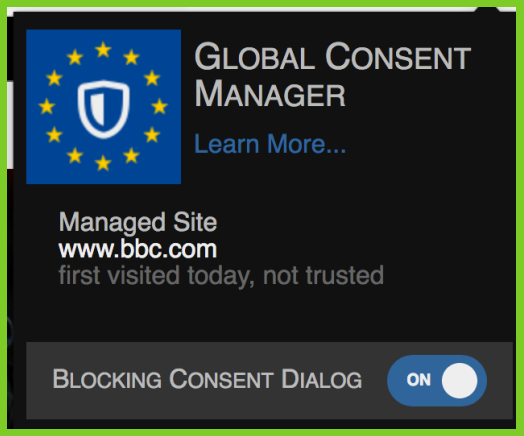
"What are you doing, Dave?"
--HAL9000, from 2001 A Space Odyssey

Social Computing, Artificial Intelligence and Machine Learning

augur

Mission hydrosci

CHAOSS



Sean P. Goggins, Phd.
Associate Professor, Computer Science
University of Missouri

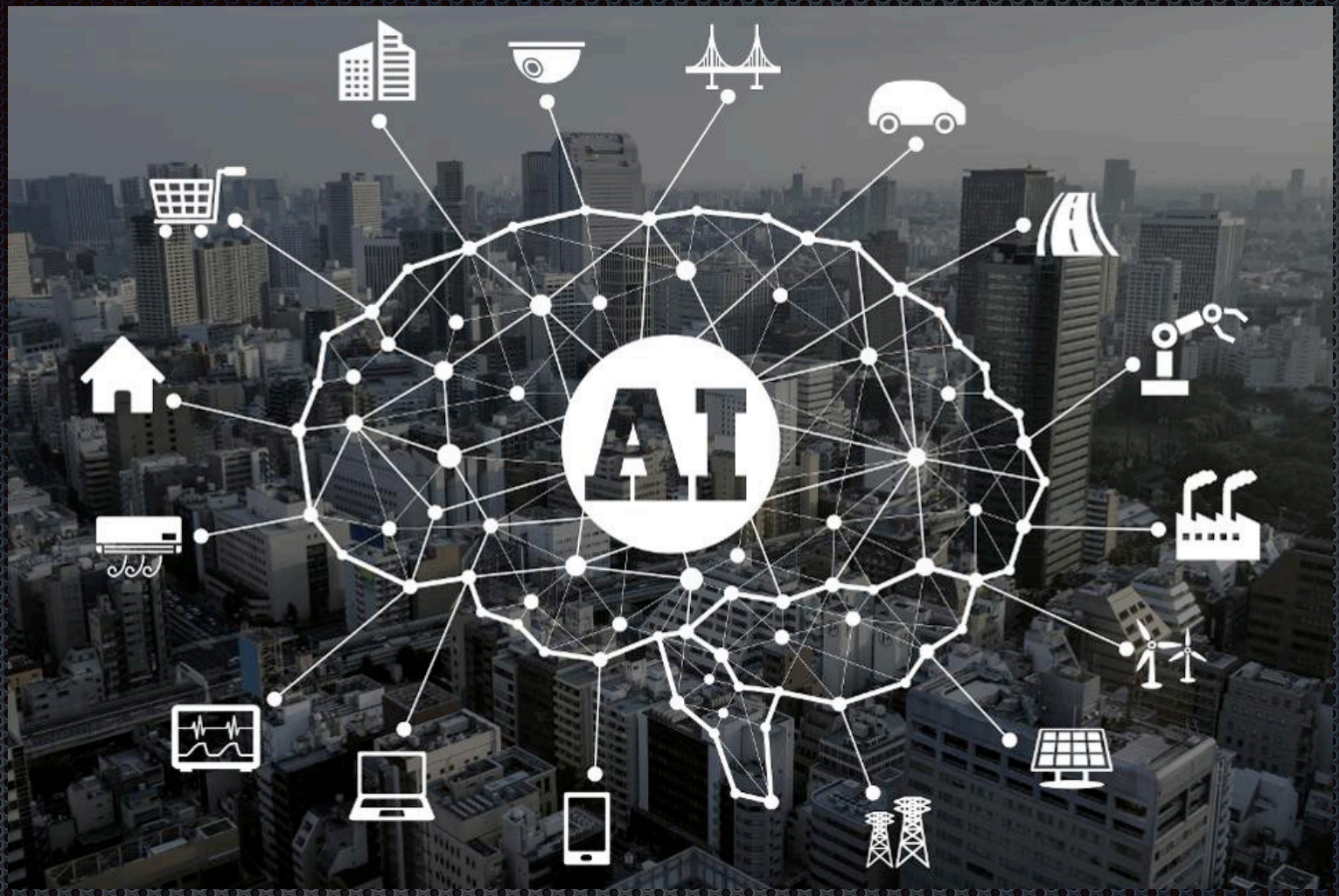
@SociallyCompute on Twitter
@ComputationalMystic on Instagram
@Sociotechnika on Flickr
@sgoggins on GitHub

October 23, 2018

http://www.seangoggins.net

Being a Brewers fan involves love without expectation. So, when they do well and don't win it all, its especially hard.

Low Expectations meet "being amazed"



Low Expectations meet “being amazed”

“Joe Buck doesn't show any enthusiasm for his job even while simultaneously exhibiting clear biases against certain athletes and teams. Buck also seems to feel as though that everyone watching the games that he talks over wants to hear what his opinion on issues that affect him personally such as times when athletes have snubbed him or when someone pretending to moon a crowd of Packers fans offends him.”

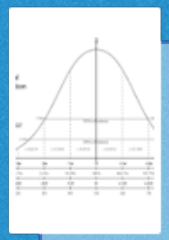
--Business Insider, on Why Buck is the Worst Sportscaster on Earth

Brewers fans want *“RI” in Sportscasting

*“RI”: “Real Intelligence”

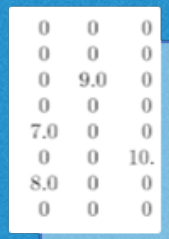
Data Science Algorithms

Intrinsic Data Quality



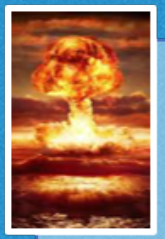
Statistics

- Organized
- Complete



Data Mining

- Semi-Structured
- Sparse Matrices

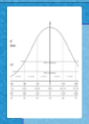


Machine Learning

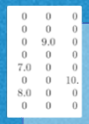
- Unstructured
- Unbounded

Data Science

Intrinsic Data Quality



Statistics
- Organized
- Complete



Data Mining
- Semi-Structured
- Sparse Matrices

Machine Learning
- Unstructured
- Unbounded

Algorithms

Algorithms (the rest)

Data Carpentry
80% - 90% of Work Hours

So, is ”AI” the Photoshop of Social Computing?

AI and Machine Learning Two Perspectives

ML is a Subset of AI

- Two Types of AI: Applied and Generalized (which only exists in theory)
- AI research led to the development of Machine Learning technologies

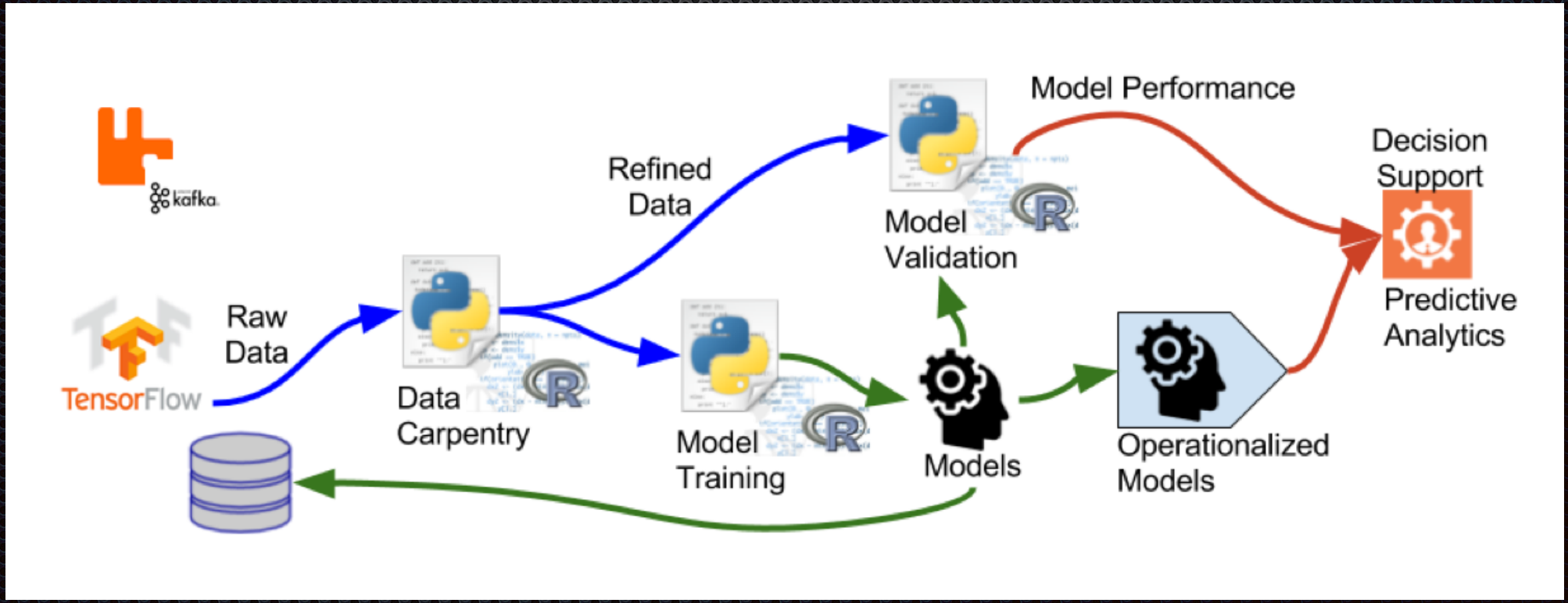
AI is A Subset of ML

- Supervised Machine Learning: Where human coded data is used to automate repetitive data analysis tasks
- Unsupervised Machine Learning (“AI”): Given a set of data and results, and possibly feedback, these applications operate with little or no explicit “training”

