2016 NECINA Financial Technology Conference

Applying Machine Intelligence in Financial Software Engineering

By Albert Ma, Chief Innovation Officer of Hengtian

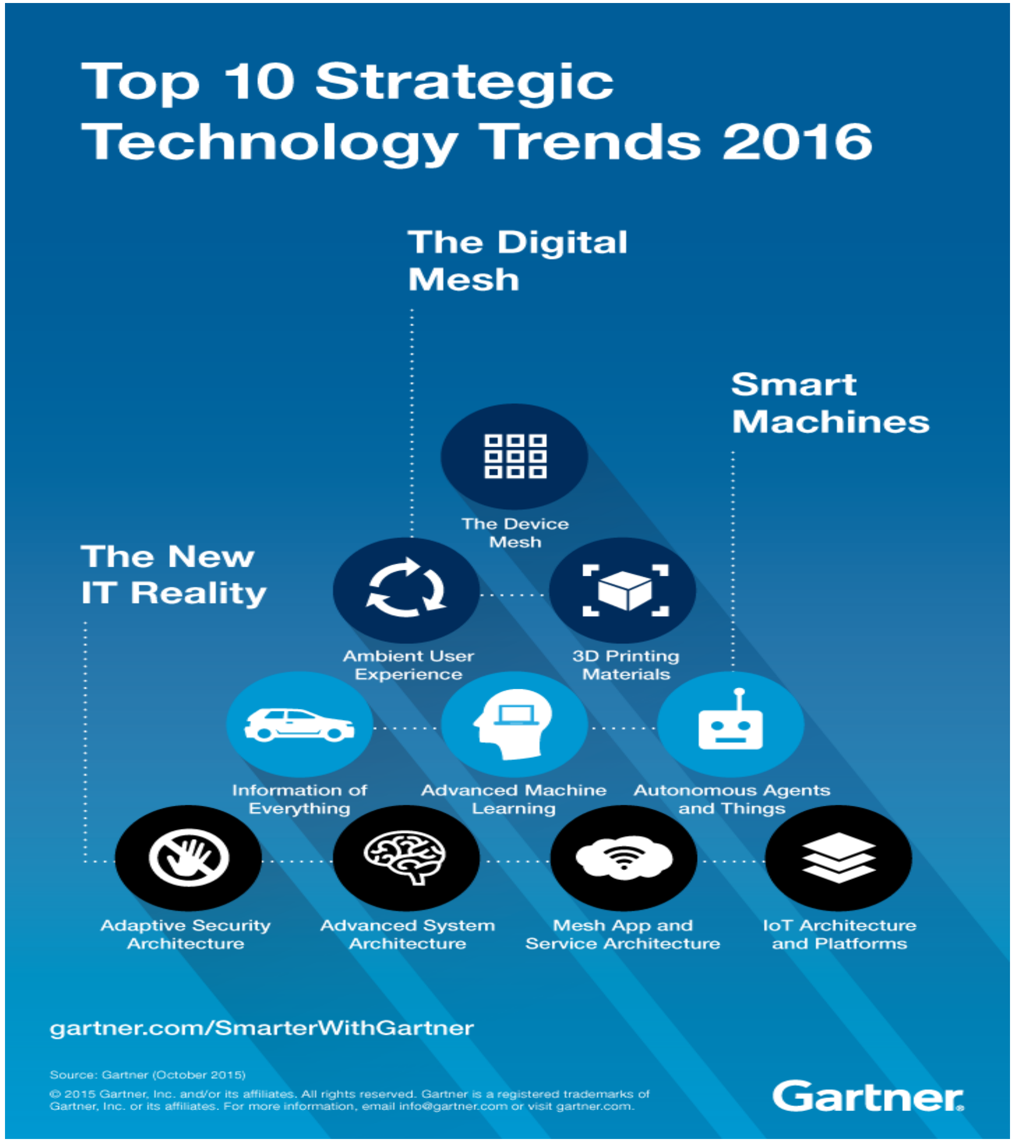
Date: April 23th, 2016

HengTian
Trusted Technology Solutions

Agenda

- Machine Intelligence Landscape
- Business Values
- Auto Programming
- DeepMorpho – Research for the Future
- Summary

HengTian
Trusted Technology Solutions



Top 10 Strategic Technology Trends 2016

The Digital Mesh

- The Device Mesh
- Ambient User Experience
- 3D Printing Materials

Smart Machines

- Information of Everything
- Advanced Machine Learning
- Autonomous Agents and Things

