Pure knowledge with the best Big Data experts

bigdata
TECHNOLOGY WARSAW SUMMIT

getindata

EVENTION
CZAS ZAANGAŻOWANY

Dear Sirs,

Big Data Technology Warsaw Summit 2017 was the biggest Big Data Technology event in Central Europe so far. It was an exciting one-day conference with purely technical content in the fields of Big Data analysis, scalability, storage and search. All presentations were given by true practitioners who work at top data-driven companies like Facebook, Uber, Spotify, Google, Criteo and SkyScanner. They shared their knowledge, recommendations, tools, models, successes and failures.

We proudly hosted over 400 participants from many European countries and companies that leverage Big Data in production use-cases. The highly technical and vendor-neutral content could be possible due to the absence of a single big vendor as the main sponsor or organizer of the conference who sets the agenda. Four technical tracks were covered the most important and up-to-date aspects of Big Data, including deep learning, real-time stream processing and the cloud. Over 20 round-table discussions allowed to exchange opinions and experiences on specific topics. Delegates could also take part in three technical and practical workshops a day before the conference.

Big Data Technology Warsaw Summit 2017 was the third such event organized by Evention and GetInData. All of them were a great success. I invite you the next edition that will take place on 22nd of February 2018, Warsaw, Poland – www.bigdatatech.pl!

In the meantime please take your time for the lecture of this report based on most interesting content from Big Data Technology Warsaw Summit 2017.

Przemysław Gamdzyk
Meeting Designer and CEO, Evention

EVENTION
CZAS ZAANGAŻOWANY

Evention

Evention is a company that specializes in increasing the value of ICT business meetings. We strongly believe that business events are integral and irreplaceable factor when it comes to the creation and maintanance of relations and improvement of communication between companies and people that create them. We are constantly searching for innovative business meeting formulas to adress current needs, expectations and aspirations of ICT managers.

getindata

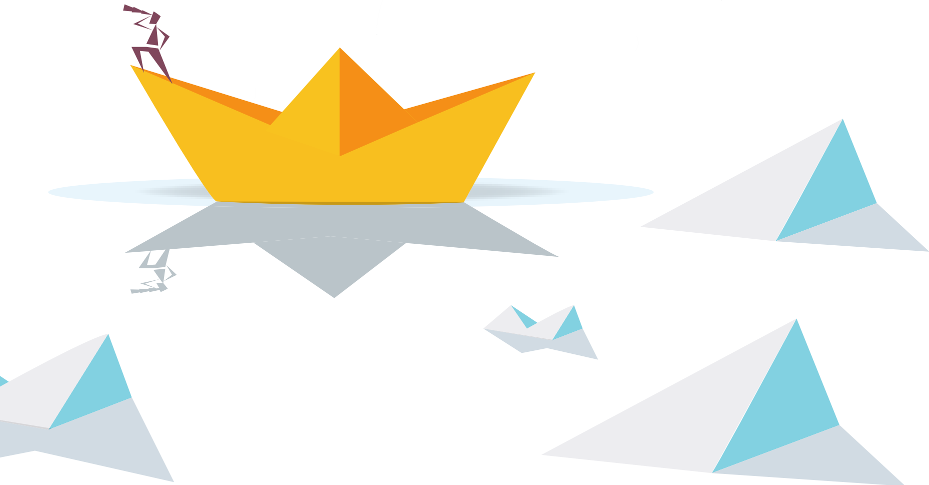
GetInData

GetInData was founded by former Spotify data engineers in 2014. We are experienced and passionate Big Data experts with proven track of records. Our mission is to help data-oriented organizations to succeed using open-source Big Data technologies such as Hadoop, Flink, Spark, Kafka by providing outsourcing, consulting and training services. We've been already contracted by tens of companies ranging from fast-growing European startups to global corporations in pharmacy, FMCG, banking and media sectors.

Content

The Biggest Big Data Tech in Central Europe

Big Data Technology Warsaw Summit 2017 was the third edition of the first truly international conference in Poland. The event was a huge success. Over 400 specialists - data scientists, data engineers, IT managers, application developers, and system administrators - registered to participate in the main event, and over 100 took part in the technical workshops earlier the same week. For the participants, it was a unique chance to get to know how data analysis is handled in the world's leading tech companies, like Google, Facebook, Cisco, SkySkanner, Fandom or Avito.



- There are not many tech big data events in Poland, and in Europe either. We wanted to create a conference similar the events in Berlin, Sevilla, and London - said Przemysław Gamdzyk, CEO & Meeting Designer at Evention, one of the conference organizers.

About 60 speakers invited by an international conference committee shared their practical knowledge during four days of workshops, thematic sessions, and roundtable debates. - Our most important goal for creating this event is knowledge sharing - said Adam Kawa CEO and Co-founder, GetInData.

There are not many tech big data events in Poland, and in Europe either. We wanted to create a conference similar the events in Berlin, Sevilla, and London.

PRZEMYSŁAW GAMDZYK,
EVENTION

Getting inside knowledge from world's leading companies

The conference was a unique chance to learn directly from the experiences of data scientists, system administrators, developers and data engineers from world's biggest companies: Google, Facebook, Cisco, H20.ai, SAS Institute, SkySkanner, Fandom, Avito, AB Initio, Criteo, PubMatic, Findwise, GetInData, and Allegro.

Dedicated tracks according to different needs

During the main conference day, the participants could choose from four dedicated tracks.

Operations & Deployment was dedicated to system administrators and people with DevOps skills who are interested in technologies and best practices for planning, installing, managing and securing their Big Data infrastructure in enterprise environments – both on-premise and the cloud.

Data Application Development was a place for developers to learn about tools, techniques and innovative solutions to collect and process large volumes of data. It covers topics like data ingestion, ETL, process scheduling, metadata and schema management, distributed data stores and more.

Analytics & Data Science included real case-studies demonstrating how Big Data is used to address a wide range of business problems. You can find here talks about large-scale Machine Learning, A/B tests, visualizing data as well as various analysis that enable making data-driven decisions and feed personalized features of data-driven products.

> *Our most important goal for creating this event is knowledge sharing*
>
> ADAM KAWA,
> GETINDATA

Finally, **Real-Time Processing** covered technologies, strategies, and use cases for real-time data ingestion and deriving real-time actionable insights from the flow of events coming from sensors, devices, users, and front-end systems. Making profit on big data

Learning the business directly from the best industry insiders

The participants could learn the business aspect of big data and machine learning from industry insiders. Sascha Schubert, Advisory Business Solutions Manager at SAS Institute, analytics software developer company from North Carolina, shared the detailed case studies on how machine learning can be used in different industries.

Paweł Goduła, senior data scientist at BCG Gamma, an advanced analytics and big data branch of Boston Consulting Group, talked about his cooperation with colleagues consultants, and how analytics can be used to increase the profits of his clients.

System administrators could get some inside information on big data architecture from Nikolay Golov, chief data warehousing architect at Avito, world's second largest classified advertisements website with 35 million unique monthly visitors.

Getting to know the newest technology

Krzysztof Baczyński, Cisco Big Data Lead for Poland, and Kamil Ciuszko, CEO of Alterdata, explained how to create an effective, scalable and easily manageable environment for big data processing. They presented a case of a real-time big data analytics related to location tracking using automated and scalable Cisco platform.

Participants could learn how to effectively use C-store DBMS analytics platform as well as Cisco Validated Design for Big Data architecture which combines tools as Cisco UCS (Unified Computing System), Cisco ACI (Application Centric Infrastructure) and UCS Director for Big Data which provides a single-touch solution that automates Hadoop deployment and provides a single management pane across both physical infrastructure and Hadoop software.

Learning from the experience of people running huge projects

Stuart Pook, Senior DevOps Engineer at Criteo, NASDAQ traded French personalized retargeting company talked about his company's experience with a Hadoop cluster with 39 PB raw stockage, 13404 CPUs, 105 TB RAM, 40 TB data imported per day and over 100000 jobs per day. This cluster was critical in both stockage and compute but without backups. After many efforts to increase their redundancy, Criteo moved to two clusters that combined have more than 2000 nodes, 130 PB, two different versions of Hadoop and 200000 jobs per day but these clusters do not yet provide a redundant solution to our all storage and compute needs.

In his presentation, he disclosed the choices and issues Criteo solved in creating a 1200 node cluster with new hardware in a new data center. Some of the challenges involved in running two different clusters in parallel will be presented.

Software developers had a chance to listen to Robin Tweedie and Arthur Vivian, software engineers in Sky-Scanner, world's leading travel metasearch engines with monthly 60 million users. They talked about approach Skyscanner took to enable every decision in the company to be data based. They shared the lessons learned when using technologies like: Kafka, Logstash, Elasticsearch, Secor, AWS Lambda with Amazon S3, Samza, Protocol Buffers, and others.

A look inside data analytics processes

Michael Hausenblas, Developer Advocate, at Mesosphere, a San Francisco-based company building operating systems for data centers based on Apache Mesos discussed options to operate elastic data pipelines with modern, cloud-native platforms such as DC/OS with Apache Mesos, Kubernetes, and Docker Swarm.

Mark Pybus, Head of Data Engineering at Sky Bet, one of the largest UK online bookmakers, explained how their Hadoop platform addresses two common problems in the gambling industry – knowing your current liability position and helping potential irresponsible gamblers before they identify themselves.

Participants could learn Sky Bet's experiences replacing a traditional data warehouse with Hadoop. How

Big Data Technology Warsaw Summit 2017 in numbers

Over **400**
specialists

Over **50**
companies taking part

Over **27**
presentations and lectures

over **100**
technical workshops

the architecture met the needs of sportsbook traders to be able to manage liabilities in a competitive and high-frequency environment and how that led to decommissioning the legacy data warehouse.

Polish companies were represented by Allegro Group among others. Piotr Guzik, Software Engineer at Allegro, explained how his company managed to detect anomalies, such as heavy web traffic after the successful commercial event, using their own simple model. He disclosed why Allegro moved from R language to a working solution in Scala.

Mariusz Strzelecki, Senior Data Engineer at Allegro shared his experience with Hadoop applications development using tools like MapReduce, Scala and a lot of APIs for submitting, scheduling and monitoring jobs. And of course is a Kerberos expert.

Scalable big data solutions in science

Big Data Technology Warsaw Summit 2017 had also some interesting presentations about the use of data engineering in science. Marek Wiewiórka, Solution Architect at GetInData talked about genomic population studies, and how they incorporate storing, analyzing and interpretation of various kinds of genomic variants as its central issue. When thousands of patients sequenced exomes and genomes are being sequenced, there is a growing need for efficient database storage systems, querying engines and powerful tools for statistical analyses. Scalable big data solutions such as Apache Impala, Apache Kudu, Apache Phoenix or Apache Kylin can address many of the challenges in large-scale genomic analyses.

