DETECTION OF POTENTIALLY DANGEROUS ACTIVITIES FROM LOGS OF MOBILE DEVICES USING MACHINE LEARNING TECHNIQUES

DEPARTMENT OF PARALLEL AND DISTRIBUTED INFORMATION PROCESSING
INSTITUTE OF INFORMATICS, SLOVAK ACADEMY OF SCIENCES

Tempest
IT makes sense

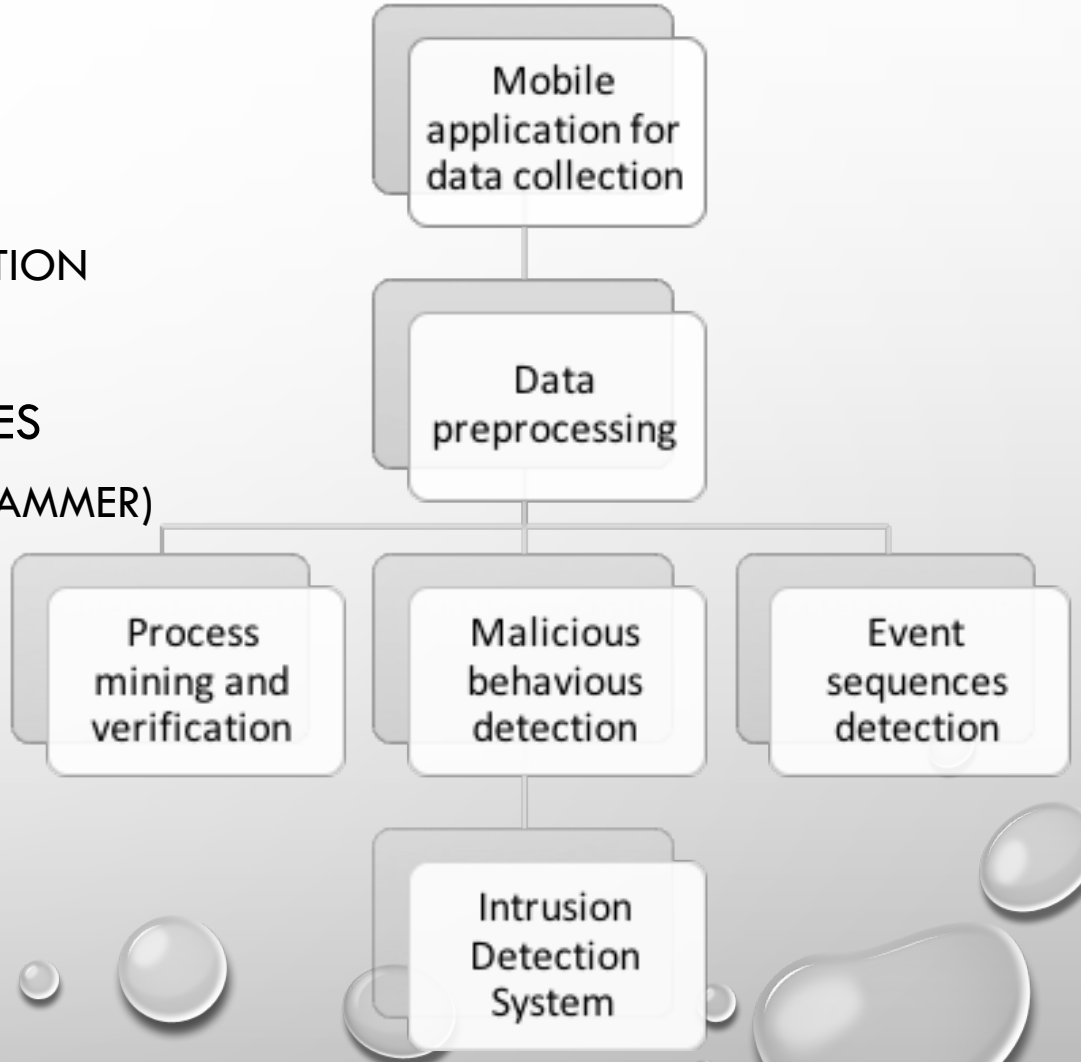
IBM
IBM Slovakia

AIMS AND REQUIREMENTS (IBM)

- **CYBER-SECURITY** RESEARCH FOR MOBILE DEVICES, **BIG DATA TECHNOLOGY ORIENTED**
- RESPONSIBILITIES
  - IBM, TEMPEST A.S.: MOBILE DEVICES, DATA COLLECTIONS, QUEST DEFINITIONS
  - **IISAS TEAM:** DATA MINING USING MACHINE LEARNING TECHNIQUES – **MALWARE DETECTION**
- **SIX MONTHS (6.2015-11.2015)**
  - ALONG WITH DATA COLLECTIONS, INCLUDED COMPLETE DOCUMENTATIONS
  - **PILOT RESEARCH AND DEVELOPMENT WITH RESULT EVALUATIONS**
  - PROOF OF THE CONCEPT, DEFINITIONS, ALGORITHMS, THEOREMS, ANALYZING, EXPERIMENTS ...

