Image source: Wikipedia

MACHINE LEARNING IN ENGINEERING

Joel Graff, P.E.

Contact Information

Joel Graff, PE

Illinois Department of Transportation, District 2

819 Depot Ave,

Dixon, IL 61021

Ph: 815-284-5383 (work)

Email: joel.graff@illinois.gov

LinkedIn: www.linkedin.com/in/joelcgraff

IDOT is Hiring!

- Maintenance workers, technical / non-technical staff, civil engineers
- No permanent postings
- Watch www.idot.gov (Employment tab)

Overview

- Neural Networks
- Collecting & Modeling Data
- Case Studies
- So What?

Engineering Applications

- Predicting slope failure
- Fault diagnosis in HVAC systems
- Estimating open channel flows
- Predicting pavement transverse crack lengths
- Optimizing industrial design processes
- Optimizing construction scheduling
- Assessing contractor / worker effectiveness

IBM 702 Mainframe used in early AI research (image source: Wikipedia)

1950
Alan Turing publishes landmark paper on "thinking machines"

1956 -1966
The "Golden Era" of AI

1966 - 1974
Funding decline

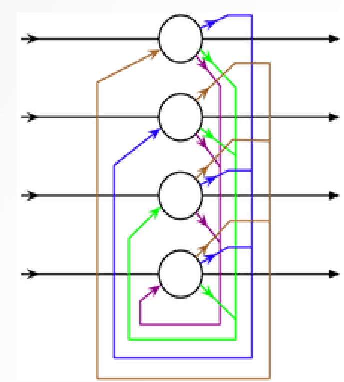
1956 - 1974 | 1980 - 1987 | 1993 - Present

Early 1980's
Autonomous vehicles successfully tested in Germany and Europe

1982
John Hopfield proves the first neural network

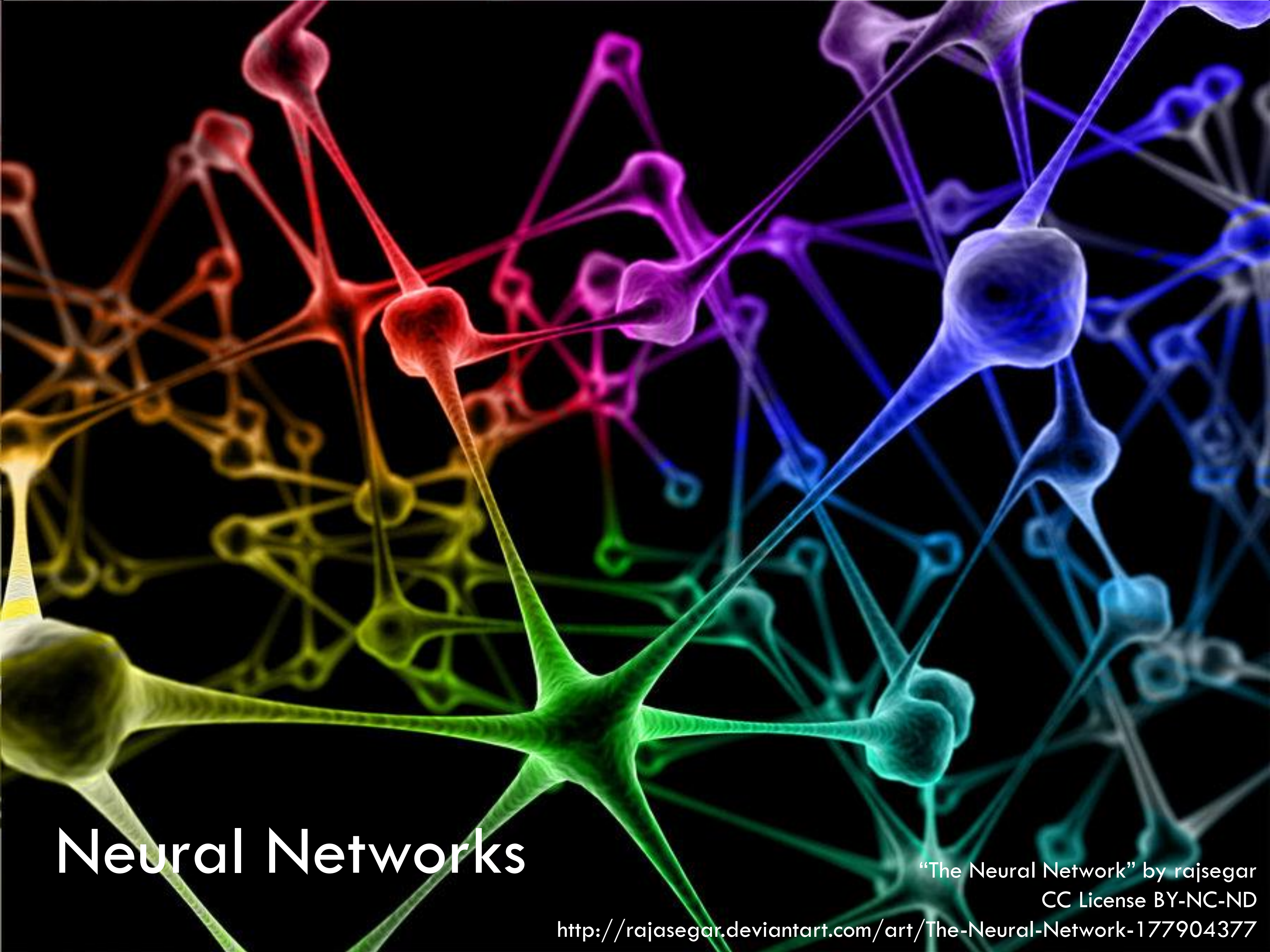
1980 - 1985
Expert systems become commercially viable

2011
Watson defeats two Jeopardy! champions for a $1 million prize



A Hopfield network
image source: Wikipedia

Neural Networks



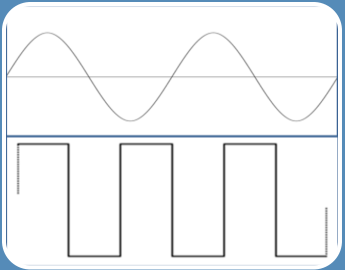
“The Neural Network” by rajsegar

http://rajasegar.deviantart.com/art/The-Neural-Network-177904377

Neural Networks

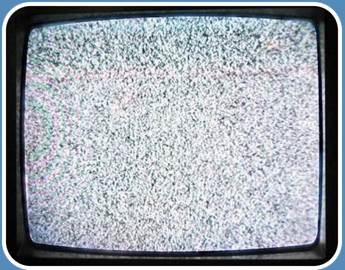
Brain physiology

- Neurons and synapses
- Pattern recognition



Classification / Regression

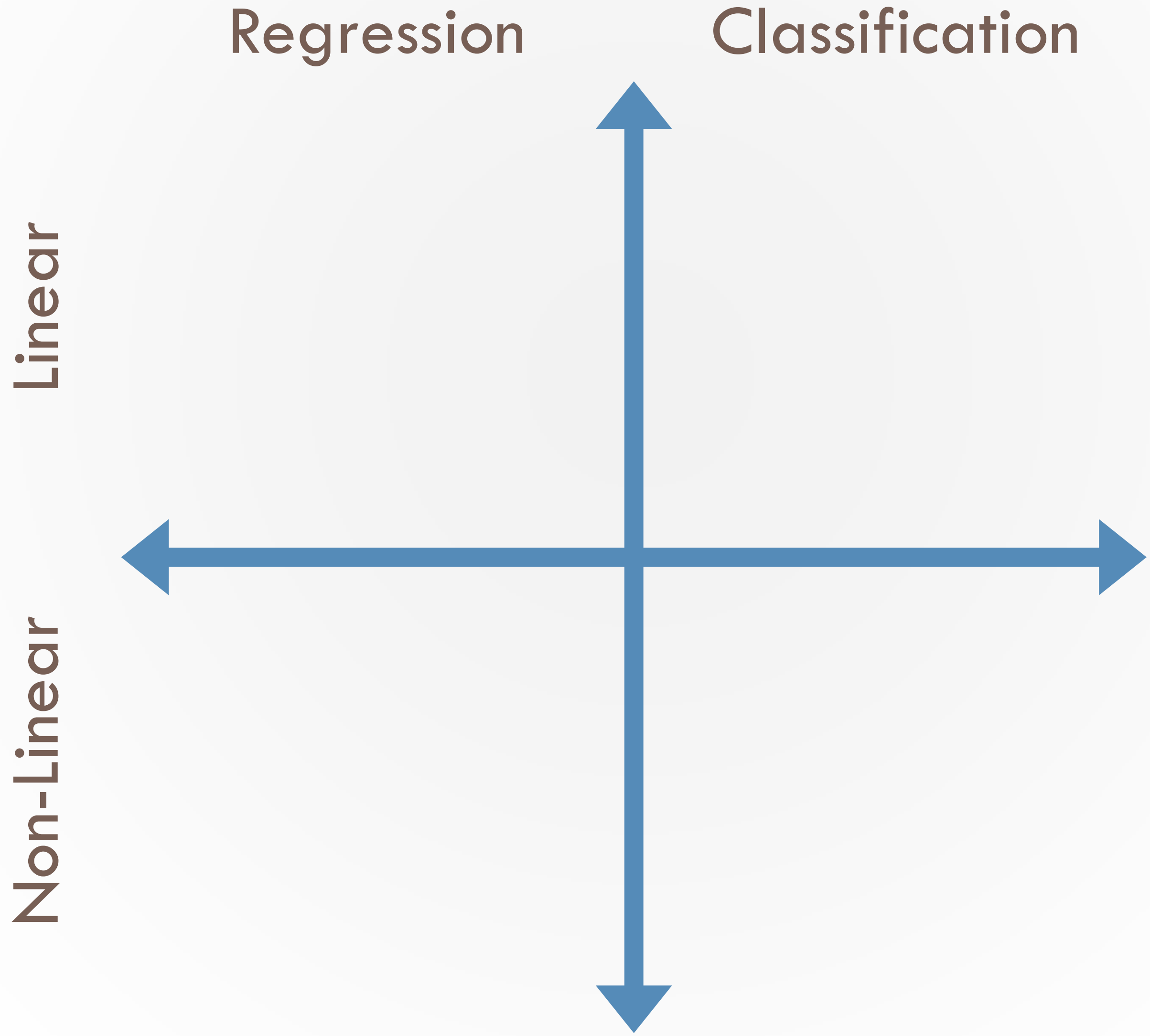
- Disease classification
- Stock price prediction



