Algorithm Optimization for Data Mining in Oil and Gas Exploration and Development

油气勘探开发常用数据挖掘算法优选

李大伟

中国石油勘探开发研究院

PetroChina

OUTLINE

1. Background
2. Big data in petroleum industry
3. Data mining in petroleum industry
4. Case study
5. Conclusions

1 Background

(1) Self-introduction

Dawei Li (English name: David Lee)

- Petroleum geologist
- Got doctoral degree from China University of Geosciences (1996)
- Worked as a post-doctor in PetroChina (2001-2003)
- Managers of some large IT systems (2004-Present)
- Published about 50 technical papers and 3 books

PetroChina

1 Background

(2) E&P informatization

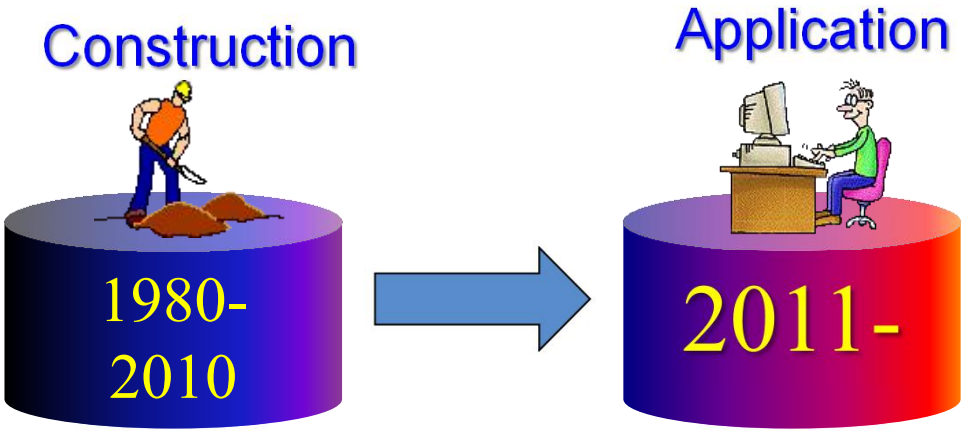
As the upstream of petroleum industry, E&P plays a very important role in the whole business flow of oil companies. Most data (about 90%, seismic, logging, etc.) of petroleum industry are in E&P.

By E&P informatization, management and application levels of E&P data can be improved greatly, and the efficiency of E&P research etc. can also be advanced efficiently.

Thus E&P informatization is very important to oil companies.

1 Background

(3) Informatization of PetroChina



Guided by four 5-years' IT plans (2000 → 2005 → 2010→ 2015→), PetroChina informatization has changed from construction stage to application stage since 2010.

For example:

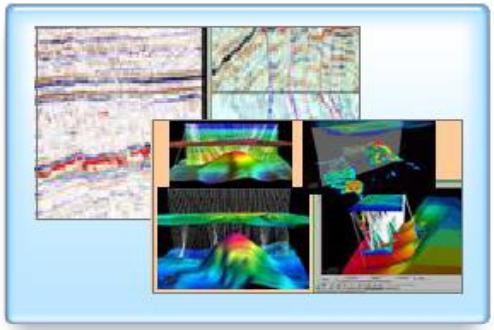
An oilfield of PetroChina: It has built a large E&P data center, including 18 types of E&P databases, about 100000 wells, 40 TB data, providing 80% data and more than 30% efficiency increase for E&P research.

