



SOCIETY OF ACTUARIES



66 PD Advanced Analytics: Building Your Toolbox

**Moderator:**

Trevor Fast, FSA, FCA, MAAA

**Presenters:**

Jeremy Dylan Achin

Jeff T. Heaton

Syed Muzayan Mehmud, ASA, MAAA, MAAA

Trevor James Fast, FSA, FCA, MAAA

Advanced Analytics: Building Your Toolbox

# Predicting Diabetes Using Gradient Boosted Tree Models

Jeremy Achin

Data Scientist @ DataRobot

[jeremy@datarobot.com](mailto:jeremy@datarobot.com)

June 24, 2014

# Outline

1. Health Analytics: Predicting  
Diabetes

1. Introduction to Gradient  
Boosting Machines (GBM)

# Predicting Diabetes

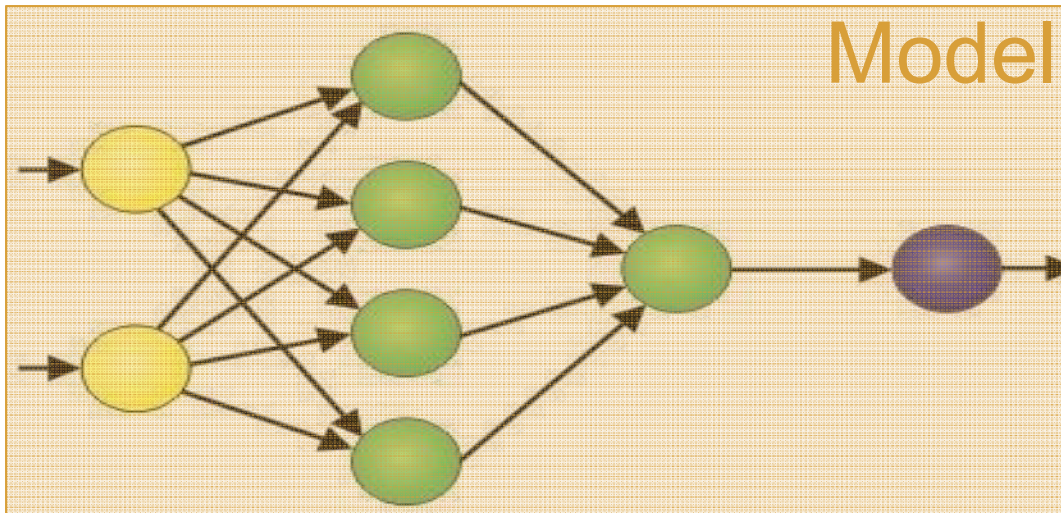


kaggle



**Weight**

**Age**








**Probability of  
Diabetes**



# 9,948 Patients










## SyncPatient

 PatientGuid	UNIQUEID	NOT NULL
 Gender	NVARCHAR(1)	NOT NULL
 YearOfBirth	SMALLINT	NOT NULL
 State	NVARCHAR(2)	NOT NULL
 PracticeGuid	UNIQUEID	NOT NULL



# 78,864 Prescriptions











## SyncPrescripti

 PrescriptionGuid	UNIQUEID	NOT NULL
 PatientGuid (FK)	UNIQUEID	NOT NULL
 MedicationGuid (FK)	UNIQUEID	NULL
 PrescriptionYear	SMALLINT	NULL
 Quantity	NVARCHAR(50)	NOT NULL
 NumberOfRefills	NVARCHAR(50)	NULL
 RefillsNeeded	BIT	NULL
 GenericAllowed	BIT	NULL
 UserGuid	UNIQUEID	NOT NULL

W

# 66,487 Medications


















SyncMedication

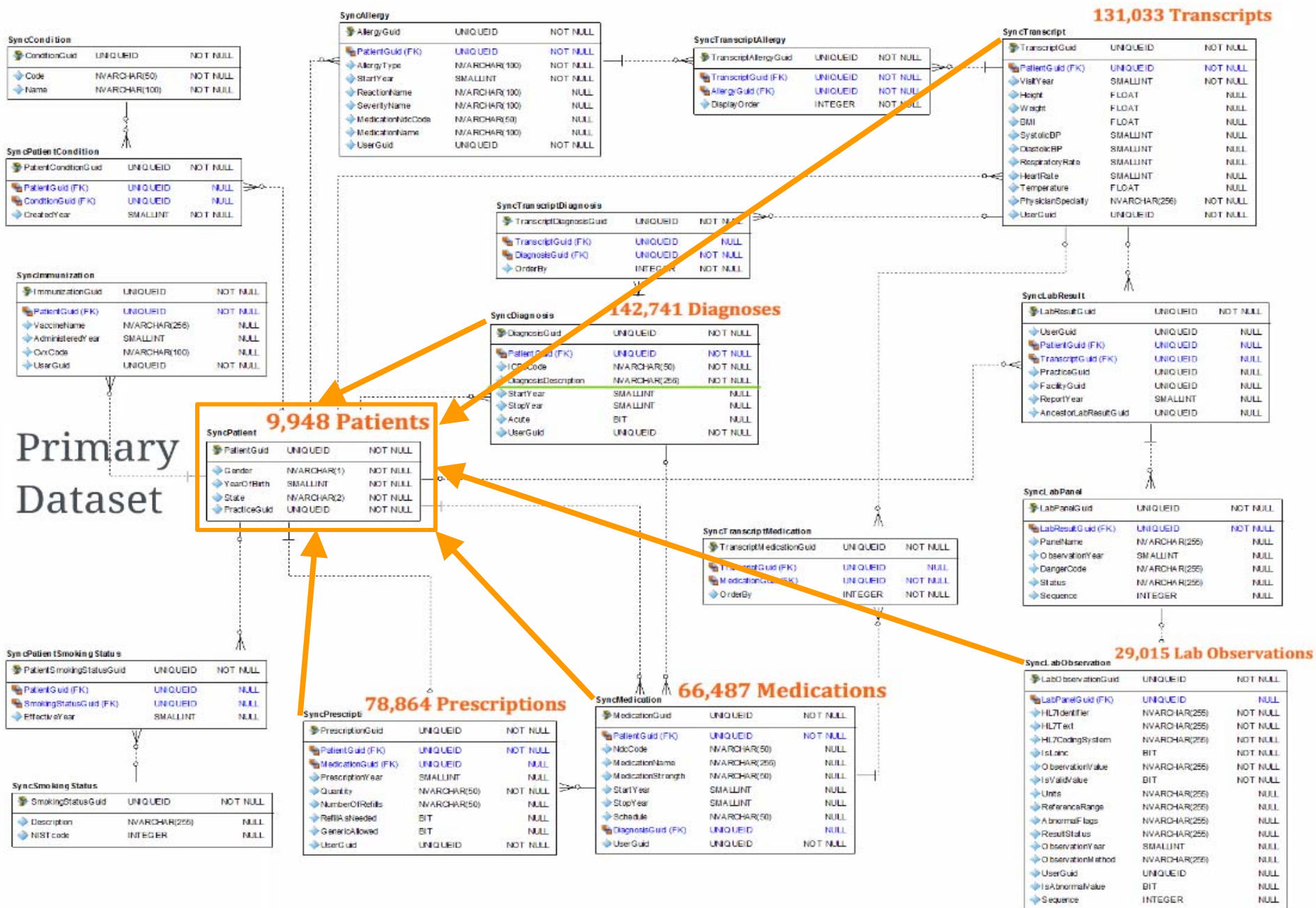
 MedicationGuid	UNIQUEID	NOT NULL
 PatientGuid (FK)	UNIQUEID	NOT NULL
 NdcCode	NVARCHAR(50)	NULL
 MedicationName	NVARCHAR(256)	NULL
 MedicationStrength	NVARCHAR(50)	NULL
 StartYear	SMALLINT	NULL
 StopYear	SMALLINT	NULL
 Schedule	NVARCHAR(50)	NULL
 DiagnosisGuid (FK)	UNIQUEID	NULL
 UserGuid	UNIQUEID	NOT NULL





# 29,015 Lab Observations

SynCLabObservation

 LabObservationGuid	UNIQUEID	NOT NULL
 LabPanelGuid (FK)	UNIQUEID	NULL
 HL7Identifier	NVARCHAR(255)	NOT NULL
 HL7Text	NVARCHAR(255)	NOT NULL
 HL7CodingSystem	NVARCHAR(255)	NOT NULL
 IsLoinc	BIT	NOT NULL
 ObservationValue	NVARCHAR(255)	NOT NULL
 IsValidValue	BIT	NOT NULL
 Units	NVARCHAR(255)	NULL
 ReferenceRange	NVARCHAR(255)	NULL
 AbnormalFlags	NVARCHAR(255)	NULL
 ResultStatus	NVARCHAR(255)	NULL
 ObservationYear	SMALLINT	NULL
 ObservationMethod	NVARCHAR(255)	NULL
 UserGuid	UNIQUEID	NULL
 IsAbnormalValue	BIT	NULL
 Sequence	INTEGER	NULL



**Good Model**  
**Log Loss: 0.41**

103	↓13	Jason Karpeles			
104	↓13	nlubchenco			
105	↓13	Zach A.			
106	↓13	Opensandwich			
107	↓13	Vincent Wong	<a href="#">0.40506</a>	12	Thu, 09 Aug 2012 02:57:57 (-3d)
108	↓12	datakore	<a href="#">0.40555</a>	2	Sun, 05 Aug 2012 17:27:50 (-4.9d)
		<b>Random Forest Benchmark</b>	0.40559		
110	↓12	InvincibleGuy	<a href="#">0.40559</a>	1	Sat, 14 Jul 2012 00:49:32
111	↓12	anything_you_like 	<a href="#">0.40559</a>	2	Sun, 15 Jul 2012 15:45:24 (-24.1h)
112	↓12	skrecok	<a href="#">0.40559</a>	3	Tue, 24 Jul 2012 09:45:57 (-6.8d)
113	↓12	Seung Hyup Hyun	<a href="#">0.40559</a>	1	Thu, 23 Aug 2012 01:07:55
114	new	Odile-the-shrew	<a href="#">0.40559</a>	2	Thu, 06 Sep 2012 15:25:56 (-0h)
115	↑6	HairyFotr	<a href="#">0.40559</a>	9	Fri, 07 Sep 2012 00:15:38
116	↓13	ethan	<a href="#">0.40568</a>	6	Mon, 06 Aug 2012 14:19:27 (-39.3h)

**(109th Place)**



#	Δ1w	Team Name <small>* in the money</small>	Score <small>?</small>	Entries	Last Submission
1	↑3	J.A. Guerrero *	<a href="#">0.31490</a>	58	Mon, 10 Sep 2012 19:00:40 (-6.9d)
2	↓1	__mtb__ *	<a href="#">0.31778</a>	42	Mon, 10 Sep 2012 20:01:18 (-29.5h)
3	↑3	An apple a day *	<a href="#">0.32426</a>	80	Mon, 10 Sep 2012 11:43:29 (-35.7d)
4	↓2	DataRobot	<a href="#">0.32457</a>	53	Mon, 10 Sep 2012 11:03:19
5	-	Indy Actuaries	<a href="#">0.32871</a>	13	Mon, 10 Sep 2012 20:43:38
6	↓3	n_m	<a href="#">0.33274</a>	32	Fri, 10 Sep 2012 20:43:38
7	↑7	kg	<a href="#">0.33836</a>	23	
8	-	Diogenes Club	<a href="#">0.33929</a>	27	



**Great Model**  
Log Loss: 0.32

(4th place)

# Difference between Good and Great

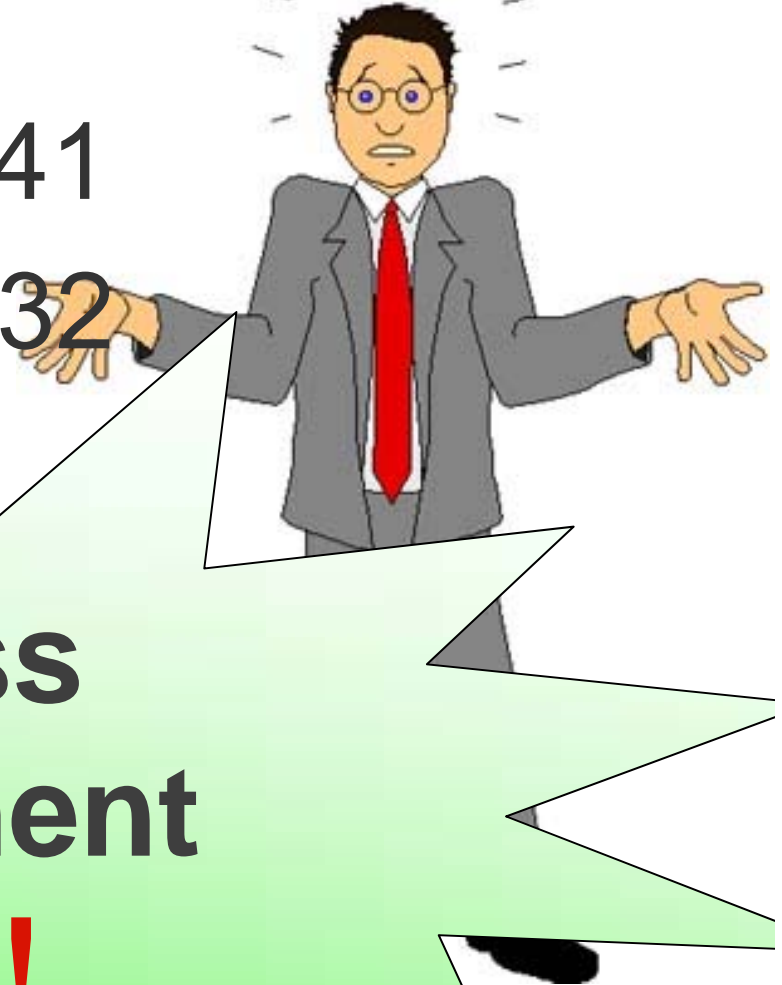
- Good: Log Loss = 0.41
- Great: Log Loss = 0.32



