USING LISP-MINER IN THE COMMERCIAL SPHERE

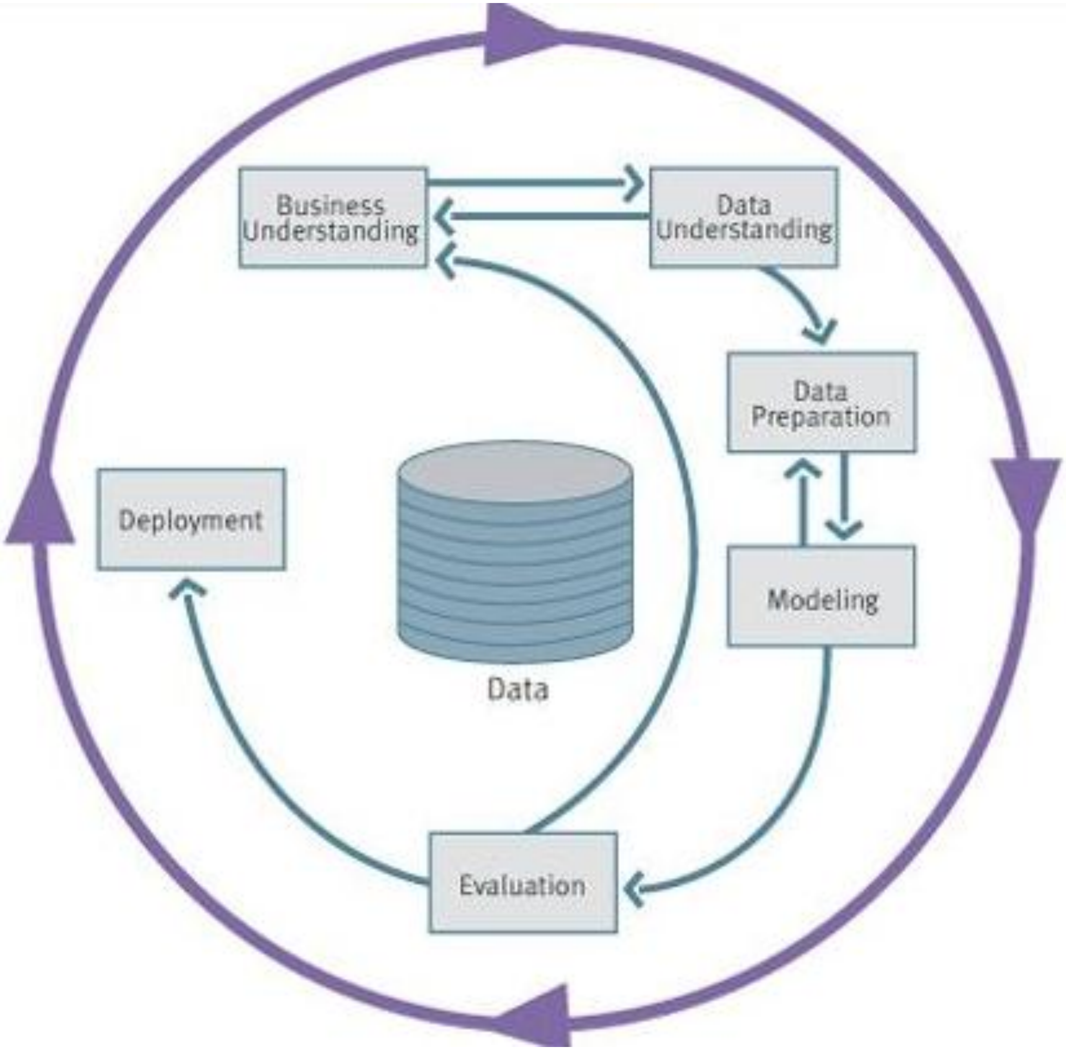
Viktor Nekvapil

Contents

- Data mining in CRM: the case of a major logistic company
- Cooperation with the market research company

DATA MINING IN CRM: THE CASE OF A MAJOR LOGISTIC COMPANY

Project overview



Iteration 1 (IT1)

- □ Business understanding
- □ Data understanding
- □ Data preparation
- □ Modelling
- □ Evaluation

Iteration 2 (IT2)

- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

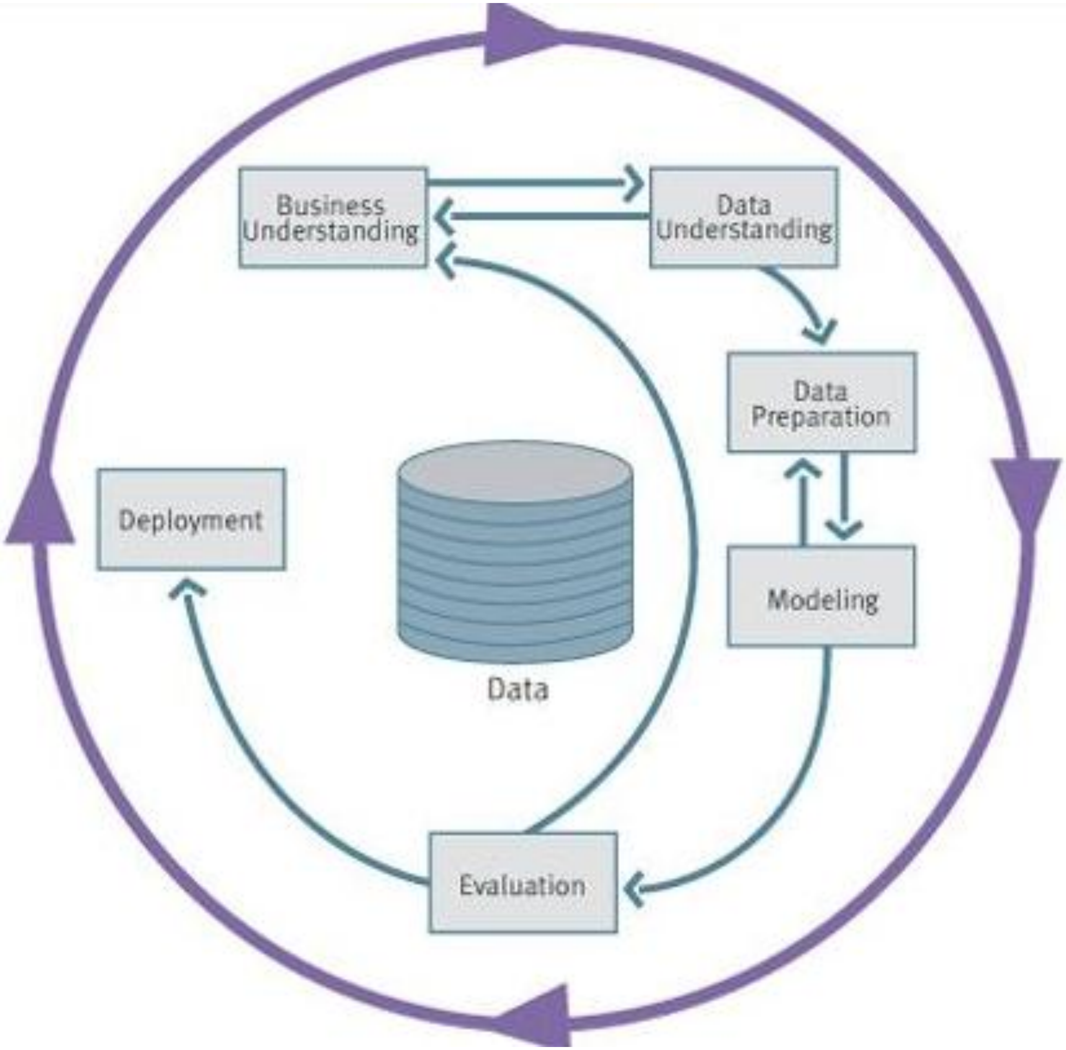
Goals of the project

- Analyse the given data using the LISp-Miner system in compliance with the aims of the case study
- Propose directions of the use of the LISp-Miner system when solving a similar data mining task
- Propose a simple and understandable way to present results of the LISp-Miner system

Getting in touch with the company

- Contact arranged by the Opti Solutions
- Four visits in the company
  1. Initial meeting – domain knowledge
  2. Processes
  3. Data was obtained
  4. Meeting after first analysis
- Email and phone communication

Project overview



Iteration 1 (IT1)

- □ **Business understanding**
- □ Data understanding
- □ Data preparation
- □ Modelling
- □ Evaluation

Iteration 2 (IT2)