The presentation covered some of the lessons learned from the project aiming at creating a data warehousing solution for storing and analyzing genomic variants information at Department of Medical Genetics Warsaw Medical University.

Ashish Tadose, Senior Data Architect at PubMatic, online advertising software company from India, explained how AdTech companies need to address data increase at breakneck speed along with customer demands of insights & analytical reports. PubMatic receive billions of events and several TBs of data per day from various geographic regions. This high volume data needs to be processed in real-time to derive actionable insights such as campaign decisions, audience targeting and also provide a feedback loop to AdServer for making efficient ad serving decisions. Ashish Tadose shared how PubMatic designed and implemented these scalable low latency real-time data processing solutions for our use cases using Apache Apex.

These are just a small part of the speakers on the Big Data Technology Warsaw Summit 2017. The participants could talk to them during roundtable sessions that took place at 11 tables in two rounds.

In this raport, we present the most interesting presentations that took place at Big Data Technology Warsaw Summit 2017.

Firat Tekiner
AB Initio

Paweł Goduła,
BCG Gamma

Nikolay Golov,
Avito

Nelson Arape,
Spotify

Arthur Vivian,
SkyScanner

Stuart Pook,
Criteo

Mark Pybus,
Sky Bet

Mariusz Strzelecki,
Allegro

Olaf Piotrowski,
Allegro

Ashish Tadose,
PubMatic

Discussion panel

The big data Hadoop revolution in big companies

Moderator:
Krzysztof Zarzycki, Big Data Architect and Co-founder, GetInData

Participants:
Grzegorz Bartler, Head of Business Intelligence Department, Polkomtel, Cyfrowy Polsat
Dr hab. Piotr Gawrysiak, Chief Data Scientist, mBank S.A.
Derek Yeung, Head of Platform Engineering, Nordea
Olaf Piotrowski, Chief Data Officer, Allegro

KZ: We would like to discuss the problem of big data. Could you tell us about the big data projects you run in your organizations?

GB: I'm not a fan of the term big data. It's just technology, and the term is misleading. It's similar to Linux. When it was started, just a few people knew it, now it's common. We can observe the same process with Hadoop framework. We started using it because of the cost. It was so cheap.

In our organization, we decided to use Hadoop for processing revenue insurance. At the moment we are processing all the calls from our customers. At the beginning, we wanted to be sure if we are not losing any revenue. It was a raw Cost Estimating Relationships (CER) tool. This was one of the first projects in the area. Now we use it in customer oriented projects. We're checking if customers are doing a particular event.

PG: I am a data scientist. We rely on the software and architecture, this is how it works. The problem is, most of the stuff we do is not visible to the customer. Our main motivation was not to reduce costs but to get some data lineage, and bring order to the data warehouse. This gives us a lot of tools to play with the data. Credit risk analysis is introduced. Payment assistant is visible to everyone using our mobile app.

DY: I've been doing big data for past 6-7 years. Now we are utilizing it to help us in many aspects: marketing, big data capabilities. Our mission is to make big data as easy as possible for the bank.

Using big data for financial services is much harder than in startup. We have to adapt to many restrictions: compliance, security measures. We have to do all probation, ground works in perfect order. Right now we have almost done the ground work, now we are realizing the benefits. We utilize predictive and analytics measures.

OP: We started using Hadoop in 2012 as an internal IT project. What can happen if we tried it? The first business case came from Google. We started our internal tracking system, it's difficult to separate big data project and product development. For Allegro microservices Hadoop is the place where we practice and start most of our analytics. We measured the quality of the images on Allegro platform, and now we use deep learning methods to score the images.

KZ: Did it succeed, when measuring the return on investment?

I am a data scientist. We rely on the software and architecture, this is how it works. The problem is, most of the stuff we do is not visible to the customer. Our main motivation was not to reduce costs but to get some data lineage, and bring order to the data warehouse. This gives us a lot of tools to play with the data.

DR HAB. PIOTR GAWRYSIAK,
MBANK S.A.

> *We started using Hadoop to reduce costs, and we stopped using it after two years. The tricky part about it is that software and hardware are really cheap, but people being able to work with the technology are scarce and expensive. We simply couldn't get people in Warsaw that could do the job. Big Data is not really changing heavily the way we are working.*
>
> **GRZEGORZ BARTLER,**
> POLKOMTEL, CYFROWY POLSAT

GB: We started using Hadoop to reduce costs, and we stopped using it after two years. The tricky part about it is that software and hardware are really cheap, but people being able to work with the technology are scarce and expensive. We simply couldn't get people in Warsaw that could do the job. Big Data is not really changing heavily the way we are working.

There are not many people in the market. To make my point clear: the ROI of the project was positive, but all the costs of hiring specialists are mounting. We have a data lake, enterprise data framework, and Hadoop. If our analytical people want, they can enable the analysis for them.

PG: In our case, if I look at my projects, I cannot think about many failures. In the banking industry software is expensive. In organizations this big you need time to see the real ROI. When someone tells me that a 2-year-old piece of software is a legacy, I strongly disagree. The real fruit will be visible in the much longer term. I'd measure it in a long time perspective, like 10 years.

For us, it is not only about implementing the cost-cutting technology. I consider it an enabling technology, a driver of change. Some of our services were impossible without Hadoop architecture. It would be impossible technically.

Even if we are not dealing with the truly big data, we are not dealing with the kind of data measured in hundreds of petabytes, it's still quite a lot.

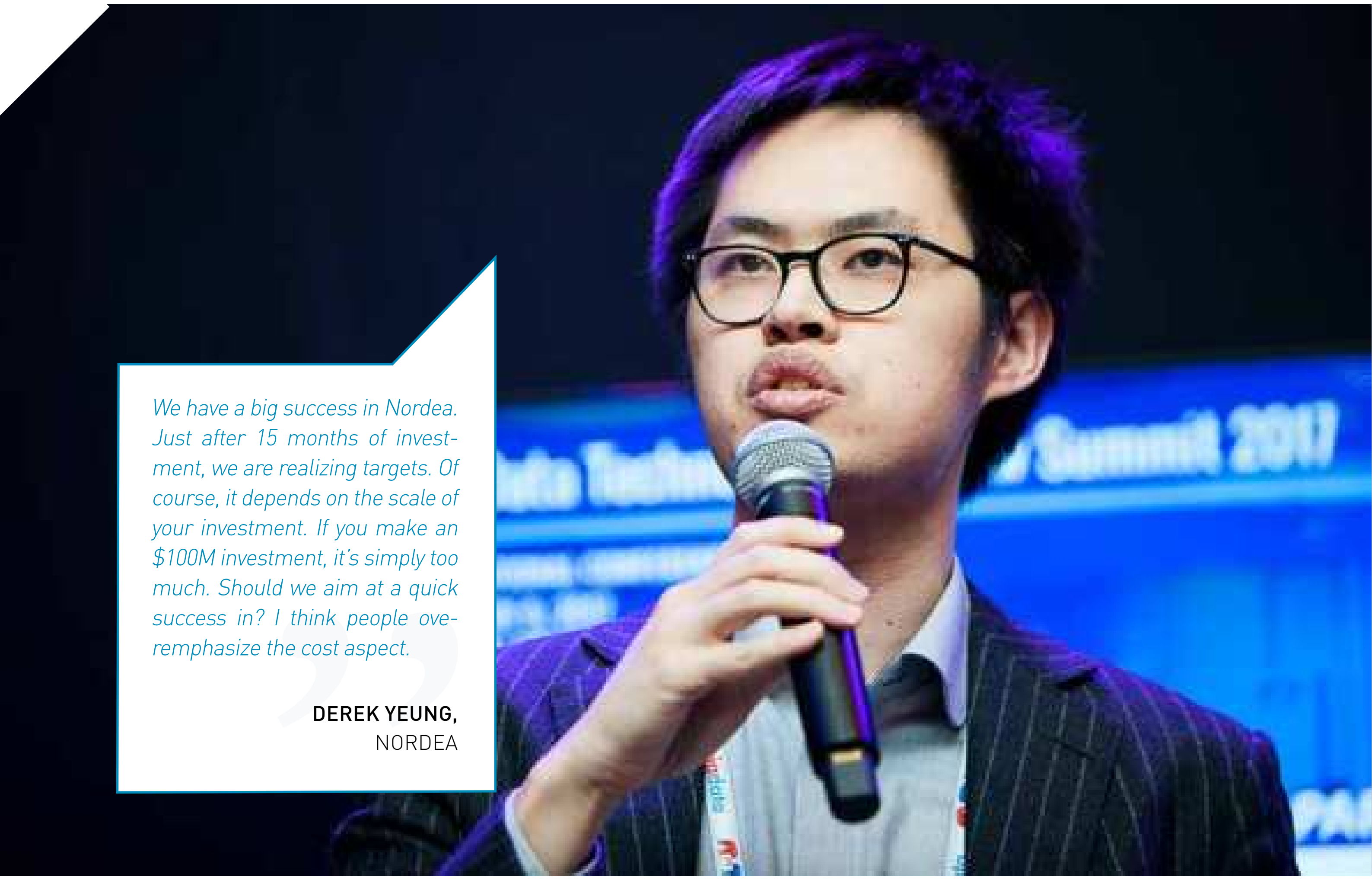
Some companies decide to run big data projects, a competitor is doing it. I'd rather consider a different motivation. We should do it because the client requires it.

DY: We have a big success in Nordea. Just after 15 months of investment, we are realizing targets. Of course, it depends on the scale of your investment. If you make an $100M investment, it's simply too much. Should we aim at a quick success in? I think people overemphasize the cost aspect.

Every data project is a long-term investment, but the technology should be as up to date as possible. Let's take data

Hadoop has so many advantages, but the most important one is that, provided you have the right people, you have an opportunity to move. You stay flexible.

OLAF PIOTROWSKI,
ALLEGRO



We have a big success in Nordea. Just after 15 months of investment, we are realizing targets. Of course, it depends on the scale of your investment. If you make an $100M investment, it's simply too much. Should we aim at a quick success in? I think people ove-remphasize the cost aspect.

DEREK YEUNG,
NORDEA

warehousing as an example. If you are using an architecture that is 10 years old, more complex, hard to change. With Hadoop, we gain agility. We can start thinking about a change in terms of weeks, instead of quarters, we used to. This agility itself is a big win.

