Practical Machine Learning

A Use Case for Data Cleansing and Object Identification in Market Research

Prof. Dr. Thomas Ruf (Kynetec Germany GmbH)

Practical Machine Learning

A Use Case for Data Cleansing and Object Identification in Market Research

Prof. Dr. Thomas Ruf (Kynetec Germany GmbH)

Outline

- Kynetec business overview
- Panel-based market research
- Current data production workflow
    - Preparation of incoming data for matching
    - Master data coding
    - Transactional data matching
- A Use Case for Machine Learning?!

Outline

- ➡ Kynetec business overview
- Panel-based market research
- Current data production workflow
  - ○ Preparation of incoming data for matching
  - ○ Master data coding
  - ○ Transactional data matching
- A Use Case for Machine Learning?!

kynetec

kynetec

clearer vision smarter decision

Global market research in animal health and agriculture

As **global leaders in market research** for **animal health** and **agriculture**, Kynetec helps companies around the world understand the dynamics of their marketplace, turning research into business opportunities and enabling clients to create winning strategies. We conduct **tracking studies** for monitoring **market trends**, **customized research** for answering unique business challenges and provide **market forecasts** to support your long-term vision, bringing you closer to your customers.

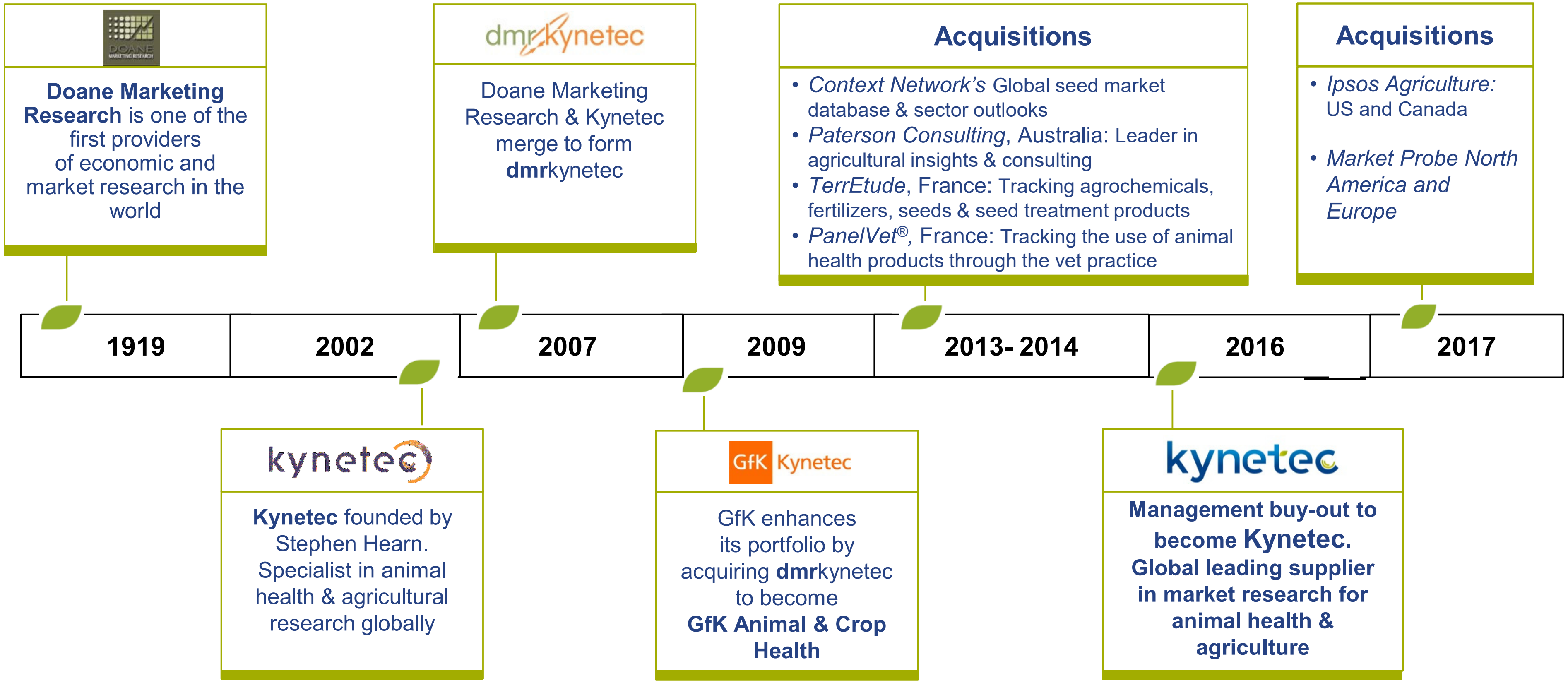
Conducting research in more than **80** countries

Connecting with **100000s** vets, farmers and pet owners

Processing more than **1bn** data points annually

Global team of more than **650** research professionals

Long history in understanding animal health & agriculture



1919

Doane Marketing Research is one of the first providers of economic and market research in the world

2002

Kynetec founded by Stephen Hearn. Specialist in animal health & agricultural research globally

2007

Doane Marketing Research & Kynetec merge to form **dmr**kynetec

2009

GfK enhances its portfolio by acquiring **dmr**kynetec to become **GfK Animal & Crop Health**

2013- 2014

Acquisitions

- *Context Network's* Global seed market database & sector outlooks
- *Paterson Consulting*, Australia: Leader in agricultural insights & consulting
- *TerrEtude*, France: Tracking agrochemicals, fertilizers, seeds & seed treatment products
- *PanelVet*®, France: Tracking the use of animal health products through the vet practice

2016

Management buy-out to become Kynetec. Global leading supplier in market research for animal health & agriculture

2017

Acquisitions

- *Ipsos Agriculture:* US and Canada
- *Market Probe North America and Europe*

Delivering exceptional quality research in all sectors of our industry, covering all information needs

Client challenges ➔ Our thinking

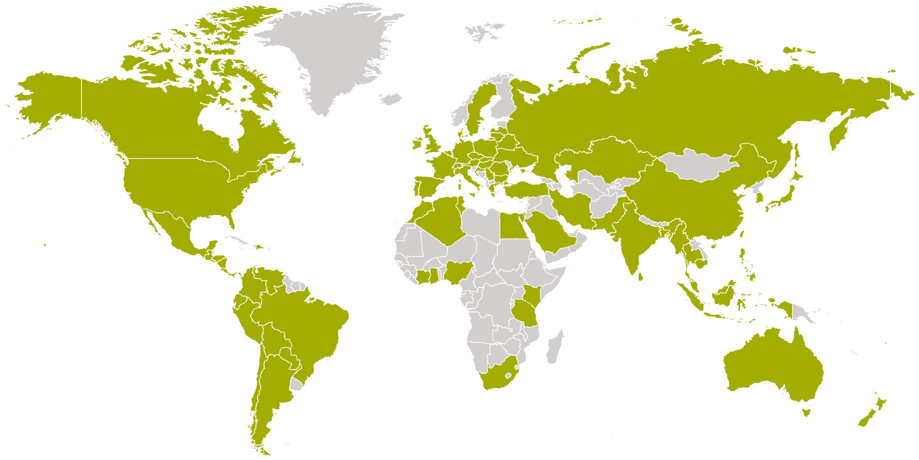
Market opportunity & innovation	Message management: Resonance, reach, recall
Brand & customer experience	Forecasting
Digital strategy & intelligence	Launch readiness & evaluation
Value & pricing optimization	Segmentation & positioning
User experience	Concept testing
Retail sales tracking	Market dynamics & competitive landscaping
Product design optimization	Packaging design optimization

Sector expertise

- Animal health
- Animal nutrition
- Animal genetics
- Crop protection
- Seed & seed treatment
- Plant nutrition
- Water management
- Soil management
- Machinery
- Amenity

Connecting with clients & markets globally

Present in all major regions, and in most major markets

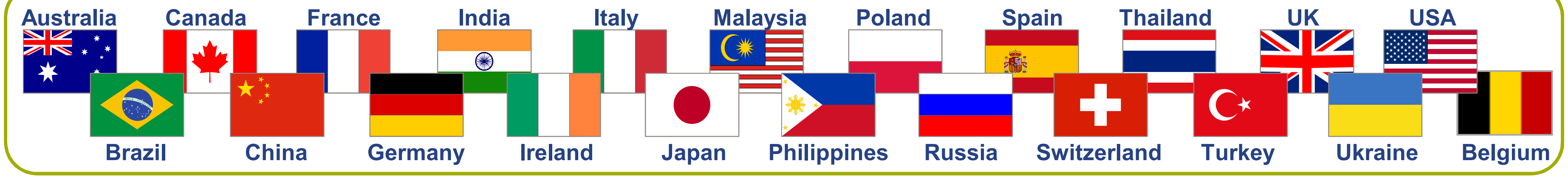


Global Reach & local presence

- Conducting research in more than 80 countries
- Employees present in 22 major agriculture & animal health countries

