Data Management for Statistical Analysis/Machine Learning

Finlay Maguire

February 6, 2018

Faculty of Computer Science, Dalhousie

Table of contents

Dangers of Spreadsheets

JP Morgan ‘London Whale’

that decision, further errors were discovered in the Basel II.5 model, including, most significantly, an operational error in the calculation of the relative changes in hazard rates and correlation estimates. Specifically, after subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the modeler had intended. This error likely had the effect of muting volatility by a factor of two and of lowering the VaR,

addition, the price-testing process relied on the use of spreadsheets that were not vetted by CIO VCG (or Finance) management, and required time-consuming manual inputs to entries and formulas, which increased the potential for errors.

During the review process, additional operational issues became apparent. For example, the model operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another. In addition, many of

[JPMorgan and Chase, 2013]

JP Morgan ‘London Whale’

London Whale scandal to cost JP Morgan $920m in penalties

US's biggest bank to pay penalties to US and UK regulators for 'unsound practices' relating to $6.2bn losses last year

https://www.theguardian.com/business/2013/sep/19/jp-morgan-920m-fine-london-whale

Data Mangling in Bioinformatics

pressor DEC1 [Deleted in Esophageal Cancer 1] [3] was being converted to '1-DEC.' Figure 1 lists 30 gene names that suffer an analogous fate.

[Zeeberg et al., 2004]

> 60,770. For example, the RIKEN identifier "2310009E13" was converted irreversibly to the floating-point number "2.31E+13." A non-expert user might well fail to notice that approximately 3% of the identifiers on a microarray with tens of thousands of genes had been converted to an incorrect form, yet the potential for 2,000 identifiers to be transmogrified without notice is a considerable concern.

[Zeeberg et al., 2004]

Data Mangling in Bioinformatics

Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

[Ziemann et al., 2016]

Austerity as a macroeconomic policy

[Reinhart and Rogoff, 2010a, Reinhart and Rogoff, 2010b]

Further, RR (2010B) was the only evidence cited on the consequences of high public debt on economic growth in the 2013 US Federal Budget plan proposed by Republican Paul Ryan, which was passed in the House of Representatives. Congressman Ryan's 'Path to Prosperity' proposal reports that RR's research 'found conclusive empirical evidence that gross debt (meaning all debt that a government owes, including debt held in government trust funds) exceeding 90 percent of the economy has a significant negative effect on economic growth' (Ryan, 2013, p. 78). George Osborne, the UK Chancellor of the Exchequer, and Olli Rehn, the leading economic official of the European Commission, are other leading policy makers who have frequently cited the RR work as significantly influencing their thinking. Indeed, Paul Krugman observed in June 2013 that 'Reinhart–Rogoff may have had more immediate influence on public debate than any previous paper in the history of economics' (Krugman, 2013).

[Herndon et al., 2014]

Austerity as a macroeconomic policy

3.2 Spreadsheet coding error

In addition to these deliberate data exclusions by RR, a coding error in the RR working spreadsheet also unintentionally excludes five countries entirely (Australia, Austria, Belgium, Canada and Denmark) from all parts of the analysis.[9] The error appears in the calculations of both mean and median GDP growth with the 1946–2009 sample as well as with the mean and median GDP growth for the sample over the 220-year period 1790–2009. The omitted countries are selected alphabetically. It is clear from the spreadsheet itself that these are random exclusions. RR have since acknowledged this to be the case (RR, 2013A, 2013B, 2013C).

[Herndon et al., 2014]

Austerity as a macroeconomic policy

	Public debt/GDP category			
	≤30%	30–60%	60–90%	>90%
Recalculated results				
All data with country-year weighting	4.2	3.1	3.2	2.2
Replication elements				
Separate effects of RR calculations				
Spreadsheet error only	4.2	3.0	3.2	1.9
Selective years exclusion only	4.2	3.1	3.2	1.9
Country weights only	4.0	3.0	3.0	1.9
Interactive effects of RR calculations				
Spreadsheet error + selective years exclusion	4.2	3.0	3.2	1.7
Spreadsheet error + country weights	4.1	2.9	3.4	1.4
Selective years exclusion + country weights	4.0	3.0	3.0	0.3
Spreadsheet error + selective years exclusion + country weights	4.1	2.9	3.4	0.0
Spreadsheet error + selective years exclusion + country weights + transcription error	4.1	2.9	3.4	−0.1
RR published results				
RR (2010A, 2010B, Figure 2) (approximated)	3.8	2.9	3.4	−0.1
RR (2010B, Appendix Table 1)	4.1	2.8	2.8	−0.1

[Herndon et al., 2014]

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.
- Mouchel (Outsourcing specialist) lost £4.3M due to pension deficit error.

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.
- Mouchel (Outsourcing specialist) lost £4.3M due to pension deficit error.
- The UK Security Service (‘MI5’) bugged the wrong telephones 1,061 times.

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.
- Mouchel (Outsourcing specialist) lost £4.3M due to pension deficit error.
- The UK Security Service (‘MI5’) bugged the wrong telephones 1,061 times.
- Oxford University History Faculty mixed up test scores and applicants.

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.
- Mouchel (Outsourcing specialist) lost £4.3M due to pension deficit error.
- The UK Security Service ('MI5') bugged the wrong telephones 1,061 times.
- Oxford University History Faculty mixed up test scores and applicants.
- See `http://www.eusprig.org/horror-stories.htm` for a non-comprehensive list!

And many more.

