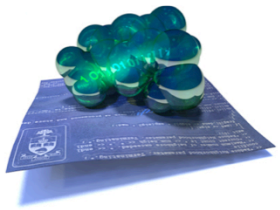
A Bioinformatics Course

Cargo Cult



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Modeling is the fundamental task of bioinformatics.

"Computational biology"

Looking beyond data management, bioinformatics is a way to study biology. This aspect – which I like to refer to as "Computational Biology" – has a lot more to do with modeling, and with the question of '''understanding''' biology, than managing large amounts of data.

Understanding biology means being able to abstract from apparent complexity in order to interpret observations in the framework of simple, fundamental principles. Such understanding should allow us to make precise, confident predictions.

This involves abstraction, and working with abstractions means we are working with models.

Obviously, models can be right or wrong ...

Modeling an observation allows us to make predictions, but if the model is appropriately constructed, the model can also allow us to test and evaluate ideas about the mechanism behind the observation. Note that striving for explanatory and/or mechanistic models is not what "model" generally means in statistics. There, all that is required is a descriptively adequate model. Thus models can be predicitve, but not explanatory, they can be explanatory, while failing to make accurate predictions, and all of this in different shades and combinations. And they can fail to make correct predictions, which may allow us to learn something about the observations too. Or, to paraphrase: models can be wholly wrong, the can be right for the wrong reasons, and they can be right.

Obviously, models can be right or wrong ...

... but you must always ensure they are relevant!

Nothing can be learned from models that are not relevant to the questions we are asking. And it is often not trivial to figure out that a model is in fact irrelevent. It may look perfectly fine – until you take a closer look. The question we must always keep in mind is: how do we preserve the connection of our models to the biology they describe.

Otherwise, we are pursuing *cargo cult* bioinformatics.

Cargo Cult Science

> [Cargo cult] derives from the belief which began among Melanesians in the late 19th and early 20th century that various ritualistic acts such as the building of an airplane runway will result in the appearance of material wealth, particularly highly desirable Western goods (i.e., "cargo"), via Western airplanes. (Wikipedia:Cargo Cult)

We use *Cargo cult* as a metaphor for an activity that outwardly appears useful, often impressively so, but has no prospect of contributing to progress due to a fatal flaw in its assumptions.

Even the most elaborate bamboo airplane on an isalnd jungle landing strip represents

Cargo Cult Science

[...] In the South Seas there is a cargo cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas--he's the controller--and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call [some examples of pseudoscience] cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

Now it behooves me, of course, to tell you what they're missing. But it would be just about as difficult to explain to the South Sea Islanders how they have to arrange things so that they get some wealth in their system. It is not something simple like telling them how to improve the shapes of the earphones. [...].

Richard Feynman

The term "cargo cult science" was coined by Richard Feynman in a Caltech commencement address, where he applied it to examples of pseudoscience.

We can apply it to Bioinformatics–or really, any scientifc discipline–when we ask: how does the outcome of our inquiry advance our goals of scientific insight.

