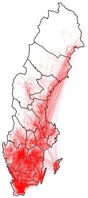
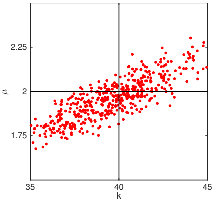
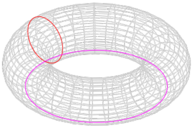
A case study of

Data-driven computational modeling in Epidemics:

bringing the dirt to the classroom







Stefan Engblom

Div of Scientific Computing, Dept of Information Technology, Uppsala University, Sweden

LTH Seminar

Lund, March 6th, 2017

Outline

1. The Case: national-scale epidemics
   - VTEC
   - Computational modeling
   - Parameter estimation and outcomes

2. Outlook
   - The role of the Lax principle
   - Model of models: the role of test equations

3. Bringing holistic computing into the classroom
   - Why design
   - A worked example: Applied finite elements

Summary

The Case: national-scale epidemics

- Ongoing research to better understand the spread of verotoxinogenic *E. coli* O157:H7 (VTEC O157:H7) in the Swedish cattle population
- *Zoonotic pathogen* (animal $\rightarrow$ human) of great public health interest, causing enteroheamorrhagic colitis (EHEC) in humans ($\sim$500 cases anually in Sweden, cost per case $\sim$24kSEK)

The Case: national-scale epidemics

- Ongoing research to better understand the spread of verotoxinogenic *E. coli* O157:H7 (VTEC O157:H7) in the Swedish cattle population
- *Zoonotic pathogen* (animal $\rightarrow$ human) of great public health interest, causing enteroheamorrhagic colitis (EHEC) in humans ($\sim$500 cases anually in Sweden, cost per case $\sim$24kSEK)
- "Understand" means to determine the dominating mechanisms in the dynamics, evaluate the effect of counter measures, investigate *"what ifs"*...
- Substantial amount of data available:
  - individual-level cattle data from 2005 and onwards ("events")
  - geographical and meteorological data
  - longitudal studies of farms

VTEC epidemics

Sverige

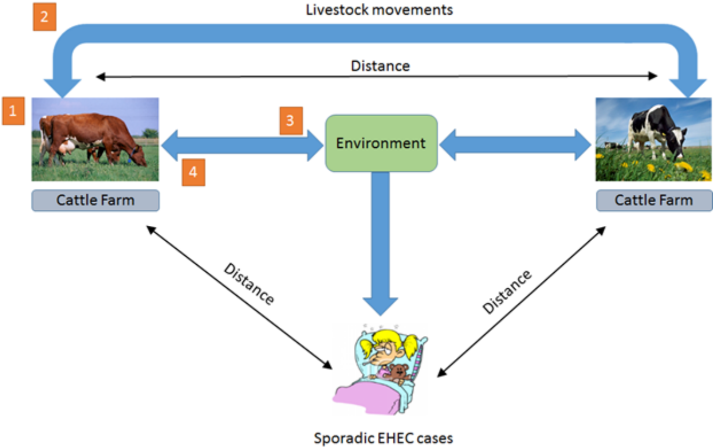
Flera barn sjuka i nytt utbrott bakteriesmitta

Sexton nya fall av ehec-smitta har konstaterats i Sverige. Sex av de
är barn och tre har fått den allvarliga följdsjukdomen hus. Merpart
i Stockholmsområdet, men det är ännu oklart vad som orsakat utb
– Den här sorten kan ge mer allvarliga komplikationer, säger Ande
Folkhälsomyndigheten.

Av Ninna Bengtsson
28 okt, 2016
Spara artikel

VTEC epidemics

in short



Infected animals show no signs of the disease!

Event data

by European Union law

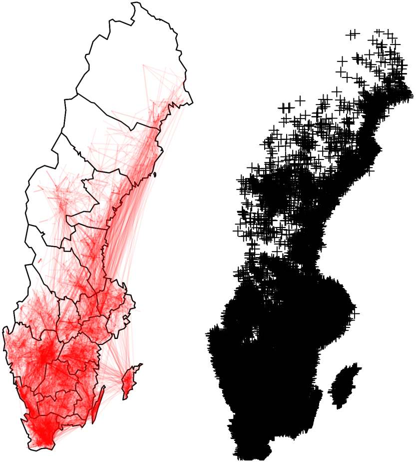
REPORTER	WHERE	ABATTOIR	DATE	EVENT	ANIMALID	BIRTHDATE
83466	83958	0	2009-10-01	2	SE0834660433	1997-04-04
83958	83466	0	2009-10-01	1	SE0834660433	1997-04-04
83958	83829	0	2012-03-15	2	SE0834660433	1997-04-04
83829	83958	0	2012-03-15	1	SE0834660433	1997-04-04
83829	83958	0	2012-03-15	4	SE0834660433	1997-04-04
54234	83829	0	2012-04-11	1	SE0834660433	1997-04-04
83829	54234	0	2012-04-11	2	SE0834660433	1997-04-04
83829	83958	0	2012-04-11	5	SE0834660433	1997-04-04

