MATH4280
Innovation and Design in Big Data Analytics

Eric Chung

Department of Mathematics

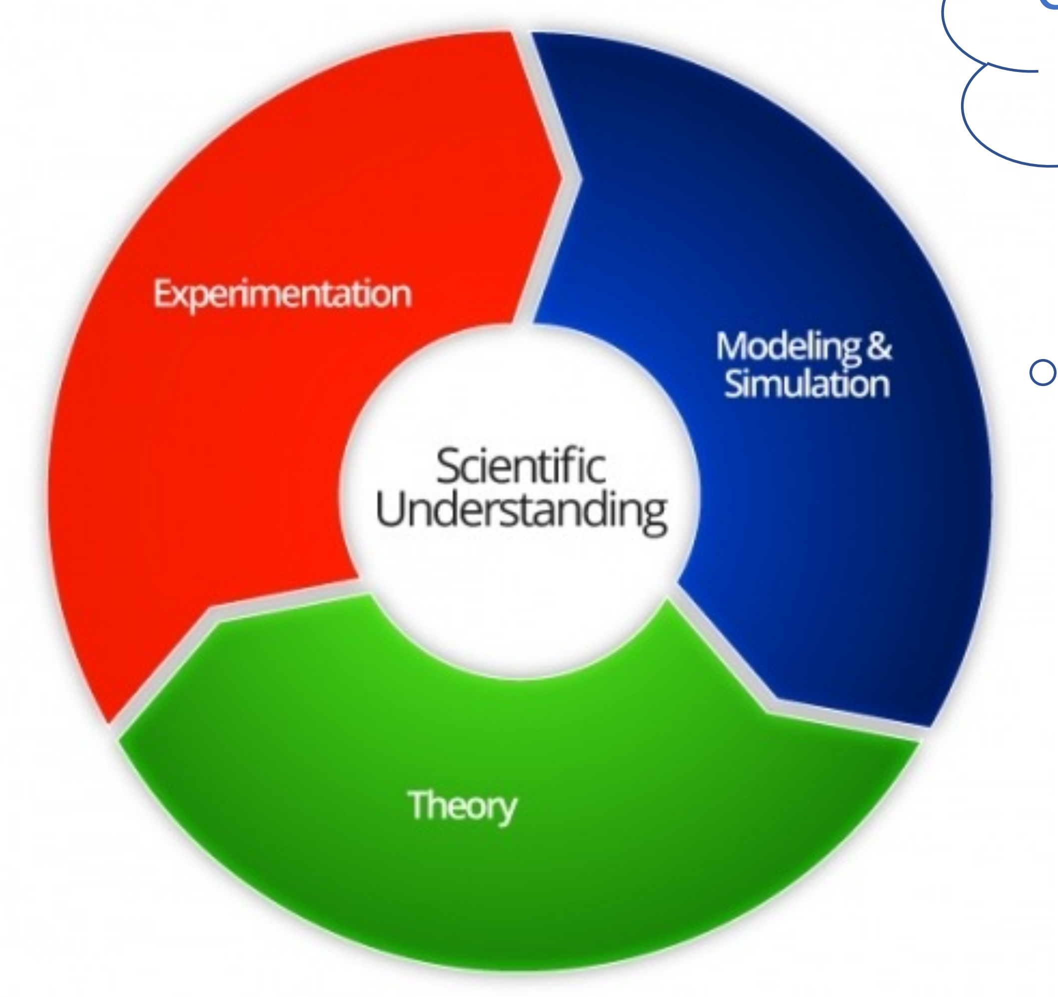
Chinese University of Hong Kong

Basic information

- Instructor: Eric Chung
- Office: LSB205
- Email: tschung@math.cuhk.edu.hk

- TA: Wong Po Chai and Mandy Man
- Office: LSB222C
- Email: pcwong@math.cuhk.edu.hk, mhyman@math.cuhk.edu.hk

Three pillars of science



Computational techniques

07 Mar 2019

Taken from the March 2019 issue of *Physics World* where it appeared under the title "The third pillar". Members of the Institute of Physics can enjoy the full issue via the *Physics World* app.

Computing has quickly evolved to become the third "pillar" of science. But to reap its true rewards, researchers need software code that is flexible and can be easily adapted to meet new needs, as Benjamin Skuse finds out

"Computation fills in a gap between theory and experiment," says David Ham, a computational scientist at Imperial College London in the UK. "A computation tells you what the consequences of your theory are, which facilitates experimentation and observation work because you can tell what you are supposed to look for to judge whether your theory is valid."

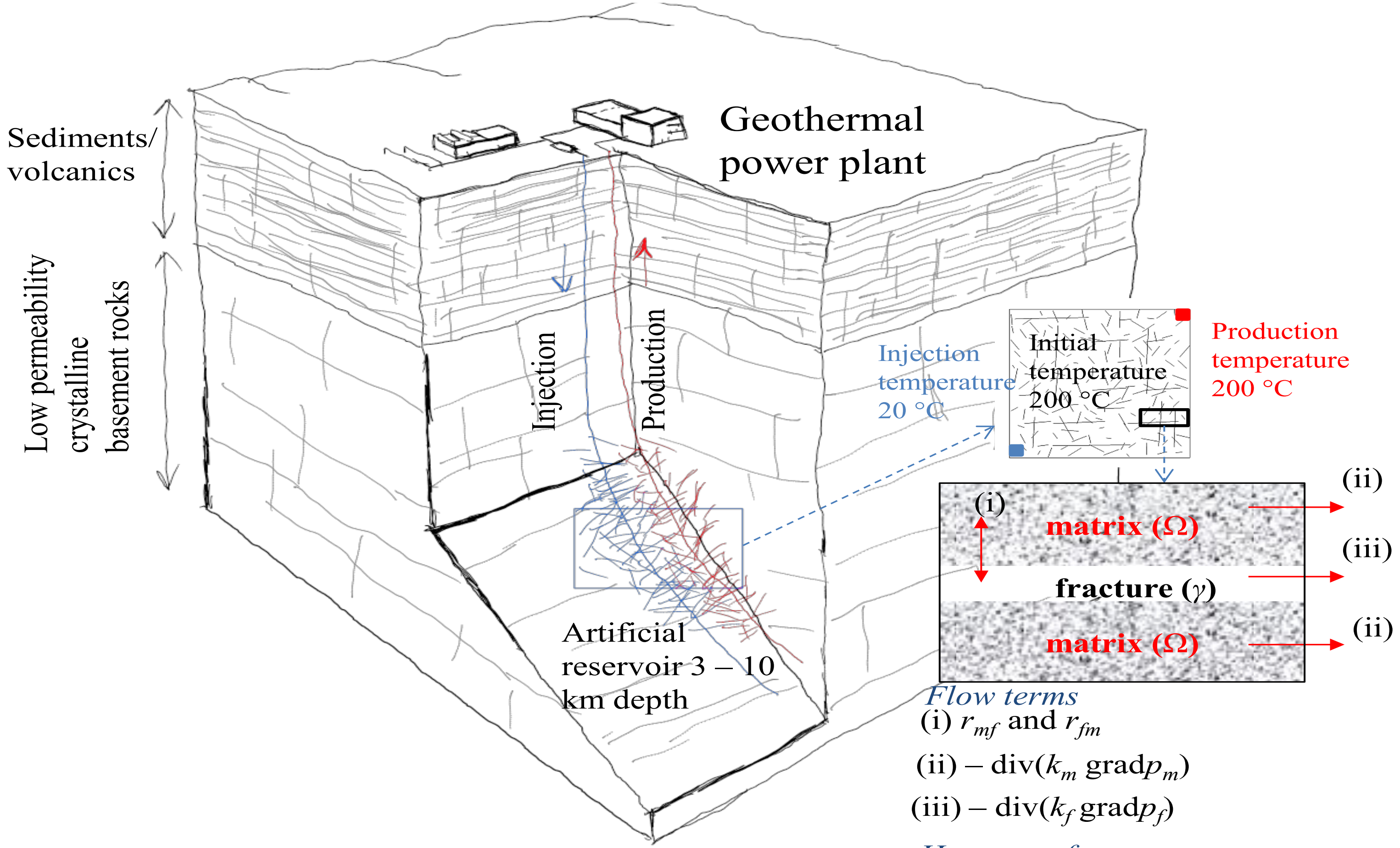
From Physics World

> Computation is not just an extra tool. It is a new way of doing science, irrevocably changing how scientists learn, experiment and theorize

