From Physical Modelling to Big Data Analytics: Examples and Challenges

Ben Leimkuhler
University of Edinburgh

ExTASY
Extensible Tools for Advanced Sampling and Analysis

RULE

EPSRC/NSF

ERC

ATI Summit on Big Data in the Physical Sciences

Outline

Part I: **Challenges**

Ideas from the Alan Turing Institute Scoping Workshop
"The Challenges of Data Intensive and Extreme Scale Numerical Simulation in Physics, Materials Science and Chemistry" (Jan 5-6, 2016, British Library)

Part II: **Example**

Simulations of physical states using noisy forces

$$\Downarrow$$

Inference and machine learning algorithms

I. Challenges

ATI Scoping Workshops

sticky notes

researchers

~30 white papers

facilitator

ATI Scoping Workshops

Examples (particularly Physical Sciences relevant)

- Theoretical and Computational Approaches to Large Scale Inverse Problems

- Partial Differential Equations for Modelling, Analysing and Simulating Data Rich Phenomena

- Big Data in Geoscience

- **The Challenges of Data Intensive and Extreme Scale Numerical Simulation in Physics, Materials Science and Chemistry**

The Challenges of Data Intensive and Extreme Scale Numerical Simulation in Physics, Materials Science and Chemistry

Main Organizers: Gabor Csanyi (Cambridge, Engineering), Detlef Hohl (Shell), Stephen Jarvis (Warwick, Computer Science), Ben Leimkuhler(Edinburgh, Mathematics), Mark Parsons (Edinburgh, EPCC)

Themes

Extreme scale data-computing (exascale)
Numerics for data science
Data-centric materials and chemistry modelling

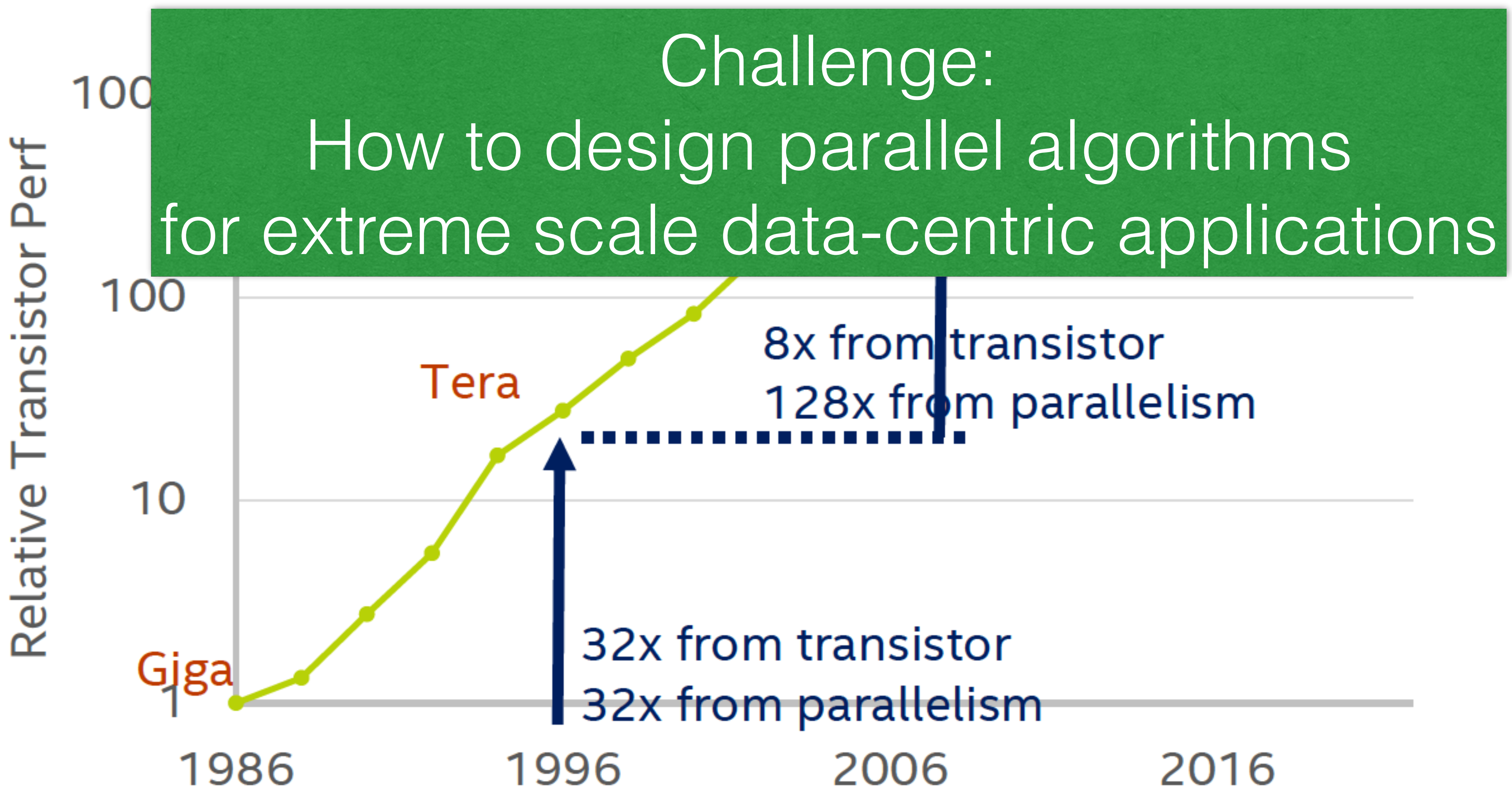
Participants

HPC scientific simulation community
Parallel computing experts
Mathematicians
Data scientists
Physical modellers (esp. materials/fluids)
Industry: Intel, Shell, Dassault (Biovia), Rolls Royce,..

