– DataSys 2020 International Expert Panel –

— Methodologies and Methods —

# Information Processing

DataSys Congress 2020 / INFOCOMP, AICT, ICIW, SMART, IMMM, MOBILITY, SPWID, ACCSE, ICIMP

The Tenth International Conference on Advanced Communications and Computation (INFOCOMP 2020)

IARIA

DataSys Congress 2020

September 27 – October 1, 2020, Lisbon, Portugal

# DataSys Expert Panel: . . . Information Processing . . .

## Panelists and Contributors

- *Claus-Peter Rückemann* (Chair/Moderator),
  Westfälische Wilhelms-Universität Münster (WWU);
  KiM, DIMF, Germany Leibniz Universität Hannover;
- **Yanting Li**, *[Panelist and Contributor]*
  Shaoguan University, P. R. China
- **Irfan Khan Tanoli**, *[Panelist and Contributor]*
  University of Beira Interior, Covilhã, Portugal
- **Alexander Sim**, *[Panelist and Contributor]*
  Lawrence Berkeley National Laboratory, USA
- **Claus-Peter Rückemann**, *[Panelist and Contributor]*
  WWU Münster / DIMF / Leibniz Universität Hannover, Germany

DataSys Congress 2020 / INFOCOMP 2020: `http://www.iaria.org/conferences2020/INFOCOMP20.html`
Program: `http://www.iaria.org/conferences2020/ProgramINFOCOMP20.html`

# DataSys Expert Panel: . . . Information Processing . . .

## Panel foci, statements, topics, and preview:

- We should be aware information science fundaments and knowledge complements when dealing with procedural knowledge, especially information processing!
- We need to increase the deployment of multi-dimensional aspects of knowledge complements when dealing with information!
- We need to gather expertise on intrinsic and extrinsic information properties!
- Conceptual information can be beneficial for information processing!
- Structural information can be beneficial for information processing!
- Semantic information can be beneficial for information processing!
- Randomised sampling can be beneficial for efficiency of processing!
- Statistical pattern detection can be beneficial for information processing!
- When planning for information processing, we should always ensure to enable options, measures, and alternative methods also to be applicable for the purpose!
- Example case scenarios: Information processing for knowledge mining, semantic information processing, streaming data analysis, lossy compression, . . .

# DataSys Expert Panel: . . . Information Processing . . .

## Pre-Discussion-Wrapup:

- **Knowledge:** What are the differences between data, information, knowledge?
- **Focus:** Why are we processing information?
- **What do we understand by** 'optimal' information gathering?
- **What do we understand by** 'optimal' information filtering?
- **What do we understand by** information compression, . . .?
- **What is the difference** between methodologies and methods?
- **How can we address** structure, semantics, meaning, . . .?
- **Why is (natural) language unique?**
- **What are intrinsic properties when dealing with formalised approaches?**
- **What are the differences between 'complex' and 'complicated'?**
- **Networking:** Discussion! Open Questions? Suggestions for next Expert Panel?

# DataSys Expert Panel: Table of Presentations, Attached

**Panelist Presentations:** (presentation order, following pages)

– DataSys 2020 International Expert Panel on Information Processing –

# The Information Science Paragon:

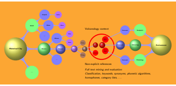
## Approaches to Universality, Consistency, and Long-term Sustainability

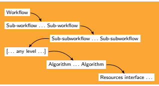
## A Prehistory to Future Case

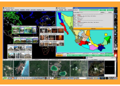
DataSys Congress 2020 / INFOCOMP, AICT, ICIW, SMART, IMMM, MOBILITY, SPWID, ACCSE, ICIMP
The Tenth International Conference on Advanced Communications and Computation (INFOCOMP 2020)
September 27 – October 1, 2020, Lisbon, Portugal

Dr. rer. nat. Claus-Peter Rückemann[1,2,3]

[1] Westfälische Wilhelms-Universität Münster (WWU), Münster, Germany
[2] Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF), Germany
[3] Leibniz Universität Hannover, Hannover, Germany
ruckema(at)uni-muenster.de





## Information: CV, lectures, studies, materials, research, and networking

**Curriculum Vitae:**

http://www.user.uni-hannover.de/cpr/x/rueckemann/en/

**Publications, lectures, and materials:**

http://www.user.uni-hannover.de/cpr/x/rueckemann/en/#Publications

http://www.user.uni-hannover.de/cpr/x/frodi/en/#Courses

**Congresses and venues:**

http://www.user.uni-hannover.de/cpr/x/rwerkr/en/

## Research

Dr. Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster
E-Mail: ruckema(at)uni-muenster.de

Chair of the Board on Advanced Computing and Emerging Technologies and
Chair of the Symposia Board, International Academy, Research, and Industry Association;
Chair of the Board of Trustees, Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung;
General Chair and Chair of the Steering Committee of
The International Conference on Advanced Communications and Computation (INFOCOMP);
General Chair and Chair of the Steering Committee of
The International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing);
Director GEXI Consortium; Head of research LX Foundation; Senior Member of Knowledge in Motion long-term project;
Fellow Member of the Int. HPC and Artificial Intelligence Advisory Council; Member of the Indexing Committee Board, IARIA;
Westfälische Wilhelms-Universität Münster (WWU);
Senior Lecturer Information Science, Security, and Computing at Leibniz Univ. Hannover; IARIA Fellow.

## Knowledge

- "Knowledge is created from a subjective combination of different attainments as there are intuition, experience, information, education, decision, power of persuasion and so on, which are selected, compared and balanced against each other, which are transformed, interpreted, and used in reasoning, also to infer further knowledge.
- Therefore, not all the knowledge can be explicitly formalised.
- Knowledge and content are multi- and inter-disciplinary long-term targets and values.
- In practice, powerful and secure information technology can support knowledge-based works and values."

**Citation:** *Rückemann, C.-P.; Hülsmann, F.; Gersbeck-Schierholz, B.; Skurowski, P.; and Staniszewski, M. (2015): Post-Summit Results, Delegates' Summit: Best Practice and Definitions of Knowledge and Computing; Sept. 23, 2015, The Fifth Symposium on Advanced Computation and Information in Natural and Applied Sciences (SACINAS), The 13th Internat. Conf. of Numerical Analysis and Applied Mathematics (ICNAAM), Sept. 23–29, 2015, Rhodes, Greece.*
*URL: http://www.user.uni-hannover.de/cpr/x/publ/2015/delegatessummit2015/rueckemann_icnaam2015_summit_summary.pdf*
*DOI: 10.15488/3409*

**Delegates and contributors:** *Claus-Peter Rückemann, Friedrich Hülsmann, Birgit Gersbeck-Schierholz, Knowledge in Motion / Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF), Germany;Przemysław Skurowski, Michał Staniszewski, Silesian University of Technology, Gliwice, Poland; International EULISP post-graduate participants, ISSC, European Legal Informatics Study Programme, Leibniz Universität Hannover, Germany*

**Complements of Knowledge and Corresponding Sample Implementations:**

- Factual Knowledge ⇔ Numerical data, data ...
- Conceptual Knowledge ⇔ Classification ...
- Procedural Knowledge ⇔ Computing ...
- Metacognitive Knowledge ⇔ Experience ...
- ...

(Source: Aristotle, 350 B.C.E.; Anderson & Krathwohl, 2001; SACINAS Delegates' Summit, 2015–2019)

## Conceptual knowledge references (Source: Excerpts from The Prehistory and Archaeology Knowledge Archive (PAKA), DIMF, 2020.)

| *Code / Sign Ref.* | *Verbal Description (EN)* |
|---|---|
| UDC:0 | Science and Knowledge. Organization. Computer Science. Information. Documentation. Librarianship. Institutions. Publications |
| UDC:1 | Philosophy. Psychology |
| UDC:2 | Religion. Theology |
| UDC:3 | Social Sciences |
| UDC:5 | Mathematics. Natural Sciences |
| UDC:6 | Applied Sciences. Medicine, Technology |
| UDC:7 | The Arts. Entertainment. Sport |
| UDC:8 | Linguistics. Literature |
| UDC:9 | Geography. Biography. History |
| UDC:001 | Science and knowledge in general |
| UDC:113 | General laws of nature. Transformation and transience of matter. Origin of the universe. Creation. Cosmogony |
| ... | |
| UDC:903 | Prehistory. Prehistoric remains, artefacts, antiquities |
| UDC:902 | Archaeology |
| UDC:904 | Cultural remains of historical times |
| UDC:93/94 | History |
| ... | |
| UDC:001.18 | Future of knowledge |

Multilingual Universal Decimal Classification Summary, 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php; Creative Commons Attribution Share Alike 3.0 license, 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ (first release 2009, subsequent update 2012).

