The digital transition: applications of machine learning to marketing, engineering sciences, and medicine

Nicolas Vayatis, ENS Cachan

June 30, 2014

A Center for Data Science - Why now?

Technology

Society

(Data) Science

???

In this talk

Three examples

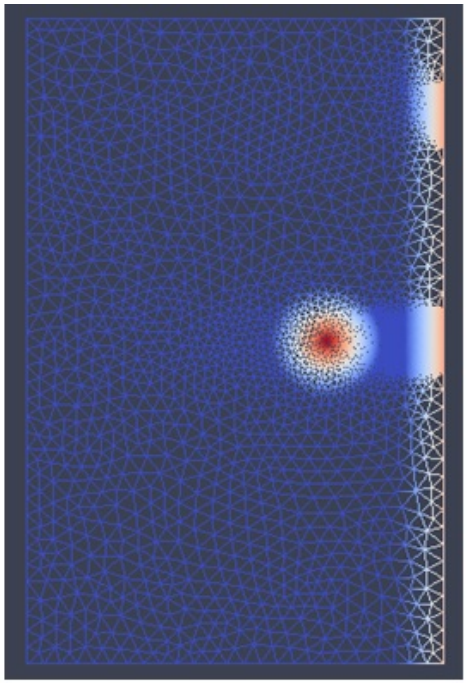
- Tsunami modeling
  - experimental design, sequential optimization, active learning
- Viral marketing and epidemics
  - diffusion process, complex networks, graph theory
- Routine motion quantification for medicine
  - signal processing, exploratory data analyis, classification methods

Challenges

- Automated (help to) decision making
- Computer-aided data exploration for science and health care

Example 1 - Experimental design for tsunami simulation

Tsunami Wave Basin at Oregon State University



Adaptive mesh grid of the VOLNA solver
[Dutykh et al., 2011]

Joint work with: Emile Contal, , Frédéric Dias*, Themis Stefanakis*, Costas Synolakis*, David Buffoni*, Alexandre Robicquet*

Experimental design - Goals and constraints

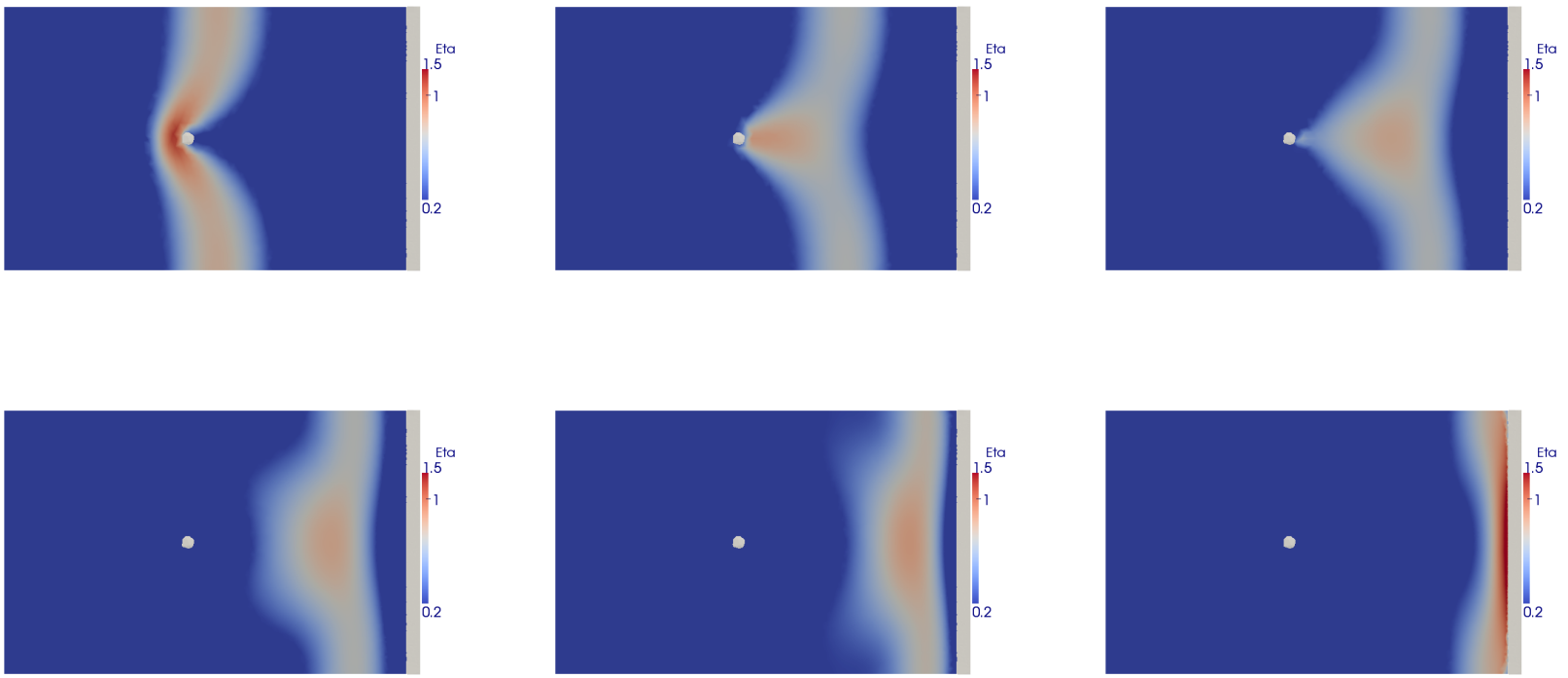
Possible goals

- Analysis/Control of the system
- Inverse problem and complex system design
- Optimization of the output $\Leftarrow$

