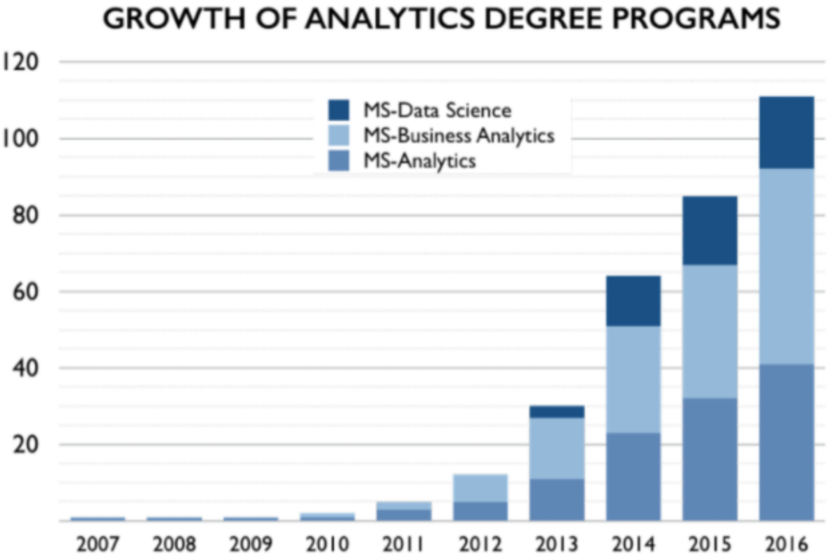
Venkat Gudivada, East Carolina University, Greenville, NC, USA
Torsten Ullrich, Fraunhofer Austria Research GmbH, Austria
Maaike de Boer, TNO & Radboud University, the Netherlands
Nuccio Piscopo, Engineering Ingegneria Informatica S.p.A., Italy
Jolon Faichney, Griffith University, Australia
Florence Nicol, ENAC, France

Evolution of Data Analytics

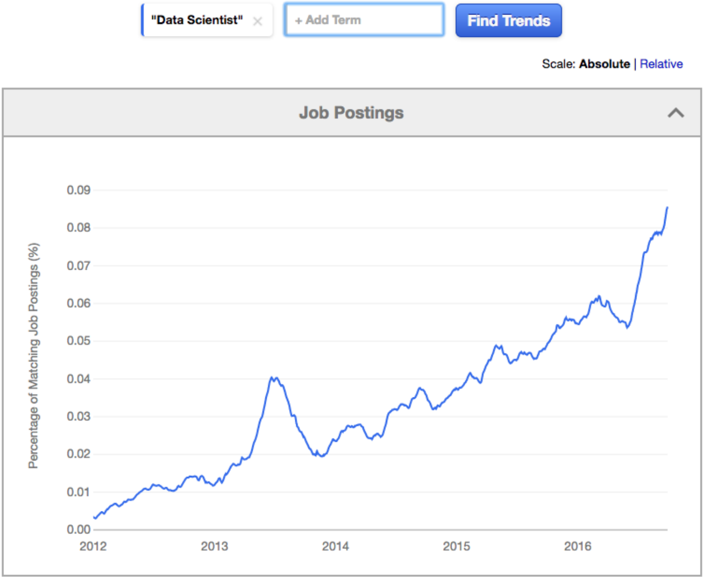
- SQL Analytics: RDBMS, OLTP, and OLAP
- Business Analytics: Business Intelligence (BI), Data Warehousing (OLAP Cubes, OLAP Servers), and Data Mining
- Visual Analytics
- Big Data Analytics
- Cognitive Analytics
- Traffic Analytics, Text Analytics, Spatial Analytics, Risk Analytics, and Graph Analytics
- Data Science

Types of Data Analytics

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics


GROWTH OF ANALYTICS DEGREE PROGRAMS
120
100
80
60
40
20
MS-Data Science
MS-Business Analytics
MS-Analytics
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

