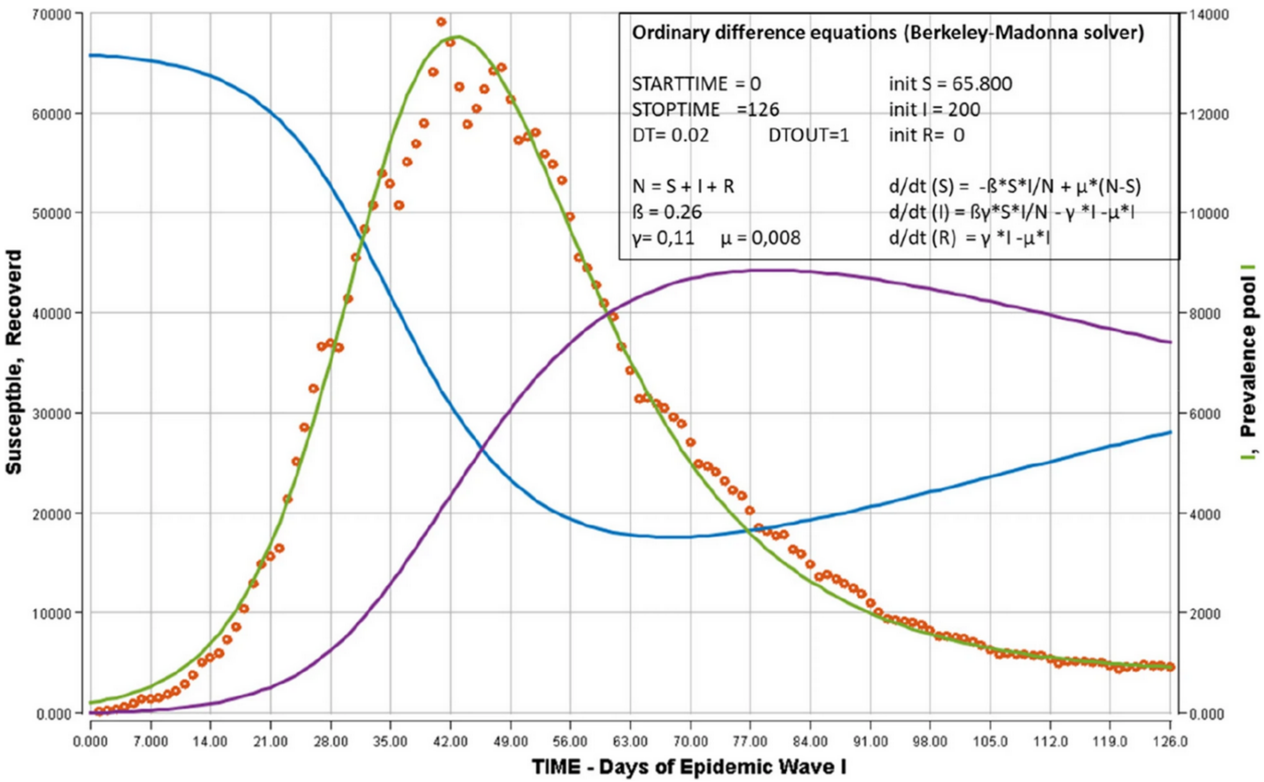
Foundations of Data Science

Lecture 1: Build, compute, critique, repeat

Prof. Gilles Louppe
g.louppe@uliege.be

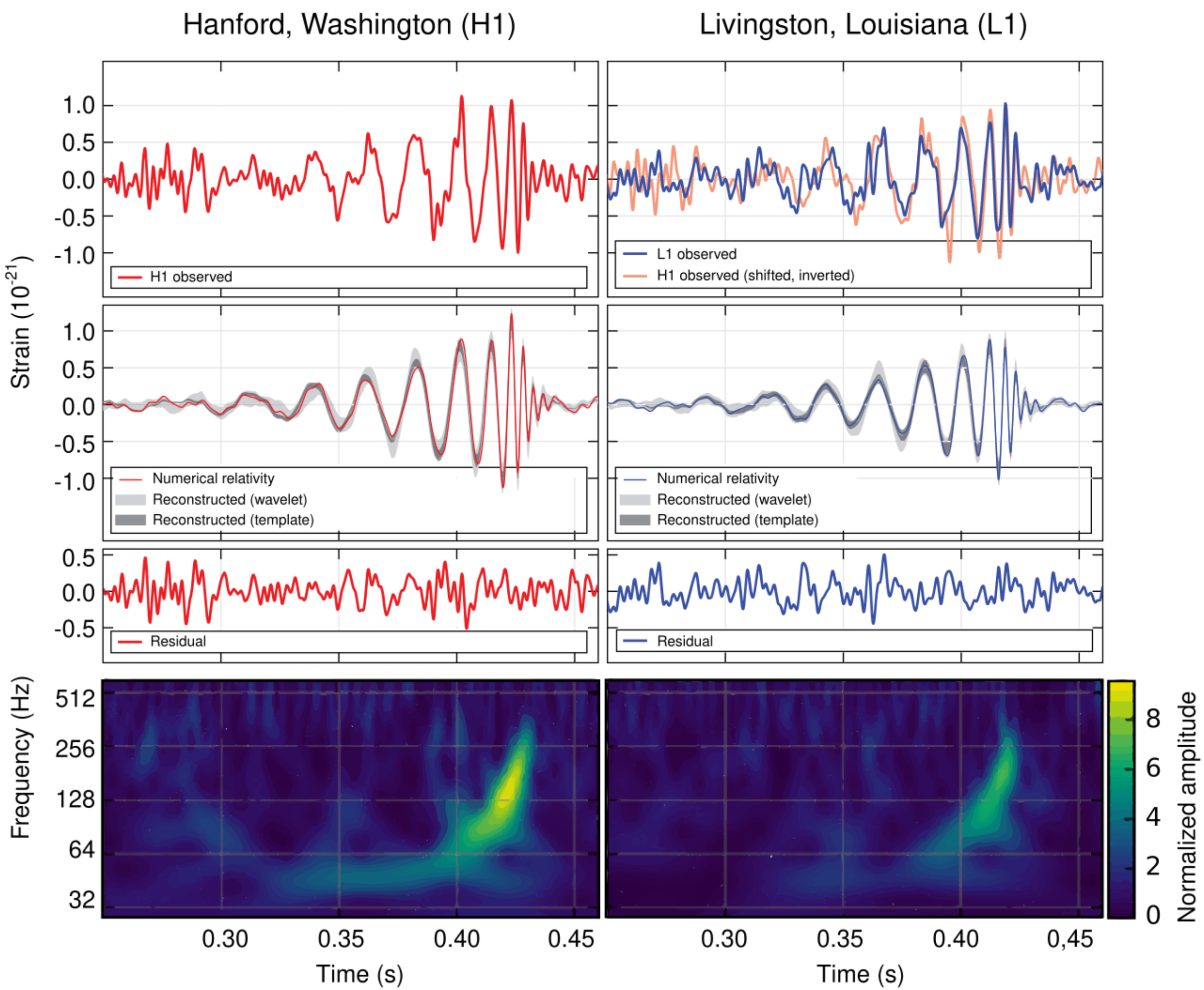
COVID-19 epidemiological models



Open SIR model fitted to observations of the first Covid-19 epidemic wave – Belgium. The evolution of daily prevalence pool I was estimated from the daily incidence of reported cases (red circles). These observed values were fitted by ordinary differential equations (ODE) to SIR model. Left vertical axis: modelized Susceptible compartment (blue); modelized Recovered compartment (violet). Right axis: modelized daily prevalence pool I compartment. The insert contains the equations of ODE in Berkeley-Madonna language, the initial values of S, I and R compartments, the ß and γ transition parameters. DT signifies that the fitting of model to observations with ODE was operated for each Δ time = 0,02 days. DTOUT indicates the Δ time for output printing

