Multiscale Representation of High Dimensional Data

Mathematics of Data Analysis in Cybersecurity

ICERM

October 2014

Linda Ness

lness@appcomsci.com

Collaborators:

Devasis Bassu, Rauf Izmailov, David Shallcross – ACS

Peter Jones, Vladimir Rokhlin – Yale

Cybersecurity Research Appendix contributors: Dov Gordon, Giovanni di Crescenzo

Abstract

- Mathematical structure theorems from measure theory and linear algebra have proven to be useful for representing high-dimensional data. They have provided canonical domain-agnostic representations of multi-scale structures implicit in the data which can be compared , further analyzed and used in decision algorithms. I will illustrate this approach using data from several different types of systems including networked computing systems, an important example of cyber systems, and image systems, which are increasingly important for autonomous systems.
- This mulitscale representation research is a collaboration with Devasis Bassu, David Shallcross , and Rauf Izmailov (Applied Communication Sciences) and Peter Jones and Vladimir Rokhlin (Yale University).
- Relevance to cyber security: Cyber-security events occur when networked computing systems are successfully used for purposes not intended by their owners, administrators, designers, and vendors – e.g. for exfiltration of data, to enable control by unauthenticated users, to exploit unknown capabilities of the system. The design of the system components specifies the types of data that may be used, stored, and communicated by the system. Cyber-security events are difficult to detect using data analysis because the types of the data are unchanged but the instances of these types of data or the sequences of instances of these types of data are anomalous (and systems typically do not maintain an adequate description of normal usage). The instances of these collections of data types or sequences of these data types are high-dimensional and increasingly encrypted (so will be observable only as highly complex randomized structures). Mathematics provides a rich source of structures that can be exploited to provide canonical representations of more detailed structures in this high-dimensional data. Canonical representations will enable comparison of the structure of instances from different system locales and times. Additional domain knowledge will likely be required to determine the semantics of the anomaly – e.g. whether the anomaly is caused by changes in normal usage, planned evolution of the system, or more cyber events. Data is not only voluminous, but also complex in structure. Mathematics can help describe the complex structure.

Outline of Talk

- Introduction
- Relevance to Cybersecurity
- Multiscale Representations and Applications
  - Product Formula Representation
  - Heat Kernel Coordinate Representation
  - Multiscale SVD Representation
- Problems for Mathematical Cybersecurity Research
  - Some ACS Cybersecurity Research

Introduction

- Point of View:
  - Many data sets are sets of streaming sets of vectors
  - Their mathematical structure: (streaming) discrete metric-measure spaces
  - Automatically compute canonical multi-scale representations from the data
    - Compute representations guaranteed to exist by mathematical theorems
  - Canonical representations are domain agnostic so can be fused to provide a description of “what’s normal” – and hence “what’s anomalous”
  - Representations provide automatically generated canonical features for use in decision algorithms (e.g. machine learning)
- Approach: Algorithmize and apply mathematical representation theorems
  - AFOSR: Applications to Network Dynamics of Positive Measures and Product Formalisms: Analysis, Synthesis, Visualization and Missing Data Approximation
  - ONR: “Fast Multiscale Methods for Information Representation and Fusion”,
    - Randomized Singular Value Decomposition (RSVD), Randomized Approximate Nearest Neighbors (RANN) , Multiscale Singular Value Decomposition (MSVD), Multi-scale Heat Kernel Coordinates (HKC), Heat Kernel Function Estimation (HK-FE).
  - Both efforts were collaborations with Peter W. Jones and Vladimir Rokhlin (Yale)

Relevance to cyber-security

- Cyber-security events typically occur when networked computing systems are successfully used for purposes not intended by their owners, administrators, designers, and vendors, e.g.
  - exfiltration of data
  - control by unauthenticated users
  - exploit capabilities of the system not typically used and unknown to most
- System designs specifies the types of data in the system
  - Input, output
  - Stored internal state and data
- In cyber-security events the types of data are unchanged, but the instances of some of the data or of sequences of some of the data are anomalous
  - the spatio-temporal scale at which the event can be detected is unknown
- Systems currently do not maintain an adequate self-description of “normal”
  - Difficult: data is distributed, data is high-dimensional, voluminous, and complex
  - How to distinguish: new usage, system evolution, black swan events, cyber-attacks,

Product Formula Representation

6

The Product Formula

- **Theorem (F,K,P)**: A Borel probability measure μ on [0,1] has a unique representation as

$$\prod(1 + a_l h_l) ,$$

where the coefficients a_l are$\in$ [-1,+1]. Conversely, if we choose any sequence of coefficients $a_l \in$ [-1,+1], the resulting product is a Borel probability measure on [0,1].

Note: For general positive measures, just multiply by a constant. Similar result on $[0,1]^d$.

Note: See "The Theory of Weights and the Dirichlet Problem for Elliptic Equations" by R. Fefferman, C. Kenig, and J.Pipher (Annals of Math., 1991)

Additional Product Formula References

A similar dyadic analysis method was used to analyse various classes of weight functions on Euclidean space. One reference

- (P. Jones, J. Garnett) *BMO from dyadic BMO.* Pacific J. Math. (1982)
- Use dyadic representation, analyze functions/weights and average over translations of the dyadic grid.

Some Haar-like functions

"The Theory of Weights and the Dirichlet Problem for Elliptic Equations" by R. Fefferman, C. Kenig, and J.Pipher (Annals of Math., 1991). We first define the "L^∞ normalized Haar function" h_I for an interval I of form $[j2^{-n},(j+1)2^{-n}]$ to be of form

$h_I = +1$ on $[j2^{-n},(j+1/2)2^{-n})$

and

$h_I = -1$ on $[(j+1/2)2^{-n},(j+1)2^{-n})$.

The only exception to this rule is if the right hand endpoint of I is 1. Then we define

$h_I(1) = -1$.

Relative “Volume”

The coefficients a_I are computed simply by computing relative measure (“volume”) on the two halves of each interval I. Let L and R = left (resp. right) halves of I. Solve:

$$\mu(L) = \tfrac{1}{2}(1 + a_I)\,\mu(I)$$

$$\mu(R) = \tfrac{1}{2}(1 - a_I)\,\mu(I)$$

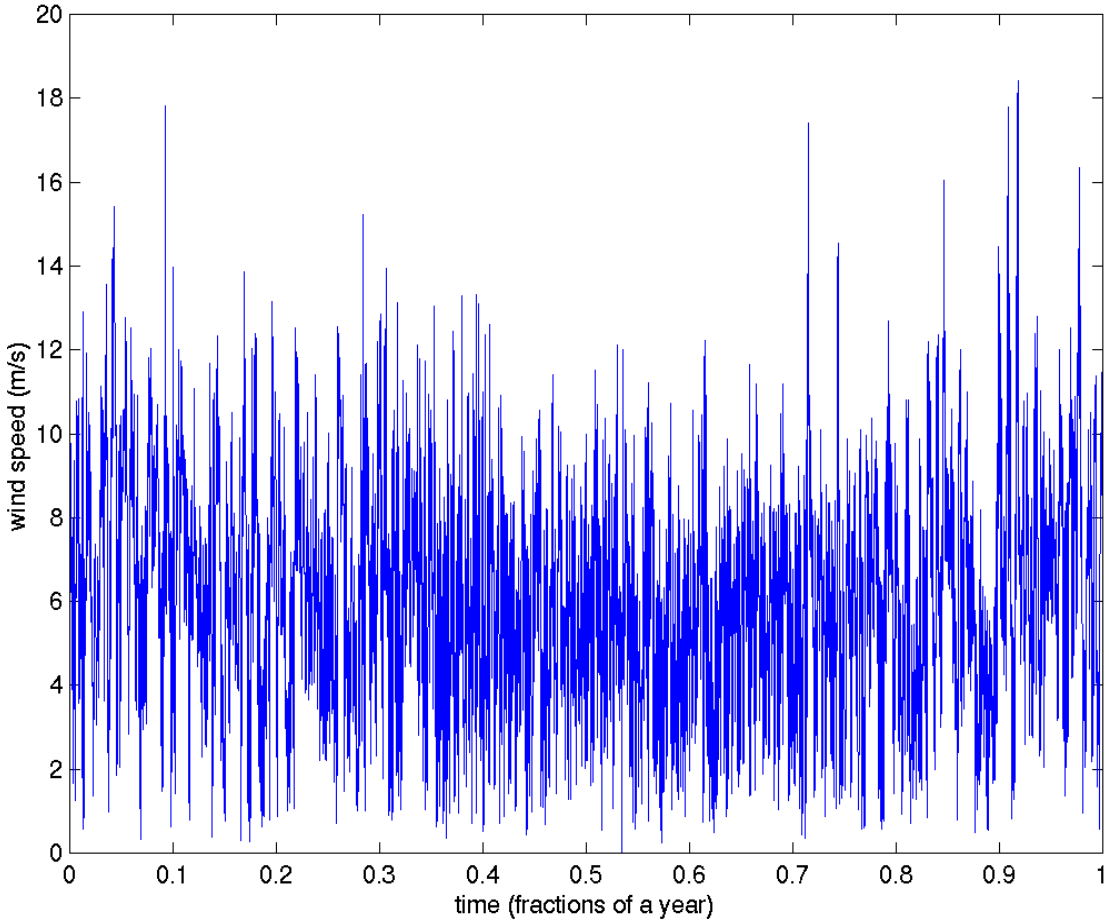
Then $-1 \le a_I \le +1$ because μ is nonnegative.