“AI Algorithms

- Neural Networks (Supervised)
- Genetic Algorithms (Unsupervised)
- Reinforcement Learning (Unsupervised)

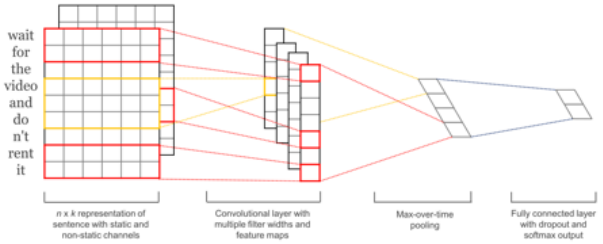
Neural Networks



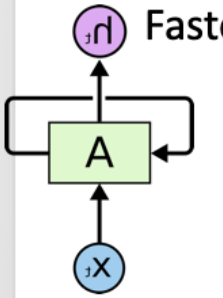
Neural Networks

Neural Network Algorithms Scully and Mulder

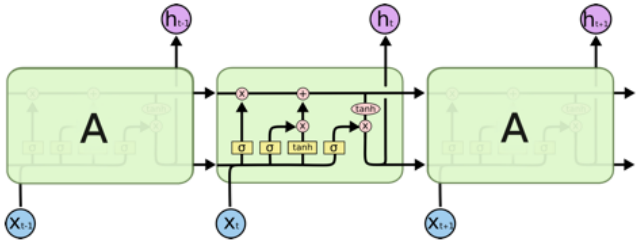
- Convolutional Neural Network (Mulder):
  - Use trigrams through 5-grams. Is a less sophisticated language processing approach
  - But they work faster with a trained data set
  - Generally used for image processing
    - But have found reasonable effectiveness with language tasks
    - Which is why, in this example, we use both



- Gated, Recurrent Neural Network (gru or LSTM) (Scully):
  - Considers all words in a phrase from left to right. More natural speech oriented
    - Listening to words as they occur in a sequence of text
    - From beginning of a phrase to the end
  - Faster to a trained result, slower to run



Recurrent neural networks have loops



But, as they get larger, the distance between nodes of Learned information is greater. LTSM manages this problem By systematically deciding what information to throw away

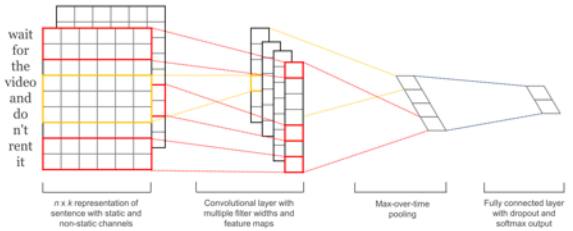
Neural Networks

Neural Network Algorithms Scully and Mulder

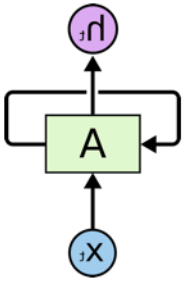
python train_mulder.py --data_file data/haterz.csv --embedding_file None --num_classes 3 --embed_dim 100

Convolutional

mulder specific parameters:
•num_layers : The number of fully connected convolution layers in the network.
•filter_sizes : Size of the convolution filer windows.
•num_filters : Number of filters.
•l2_regulizer_lambda : L2 regulizer lambda value.
•activation_function : Activation function for connected layers.



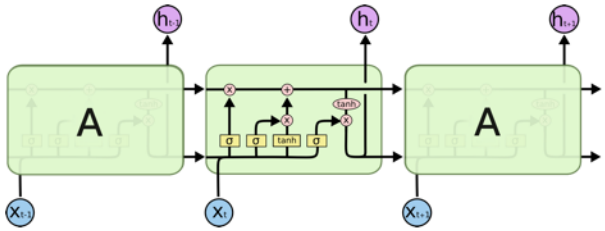
python train_scully.py --data_file data/haterz.csv --embedding_file None --num_classes 3 --embed_dim 100



Gated, Recurrent

scully specific parameters:
•hidden_dim : Number of units in the hidden dimension.
•num_layers : The number of gated cells in the network.
•cell_type : The type of cell between the options of "LSTM" and "GRU".

Recurrent neural networks have loops



But, as they get larger, the distance between nodes of Learned information is greater. LTSM manages this problem By systematically deciding what information to throw away

Neural Networks (Supervised)

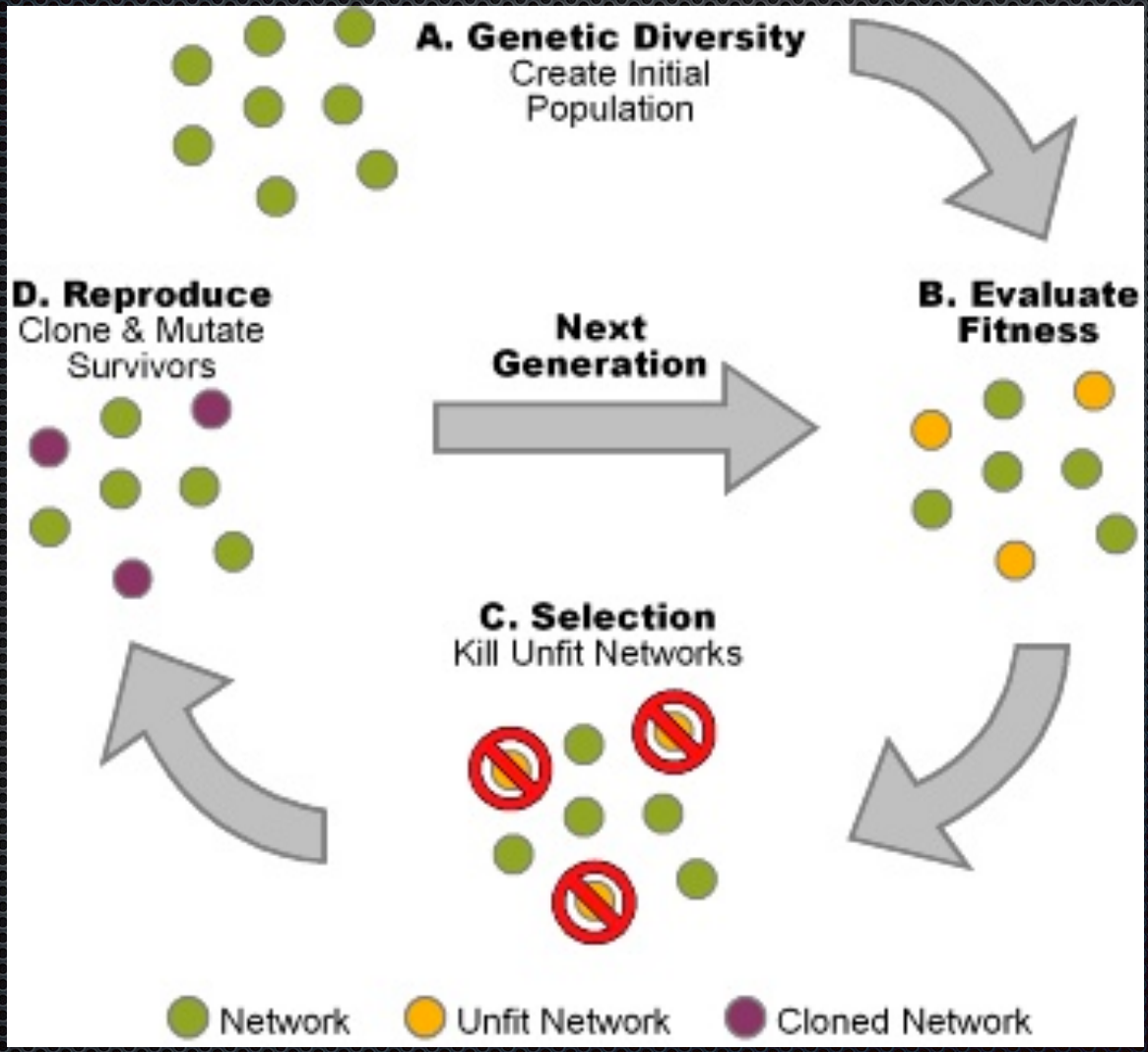
1. Identify Data to Gather: Where is the Signal?
    - Twitter
    - Geographic Streaming, 50 Cities
2. What are the signals you are looking for
    - Changes in Sentiment About Immigrants
    - Racial Slurs
3. Data Cleaning and Shaping
    - Identify Slur Language; bigrams, trigrams
        - Examples:

    |%eight%|%ball%|%eskimo%|%eyetie%|%flip%|%fob%|%fritz%|%gable%|%gaijin%|%外人%|%gin%|%gin%|%jockey%

