Addressing Challenges of Data-Intensive Research

Maciej Remiszewski
Director of Development & Collaboration
ICM Interdisciplinary Centre for Mathematical and Computational Modelling
UNIVERSITY OF WARSAW
Outline

Three stories:

- ComputationalSciences@ICM
- OpenScience@ICM
- DataScience@ICM
Interdisciplinary Centre for Mathematical and Computational Modelling

- Launched in 1993 as COMPUTATIONAL SCIENCES CENTRE running SIMULATIONS OF COMPLEX PROBLEMS using MATHEMATICAL MODELLING and SUPERCOMPUTERS

- Truly INTERDISCIPLINARY TEAM approaching 200 scientists & developers working in areas like:

  air transportation, bioinformatics, climate modelling, computer assisted medicine, cosmology, digital libraries, drug discovery, epidemiology, agriculture, high energy physics, machine learning, materials science, neurobiology, social network analysis, numerical weather prediction, (...)

ICM Interdisciplinary Centre for Mathematical and Computational Modelling UNIVERSITY OF WARSAW
Numerical Weather Prediction
Natural Environment Modelling Group
http://meteo.icm.edu.pl

Three NWP models in operation:
- Unified Model (Met Office)
- COAMPS (Naval Research Laboratory)
- WRF (internal use)

Public weather forecasting service
open-access graphical weather forecast
http://meteo.pl

Focus on Application Research
- energy generation (wind & solar)
- forest fire risk (Polish Forest Governance)
- flash-frost prediction (agriculture)
- turbulence forecast (airline industry)
multi-scale individual based model
sub-cellular, cellular and tissue scale

hybrid formulation
discrete cells, continuous environment

highly scalable
MPI + OpenMP parallelisation
optimised for execution on MPP systems

Large Scale Parallel Simulations of 3-D Cell Colony Dynamics
Maciej Cytowski, Zuzanna Szymańska

Abstract
Biological processes are inherently very complex and involve many unknown relationships and mechanisms at different scales. Despite many efforts, one still cannot explain all the observed phenomena and, if necessary, make any desirable changes in the dynamics. Recently, it has become apparent that the opportunity lies in complementing the traditional, heuristic experimental approach with mathematical modelling and computer simulations. Achieving a simulation scale that corresponds for instance to clinically detectable tumour sizes is still a huge challenge, however it is necessary to understand and control complex biological processes. In this paper we present a novel high performance computational approach allowing simulations of 3D cell colony dynamics in previously unsuitable tissue scale. Due to the high parallel scalability we are able to simulate cell colonies composed of 10⁷ cells, which allows for instance to describe tumour growth in its early clinical stage.

Keywords: mathematical modelling, cellular processes, parallel computing.

1 Introduction
Experimental methods alone are very often not sufficient to build a consistent, systematic theory which is capable of describing biological phenomena. Exploring all the complexities of biological systems may require many experiments to be performed. Moreover, a full understanding of the governing mechanisms with the use of experimental method alone is nearly impossible due to their highly non-linear nature. One of the tools extensively used to complement the traditional, experimental approach is mathematical modelling, which aims at the mathematical representation of biological processes using a variety of analytical and computational techniques. Such models have proved to be very successful in many applications e.g. systems biology, bioinformatics, molecular dynamics, neurobiology, biological tissue and cell modelling. In particular, numerous mathematical models have been proposed to study biological processes related to cell

“M. Cytowski and Z. Szymańska are with the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland e-mail: m.cytowski@icm.edu.pl, z.szymanska@icm.edu.pl”
Visual Analysis Lab

**Competence areas**
- Visualisation
- Scientific Visualisation
- Large Scale Problems
- Visual Analysis
- Computer Assisted Medicine
- Medical Image Analysis
- Diagnosis Support
- Radiology & Surgery Support
- Research & Development Projects
- Software Development
- Visualisation Infrastructure

**VISNOW**
- Generic visualisation framework
- Developed at ICM (open source)
- [http://visnow.icm.edu.pl](http://visnow.icm.edu.pl)
“PCJ is a library for Java language that helps to perform parallel and distributed calculations. The current version is able to work on multicore systems connected with the typical interconnect such as ethernet or InfiniBand providing users with a uniform view across nodes.”