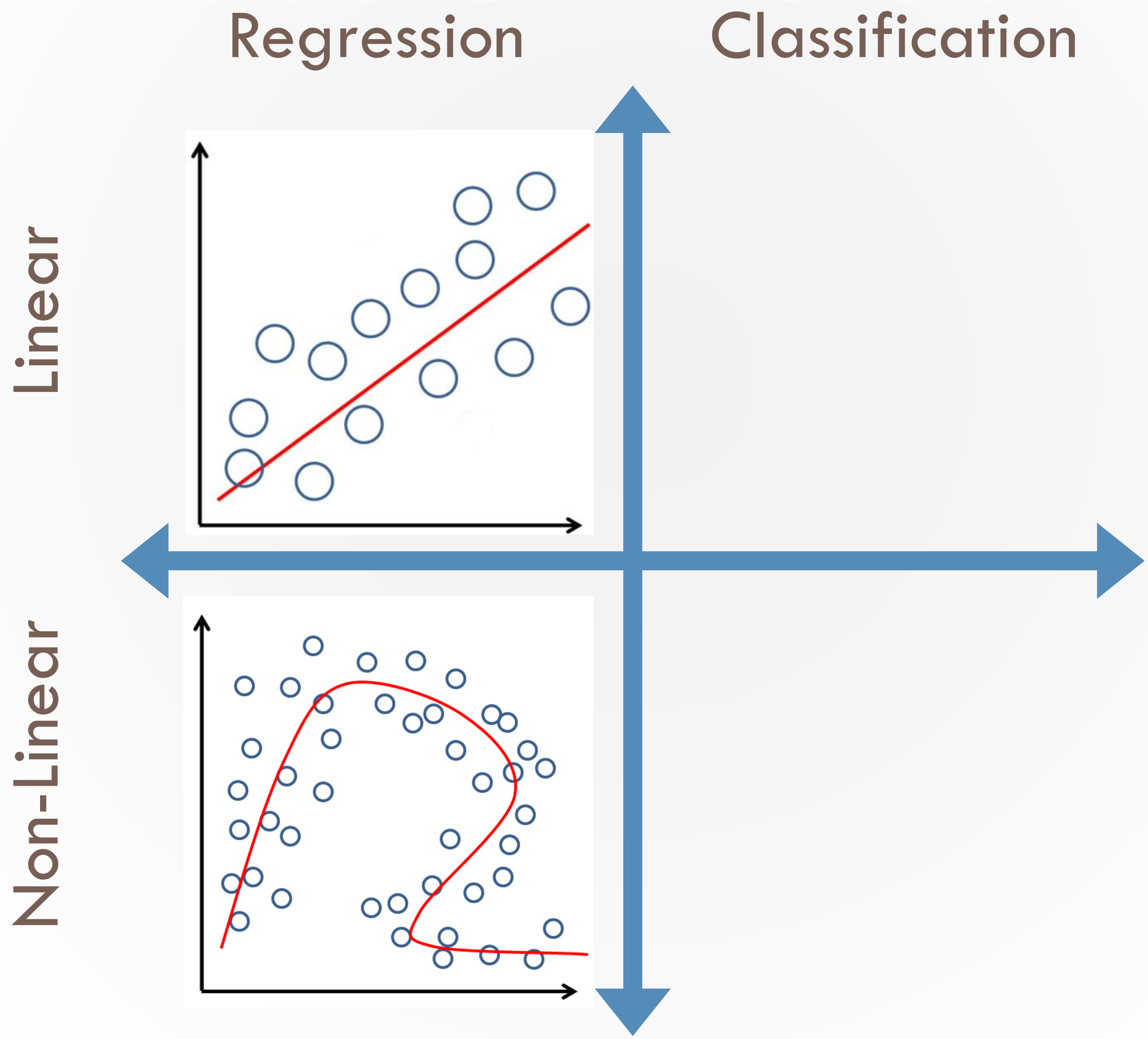
Noisy / Complex data

- Missing, incorrect, or irrelevant information
- Linear / non-linear

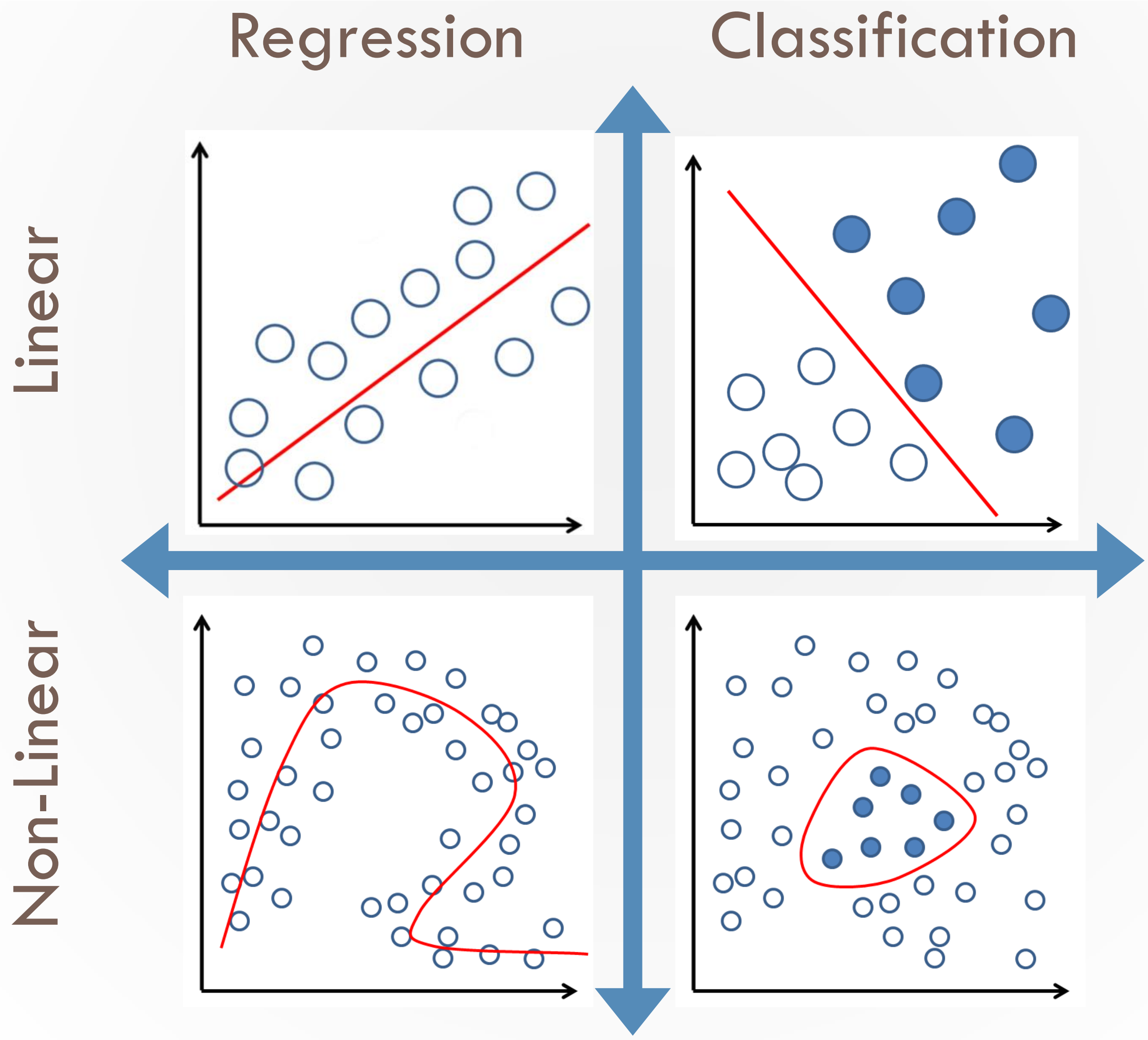
Problem Type / Complexity Matrix



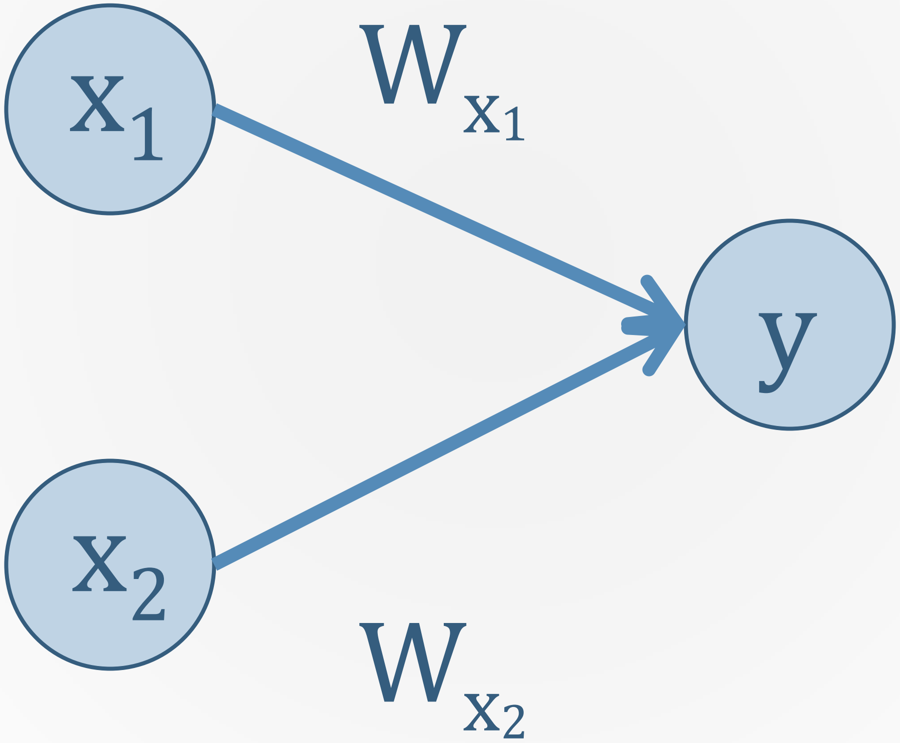
Problem Type / Complexity Matrix



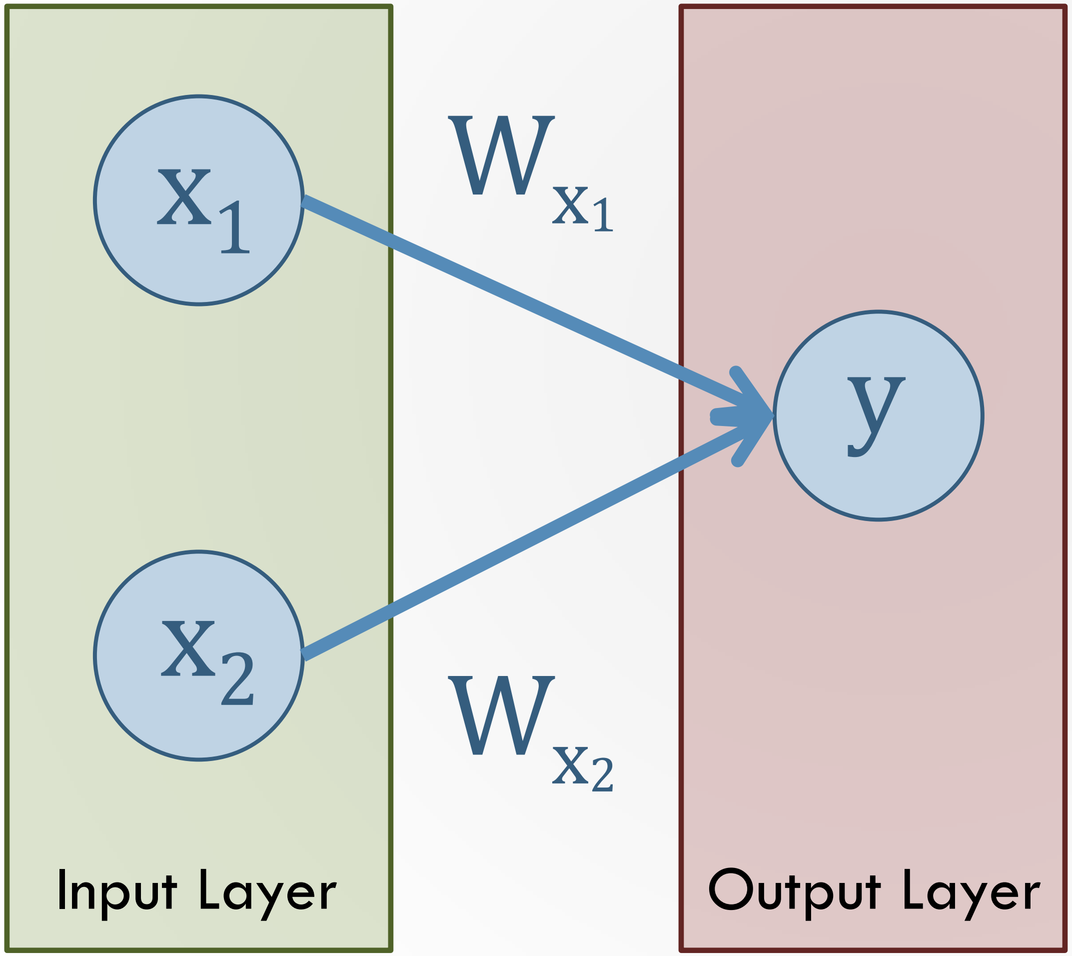
Problem Type / Complexity Matrix



Artificial Neural Network *(Linear)*

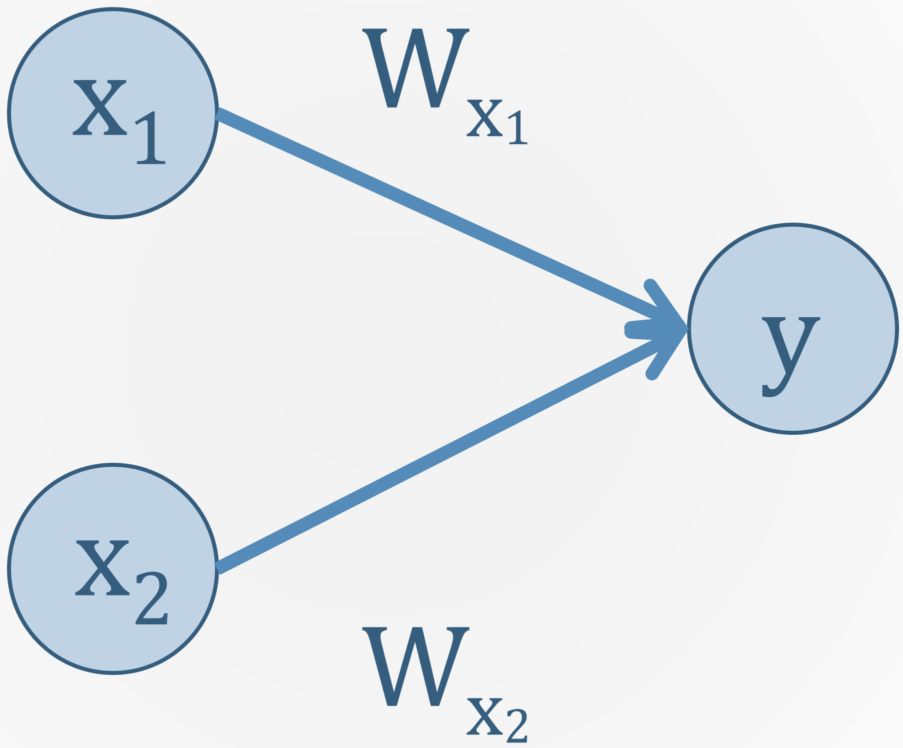


Artificial Neural Network *(Linear)*



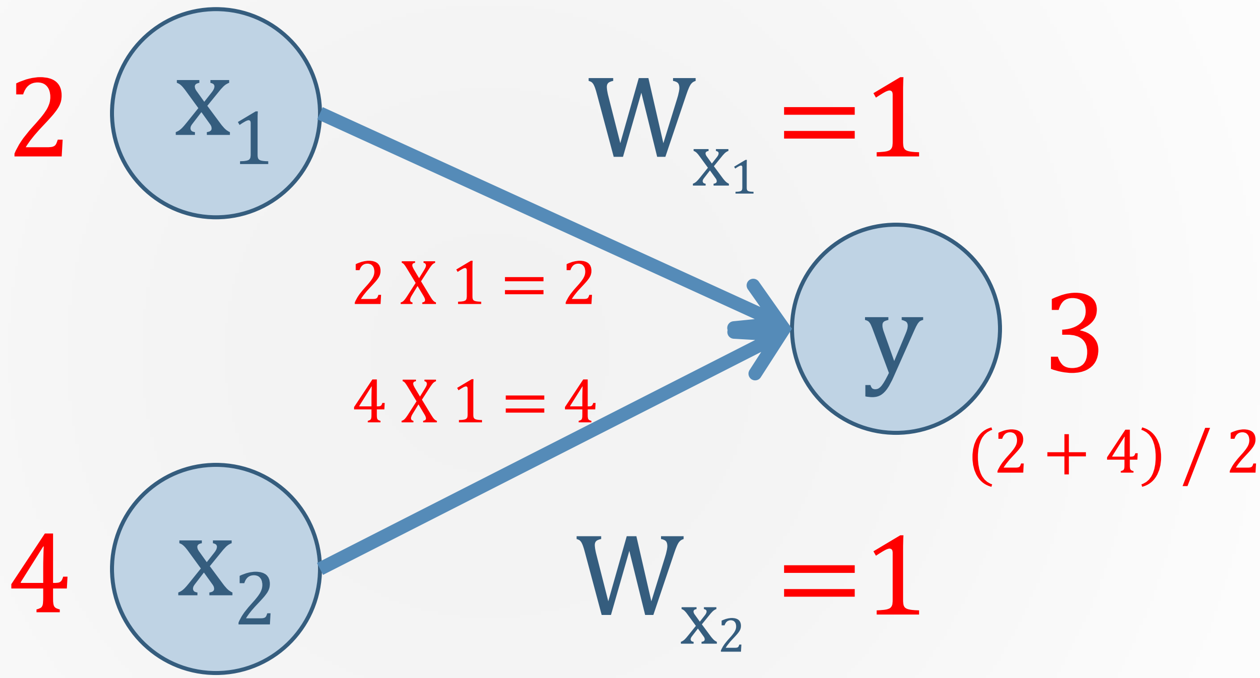
Fully-connected feed forward network

Artificial Neural Network *(Linear)*



$$\left(\frac{1}{n}\right)\sum_{1}^{n} x_n \times w_n$$