Importance of computational techniques

- Mathematical models arise in many applications
- Advanced imaging techniques produce high resolution media properties, e.g. heterogeneities, geometries, …
- Good: more information, more accurate predictions
- Bad: difficult to compute due to complicated solution pattern
- Accurate and efficient simulations are challenging, due to the complexity of the problems
- A good computational technique can help engineers and domain scientists to make predictions, and aid in production or business plans

Geothermal energy



Flow terms

(i) r_{mf} and r_{fm}

(ii) $-\,\mathrm{div}(k_m\,\mathrm{grad}p_m)$

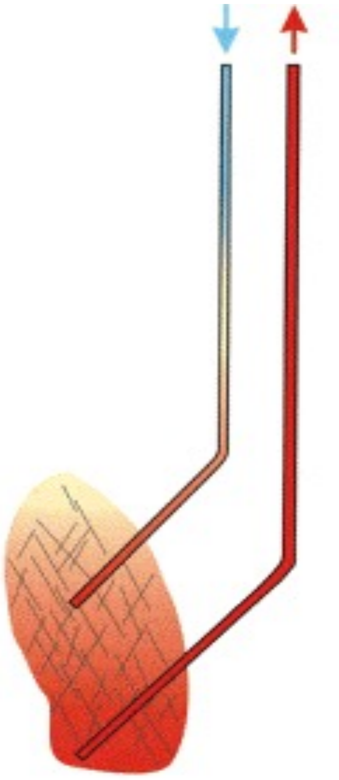
(iii) $-\,\mathrm{div}(k_f\,\mathrm{grad}p_f)$

Heat transfer terms

(i) L_{mf} and L_{fm}

(ii) $(c\rho)_m\mathrm{div}(q_m T_m) - \mathrm{div}(\lambda_m\,\mathrm{grad}p_m)$

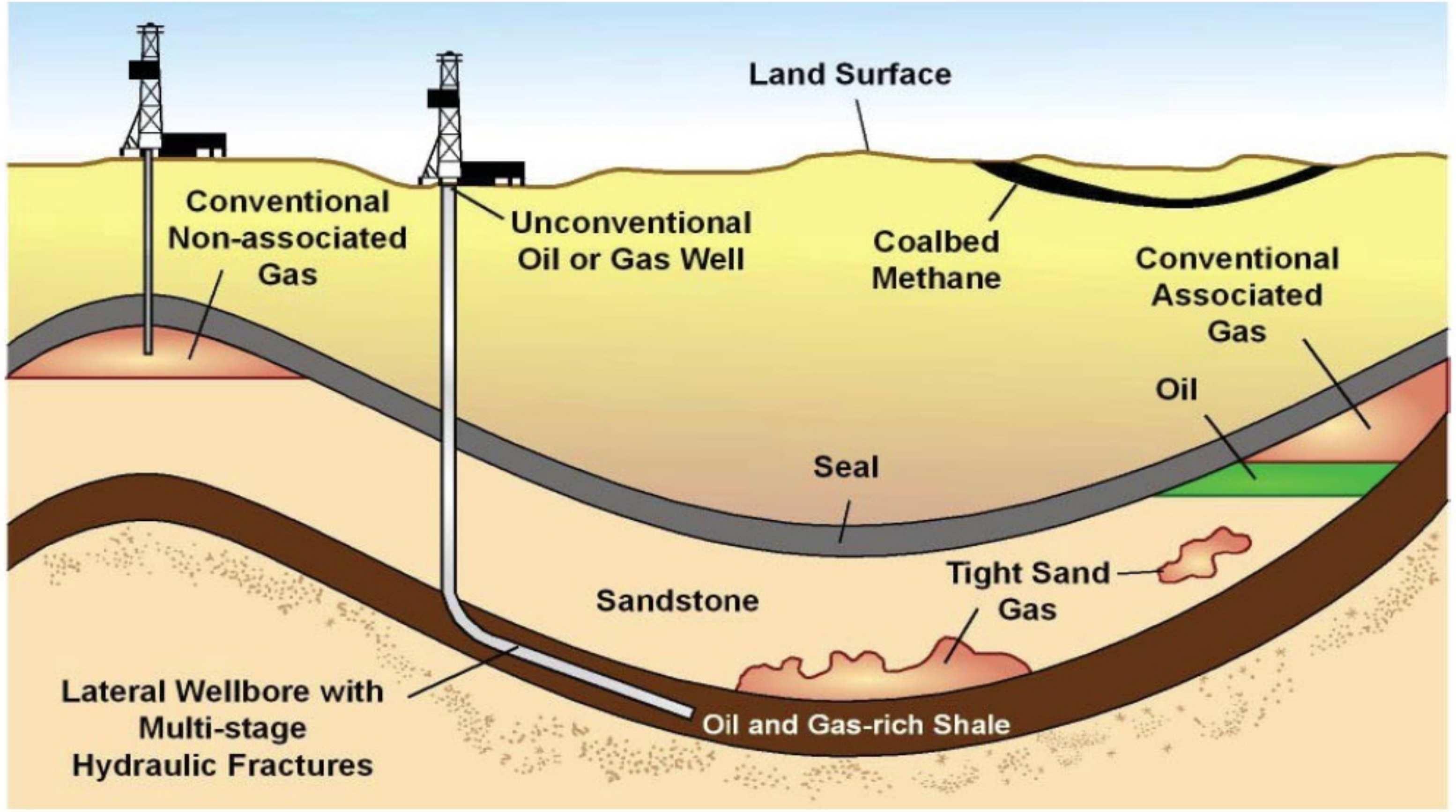
(iii) $(c\rho)_f\mathrm{div}(q_f T_f) - \mathrm{div}(\lambda_f\,\mathrm{grad}p_f)$