More than 650 talented professionals

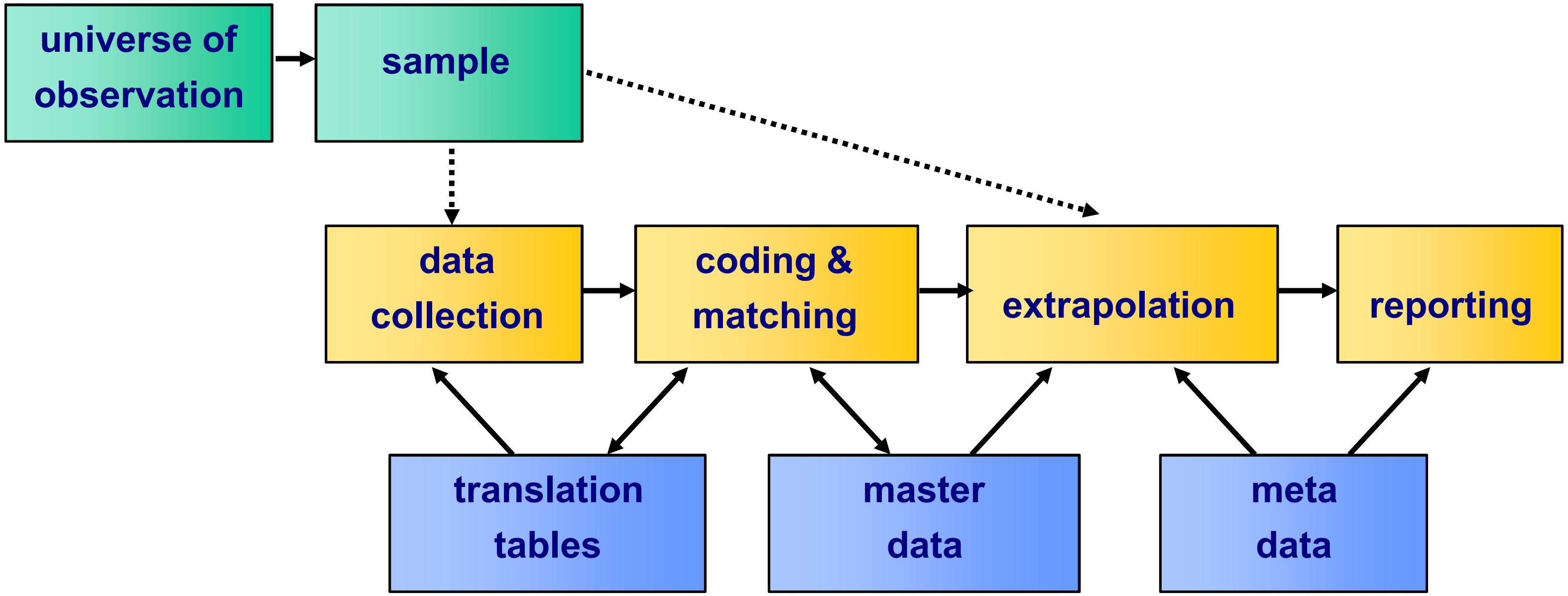
- 300 research professionals
- 300 skilled interviewers
- >80 research partners



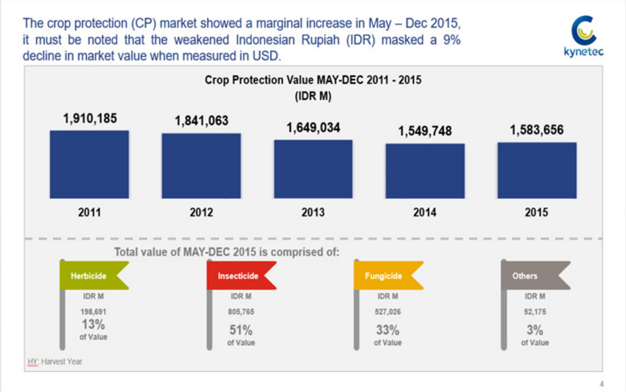
Outline

- Kynetec business overview
- → Panel-based market research
- Current data production workflow
  - Preparation of incoming data for matching
  - Master data coding
  - Transactional data matching
- A Use Case for Machine Learning?!

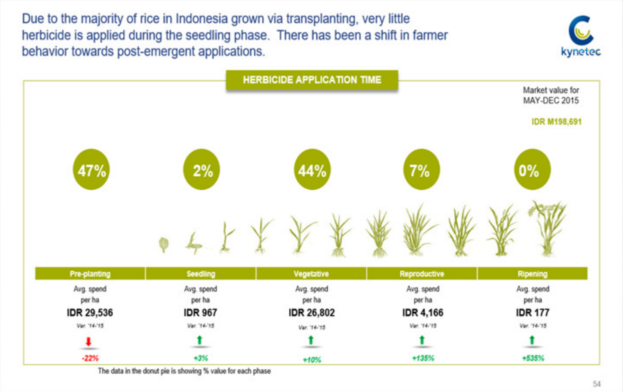
Panel market research methodology



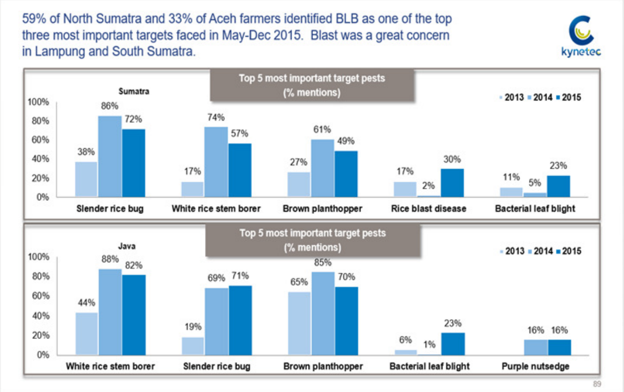
In many regions, we provide market insight reports that cover key categories segmented by defined regions.



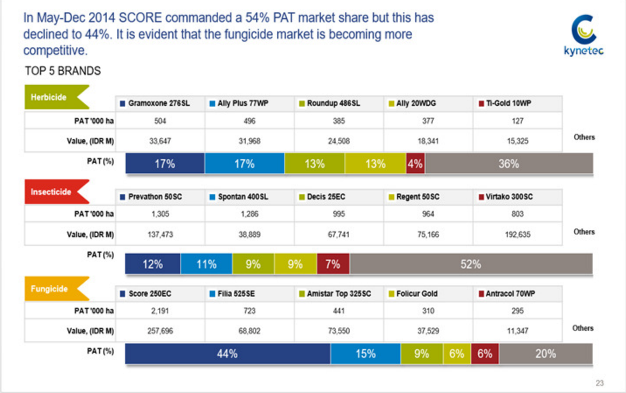
The crop protection (CP) market showed a marginal increase in May – Dec 2015, it must be noted that the weakened Indonesian Rupiah (IDR) masked a 9% decline in market value when measured in USD.



Due to the majority of rice in Indonesia grown via transplanting, very little herbicide is applied during the seedling phase. There has been a shift in farmer behavior towards post-emergent applications.



59% of North Sumatra and 33% of Aceh farmers identified BLB as one of the top three most important targets faced in May-Dec 2015. Blast was a great concern in Lampung and South Sumatra.



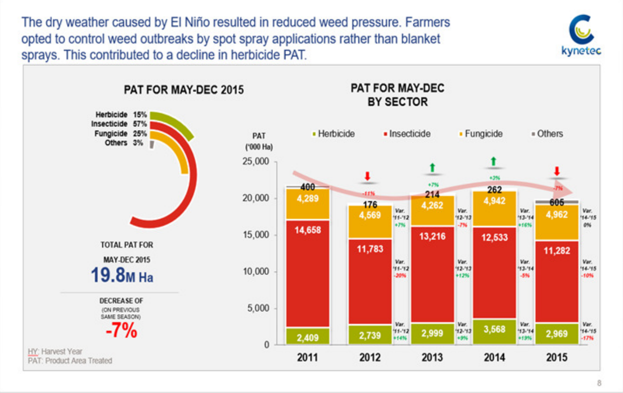
In May-Dec 2014 SCORE commanded a 54% PAT market share but this has declined to 44%. It is evident that the fungicide market is becoming more competitive.

TOP 5 BRANDS

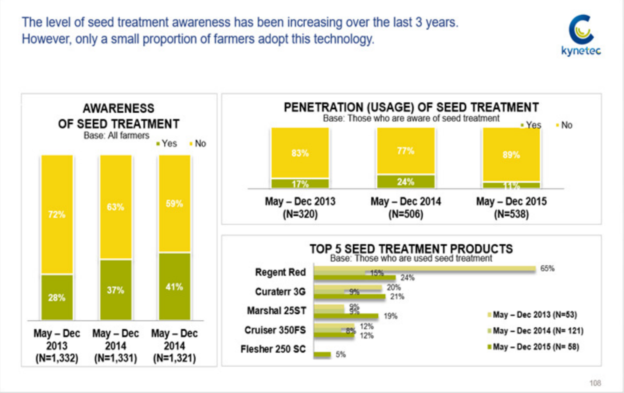
Herbicide	Gramoxone 276SL	Ally Plus 77WP	Roundup 486SL	Ally 20WDG	Ti-Gold 10WP	
PAT '000 ha	504	496	385	377	127	
Value, (IDR M)	33,647	31,968	24,508	18,341	15,325	Others
PAT (%)	17%	17%	13%	13%	4%	36%

Insecticide	Prevathon 50SC	Spontan 400SL	Decis 25EC	Regent 50SC	Virtako 300SC	
PAT '000 ha	1,305	1,286	995	964	803	
Value, (IDR M)	137,473	38,889	67,741	75,166	192,635	Others
PAT (%)	12%	11%	9%	9%	7%	52%

Fungicide	Score 250EC	Filia 525SE	Amistar Top 325SC	Folicur Gold	Antracol 70WP	
PAT '000 ha	2,191	723	441	310	295	
Value, (IDR M)	257,696	68,802	73,550	37,529	11,347	Others
PAT (%)	44%	15%	9%	6%	6%	20%



The dry weather caused by El Niño resulted in reduced weed pressure. Farmers opted to control weed outbreaks by spot spray applications rather than blanket sprays. This contributed to a decline in herbicide PAT.



The level of seed treatment awareness has been increasing over the last 3 years. However, only a small proportion of farmers adopt this technology.

