Introduction to Applied Machine Learning in R

Dr. Brian J. Spiering

Introduction to Applied Machine Learning in R

Introduction

$$P(Z_{(m,n)} = k|Z \qquad W; \alpha, \beta)$$

$$\propto P(Z_{(m,n} \qquad \beta)$$

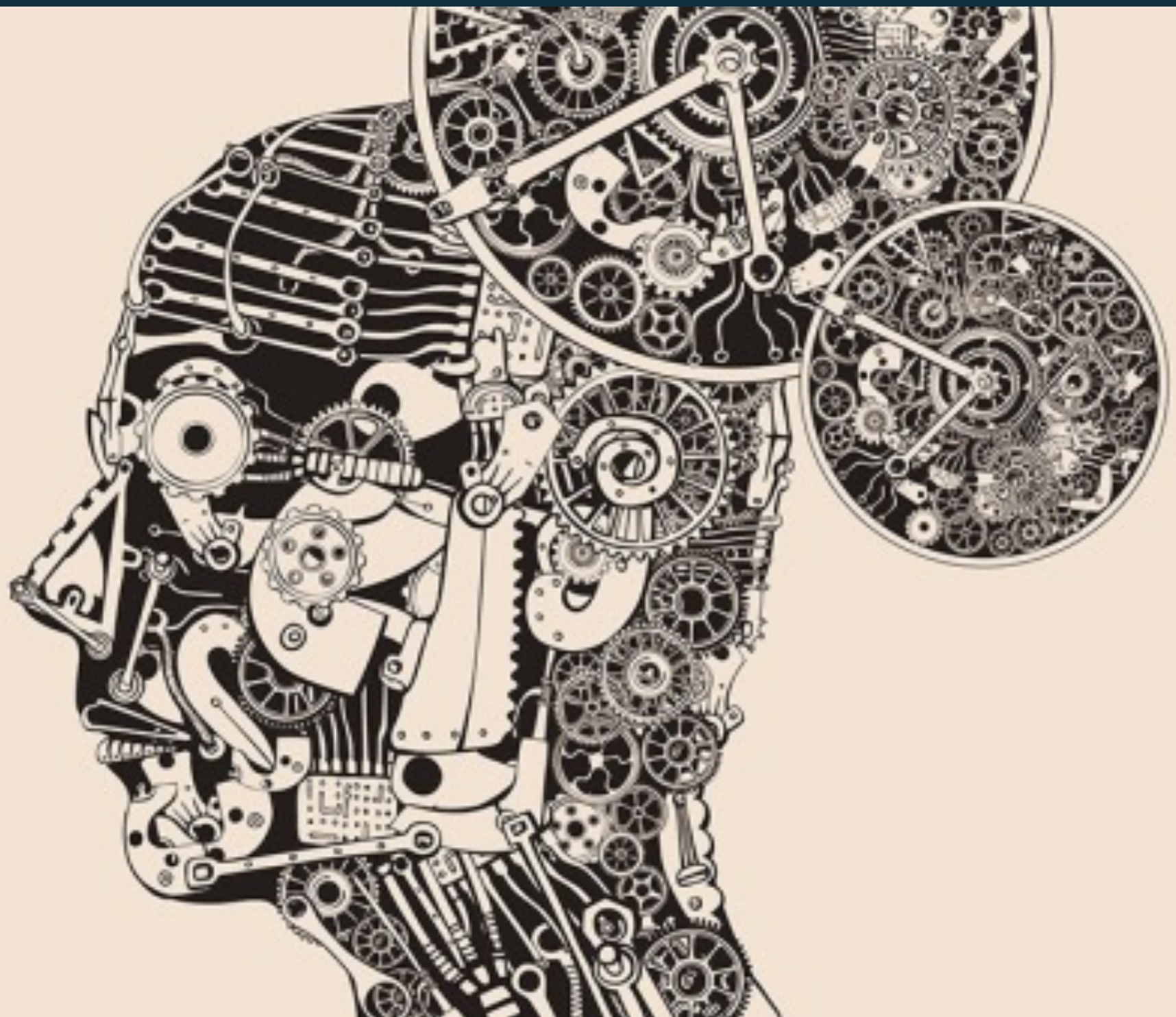
$$= \left(\Gamma \qquad \frac{\sum_{i=1} \alpha_i}{\prod_{i=1} \Gamma(\alpha_i)} \right)^{M} \prod_{j \neq n} \frac{\prod_{i}^{K} \qquad + \alpha_i)}{\qquad + \alpha_i)}$$

$$\frac{\left(\sum_{r=1}^{V} \beta_r\right) \qquad}{\prod_{r=1}^{V} \Gamma(\beta \qquad} \prod_{r \neq v} \Gamma\left(n^{i}_{(\cdot),r} \qquad\right)$$

$$\times \qquad \frac{\qquad + \alpha_i)}{\qquad + \alpha_i)} \prod_{i=1}^{K} \frac{\Gamma(\qquad + \beta_v)}{\Gamma(\qquad n^{i}_{(\cdot),r} + \beta_r)}$$

$$\propto \frac{\prod_{i=1}^{K} \Gamma \qquad \alpha_i) \qquad}{\Gamma\left(\sum_{i=1}^{K} n^{i}_{m,(\cdot)} \qquad\right)} \qquad \frac{\qquad}{\Gamma\left(\sum_{r=1}^{V} n^{i}_{(\cdot),r} + \beta_r\right)}.$$

Applied (vs. Theoretical)

