Reproducible Research: Where to Begin With?

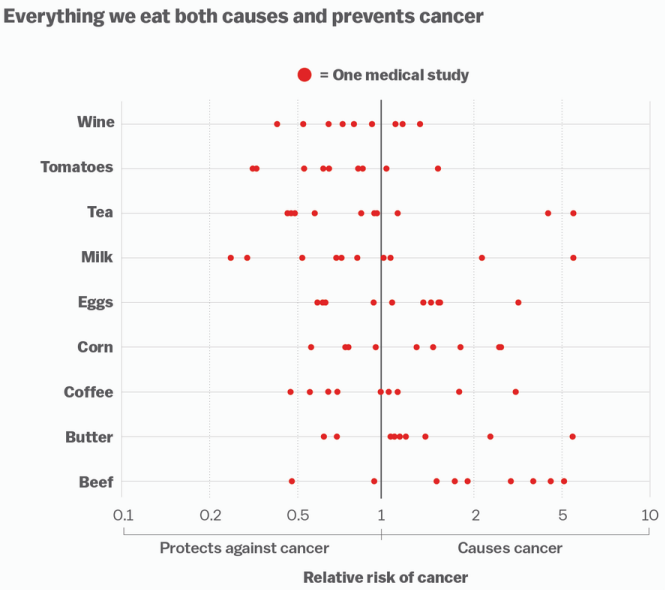
Arnaud Legrand
CNRS, Inria/POLARIS, University of Grenoble

June 21, 2016 – Journées Scientifiques Inria, Rennes

Inconsistencies

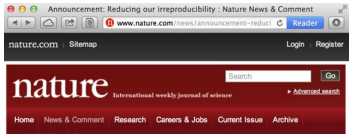
Is everything we eat associated with cancer? A systematic cookbook review, Schoenfeld and Ioannidis, *Amer. Jour. of Clinical Nutrition*, 2013.

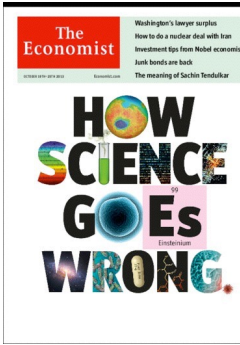
Inconsistencies



Public evidence for a Lack of Reproducibility

- J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005.
- *Lies, Damned Lies, and Medical Science*, The Atlantic. Nov, 2010

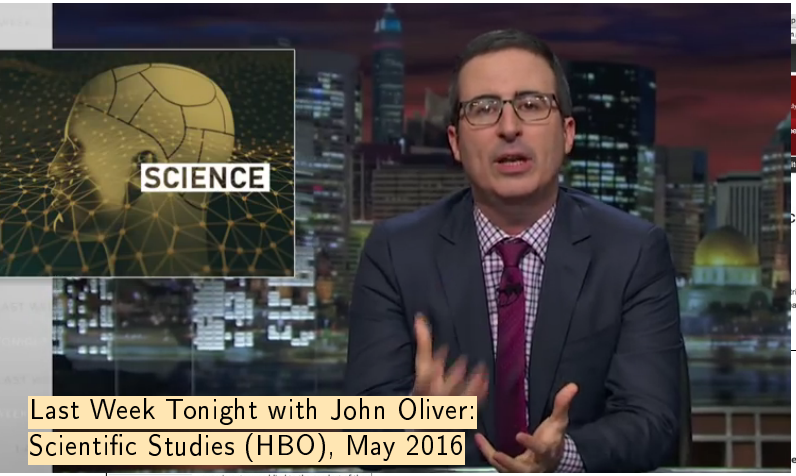




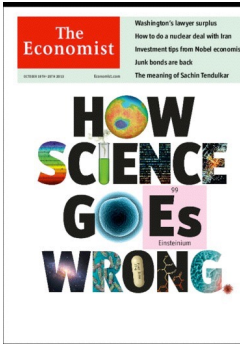
Courtesy V. Stodden, SC, 2015

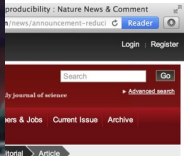
Public evidence for a Lack of Reproducibility

- J.P. Ioannidis. *Why Most Published Research Findings Are False* PLoS Med. 2005.
- *Lies, Damned Lies, and Medical Science*, The Atlantic. Nov, 2010