## Status:

- **Information science fundaments and background**
  ... are often neither understood and taught nor practically respected.
- **Society, academy, education**
  are commonly reacting with simplification and training.

## Vision and future:

- **Insight** that knowledge is not the output or result of a tool.
- **Insight** that the fundament of any Turing machine / computer is formalisation (going along with abstraction and reduction).
- **Insight** that 'information processing' should be addressed via context of knowledge complements. This should also be true whenever addressing information, e.g., with gathering, filtering, compression of information but also whenever discussing hermetical criteria like quality and optimisation.
- **Require education** for solid fundaments, information science.

## Status:

- **Information science fundaments and background**
  ... are often neither understood and taught nor practically respected.
- **Society, academy, education**
  are commonly reacting with simplification and training.

## Vision and future:

- **Insight** that knowledge is not the output or result of a tool.
- **Insight** that the fundament of any Turing machine / computer is formalisation (going along with abstraction and reduction).
- **Insight** that 'information processing' should be addressed via context of knowledge complements. This should also be true whenever addressing information, e.g., with gathering, filtering, compression of information but also whenever discussing hermetical criteria like quality and optimisation.
- **Require education** for solid fundaments, information science.

## Conclusions on information processing:

- ‘Solid’ information processing requires a ‘solid’ understanding of information science fundaments.
- Procedural implementations should not be done without serious consideration of other knowledge complements, respective methodologies, and structural fundaments.
- Learn how to decide on multi-dimensional aspects of knowledge complements when dealing with information.
- Learn how to decide on intrinsic and extrinsic properties.
- Education is not training. Learning is not using tools.
- Consistency may not be modern but does it hurt?
- Foster information science education and its practice.
- Educational Taylorism is not an (future-oriented) option.

# Randomized Sampling

Yanting Li
Shaoguan University
yanting8015@sgu.edu.cn

**Information Processing Panel, IMMM 2020** IARIA

# Introduction of the presenter

- **Dr. Yanting Li**
  - Technical consultant
  - Associate professor of Shaoguan University
  - In charge of the natural language processing research lab at Shaoguan University
  - Mainly intake lectures of cloud computing, algorithm design and data analysis
- **Research interests**
  - Text data mining
  - Document abstraction
  - Information extraction
- **Contact:** yanting8015@hotmail.com or yanting8015@sgu.edu.cn

# Motivations

- Motivations
  - Various applications of community extraction in network analysis
  - Reducing the time cost and memory cost
  - Improve the sampling precision

- Key idea
  - Triangle is the smallest and densest community
  - Correlate the sampling of edges that the third edge will be sampled if two edges of a triangle are sampled

# Randomized Sampling

## How to generate the random values for coloring nodes?

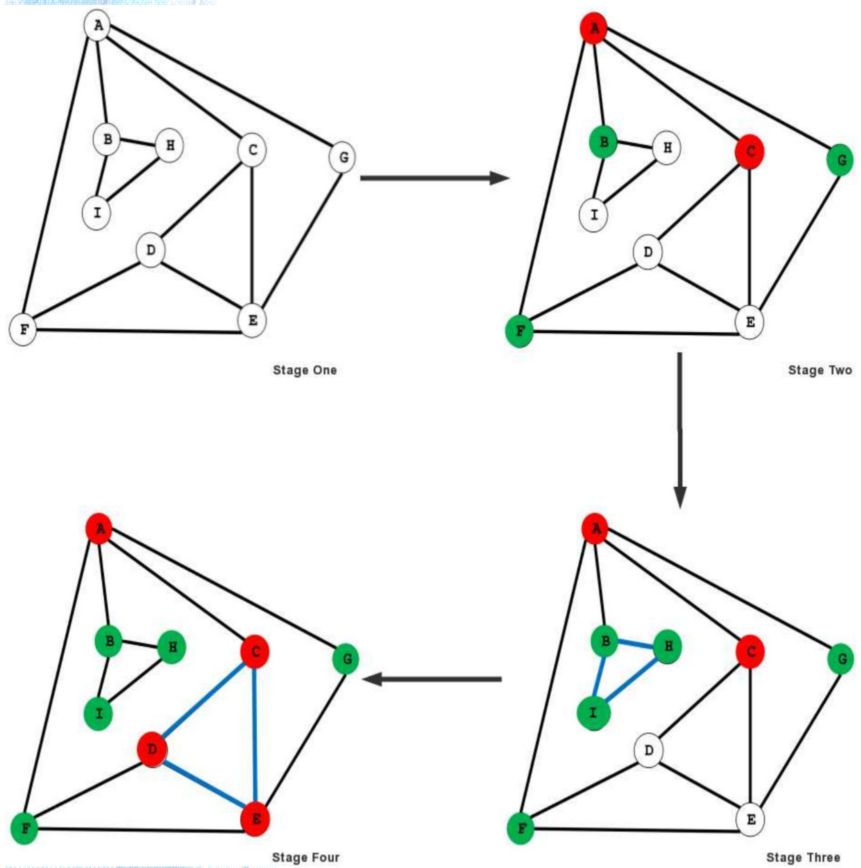
- Assume that the range of random values has finite expectations and variances mathematically. The generation of $R_v$ can be gained.

$$x_n + 1 = (\frac{x_n^{\ 2}}{10^s})(\mathrm{mod}\, 10^{2s})$$

- The ($X_n$ + 1) is an iterative operator, and ($R_v$ + 1) is the random value $R_v$ that needs to be generated every time. The $s$ is the shifting of $X_n$ square metre for generating new random value.

$$R_v + 1 = \frac{x_n + 1}{10^{2s}}$$

# Randomized Sampling

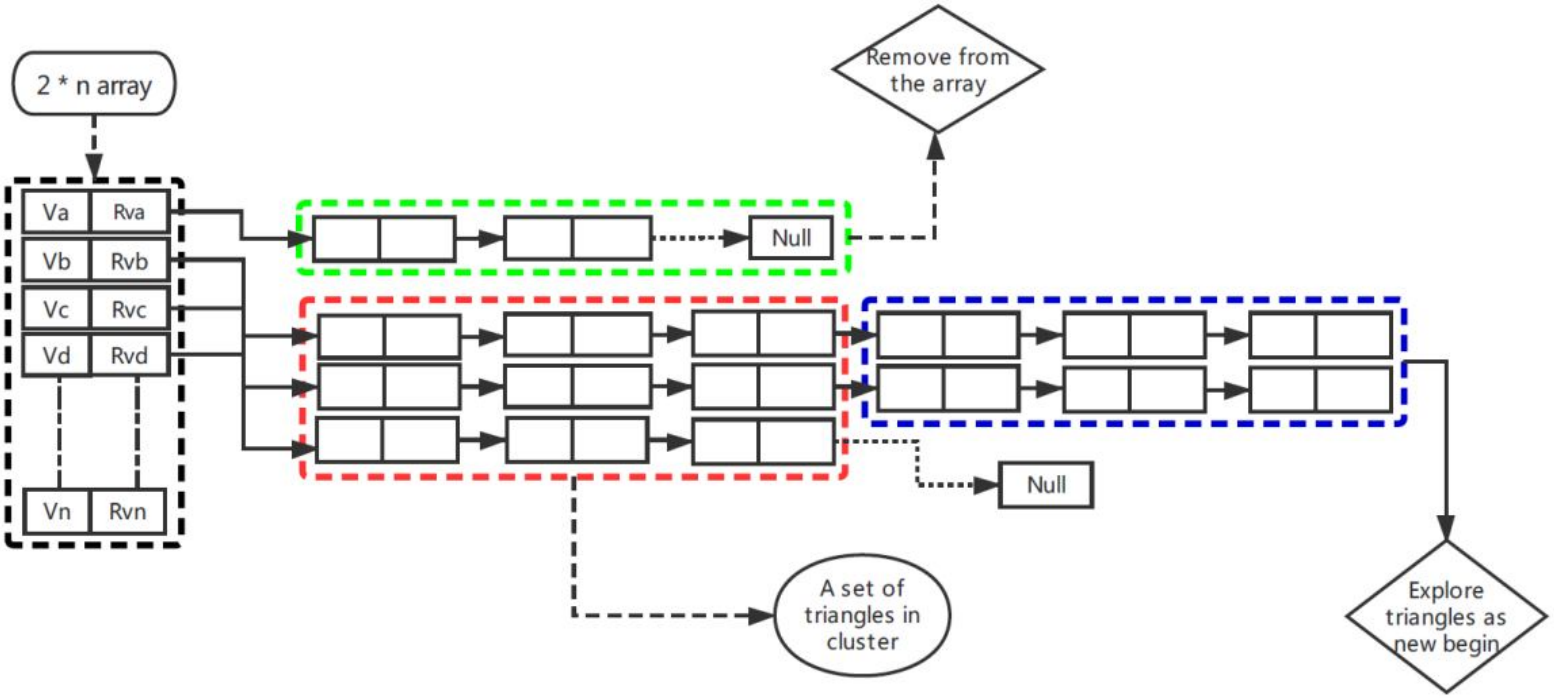



## How the randomized sampling algorithm works?

- The Breadth-First Search is employed for graph traverse
- The random value is given for coloring a node once the node is visited
- A triangle formed by three monochromatic edges is the samllest community in graph *G*