"Data Scientist" Job Trends



Job Level	Region	N	Base Salary 25%	Median	Mean	75%
Individual Contributor, Level 1	Northeast	25	$85,000	$95,000	$96,840	$100,000
	Middle U.S.	34	$83,125	$94,500	$92,662	$102,250
	West Coast	16	$94,250	$104,000	$100,813	$114,000
Individual Contributor, Level 2	Northeast	31	$116,750	$125,000	$127,000	$140,000
	Middle U.S.	42	$103,500	$123,000	$122,095	$135,000
	West Coast	36	$122,300	$130,000	$131,650	$140,000
Individual Contributor, Level 3	Northeast	27	$145,000	$155,000	$154,889	$170,000
	Middle U.S.	25	$135,000	$150,000	$150,960	$170,000
	West Coast	23	$147,000	$155,000	$164,957	$185,000
Manager, Level 1	Northeast	11	$131,000	$145,000	$148,545	$167,500
	Middle U.S.	17	$120,000	$125,000	$130,971	$147,500
	West Coast	8	$135,250	$142,500	$141,063	$147,375
Manager, Level 2	Northeast	20	$168,750	$185,000	$187,200	$200,000
	Middle U.S.	19	$175,000	$187,000	$186,211	$200,000
	West Coast	22	$185,000	$199,000	$202,773	$221,250
Manager, Level 3	Northeast	6	$228,750	$240,000	$260,000	$292,500
	Middle U.S.	5	$220,000	$220,000	$227,800	$240,000
	West Coast	7	$245,000	$253,000	$294,143	$293,000

@BurtchWorks #BWstudy

Highest rated for job satisfaction in Data Science

Data Science

Johns Hopkins University

Enroll Now

#1 MOST SATISFYING JOB

DATA SCIENTIST

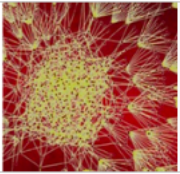
Median base salary: $110,000

Openings on Glassdoor: 4,100+

Big Data

UC San Diego

Enroll Now



#2 MOST SATISFYING JOB

DATA ENGINEER

Median base salary: $106,000

Openings on Glassdoor: 2,500+

Strategic Business Analytics

ESSEC Business School

Enroll Now

#3 MOST SATISFYING JOB

STRATEGY MANAGER

Median base salary: $130,000

Openings on Glassdoor: 1,800+

Data Challenges for Data Analytics

- Data quality
- Data provenance
- Differential privacy
- Big data-driven machine learning applications pose unique challenges

Machine Learning Challenges

- Data sparsity in feature space
- Data correlations
- Parallelization
- Decision trees, Bagging/Bootstrapped Aggregation, Random Forests, and Boosted Trees

Computing Challenges for Data Analytics

- High volume data
- Streaming data
- Real-time analytics
- In-memory analytics
- Incremental computation

Panel Summary - Data Analytics Challenges

- Data quality, differential privacy, and provenance
- Data heterogeneity
- Information extraction from multimedia big data
- Reproducibility of analysis
- Leveraging open and linked data
- Functional data analysis to overcome the inadequacy of multivariate statistical techniques