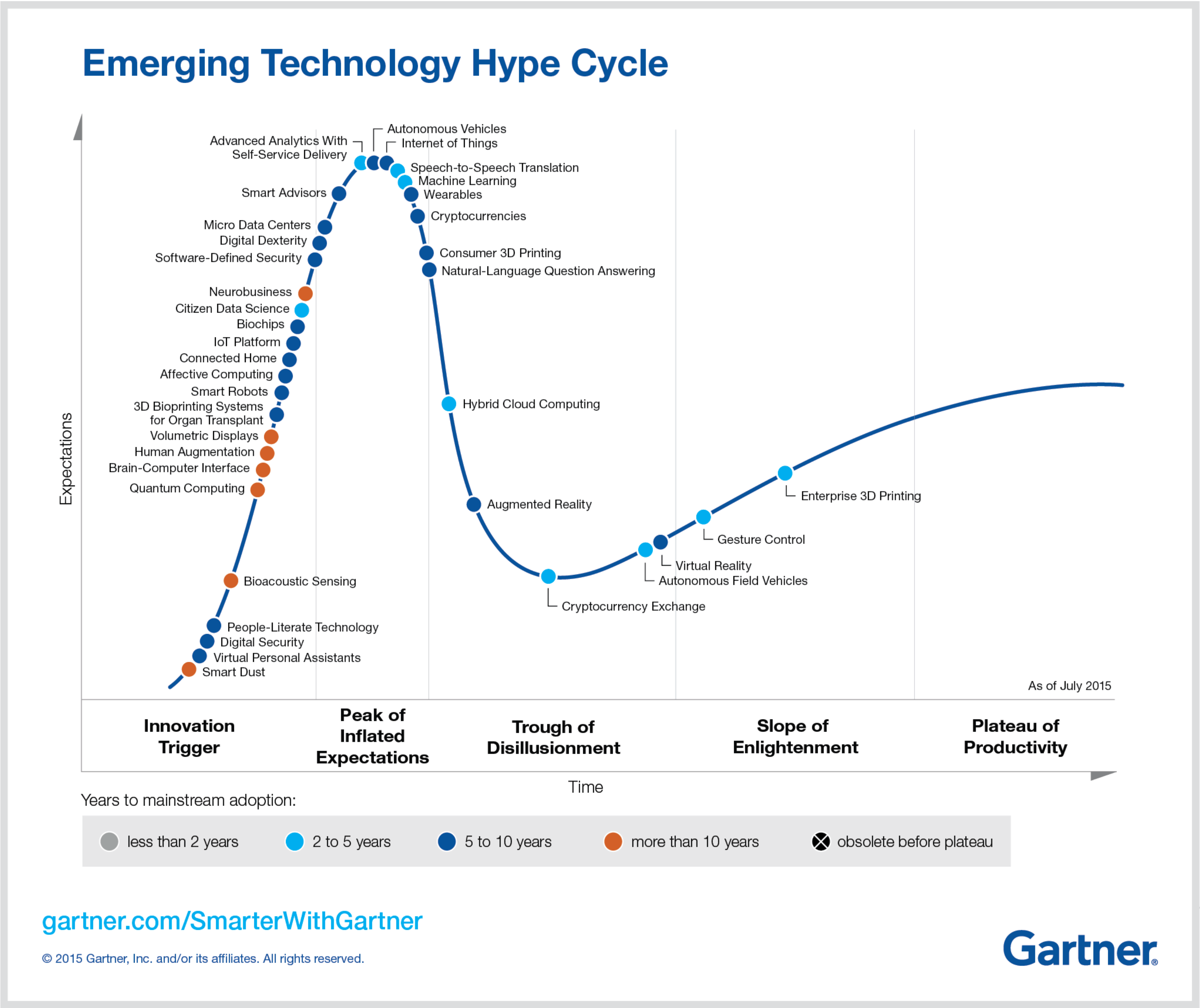
The New IT Reality

- Adaptive Security Architecture
- Advanced System Architecture
- Mesh App and Service Architecture
- IoT Architecture and Platforms

gartner.com/SmarterWithGartner

Source: Gartner (October 2015)