- Omission of a minus sign which cost Fidelity Magellan Fund 2.45 billion US dollars (in 1995).
- The London 2012 oversold synchronised swimming by 10,000 tickets.
- Kern County, California lost records of taxable property worth 1.26 billion US dollars.
- Mouchel (Outsourcing specialist) lost £4.3M due to pension deficit error.
- The UK Security Service (‘MI5’) bugged the wrong telephones 1,061 times.
- Oxford University History Faculty mixed up test scores and applicants.
- See `http://www.eusprig.org/horror-stories.htm` for a non-comprehensive list!
- Meta-review suggests that **88%** of all spreadsheets have errors [Panko, 2008].

Why are Spreadsheets Dangerous?

- Difficult to version control

Why are Spreadsheets Dangerous?

- Difficult to version control
- Non-linear dependencies (formulas all over the place)

Why are Spreadsheets Dangerous?

- Difficult to version control
- Non-linear dependencies (formulas all over the place)
- Hidden formatting/data.

Why are Spreadsheets Dangerous?

- Difficult to version control
- Non-linear dependencies (formulas all over the place)
- Hidden formatting/data.
- Automated formatting of data

Why are Spreadsheets Dangerous?

- Difficult to version control
- Non-linear dependencies (formulas all over the place)
- Hidden formatting/data.
- Automated formatting of data
- **A billion users** (according to Microsoft).

Tidy Data

So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)

So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)
- No manual editing after initial entry

So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)
- No manual editing after initial entry
- Access via scripting (e.g. ‘dplyr’, ‘pandas’, etc.)

So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)
- No manual editing after initial entry
- Access via scripting (e.g. ‘dplyr’, ‘pandas’, etc.)
- Version Control (e.g. git, dat) of analysis scripts and dataset

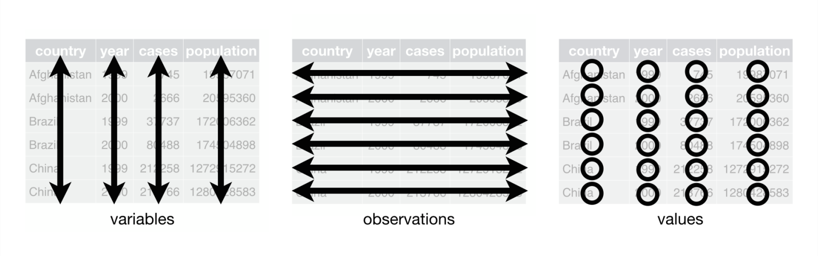
So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)
- No manual editing after initial entry
- Access via scripting (e.g. ‘dplyr’, ‘pandas’, etc.)
- Version Control (e.g. git, dat) of analysis scripts and dataset
- Tidy Data formatting

So what is the solution?

- Simple text based Machine-readable formats (e.g. CSV/TSV instead of XLSX)
- No manual editing after initial entry
- Access via scripting (e.g. ‘dplyr’, ‘pandas’, etc.)
- Version Control (e.g. git, dat) of analysis scripts and dataset
- Tidy Data formatting
- Consistent datatypes

Tidy Data



[Wickham, 2014]

‘dplyr’

- Seven ‘verbs’ that can be combined to do pretty much any operation

‘dplyr’

- Seven ‘verbs’ that can be combined to do pretty much any operation
- Intuitive: say you want to select rows ‘a’ and ‘b’ from a table then filter out values less than 10

‘dplyr’

- Seven ‘verbs’ that can be combined to do pretty much any operation
- Intuitive: say you want to select rows ‘a’ and ‘b’ from a table then filter out values less than 10
- `table %>% select('a', 'b') %>% filter(a >= 10)`

‘dplyr’ example

Source: local data frame [471,949 x 9]

	FL_DATE	CARRIER	ORIGIN	ORIGIN_CITY_NAME	ORIGIN_STATE_ABR	DEP_DELAY	DEP_TIME	ARR_DELAY	ARR_TIME
	(date)	(chr)	(chr)	(chr)	(chr)	(dbl)	(chr)	(dbl)	(chr)
1	2014-01-25	EV	LFT	Lafayette, LA	LA	NA		NA	
2	2014-01-30	EV	LFT	Lafayette, LA	LA	NA		NA	
3	2014-01-24	EV	LFT	Lafayette, LA	LA	NA		NA	
4	2014-01-01	EV	LFT	Lafayette, LA	LA	-12	0636	-23	0733
5	2014-01-03	EV	LFT	Lafayette, LA	LA	191	0959	181	1057
6	2014-01-04	EV	LFT	Lafayette, LA	LA	12	0700	15	0811
7	2014-01-05	EV	LFT	Lafayette, LA	LA	6	0654	1	0757
8	2014-01-09	EV	LFT	Lafayette, LA	LA	1	0656	2	0804
9	2014-01-13	EV	LFT	Lafayette, LA	LA	-9	0651	-18	0749
10	2014-01-12	EV	LFT	Lafayette, LA	LA	1	0656	-2	0800
..	...	...	...	...	...	...	...	...	...

https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e

- Say we want to know how many United Airline (UA) or American Airline (AA) flights leave from NYC