**Log Loss  
Improvement  
= 0.09!!!**

# Difference between Good and Great

- Good: Log Loss = 0.41
- Great: Log Loss = 0.32



**Log Loss  
Improvement  
= 0.09!!!**



# Meaningful Model Comparison

Great  
Prediction

Good  
Prediction

Ground  
Truth

A - B

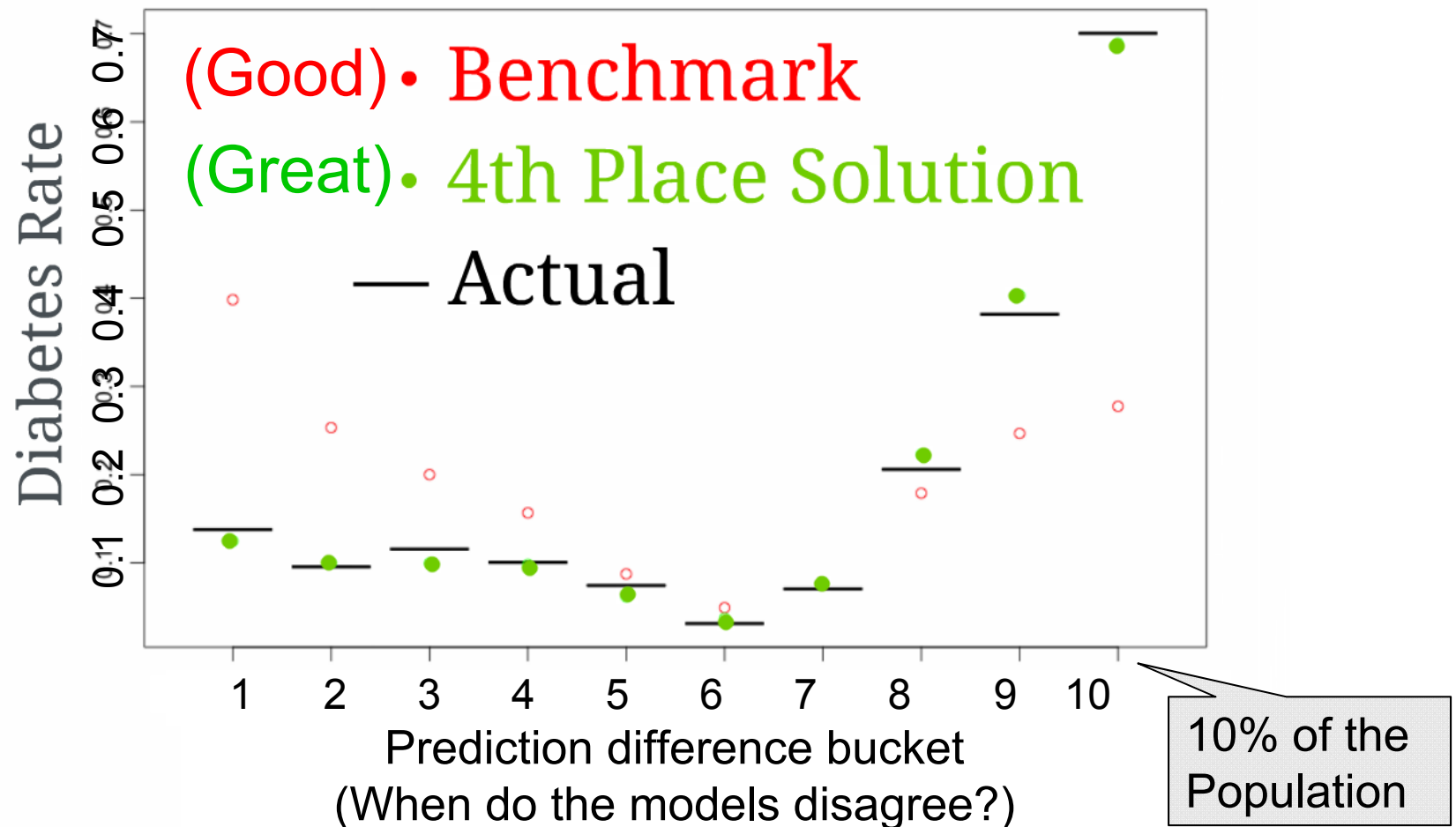
	A	B	C	D	F	G	H	I
1	Model 1	Model 2	Actual	pred_diff	Bucket		Total Obs	9948
2	0.0930961	0.97076126	0	-0.8776652	1		buckets	10
3	0.04639884	0.881745082	0	-0.8353462	1		Obs per bucket	994.8
4	0.05531521	0.864625378	0	-0.8093102	1			
5	0.13676324	0.943655042	0	-0.8068918	1			
6	0.02168173	0.823185683	0	-0.8015039	1			
7	0.03684564	0.81898334	1	-0.7821377	1			
8	0.17283439	0.954203549	0	-0.7813692	1			
9	0.18117228	0.960672761	1	-0.7795005	1			
10	0.17332438	0.946835886	0	-0.7735115	1			
11	0.20196339	0.962739608	0	-0.7607762	1			
12	0.19408966	0.952755667	0	-0.758666	1			
13	0.07146744	0.827419322	0	-0.7559519	1			
14	0.13819089	0.892337068	0	-0.7541462	1			
15	0.17846479	0.929543084	1	-0.7510783	1			
16	0.11681455	0.861886648	1	-0.7448721	1			

Sort &  
Bucket

When do the models disagree?