OP: For us, it's an enabler. It was impossible to run many services when we didn't track data, and try to find patterns in user behavior. Now it's possible. Still, it's a long learning process. Doing these projects over the years we build know-how and culture. In Allegro we have fantastic data engineers, who are multifunctional and are able to find solutions that cannot be delivered by any commercial offer. In the long term, this is our win.

PG: I agree. It's an enabler, it's not just about money and costs anymore.

GB: I don't agree. We have to consider business indicators, and costs are a priority. Every big data expense has to be validated by a business case. We cannot forget about measuring the ROI.

I consider Hadoop as an enabler. I can do some things I couldn't do a few years ago, but mainly thanks to the cost reduction. In the past, I wasn't even checking the possibility of some actions because the infrastructure was so expensive. Now I can react faster, and this way I can do it.

As for the analytical tools, I won't replace SAS with Hadoop to be able to do more things. My company uses hundreds of analytical processes.

PG: In some areas, we were simply not able to run the analysis that is possible now. We can do it thanks to the advancement of software, hardware, and algorithms.

GB: Ten years ago I built real-time analytical scoring models using Oracle. It was possible with the architecture and performance capabilities.

PG: Banking sector is not a good example of the cost reduction possibilities. For one part, we cannot use the cloud as freely as other industries. Small startups can compete with established players taking advantage of each novelty. When we built our own Hadoop platform we chose Cloudera. The reason was it had most sophisticated support and security features.

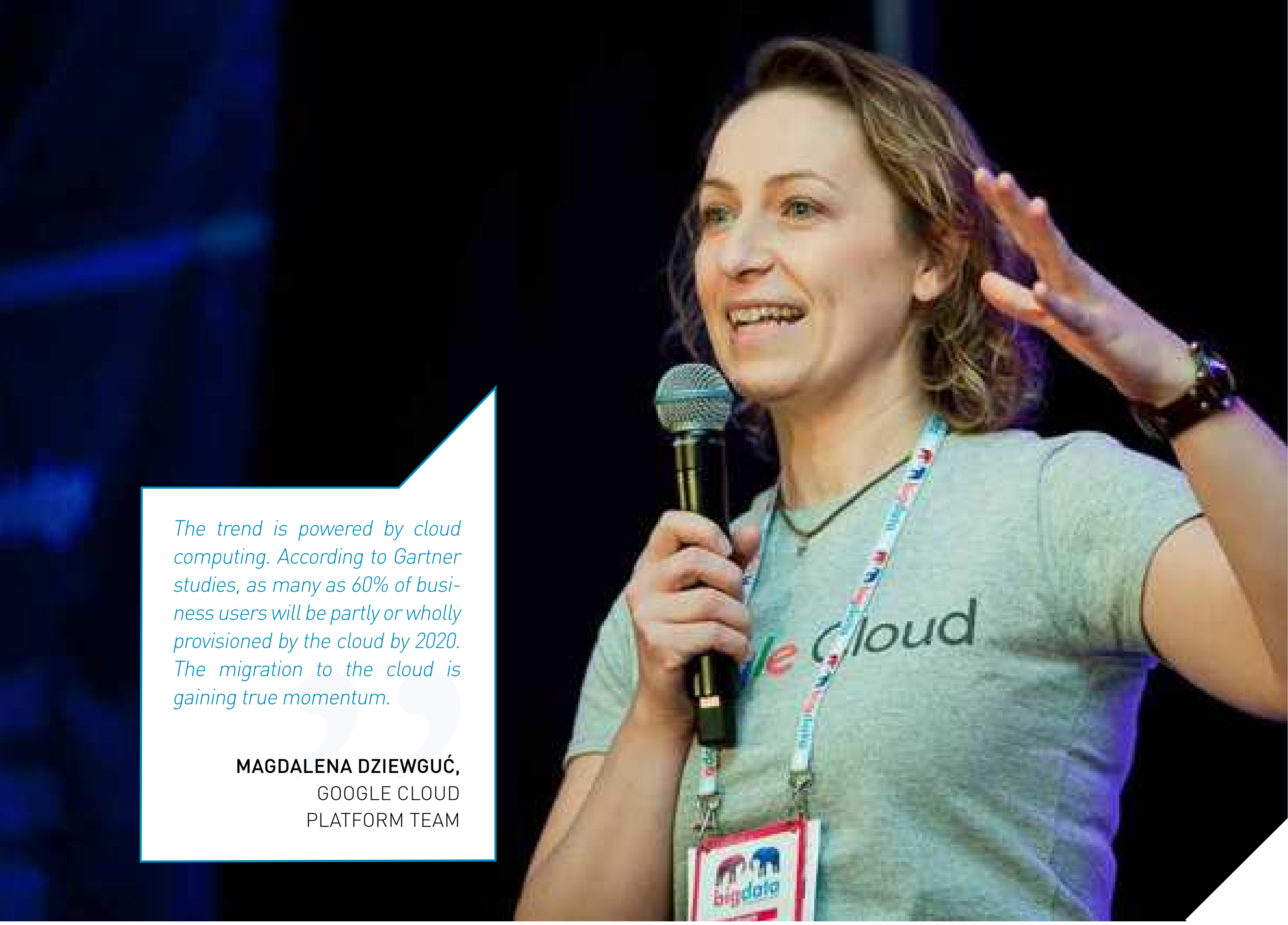
DY: At the same time cloud is getting more popular. Even financial services industry they are starting to using it. There are some banks in the UK that move to cloud. We have to realize, that two years ago banks would not even consider moving big data platforms to cloud. Now the costs are so low, the processes are much simpler, no expertise is required, that cloud is becoming not only more mature but tempting also for our sector. The focus should be on how you build around your big data expertise. Things are changing very quickly, and I wouldn't be surprised if we could move everything to cloud, I'd say in the next couple of years.

OP: Hadoop has so many advantages, but the most important one is that, provided you have the right people, you have an opportunity to move. You stay flexible.

GB: Now I do agree. Flexibility is the key since it gives us the ability to choose. Flexibility is the key point, not the cost because the cost is related to the type of coverage.

Google Team's Vision of Data Horizon in 2017

The conference was opened by They talked about the state of machine learning and the new possibilities of data analysis with the use of cloud infrastructure. Google Cloud is the company's 8th core service after search, YouTube, Google Maps, Chrome, Android, Google Play, and Gmail.



> *The trend is powered by cloud computing. According to Gartner studies, as many as 60% of business users will be partly or wholly provisioned by the cloud by 2020. The migration to the cloud is gaining true momentum.*
>
> **MAGDALENA DZIEWGUĆ,**
> GOOGLE CLOUD
> PLATFORM TEAM

Google's team argued Poland is an important region for the development of this newest strategic service Google Cloud. - We are engaging Polish developers. We have two offices. One in Wrocław, which is more customers centered, and now we are building up a team in Warsaw. - Michał Sapiński is a member of the engineering team in Warsaw.

2016 was a turning point for data analytics

From the Google team perspective, there were two groundbreaking events in 2016 for big data progress. - The first one was the success of Pokemon Go and the second AlphaGo - noticed Magda Dziewguć at her keynote speech: - The people that created Pokemon Go were expecting it to grow rapidly. They had estimated the traffic would reach 30,000 queries per second in a few months. In reality, the app proved to be so popular, it generated 750,000 queries pers second in just two weeks. Exponential growth that follows successful projects in today's reality means everyone needs to be ready for such a traffic. A very strong engine is needed - she added.

In March 2016 a computer program developed by Google DeepMind won a five-game match of Go with Lee Sedol, one of the top Go players in the world. - The game changed the perspective for what's possible in artificial intelligence and machine learning - said Magda Dziewguć.

- The trend is powered by cloud computing. According to Gartner studies, as many as 60% of business users will be partly or wholly provisioned by the cloud by 2020. The migration to the cloud is gaining true momentum.

Cloud solutions are the future

Magda Dziewguć introduced two Google managers responsible for Google Cloud: Diane Greene, who became the head of cloud business after Alphabet (Google's parent company) acquired her startup Bebop, a development platform for enterprise applications, and Urs Holzle, Google's employee #8, who is senior vice president for technical infrastructure at Google.

The speakers underlined the scale of Google Cloud project. - It's an effect of 15 years of the biggest infrastructure investment on the planet. Google is the owner of the biggest backbone on the planet, that covers roughly one-third of the total information

Pure knowledge with the best Big Data experts
big data
Zero produ
alterdata.io
Big Data Technology Warsaw Summit 2017

traffic - explained Magda Dziewguć. - The infrastructure really matters in big data project. It provides security and speed.

Big data is getting easier

Michał Sapiński, a software engineer from Google, talked about the cloud tools available for data processing in the cloud. - Making predictive models preparing data, capturing the data. These are all important tasks. Google's BigQuery is an analytics data warehouse that works really fast. The infrastructure makes the difference - he said. - It's able to perform fast queries, classification program after, and apply gradient descent algorithm.

Michał Sapiński explained machine learning process on the example Google's image recognition. - The neural network is a function that can learn. A pixel has three features. The algorithm can learn a complex hierarchy of these features, recognize complex images, and write a sentence about what it "sees" - he explained. - Merging two algorithms, we are able to recognize and caption images automatically. The image recognition algorithms can recognize abstract shapes, like Yoda toy made of wool.

It was impossible a few years ago. Speech recognition, image recognition, machine translation, molecular activity prediction, road hazard detection, optical character recognition - these are the machine learning applications that became possible thanks to better infrastructure and better algorithms.

- Alpha Go has beaten top two go players in the world. This game is more complex than chess by about 600 orders of magnitude. The winning move #37 in the second game was especially astonishing - said Michał Sapiński.

Building an open source community

Michał Sapiński discussed TensorFlow, which is an open source software library for machine learning developed by Google, and used in its commercial products such as speech and image recognition. - You probably know about its use in Google Search or Google Photos but did you know that as much as 10 percent of emails sent with Gmail is written by a machine - noticed Magda Dziewguć.

TensorFlow has been able to create a vibrant community. For example, it is used by a Japanese farmer to classify cucumbers. It can precisely sort 10 classes of the vegetables. - The great thing about TensorFlow is that the community gives us ready models we can reuse - noticed Michał Sapiński. - We can take this cucumber model, cut the last layers, and retrain it using all the knowledge. The community around TensorFlow is strong - he added.

The big data revolution is happening. - You don't need your own infrastructure, don't have to employ many people in order to realize machine learning projects - underlined Magda Dziewguć.