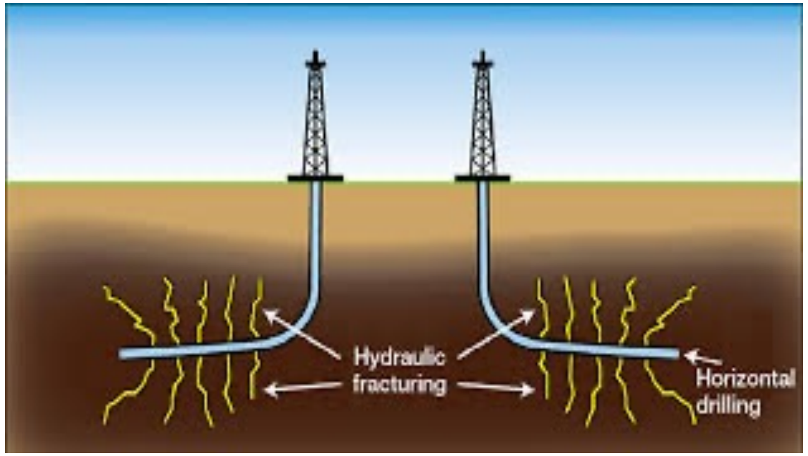
Fracture media

Unconventional oil and gas

The Geology of Conventional and Unconventional Oil and Gas

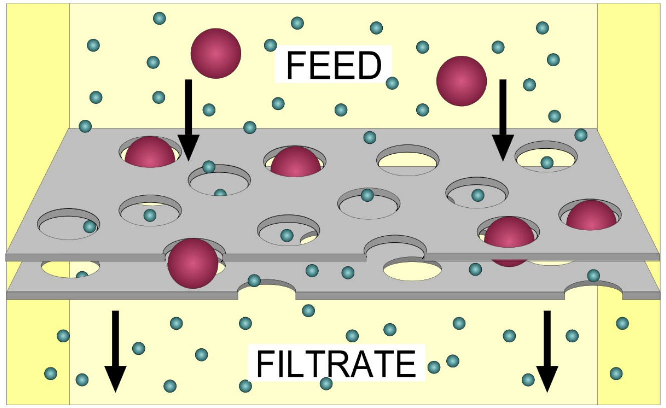


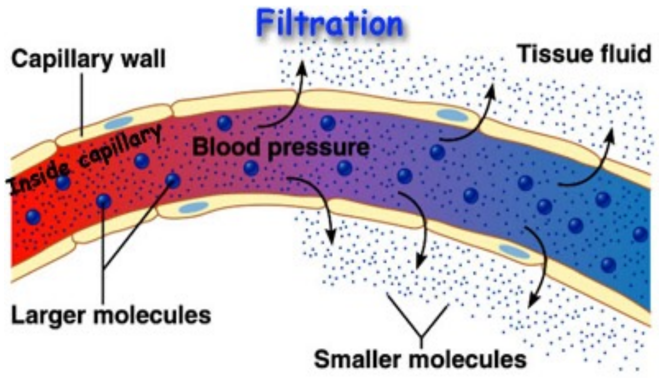
Source: EIA

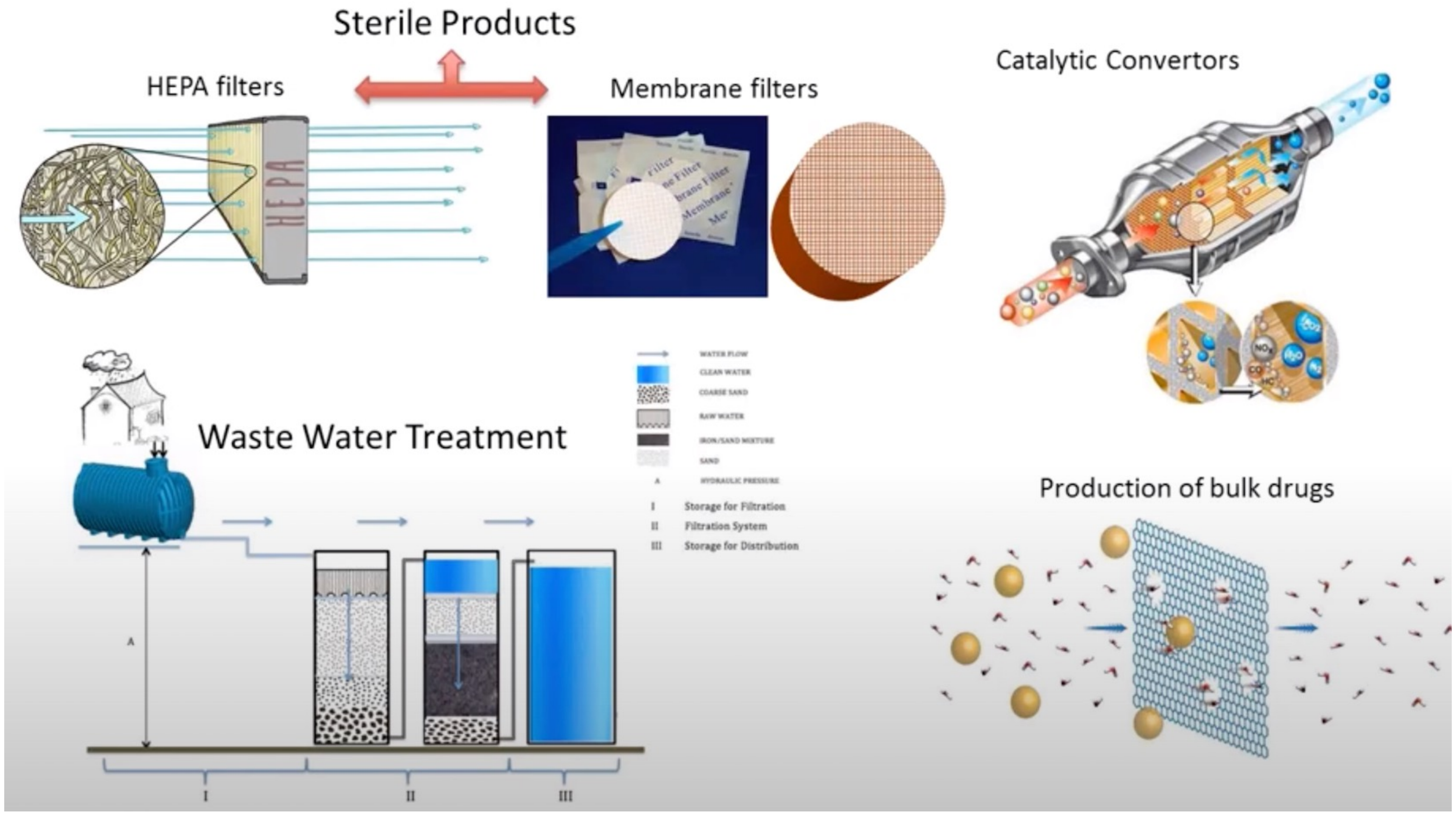


E.g. Shale oil and gas

Filtration

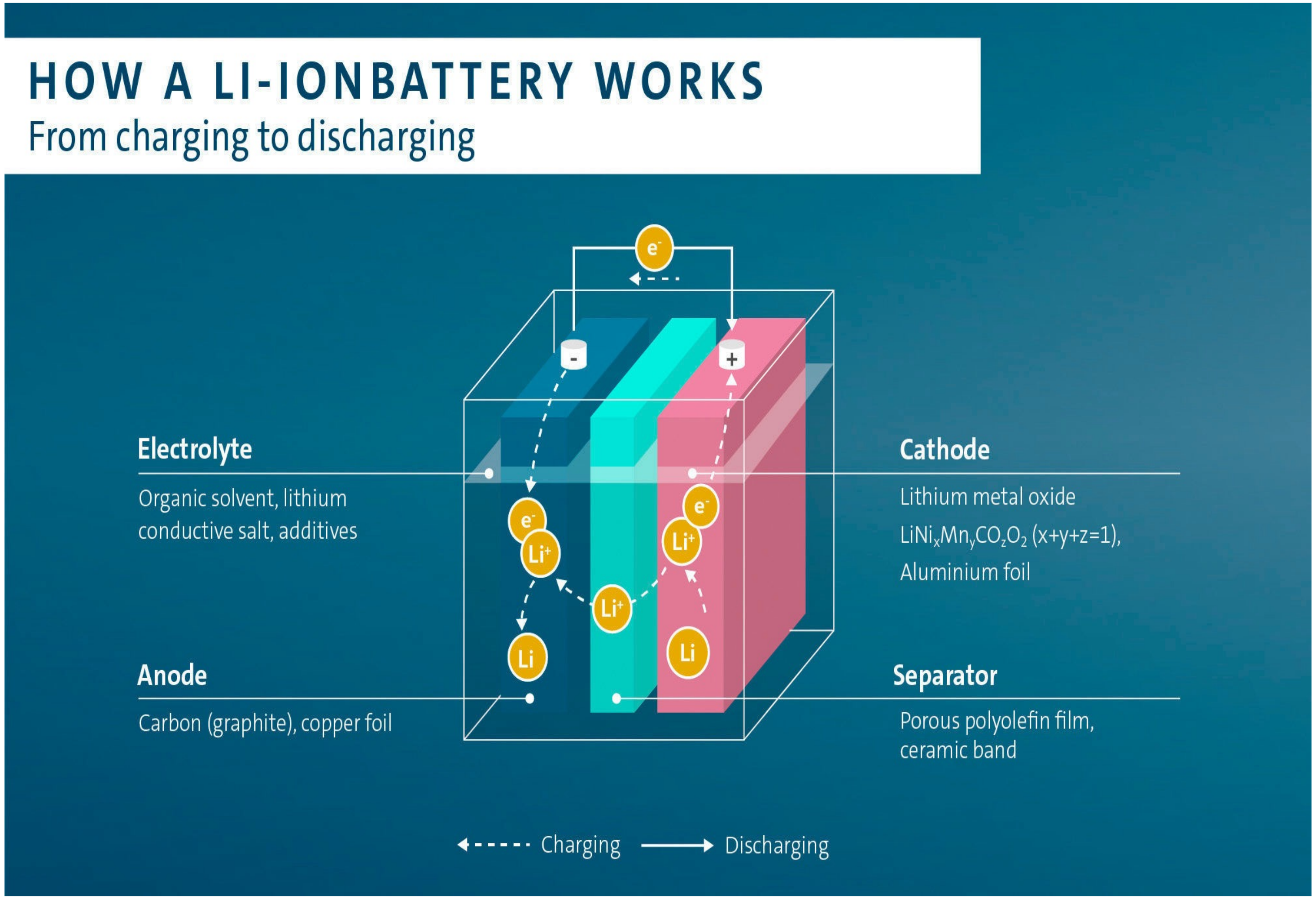




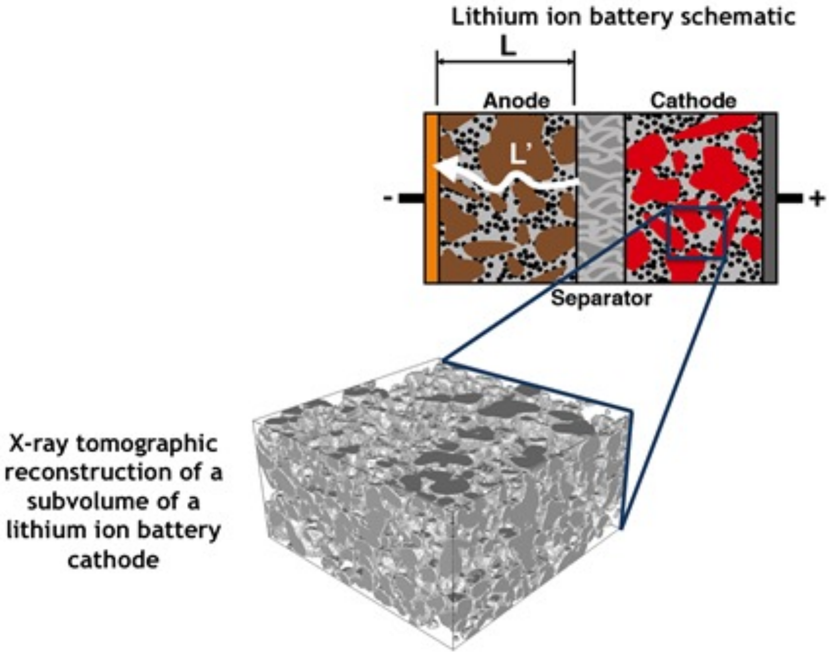