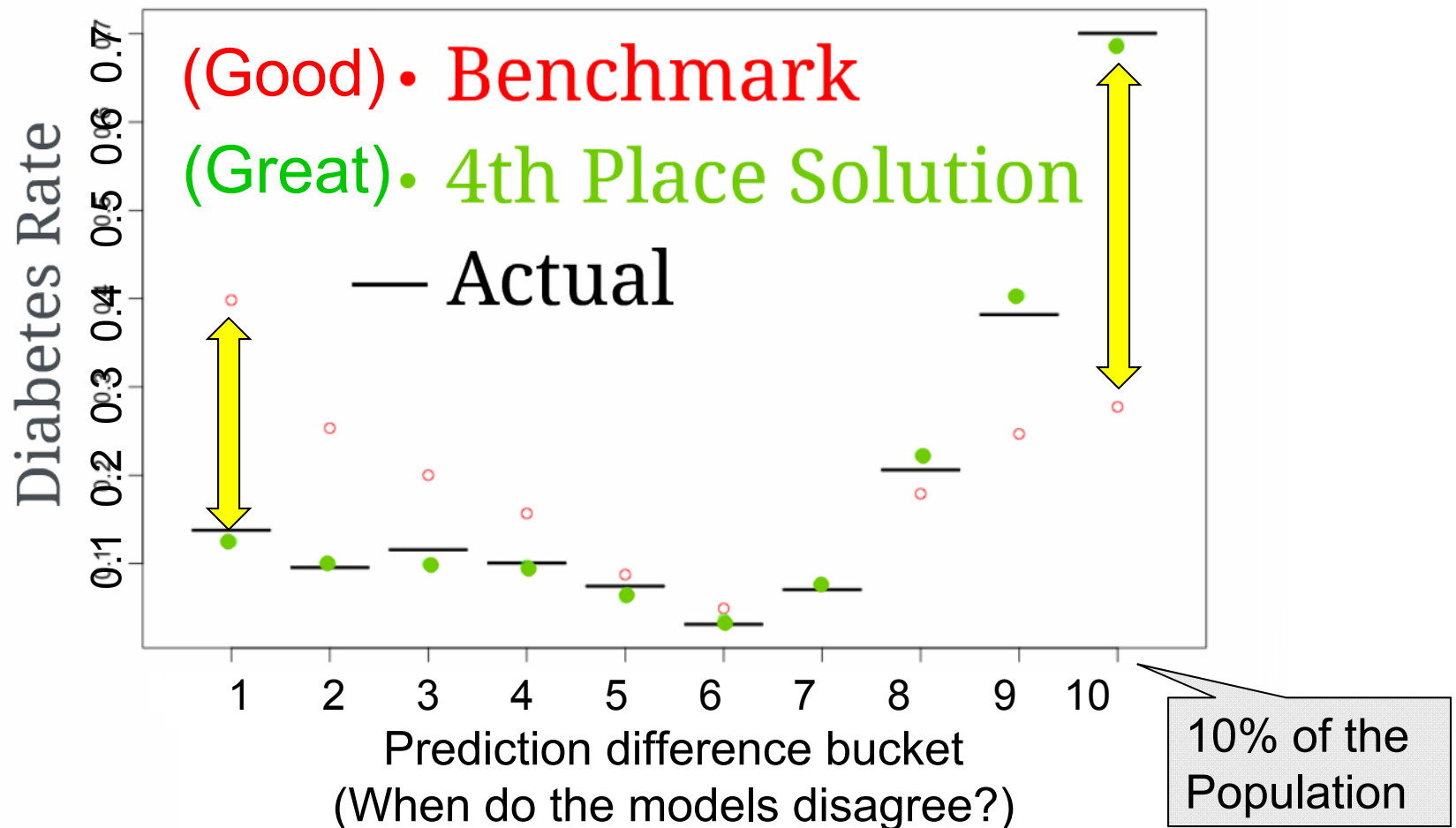
# Meaningful Model Comparison

## Dual Lift (10 Buckets)



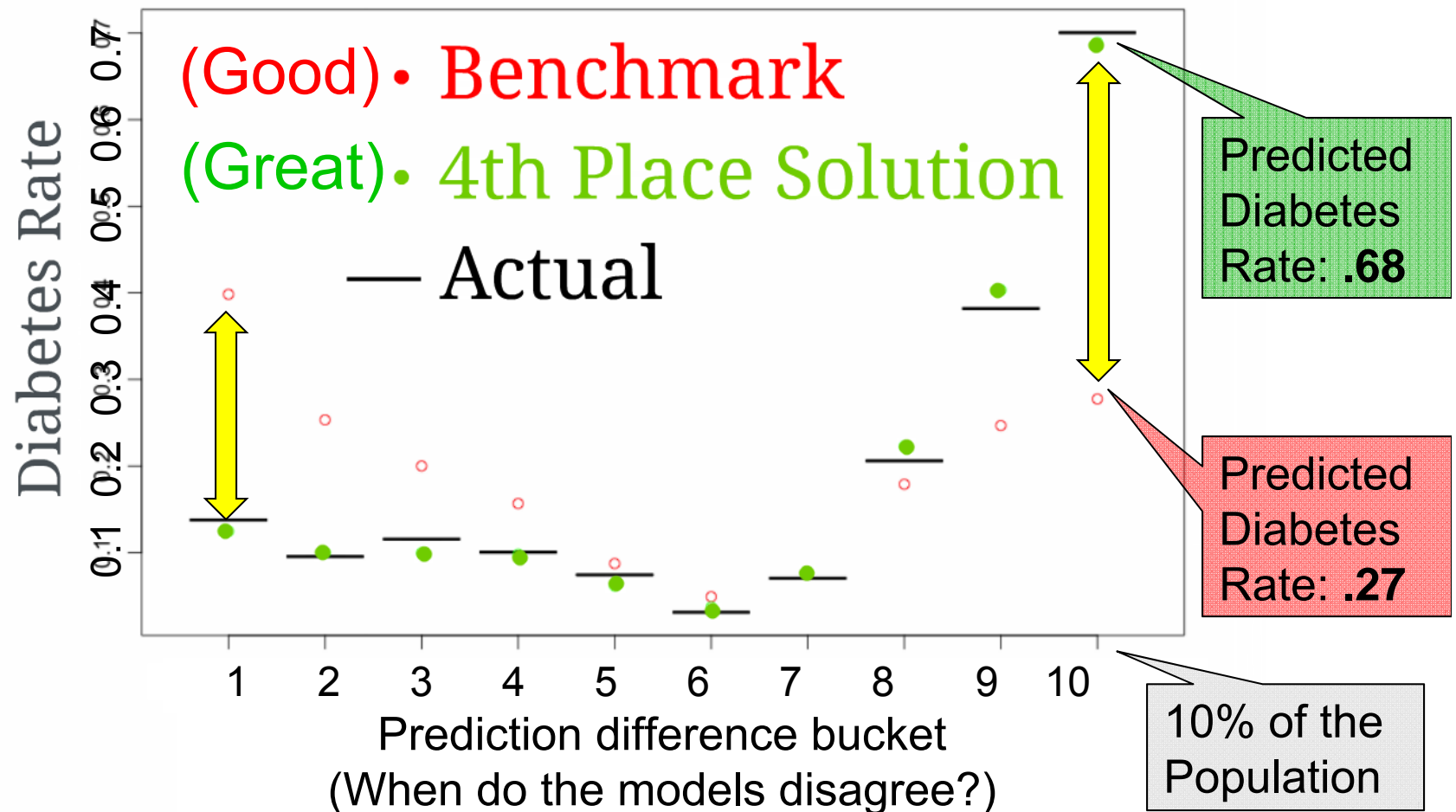
# Meaningful Model Comparison

## Dual Lift (10 Buckets)

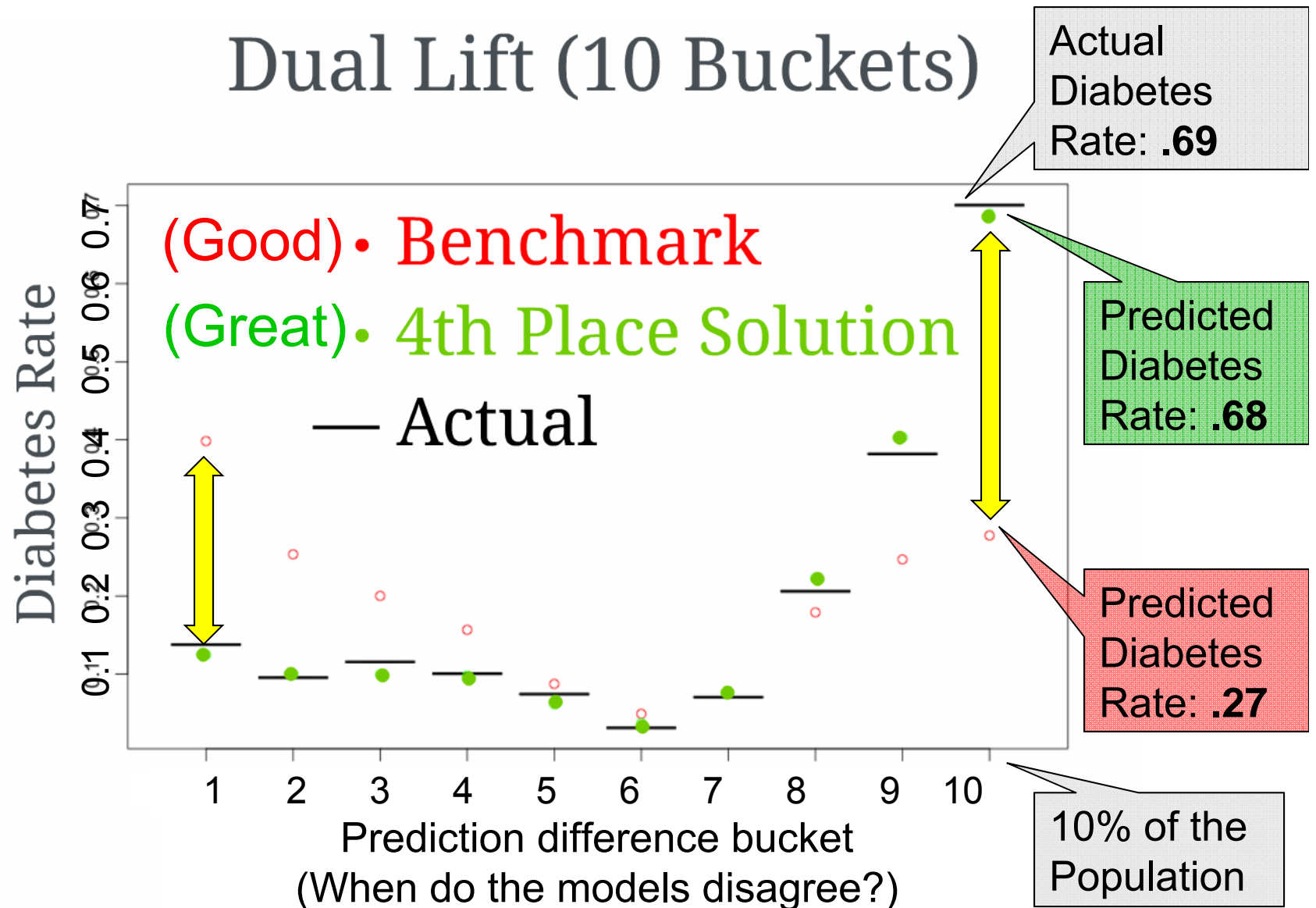


# Meaningful Model Comparison

## Dual Lift (10 Buckets)

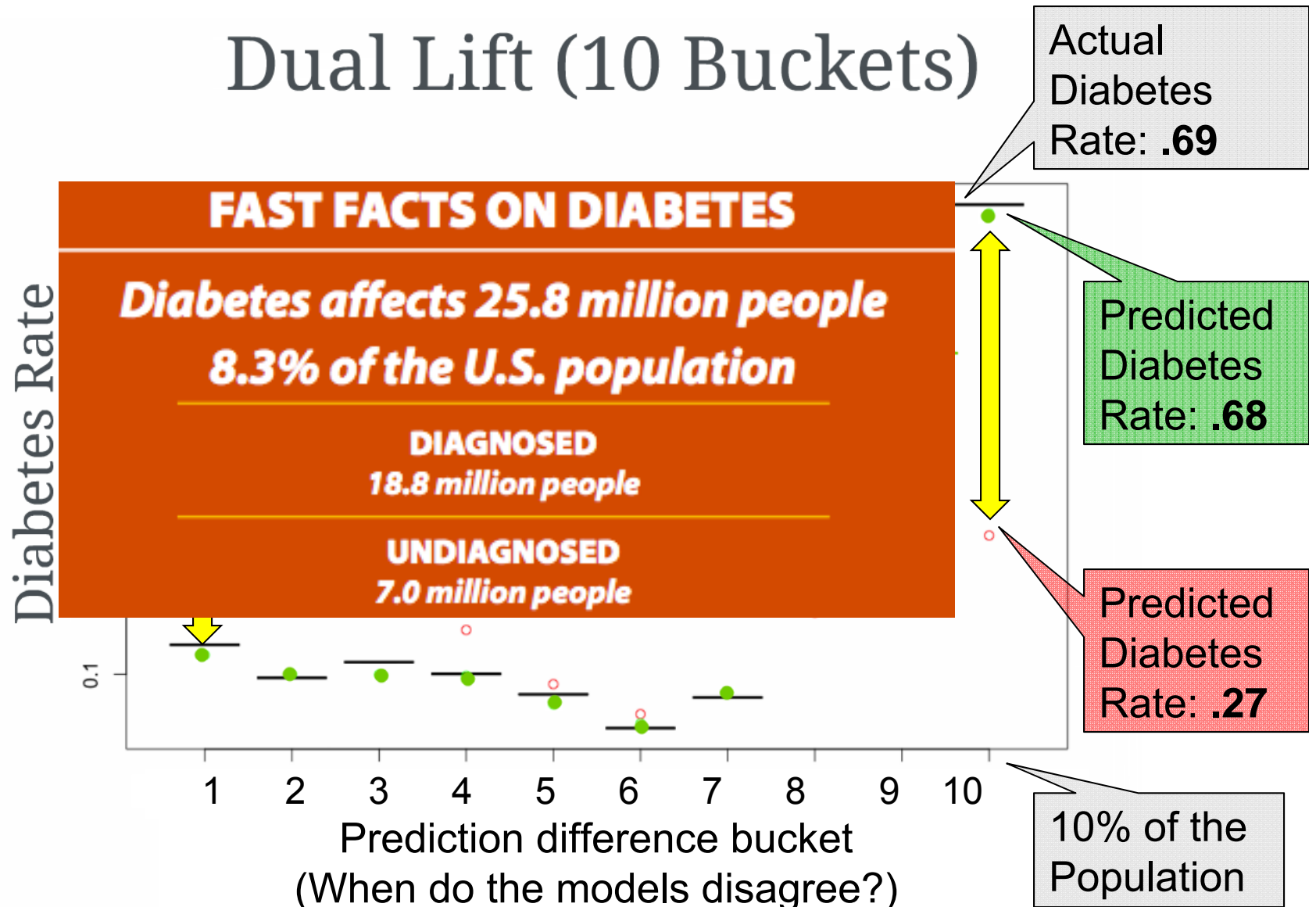


# Meaningful Model Comparison



# Meaningful Model Comparison

## Dual Lift (10 Buckets)





# Winning Diabetes Algorithms

- First Place (Jose A. Guerrero):

Boosted Trees (GBM) + Random Forest

- Second Place (Matt Berseth):

Boosted Trees (GBM)

- Third Place (Shashishekhar Godbole):

Boosted Trees (GBM) + Random Forest +  
+ Neural Networks

- Fourth Place (DataRobot):

Boosted Trees (GBM) + Support Vector Machine +  
+ Generalized Linear Mixed Model + Random  
Forest

# Winning Diabetes Algorithms

- First Place (Jose A. Guerrero):

**Boosted Trees (GBM)** + Random Forest

- Second Place (Matt Berseth):

**Boosted Trees (GBM)**

- Third Place (Shashishekhar Godbole):

**Boosted Trees (GBM)** + Random Forest +  
+ Neural Networks

- Fourth Place (DataRobot):

**Boosted Trees (GBM)** + Support Vector Machine +  
+ Generalized Linear Mixed Model + Random  
Forest

**Why does GBM work so well?**

# Boosted Trees (GBM)

- Single most powerful algorithm out there right now due to the way it automatically captures **nonlinearity** and **interactions**.

# Boosted Trees (GBM)

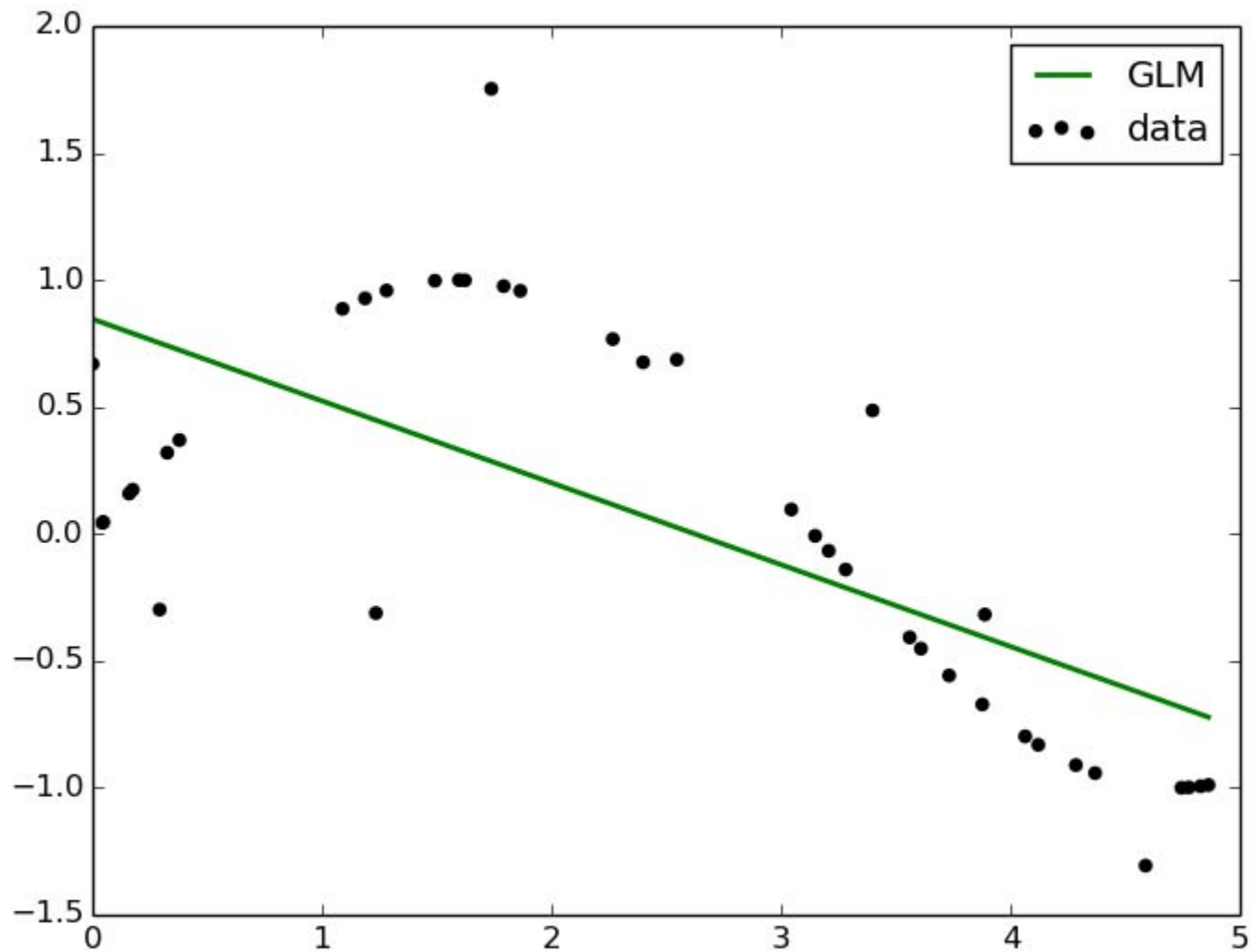
- Single most powerful algorithm out there right now due to the way it automatically captures **nonlinearity** and **interactions**.