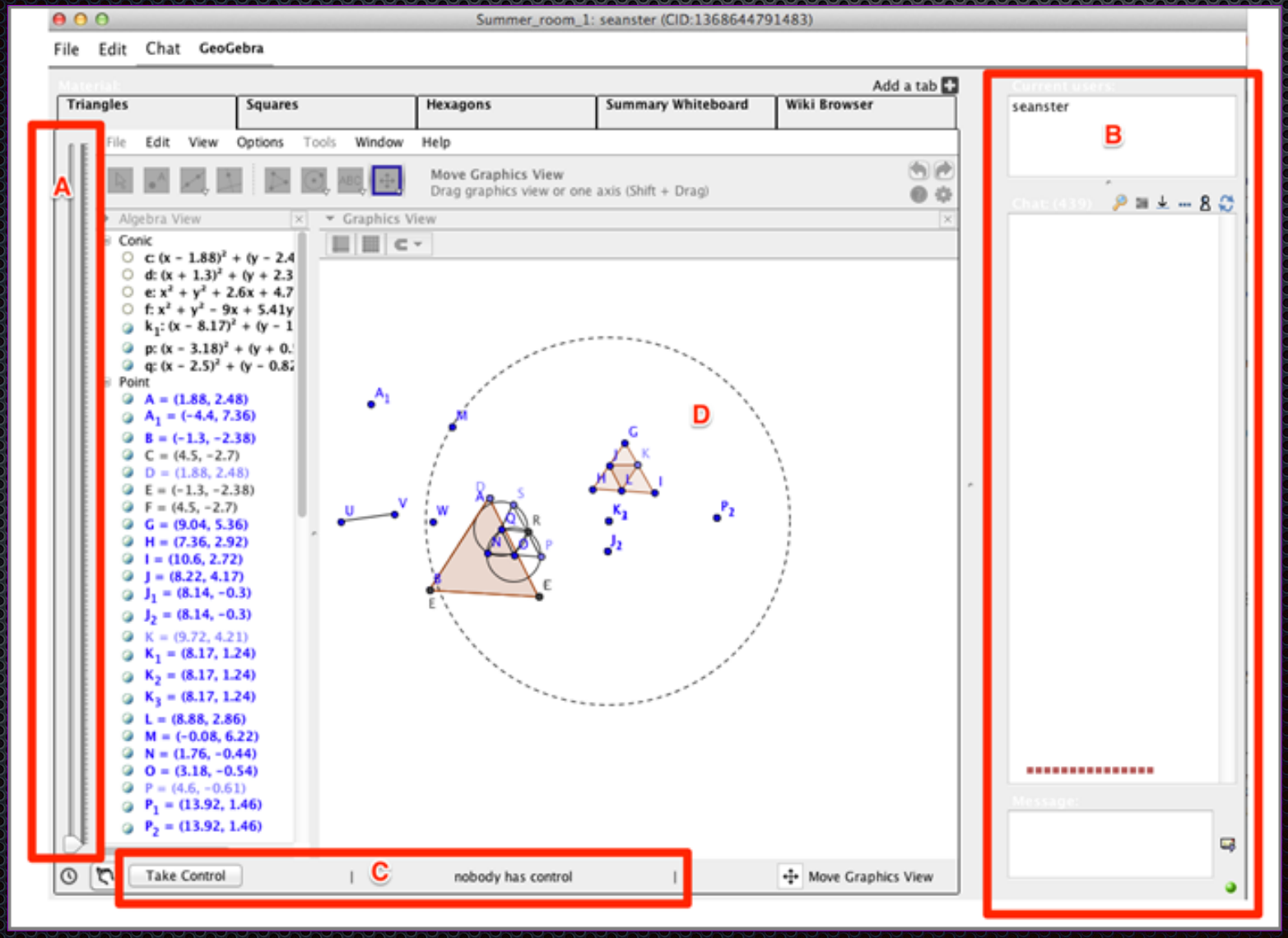
    - Iterate, eliminating slur words that are used in other ways
        - Example: SELECT text_clean
        FROM public.slurs
        WHERE LOWER(text_clean) ~* '\y(alien|hajji|hadji|haji|kebab|towel head)s?\y'
        -

