



Data Base Possibilities



Dr. Masakatsu Murakami is a Professor at the Ministry of Education's Institute of Statistical Mathematics, Graduate University for Advanced Studies. Born in 1945 in Nanking, China, he earned his Ph.D. in Statistics, specializing in social science. His research achievements include quantitative analyses of Nichiren's book on Buddhism, anthropological data and various artistic works.

Virtually any kind of information, from classics in Japanese literature to breakthroughs in genetic engineering, can now be stored as digital data. Not surprisingly, the volume of such data is increasing exponentially. Can a change in quantity affect the way we perceive qualities within a database? Is a new research method emerging? To explore these and other questions, we interviewed researchers who are entering the thick forests of data surrounding two very different subjects: The Tale of Genji and genetic engineering.

Digitizing manuscripts
for new literary perceptions
**The moment an imitation
is revealed**

Combining the ancient classical Japanese story of *The Tale of Genji* with walking up a hill beside foreigners in Tokyo's modern Minami-Azabu district seemed at first a mismatch. But by adding one more card to this hand—the

nearby Ministry of Education's Institute of Statistical Mathematics—the situation made sense. Located near Arisugawa Park, this laboratory is where Dr. Masakatsu Murakami is involved in digitalizing literature—quantifying a literary work, if you will—to help quantify the writing styles of particular authors. By analyzing the sentence structure of a book, for example, he can help determine whether it is authentic. If you were asked to determine the authenticity of a work attributed to a famous author, how would you do it? What if the only manuscript you had was a copy of the original made during a later period so you couldn't judge based on paper quality or handwriting? You might think about focusing on unique words favored by the author, which, until recently, was one of the standard methods.

“Unfortunately, special words don't occur too frequently. And if a manuscript was copied by many people and in different eras, the odds are good that certain words were added during that process. This is why when we started applying principles of statistics we used an opposite methodology, instead, focusing on the most frequently used words. Even if a copyist added or subtracted special words, the ratio of frequently used words to overall text

volume would not change.”

Basic words like auxiliary verbs, particles, or common pronouns for “people” or “things” appear regardless of the topic of the sentence. How many times do they appear?

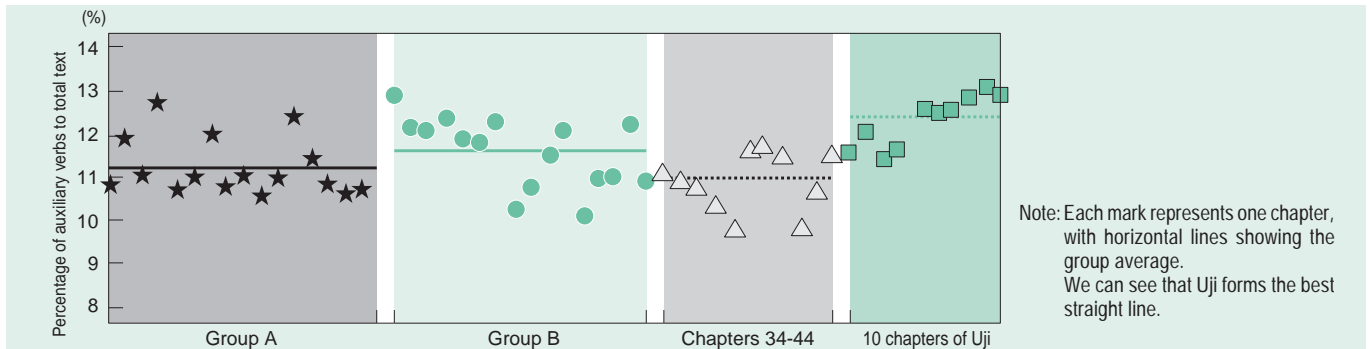
“The important thing is not how many times, but the percentage of total word volume. Also, you don't have to focus on only one word, but can consider whole word categories like nouns, auxiliary verbs, or prepositions. In the past, even if we had wanted to obtain such information it would have been impossible. We needed each word in the document to be input as a separate piece of data in order to calculate its frequency of appearance.”

As this method only takes into account words that don't have much meaning on their own, is it really a reliable way of determining authenticity?

