FIT5145 Introduction to Data Science

Module 5

Data Analysis Process

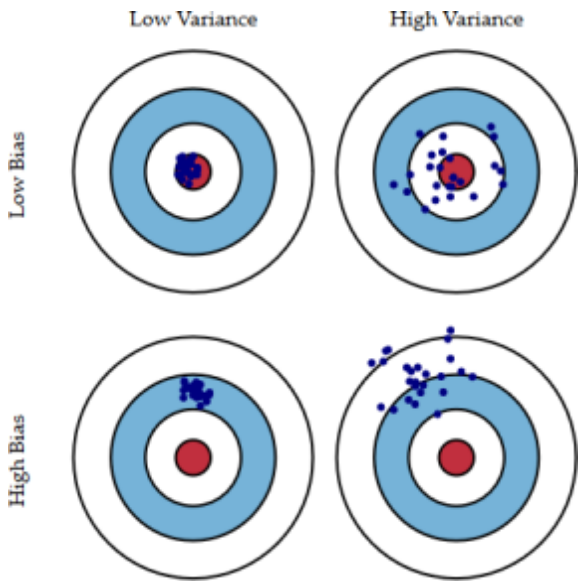
2019 Lecture 10

Monash University

Discussion: Bias Variance

From Wikipedia:

- The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).



Discussion: Investigating Twitter data in the shell

We have analysed a **large data file** from Twitter in the shell during the tutorial:

- Aim: understand what data the file contained, how we could reformat the data for further analysis
- Many **different types of columns:**
  - text, dates, locations, even code containing data structures
- real data: lots of missing data, errors, ...
- shell commands like *grep* and *cut* simplify the inspection and manipulation of the data

Unit Schedule: This Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5	**9**	**data analysis theory**
	10	**data analysis process**
6.	11	issues in data management
	12	GUEST SPEAKER & EXAM INFO

Online Experiments

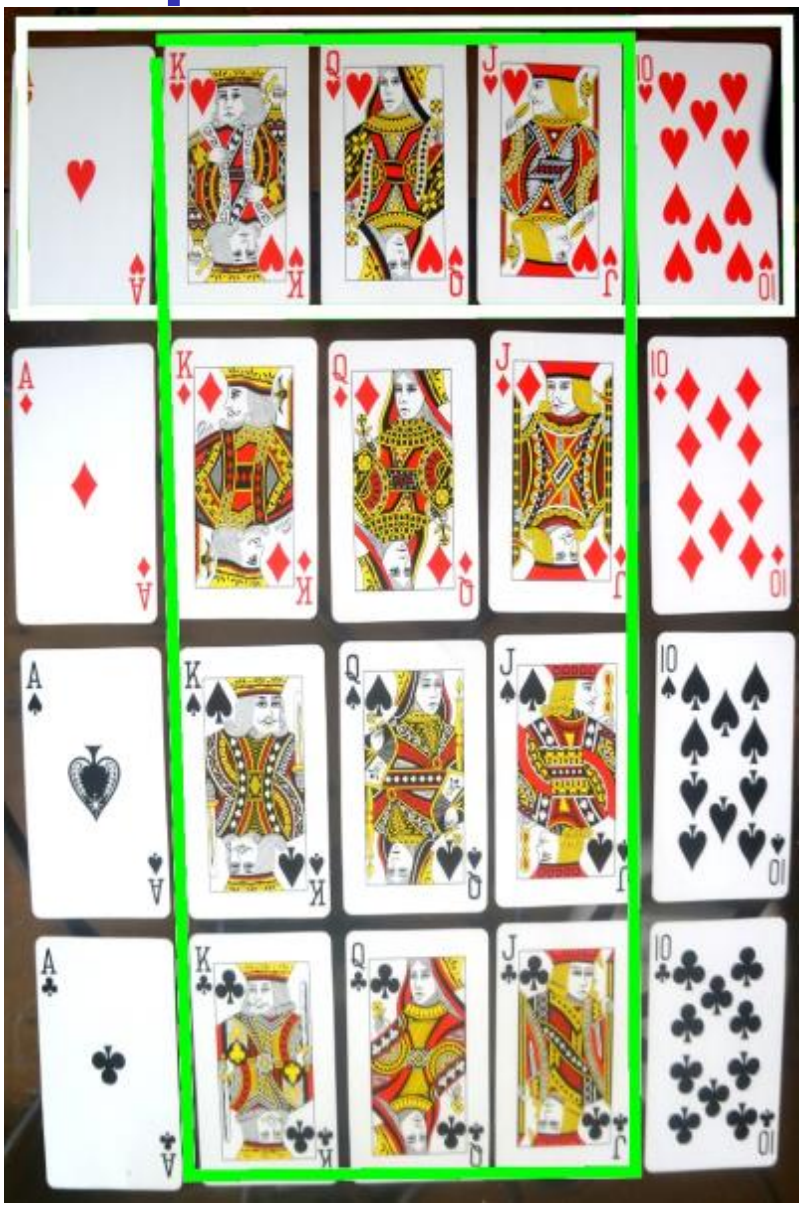
The need for A/B Testing

dependence and correlation between variables
correlation does not imply causation
A/B tests

Dependence

- statistical notion of *independence* says:
  - event A and event B are independent if knowing whether A occurred provides no information about whether B will occur or not
    - i.e. knowing A doesn't change the probability of B
  - variables are dependent if they are "not independent"
- Examples:
  - "rained last night" and "lawn is wet this morning" are dependent
  - "rained last night" and "ate cornflakes for breakfast" are independent
- Note that dependence can vary with context
  - see next slide for an example