Broad selection of syndicated studies to meet client needs

VetTrak™

VetTrak Insight™

VetTrak™
BDI/CDI Score

PetInsight™

SpecialtyTrak™

ParaTrak™

Consumer
Perceived
Recommendations

Pet Care
Retail
Trackers

VetTrak
Forecaster™

VetTrak EU™

Sigma AH™

PressTrak™
Vet

Global Digital
Trackers Studies/CEE

VetTrak™ Geo

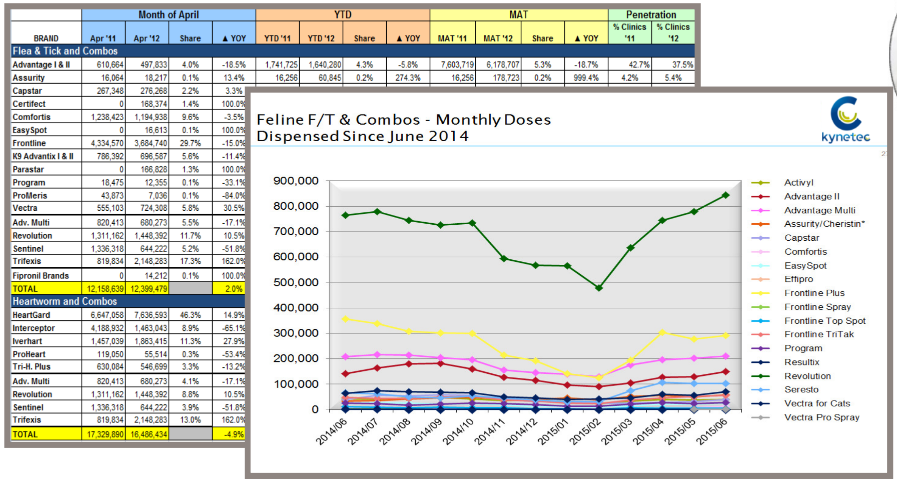
VetTrak™

Overview

- Tracks product consumption (dispensed doses and revenue) and inventory levels for key parasiticides and vaccines in the veterinary channel

Content

- Flea/tick, heartworm, GI worms, vaccines, NSAIDs, & Anti-infectives
- Metrics: Current Period, MAT, and YTD
- Includes up to10 years' worth of historical data



	Month of April				YTD				MAT				Penetration	
BRAND	Apr '11	Apr '12	Share	▲ YOY	YTD '11	YTD '12	Share	▲ YOY	MAT '11	MAT '12	Share	▲ YOY	% Clinics '11	% Clinics '12
Flea & Tick and Combos														
Advantage I & II	610,664	497,833	4.0%	-18.5%	1,741,725	1,640,280	4.3%	-5.8%	7,603,719	6,178,707	5.3%	-18.7%	42.7%	37.5%
Assurity	16,064	18,217	0.1%	13.4%	16,256	60,845	0.2%	274.3%	16,256	178,723	0.2%	999.4%	4.2%	5.4%
Capstar	267,348	276,268	2.2%	3.3%										
Certifect	0	168,374	1.4%	100.0%										
Comfortis	1,238,423	1,194,938	9.6%	-3.5%										
EasySpot	0	16,613	0.1%	100.0%										
Frontline	4,334,570	3,684,740	29.7%	-15.0%										
K9 Advantix I & II	786,392	696,587	5.6%	-11.4%										
Parastar	0	166,828	1.3%	100.0%										
Program	18,475	12,355	0.1%	-33.1%										
ProMeris	43,873	7,036	0.1%	-84.0%										
Vectra	555,103	724,308	5.8%	30.5%										
Adv. Multi	820,413	680,273	5.5%	-17.1%										
Revolution	1,311,162	1,448,392	11.7%	10.5%										
Sentinel	1,336,318	644,222	5.2%	-51.8%										
Trifexis	819,834	2,148,283	17.3%	162.0%										
Fipronil Brands	0	14,212	0.1%	100.0%										
TOTAL	12,158,639	12,399,479		2.0%										
Heartworm and Combos														
HeartGard	6,647,058	7,636,593	46.3%	14.9%										
Interceptor	4,188,932	1,463,043	8.9%	-65.1%										
Iverhart	1,457,039	1,863,415	11.3%	27.9%										
ProHeart	119,050	55,514	0.3%	-53.4%										
Tri-H. Plus	630,084	546,699	3.3%	-13.2%										
Adv. Multi	820,413	680,273	4.1%	-17.1%										
Revolution	1,311,162	1,448,392	8.8%	10.5%										
Sentinel	1,336,318	644,222	3.9%	-51.8%										
Trifexis	819,834	2,148,283	13.0%	162.0%										
TOTAL	17,329,890	16,486,434		-4.9%										

Feline F/T & Combos - Monthly Doses Dispensed Since June 2014

Methodology

- US: survey-based clinic inventory data is blended with data extracted from a panel of practice management systems
- Mixed mode survey (n=~800) / PMS (n=~8,000)
- Projected to the universe of independent companion animal clinics (N=~25,000)

Deliverables

- Monthly using i-map3™ platform
- Powerpoint summary report
- Excel based pivot table reports
- Available 20 working days after month end

Benefits

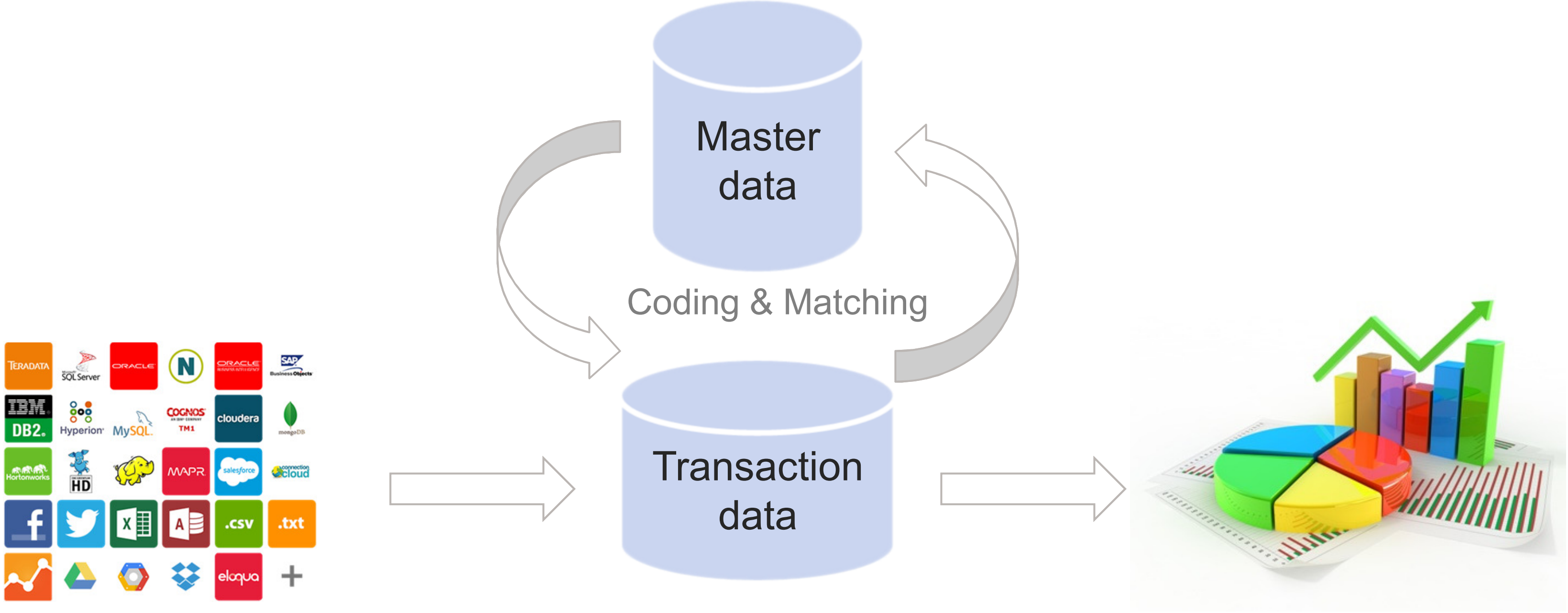
- Provides ongoing tracking of key metrics needed to manage your brand in the vet channel
- Tracks sales out of vet clinic – doses and dollars

Outline

- Kynetec business overview
- Panel-based market research
- ➡ Current data production workflow
  - ○ Preparation of incoming data for matching
  - ○ Master data coding
  - ○ Transactional data matching
- A Use Case for Machine Learning?!

Panel-based market research

Simplified Data Production Workflow



Master data

Coding & Matching

Transaction data

Data Production for Panel Research

Key challenge: distinguish noise from data

Nature of transactional data

Transactional data are sourced from incontrollable sources and may contain arbitrary references to products and services applied in a Vet practice or clinic

Lining transactional data to master data

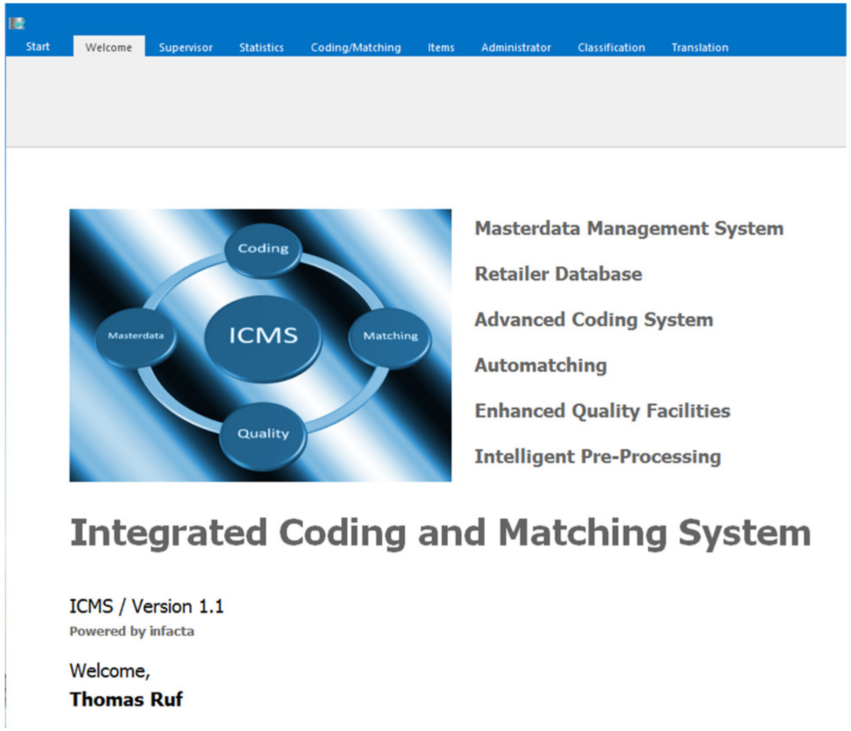
To be able to derive globally meaningful insights from transactional data, single transactions need to be linked to harmonizid master data

Huge data volumes = huge problem?

