U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

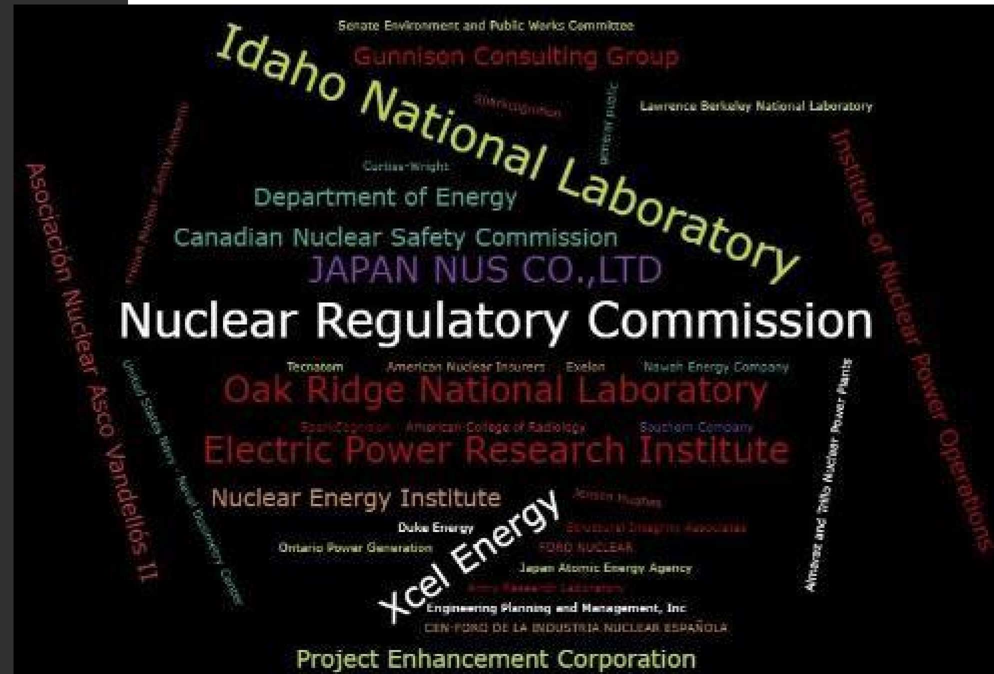
DATA SCIENCE AND ARTIFICIAL INTELLIGENCE REGULATORY APPLICATIONS WORKSHOPS

Opening Remarks

Theresa Lalain, Ph.D.
Deputy Director, Division of Systems Analysis
Office of Nuclear Regulatory Research

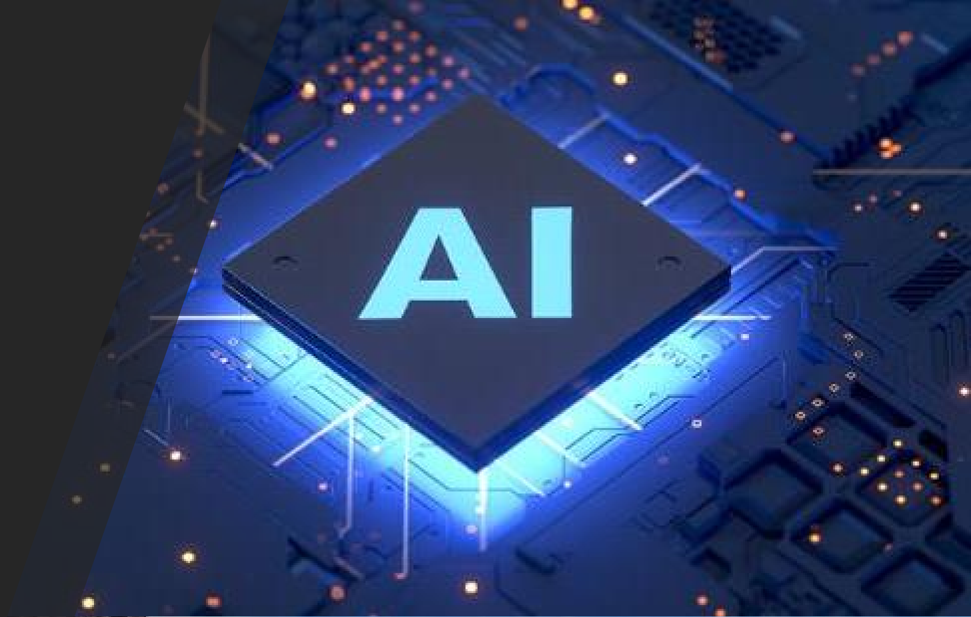
WELCOME

- Over 250 registered attendees
- Participation from U.S., Canada, Spain, France, UAE, and Japan



Regulatory Purpose

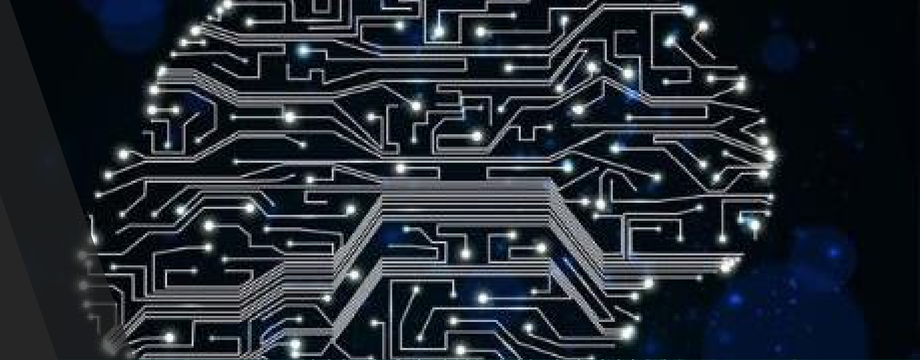
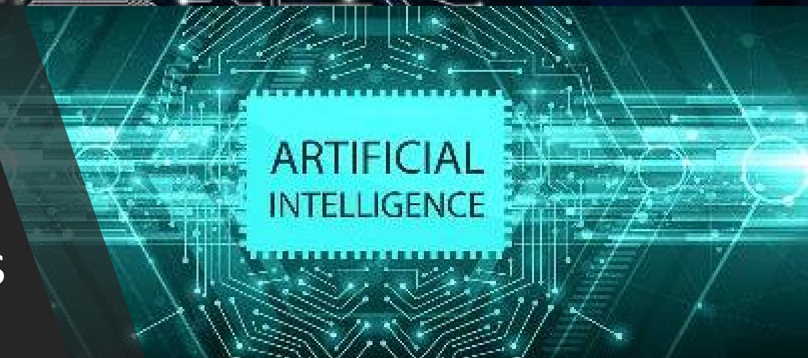
- NRC recognizes a need to use data analytics for regulatory enhancements as part of its effort to become a modern, risk-informed regulator[1]
- The nuclear industry is investigating and using AI applications; therefore, the NRC must be prepared to understand and evaluate the technology



[1] "The Dynamic Futures for NRC Mission Areas," (ML19022A178)

Data Science and Artificial Intelligence Overview

- Artificial Intelligence (AI)
  - Build “intelligent” smart machines
- Machine Learning (ML)
  - Learn from data and deliver predictive models
- Natural Language Processing (NLP)
  - Process and analyze large amounts of natural language data
- Deep Learning (DL)
  - ML methods based on artificial neural networks

Engagement and Initiatives

National Laboratories

FEDERAL REGISTER
The Daily Journal of the United States Government

SHARING KNOWLEDGE AND SEEKING STAKEHOLDER INPUT

LEVERAGING RESEARCH ACTIVITIES

EPRI | ELECTRIC POWER RESEARCH INSTITUTE

LWRS
LIGHT WATER REACTOR SUSTAINABILITY

NIST

arpa·e
CHANGING WHAT'S POSSIBLE

NRC is engaging and participating with external entities to best prepare for AI impacts on regulatory processes and decisionmaking

Upcoming Workshops

- Current Topics
  - AUGUST 2021
- Future Focused Initiatives
  - SEPT/OCT 2021

June 29, 2021

Ronald Laurids Boring, PhD, FHFES

Introduction to Artificial Intelligence (AI) and Some of Its Basic Terminology

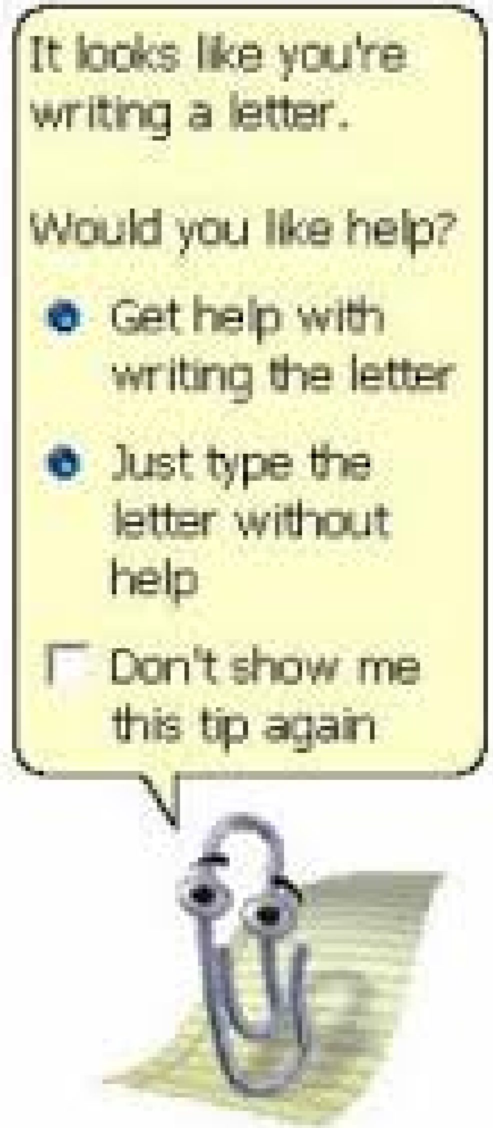
U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

INL Idaho National Laboratory

Is this AI?

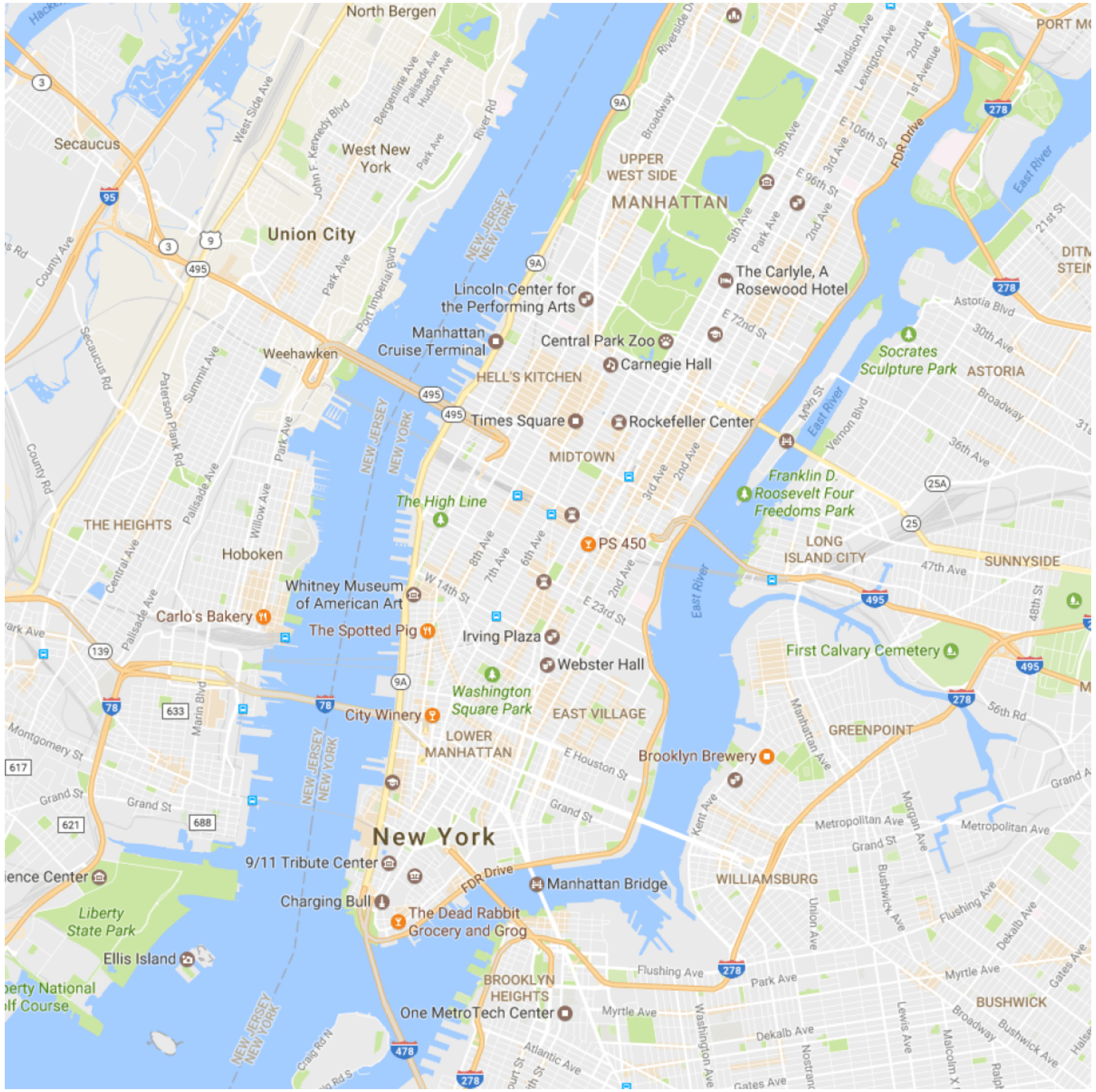
Decoder Ring

Is this AI?



Microsoft Clippy

Is this AI?

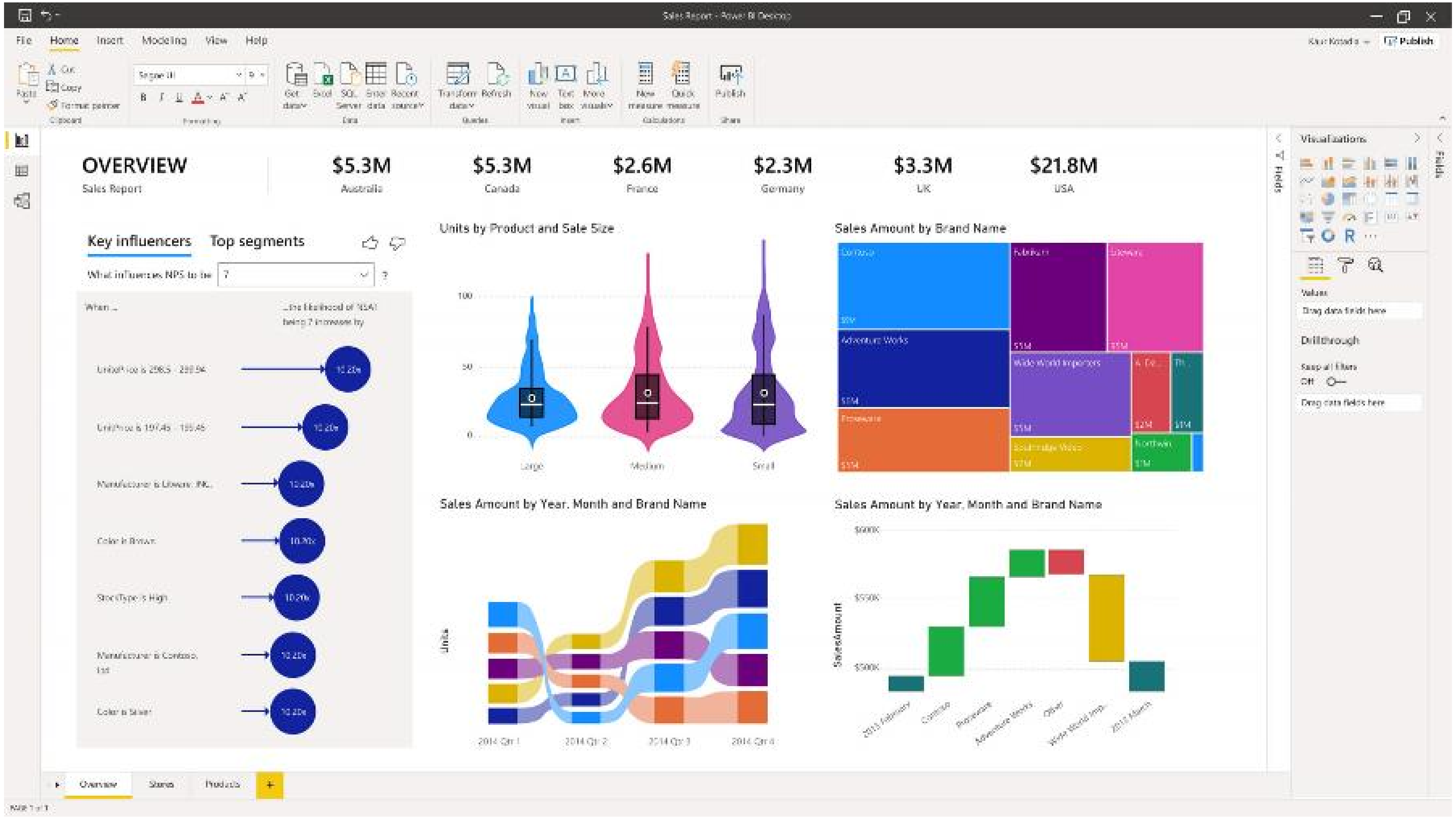


Google Maps

Is this AI?

Apple Watch Heartrate Monitor

Is this AI?



Microsoft Power BI Sample Dashboard

Is this AI?

NuScale Power Control Room Simulator

They All Feature Applications of AI

Let’s Look at Some of the History and Technology Underlying AI

It all began in

1956

1956 Was a Watershed Year

- **Two Congressional Hearings on Automation**
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- **Two Congressional Hearings on Automation**
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- Two Congressional Hearings on Automation
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- **Two Congressional Hearings on Automation**
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of **learning** or any other feature of **intelligence** can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- **Two Congressional Hearings on Automation**
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- **Two Congressional Hearings on Automation**
- **Dartmouth Summer Workshop on Artificial Intelligence**
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- **Symposium on Information Theory at MIT on September 11, 1956**
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- **Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems**
  - Deconstructing human thinking into information allowed us to make computer models of it

1956 Was a Watershed Year

- Two Congressional Hearings on Automation
- Dartmouth Summer Workshop on Artificial Intelligence
  - "We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."
  - Birth of AI, featuring founders like Marvin Minsky, John McCarthy, Claude Shannon, Allen Newell, and Herb Simon
- Symposium on Information Theory at MIT on September 11, 1956
  - Birthplace of information processing theory and study of cognition
  - Featured George Miller, Noam Chomsky, Allen Newell, and Herb Simon, and others
- Birth of AI and cognitive psychology occurred at the same time, because they were interested in the same problems
  - Deconstructing human thinking into information allowed us to make computer models of it

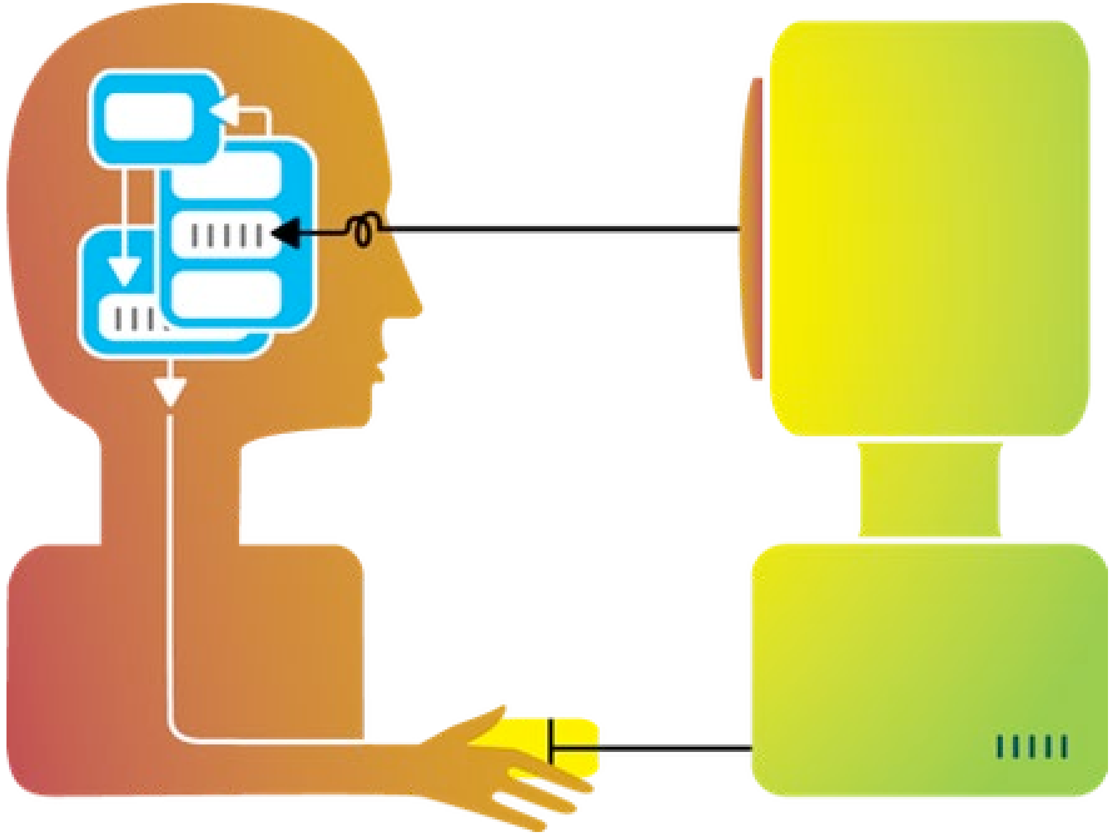
Big Picture in Information Processing

Human-System Interface (HSI)

- Computer output = human sensation and perception
- Human action = computer input
- It's a feedback loop

Each step also represents a form of intelligence that may be modelled artificially

- Perception: Pattern recognition, computer vision, natural language processing
- Knowledge: Expert systems
- Actions and behaviors: Automated controllers



