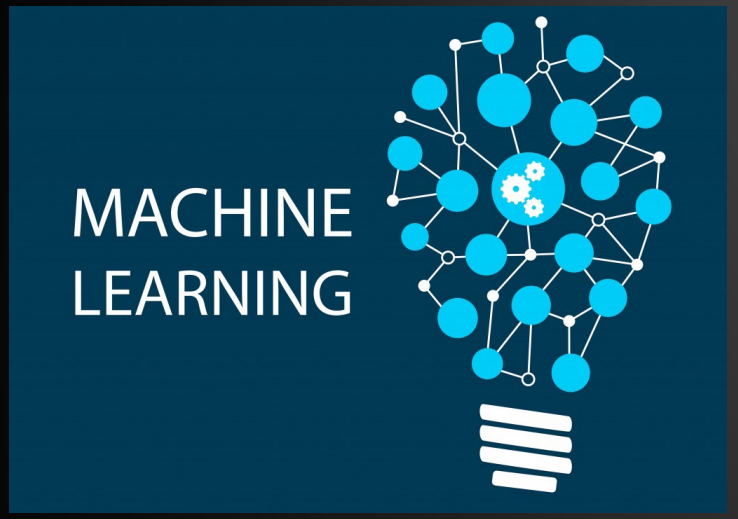
INTRODUCTION TO



TUE VU (PHD)

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CCIT\CITI

TAKE YOUR RESEARCH TO THE NEXT LEVEL

With the **FREE** high performance computing resources and training available at Clemson, you can take your work to the next level. CCIT's Cyberinfrastructure Technology Integration group will show you how.

Get in-person training and regular workshops

Learn to use Palmetto, one of the world's top supercomputers

Don't know much about programming? We'll teach you!

RESEARCH COMPUTING CAN HELP YOUR WORK IN:

Data Science | Big Data | Simulation Modeling | Visualization

CLEMSON COMPUTING AND INFORMATION TECHNOLOGY

Get started

ITHELP@clemson.edu

citi.sites.clemson.edu

CITI Cyberinfrastructure Technology Integration

Training Workshops

- Intro to Linux
- Intro to Research Computing
- Intro to Hadoop on the Cypress Cluster
- Intro to Programming with Python
- Intro to Big Data Analytics (Python)
- Intro to Data Science using R
- Machine Learning using R
- Machine Learning using Python
- GIS Training
- Certificate

CLEMSON

Cyber-Infrastructure Technology Integration

ADVANCED COMPUTING & DATA SCIENCE

CERTIFICATE OF ATTENDANCE

This acknowledges that

Joseph Ligato

has completed 15-hour workshop in
Data Science & Machine Learning with R programming language
(July 2019)

Tue Vu, PhD
Workshop Facilitator

Dustin Atkins
Executive Director CITI

OUTLINES

1. Introduction to Machine Learning
2. Why R
3. Types of Machine Learning
4. Applications DS&ML to Psychology
5. Discussions
6. Hands-on sessions

1. Introduction to Machine Learning

What is Machine Learning?

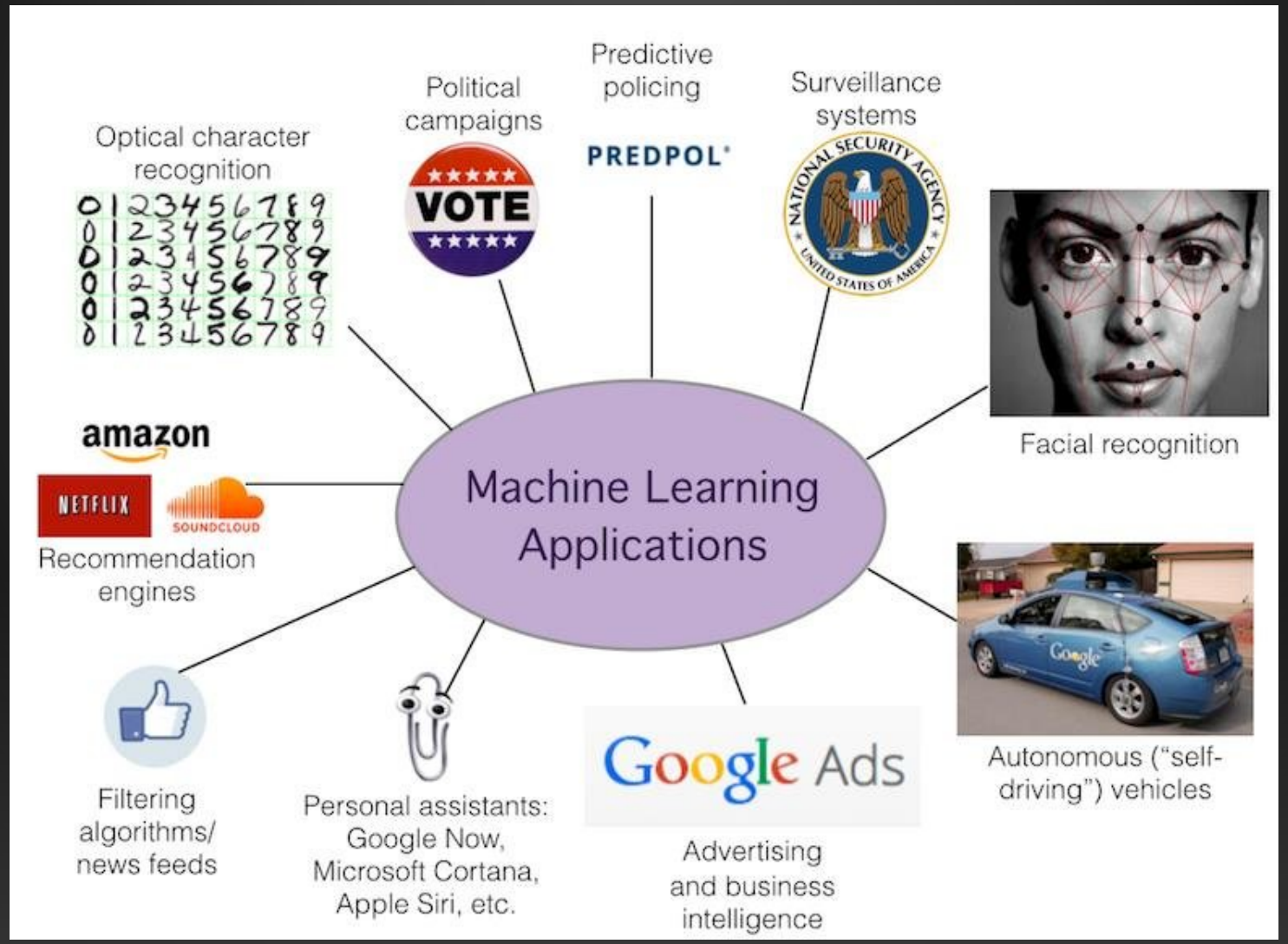
Arthur Samuel: Stanford

Field of study that gives computers the ability to learn without being explicitly programmed

Tom Mitchel: CMU

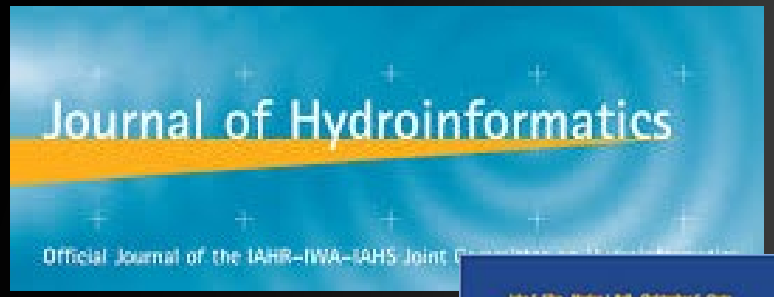
The field of Machine Learning seeks to answer the question:
How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

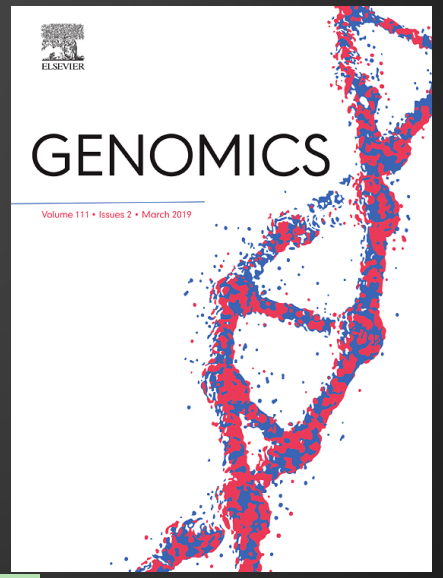
1. Introduction to Machine Learning

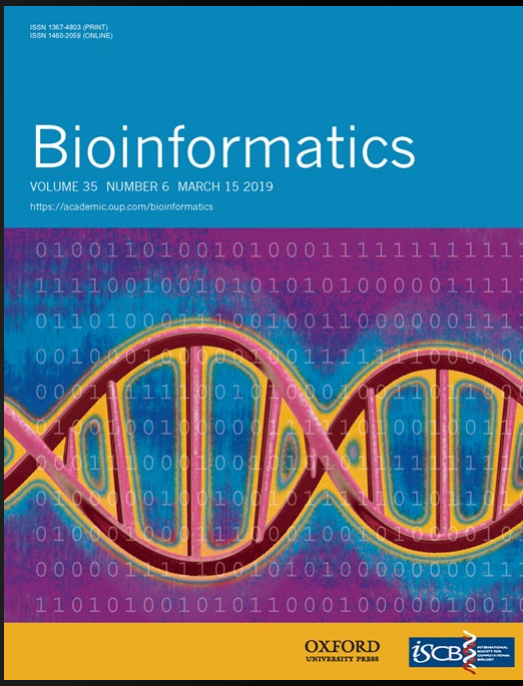


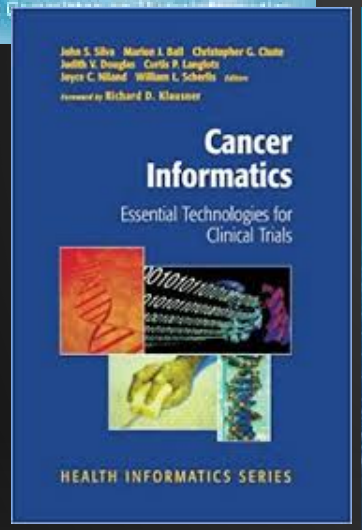
https://www.researchgate.net/publication/323108787_Introduction_to_Machine_Learning

