南京大學 NANJING UNIVERSITY

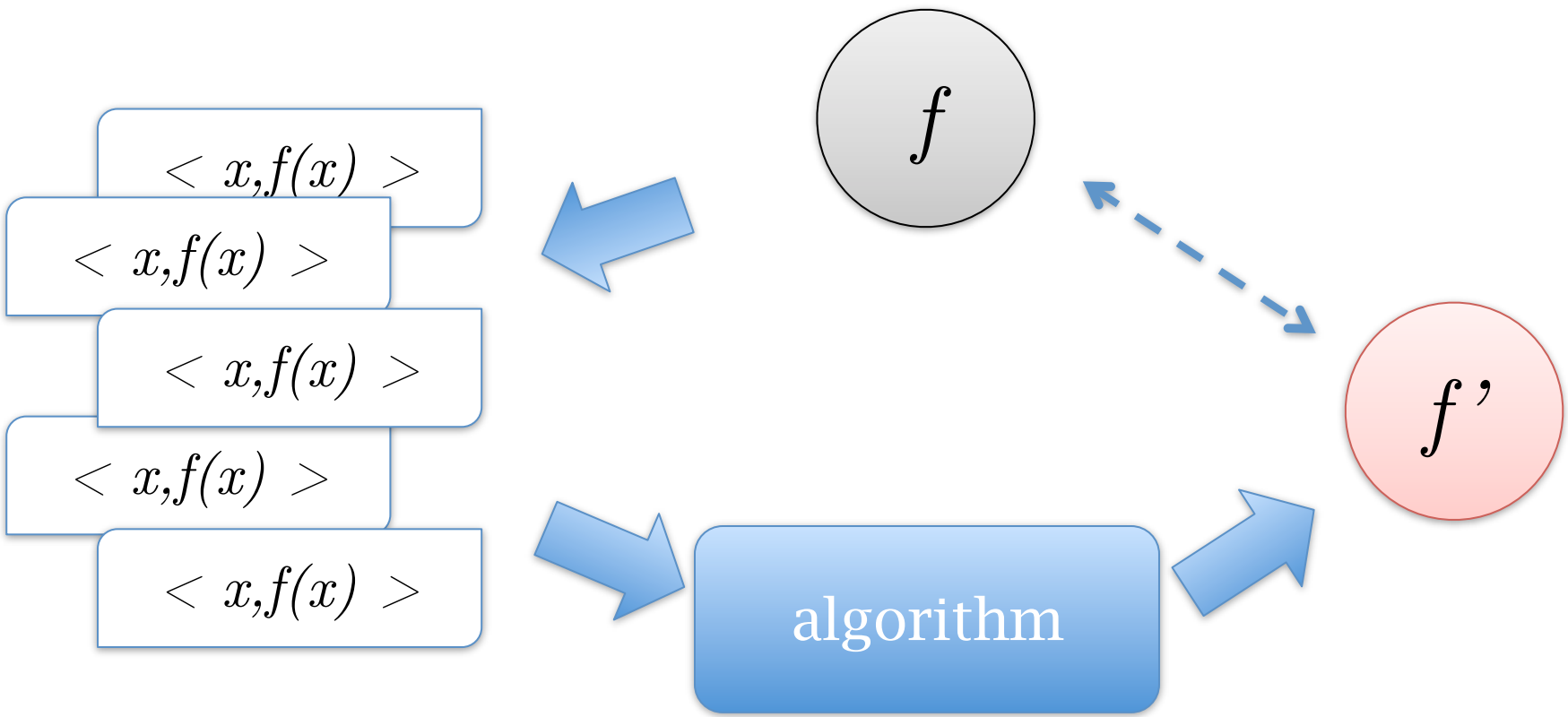
Data Mining for M.Sc. students, CS, Nanjing University