In the work with big amounts of data, it is essential to minimize the need for manual work. For this reason, a specialized coding and matching solution is needed, supporting both master data management and the actual matching process. ICMS provided by Infacta is our tool of choice.

Current data production system

Integrated Coding and Matching System (ICMS)



Masterdata Management System

Retailer Database

Advanced Coding System

Automatching

Enhanced Quality Facilities

Intelligent Pre-Processing

Integrated Coding and Matching System

ICMS / Version 1.1
Powered by infacta

Welcome,
Thomas Ruf

- ICMS developed by Infacta GmbH, Germany
- Originally created for GfK hospital panel
- Main competitive advantages of Infacta:
  - 30+ years of in-depth panel production knowledge
  - Agile, pragmatic software development approach
  - Strategic technology partnerships
  - Emphasis on managing the production workflow
  - Specific component developments possible
- No licence fees, compensation based on managed system usage

Outline

- Kynetec business overview
- Panel-based market research
- Current data production workflow
  - ➡ Preparation of incoming data for matching
  - Master data coding
  - Transactional data matching
- A Use Case for Machine Learning?!

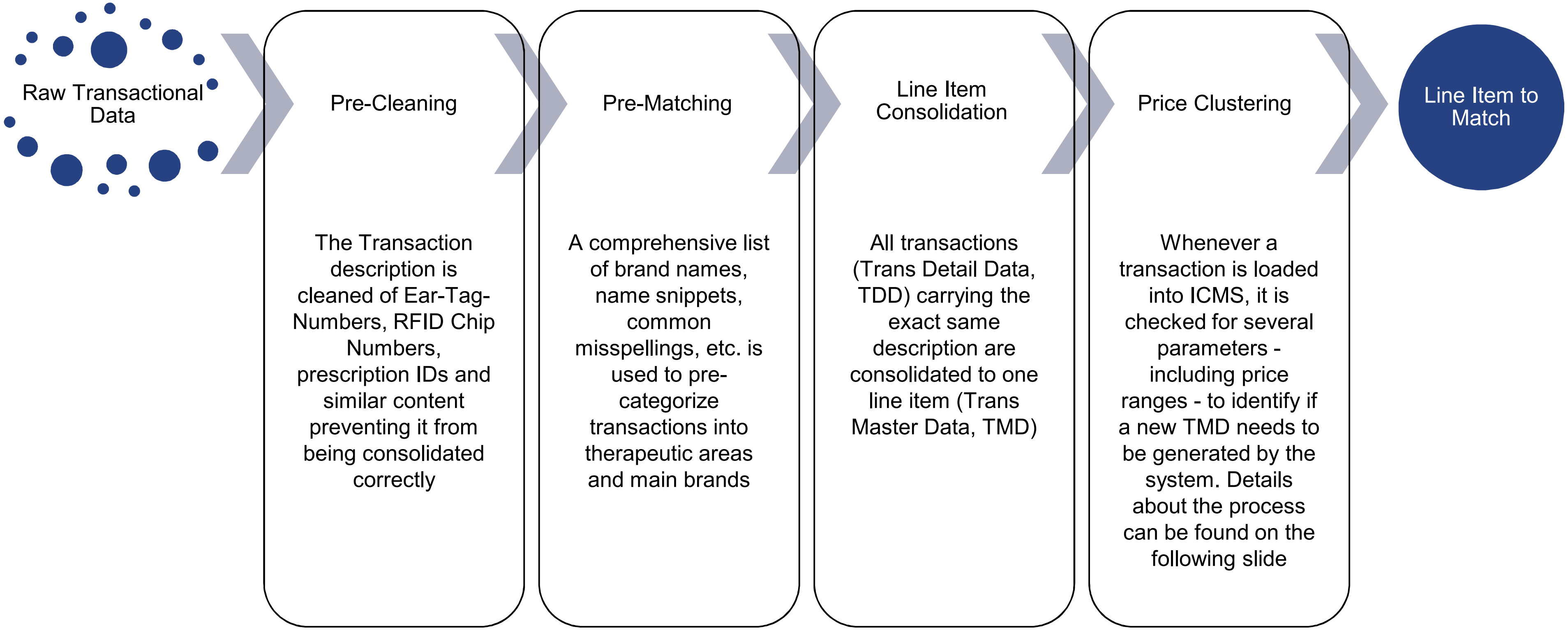
Status Quo

Transactional data of several data sources are the basis of the work in ICMS

Source	# of Transactions
Provider 1	125.462.534
Provider 2	56.847.183
Provider 3	2.203.730
Provider 4	490.590.998
Total	**675.104.446**

Matching close to 700mio transactions is not feasible in an efficient way, so we need to apply methods of data concentration

Simplified Process of Cleaning pre ICMS



Raw Transactional Data

Pre-Cleaning

The Transaction description is cleaned of Ear-Tag-Numbers, RFID Chip Numbers, prescription IDs and similar content preventing it from being consolidated correctly

Pre-Matching

A comprehensive list of brand names, name snippets, common misspellings, etc. is used to pre-categorize transactions into therapeutic areas and main brands

Line Item Consolidation

All transactions (Trans Detail Data, TDD) carrying the exact same description are consolidated to one line item (Trans Master Data, TMD)

Price Clustering

Whenever a transaction is loaded into ICMS, it is checked for several parameters - including price ranges - to identify if a new TMD needs to be generated by the system. Details about the process can be found on the following slide

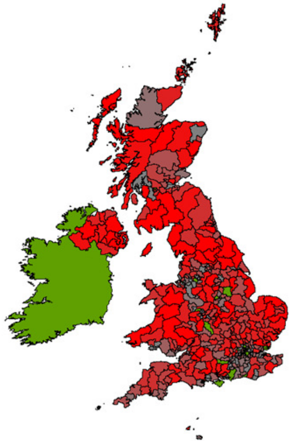
Line Item to Match

Outline

- Kynetec business overview
- Panel-based market research
- Current data production workflow
  - ○ Preparation of incoming data for matching
  - → ○ Master data coding
  - ○ Transactional data matching
- A Use Case for Machine Learning?!

Example: Frontline combo spot-on, large dogs (6 x 2.68ml)

CEESA1	H INSECTICIDES/ECTOPARASITICIDES (EARS EXCL. ANTIFUNGAL INCL.)	23 categories
CEESA2	H01 INSECTICIDES/ECTOPARASITICIDES (EARS EXCL. ANTIFUNGAL INCL.) PETS	101
CEESA3	H01A INSECTICIDES/ECTOPARASITICIDES (EARS EXCL. ANTIFUNGAL INCL.) PETS Flea and ticks	239
CEESA4	H01A03 INSECTICIDES/ECTOPARASITICIDES (EARS EXCL. ANTIFUNGAL INCL.) PETS Flea and ticks Drops	285
CEESA5	H01A03A INSECTICIDES/ECTOPARASITICIDES (EARS EXCL. ANTIFUNGAL INCL.) PETS Flea and ticks Drops Dogs	152
ATC2	C1 SML ANIMAL ECTOPARASITIC	80
ATC3	C1A SML ANIMAL ECTOPARASITIC	123
ATC4	C1A1 SML ANIMAL ECTOS-TOPICAL	203
Product group	FRONTLINE COMBO DOG	c. 6,000 product groups
Product	FRONTLINE COMBO SPOT ON LDOG 6X2.68ML	c. 24,000 SKUs
Manufacturer	MERIAL AH	c. 700 manufacturers
Form	SPOT ON	128 formulations
Mode of admin.	TOPICAL	21 modes of administration
Legal category	POM-V	4 legal categories

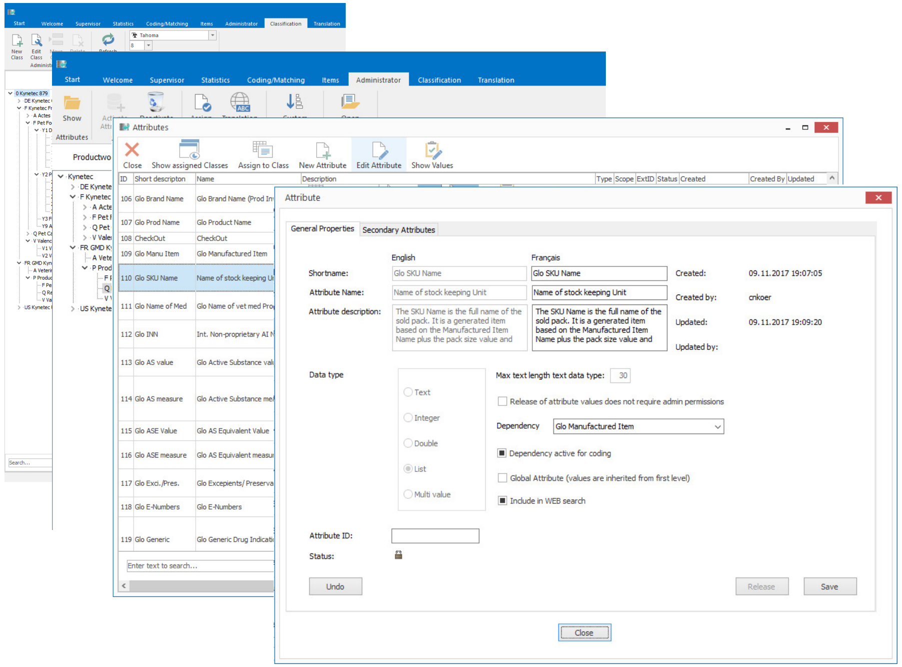


FRONTLINE COMBO SPOT ON LDOG 6X2.68ML
MAT Current Year GBP

VetTrak UK

Attribute Management

General Process (outlined in detail on following slides)



ID	Short descripton	Name
106	Glo Brand Name	Glo Brand Name (Prod In
107	Glo Prod Name	Glo Product Name
108	CheckOut	CheckOut
109	Glo Manu Item	Glo Manufactured Item
110	Glo SKU Name	Name of stock keeping U
111	Glo Name of Med	Glo Name of vet med Pro
112	Glo INN	Int. Non-proprietary AI N
113	Glo AS value	Glo Active Substance val
114	Glo AS measure	Glo Active Substance me
115	Glo ASE Value	Glo AS Equivalent Value
116	Glo ASE measure	Glo AS Equivalent measu
117	Glo Exci./Pres.	Glo Excepients/ Preserva
118	Glo E-Numbers	Glo E-Numbers
119	Glo Generic	Glo Generic Drug Indicati

Enter text to search...