Genetic Algorithms

AI

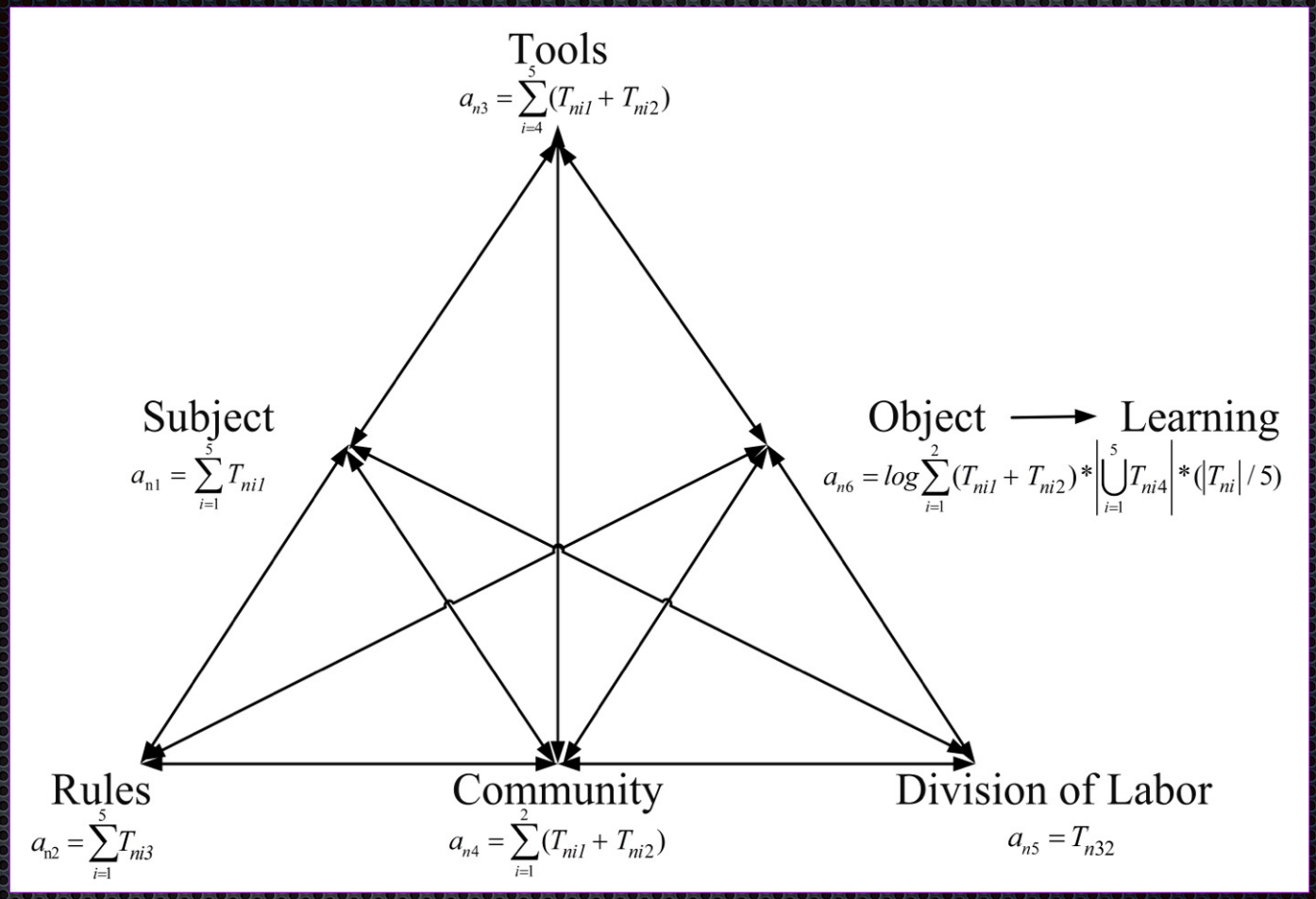


Genetic Algorithms



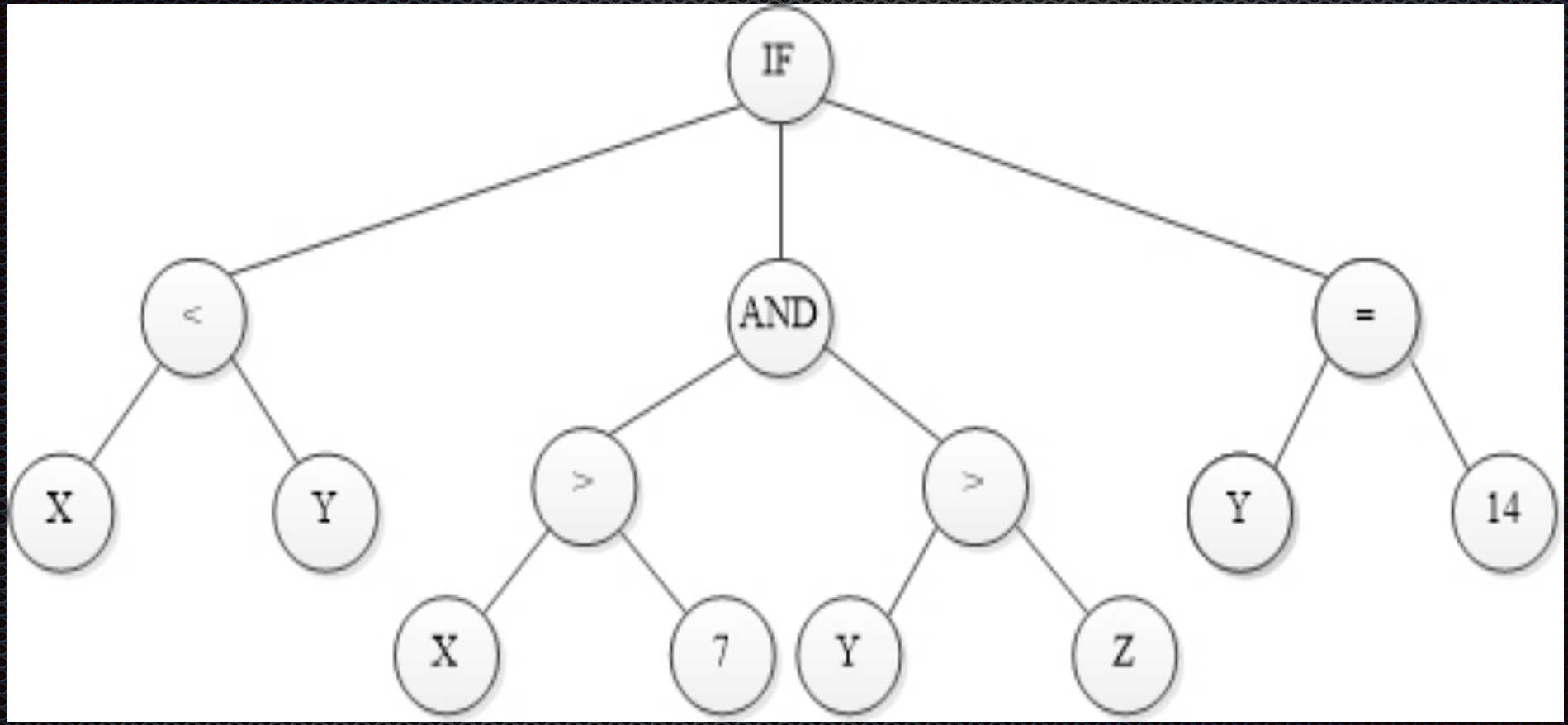
Genetic Algorithms

AI



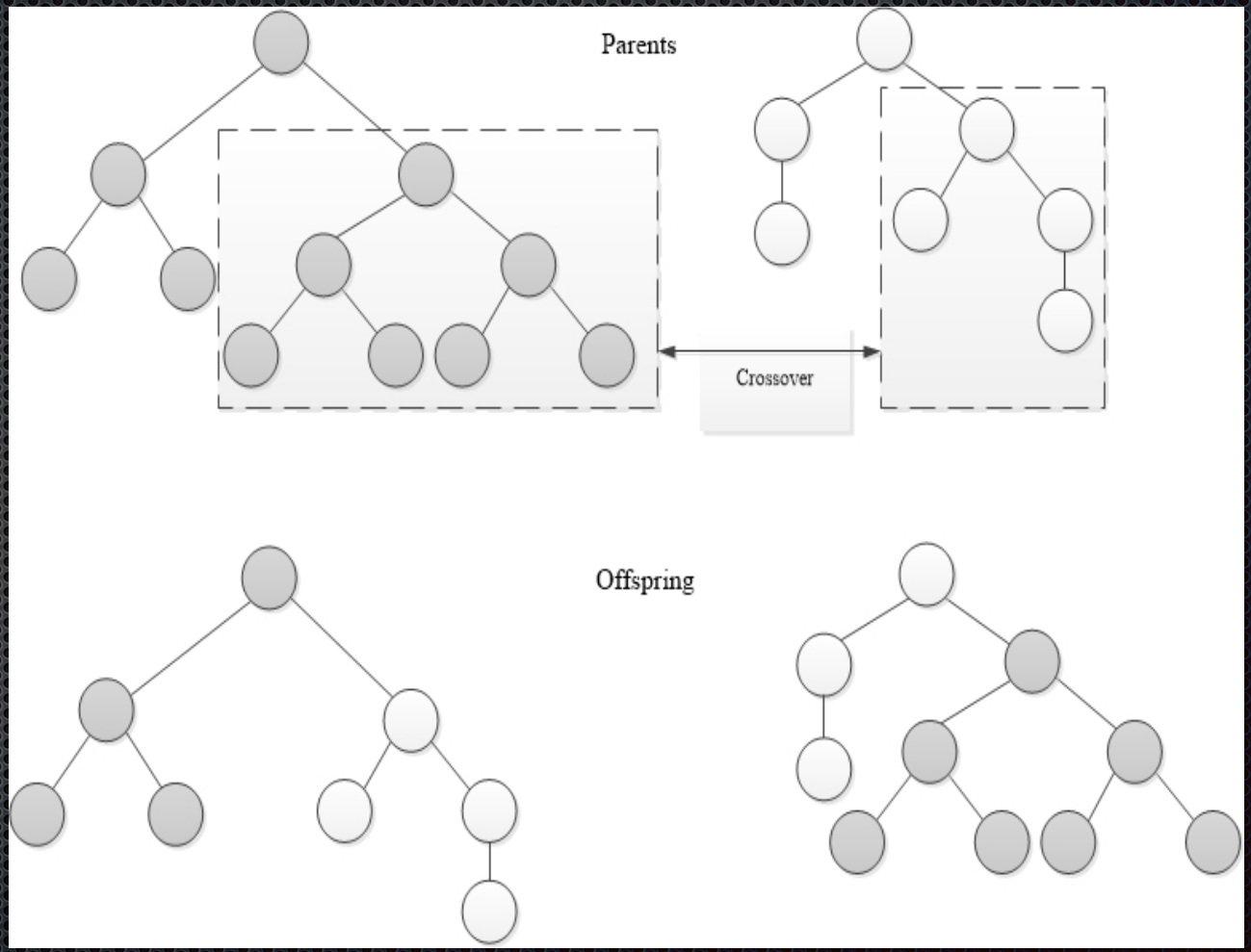
Activity theory quantification model for individual participation in CSCL. a_n represents the activity vector of student n. In the event type level, T_{ni} denotes the event type *i* of student n, meaning that there are five event types in total. Specifically, these five event types are *Awareness* (*i*=1), *Chat* (*i*=2), *Geogebra* (*i*=3), *System* (*i*=4), and *Wb* (*i*=5). In the measurement level, T_{nij} represents the value of measurement aspects *j* in event type *i* of student n, denoting four measurement aspects, *Individual* (*j*=1), *Group* (*j*=2), *Action Types* (*j*=3) and *Module Set* (*j*=4).