Why Use this Representation Instead of Wavelets?

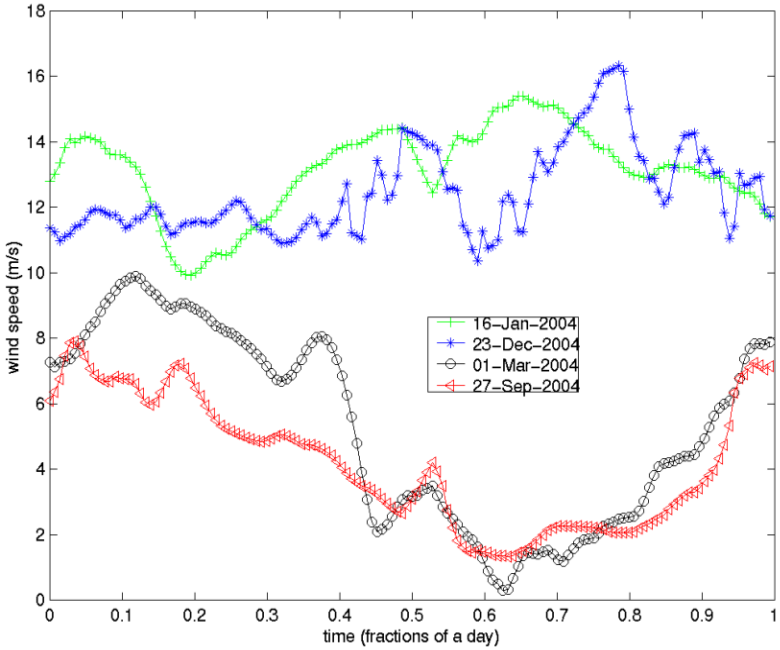
- The coefficients measure relative measure instead of measure.
- Wavelet approximations of positive functions are not necessarily positive
- All scales and locations are counted as “equal”.
- Can easily detect large changes in relative volume
- Relative coefficient representation is unique and normalized
- Can easily synthesize the measure
- Can randomly sample measures

Wind Data -

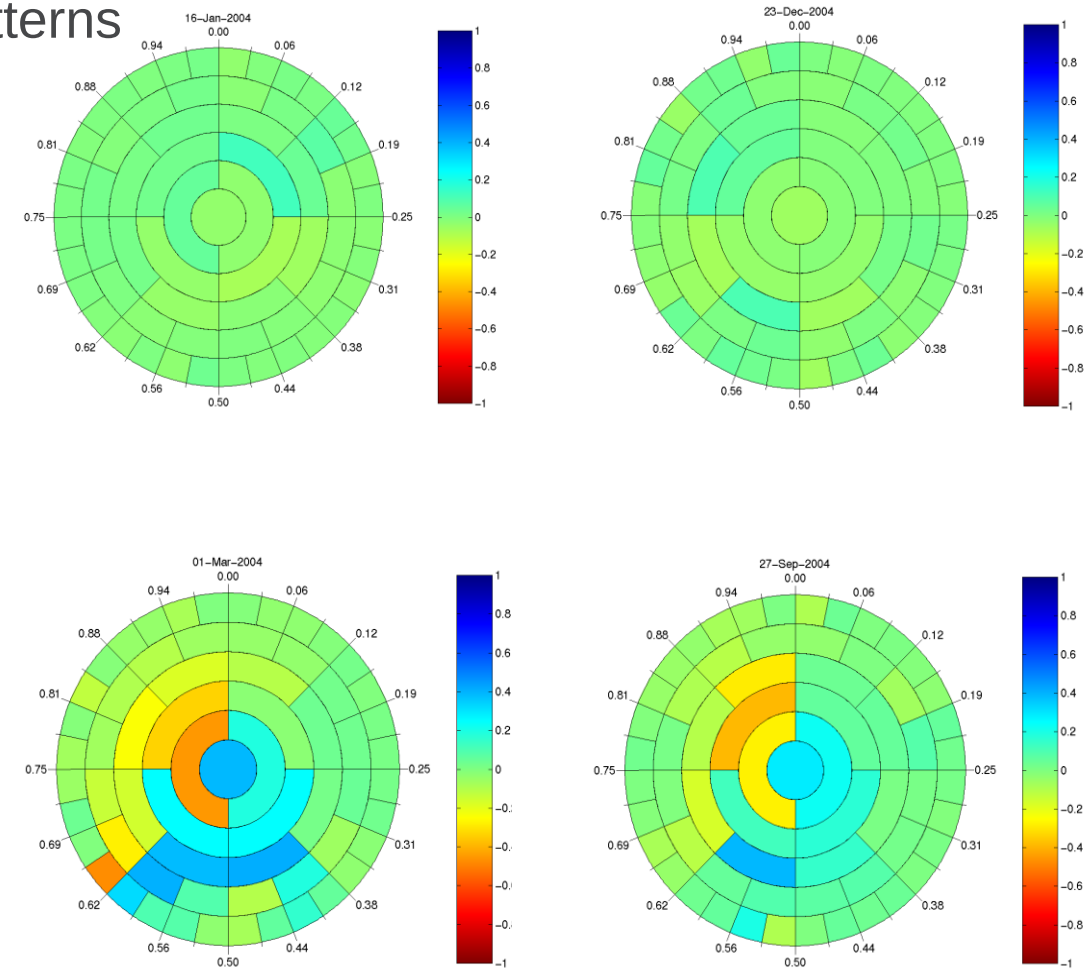
- Dataset from NREL gives wind speeds and potential wind turbine power levels for a large number of locations and elevations across the U.S., every 10 minutes for three years
  - "These wind power data ("Data") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy ("Alliance") for the U.S. Department Of Energy ("DOE"). "
- We looked at wind speeds for a single year, single location and elevation.
- We can compare the product coefficients for each day with the original time series.
- **Product coefficients provide normalized representations of multi-scale wind patterns**



