COMMENT ON DATA PROTECTION STATUTES AND BIOINFORMATIC DATABASES

Boston University School of Law
Law & Economics Series

2002

Wendy J.Gordon
Boston University School of Law

SYMPOSIUM ON BIOINFORMATICS AND INTELLECTUAL PROPERTY LAW

APRIL 27, 2001—BOSTON, MASSACHUSETTS

DATA PROTECTION STATUTES AND BIOINFORMATIC DATABASES

PROFESSOR DOGAN:

We have heard from the lawyer, the advocate's perspective of some of the legal issues involved in database protection and bioinformatics, and now we are going to hear an academic perspective on these issues. Professor Dennis Karjala is a professor at the Arizona State University College of Law with an interesting background. He has a Ph.D. in electrical engineering and taught in that field before going to law school at Boalt, and he is an internationally renowned expert on copyright law and computer law issues. Professor Karjala is going to talk to us about database protection issues. His presentation will be followed by some comments from Professor Simon Kasif, who is a professor of bioinformatics and biomedical engineering here at BU and also holds positions at Cambridge Research Lab and MIT. Professor Kasif has an interesting background in artificial intelligence and parallel and distributed computing algorithms. Professor Wendy Gordon, who needs no introduction to this audience at the School of Law, is one of the most renowned experts in copyright law and intellectual property law, and we will look forward to hearing her comments after those of Professor Kasif.

PROFESSOR DENNIS KARJALA:

Thank you very much Stacey, and thank you to all of you, and to the sponsors for inviting me. It is a pleasure to be here. I have been assigned the topic of database protection in the context of bioinformatics. I thought I would try to give a short introduction to the problem – why we are talking so much about database protection today, and particularly the proposals for new legislation – by saying a little bit today about scientific and bioinformatic databases generally. I will tell you what I know about them in the light of the goals of that I see for database protection. If time permits, I will compare some of the existing proposals for specific, statutory database protection outside the paradigms of traditional patent, copyright, and trade secret.

As I am sure all of you know, in the United States we protect compilations (if at all) under copyright law. For many years copyright protection, at least in the United States, was applied pursuant to the so-called sweat-of-the-brow theory of originality. This approach protected factual compilations in a kind of cumbersome and unpredictable manner. Although it lacked theoretical justification, it was, in my mind, serviceable in the days when most of the compilation litigation in copyright was over things like telephone books. In applying copyright to the protection of factual information, courts were trying to prevent market failure or misappropriation, where people spent a lot of time and money compiling data and, once it was available, others could take it very easily. This was a problem before the electronic age. It was exacerbated by the ease of copying works in digital form. The *Feist*[1] case, as everybody knows, changed the picture quite a bit. In *Feist*, the Court concluded that the copyright definition of compilation did not cover information as such, but rather only creative selections and arrangements of information. And by way of strong dictum, *Feist* raised this analysis to a constitutional level, suggesting that Congress actually lacks the power to protect writings that are not creative.

An important result of *Feist* was that facts are not protected by copyright. Consequently, under *Feist*, comprehensive, nonselective databases are simply not eligible for protection unless they are creatively arranged. Electronic databases, of course, have no real arrangement – if you are downloading just the data from them, you are not taking the arrangement or the computer program that does the organization. One side of the problem, then, is that *Feist* opens up these informational databases to the danger of electronic theft or misappropriation. The other side – this is the side that has worried me even more from the beginning with *Feist* – is that courts have strong negative reactions to what they view as theft, misappropriation, or reaping where someone else has sown. This activity engenders all sorts of negative reactions and colorful pejorative language. Courts try to find ways not to allow what they perceive as theft or misappropriation, and they have been trying to find ways around *Feist*.

So, on the one hand, *Feist* could leave some very expensive and very difficult to create, if intellectually noncreative, databases fully unprotected by the law. On the other hand, if the courts go too far in trying to find methods of protection, they end up protecting creative systems and methodologies for presenting information. I think that is equally dangerous. There are many examples now in traditional copyright where the courts are doing exactly that. The basis of the movement for database protection statutes is the former. Database producers are saying, "We do not have enough protection; *Feist* has left us out in the cold." The rest of us are worried about the latter, saying that maybe there is a need for some protection to fill some of the gaps created by *Feist*, but we do not want to go too far and actually end up giving intellectual property rights to factual information itself.

---

[1] Feist Pubs., Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991).

I will not dwell on the constitutionality question today, but a number of scholars have construed *Feist*'s dictum that creativity is a constitutional requirement to render much database legislation unconstitutional. If *Feist*'s dictum is in fact intended to be taken seriously, that is, Congress may not even have the power to adopt at least some forms of database protection legislation under the intellectual property clause. To protect such a work requires creativity and authorship, and that is not there by definition. The Commerce Clause, moreover, may not be available to override an express limitation on a specific power. Finally, there may be some First Amendment limitations as well unless the statute is properly tailored.

