University of BRISTOL

Automatic Identification of Veterinary Medicines in Treatment Records

Jon Massey MSc
Bristol Veterinary School

Why

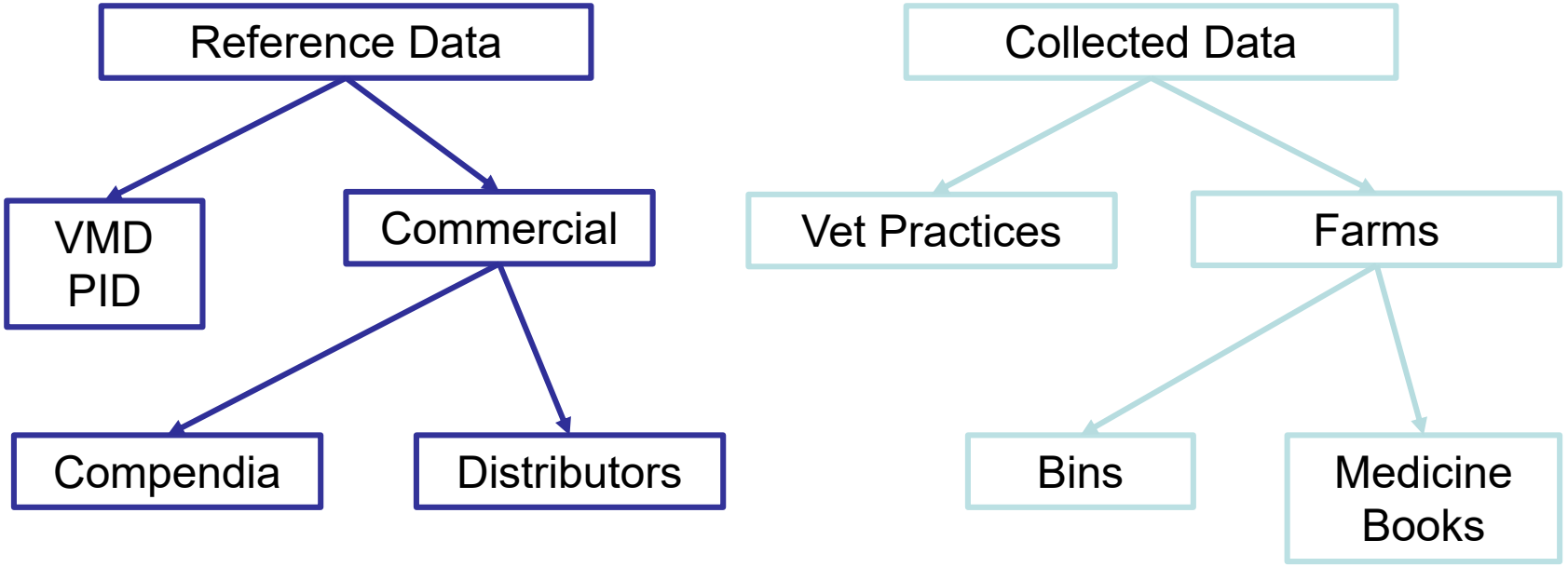
- Surveillance
- Research

- Manual entry slow & error prone

Veterinary Medicines in UK

- Regulated by Veterinary Medicines Directorate (VMD) – executive agency of DEFRA
- **POM-V/POM-VPS**/AVM-GSL
- Can only be prescribed by a vet to an animal *under their care*
- Sale to, administration by farmers