Flow and transport through porous media

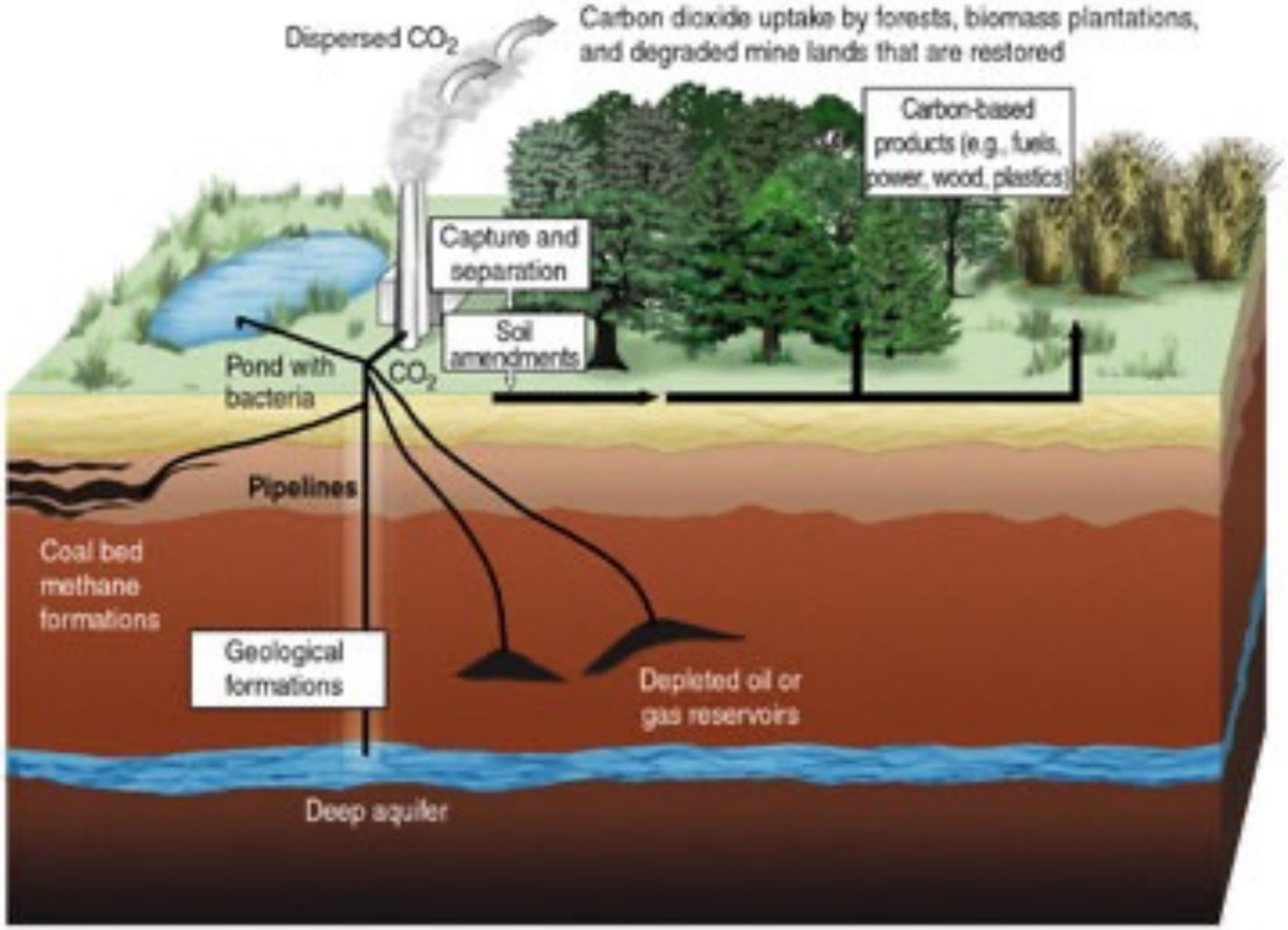
Li-ion battery

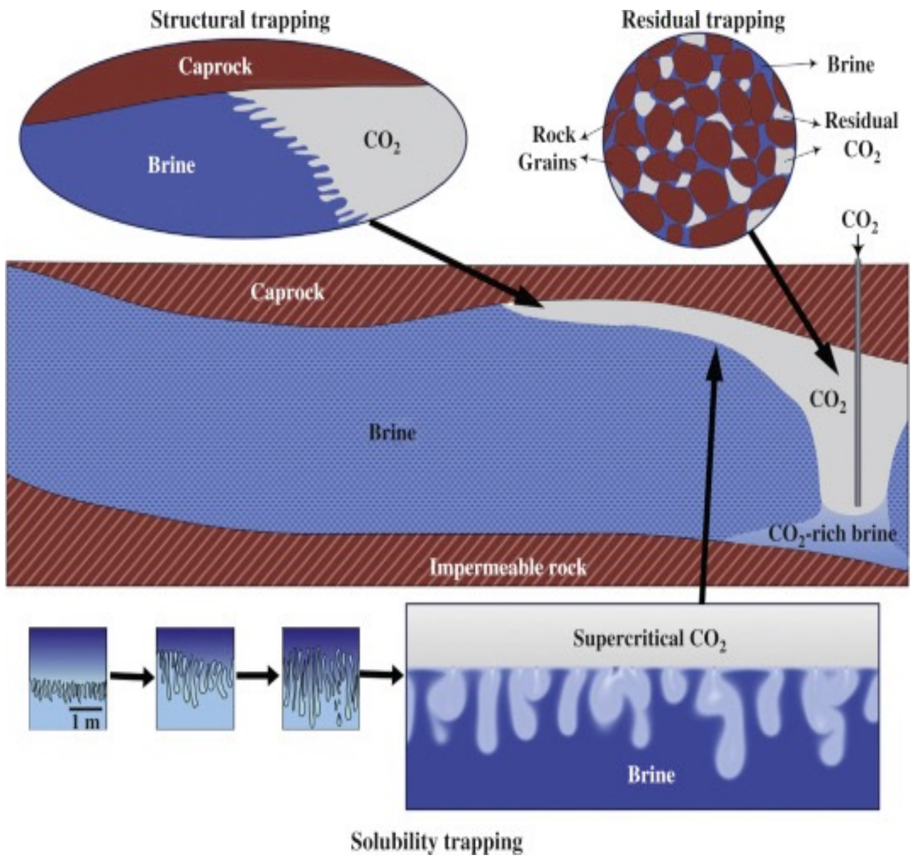


E.g. Electric cars



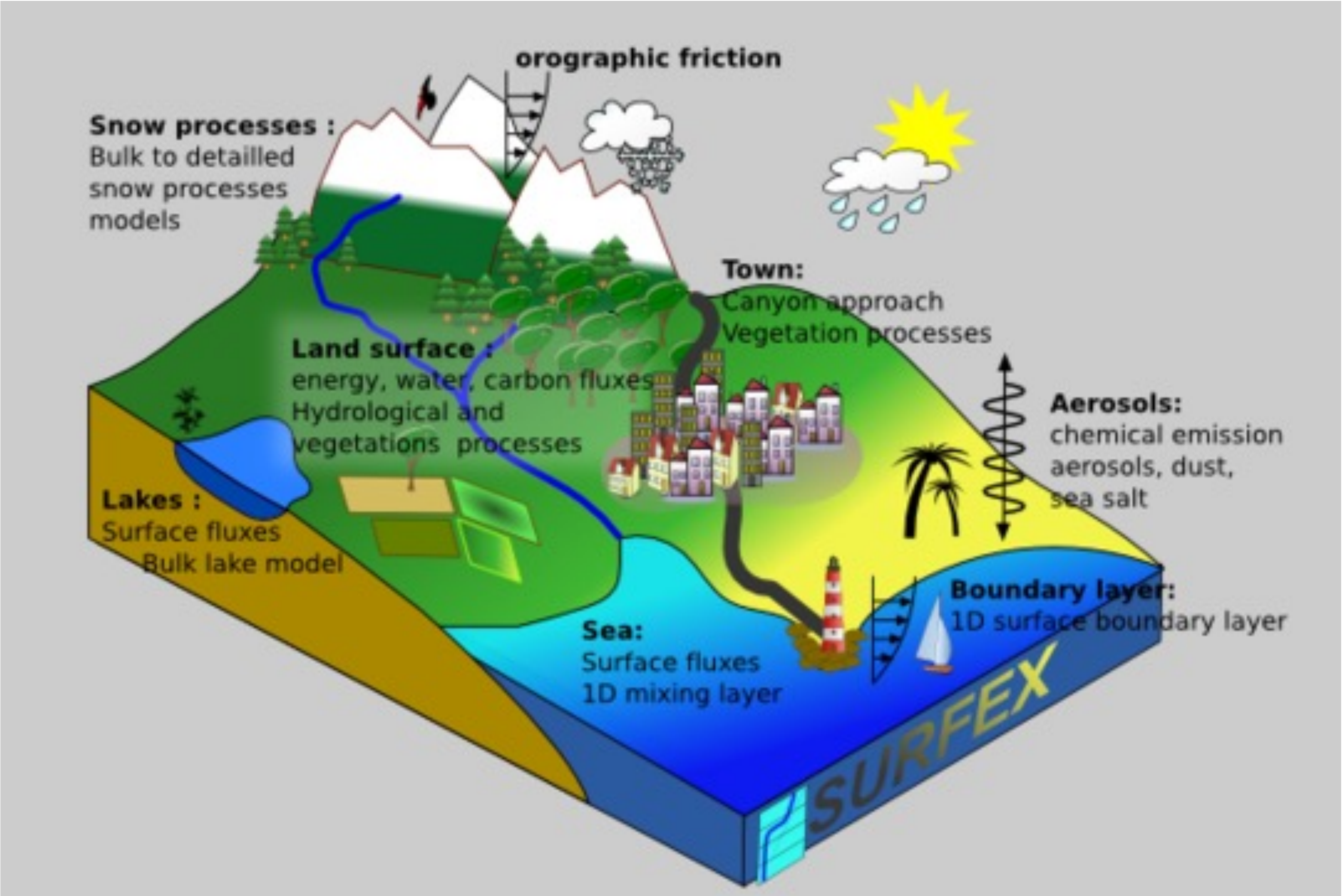
Carbon sequestration





Multiphase flow

Weather forecasting



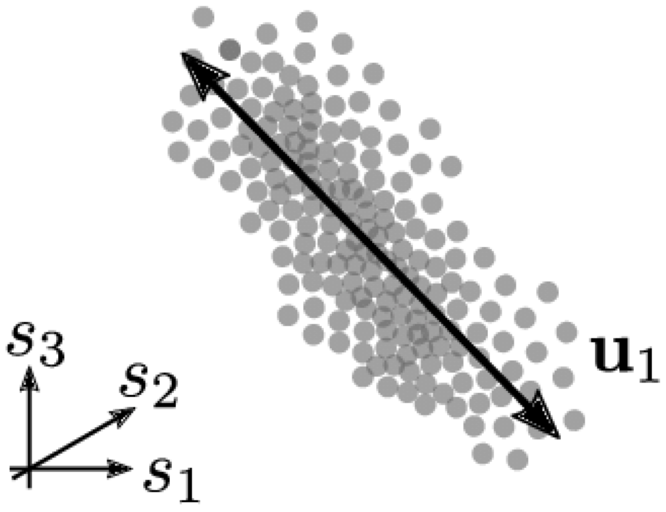
Coupled transport, wave, fluid models

