11.S965

Tuesdays
1:30–3:30

Building 5
Room 233

DATA SCIENCE *and* MACHINE LEARNING *for* REAL E$TATE

MIT
Real Estate
Innovation
Lab

Dr. Andrea Chegut
Instructor
Yair Titelboim
Teaching Assistant

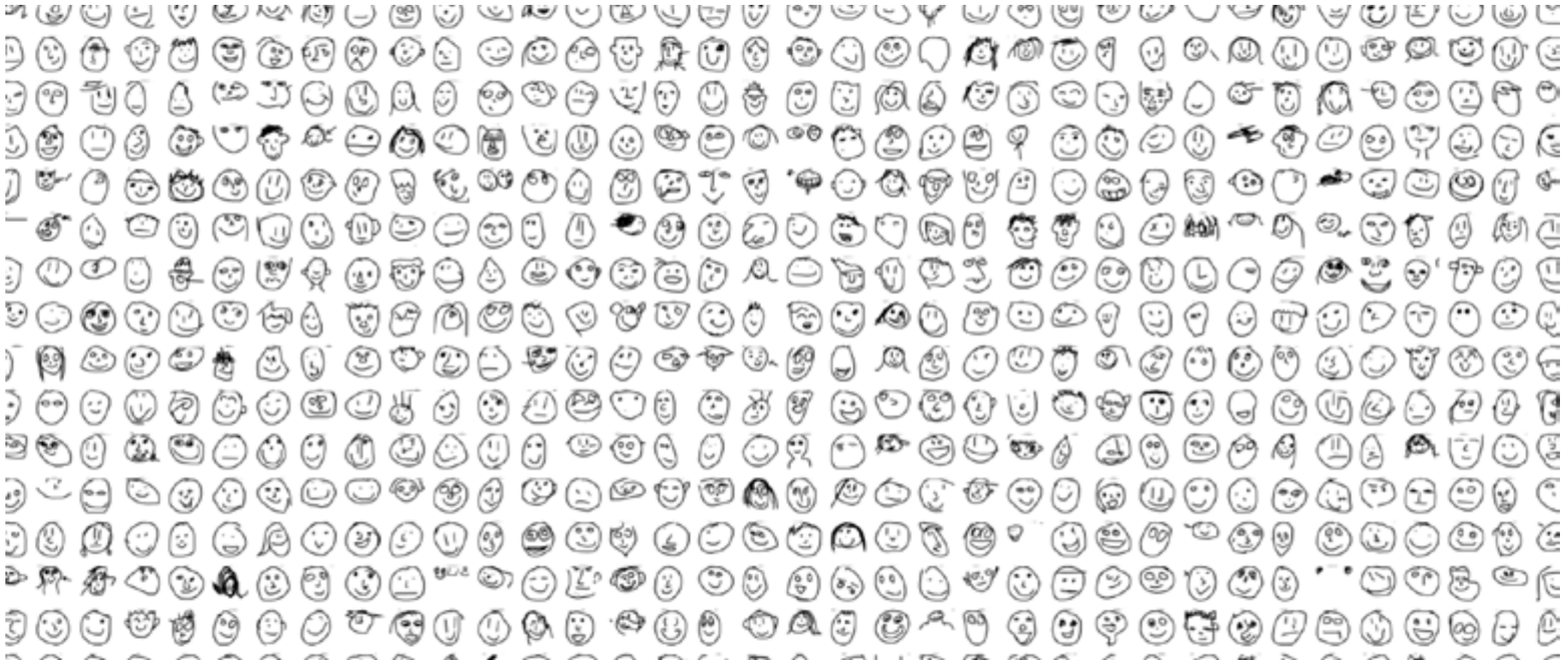
Core skills for development,design and planning are shifting to encompass analytics in data science and machine learning. This seven week mini-course aims to introduce you to the principles of data science and machine learning that are impacting the domain of real estate today. In the course, we will hear from data scientists across technology companies, learn core data science in R, and produce predictive analytics using machine learning techniques. The class is intended for students with some knowledge of data science, but are seeking to learn more. Core knowledge of R is welcome

a big why

Your Best Teammate Might Someday Be an Algorithm...

A new program from Google seeks ways for AI systems to work more effectively with humans.

by Will Knight

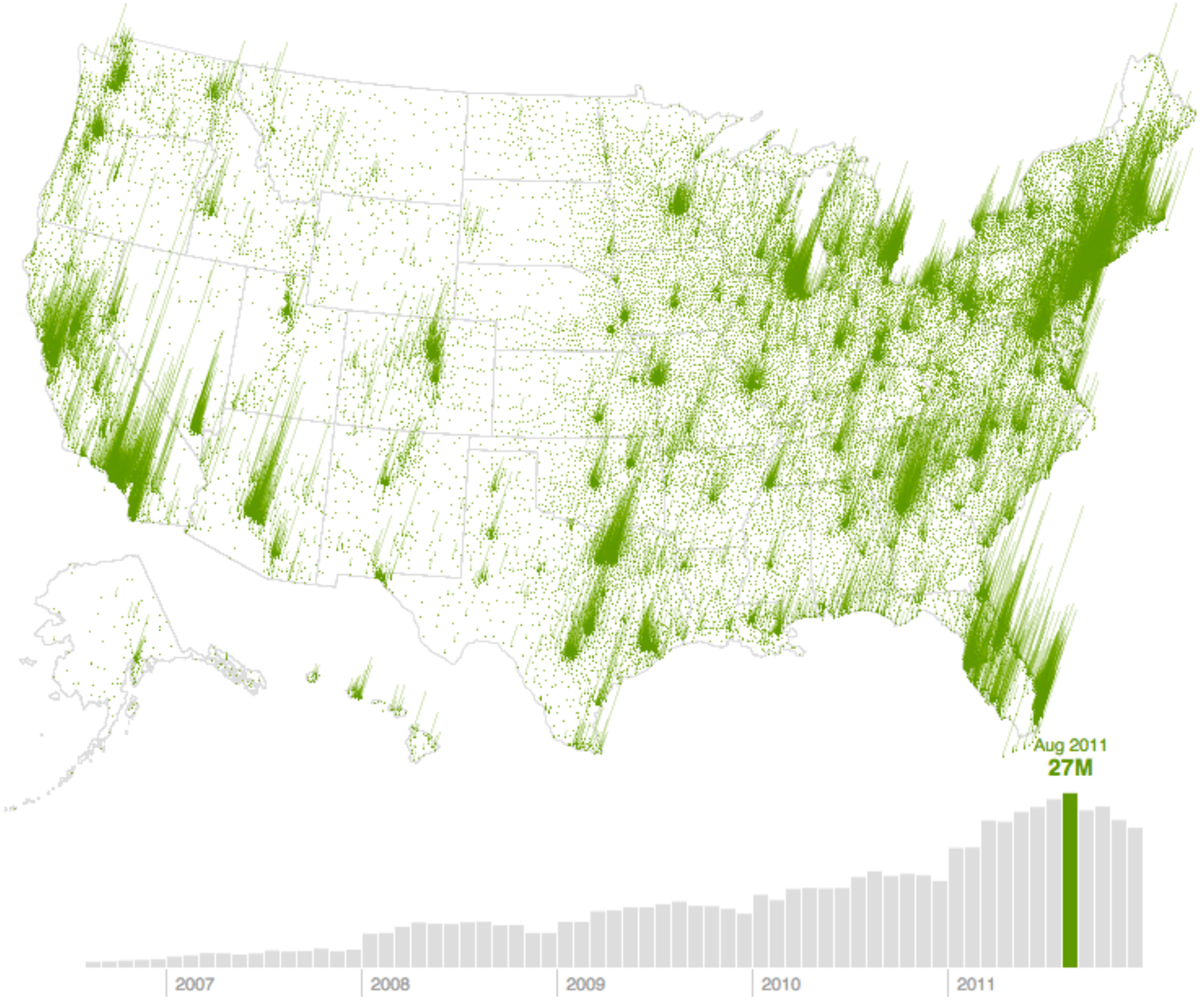


A snapshot of the Facets Dive x Quick, Draw! tool, one of the new interactive interfaces designed by Google PAIR.

