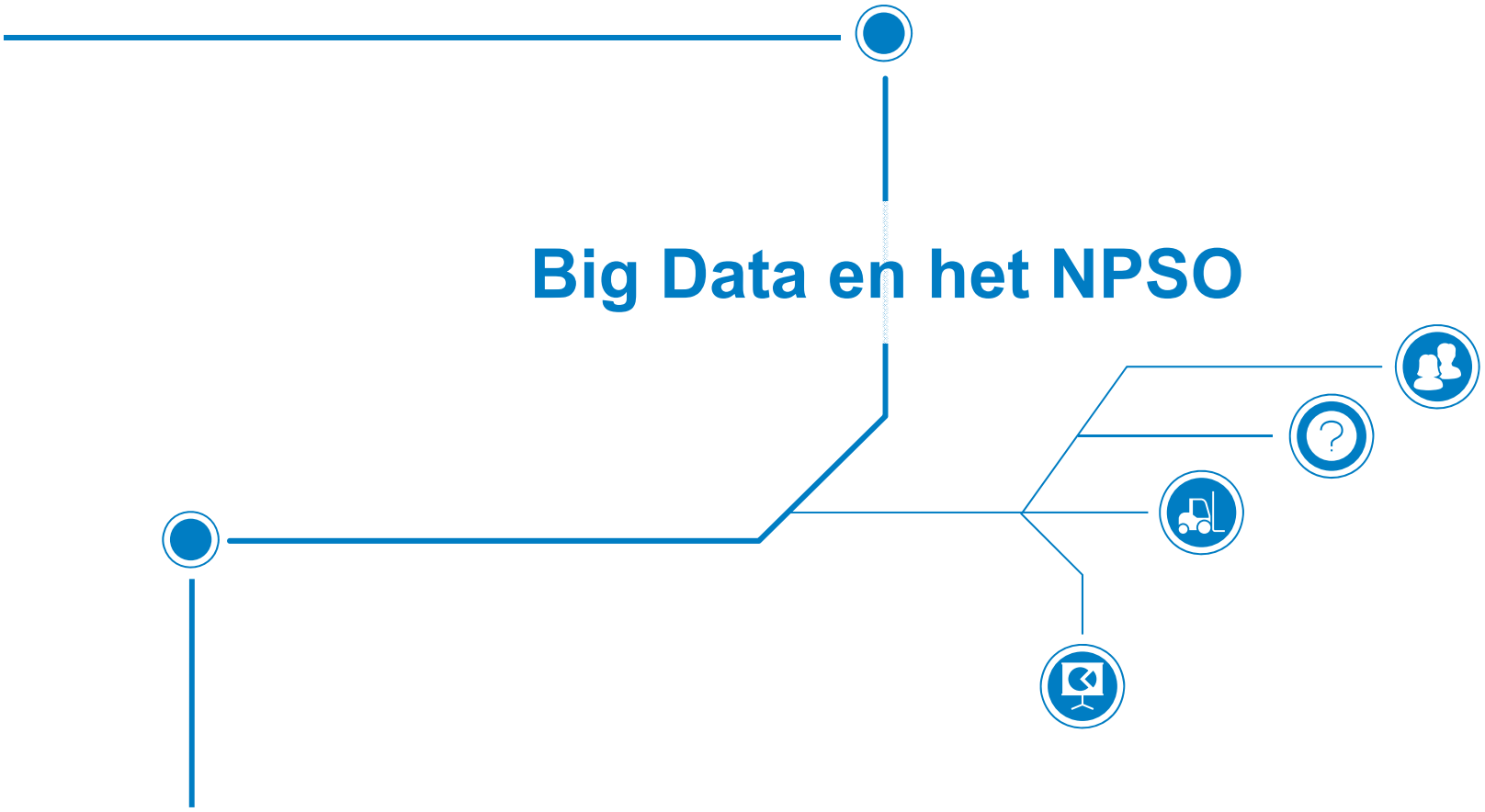
Big Data en Marktonderzoek: Opportuniteiten en gevaren

Dr. Istvan Hajnal, GfK België, NPSO Mei 2017

@IstvanHajnal

Big Data en het NPSO



NPSO

Volgens de website

- … Big data speelt in toenemende mate een rol in beleidsvorming en zeker in ook de survey wereld. Nieuwe toepassingen om gevoelens, opinies en meningen van personen te meten, zoals opinion mining, online sentiment expression, gebruik van Twitter, etc. worden in toenemende mate gebruikt.
- … vragen rondom het nut en de bruikbaarheid van het gebruik van big data.

- **Ambigue houding, die je ook terugvindt bij marktonderzoek**

Is dit terecht? En is dit altijd zo geweest?

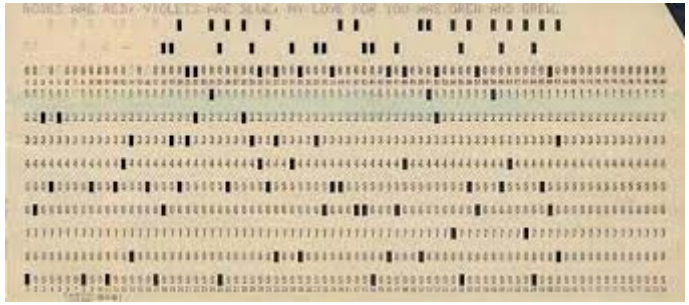
De “Heilige drievuldigheid” van het NPSO

De “Heilige drievuldigheid” van het NPSO

“Men spreekt van Big Data wanneer men werkt met een of meer datasets die te groot zijn om met reguliere databasemanagementsystemen onderhouden te worden”


Official
Statistics
Tax Incentive
NOTE !!

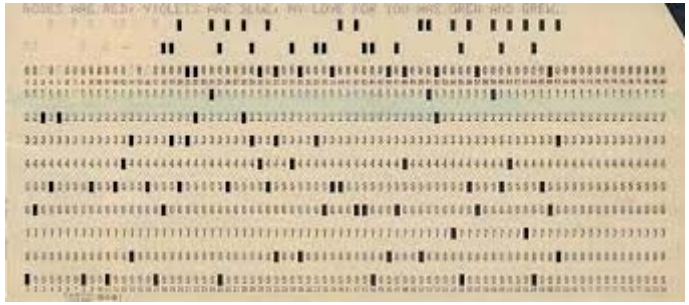
Herman Hollerith



(bron: wikipedia)

- 1860-1929
- De eerste baan van Hollerith was bij het bureau dat de volkstelling in de VS van 1880 moest uitvoeren
- Ontwikkelde een machine die het met behulp van ponskaarten mogelijk maakte grote hoeveelheden gegevens geautomatiseerd te verwerken.
- Hollerith heeft de Tabulating Machine Company opgericht
  - Later werd dit de Computer Tabulating Recording Company
  - En nog later werd de naam veranderd in …. International Business Machines Corporation (IBM).

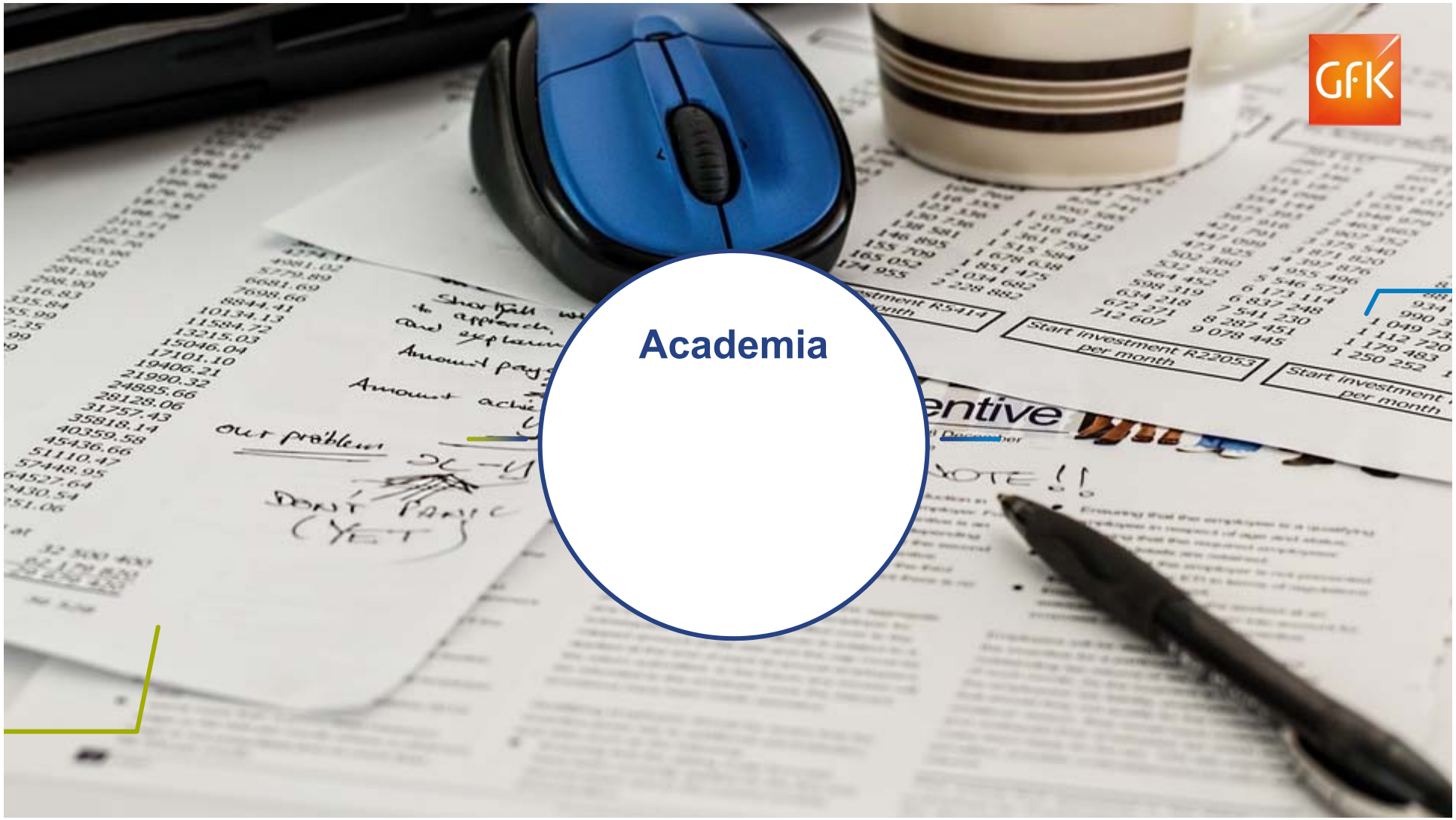
Herman Hollerith



(bron: wikipedia)

- 1860-1929
- De eerste baan van Hollerith was bij het bureau dat de volkstelling in de VS van 1880 moest uitvoeren
- Ontwikkelde een machine die het met behulp van ponskaarten mogelijk maakte grote hoeveelheden gegevens geautomatiseerd te verwerken.
- Hollerith heeft de Tabulating Machine Company opgericht
  - Later werd dit de Computer Tabulating Recording Company
  - En nog later werd de naam veranderd in …. International Business Machines Corporation (IBM).