Attribute

General Properties | Secondary Attributes

	English	Français
Shortname:	Glo SKU Name	Glo SKU Name
Attribute Name:	Name of stock keeping Unit	Name of stock keeping Unit
Attribute description:	The SKU Name is the full name of the sold pack. It is a generated item based on the Manufactured Item Name plus the pack size value and	The SKU Name is the full name of the sold pack. It is a generated item based on the Manufactured Item Name plus the pack size value and

Created: 09.11.2017 19:07:05
Created by: cnkoer
Updated: 09.11.2017 19:09:20
Updated by:

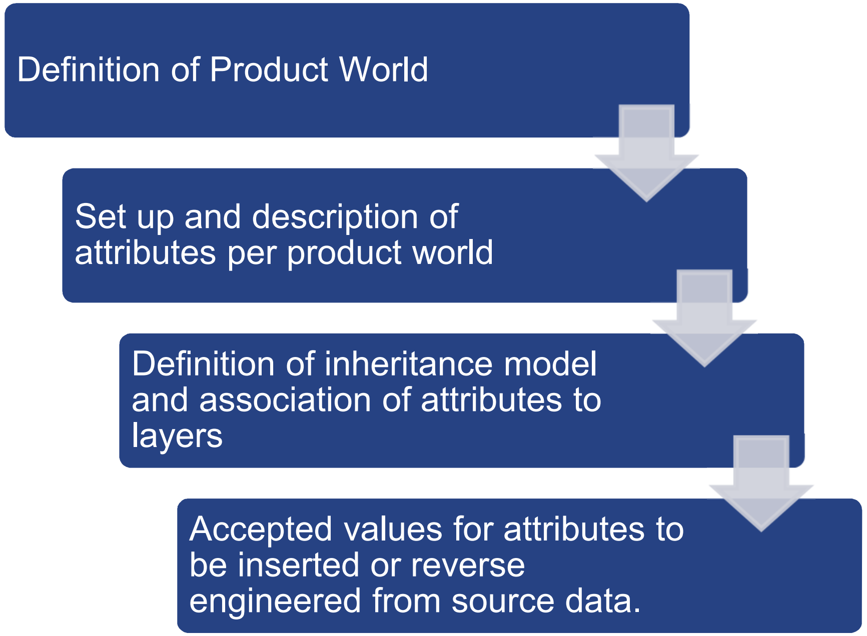
Data type
○ Text
○ Integer
○ Double
● List
○ Multi value

Max text length text data type: 30
□ Release of attribute values does not require admin permissions
Dependency: Glo Manufactured Item
☒ Dependency active for coding
□ Global Attribute (values are inherited from first level)
☒ Include in WEB search

Attribute ID:
Status:

Undo | Release | Save

Close



Definition of Product World

Set up and description of attributes per product world

Definition of inheritance model and association of attributes to layers

Accepted values for attributes to be inserted or reverse engineered from source data.

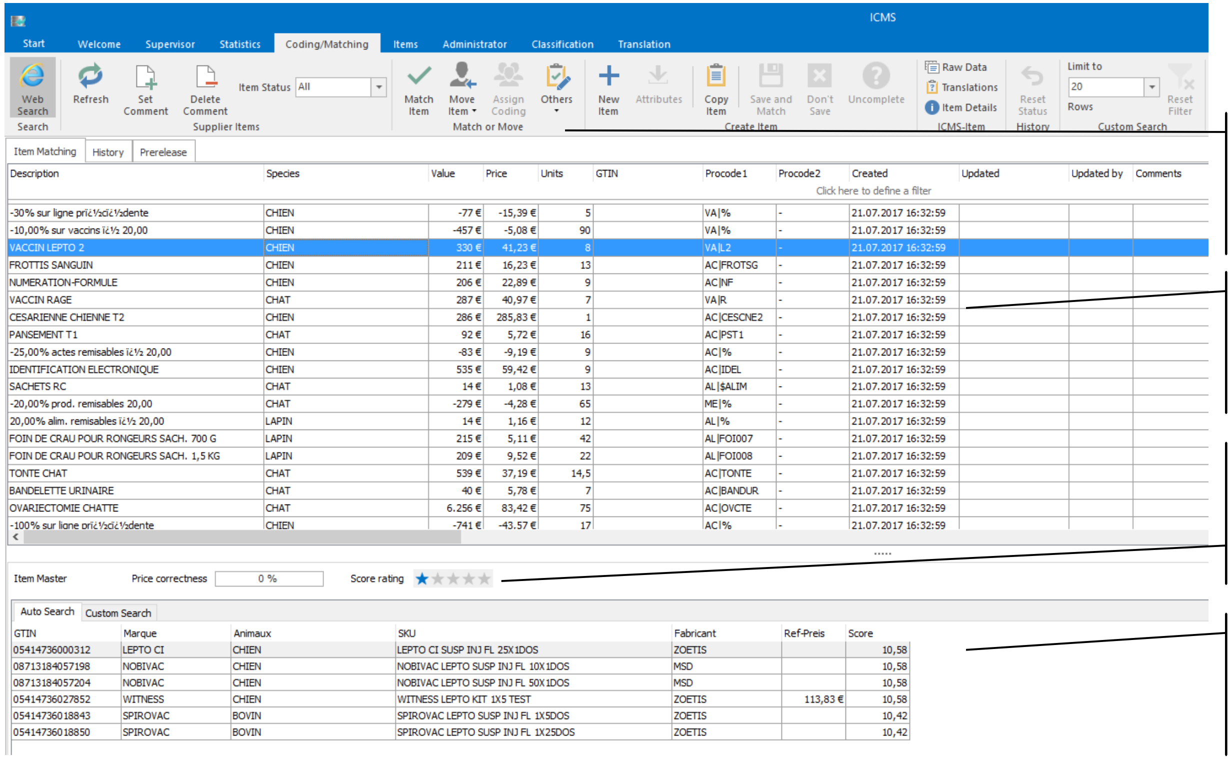
Outline

- Kynetec business overview
- Panel-based market research
- Current data production workflow
    - ○ Preparation of incoming data for matching
    - ○ Master data coding
    - ○ Transactional data matching
- A Use Case for Machine Learning?!

Matching transactional data

In many cases, transactional data doesn’t come in a format ready to use.

For this reason, additional matching to meaningful master data is required to deliver insightful information from unharmonized data.



Item Status: All

Limit to: 20 Rows

Description	Species	Value	Price	Units	GTIN	Procode1	Procode2	Created	Updated	Updated by	Comments
-30% sur ligne prï¿½cï¿½dente	CHIEN	-77 €	-15,39 €	5		VA|%	-	21.07.2017 16:32:59			
-10,00% sur vaccins ï¿½ 20,00	CHIEN	-457 €	-5,08 €	90		VA|%	-	21.07.2017 16:32:59			
VACCIN LEPTO 2	CHIEN	330 €	41,23 €	8		VA|L2	-	21.07.2017 16:32:59			
FROTTIS SANGUIN	CHIEN	211 €	16,23 €	13		AC|FROTSG	-	21.07.2017 16:32:59			
NUMERATION-FORMULE	CHIEN	206 €	22,89 €	9		AC|NF	-	21.07.2017 16:32:59			
VACCIN RAGE	CHAT	287 €	40,97 €	7		VA|R	-	21.07.2017 16:32:59			
CESARIENNE CHIENNE T2	CHIEN	286 €	285,83 €	1		AC|CESCNE2	-	21.07.2017 16:32:59			
PANSEMENT T1	CHAT	92 €	5,72 €	16		AC|PST1	-	21.07.2017 16:32:59			
-25,00% actes remisables ï¿½ 20,00	CHIEN	-83 €	-9,19 €	9		AC|%	-	21.07.2017 16:32:59			
IDENTIFICATION ELECTRONIQUE	CHIEN	535 €	59,42 €	9		AC|IDEL	-	21.07.2017 16:32:59			
SACHETS RC	CHAT	14 €	1,08 €	13		AL|$ALIM	-	21.07.2017 16:32:59			
-20,00% prod. remisables 20,00	CHAT	-279 €	-4,28 €	65		ME|%	-	21.07.2017 16:32:59			
20,00% alim. remisables ï¿½ 20,00	LAPIN	14 €	1,16 €	12		AL|%	-	21.07.2017 16:32:59			
FOIN DE CRAU POUR RONGEURS SACH. 700 G	LAPIN	215 €	5,11 €	42		AL|FOI007	-	21.07.2017 16:32:59			
FOIN DE CRAU POUR RONGEURS SACH. 1,5 KG	LAPIN	209 €	9,52 €	22		AL|FOI008	-	21.07.2017 16:32:59			
TONTE CHAT	CHAT	539 €	37,19 €	14,5		AC|TONTE	-	21.07.2017 16:32:59			
BANDELETTE URINAIRE	CHAT	40 €	5,78 €	7		AC|BANDUR	-	21.07.2017 16:32:59			
OVARIECTOMIE CHATTE	CHAT	6.256 €	83,42 €	75		AC|OVCTE	-	21.07.2017 16:32:59			
-100% sur ligne prï¿½cï¿½dente	CHIEN	-741 €	-43,57 €	17		AC|%	-	21.07.2017 16:32:59			