‘dplyr’ example

```
Source: local data frame [471,949 x 9]

      FL_DATE CARRIER ORIGIN ORIGIN_CITY_NAME ORIGIN_STATE_ABR DEP_DELAY DEP_TIME ARR_DELAY ARR_TIME
       (date)   (chr)  (chr)            (chr)            (chr)     (dbl)    (chr)     (dbl)    (chr)
1  2014-01-25      EV    LFT    Lafayette, LA               LA        NA                 NA
2  2014-01-30      EV    LFT    Lafayette, LA               LA        NA                 NA
3  2014-01-24      EV    LFT    Lafayette, LA               LA        NA                 NA
4  2014-01-01      EV    LFT    Lafayette, LA               LA       -12     0636       -23     0733
5  2014-01-03      EV    LFT    Lafayette, LA               LA       191     0959       181     1057
6  2014-01-04      EV    LFT    Lafayette, LA               LA        12     0700        15     0811
7  2014-01-05      EV    LFT    Lafayette, LA               LA         6     0654         1     0757
8  2014-01-09      EV    LFT    Lafayette, LA               LA         1     0656         2     0804
9  2014-01-13      EV    LFT    Lafayette, LA               LA        -9     0651       -18     0749
10 2014-01-12      EV    LFT    Lafayette, LA               LA         1     0656        -2     0800
..        ...     ...    ...              ...              ...       ...      ...       ...      ...
```

https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e

- Say we want to know how many United Airline (UA) or American Airline (AA) flights leave from NYC
- `flights %>% filter(CARRIER %in% c(“UA”, “AA”)) %>% count(CARRIER)`

'dplyr' example

```
Source: local data frame [2 x 2]

  CARRIER     n
    (chr) (int)
1      AA 45401
2      UA 39225
```

https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e

Machine Learning

Overview

- What is Machine Learning?

Overview

- What is Machine Learning?
- Unsupervised Learning

Overview

- What is Machine Learning?
- Unsupervised Learning
    - Clustering

Overview

- What is Machine Learning?
- Unsupervised Learning
    - Clustering
    - Dimensionality Reduction

Overview

- What is Machine Learning?
- Unsupervised Learning
    - Clustering
    - Dimensionality Reduction
- Supervised Learning

Overview

- What is Machine Learning?
- Unsupervised Learning
    - Clustering
    - Dimensionality Reduction
- Supervised Learning
    - Regression

Overview

- What is Machine Learning?
- Unsupervised Learning
    - Clustering
    - Dimensionality Reduction
- Supervised Learning
    - Regression
    - Classification

What is Machine Learning?

https://xkcd.com/1838/

What is Machine Learning?

- Algorithms which can learn from data

What is Machine Learning?

- Algorithms which can learn from data
- Artificial Intelligence?

What is Machine Learning?

- Algorithms which can learn from data
- Artificial Intelligence?
- Automatic pattern recognition

What is Machine Learning?

- Algorithms which can learn from data
- Artificial Intelligence?
- Automatic pattern recognition
- ‘Rebrand’ of statistics

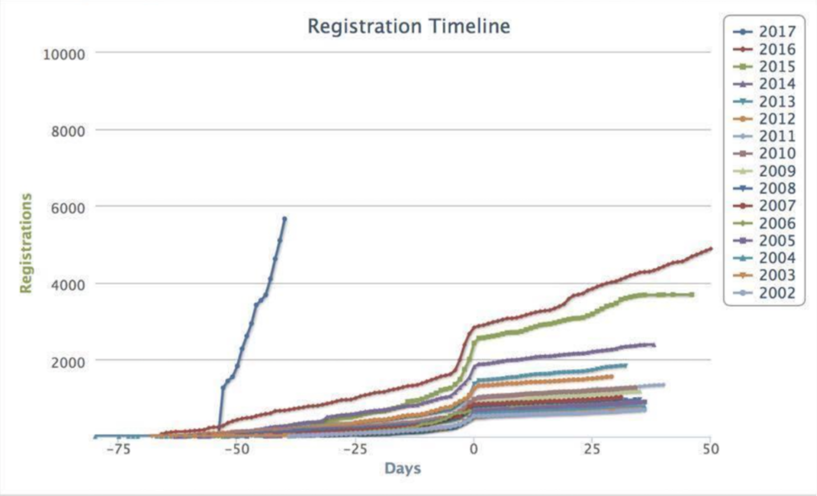
What is Machine Learning?

- Algorithms which can learn from data
- Artificial Intelligence?
- Automatic pattern recognition
- ‘Rebrand’ of statistics
- Applied laziness!

What is Machine Learning?

- Algorithms which can learn from data
- Artificial Intelligence?
- Automatic pattern recognition
- ‘Rebrand’ of statistics
- Applied laziness!
- Important tool to deal with large amounts of data (e.g. ‘big data’)

Massive Explosion of Popularity



Lots of Economic Activity

Economist via `https://greydanus.github.io/2017/12/23/nips/`

Big Data is Big Money

- Google DeepMind’s “staff costs” were $138 million for 400 employees. $345,000 per employee [Metz, 2017].

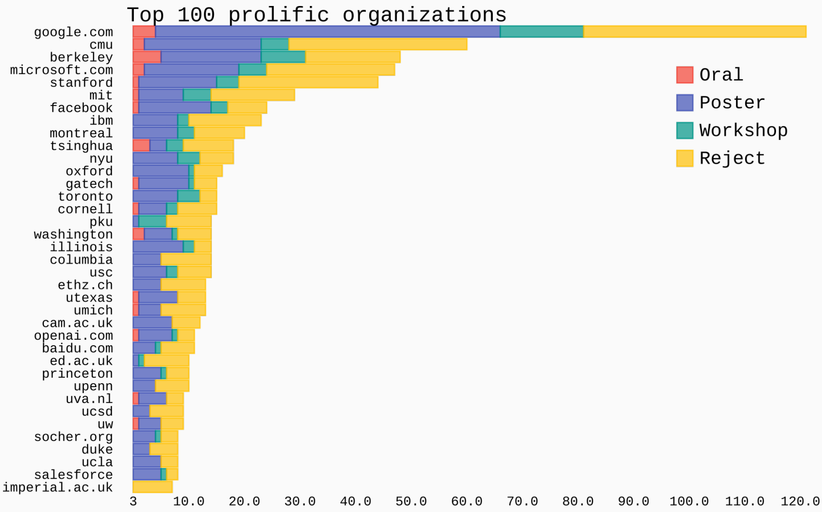
Big Data is Big Money

- Google DeepMind's "staff costs" were $138 million for 400 employees. $345,000 per employee [Metz, 2017].
- Ph.D. candidate job offer over $1 million a year [Markoff and Lohr, 2016].

