Questionnaire about the skills necessary for people working with Big Data in the Statistical Organisations

Preliminary results of the survey (19.08 2014)

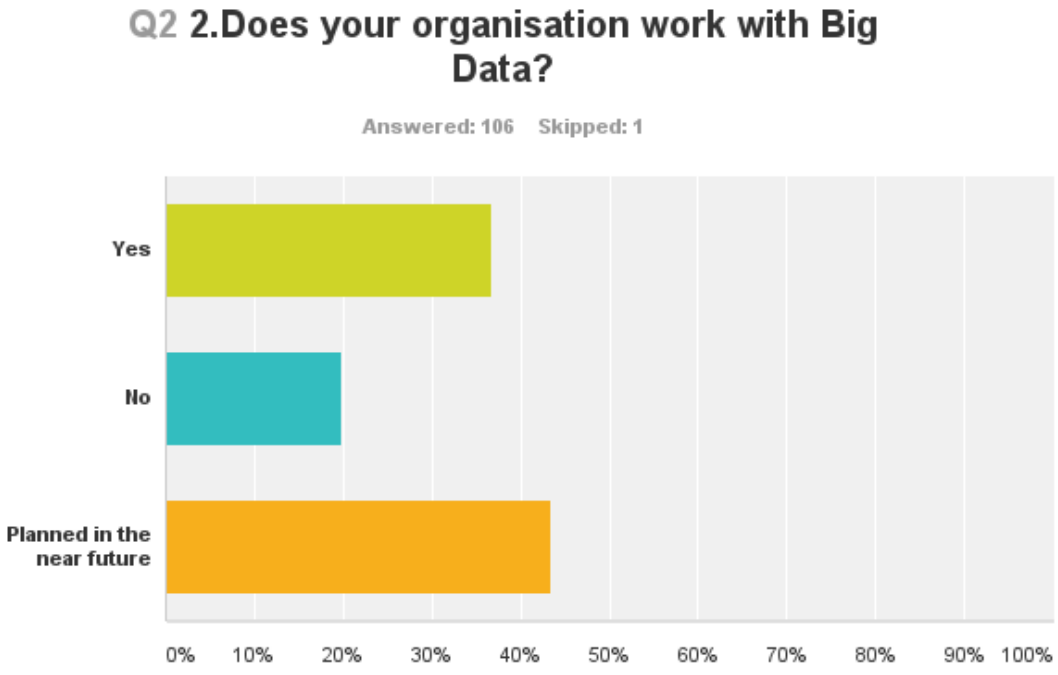
More detailed analysis will be prepared by October 2014

Total of 137 were received, because some of the responses were not complete 107 responses were used for the analysis.

Detailed results by question:

Question 2: Does your organisation work with Big Data?

Answer Options	Response Percent	Response Count
Yes	36.8%	39
No	19.8%	21
Planned in the near future	43.4%	46
Comments		32
	answered question	106
	skipped question	1



Comments:

1	We are just in the beginning
2	We will organize trainings on big data in statistics
3	Currently we are working in some experiments and probe of concept
4	We are a few people following the topic, but has no specific projects or programmes at the moment.
5	With the announcement of a national data sharing and accessibility policy, our government has defined the objective to facilitate access to the Government owned shareable data generated using public funds in machine readable format across the country in a pro-active and periodically updatable manner. A large quantum of the Government data which is currently generated by various government organizations and institutions in the country remains inaccessible to civil society, although most of such data may be non-sensitive in nature and could be used by public for scientific, economic and social developmental purposes. Thus, started toward collating Big Data.
6	The sources of official statistics in our country: 1. Statistical surveys conducted by the national statistical organisation 2. Statistical surveys conducted by other producers of statistics (National Bank, Ministry of Finance, Ministry of Education, Ministry of Health and other) 3. Administrative data
7	A first working group has been constituted within the Institute to analyse this possibility
8	We don't have yet enough experience and capacity to work with big data
9	As in pilots, not yet for publication/production purposes.
10	Except perhaps with scanned data for prices
11	Trying to get access to electricity consumption data (hourly data per household)
12	We would propose to simplify the classification and only make a distinction between methodological and IT skills.
13	for pilot projects, yet
14	For pilot projects, yet.
15	Early phase
16	RnD-phase
17	Hard to say - define big data
18	Some experimental work taking place to my knowledge
19	Developing area
20	We have plans to create a data warehouse under IPA project
21	There are some intentions to learn more about Big Data but no current projects.
22	We are, however participating in the UNECE Big Data Project
23	I am interested in statistical methodology, mainly with data quality
24	We work with some large administrative data sources but I would expect that some of the new sources will require quite different approaches from a technology angle.
26	We have established a 12 month project to understand the potential but also the challenges of using big data within official statistics - still in a research phase
27	The Production Department that I represent, does not work with Big Data
28	we have just started a Big Data project
29	We intend to develop a pilot for using Big Data as a source of statistical information in 2015 - 2017
30	Supermarket scanner data is in use for quite a while
31	I work in my own research which is not strictly connected to what my office does
32	Evaluation of technology and possibilities
33	Started experiments
34	Analysis and POC in progress
35	We did already some pilots, but I'm not exactly informed.
36	Just commenced work - at very early stage

Question 3: How important do you think are those skills for working with Big Data? Please rate them from 1 (not important) to 5 (very important)

