
# Knowledge Discovery and Data Mining (KDD) 2016

***August 13 - 17, 2016 | San Francisco, California***

# Program

- Keynotes
- Plenary Panel
- Applied Data Science Invited Talks & Panels
- Hands-On Tutorials
- Accepted Papers Presentation
- Tutorials
- Workshops
- VC Office Hours

# KDD 2016


Is Data Science
a Product
or a Technique?
Is Data Science
a Product
or a Technique?
Is Data Science
a Product
or a Technique?
Is Data Science
a Product
or a Technique?

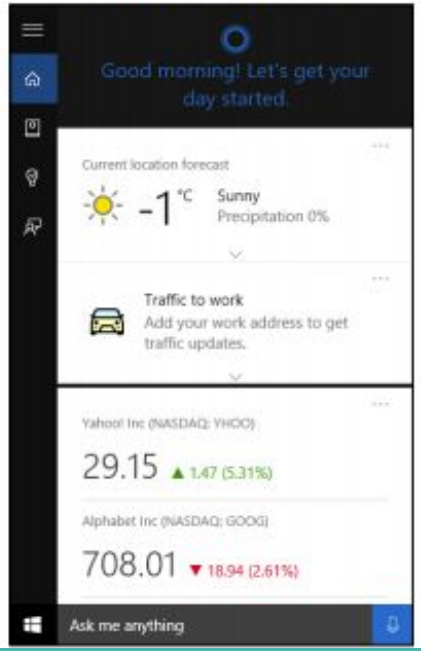
# Whitfield Diffie Talk

- Do you know Diffie–Hellman key exchange?
- Win Turing Award (2015)
  - The ACM A.M. Turing Award is an annual prize given by the Association for Computing Machinery (ACM) to "an individual selected for contributions of a technical nature made to the computing community"
- Problem now: Cryptography is threatened by quantum technology!

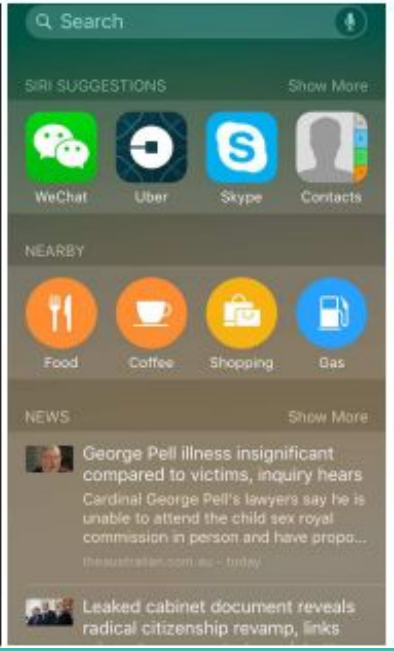
# Contextual Intent Tracking for Personal Assistants - Best student paper award

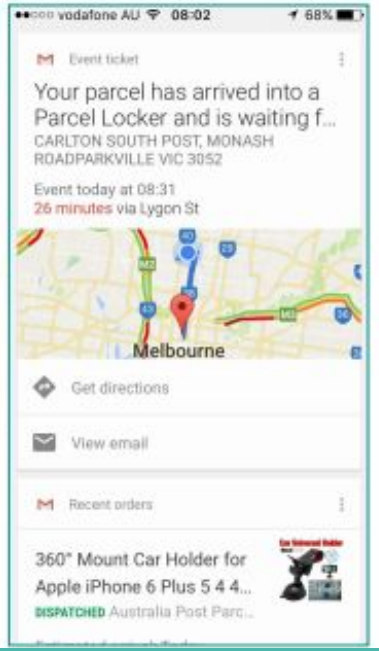
Microsoft Cortana Apple's Siri Google Now