Source: MIT Technology Review, July 2017 by Will Knight

Where are people looking for homes

In August 2006, real estate search site Trulia had 609,000 visitors. Five years later, there were 27 million. Trulia's most recent visualization shows this growth (bottom bar graph) and where people are searching for homes (map). Press play and watch it go. It's pretty much population density, but for me, the method is more interesting than the material in this case.



Source: Trulia via @shashashasha, Flowing Data, https://flowingdata.com/2012/01/06/where-people-are-looking-for-homes/

a shared why

A MAP OF EVERY BUILDING IN AMERICA

In some cases, the building shapes generated by Microsoft's automated process do not match the existing building footprints exactly. We manually corrected as many of these mistakes as we found, or, where available, replaced the shapes using more precise local data sets. Data was unavailable for much of Alaska.

Source: https://www.nytimes.com/interactive/2018/10/12/us/map-of-every-building-in-the-united-states.html

our why

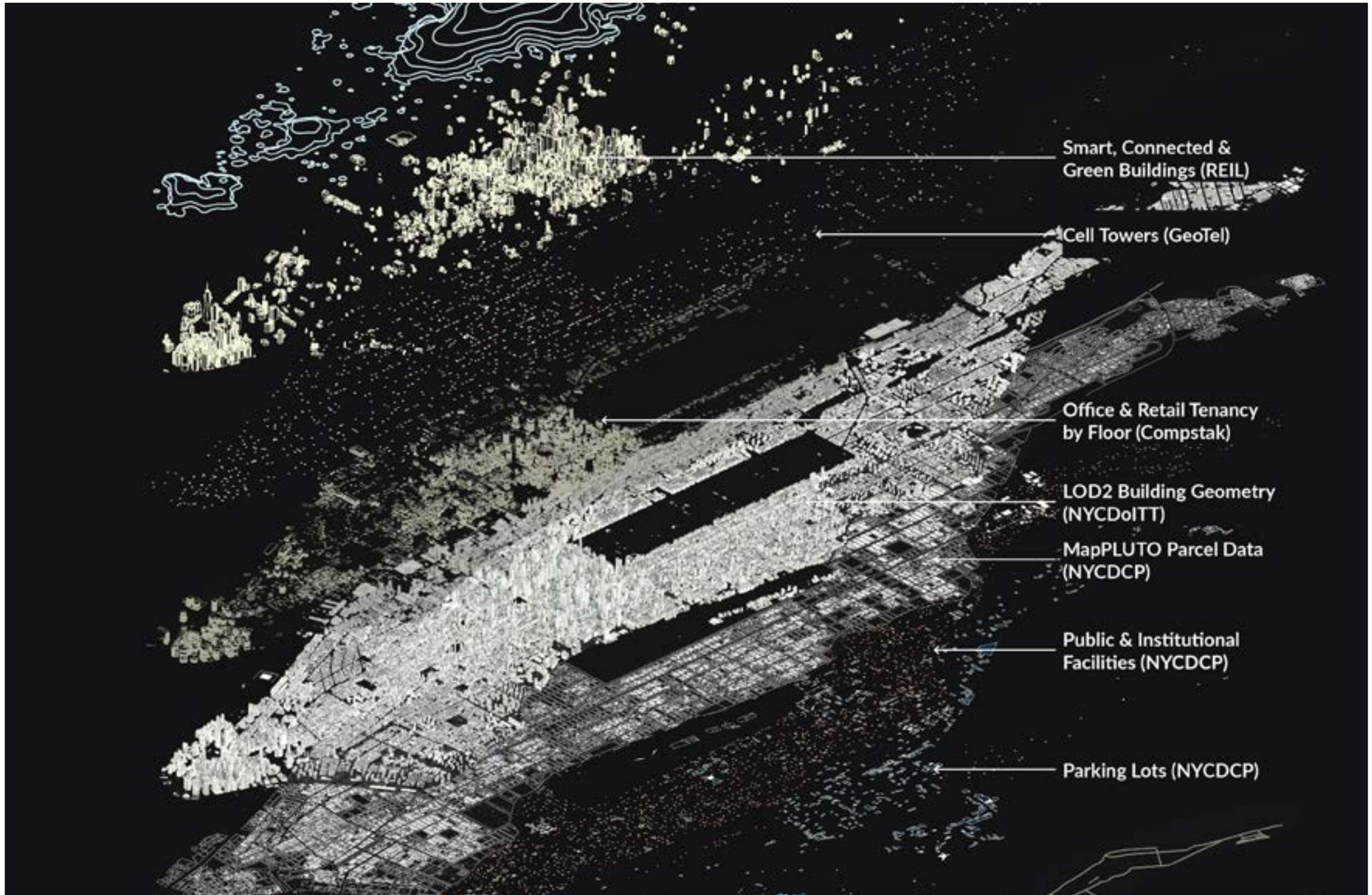
Wide Data - A Geometric, Geospatial Relational Database of NYC

The MIT Real Estate Innovation Lab is working with public and private data providers to create a wide data approach for linking design and innovation to financial performance in the built environment.

In this R&D project we are exploring the data science that connect design to the capital stack. This means combining geometry, to geospatial and relational database structures to create insights about the value of innovation in the built environment.

The data spans over 15 years with over 3,000 variables across 200 datasets and 18 data providers. Our lab approaches financial performance and economic growth questions from an interdisciplinary analysis approach, where design and planning metrics carry just as much weight as financial and economic performance.

Source: MIT Real Estate Innovation Lab



our why

Real Estate Analytics for 21st Century Cities

The real estate asset class is deeply physical and also contextual. Understanding supply, demand and pricing characteristics relies on understanding the physical nature of the building, its relationship to other characteristics of the city and the abstract elements of supply and demand of its stakeholders.

Data has become more relevant than ever to deconstruct what we can and cannot measure.

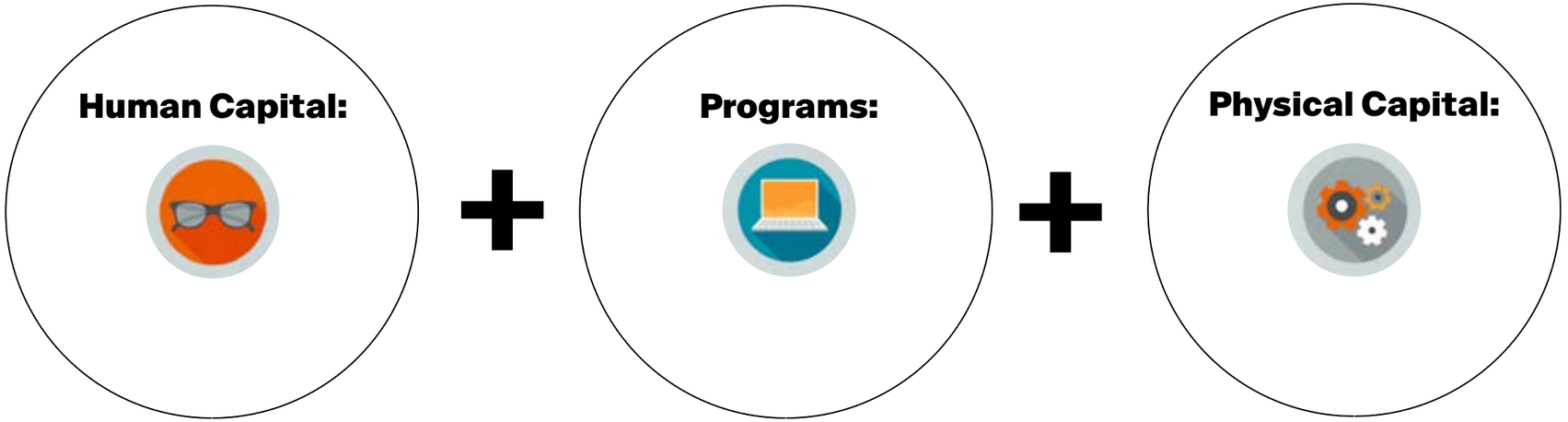
Source: MIT Real Estate Innovation Lab

so we are going to need to develop some skills

Data Science and Machine Learning for Real Estate

Skills, technologies and physical capital are even more diverse than in the traditional data science and machine learning domain.