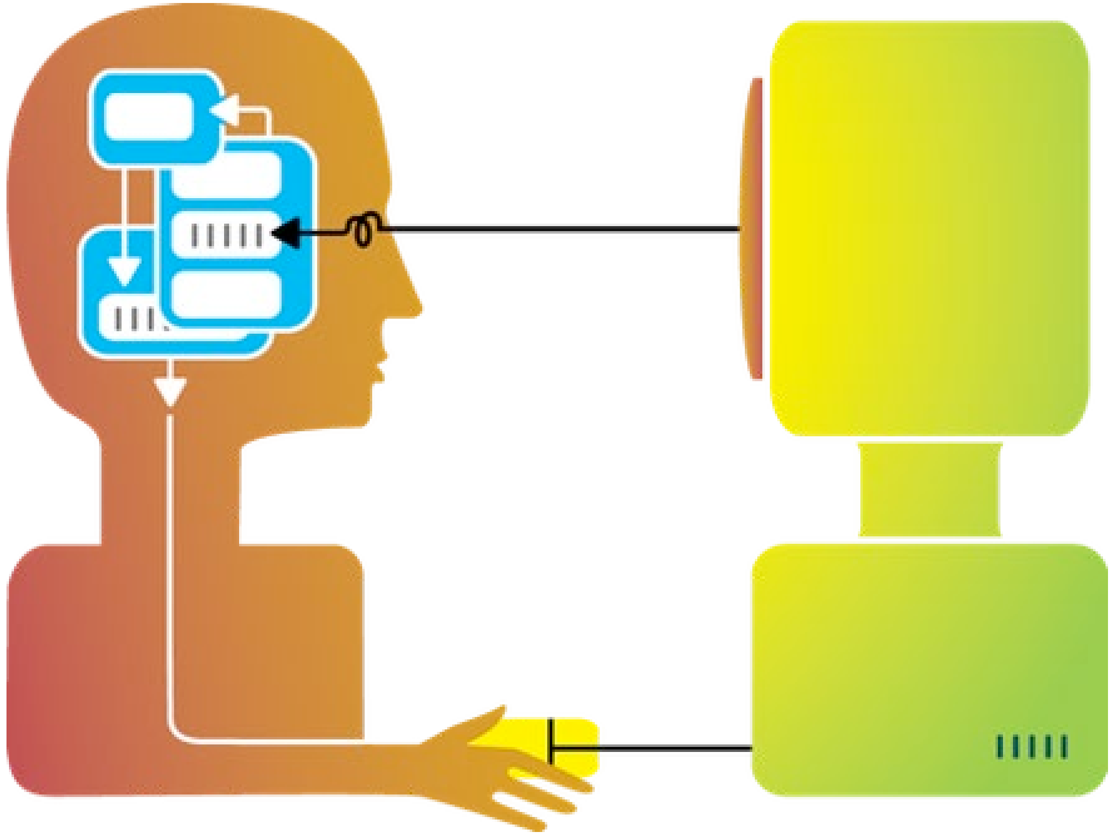
Big Picture in Information Processing

Human-System Interface (HSI)

- Computer output = human sensation and perception
- Human action = computer input
- It's a feedback loop

Each step also represents a form of intelligence that may be modelled artificially

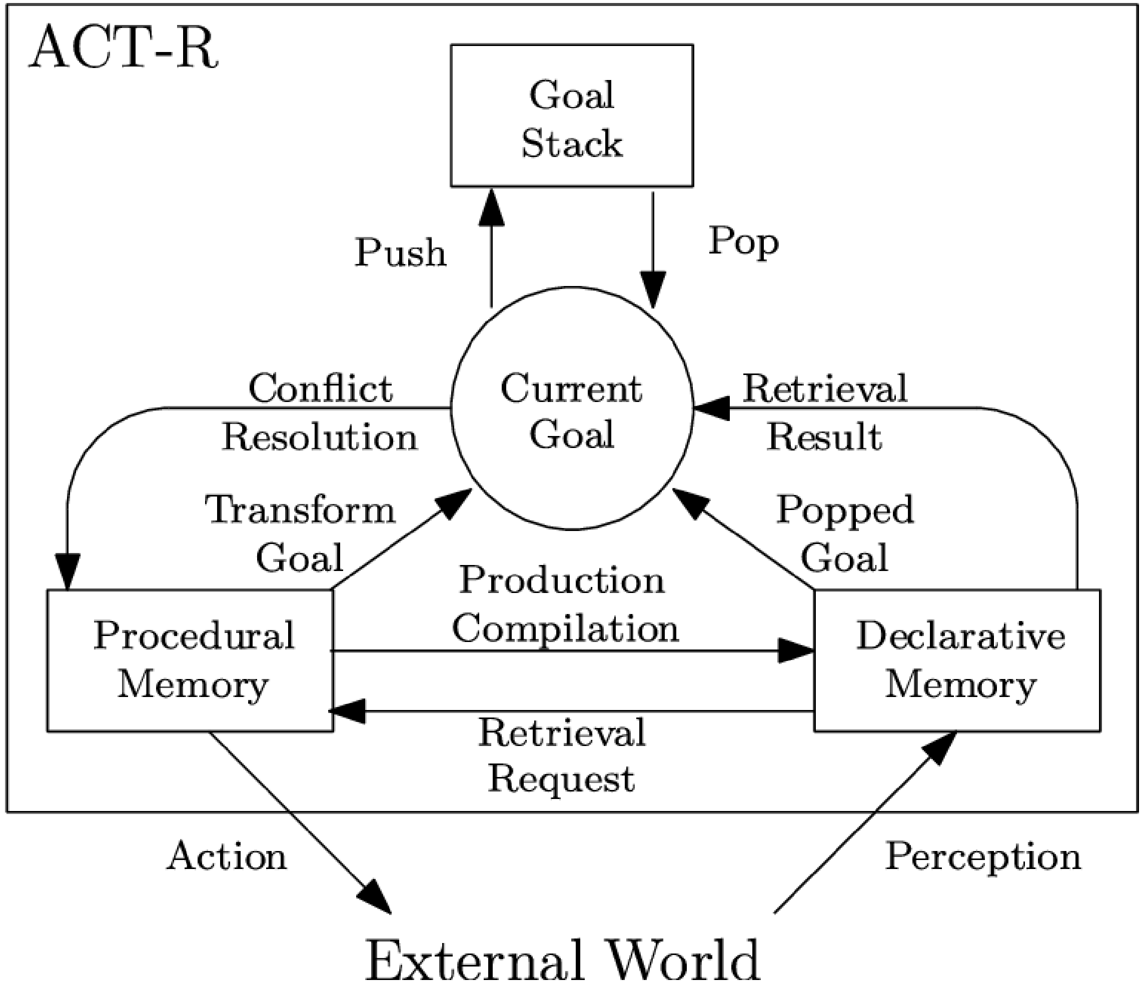
- Perception: Pattern recognition, computer vision, natural language processing
- Knowledge: Expert systems
- Actions and behaviors: Automated controllers



How Does AI Work?

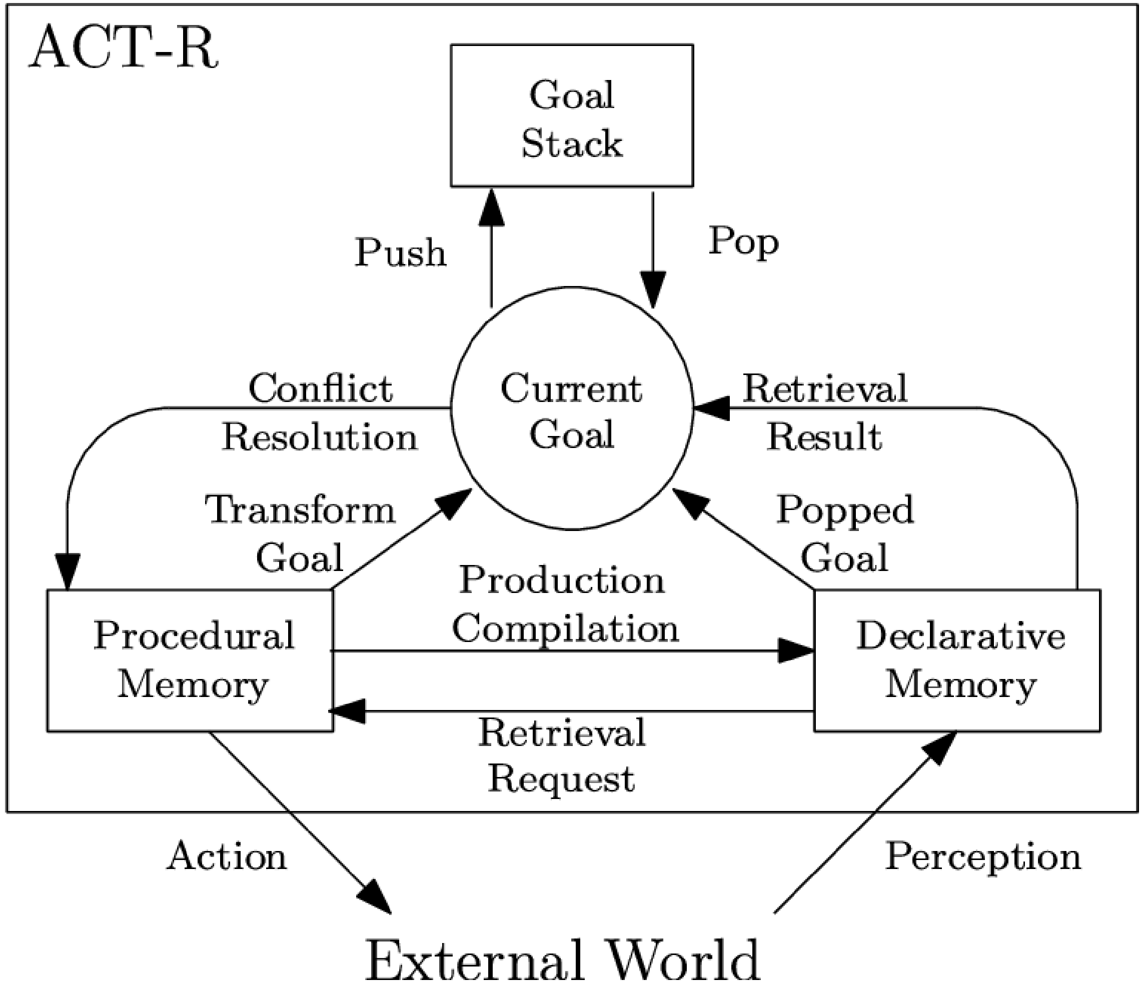
Two Types of AI

- **Good Old-Fashioned AI (GOFAI)**
  - Symbolic logic systems to represent basic elements of human thought like language, numbers, or goals
  - **Production systems** featuring if-then logic
    - General Problem Solver created by Newell and Simon in 1959
  - **Cognitive modeling architectures**
    - Systems like Soar and ACT-R with a heavy emphasis on how humans accomplish goals
  - Much of focus is not to create learning but to capture human-like intelligence related to how humans carry out decisions and actions



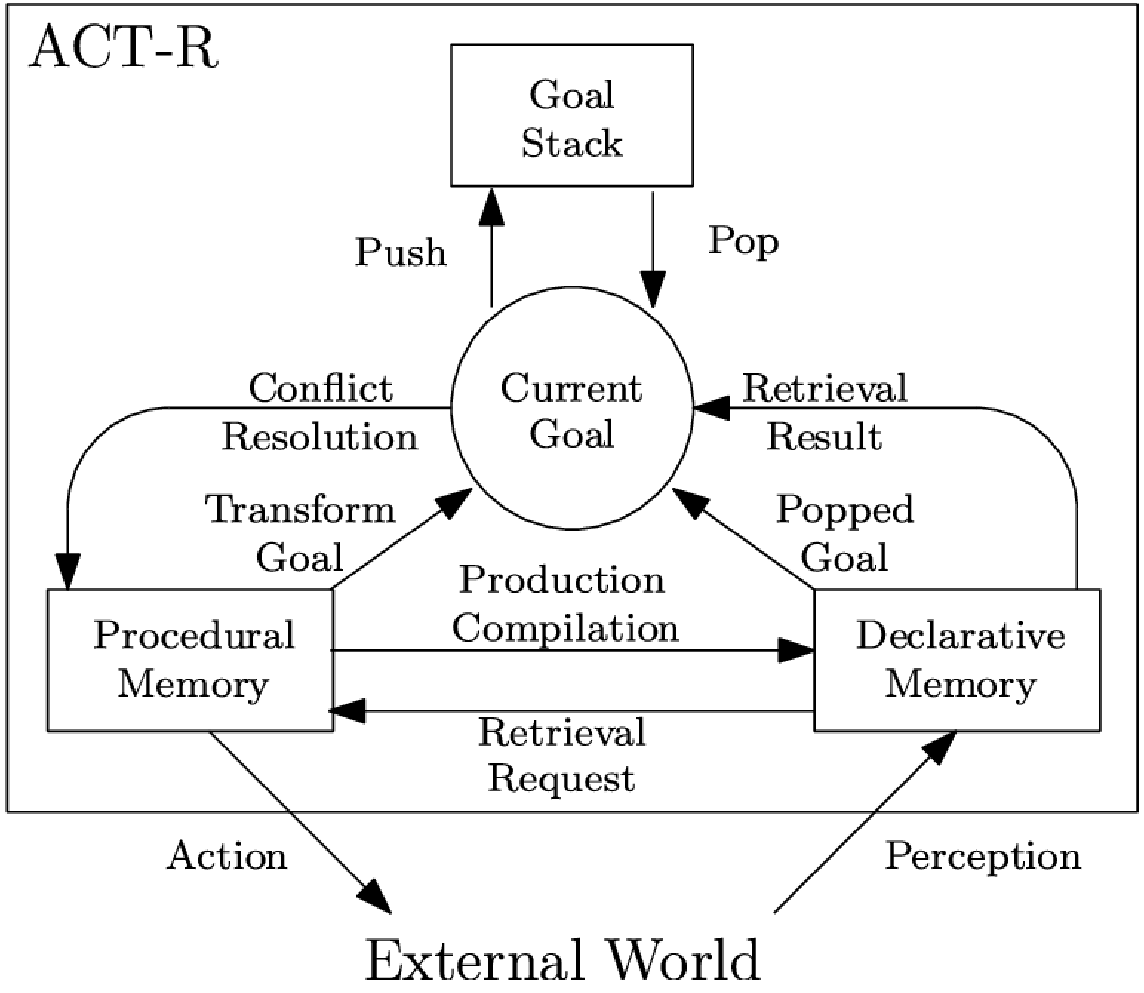
Two Types of AI

- **Good Old-Fashioned AI (GOFAI)**
  - Symbolic logic systems to represent basic elements of human thought like language, numbers, or goals
  - **Production systems** featuring if-then logic
    - General Problem Solver created by Newell and Simon in 1959
  - **Cognitive modeling architectures**
    - Systems like Soar and ACT-R with a heavy emphasis on how humans accomplish goals
  - Much of focus is not to create learning but to capture human-like intelligence related to how humans carry out decisions and actions



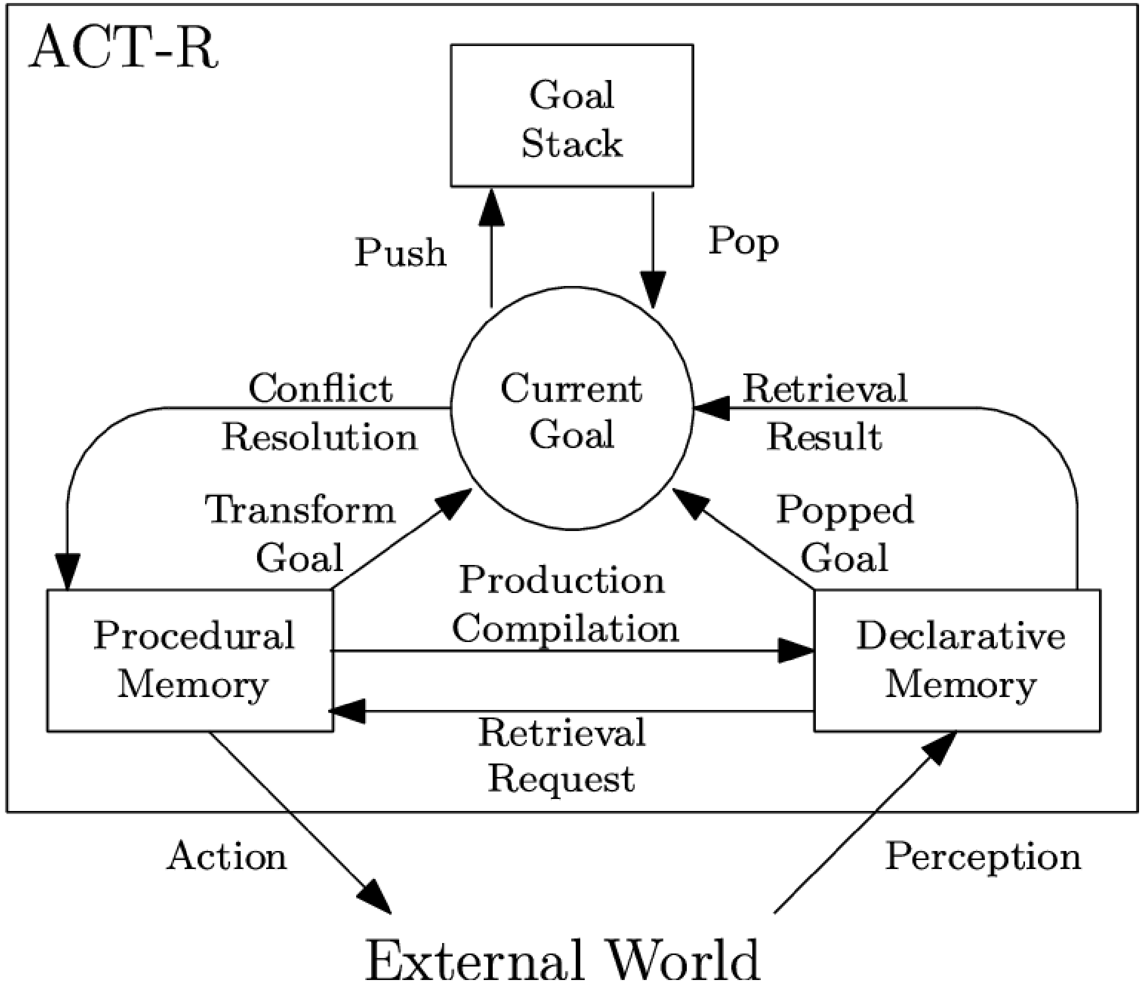
Two Types of AI

- **Good Old-Fashioned AI (GOFAI)**
  - Symbolic logic systems to represent basic elements of human thought like language, numbers, or goals
  - **Production systems** featuring if-then logic
    - General Problem Solver created by Newell and Simon in 1959
  - **Cognitive modeling architectures**
    - Systems like Soar and ACT-R with a heavy emphasis on how humans accomplish goals
  - Much of focus is not to create learning but to capture human-like intelligence related to how humans carry out decisions and actions



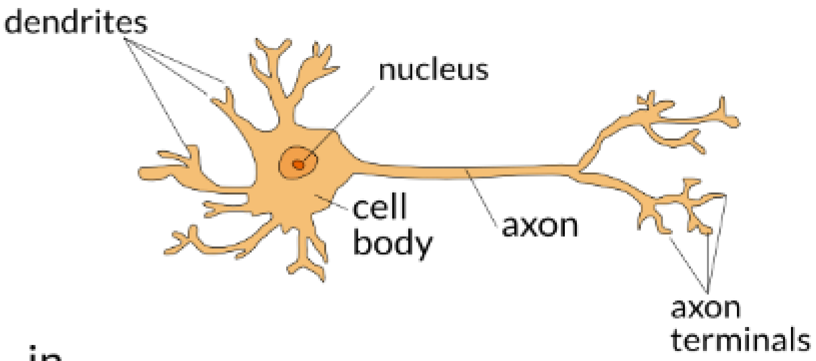
Two Types of AI

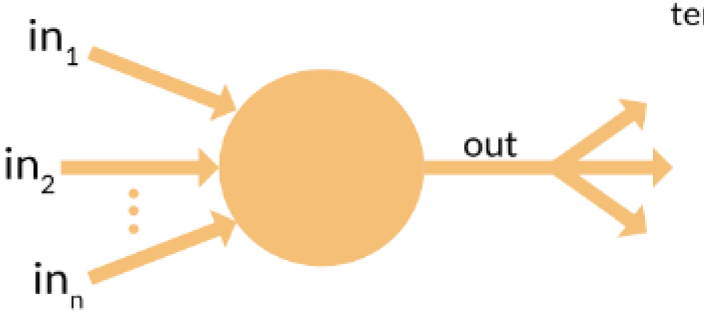
- **Good Old-Fashioned AI (GOFAI)**
  - Symbolic logic systems to represent basic elements of human thought like language, numbers, or goals
  - **Production systems** featuring if-then logic
    - General Problem Solver created by Newell and Simon in 1959
  - **Cognitive modeling architectures**
    - Systems like Soar and ACT-R with a heavy emphasis on how humans accomplish goals
  - Much of focus is not to create learning but to capture human-like intelligence related to how humans carry out decisions and actions



Two Types of AI

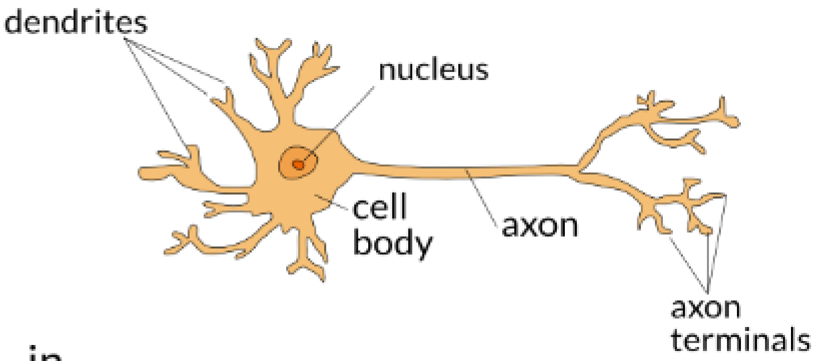
- **Neural Networks**
  - **Perceptron** developed in 1958 as approximation of single-cell neuron
  - By 1960s, mathematical algorithms like backpropagation developed to allow perceptrons to learn through training
    - **Machine learning**
  - Multiple perceptrons chained together to create neural networks
    - More layers of neural networks chained to together to create **deep learning**
    - Facilitated by greater availability of parallel computing (e.g., graphical processing units)