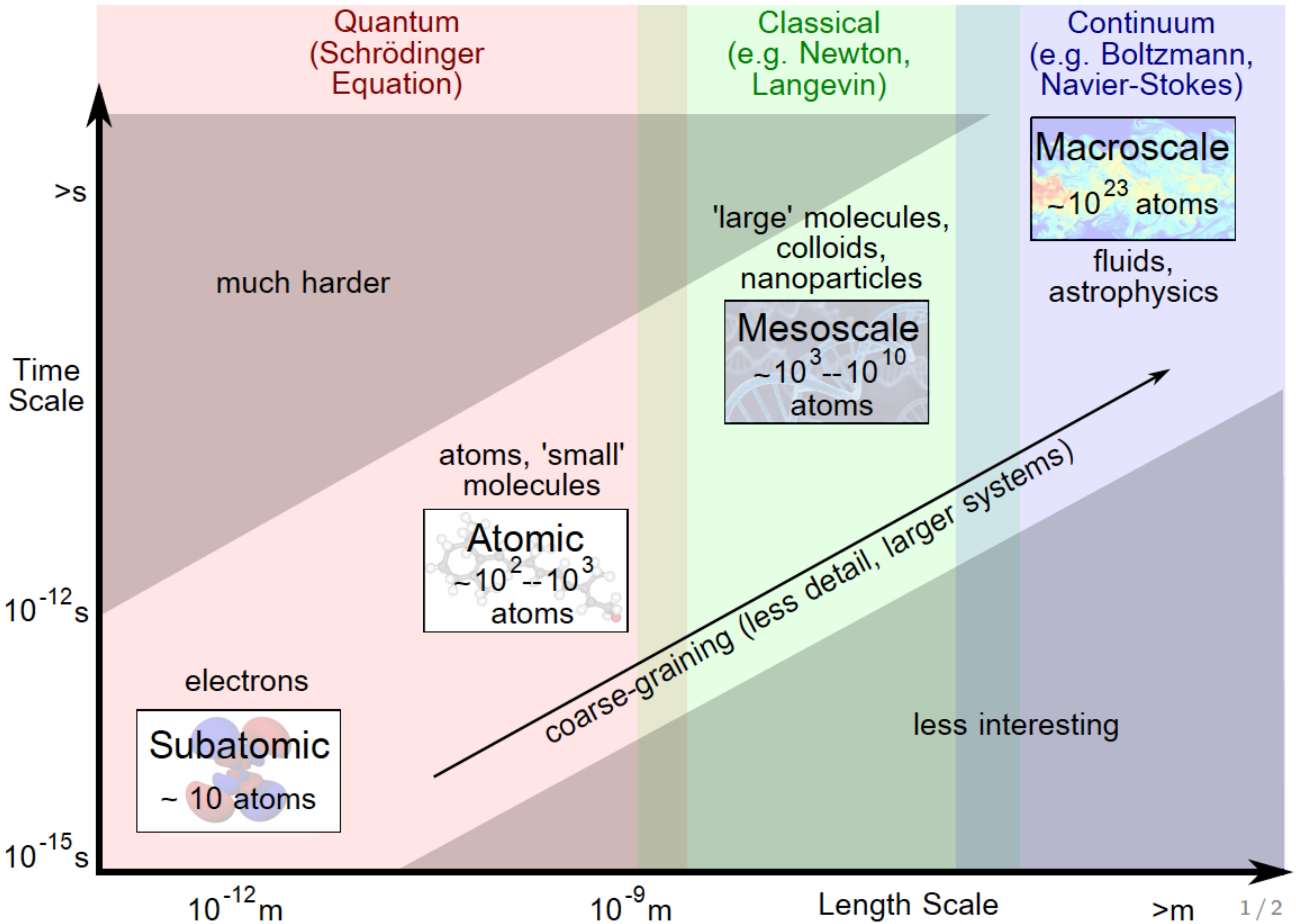
Increasing reliance on parallelism for HPC gains (as opposed to improvements from transistor) & Importance of energy considerations

Intel Exascale Labs

Implications to HPC Roadmap

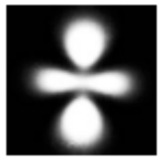


Challenge:
How to design parallel algorithms
for extreme scale data-centric applications