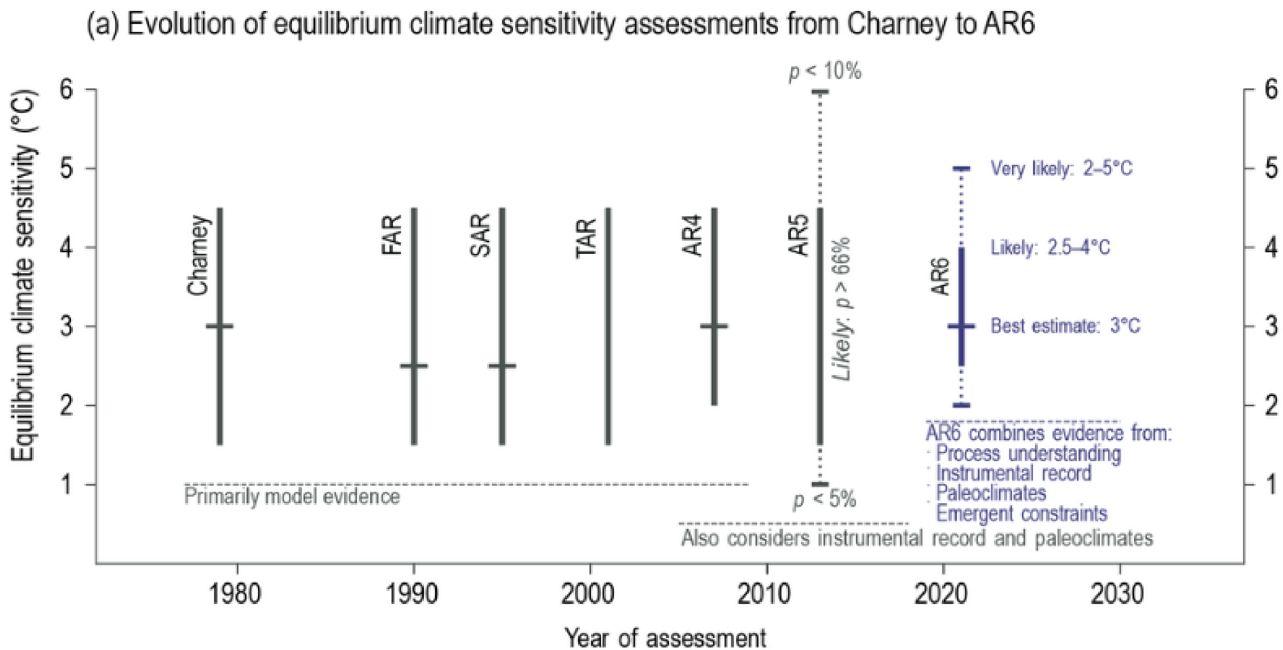
Predictions of the first wave of COVID-19 in Belgium, 2020. Many models failed to predict the peak and duration of the wave.

Gravitational wave detection



First direct detection of gravitational waves, LIGO, 2015. Many false alarms before the first confirmed detection.

Equilibrium climate sensitivity



What is the Earth's temperature increase if we double atmospheric CO2? Over 40 years, uncertainty narrowed from 3°C range to 1.5°C range.

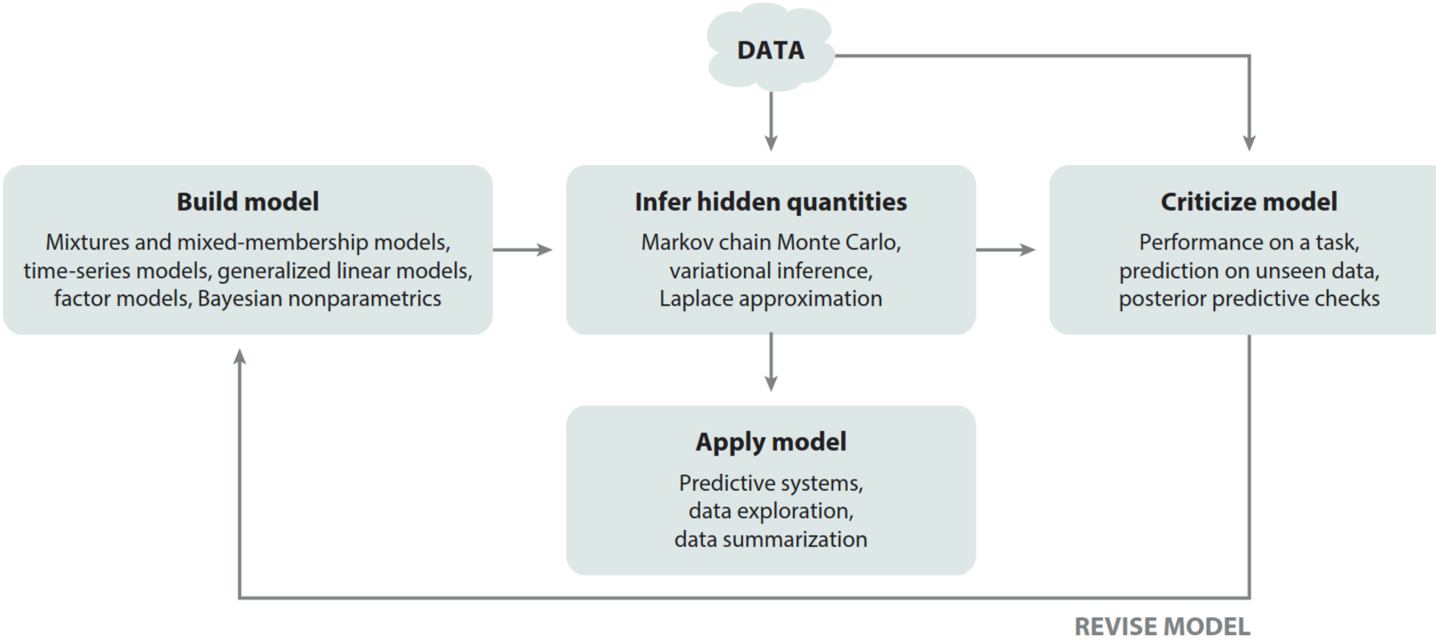
Box's loop

All models are wrong, but some are useful. -- George Box.

What is data science?

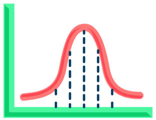
Data science is the discipline of **extracting knowledge** from data through the iterative application of the **scientific method**.

Box's loop



Scientific inquiry as an iterative process: build, compute, critique, repeat.

—
Credits: Blei, 2014.



Step 1: Build

The first step to understanding a phenomenon is to **build a model** of it, as a simplified representation that captures its essential aspects.

- A model specifies assumptions about the data generating process.
- It encodes domain knowledge and constraints.
- It is formulated within an appropriate mathematical abstraction.
- It defines what you observe and what you do not observe but assume exists.

Step 2: Compute

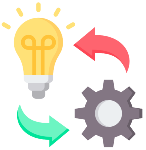
The next step is to **compute** what the data tells you about the phenomenon of interest under the assumptions of your model.

- Fitting the model to data involves solving an optimization problem.
- Inference is used to answer questions about unobserved quantities.
- Prediction is used to answer questions about future or unseen data.

Step 3: Critique

The third step is to **critique** the model and its predictions, to assess whether they are consistent with the data and domain knowledge.

- Compare predictions to observed data.
- Diagnose model fit and identify discrepancies.
- Scientific diagnostics: does it make sense?



Step 4: Repeat

What you learn from the critique step informs how to **repeat** the process.

- Add complexity to the model to address its shortcomings.
- Simplify the model to improve interpretability.
- Change the model to explore alternative hypotheses.

Why this approach matters?

The Fourth paradigm of science emphasizes the importance of data-intensive scientific discovery. However, data alone is not enough.

- Data without theory leads to spurious correlations.
- Theory without data leads to ungrounded speculation.
- Together, they enable robust scientific inquiry.

This principled approach contrasts with ad-hoc data analysis practices that often lead to unreliable results.

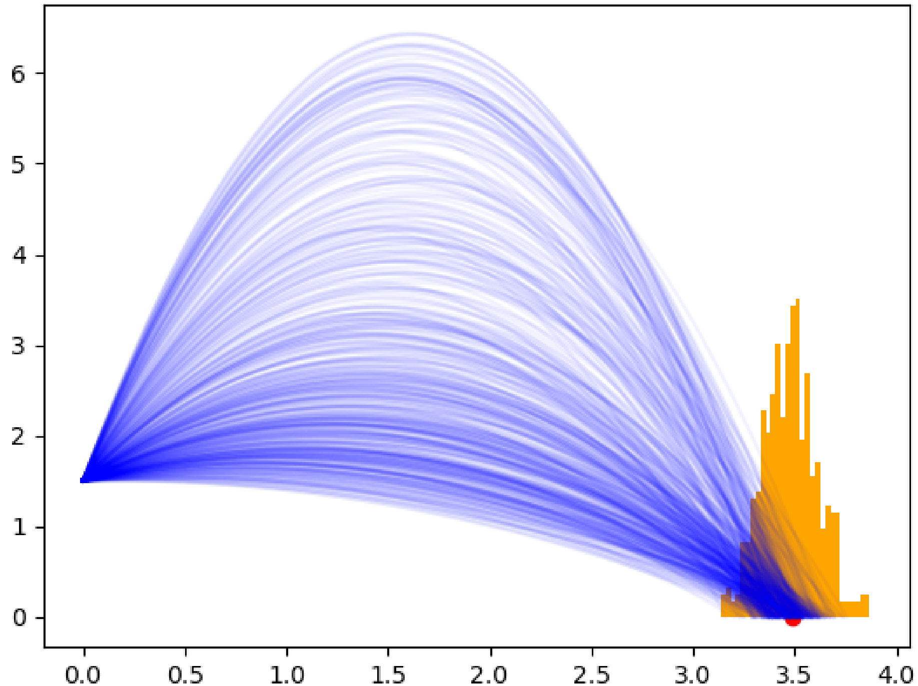
- Throw algorithms at data and see what sticks.
- Focus only on prediction accuracy.
- Treat models as black boxes.
- Stop at the first somewhat satisfactory result.