Fall, 2013, Yang Yu

Lecture 1: Introduction

http://cs.nju.edu.cn/yuy/course_dm13ms.ashx

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

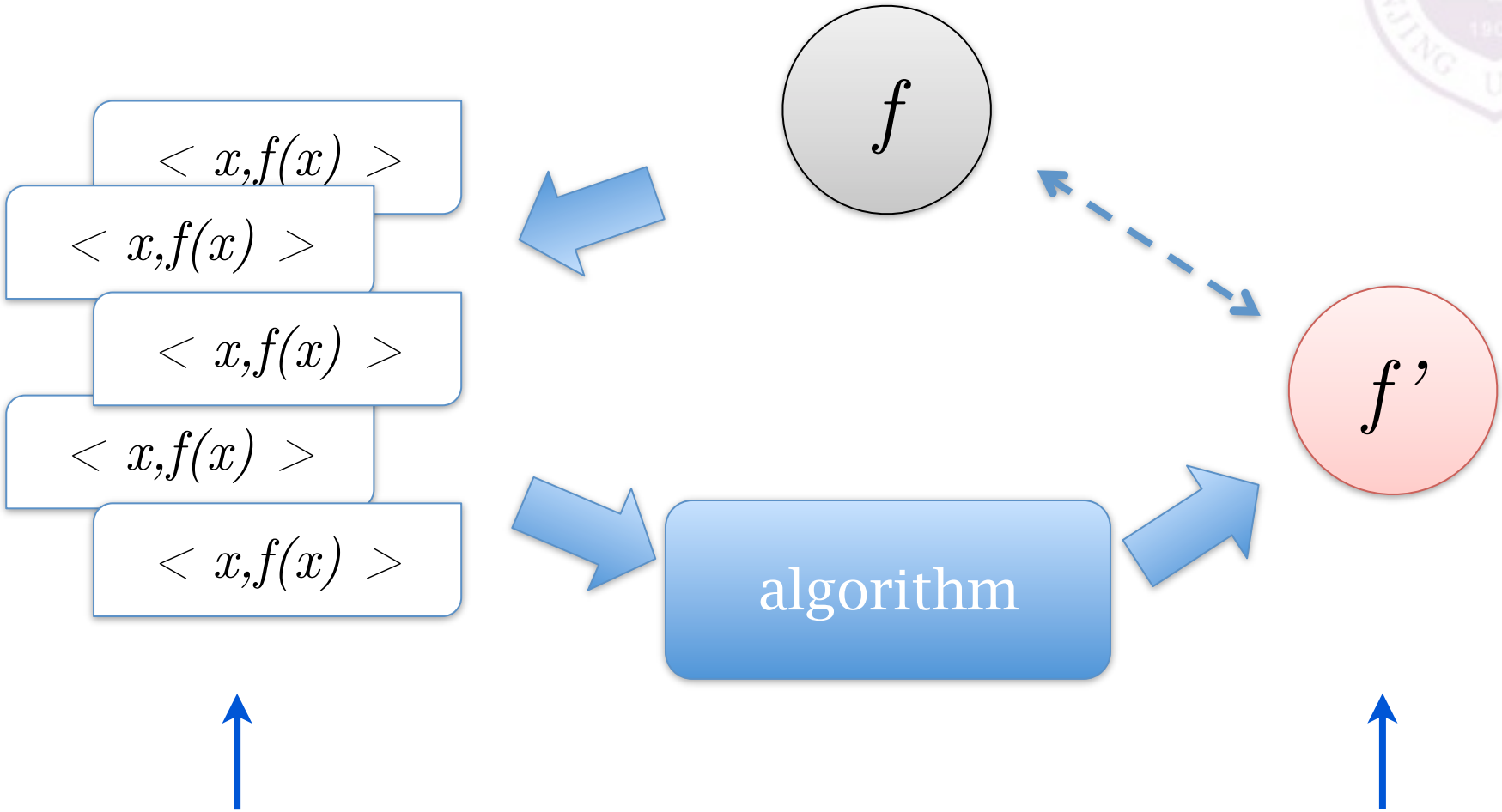
Understandable: decision maker oriented

Useful: mining results should be useful to the users

Observational data v.s. experimental data

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

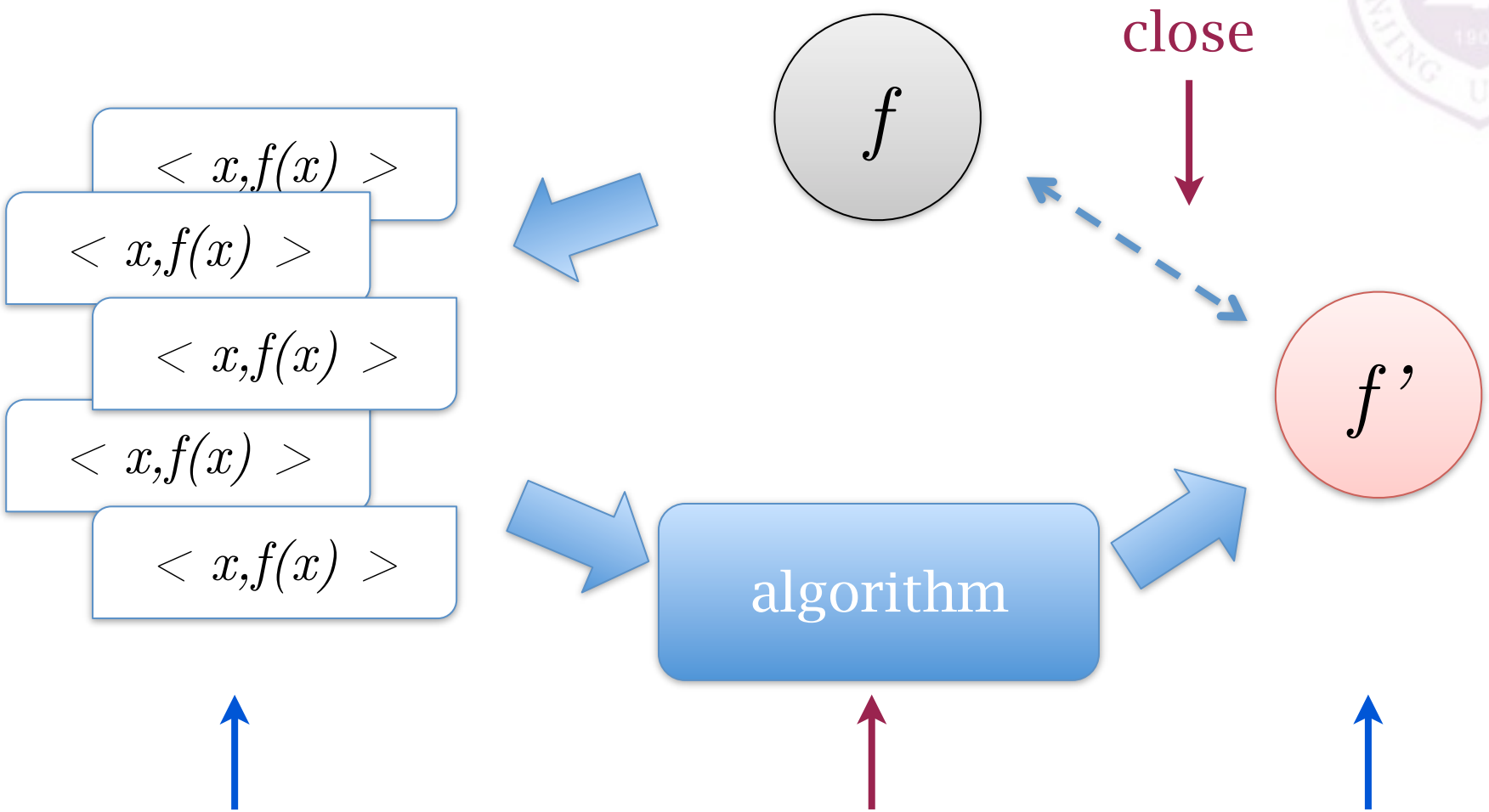
Data mining factors



large
observational

unsuspected
novel
understandable