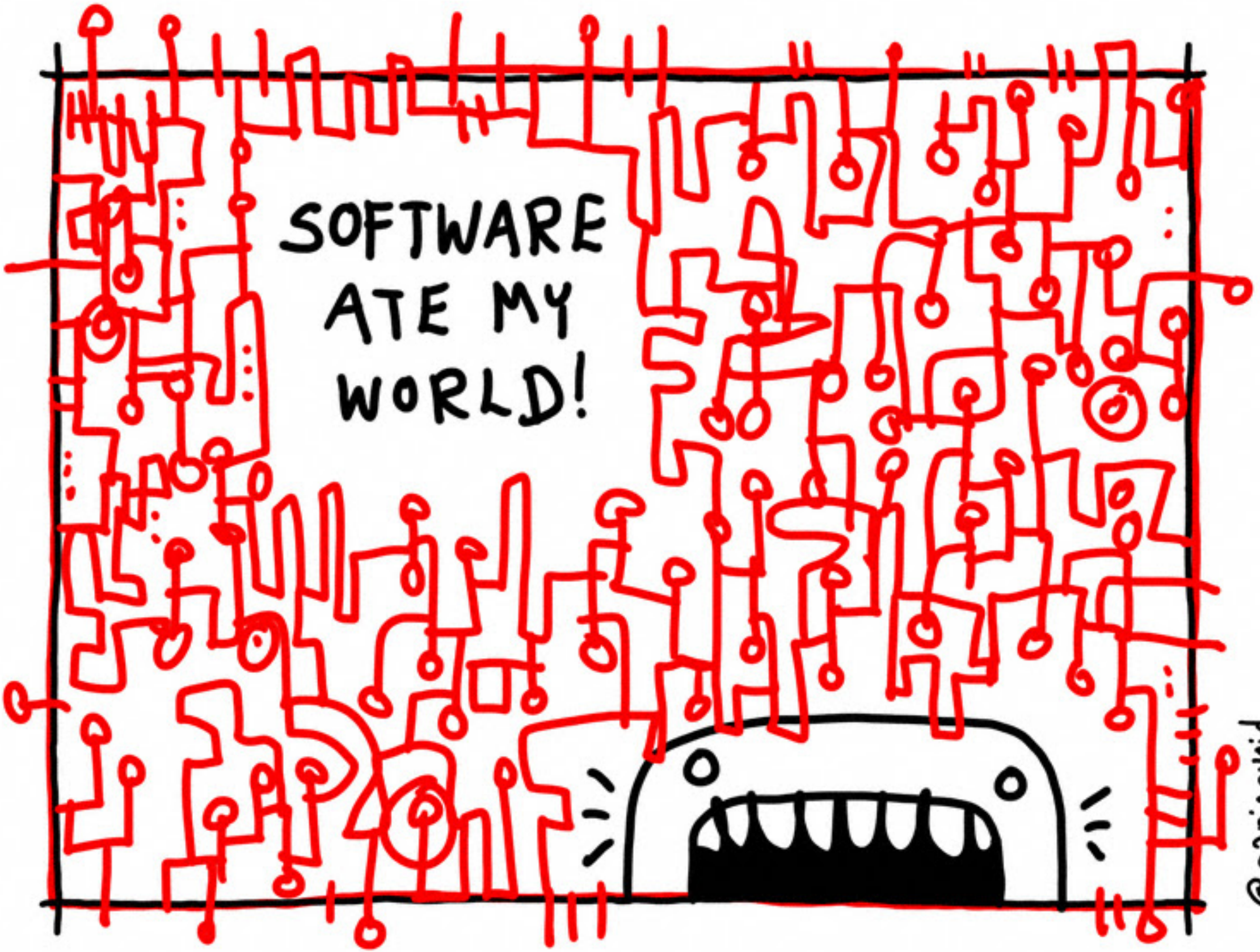


What I talk about when I talk about machine learning

(apologies to Raymond Carver & Birdman)

machine learning (ml):

Programming computers to automatically learn and generalize from examples


SOFTWARE
ATE MY
WORLD!
@gapingvoid

Software might eat the world,
but algorithms will digest it

-Bruno Aziza

Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

ARTIFICIAL INTELLIGENCE
IBM WATSON · MetaMind · Numenta · ai·one · Cycorp · Microsoft Research · nara · Reactor · SI SCALED INFERENCE

DEEP LEARNING
vicarious · Vision Factory · facebook · Google · SKYMIND · Baidu IDL · ersatz · SignalSense

MACHINE LEARNING
rapidminer · context relevant · 0xdata H2O · DATARPM · LitSgniter · SPARKBEYOND · Azure ML · yhat · wise.io · Sense · GraphLab · Alpine · tonian

NLP PLATFORMS
cortical.io · idibon · LUMINOSO · wit.ai · Maluuba

PREDICTIVE APIS
AlchemyAPI · MINDOPS · Google · bigml · indico · ALGORITHMIA · Expect Labs · PredictionIO

IMAGE RECOGNITION
clarifai · MADBITS · DNNresearch · DEXTRO · VISENZE · lookflow

SPEECH RECOGNITION
GRIDSPACE · popup archive · NUANCE

RETHINKING ENTERPRISE

SALES
Preact · AVISO · RelateIQ · NGDATA · CLARABRIDGE · FRAMED · infer · ATTENSITY · causata

SECURITY / AUTHENTICATION
CROSSMATCH · conjur · EYEVERIFY · BITSIGHT · CYLANCE · AREA 1 SECURITY · bionym

FRAUD DETECTION
sift science · SOCURE · ThreatMetrix · feedzai · Brighterion · VERAFIN