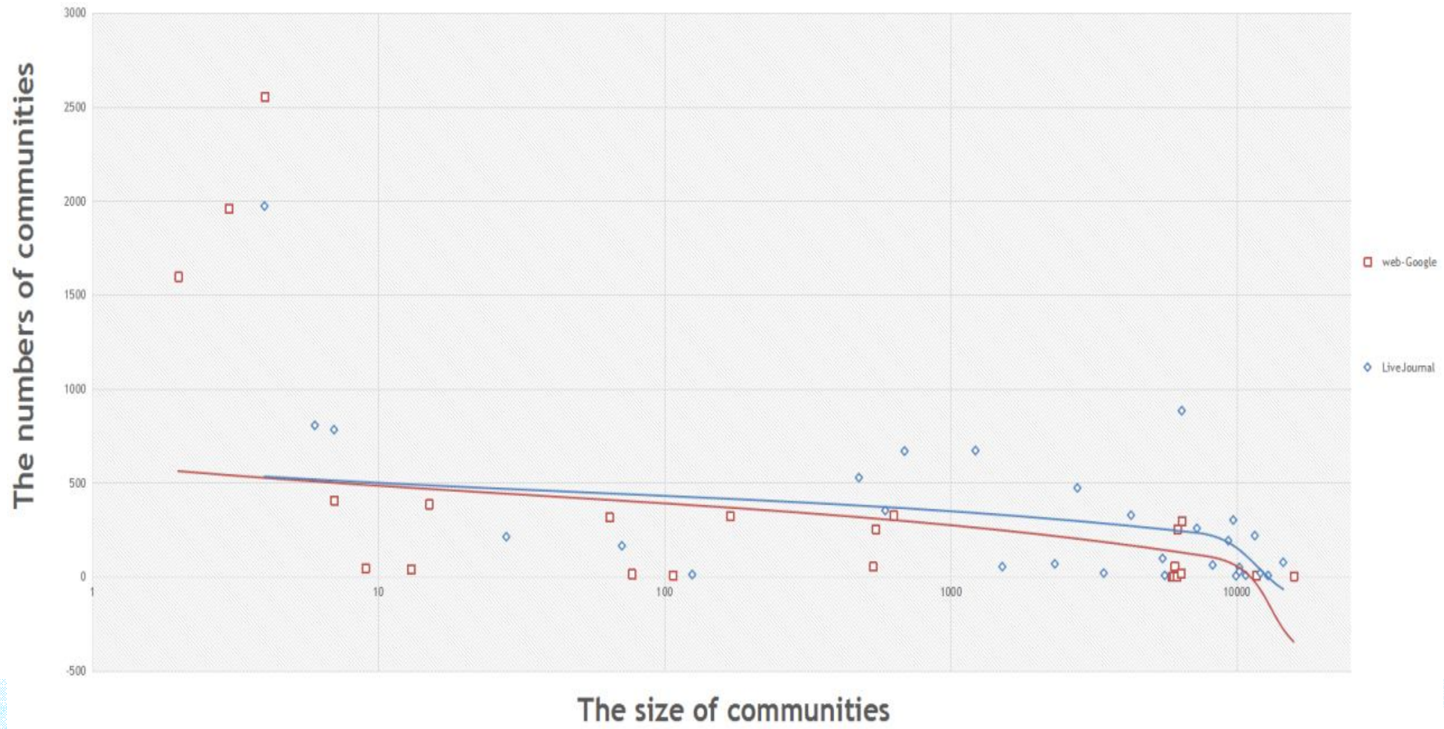
# Data Structure for Implementation

- **A (2 ∗ n) array list is employed for building the storage of all nodes in $V_G$ and their corresponding random values**



# Some Results

- Datasets for experiment (released by SNAP)
  - web-Google
  - com-LiveJournal
- Observation of communities distribution (based on size)



# Some Results

- This experiment records two experimental results
  - The maximum numbers of communities

| Dataset | Randomized Sampling | Reservior Sampling | Graph Priority Sampling |
|---|---|---|---|
| web-Google | 230018 | 13941 | 133925 |
| com-LiveJournal | 8632 | 7039 | 7780 |

  - The maximum density of samples

| Dataset | Randomized Sampling | Reservior Sampling | Graph Priority Sampling |
|---|---|---|---|
| web-Google | 0.92 | 0.85 | 0.836 |
| com-LiveJournal | 0.87 | 0.69 | 0.776 |

# Summary

- The randomized sampling algorithm combines the benifits of node-based sampling and edge-based sampling
  - **Triangle is the shortest complete cycle**
- Fast and less memory usage for real time execution
  - **Edge sampling**
  - **Triangle counting**
- Various applications of randomized sampling
  - Data clustering
  - Density analysis for networks
  - …

# Some Publications

**[Community Extraction]**

1. Yanting Li, Tetsuji Kuboyama, and Hiroshi Sakamoto. "Truss Decomposition for Extracting Communities in Bipartite Graphs." Proceedings of the 3rd International Conference on Advances in Information Mining and Management, 2013.

2. Yanting Li, Koji Maeda, Tetsuji Kuboyama, and Hiroshi Sakamoto. "An Extension of Community Extraction Algorithm on Bipartite Graph." The International Journal of Advances in Computer Science & Its Applications, Vol 4 (4), 2014.

**[Triangle]**

1. Kai Cheng, Yanting Li, and Xin Wang. "Single Document Summarization Based on Triangle Analysis of Dependency Graphs." Proceedings of Network-Based Information Systems (NBiS), 2013.

# SEMANTIC PROCESSING

Irfan Khan Tanoli
University of Beira Interior
Covilha, Portugal.
irfan.khan.tanoli@ubi.pt

Sebastião Pais
University of Beira Interior
Covilha, Portugal.
sebastiao@ubi.pt

Information Processing Panel, INFOCOMP 2020

# PRESENTER

- Irfan Khan Tanoli
  - Post-Doc Researcher at University of Beira Interior.
  - Receive PhD Degree at Gran Sasso Science Institute, L'Aquila, Italy.
  - Received the M.S degree in software engineering from Technical University of Madrid, Madrid, Spain.
  - Received the B.S degree in computer science from Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi, Pakistan.
  - Current Research Interest
    - Natural Language Processing
    - Controlled Natural Processing
    - Semantic Analysis
    - Software Engineering
    - Machine Learning
  - Current Research Project
    - Moves Project (http://moves.di.ubi.pt/)

IARIA

# PRESENTER

- Sebastião Pais
  - Professor at the Computer Science Department, the University of Beira Interior.
  - Researcher at NOVA-LINCS and GREYC Laboratory.
  - Received the PhD degree from MINES ParisTech - PSL, Paris.
  - Current research and teaching interests:
    - Artificial Intelligence.
    - Statistical Natural Language Processing.
    - Lexical Semantics.
    - Machine Learning.
    - Unsupervised and Language-Independent Methodologies.
  - Current Research Project
    - Moves (http://moves.di.ubi.pt/)
    - C4 - Cloud Computing Competence Centre (c4.ubi.pt)

# SEMANTIC PROCESSING

- Concerned with meaning of the sentence
- Many words have several meanings mean:
  - Verb to signify
  - adjective unpleasant or cheap
  - noun statistical average
- This is known as lexical ambiguity.
- To remove this, each word is associated with the context word senses, also called semantic markers

# SEMANTIC PROCESSING

- Location
- Physical-Object
- Animate-Object
- Abstract-Object

For example – **at** requires a time or a location as its object.

- The verb **hate** prefers a subject that is animate.
- John hates the cold – conveys the feeling of person.
- My lawn hates the cold – no animate object.

Other methods for semantic processing are

- Semantic grammars
- Case grammars
- Conceptual parsing
- Approximately compositional semantic interpretation.

# SEMANTIC GRAMMAR

- Encodes semantic information in syntactic grammar.
- Uses context-free rewrite rules with non terminal semantic constituents as attribute, object etc.
- The choice of non terminals and production rules are governed by semantic as well as syntactic information.
- The grammar rules are designed around key semantic concepts.