useful

Data mining factors

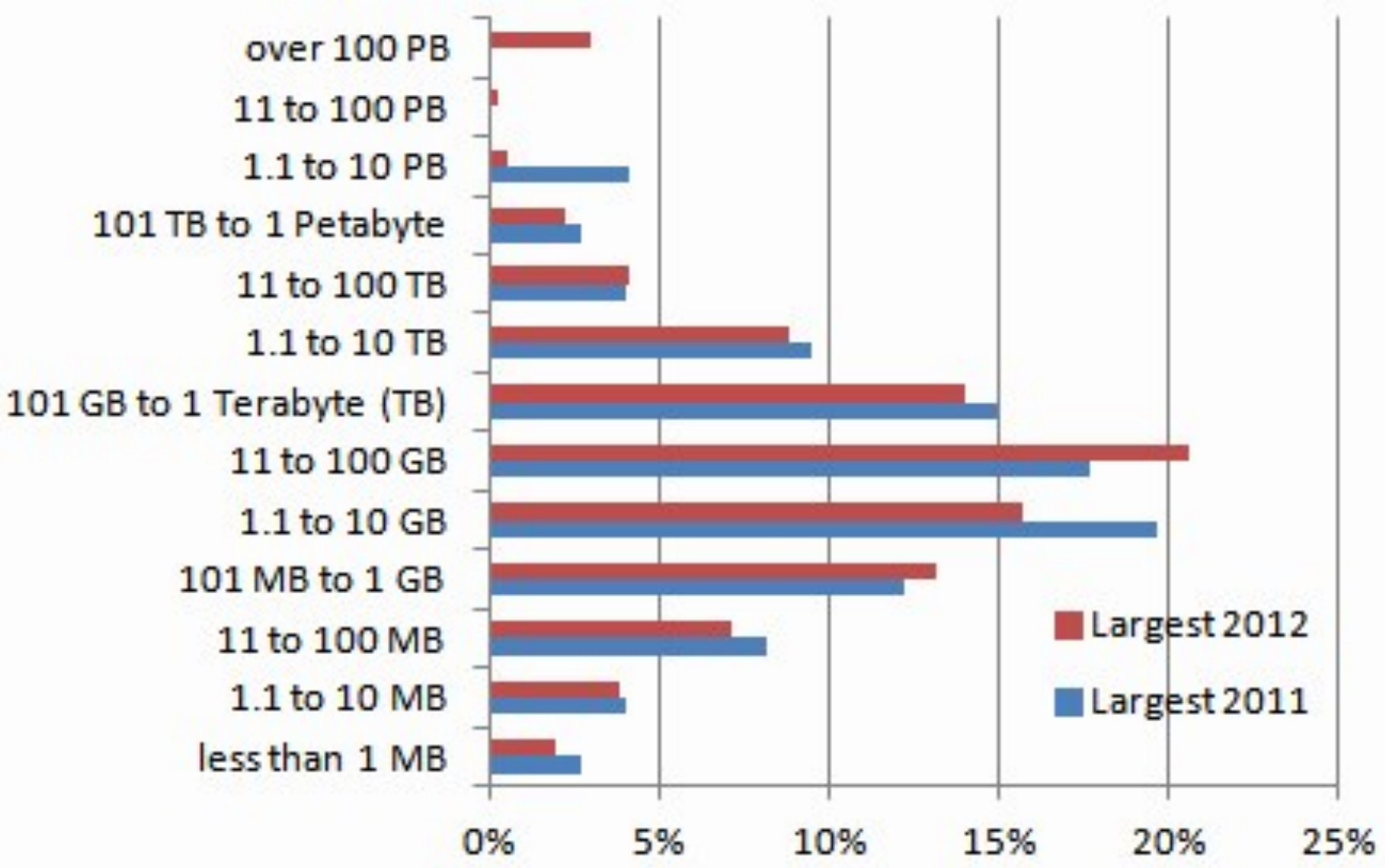


large
observational

efficient
(space & time)

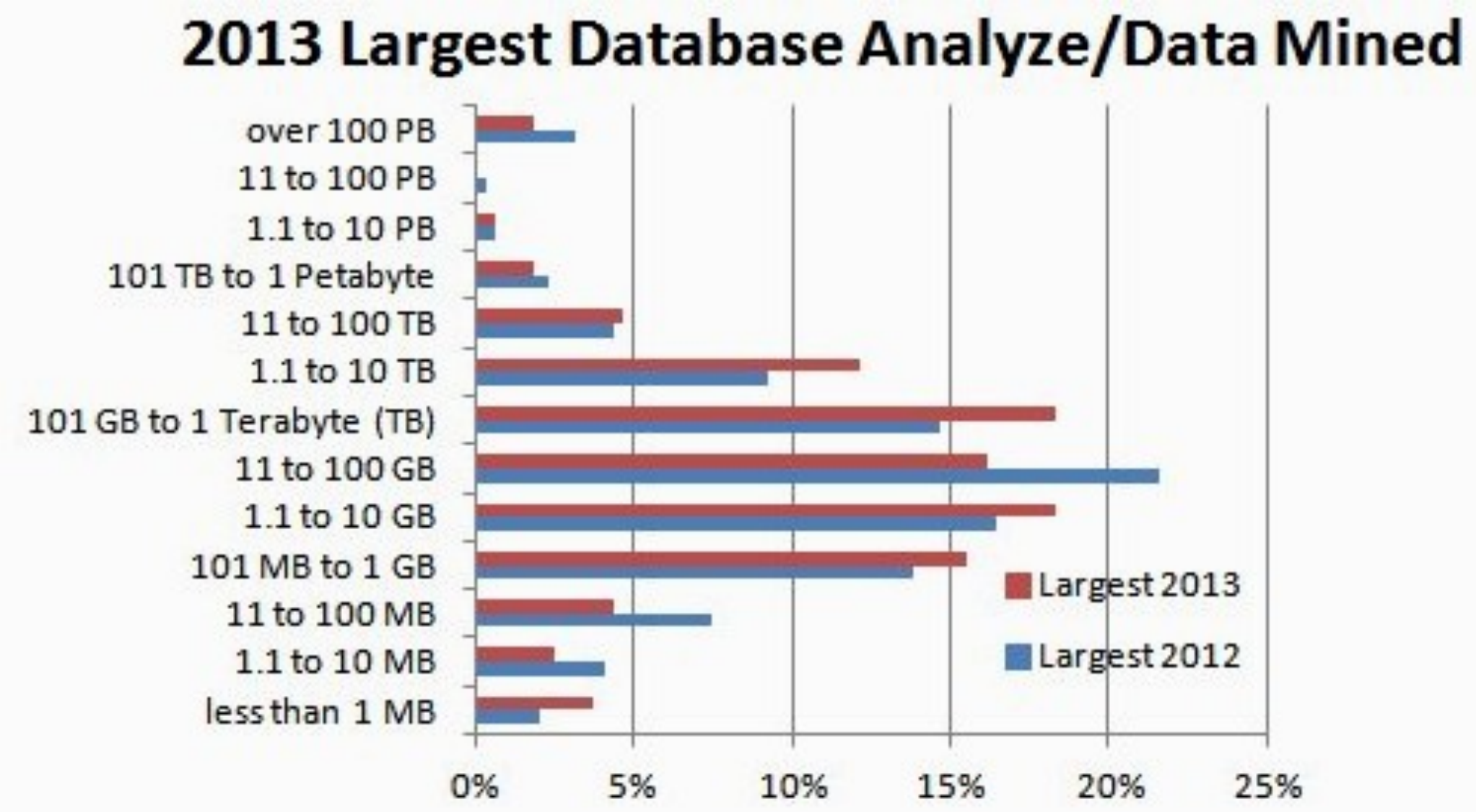
unsuspected
novel
understandable
useful

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

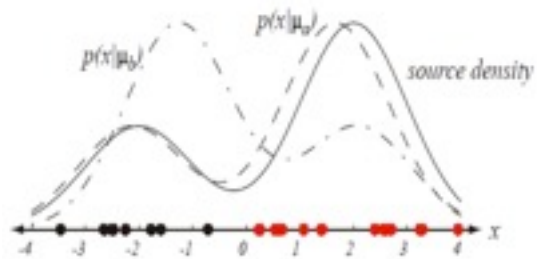
interactive and visualized

how to visualize high dimensional data?

Descriptive modeling

describe a data set

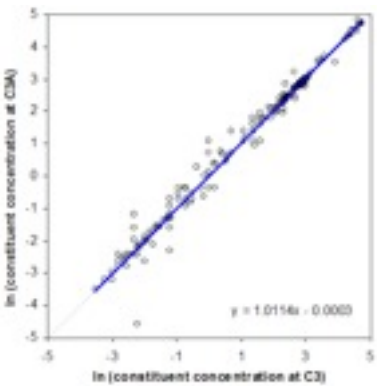
how to characterize general properties of a dataset