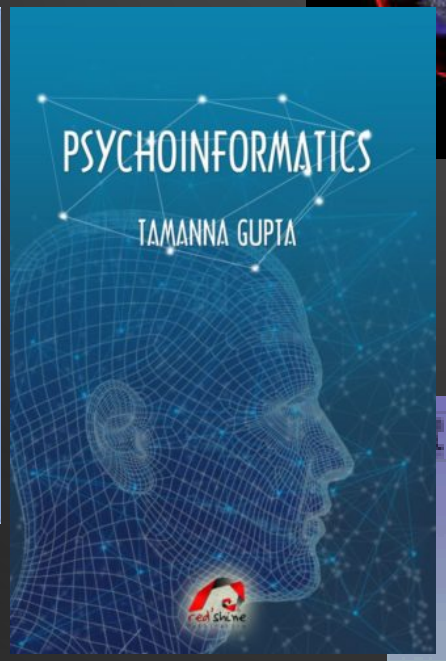
1. Introduction to Machine Learning

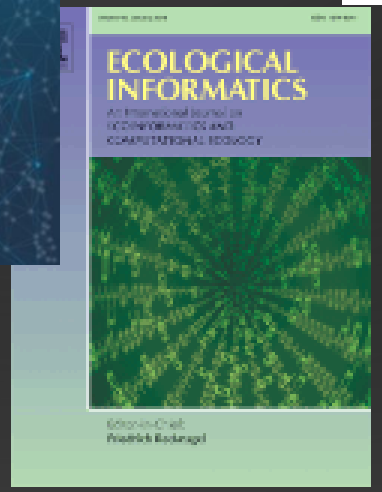


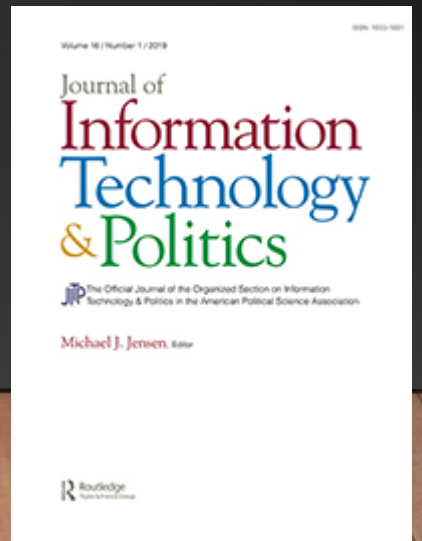


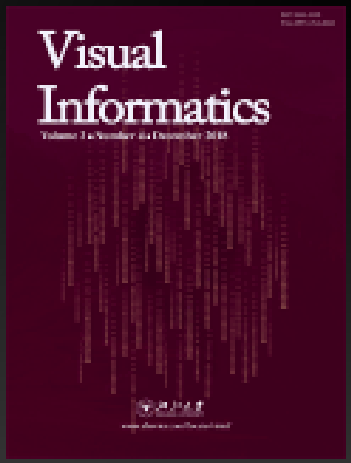






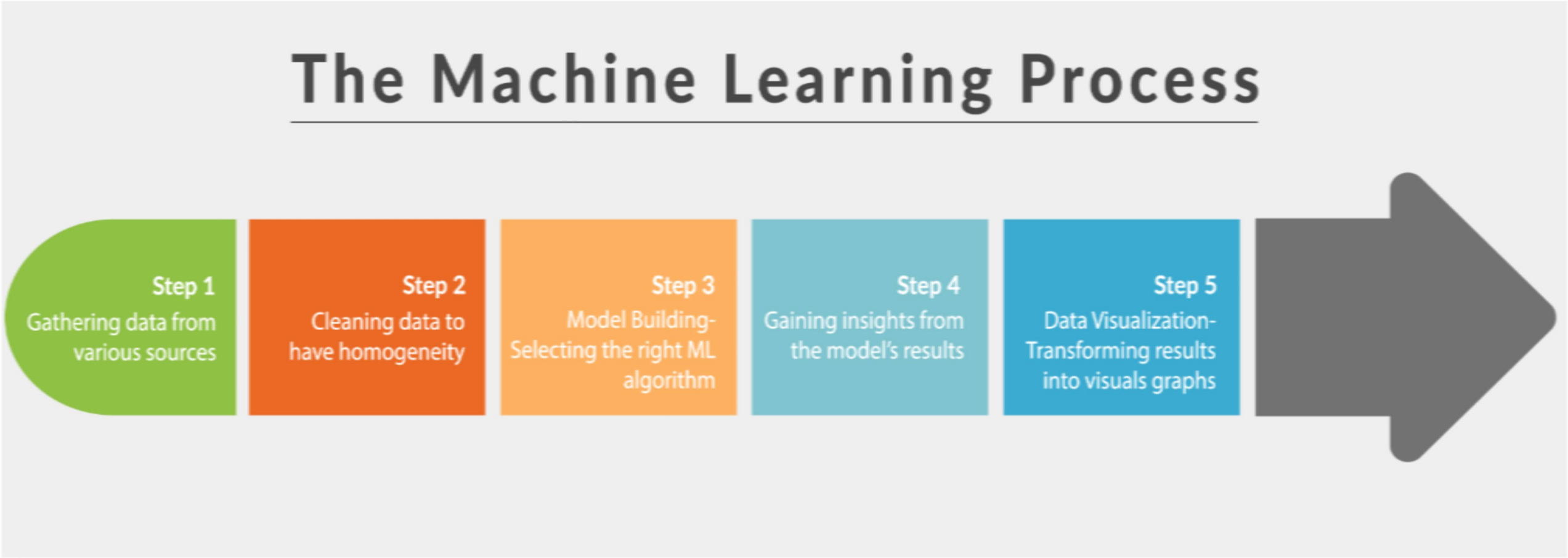




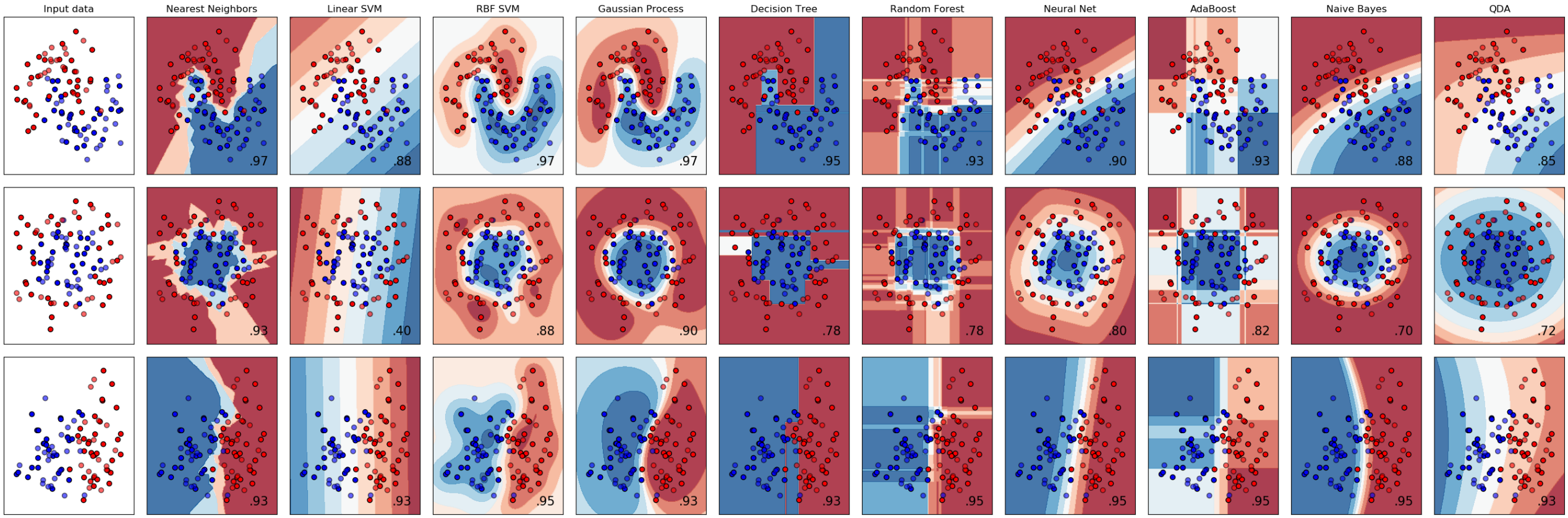


CITI
cyberinfrastructure Technology Integration

CLEMSON
COMPUTING AND
INFORMATION TECHNOLOGY

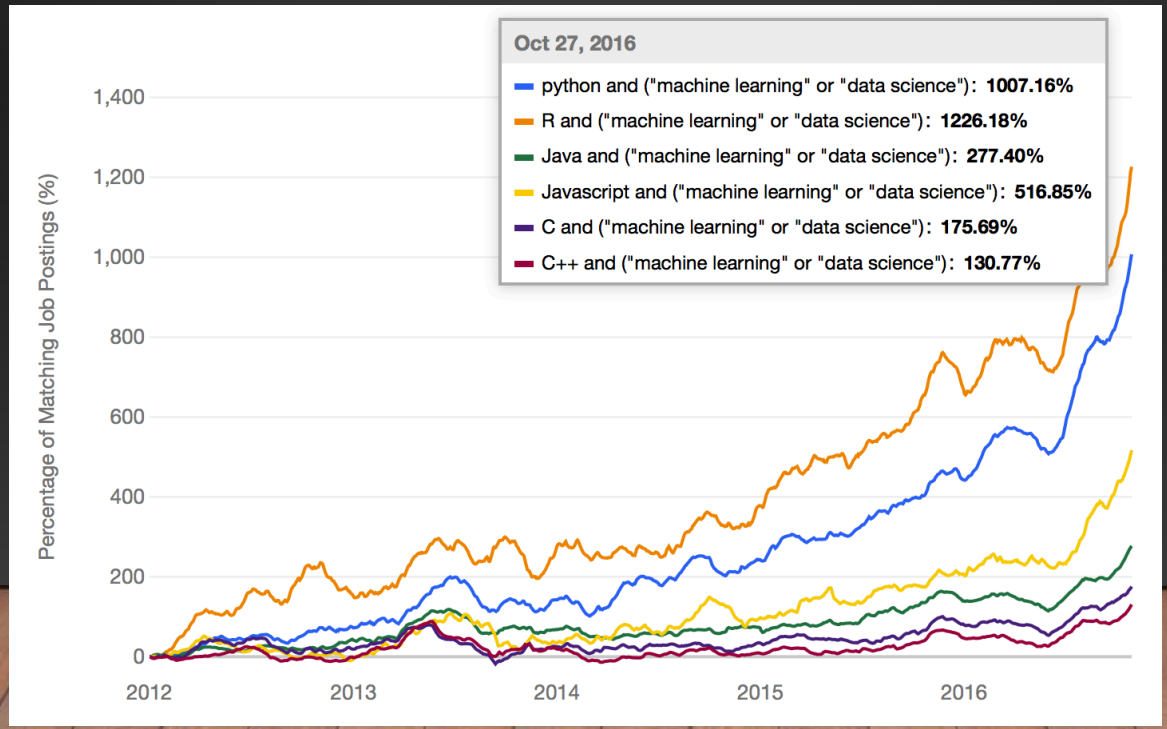

The Machine Learning Process
Step 1
Gathering data from various sources
Step 2
Cleaning data to have homogeneity
Step 3
Model Building- Selecting the right ML algorithm
Step 4
Gaining insights from the model's results
Step 5
Data Visualization- Transforming results into visuals graphs

1. Introduction to Machine Learning



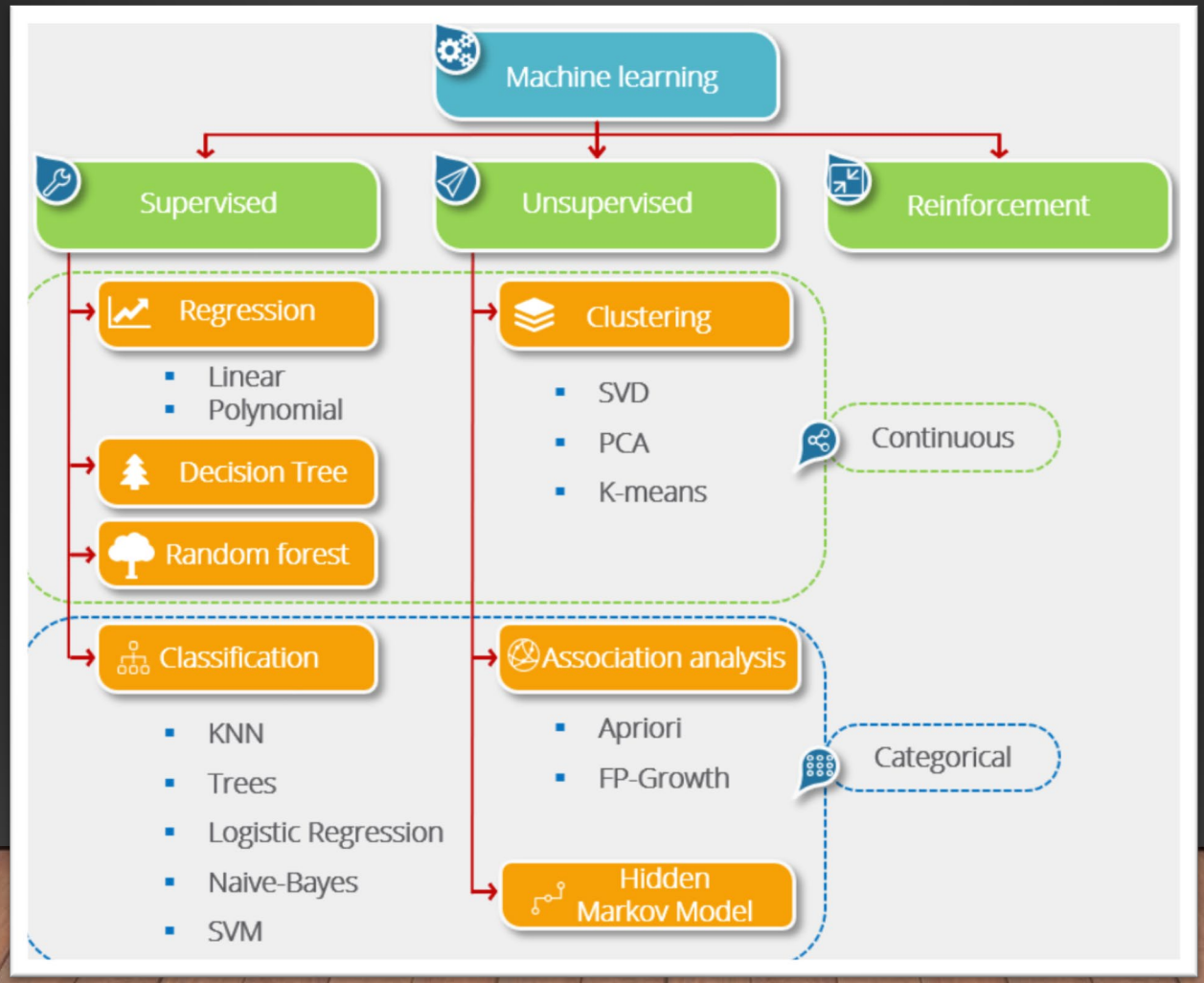
2. Why R?

- **R is used by the best data scientists in the world**. In surveys on Kaggle (the competitive machine learning platform), R is by far the most used machine learning tool.
- **R is powerful because of the breadth of techniques it offers**. The platform has more techniques than any other that you will come across.