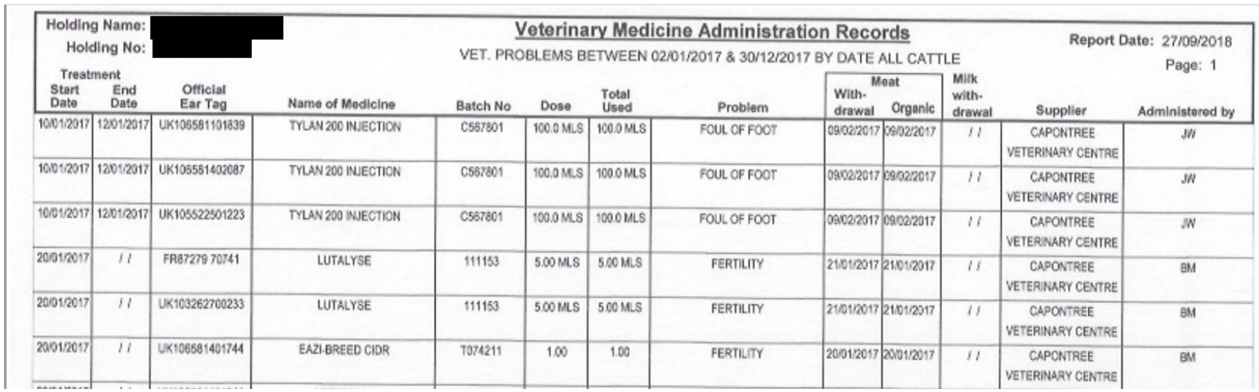
Vet Medicine Data



Collected Medicine Data

- Amount estimation
  - Unspecified units
  - Proxies for physical quantity (price)
- Entity Recognition
  - IDs (VM No, GTIN, Supplier Code) often not available

Example collected data



Holding Name: [redacted]
Holding No: [redacted]

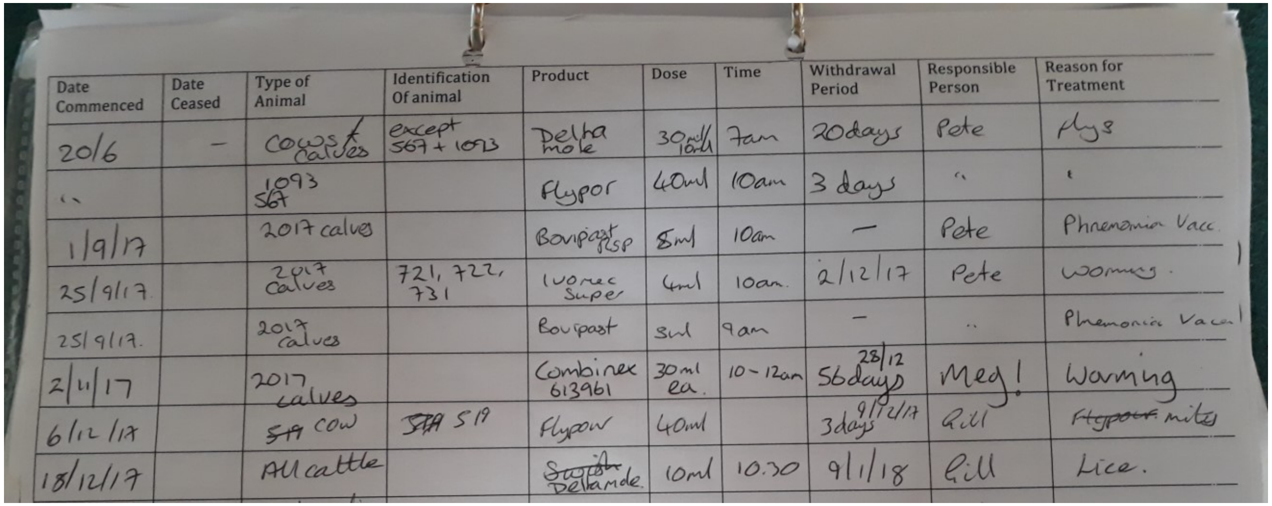
Veterinary Medicine Administration Records