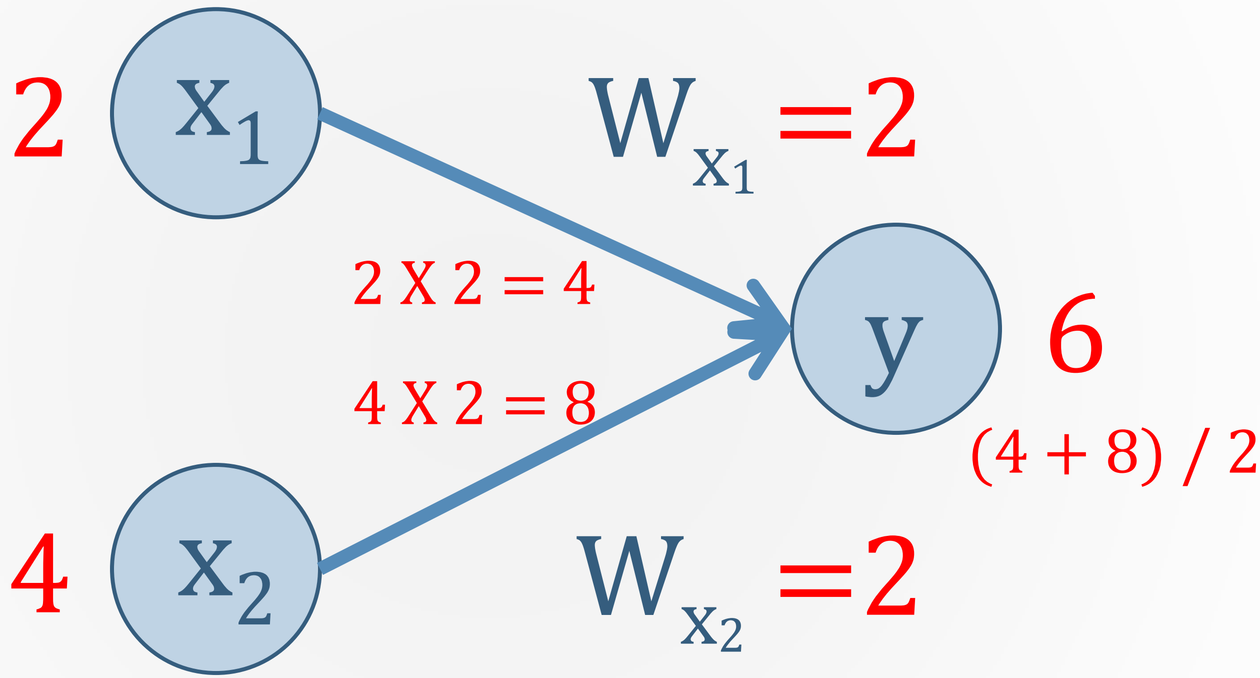
Artificial Neural Network *(Linear)*



$$\left(\frac{1}{n}\right)\sum_{1}^{n} x_n \times w_n$$

Simple Average

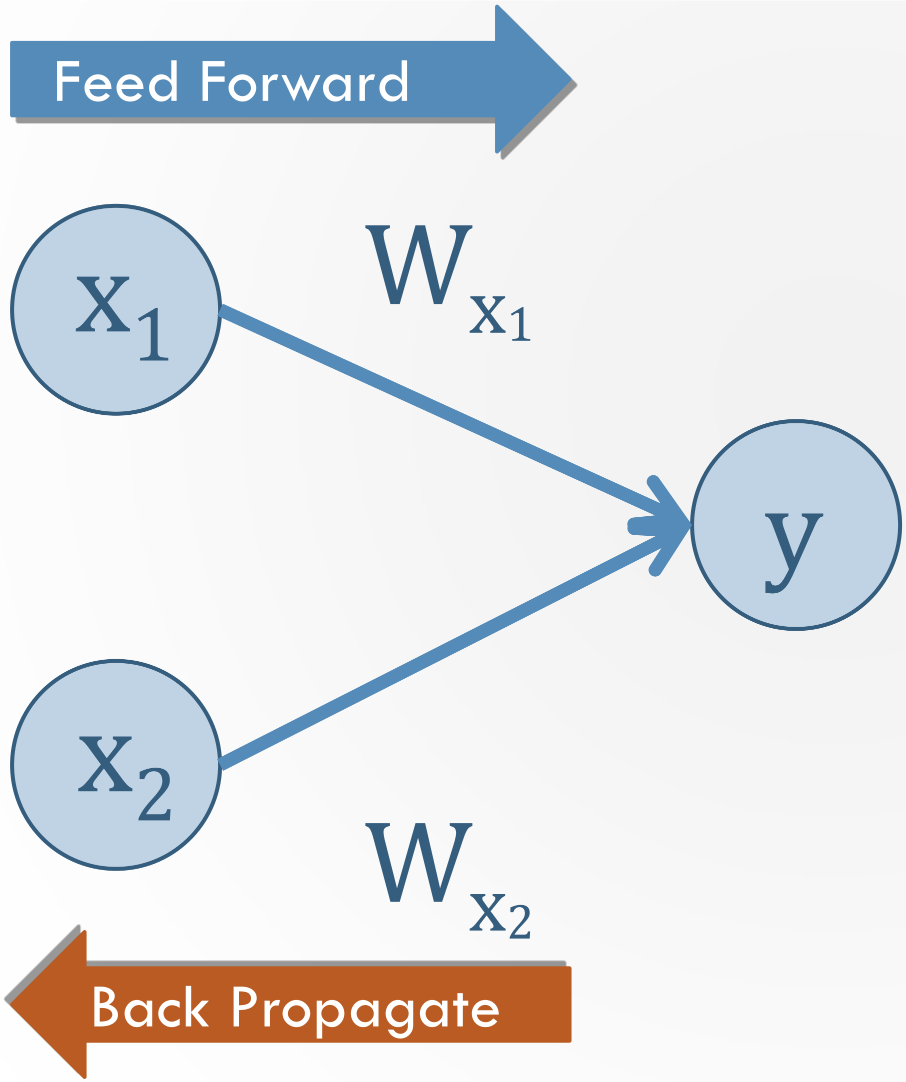
Artificial Neural Network *(Linear)*



$$\left(\frac{1}{n}\right)\sum_{1}^{n} x_n \times w_n$$

Summation

Supervised Learning

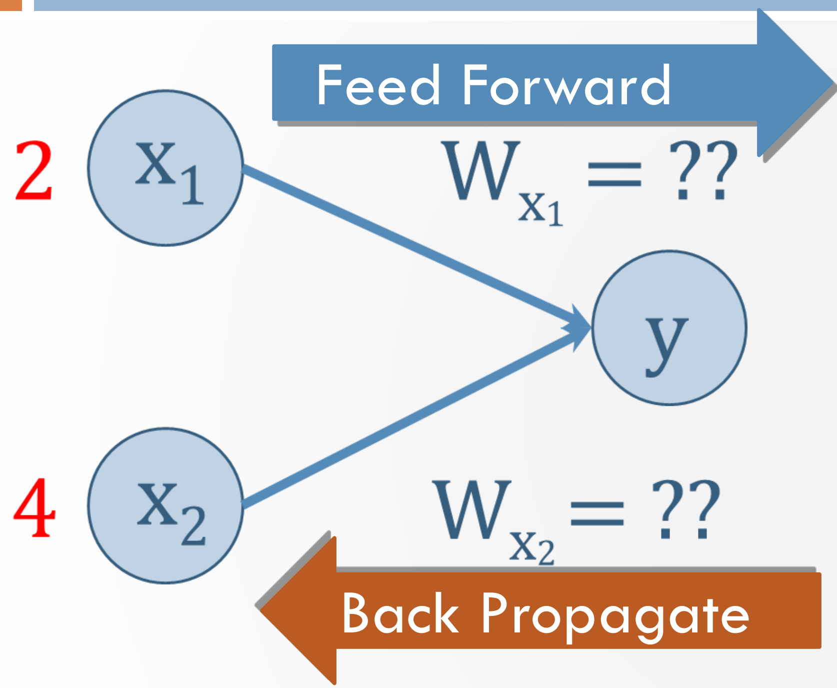


Prediction

\- Expected Result

Error

Supervised Learning



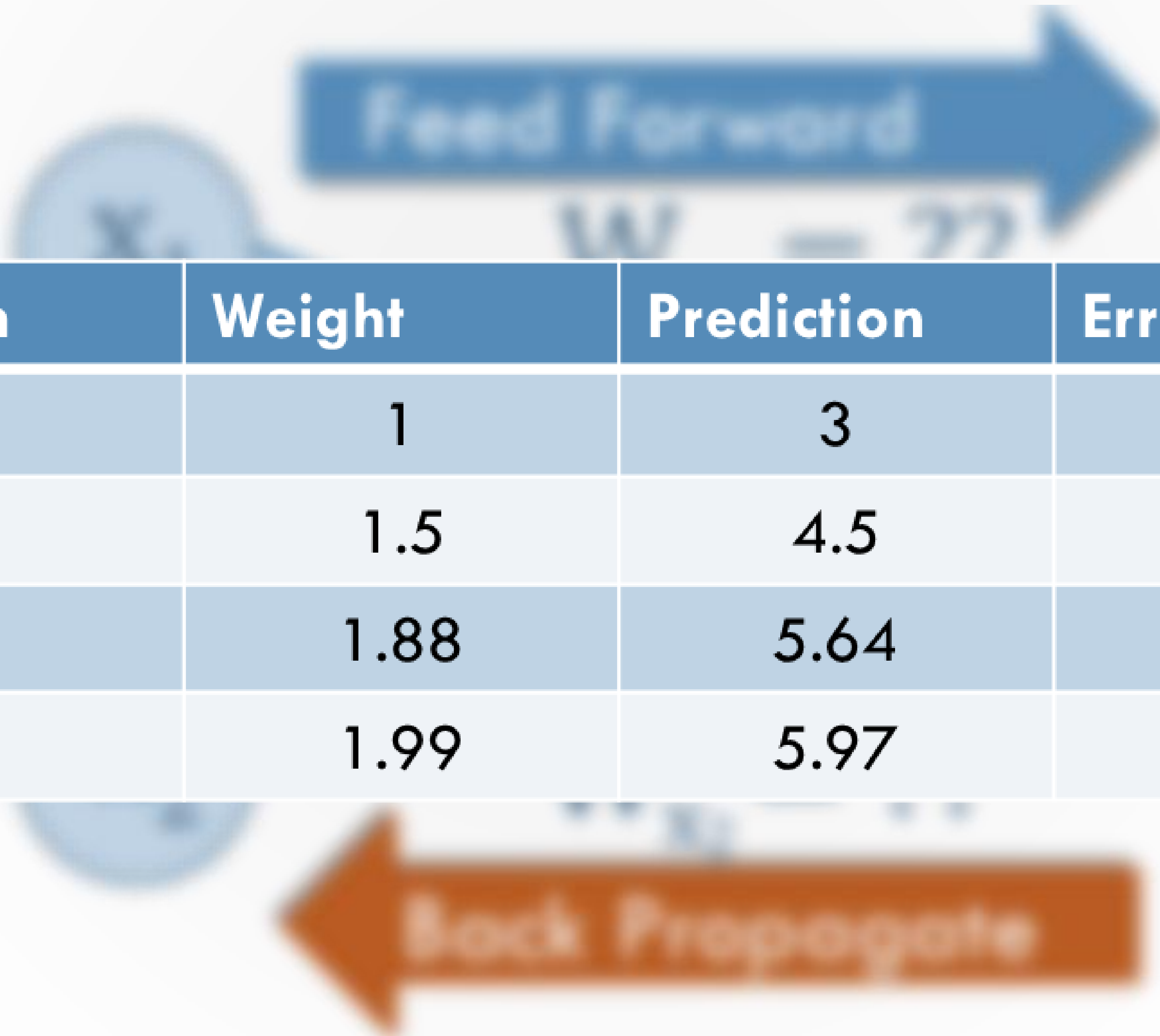
Update Rule

Increase / decrease weights by prediction error

Convergence

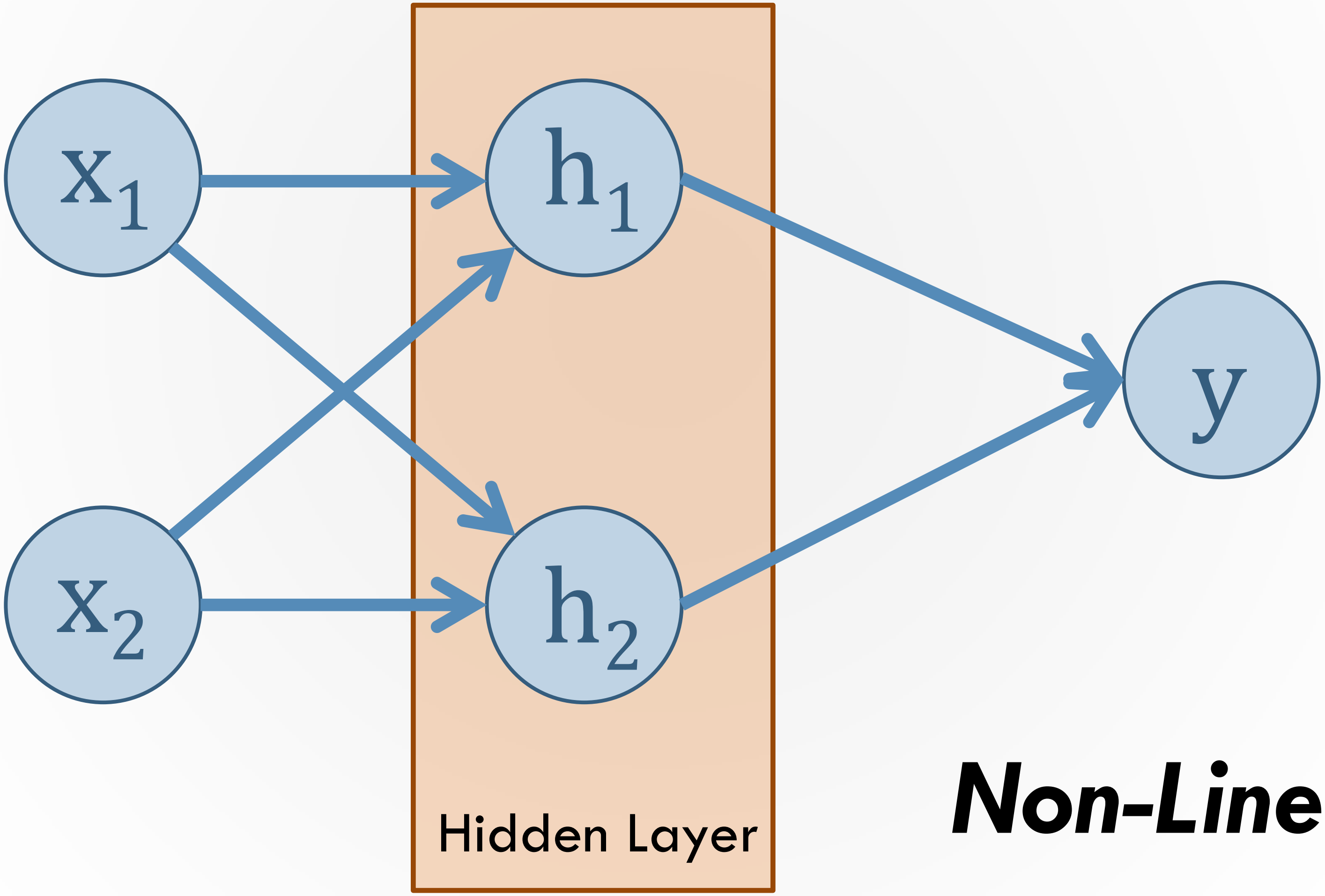
Network error minimizes, weights stabilize

