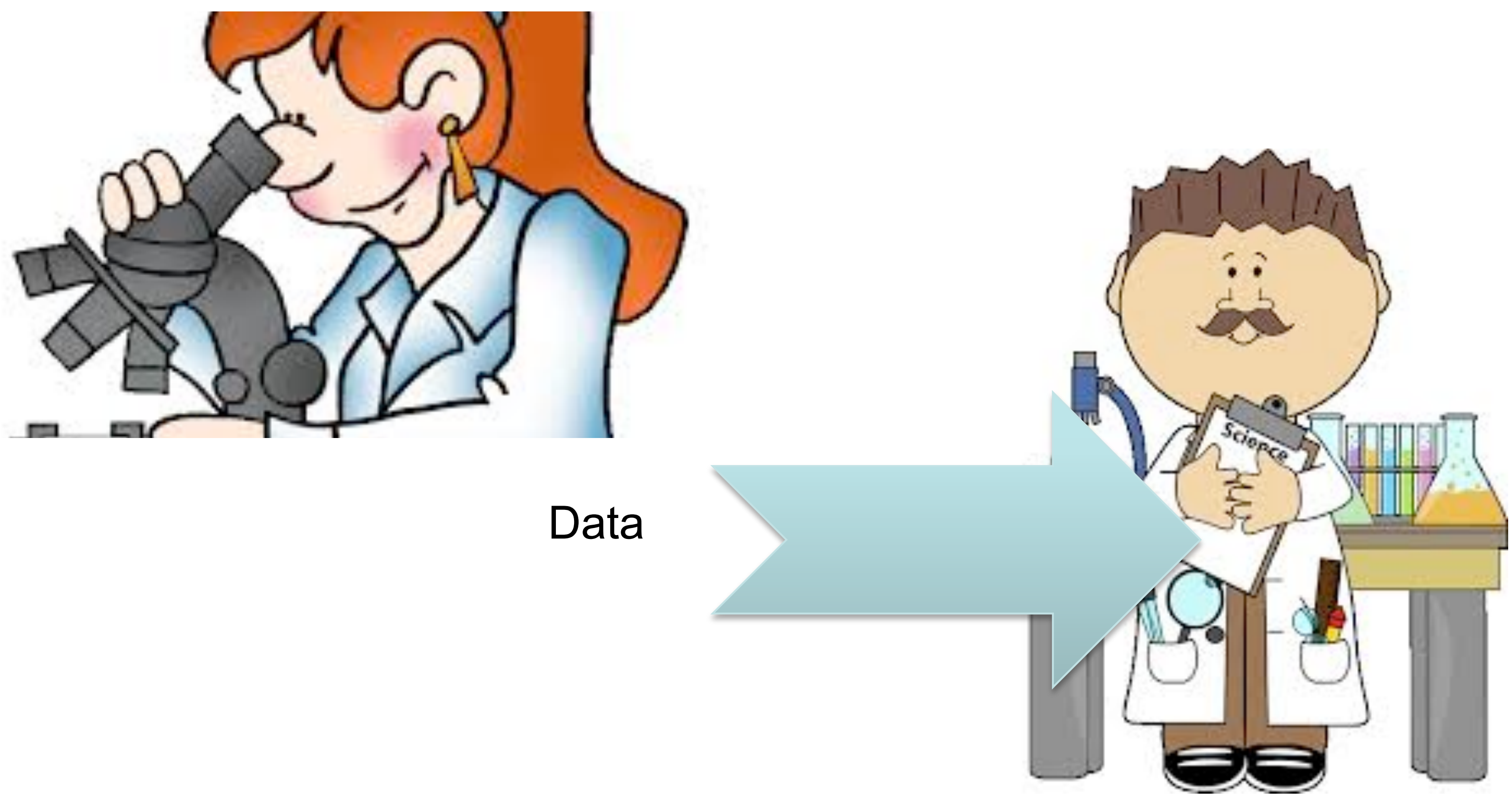
From Big Data to Quality Data: What is the emerging sensor and network technology going to deliver next?

Leon Reznik, PhD
Professor of Computer Science and Computing Security
Rochester Institute of Technology
102 Lomb Memorial Drive Rochester NY 14623 USA
Ph.: 585 475 7210 Fax: 585 475 7100 email: lr@cs.rit.edu
http: //www.cs.rit.edu/~lr

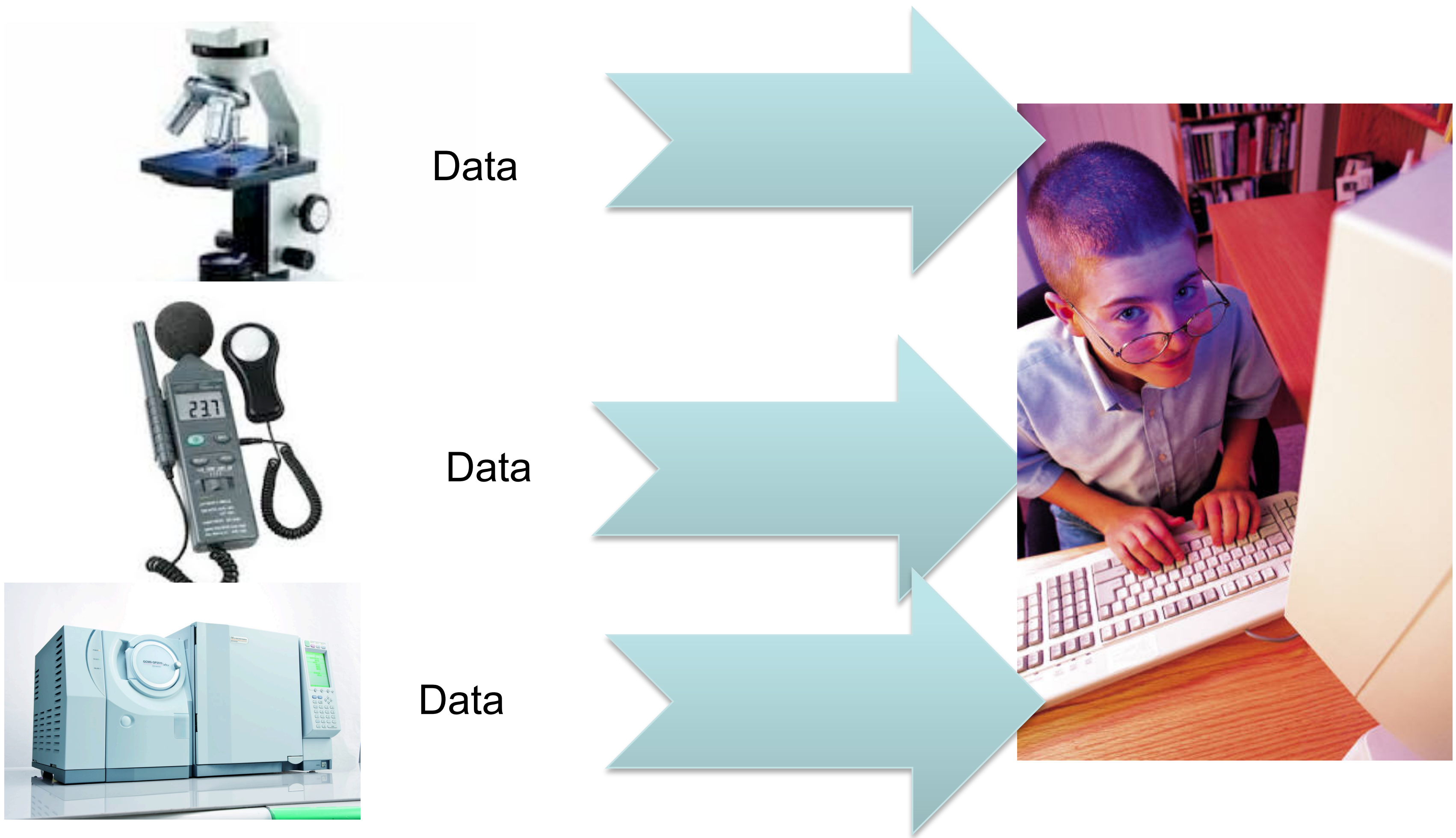
In collaboration with:
Elisa Bertino (Purdue U),
Azer Bestavros (Boston U)
Justin Cappos (New York U),
Albert Rafetseder (U of Vienna),
Yanyan Zhuang (UBC&NYU)

Computer Science @ RIT
B.Thomas Golisano College of Computing & Information Sciences

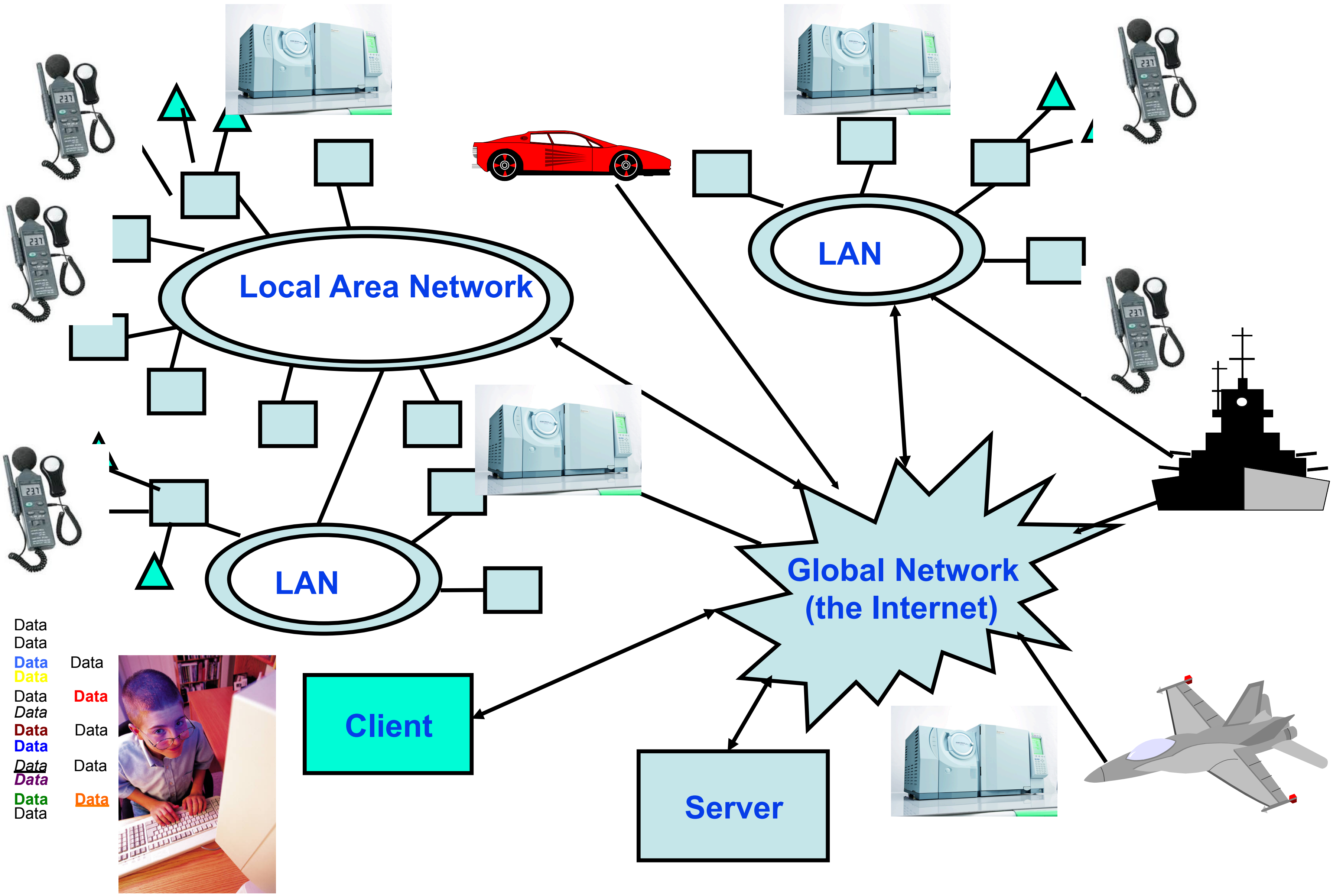
Data: old pictures



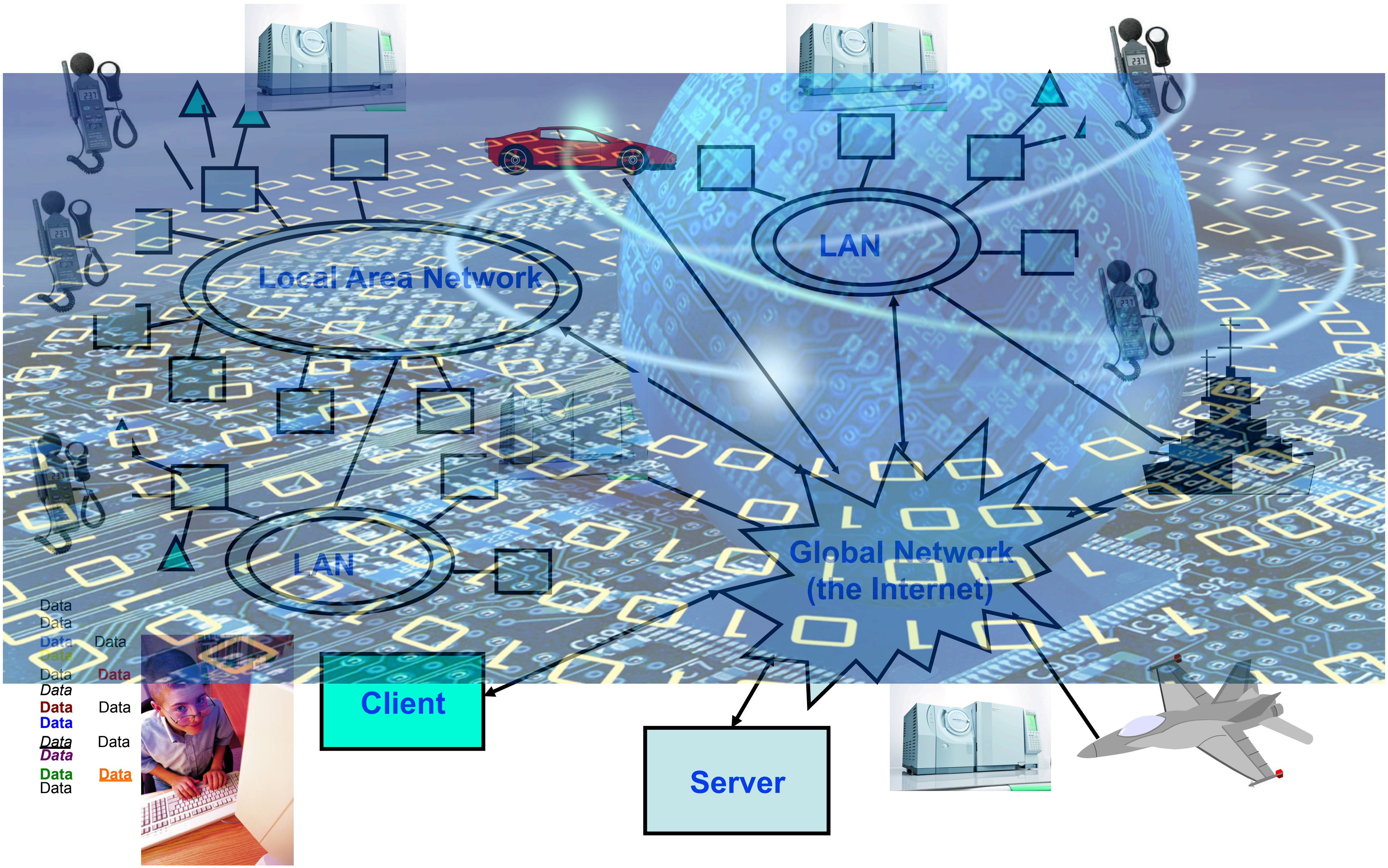
Data: recent pictures



Data: Current picture



What is Big Data?



