Intelligent Data

Tech Symposium for Machine Learning and Artificial Intelligence

**Overview of UW's Data Science Institute**   
Dr. Sarah Stone   
eScience Executive Director   
[exec-director@escience.washington.edu](mailto:exec-director@escience.washington.edu)

**Technology of Preparing Data for Machine Learning**   
Dr. Daniela Braga   
CEO   
[daniela@definedcrowd.com](mailto:daniela@definedcrowd.com)

1269 120th Ave NE, Bellevue, WA 98005

**Data visualization through dimensionality reduction:   
Self-Organizing Map technique**   
Prof Yutaka Oya, Tohoku University   
[ohuchi@uw.edu](mailto:ohuchi@uw.edu)

**Hyper-Personalized Messaging  
with Structured Data and Chatbots**   
Phil Gordon  
CEO   
[phil@chatbox.com](mailto:phil@chatbox.com)

**Apache Spark compatible fast data processing  
on CUDA GPUs**   
Shannon Smith  
COO FastData.io   
[shannon@fastdata.io](mailto:shannon@fastdata.io)

601 Union Street, 42 Floor | Seattle, WA 98101

Amir Szekley, Chief Architect

[amir@fastdata.io](mailto:amir@fastdata.io)

**ML Model Development and Importance of Data**   
Rajeev Dutt  
CEO Dimensional Mechanics   
[rajeev.dutt@dimensionalmechanics.com](mailto:rajeev.dutt@dimensionalmechanics.com)

**Machine Learning In The Real World**  
Matthew Calamatta  
CTO Matchbox Mobile   
[matthew.calamatta@matchboxmobile.com](mailto:matthew.calamatta@matchboxmobile.com)

ChatBot

1 206966 6123

Text “AI”

**1st talk**

eScience Institute

Advancing Data-Intensive Discovery in all fields

Sarah Stone

**Society Facing Data Deluge**

Science

Industry:

Government: records, sensors

**Data Science challenges go beyond size**

Data in TB, GB and MB range

Idea of 4th paradigm of scientific discovery – a big push on computation recently, now data deluge to advance discovery

Scientific Computing with Python

Software Carpentry

Community Data Science Workshops

Hack weeks

Astro Hack Week 2018

Neurohackademy

Geohackweek 2018

Oceanhackweek 2018

Data Science (Discovery Spurs) bridges Scientific Theme Area and Data Science Methodologies

**Winter 2018 Incubator**

Experimental diffusion

Political twitter project summary and goal

**eScience Institute**

Data Science for Social Good

**Summer 2018**

Seattle Mobility Index Project – Seattle Department of Transportation

* Marketplace of destination – groceries, office spaces, etc.

Automatic damage detection post-hurricane

**West Big Data Innovation Hub**

California Water Data Challenge

**2nd talk**

Daniela Braga, Ph. D. Founder and CEO of DefinedCrowd

**Industry 4.0 and AI revolution**

AI, cybersecurity, cellphones, etc.

Everybody has an AI roadmap

**AI Applications**

Call centers

Self-driven cars

Shopping attendants

Medical diagnosis and surgeries

Borders

Education and learning

Robots taking care of elderly *(UK)*

House tasks

AI = mimicking the human brain

* Communication
* Reason
* Vision

Article: How Crowds of Humans Are Making AI Systems Scary-Smart

<https://futurism.media/how-crowds-of-humans-are-making-ai-systems-scary-smart>

“It Takes A Village To Raise an AI”

Trained before it can be interact properly

AI needs dataset, machine learning systems, and actual human beings to interact with.

**What it takes to build an AI**

High Quality Training Data + Machine Learning Algorithms = AI

*Gold rush to AI*

*Very few try to solve AI problem in enterprise scale*

Data scientists spent 80% preparing and cleaning data. 20% of their time complaining about their data.