Genetic Algorithms



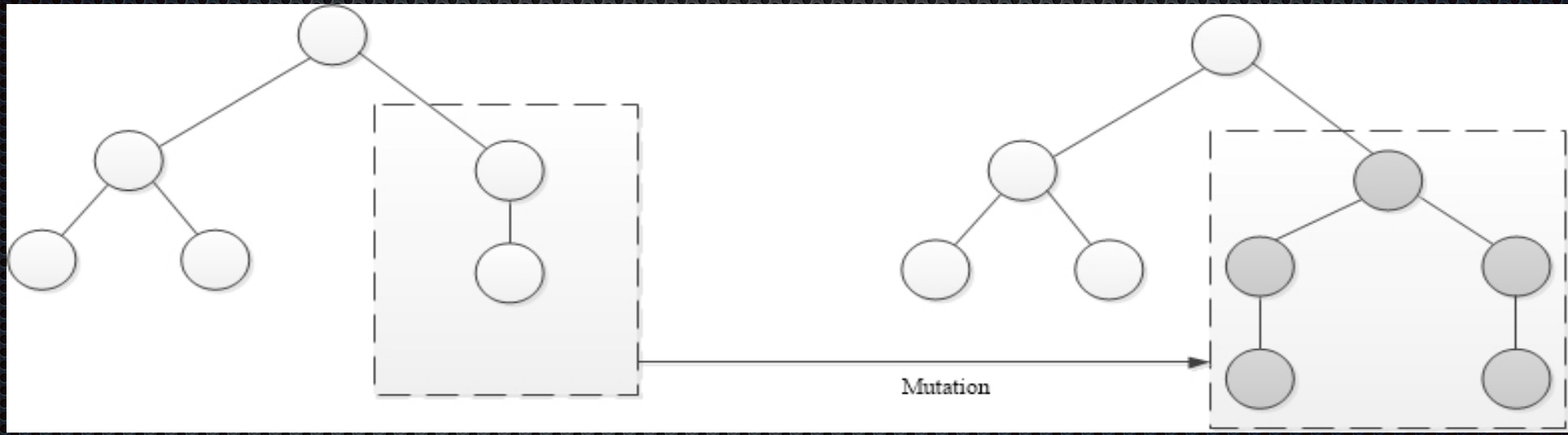
Genetic Algorithms

Crossover
Selection
Mutation



Genetic Algorithms

Mutation



Genetic Algorithms

<Rule> :: =
IF <antecedent> THEN <consequent>
<antecedent>:: =
<condition> AND <condition>
<condition> | <condition>
<consequent>:: =
IS A <class label>
<condition>:: =
    <attribute> <rel operator> <value>
<attribute>:: =
    <Subject> <Rules> <Tools> <Community> <Division of Labor> <Object>
<rel operator>:: =
    = | ≠ | > | ≥ | < | ≤
<value>:: =
    Value in each corresponding domain
<class label>:: =
    EXCELLENT|GOOD|AVERAGE|SUFFICIENT | FAIL

Genetic Algorithms

AI

Table 2
Confusion matrix.

Actual \ Predict	Positiv
Positive	True posi A
Negative	False nega C

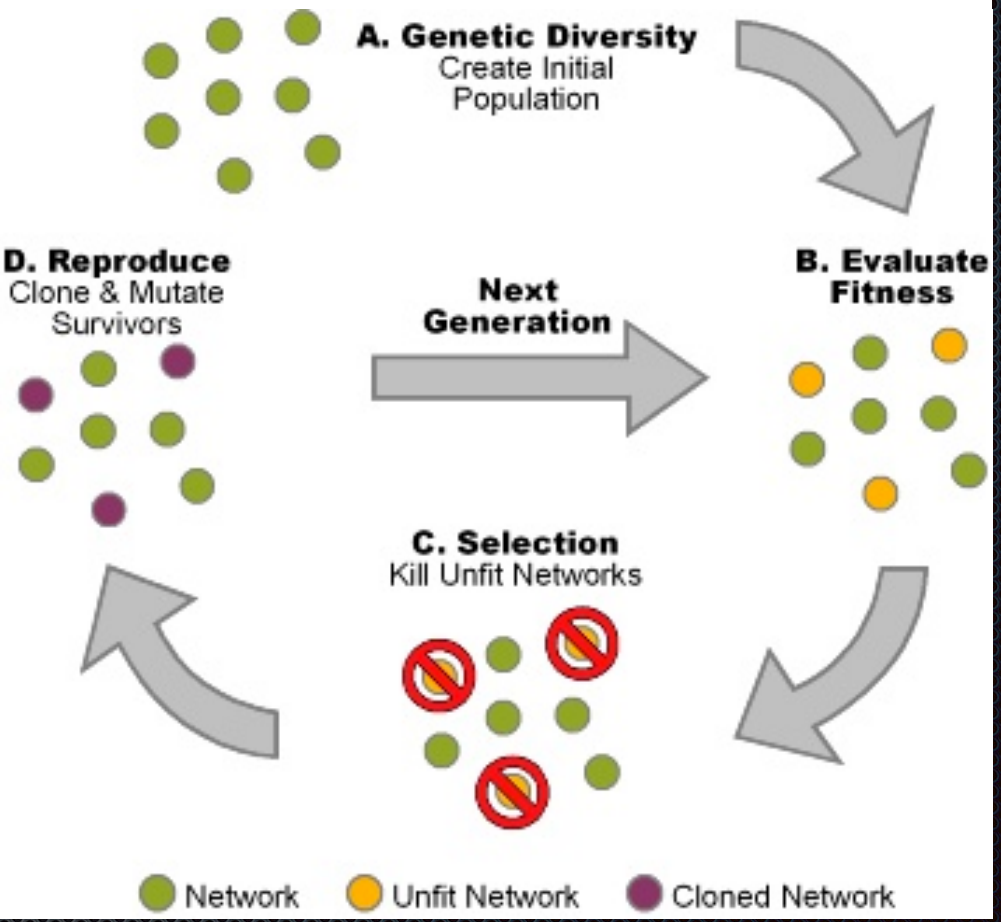
In Table 2, A is the number of correct predictior
predictions that an instance is negative; C is the number o
the number of correct predications that an instance is nega

Sensitivity is the proportion of actual positives w

Sensitiv

Specificity is the proportion of actual negatives w

Specific

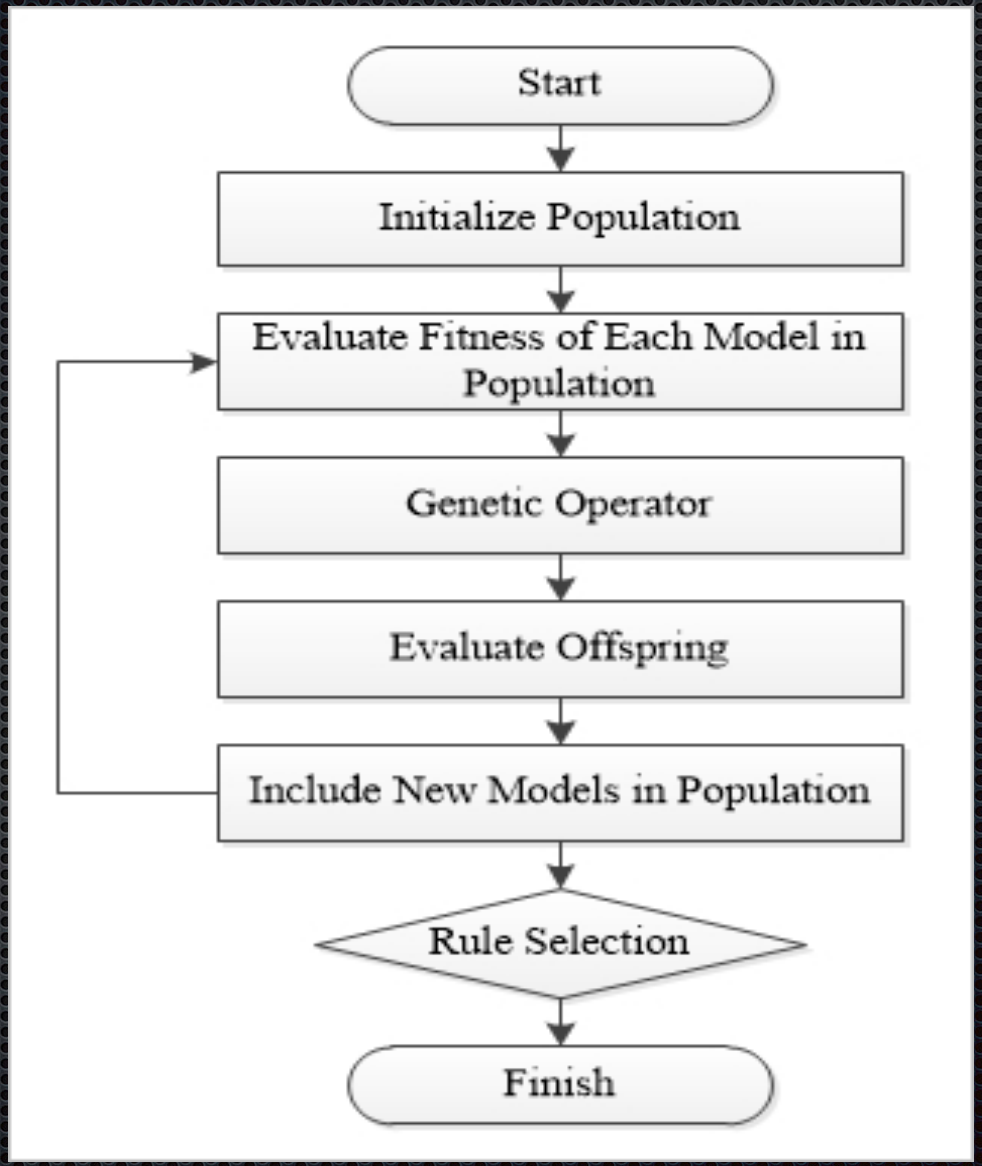


Then in order to maximize the accuracy and prevent problems associated with imbalanced data, the fitness function is calculated as the product of sensitivity and specificity.