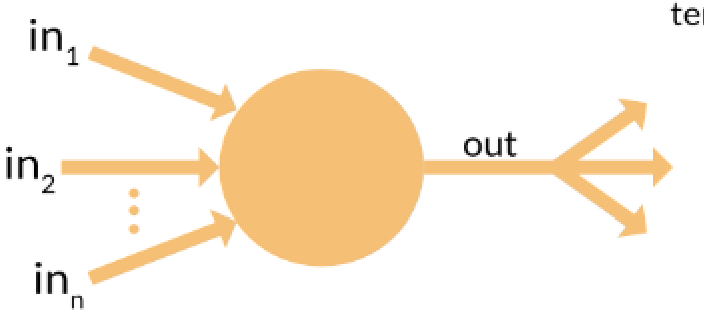




Two Types of AI

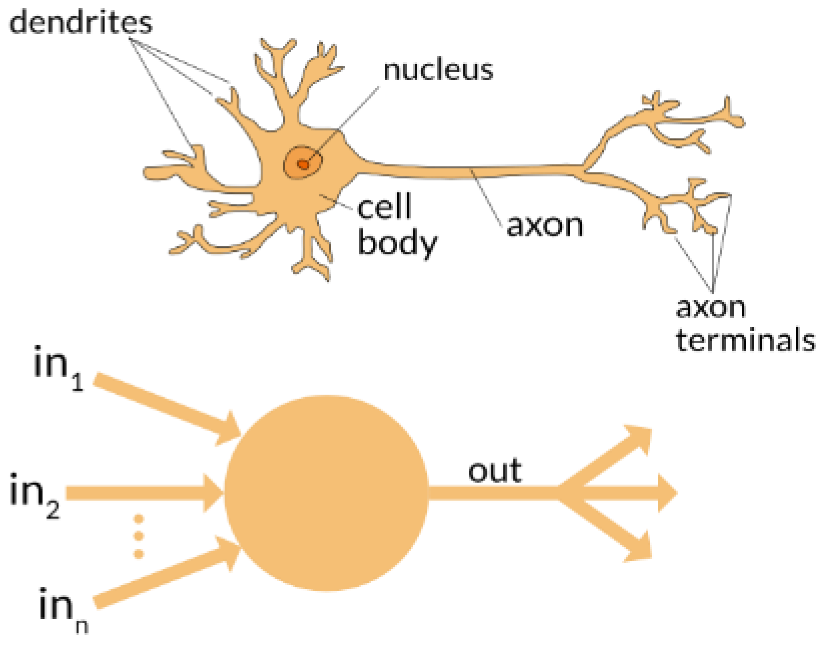
- **Neural Networks**
  - **Perceptron** developed in 1958 as approximation of single-cell neuron
  - By 1960s, mathematical algorithms like backpropagation developed to allow perceptrons to learn through training
    - **Machine learning**
  - Multiple perceptrons chained together to create neural networks
    - More layers of neural networks chained to together to create **deep learning**
    - Facilitated by greater availability of parallel computing (e.g., graphical processing units)





Two Types of AI

- **Neural Networks**
  - **Perceptron** developed in 1958 as approximation of single-cell neuron
  - By 1960s, mathematical algorithms like backpropagation developed to allow perceptrons to learn through training
    - **Machine learning**
  - Multiple perceptrons chained together to create neural networks
    - More layers of neural networks chained to together to create **deep learning**
    - Facilitated by greater availability of parallel computing (e.g., graphical processing units)



Two Types of AI

- **Different Uses**
  - GOFAI is good at following rules and making decisions
  - Neural networks are good at pattern recognition when trained

- **Self-Driving Vehicle Example**
  - Use GOFAI for the rules of the road
    - Procedural knowledge
    - Control automation
  - Neural networks used to recognize the world
    - The eyes on the road
    - Information automation

Two Types of AI

- **Different Uses**
  - GOFAI is good at following rules and making decisions
  - Neural networks are good at pattern recognition when trained

- **Self-Driving Vehicle Example**
  - Use GOFAI for the rules of the road
    - Procedural knowledge
    - Control automation
  - Neural networks used to recognize the world
    - The eyes on the road
    - Information automation

Very Briefly Noted

Some Key Applications of AI in Nuclear Industry

Key Applications of AI in Nuclear Industry

Automation

- Control automation: Using AI to control a system (or a plant, such as might be the case in a microreactor)
- Information automation: Using AI to intelligently gather information that operator needs
  - Detection of problems such as early warning systems and condition monitoring

Prediction

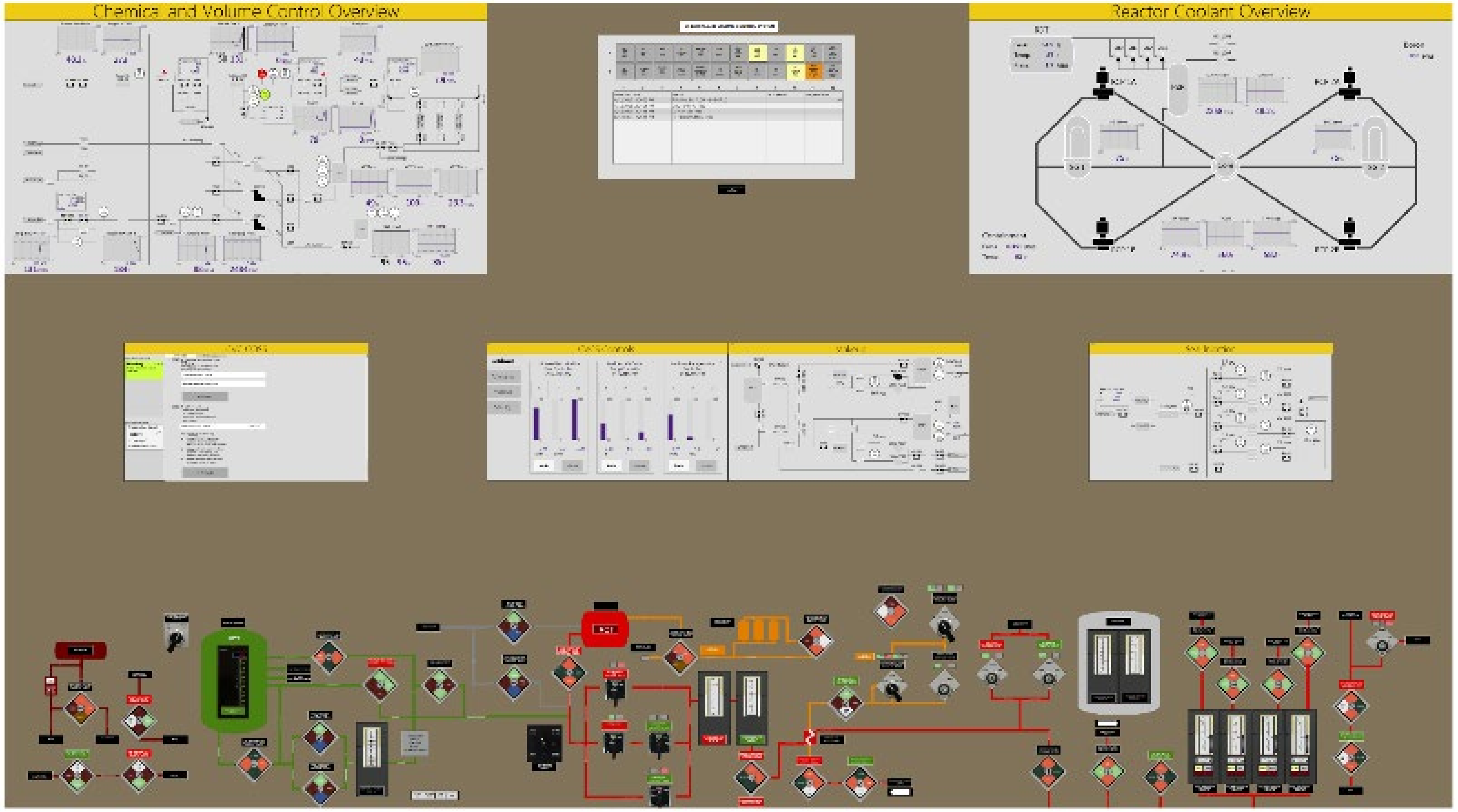
- Predictive—instead of *prescriptive*—maintenance systems

Human-System Interface

- Smart notification systems like alarm filtering
- Natural language processing for hands-free interactivity

Example Possible Automation in Nuclear Power

Information Automation *(Top)*, Control Automation *(Middle)*, and Analog Control *(Bottom)*



Computerized Operator Support System (INL)

Key Applications of AI in Nuclear Industry

Automation

- Control automation: Using AI to control a system (or a plant, such as might be the case in a microreactor)
- Information automation: Using AI to intelligently gather information that operator needs
  - Detection of problems such as early warning systems and condition monitoring

Prediction

- Predictive—instead of *preventative*—maintenance systems

Human-System Interface

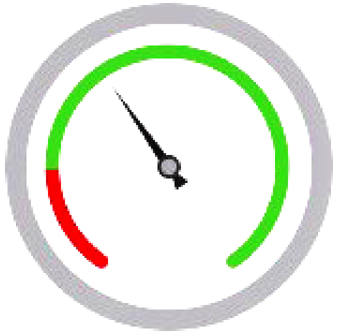
- Smart notification systems like alarm filtering
- Natural language processing for hands-free interactivity

Predictive Maintenance

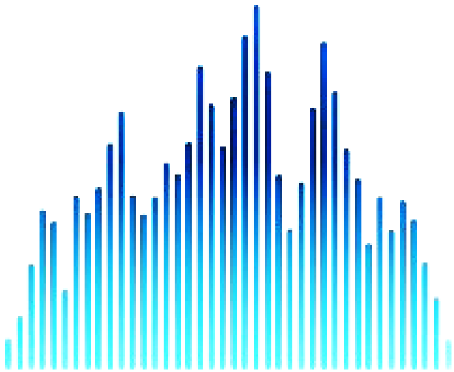
- Look for signs of performance degradation through sensor data
  - Catch parts that are failing sooner than anticipated
  - Leave perfectly good parts in operation
- Anomaly detection using machine learning
- Convey information to human

VIBRATIONS

TEMPERATURE



PRESSURE



NOISE

Key Applications of AI in Nuclear Industry

Automation

- Control automation: Using AI to control a system (or a plant, such as might be the case in a microreactor)
- Information automation: Using AI to intelligently gather information that operator needs
  - Detection of problems such as early warning systems and condition monitoring

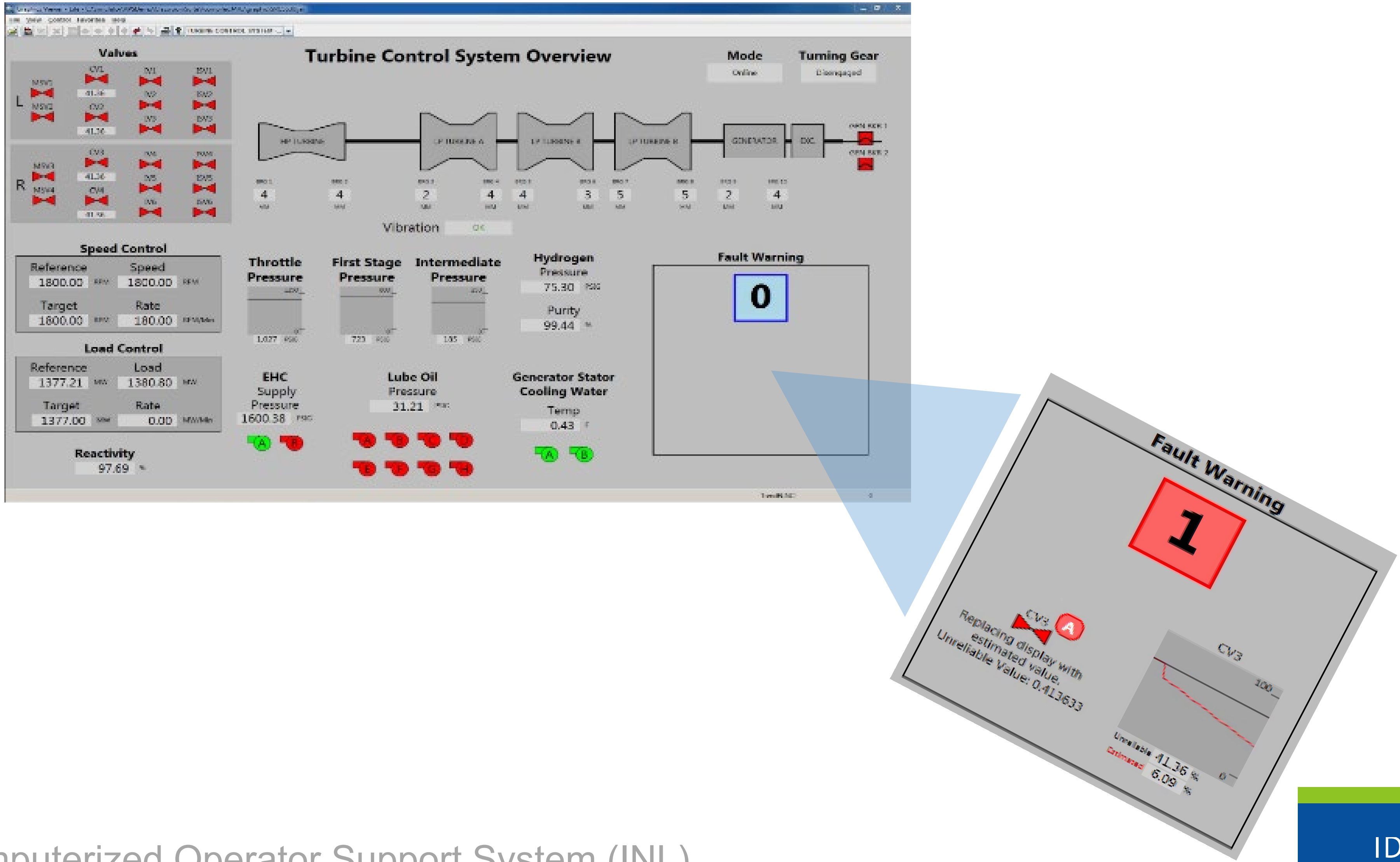
Prediction

- Predictive—instead of *preventative*—maintenance systems

Human-System Interface

- Smart notification systems like alarm filtering
- Natural language processing for hands-free interactivity

Example Smart Notification System



Computerized Operator Support System (INL)

Who Knows What the Future Will Bring, But AI Will Be Part of It!

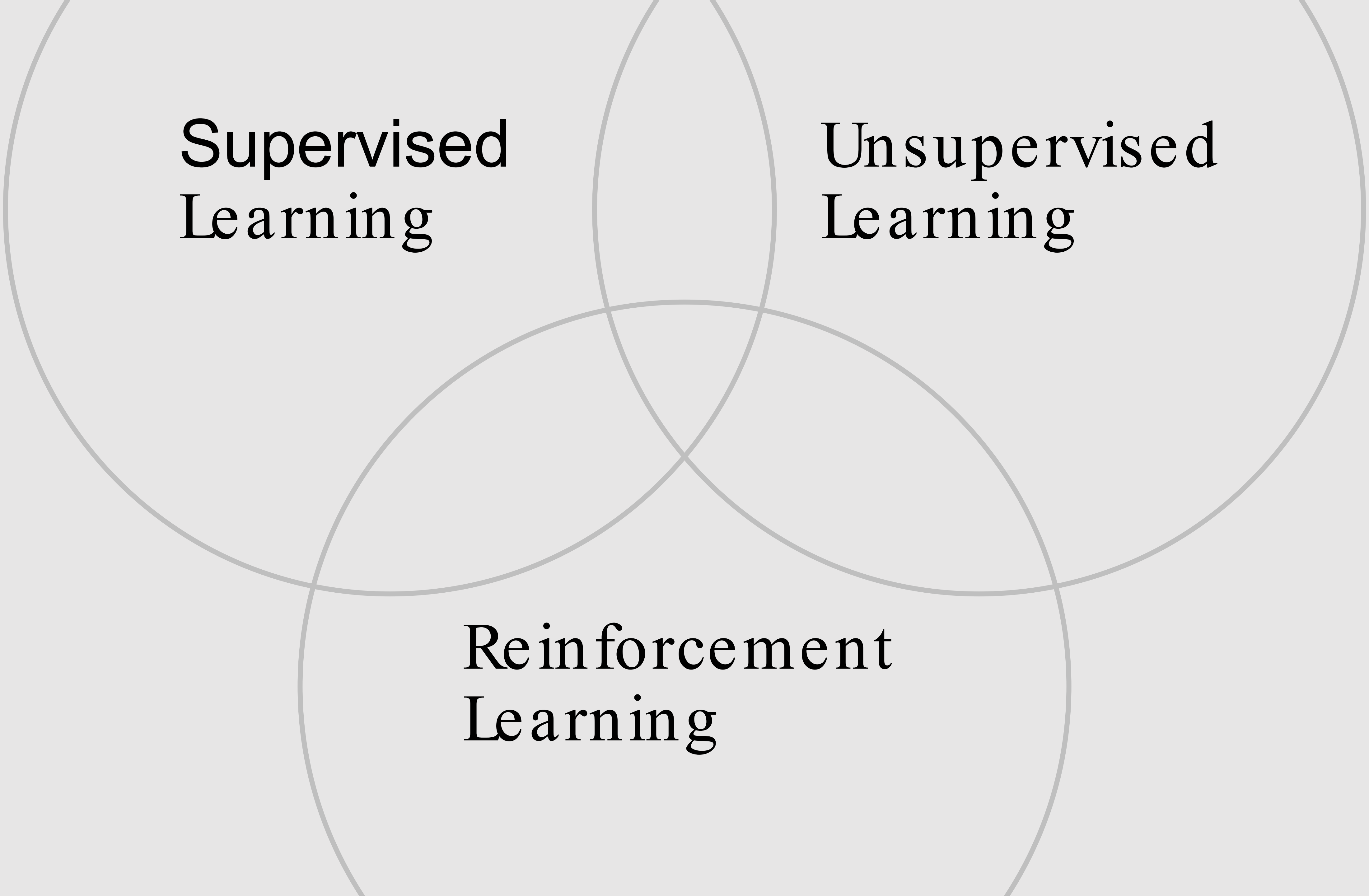

IDAHO NATIONAL
LABORATORY
70th
Anniversary
ronald.boring@inl.gov

Introduction to Machine Learning

Dr. Mark Fuge
Univ. of Maryland, College Park
(301) 405-2558
fuge@umd.edu
ideal.umd.edu

What I hope you get from today:

1. What is Machine Learning?
2. When is it helpful?
3. When is it not helpful?
4. Where do you go from here?


Supervised Learning
Unsupervised Learning
Reinforcement Learning

Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

Anomaly Detection

System identification

Optimal Control

Optimization

Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

Anomaly Detection

System identification

Optimal Control

Optimization

1. What is the problem that needs solving?
2. How can machine learning help?
3. How do we know it is working?
4. When does it break down?

Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

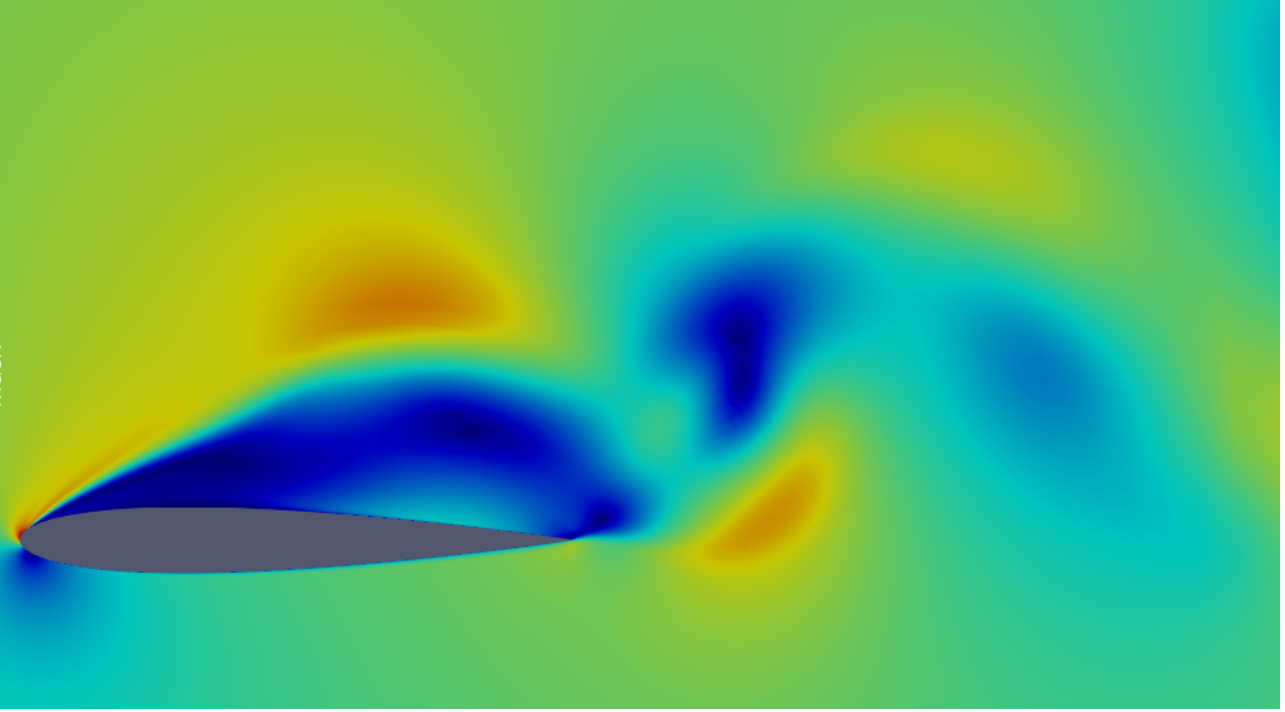
Generative Design

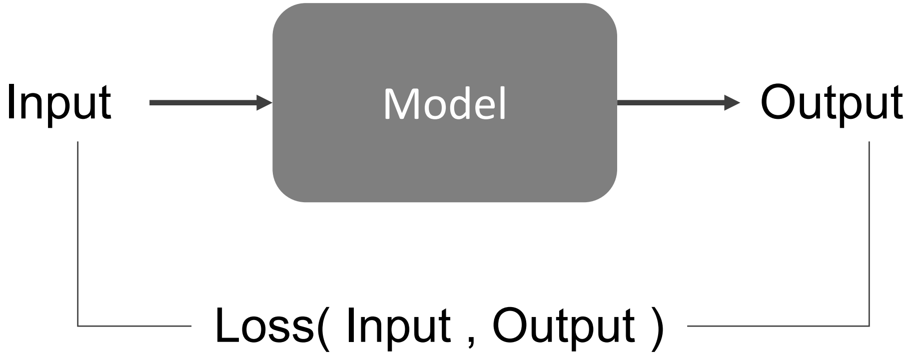
Anomaly Detection

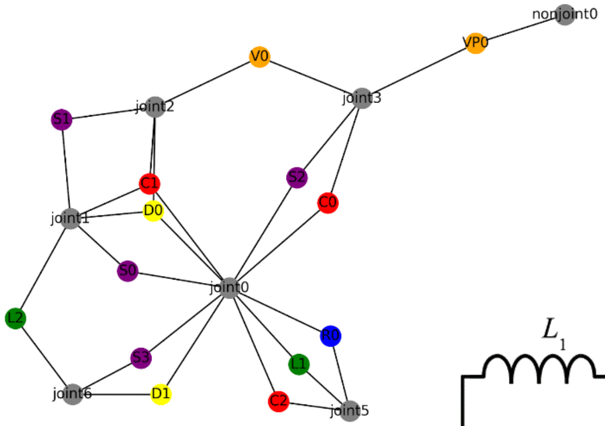
System identification

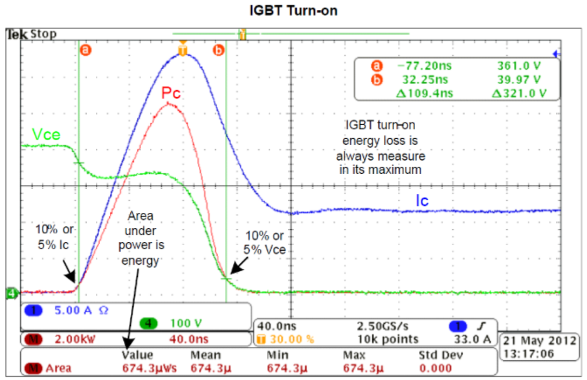
Optimal Control

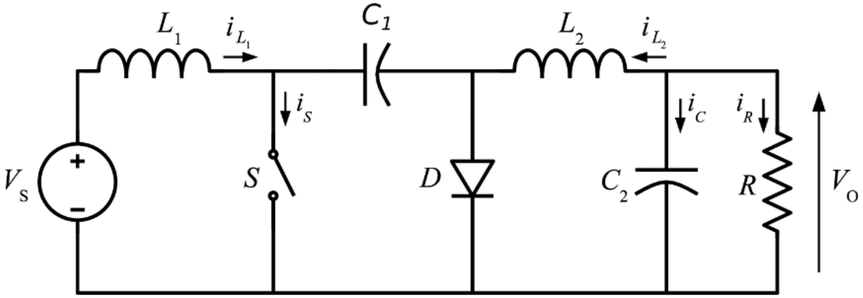
Optimization

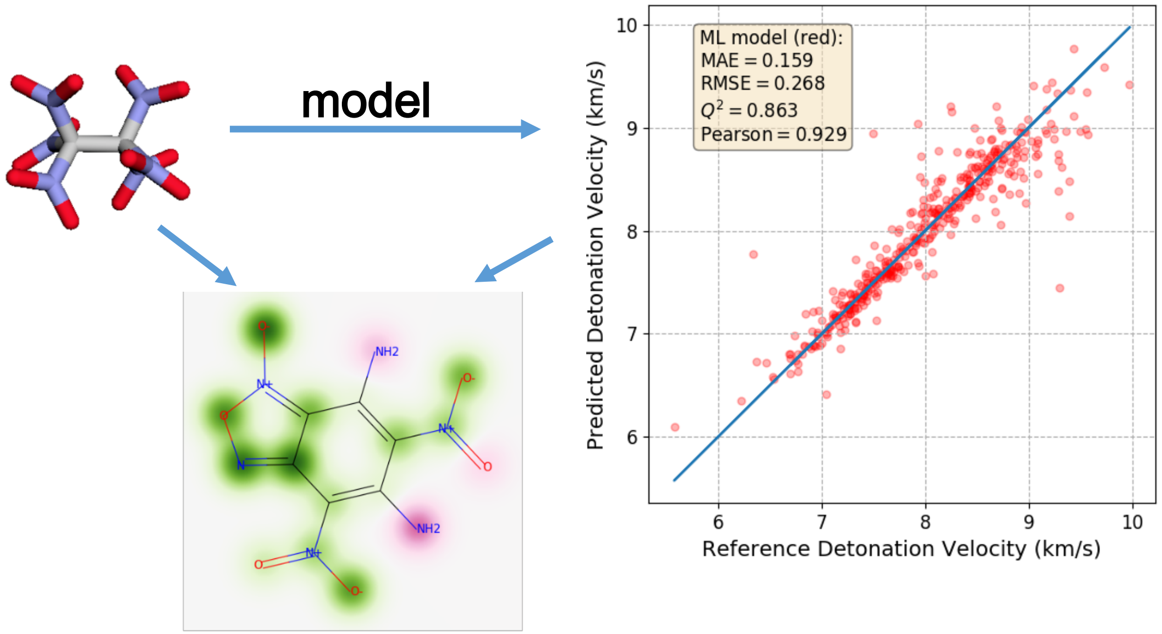



Input
Model
Output
Loss(Input , Output)


nonjoint0
VP0
V0
joint3
joint2
S1
C1
D0
S2
C0
joint1
S0
joint0
L2
B0
S3
L1
joint6
D1
C2
joint5


IGBT Turn-on
Vce
Pc
Ic
IGBT turn-on energy loss is always measure in its maximum
Area under power is energy
10% or 5% Ic
10% or 5% Vce
5.00 A Ω
100 V
2.00kW
40.0ns
2.50GS/s
10k points
33.0 A
21 May 2012
13:17:06
Area
Value 674.3μWs
Mean 674.3μ
Min 674.3μ
Max 674.3μ
Std Dev 0.000


L_1
i_{L_1}
C_1
L_2
i_{L_2}
i_S
i_C
i_R
V_S
S
D
C_2
R
V_O


model
NH2
NH2
ML model (red):
MAE = 0.159
RMSE = 0.268
Q^2 = 0.863
Pearson = 0.929
Predicted Detonation Velocity (km/s)
Reference Detonation Velocity (km/s)

Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

Anomaly Detection

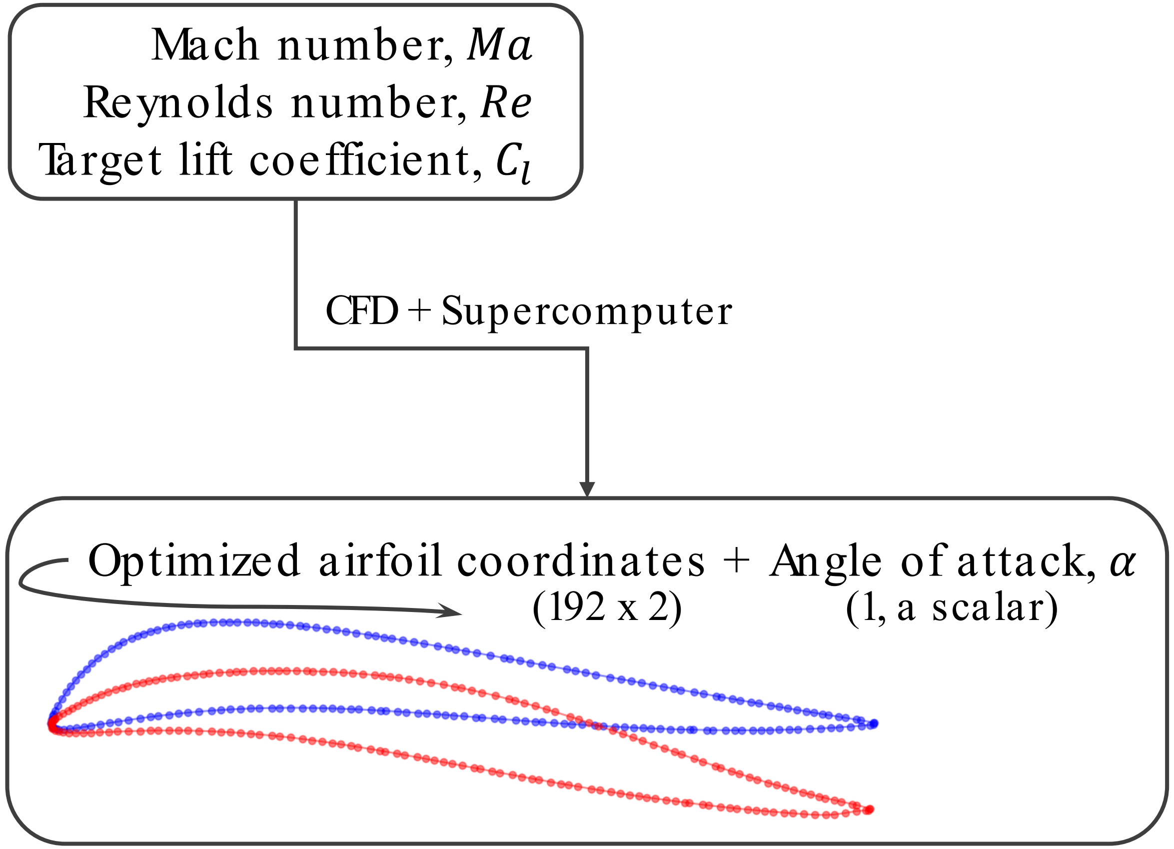
System identification

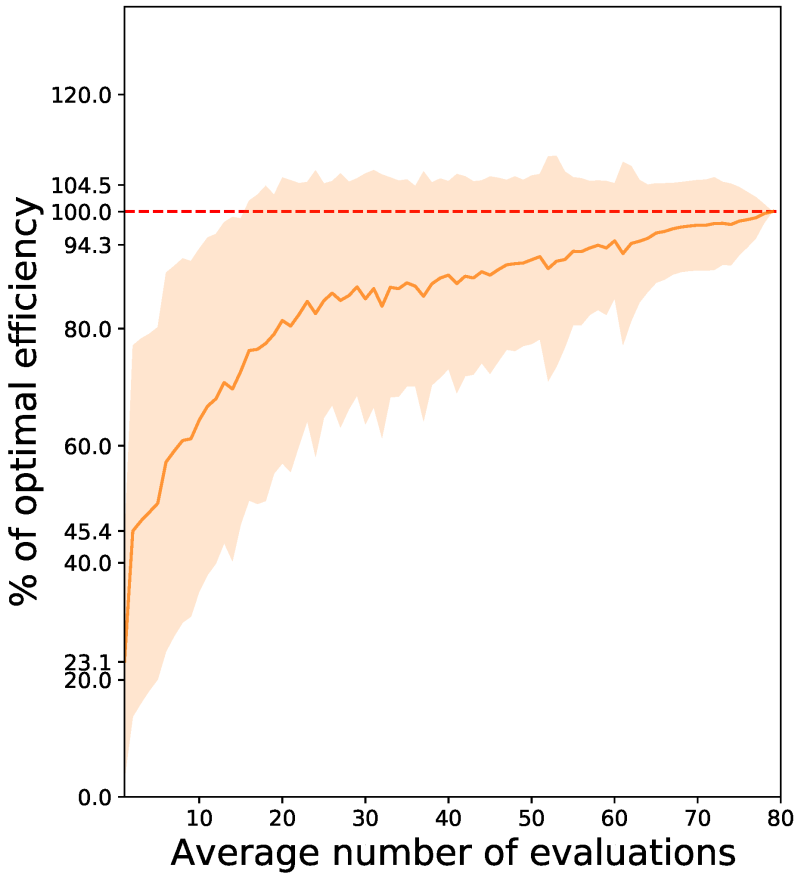
Optimal Control

Optimization

Example: ARPA-E DIFFERENTIATE Program

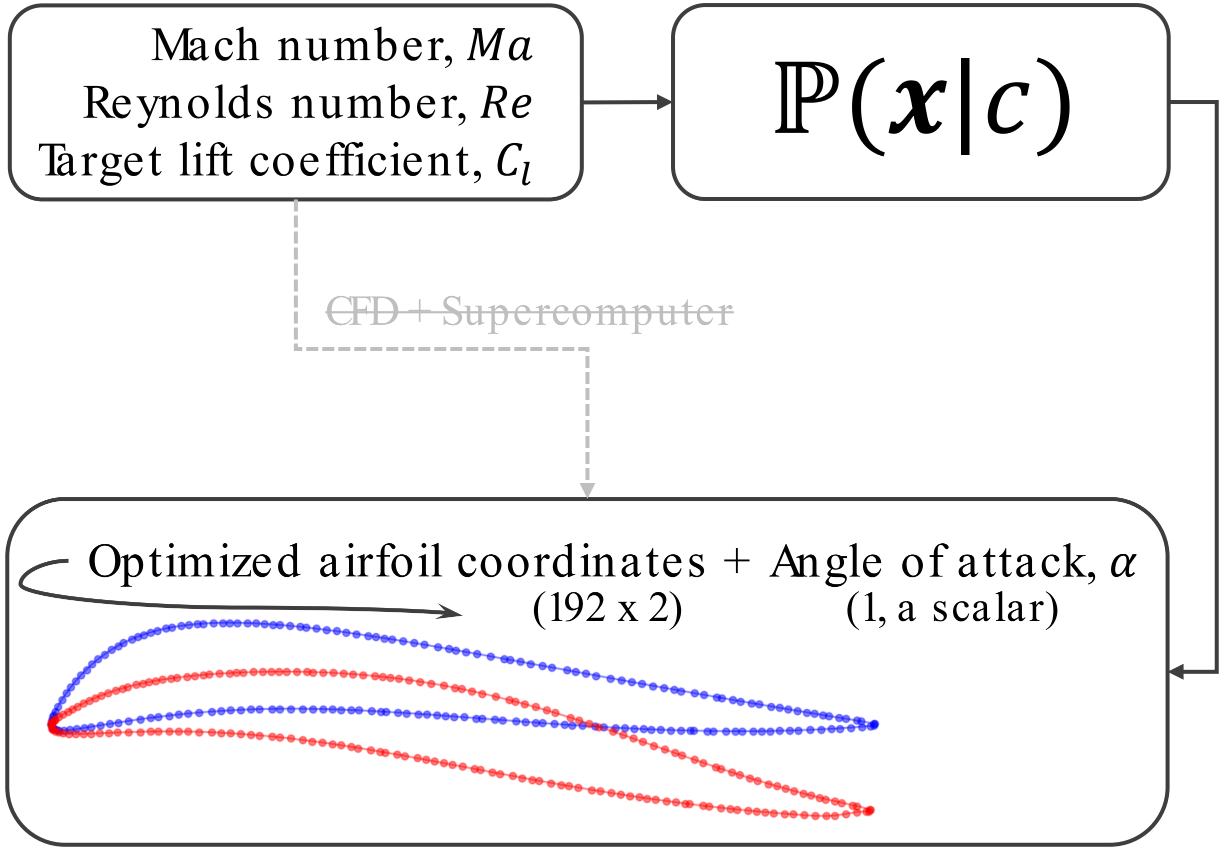
Inverse Design of Aero & Heat Transfer Surfaces

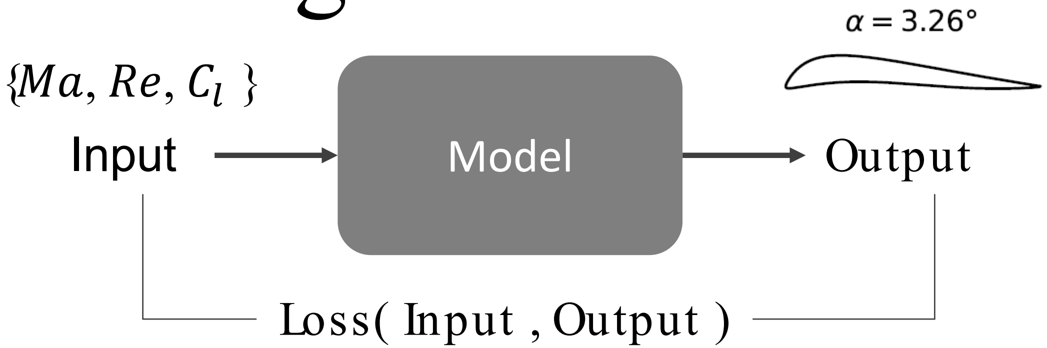


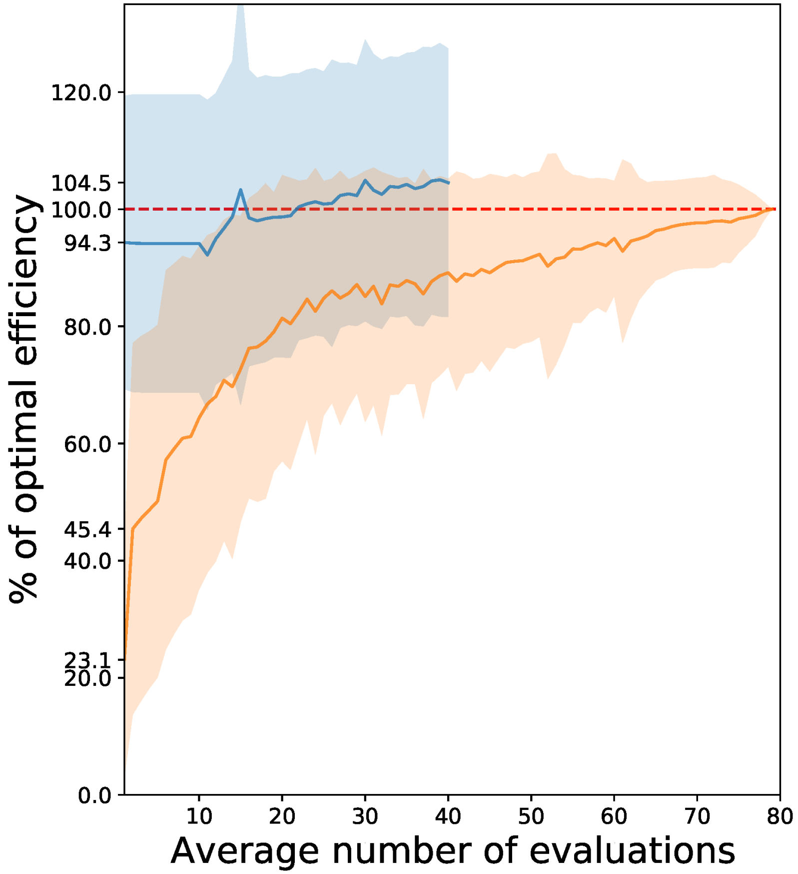


Example: ARPA-E DIFFERENTIATE Program

Inverse Design of Aero & Heat Transfer Surfaces







Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

Anomaly Detection

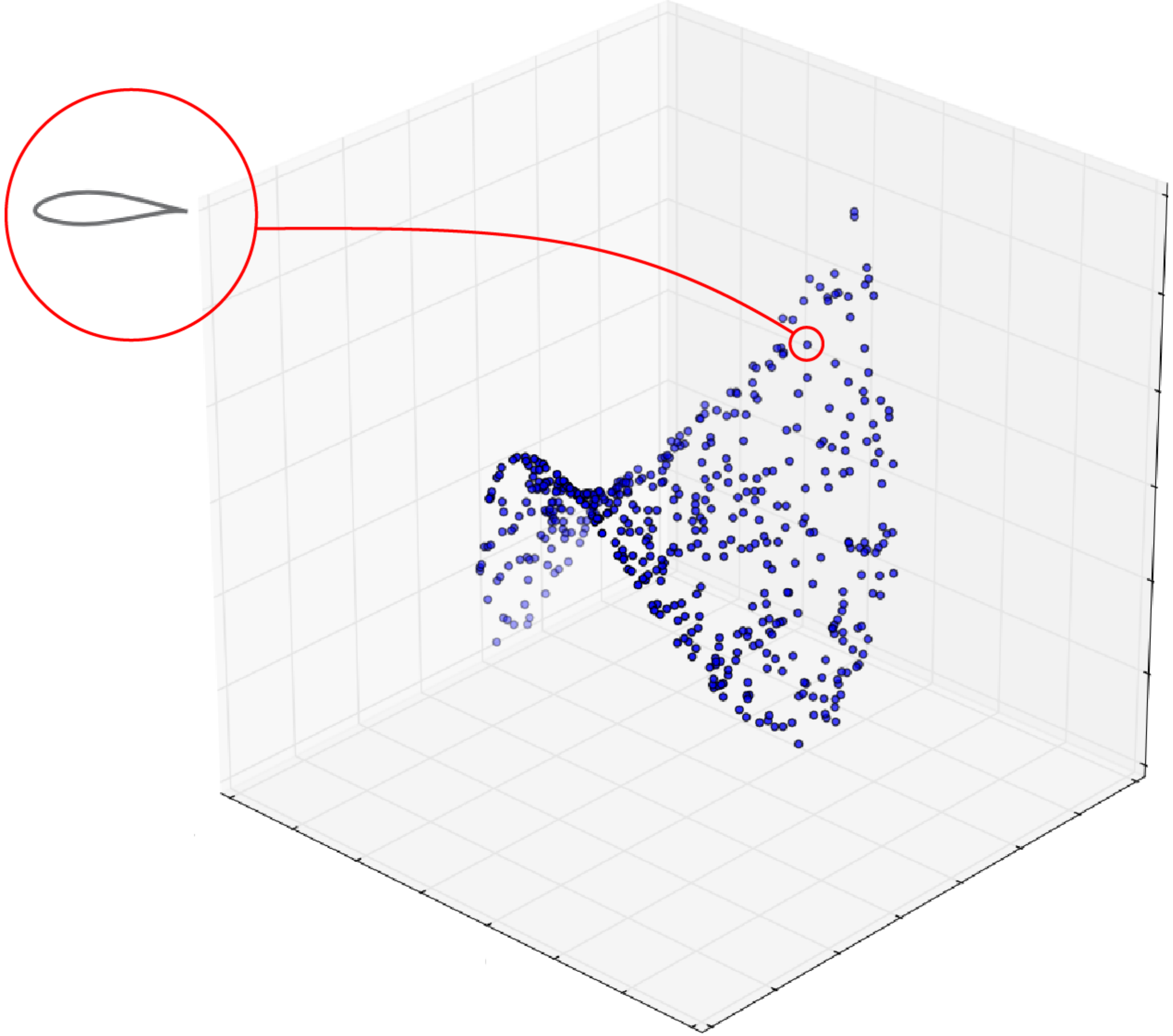
System identification

Optimal Control

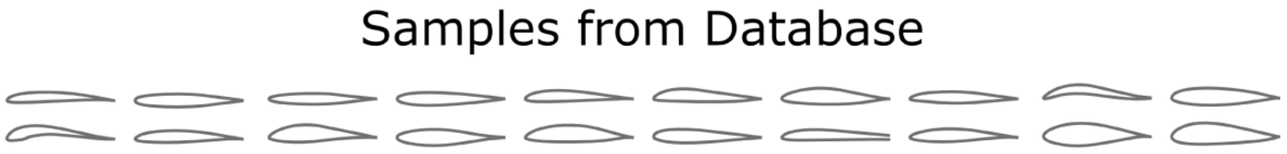
Optimization

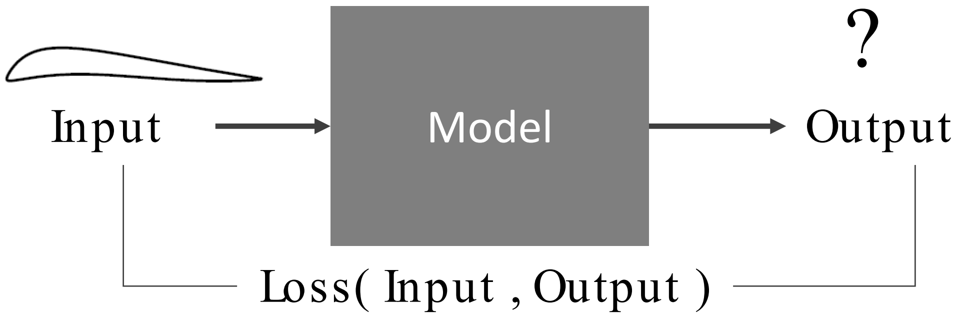
Problem: Original airfoil representation (~100 coordinates) is too large to be useful.