“It's possible for an author to opt for different sentence lengths or certain unique words. But with small words like auxiliary verbs or particles, writers tend to use them unconsciously, without any special intention in mind. This is why they work well in helping us identify the unique characteristics of a writer. We focus on common, frequently used words whose ratio to total word volume is fairly stable—and then we look for discrepancies.”

Group	Number of chapters	Chapters contained in group
A	17	1, 5, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 32, 33
B	16	2, 3, 4, 6, 15, 16, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
C	11	34, 35, 36, 37, 38, 39, 40, 41 ----- 42, 43, 44
D	10	45, 46, 47, 48, 49, 50, 51, 52, 53, 54

■ Figure 1 54 Chapters of *The Tale of Genji* Categorized into Four Groups



■ Figure 2 Ratio of Auxiliary Verbs to Total Number of Words

This was precisely the method used in evaluating the authorship of the classic Japanese novel *The Tale of Genji*.

Theories on detecting inauthentic literature, considerations of chapter sequence
The Tale of Genji, approaching the final chapter of a mystery

The Tale of Genji, the most famous literary work of Japan’s Heian Period (794-1192 A.D.), is a 54-chapter novel chronicling court life through the many love affairs of Hikaru Genji. Authorship is attributed to Shikibu Murasaki, or Lady Murasaki, however, an original manuscript has never been found. Because only copies survived, speculation that certain words had been altered started as far back as the Kamakura Period, which followed. To add to the confusion, in the last ten chapters (collectively known as *Uji*) the main character is succeeded by his son Kaoru. This has led some scholars to doubt they were even written by the same author, a subject that has been extensively debated since the end of the 15th century.

Question: Is *Uji* truly different from the first 44 chapters?

Dr. Murakami and his associates input the entire text of *The Tale of Genji* into a database, categorizing all the words by grammatical function. This task took five years and comprised entering

376,000 words. “Even though there already existed a large database containing sentences, we were not able to use it for quantitative analysis. To do that, we needed to categorize sentences by individual words. In addition, we put all verb conjugations, adjective forms, etc. into the same category.”

Dr. Murakami then split the chapters into four groups. (See figure 1.) Chapters 1-33 are a nonsequential mix from Group A and Group B, which cluster different story types. After determining the percentage of auxiliary verbs, he compared *Uji* with the other three groups. The *Uji* group scored 12.4 percent to each of the other groups’ 11.3 percent, a statistically significant difference. Even on a chapter-by-chapter basis, he found *Uji* to be unique in that it was the only group where this ratio increased in each succeeding chapter. (See figure 2.)

“The fact that many people felt *Uji* to be different is now confirmed by statistical data. If you noticed the difference on your own, however, you have a keen sense of observation.”

Does this prove *Uji* was written by a different author?

“We really don’t know for sure. During an earlier analysis of Nichiren’s famous text on Buddhism, we discovered that with age he began to use particles less and less. (Dr. Murakami’s analysis takes into account the maturation of an author’s writing style). Your writing can

mature as you age. After reviewing our statistical data, the famous writer and female monk Jakucho Setouchi suggested that when Lady Murasaki became a monk, she underwent a spiritual transformation that changed her writing style.”

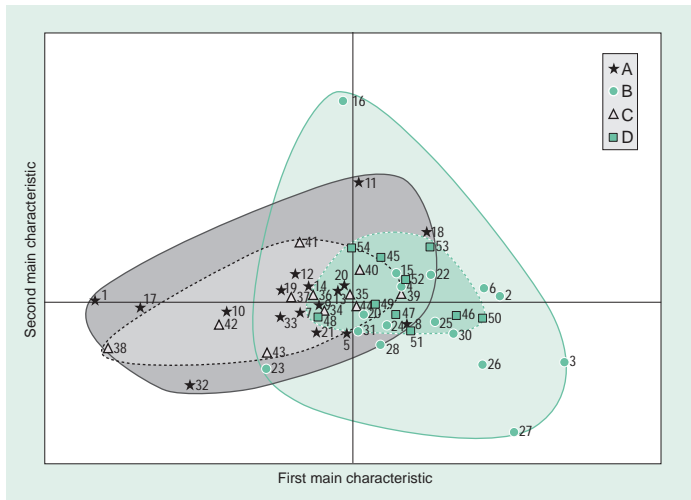
We can further broaden our perspective if we ponder the authorship of *The Tale of Genji* from yet another vantage point, the sequence in which chapters were written.

