KPMG

Consultant's view on HEP analysis software

(By an ex-physicist turned data scientist)

Contents

About me - Max Baak

Background in HEP (2002-2015)

- **2007: PhD @ Nikhef, on research at BaBar experiment (SLAC)**
- **2007 – 2015: ATLAS experiment**
  - 2008 – 2015: research fellow then staff at CERN
  - DQ, SUSY searches, Gfitter, statistics
- **Soft spot for analysis tools. Well-educated in ROOT (RooFit).**

Current affiliation

- **Consultant @ KPMG, since March 2015.**
- **In Big Data & Advanced Analytics team**
- **Title: “Chief data scientist”**

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (‘KPMG International’), a Swiss entity. All rights reserved.

Big Data @ KPMG

The company

Accounting (audit)

Consulting

- Big Data & Adv. Analytics team
  - ➢ Offer data and analytics services to organisations.

The Big Data team

Team consists of:

- ~20 people
- data scientists,
  data engineers / architects, and
  software developers
- backgrounds in physics and computer science
- many ex-Nikhef / CERN

The work

Work at clients:

- Advice companies on setting up data / analytics strategy
- Convert data from various sources into *insights* to drive better decision-making.
- Often requires processing of large data sets and statistical analysis.
  - ➢ Installation of cluster
  - ➢ Onboard data
  - ➢ “Proof of concept” analyses
  - ➢ Put analyses into production

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (‘KPMG International’), a Swiss entity. All rights reserved.

Credit card fraud detection and prevention

Background

A large international payment service provider that processes terminal and acquiring transactions for different debit and credit card schemes, asked KPMG to help set up a data lake and build advanced analytics algorithms to reduce fraud & dispute losses for issuing and acquiring customers.

Requirements and challenges

- Analyse >4 billion transactions (2 years) and >800,000 cases of fraud
- Enable Big Data capability build-up at client's own data scientists team

Our solution

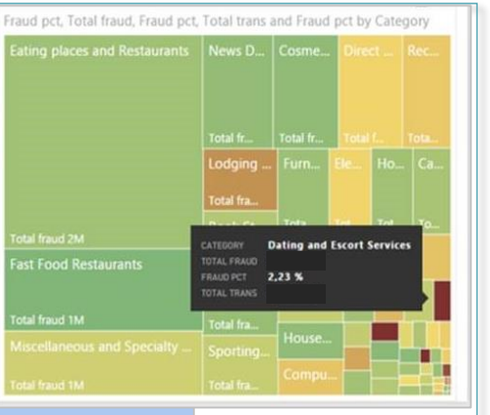
- Combine large (100 TB) data sets in a Hadoop cluster to build a complete and continuous fraud detection model (Common Point of Purchase, CPP; Common Point of Fraud, CPF and probability of fraud per card)
- Apply Machine Learning through KPMG's open source ESKAPADE framework, to identify fraud patterns at merchant, terminal & card-level
- Rank 600 existing fraud rules & select features for automated rule generation

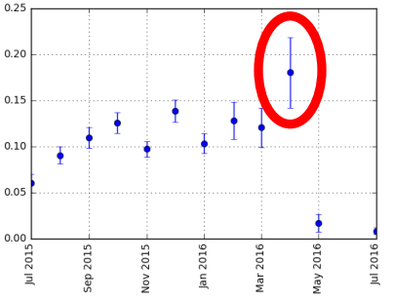
Business value

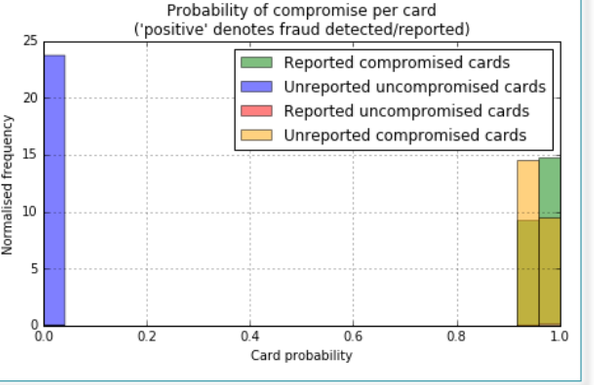
- Quick-win: fraud-reducing initiatives have already been enacted during the Proof-of-Concept phase.
- Increase in fraud detection by factor of two, with an estimated initial reduction of 10 million EUR in customer fraud losses per year.
- Reduce churn of large corporate clients due to improved service.