PetroChina

OUTLINE

1. Background
2. Big data in petroleum industry
3. Data mining in petroleum industry
4. Case study
5. Conclusions

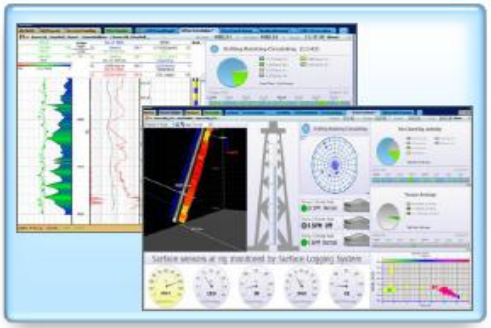
2 Big data in petroleum industry



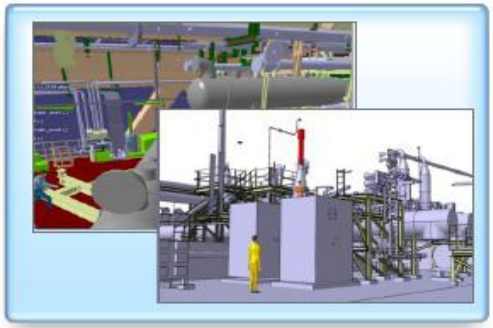
Seismic and geology modeling data

Production data

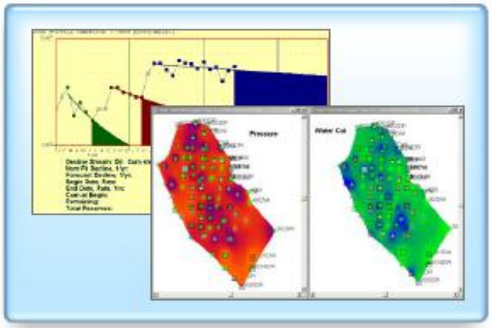
Monitoring data



Well logging data



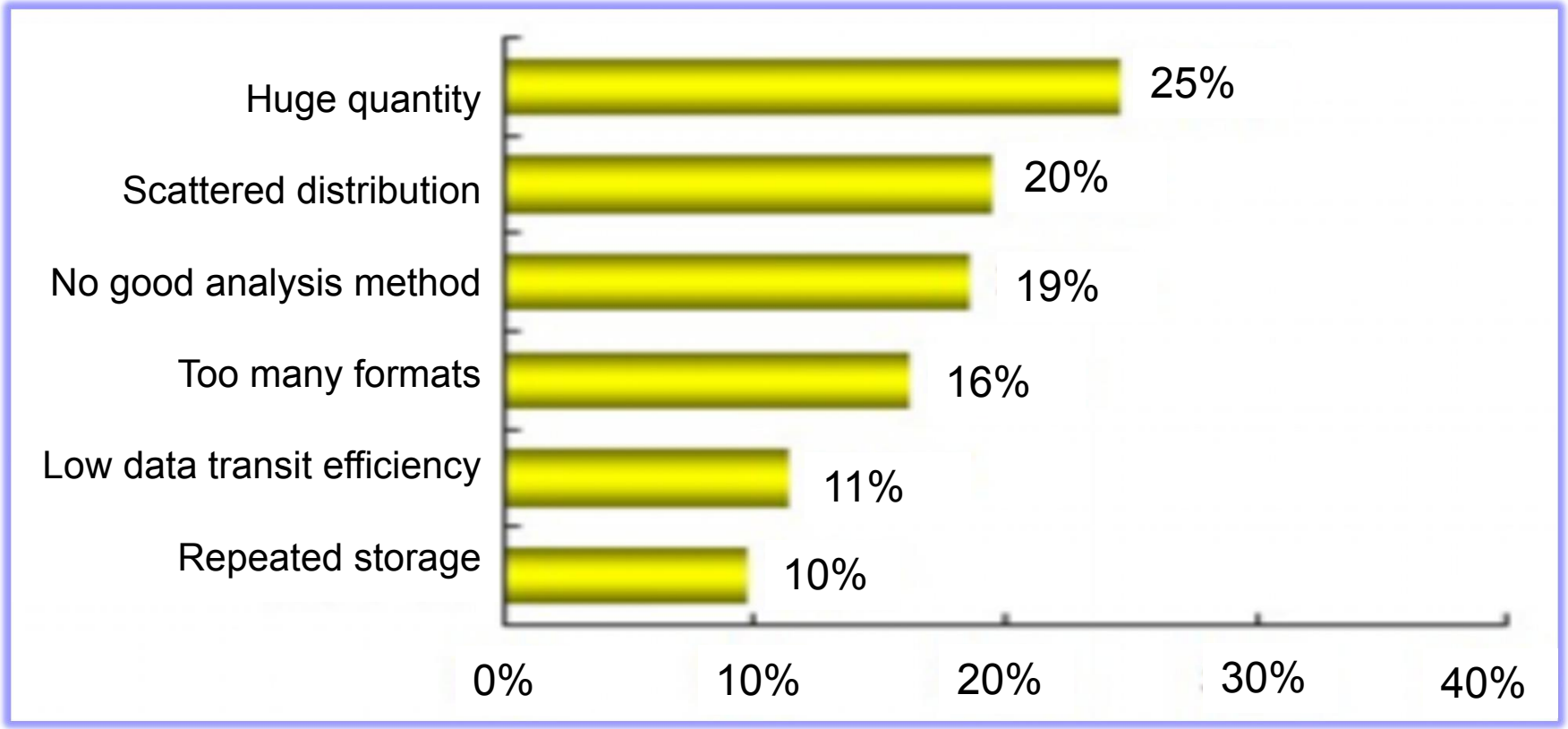
Refining data



Numerical simulation data

Major data sources in petroleum industry

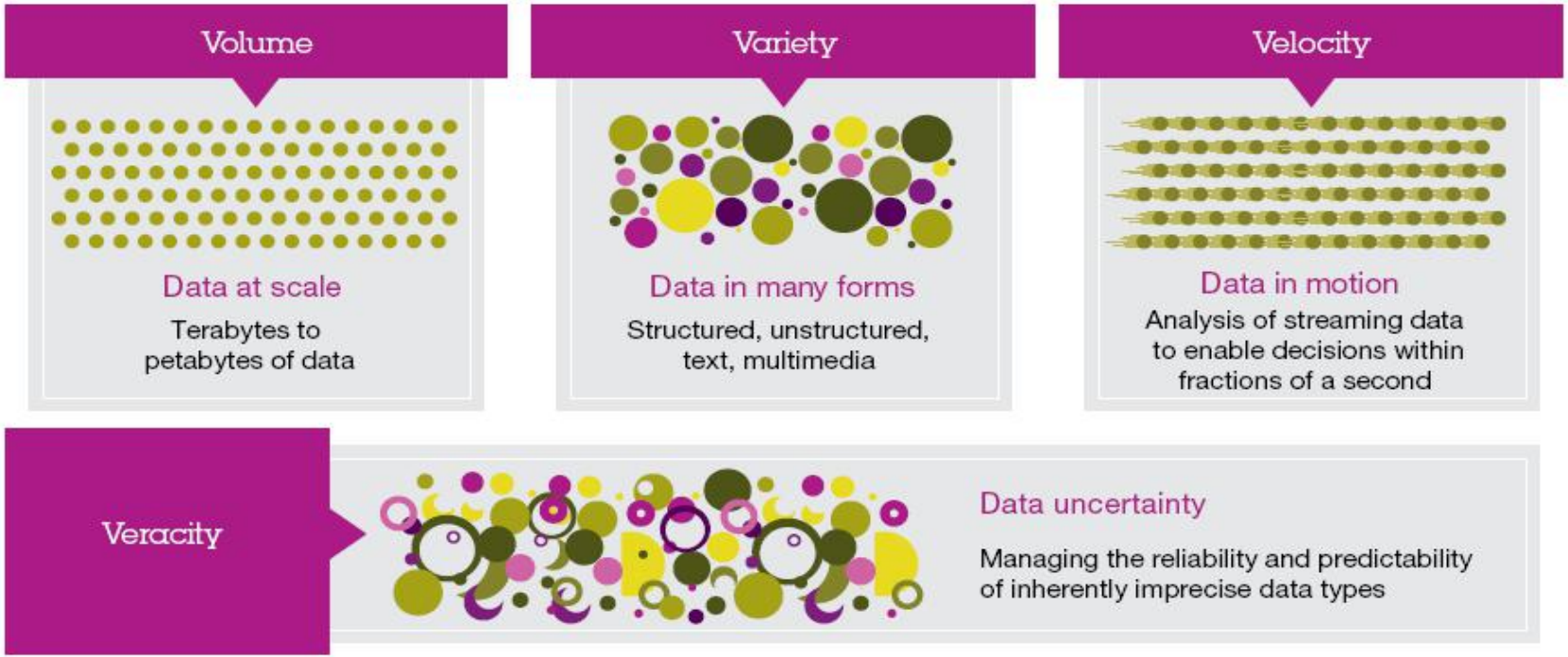
2 Big data in petroleum industry



Major data problems in petroleum industry

2 Big data in petroleum industry

According to the "4Vs" of big data by IBM, petroleum industry has already entered "Big Data" era.



Example: a 3D seismic survey data in a block of Africa owned by PetroChina has 40 000 GB.

2 Big data in petroleum industry

- PetroChina
  - 70 large IT systems
  - About 2000 TB data in DBs
  - Applications: data query, download, simple calculations and reports etc.
  - The deep value of data has not been fully utilized

12th five-year IT plan of PetroChina (after confidential treatment)

PetroChina

2 Big data in petroleum industry

- We have invested much in petroleum informatization.
- How to fully utilize the value of the IT systems and data assets has become a critical problem for oil companies.

There are many solutions for this problem.

Data mining (DM) is one of the good solutions.

PetroChina

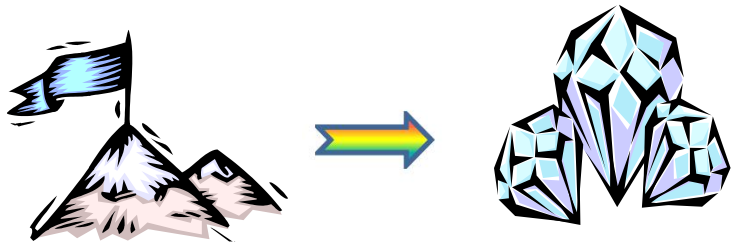
OUTLINE

1. Background
2. Big data in petroleum industry
3. Data mining in petroleum industry
4. Case study
5. Conclusions

3 DM in petroleum industry

Data mining (DM) is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems (Wikipedia).