Real Estate is a physical asset, made up of a complex bundle of abstract processes and physical attributes and brought together by a random, semi-permant group of actors.



	Human Capital:	Programs:	Physical Capital:
In general:	• Data Scientists • Statisticians • Econometricians	• Relational Database • Statistical Program (R, Python, Stata)	• Cloud Computing • Local Workstations
For 21st century real estate:	• Computational Architects • Geospatial Scientists • Urban Econometricians	• Geospatial & Geometry Processing • Custom CAD-GIS interoperability • Post GIS • Web-mapping FrontEnd	• Advanced Graphics Cards • Large Data Storage Capacity • Larger Computing RAM

a little expertise and practice to get the skills flowing...

This class

Date	Lecture	Exercise	Readings	Logistics
April 02	**Introductory Lecture - Data Science and Machine Learning Principles** - by `Dr. Andrea Chegut`	**Data Science ToolKit part one:** getting started with R Studio (TA: Yair Titelboim)	`Read` Data Science by John D. Kelleher and Brendan Tierney (a part of the MIT Press Essential Knowledge Series) - Chapters 1-3	- `Form` Groups `Install` R
April 09	**Industry Insights from the CRE Data Buffett** - by `Steve Weikal` Head of Industry Relations and REI Lab CRE Tech Lead, Center for Real Estate	**Data Science ToolKit part two:** intro to the `tidyverse` (TA: Yair Titelboim).	`Read` Data Science by John D. Kelleher and Brendan Tierney (a part of the MIT Press Essential Knowledge Series) - Chapter 5	- `Practice` Data Science Tools in R using TidyVerse.
April 16	**Data Science from a Source** - by Vice President of Data Science, CompStak, `Wayne Yu`	**Clustering and Anomaly Detection:** detecting with `tidyverse, cluster` and and `clusplot` (TA: Yair Titelboim and Dr. Andrea Chegut).	`Focus` on Assignment	- `Submit` TidyVerse graphical outcomes and explanations of outcomes by April 23, 2019.
April 23	**Machine Learning vs. Econometric Tools** - by `Mossimo Young`, Director, Head of Data Solutions, Investment Management BNY Mellon	**Machine Learning Toolkit (Price Prediction) part one:** prototyping with `tidyverse` and `factoextra` (TA: Yair Titelboim)	`Read` Machine Learning by Ethem Alpaydin (a part of the MIT Press Essential Knowledge Series) - Chapters 1 and 2	- `Submit` Clustering and Anomaly Detection exercise and explanations by April 30, 2019.
April 30	**Machine Learning Applications** - by `Dr. Alex van De Minnie`, Head of the MIT Price Dynamics Platform, MIT Center for Real Estate	**Machine Learning Toolkit (Price Prediction) part two:** introducing the `Caret` package (TA: Yair Titelboim)	`Read` Machine Learning by Ethem Alpaydin (a part of the MIT Press Essential Knowledge Series) - Chapters 3 and 4	- `Practice` Machine Learning Price Predictions.
May 07	**Machine Learning Applications** - by `John Poulin`, SVP of Technology, Real Capital Analytics	**ML Algos part two:** Implementing predictive performance, `Neural Network` Models in R	`Focus` on Assignment	- `Submit` Machine Learning Predictive Performance Exercise by May 14, 2019.
May 14	**The Ethics and Responsibility of Data Science and Machine Learning for Real Estate** - by `Dr. Andrea Chegut`	**Presentations(discussion) and summary**	`Focus` on Assignment	- `Submit` One page predictive summary utilizing data to tell a predictive story about commercial real estate by May 25, 2019.

data science and machine learning tour guides

your instructors

INSTRUCTOR:
Dr. Andrea Chegut

Research Scientist
Director of the MIT Real Estate Innovation Lab
Head of Research DesignX
Research Coordinator Center for Real Estate

Financial Economist by training, practicing asset valuation models for innovation in real estate

TEACHING ASSISTANT
Yair Titelboim

Computational Planner
Lead Researcher MIT Technology Tracker - Live Cataloging Division

Architect by training, practicing data scientist for planning, design and asset valuation...oh and scrapping for measuring technological change

MUSE:
Greg

Some books for core knowledge

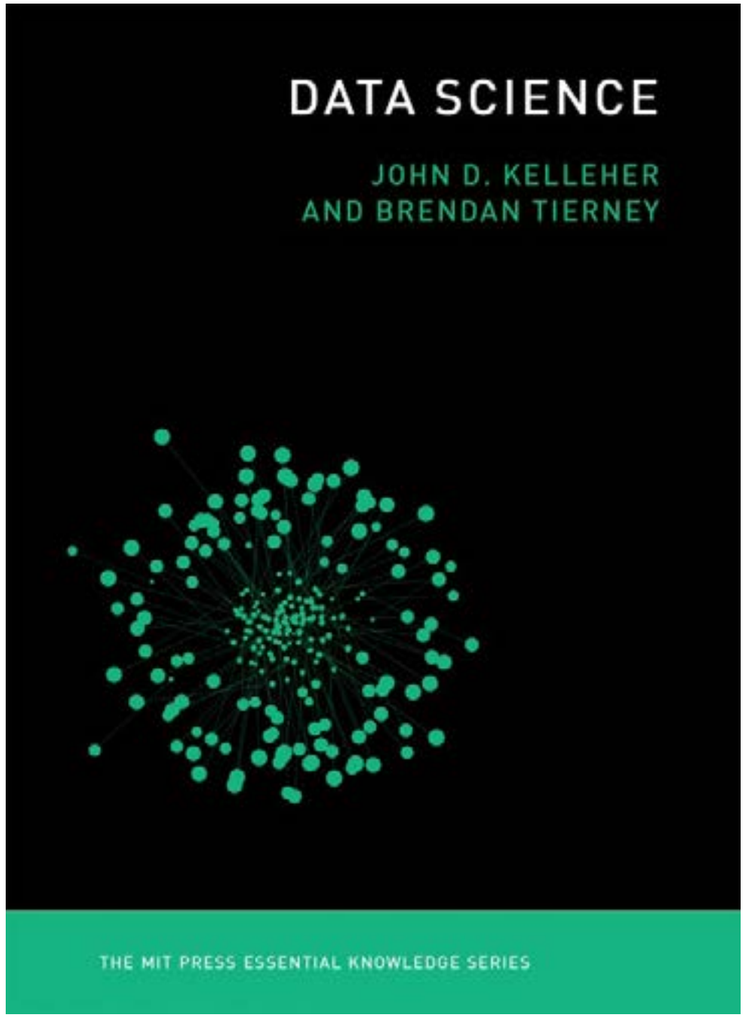
Buy your books here:

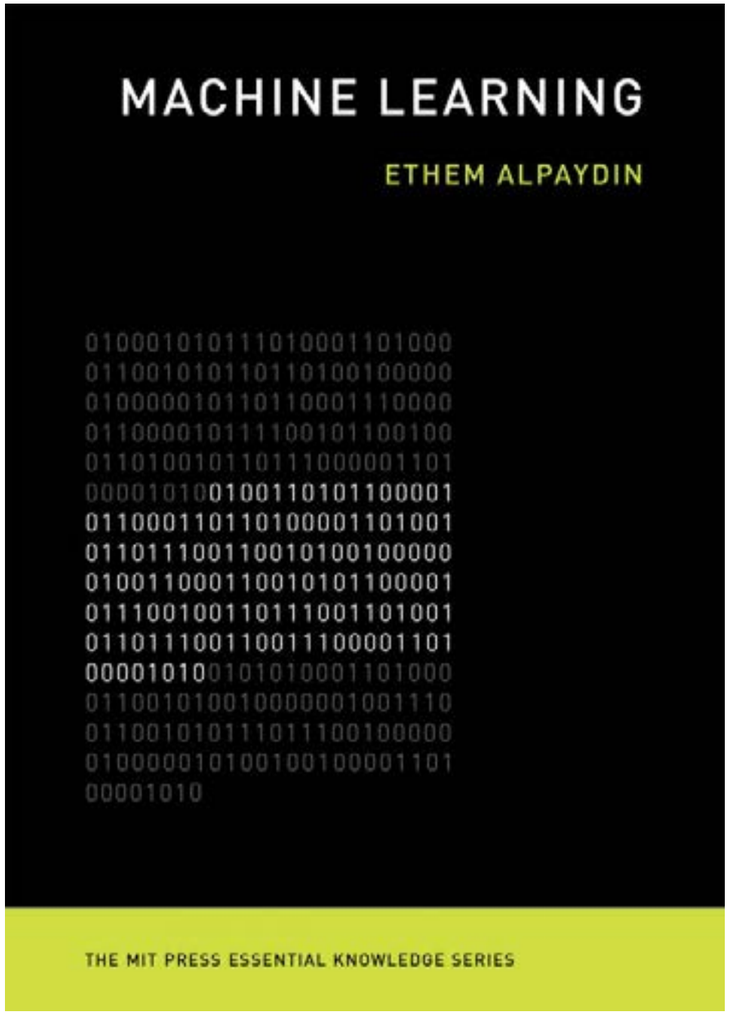
https://mitpress.mit.edu/books/data-science

https://mitpress.mit.edu/books/machine-learning

Optional readings:

Please see the list of optional readings on the Stellar site. Academic and industry papers highlighting the developments of machine learning applications in real estate.





Data Scientists, Machine Learning Experts, Technologists and Econometricians

Steve Weikal

Head of Industry Relations and REI Lab CRE Tech Lead, Center for Real Estate

Dr. Alex van De Minnie

Head of the MIT Price Dynamics Platform, MIT Center for Real Estate

Wayne Yu

Vice President of Data Science, CompStak

Mossimo Young

Director, Head of Data Solutions, Investment Management BNY Mellon

John Poulin

SVP of Technology, Real Capital Analytics

Dr. Schery Bokhari

Head of Research, RedFin

Data Science Programming Skills

Core skills require knowledge of relational databases and datasets.

However, at minimium knowledge of Python or R are necessary for moving foward with interactive components.

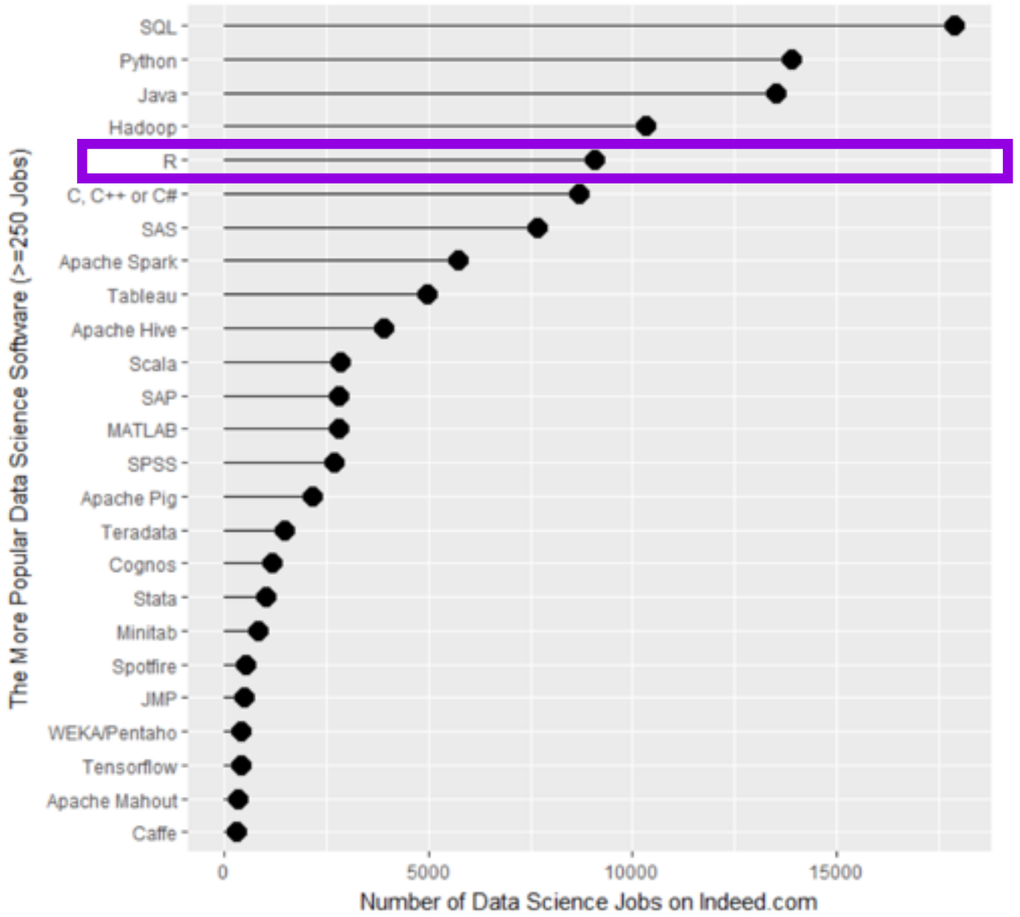


Figure 1a. The number of data science jobs for the more popular software (those with 250 jobs or more, 2/2017).