Supervised Learning



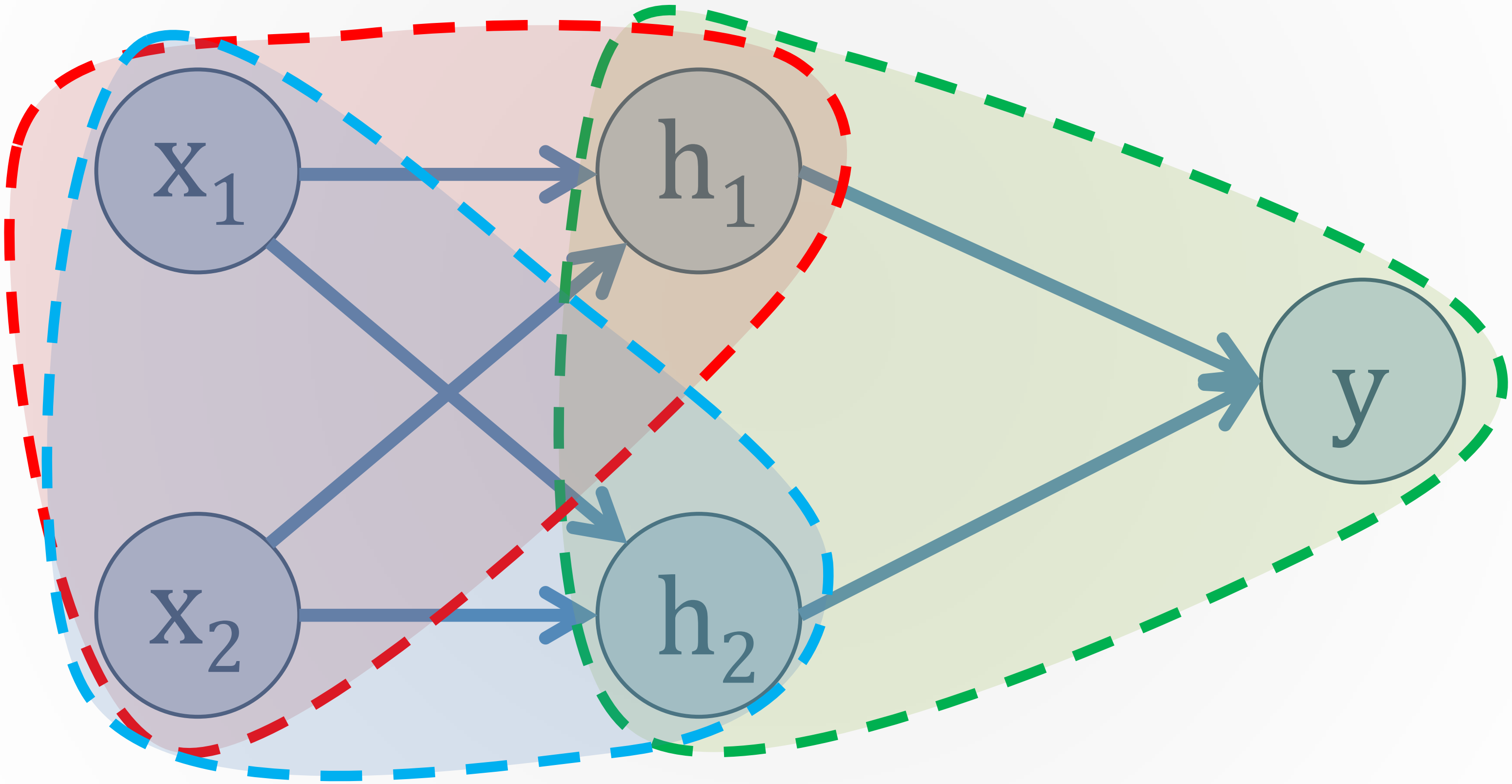
Iteration	Weight	Prediction	Error
1	1	3	50%
2	1.5	4.5	25%
3	1.88	5.64	6%
4	1.99	5.97	0.5%

Artificial Neural Network *(Non-Linear)*



Non-Linear

Artificial Neural Network *(Non-Linear)*



Linear Network Composition

In 2012, researchers at Google Brain created a network of 16,000 computer processors with over 1 billion connections.

They then trained this network by showing it screen captures from 10 million randomly selected YouTube videos over three days.

? ?

At the end of the experiment, researchers discovered the network was able to recognize two things in particular.

Can you guess what they were?

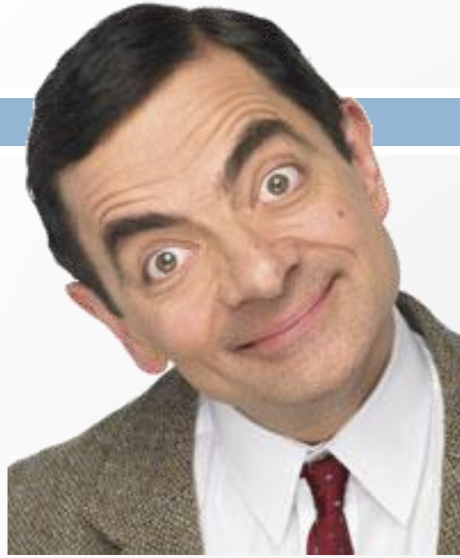


Image source: http://www.chroniclelive.co.uk/

Image source: www.twitter.com/realgrumpycat

At the end of the experiment, researchers discovered the network was able to recognize two things in particular.

Can you guess what they were?

Data collection & Modeling

Data Sets

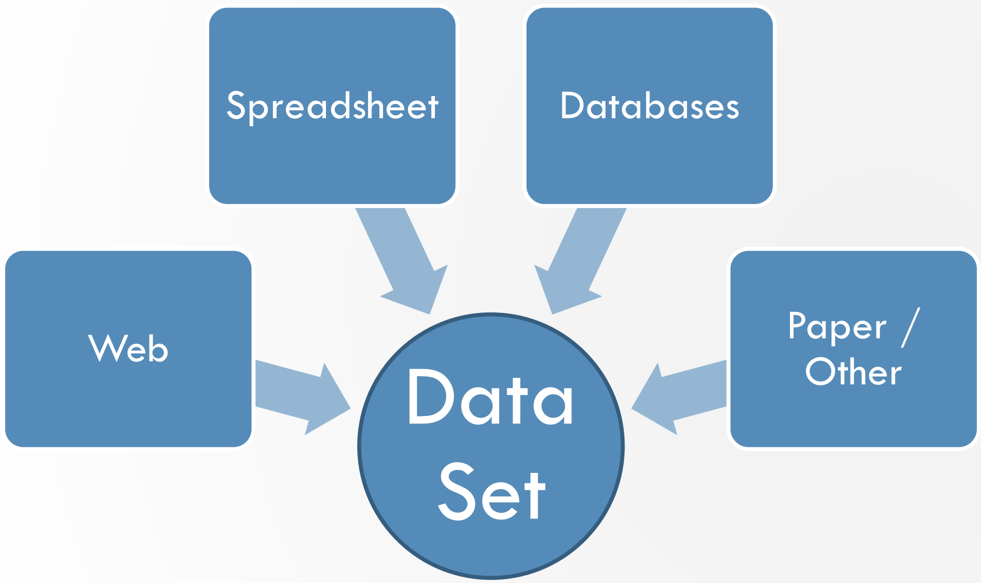
- Data Set
  - Roles
    - Target
      - What we're trying to predict
    - Features / Predictors
      - Describes the characteristics of the dataset
  - Types
    - Numeric
      - {3.14159, 1.333, 42.0}
    - Unordered Categorical
      - {Atlanta, Dallas, Chicago}
    - Ordered Categorical
      - {Low, Medium, High}

Data Sets - Example

Housing Market Price Predictions

Feature	Range	Type
Bathrooms	1, 1.5, 2, 2.5, …	Numeric
Bedrooms	1, 2, 3, …	Numeric
Square Feet	100 – 10,000+	Numeric
Crime Rate	Low, Medium, High	Ordered
City	Chicago, Atlanta, New York, …	Unordered
Price	0 - $1,000,000+	Numeric

Data Sources



"He who has the most data, wins."

Data collection can be very time consuming!

Data set sizes:

- 10 – 100 million
- 500 – 10,000 typical

The R and Python languages are well-suited for retrieving and managing data.

Data Preparation

✓ *Clean Data*

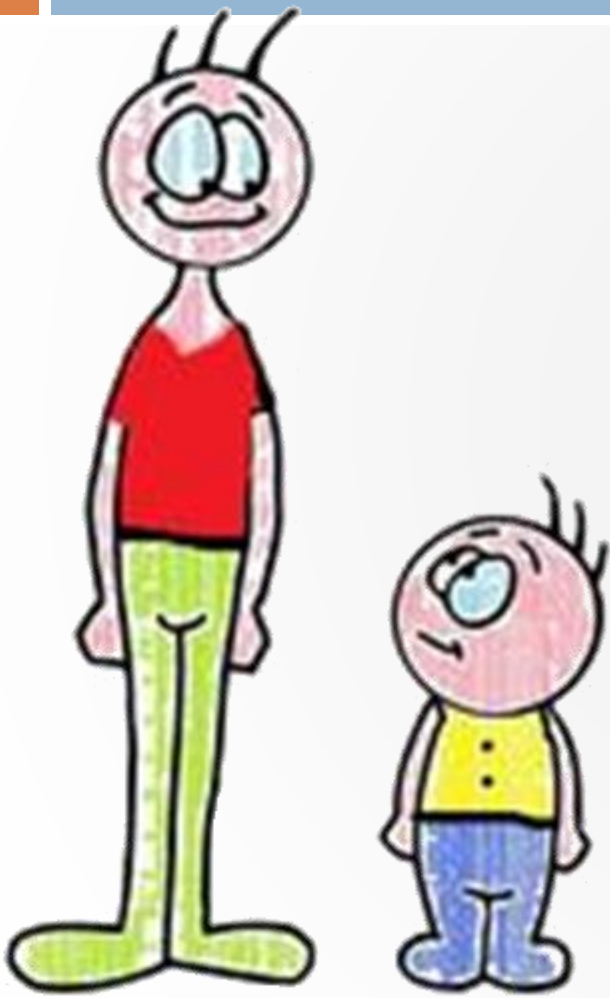
- No missing / incorrect values
- No misspelled categorical values
- No mixed data types

✓ *Tabular Layout*

- Features and targets in columns
- Each row is an "observation"
- Avoid duplicate records

county	district	date	pi.number	pi.c
WILL	01	01/18/2002	63000000	146
WILL	01	01/18/2002	63000000	146
WINNEBAGO	02	01/18/2002	63000000	136
STEPHENSON	02	01/18/2002	63000000	375
WHITESIDE	02	01/18/2002	63000000	446
COOK	01	01/18/2002	63000000	126
WILL	01	01/18/2002	63000000	146
COOK	01	01/18/2002	63000000	276
COOK	01	01/18/2002	63000000	276
COOK	01	01/18/2002	63000000	126
COOK	01	01/18/2002	63000000	276
COOK	01	01/18/2002	63000000	276
COOK	01	01/18/2002	63000000	126
COOK	01	01/18/2002	63000000	126

Data Preparation



Normalization

- Values may vary by several orders of magnitude
- Larger values have greater influence
- Normalization constrains feature value ranges to the same values.
- [0,1] and [-1,1] are common ranges.
- Generally, ~[-3, 3] is acceptable.

Prediction

"Prediction is very difficult, especially about the future."

- Niels Bohr

Image source: http://info.iqms.com/IQMS-Manufacturing-erp-expertise/bid/102603/IQMS-Quality-Assurance-Predictions-With-Carnac-the-Magnificent

