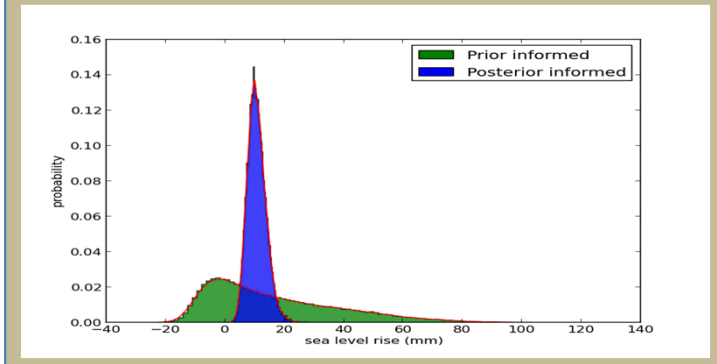
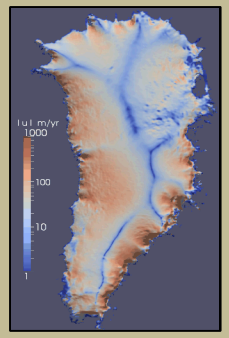
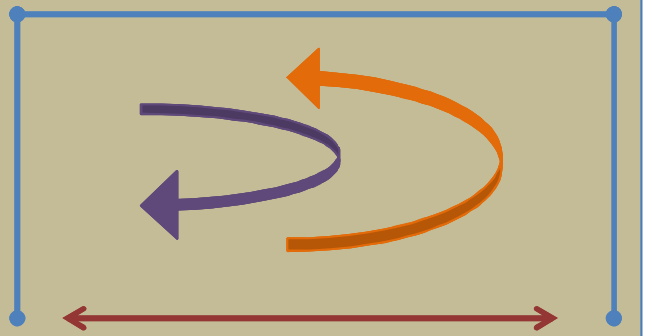
Exceptional service in the national interest

SAND2016-2884PE

Sandia National Laboratories







Science-Based Computational Modeling: *New Questions, New Challenges, and Big Data Problems*

James R. Stewart

March 28, 2016

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Acknowledgments

Thanks to many for providing material for this presentation

Brian Adams

Scott Collis

Marta D'Elia

Mike Eldred*

John Jakeman

Drew Kouri

Mauro Perego

Eric Phipps

Denis Ridzal

Laura Swiler

Irina Tezaur

Bart van Bloemen Waanders

Greg von Winckel

**Michigan Alum*

What is different?

Sandia National Laboratories

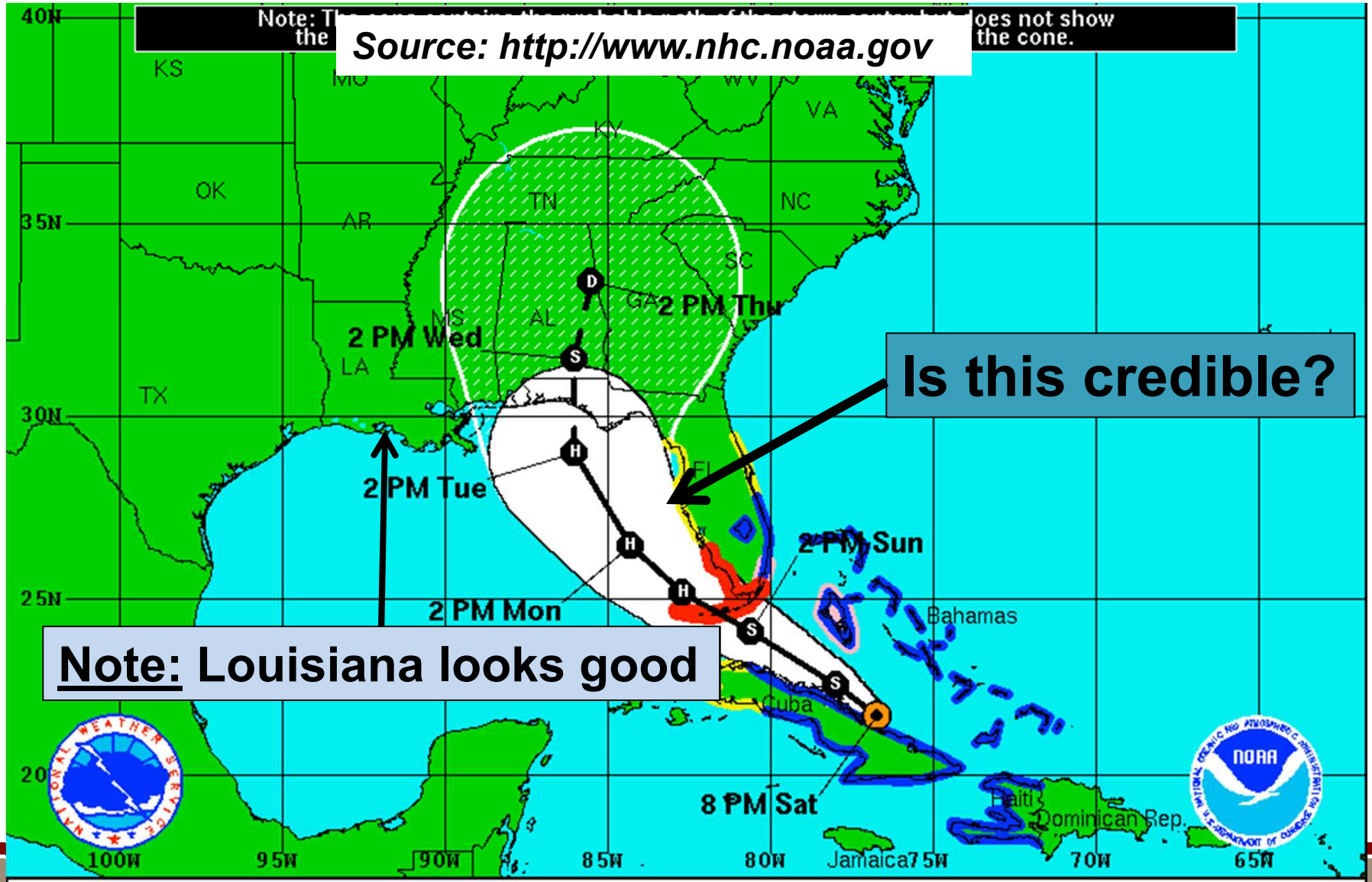
We're not solving the same problems today that we solved 10 years ago!

- Different questions are being asked
  - *What is the uncertainty in my answer?*
  - *What is the optimal design?*
  - *How credible is my result?*
  - *What are the most important parameters affecting my result?*
  - *Can I improve my models?*
  - *How can I impact decisions (e.g., design, economic, safety, etc.)*?
- Different computer hardware
- Big data!

  -> Big data problems

Tropical Storm Isaac Forecast (8 PM Sat, 8/25/12)

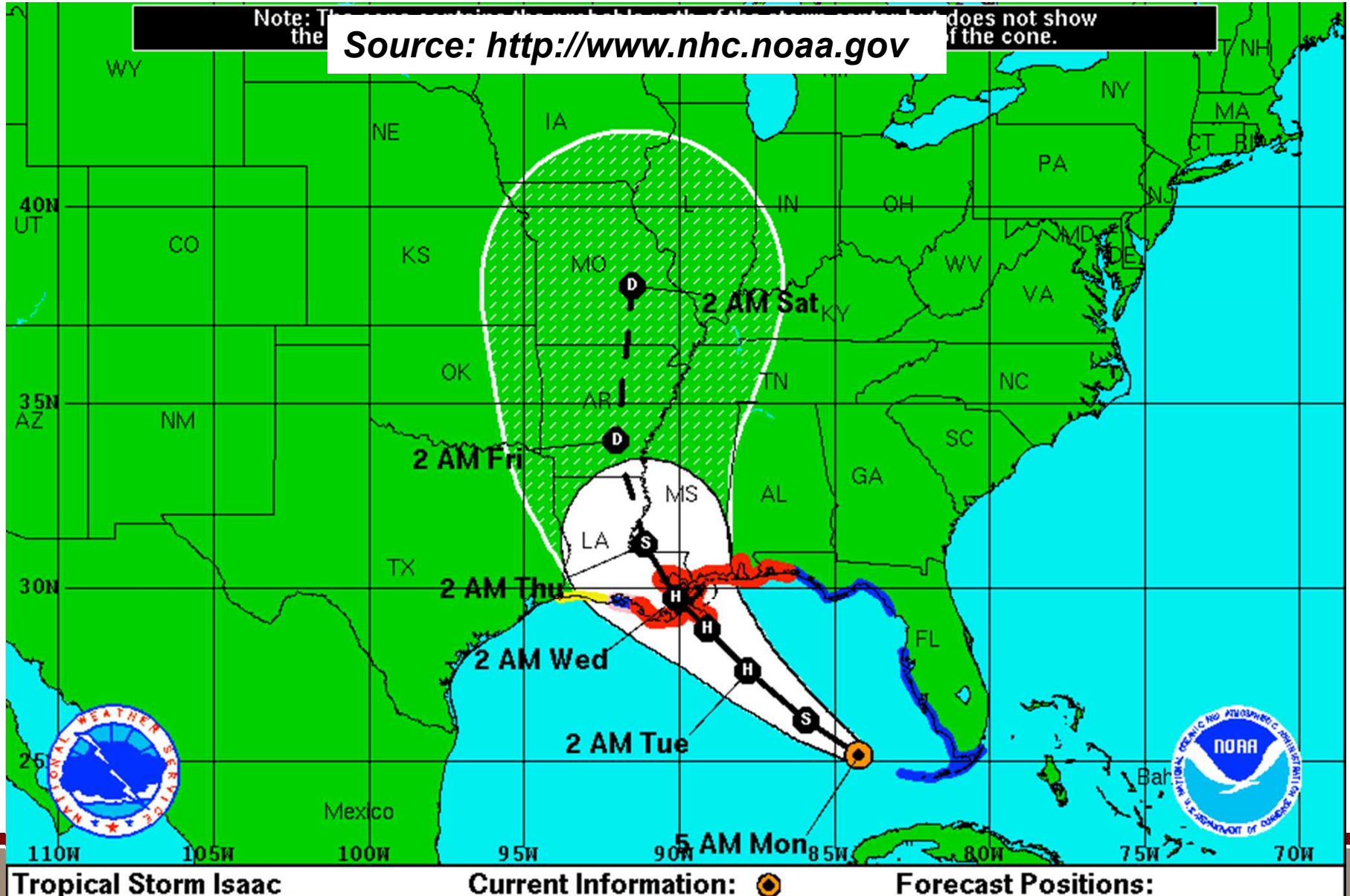
"Cone of Uncertainty"



