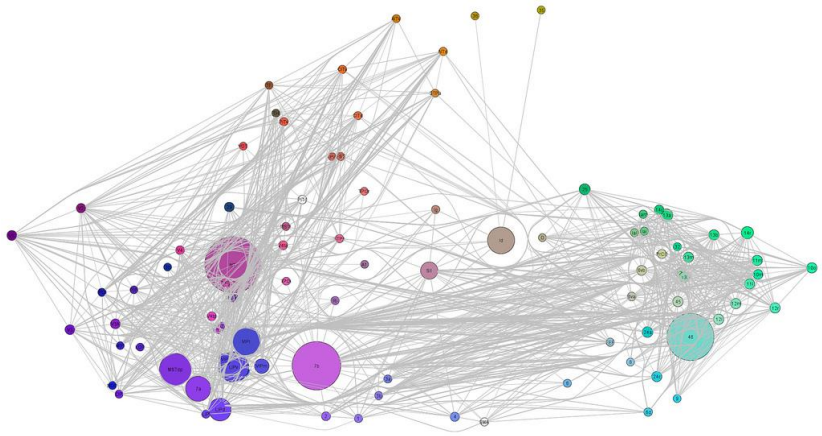
What questions should we ask?

Anders Sandberg, Future of Humanity Institute, Oxford Martin School, Oxford University

Talk given at the second 2014 Oxford symposium on Big Data Science in Medicine, December 8 2014



Introduction

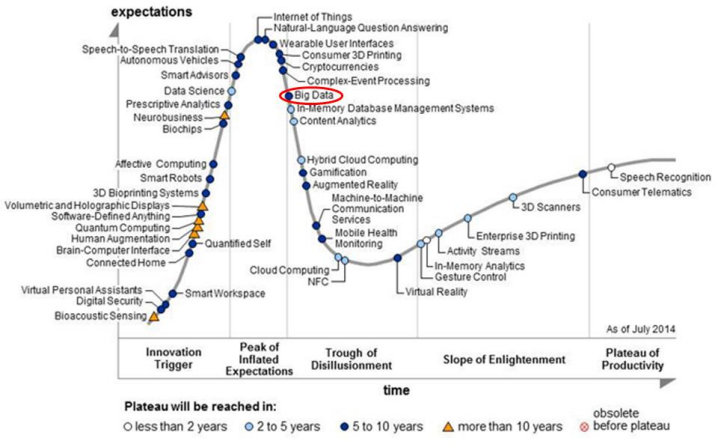
> "The Answer to the Great Question... Of Life, the Universe and Everything... Is... Forty-two," said Deep Thought, with infinite majesty and calm.
>
> "Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"
>
> "I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."
>
> — Douglas Adams, The Hitchhiker's Guide to the Galaxy

Do you understand what the question is, and what type of answer would satisfy you? We are often confused about what we want to know. Then we get annoyed when we are told something that is not worth knowing, something we do not want to know – or told that we didn't know what to ask for.

This talk is going to be about asking the right questions. It is more abstract and philosophical than the other talks at this meeting. This might make it less readily applicable to your problem or dataset, but has the virtue of making it applicable to a broad range of problems. As I hope to show, starting out with a 10,000 feet view is a good start for actually answering the questions than matter.

The Big Data hype curve

The 2014 Gartner hype cycle curve is revealing:



Source: Gartner (August 2014)

This curve was presented in August. By now Big Data has slid further down the slope. But I suspect we here are the kind of people who realize that ten years down the line Big Data will still be around because it is actually useful and powerful. Large datasets and statistical processing do have amazing power. In ten years' time we will be even better at using it. We might not call it Big Data – just like artificial intelligence we might switch name to avoid those embarrassing early 10's claims.

Right now many critics are criticising Big Data. This is not a coincidence. A few years back there was little to discuss, then the technology started to take off and people got excited and hyped it endlessly: it was new, people had research projects and start-ups to sell. Critics take a bit longer to wake up, but once they are going they have plenty of easily criticised hype fodder.

