Discovering Persistent Change Windows (PCW) in Spatiotemporal Climate Datasets

Shashi Shekhar

Students:

Xun Zhou

Zhe Jiang

Joseph F. Knight

Stefan Liess

Peter K. Snyder

3rd Annual Workshop of NSF Expeditions in Computing:
Understanding Climate Change: A Data Driven Approach.
Evanston, IL, August 15th-16th, 2013
Sponsor: NSF CISE/EIA

Change Windows

- **Time-Window of Change**, e.g., interesting interval in <u>a</u> time series

- **Space-window of Change (**e.g., Region or Sub-path **):**


Desertification


Deforestation


Urban sprawl

- **Spatio-Temporal Window of Change**:


2001

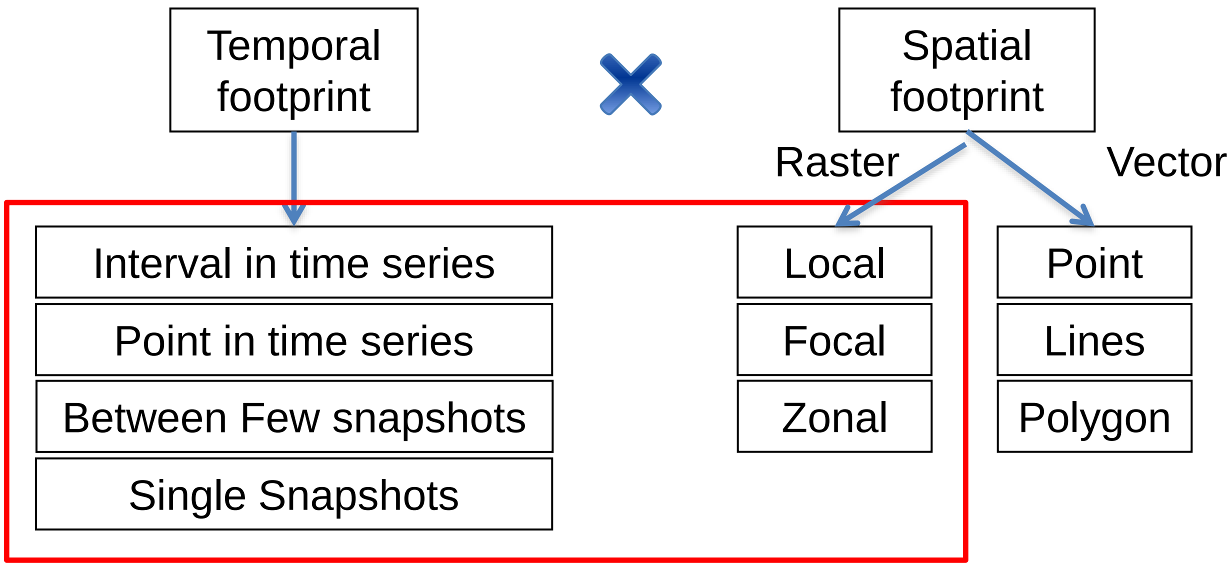

2006

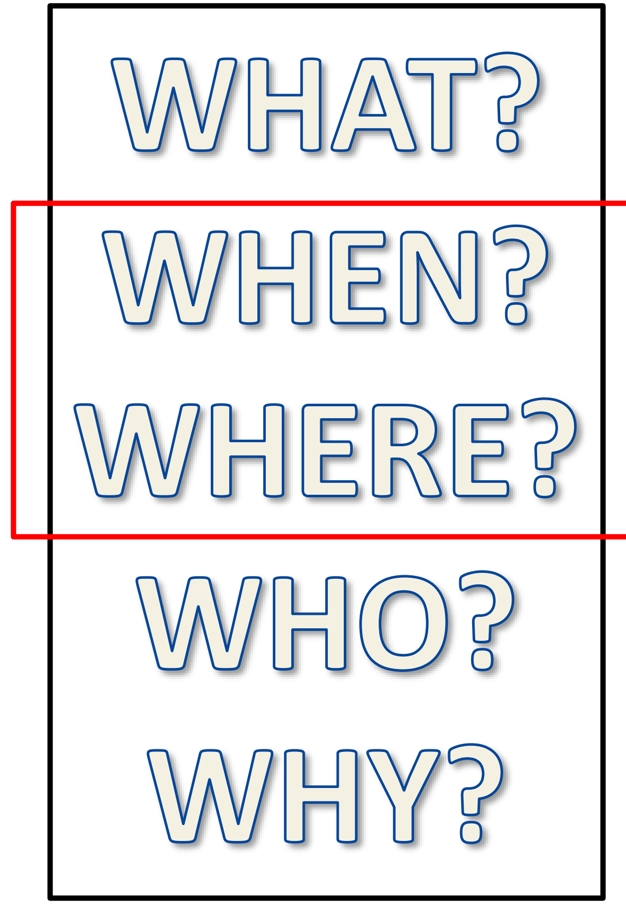

2012

Irrigation in Saudi Arabia (Google Time Lapse[5])

Spatiotemporal (ST) footprint of changes

- "Where" and "when" a change occurs?
- A taxonomy of ST change footprint:



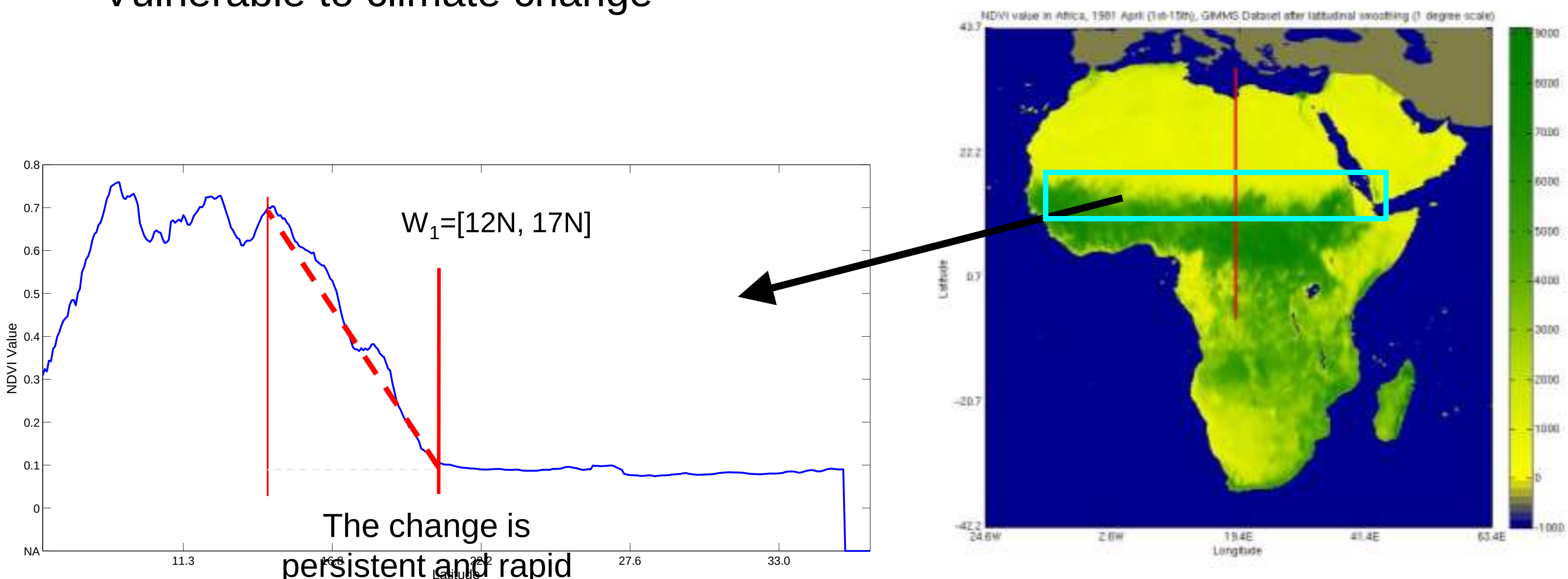


Spatiotemporal change footprint (raster)

		Temporal				
		Snapshot	Few Snapshots	Time series (point)	Time series (interval)	Time series collections
Spatial	Local					
	Focal					
	Zonal					

