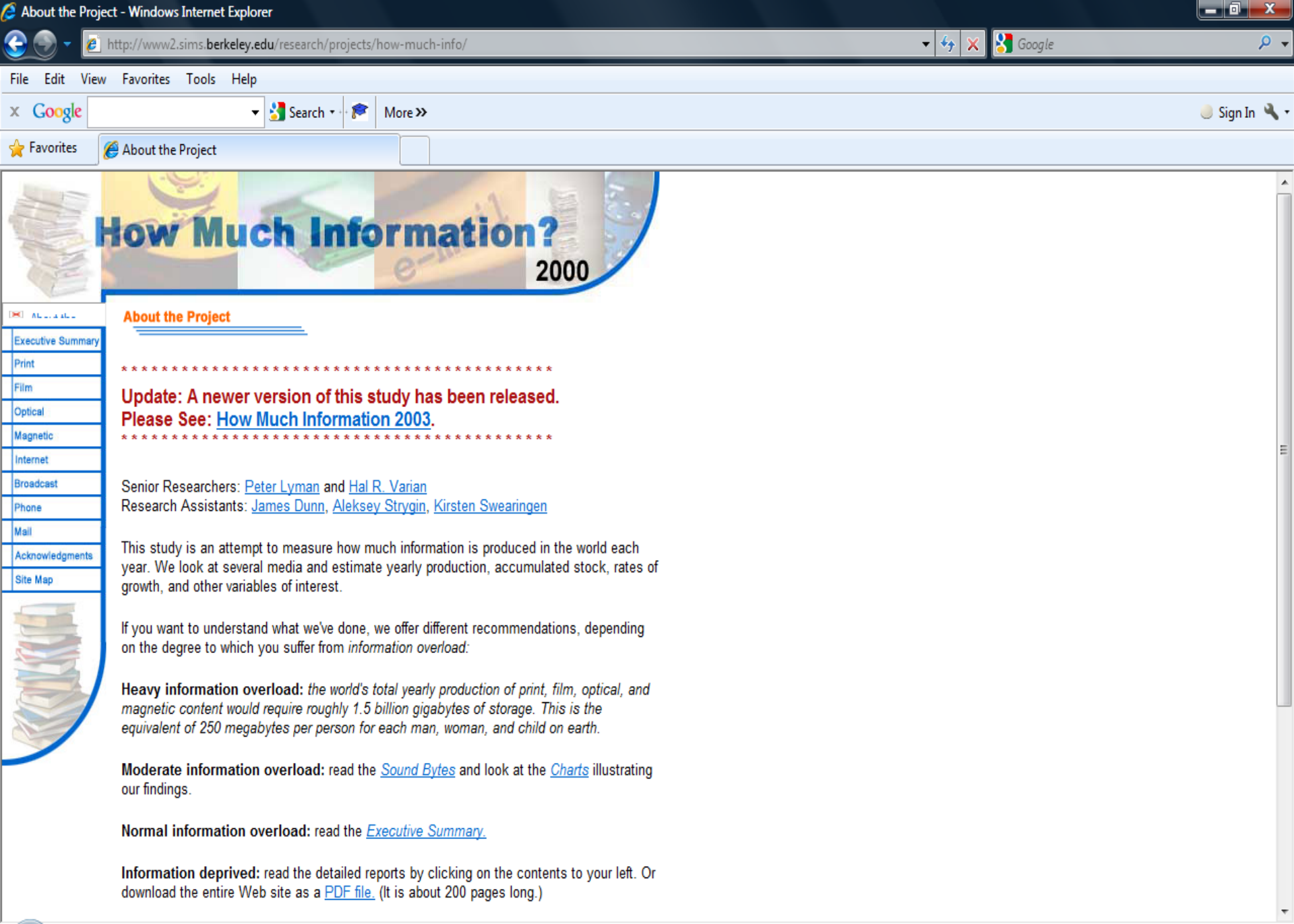
Veri Madenciliği

Umut Al

umutal@hacettepe.edu.tr

Veri Madenciliğinin Ortaya Çıkışı

- “data dredging”
- “data fishing”
- Farklı alanların etkisi
  - İstatistik
  - Bilgisayar bilimleri
  - Otomasyon
- Geleneksel tekniklerin yetersizliği
- Veri artış hızı


About the Project - Windows Internet Explorer
http://www2.sims.berkeley.edu/research/projects/how-much-info/
File Edit View Favorites Tools Help
Search
More
Sign In
Favorites
About the Project
How Much Information?
2000
About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map
Update: A newer version of this study has been released.
Please See: How Much Information 2003.
Senior Researchers: Peter Lyman and Hal R. Varian
Research Assistants: James Dunn, Aleksey Strygin, Kirsten Swearingen
This study is an attempt to measure how much information is produced in the world each year. We look at several media and estimate yearly production, accumulated stock, rates of growth, and other variables of interest.
If you want to understand what we've done, we offer different recommendations, depending on the degree to which you suffer from information overload:
Heavy information overload: the world's total yearly production of print, film, optical, and magnetic content would require roughly 1.5 billion gigabytes of storage. This is the equivalent of 250 megabytes per person for each man, woman, and child on earth.
Moderate information overload: read the Sound Bytes and look at the Charts illustrating our findings.
Normal information overload: read the Executive Summary.
Information deprived: read the detailed reports by clicking on the contents to your left. Or download the entire Web site as a PDF file. (It is about 200 pages long.)

How Much Information? 2009
Report on American Consumers

Roger E. Bohn
James E. Short

Global Information Industry Center
University of California, San Diego

Date of Publication: December 2009

Last Update: January 2010

Counting Very Large Numbers

Byte (B)	=	1 byte	=	1	=	One character of text
Kilobyte (KB)	=	10^3 bytes	=	1,000	=	One page of text
Megabyte (MB)	=	10^6 bytes	=	1,000,000	=	One small photo
Gigabyte (GB)	=	10^9 bytes	=	1,000,000,000	=	One hour of High-Definition video, recorded on a digital video camera at its highest quality setting, is approximately 7 Gigabytes
Terabyte (TB)	=	10^{12} bytes	=	1,000,000,000,000	=	The largest current hard drive
Petabyte (PB)	=	10^{15} bytes	=	1,000,000,000,000,000	=	AT&T currently carries about 18.7 Petabytes of data traffic on an average business day
Exabyte (EB)	=	10^{18} bytes	=	1,000,000,000,000,000,000	=	Approximately all of the hard drives in home computers in Minnesota, which has a population of 5.1M
Zettabyte (ZB)	=	10^{21} bytes	=	1,000,000,000,000,000,000,000		

How Much Information? 2009 Report on American Consumers

Figure 1: Information Flows In A Home

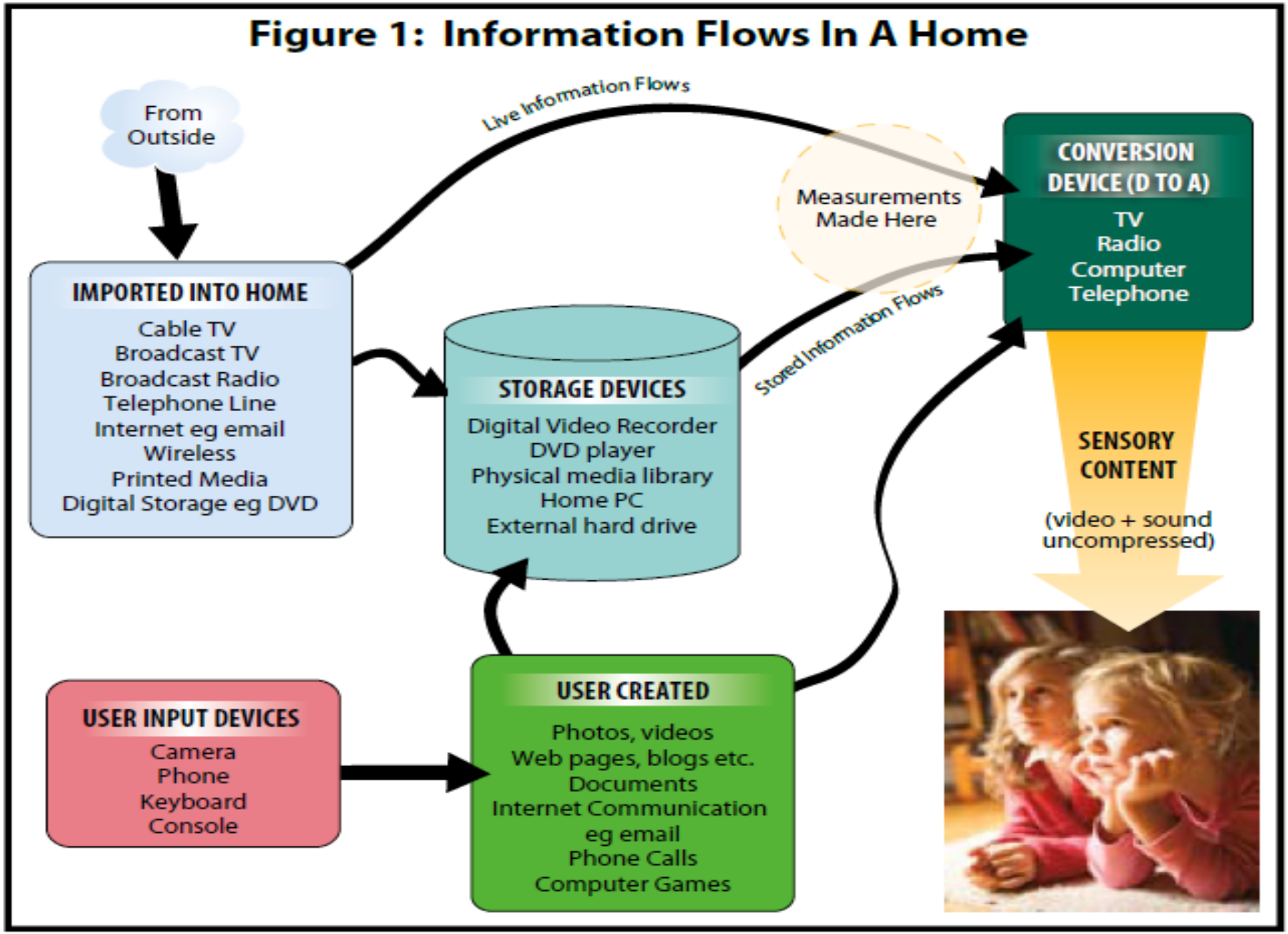


Figure 10: Shares of Information in Different Formats

Per Average American, Per Day

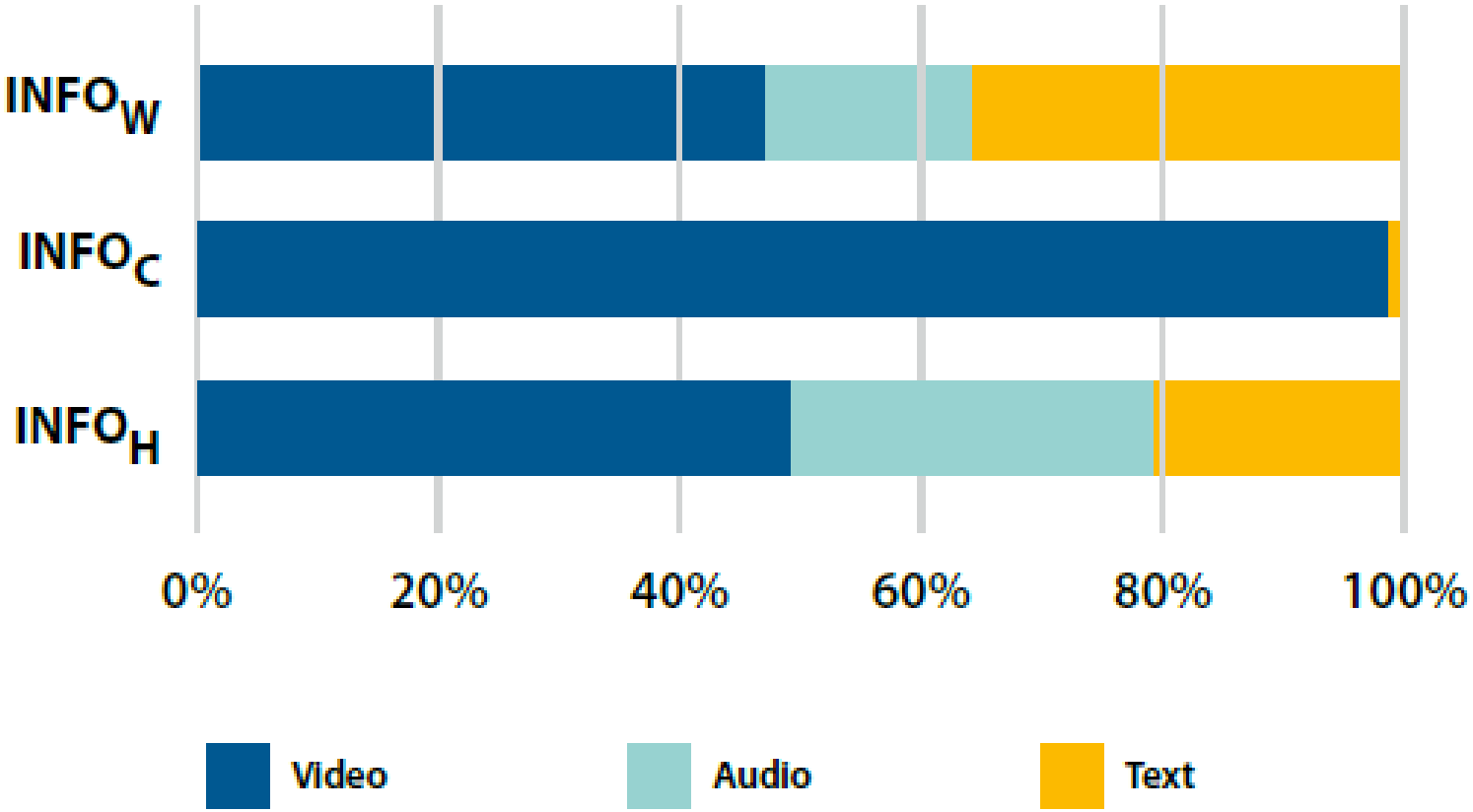


Table 9: Summary of Information for Major Groups

ACTIVITY	Total per year (entire population)			% of Total			Per average American per day		
	$INFO_H$ in hours	$INFO_C$ in bytes	$INFO_W$ in words	% Hrs	% Bytes	% words	Hours	Giga-bytes	Words
All TV	5.30E+11	1.27E+21	4.86E+15	41.62%	34.77%	44.85%	4.91	11.75	45,100
Radio	2.39E+11	1.10E+19	1.15E+15	18.79%	0.30%	10.59%	2.22	0.10	10,645
Phone	7.89E+10	1.36E+18	5.68E+14	6.20%	0.04%	5.24%	0.73	0.01	5,269
Print	6.49E+10	6.72E+17	9.34E+14	5.09%	0.02%	8.61%	0.60	0.01	8,659
Computer	2.08E+11	8.69E+18	2.93E+15	16.35%	0.24%	26.97%	1.93	0.08	27,122
Computer Games	1.00E+11	1.99E+21	2.65E+14	7.86%	54.62%	2.44%	0.93	18.46	2,459
Movies	3.24E+09	3.56E+20	2.14E+13	0.25%	9.78%	0.20%	0.03	3.30	198
Recorded music	4.88E+10	8.85E+18	1.20E+14	3.83%	0.24%	1.11%	0.45	0.08	1,112
TOTALS	***1.27E+12***	***3.64E+21***	***1.08E+16***	***100.00%***	***100.00%***	***100.00%***	***11.80***	***33.80***	***100,564***

5.3E+11 means $5.3 \times 10^{11} = 530,000,000,000$

Dünyayı artık veri analizi yönetiyor

8 Mayıs 2012 | A A

İstanbul'un trafik sorunu nasıl çözülecek? ABD'de kargo kamyonları neden hiç sola sapmaz? Hangi ülkede süpermarketler indirim yapılacak ürünleri Facebook'taki takipçilerine sorar? Peki, dünya 2030'da nasıl bir yer olacak?

İrlanda'nın başkenti Dublin'de, geçtiğimiz günlerde, Teradata Universe 2012 konferansı için bir araya gelen 982 kişi bu soruların cevaplarını ve bunu mümkün kılacak teknolojiyi tartıştı: Veri analizi.

Konferanstan bahsetmeden önce birkaç adım geri gidelim. 'Veri analizi' dediğimiz şey yeni bir kavram değil. Nüfus sayımlarından telefon rehberlerine, hayatımız verilere dayanarak oluşturulan yapılar içinde geçiyor.

Ancak teknolojinin gelişmesi ve ucuzlamasıyla bu geleneksel verilerin yanına çok daha kapsamlı ve gelişmiş veri tabanları eklenmeye başladı. Google aramaları, internet üzerinden yaptığımız alışverişler, yazışmalar, cep telefonları, sosyal medya şimdiden, hayatımızla ilgili en küçük detayları bile gönüllü olarak yüklediğimiz birer veri tabanı haline geldi. Örneğin hangi restoranda, hangi tarihte, hangi arkadaşlarımızla yemek yediğimiz, Facebook'ta fotoğraflarıyla belgelenebiliyor.

KLOZETTEKİ DOKTOR

Buna bir de çok yakın bir gelecekte arabamızın koltuğuna, hatta çorabımızın içine yerleştirilecek sensorlar sayesinde elde edilecek verileri eklevin. Diyelim ki banyomuzdaki

DİĞER HABERLER

Yunanistan'da deneme sırası solda

Yunanistan'da uluslararası kurtarma planına destek veren partilerin koalisyon hükümeti kurmayı ...

Sorunlar

- Farklı veri kanalları
- Veri çeşitliliği
- Veri kalitesi
- Veri temizliği
- Anlık veri
- Birbiri ile bağlantılı veriler (etkileşimlilik)
- Veri miktarı * analist sayısı

http://www.ctg.albany.edu/publications/reports/data_quality_tools/data_quality_tools.pdf - Windows Internet Explorer

http://www.ctg.albany.edu/publications/reports/data_quality_tools/data_quality_tools.pdf

File Edit Go To Favorites Help

Google Search More

Sign In

Favorites http://www.ctg.albany.edu/publications/reports/...

1 / 12 100% Collaborate Sign Find

Data Quality Tools for Data Warehousing – A Small Sample Survey

Using Information in Government Program

Quantitative Data Cleaning for Large Databases

Joseph M. Hellerstein*
EECS Computer Science Division
UC Berkeley
http://db.cs.berkeley.edu/jmh

February 27, 2008

1 Introduction

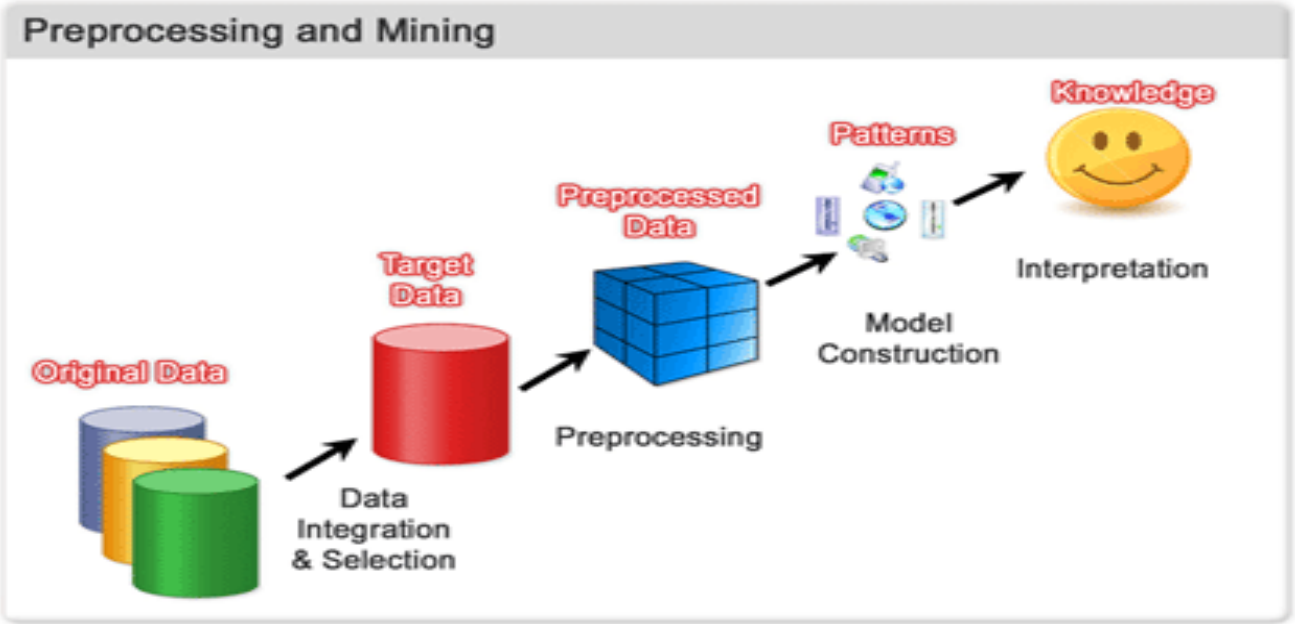
Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission. Data analysis typically drives decision-making processes and efficiency optimizations, and in an increasing number of settings is the *raison d'etre* of entire agencies or firms.

Despite the importance of data collection and analysis, data *quality* remains a pervasive and thorny problem in almost every large organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. As a result, there has been a variety of research over the last decades on various aspects of *data cleaning*: computational procedures to automatically or semi-automatically identify – and, when possible, correct – errors in large data sets.

In this report, we survey data cleaning methods that focus on errors in *quantitative* attributes of large databases, though we also provide references to data cleaning methods for

Tahmin ve Tanımlama Fonksiyonu

- Eldeki verilerle gelecekteki durumu yordama
- Yoruma açık olmayan kalabalık veri setlerinden yorumlanabilir örüntüleri bulma



Quakefinder: A Scalable Data Mining System for Detecting Earthquakes from Space

Paul Stolorz and **Christopher Dean**
Jet Propulsion Laboratory
California Institute of Technology
(pauls,ctdean)@aig.jpl.nasa.gov

Abstract

We present an application of novel massively parallel datamining techniques to highly precise inference of important physical processes from remote sensing imagery. Specifically, we have developed and applied a system, Quakefinder, that automatically detects and measures tectonic activity in the earth's crust by examination of satellite data. We [illegible] Quakefinder by enabling the automatic detection and measurement of earthquake faults from satellite imagery.

The system, Quakefinder, is applied here to the analysis of data collected by the French SPOT satellite. SPOT is a push-broom detector that collects panchromatic data at 10 meter resolution from a satellite in sun-synchronous orbit around the earth. In our appli-

Kullanım Alanları

- Bilgi erişim
- Pazarlama
- Müşteri ilişkileri yönetimi
- Haberleşme
- Sağlık
- E-ticaret
- …

Veri Madenciliğinde Kullanılan Yöntemler

- Sınıflandırma
- Kümeleme
- Ayırma analizi
- Regresyon
- Varyans analizi
- Yapay sinir ağları
- Karar ağaçları
- …

Avrupa Birliği Ülkeleri ve Türkiye'nin Yayın ve Atıf Performansı

Umut Al*

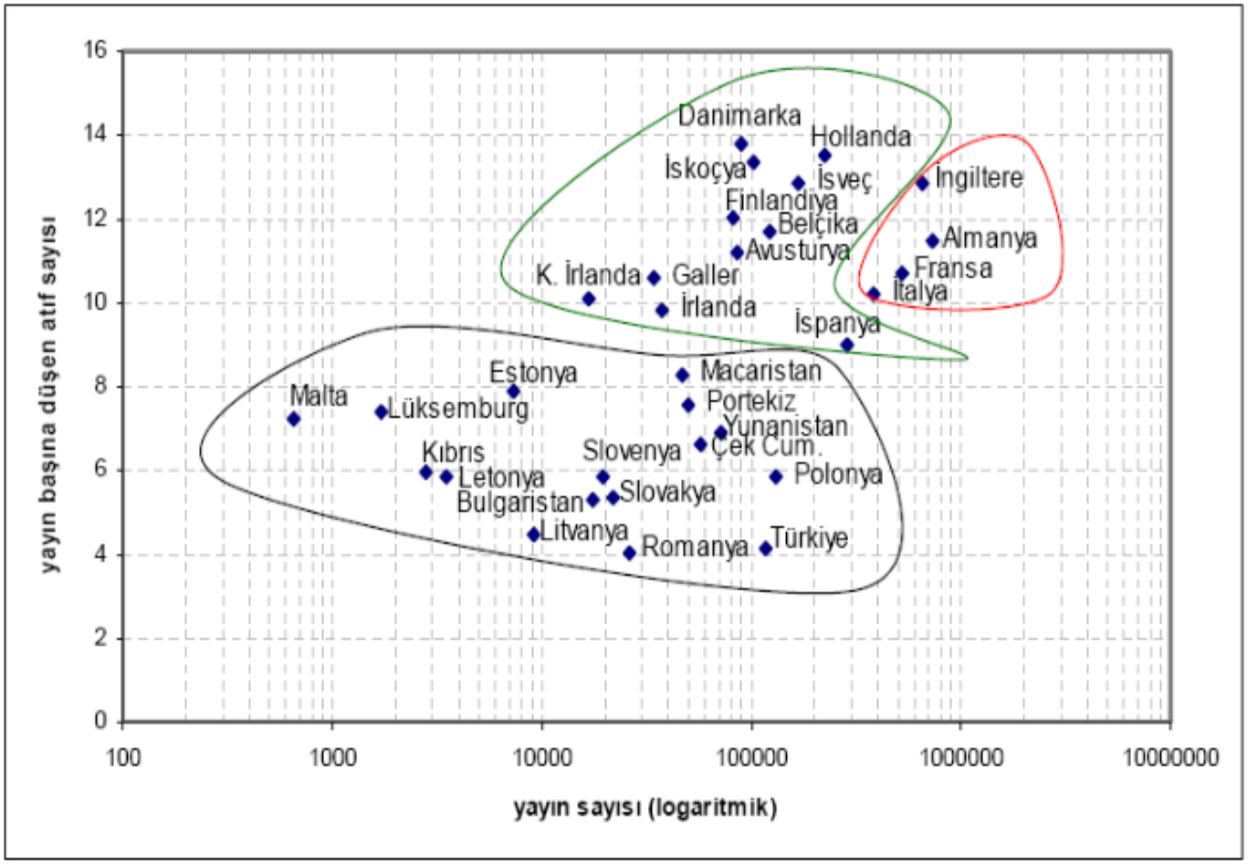
Özet

Bilimsel yayınların dinamiğine ilişkin araştırmalar bilim dünyasının ilgisini çekmektedir. Bu tip çalışmaların sayısı son yıllarda giderek artmaktadır. Bilimsel yayınların etkinliği genellikle bibliyometrik çalışmalarla ortaya konulmaktadır. Gerçekleştirilen bibliyometrik araştırmalarda veri kaynağı olarak atıf dizinlerinden yararlanılmaktadır. Çalışmanın verileri Essential Science Indicators adlı kaynaktan elde edilmiştir. Essential Science Indicators farklı alanlarda ülkelere yönelik yayın ve atıf verilerini içermektedir. Bu çalışmada Türkiye'nin göstermiş olduğu yayın ve atıf performansı değerlendirmekte, Avrupa Birliği ülkeleri ile yapılan çeşitli karşılaştırmalara yer verilmektedir. Araştırmada yayın ve atıf performansı birbirine benzer ülkelerin hangileri olduğunu saptayabilmek için kümeleme analizinden yararlanılmıştır. Bulgular İngiltere, Almanya, Fransa ve İtalya'nın Avrupa Birliği ülkeleri arasında en üst düzey yayın ve atıf performansına sahip olduğunu göstermektedir. Türkiye ise araştırma kapsamındaki tüm alanlarda atıf performansı düşük ülkelerin bulunduğu grupta yer almaktadır.

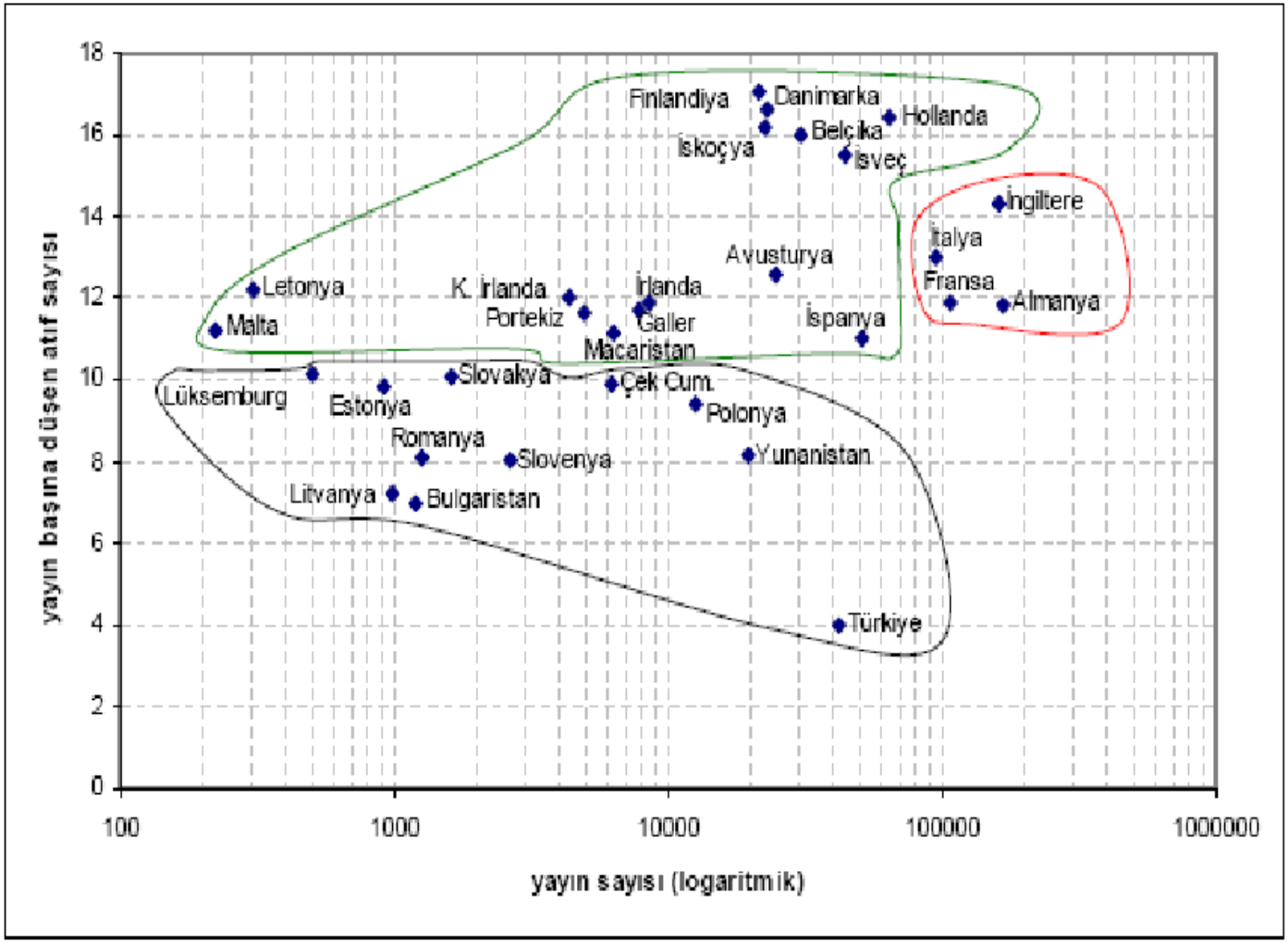
Anahtar Kelimeler: Avrupa Birliği ülkeleri, atıf etkisi, bibliyometri, yayın performansı.

Tablo 1. Avrupa Birliği ülkelerinin yayın ve atıf sayıları

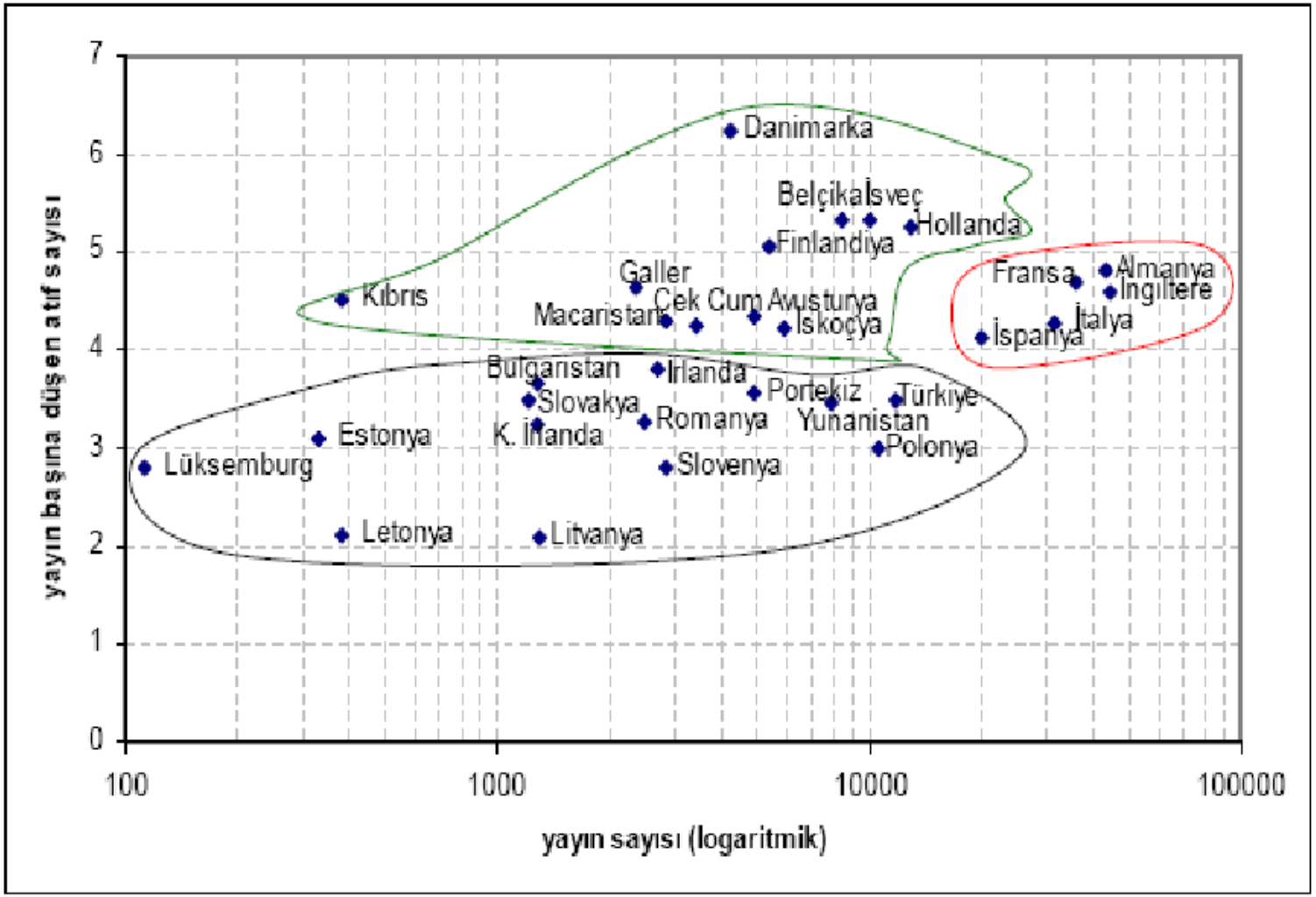
Ülke	Yayın sayısı	Atıf sayısı	Yayın başına düşen atıf sayısı
Almanya	732.911	8.409.979	11,47
İngiltere	652.095	8.385.007	12,86
Fransa	525.128	5.631.061	10,72
İtalya	384.287	3.924.702	10,21
İspanya	288.577	2.589.912	8,97
Hollanda	224.614	3.036.523	13,52
İsveç	167.176	2.150.929	12,87
Polonya	131.120	766.166	5,84
Belçika	122.476	1.428.814	11,67
İskoçya	101.811	1.359.882	13,36
Danimarka	88.472	1.219.245	13,78
Avusturya	85.522	956.342	11,18
Finlandiya	82.658	995.738	12,05
Yunanistan	71.189	492.508	6,92
Çek Cumhuriyeti	57.296	377.910	6,60
Portekiz	49.681	376.682	7,58
Macaristan	46.817	388.410	8,30
İrlanda	37.271	365.909	9,82
Galler	33.956	358.901	10,57
Romanya	26.365	106.199	4,03
Slovakya	21.986	117.761	5,36
Slovenya	19.696	114.832	5,83
Bulgaristan	17.415	91.814	5,27
Kuzey İrlanda	16.744	168.746	10,08
Litvanya	9058	40.468	4,47
Estonya	7347	57.991	7,89
Letonya	3498	20.485	5,86



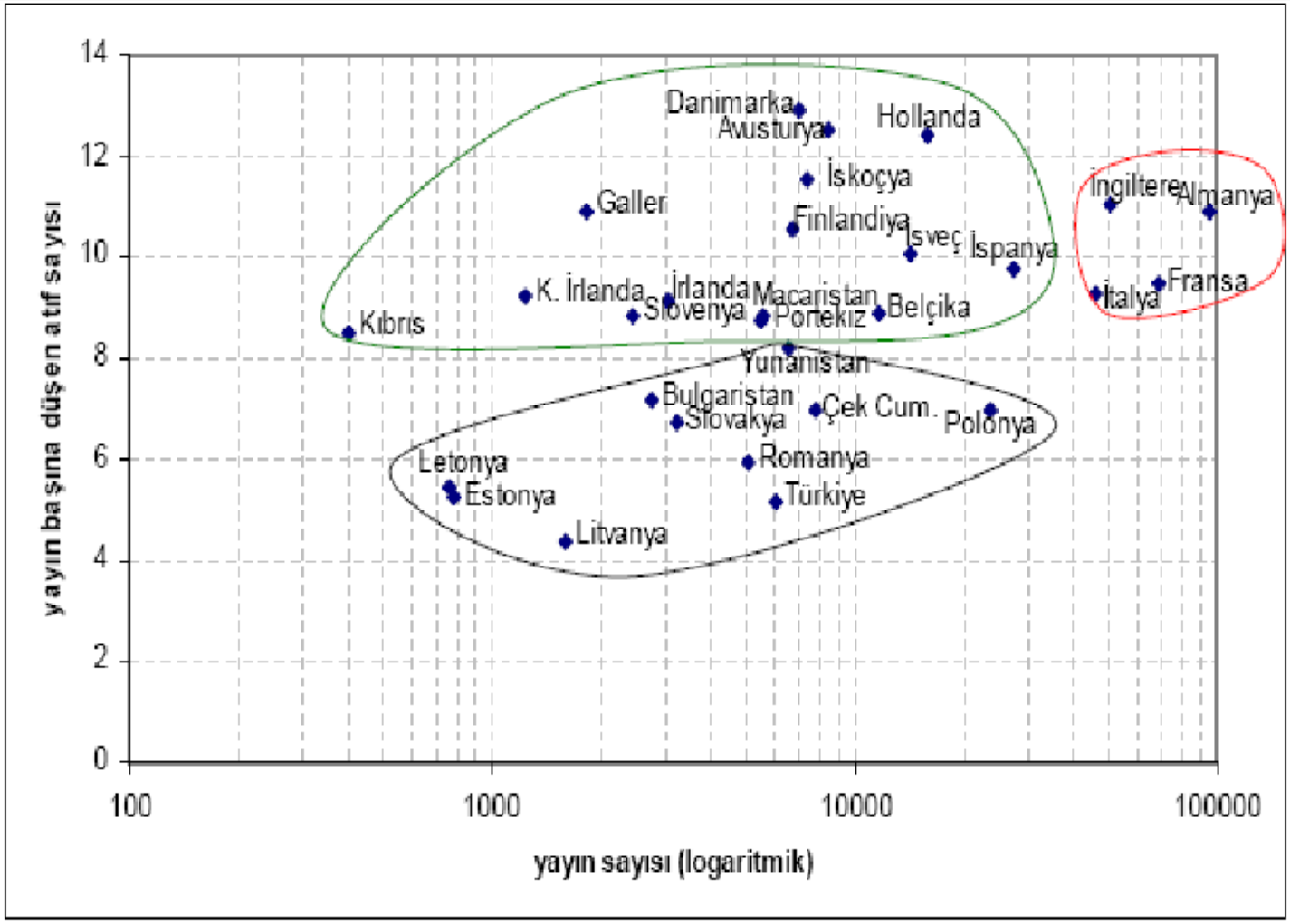
Şekil 1. Yayın ve atıf performansları itibariyle ülkelerin yer aldığı gruplar



Şekil 2. Klinik tıp alanında ülkelerin gruplara dağılımı



Şekil 4. Mühendislik alanında ülkelerin gruplara dağılımı



Şekil 6. Fizik alanında ülkelerin gruplara dağılımı

Tablo 2. Ülkelerin yayın ve atıf sayıları itibariyle yer aldıkları grup sayıları

Ülke	Yayın ve atıf sayısı yüksek	Atıf sayısı yüksek	Atıf sayısı düşük	Toplam alan sayısı
İngiltere	20	1	-	21
Almanya	18	3	-	21
Fransa	17	3	1	21
İtalya	13	4	4	21
İspanya	7	8	6	21
Hollanda	-	18	3	21
İsveç	-	18	3	21
İskoçya	-	17	4	21
Danimarka	-	17	4	21
Belçika	-	16	5	21
Galler	-	15	6	21
Avusturya	-	14	7	21
Finlandiya	-	14	7	21
İrlanda	-	12	9	21
K. İrlanda	-	11	10	21
Macaristan	-	7	14	21
Estonya	-	4	17	21
Portekiz	-	4	17	21
Letonya	-	4	12	16
Kıbrıs	-	4	8	12
Lüksemburg	-	3	11	14
Yunanistan	-	2	19	21
Malta	-	2	3	5
Çek Cumhuriyeti	-	1	20	21
Slovenya	-	1	20	21
Bulgaristan	-	-	21	21
Litvanya	-	-	21	21
Polonya	-	-	21	21
Romanya	-	-	21	21
Slovakya	-	-	21	21
Türkiye	-	-	21	21

Tartışma

- Bilgi toplumu
- Bilgi toplumunun özellikleri
- Sayısal uçurum
- Verilerin e-devlet uygulamalarındaki rolü
- Ticarette veri madenciliğinin kullanımı
- Bilgi ekonomisi