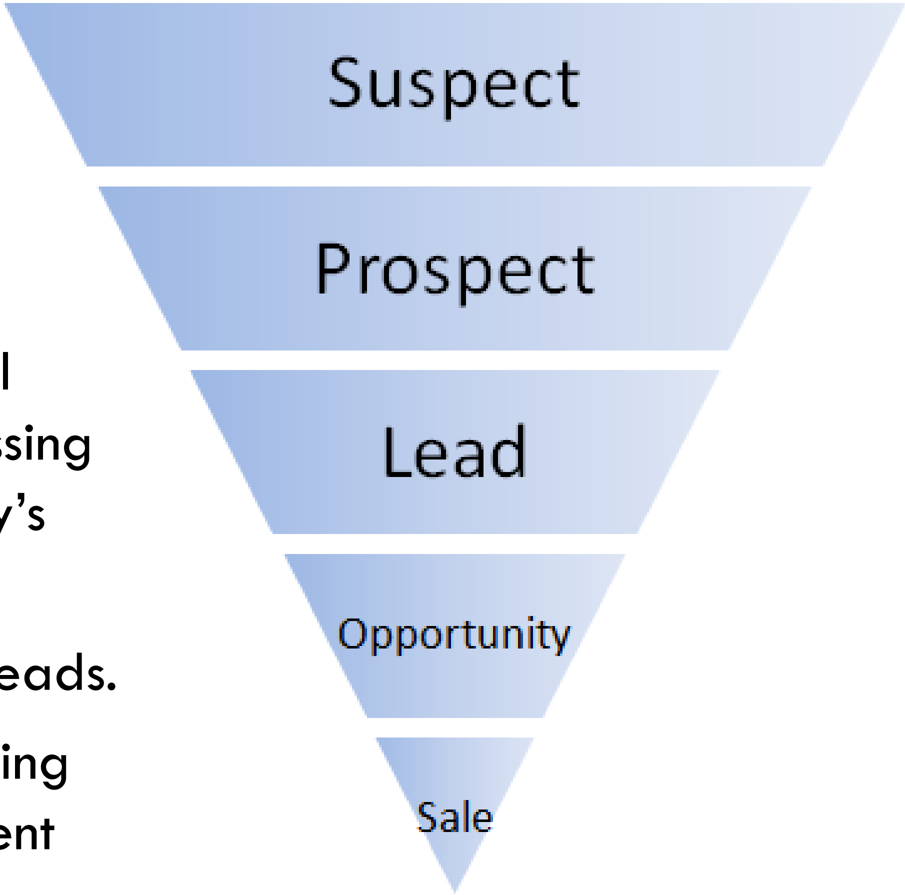
- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

IT1 – BUSINESS UNDERSTANDING

- CRM – Lead management

previously unknown domain

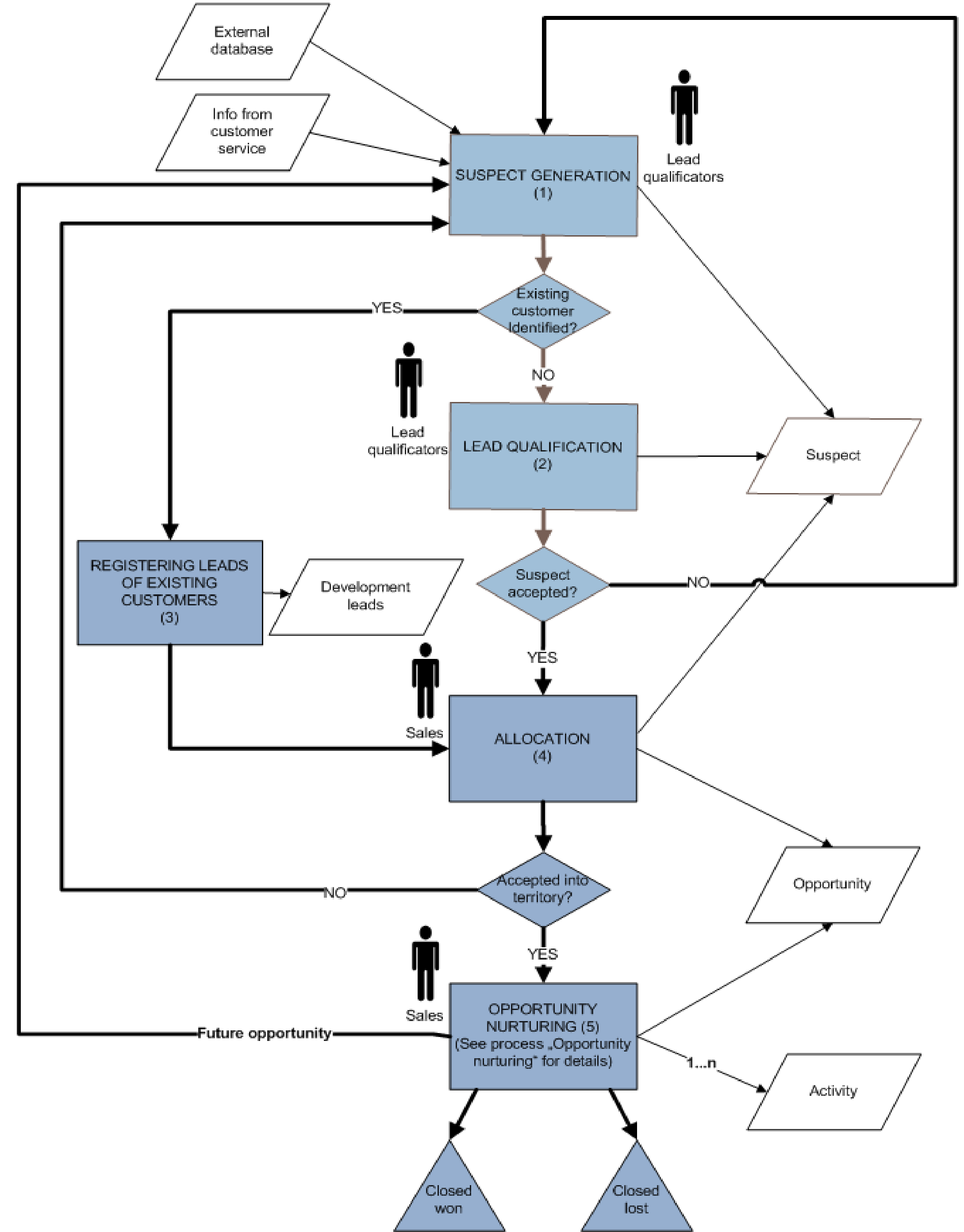
- □ Suspect – organisation that is believed to fit to the company's customer profile
- □ Prospect – indication of potential opportunity; organisation expressing some level of interest in company's product.
- □ Lead – qualified prospects are leads.
- □ Opportunity – qualified lead being processed by the sales department



Processes

- No documentation available
- All information had to be obtained from domain experts
- Took lots of time and effort

Process: Lead management



Important indicators

- Important from the business point of view
- Experts make decisions according to them
- Assumption: important also in the analysis
- Examples
  - Committed revenue
  - Potential revenue
  - Closing ratio =

$$\frac{\text{closed lost opp} + \text{closed won opp}}{\text{closed lost opp} + \text{closed won opp} + \text{future opp} + \text{open opp}}$$

Business and DM objectives

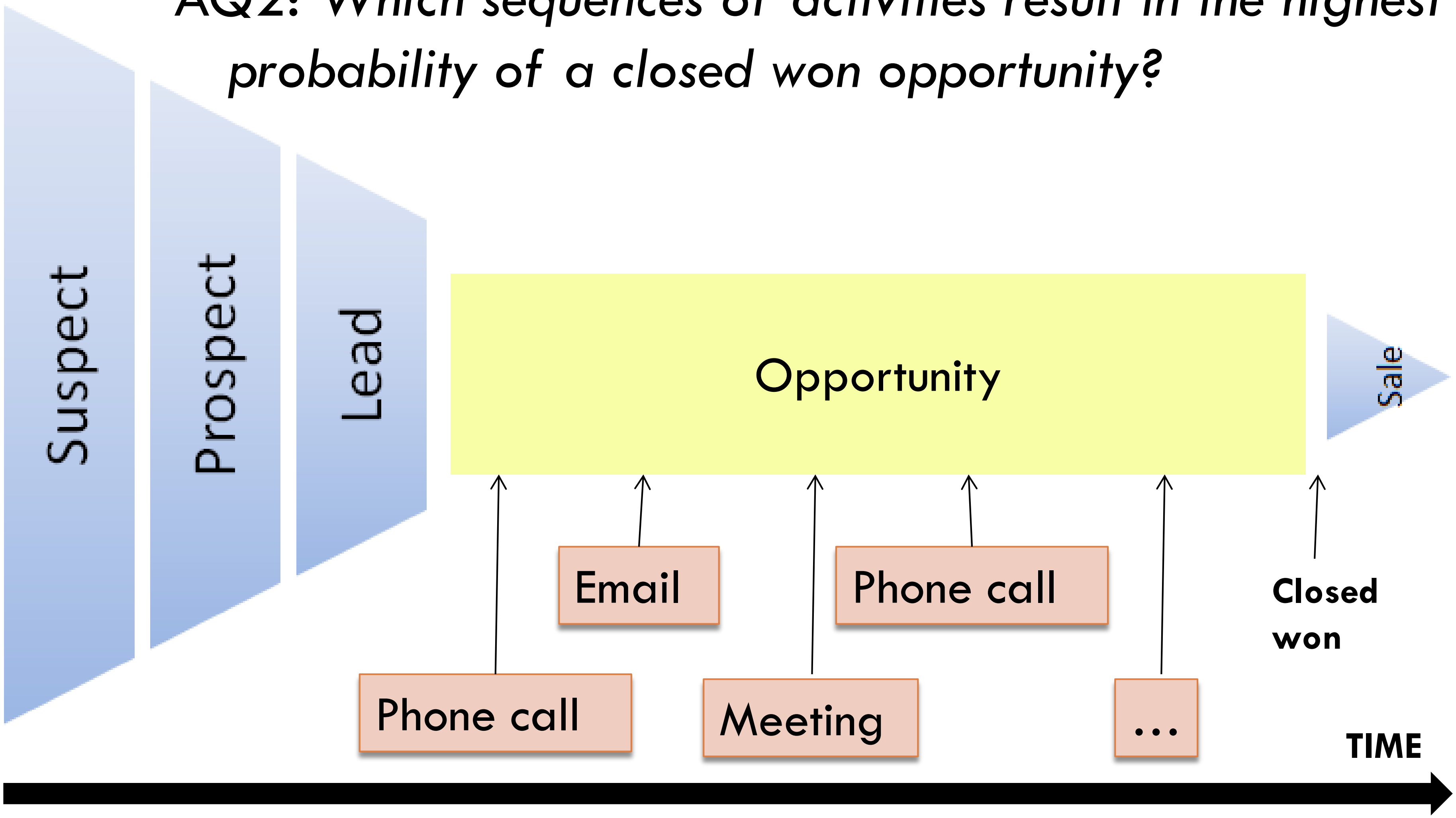
Business – change of internal processes of the company (increase the number of closed won opportunities)