Last Week Tonight with John Oliver: Scientific Studies (HBO), May 2016





Courtesy V. Stodden, SC, 2015

Quick poll

1. Have you ever tried to reproduce some research results ?

Quick poll

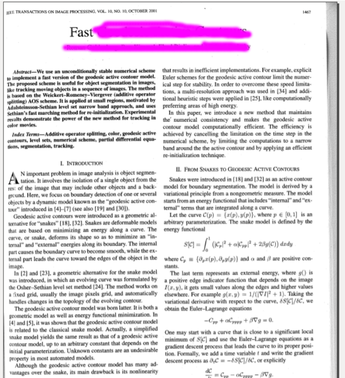
1. Have you ever tried to reproduce some research results ?
2. Have you ever failed ?

Quick poll

1. Have you ever tried to reproduce some research results ?
2. Have you ever failed ?

Article typique en traitement d'image

fichier PDF


IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 10, NO. 10, OCTOBER 2001 — 1467

Fast [redacted]

Abstract—**We use an unconditionally stable numerical scheme to implement a fast version of the geodesic active contour model. The proposed scheme is useful for object segmentation in images, like tracking moving objects in a sequence of images. The method is based on the Weickert–Romeney–Viergever (additive operator splitting) AOS scheme. It is applied at small regions, motivated by Adalsteinsson-Sethian level set narrow band approach, and uses Sethian's fast marching method for re-initialization. Experimental results demonstrate the power of the new method for tracking in color movies.**

Index Terms—**Additive operator splitting, color, geodesic active contours, level sets, numerical scheme, partial differential equations, segmentation, tracking.**

I. Introduction

An important problem in image analysis is object segmentation. It involves the isolation of a single object from the rest of the image that may include other objects and a background. Here, we focus on boundary detection of one or several objects by a dynamic model known as the "geodesic active contour" introduced in [4]–[7] (see also [19] and [30]).

Geodesic active contours were introduced as a geometric alternative for "snakes" [18], [32]. Snakes are deformable models that are based on minimizing an energy along a curve. The curve, or snake, deforms its shape so as to minimize an "internal" and "external" energies along its boundary. The internal part causes the boundary curve to become smooth, while the external part leads the curve toward the edges of the object in the image.

In [2] and [23], a geometric alternative for the snake model was introduced, in which an evolving curve was formulated by the Osher–Sethian level set method [24]. The method works on a fixed grid, usually the image pixels grid, and automatically handles changes in the topology of the evolving contour.

The geodesic active contour model was born latter. It is both a geometric model as well as energy functional minimization. In [4] and [5], it was shown that the geodesic active contour model is related to the classical snake model. Actually, a simplified snake model yields the same result as that of a geodesic active contour model, up to an arbitrary constant that depends on the initial parameterization. Unknown constants are an undesirable property in most automated models.

Although the geodesic active contour model has many advantages over the snake, its main drawback is its nonlinearity that results in inefficient implementations. For example, explicit Euler schemes for the geodesic active contour limit the numerical step for stability. In order to overcome these speed limitations, a multi-resolution approach was used in [34] and additional heuristic steps were applied in [25], like computationally preferring areas of high energy.

In this paper, we introduce a new method that maintains the numerical consistency and makes the geodesic active contour model computationally efficient. The efficiency is achieved by cancelling the limitation on the time step in the numerical scheme, by limiting the computations to a narrow band around the the active contour and by applying an efficient re-initialization technique.

II. From Snakes to Geodesic Active Contours

Snakes were introduced in [18] and [32] as an active contour model for boundary segmentation. The model is derived by a variational principle from a nongeometric measure. The model starts from an energy functional that includes "internal" and "external" terms that are integrated along a curve.

Let the curve $C(p) = \{x(p), y(p)\}$, where $p \in [0,1]$ is an arbitrary parameterization. The snake model is defined by the energy functional

$$S[C] = \int_0^1 \left(|C_p|^2 + \alpha |C_{pp}|^2 + 2\beta g(C)\right) dx dy$$

where $C_p \equiv \{\partial_p x(p), \partial_p y(p)\}$ and α and β are positive constants.