RESEARCH INTERESTS - BUSINESS UNDERSTANDING

- **IISAS TEAM: DOC. ING. LADISLAV HLUCHÝ, PHD.**
  - DEPARTMENT OF PARALLEL AND DISTRIBUTED INFORMATION PROCESSING
- DATA MINING USING MACHINE LEARNING TECHNIQUES
  - **MALICIOUS BEHAVIOR DETECTION** (G. NGUYEN, P. KRAMMER)
  - EVENT SEQUENCES DETECTION (Š. DLUGOLINSKÝ)
  - PROCESS MINING AND VERIFICATION (O. HABALA)
  - INTRUSION DETECTION SYSTEM (V. TRAN)
  - INFRASTRUCTURE (M. ŠELENG)



RAW LOGS

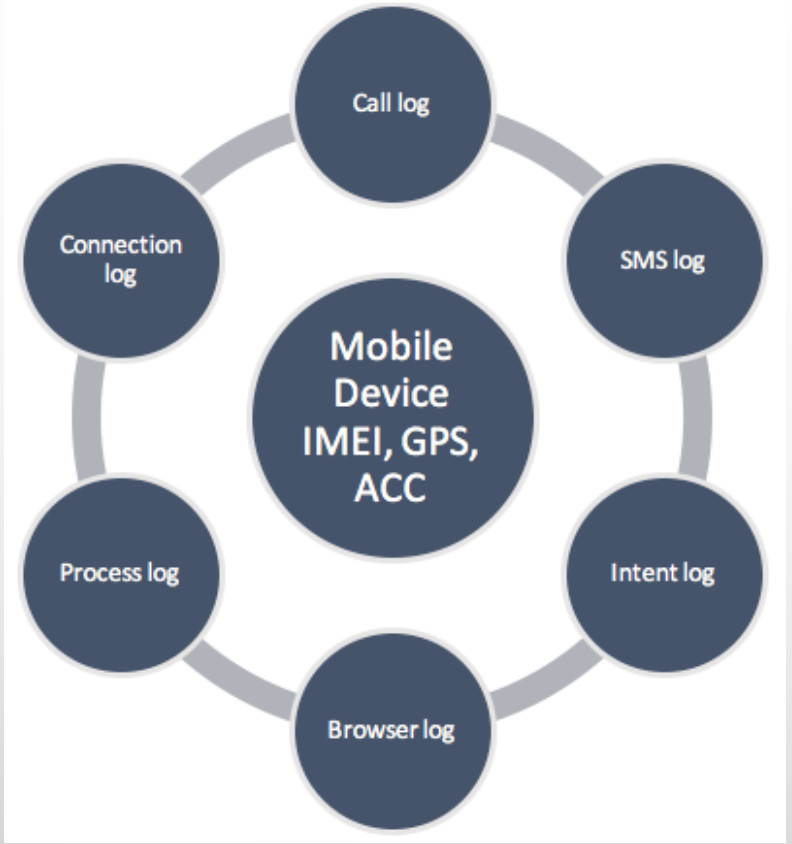

20150908-23:35:55:523§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.hlavnespravy.sk/§§Hlavné správy - Konzervatívny denník§§20150908-18:12:02:176§§{85.248.228.27}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:526§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.pravda.sk/§§Pravda.sk§§20150908-21:50:08:171§§{217.67.31.48}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:528§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://spravy.pravda.sk/ekonomika/clanok/366526-regiojet-ma-opat-problem-s-licenciou-na-autobusovu-linku/§§RegioJet má opäť problém s licenciou na autobusovú linku - Ekonomika - Správy - Pravda.sk§§20150903-17:37:57:394§§{217.67.31.48}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:531§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://spravy.pravda.sk/ekonomika/clanok/366469-slovenske-rajciny-po-cely-rok-domaci-zeleninari-sa-spajaju/§§Slovenské rajčiny po§§20150903-17:39:22:851§§{217.67.31.48}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:534§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://spravy.pravda.sk/domace/clanok/366523-kiska-telefonoval-s-ukrajinskym-prezidentom-porosenkom/§§Kiska telefonoval s ukrajinským prezidentom Porošenkom - Domáce - Správy - Pravda.sk§§20150903-17:43:49:506§§{217.67.31.48}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:537§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://debata.pravda.sk/debata/366523-kiska-telefonoval-s-ukrajinskym-prezidentom-porosenkom/§§Pravda.sk - Debata - Kiska telefonoval s ukrajinským prezidentom Porošenkom§§20150903-17:43:48:572§§{217.67.31.59}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:540§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://debata.pravda.sk/debata/366523-kiska-telefonoval-s-ukrajinskym-prezidentom-porosenkom/?view_mode=vlakna&ordering=od_najnovsieho&strana=2§§Pravda.sk - Debata - Kiska telefonoval s ukrajinským prezidentom Porošenkom§§20150903-17:43:38:565§§{217.67.31.59}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:543§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://spravy.pravda.sk/svet/clanok/366525-orban-sokoval-nemeckych-politikov-imigracna-vlna-je-vas-problem-vyhlasil/§§Orbán