Anomaly Detection Using Deep Learning

Dr Jolon Faichney

School of Information and Communication Technology

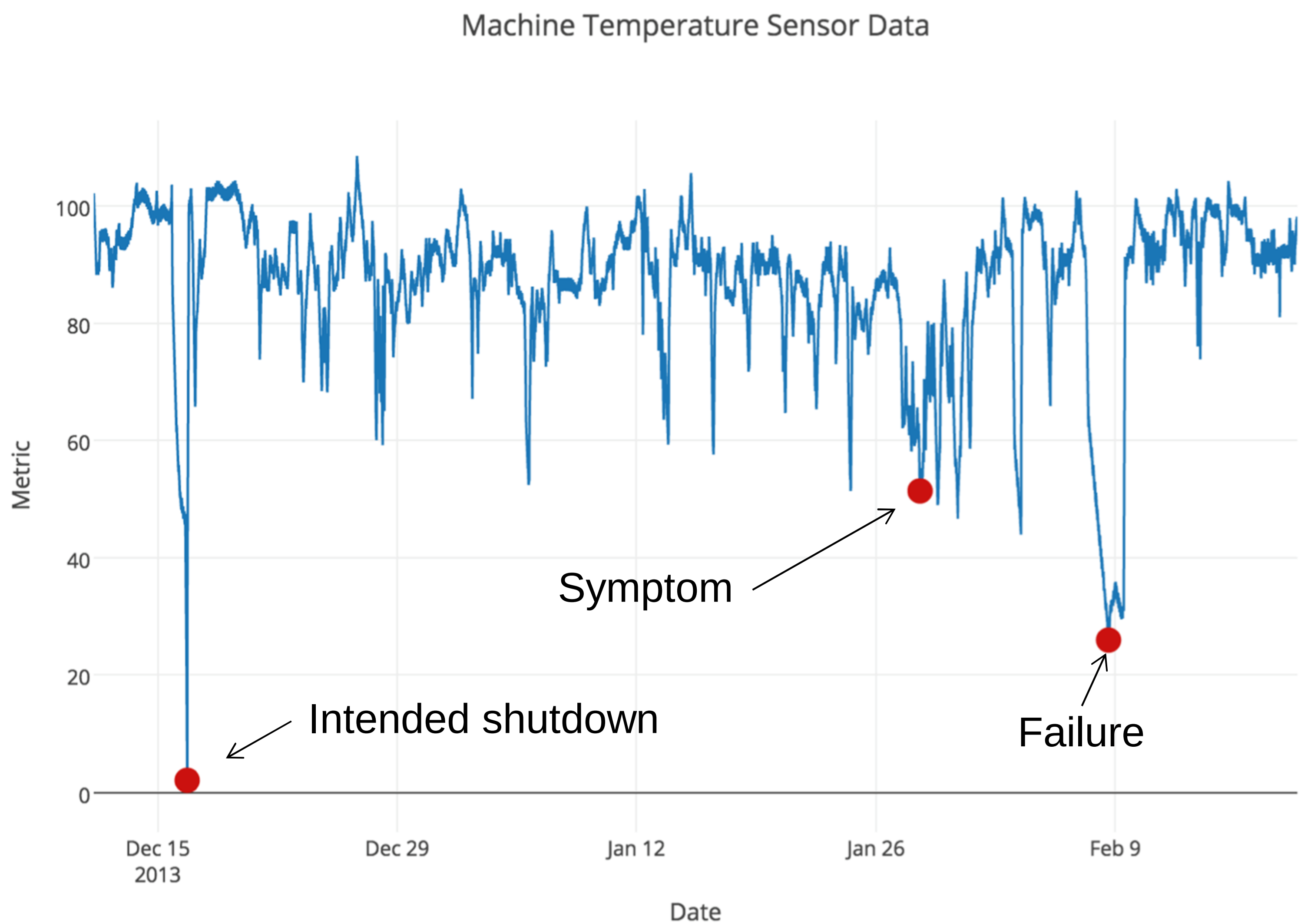
Griffith University, Australia

j.faichney@griffith.edu.au

What is Anomaly Detection?