3.1 IT skills, ability to use the following programs, languages, technologies and software such as...

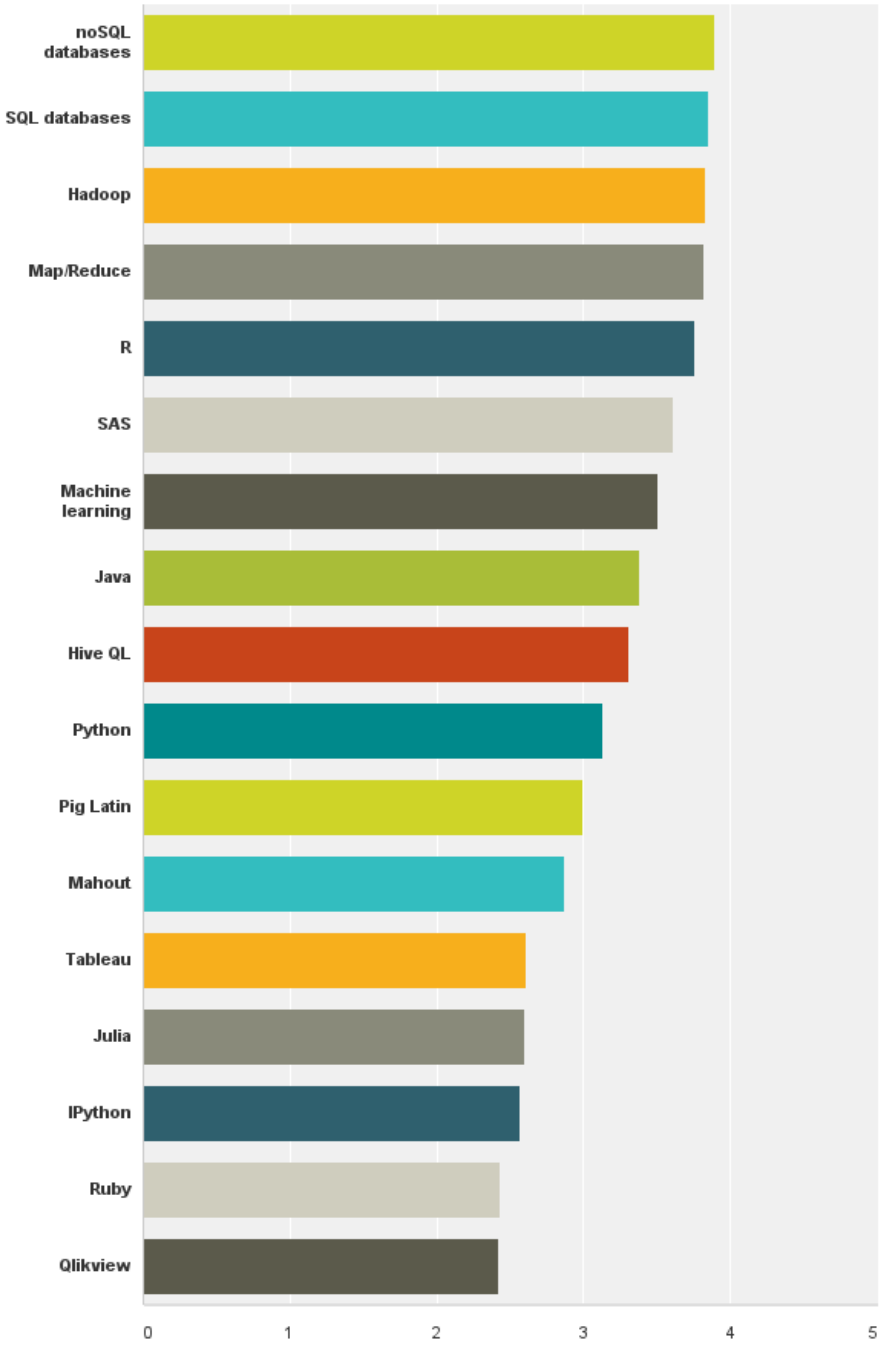
Answer Options	1	2	3	4	5	Rating Average	Response Count
noSQL databases	5	4	10	36	24	3.89	79
SQL databases	4	8	13	35	28	3.85	88
Hadoop	9	7	13	20	39	3.83	88
Map/Reduce	8	1	17	18	30	3.82	74
R	5	8	15	35	25	3.76	88
SAS	8	4	22	30	21	3.61	85
Machine learning	7	9	20	21	21	3.51	78
Java	6	16	19	23	18	3.38	82
Hive QL	10	3	20	26	9	3.31	68
Python	8	15	23	19	11	3.13	76
Pig Latin	13	9	13	23	6	3.00	64
Mahout	11	18	18	15	8	2.87	70
Tableau	17	14	19	12	5	2.61	67
Julia	11	17	25	6	4	2.60	63
IPython	13	17	22	6	5	2.57	63
Ruby	14	21	19	10	1	2.43	65
Qlikview	20	15	24	10	2	2.42	71
Other (please specify)							18
					answered question		100
					skipped question		7

Other answers:

1 IT depends on the kind of source and its usage
2 Scala (5), Spark (5), D3 (or another visualization tool, 5)
4 Scala - 3, Java Script - 4
5 Hadoop administration, cluster management
6 STATISTICA
7 Other in memory tools like qlickview like sas visual analytic
8 SPSS
9 Don’t work directly on this side and it is a developing area, so identified what I have heard from staff in the Statistical side of the House
10 Note the above responses are preliminary. We have licensed Splunk and will look at wider applicability than initial business case in Security monitoring.
11 Not able to specify, as we are not working with Big Data
12 SPSS, Statistica
13 Linux (5), HDFS (4), Business Intelligence, Enterprise Architecture, Data management, Systems administrators with skills in administering, monitoring and supporting specific big data platforms
14 I'm not an IT worker and therefore it is difficult to answer to this question.
15 Scala, Linux, NB score of 1 above if not open source
16 Bash, awk and C/C++ (low-level and very fast!)
17 ETL software
18 No knowledge on IT skills required

Q3 3. How important do you think are those skills for working with Big Data? Please rate them from 1 (not important) to 5 (very important) 3.1 IT skills, ability to use the following programs, languages, technologies and software such as...

Answered: 100 Skipped: 7



3.2 Statistical skills, such as...

Answer Options	1	2	3	4	5	Rating Average	Response Count
Methodology for processing Big Data	4	1	3	23	72	4.53	103
Data mining	1	0	11	37	49	4.36	98
Standards for processing Big Data	5	3	6	44	45	4.17	103
Data management skills including documentation, registration, access control	6	5	22	35	32	3.82	100
Strong/power user of software such as Excel, SAS, SPSS	5	11	23	28	35	3.75	102
Ability to work with text analytics	2	7	29	44	17	3.68	99
Other (please specify)							7
						answered question	106
						skipped question	1

Other answers:

1 Assessment from a more general point of view and might vary individually

2 Mathematical statistics skills (5), multivariate data analysis (5)

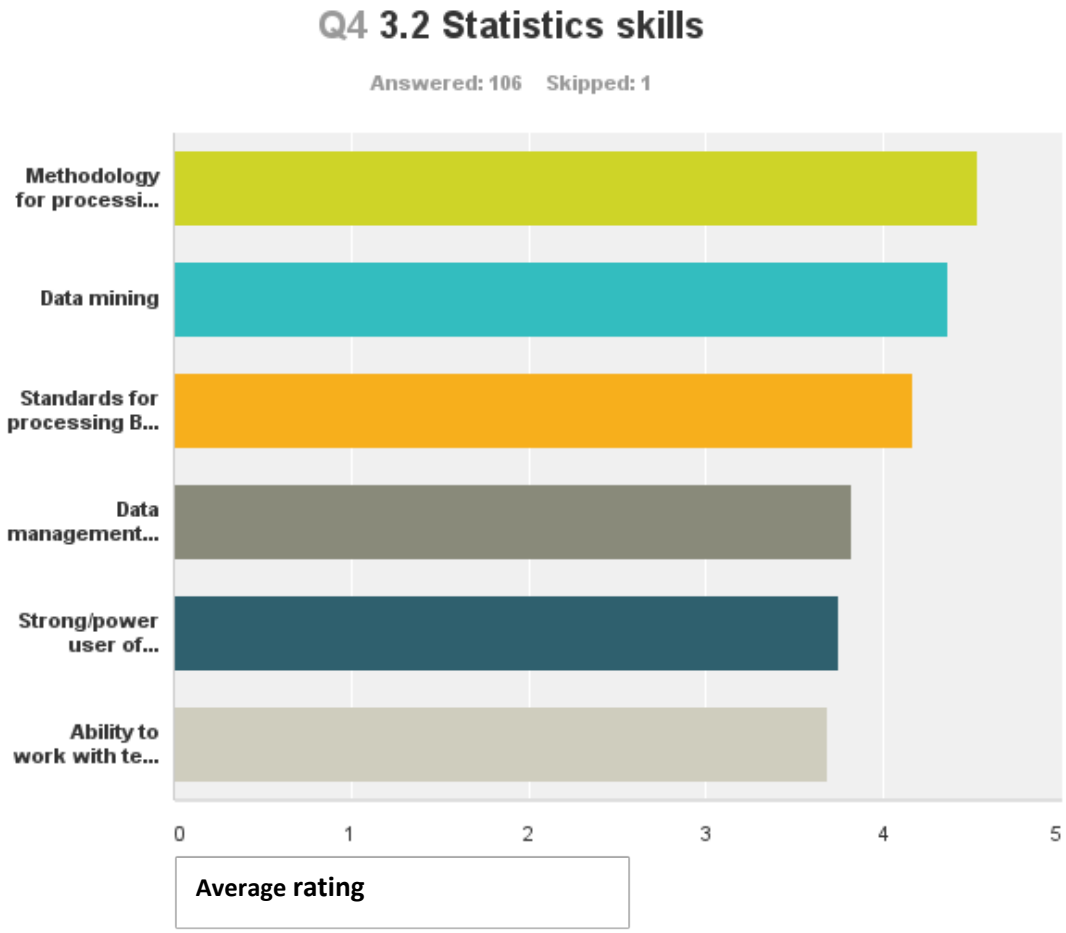
3 Statistical thinking, TQM, data quality