Item Master Price correctness 0 % Score rating ★☆☆☆☆

GTIN	Marque	Animaux	SKU	Fabricant	Ref-Preis	Score
05414736000312	LEPTO CI	CHIEN	LEPTO CI SUSP INJ FL 25X1DOS	ZOETIS		10,58
08713184057198	NOBIVAC	CHIEN	NOBIVAC LEPTO SUSP INJ FL 10X1DOS	MSD		10,58
08713184057204	NOBIVAC	CHIEN	NOBIVAC LEPTO SUSP INJ FL 50X1DOS	MSD		10,58
05414736027852	WITNESS	CHIEN	WITNESS LEPTO KIT 1X5 TEST	ZOETIS	113,83 €	10,58
05414736018843	SPIROVAC	BOVIN	SPIROVAC LEPTO SUSP INJ FL 1X5DOS	ZOETIS		10,42
05414736018850	SPIROVAC	BOVIN	SPIROVAC LEPTO SUSP INJ FL 1X25DOS	ZOETIS		10,42

Selection of key functionality

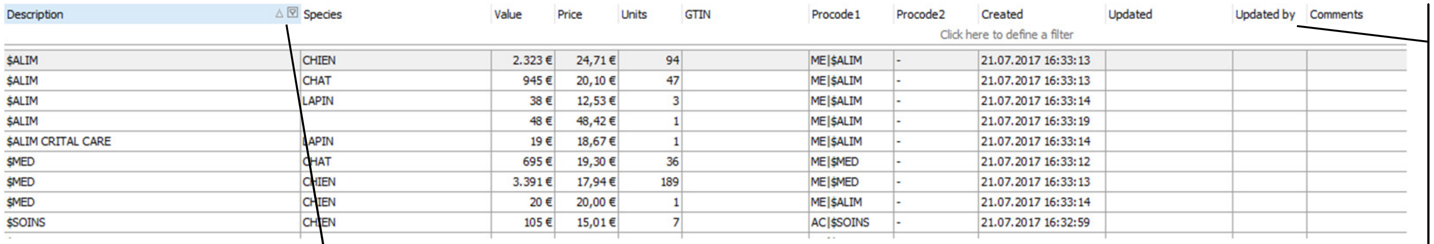
Transactional data

Indicators for the quality of the proposed master data item

Master data items

Matching transactional data

Transactions window

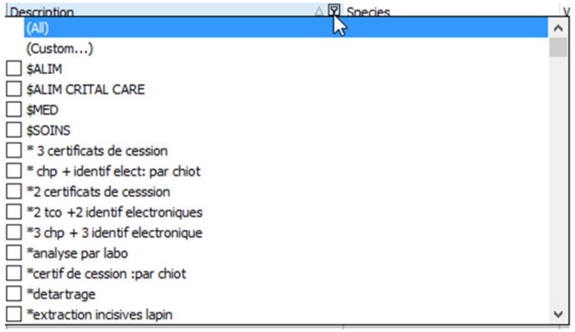


Click here to define a filter

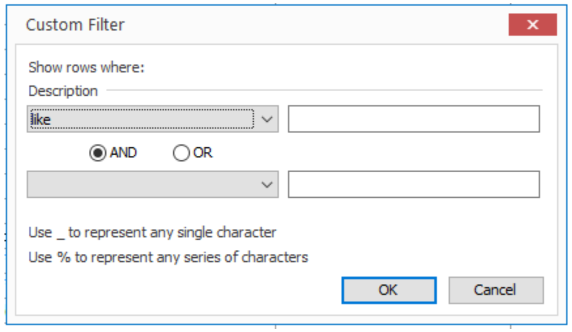
Description	Species	Value	Price	Units	GTIN	Procode1	Procode2	Created	Updated	Updated by	Comments
$ALIM	CHIEN	2.323 €	24,71 €	94		ME|$ALIM	-	21.07.2017 16:33:13			
$ALIM	CHAT	945 €	20,10 €	47		ME|$ALIM	-	21.07.2017 16:33:13			
$ALIM	LAPIN	38 €	12,53 €	3		ME|$ALIM	-	21.07.2017 16:33:14			
$ALIM		48 €	48,42 €	1		ME|$ALIM	-	21.07.2017 16:33:19			
$ALIM CRITAL CARE	LAPIN	19 €	18,67 €	1		ME|$ALIM	-	21.07.2017 16:33:14			
$MED	CHAT	695 €	19,30 €	36		ME|$MED	-	21.07.2017 16:33:12			
$MED	CHIEN	3.391 €	17,94 €	189		ME|$MED	-	21.07.2017 16:33:13			
$MED	CHIEN	20 €	20,00 €	1		ME|$ALIM	-	21.07.2017 16:33:14			
$SOINS	CHIEN	105 €	15,01 €	7		AC|$SOINS	-	21.07.2017 16:32:59			

All relevant information of the transactional data can be displayed to enhance coding quality through more reference information

In addition to this, all columns can make use of filters and customized searches



Description
Species
(All)
(Custom...)
□ $ALIM
□ $ALIM CRITAL CARE
□ $MED
□ $SOINS
□ * 3 certificats de cession
□ * chp + identif elect: par chiot
□ *2 certificats de cesssion
□ *2 tco +2 identif electroniques
□ *3 chp + 3 identif electronique
□ *analyse par labo
□ *certif de cession :par chiot
□ *detartrage
□ *extraction incisives lapin



Custom Filter

Show rows where:

Description

like

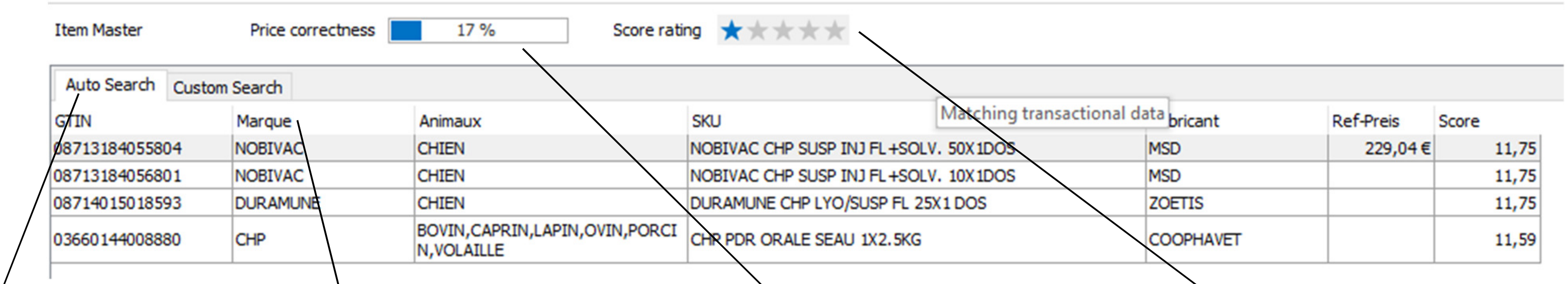
● AND ○ OR

Use _ to represent any single character

Use % to represent any series of characters

Matching transactional data

Master Data window



Item Master   Price correctness 17 %   Score rating ★☆☆☆☆

GTIN	Marque	Animaux	SKU	bricant	Ref-Preis	Score
08713184055804	NOBIVAC	CHIEN	NOBIVAC CHP SUSP INJ FL+SOLV. 50X1DOS	MSD	229,04 €	11,75
08713184056801	NOBIVAC	CHIEN	NOBIVAC CHP SUSP INJ FL+SOLV. 10X1DOS	MSD		11,75
08714015018593	DURAMUNE	CHIEN	DURAMUNE CHP LYO/SUSP FL 25X1 DOS	ZOETIS		11,75
03660144008880	CHP	BOVIN,CAPRIN,LAPIN,OVIN,PORCIN,VOLAILLE	CHP PDR ORALE SEAU 1X2.5KG	COOPHAVET		11,59

The auto search is triggered whenever a transactional data entry is selected. It displays the best results of the search using google full text search algorithms.

Columns of additional reference data can be selected based on the full list of master data attributes. Additionally data based on recent translations of transactions to master data items such like reference price or date of latest link are available