Conclusie: De IBM Mainframe, de PC, Watson, e.d. komen, met wat goede wil, van een spin-off van Official Statistics


Academia

Academia

Alan Turing

- Computer pionier
- De Turingmachine
- Ontcijfering Enigma
- Princeton – Cambridge - Manchester

Academia

Geoffrey Hinton

- 1947-
- Cognitieve psycholoog en computerwetenschapper
- Universiteit van Toronto
- Expert in neurale netwerken
- Grote invloed op artificiële intelligentie en “deep learning”
- En dus op de verwerking van Big Data


Markt
Onderzoek
bureaus

Marktonderzoeksbureaus

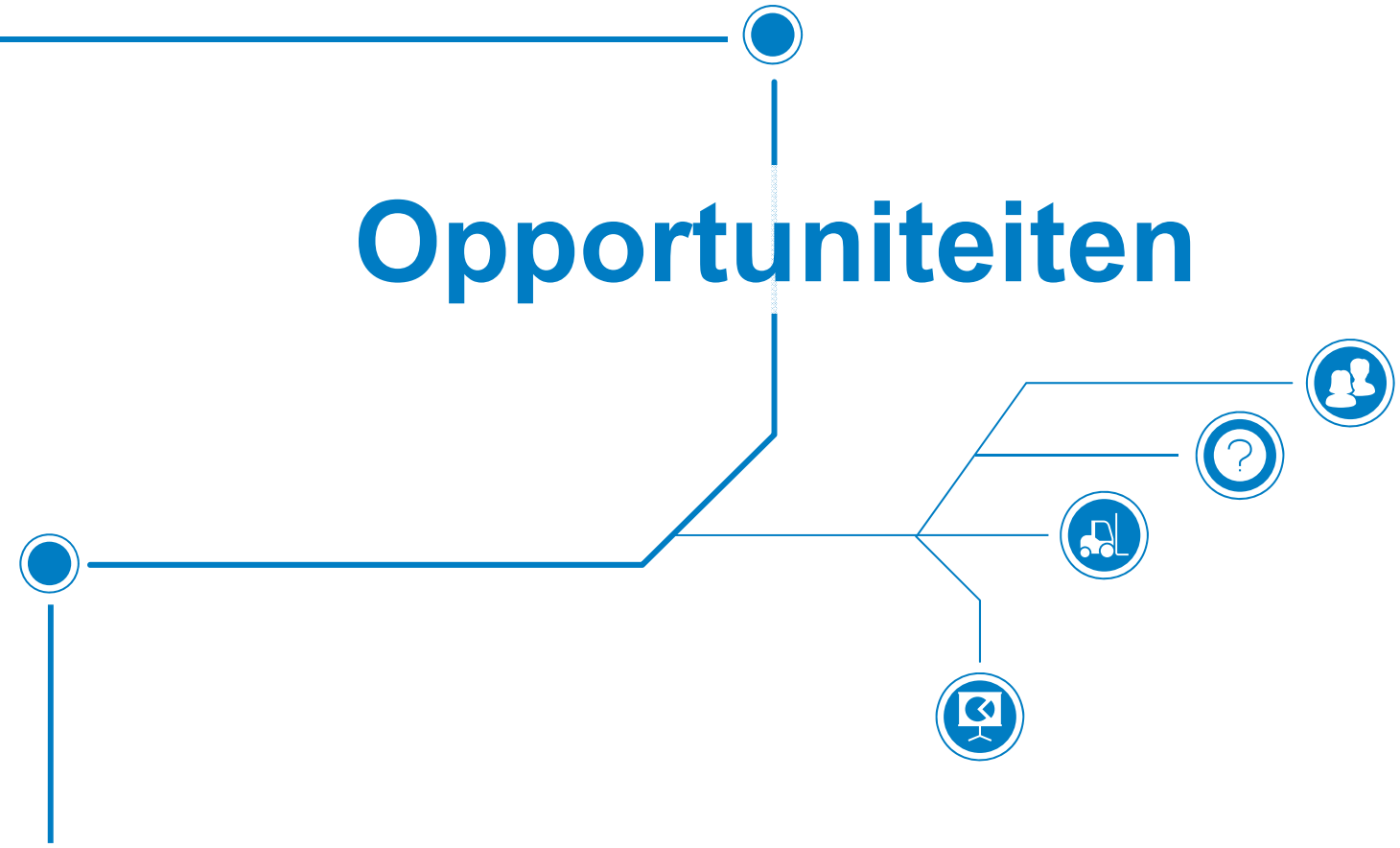
Weinig gekend feiten:

- Het eerste privé-bedrijf dat ooit een computer bestelde was een marktonderzoekbureau!
- In de jaren 60 waren marktonderzoekers pionieers in de eerste programmeertalen

De “Heilige drievuldigheid” van het NPSO

In het verleden heeft deze “Heilige drievuldigheid” korter op de bal gespeeld dan vandaag. We zijn “laggers” geworden.

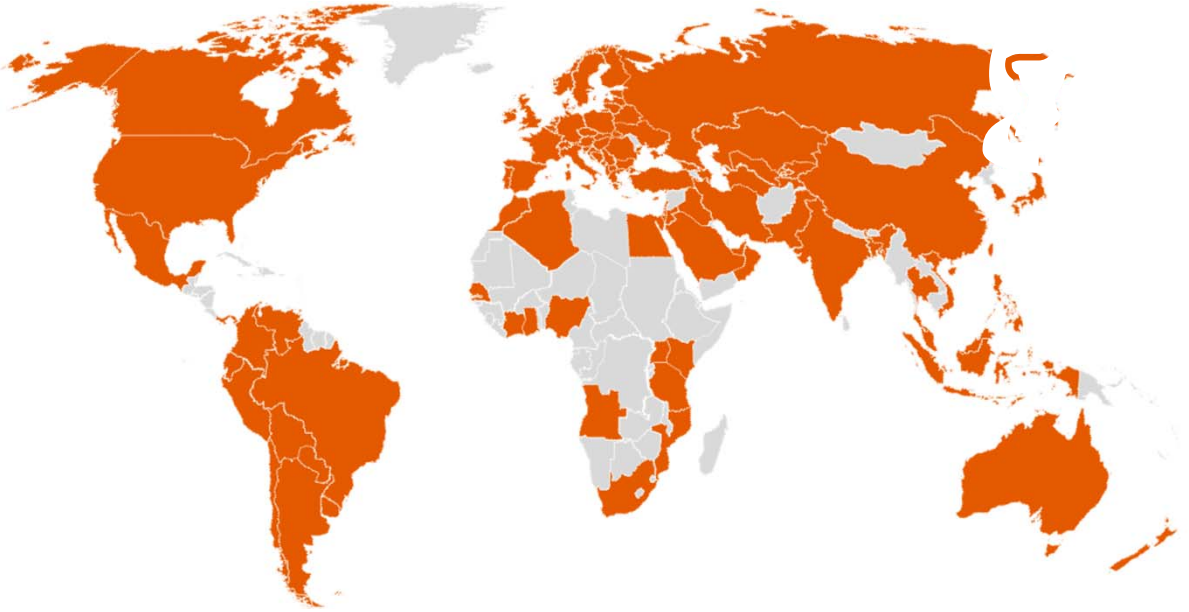
Opportuniteiten



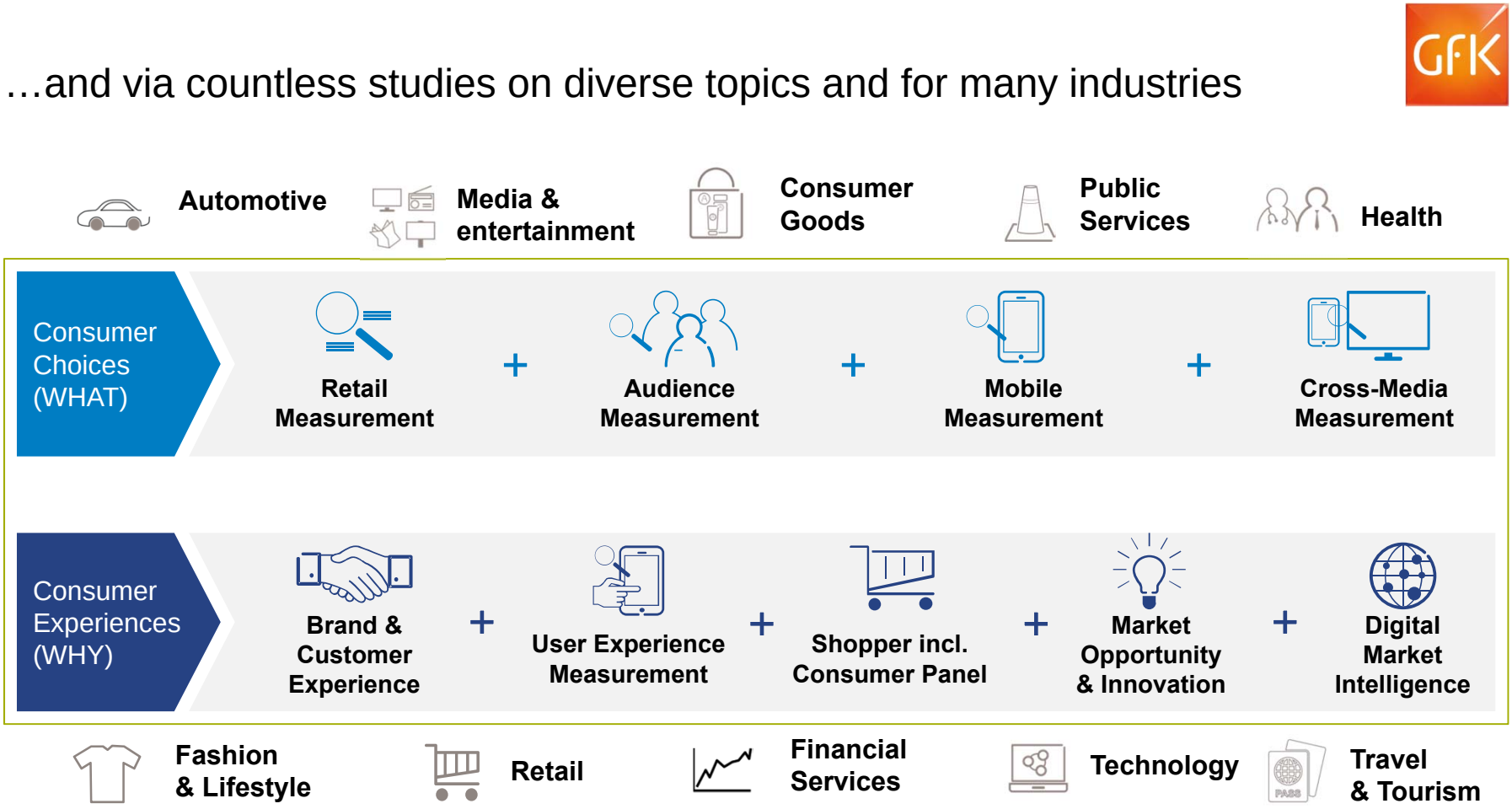
Big Data & Data Science @GfK

Taken from “How to Surf a Data Lake without Drowning?” a presentation by Ralph Wirth and Anna Machens on Predictive Analytics World London, 29th October 2015