Spatial Sub-path of Change

- Spatial footprint of Change
  - Ex. Sahel – sharp change in vegetation cover
  - Transition between ecological zones (ecotones)
  - Vulnerable to climate change

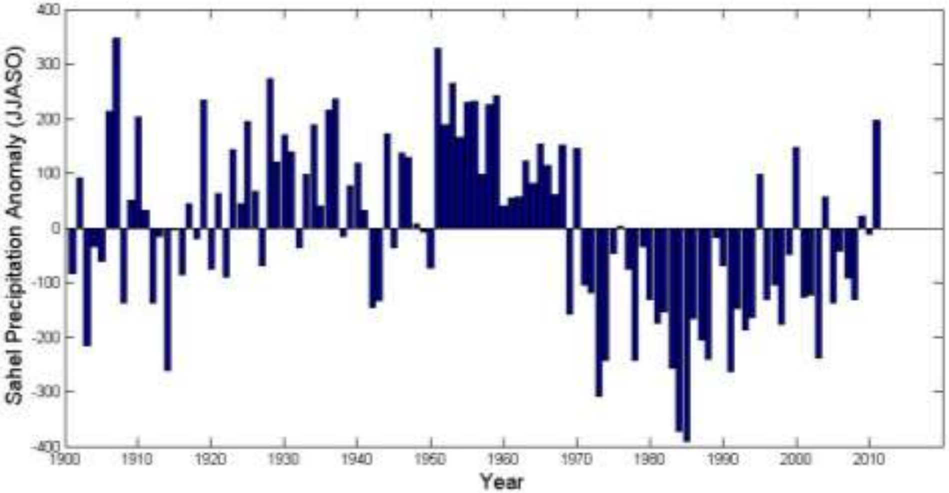


A snapshot of vegetation cover in Africa [6]

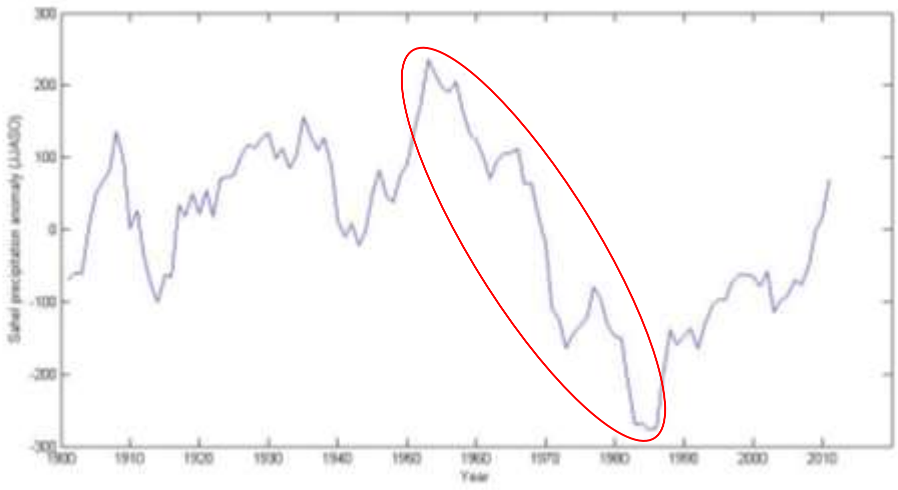
Temporal Sub-path of Change

- Time footprint of Change
  - Abrupt shift in precipitation, temperature, etc.
  - Climate change detection.

Raw Sahel precipitation anomaly (JJASO)[7]



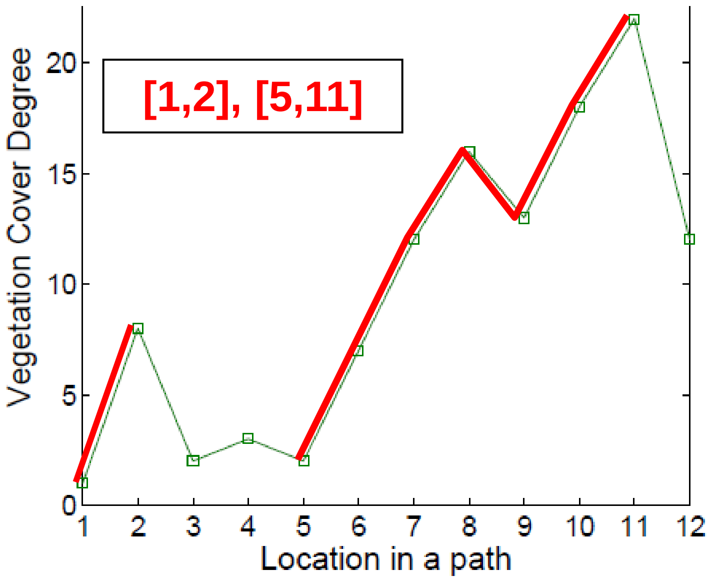
Smoothed Sahel precipitation anomaly (JJASO)



Computer Sc. Problem: Interesting Sub-path Query (ISQ)

- Input
  - A statistical interest measure & thresholds.
  - A path and its attribute
- Output
  - All dominant interesting sub-path
- Constraints
  - Correctness & completeness
  - Automation & scalability to large datasets
  - Algebraic Interest Measure,e.g., average slope

Average change (slope) ≥ 3.5



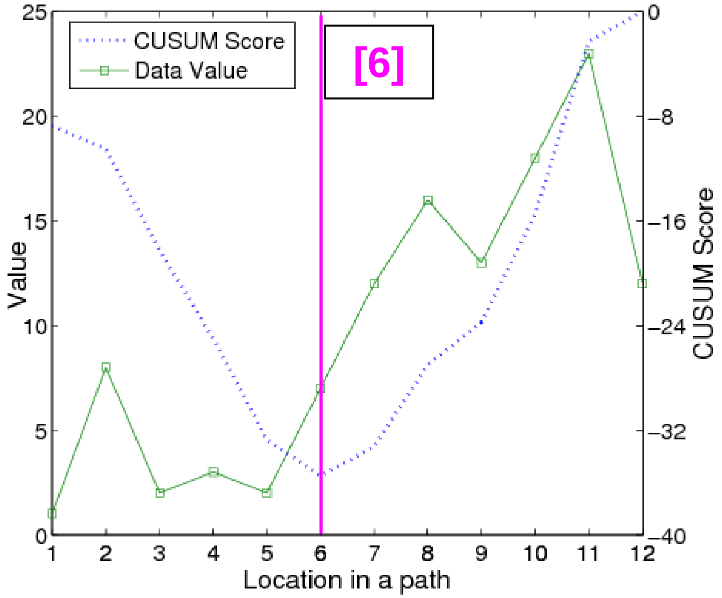
Related Work, Its Limitations, Novelty of Our Approach

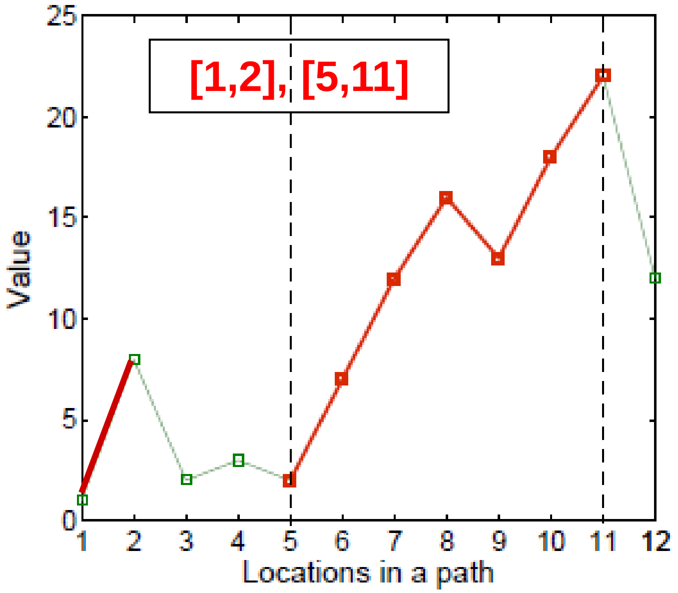
Interesting sub-region query

Change-points
e.g., CUSUM[8]

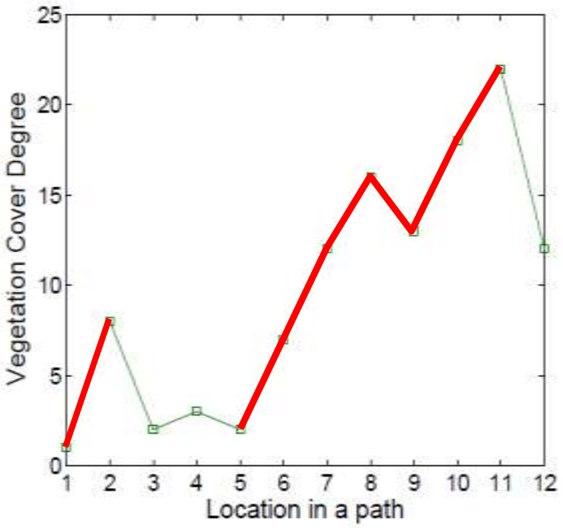
sub-paths
e.g., ISQ (Our Work)

sub-regions
(PCW) (Our new Work)





Naive Approach does not scale!