The last term represents an external energy, where $g()$ is a positive edge indicator function that depends on the image $I(x,y)$, it gets small values along the edges and higher values elsewhere. For example $g(x,y) = 1/(|\nabla I|^2 + 1)$. Taking the variational derivative with respect to the curve, $\delta S[C]/\delta C$, we obtain the Euler–Lagrange equations

$$-C_{pp} + \alpha C_{pppp} + \beta \nabla g = 0.$$

One may start with a curve that is close to a significant local minimum of $S[C]$ and use the Euler–Lagrange equations as a gradient descent process that leads the curve to its proper position. Formally, we add a time variable t and write the gradient descent process as $\partial_t C = -\delta S[C]/\delta C$, or explicitly

$$\frac{dC}{dt} = C_{pp} - \alpha C_{pppp} - \beta \nabla g.$$

ce qu'on peut faire avec :

- ✔ lire les formules
- ✔ croire les résultats
- ✘ vérifier les résultats
- ✘ reproduire les résultats
- ✘ voir les images en détail
- ✘ voir les graphes en détail

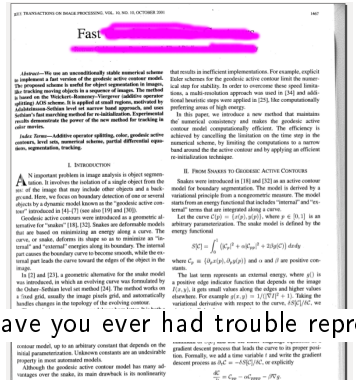
Courtesy of Enric Meinhardt-Llopis, CANUM 2016

Quick poll

1. Have you ever tried to reproduce some research results ?
2. Have you ever failed ?

Article typique en traitement d'image

fichier PDF



Fast[redacted]

Abstract—We use an unconditionally stable numerical scheme to implement a fast version of the geodesic active contour model. The proposed scheme is based on additive operator splitting, thus tracking multiple objects in a sequence of images. The method relies on the Weickert–Romeny–Viergever (additive operator splitting) AOS scheme. It is applied at small regions, motivated by Adalsteinsson–Sethian level set narrow band approach, and uses Sethian's fast marching method for re-initialization. Experimental results demonstrate the power of the new method for tracking in color movies.

Index Terms—Additive operator splitting, color, geodesic active contours, level sets, numerical scheme, partial differential equations, segmentation, tracking.

I. Introduction

An important problem in image analysis is object segmentation. It involves the isolation of a single object from the rest of the image that may include other objects and a background. Here, we focus on boundary detection of one or several objects by a dynamic model known as the "geodesic active contour" introduced in [4]–[7] (see also [39] and [30]).

Geodesic active contours were introduced as a geometric alternative for "snakes" [18], [32]. Snakes are deformable models that are based on minimizing an energy along a curve. The curve, or snake, deforms its shape so as to minimize an "internal" and "external" energies along its boundary. The internal part causes the boundary curve to become smooth, while the external part leads the curve toward the edges of the object in the image.

In [2] and [23], a geometric alternative for the snake model was introduced, in which an evolving curve was formulated by the Osher–Sethian level set method [24]. The method works on a fixed grid, usually the image pixels grid, and automatically handles changes in the topology of the evolving contour.

II. From Snakes to Geodesic Active Contours

Snakes were introduced in [18] and [32] as an active contour model for boundary segmentation. The model is derived by a variational principle from a nongeometric measure. The model starts from an energy functional that includes "internal" and "external" terms that are integrated along a curve.

Let the curve $C(p) = (x(p), y(p))$, where $p \in [0, 1]$ is an arbitrary scale-normalization. The snake model is defined by the energy functional

$$S[C] = \int_0^1 \left(|C_p|^2 + \alpha |C_{pp}|^2 + \beta |\eta(C)| \right) dp \, dq$$

where $C_p \equiv (\partial_p x(p), \partial_p y(p))$ and α and β are positive constants.

The last term represents an external energy, where $\eta(\cdot)$ is a positive edge indicator function that depends on the image $I(x, y)$, it gets small values along the edges and higher values elsewhere. For example $\eta(x, y) = 1/(|\nabla I|^2 + 1)$. Taking the variational derivative with respect to the curve, $\delta S[C]/\delta C$, we

contour model, up to an arbitrary constant that depends on the initial parametrization. Unknown constants are an undesirable property in most automated models.