Box's loop encourages

- transparent reasoning about assumptions,
- understanding mechanisms behind data, not just correlations,
- honest assessment of model limitations,
- continuous improvement and learning.

Today's example: Projectile motion

A ball is thrown and lands at some measured distance x. What can infer about the initial velocity v and angle α of the throw?



Let's code!

- Open a Jupyter notebook and follow along (or check `nb01.ipynb` after the lecture).
- Implement Box's loop step by step:
    1. Build a simple physical model of projectile motion.
    2. Compute estimates of v and α from data.
    3. Critique the model fit and predictions.
    4. Repeat by refining the model to account for more realistic factors.

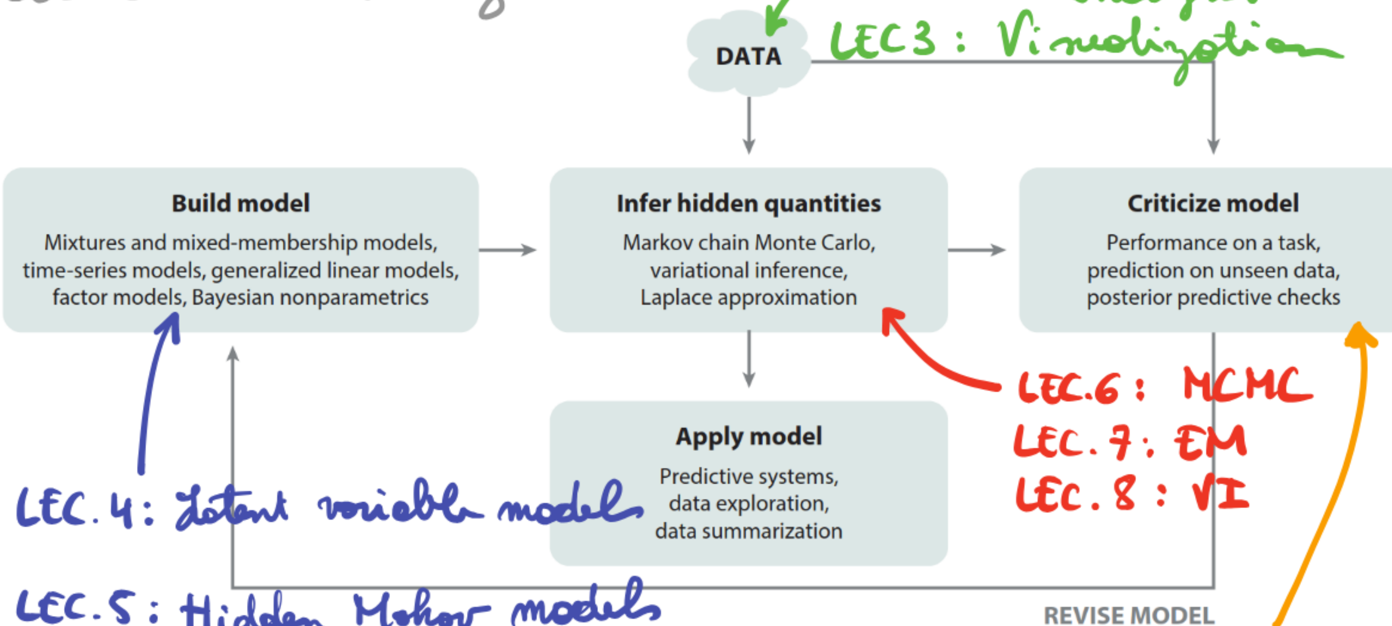
DATS0001

Outline

- Lecture 1: Build, compute, critique, repeat
- Lecture 2: Data and exploratory analysis
- Lecture 3: Visualization
- Lecture 4: Latent variable models
- Lecture 5: Hidden Markov models
- Lecture 6: Markov Chain Monte Carlo
- Lecture 7: Expectation-Maximization
- Lecture 8: Variational Inference
- Lecture 9: Model criticism and validation
- Lecture 10: Case study

LEC. 1: Box's loop
LEC 10: Case study

LEC2: Data exploratory analysis
LEC3: Visualization



LEC. 4: Latent variable models
LEC. 5: Hidden Markov models

LEC.6: MCMC
LEC. 7: EM
LEC. 8: VI

LEC9: Model criticism and comparison

My mission

Teach you how to think like a scientist in the age of data.

By the end of this course, you will be able to explore, model, and reason about data in a principled way, but also to communicate your findings effectively.

The end.