šokoval nemeckých politikov. Migračná vlna je váš problém, vyhlásil - Svet - Správy - Pravda.sk§§20150903-17:44:04:015§§{217.67.31.48}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:546§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://debata.pravda.sk/debata/366525-orban-sokoval-nemeckych-politikov-imigracna-vlna-je-vas-problem-vyhlasil/§§Pravda.sk - Debata - Orbán šokoval nemeckých politikov. Migračná vlna je váš problém, vyhlásil§§20150903-17:45:46:794§§{217.67.31.59}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:549§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.aktuality.sk/§§Aktuálne spravodajstvo na Slovensku a vo svete | Aktuality.sk§§20150903-17:52:43:075§§{91.235.52.35}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:552§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.aktuality.sk/clanok/303370/v-konflikte-na-juhovychode-ukrajiny-zahynulo-uz-vyse-6400-civilistov/§§V konflikte na juhovýchode Ukrajiny zahynulo už vyše 6400 civilistov§§20150903-17:50:10:165§§{91.235.52.35}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:555§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.aktuality.sk/clanok/303368/narodne-gmo-zakazy-europoslanci-z-polnohospodarskeho-vyboru-odmietaju/§§Národné GMO zákazy europoslanci z poľnohospodárskeho výboru odmietajú§§20150903-17:50:47:203§§{91.235.52.35}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:557§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.cas.sk/§§ČAS.sk - Najčítanejší denník§§20150908-18:24:55:167§§{91.235.52.131}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:560§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://www.pluska.sk/§§Správy z domova a zahraničia, šoubiznis, krimi a šport | Pluska.sk§§20150903-17:59:19:058§§{85.248.116.198;85.248.116.194}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:563§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://195.28.96.161/apps§§§§20150908-13:02:08:766§§}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:566§§location_unavailable§§-9.107543;-0.277727;2.547431§§http://195.28.96.161/apps/§§§§20150908-18:15:29:647§§}§§80§§http§§4b09f848cd21186f
20150908-23:35:55:569§§location_unavailable§§-9.107543;-0.277727;2.547431§§https://www.google.sk/search?q=facebook&oq=face&aqs=chrome.0.0j69i57j0l2.1988j0j4&client=tablet-android-lenovo&sourceid=chrome-mobile&espv=1&ie=UTF-8§§facebook - Hľadať Googlom§§20150903-18:31:47:034§§{173.194.112.127;173.194.112.120;173.194.112.111;173.194.112.119;2a00:1450:4001:803::101f}§§443§§https§§4b09f848cd21186f