Windows 10 Win Phone







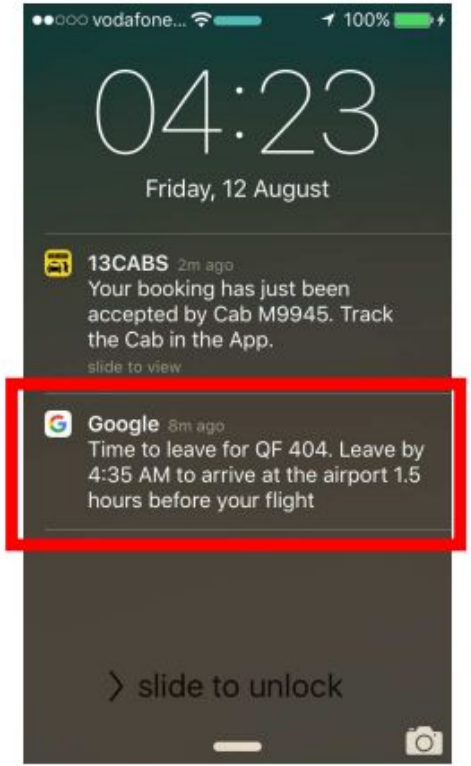



# Intelligent Personal Assistants

Morning: Email


Evening: Music


Travel Reminder



# What Users Intend to Know/Do

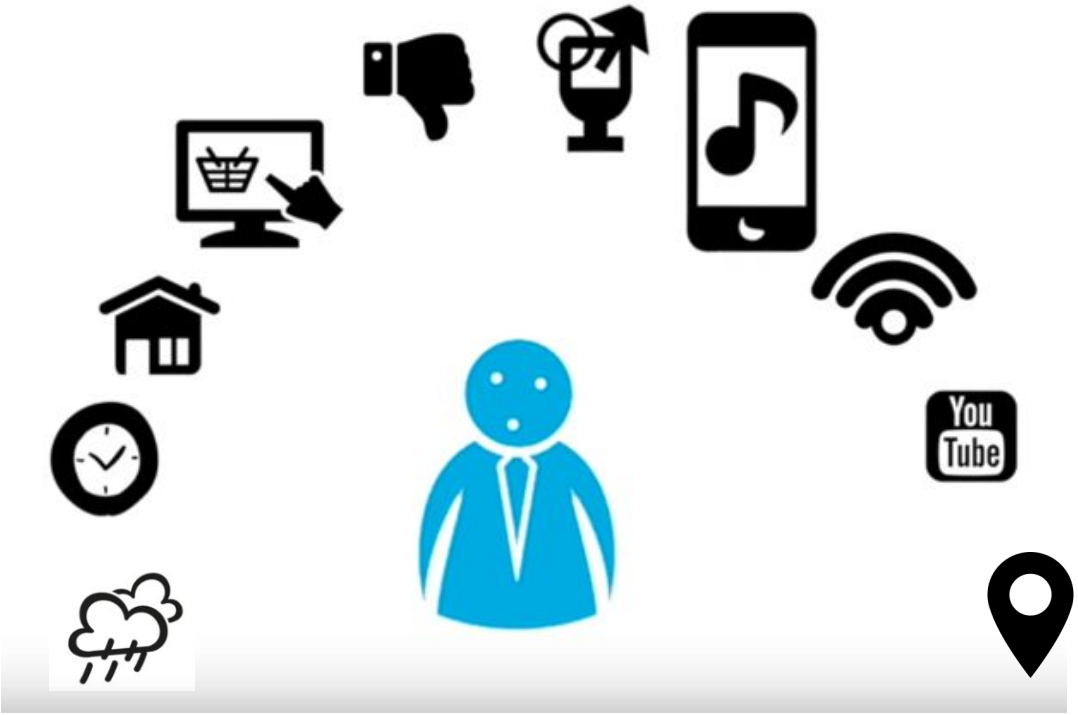
## Focused Recommendation/Notification

- **Limited** display sizes show limited content
- Push **one** notification or remind **one** task

## Track Users' Intent

- What users intend **to know**: information intent
- What users intend **to do**: task-completion intent

# Contextual Intent Tracking for Personal Assistants



# Contextual Intent Tracking for Personal Assistants

# Results

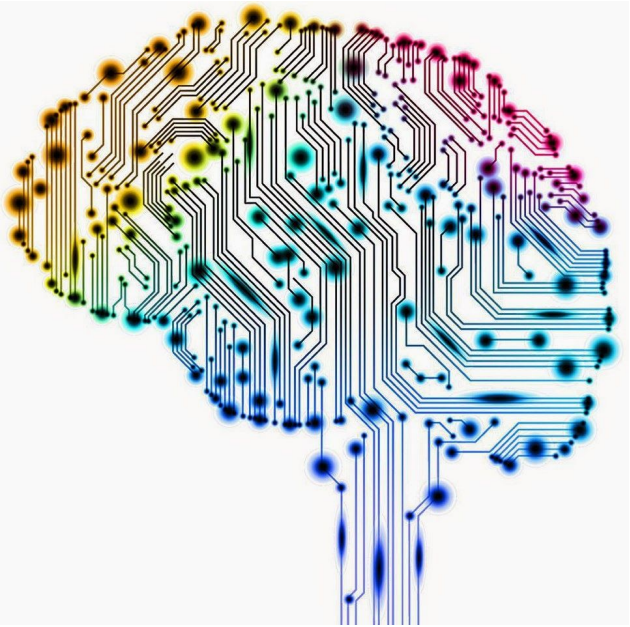
| | Intent | Triggers |
|---|---|---|
| | Message | Between 5:30 p.m. and 7:30 p.m., weekday, arriving at a food and drink venue |
| | Music | Later than 6:30 p.m., using browsers |
| | Taxi | Later than 8:30 p.m., weekday, distance to office $>$ 8km, leaving a supermarket |
| | Reserva-tion | Earlier than 6:30 p.m., Sunday, playing computer games for a long time |
| | News | Between 6:00 a.m. and 10:00 a.m., Friday, or weekends, distance to office $>$ 10km |

# "Why Should I Trust You?" Explaining the Predictions of Any Classifier By Marco Tulio Ribeiro


Github code

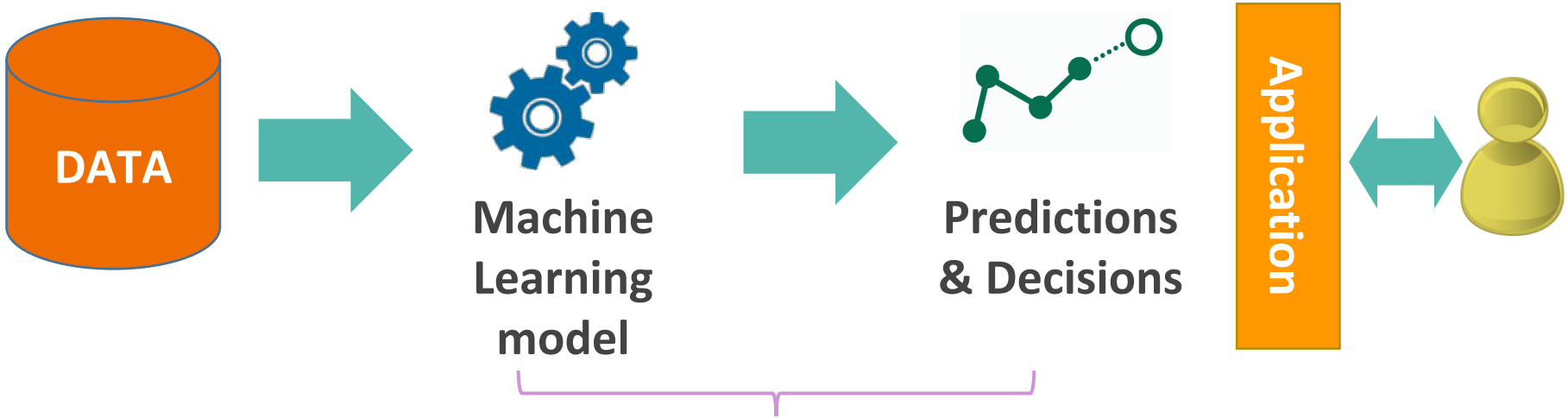
# Machine learning nowadays





Source

# How to build an application with ML




**TRUST CHALLENGE**

- Is model really working?
- Convince myself and others?