Four Specific Days of Wind –



Product coefficients specify multi-scale patterns

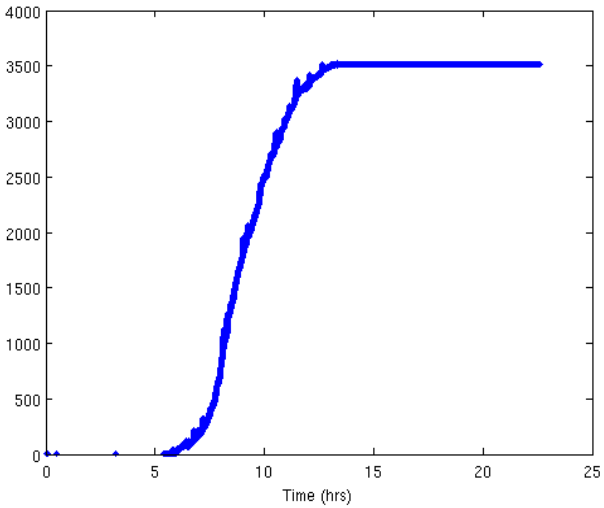


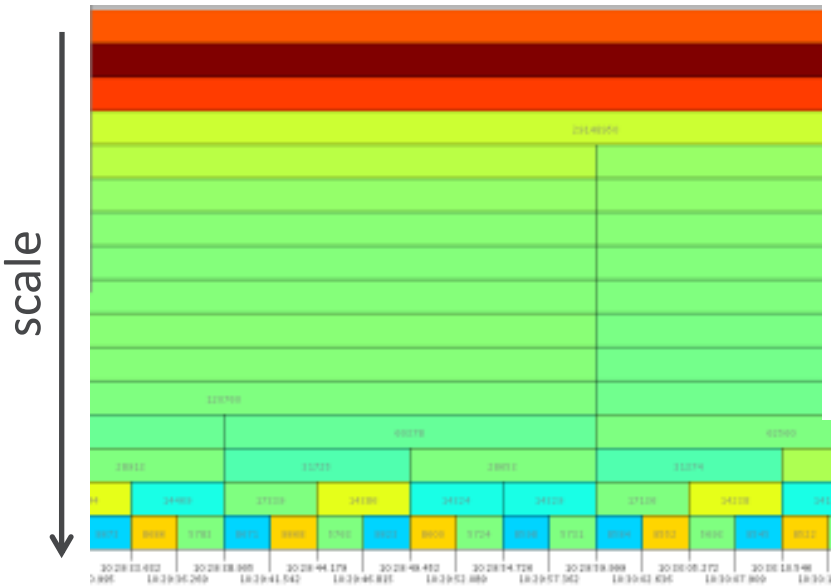
- The Jan 16 and Dec 23 wind patterns have relatively little variation, so product coefficients are small
- The Mar 1 and Sep 27 wind patterns have minimum in early afternoon,
  - so the scale 0 coefficients are > 0,
  - the first scale 1 coefficient is > 0,
  - second scale 1 coefficient is < 0.

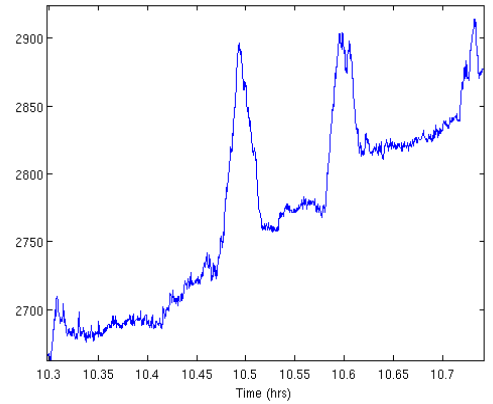
13

Product Coefficient Representation of Network Measurements

- Network measurements over the span of ~24 hours; ~65k sources
- Overall ramping behavior
- Inherent bursty behavior revealed at select scales







Note: Colors represents amount of function variation

14

Observation

- Product Coefficient Vectors provide a unique high-dimensional representation of relative volume in a dyadic decomposition of a unique cube
  - Unique for dimension 1
  - Unique for higher dimensions relative to a choice of Haar basis
  - Applicable to positive measures
- Issue:
  - The set of product coefficient vectors representing windowed measurements determines a high-dimensional empirical measure (data set).
    - Computing a product coefficient representation of this high-dimensional measure is theoretically possible but prohibitive due to complexity.
    - The intrinsic dimension of the measure is probably much lower than that of the ambient space.
    - It cannot be assumed that the support will lie on or near a manifold.
  - An additional automated method is needed to represent the structure in this and other high-dimensional measures.

Local Heat Kernel Coordinate Representation

16

Existence of Good Local Coordinate Systems

- Theorem (Jones, Maggioni, Schul)
  - For a domain D or manifold M of dimension d and volume = 1, the Laplace Eigenfunctions form a universal coordinate system on embedded balls B(x,r).
  - There exist exactly d eigenfunctions that blow up B(x,εr) to at least size 1 and with low distortion. The epsilon is universal
- Hence "the eigenfunctions find the manifold"
- The eigenfunctions need not be the first d eigenfunctions.
- Also - the eigenfunctions whose eigenvalues > c*(in-radius(M)) $^{-2}$ form a global chart and Weyl's theorem computes the number of these functions
- References:
  - P. Jones, M. Maggioni, R. Schul, "Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels". PNAS, vol. 105, no. 6 (2008), pp. 1803-1808.
  - P. Jones, M. Maggioni, R. Schul, "Universal Local Parametrizations via Heat Kernels and Eigenfunctions of the Laplacian" , to appear in Ann. Acad. Sci. Fennica
  - http://icerm.brown.edu/materials/Slides/sp-s14-w4/Eigenvectors,_Heat_Kernels,_and_Low_Dimensional_Representation_of_Data_Sets_%5D_Peter_Jones,_Yale_University.pdf

Heat Kernel Coordinates

- A different method of computing coordinates
- Justified by Jones corollary to Varadhan's Lemma
- Logs of ratios of heat kernels using d pairs of "anchor points"
  - In Euclidean space recover usual coordinates
- Use a small number of eigenfunctions to select d anchor points
- Provide coordinates on a ball of radius c*(in-radius), c universal (Jones)
- Ph.D. thesis by Michal Tryniecki, Yale University, 2013 (Hyperbolic Space)
- Advantage
  - automated computation, no manual selection of eigenvectors as for diffusion coordinates
- Validation experimentation on-going and publication in progress

Applications of Diffusion Coordinate Representations (Coordinates and Scales picked manually)

Diffusion Geometry Methodology Summary

The idea behind Diffusion Geometry is to put a new, nonlinearly defined representation of the data set. “Cartoon Version”:

Step 1. Embed the data set in $\mathbb{R}^d$.

Step 2. Choose a value of σ to define a length scale. Build the heat kernel

$W = (\exp\{ - |x_i - x_j|^2/\sigma\})$ and normalize to a random walk $D^{-1}W$, D = row sum

Step 3. Compute the Eigenvectors $\{\Phi_k\}$.

the right singular vectors

Step 4. Manually choose a small number of eigenfunctions, e.g. Φ_3 , Φ_4 , Φ_7 . The new data set representation is given by the image

$x_i \rightarrow \{\Phi_3(x_i), \Phi_4(x_i), \Phi_7(x_i)\}$

Step 5: Why do it? It could be helpful where PCA works poorly. It computes “local affinities” and builds coordinates from that information.

Mathematical Property: The Euclidean distance between two points in the full embedding is the averaged “random walk” distance between the original points. An explicit proof of this is given in [Coifman, Lafon 2006].

Diffusion Geometry References

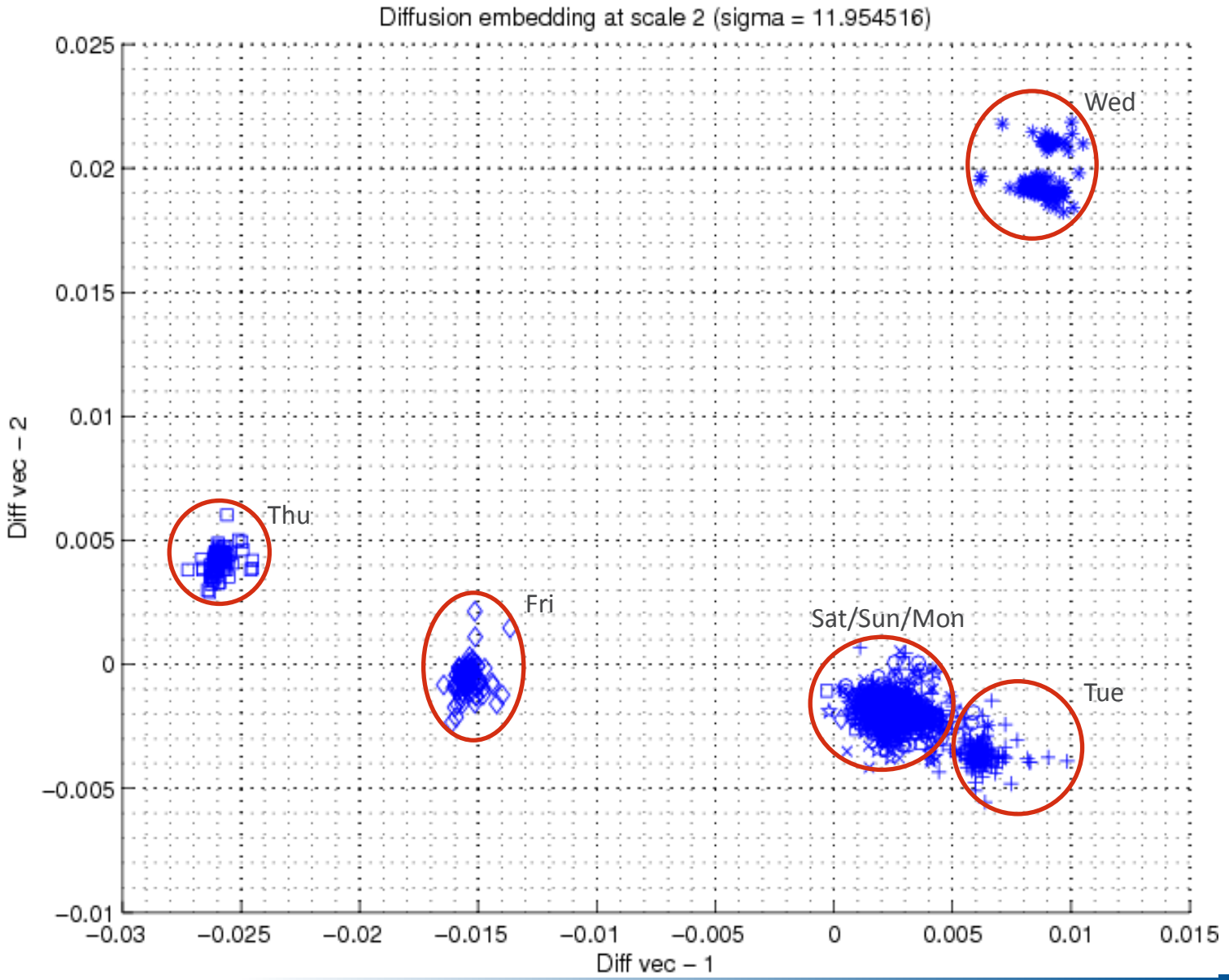
- Tutorial: Mauro Maggioni's Homepage. Click on "Diffusion Geometry".
- Why can it work? See:

  "Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels". (P. Jones, Mauro Maggioni and Raanan Schul)
  PNAS, vol. 105 no. 6 (2008), pages 1803-1808

- An Application: [Rohrdanz, Zhen, Maggioni, Clementi, "Determination of Reaction Coordinates Using Locally Scaled Diffusion Map" 2011] .

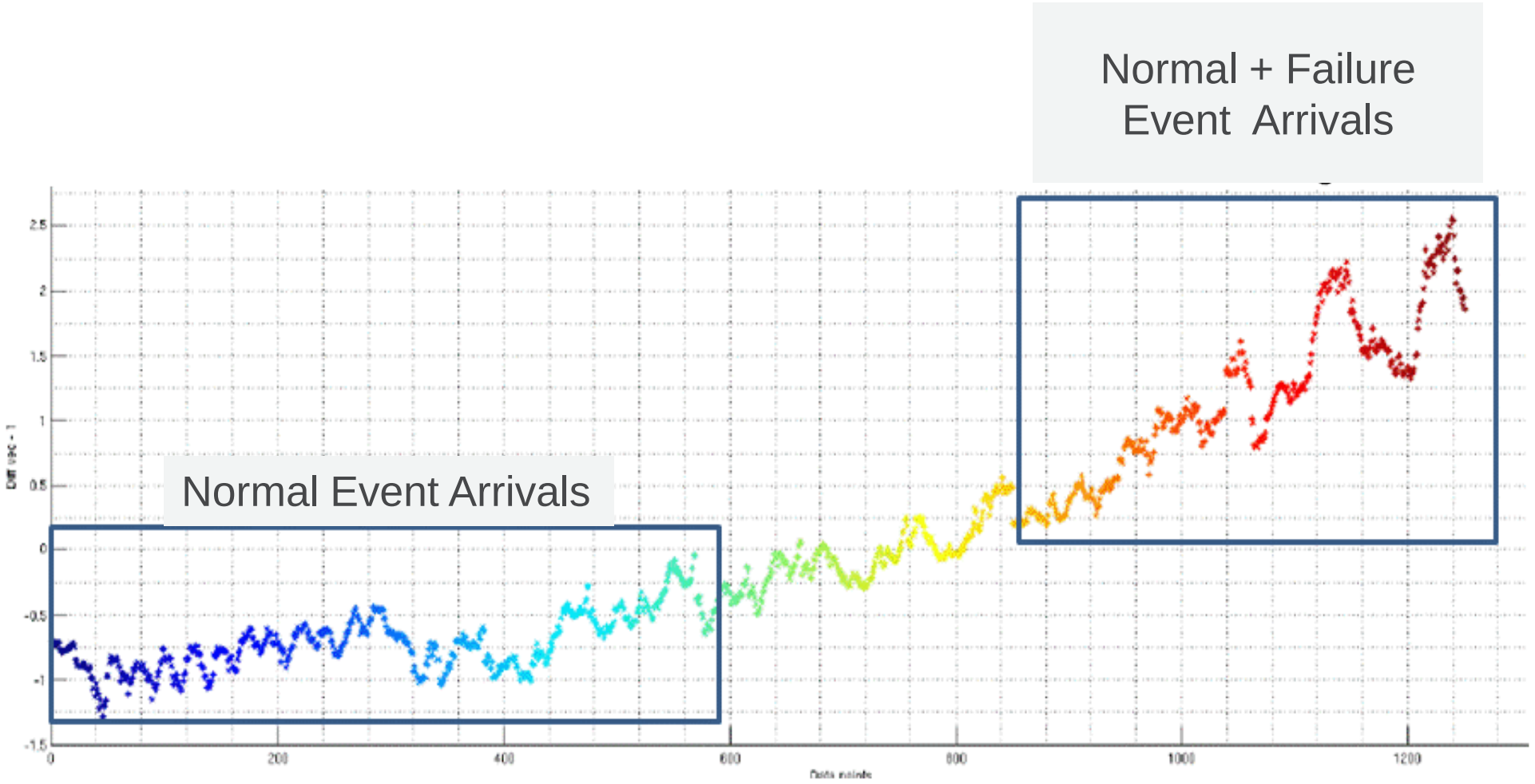
Low-Dimensional Model of Operational Regimes

- Daily profile
  - 182 antennas
  - 14 days
  - @ 15 mins
- Analysis
  - ~2550 points
  - 96 dim.
    a) PDPM
    b) Diffusion Map



22

Detection of Subtle Regime Change



Simulation Scenario proposed by Ron Skoog (Telcordia) to Model Actual Network Cascading Failure