Big Data: definitions

Big Data as the three Vs:
Volume, Velocity, and Variety – most well-known definition, first coined by Doug Laney of Gartner (source: D. Laney 3D Data Management: Controlling Data Volume, Velocity, and Variety in Application Delivery Strategies, Meta Group, 6 February 2001)

Photo source: Forbes, available at http://www.forbes.com/sites/gartnergroup/2012/05/22/infonomics-the-practice-of-information-economics/

1. Big Data: Volume

- **According to IBM, 2.5 exabytes - that's 2.5 billion gigabytes (GB) - of data was generated every day in 2012. [source: Matthew Wall Big Data: Are you ready for blast-off? BBC News March 3, 2014 at http://www.bbc.com/news/business-26383058]**

According to IDC, in 2011 we created 1.8 zettabytes **(or 1.8 trillion GBs) of information, which is enough data to fill 57.5 billion 32GB** Apple iPads. **That's enough iPads to build a Great iPad Wall of China twice as tall as the original. In 2012 it reached 2.8 zettabytes and IDC now forecasts that we will generate 40 zettabytes (ZB) by 2020.**

- **to put the data explosion in context, consider this. Every minute of every day we create:**
- **- More than 204 million email messages**
  - Over 2 million Google search queries
  - 48 hours of new YouTube videos
  - 684,000 bits of content shared on Facebook
  - More than 100,000 tweets
  - $272,000 spent on e-commerce

[source: Webopedia, How much data is out there? Updated March 3, 2014 at ttp://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html]

Volume:
terrabytes to exabytes of data to process

Data
Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data

Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data
Data Data Data

Moore's law: Doubling amount each year

2. Big Data: Variety

Data today comes in all types of formats - from a standard productions systems or transaction databases to OLAP (Online Analytical Processing) cubes. There are emails, stock or financial data and huge percentage of non-numerical data. Currently, there is a lot of new data formats and what is worse, a lot of data is even in the unstructured forms (images, audio, tweets, text messages, server logs, and so on)

- Relational data (tables, transaction, legacy data)
- Text data (web)
- Semi-structured data (XML)
- Graph data
- Streaming data

Variety:
data coming in different formats

Data
Data
Data Data
Data
Data **Data**
Data
Data Data
Data
Data Data
Data
Data **Data**
Data

2. Big Data: Variety

- Currently most data – from social media applications
- **But** sensor (machine) generated data is a much bigger story.
- **Fact**: "Boeing jet engines can produce 10 terabytes of operational information for every 30 minutes they turn. A four engine jumbo jet can create 640 terabytes of data on just one Atlantic crossing, multiply that by the more than 25,000 flights flown each day"[*Source* H.Kotadia Big Data: The Coming Sensor Data Driven Productivity Revolution published July 21, 2012 at http://hkotadia.com/archives/5000]
- **Opinion**: Within the next three to five years, I expect to see sensor data hit the crossover point, with unstructured data generated by social media. From there, the former will dominate by factors; not just by 10-20 percent, but by 10-20 times that of social media. Source:[K.Kwang Sensor data is data analytics' future goldmine, ZDNet, JUne 11, 2010, http://www.zdnet.com/sensor-data-is-data-analytics-future-goldmine-2062200657/]

3. Big Data: Velocity

- Sensor and network technological developments produce the data at an unprecedented speed
- Need to process these data in real time

Velocity:
streaming data msec to sec to respond

Data
Data
Data Data
Data
Data **Data**
Data

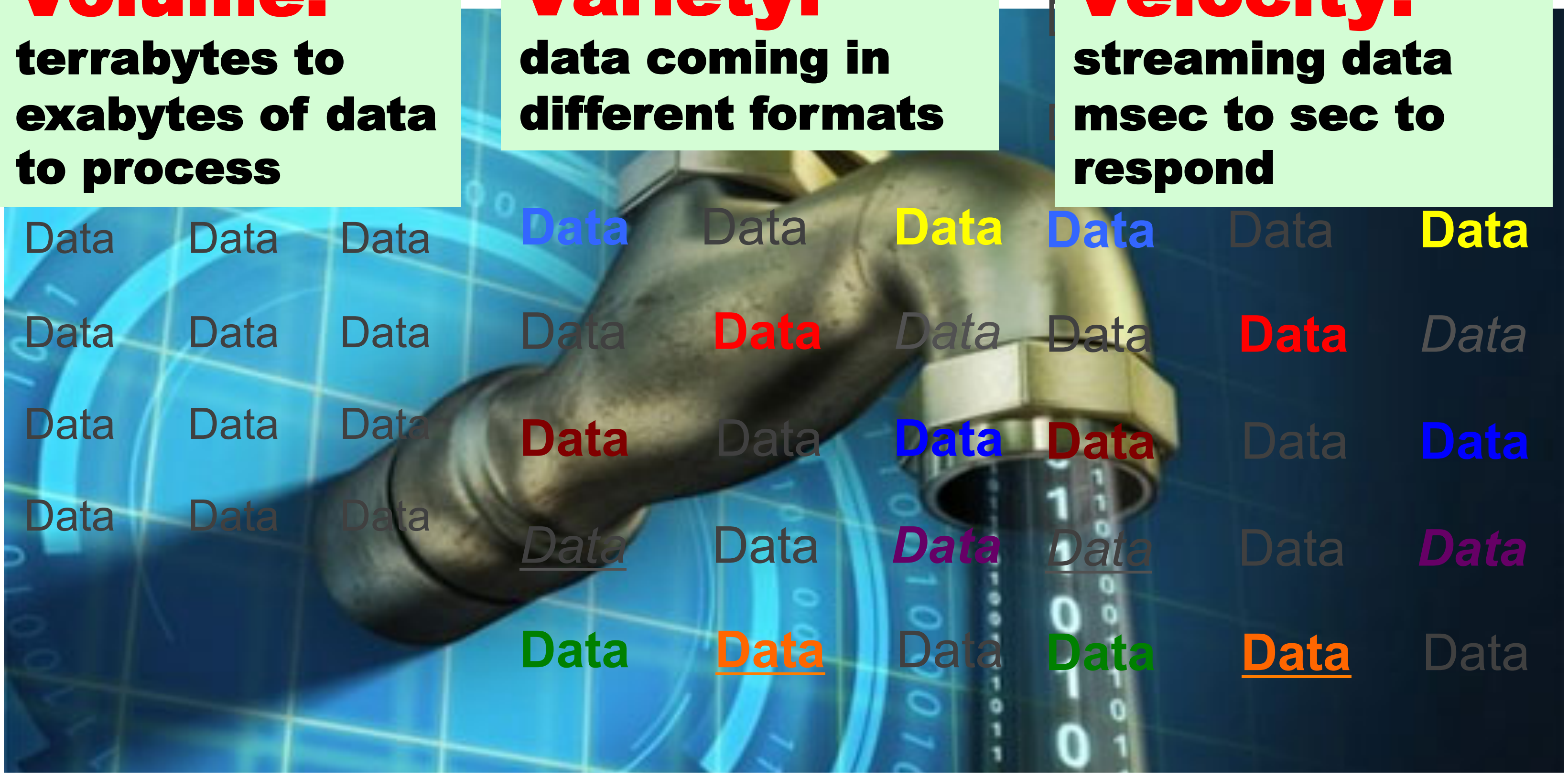
Data Data
Data
Data Data
Data
Data **Data**
Data

Big Data definition: V^3

Volume: terrabytes to exabytes of data to process

Variety: data coming in different formats

Velocity: streaming data msec to sec to respond



most well-known definition, first coined by Doug Laney of Gartner (source: D. Laney 3D Data Management: Controlling Data Volume, Velocity, and Variety in Application Delivery Strategies, Meta Group, 6 February 2001)

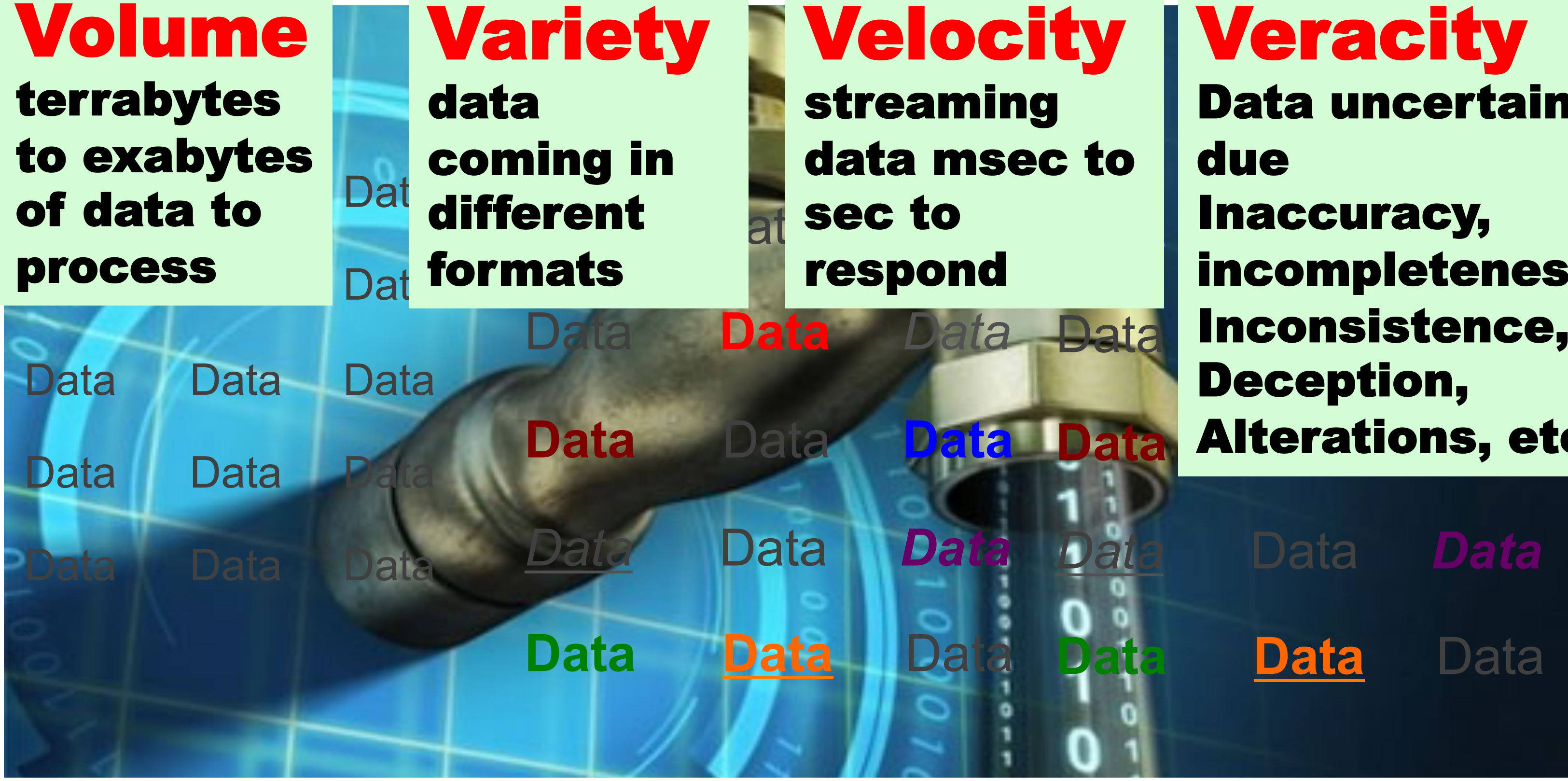
Big Data: from V^3 to V^4

Volume
terrabytes to exabytes of data to process

Variety
data coming in different formats

Velocity
streaming data msec to sec to respond

Veracity
Data uncertainty due Inaccuracy, incompleteness, Inconsistence, Deception, Alterations, etc.



Big Data: other definitions

from Big Data For Dummies by J. Hurwitz,F. Halper, M. Kaufman John Wiley & Sons, March 2013

Big Data is the capability to manage the huge volume of disparate data at the right speed, and at the right time frame to allow real-time analysis and reaction

to M. Adrian, "It's going mainstream, and it's your next opportunity," *Teradata Magazine,* pp. 38-43, 2011

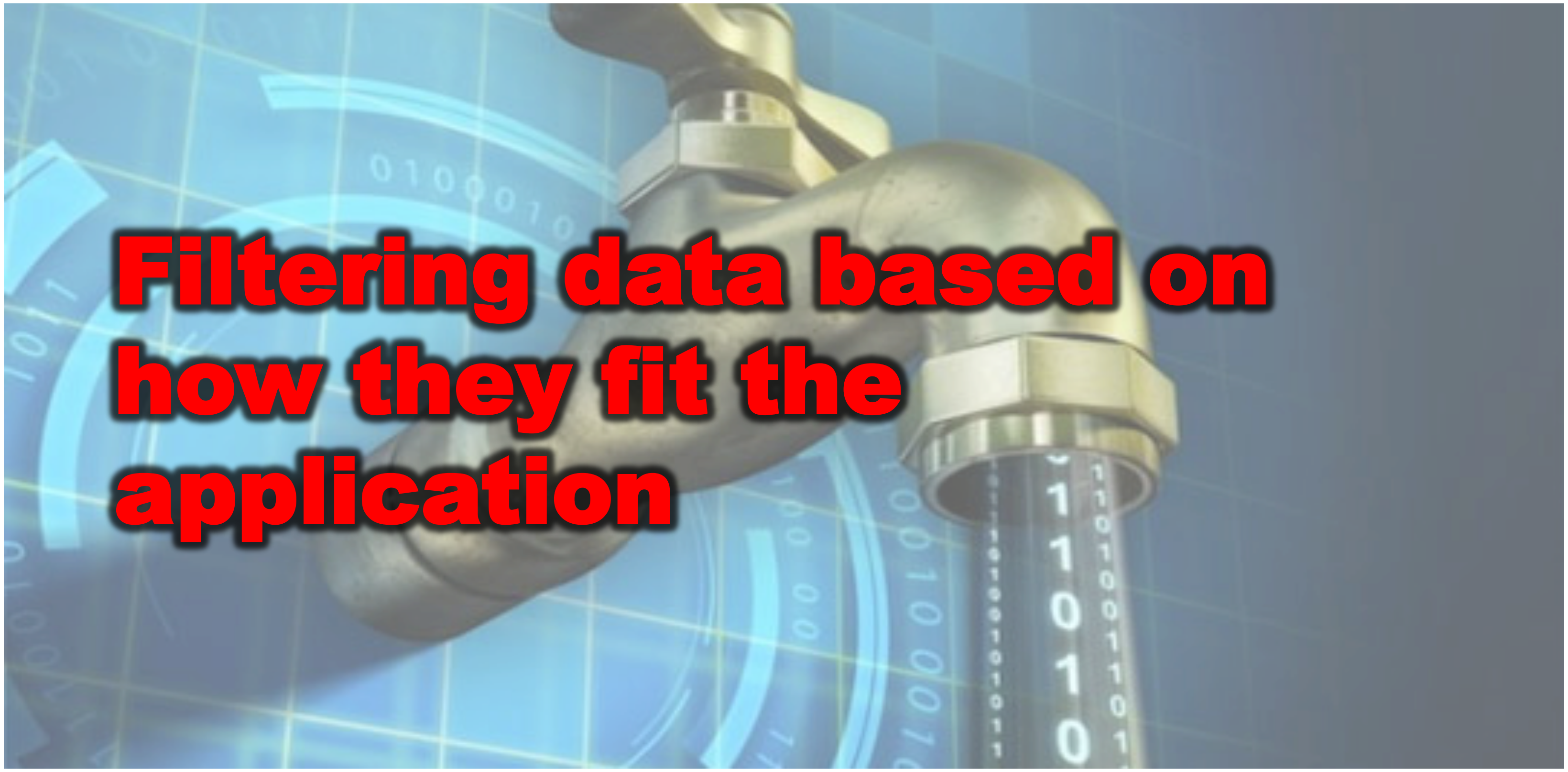
Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage and process it within a tolerable elapsed time for its user population

and Wikipedia

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Where is the Big Data solution?

Where is the Big Data solution?



What is Data Quality (DQ)

- How to optimize decision making and planning?
- Metrics that engender trust in the processes that use/consume the data
  - How accurate are the data? (this is about measurement errors and noise)
  - How complete are the data ? (this is about missing data due to equipment failures)
  - How timely are the data? (this is about time delay)
  - Are the data valid? (this is about time expiration or safety)
  - Are the data authentic? (this is about origin and provenance)
  - How big is the chance that the data have been maliciously altered? (This is about security)
  - ...

This slide was originally produced by A.Bestavros(Boston U), modified by L.Reznik

What is Data Quality?

Data are of high quality "if they are fit for their intended uses in operations, decision making and planning" (J. M. Juran)....source: en.wikipedia.org/wiki/Data_quality

How to integrate all of the metrics into an indicator that demonstrates the data suitability for a specific task and facilitates decision making?

What is the role of technology?

Today, there are more things connected to the Internet than people on the planet.

Applications of nano-technological devices significantly increase an amount of data of rather poor quality.

There are more things already connected to the Internet than people on the planet. CISCO IBSG group estimates the number of connected devices at 25B in 2015 and 50B by 2020 [source: D. Evans, "Internet of Things. How the next evolution of Internet is changing everything," ed: CISCO IBSG group, 2011]

Current developments fusing multiple data sources with various quality data and creating big data collections as well as studies in novel areas such as nano-engineering and technology have substantially advanced the requirements on DQ.

What is the NEXT role of technology?

From current

EXTENSIVE development : FASTER collecting, communicating and processing MORE and MORE DIVERSE data (V^3)

To future

INTENSIVE developments: effective and efficient data collection and presentation where and when they are needed.

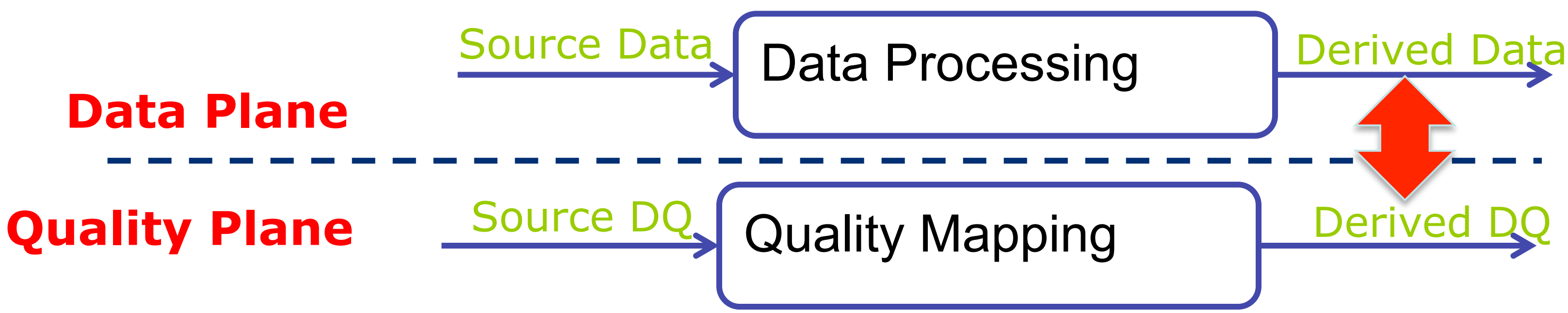
The Internet of Things (IoT) has the potential to transform how and when decisions are made throughout business and our daily lives, but only if that data can be processed and analyzed effectively, and more importantly if data are of high quality.