Constraints

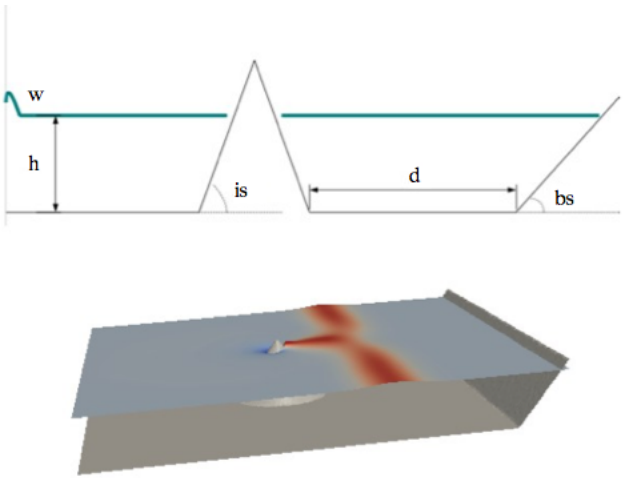
- Many input variables (i.e. parameters that drive the simulation)
- High cost of one experiment (gives one data point)
- Overall budget constraints: time and resources

Tsunamis amplification phenomena



Numerical simulations of a tsunami amplification generated by a conical island

Tsunami modeling example - Simulation setup