Big Data is Big Money

- Google DeepMind's "staff costs" were $138 million for 400 employees. $345,000 per employee [Metz, 2017].
- Ph.D. candidate job offer over $1 million a year [Markoff and Lohr, 2016].
- Apple, Google, Microsoft, Intel, Uber, Facebook, Amazon all running their own labs [Metz, 2017].

Tech Industry Heavy



modified from `https://prlz77.github.io/iclr2018-stats-3/`

Possibly Over-hyped

Let the Gradient Flo

Celebrate NIPS 2017 with Intel AI

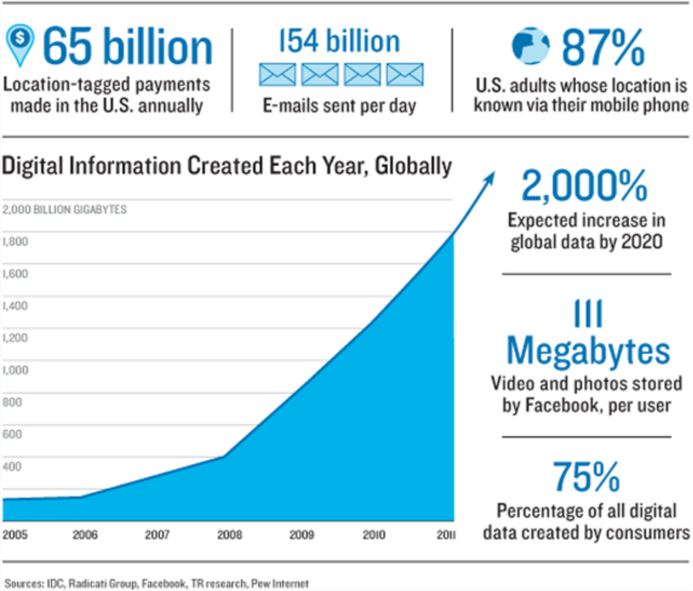
Join us for an exclusive party – and a surprise reveal.

Giveaways, buskers, acrobats, DJ Nostalgia B and a special performance by Flo Rida!

When	Where
Tuesday, December 5th	The Loft on Pine
9:00 PM – 12:00 AM	230 Pine Avenue
Door open at 9:00 PM	Long Beach, CA 90802
Show up early - space is limited	Near the Long Beach Convention Center

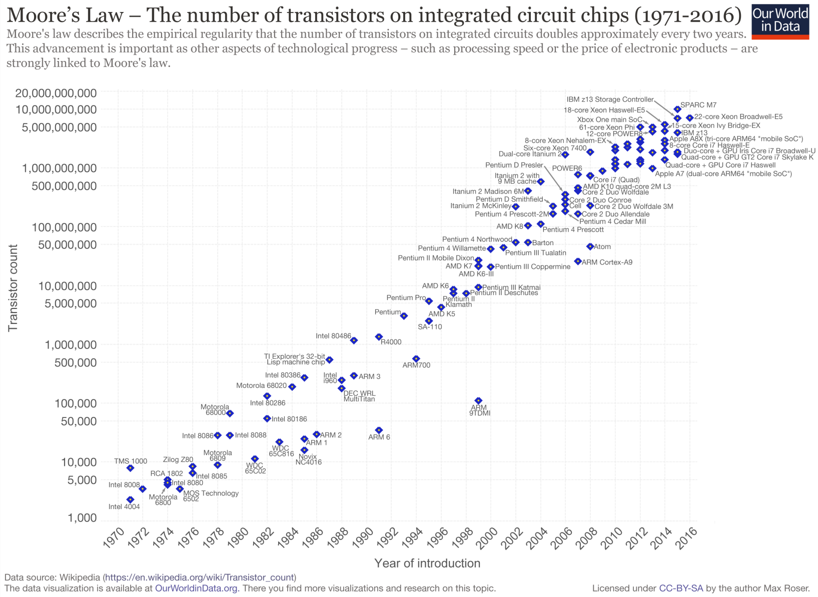
https://greydanus.github.io/2017/12/23/nips/

Why this explosion? Data



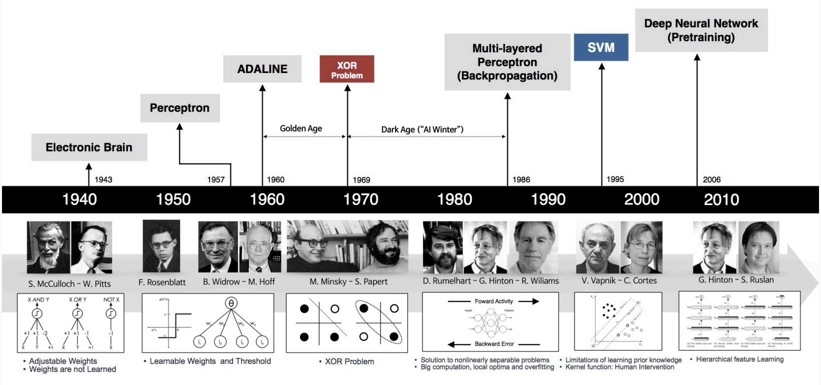
http://allennance.com/2013/07/dataexplosion

Why this explosion? Computing Power



https://ourworldindata.org/wp-content/uploads/2013/05/Transistor-Count-over-time.png

Why this explosion? Algorithms



https://www.slideshare.net/deview/251-implementing-deep-learning-using-cu-dnn/4

Unsupervised Learning

What is Unsupervised Learning?