Card Probability Analysis

Which cards are most likely to be stolen considering the locations they passed?







Customer store card use: patterns, modelling, and reporting

Background

- Large multinational chain of retail stores.
- Over 80 million gift and loyalty cards distributed to customers globally.
- Client's live database system could not provide the insights required to declare liability correctly for profit and loss calculation.
- Large group of customers will never spend some portion of their open balance.

Requirements / Challenges

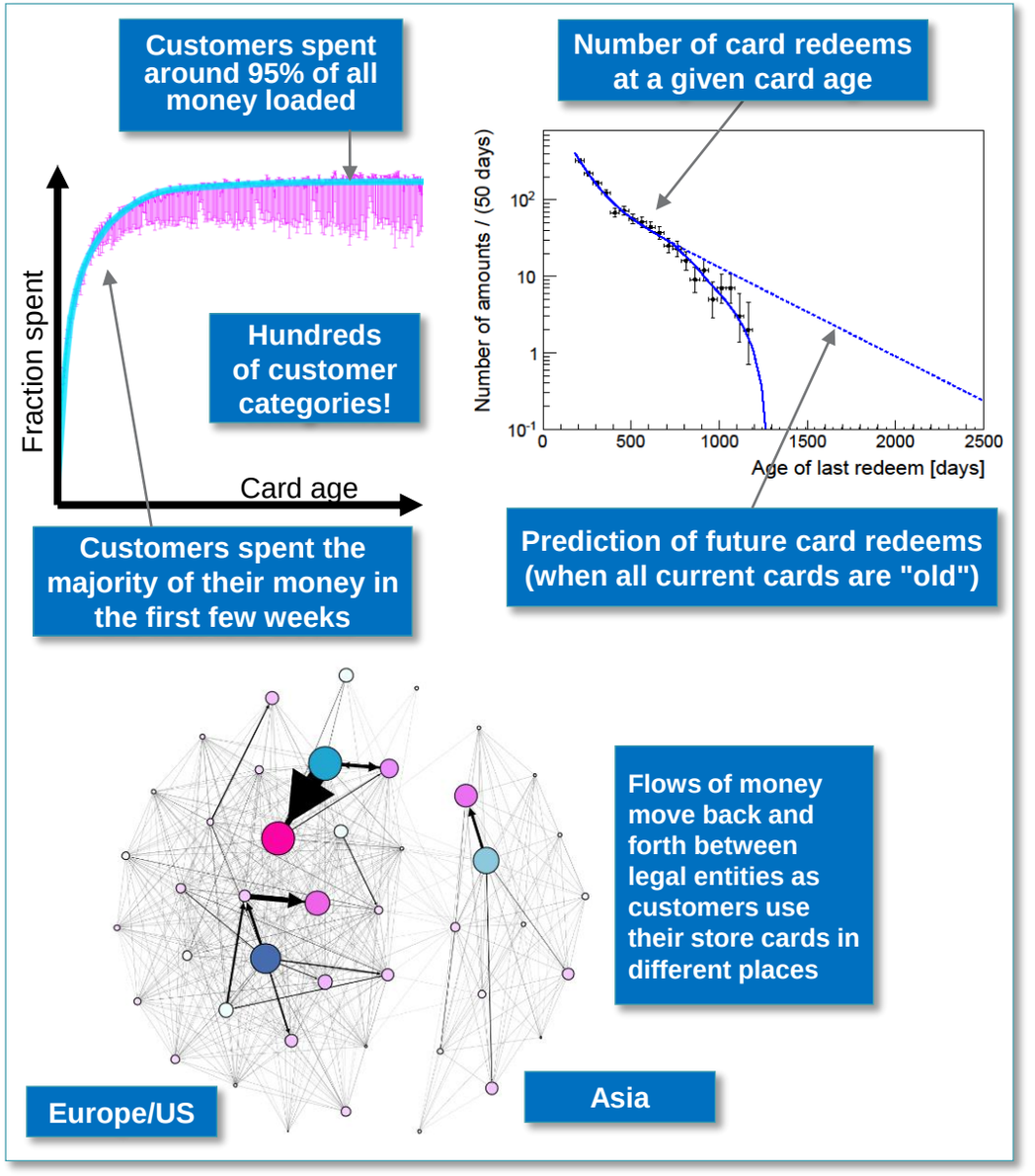
- Fifty different legal entities across several countries used the customer cards in different ways.
- Different groups of customers used cards in different ways.