$$Fitness = \frac{A * D}{(A+C)*(B+D)}.$$

Genetic Algorithms

AI



Genetic Algorithms

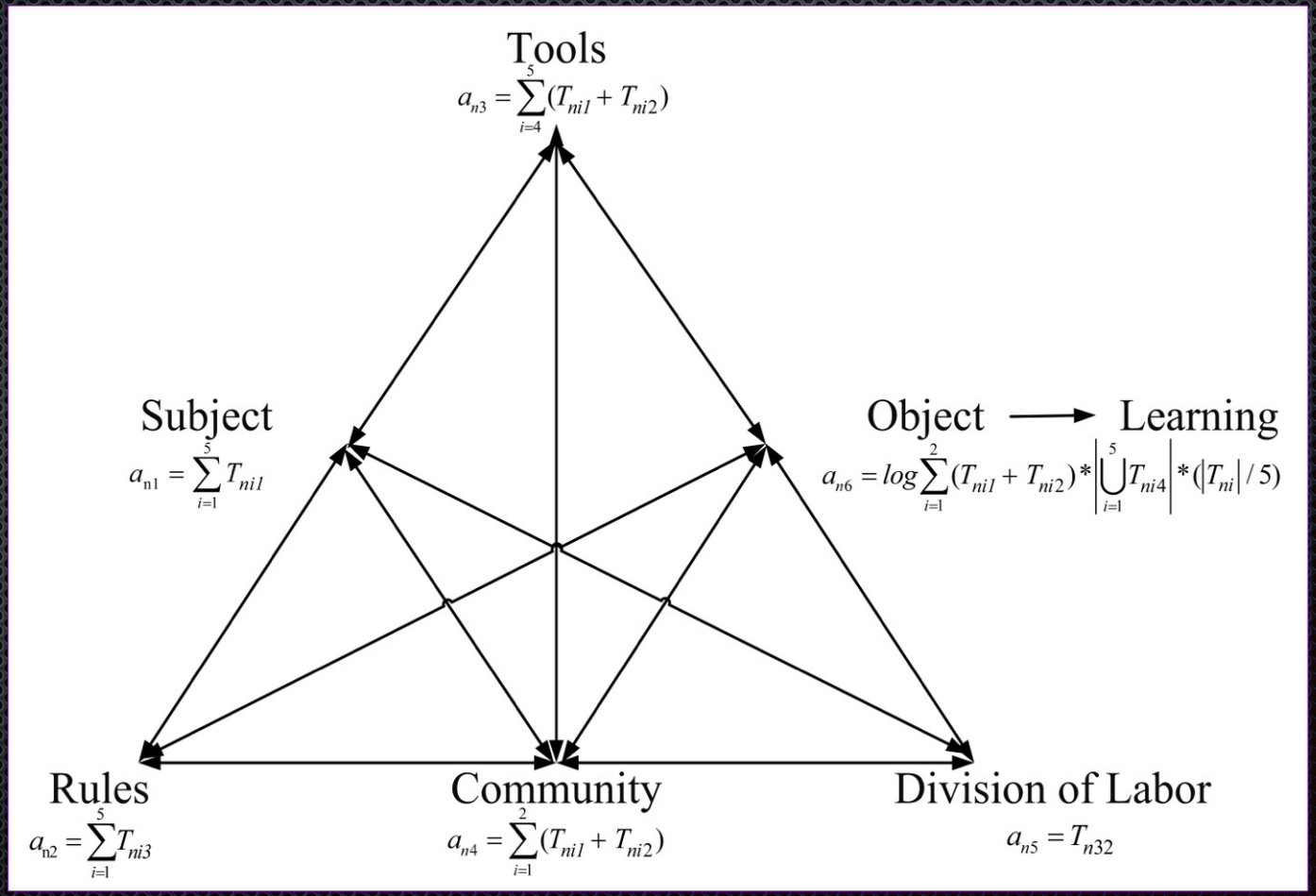
AI

Result \ Algorithm	Prediction Result	Easy to Understand Model			Difficult to Understand Model		
		Rule-based Model		Decision-Tree	Statistical Model	Artificial Neural Network	Bayesian Network
		GP-ICRM	NNge	RandomTree	Logistic Regression	Perceptron	Naïve Bayes
Fitness	Overall Prediction	80.2%	76.2% **(.000)**	72.6% **(.000)**	77.1% **(.000)**	36.6% **(.000)**	77.7% **(.000)**
	At-Risk Prediction	89.5%	82.1% **(.000)**	82.1% **(.000)**	81.1% **(.000)**	42.1% **(.000)**	94.7% (.784)
Sensitivity	Overall Prediction	80.3%	76.8% **(.000)**	72.7% **(.000)**	77.2% **(.000)**	38.9% **(.000)**	78.2% (.018)
	At-Risk Prediction	85.0%	76.2% **(.000)**	76.2% **(.000)**	78.9% **(.000)**	66.7% **(.000)**	90.0% **(.000)**
Specificity	Overall Prediction	80.3%	76.2% **(.000)**	73.0% **(.000)**	77.0% **(.000)**	36.4% **(.000)**	77.9% **(.000)**
	At-Risk Prediction	94.4%	88.9% (.028)	88.9% **(.000)**	83.3% **(.000)**	30.8% **(.000)**	100% **(.000)**

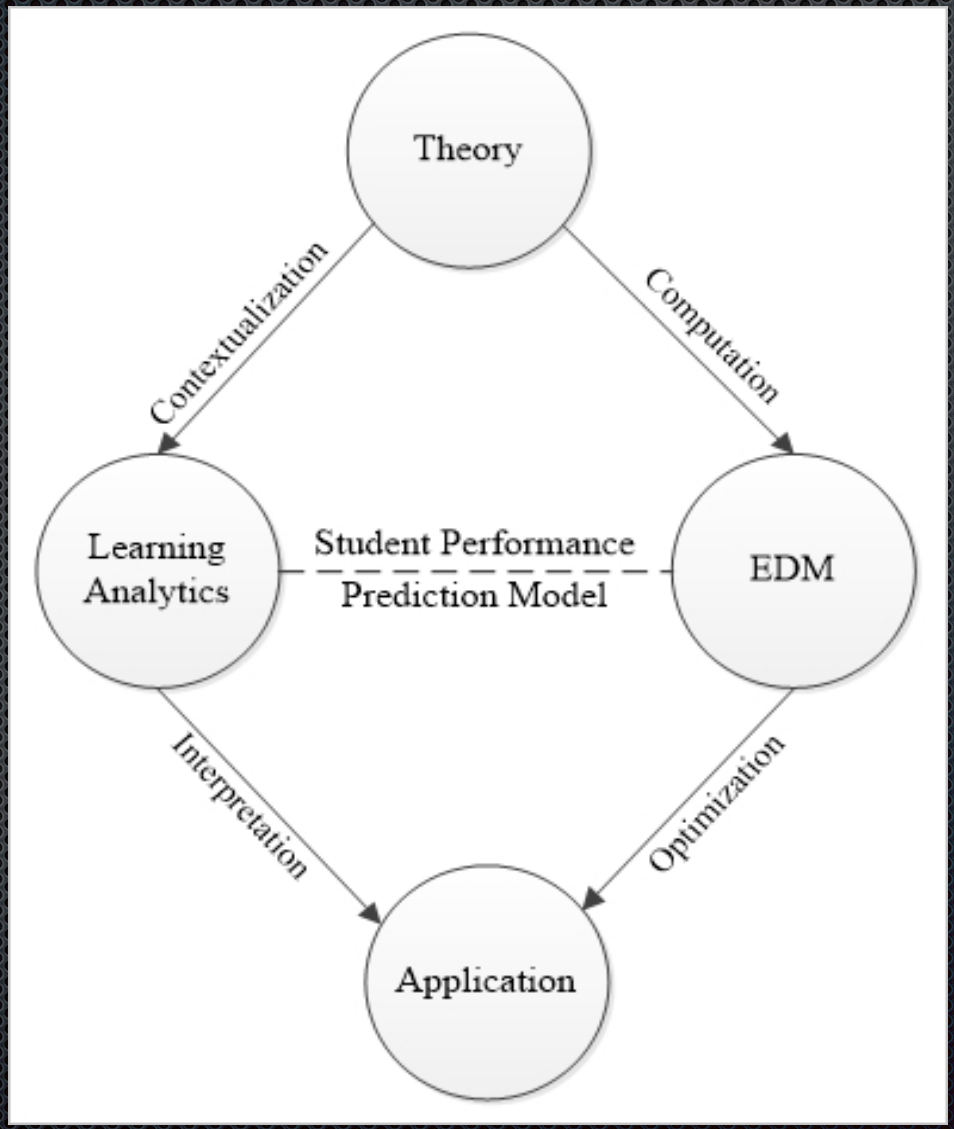
Genetic Algorithms

AI

There's a subtle piece to why the paper this (part of the talk) is based on is my most cited … It applies genetic algorithms using theory