Naive approach

Phase 1: Evaluate interest measure for all $O(N^2)$ sub-intervals

Phase 2: Identify dominant sub-paths (compare sub-path pairs)

Complexity: $O(n^4)$ for a path,

where n = number of locations on the path

Example: Find intervals of high gradient along all longitudes

using 0.07 degree resolution dataset

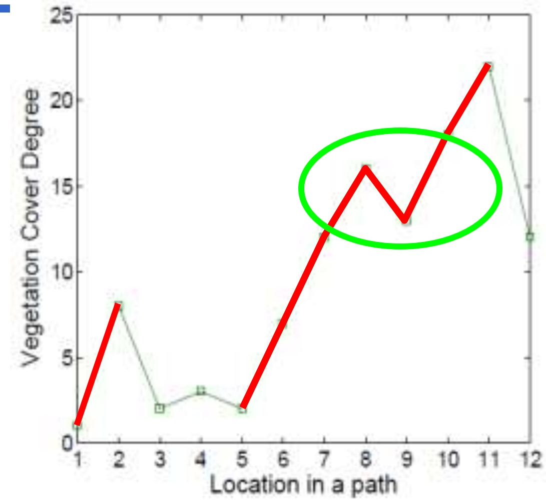
=> 10^3 locations per path, 10^2 longitudes,

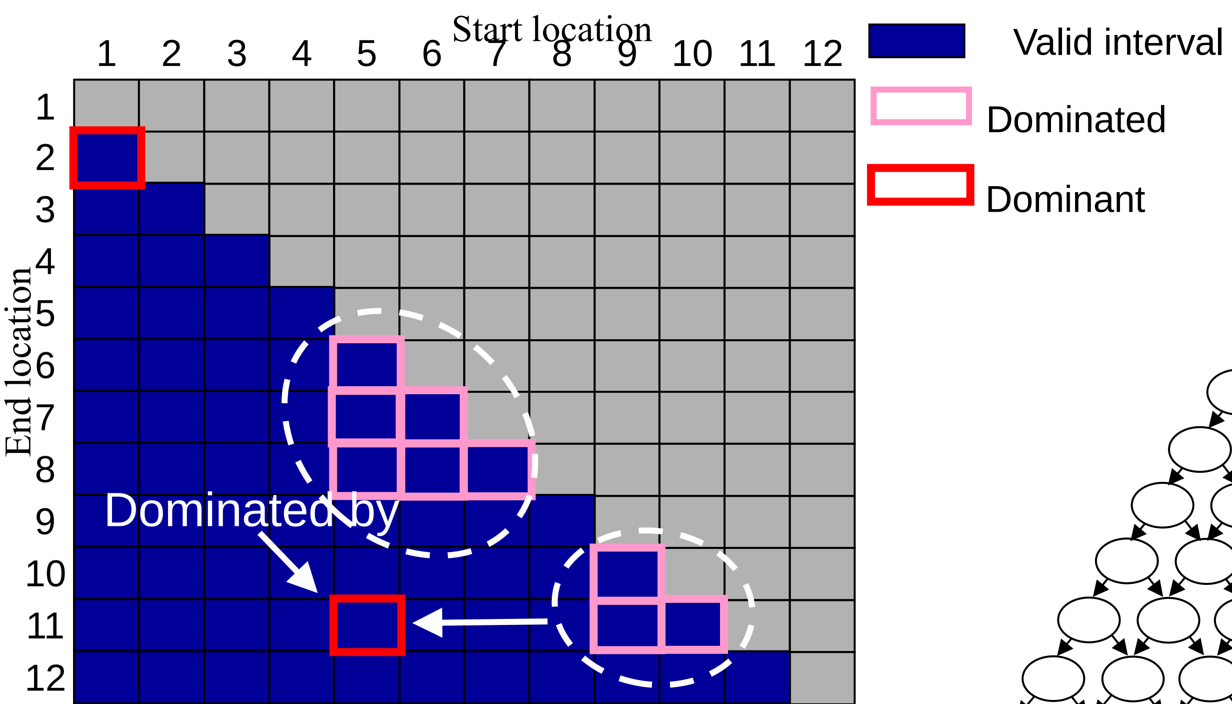
=> 10^{14} computations

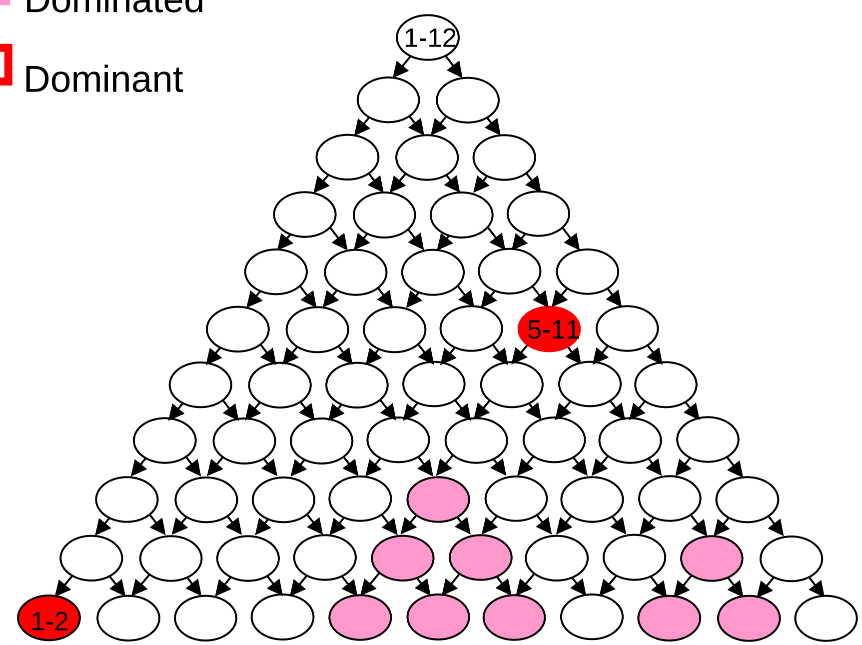
Window footprint	1-D (path)	2-D (spatial)	3-D (spatiotemporal)
# locations	10^3	10^6	10^9
# windows	10^8	10^{12}	10^{18}
computation	10^{14}	10^{24}	10^{36}
Search spaces	$O(n^2)$	$O(n^4)$	$O(n^6)$

Computational Structure of ISQ

- Not Dynamic Programming!
- GRID-DAG (Directed Acyclic Graph)
  - Node = sub-paths
  - Edge = Dominance relationship
  - Interest Measure (node) = f (leftmost & rightmost leafs)
- Q? Which traversal order avoids unnecessary work?