Feist was warmly welcomed, at the time, by many of my colleagues in academe who study intellectual property. I was always somewhat nervous about it for the reason I just mentioned: I thought it was going to induce, and I think it has induced, courts to become overly protective in areas of functional methodologies that really do not belong in copyright at all. So I think *Feist* has been a problem, but for those who are interested in bioinformatics in particular, it may be what was needed. I do not think the old sweat-of-the-brow theory of originality was really up to the task of balancing incentives for creation and the needs of users in any specific community, and, particularly in the scientific community. As we just heard from Tom Meyers, the kinds of bioinformatic information being created are complex. The information itself is dispersed, and we need much creative thought about methodologies for putting all this information together and discovering new scientific results.

I do not know that the sweat-of-the-brow theory, which as I say, was developed in the days when the biggest problems were basically telephone books, was really up to the job. And what *Feist* is doing is forcing legislative consideration. We have had consideration in Congress now for the past four years. I do not think anything has yet been introduced in the current Congress, but there are certainly threats. Nothing has been more than a threat thus far, but Representative Coble, Chair of the House Judiciary Committee's Subcommittee on Courts and Intellectual Property, has said that he will introduce legislation shortly if the two sides who are now arguing do not get together. The chairs of the House Judiciary and Commerce Committees have also said there will be database legislation this year. So I think there will be action. With proper database protection legislation, science may continue to progress in the way it always has with the benefits of full and free information exchange and yet still have the necessary incentives to create the bioinformatic databases and other methodologies that may well be necessary to the progress of science. The question here is, what would a "proper" database protection statute look like?

Scientific databases are often very large and very complex. They are becoming indispensable in many fields of science, and not just biology. Important results, as Tom mentioned earlier, often depend on finding connections between apparently isolated facts. There are a few needles in very, very large data haystacks. Much of the science that is being conducted today seems to derive from the ability to combine data that is dispersed

throughout many sources. Apparently there are many biological databases today, including twelve or fifteen large ones in connection with the Human Genome Project. This latter may not sound like a large number, but scientists have been trying to coordinate data exchange among the people engaged in the Human Genome Project, and even purely as a bioinformatics problem, there have been many difficulties with format incompatibilities and so forth. The Internet – I think Tom has mentioned this problem – has exacerbated the problem because many individuals now publish their data as soon as they get it, and they are published all over the world and in nonstandard formats and conventions.

So, we have all this data out there in need of coordination. Tom distinguished between data and information, and, that is a distinction we must bear in mind. We, that is, the scientific community, must have databases that are, in fact, usable, and we cannot expect everyone to be as fully conversant with the informatics aspect of bioinformatics as the computer scientists. Thus, we need databases that biologists from a wide range of fields can use without wasting time trying to get up to date on the information technology. In other words, they are interested in data, not information. The information technologists who create the databases also have an interest in not tying up data, as opposed to information, too strongly. They might get patents for all sorts of great acquisition, retrieval, and presentation inventions (information), but if they do not have any data, they will not have much of a database. These patents are not going to make much money unless data is available. Thus, if people have intellectual property rights down to the level of the data, downstream problems for later use and combination will affect not only the biologists seeking to use the databases to make important combinatorial discoveries but also the rightsholders in the technologies underlying the databases.

Another potential problem of data protection is that scientists may become reluctant to share data as they have in the past. All lawyers follow the rule that if you do not know what you are doing, you keep things close to your vest. Scientists, who are used to the free exchange of information, may start saying, "Well, I have some data here that does not seem to be of much use right now, but it is very interesting, so who knows? Maybe there could be some use for it, so I better not release it." There is also the argument on the other side, though, that proprietary rights may not interfere very much if economic forces are working properly – that is, proprietary rights may not slow down the exchange – because people will find that even if they have proprietary rights, they can make more money by pooling and sharing than by hoarding information. Where one party has just a small piece and a hundred other people each have just a small piece, however, nothing is accomplished unless everyone puts it together in collaboration. If that is the way things develop, that will be fine, but I am not so sure we can just start giving a lot of statutory protection and hope things will develop that way. Free exchange of scientific information has been the norm from the beginning of science. We should hesitate before establishing rules that threaten to change that norm.

Bioinformatics, in particular, requires the integration of data from a wide range of sources. This is another point that Tom made: Biology is increasingly a computer game, and we have the biology and informatics people working hand-in-hand as developments in one field feed the other back and forth. It is a very complex game now, and development of the necessary databases is itself dependent on access to data from diverse sources, so we have to make sure the data is available. Building on existing results and including them in new and improved databases is, to me, a simple and obvious way of building ever better databases, but achieving that result through legal incentives requires some thought. If we start giving database rights, we will begin with a few databases in which people have rights. When people later think of ways to build bigger and better databases, they will have to negotiate for the information from the first generation of database builders. It could well be that the vast amounts of information that are arising out of bioinformatics will create an anti-commons problem, with too many people holding too many rights in too many small pieces. That, too, is something we have to think about.