Gartner



Emerging Technology Hype Cycle

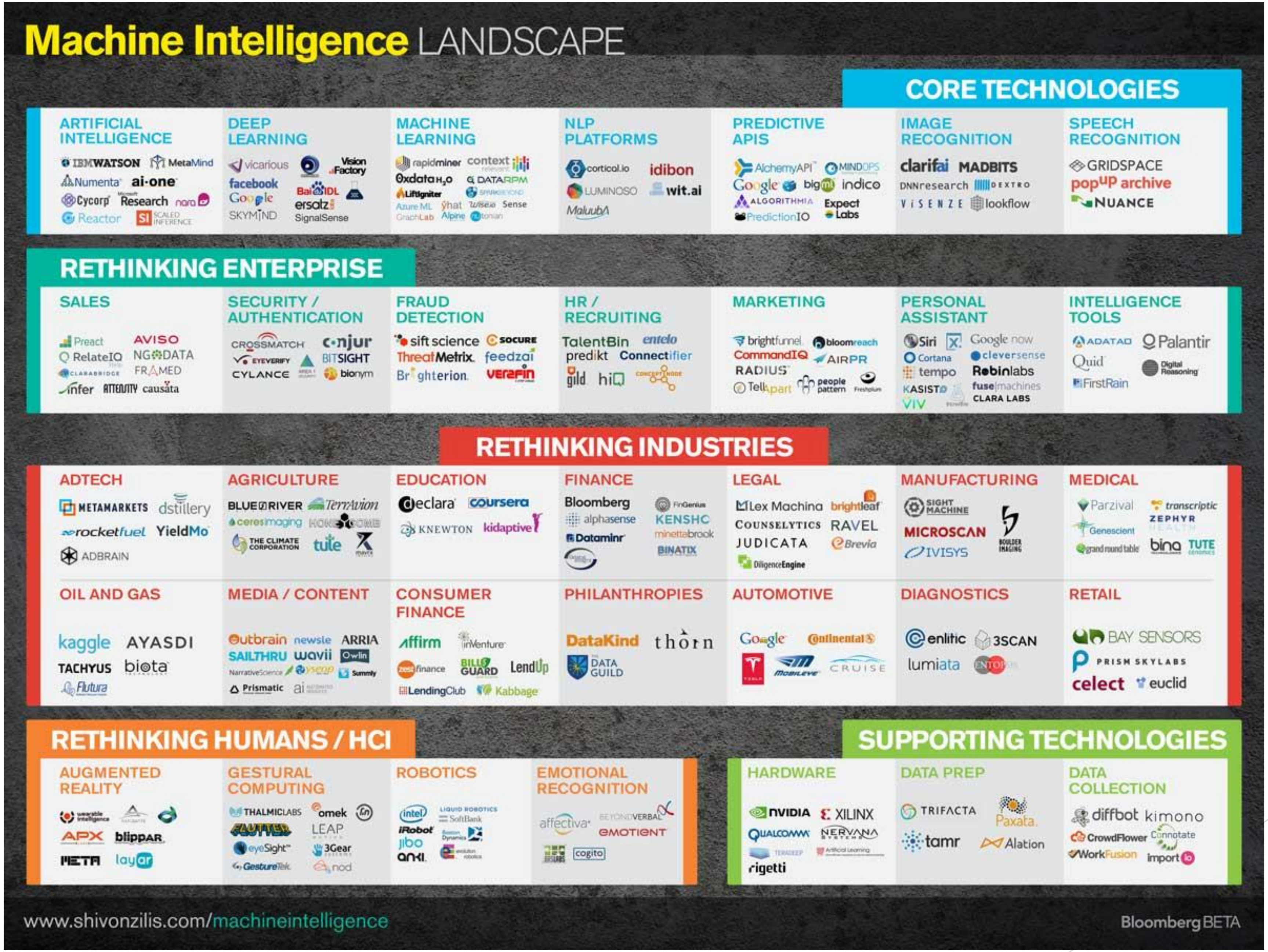
Years to mainstream adoption:

● less than 2 years ● 2 to 5 years ● 5 to 10 years ● more than 10 years ⊗ obsolete before plateau

gartner.com/SmarterWithGartner

Gartner

HengTian
Trusted Technology Solutions



Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

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

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

MACHINE LEARNING
rapidminer, context relevant, 0xdata H2O, DATARPM, Liftigniter, SPARKBEYOND, Azure ML, yhat, Wise.io, Sense, GraphLab, Alpine, Nutonian

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, CONCEPT NODE

MARKETING
brightfunnel, bloomreach, CommandIQ, AIRPR, RADIUS, TellApart, people pattern, Freshplum

PERSONAL ASSISTANT
Siri, X, Google now, Cortana, cleversense, tempo, Robinlabs, KASISTO, fuse machines, VIV, CLARA LABS

INTELLIGENCE TOOLS
ADATAO, Palantir, Quid, Digital Reasoning, FirstRain

RETHINKING INDUSTRIES

ADTECH
METAMARKETS, dstillery, rocketfuel, YieldMo, ADBRAIN

AGRICULTURE
BLUE RIVER, TerrAvion, ceres imaging, HONEYCOMB, 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, Genescient, ZEPHYR HEALTH, 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 AUTOMATED INSIGHTS

CONSUMER FINANCE
Affirm, inVenture, zest finance, BILLGUARD, LendUp, LendingClub, Kabbage

PHILANTHROPIES
DataKind, thorn, DATA GUILD

AUTOMOTIVE
Google, Continental, TESLA, MOBILEYE, CRUISE

DIAGNOSTICS
enlitic, 3SCAN, lumiata, ENTOPSIS

RETAIL
BAY SENSORS, PRISM SKYLABS, celect, euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence, NAVDY, APX, blippar, META, layar

GESTURAL COMPUTING
THALMICLABS, omek, FLUTTER, LEAP MOTION, eyeSight, 3Gear SYSTEMS, 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

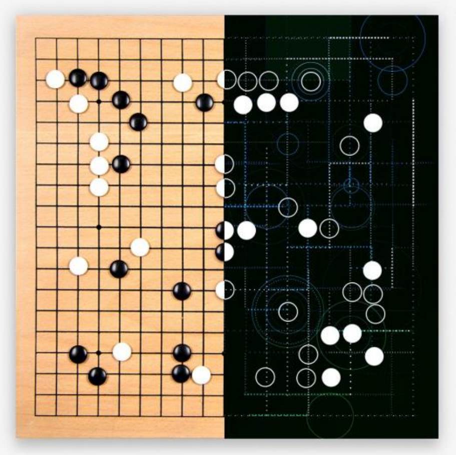
The Trends of Machine Intelligence

- Two Drivers: Raw Data and Data Model
- Machine Learning becomes a buzz word for business (InfoWorld)
- Big Data is assumed in Machine Learning applications (Gartner)
- Wall Street is gearing up with Machine Learning in fixed income, blockchain, predictive analytics etc. (McKinsey)
- Machine Learning is full of contradiction (Thomas Frey, The Da Vinci Inst.)