As a global market research company…



…and via countless studies on diverse topics and for many industries



In an ideal world, Data Scientists could easily identify all relevant data assets for specific research questions and directly access them

Example: Potential predictors in GfK data for sales of Jacobs filter coffee

…

Past purchases of Jacobs and competitive brands (Consumer Panel)

Demographics, values, attitudes, lifestyles of Jacobs and competitive brand buyers (Consumer Panel)

Media usage of Jacobs and competitive brand buyers (Crossmedia Link Panel)

Sales of filter coffee machines vs. other systems like capsule machines (Retail Panels)

Jacobs brand KPIs (Brand tracker)

Jacobs advertising awareness (Advertising tracker)

Consumer price sensitivity and price trends (Consumer Panel, ad hoc studies)

Regional purchase power (GeoMarketing data)

Relevant coffee-related trends (Roper Trends database)

Customer Satisfaction (Satisfaction trackers)

Social Media Momentum & current "hot topics" (SMI study data base)

User Experience of Online Shops that sell Jacobs coffee (User Experience database)

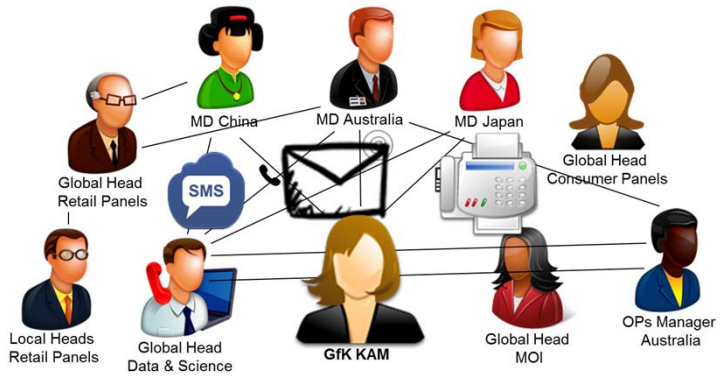
Evaluation of Jacobs' and competitors' new products (product test database)

With a multi-national, heterogeneous data landscape, this is not a given

Data storage is not fully integrated, no metadata taxonomy covers all data facets

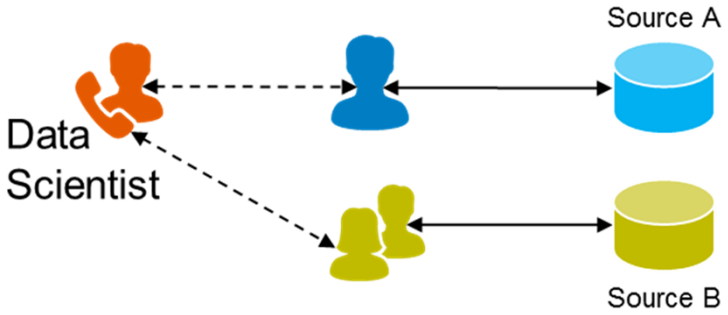
Imperfect data transparency

In some situations, a lot of communication is needed to find out what suitable data we have for a specific research problem.

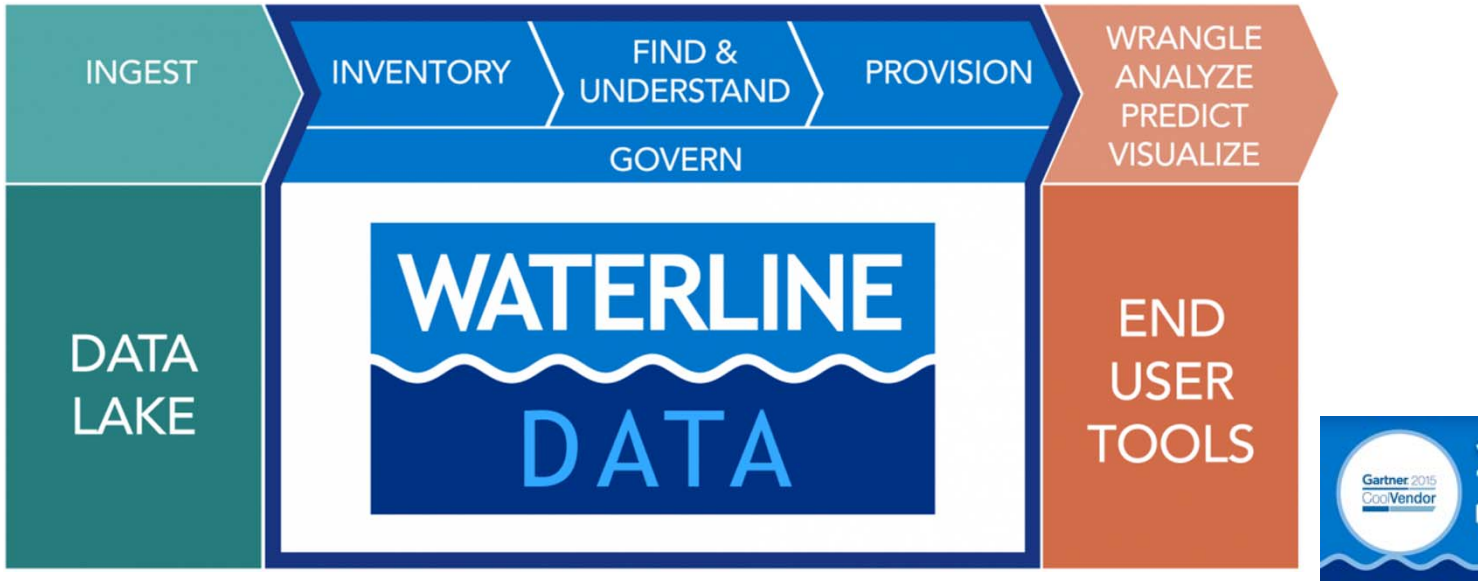


No direct access to full data landscape

Some data is stored in silos, so that access has to be handled by the respective data managers.

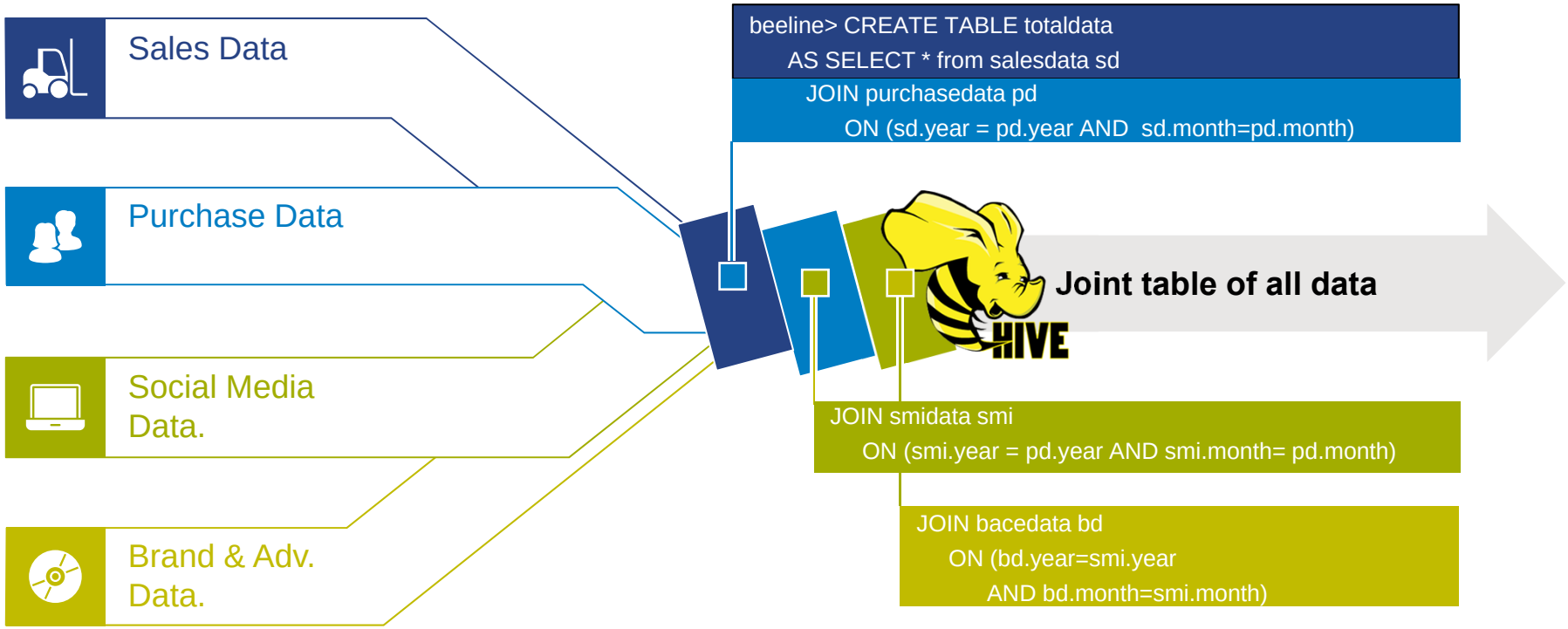


We test and use metadata management tools to find relevant data and get a first understanding of data sets