- You have a pile of data and you want to find patterns in it.

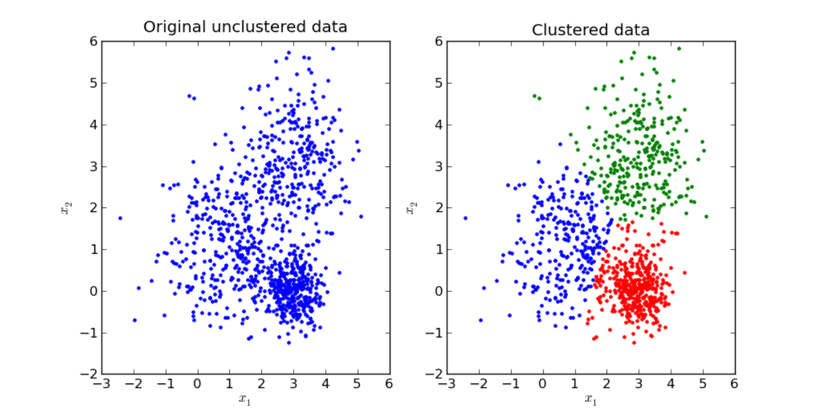
What is Unsupervised Learning?

- You have a pile of data and you want to find patterns in it.
- These patterns can be used to find groupings within the data.

What is Unsupervised Learning?

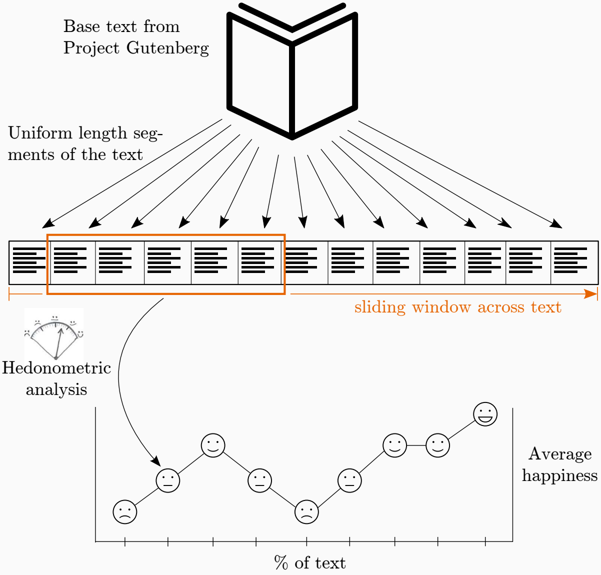
- You have a pile of data and you want to find patterns in it.
- These patterns can be used to find groupings within the data.
- Find a simpler/smaller version of the same data.

Clustering



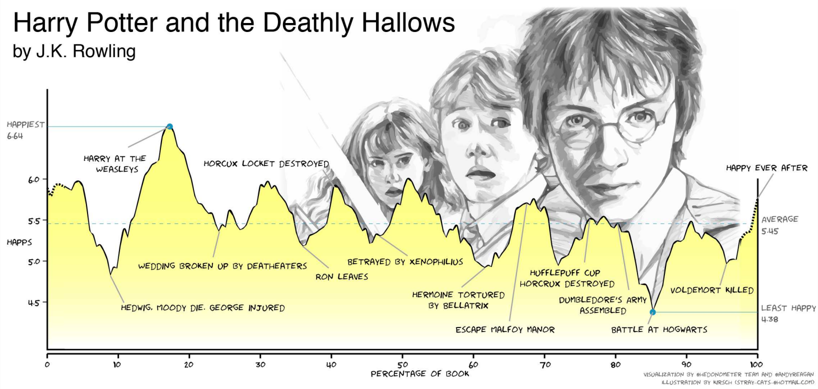
https://mubaris.com/2017/10/01/kmeans-clustering-in-python/

Story Emotional Arc Analysis



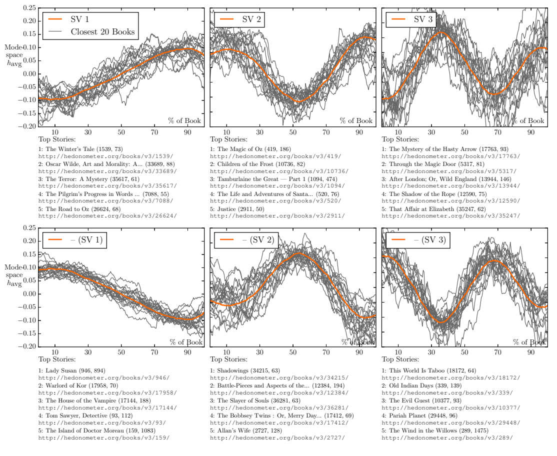
[Reagan et al., 2016]

Harry Potter Example



[Reagan et al., 2016]