# SEMANTIC GRAMMAR

Consider the sentence

Example – 1 : I want to print Bill's .init file

In this the semantic action is a command action.

The grammar rules for the above sentence are

S → I want to Action

Action → Print File

File → File-name/File1

File1 → User's File2

File2 → Ext. File

Ext. → .init /.txt/.lsp

User → Bill

# SEMANTIC GRAMMAR

Example – 2 : What is the extension of file ?

S → What is File-property of file ?

In – this the semantic action is a query.

File-Property → The File-Prop

File-Prop → extension/protection/ creation date/owner

Applications Include

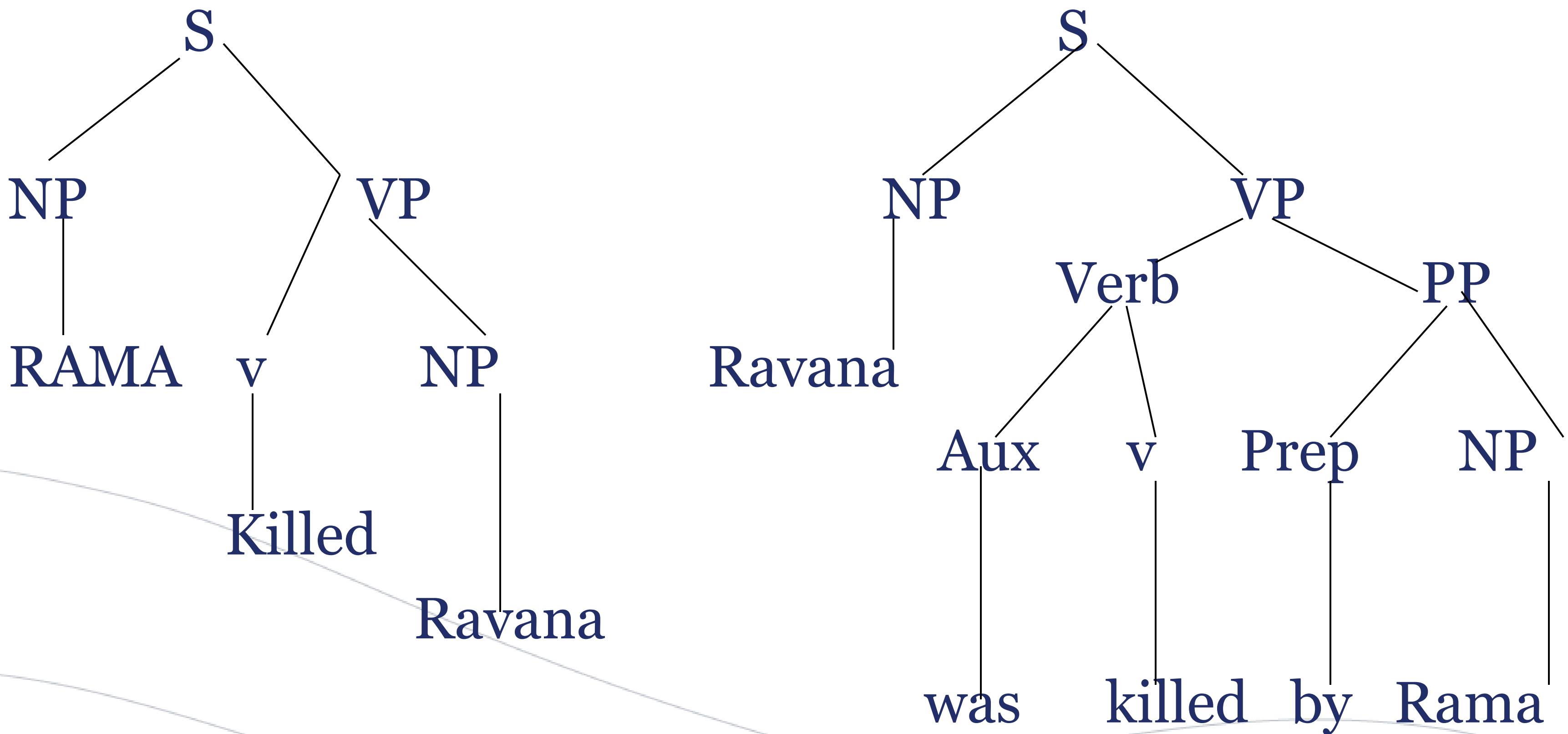
- Lifer - a Data Base query system for navy and
- Sophie - a Tutorial System to teach the debugging of circuit faults.
- Queries in LIFER include
  - "What is the name and location of the carrier nearest to New York ?"
  - "Who commands the KENNEDY?"

IARIA

# TRANSFORMATIONAL GRAMMARS

- Generative grammars produce different structures for sentences having the same meaning.
- Example :
  - Active and passive forms of a sentence Rama killed Ravana
  - Ravana was killed by Rama

# TRANSFORMATIONAL GRAMMARS




Structures for above active and passive sentences

# CASE GRAMMARS

- American linguist Charles J. Fillmore extended transformational grammars to include more semantic aspects.
- Case is used to extract the meaning of sentences.
- Mood : Group of forms of a verb indicating a fact, a possibility or a condition.
  - S → M + P
  - M – Modality constituent composed of mood, tense, aspect, negation etc.
- Mood represents verb category as indicative, imperative, or subjunctive (question, condition or a wish)
  - P – Consists of one or more cases
  - P → C1 + C2 + C3 + …. Ck

# CASE GRAMMARS

Examples of Cases :

- The case of an instigator of action (agentive case)
- The case of an instrument or object used in an action (instrumental case)
- The case of an object receiving the action (objective case)
- The case of location of an event (locative case)
- The case of an entity effected by an action (dative case) .

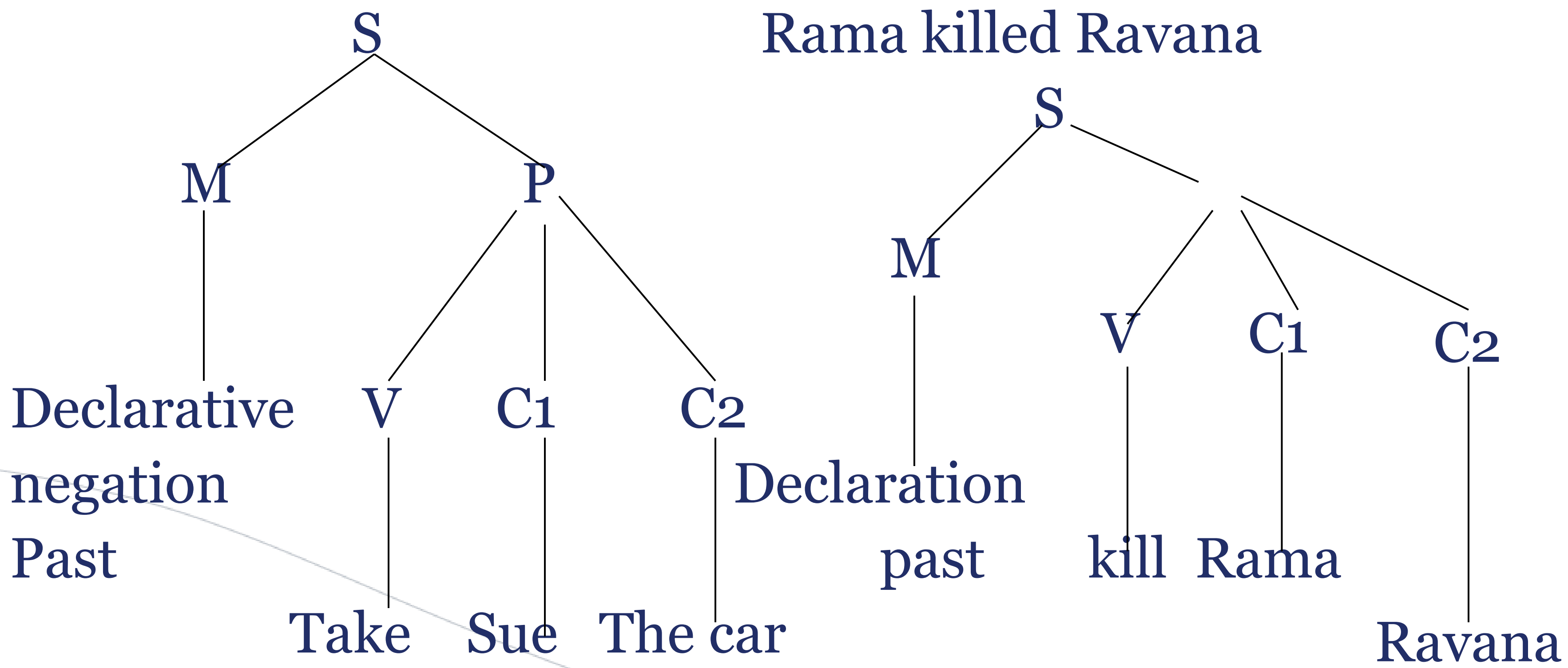
# CASE GRAMMARS

"The soldier struck the suspect with the rifle butt"

- The soldier is the agentive case.
- The suspect is the objective case.
- The rifle butt is the instrumental case.
- Case frames are provided for verbs to identify allowable cases.
- Struck [ Objective (Agentive) (Instrumental) ]
- Verb struck must occur in sentences with a noun phrase in the objective case and optionally with noun phrases in the agentive and instrumental cases.