33 Hours Later! (5 AM Mon, 8/27/12)

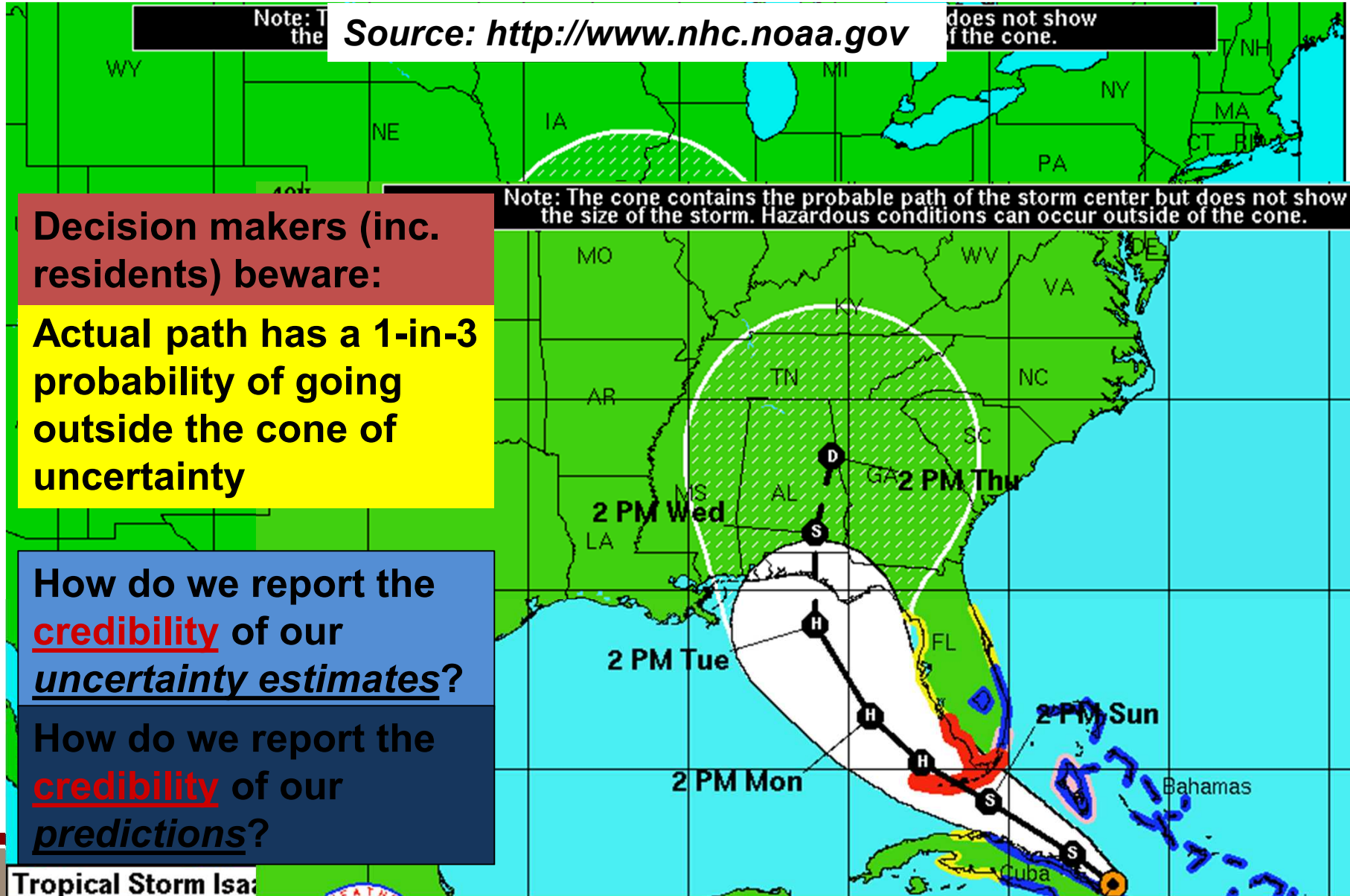
Storm Headed Straight for Louisiana

Sandia National Laboratories



Overlay of Saturday/Monday Forecasts

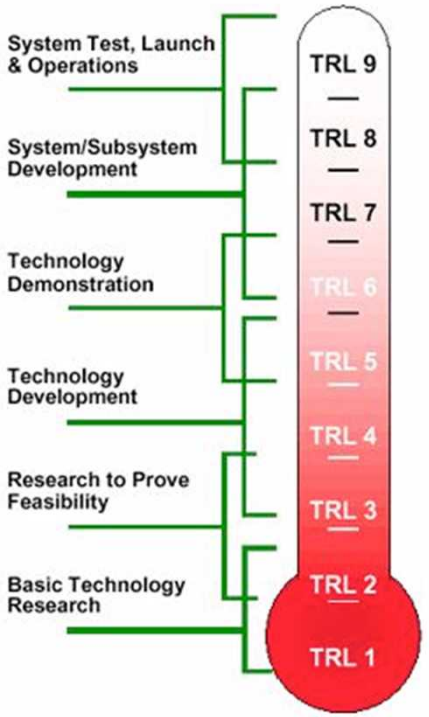
Sandia National Laboratories



Credibility Measures – Do Any Exist?
(…that apply to computational simulation)

Sandia National Laboratories

- TRL – Technical Readiness Level
  - Widely used to *assess maturity of evolving technologies* prior to incorporation into system
  - Many definitions (DOE, DoD, NASA, …)
- CMMI – Capability Maturity Model Integration
  - Carnegie Mellon Software Engineering Institute (SEI)
  - Models, appraisal methods, training *to improve process performance* (typically software development)
- Predictive Maturity *(for computational mod/sim)*
  - CAS – Credibility Assessment Scale (NASA)
  - PMI – Predictive Maturity Index (Los Alamos)
  - **PCMM – Predictive Capability Maturity Model** (Sandia)



NASA TRL Definitions

Science-Based Discovery & Prediction

It's simple, right?

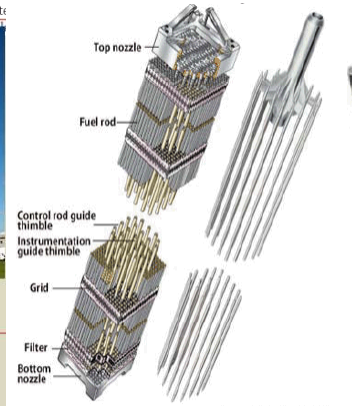
Theory
(Math Models)

Science-based engineering models typically involve:

- Multiscale
- Multiphysics
- Time-dependent, nonlinear PDEs
- Complex materials
- Turbulence
- Nonlocal effects
- Contact
- Complex chemistry
- Phase change
- Closure laws
- ***Uncertainty representation***
- ***Stochastic models***
- ***Manufacturing and design (inc. risk and uncertainty)***
- ***Decision support***
- Etc…

This wind tunnel model of F-35 features an optimized external fuel tank.

Horizontal Axis Wind Turbine ***Vertical Axis Wind Turbine***



Source: www.mhi.jp.uk

PWR Fuel Assembly and Control Rod

It's simple, right?

Sandia National Laboratories

Software
(Computational Models)

Our science & engineering codes typically require:

- Advanced discretizations (finite element, finite volume, etc)
- Geometry/mesh I/O and data structures
- Nonlinear/linear solvers
- Time integrators
- Physics-coupling modules *(data locality)*
- Adaptivity
- Performance/scalability
- Third-party libraries
- Layers of software testing
- ***Adjoints***
- ***Embedded optimization and UQ***
- ***Embedded meshing and geometry***
- Etc.

It's simple, right?

Software
(Computational Models)

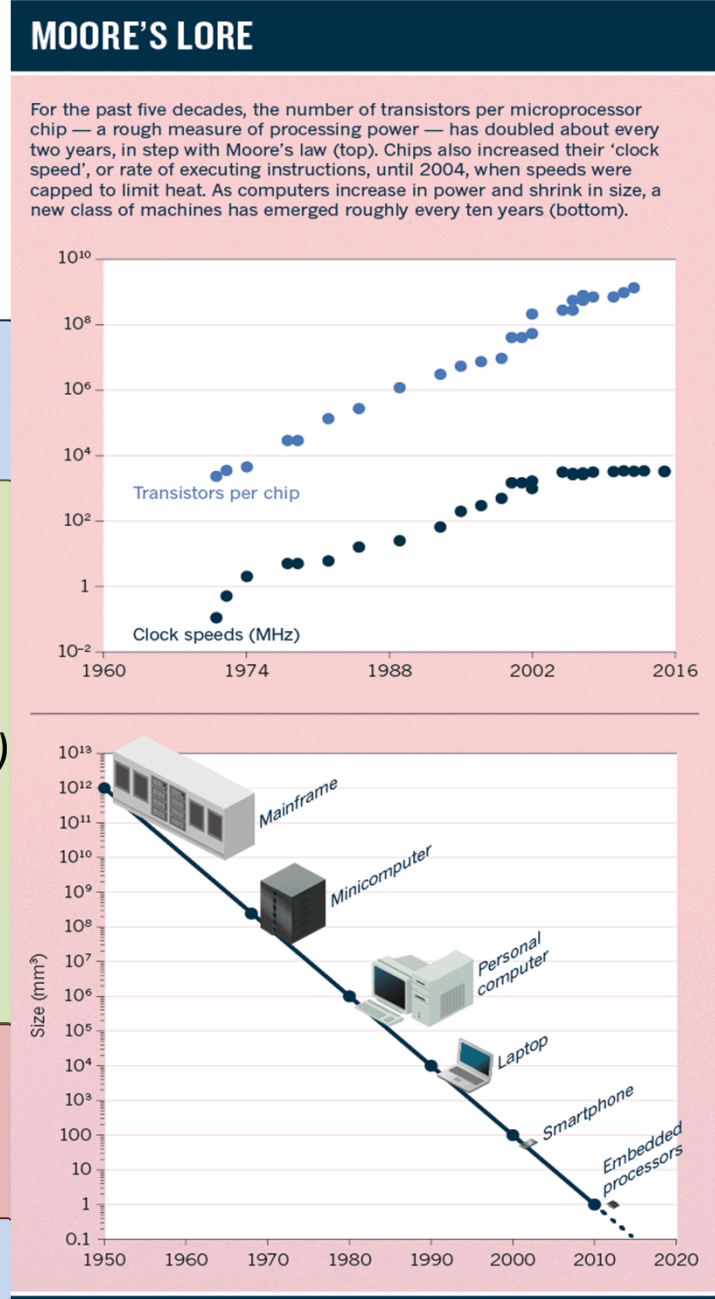
Toward Exascale:

- New programming models
  - Layers of parallel execution
  - MPI *(between nodes)* + threads *(on node)*
  - Asynchronous Many Task (AMT) *(in between)*
- New challenges
  - ***Minimizing data movement!***
  - Resilience/fault tolerance
  - Performance portability

Beyond Exascale (Beyond Moore):

- Quantum computing
- Neuromorphic (brain-inspired) computing

- ***Embedded optimization and UQ***
- ***Embedded meshing and geometry***
- Etc.



Reprinted by permission from Macmillan Publishers Ltd: Nature News, "The chips are down for Moore's Law," copyright 2016.

