For Today’s Graduate, Just One Word: Statistics

By [STEVE LOHR](http://topics.nytimes.com/top/reference/timestopics/people/l/steve_lohr/index.html?inline=nyt-per) Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

“People think of field archaeology as Indiana Jones, but much of what you really do is data analysis,” she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by 2012, according to a projection by IDC, a research firm.

Yet data is merely the raw material of knowledge. “We’re rapidly entering a world where everything can be monitored and measured,” said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology’s Center for Digital Business. “But the big problem is going to be the ability of humans to use, analyze and make sense of the data.”

The new breed of statisticians tackle that problem. They use powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data. The applications are as diverse as improving Internet search and online advertising, culling gene sequencing information for cancer research and analyzing sensor and location data to optimize the handling of food shipments.

Even the recently ended Netflix contest, which offered $1 million to anyone who could significantly improve the company’s movie recommendation system, was a battle waged with the weapons of modern statistics.

Though at the fore, statisticians are only a small part of an army of experts using modern statistical techniques for data analysis. Computing and numerical skills, experts say, matter far more than degrees. So the new data sleuths come from backgrounds like economics, computer science and mathematics.

They are certainly welcomed in the White House these days. “Robust, unbiased data are the first step toward addressing our long-term economic needs and key policy priorities,” Peter R. Orszag, director of the Office of Management and Budget, declared in a speech in May. Later that day, Mr. Orszag confessed in a blog entry that his talk on the importance of statistics was a subject “near to my (admittedly wonkish) heart.”

I.B.M., seeing an opportunity in data-hunting services, created a Business Analytics and Optimization Services group in April. The unit will tap the expertise of the more than 200 mathematicians, statisticians and other data analysts in its research labs — but that number is not enough. I.B.M. plans to retrain or hire 4,000 more analysts across the company.

In another sign of the growing interest in the field, an estimated 6,400 people are attending the statistics profession’s annual conference in Washington this week, up from around 5,400 in recent years, according to the American Statistical Association. The attendees, men and women, young and graying, looked much like any other crowd of tourists in the nation’s capital. But their rapt exchanges were filled with talk of randomization, parameters, regressions and data clusters. The data surge is elevating a profession that traditionally tackled less visible and less lucrative work, like figuring out life expectancy rates for insurance companies.

Ms. Grimes, 32, got her doctorate in statistics from Stanford in 2003 and joined Google later that year. She is now one of many statisticians in a group of 250 data analysts. She uses statistical modeling to help improve the company’s search technology.

For example, Ms. Grimes worked on an algorithm to fine-tune Google’s crawler software, which roams the Web to constantly update its search index. The model increased the chances that the crawler would scan frequently updated Web pages and make fewer trips to more static ones.

The goal, Ms. Grimes explained, is to make tiny gains in the efficiency of computer and network use. “Even an improvement of a percent or two can be huge, when you do things over the millions and billions of times we do things at Google,” she said.

It is the size of the data sets on the Web that opens new worlds of discovery. Traditionally, social sciences tracked people’s behavior by interviewing or surveying them. “But the Web provides this amazing resource for observing how millions of people interact,” said Jon Kleinberg, a computer scientist and social networking researcher at Cornell.

For example, in research just published, Mr. Kleinberg and two colleagues followed the flow of ideas across cyberspace. They tracked 1.6 million news sites and blogs during the 2008 presidential campaign, using algorithms that scanned for phrases associated with news topics like “lipstick on a pig.”

The Cornell researchers found that, generally, the traditional media leads and the blogs follow, typically by 2.5 hours. But a handful of blogs were quickest to quotes that later gained wide attention.

The rich lode of Web data, experts warn, has its perils. Its sheer volume can easily overwhelm statistical models. Statisticians also caution that strong correlations of data do not necessarily prove a cause-and-effect link.

For example, in the late 1940s, before there was a polio vaccine, public health experts in America noted that polio cases increased in step with the consumption of ice cream and soft drinks, according to David Alan Grier, a historian and statistician at George Washington University. Eliminating such treats was even recommended as part of an anti-polio diet. It turned out that polio outbreaks were most common in the hot months of summer, when people naturally ate more ice cream, showing only an association, Mr. Grier said.