Although the geodesic active contour model has many advantages over the snake, its main drawback is its nonlinearity that results in inefficient implementations. For example, explicit Euler schemes for the geodesic active contour limit the numerical step for stability. In order to overcome these speed limitations, a multi-resolution approach was used in [34] and additional heuristic steps were applied in [25], like computationally preforming areas of high energy.

In this paper, we introduce a new method that maintains the numerical consistency and makes the geodesic active contour model computationally efficient. The efficiency is achieved by cancelling the limitation on the time step in the numerical scheme, by limiting the computations to a narrow band around the the active contour and by applying an efficient re-initialization technique.

gradient descent process that leads the curve to its proper position. Formally, we add a time variable t and write the gradient descent process as $\partial_t C = -\delta S[C]/\delta C$, or explicitly

$$\frac{\partial C}{\partial t} = C_{pp} - \alpha C_{pppp} - \beta \nabla \eta$$

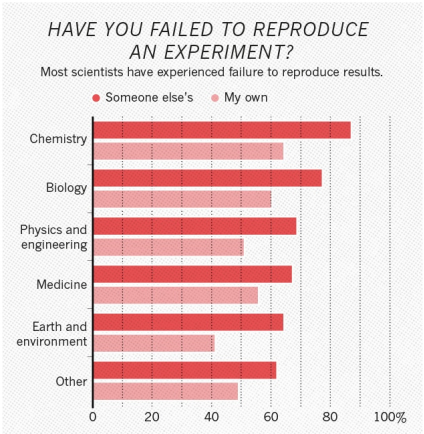
- ✔ lire les formules
- ✔ croire les résultats
- ✘ vérifier les résultats
- ✘ reproduire les résultats
- ✘ voir les images en détail
- ✘ voir les graphes en détail

Courtesy of Enric Meinhardt-Llopis, CANUM 2016

3. Have you ever had trouble reproducing the work of one of your student?

Why are scientific studies so difficult to reproduce?

1,500 scientists lift the lid on reproducibility, Nature, May 2016

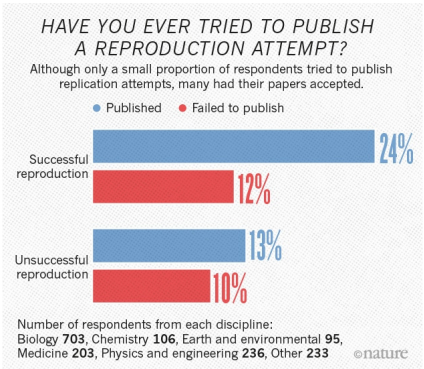


1934: Karl Popper introduces the notion of falsifiability and crucial experiment and puts reproducing the work of others at the core of science

> *Reproducibility of experimental results is the hallmark of science* *[Drummond, 2009]*

Why are scientific studies so difficult to reproduce?

1,500 scientists lift the lid on reproducibility, Nature, May 2016

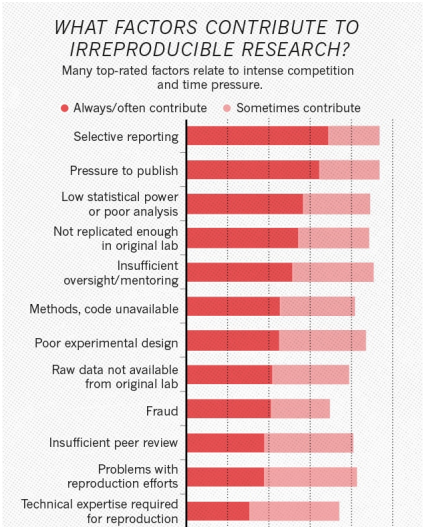


1934: Karl Popper introduces the notion of falsifiability and crucial experiment and puts reproducing the work of others at the core of science

> *Reproducibility of experimental results is the hallmark of science* *[Drummond, 2009]*

Why are scientific studies so difficult to reproduce?