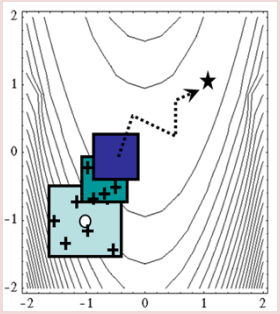
It's simple, right?

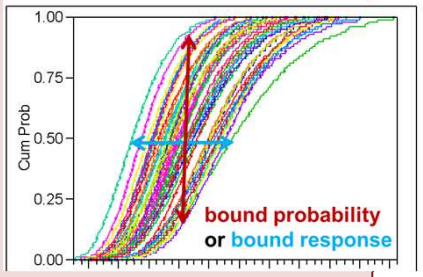
Sandia National Laboratories

Simulation
(Numerical Results /"Simulated Data")

Our numerical simulations typically require:

- Boundary/Initial conditions *(w/ uncertainties)*
- Choice of models & model parameters *(w/ uncertainties)*
- Choice of numerical algorithms & solvers
- Solver & geometric tolerances
- Mesh & time step convergence
- ***Optimization and inversion***
- ***Sensitivity analysis***
- ***Reduced-order modeling***
- ***Uncertainty quantification***
- ***Quantity-of-interest extraction***
- ***Data storage, retrieval, visualization***
- ***New (disruptive) analysis workflows***
- Etc.





(Actually, it's *complex*!)

Sandia National Laboratories

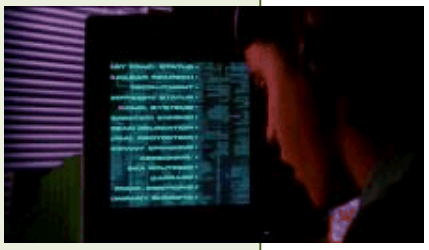
Experiments
(Field/Test Data)

Tests and experiments face new challenges:

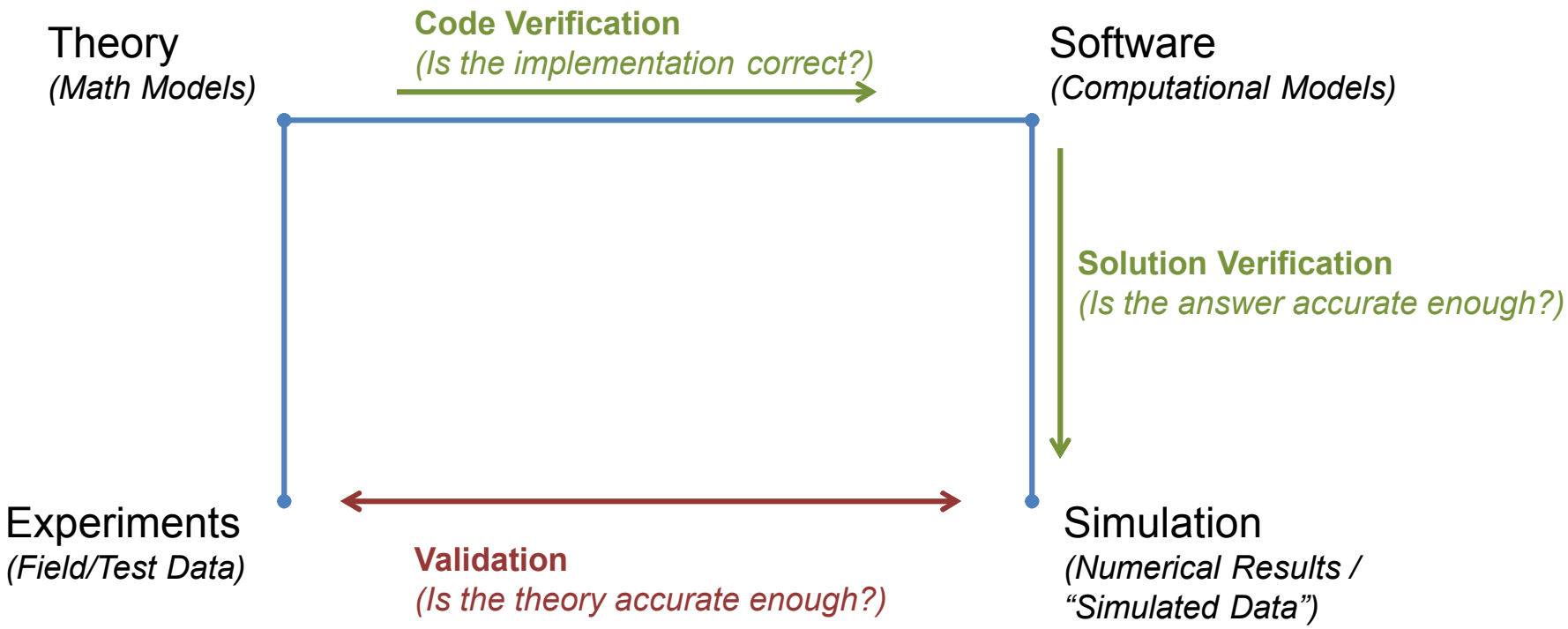
- Sparse/incomplete data
- Noisy data
- Poorly characterized uncertainties
- Quantity-of-interest extraction
- Data storage, retrieval, visualization
- Cost, safety
- Etc.

…along with new technology and opportunities:

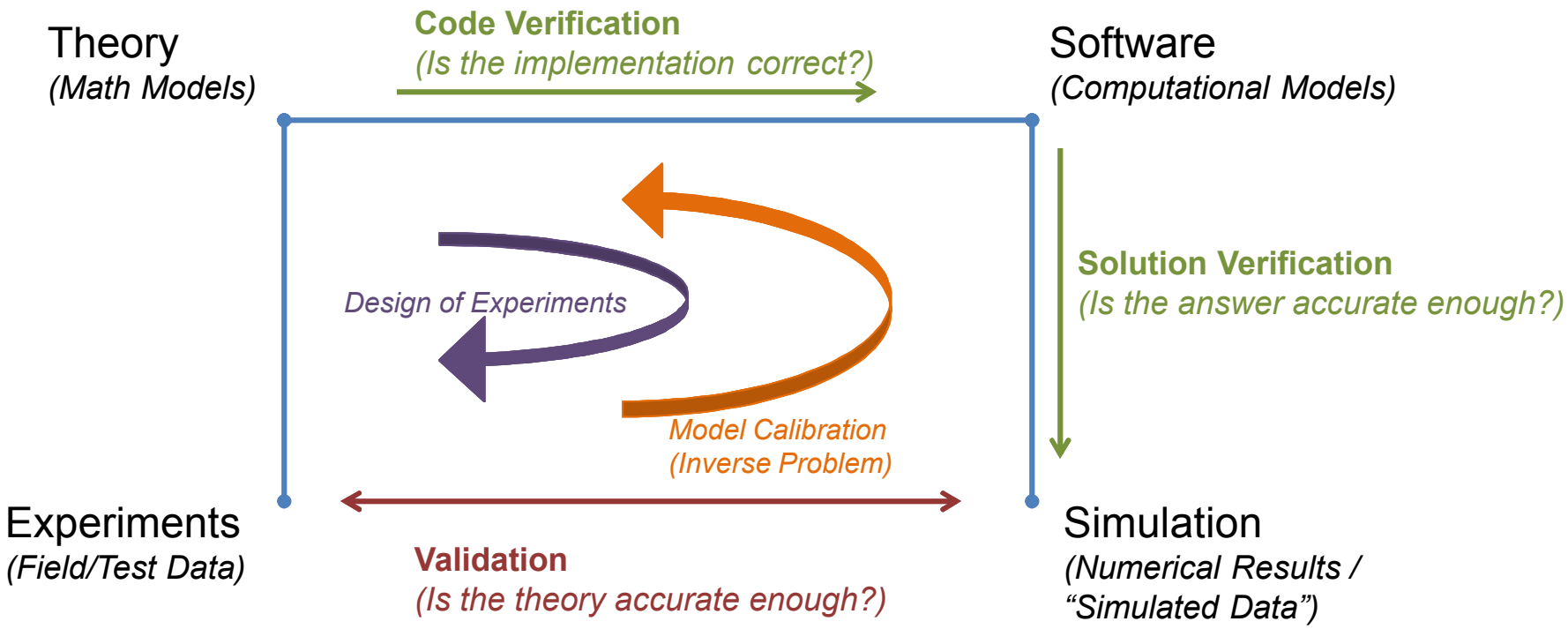
- Embedded sensors
- Non-invasive tests
- Remote sensing
- ***Data overload (and at the wrong scales…)!***
- Etc.



V&V Connects the Dots…



Using Data to Improve Models & Experiments



Design of Experiments: *Simulation data steers new experiments*

Model Calibration: *Experimental data used to estimate model parameters*

New challenges in ***optimization*** and ***data science***

Sandia National Laboratories

Beyond Forward Simulation…

Our systems will be "swimming in data"

Big Data Problems

Sandia National Laboratories

Beyond Forward Simulation…

Our systems will be "swimming in data"

(Big Data) Problems

In some cases we'll be "data starved"

Big (Data Problems)

Back to the Question of Models…

Sandia National Laboratories

Ask an engineer…

- "It's all about the *(science-based)* models"

From dictionary.com:

Engineering

The art or science of making practical application of the knowledge of pure sciences, as physics or chemistry, as in the construction of engines, bridges, buildings, mines, ships, and chemical plants.

Back to the Question of Models…

Sandia National Laboratories

Ask an engineer…

- "It's all about the *(science-based)* models"

Ask a statistician…

- "It's all about the data"

The "truth" is…

- *(…later…)*

Simulations today must go *"Beyond Forward Solve…"*

Sandia National Laboratories

Simulation
(Numerical Results /"Simulated Data")

Our numerical simulations typically require:

- Boundary/Initial conditions *(w/ uncertainties)*
- Choice of models & model parameters *(w/ uncertainties)*
- Choice of numerical algorithms & solvers
- Solver & geometric tolerances
- Mesh & time step convergence
- ***Optimization and inversion***
- ***Sensitivity analysis***
- ***Reduced-order modeling***
- ***Uncertainty quantification***
- ***Quantity-of-interest extraction***
- ***Data storage, retrieval, visualization***
- ***New (disruptive) analysis workflows***
- Etc.