The question is, then, how do we insure continued database development and improvement, once we start giving database rights? I think we must remember what the goal of database protection legislation is. That goal is the narrow one of curing the market failure – if there is a market failure – that was created by *Feist*. I would recommend covering only databases that are offered as databases for commercial purposes, and I would limit the coverage to databases that are truly comprehensive and cohesive bodies of information, not just traditional scientific information that may in fact be a collection of inter-related information. We want to minimize the effect, obviously, on the traditional exchange of information among scientists. We also want to minimize the effect on intellectual property law generally, because the problem (if there is one) that we are trying to solve is quite narrow. We want to make sure that "databases" do not swallow up all of copyright law, for example, which has been carefully honing its balances between creation incentives and free use for several centuries. Therefore, we should exclude from database coverage everything that is covered by patent and copyright, with the exception of traditional compilations. I am not so sure that the existing proposals, certainly not those in Europe, do that, although I do not have time today to analyze the statutory language (which may change, in any event).

Finally, we need to deal with the question of downstream uses and non-commercial copying. I do not have too much time left, so let me just give you a brief introduction to what is happening with database protection in Europe, where an actual Directive is in place in the European Union. Discussion of one recent development, I hope, will show some of the real problems that can come from database legislation. The main point I wish to emphasize about the EU directive is that it provides a right for protection in databases that show a "qualitative or quantitative" investment, and prohibits extraction that is "qualitatively or quantitatively" substantial. These words, "qualitatively or quantitatively," have been driving me crazy. They are the worst parts of the EU directive, and we will see why in just a minute.

The case I will discuss is a very recent one out of the UK.[2] It does not involve scientific databases, but the reasoning of the court, I think, shows that our worst fears about the EU directive may be realized. The plaintiff was the British Horseracing Board, which collects data on all horse racing in England. They have mountains of data in their database. The defendant, William Hill, is the largest bookmaking organization in England. It has shops all over the place, which were not involved in the suit, and they are now allowing betting over the Internet. William Hill obtains information on today's races from a satellite feed that itself is licensed by the plaintiff British Horseracing Board. The satellite feed gets the information directly from the plaintiff's computer, and then, William Hill gets the information from the satellite feed for screen display showing what horses are running at what tracks and at what times.

Now, believe it or not, the court held that without getting a license from the British Horseracing Board, William Hill was violating the EU Database Directive and violating the plaintiff's database rights. The court found that there was systematic arrangement of the database, which is a necessary condition under the Directive for protection of the database. The court goes on to say, however that it is the data that is protected and not the arrangement. Thus, the condition for protection is fully divorced from the protection that is recognized when the condition is satisfied. Even though William Hill only took a tiny portion of the total amount of data stored, it relied on the completeness and accuracy of the information, as of course would anyone who was going to make a bet. So, qualitatively, this was very important information, and that was enough for the court. Taking the information indirectly from the satellite service rather than directly from the database does not change anything because the use is substantial. Even the fact that the same information was published in the Racing Form makes no difference. As a last resort, William Hill offered to just number the horses rather than name them, and omit giving the race times by saying first race, second race, third race, and so on. The court concluded that even doing that would violate the Directive because defendant would still essentially be getting the information out of the protected database.

As I read this decision, it gives a complete monopoly on information that only the plaintiff Board can create. The Board makes the decisions on who races where, so it is operating a what has been called a "single source" database. But what kind of incentive is needed to create this kind of database? The Board must create this information in any event. We do not need to give the Board an incentive to tell us who is racing today. Its business is horseracing, and nobody is going to place bets on races unless they know who is running, where, and at what time. This is information that is going to be produced independent of any database right. The court's interpretation results

[2] British Horseracing Board v. William Hill (High Court of Justice, Chancery Division, Patents Court, Feb. 9, 2001).

in a property right in information that does not add anything to the intellectual property equation.

It is extremely important for the future of bioinformatics that we consider this question of property rights in scientific information very carefully. We may need some sort of intellectual property protection in order to supply the incentives to make the necessary investments, but we have to be very, very careful that we do not go as far as the court in *British Horseracing Board.*

Thank you. (*applause*)

PROFESSOR DOGAN:

We will now hear from Professor Kasif and Professor Gordon, and then we will take questions at the end.

PROFESSOR SIMON KASIF:

Good morning. Thank you very much for inviting me. I will start my short presentation with a couple of apologies. First, I am not very well educated in the law, and I find myself a little bit humbled talking to a group of people who are very smart and have thought about these issues in detail. All I know about the law is watching Ally McBeal here and there. (*laughter*) So that is my first apology. The second is that I was told this meeting was very informal. I do own a tie, so – (*laughter*) I was also told by the organizers that I should give some general comments rather than focusing on a specific topic. I will therefore make some very high-level comments about bioinformatics and the legal challenges it creates.

One important point to remember is that work in biology and computer science relevant to this field has been going on for a long time. However, a real revolution happened in 1995 when Craig Venter and his group published a paper in *Science*[3] on their high-throughput – some call it a shotgun – sequencing of the *Haemophilus influenza* bacterium, and it really changed the biotechnology world. I have been working with TIGR and other major sequencing institutes for many, many years, and I watched this revolution unfold, which triggered the information explosion we are seeing today.