4 Not able to specify, as we are not working with Big Data

5 Analytic modelling techniques such as predictive modelling; subject matter knowledge

6 Since there is no standard yet, the ability to think outside the box is very important (there are no standard recipes yet)

7 Data integration skills

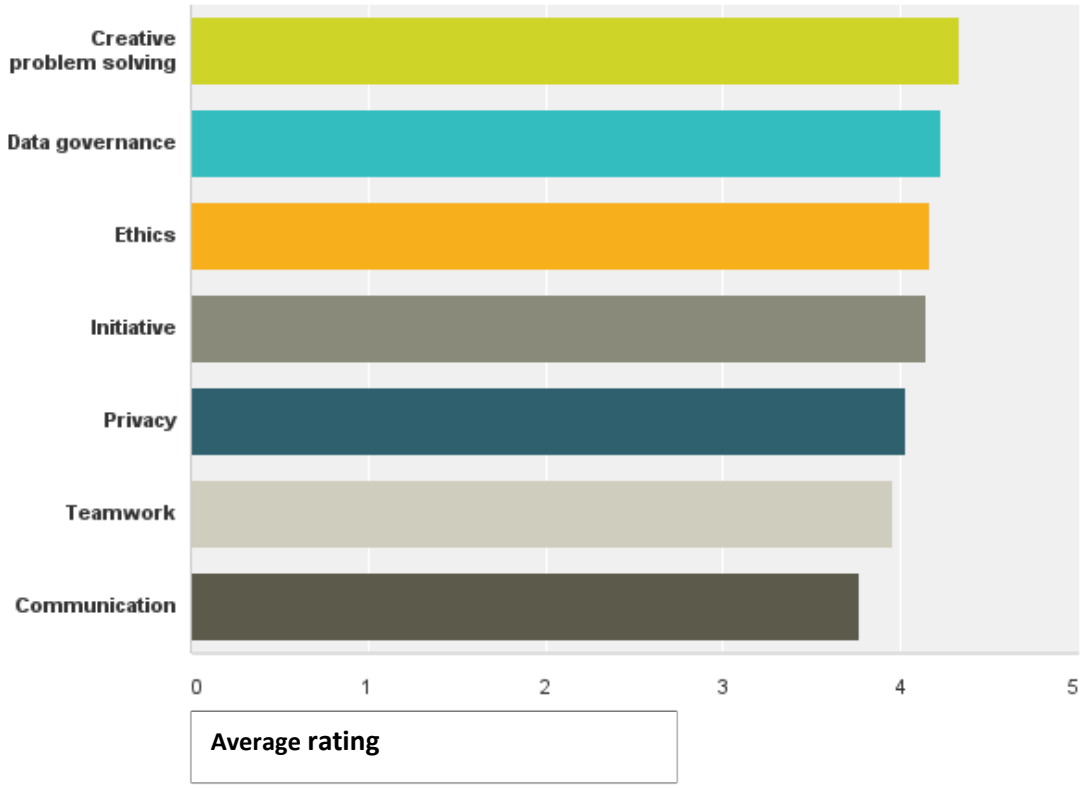


3.3 Other skills, such as...

Answer Options	1	2	3	4	5	Rating Average	Response Count
Creative problem solving	2	2	8	39	52	4.33	103
Data governance	1	1	16	37	44	4.23	99
Ethics	1	3	17	37	43	4.17	101
Initiative	2	5	12	39	42	4.14	100
Privacy	3	6	20	27	44	4.03	100
Teamwork	1	8	20	40	35	3.96	104
Communication	2	8	33	30	31	3.77	104
Other (please specify)							8
						answered question	105
						skipped question	2

Q5 3.3 Other skills

Answered: 105 Skipped: 2



Question 4: Which of the following skills you already have in your organisation and at what level?