HR / RECRUITING
TalentBin · entelo · predikt · Connectifier · gild · hiQ · CONCEPTNODE

MARKETING
brightfunnel · bloomreach · CommandIQ · AIRPR · RADIUS · Tellapart · people pattern · Freshplum

PERSONAL ASSISTANT
Siri · X.ai · Google now · Cortana · cleversense · tempo · Robinlabs · KASISTO · fuse|machines · VIV · CLARA LABS

INTELLIGENCE TOOLS
ADATAD · Palantir · Quid · Digital Reasoning · FirstRain

RETHINKING INDUSTRIES

ADTECH
METAMARKETS · dstillery · rocketfuel · YieldMo · ADBRAIN

AGRICULTURE
BLUE RIVER · TerrAvion · ceresimaging · HARVEST CROO · THE CLIMATE CORPORATION · tule

EDUCATION
declara · coursera · KNEWTON · kidaptive

FINANCE
Bloomberg · FinGenius · alphasense · KENSHO · Dataminr · minettabrook · BINATIX

LEGAL
Lex Machina · brightleaf · COUNSELYTICS · RAVEL · JUDICATA · Brevia · DiligenceEngine

MANUFACTURING
SIGHT MACHINE · MICROSCAN · BOULDER IMAGING · IVISYS

MEDICAL
Parzival · transcriptic · ZEPHYR HEALTH · Genescient · grand round table · bina · TUTE GENOMICS

OIL AND GAS
kaggle · AYASDI · TACHYUS · biota · Flutura

MEDIA / CONTENT
Outbrain · newsle · ARRIA · SAILTHRU · wavii · Owlin · NarrativeScience · yseop · Summly · Prismatic · ai

CONSUMER FINANCE
Affirm · inVenture · zestfinance · BILLGUARD · LendUp · LendingClub · Kabbage

PHILANTHROPIES
DataKind · thorn · THE DATA GUILD

AUTOMOTIVE
Google · Continental · Tesla · MobileEye · CRUISE

DIAGNOSTICS
enlitic · 3SCAN · lumiata · ENTOPSY

RETAIL
BAY SENSORS · PRISM SKYLABS · celect · euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence · METAIO · APX · blippAR · META · layar

GESTURAL COMPUTING
THALMICLABS · omek · FLUTTER · LEAP MOTION · eyeSight · 3Gear · GestureTek · nod

ROBOTICS
intel · LIQUID ROBOTICS · SoftBank · iRobot · Boston Dynamics · jibo · anki · evolution robotics

EMOTIONAL RECOGNITION
affectiva · BEYOND VERBAL · EMOTIENT · JIBO LABS · cogito

SUPPORTING TECHNOLOGIES

HARDWARE
NVIDIA · XILINX · QUALCOMM · NERVANA SYSTEMS · TERADEEP · Artificial Learning · rigetti

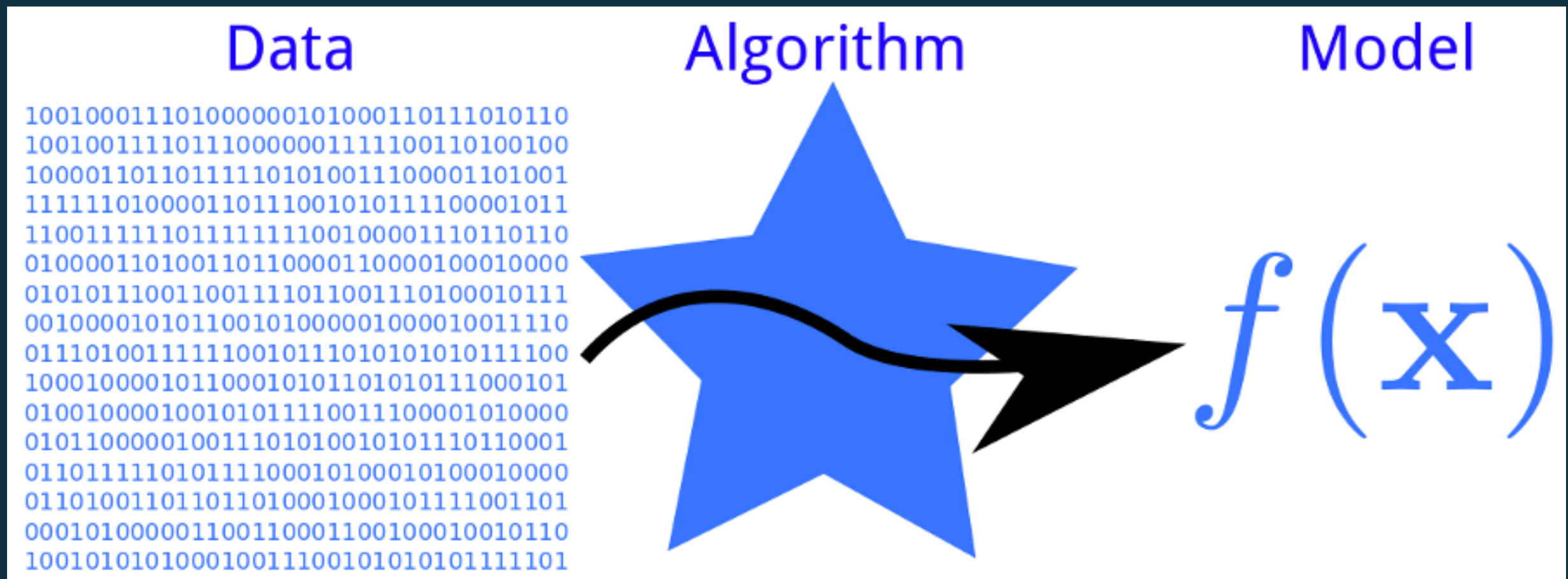
DATA PREP
TRIFACTA · Paxata · tamr · Alation