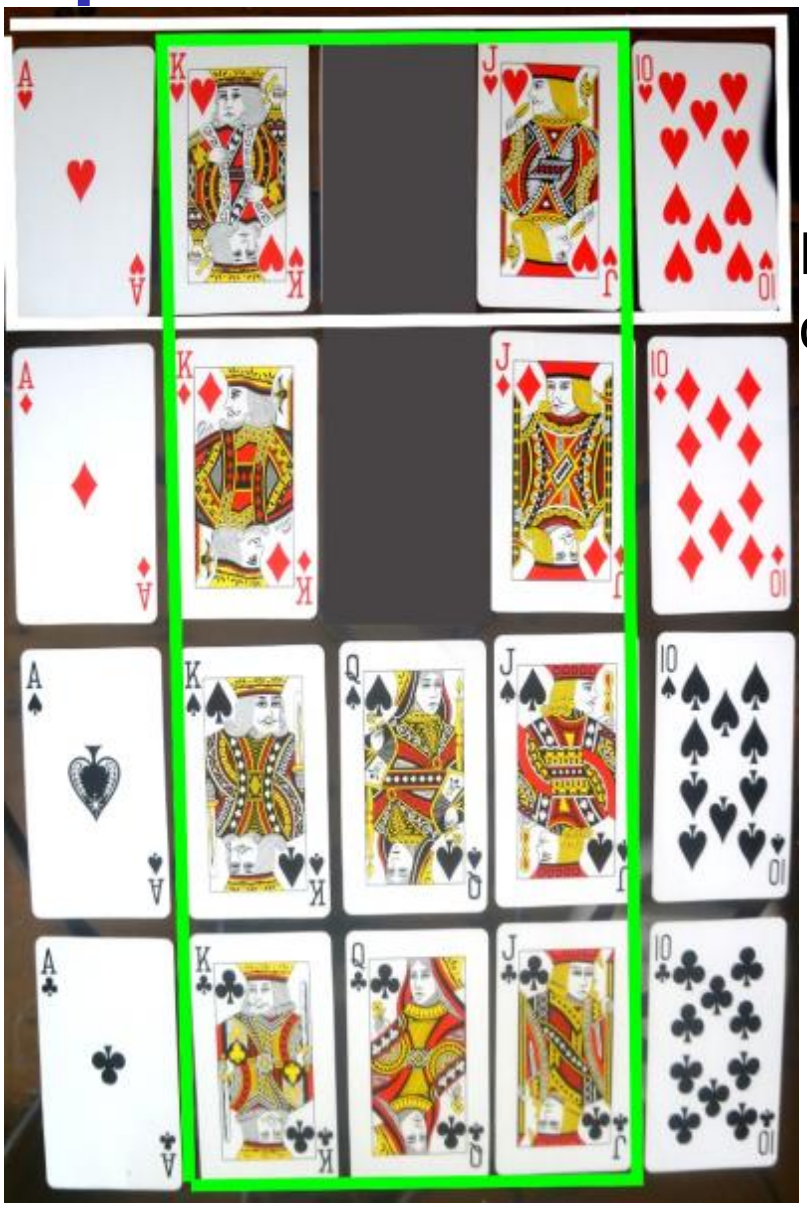
Independence



Example of Independence:

- green boundary is "royalty"
- white boundary is "hearts"
- "royalty" and "hearts" are independent
  - since the relative proportions of inside/outside green boundary are the same irregardless of whether you are inside/outside white boundary
  - so probability of green doesn't change whether we know white or not.

Dependence



By removing the two red queens, we create an example of Dependence:

- green boundary is "royalty"
- white boundary is "hearts"
- "royalty" and "hearts" are now dependent
  - relative proportions of inside/outside green boundary now change whether you are inside/outside white boundary

Correlation

- if variables are continuous, i.e. real valued rather than binary, dependence between the variables is usually referred to as “correlation”
- statistical notion of *correlation* usually measures the “linear” dependence between variables

Causality

- a relationship between an event (the cause) and another event (the effect), where the cause is responsible for the effect
- long history in philosophy and physics, see *Causality*
- measured by intervention:

  > *To test if A causes B, we hold every other variable fixed, then force A to have a certain value in an observed population, and observe B. If, in this situation, B is dependent on A (or B is changed by A), then it follows that A causes B.*

- intervention forces A to be a given value, then we observe B
- fundamental concept for *clinical trials* in science/medicine

FLUX Question

Correlation does not imply causation.