A critique well worth reading is Gary Marcus and Ernest Davies "Eight (no, nine!) problems with Big Data" (*The New York Times*, 6 April 2014). As you read it, you will shake your head at how stupid everybody else is. That is fine, as long as you make sure you do not repeat the mistakes yourself now when they have been made explicit. Some of their points are:

- Finding correlations is easy, finding *meaningful* correlations is another thing. Bogus correlations are trivial to get.
- Learning what you want often requires understanding of domain, not just data crunching.
- Avoid gameable data (is somebody motivated to try to push results in some direction?) and lack of robustness (is the answer going to *remain* true?)
- Is the data really data and not just feedback from yourself?
- Scientific precision to imprecise questions make answers look like they are meaningful.
- Big Data is great at analysing common things, but gets into trouble when dealing with rare stuff – which often matters (bigger data is not always better).

This essay will deal a bit with why these things are problems, and what kind of questions get around them.

What are questions?

As a philosopher I cannot resist attacking the question "what *are* questions?"

It turns out that there is a fair bit of philosophy about questions, most of it rather useless for our present purposes. I think a good-enough answer is that questions are requests for information. We can distinguish between three main kinds:

1. Descriptive questions: what is X? (How does it exist? What properties does it have?)
2. Relational questions: how is X related to Y?
3. Causal questions: how does X affect Y?

Of course, there are many other kinds of questions – rhetorical questions, requests to do something, philosophical conundrums – but we are interested in information.

In particular, a question is *about* something and has a *kind* of desired answer. "42" is not the desired kind of answer to the meaning of life. It is also fairly unclear what the question really is about: one can interpret it in many ways.

Serendipity

Sometimes this is a great start. We are fishing for serendipity: we don't know what the question or the answer should be, and just play around observing what happens.

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'
> — Isaac Asimov

Unfortunately, fishing for serendipity is *mostly* annoying and a waste of resources. If you mix stuff in the lab randomly you are far more likely to run out of chemicals rather than get a Nobel prize. Or even get a decent explosion. But when serendipity happens it is often awesome enough that we forget that it is very rare it succeeds. We have all heard about Fleming's penicillin mould-infested petri dish, but how many of the millions of other failed petri dishes can we name?

One can set things up so serendipity is more likely, one should be open to play and creativity, but it is not a reliable way of achieving something.

Useless questions and answers

> To exhibit the perfect uselessness of knowing the answer to the wrong question.
> — Ursula K. Le Guin

Another situation is when we don't know what the question should be, or we don't know what kind of answer is right.

This kind of mistake seems to be fairly common in the use of Big Data. It is at the root of false precision: an imprecise question is formulated (Marcus gives "who are the most influential people?" as an example), operationalized in some way, and a clear answer is produced – except that the operationalization is rather loose and introduces a whole host of assumptions, which are then hidden by the apparently clear answer. The actual question asked became different from what *we* asked, and

the answer type different from what we think it is. Just because the question sounds clear and the answer is clear doesn't mean it is meaningful.

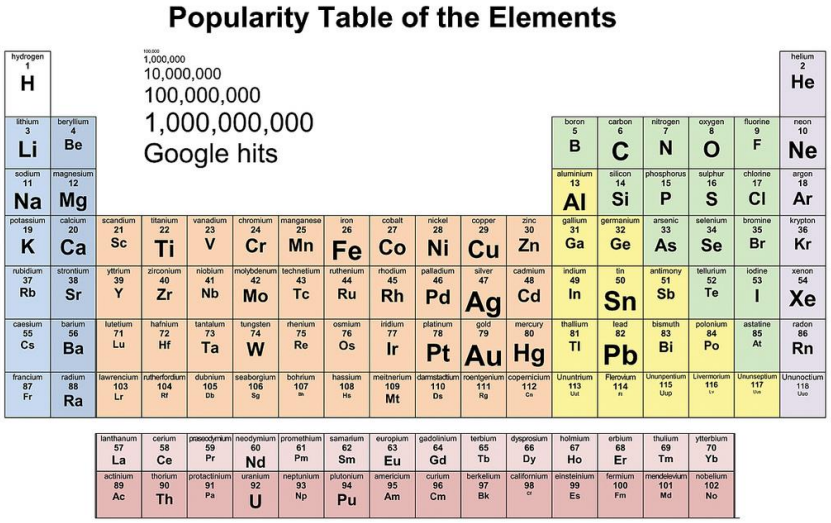
Answers can also be true but useless. The question "Who is president in the US" has the true answer "Somebody taller than 3 inches." It is true but doesn't convey the information we are hoping for. Much of the usefulness lies in being precise (the answer is relevant) and complete (the answer doesn't leave out too much important information). Or, put differently: how much noise are you willing to get into your answer, and how many positive cases are you willing to overlook? This depends on what you aim to do and the bigger context.