4.1 IT skills, such as...

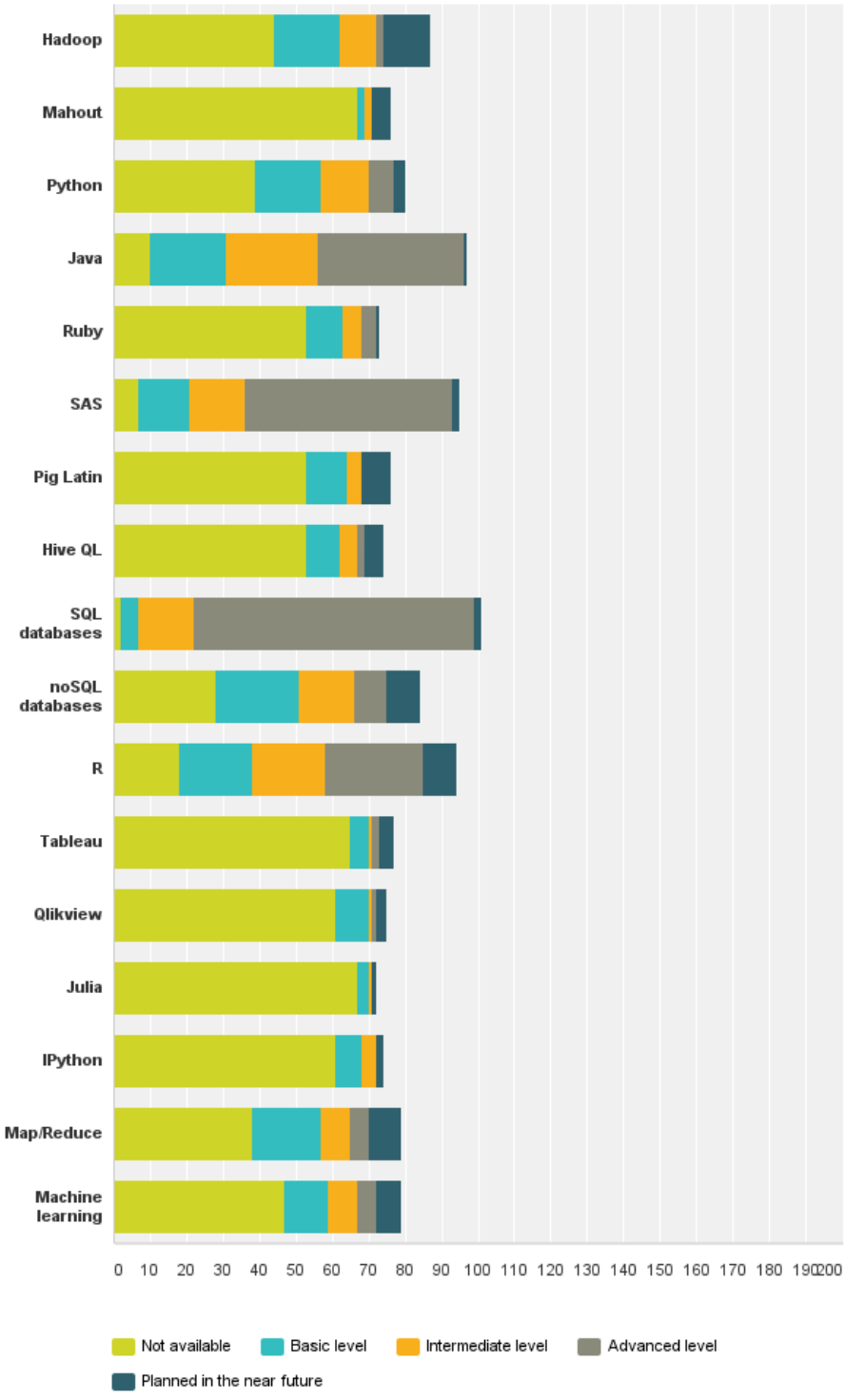
Answer Options	Not available	Basic level	Intermediate level	Advanced level	Planned in the near future	Response Count
Hadoop	44	18	10	2	13	81
Mahout	67	2	2	0	5	76
Python	39	18	13	7	3	77
Java	10	21	25	40	1	92
Ruby	53	10	5	4	1	71
SAS	7	14	15	57	2	89
Pig Latin	53	11	4	0	8	72
Hive QL	53	9	5	2	5	72
SQL databases	2	5	15	77	2	95
noSQL databases	28	23	15	9	9	79
R	18	20	20	27	9	90
Tableau	65	5	1	2	4	75
Qlikview	61	9	1	1	3	74
Julia	67	3	1	0	1	72
IPython	61	7	4	0	2	74
Map/Reduce	38	19	8	5	9	77
Machine learning	47	12	8	5	7	76
Other (please specify)						15
					answered question	99
					skipped question	8

Other answers:

1 Several tools might exist but relevance must be checked for individual application
2 Scala (basic, planned), Spark (basic, planned), D3 (basic, planned)
3 SPSS - advanced
4 Scala - 3 Java Script - 3
5 STATISTICA - maps and analyzes (advanced level)
6 SPSS level 4
7 SAS is considered very important in our organisation. We have just completed 3 courses in R for staff interested. Again this is not my side of the house, and not aware of some of those headings.
8 SPSS
9 Linux
10 Rhadoop is also available on the basic level.
11 I miss GPGPU programming abilities (CUDA C or Open-CL for example)
12 R is used for specific purposes and not widely.

Q6 4. Which of the following skills you already have in your organisation and at what level? 4.1 IT skills

Answered: 99 Skipped: 8

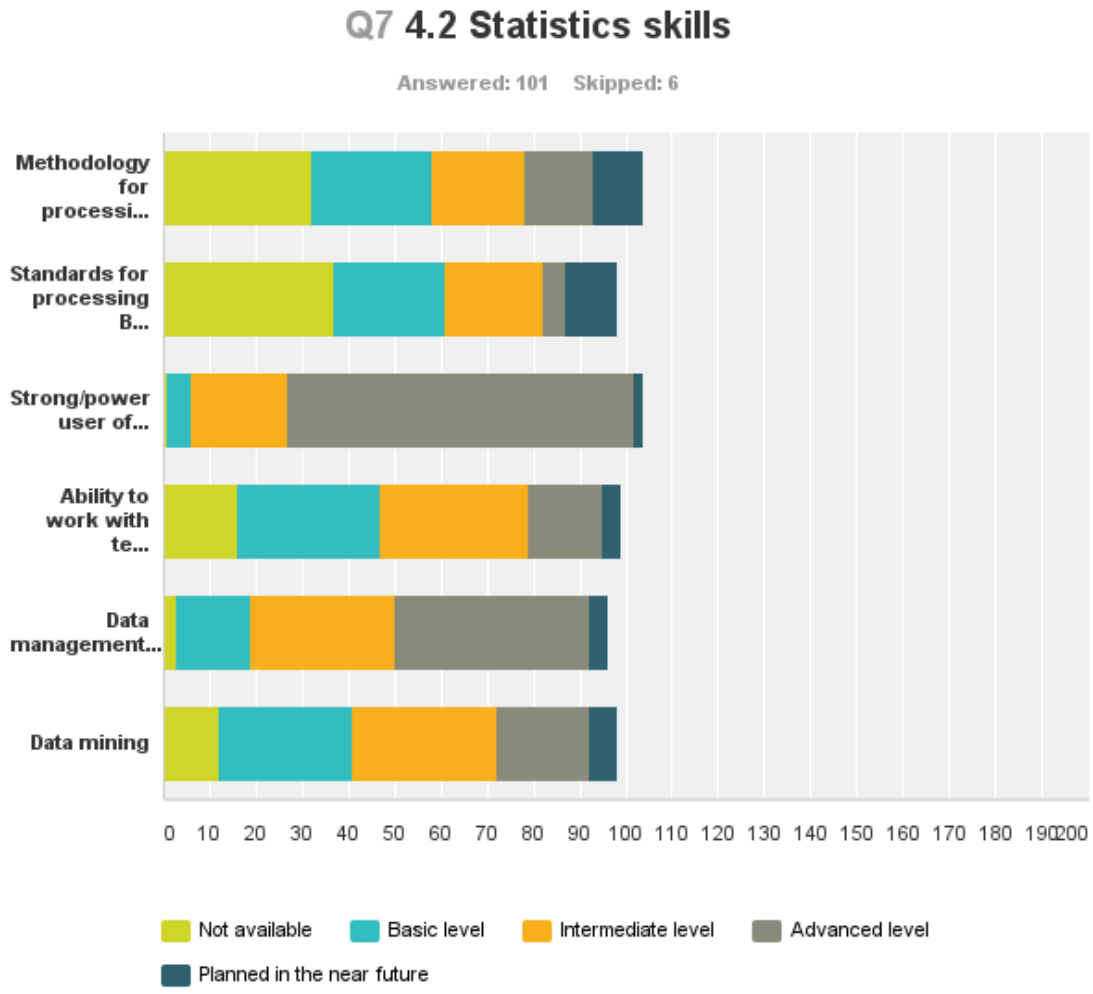


4.2 Statistical skills, such as...

Answer Options	Not available	Basic level	Intermediate level	Advanced level	Planned in the near future	Response Count
Methodology for processing Big Data	32	26	20	15	11	97
Standards for processing Big Data	37	24	21	5	11	96
Strong/power user of software such as Excel, SAS, SPSS	1	5	21	75	2	97
Ability to work with text analytics	16	31	32	16	4	96
Data management skills including documentation, registration, access control	3	16	31	42	4	94
Data mining	12	29	31	20	6	93
Other (please specify)						8
					answered question	101
					skipped question	6

Other answers:

1	Mathematical Statistics (Advanced, Planned), Multivariate Data Analysis (Intermediate, planned),
2	SAS is not available
3	Advanced user of Excel
4	Not too sure about this as its not my domain, hence middle markings
5	We have a big experience to work with administrative registers
6	Data Scientist
7	One may wonder if methodology for processing Big Data is already at an advanced level. I think our organization is approaching that.