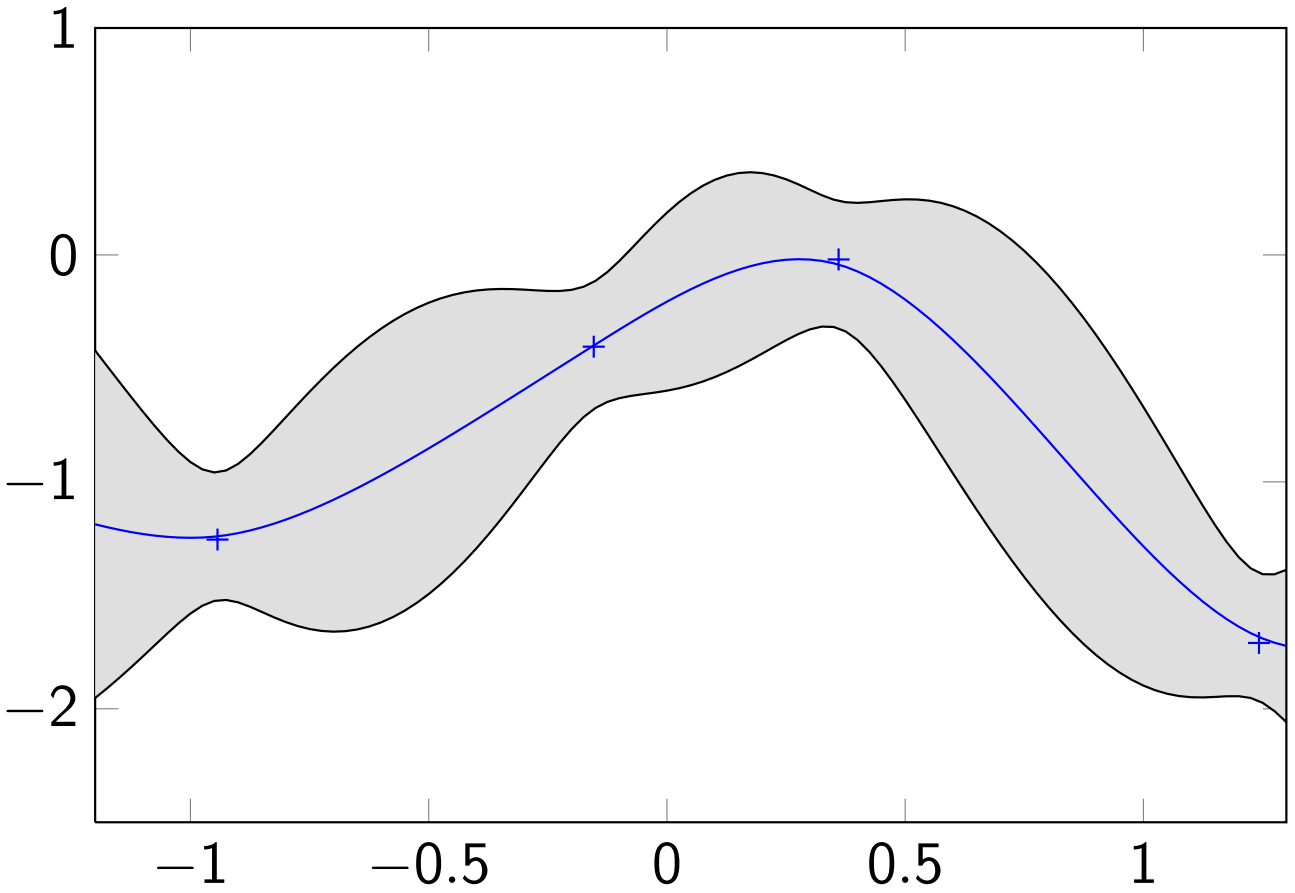


Five parameters modelling the geometry stored in a vector x

Goal

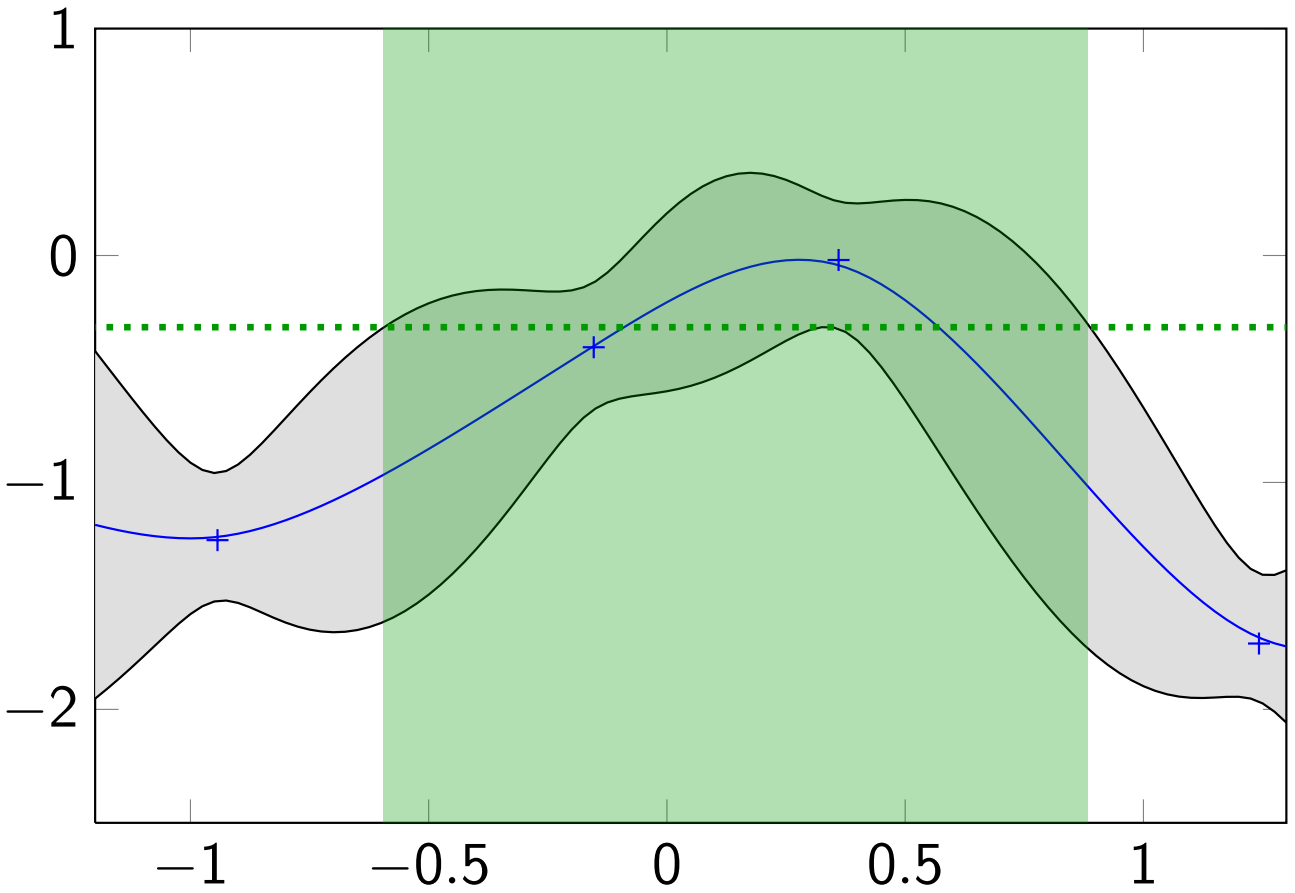
- Denote by f the unknown function relating topographic parameters x to runup amplification
- Consider access to $K \geq 2$ processors with time horizon $T \geq 2$
- Find the maximal value of f with T batches of size K