The most important causal event to focus on from the legal point of view here is the radical change that this genomic revolution created in a biology laboratory. In the past the classical approach taken by many biologists, even very famous ones with a lot of resources, was reductionism, focusing on research dedicated to a single gene or protein. A big lab was working on something like a P450, or other proteins that have important functions such as oncogenes, G-proteins, receptors, transcription factors, or enzymes. This approach was popular for years and years and indeed resulted in numerous

[3] *See* R. D. Fleischmann, et al., *Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd*, SCIENCE, July 28, 1995, at 496.

significant discoveries produced at a relatively slow pace. Even Craig Venter, before he switched to the new high-throughput biology mode, was essentially working on a single protein.

The previous classical approach to biology research naturally implies that the number of patents we can file is obviously limited. The first genome sequenced (*H. Influenza*) is quite long, and, generally, in a typical bacterial genome you might be filing patents on 4000 genes. If you basically switch gears and focus on eukaryotic genomes like the human genome, you might be filing patents on 40,000 to 100,000 genes. Therefore, sequencing and other high-throughput profiling technologies affect the level of engagement of this particular community, I think that is where the legal community might want to think about its ability to track and legally support this rapid pace of new inventions and discoveries.

The high-throughput capability of doing sequencing is one element of this new reality, and the second element is the computing power that has now made the process of interpretation of data largely automatic. I was a member of the International Genome Consortium at the Whitehead Institute at MIT that did the human genome analysis in a "competitive" effort to the one at Celera. The majority of the effort to produce the annotation of the genome, was done more or less automatically by computers. So again you have this extremely high-throughput information pipeline that goes from biology into computers and generates many new potential genomic territories where you can place your legal markers and establish turf. I think that's really the key issue for this community to worry about.

Now I want to pose a few questions that I personally find very challenging, though I am not sure how to answer them in a way that is simple and satisfying to a large number of scientists, medical doctors, and most importantly people that actually need the cure that is provided by some of the drug targets.

The questions are: What is a gene? What is the function of a gene? What is a genome? Are there any laws in biology? And what is a bioinformatics database? I will go through these questions quickly. Sometimes we are used to thinking by analogies; the best analogy I could come up with was this: The human genome looks a little bit similar to the Web. Why is that? It is very large and noisy. It is a little bit about sex. (*laughter*) It gets attacked by viruses. Portions of it are copied continuously. It needs lots of machines for analysis. The more you know, the more confusing it is. When you know a little bit, you start a dot-com. And the legal implication is that it is a mess.

So I do not know if this gives you any kind of legal precedent (*laughter*), but I find these seemingly disparate phenomena very similar. So what is a gene? There are many confusing definitions, and of course in high school, you hear the word gene and you associate it with an inherited trait – the gene for blue eyes, right? – which is highly misleading actually. I have a daughter who has red hair, and neither I nor my wife have red hair. Initially, I was quite worried. (*laughter*), so I went to my colleagues, who are biologists, and they explained to me that color of hair is a very complicated complex but that I was probably okay. (*laughter*) Anyway, the definition of a gene is rather

complicated. In the context of filing a patent on a gene, I want to point out a few technical matters we will go into next. Here is a very simple picture of what is called the central dogma in biology. One of the most important images in biology corresponds to an image in a region in a genome, which we call a gene, and we find many such typical pictures on the human genome.

The interesting thing to notice about this picture is that it has regions within the so-called gene that contain sequences that code for a protein, and they are called exons. Somehow biology is very clever, being able to cut out the unimportant regions called introns and glue (ligate) together the exons to produce a single sequence which then can be translated into protein, which you already know is one of the building blocks of life. Genes are sequences, but when you say something is a gene, it is a little bit confusing, because if you look at this picture, you have this long region that has exons and introns, and the exons are put together by biologic mechanisms to produce a protein. But here is the real complication. Most of you have probably read that both teams that worked on the human genome found less than 30,000 genes, and it was a big surprise. What happened?

One of the plausible explanations has to do with a process called alternative splicing, and alternative splicing means basically that there are many different possibilities for the biological mechanism to cut the exons out and glue them together to produce different forms of proteins. In particular it is more likely to occur in human beings than in lower-order eukaryotes. Our brain, for example, has many, many neural receptors that are very commonly alternatively spliced, and, on average we expect every gene to have three or more alternative splicing mechanisms. Thus, there are three alternative proteins that emerge from a single gene that could each have a different function during development, during different ages, during different evolutions of the cell, and so on.

As we move forward, this is an important point to remember. Consider a microbial genome. It has coding regions and intermediate regions, called intergenic regions (no introns). In microbes, it is pretty simple: There is one region that codes for information, which may be called a gene. In human beings, there is a more complicated picture with regions coding for proteins and the intermediate regions do not, but, surprisingly, you can put many patterns together in an alternatively spliced coding regions in many different ways, and as a result you can get many different answers what a gene might be.

