Slides: kndtran.com

Data Science and Machine Learning

KHOI-NGUYEN TRAN

POST DOCTORAL RESEARCHER, IBM RESEARCH AUSTRALIA (2016-NOW)

PHD IN COMPUTER SCIENCE, THE AUSTRALIAN NATIONAL UNIVERSITY (2015)

RMIT VIETNAM – 17 JULY 2017

Slides: kndtran.com

About me

Ho Chi Minh City

- Born here

Canberra, Capital of Australia

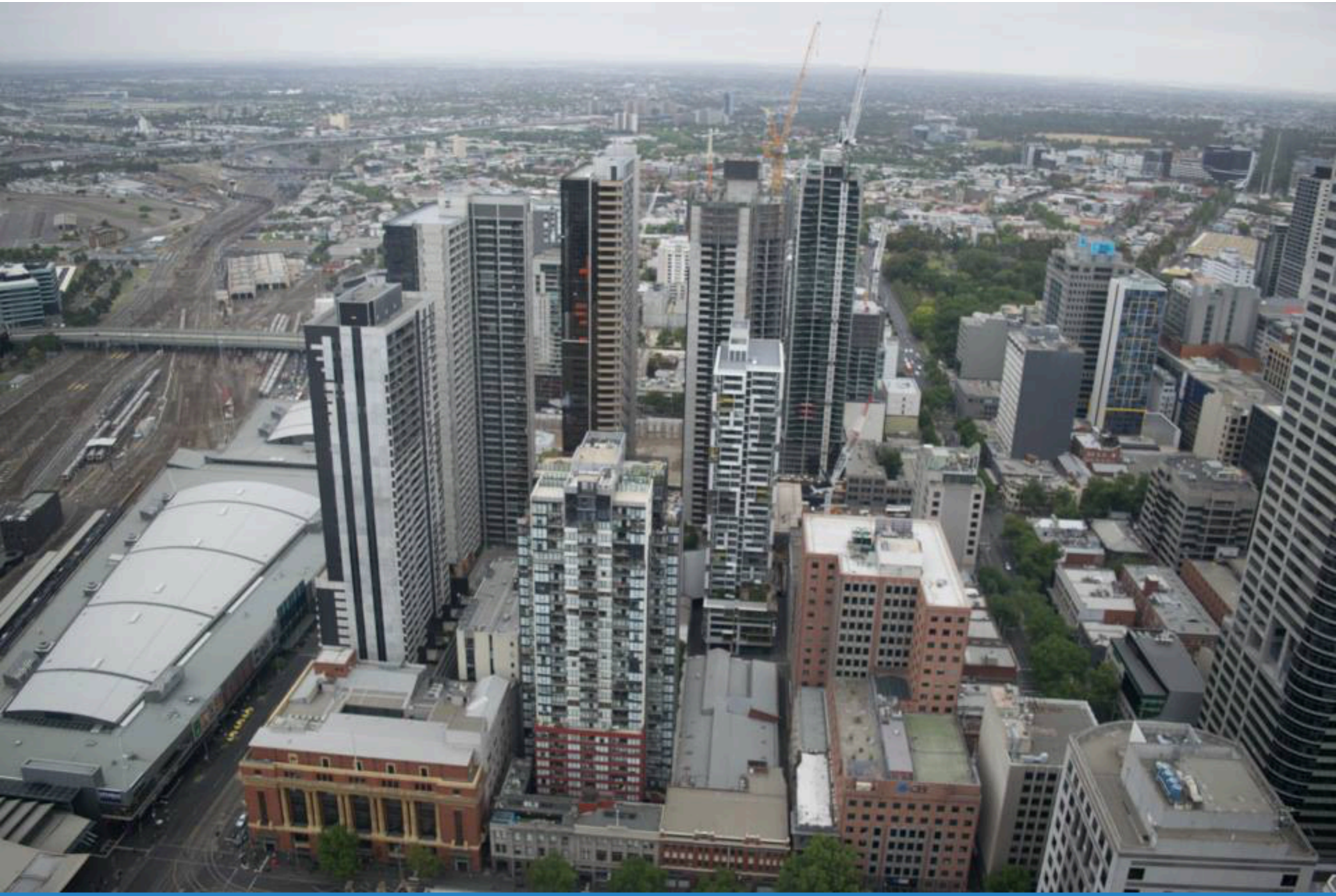
- Migrated here when I was very young
- Studied primary school and high school
- ANU – Bachelor of Computer Science with First Class Honours
- ANU – PhD in Computer Science
  - Thesis topic: Detecting Vandalism on Wikipedia across Multiple Languages

Melbourne

- Data Scientist role with the Australian Public Service
- Post Doctoral Researcher at IBM Research Australia

Image from https://move.ru/images/admin_uploads/14640769491.jpg

Image from http://d1w99recw67lvf.cloudfront.net/photos/large_Canberra_hero.jpg



IBM

Marketplace

IBM Research

Research areas | Work with us | About us | Blog

Featured research areas

Blockchain

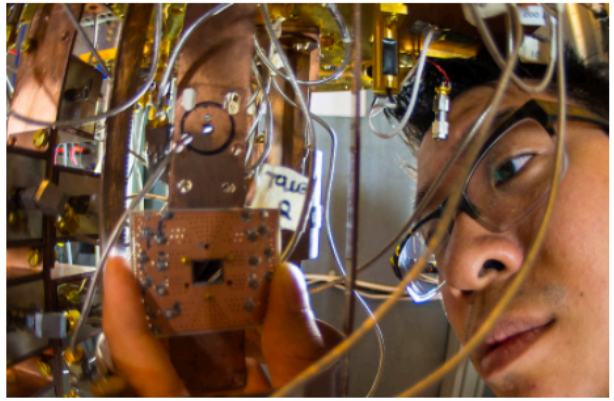
Blockchain is poised to do for transactions what the Internet did for information. Distributed ledger technology based on the IBM Research-backed Hyperledger Project has the potential to build trust into every transaction and remove barriers to doing business globally.

Learn more

AI and Cognitive Computing

Today we are at the start of a new technological era fueled by artificial intelligence. At IBM Research, we're relentlessly focused on using AI to augment human intelligence and decision-making, building cognitive systems that reason, draw insights and learn from data in ways no other organization can match.

Learn more



Quantum Computing

Quantum computing is a radical new computing model that harnesses the power of nature to address problems unsolvable with today's systems. To allow the scientific community to explore the possibilities, we launched in 2016 the first quantum computing platform on the cloud, the IBM Quantum Experience.

Learn more

Screenshot from http://www.research.ibm.com/

Overview

Data Science

- Importance, illustrative example, and the frontiers
- IBM, Australia, Vietnam
- Where do you start?
- Questions

Machine Learning

- (Same topics as above)

The Future

- Cognitive Computing
- Final Questions

The Big Picture

Extended slides

- kndtran.com

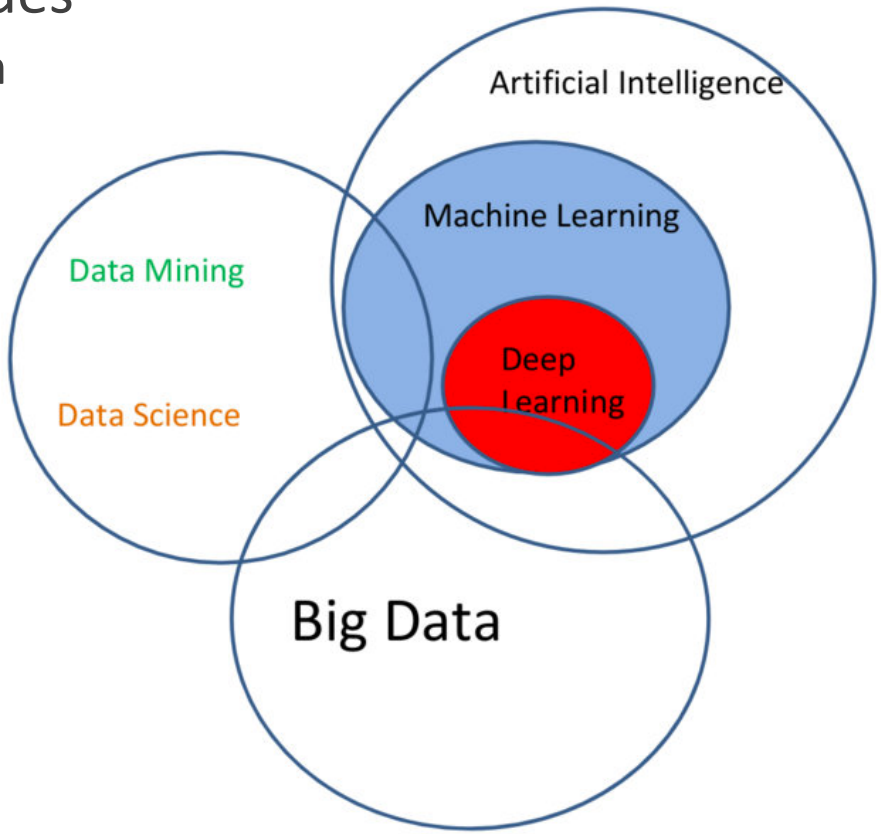


Image from http://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html

Definitions

Big Data

- Primarily focuses on the management of extremely large amounts of data

Data Science

- Primarily focuses on manipulation of data to extract knowledge

Machine Learning

- Primarily focuses on developing algorithms to find patterns in data with the aim of grouping or classification

Deep Learning

- New advances in hardware has allow artificial neural networks to drastically increase in complexity, advancing the problems machines can solve

Artificial Intelligence

- Primarily focuses on creating autonomous learning algorithms or agents that can adapt to a variety of environments

Data Science

Importance of Data

Raw material of the information age

90% of all data produced by humanity was in the last 2 years

Infinite and immediately accessible resource

Power comes from controlling, owning, or extracting meaning from data

Basis for industries and businesses of the future

Changing thinking of Data

Historically, data was made for humans

- Recorded in books, paper, spreadsheets
- Each person creates their own information

Nowadays, data is created for computers

- Generated/query from APIs
- Data exchange formats: CSV, JSON, etc

Data Science: the art of turning data created for computers into information for humans

Data on the Internet

Internet Live Stats – one second on the internet

- http://www.internetlivestats.com/
- Tweets: ~7.6K
- Internet traffic: ~46K GB
- Google searches: ~60K
- Youtube video views: ~70K
- Emails sent: ~2.6 million

Live changes on MediaWiki (owner of Wikipedia)

- https://codepen.io/ottomata/pen/VKNyEw/

Need for Data Scientists

Data is meaningless on its own.

Transform data into information, especially actionable information and information for making decisions

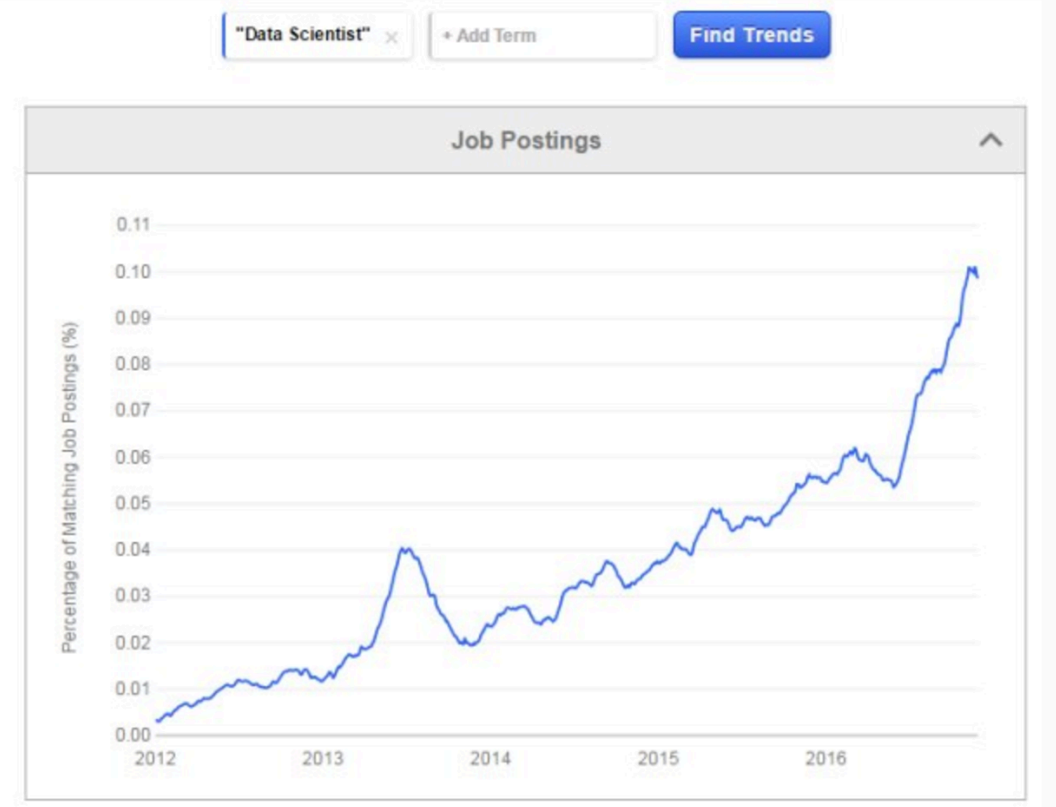
Internet of Things

- Trillions of devices will be connected to the internet
- Each device will generate their own data

We need people who can work with all this data, and there is demand

Need for Data Scientists (USA)

Indeed job trends also show continuing growth in the demand for Data Scientists:



Screenshot from http://www.kdnuggets.com/2017/01/glassdoor-data-scientist-best-job-america.html

Need for Data Scientists (USA)

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

View Jobs

Half of the top 10 jobs are related to Analytics, Big Data, and Data Science !

Rank	Title	Job Score	Job Satisfaction	Median Base Salary
1	Data Scientist	4.8	4.4	$110,000
2	DevOps Engineer	4.7	4.2	$110,000
3	Data Engineer	4.7	4.3	$106,000
5	Analytics Manager	4.6	4.1	$112,000
7	Database Administrator	4.5	3.8	$93,000

Compared to Glassdoor 2016 post where Data Scientist was also no. 1 job in USA, we note that the median Salary has declined from $117 to $110, but the number of listed job openings has increased from 1,700 to 4,200.

Screenshot from http://www.kdnuggets.com/2017/01/glassdoor-data-scientist-best-job-america.html