The Business Perspectives of Machine Learning

The Big Data Technology Warsaw Summit 2017 participants could learn how the biggest companies, such as semiconductor manufacturers, use big data analysis to improve their processes. Sascha Schubert, the Advisory Business Solutions Manager at SAS Institute, business intelligence and data management software and services provider with over 40 years history, shared some of the experience acquired over the decades.

- We are seeing the beginning of what's possible with the application of analytics in commercial organizations - said Sascha Schubert. During his keynote, he talked about the practical use of big data in different industries. - I have been in the industry for quite a time, and I have seen the evolution of analytics that started in small rooms at the perimeters of the organization, and now become the business model of many big companies.

The Big Data conference participants could learn the details of an impressive case study, where a combination of SAS and open source analytics helped a semiconductor manufacturer optimize their production process.- The semiconductor industry is the backbone of the digital age. The sector innovations drive the ability to do more on ever smaller machines, but perhaps equally important is the ability to optimize the manufacturing processes - said Sascha Schubert. - In the digital printing of semiconductor components, 1 in a billion failure rate for droplets may sound like an acceptable rate. This is less so when you consider that up to 50 million droplets can be pushed per second, leading to an unacceptable defect rate of one failure every 20 seconds - he explained.

Towards prescriptive analytics

Sascha Schubert listed four types of analytics useful for business:

- **Descriptive** - providing information on what happened.
- **Diagnostics** - giving information on why it happened.
- **Predictive** - forecasting what will happen.
- **Prescriptive** - giving advice on what should be done.

The last one is the most category, that includes both decision support and decision automation. - It's similar to Alpha Go mentioned in Google Team's keynote. The machine is taking the action, and we are just controlling the process. These include self-driving cars - said Sascha Schubert. - This may mean a real-time decision support, where the machine recommends you take an action, i.e. overtake a car. Or real-time decision automation, where the car drives on its own, and we can only see the effects of its decisions. If you want to listen to the data in your organization, you need to eliminate the garbage.

Why machine learning so hot now?

- It's a perfect storm out there - said Sasha Schubert. - Not until recently did we have the data and computing power that enables practical business implementations. Now we have computers, distributed networks, cloud infrastructure. We have much more data to analyze, the volume is growing as companies gather new types of data. The third part is algorithms. Huge sets of data, and ready algorithms to understand the patterns and use it perfectly - said, Sasha Schubert.

Schubert talked about the main areas big data and machine learning are being used at. - Automated next best offers are a good example. Imagine you go to a store, but the transaction with your credit card is denied because you hit your limit. The bank can merge the information, and make you a credit offer on the spot. It's real-time machine decision making.

The list of potential commercial uses of machine learning is long, and still, new innovative companies are adding new areas to the list. At the moment algorithms are used widely for:

- detecting fraud
- customer segmentation
- online recommendations
- claims management for insurance companies
- targeted acquisition
- real-time ad placements
- natural language processing
- identifying network intrusion

Insurance companies are using image recognition to automate claims management. - A customer may take a picture of a damaged car with a smartphone, upload it to the insurer's system and receive a real-time estimation. The insurance company can find an optimal damage process - explained Sascha Schubert.

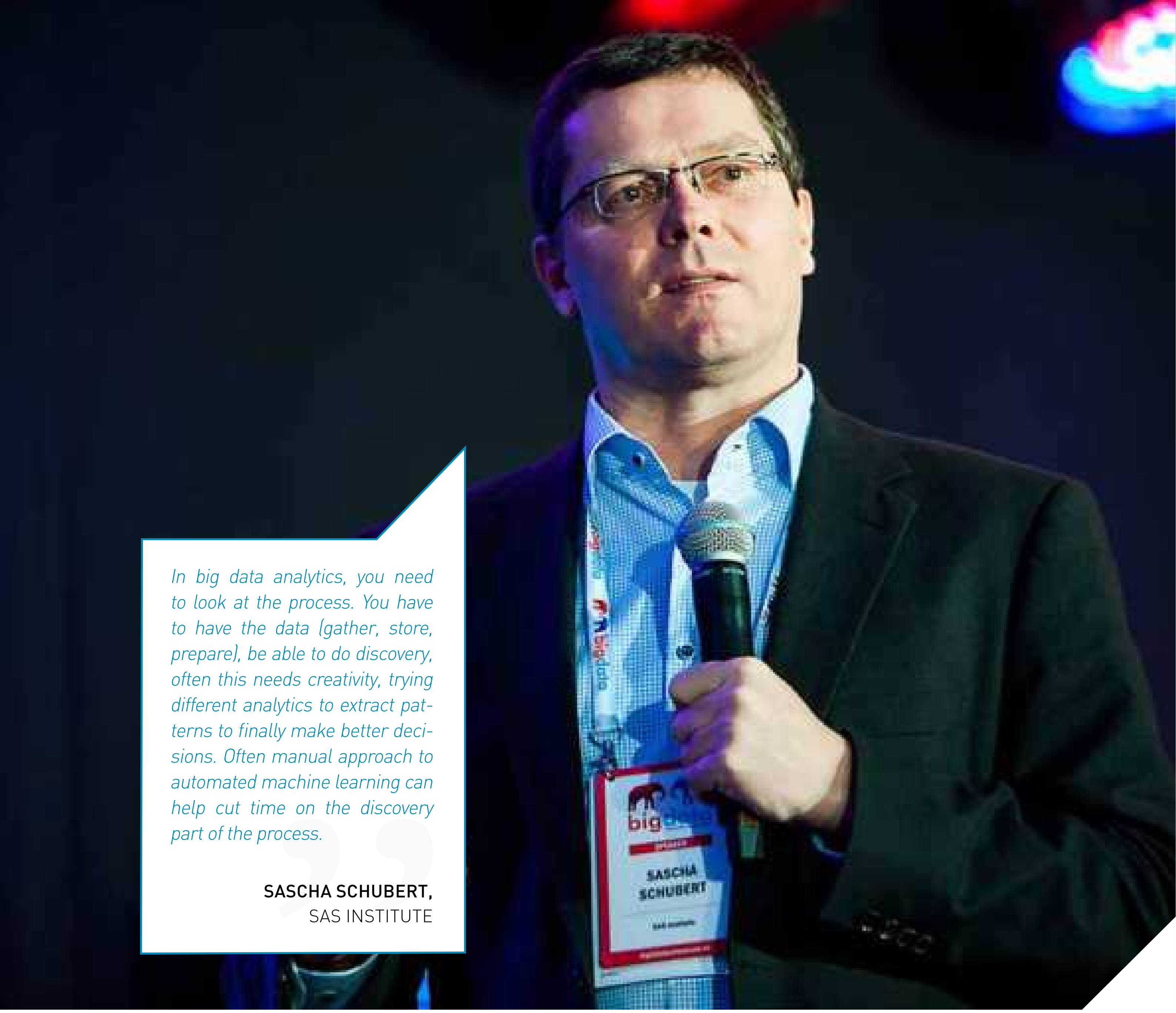
Preemptive maintenance in manufacturing

Another point on the list is improving production quality and efficiency. - For manufacturers a success means creating a product that meets market needs, is reliable, and profitable - said Sascha Schubert. As he underlined, these companies have to make provisions for quality issues. Limiting the number of defects brings manufacturers an immediate benefit.

In the case presented by Sascha Schubert, an automatic quality control using efficient in-memory processing helped the semiconductor manufacturer reduce time to run weekly quality control checks. - The company introduced early warning for wafer yield quality by processing images of wafers use machine learning to identify flaws in wafers. They used pattern recognition to reduce the issues - said Sascha Schubert.

The analytical process was divided into phases: take data, create metadata, sample (2/25 groups), calculate correlations. While in-memory analytics process: input, correlations lead to 10-fold improvement of the process speed.

- The company uses the brightness of the image pixels for analysis. When comparing two images of a matrix of pixels, there are more than 96,000 variables - explained Sascha Schubert. - Then we look at a how a good wafer looks like. You can relate the image we analyze to a topographic map - he added. - We built patch dictionaries, and did unsupervised learning with k-means clustering technique - explained Sascha Schubert.



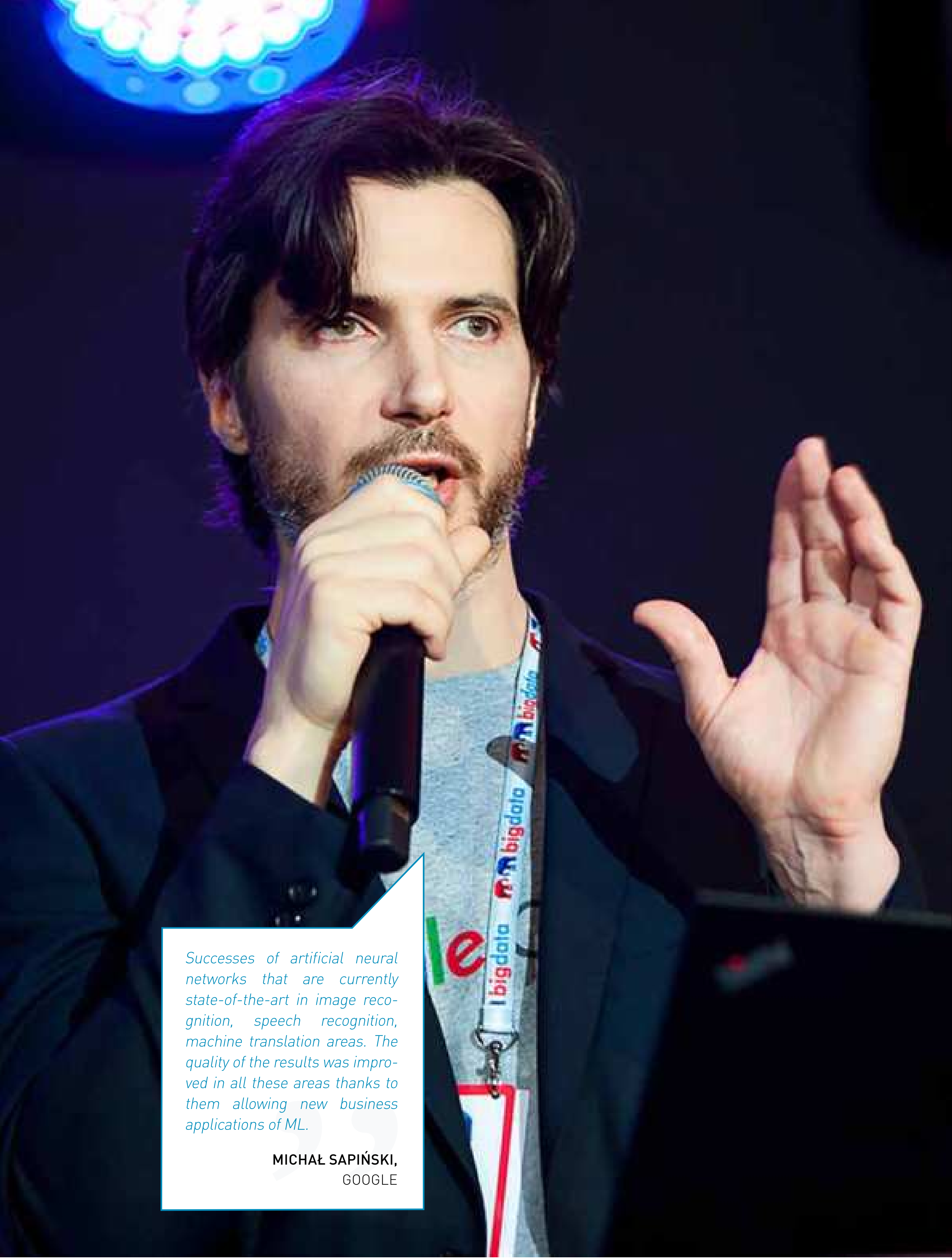
In big data analytics, you need to look at the process. You have to have the data (gather, store, prepare), be able to do discovery, often this needs creativity, trying different analytics to extract patterns to finally make better decisions. Often manual approach to automated machine learning can help cut time on the discovery part of the process.