IARIA

# CASE GRAMMARS




Case grammar tree representation for the sentence "Sue did not take the car" and "Rama killed Ravana"

# CONCEPTUAL PARSING

- Finds the structure and meaning of a sentence in one step.
- Uses a dictionary that describes the meaning of words as conceptual dependency (CD) structures.
- CD representation involves a syntactic processor that extracts the main noun and verb and determines the aspectual class of the verb based on the environment in which a verb can appear.
  - Example : The dictionary entries for want.
  - Wanting : something to happen (stative) (It must rain)
  - Wanting an object (Transitive) (The little boy wants a toy)
  - Wanting a person (intransitive) (The boss wanted his subordinate to work hard)

IARIA

# CONCEPTUAL PARSING

Consider the sentence

"John wanted Mary to go to the store".

O Marry the store

cf

Mary ⇔ ptrans

⇑I

John ⇔ pleased

< PTRANS --- Go >

⇑I relationship between two conceptualities

PP ⇔ PA means PP (picture producer) has an attribute PA (picture aider)

C condition f future, ptrans – go, ⇑I Relationship between two conceptualizations, ⇔ agent-verb relation

- The dictionary entry of the verb is chosen
- The conceptual processor analyses the sentence and fills the empty slots.

IARIA

# COMPOSITIONAL SEMANTICS

- Also known as Montague semantics.
- For every step in syntactic parsing process there is a corresponding step in semantic interpretation.
- Semantic interpretation rules are applied to each syntactic constituent and an interpretation for the sentence is produced
- consider the sentence

“I want to print Bill’s .init file”

# COMPOSITIONAL SEMANTICS

- Knowledge base (KB) for the sentence
- User

    is a : Person

    name : must be <string>
- Printing

    is a : Physical event

    agent : must be < animate>

    Object : must be <state or event>
- Wanting

    is a : Mental-event

    agent : must be <animate>

    performer : must be <animate or program>

    object : must be <event>

# COMPOSITIONAL SEMANTICS

- Compositional semantic rules describe the mapping of the verbs in terms of events in the KB.

Want → unit instance : wanting

Subject : agent : $RM_i$ ;

Object : object : $RM_j$ ;

($RM_i$ ; $RM_j$ are Reference Markers)

# COMPOSITIONAL SEMANTICS

- Print → unit
  instance : printing
- Subject : agent : RM i
- Object : object : RM j
- Init
  modifying NPI → unit for NP1 plus

  extension . init
- Possessive marker → unit for NP2 plus owner : NP1
  (NP1's NP2)
- "file"→ unit
  instance : File-structure
- "Bill" → unit
  instance : person
  first-name : Bill

(NP1, NP2 are NOUN PHRASE 1 and NOUN PHRASE 2)

# COMPOSITIONAL SEMANTICS

- The semantic rules implicitly make available to the semantic processing system all the information contained in the KB.
- For example, for the verb want, combining mapping knowledge in semantic rule with KB constraints,

unit

instance : wanting

agent : RM i

must be < animate >

object : RM j

must be < state or event >

# COMPOSITIONAL SEMANTICS

- There are limitations in qualified expressions
  - " John only eats meat on Friday and Mary does too".

can be interpreted as

  - Meat is the only thing that John eats on Friday.

- The control between syntactic and semantic processors can be:
  - apply semantic interpretation to a syntactic constituent.
  - parse entire sentence and interpret the whole thing.

IARIA

# DISCOURSE AND PRAGMATIC PROCESSING

- **Discourse** means written or spoken communication or a formal discussion of debate
- **Pragmatics** is a subfield of linguistics which studies the ways in which context contributes to meaning
- To recognize relationships among sentences, a great deal of knowledge about the world is required.
- Some examples of relationships between phrases and parts of discourse contexts
- Identical entities :
  - Bill had red balloon
  - John wanted it red balloon
- Parts of entities

  Sue opened the book she just bought

  The title page was torn (refers to the page of the book)
- Parts of action :
  - John went on a business trip to New York
  - He left on an early morning flight

Taking a flight refers to the action of going on a trip

# DISCOURSE AND PRAGMATIC PROCESSING

- Names of individuals :
  - Dave went to the movie person’s name

- Casual chains :
  - There was big snow storm yesterday
  - The schools were closed today
  - snow storm was the reason for closing of schools.

- Planning Sequence :

  Salley wanted a new car.

  *Getting a job due to desire for a new car*

  She decided to get a job.

Illocutionary force : It sure is cold in here

*intended effect may be expecting to close the window or turn up the thermostat.*

# DISCOURSE AND PRAGMATIC PROCESSING

- Implicit Presumptions
  - Did joe clear CSIO1
- Presuppositions Include
  - CSIO1 is a valid course
  - John is a student
  - John took the course
- Programs to understand such contexts require large knowledge bases or strong constraints on the domain of discourse to limit the KB.
- The way the knowledge is organized is critical to the success of the understanding program.

# KINDS OF KNOWLEDGE IN DISCOURSE AND PRAGRAMATIC PROGRAMMING

- Four kinds of knowledge can be identified
  1. The current focus of the dialogue.
  2. A model of each participant's current beliefs
  3. The goal driven character of dialogue
  4. The rules of conversation shared by all participants.
- The goal is to reason about objects, events, goals, beliefs, plans and likelihoods into NLU.

# USING FOCUS IN UNDERSTANDING

- There are two tasks :
  - Focus on the relevant parts of the KB.
  - Use that knowledge to make connections among things that were said.
- Some mechanisms for focusing.
  - Using appropriate scripts such as hotel script.
  - By giving highly simplified instruction
  - To make the cake, combine all ingredients pour them into the pan, and bake for 30 mns.
- Use phrases (explicitly) such as “on the other hand” to return to an earlier topic or “a second issue is” to denote the continuation of a topic.

# HOW TO USE THE FOCUSED KNOWLEDGE ?

- Any object in a KB relates somehow to almost to any other. Some highly important relations include physical-part-of, temporal part-of, and element-of.
- Consider "sue opened the book she just bought"
  - "The title page was torn"
- In this physical-part-of relates the title page to the book that is in focus.

# THANK YOU

# Statistical Pattern Detection with Locally Exchangeable Measures

Alex Sim

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

In collaboration with K. Wu, D. Lee, J. Choi,
O. Del Guercio, K. Gibson, R. Orozco, K. Hu