Multiscale SVD Representation

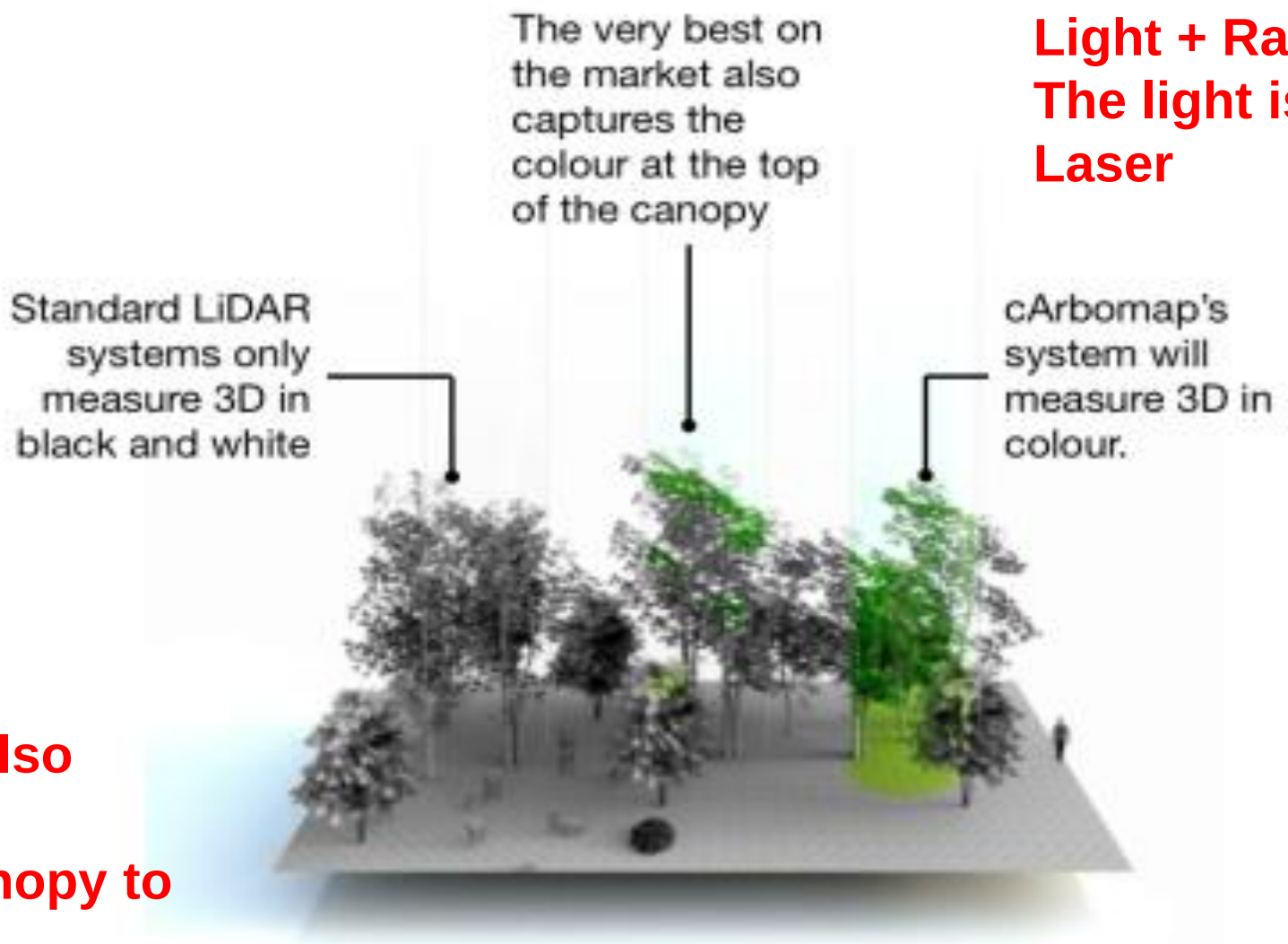
24

Multiscale SVD mathematics

- Approach – embed data using multi-scale SVD features
  - Exploiting geometric measure theoretic techniques developed by Jones
  - David Semmes – “Analysis of and on Uniformly Rectificable Sets” 1993
  - Jones - “Rectifiable Sets and the Traveling Salesman Problem”, 1990
- Topic of current research
- We are preparing a paper to illustrate the mathematical methodology in a series of examples.
- We have applied this to image data:
  - Multiscale SVD representation and heat kernel coordinates for LIDAR image
  - Used automatically generated MSVD features to improve decision algorithms (SVM classification)
  - Demonstrated improved classification by combining MSVD features and topology features on a LIDAR image and simple planar singularities (recent joint work with Bendich, Gasparovich, Harer)

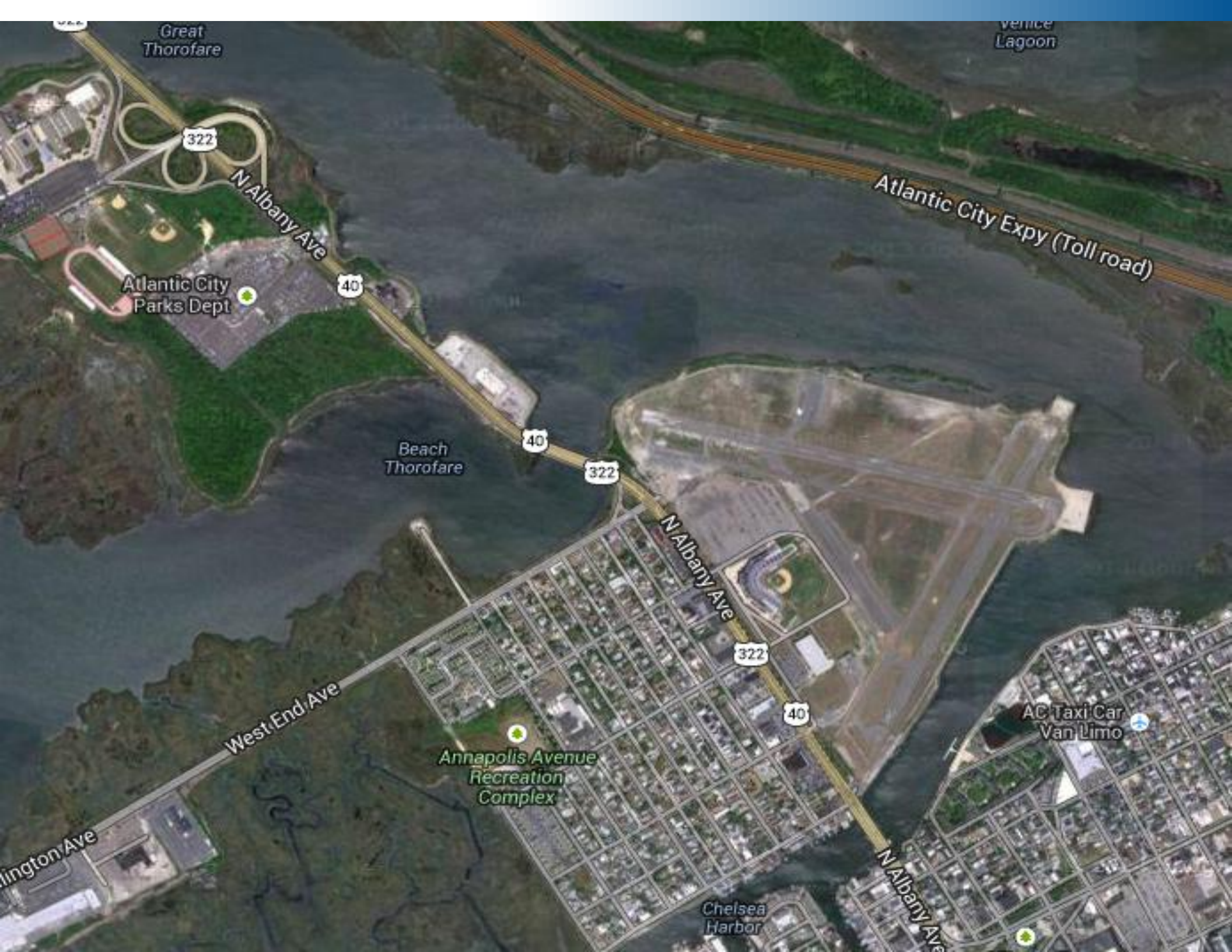
Multi-Spectral LIDAR

Portmanteau of Light + Radar
The light is from a Laser



This can also penetrate Forest Canopy to show structures below.

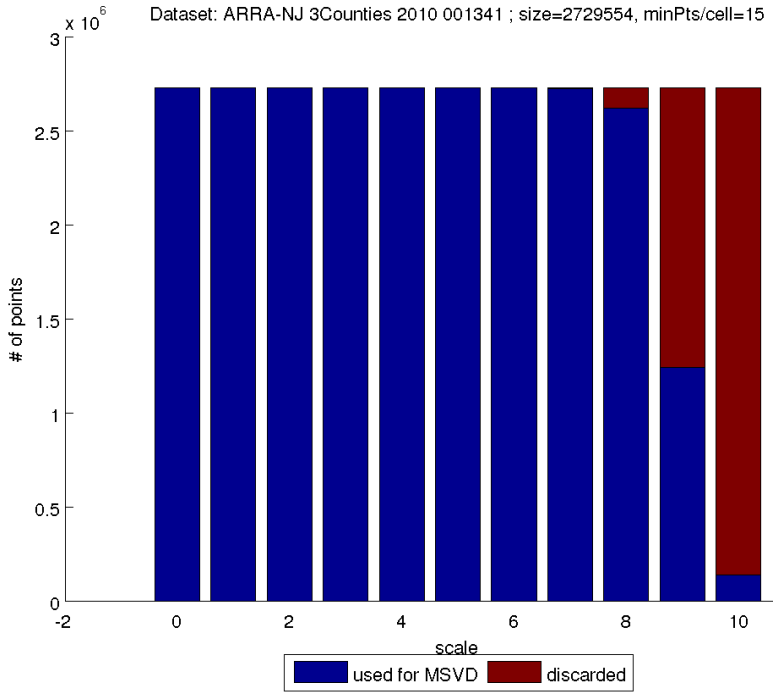
5


Great Thorofare
Venice Lagoon
322
N Albany Ave
Atlantic City Expy (Toll road)
Atlantic City Parks Dept
40
Beach Thorofare
322
40
N Albany Ave
322
40
West End Ave
AC Taxi Car Van Limo
Annapolis Avenue Recreation Complex
lington Ave
Chelsea Harbor
N Albany Ave

