Efficient Data Reduction and Summarization

Ping Li

Department of Statistical Science

Faculty of Computing and Information Science

Cornell University

September 16, 2008

Research Goals

1. Developing fundamental algorithms for processing massive data
   - Massive high-dimensional data
   - High-speed dynamic streaming data
   - Massive sparse data
2. Applying these algorithms to data analytics and visualization
   - Scalable visualization algorithms.
   - Scalable Machine Learning algorithms.
   - Real-time network flow measurement algorithms.
3. Training Graduate/Undergraduate students

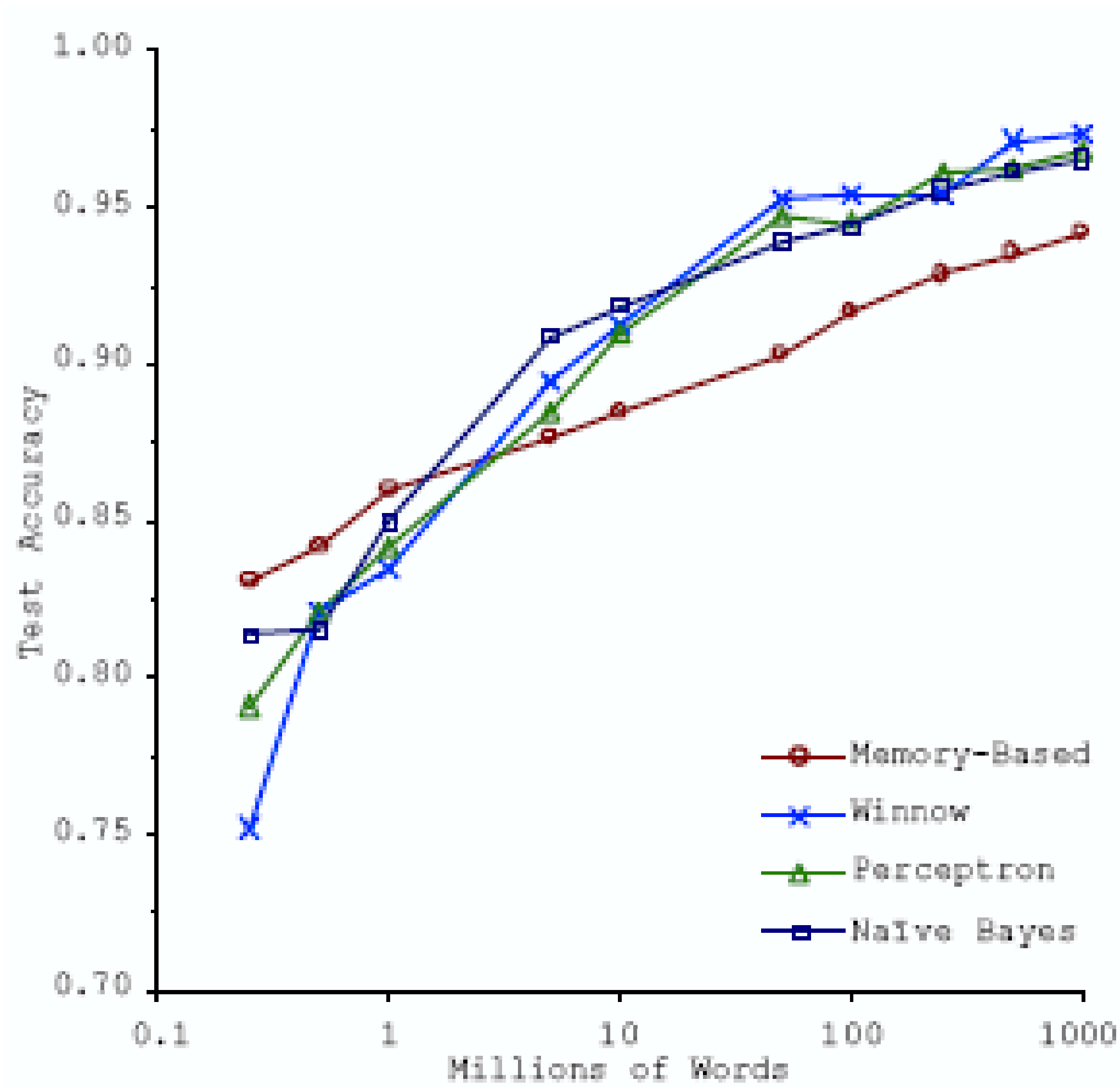
Some of the research goals were not included in the original proposal.