Solution

- Discovered several factors with a significant influence on customer spending behavior.
- Categorized customers to produce a stable and accurate predictive model with known and quantified uncertainties.

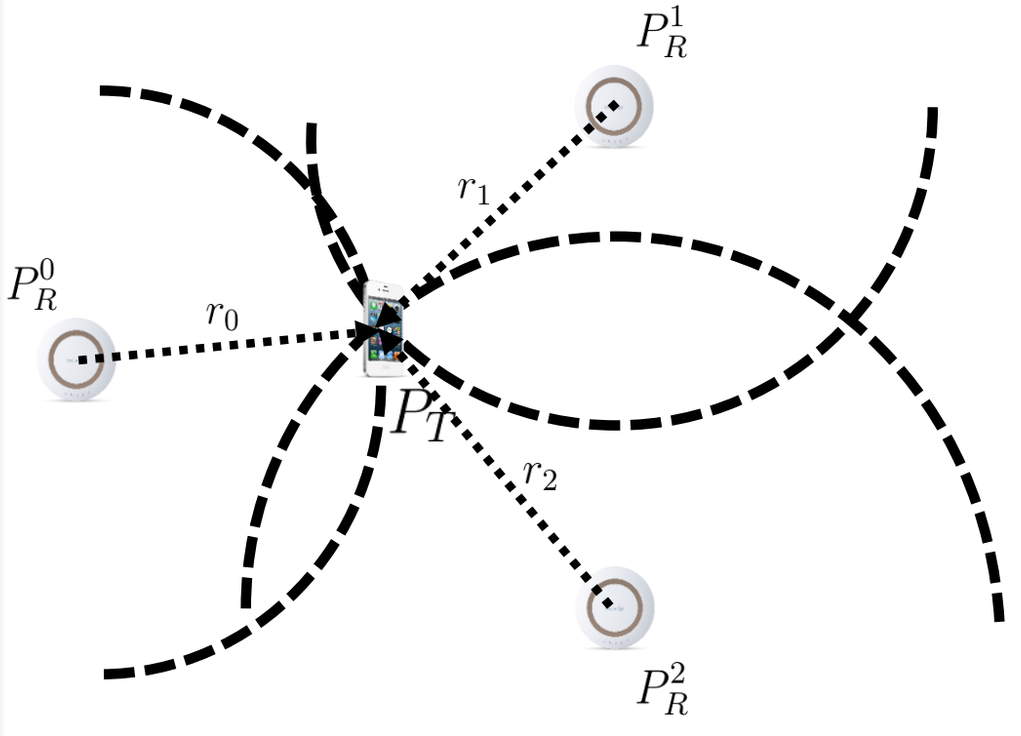
Benefits

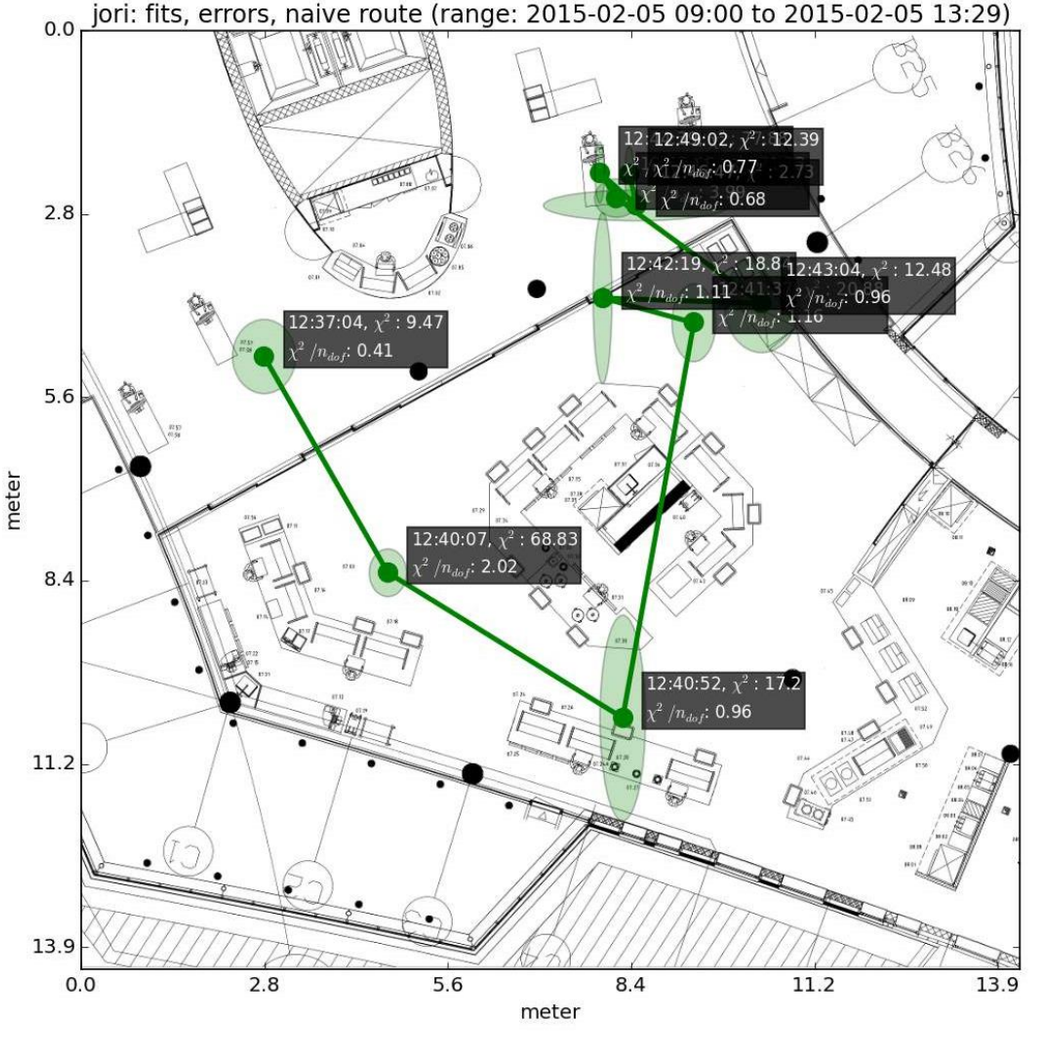
- Complete and accurate reporting of more than 100 M € open balance.
- Evaluation and accounting of cross-legal-entity spending.
- Accurate predictions of the amounts that will not be spent in the future



© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

Wifi tracking: tracking a smart phone





© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

Wifi tracking

Business Need

An event organizer wants to measure crowd density and flow in order to improve the safety of event visitors, optimize on security personnel utilization and improve general visitor experience. Furthermore, they want to interact with visitors during the event.

Solution in use

LAS is used for visualizing crowd density and movements in (near) real-time. Using this information, certain entrances might be closed or opened temporarily, or personnel might be instructed to guide the public to a certain area.