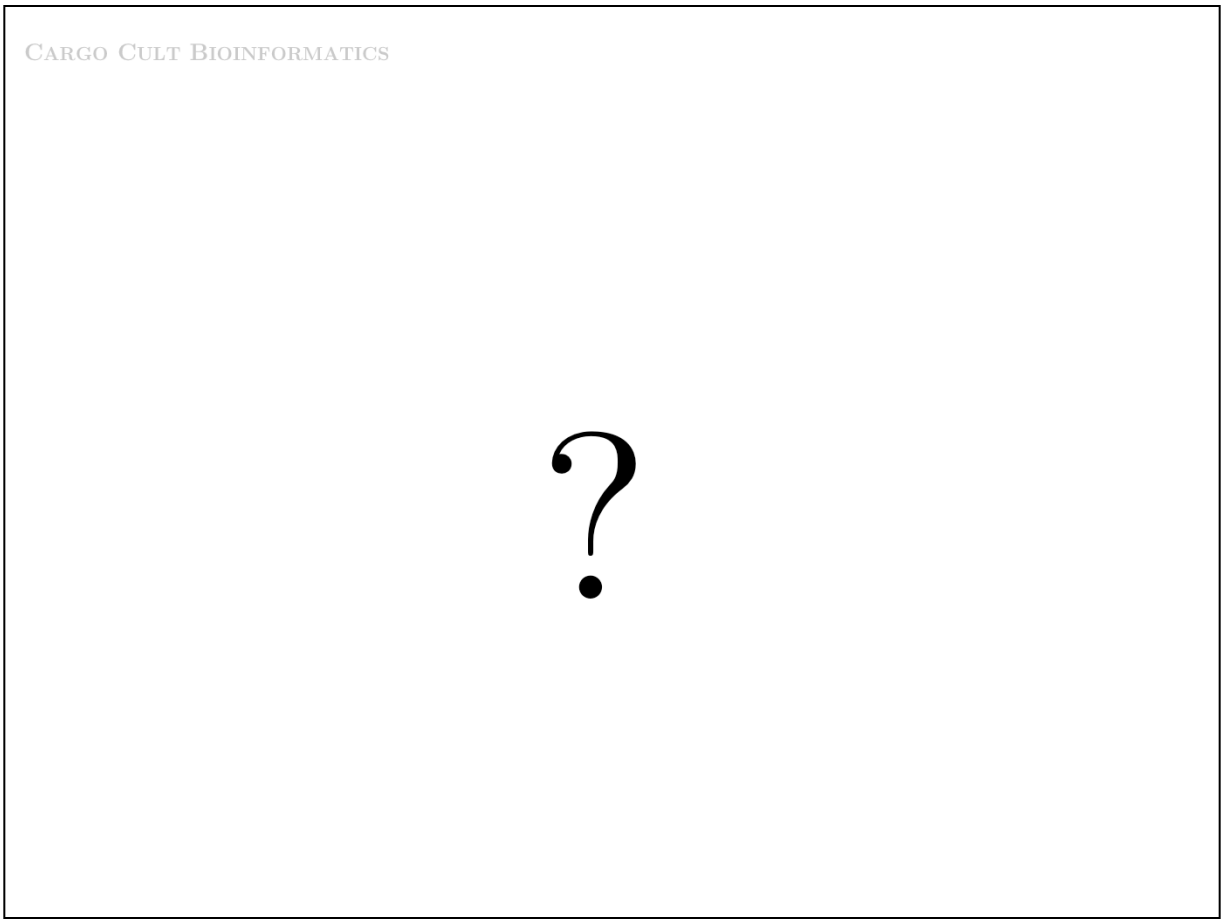
Avro Lancaster, WikiMedia Commons

In one example, Abraham Wald, a statistician, provided a crucial insight about armoring Lancaster bombers, which is considered seminal to *operational research.*

> [In the second World War ...] Researchers from the Center for Naval Analyses had conducted a study of the damage done to aircraft that had returned from missions, and had recommended that armor be added to the areas that showed the most damage. Wald noted that the study only considered the aircraft that had survived their missions—the bombers that had been shot down were not present for the damage assessment. The holes in the returning aircraft, then, represented areas where a bomber could take damage and still return home safely. Wald proposed that the Navy instead reinforce the areas where the returning aircraft were unscathed, since those were the areas that, if hit, would cause the plane to be lost.
>
> (Wikipedia:Abraham Wald)

The Cargo Cult approach here was exemplified by the original study. It had aspects of engineering (evaluating damage), it had aspects of statistics (establishing components with a high risk of being damaged), it asked questions of vital importance. It looked really good at the outset. But the study got it all wrong by failing to understand the implications of the statistics in context.

Good science is all about asking the right question.



Where do we find "cargo cult" in bioinformatics?

Anything that ...