- **R is state-of-the-art because it is used by academics**. One of the reasons why R has so many techniques is because academics that develop new algorithms are developing them in R and releasing them as R packages. This means that you can get access to state-of-the-art algorithms in R before other platforms.
- **R is free because it is open source software**. You can download it right now for free and it runs on any workstation platform you are likely to use.
- **R is a great tool for researcher.** PhD students and researchers need lots of statistics for their studies and publications



3. Types of Machine Learning

- Supervised Learning – Train Me! (target are dependent variables)
- Unsupervised Learning – I am self sufficient in learning
- Semi-supervised Learning: combination of both methods, when cost to label are high
- Reinforcement Learning – My life My rules! (Hit & Trial)

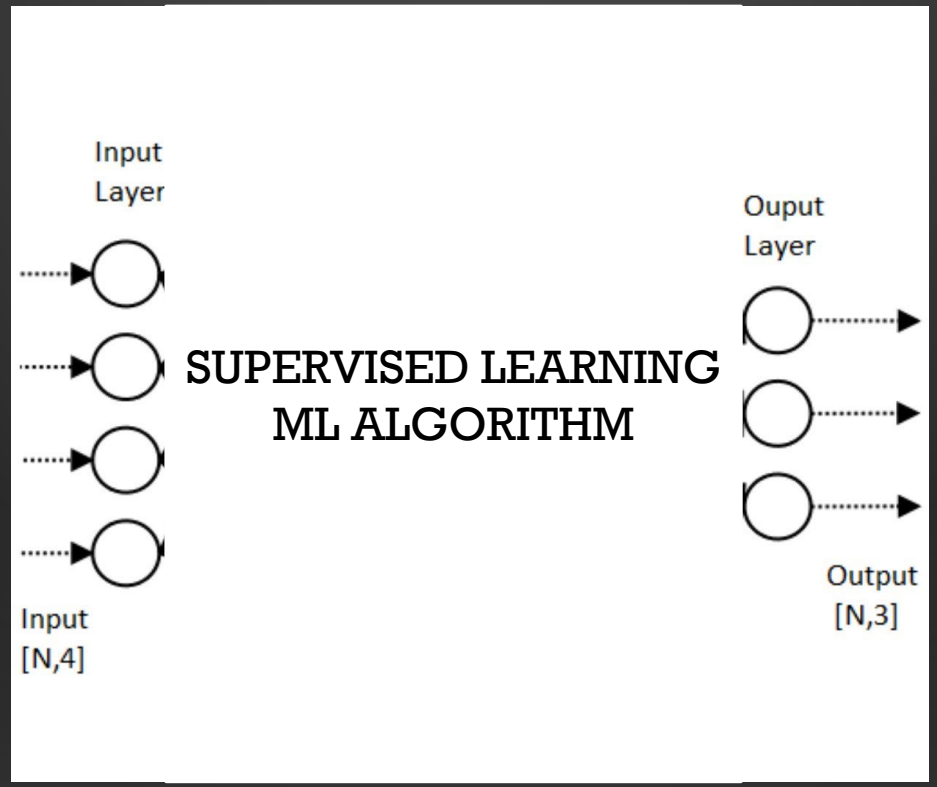


3. Types of Machine Learning

3. Types of Machine Learning

Terminology

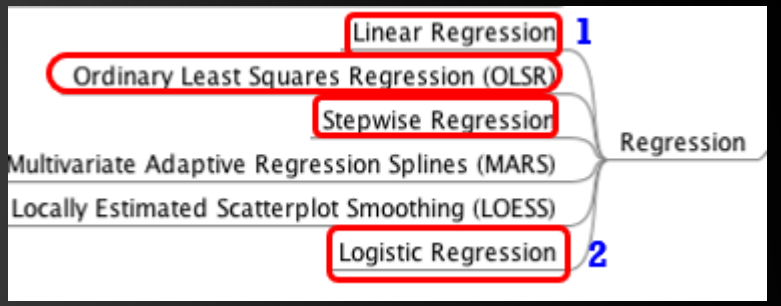
- Input variables
- Independent variables
- Predictors
- Features
- Input Field