Data Science Lifecycle

Common workflow of a data scientist

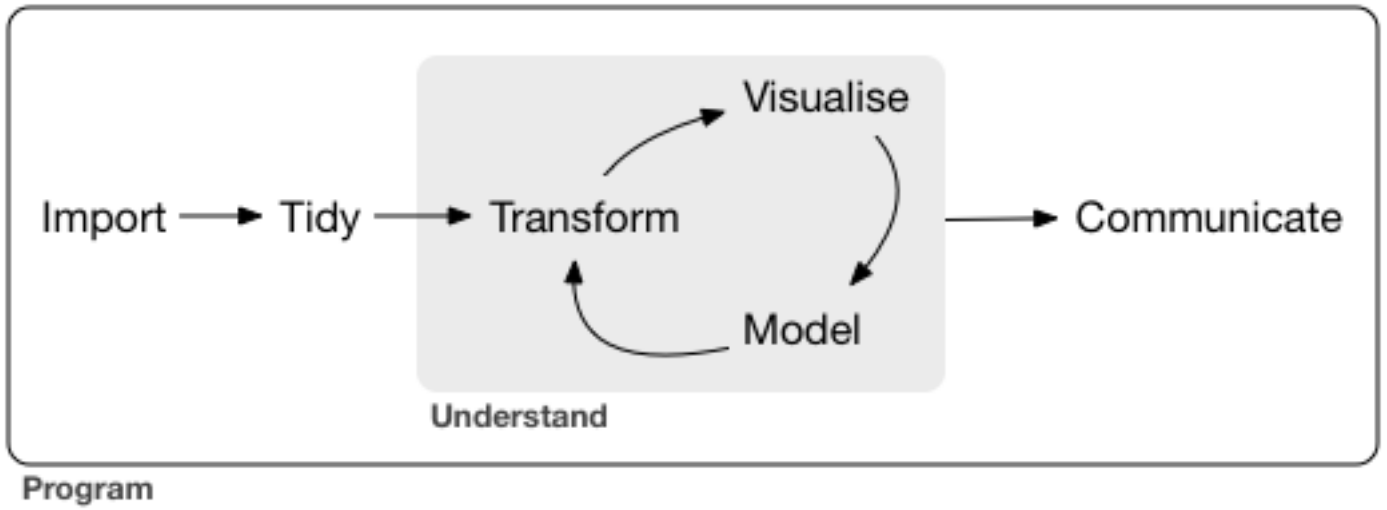


Image from http://r4ds.had.co.nz/introduction.html

Data Wrangling

Transforming data for humans into data for computers

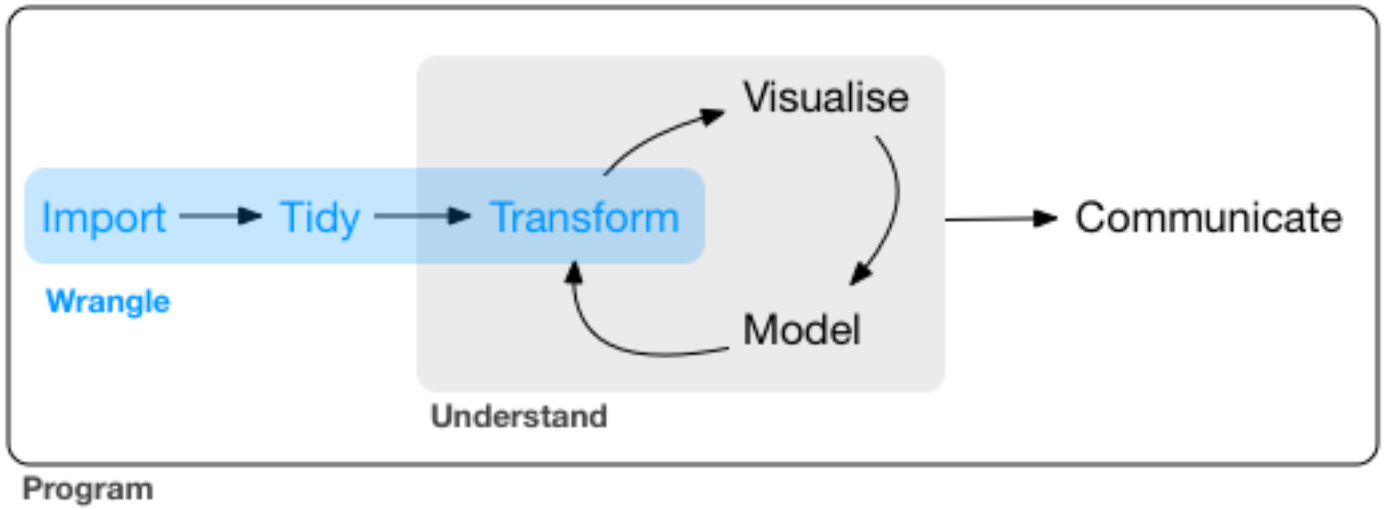
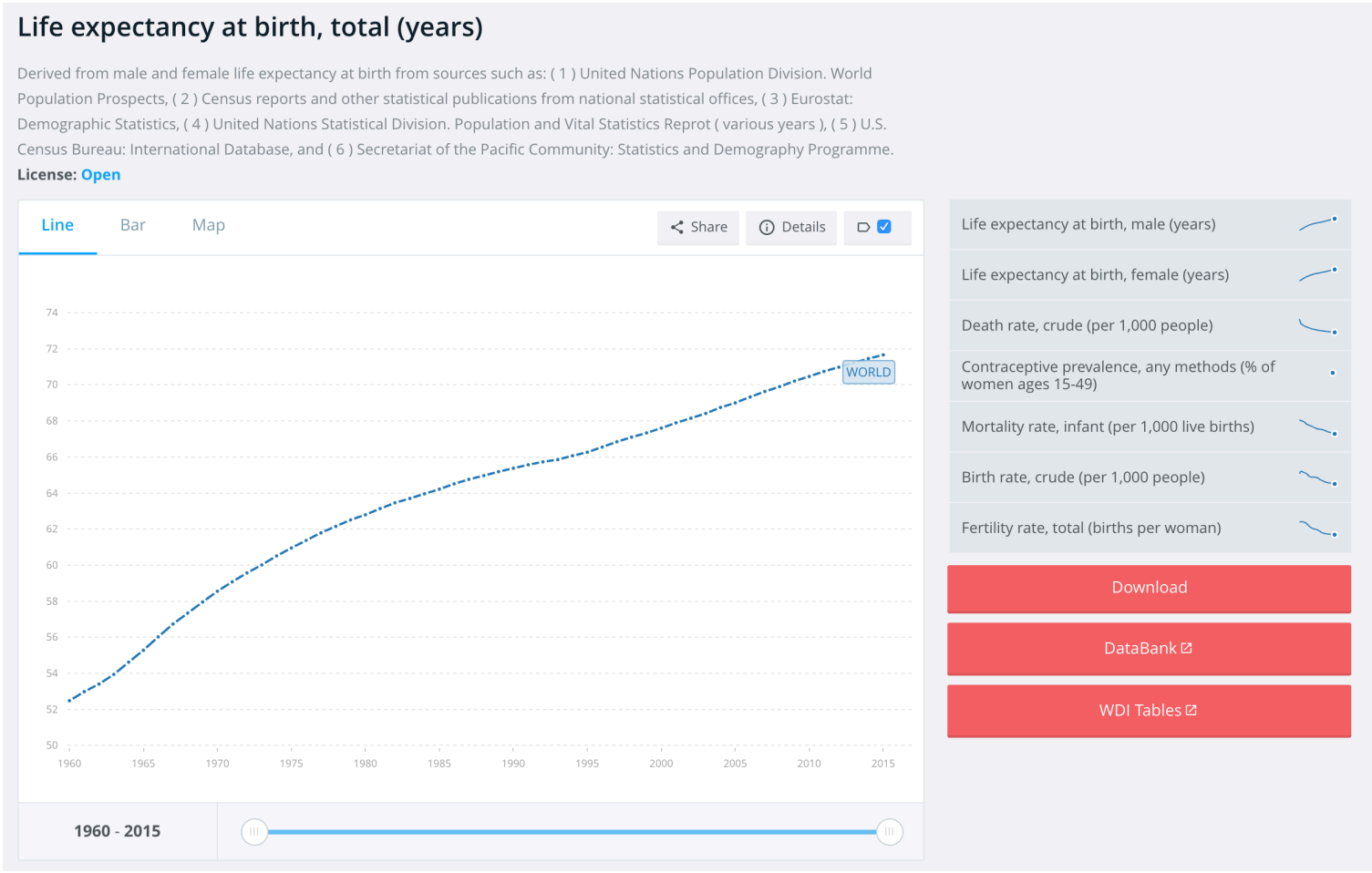


Image from http://r4ds.had.co.nz/wrangle-intro.html

Example – World Bank Data



Data from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

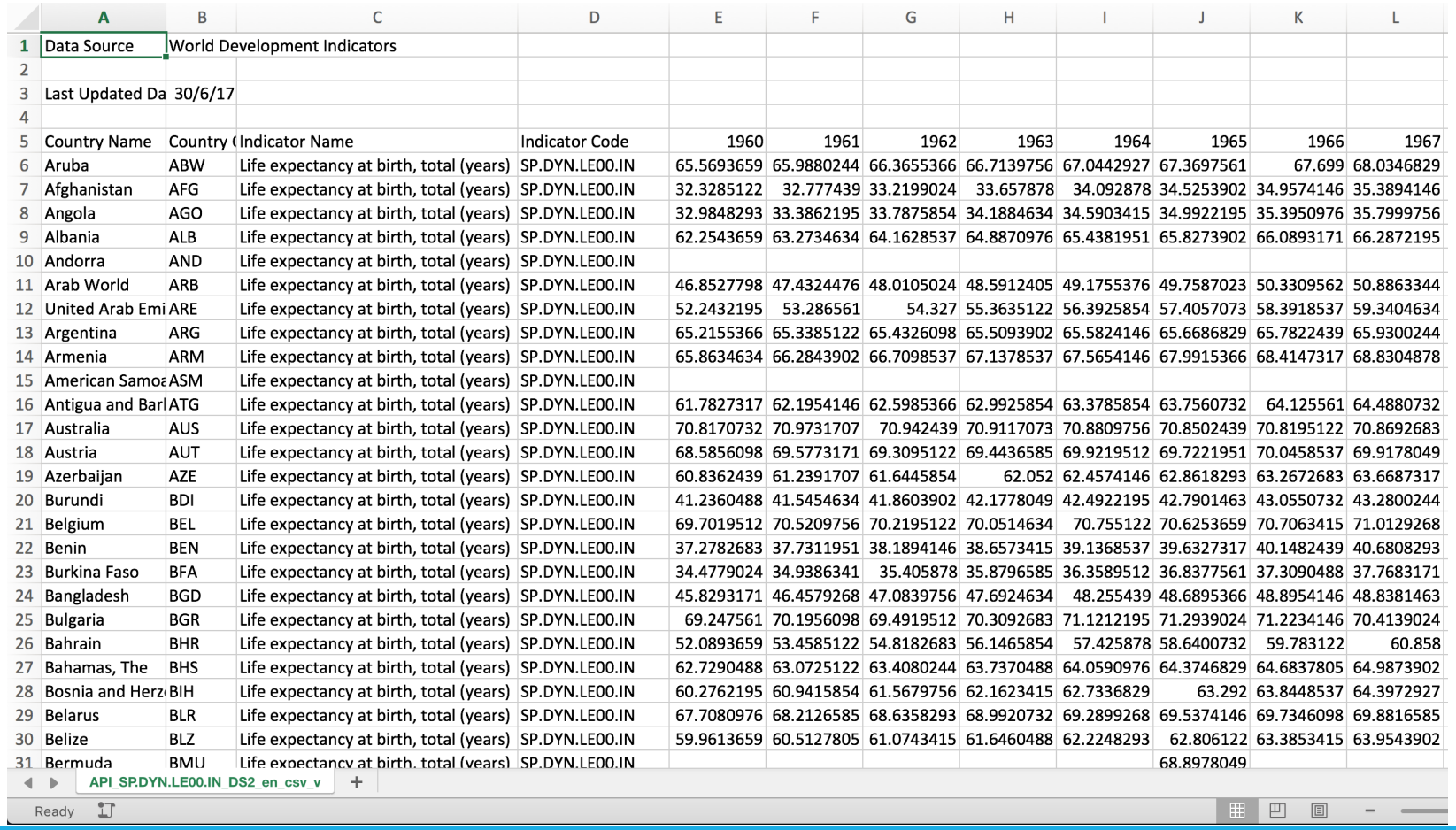
Example – World Bank Data

All Countries and Economies

Country	1960	2015	
Zimbabwe	52	59	
Zambia	45	61	
Yemen, Rep.	34	64	
West Bank and Gaza		73	
Virgin Islands (U.S.)	64	80	
Vietnam	59	76	

Data from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

Example – Importing Data



	A	B	C	D	E	F	G	H	I	J	K	L
1	Data Source	World Development Indicators										
2												
3	Last Updated Da	30/6/17										
4												
5	Country Name	Country (	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965	1966	1967
6	Aruba	ABW	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	65.5693659	65.9880244	66.3655366	66.7139756	67.0442927	67.3697561	67.699	68.0346829
7	Afghanistan	AFG	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	32.3285122	32.777439	33.2199024	33.657878	34.092878	34.5253902	34.9574146	35.3894146
8	Angola	AGO	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	32.9848293	33.3862195	33.7875854	34.1884634	34.5903415	34.9922195	35.3950976	35.7999756
9	Albania	ALB	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	62.2543659	63.2734634	64.1628537	64.8870976	65.4381951	65.8273902	66.0893171	66.2872195
10	Andorra	AND	Life expectancy at birth, total (years)	SP.DYN.LE00.IN								
11	Arab World	ARB	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	46.8527798	47.4324476	48.0105024	48.5912405	49.1755376	49.7587023	50.3309562	50.8863344
12	United Arab Emi	ARE	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	52.2432195	53.286561	54.327	55.3635122	56.3925854	57.4057073	58.3918537	59.3404634
13	Argentina	ARG	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	65.2155366	65.3385122	65.4326098	65.5093902	65.5824146	65.6686829	65.7822439	65.9300244
14	Armenia	ARM	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	65.8634634	66.2843902	66.7098537	67.1378537	67.5654146	67.9915366	68.4147317	68.8304878
15	American Samoa	ASM	Life expectancy at birth, total (years)	SP.DYN.LE00.IN								
16	Antigua and Barl	ATG	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	61.7827317	62.1954146	62.5985366	62.9925854	63.3785854	63.7560732	64.125561	64.4880732
17	Australia	AUS	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	70.8170732	70.9731707	70.942439	70.9117073	70.8809756	70.8502439	70.8195122	70.8692683
18	Austria	AUT	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	68.5856098	69.5773171	69.3095122	69.4436585	69.9219512	69.7221951	70.0458537	69.9178049
19	Azerbaijan	AZE	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	60.8362439	61.2391707	61.6445854	62.052	62.4574146	62.8618293	63.2672683	63.6687317
20	Burundi	BDI	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	41.2360488	41.5454634	41.8603902	42.1778049	42.4922195	42.7901463	43.0550732	43.2800244
21	Belgium	BEL	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	69.7019512	70.5209756	70.2195122	70.0514634	70.755122	70.6253659	70.7063415	71.0129268
22	Benin	BEN	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	37.2782683	37.7311951	38.1894146	38.6573415	39.1368537	39.6327317	40.1482439	40.6808293
23	Burkina Faso	BFA	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	34.4779024	34.9386341	35.405878	35.8796585	36.3589512	36.8377561	37.3090488	37.7683171
24	Bangladesh	BGD	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	45.8293171	46.4579268	47.0839756	47.6924634	48.255439	48.6895366	48.8954146	48.8381463
25	Bulgaria	BGR	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	69.247561	70.1956098	69.4919512	70.3092683	71.1212195	71.2939024	71.2234146	70.4139024
26	Bahrain	BHR	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	52.0893659	53.4585122	54.8182683	56.1465854	57.425878	58.6400732	59.783122	60.858
27	Bahamas, The	BHS	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	62.7290488	63.0725122	63.4080244	63.7370488	64.0590976	64.3746829	64.6837805	64.9873902
28	Bosnia and Herz	BIH	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	60.2762195	60.9415854	61.5679756	62.1623415	62.7336829	63.292	63.8448537	64.3972927
29	Belarus	BLR	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	67.7080976	68.2126585	68.6358293	68.9920732	69.2899268	69.5374146	69.7346098	69.8816585
30	Belize	BLZ	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	59.9613659	60.5127805	61.0743415	61.6460488	62.2248293	62.806122	63.3853415	63.9543902
31	Bermuda	BMU	Life expectancy at birth, total (years)	SP.DYN.LE00.IN						68.8978049		

API_SP.DYN.LE00.IN_DS2_en_csv_v

Ready

Data from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

Code is here: http://ibm.biz/WBExample

Example – Tidy and Transform

```
> dt %>% print(n=15)
# A tibble: 3,030 x 6
       Country                      Region  Year Population Life.Expectancy Health.Expenditure
        <fctr>                       <chr> <dbl>      <dbl>           <dbl>              <dbl>
 1      Albania       Europe & Central Asia  1995    3141102        71.87029           28.22459
 2      Algeria Middle East & North Africa  1995   28291591        68.46588           62.05589
 3       Angola          Sub-Saharan Africa  1995   12105105        42.05093           20.74863
 4    Argentina  Latin America & Caribbean  1995   34855160        72.62363          615.41426
 5      Armenia       Europe & Central Asia  1995    3223173        68.62046           25.73810
 6    Australia         East Asia & Pacific  1995   18072000        77.82927         1564.94429
 7      Austria       Europe & Central Asia  1995    7948278        76.71561         2876.07979
 8   Azerbaijan       Europe & Central Asia  1995    7685000        64.57583           18.03310
 9 Bahamas, The  Latin America & Caribbean  1995     279774        70.05637          839.60625
10      Bahrain Middle East & North Africa  1995     559069        73.14702          481.88143
11   Bangladesh                  South Asia  1995  117486952        62.12268           11.37606
12     Barbados  Latin America & Caribbean  1995     263416        74.95937          450.10554
13      Belarus       Europe & Central Asia  1995   10194000        68.46098           69.12994
14      Belgium       Europe & Central Asia  1995   10136811        76.84073         2136.33351
15       Belize  Latin America & Caribbean  1995     216500        73.12761          120.49518
# ... with 3,015 more rows
> 
```

Data from https://github.com/jcheng5/googleCharts/tree/master/inst/examples/bubble