SASCHA SCHUBERT,
SAS INSTITUTE

In business, deployment is the key

As a result, a machine learning pipeline was created. - In big data analytics, you need to look at the process. You have to have the data (gather, store, prepare), be able to do discovery, often this needs creativity, trying different analytics to extract patterns to finally make better decisions - Sascha Schubert told at the conference keynote. - Often manual approach to automated machine learning can help cut time on the discovery part of the process.

However as Sascha Schubert underlined in the commercial implementations of big data one should always be focused on the outcome. - In business, deployment is the key. If you cannot use it in production, it was just a nice research. The companies look how fast it can be implemented, and how valuable is the process, how much money we are able to save or earn - he said.



> *Successes of artificial neural networks that are currently state-of-the-art in image recognition, speech recognition, machine translation areas. The quality of the results was improved in all these areas thanks to them allowing new business applications of ML.*
>
> **MICHAŁ SAPIŃSKI,**
> GOOGLE

To make computers a little less dumb, and a little more useful

Iinterview with Michał Sapiński, Software Engineer, Google

Why the AlphaGo victory over one of the best professional Go player should be considered such a breakthrough?

Rules of the game are very simple, but the number of game configurations is vast, hundreds of orders of magnitude more than chess. It is impossible to compute all possible configurations - there are more than the estimated number of atoms in the universe. AlphaGo used deep neural nets to develop the ability to 'see' the value of the game position.

What caused the recent machine learning boom?

Successes of artificial neural networks that are currently state-of-the-art in image recognition, speech recognition, machine translation areas. The quality of the results was improved in all these areas thanks to them allowing new business applications of ML.

What are the jobs engineers do at Google in Poland?

The engineering team in Warsaw is involved in developing projects in the area of Google Cloud Platform, Kubernetes, Borg (cluster management system used internally by Google) and others. There are currently over 150 engineers working in Warsaw office.

What are the advantages of using TensorFlow? What is the best way to start working with it?

One of the main strengths of TensorFlow is its flexibility allowing researching new ML model architectures and easy model deployment at the same time. It is also the most popular open source machine learning framework at GitHub at the moment.

The best way to start working with TensorFlow is to visit www.tensorflow.org.

What are the main challenges for achieving true AI?

Researchers seem to agree that they don't know as there are many obstacles we don't even see at the moment. At this point, it's impressive if you can make computers a little less dumb, and a little more useful.

Still, there is a big progress in the area happening at the moment.

How to Verify the Effectiveness of a Change

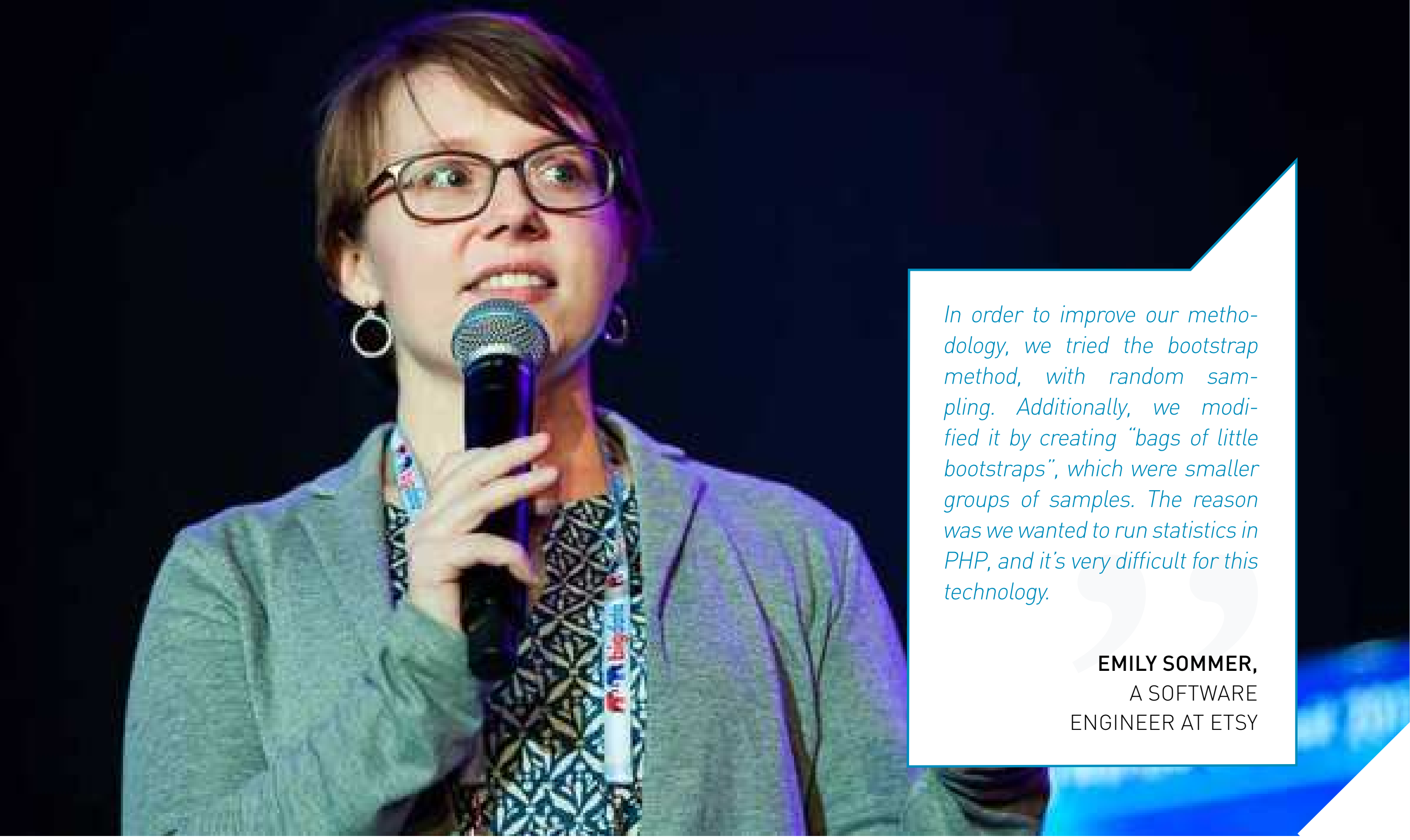
Emily Sommer, a software engineer at Etsy, a p2p e-commerce service with over 54 million registered users and about 20 million active buyers, spoke in her keynote about the experiments she runs on the website's UX.

Etsy team found out the experimentation methods they had used are broken and needed fixing. Surprisingly they way to fix it was to run A/A tests instead of A/B tests. What helped them improve their analysis was a statistical method called bootstrapping.

The value of A/A tests

- We run A/B test on almost all new features. That's how we evaluate them - said Emily Sommer. - We just flip a coin and divide our users into two groups. Everybody has a 50 percent chance to see the new feature. We measure the effect and check if the result is statistically significant. The more spread up the data is the less precise we can get.

One day Emily decided to run some A/A tests to check if the methodology she used was correct. - We split the groups as usually but made no difference in the interface. We ran a bunch of a/a test, and saw tons of significant metrics - Emily Sommer told the participants. - We fixed some issues with login on native apps, but the results were still bad. Our tests showed different results on the same interface.



In order to improve our methodology, we tried the bootstrap method, with random sampling. Additionally, we modified it by creating "bags of little bootstraps", which were smaller groups of samples. The reason was we wanted to run statistics in PHP, and it's very difficult for this technology.

EMILY SOMMER,
A SOFTWARE
ENGINEER AT ETSY

The reason was we weren't following statistical rules.

Etsy's experience with the bootstrap method

Etsy team have found some issues. It turned out the data points were wrong. One user could have many sessions, and they treated them as independent data. - With millions of users with about 50 visits in a week, it becomes a huge problem. - In order to improve our methodology, we tried the bootstrap method, with random sampling. Additionally, we modified it by creating "bags of little bootstraps", which were smaller groups of samples. The reason was we wanted to run statistics in PHP, and it's very difficult for this technology. - said Emily Sommer. - The data analyst I was working with proposed a bag of little bootstraps algorithms. We took a sample (our bag), we select a subset than we bootstrap within this sample. This lets us draw conclusions with smaller sets. We are doing 50 bags and 50 bootstraps in each bag.

Etsy's methodology is suitable for distributed system because it requires access only to small portions it's faster for us. It proved to be correct. Etsy took existing data and simulated A/A tests. They saw much lower false positive rate.

How does Emily Sommer plan to use the discovery? - We want to try hyperparameter optimization (poison bootstrap). Instead of pulling samples each time, it's a common way to do it. Right now it's our native app experiment. We want to move to larger experiments. Power calculations and more complicated metric - she explained the participants.

A data scientist in a consulting company

Paweł Goduła, the senior data scientist at BCG Gamma, told the conference participants how to bring money to the table with data science. He disclosed some details on practical examples of data science "in action" from recent BCG Gamma projects. He also explained when to use Linear Regression vs XG Boost in business applications.

Paweł Goduła graduated from Warsaw School of Economics with a degree in quantitative methods. He later did an MBP in Singapore, where he stayed for a year and a half. In the free time, he started competitions on Kaggle, where companies give their data to the community asking for help in solving a particular problem. For example, Airbnb could challenge the community to predict the next booking of a particular user.