DATA COLLECTION
diffbot · kimono · CrowdFlower · Connotate · WorkFusion · import.io

www.shivonzilis.com/machineintelligence

Bloomberg BETA

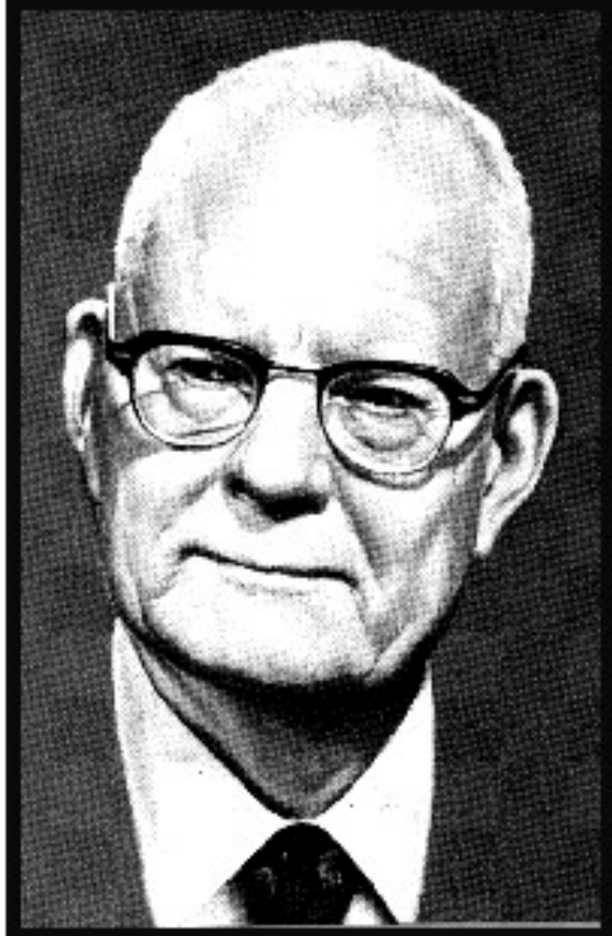
ml process


Data
Algorithm
Model
1001000111010000001010001101110101101001001111011100000011111001101001001000011011011111101010011100001101001111111010000110111001010111100001011100111111011111111100100001110110110100000110100110110000011000010001000001010110011001111011001111010001011100010110000101011001010000010000100111100100001010110011111100101110101010111100100010000101100010101101010111000101001000010010101111001110000101000001010110000010011101010010101110110001011011110101111000101000101000100001011010011011011010001000101111001101001010001001110010101010111110

Take a step back.
Look at the bigger picture.

