جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Tile Low-Rank Approximation of Maximum Large-Scale Likelihood Estimation on Manycore Architectures

ExaGeoStat: A Framework for Large-Scale Weather and Climate Prediction using Machine Learning

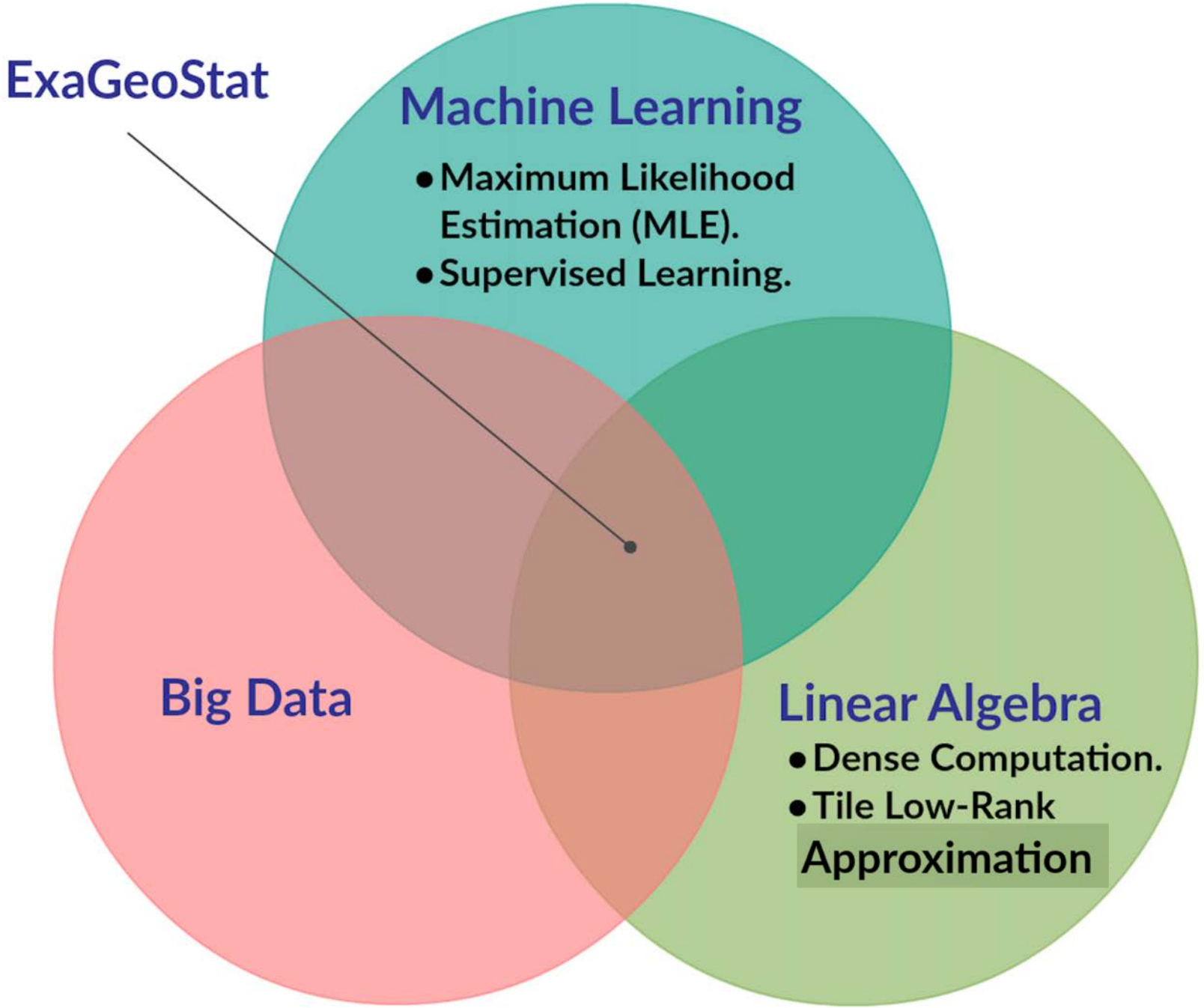
SAMEH ABDULAH, HATEM LTAIEF, YING SUN, MARC GENTON AND DAVID KEYES

EXTREME COMPUTING RESEARCH CENTER,
KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Motivation