数据挖掘是在大型数据集中发现模式的计算过程，涉及机器学习、统计和数据库系统。

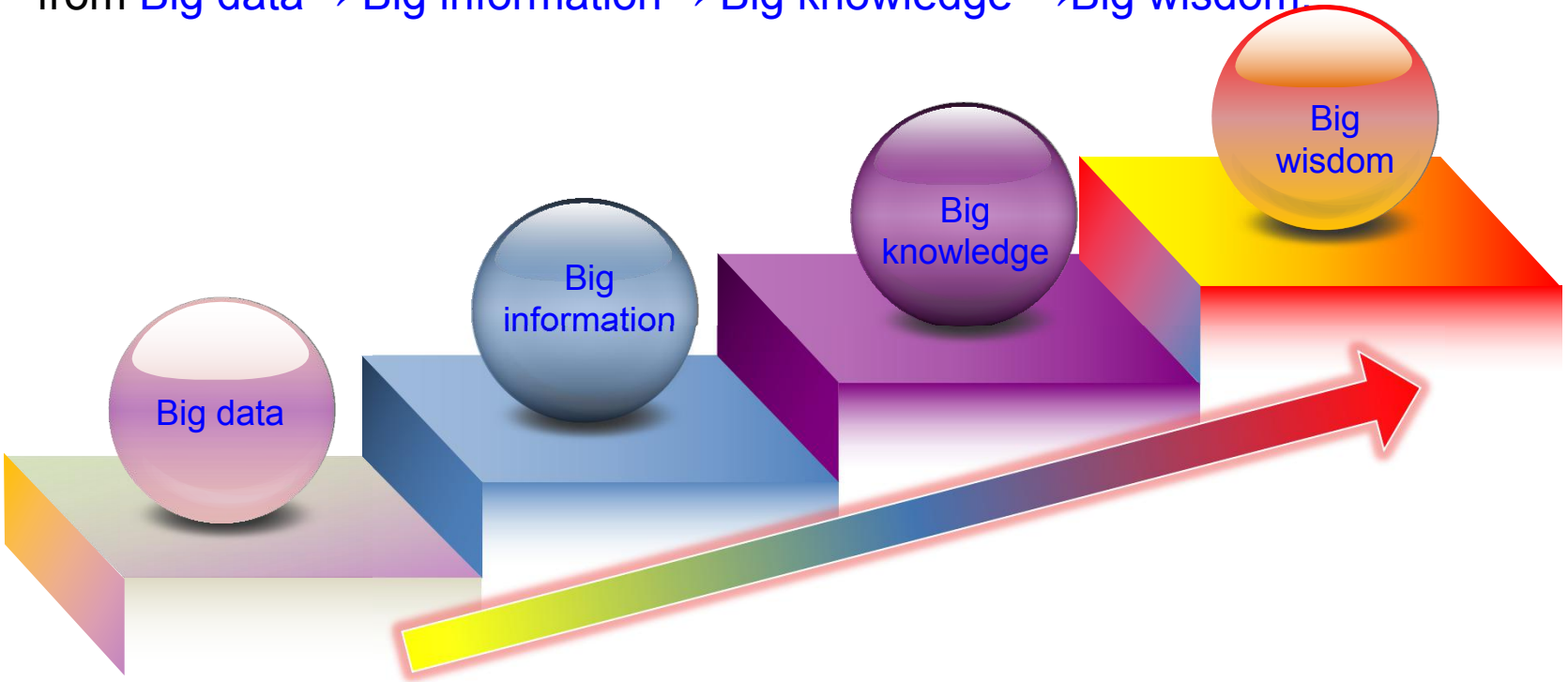


3 DM in petroleum industry

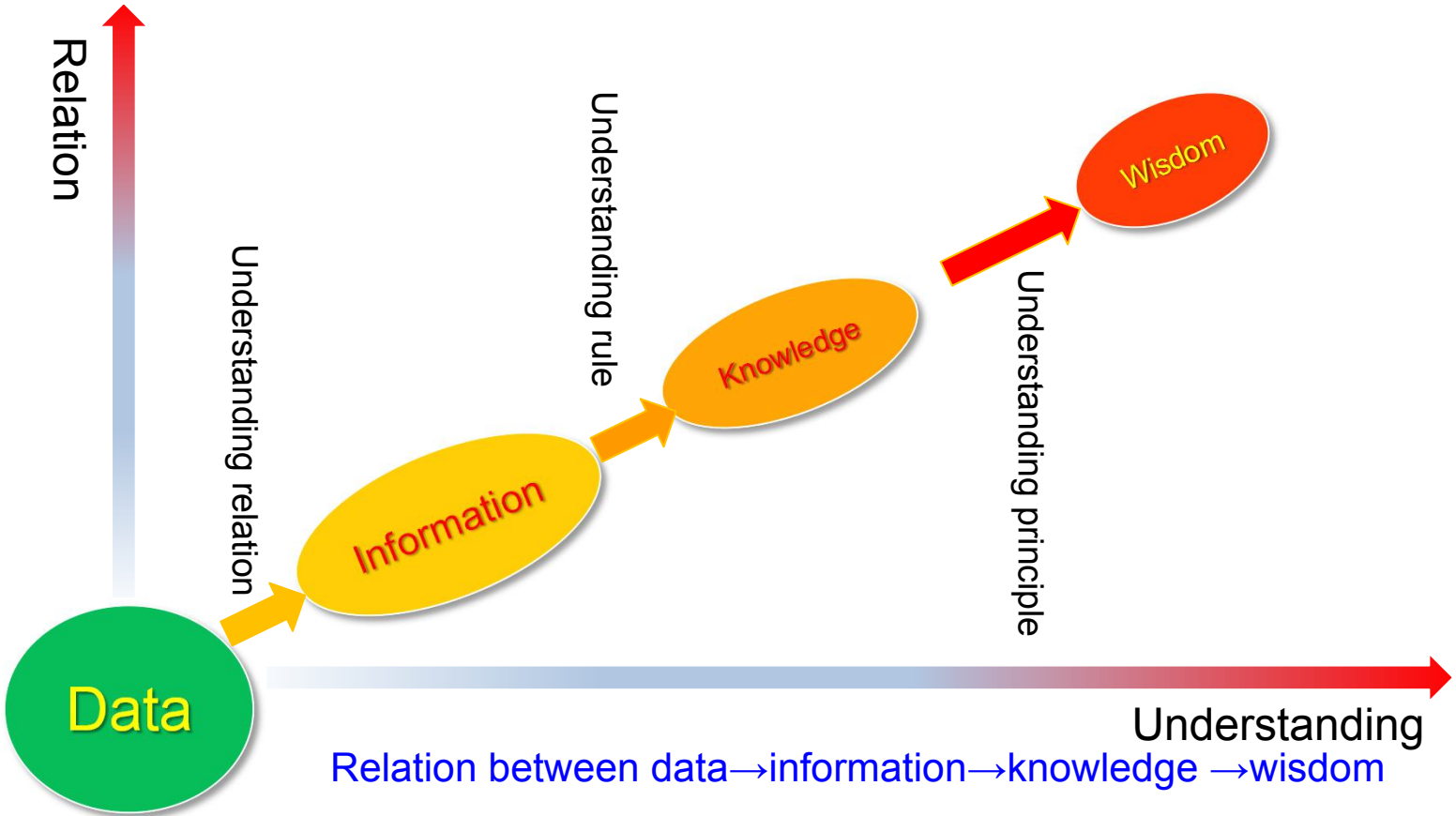
DM objects：Pure data, txt, graph, spatial data, audio data, video data, web data etc.

3 DM in petroleum industry

DM can fully utilize the deep value of data assets, and achieve leaps from Big data→ Big information→ Big knowledge →Big wisdom.