Cross-Validation

Generalization

The ability to accurately predict using previously-unseen data

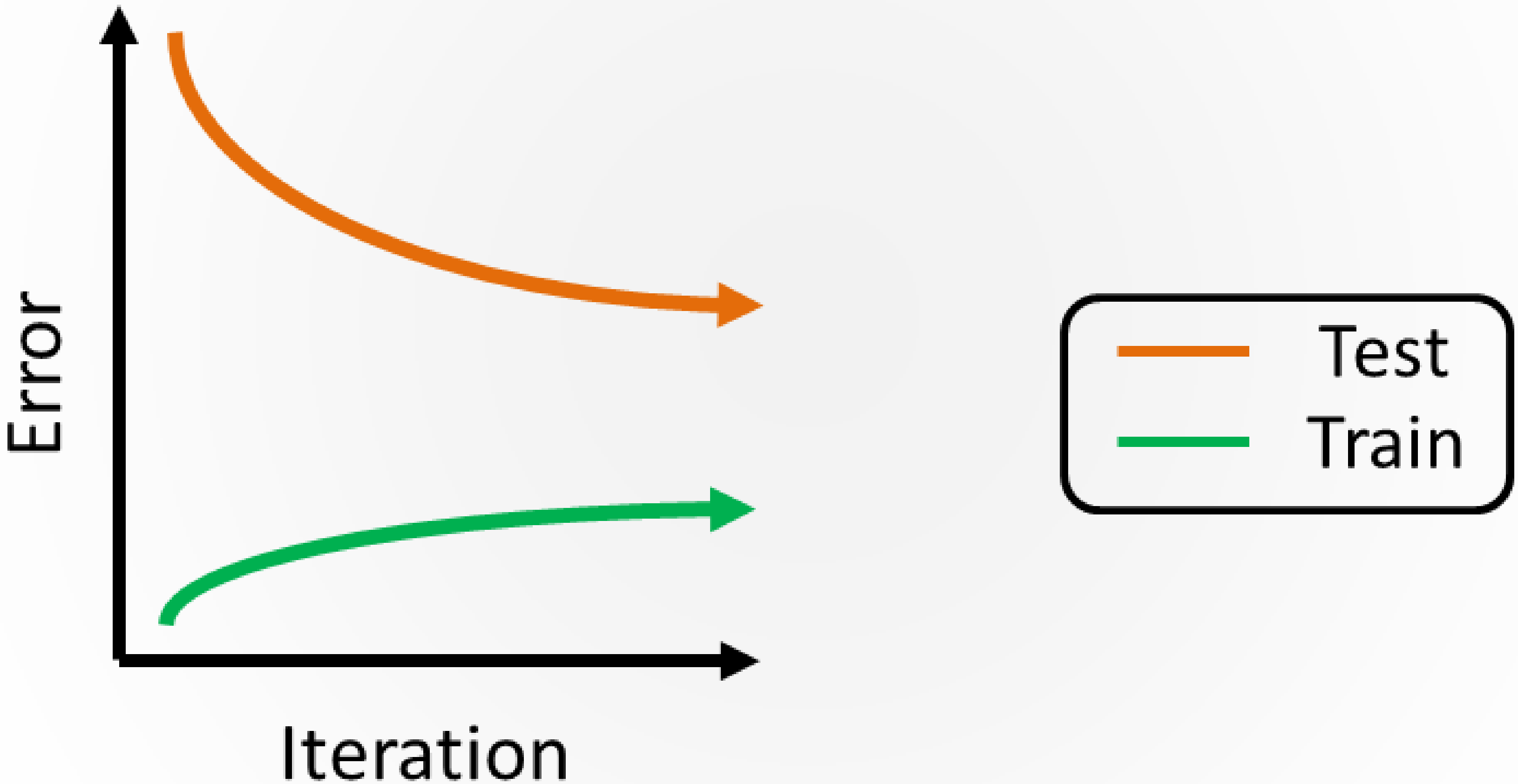
Cross-Validation

Measures how well a model generalizes

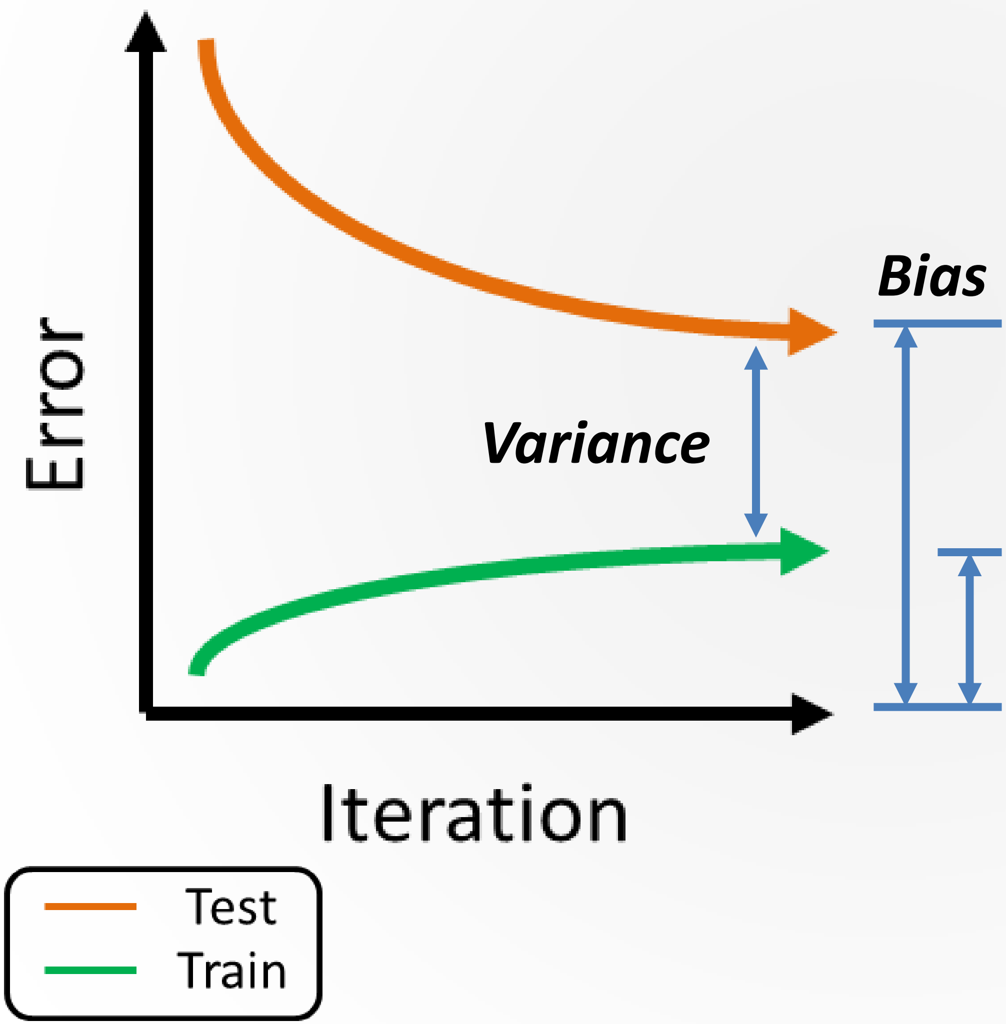
Steps:

1. Split original data set into train and test sets (80/20).
2. Train the model with the larger portion
3. Predict with both the training and testing data.
4. Measure and compare the prediction error in both data sets

Cross-Validation



Cross-Validation



Bias

The total error in each data set

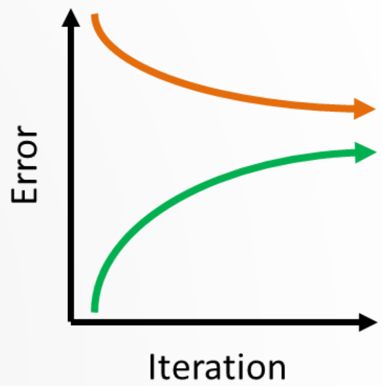
Variance

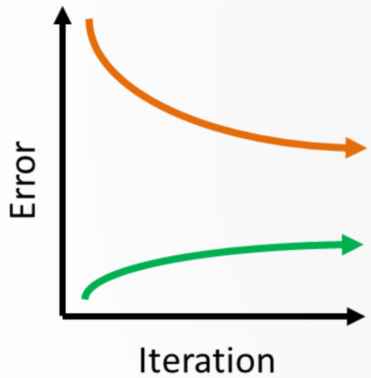
The difference in bias between the data sets

Cross-Validation

Underfit (high bias)

- Does not predict well on either data set
- Need more data, features, better algorithm



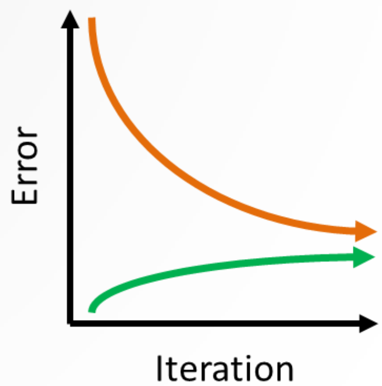


Overfit (high variance)

- Predicts well on the training data, but not the testing data
- Need fewer features, less-powerful algorithm

Good fit

- The network generalizes well on data it has not seen
- Performance on both data sets is similar
- Overall error is low



Measure of Success

A “successful” prediction is a context-specific, subjective determination.

It is not about necessarily about accuracy.

Measure of Success

A meaningful, context-specific statement of how successfully the model predicts.

"On average, the model predicts within __% of the actual value, __% of the time."

Case Studies

Overview

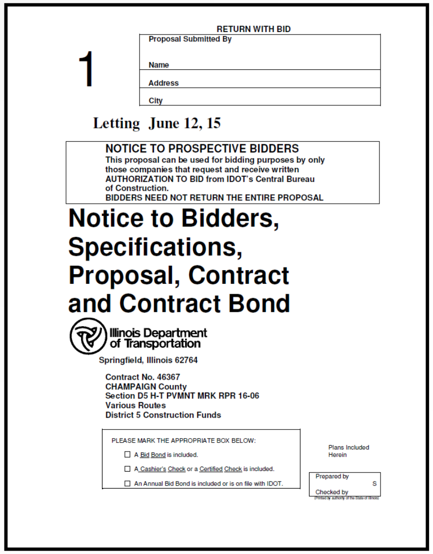
Image source:
http://info.admet.com/blog
/topic/compression-test

Compressive Strength of Concrete Samples

Given a concrete sample's mix design and age,
can we accurately estimate is compressive strength?

Project Cost Estimation

Given a history of contract bid tabulations,
can we accurately estimate unit prices of
contract payitems?



RETURN WITH BID

1

Proposal Submitted By

Name

Address

City

Letting June 12, 15

NOTICE TO PROSPECTIVE BIDDERS
This proposal can be used for bidding purposes by only those companies that request and receive written AUTHORIZATION TO BID from IDOT's Central Bureau of Construction.
BIDDERS NEED NOT RETURN THE ENTIRE PROPOSAL

Notice to Bidders,
Specifications,
Proposal, Contract
and Contract Bond

Illinois Department of Transportation

Springfield, Illinois 62764

Contract No. 46367
CHAMPAIGN County
Section D5 H-T PVMNT MRK RPR 16-06
Various Routes
District 5 Construction Funds

PLEASE MARK THE APPROPRIATE BOX BELOW:
☐ A Bid Bond is included.
☐ A Cashier's Check or a Certified Check is included.
☐ An Annual Bid Bond is included or is on file with IDOT.

Plans Included Herein

Prepared by

Checked by

Concrete Compressive Strength

Data Profile

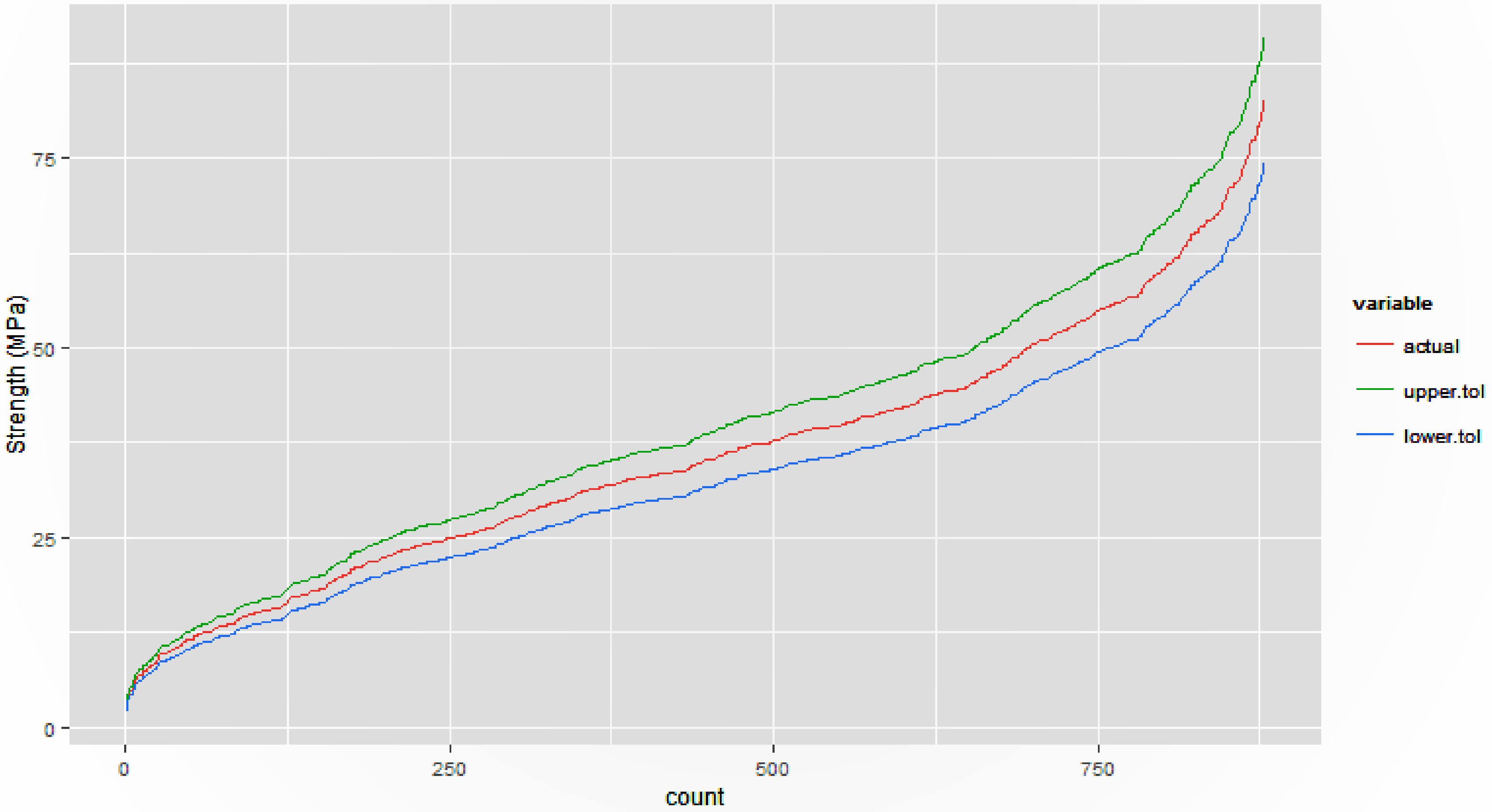
- ❖ Source: University of California, Irvine (UCI) website
- ❖ 1,030 samples (metric units)
- ❖ Non-Linear
- ❖ Features:
  1. Cement
  2. Slag
  3. FlyAsh
  4. Water
  5. Superplasticizer
  6. Coarse Aggregate
  7. Fine Aggregate
  8. Age (days)

A1 Cement

	A	B	C	D	E	F	G	H	I	J
1	Cement	Slag	FlyAsh	Water	Super	CA	FA	Age	Target	
2	540	0	0	162	2.5	1040	676	28	79.99	
3	540	0	0	162	2.5	1055	676	28	61.89	
4	332.5	142.5	0	228	0	932	594	270	40.27	
5	332.5	142.5	0	228	0	932	594	365	41.05	
6	198.6	132.4	0	192	0	978.4	825.5	360	44.3	
7	266	114	0	228	0	932	670	90	47.03	
8	380	95	0	228	0	932	594	365	43.7	
9	380	95	0	228	0	932	594	28	36.45	
10	266	114	0	228	0	932	670	28	45.85	
11	475	0	0	228	0	932	594	28	39.29	
12	198.6	132.4	0	192	0	978.4	825.5	90	38.07	
13	198.6	132.4	0	192	0	978.4	825.5	28	28.02	
14	427.5	47.5	0	228	0	932	594	270	43.01	
15	190	190	0	228	0	932	670	90	42.33	
16	304	76	0	228	0	932	670	28	47.81	
17	380	0	0	228	0	932	670	90	52.91	
18	139.6	209.4	0	192	0	1047	806.9	90	39.36	
19	342	38	0	228	0	932	670	365	56.14	
20	380	95	0	228	0	932	594	90	40.56	
21	475	0	0	228	0	932	594	180	42.62	
22	427.5	47.5	0	228	0	932	594	180	41.84	
23	139.6	209.4	0	192	0	1047	806.9	28	28.24	
24	139.6	209.4	0	192	0	1047	806.9	3	8.06	
25	139.6	209.4	0	192	0	1047	806.9	180	44.21	