DM – 2 analytical questions

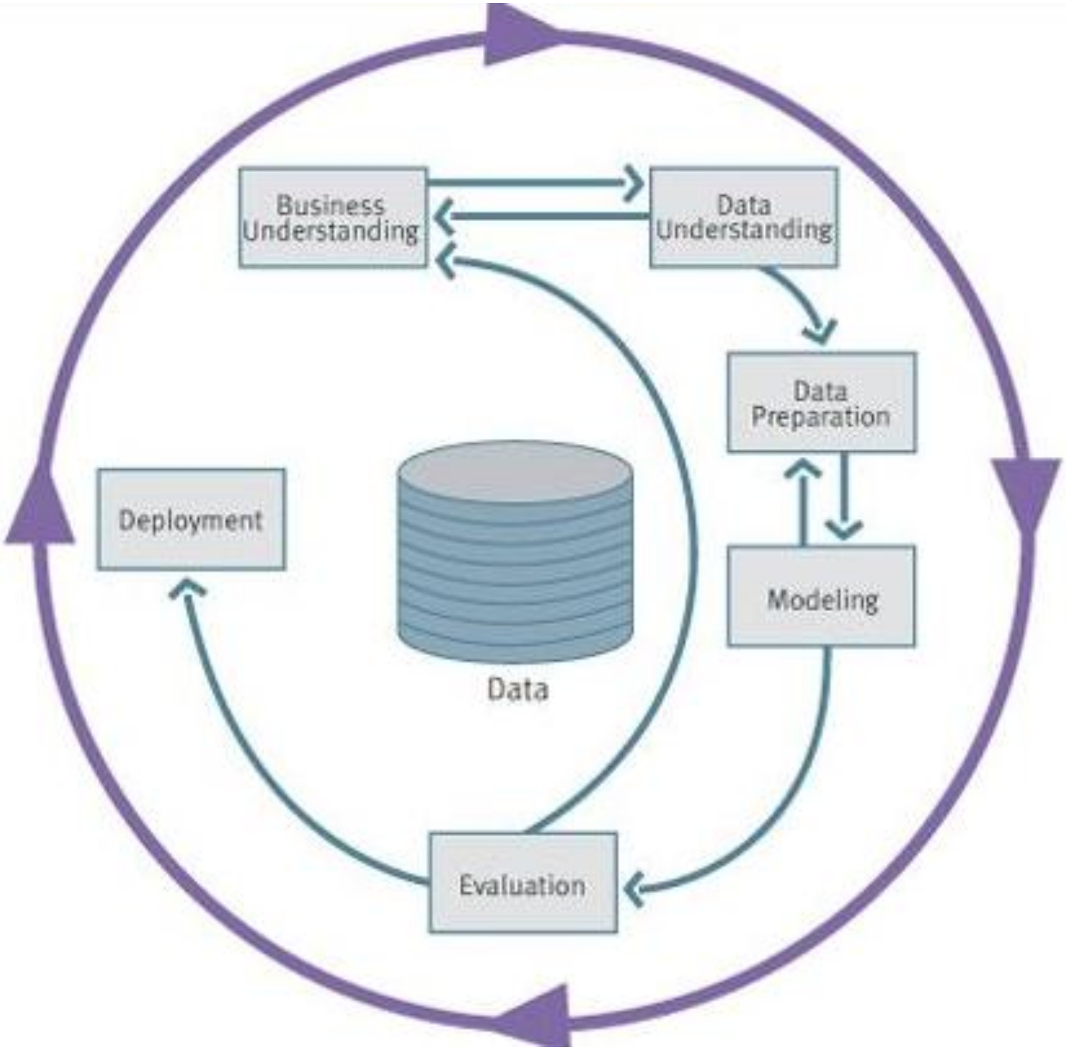
- □ Which combinations of salesman and lead source have the highest revenue / closing ratio / share of closed won opportunities?
- □ Which sequences of activities result in the highest probability of a closed won opportunity?

Analytical question 2

AQ2: Which sequences of activities result in the highest probability of a closed won opportunity?



Project overview



Iteration 1 (IT1)

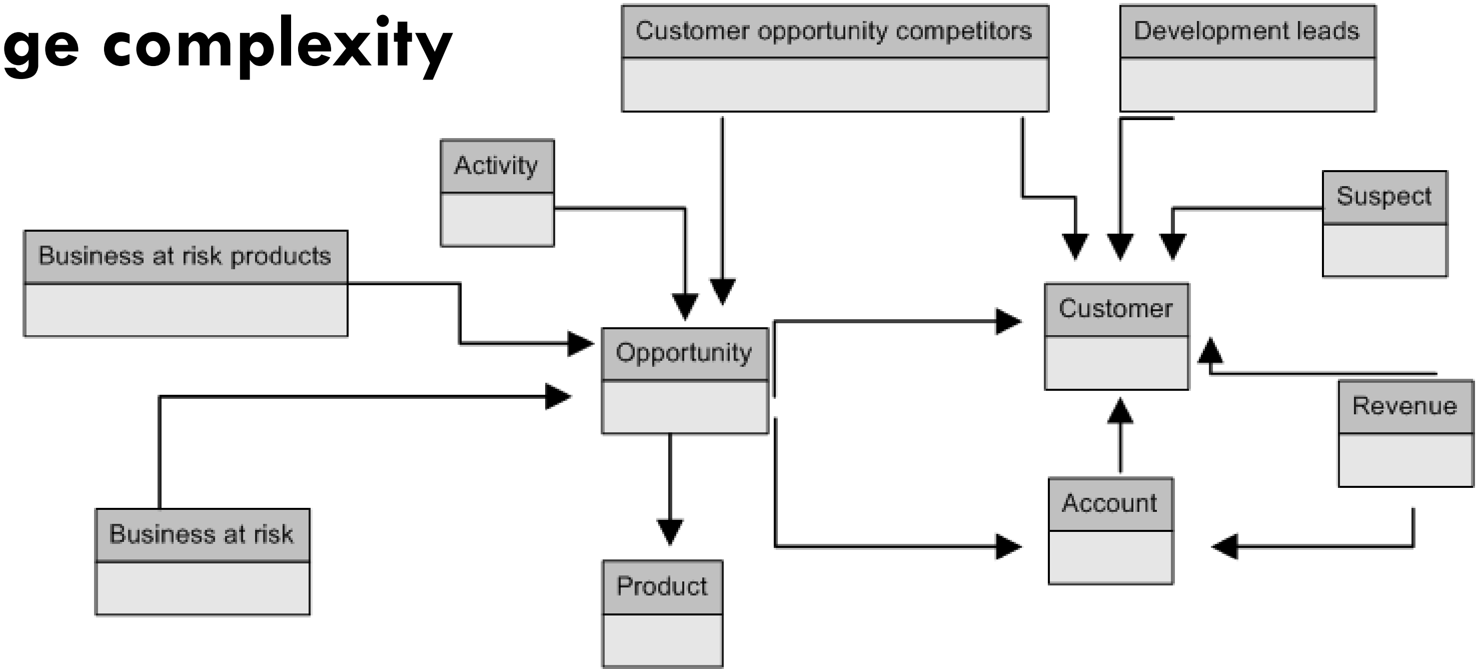
- □ Business understanding
- □ **Data understanding**
- □ Data preparation
- □ Modelling
- □ Evaluation