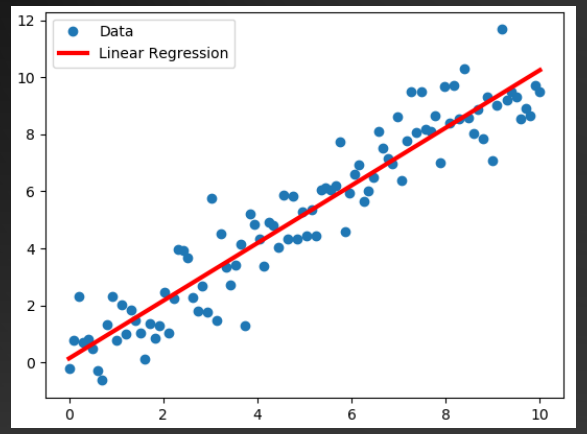
- Output variables
- Dependent variables
- Predictand
- Target variables
- Outcome Field

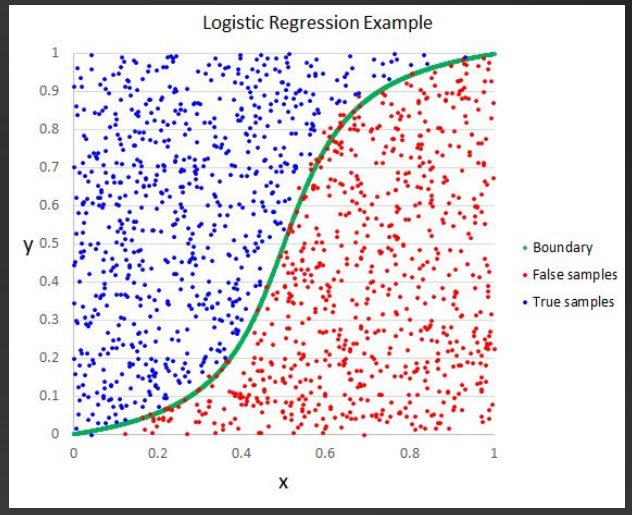
3. Types of Machine Learning

Regression based methods

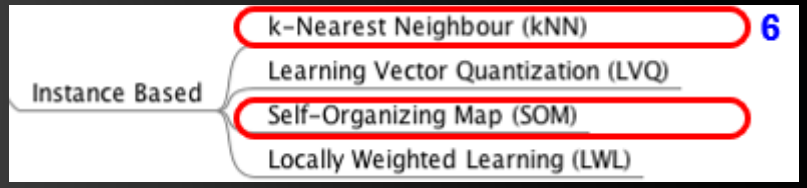


- Most popular & widely used in research for engineer
- Easy to explain and apply
- The relationship between dependent variable and set of independent variables is estimated by probabilistic method/error function minimization



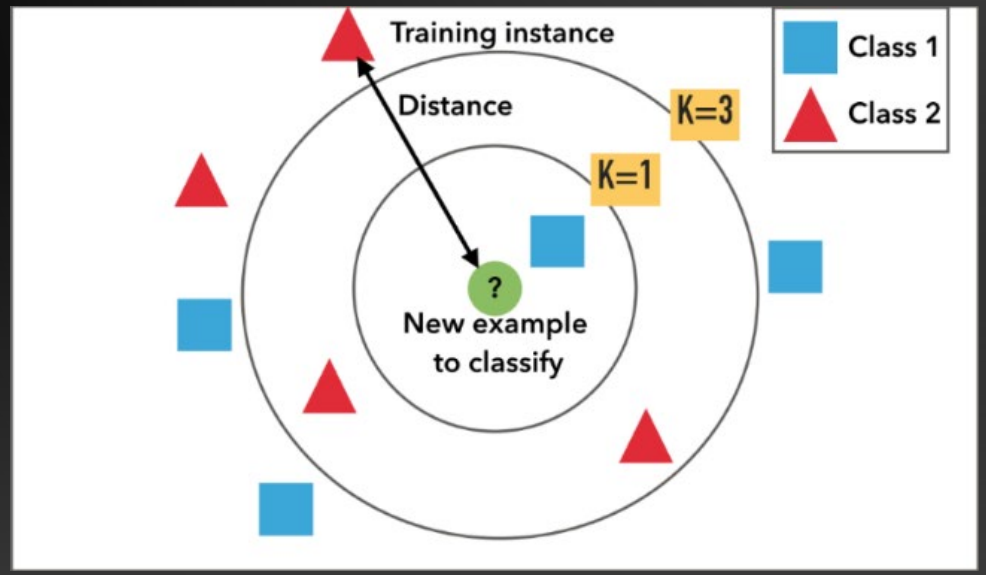


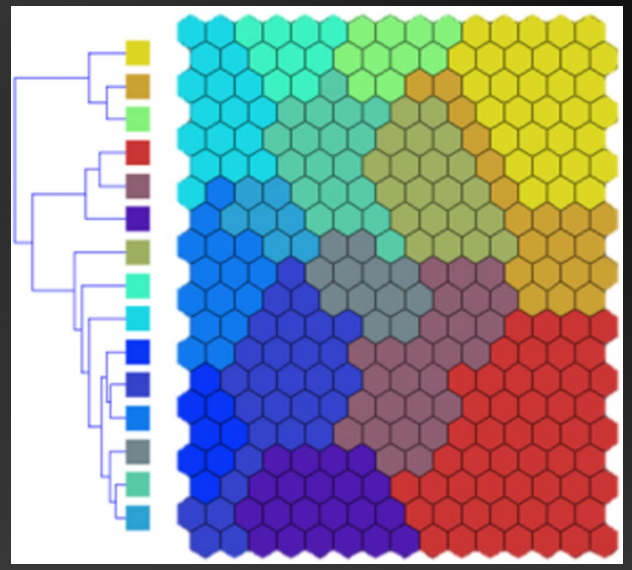
3. Types of Machine Learning



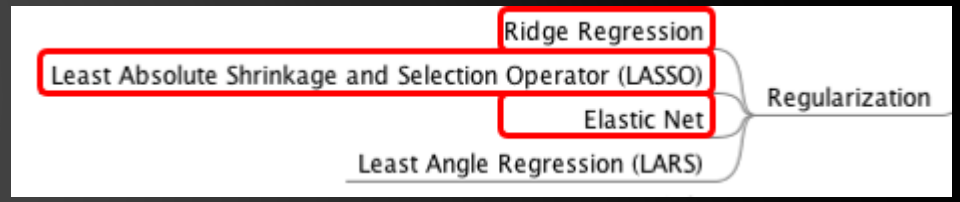
Instance based methods

- So called Distance-based, event-based or memory-based learning
- Self-learning and create a metric to identify whether an object belongs to the class of interest or not
- Learn from sets of events/instances captured in the data



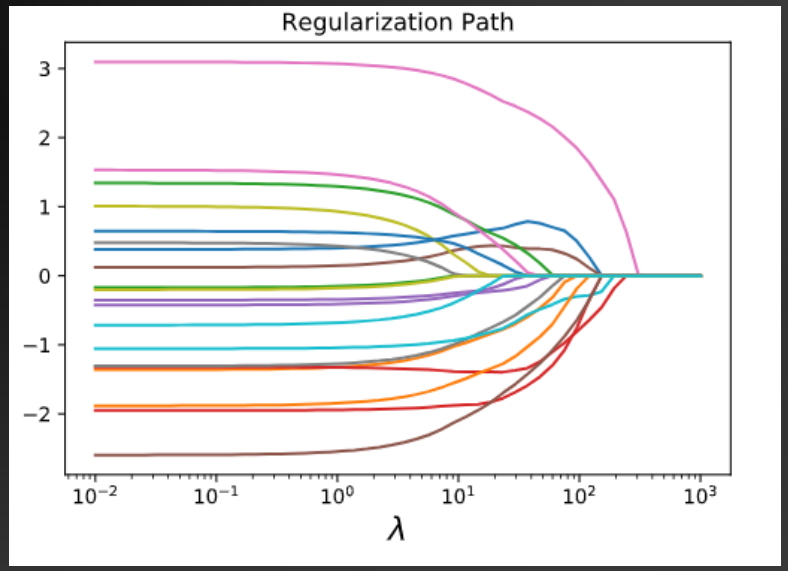


3. Types of Machine Learning



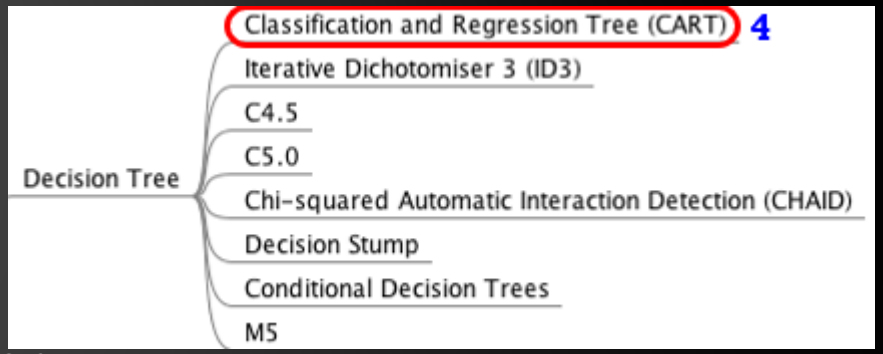
Regularization methods

- An extension of regression methods
- Introduce a penalization term to the loss function for balancing between complexity of model and improvement in results
- Powerful dealing with large number of input dataset