Ex.1: Former Atlantic City International Airport, NJ

An Example by D. Bassu

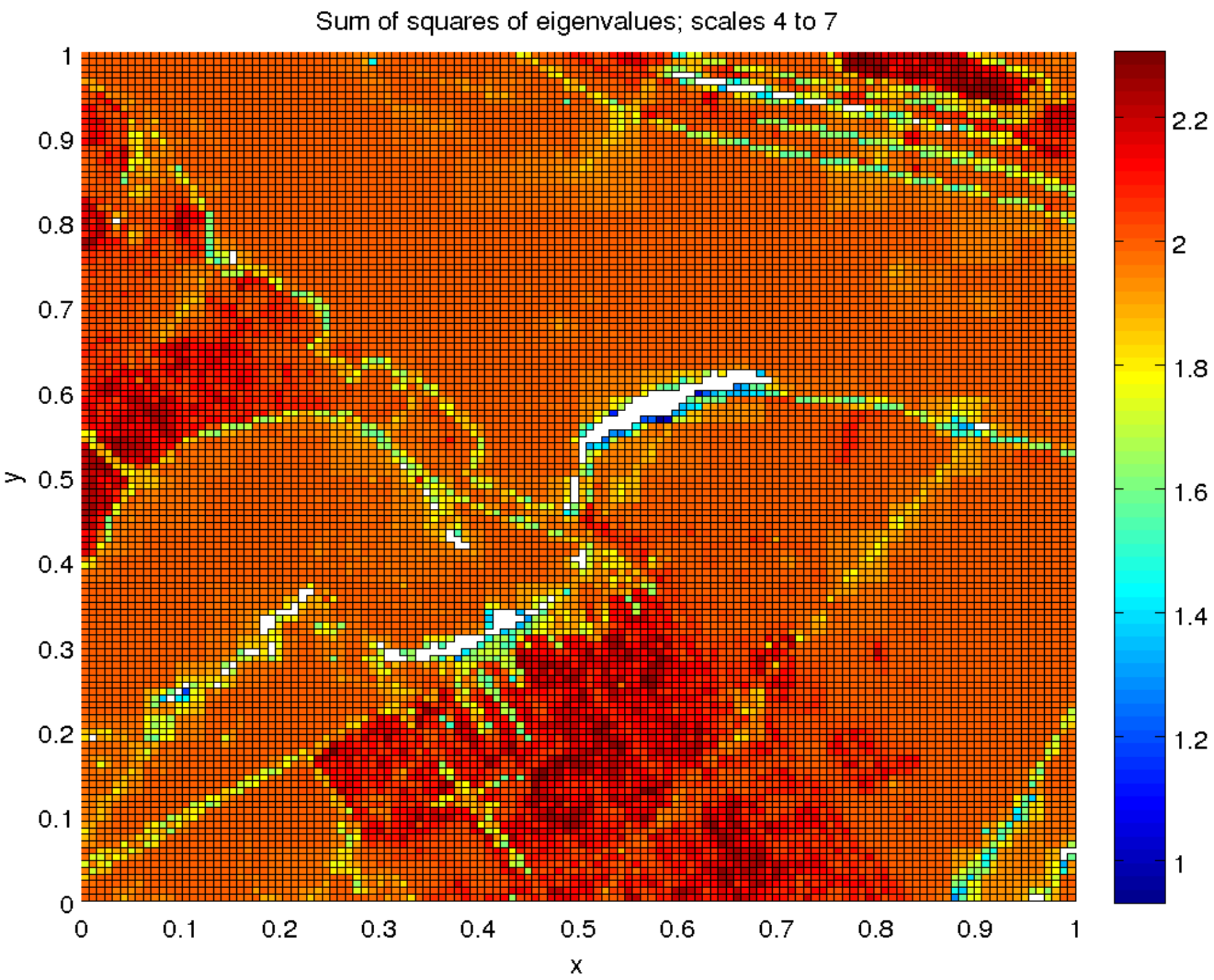
- LIDAR dataset comprising 2,729,554 points provided by USGS Earth Explorer tool (collected April 2010)
- Square region covering most of the former Atlantic City Airport, NJ with side length ~ 5,000 meters

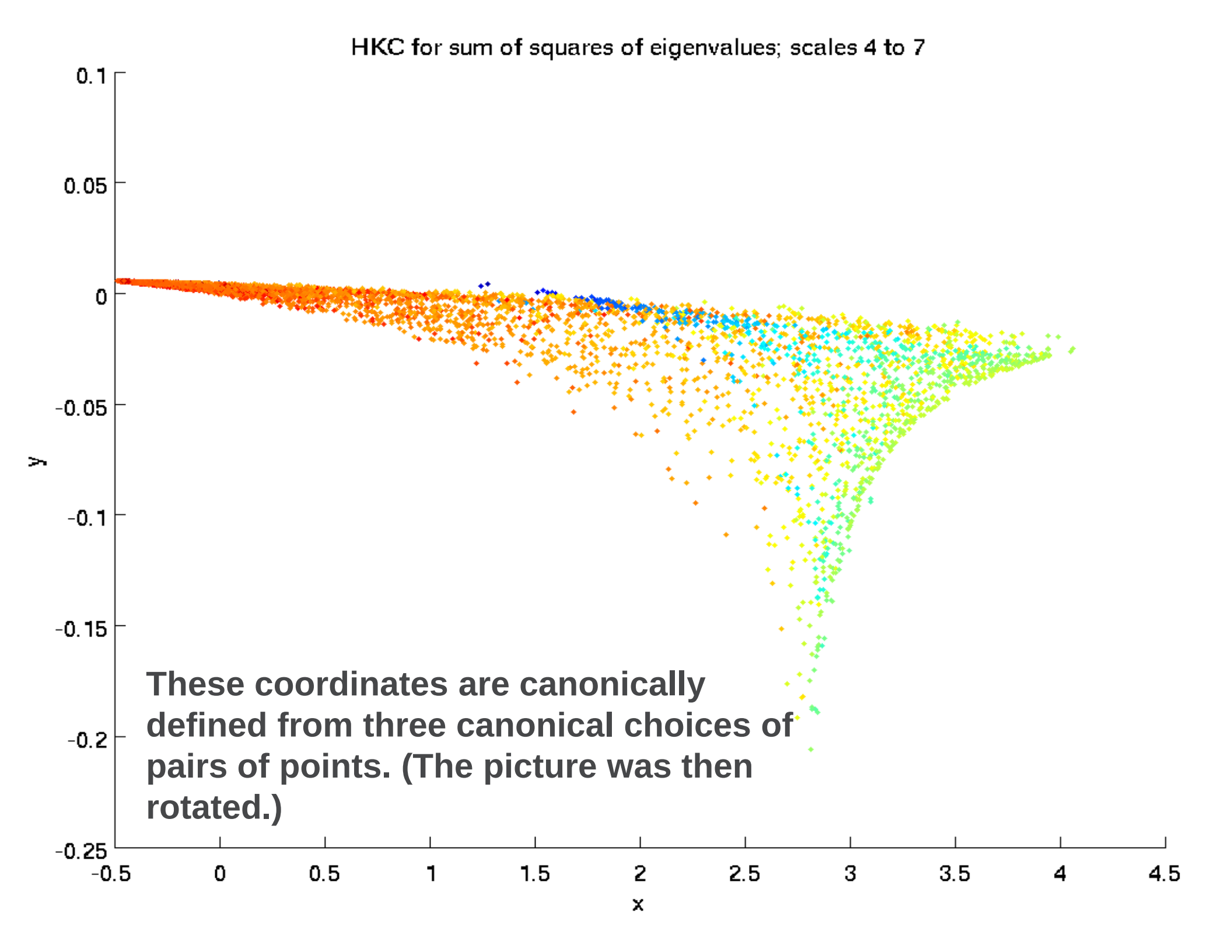


Points get isolated beyond 2^{-7} resolution (dyadic grid)

