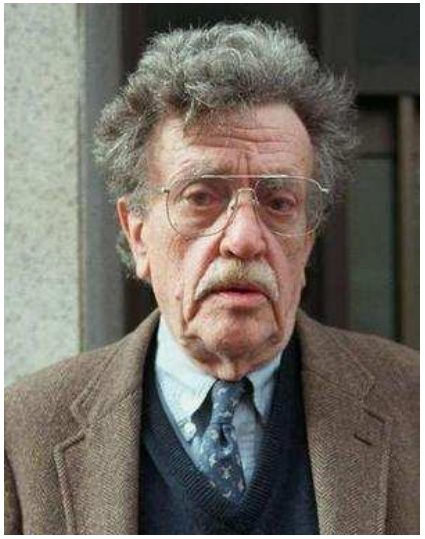
Building Virtual Organizations

Kurt Vonnegut in Cat's Cradle defined a **granfalloon** as an organization of people who think they have some common purpose, but actually don't (e.g., GE, the DAR, and the Communist Party).

In contrast, a **karass** is a group of people who, often unbeknownst to them, are working together towards some greater end (e.g., Wikipedia contributors, workers in an economic system, functional families).

The CDI solicitation appears concerned that the modern scientific community is a granfalloon, but should operate more like a karass.



A famous example of virtual organizations arises in Epstein and Axtell's (Growing Artificial Societies, MIT Press, 1996). They describe a "sugarscape" of interacting agents. The rules for the agents generate interestingly complex emergent behavior that mirrors human activity.

Joshua Epstein

The simplest rules tell the agent to move across the sugarscape, searching for food. Other rules describe mating, then small group interaction.

The upshot is that with a very small and transparent set of rules, which are easy to adjust and validate, one can mimic demography, epidemics, migration, barter economies, division of labor, and so forth.

Robert Axtell

This is one famous example of the agent-based models that have transformed modern simulation.

Statisticians have not been attentive to agent-based modeling. One wants a principled understanding of the properties of such models.

In particular, one would like to know:

- What is the essential dimension of an agent-based model?
- When are two simulation models "close"?
- How does one quantify uncertainty in such models?

Consider the Kermack-McKendrick model for epidemic spread:

$$\begin{aligned} \frac{dI}{dt} &= \lambda IS - \gamma I \\ \frac{dS}{dt} &= -\lambda IS \\ \frac{dR}{dt} &= \gamma I \end{aligned}$$

Here $I(t)$ is the number of infected people, $S(t)$ is the number of susceptible people, and R(t) is the number of people who have recovered and are immune. The λ is the infection rate and γ is the recovery rate.

The Kermack-McKendrick SIR model assumes a closed population and constant interaction rates. It is very simple but still widely used as a starting point for epidemiological forecasting. It consists of three coupled nonlinear ODEs with two parameters, so the dimension is 2.

The same dynamics can be obtained from an agent-based model. Epstein and Axtell's sugarscape does this, and the strategy for modeling the closed population is clear. The rules for the agents can be simple or complex—they dictate how frequently they contact (infect) each other and whom they infect.

The agent-based model must have an essential dimension of 2, corresponding to the λ and γ in the SIR model (perhaps with some second-order complexity to capture geography). But it is entirely unclear how to use the rule set to infer the dimension.

Statistical mechanics, or perhaps many simulation runs coupled with Principal Components Analysis, may offer a way forward.

The CDI Solicitation

The three themes of the CDI solicitation are:

- enhancing human cognition by generating knowledge from digital data;
- understanding complexity in complex systems
- enhancing discovery by building virtual organizations.

It seems clear that the NSF wants to think about the creation of new infrastructure for scientific process.

Previous transformative breakthroughs include:

- email, which has enabled international collaborations
- search engines, which have largely replaced the cumbersome old libraries
- Wikipedia, which offers a new model for the creation of scientific value.