Heterogenous Parallel and Distributed Computing with JAVA (HPCDJ)

2014 HPC Challenge Award for Most Elegant implementation
HPC infrastructure at ICM

Capacity Systems
- thin nodes (2-socket Intel)
- fat nodes (4-socket AMD)
  shared within PL-GRID

Capability Systems
- MPP (BlueGene/Q & P)
- coarse-grain (POWER775)
- accelerated (NVIDIA)
providing tools and advocating for Open Access approach in science

close collaboration with key players in this field in Poland and Europe

participation in OpenAIREplus and FOSTER projects

operation (lead) of Open Science Platform which aggregates several projects and services, including:

 Virtual Library of Science: access to scientific publications

 CeON repository: an open-access repository for Polish research output

 CeON aggregator: aggregating the content of Polish repositories

 Open the Book: access to digitised versions of Polish science literature

conferences, public meetings and lectures promoting Open Access issues
ICM provides access to scientific publications and search tools for the Polish scientific community to support research and development.
Joint effort of 50 partner organisations

Supporting EC’s Open Access policies and mandates

Funded through FP7 & H2020 projects, planned to become a sustainable legal entity within the next year

ICM is one of the leading partners (Project Steering Committee member), responsible for 2 work packages

- Knowledge Extraction Services
- Operation and Maintenance
design and build a digital library service that will collate the current distributed content

help plan the long term preservation of digital mathematical literature

create a document network as integrated and transnational as the discipline itself

12 partners

ICM acting as technological lead
Open Research Data: Implications for Science and Society

Topics covered will include:

- DATA SHARING & PUBLISHING policies, strategies and incentives
- TOOLS & METHODOLOGIES for opening data
- RE-USE OF DATA for science and society: opportunities and challenges


Open Science Platform

Conference

Open Research Data: Implications for Science and Society
Warsaw, May 28–29, 2015

Celina Ramjoué, European Commission
Tim Smith, CERN
Martin Hamilton, Jisc
Kevin Ashley, Digital Curation Center
Mark Parsons, Research Data Alliance
Giulia Ajmone-Marsan, OECD

Online registration and submission of abstracts for short presentations are now open.

http://opendataconference2015.pon.edu.pl
4th Paradigm

• Empirical
describing natural phenomena
(last millenia)

• Theoretical
building models and generalisations
(last centuries)

• Computational
simulating complex phenomena
(last decades)

• Data Exploration
“data-intensive” scientific discovery
(last years)
Consequences of Paradigm Shift

- need for new types of RESEARCH INFRASTRUCTURES
  - data-intensive research
  - sharing and publishing data sets (Research Data Alliance)

- need for new METHODS & ALGORITHMS
  - MapReduce algorithms
  - in-memory processing

- need for new SKILLS & ABILITIES
  - Big Data processing
  - Machine Learning
  - Information Visualisation
Data Science is a strategic development direction for ICM.

The OCEAN programme establishes a data science centre of excellence at ICM featuring interdisciplinary research teams along with a new facility and state-of-the-art infrastructure.
Judical Decision Analysis System (SAOS)
GOAL: gain insight into law practice in Poland
DATA: 200+ Polish courts using various data formats (PDF, DOC, HTML tables, XML, etc.) in discrete, unaligned systems.

Bovine milk performance prediction
GOAL: better milk yields better chocolate :)
DATA: 80+M records describing animals (herd, genes) and milking events (milk quantity and composition) collected over the last 20 years.
Computational Social Sciences Lab

- Quantitative analysis of structures, processes and social phenomena
- Solutions based on research
The Crunchers of Big and Complex Data

Skills

- Statistical Modelling
- Deep Data Digging
- Data Visualisation

Applications

- Bioinformatics / Cancer research
- Energy Demand Forecasting
- Design of Scalable ML Algorithms
Data Science Warsaw

“Data Science Warsaw is a community of data scientists based in Warsaw. We are a non-profit professional organisation dedicated to the free, open, dissemination of data science. We meet to discuss the tools, methods and technologies used to ingest, transform, explore, analyze; visualize data, obtain predictive; prescriptive insight, develop data products, and exploit business opportunities from data products.

The organizers of the Data Science Warsaw meeting are ICM and Foundation DataSci.”
Collaboration with Industry

- The INDUSTRY/BUSINESS World has a lot of data almost as many questions, but very few Data Scientists.
- The SCIENCE World has the people, skills and infrastructure, and is looking for new challenges.
- It seems like we are VERY MUCH NEEDED:
  - understanding and extracting value from data
  - statistics, machine learning, visualisation
  - big data processing (Hadoop, SPARK)
  - Project methodologies (CRISP-DM)
- ENGAGE IN DISCUSSIONS WITH COMERIAL CLIENTS!
Thank you for your time & attention!

Come to the dark side...
...We have cookies

Maciej Remiszewski
maciej.remiszewski@icm.edu.pl
pl.linkedin.com/in/remiszewski