- Frank Underwood

Analytics Workflow

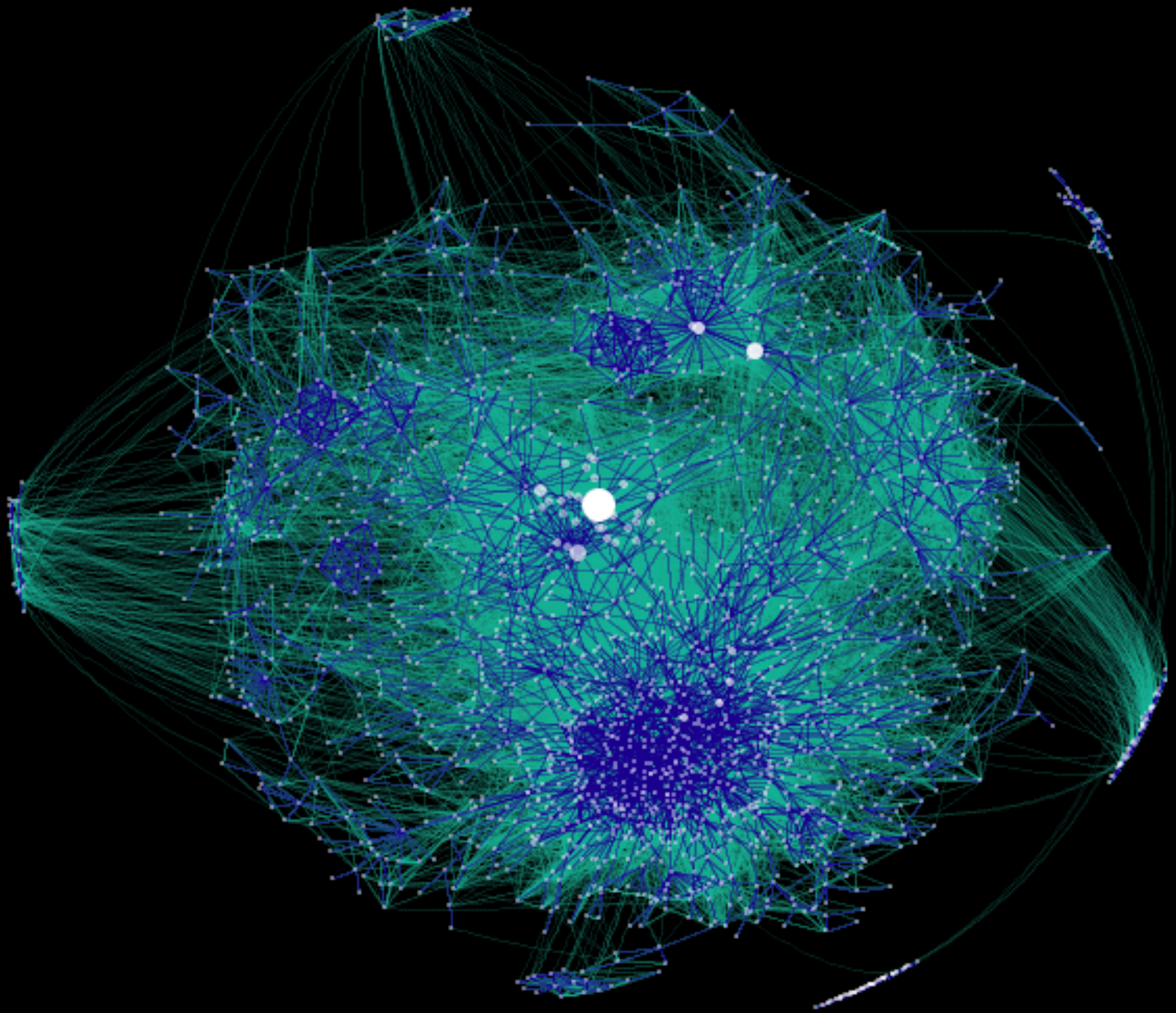


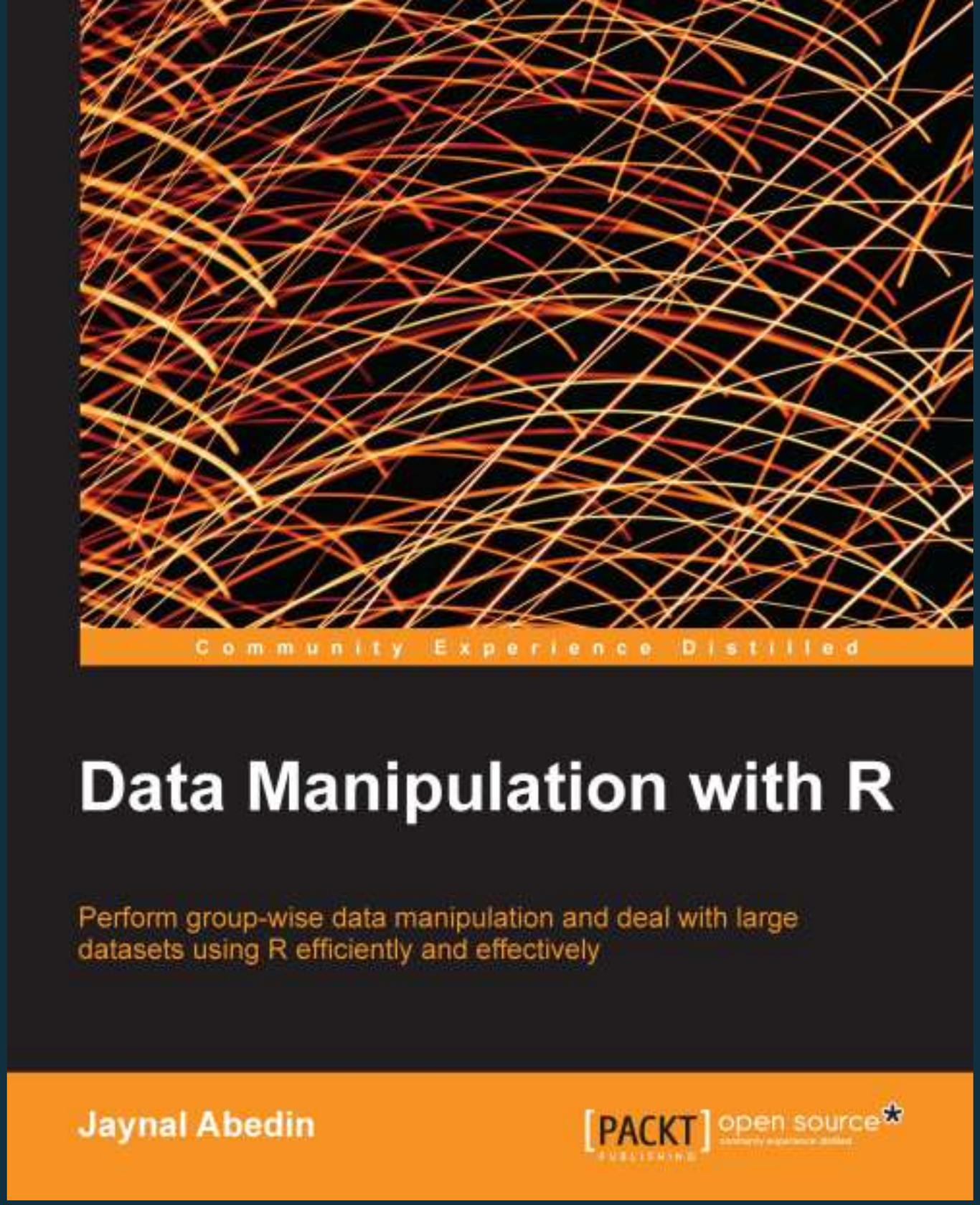
If you do not know how to ask the right question,
you discover nothing.