If the data explosion magnifies longstanding issues in statistics, it also opens up new frontiers.

“The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd,” said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. “And that makes it easier for humans to do what they are good at — explain those anomalies.”

*Andrea Fuller contributed reporting.*

=============================================================

Carrie Grimes 在哈佛主修人類學和考古學，曾在宏都拉斯的雨林中經歷過一場冒險，她透過標記古文物的出土位置來研究馬雅人的居住地。「在大眾眼中的考古學大多來自電影中印第安那瓊斯冒險犯難的場景，可是實際上的考古學大多都在做資料分析。」她這麼描述著，而她也沉浸在她所謂的全是電腦與數學的領域中。如今Grimes女士從事另外一種挖掘的工作。她現在是Google的統計分析師，成天面對成堆的資料，並運用各種統計分析方法去找到能改善公司的搜尋引擎的方法。

**未來十年最迷人的工作將會是統計學家**

一般人對網路世代的印象大多是成天與電腦為伍、足不出戶的阿宅，而Grimes女士正是屬於網路世代的統計學家，但是她和其他同屬網路世代的優秀同伴們一樣，致力扭轉大眾對網路世代的刻板印象，而大家也發現他們的能力不僅被這個社會密切的需要，而且也越來越搶手。

「我一直說在未來的十年裡，最迷人的工作將會是統計學家，這並不是開玩笑的。」Google的首席經濟學家Hal Varian這麼強調著。

拜近年來的資訊爆炸所賜，它造就了這社會對統計學家需求的成長。對一位剛畢業的統計博士來說，在美國頂尖的公司第一年的年薪就可以有125,000美金（約四百萬台幣）在未來的資訊和網路科技的領域裡，還有許多有潛力的資料等著人們去研究與分析，舉凡感應訊號、監視錄影帶、社會網路脈絡、公眾資料記錄等等，都在其範疇內。根據IDC市場研究公司的預測，這股數位資料的浪潮在未來只會加速的洶湧，在2012年更會達到現今五倍的水準。

**分析資料是統計的核心價值**

然而成堆的資料並不等於有用的知識。任職於麻省理工學院電子商務中心的經濟學家兼董事Erik Brynjolfsson指出：「在現今的世界，資料的取得是可以很迅速且容易的，幾乎所有的事物都可以被監測與量化成所需的資料。所以目前對人類來說，最大的問題是該如何去分析這些資料，並從中整理出我們所關心的資訊。」

因應這股浪潮，新興的統計學家們也隨之興起。他們利用高性能的電腦和精密的數學模型去處理成堆的資料，試圖從中尋找到有意義的樣本和珍貴的資訊。關於這方面的應用可以說是五花八門，從搜尋引擎最佳化、有效的線上廣告模式、基因序列中癌細胞的篩檢到糧食配送的最佳化等等，都可以是統計應用的範圍。

全美最大的DVD租借網站Netfilx公司在前陣子舉行了一場競賽，只要能夠提出有效改善他們公司的電影推薦系統的方法，就能抱走一百萬美金的獎賞，這場競賽可以說是用現代統計方法當作武器來廝殺的戰場。

**更多學者專家紛紛投入統計的懷抱**

在這股數位資料的浪頭上，正統的統計學家僅僅只是這些先行者的一小部分而已。專家說計算與數值分析的能力遠比學位重要多了，有更多來自不同背景的學者專家們，有經濟學家、電腦科學家、數學家等等，都紛紛去擁抱最新的統計資料分析技巧。

這些學者專家們在現在的社會中，不管到哪個領域都是受歡迎的，就連白宮也不例外。美國行政管理和預算局的一位經理Peter R. Orszag在今年五月的一場演講中指出：「穩健與不偏的資訊將會是我們在制定長期的經濟需求政策中最優先考慮的關鍵點。」而日後他也在自己的部落格寫下「統計是多麼的貼近我的心」的字句，也再次闡述了關於統計的重要性。