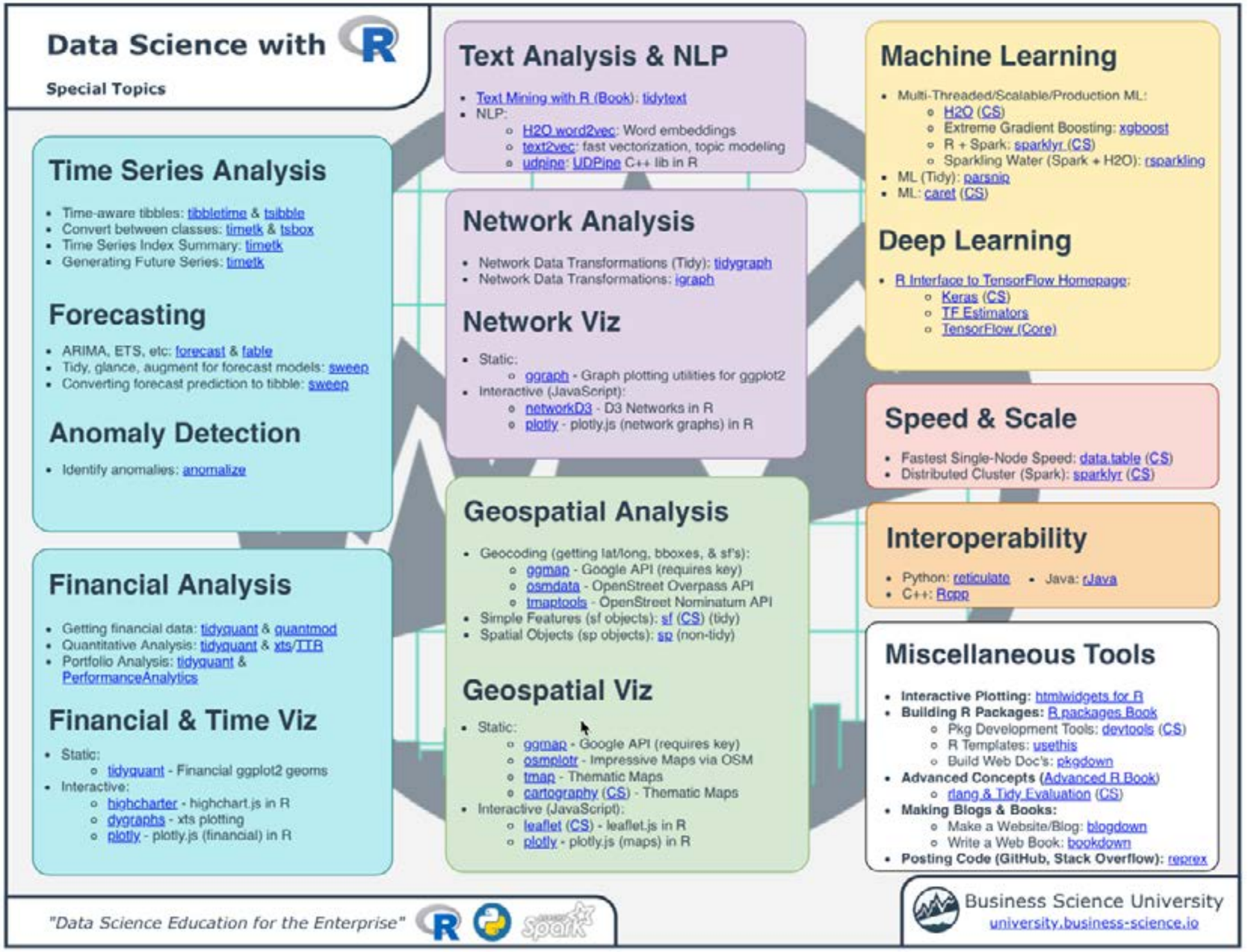
Source: Indeed.com

For this class, we are going to deploy R.

It is open source, with loads of connectivity between statisticans, geographers, economists and advanceing data science and machine learning techniques.

The R Tool Kit

You will learn the tidy skills...



and some guided nudges to get the job done...

You need to submit assignments

Assignments are due one week after lesson.

TO BE EXECUTED IN TEAMS!!!!

(share the work don't divide and conquer, talk through it togeth-er)

April 23, 2019
April 30, 2019
May 14, 2019
May 25, 2019

Date	Exercise	Readings	Logistics
April 02	**Data Science ToolKit part one:** getting started with R Studio (TA: Yair Titelboim)	`Read` Data Science by John D. Kelleher and Brendan Tierney (a part of the MIT Press Essential Knowledge Series) - Chapters 1-3	- `Form` Groups `Install` R
April 09	**Data Science ToolKit part two:** intro to the `tidyverse` (TA: Yair Titelboim).	`Read` Data Science by John D. Kelleher and Brendan Tierney (a part of the MIT Press Essential Knowledge Series) - Chapter 5	- `Practice` Data Science Tools in R using TidyVerse.
April 16	**Clustering and Anomaly Detection:** detecting with `tidyverse, cluster` and and `clusplot` (TA: Yair Titelboim and Dr. Andrea Chegut).	`Focus` on Assignment	- `Submit` TidyVerse graphical outcomes and explanations of outcomes by April 23, 2019.
April 23	**Machine Learning Toolkit (Price Prediction) part one:** prototyping with `tidyverse` and `factoextra` (TA: Yair Titelboim)	`Read` Machine Learning by Ethem Alpaydin (a part of the MIT Press Essential Knowledge Series) - Chapters 1 and 2	- `Submit` Clustering and Anomaly Detection exercise and explanations by April 30, 2019.
April 30	**Machine Learning Toolkit (Price Prediction) part two:** introducing the `Caret` package (TA: Yair Titelboim)	`Read` Machine Learning by Ethem Alpaydin (a part of the MIT Press Essential Knowledge Series) - Chapters 3 and 4	- `Practice` Machine Learning Price Predictions.
May 07	**ML Algos part two:** Implementing predictive performance, `Neural Network` Models in R	`Focus` on Assignment	- `Submit` Machine Learning Predictive Performance Exercise by May 14, 2019.
May 14	**Presentations(discussion) and summary**	`Focus` on Assignment	- `Submit` One page predictive summary utilizing data to tell a predictive story about commercial real estate by May 25, 2019.

A SHARED CONCEPTUAL FRAMEWORK

...extracting non-obvious and useful patterns from large data sets.

-Kelleher and Tierney

Data Science is used for just that...

Data science encompasses a set of principles, problem definitions, algorithms, and processes...but also takes up other challenges capturing, cleaning and transforming...data.

-Kelleher and Tierney

But what are data?
And what is a dataset?

a datum or piece of information is an abstraction of a real-world entity (person, place, object, event, emotion, values, etc.)...establish data attributes to form a dataset.

-Kelleher and Tierney

data attributes

Data classification is a fundamental skill. Learning how to classify data or to work with categorical and numerical data for analytics and display will drive the whole of your data science experience.