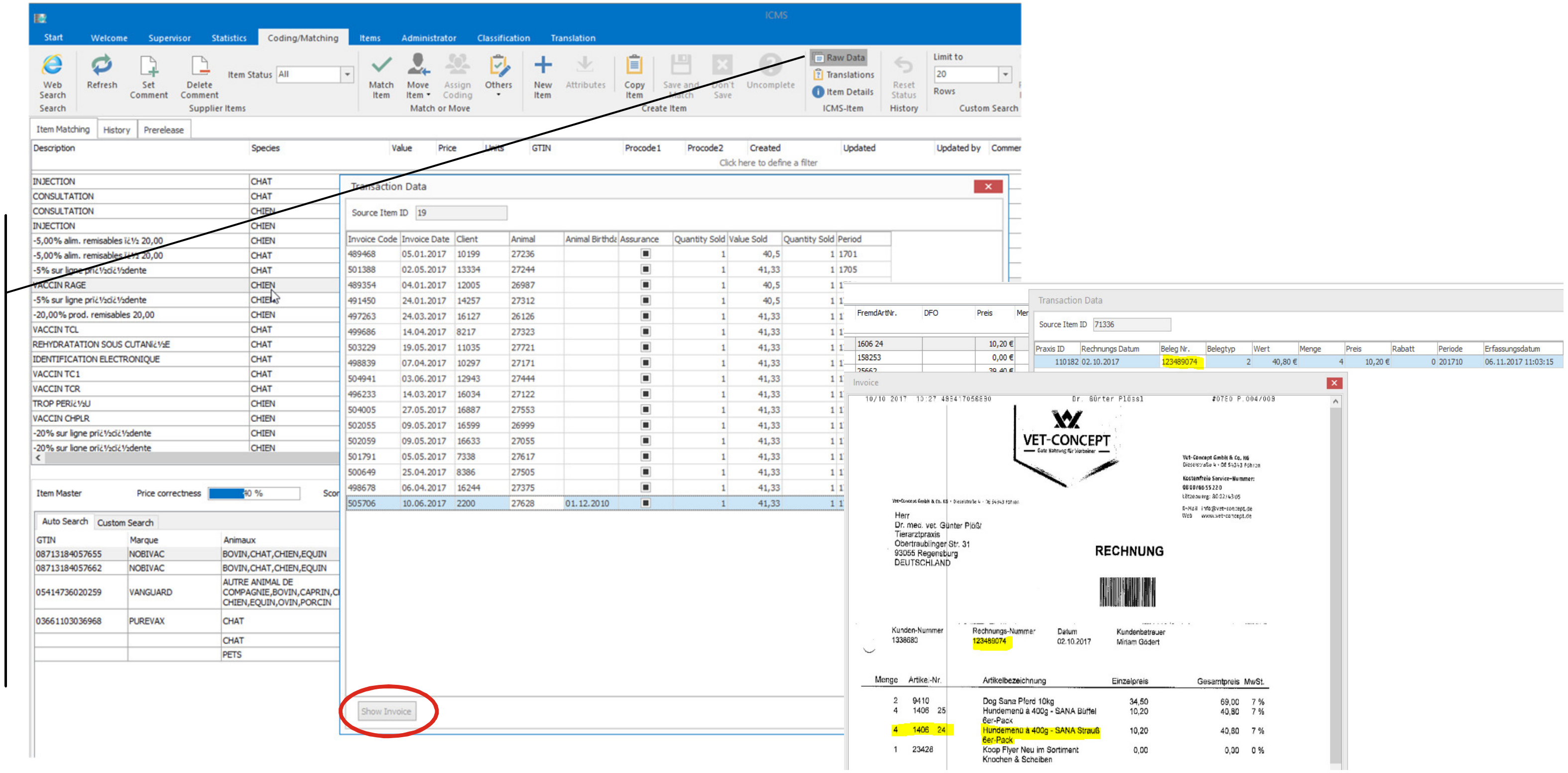
An indicator of the correctness of the price of transaction versus the master data item is available to be able to further qualify translations. In this case the quality it too low to establish a direct link.

Also a quality indicator of the strength of the overlap of the full text search is provided. Here first experiences need to show if an automated matching from a score threshold could be implemented to minimize manual work.

Matching transactional data

Data Source Investigation

A click on Raw Data shows all transactions with the same transaction text. Systems that are based on invoice data can directly show the scan of the invoice with a click on show invoice.



Description	Species
INJECTION	CHAT
CONSULTATION	CHAT
CONSULTATION	CHIEN
INJECTION	CHIEN
-5,00% alim. remisables ï¿½ 20,00	CHIEN
-5,00% alim. remisables ï¿½ 20,00	CHAT
-5% sur ligne priï¿½cï¿½dente	CHAT
VACCIN RAGE	CHIEN
-5% sur ligne priï¿½cï¿½dente	CHIEN
-20,00% prod. remisables 20,00	CHIEN
VACCIN TCL	CHAT
REHYDRATATION SOUS CUTANï¿½E	CHAT
IDENTIFICATION ELECTRONIQUE	CHAT
VACCIN TC1	CHAT
VACCIN TCR	CHAT
TROP PERï¿½U	CHIEN
VACCIN CHPLR	CHIEN
-20% sur ligne priï¿½cï¿½dente	CHIEN
-20% sur ligne priï¿½cï¿½dente	CHIEN

GTIN	Marque	Animaux
08713184057655	NOBIVAC	BOVIN,CHAT,CHIEN,EQUIN
08713184057662	NOBIVAC	BOVIN,CHAT,CHIEN,EQUIN
05414736020259	VANGUARD	AUTRE ANIMAL DE COMPAGNIE,BOVIN,CAPRIN,C CHIEN,EQUIN,OVIN,PORCIN
03661103036968	PUREVAX	CHAT
		CHAT
		PETS

Transaction Data

Source Item ID 19

Invoice Code	Invoice Date	Client	Animal	Animal Birthda	Assurance	Quantity Sold	Value Sold	Quantity Sold	Period
489468	05.01.2017	10199	27236		■	1	40,5	1	1701
501388	02.05.2017	13334	27244		■	1	41,33	1	1705
489354	04.01.2017	12005	26987		■	1	40,5	1	1
491450	24.01.2017	14257	27312		■	1	40,5	1	1
497263	24.03.2017	16127	26126		■	1	41,33	1	1
499686	14.04.2017	8217	27323		■	1	41,33	1	1
503229	19.05.2017	11035	27721		■	1	41,33	1	1
498839	07.04.2017	10297	27171		■	1	41,33	1	1
504941	03.06.2017	12943	27444		■	1	41,33	1	1
496233	14.03.2017	16034	27122		■	1	41,33	1	1
504005	27.05.2017	16887	27553		■	1	41,33	1	1
502055	09.05.2017	16599	26999		■	1	41,33	1	1
502059	09.05.2017	16633	27055		■	1	41,33	1	1
501791	05.05.2017	7338	27617		■	1	41,33	1	1
500649	25.04.2017	8386	27505		■	1	41,33	1	1
498678	06.04.2017	16244	27375		■	1	41,33	1	1
505706	10.06.2017	2200	27628	01.12.2010	■	1	41,33	1	1

FremdArtNr.	DFO	Preis	Mer
1606 24		10,20 €	
158253		0,00 €	
25662		39,40 €	

Transaction Data

Source Item ID 71336

Praxis ID	Rechnungs Datum	Beleg Nr.	Belegtyp	Wert	Menge	Preis	Rabatt	Periode	Erfassungsdatum
110182	02.10.2017	123489074	2	40,80 €	4	10,20 €	0	201710	06.11.2017 11:03:15

Invoice

10/10 2017 10:27 493417056890 Dr. Günter Plössl #0780 P.004/009

VET-CONCEPT
Gute Nahrung für Vierbeiner

Vet-Concept GmbH & Co. KG
Dieselstraße 4 · DE 54343 Föhren

Kostenfreie Service-Nummer:
0800/6655220
Ltzeduerg: 80021305

E-Mail info@vet-concept.de
Web www.vet-concept.de

Vet-Concept GmbH & Co. KG · Dieselstraße 4 · DE 54343 Föhren

Herr
Dr. med. vet. Günter Plößl
Tierarztpraxis
Obertraublinger Str. 31
93055 Regensburg
DEUTSCHLAND

RECHNUNG

Kunden-Nummer	Rechnungs-Nummer	Datum	Kundenbetreuer
1338680	123489074	02.10.2017	Miriam Gödert

Menge	Artikel-Nr.	Artikelbezeichnung	Einzelpreis	Gesamtpreis	MwSt.
2	9410	Dog Sana Pferd 10kg	34,50	69,00	7 %
4	1406 25	Hundemenü à 400g - SANA Büffel 6er-Pack	10,20	40,80	7 %
4	1406 24	Hundemenü à 400g - SANA Strauß 6er-Pack	10,20	40,80	7 %
1	23428	Koop Flyer Neu im Sortiment Knochen & Scheiben	0,00	0,00	0 %

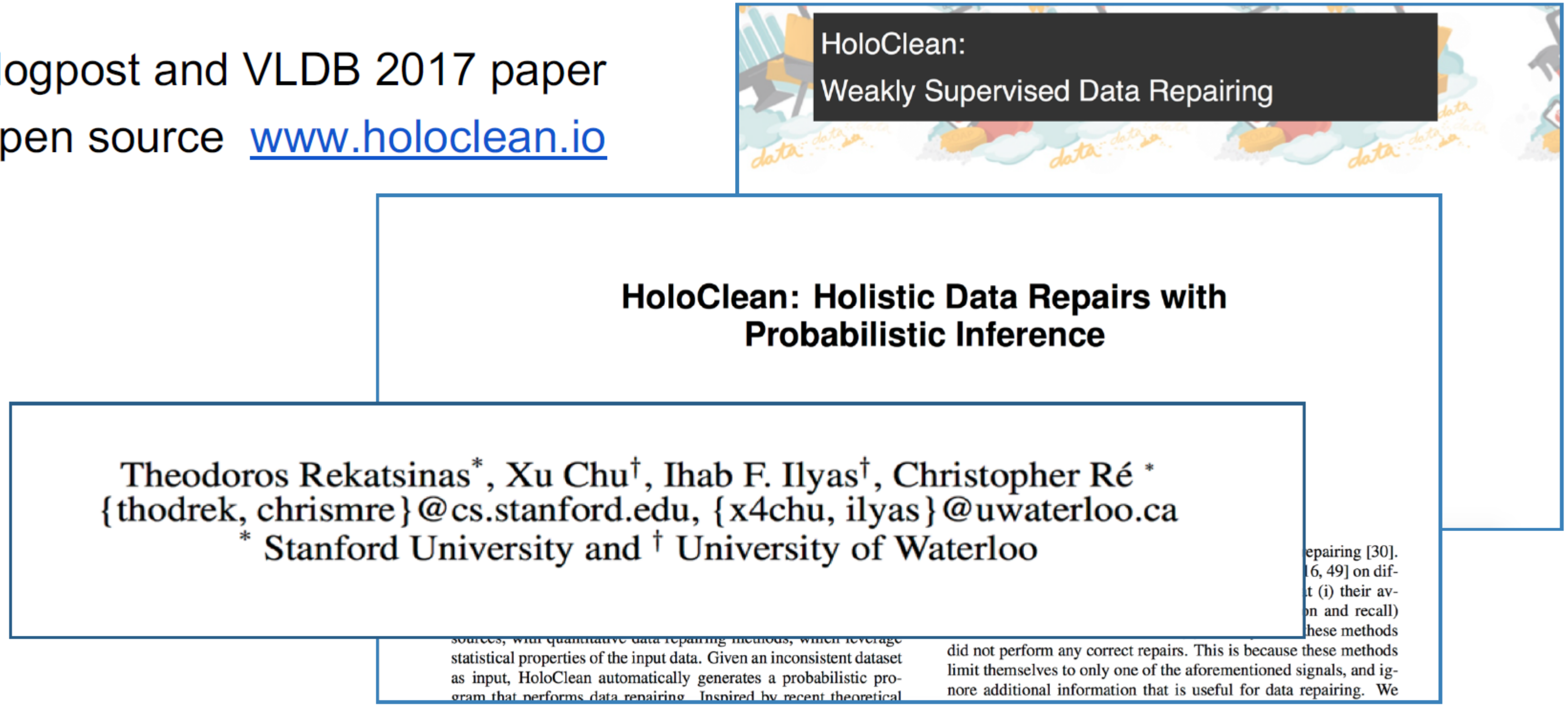
Outline

- Kynetec business overview
- Panel-based market research
- Current data production workflow
  - Preparation of incoming data for matching
  - Master data coding
  - Transactional data matching
- ➡ A Use Case for Machine Learning?!