HengTian
Trusted Technology Solutions

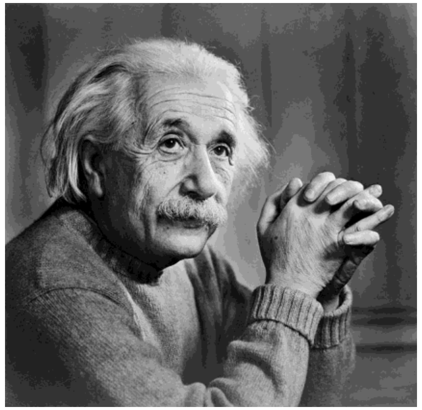
Source: http://www.dataversity.net/the-future-of-machine-learning-trends-observations-and-forecasts/

It is the year of artificial intelligence !

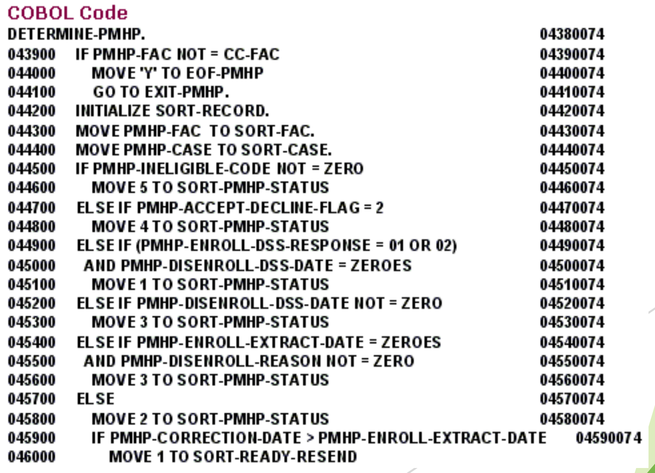


When DeepMind AlphaGo
win Lee Sedol…....

Can computer read source
codes like a human ?



?



COBOL Code

```
DETERMINE-PMHP.                                                        04380074
043900   IF PMHP-FAC NOT = CC-FAC                                      04390074
044000      MOVE 'Y' TO EOF-PMHP                                       04400074
044100      GO TO EXIT-PMHP.                                           04410074
044200   INITIALIZE SORT-RECORD.                                       04420074
044300   MOVE PMHP-FAC  TO SORT-FAC.                                   04430074
044400   MOVE PMHP-CASE TO SORT-CASE.                                  04440074
044500   IF PMHP-INELIGIBLE-CODE NOT = ZERO                            04450074
044600      MOVE 5 TO SORT-PMHP-STATUS                                 04460074
044700   ELSE IF PMHP-ACCEPT-DECLINE-FLAG = 2                          04470074
044800      MOVE 4 TO SORT-PMHP-STATUS                                 04480074
044900   ELSE IF (PMHP-ENROLL-DSS-RESPONSE = 01 OR 02)                 04490074
045000    AND PMHP-DISENROLL-DSS-DATE = ZEROES                         04500074
045100      MOVE 1 TO SORT-PMHP-STATUS                                 04510074
045200   ELSE IF PMHP-DISENROLL-DSS-DATE NOT = ZERO                    04520074
045300      MOVE 3 TO SORT-PMHP-STATUS                                 04530074
045400   ELSE IF PMHP-ENROLL-EXTRACT-DATE = ZEROES                     04540074
045500    AND PMHP-DISENROLL-REASON NOT = ZERO                         04550074
045600      MOVE 3 TO SORT-PMHP-STATUS                                 04560074
045700   ELSE                                                          04570074
045800      MOVE 2 TO SORT-PMHP-STATUS                                 04580074
045900      IF PMHP-CORRECTION-DATE > PMHP-ENROLL-EXTRACT-DATE     04590074
046000         MOVE 1 TO SORT-READY-RESEND
```