Question: Was *The Tale of Genji* written in sequence?

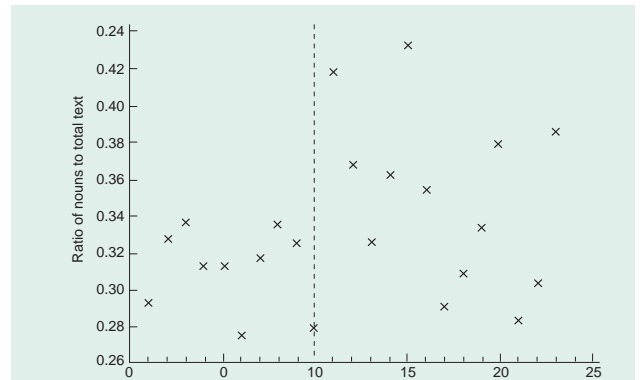
There is a theory that it was not, with the chapters in Group B later inserted into Group A. To investigate this, Dr. Murakami selected the 21 most frequently used words in *The Tale of Genji*, calculated the percentage of each word in each chapter and visualized his results in a 21-sector matrix. (See figure 3.)

“The closer the sector, the more similar the usage of the 21 words. You can see overlap between Group A and Group C as well as Group B and Group D. On the basis of this alone, it would be natural to conclude that the chronology of authorship was A, C, B, D rather than A, B, C, D.”

This would lend support to the theory that another author wrote *Uji* (Group D), especially if you read D immediately after C. But if you read D after B, the authorship might not seem so different.



■ Figure 3 54-Chapter Analysis for Two Main Characteristics (relational matrix)



The ratio of nouns to total number of words in 23 ransom letters sent to police by kidnapers of the president of Glico candy company (1984). There has been speculation on whether they were all written by the same person. "I think these sentences are too short to analyze," said Dr. Murakami. "But what do you think? There are more nouns starting in letter 10. Is this due to a new writer, or were the latter letters written by two people?"

■ Noun Ratio Contained in Letters Written by Kidnappers of Glico Company President

"We must work closely with literary scholars to keep a balanced perspective on *The Tale of Genji*. Our main strength is that we possess a vast database containing the entire text of *The Tale of Genji*, and we can use it to analyze many types of new theories."

This database will soon be available on CD-ROM.

"And someone might be able to come up with a more sophisticated analysis than ours."

Dr. Murakami focuses his analysis on words that were written without conscious intention. Similarly, he determines whether Buson's sumi-e (traditional B&W brush paintings) are authentic by focusing on the small leaves or on trees in the background. With some hesitation, Dr. Murakami pulled from his shelves a book rife with bookmarks.