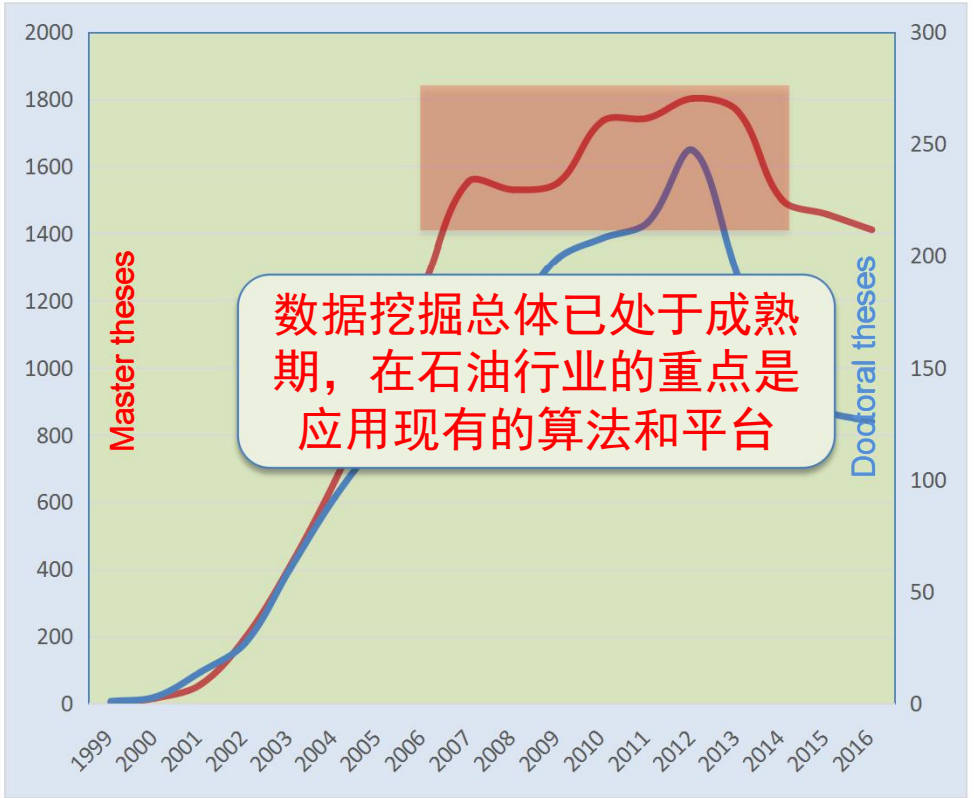


3 DM in petroleum industry



3 DM in petroleum industry

Statistics of published papers in various sectors related to data mining in Chinese (Dawei Li, 2016)

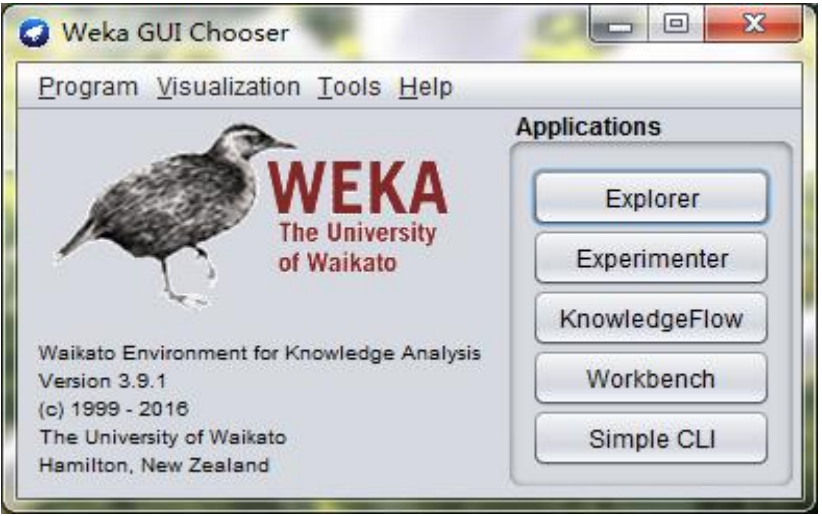


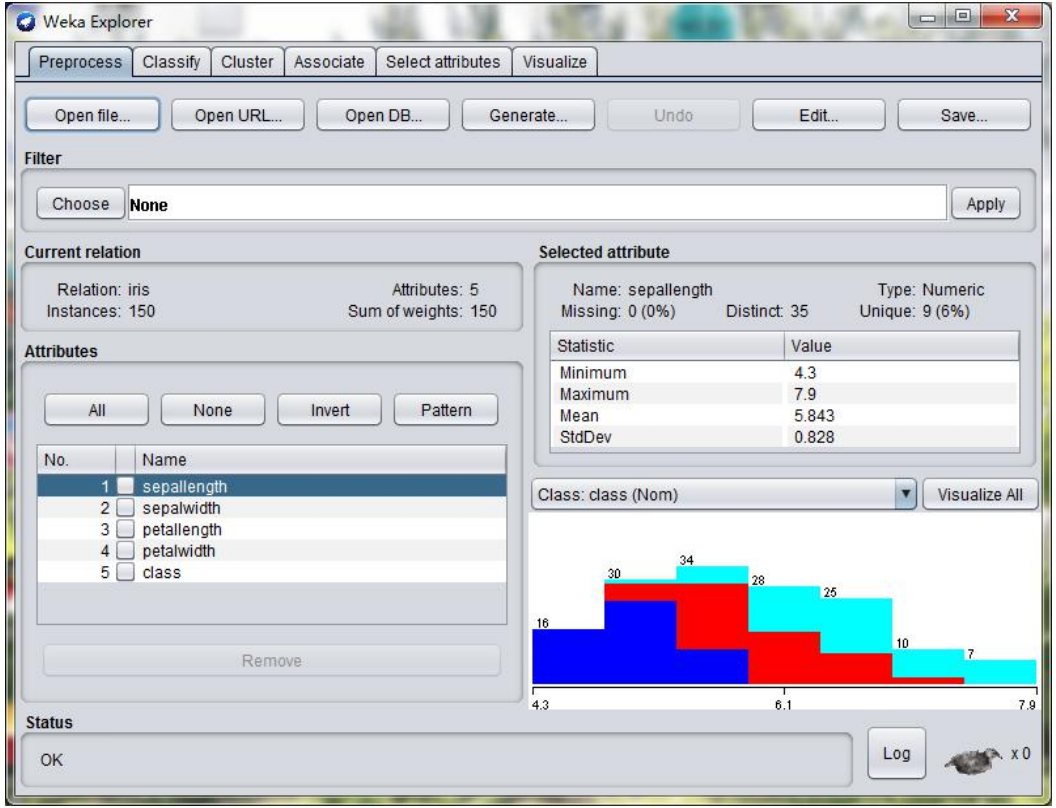
Statistics of master and doctoral theses related to data mining in Chinese