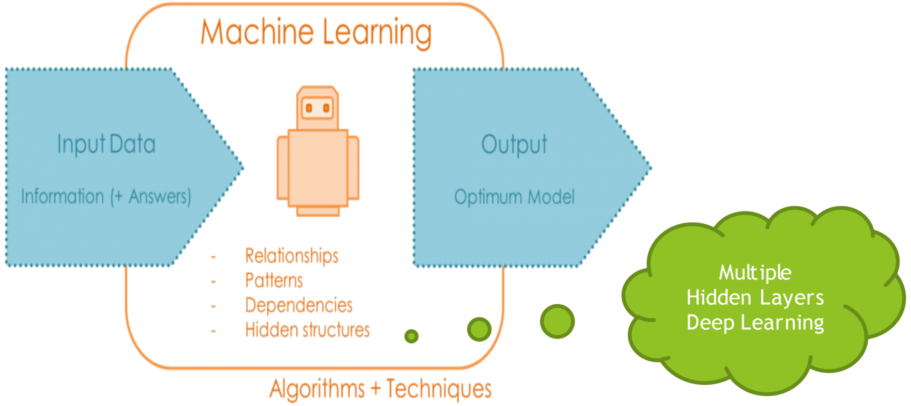
HengTian
Trusted Technology Solutions

Auto Programming Samples

- Tweet to program (http://www.wolfram.com)
- Natural Language Programming (http://www.pegasus-project.org)
- Cognitive Computing (http://www.ibm.com/watson)
- Statistical Machine Translation (https://youtube.com/watch?v=aRSnl5-7vNo)
- Pliny Big Code Analytics (http://pliny.rice.edu/index.html)
- Code Transplant (http://crest.cs.ucl.ac.uk/autotransplantation/MuScalpel.html)
- Bug Repair - MIT CodePhage (http://news.mit.edu/2015/automatic-code-bug-repair-0629)
- Malware Prevention - Deep Instinct (http://www.deepinstinct.com/#/what-we-do)

HengTian
Trusted Technology Solutions

Some Basics of Machine Learning



Supervised Learning:

Predicting values. **Known** targets.
User inputs correct answers to learn from. Machine uses the information to guess new answers.

REGRESSION:
Estimate continuous values
(Real-valued output)

CLASSIFICATION:
Identify a unique class
(Discrete values, Boolean, Categories)

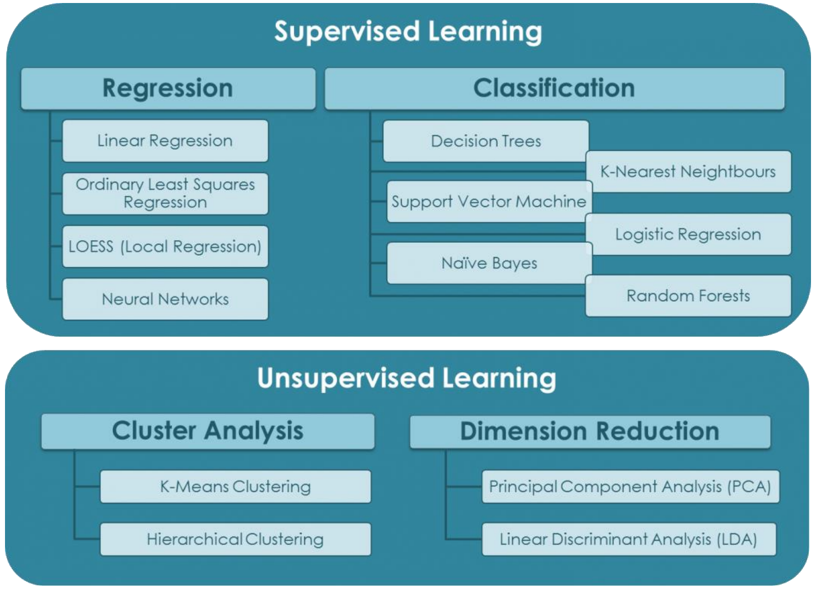
Unsupervised Learning:

Search for structure in data. **Unknown** targets.
User inputs data with undefined answers. Machine finds useful information hidden in data.

Cluster Analysis
Group into sets

Density Estimation
Approximate distributions

Dimension Reduction
Select relevant variables



Supervised Learning

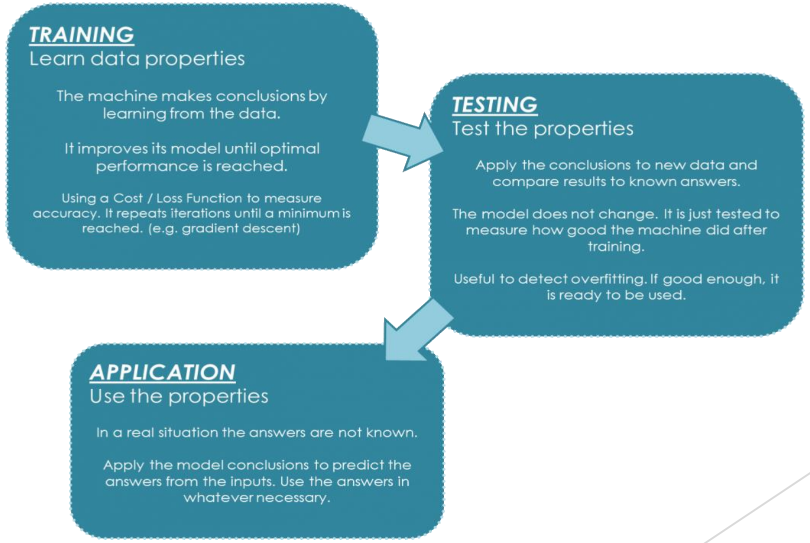
Regression
- Linear Regression
- Ordinary Least Squares Regression
- LOESS (Local Regression)
- Neural Networks

Classification
- Decision Trees
- K-Nearest Neightbours
- Support Vector Machine
- Logistic Regression
- Naïve Bayes
- Random Forests

Unsupervised Learning

Cluster Analysis
- K-Means Clustering
- Hierarchical Clustering

Dimension Reduction
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)



TRAINING
Learn data properties

The machine makes conclusions by learning from the data.