# If we don’t understand our model

Hard to improve

to move to other classfier

Getting fired!

# Accuracy problems - Example

20 Newsgroups subset – Atheism vs Christianity

94% accuracy!!!

Predictions due to email addresses, names,…

Test on recent dataset, accuracy only 57%

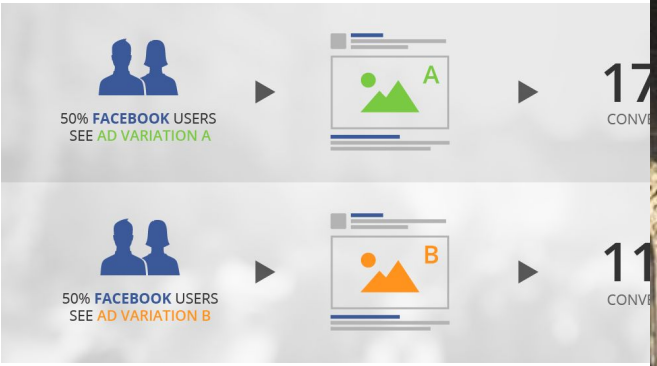
# How we try to gain trust?

Interpretable

Accuracy

A/B Testing

Voodoo



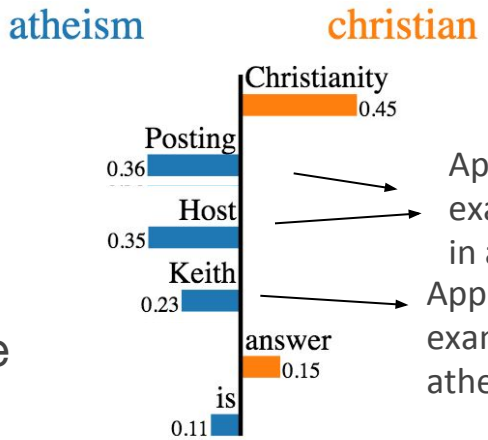

- “Almost” gold standard, but…
- Slow, expensive, tricky to interpret properly [Kohavi et al, KDD2012]

- AKA gut feeling, “I’m the expert”, looks good,…

# What an explanation looks like

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note




Appear in 21% of training examples, almost always in atheism

Appears in 11% of training examples, **always** in atheism

Why did this happen? How do I fix it?

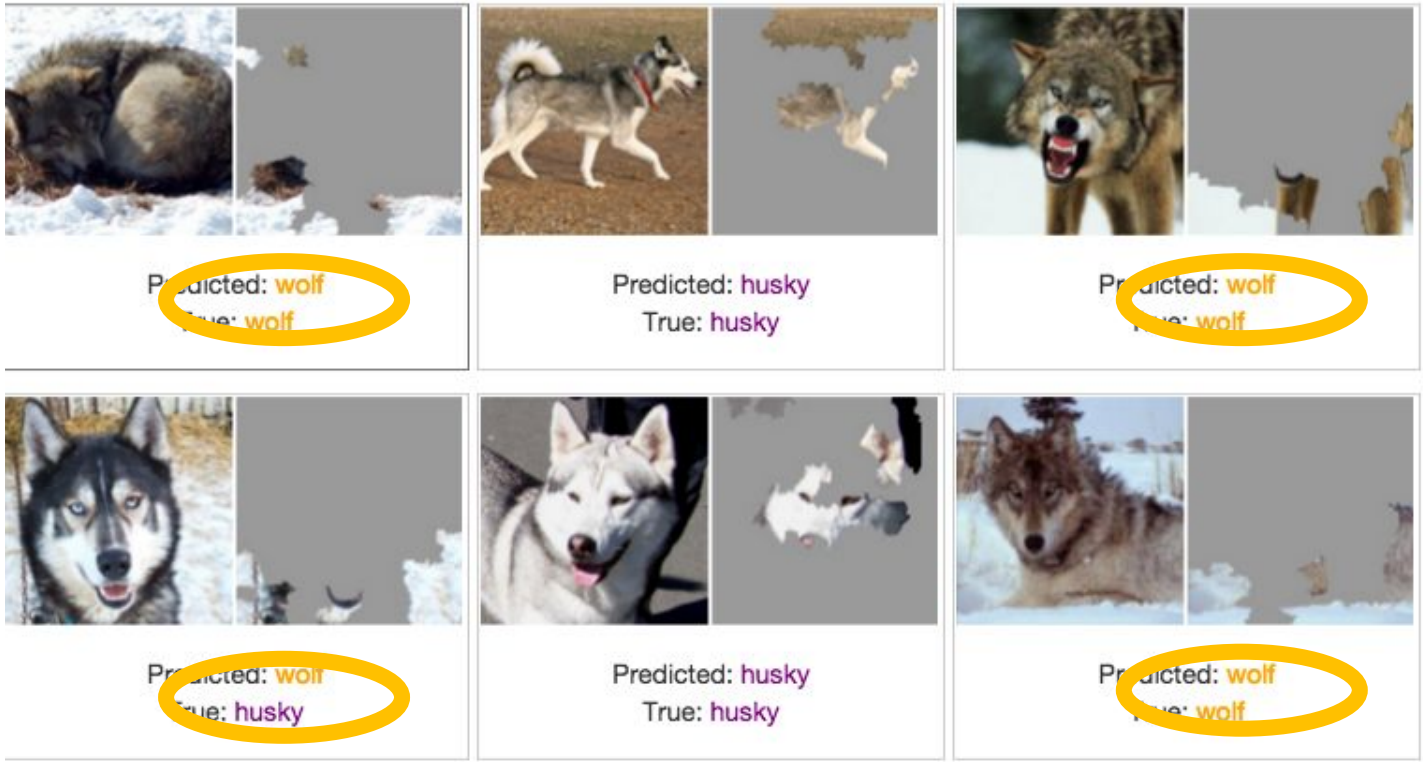
→ Will not generalize
→ Don't trust this model!

# Train a neural network to predict wolf vs. husky


Only 1 mistake!!!