Hive is used to aggregate and integrate data from different sources with regard to the time dimension

Integration



Big Data & Data Science @GfK Belgium

A sample of cases where the Marketing and Data Sciences team of GfK Belgium was involved in

Statistics and Methodology

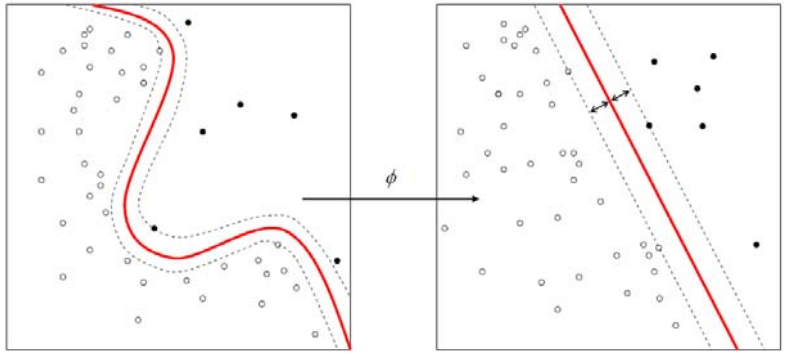
Association Rules Mining

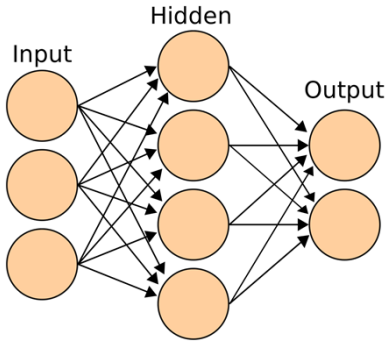
- **Association Rules Mining** is a method for discovering interesting relations between variables by identifying strong association rules in large databases.
- Association rules were introduced for discovering regularities between products in point-of-sale systems in supermarkets.
- We have applied this on Shopper data to try an identify presence of beer in a shopping basket

Statistics and Methodology

Machine Learning

- **Machine Learning** deals with algorithms that can learn from and make predictions on data. While ML is a subfield of computer science it is heavily influenced by statistics.
- Examples of such techniques are:
  - Ridge + Lasso
  - Random Forests
  - Support Vector Machines
  - AdaBoost
  - …
- Applied on predicting presence of certain product categories in shopping baskets

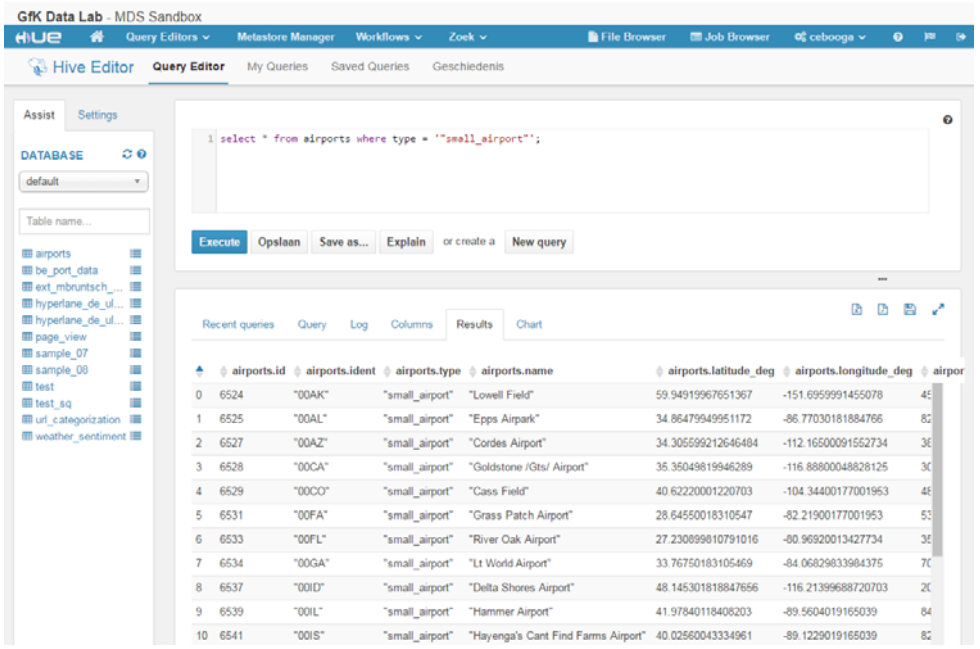




Data handling

Analysis on large GfK databases

- Analyses on Media Measurement TAM data
  - Ongoing proof-of-concept
  - Using the Big Data infrastructure of the DataLab



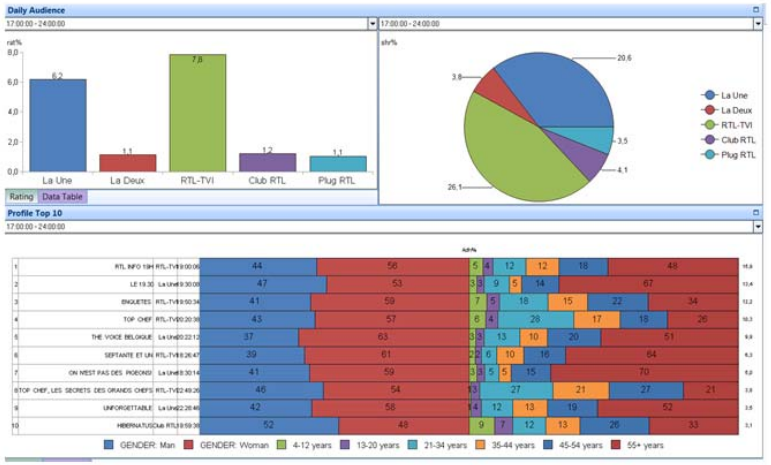
Data handling

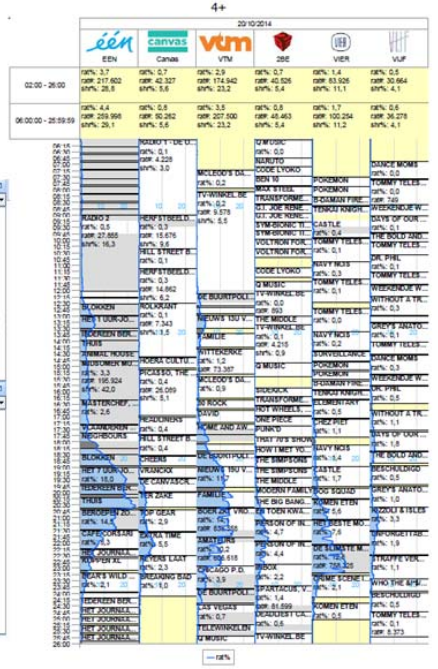
Integration of GfK data sources

- **Data Fusion**
- In cooperation with Germany
- Example: tgo.be: Data Fusion of Media Measurement (TAM) data and Shopper (CP) data
- Not available yet

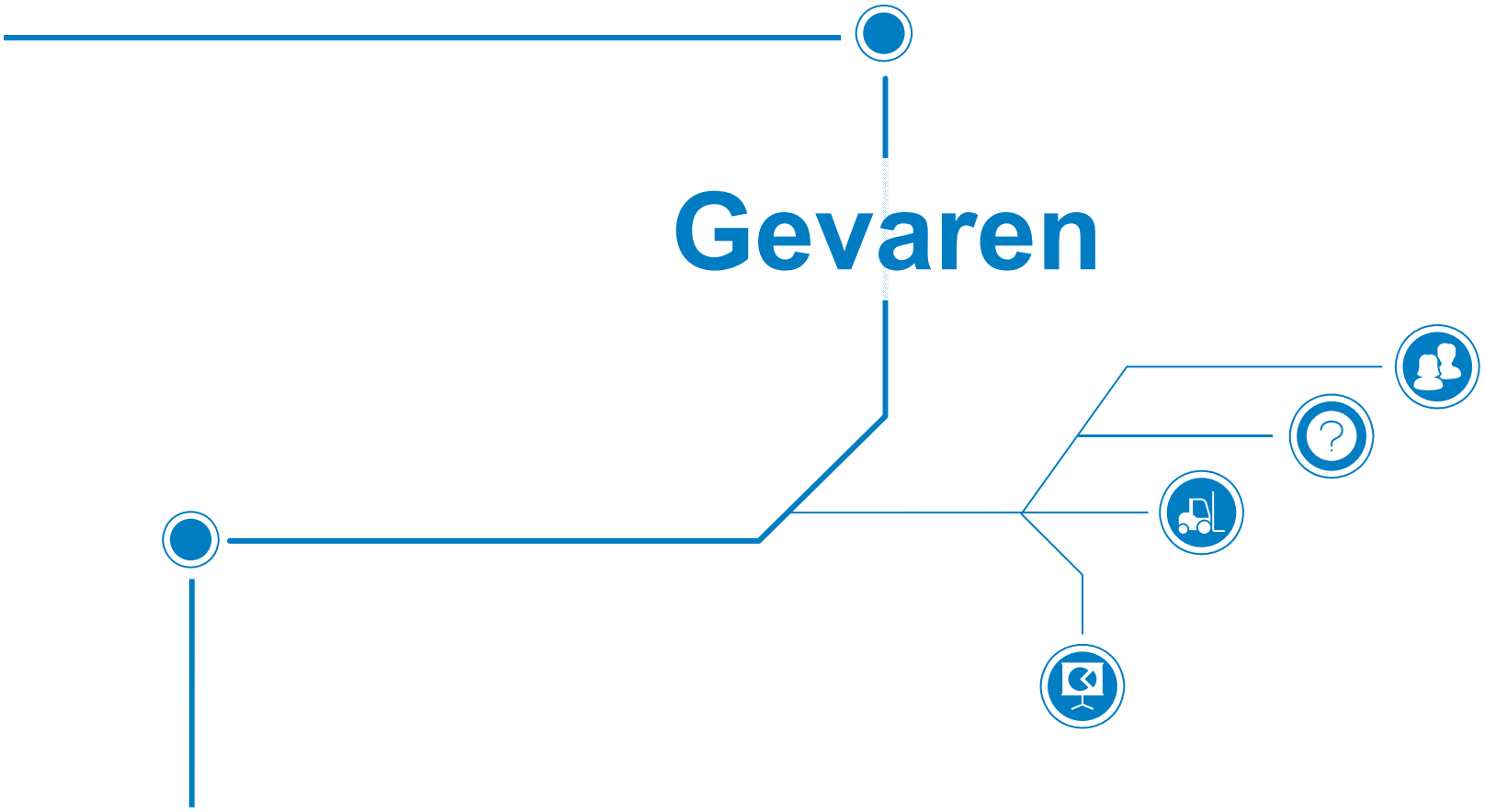
MAP software

Example output





Gevaren



Gevaren

In het kort … Bias, Bias & Bias

Maar eerst een opmerking vooraf

IH op IIEX 2016

I'm not buying the zillions of datapoints argument: it's not because there is a lot of water in the ocean that you need to drink it

1936 US Presidential Elections

Franklin Roosevelt
(Democrat)

Alf Landon
(Republican)

- Literary Digest
  - Sample size: 2.4 MILLION
  - Prediction: Roosevelt 43% : Landon 57%
- George Gallup
  - Sample size: 50,000
  - Prediction: Roosevelt 56% : Landon 44%

Steekproefgrootte is belangrijk

Representativiteit nog meer

Today
Election Forecast

AMERICA SPEAKS
THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday
The Election in Review

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

1—The American Institute of Public Opinion predicts the re-election of Franklin D. Roosevelt and John N. Garner.