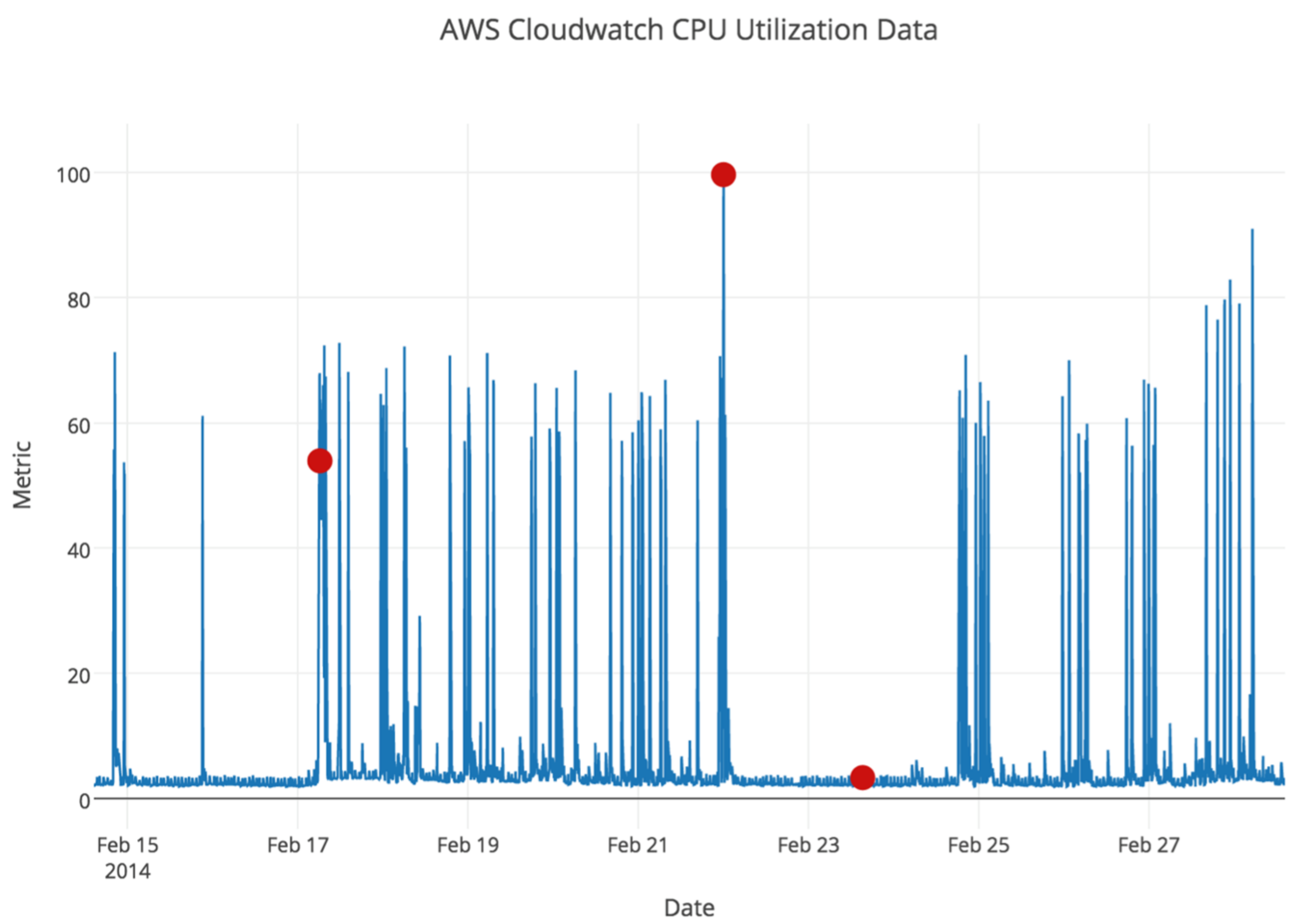
- Historically faults were detected by analysing logs
- Today, logs are too large to manually analyse in realtime
- Changes in data may indicate that a fault will occur before it has occurred
- What is considered an anomaly may change over time

Machine Temperature



Amazon Web Services

- Can you pick the anomalies?



Anomaly Algorithms

- Etsy.com
  - *Skyline*
  - A set of simple detectors and a voting scheme
- Twitter
  - *ADVec*
  - Can detect short and long term trends
- Numenta
  - *HTM*
  - Hierarchical Temporal Memory