Main ingredient - Confidence bands based on gaussian processes



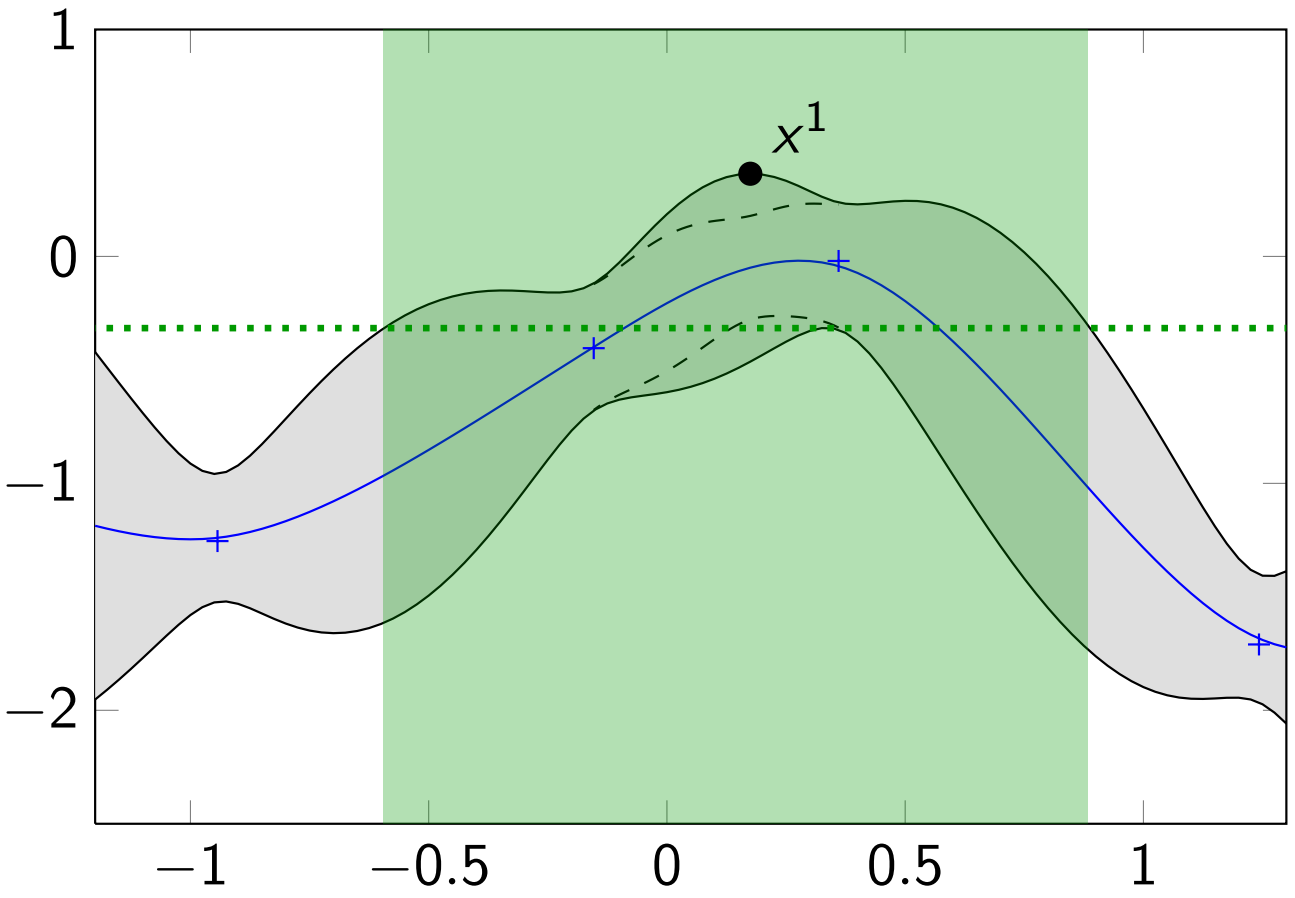
After bayesian inference obtained with four points on a 1D toy example