Now that we know, more or less, what a gene is – of course, we also know it is difficult to say what a gene exactly is – let us ask a different question: How do you assign a function to a gene? That is really the biggest mess to address, and we are dealing with it daily. As it turns out, it is actually another complicated topic. As you know, nature tends to be very lazy. It does not like to invent new things; it likes to reuse things that were already invented. As a result, genes have multiple functions in different contexts. So when you say something has a function, it is a very subtle statement, and one of the reasons people are having trouble unifying databases is that they very often do not agree on function. One person is using a neural network to predict what a gene

function is, while another person is using a different kind of predictive network, and, as a result, they get different answers and cannot unify them. Of course, there are big egos involved, which is a factor as well. There is MIT, and there is BU. It is not so easy to unify those things. (*laughter*)

One potentially useful way to think about a database of genomic information is as a database of cases, which I will explain in just a moment. Some of the most important gene laws I can think of from the point of view of intellectual property issues and patent laws. First, many genes are similar to each other. In particular, there are some gene clusters or protein clusters that have 10,000 to 20,000 members. So when you say, I have a patent on this gene, you really are defining a very large territory, and that territory is very difficult to characterize precisely because we know very little about that family. Sometimes, as we go to the third or fourth item in the family, we see only ten to thirty percent of identity among members in a sequence – ten percent identity of a very long sequence might mean they have the same function. However, sometimes one little change will make the two genes work very differently. When you patent that element, most of the time you are not aware of what a mutant is going to do, so how are you going to argue later on whether the new behavior of a mutant was already anticipated?

So basically when you talk about a gene database, the sequences on their own have very little meaning. Someone mentioned this – databases versus information. Very often what we do is we actually group those sequences together into what we call a profile, which is a descriptive way to basically describe this family like a case in the legal system. And then we use analogy and similarity methods to argue that this gene and another gene have the same function because the following similarities hold over. That is a very important argument in this field, in bioinformatics, because those similarities are studied by researchers, papers are written about those similarities just as papers are written about case-based analysis. In this context, I think the most important gene law is, perhaps, the exceptions to the law. Genomic databases are similar to databases of legal cases because we conduct analogy-driven reasoning which perhaps leads to some interesting methodological implications for both communities.

To summarize, the increase throughput of biological discovery poses many challenges to scientists and lawyers. Since many of these discoveries present important opportunities to improve the quality of life, much legal work has to be undertaken in order to make the process of moving biological discovery into practice efficiently. In other words, the future looks bright for biologists, engineers and lawyers.

Thank you. (*applause*)

PROFESSOR. WENDY GORDON:

Dennis Karjala helpfully covered most of the issues. I have a few points to add.

This is the second time that Professor Karjala and I have met on the issue of databases. I am glad to say that he seems to be more cautious about database ownership now than he appeared to be then.[4] Caution is appropriate, as I hope my preliminary remarks will help to show. I will close my remarks by discussing a philosophic dilemma inherent in the controversy.

First I turn to the practical issue of assessing what consequences database protection would produce. I do not think there is a proven need for giving persons who produce noncreative databases legal protection against copying. At most, there may be need for the law to give database-makers a limited tort right to restrain, for a short periods of time, wholesale and commercially destructive appropriation.[5] Rights that are broader, especially those that approach "ownership" rights, are very troubling. They pose a danger to the spread of knowledge, and I suspect will yield little countervailing benefit.

I have some hypotheses about why, aside from considerations of *Realpolitik,* Congress might keep toying with statutes which would give ownership rights in databases. One hypothesis is that many on the Hill share the plausible, alluring, but incorrect assumption that privatizing ownership always results in increasing wealth.

Privatization may reliably increase the wealth of some individuals, but the private capture of wealth is only part of the story. More important is the *overall* total of social welfare. Except in unusual cases that have special justification, as a society we do not want to increase private individuals' wealth at the expense of everyone else. American copyright law already protects creative arrangements and creative selection of data. What it does not do is give copyright protection for data itself. To change that rule – to privatize public-domain information – could cause a net decrease in social wealth because it would raise the price of follow-on innovation and in some cases could even prevent later developments.

When the suggestion is made to give ownership in data, therefore, Congress should not rely on intuition. It must demand empirical proof that this legal change will effect a net benefit socially.

I think the pro-ownership attitude is also partially driven by sibling rivalry, the notion that Germany and France have something that we do not have. They have the EU Database Directive,[6] and so by reflex, some of us want it too. But I think the EU countries are, in the long run, going to be worse off with their Database Directive than they would have been without it. We should not emulate an action that would be burdensome. We need to be smarter than

---

[4] *See* Dennis S. Karjala, *Copyright Symposium Part II: Copyright and Misappropriation,* 17 U. DAYTON L. REV. 885 (1992).

[5] For an alternative formulation of a database tort, *see* Wendy J. Gordon, *On Owning Information: Intellectual Property and the Restitutionary Impulse,* 78 VA. L. REV. 149, 222-23 (1992).

[6] Also important is the related issue of reciprocity. *See* Amy C. Sullivan, *When the Creative is the Enemy of the True: Database Protection in the U.S. and Abroad,* 29 AIPLA Q. J. 317, 325 & n.93 (2001).

those boys who were bamboozled into taking over Tom Sawyer's fence-painting chores.