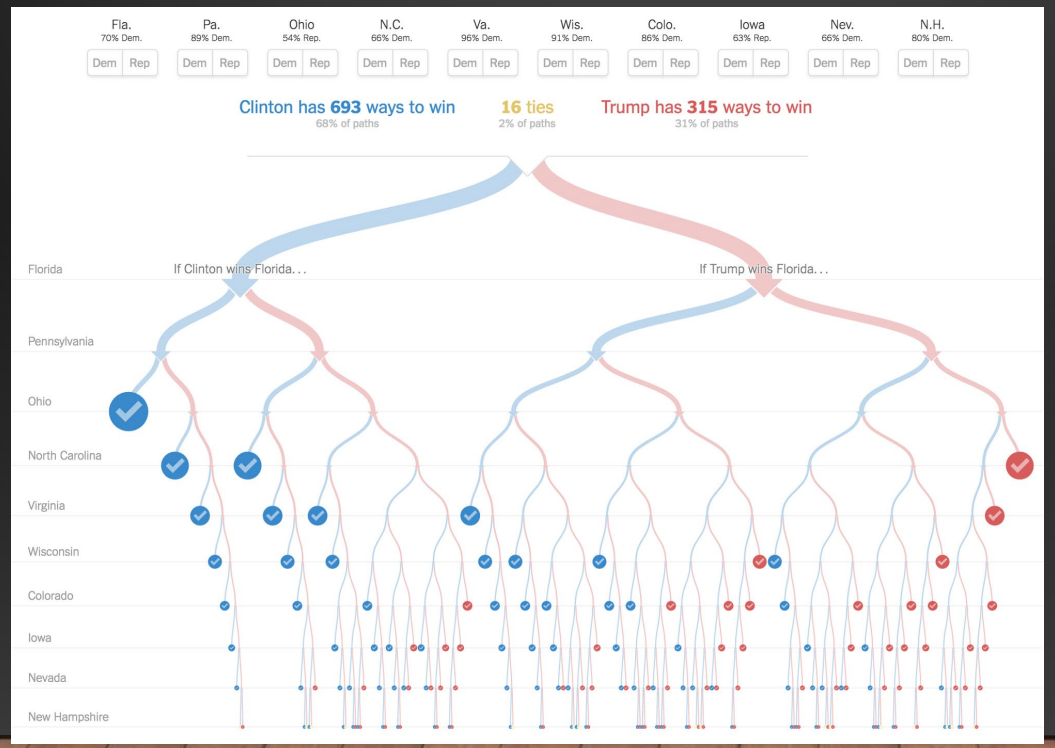


3. Types of Machine Learning

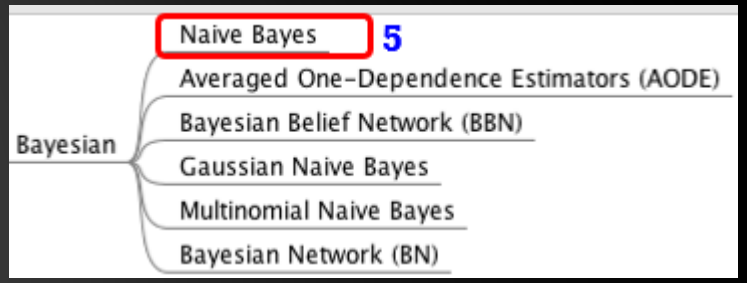
Tree-based algorithms



- Sequential conditional rules applied on the actual data
- Rules are applied serially and a classification decision is made when all conditions are met
- Fast and distributed algorithm

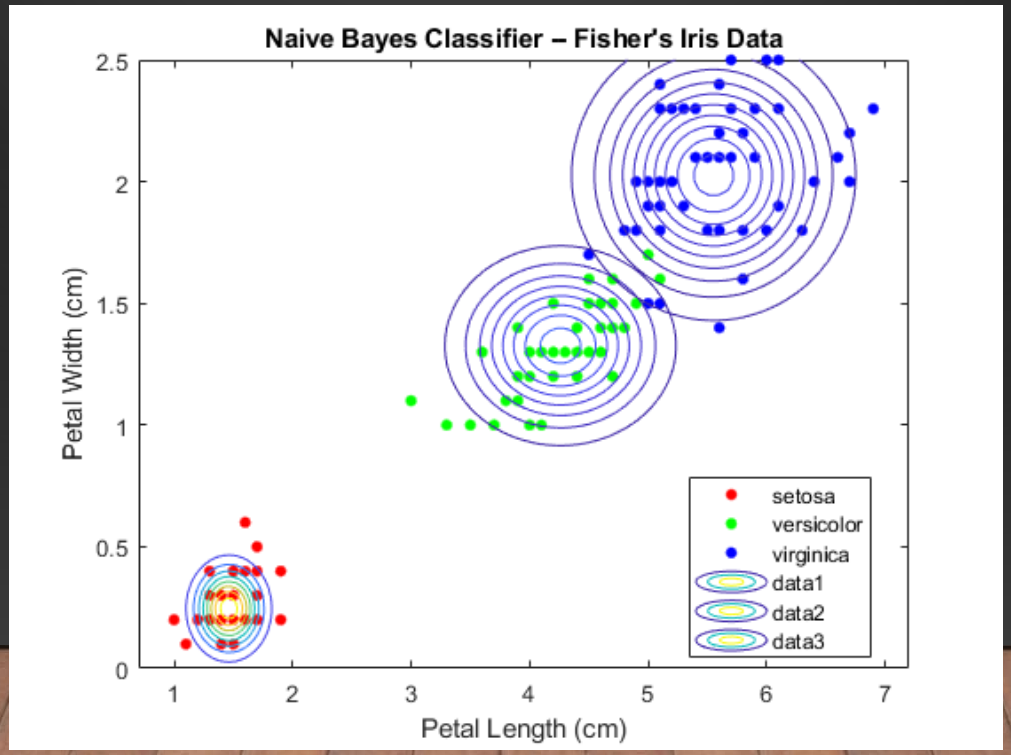


3. Types of Machine Learning

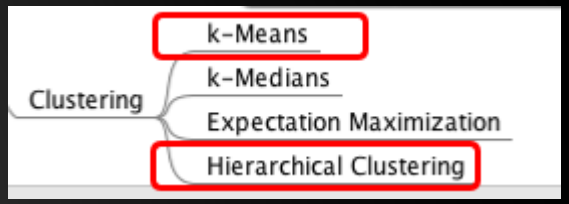


Bayesian Algorithms

- Work based on Bayes Theorem using prior and post distribution
- The machine does not learn from iterative process but using inference from distribution of variable
- Used in most classification and inference testing

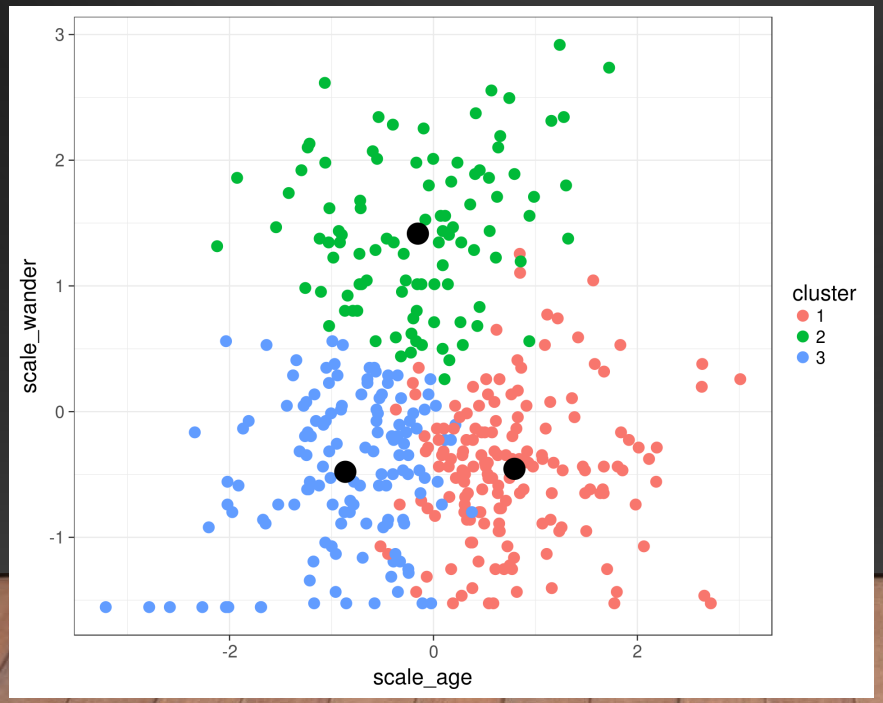


3. Types of Machine Learning

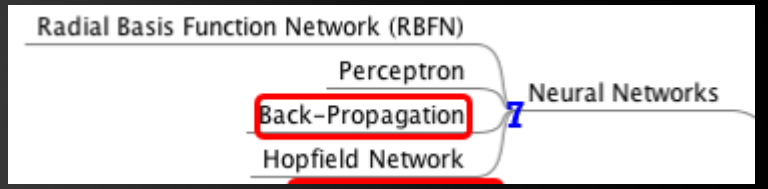


Clustering Algorithms

- Principle of maximization of intra-cluster similarities and minimization inter-cluster similarities
- The measure of similarities determines how the clusters need to be formed
- Unsupervised algorithm: group the data for maximum commonality