Main idea - Relevant region



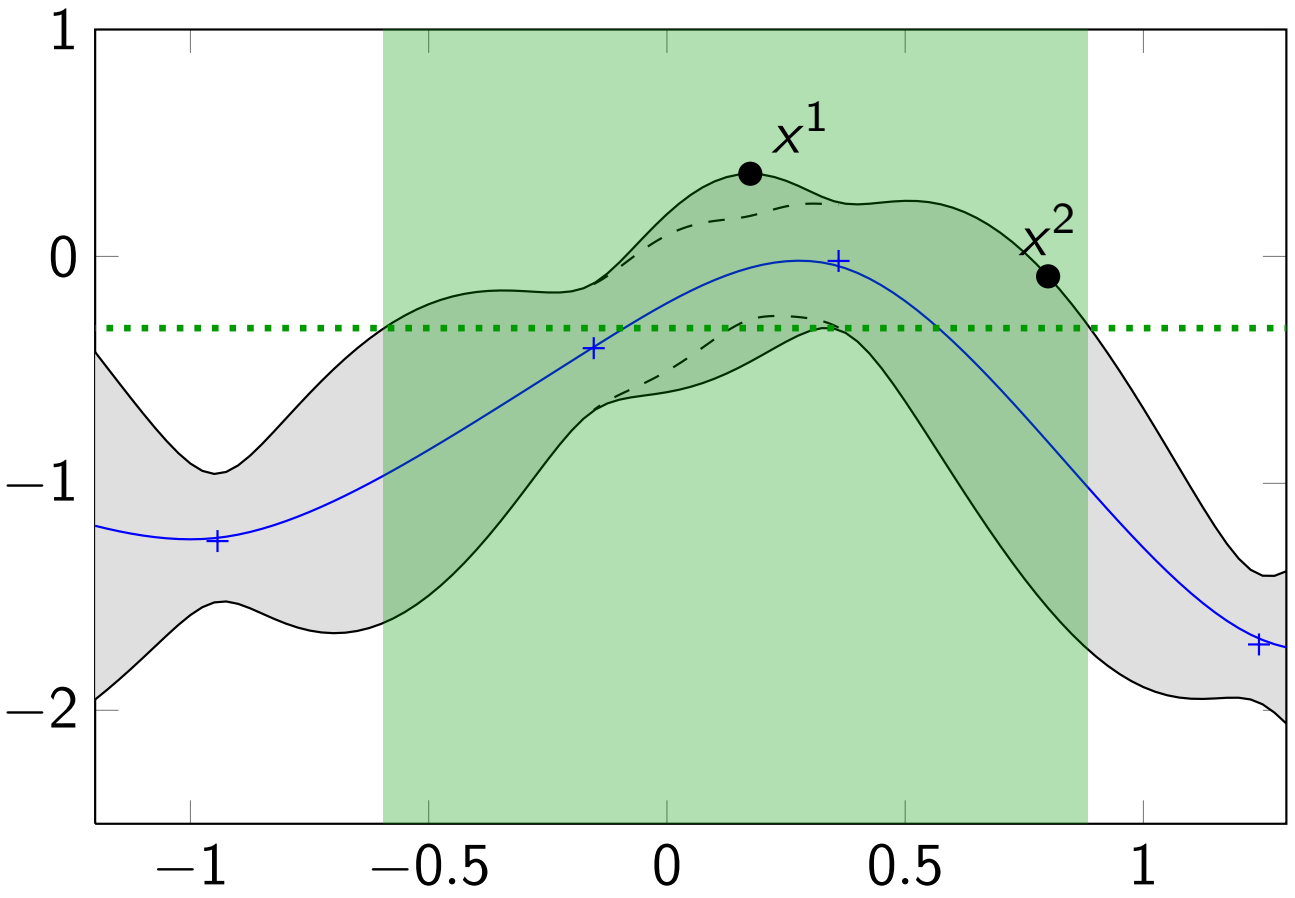
Based on the level set corresponding to the max of the lower bound

The GP-UCB-PE algorithm [Contal *et al.*, 2013]



UCB = Upper-Confidence-Bound $\Rightarrow$ Exploitation (1 point out of K)
PE = Pure exploration $\Rightarrow$ Exploration ($K-1$ remaining points in the batch)

The GP-UCB-PE algorithm [Contal *et al.*, 2013]



UCB = Upper-Confidence-Bound $\Rightarrow$ Exploitation (1 point out of K)
PE = Pure exploration $\Rightarrow$ Exploration ($K-1$ remaining points in the batch)

Theoretical Analysis

Theorem (Contal *et al.*, 2013)

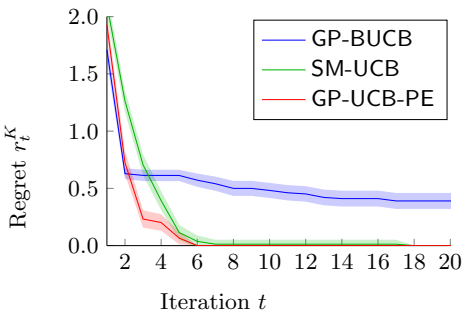
Consider $f \sim \mathcal{GP}(0, k)$ *with* $k(x, x) \leq 1$ *for all* x*, and* $x^\star = \text{argmax}_{x \in \mathcal{X}} f(x)$*, then we have, with high probability:*

$$R_T^K \doteq \sum_{t=1}^{T} \left(f(x^\star) - \max_{1 \leq k \leq K} f(x_t^k) \right) = \mathcal{O}\left(\sqrt{\left(\frac{T}{K} \right) \gamma(T, K) \log T} \right)$$