A. TRUE

B. FALSE

Correlation does not imply Causation

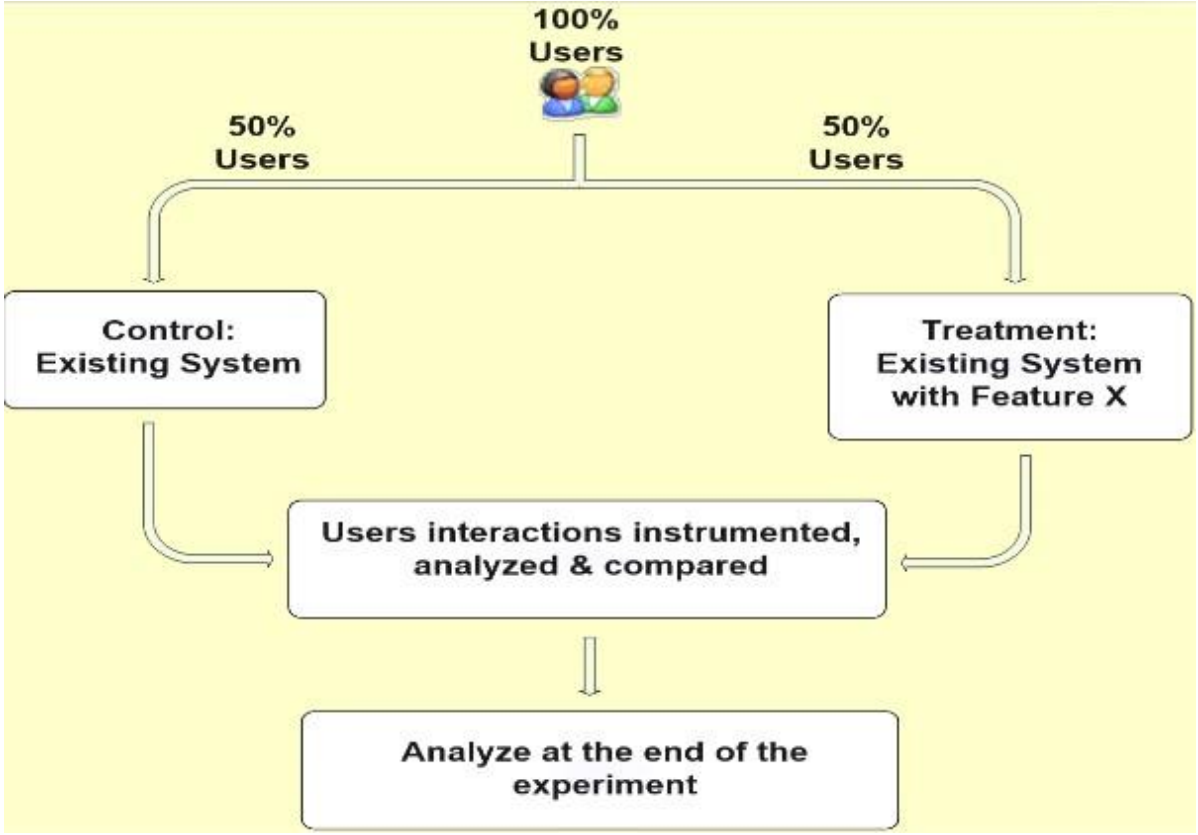
See *correlation does not imply causation* (Wikipedia).

Examples of the fallacy:

> *The faster windmills are observed to rotate, the more wind is observed. Therefore wind is caused by the rotation of windmills.*
>
> *Sleeping with one's shoes on is strongly correlated with waking up with a headache. Therefore, sleeping with one's shoes on causes headaches.*

See also *Spurious Correlations* for more.

Controlled Experiments



- crucial for internet companies testing alternatives
- see also *A/B testing* and *clinical trials* (adds blind controls)

Preprocessing Data
For building a Predictive Model

imputing missing values

Imputation

ID	Age	Amount	Duration	Job	Housing	Marital	Default
001	43	$200,000	240	A	apartment	yes	no
002	27	$150,000	280	A	apartment	no	?
003	?	$180,000	240	B	house	yes	no
004	42	$200,000	240	?	apartment	yes	no
005	31	$300,000	240	C	house	yes	no

- here we have the housing loan prediction problem
- record 002 has the target variable (*Default*) missing
  - cannot be used by standard learning algorithms
- record 003 has *Age* missing, record 004 has *Job* missing
  - if we "fill in" the missing variables using imputation then these records can be used

Theory of Data Analysis

Characterizing Learning

broad characterisations for general discussion

Characterizing Learning

Prediction: Is the task a simple prediction?

Dynamic: Does the task repeat over space or time? (GPS, game playing)

Missing data: Do some of the variables missing have missing data? (note they cannot be 100% missing)

Latent variables: Are there latent variables? *e.g.*, a segmentation task. Note the target variable for a prediction task cannot be latent.

Optimisation: Does evaluation/prediction require optimisation *after* statistical inference (*i.e.* after prediction)?

latent variable ::= variable whose value never appears in any data

Types of Data Analysis

"Six types of analyses every data scientist should know", by Jeffrey Leek

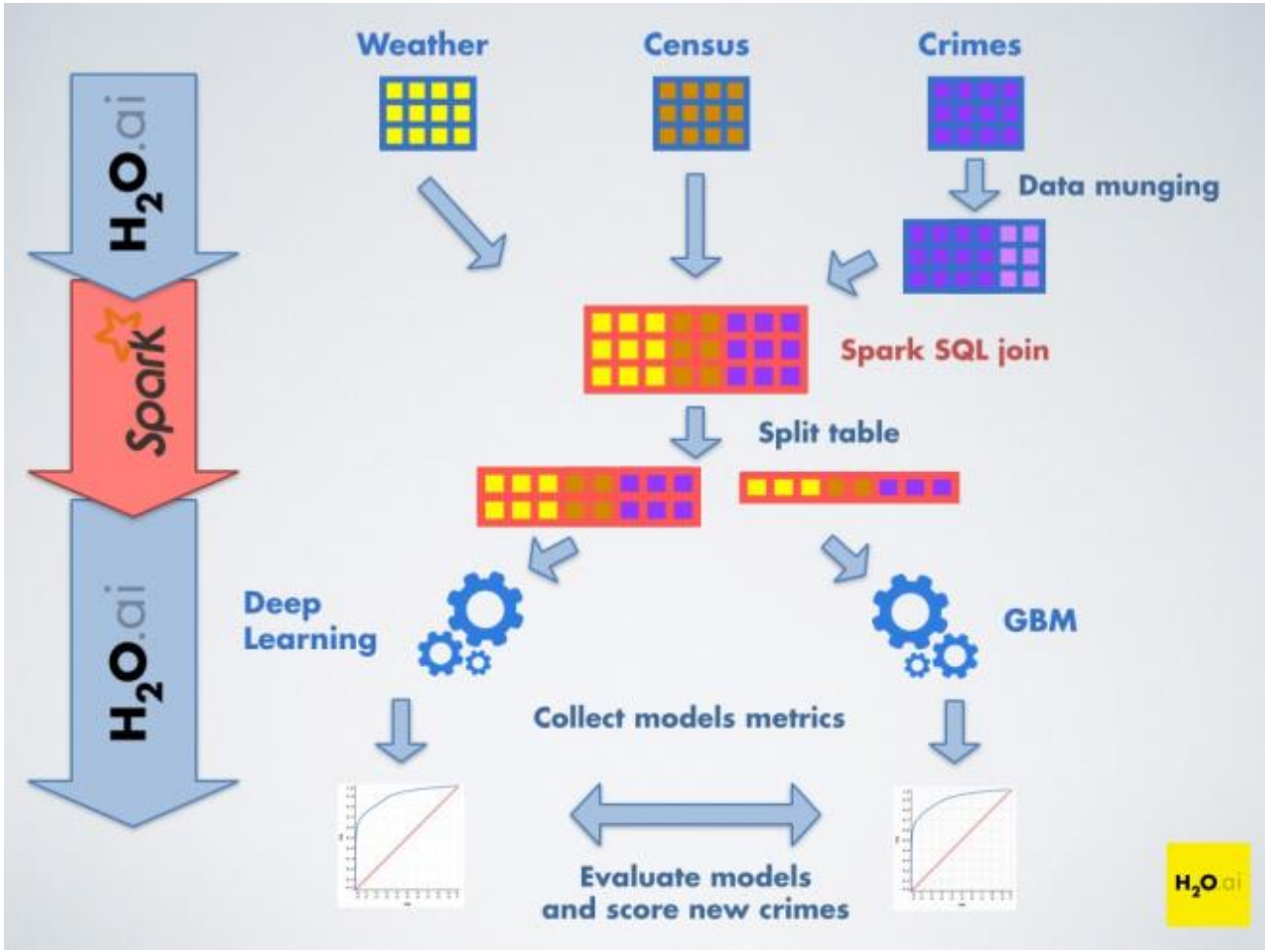
1. Descriptive (quantitatively describe data)
2. Exploratory (explore relationships between variables)
3. Inferential (infer values of unknown variables)
4. Predictive (predict future values)
5. Causal (determine if a causal relationship exists)
6. Mechanistic (explain causal relationships)