(W. Edwards Deming)

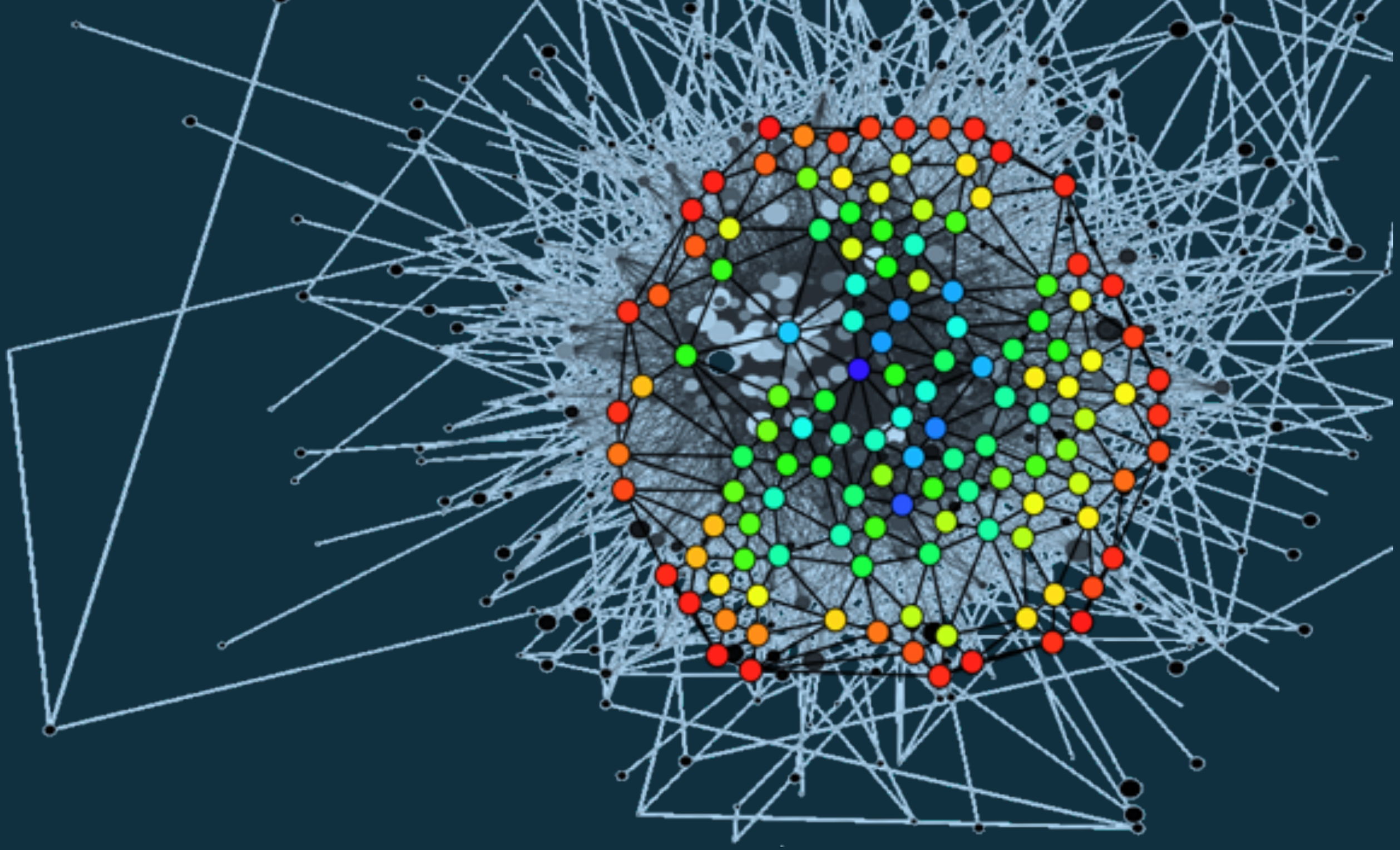
1. Ask the right question

2. Find the right data





3. Process the data



4. Model the data

5. Share the results

Analytics Workflow

1. Ask the right question
2. Find the right data
3. Process the data
4. Model the data
5. Share the results

ml =

representation + evaluation + optimization

Representation

We need to represent the data and model in a way computers can understand.

Evaluation

Separate good models from bad models

Optimization

A method to search among models.

Efficiently search for effective models.