Emotional Arc Clusters

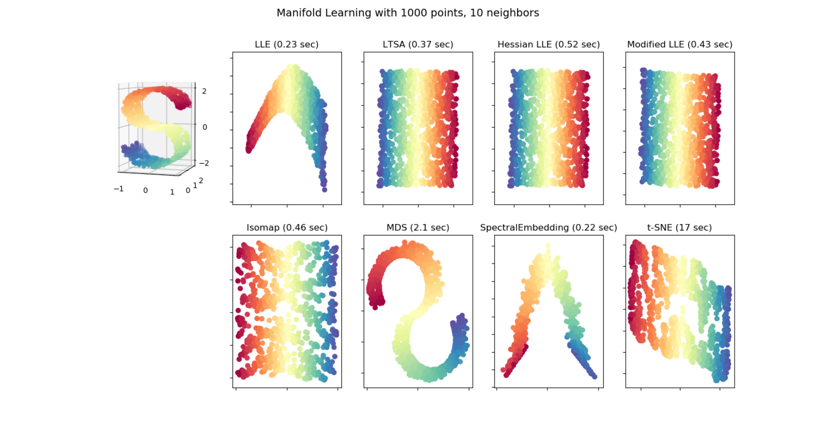


[Reagan et al., 2016]

"Rags to riches" (rise), "Man in a hole" (fall-rise), "Cinderella" (rise-fall-rise)

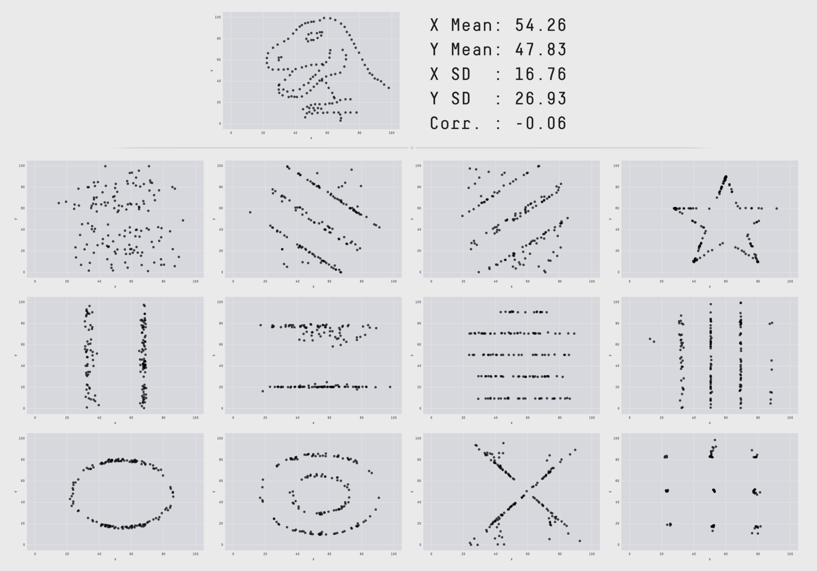
"Tragedy" (fall), "Icarus" (rise-fall), "Oedipus" (fall-rise-fall)

Dimensionality Reduction



http://scikit-learn.org/stable/auto_examples/manifold/plot_compare_methods.html

Aside: Visualisation is important



http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

Supervised Learning

What is Supervised Learning?

- You have labelled data.

What is Supervised Learning?

- You have labelled data.
- You want to predict the label for new data that isn't labelled.

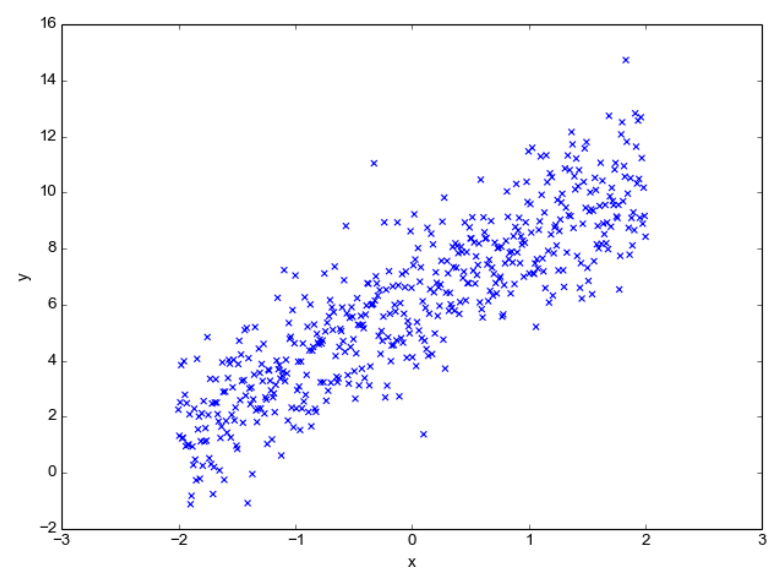
What is Supervised Learning?

- You have labelled data.
- You want to predict the label for new data that isn't labelled.
- Those labels are another number: regression.

What is Supervised Learning?