**Information Processing Panel, INFOCOMP 2020**

# Introduction

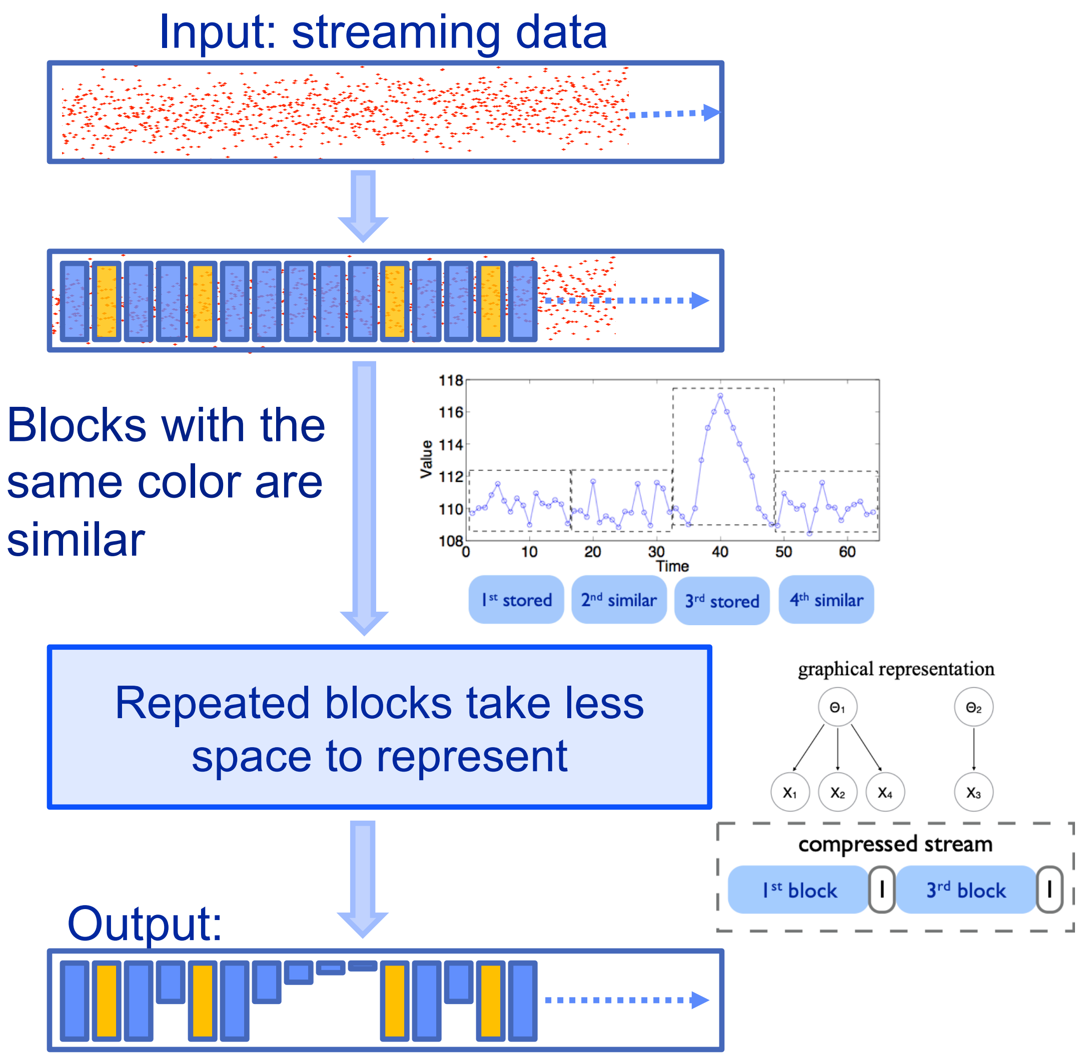
- **Alex Sim**
  - Senior Computing Engineer
  - Lawrence Berkeley National Laboratory, USA
  - Has been working with science applications and communities, mostly for data management, dynamic resource management, data analysis, HPC I/O optimization, and distributed workflow optimization
  - Current research interests
    - Performance modeling and prediction, anomaly detection and classification, edge computing, statistical learning, and various aspects of machine learning, especially on federated and distributed learning
  - Senior member of IEEE
  - Contact: asim@lbl.gov, http://www.lbl.gov/~asim

# Motivation/Observations

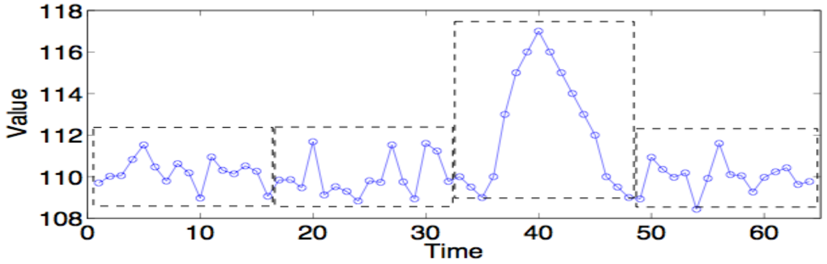
- **Motivation**
  - Large streaming data needs a lot of storage
  - Statistical analysis is needed on big data
  - Challenges in many scientific measurements
    - Floating-point numbers are known to be hard to compress
    - “Random” fluctuations are hard to compress
  - Exact compression of big streaming data is intractable, in general
    - Alternative: Linear random sampling, e.g. 1 out of 1000 records
      - It is not scalable for high-rate multiple streaming data
      - There is no guarantee of reflecting the underlying data distribution
- **Observations**
  - Large streaming data tend to show redundant data patterns
  - Many conventional statistical methods are based on a specific assumption (exchangeability)

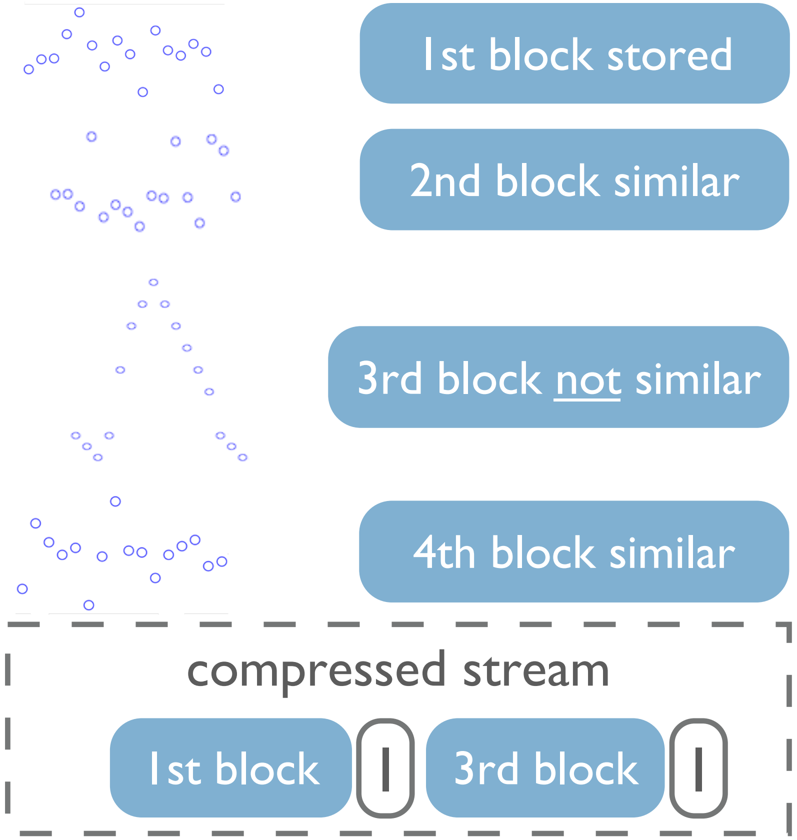
# Locally Exchangeable Measures – New Perspective on Data Compression

- Random-looking sequence of values are hard to compress, can we do something about it?
  - IDEALEM (Implementation of Dynamic Extensible Adaptive Locally Exchangeable Measures)
  - SSDBM2016, 2017, BigData2018, DCC2019, SNTA2019
- Relaxing order of values opens up new horizon on data compression
  - Information loss due to compression has been generally measured by Euclidean distance ($L^2$-norm) between original data and reconstructed data with MSE/SNR criteria
    - High entropy (nearly random) data and floating-point values are hard to compress
  - Limitation: order of values not preserved
    - Is the order of values really important?
    - Devices such as sensors often measure random fluctuations
    - Exact reproduction of random fluctuations is not necessary



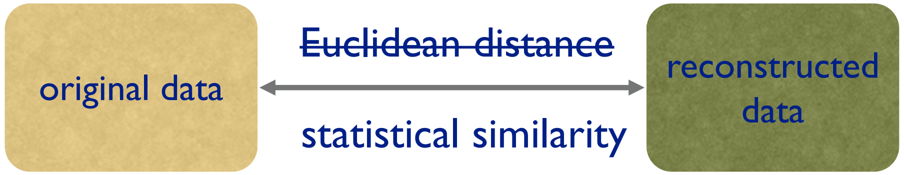
# How it works



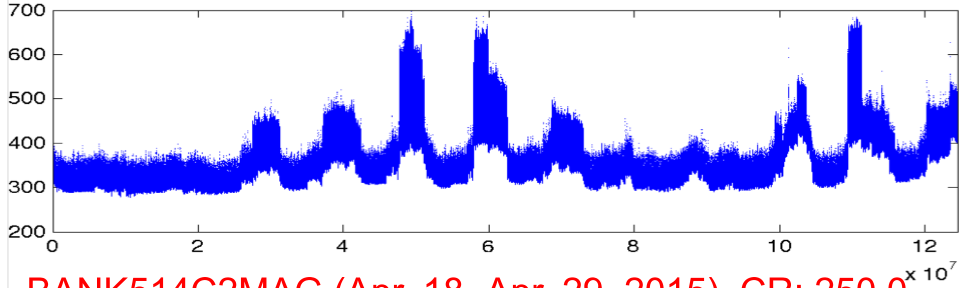

**Breaks an incoming data stream into blocks of a fixed size**

- **Represents similar blocks with the one that appears earlier in the sequence**
- **Similarity here is based on statistical measure**
  - **Not on Euclidean distance**
  - **Kolmogorov-Smirnov test (KS test)**