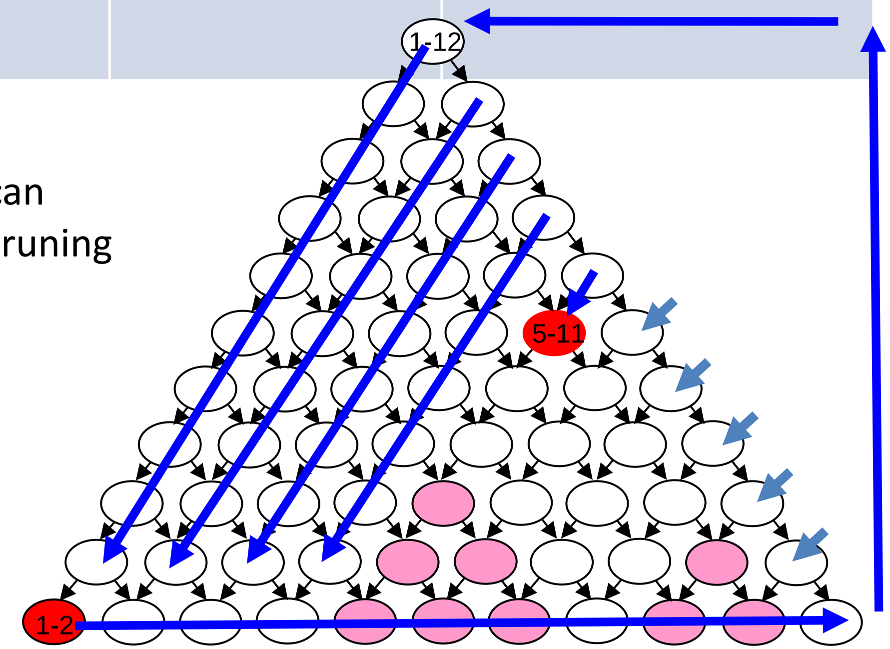






A Comparison of Techniques for Traversing G-DAG

	DFS or BFS	Bottom-Up with in-Row Pruning (BURP)	BFS with Sub-Graph Pruning (BSGP)	
A: Avoid Redundant leaf visits	No	Yes	Yes	
B: Avoid Unnecessary visits to dominated non-leafs	No	No	Yes	
C: Memory need for **B** (n = number of locations in path)	$O(n^2)$	$O(n)$	$O(n^2)$	



Insights:

- Interest measure is a algebraic function => leaf scan
- Dominance = partial order among sub-paths => pruning
- The partial order is a **Grid-DAG** =>
  O(1) memory traversal & **pruning**
  via row-wise scan (of non-leaf)

Theoretical and Experimental Evaluations

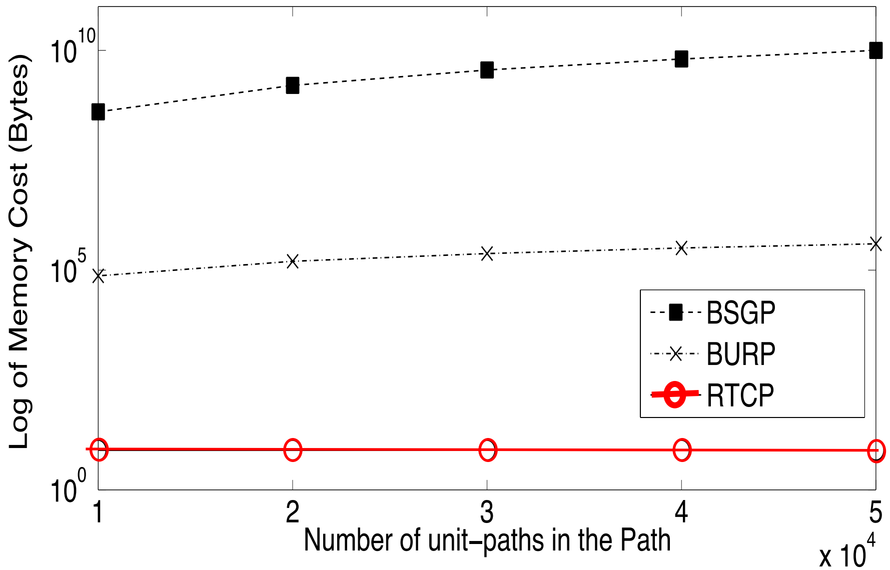
- Theoretical Evaluation:
  - RTCP is Correct and Complete
  - Correct: All the reported sub-paths are qualifying dominant sub-paths
  - Complete: All the dominant interesting sub-paths are reported
- Experimental Evaluation
  - RTCP is faster than competitions
  - RTCP needs less memory than competition

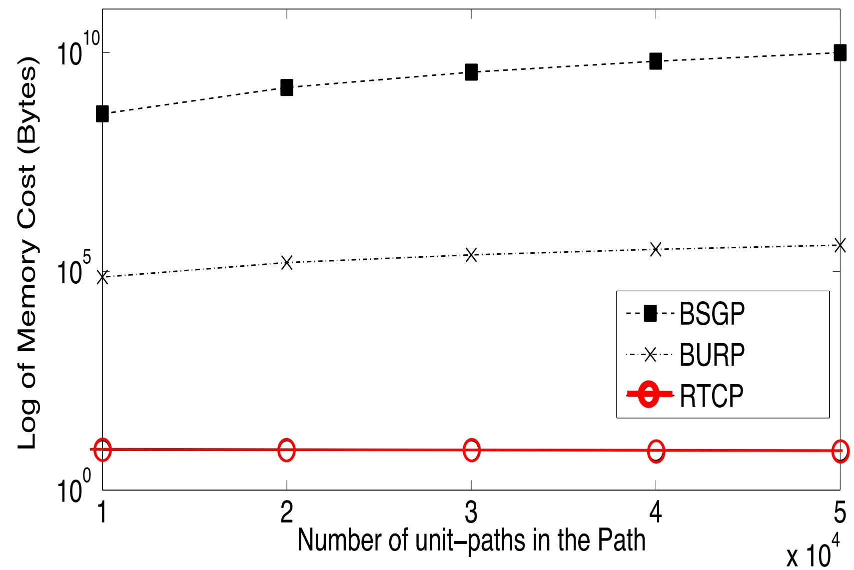
Effect of Dataset Size on Memory Needs

- Setup: Pattern length ratio (PLR) is fixed at 0.1

  Dataset – synthetic (left), real (right)

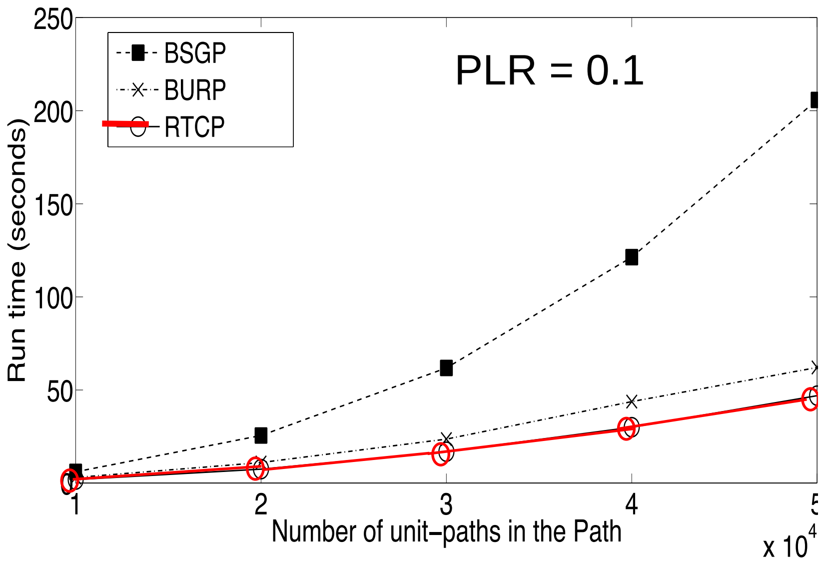
- Trends: RTCP has smaller memory cost than competitions
  - Memory cost are not sensitive to pattern length ratio