1,500 scientists lift the lid on reproducibility, Nature, May 2016



1934: Karl Popper introduces the notion of falsifiability and crucial experiment and puts reproducing the work of others at the core of science

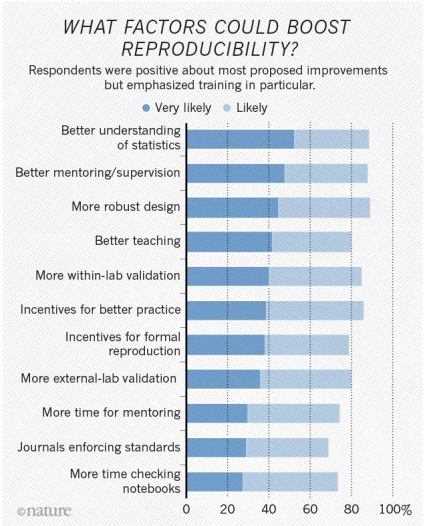
> *Reproducibility of experimental results is the hallmark of science* *[Drummond, 2009]*

Key factors

- publication pressure, mentoring, ...
- selective reporting, poor analysis
- code/raw data unavailable

Why are scientific studies so difficult to reproduce?

1,500 scientists lift the lid on reproducibility, Nature, May 2016



1934: Karl Popper introduces the notion of falsifiability and crucial experiment and puts reproducing the work of others at the core of science

> *Reproducibility of experimental results is the hallmark of science* *[Drummond, 2009]*

Key factors

- publication pressure, mentoring, ...
- selective reporting, poor analysis
- code/raw data unavailable

What can be done?

- better teaching/understanding of stats, better designs
- incentives for better practices

All this is about Natural Sciences. Should we care ?

Yes. Computer Science is young and inherits from Mathematics, Engineering, Nat. Sciences, ...

Model $\neq$ **Reality**.

All this is about Natural Sciences. Should we care ?

Yes. Computer Science is young and inherits from Mathematics, Engineering, Nat. Sciences, ...

Model $\neq$ **Reality**. Although designed and built by human beings, computer systems are so complex that mistakes easily slip in...

- Experiments: Mytkowicz, Diwan, Hauswirth, Sweeney. Producing wrong data without doing anything obviously wrong!. SIGPLAN Not. 44(3), March 2009

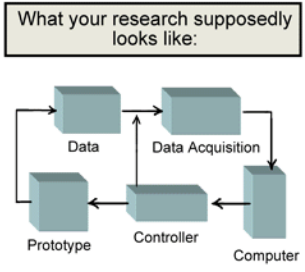


Figure 1. Experimental Diagram

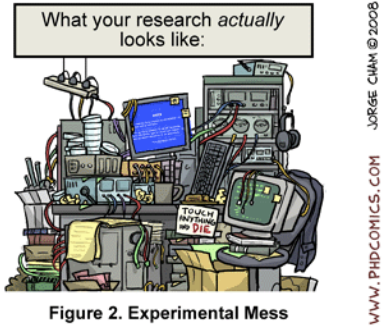


Figure 2. Experimental Mess

All this is about Natural Sciences. Should we care ?

Yes. Computer Science is young and inherits from Mathematics, Engineering, Nat. Sciences, ...

Model $\neq$ **Reality**. Although designed and built by human beings, computer systems are so complex that mistakes easily slip in...

- Experiments: Mytkowicz, Diwan, Hauswirth, Sweeney. Producing wrong data without doing anything obviously wrong!. SIGPLAN Not. 44(3), March 2009

- Statistics: Trouble at the lab, The Economist 2013

  According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting". *Sandy Pentland, MIT*

All this is about Natural Sciences. Should we care ?

Yes. Computer Science is young and inherits from Mathematics, Engineering, Nat. Sciences, ...

Model $\neq$ **Reality**. Although designed and built by human beings, computer systems are so complex that mistakes easily slip in...