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

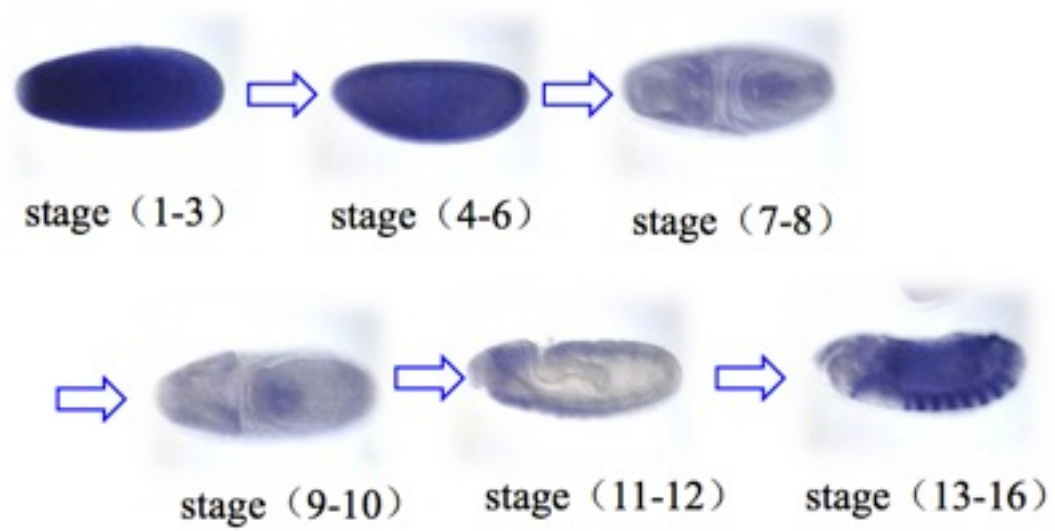
GSM

CDMA

Example: mining network intrusion patterns

recognize intrusion accesses

Example: Mining biology data



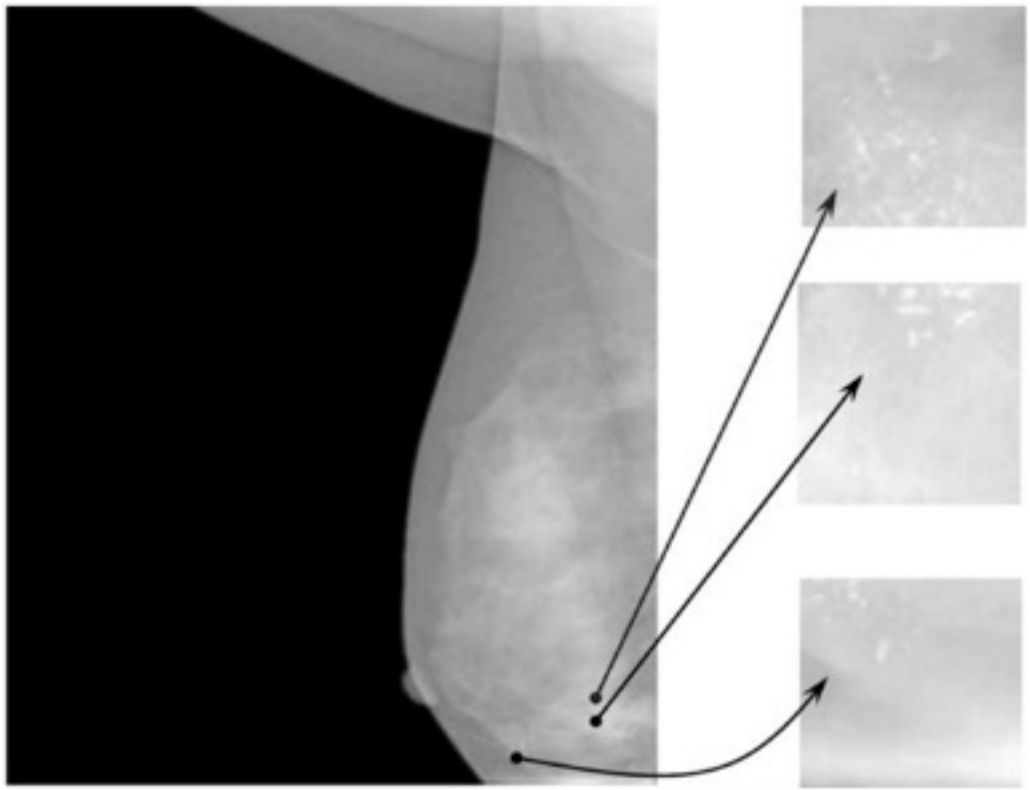
Finding key genes

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

Example: Mining usage data

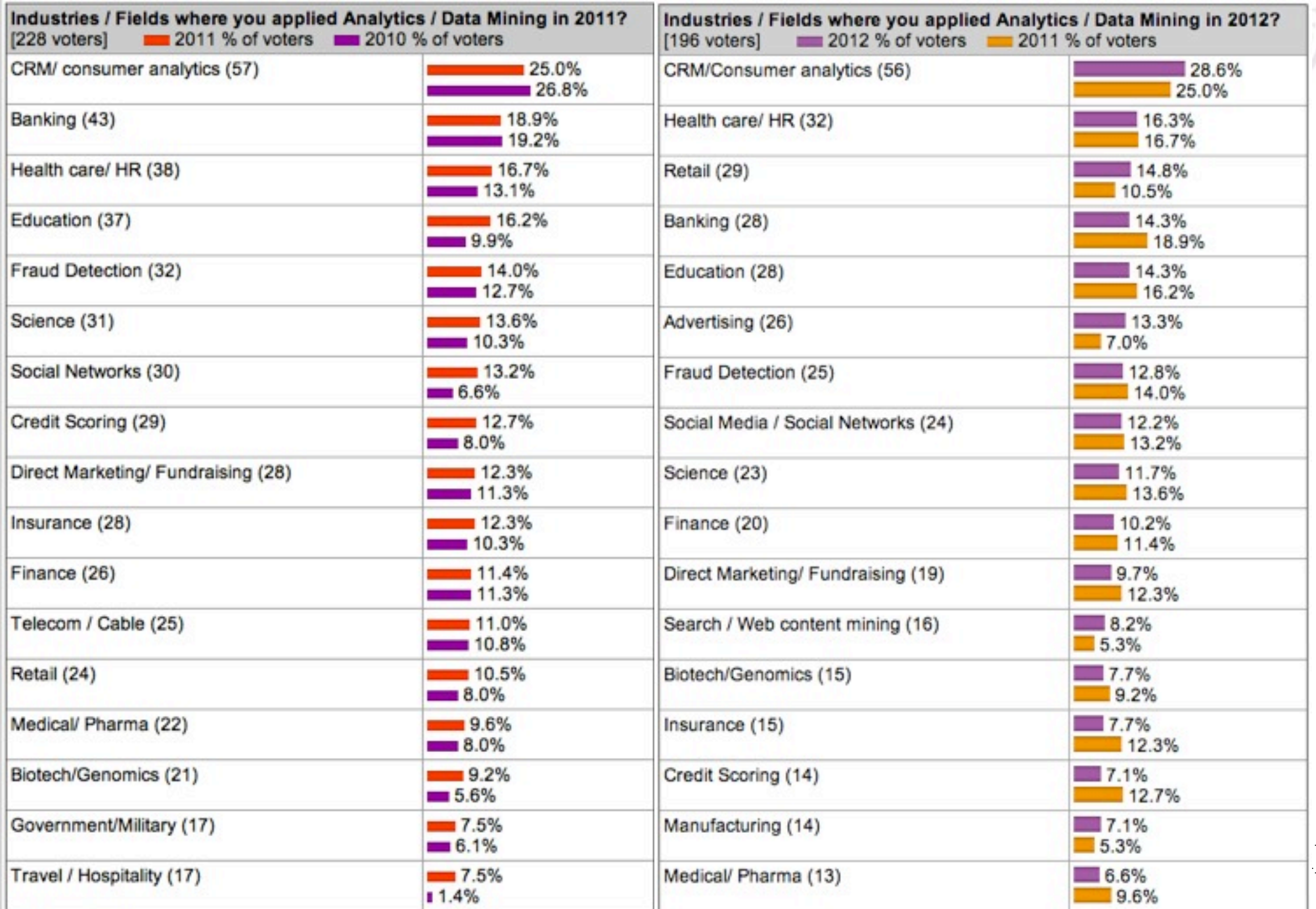
Mining usage data to allow natural human-computer interaction