Big Data tempts us by allowing us to ask questions about corpuses that cover any topic and any kind of input variable, not just the obviously relevant ones – the dream of serendipity again! – which of course means that we open ourselves for potentially worse answers since we get more noise. We can defend ourselves statistically (Bonferroni correction etc.) but we can also defend ourselves by being careful with relevance: what should be included in our exploration, and what shouldn't? In fact, asking questions about what is *relevant* is often a valuable start for an investigation – especially if we can use Big Data to do it effectively. Discovering robust links or relevance matters!

Good answers are often maximally compressed: they contain the information we want in a succinct form, and can still be unfolded into explanations of the data or allow mental (or practical) simulation of possible scenarios. Bad answers are often too complex or don't lead to anything else.

An overly detailed ingredient list for a loaf of bread made with enriched flours.

What questions are valuable?



But even if the question and kind of answer are clear, should we even be asking it? We can potentially ask infinitely many questions, but only have time and resources for actually asking a few. It becomes a matter of priority.

Maximal information or value?

One obvious aim would be to look for the most informative questions: they reduce uncertainty in a domain maximally. Maximize information gain and you will narrow down the possibilities optimally. A large literature of machine learning and data science can help us here.

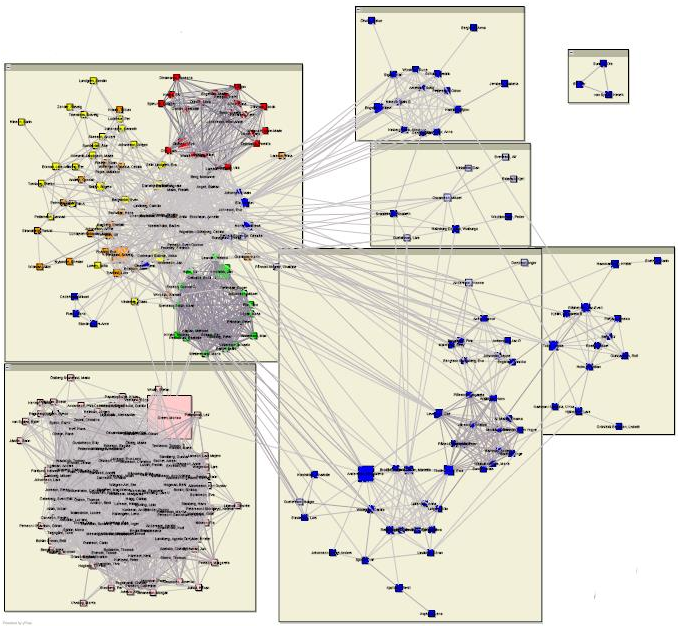
But something else may be more significant: maximize the expected value difference. Knowing some answers is worth far more than knowing others. Learning that the office is on fire trumps most other building usage data, despite being unlikely. So by this heuristic we should try to ask the most important questions first.

A case in point is the lack of balance between diagnosis and therapy. Over the past years we have become far better at diagnosing medical conditions – but for many of them there is precious little we can do about them. That suggests that in many medical specialities the value of answering questions about therapy is increasing compared to the value of diagnosis questions.

Data doesn't speak for itself

These heuristics all requires a prior, and can be quite hard if you do not know the domain well. What problems are hard? What answers are likely to be noisy, biased or statistically insignificant? Worse, what answers will actually tell you anything?