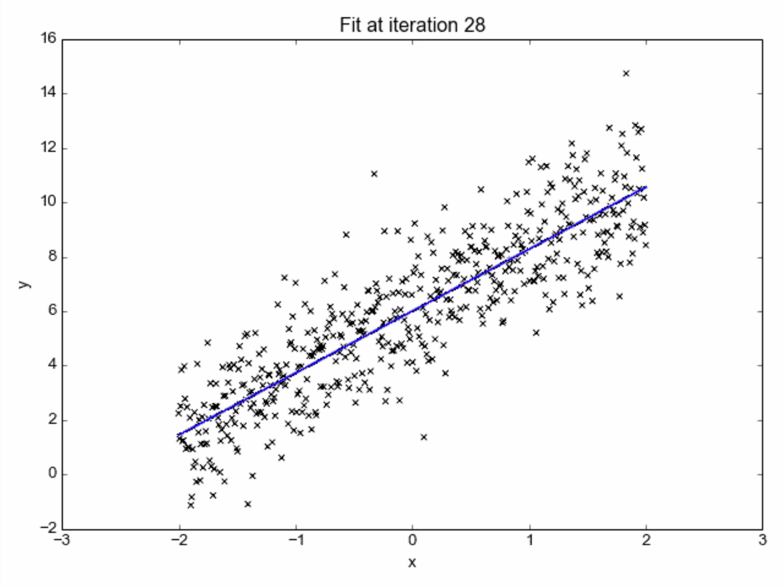
- You have labelled data.
- You want to predict the label for new data that isn't labelled.
- Those labels are another number: regression.
- Those labels are classes: classification.

Regression



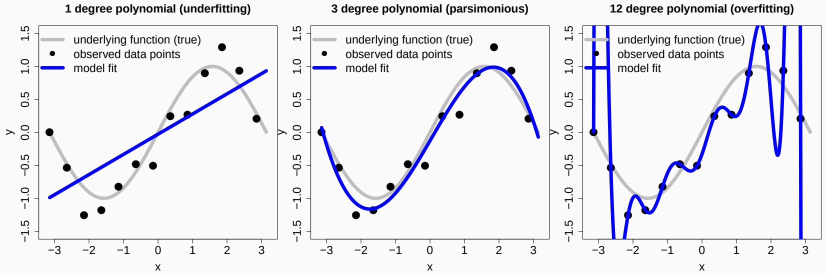
https://eli.thegreenplace.net/2016/linear-regression/

Fitting a simple line

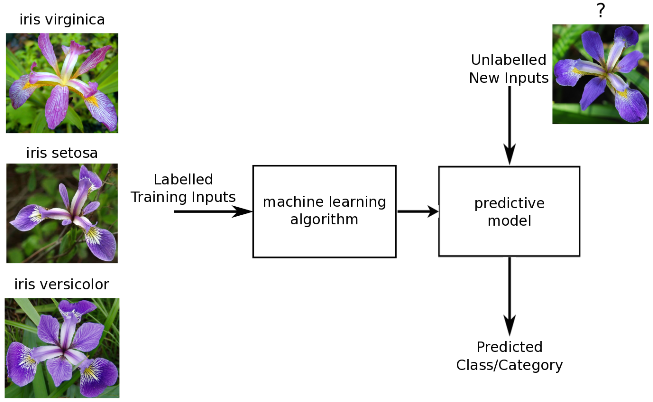


https://eli.thegreenplace.net/2016/linear-regression/

Which is the best model?

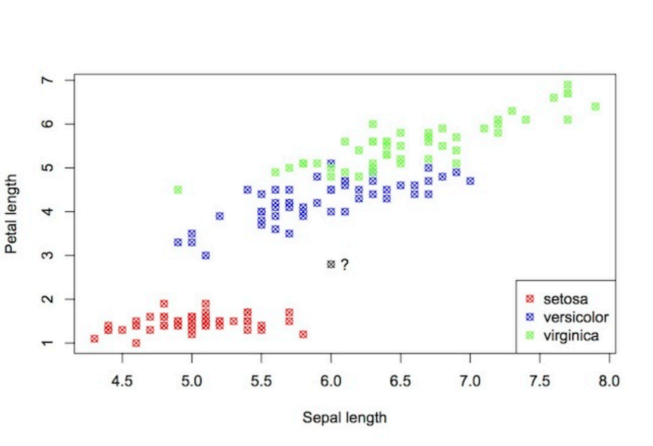


Classification

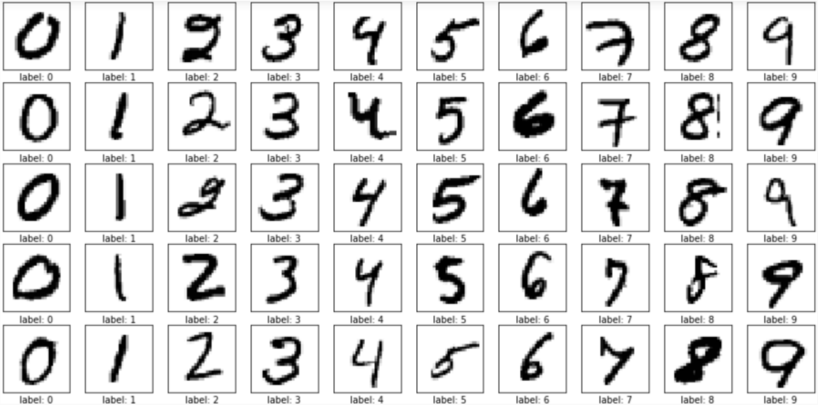


Assigning a set of new observations to a predefined category/class, using a predictive model trained on observations whose category/class is known

Iris Classification Features



http://www.ashbooth.com/wp-content/uploads/2014/07/class2.jpg



MNIST dataset

Dataset selection for ML

So you want to use ML?

- Define your problem.

So you want to use ML?

- Define your problem.
- Identify the type of data that might help solve this.

So you want to use ML?