Top data mining fields

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

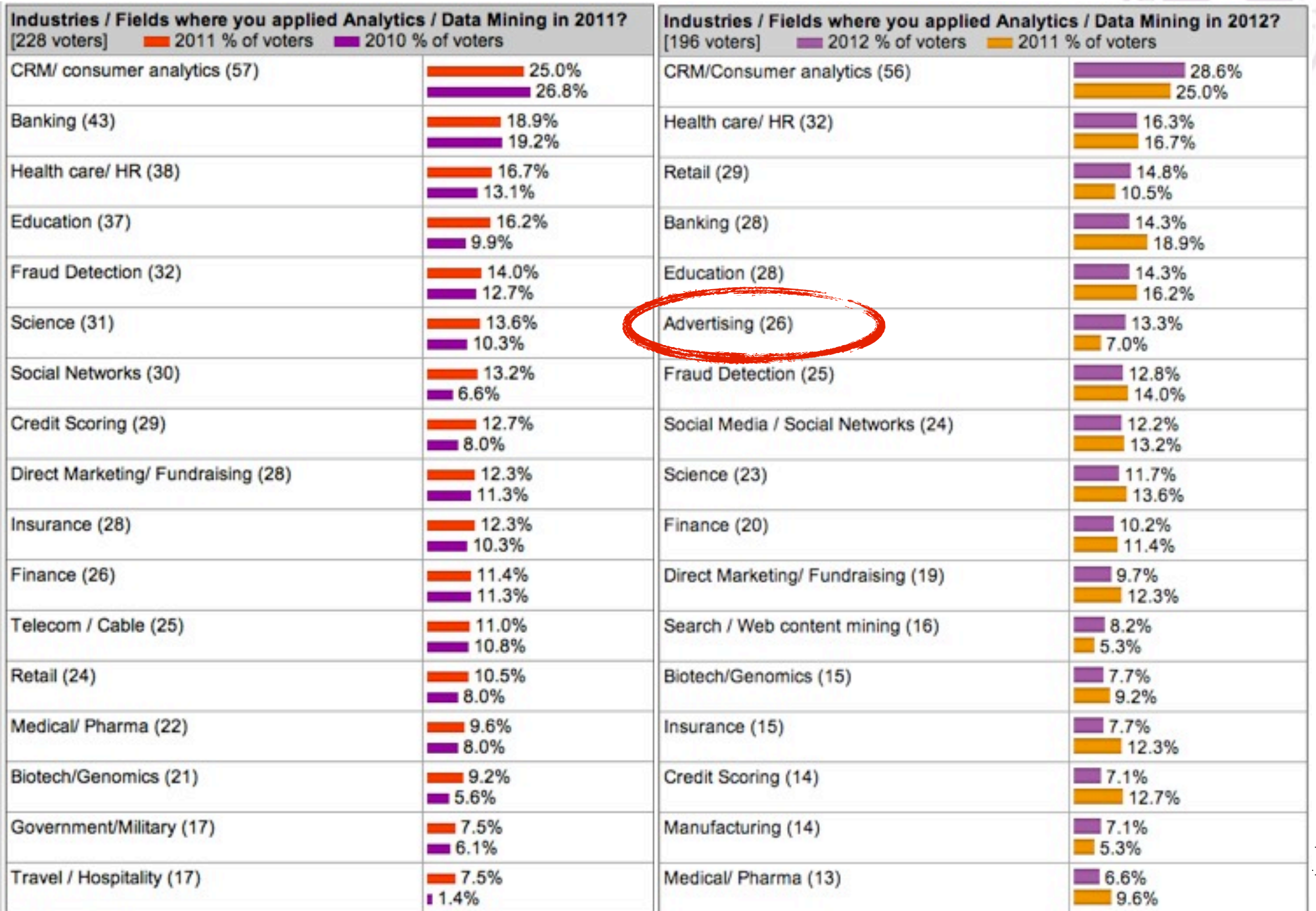
Top data mining fields



Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters] — 2011 % of voters — 2010 % of voters		Industries / Fields where you applied Analytics / Data Mining in 2012? [196 voters] — 2012 % of voters — 2011 % of voters	
CRM/ consumer analytics (57)	25.0% / 26.8%	CRM/Consumer analytics (56)	28.6% / 25.0%
Banking (43)	18.9% / 19.2%	Health care/ HR (32)	16.3% / 16.7%
Health care/ HR (38)	16.7% / 13.1%	Retail (29)	14.8% / 10.5%
Education (37)	16.2% / 9.9%	Banking (28)	14.3% / 18.9%
Fraud Detection (32)	14.0% / 12.7%	Education (28)	14.3% / 16.2%
Science (31)	13.6% / 10.3%	Advertising (26)	13.3% / 7.0%
Social Networks (30)	13.2% / 6.6%	Fraud Detection (25)	12.8% / 14.0%
Credit Scoring (29)	12.7% / 8.0%	Social Media / Social Networks (24)	12.2% / 13.2%
Direct Marketing/ Fundraising (28)	12.3% / 11.3%	Science (23)	11.7% / 13.6%
Insurance (28)	12.3% / 10.3%	Finance (20)	10.2% / 11.4%
Finance (26)	11.4% / 11.3%	Direct Marketing/ Fundraising (19)	9.7% / 12.3%
Telecom / Cable (25)	11.0% / 10.8%	Search / Web content mining (16)	8.2% / 5.3%
Retail (24)	10.5% / 8.0%	Biotech/Genomics (15)	7.7% / 9.2%
Medical/ Pharma (22)	9.6% / 8.0%	Insurance (15)	7.7% / 12.3%
Biotech/Genomics (21)	9.2% / 5.6%	Credit Scoring (14)	7.1% / 12.7%
Government/Military (17)	7.5% / 6.1%	Manufacturing (14)	7.1% / 5.3%
Travel / Hospitality (17)	7.5% / 1.4%	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		Industries / Fields where you applied Analytics / Data Mining in 2012? [196 voters] — 2012 % of voters — 2011 % of voters	
CRM/ consumer analytics (57)	25.0% 26.8%	CRM/Consumer analytics (56)	28.6% 25.0%
Banking (43)	18.9% 19.2%	Health care/ HR (32)	16.3% 16.7%
Health care/ HR (38)	16.7% 13.1%	Retail (29)	14.8% 10.5%
Education (37)	16.2% 9.9%	Banking (28)	14.3% 18.9%
Fraud Detection (32)	14.0% 12.7%	Education (28)	14.3% 16.2%
Science (31)	13.6% 10.3%	Advertising (26)	13.3% 7.0%
Social Networks (30)	13.2% 6.6%	Fraud Detection (25)	12.8% 14.0%
Credit Scoring (29)	12.7% 8.0%	Social Media / Social Networks (24)	12.2% 13.2%
Direct Marketing/ Fundraising (28)	12.3% 11.3%	Science (23)	11.7% 13.6%
Insurance (28)	12.3% 10.3%	Finance (20)	10.2% 11.4%
Finance (26)	11.4% 11.3%	Direct Marketing/ Fundraising (19)	9.7% 12.3%
Telecom / Cable (25)	11.0% 10.8%	Search / Web content mining (16)	8.2% 5.3%
Retail (24)	10.5% 8.0%	Biotech/Genomics (15)	7.7% 9.2%
Medical/ Pharma (22)	9.6% 8.0%	Insurance (15)	7.7% 12.3%
Biotech/Genomics (21)	9.2% 5.6%	Credit Scoring (14)	7.1% 12.7%
Government/Military (17)	7.5% 6.1%	Manufacturing (14)	7.1% 5.3%
Travel / Hospitality (17)	7.5% 1.4%	Medical/ Pharma (13)	6.6% 9.6%