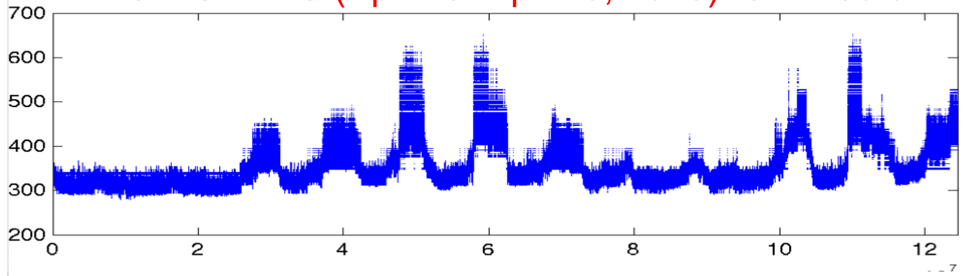


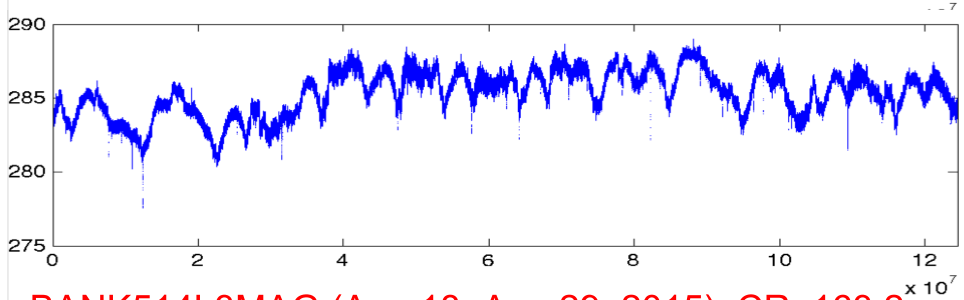
# Examples on Power Grid Monitoring (µPMU) data

- **For µPMU data,**
  - Compression will reduce the data volume to be sent around the data network
  - Compression will remove redundant information and make it easer to locate the interesting information
- **Characteristics of µPMU Measurements**
  - Numerical values: voltage, current, phase angles for voltage and currents
  - Typically have a lot of "random" "small" fluctuations that are considered normal for the electric power grid system
  - Occasionally, has relatively "large" changes that require attention or intervention

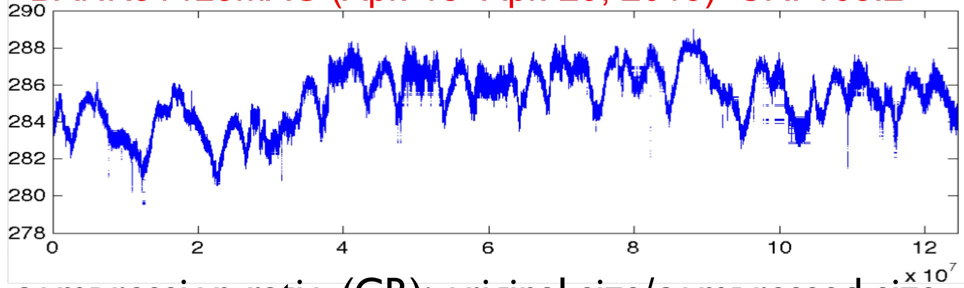


BANK514C2MAG (Apr. 18~Apr. 29, 2015) CR: 250.0



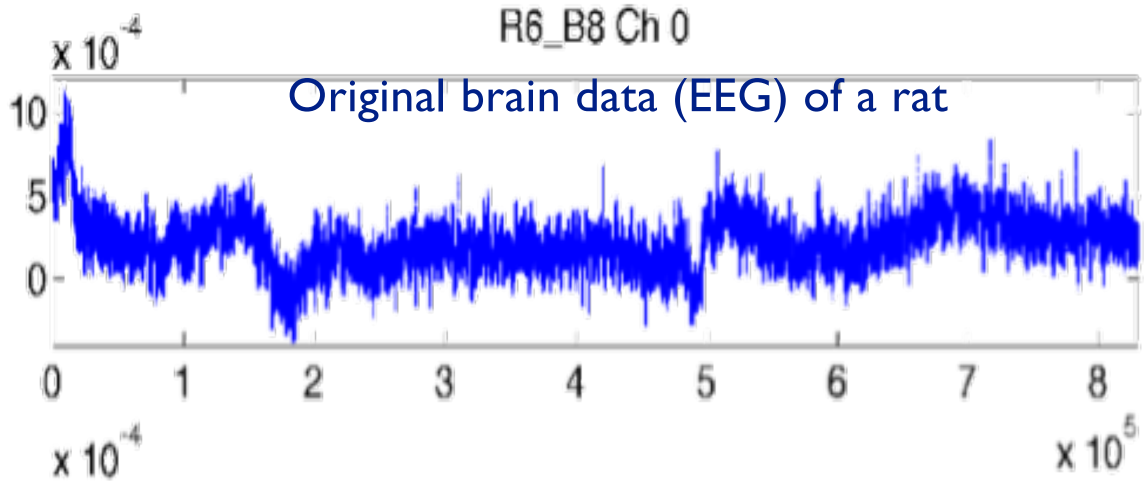


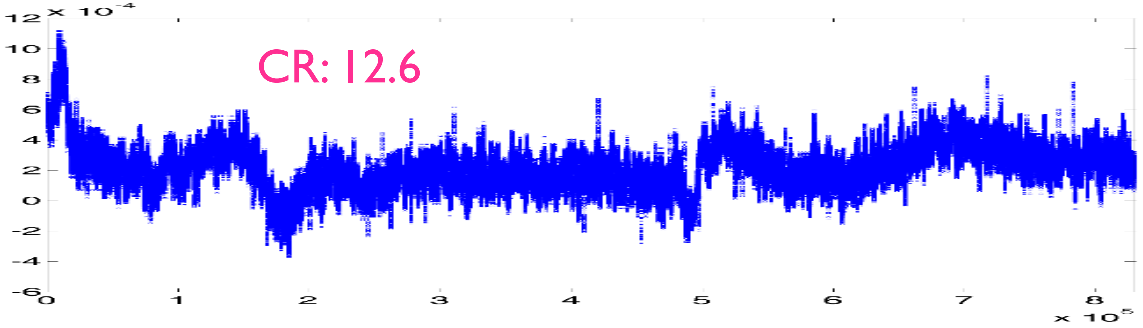
BANK514L3MAG (Apr. 18~Apr. 29, 2015) CR: 163.2

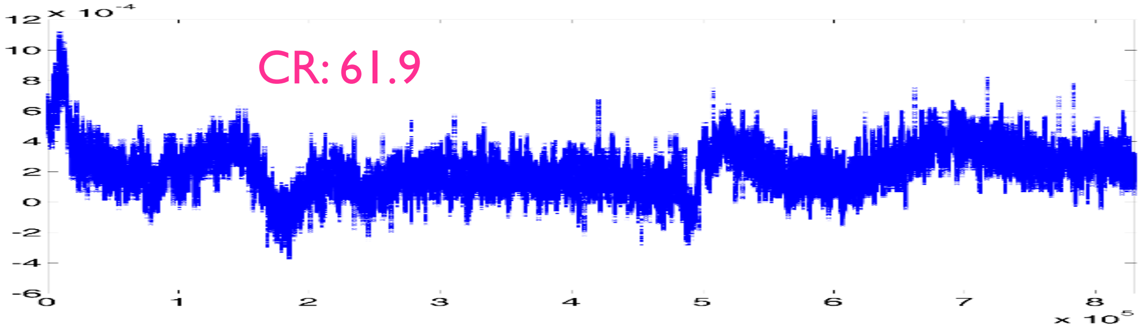


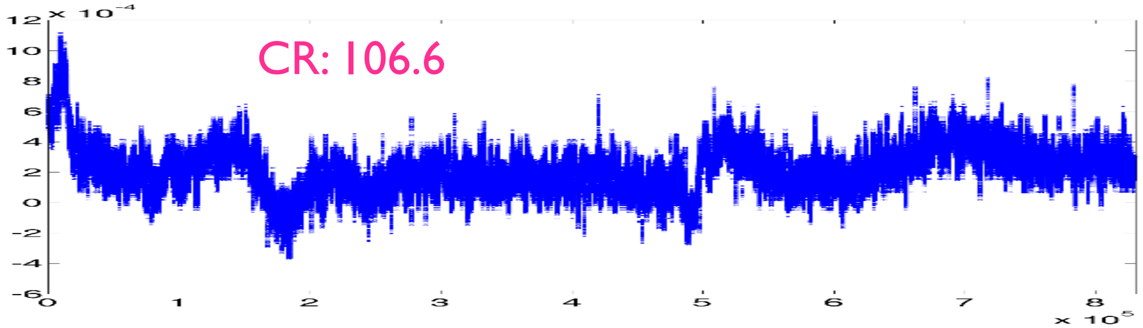
compression ratio (CR): original size/compressed size