* * *

2—The Institute's latest presidential poll indicates that Roosevelt will receive approximately 56% of the major party vote (minor parties eliminated), to 44% for Alfred M. Landon and Frank Knox. In 1932 the President received 59.1% of the major party vote.

* * *

3—With minor parties included, President Roosevelt's percentage of the total popular vote will be approximately 54%, to 46% for Landon.

* * *

4—The President will receive a minimum of 315 electoral votes. The number necessary to win is 266. Should last-minute shifts in the group of states where the race is nip-and-tuck give this entire group to Roosevelt, he would receive more electoral votes than in 1932, when he polled 472.

* * *

5—William Lemke, candidate of the Union Party, will poll fewer than 1,500,000 popular votes, and carry no state.

* * *

6—Norman Thomas, Socialist candidate, will poll about half as many votes as in 1932, when he received 884,000.

Election Will Test Clashing Poll Methods

Geldt zeker ook voor Big Data

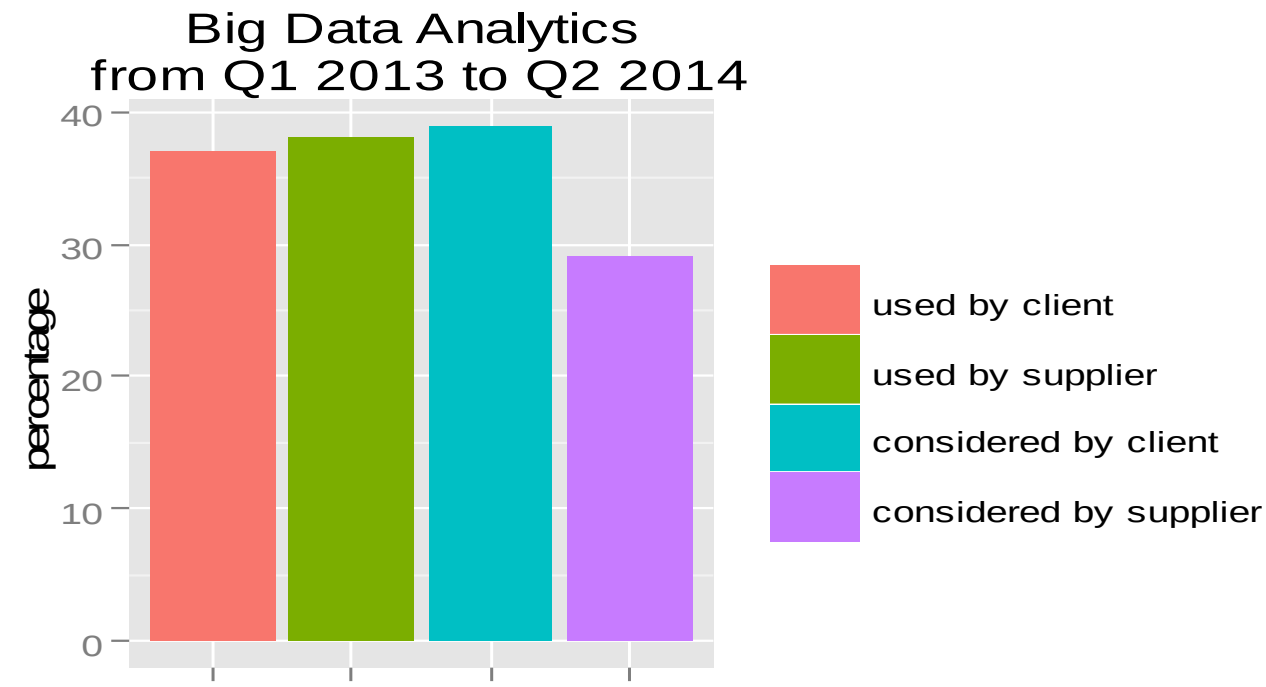
Nog vaak verkeerd begrepen bij Big Data gebruikers

“Of course samples can and are biased as well. But there is a difference: Samples are constructed specifically with a research question in mind, and often are designed to be unbiased. Big data or other sources of data are often created for other reasons than research questions. As a consequence big data might have some disadvantages that are not offset by its bigger size”

Zie ook http://allthingsdatascience.blogspot.be/2016/12/small-samples-versus-alternative-big.html over een discussie rond kijkcijfers in België

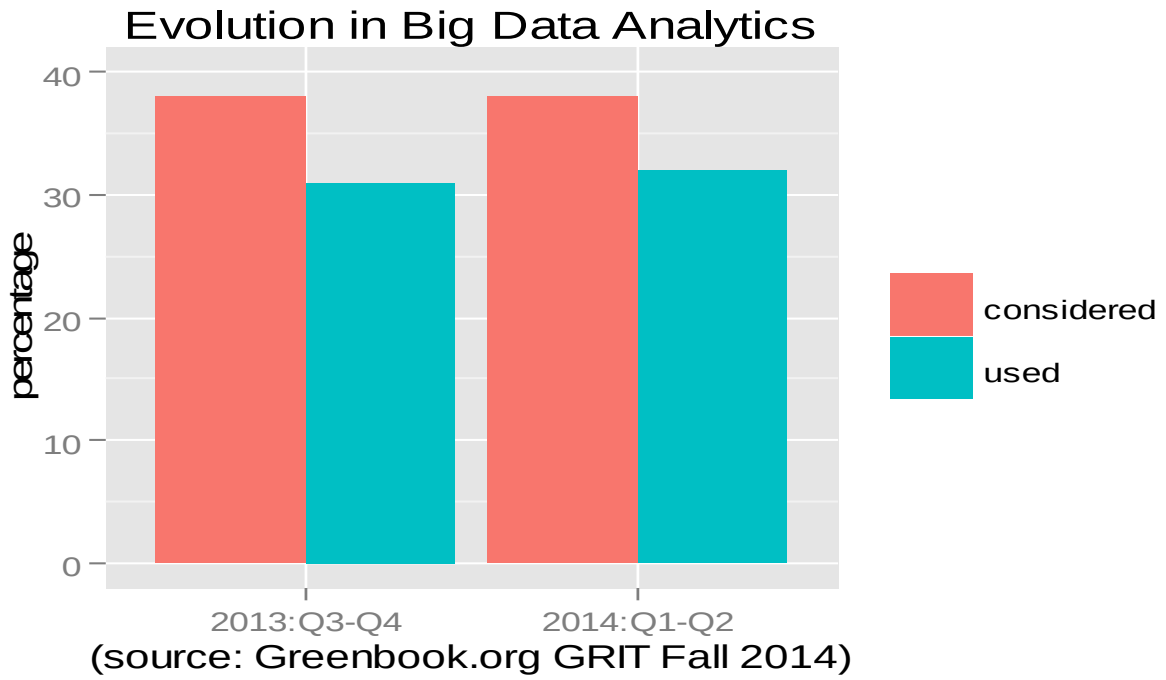
Conclusies voor marktonderzoek

Big Data: today



(source: Greenbook.org GRIT Fall 2014)

Big Data: evolution



(source: Greenbook.org GRIT Fall 2014)

Big Data fatigue?

Mijn conclusies

IIEX 2015

MRX seems to be bipolar when it comes to Big Data

My advice to suppliers of Market Research:

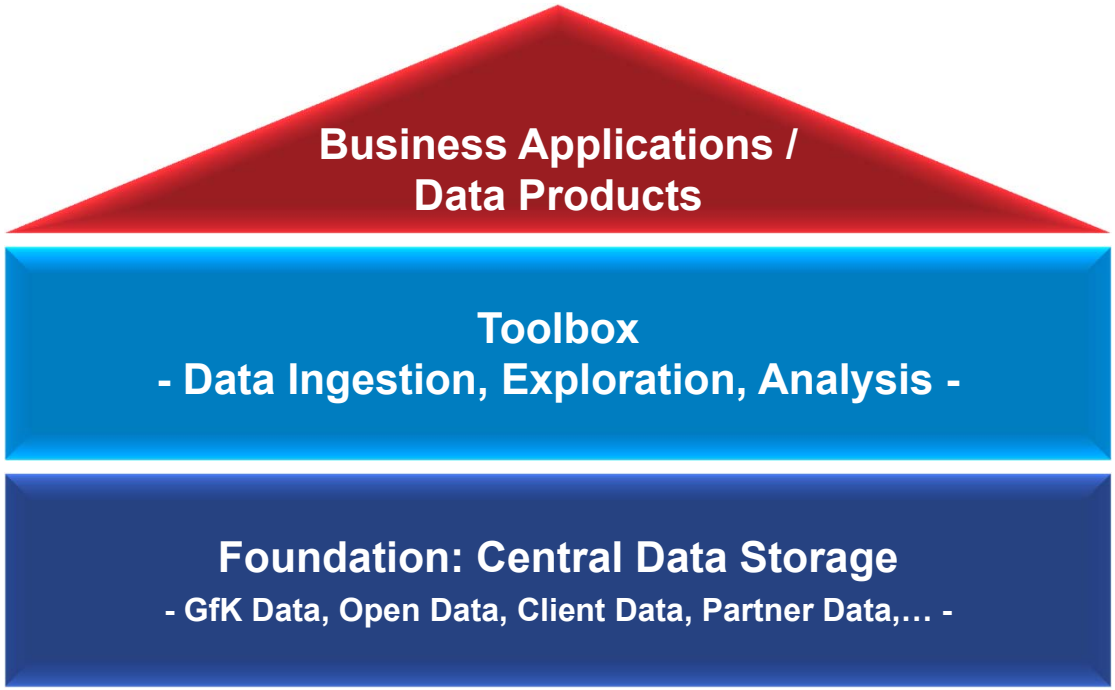
- Don’t look at Big Data as just a fad or hype
- Don’t look at Big Data as a threat to Market Research
- But embrace it as a new (business) reality
- Learn how to process large amounts of data