Usually requires an ensemble of computations

Optimization and Inversion:
ROL: Rapid Optimization Library

Sandia National Laboratories

- ROL is a **Trilinos package for large-scale continuous optimization**, a.k.a. nonlinear programming (NLP).
- Released in *Trilinos 12.2* in August, 2015

Information	Size of parameter space				Methods
	1	10	10^3	10^6	
Function samples (incl. finite diff's)	100	10,000	∞	∞	Global search or steepest descent
Analytic gradients (hand-coded or AD)	50	100	200	1,000	Quasi-Newton
Analytic Hessians (hand-coded or AD)	50	50	50	50	Newton-Krylov

- Derivative-based methods
- **Embedded and matrix-free methods:**
  - Direct access to application data structures: vectors, etc
  - Direct use of application methods: (non)linear solvers, etc

ROL Supports Four Basic Problem Types

Generic PDE-constrained problem

$$\min_{z,u} J(z,u) \qquad c(z,u) = 0 \qquad z_\ell \leq z \leq z_u$$

simulation variable

optimization variable

Type-U

Unconstrained problems

$$\min_x f(x)$$

Type-B

Bound constrained

$$\min_x \; f(x) \quad \text{subject to} \quad a \leq x \leq b$$

Type-E

Equality constrained

$$\min_x \; f(x) \quad \text{subject to} \quad c(x) = 0$$

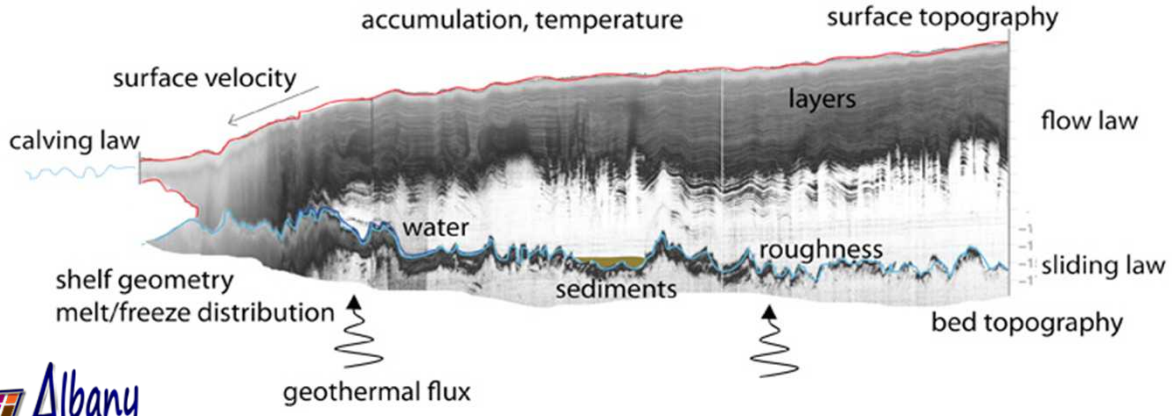
Type-EB

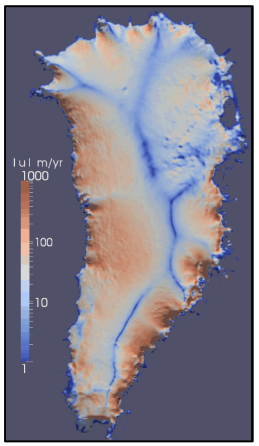
Equality + Bounds

$$\min_x \; f(x) \quad \text{subject to} \quad c(x) = 0, \quad a \leq x \leq b$$

Example: Greenland Ice Sheet Modeling

Balance between accumulation of new ice and loss of ice to the ocean is influenced by





Greenland surface ice velocity

- Sandia's ***Albany/FELIX**** First Order Stokes solver; QoI: ***Ice Surface Velocity***
- NCAR's ***Community Ice Sheet Model (CISM)****;* QoI: ***Sea Level Rise***

There are several sources of uncertainty, most notably:

- Climate forcings (e.g., surface mass balance).
- Basal friction ➔ *our initial focus*
- Bedrock topography
- Geothermal heat flux
- Model parameters

**FELIX*="Finite Elements for Land Ice eXperiments"

Large-Scale Optimization with ROL:
Inverting for the Basal Friction

Sandia National Laboratories

GOAL: Invert for unknown model parameters (**basal friction**) to match observational data (**observed surface velocity of ice**).

STRATEGY: Solve large scale PDE constrained optimization problem.

ALGORITHM: Reduced gradient method, where gradient is computed using adjoints and optimization is solved using BFGS.

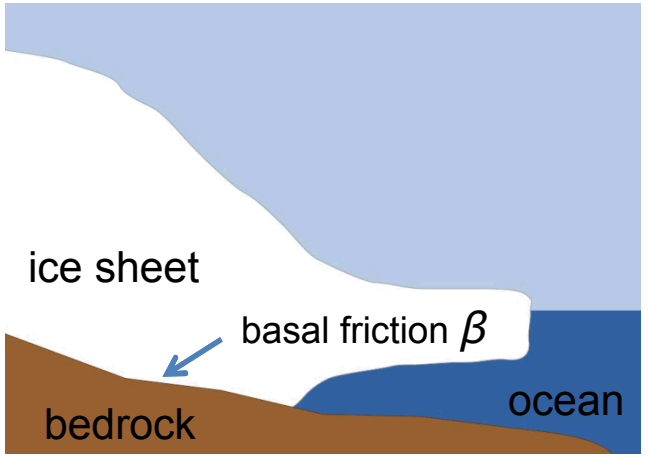
TOOLS: ROL for optimization and Albany/FELIX for assembly, Trilinos for solvers.

SIGNIFICANCE: Basal friction determines ice sheet sliding and affects ice sheets mass balance and sea level rise.

OPTIMIZATION PROBLEM:

$$\min \mathcal{J}(\boldsymbol{\beta}) = \int_{\Sigma} \frac{1}{\sigma_u^2} \left|\mathbf{u}(\boldsymbol{\beta}) - \mathbf{u}^{\text{obs}}\right|^2 ds + \alpha \int_{\Sigma} |\nabla \boldsymbol{\beta}|^2 ds$$

$\mathbf{u}(\boldsymbol{\beta})$ → ice velocity solution of the flow model (First Order Stokes) as a function of the basal friction $\boldsymbol{\beta}$



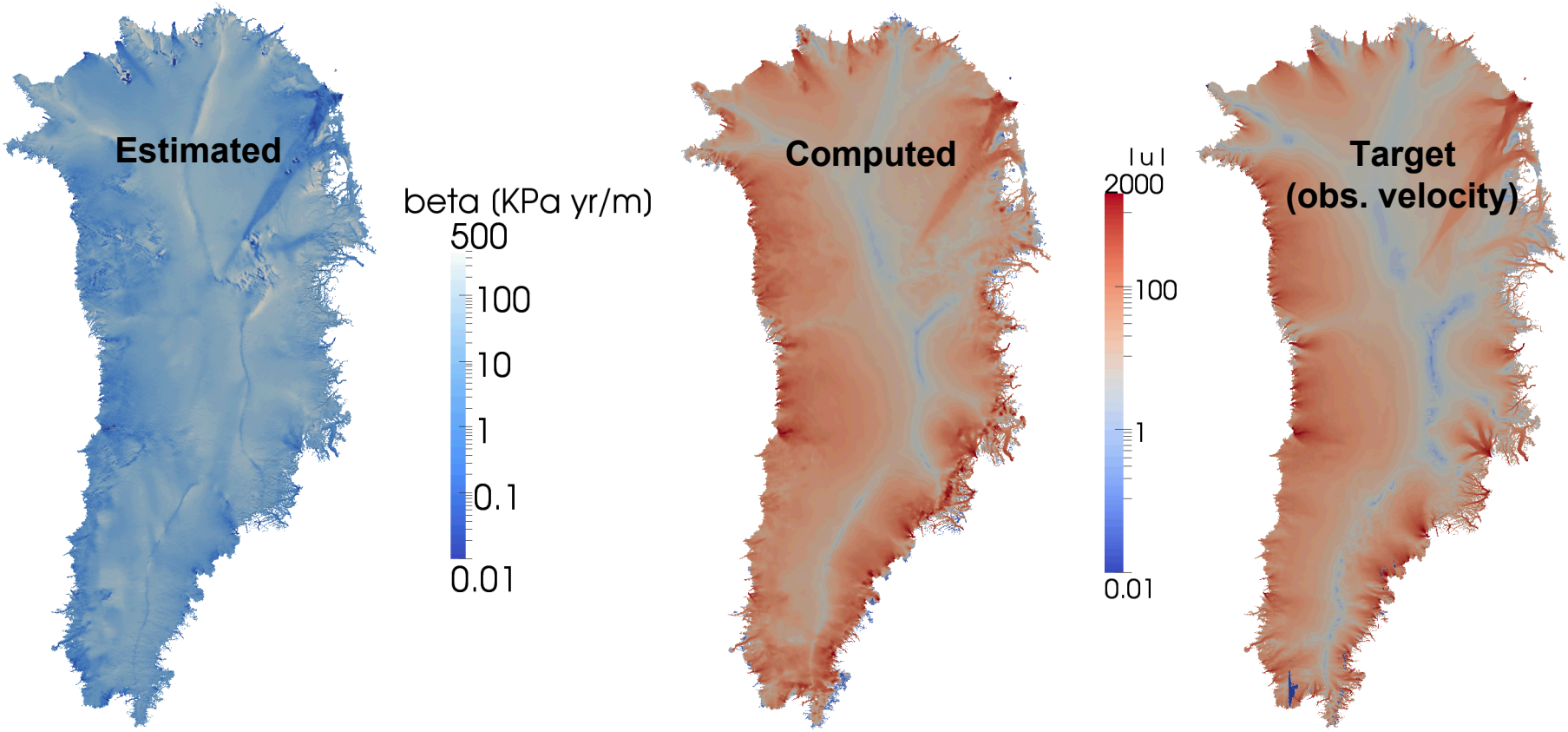
ROL Albany Trilinos

Results: Basal Friction Inversion for Greenland Ice Sheet

ROL ALGORITHM:
- Limited Memory BFGS
- Backtrack line search

Problem size:
Unknowns: 36 M
 # Parameters: 1.6 M

Geometry: *Morlighem et al., Nature Geo., 2014*
Resolution: 1km.