- extends SAS's "analytic levels" with a some nuances
- introducing inference and causality
- the level of characterisation we want for the Assessment

Tools for the Data Analysis Process
(ePub section 5.4)

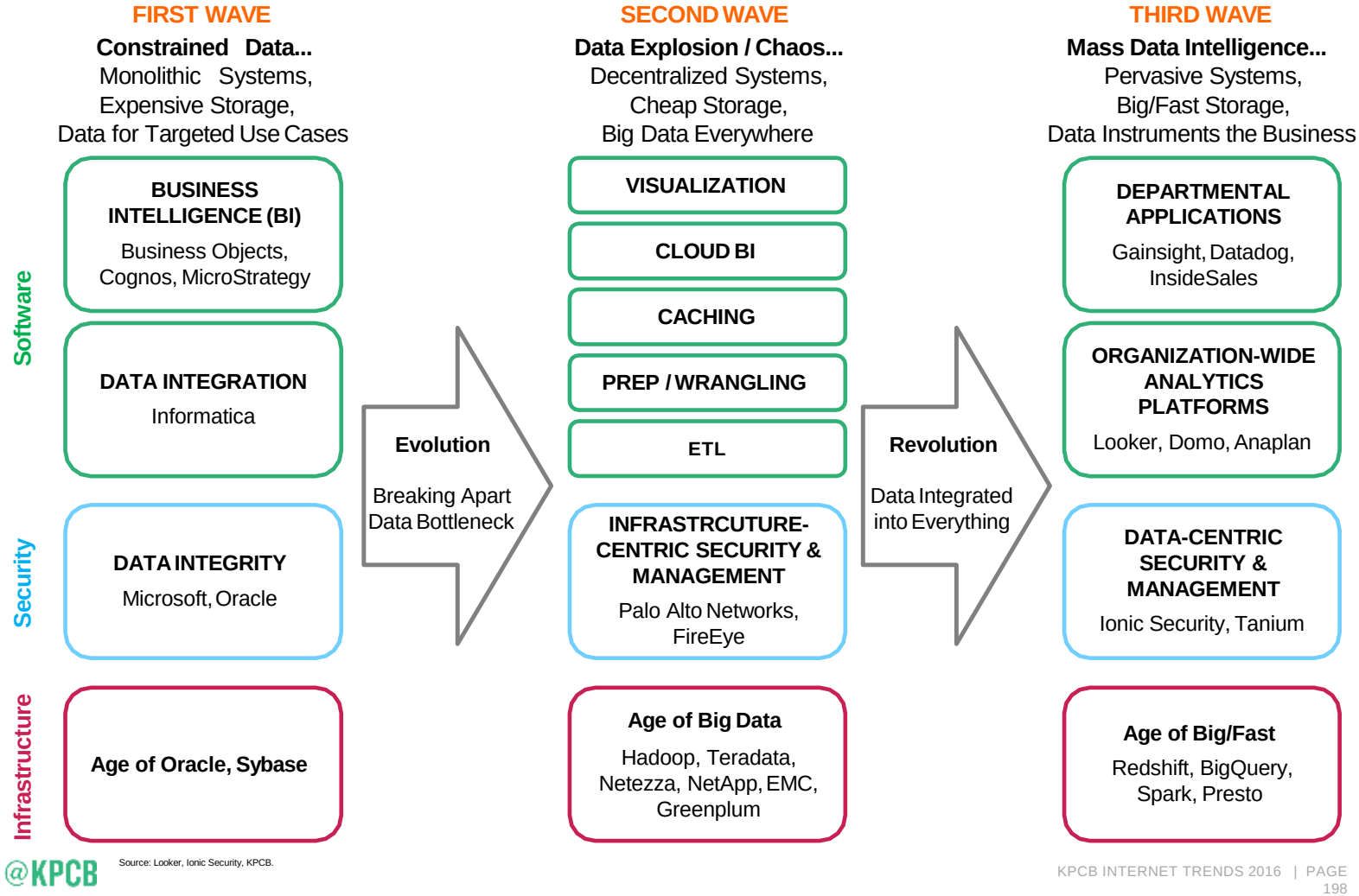
popular software and prototyping

H2O ML Platform



Rapid, reliable engineering of data analysis requires good platforms and systems. Shown is *ML platform for H2O*.