Data Quality Management

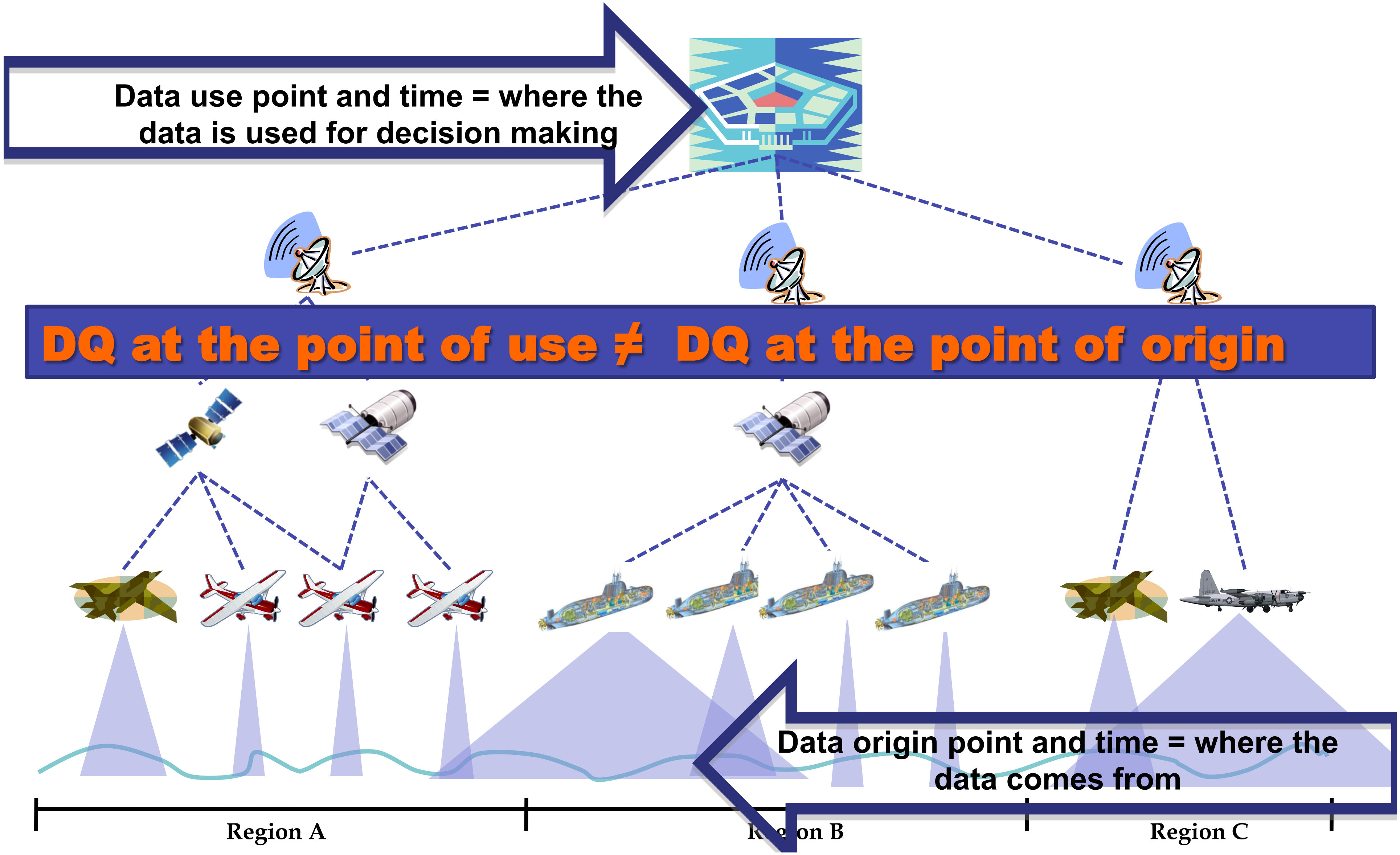
DQ Management

- When new data are received (e.g. by measurement, communicating or computing over existing data) it is necessary to also derive DQ metrics for the new data.
- To maintain proper DQ, it is necessary to have a disciplined approach to the inference and processing of DQ metrics.

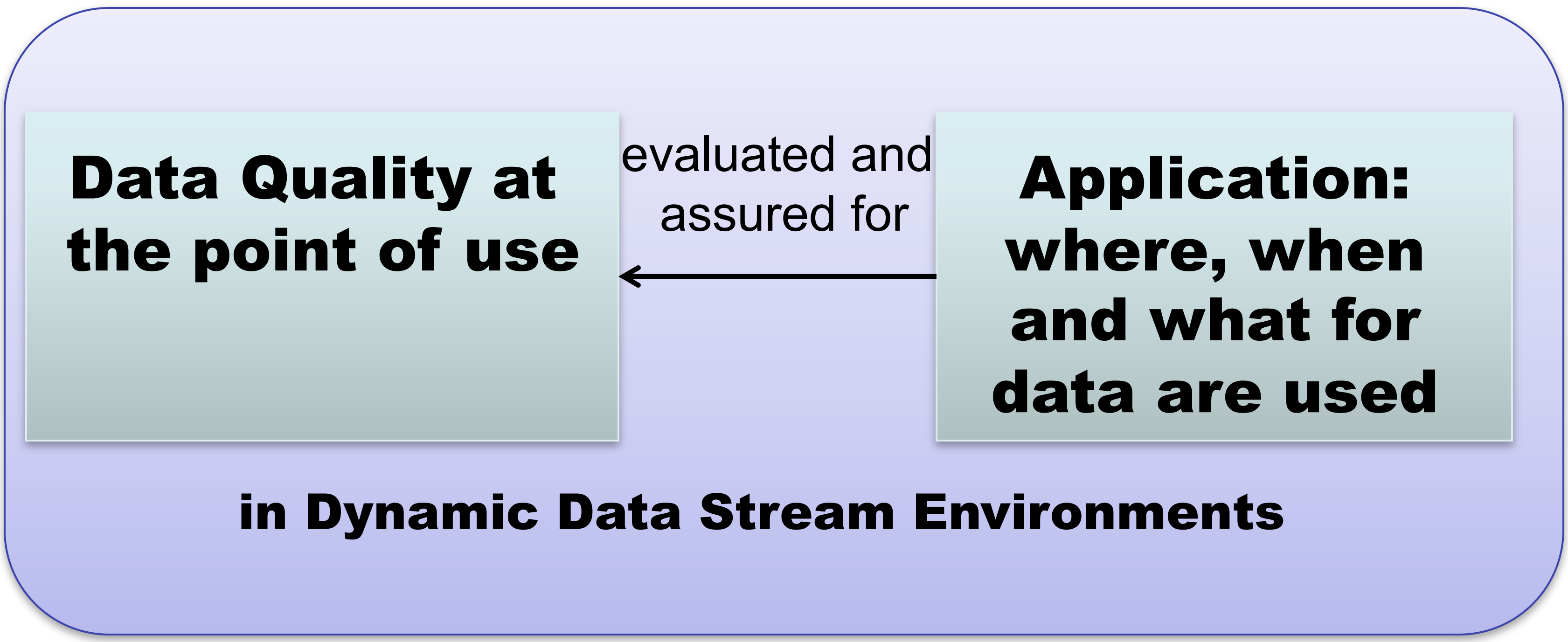
Data Quality Management



An Example Application: Battlefield Monitoring Sensor Network



Focus of Our Work



Which problems do we address?

1. DQ evaluation: Given the data collection scheme and the data streams flows, the specified DQ composition and calculus of various metrics, **evaluate DQ at the decision making point**

2. DQ verification: under the same conditions, **verify if the DQ reaches a certain level**

3. DQ assurance: given the same conditions and resource availability, **determine which conditions and resources need to be modified to reach a certain DQ level**

How do we target the application?

1. Task (application) determines the data collection scheme and data stream flows → Use the data provenance schemes for an integral DQ composition and calculus in DQ evaluation, verification and assurance

2. Task (application) determines the DQ metrics composition and calculus → Target DQ metrics composition at the mission decision making

3. Task (application) determines the resource availability → Resource availability is a key factor in DQ assurance

How do we do it?

Our Solution:

A Cyclic Distributed Hierarchical Framework for Data Quality Evaluation and Assurance

How do we do it?

Part 1: **DQ metrics content or which metrics to choose to evaluate DQ?**

Part 2: **DQ integral composition and calculus or how to integrate individual DQ metrics for decision making?**

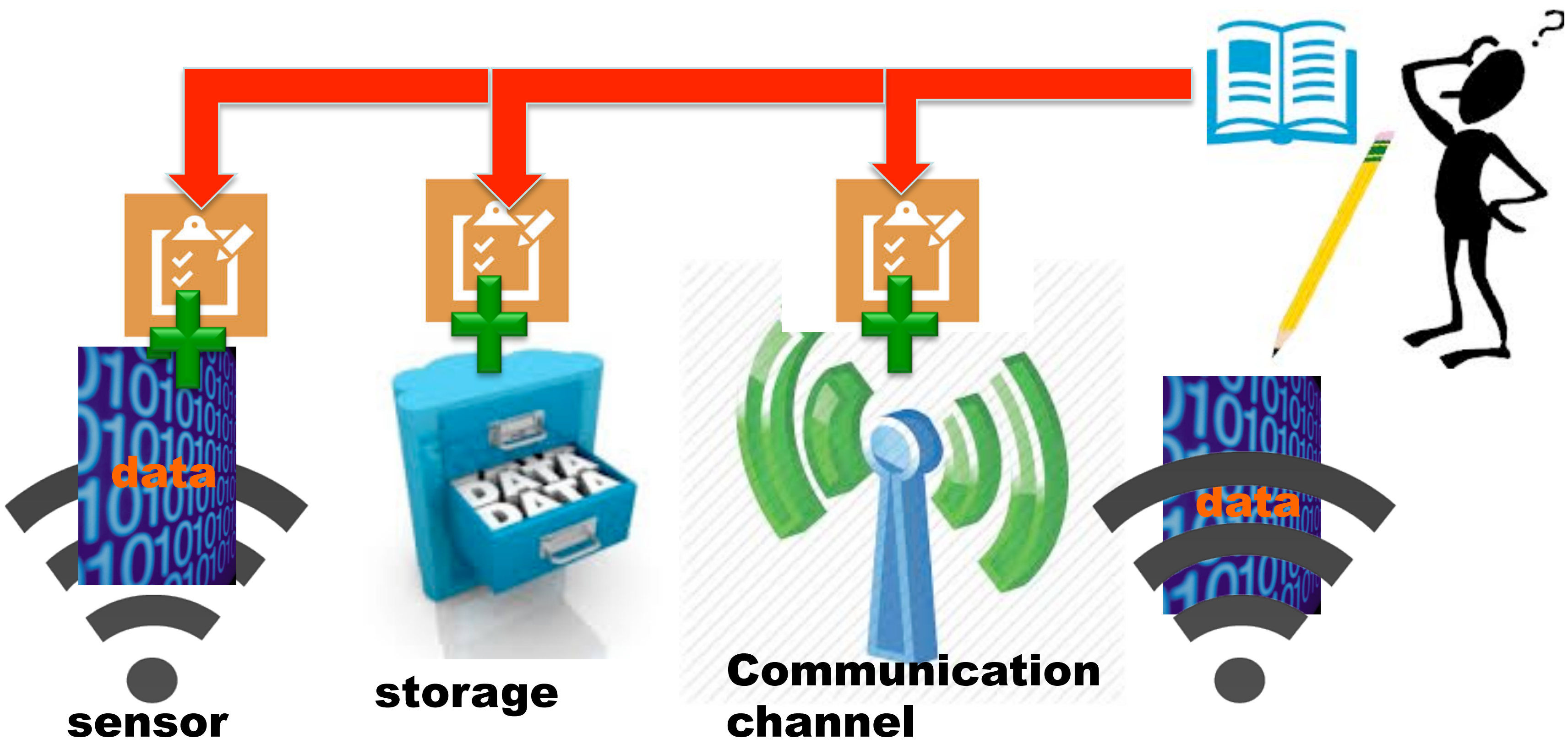
Part 3: **Dynamic DQ evaluation and assurance based on data provenance or how to take into account various data streams and their merging in DQ evaluation? *(note the difference with metrics composition only in part 2)***

Part 4: **System survivability and assurance with DQ or how to assure that the given DQ level is reached under the specified condition or which conditions are needed to reach this level?**

How do we do it?

And now a bit more details about our methods

1. DQ metrics definition and assignment



Part 1: *DQ metrics choice*

Major research challenges:

- demonstrate metrics that are important to the range of possible tasks
- measures must yield quantifiable information (percentages, averages, numbers or order scales)
- data that supports the metrics need to be readily obtainable in military data systems
- metrics must be useful for evaluating and assuring overall DQ and tracking QE/QA system performance

Research results:

L. Reznik, "Integral Instrumentation Data Quality Evaluation: the Way to Enhance Safety, Security, and Environment Impact," presented at the 2012 IEEE International Instrumentation and Measurement Technology Conference, Graz, Austria, May 13-16, 2012, 2012.

G. P. Timms, P. A. J. de Souza, L. Reznik, and D. V. Smith, "Automated Data Quality Assessment of Marine Sensors," Sensors, vol. 11, pp. 9589-9602, 2011.

Metrics content: generic sample

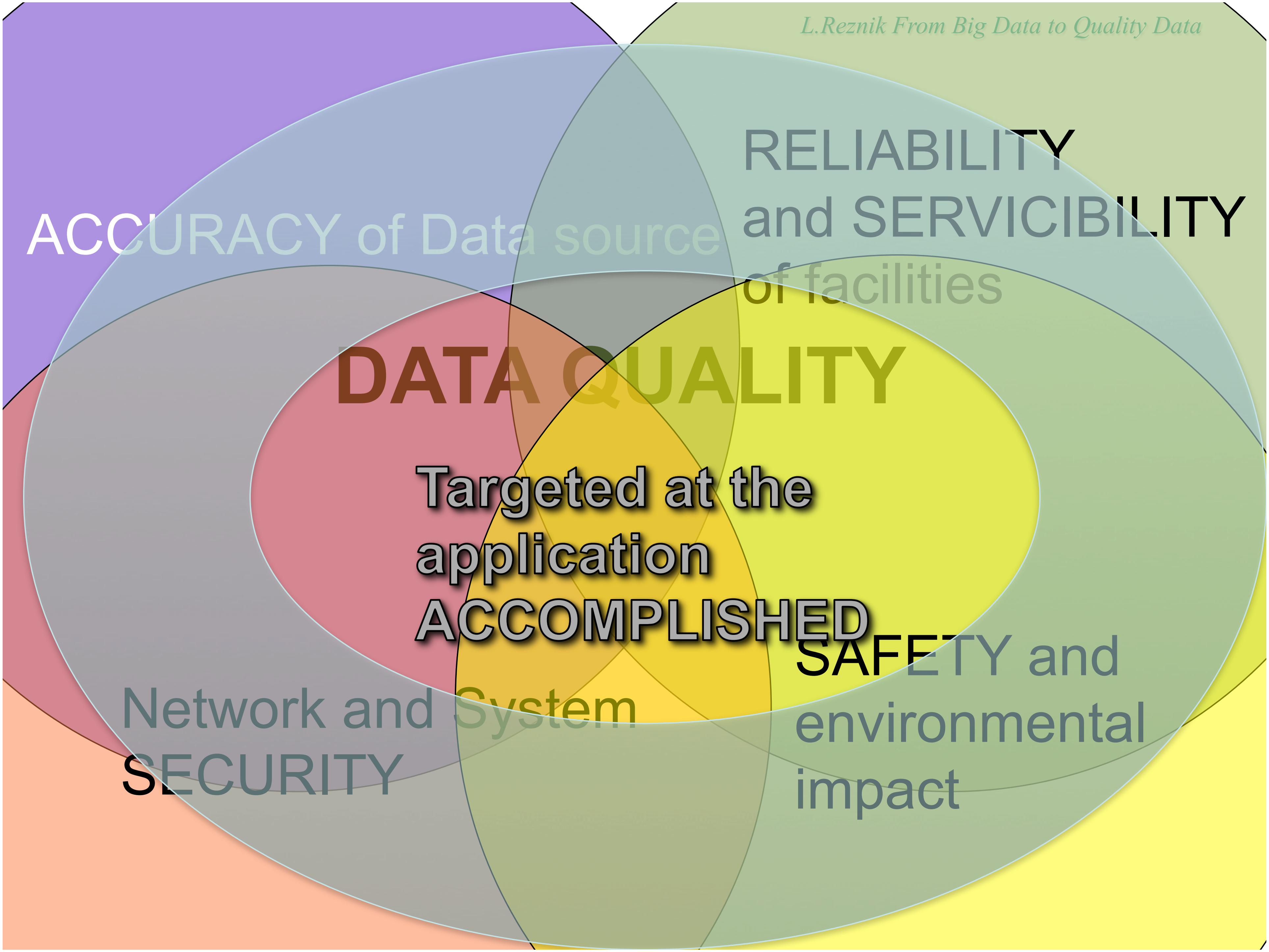
Generic Attribute Name	DQ indicator/group (Figure1)	Description
Time-since-Manufacturing	Maintenance/reliability	The measure of the age of the device
Time-since-Service	Maintenance/reliability	The measure of the days since last service was performed in accord with the servicing schedule
Time-since-Calibration	Calibration/reliability	The measure of the days since last calibration was performed in accord with the calibration schedule
Temperature Range	Application/performance	The measure of temperature range within which the device will provide optimum performance
Physical Tampering Incidences	Physical security/security	The number of reported incidents that allowed unauthorized physical contact with the device
System Breaches	Access control/security	The measure of the number of unauthorized accesses into the system, denial of service attacks, improper usage, suspicious investigations, incidences of malicious code.
System Security	Security/security	Measures presence of intrusion detection systems, firewalls, anti-viruses.
Data Integrity	Vulnerabilities/securities	Number of operating system vulnerabilities that were detected.
Environmental Influences	Environment/environment	Number of incidences reported that would subject the device to mechanical, acoustical and triboelectric effects.
Atmospheric Influences	Environment/environment	Number of incidences reported that would subject the device to magnetic, capacitive and radio frequencies.
Response Time	Signals/reliability	Time between the change of the state and time taken to record the change

Metrics content: specific sample

Device Name	Application specific Quality indicator	Description
Electric / Power Meters	Foucalt Disk	Check to verify the material of the foucalt disk.
	Friction Compensation	Difference in the measure of initial friction at the time of application of the compensation and the current friction in the device.
	Exposure to Vibrations	Measure of the number of incidences reported which would have caused the device to be subjected to external vibrations
Water Meters	Mounting Position	The measure of the number of days since regulatory check was performed to observe the mounting position of the device.
	Environmental Factors	Number of incidences reported which may have affected the mounting position of the device.
	Particle Collection	Measure of the amount of particle deposition.