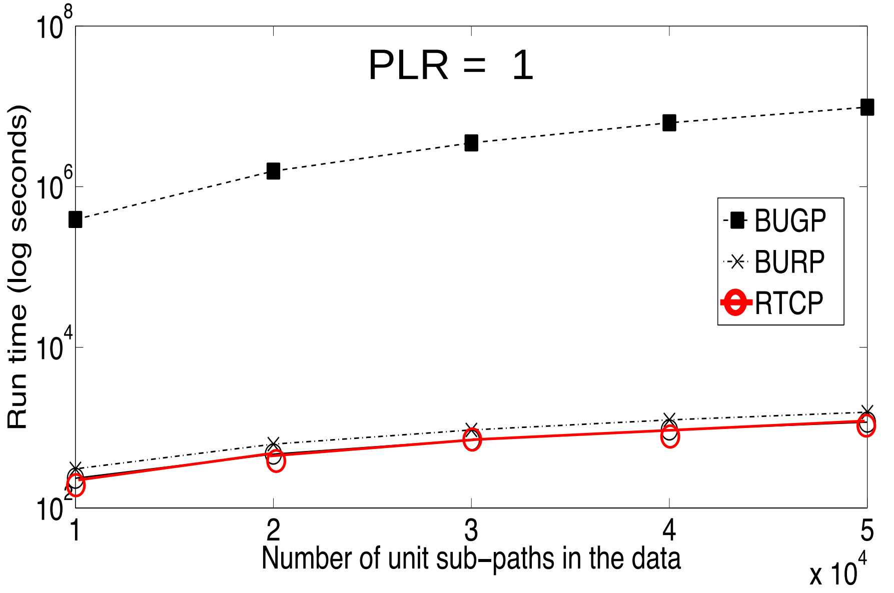




Effect of Data Size on Computational Cost

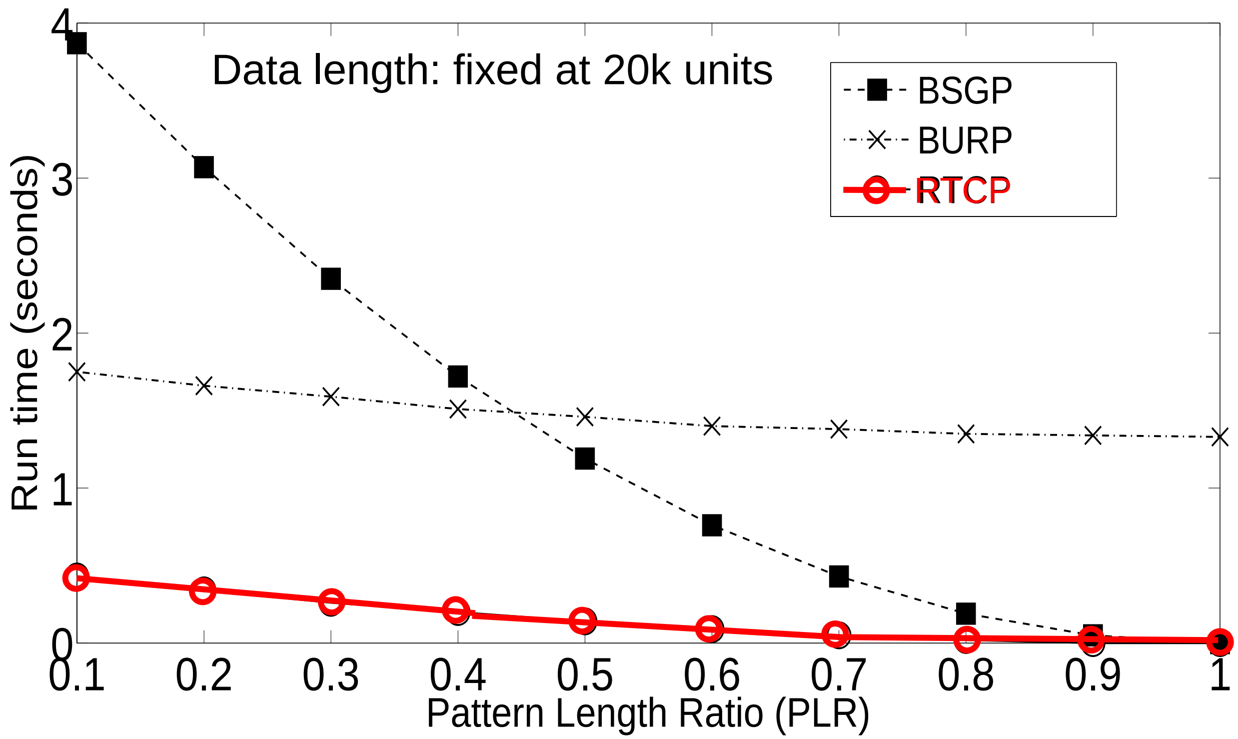
- Setup:
  - Data: GFDL-CM2.0 coupled model realization (1861-2000, entire world)
  - Global weighted average of max daily temperature
  - Data length: 51100 :: Pattern Length Ration (PLR) = either 0.1 or 1
  - Note difference in scale across 2 plots
- Trends: RTCP is faster than competitions





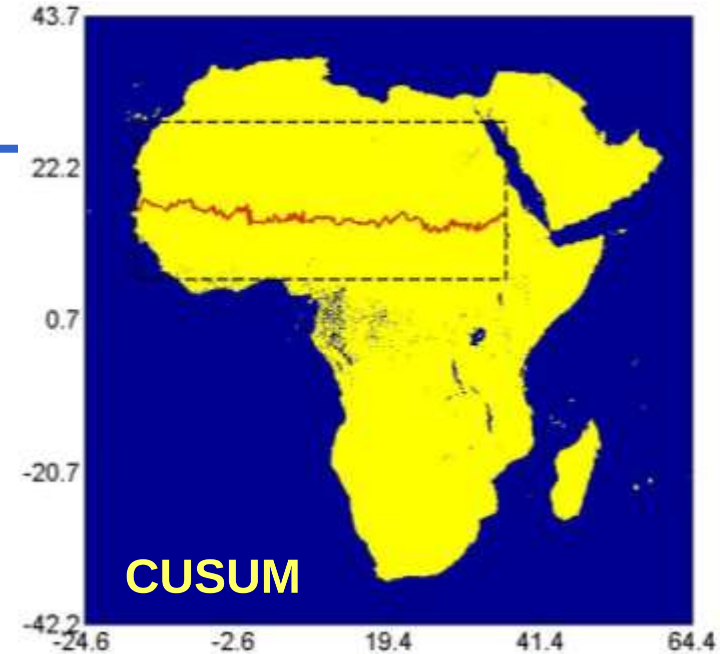
Effect of Pattern Length

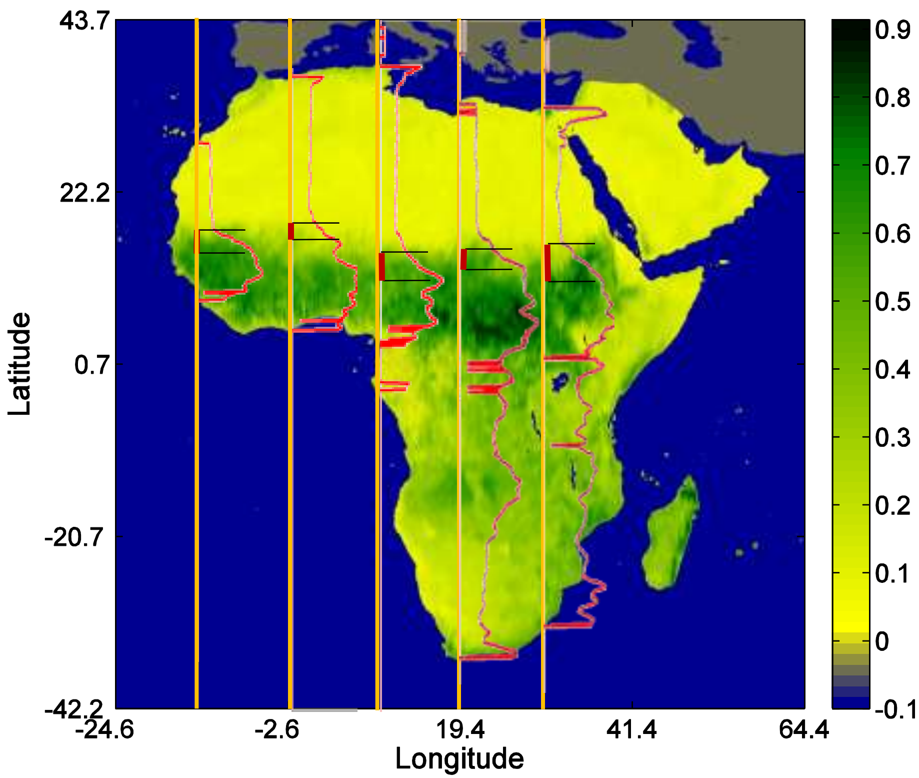
- Pattern Length Ratio =
  - Ratio of length of longest interesting sub-path and the length of the entire path,
  - between 0 and 1.
- Synthetic data: generated with Gaussian distributed unit values.
- Trend: RTCP is faster than competitions with any pattern length