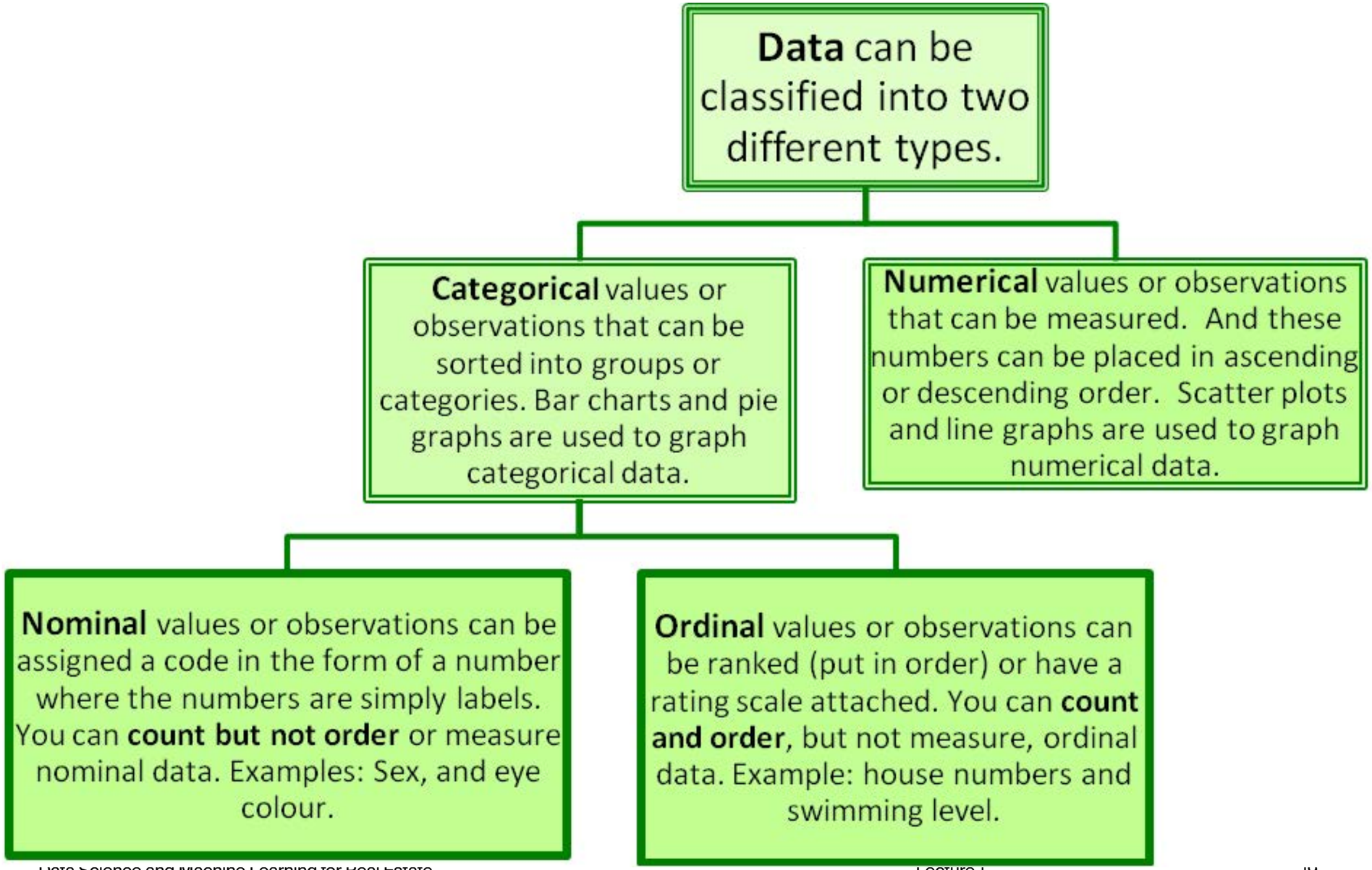
Numerical data is ideal, but it does not always exist to represent the real data generating process.

However, categorical variables are the most easy to develop.

IMPORTANT

The data type of an attribute of an attribute (numeric, ordinal, nominal) affects the methods we can use to analyze and understand the data.

-Kellher and Tierney, pg. 44



Data Science is really about the wrangling, cleaning and tidying of data...

Garbage in...garbage out.

-David Geltner

IMPORTANT

Two characteristics of data science cannot be overemphasized:

(a) for data science to be successful we need to pay a great deal of attention to how we create our data (in terms of both the choices we make in designing the data abstractions and the quality of the data captured by our abstraction processes), and

(b) we also need to sense check the results of the data science process that is, we need to understand that just because the computer identifies a pattern in the data this doesn't mean that it is identifying a real insight in the processes we are trying to analyze.

-Kellher and Tierney, pg. 47

There are all sorts of rules about data science and analysis and then the data just goes and breaks them...

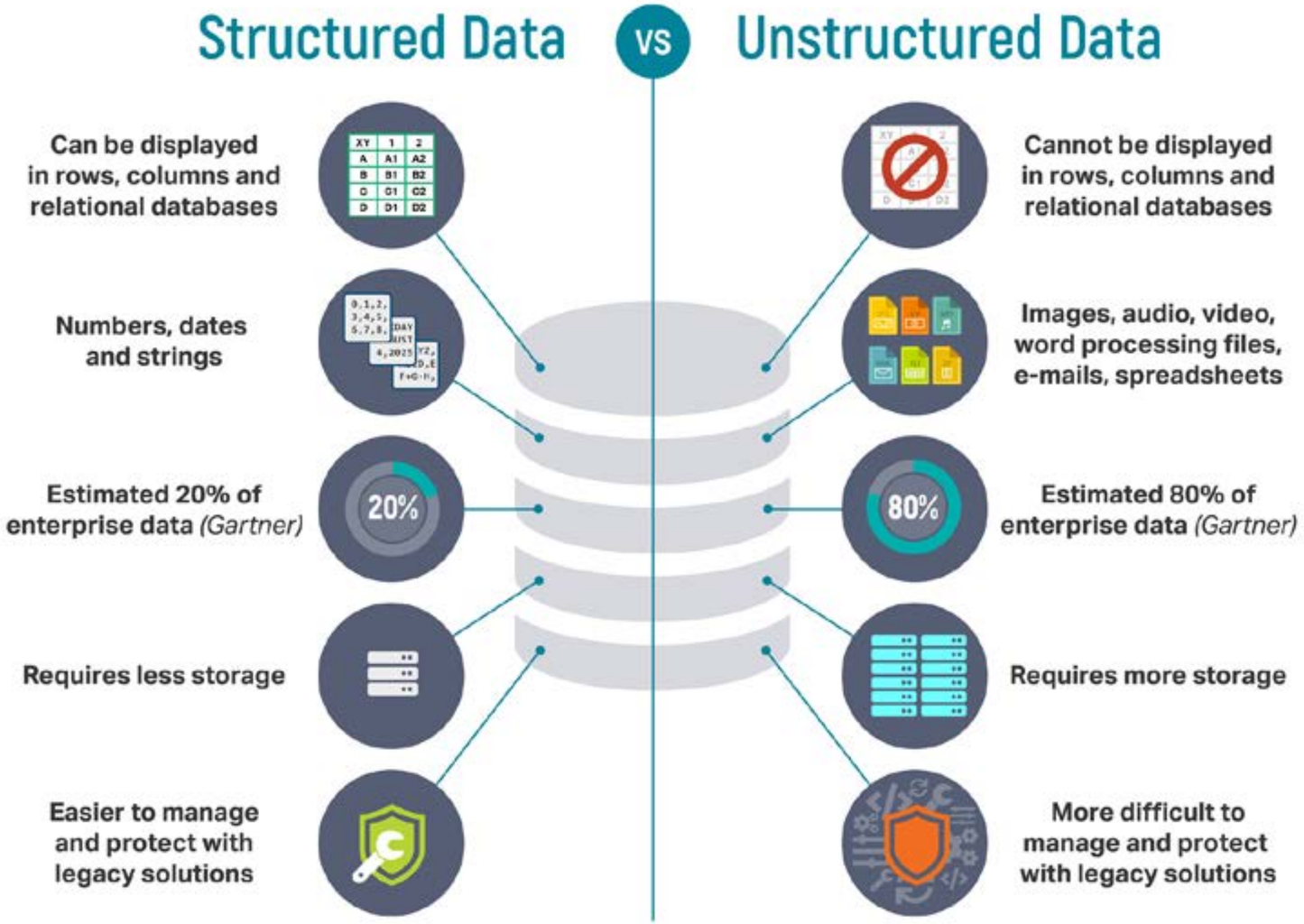
Structured vs. Unstructured Data

IMPORTANT

Structured data are data that can be stored in a table, and every instance in the table has the same structure (i.e., set of attributes)

Unstructured data are data where each instance in the data set may have its own internal structure, and this structure is not necessar-ility the same in every instance. (e.g., webpages)

-Kellher and Tierney, pg. 48



We are doing all of this to identify a dataset that hopefully captures features of the data generating process we either want to EXPLAIN or PREDICT from.

Structured Data Forms a Matrix or Data Table

What is it about this dataset that looks good so far?