Iteration 2 (IT2)

- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

IT1 – DATA UNDERSTANDING

- □ 22 tables („extracts“) available
- □ each containing on average about 20 columns (fields)
- □ no description of the meaning of the columns
- □ -> **huge complexity**



IT1 – Data understanding

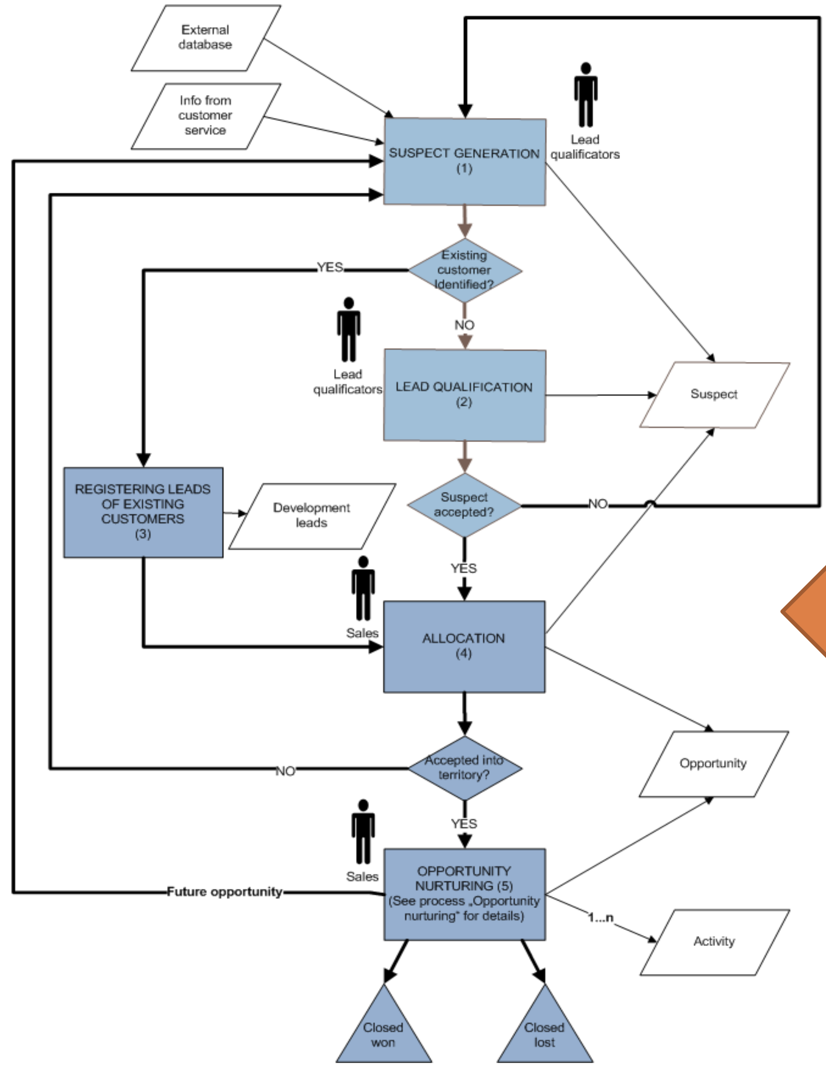
- Opportunity and Activity extract identified as promising for answering both analytical questions
- Only Opportunity extract available at the moment
- => only first analytical question is solved in Iteration 1

Example – Extract „Opportunity“

#	Field Name	Data Type	Length	LOV	Comments
1	ACTUAL CLOSE DATE	Date	7		Date format will be DD-MM-YYYY
2	[TOTAL] COMMITTED REVENUE	Number	22		
3	CUSTOMER NAME	Char	100		
4	PIPELINE 2 ENTERED	Date	7		Date format will be DD-MM-YYYY
5	PIPELINE 3 ENTERED	Date	7		Date format will be DD-MM-YYYY
6	PIPELINE 4 ENTERED	Date	7		Date format will be DD-MM-YYYY
7	PIPELINE 5 ENTERED	Date	7		Date format will be DD-MM-YYYY
8	PIPELINE 6 ENTERED	Date	7		Date format will be DD-MM-YYYY
9	PIPELINE 7 ENTERED	Date	7		Date format will be DD-MM-YYYY
10	PIPELINE 8 ENTERED	Date	7		Date format will be DD-MM-YYYY
11	PIPELINE 9 ENTERED	Date	7		Date format will be DD-MM-YYYY
12	EXPECTED CLOSE DATE	Date	7		Date format will be DD-MM-YYYY
13	GSFA CUSTOMER ID	Char	50		
14	LEAD ORIGINATOR	Char	15		
15	LEAD SOURCE	Char	100		
16	OPPORTUNITY CREATED BY LOGIN ID	Char	15		
17	OPPORTUNITY CREATED DATE	Date	7		Date format will be DD-MM-YYYY
18	OPPORTUNITY ID	Char	15		
19	OPPORTUNITY NAME	Char	100		
20	OPPORTUNITY STAGE	Char	30	Yes	
21	OPPORTUNITY TYPE	Char	30	Yes	
22	PIPELINE LAST UPDATE BY	Char	15		
23	PIPELINE LAST UPDATED DATE	Date	7		Date format will be DD-MM-YYYY
24	[TOTAL] POTENTIAL REVENUE	Number	22		
25	PREVIOUS PIPELINE STAGE	Char	50		
26	REASON FOR LEAD	Char	30	Yes	
27	REASON [WON]/LOST	Char	30	Yes	
28	SALES TEAM	Char	50		
29	SALES TERRITORY CODE	Char	50		
30	PRIMARY [MAIN] COMPETITOR	Char	75		
31	OPPORTUNITY LAST UPDATED BY	Char	50		
32	GSFA ACCOUNT ID	Char	50		
33	ACCOUNT NUMBER	Char	15		
34	SOURCE TYPE (lead)	Char	30	Yes	

Mapping data on the processes

Process: Lead management



#	Field Name	Data Type	Length	LOV	Comments
1	ACTUAL CLOSE DATE	Date	7		Date format will be DD-MM-YYYY
2	[TOTAL] COMMITTED REVENUE	Number	22		
3	CUSTOMER NAME	Char	100		
4	PIPELINE 2 ENTERED	Date	7		Date format will be DD-MM-YYYY
5	PIPELINE 3 ENTERED	Date	7		Date format will be DD-MM-YYYY
6	PIPELINE 4 ENTERED	Date	7		Date format will be DD-MM-YYYY
7	PIPELINE 5 ENTERED	Date	7		Date format will be DD-MM-YYYY
8	PIPELINE 6 ENTERED	Date	7		Date format will be DD-MM-YYYY
9	PIPELINE 7 ENTERED	Date	7		Date format will be DD-MM-YYYY
10	PIPELINE 8 ENTERED	Date	7		Date format will be DD-MM-YYYY
11	PIPELINE 9 ENTERED	Date	7		Date format will be DD-MM-YYYY
12	EXPECTED CLOSE DATE	Date	7		Date format will be DD-MM-YYYY
13	GSFA CUSTOMER ID	Char	50		
14	LEAD ORIGINATOR	Char	15		
15	LEAD SOURCE	Char	100		
16	OPPORTUNITY CREATED BY LOGIN ID	Char	15		
17	OPPORTUNITY CREATED DATE	Date	7		Date format will be DD-MM-YYYY
18	OPPORTUNITY ID	Char	15		
19	OPPORTUNITY NAME	Char	100		
20	OPPORTUNITY STAGE	Char	30	Yes	
21	OPPORTUNITY TYPE	Char	30	Yes	
22	PIPELINE LAST UPDATE BY	Char	15		
23	PIPELINE LAST UPDATED DATE	Date	7		Date format will be DD-MM-YYYY
24	[TOTAL] POTENTIAL REVENUE	Number	22		

Mapping data on the processes

#	Field name	Stage / state
1	ACTUAL CLOSE DATE	Closed won / closed lost / future opportunity
2	[TOTAL] COMMITTED REVENUE	ALLOCATION (4)
3	CUSTOMER NAME	ALLOCATION (4)
4	PIPELINE 2 ENTERED	ESTABLISHING FIRST CONTACT (5. 1) / established: yes
5	PIPELINE 3 ENTERED	PRICE OFFER (5. 2) / accepted: yes
6	PIPELINE 4 ENTERED	SHIPMENT AGREEMENT (5. 3) / agreed: yes
7	PIPELINE 5 ENTERED	IMPLEMENTATION (5. 4) / implemented: yes
8	PIPELINE 6 ENTERED	Irrelevant – not in the process schema
9	PIPELINE 7 ENTERED	Irrelevant – not in the process schema
10	PIPELINE 8 ENTERED	OPPORTUNITY NURTURING (5) / closed lost
11	PIPELINE 9 ENTERED	OPPORTUNITY NURTURING (5) / future opportunity
12	EXPECTED CLOSE DATE	ALLOCATION (4), further
13	GSFA CUSTOMER ID	ALLOCATION (4)
14	LEAD ORIGINATOR	ALLOCATION (4)
15	LEAD SOURCE	ALLOCATION (4)
16	OPPORTUNITY CREATED BY LOGIN ID	ALLOCATION (4)
17	OPPORTUNITY CREATED DATE	ALLOCATION (4)
18	OPPORTUNITY ID	ALLOCATION (4)
19	OPPORTUNITY NAME	ALLOCATION (4)
20	OPPORTUNITY STAGE	ALLOCATION (4), OPPORTUNITY NURTURING (5)

Data description – Opportunity extr.