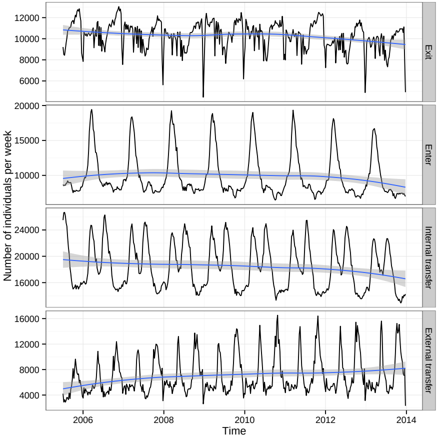
Total: 18 649 921 reports and 37 221 holdings

Events

- Exit (death, n=1 438 506)
- Enter (birth, n=3 479 000)
- Internal transfer (ageing, n=6 593 921)
- External transfer (transport between holdings, n=732 292)

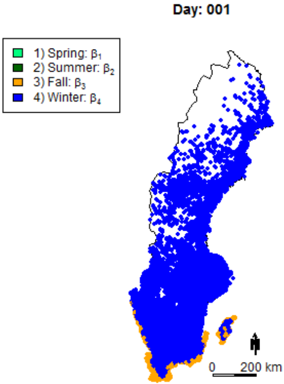
Event data




Exit
Enter
Internal transfer
External transfer
Number of individuals per week
Time
2006
2008
2010
2012
2014

Meterological data

by SMHI



Best practise

Modeling and parametrization in epidemics

Typically:

1. Highly coarse-grained models, e.g. "mosaic" ODEs, combined with rule of thumbs, and various types of data averaged and understood as fluxes and source terms
2. For a parameter proposal, a *single* model run is used to obtain a residual of some kind wrt some measured data
3. What parameter combination minimizes $\|residual\|_{some\ norm}$?

Best practise

Modeling and parametrization in epidemics

Typically:

1. Highly coarse-grained models, e.g. "mosaic" ODEs, combined with rule of thumbs, and various types of data averaged and understood as fluxes and source terms
2. For a parameter proposal, a *single* model run is used to obtain a residual of some kind wrt some measured data
3. What parameter combination minimizes $\|residual\|_{\text{some norm}}$?

-On the one hand, what is the uncertainty of the model so obtained?

-On the other hand, the problem is not easy! The *topology* of the model might be the research question. Data is scarce and expensive to collect...

Forming a model

a priori thoughts

The dynamics/epidemics is quite likely stochastic, nonlinear, spatially inhomogeneous...

Designing/understanding computational models: either we do

- "mosaic approach" relying on fingerspitzengefühl...
- or, relying on the **Lax principle**: *if the numerical physics $\approx$ the wanted "true" physics (consistency), then the numerical solution $\to$ the true solution (convergence) IFF the numerical physics is stable*

Local model

"SIS_E"

Model states: **S**usceptible, **I**nfected

State transitions

$$I \longrightarrow S \text{ at rate } \propto I(t)$$
$$S \longrightarrow I \text{ at rate } \propto S(t)\varphi(t)$$

80% of the holdings consist of <100 individuals. A suitable model for (S, I) is therefore a *continuous-time Markov chain*.

Local model

"SIS_E"

Model states: **S**usceptible, **I**nfected

State transitions

$$I \longrightarrow S \text{ at rate } \propto I(t)$$
$$S \longrightarrow I \text{ at rate } \propto S(t)\varphi(t)$$

80% of the holdings consist of <100 individuals. A suitable model for (S, I) is therefore a *continuous-time Markov chain*.