Quantum (Schrödinger Equation)
Classical (e.g. Newton, Langevin)
Continuum (e.g. Boltzmann, Navier-Stokes)
>s
Time Scale
10^{-12}s
10^{-15}s
much harder
Macroscale
$\sim 10^{23}$ atoms
fluids, astrophysics
'large' molecules, colloids, nanoparticles
Mesoscale
$\sim 10^{3}$--10^{10} atoms
atoms, 'small' molecules
Atomic
$\sim 10^{2}$--10^{3} atoms
electrons
Subatomic
~ 10 atoms
coarse-graining (less detail, larger systems)
less interesting
10^{-12}m
10^{-9}m
Length Scale
>m

Machine Learning Quantum Mechanics

Gabor Csanyi (Cambridge)



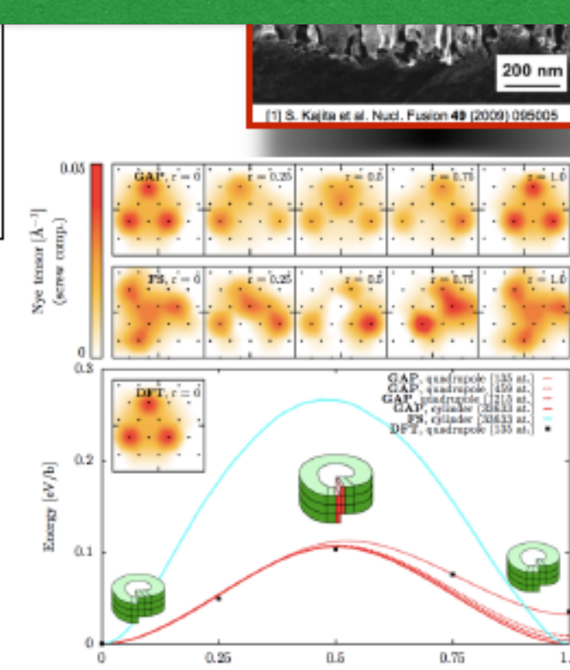
First principles simulation is extremely successful in materials science and chemistry

Traditionally $O(N^3)$ or worse
100 atoms ~ 100 CPU hours

$$-i\hbar\frac{\partial\Psi}{\partial t} = \mathcal{H}\Psi$$

Challenge:
How to build efficient molecular algorithms that can learn forcefields on-the-fly with prescribed accuracy

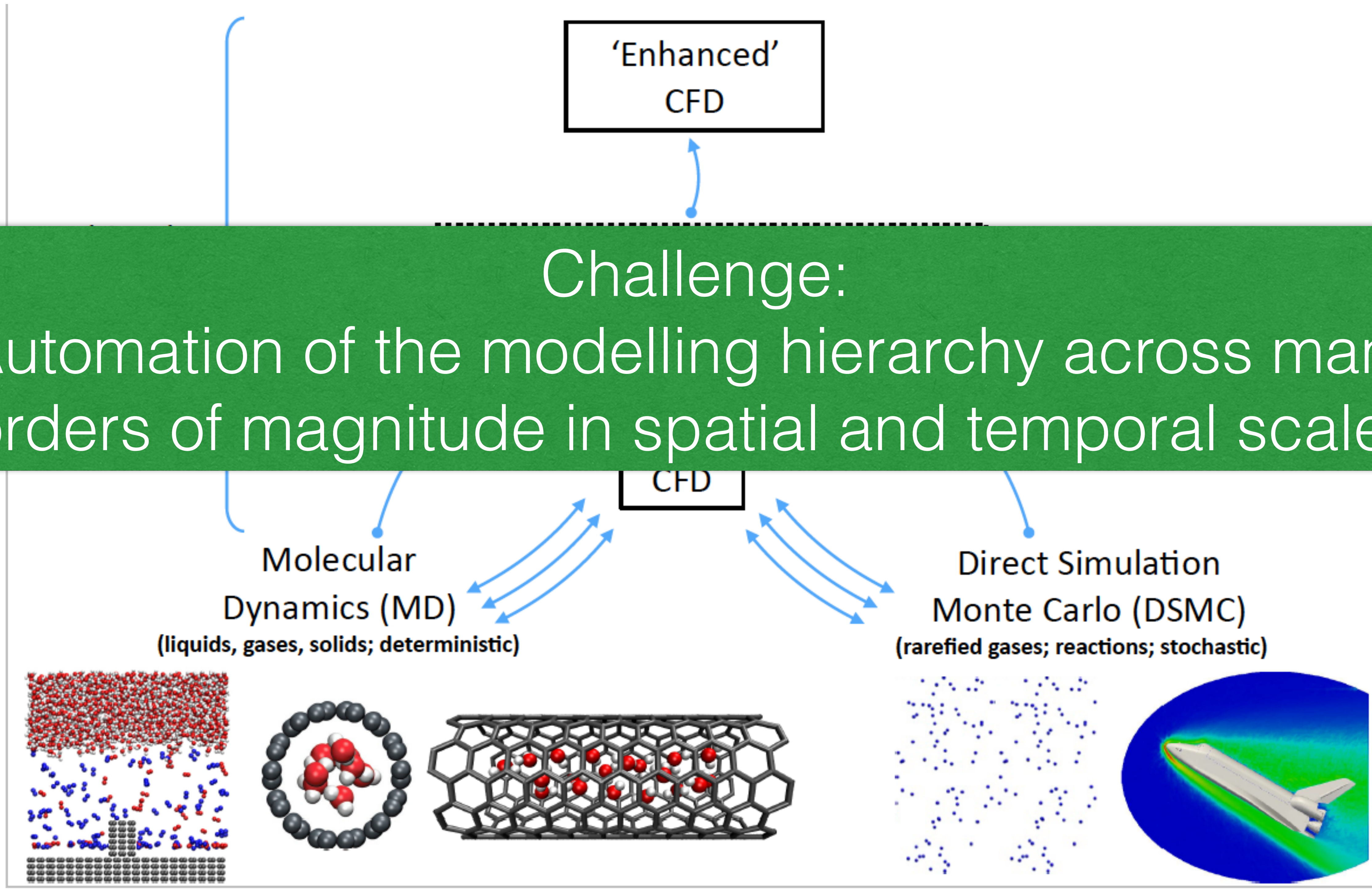
Force Fields - O(N)
(interatomic potentials)
10 ms / atom / core