4 stages of ML life cycle

1. Model ideation

* Raw & Unstructured Data
* OR no data

1. Model bootstrapping

* Data cleanup & annotation
* If no data, data collection

1. Model tuning / feedback loop of continuous movements

* Model training loop
* Model prediction
* Human Annotation/correction

1. Model customization

* Data collection (*may happen again, if you move into a different domain*)
* Model training loop
* Model prediction
* Human annotation/correction

**Personal assistant**

Under the hood

**Dialogue system flow**

* Automatic Speech Recognition (*orthographic transcription of voice signal*)
* Natural Language Processing
* Dialogue and services manager <= => External services
* Response generation
* Text-to-speech

**How AI works**

Use case:  
IBM client in Canada

Partnership program that allows the company to help their clients, customizing Watson on their base line

Word error rate

50% WER in ASR output

WER = (S + D + I) / N = ( ( S + D + I ) / (S + D + C) )

S is the number of substitutions

D is the number of deletions

I is the number of insertions

C is the number of correct words

N is the number of words in references

*Limit of 50 hours for acoustic training*

*Can we show any improvement with custom data?*

*We had to fake or mimic the data*

*Adapted the content with the real call center scenario*

*With French-Canadian speakers*

**Towards an AM model customization – step by step**

7 hours of data collection

Step 0: data preparation:

* Client product analysis for customer service
* Scenarios creation

Step 1: scripts generation (6208 sentences)

Step 2:

* Sourcing of (15) French Canadian speakers (10 female and 5 male)
* Recordings in mobile and desktop (distribution of acoustic environments)

Step 3: Text Audio validation

*The audio signal needs to map in time, exact with the orthographic transcription*

Trial version of definedcrowd

Step 1: Scripts creation

Step 1 on Neevo: text/scripts collection

*Impact of language translation*

Step 2: Voice creation

Step 2: Voice recordings

Step 2 on Neevo: Voice recordings

*PII gender etc.*

Step 3: Text-Audio Validation

Step 4: Watson speech to text customization

Step 4” Watson speech to text customization

Total length of upload audio

358,08 minutes – roughly 6 hours

From: 52% WER to 23% WER

*Machine learning accuracy is linear, it will plateau*

E2E Text to Speech Service – Scripted Speech

1 Access service

2 configure data and model settings

3 ….

*The only with 46 languages support*

Neevo community overview

Confidential Agreement

* NDA
* MSA
* Independent Contractor Agreement

Testing

Segmented based on metadata

Tracking

Scale because of crowd sourcing

Machine learning - > AI knowledge < - Crowd

*We have SLAs with clients, for quality.*

**Quality is everything**

Crowd Quality

Machine Learning & Quality Control

Task Design

*Getting the crowd is hard in other country in terms of paying them.*

*Hard payment route*

*Takes 2 weeks to ramp a crowd in a region*

*Every language have patterns*

*Acoustics and pragmatic*

*Other languages require more patterns to create a model*

*Romance languages and syntactic*

*All goes down to the data*

**3rd talk**

Self-organizing Map Technique: Data Visualization through Dimensionality Reduction

Speaker: Yutaka Oya

Traditional-way of presenting data

* Date-Base
* Properties
* Materials

**Visualization Techniques**

Data Set

* Precise Quantitative Information

Data Set

* Understanding Features

Density, etc.

**Co-occurrence Network Connectivity and Frequency of Each Word**

*Sample in a literature*

*Sentences are created by the network structure*

*Finding the main character in the literature*

**Ashby Maps**

Materials Distributes on Two-Dimensional Space

Graph: Elasticity X Density

Foams, Elastomers, Polymers, Metals, Natural materials, Composites, Ceramics

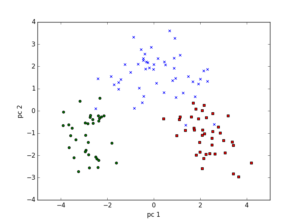
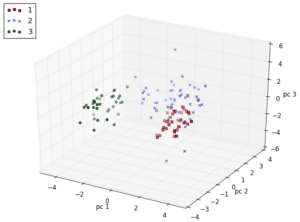
Understanding the relationship between Materials and Properties