Flights data table in R.

```
flights <- fread("flights14.csv")
flights
#         year month day dep_delay arr_delay carrier origin dest air_time distance hour
#      1: 2014     1   1        14        13      AA    JFK  LAX      359     2475    9
#      2: 2014     1   1        -3        13      AA    JFK  LAX      363     2475   11
#      3: 2014     1   1         2         9      AA    JFK  LAX      351     2475   19
#      4: 2014     1   1        -8       -26      AA    LGA  PBI      157     1035    7
#      5: 2014     1   1         2         1      AA    JFK  LAX      350     2475   13
#     ---
# 253312: 2014    10  31         1       -30      UA    LGA  IAH      201     1416   14
# 253313: 2014    10  31        -5       -14      UA    EWR  IAH      189     1400    8
# 253314: 2014    10  31        -8        16      MQ    LGA  RDU       83      431   11
# 253315: 2014    10  31        -4        15      MQ    LGA  DTW       75      502   11
# 253316: 2014    10  31        -5         1      MQ    LGA  SDF      110      659    8
dim(flights)
# [1] 253316     11
```

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Machine learning focuses on the design and evaluation of algo-rithims for extracting patterns from data.

-Kelleher and Tierney

Machine learnig and prediction is possible because the world has regularities.

-Alpaydin

many tools to arrive at categorization and prediction of an experience

Methods of Machine Learning

Source: Machine Learning Mastery

Econometrics vs. Machine Learning

We have often been looking for causaility or trying to approximate it, but now the movement is shifting towards just predict as nothing is really exactly causal....

Important

It is more than likely correlated or systemic, but not exacting.

BUT nor is the prediction. I have an expectation or probability that something will happen.

Source: Breiman (2001)

Econometrics vs. machine learning

	Econometrics	Machine learning
Approach	statistical: data generating process	algorithmic model, DGP unknown
Driver	theory	fitting the data
Focus	hypothesis testing & interpretability	predictive accuracy
Model choice	parameter significance & in-sample goodness of fit	cross-validation of predictive accuracy on partitions of data
Strength	understand causal relationships & behavior	prediction

See Breiman (2001) and Matt Bogard's blog

The distinction cannot be more simple than what the tool is that you need to employ.

Approximate the past OR predict the future

IMPORTANT

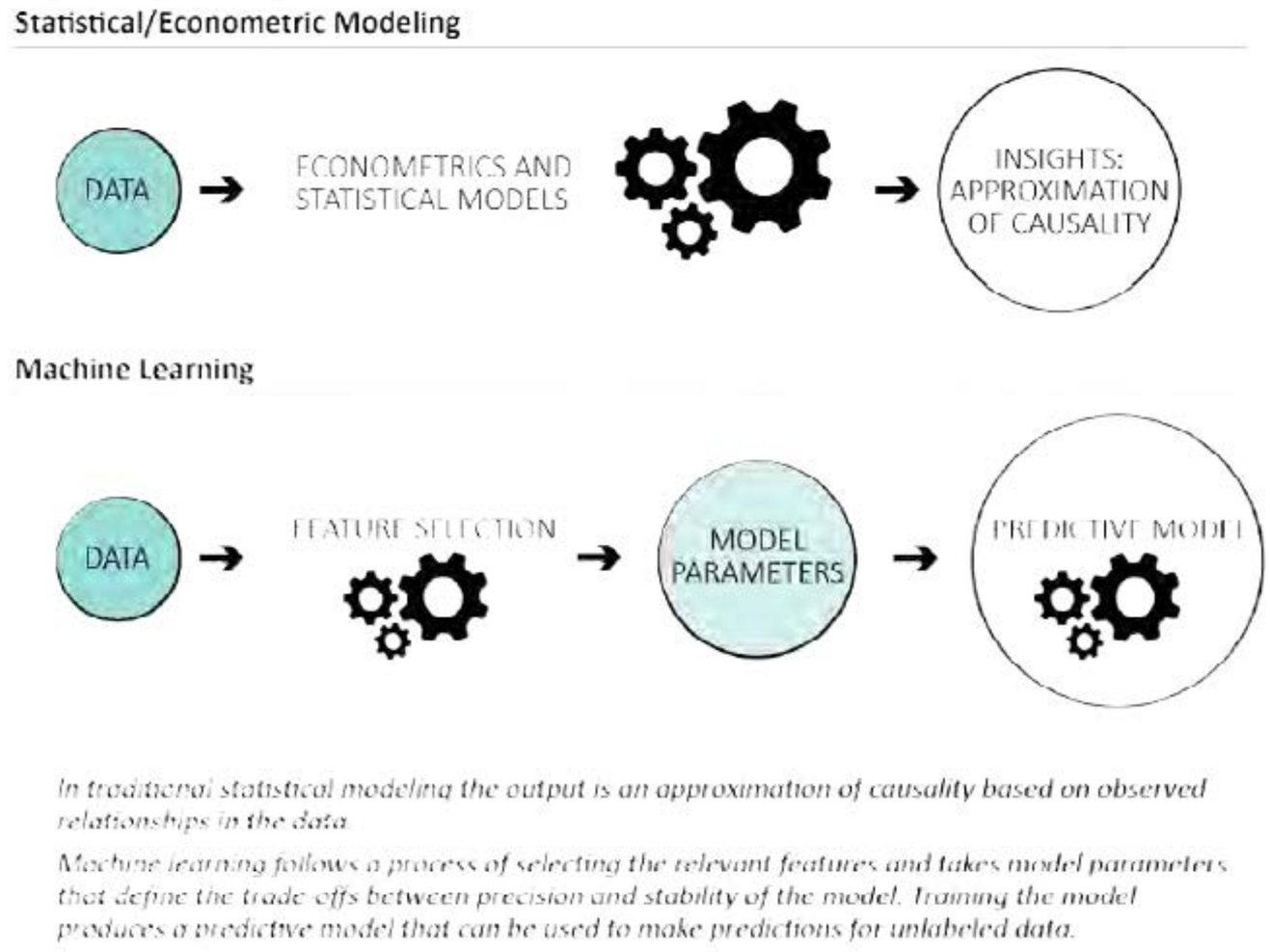
Do you need to create a value proposition, estimate a policy impact or deconstruct what drives, demand, supply or prices...

Econometrics

Do you need to forecast the future, predict the risk and return of asset, or predict a future purchase decision?

Machine Learning

Source: Machine Learning and Artifical Intelligence in Real Estate by Jenny Conway



In traditional statistical modeling the output is an approximation of causality based on observed relationships in the data.

Machine learning follows a process of selecting the relevant features and takes model parameters that define the trade-offs between precision and stability of the model. Training the model produces a predictive model that can be used to make predictions for unlabeled data.

To employ either, we must understand elements of statistics

Randomness and probability

Ex - Ante

Can you say something about the probability of something happening?

Probabilistic Outcome

e.g., coin toss

Can you not say something about the probability?

Estimation

e.g., using data to either estimate the predicted outcome or explain previous outcomes

We expect consumers in general to follow certain patterns in their decisions, depending on factors such as the composition of their household, their tastes, their income and so on. Still there are always additional random factors that introduce variance: vacation, change in weather, advertising, etc.
- Alpaydin, pg. 34

In statistics, we call these Omitted Variable Biases OR Endogeneity

To employ either, we must understand elements of statistics

Supervised Learning

Ex - Ante

Can you say something about the probability of something happening?

Probabilistic Outcome

e.g., coin toss

Can you not say something about the probability?