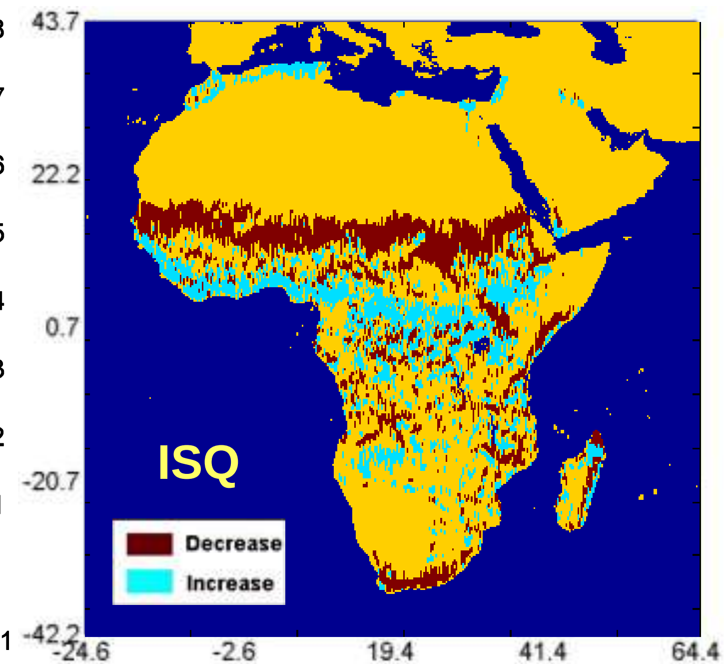


Case Study

- Data: NDVI by GIMMS [4], Africa, 1981 August.
  - Resolution: 8km. Smoothed within 1x1 degree.
- Path: along each longitude (south → north)
- Interest measure: (Slope) Sameness degree $\frac{AVG\{\Delta\}}{AVG_{\geq\alpha}\{\Delta\}}$
  - Δ : unit slope
- Thresholds: α= 20% percentile, SD ≥0.5

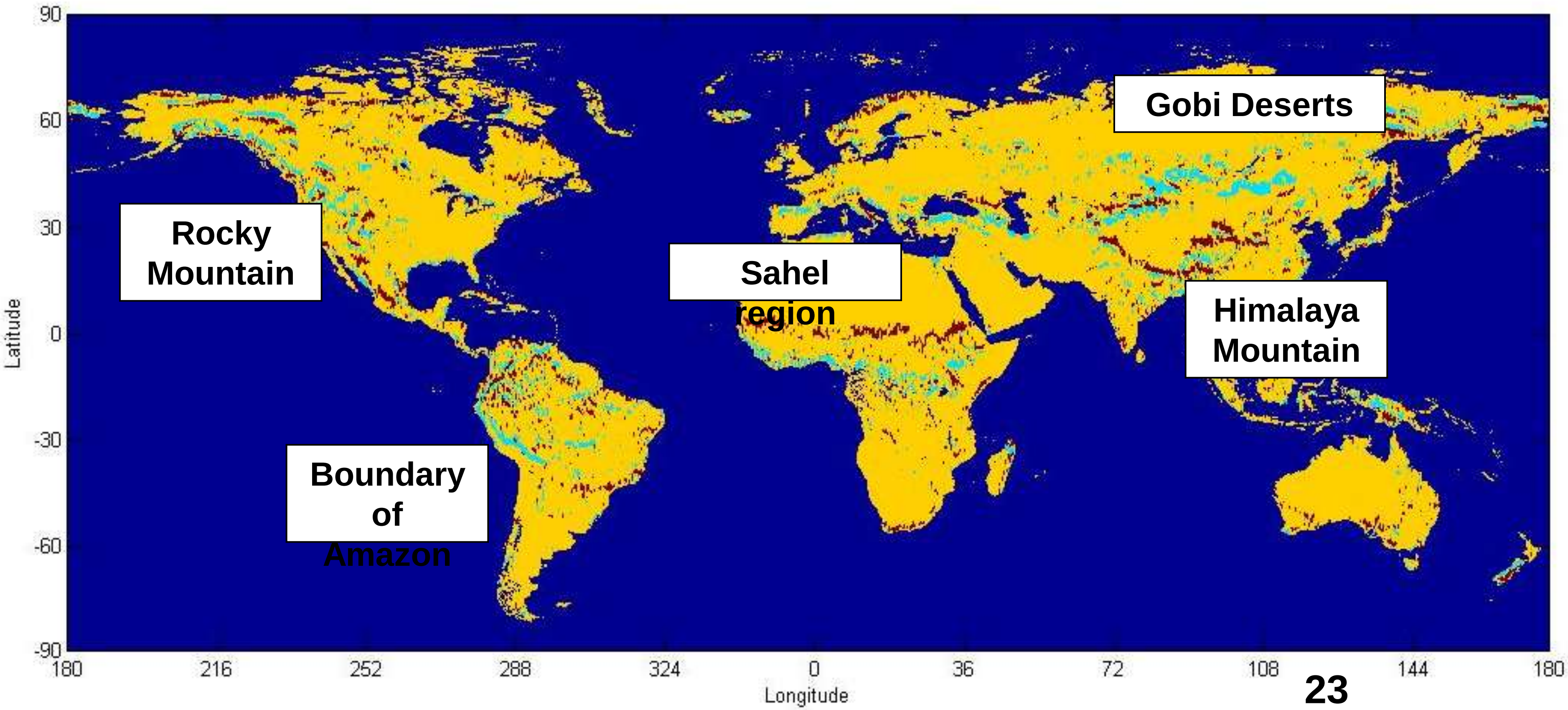






Case Study: Global Data and Ecotones

- NDVI of the entire world
- Aug 1-15 1981, 0.07 degree (8km) resolution
- $a = 10\%$, **SD** ≥ 0.5



Contributions : Summary

- Formalize Interesting Sub-path Query (ISQ) problem
- Computational structure of efficient 1D interval enumeration
  - Grid-DAG traversal strategy
  - Bottom-up with in-row pruning (BURP) [1], BFS with sub-graph pruning (BSGP)[2]
  - Row-wise with column-pruning (RTCP)[3]
- Evaluation : Experiments, Case Studies
- Next Steps
  - GPU platform
  - Efficient 3D region enumeration
  - submitted to ACMGIS 2013.
  - Visit poster session for details

	1-D (path)	2-D (spatial)	3-D (spatio-temporal)
# locations	10^3	10^6	10^9
# windows	10^8	10^{12}	10^{18}
computation	10^{14}	10^{24}	10^{36}
Search spaces	$O(n^2)$	$O(n^4)$	$O(n^6)$

[1] From our early work: J. Kang, et al, Discovering Flow Anomalies: A SWEET Approach. In IEEE conference on Data Mining (ICDM 2008).
[2] Published in ACM SIGSPATIAL GIS 2011.
[3] Manuscript to be submitted to IEEE Transaction on Knowledge and Data Engineering (TKDE).