- Data is plentiful and “cheap” to generate
- Need hard accuracy guarantees
- Need error prediction
- Multiple scales of interactions
- Interpolation to 10^{-3} - 10^{-4} accuracy

Data-Centric Multiscale Modelling

Matthew Borg/Jason Reese (Edinburgh)



Challenge:
Automation of the modelling hierarchy across many orders of magnitude in spatial and temporal scales

Modelling of flows, rock properties, seismic imaging

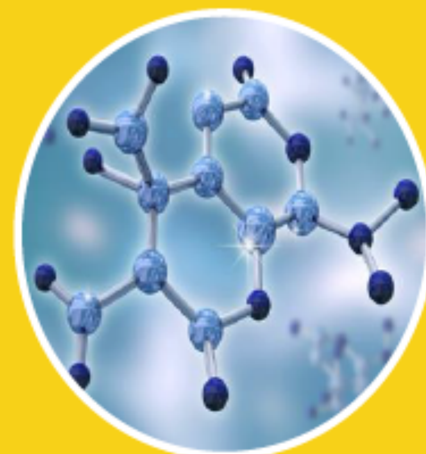
Shell Technology Centres

Geo Labs

Parallel Seismic Imaging

Micro-seismic

Induced Seismic



Sim Labs

Computational Chemistry

Comp. Material Science

Flow Compute

Digital Rock

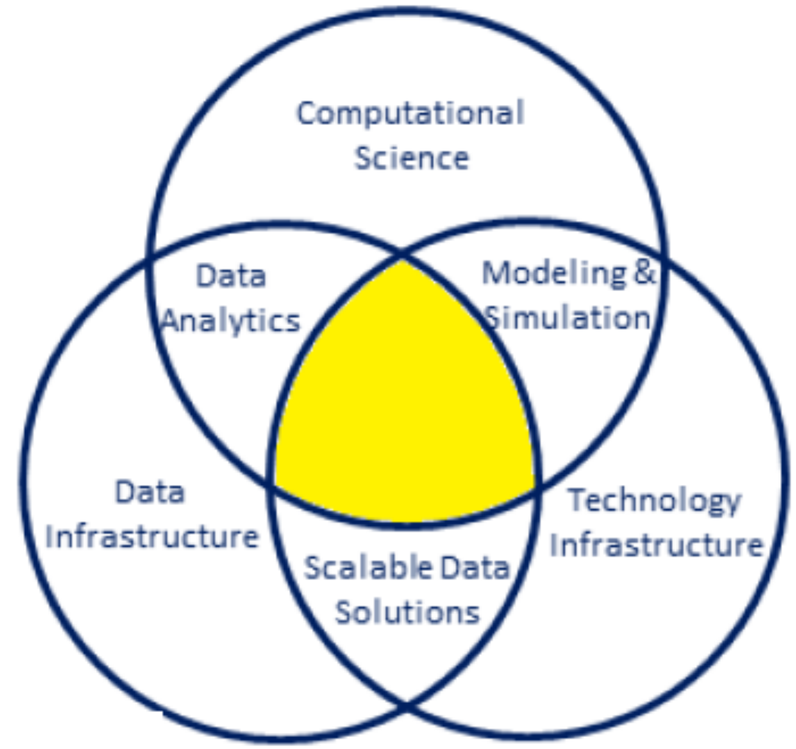
Data Labs

Adv. Analytics & Machine Learning

Sparse Modeling

Extreme statistics

High Performance Computing



Highlights

Spirited discussions about critical themes and **grand challenge problems**, e.g.

- Complex flows (unsteady, non-Newtonian fluids)
- Predictive molecular biology (e.g. rational drug design)
- Virtual materials laboratory

Focus on enabling technologies, e.g.