- Maximum Likelihood Function
  - An important statistical technique for predicting unknown measurements in climate and environmental applications
- Weather and climate data are usually big!
  - Prohibitive computational Cost and memory requirements
  - Two real examples: Soil Moisture and Wind Speed
- Exascale GeoStatistics (ExaGeoStat)
  - A framework which exploits machine learning, statistical modeling and forecasting, and the state-of-the-art linear algebra techniques to handle large-scale geostatistics data


ExaGeoStat
Machine Learning
• Maximum Likelihood Estimation (MLE).
• Supervised Learning.
Big Data
Linear Algebra
• Dense Computation.
• Tile Low-Rank Approximation

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

ExaGeoStat Components

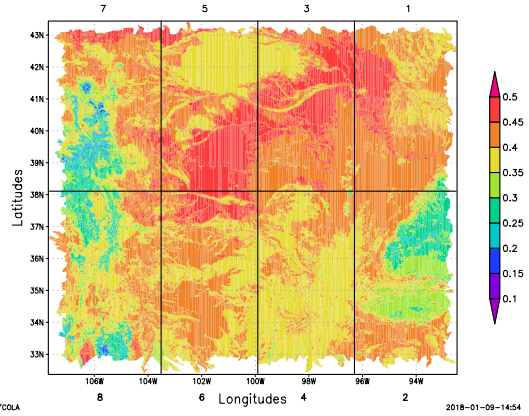
- Synthetic Dataset Generator
  - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages
- Maximum Likelihood Estimator (MLE)
  - Evaluate the maximum likelihood function on large-scale geospatial datasets
  - Support full machine precision accuracy (full-matrix) and Tile Low-Rank (TLR) approximation
- ExaGeoStat Predictor
  - Predict unknown measurements on known geospatial locations by leveraging the MLE estimated parameters

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

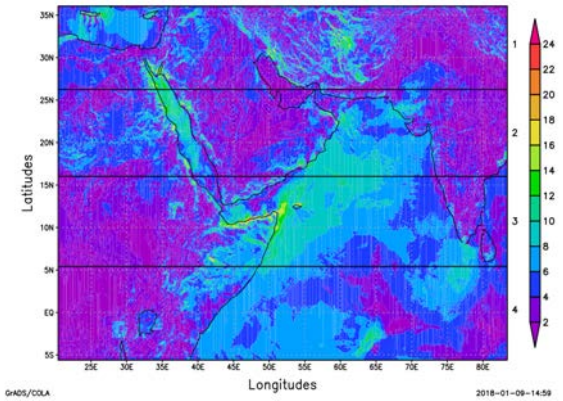
Outcomes

- Hardware-agnostic framework
- 2M spatial locations (160 TB memory requirement)
- 96 % prediction accuracy on real datasets
- Large synthetic spatial data generation tool
- R-Wrapper package

Soil Moisture Data



Wind Speed Data



جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact Computation using Chameleon dense Linear algebra library and StarPU runtime system
  - Support real and synthetic geospatial datasets
  - Soil moisture dataset at Mississippi basin area
- Today:
  - ExaGeoStat supports
    - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation library and StarPU runtime system
    - Performance results of TLR-based computations on shared and distributed-memory systems attain up to 13X and 5X speedups
    - Support NetCDF Format
    - Win Speed dataset at Middle-East

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

COME AND SEE US!

TILE LOW-RANK APPROXIMATION OF MAXIMUM LARGE-SCALE LIKELIHOOD ESTIMATION ON MANYCORE ARCHITECTURES

SAMEH ABDULAH, HATEM LTAIEF, YING SUN, MARC GENTON AND DAVID KEYES
EXTREME COMPUTING RESEARCH CENTER, KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

EXAGEOSTAT FRAMEWORK

- The Exascale GeoStatistics project (ExaGeoStat) is a parallel high performance unified framework for computational geostatistics on many-core systems.
- The project aims at optimizing the likelihood function for a given spatial data to provide a way to predict missing observations in the context of climate/weather forecasting applications using both full-tile and Tile Low-Rank (TLR) approximation techniques.
- This machine learning framework proposes a unified simulation code structure to target various hardware architectures, from commodity x86 to distributed-memory systems.
- ExaGeoStat enables statisticians to tackle computationally challenging scientific problems at large-scale, while abstracting the hardware complexity, through state-of-the-art high-performance linear algebra software libraries (i.e., Chameleon and HiCMA).