20150907-05:33:16:728§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§754§§com.asus.launcher§§10008§§0%§§0§§0§§100§§0§§17MB§§4MB§§18MB§§10MB§§1016kB§§10MB§§7MB§§19MB§§26MB§§35MB§§25MB§§3565840674137 43
20150907-05:33:16:851§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§850§§com.google.process.gapps§§10019§§0%§§0§§893§§200§§2§§13MB§§5MB§§13MB§§4MB§§1MB§§4MB§§5MB§§13MB§§7MB§§23MB§§19MB§§356584067413743
20150907-05:33:16:918§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§1448§§com.asus.contacts§§10005§§0%§§0§§0§§130§§0§§7MB§§4MB§§7MB§§3MB§§1MB§§3MB§§4MB§§18MB§§10MB§§16MB§§24MB§§356584067413743
20150907-05:33:16:980§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§838§§com.google.android.gms§§10019§§0%§§1§§0§§400§§0§§10MB§§4MB§§10MB§§6MB§§1MB§§6MB§§9MB§§13MB§§11MB§§25MB§§19MB§§356584067413743
20150907-05:33:17:067§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§6361§§android.process.acore§§10005§§0%§§0§§850§§200§§1§§3MB§§5MB§§3MB§§2MB§§1MB§§2MB§§1MB§§14MB§§2MB§§7MB§§20MB§§356584067413743
20150907-05:33:17:115§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§8624§§com.asus.sitd.whatsnext§§10116§§0%§§1§§0§§400§§0§§5MB§§4MB§§5MB§§1MB§§1MB§§1MB§§2MB§§14MB§§2MB§§8MB§§20MB§§356584067413743
20150907-05:33:17:176§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§893§§com.google.android.gms.persistent§§4MB§§7MB§§13MB§§9MB§§25MB§§19MB§§356584067413743
20150907-05:33:17:176§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§12578§§com.google.android.apps.fitness§§1MB§§2MB§§13MB§§6MB§§8MB§§20MB§§356584067413743
20150907-05:33:17:176§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§12234§§com.facebook.katana§§10067§§0%§§
20150907-05:34:26:950§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§192.168.1.110§§33135§§216.58.209.202§§443§§08§§10019§§com.google.uid.shared§§tcp6§§356584067413743
20150907-05:34:26:955§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§192.168.1.110§§53701§§216.58.209.202§§443§§08§§10019§§com.google.uid.shared§§tcp6§§356584067413743
20150907-05:34:26:960§§17.25484552238640;48.09291989707780§§-0.306458;-0.296881;9.538500§§192.168.1.110§§53902§§173.194.122.9§§44

METHODOLOGY AND THE FIRST INSIGHT INTO DATA

- **CRISP-DM** CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING
  - **DEVELOPMENT PHASE**
  - DEPLOYMENT PHASE
- **MOBILE LOGS = HUMAN-GENERATED CLASS**
  - HIGH POTENTIAL IN BIG DATA 3V (VOLUME, VELOCITY, VARIETY)
  - EXTREMELY NOISY FOR SPECIFIC RESEARCH PURPOSES
  - LOW OCCURENCES OF MALWARE RELATED ACTIVITIES
  - EVOLVING CHARACTERISTICS E.G. APPLICATIONS ON DEVICES ARE DYNAMICALLY INSTALLED/UNINSTALLED WITHOUT ANY RESTRICTIONS
- PRIVACY PRESERVING DATA MINING DUE TO CONTRACT CONFIDENTIALITY AND PERSONAL SENSITIVE INFORMATION

EXPLORATORY DATA ANALYSIS (EDA)