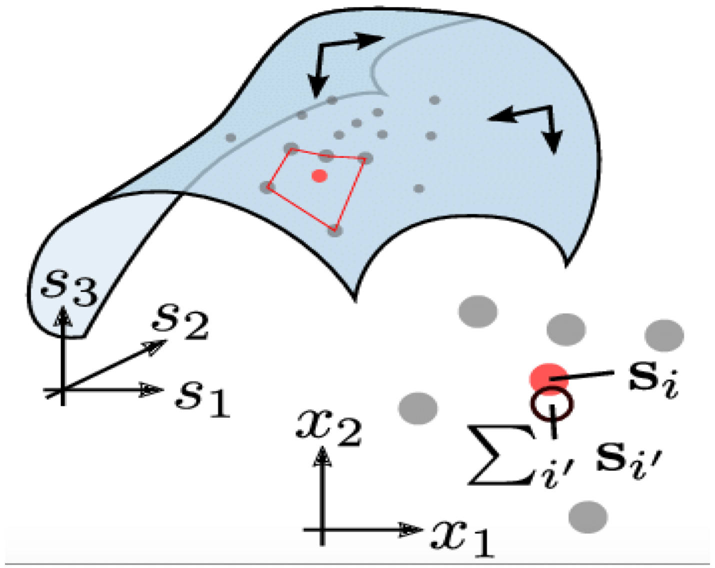
Data driven computational science

- Data driven technique is a new way to model, predict and control complex systems
- Many modern scientific and engineering problems do not have suitable empirical models or analytical models
- The success of data driven approaches is based on the availability of vast quantities of data
- Next, we briefly discuss some key ideas