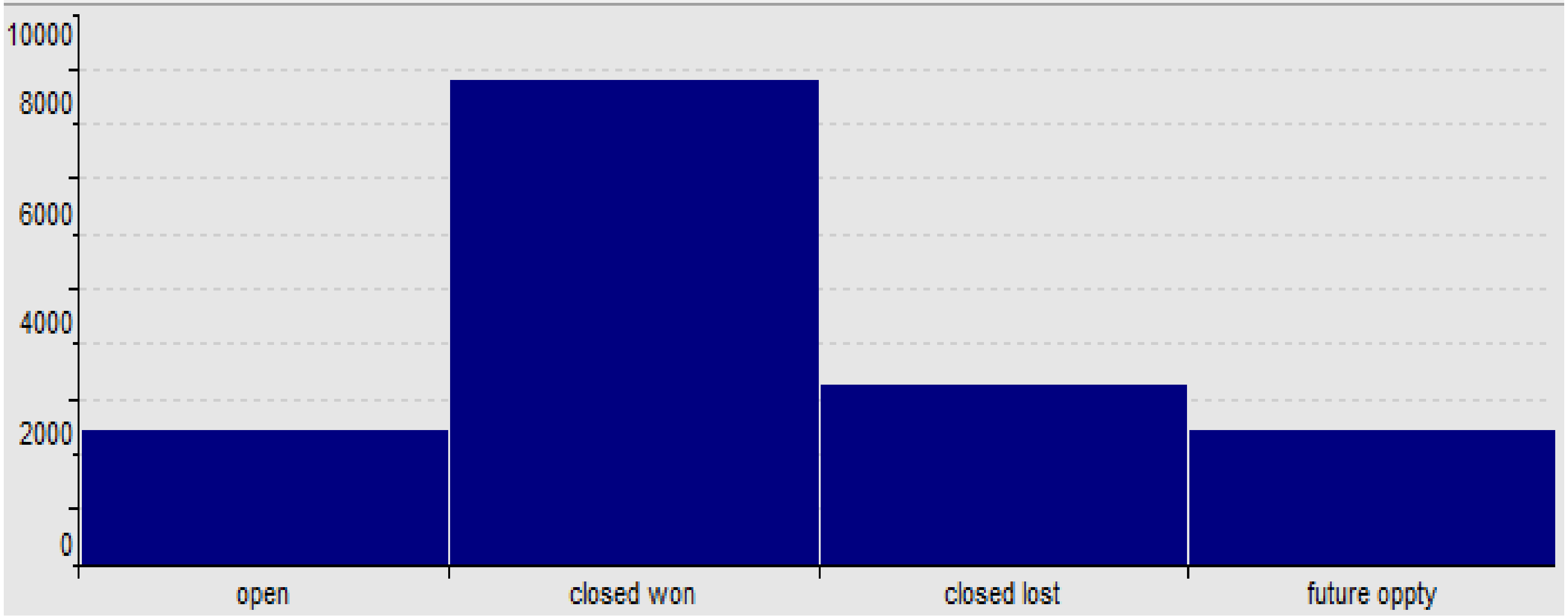
#	Column	data type	range of values	No. of missing values	% of missing values	meaning of missing value	Remarks / meaning of the column
1	Actual Close Date	date	2005 – 2012	8009	47.16%	not closed yet	
2	Committed Revenue	int	0 – 12 714 000	0	0.00%	-	
3	Pipeline 2 Entered	date	2005 – 2012	990	5.83%	opp was/is not in the stage	first contact established
4	Pipeline 3 Entered	date	2005 – 2012	2650	15.61%	opp was/is not in the stage	price offer
5	Pipeline 4 Entered	date	2005 – 2012	6221	36.64%	opp was/is not in the stage	shipment agreement
6	Pipeline 5 Entered	date	2005 – 2012	6462	38.05%	opp was/is not in the stage	implemented
7	Pipeline 6 Entered	date	2005 – 2012	7131	41.99%	opp was/is not in the stage	first consignment
8	Pipeline 7 Entered	date	2005 – 2012	12965	76.35%	opp was/is not in the stage	shipped to profile
9	Pipeline 8 Entered	date	2005 – 2012	13460	79.27%	opp was/is not in the stage	unable to gain
10	Pipeline 9 Entered	date	2005 – 2012	12157	71.59%	opp was/is not in the stage	future opportunity
11	Expected Close Date	date	2004 – 2013, 2015	15	0.09%	value not known / omitted	
12	GSFA Cust ID	char	10221 distinct values	0	0.00%	-	ID of a customer
13	Lead Originator	char	200 distinct values	13320	78.44%	value not known / omitted	
14	Lead Source	char	24 distinct values	114	0.67%	value not known / omitted	
15	Oppty Created Date	date	2004 – 2012	0	0.00%	-	
16	Oppty ID	char	16121 distinct values	0	0.00%	-	
17	Oppty Stage	char	11 distinct values	0	0.00%	-	*
18	Oppty Type	char	6 distinct values	0	0.00%	-	
19	Pipeline Last Upd Date	date	2007 – 2012	29	0.17%	not updated yet	
20	Potential Revenue	int	0 – 59 332 000	68	0.40%	value not known / omitted	

Data selection – Opportunity extr.

#	column	data type	range of values	No. of missing values	% of missing values	meaning of missing value
2	Committed Revenue	int	0 – 12 714 000	0	0.00%	-
14	Lead Source	char	24 distinct values	114	0.67%	value not known / omitted
17	Oppty Stage	char	11 distinct values	0	0.00%	-
18	Oppty Type	char	6 distinct values	0	0.00%	-
20	Potential Revenue	int	0 – 59 332 000	68	0.40%	value not known / omitted
21	Prev Pipeline Stage	char	12 distinct values	646	3.80%	newly created opportunity
24	Territory	char	54 distinct values	0	0.00%	-
26	Lead Source Type	char	9 distinct values	114	0.67%	value not known / omitted
28	New	char	2 distinct values	0	0.00%	-
29	Nr Of Shpts	int	0 – 1 800 000	0	0.00%	