Data Exploration

Transforming data for humans into data for computers

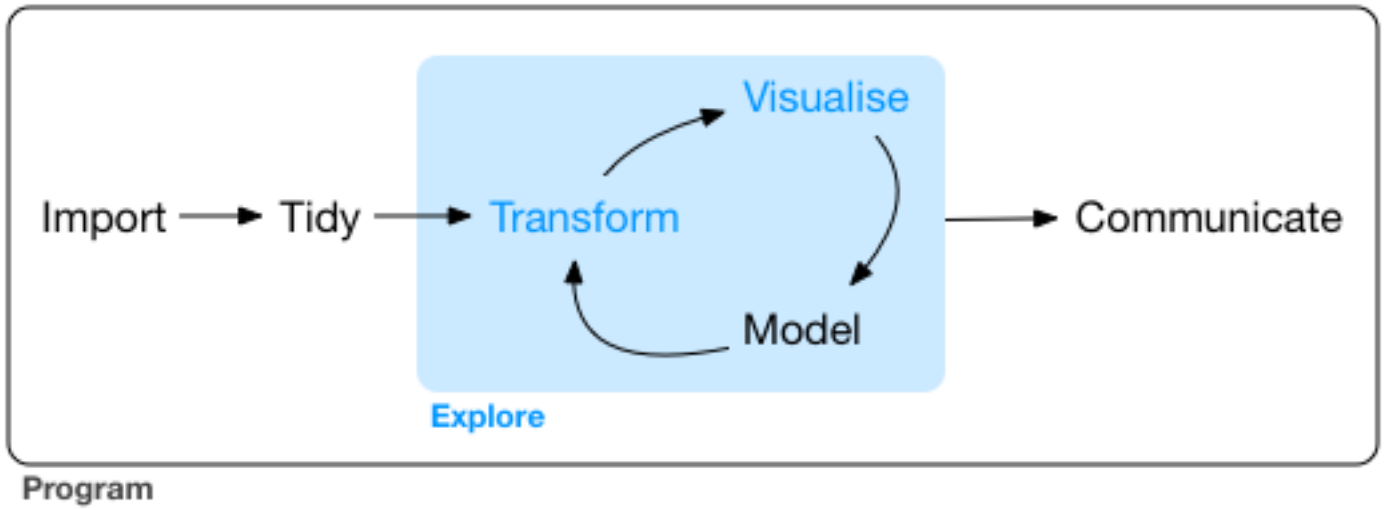
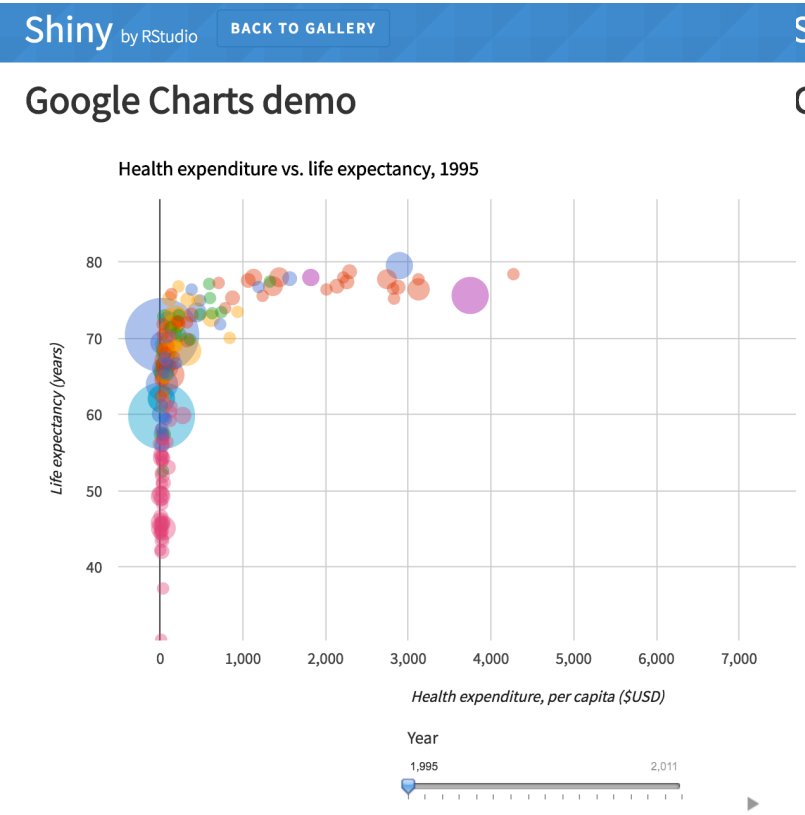
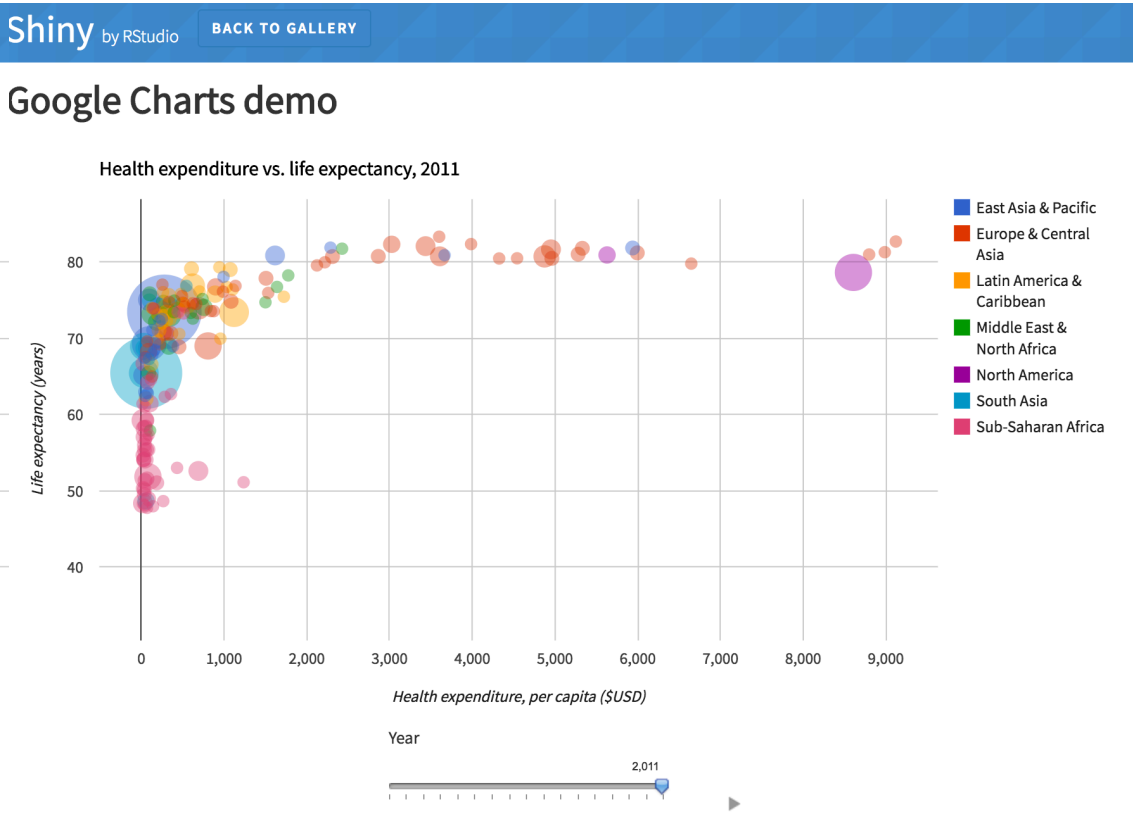


Image from http://r4ds.had.co.nz/wrangle-intro.html

Example – Visualising





Screenshot of https://shiny.rstudio.com/gallery/google-charts.html

Communicate Your Data

Transforming data for humans into data for computers

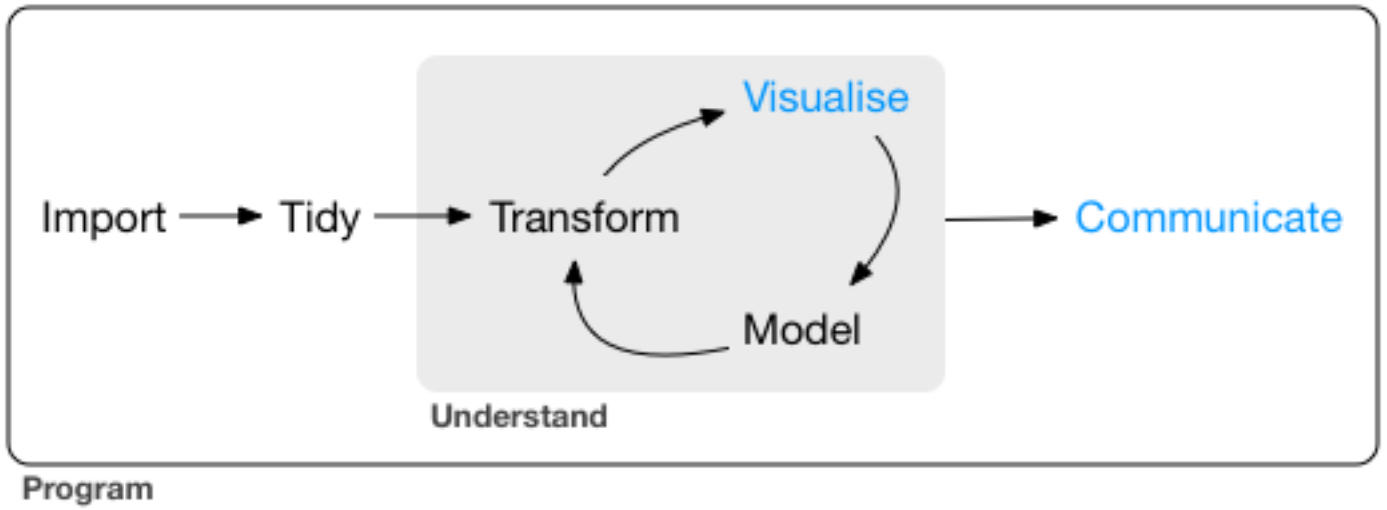


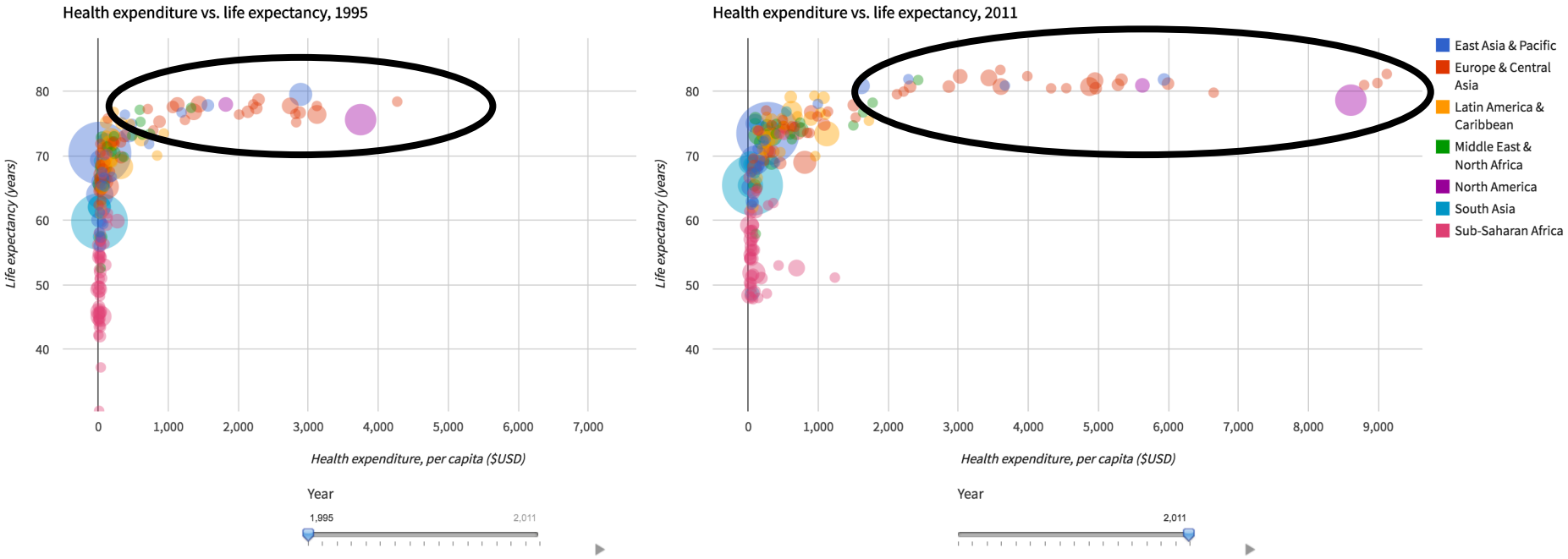
Image from http://r4ds.had.co.nz/communicate-intro.html

Example – Communicate

Shiny by RStudio BACK TO GALLERY

Google Charts demo

The US and European countries have increased healthcare spending dramatically, but they have reached the limit of human life expectancy.



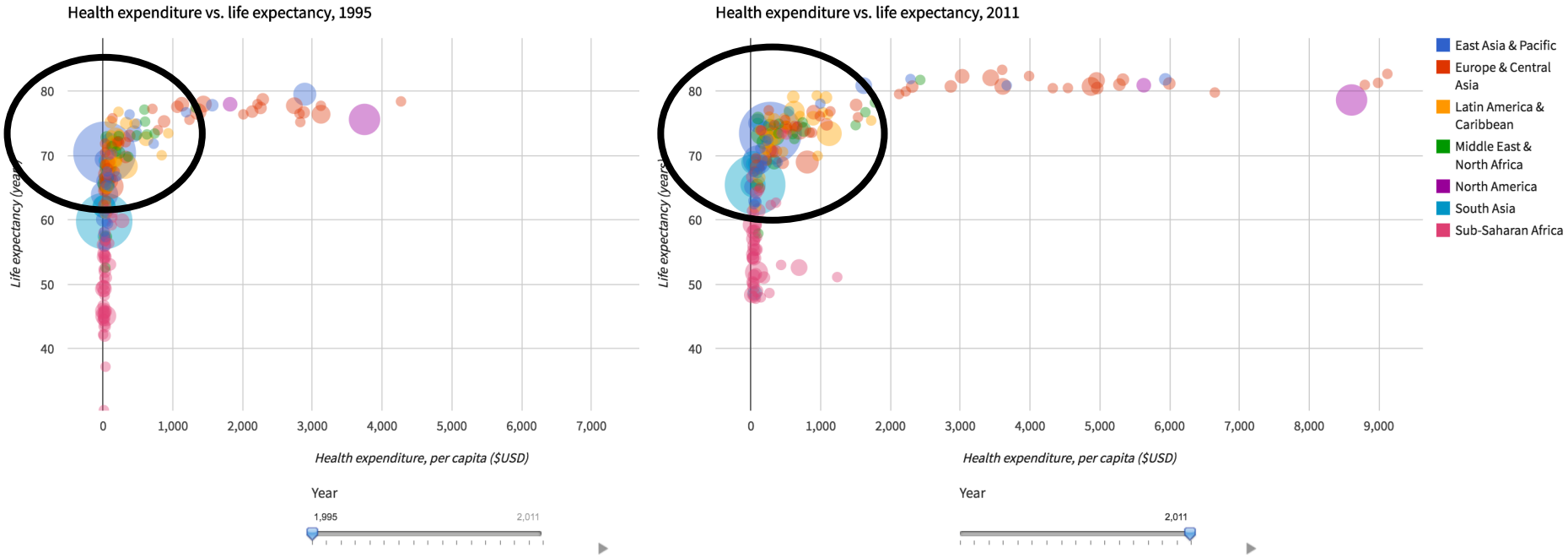
Screenshot of https://shiny.rstudio.com/gallery/google-charts.html

Example – Communicate

Shiny by RStudio BACK TO GALLERY

Google Charts demo

In contrast, developing countries have increased their life expectancy with relatively minor increases in healthcare spending.