**Hierarchical Clustering 1-D Storing depends on Similarities**

*Hierarchy has indirect proportion with similarities*

**K-Means Method Clustering method for Dispersed data**

**Principal component analysis Dimensionally Reduction**



Converting 3D -> 2D

**Summary of Visualization Techniques**

* Dimensionally Reduction
* Clustering
* 1 to 1 Relation between Materials and Properties
* Sorting Depends on Similarity

**Overview of Self Organizing Maps (SOM)**

Low to high heatmaps

* Density
* Glass Temperature
* Yield Stress
* Young’s Modulus
* Tri-Ax Strength
* Cure Rate

Above will be summarized in 1 heatmap

**Overview of SOM**

7 Properties (7 dimension)

Dimensionally Reduction: 7D -> 2D

Color Clustering

Similar (same color)

Dissimilar (different color)

Clustering and Sorting depending on Similarity

**Algorithm of Self Organization**

**Algorithm of SOM: Data Vectorization**

*I* (1-12): Materials *a ~ f* : Properties

**Two-Dimensional Mesh Structure**

Each connection can deform

From deformation of mesh to clusterization

**Clusterization of Materials**

Mesh + Cluterization + Self Organizing Map

Black solid lines denote clustered

Heatmaps

Density, heat capacity, melting point, boiling point, vapor pressure, surface tension, viscosity, thermal conductivity

**Computer Aided Engineering: Air Plane**

Base Resin

Curing Agent

Mixing Chemical Reaction

Cross-linked

Which combination is the best?

**Optimization of Thermosetting Resins**

Thousand of species to 2D

Cluster map is divided by Curing Agents and Base Resin

Materials in Group1, and Group 2 have higher Mechanical Properties

**Optimization of Wing Shape**

* Min bending
* Min pitching
* Low moment
* ….

**Working on Project: Optimization of Ceramic**

**4th talk**

**Hyper-Personalized Messaging  
with Structured Data and Chatbots**   
Phil Gordon  
CEO   
[phil@chatbox.com](mailto:phil@chatbox.com)

Another paradigm shift. Instead of bloated websites.

Chatbot interface.

Messaging is the killer app

5+ billion text messages

9/10 people prefer to message instead of call

Static -> Dynamic -> Mobile -> Responsive -> Hyper-Personalized

Google RCS

Apple Business Chat

Data and AI for personalized, targeted products, services, experiences

*Not 1 size fits all, its 1 customer at a time.*

6 Key Technologies

* Omni-Channel
* Automation
* Instant Apps
* Integrations
* Agent Tools
* Analytics

Build Once, Deploy to Any Channel

**Automate through Natural Language Processing (“NLP”)**

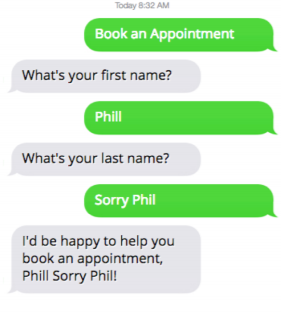
Intent-Slot Engines

FAQ Engines

**BIG PROBLEM**

With intent-slot engines

* Really bad at structured data capture



“What time is a good time to call?”

Checked all of the intents with Amazon LEX and Google API.AI (Google DialogFlow)

**The Scourge of Combined Probability**

**Instant Apps: Dynamically Generated by AI**

Generative AI will generate an instant app

* Profile
* SaaS Apps
* Business Logic
* Data Sources/APIs
* Machine Learning

**Agent Tools**

**Analytics**

Message Analytics, *how many times rebooked a flight.*

*Bots are never going to be good with structured data*

**Demo**

Customer Says “Feedback”

Train NLP to Determine Intent

Fulfill with Instant App

Feedback in Google Sheets

*ChatBox*

*Helps the AI Chatbots*

*We are not providing the AI on our own*

*We are the glue*

*Bring your own bot engine, we provide the fulfillment*

**5th talk**

FASTDATA

Data Processing at the Speed of Thought

**Paradigm Shift**

Batch Processing

COLLECT -> STORE -> PROCESS

FASTDATA

Time-Critical Decisions

@Real-time

Batch Processing Intelligence

@Seconds/MinutesHours

Long-running map reduce jobs for business analytics

@Days/Weeks/Months/Years

Problem: Currently you only have two

* Fast
* Big
* Efficient

Current Limitations

CPU + Java

Fail Points:

* Too much power & space
* Never designed for real-tome
* Rail & Stack

Plasma Engine

The first GPU-native software engine to fully leverage the Massively Parallel Processing Power of NVIDIA

Up to 1000X Faster

fastdata.io’s algorithm

Plasma Engine faster than Apache Spark

FDIO = 114,000,000 rows/sec

Apache Spark = 118,000 rows/sec

**Demo**

*We ship our software as a docker container*

*Apache Spark Demo uses CPU only*

*Fastdata uses minimal CPU and mostly on GPU*

**Fastdata.io Advantage**

Highly Efficient

Native GPU Processing

* Industry first over 80% of processing executed on GPU

Apache Spark 2.0 APIs

* Structured Streaming
* DataFrame

*Fast data collection but usual problem is slow connection*

**Easy Integration**

No Rip & Replace Required

Step 1: Identify data stream

Step 2

Step 3

**Customer ROI**

Saving you 9.39 hours of TB

*If Uber uses real-time data, they’ll lose a lot of money or charge you $1000.*

*If its network bound or I/O bound, then it is not better.*

**6th talk**

The Tyranny of Data

Elephant in the Room: Data

“Quality data is the truth, the whole truth, and nothing but the truth” – Ralph Kimball

5 elements of data quality

* Accuracy
* Issue: Precision
* Completeness
* Issue: Existence, Comprehensiveness
* Consistency
* Issue: Internal Consistency and External Consistency
* Uniqueness
* Issue: Referential Integrity
* Timeliness
* Issue: Integrity Across Time, Mutability of Data

Barriers to collecting data:

* Security Issues (access & authorization)
* Compliance
* Privacy
* Boundaries
* Compatibility

**Machine Learning & Deep Learning**

All machine learning problems can be mapped onto a deep learning problem (Universal Function Approximator Theorem)

*If data is poor, not able to represent reality, then it will be garbage.*

*Incredibility expensive to invest in data*

**Deep Learning Process**

* Ask the right question
* Select/curate data
* Pre-process
* Transform data (ML process)
* Train (ML process)
* Validate (ML process)
* Integrate (ML process)
* Test

*A false positive for hotdogs will have a lower positive.*

*Don’t start the question with data, start with the business question*

**Pre-process**

Format - > Clean - > Sample

**Transform**

Scale - > Decompose - > Aggregate

**Bad habits for machine learning**

*Machine learning should be the end point not the start point*

Not normalizing data

Not visualizing data

**The NeoPulse Framework**

Enterprise solution for AI

NeoPulse AI Studio

Portable Inference Models

NeoPulse Query Runime

NeoPulse Modeling Language

Using CSV file

Create the .nml script

Export a PIM file – the executable

*What the oracle (auto) does, is picks the best model for the problem it is trying to solve*

*How to solve bias in the model? Not doing this.*

*No human intervention on algorithms picking*

*Model optimization is not that important compared to hyper parameters.*

**7th talk**

Matchbox Mobile

Machine Learning in the Real World

**30 minutes in 30 seconds**

Book: Start with Why?