Hierarchical Temporal Memory

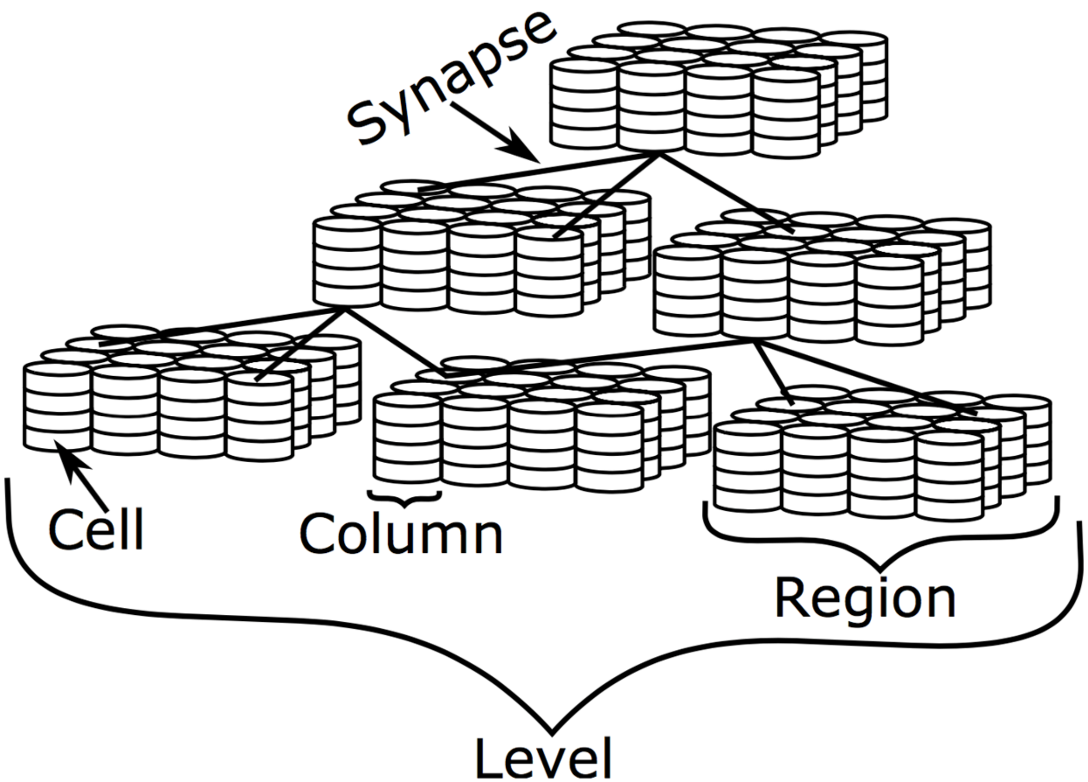


Fig. 1: Depiction of HTM, showing the various levels of detail.

Anomaly Data Set

- NAB – Numenta Anomaly Benchmark
  - AWS CloudWatch
  - Machine Temperature Sensor
  - NYC Taxi
  - Tweets
  - Traffic
  - AdExchange
  - Artificial Data

Results

Detector	Standard	Reward Low FP	Reward Low FN
Numenta HTM	64.7	56.5	69.3
Twitter AdVec	47.1	33.6	53.5
Template Matching	**41.02**	**43.15**	**38.44**
Etsy Skyline	35.7	27.1	44.5
Random	16.8	5.8	25.9
Null	0	0	0

Topics for Discussion

- Can machines reliably find anomalies?
- Can machine learning be implemented for real time anomaly detection at levels of scale?

THE FUTURE OF MULTIMEDIA SYSTEMS

Panel on Data Analytics and Computing Challenges | Maaike de Boer

TNO innovation for life

MULTIMEDIA SYSTEMS NEED TO BE SELF-EXPLAINABLE DESPITE OF (POSSIBLE) LOWER PERFORMANCE

High performing deep learning systems
vs.
Lower performing explainable systems
Or can we use the best of both (and how)?

SCALABLE SOLUTIONS

Assume a user query in a multimedia system has no match to pre-trained detectors (words used to index an item with)

What to do?

- We should pre-train as many concept detectors as possible (opposed to a few high-performing detectors) to have some match
- We should focus on semantic decomposability of a query
- Other suggestions?