This is one of the big myths of Big Data, that the data will speak for itself, that theory is unnecessary. It assumes that as soon as you process it you will see a pattern that will tell you what you should be doing with it. But how do you process it? You could cluster or categorize it, but doing that will not necessarily show you patterns you can understand. Worse, the right settings for your algorithm often depend on the domain. And of course, theories are causal: without them you will be assuming a correlation-causation link.



A few years back I experimented in mining political data. It was not hard to cluster to find structure: it is plentiful when you look at the social network formed by co-authored bills or parliamentary responses. But what did it tell me? Beyond the obvious patterns (members of the same party tend to cluster together, some parties have different internal clustering hierarchies from each other) the patterns were still mere data to me. I needed to find a political scientist to explain what was actually interesting and new, what was obvious to anybody who cared about politics, and what was merely an artefact. No amount of pure data mining could tell me that. In fact, the hard question turned out to be how to find a political scientist to work with, not processing the data.

Having a theory might bias you, but you have something to search with and make a preliminary ranking of your questions. As Karl Popper pointed out, testing your theory is important: we can never prove that a theory is right, but we can sometimes disprove it. Falsification is useful: disproving large parts of hypothesis space is a good thing since it lets you focus on better parts.

Time and a little theory of problems

Nick Bostrom wrote "A little theory of problems":

> There are many problems in the world. Not all of them ought to be solved.
>
> Important problems are those for which the value of a solution is either large and positive or large and negative.
>
> Not all important problems ought to be solved.
>
> We can distinguish positive-value problems (some of which are high-value, others low-value) from negative-value problems.
>
> Not all important positive-value problems ought to be addressed.
>
> Elastic problems are those whose solution can be found significantly sooner with one extra unit of effort.
>
> We ought to address high-value high-elasticity problems.
>
> *"Discoveries" are acts that move the arrival of some information from a later point in time to an earlier point in time.*
>
> *The value of a discovery does not equal the value of the solution discovered. The value of a discovery equals the value of having the solution moved from the later time at it would otherwise have arrived to the time of the discovery*.
> — Nick Bostrom

(Italics mine). In the long run we will know everything. The answers that matter are the ones where it matters that we know them *now*.

A kind of question that is good to answer early by this account is in which direction to go. If you are moving in the wrong direction you will be happier if you find out about it early rather than late. And if you cannot tell if you are going in the right direction, the first questions should be how you can find out, and if not, whether you should be doing what you are doing at all.

This is why "fail fast" strategies are so useful: knowing that we cannot succeed with something is better learned early than after a lot of resources have been wasted. Falsification of theories means we can safely ignore them henceforth: whatever the goal is, it was at least not in that direction.

I suspect that in many domains we are both overconfident in our ability to plan ahead and too rigid in our planning, making us pivot too little and too late when our past guesses turn out to be wrong

Sometimes waiting with a question makes sense: if computing advances exponentially, a certain computation can be finished earlier if you start later and buy the far better computer there will be in the future. This of works best if you can accurately estimate how fast it will run your problem. Some questions may require information or technology that does not exist yet. But if you figure out what it is, you can hide and wait for it to show up, pouncing at the earliest opportunity.

Metaplanning

How meta should we be when asking questions? Should we ask questions about asking questions about asking questions, or is this merely philosophical entertainment?

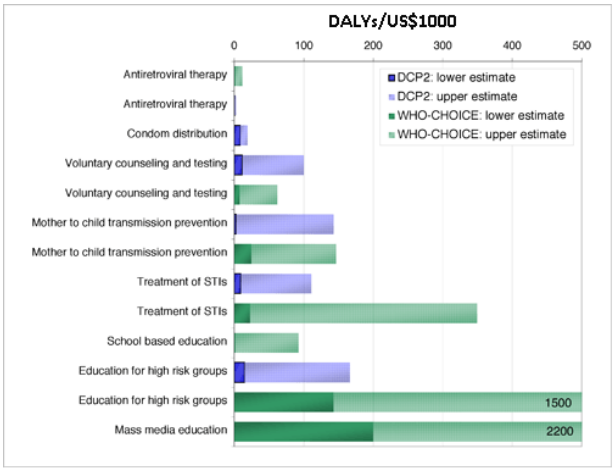


Figure 1: Cost-effectiveness (DALYs/US$1000) in fighting HIV/AIDS, http://www.givingwhatwecan.org

An important fact of life is that importance is typically skew distributed: the most important task you could do is often several times more important than the second most important task, and they can in turn be orders of magnitude above the median task. This appears to hold true for emissions reductions, healthcare interventions, disasters and many other things.