Environmental infectious pressure (plain ODE)

$$\frac{d\varphi}{dt} = \frac{I(t)}{S(t)+I(t)} - \beta(t)\varphi(t) + (...)$$

Global model

Stochastic reaction-transport framework

Put $\mathbb{X}_t^{(i)} = [S_{ij}, I_{ij}, \varphi_i]_t^T$ for $j \in \{calves, young\ stock, adults\}$ and $i = 1, \ldots, \sim$40,000 holdings.

$$d\mathbb{X}_t^{(i)} = \underbrace{\mathbb{S}\boldsymbol{\mu}^{(i)}(dt)}_{\text{local } SIS_E\text{-model+local events}} \underbrace{- \sum_{j \in C(i)} \mathbb{C}\boldsymbol{\nu}^{(i,j)}(dt) + \sum_{j;\, i \in C(j)} \mathbb{C}\boldsymbol{\nu}^{(j,i)}(dt)}_{\text{global events+physics}}.$$

Data now goes into all these forward operators.

The above general framework is implemented in SimInf (GitHub).

Numerical split-step method

Set-up

Local physics first, then global;

$$\tilde{\mathbb{X}}_{n+1}^{(i)} = \mathbb{X}_n^{(i)} + \int_{t_n}^{t_{n+1}} \mathbb{S}\boldsymbol{\mu}^{(i)}(\tilde{\mathbb{X}}^{(i)}(s);\ ds),$$

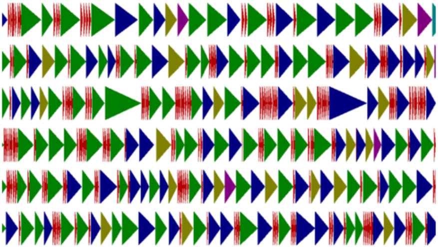
$$\mathbb{X}_{n+1}^{(i)} = \tilde{\mathbb{X}}_{n+1}^{(i)} - \int_{t_n}^{t_{n+1}} \sum_{j \in C(i)} \mathbb{C}\boldsymbol{\nu}^{(i,j)}(\mathbb{X}^{(i)}(s);\ ds)$$
$$+ \int_{t_n}^{t_{n+1}} \sum_{j;\, i \in C(j)} \mathbb{C}\boldsymbol{\nu}^{(j,i)}(\mathbb{X}^{(i)}(s);\ ds)$$

Assume (certain assumptions). Then

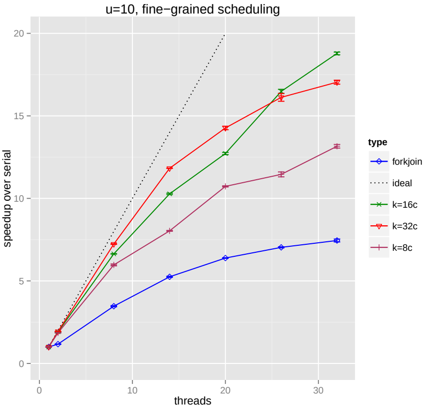
- $\mathbb{E}[\sup_{t_n \in [0,t]} \|\mathbb{X}_n\|_{\boldsymbol{l}}^p]$ bounded, any $p \geq 1$ (stability)
- $\mathbb{E}[\|\mathbb{X}_n - \mathbb{X}(t_n)\|^2] = O(h),\ h = max_n(t_{n+1} - t_n)$ (convergence)

Parallel implementation

Dependency-aware scheduling via task-based framework

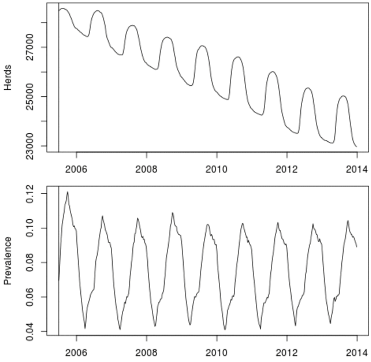


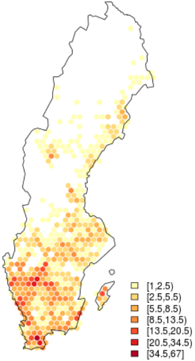
6 core task execution trace; red tasks are dependent steps (requiring thread synchronization).



Sample simulation

∼9 years of actual data





($\sim 10^8$ data base events plus $\sim 10^9$ infectious events during 9 years simulated in 15s on a desktop)

Feasibility of parameter estimation

Synthetic data ("inverse crime")

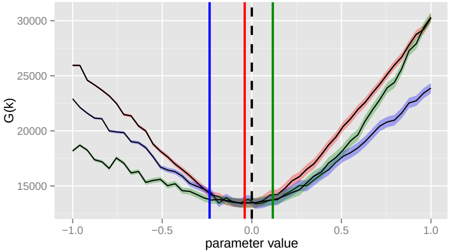
Setup: determine $\hat{k} = \arg\min_k G(k)$,

$$G(k)^2 = M^{-1} \sum_{i=1}^{M} \|\mathcal{F} \circ \mathbb{X}^{(i)}_{\text{simulated}}(k) - \mathcal{F} \circ \mathbb{X}_{\text{input}}(k^*)\|^2,$$

$\mathcal{F}$ a "summary statistics" / "measurement filter" (...)