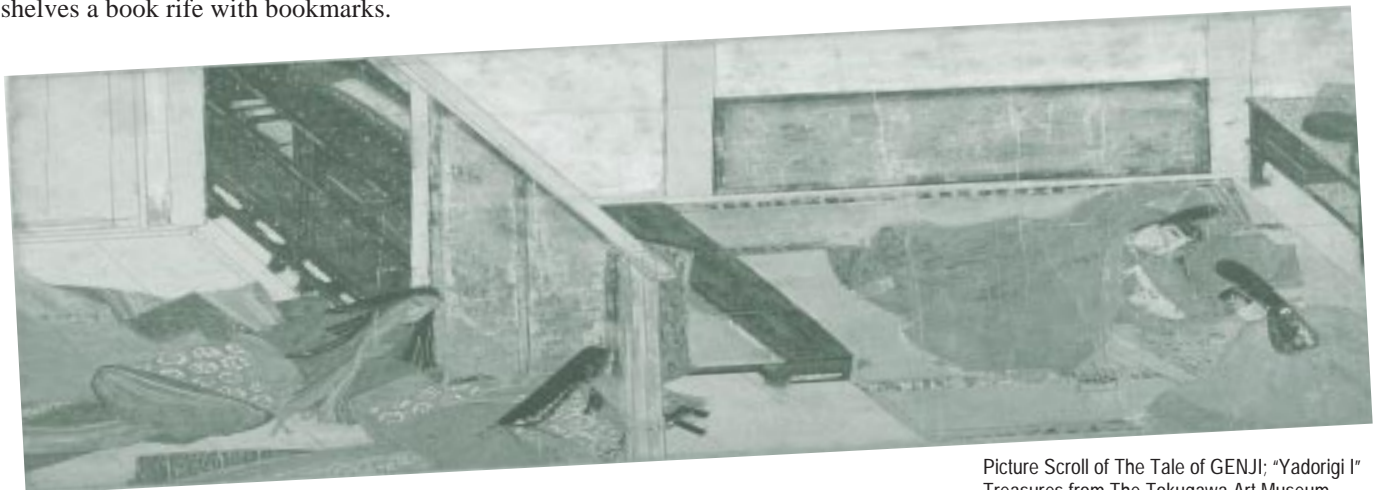
"I am creating a database of a novel by (Nobel laureate) Yasunari Kawabata as it has been rumored that some works attributed to him were actually written by his students. I also work in a number of other areas, including pornographic novels, an area where I am certain famous authors sometimes wrote under a penname. Some publications of this genre are rumored to have been written by famed novelist Yukio Mishima. Of course, the main focus of our work is still in supporting serious researchers analyzing academic, philosophical and religious works."

Someday we might see a scholar of literature walking up that hill in Minami-Azabu carrying a pornographic novel (for analysis).

A new trend: mining for data
Beer and diapers
lead the way

"In America, it seems like if something looks like the next 'big thing,' everyone jumps on the bandwagon. Take data mining, for example. It became incredibly popular within a year of its inclusion as a division of the Association for Computing Machinery."

In contrast to the stately, traditional atmosphere I expected from The University of Tokyo, historical Hongo Campus, Dr. Morishita's office was quite modern, complete with freshly painted white walls. He explained that he had just moved there in September before pull-



Picture Scroll of The Tale of GENJI; "Yadorigi I" Treasures from The Tokugawa Art Museum

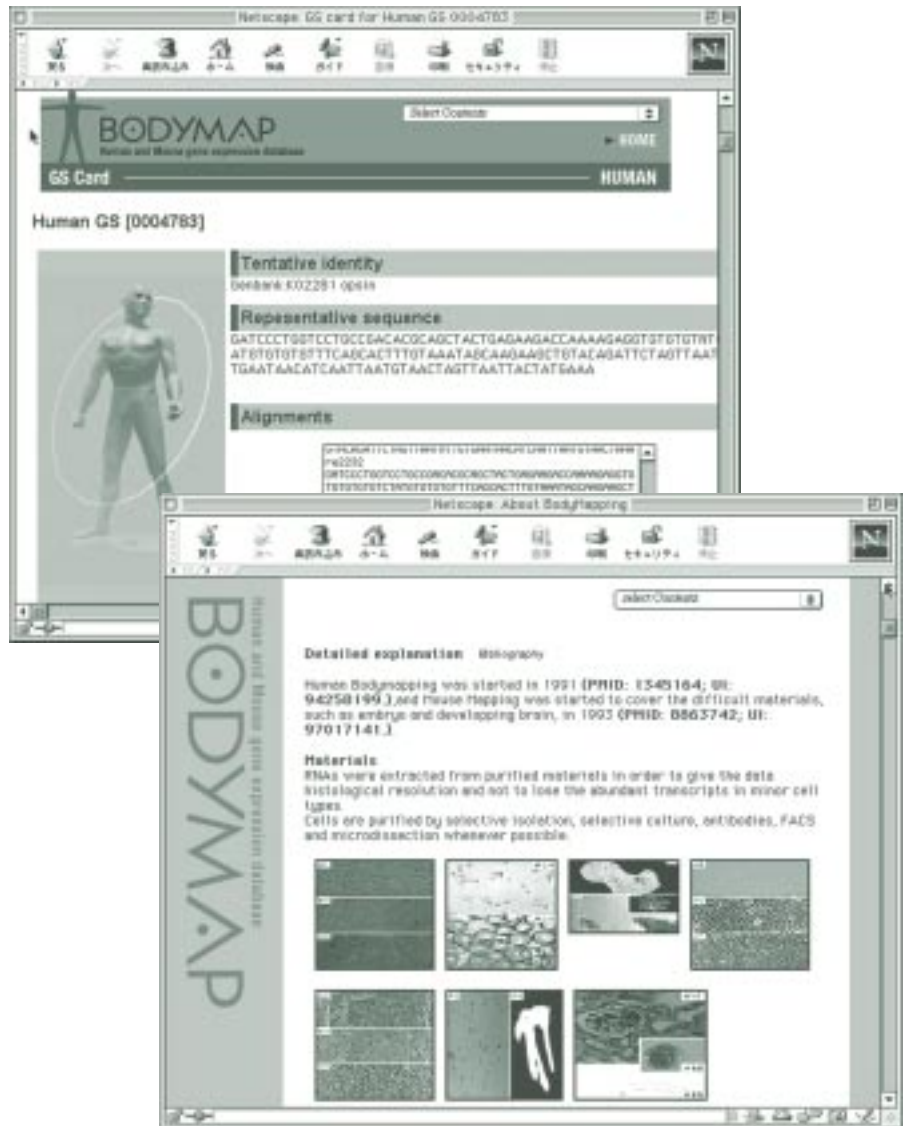
ing out a well-used personal computer. He used it to show me a report summarizing the data mining research he had conducted overseas during the past year. "Let's start by talking about IBM's Almaden Research Center, where its data mining group leader Rakesh Agrawal first published his research findings in 1993. That's when it really all began."