Concrete Compressive Strength

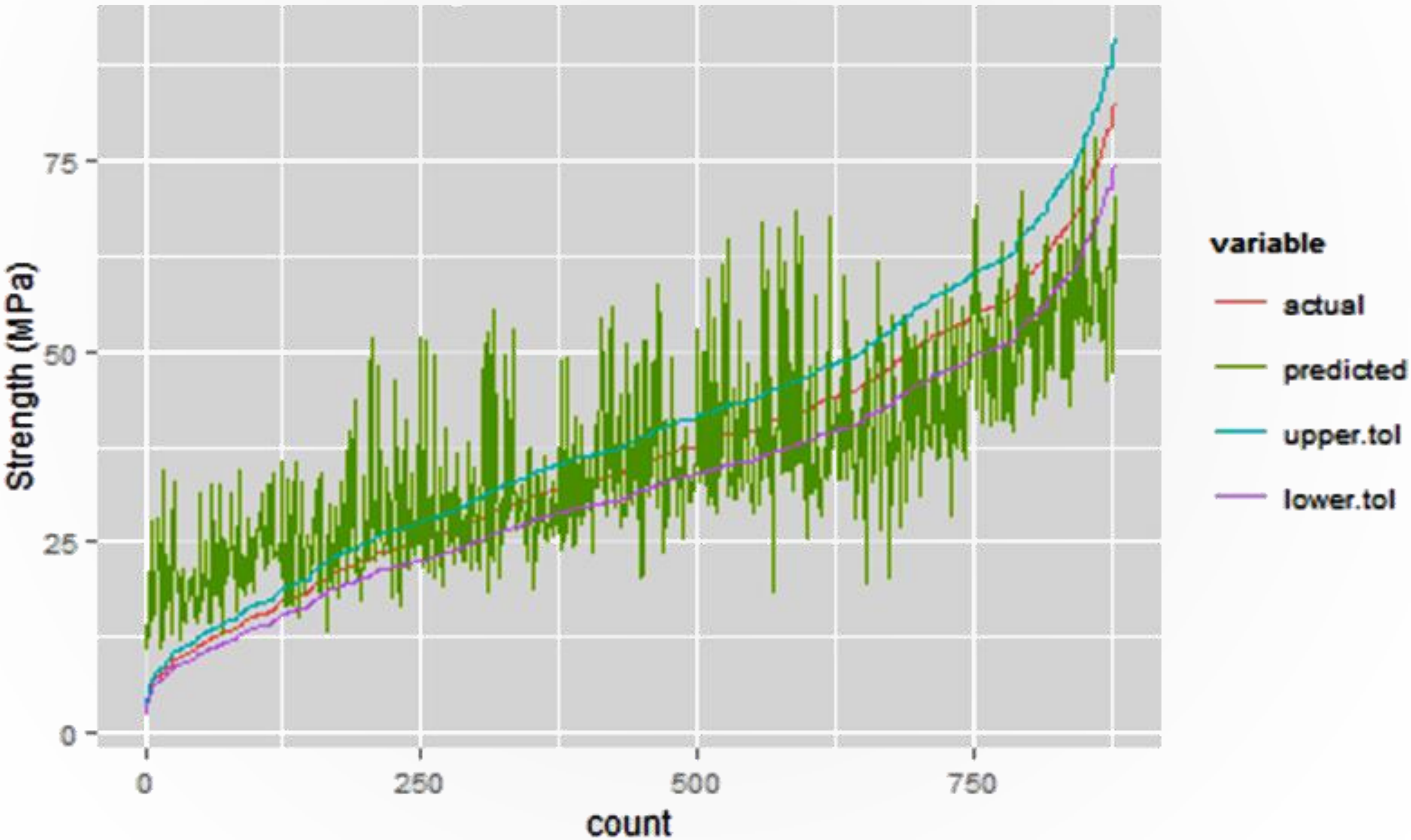
Predictions



Concrete Compressive Strength

Predictions

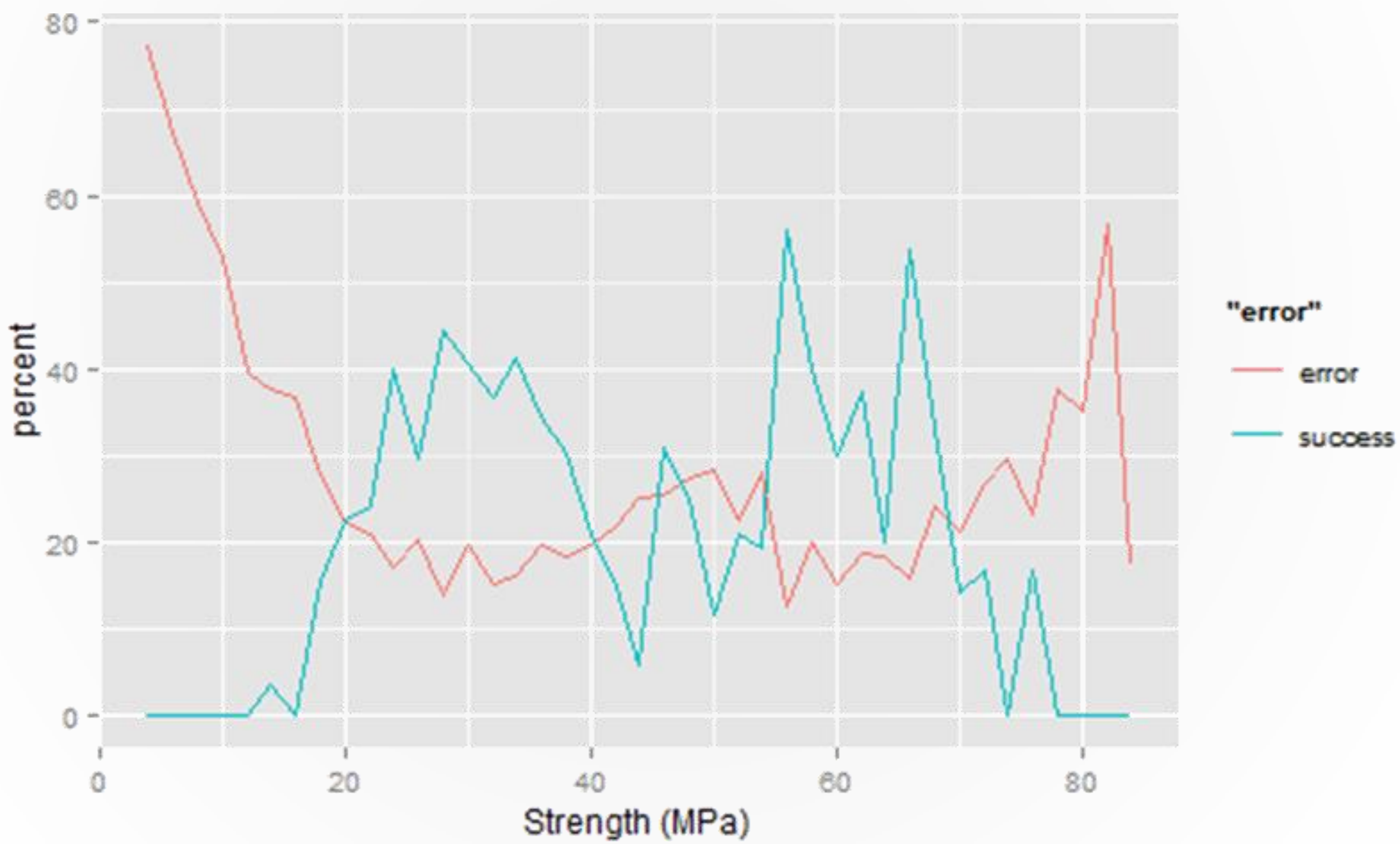
Generalized Linear Regression (train)



Concrete Compressive Strength

Predictions

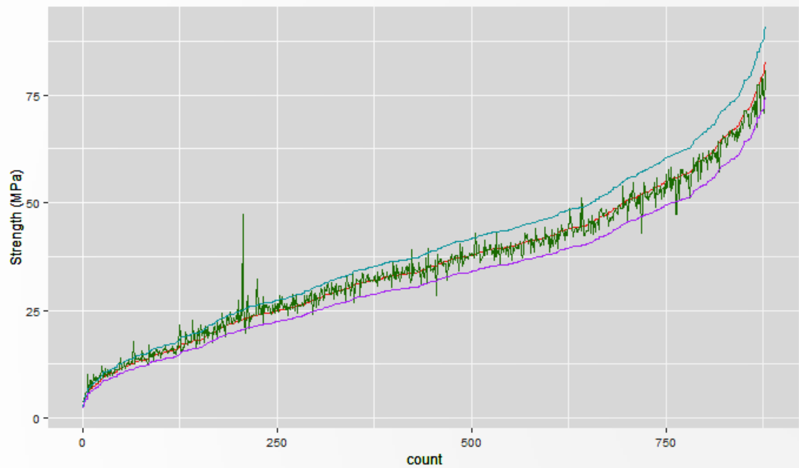
Generalized Linear Regression (train)

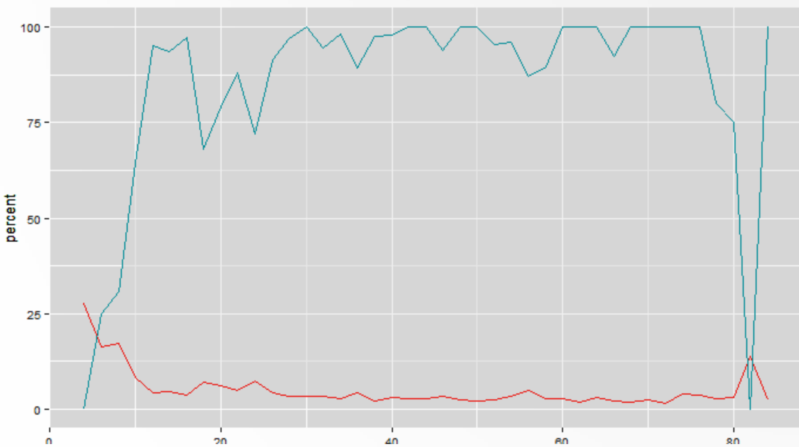


Concrete Compressive Strength

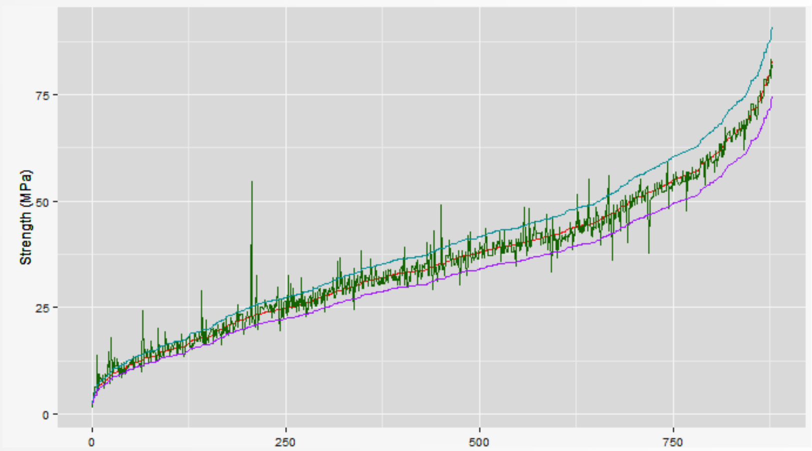
Predictions

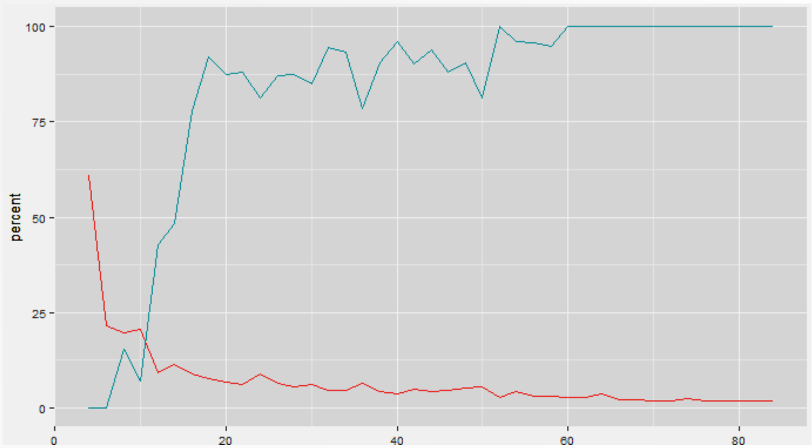
Random Forest (training)





Support Vector Machine (training)





Concrete Compressive Strength

Model Results

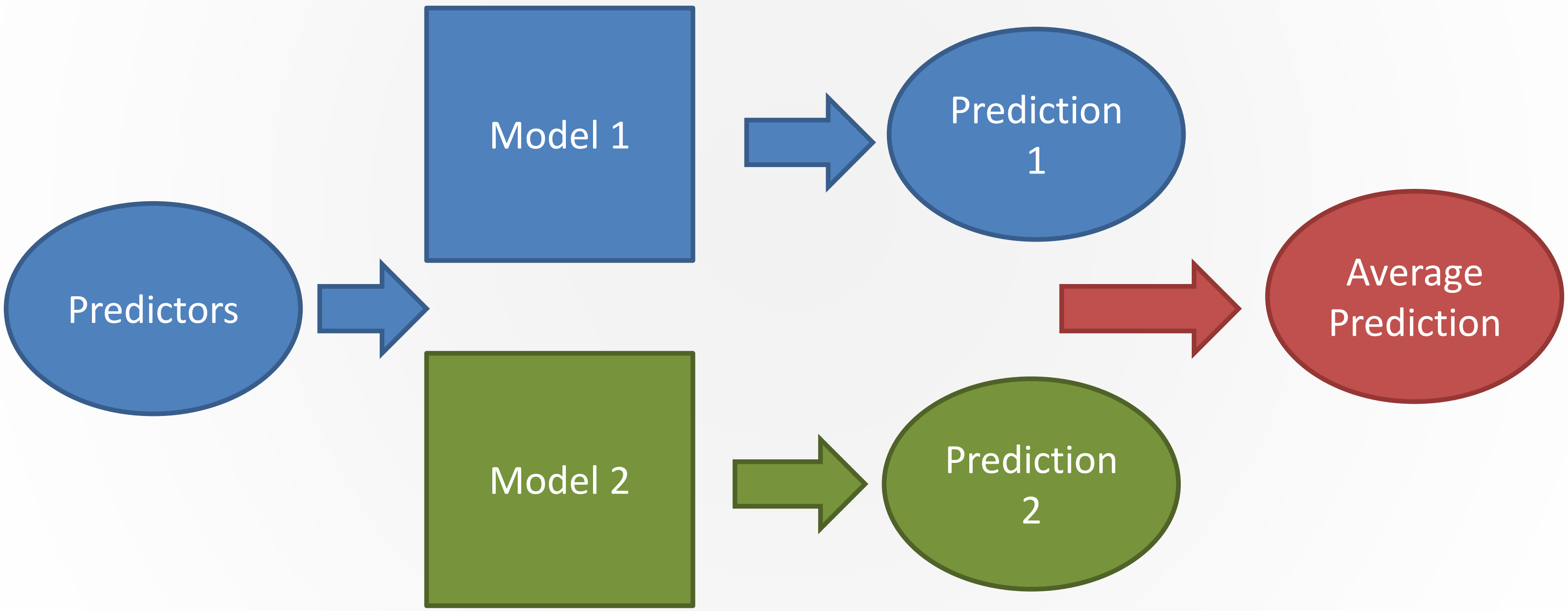
Model	Test Success[1]
Generalized Linear Regression (GLM)	18%
Support Vector Machine (SVM)	52%
Random Forest (RF)	61%
Ensemble[2]	60%
Chained Ensemble[3]	87%

[1]Test Success: Percentage of time model is at least 90% accurate on previously-unseen data.