If we are going to have database protection, however, I think Professor Karjala well indicated some of the main things about which we have to be careful. First of all, in terms of what counts as an actionable taking from a database, it should not be just "any" taking or one that is so flexibly defined, quantitatively or qualitatively, that the law has a chilling effect on creating databases. People who use data can generate large benefits for all of us. We do not want vague laws to discourage beneficial behavior. If there is going to be any database protection at all, it should be clearly defined to operate only against wholesale, large-scale, commercial misappropriation.

Second, in terms of what kind of databases should be covered, there need to be limits on subject matter. As Professor Karjala indicated indirectly through his discussion of the British horseracing case, one of the things that should not be given over to private ownership is data that comes from only one source.

For ordinary data, ownership means that potential users must either purchase access rights from the owner, or "re-invent the wheel" by gathering the desired information themselves. Among other things, this second option places a ceiling on the price that the 'owner' of a database can charge: He cannot charge more than it would cost a stranger to gather the material himself. But with sole-source data, this salutary ceiling can be absent. If our law were to give to the sources of sole-source data the right to own it, that could foreclose all other persons from the possibility of gathering the data independently. Thus, unique monopoly problems would ensue from protection.

Also, the producer of unique information, the source of the data, often has no incentive difficulties. If his activity generates data that is publicly valuable, then he often will have an activity that is valuable in itself and that generates revenues. Thus, football teams generate data – game scores – and also generate much revenue from charging admission to the games and for licenses to broadcast the games. Whether or not the teams can own their scores, their valuable activities will continue, and for them, the incremental incentive effects of data ownership are therefore small.

Most important are the issues of scientific progress and free speech. From the perspective of the First Amendment I can imagine few more pernicious dangers than putting a fact into the ownership of only one person, or allowing an entity who generates a fact (whether the entity be a football league or a government) to control how it is used.

In the United States, the National Football League did sue the State of Delaware once for daring to make a lottery that made reference to the NFL game results.[7] But in our country, the sports league lost their litigation, unlike the British Horseracing Board.

In discussing this issue of sole source, Professor Karjala highlighted many of the goals that can be served and disserved by database protection. I am

[7] National Football League v. Governor of State of Del., 435 F. Supp. 1372 (D. Del. 1977).

hoping that this afternoon, Professor Arti Rai along with the virtual godmother of the field, Professor Rebecca Eisenberg, will speak further on how the law might minimize the negative effects on traditional information exchange among scientists. For myself, in the short time remaining, let me add a simplifying model to illustrate a philosophic tension that underlies many issues in data policy.

First of all, as Professor Karjala observed, there is a great desire among courts to punish a defendant who appears to be a bad actor. I suggest that the most relevant type of bad actor can be seen in the writings of two quite different philosophers: John Locke and Emmanuel Kant. That bad actor is someone who is motivated to act solely because he sees someone else who is vulnerable, and he wants to take advantage of that vulnerability.

Often what we do happens to hurt someone else. Someone who enters a race wants to win, and if she does win she will deprive someone else of the chance to be first. But no one considers that kind of harm morally wrong. Much of law and morality seeks to judge what harms constitute wrongs and which do not. The question is difficult when someone who has an independently-determined course of behavior happens to hurt someone else, but it seems clear to all of us that someone acts wrongfully when he chooses a course of behavior only because he wants to make himself better off at another's expense. Such an actor not only prefers his own welfare to another's, but also is willing to sacrifice that other person to his own ends.

In Locke's terms, this bad actor is the "covetous" person who is not "industrious" and "rational" but merely seeks to "take advantage of another's pains."[8] For Kant, this bad actor is a person who fails to live by universalizable moral tenets: He treats another in a way he would not want himself to be treated.[9]

What do we call this bad actor? I do not want to call him a "free rider" because we all ride freely on the culture we did not create. Free riding can be a wonderful thing, as Jefferson pointed out: We can all light our candles from your taper without diminishing your light. We all free ride without harming you, and thereby bring illumination to the world.

Sometimes – and only rarely – should we condemn those who "reap without sowing." What makes a civilization is the accretion of knowledge and skills that let most of the civilization's members "reap without sowing" most of the time. We need therefore to avoid the "free rider" label as drastically overbroad.

---

[8] JOHN LOCKE, TWO TREATISES OF GOVERNMENT (Peter Laslett ed., Cambridge University Press 1988) (3d ed. 1698, corrected by Locke), *as discussed in* Wendy J. Gordon, *A Property Right in Self-Expression: Equality and Individualism in the Natural Law of Intellectual Property*, 102 YALE L. J. 1533, 1535-78 (1993).

[9] On Kantian notions of doing unjustifiable harm, *see* Wendy J. Gordon, *Truth and Consequences: The Force of Blackmail's Central Case*, 141 U. PA. L. REV. 1741, 1758-66 (1993).

Instead let us call the bad actor I see in both Locke and Kant an "opportunist." I think it is the image of this opportunist and the desire to strip him of his perceived unjust enrichment that motivates many courts and policymakers.[10] And if the world of intellectual property had only two parties – the hardworking laborer and the opportunistic second comer – the moral case against the opportunist would be inarguable. However, in the real world, it is never just two people who are involved. There are also customers and the larger public.