VET. PROBLEMS BETWEEN 02/01/2017 & 30/12/2017 BY DATE ALL CATTLE

Report Date: 27/09/2018
Page: 1

Treatment Start Date	Treatment End Date	Official Ear Tag	Name of Medicine	Batch No	Dose	Total Used	Problem	Meat With-drawal	Meat Organic	Milk with-drawal	Supplier	Administered by
10/01/2017	12/01/2017	UK106581101839	TYLAN 200 INJECTION	C567801	100.0 MLS	100.0 MLS	FOUL OF FOOT	09/02/2017	09/02/2017	/ /	CAPONTREE VETERINARY CENTRE	JW
10/01/2017	12/01/2017	UK106581402087	TYLAN 200 INJECTION	C567801	100.0 MLS	100.0 MLS	FOUL OF FOOT	09/02/2017	09/02/2017	/ /	CAPONTREE VETERINARY CENTRE	JW
10/01/2017	12/01/2017	UK105522501223	TYLAN 200 INJECTION	C567801	100.0 MLS	100.0 MLS	FOUL OF FOOT	09/02/2017	09/02/2017	/ /	CAPONTREE VETERINARY CENTRE	JW
20/01/2017	/ /	FR87279 70741	LUTALYSE	111153	5.00 MLS	5.00 MLS	FERTILITY	21/01/2017	21/01/2017	/ /	CAPONTREE VETERINARY CENTRE	BM
20/01/2017	/ /	UK103262700233	LUTALYSE	111153	5.00 MLS	5.00 MLS	FERTILITY	21/01/2017	21/01/2017	/ /	CAPONTREE VETERINARY CENTRE	BM
20/01/2017	/ /	UK106581401744	EAZI-BREED CIDR	T074211	1.00	1.00	FERTILITY	20/01/2017	20/01/2017	/ /	CAPONTREE VETERINARY CENTRE	BM

Consultation Date	Item Name
03/08/2017	F Bimoxyl La Injection (100)
03/08/2017	METACAM INJ 2% L/A
03/08/2017	Denagard
15/08/2017	Willcain
17/08/2017	Denagard
17/08/2017	Lincocin Sterile Solution
18/08/2017	Engemycin Aerosol 200ml
31/08/2017	Bovipast RSP (10)
31/08/2017	Denagard
31/08/2017	Engemycin Aerosol 200ml
08/09/2017	Bimoxyl La Injection (100)
14/09/2017	Depocillin Injection (100)
14/09/2017	Engemycin Aerosol 200ml
14/09/2017	Lincocin Sterile Solution
14/09/2017	Euthatal Inj 100ml
21/09/2017	Bovipast RSP (10)
29/09/2017	Ketofen 10% Injection (100)
20/10/2017	Lincocin Sterile Solution



Date Commenced	Date Ceased	Type of Animal	Identification Of animal	Product	Dose	Time	Withdrawal Period	Responsible Person	Reason for Treatment
20/6	–	cows + calves	except 567 + 1093	Delta mole	30ml/10ml	7am	20days	Pete	flys
"		1093 567		Flypor	40ml	10am	3 days	"	"
1/9/17		2017 calves		Bovipast RSP	8ml	10am	–	Pete	Phnenomia Vacc.
25/9/17.		2017 calves	721, 722, 731	Ivomec Super	4ml	10am.	2/12/17	Pete	worming.
25/9/17.		2017 calves		Bovipast	3ml	9am	–	"	Phenomia Vacc
2/11/17		2017 calves		Combinex 613961	30ml ea.	10–12am	56days 28/12	Meg!	Worming
6/12/17		~~519~~ cow	~~519~~ 519	Flypow	40ml		3days 9/12/17	Gill	~~Flypour~~ mites
18/12/17		All cattle		~~Swish~~ Dellamde.	10ml	10.30	9/1/18	Gill	Lice.

Vet Medicine Name Matching

- Many names for same medicine
- Abbreviations
- Misspeelings
- Name composed of multiple tokens
  - Semantics of tokens
  - Importance of tokens to match score

Example names

Alamycin LA 200 mg/ml Solution for Injection

Brand

Long Acting

Concentration

Pharmaceutical form

Some of these terms are more useful than others

String similarity metrics

- Edit distances
  - Steve => Stave :1
  - Line => Liner :1
  - Slide => Sledding :5

- Edit based similarity

Multi-token edit distances

- Monge Elkan
  - Tokenise, pairwise Levenshtein, mean similarity of most similar pairs
- Fails to take account of intra- and inter-token positional effects
  - Misspellings/abbreviations affect latter part
  - Latter tokens *usually* less important

Declining importance

Flunixin 50 mg/ml Solution for Injection for Cattle, Horses and Pigs

TETRA DELTA MC 24X10ML (1) [DISCOUNT 15%]

50X I MARBOCARE SOLN FOR INJECTION 100ML (1) [DISCOUNT 15%]

Positionally-weighted Monge Elkan

- Tokenise, pairwise *Jaro-Winkler*, *weighted* mean similarity of most similar pairs
- How sharply to decline waiting?
- How to weight string A token position vs string B token position?