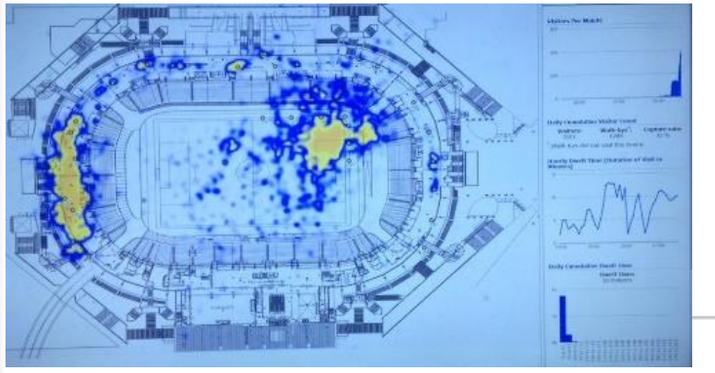
By using a mobile app and beacons, custom messages can be sent to visitors on the program and upcoming activities or by vendors with specific promotions.

Products & Services

- Real-time messaging
- Real-time analysis (i.e. heatmap)
- Qlik Sense dashboards for analysis of historical and real-time data and integration with beacon functionality

Benefits

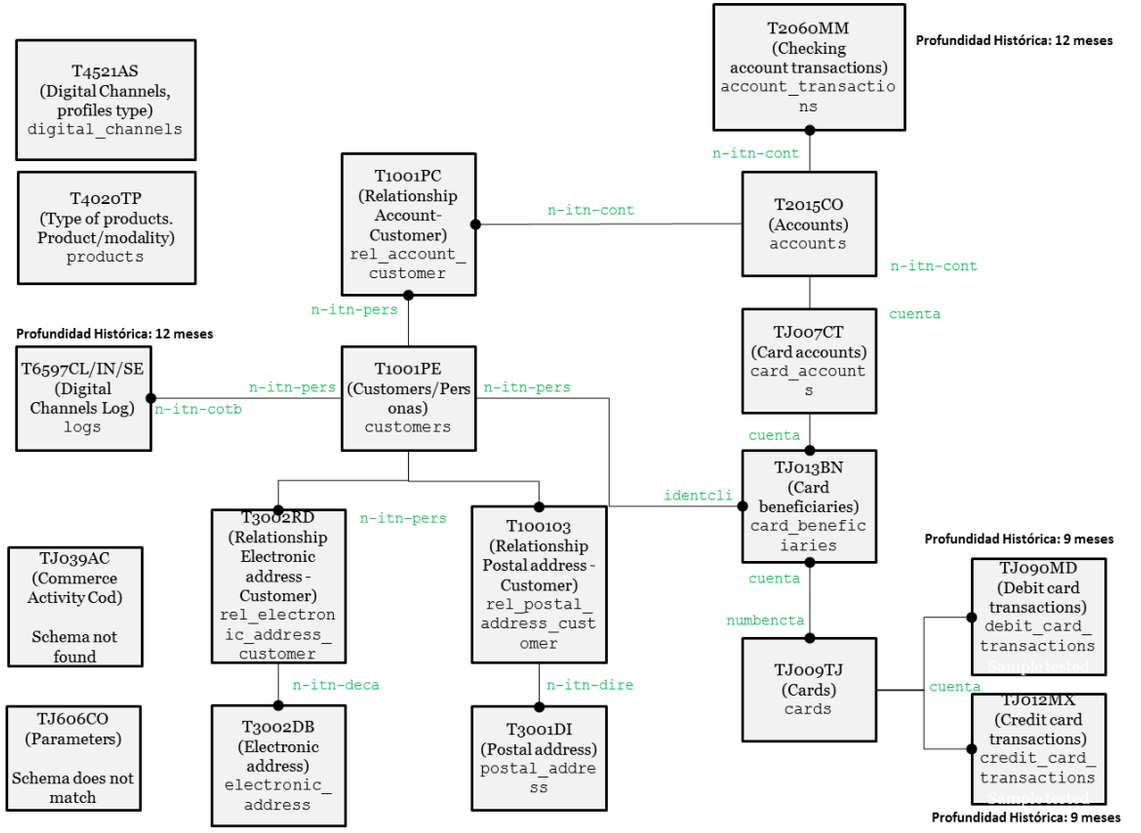
- Increased safety of event visitors
- Better utilization of resources (i.e. event venue and personnel)
- Real-time interaction with visitors to increase customer experience
- Promotions by vendors allow for higher fees for promotional activities



© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

Typical datasets

- **Quite different from HEP**
  - Typically no filtering on data.
- **Relational data**
  - No concept of “event” that packages all information related to one record.
- **Structureless data**
- **Security is big issue**
  - Often personal information
- **Categorical variables**
  - Cannot be logically ordered

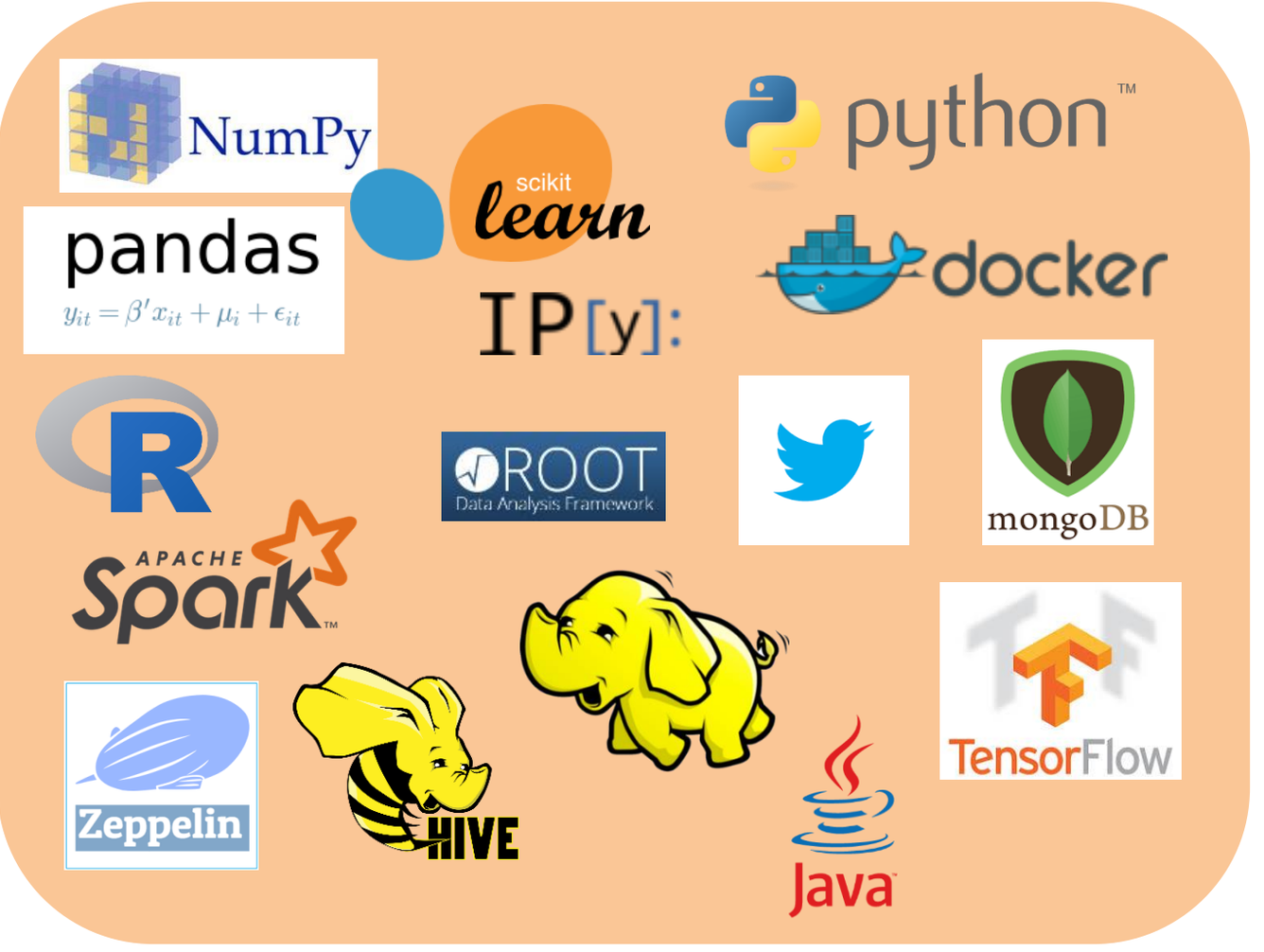


© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (‘KPMG International’), a Swiss entity. All rights reserved.

Our analysis tool stack ("KAVE")

All open source – no need to buy any analysis software

- **Our projects are very diverse.**
- **Each project typically requires some new (analysis-) software package(s).**
- **Before we build something new, check if it already exists.**
  - It always does!



© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

My impressions of (other) data analysis tools

- Very impressed by open-source community.

- Many many great data processing & analysis tools out there.

  - For all your needs: databases, persistence, data processing, data analysis.

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

KPMG

Questionnaire

Most popular analysis toolkit used in business?

A. SAS

B. R

C. Python scientific libraries

D. Spark

E. ROOT

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

What is the best programming language to learn to get a data science job?