I.B.M.也看見了隱藏在收集資料的服務裡的龐大商機，在今年四月成立了商業分析與最佳化服務組織。在組織的研究室中擁有超過兩百位的數學家、統計學家及其他資料分析的專家，然而這個數目遠遠低於I.B.M.的需求，他們計畫為整個公司重新培訓或雇用超過4,000位的分析專家。

統計領域的興起早有跡象可循，根據美國統計學會的資料來看，這個星期在華盛頓舉行的統計專業人才年度會議就有約6,400人出席，比起往年約5,400人左右的出席人數多出了整整一千人。而這些出席者不管男女老幼，看起來就跟首都裡其他的觀光客沒什麼兩樣，但是從他們全神貫注的討論對隨機化、參數、迴歸及資料叢集的神情來看，這股數位資料的浪潮已經讓那些專業人士愈來愈重視在傳統上屬於能見度較低且無法獲利的工作上，保險公司越來越重視壽險就是一個很好的例子。

**網路的出現讓統計有更大的揮灑空間**

讓我們把焦點回到一開始的Grimes女士身上，現年32歲的她，在2003年從史丹佛大學拿到統計學博士的學位，在隔年就加入了Google。現在隸屬於一個250人的統計分析團隊裡，並運用各種的統計模型去改善公司的搜尋引擎。

舉例來說，Grimes女士曾參與了搜尋引擎機器人的演算法最佳化工作，那個機器人會在網際網路中漫遊，並定時更新搜尋引擎的索引。Grimes的工作就是找到一個適當的模型，讓機器人拜訪經常更新的頁面次數會比那些靜態內容的頁面的次數來得多。

Grimes解釋這個工作的最終目的就是希望得到在運算或者網路上的效能改善，即使只有2%的改善就整體而言是很巨大的，畢竟Google面對的是成千上萬筆資料，累加下來的績效是很可觀的。

在網路上那些巨大的資料集的也讓資料探勘進入了新世界。傳統的社會科學家要研究人類行為的時候，得透過實際面談來調查受訪者。一位康乃爾大學的電腦科學家和社會網路研究者Jon Kleinberg就說：「網際網路提供了驚人的資源，讓我得以觀察上百萬的人們是如何互動的。」

在Kleinberg剛發表的研究報告中指出，他和另外兩位同事依照網路上的流程，使用搜尋與特定字彙相關的新聞標題的演算法，在2008年的美國總統大選期間，追蹤了一百六十萬筆的新聞網站和部落格的頁面內容。他們發現平均來說傳統傳媒的資訊會領先那些部落客2.5個小時，但是因為部落格可以被到處引用的特性，進而讓資訊更快速的傳播並得到世人的注意。

**正確的統計分析會讓事半功倍**

雖然網路上蘊藏著豐富的資料，但是專家也警告那些資料是有風險存在的，那些資料量可能會輕易得讓既有的統計模型不堪使用。統計學家也提出警告，一些看起來有強烈關連性的資料，實際上來說並不是真的有因果關係。

舉例來說，在40年代小兒麻痺疫苗出現前，一位美國喬治華盛頓大學的統計學家兼公共衛生專家David Alan Grier宣稱：「小兒麻痺患者的增加是因為飲料及冰淇淋的消費量的上升。」減少飲料及冰淇淋的攝取甚至成為抗小兒麻痺療程的一部份。而他僅僅是依據小兒麻痺患者的激增多出現在炎熱的夏季，而這個季節的人們會食用更多的冰淇淋，就得出這項結論。

資訊爆炸不僅會擴展一些統計上陳年的議題，也會開啟更多的新領域。

一位在I.B.M.從事醫藥資料探勘的研究人員Daniel Gruhl就說：「現在最關鍵的事就是讓電腦去做他擅長的事，也就是處理並收集成堆的資料。對人類來說只需要專注在如何去解釋那些異常現象就好。」

*Translate by Keith*

原文出處：<http://www.nytimes.com/2009/08/06/technology/06stats.html>