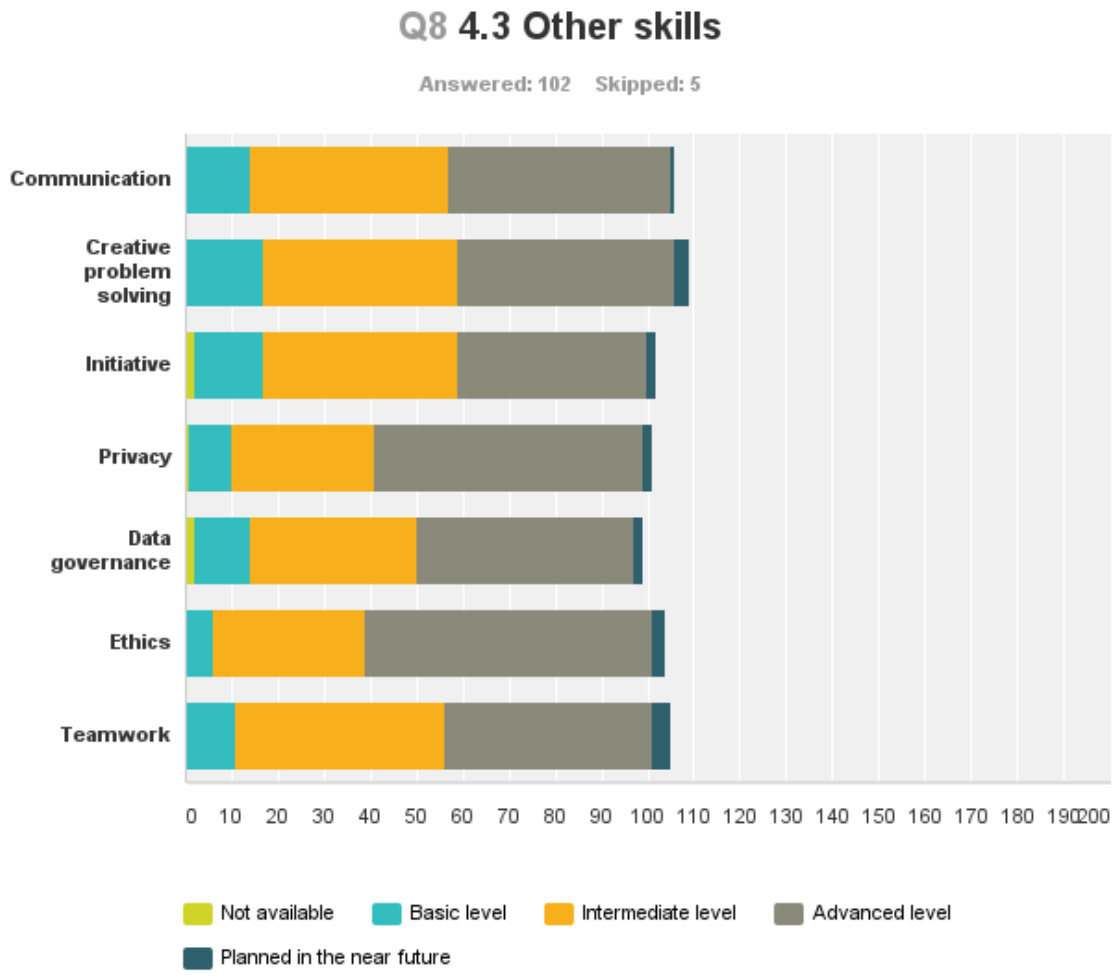


4.3 Other skills, such as...

Answer Options	Not available	Basic level	Intermediate level	Advanced level	Planned in the near future	Response Count
Communication	0	14	43	48	1	101
Creative problem solving	0	17	42	47	3	101
Initiative	2	15	42	41	2	98
Privacy	1	9	31	58	2	99
Data governance	2	12	36	47	2	97
Ethics	0	6	33	62	3	99
Teamwork	0	11	45	45	4	101
Other (please specify)						5
					answered question	102
					skipped question	5

Other answers:

1	All skills are available at high level for the traditional system of generating and releasing statistics; adoption and modifications for big data purposes seem possible
2	Except the listed skills, our staff periodically participate in training courses in such areas as: 1. The acquisition of knowledge in economics, law, improvement of management skills 2. Improving professional knowledge in statistics 3. Improvement of knowledge on modern information technologies 4. The acquisition and improvement of foreign language skills
3	Difference between levels is not clear when applied to listed skills. How does intermediate level communication differ from basic level communication in practice?
4	Of course we have skilled in these areas but not when it comes to Big Data...
5	The use of visualisation methods for Big Data is missing here. Because of the amount of data, visualisation methods are essential in getting insight in the effect of the various analyses steps.



Question 5: Please indicate in which areas you have training in your statistical organisation and indicate if you have training materials that you can share or recommend? (Training materials include: books, internet resources, training materials developed in the Statistical Organisations, etc).

5.1 IT skills, such as...

Answer Options	Training	Training materials that you can share or recommend	Response Count
SAS	62	10	63
SQL databases	54	5	54
Java	32	4	32
R	28	8	31
Hadoop	7	3	8
Python	7	1	8
noSQL databases	7	1	8
Map/Reduce	6	3	7
Pig Latin	4	2	5
Hive QL	4	2	5
Machine learning	3	2	5
Ruby	2	1	3
Qlikview	2	1	3
IPython	2	1	3
Mahout	1	1	2
Tableau	1	1	2
Julia	1	1	2
If you have any training material, please provide us the title or the link to the website			14
		answered question	80
		skipped question	27