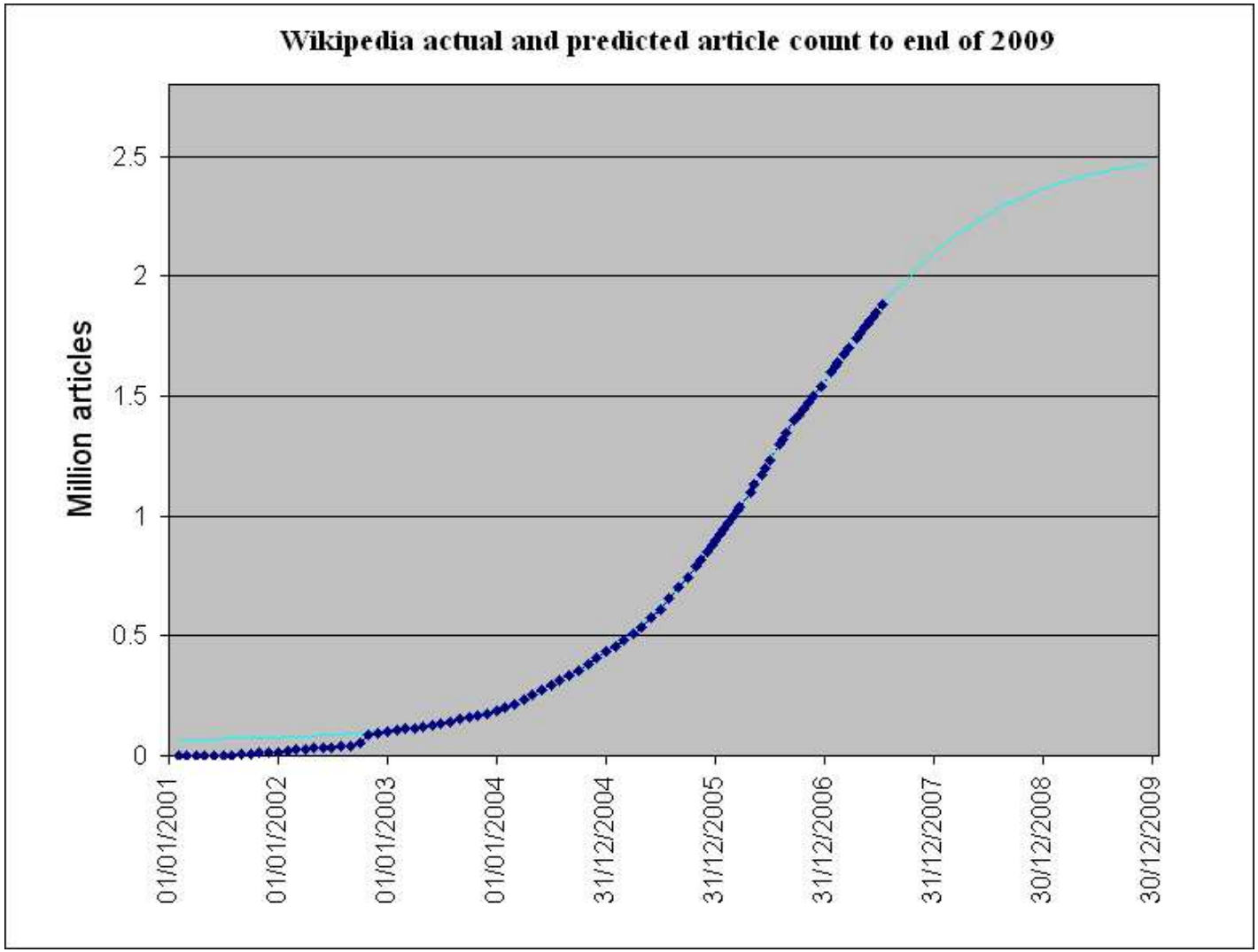
The latter has unrealized potential.

The Wikipedia began in 1999, conceived by Jimbo Wales (with help from Larry Sanger), and went public on 1/15/2001. Its innovation is to encourage highly distributed collaborative construction and revision of content.

Jimbo Wales

Key facts about Wikipedia are:

- It uses wiki-ware to facilitate collaborations and the GNU Free Documentation License to avoid legal problems with ownership.
- Its quality and accuracy are enforced by the user community (and according to Nature, **438**, 900-901, it is more accurate than the Encyclopedia Britannica).
- All the smart young scientists in every field use it a lot—it has become new research infrastructure.
- It has an open-access record of every change ever made, and who did it.
- It has internal and external links, and is a network model for the current state of human knowledge.
- It is the ninth most popular Internet site, and it grew very quickly.


Wikipedia actual and predicted article count to end of 2009
Million articles
2.5
2
1.5
1
0.5
0
01/01/2001
01/01/2002
01/01/2003
01/01/2004
31/12/2004
31/12/2005
31/12/2006
31/12/2007
30/12/2008
30/12/2009

In terms of the goals of the CDI, research on the Wikipedia seems like a good fit. Regarding

enhancing human cognition by generating knowledge from digital data

consider Marcus Hutter's prize of €50,000 to whomever can compress the first 10^9 characters of the 1/1/2006 English Wikipedia down to the Shannon limit.

The point of the prize is that strong compression of Wikipedia is equivalent (sort of) to the problem of describing artificial intelligence. There is a bit of controversy about this, but many computer scientists believe that winning the Hutter Prize would be a major breakthrough in understanding how brains organize information.

Marcus Hutter

As a second front on this same CDI theme, note that one could use statistical models for dynamic networks to predict the holes in Wikipedia. Success in that would enable researchers to identify the gaps in human knowledge. (**yes, I know this sounds a lot like science fiction...**)

The strategy would be to use the archived time slice of Wikipedia from, say, 1/1/04 and look at the new entries in 1/2/04. Repeat this process for each month. That provides data on where historical holes existed and were filled.

The explanatory variables in this effort would include local linkage structure, along with local latent space structure and bag-of-words models. From that, it is plausible to hope that one could build a model for in-fill.

As a related note, there is also interest in characterizing the Wikipedia connectivity structure of different disciplines, and understanding how and why they differ.

Regarding the NSF CDI theme of

understanding complexity in complex systems

the correspondence here is closely related to the strategy for generating knowledge from digital data. One would use models for dynamic networks to understand this complexity, and then use those models to identify in-fill opportunities.

Regarding the NSF CDI theme of

enhancing discovery by building virtual organizations

the Wikipedia is both an example of a revolution in virtual organization and it has become a key piece of infrastructure for the development of new scientific collaboration.

But the Wikipedia was not built by scientists, and the next breakthrough relies upon the development of new functionality.

Breakthrough functionality:

- How can one search for a theorem or a chemical formula or a word that means "red and green at the same time"?
- How can one link software tightly to articles, so that readers of an NSF-Wikipedia can dump their data or their physics problem or their epidemic parameters directly into the package?
- How can one build in dynamic graphics, sonics, etc.?
- How can the community police an NSF-Wikipedia?

These are daunting but not impossible. But progress on this would affect all branches of science in fundamental ways. And a major portion of the toolkit for some of these problems is explicitly statistical.