Table 1. Metrics examples: specification and calculation

Group	Security	Security	Measurement Accuracy	Process safety and performance
Measure ID	**Physical security incidents**	**Cryptographic system and communication protection**	**Measurement uncertainty**	**Process safety risk**
Goal	Strategic: Ensure an environment of comprehensive security and accountability for personnel, facilities, and products Security: Integrate physical and security protection mechanisms to ensure appropriate protection of the data sources and facilities	Strategic: accelerate the development and use of a quality information infrastructure Security: allocate sufficient resources to adequately protect an information infrastructure in a application	Ensure collecting high accuracy data from the data sources	Improve process safety by reducing such risks as equipment damage or business interruption due to wrong data use
Measure	Percentage (%) of physical security incidents involving unauthorized entry to facilities	Percentage of mobile computers and other devices that perform all cryptographic operations as recommended for this application	Probability of the type A uncertainty value (random measurement error) being within the range specified by the application for this data source	Probability that equipment damage or business interruption will not increase a specified level
Calculation formula or algorithm	(number of physical security incidents involving unauthorized entry to facilities/total number of access to facilities) *100	(number of mobile computers and other devices that perform all cryptographic operations as recommended for the application/total number of mobile computers and other devices)*100	Probability above is calculated based on the standard deviation empirical estimate or other probability characteristics available	Probability above is calculated based on the business process analysis, equipment costs, expert evaluations
Target	Should be a high (or low) percentage defined by the application			
Implementation evidence	How many physical security incidents involving unauthorized entry to facilities occurred over specified period? How many total entries to facilities occurred over specified period?	How many mobile computers and devices are employed in the application? How many mobile computers and devices employ cryptography as prescribed? How many mobile computers and devices have cryptography implementation waivers?	Measurement results	Equipment costs, process risk evaluations, etc.
Frequency	Collection: defined by the organization Reporting: defined by the application		Collection: defined by the application Reporting: defined by the application	
Responsible parties	Information owner: defined by the organization (e.g. physical security officer) Information collector: defined by the organization (e.g. computer security incident response team)	Information owner: defined by the organization (e.g. information security officer) Information collector: defined by the organization (e.g. system administrators)	Information collection and measurement team	Information owner: Process safety team. Operations and maintenance team, Accounting team Information collector: defined by the organization
Information source	Physical security incident report, physical access control logs	System and network security plans	Measurement results database, sensor networks or systems	Operations, accounting and other databases
Report format	Percentage and/or pie charts		As specified by application	
Documentation	NIST SP 800-53	NIST SP 800-53	Guide to the expression of uncertainty in measurement	ANSI/ISA-99.00.01


L.Reznik From Big Data to Quality Data
RELIABILITY
and SERVICIBILITY
of facilities
ACCURACY of Data source
DATA QUALITY
Targeted at the
application
ACCOMPLISHED
SAFETY and
environmental
impact
Network and System
SECURITY

Project content part 2: *Composition and calculus for DQ evaluation and assurance*

Major research challenge:

- developing a formal description for DQ compositions and an operational calculus oriented towards applications
- Adjsting calculus procedures to a particular application

Prior research results:

L.Reznik and E.Bertino Data Quality Evaluation: Integrating Security and Accuracy, **CCS '13:** Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, Berlin, November 2013

S.E.Lyshevsky and L.Reznik Information-theoretic estimates of communication and processing in nanoscale and quantum optoelectronic systems, 2013 IEEE XXXIII International Scientific Conference on Electronics and Nanotechnology (ELNANO), 16-19 April 2013, pp. 33-37

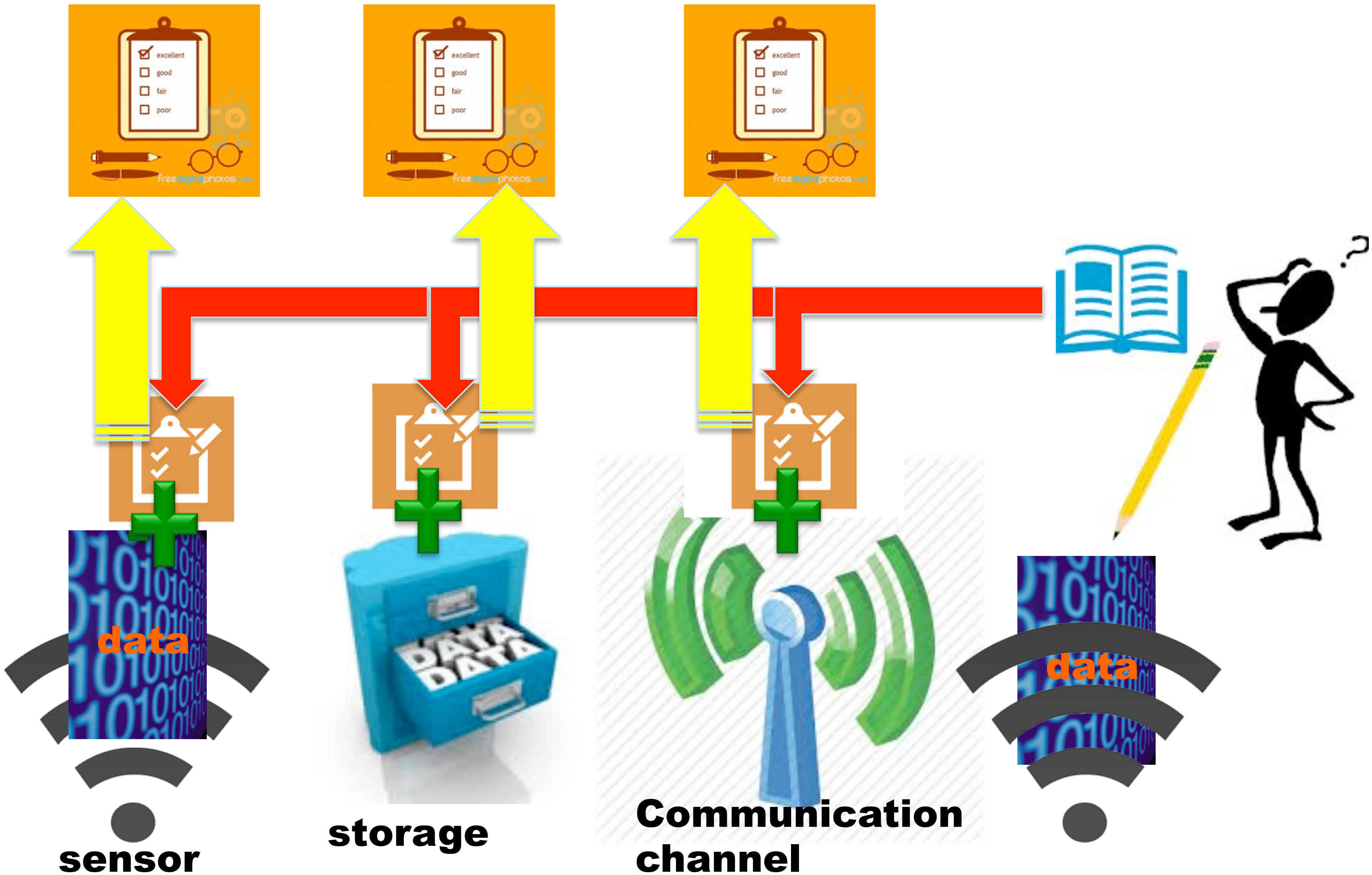
J. Podpora, L. Reznik, and G. Von Pless, "Intelligent Real-Time Adaptation for Power Efficiency in Sensor Networks," Sensors Journal, IEEE, vol. 8, pp. 2066-2073, 2008.

L. Reznik and G. Von Pless, "Neural networks for cognitive sensor networks," in Neural Networks, 2008. IJCNN 2008 pp. 1235-1241.

L. Reznik, V. Kreinovich, and S. A. Starks, "Use of fuzzy expert's information in measurement and what we can gain from its application in geophysics," in Fuzzy Systems, 2003. FUZZ '03. The 12th IEEE International Conference on, 2003, pp. 1026-1031 vol.2.

2. DQ indicators calculation and integration

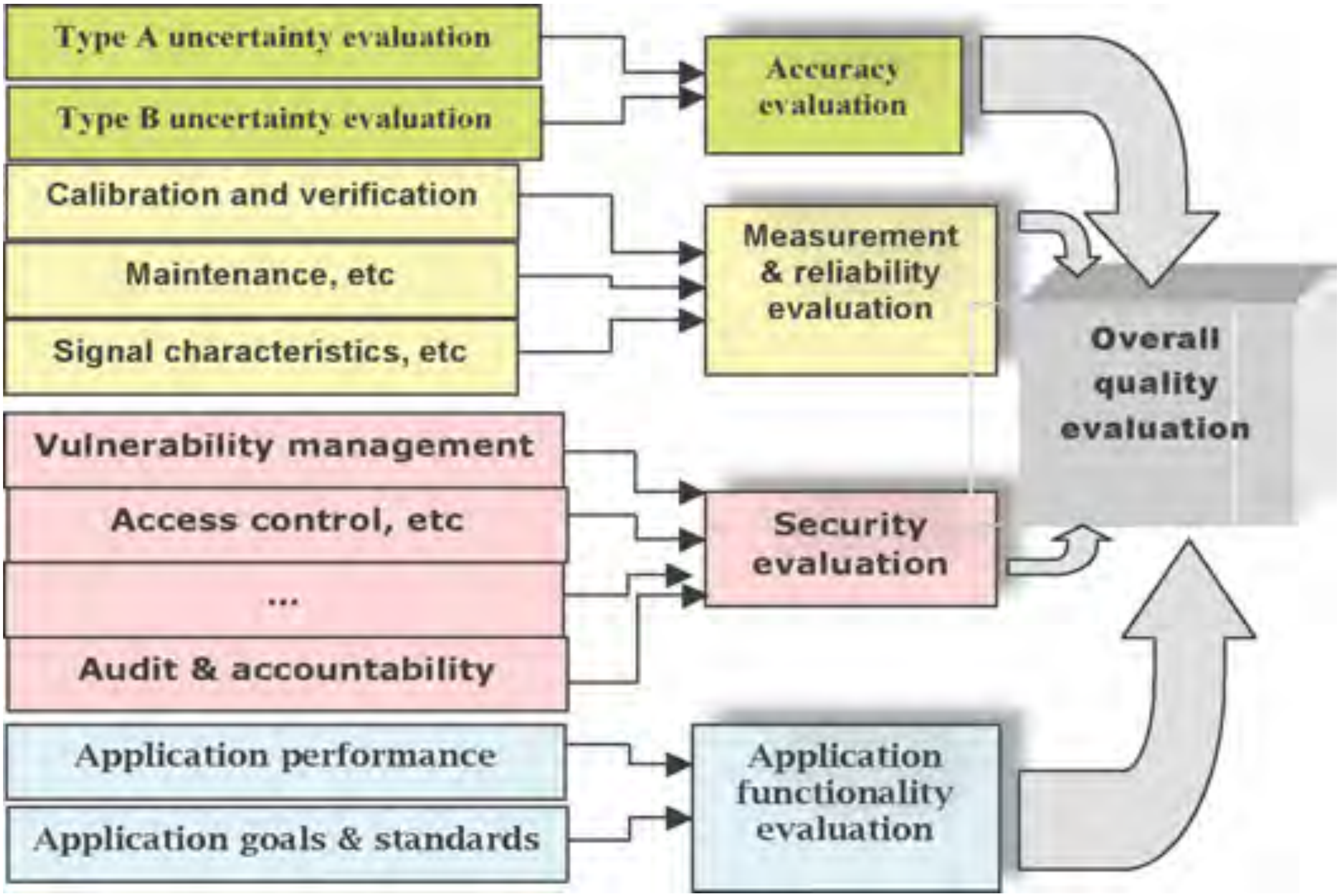
1.



DQ metrics calculation

Generic Attribute Name	Series Name	Formula to Compute Attribute Quality Score
Time-since-Manufacturing	Linear Series	QS = MaxScore – (rate * number of years after manufacturing)
	Exponential Series	QS = ((Initial Drop * MaxScore)*100) - e ^(X), where, X = rate *(number of years after manufacturing -1)
	Step Drop	QS = MaxScore – (rate * (number of years after manufacturing / number of permissible years between drops))
Calibration Date	Linear Series	QS = MaxScore – (rate * number of months passed after the applicable calibration date)
	Exponential Series	QS = ((Initial Drop * MaxScore)*100) - e ^(X), where, X = rate *(number of months passed after the applicable calibration date -1))
	Step Drop	QS = MaxScore – (rate * (number of months after calibration / number of permissible months between calibrations))
Physical Tampering	Linear Series	QS = MaxScore – (rate * number of incidents of physical tampering reported)
	Exponential Series	QS = ((Initial Drop * MaxScore)*100) - e ^(X) , where, X = rate *(number of incidents of physical tampering reported - 1))
	Step Drop	QS = MaxScore – (rate * (number of incidents of physical tampering reported / number of permissible incidents of physical tampering between drops))

QE/QA



Type A uncertainty evaluation
Type B uncertainty evaluation
Accuracy evaluation
Calibration and verification
Maintenance, etc
Signal characteristics, etc
Measurement & reliability evaluation
Overall quality evaluation
Vulnerability management
Access control, etc
...
Audit & accountability
Security evaluation
Application performance
Application goals & standards
Application functionality evaluation

DQ metrics integration: Component model composition

$$[A]\bullet[B]\Rightarrow[A\otimes B]$$

$$[A\otimes B]\bullet[C\otimes D]\Rightarrow[A\otimes B\otimes C\otimes D]$$

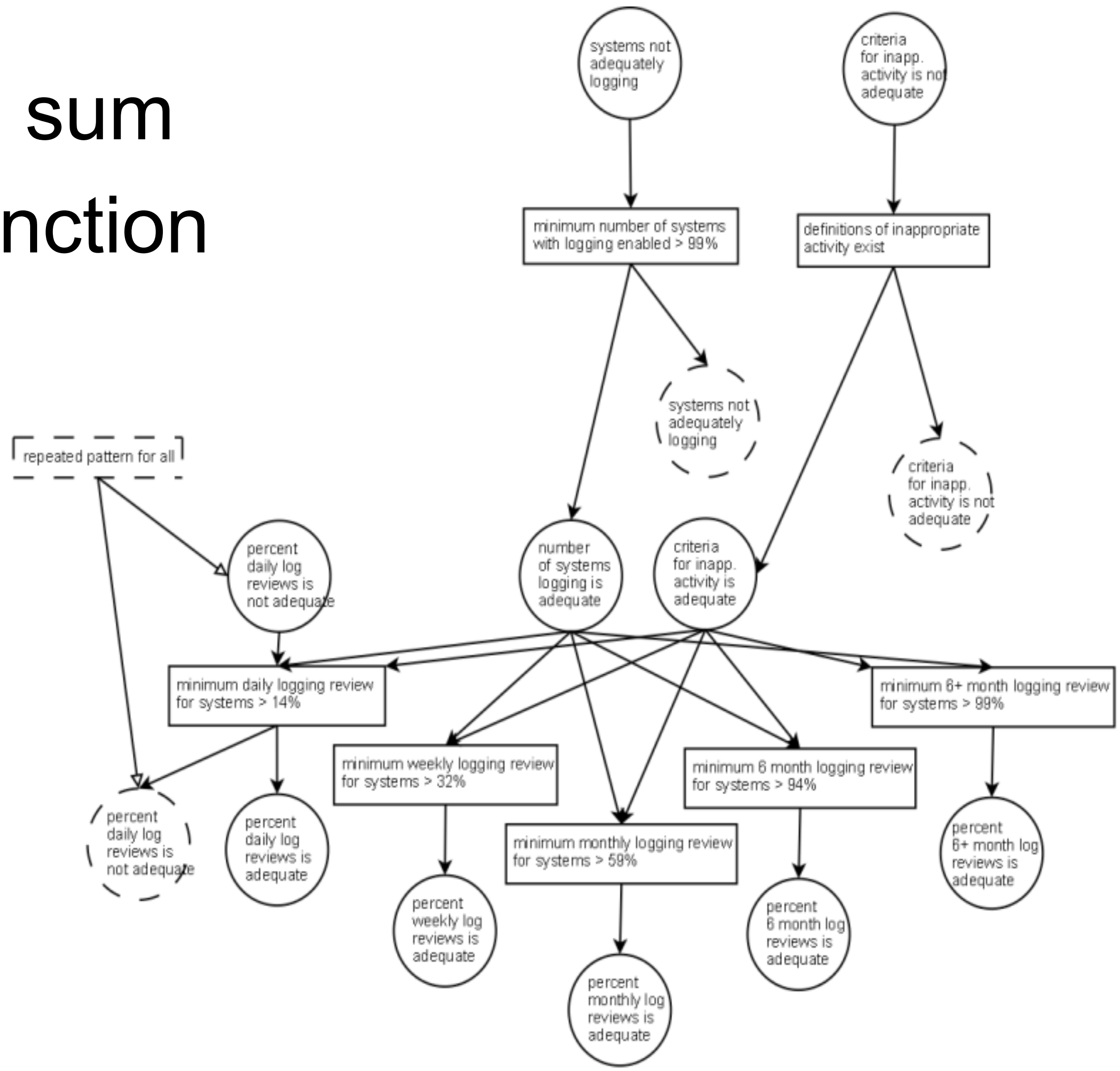
$$[A\otimes B\otimes C\otimes D]\bullet[E\otimes F\otimes G\otimes H]\Rightarrow[A\otimes B\otimes C\otimes D\otimes E\otimes F\otimes G\otimes H]$$