Additional funding will be sought from other sources.

The Era of Modern Massive Data

- *There is no data like more data* (Mercer at Arden House, 1985)

- *More data is more important than better algorithms* (Banko & Brill, ACL 2001)



1.00
0.95
0.90
0.85
0.80
0.75
0.70
Test Accuracy
0.1
1
10
100
1000
Millions of Words
Memory-Based
Winnow
Perceptron
Naïve Bayes

Workshop on Algorithms for Modern Massive Data Sets

Highly successful workshop Funded by NSF and Yahoo!

- MMDS 2006, June, Stanford University

  Ping Li, Trevor Hastie,

  Efficient L2 and L1 dimension reduction in massive databases

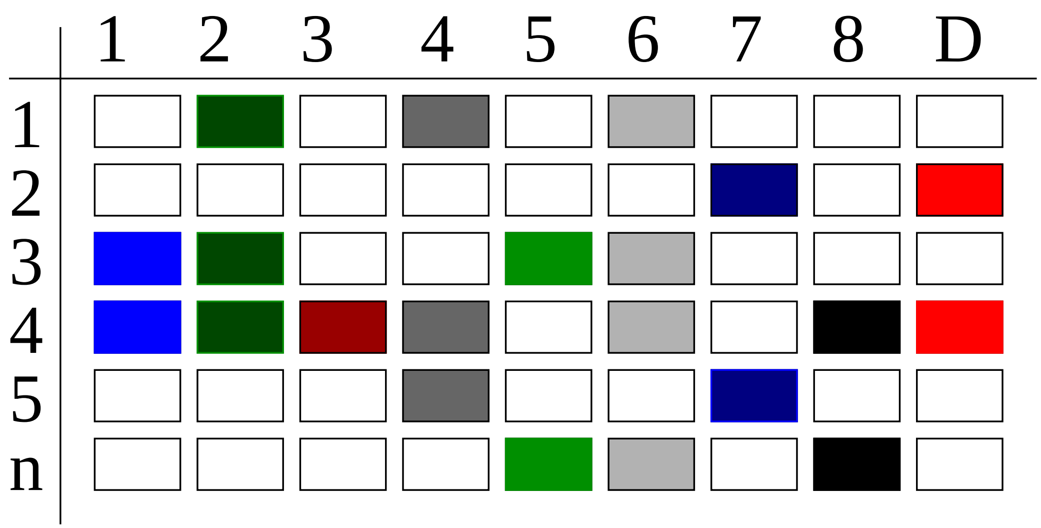
- MMDS 2008, June, Stanford University

  Ping Li,

  Compressed Counting and Stable Random Projections

The Data Matrix

Data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$: n rows and D columns.



Examples: Term-doc matrix, Image-pixel matrix, etc.

Characteristics of Modern Data Matrix

- Massive eg, both n, $D \approx 10^{10}$
- Dynamic eg, high-speed data streams
- Often Sparse eg, text data

Massive Data Summarization and Some Challenges

Summarization is fundamental in learning, visualization, and linear algebra.

- Summary statistics of individual rows (or columns)

  eg, αth moment $\sum_{i=1}^{D} |u_i|^\alpha$, entropy, etc.

- Summary statistics between rows (or columns)

  eg, dot products, αth distance $\sum_{i=1}^{D} |u_i - v_i|^\alpha$, χ^2 distance, etc.

Some challenges

- Memory intensive Loading $\mathbf{A} \in \mathbb{R}^{n\times D}$ may be infeasible.

  Loading all pairwise (eg, n^2) distances of $\mathbf{A}$ can be easily infeasible.

- CPU intensive

- Dynamic updating

From Exact Answers to Approximations

(Good) Approximate summary statistics (eg distances) often suffice

- Visualization systems only need a certain resolution.
- Good (robust) algorithms are stable even using approximate inputs.