- Solving this type of problems is fun. It was enough to convince me from computer gaming - says Paweł Goduła. - I was playing Starcraft II and Heroes of Might and Magic professionally. Kaggle made me change my habits. It was still competitive, but more productive.

As a data specialist, you want to be at the intersection of statistical knowledge with business

According to Paweł Goduła, working as a data scientist for a consulting company is complex. - You work with consultants, who care

On today's job market you want to be in the intersection: match statistical knowledge with business. Statistical and business savvy. There are few people who can do it available, so you can imagine the implications it will have for you.

PAWEŁ GODUŁA,
BCG GAMMA

about earning money. They ask me "Pawel is this model going to bring us any money?" or "When the model will be ready?" - said Paweł Goduła. - On today's job market you want to be in the intersection: match statistical knowledge with business. Statistical and business savvy. There are few people who can do it available, so you can imagine the implications it will have for you - he added.

Data scientist's areas of expertise are computer science, artificial intelligence, statistics, machine learning, and applied mathematics. While consultants specialized in analytics have experience in consulting industry and domain experience in sectors and functions.

Solving a problem for a coffee chain

We are working or a large chain of coffee houses with over 20 million customers in hundreds of locations and thousands of stock keeping units. The company found out that traditional source of revenue stopped working. Traditionally they would add a coffee house, at some point a next coffee house give less and less incremental revenue. If we cannot extend the chain, it's hard to extend the customer base.

In this case, the best way to grow revenues is to increase the value of each customer. Let's personalize the offer and content and adjust it to personal needs and preferences.

This was the goal.

We had data from loyalty program and mobile application, where you could order favorite coffee. Every transaction is tied to a loyalty card. We looked at the historical behavior of "Annie". We analyzed it day by day. We analyzed Annie's reaction to any type of offer. We knew her the geo-location patterns from the app, so we could apply weather data. Once we did that, we had command center of all customers.

We distil all the captured, transformed and predicted signals into accessible command centers that allow for:

- Exploration: understand trends, build hypotheses
- Exploitation: leverage insights and craft tailored offers/messages/initiatives

All the information about the customer in one place, with metrics derived by our analytics hub:

- Propensity to buy a specific product
- Preferred location (and location type)
- Propensity to redeem offers and be engaged ("game score")
- Risk of attrition
- Headroom potential

Personalized offers are a combination of six elements: the product, offer type, reward level, timing, context, and channel.

Making the optimal upsell offer

- Of Course, the product is very important, as is the type of offer. Some people want a discount, some want a next one cheaper, some want to collect points. You can A/B test this preferences, and the details - explained Paweł Goduła. - There are lots of possibilities. Do I give 10 or 20 points to this person? How much potential do you see in a client? If you want to be really smart, you see the potential of their friends, you want to invest in influencers

According to Paweł Goduła timing of an offer is almost as powerful as a product. An offer for a sandwich half an hour before customer's lunch break can prove to be a game-changing detail. The same with context. It's better to offer frappe on hot days, and hot chocolate when it's cloudy. - In order to achieve the best offer, these elements should play as a team, not as superstars. We need a model that can do all the six things together - Paweł Goduła explained the conference participants.

In the BCG Gamma product recommendation model, the analysts had a past data of purchases for every client. - We picked a singular value decomposition (SVD) model that won Netflix competition. How much client nr 1 likes product 1 based on past data. The objective is to find a secret space, where the client and the product features are aligned. Cross-validation gives you the best module quality - explained Paweł Goduła.

How to deal with very scarce data?

The audience could also participate in solving a real life challenge. - Let's assume we would like to train a model for clients visiting a coffee shop in Krakowskie Przedmieście street in Lublin, my hometown. The problem is we have just history on two clients there. This means our recommendations will be super random - said Paweł Goduła. - How do we fix it? We get the data from all the coffee houses in Poland. The problem is the population may not be homogenous. Is the taste in Lublin the same as in Zakopane? As you move away from Krakowskie Przedmieście street, we get more data, but the data is less relevant - he added.

There's a trade-off between the amount of data and relevance. - We tried to adjust the model looking for the most accurate, which in this case proved to be taking the data from the coffee shops in the same city. With multiple testing of this approach, we achieved 70-80% accuracy in predicting next buy on the test set - said Paweł Goduła.

People from Lublin share the same context. - We don't have to discuss these issues, we can calculate all of them and check, which is the best. When we took the data on Lublin level it gave us the best result. In every model I was training (majority), you need to be able to check the optimal and set the optimal size of the training set to maximize the accuracy - said Paweł Goduła.

How data analysis can make a difference

In the discussed business case BCG Gamma was able to quadruple the revenues from clients loyalty programs. - It was one of the most successful cases we did. We changed fundamentally how the company works - said Paweł Goduła.

Instead of 4-week lag to read the results from offers, we introduced one-day turnaround, with automated templates instead of manual creative work to craft each offer. The company sent 30 different personalized emails a week. We made it possible to send over 350,000 variants per week. Finally, the offers used to be chosen based on the past experience, and now are optimized automatically.

Scalable Analytics for Microservices Architecture

Avito, the third biggest classified ads website in the world, following US Craigslist and 58.com from China. Avito is not a monolith project but comprises dozens specialized vertical sites and applications. Nikolay Golov, Chief Data Warehousing Architect, who was responsible for the process from the beginning spoke at the Big Data Technology Warsaw Summit on how he and his team prepared the infrastructure for data analysis.

The introduction of microservice architecture in Avito spawned hundreds of new services. In this situation is was critical to implement common business intelligence infrastructure, one that would be able to collect, process, combine and analyze data from all those microservices and persistent to constant changes.

Avito Analytics is based on HP Vertica MPP database, highly normalized data lake and an asynchronous event bus. Those tools give Avito the ability to use all types of Machine Learning and Reporting tools, manage sites, applications and microservices.

- We were very small, then we grew, we added new cities, more verticals: cars, businesses, and more - said Nikolay Golov. - At the beginning, it was only for selling pets. In the middle of 2013, Avito bought OLX and other biggest competitors. We switched from traffic growth to monetization.

Avoiding data swamp

In 2013 Avito created a business intelligence unit and started analytics. - At the beginning, you don't know the future traffic, the future data sources, and future data monetization tools. You don't know how your data will be used. You need to build an infrastructure that will survive such uncertainties - Nikolay Golov told the conference participants. - We decided not to do data lake, we needed full SQL support. We will have data from different sources but has to be analyzed combined. We needed normalization with simple data models.

While a traditional data lake is Hadoop based and schemaless, the solution implemented by Avito kept full SQL support, single data model and normalization of all incoming data. - You need a schema on reading because different tools see the same data to the scheme. It's easy to convert data lake to data swamp - said Nikolay Golov. That is why Avito created a normalized data lake. It will be able to take all possible data count.

- We did anchor modeling. It's an agile modeling technique using the sixth normal form for structurally and temporally evolving data. It provides the maximum level of harmonization. We used it to store our data. These are real world figures on our data. All attributes are stored (payments, users, articles, etc.) - said Nikolay Golov.

In Avito's historized data there are approximately 1000 tables, and some of them contained up to 2 billion records and were loaded each hour. While Avito's ClickStream data, which sometimes recorded even mouse moves, has more or less 90 tables with 190 bln records, loaded each 15 minutes and performing about 1 billion actions daily. - Surprisingly it works - joked Nikolay Golov.

Dealing with a huge amount of data

Avito started to add new data streams, new analytical tools, and

> *When Avito started implementing microservices, they started with a few hundreds of tables. Now they have more than 2000 tables. The model discussed by Nikolay Golov is used only to store data. Data analysis is done by data science teams.*
>
> **NIKOLAY GOLOV,**
> AVITO ANALYTICS

new services in 2013. The number of systems grew from three in 2013 to 29 in 2016. - Now we have stopped counting. We have 1 billion events. Our number of sources is growing, the volume of data is growing, cluster size is growing, but our vertical cluster is 17 servers. 14 in the main cluster. This is the basic idea - explained Nikolay Golov.

Microservices add new levels of complexity. - There is one big problem of microservices architecture. Even with Polyglot Persistence, each one must have its own database. Imagine problems with getting data from spawning and dying microservices - said Nikolay Golov.

Avito's solution was event stream processing (ESP). Each microservice just throws events into a pool, and the pool knows who subscribed for this type of event. - We tried Kafka, NSQ, and decided to use our own approach. The idea is you implement a new microservice, you register your event in a universal event register after it is only one required operation. Once you implemented login, it will start throwing them to event processing, data will move it to normalized model. It will appeal to new records - explained Nikolay Golov to the Big Data Technology Warsaw Summit audience.

When Avito started implementing microservices, they started with a few hundreds of tables. Now they have more than 2000 tables. The model discussed by Nikolay Golov is used only to store data. Data analysis is done by data science teams.

One System One Architecture Many Applications

AB Initio, a general-purpose data processing, and metadata management platform. AB Initio has a single architecture for processing data on HDFS, regular files, database tables, message queues, web services, and metadata. This architecture enables virtually any technical or business rule to be graphically defined, shared, and executed in a timely manner.

- It is a true big data architecture, being able to process data in parallel across multiple processors, even processors on different servers such as Hadoop - Firat Tekiner, a data scientist and big data architect at AB Initio told the conference participants. It can run the same rules in batch and real-time and within a service-oriented architecture. It is fully production ready and supports distributed checkpoint restart with application monitoring and alerting. And it enables end-to-end metadata to be collected, versioned, and analyzed by nontechnical users.

The challenges for making Hadoop fully operational

AB Initio delivers a rich set of software products that work together in a way that makes it easy to rapidly develop big data systems. The building block of these systems is the AB

> *AB Initio does everything with a graphic interface. Working with it is like building with Lego bricks. We want to make it as simple for our customers as possible.*
>
> **FIRAT TEKINER,**
> AB INITIO

Initio graph, which combines AB Initio processing components, third-party programs, and any necessary custom codes into a high-performance parallel and distributed application.

There are challenges for making Hadoop environment fully operational. - For 2015-2016 I've been doing serious data mining. What I found, I was spending the most time on preparing the data. We try to bridge this gap by joining sparse data sources together said Firat Tekiner.

AB Initio does everything with a graphic interface. Working with it is like building with Lego bricks. We want to make it as simple for our customers as possible. - Everyone wants to create a data lake. Everyone is populating databases, but they have no idea what's inside. Every data is cataloged not only in Hadoop but across the environment. The business side doesn't care where the app is running, they want to have a report in the morning - said Firat Tekiner.