... How about a network-scale application ??? …

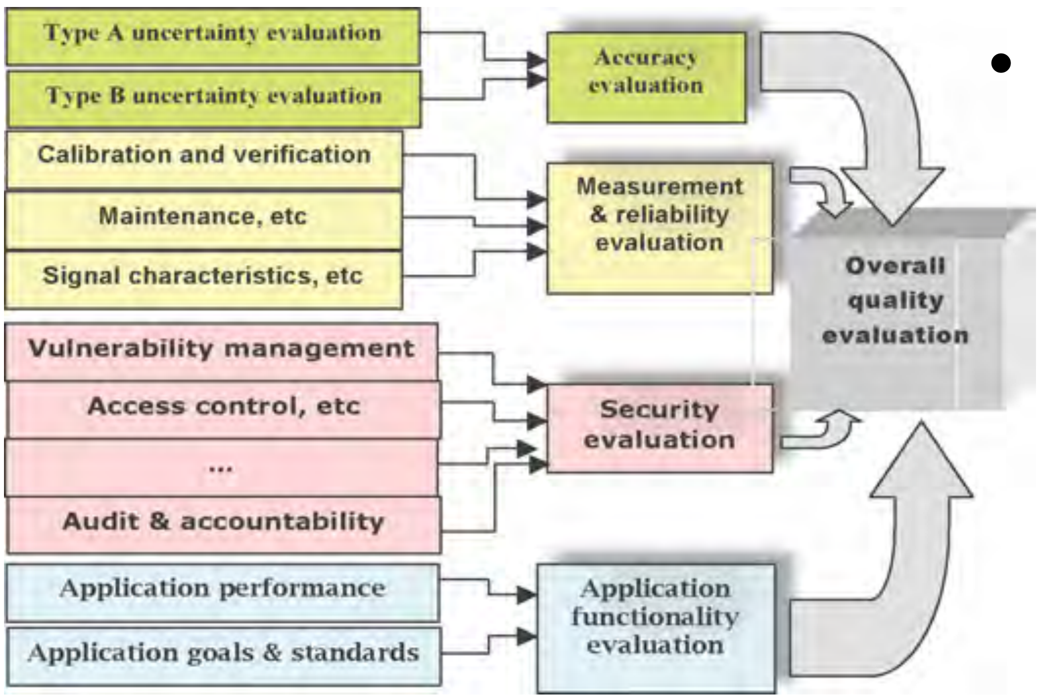
This slide was originally produced by A.Bestavros(Boston U), modified by L.Reznik

DQ metrics integration calculus

- 1. weighted sum
- 2. logical function

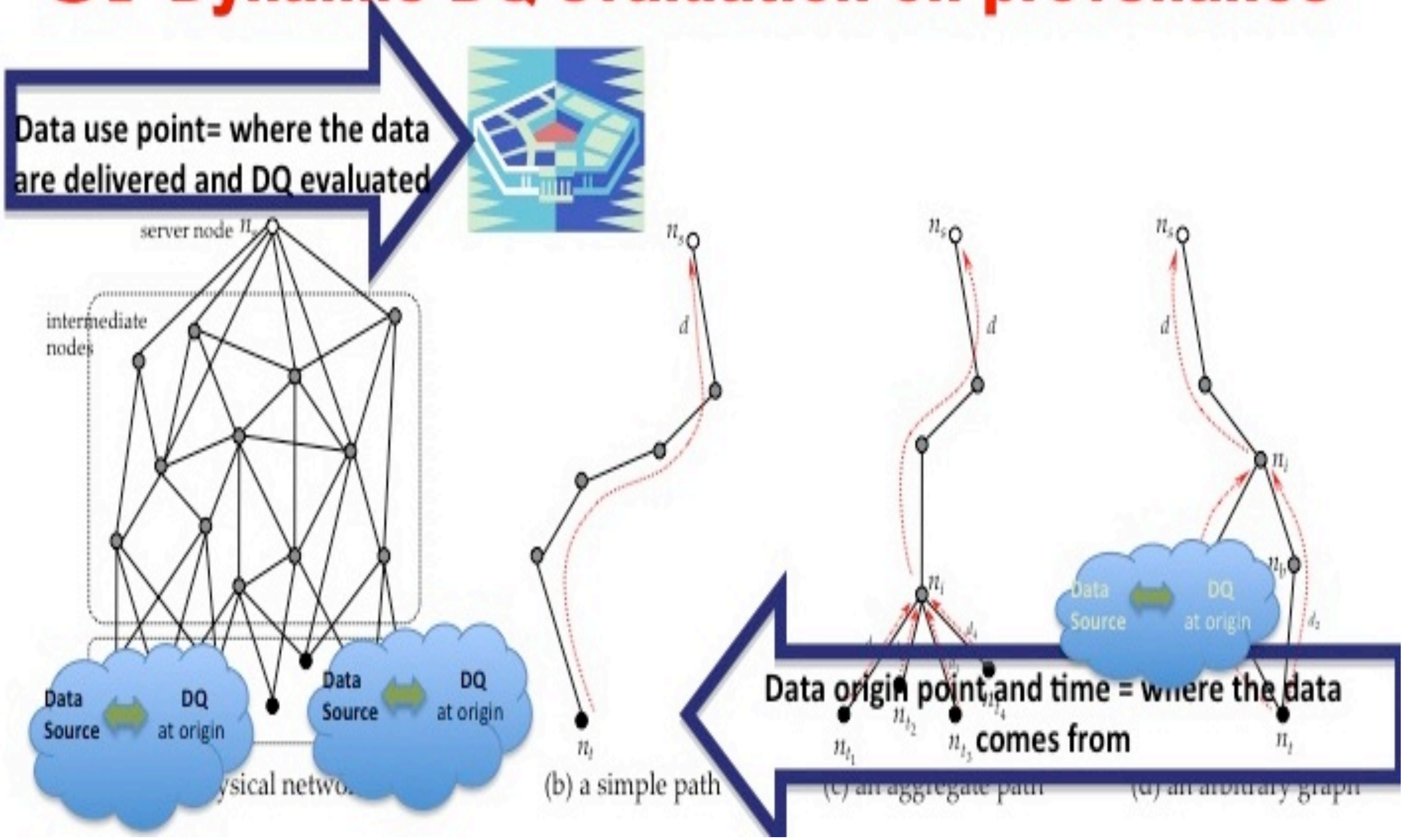


DQ metrics integration



- Integrating various metrics of data accuracy, security and safety
  - Produce overall evaluation
  - Develop recommendations for improvement
  - Integrate for efficient and effective mission accomplishment

3. Dynamic DQ evaluation on provenance



Part 3:
Dynamic DQ evaluation and assurance based on data provenance techniques

What is Provenance?

- In general, the origin, or history of something is known as its **provenance**.
- In the context of computer science,

  data provenance refers to information documenting how data came to be in its current state - where it originated, how it was generated, and the manipulations it underwent since its creation.

Project content part 3: Dynamic DQ evaluation and assurance based on data provenance techniques

Major research challenge:

- reflecting the system mobility and dynamic nature in the DQ evaluation by developing data structures and algorithms, which dynamically modify the DQ evaluation.

Prior research results:

•**L.Reznik and E.Bertino** Data Quality Evaluation: Integrating Security and Accuracy, **CCS '13:** Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, Berlin, November 2013

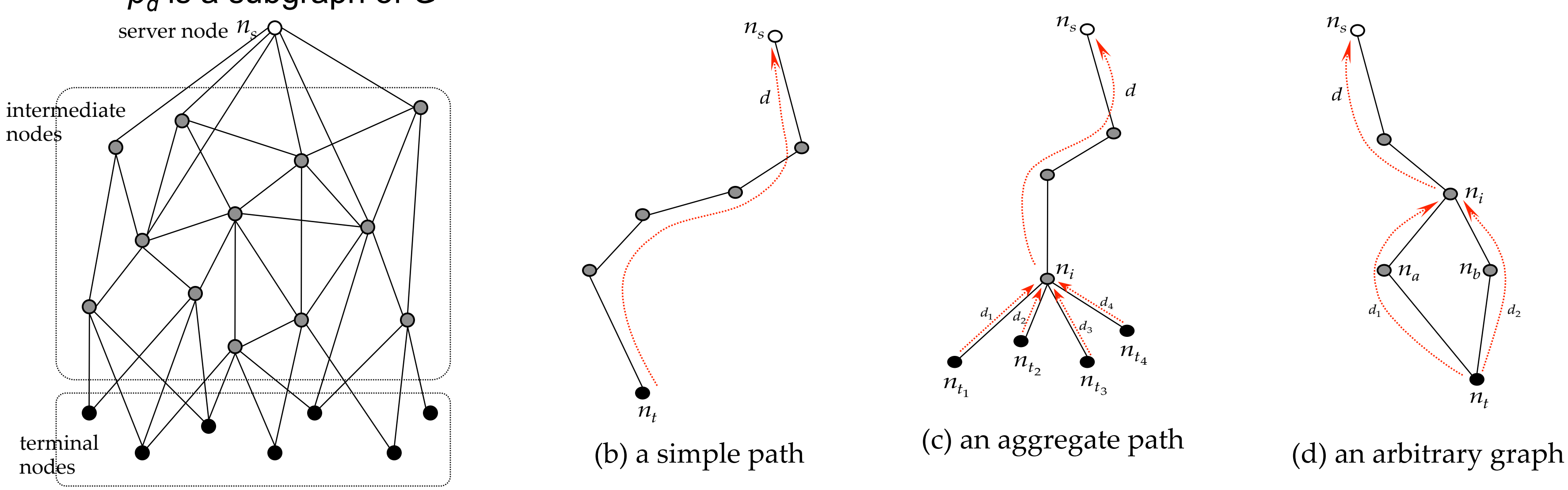
•H.S. Lim, Y.S. Moon, and E. Bertino, "Provenance-based Trustworthiness Assessment in Sensor Networks," presented at the 7th Workshop on Data Management for Sensor Networks (DMSN), in conjunction with VLDB, DMSN 2010, Singapore, 2010.

•S. Sultana, M. Shehab, and E. Bertino, "Secure Provenance Transmission for Streaming Data, IEEE Transactions on Knowledge and Data Engineering,, vol. PP, pp. 1-1, 2012.

•Dai, H.-S. Lim, E. Bertino, and Y.-S. Moon, "Assessing the trustworthiness of location data based on provenance," presented at the 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009, Seattle, WA, USA, 2009

Modeling Sensor Networks and Data Provenance

- A sensor network be a graph, $G(N,E)$
  - $N = \{ n_i | n_i$ is a network node of which identifier is $i \}$: a set of sensor nodes
    - a *terminal node* generates a data item and sends it to one or more intermediate or server nodes
    - an *intermediate node* receives data items from terminal or intermediate nodes, and it passes them to intermediate or server nodes
    - a *server node* receives data items and evaluates continuous queries based on those items
  - $E = \{ e_{i,j} | e_{i,j}$ is an edge connecting nodes n_i and n_j.$\}$: a set of edges connecting sensor nodes
- A data provenance, p_d
  - p_d is a subgraph of G



(a) a physical network (b) a simple path (c) an aggregate path (d) an arbitrary graph

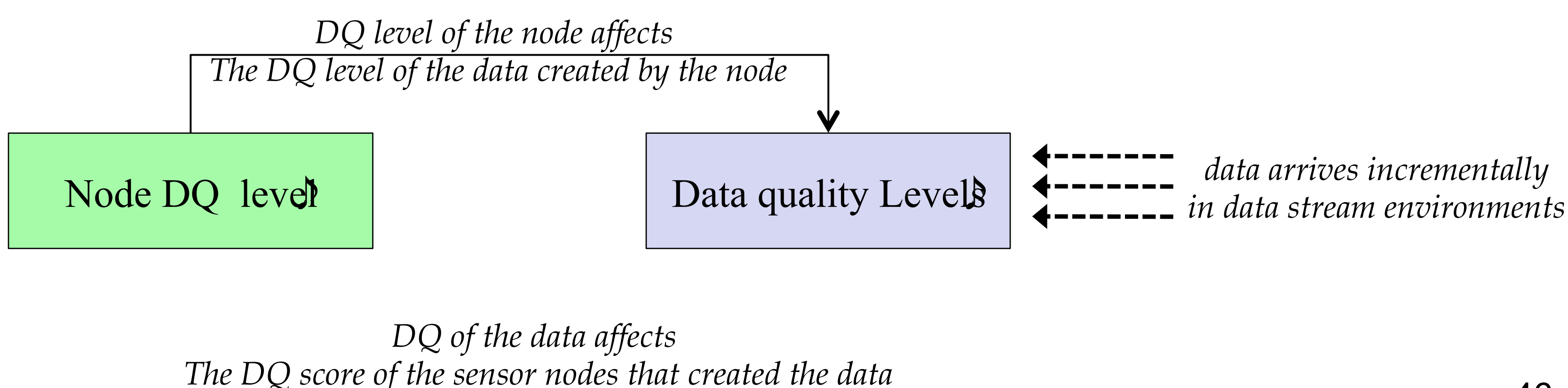
This slide was originally produced by E.Bertino (Purdue), modified by L.Reznik

Evaluating DQ → Computing DQ Levels

- DQ levels: *quantitative* measures of quality
  - **Data quality levels**: indicate about how much we can trust the data items
  - **Node DQ levels**: indicate about how much we can trust the sensor nodes collect quality data

➡ Levels provide an indication about the quality of data items/sensor nodes and can be used for comparison or ranking purpose

- ***Interdependency*** between data and node DQ levels

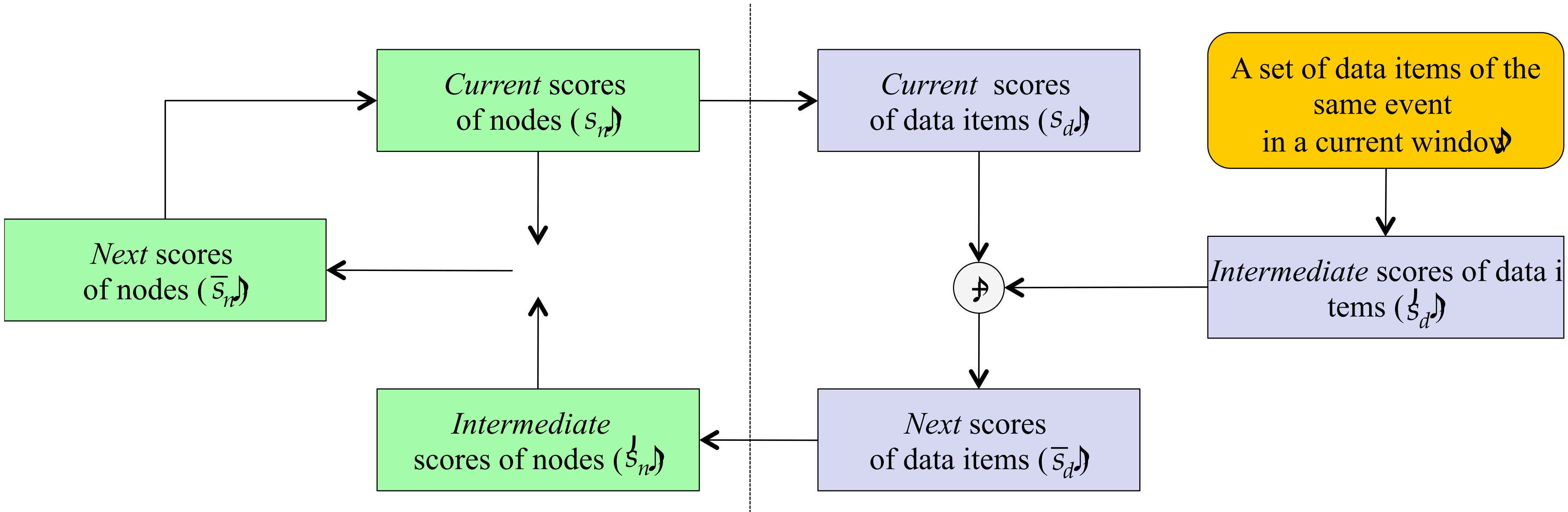


This slide was originally produced by E.Bertino (Purdue), modified by L.Reznik

Table 2. Metrics calculus (sample) for data streams fusion

Measure ID	Physical security incidents	Cryptographic system and communication protection	Measurement uncertainty
Formula for the data from one stream	P_i/A_i, where P_i is the number of physical security accidents in the i-th part of the network and A_i is the total number of access attempts in this part	C_i/D_i, where C_i is the number of computers and other devices that perform all cryptographic operations as recommended in the i-th part of the network and D_i is the total number of devices in this part	Standard deviation estimate is recommended as the probability in Table 1 could be calculated from it
Fusion of n data streams formula	$\text{Total} = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} A_i}$	$\text{Total} = \frac{\sum_{i=1}^{n} C_i}{\sum_{i=1}^{n} D_i}$	$s(z) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - z)^2}$, where Z_i is measurement from the i-th data stream and z is average from n streams
Relationship	$\min_i {}^{P_i}/_{A_i} \le \text{Total} \le \max_i {}^{P_i}/$	$\min_i {}^{C_i}/_{D_i} \le \text{Total} \le \max_i {}^{C_i}/$	$s(z) \le \min_i s(Z_i)$
Relationship if all n fused data have the same quality	Total =Pi/Ai	Total=Ci/Di	$s(z) = {}^{s(Z)}/_{\sqrt{n}}$