# Examples on EEG data



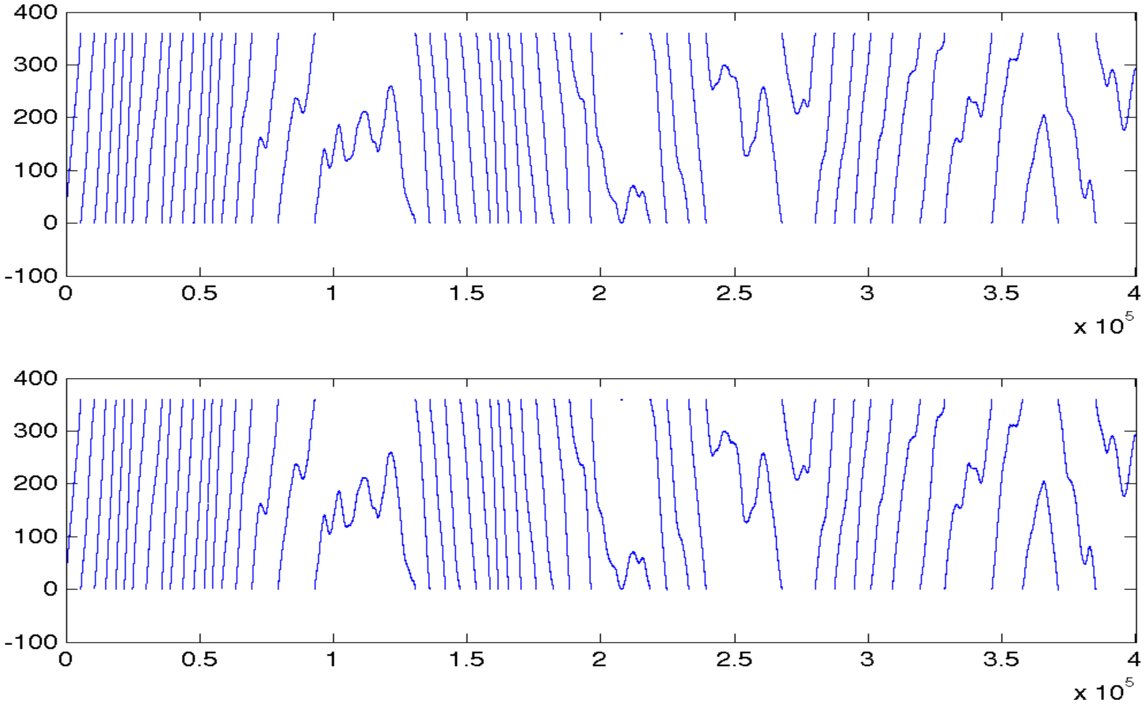



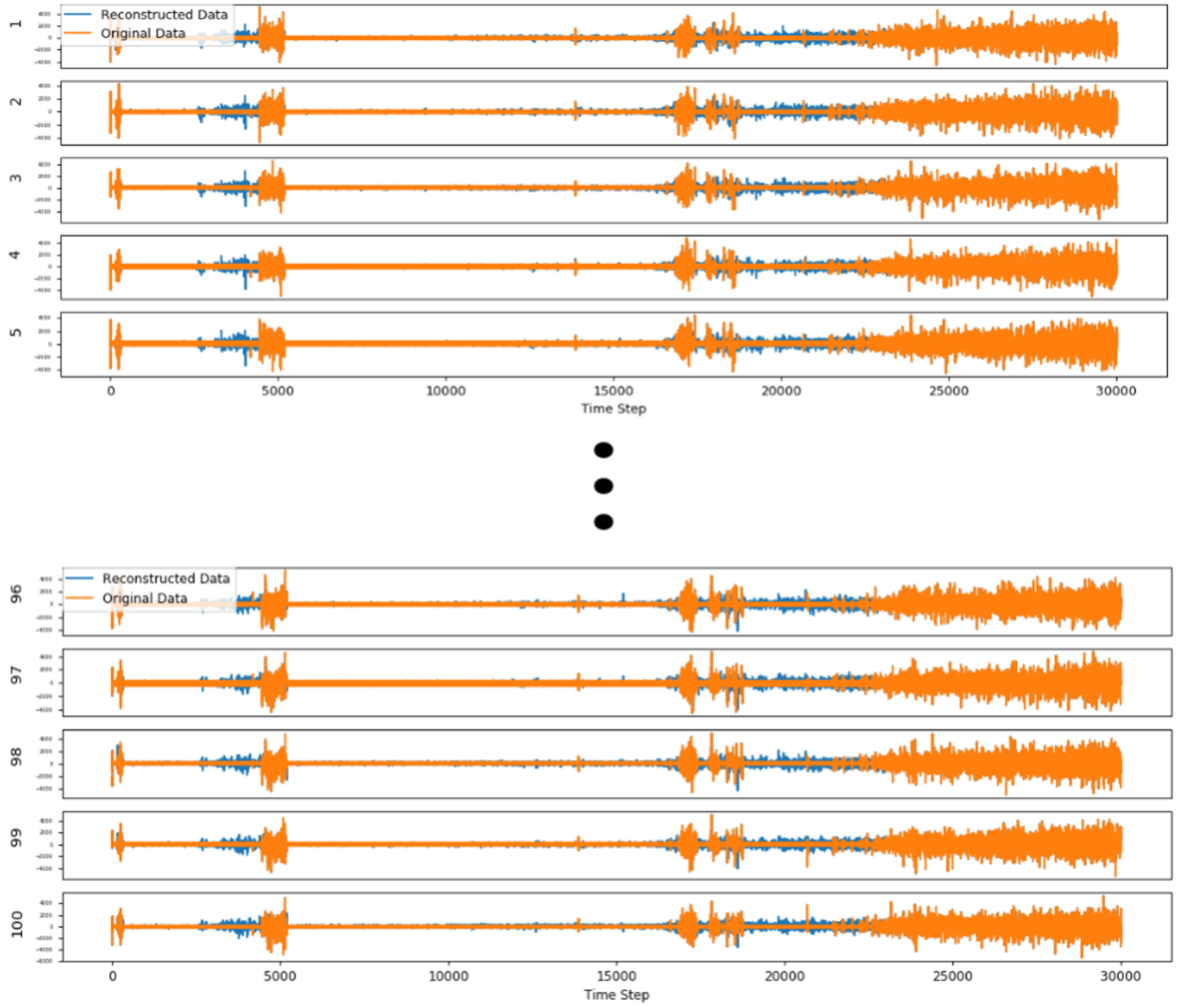




compression ratio (CR): original size/compressed size

# Examples on non-stationary data and multi-dimensional data



Phase angle μPMU data
A6BUS1C1ANG (Apr. 18~Apr. 29, 2015)   CR: 56.56




Distributed Acoustic Sensing dataset with 100 dimensions

# Examples on images

ORIGINAL PHOTO (2560 x 1440 = 3,686,400 pixels)
Each pixel has three dimensions of color (RGB)
(image courtesy by Nina Fox)

**Compression ratio = 7.71**

**Compression ratio = 19.61**

**Compression ratio = 57.65**

# Examples on video

# More application areas

- **Statistical analysis enables estimating future events in various applications. For example,**
  - Financial market analysis
  - Environmental study (e.g. extreme weather, climate change)
  - Energy usage analysis
  - Social network media analysis
  - Traffic analysis
  - System performance monitoring analysis

# Summary

- **IDEALEM is a new class of compression methods**
  - **Similarity-based Compression with Multidimensional Pattern Matching**
    - **Measures distance based on statistical similarity**
    - **Not traditional Euclidean distance ($L^2$-norm)**
- **IDEALEM can reduce data volume by more than 100-fold, while retaining key features from original data**
  - **Applicable to large, high frequency streaming data as well as large offline data archives**
  - **Provides accurate statistical analysis without loosing the underlying data distribution**
  - **A promising alternative to leading lossy compression algorithms**
  - **Applies to photos and videos, in addition to scientific multidimensional floating point data**
- **Fast enough execution time and small memory footprints to be used on resource limited devices for real time compression**