It improves its model until optimal performance is reached.

Using a Cost / Loss Function to measure accuracy. It repeats iterations until a minimum is reached. (e.g. gradient descent)

TESTING
Test the properties

Apply the conclusions to new data and compare results to known answers.

The model does not change. It is just tested to measure how good the machine did after training.

Useful to detect overfitting. If good enough, it is ready to be used.

APPLICATION
Use the properties

In a real situation the answers are not known.

Apply the model conclusions to predict the answers from the inputs. Use the answers in whatever necessary.

Source: http://quantdare.com/2016/03/machine-learning-a-brief-breakdown/

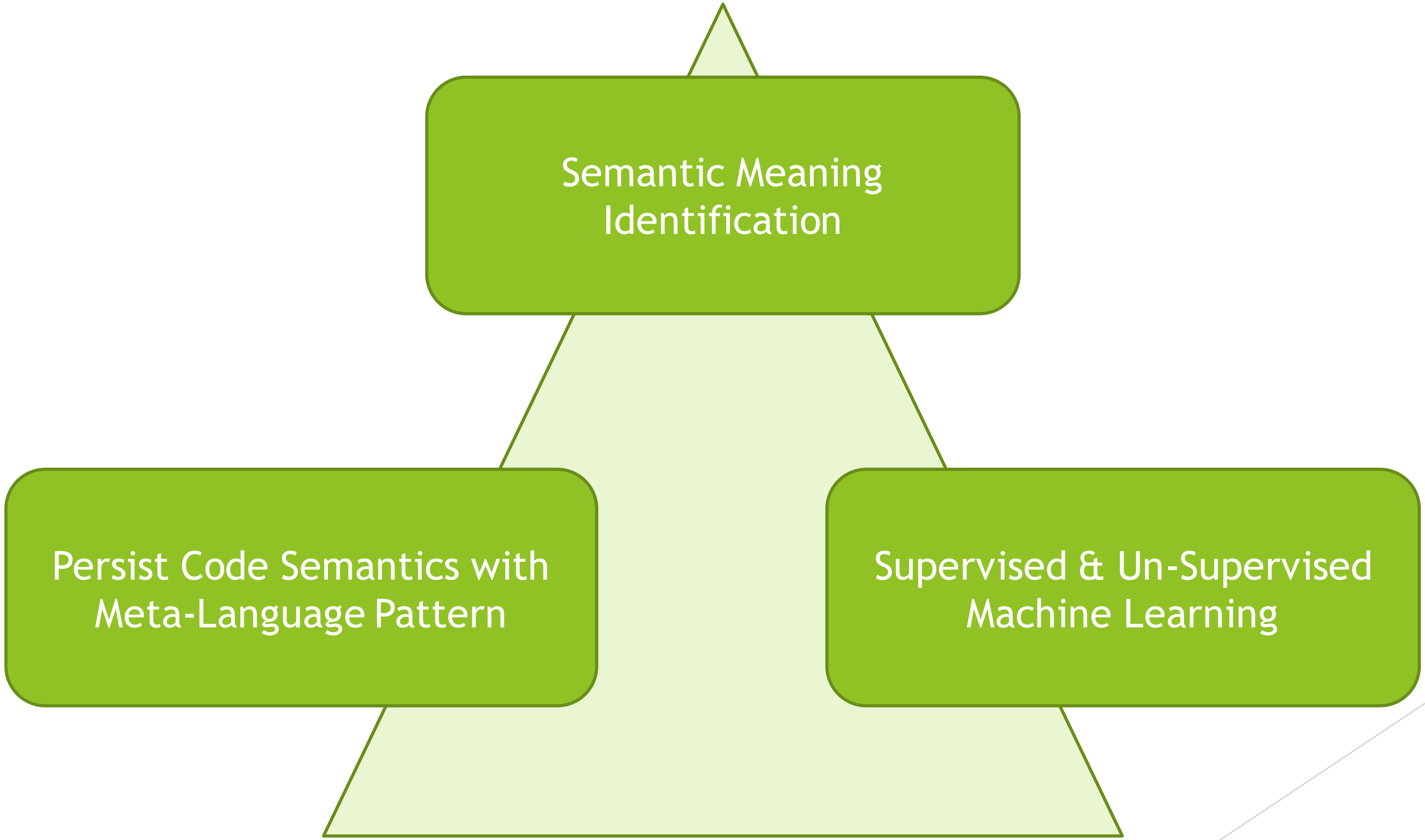
HengTian
Trusted Technology Solutions

Why learning from source code is difficult ?