]

Data types in mining tasks

"Flat" data: vectors and matrix

Show 10 entries Search:

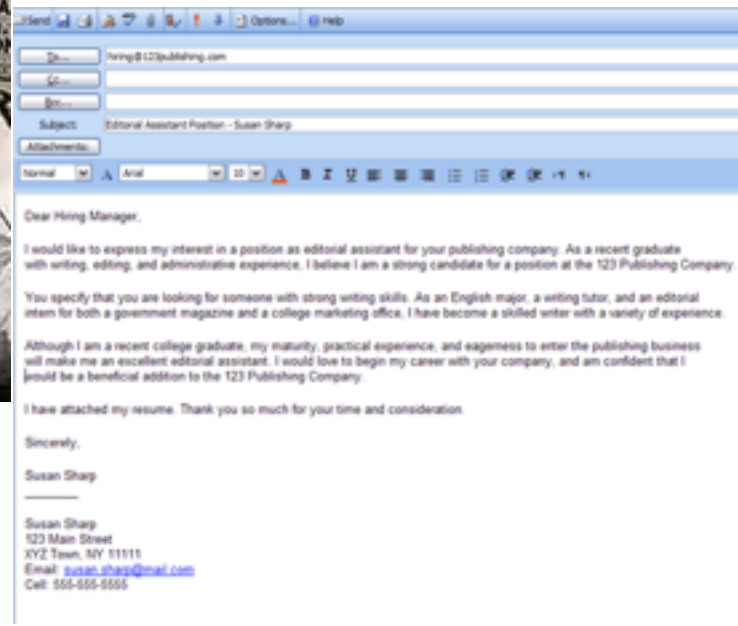
id	words	fog	kincaid	flesch	angel	animal	aristocracy	art	astronomy	beauty	being	cause	chance	change	citizen	constitution
aeschylus-agamemnon-1860	14951	8	6	80	0	0	0	3	0	0	118	1	2	0	0	0
aeschylus-persians-1782	8372	14	11	62	0	0	0	0	0	2	0	1	0	0	0	0
aeschylus-prometheus-2549	10070	10	8	68	0	0	0	19	0	0	194	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

