Big data in the context of Euskalmet activities

Gaztelumendi S. [1,2], Gomez de Segura J.D. [2], Otxoa de Alda K. [1,2], Orue J. [1,2], Torres J. [1,3], Aranda J.A. [1,3].

1 - Basque Meteorology Agency (EUSKALMET). Parque tecnológico de Álava. Avda. Albert Einstein 44 Ed. 6 Of. 303, 01510 Miñano, Álava, Spain
2 - TECNALIA, Meteo Area. Parque tecnológico de Álava. Avda. Albert Einstein 28, 01510 Miñano, Álava, Spain.
3 - Basque Government, Security Department , Directorate of Emergencies and Meteorology. C/ Portal de Foronda 41, 01010 Vitoria-Gasteiz, Álava, Spain.

Introduction

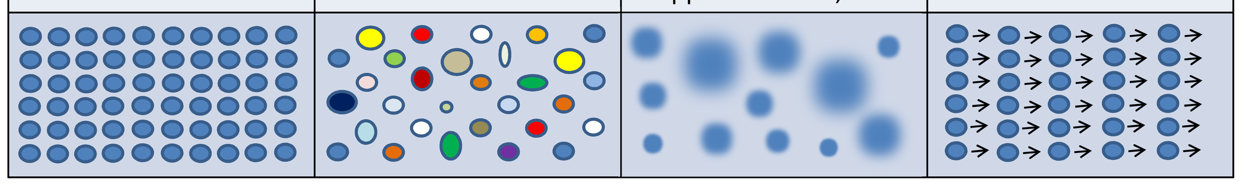
Traditional operational meteorology rely on big amount of data, from different simulation models, observation sources and different man-based products. This classical scenario involves large amounts of more or less structured data. Nowadays, connected sensors are becoming ubiquitous and social networks offer valuable real-time geolocalised information. Big data may offer a greater insight and result in better and new products for end-users (including forecasts improvements), although some challenges have to be faced before the technology can be used in current procedures.

There are many definitions of Big Data (Manyika et al 2011, Gartner 2011,2014, NESSI 2014, Kitchin 2014, IBM 2014). Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. Users of these technologies understand that big data is best described by today's greater volume of data, the new types of data and analysis, or the emerging requirements for more real-time information analysis (Schroeck et al 2012). Here we understand that "big data" cover the technologies and techniques that allow to extract value from a set of data that is too complex to be processed by traditional means in an acceptable way. The complexity comes from the multiple dimensions of data (Ishwarappa 2015, NESSI 2014, Manyika et al 2011 Gartner 2014, IBM 2014), and particularly from Volume, Variety, Veracity or Velocity, the 'so call four "Vs" of big data (see Fig 1.).

The final goal in the case of Basque Meteorology Agency (Euskalmet) is to be able to extract trends and outliers by analysis of a combination of new and traditional data sources using "big data". In this work we present preliminary aspects dealing with big data in the context of Euskalmet. First we analyze data used in Euskalmet and its possible future evolution. Secondly we briefly introduce how data are managed today in Euskalmet and some techniques and technologies already available in big data ecosystems. Finally some conclusions from Euskalmet context and possible consideration for future are mentioned.

Fig.1. Four dimensions of big data

VOLUME	VARIETY	VERACITY	VELOCITY
Data at scale	Data in many forms	Data in doubt	Data in motion
Terabytes to Exabytes data	Structured, unstructured, text/numbers multimedia	Uncertainty due to lack of predictability, in consistency, incompleteness, ambiguities, approximations, etc.	Streaming data