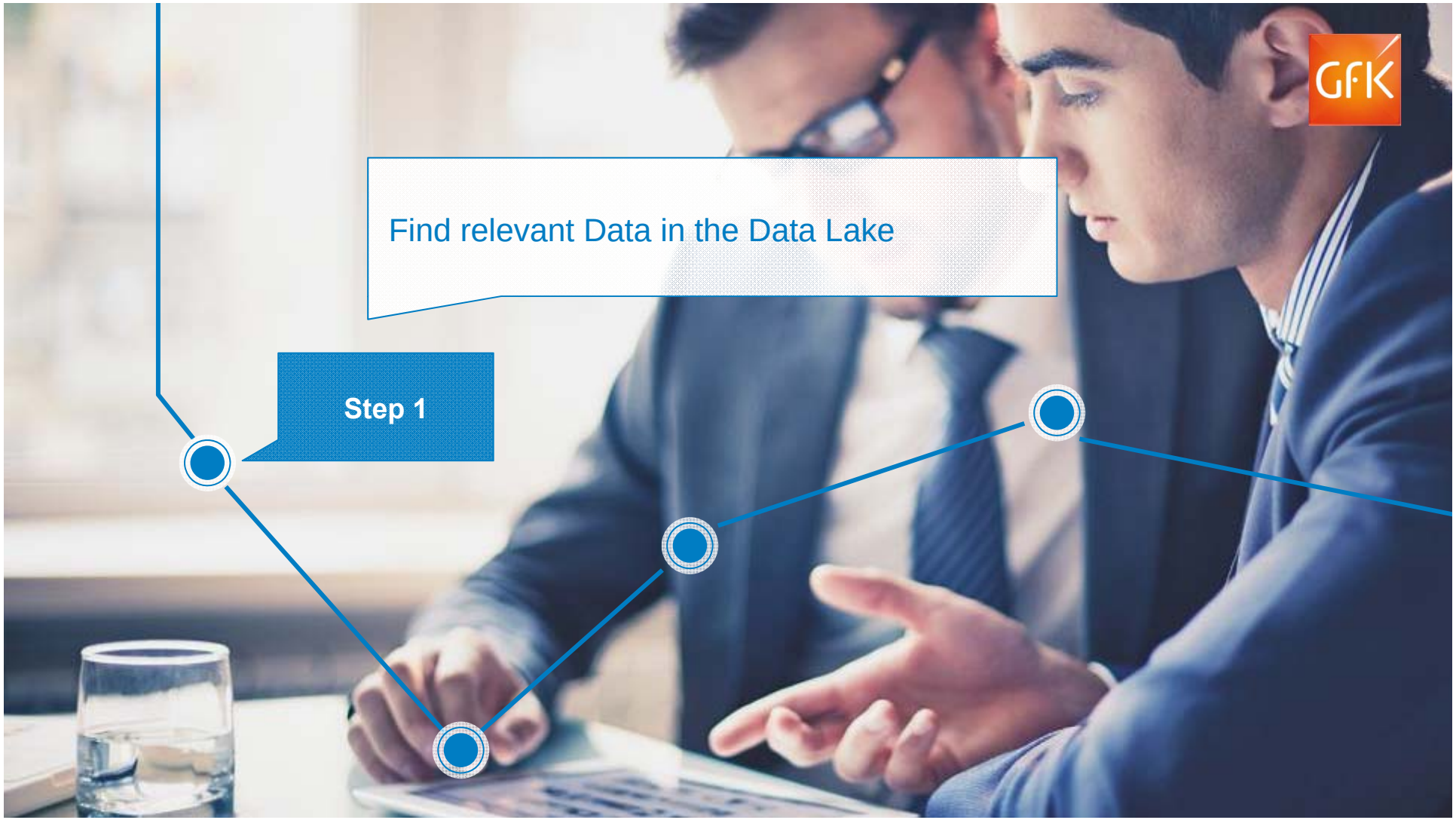

Dank u

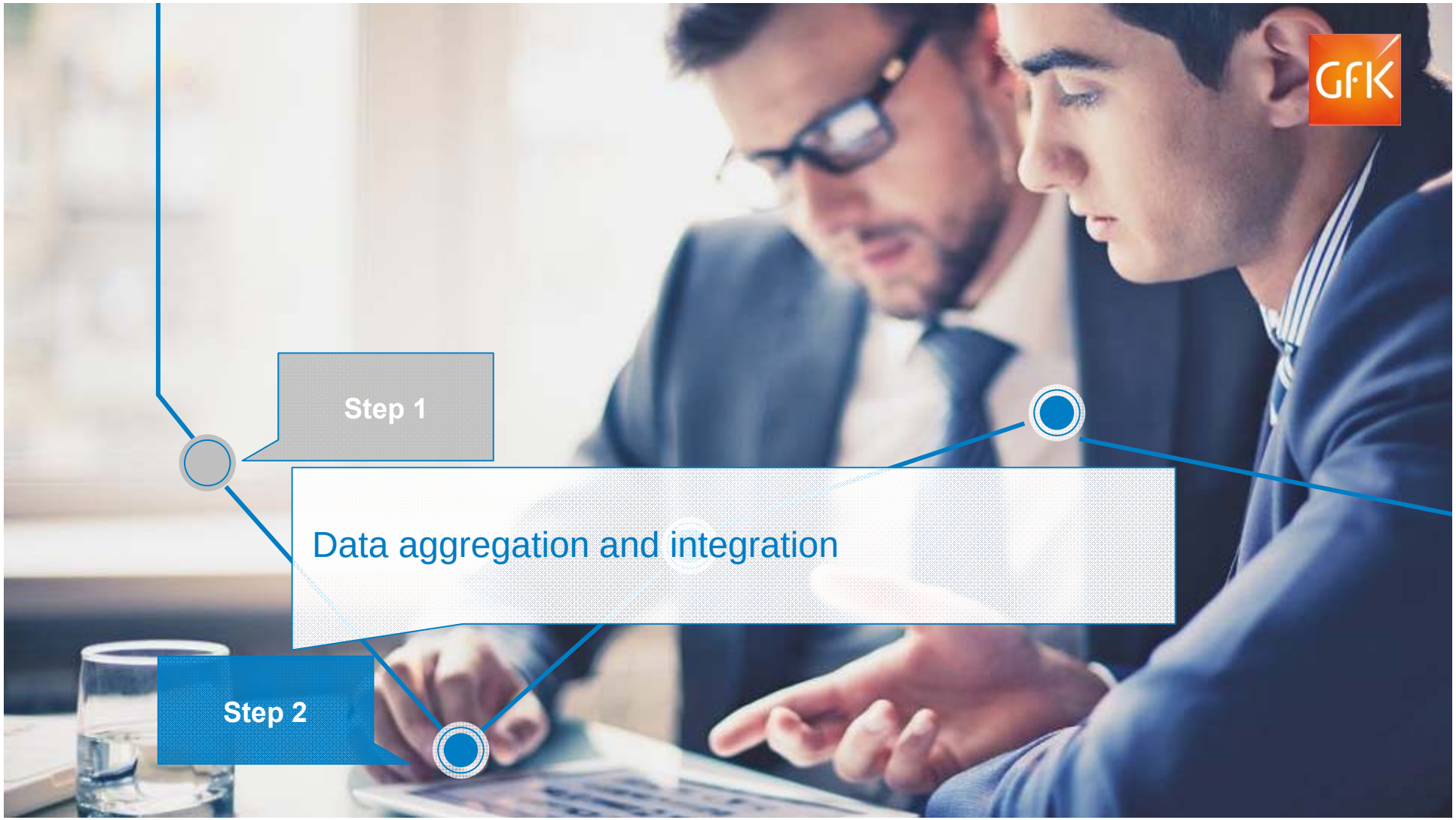

Backup Slides

Our Data Lake combines storage for all our major data assets
with tools for (Big) data ingestion, exploration, and analysis