Multi-Scale SVD

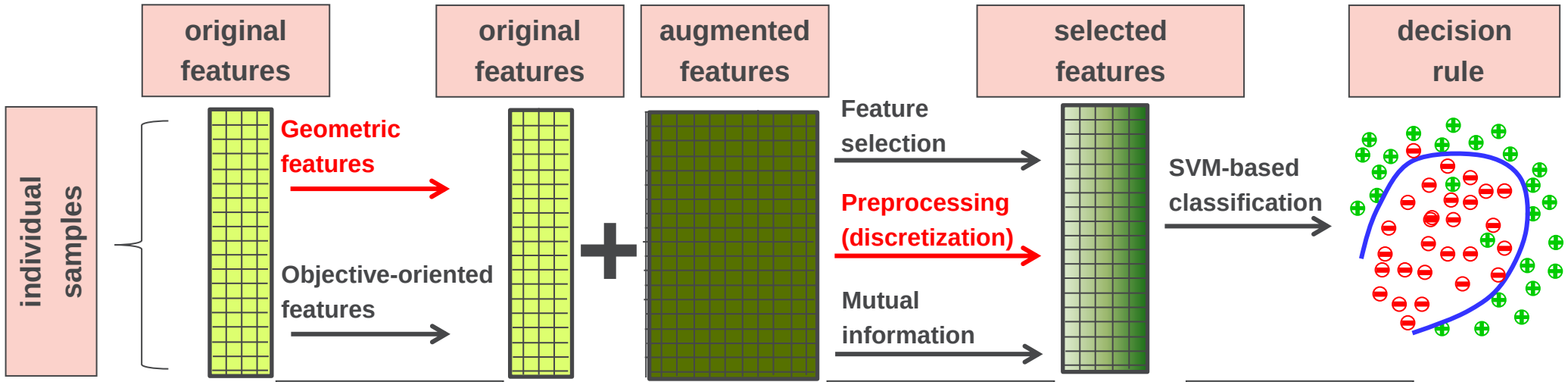
- Raw LIDAR data set is not 3 D
- Application to this data set
  - Compute SVD on four scales (for each point).
  - Square the top three (normalized!) eigenvalues
  - Feature vector V of length 12


Sum of squares of eigenvalues; scales 4 to 7
y
x
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
1.2
1.4
1.6
1.8
2
2.2


HKC for sum of squares of eigenvalues; scales 4 to 7
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
y
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
x
These coordinates are canonically defined from three canonical choices of pairs of points. (The picture was then rotated.)

Generating MSVD Features for Classification

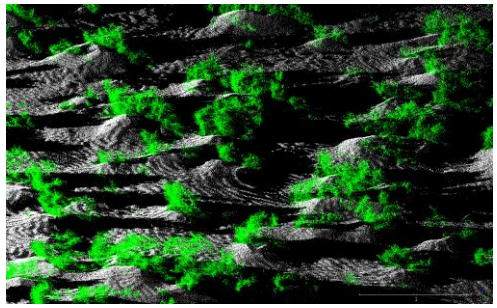
- Supported by ONR Program N00014-10-C-0176 “Fast Multiscale Algorithms for Information Representation and Fusion”.
- “Centralized Multi-Scale Singular Vector Decomposition for Feature Construction in LIDAR Image Classification Problems”, Applied Imagery Pattern Recognition Workshop 2012, D.Bassu, R.Izmailov, A.McIntosh, L.Ness, D.Shallcross
- General data analysis approach:
  - Augmenting original features (dimensions) with potentially relevant new features
  - Examining the whole space of features (original + augmented) to select the most relevant ones for the problem at hand
  - Preprocess the data to a more generalizable and robust form
  - Applying one of the best-in-class classification rules (SVM)



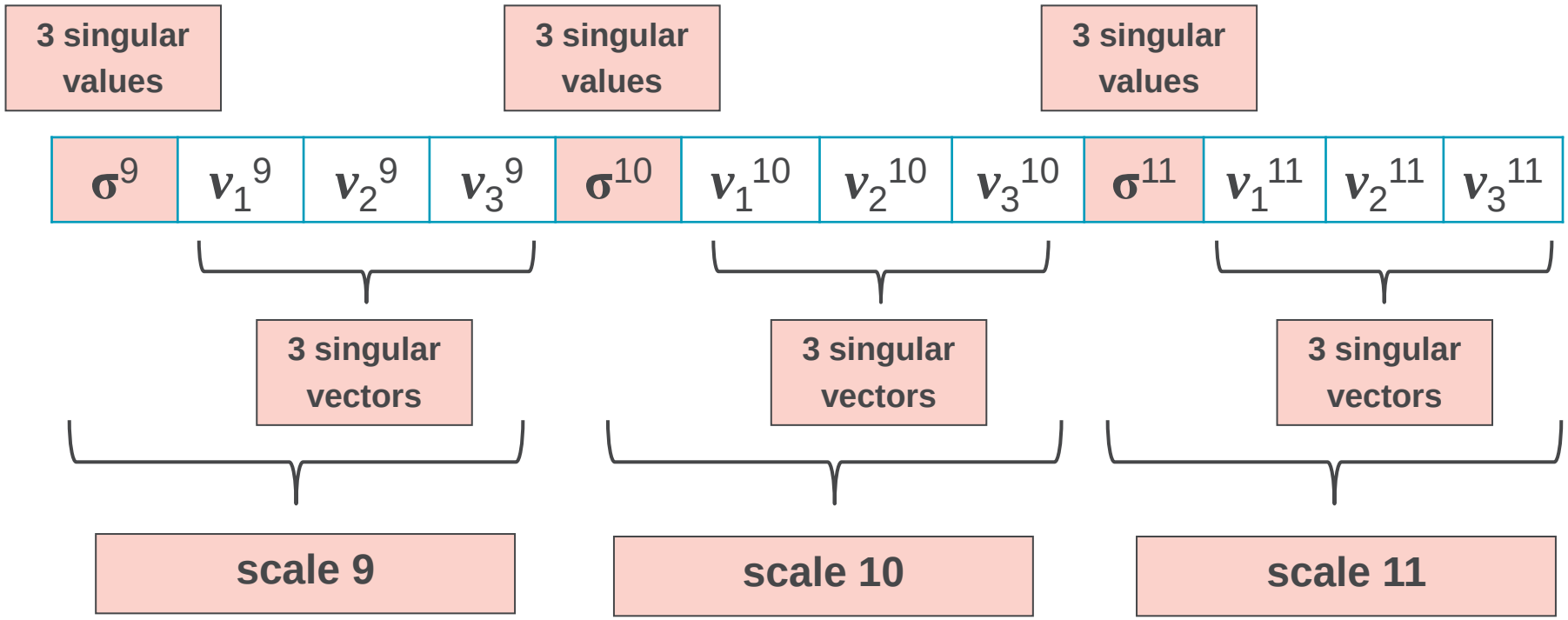
Augmented Features for LIDAR Dataset

- Original features (3 dimensions)

x	y	z



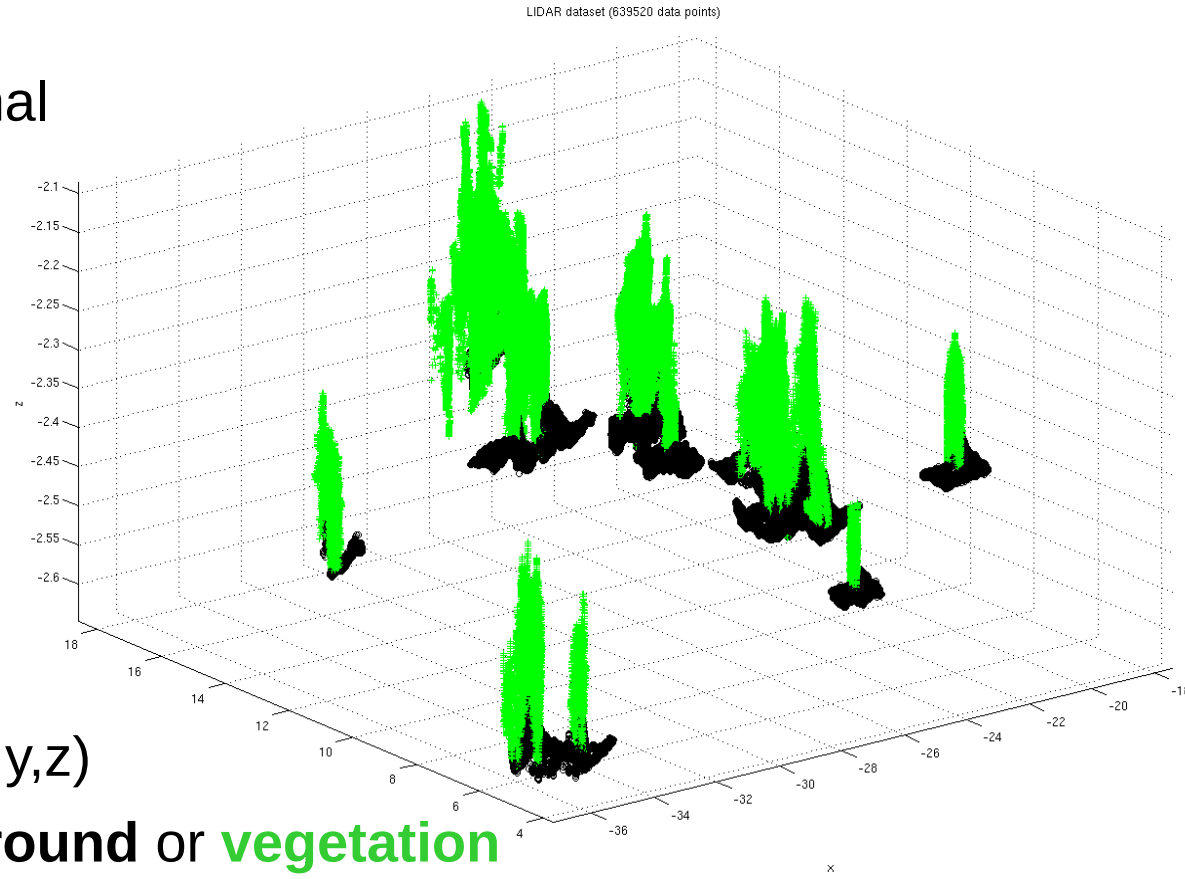
- Augmented features (additional 36 dimensions)