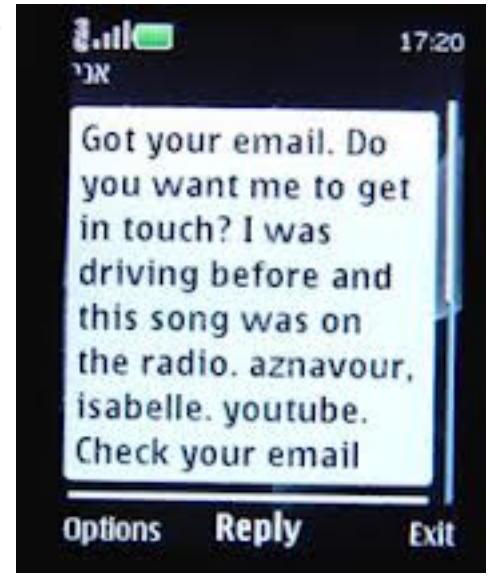
Showing 1 to 10 of 222 entries

First Previous 1 2 3 4 5 Next Last

Data types in mining tasks

Text data





Data types in mining tasks

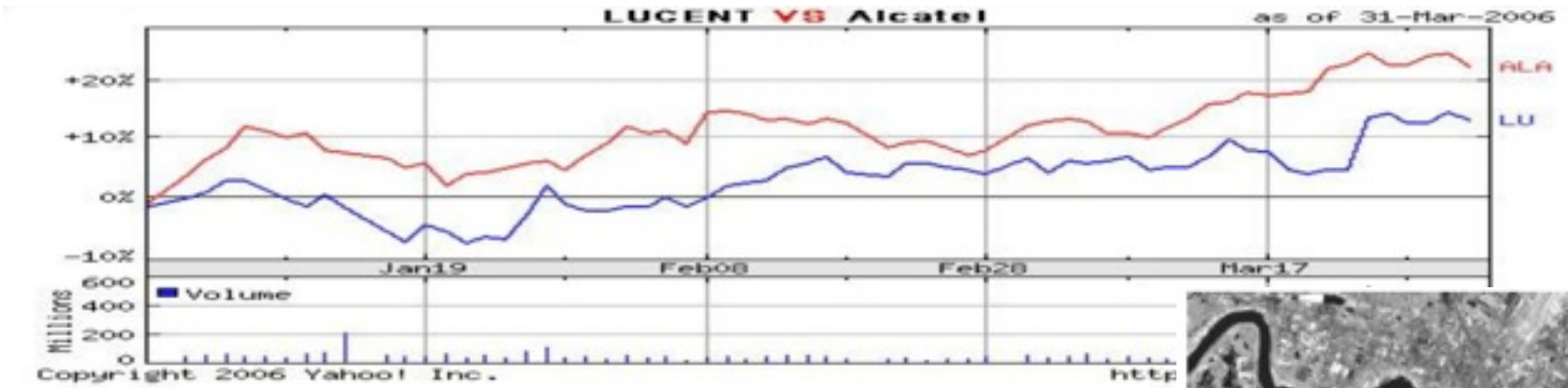
Structured data

Data types in mining tasks

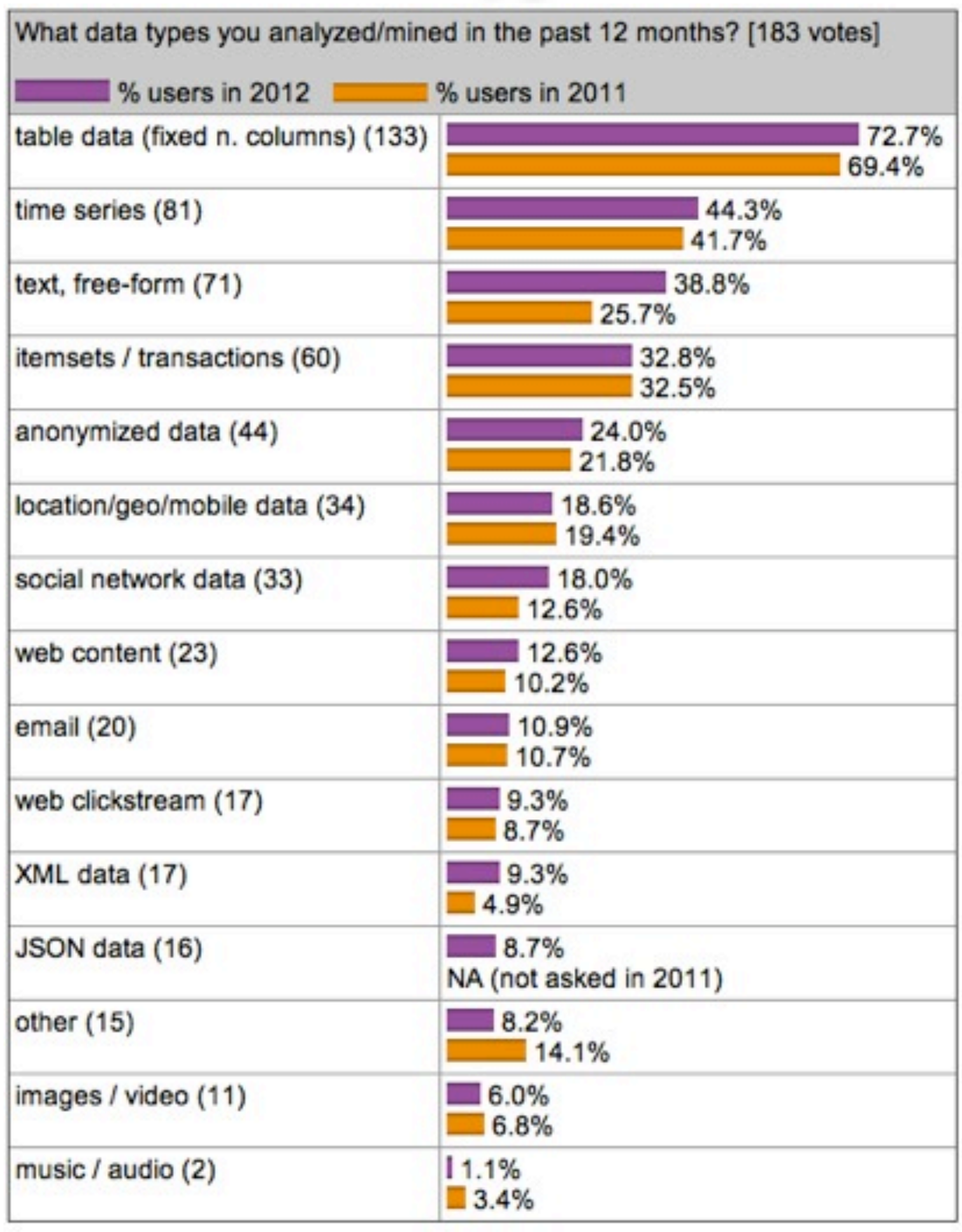
Multi-media data

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):

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):

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):

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):

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):

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):

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):

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

Who needs data mining

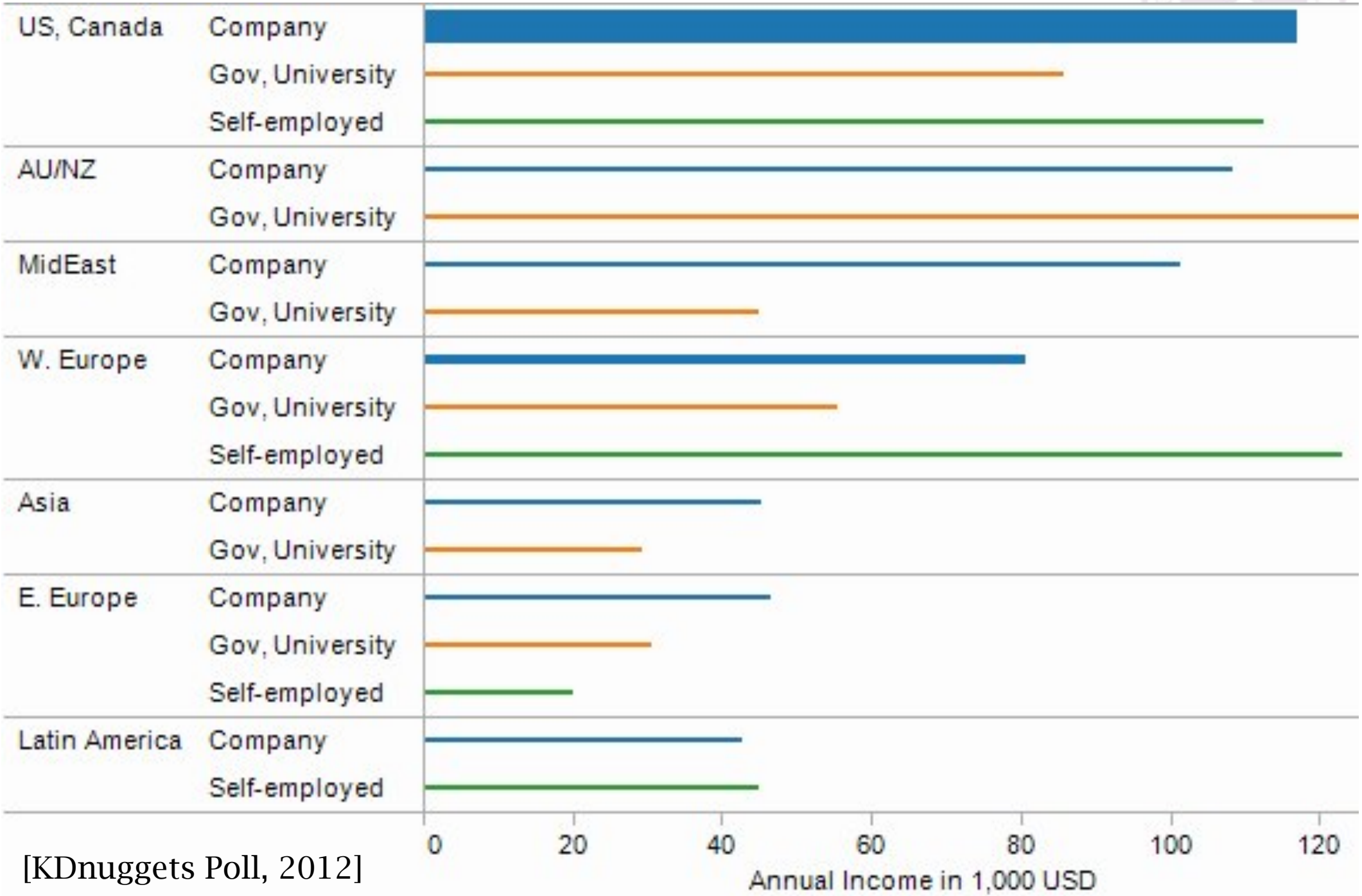
Sponsors of ACM SIGKDD 2012 (Companies):

Adobe

Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):