Weighting & parameters

$$Weight_{a/b} = tokenIndex_{a/b}{}^{weightingParameter_{a/b}}$$

$$Weight_{combined} = (W_a \,*(1- P_{ratio})) + (W_b \,*P_{ratio})$$

$$Score_{total} = \frac{\sum_{i=1}^{i_{max}} \frac{S_i}{W_{ci}}}{\sum_{i=1}^{i_{max}} \frac{S_i}{W_{ci}}}$$

Generation	ABWeightRatio	AWeight	BWeight	Threshold	SuccessRate
1	0.125	5.48	2.39	0.697	0.720472441
2	0.16	5.34	0.22	0.711	0.728346457
3	0.1	5.32	1.35	0.681	0.736220472
4	0.629	5.72	0.29	0.703	0.744094488
6	0.1	5.72	0.31	0.683	0.748031496
7	0.116	5.72	0.15	0.701	0.751968504
8	0.165	6.04	0.05	0.683	0.755905512

Positional effect of test string less pronounced – more laconic, only include what's necessary

Results

- 780 treatment/sale records
- Mixed sources – dairy farms, beef farms, veterinary software
- 72% accuracy
- Domain-specific parameters required

Using semantics as matching features

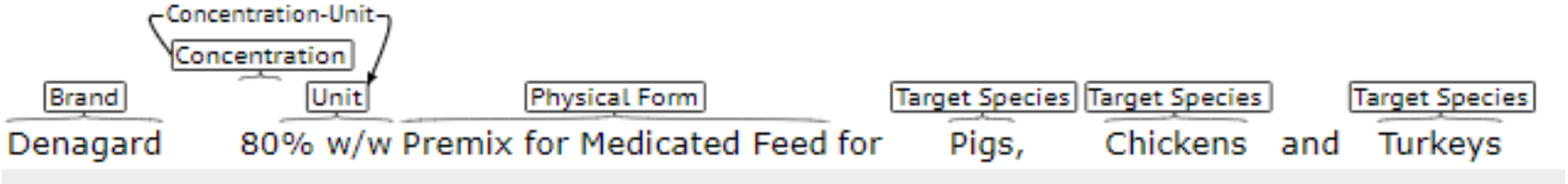
- Token index is only rough proxy for semantics
- Assumptions re: token order
- What if we could use explicit semantics?

Relevant Natural Language Processing Techniques

- Part-of-speech tagging
  - *What sort of word is this*
- Named Entity Recognition
  - *To what sort of thing does this refer?*

Annotating training data

- All currently licensed antimicrobials n=560



- Brand
- Concentration
- Unit
- Sub-Brand
- Duration-of-Action
- Lactation-Phase
- Physical-Form
- Target-Species
- Active-Ingredient

Conditional Random Field Classifier

- Statistical model
- Sequence model – state of a sample + neighbours
- Good at predicting labels for sequential data

Feature function

- Proper case
- First/last word in sentence
- Previous/Next word
- [letter][punctuation][letter]
- Numeric
- Alphameric
- Prefixes/suffixies

PoS tagging AM Names

- 80:20 training:test

Tag	precision	recall	f1-score	support
Brand	1.000	1.000	1.000	112
Sub-Brand	0.875	0.808	0.840	26
Concentration	0.966	0.977	0.971	87
Unit	1.000	1.000	1.000	107
Physical-Form	0.989	1.000	0.994	261
Ignore	0.994	1.000	0.997	174
Target-Species	1.000	0.992	0.996	131
Lactation-Phase	1.000	1.000	1.000	9
Active-Ingredient	1.000	0.600	0.750	5
Duration-of-Action	1.000	1.000	1.000	4
---	---	---	---	---
accuracy	-	-	0.989	916
macro-avg	0.982	0.938	0.955	916
weighted-avg	0.989	0.989	0.989	916