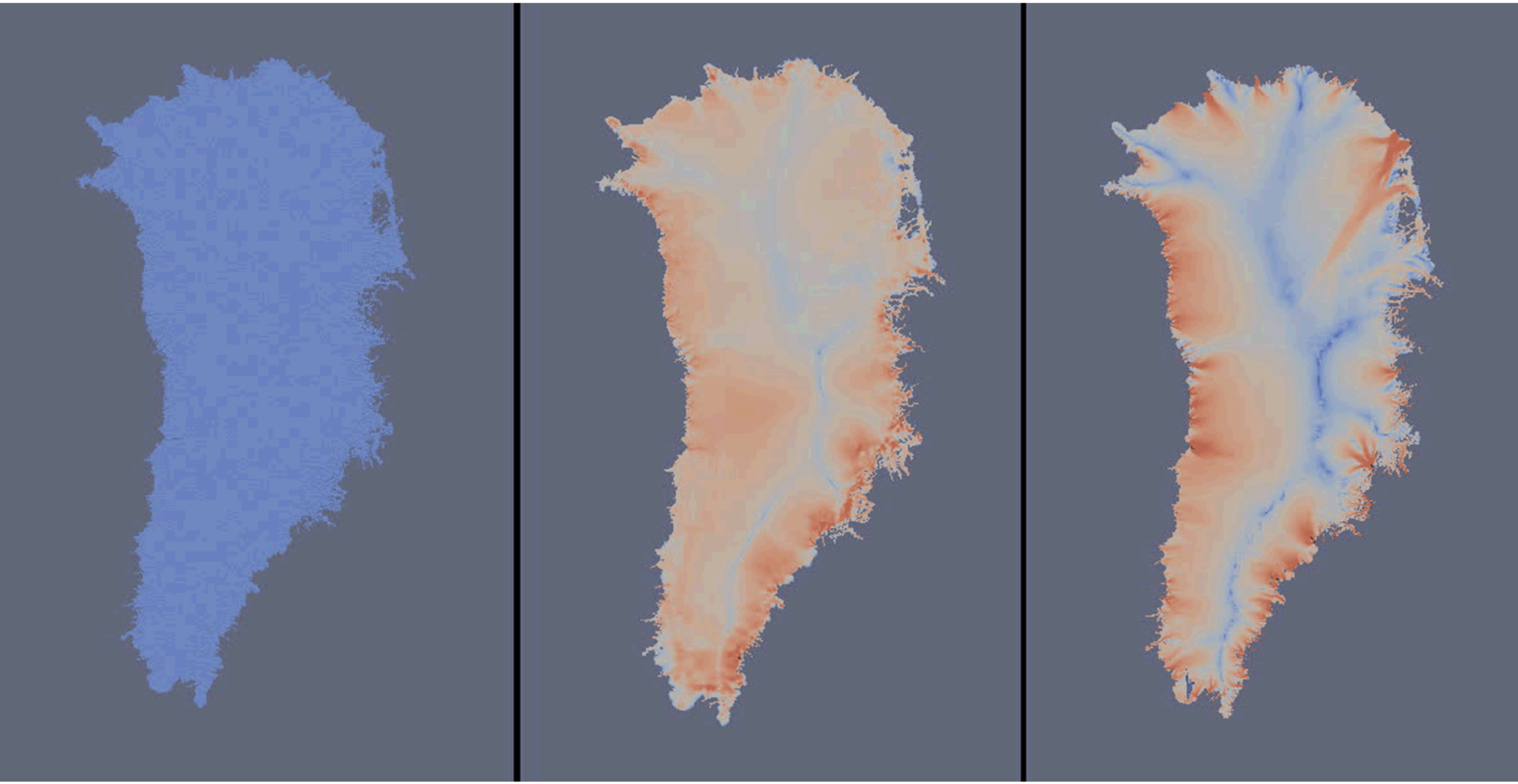


Recovered basal friction *(kPa yr/m)*

Surface velocity magnitude *(m/yr)*

1st 25 iterations of optimization

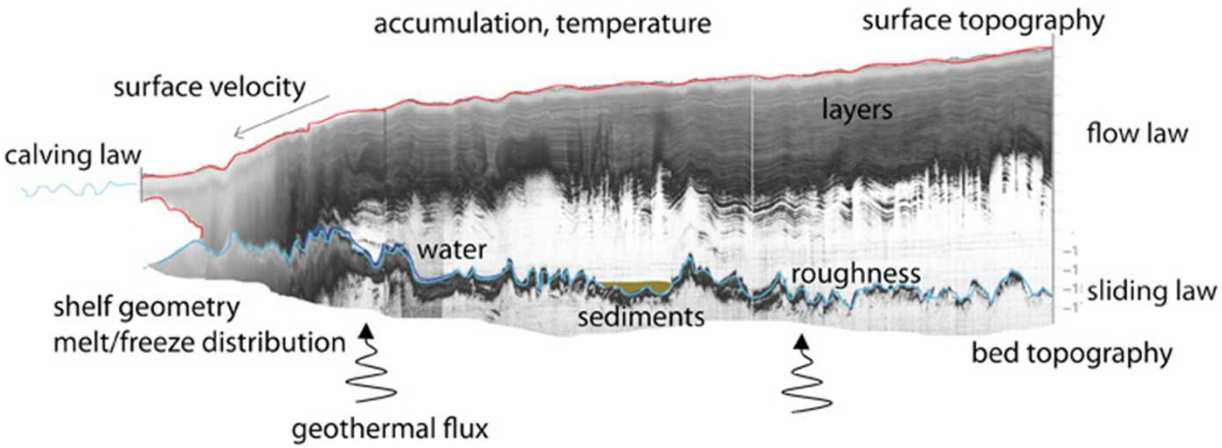
(121 total for convergence)

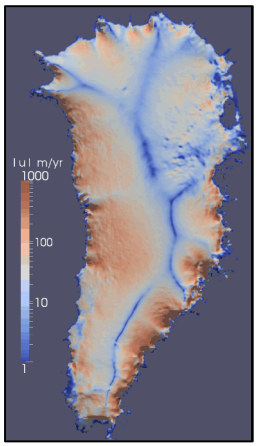


Greenland Ice Sheet Modeling:

Consider the Uncertainty in Basal Law

Balance between accumulation of new ice and loss of ice to the ocean is influenced by





Greenland surface ice velocity

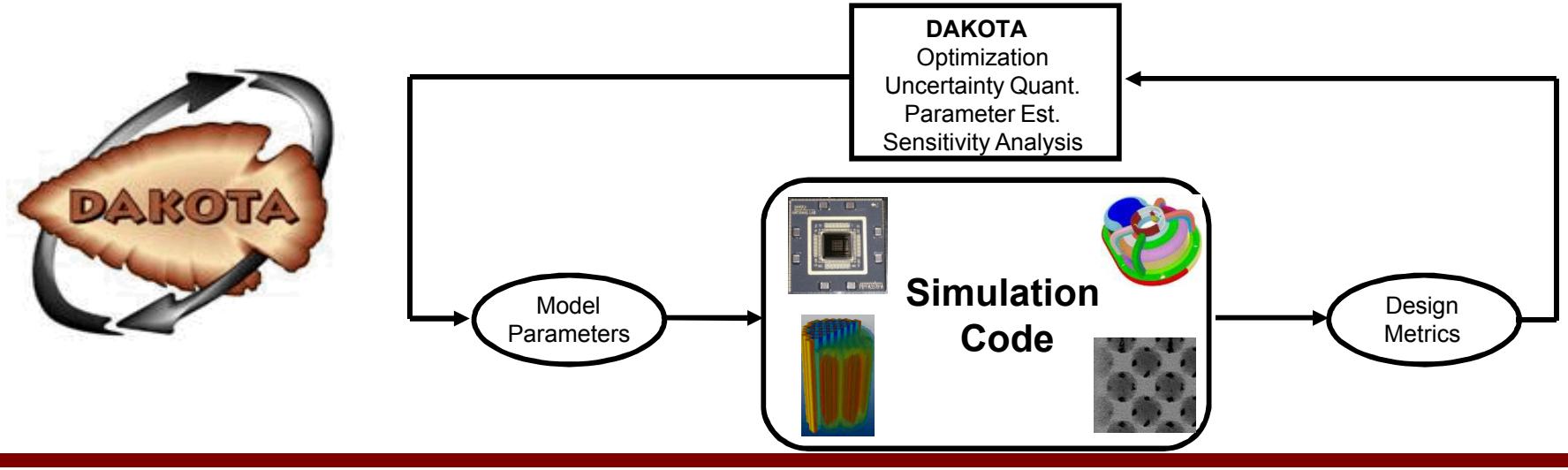
Procedure:

Start with the basal friction from ROL's deterministic inversion, then do the following:

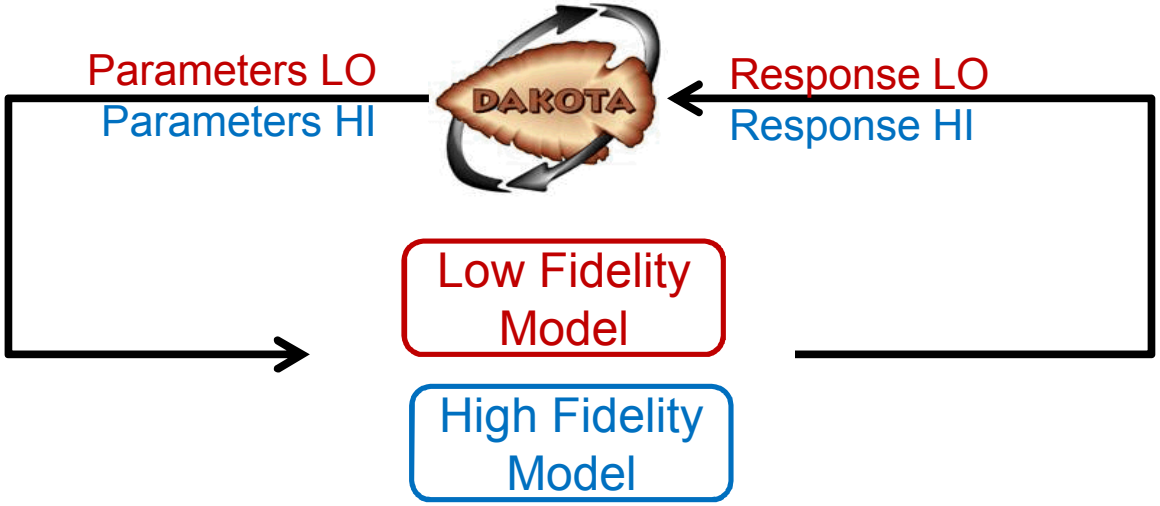
1. Spatial dimension reduction via Karhunen-Loève expansion (KLE)
2. Statistical emulation via polynomial chaos expansion (PCE)
3. Emulator-based **Bayesian inference** using preconditioned Markov Chain Monte Carlo (MCMC)

Dakota

- **Dakota:** *Design exploration and UQ for science and engineering*
- A suite of iterative mathematical and statistical methods that interface to computational models.
- Dakota makes iterative parametric analysis practical for *black-box simulations* to answer questions regarding:
  - **Sensitivity:** Which are the crucial factors/parameters?
  - **Uncertainty:** How safe, reliable, or robust is my system?
  - **Optimization:** What is the best performing design or control?
  - **Calibration:** What models and parameters best match data?