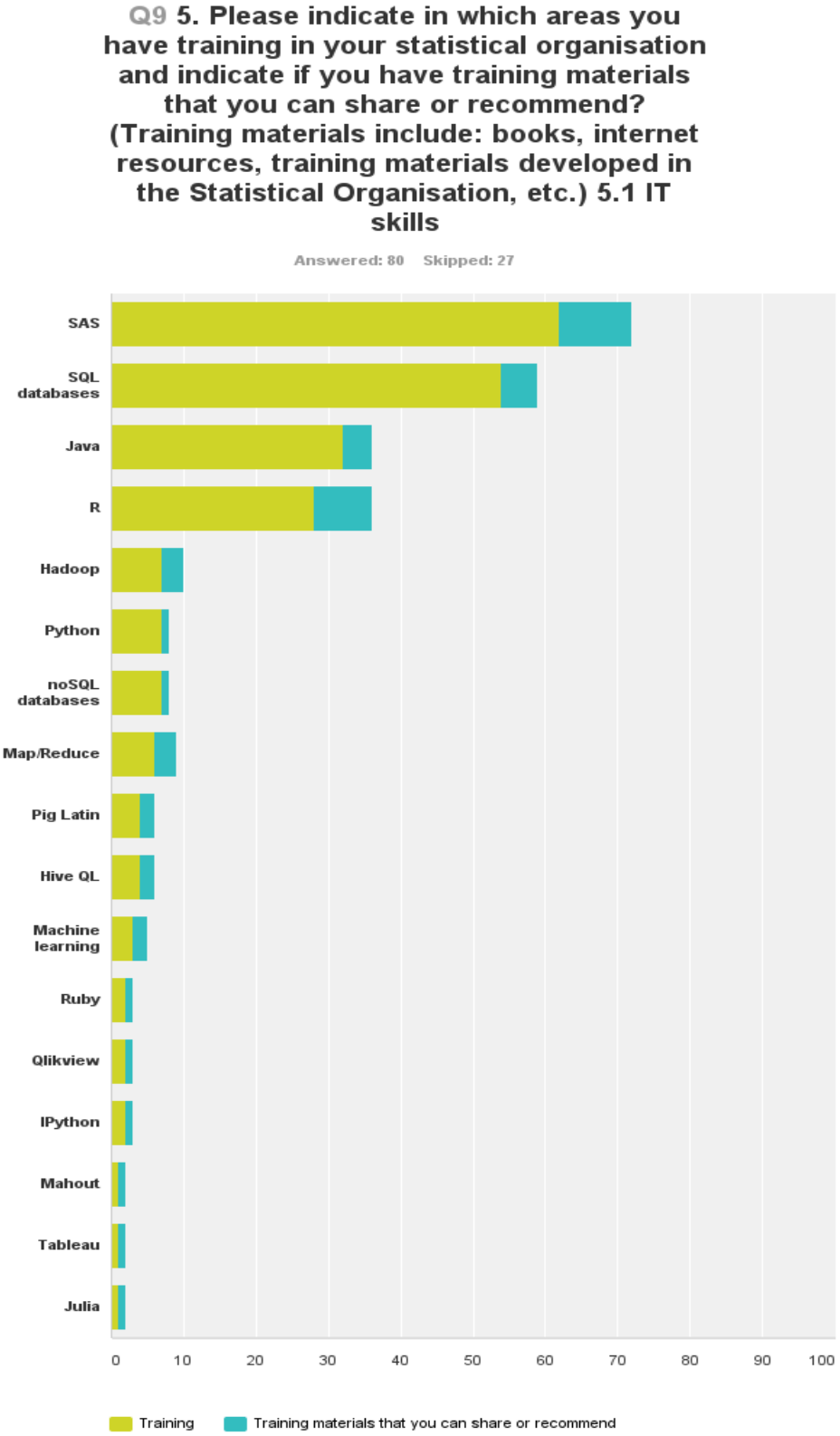
If you have any training material, please provide us the title or the link to the website:

1 There are no extended big data projects running until now
2 I need to follow up and see if our training partners are willing to share training materials.
3 Material is in the form of PPT slides, just note that it's in our national language
4 To R: various readers and best practices developed by our employees. Mainly in our national language.
5 Training material is not specific to the context of big data (standard SAS training).
6 No training!
7 1. Beginning T-SQL with MS SQL Server 2005 and 2008, Paul Turley, Dan Wood 2. ACCESS 2010 Bible, Michael R. Groh 3. MS SQL Server 2008 Bible, Paul Nielsen, Uttam Parui
8 Internal training material available in our national language for SAS and SQL.
9 SPSS
10 We don't currently have training material for big data in any of these areas. We would have some material for Java and potentially R but not specific to big data processing\analytics.

part of our Big Data project is to compile such a list from internet resources. Not currently
11 available but will be at a later date

For many IT-skills on-line training sources are available (see the Coursera website). We are
12 very experienced in using R to analyse Big Data