A consequence is that if your list of priorities is not in the correct order, you will focus on something less valuable than you could, suffering a big, possibly dominant, opportunity cost.

Results in decision theory show that optimal meta-level problem solving can at most give savings equal to half the expected utility difference between the two alternatives considered – but if the alternatives differ *enormously*, spending effort on this meta-level work is hugely valuable.

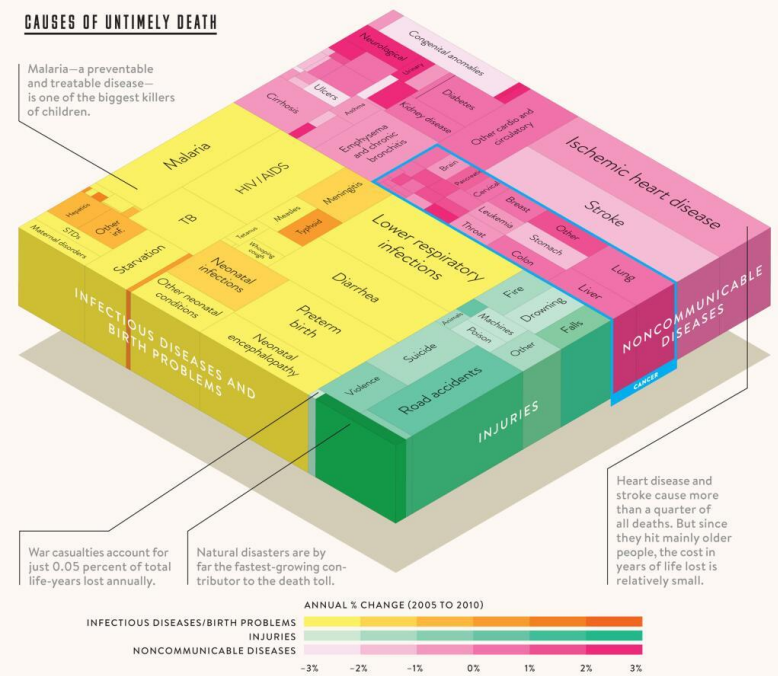
It is my suspicion that in many domains we spend far too little effort on asking meta-questions before setting out to actually solve problems. If you suspect that importance is skew in your domain you should consider spending more work on refining your questions, method and aim than you would intuitively do.

Applying this

What do these considerations tell us about Big Data and medicine?

Figuring out where we ought to be going, whether as the Big Data field or with our individual projects is obviously most essential. Given the importance of individual health multiplied by billions of lives, we should likely give this much more thought than we currently do.

In particular, **can we discover new methods**? Historically, we know science has advanced the fastest when new domains of nature were made available through new instruments. Building good instruments, the digital counterpart of the microscope or spectroscope that make answering medical questions easier is likely to speed up progress. And selling them to the people flocking to the new field is going to be profitable.



http://vizhub.healthdata.org/gbd-compare/

The map above suggests several strategies for choosing a target. We may aim at the big losses of life-years such as stroke, diarrhoea and heart disease. But some are likely more elastic than others: find out which – or if you could *change* the elasticity of the problem!

We also want to **tackle the big things as early as possible**. One reason is to have more research time to deal with them; another is to defuse potential problems. One example is predictably upcoming illnesses like **diabetes**, which will become significantly more problematic on the global scene as the first generation exposed to affluent diets in China and India get older, and **antibiotics resistance**, which predictably will emerge as pathogens evolve.