Using $M \in \{10, 20, 40\}$ simulations for G and $N = 20$ iterations of an optimization routine:

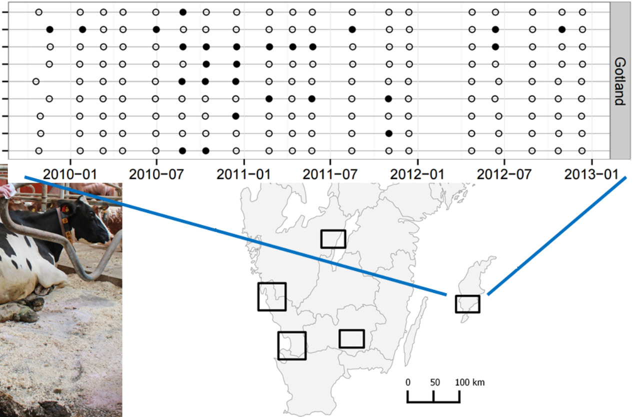
M	Residual	12 cores	32 cores
10	0.174	46.6 min	30.2 min
20	0.090	94.2 min	61.5 min
40	0.036	189.3 min	123.7 min



Parameter estimation

Real data

126 holdings sampled regularly during 38 months; $\sim$17 swipe samples per group of 3 animals. Probability(test positive$|n$ individuals infected), $n \in \{0, 1, 2, 3\}$ estimated via detailed studies *a priori*.



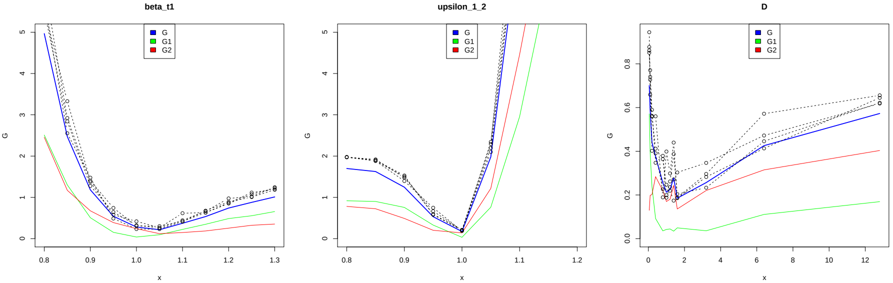
Parameter estimation

Real data, but *after* testing the equivalent synthetic situation first!

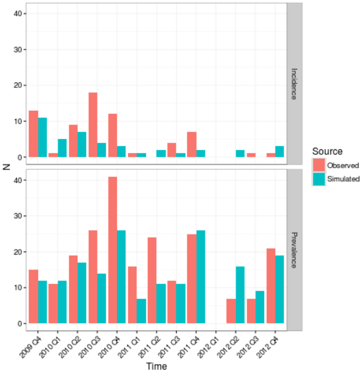
Setup: determine $\hat{k} = \arg\min_k G(k)$,

$$G(k)^2 = M^{-1} \sum_{i=1}^{M} \|\mathcal{F} \circ \mathbb{X}^{(i)}_{\text{simulated}}(k) - \mathcal{F}^*_{\text{measured}}\|^2,$$

$\mathcal{F}$ is now the probabilistic map from state $\mathbb{X}$ to sample $\{0, 1\}$.

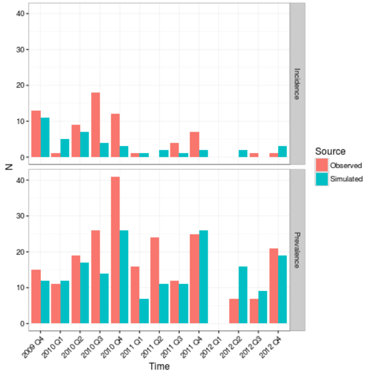


Outcome



- On the one hand, "an answer", a parametrized model
- More importantly, and usually from mistakes/misfits: a better understanding of the dynamics, of the interplay between parameters, an efficient procedure to find optimal models among suggestions...