Screenshot of https://shiny.rstudio.com/gallery/google-charts.html

Data Modelling

More on this in the Machine Learning section

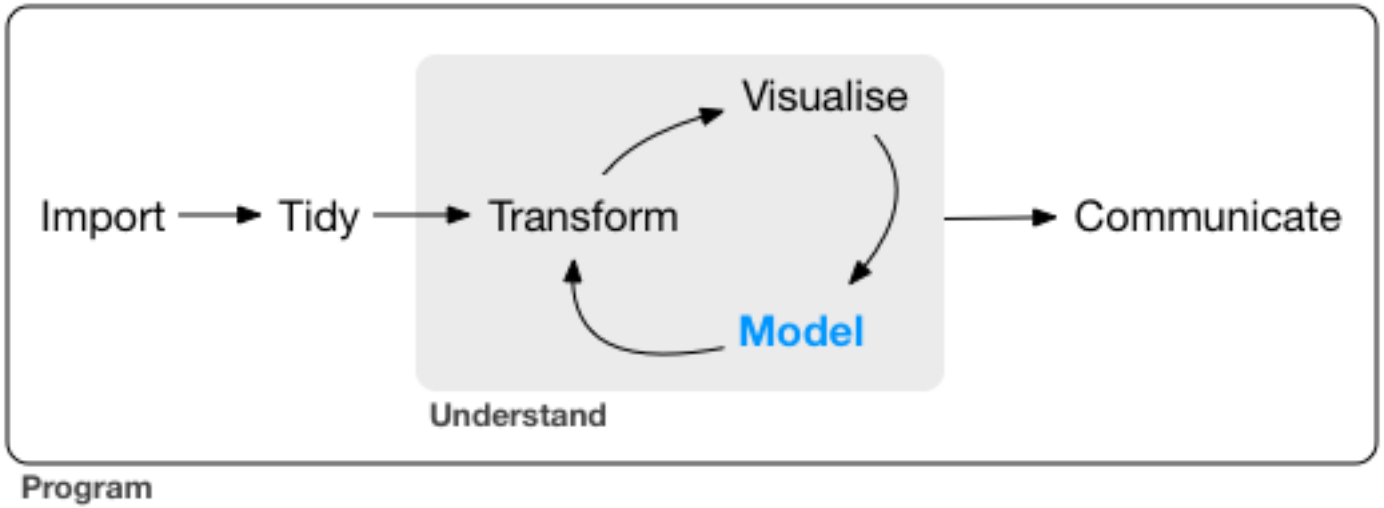


Image from http://r4ds.had.co.nz/model-intro.html

Data Science – Frontiers

Data to Decisions

- Real-time dashboards for big data
- Data Science combined with Decision Science
- Data influence our decisions
- Machines making decisions on our behalf

Internet of Things

- 100s of billion of devices connected to the internet
- E.g. fridges, washing machines, coffee makers, alarm clocks, etc
- People-People, People-Things, Things-Things
- IBM Internet of Things: https://www.ibm.com/internet-of-things/

Data Science – IBM

IBM Data Science Experience (DSE)

- https://datascience.ibm.com/

Data Science platforms

- IBM Watson on the Bluemix cloud: http://bluemix.net (free 30 day trial)
- 12 month trial for academic email addresses: https://ibm.onthehub.com

IBM DSE Notebook for World Bank Example

- http://ibm.biz/WBExample
- Find this presentation here: kndtran.com

Data Science – Australia

Melbourne Data Science meetup

- https://www.meetup.com/en-AU/Data-Science-Melbourne/
- http://www.datasciencemelbourne.com/
- Annual Datathon: http://www.datasciencemelbourne.com/datathon/

Open data initiatives – open data and open knowledge foundations

- http://data.gov.au/
- https://data.melbourne.vic.gov.au/

Govhack

- https://govhack.org/

Increasing demand for data scientists

Data Science – Vietnam

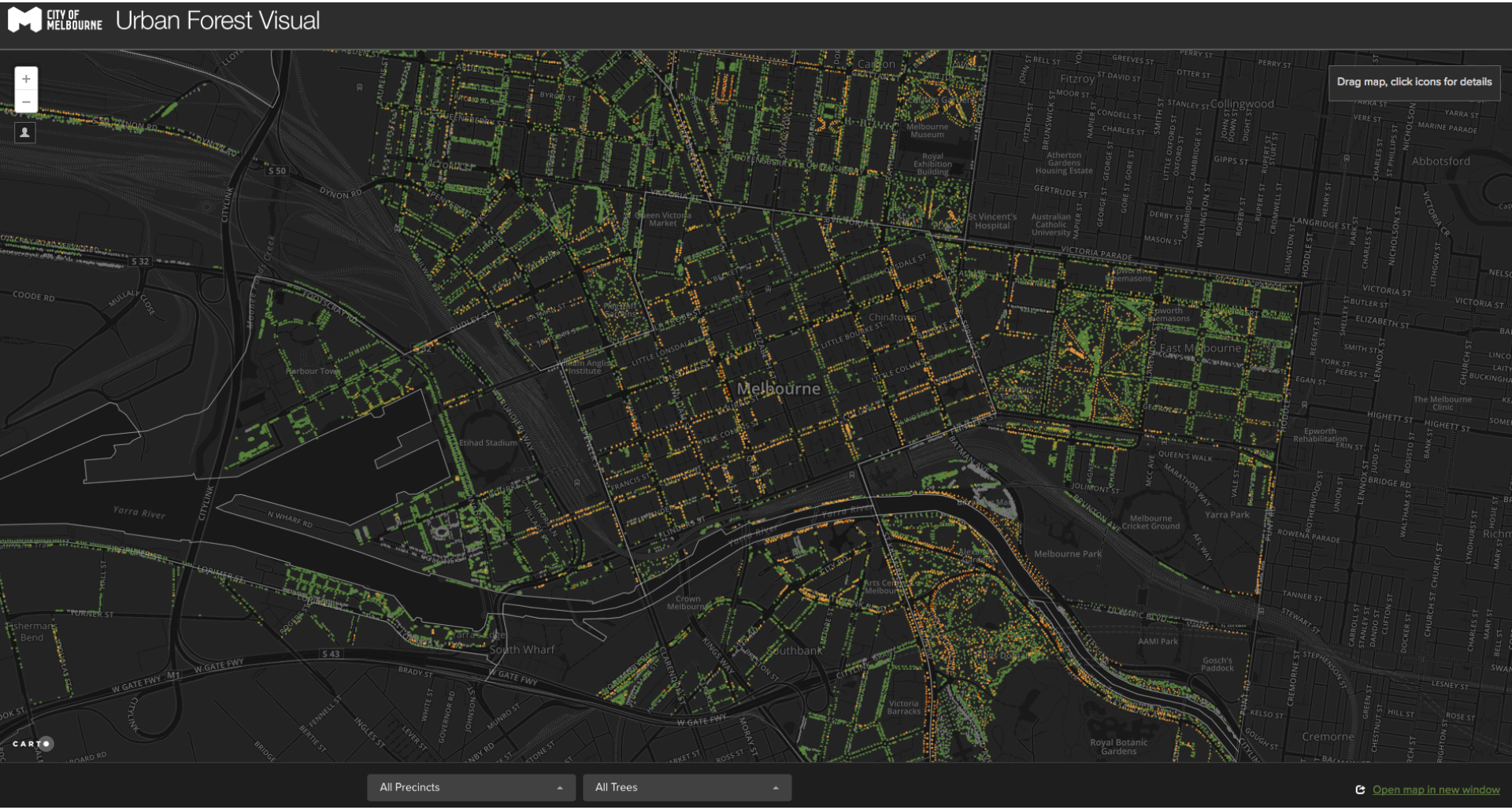
I need your help.

What is happening here? What companies use/need data scientists?

Connect and tell me what you know!

General observations

- Generate more data
- Collect data
- Open up data
- Demand data generation platforms
- Data hackathons?
- Connect to the Internet



Screenshot from http://melbourneurbanforestvisual.com.au/bigmap.html

Where do you start?

Programming Languages
- R, Python, SQL

Resources
- News: http://www.kdnuggets.com
- Data Sets: https://www.kaggle.com
- Data Science platforms
  - IBM Watson on the Bluemix cloud: http://bluemix.net (free 30 day trial)
  - 12 month trial for academic email addresses: https://ibm.onthehub.com
- R for Data Science book (online and free)
  - http://r4ds.had.co.nz
- Python for Data Science (Anaconda package)
  - https://www.datacamp.com/learn-python-with-anaconda

Focus on Visualisations and Storytelling
- https://github.com/d3/d3/wiki/Gallery

Social Network for data people
- https://data.world

Data Science Questions?

kndtran.com

Machine Learning

Importance

Learn and make predictions on data

Collection of statistical algorithms

Automatically discover insights from data

Some tasks are simply too difficult to program a solution

For some tasks, with a certain amount of data, machines will learn more from data than from experts

Machines don’t get tired

Why you need to know?

Businesses are realising some of their work can be automated

Costs of cloud services are dropping

Machine learning moved onto the cloud

Deep Learning is the latest revolution in Machine Learning

Unlocking even more complex tasks

Businesses want more tasks to be automated

Need people who can apply machine learning to business processes

ML around you

Spam email filters

- This problem is now essentially solved

Recommendations

- Search
- Advertising
- Purchases

Sentiment on feedback

- Customer reviews
- Social media

Machine Learning Lifecycle

Extending the Data Science lifecycle

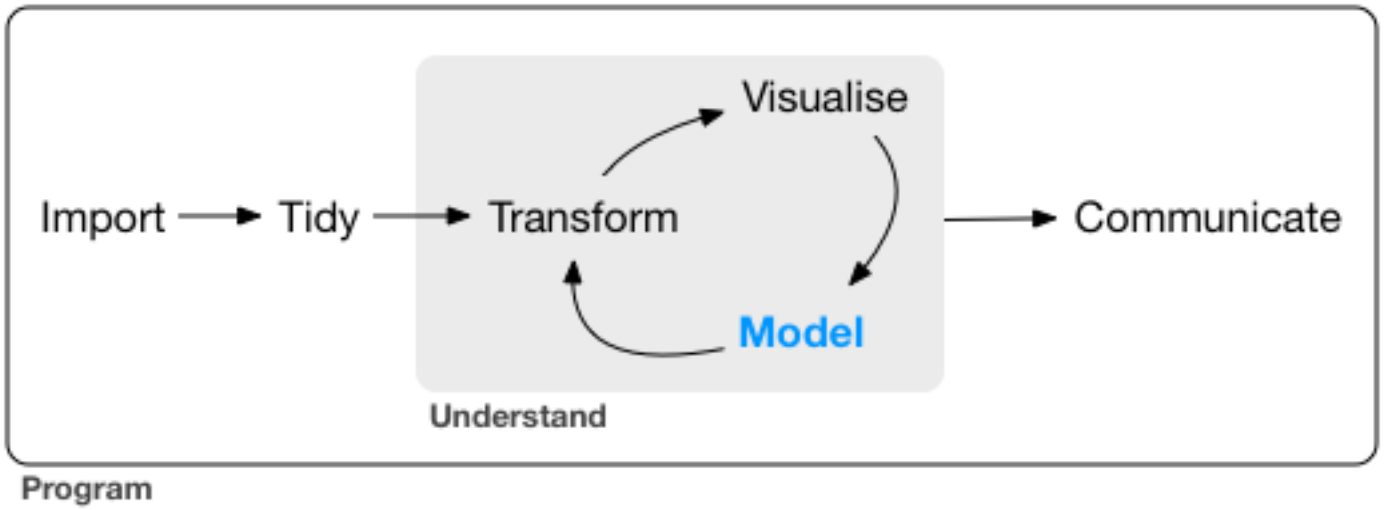
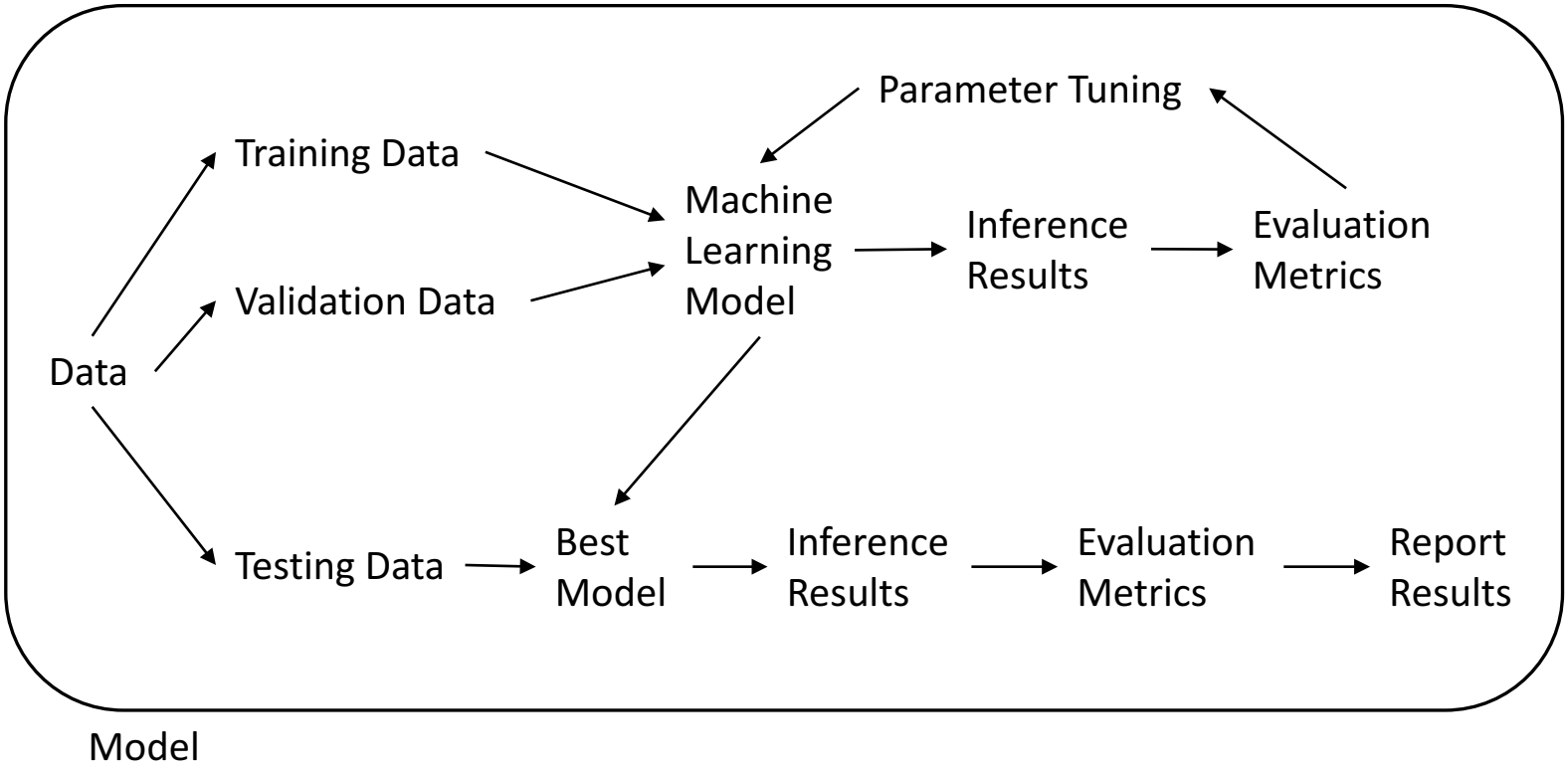


Image from http://r4ds.had.co.nz/model-intro.html

Machine Learning Lifecycle



Example

Distinguishing homes in New York and in San Francisco

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Find this presentation here: kndtran.com