- Learning strategies for large scale sampling
- Data-intensive scale-bridging/coupling strategies
- Data-centric, hardware aware, extreme computing
- Data Fusion workbench

II. Example

Molecular Dynamics

$$H(x,p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + U(x_1, x_2, \ldots, x_N)$$

e.g empirical potentials

- Newton's equations + stochastic perturbations
- Computationally intensive
- Typical computations: averages in a statistical ensemble
- "Fast" force components, hence very small timesteps
- Lengths of simulations extremely limited
- Substantial share of worldwide supercomputing

Problem: use stochastic dynamics to accurately sample a distribution with given positive smooth density

$$\rho \propto \exp(-U)$$

in case the force $-\nabla U$ can only be computed approximately

Examples:

Multiscale models
several flavors of hybrid **ab initio MD Methods**
learning-based **QM/MM** methods [w./ G. Csanyi and others]
…Many applications in **Bayesian Inference &**
Big Data Analytics

What to do about the force error?

Langevin Dynamics

$$\mathrm{d}x = M^{-1} p \mathrm{d}t$$

$$\mathrm{d}p = -\nabla U \mathrm{d}t - \gamma M^{-1} p \mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathrm{dW}$$

With Periodic Boundary Conditions and smooth potential, ergodic sampling of the canonical distribution with density

$$\rho \propto e^{-\beta[p^T M^{-1} p/2 + U(x)]}$$

Splitting Methods for Langevin Dynamics

$$\mathscr{L} = \mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{B}} + \mathscr{L}_{\mathrm{O}}$$

$$\mathscr{L}_{\mathrm{A}} = (M^{-1}p) \cdot \nabla_x$$

$$\mathscr{L}_{\mathrm{B}} = -\nabla U(x) \cdot \nabla_p$$

$$\mathscr{L}_{\mathrm{O}} = -\gamma(M^{-1}p) \cdot \nabla_p + \gamma\beta^{-1}\Delta_p$$

$$e^{h\hat{\mathscr{L}}_{\mathrm{BAOAB}}} = e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{h\mathscr{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}}$$

Expansion of the invariant distribution

$$[\mathscr{L}^\dagger + h^2 \mathscr{L}_2^\dagger + \ldots] e^{-\beta(H + h^2 f_2 + \ldots)} = 0$$

Leading order:

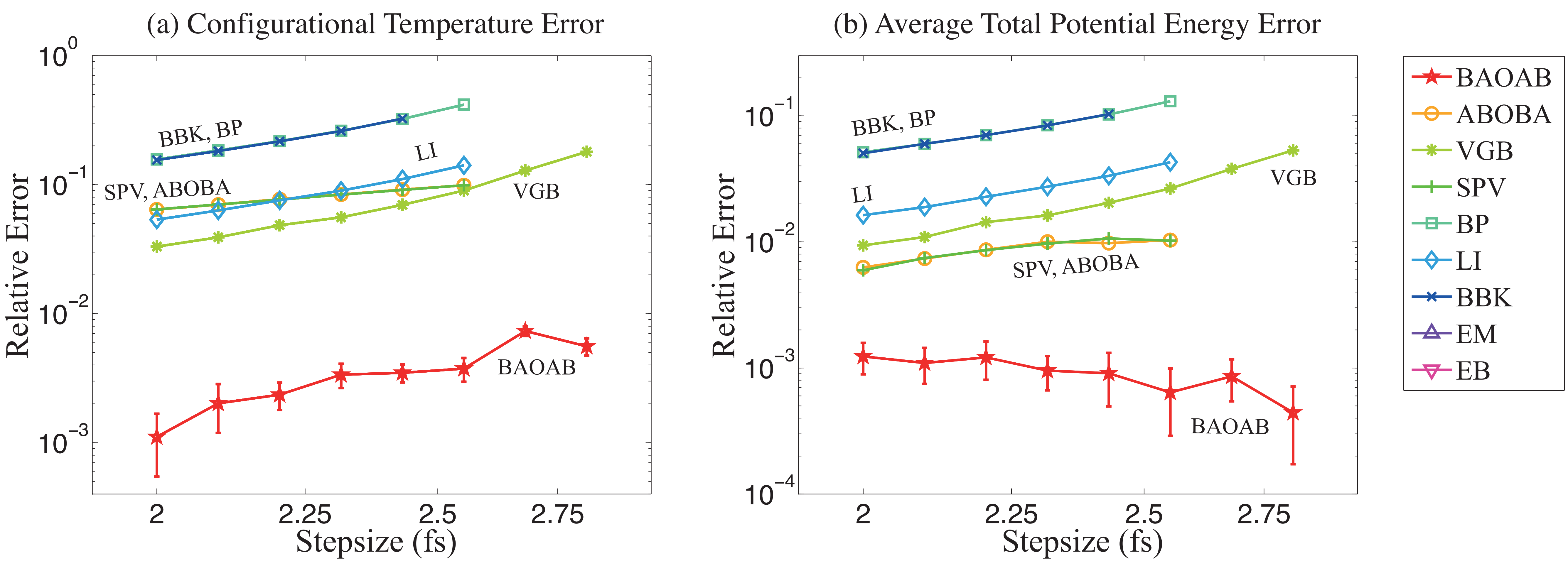
$$\mathscr{L}^\dagger(\rho_{\mathrm{can}} f_2) = \beta^{-1} \mathscr{L}_2^\dagger \rho_{\mathrm{can}}$$

L. & Matthews, AMRX, 2013
L., Matthews, & Stoltz, IMA J. Num. Anal. 2015

- detailed treatment of all 1st and 2nd order splittings
- estimates for the operator inverse and justification of the expansion
- treatment of nonequilibrium (e.g. transport coefficients)

Improvements for "real" molecular systems

for alanine dipeptide in flexible TIP3P



BAOAB much better than alternatives for relevant configurational quantities…

but….