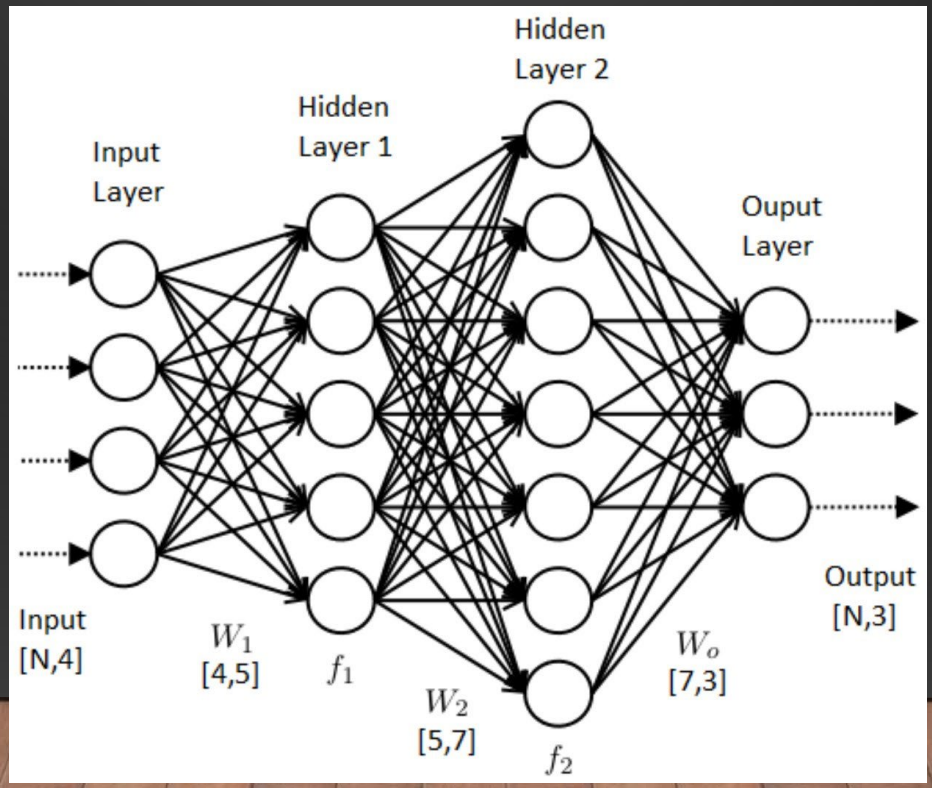


3. Types of Machine Learning



Artificial Neural Networks (ANN)

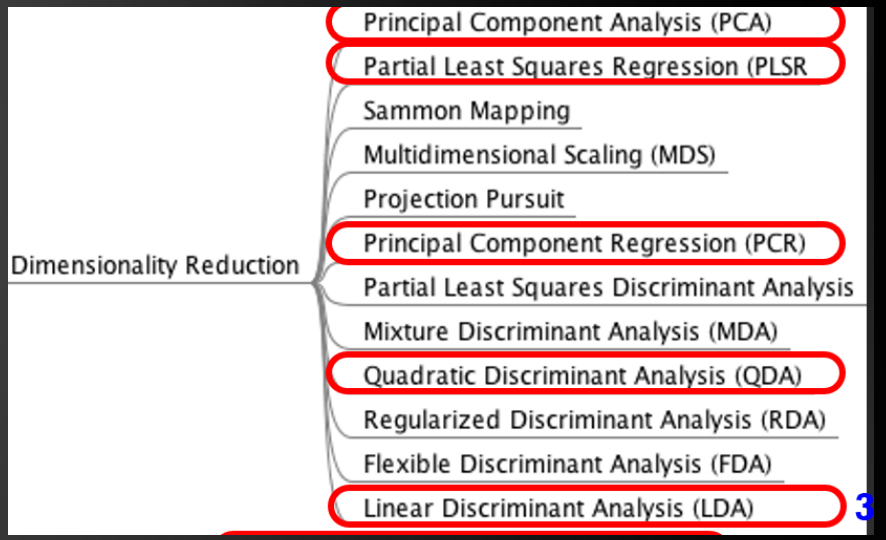
- Inspired by the biological neural networks
- Powerful to learn non-linear relationships
- Recognize higher order relationships among variables
- Used in both supervised/unsupervised learning

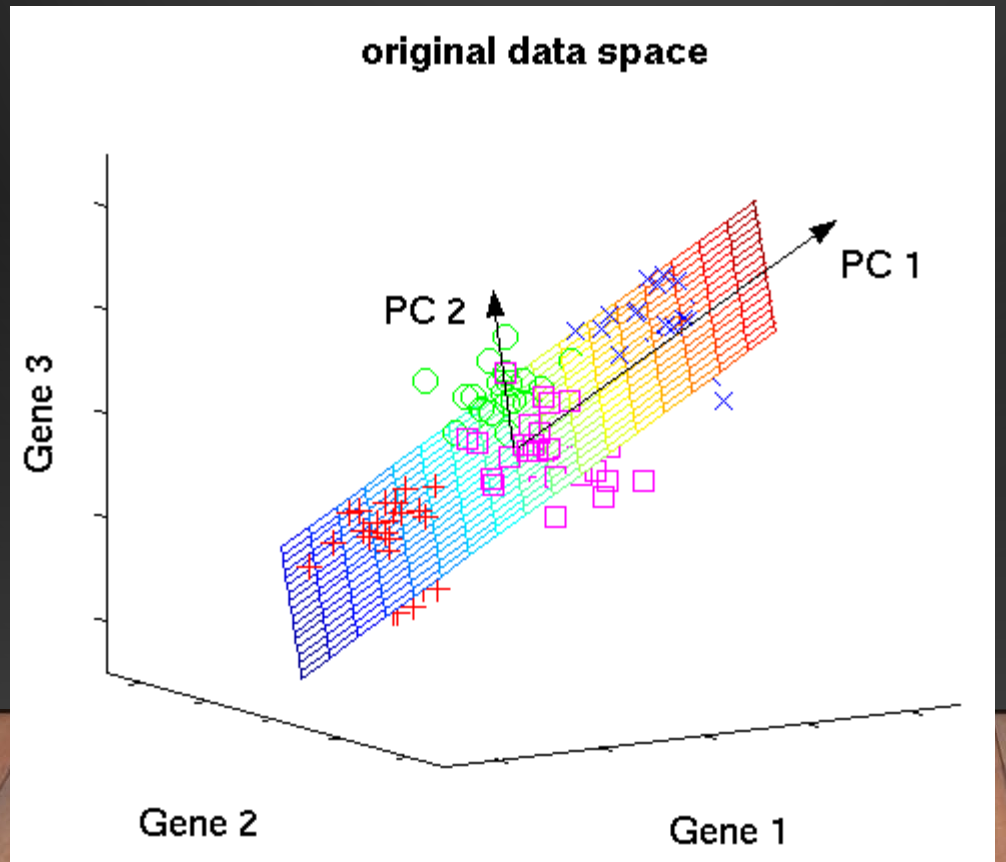


3. Types of Machine Learning

Dimensionality Reduction Algorithm

- Essential method to amplify the signal in data by various transformation
- Reduce number of independent variables (inputs)
- To be applied before modeling

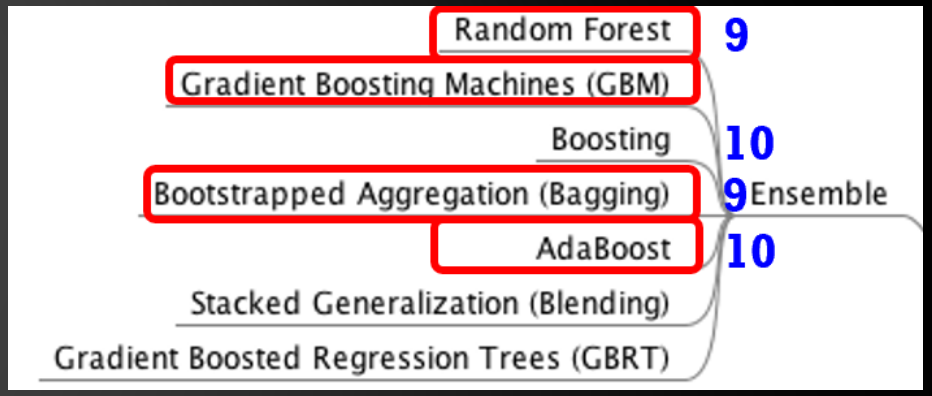




3. Types of Machine Learning

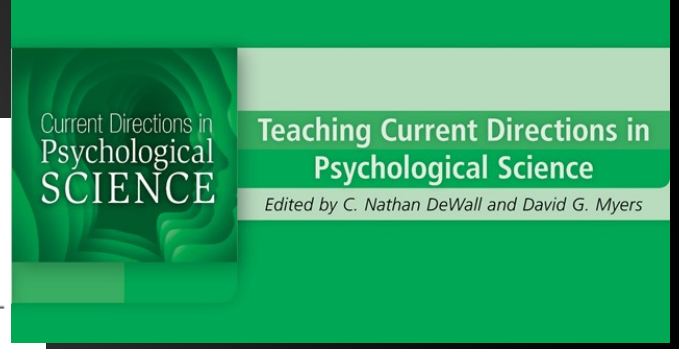
Ensemble Learning

- Combination of results from different ML approach
- Very popular as they have ability to provide superior results
- Possibility to break into independent model to train a distributed network