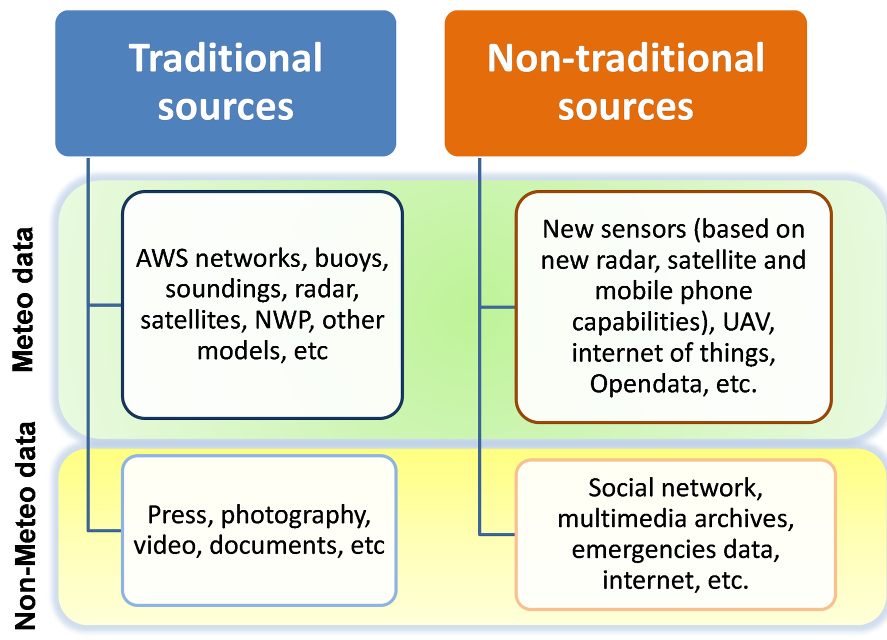


Fig. 2.. Data classification based on traditional/non-traditional Meteo/non-meteo attributes

Data characteristics

In a Meteorological Service a large amount of information is generated, collected and accumulated in operational, research and management activities (Hoffman et al 2011, Overpeck et al 2011). These data can be categorized as Big data, not only to the extent that large volumes of data are available (terabytes to exabytes) but also because their variety (presented in different ways) and veracity (often present uncertainties, inconsistencies or are ambiguous).

In the case of a Regional Weather Service, the volume of data, depending on the type of facility and services provided (mainly with origin in instrumentation and Local Area Models), will be several orders of magnitude less that in the case of bigger Meteorological Centers with larger coverage (even global). In any case, advanced services and proximity to the user introduce variety and veracity factors as the essence of "big data" at local and regional scale. Data handled in the case of Euskalmet are of similar nature to those handled in any weather center at regional level. Data can be classified in different ways according to structuration, type, format, volume, access, source, latency, generation, etc. In Fig.2. we classified data used in operational meteorology in two main data groups; traditional (already existing in the twentieth century) and nontraditional (XXI century), distinguishing between purely meteorological and non-meteorological data. If we consider data structuration; meteo sensors data are structured and usually stored in relational database; NWP, Satellite and Radar are highly structured and suitable for computer processing but stored in native format; Video/images are of well-structured nature but difficult to process; Social networks data and others natural language products are unstructured and difficult to process. In Fig. 3 we group the data considering its structuration and if is generated automatically with machines or by human hand. One of the challenges that arises is how to extract value from non-meteo data not used regularly and particularly how to extract value in operational meteorology from data loosely structured and often ungoverned.

Looking at current day, in Tab. 1 we present some general characteristics of main data handled by Euskalmet. Note on one hand, the huge amount of modeling data and the importance of remote sensing data, on the other hand, the low volume involved in today human based products (includes social network). The volume of structured data today archived in SQL Data Bases (DB) are no more than 1% of the total volume and comes mainly from Euskalmet record responsibilities of Basque weather (Fig 3) .

Looking to the future, several cross-cutting trends have fueled growth in data generation and will continue to propel the rapidly expanding pools of data. These trends include growth in traditional transactional databases, continued expansion of multimedia content, increasing popularity of social media, and proliferation of applications of sensors in the Internet of Things (Manyika et al 2011, NESSI 2014). In the meteo-climatic context, we must be prepared for huge data generation capabilities from new NWP (multi-purpose, multi-scale, hyper-resolution, super-ensembles...) and new remote sensing (mainly new satellite and radar capabilities), the increasing use of multimedia (particularly videos from surveillance), the rapid adoption of smartphones driving up the usage of social networking (unstructured data, natural language, etc.) and new sensor and devices with meteo capabilities embedded in the physical world and connected by networks to computing resources (vehicles, mobile phone, UAV, etc.).