Data Exploration

Transforming data for humans into data for computers

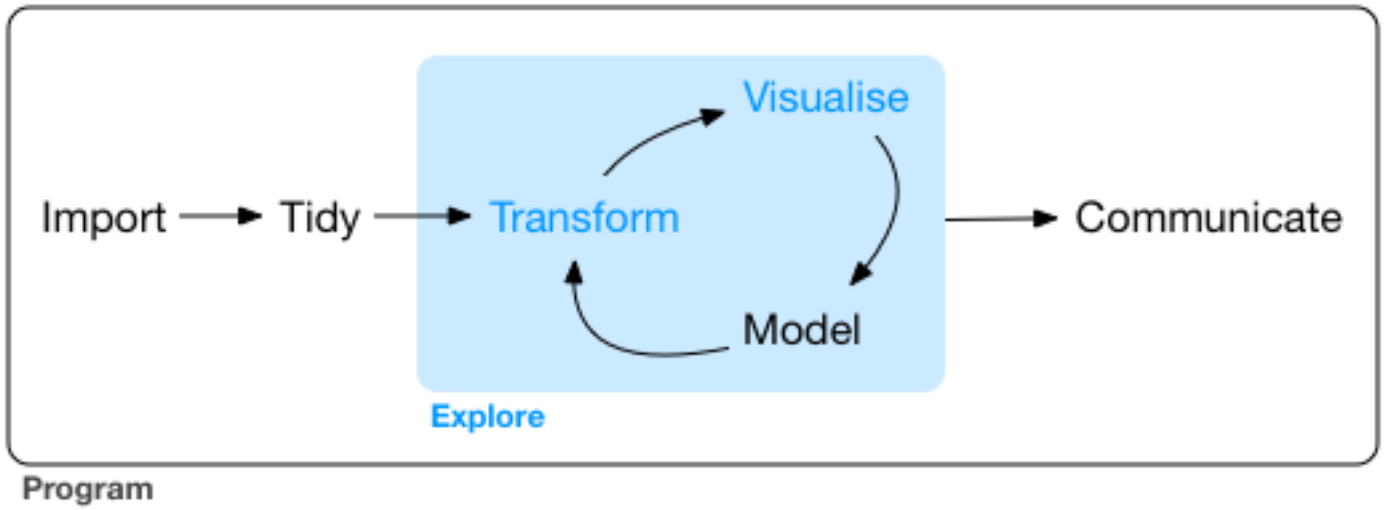
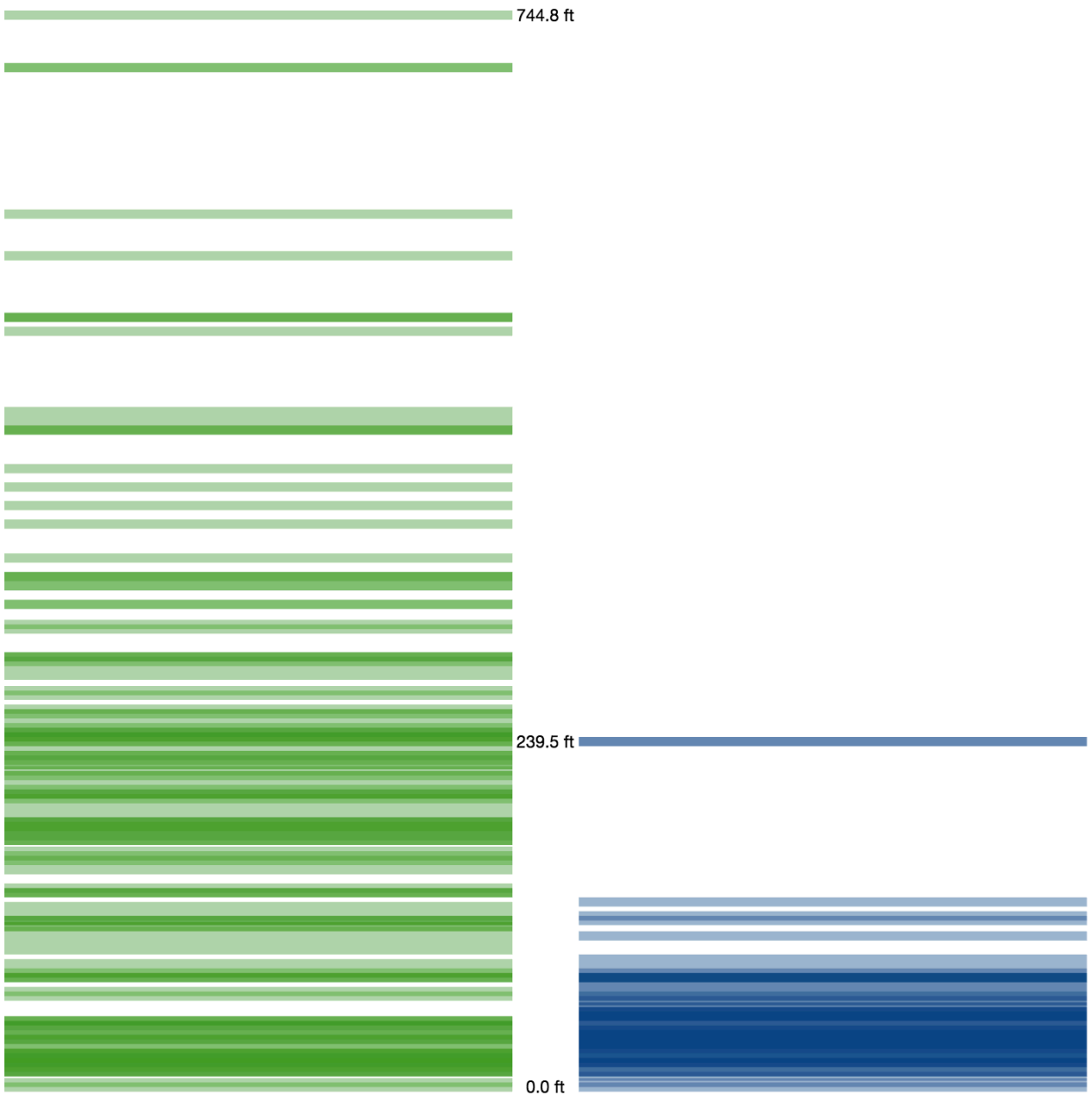
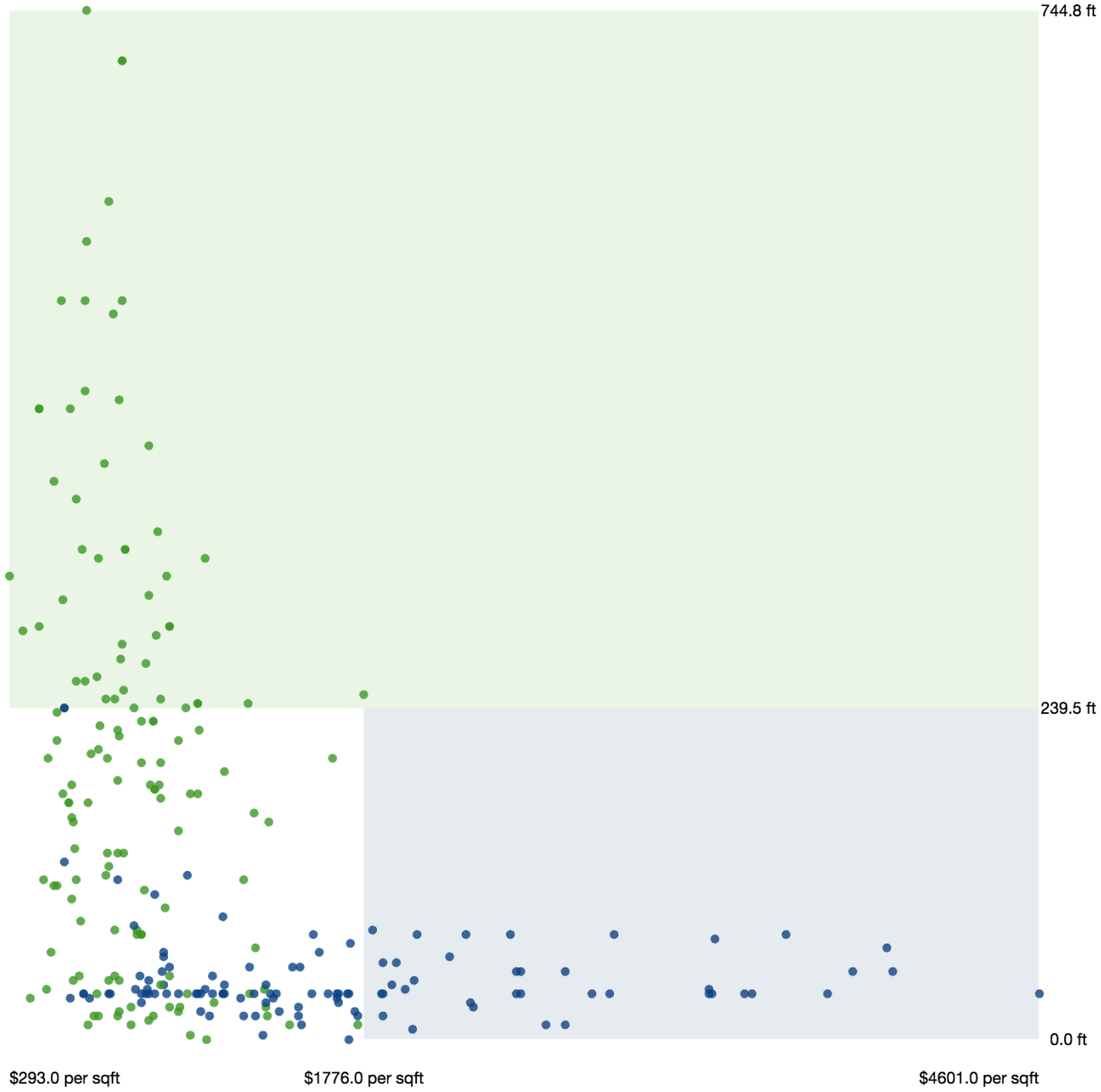


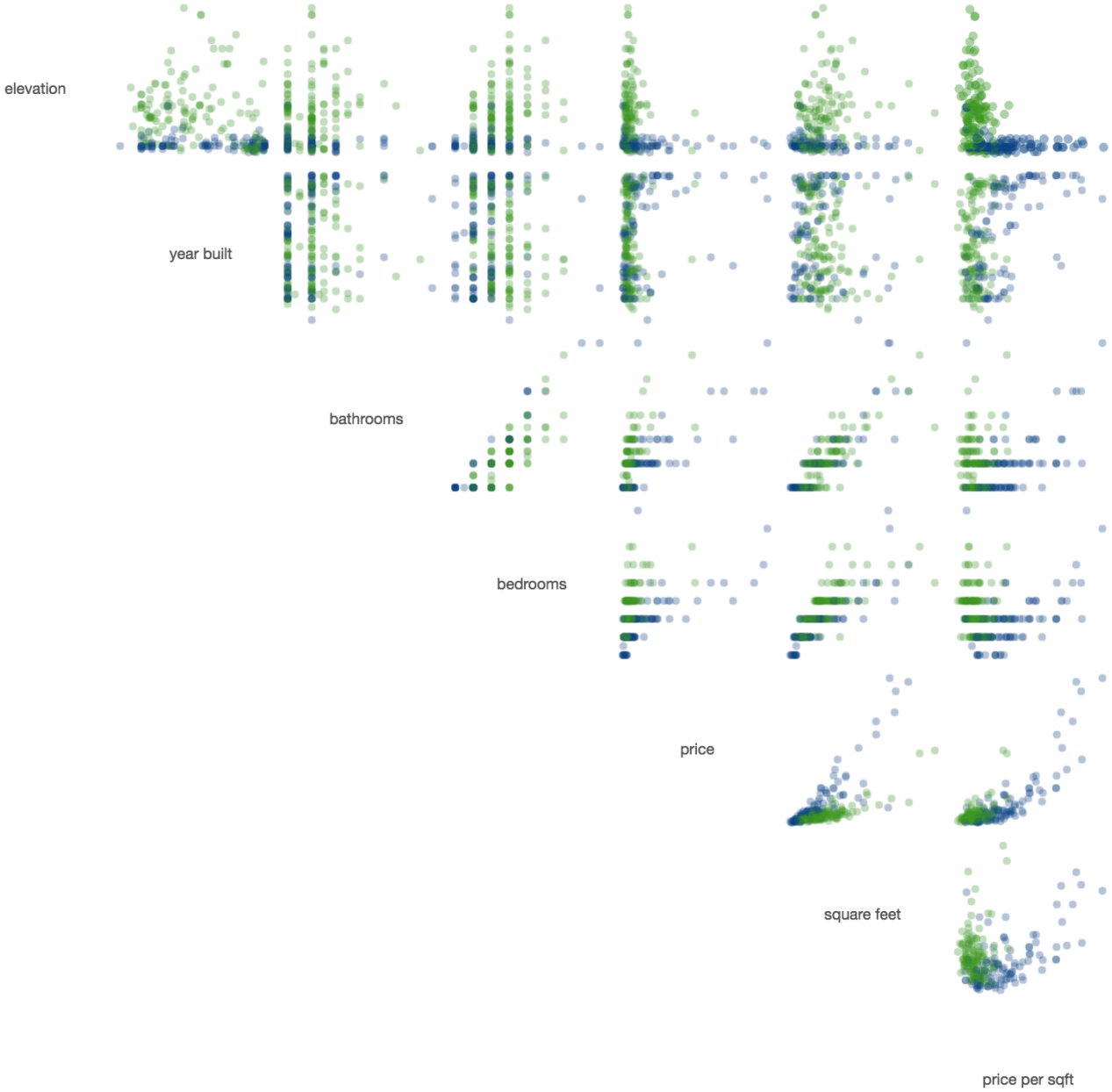
Image from http://r4ds.had.co.nz/wrangle-intro.html



Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

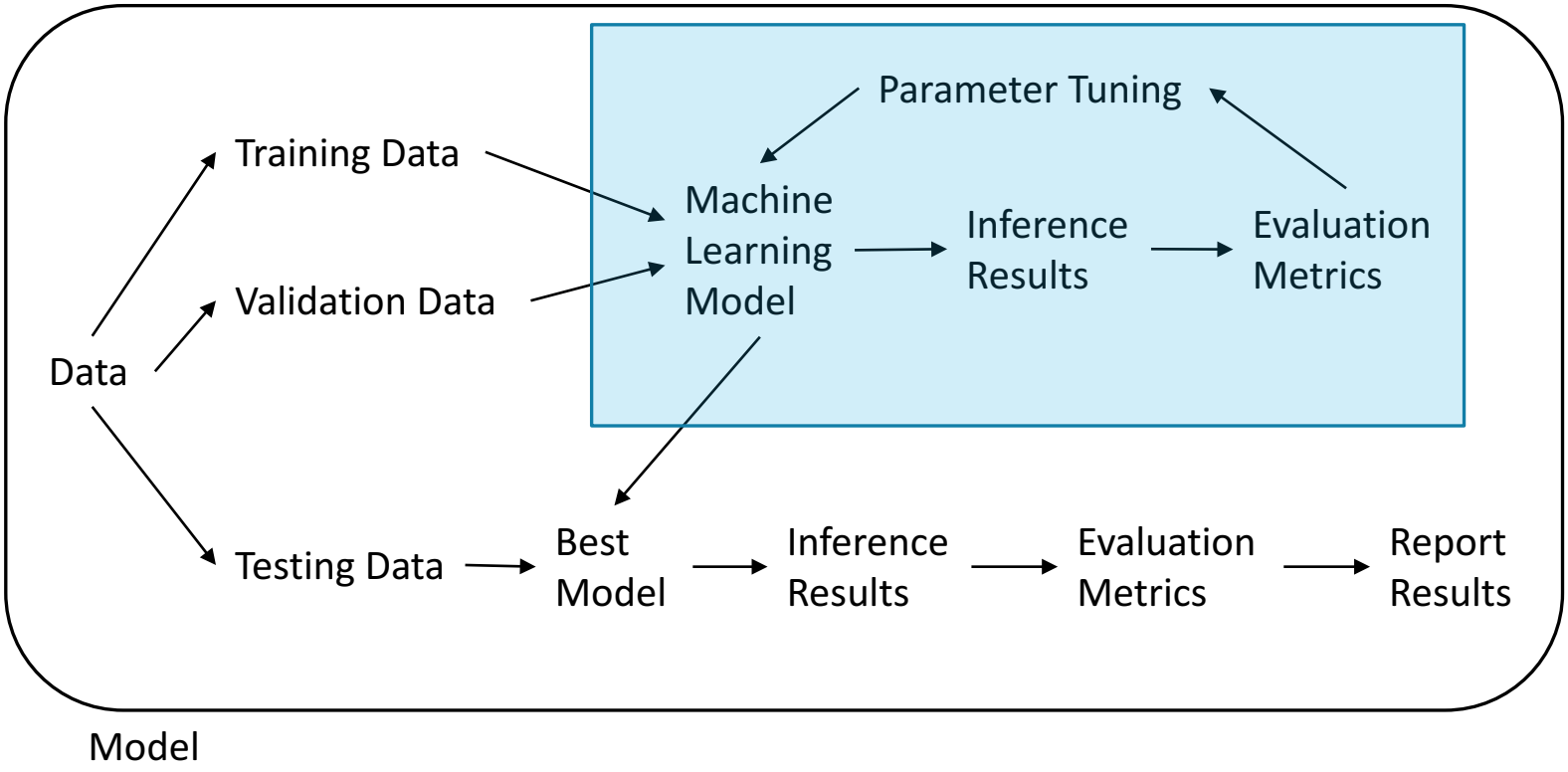


Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Machine Learning Lifecycle



Decision Tree

A simple and intuitive machine learning model

With many features, choose the best split in the feature that partitions the data set