Do you trust this model?
How does it distinguish between huskies and wolves?

# Explanations for neural network prediction




We've built a great snow detector… ☹

# Three must-haves for a good explanation

| | |
|---|---|
| Interpretable | Humans can easily interpret reasoning |
| Faithful | Describes how this model actually behaves |
| Model agnostic | Can be used for ***any*** ML model |

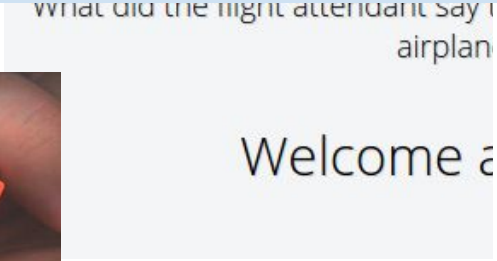
# DopeLearning: A Computational Approach to Rap Lyrics Generation By Eric Malmi

- Miscellaneous Topics
- Computational Creativity : (also known as artificial creativity, mechanical creativity or creative computation) is a multidisciplinary endeavour that is located at the intersection of the fields of artificial intelligence, cognitive psychology, philosophy, and the arts. - Wikipedia

# Computational Creativity

- Joke generator: dadjokegenerator
- Poetry generator: poemgenerator
- Music generator


## computer like human





http://weknowyourdreams.com/images/music/music-04.jpg

http://weknowyourdreams.com/images/music/music-04.jpg

# Rap Lyrics

She said "Some days I feel like **s**t**,
Some days I wanna **quit**,
and just be normal for a **bit**,"
I don't understand why you have to always be **gone**,
I get **along**
but the trips always feel so **long**,
And, I find myself trying to stay by the **phone**,
'Cause your voice always helps me to not feel so **alone**,
....

***Fort Minor - Where'd you go***

# deepbeat

| | |
|---|---|
| **Everybody got one** | *(2 Chainz - Extremely Blessed)* |
| **And all the pretty mommies want some** | *(Mos Def - Undeniable)* |
| **And what i told you all was** | *(Lil Wayne - Welcome Back)* |
| **But you need to stay such do not touch** | *(Common - Heidi Hoe)* |
| **They really do not want you to vote** | *(KRS One - The Mind)* |
| **what do you condone** | *(Cam'ron - Bubble Music)* |
| **Music make you lose control** | *(Missy Elliot - Lose Control)* |
| **What you need is right here ahh oh** | *(Wiz Khalifa - Right Here)* |
| **This is for you and me** | *(Missy Elliot - Hit Em Wit Da Hee)* |
| **I had to dedicate this song to you Mami** | *(Fat Joe - Bendicion Mami)* |
| **Now I see how you can be** | *(Lil Wayne - How To Hate)* |
| **I see u smiling i kno u hattig** | *(Wiz Khalifa - Damn Thing)* |
| **Best I Eva Had x4** | *(Nicki Minaj - Best I Ever Had)* |
| **That I had to pay for** | *(Ice Cube - X Bitches)* |
| **Do I have the right to take yours** | *(Common - Retrospect For Life)* |
| **Trying to stay warm** | *(Everlast - 2 Pieces Of Drama)* |

# DopeLearning

- Lyrics created by dopelearning
- DopeLearning learn to sing

# Plenary Panel Is Deep Learning the New 42?

**Pedro Domingos**
Professor
Univ. of Washington

**Nando de Freitas**
Professor
Oxford University

**Isabelle Guyon**
Professor
Université Paris-Saclay

**Jitendra Malik**
Professor
Univ. of California at Berkeley

# Plenary Panel Is Deep Learning the New 42?

Why Deep Learning?