A Cyclic Framework for Computing DQ Levels



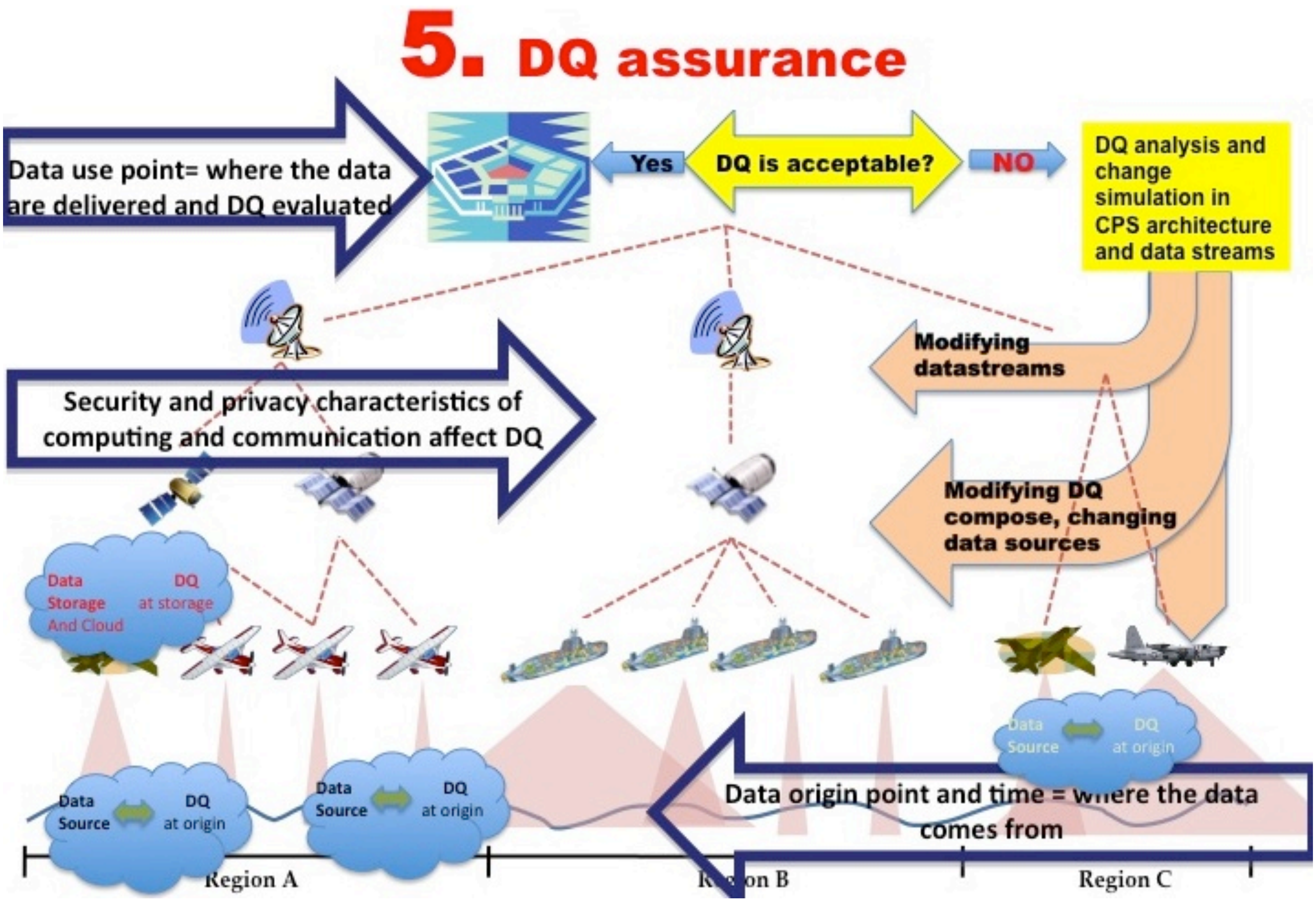
- DQ level of a data item *d*
  - The *current* level of *d* is the score computed from the current scores of its related nodes.
  - The *intermediate* DQ level of *d* is the score computed from a set (d ∈) *D* of data items of the same event.
  - The *next* level of *d* is the score computed from its current and intermediate scores.
- Quality level of a sensor node *n*
  - The *intermediate* level of *n* is the score computed from the (next) scores of data items.
  - The *next* level of *n* is the score computed from its current and intermediate scores.
  - The *current* level of *n*, is the score assigned to that node at the last stage.

This slide was originally produced by E.Bertino (Purdue), modified by L.Reznik

Planned


4. DQ re-adjustment through self-learning
DQ analysis and re-adjustment through self-learning
Data and DQ at destination
Data and DQ at destination
Data and DQ at destination
DQ analysis and re-adjustment by comparison against test cases
Test cases
Data use point= where the data are delivered and DQ evaluated

4. DQ re-adjustment through self-learning


5. DQ assurance
Data use point= where the data are delivered and DQ evaluated
Yes
DQ is acceptable?
NO
DQ analysis and change simulation in CPS architecture and data streams
Modifying datastreams
Security and privacy characteristics of computing and communication affect DQ
Modifying DQ compose, changing data sources
Data Storage And Cloud
DQ at storage
Data Source
DQ at origin
Data Source
DQ at origin
Data Source
DQ at origin
Data origin point and time = where the data comes from
Region A
Region C

Part 4:
System survivability and assurance with DQ

Major research challenges:

- developing DQ metrics and calculus procedures based on possible reactions to internal changes and external inputs, e.g. malicious alterations and attacks
- designing survivability assurance procedures based on DQ evaluation and possible changes

Prior research results:

•**J.Bacaj and L.Reznik** Signal Anomaly Based Attack Detection in Wireless Sensor Networks, **CCS '13:** Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, Berlin, November 2013

L. Reznik, Von Pless, G.; Al Karim, T. *Distributed Neural Networks for Signal Change Detection: On the Way to Cognition in Sensor Networks*, IEEE Sensors Journal, Volume 11 , Issue 3, March 2011, pp. 791-798

L. Reznik, M. J. Adams, and B. Woodard *Intelligent Intrusion Detection Based on Genetically Tuned Artificial Neural Networks,* Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.14, No.6 pp. 708-713, 2010

•M. Guirguis, J. Tharp, A. Bestavros, and I. Matta, "Assessment of Vulnerability of Content Adaptation Mechanisms to RoQ Attacks," in Networks, 2009. ICN '09. Eighth International Conference on, 2009, pp. 445-450.

•M. Guirguis, A. Bestavros, I. Matta, and Z. Yuting, "Reduction of Quality (RoQ) Attacks on Dynamic Load Balancers: Vulnerability Assessment and Design Tradeoffs," in INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE, 2007, pp. 857-865.

DQ as Safety Constraints

- A system/process is safe if minimal DQ constraints are met. Examples:
  - A control system is safe if the delay of the feedback signal is less than 10 msec and the noise to signal ratio is less than 5%.
  - A target recognition system is trustworthy if the native resolution of the video sensor is at least 720p.
  - A collision avoidance system is safe if GPS location accuracy is within 1 foot and delay is less than 0.1 sec.
- Assuming that DQ constraints on inputs hold, the DQ on the output is assured.

This slide was originally produced by A.Bestavros(Boston U), modified by L.Reznik

Compositional DQ analysis

- It’s good to be able to show that a system is safe given minimal DQ specs of its inputs
- It’s better to be able to show that the composition of DQ-safe systems is also safe
- It’s even better to be able to derive the DQ specs necessary for a system to be safe

➔ Need to do the above at scale!

IMPLEMENTATION ISSUES

Seattle Testbed Platform

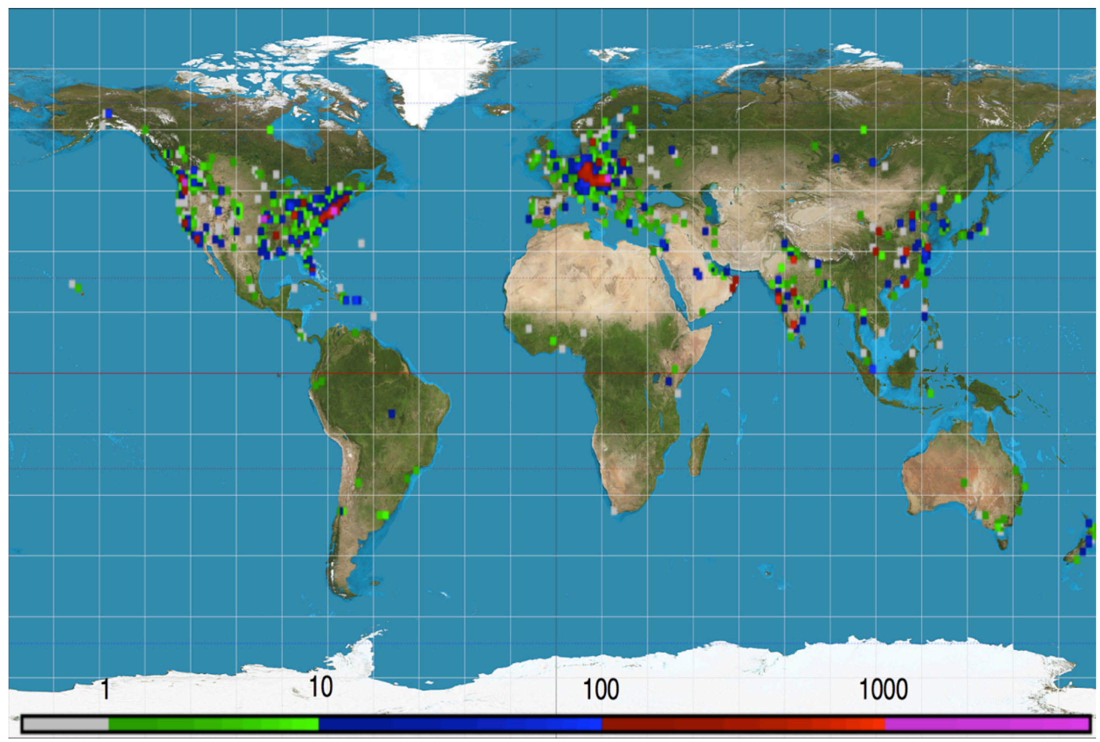
Tens of thousands of smartphones, tablets, laptops, ... used in various educational projects

- Networking, operating systems, security (~60 classes)
- Thousands of researchers at hundreds of universities

- Free
- Easy to learn
- Quick to get started

More info at

https://seattle.poly.edu



This slide was originally prepared by J. Cappos (NYU), A. Rafetseder (UVienna), Y. Zhuang(UBC & NYU), modified by L.Reznik

Based on Seattle platform

Sensibility: smart phone sensing for science and other applications

Offered scientists an easy way to collect real life data

Samples real data from smart phone sensors

- Users already agreed to participate
- Non-intrusive, not disturbing other daily activities
- Security and privacy issues considered

This slide was originally prepared by J. Cappos (NYU), A. Rafetseder (UVienna), Y. Zhuang(UBC&NYU), modified by L.Reznik

Sensibility Testbed

- **Sensing capabilities**

GPS

ambient light

Cellular

rotation

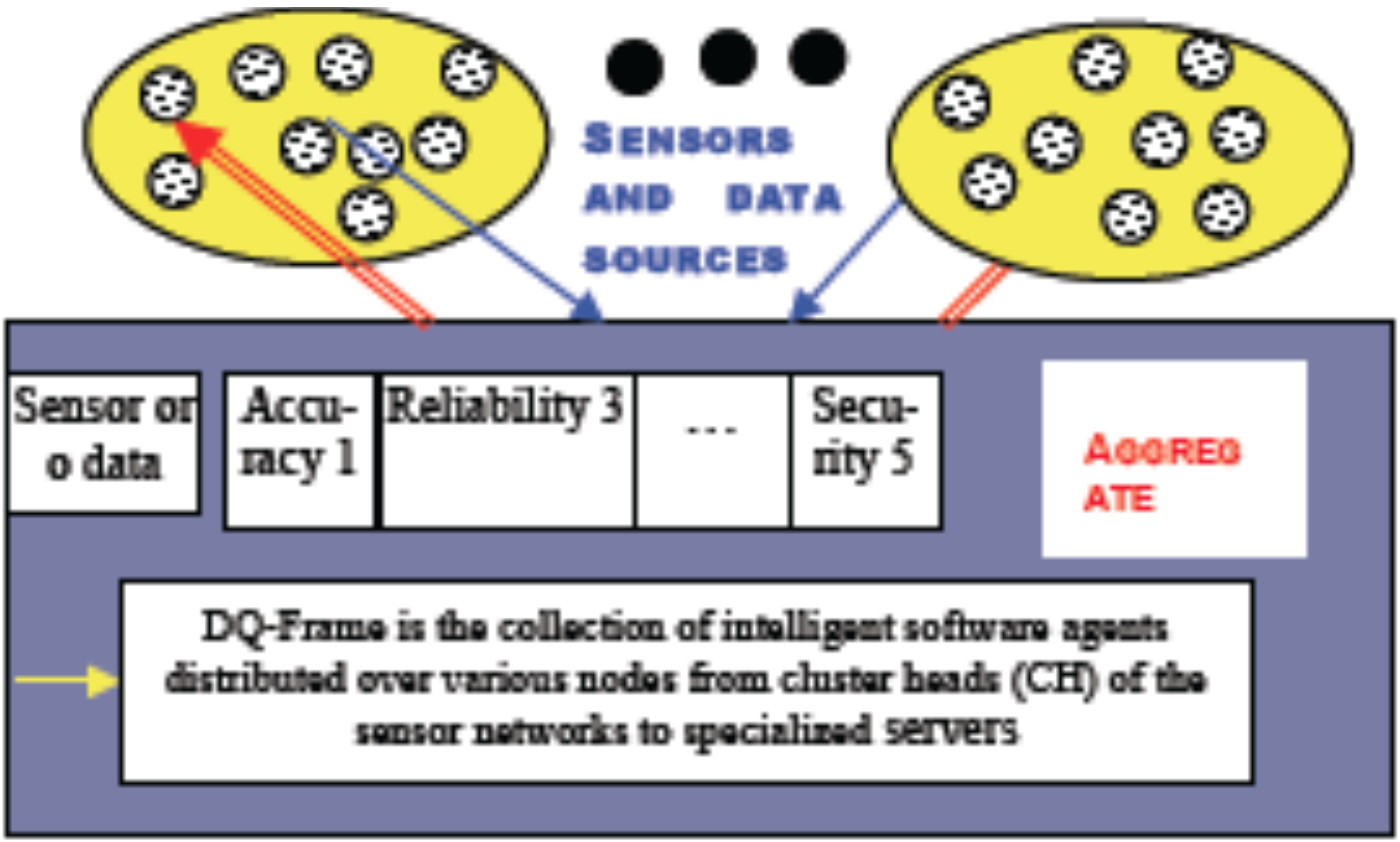
WiFi

Pre-release (alpha)

http://sensibilitytestbed.c

This slide was originally prepared by J. Cappos (NYU), A. Rafetseder (UVienna), Y. Zhuang(UBC&NYU), modified by L.Reznik