Paradigms in ml

- Supervised Learning
- Unsupervised Learning

Supervised Learning

Infer a function from labeled data

Unsupervised Learning

Find hidden structure in unlabeled data

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Fit a line to the data

R code interlude

Pros & Cons

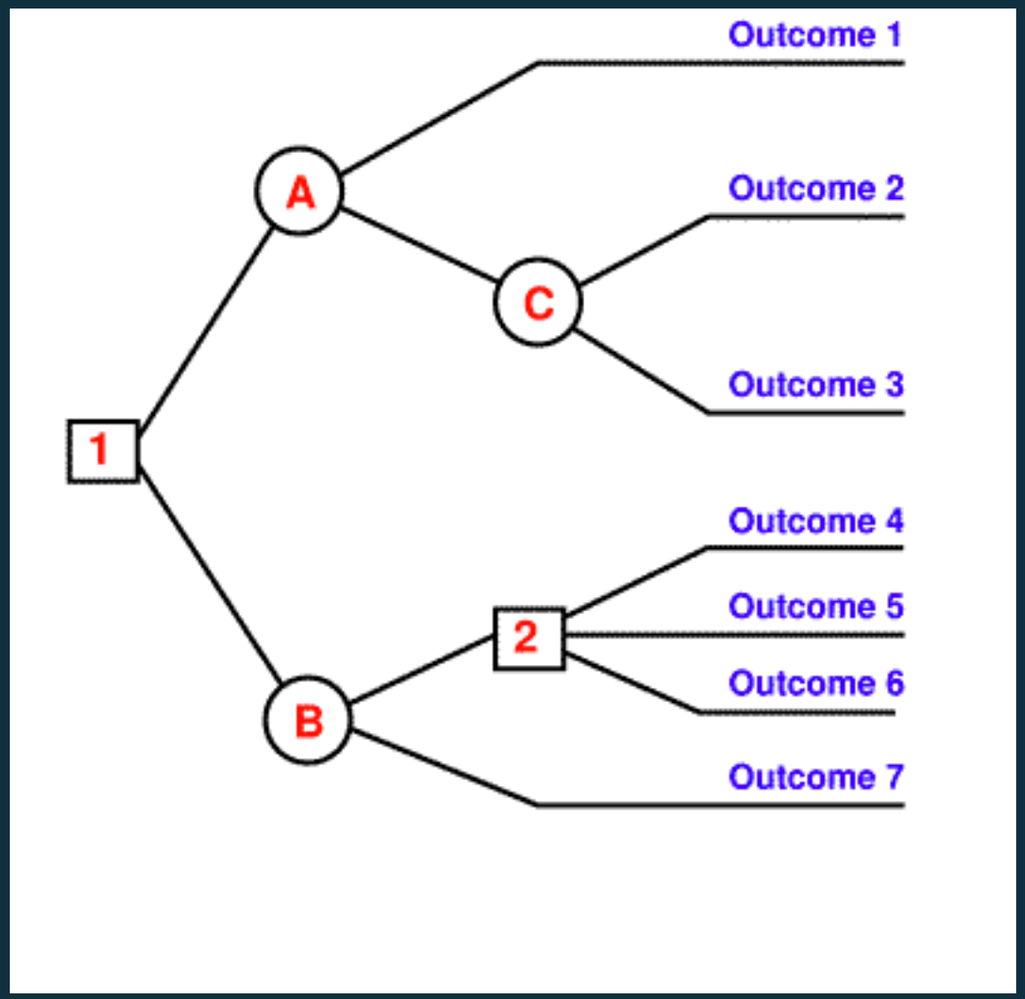
+ Easy to fit

+ Easy to interpret

- Life is nonlinear

- Limited to predicting numeric output

Decision Trees



A series of branches that divide the data

R code interlude

Pros & Cons

+ White box model

+ Can fit most types of data

- Fragile

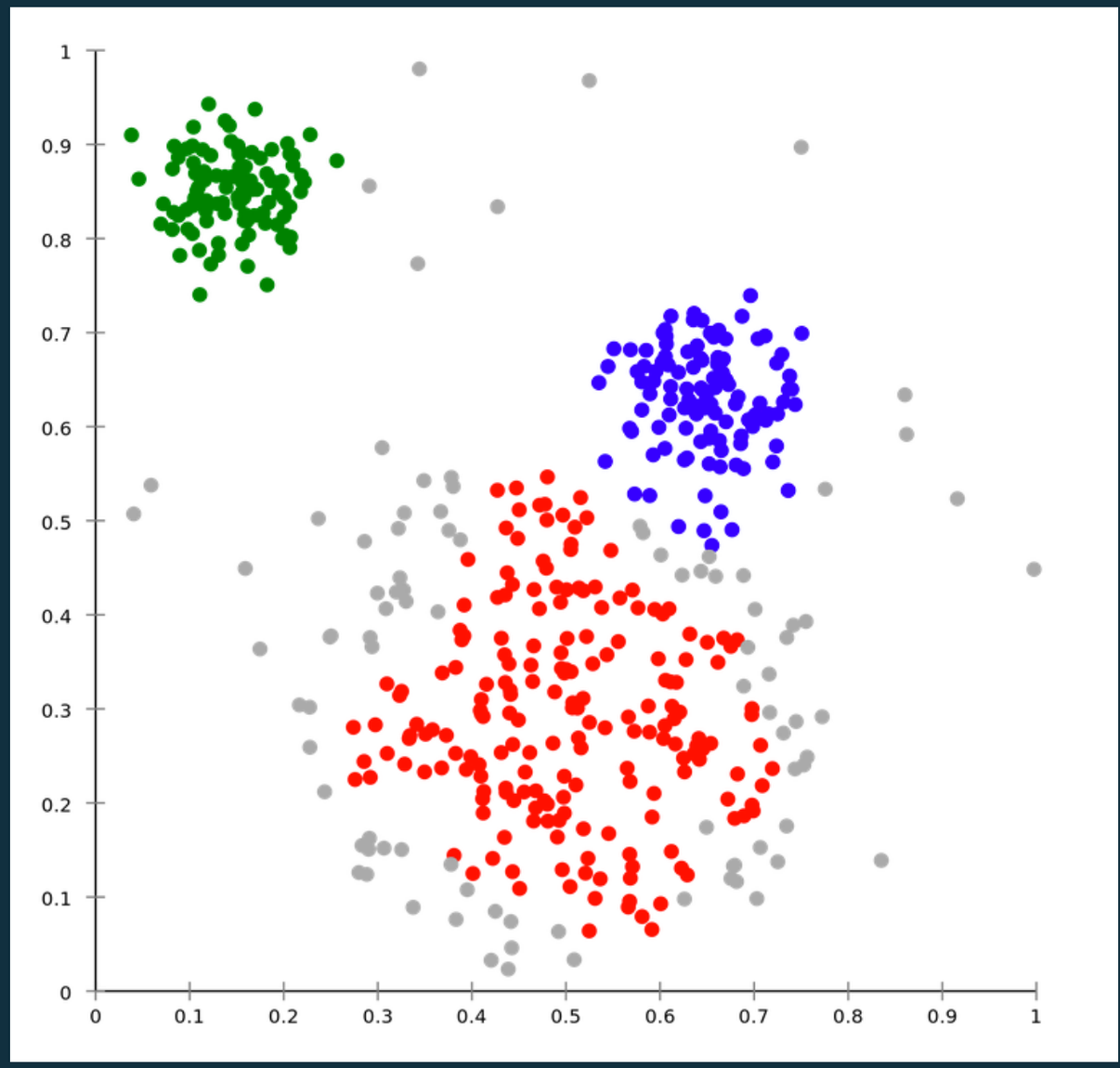
- Overfitting

- Out-of-sample prediction


www.PizzaAmoreWoodfire.com

Clustering

Grouping data so that data in the same group is more similar to each other than to data in the other groups