Projections for the future, always a challenging field, are difficult but necessary to establish potential scenarios for planning. In Fig. 4 and 5 we can see possible Euskalmet evolution for near and long future. Note the exponential data volume tendency (see Fig. 4) and how at any time data volume is driven by NWP necessities but data from new instrumentation and social networks became also considerable in future (Fig. 5). Although large data volumes are expected, we handled thousand times less information than large meteorological centers like MetOffice (Nelson 2013).

DATA SOURCE	LATENCY	MONTHLY VOLUM (Gb)	FORMAT	STORAGE
Global Models	6 hours	150	,grib2. ,netcdf,.grd	dir
LAM Models	6 hours	180	,grib2. ,netcdf,.grd	dir
Other numerical models	6 hours	30	,grib2. ,netcdf,.grd	dir
MOS and stattiscal guidance products	6 hours	3	.txt, netcdf	dir/DB
AWS Mesonetwork	10 min	0,72	,dat	dir/DB
Buoys	1 hour	0,001	.txt	dir/DB
Synop and others	1 hour	0,1	.txt	dir
Disdrometers	1 min	0,013	.dat	dir
Kapildui Radar	10 min	9,5	.ele .vol	dir
Punta Galea wind profiler	30 min	0,03	.cns	dir
Coastal Radar	1 hour	4	.ruv .tuv .wls .cs4	dir
Lightimg System	under 1 min	10,6	.raw,.txt,.bin,.dat	dir
Meteosat	15 min	99	.XPIF	dir
Radiosoundings	24 hours	0,012	.txt	dir
Operational human products	at least 12 hours	2	,pdf, html, ..	dir/DB
Twitter	not fixed	5	.txt	dir/DB
Web and intranet	at least 12 hours	1	html, xml, ..	dir/DB
Press	daily	5	.pdf	Dir
Other data	variable	0,5		dir/DB

Tab.1. Data characteristics in Euskalmet case

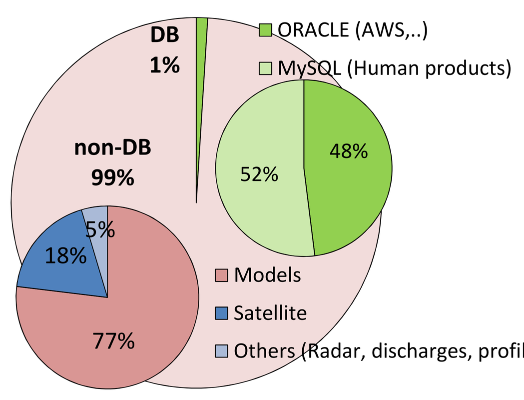


Fig. 3.. Archived data distribution in Euskalmet at present day.

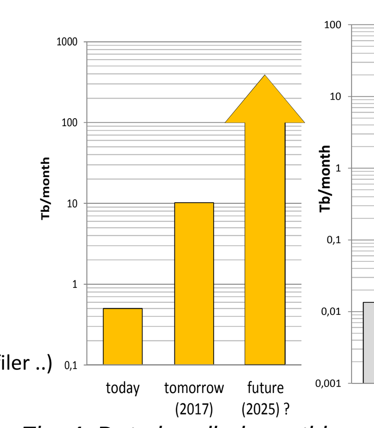


Fig. 4. Data handled monthly by Euskalmet for all sources

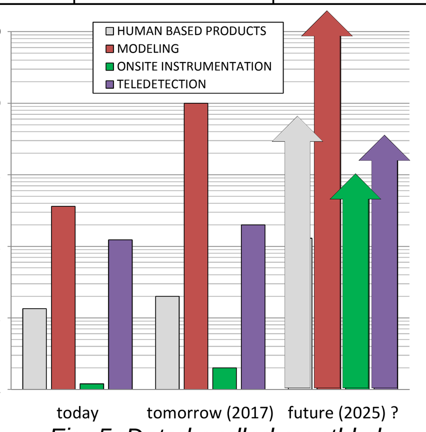


Fig. 5. Data handled monthly by Euskalmet depending on source

Techniques and technologies

Once we have analyzed some characteristics of common meteo-climatic data, don't forget that we need to put intelligence on raw data in order to convert it in information to obtain knowledge and extract value. For this purpose support from adequate technologies and efficient techniques are essential (Lavalle et al 2011). Specific hard and soft are needed for archive, manage and analyze data (Singh et al 2014). All Meteorological Services has today, in different degree, this capabilities for its actual data requirements including HPC (in some degree) and specific mathematical and analytical models.