Mr. Agrawal developed software capable of quickly analyzing huge volumes of sales data, which led to a revelation that consumers tend to buy paper diapers and canned beer together. This took everyone by surprise, and data mining quickly became a hot topic of discussion. Basically, data mining is software technology that enables researchers to uncover hidden patterns and phenomena through statistical analysis of immense volumes of data.

"Almaden Research Center retained their lead in this field and went on to develop many powerful data mining technologies. Many strong contributors in the field once worked at this center. Ashish Gupta, for example, established a venture company named Junglee, which he later sold to Amazon online retailer."

Dr. Morishita became friends with Mr. Gupta, who is only in his early 30s, when they were both at Stanford University. "Microsoft Corporation then began to invest in data mining and attracted many key researchers, including Fields and Turing Prize winners. An established leader in database technology, Stanford University has also been active in data mining research under the pioneering leadership of Dr. Ullman. Within his group, a student in his 20s, Sergey Brin, developed the research engine 'google,' which became tremendously popular." Is data mining now a major trend? Would previous analysis techniques have been able to identify the link between diapers and canned beer?

"It's not impossible to perform such analyses (using older methods) if you put in a tremendous amount of effort. Compared with the past, however, there are now huge volumes of available data. We now have several hundred thousand products and product combinations and tens of millions of pieces of sales data.



Dr. Kosaku Okubo of Osaka University's Institute for Molecular and Cellular Biology has posted his research results on a website to help support research on gene function. Dr. Morishita supports his work by providing data mining technology, including a search engine. They are cooperating to try and establish the BODYMAP website as a "Mecca" for information on gene function. <http://bodymap.ims.u-tokyo.ac.jp/>

If we are to uncover significant relationships within this group, it is impractical to try and achieve this using earlier methods."

How does data mining handle these huge volumes of data?

"How about if I explain this using familiar objects as analogies? If I'm the CPU, this desk is the main memory, and a library located far away is the secondary hard drive, making a trip to the library every time I needed a book would waste my time. If you need a dictionary frequently, you would keep it on your

desk, which has limited space. To limit trips to the library, you need to decide the most important books to keep close at hand. Data mining is used to determine the most useful books to keep on your desk. Importing data to the CPU from the main memory is 10,000 to 100,000 times faster than from the secondary hard drive. So there's a major difference between systems that utilize this concept and those that don't." Since data mining enables us to process immense volumes of data, we can not only identify products typically pur-



Dr. Shinichi Morishita is Associate Professor in the Department of Information Science, University of Tokyo. He also holds instructorships in Frontier Information Complexity Science and Engineering at the Graduate School of Frontier Science and at The Institute of Medical Science, both at The University of Tokyo. Born in 1960 in Tokyo, he previously worked at IBM Japan's Tokyo Research Laboratory and later at The Institute of Medical Science at The University of Tokyo before assuming his current responsibilities.

chased together, powerful statistical methods can also be applied to other fields. Consider, for example, the standard search engine. It navigates through vast amounts of data, finding key words and even sorting them according to related key words. This is an extension of the same kind of analysis used to identify the sales connection between beer and diapers.