The hazard of getting lost in the data

As he explained, companies have problems with moving everything within their data lake. They don't have to do it. They can keep them in the same place, it takes the time to build this system, the data is dynamic. - Everyone wants to look for a broader data. They want 7 years instead of one year. That is why they need a solution for data integration challenge - said Firat Tekiner. - You can run a query, we run it through the data sources, and you get results.

- It took one of our clients, a media company, 24-months to develop a big data solution. They called us in, and little proof of concept. We joined their XML data with their two databases. They have been running this in production for 4 years, they had bad records and they didn't notice. It was costing business real money – they didn't look into the text files. We captured this in a very operational, graphical way.

According to Firat Tekiner, the biggest barrier for the companies to adapt AB Initio solutions is that it is not an open source tool. The mindset is difficult to overcome. But we think that the benefits outweigh using something else. Productivity is much higher using our solutions compared to open source.

Making Deep Learning Accessible to Everyone

Jo-fai Chow, Data Scientist at H2O.ai talked about the motivation and benefits of Deep Water. He showed how to build and deploy deep learning models with or without programming experience using H2O's R/Python/Flow (Web) interfaces.

Deep Water is H2O's integration with multiple open source deep learning libraries such as TensorFlow, MXNet, and Caffe. On top of the performance gains from GPU backends, Deep Water naturally inherits all H2O properties in scalability, ease of use and deployment.

- Deep Learning is only one part of a bigger system. H2O.ai data science platform combines many different interfaces. As a result, you get one web interface to build a model quickly, just by clicking. I use Python, but Java users have their own interface. It's one of top 10 machine learning platforms - said Jo-fai Chow.

He explained in detail the structure of Deep Water. He presented it as a set of four most powerful deep learning tools.

First one is TensorFlow backed by Google. It's hugely popular open source machine learning framework by Google.

- Python / C++ API
- TensorBoard
- Data Flow Graph Visualization

> *H2O's goal is to put together all the powerful deep learning tools into one platform. - This is just a beginning of Deep Water. We allow users to have more choices. You can use H2O to tackle more problems. It opens up the opportunity to tackle more problems with machine learning.*
>
> **JO-FAI CHOW,**
> H2O.AI

- Multi CPU / GPU
- v0.8+ distributed machines support
- Multi devices support
- desktop, server and Android devices
- Image, audio and NLP applications
- HUGE Community
- Support for Spark, Windows

The second one is MXNet for Deep Learning. It's open source, too. It's portable, efficient and scalable. In the end of 2016, it was chosen by Amazon as Deep Learning Framework.

The third one is Convolution Architecture For Feature Extraction (CAFFE). - It is also open source, so you can reuse someone's model without reloading it - explained Jo-fai Chow.

- Pure C++ / CUDA architecture for deep learning
- Command line, Python, and MATLAB interface
- Model Zoo
- Open collection of models

- Finally, the fourth framework is our own H2O Deep Learning. It supports both supervised learning, where you can create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations, and unsupervised learning, where autoencoders can detect anomalies by finding outliers using a nonlinear dimensionality reduction using deep learning - said, Jo-fai Chow. - Why do we integrate all the tools into one single platform? We want to give the power to people. Everyone can use it. TensorFlow, MXNet, Caffe and H2O democratize the power of deep learning. H2O platform democratizes artificial intelligence & big data science - he added.

Jo-fai Chow explained that his company goal is to put together all the powerful deep learning tools into one platform. - This is just a beginning of Deep Water. We allow users to have more choices. You can use H2O to tackle more problems. It opens up the opportunity to tackle more problems with machine learning - he explained.

The participants could get familiar with Doop Water's user interface, and its possibilities. - The user has just one interface, while there's a lot underneath - said Jo-fai Chow. - We'll be able to use all the structures for different networks. We can easily stack the models together. This is quite powerful - he added.

Convolutional Neural Networks enabling Image, video, speech recognition

Recurrent Neural Networks enabling natural language processing, sequences, time series, and more

Hybrid Neural Network Architectures enabling speech to text translation, image captioning, scene parsing and more

By utilizing the one interface, H2O.io lower the entrance barrier for new users. Jo-fai Chow discussed the networks available in Deep Water.

- LeNet
- AlexNet
- VGGNet
- Inception (GoogLeNet)
- ResNet (Deep Residual Learning)
- Build Your Own

Big Data Technology Warsaw Summit 2017 had a chance to talk with Jo-fai Chow about his company's newest product, as Deep Water has been just released. - To complement existing offerings like Sparkling Water and Steam, H2O.ai is releasing Deep Water, a new tool to help businesses make deep learning a part of everyday operations - Jo-fai Chow said.

Data Engineering in Facebook Ads teams

How does Facebook process its data? What does data engineer job look like in one of the world's largest corporations? At Big Data Technology Warsaw Summit 2017 you could talk about it with Paweł Koperek, Data Engineer at Facebook.

Facebook serves ads from over 4 million advertisers to more than a billion people each day. - Every day we face the challenge of building the best products to such a large user base. In order to focus the right ones, we have to make well-informed decisions, which we can prove with data - Paweł Koperek told the conference participants.

This is why making information easily accessible and understandable is crucial for the success of the whole team. This talk provides an overview of how Facebook uses data to run the Ads products teams. Paweł Koperek discussed embedding data engineers work within engineering teams, their impact on the product, have a look at techniques which help with standardization and organization of metrics to manage the complexity of data in a scalable way.

Managing data from over billion daily users

- What is Facebook scale? We are a whole family of apps and products, they are being used by billions of people monthly: Facebook by 1.86B, Messenger by 1B, WhatsApp by 1.2B, and Instagram by 600M each month. Average daily number of users is more than a billion - said Paweł Koperek. - The best feature of all of this is that you get it for free. We have to invest in maintenance, financial investment, this brings us to ads, this is the way we try to make money. At first, users are a little bit annoyed that you have to see an add, as a business user you get the opportunity to advertise.

In Facebook, there are multiple teams that try to work on new ads format. We do a lot of research and analysis, the volume of data is huge, without data engineering it's hard to answer the simplest question. Data engineers report directly to the general manager of the product. We are accountable for making an impact on a product.

How data scientists and engineers set the tone in Facebook

Data scientists, they enable us to maximize the impact we have. Data scientists are among the first people, who can state the health of a product. "You're losing users" – we can read this kind of signals. In many cases, it's data scientist or engineer that sets the tone.

Because we ground all our opinions and recommendations on hard data, our claims are difficult to question, if we can prove a theory, we have a strong position in the organization - said Paweł Koperek. - If you have a piece of working software or a proven with data theory it's hard to question it.

Paweł Koperek has two golden rules. First, be open, since it's easy to become biased. Second, move fast, since all the products in Facebook are in customer space, they come and go, the context can be switched easily.

The right interface is the biggest challenge

The biggest challenge is creating the right "interface" to data. This enables the people to answer the questions. It has to scale with the number and complexity of the questions. The technique I use in my everyday job.

What is Facebook scale? We are a whole family of apps and products, they are being used by billions of people monthly: Facebook by 1.86B, Messenger by 1B, WhatsApp by 1.2B, and Instagram by 600M each month. Average daily number of users is more than a billion.

PAWEŁ KOPEREK,
FACEBOOK

- There's nothing unexpected in the diagram of our warehouse. These people (product managers, software engineers) should be able to understand the interface. Let's see how we can do it - explained Paweł Koperek. - Let's imagine we are in a multinational company selling multiple products online. We try to increase the revenues. There is a team that has a hypothesis: in Europe our customers are choosy, if we could optimize the website, we can squeeze some more from there. Data engineers build tables, start transactional data, weblogs (sessions, pages), we limit it to Europe, we merge it, combine together, then we reorient it to the user level. We'll find out it the theory we set is true - said Koperek.

Data is the New Oil

Krystian Mistrzak and Thejas Murthy, data engineers at Fandom Powered by Wikia, gave conference participants an overview of comparisons of existing tools and emphasized on why they chose Airflow, and how Airflow is being used to create a stable and reliable orchestration platform to enable non-data engineers to seamlessly access data by democratizing data. They also shared some tricks and best practices of developing workflows with Airflow and showed how they use it.

Fandom powered by Wikia is the largest entertainment fan site in the world, with more than 360,000 fan communities and a global audience of over 190 million monthly unique users. Being the largest entertainment site, Wikia generates massive volumes of data, which varies from clickstream, user activities, API requests, ad delivery, A/B testing and much more. The big challenge is not just the volume, but the orchestration involved in combining various sources of data with various periodicity and volumes.

In order to make sure the processed data is available for the consumers within the expected time, and help the analytic team to get the right insights at right time Fandom engineers decided to use Apache Airflow.

Foreseeing unexpected surges in traffic

- When you think of the process the crude oil needs to go to become fuel for cars, you can divide it into four phases: extract, transform, load, use it - Krystian Mistrzak told the audience. - The same applies to the data. Data is the new oil, it makes sense from not only business perspective.

Data engineers working in Fandom had lots of valuable experience to share. - The business problem is silent but violent. Consider the traffic peak generated by the last episode of the most recent season of "Game of Thrones" aired in June 2016. There were over 9M pageviews that day. The sales guy responsible for running campaigns and selling ads predicted 3.5M based on the history data for the peaks. He predicted 3.5M instead of 9M - said Krystian Mistrzak and missed the opportunity for additional 50k USD of revenue.

The main challenges for data engineers

Fandom engineers extended Airflow with built in SLA mechanism. - We did it to create automatic Jira tickets for the team. Everybody wants data, what is more, and more challenging when the volume is growing. Business wants at once to query different data sources - said Krystian Mistrzak. As he explained, a reliable data pipeline needs to be monitored, robust, maintainable, scheduled and collaborative

- A rotten apple spoils the barrel, you may have a perfect notification system, you know the people, you know the nodes, but adding new nodes or changing configuration is time-consuming. When the time to take action is too long it spoils everything - said Krystian Mistrzak.

Fandom's best practices

- Everyone wants data, and it shouldn't be delayed, especially if it leads directly to a loss in revenue - said Thejas Murphy. This business problem can be reduced to a technical problem, which is querying different data sources at scale.