Final objective: Valuable business applications that leverage our data landscape

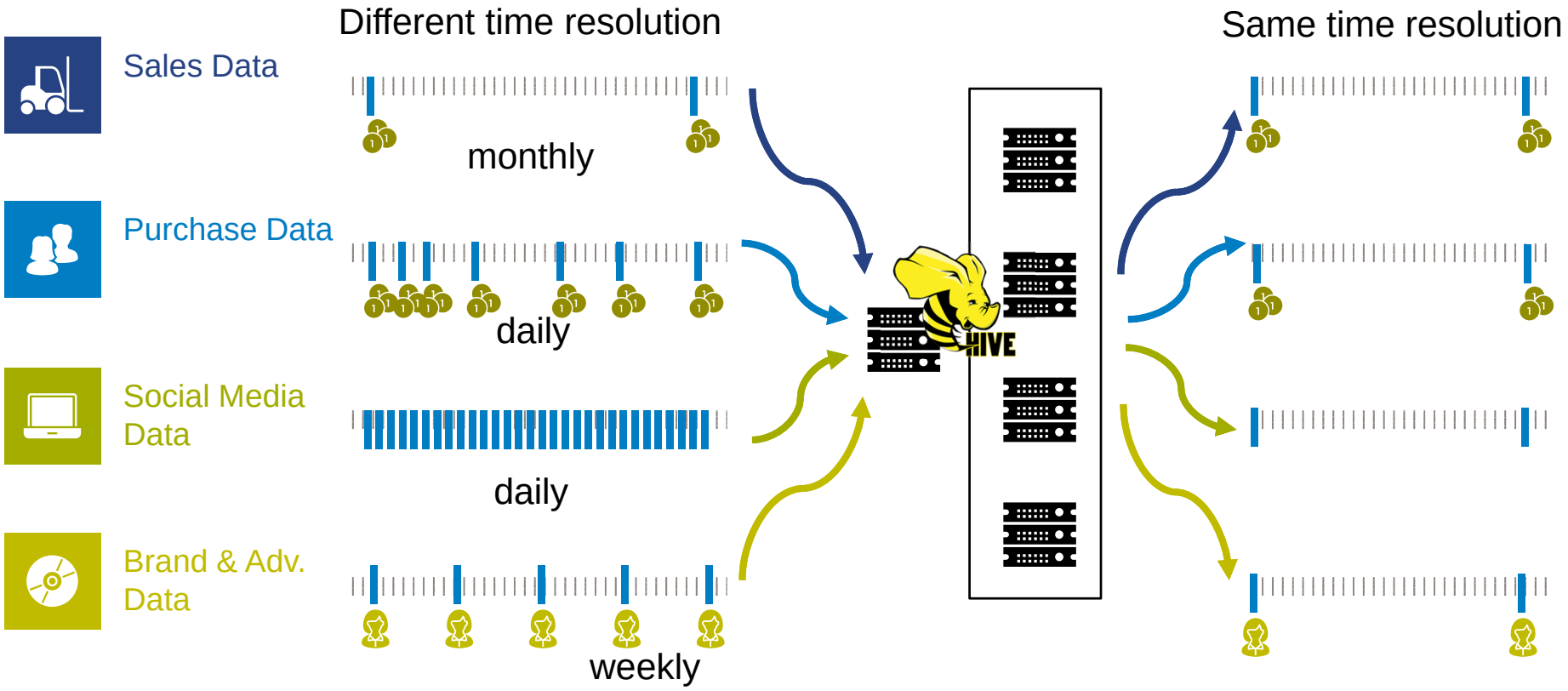


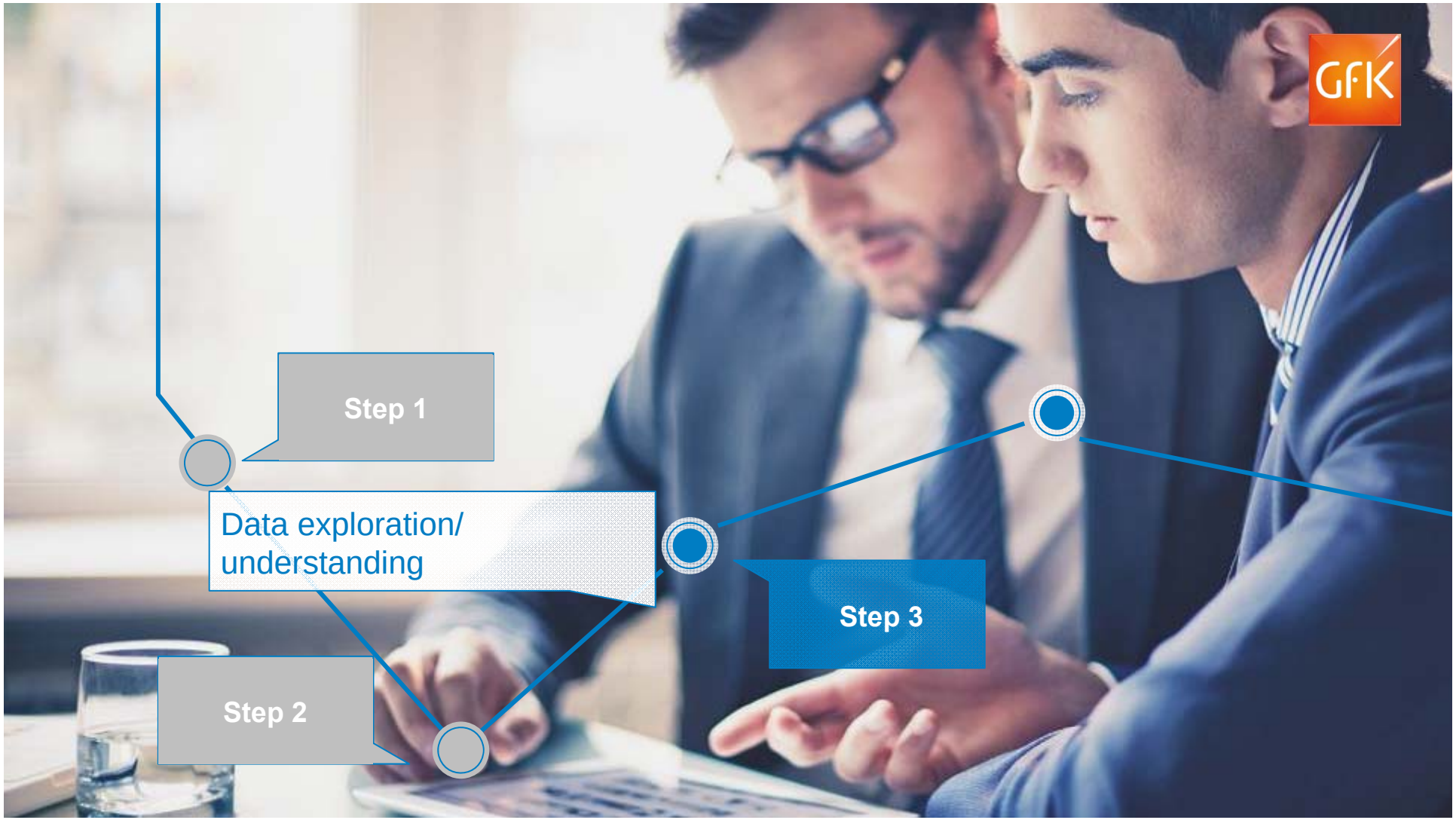

Find relevant Data in the Data Lake
Step 1


Step 1
Data aggregation and integration
Step 2

Hive is used to aggregate and integrate data from different sources with regard to the time dimension

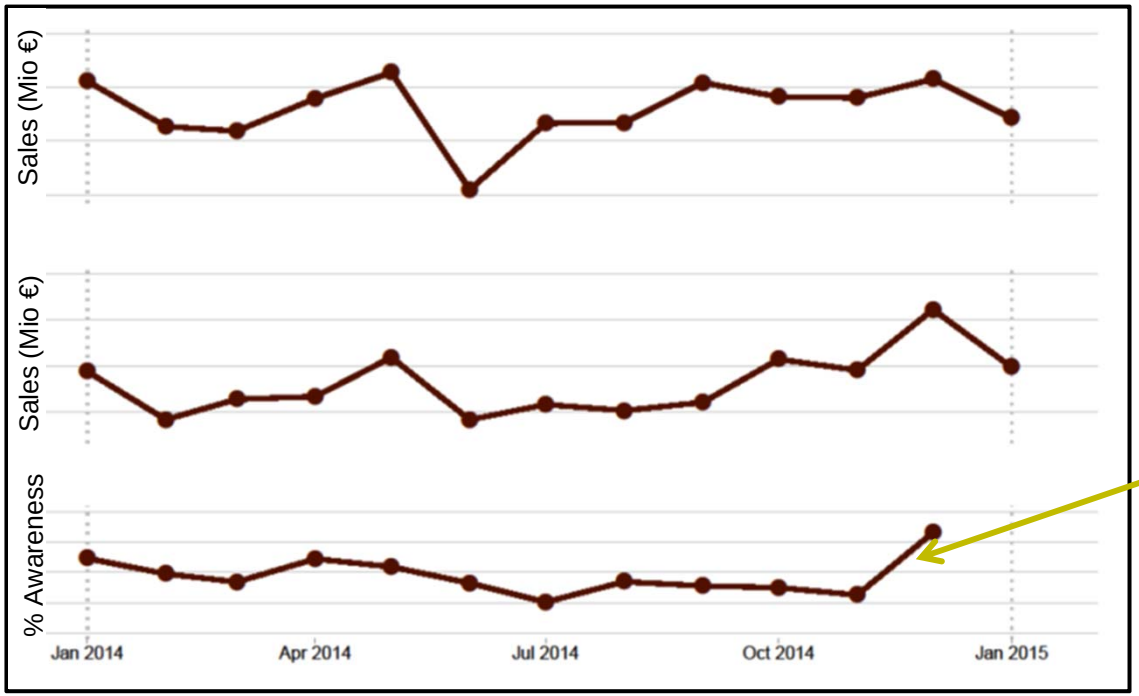
Aggregation




Step 1
Data exploration/
understanding
Step 3
Step 2

Data Scientists use preferred tools (e.g., R, Python) to explore data and understand relationships

For example, check for correlations of KPIs and insights from different data sources, …



Product purch. data from Consumer Panel

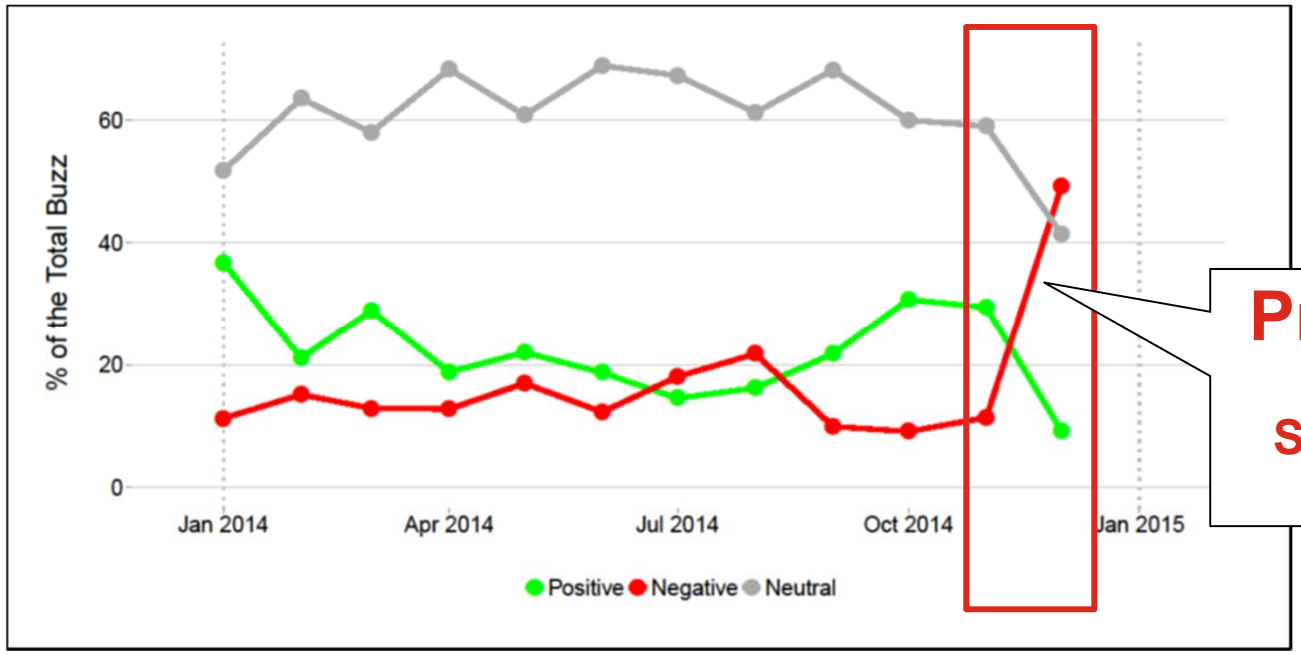
Sales of related technical enablers from Retail Panel

Advertising pressure ↑?