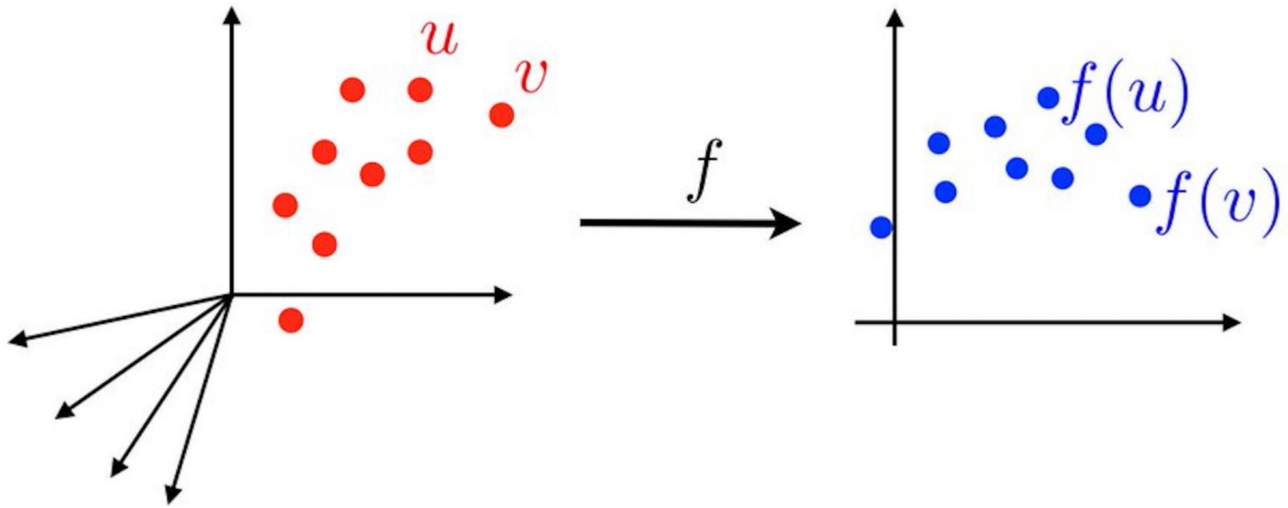
Dimensionality reduction

- Many complex systems exhibit dominant low-dimensional pattern
- Gives compact representation for modeling and control





The Johnson-Lindenstrauss Lemma



Given $0 < \varepsilon < 1$, a set X of m points in $\mathbb{R}^N$, and a number $n > 8\ln(m)/\varepsilon^2$, there is a linear map $f : \mathbb{R}^N \to \mathbb{R}^n$ such that

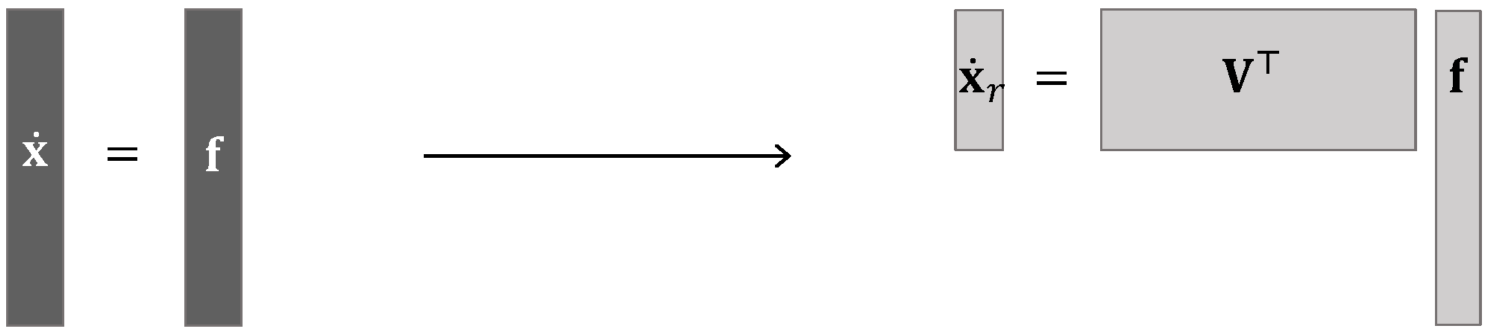
$$(1-\varepsilon)\|u-v\|^2 \le \|f(u)-f(v)\|^2 \le (1+\varepsilon)\|u-v\|^2$$

for all $u, v \in X$.

Data-driven coordinate transformation

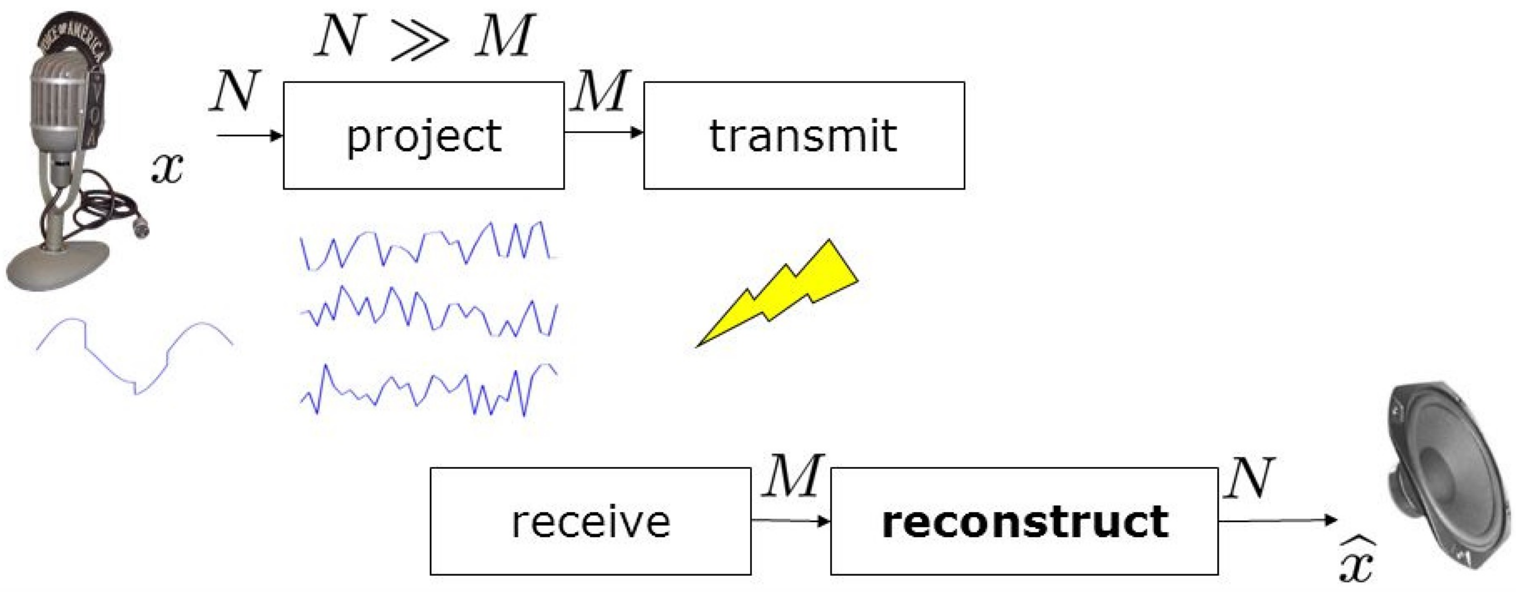
- Finding a coordinate system that can simplify the problem
- The basis is data-driven, and is tailored made
- It is a foundation for reduced order modeling

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \qquad\qquad \dot{\mathbf{x}}_r = \mathbf{V}^\top \mathbf{f}(\mathbf{V}\mathbf{x}_r)$$



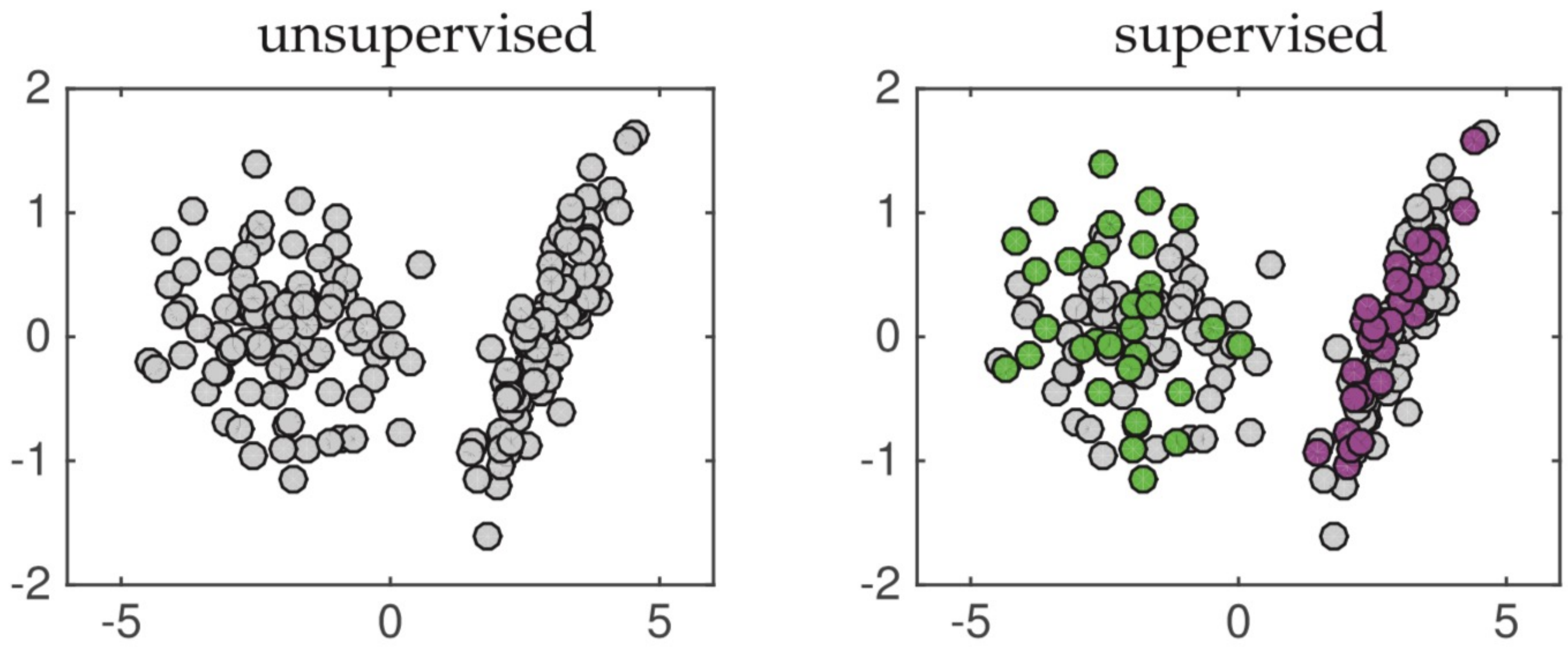
Compressed sensing

- Structures in data imply that the data admits a sparse representation in a suitable coordinate system
- Instead of collecting high dimensional measurement and then compressing, one can perform compressed measurement



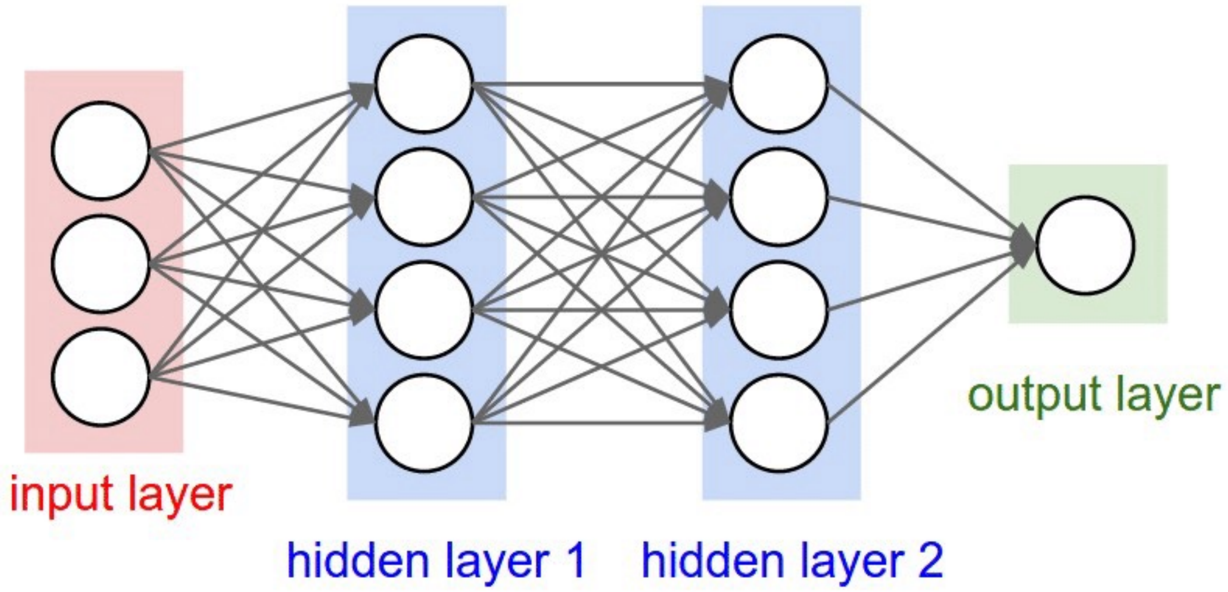
Machine learning

- Supervised vs unsupervised learning
- Supervised learning uses labelled data to extract pattern
- Unsupervised learning finds pattern without using labelled data



Neural networks

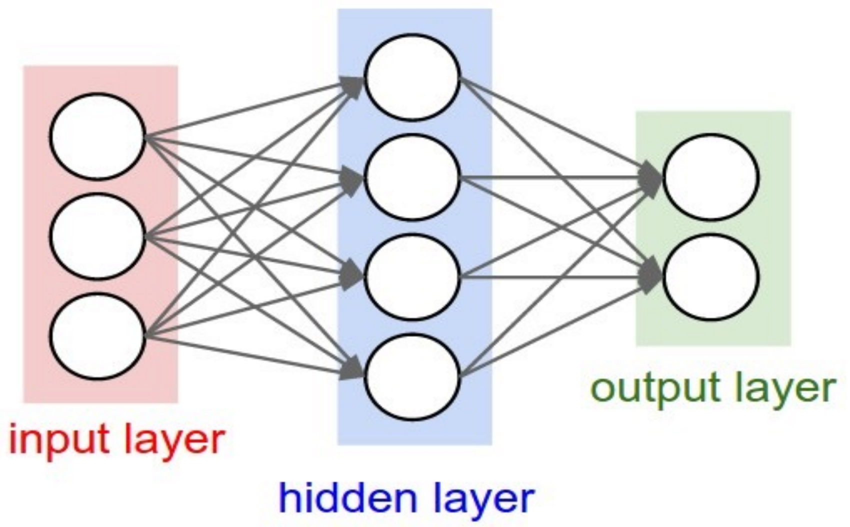
- Universal approximation theorem: neural network can approximate any continuous function on a compact set
- It gives an efficient way to compute a complicated nonlinear function
- Learning of unknown parameters in computational models