Estimation

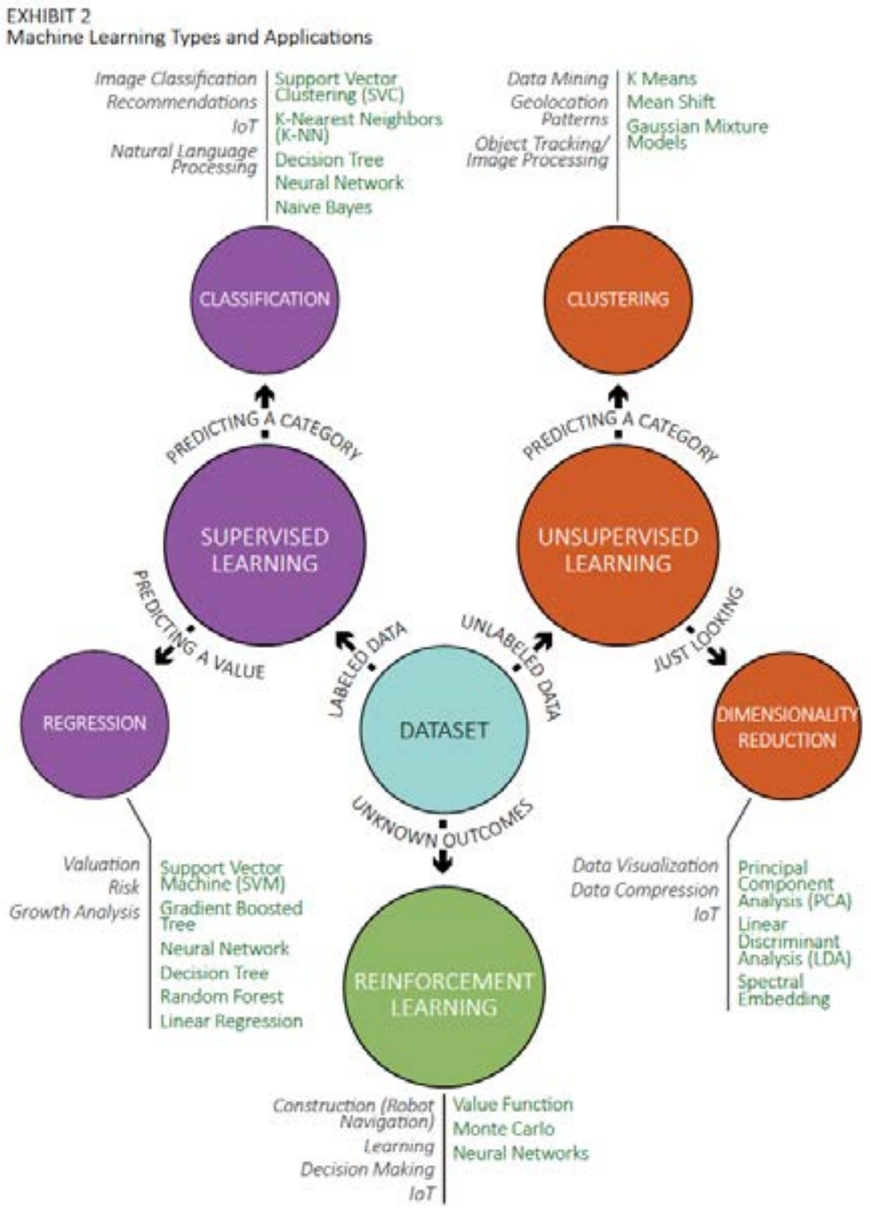
e.g., using data to either estimate the predicted outcome or explain previous outcomes

The task of estimating an output value from a set of input values is called a regression in statistics...

and in machine learning a regression is one type of supervised learning.

Machine learning has grown over the last 50 years into an explosive paradigm shift in how we form expectations...

Methods of Machine Learning by Learning Type



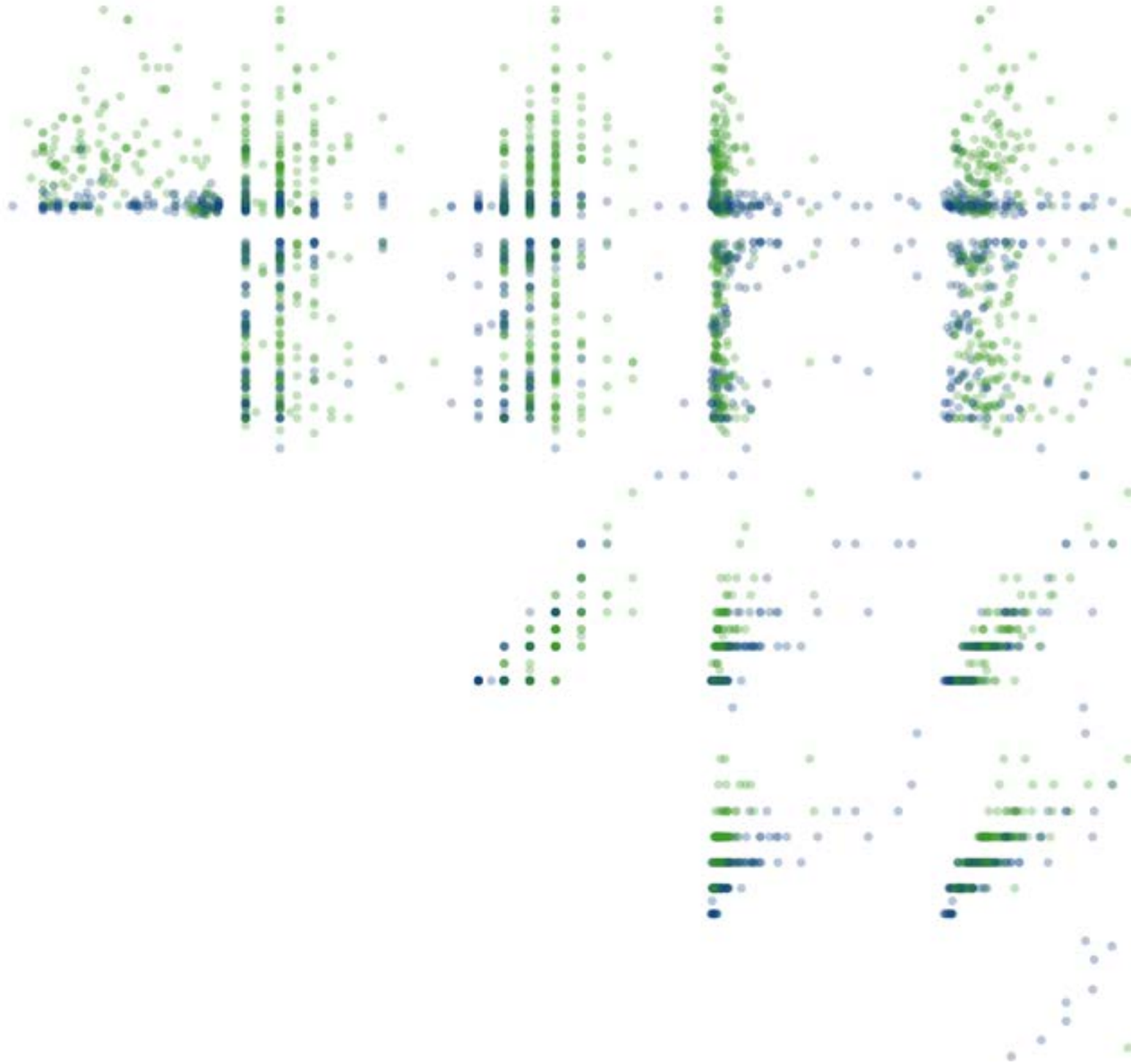
Source: Machine Learning and Artif-ical Intelligence in Real Estate by Jenny Conway

A visual introduction to machine learning

English

In machine learning, computers apply **statistical learning** techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions.

Keep scrolling. Using a data set about homes, we will create a machine learning model to distinguish homes in New York from homes in San Francisco.



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

Now to get prepared...

Install R and Form Groups with Yair