The Manifold Hypothesis

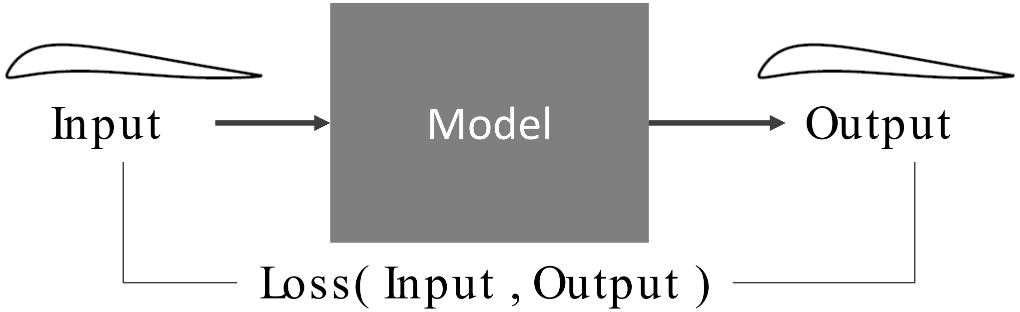


Example: Learning Airfoil Manifolds

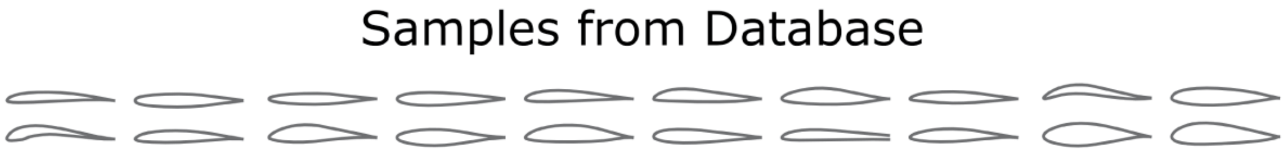


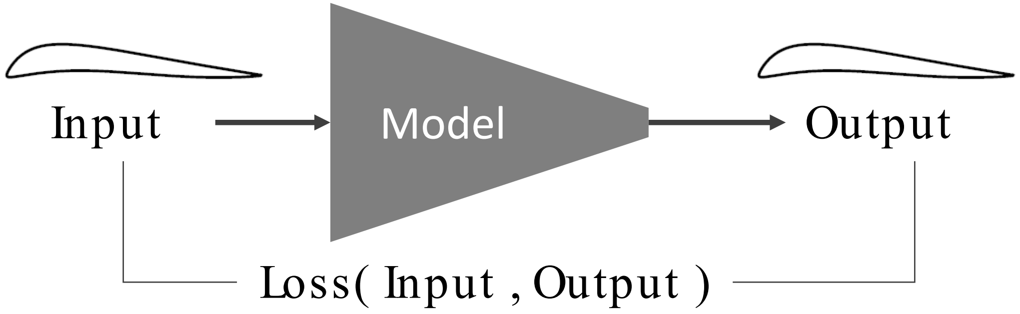


Example: Learning Airfoil Manifolds



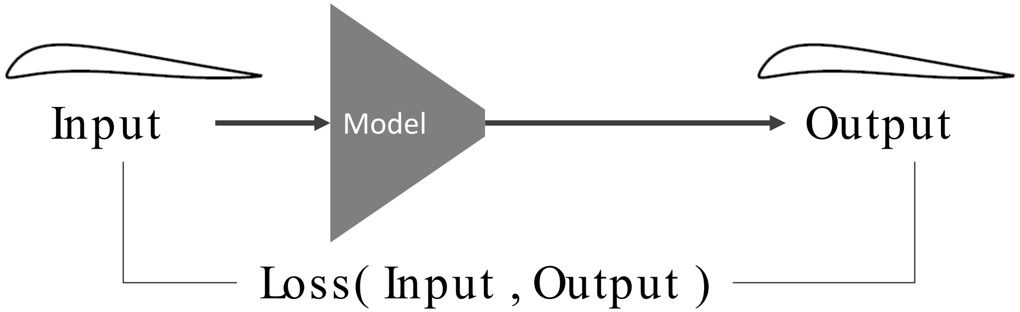
Example: Learning Airfoil Manifolds



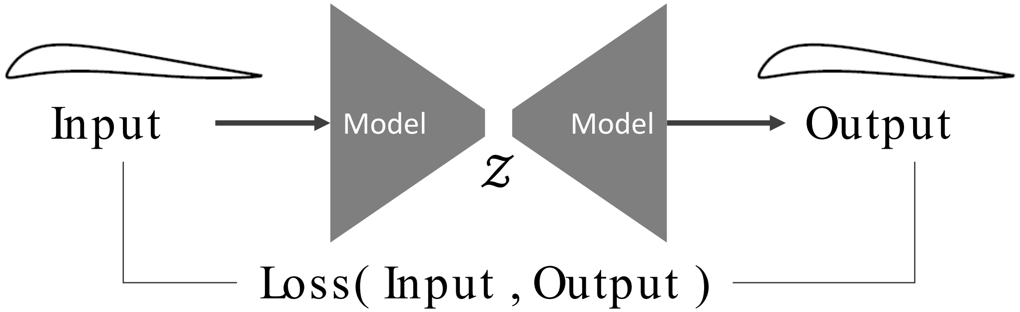


Example: Learning Airfoil Manifolds

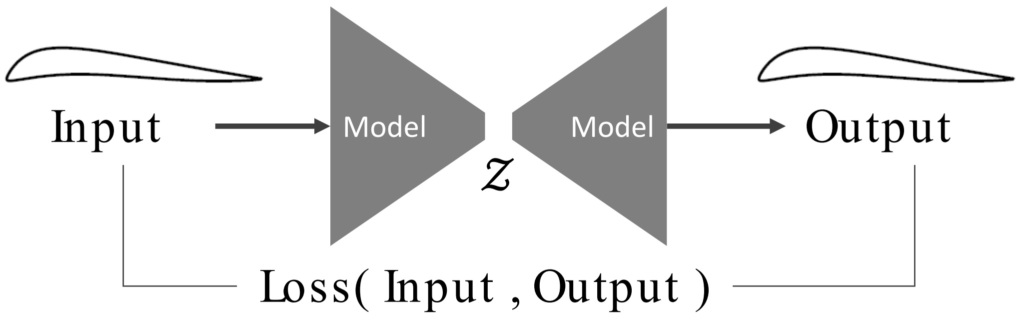
Samples from Database

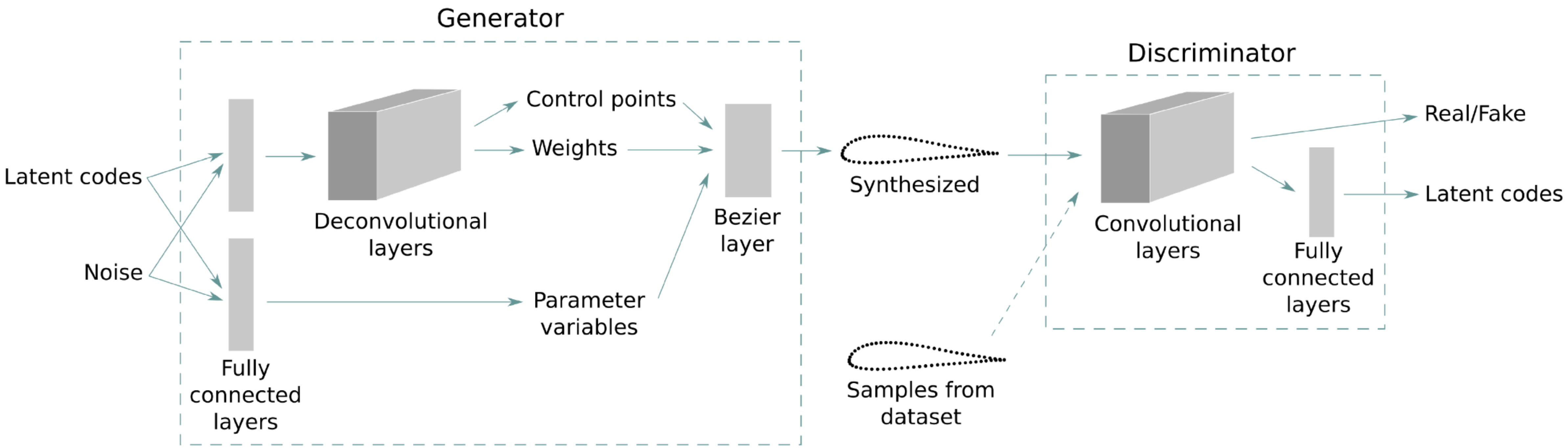


Example: Learning Airfoil Manifolds

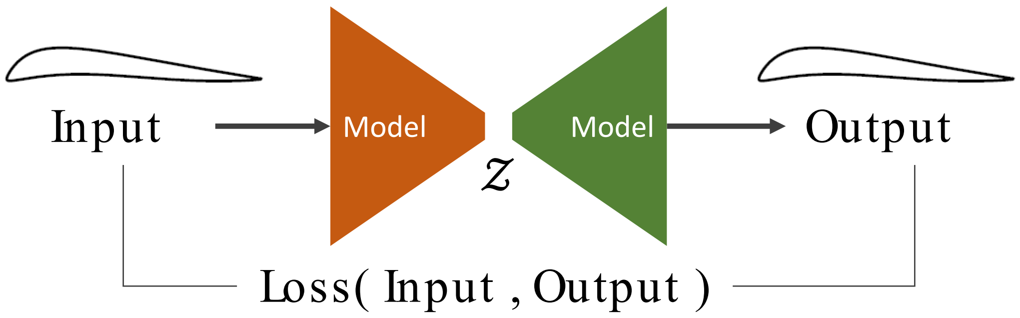


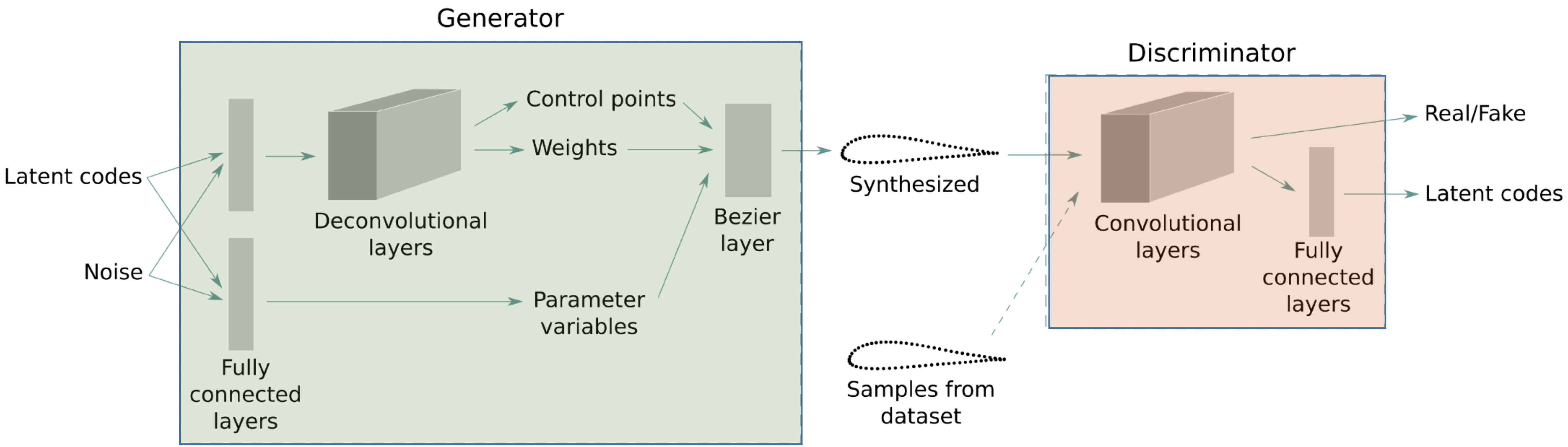
Example: Learning Airfoil Manifolds



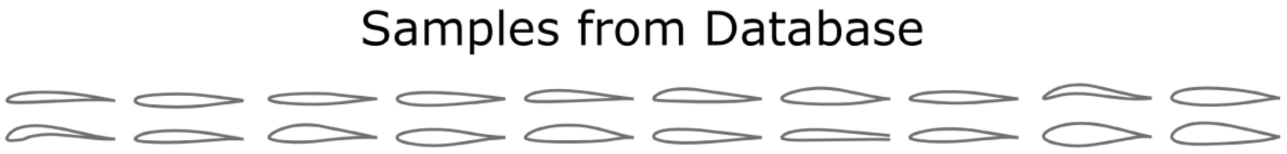


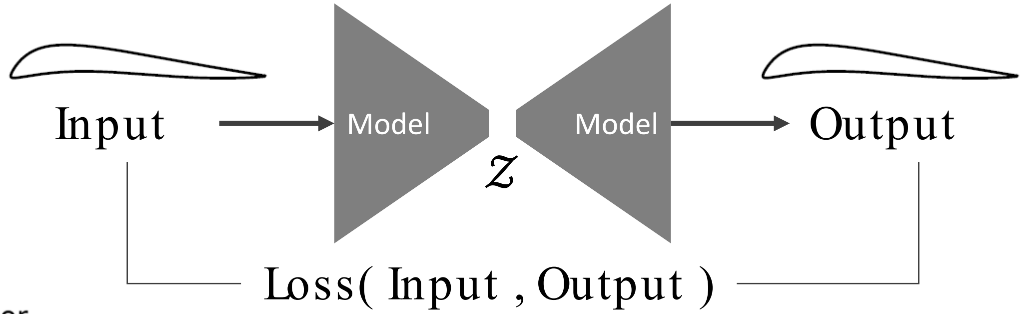
Example: Learning Airfoil Manifolds

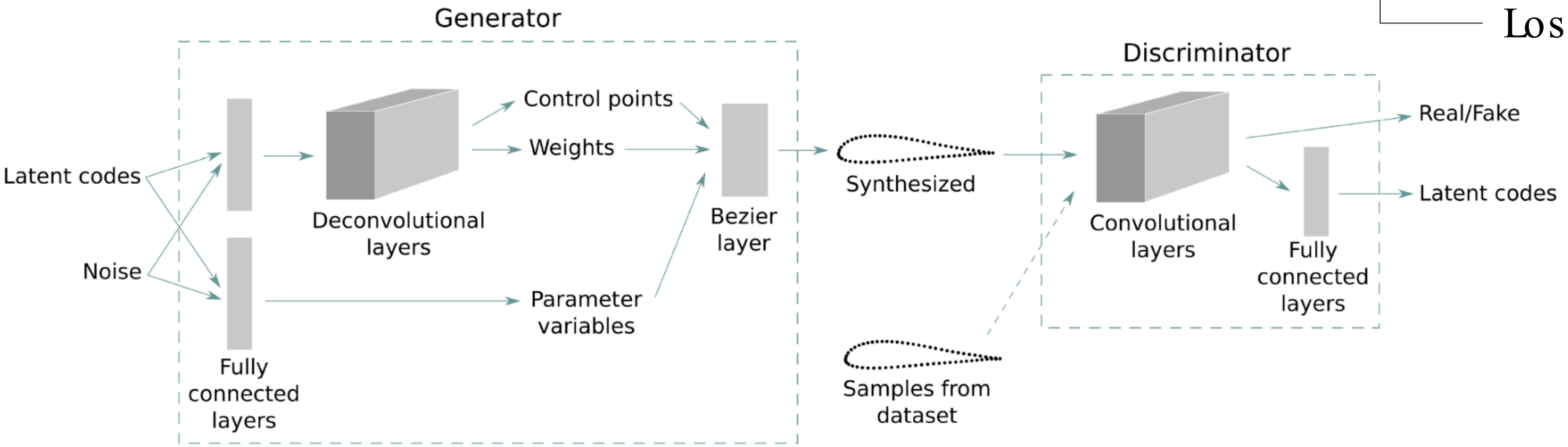


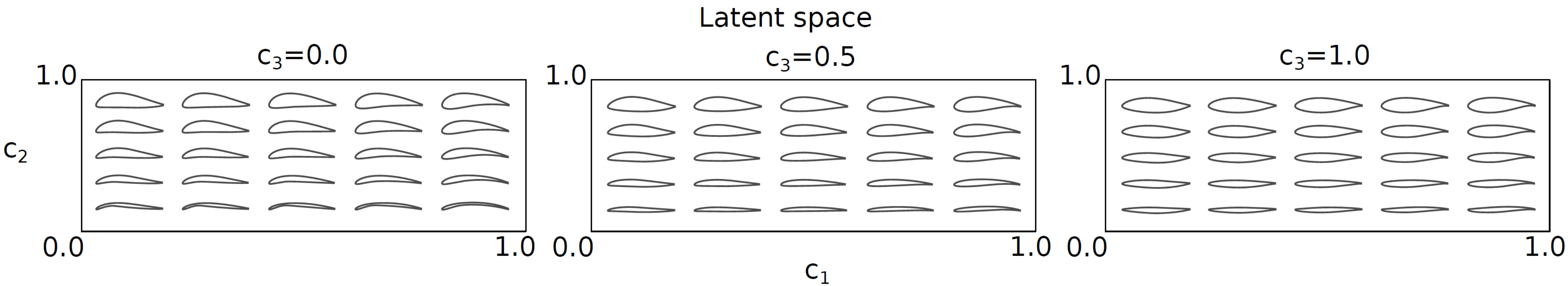


Example: Learning Airfoil Manifolds

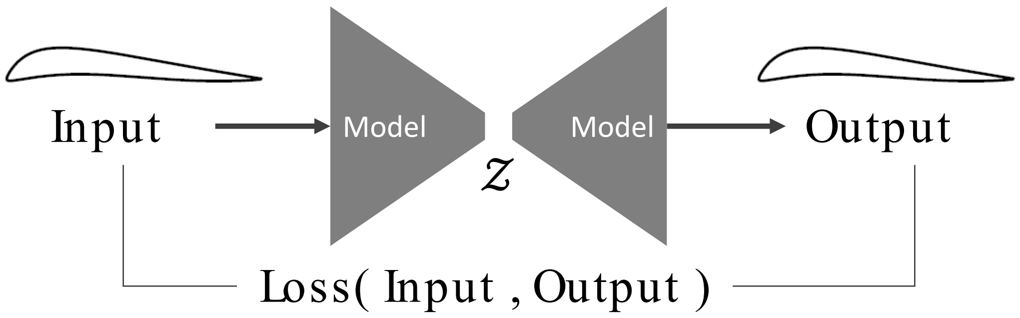


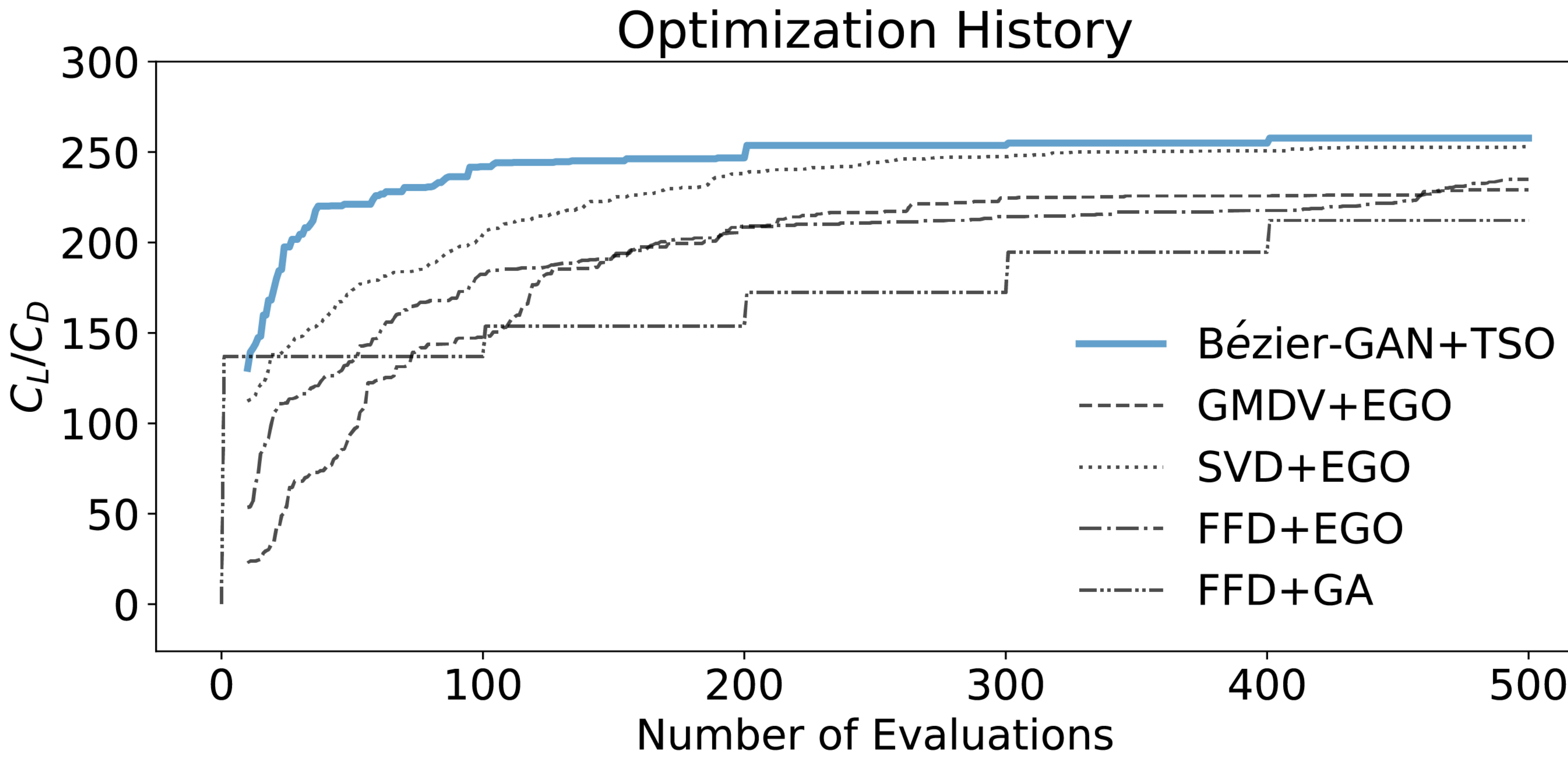






Example: Learning Airfoil Manifolds





Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

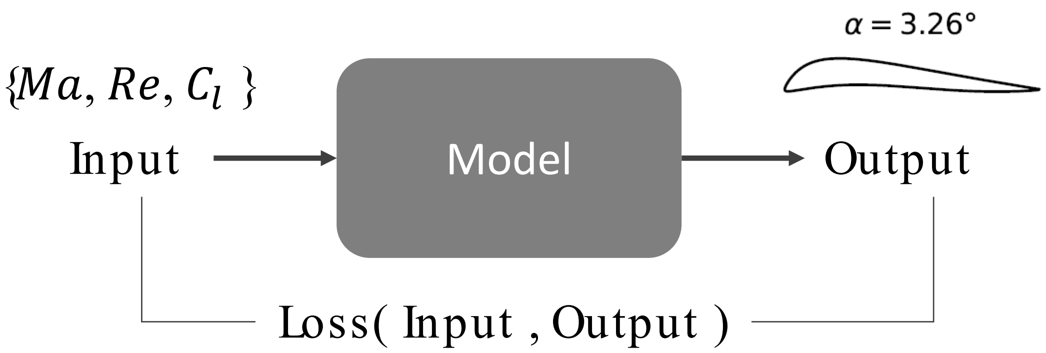
Anomaly Detection

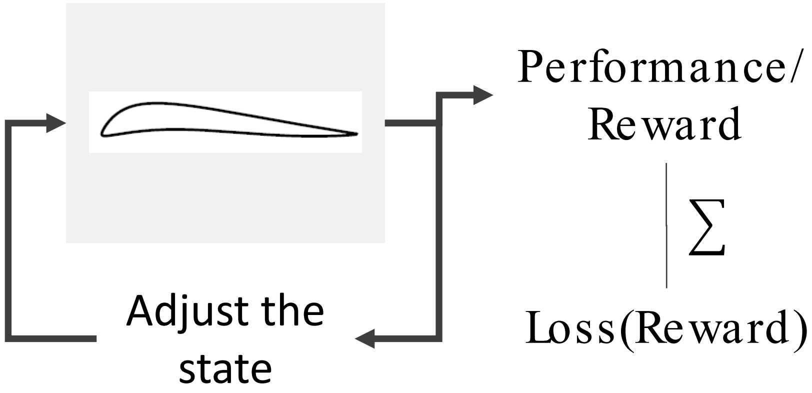
System identification

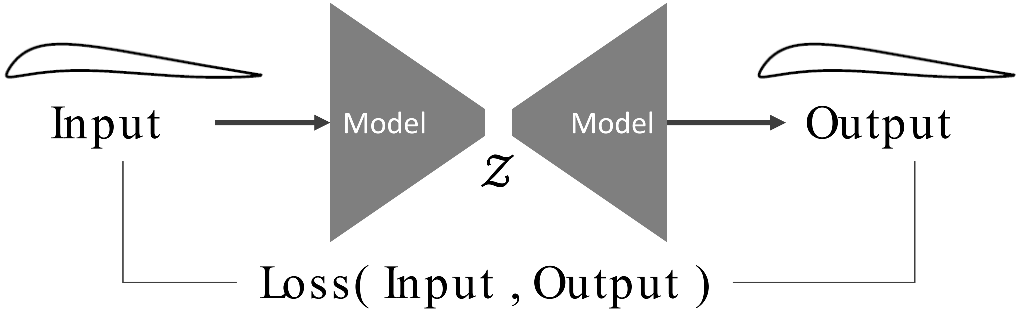
Optimal Control

Optimization

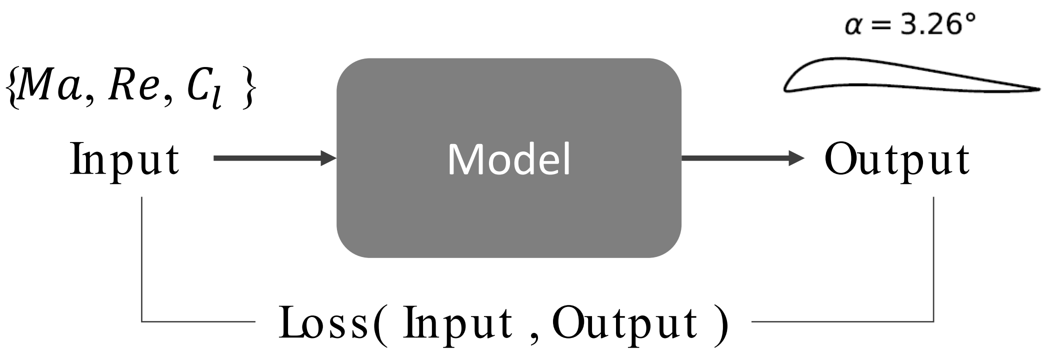
Reinforcement Learning vs Other models

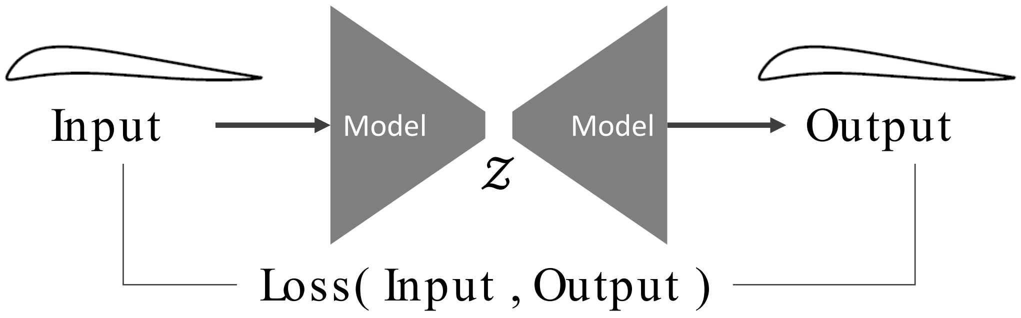


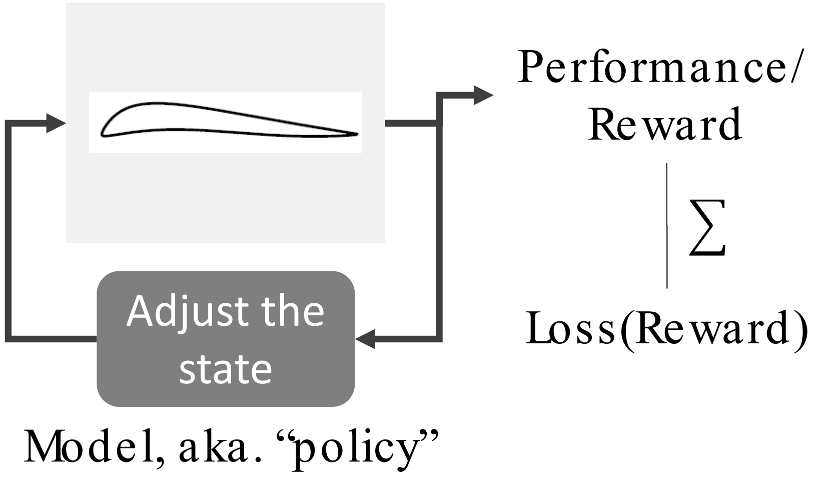




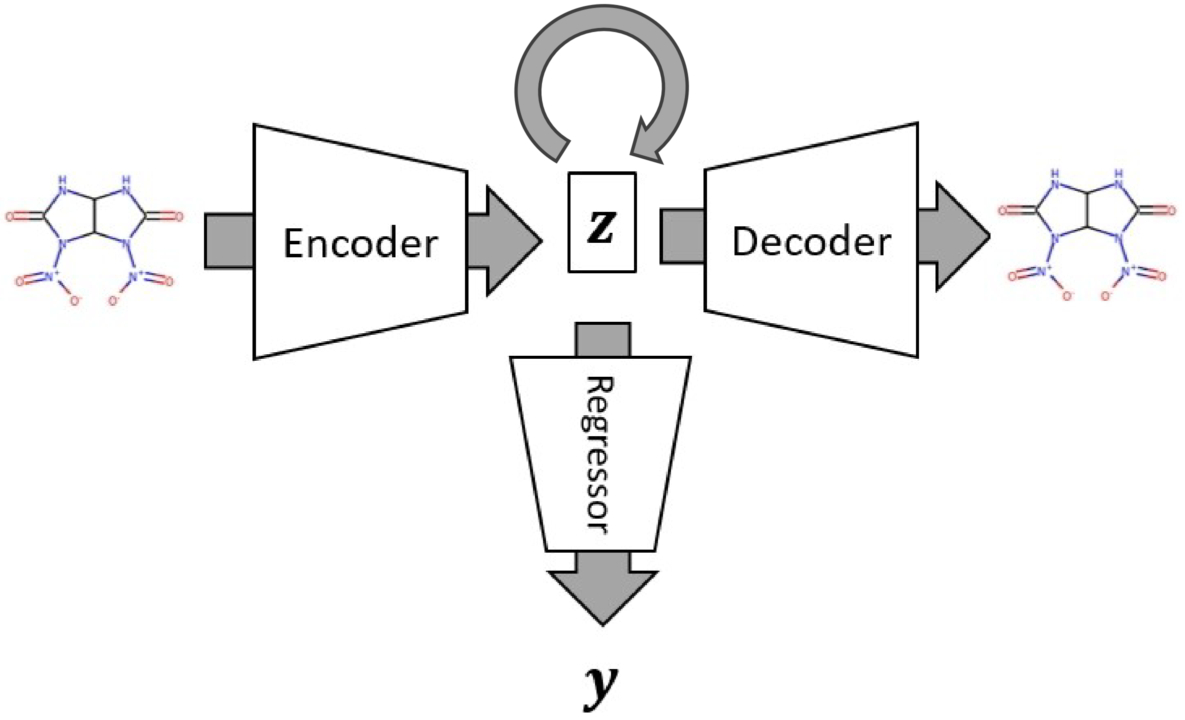
Reinforcement Learning vs Other models

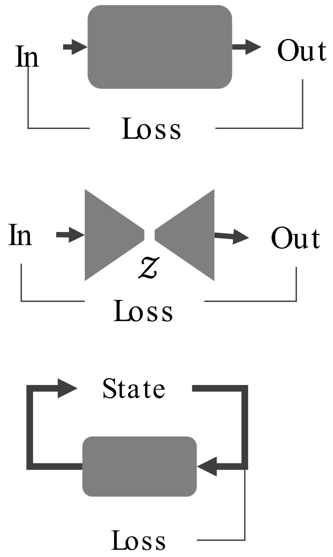




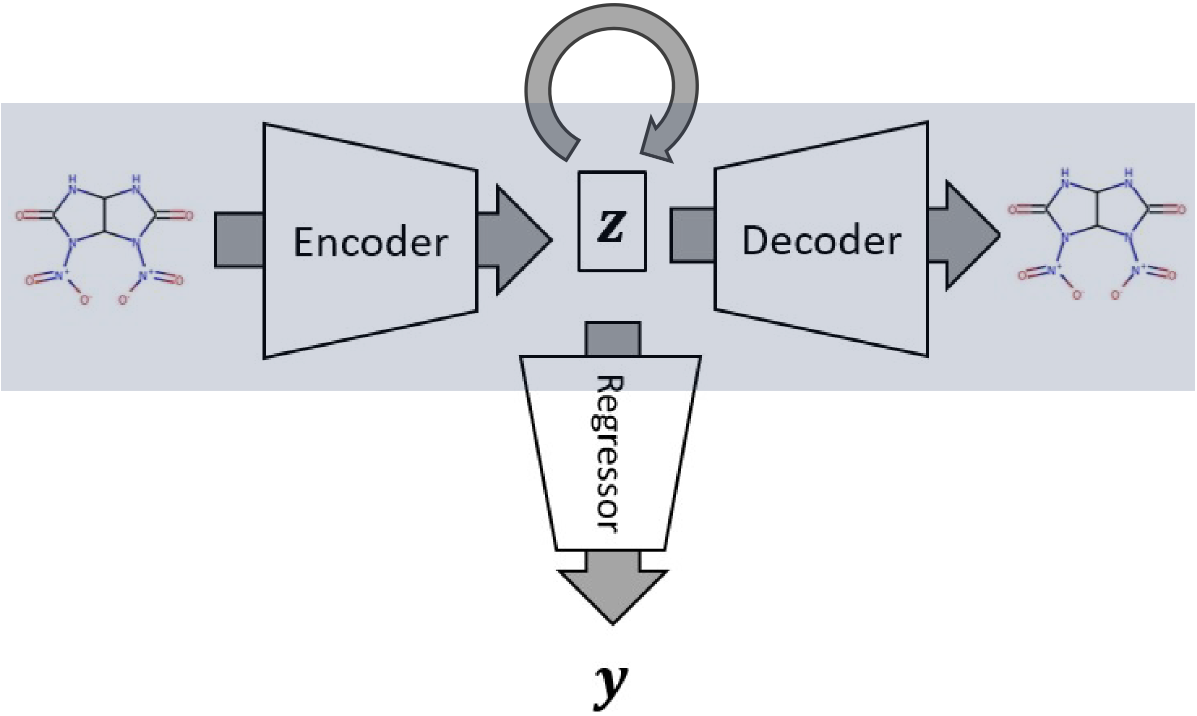


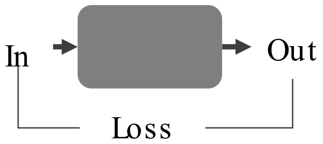
Example: Optimizing Molecular Properties

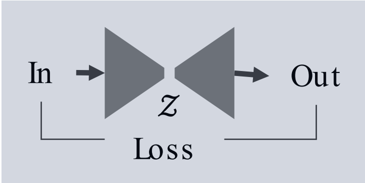


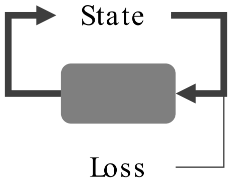


Example: Optimizing Molecular Properties

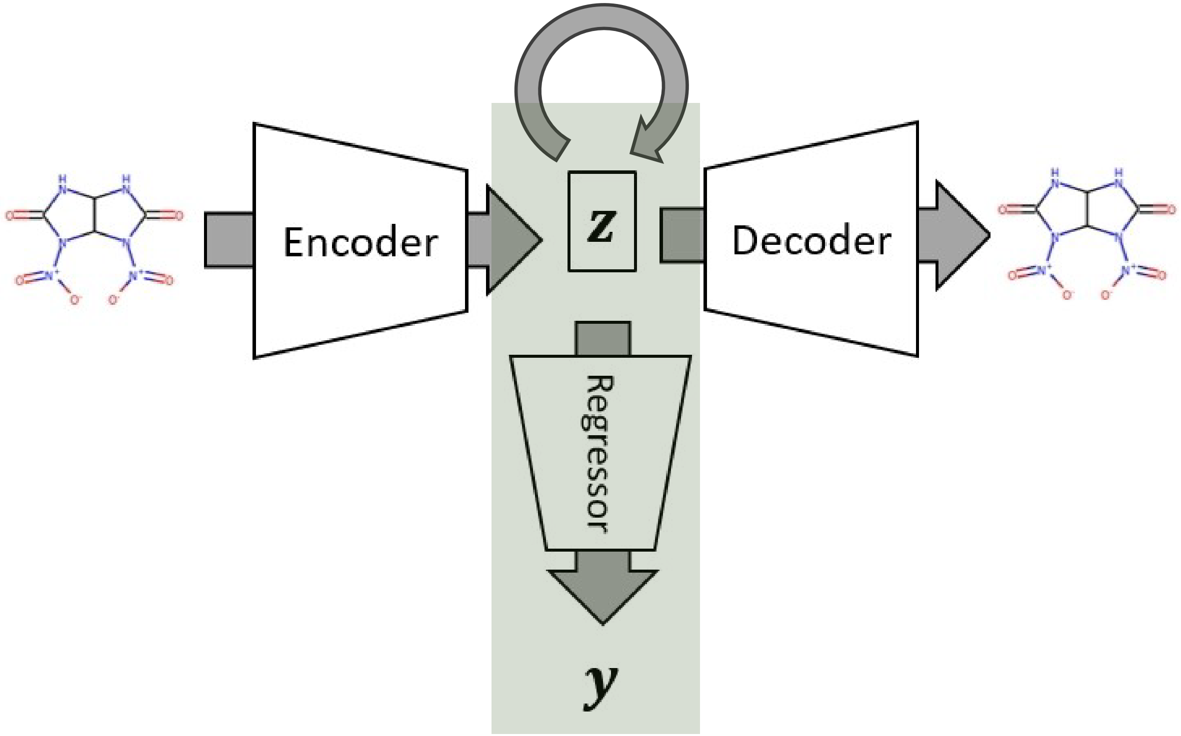


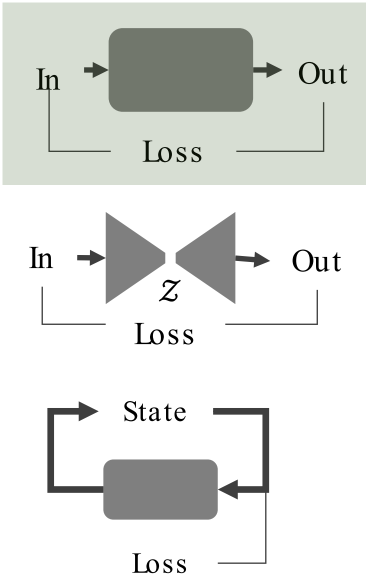




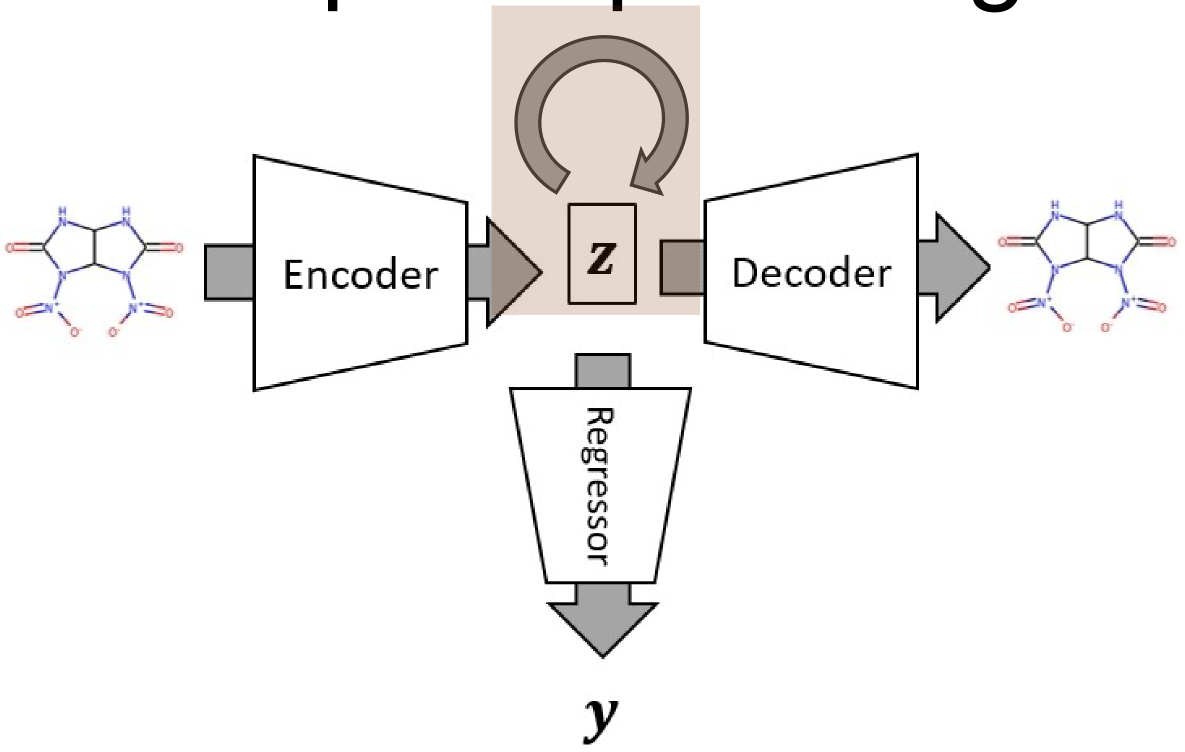


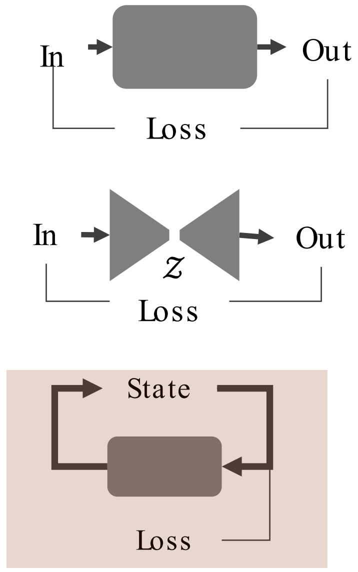
Example: Optimizing Molecular Properties