13 https://www.coursera.org/course/ml

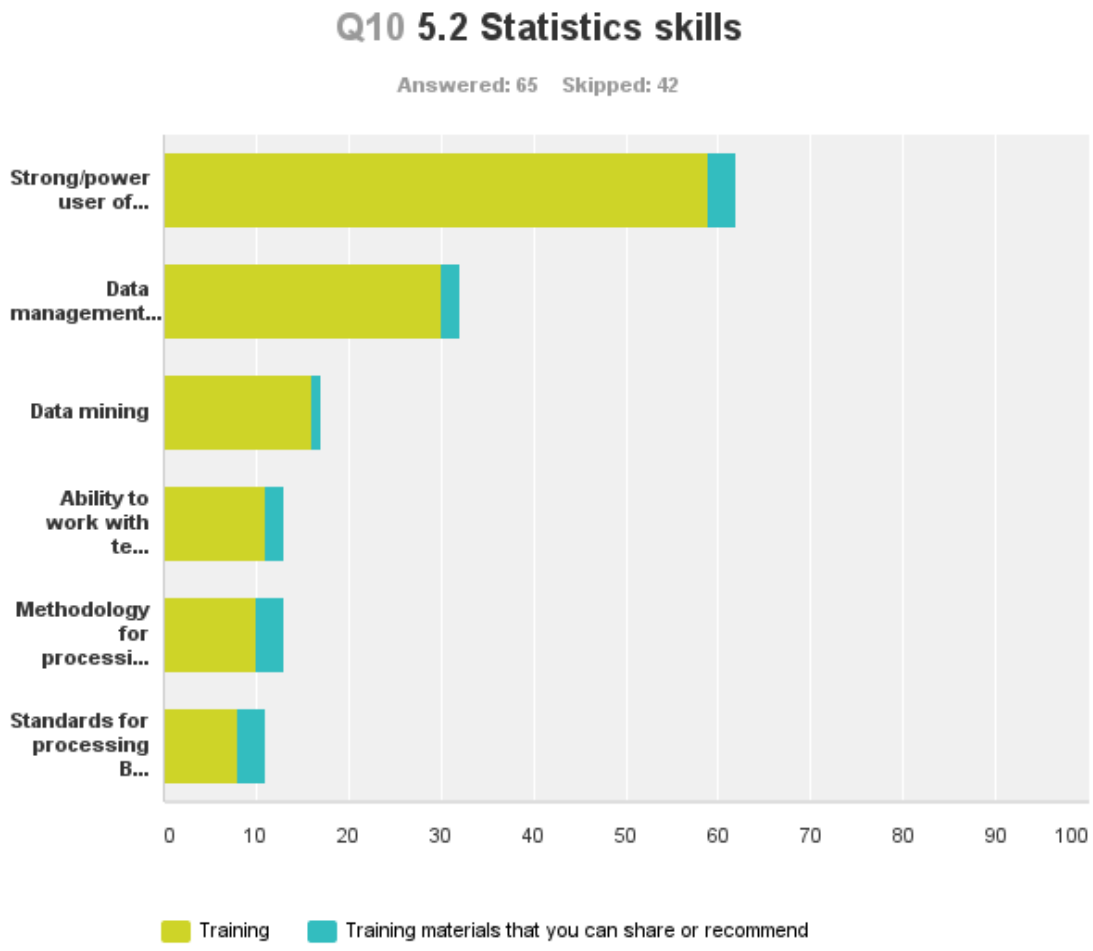


5.2 Statistical skills, such as...

Answer Options	Training	Training materials that you can share or recommend	Response Count
Strong/power user of software such as Excel, SAS, SPSS	59	3	60
Data management skills including documentation, registration, access control	30	2	30
Data mining	16	1	16
Ability to work with text analytics	11	2	12
Methodology for processing Big Data	10	3	12
Standards for processing Big Data	8	3	10
If you have any training material, please provide us the title or the link to the website			9
		answered question	65
		skipped question	42

If you have any training material, please provide us the title or the link to the website:

1	Cooperation with the ESS is planned; We take part in a Eurostat Working Group on Big data
2	No training, except Excel
3	1. EXCEL 2010 Bible 2. EXCEL 2010 PowePivot for the Data Analyst 3. The Excel Analyst's Guide to Access, Michael Alexander

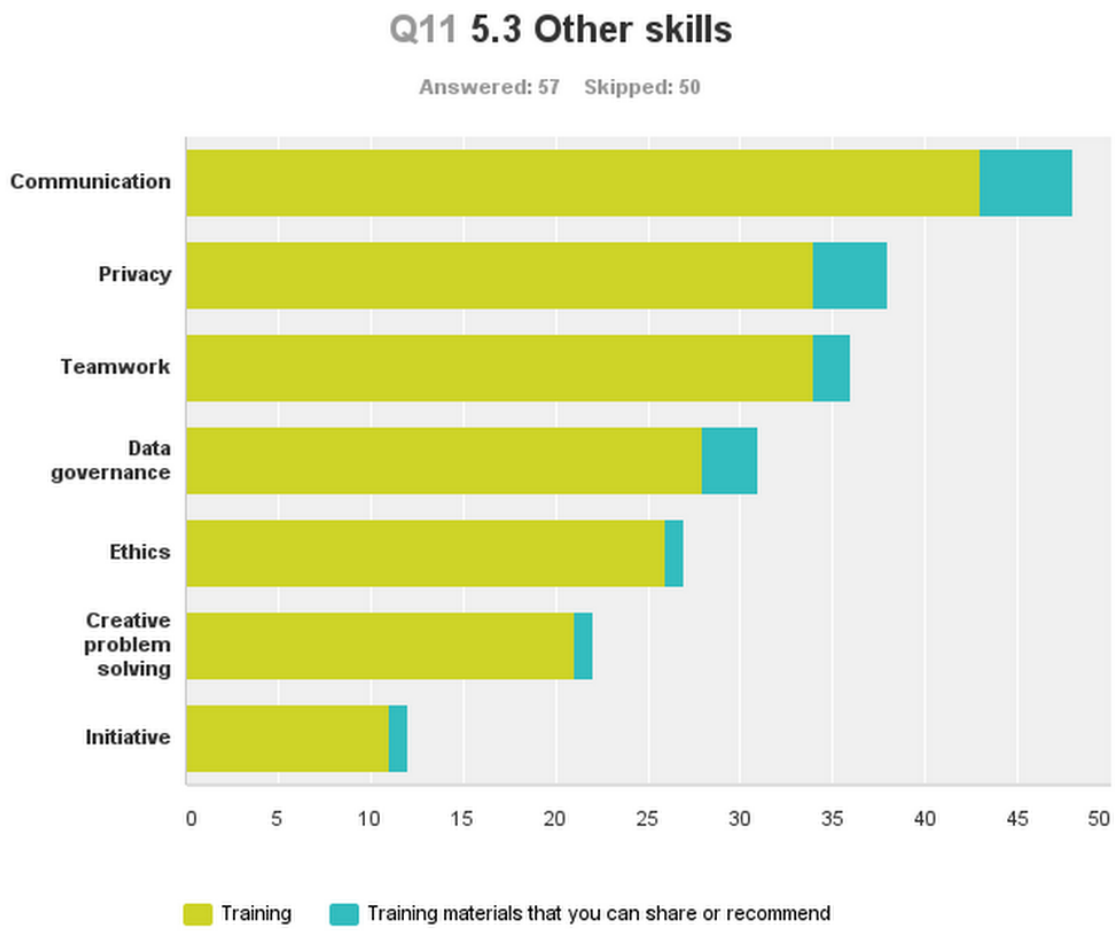


5.3 Other skills, such as...

Answer Options	Training	Training materials that you can share or recommend	Response Count
Communication	43	5	45
Privacy	34	4	35
Teamwork	34	2	35
Data governance	28	3	29
Ethics	26	1	27
Creative problem solving	21	1	21
Initiative	11	1	12
If you have any training material, please provide us the title or the link to the website			8
		answered question	57
		skipped question	50

If you have any training material, please provide us the title or the link to the website:

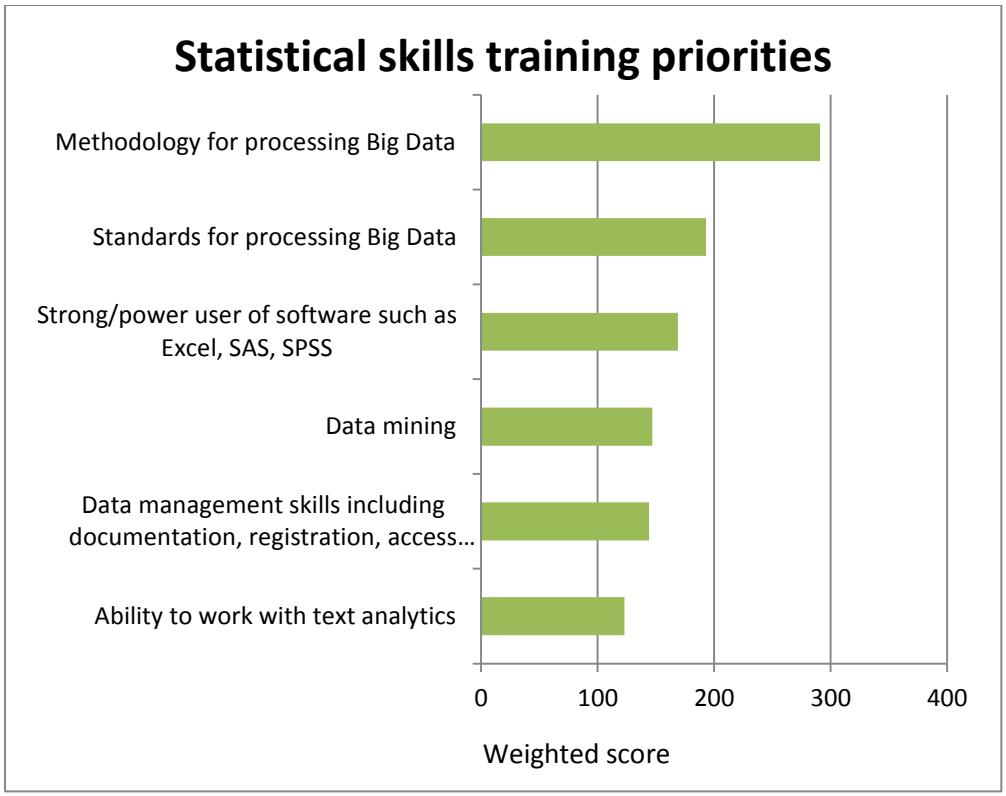
1	All skills are provided at high level for traditional statistical system, adoptions and specifications to big data seem possible
2	The Argus documentation.
3	Lean SixSigma by UNC Plus Delta
4	We could share material on the ESTP course on Big Data
5	We have office notices and also courses developed on Confidentiality. For communication we have improved the user satisfaction survey and developed documents: how to write a press release, how to organize a press conference. For Ethics, we have translated the ISI's Declaration on Professional Ethics into our national language and posted on its website, approved by the Resolution of the State Council on Statistics to apply it in the activity of our organisation and to suggest to others dealing with the statistical activity to follow
6	the Declaration.
7	I wonder if the two topics that are not checked can be trained with a program.
8	See also ESTP Training programme