Advertising Awareness from Advertising Tracker

Data Scientists use preferred tools (e.g., R, Python) to explore data and understand relationships

… dive into the “why” behind some data phenomena …



Social Media Monitoring

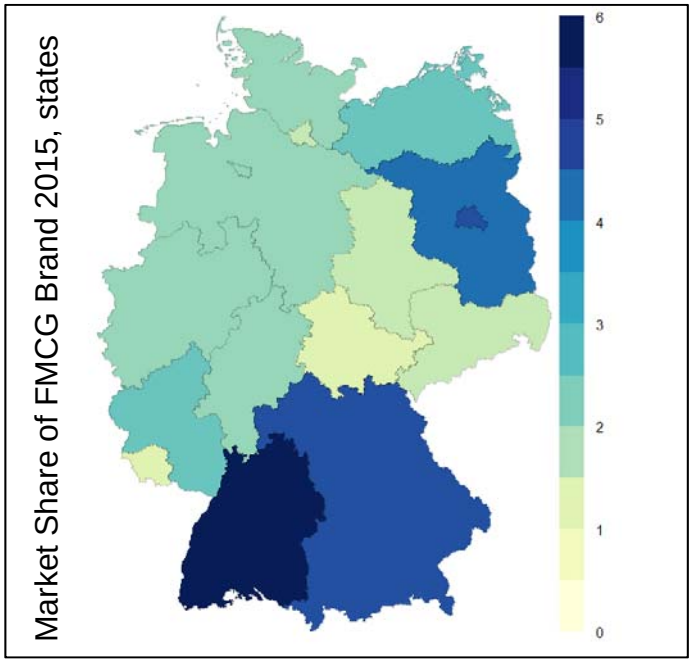
Price increase

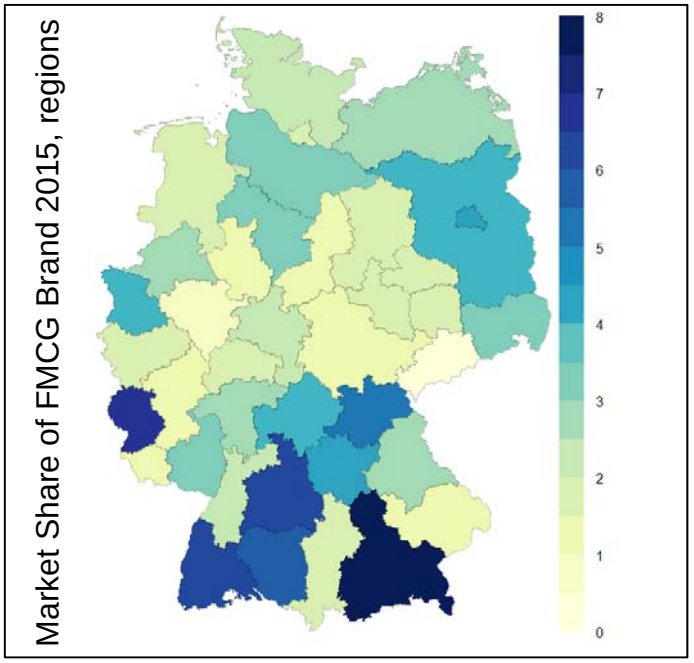
Minimum wage

Smaller Christmas packages

Data Scientists use preferred tools (e.g., R, Python) to explore data and understand relationships

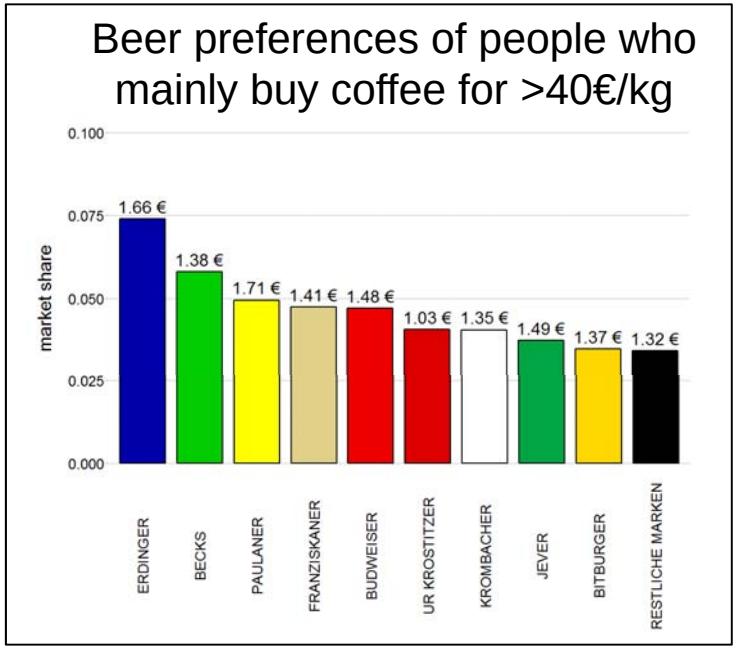
… explore regional sales patterns…



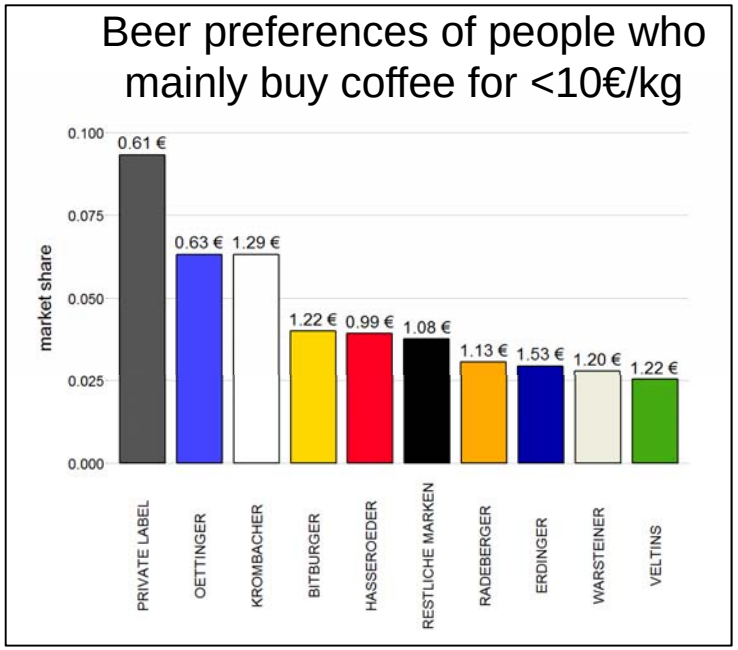


The Data Lake allows for quick and agile hypothesis-testing across data sets and product groups

Do people who buy premium coffee also prefer more expensive beer brands?



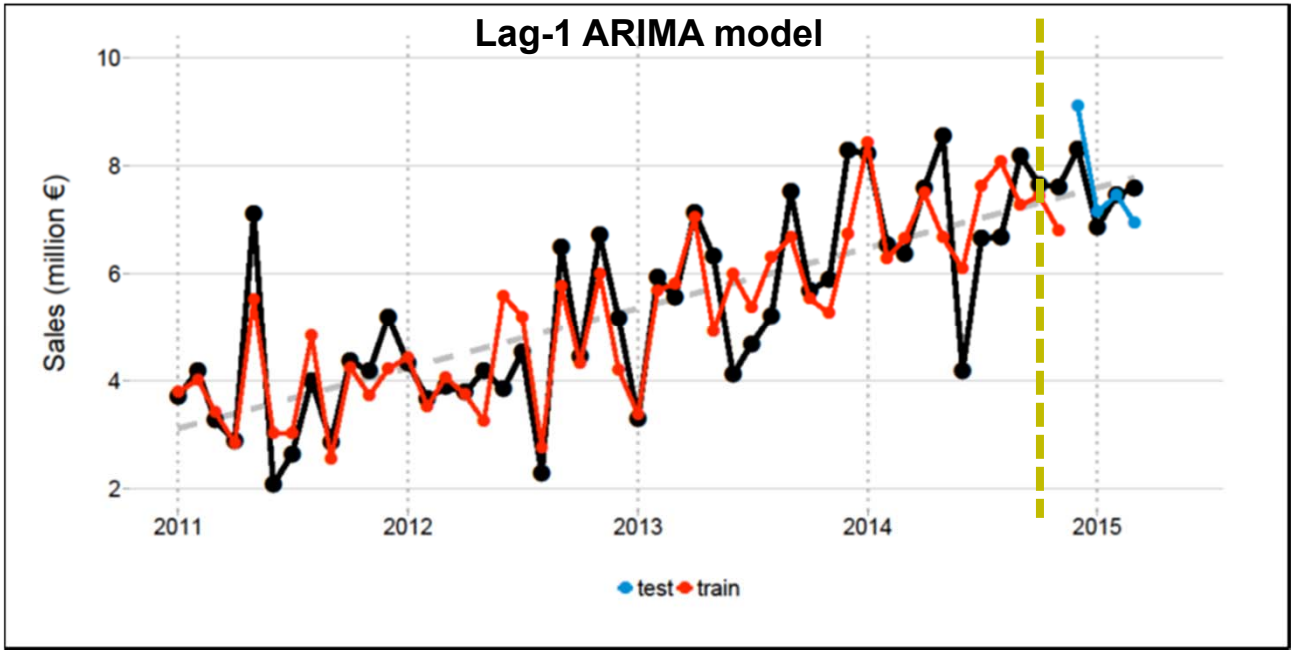
YES!




Modeling
Step 1
Step 2
Step 3
Step 4

Finally, predictive modeling benefits from direct data access and freedom to explore different approaches

Prediction of Nespresso sales in 2015 – utilizing the full four years of coffee purchases…



Finally, predictive modeling benefits from direct data access and freedom to explore different approaches

…and a well-performing integrated model using only two years of information from different data sources

Cons. Panel **Single Serving Coffee Sales**

Retail Panel **Single Serving Machine Sales**

Ext. Data **World market coffee prices**