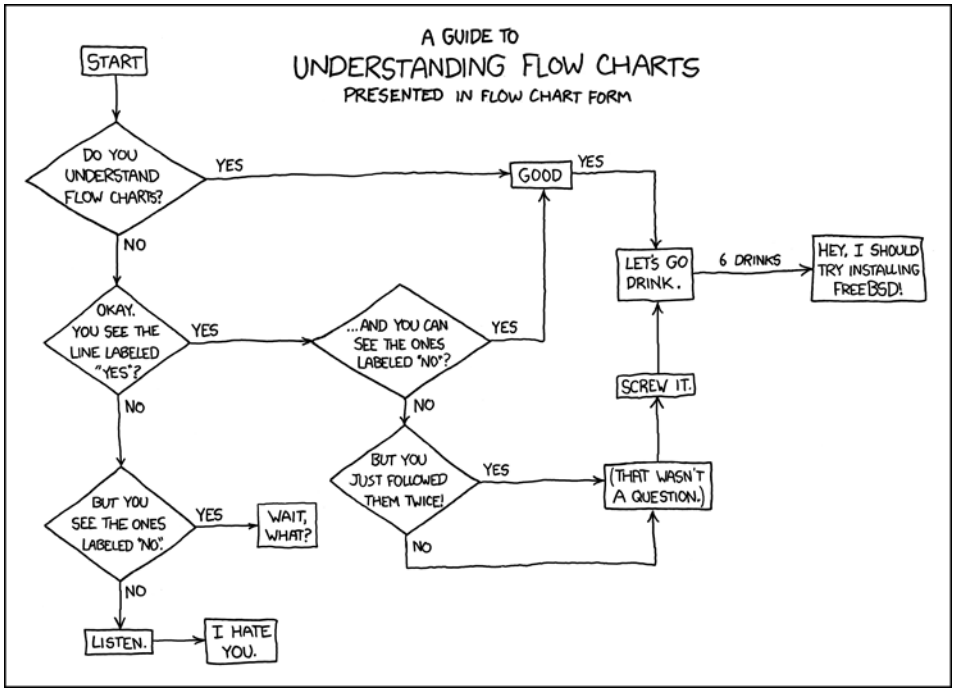
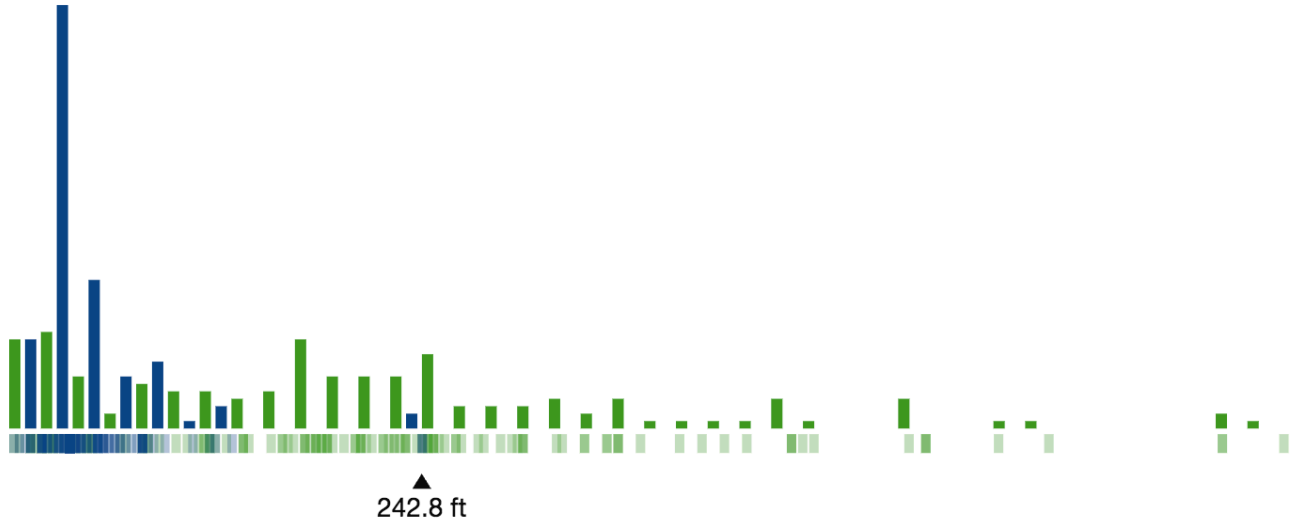
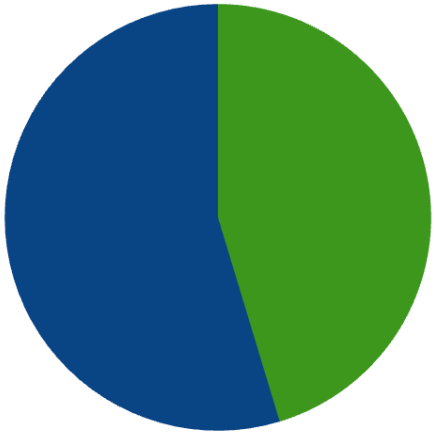


Image from https://xkcd.com/518/

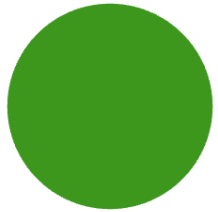




111
NY

92
SF

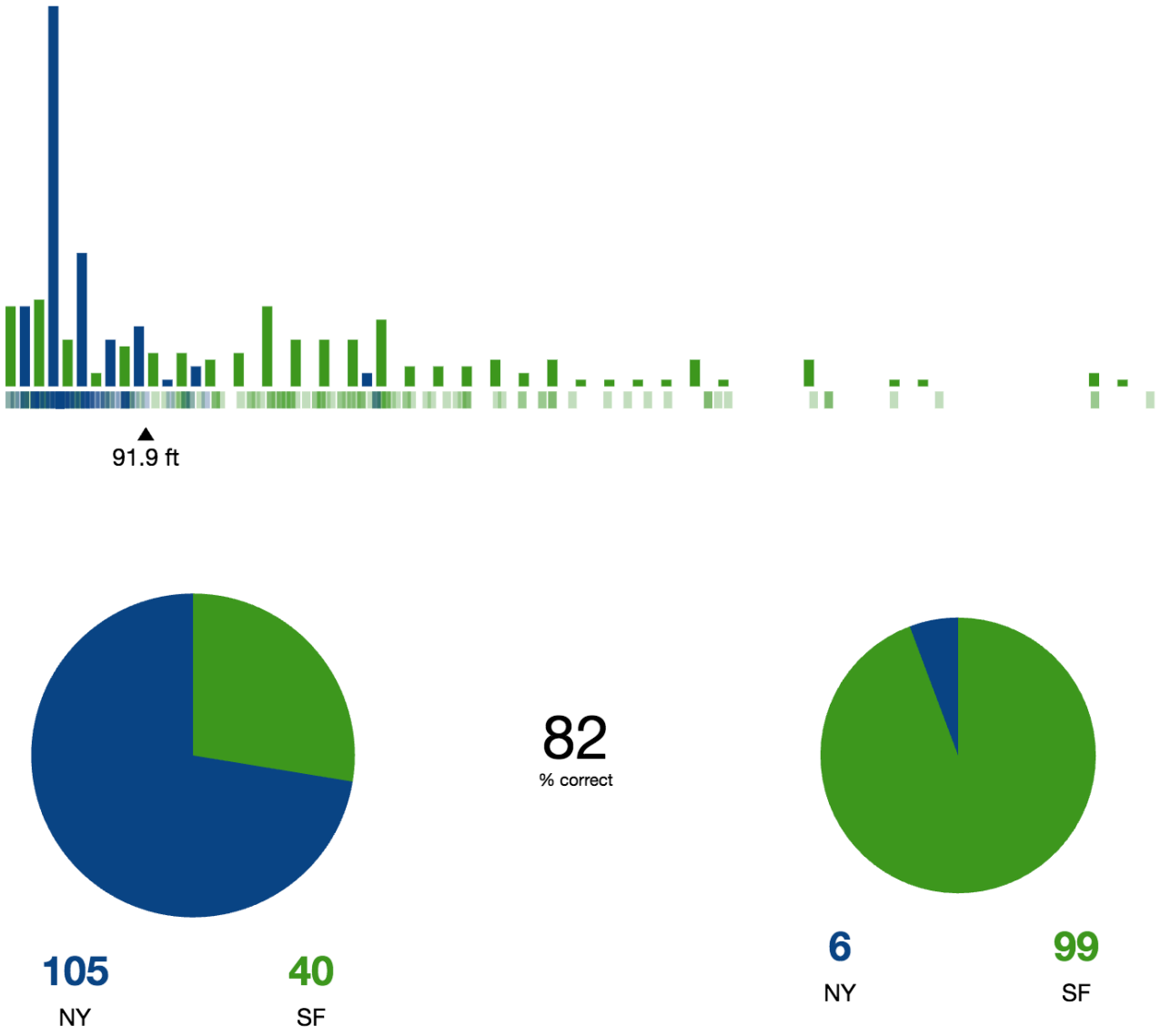
63
% correct



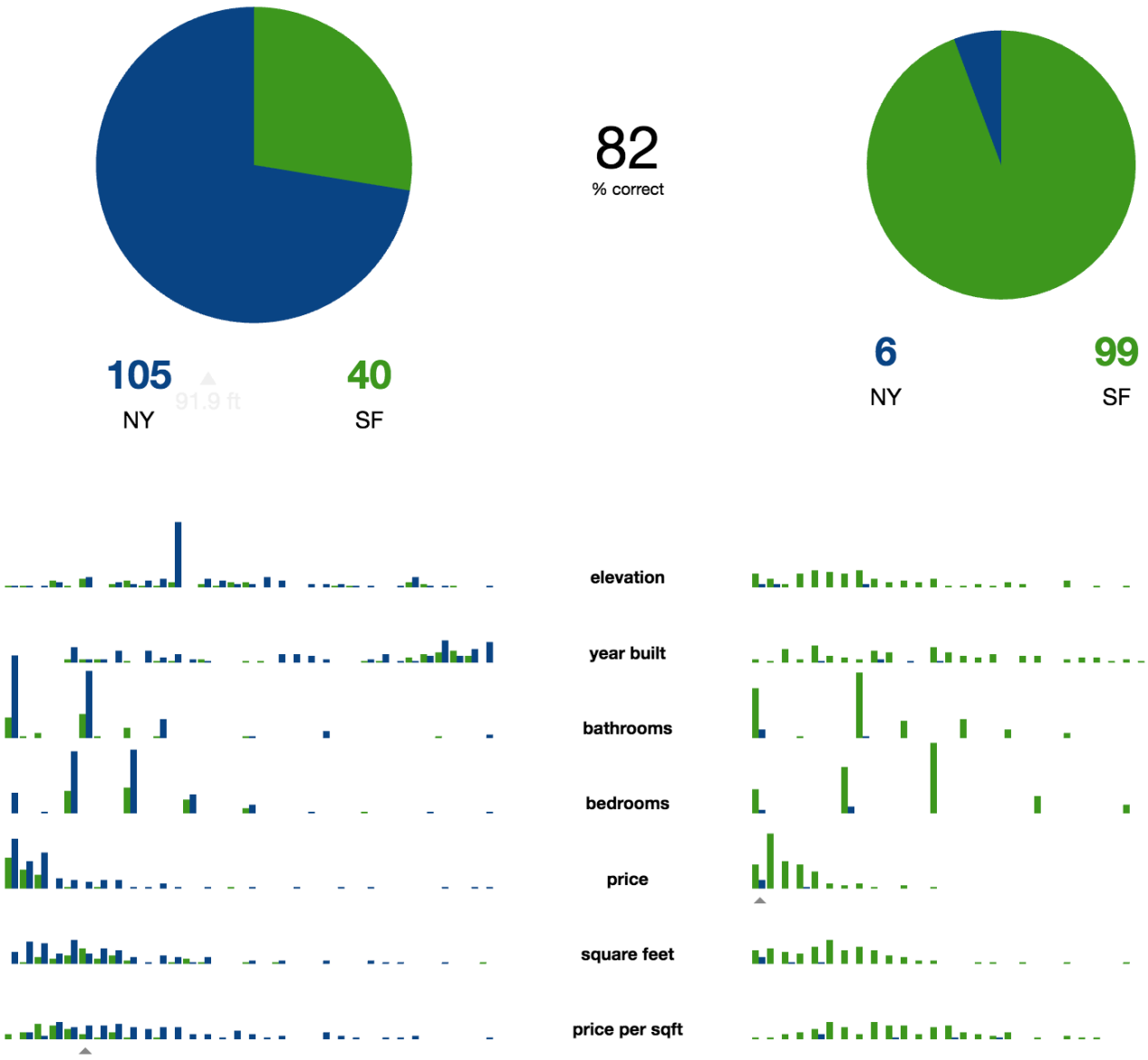
0
NY

47
SF

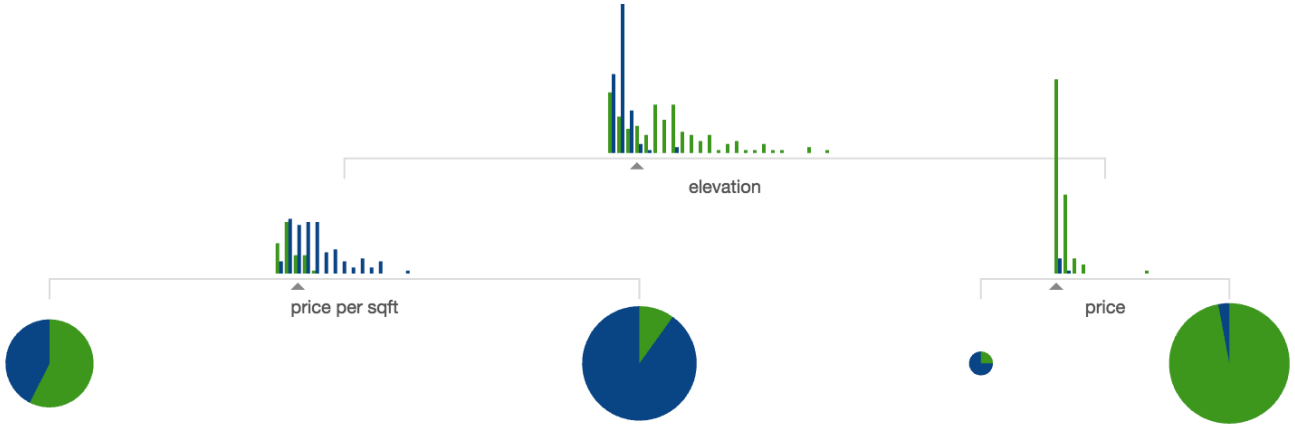
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



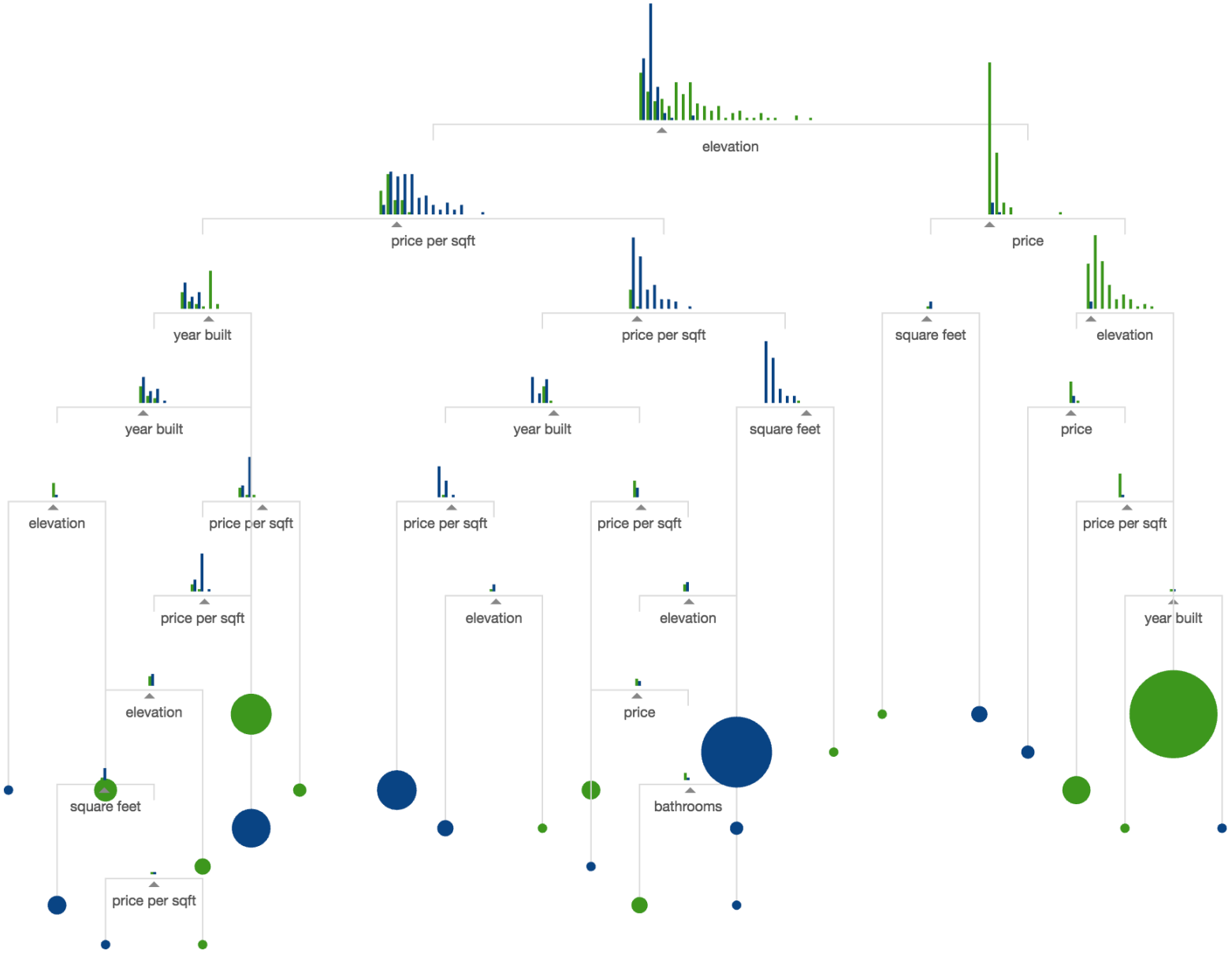
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