PetroChina

3 DM in petroleum industry

常用数据挖掘平台：Weka, SPSS, Rapid Miner, Matlab, TipDM…

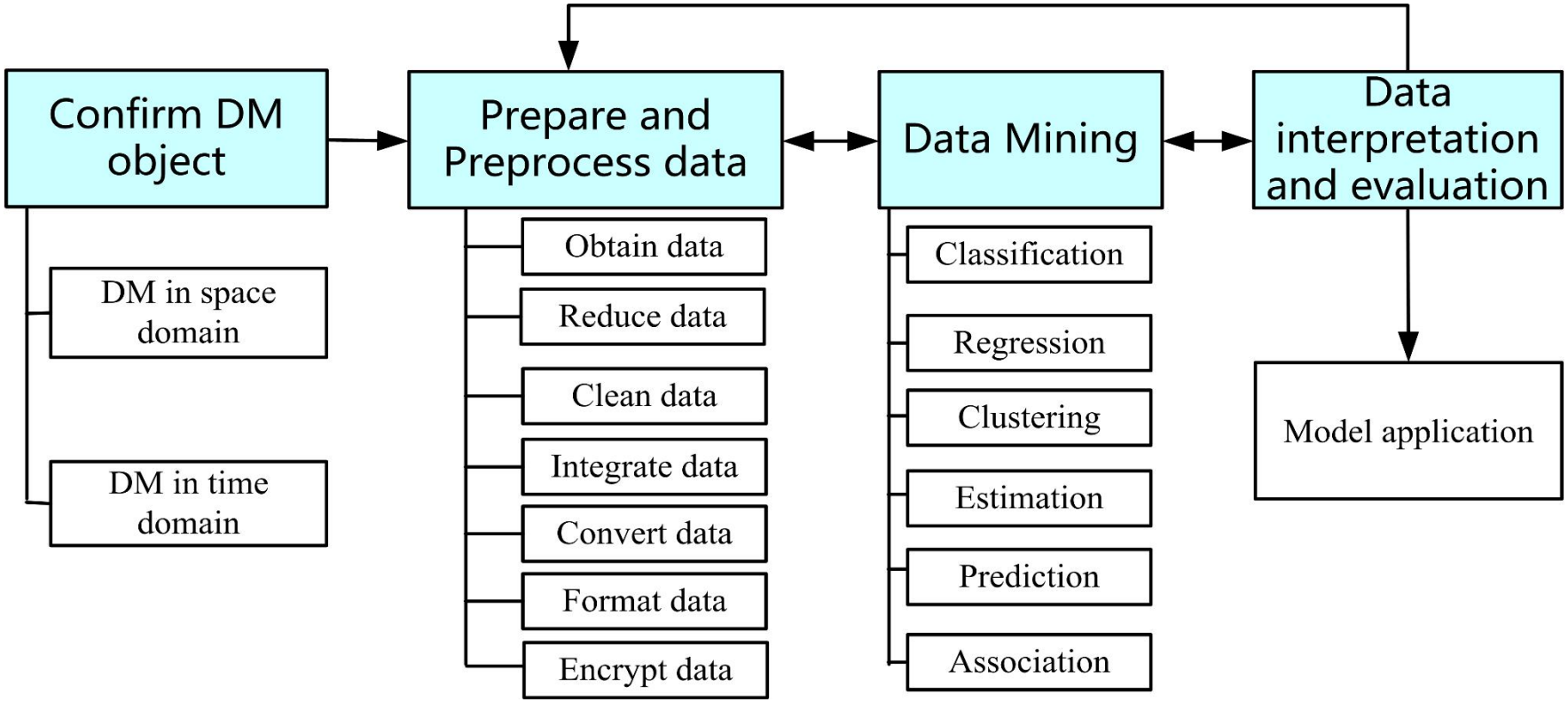




Weka（3.9.1版本）中，共3大类、约130种不同的算法，其中关联算法6种、聚类算法12种、分类算法110种，预处理算法约79种。

PetroChina

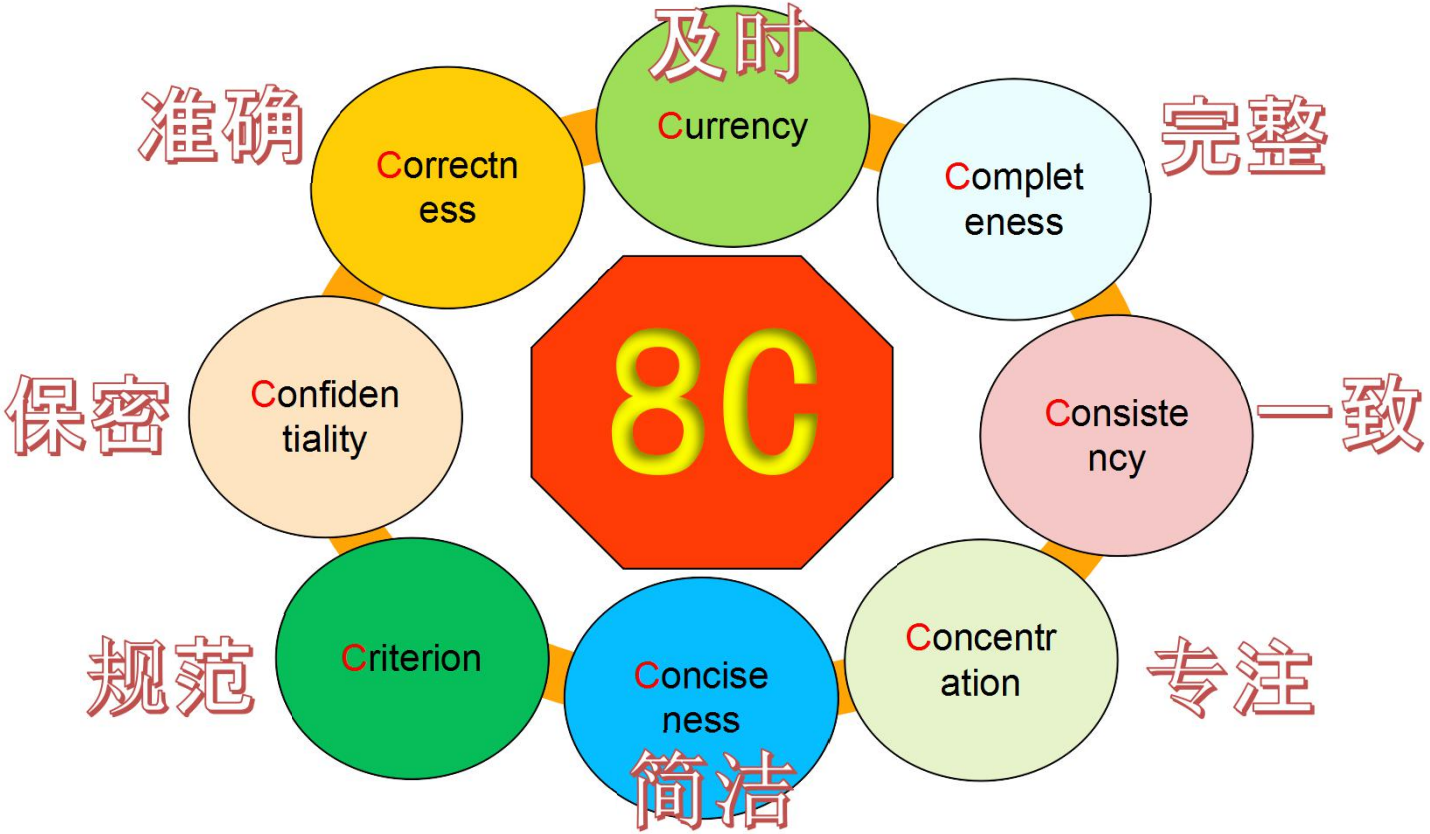
3 DM in petroleum industry



Flow chart of data mining

PetroChina

3 DM in petroleum industry



"8C" standard of data mining (Dawei Li, 2017)