EXAGEOSTAT V0.1.0

Synthetic Dataset Generator

- Generate large-scale geo-spatial datasets which can be separately used as benchmark datasets for other software packages.

Maximum Likelihood Estimator (MLE)

- Evaluate the maximum likelihood function on large-scale geo-spatial datasets.
- Support full machine precision accuracy (Full-tile) and Tile Low-Rank (TLR) approximation.
- Support both synthetic and real datasets.
- Support NetCDF format.

ExaGeoStat Predictor

- Predict unknown measurements on known geospatial locations by leveraging the MLE estimated parameters.

SOFTWARE RELEASE

- C-based ExaGeoStat Package can be found at: https://github.com/ecrc/exageostat
- R-Wrapper ExaGeoStat Package can be found at: https://github.com/ecrc/exageostatR

DATASET GENERATOR

- Generate 2D spatial Locations using uniform distribution based on the Matérn matrix kernel:

$$C(r;\boldsymbol{\theta}) = \frac{\theta_1}{2^{\theta_3-1}\Gamma(\theta_3)}\left(\frac{r}{\theta_2}\right)^{\theta_3} K_{\theta_3}\left(\frac{r}{\theta_2}\right)$$

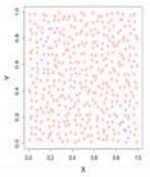
- Cholesky factorization of Covariance Matrix:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = V \cdot V^T$$

- Z observation Vector Generation:

$$\vec{Z} = V \cdot \vec{e}, \quad \vec{e} \sim N(0,1)$$

- An example of 400 points irregularly distributed in space, with 362 sampled points (o) and 38 non-sampled points (×).

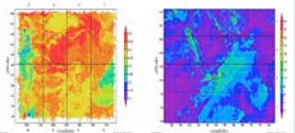


MAXIMUM LIKELIHOOD ESTIMATOR

- Definition of the MLE function:

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}Z^T\Sigma(\theta)^{-1}Z$$

- Two examples of real datasets (wind speed, Middle-East and soil moisture, Mississippi River Basin, USA).



EXAGEOSTAT PREDICTOR

- The Associated Conditional Distribution where Z_1 represents unknown measures:

$$\mathbf{Z}_1|\mathbf{Z}_2 \sim N_m(\mu_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Z}_2 - \mu_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

SYNTHETIC DATA ACCURACY VERIFICATION

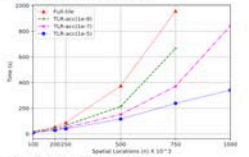
Data Corr.	Average Mean Square Error (MSE)							
	Est. θ_1		Est. θ_2		Est. θ_3		Prediction	
	T(-7)	FT	T(-7)	FT	T(-7)	FT	T(-7)	FT
Strong	0.192	0.085	0.134	0.112	0.003	0.003	0.012	0.012
Medium	0.151	0.071	0.040	0.016	0.004	0.004	0.036	0.036
Low	0.025	0.022	0.027	0.027	0.003	0.004	0.120	0.120

REAL DATA PREDICTION MSE

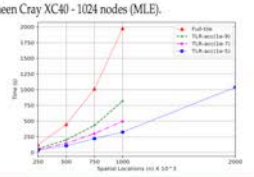
Real Data Examples	Average Prediction Mean Square Error (MSE)			
	T(-5)	T(-7)	T(-9)	FT
Wind Speed	0.0425	0.0426	0.0427	0.0426
Soil Moisture	0.0477	0.0477	0.0477	0.0477

PERFORMANCE ON DISTRIBUTED MEMORY

- Shaheen Cray XC40 - 256 nodes (MLE).

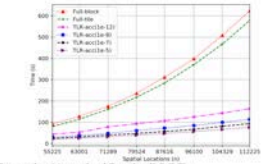


- Shaheen Cray XC40 - 1024 nodes (MLE).

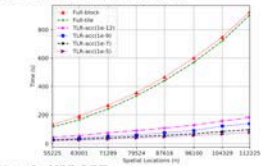


PERFORMANCE ON SHARED MEMORY

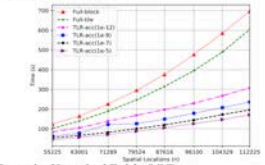
- Two-socket 18-core Intel Haswell (MLE).



- Two-socket 14-core Intel Broadwell (MLE).



- 64-core Intel KNL (MLE).



- Two-socket 28-core Intel Skylake (MLE).