Evolution of the Data Platform, 1990 – 2016



Common Software

access: SQL, Hadoop, MS SQL Server, PIG, Spark

wrangling: common scripting languages (Python, Perl)

visualisation: Tableau, Matlab, Javascript+D3.js

statistical analysis: Weka, SAS, R

multi-purpose: Python, R, SAS, KNIME, RapidMiner

cloud-based: Azure ML (Microsoft), AWS ML (Amazon)

KDnuggets on the R vs. Python debate

Mapping Big Data

See *"Mapping Big Data: A Data-Driven Market Report"* by Russell Jurney, published by O'Reilly 2015. See Table 1-6.

Cluster	**Company**
Old Data Platforms	IBM, Microsoft, Oracle, Dell, Netapp
Servers	Intel, SUSE, MSC Software, NVidia
Analytic Tools	Tableau, Teradata, Informatica, Talend, Actian
New Data Platforms	Cloudera, Hortonworks, MapR, Datastax, Pivotal
Enterprise Software	HP, SAP, Cisco, VMWare, EMC
Cloud Computing	Amazon Web Svcs., Google, Rackspace

Note the Enterprise Software segment developing good connections with all others, but already has strong connections with Old Data Platforms.

Machine Learning Platforms

Most of the big internet companies offer cloud-based machine learning tools and processing pipelines. Open source projects also offer ML tools.

- *Amazon Web Services ML*
- *Google Cloud ML*, note they are heavily invested in deep neural networks
- *Scikit-Learn* is Python based on standard scientific Python libraries
- *The R Project*
- *TensorFlow™* software for CPU/GPU dataflow, especially deep neural networks
- *Apache Mahout* is a distributed linear algebra framework
- also many startups looking at the space

Scripting Languages

see Wikipedia entry *scripting languages*:

- no formal or universally agreed definition
- often interpreted and are high-level programming languages
- automating tasks originally done one-by-one by hand
- also, extension language, control language

e.g. bash, Perl, Python, R, Matlab, ...

kinds: glue languages (connecting software components), GUI scripting, job control, macros, extensible languages, application specific, ...

- an *endless discussion on StackExchange*

Rapid Prototyping

see Wikipedia entry *software prototyping*:

- software development for data science projects is often (almost) one-off ... get the results, but ensure it is reproducable
- not standard software engineering, not “waterfall model”, not “agile”
- little requirements analysis
- the results are tested, not the software and its full capability
- development speed and agility are important
- hence use of scripting languages

Rapid Prototyping Examples

- putting together a processing pipeline
- testing out different alternatives
- trying “cheap hacks” for data cleaning to test ideas before investing more effort
- glueing in custom software that might be difficult or particular to use

  e.g. natural language or image processing tools or web services
- running what is usually GUI or internet API systems in command line mode
- modifying/restarting a processing pipeline

FLUX Question

Scripting languages are ideal for rapid prototyping, so are often used for it.

A. TRUE

B. FALSE

Data Analysis
Meta Case Studies
(ePub section 5.7)

general considerations about data analysis

- Google flu trends
  - case study on use of proxy data
- scientific method
  - is Data Science writ large, so what can we learn
- drug interactions
- what is hard?
  - the machine learning or statistical step is often not

Data Analysis Meta Case Studies
Google flu trends

case study on use of proxy data

Google Flu Trends

Google Flu Trends (2min YouTube video)

- U.S. Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS) provide data with 2 week lag
- CDC has 9 surveillance regions in US and report "influenza-like illness" (ILI) visits weekly
- Google researchers
  - selected top 45 queries that predicted *ILI visits* across regions in 2003-2008
  - built a linear model on these 45 queries to predict *ILI* visits 2003-2007
  - tested it against 2007-2008 data

Google Flu Trends, cont.