... sequence alignment algorithm

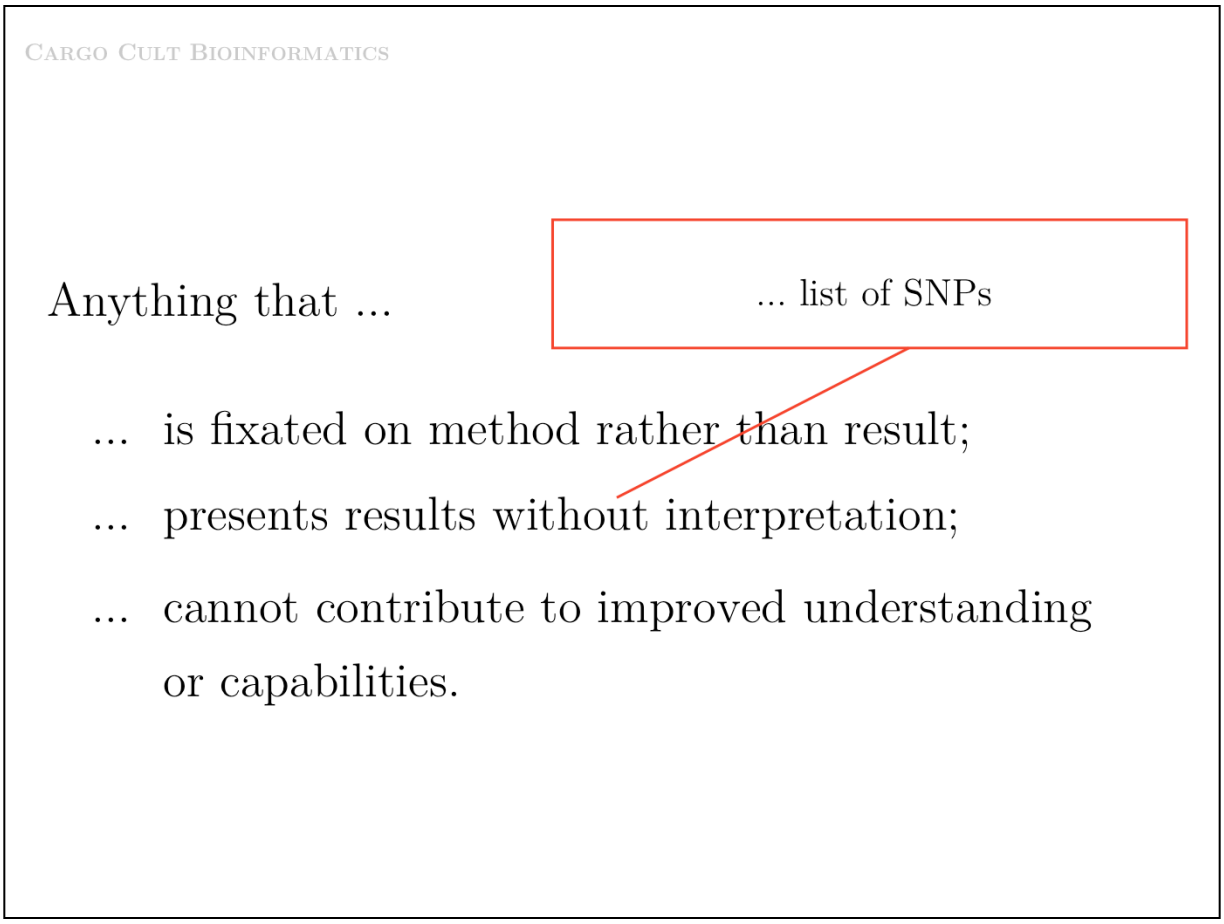
... is fixated on method rather than result;

... presents results without interpretation;

... cannot contribute to improved understanding or capabilities.