A. Python

B. Java

C. R

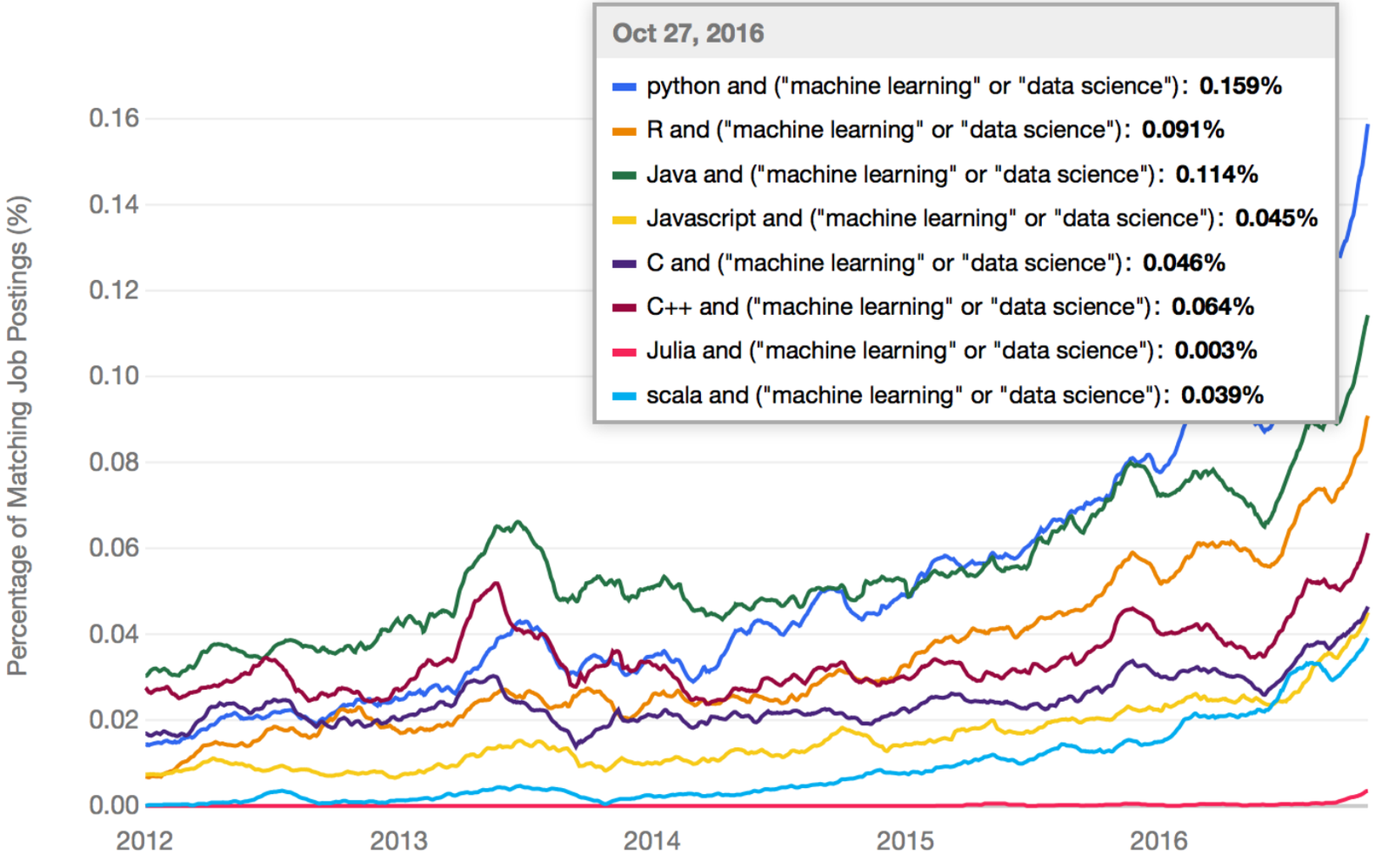
D. C / C++

E. Scala

F. Julia

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

What programming language to learn to get a data science job?



From KDNuggets

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

In how many of our past 25 projects do you think we have used ROOT?

A. 0 x

B. 1 x

C. [2, 5] x

D. [6, 10] x

E. > 10 x

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

KPMG

Vision

Impressions of HEP & open-source community

- CERN has a great name. Fantastic reference!
- High-energy physicists are considered the smartest people in the world.
- Great pity to see that (all intellectual effort put into) HEP analysis software is barely picked up outside of HEP community.
  - (Why is that?)
- Would be if great if HEP s/w contributions would make the same impact as reputation of the field!

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

Retrospective on ROOT

Law of the handicap of a head start :

“theory that suggests that an initial head start in a given area may result in a handicap in the long term.”

- ROOT has been state-of-the-art for a long time, but …
- … thanks to rapid development in open-source community in past 10 yrs …
- … has by now been taken over in multiple areas ...
- … community

Q: If you could start from scratch, would there still be a need to develop ROOT?

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (‘KPMG International’), a Swiss entity. All rights reserved.

My vision (as an outsider)

- HEP community has an obligation to society to actively use and contribute to existing open-source tooling.
  - Instead of developing yourselves.

- Use existing open-source where you can, and contribute to it where it needs to be improved
  - *E.g. hadoop, spark, google protobuf (persistence), scientific python*
    - *Thanks to google!*
  - You can make a great impact!

- Nice to have: if HEP community could educate its PhD students & post-docs in analysis tools that are also used by the rest of the world.
  - Remember: 95+% leave the field …

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.

Tips for migration of ROOT

Goal of ROOT should be (imho): become the serious competitor of R !

1. Split up and make a light-weight version of ROOT
   - Leave only analysis components.
2. Replacements:
   - Persistence → protobuf, capn' proto, avro
   - Xrootd → spark
3. Actively support **& test** installing with: homebrew, conda (← personal request)
   - E.g. Installation together with anaconda is a real pain.
4. Make python support as important as c++
   - Accept contributions in python (and dependencies on numpy)
5. Actively support spin-off activities: sparkroot, protobuf, anaconda installation, rootpy.
   - They usually die once PhD student finishes …
   - Absorb them to make ROOT future proof!

KPMG

© 2016 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ('KPMG International'), a Swiss entity. All rights reserved.