Genetic Algorithms



AI To this Moment

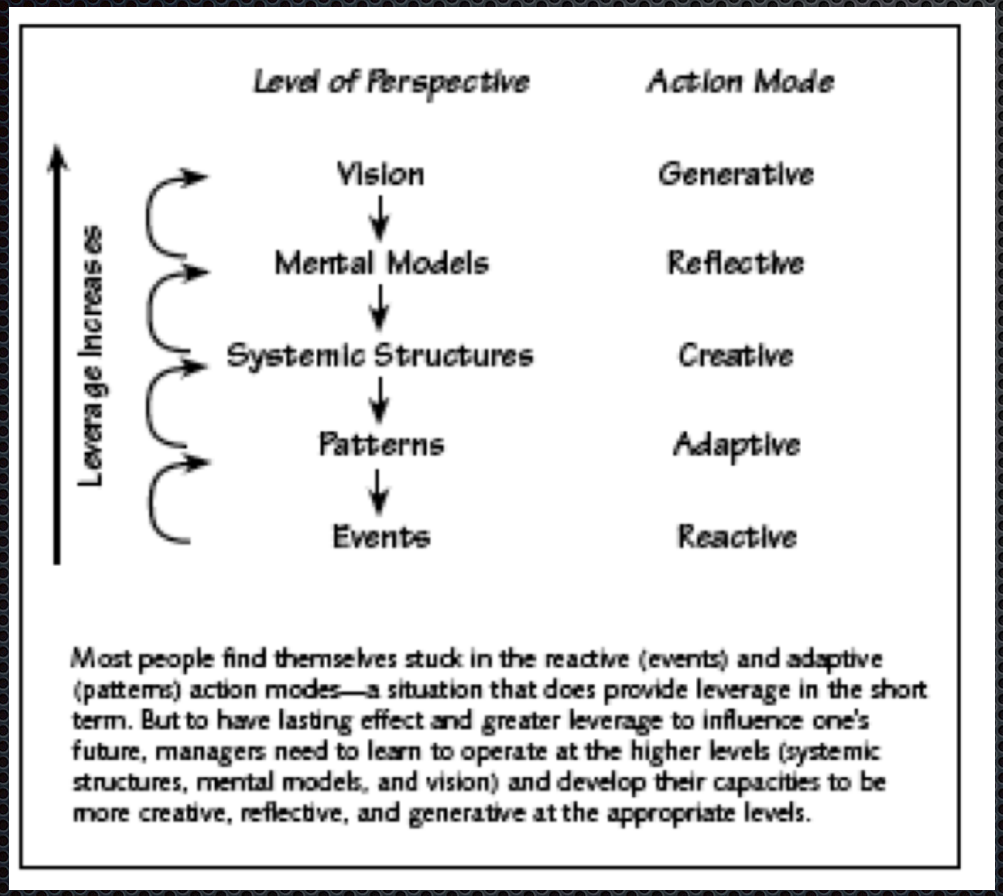
- Three Algorithms: Neural Networks and Genetic Covered.
- Reinforcement Learning Remains
- Why are we learning about this?

An Interlude: Our Obsession with Data.

There's this little thing of Data becoming part of what we all do

But, Unless the Data Informs, its Useless. And Data Exists to Help Tell a Story.

Here's How (Leading to "Who to Trust")



Most people find themselves stuck in the reactive (events) and adaptive (patterns) action modes—a situation that does provide leverage in the short term. But to have lasting effect and greater leverage to influence one's future, managers need to learn to operate at the higher levels (systemic structures, mental models, and vision) and develop their capacities to be more creative, reflective, and generative at the appropriate levels.

Systems Thinking

What is a Data Futurist?

Black Mirror & The X-Files

What is a Data Futurist?

Black Mirror & The X-Files

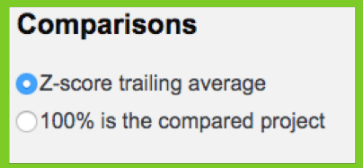
Engagement

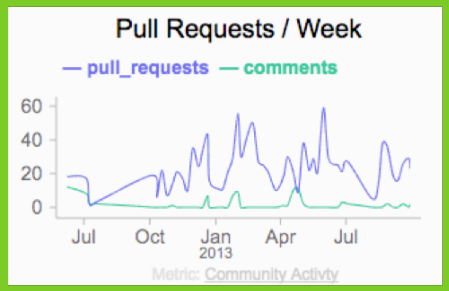
Organizational Health

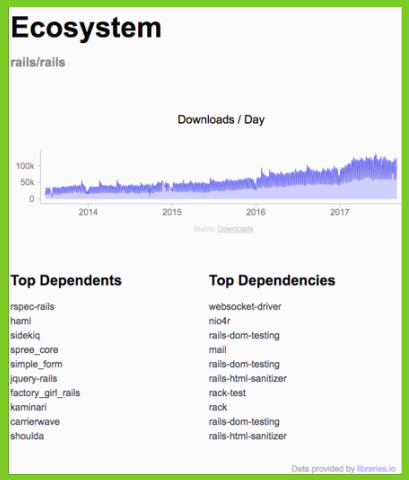
Augur.Software

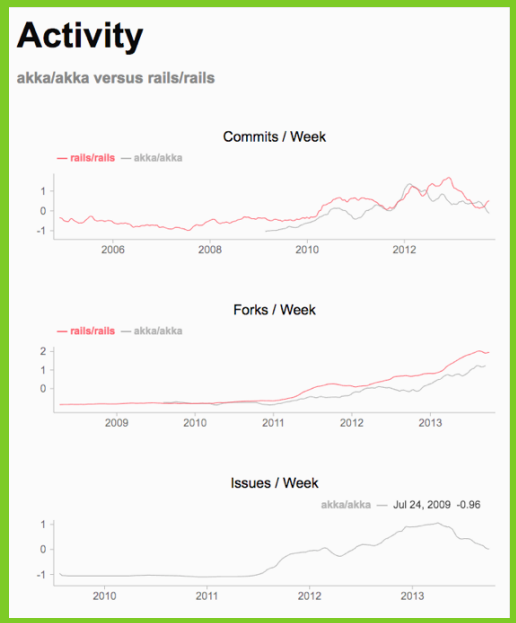
Prototyping Human Centered Metrics:

Enable Comparisons, Make Trends Central to the Use Experience

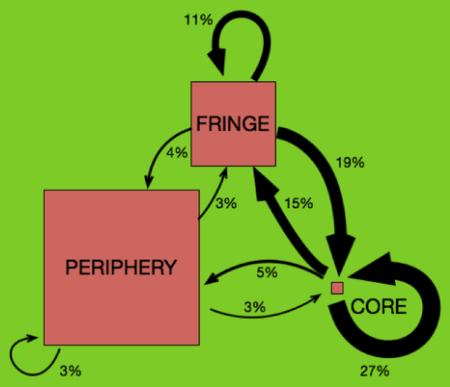




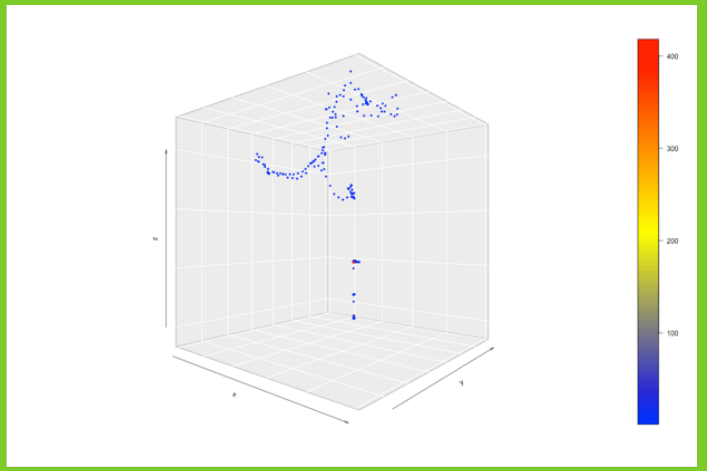




https://github.com/OSSHealth/ghdata



Who Does What with Your Data?