If the opportunistic bad actor is restrained, a valuable kind of dissemination and creative use may be restrained. Whatever the opportunist's motives, she may serve the public good. She may make data available at lower cost or otherwise make it capable of reaching additional persons or being put to new uses by customers. Therefore, there can be conflicts between the morality of opposing opportunism and the consequentialist morality that seeks to increase the amount of welfare in the world.

To envisage this, consider ordinary competition. If you develop a new market and come up with something that is not patented but serves a great public need, Mr. A may be lured to your market simply by the promise of earning a supernormal profit. By copying you, he can earn some easy dollars. You have done all the marketing research, you have taken risks, and you have discovered this great niche. And the only reason Mr. A is going to follow on is because Mr. A thinks there is a profit he can obtain by selling a little more cheaply than you and by taking from your market. Mr. A certainly looks like an opportunist. But Mr. A is also the person who is going to drive the price down and help consumers get a valuable product much more easily.

Judges in intellectual property disputes tend to focus on the two most salient parties: the hardworking author or database gatherer, on the one hand, and, on the other, a defending party who may appear to be motivated by grabbing a quick buck. The latter looks like an opportunist, and a desire often arises to make the opportunist "disgorge" what is seen as his "unjust enrichment." [11] There is a problem, however, with giving data protection out of a desire to punish the opportunism seen in the two-person model: the effect on the larger public.

---

[10] *See generally* Gordon, *supra* note 5.

[11] The highest court of the State of Illinois gave protection to the Dow Jones Average based on a theory of misappropriation. Board of Trade of City of Chicago v. Dow Jones & Co., Inc., 98 Ill. 2d 109 (1983). The dissent noted:

> The majority is swayed by what it sees as 'unjust' enrichment – the Board of Trade's plan to earn a profit by the free use of an idea developed by Dow Jones at considerable cost. I do not regard this use as 'unjust' in the least. The Board of Trade proposed to use information that Dow Jones had freely allowed the public to acquire in a business that Dow Jones has not shown the slightest interest in pursuing. If 'unjust enrichment' has become the only element for the tort of misappropriation in Illinois, I fear that there will be few commercial ideas and little information left in the public domain.

Id. at 127 (Simon, J, dissenting).

Our social dialogue has many ways of integrating two-person and multi-person models of morality and of integrating consequentialist and non-consequentialist approaches. We must not hide behind the rhetoric of unjust enrichment or "reaping and sowing" to shirk these difficult tasks.

Thank you. (*applause*)

QUESTION & ANSWER SESSION

MODERATOR:

Questions from the floor? Yes.

Q: AUDIENCE:

Professor Gordon, what is the incentive, as the individual compiling the database to disseminate that information to the public?

A: *PROFESSOR. GORDON*:

That is a wonderful question. It is most useful when turned around: Given the lack of formal protection for noncreative databases now in the United States, why do we have so many databases? One answer might be that lead time, physical protections, and already-existing interstitial legal protections provide enough shelter for incentives to survive. Thus, a database producer can sometimes use encryption, physical walls, and the law of tangible property or contract to assist him. Some states might even be willing to give the database compiler a misappropriation cause of action. In addition, of course, federal copyright protects creatively arranged or selected databases. And perhaps most importantly, the authorized producer has advantages in being able to produce timely, guaranteed, and conveniently-packaged material.

Another strong reason why we have a great deal of data generated at the moment, despite the absence of ownership in pure data, is the existence of incentives in the marketplace other than revenues from sales. These alternative incentive sources include government funding, tenure, and the beneficial reputational side effects that a data-generator receives from publicizing her research.

I am still waiting for the empirical case supporting database ownership to be proven. I am open to the possibility that it can be proven, and if it were, I might change my view in some respects, but I am not yet persuaded. I will give you one example of the many kinds of alternative incentives that exist.

Back in the early days of *Feist*, Professor Karjala and I were at a conference addressing what was going to happen, now that the Supreme Court had ruled that no copyright subsists in the alphabetical and comprehensive listings of phonebook white pages. One industry representative noted that despite the

new decision that there continues to be a reliance on licensing to collect information for databases.[12] So in many instances, the *Feist* decision has not changed the business of licensing data access. Royalties and incentives were present even in the absence of ownership.

Now, I do not want to overstate the force of one example. Some observers point to *ProCD, Inc. v. Zeidenberg*[13] to suggest that sometimes the absence of legal protection will result in potentially destructive copying of databases. I am willing to concede the possibility, but only as a bare possibility. In my view, even in the instance of *ProCD*, neither the plaintiff nor the judge made a strong economic case for data protection for reasons I have explored elsewhere.[14]

In sum, there already exist clearly significant incentives that can keep money flowing to database creators.