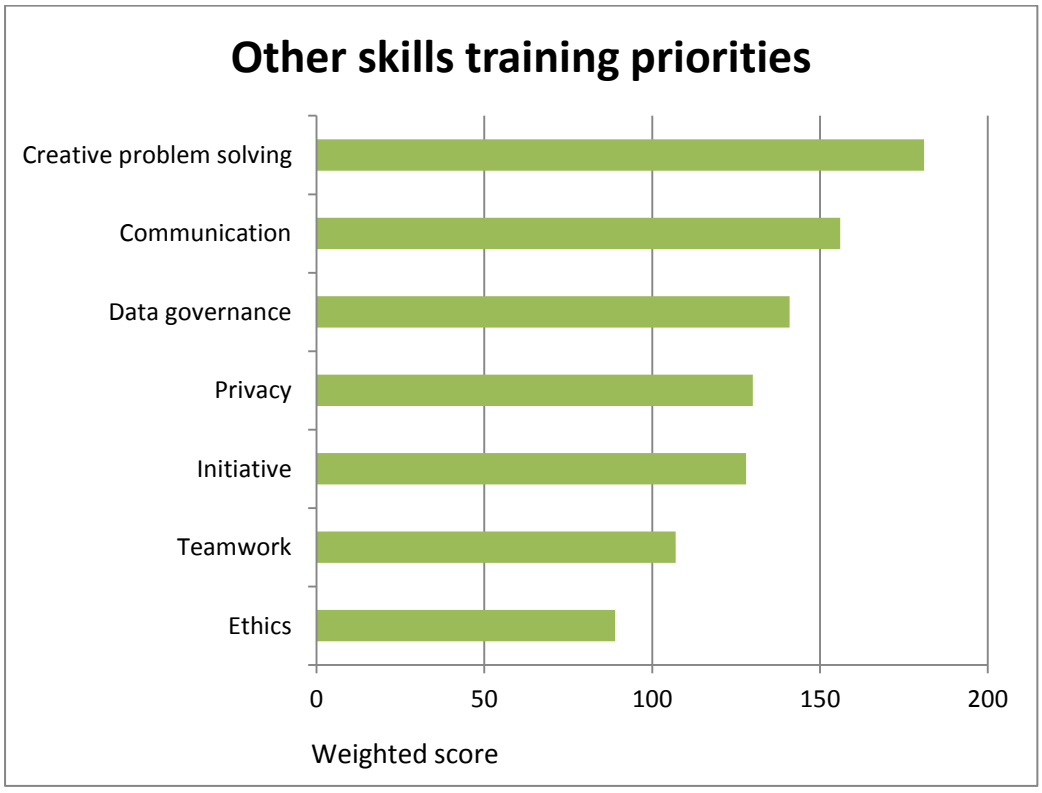
Question 6: Please indicate top 5 priorities for training for your statistical organisation across all areas: IT, Statistics and other (by marking them 1-5, where 1 is the highest)

6.1 IT skills

6.2 Statistical skills



6.3 Other skills



Question 7: Other comments/suggestions

2	It is really not easy to answer this questionnaire. Very different things are put at the same level. The answer represent my idea, and in no case those of my organisation as a whole.
3	We consider acute the study of Big Data problems and are interested in acquiring new knowledge in this field
4	I answer as an individual with limited background in big data but experience of large datasets of survey/census data
5	Encourage the use of a monthly or quarterly newsletter giving the latest developments and innovations in the area. This could also give examples and which tools were used in these cases.
6	We would recommend different trainings for different levels or purposes. There should be courses for Big Data for managers to give an overview on methodologies and tools and mainly to raise awareness about organisational and management issues induced by the use of Big Data sources. There should be training on methodology and IT tools for more technically oriented staff to be able to use these methodologies. Our fifth priority for training would be methodology on statistical learning.
7	As a HR coordinator I could not interpret certain questions of the survey, because I do not have enough information on the IT field.
8	In my opinion, people who will work with big data should have some Java skills. Because Hadoop environment is based on Java and so for example all exception messages are in Java.To better understand the hadoop environment, java is the essential skill.
9	Good luck!
10	Big data is not the issue. Administrative data governance and a willingness to accept/adopt machine learning are the issues.
11	It's too early to priorities the technologies and training we will adopt. We are expecting to enter a phase of experimentation\exploration and learning in the next 12 months or so with the aim of having a better understanding of what tools\approaches we should invest in.
12	It is important not to underestimate the methodological component associated with Big Data. Once the IT component is more well understood, the relevance of the data as well as the accuracy and representativity issues need to be explored. -the computational aspect of Big Data (algorithmic approaches, computer optimizations, visualization) needs to be considered, as well as efficient techniques to visualize, pattern match, analyze, manage. There is an Information Architecture-related component as well (structural, semantic, ontological) -There are other software tools not mentioned in the questionnaire (Web Crawlers, Big Data platforms such as those from 1010data, Amazon, Cloudera, HP, Hortonworks, Actian, Teradata, SAP, Pivotal, MapR, kognitio, infobright, InfiniDB and IBM.
13	Big Data teams should be made up of people having skills from both IT and Methodology. People with good skills in both areas are very rare, hence expensive!
14	It would be desirable to have a training on work with big data.
15	Some of the areas on which questions were asked are just emerging (Statistical methodology for Big Data). Which makes it very difficult to answer these questions. We are learning by doing, which forms the basis of a training program that will be set up in the near future.
16	Keep us updated about the survey's results!
17	In general, I'm very interested in the subject of Big Data. However, so far I'm not really experienced, there are other colleagues in my organisation who are more the experts. I would appreciate it to be informed on further developments in this project.