Data mining can also be used to predict changes in the price of gold, identify companies heading toward bankruptcy and more. If you analyze personal savings account type, blood type and other factors, you can identify individuals with a tendency to default on loan payments.

"In the past, human beings selected the items they felt had a potential link that was worth investigating. But since data mining can analyze tremendous volumes of data, we can now input unrelated data and test it to see if, in fact, a relationship does exist."

Why not apply data mining to horse racing? You could analyze blood characteristics, genetic heritage, track record, sweat type and more.

"Sweat? (ha-ha). This may be possible, but it could be tough to reveal hidden trends in a field where so many others have already performed extensive analyses. In my opinion, data mining is most effective in areas where the data doesn't appear to have been extensively analyzed, and where there exists a large volume of data. From this point of view,

cluster gene research is ideal for data mining."

Deciphering gene distributions What will we find next?

Homo Sapiens, as a global species, is currently engaged in an international effort to decipher all three billion combinations of human genes by the year 2003. Different countries are handling specific groups of genes, and if the project is completed it will be a major milestone in the life sciences.

Still, even if we can decipher the genetic alphabet, we won't understand the roles played by each individual gene. Being able to read the English alphabet doesn't mean you can understand English.

"Osaka University's Dr. Kosaku Okubo has been collecting data for nine years at its Institute for Molecular and Cellular Biology. He thinks it is important to determine each letter of the genetic alphabet, but even more important to understand which cells contain which genes, and what their function is in the human body. He has tremendous foresight and we are working to support him with data analysis."

I've heard that with advanced genetic technology we can determine whether specific genes exist within a particular cell. What would be the purpose of analyzing such information?

"I think we've named about 15 percent of human genes. I've heard that 60,000 kinds of cDNA have been categorized into the genetic alphabet, but for 80 percent of we still don't know what role play. It's natural to want to categorize them by similar characteristics, and this is achieved using a technology called clustering."

Genes are not simply categorized by genetic letter. Various research projects have determined the distribution of certain genes in rodents as a percentage of total genes within vital organs such as the brain, heart and lungs. As the rodents developed from fetus to adult, the percentage of certain genes present in these organs was observed for changes.

"With 30 different types of data, you could conceptually apply a 30-D vector

space for categorizing each gene. Since it's hard to visualize a 30-D vector space, however, we clustered them into some groups."

Data mining is what first created the possibility of performing a 30-D data analysis.

"When we looked at the results of clustering, we saw areas where totally different genetic letters fell into the same cluster. This suggested to us that dissimilar genes team up to perform specific functions."

To fully understand a gene's function, we must also know the role of its group. Data mining is ideal for investigating the cooperative function of genes within a group.

"Right. On the other hand, most biologists still limit the number of gene types they analyze (as opposed to looking at a grouping). Beginning year 2000, I hope to display my research on a website so that others will be able to utilize the results of my cluster analysis. I also hope researchers will start thinking about experiments to reveal what would happen if you destroyed one gene within a group."

There is still a fundamental question that remains unanswered. Did anyone determine why shoppers commonly bought canned beer along with paper diapers?

"Some have explained it this way: the mother is not at home on the weekend, and the father is taking care of the kids, and he tends to buy beer and diapers the same trip. Data mining doesn't explain causative factors yet, but it would be quite interesting if it could. Regarding the global human gene project, I think it would be wonderful if we could cooperate together and expand our understanding of a common truth."



Researchers analyzing *The Tale of Genji* as well as those trying to crack the genetic code hope to share the results of their research as soon as possible. Meticulous databases, as in *The Tale of Genji*, and advanced data mining technologies are both powerful tools that can expand our knowledge. But humanity is now being tested by these new powers. How will large these vast databases be utilized? That is our newest challenge.