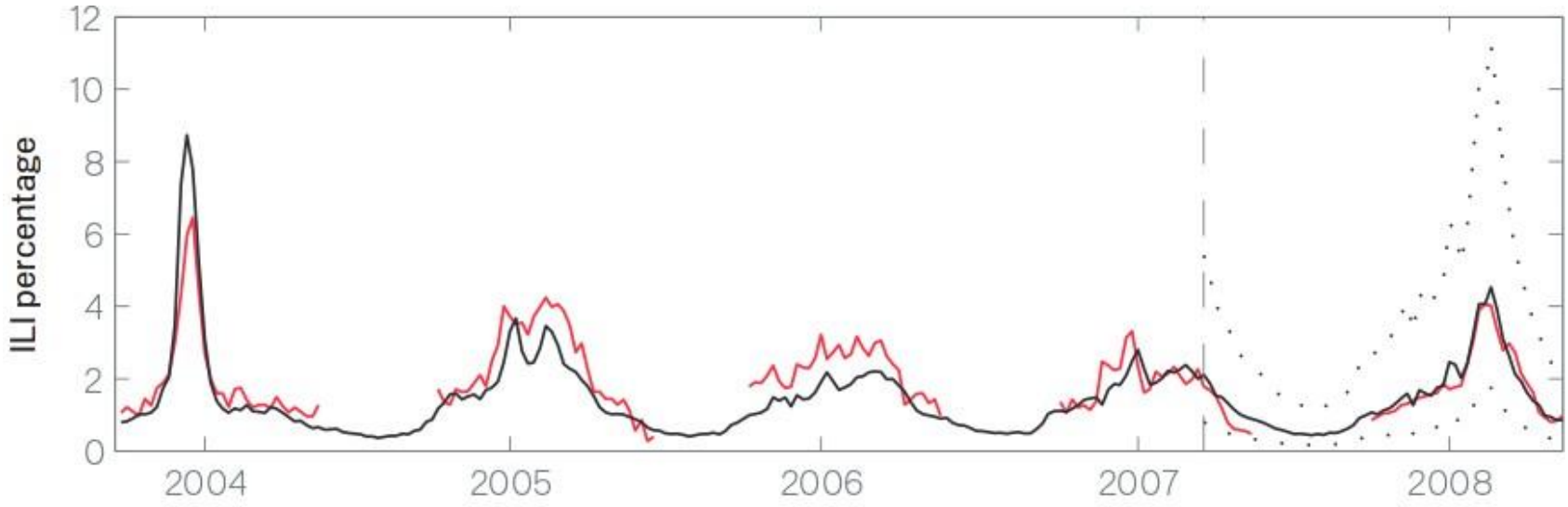


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Google Flu Trends: Critique

see *Science* March 2014,
"The Parable of Google Flu: Traps in Big Data Analysis"

- queries used were not disclosed
- didn't include CDC data as lagged variables (standard time series approach)

Moreover

- "flu" and "flu-like" are quite distinct
- behaviour of these proxy variables, queries by the public, may be affected by publicity around flu pandemics (2009, 2013)
- general uncertain nature of proxy variables

Data Analysis Meta Case Studies
Scientific method

is Data Science writ large, so what can we learn

Scientific Method in Medicine

- *John Oliver on Scientific Studies* in *Fortune*, May 2016
- *"How science goes wrong"* on *The Economist*, 2013
- *"Battling Bad Science"* a TED talk by Ben Goldacre, 2011
- *"The Truth Wears Off"* by Jonah Lehrer in *The New Yorker, 2010*
- *"Richard Smith: Time for science to be about truth rather* than careers" blog on *BMJ*, 2013
- *"Offline: What is medicine's 5 sigma?"* by Richard Horton on *The Lancet*, 2015
- *"The 10 stuff ups we all make when interpreting research"* by Will J Grant and Rod Lamberts in *The Conversation*, 2015.

Broadly:

- ~~industry coercion~~, ~~academic games~~, ~~press spin~~
- errors in application of scientific method

Scientific Method in Medicine

Major applications errors are:

- misuse of significance testing
- correlation does not imply causation
- not checking/testing the true costs
- inadequate reproducibility *e.g.*, difficult to repeat
- selection bias

Significance Testing Primer

Hypothesis: “a daily aspirin reduces heart attacks in older men”

1. experimental design: get 100 subjects (men), give 50 daily aspirin for 5 years, give another 50 a placebo
   - (but assign men to groups randomly and blind)
2. results: **HA** = count of aspirin group having heart attacks, **HP** = count of placebo group having heart attacks
3. statistical computation: compute (statistical) *surprise* in getting counts (HA,HP) *assuming* that a daily aspirin had no effect on heart attacks
   - *surprise* is measured in units of chance; 1:1000 is quite surprising, 1:10 is mildly surprising, *etc.*
   - statistician’s call this a *P-value*
4. evaluation:
   - if *surprise* is high enough (“significant”) then get to publish paper in prestigous medical journal
   - if *surprise* is low/moderate, then quietly ignore results

Significance Testing Primer

HA = count of aspirin group having heart attacks,
HP = count of placebo group having heart attacks

- HP=4, HA=3, results similar, little surprise, no result!
- HP=10, HA=1, results very different, so significance needs to be tested properly
- more data, more likely to get significant result
- stronger the effect, more likely to get significant result

Significance Testing Errors

Significance chasing: repeat many experiments until you get significance

- *"I Fooled Millions Into Thinking Chocolate Helps Weight Loss."* by John Bohannon,

In parallel: multiple teams trying different experiments until one gets significance

Ignoring negative results: a variation on the above; similar to repeated testing until success

The decline effect: a variation on the above, as *some* negative results get recorded, eventually the original (flawed) positive result gets overturned

Inadequate repeatability: means subsequent teams cannot check your results, so you're inital inadequate significance testing doesn't get retested

Significance Testing

- be careful with P-values and significance levels: use strong significance levels and don't "repeat until success"
- record negative results
- ensure repeatability by properly recording experimental methodology and data processing

Error: Correlation versus Causation

- we considered this earlier
- happens when medical experts use observational data to draw conclusions, *e.g.*, epidemiological data
- methods for testing/estimating causation from data is currently a research agenda in discovery science
- "intervention" is a basic part of *double blind trials* (a major experimental standard)

Error: Not Checking True Costs

Disclaimer: we are not medical experts

- justification for statin drugs: lowering cholesterol, therefore improves cardiovascular mortality
- *Arch Intern Med.* 2010 Jun 28;170(12):1032, "Cholesterol lowering, cardiovascular diseases, and the rosuvastatin-JUPITER controversy: a critical reappraisal," de Lorgeril *et al.*
  - rosuvastatin lowered cardiovascular mortality yes
  - but no significant decrease in overall mortality
- "cardiovascular mortality" is used as *surrogate endpoint*
- in this case it is apparently not reliable for the real measure of quality, overall mortality

Get Rich Quick with Machine Learning!

Why not train a machine learning algorithm to do prediction on the stock market?