AND

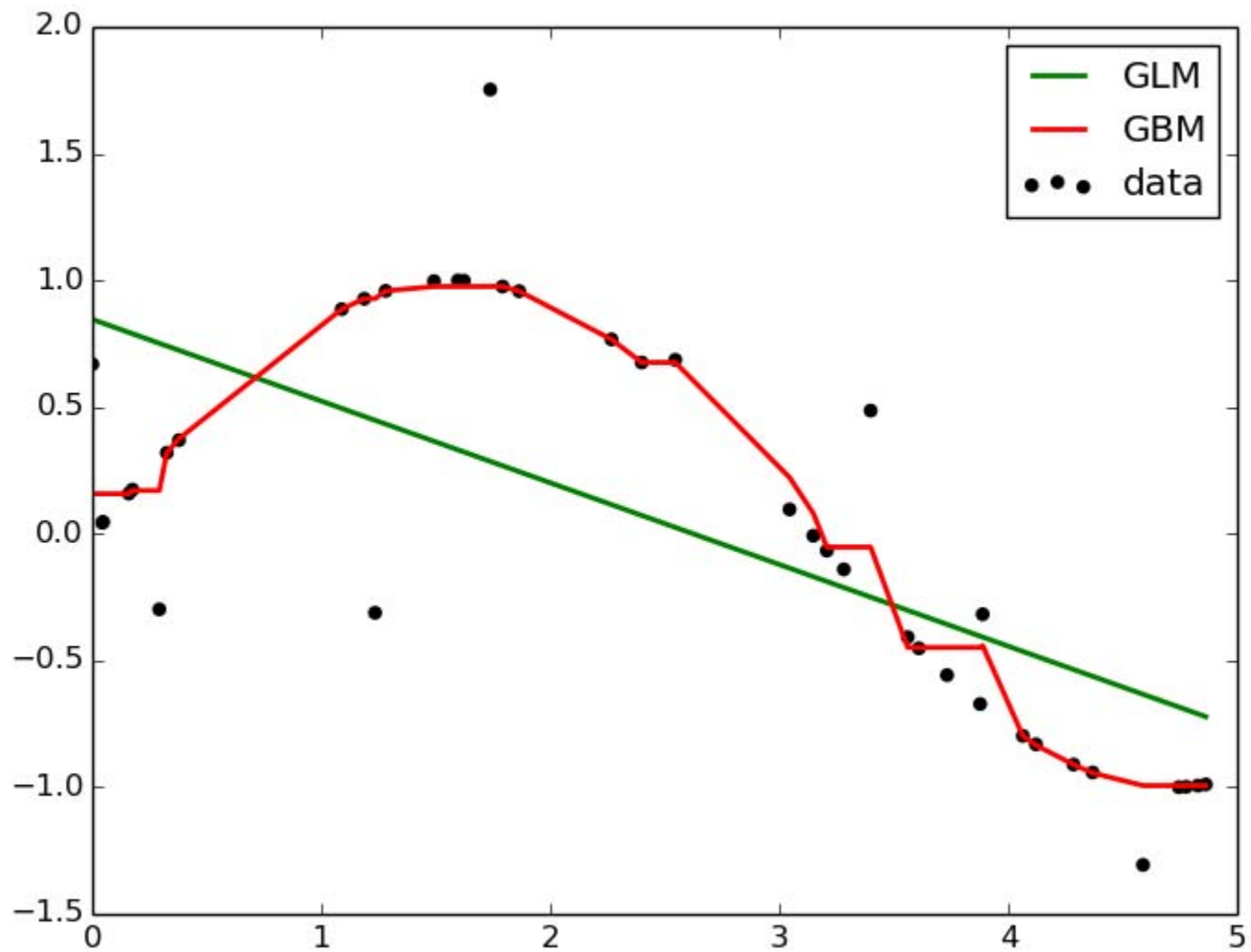




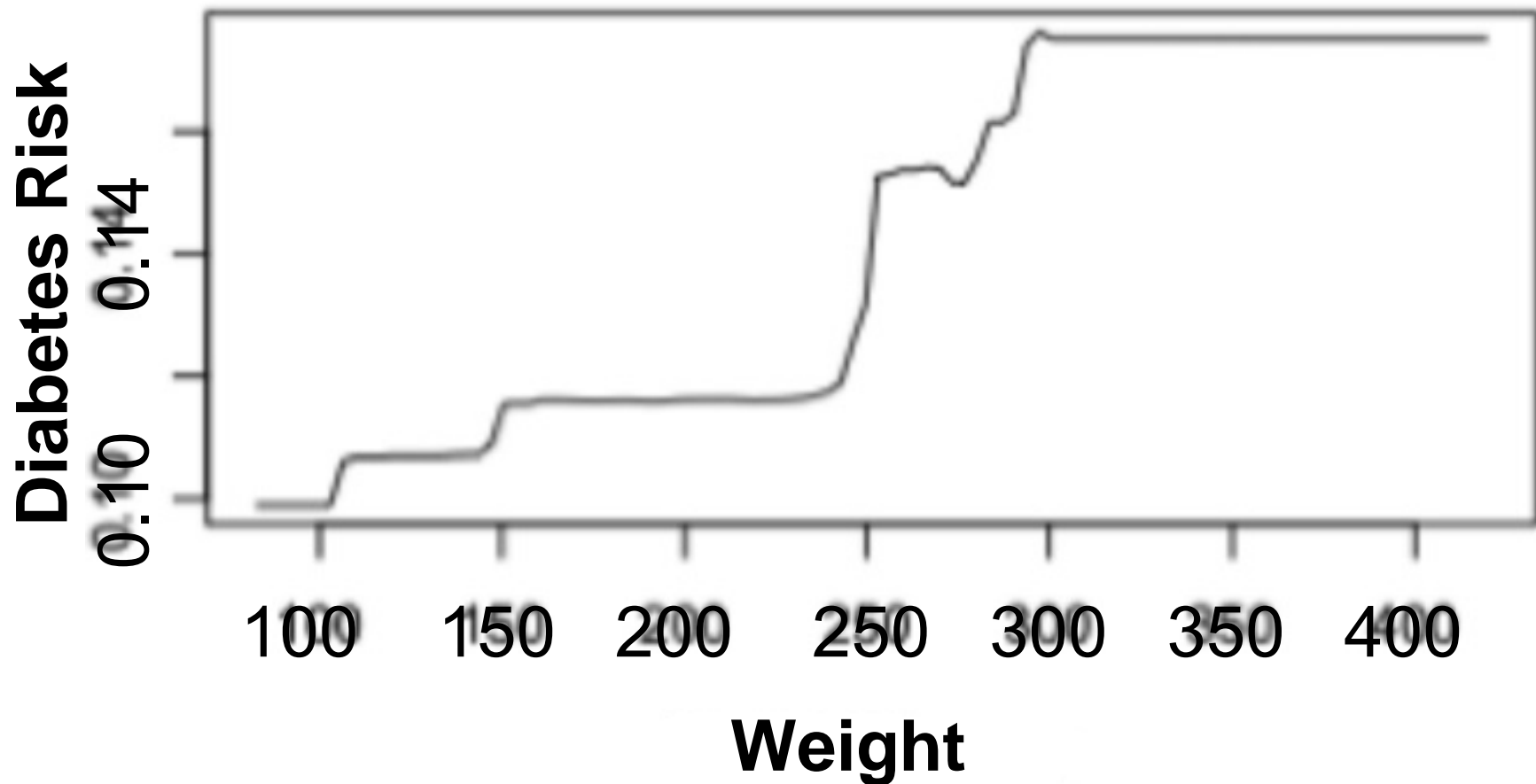
# Nonlinearity



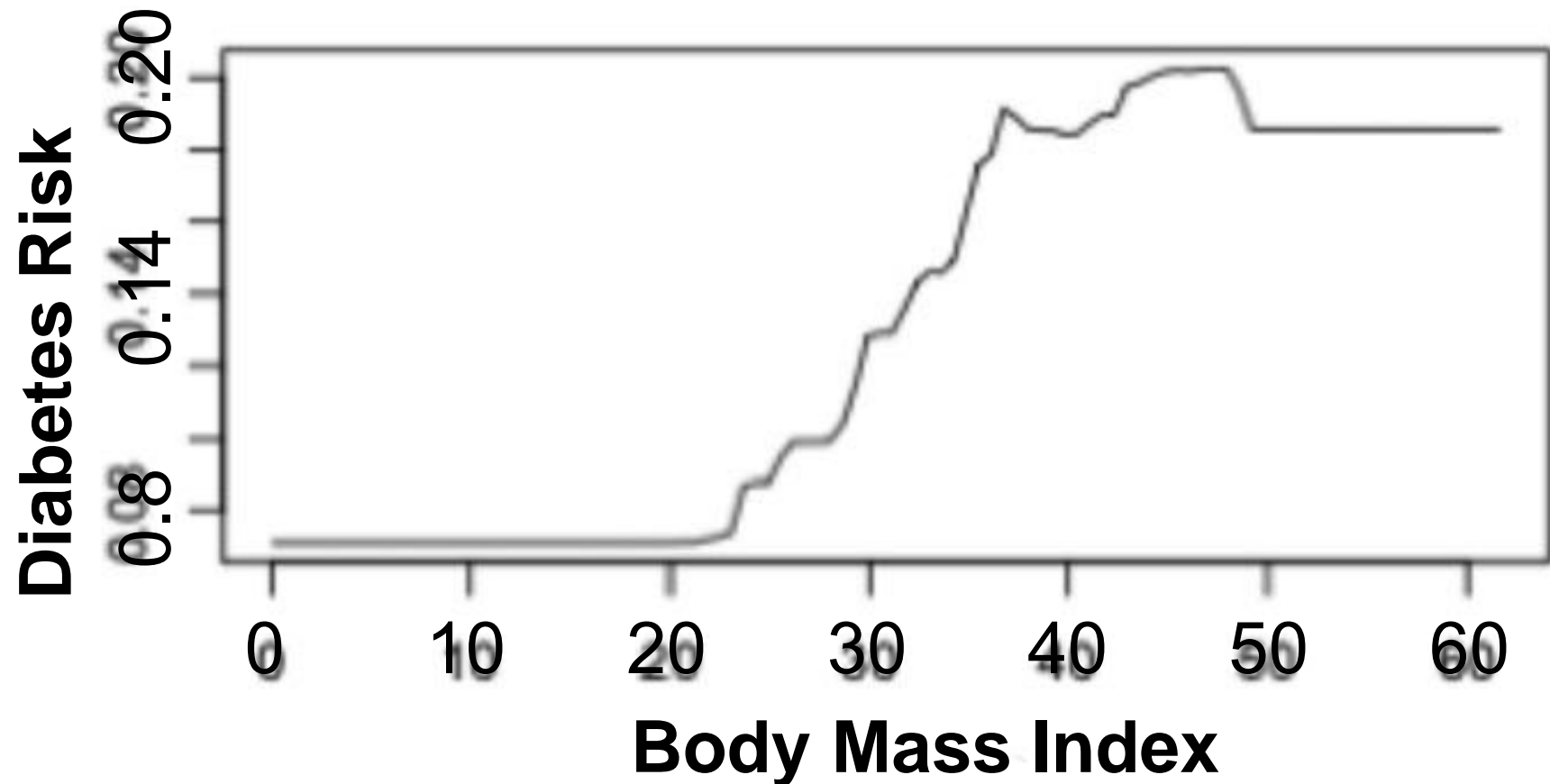
# Nonlinearity



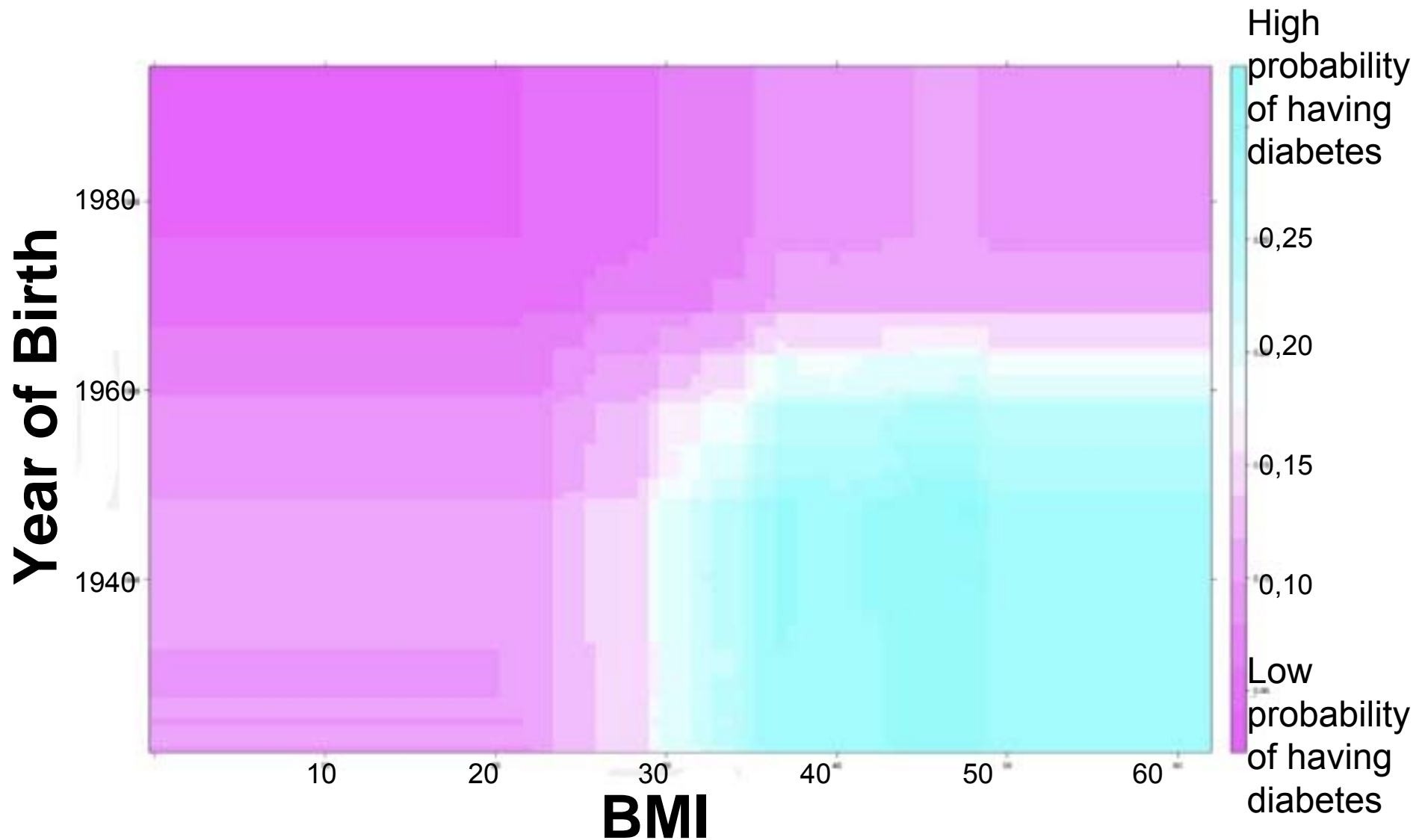
# GBM Capturing Nonlinear Effect of Weight on Diabetes Risk



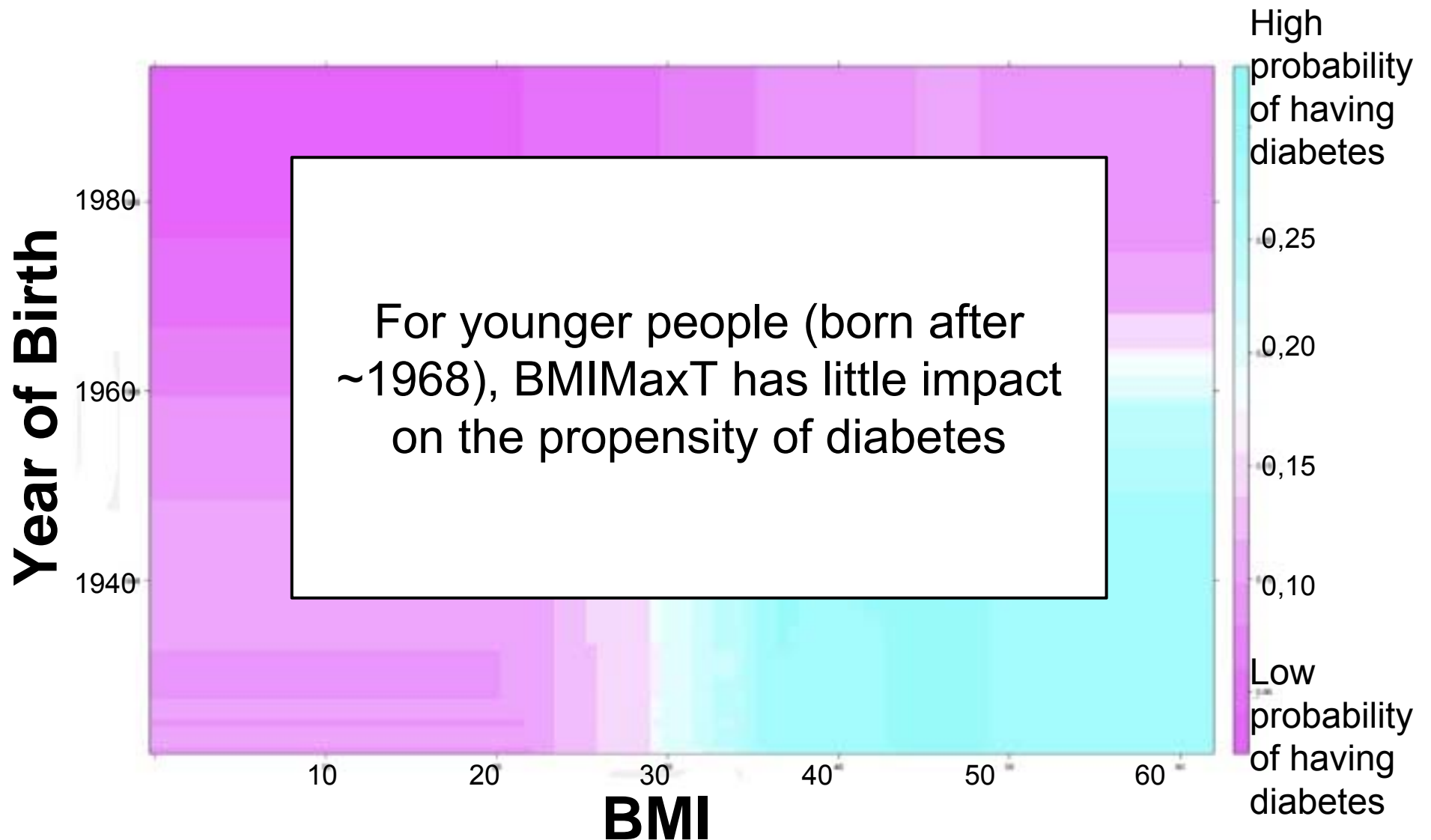
# GBM Capturing Nonlinear Effect of BMI on Diabetes Risk