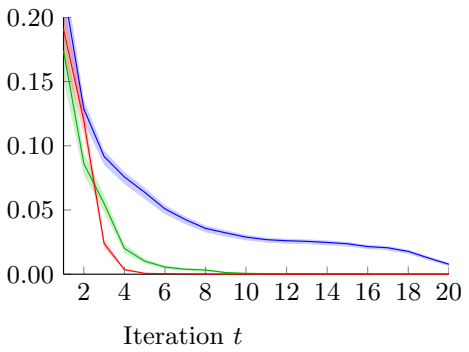
Parameter γ accounts for the information gain

- Linear kernel: $\gamma(T, K) = \mathcal{O}(d \log TK)$
- RBF kernel: $\gamma(T, K) = \mathcal{O}\big((\log TK)^{d+1}\big)$
- Matérn kernel: $\gamma(T, K) = \mathcal{O}\big((TK)^{\alpha} \log TK\big)$, $\alpha = \frac{d(d+1)}{2\nu + d(d+1)} \leq 1$

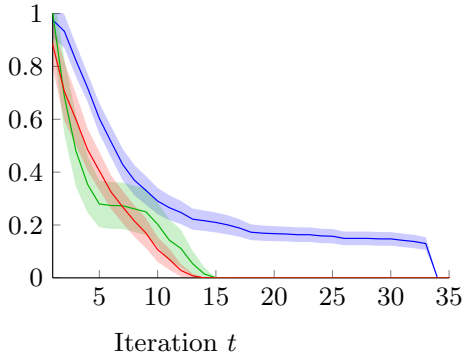
Results: mean instantaneous batch regret and confidence interval over 64 experiments