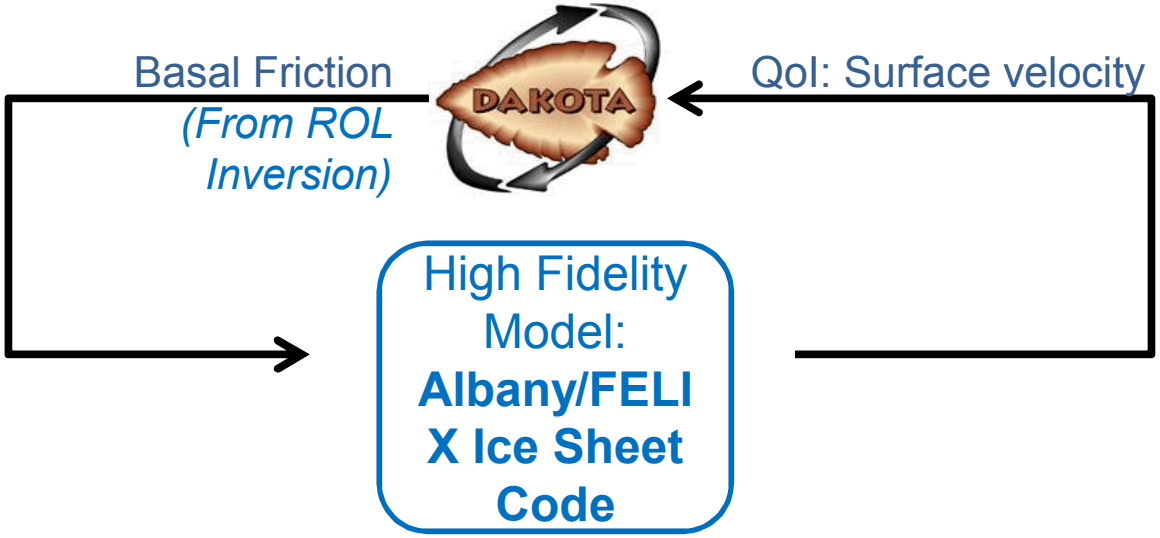


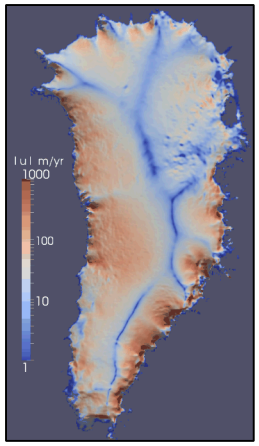
Multi-Fidelity UQ with Dakota



- High fidelity simulation is expensive – want fewer of these
- Low fidelity model is (relatively) cheap – run more of these
- Cycle between them to converge on (high-fidelity) statistics at lowest cost

Construction of PCE Emulator





- Dakota constructs PCE of Quantity of Interest (surface velocity)
  - 3rd-order polynomial
  - Required 286 Albany/FELIX solves
- ***This becomes the low-fidelity model*** (our "emulator")

KLE of Spatial Random Field for Basal Friction

Karhunen-Loeve expansion (KLE):

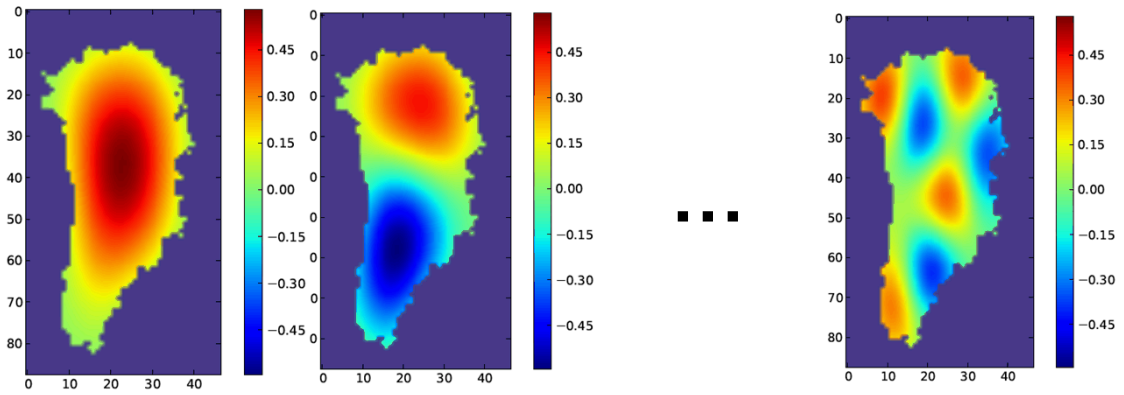
$$\beta(x,\omega) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\,\xi_i(\omega)\,\phi_i(x)$$

Assume analytic spatial covariance kernel (***Gaussian***) for random field

$$C(r_1, r_2) = e^{-(r_1 - r_2)^2 / L^2}$$

and integrate over domain for eigenvalues, eigenvectors (λ_i, ϕ_i).
Length scale (L) balances feature resolution vs. #modes required in truncated KLE.

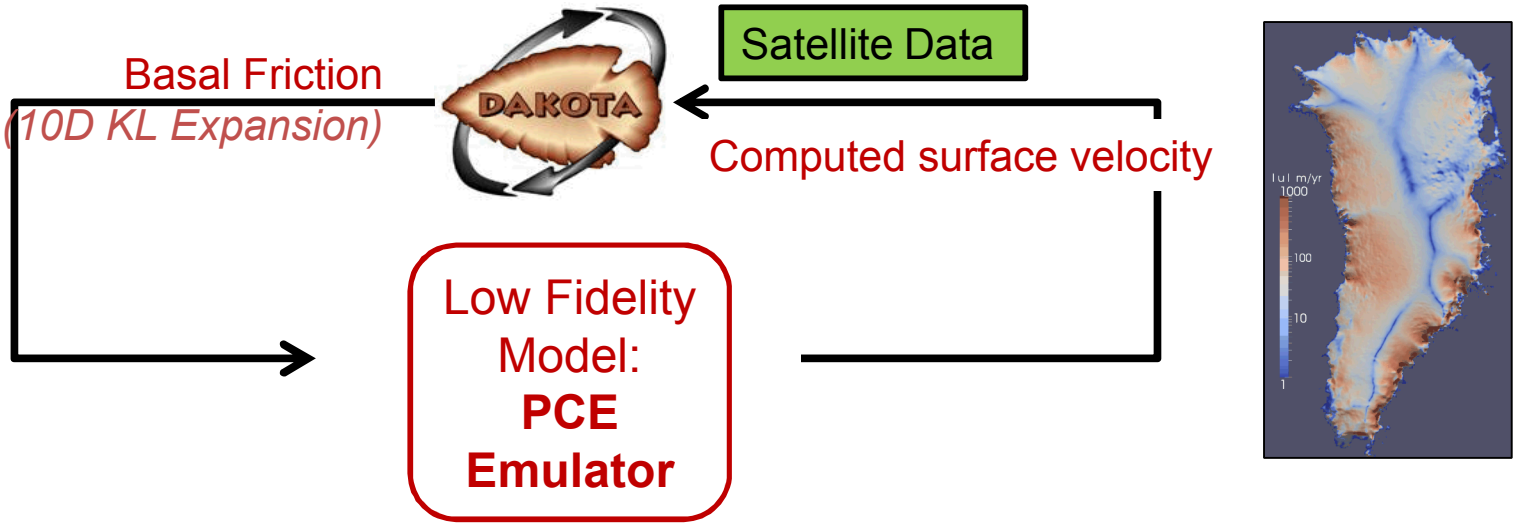
KLE modes (first 10):



Dimension-reduced inference of KLE coefficients:

Assume uniform priors on KLE coefficient distributions of KLE coefficients $\xi_i \sim [-1,1]$

Bayesian Calibration of Basal Friction



- Emulator-based Bayesian inference using preconditioned MCMC
- Uses Dakota's QUESO package developed by UT-Austin

$$\pi(\theta \mid d) \propto \pi(\theta) L(d \mid \theta)$$

Posterior distribution

Model parameters

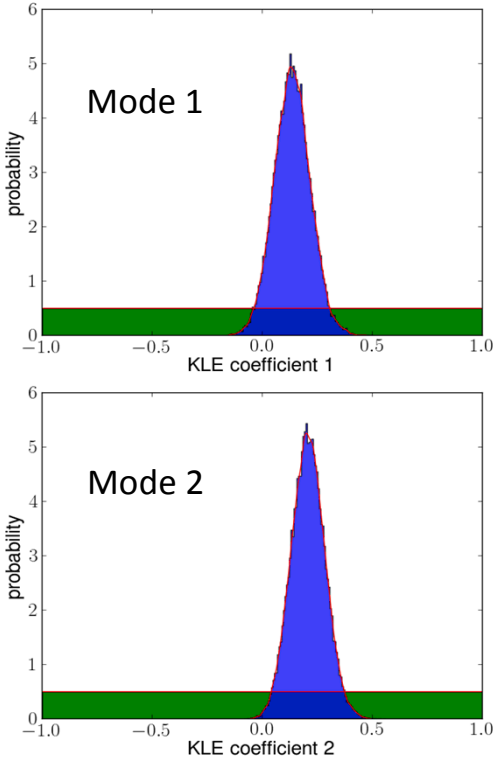
Observed Data

Prior parameter distribution

Likelihood function which Incorporates the model

Prediction of Sea-Level Rise using Calibrated Basal Friction

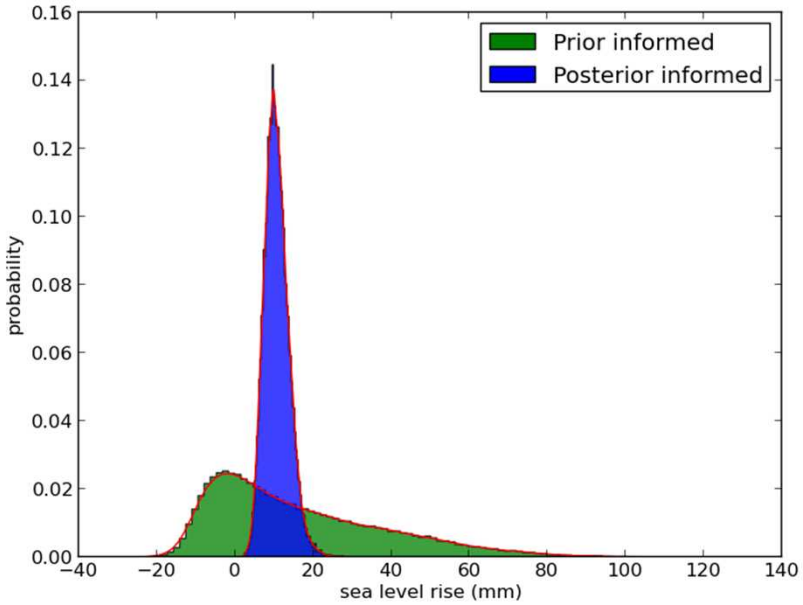
KLE mode priors (green) and (calibrated) posteriors (blue)



Forward

Propagation (NCAR's **CISM**: Community Ice Sheet Model)