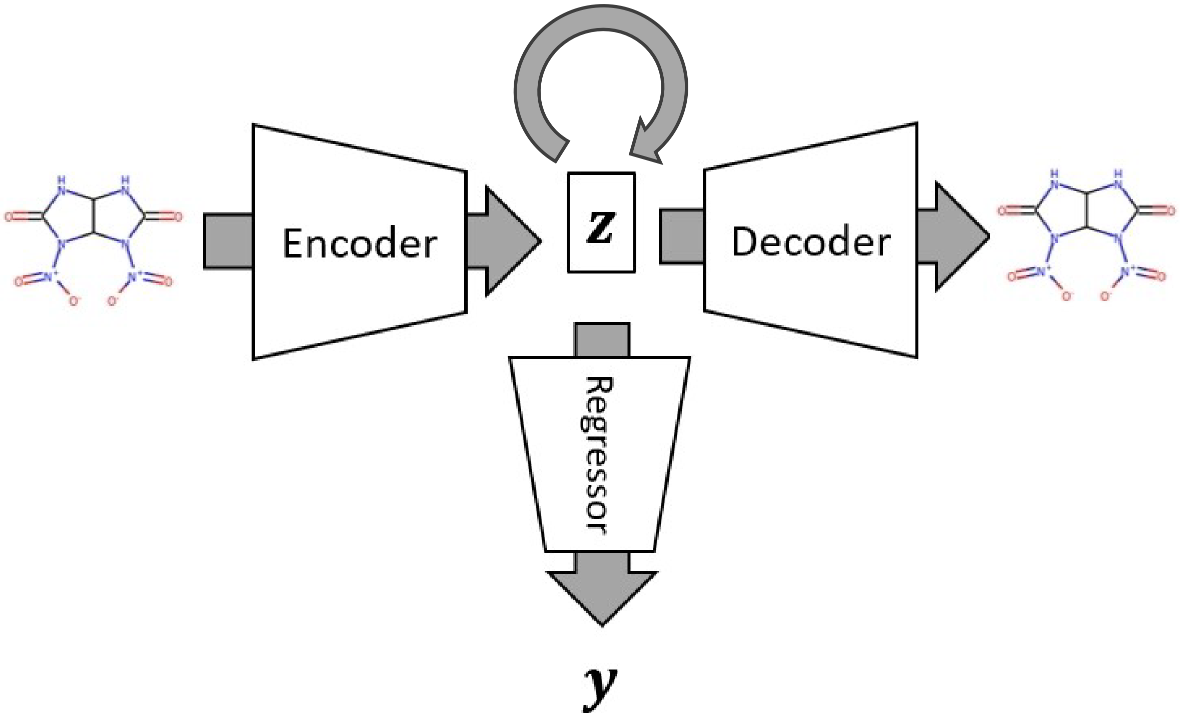


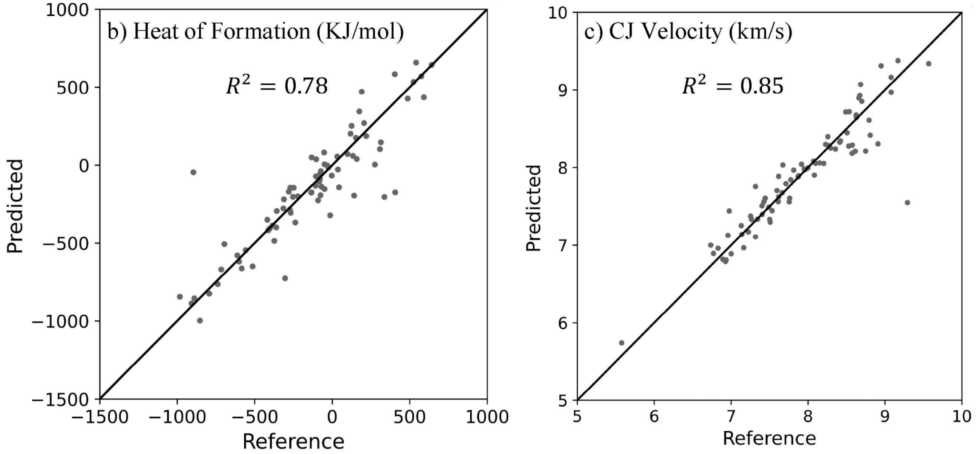
Example: Optimizing Molecular Properties

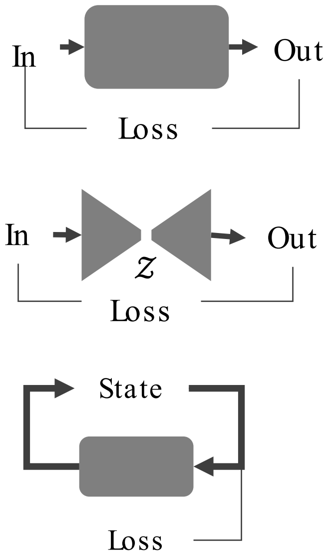


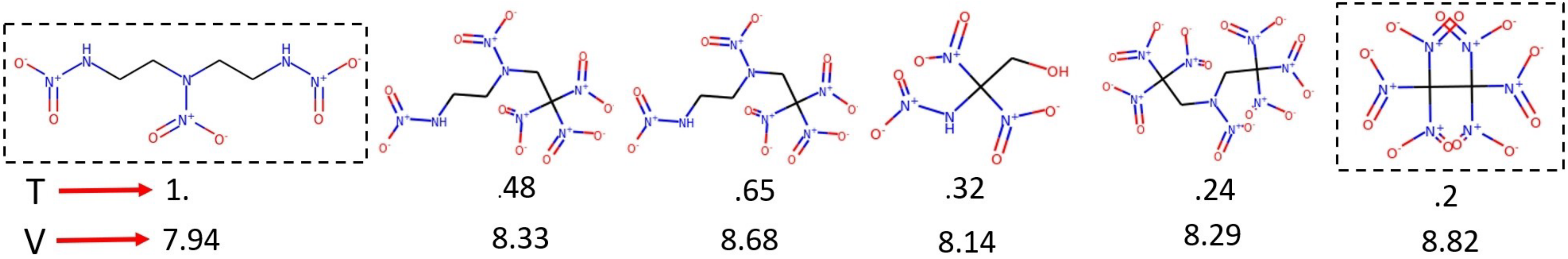


Example: Optimizing Molecular Properties









Types of ML

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Typical Engineering or Science Tasks

Reduced Order Models

Multi-Fidelity / Coarse-graining

Inverse Problems/ Design

Forecasting/ Prognostics

Generative Design

Anomaly Detection

System identification

Optimal Control

Optimization

Where do you go from here?

Technical Challenges

How do we create, collect, and share benchmark datasets?

How do we best combine existing Engineering knowledge with ML techniques?

How do we perform Verification and Validation?

What are appropriate Standards for such models?

What are the key Figures of Merit we should be optimizing in such systems?

Socio-Economic Challenges

How do we estimate the economic Return on Investment for ML techniques or datasets?

How do we protect IP or Privacy in trained models?

What regulatory frameworks do we need for verification of safety critical or other systems?

How should we train our workforce differently to leverage these techniques?

For more details see:

- JMD Editorial: ML in Engineering Design: http://ideal.umd.edu/papers/paper/ml-eng-design-jmd
- Summary of Data-Driven Design workshop: http://ideal.umd.edu/papers/paper/d3-implications

Where do you go from here?

What can you do?

Continue your education in these areas, or for those of your workforce.

Reach out to researchers and domain experts for new technical challenges we can resolve in these areas.

Provide guidance to policy and regulatory bodies on how these techniques might be managed.

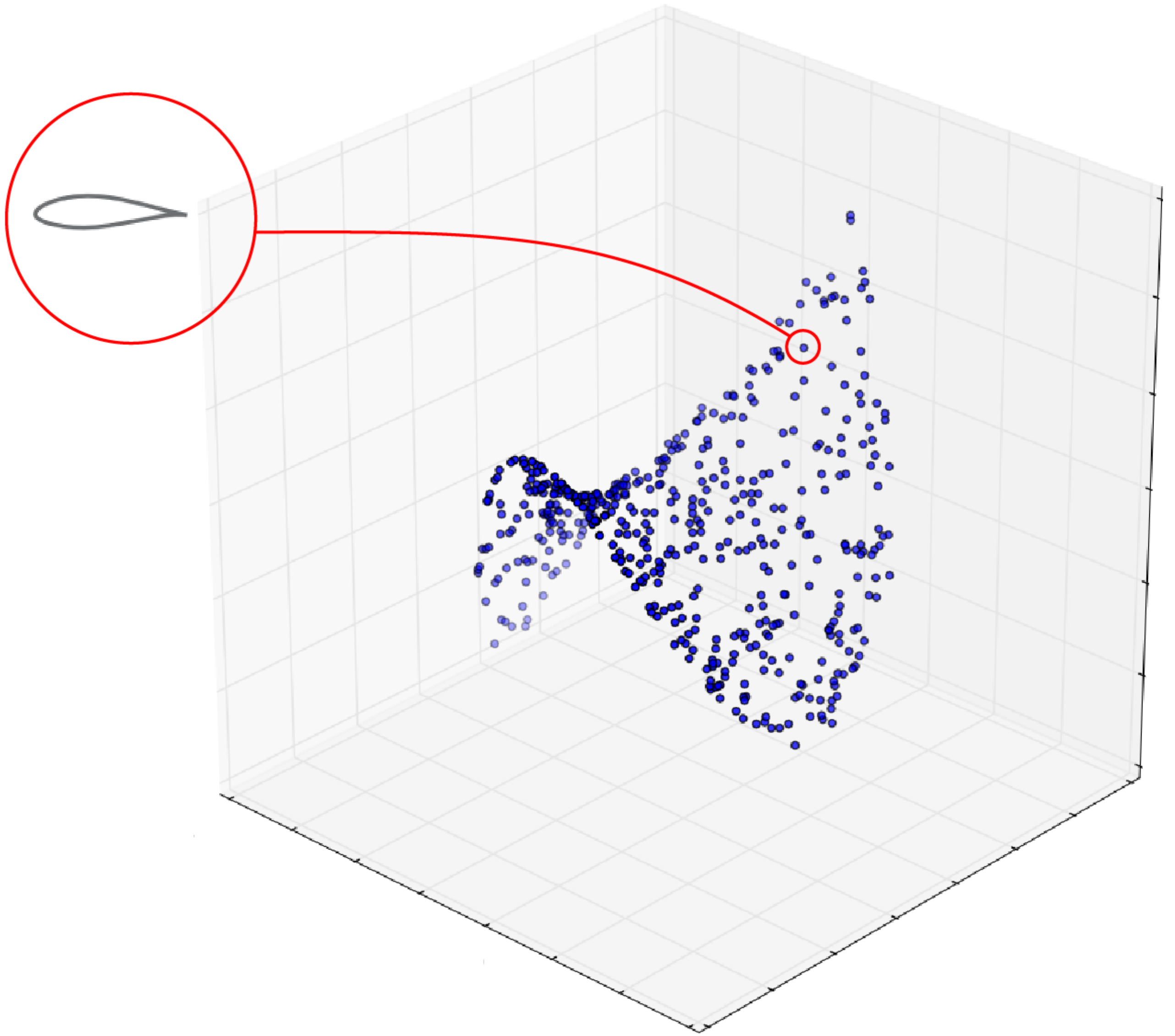
Advocate for additional studies of impact in these areas.

Thank you

Dr. Mark Fuge
Univ. of Maryland, College Park
(301) 405-2558
fuge@umd.edu
ideal.umd.edu

Backup Slides

What are Generative Models doing?

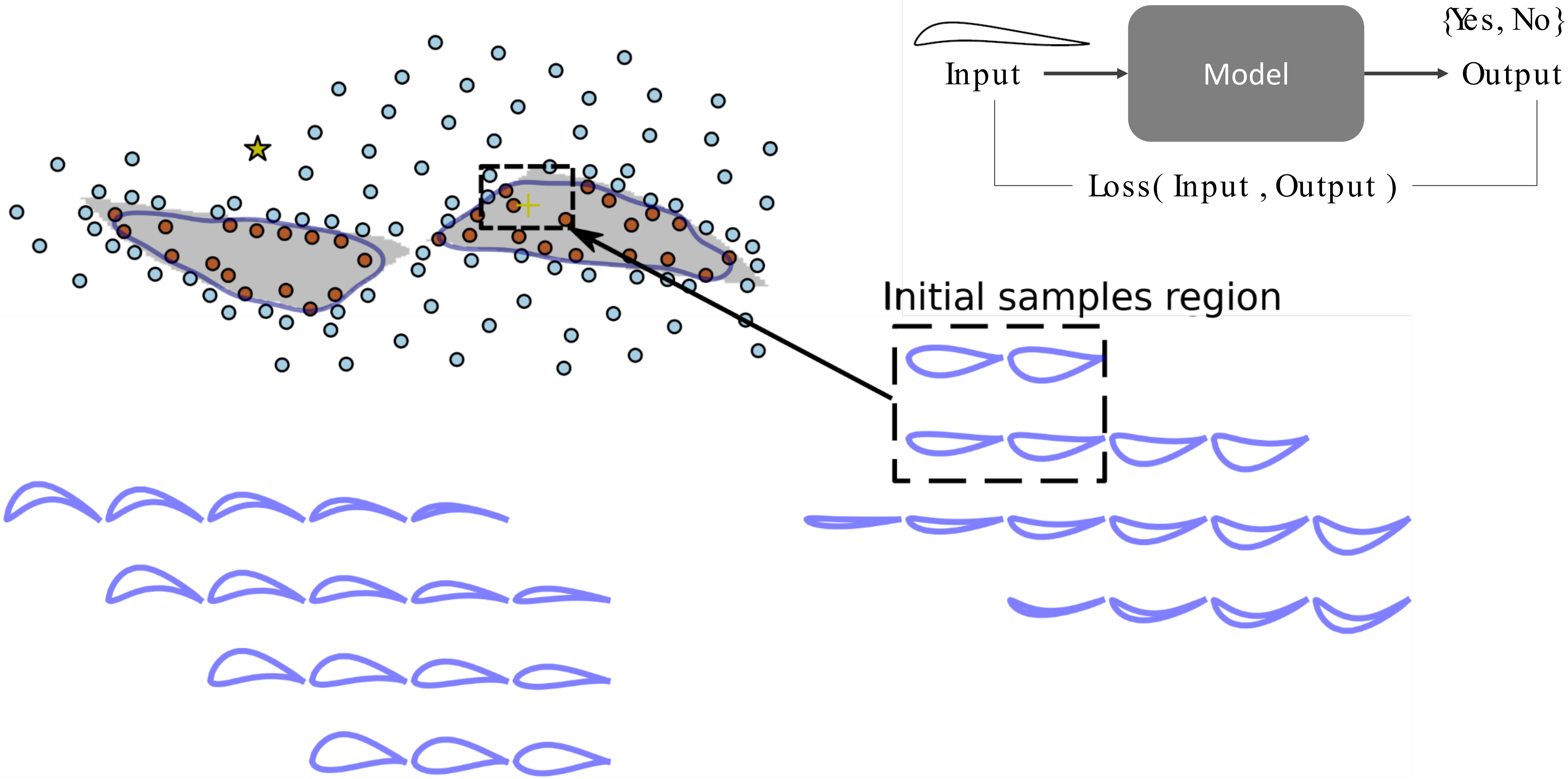


$$f: \mathcal{Z} \rightarrow \mathcal{X} \qquad \mathbb{P}(\boldsymbol{x}|\boldsymbol{z})$$

$$f^{-1}: \mathcal{X} \rightarrow \mathcal{Z} \qquad \mathbb{P}(\boldsymbol{z}|\boldsymbol{x})$$

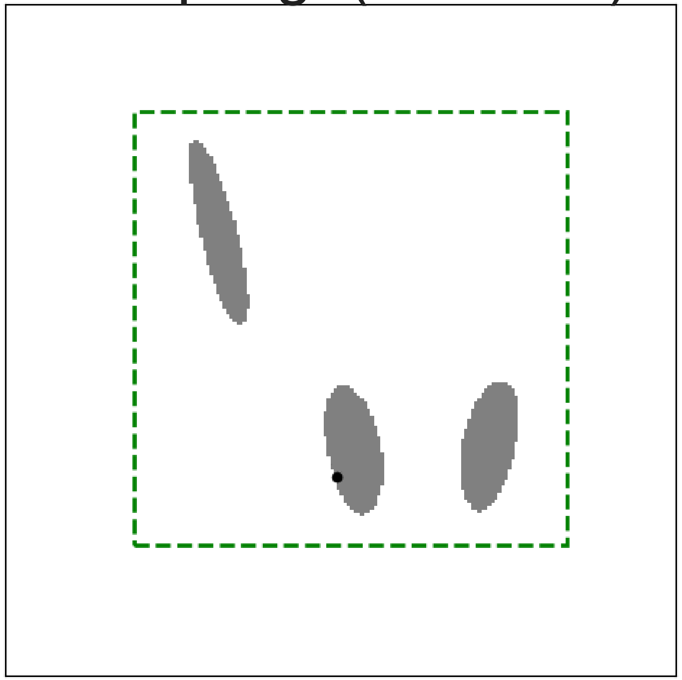
$$\log \mathbb{P}(\boldsymbol{x}) = \log \mathbb{P}(\boldsymbol{z}) + \log|\det \nabla_{\boldsymbol{x}} f^{-1}(\boldsymbol{x})|$$

Example: Identifying Feasible Performance Regions

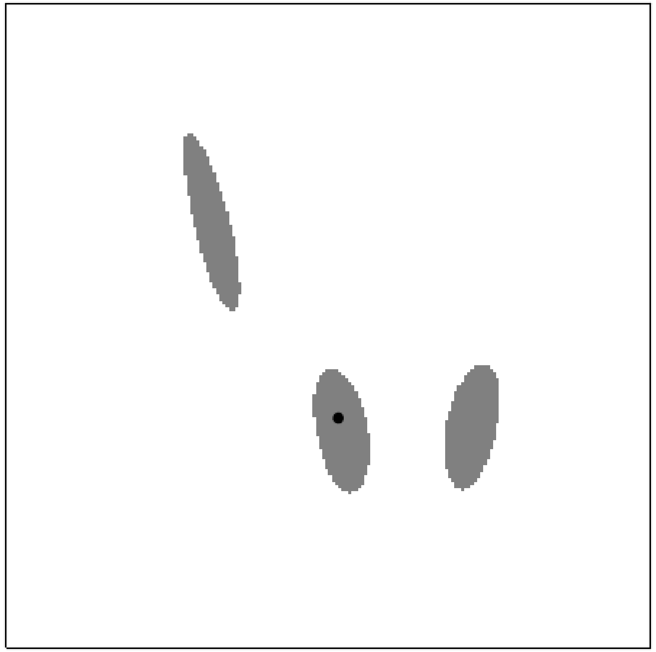


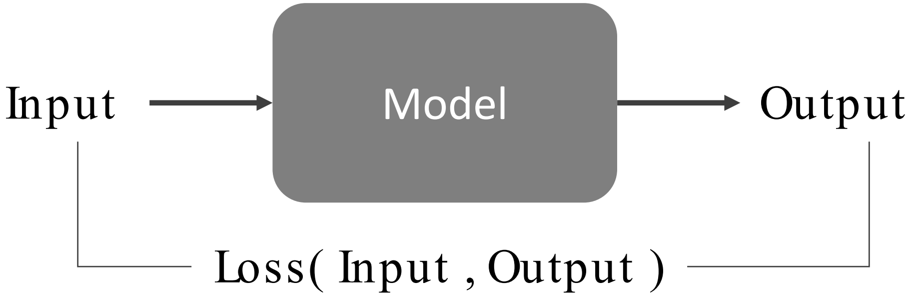
Example: Identifying Feasible Performance Regions

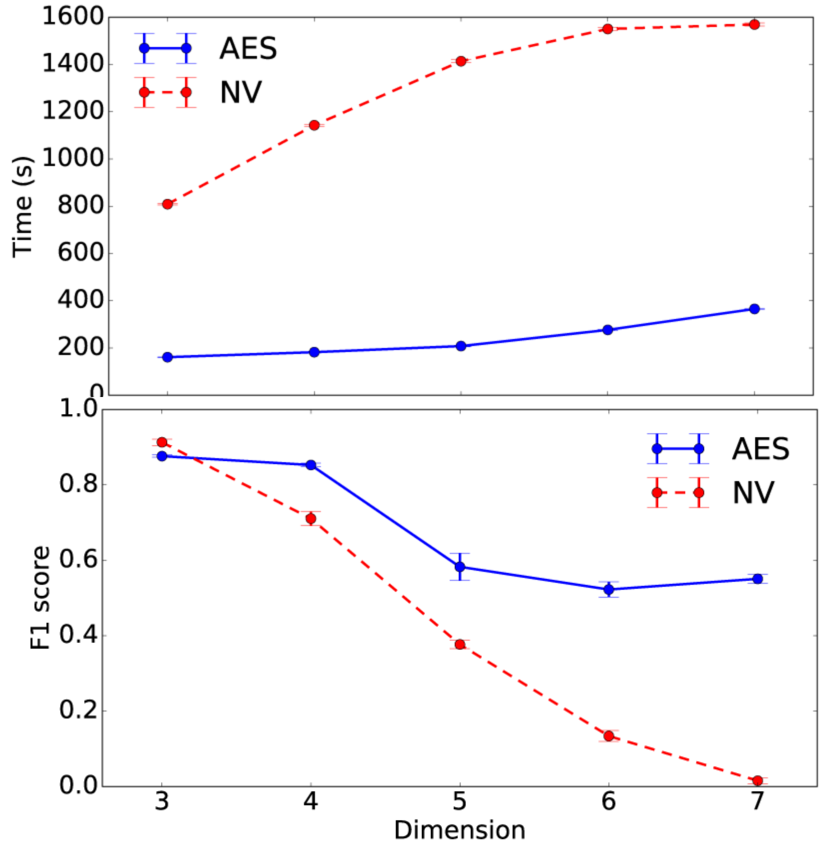
Conventional adaptive sampling (Straddle)