# GBM Automatically Detects Feature Interactions



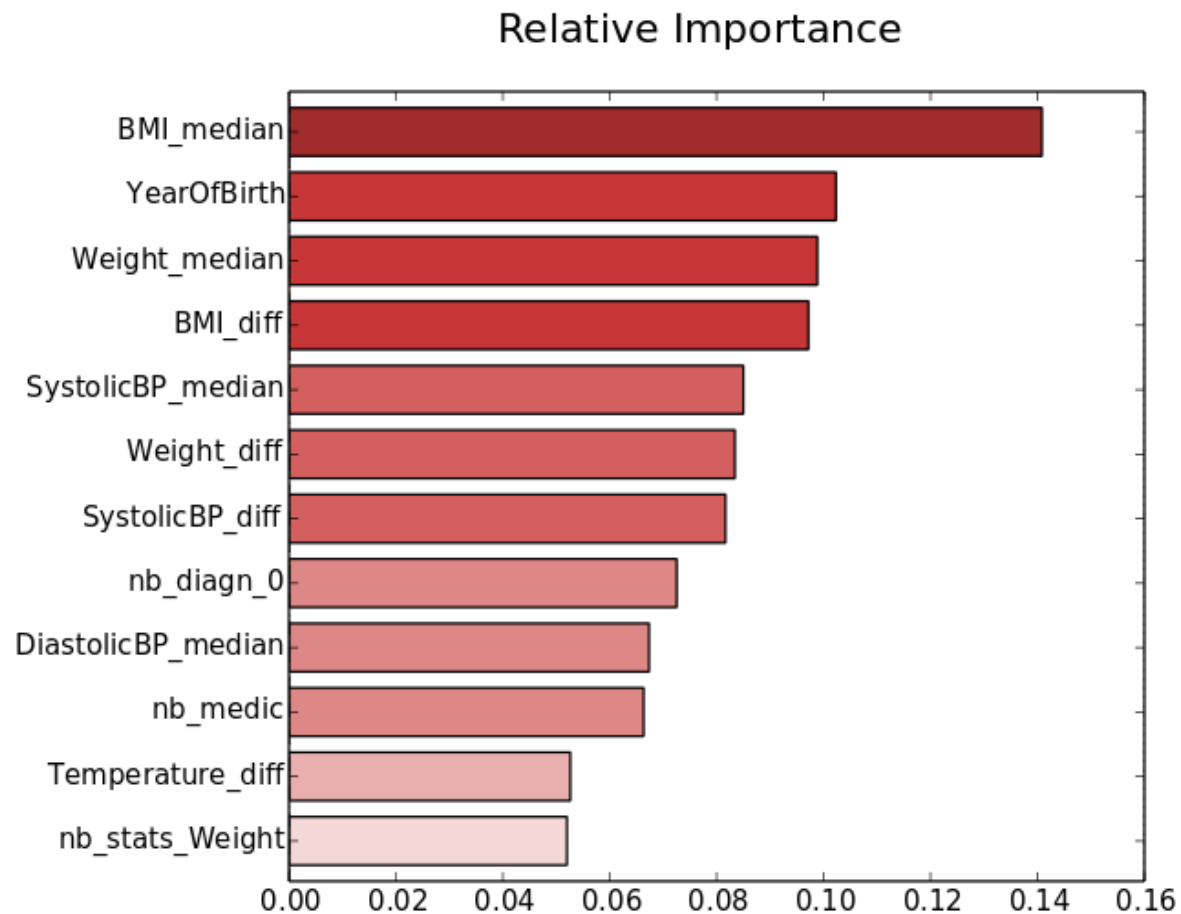
# GBM Automatically Detects Feature Interactions





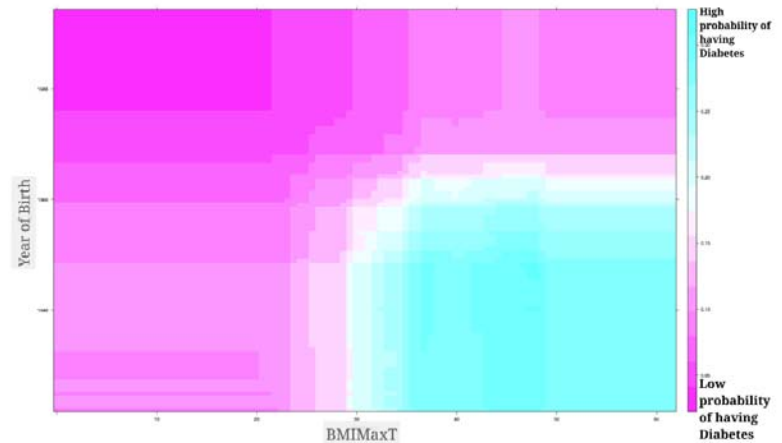
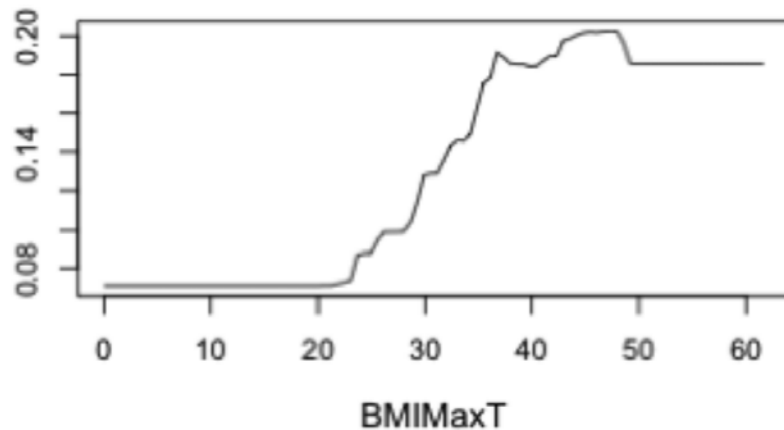
# Model Interpretation

- What are the most important variables?



# Model Interpretation

- How do features interact with the response?
  - Partial dependence plots



# Other GBM Success Stories

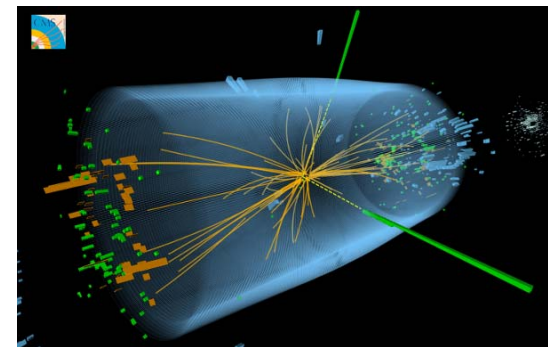
Web page ranking  
Google, Bing, Yandex



Data Mining Competitions  
Health, Energy, Pharma,  
Advertising, E-commerce, ...



Research  
GBM @ CERN



**How does GBM  
work?**

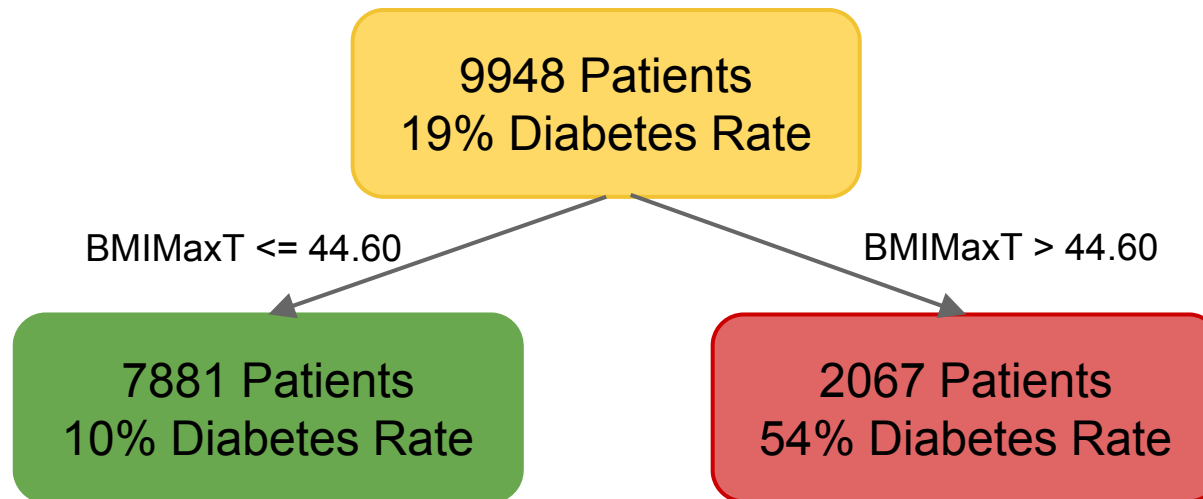
# Decision Trees

- Recursive partitioning

9948 Patients  
19% Diabetes Rate

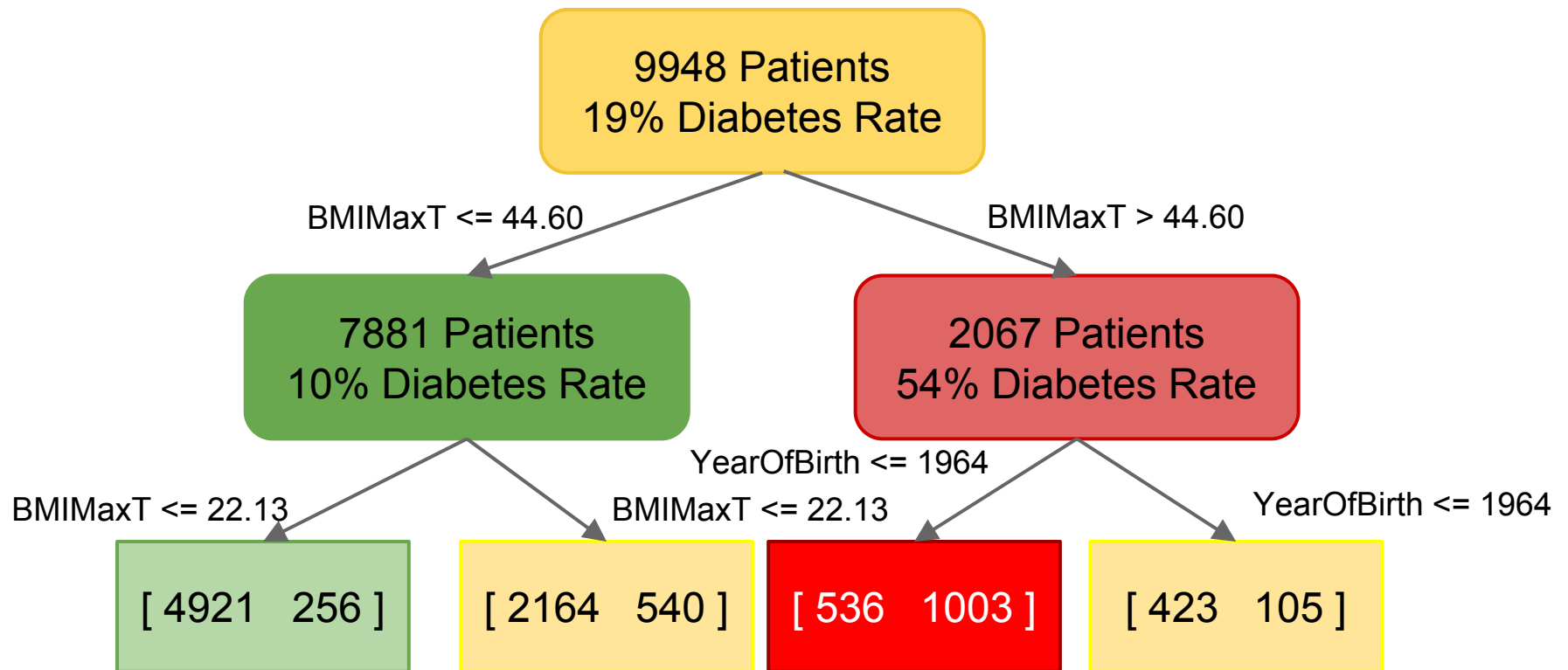
# Decision Trees

- Recursive partitioning



# Decision Trees

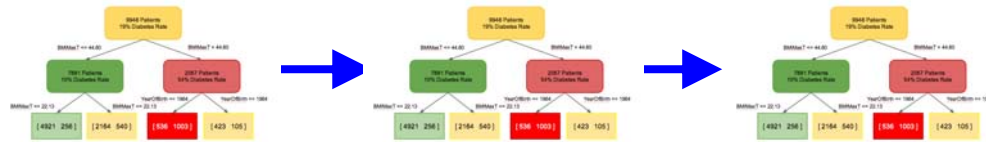
- Recursive partitioning





# Gradient Boosting

- Tree ensemble fitted in a forward stage-wise manner

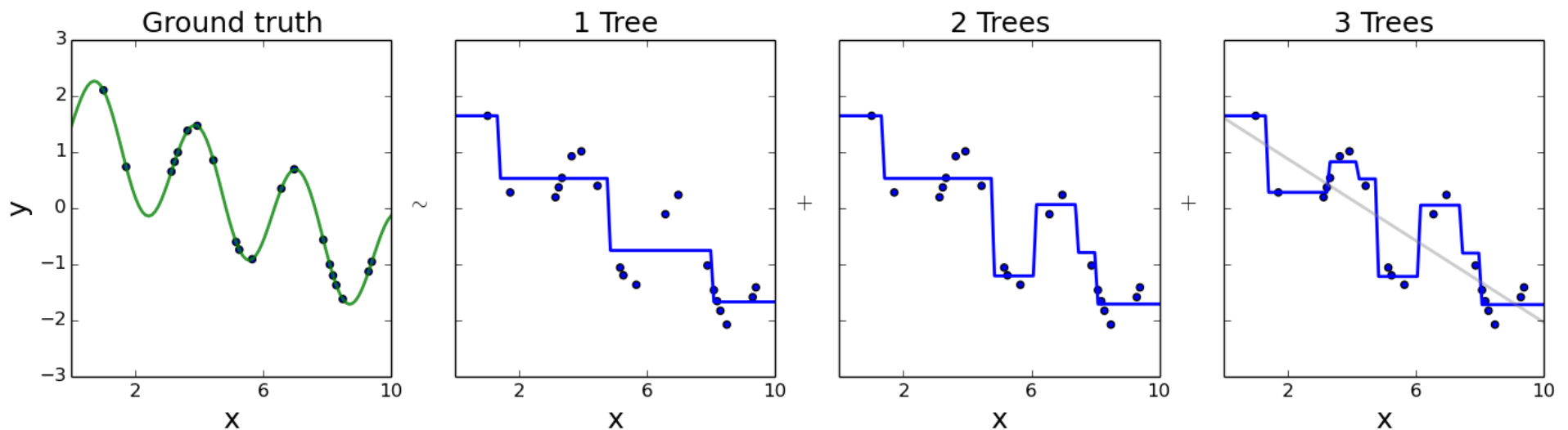


# Gradient Boosting

- Tree ensemble fitted in a forward stage-wise manner



- Intuition: Residual Fitting



# Free GBM Software



<http://scikit-learn.org/dev/modules/ensemble.html#gradient-tree-boosting>

<http://cran.r-project.org/web/packages/gbm/index.html>



# GBM Tutorial

<http://www.datarobot.com/blog/gradient-boosted-regression-trees/>

DataRobot

WHAT IS DATAROBOT?

