Establishing a data cyberinfrastructure for combustion research and design

Phil Westmoreland, NC State Univ
Professor, Chemical & Biomolecular Eng;
Executive Director, Institute for Computational Science and Engineering
Two pertinent quotes:

- Stewart Brand: “Information wants to be free. It also wants to be expensive.”

- Albert Einstein: “An experiment is something everybody believes, except the person who made it.”
Consider these perspectives:

- Data represent reality.
- No data are exact except counted integers, and they can still be uncertain.
- Computed results are data, too.
- Providing data has always been virtual: journals, dissertations, letters, punch cards.
- Sharing data is the foundation of advancing science.
Larger context:

- Why? **Fuel-based energy** is a big reason.
  - “Energy costs and other expenses are falling” including “natural-gas prices -- down 60% in the past year, thanks to surging U.S. shale-gas production.”
  - “Petroleum and coal products represent 10% of U.S. manufacturing production, up from 3% in 2000.”
  - Can’t ignore sustainability or CO₂ emissions, but it buys us time – if used efficiently and cleanly.
MACCCCR-sponsored NRC report led by Mitch Smooke.

- Its first recommendation was:
  - "A unified combustion cyberinfrastructure should be constructed that efficiently and effectively connects with and enables the movement of data and the sharing of software tools among the different research communities contributing to engine and combustion research and development."
  - If started from scratch, a 5-year project to create comprehensive combustion CI; up to 75 FTEs.
  - Instead, can we adapt available cyberinfrastructure?
Workshop Dec 12-13, 2011.

- 2010 Fuel Summit discussion led to white paper, proposing action-plan-writing workshop.
  - Goal: An action plan for feasible CI development for minimum to larger funding levels.

- NSF funded the workshop at NCSU’s Institute for Computational Science and Engineering.
  - 12 people; analyze issues, propose actions.
  - Documents were developed and made available to MACCCR via Google Docs (cloud).
Different contributions with different perspectives.

- Pam Chu, NIST
- Med Colket, United Technologies Research Center
- Michael Frenklach, UC Berkeley
- Bill Green, MIT
- Mike Burke, Argonne
- Jeffrey Manion, NIST
- Phil Smith, University of Utah
- Mitch Smooke, Yale
- Mani Sarathy, Lawrence Livermore*
- Wing Tsang, NIST
- Charlie Westbrook (Livermore, Sandia, Rxn Design, Comb Inst)
- Phil Westmoreland, NC State
Quickly agreed on larger goals:

- Achieve both usability and usage.
- Set up CI that can ultimately serve a wide range of combustion and other science.
- Set up sustainable governance and operation.
- Propose a spectrum of activity levels, from the minimum to the NRC-proposed level.
Identify key components.

- Software
  - Prediction
  - Interpretation
- Data
  - Data mining
  - Parametric database and metadata
- Hardware
  - Server hardware
  - Distributed hardware
- Data from new expts
- User-community support
Community-building; How?

- Requires breadth eventually; not just kinetics.
- Start with a small problem, tackled hard.
- Create user-community support.
- Change data culture: Share fast, get credit.
- International community, but US-led => US benefit.
- Build on PrIMe but fix on current version as a Version 2.0; integrate with NIST Webbook.
Key tasks and roles identified.

(1) **Must curate the data.**
- More than storage.
- Extensions of PrIMe’s XML formats.
- Completeness and consistency – assess/assign uncertainty.
- Adequate metadata and live URL links.
- Couple with databases like the NIST Chemistry Workbook.
- Aid the data contributors and access by 3rd-party codes.

- Identified NIST as a natural lead.
  - As with each task, need buy-in from the team/community.
  - As with each task, this lead must cede some authority to the team/community leadership (negotiated).
Possible key tasks and roles identified.