- Only scales 9,10,11 were used: others were not informative (either contained the whole dataset (full), or mostly empty)

LIDAR Dataset Classification

- Problem: classify 3-dimensional LIDAR data into two classes:
  - **ground**
  - **vegetation**
- Details:
  - Given 10 clouds of points covering different areas
  - Total 639,520 of 3-dimensional samples (x,y,z)
  - Each sample is labeled **ground** or **vegetation**



- Classification metric 1: minimum of type I (misclassify **vegetation** as **ground**) and type II (misclassify **ground** as **vegetation**) error rates
- Classification metric 2: minimum of *sensitivity (100% minus type I error rate)* and *specificity (100 % minus type II error rate)*
- Classification metric 1 of 5% reported by reference paper of Brodu and Lague.

LIDAR Dataset Classification Results

Preprocessing		coordinates (x,y,z)		MSVD scales used				Classification Metrics		
Continuous (just scaling)	10-discretization	included	excluded	9	10	11	9,10,11	Sensitivity %	Specificity %	Error %
Y		Y		Y				98	93	7
	Y	Y		Y				98	96	4
Y			Y	Y				95	95	5
	Y		Y	Y				95	95	5
Y		Y			Y			98	93	7
	Y	Y			Y			95	98	5
Y			Y		Y			95	98	5
	Y		Y		Y			92	97	8
Y		Y				Y		93	90	10
	Y	Y				Y		86	96	14
Y			Y			Y		97	80	20
	Y		Y			Y		84	77	23
Y		Y					Y	97	97	3
	Y	**Y**					**Y**	**98**	**98**	**2**
Y			Y				Y	97	98	3
	Y		Y				Y	97	97	3

- Extensive experiments were carried out:
  - Including and excluding original features ((i.e., x,y,z) LIDAR coordinates)
  - Various combinations of MSVD-augmented features (scales 9,10,11)
  - With and without discretization preprocessing
- Results show that:
  - Combination of several localized scales with original coordinates gives best performance
  - 10-discretization usually improves classification performance
  - The proposed approach reduces classification error to 2% (from 5% in the original study)

Some Current Research

- Recently, we have experimented with enlarging the feature set to include topological features (relative persistent homology) in addition to non-normalized MSVD features.
  - Bendich, Gasparovic, Harer, Izmailov, Ness, “Multi-Scale Local Shape analysis and Feature Selection in Machine Learning Applications”
  - Sampling result showed that adding relative persistent local homology features to MSVD features improved the classification results on the LIDAR data sets

Suggestions for Mathematical Security Research

- Characterize what can and cannot be observed using mathematical features defined on observed data from dynamically evolving networked computing systems - even when some of the observed data is encrypted
- A Cryptography inspired Science of Cyber-Security (Giovanni di Crescenzo)

Some ACS Theory-Based Security Research

Secure Multi-party Computation:

2 or more parties with private input **interact** to compute a function or program on their inputs. Provably, the result of the computation reveals nothing but the output.

- Multi-Party Computation of Polynomials and Branching Program without Simultaneous Interaction, Gordon et al., Eurocrypt 2013.
  Here the parties interact with a central server, and do not need to talk to one another. We still require the security against a malicious, colluding server.
- Secure two-party computation in sublinear (amortized) time. Gordon et al., ACM Conference on Computer and Communications Security 2012.
  This work moves the computation from the circuit model to a RAM model of computation, using a primitive called **Oblivious RAM**. ORAM hides access patterns to memory, while avoiding the trivial solution of a full linear scan on the memory.

Some ACS Theory-Based Security Research

Functional Encryption

Decryption keys restrict a user to learning some **function** of the plaintext. This allows non-interactive computation on encrypted data.

- Multi-Input Functional Encryption, Gordon et al., Eurocrypt 2014
  This enables the holder of the key to compute on multiple ciphertexts, even when generated by different sources. Among other applications, this gives the first fully-secure solution to the problem of order-preserving encryption, allowing a server to sort and search encrypted data.

Program Obfuscation

A provable method for hiding code, allowing a user to execute a circuit on arbitrary inputs, while learning nothing about the circuit beyond the corresponding output values.

- A General Approach to Withstanding Leakage on Key Update, Dachman-Soled et al., in submission.
  This paper demonstrates a connection between program obfuscation and leakage-resilient cryptography, where security holds even when the adversary can learn functions of the secret key.
- On the Relationship between Functional Encryption, Obfuscation, and Fully Homomorphic Encryption. Alwen et al., IMA Int. Conf on Cryptography and Coding, 2013.
  This work explores the theoretical connections between the mentioned primitives.

Suggested Research Topic: A Cryptography-Inspired Science of Cybersecurity

Giovanni di Crescenzo

A science of cybersecurity

- A science of cybersecurity should allow us to:
  - Answer natural questions about the security of a system
    - E.g.: How much do I have to spend, in terms of (in)efficiency parameters A,B,C,etc. to obtain given levels X,Y,Z,etc. of security?
  - Define rigorous modeling (and solution) approaches
    - Each of these approaches has its own way to define a model of the system, system threats, attacker resources and methodology, solutions methods and costs, and when the system is declared secure (wrt to such threats, attackers, solutions)
  - Rigorously specify security solutions within these models
  - Rigorously prove solutions to be secure within the model
  - Validate solutions costs and assumptions from concrete data collected from real system behavior and newly discovered vulnerabilities and refine modeling based on these
  - Building security-preserving solution architectures from solution components and their interaction

Why a cryptography-inspired science of cybersecurity?

- After years of investigations, modern cryptography has transitioned from its early years mostly made of breakthrough ideas (e.g., the RSA cryptosystem and the Diffie-Hellman key agreement protocol), into a scientific discipline with rigorous models, security notions and theorems about practical solutions to real-life problems (e.g., producing statements of the type "if modeling assumptions X are true then the studied system satisfies security notion Y", for rigorously formulated X and Y).
- Today, as more cybersecurity solutions get proposed, we still lack models and theories to rigorously evaluate, compare and improve their properties and practical impact, a state of affair similar to the early years of modern cryptography.
- Our vision is that cybersecurity needs a transition process and technical treatment similar to what already happened to cryptography.

The cryptography-inspired approach

- In line with opinions such as “The field of cryptography comes close to exemplifying the kind of science base we seek. [S12]”, we propose to expand the scientific approach of modern cryptography into a science of cybersecurity.
- Our focus area is detection of attacks to a given system. Our goal is to produce rigorous models, theories and solutions that allow us to answer natural cybersecurity questions about a given system; i.e., finding multi-dimensional tradeoffs between system assumptions, attacker resources, detected attacks, solution correctness and solution performance of both known and new solutions. Our cryptographic modeling approach, in a nutshell, consists of a suitable applications of the following 3 basic requirement principles to detection areas:
  - 1) representation expressivity (i.e., rigorously formulated requirements demanding that the chosen mathematical representation for system assumptions, attacker resources and attacks, is general enough to capture a large class of attacks to the system);
  - 2) conditional detection correctness (i.e., rigorously formulated requirements demanding that given a sufficiently expressive representation, it is possible to detect of a large class of attacks and/or attacker features); and
  - 3) detection performance (i.e., rigorously formulated requirements demanding that detection of attacks and/or attacker features is performed with practically acceptable computing resources).
- We see applicability to cybersecurity areas: Intrusion detection, APT detection, digital forensics, IP traceback, Botnet topology detection, etc.

Powerhouse Research. Practical Solutions.