- **Computer Vision**
  Reduce error
  rate significantly

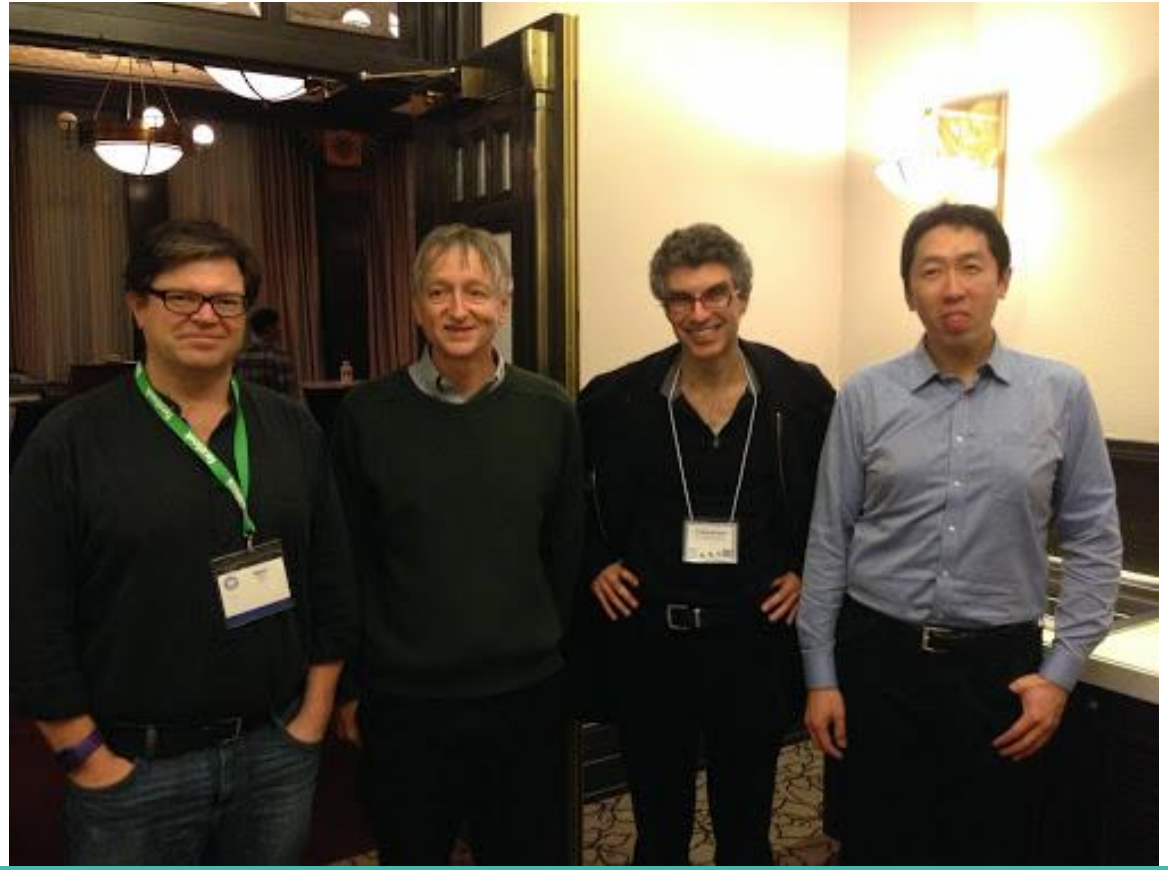
- **Speech**
  Google Voice Search

# Plenary Panel Is Deep Learning the New 42?

Why Deep Learning Succeed?

1. Big labelled data
2. GPU (thanks gamers)
3. ANN innovation (thanks Geoffrey Hinton)

# Plenary Panel Is Deep Learning the New 42?



# Plenary Panel Is Deep Learning the New 42?

Where will traditional ML continue to beat DL?

1. Interpretability
2. Not a silver bullet
3. Small size of data
4. Diversities

# Plenary Panel Is Deep Learning the New 42?

Is there preference cascade for deep learning?

**Yes, but the hype must be stir into the right direction**

# Plenary Panel Is Deep Learning the New 42?

Will consumptions of energy limit the development of deep learning?

1. Neuromorphic chips
2. Optimize algorithm

# Plenary Panel Is Deep Learning the New 42?

Is there such a thing as Repugnant Data or Repugnant Machine Learning?

**YES**

1. Redlining
2. Machine bias

**SOLUTIONS**

1. Final decision depends on human
2. Educate

# Standards in Predictive Analytics In the Era of Big and Fast Data

WEKA
The University of Waikato

scikit learn

rapidminer

R

Gephi

TensorFlow

Caffe

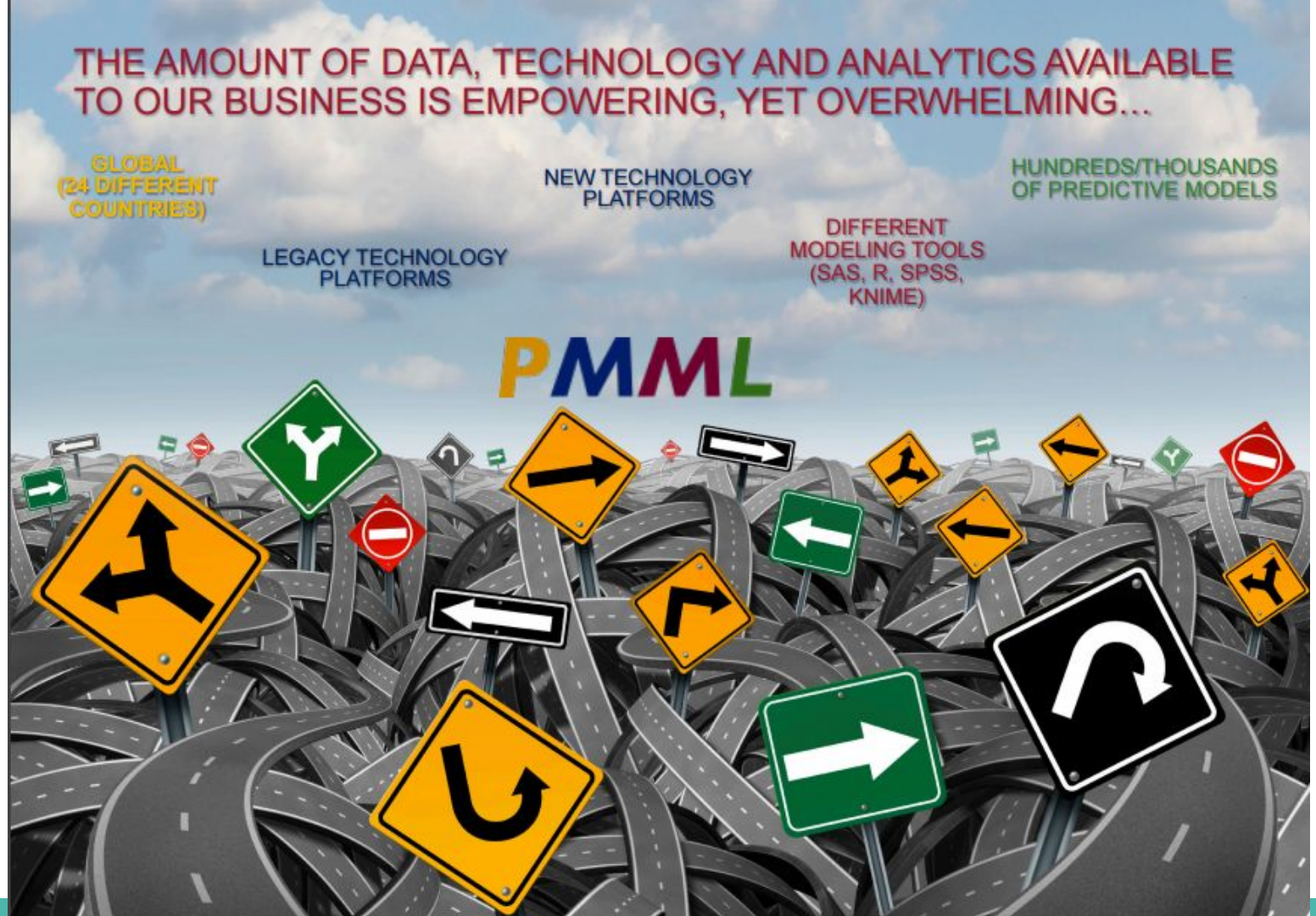
Spark
MLlib

Microsoft
Cognitive Services

mahout

python™
Natural Language Analyses with NLTK

MATLAB


THE AMOUNT OF DATA, TECHNOLOGY AND ANALYTICS AVAILABLE TO OUR BUSINESS IS EMPOWERING, YET OVERWHELMING…
GLOBAL (24 DIFFERENT COUNTRIES)
NEW TECHNOLOGY PLATFORMS
HUNDREDS/THOUSANDS OF PREDICTIVE MODELS
DIFFERENT MODELING TOOLS (SAS, R, SPSS, KNIME)
LEGACY TECHNOLOGY PLATFORMS
PMML

# Standards in Predictive Analytics In the Era of Big and Fast Data

PMML
Predictive Model Markup Language

## WRITE ONCE, RUN ANYWHERE

- PMML
  Predictive Model Standardization
  Developed by DMG, supported by **30 organizations**.