HoloClean

Repair as an Inference Problem

- Blogpost and VLDB 2017 paper
- Open source www.holoclean.io



HoloClean:

Weakly Supervised Data Repairing

HoloClean: Holistic Data Repairs with Probabilistic Inference

Theodoros Rekatsinas[*], Xu Chu[†], Ihab F. Ilyas[†], Christopher Ré [*]
{thodrek, chrismre}@cs.stanford.edu, {x4chu, ilyas}@uwaterloo.ca
[*] Stanford University and [†] University of Waterloo

sources, with quantitative data repairing methods, which leverage statistical properties of the input data. Given an inconsistent dataset as input, HoloClean automatically generates a probabilistic program that performs data repairing. Inspired by recent theoretical

epairing [30]. 16, 49] on dift (i) their avon and recall) these methods did not perform any correct repairs. This is because these methods limit themselves to only one of the aforementioned signals, and ignore additional information that is useful for data repairing. We

HoloClean [VLDB’17]

Repair as an Inference Problem

Input

Dataset to be cleaned

	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60609
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60609
t4	Johnnyo’s	3465 S Morgan ST	Cicago	IL	60608

Denial Constraints

c1: DBAName → Zip

c2: Zip → City, State

c3: City, State, Address → Zip

Matching Dependencies

m1: Zip = Ext_Zip → City = Ext_City

m2: Zip = Ext_Zip → State = Ext_State

m3: City = Ext_City ∧ State = Ext_State∧
∧ Address = Ext_Address → Zip = Ext_Zip

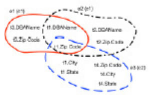
External Information

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608
1208 N Wells ST	Chicago	IL	60610
259 E Erie ST	Chicago	IL	60611
2806 W Cermak Rd	Chicago	IL	60623

The HoloClean Framework

1. Error Detection Module

- Use integrity constraints
- Leverage external data
- Detect outliers
- Identify possible repairs



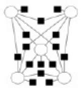
2. Compilation Module

- Automatic Featurization
- Statistical analysis and candidate repair generation
- Compilation to factors/tensors

3. Repair Module

PYTORCH

- Ground probalistic model
- Statistical learning (weights)
- Probabilistic inference



Output

Proposed Cleaned Dataset

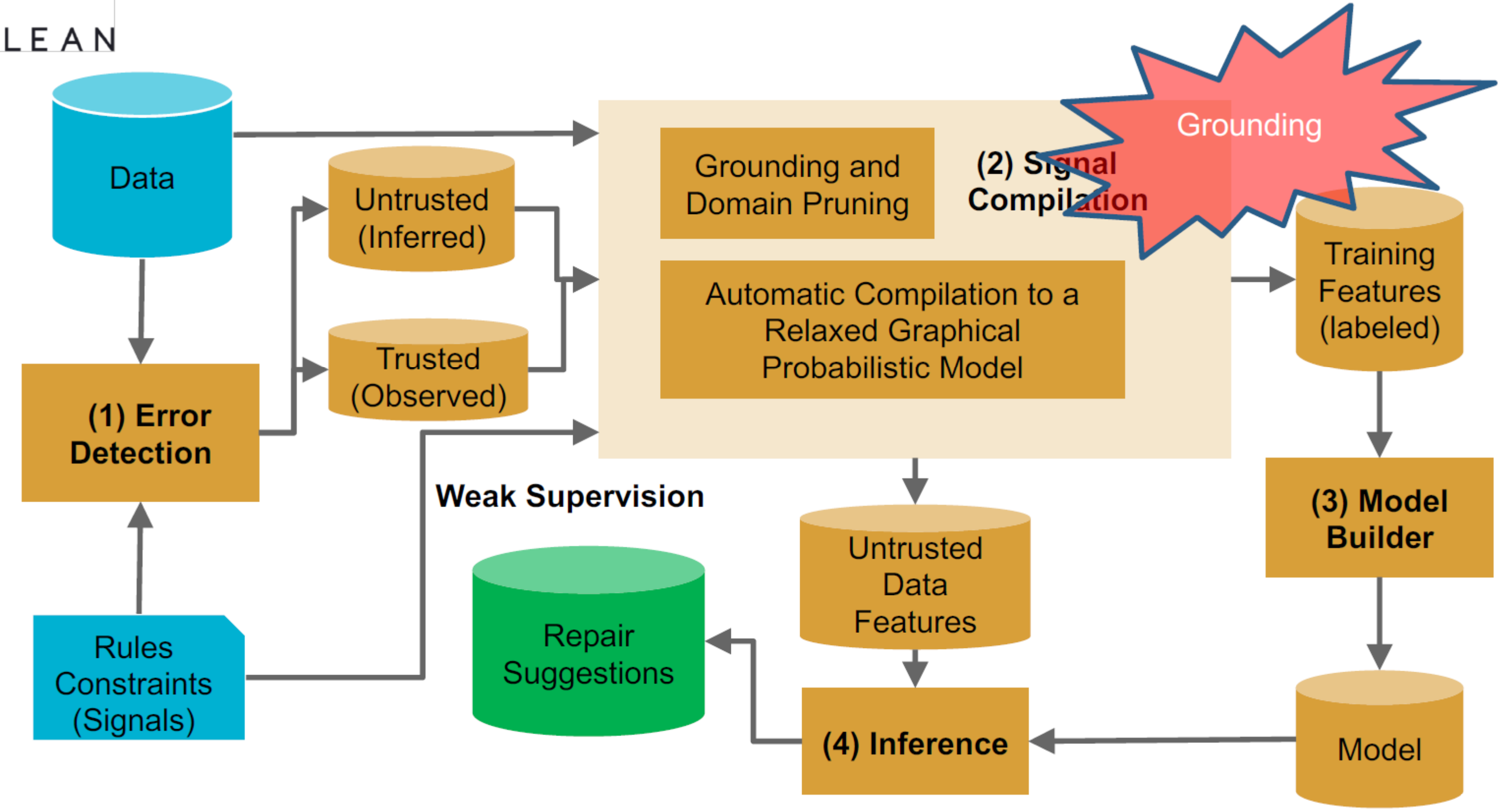
	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t4	**John Veliotis Sr.**	3465 S Morgan ST	**Chicago**	IL	60608

Marginal Distribution of Cell Assignments

Cell	Possible Values	Probability
t2.Zip	60608	0.84
	60609	0.16
t4.City	Chicago	0.95
	Cicago	0.05
t4.DBAName	John Veliotis Sr.	0.99
	Johnnyo’s	0.01

UNIVERSITY OF WATERLOO

HoloClean Architecture



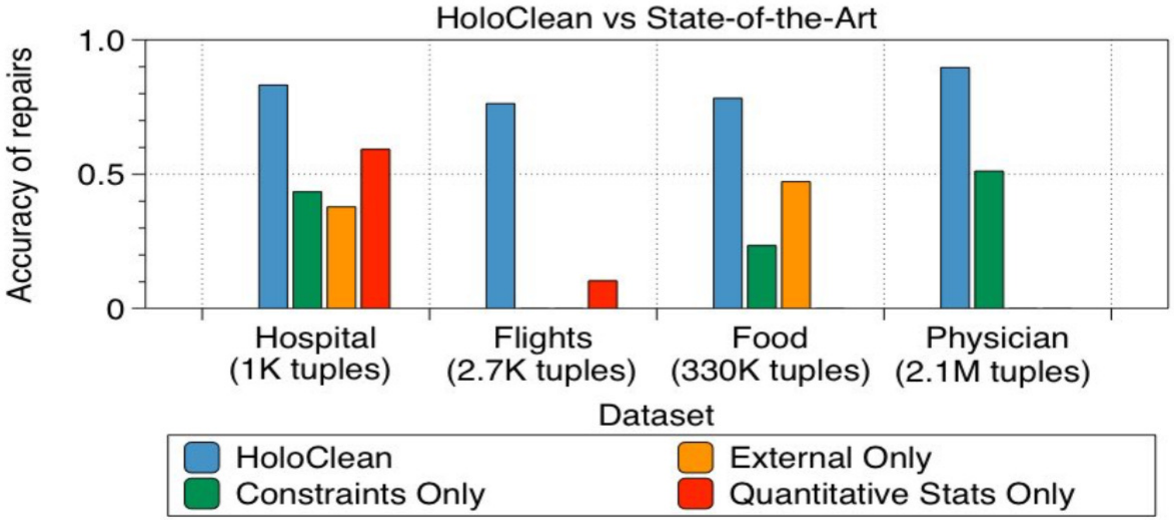
HoloClean Key Ideas

- ***Domain Pruning***: Limit the active domain for random variable
  - Values that appear in the table
  - Use co-occurrence to prune even more aggressively
  - Borrow ideas from missing value imputation

- ***Tying Weights:*** *Learn one weight per constraint not per factor (Templated MLNs)*

- ***Constraint Relaxation***: Relax constraints over sets of random variables to features over independent random variables.

HoloClean Results



HoloClean: our approach combining all signals and using inference
Holistic[Chu,2013]: state-of-the-art for constraints & minimality
KATARA[Chu,2015]: state-of-the-art for external data
SCARE[Yakout,2013]: state-of-the-art ML & qualitative statistics

Results of the application of HoloClean to Kynetec data matching

See you next year…