Currently, in the Euskalmet case, we manage data stored on different servers depending mainly on their type. Structured data from different sensors and operational activities are stored in tables within relational databases (see Fig 4), unstructured data such as emails or texts, are stored in their native formats. Data generated by machines (NWP, Remote sensing) without human intervention (computers, processes, applications), and audio/video/images are stored in different directories within specific computers disks that facilitate its further dissemination. Exploitation and data analytic is done based in multiples ad-hoc solutions carry out for different purposes (forecast, surveillance, monitoring, validation, integration, management, dissemination, etc). Those solutions are based in different techniques (numerical modeling, statistical techniques, visualization, etc.) and are implemented using a variety of languages and tools (Fortran, C++, IDL, R, Python, Matlab, etc.).

In the future, the increasing volume and complexity in data managed by Euskalmet (see Fig 4), will require the implementation of new techniques and technologies (tools) suitable for new data incoming (i.e. be ready for Big data).

Today there are many big-data tools (hard and soft) in the market . Dealing with hard, among others, high-performance computers and some cloud capabilities are required (more in Singh et al 2014). Focusing on soft, MPP databases, search-based applications, distributed file systems, distributed databases, cloud-based, parallel data analysis platforms, data programming languages and more specialized tools for specific data domains are needed (e.g. Singh et al 2014). Commercial solutions from different companies (Oracle Corporation, IBM, Microsoft, Software AG, SAP, EMC, HP, Dell, etc) are available , and also some open-source (see Fig. 7). Data without insight is an unrealized resource. Big Data analytics is the application of analytic capabilities (descriptive, diagnostic, predictive and prescriptive) on complex datasets. Big Data analytics can be considered as a extension of data, text, network and mobile analytics (Chen et al 2012). Includes different technologies like A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualisation (Mkinsey 2011). At the end grounded mostly in data mining and statistical analysis (well known tools in meteo-climatic business).