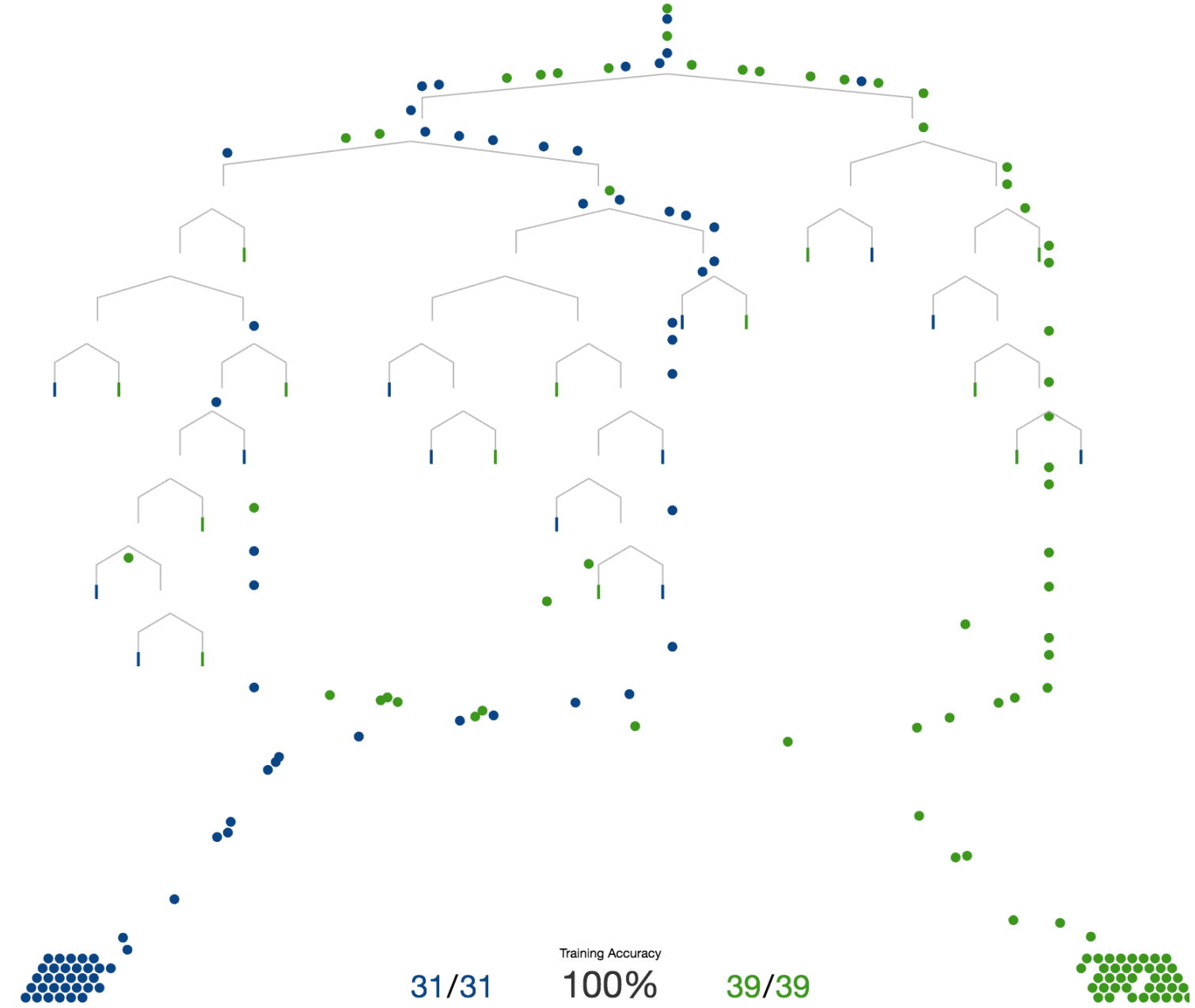
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



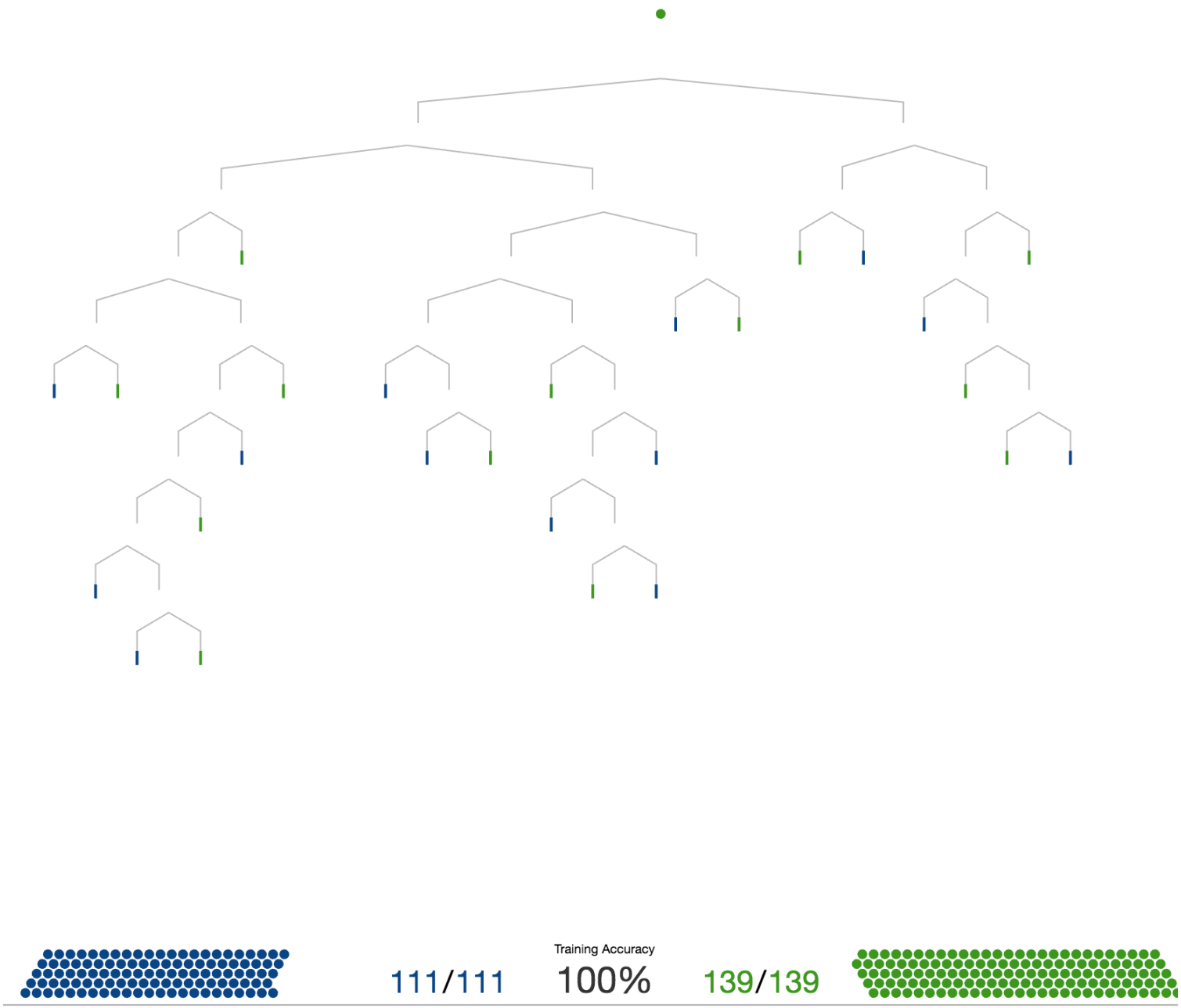
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

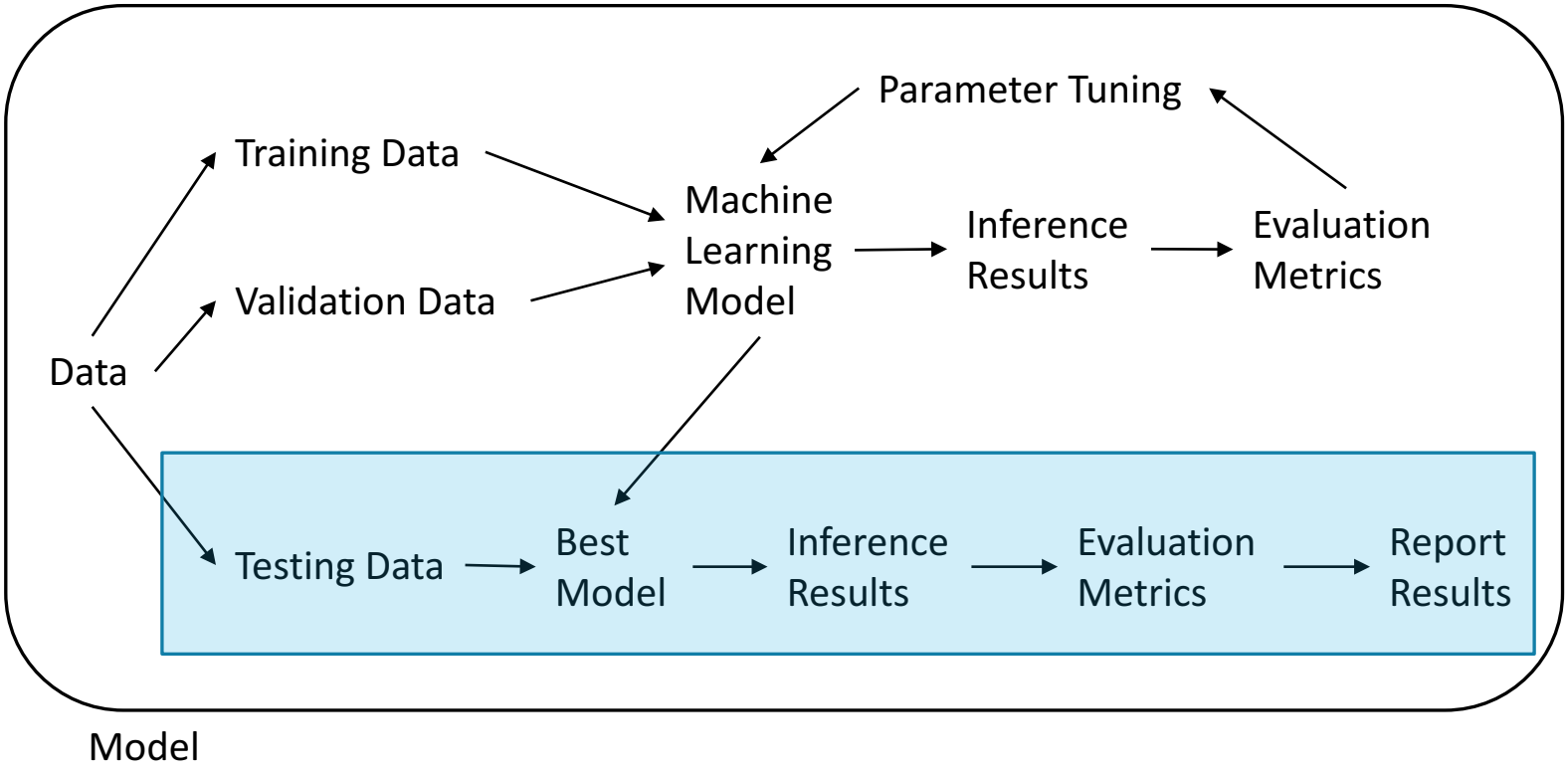


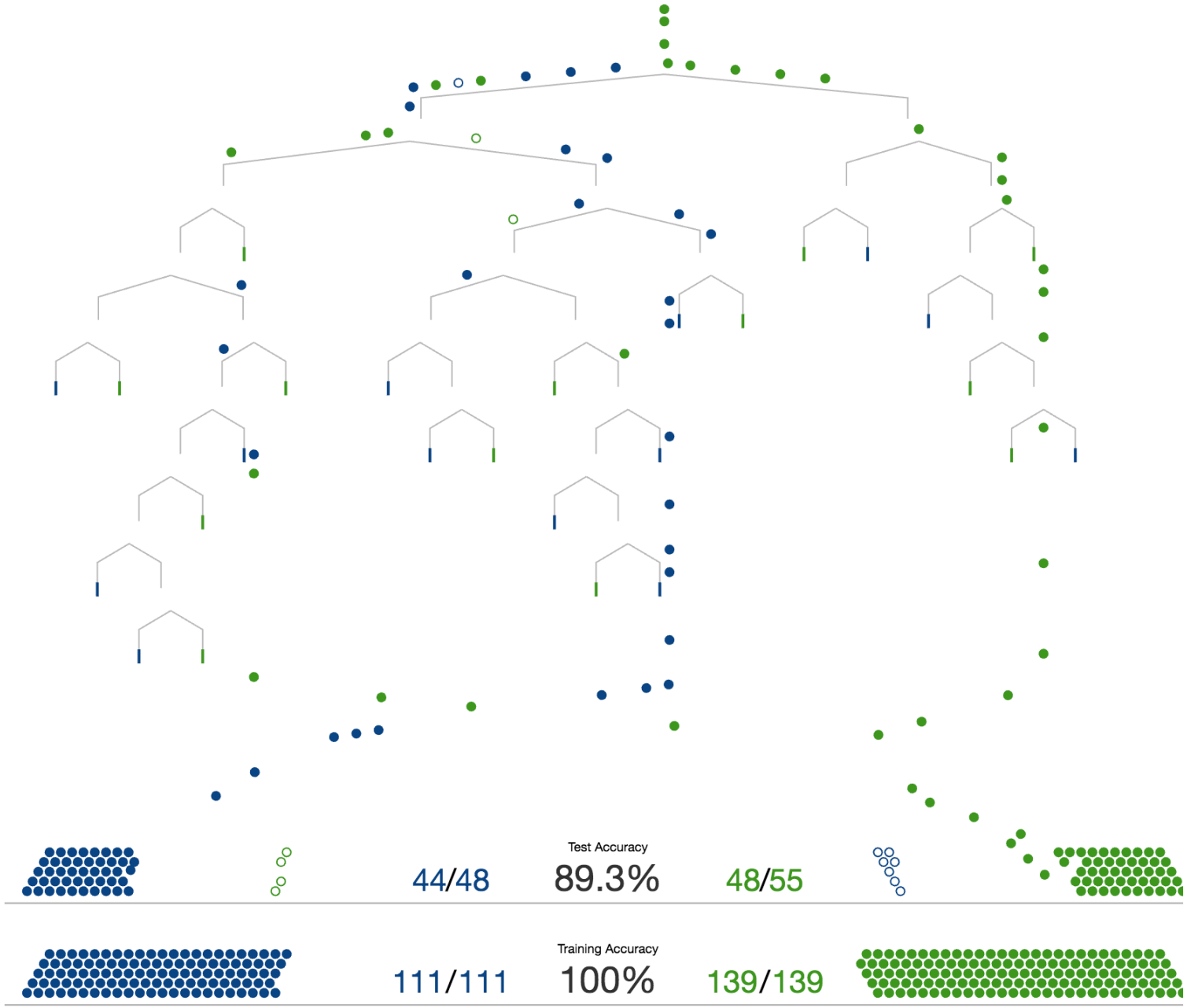
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