PetroChina

3 DM in petroleum industry

石油工业信息化大数据“8C”标准研究

Research of Big Data "8C" Standard for Petroleum Industry Informatization

■ 中国石油勘探开发研究院[1] 东北石油大学计算机与信息技术学院[2]
李大伟[1] 袁 满[2]

摘 要 在大数据经典“4V”标准基础上，针对石油工业信息化建设周期长、风险高、实施过程复杂等特点，提出大数据的正确性、及时性、完整性、一致性、保密性、规范性、简洁性、专注性的“8C”标准理念，对石油工业信息化的数据资源建设、应用和管理水平的提升具有指导意义。

关键词 石油工业信息化 数据挖掘 “4V”标准 “8C”标准

Abstract: Based on the classic “4V” standard, this paper puts forward the concept of big data's “8C” standard of correctness,currency,completeness,consistency,confidentiality,criterion,conciseness, concentration in view of the long construction cycle, high risk and complex implementation process of petroleum industry informatization, It has guiding significance for the construction, application and promotion of management level for the information resources for petroleum industry.

Keywords: petroleum industry informatization; data mining; “4V” standard; “8C “standard

1 引言

中国石油工业的信息化起步较早，经过30多年的发展，国内几大石油公司的信息化水平已经取得了长足的发展，并一直走在国企的前列，成为推进整个行业发展的巨大动力[1]。石油工业信息化经过分散建设、集中建设和集成应用三大阶段的发展，已经向全面共享阶段推进（中国石油于2016年明确提出了建设“共享中国石油”的奋斗目标）。但是，根据目前的情况，中国石油工业信息化所取得的成就和效果与投入的人力、物力和财力等还存在一定的差距[2]，这既有技术方面的原因，也有管理方面的原因。其中很重要一条，就是在信息化建设过程中，没有对信息化最核心和最基础的内容——数据给以全面正确的理解和管控，因此也就没有得到充分的利用。所以，如何理解石油工业的信息化，特别是如何提升石油工业信息化之数据资源建设、应用和管理水平，具有重要的现实意义。

2 大数据经典的“4V”标准

根据笔者的理解和大数据定义的一般标准，石油工业早已迈入了大数据时代，但有人对此持不同意见，认为石油工业还根本没有所谓的“大数据”。本节以国内最大的石油公司——中国石油（CNPC）为例，用IBM所提出的“大数据”最经典的“4V”标准对此作简要分析和评价。

（1）数量（Volume）：指数据巨大，从TB级别跃升到PB级别。以中国石油为例，经过“十五”以来的建设，已经有大约70个统一的大型信息系统完成了建设并上线运行，仅其中的“勘探与生产技术数据管理系统”（A1系统）就管理着约1 500 TB的数据和约32万口井的结构化数据。而尚未入库、分散在各个单位和个人手中的数据更是难以统计。因此，中国石油可管理的数据满足大数据在数量方面的要求。

（2）多样（Variety）：即数据类型繁多，不仅包括

项目来源：国家“十二五”油气重大科技专项“全球剩余油气资源研究及油气资产快速评价技术”，项目编号：2011ZX050。

严格的安全策略不利于用户使用的方便性，与信息化的开放性、共享性相悖。所以保密和共享是信息化过程中需要平衡好的一对矛盾。如果处理不好两者之间的关系，会严重影响信息化的进程，造成数据所属单位不敢将所拥有的数据存入数据库共享，数据库管理人员也不敢共享入库的数据，最终势必造成库中的数据不全面、不及时等问题。

3.6 规范性（Criterion）

规范性是指信息化过程中所管理和共享使用的数据要遵循相关的规范和标准，这样才能最大化地使数据在不同业务流程中畅通无阻，使信息化的效益实现最大化。

石油工业因其业务流程长、业务环节多、专业应用软件繁杂等原因，造成数据在内容和形式上的不统一。尽管我们已经有各种各样的行业、企业标准，但有些标准本身就存在不标准、可行性差等问题，从而影响了这些标准的推广使用。所以，要保证数据的规范性，一方面要加强相关标准规范的研制、更新和推广[3]，另一方面也要加强相关人员的标准意识，在数据全生命周期中贯彻标准管理和使用。

3.7 简洁性（Conciseness）

简洁性是指信息化过程中，要根据应用的需要，尽量将最需要的数据管好、用好，而不是一味求多、求全。

石油工业由于历史悠久，分布广泛，因此所产生的数据在时间和空间上都有非常广的分布，这其中必然存在数据过时、以后数据不再使用、数据冗余等情况。例如一个勘探区块，早期可能使用了重、磁、电、二维地震等分辨率较低的勘探技术，所得的数据可靠性也比较低。后期随着勘探开发的需要和技术的进步，使用测井、三维地震（甚至四维地震），那么早期产生的这些数据就基本没有使用价值了，就可以暂缓入库甚至废弃不用。对于海外油气勘探开发来说，在评价一个区块的新项目时，可能会花费巨额费用购买国外的资料，但经过评价，该区块根本没有勘探开发前景，对于这些区块的资料，无需再花许多精力和财力去收集、整理和入库。

3.8 专注性（Concentration）

专注性是指信息化过程中，要根据具体的业务和需求，有针对性地专注于收集、管理所需要的数据资源。

总管石油行业有非常长的业务链，但具体到某一个部门的某一个管理人员或研究人员，他所关心和使用的数据类型还是相对比较少和比较集中的。所以在石油工业信息化过程中，在总体规划、统一管理、统一标准等的大前提下，各单位根据自己的具体需求做好做精某一专业领域的数据收集、入库、共享。否则，如果一味求统一，势必造成建设周期过长、不同单位的具体需求无法满足，影响信息化进程。但各单位要遵循统一的相关标准和规范，这样才不至于造成各个系统之间数据资源无法共享。

测井是油气勘探开发中非常重要而且专业性很强的一大类数据，可以通过专门的测井数据库建设以更好地支持相关的应用[4]。如中国石油海外测井技术支持中心所建立的海外测井库，专注于收集、整理、管理和共享使用中国石油海外业务有关的一万多口井的测井、录井、试油等与井相关的数据和文档，并基于该数据库和中国石油统一的数据模型，研制了井筒数据采集、处理、解释、评价一体化系统，同时该系统正在规划与中国石油相关系统（A1系统、海外勘探开发信息管理系统等）的互联互通，从而可以实现资料共享，避免重复建设。

4 结语

石油工业作为国民经济支柱产业之一，早已进入大数据时代。在数据资源建设时，不仅应当严格遵循相关的标准和规范，还应当从本文提出的“8C”标准等方面加强管理，为石油工业的信息化打下牢固的数据基础。

要达到“8C”标准，需要在石油工业某一业务领域的数据全流程和生命周期中，严格按照相关标准、规范，做好各方面的数据处理[5]，必要时还需要对某些信息系统的底层、模型和架（下转第61页）

PetroChina

3 DM in petroleum industry

Common DM algorithms

Regression:

- Multiple regression analysis (MRA)
- Error back-propagation neural network (BPNN)
- Regression of support vector machine (*R*-SVM)

Classification:

- Classification of support vector machine (*C*-SVM)
- Naïve Bayesian (NBAY)
- Bayesian discrimination (BAYD)