A: *PROFESSOR. KARJALA*:

Just to add to the *ProCD* discussion – I am glad Wendy mentioned it – I think that is an example in which some form of statutory protection might be necessary or at least helpful. It is the same argument that I suggested in my talk for some very limited form of data protection statute: If we do not do it with a limited statute, we may find the courts doing it much more expansively, and *ProCD v. Zeidenberg* is a good example. In *ProCD*, protection of a digital telephone book was based on a shrink-wrap license. Protection on that ground is very dangerous. The *ProCD* court emphasized the amount of money and effort that went into creating the digital directory, which was electronically copied and offered free over the Internet. Thus, by remedying a perceived misappropriation that, if allowed, would have acted as a disincentive to invest in digital telephone directories, the court greatly broadened the rights of software suppliers who seek to control uses of their software – even uses that copyright law permits – pursuant to shrinkwrap or clickwrap licenses. That, I think, is very dangerous because the public interest is out of the picture

---

[12] *See* Steven J. Metalitz, *Copyright Symposium Part II: Presentation by Steven J. Metalitz, Esq.*, 17 U. Dayton L. Rev. 775, 783 (1992). The author continued:

> Within weeks after the [*Feist*] decision, information crossed my desk about a new CD-ROM product that consisted of scanning white pages onto CD-ROM. That product has proven successful and fills a market need. But at the same time the licensing agreement, even for telephone directories, is essential if you want the best possible product. If you want the most up-to-date listings, you do not want to wait until the directory comes out, you want to license the tapes of updates immediately, and for a variety of other reasons, while you may be able to make a non-infringing product without a licensing agreement, you are, in most cases, going to make a better product with a licensing agreement.

Id.

[13] 86 F.3d 1447 (7th Cir. 1996).

[14] *See* Wendy J. Gordon, *Intellectual Property As Price Discrimination: Implications For Contract*, 73 CHI-KENT L. REV. 1367 (1998).

completely, and the whole situation is reduced to two-party analysis. That is not the way intellectual property historically has worked, and, in my opinion, it is not the way it should work. Defining rights or remedies by means of legal rules that do not take the public interest into consideration will result in long-term losses that greatly outweigh whatever short-term gains come from the increased incentive. A properly limited statute can obviate this kind of judicial legislation.

Q: *AUDIENCE*:

I have a factual question for Dr. Kasif. In your work with bioinformatics databases, how much invaluable information that you deal with is in the public domain versus how much is in proprietary databases, either of your own creation or ones from which you purchased or licensed information for your institution's use?

A: *PROFESSOR. KASIF*:

That is a loaded question because, as you well know, there are different ways of answering those questions. For example, this is different from a public domain versus private access issue. I think what you will find across the community of scientists, there has been a split on this issue. The main reason there is a split is the following. It depends what you want out of the others. Most drug companies, no matter how big they are, actually chase very few drug targets. They probably know that the top pipeline is very expensive to use. We do not have resources to it. So the question is, why do they need this enormous database to chase one drug target? For the various reasons, basically they can be useful. But I think the answer to your question truly depends on the scientist's own application. Some information could be in public areas, and some different type information might not be in public areas. So another kind of database that may become very, very important are SNP databases. SNP databases are databases of single nucleotide polymorphisms. These are, for example, very important in cancer studies. A number of companies are doing database compilations of these SNPs.

Q: *AUDIENCE*:

I think it is very important to consider incentives in this analysis.

A: *MODERATOR*:

There is the additional question, of course, of whether those private databases exist, in part, through contract law and, whether the arrangements between the enterprises that rely on those databases are enforceable through existing contract law, or, whether sui generis protection is necessary or preferable, perhaps, to that contract law.

A*: PROFESSOR KARJALA*:

That sort of contract is often the problem because it reduces the deal to a two-party agreement. The public interest in free flow of information may not be served.

A: *MODERATOR*:

Right, so that gets back to your point that, perhaps, from the public access perspective, sui generis protection, if cabined, might be preferable.

Q: *AUDIENCE*:

I just have a follow-up to that. One important limitation on contract protection is independent creation. Independent creation is a risk to the protection the client is getting, but it is also a limitation on the risk to the public of these private agreements. If private agreements are too draconian, well, then, someone else can identify a market niche there, or, they can be less draconian and provide a competing database.

A: *PROFESSOR KARJALA*:

Well, of course, I do not think there are any database statutes in existence or proposed that would not create a defense for independent creation. The problem, however, is the way people actually work with an existing database. It does not make sense constantly to re-invent the wheel. By its very nature, progress will come from people's reading, understanding, and using other people's data. It is only after they have found something valuable that they could say, "Oh, gee, I could have gathered this particular information myself." But by that time, they have already used the existing data. In other words, until you know what is valuable, you do not know what data "independently" you should collect, and you only find out what data is valuable by looking at what the rest of the scientific community has already done. At that point, the data is no longer "independent." We should not adopt legal rules that change this approach to doing science by making it more difficult to find and use information already discovered by others.

Q: *AUDIENCE*:

We do have a SNP Consortium,[15] for example, and this is an ongoing private effort to create SNP collections. They say, "Whoa, we do not want to

[15] *See* SNP CONSORTIUM LTD., SINGLE NUCLEOTIDE POLYMORPHISMS FOR BIOMEDICAL

go there. We are instead going to pool our resources and create our own independent database." This really is a significant restriction on the ability of private database owners to market their information.

RESEARCH, *at* http://snp.cshl.org (last modified Jan. 2, 2002).