Software tools

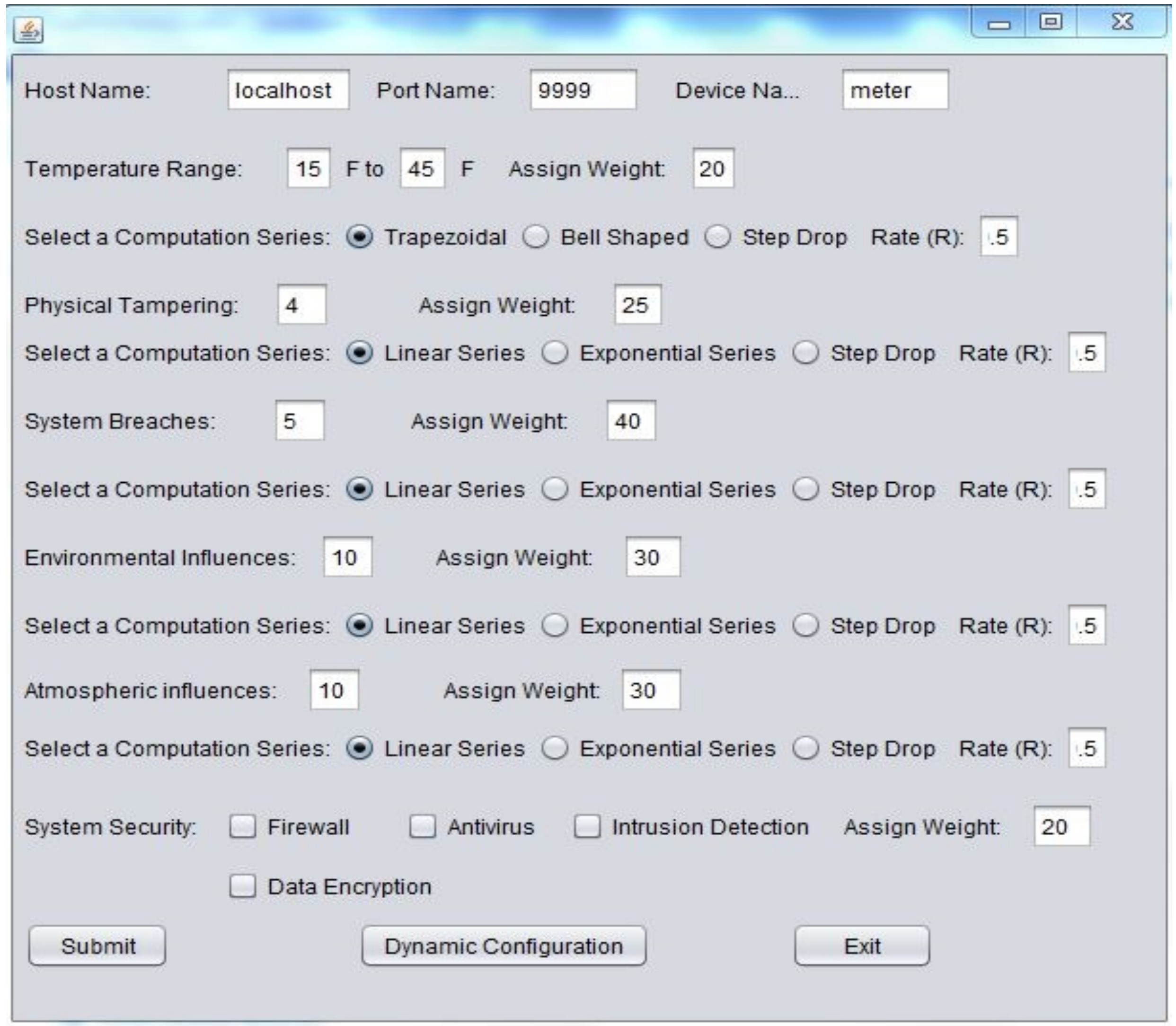


Software tools

1. DQ metrics assignment and configuration
2. DQ calculus
3. DQ dynamic calculation based on the provenance
4. DQ analysis and assurance

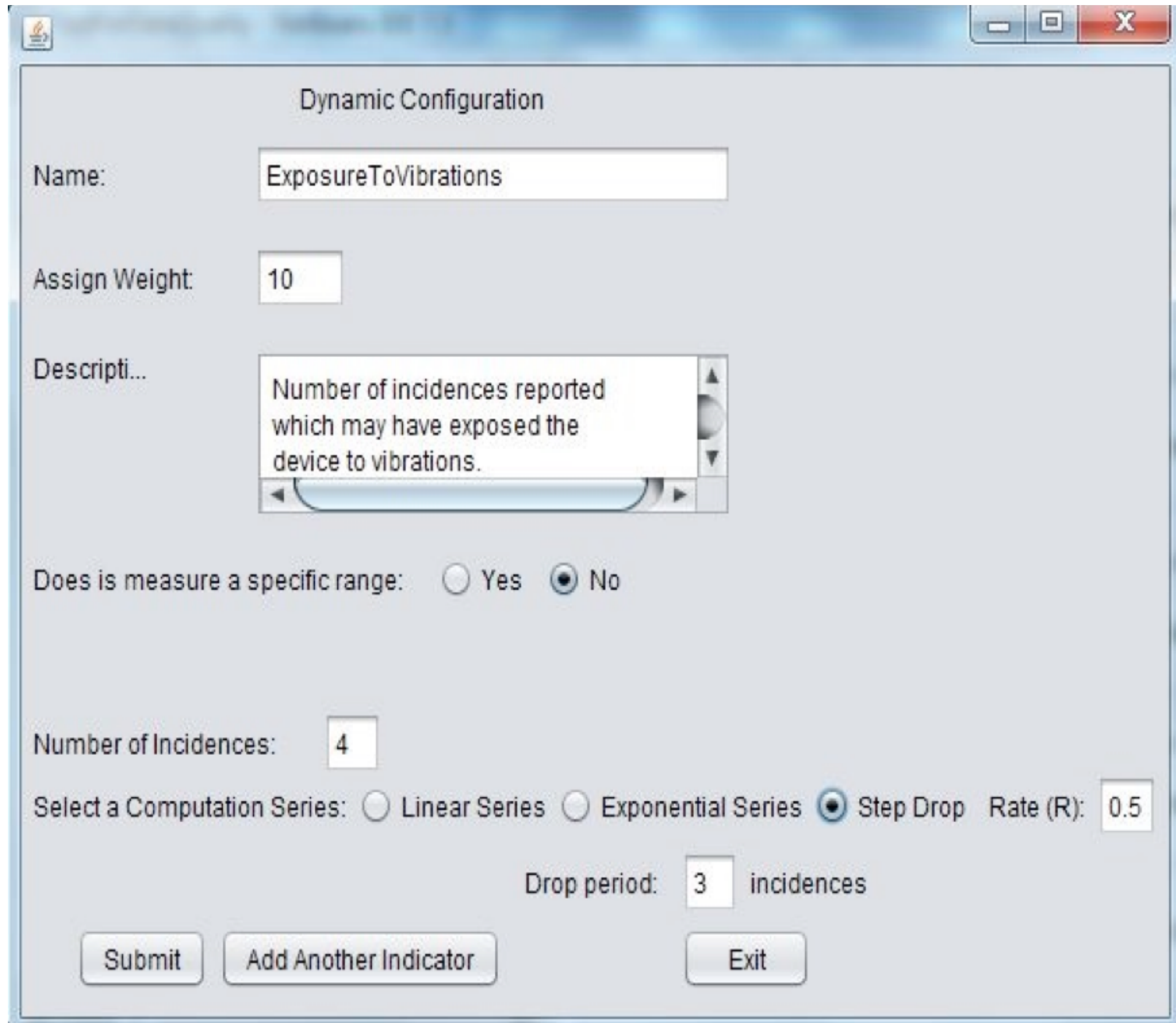
Software tools

DQ assignment and configuration: generic metric configuration interface portal

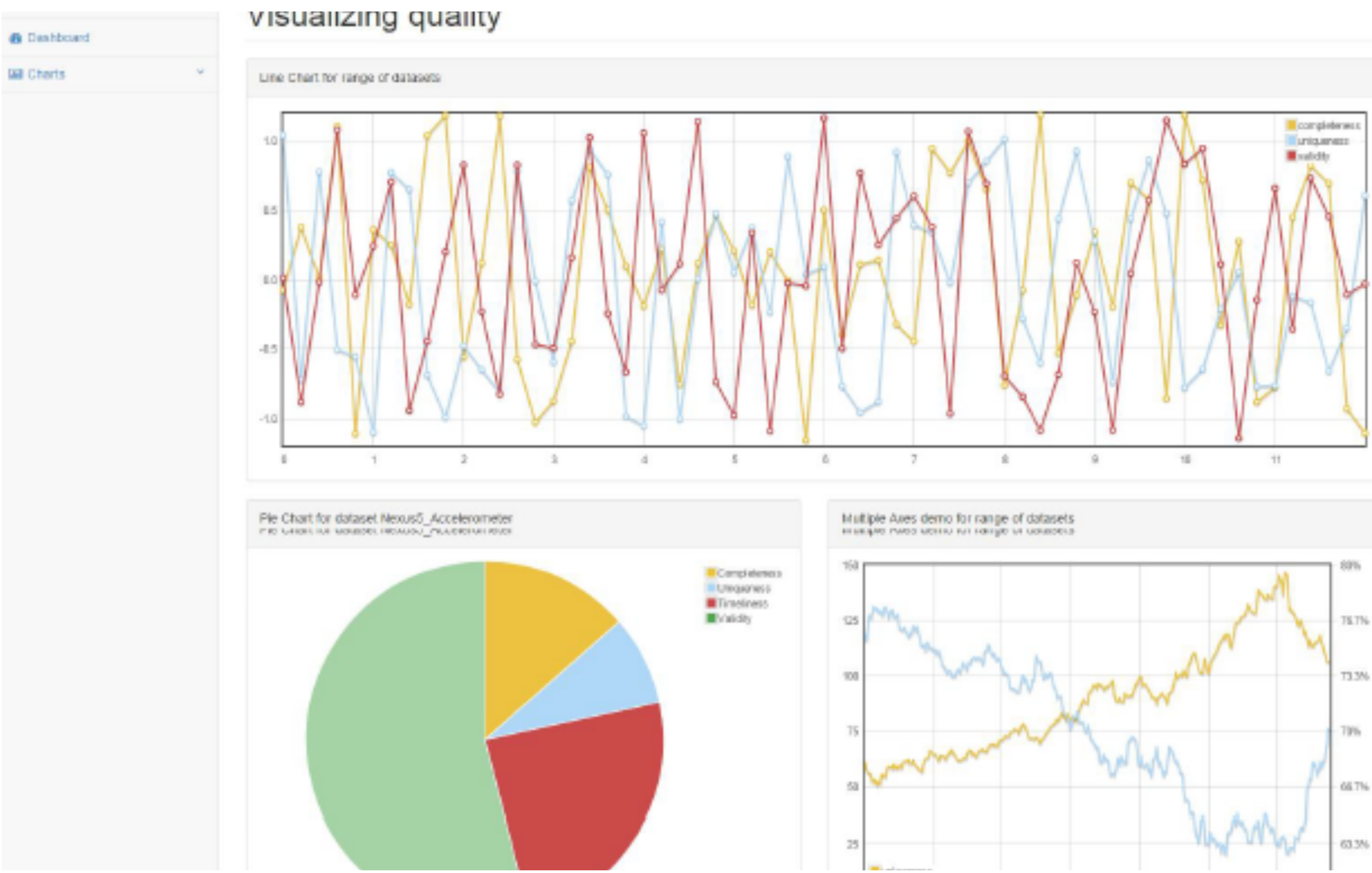


Software tools

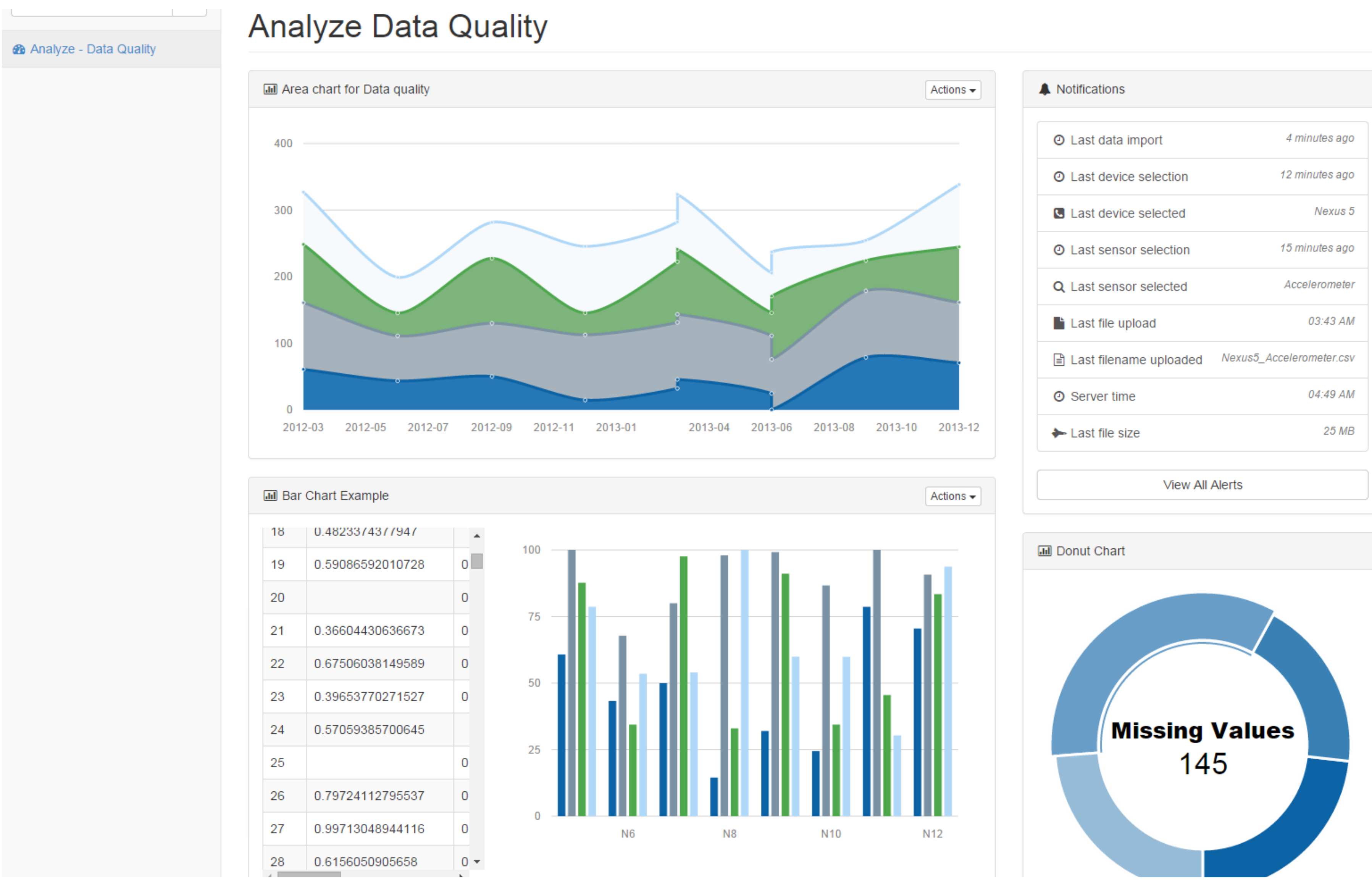
DQ assignment and configuration: specific metric configuration interface portal