R code interlude

Pros & Cons

+ Powerful and common unsupervised model

+ Good for a quick insights into data

- Difficult to identify or define clusters

- Slow to fit

ROAP
(Random Acts of Pizza)

Suggested Resources: general ml

1. Data Mining and Analysis

2. Machine Learning: A Probabilistic Perspective

Suggested Resources: R

1. Try R from Code School
2. R in a Nutshell
3. The Art of R Programming

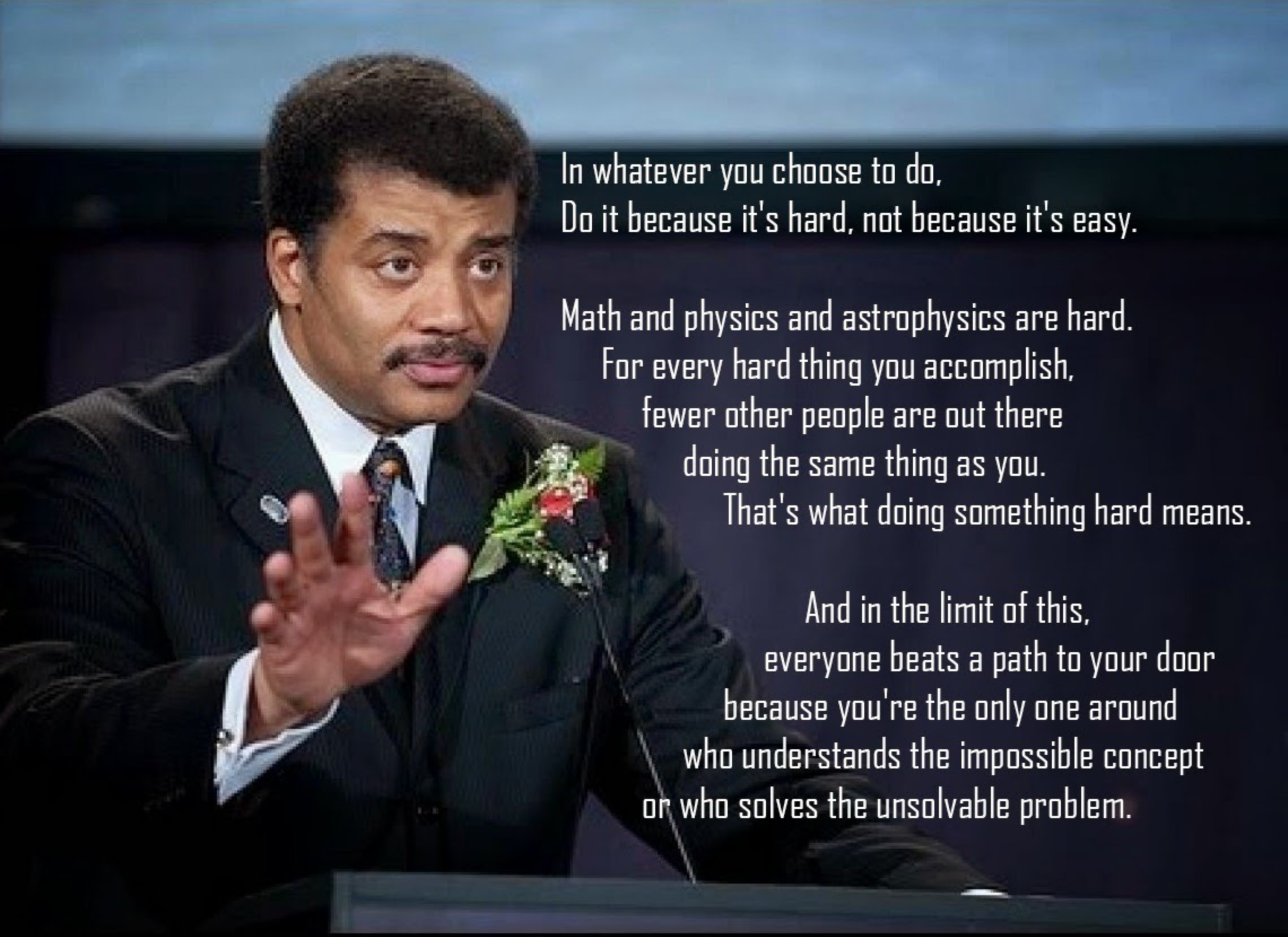
Suggested Resources: ml in R

- An Introduction to Statistical Learning: with Applications in R
  - Book
  - Video & Slides
- Applied Predictive Modeling

LiveCareer

is Hiring

drop me a line.


In whatever you choose to do,
Do it because it's hard, not because it's easy.
Math and physics and astrophysics are hard.
For every hard thing you accomplish,
fewer other people are out there
doing the same thing as you.
That's what doing something hard means.
And in the limit of this,
everyone beats a path to your door
because you're the only one around
who understands the impossible concept
or who solves the unsolvable problem.