Outcome



- On the one hand, "an answer", a parametrized model
- More importantly, and usually from mistakes/misfits: a better understanding of the dynamics, of the interplay between parameters, an efficient procedure to find optimal models among suggestions...

Finding #1: decay $\beta = \beta(t)$ required in the Swedish climate.

Finding #2: a mathematical analysis reveals a finite-time extinction in the stochastic model, contrary to a corresponding deterministic model.

"The purpose of computing is insight, not numbers." (R. Hamming)

The role of the Lax principle

Complex modeling situations

On the one hand, convergence to a well-defined "truth" from consistency and stability is neccessary...

...but it's not enough. Parameters need to be *observable* too. And observable from data that can actually be collected!

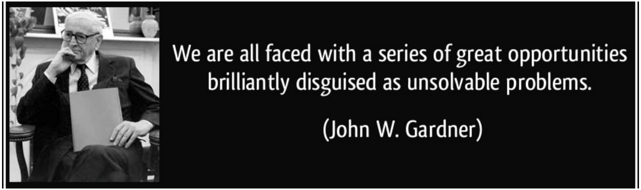
The role of the Lax principle

Complex modeling situations

On the one hand, convergence to a well-defined "truth" from consistency and stability is neccessary...

...but it's not enough. Parameters need to be *observable* too. And observable from data that can actually be collected!

$\Longrightarrow$ Understanding the limits is important, knowing what you cannot do is a good thing! *Negative reasoning* is a good take-away.



(http://izquotes.com)

The role of test equations

Model of models!

Classical model $y' = \lambda y$, one parameter $\lambda \in \mathbb{C}$

Study convergence $y_h \to y$ as $h \to 0$, but what about $y_{h,N} \to y$ with N the # observations to estimate $\lambda_N \approx \lambda$?

The role of test equations

Model of models!

Classical model $y' = \lambda y$, one parameter $\lambda \in \mathbb{C}$

Study convergence $y_h \to y$ as $h \to 0$, but what about $y_{h,N} \to y$ with N the # observations to estimate $\lambda_N \approx \lambda$?

$\Longrightarrow$ Linear birth-death model $y' = k - \mu y$, a model of a *source* and a *sink*.

Such sources/sinks typically occur in more than one place, e.g.,

$$S' \propto I - \varphi S,$$
$$\varphi' \propto I - \varphi,$$

(hence only one of the birth-constants, the "ks", can be non-dimensionalized away).

The role of test equations

Model of models!

Classical model $y' = \lambda y$, one parameter $\lambda \in \mathbb{C}$

Study convergence $y_h \to y$ as $h \to 0$, but what about $y_{h,N} \to y$ with N the # observations to estimate $\lambda_N \approx \lambda$?

$\Longrightarrow$ Linear birth-death model $y' = k - \mu y$, a model of a *source* and a *sink*.

Such sources/sinks typically occur in more than one place, e.g.,

$$\begin{aligned} S' &\propto I - \varphi S, \\ \varphi' &\propto I - \varphi, \end{aligned}$$

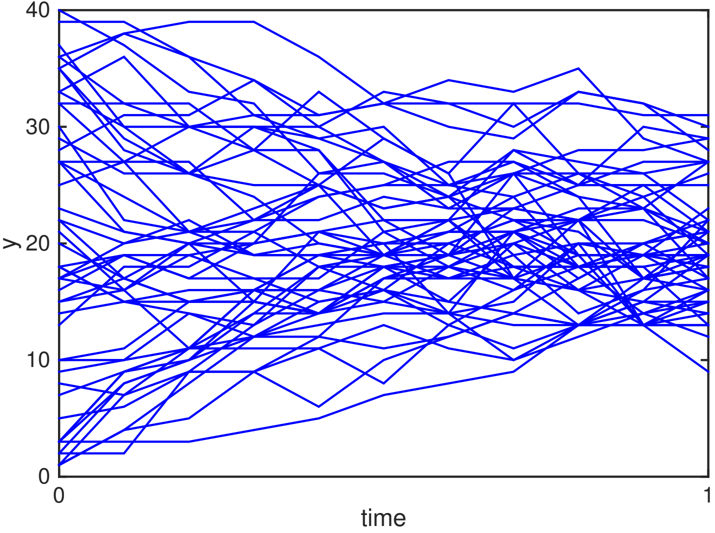
(hence only one of the birth-constants, the "ks", can be non-dimensionalized away).

-Is it doable to get (k, μ) from observations?