Annual salary of data miners



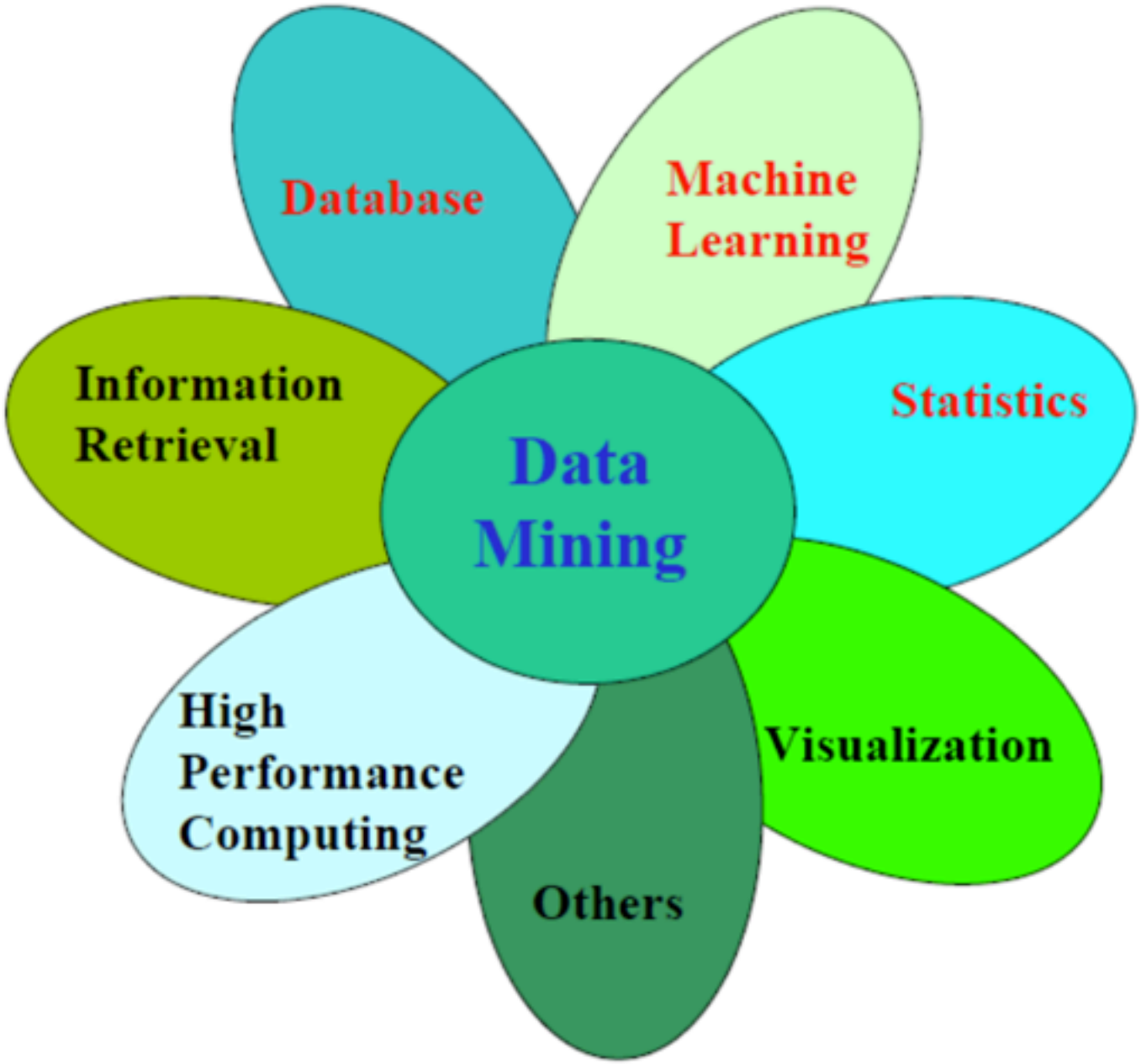
[KDnuggets Poll, 2012]

Annual salary o

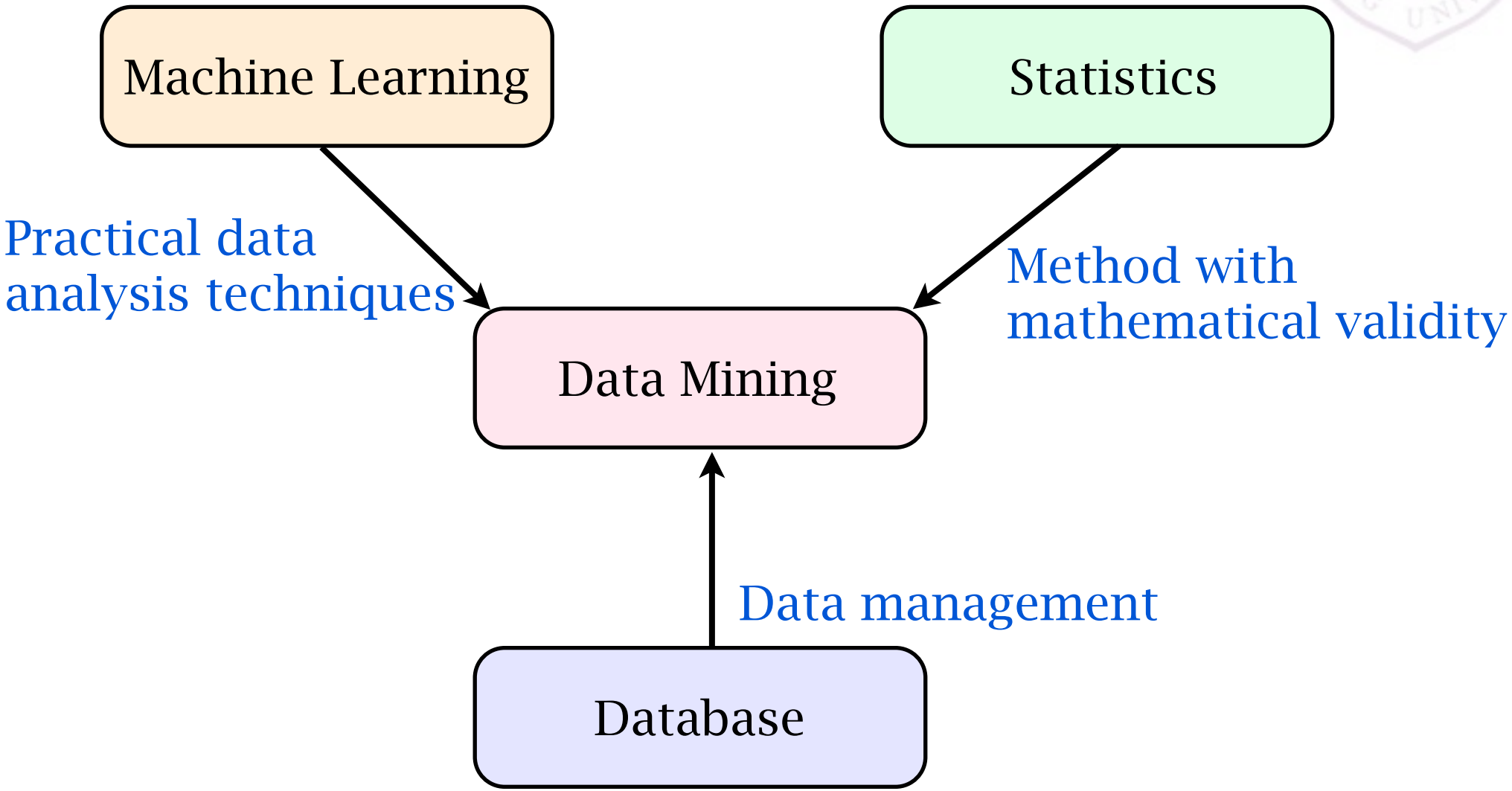
Region	Employment	2013 Avg. Salary	2012 Avg. Salary	% Change	2013 Count
US/Canada	**all**	**128.8**	**113.9**	**13.1%**	**223**
	Comp/Self	131.3	116.8	12.4%	194
	Univ/Gov	112.1	85.9	30.5%	29
Australia/NZ	**all**	108.1	111.8	-3.3%	8
	Comp/Self	112.9	108.3	4.2%	7
	Univ/Gov	75.0	127.5	na	1
W. Europe	**all**	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



[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?