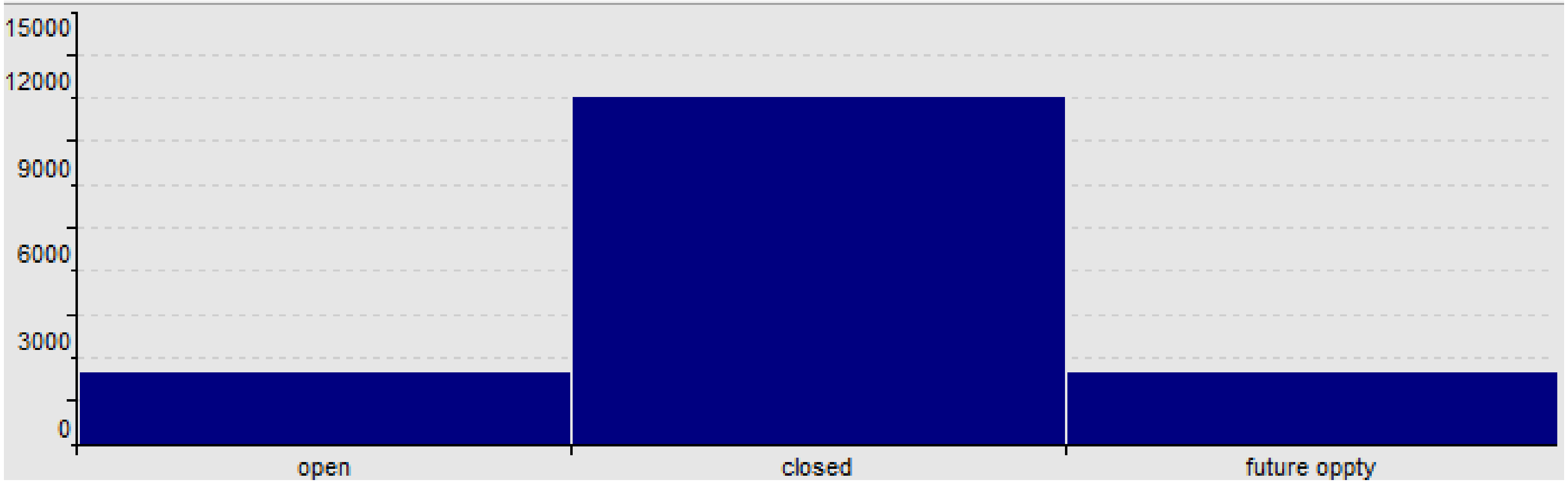
Data construction

- Derived attribute *Status* – merging of the opportunity stage columns



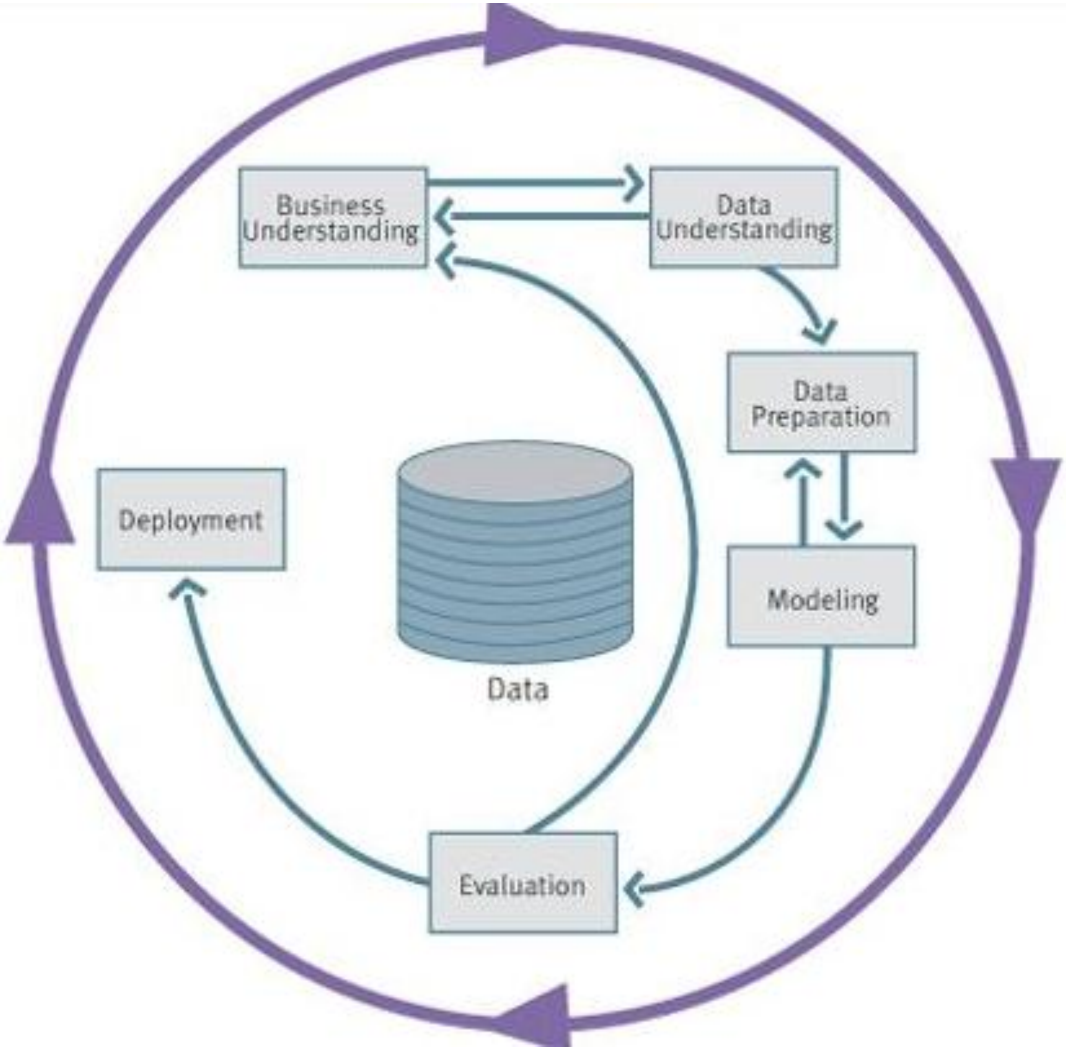
Data construction

- Derived attribute *Closed* – to compute Closing ratio
- Category *closed* in the succedent + various attributes in antecedent =>
- closing ratio = confidence of the rule



$$\frac{\text{closed lost opp} + \text{closed won opp}}{\text{closed lost opp} + \text{closed won opp} + \text{future opp} + \text{open opp}}$$

Project overview



Iteration 1 (IT1)

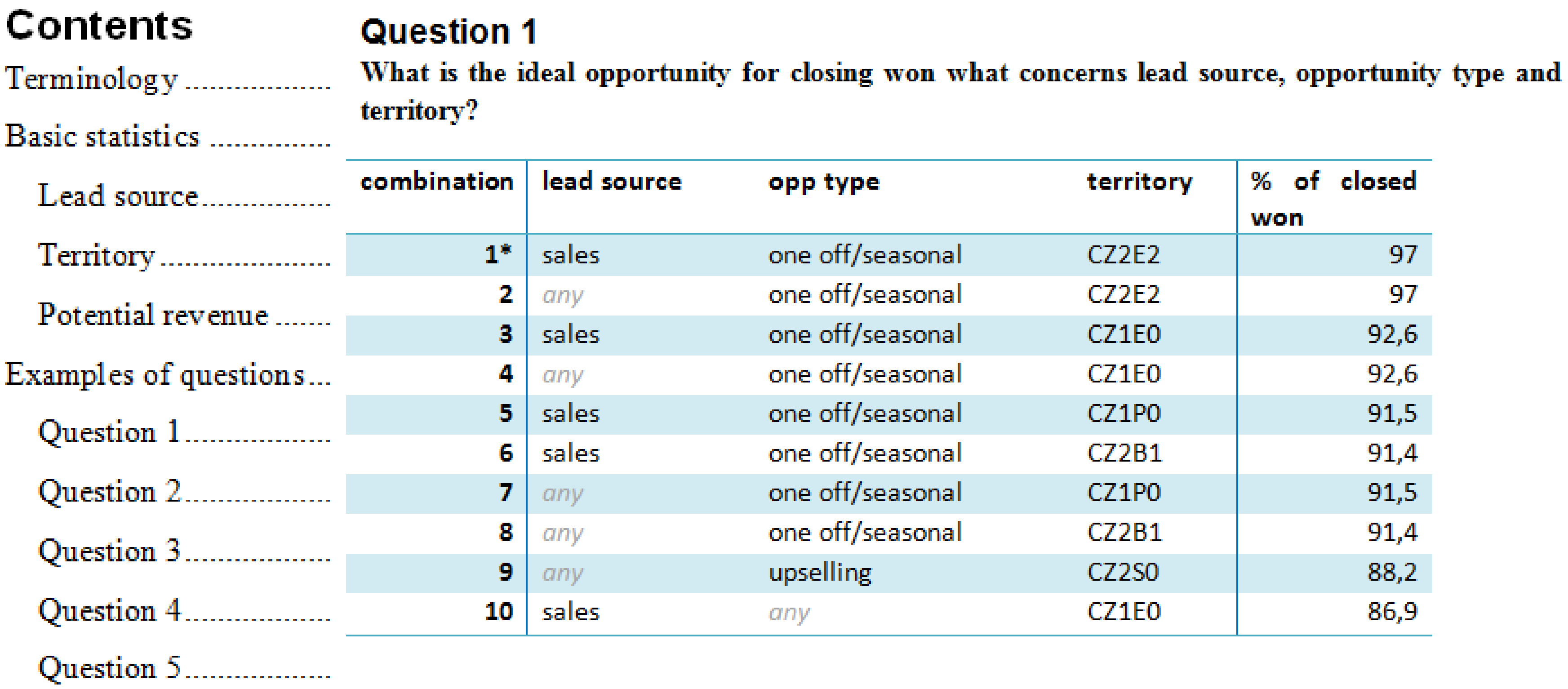
- □ Business understanding
- □ Data understanding
- □ Data preparation
- □ **Modelling**
- □ Evaluation

Iteration 2 (IT2)

- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

IT1 - MODELLING

- First analysis as a demonstration of possibilities of the LISp-Miner System



Contents

Question 1

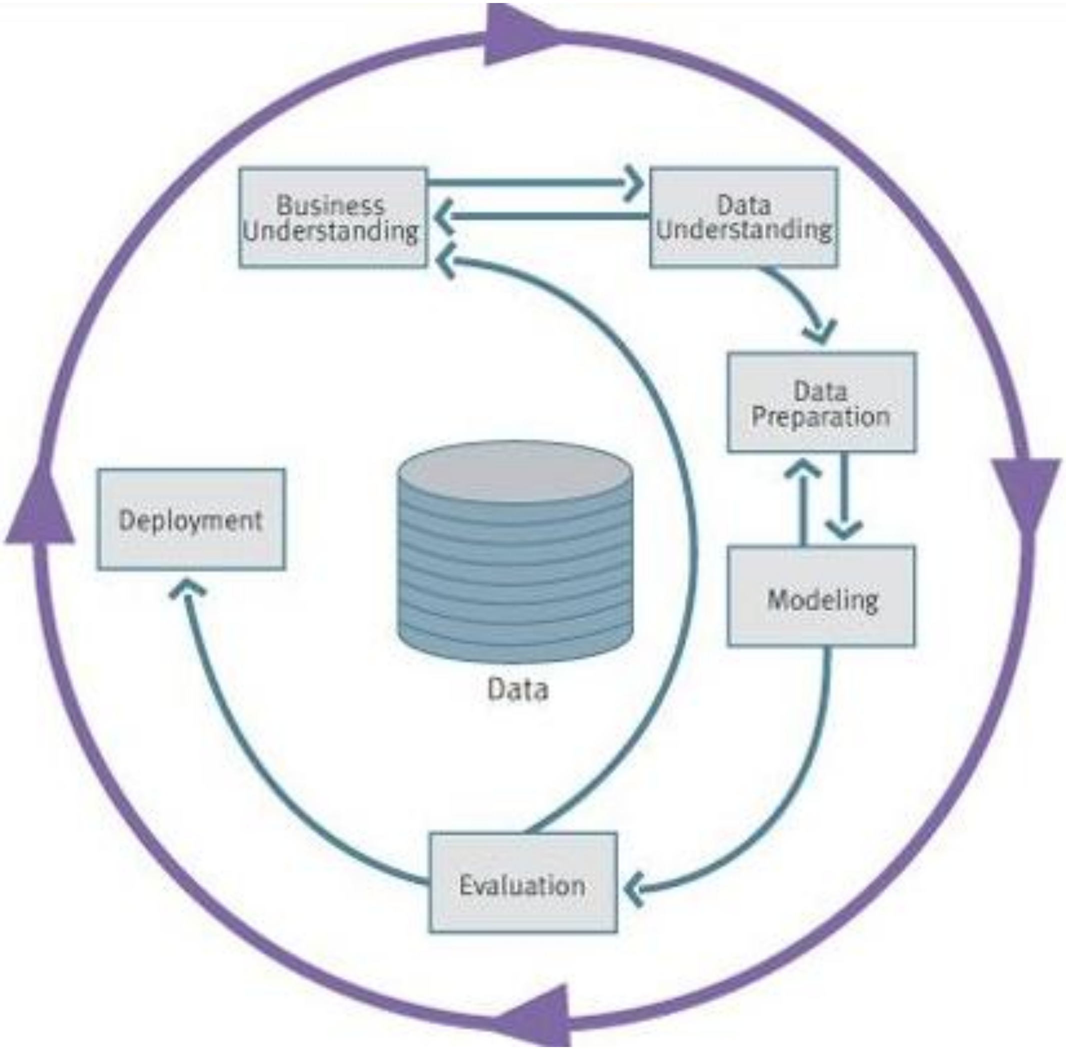
What is the ideal opportunity for closing won what concerns lead source, opportunity type and territory?

combination	lead source	opp type	territory	% of closed won
1*	sales	one off/seasonal	CZ2E2	97
2	*any*	one off/seasonal	CZ2E2	97
3	sales	one off/seasonal	CZ1E0	92,6
4	*any*	one off/seasonal	CZ1E0	92,6
5	sales	one off/seasonal	CZ1P0	91,5
6	sales	one off/seasonal	CZ2B1	91,4
7	*any*	one off/seasonal	CZ1P0	91,5
8	*any*	one off/seasonal	CZ2B1	91,4
9	*any*	upselling	CZ2S0	88,2
10	sales	*any*	CZ1E0	86,9

Facts taken into account

- Managers do not have time – the document should not be too extensive
- They are not interested in how the software works – keep it as simple as possible, hide all unnecessary technical details
- Prerequisite: managers know what the data represents – the data description is not presented, because it would extend the document to an undesirable length

Project overview



Iteration 1 (IT1)

- □ Business understanding
- □ Data understanding
- □ Data preparation
- □ Modelling
- □ **Evaluation**

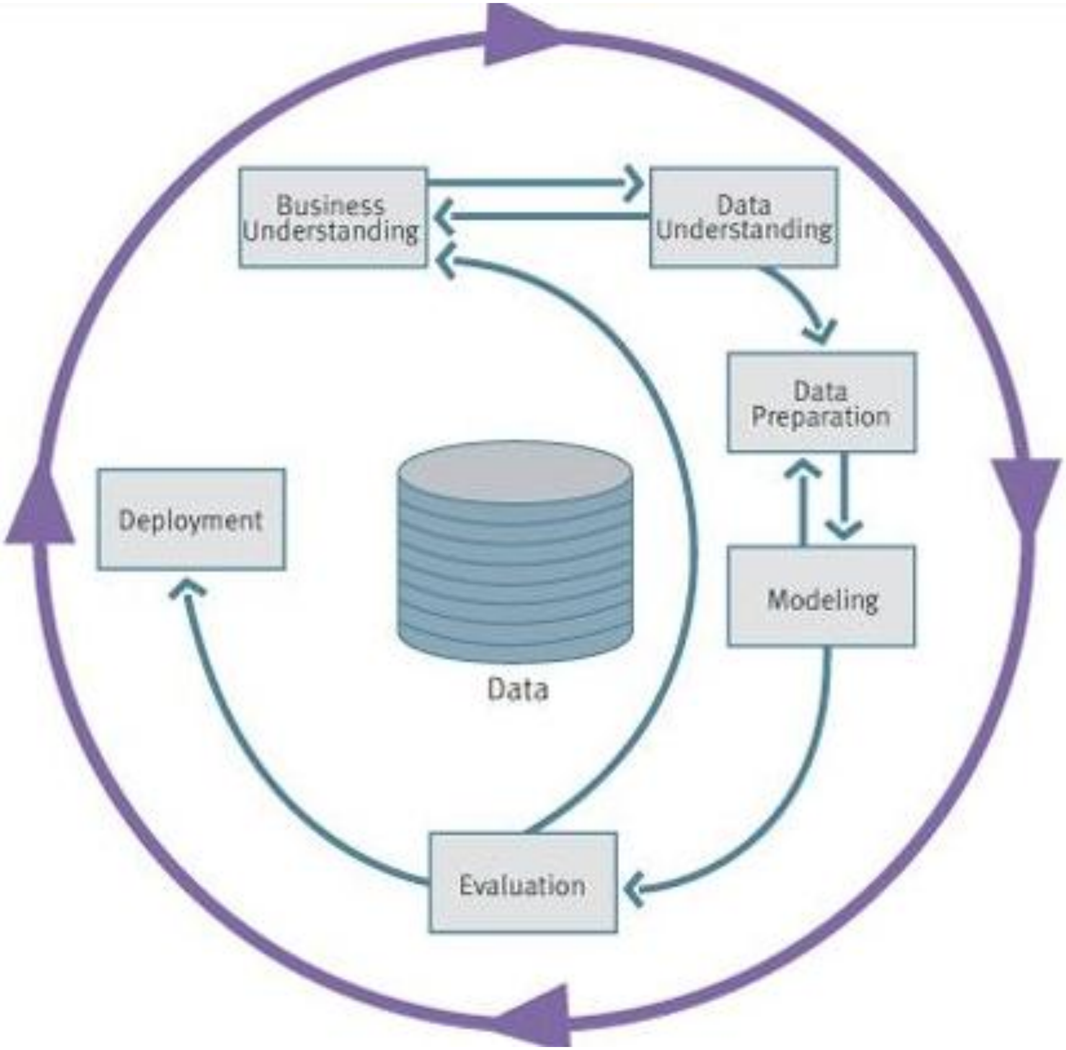
Iteration 2 (IT2)

- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

IT1 - EVALUATION

- Meeting with the experts
- The results of the first analysis are promising, however there are some inaccuracies.
- The way the results are presented to the business experts is comprehensible
- The Activity extract will be made available to answer the second analytical question

Project overview



Iteration 1 (IT1)

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation

Iteration 2 (IT2)

- After the meeting with the experts
- Includes comments and observations from the first iteration
- New portion of data obtained