In order to create such a robust pipeline, Fandom data engineers set up a set of rules for querying different data sources:

Thejas Murthy,
Fandom Powered by Wikia

- jobs are defined as DAGs (directed acyclic graphs)
- DAGs are defined in Python
- set of the tasks you want to run is prepared
- tasks are defined as operators
- DAG describes how to run a pipeline, operators what to run
- dynamic pipeline generation is enabled

Krystian Mistrzak and Thejas Murthy had some advice for the engineers working with big data pipelines:

- build a data platform that: enables data-driven decisions, is ready to scale, and is ready to evolve and change
- keep calm and code dynamic DAGs
- maintain and automate
- propagate good news fast and bad news faster

The conference participants had a unique chance to talk with Krystian Mistrzak and Thejas Murthy about data engineering at one of the world's most popular websites.

Working with different types of data is crutial

Interview with Sascha Schubert, the Advisory Business Solutions Manager at SAS Institute

How long does it take to implement prescriptive machine learning solutions in a business environment?

It depends on the industry. The financial sector is leading, along with telecommunications, and the companies that have access to people with smart devices. A lot of people do their finance on smart devices. This way the data gathering process is much faster. You also have much more direct contact with the customer. Customer intelligence is a big part of analytics. Web optimization is also very advanced. You have credit score online or real-time ad placement on websites.

Transport companies, that have sensors on their vehicles and crafts, are also gathering a lot of valuable data. Actually much more than a simple customer mobile device. Here you have a larger issue with a data storage and data preparation.

And now manufacturing is getting into this. In manufacturing, you have preemptive maintenance. Many organizations have large fleets - airlines for example. The fastest projects are the ones that have to do with smart devices and social media. I would guess maybe 18 to 24 months from inception, when the organization realizes: "we need to do something", to going live.

Then projects as complex as preemptive maintenance may take a little bit longer since there are more factors involved.

How can you know that at some point you can give the decision making to an algorithm?

You have to monitor it all the time. We have a feedback loop: raw data goes to the analytical process, then implementation. We close that loop with a monitoring process. We take look at a decision made by a machine, and we check if the decision was the right or wrong. You can have a business or analytical measures for that.

Obviously, having the feedback you have to retrain the system. You may do it with a continuous loop or from time to time. Monitoring gives you the points on when you have to go back and retrain the

That's a huge set of skills that we are looking for. That's why we see team approach working best. Having this creativity and curiosity, saying "I wanna try something", and we are giving them the tools and environment to do it quickly.

SASCHA SCHUBERT,
SAS INSTITUTE

system. Making sure that everything is working accurately is what you do before you decide to let the machine make some actual decisions.

When can I make the decision to trust the algorithm?

You have to estimate the accuracy of the model, let's say 80-90 percent in cases that the error is not that critical in preemptive maintenance might be 99 percent. If your indicator is "this piece of equipment will break down" in the case of an error, it should be very high.

What kind of jobs are most wanted, what kind of specialists are you looking for?

First, data scientists, definitely. Lots of our customers are moving to Hadoop, so people that can work with data on Hadoop. People that have the skills to turn raw, huge data sets of different types - not only structural data, but also social media data, web data, text data, image data - into data that you can analyze, and then be able to apply analytical techniques and tools to extract the value from the data.

People that have the ability to tell the story are also very valuable for us. Being able to convince business that analytics can help them to make things better to optimize the decision process or maybe even automate the decision process.

That's a huge set of skills that we are looking for. That's why we see team approach working best. Having this creativity and curiosity, saying "I wanna try something", and we are giving them the tools and environment to do it quickly.

Programming has become very important again. Being able to program either open source or SAS. Data management in Hadoop. The combination of open source and commercial tools. That's where we are moving. In general being able to use whatever tool is available, and combine them as well, because not every tool does everything.

We also focus on productivity in our organization. People spend too long in training and testing phase, moving over to deploying the system to production takes very long time.

Roundtable debates

Exchanging opinions and experiences

Parallel roundtable discussions where a unique chance for everybody to engage in the conversation on the most interesting topics. Participants had the opportunity to exchange their opinions and experiences and meet the people that share the same interests. At the Big Data Technology Warsaw Summit 2017, the debates took place at 11 different tables in two rounds

▶ Latest advances in machine learning and their impact on our industries

Hosted by **Michał Sapiński**, Software engineer, Google

The discussion was focused on the latest advancements in machine learning – mostly in the area of artificial neural networks – and their impact on the landscape of industries, tools, and IT professions. Should we expect another AI ice age or this time is different and we are on a good way to solving intelligence?

▶ How to overcome challenges which you can expect while designing and managing the environment, both software, and hardware, for big data analytics.

Hosted by **Krzysztof Baczyński**, Cisco Big Data Lead for Poland, Cisco, and **Kamil Ciuszko**, Founder and CEO, Alterdata

Participants talked about how they perceive hardware in the process of data analysis. How to combine software and hardware to perform fast queries. They discussed a case study of a company measuring customers' movement on cameras in a retailer shop. The system was related to multiple points through the GSM network.

▶ Building an EDW using the Big Data technologies – challenges and opportunities

Hosted by Marcin **Choiński**, Head of Big Data & Analytics Ecosystem, TVN Digital S.A.

How to successfully build an EDW using the Big Data technologies stack? Adapting the EDW methodologies, techniques and best practices (Kimball, Inmon, Data Vault, Anchor, Hub and Spoke) to the Big Data realities. How to plan the program, build the team, choose technologies, infrastructure (cloud vs on-prem), model and process the data, etc.

▶ Being efficient data engineer. Tools, ecosystem, skills, ways of learning

Hosted by Piotr **Guzik**, Software Engineer, Grupa Allegro

Big Data Engineer is quite a new profession. Yet, Big Data ecosystem is big and it is growing rapidly and changing fast. There are a lot of frameworks, tools which are supposed to make us efficient. Some of them can help, while others are obsolete. There are specific use-cases when we should apply different tools and approach. I would like to talk about usage of common frameworks like Spark, Kafka, Hadoop, Camus, Oozie, Airbnb Workflow and others in order to make our life easier. Participants discussed typical issues that occur in daily work and the way we handled them @Allegro. We might also talk about different ways of learning Big Data technologies.

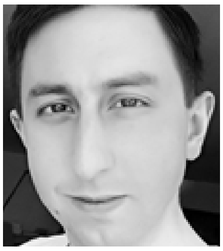
▶ Beyond pre-computed answers – interactive, sub-second OLAP queries with Druid / Kylin

Hosted by Piotr **Turek**, Big Data Software Architect, DreamLab

BigData stands for volume, velocity and last, but not the least, variety. One of the major challenges in modern data engineering is how to produce systems, which not only satisfy the need of our businesses today but are also capable of keeping up with the ever-increasing pace of evolving business requirements – at a palatable cost.

One emerging segment of Big Data technologies enabling us to build such systems are distributed OLAP engines such as Druid and Kylin. The participants talked about the ideal and not-so-ideal use-cases, success and failure stories, operational trade-offs and issues, scaling and optimizing.

▶ Fast SQL solutions for Hadoop



Hosted by Jakub **Pieprzyk**, Data Science Developer, RyanAir

Hadoop was developed as a batch processing solution but it quickly became important also for data scientists and analysts. There are plenty products that give you the opportunity to do fast ad-hoc analysis on big data like Spark, Impala, Presto or Drill, to mention just a few of them. Participants shared their experience with various "SQL on Hadoop" solutions, hear some success stories and also discuss common pitfalls.

▶ Best tools for alerting and monitoring of the clusters

Hosted by
Tomasz Sujkowski, Big Data Administrator, Agora SA

▶ Machine Learning and Big Data: the perfect solution for all problems?

Hosted by
Andrzej Dydyński, Data Scientist, Samsung

▶ Effective tools and environment for data scientists

Hosted by Artur Maliszewski, Head of Business Intelligence, Currency One

▶ How to hire data scientists?

Hosted by
Przemysław Biecek, Co-founder, SmarterPoland.pl

▶ Major challenges in project based on Hadoop environment (lack of measurable results, staffing problems, the high cost of keeping up to date source code, the necessity to deal with many different and fast changing technologies). Data Governance in BigData.

Hosted by
Konrad Hoszowski, Technical Account Manager, and **Firat Tekiner,** Data Scientist and Big Data Architect at AB Initio

▶ Real-time stream processing frameworks – available technologies, their pros & cons, deployment techniques, interesting features.

Hosted by
Fabian Hueske, Software Engineer, data Artisans

▶ **Enterprise requirements for clusters: security, audit, encryption, backups**

Hosted by **Artur Szymański**, Hadoop Administrator, Vodafone

▶ **Machine Learning and Big Data: the perfect solution for all problems?**

Hosted by **Andrzej Dydyński**, Data Scientist, Samsung

▶ **Effective tools and environment for data scientists**

Hosted by **Artur Maliszewski**, Head of Business Intelligence, Currency One

▶ **Expensive mistakes to avoid when building a data platform**

Hosted by **Piotr Kalański**, Data Engineering Team Leader, Step-Stone

▶ **Release process when deploying production data applications**

Hosted by **Paweł Cejrowski**, Big Data Engineer, Grupa Wirtualna Polska

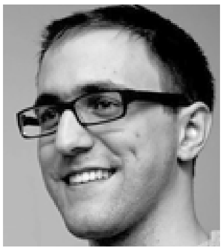
▶ **On-click deployment – how to automate the platform properly and efficiently**

Hosted by **Piotr Bednarek**, Administrator Hadoop, GetInData

▶ **Data visualization – why, how and when?**

Hosted by **Przemysław Biecek**, Co-founder, SmarterPoland.pl

▶ **Large-scale data collection and ingestion – Kylo and other projects (Gobblin, Nifi, Kafka Connect, Camus)**



Hosted by **Tomasz Domański**, Senior Data Engineer, ThinkBig (a Teradata company)

bigdata
TECHNOLOGY WARSAW SUMMIT

Pure knowledge with the best Big Data experts

VERTICA

SUNG

bigdata
TECHNOLOGY WARSAW SUMMIT

ORGANIZERS

EVENTION
CZAS ZAANGAŻOWANY

getindata
big data. experience. passion.

GENERAL PARTNER

sas THE POWER TO KNOW.

STRATEGIC PARTNER

Google Cloud

PATRONAGE

Data Science Group

DATA SCIENCE

DATAKRK

mbn

DATA SCIENCE

SWISSBIGDATA
USER GROUP / ZURICH

WDSG

CONTENT PARTNER

Ab Initio

alterdata.io

BCG
The Boston Consulting Group

cisco

Hewlett Packard
Enterprise

VERTICA

LUXOFT

TERADATA

SUPPORTING PARTNER

allegro

ataccama

codilime

Contact Singapore
contactsingapore.sg

criteo labs

OVH.pl
Innovation is Freedom

P&G

SAMSUNG

www.bigdatatechwarsaw.eu