PetroChina

OUTLINE

1. Background
2. Big data in petroleum industry
3. Data mining in petroleum industry
4. Case study
5. Conclusions

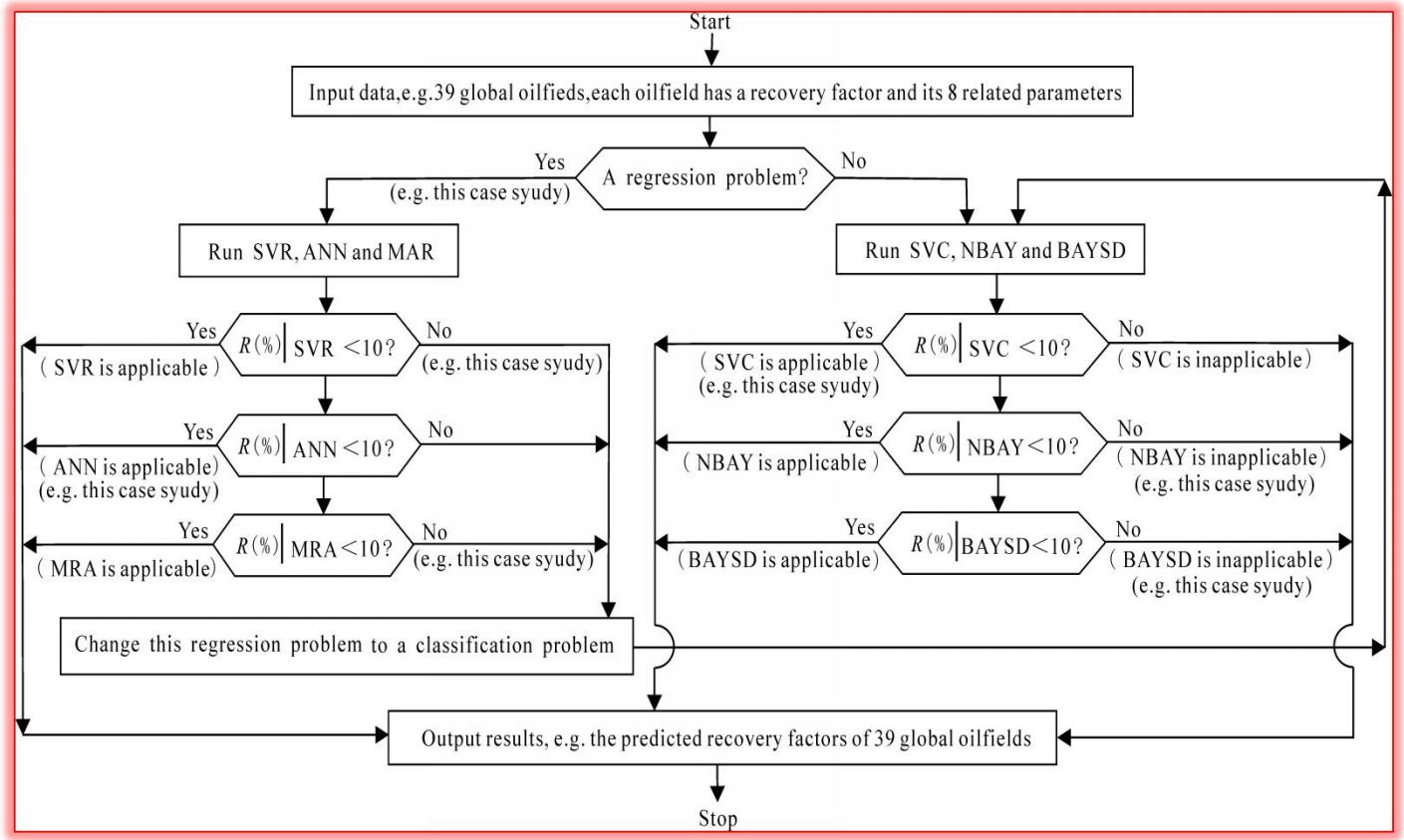
PetroChina

4 Case study

Based on a commercial global oil and gas field database bought from C&C (an American company), we analyzed the key affecting attributes for recovery factor and the applicability of common data mining algorithms.

Database name	Oil and gas field data	Static reservoir data	Dynamic reservoir data
Field numbers	55	377	423
Record numbers	1079	1446	699

4 Case study



Flow chart of data mining of case study

PetroChina

4 Case study

Major attributes of oilfield database from C&C

Oilfield code	Total production years	Current production stage code	Current production well numbers	Original in-place oil equivalent (MMBOE)	EUR oil equivalent (MMBOE)	Oil recovery factor (%)
Field 1	63	2	465	3200	311.7	12.03
Field 2	21	5	82	146	24	39.04
Field 3	46	3	193	1850	109	10
Field 4	58	5	63	405	126.8	32.59
Field 5	63	4	570	36840	1452.4	6.7
……	……	……	……	……	……	……
Field 1076	9	2	28	75	6353	43.6
Field 1077	73	2	6150	2984	37452	33.91
Field 1078	31	3	145	66.7	1960	42
Field 1079	25	1	643	94.1	11308	14.77

Current production stage code: 1--primary rejuvenating, 2--secondary peak or plateau, 3--secondary decline, 4-- secondary mature, 5-- secondary rejuvenating, 6--tertiary peak or plateau

PetroChina

4 Case study

Input data for recovery factor in 39 selected oilfields

Sample type	Sample No.	Field No.	8 parameters related to y								y*	
			x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	RF[b]	RFC[c]
Learning samples	1	F001	63	1	465	792	3200	385	311.7	16555	12.03	5
	2	F002	21	3	82	148	146	57	24	3884	10	5
	3	F003	46	5	193	308	1850	185	109	7646	32.59	3
	4	F004	58	3	63	288	405	132	126.8	2535	6.7	5
	5	F005	63	2	570	800	36840	2470	1452.4	103000	58.97	1
	6	F006	37	3	47	96	3900	2300	1830	63500	46	2
	7	F007	34	2	201	730	25000	11500	4394	447052	32.8	3
	8	F008	53	4	7	186	25	8.2	8.1	19	15.08	5
	9	F009	21	2	122	178	325	49	25	6353	6.19	5
	10	F010	59	1	2002	3632	13436	982	267.3	71470	37.18	3
	11	F011	44	3	153	430	347	129	122	4196	42.74	2
	12	F012	61	3	185	632	15667	7355	3650	335000	33.51	3
	13	F013	41	3	93	326	6900	3580	3697	211868	30.62	3
	14	F014	37	1	3200	4600	967	324	270	20580	36.26	3
	15	F015	54	1	145	350	60520	22530	7222	500000	24.39	4
	16	F016	37	4	179	437	353	128	116.6	1910	18.13	5
	17	F017	50	2	782	1218	41000	10000	7083	350000	20	5
	35	F035	74	3	2221	6000	2800	1481	1340	26676	32.14	3
Prediction samples	36	F036	20	1	23	31	549	55	31	5900	(39.04)	(3)
	37	F037	74	1	2613	4121	28437	7565	5055	246441	(51.88)	(1)
	38	F038	44	4	33	73	140	45	41.8	198	(26.6)	(4)
	39	F039	25	1	327	643	637.3	94.1	59.8	11308	(14.77)	(5)

x_1 = total production years, x_2 = current production stage code, x_3 = current producing well numbers, x_4 = total well numbers, x_5 = original in-place oil equivalent (MMBOE), x_6 = EUR (Estimated Ultimate Recovery) oil equivalent (MMBOE), x_7 = production of cumulative oil equivalent (MMBOE), x_8 = production rate of current oil equivalent (BOEPD)], y^* = RF or y^* = RFC