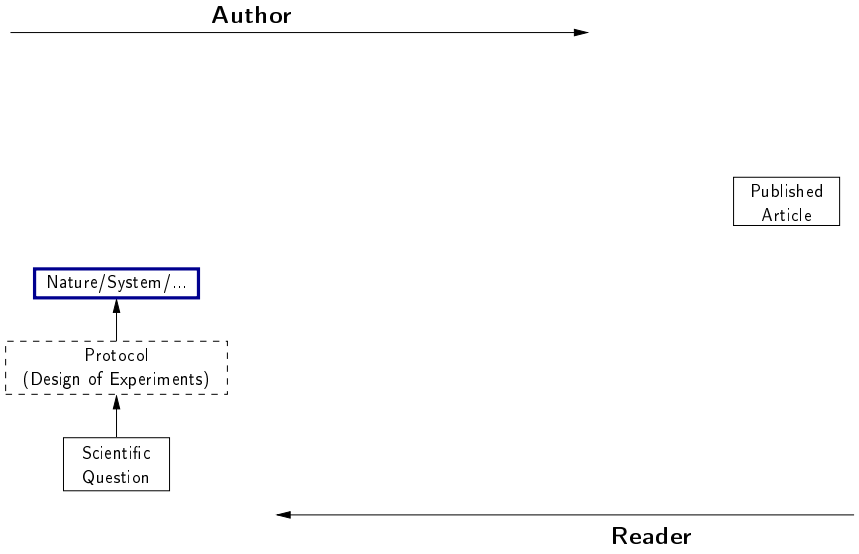
- Experiments: Mytkowicz, Diwan, Hauswirth, Sweeney. Producing wrong data without doing anything obviously wrong!. SIGPLAN Not. 44(3), March 2009

- Statistics: Trouble at the lab, The Economist 2013

  According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting". *Sandy Pentland, MIT*

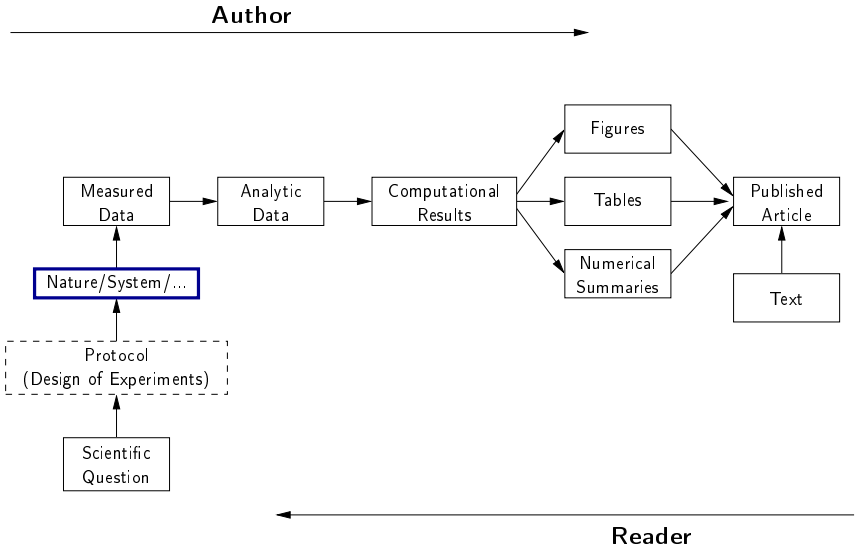
- Numerical reproducibility: change compiler, OS, machine and see what happens. Ever tried to exploit a parallel architecture ? ☺

Reproducible Research: Trying to Bridge the Gap



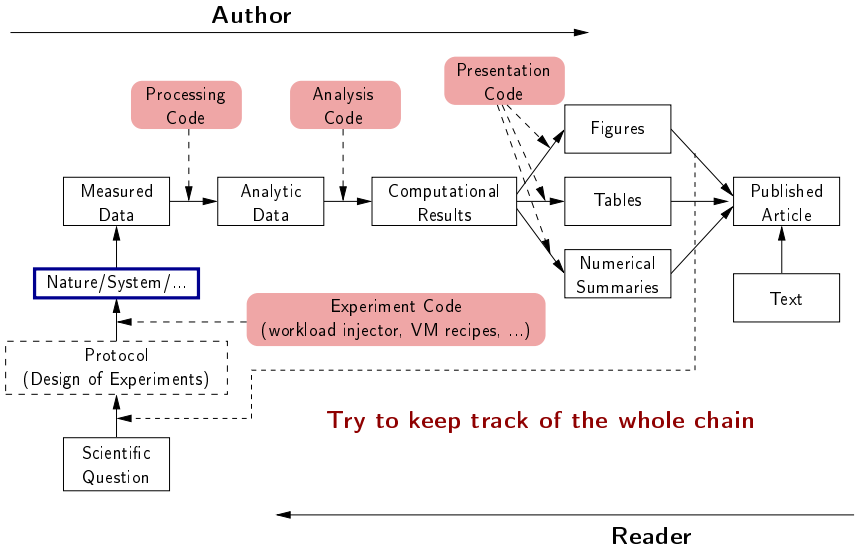
Inspired by Roger D. Peng's lecture on reproducible research, May 2014