PDF of Sea Level Rise (SLR) after 50 yrs



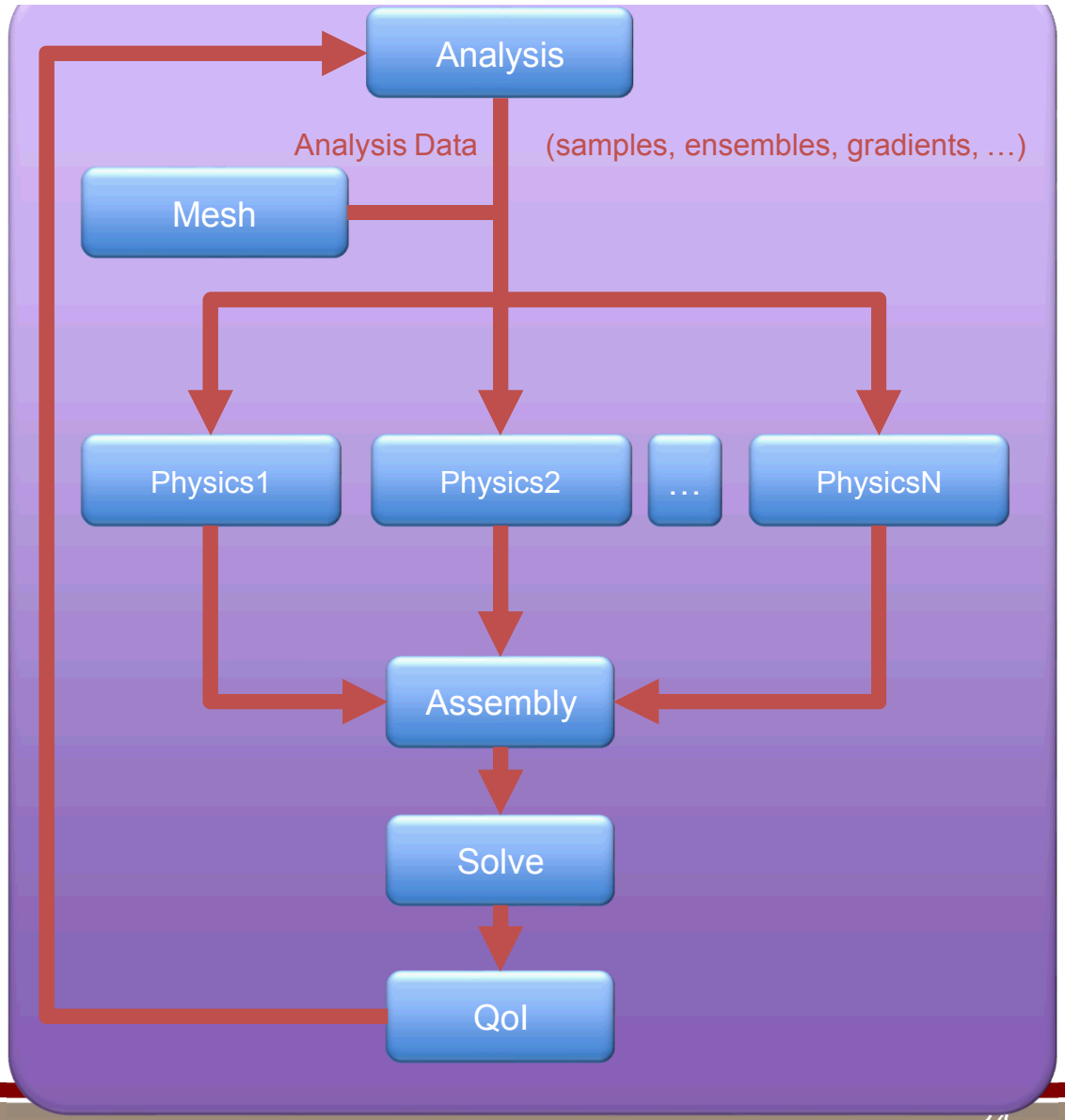
Possible paths for adaptation:

- Increase *spatial resolution* (reduce KLE truncation error by increasing # of modes)
- Improve *UQ accuracy* (reduce PCE error by increasing # of CISM simulations)

Toward Exascale: Many-Task Workflows

Kokkos, Darma, …

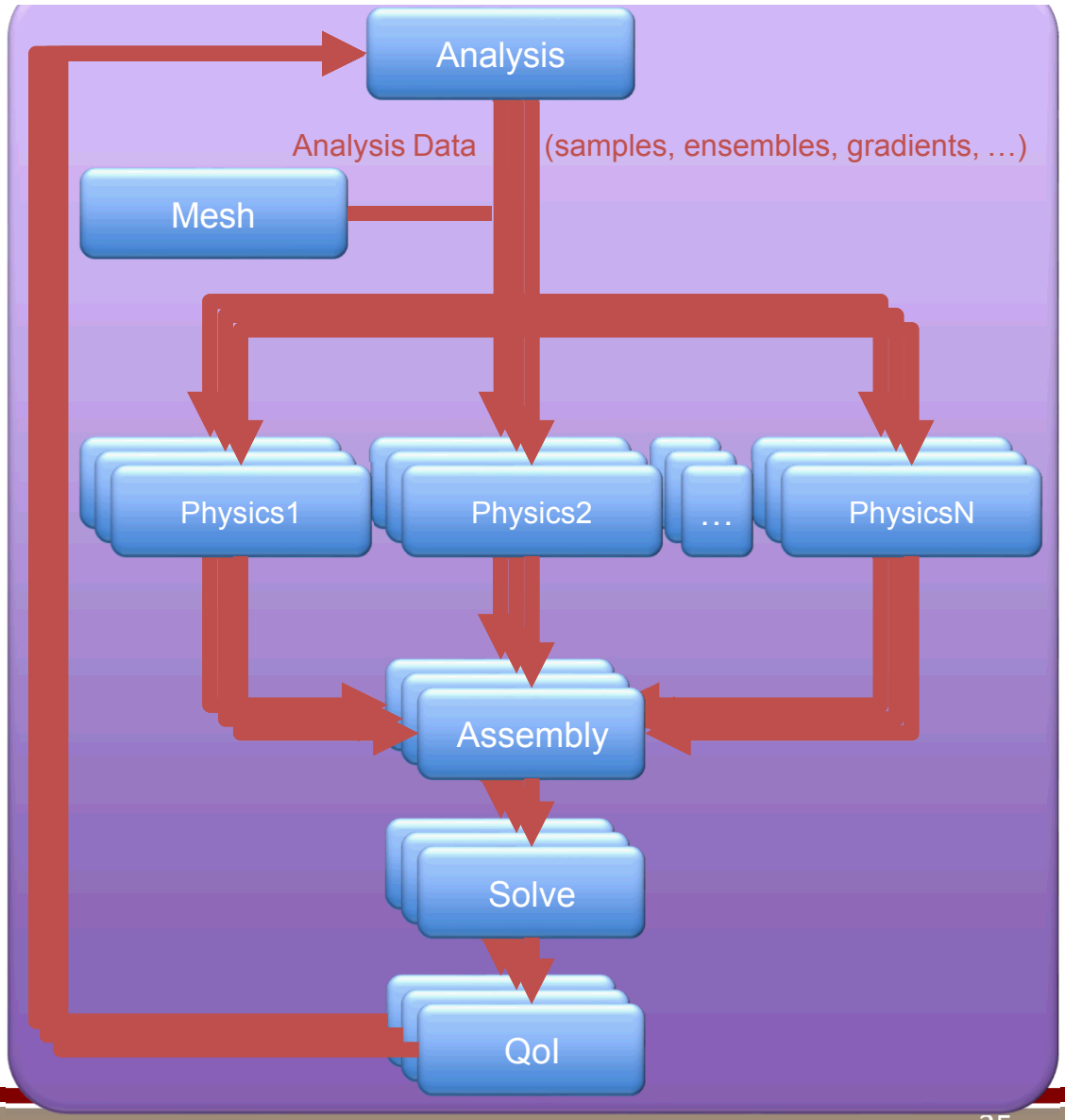
- Each box represents a family of fine-grained tasks
- Using **C++ templates and operator overloading**, we can high-level analysis data
  - Derivatives *(**Trilinos:Sacado**)*



Toward Exascale: Many-Task Workflows

Kokkos, Darma, …

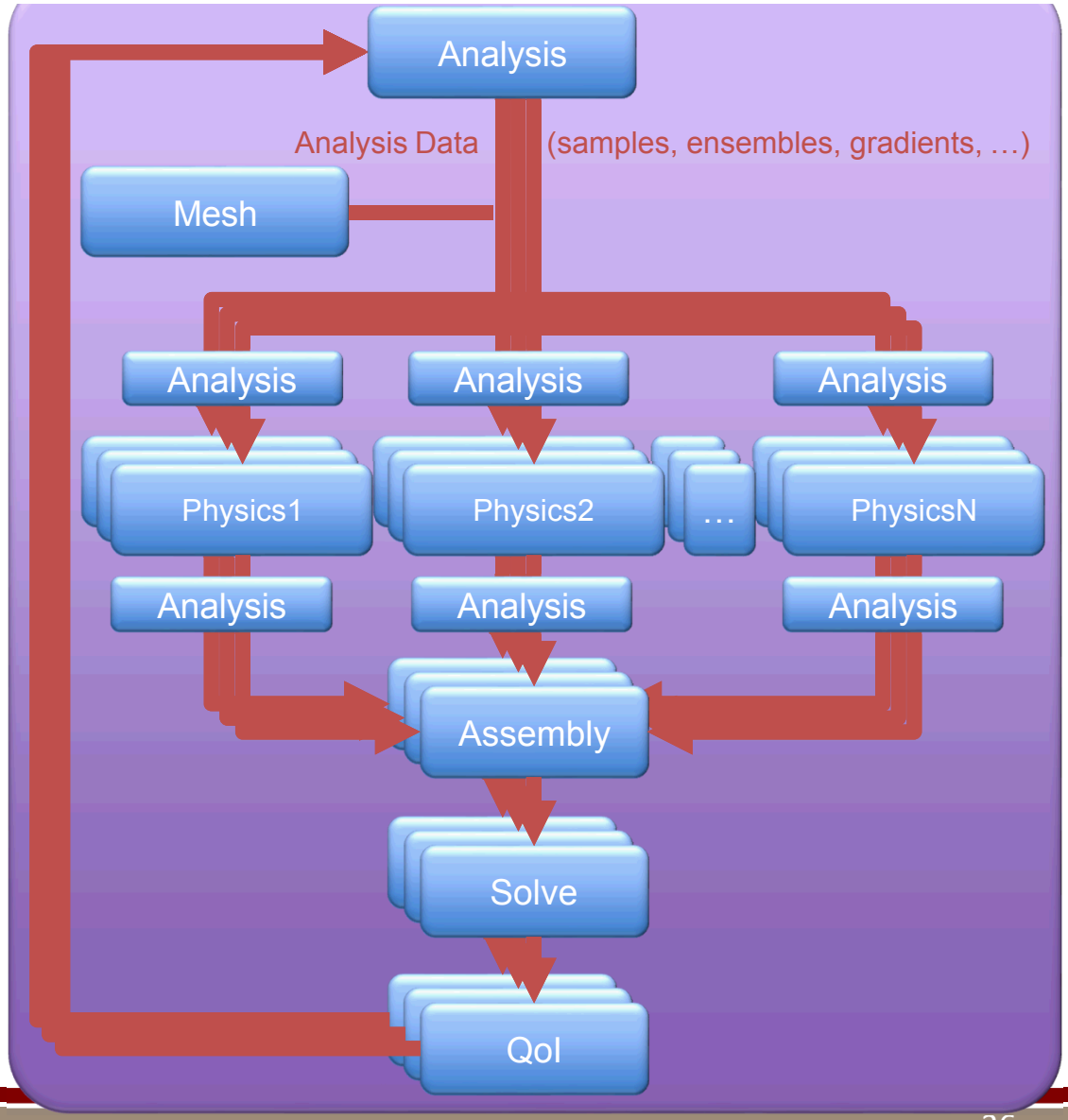
- Each box represents a family of fine-grained tasks
- Using C++ templates and operator overloading, we can high-level analysis data
  - Derivatives
  - Ensembles of UQ samples *(**Trilinos:Stokhos**)*



Toward Exascale: Many-Task Workflows

Kokkos, Darma, …

- Each box represents a family of fine-grained tasks
- Using C++ templates and operator overloading, we can high-level analysis data
  - Derivatives
  - Ensembles of UQ samples
- We can **insert analysis tasks anywhere**, e.g.,
  - LHS UQ for Physics Task 1
  - PCE UQ for Physics Task 2
  - Etc.