- Programing language syntax structure is well defined in the complier but has infinite ways of implementing the same business logic
- There is not a universal way of representing program semantic meanings
- Existing NLP representation learning algorithms are inapplicable since all of them are “flat”
- Program symbols (nodes in AST) are discrete and cannot be fed directly to a neural network
- Like NLP, it takes many years to mature multi-language corpus
- Multiple programing languages and systems interconnected together inflates the complexity permutation

HengTian
Trusted Technology Solutions

The problems to be resolved in auto programming?



Semantic Meaning Identification

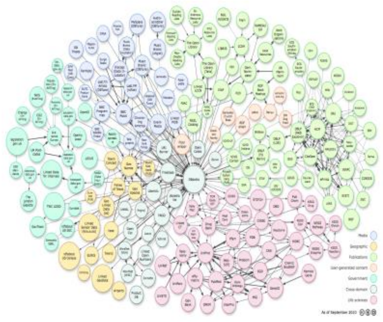
Persist Code Semantics with Meta-Language Pattern

Supervised & Un-Supervised Machine Learning

HengTian
Trusted Technology Solutions

Value Proposition – Why Machine Learning ?

1



Enterprise Code Ontology

Know your IT asset at granular level

2

Malware ?

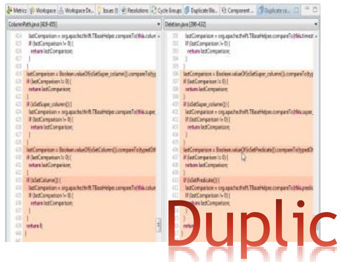
Bugs ?

System Failure ?

Back Door ?

Predictive Analytics

3



ColumnPath.java [419-455]

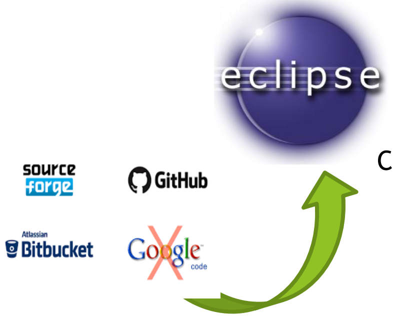
```
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.colum
    if (lastComparison != 0) {
      return lastComparison;
    }
  }
  lastComparison = Boolean.valueOf(isSetSuper_column()).compareTo(typ
  if (lastComparison != 0) {
    return lastComparison;
  }
  if (isSetSuper_column()) {
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.super_
    if (lastComparison != 0) {
      return lastComparison;
    }
  }
  lastComparison = Boolean.valueOf(isSetColumn()).compareTo(typedOth
  if (lastComparison != 0) {
    return lastComparison;
  }
  if (isSetColumn()) {
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.colum
    if (lastComparison != 0) {
      return lastComparison;
    }
  }
  return 0;
}
```

Deletion.java [386-432]

```
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.timest
    if (lastComparison != 0) {
      return lastComparison;
    }
  }
  lastComparison = Boolean.valueOf(isSetSuper_column()).compareTo(typ
  if (lastComparison != 0) {
    return lastComparison;
  }
  if (isSetSuper_column()) {
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.super_
    if (lastComparison != 0) {
      return lastComparison;
    }
  }
  lastComparison = Boolean.valueOf(isSetPredicate()).compareTo(typedOt
  if (lastComparison != 0) {
    return lastComparison;
  }
  if (isSetPredicate()) {
    lastComparison = org.apache.thrift.TBaseHelper.compareTo(this.predic
    if (lastComparison != 0) {
      return lastComparison;
```

Duplicate & Dead Codes

4



Code Transplant

HengTian
Trusted Technology Solutions

How does a programmer read the code ?

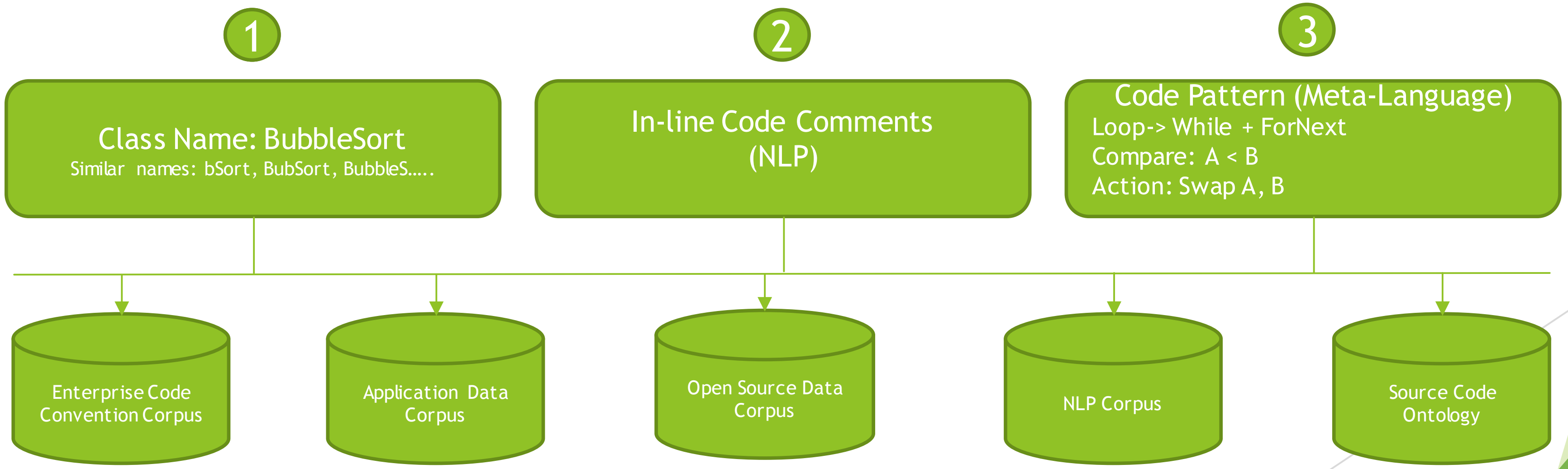
```
public static void BubbleSort( int [ ] num ) {
int j;
boolean sortFlag = true;   // initialize the flag for while loop
    int tempNum;

while ( sortFlag ) {

            sortFlag= false;    //set flag to false awaiting a possible swap
            for( j=0; j < num.length -1; j++ ){
                        if ( num[ j ] < num[j+1] ) {   // compare adjacent array elements

                                    tempNum = num[ j ];            //swap two elements
                                    num[ j ] = num[ j+1 ];
                                    num[ j+1 ] = tempNum;
                                    sortFlag = true;           //shows a swap occurred
                        }
            }
}
}
```