Reproducible Research: Trying to Bridge the Gap



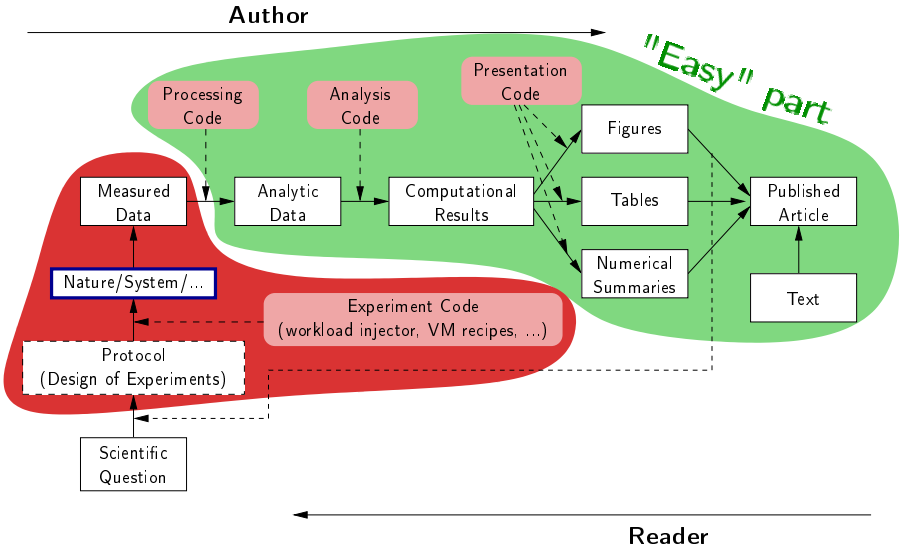
Inspired by Roger D. Peng's lecture on reproducible research, May 2014

Reproducible Research: Trying to Bridge the Gap



Inspired by Roger D. Peng's lecture on reproducible research, May 2014

Reproducible Research: Trying to Bridge the Gap

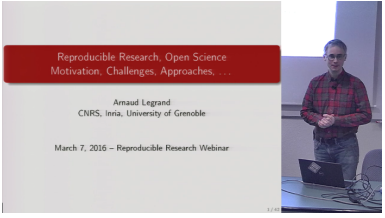


Inspired by Roger D. Peng's lecture on reproducible research, May 2014

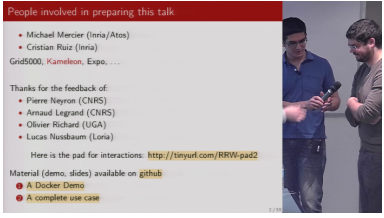
Webinars: Learning by Doing

Many different tools/approaches developed in various communities

1. Replicable article
2. Logging your activity
3. Logging and backing up your data
4. Organizing your data
5. Mastering your environment
6. Controlling your experiments
7. Making your data/code/article available



1. Introduction, Litterate programming



2. Controling your environment

3. Numerical reproducibility

4. Logging and backing up your work

What Next ?

Article typique en traitement d'image

fichier PDF

ce qu'on peut faire avec :

- ✔ lire les formules
- ✔ croire les résultats
- ✘ vérifier les résultats
- ✘ reproduire les résultats
- ✘ voir les images en détail
- ✘ voir les graphes en détail

7 / 21

It's up to us. **We** should **care** and **take the lead**

Learning is the essence of our work

$\rightsquigarrow$ Train our researchers and students

- Slight cultural changes in our relation to publication and daily practice
- Higher confidence in our (students) work $\rightsquigarrow$ definite competitive advantage
- Our research will become sound, deeper, auditable, more visible, reusable, ...

Next webinars: in October 2016

https://github.com/alegrand/RR_webinars