ENGINEERING

www.eng.it

ENGINEERING

Data Analytics and Computing Challenges

Panel at AllData Conference, Venice – April 26, 2107

Nuccio Piscopo
Data Scientist - Big Data & Analytics Competency Center
Engineering Ingegneria Informatica S.p.A.

www.eng.it

Big Data Analytics:

- Logics (data intelligence) moves to **functional programming** paradigm
- Data transfer from structure/unstructured/semi-structured runs on **dataframes**

……. **so, might data modeling change by design elements/construct?**

Metadata:

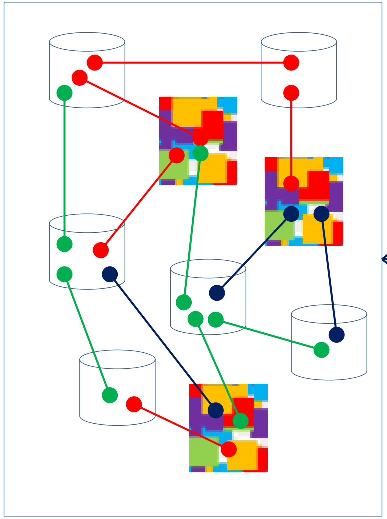
- **Vector Construct** V = (v_1, v_2, v_3, …). Vector elements map heterogeneous data topology.

Metamodel:

- **Set of vectors** covering sources morphology through explicit formal specifications of the terms and relationships in the datasource domain (ontology)

Prescriptive Metamodel Framework – Ontology vs. Vectors

Structured/Unstructured



Ontology

Functional Layer - Vector Identifier

Category: dataflow, datasource, dataset , spare source
Element: time, sourcetype, entitytype, provenance, destination, ext, …
Construct: vector F = $[f_{i,j}]$ i,j Є N

Category: record, table, spare info
Element: data records, fields record, spare field
Construct: vector R = $[r_{i,j}]$ i,j Є N

Category: Datafile size, frequency, transferring method, owner, approvals, …
Element: dataflow properties, source properties
Construct: vector P = $[p_{i,j}]$ i,j Є N

Data Mapping Layer - Metamodel

⟨**F**| = (Source, Date, Category, Type, Destination, Version, …)

⟨**R**| = (Checkdate, AthleteID, Age, Height, Weight,…)

…………………

Category: metadata
Element: vectors
Construct: metamodel Mv = $\{\mathbf{f_{i,j}}, \mathbf{r_{i,q}}, \mathbf{p_{i,l}}\}$

<source><date><category><type><destination><version>

CheckDate	AthleteID	Age	Height	Weight	PCapacity	Speciality
20150710	756843	25	185	79	5,2	100Ms
20150109	154647	34	177	75	8,2	Swimming
20150815	875643	28	182	83	5,6	High-Jump
20151121	985641	23	190	88	6,4	800Ms
20151207	867532	25	179	55	8,1	Swimming
20151116	487532	30	206	96	7,9	Volley-ball
20150928	675843	26	181	62	7,5	Pole-Vault
20151220	745301	21	188	79	6,7	200M
20151216	564732	22	180	68	6,9	Marathon
20160710	357843	26	180	75	6,5	800Ms
20160109	559647	22	187	84	7,2	Swimming
20160815	975623	28	190	90	7,7	Basket
20161121	688642	24	193	89	6,9	Pole-Vault
20161207	267634	20	187	86	7,4	100Ms
20161116	785521	21	185	83	8,0	Swimming
20160928	738311	19	190	90	7,3	High-Jump
20161220	185205	27	189	88	7,5	Basket
20161216	362759	31	183	89	7,1	Volley-ball

Prescriptive Analytics – Machine-Learning

Distillation
Data Quality
Source morphology changes
Data topology anomalies
….

Backward-Looking

Service variability
Crowding fluctuation
Variables behaviour
Data gauges instability
…..

Forward-Looking

Sources Layer

Understanding
(“**prescriptively**”)
Services Changes

Ingestion Layer

Data-Lake

Machine-Learning

Algorithms

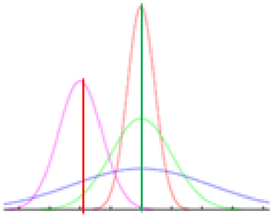
Density-Based, Connectivity, State Space,..