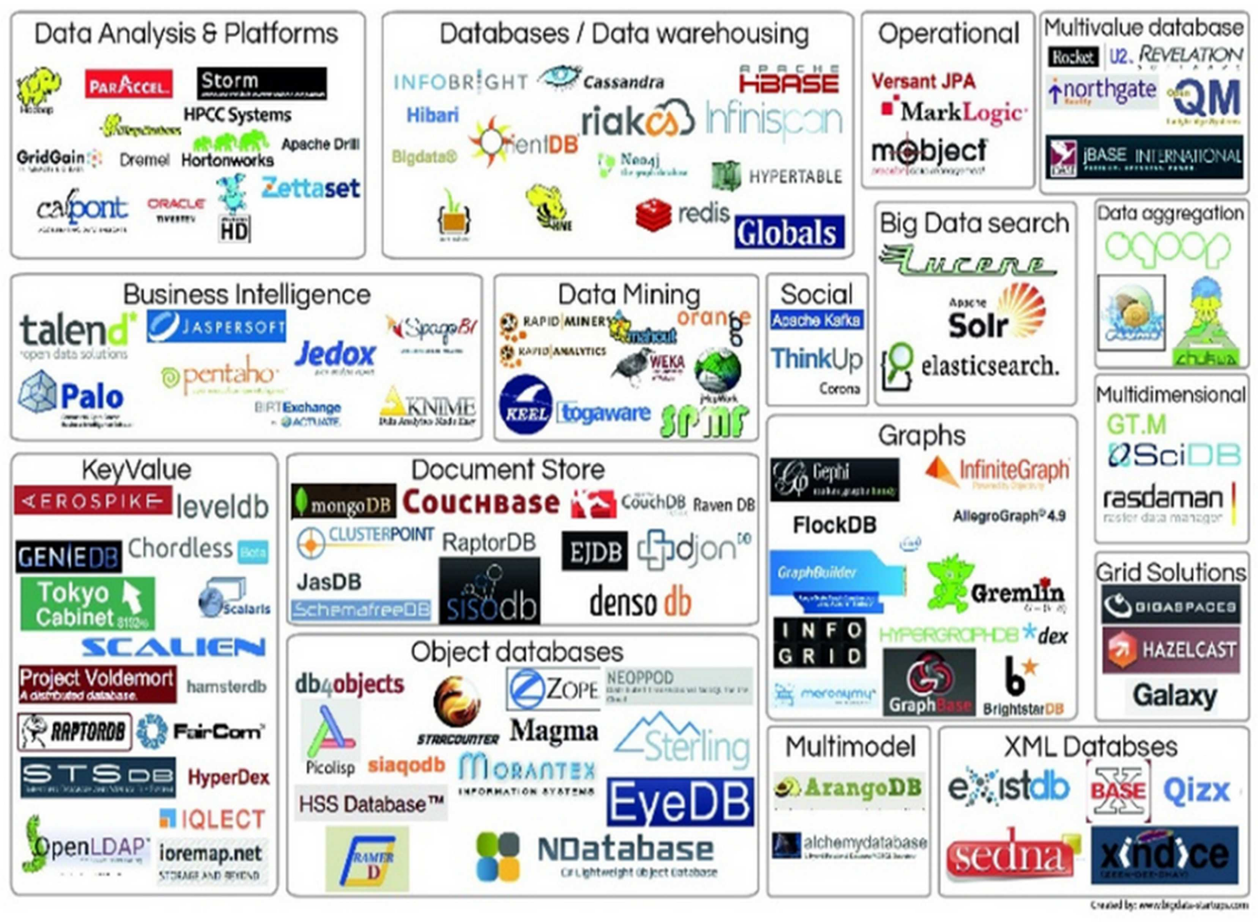


Fig. 6. Most popular open source tools grouped depending on main purpose

Conclusions

Big data, not new in weather services. The promise of data driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of Big Data (Labrinidis 2012). But in operational meteorology data driven-decision is our daily job. As a meteo service we have been using big-data concepts during last years but we didn't know. Our problems start long time ago, during data acquisition, when huge potential data volumes requires us to make decisions, usually in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Challenges continued with the need to extract valuable and actionable information with in deep analysis from different data sets considering interrelationships. Finally it remain clear that user orientation should guide all our actions and that results presentation and its interpretation by non-technical experts is crucial to extracting actionable knowledge.

Integration and big data analytics is the key. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Traditional meteo sensors data are structured an usually managed throw SQL Data Bases. NWP, Satellite or Radar are highly structured data but in native formats that usually requires specific tools in order to extract value. Other data as images and video are structured for storage and display, but not for semantic content and search. In other cases data are semi-structured or unstructured (text, mails, social networks, webs) and difficult to analyze. Big data analysis is the key question in many meteorological applications, due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed.

Criticism is still necessary. The adoption of big data in near future seems to be inevitable in many business but some critical questions are open (e.g. Boyd D. 2012).. The question is, if there is actually any new and real substance behind all these big data discussions for operational weather service is important to detect real improvements and new opportunities that potentially bring big data, but without obsession. In any case, in Euskalmet case , new sensors, video surveillance data, and social networks seems to be the data drivers for new opportunities. But keep in mind that, bigger data are not always better data, specially when dealing with social network data, rubbish is always rubbish. When dealing with social big data, don't forget that legal and privacy aspects must be considered (e.g. NESSI 2014) and that people' and 'Social media users' are not synonymous.