Software tools: DQ visualization



Software tools: DQ analysis portal



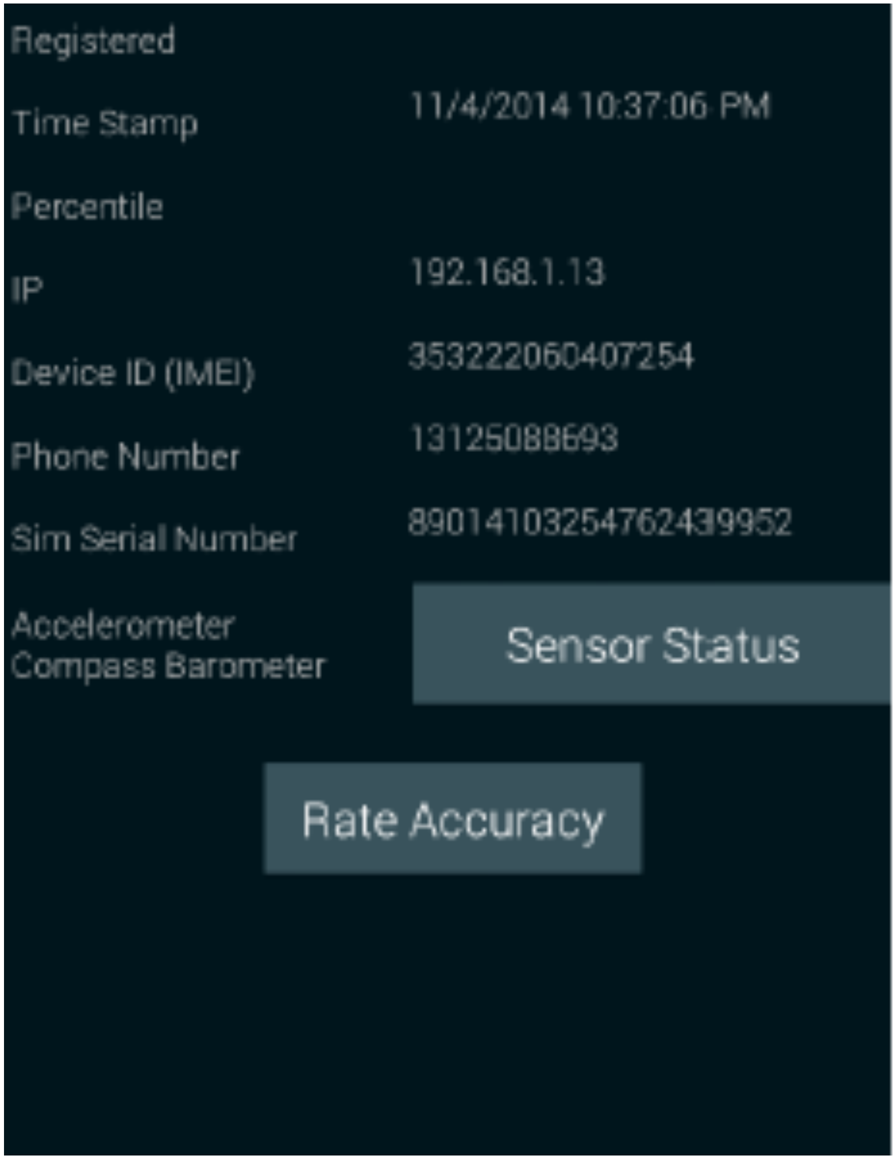
Software tools: Database portal

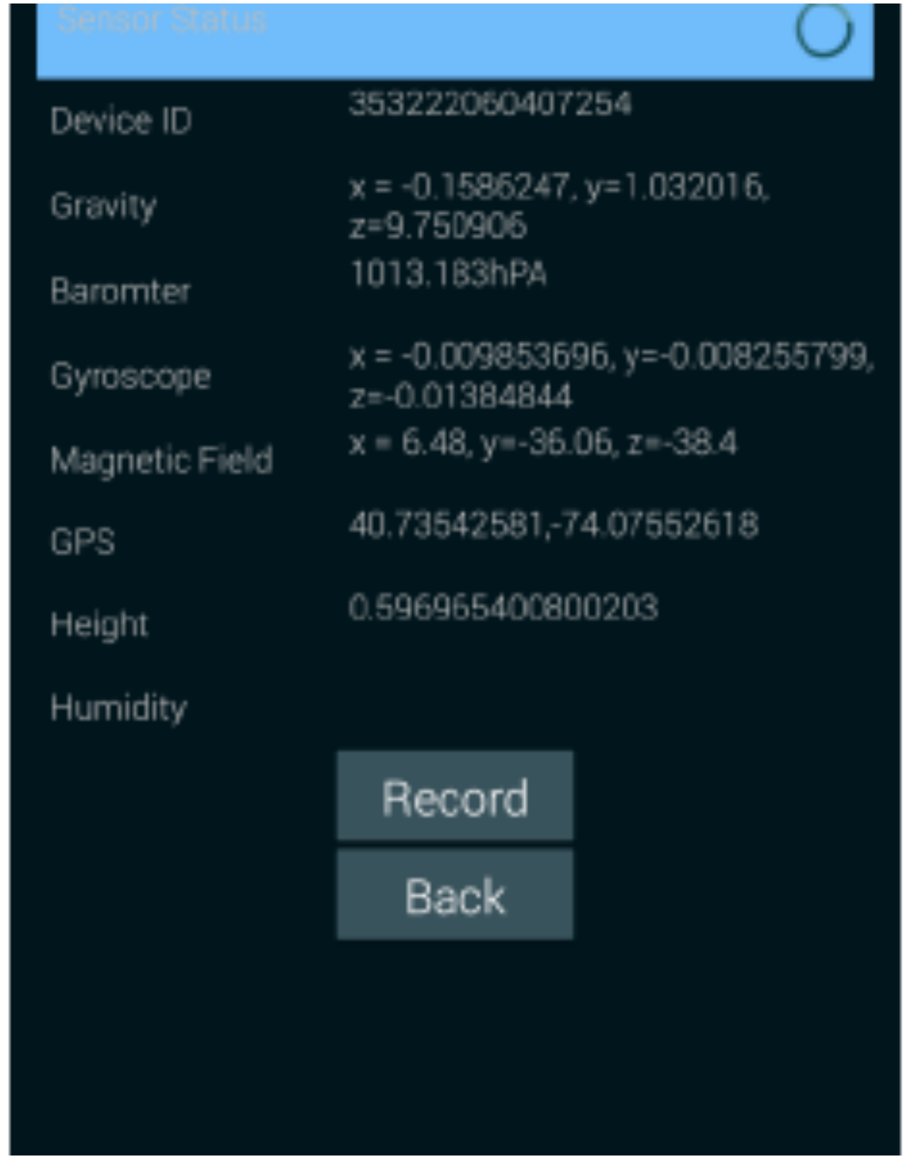
devicedata

BROWSE SCRIPT COLUMNS PERMISSIONS

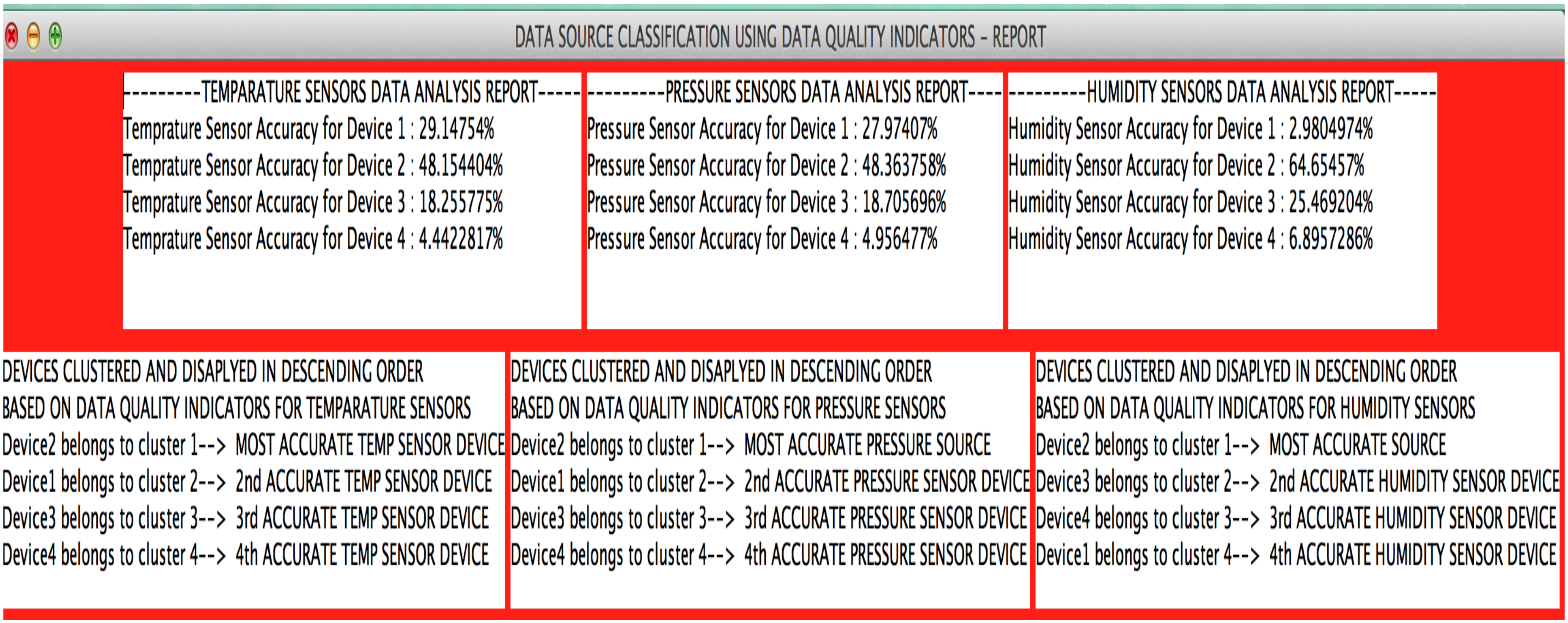
gravity_cumulative	gravity_x	gravity_y	gravity_z	error_gravity	pressure
9.806644320416053	0.26007146686315535	0.3094634175300598	9.798309421539306	-0.000055879583946 14436	1014.62
9.806644709407138	0.12720550820231438	0.3872726082801819	9.798169040679932	-0.0000552905928653244115	1014.68
9.806626815801371	1.5907947897911072	2.172760701179 5045	9.429655838012696	-0.0000731841986283 3991	1014.67
9.80664237546041	1.59655751247406	2.17391006946 5637	9.428466987609863	-0.00005762453958979563	1014.68
9.806643542433843	1.5926559295645032	2.1661062717437742	9.430893611907958	-0.0000564575661 5635378	1014.70
9.806645487589254	0.21559845358133517	0.34167487323 28415	9.798519816589355	-0.00005451261074 5405254	1014.59
9.806647045353307	0.23215997761276702	0.3072501954 0509904	9.799063805084228	-0.000052956646 69269229	1014.64
9.806645487589254	1.5988060857772828	2.16365124847 4121	9.430453491210937	-0.00005451261074 5405254	1014.71
9.806644320416053	1.596883603591919	2.17768726548 87696	9.427511596679688	-0.000055879583946 14436	1014.68
9.806639265330577	0.19892760854959487	0.33295048773288727	9.799006938934326	-0.00006075648 942184041	1014.59
9.806645098598203	1.5968572139739999	2.17550010681 15256	9.428018569946289	-0.00005490160179 6535108	1014.69
9.806643155442714	0.2036087393760681	0.33211355222911 8546	9.798902606964111	-0.0000568465572854393	1014.65
9.806643931424956	0.046016088686883 45	0.6508776605129242	9.784912109375	-0.00005636857504 3255464	1014.74
9.80663965252186	0.25063802697930143	0.3038506637117386	9.798726177215576	-0.00006034747813 96882	1014.60

Software tools: Sensor device portal





Software tools: Data category assignment based on DQ



DQ applications: Intrusion detection in sensor networks

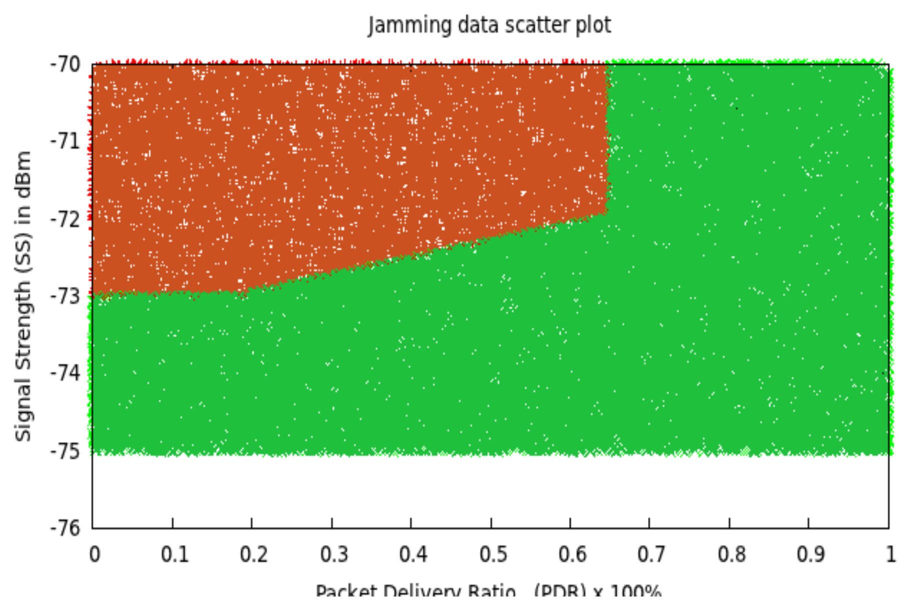
Study: detecting various attacks in sensor networks by building classifiers using various technological parameters (e.g. signal strength) and DQ metrics

Most successful study: jamming attacks – see here

Sources:

J.Bacaj and L.Reznik Signal Anomaly Based Attack Detection in Wireless Sensor Networks, **CCS '13:** Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, Berlin, November 2013

W. Xu, Ke Ma, W. Trappe, and Y. Zhang. Jamming sensor networks: Attack and defense strategies. In IEEE Network, volume 20, pages 41-47, June 2006.



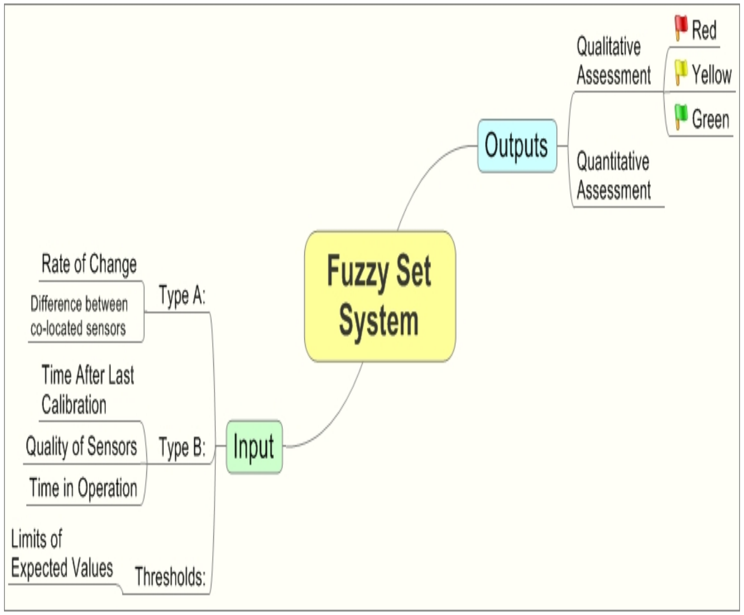
DQ applications: Fuzzy logic expert system for DQ evaluation in Tasmanian marine sensor network

A fuzzy rules-based system was implemented to assess the data quality at the sensor level. The system includes provisions for both Type A (rate of change of output values, cumulative rate of change of output values and node differences) and Type B (time since last calibration and time since last maintenance) uncertainty parameters.

Sources:

G.P. Timms, P.A. de Souza, Jr., L. Reznik and D. V. Smith Automated Data Quality Assessment of Marine Sensors, Sensors 2011, 11(10), p.9589-9602

G.P. Timms, P.A. de Souza, L. Reznik Automated assessment of data quality in marine sensor networks, IEEE International Conference OCEANS 2010 IEEE – Sydney, Australia, 24-27 May 2010, pp.1-5



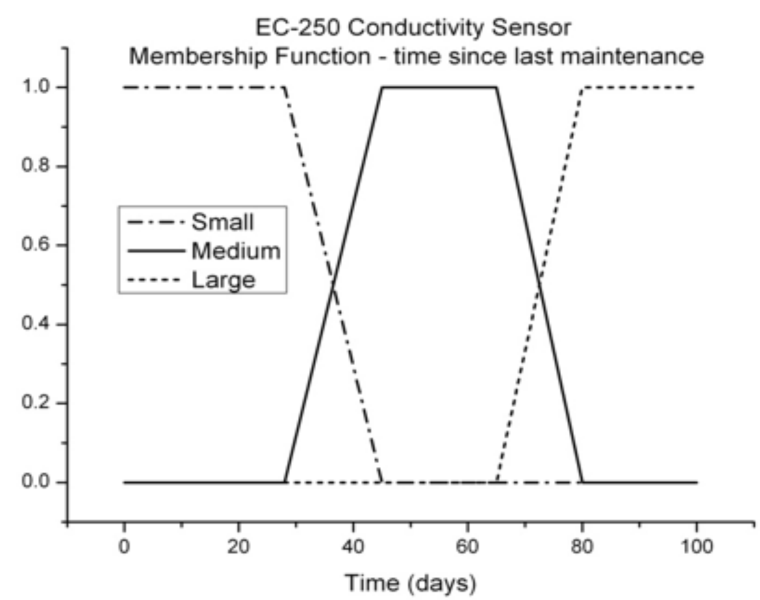
DQ applications: Fuzzy logic expert system for DQ evaluation in Tasmanian marine sensor network

An example of the membership function, for one of the temperature sensors (EC-250) and the Seabird37 conductivity, temperature and pressure sensors, is shown in Figure 2. Here, S (small), M (medium) and L (large) refer to the error introduced in the sensor output by this metric, time since last maintenance.

Sources:

G.P. Timms, P.A. de Souza, Jr., L. Reznik and D. V. Smith Automated Data Quality Assessment of Marine Sensors, Sensors 2011, 11(10), p. 9589-9602

G.P. Timms, P.A. de Souza, L. Reznik Automated assessment of data quality in marine sensor networks, IEEE International Conference OCEANS 2010 IEEE – Sydney, Australia, 24-27 May 2010, pp.1-5



DQ applications: Fuzzy logic expert system for DQ evaluation in Tasmanian marine sensor network

The system was initially applied to data collected at the CSIRO Wharf in Hobart between 25 August and 5 November 2010. This node was composed of two EC-250 temperature and conductivity sensors fixed to the wharf, one at a depth of 1.0 meters below chart datum and the other at a depth of 9.5 meters below chart datum (chart datum was at the level of the lowest possible astronomical low tide). The sensors were field calibrated in early September 2010, following their deployment on 25 August 2010.

Results:

The automatically generated error bars were expressed as a percentage of the manually-determined error bars over the 2672 datapoints for each of the four sensors.

The fuzzy system is much more successful when applied to the temperature sensors than to the conductivity sensors. In the case of the temperature sensors, the automatically generated error bars are within 50% of the manually determined error bars for approximately 80% of the time, compared with approximately 37% in the case of the conductivity sensors. The fuzzy system is also more successful at estimating error bars for the deeper sensors than for the shallower sensors.

Big and Quality
THANK YOU!

Leon Reznik, PhD
Professor of Computer Science and Computing Security
Rochester Institute of Technology
102 Lomb Memorial Drive Rochester NY 14623 USA
Ph.: 585 475 7210 Fax: 585 475 7100 email: l
http: //www.cs.rit.edu/~lr

Computer Science @ RIT
B.Thomas Golisano College of Computing & Information Sciences