(2) Need data-addition and -manipulation software.
- Begin with existing version 2.0 of PrlMe as stable platform.
- In background, develop additional features, such as:
  - Read-in of legacy Chemkin mechanisms.
  - New DataModels (data templates for types of experiments).
  - Links with other databases like the NIST Chemistry Workbook.
  - Cloud computation of response surfaces for uncertainty analysis.
- Build in additional modeling codes, like Cantera, Green’s RMG, Smooke’s flame codes, Ruscic’s ATcT, Smith’s Glacier/Banff.
- Identified Michael Frenklach as a natural CTO/coordinator.
  - As with each task, need buy-in from the team/community.
  - As with each task, this lead must cede some authority to the team/community leadership (negotiated).
Possible key tasks and roles identified.

(3) Need implementation on suitable hardware.
- Need adequate storage for this phase and scope.
- Need cloud base for the data-contribution and –use software.
- Need user support.
- Identified PRW as lead/coordinator.
  - Mostly through opportunity: NCSU’s offer of initial two years free, plus VCL cloud-based system.
  - As with each task, need buy-in from the team/community.
  - As with each task, this lead must cede some authority to the team/community leadership (negotiated).
Possible key tasks and roles identified.

(4) Need oversight and control by the community and constituencies.

- Identified two boards:
  - **Community Advisory Board**, to whom the operational leads report.
  - Community representatives as a governing board of directors.
  - Provide oversight, assessment, community feedback.
  - **Executive Advisory Board** from industry and government.
  - Provide assessment, insights, and sponsor feedback.
Developed a model for critical mass, targeting predictive combustion models.

- **Drive technical advances:**
  - Set up cloud-based data storage, data-use software, data archiving, and data curation.
  - Launch quickly by adapting existing formats, databases, and software.
  - Develop new software sustainably, incorporating new uncertainty-analysis features, experimental templates, open-access formats, and open-source software for data mining and predictive simulation.

- **Demonstrate the DCI’s power with high-impact results:**
  - Lead community- and software-driven projects in $\text{H}_2/\text{O}_2$, natural-gas, and liquid-fuel reaction sets.

- **Aid the growth in productivity of the community of users:**
  - Sustainably provide support, usage training, technical training in uncertainty quantification, and community-building projects.
Governance is a key.

Community Advisory Board: Senior investigators and R&D participants

Project administration: PI: P.R. Westmoreland, NC State Univ.

Core-software coordination, maintenance, development; 2 FTE professionals reporting to M. Frenklach, UC-Berkeley

Data operations (Curation, data-I/O support): 1 FTE professional reporting to P. Chu, NIST

Hardware operations (Cloud server and technical support): 1 FTE professional reporting to P.R. Westmoreland, NCSU

DCI application-project teams (Sr investigators, volunteers)

Engage other Pred S&E SG / SI / center awardees

DCI Software-innovation teams (Sr investigators, volunteers)

Executive Advisory Board: Industry/agency management representatives
More on this model.

- Westmoreland, Chu, Frenklach as above.
- Smooke, oversight of combustion-modeling software to be implemented via the DCI.
- Green, coordinate implementation of his Reaction Mechanism Generator code.
- Smith, soot and radiative-transport systems and for aspects of uncertainty quantification.
- Westbrook, incorporate existing chemical-kinetic reaction mechanisms, aid involvement of the international combustion community, and aid interactions with LLNL and Sandia.
- Colket, advise on and aid industrial application, utilization, and access issues.
- Ruscic, incorporate “Active Thermochemical Tables.”
- Student data entry; short courses/tutorials; HW services.
Is it just nice? Optional? No.

- Science and government confronting the reality of Big Data, analytics, and sharing data – a cultural shift.
  - Alternative is to be “on the wrong side of history” - science’s future – and to be left behind.
- Tsinghua is starting to develop an alternative.
  - Let them do it instead? International investment?
- Journals aren’t the solution.
- “Trust but verify” is the essence of science and technology – and having the data is necessary.
Data mining
Prediction
Interpretation
Parametric database and metadata
Server hardware
Data from new expts
Distributed hardware
User-community support