4. Application of Data Science and Machine Learning to Psychology

4. Application of Data Science and Machine Learning to Psychology



aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
21(6) 391–397

DOI: 10.1177/0963721412457362
http://cdps.sagepub.com
SAGE

Psychoinformatics: New Horizons at the Interface of the Psychological and Computing Sciences

Tal Yarkoni
Institute of Cognitive Science, University of Colorado Boulder

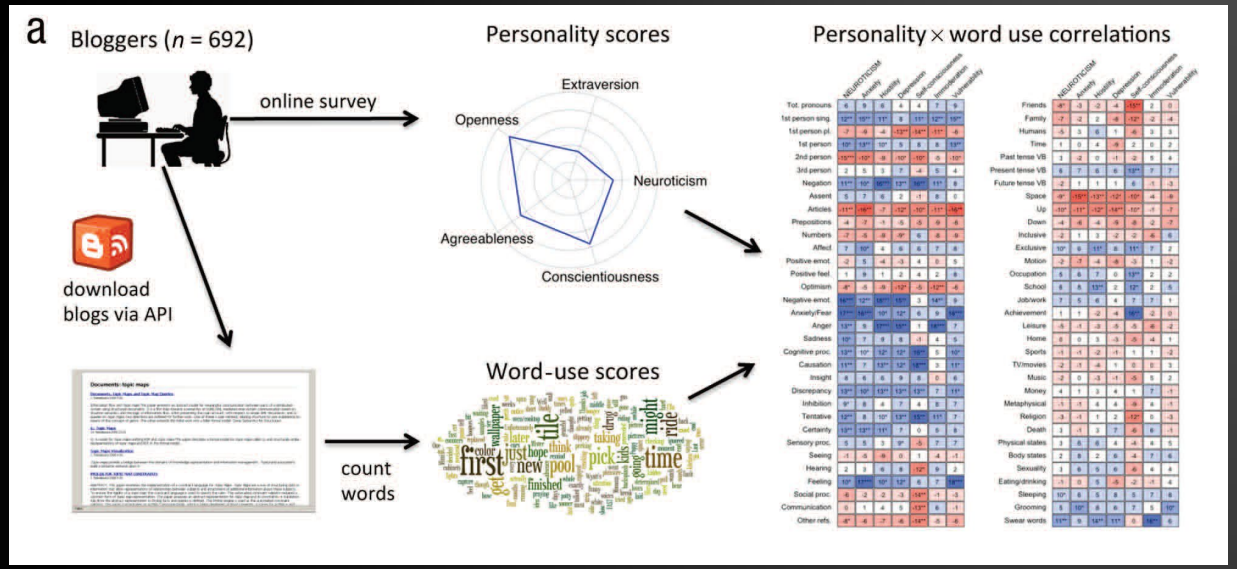
Abstract

Psychologists live in an increasingly data-rich world, and our ability to make continued progress in understanding the mind and brain depends on finding new ways to organize and synthesize an ever-expanding body of knowledge. In this article, I review current research in psychoinformatics—an emerging discipline that uses tools and techniques from the computer and information sciences to improve the acquisition, organization, and synthesis of psychological data. I focus on several areas where the application of informatics approaches has already paid large dividends, leading to advances including novel data-collection approaches, the adaptation of computational techniques and insights, the enhanced aggregation and organization of psychological data, large-scale data mining and synthesis, and improved research and publication practices. I argue that in the coming years, informatics approaches are likely to play the same instrumental role in shaping psychological research that they have already played in other fields, such as genetics and neuroscience.

Keywords

informatics, methods, data mining, information science

4. Application of Data Science and Machine Learning to Psychology



1. Download the content of blogs
2. Measure the language use
3. Measure the word/number of words use
4. Evaluate personality scores
5. Compute the correlation between personality and word

(more * denotes significant value)

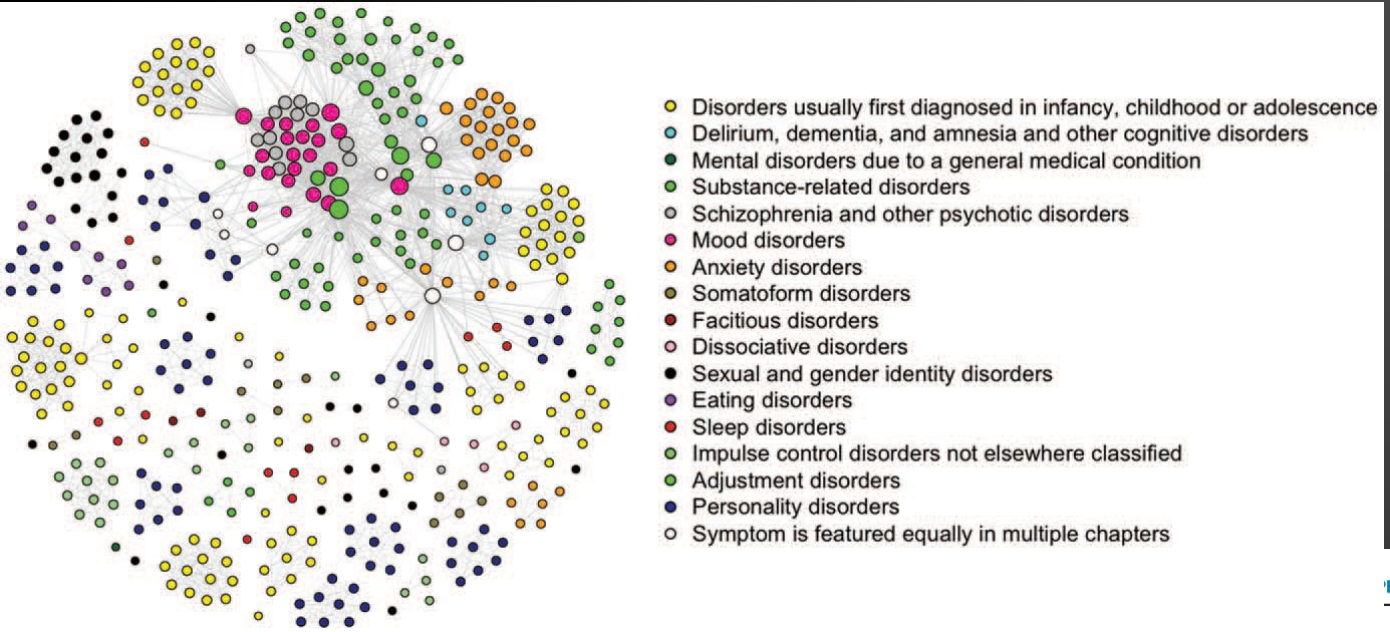
More number of bloggers, more significant

Published in final edited form as:
J Res Pers. 2010 June 1; 44(3): 363–373. doi:10.1016/j.jrp.2010.04.001.

Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers

Tal Yarkoni
University of Colorado at Boulder

4. Application of Data Science and Machine Learning to Psychology



1. Network analyses for small-world structure to symptoms
2. Clustering method in the DSM-IV network

EN ACCESS Freely available online

PLoS one

The Small World of Psychopathology

Denny Borsboom*, Angélique O. J. Cramer, Verena D. Schmittmann, Sacha Epskamp, Lourens J. Waldorp

Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

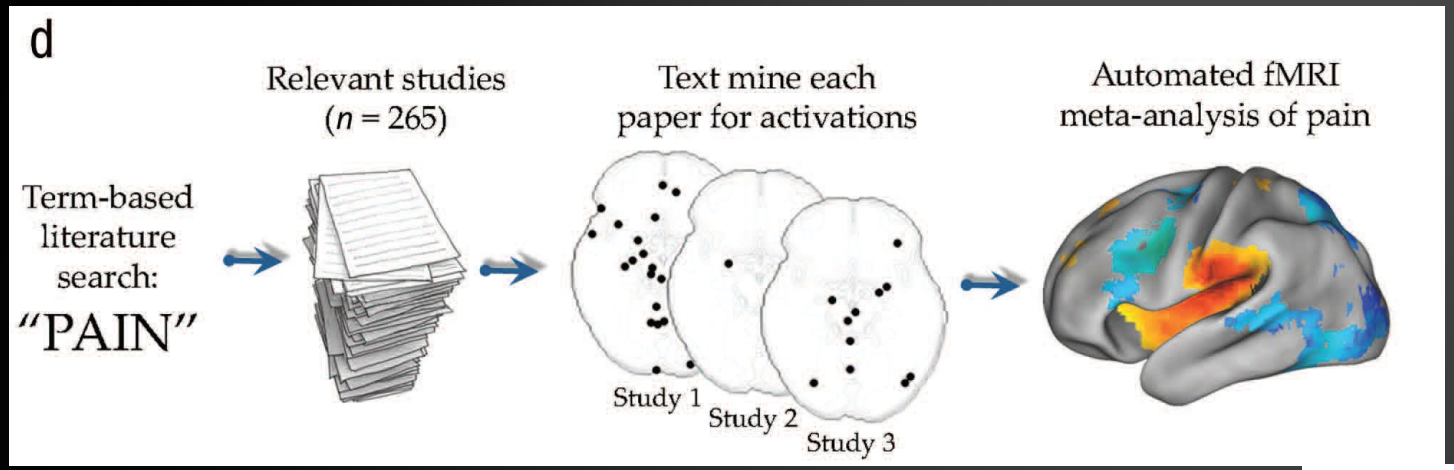
Abstract

Background: Mental disorders are highly comorbid: people having one disorder are likely to have another as well. We explain empirical comorbidity patterns based on a network model of psychiatric symptoms, derived from an analysis of symptom overlap in the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV).