- PFA

# Standards in Predictive Analytics In the Era of Big and Fast Data

- Improve Operational Efficiency & Reduce Time
  - Deploy PMML directly using ADAPA (available in AWS)

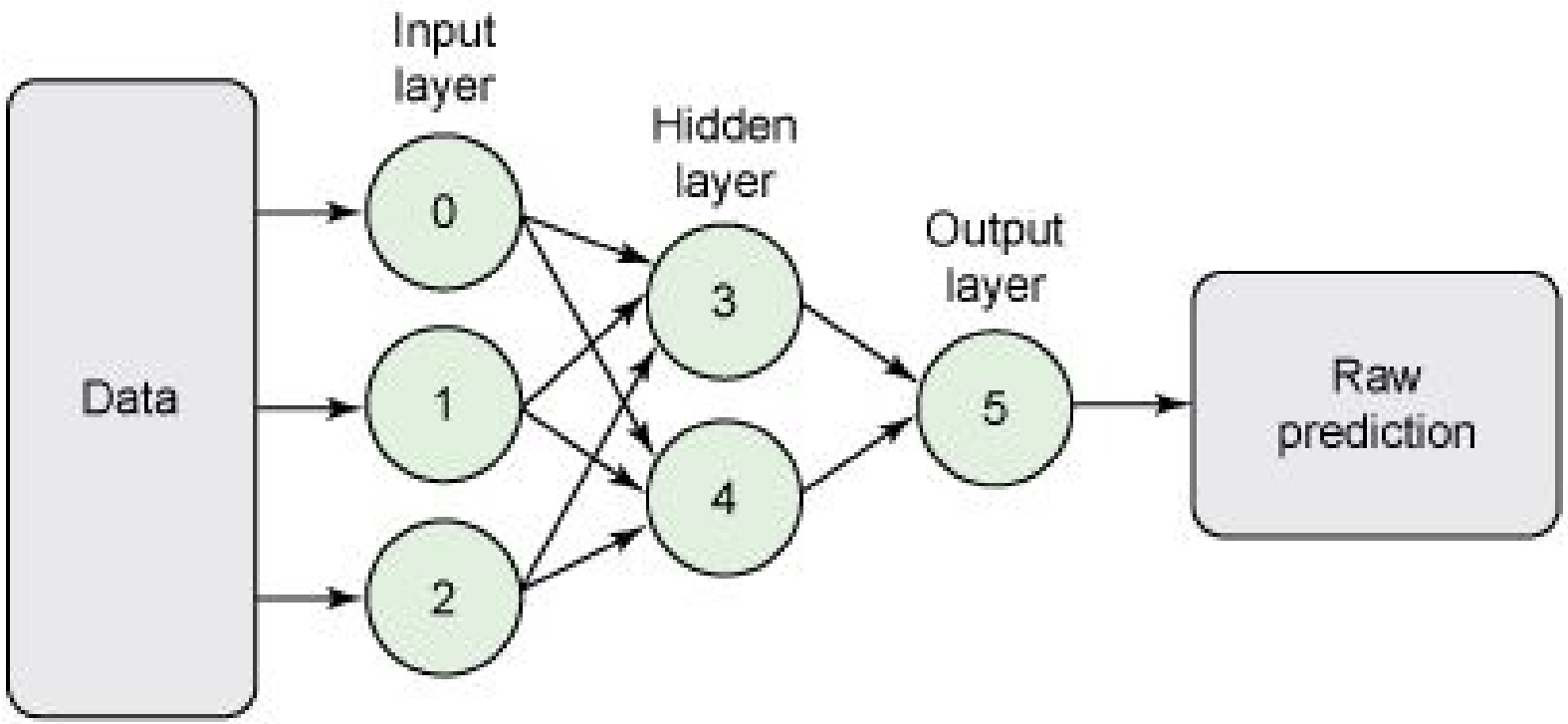
- Greater Flexibility

- Vendor-neutral, Cross-Platform Deployment of Predictive Capabilities

# PMML: Data dictionary

```
<DataDictionary numberOfFields="3">
   <DataField dataType="double" name="Value" optype="continuous">
      <Interval closure="openClosed" rightMargin="60" />
   </DataField>
   <DataField dataType="string" name="Element" optype="categorical">
      <Value property="valid" value="Magnesium" />
      <Value property="valid" value="Sodium" />
      <Value property="valid" value="Calcium" />
      <Value property="valid" value="Radium" />
   </DataField>
   <DataField dataType="double" name="Risk" optype="continuous" />
</DataDictionary>
```

# PMML: Model Definition



# PMML: Model Definition

```
<NeuralLayer numberOfNeurons="2">
    <Neuron id="3" bias="-3.1808306946637">
        <Con from="0" weight="0.119477686963504" />
        <Con from="1" weight="-1.97301278112877" />
        <Con from="2" weight="3.04381251760906" />
    </Neuron>
    <Neuron id="4" bias="0.743161353729323">
        <Con from="0" weight="-0.49411146396721" />
        <Con from="1" weight="2.18588757615864" />
        <Con from="2" weight="-2.01213331163562" />
    </Neuron>
</NeuralLayer>
```