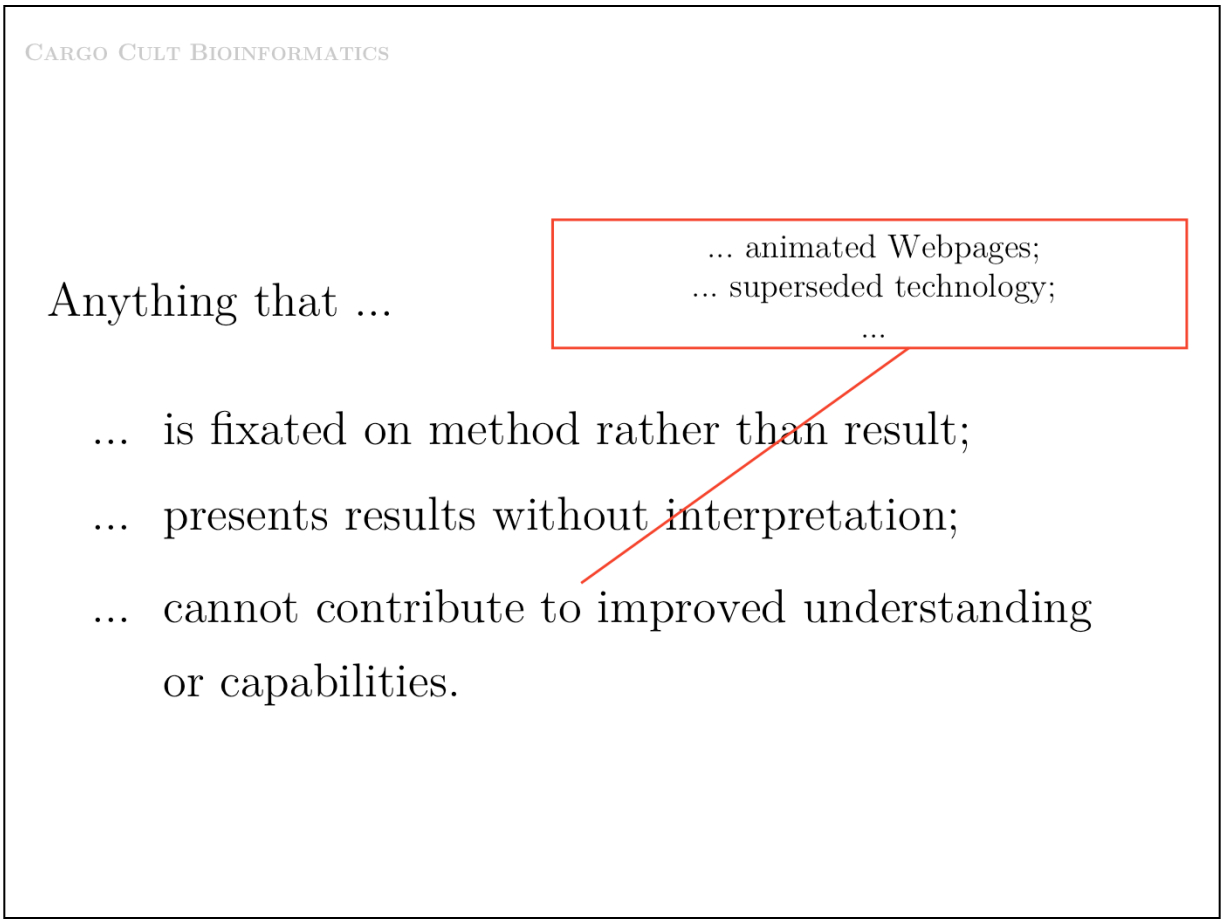
One example is the fixation on teaching algorithms.

Yes, you need to understand how your tools work, in order to apply them correctly. This does not mean however that you need to be able to build the tools yourself, even though such knowledge can fill many lectures and serve as the basis for many an exam.



Mountains of data make good press. Funders may be impressed with the sheer speed with which – it is claimed – knowledge grows. But data is only the **beginning**. Without interpretation, the workflow is incomplete. **Data does not explain itself.** And if the path to proper interpretation is not clear, there are "no airplanes in the system."

Hypothesis without data is sterile. Data without hypothesis is useless. Both need to be balanced.



Other cardinal sins include substituting flashy form for well defined semantics, or working with technology that has been proven inferior to alternatives (CLUSTAL, a multiple sequence alignment algorithm comes to mind).

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA