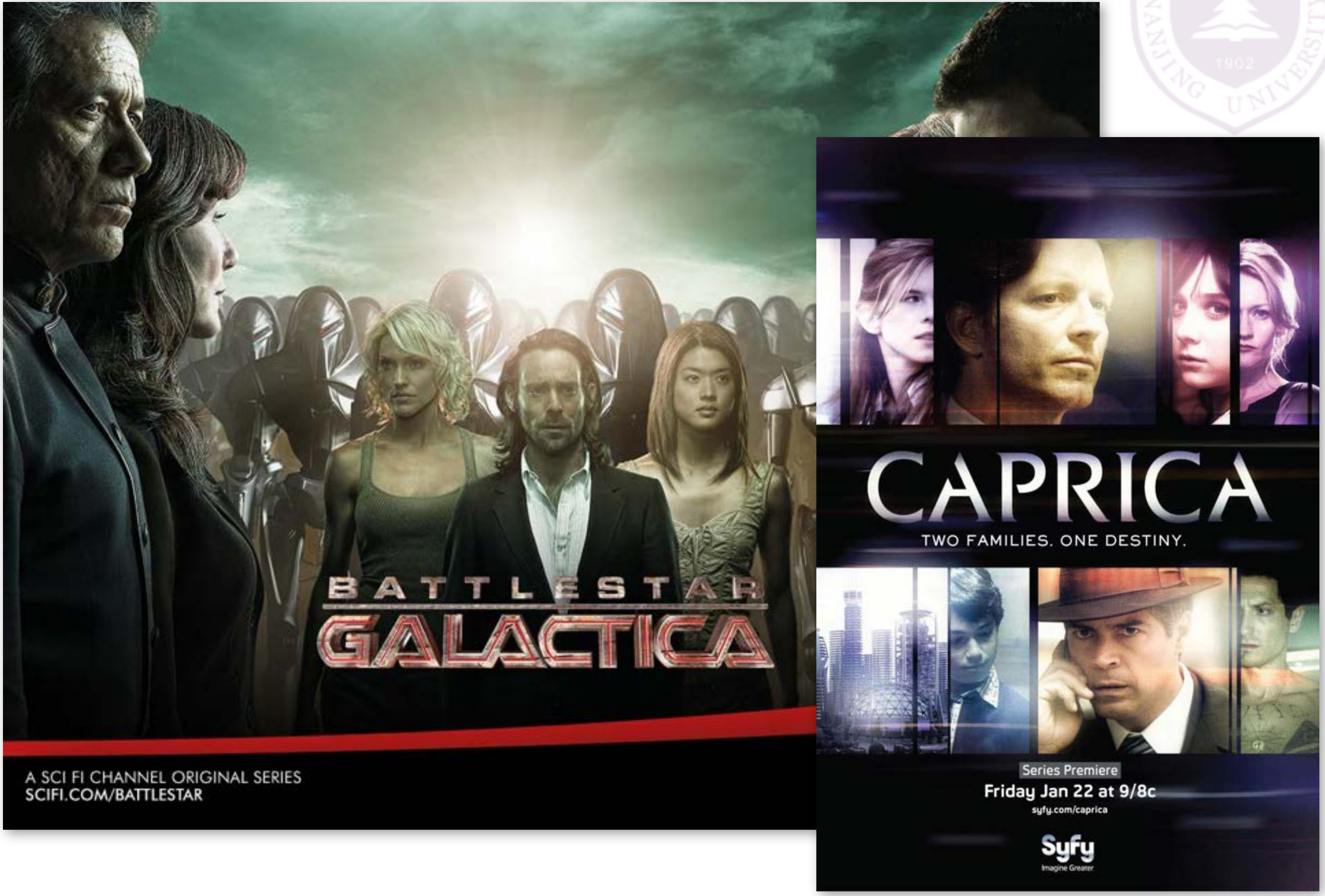
Lecture 1: Introduction

http://cs.nju.edu.cn/yuy/course_dm14ms.ashx

Start from a sci-fiction series

Start from a sci-fiction series



Syfy HD

人脑能存储近一百兆兆位的信息

The human brain contains roughly 100 terabytes of information.

真要用起来 也不算多

Not much, when you get right down to it.

关键不在于如何存储 而是提取

The question isn't how to store it. It's how to access it.

人格是无法下载的

You can't download a personality.

根本无法转换数据

There's no way to translate the data.

但我们脑中存储的信息
But the information being held in our heads

可以从其他数据库提取
Is available in other databases.

人生的轨迹 远不止足迹这么简单

People leave more than footprints as they travel through life.

医疗扫描 基因组态 心理评估

Medical scans, DNA profiles, psych evaluations,

校园活动记录 电子邮件 摄像录音

School records, e-mails, recording video/audio,

造影扫描图　遗传分型

CAT scans, genetic typing,

突触记录 监控录像

Synaptic records, security cameras,

考试成绩 购物单

Test results, shopping records,

才艺表演 球赛

Talent shows, ball games,

罚款单 餐厅账单

Traffic tickets, restaurant bills,

电话记录 音乐列表
Phone records, music lists,

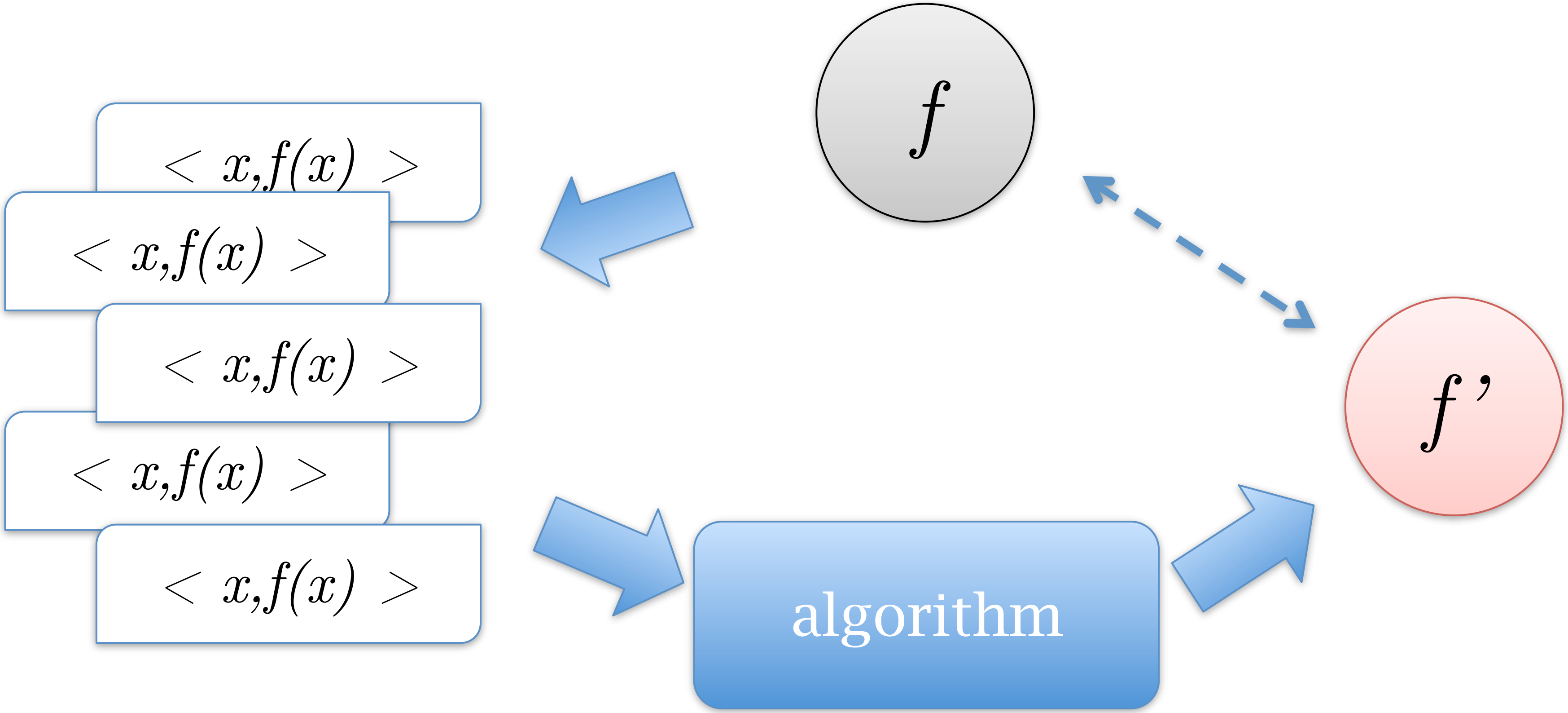
电影票 电视剧

Movie tickets, TV shows.

甚至是节育处方

Even prescriptions for birth control.

An abstract view of DM systems



Data mining

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

[D. Hand et al. , Principles of Data Mining]

数据挖掘是通过对(大规模)观测数据集的分析,寻找确信的关系,并将数据以一种可理解的且利于使用的新颖方式概括数据的方法。

Data mining factors

Large: small data needs no data mining

Unsuspected relationships: correct and significant

Novel: rediscovery of known facts is useless

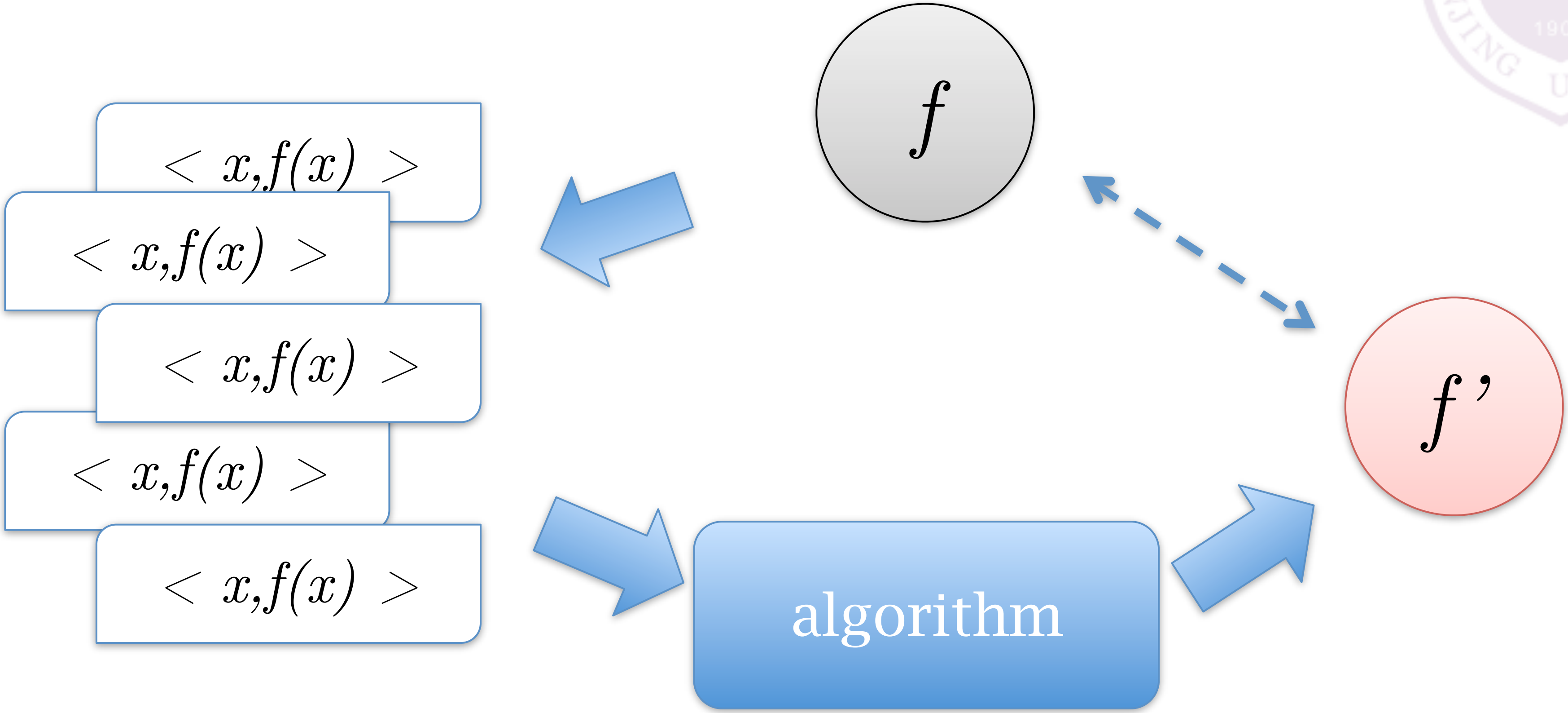
Understandable: decision maker oriented

Useful: mining results should be useful to the users

Observational data v.s. experimental data

[D. Hand et al. , Principles of Data Mining]

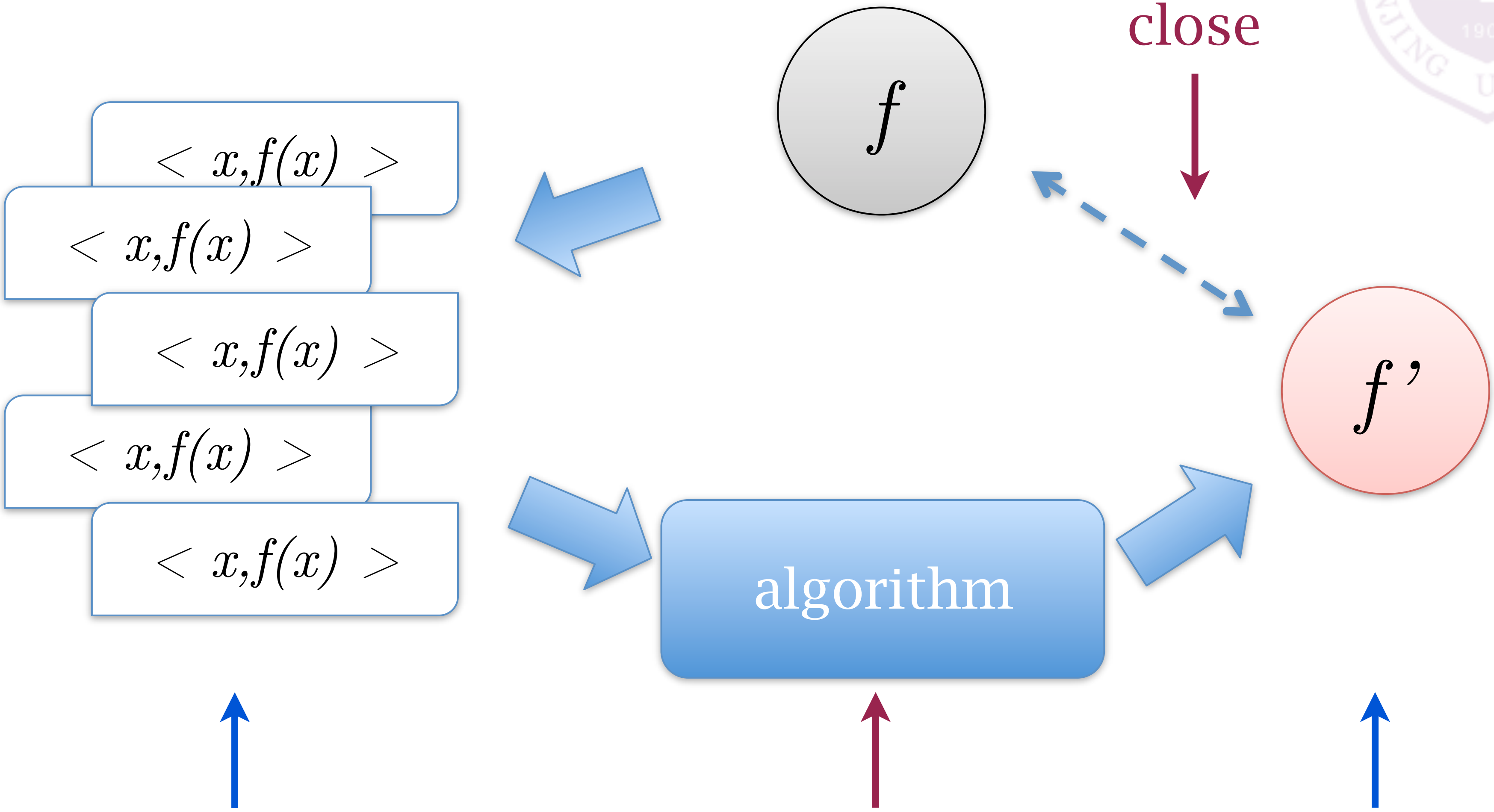
Data mining factors



large
observational

unsuspected
novel
understandable
useful

Data mining factors



$< x,f(x) >$

$< x,f(x) >$

$< x,f(x) >$

$< x,f(x) >$

$< x,f(x) >$

f

close

f ’

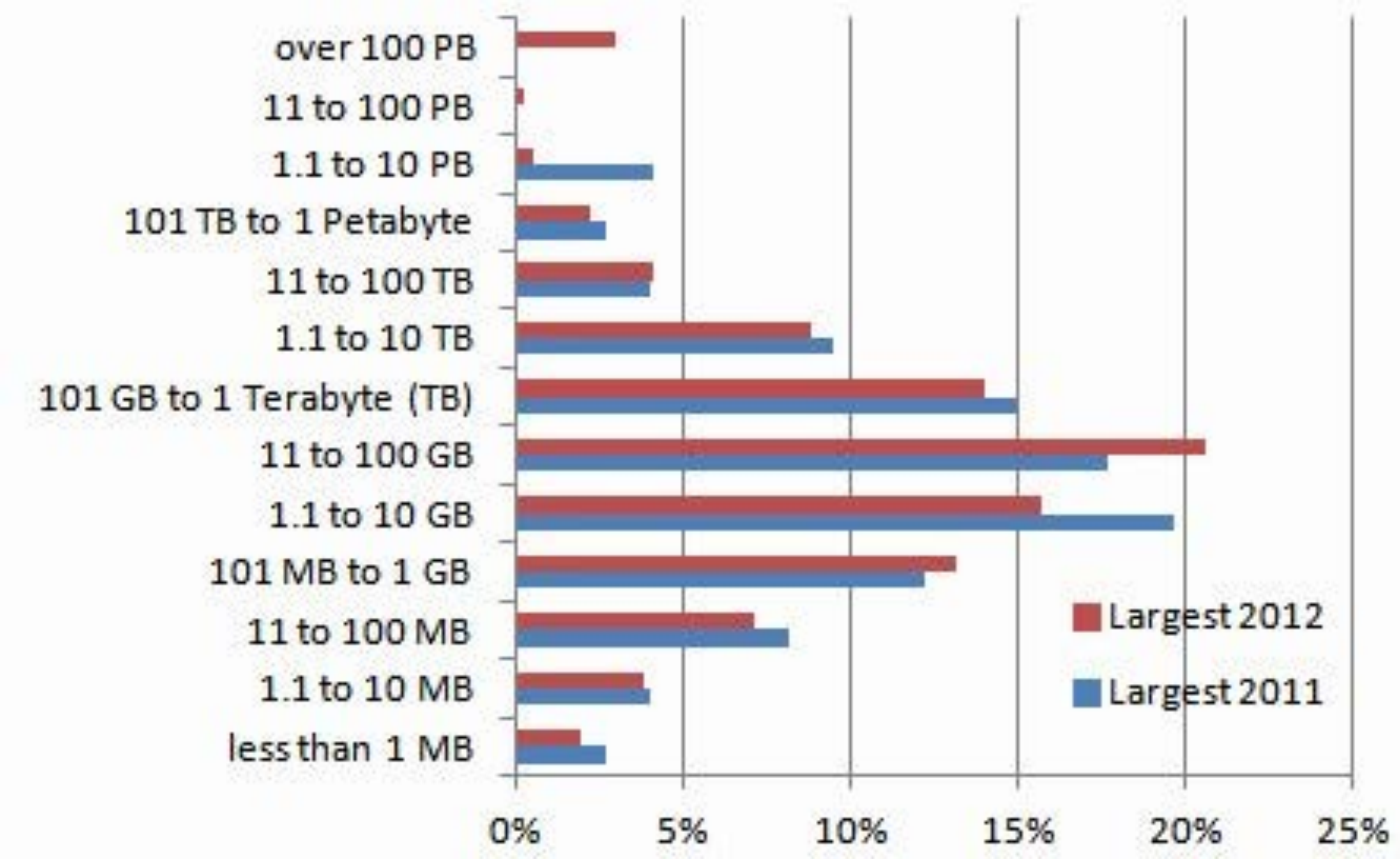
algorithm

large
observational

efficient
(space & time)

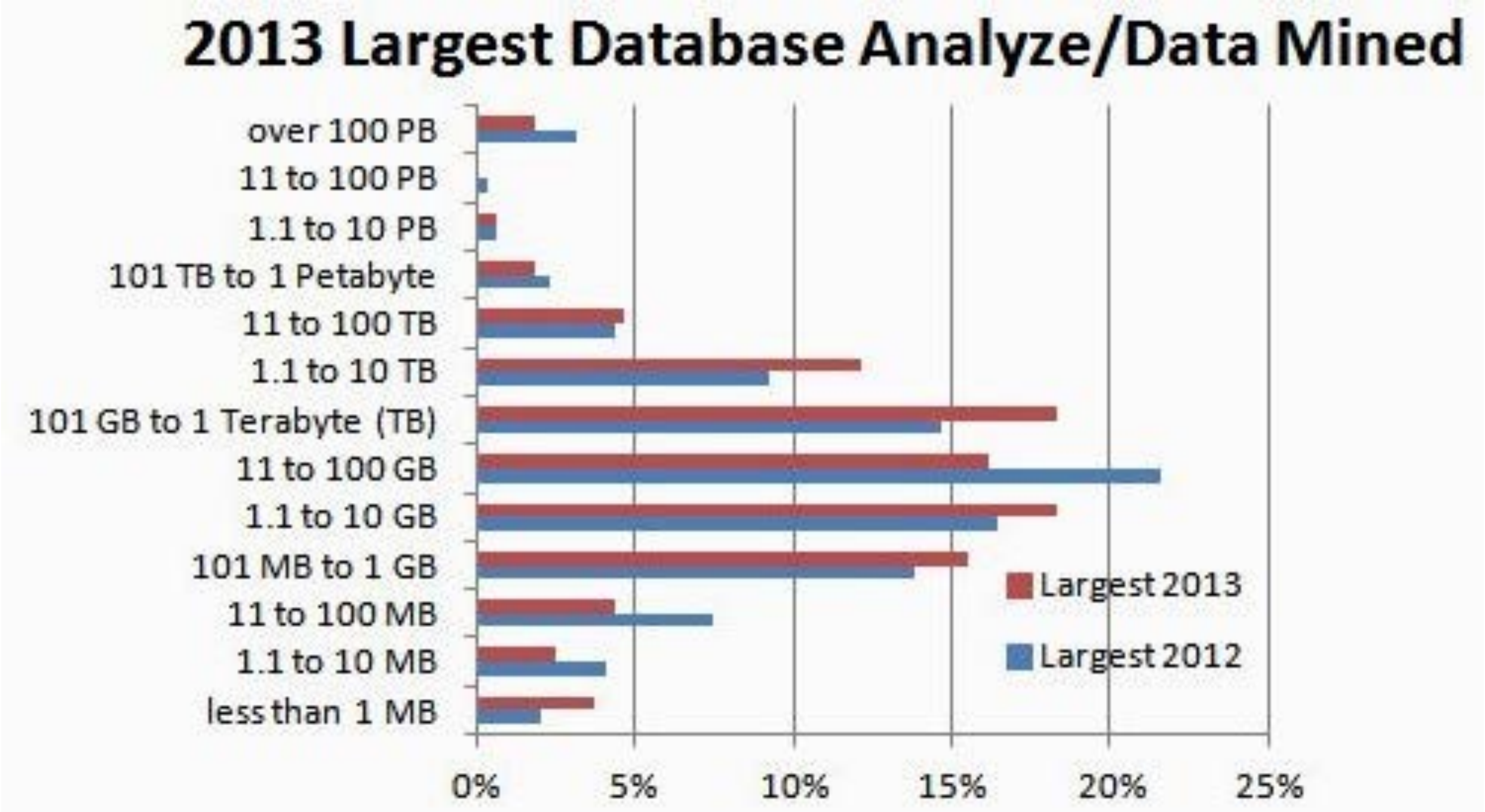
unsuspected
novel
understandable
useful

How large can the data be



[KDnuggets Poll, 2012]

How large can the data be



[KDnuggets Poll, 2013]

What can data mining do? DM Tasks

Exploratory data analysis

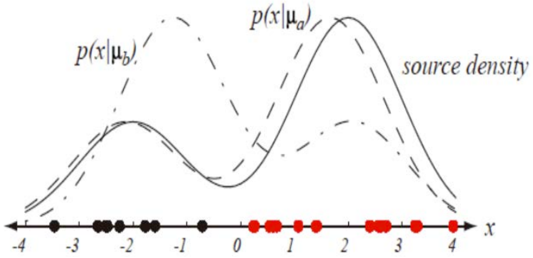
interactive and visualized

how to visualize high dimensional data?

Descriptive modeling

describe a data set

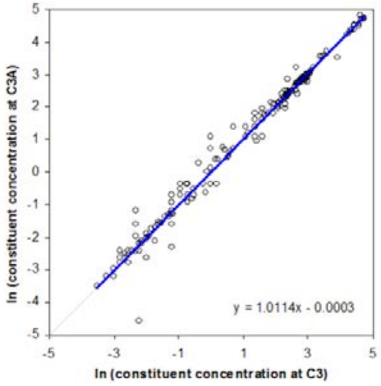
how to characterize general properties of a dataset



Predictive Modeling

perform inference from a data set

how to construct the mapping from the input space to the output space



Discovering patterns and rules

find association relationship

how to find high correlated items out of a huge data set

Retrieval by content

Example: Mining supermarket transactions

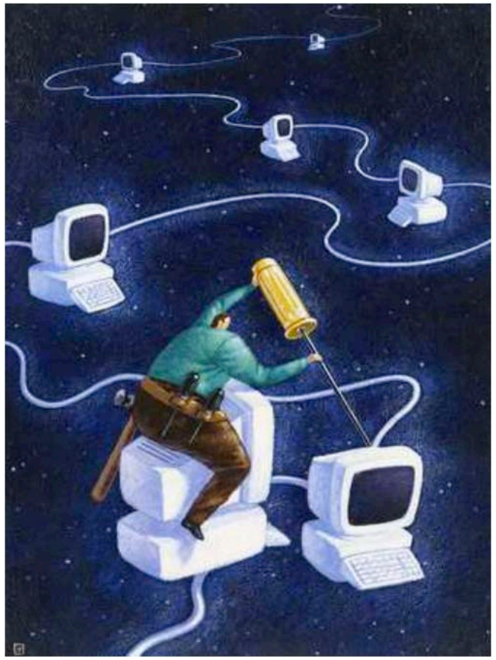
WAL★MART
ALWAYS LOW PRICES.
Always.

Example: Mining valuable customers

GSM

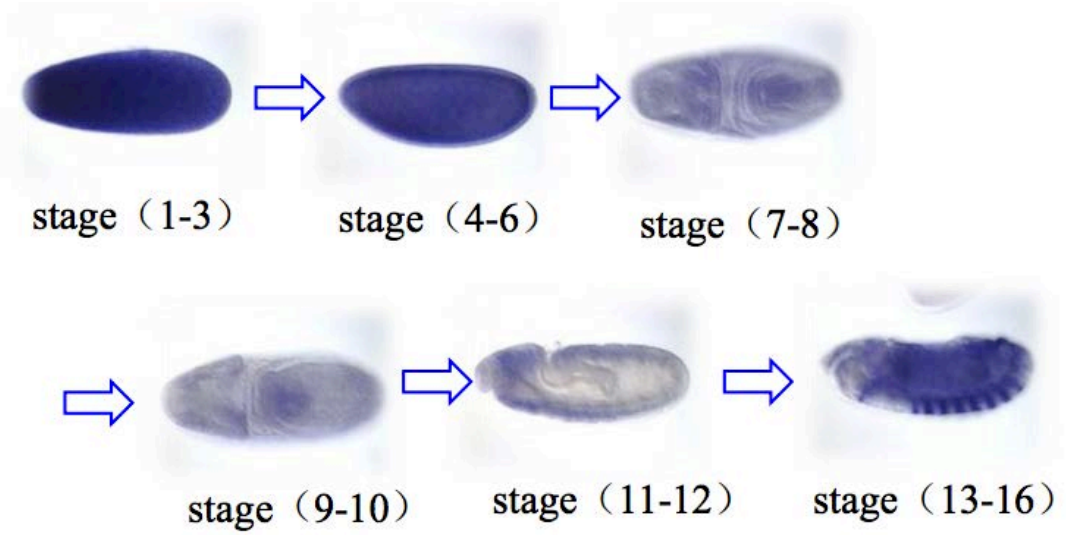
CDMA

Example: mining network intrusion patterns



recognize intrusion accesses

Example: Mining biology data



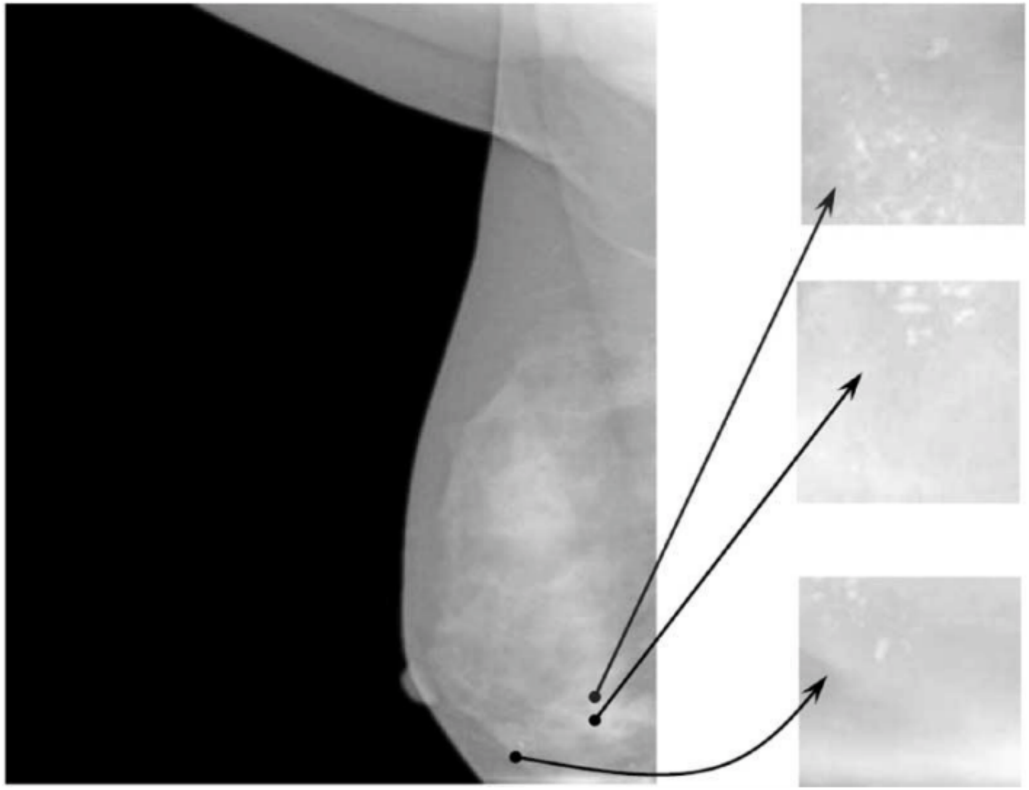
Finding key genes

Identifying gene expression patterns

Identifying gene interactions

...

Example: Mining medical data



Improving diagnosis of doctors by providing suggestions based on historical medical data

Example: Mining financial data

Fraud detection

Stock trends prediction

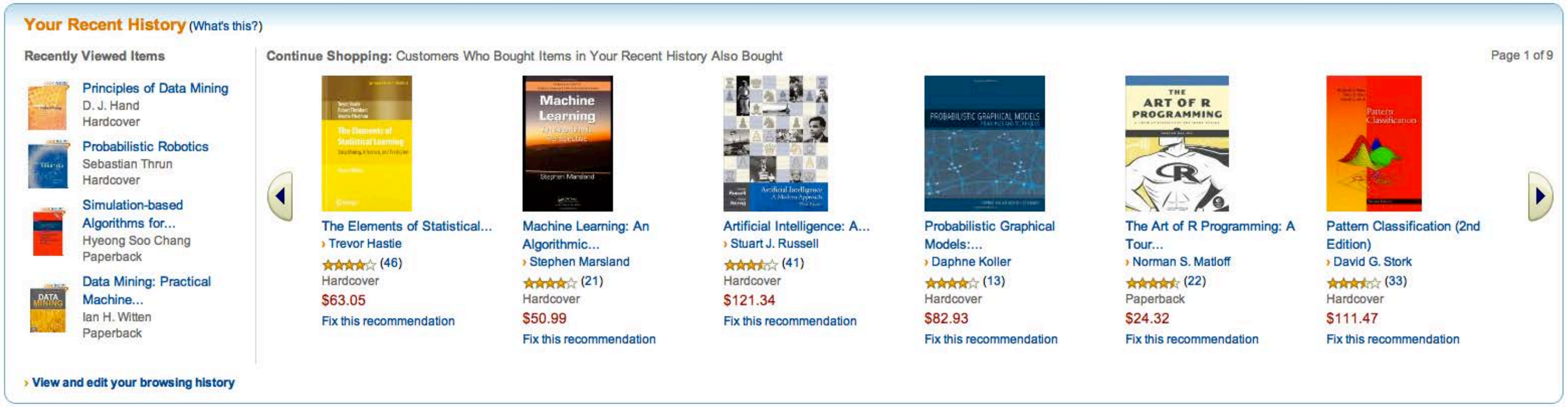
...

Example: Mining the web

Google™

bing™

amazon.com



Your Recent History (What's this?)

Recently Viewed Items

- Principles of Data Mining
  D. J. Hand
  Hardcover
- Probabilistic Robotics
  Sebastian Thrun
  Hardcover
- Simulation-based Algorithms for...
  Hyeong Soo Chang
  Paperback
- Data Mining: Practical Machine...
  Ian H. Witten
  Paperback

› View and edit your browsing history

Continue Shopping: Customers Who Bought Items in Your Recent History Also Bought

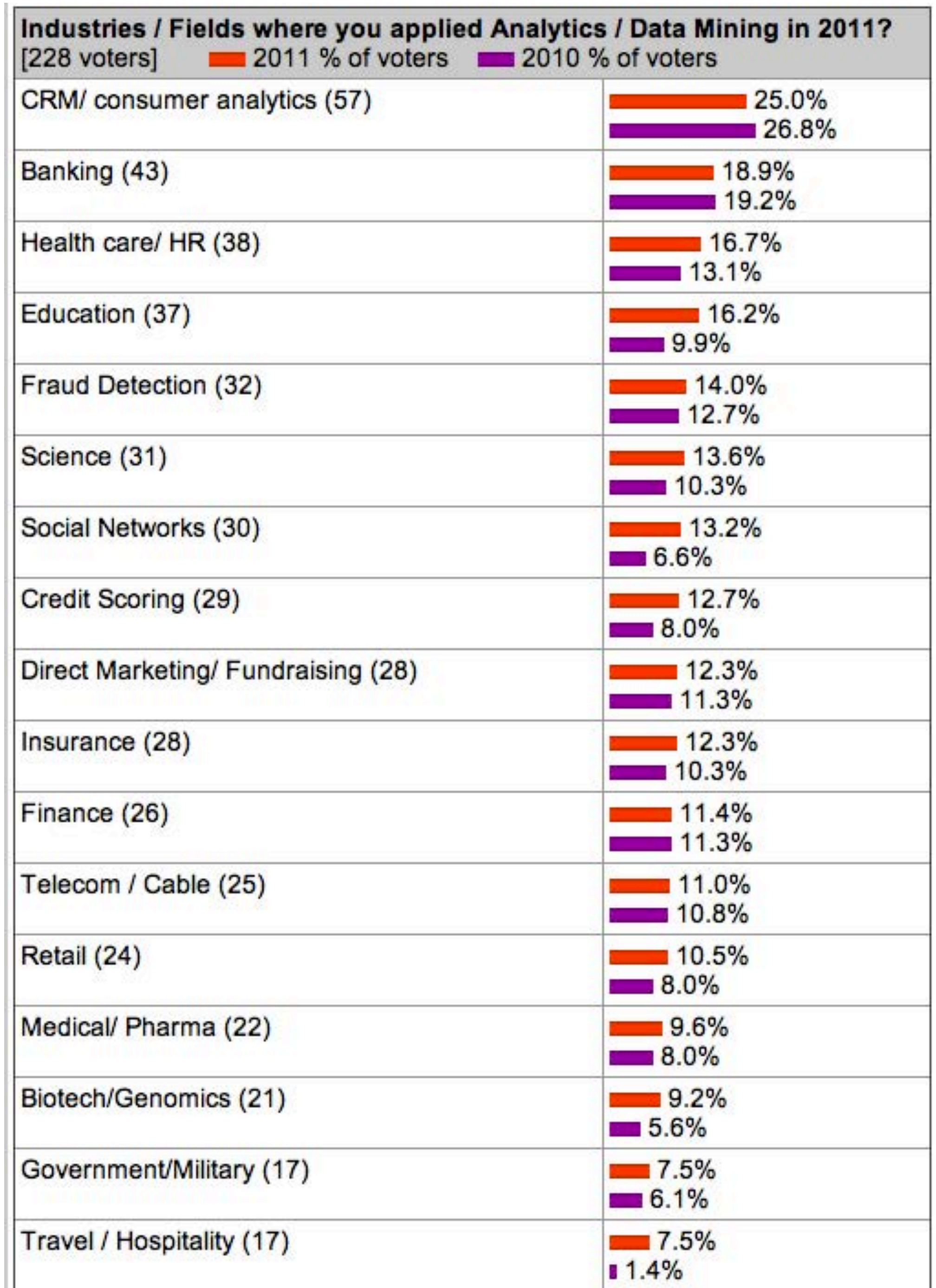
Page 1 of 9

Title	Author	Reviews	Format	Price	
The Elements of Statistical...	› Trevor Hastie	(46)	Hardcover	$63.05	Fix this recommendation
Machine Learning: An Algorithmic...	› Stephen Marsland	(21)	Hardcover	$50.99	Fix this recommendation
Artificial Intelligence: A...	› Stuart J. Russell	(41)	Hardcover	$121.34	Fix this recommendation
Probabilistic Graphical Models:...	› Daphne Koller	(13)	Hardcover	$82.93	Fix this recommendation
The Art of R Programming: A Tour...	› Norman S. Matloff	(22)	Paperback	$24.32	Fix this recommendation
Pattern Classification (2nd Edition)	› David G. Stork	(33)	Hardcover	$111.47	Fix this recommendation

Example: Mining usage data

Mining usage data to allow natural human-computer interaction

Top data mining fields

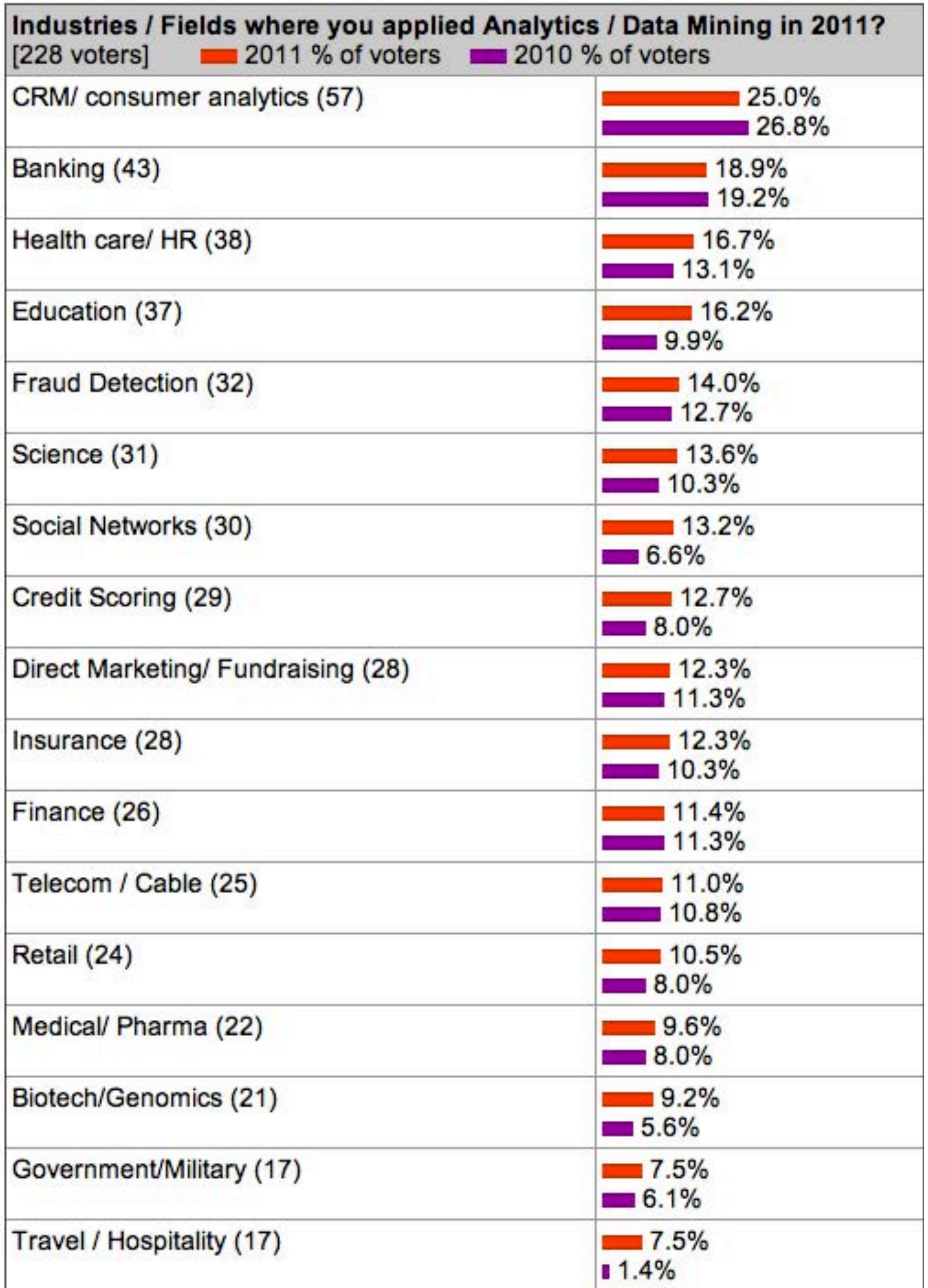


Industries / Fields where you applied Analytics / Data Mining in 2011?
[228 voters] 🟥 2011 % of voters 🟪 2010 % of voters

Field	2011 % of voters	2010 % of voters
CRM/ consumer analytics (57)	25.0%	26.8%
Banking (43)	18.9%	19.2%
Health care/ HR (38)	16.7%	13.1%
Education (37)	16.2%	9.9%
Fraud Detection (32)	14.0%	12.7%
Science (31)	13.6%	10.3%
Social Networks (30)	13.2%	6.6%
Credit Scoring (29)	12.7%	8.0%
Direct Marketing/ Fundraising (28)	12.3%	11.3%
Insurance (28)	12.3%	10.3%
Finance (26)	11.4%	11.3%
Telecom / Cable (25)	11.0%	10.8%
Retail (24)	10.5%	8.0%
Medical/ Pharma (22)	9.6%	8.0%
Biotech/Genomics (21)	9.2%	5.6%
Government/Military (17)	7.5%	6.1%
Travel / Hospitality (17)	7.5%	1.4%

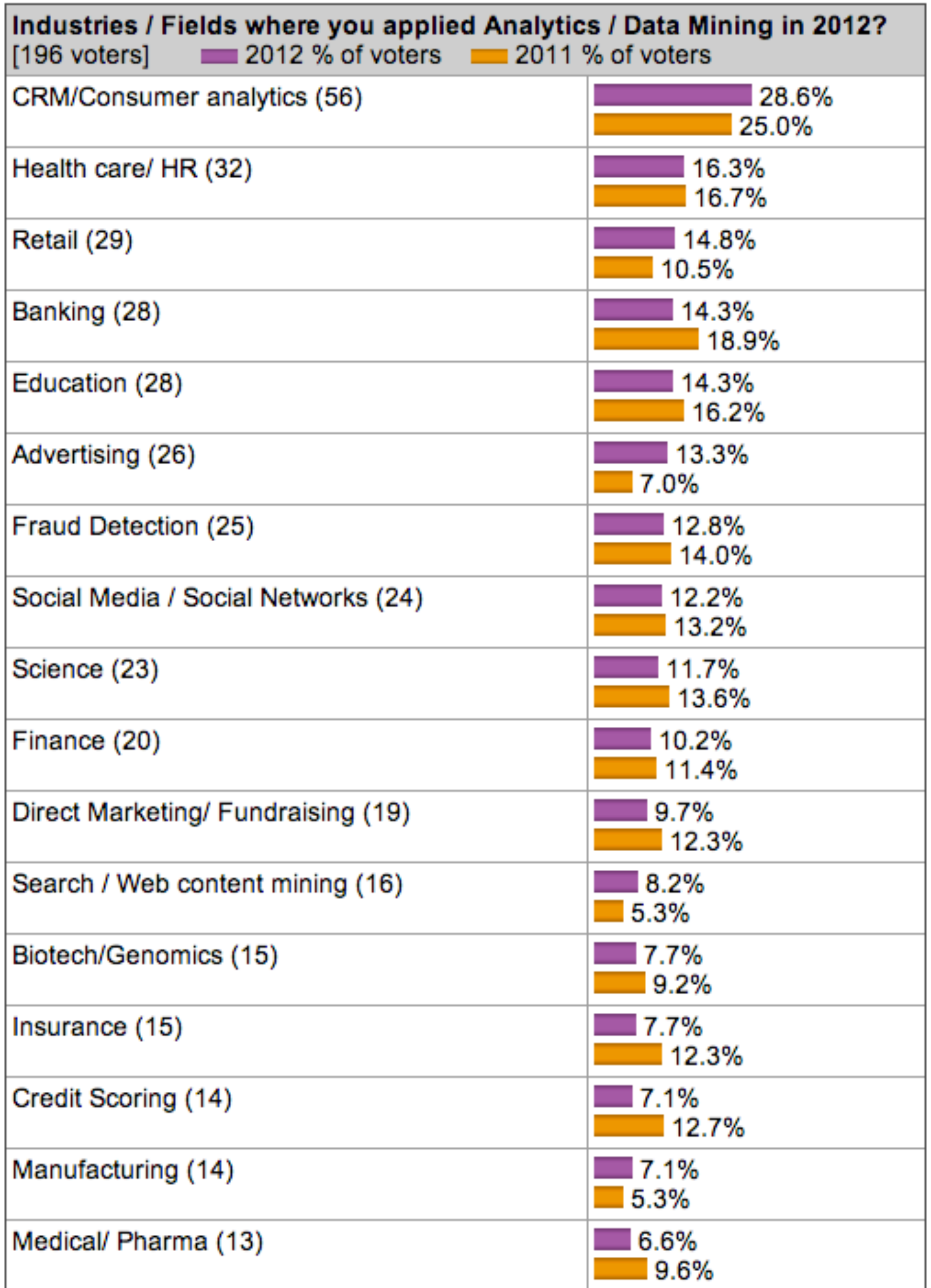
[KDnuggets Poll]

Top data mining fields



Industries / Fields where you applied Analytics / Data Mining in 2011?
[228 voters] 🟥 2011 % of voters 🟪 2010 % of voters

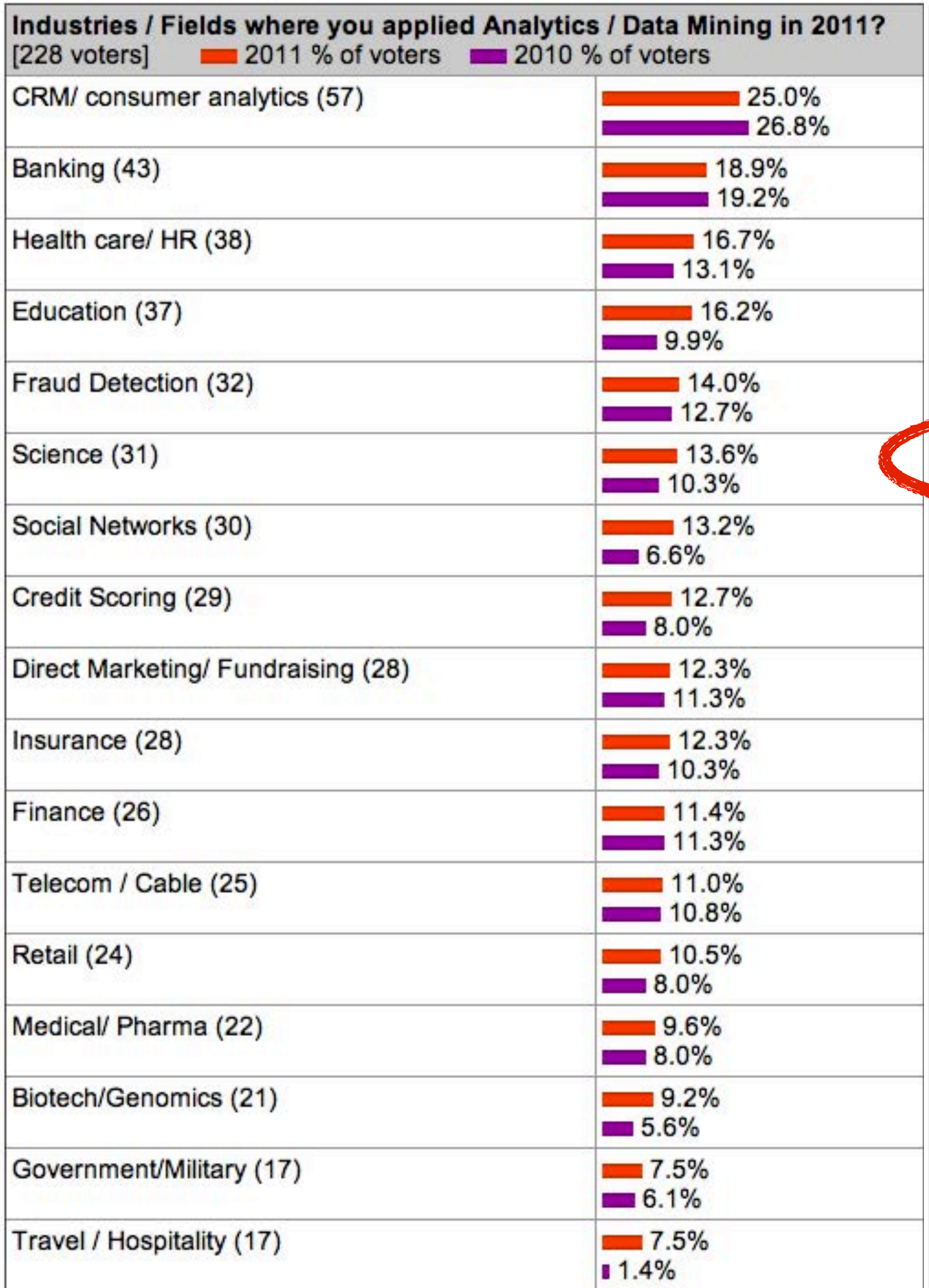
	2011 % of voters	2010 % of voters
CRM/ consumer analytics (57)	25.0%	26.8%
Banking (43)	18.9%	19.2%
Health care/ HR (38)	16.7%	13.1%
Education (37)	16.2%	9.9%
Fraud Detection (32)	14.0%	12.7%
Science (31)	13.6%	10.3%
Social Networks (30)	13.2%	6.6%
Credit Scoring (29)	12.7%	8.0%
Direct Marketing/ Fundraising (28)	12.3%	11.3%
Insurance (28)	12.3%	10.3%
Finance (26)	11.4%	11.3%
Telecom / Cable (25)	11.0%	10.8%
Retail (24)	10.5%	8.0%
Medical/ Pharma (22)	9.6%	8.0%
Biotech/Genomics (21)	9.2%	5.6%
Government/Military (17)	7.5%	6.1%
Travel / Hospitality (17)	7.5%	1.4%



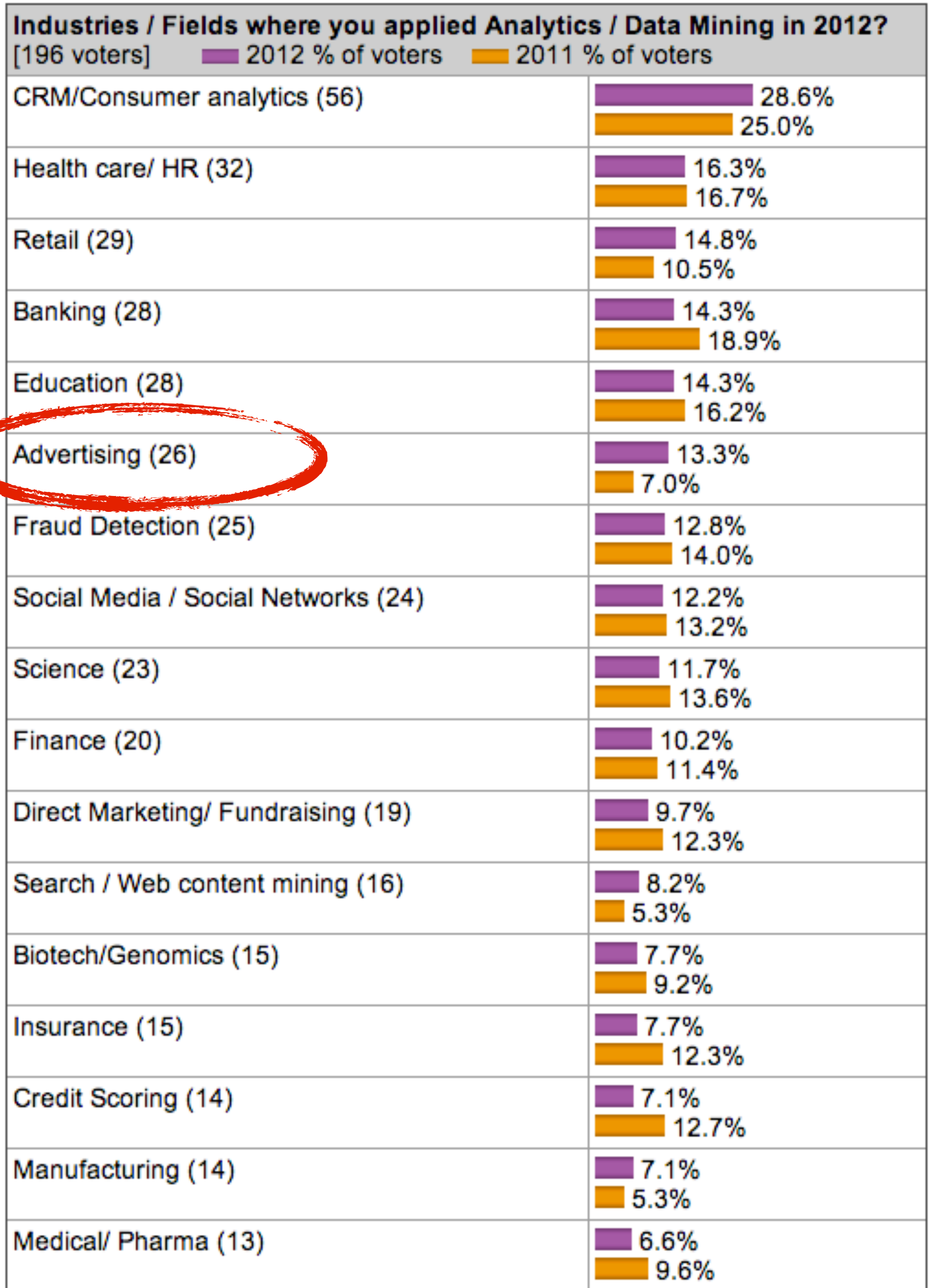
Industries / Fields where you applied Analytics / Data Mining in 2012?
[196 voters] 🟪 2012 % of voters 🟧 2011 % of voters

	2012 % of voters	2011 % of voters
CRM/Consumer analytics (56)	28.6%	25.0%
Health care/ HR (32)	16.3%	16.7%
Retail (29)	14.8%	10.5%
Banking (28)	14.3%	18.9%
Education (28)	14.3%	16.2%
Advertising (26)	13.3%	7.0%
Fraud Detection (25)	12.8%	14.0%
Social Media / Social Networks (24)	12.2%	13.2%
Science (23)	11.7%	13.6%
Finance (20)	10.2%	11.4%
Direct Marketing/ Fundraising (19)	9.7%	12.3%
Search / Web content mining (16)	8.2%	5.3%
Biotech/Genomics (15)	7.7%	9.2%
Insurance (15)	7.7%	12.3%
Credit Scoring (14)	7.1%	12.7%
Manufacturing (14)	7.1%	5.3%
Medical/ Pharma (13)	6.6%	9.6%

Top data mining fields



Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters] 🟥 2011 % of voters 🟪 2010 % of voters	
CRM/ consumer analytics (57)	25.0% 26.8%
Banking (43)	18.9% 19.2%
Health care/ HR (38)	16.7% 13.1%
Education (37)	16.2% 9.9%
Fraud Detection (32)	14.0% 12.7%
Science (31)	13.6% 10.3%
Social Networks (30)	13.2% 6.6%
Credit Scoring (29)	12.7% 8.0%
Direct Marketing/ Fundraising (28)	12.3% 11.3%
Insurance (28)	12.3% 10.3%
Finance (26)	11.4% 11.3%
Telecom / Cable (25)	11.0% 10.8%
Retail (24)	10.5% 8.0%
Medical/ Pharma (22)	9.6% 8.0%
Biotech/Genomics (21)	9.2% 5.6%
Government/Military (17)	7.5% 6.1%
Travel / Hospitality (17)	7.5% 1.4%

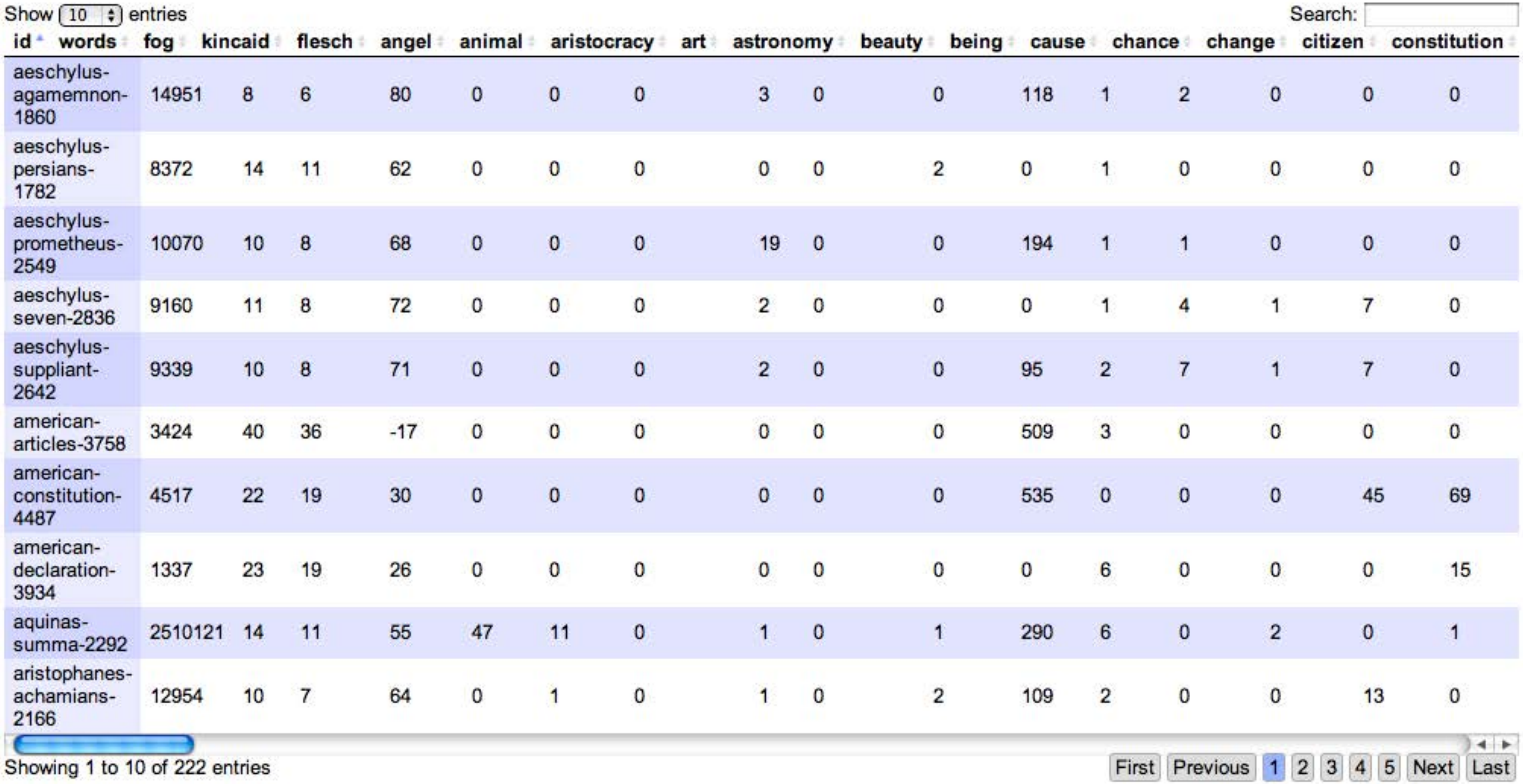


Industries / Fields where you applied Analytics / Data Mining in 2012? [196 voters] 🟪 2012 % of voters 🟧 2011 % of voters	
CRM/Consumer analytics (56)	28.6% 25.0%
Health care/ HR (32)	16.3% 16.7%
Retail (29)	14.8% 10.5%
Banking (28)	14.3% 18.9%
Education (28)	14.3% 16.2%
Advertising (26)	13.3% 7.0%
Fraud Detection (25)	12.8% 14.0%
Social Media / Social Networks (24)	12.2% 13.2%
Science (23)	11.7% 13.6%
Finance (20)	10.2% 11.4%
Direct Marketing/ Fundraising (19)	9.7% 12.3%
Search / Web content mining (16)	8.2% 5.3%
Biotech/Genomics (15)	7.7% 9.2%
Insurance (15)	7.7% 12.3%
Credit Scoring (14)	7.1% 12.7%
Manufacturing (14)	7.1% 5.3%
Medical/ Pharma (13)	6.6% 9.6%

]

Data types in mining tasks

“Flat” data: vectors and matrix

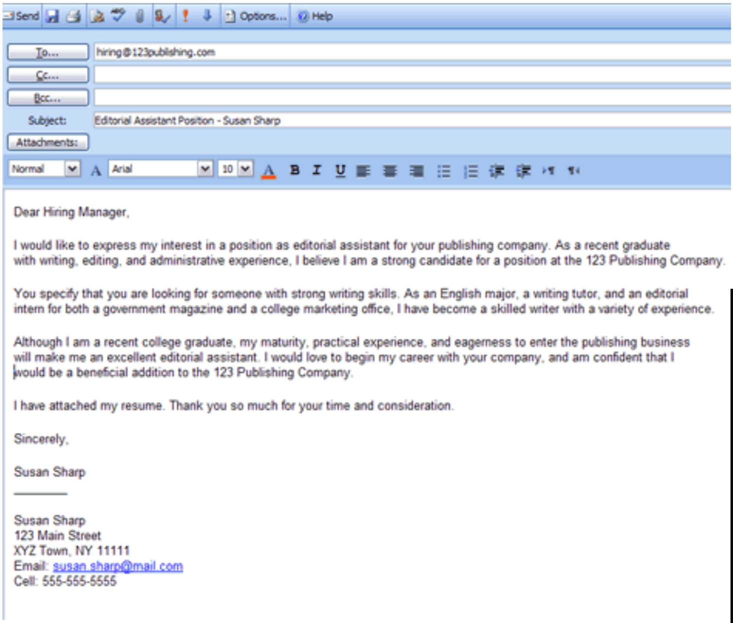


id	words	fog	kincaid	flesch	angel	animal	aristocracy	art	astronomy	beauty	being	cause	chance	change	citizen	constitution
aeschylus-agamemnon-1860	14951	8	6	80	0	0	0	3	0	0	118	1	2	0	0	0
aeschylus-persians-1782	8372	14	11	62	0	0	0	0	0	2	0	1	0	0	0	0
aeschylus-prometheus-2549	10070	10	8	68	0	0	0	19	0	0	194	1	1	0	0	0
aeschylus-seven-2836	9160	11	8	72	0	0	0	2	0	0	0	1	4	1	7	0
aeschylus-suppliant-2642	9339	10	8	71	0	0	0	2	0	0	95	2	7	1	7	0
american-articles-3758	3424	40	36	-17	0	0	0	0	0	0	509	3	0	0	0	0
american-constitution-4487	4517	22	19	30	0	0	0	0	0	0	535	0	0	0	45	69
american-declaration-3934	1337	23	19	26	0	0	0	0	0	0	0	6	0	0	0	15
aquinas-summa-2292	2510121	14	11	55	47	11	0	1	0	1	290	6	0	2	0	1
aristophanes-achamians-2166	12954	10	7	64	0	1	0	1	0	2	109	2	0	0	13	0

Showing 1 to 10 of 222 entries

Data types in mining tasks

Text data



To... hiring@123publishing.com

Cc...

Bcc...

Subject: Editorial Assistant Position - Susan Sharp

Attachments:

Dear Hiring Manager,

I would like to express my interest in a position as editorial assistant for your publishing company. As a recent graduate with writing, editing, and administrative experience, I believe I am a strong candidate for a position at the 123 Publishing Company.

You specify that you are looking for someone with strong writing skills. As an English major, a writing tutor, and an editorial intern for both a government magazine and a college marketing office, I have become a skilled writer with a variety of experience.

Although I am a recent college graduate, my maturity, practical experience, and eagerness to enter the publishing business will make me an excellent editorial assistant. I would love to begin my career with your company, and am confident that I would be a beneficial addition to the 123 Publishing Company.

I have attached my resume. Thank you so much for your time and consideration.

Sincerely,

Susan Sharp

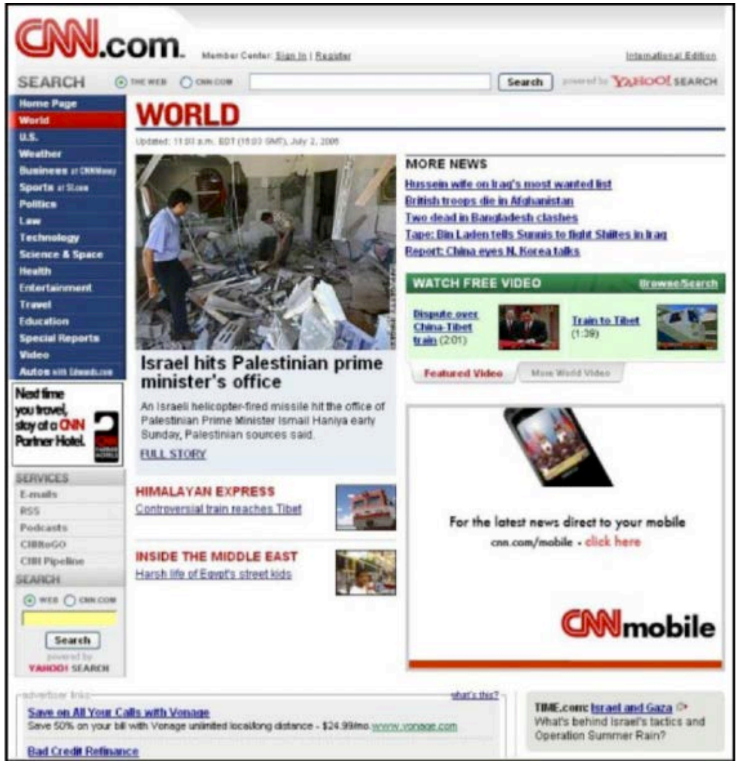
Susan Sharp
123 Main Street
XYZ Town, NY 11111
Email: susan.sharp@mail.com
Cell: 555-555-5555

אני

Got your email. Do you want me to get in touch? I was driving before and this song was on the radio. aznavour, isabelle. youtube. Check your email

Data types in mining tasks

Structured data



WORLD

Updated: 11:03 a.m. EDT (15:03 GMT), July 2, 2006

Israel hits Palestinian prime minister's office

An Israeli helicopter-fired missile hit the office of Palestinian Prime Minister Ismail Haniya early Sunday, Palestinian sources said.

FULL STORY

MORE NEWS

- Hussein wife on Iraq's most wanted list
- British troops die in Afghanistan
- Two dead in Bangladesh clashes
- Tape: Bin Laden tells Sunnis to fight Shiites in Iraq
- Report: China eyes N. Korea talks

Dispute over China-Tibet train (2:01)

Train to Tibet (1:39)

HIMALAYAN EXPRESS

Controversial train reaches Tibet

INSIDE THE MIDDLE EAST

Harsh life of Egypt's street kids

For the latest news direct to your mobile cnn.com/mobile - click here

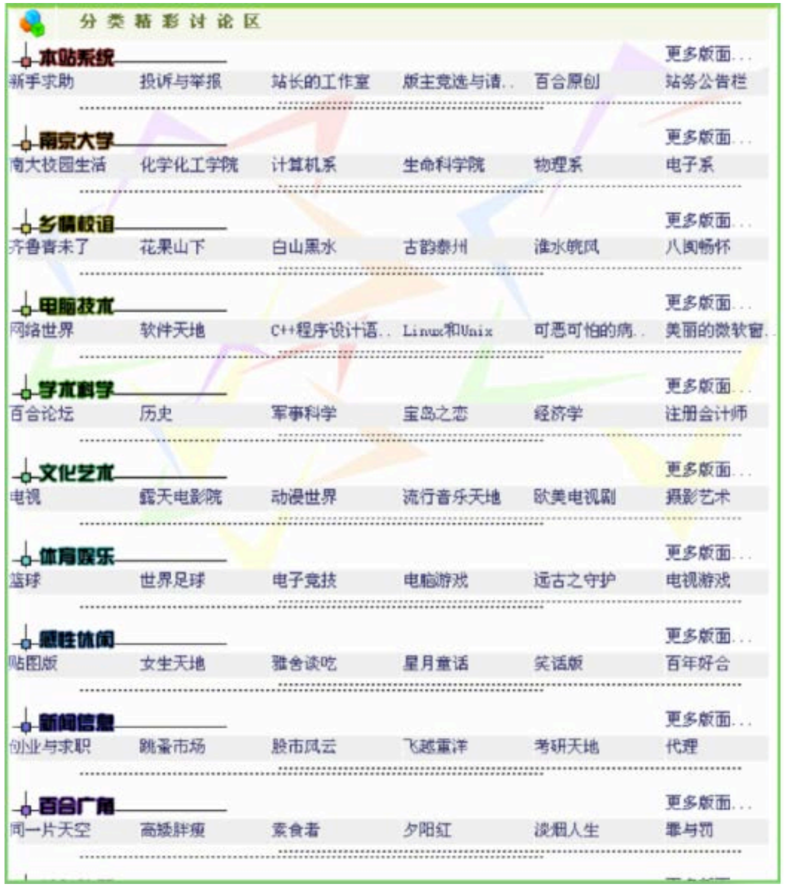
Save on All Your Calls with Vonage

Save 50% on your bill with Vonage unlimited local/long distance - $24.99/mo www.vonage.com

Bad Credit Refinance

TIME.com: Israel and Gaza

What's behind Israel's tactics and Operation Summer Rain?



分类精彩讨论区

本站系统 更多版面...

新手求助　投诉与举报　站长的工作室　版主竞选与请..　百合原创　站务公告栏

南京大学 更多版面...

南大校园生活　化学化工学院　计算机系　生命科学院　物理系　电子系

乡情校谊 更多版面...

齐鲁青未了　花果山下　白山黑水　古韵泰州　淮水皖风　八闽畅怀

电脑技术 更多版面...

网络世界　软件天地　C++程序设计语..　Linux和Unix　可恶可怕的病..　美丽的微软窗..

学术科学 更多版面...

百合论坛　历史　军事科学　宝岛之恋　经济学　注册会计师

文化艺术 更多版面...

电视　露天电影院　动漫世界　流行音乐天地　欧美电视剧　摄影艺术

体育娱乐 更多版面...

篮球　世界足球　电子竞技　电脑游戏　远古之守护　电视游戏

感性休闲 更多版面...

贴图版　女生天地　雅舍谈吃　星月童话　笑话版　百年好合

新闻信息 更多版面...

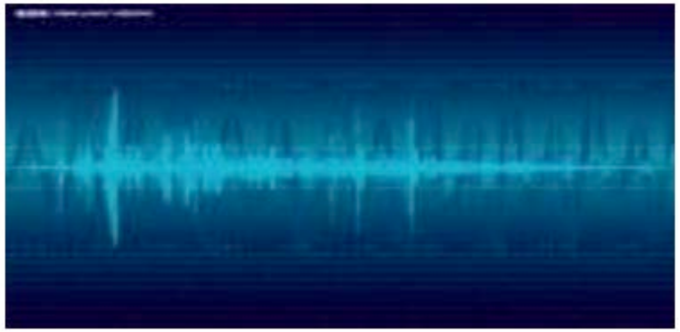
创业与求职　跳蚤市场　股市风云　飞越重洋　考研天地　代理

百合广角 更多版面...

同一片天空　高矮胖瘦　素食者　夕阳红　淡烟人生　罪与罚

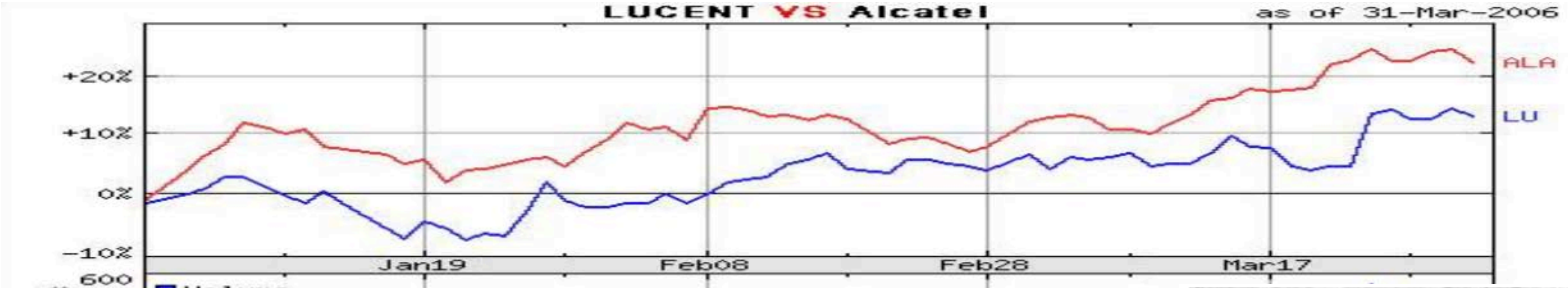
Data types in mining tasks

Multi-media data

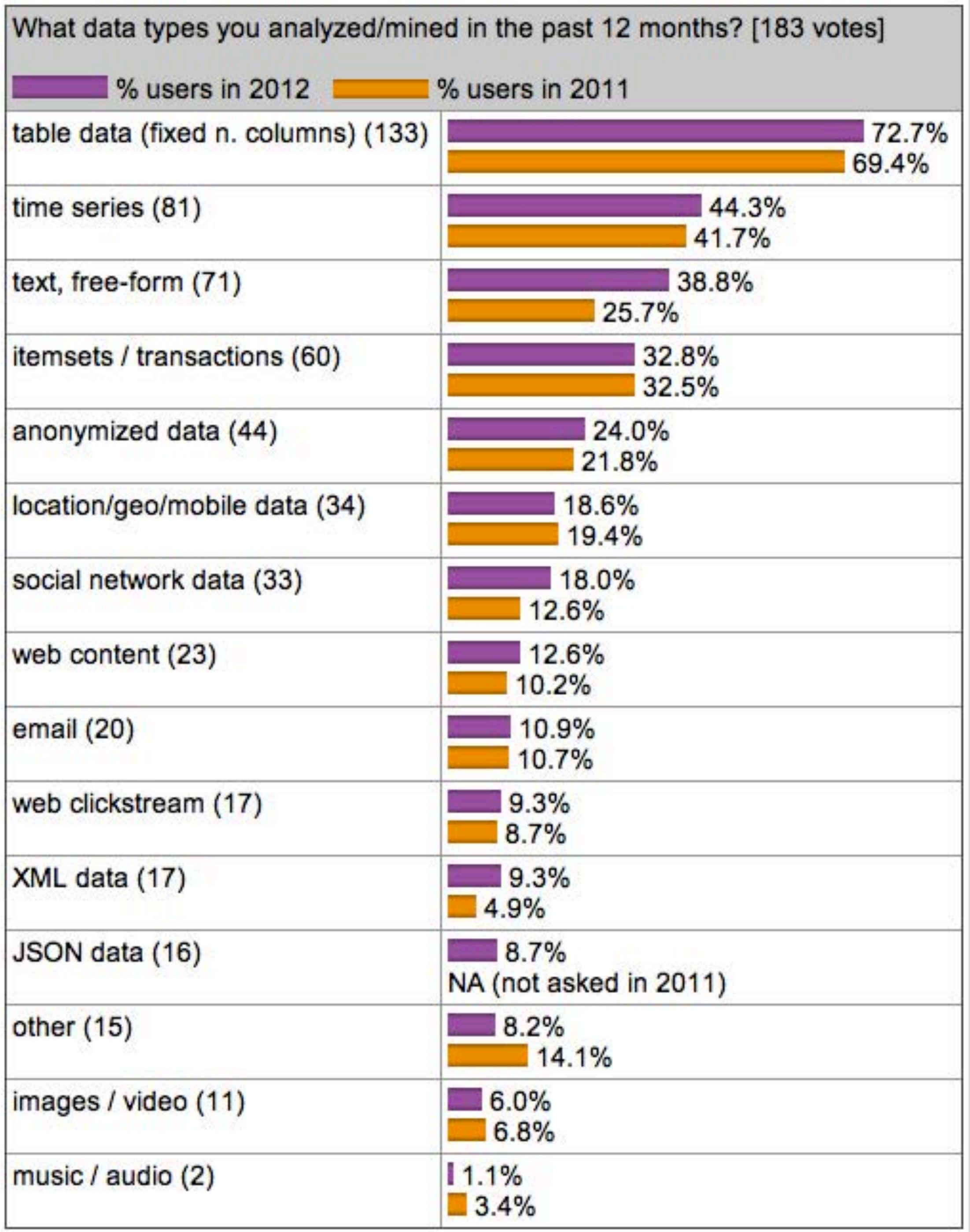


Data types in mining tasks

Temporal and spatial data



Top mined data types



[KDnuggets Poll, 2012]

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

- HUAWEI
- SOSO 搜搜
- Baidu 百度
- SAP®
- 阿里云 aliyun.com
- Microsoft®

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI
SOSO 搜搜
Bai du 百度
SAP®
阿里云
aliyun.com
Microsoft®
EMC²

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

Nokia Research Center

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

Nokia Research Center

SOSO 搜搜

PayPal of ebay

Baidu 百度

SAP

阿里云 aliyun.com

Microsoft

EMC^2

Greenplum

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebY

Deloitte.

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebay

Deloitte.

sas

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebaY

Deloitte.

sas

Adobe

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebaY

Deloitte.

sas

Adobe®

Google

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebY

Deloitte.

sas

Adobe

Google

IBM Research

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

- HUAWEI
- SOSO 搜搜
- Baidu 百度
- SAP®
- 阿里云 aliyun.com
- Microsoft®
- EMC²
- Greenplum

- Nokia Research Center
- PayPal™ of eBay
- Deloitte.
- sas
- Adobe
- Google
- IBM Research
- m6d media6degrees

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC^2

Greenplum

Nokia Research Center

PayPal™ of ebay

Deloitte.

sas

Adobe

Google

IBM Research

m6d media6degrees

facebook

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI
SOSO 搜搜
Bai du 百度
SAP®
阿里云 aliyun.com
Microsoft®
EMC²
Greenplum
Nokia Research Center
PayPal™ of ebay
Deloitte.
sas
Adobe
Google
IBM Research
m6d media6degrees
facebook
SAMSUNG
China Samsung Telecom R&D Center

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebY

Deloitte.

sas

Adobe®

Google

IBM Research

m6d media6degrees

facebook

SAMSUNG China Samsung Telecom R&D Center

CITYGRID®.com

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebY

Deloitte.

sas

Adobe®

Google

IBM Research

m6d media6degrees

facebook

SAMSUNG China Samsung Telecom R&D Center

CITYGRID®.com

hp

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Bai du 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebay

Deloitte.

sas

Adobe

Google

IBM Research

m6d media6degrees

facebook

SAMSUNG China Samsung Telecom R&D Center

CITYGRID.com

hp

OPERA SOLUTIONS

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC^2

Greenplum

Nokia Research Center

PayPal™ of ebay

Deloitte.

sas

Adobe

Google

IBM Research

m6d media6degrees

facebook

SAMSUNG China Samsung Telecom R&D Center

CITYGRID®.com

hp

OPERA SOLUTIONS

technicolor

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

HUAWEI

SOSO 搜搜

Baidu 百度

SAP®

阿里云 aliyun.com

Microsoft®

EMC²

Greenplum

Nokia Research Center

PayPal™ of ebaY

Deloitte.

sas

Adobe

Google

IBM Research

m6d media6degrees

facebook

SAMSUNG China Samsung Telecom R&D Center

CITYGRID.com

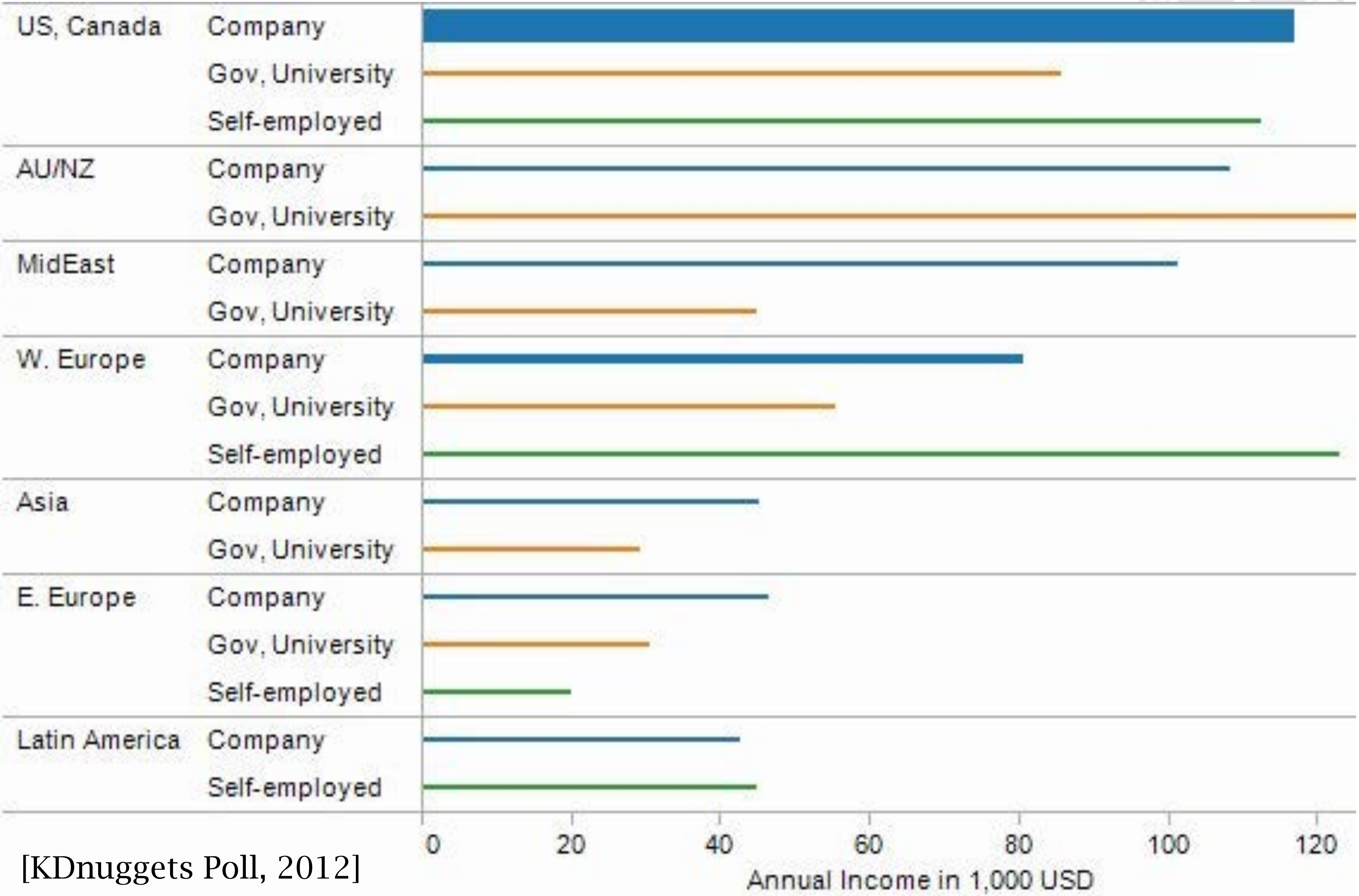
hp

OPERA SOLUTIONS

technicolor

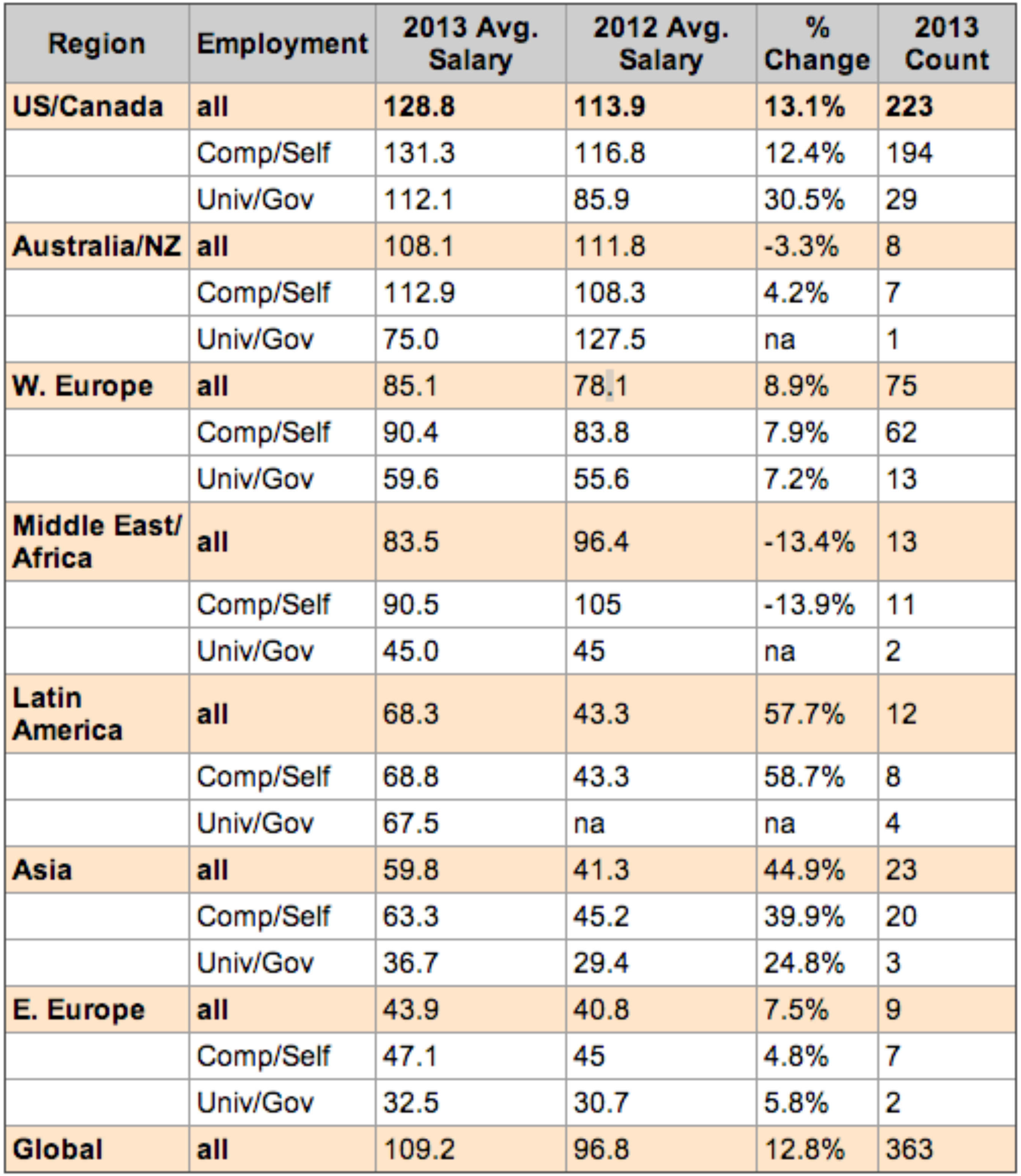
SALFORD SYSTEMS

Annual salary of data miners



[KDnuggets Poll, 2012]

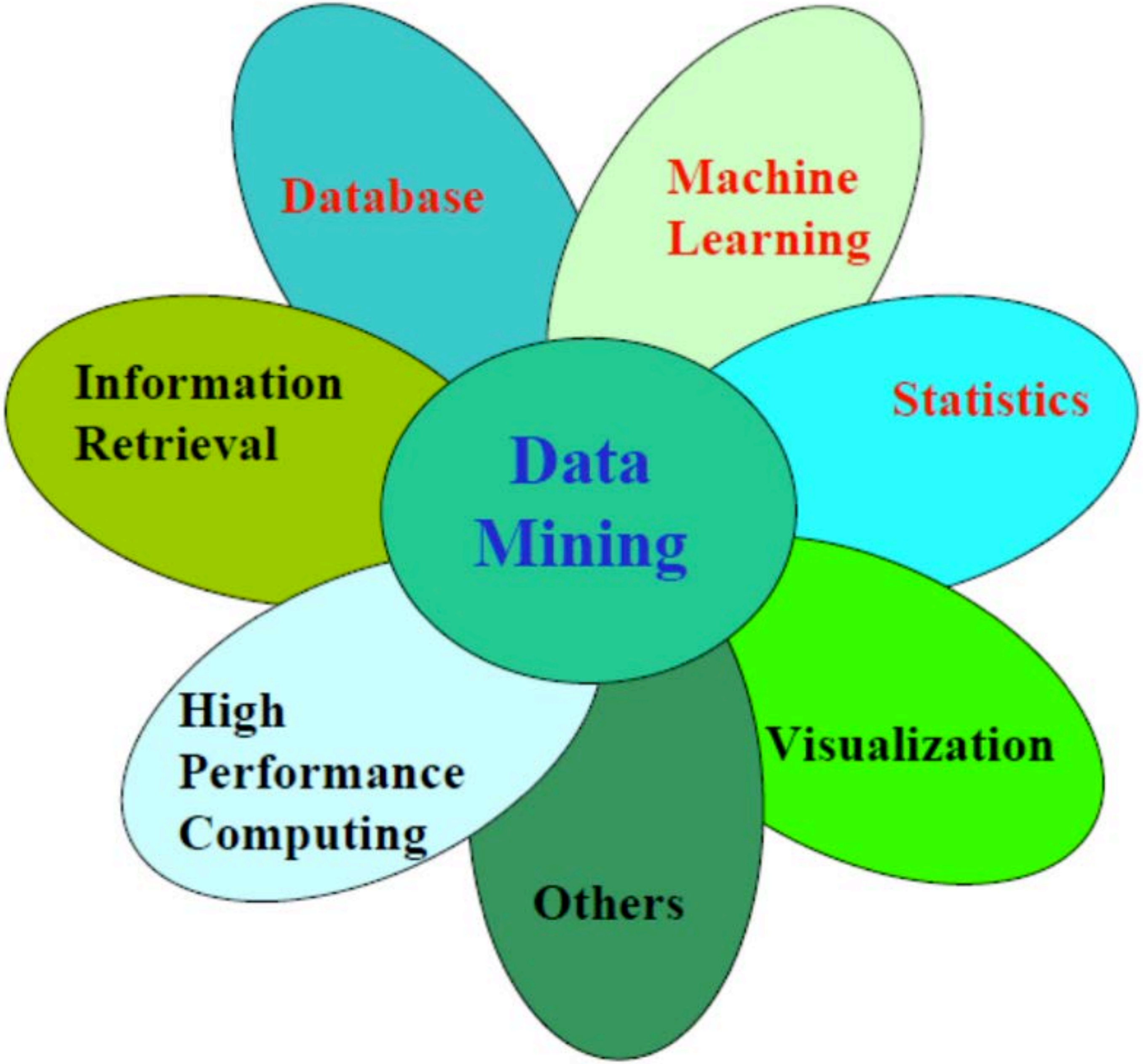
Annual salary o



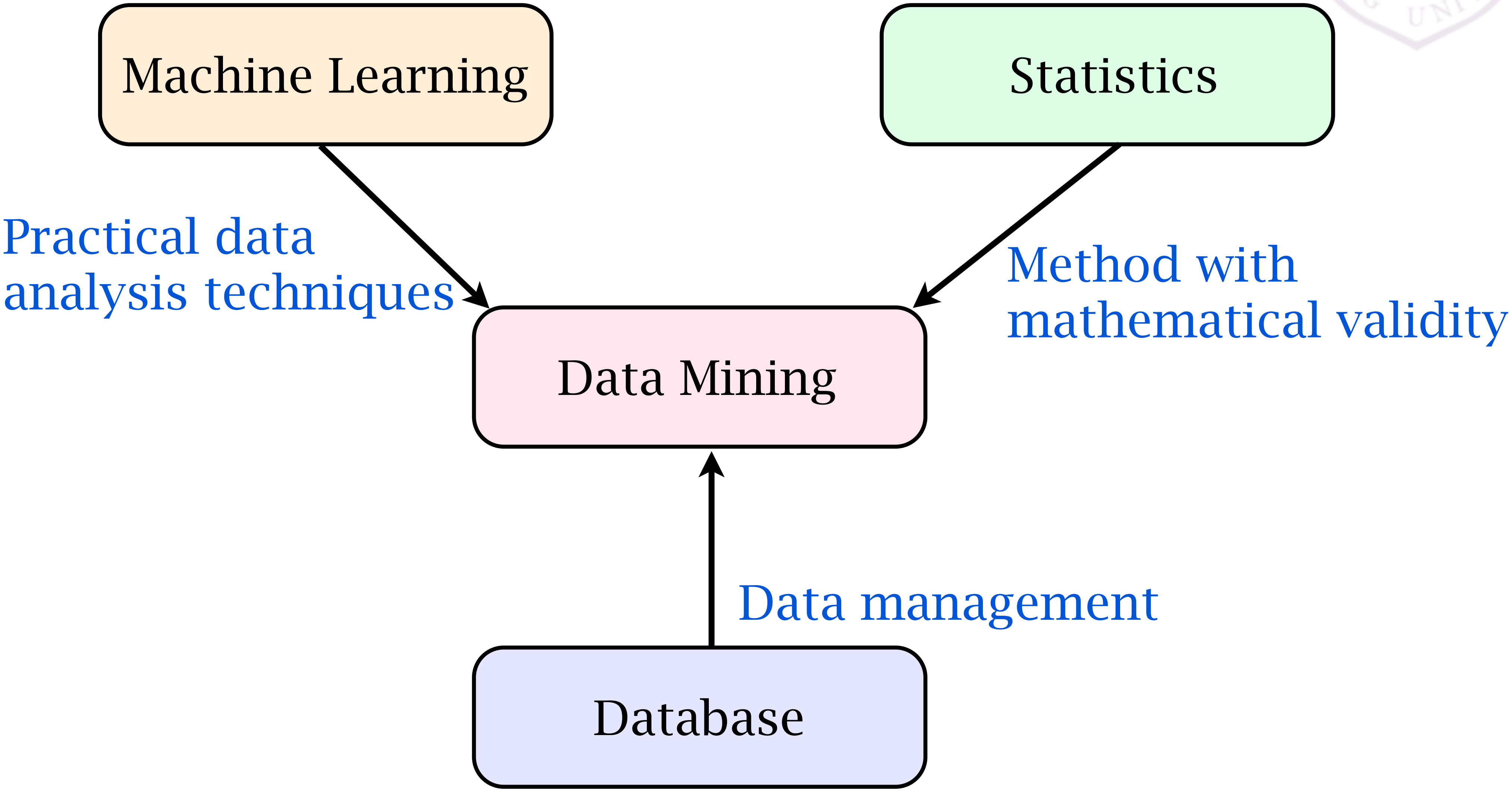
Region	Employment	2013 Avg. Salary	2012 Avg. Salary	% Change	2013 Count
US/Canada	**all**	**128.8**	**113.9**	**13.1%**	**223**
	Comp/Self	131.3	116.8	12.4%	194
	Univ/Gov	112.1	85.9	30.5%	29
Australia/NZ	**all**	108.1	111.8	-3.3%	8
	Comp/Self	112.9	108.3	4.2%	7
	Univ/Gov	75.0	127.5	na	1
W. Europe	**all**	85.1	78.1	8.9%	75
	Comp/Self	90.4	83.8	7.9%	62
	Univ/Gov	59.6	55.6	7.2%	13
Middle East/ Africa	**all**	83.5	96.4	-13.4%	13
	Comp/Self	90.5	105	-13.9%	11
	Univ/Gov	45.0	45	na	2
Latin America	**all**	68.3	43.3	57.7%	12
	Comp/Self	68.8	43.3	58.7%	8
	Univ/Gov	67.5	na	na	4
Asia	**all**	59.8	41.3	44.9%	23
	Comp/Self	63.3	45.2	39.9%	20
	Univ/Gov	36.7	29.4	24.8%	3
E. Europe	**all**	43.9	40.8	7.5%	9
	Comp/Self	47.1	45	4.8%	7
	Univ/Gov	32.5	30.7	5.8%	2
Global	**all**	109.2	96.8	12.8%	363

[KDnuggets Poll, 2013]

Cross-disciplines of data mining



Three perspectives of data mining


Machine Learning
Statistics
Practical data analysis techniques
Method with mathematical validity
Data Mining
Data management
Database

[Z.-H. Zhou, AIJ’03]

习题

为何数据挖掘强调挖掘大数据集?

为何强调数据挖掘结果的可理解性?

数据挖掘是否只处理表格数据?

数据挖掘与统计有哪些区别?

视频

Learning from Data: to section 3

Machine Learning Foundation: to section 2