- Define your problem.
- Identify the type of data that might help solve this.
- Work out what format you are collecting.

So you want to use ML?

- Define your problem.
- Identify the type of data that might help solve this.
- Work out what format you are collecting.
- Balanced data collection.

So you want to use ML?

- Define your problem.
- Identify the type of data that might help solve this.
- Work out what format you are collecting.
- Balanced data collection.
- Data leakage.

Balanced data collection

- Say you have 95 examples of class A and 5 example of class B.

Balanced data collection

- Say you have 95 examples of class A and 5 example of class B.
- How accurate is a classifier that just says all are class A?

Balanced data collection

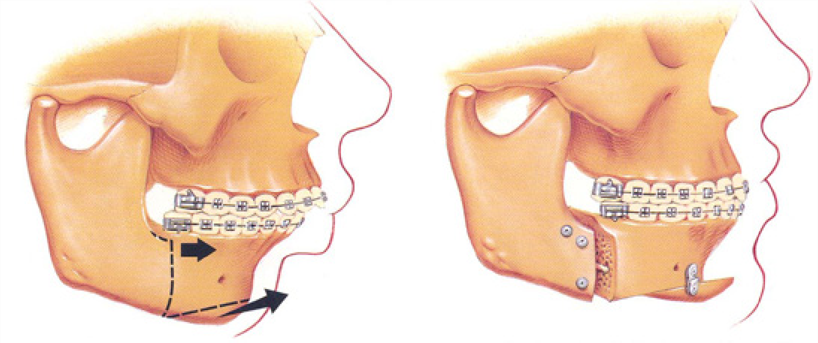
- Say you have 95 examples of class A and 5 example of class B.
- How accurate is a classifier that just says all are class A?
- 95% (not bad in a lot of cases).

Balanced data collection

- Say you have 95 examples of class A and 5 example of class B.
- How accurate is a classifier that just says all are class A?
- 95% (not bad in a lot of cases).
- Obviously an extreme example but a common problem.

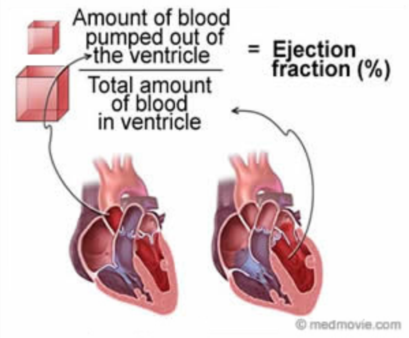
ML is lazy: Apocryphal Tank Example

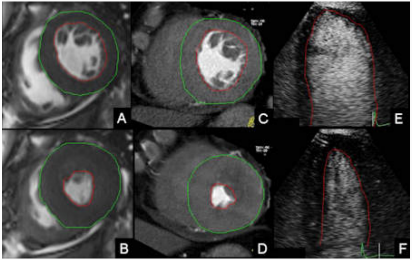
ML is lazy: Maxillofacial Surgery Success Rate



http://thazhathdentalclinic.com/oral-and-maxillofacial-surgery.html

ML is lazy: Ejected Fraction Estimation





NDSBII Dataset

Axioms of Machine Learning

- Crap in, crap out.

Axioms of Machine Learning

- Crap in, crap out.
- ‘If you torture data long enough it will confess to anything’: Ronald Coase

Axioms of Machine Learning

- Crap in, crap out.
- ‘If you torture data long enough it will confess to anything’: Ronald Coase
- ‘A sufficiently elaborate analysis process can always lend an air of legitimacy’: Chris Laws

Conclusion

Take-homes

- Extreme care must be taken when using excel.

Take-homes

- Extreme care must be taken when using excel.
- Use tidy data practices and always use version control.

Take-homes

- Extreme care must be taken when using excel.
- Use tidy data practices and always use version control.
- ML is a big topic and very powerful.

Take-homes

- Extreme care must be taken when using excel.
- Use tidy data practices and always use version control.
- ML is a big topic and very powerful.
- Effective ML requires careful data management.

Take-homes

- Extreme care must be taken when using excel.
- Use tidy data practices and always use version control.
- ML is a big topic and very powerful.
- Effective ML requires careful data management.
- ML is inherently lazy so take care with the input.

Questions?

References i

Herndon, T., Ash, M., and Pollin, R. (2014).
Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff.
Cambridge Journal of Economics, 38(2):257–279.

JPMorgan and Chase (2013).
Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses.
page 129.

Markoff, J. and Lohr, S. (2016).
The Race Is On to Control Artificial Intelligence, and Tech's Future.

Metz, C. (2017).
Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent.

Panko, R. R. (2008).
What we don't know about spreadsheet errors.
Journal of End User Computing, 10(2):15–21.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016).
The emotional arcs of stories are dominated by six basic shapes.
CoRR, abs/1606.07772.

Reinhart, C. M. and Rogoff, K. S. (2010a).
Growth in a Time of Debt.
American Economic Review: Papers & Proceedings, 100(322):1–10.

Reinhart, C. M. and Rogoff, K. S. (2010b).
Growth in a Time of Debt.
NBER Working Paper Series, (15639).

- Wickham, H. (2014).
  Tidy Data.
  Journal of Statistical Software, 59(10).

- Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C., and Weinstein, J. N. (2004).
  Mistaken identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics [1].
  BMC Bioinformatics, 5:1–6.

- Ziemann, M., Eren, Y., and El-Osta, A. (2016).
  Gene name errors are widespread in the scientific literature.
  Genome Biology, 17(1):17–19.