What to do about the force error?

$$\tilde{F}(x) = -\nabla U(x) + \eta(x)$$

a sampling error… it seems natural to take

$$\eta(x) \sim \mathcal{N}(0, \sigma(x))$$

and also, at least in the first stage, to assume $\sigma(x) \approx \sigma$

$$\begin{aligned} h\tilde{F}(x) &= -h\nabla U(x) + h\eta \\ &= -h\nabla U(x) + \sqrt{h}(\sqrt{h}\eta) \end{aligned}$$

Like discretizing a stochastic differential equation with $O(h)$ variance!

The Adaptive Property

Jones & L. 2011

Applying Nosé-Hoover Dynamics to a system which is driven by white noise restores the canonical distribution.

Adaptive (Automatic) Langevin

$$\mathrm{d}x = M^{-1}p\mathrm{d}t$$

$$\mathrm{d}p = -\nabla U \mathrm{d}t - \sqrt{h}\sigma \mathrm{d}W - \xi p \mathrm{d}t + \sigma_A \mathrm{d}W_A$$

$$\mathrm{d}\xi = \mu^{-1}[p^T M^{-1} p - n\beta^{-1}]\mathrm{d}t$$

$$\tilde{\rho} = e^{-\beta[p^T M^{-1} p/2 + U(x)]} \times e^{-\beta\mu(\xi-\gamma)^2/2}$$ ergodic!

Shift in auxiliary variable by $\gamma = \dfrac{\beta(h\sigma^2 + \sigma_A^2)}{2\mathrm{Tr}(M)}$

Discretization

[With X. Shang, 2015]

generator: $\mathscr{L} = \mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{B}} + \mathscr{L}_{\mathrm{O}} + \mathscr{L}_{\mathrm{D}}$

$$\mathscr{L}_{\mathrm{A}} = (M^{-1}p) \cdot \nabla_x$$

$$\mathscr{L}_{\mathrm{B}} = -\nabla U(x) \cdot \nabla_p + \frac{h\sigma^2}{2}\Delta_p$$

$$\mathscr{L}_{\mathrm{O}} = -\xi p \cdot \nabla_p + \frac{\sigma_{\mathrm{A}}^2}{2}\Delta_p$$

$$\mathscr{L}_{\mathrm{D}} = G(p)\frac{\partial}{\partial \xi}$$

define related operator by composition, e.g. **BADODAB**

$$e^{h\hat{\mathscr{L}}} = e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{D}}} e^{h\mathscr{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{D}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}}$$

typically anticipate 2nd order (IM)