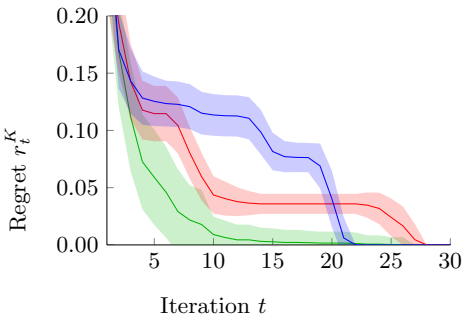
(a) Generated GP



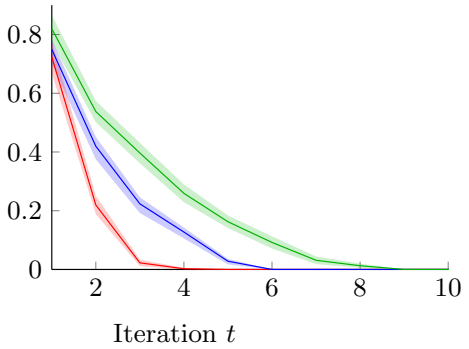
(b) Himmelblau



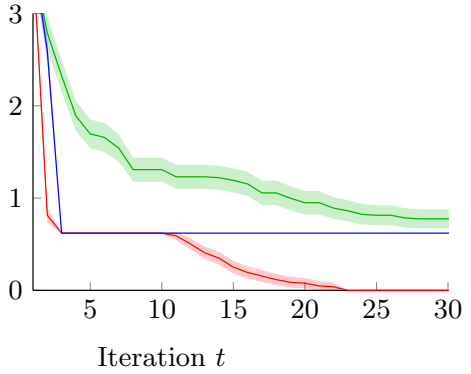
(c) Gaussian mixture



(d) Mackey-Glass

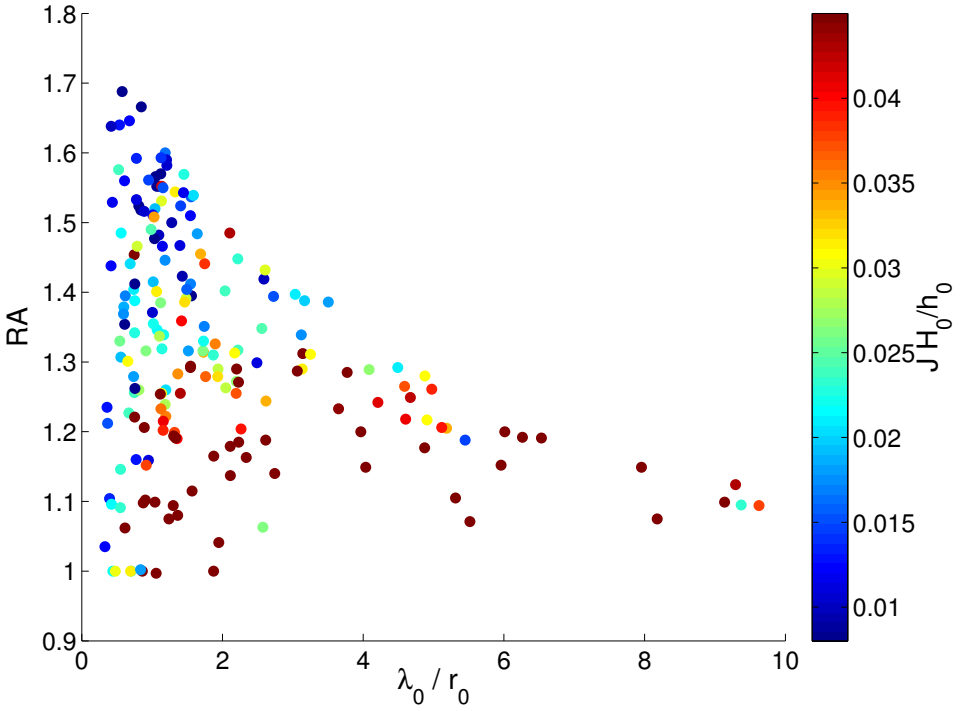


(e) Tsunamis



(f) Abalone

Proof of runup amplification and physical priors



Run-up amplification (RA) as a function of the wavelength to the island radius (at its base) ratio. The color code indicates the surf similarity (Iribarren number) computed with the beach slope and multiplied with the relative wave amplitude (wave amplitude to water depth ratio).

Example 2 - Viral marketing

Social data

Transactional data

Joint ongoing work with: Argyris Kalogeratos, Kevin Scaman
but also with: Rémi Lemonnier, Emile Richard*

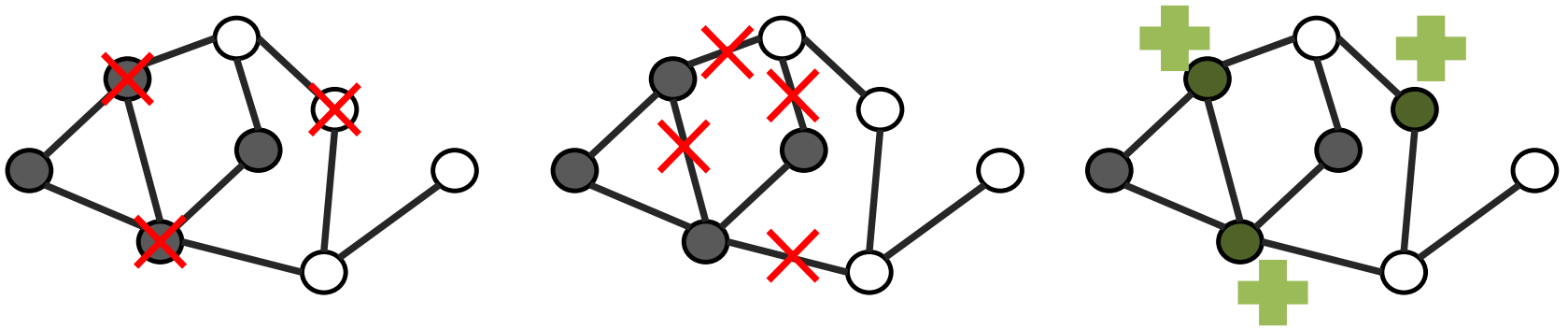
... or epidemic control

Source: The Hidden Geometry of Complex, Network-Driven Contagion Phenomena

Dirk Brockmann and Dirk Helbing

Science, 13 December 2013: 342 (6164), 1337-1342.

Acting on the network



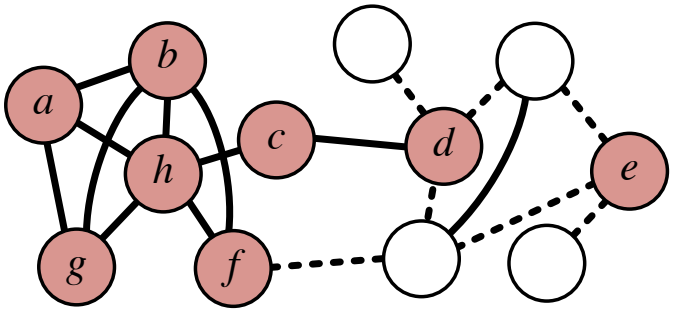
Many different ways to control a diffusion process

- Removal of a set of nodes (vaccination or quarantine)
- Removal of a set of edges (cancel flights)
- *Resource allocation* (antidotes) $\Leftarrow$

- The types of action available to authorities affect the design of optimal control strategies

Dynamic Treatment Allocation

Example on a toy network



- The red nodes are infected, the dashed edges are *infectious*
- Node h is the most central
- Node e and d are the most viral
- Node e is the safest