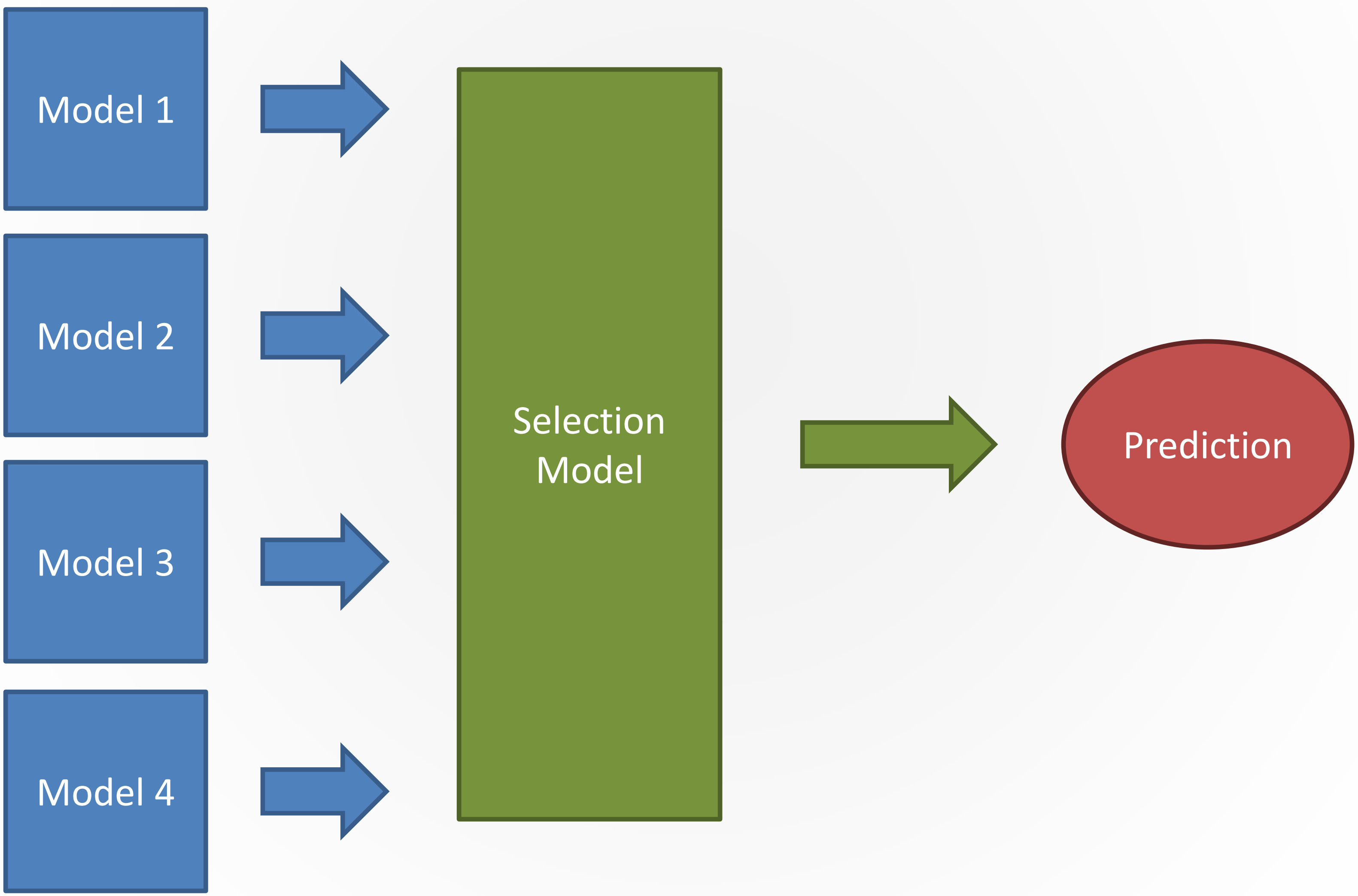
[2]Ensemble: Combination of SVM and RF only.

[3]Chained Ensemble: Predictions of one ensemble are used as inputs to another.

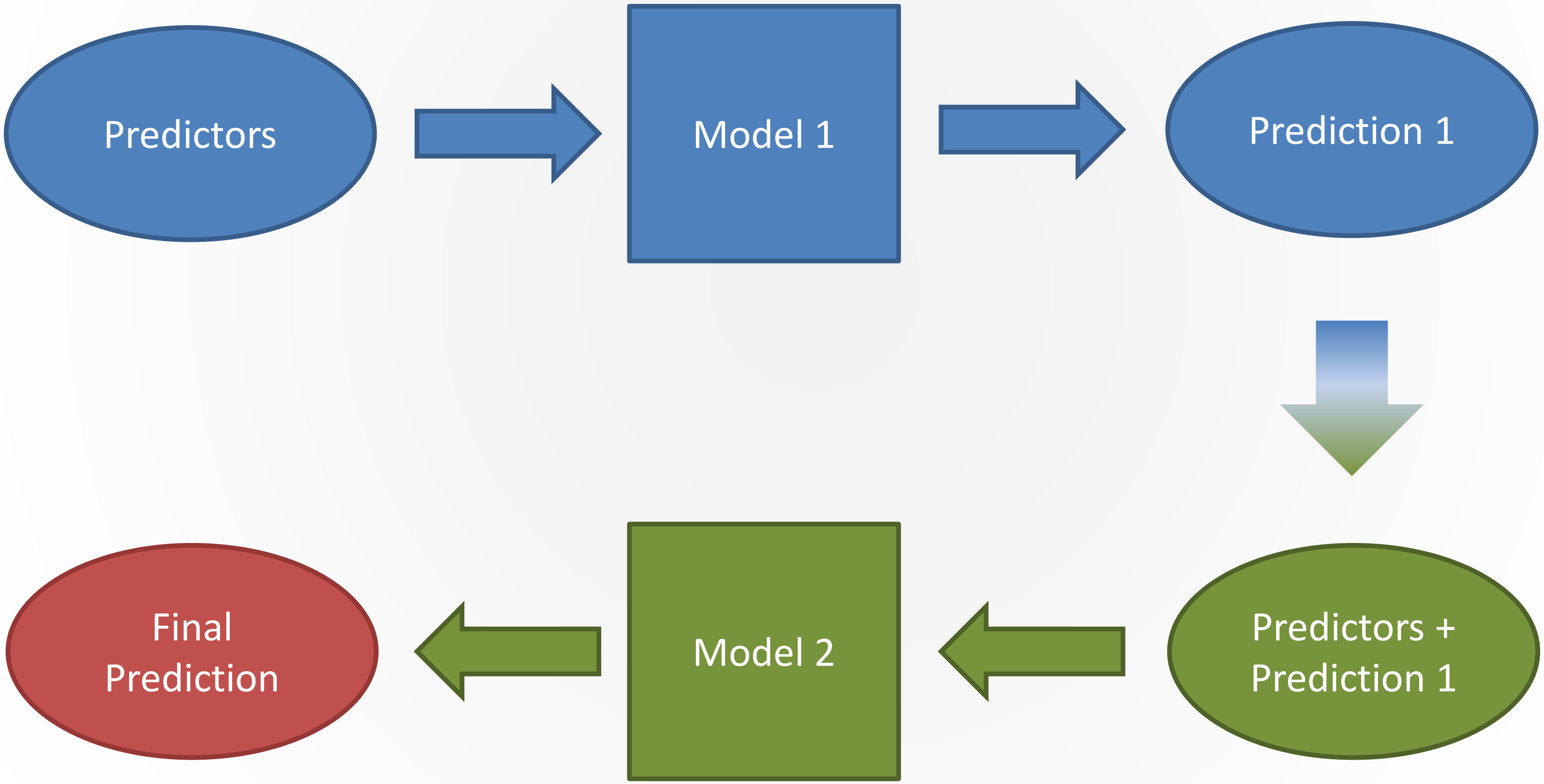
Ensembles - Averaging



Ensembles – Mixture of Experts

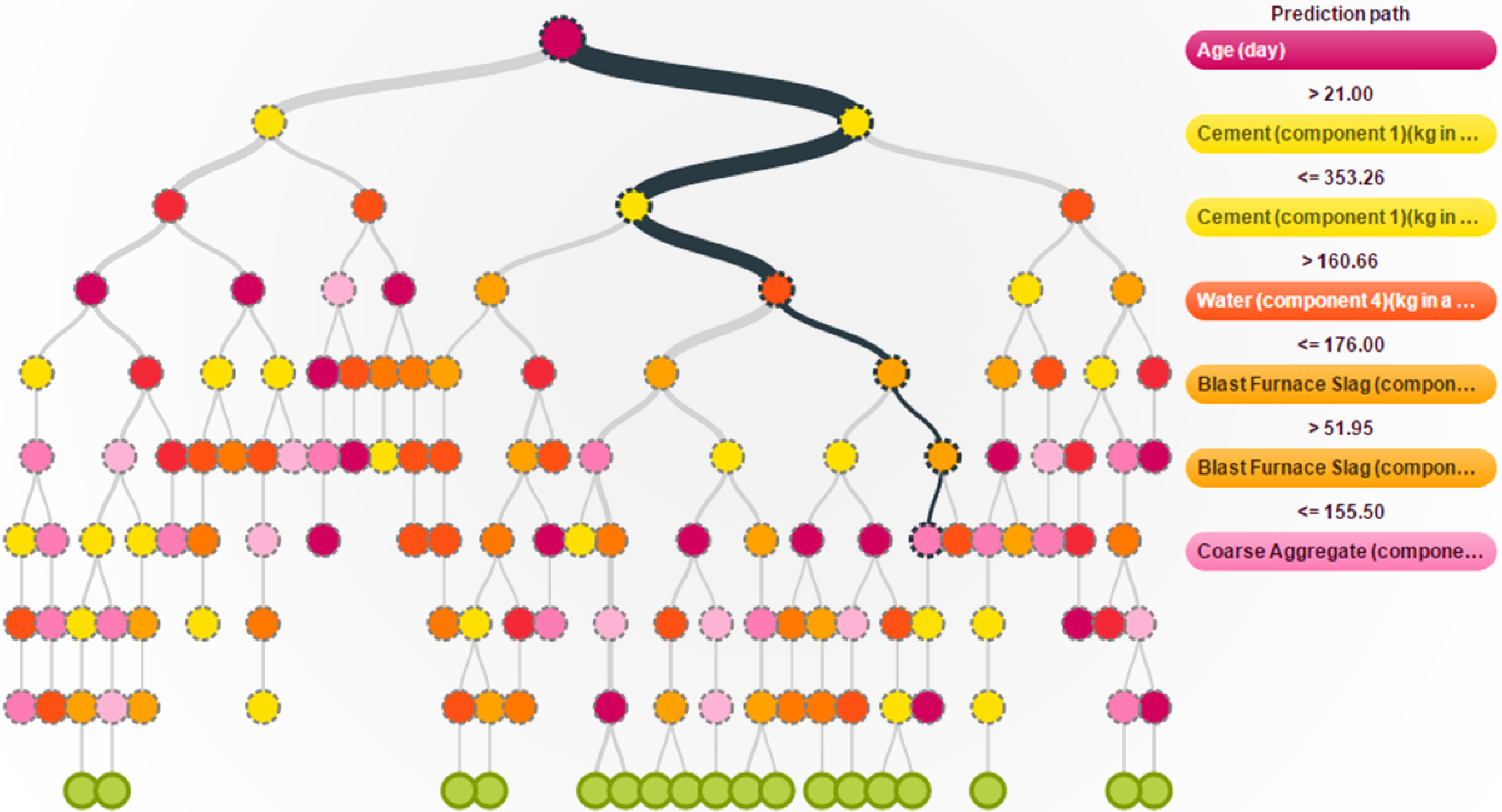


Ensembles - Chained



Concrete Compressive Strength

Feature Importance



Cost Estimation

RETURN WITH BID

Proposal Submitted By

Name

Address

City

1

Letting June 12, 15

NOTICE TO PROSPECTIVE BIDDERS
This proposal can be used for bidding purposes by only those companies that request and receive written AUTHORIZATION TO BID from IDOT's Central Bureau of Construction.
BIDDERS NEED NOT RETURN THE ENTIRE PROPOSAL

Notice to Bidders, Specifications, Proposal, Contract and Contract Bond

Illinois Department of Transportation

Springfield, Illinois 62764

Contract No. 46367
CHAMPAIGN County
Section D5 H-T PVMNT MRK RPR 16-06
Various Routes
District 5 Construction Funds

PLEASE MARK THE APPROPRIATE BOX BELOW:

- ☐ A Bid Bond is included.
- ☐ A Cashier's Check or a Certified Check is included.
- ☐ An Annual Bid Bond is included or is on file with IDOT.

Plans Included Herein

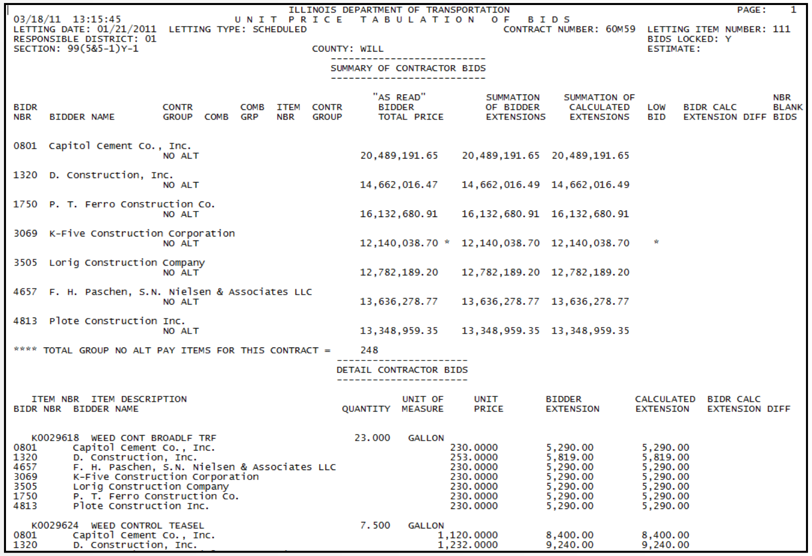
Prepared by S

Checked by

(Printed by authority of the State of Illinois)

Cost Estimation
Data Profile

- Almost 2.5 million records (2002 – 2014)
- Steel Plate Beam Guardrail had 9,580 records
- Nine key features including quantity, time of year, location, and various cost indices



ILLINOIS DEPARTMENT OF TRANSPORTATION PAGE: 1
03/18/11 13:15:45 UNIT PRICE TABULATION OF BIDS
LETTING DATE: 01/21/2011 LETTING TYPE: SCHEDULED CONTRACT NUMBER: 60M59 LETTING ITEM NUMBER: 111
RESPONSIBLE DISTRICT: 01 BIDS LOCKED: Y
SECTION: 99(5&5-1)Y-1 COUNTY: WILL ESTIMATE:

SUMMARY OF CONTRACTOR BIDS

BIDR NBR	BIDDER NAME	CONTR GROUP	COMB	COMB GRP	ITEM NBR	CONTR GROUP	"AS READ" BIDDER TOTAL PRICE	SUMMATION OF BIDDER EXTENSIONS	SUMMATION OF CALCULATED EXTENSIONS	LOW BID	BIDR CALC EXTENSION DIFF	NBR BLANK BIDS
0801	Capitol Cement Co., Inc.	NO ALT					20,489,191.65	20,489,191.65	20,489,191.65			
1320	D. Construction, Inc.	NO ALT					14,662,016.47	14,662,016.49	14,662,016.49			
1750	P. T. Ferro Construction Co.	NO ALT					16,132,680.91	16,132,680.91	16,132,680.91			
3069	K-Five Construction Corporation	NO ALT					12,140,038.70 *	12,140,038.70	12,140,038.70	*		
3505	Lorig Construction Company	NO ALT					12,782,189.20	12,782,189.20	12,782,189.20			
4657	F. H. Paschen, S.N. Nielsen & Associates LLC	NO ALT					13,636,278.77	13,636,278.77	13,636,278.77			
4813	Plote Construction Inc.	NO ALT					13,348,959.35	13,348,959.35	13,348,959.35			

**** TOTAL GROUP NO ALT PAY ITEMS FOR THIS CONTRACT = 248

DETAIL CONTRACTOR BIDS

ITEM NBR / BIDR NBR	ITEM DESCRIPTION / BIDDER NAME	QUANTITY	UNIT OF MEASURE	UNIT PRICE	BIDDER EXTENSION	CALCULATED EXTENSION	BIDR CALC EXTENSION DIFF
K0029618	WEED CONT BROADLF TRF	23.000	GALLON				
0801	Capitol Cement Co., Inc.			230.0000	5,290.00	5,290.00	
1320	D. Construction, Inc.			253.0000	5,819.00	5,819.00	
4657	F. H. Paschen, S.N. Nielsen & Associates LLC			230.0000	5,290.00	5,290.00	
3069	K-Five Construction Corporation			230.0000	5,290.00	5,290.00	
3505	Lorig Construction Company			230.0000	5,290.00	5,290.00	
1750	P. T. Ferro Construction Co.			230.0000	5,290.00	5,290.00	
4813	Plote Construction Inc.			230.0000	5,290.00	5,290.00	
K0029624	WEED CONTROL TEASEL	7.500	GALLON				
0801	Capitol Cement Co., Inc.			1,120.0000	8,400.00	8,400.00	
1320	D. Construction, Inc.			1,232.0000	9,240.00	9,240.00	

- Index Sources
  - IDOT – Bid tabs
  - IDOL - Prevailing wage
  - FBLS - Steel and gas indices
  - FHWA -NHCCI index

Illinois Department of LABOR

U.S. BUREAU OF LABOR STATISTICS

U.S. Department of Transportation Federal Highway Administration

Illinois Department of Transportation

Cost Estimation

Data Source

ILLINOIS DEPARTMENT OF TRANSPORTATION PAGE: 1

03/18/11 13:15:45 U N I T P R I C E T A B U L A T I O N O F B I D S

LETTING DATE: 01/21/2011 LETTING TYPE: SCHEDULED CONTRACT NUMBER: 60M59 LETTING ITEM NUMBER: 111

RESPONSIBLE DISTRICT: 01 BIDS LOCKED: Y

SECTION: 99(5&5-1)Y-1 COUNTY: WILL ESTIMATE:

SUMMARY OF CONTRACTOR BIDS

BIDR NBR	BIDDER NAME	CONTR GROUP	COMB	COMB GRP	ITEM NBR	CONTR GROUP	"AS READ" BIDDER TOTAL PRICE	SUMMATION OF BIDDER EXTENSIONS	SUMMATION OF CALCULATED EXTENSIONS	LOW BID	BIDR CALC EXTENSION DIFF	NBR BLANK BIDS
0801	Capitol Cement Co., Inc.	NO ALT					20,489,191.65	20,489,191.65	20,489,191.65			
1320	D. Construction, Inc.	NO ALT					14,662,016.47	14,662,016.49	14,662,016.49			
1750	P. T. Ferro Construction Co.	NO ALT					16,132,680.91	16,132,680.91	16,132,680.91			
3069	K-Five Construction Corporation	NO ALT					12,140,038.70 *	12,140,038.70	12,140,038.70	*		
3505	Lorig Construction Company	NO ALT					12,782,189.20	12,782,189.20	12,782,189.20			
4657	F. H. Paschen, S.N. Nielsen & Associates LLC	NO ALT					13,636,278.77	13,636,278.77	13,636,278.77			
4813	Plote Construction Inc.	NO ALT					13,348,959.35	13,348,959.35	13,348,959.35			

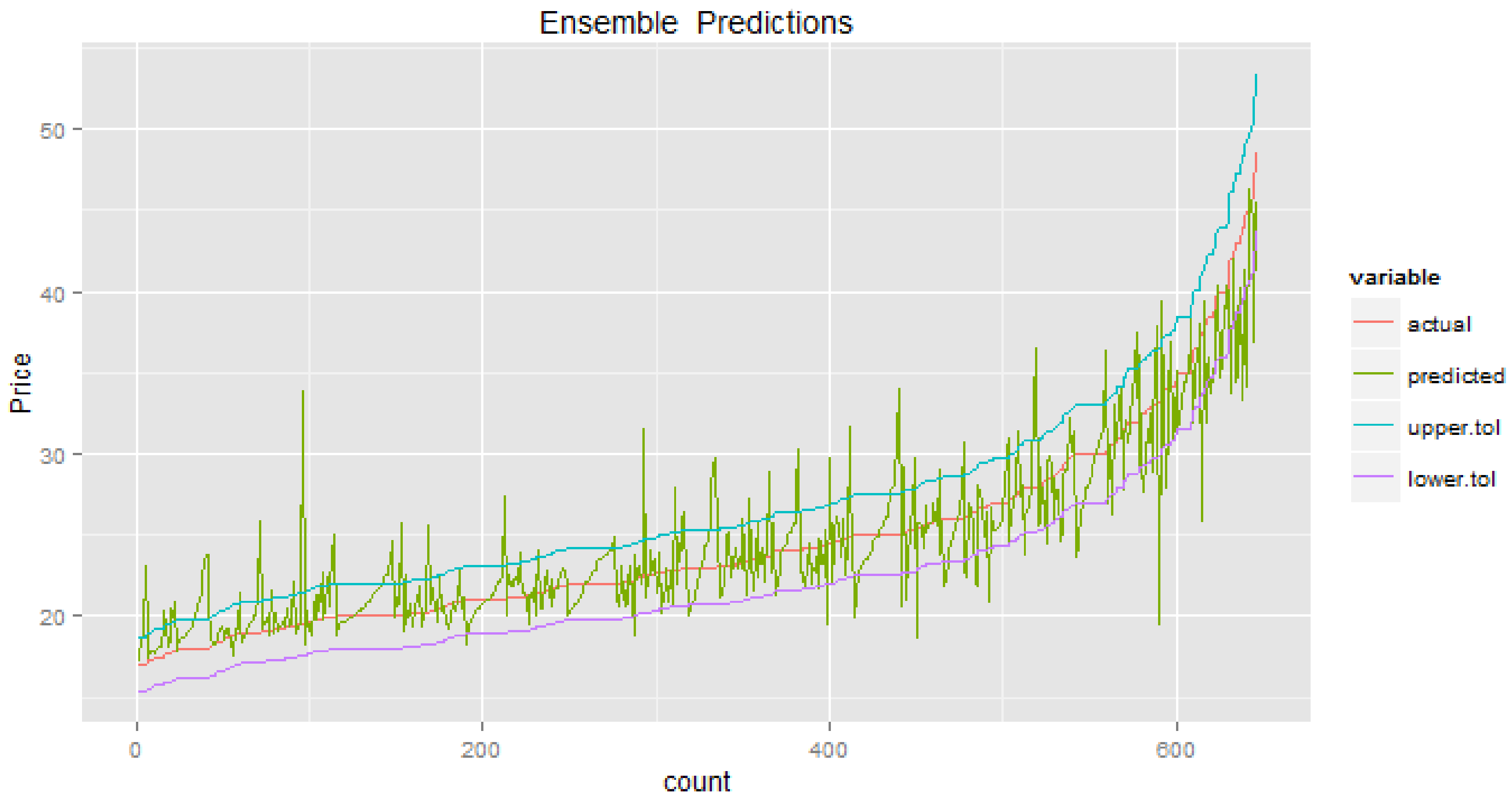
**** TOTAL GROUP NO ALT PAY ITEMS FOR THIS CONTRACT = 248

DETAIL CONTRACTOR BIDS

ITEM NBR / BIDR NBR	ITEM DESCRIPTION / BIDDER NAME	QUANTITY	UNIT OF MEASURE	UNIT PRICE	BIDDER EXTENSION	CALCULATED EXTENSION	BIDR CALC EXTENSION DIFF
K0029618	WEED CONT BROADLF TRF	23.000	GALLON				
0801	Capitol Cement Co., Inc.			230.0000	5,290.00	5,290.00	
1320	D. Construction, Inc.			253.0000	5,819.00	5,819.00	
4657	F. H. Paschen, S.N. Nielsen & Associates LLC			230.0000	5,290.00	5,290.00	
3069	K-Five Construction Corporation			230.0000	5,290.00	5,290.00	
3505	Lorig Construction Company			230.0000	5,290.00	5,290.00	
1750	P. T. Ferro Construction Co.			230.0000	5,290.00	5,290.00	
4813	Plote Construction Inc.			230.0000	5,290.00	5,290.00	
K0029624	WEED CONTROL TEASEL	7.500	GALLON				
0801	Capitol Cement Co., Inc.			1,120.0000	8,400.00	8,400.00	
1320	D. Construction, Inc.			1,232.0000	9,240.00	9,240.00	

Cost Estimation

Predictions



Cost Estimation

Model Results

Model	Test Success[1]
Generalized Linear Regression (GLM)	55%
Support Vector Machine (SVM)	71%
Random Forest (RF)	74%
Ensemble[2]	74%
Chained Ensemble[3]	89%

[1]Test Success: Percentage of time model is at least 85% accurate on previously-unseen data.

[2]Ensemble: Combination of SVM and RF only.

[3]Chained Ensemble: Predictions of one ensemble are used as inputs to another.


So, your support vector machine has high bias and low variance?
Tell me more...

So What?

What can this technology really do for us?

The answer lies in asking the right question.

So What?

"Given _______,
can we determine ______
with ______ accuracy?"

So What?

The question has three key ingredients:

- The Givens (features / predictors)
- The Goal (target / prediction)
- The Accuracy (measure of success)

So What?

Using this question you can determine:

- What data you need
- The types of the data
- How successful the algorithm needs to be

A single dataset may answer a variety of different questions. We can use these three elements to determine those questions

So What?

Construction Scheduling

*Given the **strength and mix design,***
*can I determine the **time it will take to cure***
*with **95%** accuracy?*

So What?

QC / QA

Given the ***compressive strength and cure time*** *can I determine the* ***most valid mix design*** *with* ***90%*** *accuracy?*

So What?

A trained model ***runs instantaneously*** and has ***flexible deployment*** options:

- Spreadsheet or database backend
- iPhone or Android mobile app
- Web app

So What?

bigml

✓ Provides an online interface

✓ Users can sign up for a free account.

✓ Provides tools to prepare data and analyze model results

✓ Numerous resources are available online to begin learning.

Requires a significant time commitment, algebra-level math skills and basic grasp of elementary statistics

So What?

Most problems don't require complex ML solutions to get useful results…

- Mean
- Median
- Standard Deviation
- Distribution
- Confidence Interval

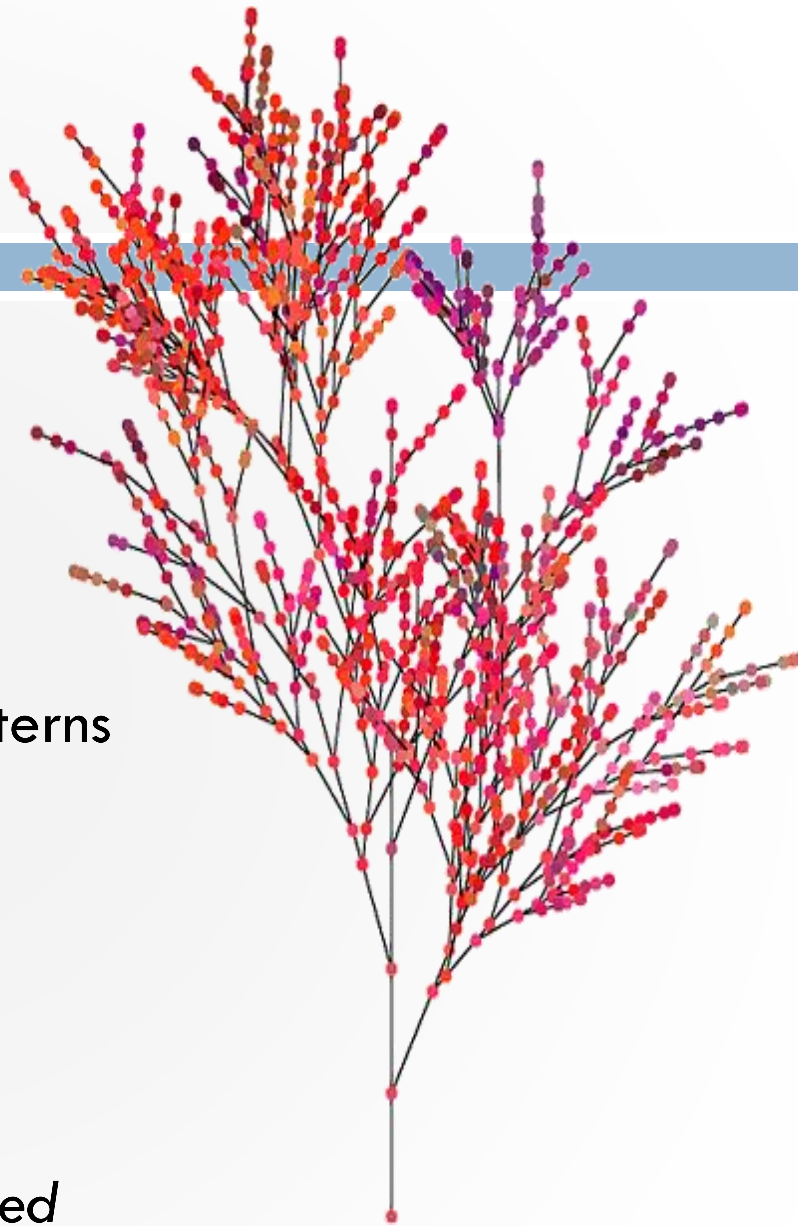
Conclusion

Scope

- Small to medium-sized data sets
- Strong feature-target correlations
- Where there's data, there are patterns

Implementation

- *Up-front time commitment*
- *Instantaneous feedback*
- *Predictive accuracy is well-established*



"Luc's random forest"
http://2things.tumblr.com/post/28394765/lucs-random-forest

Additional Resources

BigML.com (http://www.bigml.com)
On-line machine learning and data visualization tools

The R Project (http://www.r-project.org/)
Free scripting language for statistical computing and graphics

Coursera (http://www.coursera.org)
Free on-line college-level courses in technology and other topics

UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/)
Wide range of data sets for machine learning applications