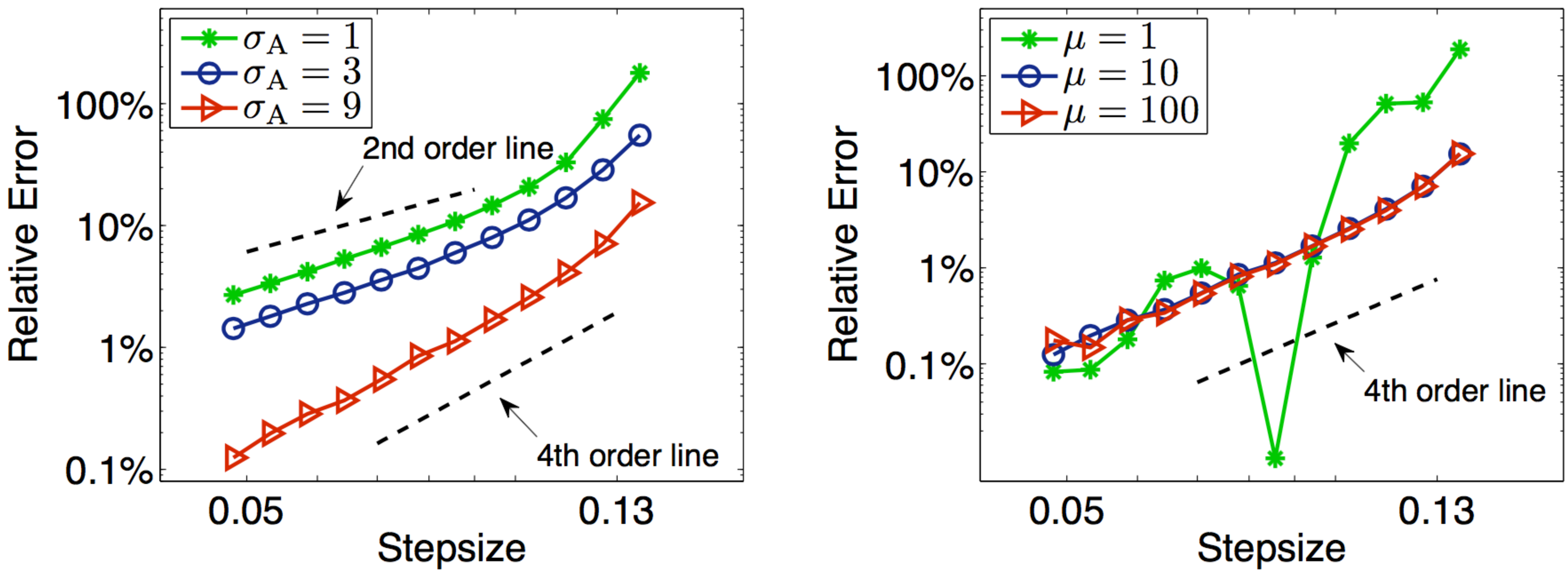
Superconvergence

BAOAB, in the high friction limit, gives a superconvergence property for configurational quantities.

By taking large $\gamma \propto \sigma_A^2$ and $\mu \propto \sigma_A^2$ we can make BADODAB behave like BAOAB in the high friction limit after averaging over the auxiliary variable.

Effectively the extra driving noise implements a projection to the case of Langevin dynamics, **but large driving noise also implies large friction so restricted phase space exploration** (even if better accuracy). So caution is needed…

500 Lennard-Jones Particle MD



- Fourth order convergence to the invariant measure
- Large friction ($\hat{\gamma} \propto \sigma_A^2$) and thermal mass ($\mu$) limits
- Only one force calculation required at each step

Bayesian Learning Application

Find best choice of parameters q given observations X

$$X = \{x_1, x_2, \ldots x_N\}$$

Challenges: data set very large

Ex: Netflix: 480000 users, 17000 ratings ⇒ 100M ratings!

Posterior probability density (from Bayes' Theorem):

$$p(q|X) \propto \exp(-U(q)), \qquad U(q) = -\log p(X|q) - \log p(q)$$

Data Scientist Thomas Bayes, U of Edinburgh, Class of 1721

Use Maximum Likelihood Estimate/"Subsampling":

$$\log p(X|q) \approx \frac{N}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \log p(x_i|q) \qquad \tilde{N} << N$$

Bayesian Logistic Regression

$$\pi(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)$$

f: logistic function

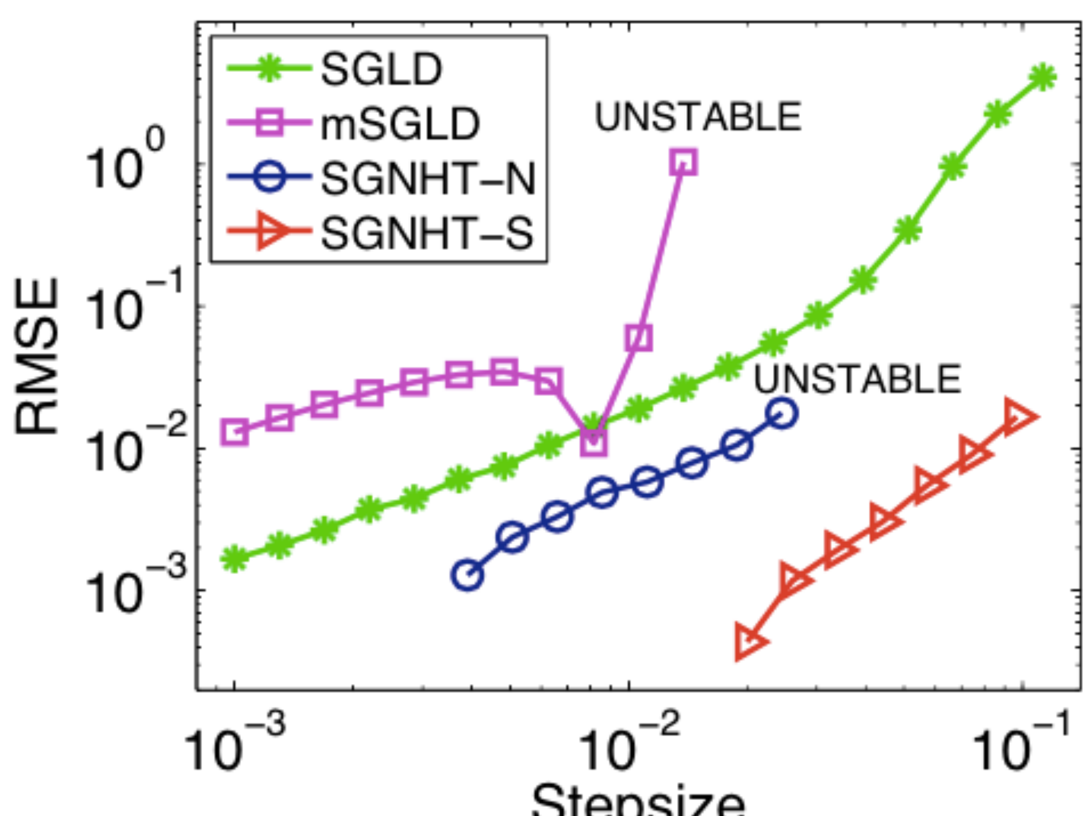
covariates e.g. age, income, …

data e.g. voting intention

posterior parameter distribution

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\beta}\|^2\right) \prod_{i=1}^{N} f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)$$

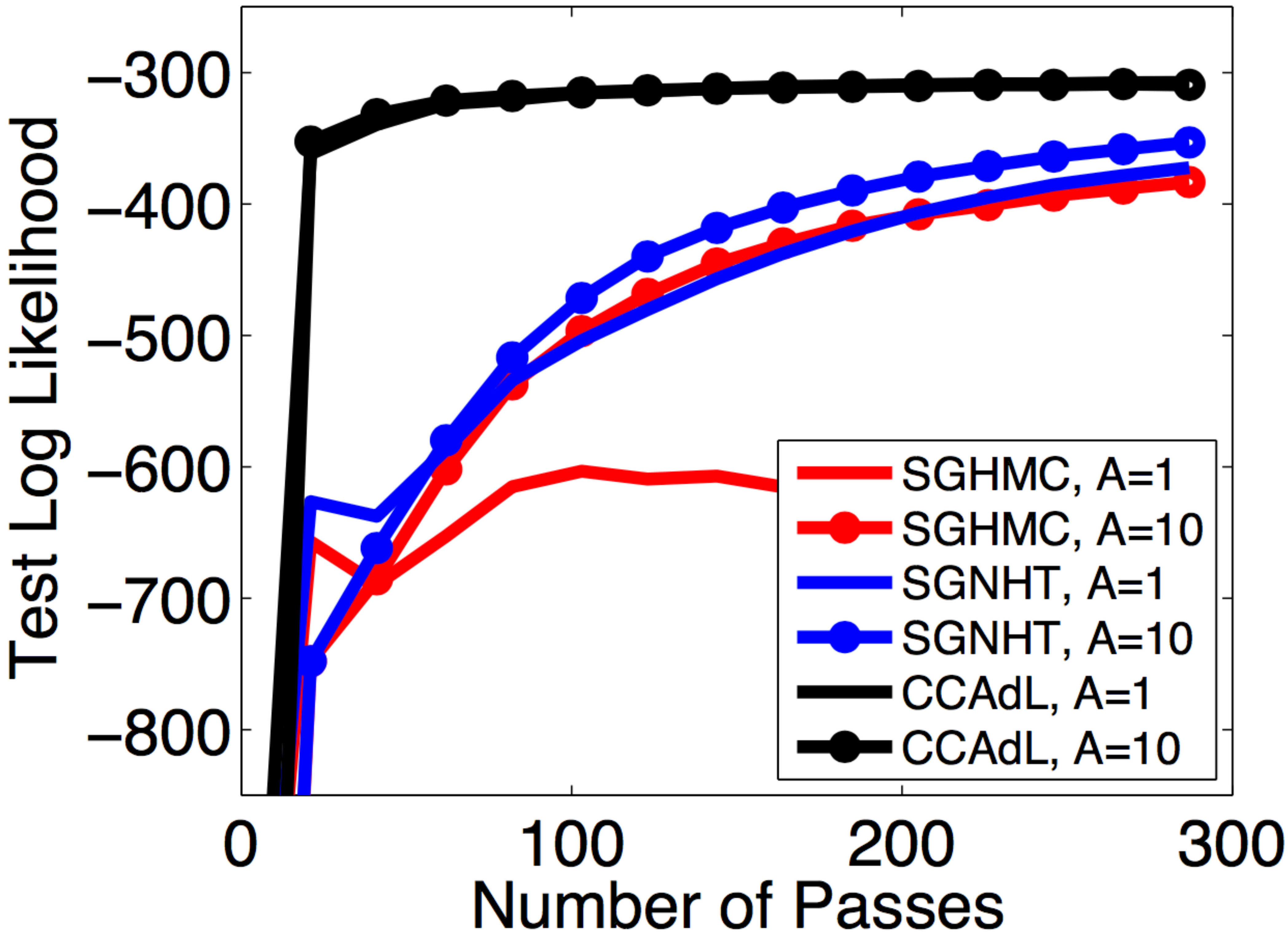
Gaussian prior



Covariance-Controlled Adaptive Langevin Dynamics

If we assume that we can obtain a covariance estimator then we can use this to enhance the accuracy of the SDEs.

CCAdL=
“Covariance Controlled Adaptive Langevin Dynamics” incorporates such a correction term together with an adaptive Langevin thermostat…



Binary classification of handwritten digits 7 and 9.

Conclusions

There are many open and interesting challenges in data-centric scientific simulation.

Much of the interest lies in the interfaces between scale regimes and in incorporating data from experiment and observation

Sometimes, methods developed for solving large scale physical applications can find new uses in the world of big data analytics.