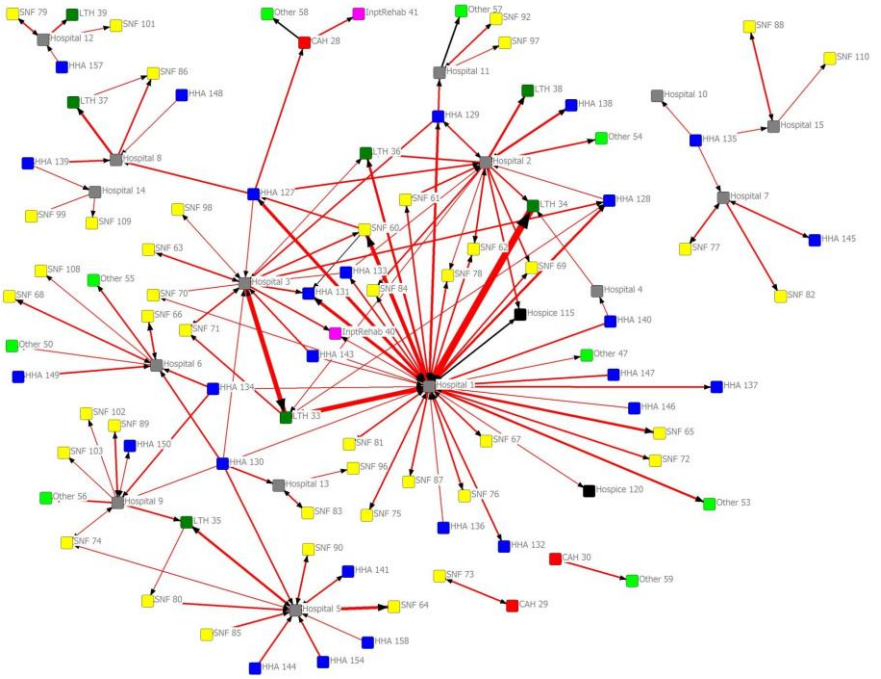
Anything that can grow exponentially should have priority over things that merely grow linearly – getting hold of it early gives vastly better payoffs than trying later.

Attacking root causes and extreme value outliers also makes sense. **Ageing** is a root cause for a large number of chronic, important diseases: slowing or stopping ageing, or breaking its link to the age-

related illnesses would have a great impact. **Preventative medicine and improved immune systems** are clearly high value targets since improved health by definition blocks a large number of conditions. Anything **multiplying human capital** (such as cognitive enhancement or improved training) has large network effects and allows better future decisionmaking.

Pandemics represent rare but potentially deadly problems. Most years influenza is not a problem, but occasionally something like the 1917 influenza comes along killing tens of millions. Outbreaks of SARS or Ebola shows the tremendous risks and costs of our current limited pandemic readiness. The importance here lies in the tail risk: while most instances are manageable, *occasionally* they represent global catastrophic risks, and that make finding a robust solution more important than it looks.

Another high priority question is where the **bottlenecks** are in our medical systems. In most systems there is one step that is the rate limiting step, and improvements elsewhere typically have little effect. Fix the medical bottlenecks and the system will become more cost-effective and better able to produce health. Conversely, **iatrogenic illness**, illness caused by treatment, is already a huge problem and will be growing in the future. Safer medicine means more medicine can be done. Finding ways of tackling the adverse effects of medical treatment can be as important as inventing new treatments.



Shreveport, Louisiana Hospital Referral Region Provider Network Analysis

Moral issues

A final consideration: consider ethics.

Some questions *should not* be answered. The main example in Big Data is privacy breaching questions. Actually preventing them from being answerable when the data is available is the target of privacy engineering. But to start caring about implementing privacy protections you need to notice that there could be potential privacy breaches. In many domains this is not entirely obvious, and people get nasty surprises.

Conversely, there are other questions that *should* be answered. We have a moral obligation to try to answer important questions that help the world.

Consider that the costs of gadgets and pills tend to come down exponentially over time, as improvements are made in manufacturing and organizational learning, and R&D costs are diffused over many products. Meanwhile the costs of services tend to remain constant or even grow, since they include the salaries of people doing the service. As long as healthcare is mainly a service it will be expensive and not everyone can access it. The more the essential parts of medicine can be done by pills and gadgets, the more all of mankind can afford it. We have a *moral* reason to automate it as much as possible!

There is no more motivating answer to the question "Why am I doing this?" than "Because it will help humanity as a whole."