Simple random sampling (eg using a few columns) is not enough

- Not accurate.
- Not suitable for sparse data.

Three Basic Approximation Techniques

1. Symmetric Stable Random Projections

   Computing αth distances ($0 < \alpha \leq 2$) of data matrix.

   $\alpha = 2$: Euclidean distance. $\alpha = 0$: Hamming distance.

   Applicable to dynamic streaming data in Turnstile model.

2. Compressed Counting (CC) (Skewed Stable Random Projections)

   Computing αth moments ($0 < \alpha \leq 2$) of data stream in strict-Turnstile model.

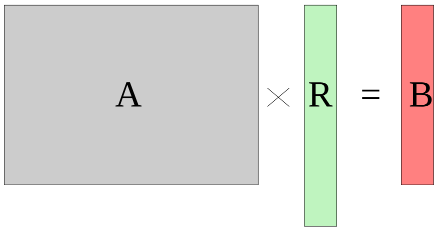
3. Conditional Random Sampling (CRS) One-sketch-for-all

   Computing any type of distances or moments

   Applicable to more general dynamic data

   Only works well in sparse data

Symmetric Stable Random Projections



- Original data matrix $\mathbf{A} \in \mathbb{R}^{n\times D}$: n rows and D columns,
  Massive, eg, both n, $D = O\left(10^{10}\right)$.
  Possibly dynamic, according to the Turnstile model.

- Projection matrix $\mathbf{R} \in \mathbb{R}^{D\times k}$: D rows and k columns, $k \ll n, D$
  Entries are samples of a symmetric α-stable distribution.
  $\alpha = 2$: Normal distribution. $\alpha = 1$: Cauchy distribution.

- Projected matrix $\mathbf{B} \in \mathbb{R}^{n\times k}$: n rows and k columns
  Viewed as a sketch of $\mathbf{A}$, which may be discarded.

Symmetric α-Stable Distributions

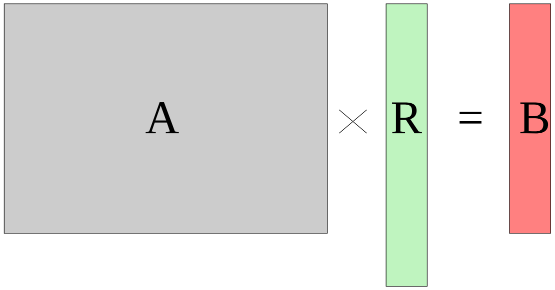
Denoted by $S(\alpha, d)$, where $0 < \alpha \leq 2$.

Two random variables $Z_1 \sim S(\alpha, 1)$ and $Z_2 \sim S(\alpha, 1)$.

For any constants C_1 and C_2

$$Z = C_1 \times Z_1 + C_2 \times Z_2 \sim S\left(\alpha, |C_1|^\alpha + |C_2|^\alpha\right)$$

For example, weighted sum of normals is also normal ($\alpha = 2$).



Therefore, the projected matrix $\mathbf{B}$ contains information about

1. αth moment, $\sum_{i=1}^{D} |u_i|^\alpha$, of each row of $\mathbf{A}$.
2. αth distance, $\sum_{i=1}^{D} |u_i - v_i|^\alpha$, between any two rows of $\mathbf{A}$.

Applications of Symmetric Stable Random Projections

- Data visualization algorithms

  Multi-dimensional scaling (MDS) requires a pairwise similarity matrix.

- Machine Learning algorithms

  SVM (support vector machine) requires a $O(n^2)$ pairwise distance matrix.

- Information retrieval

  Finding (filtering) nearly duplicate docs (often measured by distance)

- Databases

  Estimating join sizes (dot products) for optimizing query execution.

- Dynamic data stream computations

  Estimating summary statistics for visualizing/detecting anomaly real-time

An incomplete list of references:

- Vempala 2004. A monograph focused on $\alpha = 2$.
- Alon, Matias, and Szegedy, 1996, STOC
- Indyk, 2006, JACM
- Li, Hastie, and Church, 2006, KDD
- Li, Hastie, and Church, 2006, COLT
- Li, 2007, KDD
- Li, Hastie, and Church, 2007, COLT
- Li and Hastie, 2008, NIPS
- Li, 2008, SODA

A lot have been done, and a lot more to do!

1. Theory
   - Statistically optimal recovery (estimation) methods
   - Computationally efficient estimation methods.
2. Applications
   - Building scalable data visualization algorithms (eg, MDS).
   - Building scalable machine learning algorithms (eg, SVM).
3. Connection to Compressed Sensing (CS)

   CS uses $\alpha = 2$ (normal) random projections.

   Can we use general αth projections for sparse signal recovery?

Compressed Counting (CC)

A new methodology recently invented

- Preliminary results: *Li, Compressed Counting, SODA 2009*.
- Based on skewed stable random projections.
- Applicable to dynamic data streams following strict-Turnstile model.
- Achieving an "infinite" improvement over symmetric projections when $\alpha \approx 1$.
- Applications in estimating entropy real-time for network anomaly detections.

Turnstile Data Stream Model

At time t, an incoming element : $a_t = (i_t, I_t)$

$i_t \in [1, D]$ index, I_t: increment/decrement.

Updating rule : $A_t[i_t] = A_{t-1}[i_t] + I_t$

Goal : Count αth moment $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$

Strict-Turnstile model : $A_t[i] \geq 0$ always, suffices for almost all applications.

For example, the strict-Turnstile model for an online bookstore

t=0

0	0	0	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=1 arriving stream = (3, 10) user 3 ordered 10 books

0	0	10	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=2 arriving stream = (1, 5) user 1 ordered 5 books

5	0	10	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=3 arriving stream = (3, −8) user 3 cancelled 8 books

5	0	2	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

Counting: Trivial if $\alpha = 1$, but Non-trivial in General

Goal : Count $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$, where $\boxed{A_t[i_t] = A_{t-1}[i_t] + I_t}$.

When $\alpha \neq 1$, counting $F_{(\alpha)}$ exactly requires D counters. (but D can be 2^{64})

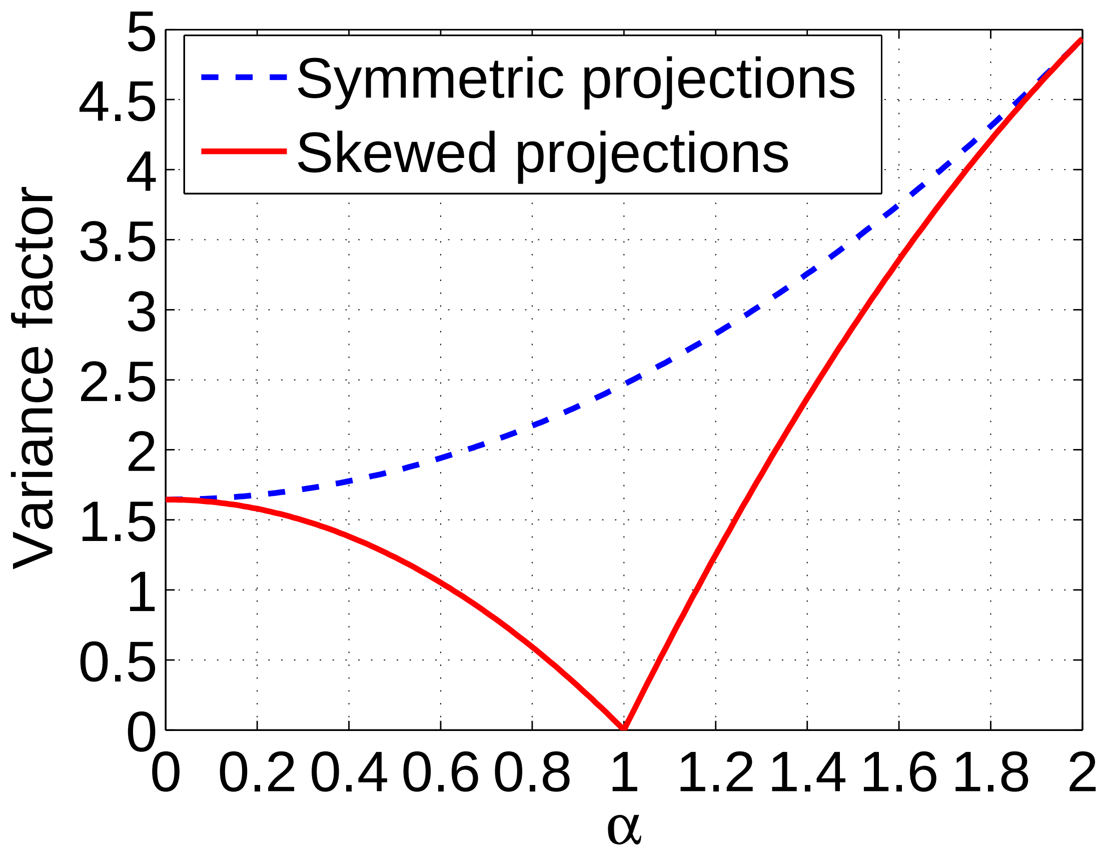
When $\alpha = 1$, however, counting the sum is trivial, using a simple counter.

$$F_{(1)} = \sum_{i=1}^{D} A_t[i] = \sum_{s=1}^{t} I_s,$$

Compressed Counting (CC) captures this intuition

Symmetric stable random projections totally ignore this fact.

Dramatic Improvement of CC, in Terms of Variances



Skewed α-Stable Distributions

Denoted by $S(\alpha, \beta, d)$. $\beta = 0$: symmetric, $\beta = 1$, maximally-skewed.

Two random variables $Z_1 \sim S(\alpha, \beta, 1)$ and $Z_2 \sim S(\alpha, \beta, 1)$.

For any constants $C_1 \geq 0$ and $C_2 \geq 0$

$$Z = C_1 \times Z_1 + C_2 \times Z_2 \sim S\left(\alpha, \beta, C_1^\alpha + C_2^\alpha\right)$$

CC works only for strict-Turnstile model.

One Application of CC: Entropy Measurement

The Shannon entropy:

- An extremely useful measurement in network data flow.

  Monitoring/visualizing network anomaly.

- Real-time measure is critical.

- It can be approximated by functions of αth moments with $\alpha \to 1$.

- Therefore, CC becomes very useful.

Research Topics for Compressed Counting

1. Theory
   - Improved estimation methods with better convergence rate as $\alpha \rightarrow 1$.
   - Computationally efficient estimation methods.
   - Computationally efficient methods for sampling skewed distributions.
2. Applications

   eg, practically efficient entropy estimation.
3. Connection to Compressed Sensing (CS)

   CC has recently attracted attention in CS community.

Limitations of Random Projections

1. Ignoring data sparsity

   eg, text data, histogram-based data

2. Applicable only to a particular αth moment.

   Different projections for different α's.

3. Not applicable to many other summary statistics

   eg χ^2 distance.

4. Applicable only to Turnstile data stream model $A_t[i] = A_{t-1}[i] + I_t$

   but real-world may need nonlinear updating rules.

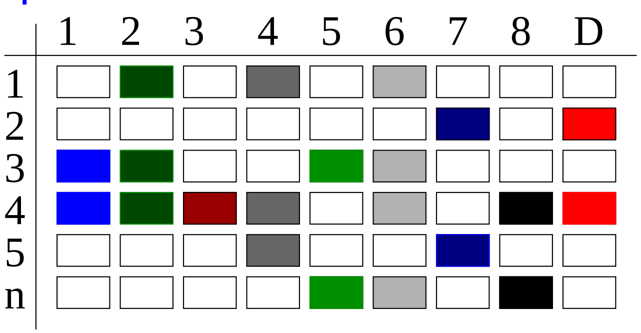
Conditional Random Sampling (CRS) provides a fix and works well in sparse data.

Conditional Random Sampling (CRS): Progress

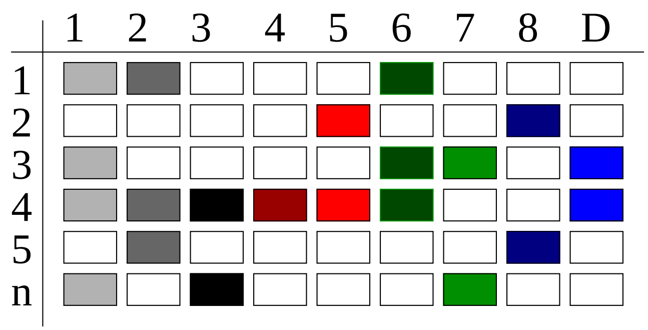
- Li and Church, EMNLP, 2005
- Li and Church, Computational Linguistics, 2007
- Li, Church, and Hastie, NIPS, 2007
- Li, Church, and Hastie, NIPS, 2009

Th Sketching Procedure of CRS

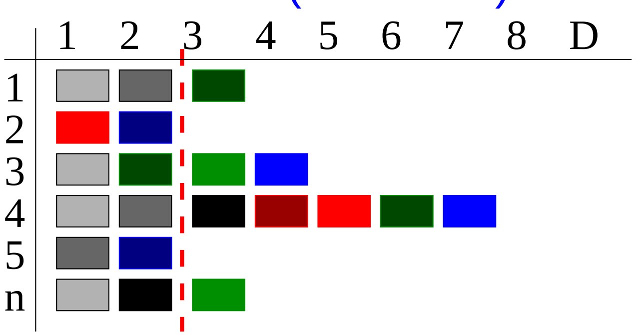
Sparse Data Matrix



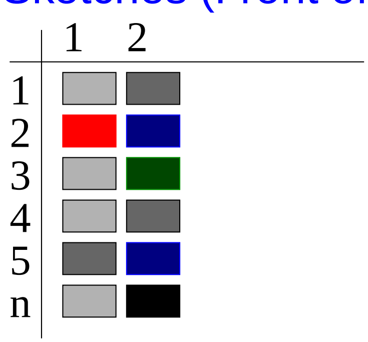
Random Permutation on Columns



Inverted Index (Nonzeros)



Sketches (Front of inverted index)



From Sketches to Random Coordinate Samples (Pairwise)

Sketches for binary (0/1) data: front of inverted index

Words														
1	1	4	5	7	11	13	15							
2	2	4	7	8	10	11	13							
3	1	3	4	5	6	9	12							
4	2	4	6	8	10	13								
n=5	1	2	3	4	5	6	7	8	9	11	12	13	14	15

Represent sketches in the (permuted) matrix

Document IDs

Words	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15=D
1	1	0	0	1	1	0	1	0	0	0	1	0	1	0	1
2	0	1	0	1	0	0	1	1	0	1	1	0	1	0	0
3	1	0	1	1	1	1	0	0	1	0	0	1	0	0	0
4	0	1	0	1	0	1	0	1	0	1	0	0	1	0	0
n=5	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1

Small sketches $\Longrightarrow$ many columns $\Longrightarrow$ random samples pairwise (?)

Research Topics of CRS

1. Theory
   - The current algorithm is basically a (very good) heuristic.

     The exact solution is a difficult classical statistical problem.
   - Improved estimation methods using side information.
2. Applications
   - Scalable data visualization algorithms.
   - Scalable machine learning algorithms.
   - Maintaining multi-way histograms.
   - General data stream applications.
3. Combining CRS with random projections

How Will This Influence FODAVA? Broader Impact?

Possibly all data analytics and visualization techniques need to address

- How to feasibly store massive data in a compact format?
- How to update the data in dynamic settings?
- How to compute summary statistics (distances) efficiently or real-time?

Broader Impact:

- Scalable machine learning
- Databases and information retrieval
- Network measurement
- (Possibly) sparse signal recovery (compressed sensing)

Plans for Helping the Development of FODAVA

- Attending conferences in visualization and massive data sets

  eg, IEEE VAST, DHS NVAC, MMDS 2010 (?)

- Introducing the basic problems/solutions to traditional statistics community

- Introducing statistical techniques to Computer Science community

- Publishing in CS conferences and statistical journals

- Collaborating with other FODAVA research teams.