Principal Findings: We show that a) half of the symptoms in the DSM-IV network are connected, b) the architecture of these connections conforms to a small world structure, featuring a high degree of clustering but a short average path length, and c) distances between disorders in this structure predict empirical comorbidity rates. Network simulations of Major Depressive Episode and Generalized Anxiety Disorder show that the model faithfully reproduces empirical population statistics for these disorders.

Conclusions: In the network model, mental disorders are inherently complex. This explains the limited successes of genetic, neuroscientific, and etiological approaches to unravel their causes. We outline a psychosystems approach to investigate the structure and dynamics of mental disorders.

4. Application of Data Science and Machine Learning to Psychology



1. Use text mining, meta-analysis & ML technique
2. Automatically conduct large scale , high quality neuroimaging to address long standing inferential problems & human subjects

Large-scale automated synthesis of human functional neuroimaging data

nature methods

Tal Yarkoni[1], Russell A Poldrack[2–4], Thomas E Nichols[5,6], David C Van Essen[7] & Tor D Wager[1]

The rapid growth of the literature on neuroimaging in humans has led to major advances in our understanding of human brain function but has also made it increasingly difficult to aggregate and synthesize neuroimaging findings. Here we describe and validate an automated brain-mapping framework that uses text-mining, meta-analysis and machine-learning techniques to generate a large database of mappings between neural and cognitive states. We show that our approach can be used to automatically conduct large-scale, high-quality neuroimaging meta-analyses, address long-standing inferential problems in the neuroimaging literature and support accurate 'decoding' of broad cognitive states from brain activity in both entire studies and individual human subjects. Collectively, our results have validated a powerful and generative framework for synthesizing human neuroimaging data on an unprecedented scale.

The development of noninvasive neuroimaging techniques such as functional magnetic resonance imaging (fMRI) has spurred rapid growth of literature on human brain imaging in recent years. In 2010 alone, more than 1,000 fMRI articles had been published[1]. This proliferation has led to substantial advances in our understanding of the human brain and cognitive function; however, it

analyses[1], our framework is fully automated and allows rapid and scalable synthesis of the neuroimaging literature. We show that this framework can be used to generate large-scale meta-analyses for hundreds of broad psychological concepts; support quantitative inferences about the consistency and specificity with which different cognitive processes elicit regional changes in brain activity; and decode and classify broad cognitive states in new data solely on the basis of observed brain activity.

RESULTS

Overview

Our methodological approach includes several steps (**Fig. 1a**). First, we used text-mining techniques to identify neuroimaging studies that used specific terms of interest (for example, 'pain', 'emotion', 'working memory' and so on) at a high frequency (>1 in 1,000 words) in the article text. Second, we automatically extracted activation coordinates from all tables reported in these studies. This approach produced a large database of term-to-coordinate mappings; here we report results based on 100,953 activation foci drawn from 3,489 neuroimaging studies published in 17 journals (Online Methods). Third, we conducted automated meta-analyses of hundreds of psychological concepts, producing an extensive set

CITI
cyberinfrastructure technology integration

4. Application of Data Science and Machine Learning to Psychology

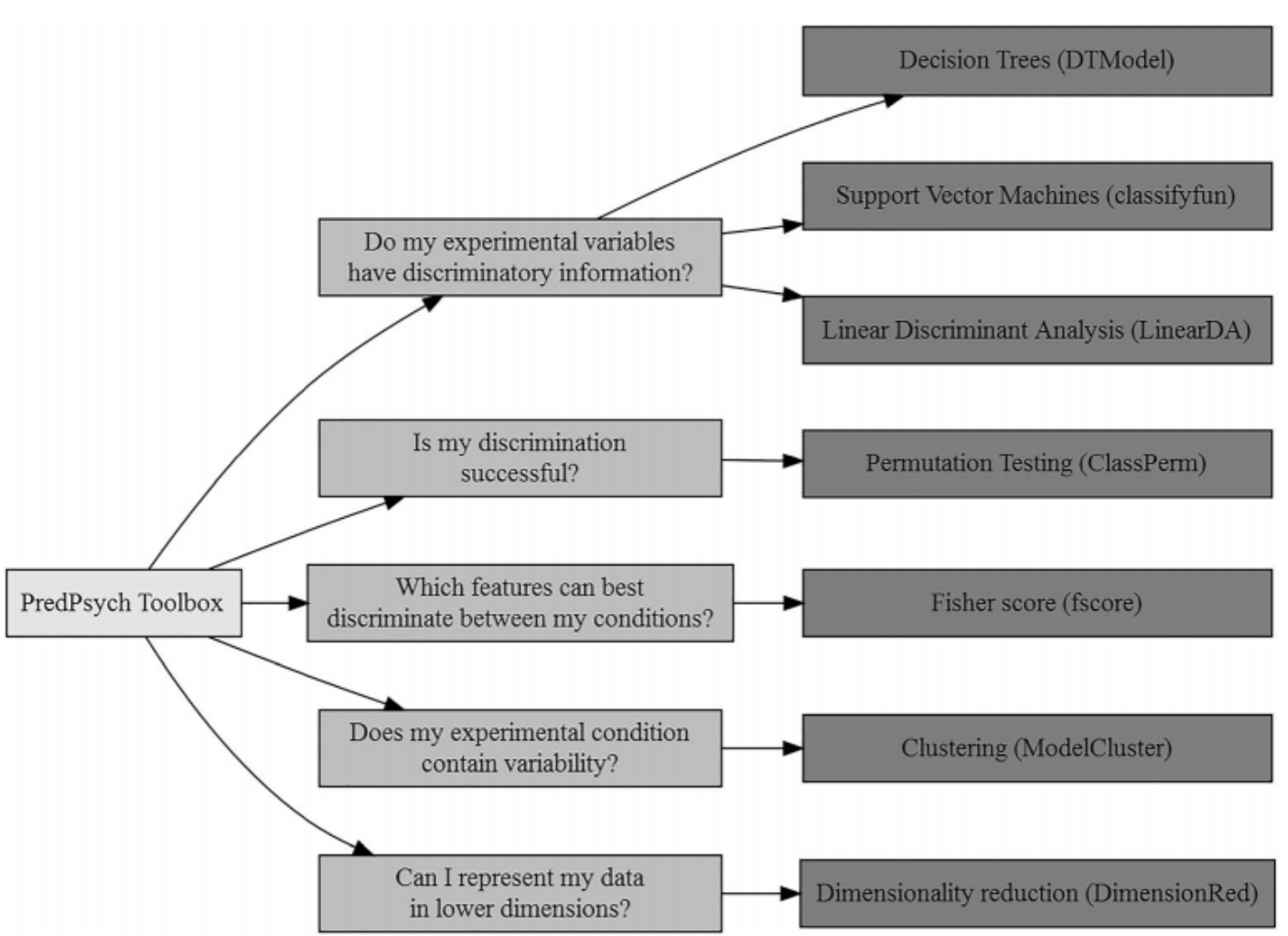
Psychology toolbox in R: "psych"

- The psych package has been developed to help psychologists do basic research.
- Many of the functions were developed to supplement a book

http://personality-project.org/r/

4. Application of Data Science and Machine Learning to Psychology

Machine learning toolbox in R:



Behav Res (2018) 50:1657–1672
https://doi.org/10.3758/s13428-017-0987-2

CrossMark

***PredPsych*: A toolbox for predictive machine learning-based approach in experimental psychology research**

Atesh Koul[1,2] · Cristina Becchio[1,2] · Andrea Cavallo[1,2]

CLEMSON
COMPUTING AND INFORMATION TECHNOLOGY

5. Discussions

1. What is your input data?
2. What is your expected output?
3. What is your usual workflow to analyse your data, i.e your favorite software?
4. How many algorithms do you use in your study? Name them.
5. How do you visualize your analyses?

6. Hands-on session

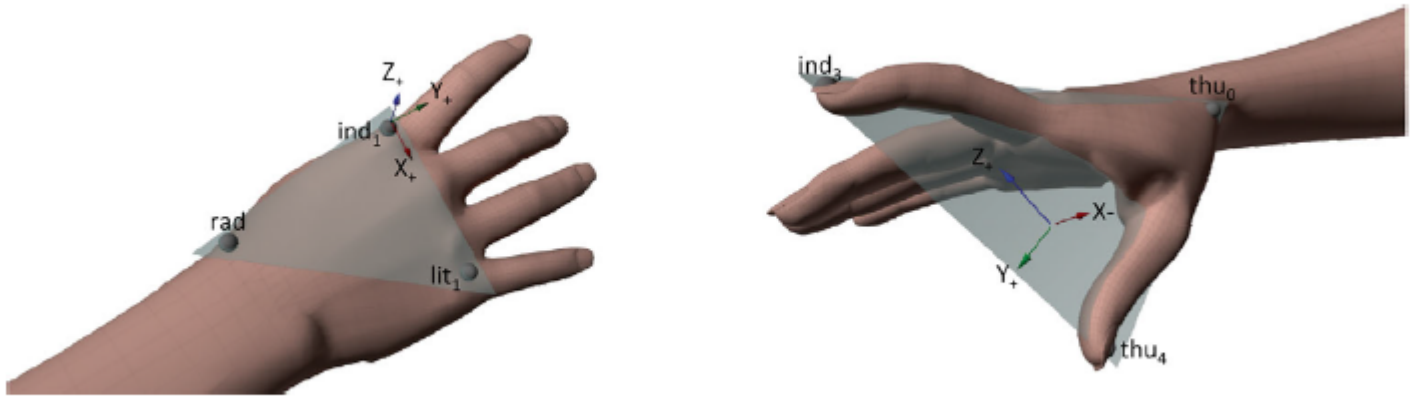


Fig. 2 Hand model for estimating kinematics variables. Schematic showing the hand model depicting global and local frames of reference used for the calculation of kinematics variables