DATA PREPARATION

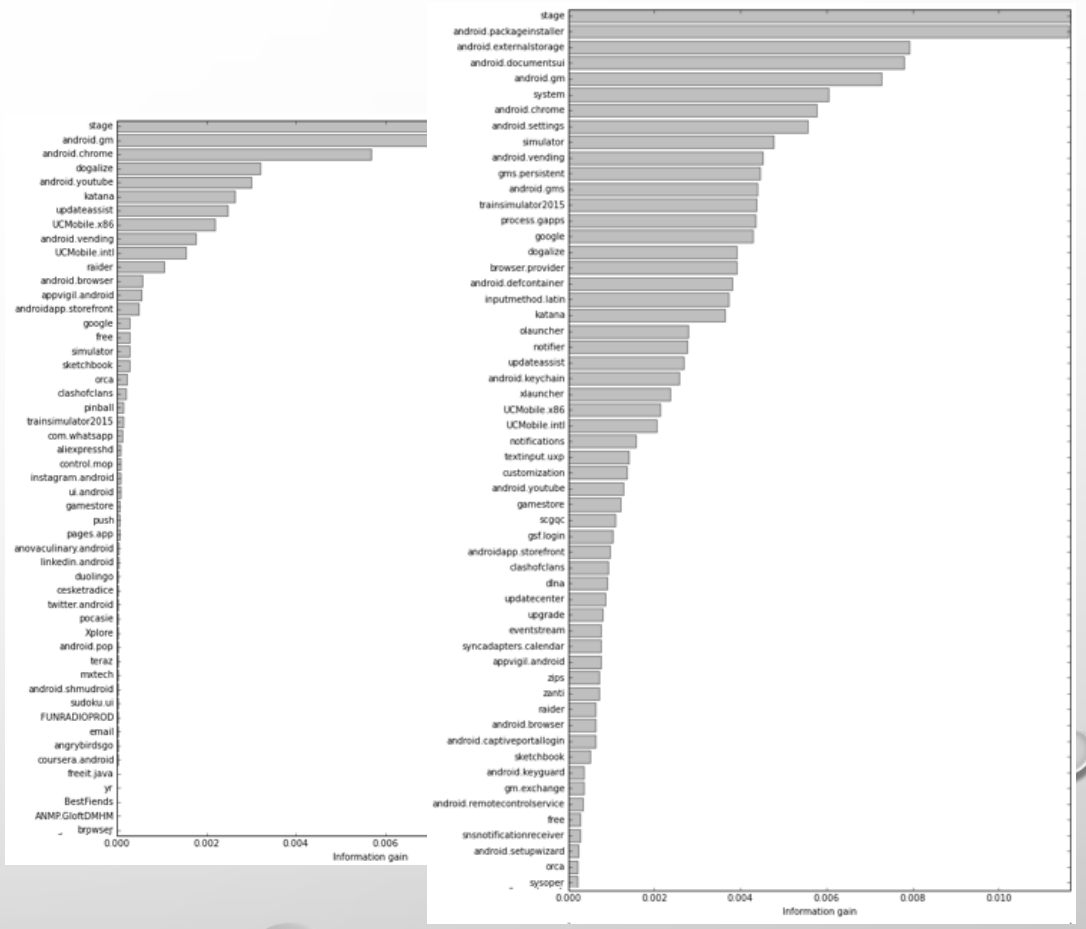
- DATA POOL FOR INCREMENTAL REALIZATION
- TIME-FRAMING SLIDING WINDOWS
- NOISE FILTERING, FILTERS CREATION
- FEATURE ENGINEERING
  - MOBILE DEVICE BEHAVIOR MODELING
  - FEATURE EXTRACTION FROM RAW MOBILE LOGS
  - FEATURE SELECTION
- DATA TRANSFORMATION FROM RAW LOGS INTO MACHINE LEARNING DATA

FEATURE SELECTION

- ENTROPY

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i)$$

- **INFORMATION GAIN (IG)**
- INTRINSIC VALUE (IV)
- GAIN RATIO (IGR) IGR = IG/IV
- THRESHOLDS, EXPLORATORY DATA ANALYSIS



MODELING: TRAINING OF PREDICTIVE MODELS

- TRADITIONAL IN-MEMORY LEARNING: OFF-LINE LEARNING
- DISTRIBUTED LEARNING: COUPLED WITH INFRASTRUCTURE
- **INCREMENTAL LEARNING: ADAPTIVE ONLINE LEARNING, ONE-PASS OVER DATA**
- SUPERVISED LEARNING, BINARY CLASSIFICATION
  - SUPPORT VECTOR MACHINE (SVM)
  - LOGISTIC REGRESSION
  - FEED-FORWARD NEURON NETWORK
- HIGHLY IMBALANCED ML DATA CLASSES: BOOSTING OF POSITIVE EXAMPLES, REGULARIZATION TO PREVENT OVERFITTING

RESEARCH RESULTS AFTER SIX MONTHS

- **MODEL PERFORMANCE EVALUATIONS**
  - SEQUENTIAL AND CONCURRENT TRAINING DESIGNS
  - ACCURACY, PRECISION, RECALL, F1, MCC, RMSE … HOLD-OUT, K-FOLD, …
  - MATTHEWS CORRELATION COEFFICIENT (MCC) <0.89, 0.92>
- **MALICIOUS BEHAVIOR DETECTION**
  - SATISFIED RESULTS TO DISTINGUISH MALICIOUS BEHAVIOR FROM NORMAL ONE
  - DETECTION FOR MALWARE TYPE WITHOUT BINDING TO CONCRETE SPECIFIC ONE
- **MORE TIME AND DATA FOR MORE REAL AND REAL-TIME DETECTIONS**