Methods

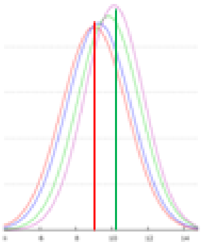
MBBEFD, SVM, SOM- Kohonen, LinkedMatrix, Cstab,, …

Prescriptive Analytics – Engine

Sources Layer



Ingestion Layer



Data-Lake

Service Layer

Prescriptive on-the-fly analytics:

Simulation by vectors metamodels
Aggregation status by dataframes
Verify variables behaviour
Verify services gauges deviations

Bulk analytics:

Compare function by prescriptive directions
Start conditional statistics
Verify deviations on mass-crowding
Trace data aggregation instability

Conclusion – Future Works

Prescriptive Metamodel Framework introduces vectors data modelling as extended construct for dataframes metamodels in Big Data systems. Analytics running on vectors metadata enables on-the-fly service gauge changes and machine-learning analytics by a new formalism. Prescriptive analytics runs both on the forward-looking and backward-looking.

Future Works:

- Consolidate the framework as Prescriptive Analytics Solution
- Extend Vector Modeling by general construct of Data Models for structured, unstructured and semi-structured information
- Extend vectors mathematical method as practice for Big Data analytics

www.eng.it

@EngineeringSpa

Engineering Ingegneria Informatica Spa

gruppo.engineering

ENGINEERING

Documentation and Traceability

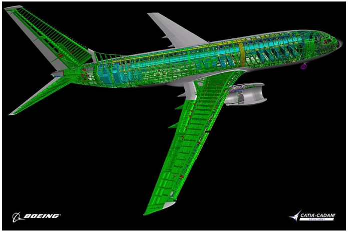
Data Analytics and Computing Challenges

Torsten Ullrich

Fraunhofer Austria Research GmbH, Visual Computing &
Technische Universität Graz, Austria

Panel on ALLDATA & MMEDIA & KESA

Documentation and Traceability



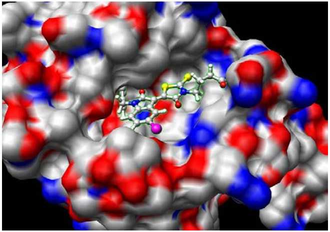


Image Sources: Prana Fistianduta, CC3.0, Wikimedia Commons; Huntster, NASA, Wikimedia Commons; Emgonzalez, Public Domain, Wikimedia Commons; Chaos, CC3.0 / GNU Free Documentation License, Wikimedia Commons

Documentation and Traceability

- Open Access, Open Data & Open Science

- Open Problems: future reproducibility

  1. physical layer / hardware layer
  2. hardware abstraction layer
  3. operating system call interface
  4. system libraries & software frameworks
  5. application layer & system environment

Documentation and Traceability

- Open Access, Open Data & Open Science
- Open Problems: future reproducibility

1. physical layer / hardware layer
2. hardware abstraction layer
3. operating system call interface
4. system libraries & software frameworks
5. application layer & system environment

Institut für ComputerGraphik und WissensVisualisierung

Documentation and Traceability

- Open Access, Open Data & Open Science
- Open Problems: future reproducibility

1. physical layer / hardware layer
2. hardware abstraction layer
3. operating system call interface
4. system libraries & software frameworks
5. application layer & system environment

Institut für **C**omputer**G**raphik und Wissens**V**isualisierung

Documentation and Traceability

- Open Access, Open Data & Open Science
- Open Problems: future reproducibility
  1. physical layer / hardware layer
  2. hardware abstraction layer
  3. operating system call interface
  4. system libraries & software frameworks
  5. application layer & system environment

Institut für ComputerGraphik und WissensVisualisierung

Documentation and Traceability

- Open Access, Open Data & Open Science

- Open Problems: future reproducibility

  1. physical layer / hardware layer
  2. hardware abstraction layer
  3. operating system call interface
  4. system libraries & software frameworks
  5. application layer & system environment

Documentation and Traceability

- Open Access, Open Data & Open Science
- Open Problems: future reproducibility

1. physical layer / hardware layer
2. hardware abstraction layer
3. operating system call interface
4. system libraries & software frameworks
5. application layer & system environment

Documentation and Traceability

- Open Access, Open Data & Open Science
- Open Problems: future reproducibility
  1. physical layer / hardware layer
  2. hardware abstraction layer
  3. operating system call interface
  4. system libraries & software frameworks
  5. application layer & system environment

Institut für ComputerGraphik und WissensVisualisierung