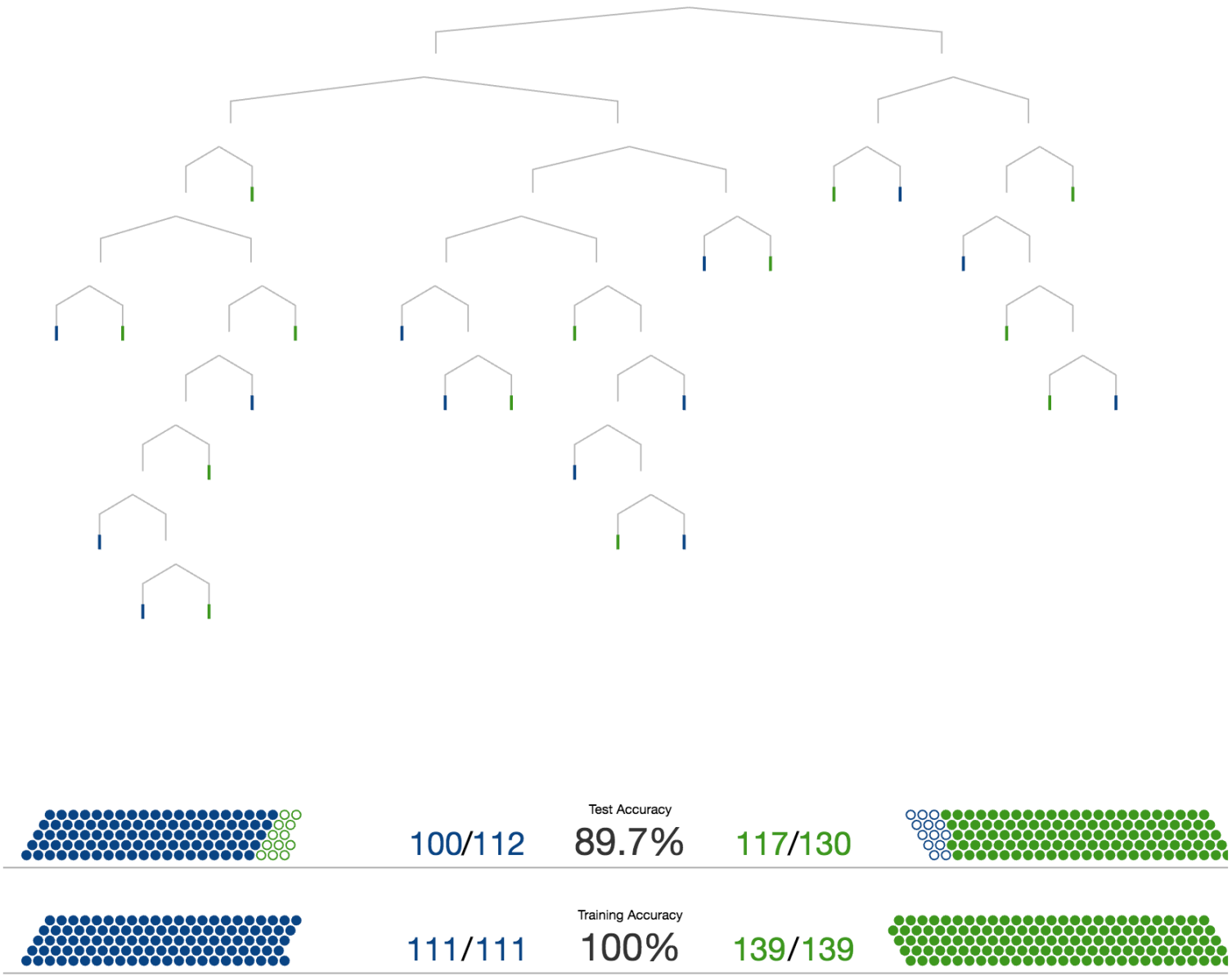
Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Machine Learning Lifecycle



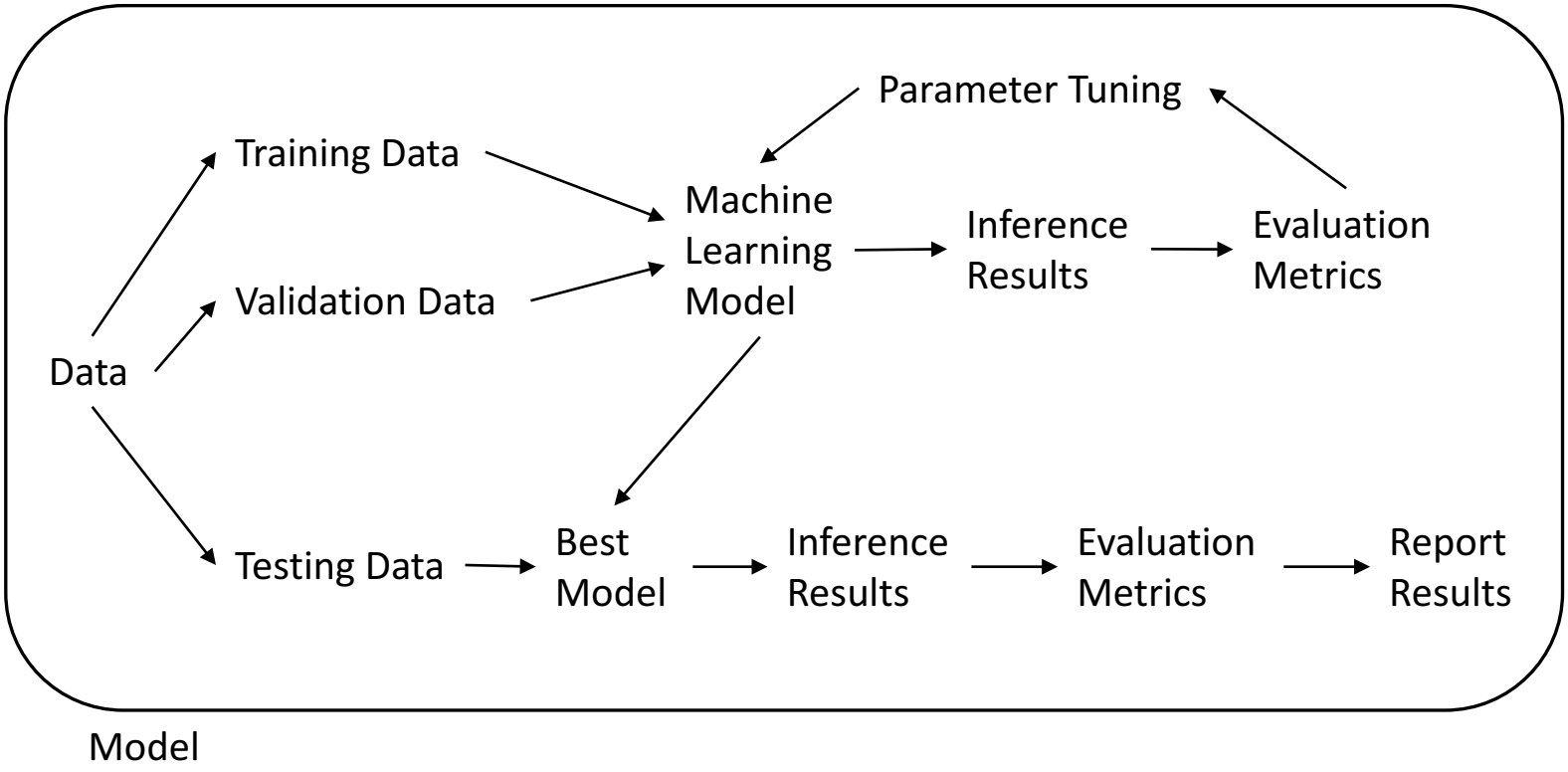


Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



Screenshot from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Machine Learning Lifecycle



Machine Learning Lifecycle

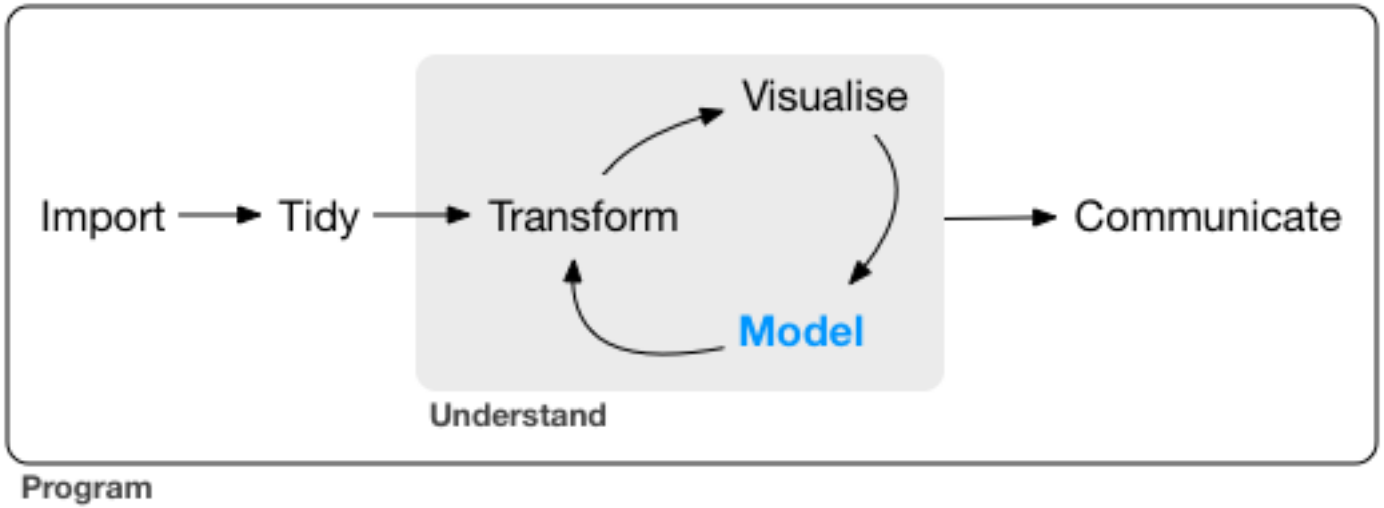


Image from http://r4ds.had.co.nz/model-intro.html

Machine Learning – Frontiers

Deep learning

- Extensions of Artificial Neural Networks
- Hardware has enabled many layers of neurons
- Solving tasks that were previously too difficult to solve

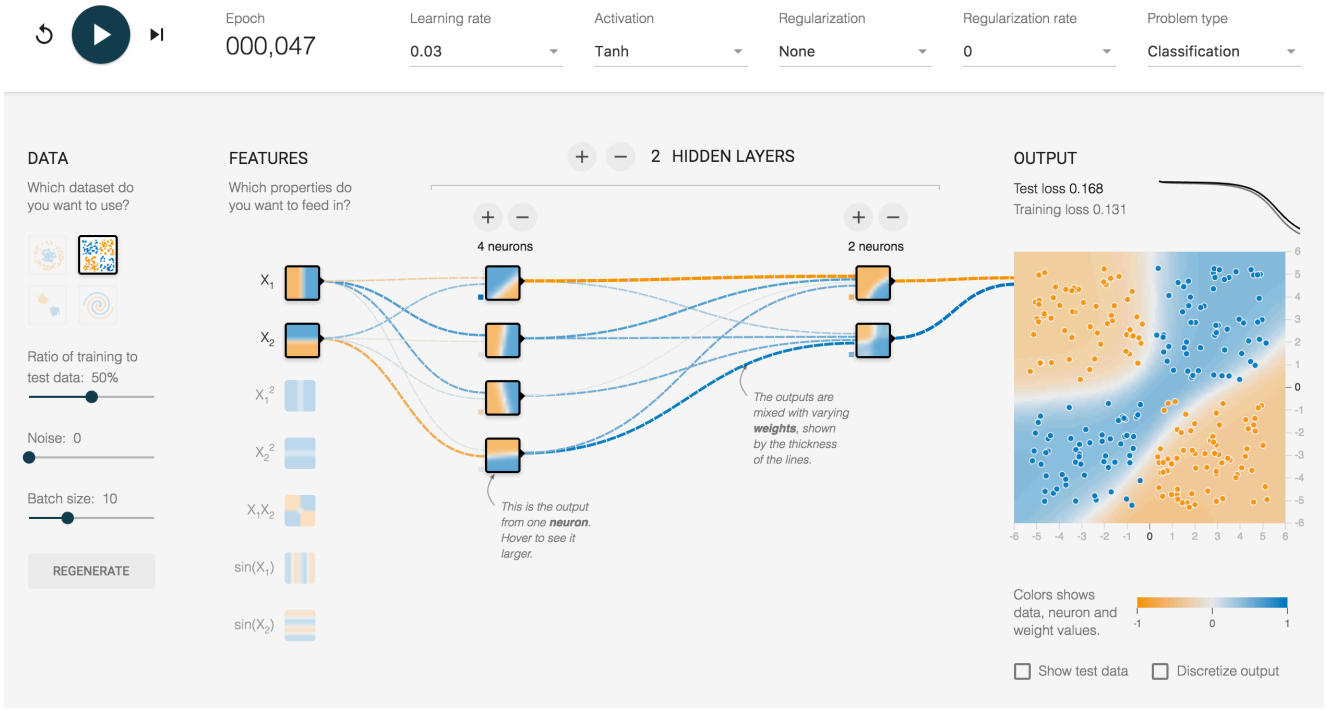
Intelligence on demand

- Cloud based services for machine algorithms
- IBM Watson Services on the IBM Bluemix Cloud
- Bring your data, access state-of-the-art machine learning models

Deep Learning

Visualisations

- http://playground.tensorflow.org/
- http://distill.pub/



Screenshot from http://playground.tensorflow.org/

Machine Learning at IBM

Intelligence on demand – IBM Watson Services

https://www.ibm.com/watson/developercloud/services-catalog.html

Language APIs
- Conversation
- Translator
- Natural Language Classifier
- Natural Language Understanding
- Personality Insights
- Tone Analyzer

Speech
- Speech to Text
- Text to Speech

Vision
- Visual Recognition

Example - Visual Recognition

JSON

Classes	Score
Pomeranian	0.89
dog	0.90
domestic animal	0.90
animal	0.90
reddish orange color	0.71

JSON

Classes	Score
Pho	0.70
soup	0.93
dish	0.93
nutrition	0.93
food	0.93

JSON

Food	Score
beef broth	0.60
broth	0.71
soup	0.80
beef noodle soup	0.57
noodles	0.57

Screenshot from https://visual-recognition-demo.mybluemix.net/

Machine Learning – Australia

Melbourne Machine Learning and AI MeetUp group

- https://www.meetup.com/Machine-Learning-AI-Meetup/

Companies are realising the potential for automating their business processes

Multiplying the productivity power of their employees with machines

Strong research in academic and commercial institutions

- RMIT has a unique strength: applied research and engagement with industry

Machine Learning – Vietnam

I need your help.

What is happening here? What companies are using ML?

Connect and tell me what you know!

General observations

- Start ups into AI
- Foreign companies looking for AI talent
- Expats?
- Local talent?

Where do you start?

Programming Languages

- Python

Resources

- Data Sets: https://www.kaggle.com
- Machine Learning on the cloud
  - IBM Watson on the Bluemix cloud: http://bluemix.net (free 30 day trial)
  - 12 month trial for academic email addresses: https://ibm.onthehub.com
- Introduction to Machine Learning course (taught by Andrew Ng)
  - https://www.coursera.org/learn/machine-learning
- Introduction to Deep Learning (taught by Geoff Hinton)
  - https://www.coursera.org/learn/neural-networks

Machine Learning Questions?

kndtran.com

The Future

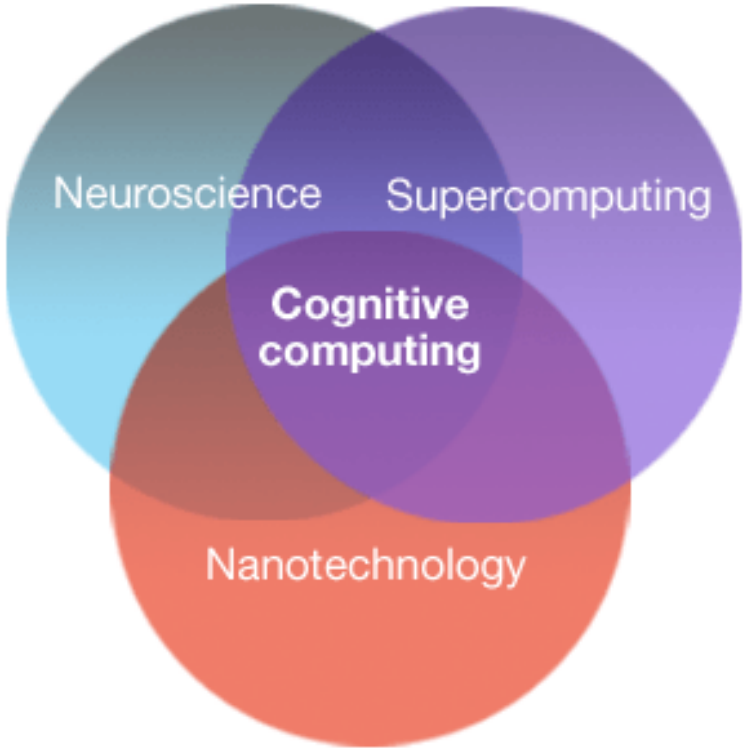
Cognitive Computing

Goals

- Simulate human thought processes in a computerised model
- Mimic the way the human brain works
- Handle massive amounts of unstructured data

Convergence of

- Data Science
- Machine Learning
- Artificial Intelligence
- Neuroscience
- Supercomputing
- Nanotechnology
- (possibly) Quantum Computing
- and many other science areas



Data from http://www.duperrin.com/english/2014/05/27/whats-cognitive-computing-ibm-watson-example/

Vietnam – Connectedness

People connected to the Internet

Country	% Population
Vietnam	52%
Australia	92%
China	52%
Japan	94%
South Korea	89%
Singapore	81%
Hong Kong	82%
Taiwan	88%

Data from http://www.internetworldstats.com/

Vietnam – Opportunities

Increase trade in data

- With neighbours and global markets
- Increase complexity of data exports
- Collect and share more data

Hub for creating automation

- Increase export of custom ML and AI algorithms
- Application of ML to business processes
- Teaching AIs about human tasks
- AI creation pipeline

Data from http://www.internetworldstats.com/

Final Questions?

Summary

Data Science

- Current state and the frontiers
- IBM, Australia, Vietnam
- Where do you start?

Machine Learning

- Current state and the frontiers
- IBM, Australia, Vietnam
- Where do you start?

The Future: Cognitive Computing

Thank you!

kndtran.com
kndtran@gmail.com