ITERATION 2

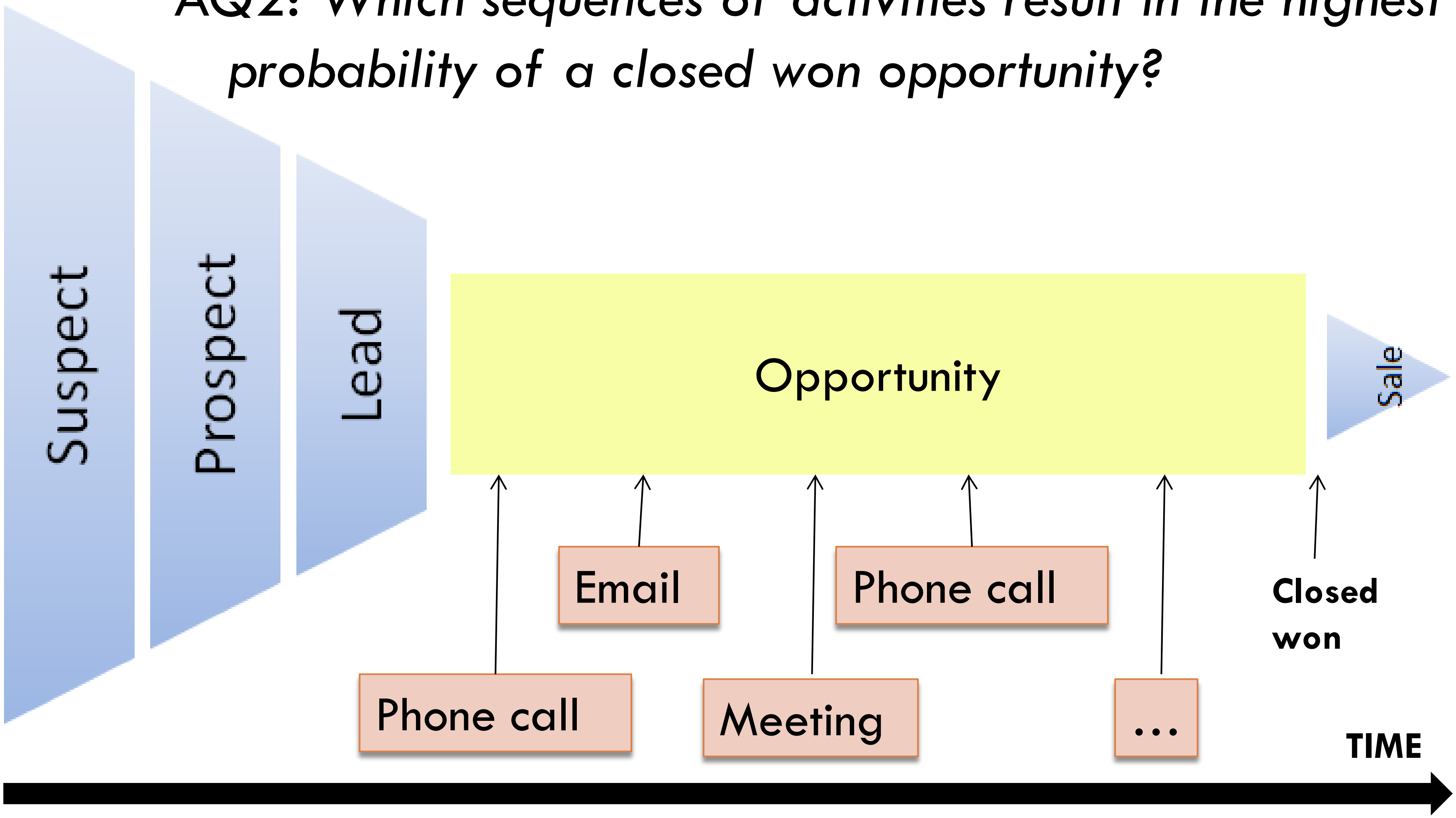
- Includes comments and observations from the first iteration
- New portion of data obtained – Activity extract
  - To answer second analytical question
- Data understanding, Data preparation made for the Activity extract

Issues in the second iteration

- Inaccuracies in processes (closed won)
- Data quality issues
- Duplicate rows in the data
- How to represent sequence of activities

How to represent sequence of activities

AQ2: Which sequences of activities result in the highest probability of a closed won opportunity?



1:n relation between opportunity and activity

How to make a single matrix suitable for analysis with LISp-Miner?

Number of activities performed during existence of an opportunity

number of activities performed	number of opportunities
0	1922
1	3270
2	3025
3	2177
4	1442
5	1058
6	750
7	547
8	358
9	292
10	252
…	…

number of activities performed	number of opportunities
…	…
88	1
89	1
96	1
109	1
114	1
121	1
134	1
141	1
sum	16121

Number of all activities	61 298
median category	2
average number of activities performed	3.80

Representing sequences of activities

- □ Maximum number of activities taken into account
- □ Type of activity
- □ Length of sequences
- □ Measuring time distance of opportunity and activity

Proposed derived attributes characterising sequence of activities

attribute	meaning	meaning of null value
no_of_act	number of activities performed during the opportunity	*no null values*
opp_A1_dist	days between creation of the opportunity and completion of the first activity	the opp has no activities performed
A1_type	type of the first activity	the opp has no activities performed
A1_A2_dist	days between completion of the first activity and completion of the second activity	the opp has less than 2 activities performed
A2_type	type of the second activity	the opp has less than 2 activities performed
A2_A3_dist	days between completion of the second activity and completion of the third activity	the opp has less than 3 activities performed
A3_type	type of the third activity	the opp has less than 3 activities performed
seq_3	sequence of the types of the first three activities	the opp has less than 3 activities performed
seq_5	sequence of the types of the first five activities	the opp has less than 5 activities performed
seq_10	sequence of the types of the first ten activities	the opp has less than 10 activities performed
reduced_3	boolean attribute expressing whether the sequence of activities was longer than 3	the opp has less than 3 activities performed
reduced_5	boolean attribute expressing whether the sequence of activities was longer than 5	the opp has less than 5 activities performed
reduced_10	boolean attribute expressing whether the sequence of activities was longer than 10	the opp has less than 10 activities performed

Creating one matrix with proposed derived attributes

→

Too complicated

IT2 – Modelling

- „Basic“rules - interesting (changeable) attributes in the antecedent and an indicator in the succedent
- More complex rules – combinations of interesting attributes in the antecedent and an indicator in the succedent

Which combinations of salesman and lead source have the highest revenue?

ANTECEDENT

More complex rule

Basic rule 1

Basic rule 2

SUCCEDENT

More complex rule

Basic rule 1

Basic rule 2

#	**Salesman**	**Lead source**	**% of opps with potential revenue higher than 70000**		
			Salesman and lead source together	Salesman alone	Lead source alone
1	KA	Sales	77.5 %	76.6 %	16.0 %
2	FS	IMP 2010	54.4 %	38.0 %	20.4 %
3	FS	Campaign Squeeze TNT	45.6 %	38.0 %	31.3 %
4	FS	Sales	37.2 %	38.0 %	16.0 %
5	Other	Sales	11.1 %	11.5 %	16.0 %
6	TS	Sales	9.2 %	9.3 %	16.0 %
7	Mic	Sales	6.3 %	6.2 %	16.0 %

IT2 - EVALUATION

- The results of the second analysis were sent to business experts
- No response

Observations, recommendations

- Business understanding
  - deployment of the analysis should be very concretely defined at the very beginning
  - formulating of business aims
  - motivation of the company
- Data understanding
  - abstract from the complexity of the data
  - no data description available - anticipate the meaning and ask for feedback when you have something to offer
  - identify indicators

Observations, recommendations

- Data preparation
  - always test the data for duplicate rows
  - consider whether answering an analytical question is worth of time and effort in the data preparation phase

Observations, recommendations

- Modelling
  1. Create basic rules – place interesting attributes in the antecedent and an indicator in the succedent
  2. Create more complex rules – combinations of interesting attributes in the antecedent and an indicator in the succedent
  3. Compare the rules generated in point 2 with those generated in point 1 – potentially interesting are those rules that have higher confidence than the basic rules

Observations, recommendations

- Ac4ft-Miner and SD4ft-Miner are too complicated for domain experts
- Motivation of the domain experts?

COOPERATION WITH THE MARKET RESEARCH COMPANY

Market research company

- Shopper behaviour – typical questions:
  - What to do to make the customer come back again?
  - Which products should be discounted?
  - Which products should be placed together in the shopping unit?
  - What new items should be introduced?

Data

- Shopping baskets
- Shopping lists

- Thousands of columns = products
- Thousands of rows = baskets, lists

Sparse binary data – example

R	K10001	RK10001	K10002	RK10002	K10003	RK10003	K10004	RK10004	K10005	RK10005	K10006	RK10006	K10007	RK10007	K10008	RK10008	K10009	RK10009	K10010
0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	1		0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1
0	1		1		0	0	0	0	0	0	1		0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	1		1		0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		1		0	0	0	0	0	0	0	0	0	0	0	0	1
0	1		1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1		1		0	0	1		1		1		0	0	0	0	1		1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1		1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1		1		1		1		1		0	0	0	0	0	0	1		0

Possible research areas

- □ Association rules?
- □ Clustering?

Bit string approach in LISp-Miner

Sparse binary data

QUESTIONS?