- very many have tried
- extremely few succeeded
  - *e.g.*, see *The Prediction Company*
  - see *Quant.Stackexchange.Com query*
- The *Efficient Market Hypothesis* coupled with "transaction costs" always beats them

Error: Not Checking True Costs

- “surrogate endpoints” now controversial in medicine
- be careful with your evaluation measure
- bring the domain experts in to help
- consider delivery/implementation costs as well
- need to consider “total cost of ownership”, not “sticker price”

Data Analysis Meta Case Studies
Drug Interactions

Some concepts are related to Google Flu Trend

Drug Interactions

Search log data contains lots of interesting patterns.

- Microsoft investigated whether it was possible to use search data to perform pharmacovigilance
- pharmacovigilance ::= monitoring the effects of medical drugs after they have been licensed for use

See Eric Horvitz's video (from 38:40 to 42:30):
"Data, Predictions, and Decisions in Support of People and Society"

- they modelled the problem as a **prediction task**:
- to determine which drug pairs interact to cause hyperglycemia

Drug Interactions, cont.

Using Web queries to determine which drugs cause interactions:

- Researchers looked at the frequency with which different drug names and terms indicating hyperglycemia (high blood sugar levels) occurred in the user search histories.
- The reporting ratio (RR) is the percentage of times the hyperglycemia terms occurred for a given pair of drugs
- Ground truth data (regarding which pairs of drugs do interact to cause side effect) was taken from the *FAERS* system (data from physicians about drug interactions observed in their patients)
- Fictional example data below:

drug 1	**drug 2**	**RR**	**truth**
dobutamine	hydrocortisone	12.6	causes
glipizide	phenotoin	9.4	causes
...	...	...	...
budesonide	formoterol	7.3	not
labetalol	sertraline	2.4	not

Drug Interactions, cont.

drug 1	drug 2	RR	truth
dobutamine	hydrocortisone	12.6	causes
glipizide	phenotoin	9.4	causes
...	...	...	...
budesonide	formoterol	7.3	not
labetalol	sertraline	2.4	not

Based on reporting ratio (RR) value a prediction could be made for each pair of drugs:

- when $(RR > \texttt{cut-off})$, predict "causes", otherwise predict "not"
- changing the $\texttt{cut-off}$ controls the accuracy of the predictions
  - higher $\texttt{cut-off}$ results in more accurate predictions but small coverage
  - lower $\texttt{cut-off}$ results in less accurate predictions but larger coverage
- this is an inevitable tradeoff when making predictions!

Prediction Outcomes

For a given `cut-off` value, we can investigate the quality of the predictions by filling in counts in the confusion matrix:

	truth=not	truth=causes
prediction=not	19	7
prediction=causes	12	24

Each pair of drugs will correspond to one of 4 situations:

- true positive ::= entry for “prediction=causes” and “truth=causes”
- true negative ::= entry for “prediction=not” and “truth=not”
- false positive ::= entry for “prediction=causes” and “truth=not”
- false negative ::= entry for “prediction=not” and “truth=causes”

Aside – True/False Positive Rates

Based on the counts of true/false positives/negatives, we can calculate the *true positive and false positive rates*:

true positive rate ::= proportion of "true"s you get right

$$::= \frac{\text{true positives}}{\text{"true"}}$$

$$::= \frac{\text{true positives}}{\text{true positives + false negatives}} = \frac{24}{24+7}$$

a.k.a. sensitivity or recall

false positive rate ::= proportion of "false"s you get wrong

$$::= \frac{\text{false positive}}{\text{"false"}}$$

$$::= \frac{\text{false positive}}{\text{false positive + true negative}} = \frac{12}{12+19}$$

a.k.a. fall-out or (1-specificity)

Ideally, we would want true positive rate high but false positive rate low.

Aside – True/False Positive Rates

We can investigate how the true & false positive rates vary as we change the `cut-off` value:

- a low cut-off means we predict "causes" more often, so true positives are larger and false positives are larger
- as we **increase** the `cut-off`:
  - the true positive rate **decrease** monotonically (in steps)
  - but the false positive rate also **decreases** monotonically
- note that for each cut-off we have a point in 2 dimensions **(true positive rate, false positive rate)**
  - at lowest possible cut-off we have (1,1)
  - at highest possible cut-off we have (0,0)

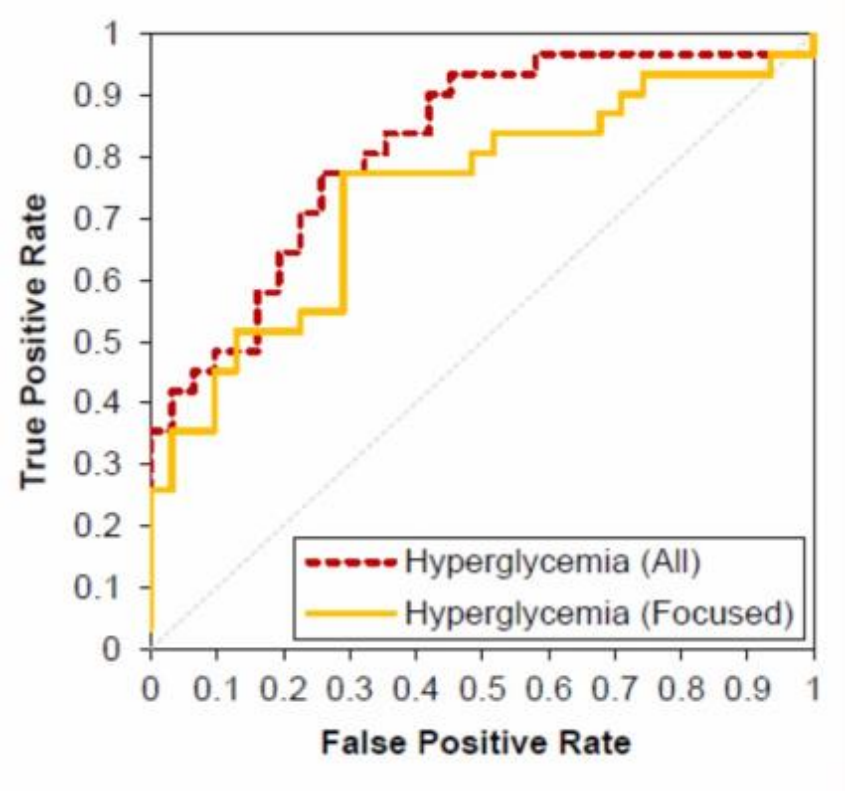
Aside – ROC Curves

Plotting the pairs of values yields the so-called ROC curve

- ROC ::= *receiver operating characteristic*

The ROC curve answers the question:

- "how do the prediction qualities change as I change my cut-off?"