# Uber ATC: Moving from Anomalies to Known Phenomena

UBER

# 1980s: CMU NavLab

- Hand made
- Many bulk sensors
- Racks of bulky computers on board

# 1995: No Hands Across America

- Pittsburgh to LA
- Over 98% autonomously
- Image based sensing
- Lane keeping functionality
- Multi layer perceptron

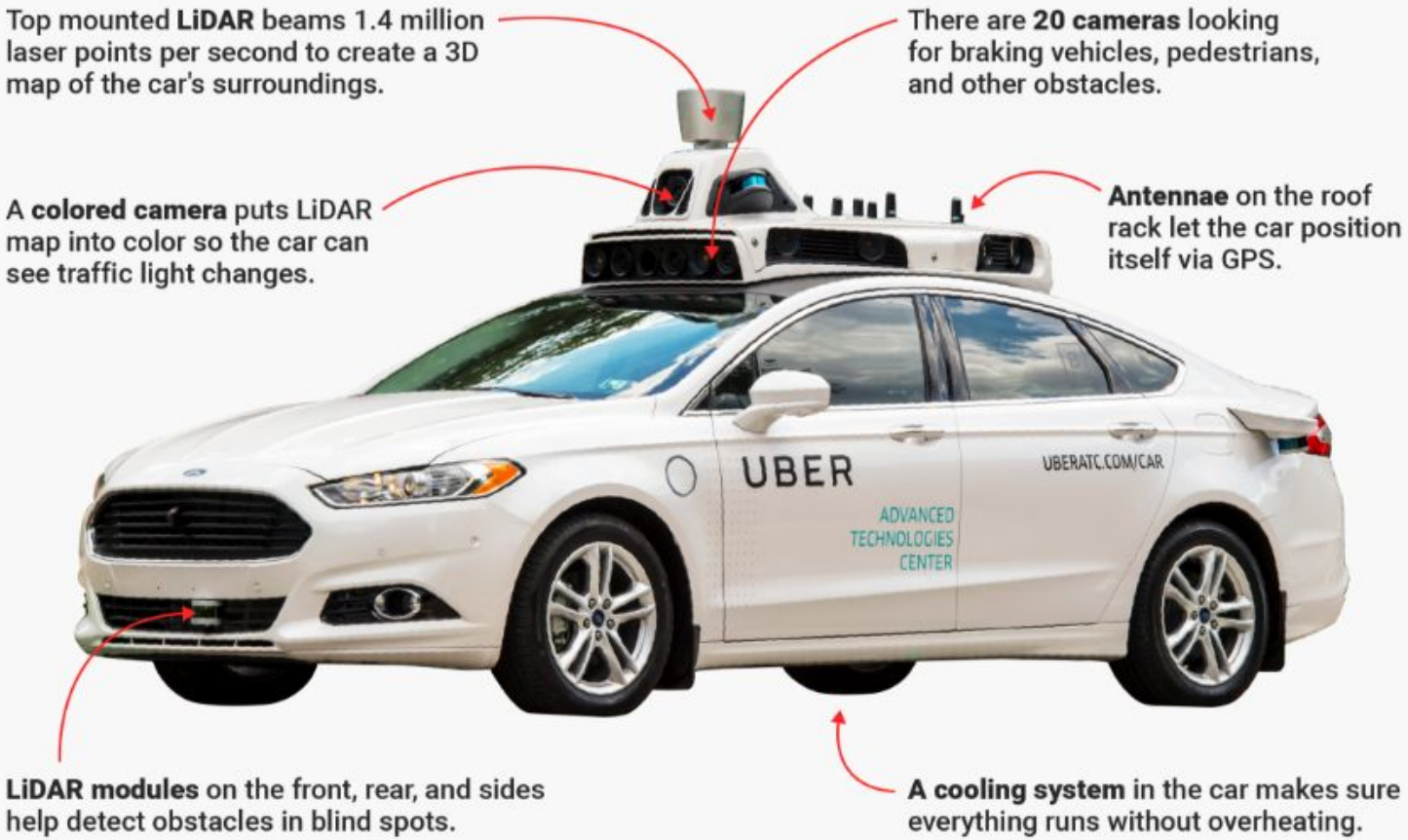
# 2000s: Crusher to APD

- Lidar, cameras
- Sense object statically
- No local map

# 2007: DARPA Urban Challenge

- Fully autonomous driving in urban environment
- Good maps
- Detect other object movement
- Google car project begins based on this project

# Uber Self-Driving Car



# Questions to Answer

Environment

- Has this vehicle encountered anything unusual?
- Do I already know what it is?
- How unusual is it?