JOIN OUR BETA

BLOG

CAREERS



## Gradient Boosted Regression Trees

© April 4, 2014 in [Machine Learning](#)

0

Gradient Boosted Regression Trees (GBRT) or shorter Gradient Boosting is a flexible non-parametric statistical learning technique for classification and regression.

This notebook shows how to use GBRT in [scikit-learn](#), an easy-to-use, general-purpose toolbox for machine learning in Python. We will start by giving a brief introduction to scikit-learn and its GBRT interface. The bulk of the tutorial will show how to use GBRT in practice and discuss important issues such as regularization, tuning, and model interpretation.

### Scikit-learn

Scikit-learn is a library that provides a variety of both supervised and unsupervised machine learning techniques as well as utilities for common tasks such as model selection, feature extraction, and feature selection.

Scikit-learn provides an object-oriented interface centered around the concept of an [Estimator](#). According to the [scikit-learn tutorial](#) "An estimator is any object that learns from data; it may be a classification, regression or clustering algorithm or a transformer that extracts/filters useful features from raw data." The API of an estimator looks roughly as follows:

```
In [1]: class Estimator(object):

    def fit(self, X, y=None):
        """Fits estimator to data. """
        # set state of ``self``
        return self

    def predict(self, X):
        """Predict response of ``X``. """
        # compute predictions of ``X``
```

# Summary

- Meaningful Model Comparison

GBM:

- One of the most powerful ML algorithms
- Not a black box
- Easy and free to use

# Questions?

[jeremy@datarobot.com](mailto:jeremy@datarobot.com)

# Society of Actuaries (SOA) – 2014 Health Meeting

## Advanced Analytics: Building Your Toolbox

# Preliminaries

## Terminology

- **Model** – A set of relationships that uses features to predict labels, and other outcomes.
- **Features** – The data provided to a model.
- **Label** – The expected numeric or categorical output from a particular model, given a set of features. Data items can be labeled or unlabeled.
- **Tag** – A text string used to subset a larger data set. One item can have multiple tags. Gmail is a good example of “tagging”.



# Preliminaries

## Terminology (continued)

- **Structured Data** – Data that is in a form that can easily be queried. Typically rows and columns that can easily be used by a model.
- **Semi-Structured Data** – Does not fit directly into rows and columns, but contains structure. XML/JSON are good examples.
- **Unstructured Data** – Textual data, often written in a human language (e.g. English). No easy way to query and extract information. Also includes images, sound & video (though these are much more difficult).

# Preliminaries

## Structured Data

1	Feature 1	Feature 2	Feature 3	Label
2	19	18	20	Good
3	1	1	3	Bad
4	10	9	11	Neutral
5	10	10	11	Neutral
6	8	11	9	Neutral
7	1	2	1	Bad
8	20	21	22	Good
9	20	22	1	Good
10	19	22	22	Good
11	20	21	21	Bad
12	20	22	22	Bad
13	10	9	9	Neutral
14	10	10	10	Neutral
15	1	1	1	???

## Unstructured Data

### Diabetes mellitus

From Wikipedia, the free encyclopedia

*"Diabetes" redirects here. For other uses, see [Diabetes \(disambiguation\)](#)*

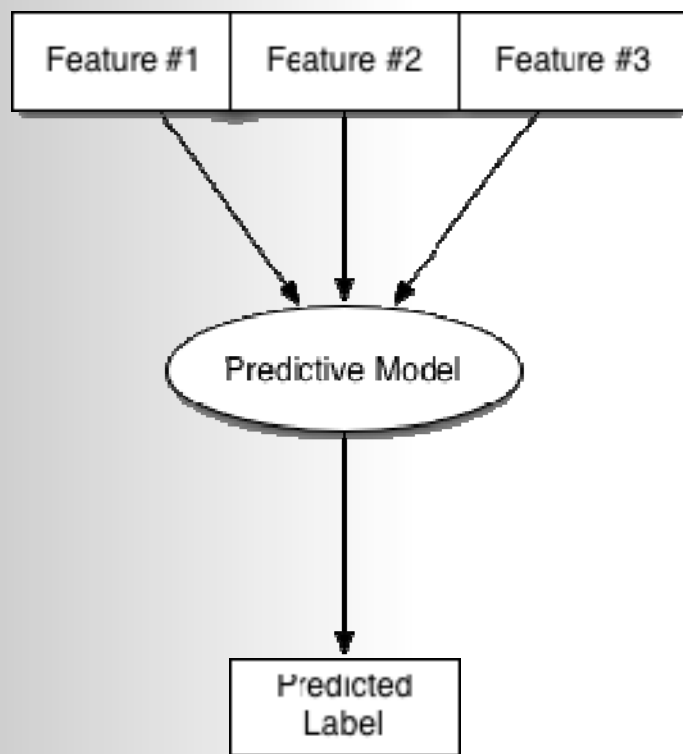
**Diabetes mellitus** (DM) or simply **diabetes**, is a group of metabolic diseases in which a person has high [blood sugar](#).<sup>[2]</sup> This high blood sugar produces the symptoms of [frequent urination](#), [increased thirst](#), and [increased hunger](#). Untreated, diabetes can cause many complications. [Acute](#) complications include [diabetic ketoacidosis](#) and [nonketotic hyperosmolar coma](#). Serious long-term complications include [heart disease](#), [kidney failure](#), and [damage to the eyes](#).

Diabetes is due to either the [pancreas](#) not producing enough [insulin](#), or because [cells](#) of the body do not respond properly to the insulin that is produced.<sup>[3]</sup> There are three main types of diabetes mellitus:<sup>[4]</sup>

- [Type 1 DM](#) results from the body's failure to produce insulin. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes".<sup>[4]</sup>

# Preliminaries

What is a model?



## Model Functions

- Models are fit/trained with known features and labels
- Hopefully models gains the ability to predict labels based on features
- Labels can be either numeric or categorical

## Common Model Types

- Random Forest
- Support Vector Machine
- Neural Network
- Generalized Linear Model (GLM)

# Preliminaries

## Model Types

- **Linear Regression** – One of the original “models” essentially draw a line through data points to predict.
- **Generalized Linear Model (GLM)** – A family of models (e.g. Logistic Regression) that extend linear regression beyond normally distributed error models.
- **Neural Network** – A model based on the human brain. Black box in the sense that output cannot be “explained.”
- **Classification and Regression Trees (CART)** – Models based on decision trees. Output can be explained.
- **Random Forest** – Many trees with specialized training method. Quickly becoming the “go to” model. Mostly a black box.

# Preliminaries

## Electronic Health Records (EHR) 101

### What is EHR?

- EHR records contain information on a patient/insured such as:
  - Medical history, immunization dates, diagnoses and medications
  - Treatment plans
  - Allergies
  - Radiology images and laboratory/test results
- Some EHR documents typically contain a variety of codes:
  - **ICD10** codes attempt to standardize diagnoses (e.g. coronary artery disease) and treatments (bypass surgery). Scheduled for adoption in USA next year.
  - **ICD9** is widespread in the USA, ICD10 is currently being adopted. SNOMED is a more international standard. Others exist, and crosswalks exist to convert between standards.
  - **UNII** is a common standard to encode substances (medications). RxNorm is more common in EHR, however Wikipedia uses UNII.
- ICD10/9 and UNII are the primary focus of this talk

# Sources of Un/Semi-structured Data

A great deal of data is unstructured

## Public Sources

- Data.gov
- Social Media
- Wikipedia
- Public Message Forums
- Open Access Journals
- UCI Machine Learning Repository
- RSS Feeds

## Proprietary Sources

- Data Marts
- Closed Journals
- News Archives
- Search Engine Queries
- Subscription Web services

## Internal Sources

- Newsletters
- Manuals
- Intranet
- E-Mail Archives

# Why use Wikipedia?

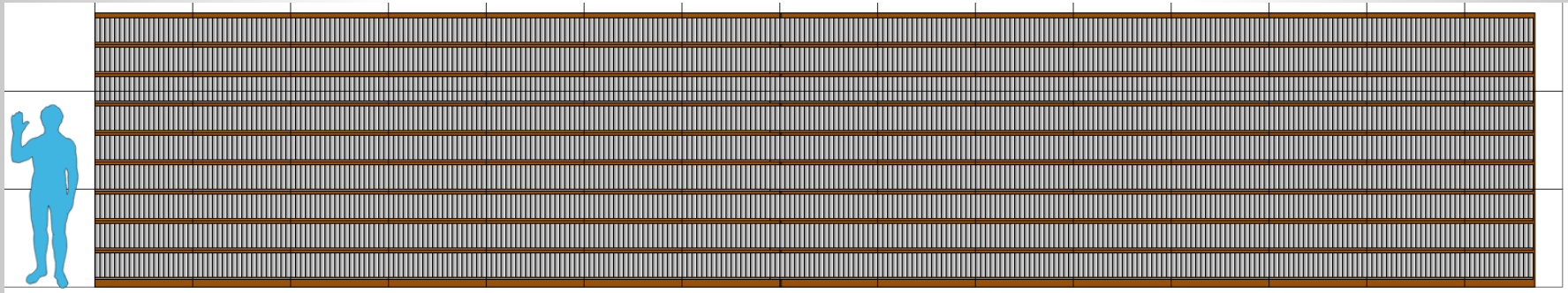
Wikipedia is one of several public sources we use

## Facts about Wikipedia Information

- Wikipedia articles are tagged according to many criteria, such as EHR code.
- Wikipedia articles are cross-linked providing insights into similarities between multiple articles.
- Tables are easily identified and extracted from Wikipedia.
- Wikipedia contains many articles with information about diagnoses, treatments and medications.
- Wikipedia text can be very valuable when used in conjunction with other data sources.
- Wikipedia is publically edited, and may not always be 100% accurate.
  - Text is often analyzed based on word counts, keyword density and document structure. This mitigates some concerns with inaccurate information.
  - However, it is still important to evaluate insights/data mined from Wikipedia. This is true of any data science project. Always involve a Subject Matter Expert (SME).

# Using Wikipedia

How big is Wikipedia (excluding images & video)



Wikipedia's size in 2010, if it were bound and printed.

Wikipedia provides a downloadable XML file containing text of current version

- Always use this offline file, do not access Wikipedia through HTTP.
- Wikipedia offline data is distributed as a very large XML file.
- XML file length: 47.46 GB (April 2014)
- Number of pages: 14,313,024
  - **Articles: 5,771,666 (this is what I mainly care about)**
  - Redirects: 6,308,100 (sometimes I care about this, for resolving links)
  - Files (media): 854,151
  - Other (templates): 1,379,107



# Using Wikipedia

Many articles in Wikipedia have tags. Some of the tags in Wikipedia are EHR codes. We are particularly interested in the following tags. They let us know what an article is about.

## Drug/Substance Codes

- CAS number
- ATC code
- PubChem
- IUPHAR ligand
- DrugBank
- ChemSpider
- UNII
- KEGG
- ChEBI
- ChEMBL

## Interventions

- ICD10-PCS
- ICD9
- MeSH
- MedlinePlus

## Diagnosis

- ICD10-CM
- ICD9
- DiseasesDB
- MedlinePlus
- eMedicine
- MeSH

# Using Wikipedia

## Embedded tags in Wikipedia (semi-structure)

Article
Talk
Read
View source
View history
Search

## Diabetes mellitus

From Wikipedia, the free encyclopedia  
(Redirected from [Diabetes](#))

*"Diabetes" redirects here. For other uses, see [Diabetes \(disambiguation\)](#).*

**Diabetes mellitus** (**DM**) or simply **diabetes**, is a group of metabolic diseases in which a person has high **blood sugar**. This high blood sugar produces the symptoms of **frequent urination**, **increased thirst**, and **increased hunger**. Untreated, diabetes can cause many complications. **Acute** complications include **diabetic ketoacidosis** and **nonketotic hyperosmolar coma**. Serious long-term complications include heart disease, kidney failure, and damage to the eyes.

Diabetes is due to either the **pancreas** not produce enough **insulin**, or because cells of the body do not respond properly to the insulin that is produced.<sup>[2]</sup> There are three main types of diabetes mellitus:<sup>[3]</sup>


- **Type 1** DM results from the body's failure to produce insulin. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes".<sup>[3]</sup>
- **Type 2** DM results from **insulin resistance**, a condition in which cells fail to use insulin properly, sometimes also with an absolute insulin deficiency. This form was previously referred to as non insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes".
- **Gestational diabetes**, is the third main form and occurs when pregnant women without a previous diagnosis of diabetes develop a high **blood glucose** level.

Prevention and treatment often involve: a **healthy diet**, **physical exercise**, not using **tobacco**, and being a **normal body weight**.<sup>[3]</sup> **Blood pressure** control and proper foot care are also important in those with the disease.<sup>[3]</sup> **Type 1** diabetes must be managed with **insulin injections**.<sup>[3]</sup> **Type 2** diabetes may be treated with medications with or without insulin.<sup>[3]</sup> Insulin and some oral medications can cause **low blood sugar**, which can be dangerous. **Pancreas transplants** have been tried in an effort to cure **type 1** diabetes with limited success. **Gastric bypass surgery** has been successful in many with severe obesity and **type 2** DM. **Gestational diabetes** usually resolves after delivery.

Globally, as of **2010**, an estimated **227 to 285** million people had diabetes, with **type 2** making up about **90%** of the cases.<sup>[4][5]</sup> This is equal to **3.3%** of the population with equal rates in both women and men.<sup>[5]</sup> In **2011**, it resulted in **1.4** million deaths worldwide making it the 8th leading cause of death.<sup>[3]</sup>

### Diabetes mellitus

Classification and external resources



Universal blue circle symbol for diabetes.<sup>[1]</sup>

ICD-10	<b>E10</b> , <b>E14</b>
ICD-9	<b>250</b>
MedlinePlus	<b>001214</b>
eMedicine	<b>med/546</b> , <b>emerg/134</b>
MeSH	<b>C18.452.394.750</b>

# Using Wikipedia

## Some uses for Wikipedia data

- Wikipedia tags most diagnosis (e.g. coronary artery disease) and treatments(e.g. bypass surgery) by ICD/drug code.
  - By statistically comparing an “unknown document” to a Wikipedia article, we can get an idea of what the unknown document’s topic is and tag it by the same tags Wikipedia uses.
  - Word structure in a Wikipedia articles can give us a general idea of the mortality and morbidity of the topic the diagnosis/procedure the article is describing
- Links between articles highlight medication and treatment relationships
- Word structure and clustering analysis can be used to group similar ICD and drug codes

# Using Wikipedia

## Making sense of unstructured Wikipedia data

- Tags can usually be found using traditional programming with regular expressions (REGEX).
- Because each Wikipedia article is tagged with multiple coding standards some degree of “code translation,” or crosswalk, can be inferred.
- Wikipedia redirects can be used to infer synonymous phrases.
- Links between articles can also be found using traditional programming.
- **Making sense of the article text is much more complex**
  - **Bag of Words Analysis:** Simply produce a histogram of word frequencies in the document. Create rules for defining the words that make it into the histogram.
  - **Bag of Words with Proximity:** Similar to bag of words, except weights words that appear closer to each other.
  - **Sentence Structure Analysis:** Attempt to locate certain sentence structures within documents.
  - **Natural Language Processing:** Attempt to extract the “knowledge” contained in the text.