Mission Hydro Sci Teacher Dashboard

All Classes

Time Grid - Progress Chart

Quest Times

Search:

player ID	Check in with ARF	Talk to Toppo	Talk to Jasper	What just happened?	Find the Team	Settling WAT247	The Largest Watershed
alexy	00:33:24	00:35:39	00:38:23	00:39:20	00:44:12	00:52:18	00:54:04
alexy	00:06:21	00:09:36	00:11:42	00:12:49	00:23:12	00:27:52	00:34:26
alla	00:04:04	00:10:08	00:14:17	00:15:15	00:29:04	00:47:45	00:48:33
alla	00:48:56	00:51:03	00:53:16	00:53:34	00:10:07	00:22:16	00:22:36
allan	00:03:18	00:11:59	00:15:29	00:17:05	00:25:50	01:11:40	01:16:09

Data and Algorithms

Warning: Every, single choice has cascading impact

“AI Algorithms

- ~~Neural Networks (Supervised)~~
- ~~Genetic Algorithms (Unsupervised)~~
- Reinforcement Learning (Unsupervised)

Reinforcement Learning

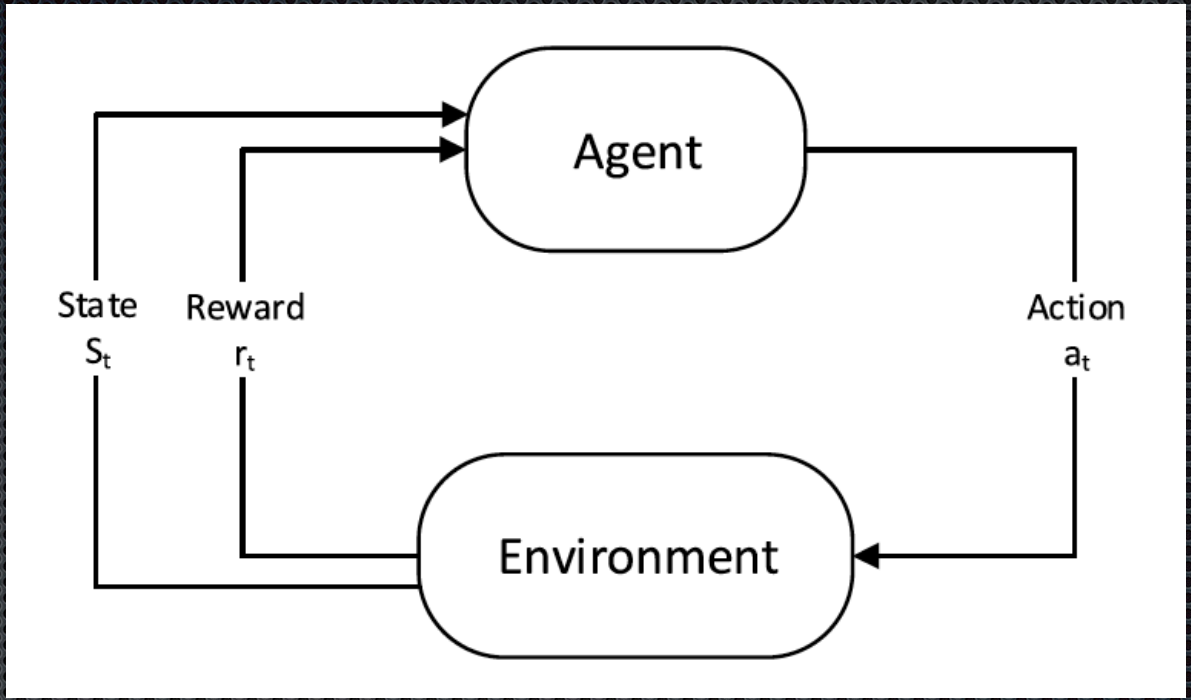
Algorithm learns to react to its environment

Closest thing to “AI”

Some view it as “beyond” unsupervised ML

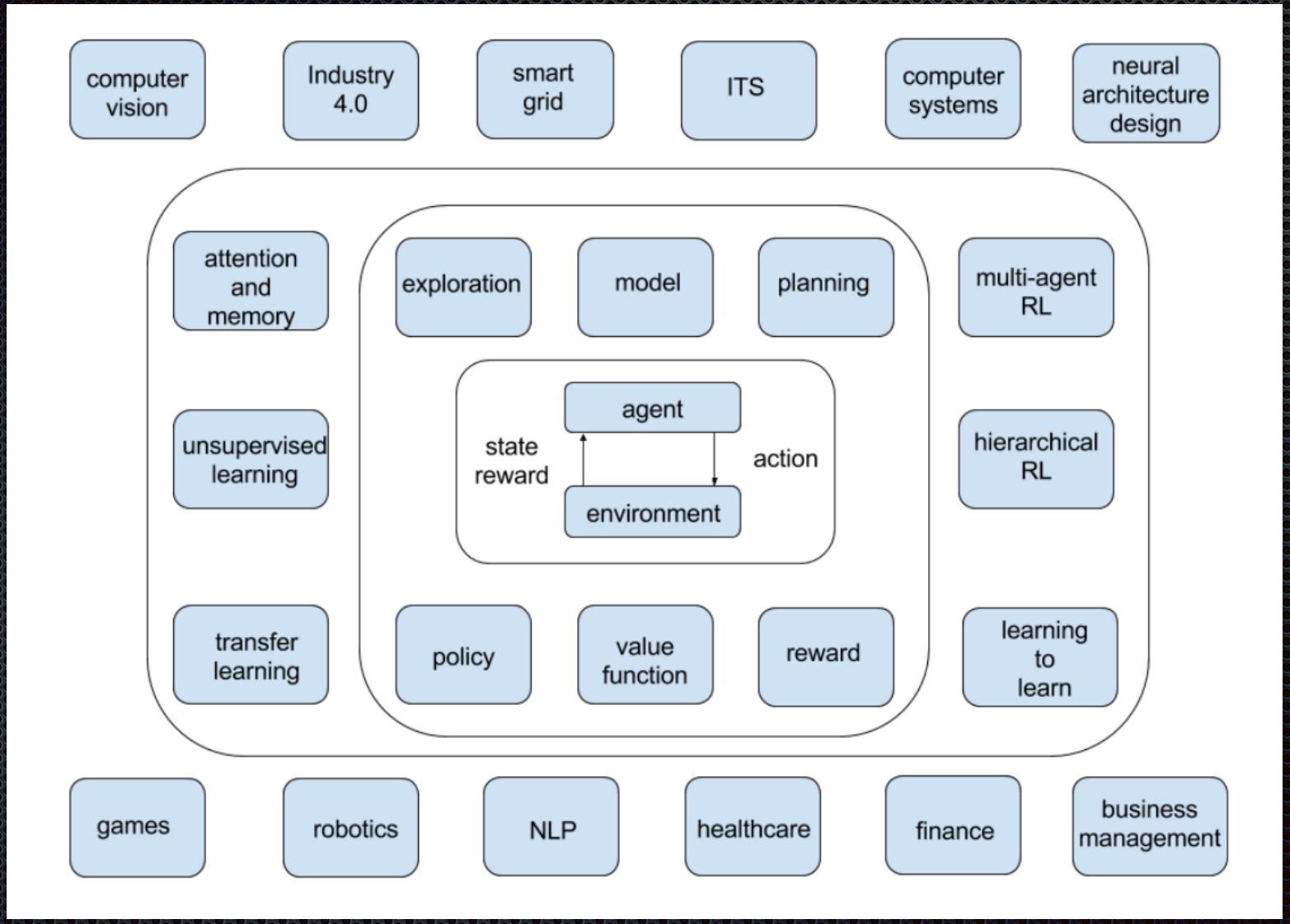
Agent Based Modeling

Reinforcement Learning



Reinforcement Learning

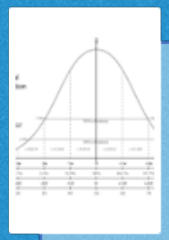
AI



http://arxiv.org/abs/1701.07274

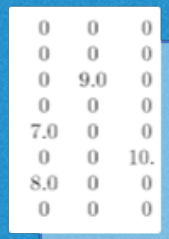
Data Science Algorithms

Intrinsic Data Quality



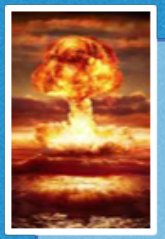
Statistics

- Organized
- Complete



Data Mining

- Semi-Structured
- Sparse Matrices

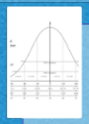


Machine Learning

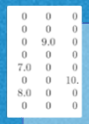
- Unstructured
- Unbounded

Data Science

Intrinsic Data Quality



Statistics
- Organized
- Complete



Data Mining
- Semi-Structured
- Sparse Matrices

Machine Learning
- Unstructured
- Unbounded

Algorithms

Algorithms (the rest)

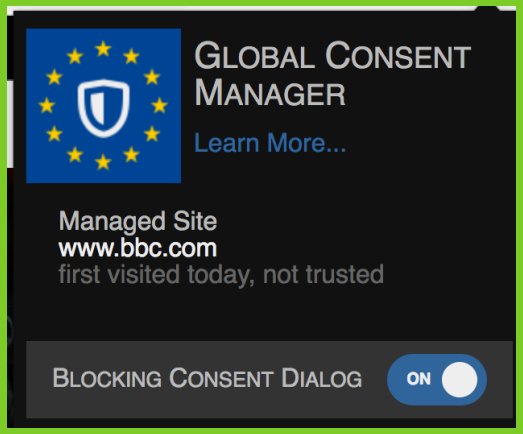
Data Carpentry
80% - 90% of Work Hours

Al: Red Sox in 4
RI: ???

Thank you!

CHAOSS

augur



Sean P. Goggins, Phd.
Associate Professor, Computer Science
University of Missouri

@SociallyCompute on Twitter
@ComputationalMystic on Instagram
@Sociotechnika on Flickr
@sgoggins on GitHub

http://www.seangoggins.net