Acceleration of Stochastic Sampling Methods

Sandia National Laboratories

GOAL: determine statistical information about an output of interest that depends on the solution of a stochastic PDE

$$-\nabla \cdot (A(\mathbf{x},\mathbf{y}))\nabla u) = f$$

Output: $F(\mathbf{y}) = \|u(\cdot,\mathbf{y})\|_2$

Quantity of interest: $\mathbb{E}[F(\mathbf{y})] = \int_\Gamma F(\mathbf{y})\rho(\mathbf{y})d\mathbf{y}$

Anisotropic diffusion problem

$\mathbf{y}$ ⟶ N-dimensional random vector with pdf ρ

A ⟶ uncertain parameter: diffusivity tensor $A(\mathbf{x},\mathbf{y}) = diag(a(\mathbf{x},\mathbf{y}), a_2, a_3)$

$$a(\mathbf{x},\mathbf{y}) = \check{a} + \hat{a}\exp\left\{\sum_{n=1}^{N}\sqrt{\lambda_n}b_n(\mathbf{x})y_n\right\}$$ truncated KL expansion

Stochastic collocation (SC) methods:

- Compute discrete approximations for a set of M samples: $\{u_h(\mathbf{x},\mathbf{y}_m)\}_{m=1}^{M}$
- Construct a polynomial over random dim: $u_h^{\mathrm{SC}}(\mathbf{x},\mathbf{y}) = \sum_{m=1}^{M} u_h(\mathbf{x},\mathbf{y}_m)\psi_m(\mathbf{y})$

FULLY NON-INTRUSIVE: use PDE solvers as black boxes

VERY EXPENSIVE for large-scale problems!

HOW TO IMPROVE PERFORMANCE??

Acceleration of Stochastic Sampling Methods

Sandia National Laboratories

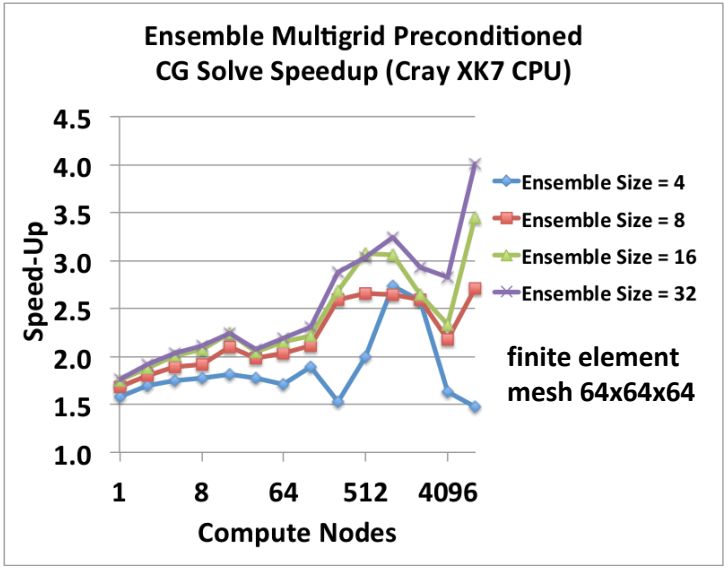
IDEA: in scientific simulations there is a huge amount of data that can be reused (computational mesh, matrix graph, …)

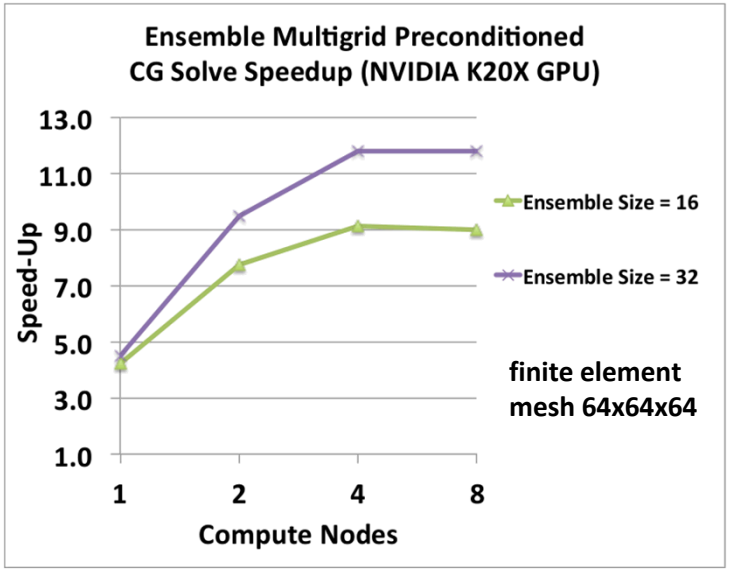
STRATEGY: **REUSE** data propagating multiple samples at a time

ALGORITHM: **EMBEDDED ENSEMBLE PROPAGATION**

- propagation of ensembles of samples of size S
- each sample dependent quantity is replaced with a length-S array

CONSEQUENCES: reduction of **COMPUTATION, MEMORY USAGE, MEMORY TRAFFIC**





ISOTROPIC Diffusion Problem → for every sample the linear solver #its is **the same**

Acceleration of Stochastic Sampling Methods

Sandia National Laboratories

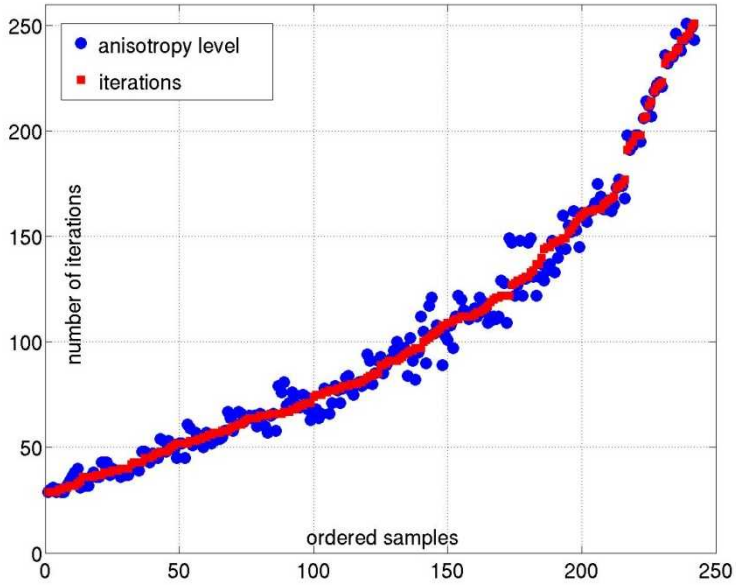
ISOTROPIC diffusion: the way samples are grouped together is not relevant

ANISOTROPIC diffusion: the convergence behavior **DEPENDS** on the sample

- the way samples are grouped **affects the convergence** of the ensemble system
- **idea:** group together samples with similar #its
- **fact:** similar **ANISOTROPY LEVEL** → similar **#ITS**

STRATEGY: **1.** order the samples for increasing anisotropy level

2. group them in ensembles of size S



anisotropy level: $\left\| \frac{\lambda_{\max}(A(\mathbf{x},\widetilde{\mathbf{y}}))}{\lambda_{\min}(A(\mathbf{x},\widetilde{\mathbf{y}}))} \right\|_{\infty}$

Anisotropic Diffusion Problem
64x64 finite element mesh
AMG Preconditioned CG

Comparison of the orderings based on the anisotropy level and the #its

Acceleration of Stochastic Sampling Methods

ASSESSING THE COMPUTATIONAL SAVING

Note: the #its of the ensemble system is always greater than the one of each sample within the ensemble → **increase in computational work** induced by the ensemble propagation

$$R = \frac{S \sum_{i=1}^{\overline{S}} \mathrm{ITS}_i}{\sum_{k=1}^{M} \mathrm{its}_k}$$

ITS: #its for the i^{th} ensemble
its: #its for the k^{th} sample

covariance	S	R ordering	R no-ordering
Gaussian	8	1.374	1.793
Gaussian	16	1.469	2.197
Gaussian	32	1.652	2.852
Exponential	8	1.274	1.448
Exponential	16	1.337	1.673
Exponential	32	1.427	1.847
γ-Exponential	8	1.217	1.503
γ-Exponential	16	1.272	1.794
γ-Exponential	32	1.384	2.223

Note: the increase in work is **mitigated** by the computational savings induced by the ensemble propagation → the achieved speed-up is reduced by a factor of R

Back to the Question of Models…

Sandia National Laboratories

Ask an engineer…

- "It's all about the *(science-based)* models"

Ask a statistician…

- "It's all about the data"

The "truth" is…

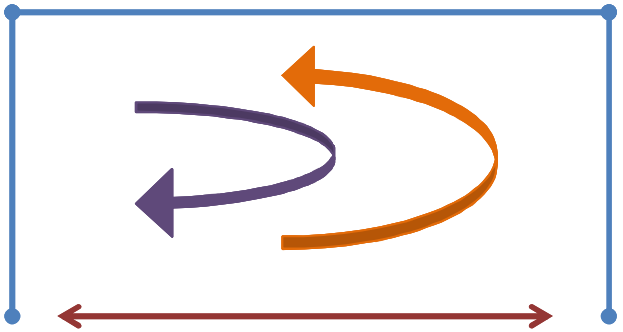
- **Both are critical to developing science-based models!**

Closing Thoughts

Sandia National Laboratories

We're not solving the same problems today that we solved 10 years ago!

- "Beyond forward solve…" – single-point solutions and workflows no longer sufficient
- Fusion of models and data



Computer architectures are changing!

- Exascale is on the horizon
- Beyond Moore's Law

Big data problems!

- New communities of experts

Big opportunities lie ahead!

https://trilinos.org

https://trilinos.org/packages/rol/

https://dakota.sandia.gov