-*Negative reasoning:* if it won't work when almost everything is known, it won't work when confronted with a more realistic situation...

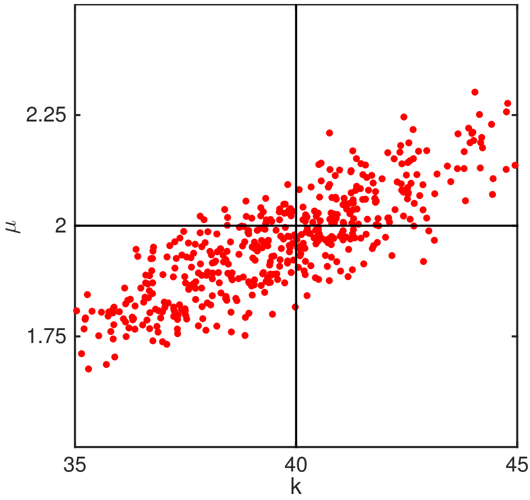
Data

$N = 50$ trajectories, sampled 10 times each



Posterior

MCMC using exact likelihood



Holistic computing to the classroom

It's all about design!

```
Design, transitive verb:
to create, fashion, execute, or construct according to plan
```

Design as a task requires:

1. a working "forward model" — *(I tend to stop here!)*
2. a way to find the parameters
3. a method to make it plausible that the design is sound
4. communication of the result

Project – Design a Medical Torus

Applied Finite Element Methods 1TD056 (5.0 hp)

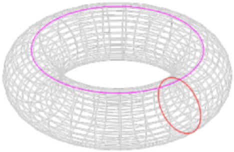
Murtazo Nazarov, Hanna Holmgren

October 28, 2016

Introduction

The overall grand project is to design a medical torus which emits a hormone through diffusion. The design variables are the shape of the torus and the initial concentration of hormone in the torus. The design constraints are the time until a certain amount of hormone has been emitted. Additional requirements are robustness to manufacturing errors and to model errors.

In this project we consider a medical torus of outer radius R and inner radius r that may be used in certain hormone treatments (see Fig. 1). It is inserted just under the skin on the inside of the upper arm where the conditions are such that the hormone will slowly diffuse out in the surrounding tissue.



A huge task

...solved in pieces

"The design constraints are the time until a certain amount of hormone has been emitted. Additional requirements are robustness to manufacturing errors and to model errors."

"Your task is to **investigate** the physics and the numerical modeling associated with this problem, **design** the torus so that it fulfills certain conditions and requirements, and **communicate** the design and the qualities of the design in a convincing way. Your results should support an expert committe in reaching a decision concerning the design of a product."

- Part A 1D simplification, show convergence of adaptive FEM
- Part B 2D simplification, implement in Matlab (assembly, solving)
- Part C 3D using FEniCS, *final design*

Evaluation comments

Student voices:

- *"The project was very enjoyable, and I liked very much that the course was so project-driven. Having a large "final project" that you build upon continually, rather than three disjoint assignments, was great and I think more courses should follow that recipe."*
- *"I gotta say, the assignments were rather hard. But I feel like I've learned so much and I think the assignments probably are the most important part of the course."*
- *"I really liked working with the project. It really sparked an interest for finite element methods for me, and I feel like I learned a lot."*

(On the teacher's side: helps to motivate what goes in or out of a course...)

Summary

- Case of national-scale computational modeling in Epidemics, incorporating large amounts of data (data bases, internet)
- Consistent modeling and the Lax principle $\Longrightarrow$ well-posedness, stability, consistency, convergence
- Efficient simulation, numerical method designed in order to expose parallelism ($\sim 10^8$ data base events plus $\sim 10^9$ infectious events during 9 years simulated in 15s on a desktop)
- $\Longrightarrow$ Parametrization of a national-scale model solved in SimInf (GitHub), interesting findings when fitting parameters to data
- At the meta-level: the actual role of stability, consistence, observability, test equations...
- Discrepancy wrt what we tend to teach
- $\Longrightarrow$ A worked example of *"PBL Light"*

Acknowledgment

Joint work with:

- Pavol Bauer (PhD-student, Uppsala university)
- Augustin Chevallier (MSc-student, ENS Cachan/INRIA Sophia Antipolis)
- Stefan Widgren (National Veterinary Institute)
- Murtazo Nazarov and Hanna Holmgren (Uppsala university)

Thanks!

Programs, Papers, and Preprints are available from my web-page.
Thank you for the attention!