Pragmatic approach. Euskalmet strategy is simple and pragmatic first identify requirements (operational and user driven) and then to incrementally upgrade our infrastructures, data sources and analytics capabilities over time. At same time, as a internal proof of concept, starting with existing data and infrastructures, we are going to implement a preliminary "low cost big data ecosystem" based on open source tools (Hadoop cluster, HDFS system, Yarn Mapreduce, Apache Mahout/Spark, Mongo DB, Elastic Logstash/Search/Kibana and R and Python). The final aim is to build new data structures and analytical capabilities based on integration of non-structured text data (from bulletin, texts, mails, press twiter, etc) and other available data in order to achieve relative near-term results improving user-oriented local services (including operational forecast).

Aknowledgements

The authors would like to thank the Emergencies and Meteorology Directorate - Security Department - Basque Government for public provision of data and financial support. We also would like to thank all our colleagues from EUSKALMET for their daily effort in promoting valuable services for the Basque community, We would also like to thank the Free Software Community and all institutions that maintain and support availability of free of charge data, models and tools for the Scientific Community.

euskalmet
agencia vasca de meteorología
euskal meteorologia agentzia

EUSKO JAURLARITZA GOBIERNO VASCO
SEGURTASUN SAILA
Larrialdiei Aurre Egiteko eta Meteorologiako Zuzendaritza
DEPARTAMENTO DE SEGURIDAD
Dirección de Atención de Emergencias y Meteorología

References

- Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". Information, Communication & Society 15 (5): 662
- Chen H, Chiang, R.H.L. Storey V.C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36 (4) (2012), pp. 1165–1188
- Gartner 2014. Answering Big Data's 10 Biggest Vision and Strategy Questions. Available athttps://www.gartner.com/doc/2822220?ref=SiteSearch&refval=&pcp=mpe
- Gartner. 2011. Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916
- Hoffman F.M. , Larson JW Mills RT ,Bjørn-Gustaf Auroop, Ganguly, Hargrove, Huang, Kumar,Vatsavai 2011 Data Mining in Earth System Science (DMESS 2011) Procedia Computer Science. Elsevier
- IBM. 2015. "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2015-10-06.
- Ishwarappa, J 2015 A Brief Introduction on Big Data 5Vs Characteristics and Hadoop TechnologyOriginal Research Article Pages 319- Anuradha Procedia Computer Science Volume 48, Pages 1-808
- Kitchin R. 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE publications.
- Labrinidis A. And Jagadish H.V. 2012. Challenges and Opportunities with Big Data. Journal Proceedings of the VLDB Endowment Volume 5 Issue 12, August 2012 Pages 2032-2033
- LaValle S. Lesser, E. Shockley, R. Hopkins M.S. and Kruschwitz N. 2011. Big Data, Analytics and the Path From Insights to Value, MITSloam Management review Vol 52 No2.
- Madden S. 2012. From Databases to Big Data IEEE Inter-net Computing, 16 (2012 June), pp. 4–6
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R. Roxburgh, C. , & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity (pp. 1–143). The McKinsey Global Institute.
- Nelson, P. 2013. Ready for the storm. Data management and backup at the Met Office. UCISA Event Big data, big backup? 11jun 2013, Birminghan.
- NESSI 2012. Big Data A New World of Opportunities. NESSI – Big Data White Paper.
- Overpeck J. T., Meehl G. A., Bony S., and Easterling D. R. 2011 "Climate Data Challenges in the 21st Century," Science, vol. 331, no. 6018, pp. 700 –702, Feb. 2011.
- Schroeck M, Shockley R, Smart J, Romero-Morales D , Tufano P 2012. Analytics: The real-world use of big data Business Analytics and Optimization. IBM Institute for Business Value Report.
- Singh D. and Reddy C.K. "A survey on platforms for big data analytics", Journal of Big Data, Vol.2, No.8, pp.1-20, October 2014.
- Singh D. and Reddy C.K. 2014 "A survey on platforms for big data analytics", Journal of Big Data, Vol.2, No.8, pp.1-20, October 2014.

Contact info:

santiago.gaztelumendi@tecnalia.com

TECNALIA

Parque Tecnológico de Bizkaia
C/ Geldo Edificio 700
E-48160 DERIO (Bizkaia) Spain
www.tecnalia.com

15th EMS Annual Meeting
12th European Conference on Applications of Meteorology (ECAM)
07 – 11 September 2015, Sofia, Bulgaria