Dynamic Treatment Allocation

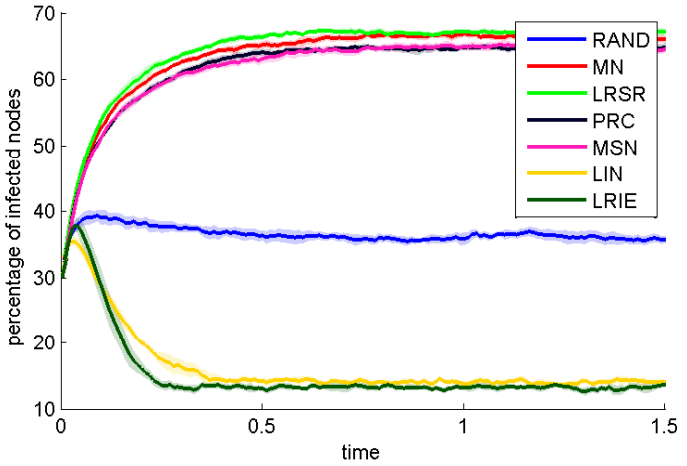
Strategy	**Score**
Random (RAND)	uniform in [0, 1]
Most Neighbors (MN)	degree
Page Rank Centrality (PRC)	PageRank score
Largest Reduction in Spectral Radius (LRSR)	$\lambda_1 - \lambda_1^{G \setminus i}$
Most Susceptible Neighbors (MSN)	$\sum_j A_{ij}(1 - X_j)$
Least Infected Neighbors (LIN)	$-\sum_j A_{ij} X_j$
Largest Reduction in Infectious Edges	$\sum_j A_{ij}(1 - 2X_j)$

The proposed LRIE - Scaman *et al.*, 2014

- Focuses on the most *viral* and *safe* nodes
- Targets nodes whose healing would minimize the number of *infectious edges*, i.e. edges between infected and susceptible nodes

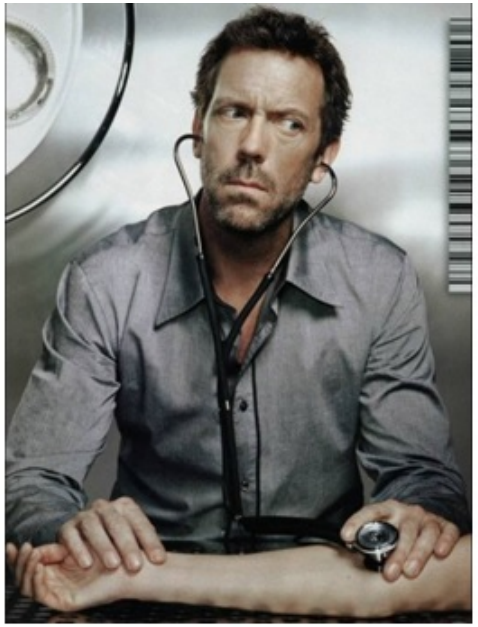
Experimental results - Scaman *et al.*, 2014

US air traffic network

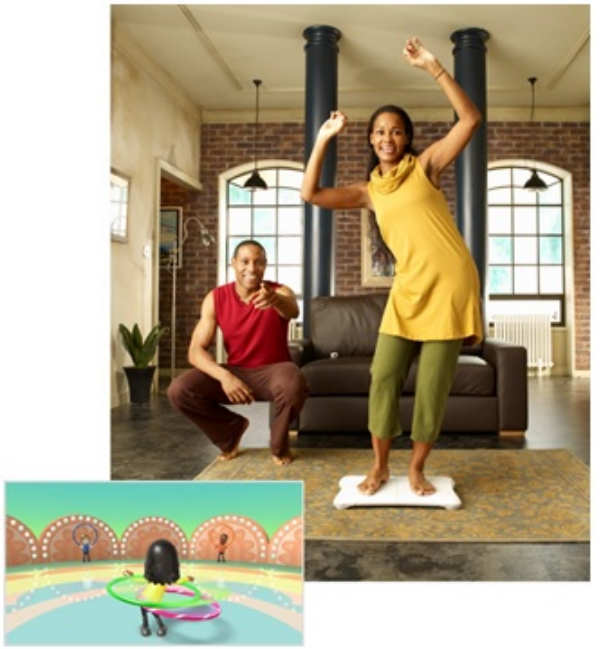


- N=1574 airports (nodes)
- effective spreading rate β/δ=2,
- treatment efficiency ρ/δ=600, b_{tot}=10 medicines.
- *Real US air traffic network* for the year 2010.
- *Large difference* between the competing strategies.
- *Persistence of the epidemic at low rates*, which is typical of scale-free networks.

Example 3 - Technology aided medicine



Conventional clinical examination



Cheap sensors for gait analysis

Joint ongoing work with: Julien Audiffren, Rémi Barrois-Müller, Emile Contal, Thomas Moreau, Laurent Oudre, Nikos Promponas, Damien Ricard, Charles Truong, Pierre-Paul Vidal,

Smartcheck - Scaling up the double loop

idfinnov
Accélérateur du Transfert de Technologies

COGNAC - G
COGNition and ACtion Group
UMR MD

Service de Santé des Armées

Acquisition

Transmission

Protocol

HMI

Network

Expertise

Processing

Some thoughts

Key elements

- Interdisciplinary research
- Replicable protocols and data standardization
- Reproducible signal processing and predictive models $\Leftarrow$
- Provide ergonomic HMIs
- Some funding for IT questions

What reproducibility of algorithms really means

- open source code
- peer-reviewed
- open to online experiments

IPOL
IMAGE PROCESSING ON LINE

Center for Data Science: Where to?

Climbing over or breaking the walls between scientific fields...

Indivividual initiatives vs. Collective (with tools)