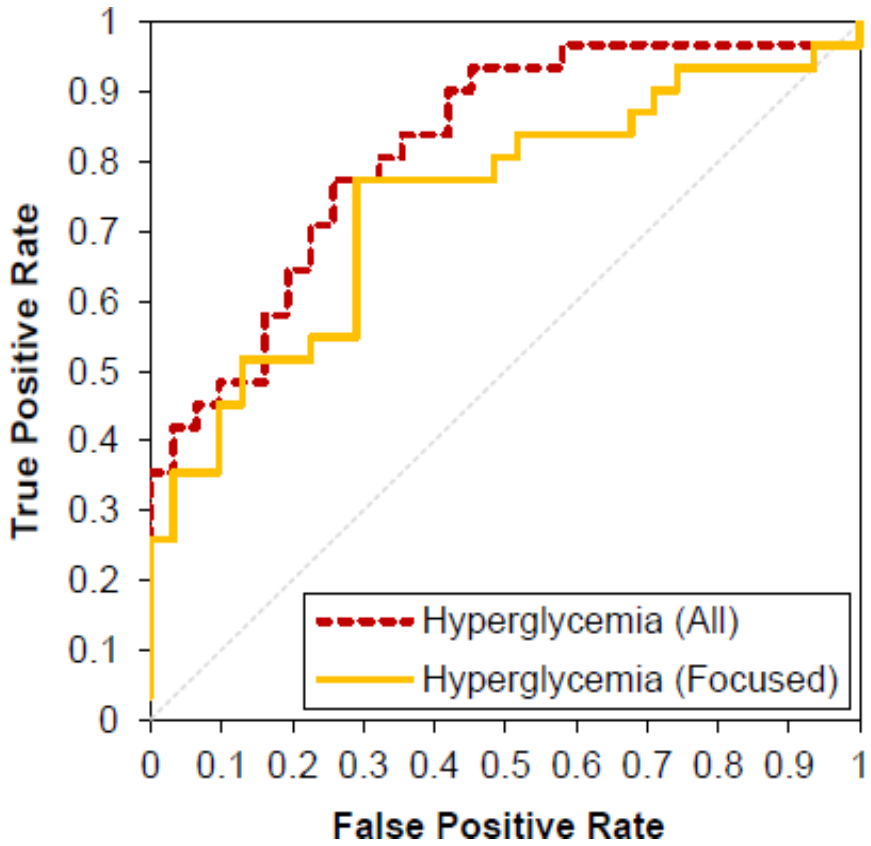


Note:
"All" is broader set of search terms for "hyperglycemia."
"Focused" is smaller but more reliable set.

Characterizing Sensor Error

Test on known interactions

- 31 true positives for hyperglycemia
- 31 true negatives for hyperglycemia



Label	*Drug 1*	*Drug 2*
TP	dobutamine	hydrocortisone
TP	dobutamine	triamcinolone
TP	dobutamine	prednisolone
TP	betamethasone	dobutamine
TP	glipizide	phenytoin
TP	dobutamine	methylprednisolone
TP	prednisolone	salmeterol
TP	salmeterol	triamcinolone
TP	betamethasone	terbutaline
TP	dexamethasone	dobutamine
TP	budesonide	salmeterol
TN	hydrochlorothiazide	tazobactam
TN	clindamycin	montelukast
TN	lamotrigine	nystatin
TN	methylprednisolone	rosuvastatin
TP	budesonide	formoterol
TN	loratadine	nystatin
TN	hydroxychloroquine	prochlorperazine
TN	labetalol	sertraline
TN	ciprofloxacin	vecuronium

Web Pharmacovigilance

Previous ROC curves show web query data can be used quite reliably to predict drug interactions causing hyperglycemia,

- *i.e.* without need for obtaining the physician data
- the web query data is being used as *proxy data* for the FAERS reports

Note that you could almost perform the RR computations yourself using Googles results estimates!

It's good to use **proxy data** when you can, but you have to understand its reliability.

Question: when would web search queries make unreliable proxy data?

e.g. see See Eric Horvitz's video (from 44:10 to 45:00 plus ...):
"Data, Predictions, and Decisions in Support of People and Society"

Data Analysis Meta Case Studies
What is Hard?

comments

The Hardest Parts

See blog *"The hardest parts of data science"* by Yanir Seroussi 23rd Nov. 2015.

Model fitting: core statistics/machine learning not usually hard (*e.g.,* many use R as a black box for this)

Data collection: can be critical sometimes, but often more routine

Data cleaning: can be a lot of work, but often more routine

Problem definition: getting into the application and understanding the real problem can be hard

Evaluation: what is measured? should multiple evaluations be done? can be hard

Ambiguity and uncertainty: invariably these occur and we need to live with them; can be hard

Unit Schedule: Next Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	9	data analysis theory
	10	**data analysis process**
6.	**11**	**issues in data management**
	12	GUEST SPEAKER & EXAM INFO