# Questions to Answer

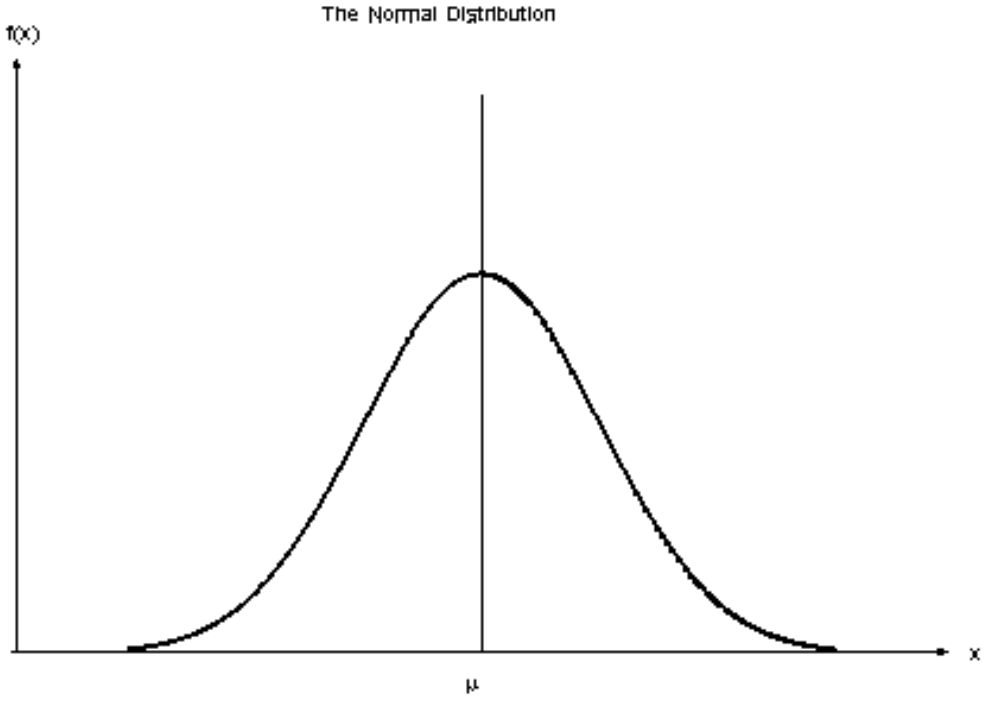
Vehicle

- Has this vehicle done anything unusual?
- Do I already know why?
- Does this affect only this car? Or a whole fleet?

Overall

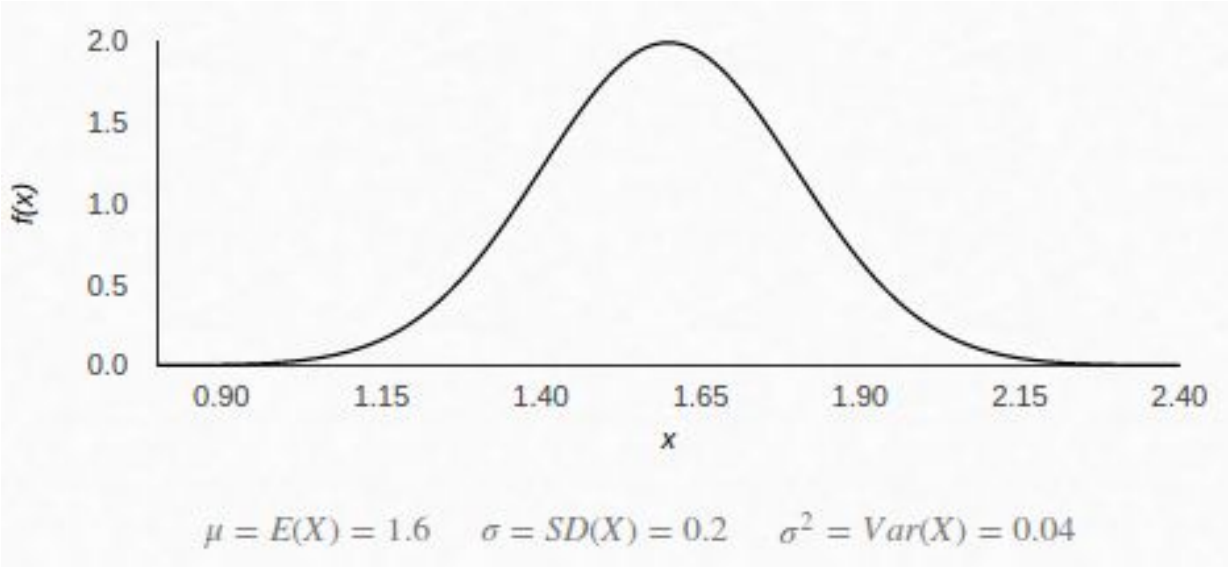
- What is the underlying phenomenon?
- What should I do about it?

# Basic Anomaly Detection




1. Learn probability distribution over typical data points
2. Evaluate the likelihood of points of interests
3. Flag those with low likelihood as “anomalous”

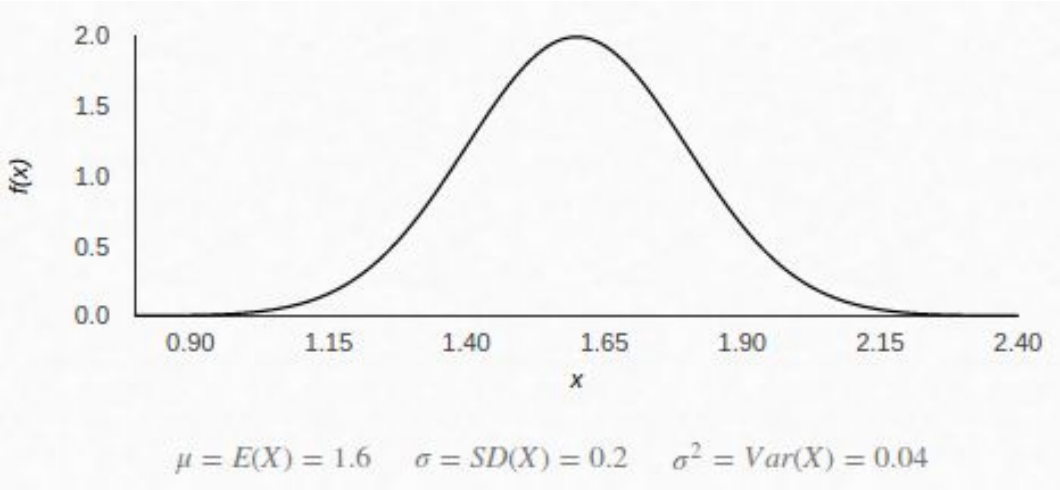
# Basic Anomaly Detection




$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$

# Basic Anomaly Detection




1. New data, A and B
   A, height = 1.4 meter
   B, height = 2 meter

2. Calculate f(A) and f(B)
   f(A) = 1.21
   f(B) = 0.27

3. Anomaly if f(X) < *e*,
   *e* = 0.4
   A is normal
   B is anomaly

# KDD 2017

## Halifax, Nova Scotia - Canada

**August 13 - 17, 2017**

# Thank You!

# Q&A

# References


- KDD 2016
- https://homes.cs.washington.edu/~marcotcr/
- http://deepbeat.org/
- http://www.acsu.buffalo.edu/~qli22/
- https://www.youtube.com/watch?v=WaZ0EL3E7XY&t=1s
- http://www.ruizhang.info/publications/KDD_2016_intent_tracking_slides.pdf