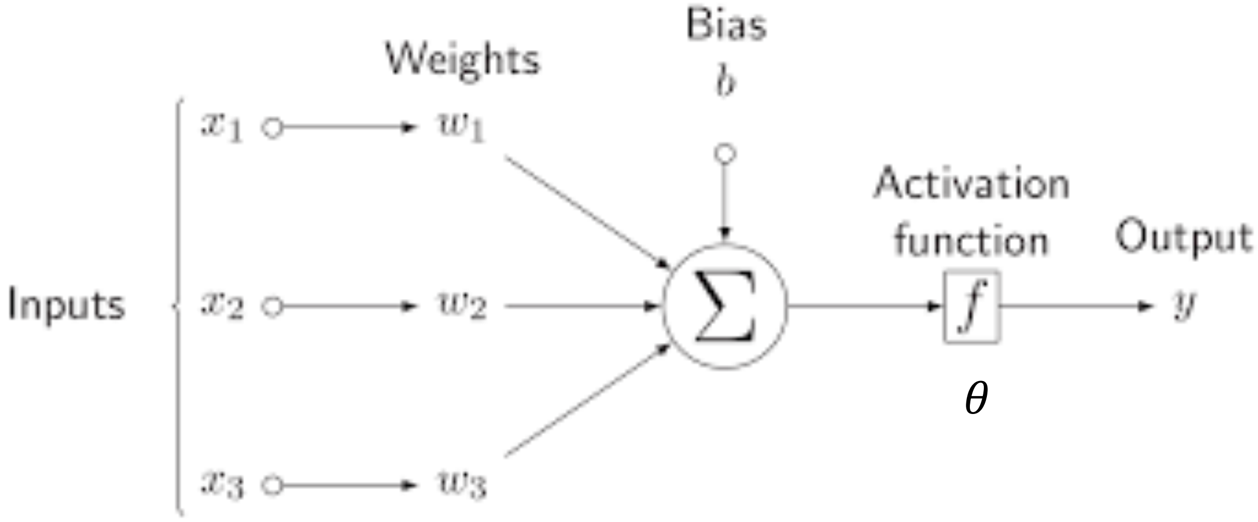


Universal approximation theorem

$$f \in C(\mathbb{D}^{(m)}), \quad \varepsilon > 0 \qquad F(x) = \sum_{i=1}^{N} v_i \theta(w_i^T x + b_i) \qquad \theta(y) = \max(0, y)$$

$$|F(x) - f(x)| < \varepsilon \quad \forall x \in \mathbb{D}^{(m)}$$

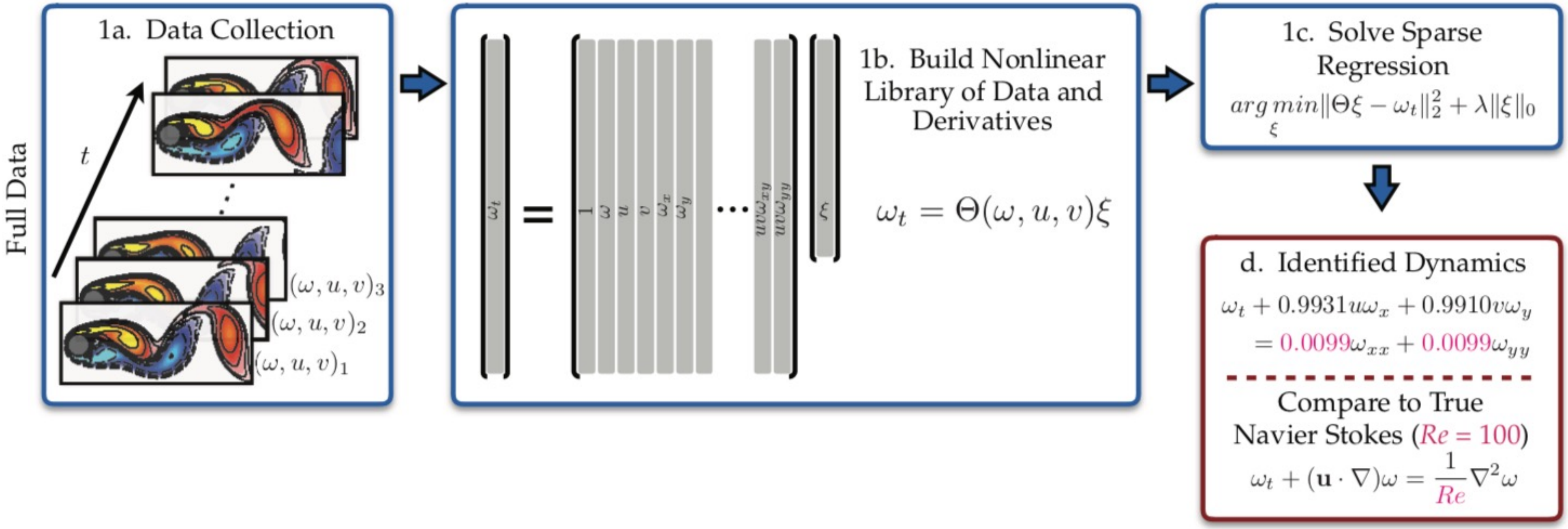




θ

Learning dynamical system

- Data is used to learn the underlying dynamics



Organization of the course

- Mondays (MMW 404)
  - Lectures
  - Overview of modern data-driven methods
  - Focus on the key ideas
  - Technical mathematical details will be discussed briefly
- Wednesdays (LSB 232B)
  - Lab sessions
  - Review on Python
  - Implementations of various methods

Goals

- An overview of modern mathematics for data driven computing
- We will focus on the ideas, instead of technical details
- Implementation of computational techniques for applications
- Develop skillsets for industry and academic research
- Let you know what’s going on in computational mathematics

Assessment scheme

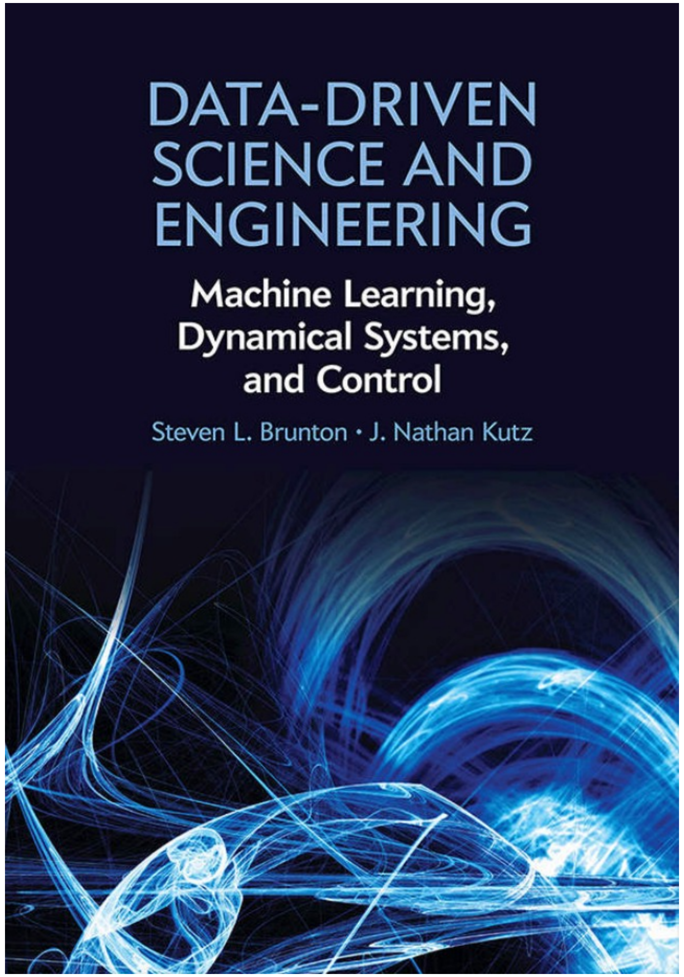
- 9 out of 10 lab exercises: 54%
- (some questions submitted after class)
- 2 open-book tests (in the form of lab exercises): 46%
- Test 1: Oct 11
- Test 2: Nov 29

Course website

- https://www.math.cuhk.edu.hk/course/2324/math4280

- Login information
  - Username: math4280
  - Password: math4280#2324*term1

Reference



DATA-DRIVEN SCIENCE AND ENGINEERING
Machine Learning, Dynamical Systems, and Control
Steven L. Brunton • J. Nathan Kutz

DEITEL DEVELOPER SERIES
python
Python
for Programmers
with introductory AI case studies
Natural Language Processing
Data Mining Twitter
IBM Watson
Machine Learning with scikit-learn
Deep Learning with Keras
Big Data with Hadoop, Spark, NoSQL and the Cloud
Internet of Things (IoT)
Python Standard Library
Data Science Libraries: NumPy, Pandas, SciPy, NLTK, TextBlob, Tweepy, Matplotlib, Seaborn, Folium and more
PAUL DEITEL • HARVEY DEITEL