PetroChina

4 Case study

Classification of recovery factor

Recovery Factor	RF (recovery factor) (%)	RFC (recovery factor classification) (integer)
Very high recovery factor	>50	1
High recovery factor	40< RF≤50	2
Intermediate recovery factor	30< RF≤40	3
Low recovery factor	20< RF≤30	4
Very low recovery factor	≤20	5

PetroChina

4 Case study

Prediction results of recovery factor regression in 39 global oilfields

Sample type	Sample No.	Recovery factor						
		y*	R-SVR		ANN		MRA	
			y	R(%)	y	R(%)	y	R(%)
Learning samples	1	12.03	30.7558	155.659	11.6	3.88	18.7	55.1
	2	10	33.2312	232.312	9.35	6.49	35.9	259
	3	32.59	32.5433	0.143	32.1	1.67	28.9	11.4
	4	6.7	31.0992	364.167	6.19	7.61	-.3.6	154
	5	58.97	32.4701	44.938	58.1	1.56	33.2	43.7
	6	46	31.862	30.735	47.2	2.52	42.5	7.62
	7	32.8	32.3528	1.363	32.9	0.169	30.1	8.21
	8	15.08	30.8125	104.327	15.2	0.629	22.6	49.7
	9	6.19	30.8843	398.939	6.19	0.0000308	16.6	168
	10	37.18	32.4168	12.811	36.4	2.05	29.3	21.3
	11	42.74	33.1068	22.539	44.4	3.87	44.2	3.34
	12	33.51	30.9235	7.719	35.0	4.53	29.7	11.4
	13	30.62	32.1243	4.913	29.5	3.60	31.9	4.23
	14	36.26	32.2356	11.099	35.6	1.88	30.7	15.4
	15	24.39	31.7768	30.286	22.4	8.10	21.8	10.6
	16	18.13	30.6217	68.901	19.5	7.55	20.6	13.7
	17	20	30.3865	51.933	16.3	18.7	18.8	6.24
	18	34.89	33.8271	3.046	32.7	6.23	40.0	14.8
	19	65.85	34.7358	47.25	67.6	2.59	69.1	4.99
	20	25.64	31.1359	21.435	26.2	1.98	20.5	19.9
	21	15.6	30.3802	94.745	17.6	12.8	18.8	20.2
	22	14.9	30.5754	105.204	15.7	5.62	19.0	27.4
	23	32.68	32.3982	0.862	32.1	1.88	29.9	8.39
	24	79.05	34.8936	55.859	79.1	0.000029	77.7	1.68
	25	35.36	32.3736	8.446	33.1	6.30	30.3	14.4
	26	34.37	32.5481	5.301	34.3	0.149	30.3	11.8
	27	63.5	33.5591	47.151	63.0	0.871	44.4	30.1
	28	11.12	30.8219	177.175	12.1	8.89	19.9	78.7
	29	13.91	33.568	141.323	12.7	8.77	40.0	188
	30	33.33	33.23	0.3	36.0	8.04	41.7	25.1
	31	37.81	32.4452	14.189	37.1	1.93	37.8	0.0841
	32	32.09	32.19	0.312	31.7	1.32	30.0	6.63
	33	52.89	33.1232	37.373	52.9	0.0276	38.6	27.1
	34	10.02	30.396	203.353	12.0	20.1	18.5	84.9
	35	32.14	32.2773	0.427	31.0	3.49	30.1	6.26
Prediction samples	36	39.04	32.2065	17.504	24.8	36.5	28.9	25.9
	37	51.88	32.7885	36.799	44.3	14.7	39.3	24.2
	38	26.6	31.6331	18.921	26.5	0.276	27.0	1.49
	39	14.77	30.447	106.141	12.9	12.6	19.8	3.3.8

4 Case study

PetroChina

Prediction results of recovery factor classification in 39 global oilfields

Sample type	Sample No.	RFC						
		y^*	C-SVC		NBAY		BAYSD	
			y	R(%)	y	R(%)	y	R(%)
Learning samples	1	5	5	0	5	0	5	0
	2	5	5	0	5	0	3	40
	3	3	3	0	5	66.7	5	66.7
	4	5	5	0	4	20	5	0
	5	1	1	0	1	0	3	200
	6	2	2	0	2	0	2	0
	7	3	3	0	3	0	3	0
	8	5	5	0	5	0	5	0
	9	5	5	0	5	0	5	0
	10	3	3	0	5	66.7	5	66.7
	11	2	2	0	2	0	2	0
	12	3	3	0	1	66.7	3	0
	13	3	3	0	3	0	3	0
	14	3	3	0	3	0	3	0
	15	4	4	0	4	0	4	0
	16	5	5	0	5	0	5	0
	17	5	5	0	5	0	5	0
	18	3	3	0	5	66.7	3	0
	19	1	1	0	1	0	1	0
	20	4	4	0	5	25	4	0
	21	5	5	0	5	0	5	0
	22	5	5	0	5	0	5	0
	23	3	3	0	5	66.7	5	66.7
	24	1	1	0	1	0	1	0
	25	3	3	0	3	0	3	0
	26	3	3	0	3	0	3	0
	27	1	1	0	5	400	3	200
	28	5	5	0	5	0	5	0
	29	5	5	0	5	0	3	40
	30	3	3	0	3	0	3	0
	31	3	3	0	3	0	3	0
	32	3	3	0	3	0	3	0
	33	1	1	0	1	0	3	200
	34	5	5	0	5	0	5	0
	35	3	3	0	3	0	3	0
Prediction samples	36	3	3	0	5	66.7	5	66.7
	37	1	1	0	3	200	5	400
	38	4	4	0	3	25	4	0
	39	5	5	0	5	0	5	0

4 Case study

PetroChina

Summary of the case study with 39 global oilfields

Problem type	Algorithm	Mean absolute relative residual error	Time consuming on PC	Conclusion
		R(%) (平均相对剩余误差的绝对值)		
Regression	SVR	68.9	3 s	Inapplicable
	ANN	5.89	30 s	Applicable
	MRA	38.4	<1 s	Inapplicable
Classification	SVC	0	5 s	Applicable
	NBAY	24.7	<1 s	Inapplicable
	BAYSD	34.5	1 s	Inapplicable

1. This program ran on a PC with configuration: HP Z230, Windows 7 (64 bit), Intel E3-1231, 3.40 GHz, 16GB RAM
2. If R > 10%, it is inapplicable.

PetroChina

OUTLINE

1. Background
2. Big data in petroleum industry
3. Data mining in petroleum industry
4. Case study
5. **Conclusions**

PetroChina

Conclusions

- Petroleum industry has entered “Big Data” era;
- Data mining (DM) is one of the good solutions to fully utilize the value of the IT systems and data assets;
- For the case study, the best algorithm is ANN for recovery factor regression, while the optimum algorithm is SVC for recovery factor classification;
- The best algorithms are different for various data sets.

Thank You !

Questions?

Email: leedw@petrochina.com.cn
Wechat: 13671142720