Accelerating ISQ with GPUs – Preliminary Results

- Motivation: Google Time-lapse like visualization of spatio-temporal change windows
  - Example: GIMMS NDVI
  - 611 snapshots (26 years every 16 days)
  - Africa, 8km resolution: 1152 pixels x 1152 pixels

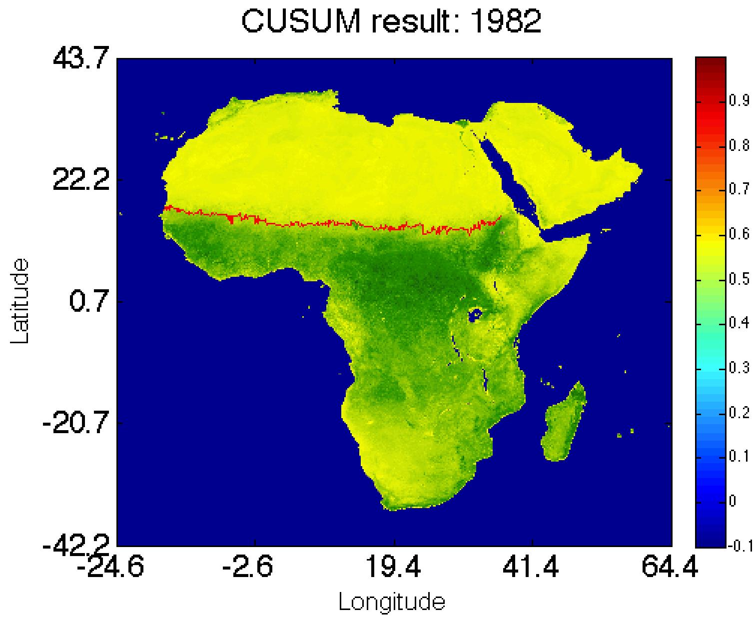
- Initial results with BURP (Bottom-up with in-row pruning)
  - Setup: Platform: CUDA, 1 thread per longitude
  - Trend: 10x speedup with 1 GPU
  - CPU time (matlab): **240 s**econds / video
  - GPU/CUDA: **19 seconds / video**

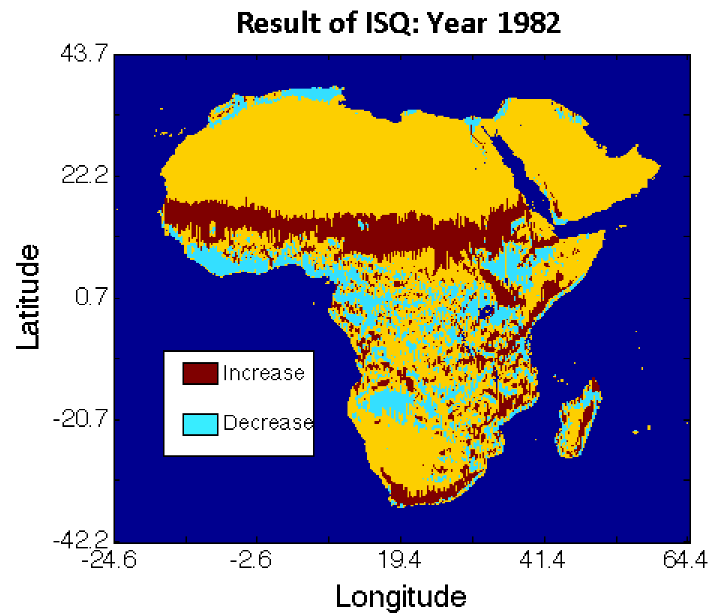
- Next steps:
  - More algorithms (e.g., RTCP), Platforms (e.g., multi-GPU), Datasets

* Collaborated with Dr. S. Prasad et al. in Georgia State Univ. Results submitted to CyberGIS AHM'13

Case Study

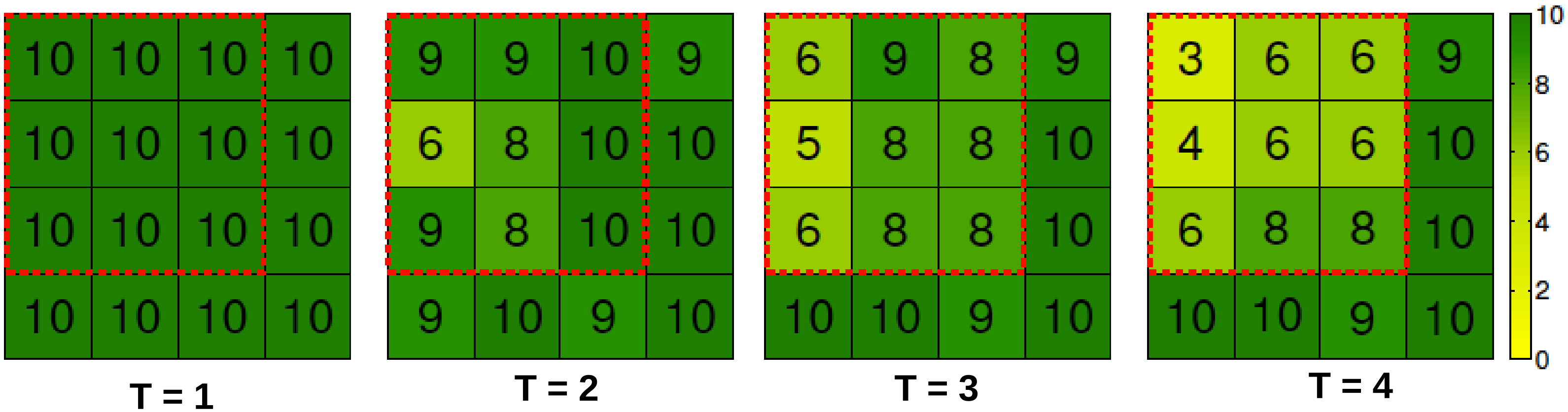
CUSUM (pre-selected area) vs. Proposed approach (a=20% quantile, SD = 0.5)





Persistent Change Window (PCW) Discovery: Problem

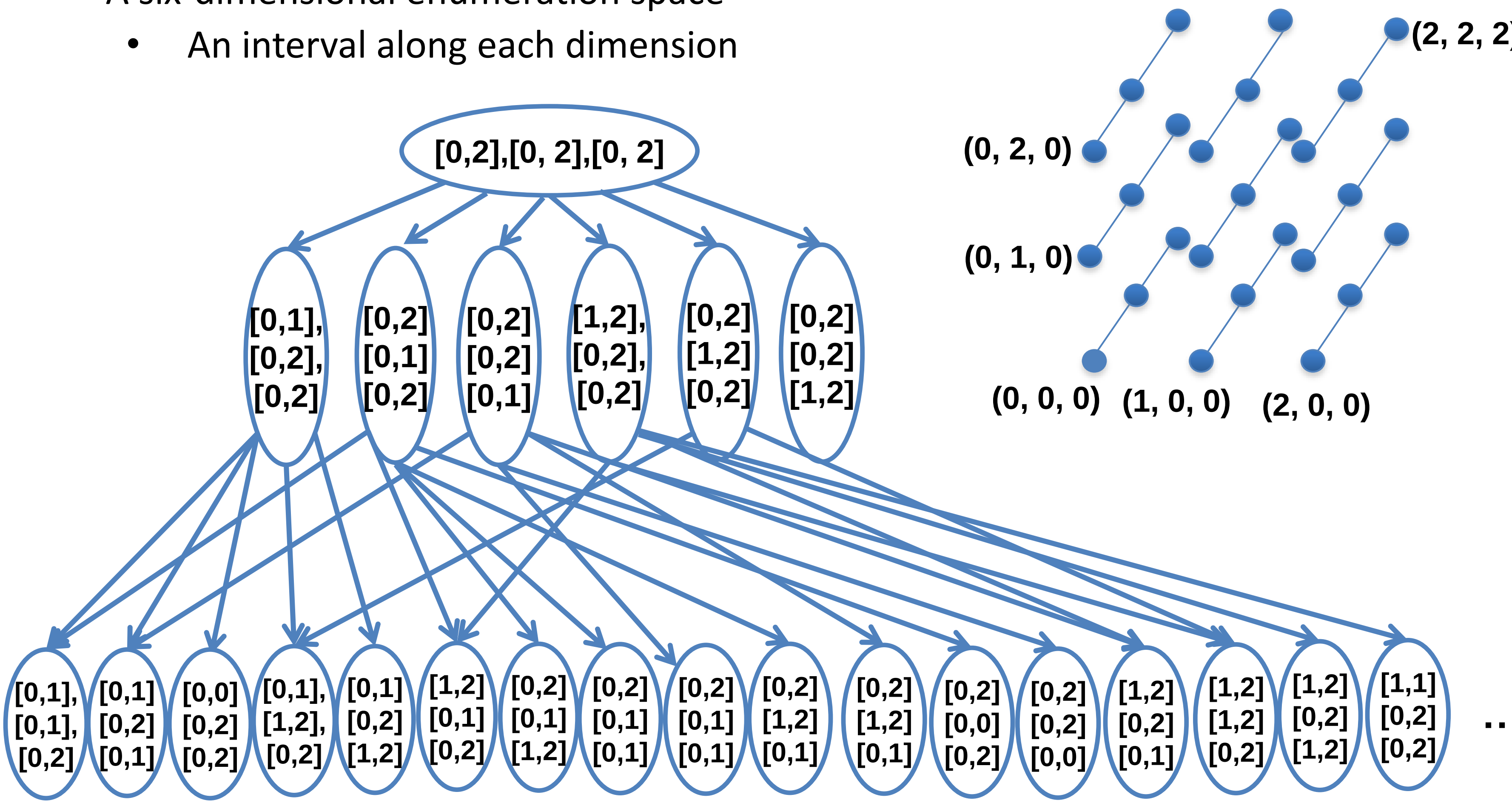
- Given:
  - A spatial time series dataset
  - An average change rate threshold
  - A spatial aggregate function (e.g., sum, average)
- Find:
  - All the dominant persistent change windows {Si, Ti}
- Constraints
  - Correctness & completeness
  - Automation & scalability to large datasets