01 Why is the most important question

02 Adventures in Applied ML

03 Three Little Lessons

**Our “Why”**

* Help people we like
* Understand new tech
* Meet like-minded explorers
* Build excellent projects
* Do no harm

*May or may not involve machine learning*

BANDAI NAMCO – Pacman

* Alexa Voice Adventure in Pacman

**How we help**

* Professional services
* Lovely people: Engineers, Designers, Data Engineers, etc.
* Bold partners
* Pioneering projects
* Relationships
* Mastering new tech, deliver the benefits

Bluetooth has an office in Kirkland

*We haven’t quite seen the use of blockchain*

*Story of a blind man and the elephant*

*Grabbed the tail – rope*

*Leg – tree*

*Trunk –*

*You approach ML like that*

*We are not creating new algorithms*

**The Real World**

* Messy & fast-moving
* Nobody knows anything
* We don’t know who to trust
* We don’t know what works
* We don’t have the money, and we don’t know how much we need
* And yet we persist

**ML & Matchbox**

* People come to us with business challenges, Why?
* To improve or automate processes
* To make better decisions and faster

**A question of sport**

Problem: “Automate our Process”

Premise: “People watching video”

Promise: “Predict future performance”

*Video of sports, look at the video, and look at the form.*

Problem: not enough data

Problem: metrics did not support the promise

Process: find out what does; bring in our friends: Microsoft and Management Foresight

Now we know what questions to ask of the video…

Hack-Fest

Pose Estimation

Points to Ponder (Next)

*Microsoft gave us a room, for VM. 3 teams running image-classification test in a hack-fest*

*And used pose estimation*

*Will even super impose on a crappy webcam video*

**Lessons Learned**

Yes, you can learn fast

1. Hack-fest, design sprint also apply to ML!
2. Success is a team sport
3. Don’t build it if you don’t have to

*How about bias? Problematic if used for getting house loan*

**Lesson 1: Structured Hack**

* Design Sprint or Hack-Fest
* 3-5 days on site
* Team of stakeholders
* Come with questions, people, data and tools
* Leave with a tested prototype and user feedback
* Take that to your investors,
* Management, peers!
* Effective and cost-effective

*You walk out with a prototype to show to stakeholders*

**Hack Demo: What It’s Worth**

* Buy and sell second-hand items on eBay
* Use your camera and MS Computer Vision to recognize the item
* Use eBay to get the price range
* 2 days to working app using trained image classifier

**Lesson 2: Teamwork**

Many skills and roles for applied ML

* Decision-makers / product managers
* Domain experts / social scientists
* Data engineers and analysts
* Applied ML enginerers
* Statisticians
* Software developers
* Project / program managers
* Reliability engineers

Google defines Decision Intelligence Engineer

* Cassie Kozyrkov

*Tells you might just get an existing app, but you should know what is it.*

*How would you know you picked the right algorithm?*

**But what about…**

* Ethicists
* Should we do this? What are the ethical/moral/social implications
* Legal Analysts
* Can we do this here? In some other country?
* Explainers and Auditors
* How will you market, explain and defend your technical solution based on ML?
* Operations Managers
* How will you put ML into the market, into operation, with human oversight and intervention?

*Software is the least responsible product in the world*

*Loan or no loan, there must be a human appeal.*

**Lesson 3: use what’s around you**

* Plumbing is easier than making steel
* Don’t build what you don’t have to
* Lots of platforms
* Lots of libraries
* Be mindful of fitness for purpose
* Beware of running costs
* Be aware of IP licensing restrictions

**Regardless, in Seattle**

* Right, people, tech. community, passion
* Microsoft, Amazon et al
* Chatbox
* DefinedCrowd
* Management Foresight
* FastData
* DimensionalMatrix

**Starter questions**

* Why are you doing this? What is the business problem?
* Is your data good enough?
* Do you have the right data?
* Do you have the right IT infrastructure?
* How much it cost to run?
* How will you explain and defend?
* How will this impact your customers, partners, suppliers, society?
* How could ML fit, and what are the alternatives?

*Use cheaper humans and say it was ML*

**Summary for all**

**Chatbox**

Platform for chatbots and instant apps

**Fastdata**

Utilizing GPU

**Matchbox**

Consultancy. What tech to use.

Tom Sato

Data

* Network with the best people

Future

* Japan-Seattle Innovation Meetup

Planet

* What happens in Seattle affects everyone globally