BRIEFLY ABOUT OTHER RESEARCH QUESTS

- **EVENT SEQUENCES DETECTION USING NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES**
  - ATTEMPT OF SEMI-SUPERVISED LEARNING, BIG DATA APPROACH OF DATA TRANSFORMATION (APACHE PIG, HIVE)
  - N-GRAMS, PERPLEXITY MODEL, LOGARITHMIC PROBABILITY, SMOOTHING TECHNIQUES …
- **PROCESS MINING AND VERIFICATION**
  - EVENTS IN FIXED TIME ORDERS WITH PETRI NET THEORY AND MACHINE LEARNING
  - RESULTS ARE ENCOURAGED, HOWEVER MORE DATA AND TIME IS NECESSARY
- **INTRUSION DETECTION SYSTEM (IDS)**
  - SUCCESSFULLY APPLIED FOR DDOS (DECODERS, RULES)
  - OFFLINE TESTING DUE TO TIME CONSTRAINTS
  - PIONEERING DEPLOYMENT FOR MALICIOUS BEHAVIOR DETECTION IN COLLABORATION WITH IDS

MACHINE LEARNING AND DATA ANALYTICAL TOOLS

- **VOWPAL WABBIT**: JOHN LANGFORD, MICROSOFT/YAHOO! RESEARCH
  - OUT-OF-CORE FAST ONLINE LEARNING: TERA-FEATURE (10^{12}) DATASET ON 1000 NODES IN 1H
  - THE HASHING TRICK 32-BIT MURMURHASH3
  - EXPLOITING MULTI-CORE CPU: PARSING OF INPUT AND LEARNING IN SEPARATE THREADS
  - COMPILED C++ CODE
- **SCIKIT-LEARN**: DAVID COURNAPEAU, INRIA (TELECOM PARISTECH) AND GOOGLE
  - WELL-MAINTAINED AND POPULAR ML TOOL, COMPREHENSIVE ALGORITHM LIBRARY
  - SINCE APRIL 2016, IT IS PROVIDED IN JOINTLY-DEVELOPED ANACONDA FOR CLOUDERA ON HADOOP
- **PYTHON, ANACONDA, NUMPY, SCIPY, PANDAS, MATPLOTLIB …**

INFRASTRUCTURES

- **HADOOP CLUSTER** - HP BLADE SYSTEM
  - 1X SERVER AND 11X CLIENT 11TB HDD, NODE SPECIFICATION: 2X INTEL® XEON® PROCESSOR E5-2620 (15M CACHE, 2.00 GHZ, 7.20 GT/S INTEL® QPI, 6X CORES, 12X THREADS) + HYPERX THREADING (24 SIMULTANEOUS TASKS PER CLIENT), 32GB RAM, 1TB HDD
- **HPC CLUSTER –** IBM SYSTEM
  - 52X IBM DX360 M3, 8X IBM DX360 M4 2X NVIDIA TESLA K20, 2X IBM DX360 M3 NVIDIA TESLA M2070, 2X X3650 M3 MANAGING SERVERS, 4X X3650 M3 DATA-MANAGING SERVERS, X3550 M4 SERVER, INFINIBAND 2X 40 GBPS (IN 52+2+2+4 NODES), 2X DS3512 WITH 72TB DISKS

RESEARCH DIRECTIONS

- DATA MINING AND MACHINE LEARNING FOR BIG DATA (VOLUME, **VELOCITY**, **VARIETY**, …)
  - USER BEHAVIORS, MONITORING SYSTEM, CYBER-SECURITY
  - TEXT, IMAGE AND SPEECH RECOGNITIONS
- RESEARCH DIRECTIONS TOGETHER WITH
  - ACCELERATED PROCESSING: GPU, APACHE SPARK/STORM, HPC AND CLOUD
  - BIG DATA PROCESSING, SEARCHING, KNOWLEDGE AND RELATION DISCOVERY, GRAPH TRAVELLING
  - MOBILE COMPUTING, INTERNET OF THINGS
  - **BASED ON DOMAIN DEMANDS**

THANK YOU FOR YOUR ATTENTION

DETECTION OF POTENTIALLY DANGEROUS ACTIVITIES
FROM LOGS OF MOBILE DEVICES
USING MACHINE LEARNING TECHNIQUES

GIANG.UI@SAVBA.SK
DEPARTMENT OF PARALLEL AND DISTRIBUTED INFORMATION PROCESSING
INSTITUTE OF INFORMATICS
SLOVAK ACADEMY OF SCIENCES

Tempest
IT makes sense

IBM
IBM Slovakia