Highlighted (3x3) over Time [1,4]: Sum(T1) = 90, Sum(T4) = 53 **Threshold = 15%**
Average decrease rate = [(90-53)/90]/3= 17.3%

Persistent Change Window (PCW) Discovery: Challenges

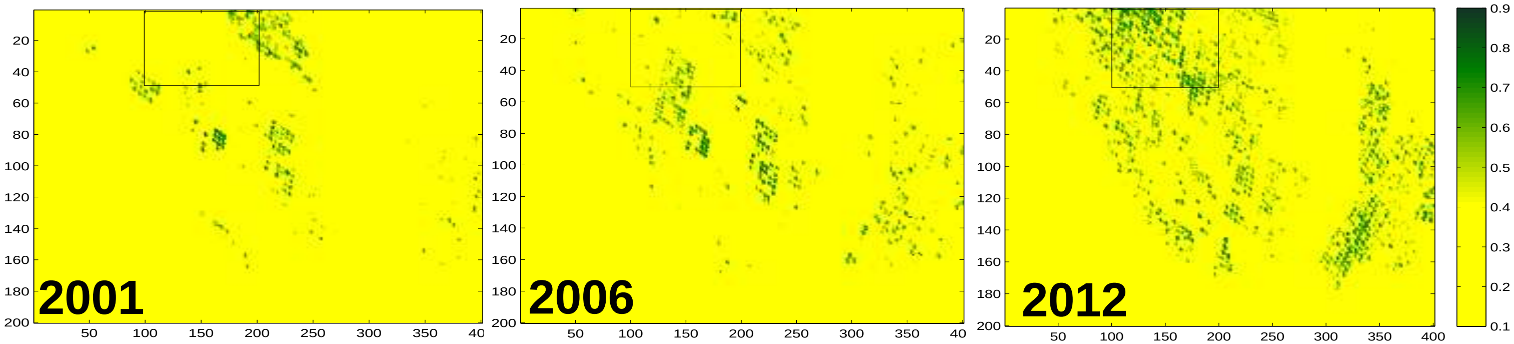
- A six-dimensional enumeration space
  - An interval along each dimension



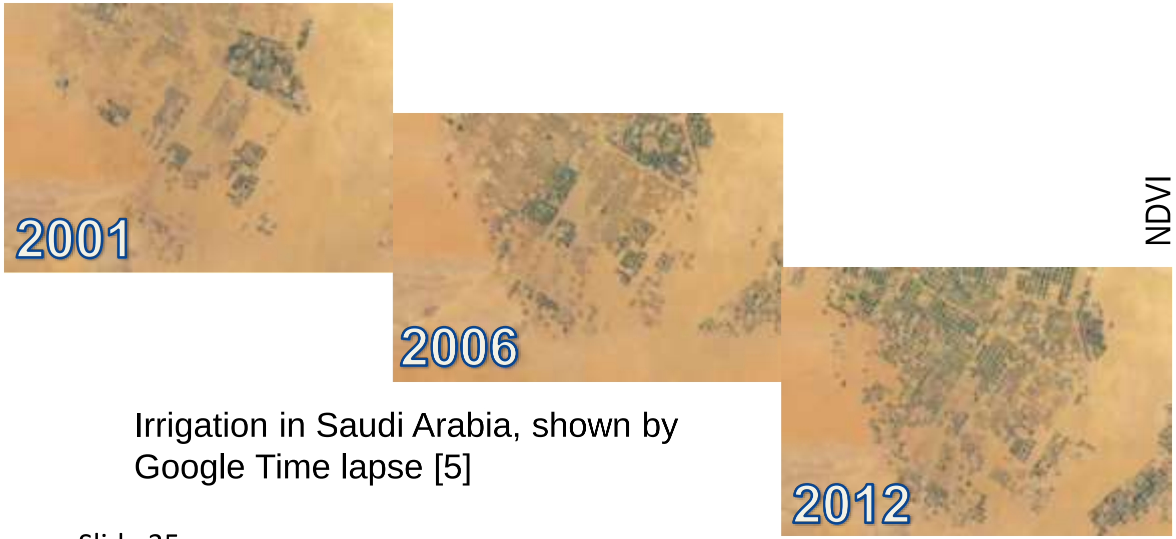
Persistent Change Window Discovery: Case Study

- Initial Results
  - Initial algorithms
  - Case Study: MODIS 250m NDVI data (16 days)
  - Time:2000-2012. Annual: July 27/28 of each year.

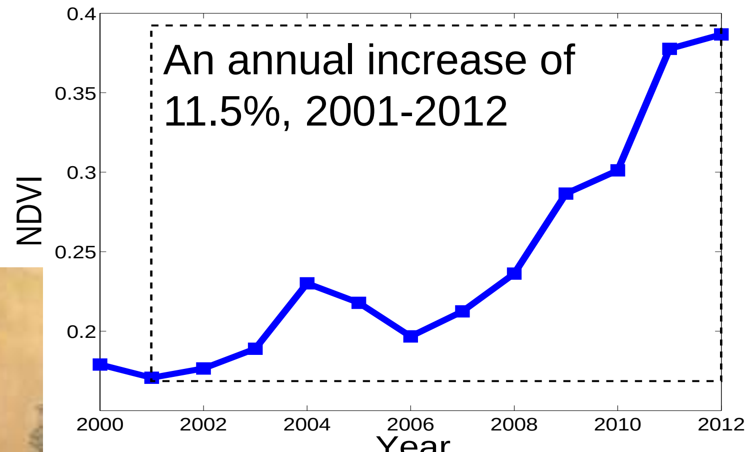
Study area



← Results of the proposed algorithm with average change rate >= 10% (outlined window)



Irrigation in Saudi Arabia, shown by Google Time lapse [5]



Average NDVI in outlined window

List of Publications and References

Contributors' Publications:

[1] Xun Zhou, Shashi Shekhar, Pradeep Mohan, Stefan Liess, Peter K. Snyder: Discovering interesting sub-paths in spatiotemporal datasets: a summary of results. GIS 2011: 44-53

(A full journal version to be submitted to IEEE Transection on Knowledge and Data Engineering)

[2] Xun Zhou, Shashi Shekhar, Pradeep Mohan. Spatiotemporal (ST) Change Pattern Mining: A Multi disciplinary Perspective. In Book: Mei-Po Kwan, Douglas Richardson, Donggen Wang and Chenghu Zhou (eds) (2013) Space-Time Integration in Geography and GIScience: Research Frontiers in the US and China. Dordrecht: Springer (in Press)

[3] Xun Zhou, Shashi Shekhar, Reem Y. Ali. Spatiotemporal (ST) Change Pattern Mining: A Multi-disciplinary Perspective. Submitted to the Wiley's Interdisciplinary Review on Data Mining and Knowledge Discovery (DMKD).

[4] Xun Zhou, Shashi Shekhar, Dev Oliver. Discovering Spatiotemporal (ST) Persistent Change Windows: A Summary of Results. Submitted to ACM SIG SPATIAL GIS 2013.

References:

[5] Google Earth Engine (Accessed: June 22, 2013).

[6] Tucker, C. J., J. E. Pinzon, M. E. Brown. Global inventory modeling and mapping studies. Global Land Cover Facility, University of Maryland, College Park, Maryland, 1981--2006.

[7]. Joint Institute for the Study of the atmosphere and Ocean(JISAO). Sahel rainfall index. http://jisao.washington.edu/data/sahel/.

[8] E. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100--115, 1954.

Google Time lapse

- **Google Time lapse main page**
- Irrigation in Saudi Arabia
- Amazon Deforestation