Tagging non-AM Names

- <Brand>Milbemax</Brand> <Physical-Form>Chewable</Physical-Form> <Physical-Form>Tablets</Physical-Form> <Ignore>for</Ignore> <Target-Species>Dogs</Target-Species> <
- <Brand>CLiKZiN</Brand> <Concentration>12.5</Concentration> <Unit>mg/ml</Unit> <Physical-Form>Pour-On</Physical-Form> <Physical-Form>Suspension</Physical-Form> <Ignore>for</Ignore> <Target-Species>Sheep</Target-Species>
- <Brand>Prid</Brand> <Sub-Brand>Delta</Sub-Brand> <Concentration>1.55</Concentration> <Unit>g</Unit> <Physical-Form>Vaginal</Physical-Form> <Physical-Form>Delivery</Physical-Form> <Physical-Form>System</Physical-Form> <Ignore>for</Ignore> <Target-Species>Cattle</Target-Species>
- <Brand>Pracetam</Brand> <Sub-Brand>200</Sub-Brand> <Unit>mg/ml</Unit> <Physical-Form>Solution</Physical-Form> <Ignore>for</Ignore> <Target-Species>Use</Target-Species> <Ignore>in</Ignore> <Physical-Form>Drinking</Physical-Form> <Physical-Form>Water</Physical-Form> <Ignore>for</Ignore> <Target-Species>Pigs</Target-Species>

Tagging real-world data

- <Brand>6X</Brand> <Sub-Brand>IHYMATIL</Sub-Brand> <Concentration>300MG/ML</Concentration> <Unit>INJ</Unit> <Concentration>100ML</Concentration> <Unit>(1)</Unit> <Concentration>[DISCOUNT</Concentration> <Concentration>15%]</Concentration>
- <Brand>7X</Brand> <Sub-Brand>IENZAPROST-T</Sub-Brand> <Concentration>5ML</Concentration> <Unit>DOSE</Unit> <Concentration>(1)</Concentration> <Unit>[DISCOUNT</Unit> <Concentration>15%]</Concentration>
- <Brand>8X</Brand> <Sub-Brand>UCEFIMAM</Sub-Brand> <Lactation-Phase>DC</Lactation-Phase> <Target-Species>DRY</Target-Species> <Target-Species>COW</Target-Species> <Target-Species>TUBES</Target-Species> <Target-Species>(1)</Target-Species>
- <Brand>8X</Brand> <Sub-Brand>UCEFIMAM</Sub-Brand> <Lactation-Phase>DC</Lactation-Phase> <Target-Species>DRY</Target-Species> <Target-Species>COW</Target-Species> <Target-Species>TUBES</Target-Species> <Concentration>4'S</Concentration>
- <Brand>8X</Brand> <Sub-Brand>UCEPRAVIN</Sub-Brand> <Lactation-Phase>DC</Lactation-Phase> <Concentration>20'S</Concentration> <Unit>(1)</Unit>
- <Brand>8X</Brand> <Sub-Brand>UORBESEAL</Sub-Brand> <Lactation-Phase>DC</Lactation-Phase> <Target-Species>SYRINGE</Target-Species> <Concentration>1'S</Concentration>

Word embeddings

- “Gazetteer” of semantic tags
- terms associated with them
- “Distance” between them

- Encoding set of information about a word into numbers

Nearest Neighbour

- Word => “place on many-dimensional map”

- Promising, but slow
- Results TBC
- Can be fooled if extraneous tags are close to tags present in gazetteer

Next steps

- Train classifier to find most similar word embeddings
- Use CRF to train not just “what sort of entity” but “which specific entity”?

Curse of Dimensionality

- Lots of features
- Lots of classes

- Not a lot of training data
- Synthetic data
- Dimensionality reduction

Synthetic data

- Same statistical distribution of features as real data
- Term subsets
- Term addition
- Term mutation (mis-spellings)
- Null records
- Additional 3050 elements so far

Thank you!