# Using Wikipedia

## Finding ICD/Drug Codes in Wikipedia

### Wikipedia Articles Found by Code Type

- ICD10-CM: 4,633
- ICD10-PCS: 122
- ICD9: 5,201
- UNII: 339
- Medical codes some very unusual conditions.

### Notes

- A single Wikipedia article is often tagged to a range of codes.
  - For “ear infection” there is a single Wikipedia article, yet codes for left/right ear, initial/subsequent occurrence, and if related to another root cause.
  - Some codes cover very unusual situations that are simply do not have Wikipedia articles.

# Using Wikipedia

## High Coverage Rates are Imposable

- Medical codes tend to be very specific, e.g. left or right arm, left or right ear, initial or subsequent encounter.
- Medical codes some very unusual conditions.

### Unusual ICD10 Codes

- V9542XA:       Spacecraft crash injuring occupant, initial encounter
- W59.22XA:     Struck by a turtle
- V91.07XA:     Burn due to water-skis on fire, initial encounter
- W22.02XA:     Hurt walking into a lamppost, initial encounter

### Unusual UNII Codes

- ZZ4SNQ91FW:       Pale Toadfish
- E06K5531Q8:       Chicken Foot, Cooked
- 466251J72G:       Pork Brain

# Bag of Words

## Approaches to unstructured textual data

### General Description

- Regular textual data cannot be directly applied to a model.
- Bag of words is one of the oldest, and a common approach for modeling textual data.
- To use a “Bag of Words” you should choose a set of important words.
- Generate a count of the words.
- The word count, or percentage weight of each word, becomes your features.

### Effectiveness

- Bag of words counts only individual words (e.g. “United States of America” is 4 different counts).
- No meaning is derived from the underlying sentences.

# Bag of Words

## Introduction to Bag of Words

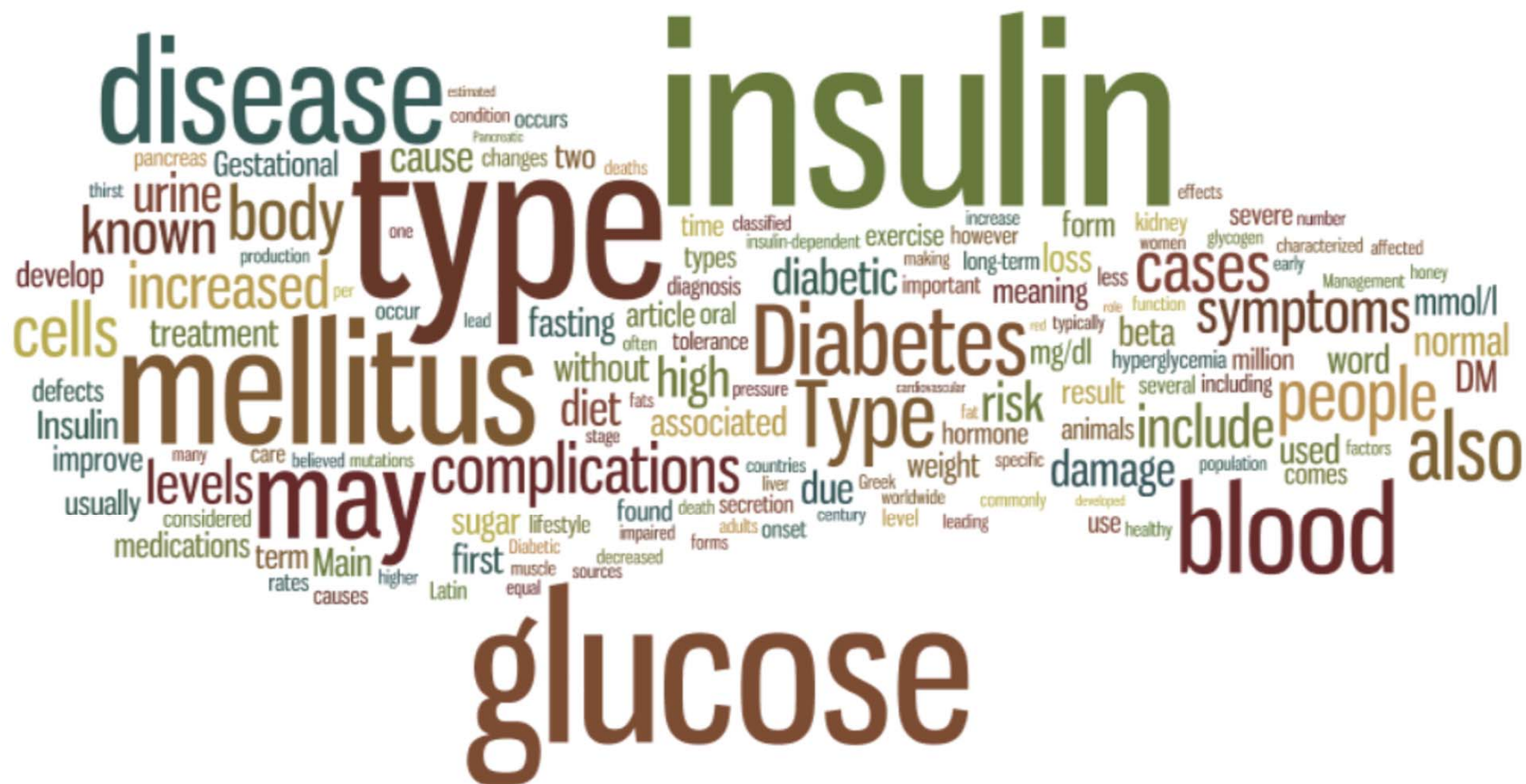
Rank	Count	Word
1	245	the
2	209	diabetes
3	149	and
4	75	type
5	66	insulin
6	49	glucose
7	48	mellitus
8	48	with
9	38	are
10	34	for
11	30	disease
12	28	blood
13	26	legend
14	25	may
15	24	which
16	23	from
17	23	risk
18	22	diabetic
19	20	first

- The table to the left is a histogram for the Wikipedia article for diabetes.
- Typically common words (e.g. the, and, are, for, etc.) are removed.
- Key words (the bag) can be designated and their normalized counts can become features for a model.
- Bag of words can be quite valuable for document classification.
- However, bag of words only looks at single word counts. How the words are used is not considered.



# Bag of Words

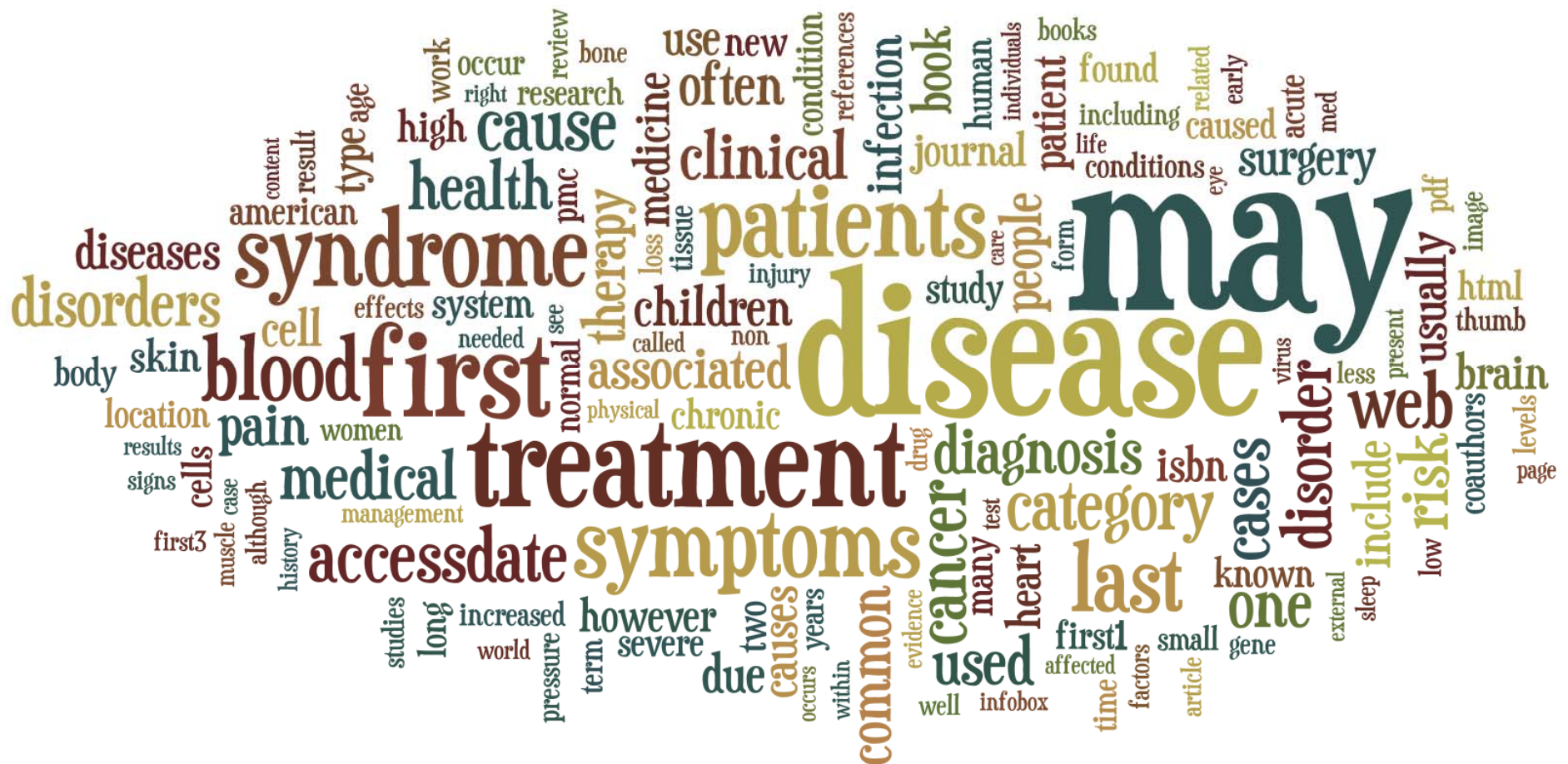
# Wikipedia Diabetes Article as a Wordle



# Bag of Words

## Wikipedia ICD Articles as a Wordle

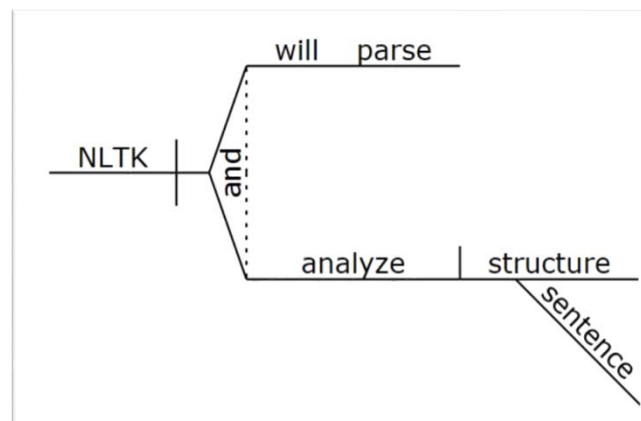
# Wikipedia ICD Articles as a Wordle



# Natural Language Processing (NLP)

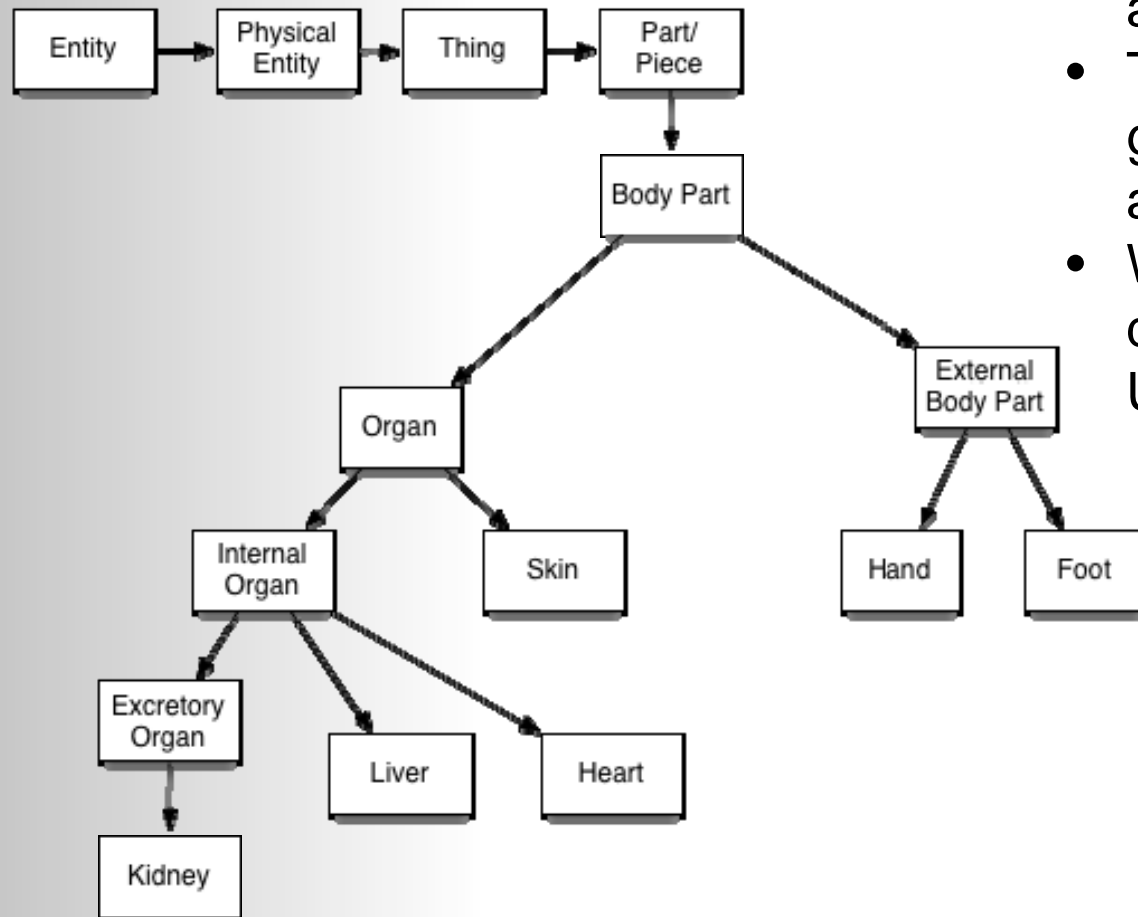
## Overview

- Natural Language Processing (NLP) attempts to extract knowledge from textual data.
- NLP is a very deep, yet ever evolving field of study.
- The Python Natural Language Tool Kit (NLTK) can be used to parse and analyze sentences.
- WordNet is a lexical database that can be used with NLTK. WordNet is used by Google and projects such as IBM Watson.
- NLTK can be used to parse and analyze sentence structure.