Bubble Sort ?



Code Signature

HengTian
Trusted Technology Solutions

Source Code Representation

- Semantic Annotation
  - Triples (‘class A’, ‘inherit from’, ‘class B’); (‘class A’, ‘has bizfunc’, ‘interest calculation’)
  - Resource Description File (RDF)
- Ontological Representation for Machine Learning
  - Vector based code block
  - Data flow
- Meta Language for Pattern Persistence
  - Regular Expression
  - Domain Specific Language (DSL)

HengTian
Trusted Technology Solutions

Learning The Source Code

Feature Engineering

Code Slicing

IDENTIFICATION DIVISION
ENVIORNMENT DIVISION
DATA DIVISION
PROCEDURE DIVISION…..

Syntax Parsing

Semantic Annotation

Data Flow Analysis

Deep Learning Architecture

Supervised Training

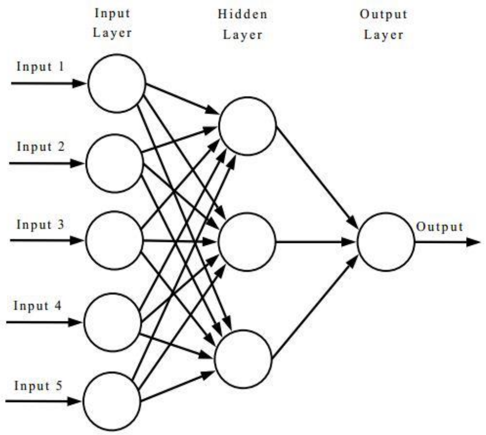
Code Pattern Expression

Meta Language

Class bSort
Swap VarA->VarB
Condition: VarA < VarB

Data & Language Corpus

Un-Supervised Training



Recurrent Neural Network

Research to the Future
Code Name: DeepMorpho

Output

Multi-System Code Ontology

Reusable Codes

Business Rules Doc

Code Transplant

Defect Prediction

HengTian
Trusted Technology Solutions

Summary

- There is not a single model that can solve all the software development problems
- Significant effort is needed to develop different language corpus and training data set. Payback occurs when there is a strategic motive and sustainable methods embedded in software development life cycle
- It is just the beginning for academic researchers and software vendors to look at applying machine intelligence into software development. The potential is huge with more and more algorithmic libraries coming out
- Accuracy will significantly improve if there is an enterprise code standard and it is being enforced properly
- As more software development counts on open source, it is a logical next step to extend machine intelligence to transplant codes from open source repositories

HengTian
Trusted Technology Solutions

Recommended Papers

- L.Mou, G.Li, Y.Liu, H.Peng, Z.Jin, Y.Xu, L.Zhang, Building Program Vector Representations for Deep Learning, Software Institute, School of EECS, Peking University
- N.Phan, D.Dou, H.Wang, D.Kil, Ontology-Based Deep Learning for Human Behavior Prediction in Health Social Networks, Computer Science, University of Oregon
- F.Long, M.Rinard, Automatic Patch Generation by Learning Correct Code, CSAIL, Massachusetts Institute of Technology
- E.Barr, M.Harman, Y.Jia, A.Marginean, J.Petke, Automated Software Transplantation, CREST, University College London
- W.Ling, E.Grefenstette, K.M.Hermann, T.Kocisky, A.Senior, F.Wang, P.Blunsom, Latent Predictor Networks for Code Generation, Google DeepMind, University of Oxford
- S. Liu, N.Yang, M.Li, M.Zhou, A Recursive Recurrent Neural Network for Statistical Machine Translation, University of Science and Technology of China

HengTian
Trusted Technology Solutions