# Natural Language Processing

## WordNet Databases

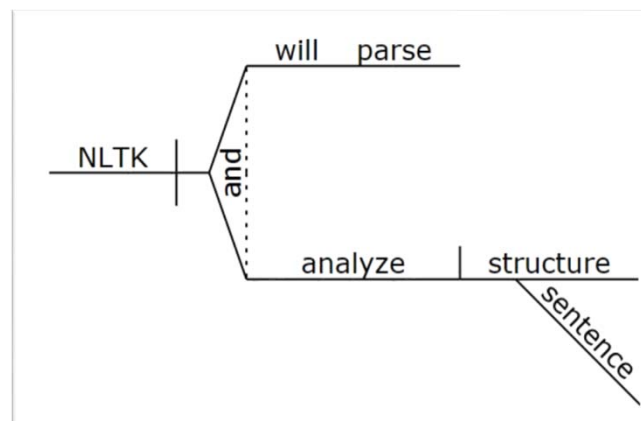


- WordNet organizes nouns and verbs into hierarchies.
- This can be very useful for generalizations with lexically analyzing text.
- WordNet was originally developed by Princeton University.

# Natural Language Processing (NLP)

## Overview

- Natural Language Processing (NLP) attempts to extract knowledge from textual data.
- NLP is a very deep, yet ever evolving field of study.
- The Python Natural Language Tool Kit (NLTK) can be used to parse and analyze sentences.
- WordNet is a lexical database that can be used with NLTK. WordNet is used by Google and projects such as IBM Watson.
- NLTK can be used to parse and analyze sentence structure.



# Deep Learning

## Key Features & Requirements

### Deep Learning Features

- Deep learning models are “deep” with many layers to learn different aspects of the data.
- Deep layers can be trained independently. This scales well.
- Deep learning requires only partially labeled data.

### Deep Learning Requirements

- Deep Belief Neural Networks (DBNN) require a fixed length input of features, just like other models.
- DBNN's require binary features.
- Because of the binary feature requirement DBNN's can be difficult to apply to some problems.

# Deep Learning

## Overview

### Shallow Learning

- Only a few layers, typical of models such as:
  - Feedforward Neural Networks
  - Support Vector Machines
  - General Linear Models
- Limited ability to create “hidden features”, such as our own eyes detecting edges and corners.

### Deep Learning

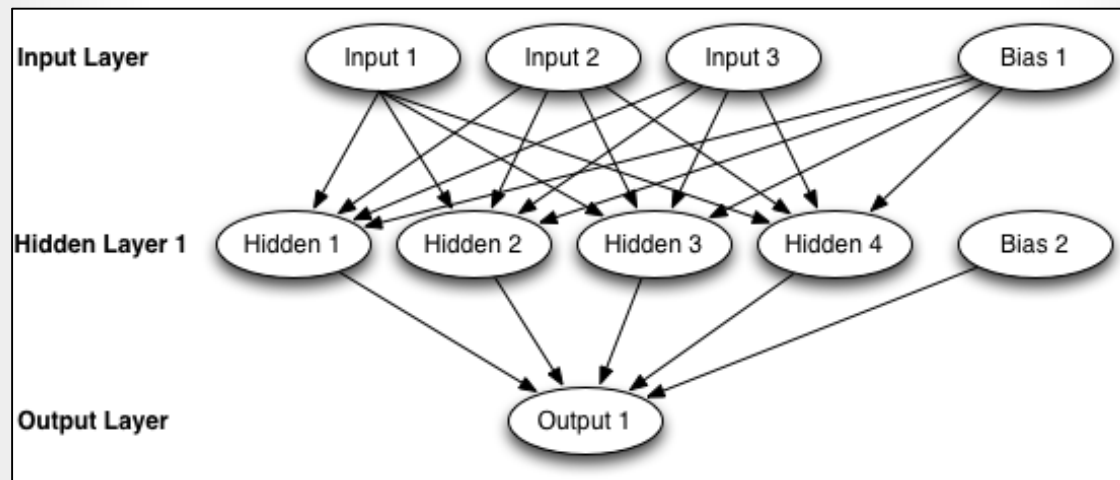
- Feature input to deep neural networks are typically binary. Output label is typically categorical.
- Deep learning algorithms typically have many layers (depth) compared to more shallow learning algorithms.
- Deep Neural Networks started a resurgence in neural network research.
- Deep learning layers are not necessarily neural networks, there is some research into deep Support Vector Machines (SVM).



# Shallow Neural Network

## Overview

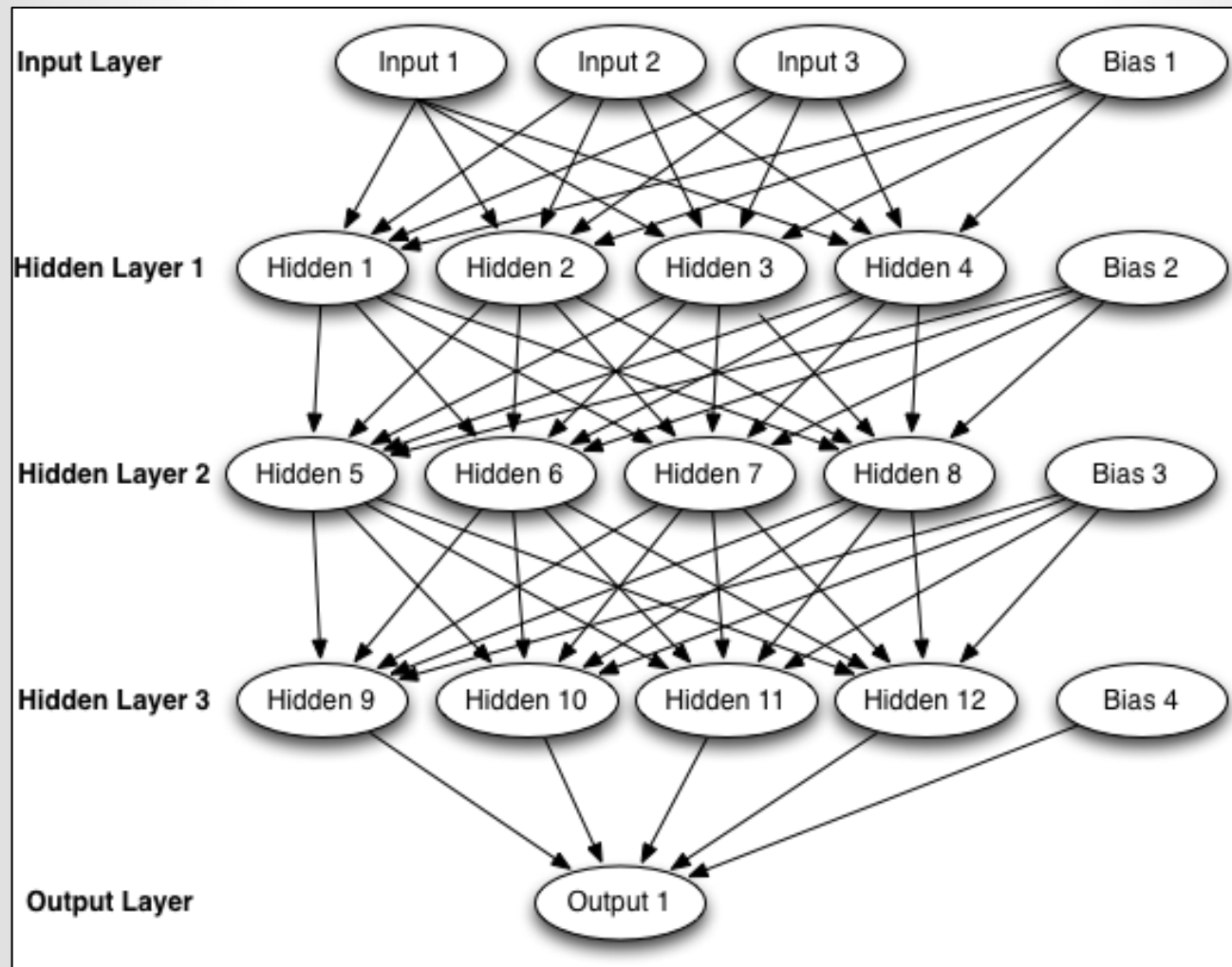
- Typical Artificial Neural Networks (ANN) have a single hidden layer.
- Sigmoidal output with weight and bias links:
  - Weights control sigmoid curvature
  - Bias controls sigmoid position
- Additional hidden layers are counterproductive for several reasons:
  - Vanishing gradient problem
  - Complexity to train





# ANN Trying to be Deep

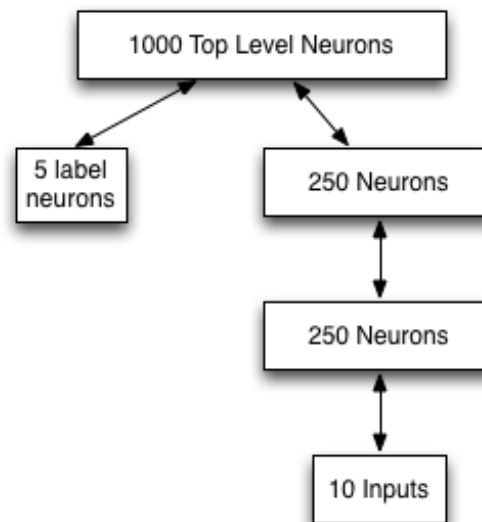
## But unsuccessful



# Deep Belief Neural Network (DBNN)

## Overview

- Multiple layers of Restrictive Boltzmann Machines (binary)
- Training data (10 inputs) trains the next layer (250 neurons) using unsupervised learning.
- Unsupervised training works its way higher until the top 1000 neurons is reached.
- Any labels we have, from the data, are used to train using supervised training.
- Binary features are one of the largest stumbling blocks to DBNN implementation.



# Conclusion

## Citations

- Samuel R. Bowman. 2013. “Can recursive neural tensor networks learn logical reasoning?”
- Richard Socher, Cliff Lin, Andrew Y. Ng, and Christopher D. Manning, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks". ICML 201
- Richard Socher, John Bauer, Christopher D. Manning and Andrew Y. Ng, "Parsing with Compositional Vector Grammars"
- [www.wordle.net](http://www.wordle.net) – For wordle generation.

# Conclusion

## Useful Tools

- R Programming Language
- Python Programming Language
- Numpy/SciPy for numeric processing (Python)
- Scikit-Learn for machine learning (Python)
- Natural Language Toolkit (NLTK) (Python)
- WordNet – For lexical meaning
- Theano – For deep learning (Python)

# #66: Building your toolbox

**Syed M. Mehmud**

Director and Senior Consulting Actuary  
Wakely Consulting Group  
[syedm@wakely.com](mailto:syedm@wakely.com)



# Building your Toolbox

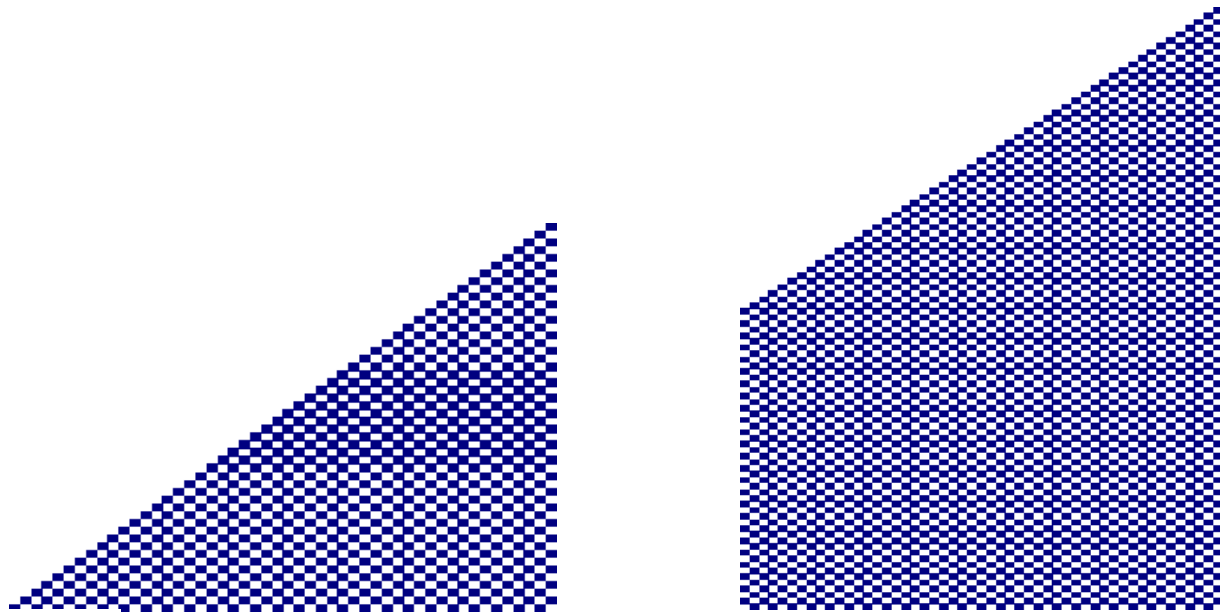
- What is a toolbox?

# Building your Toolbox

- Things you need
- Things you can live without...

# Building your Toolbox

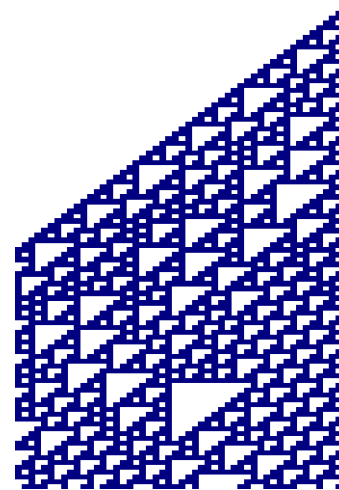
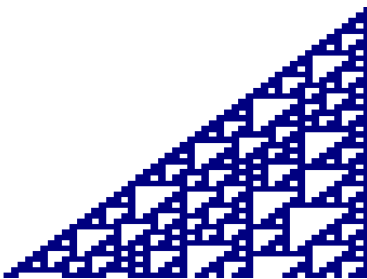
- What is complexity science?
  - Difficult to define
  - States of rule 250 at 50 and 100 time steps





# Building your Toolbox

- What is complexity science?
  - Difficult to define
  - States of rule 110 at 50 and 100 time steps



# Building your Toolbox

- What is complexity science?

***A complexity model is one where all prior states must be computed in order to observe a certain state***

# Building your Toolbox

- Example
- Model is available at:  
<http://www.soa.org/research/research-projects/health/research-complex-call-models-winner.aspx>