Active Expansion Sampling

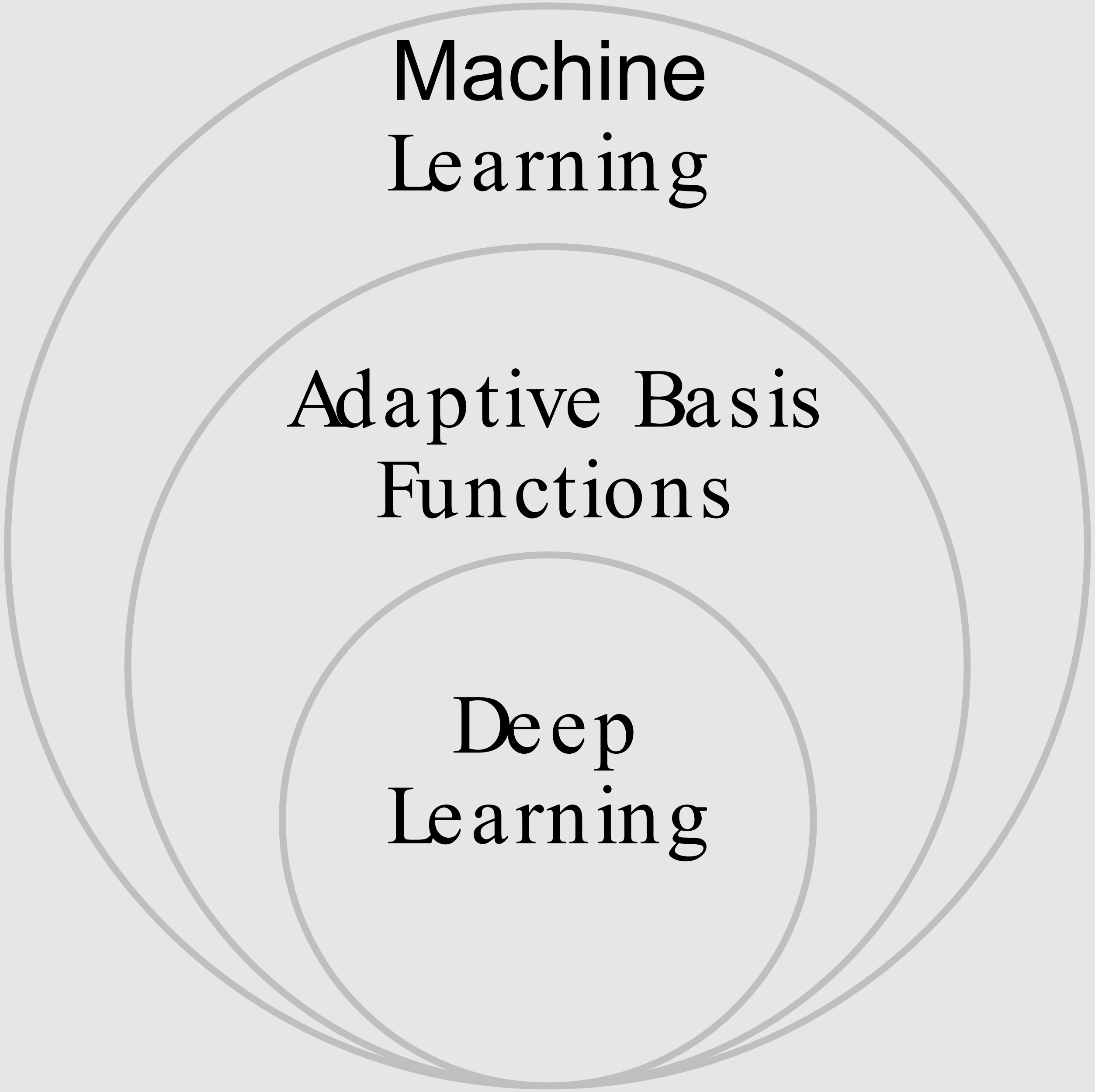


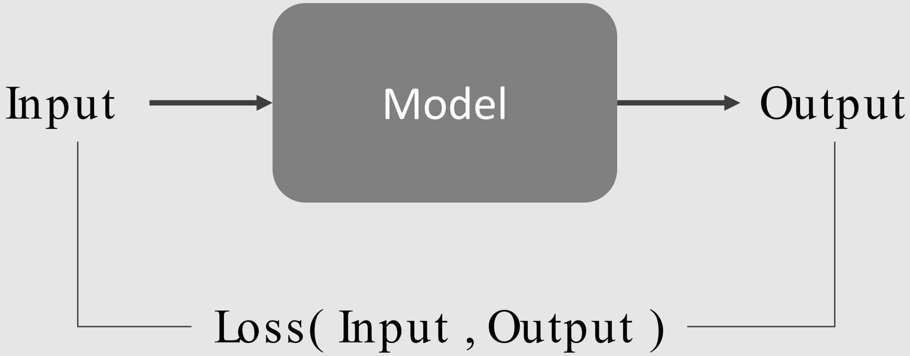




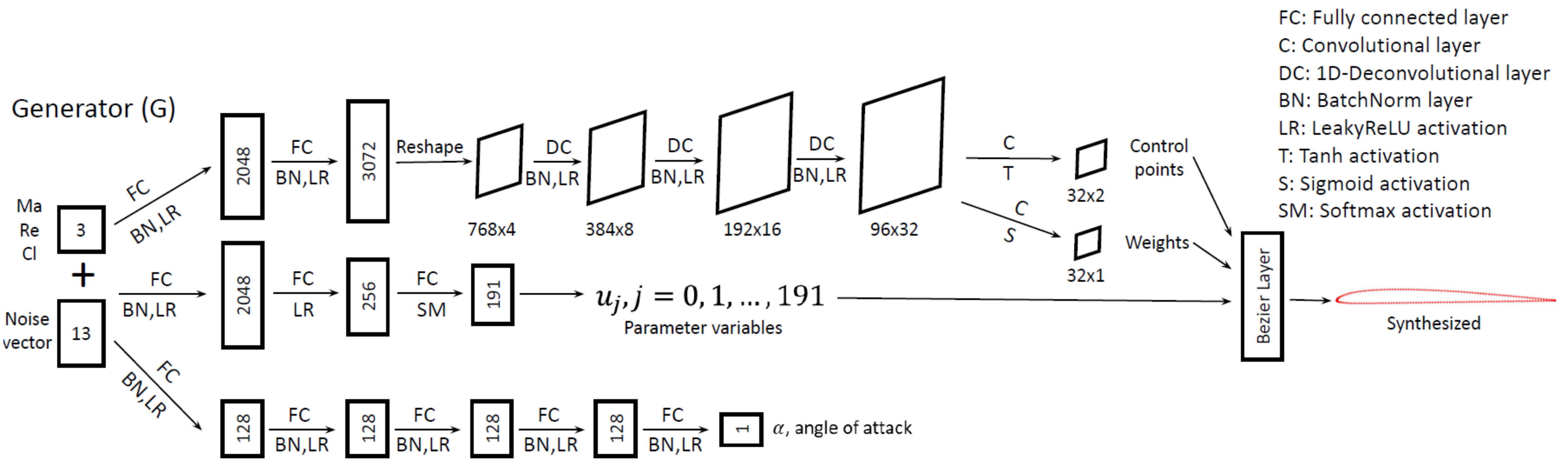
Introduction to Deep Learning

Dr. Mark Fuge
Univ. of Maryland, College Park
(301) 405-2558
fuge@umd.edu
ideal.umd.edu


Machine Learning
Adaptive Basis Functions
Deep Learning


Input
Model
Output
Loss(Input , Output)

Typical unreadable Deep Learning Slide from my research group for a recent DoE Technical Review



Let's build a Deep Learning model to predict airfoil lift

Input → Model → Output

Loss(Input , Output)

What should the input be?

(This will be our *basis function*)

Lift = Model(Input)

$$y = f(x)$$

How do you mathematically *represent* an airfoil?

How do you mathematically *represent* an airfoil?

$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \\ \\ \\ w_{199} \\ w_{200} \end{pmatrix}^T \begin{pmatrix} x_1 \\ y_1 \\ \\ \\ \\ x_{100} \\ y_{100} \end{pmatrix}$$

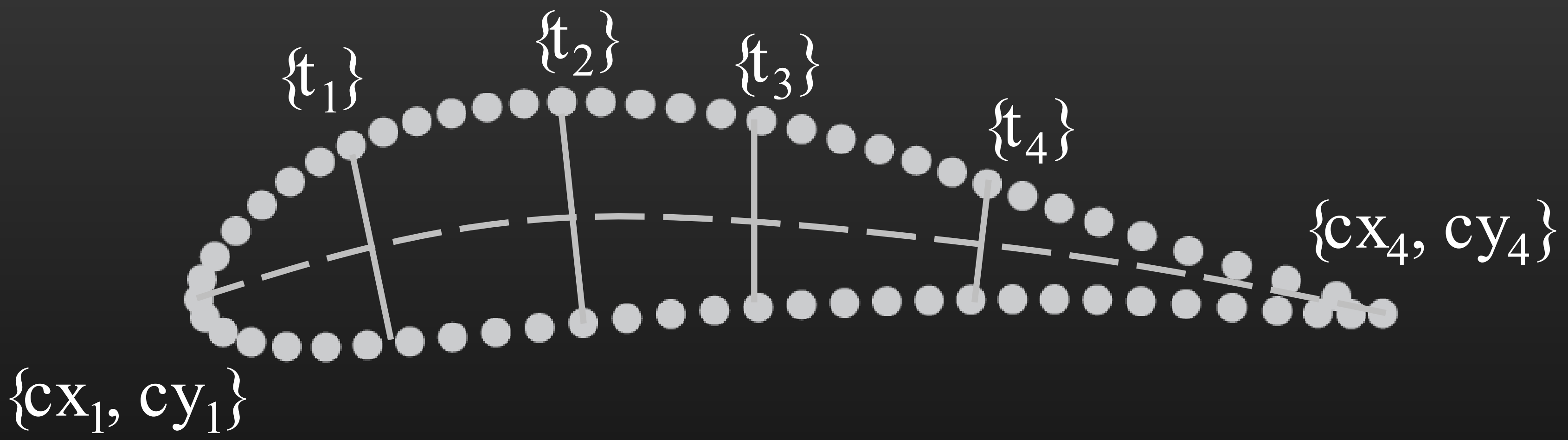
$$\begin{aligned} \text{Loss} &= (\text{Lift}_{\text{predicted}} - \text{Lift}_{\text{actual}})^2 \\ &= (\quad w^T x \quad - \text{Lift}_{\text{actual}})^2 \end{aligned}$$

Find w where:

$$\partial \text{Loss} / \partial w = 0$$

How do you mathematically *represent* an airfoil?

$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \\ \\ \\ w_{11} \\ w_{12} \end{pmatrix}^T \begin{pmatrix} cx_1 \\ cy_1 \\ \\ \\ \\ t_3 \\ t_4 \end{pmatrix}$$



Only thing we changed was the *basis*.

But the basis was fixed/ static.

What if we *adapted* or *learned* the basis?

How do you *adapt* a basis?

$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{199} \\ w_{200} \end{pmatrix}^T \begin{pmatrix} x_1 \\ y_1 \\ \vdots \\ x_{100} \\ y_{100} \end{pmatrix}$$

$$\begin{aligned} \text{Loss} &= (\text{Lift}_{\text{predicted}} - \text{Lift}_{\text{actual}})^2 \\ &= (w^T x - \text{Lift}_{\text{actual}})^2 \end{aligned}$$

How do you *adapt* a basis?

$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{199} \\ w_{200} \end{pmatrix}^T \begin{pmatrix} g(x_1) \\ g(y_1) \\ \vdots \\ g(x_{100}) \\ g(y_{100}) \end{pmatrix}$$

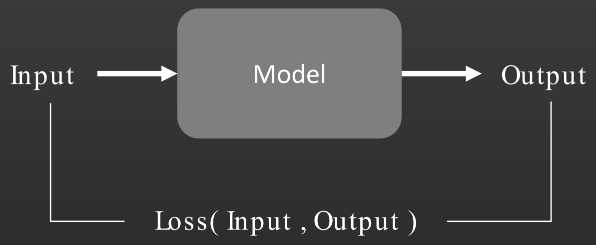
$$\text{Loss} = (\text{Lift}_{\text{predicted}} - \text{Lift}_{\text{actual}})^2$$
$$= (w^T g(x) - \text{Lift}_{\text{actual}})^2$$

How do you *adapt* a basis?

$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{199} \\ w_{200} \end{pmatrix}^T \begin{pmatrix} g(x_1) \\ g(y_1) \\ \vdots \\ g(x_{100}) \\ g(y_{100}) \end{pmatrix}$$

What is g?

Another model!



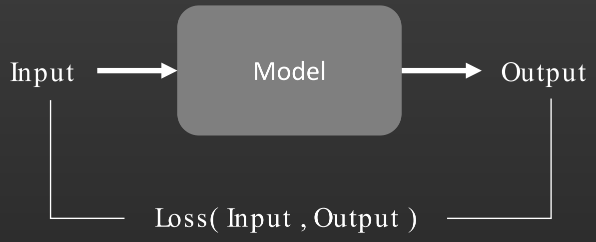
$$\begin{aligned} \text{Loss} &= (\text{Lift}_{\text{predicted}} - \text{Lift}_{\text{actual}})^2 \\ &= (\quad w^T g(x) \quad - \text{Lift}_{\text{actual}})^2 \\ &= (\quad w_1^T\{w_2^T(x)\} - \text{Lift}_{\text{actual}})^2 \end{aligned}$$

How do you *adapt* a basis?

What is g?

Another model!

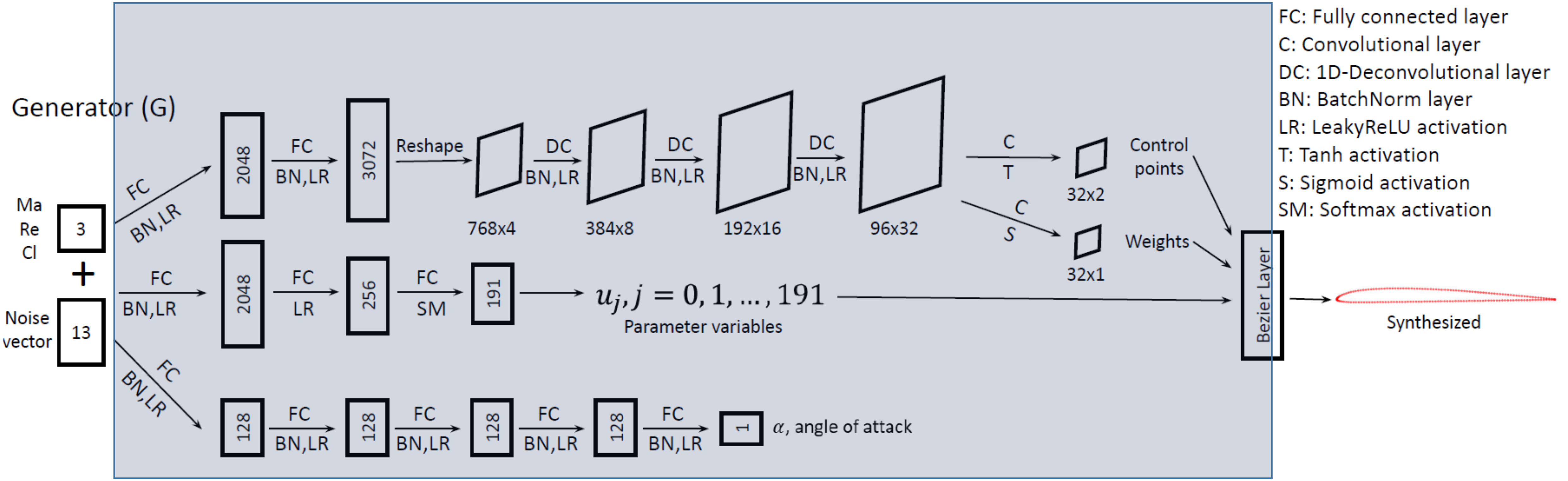
$$\text{Lift} = \begin{pmatrix} w_1 \\ w_2 \\ \\ \\ w_{199} \\ w_{200} \end{pmatrix}^T \begin{pmatrix} g(x_1) \\ g(y_1) \\ \\ \\ g(x_{100}) \\ g(y_{100}) \end{pmatrix}$$



$$\begin{aligned} \text{Loss} &= (\text{Lift}_{\text{predicted}} - \text{Lift}_{\text{actual}})^2 \\ &= (\quad w^T g(x) \quad - \text{Lift}_{\text{actual}})^2 \\ &= (\sigma(w_1^T\{\sigma(w_2^T(x))\}) - \text{Lift}_{\text{actual}})^2 \end{aligned}$$

Now we can see that the Deep Learning model is (in essence) a series of chained basis transformations!

Model g(x)



FC: Fully connected layer
C: Convolutional layer
DC: 1D-Deconvolutional layer
BN: BatchNorm layer
LR: LeakyReLU activation
T: Tanh activation
S: Sigmoid activation
SM: Softmax activation

Why use Deep Learning over other (non-Deep) approaches or not?

Advantages

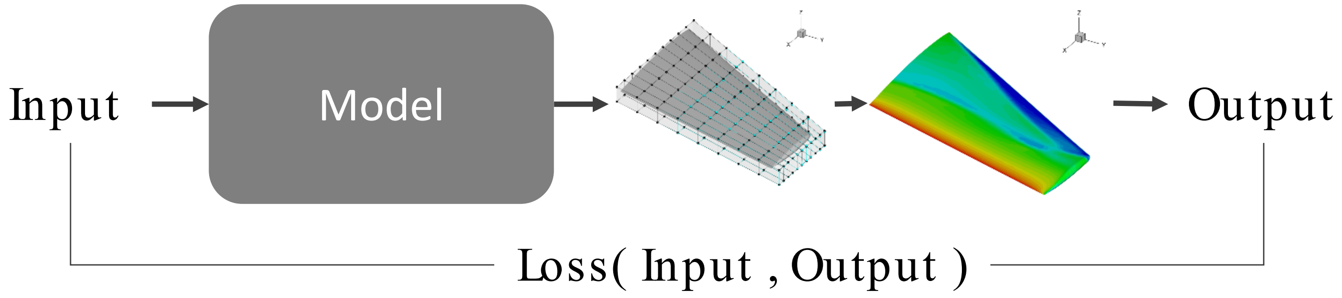
1. Fairly extensible with modern libraries
2. Plays nicely with other differentiable approaches
3. Good hardware acceleration
4. Active research community + industrial investment

Disadvantages

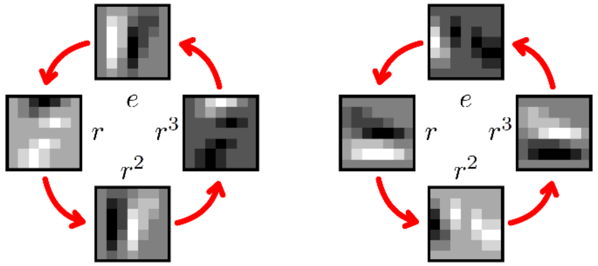
1. Certain modeling assumptions difficult to do
2. Certain architectures have difficulty converging or possess pathologies
3. Theory less developed than some other models

Opportunities and Directions

Merging of Engineering and Deep-Learning models



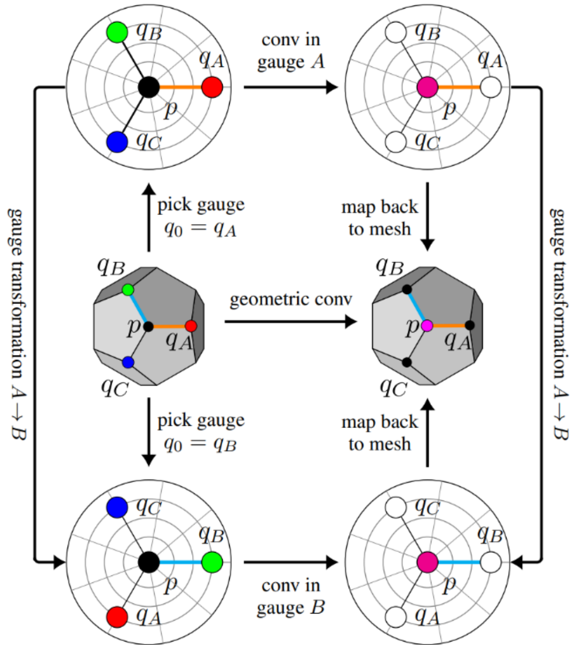
New invariances & constraints on Deep Learning models



Cohen and Welling, 2016
arxiv.org/abs/1602.07576

Generalizing Convolution

de Haan et al., 2020
arxiv.org/abs/2003.05425



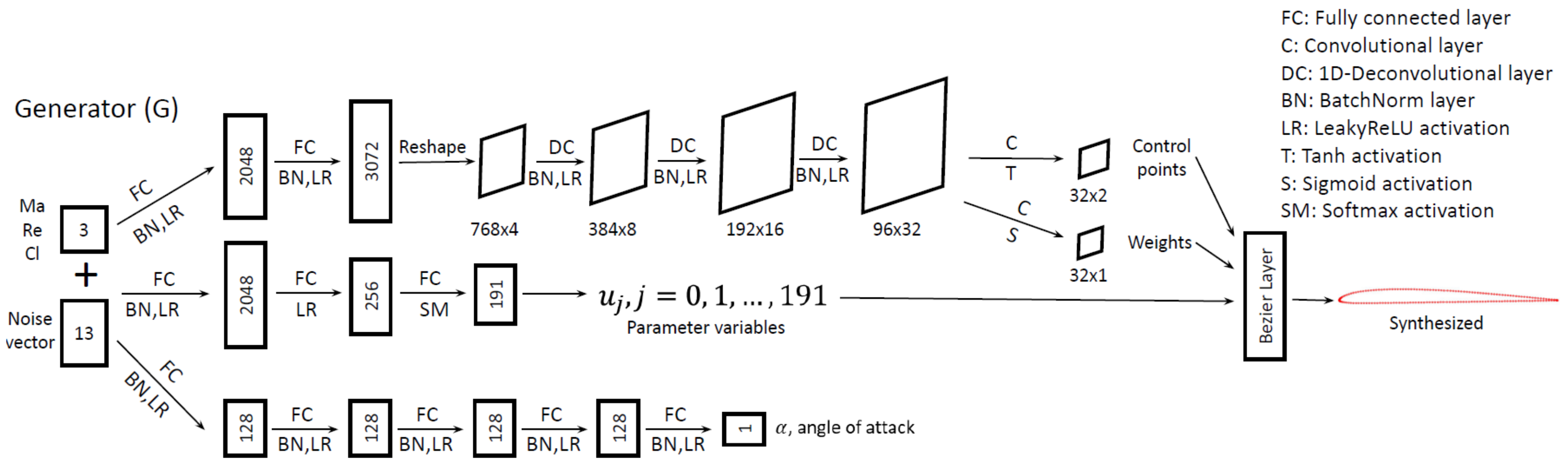
Combining Probabilistic and Deep Learning Models

Thank you

Dr. Mark Fuge
Univ. of Maryland, College Park
(301) 405-2558
fuge@umd.edu
ideal.umd.edu

Backup Slides

Now we can see that the Deep Learning model is (in essence) a series of chained basis transformations!



Minimize Sinkhorn divergence:

$$\min_G S_\lambda(p_r(\boldsymbol{x},\boldsymbol{y}) \| p_g(\boldsymbol{x}|\boldsymbol{y}) p_r(\boldsymbol{y}))$$

Loss

Conditional Formulation:

$$p(\mathbf{x},\mathbf{y}) = \int_{\hat{\mathbf{y}},\mathbf{z}} p(\mathbf{x},\mathbf{y} \mid \hat{\mathbf{y}},\mathbf{z})\, p_r(\hat{\mathbf{y}})\, p(\mathbf{z})\, d\hat{\mathbf{y}}\, dz$$

$$p(\mathbf{x},\mathbf{y} \mid \hat{\mathbf{y}},\mathbf{z}) \propto \exp - \frac{c([\mathbf{x},\mathbf{y}],[G(\mathbf{z},\hat{\mathbf{y}}),\hat{\mathbf{y}}])}{\lambda}.$$

Surrogate Log-Likelihood (SLL):

$$\log p(\mathbf{x},\mathbf{y}) \geq -\frac{1}{\lambda}\mathbb{E}_{\mathbb{P}^\star_{Z|X,Y}}\left[c([\mathbf{x},\mathbf{y}],[G^\star(\mathbf{z},\mathbf{y}),\mathbf{y}])\right]$$

$$+\mathbb{E}_{\mathbb{P}_Z}\left[\log p(\mathbf{z})\right] + H\left(\mathbb{P}^\star_{Z|X,Y}\right) + \log p_r(\mathbf{y}) + \text{const}$$

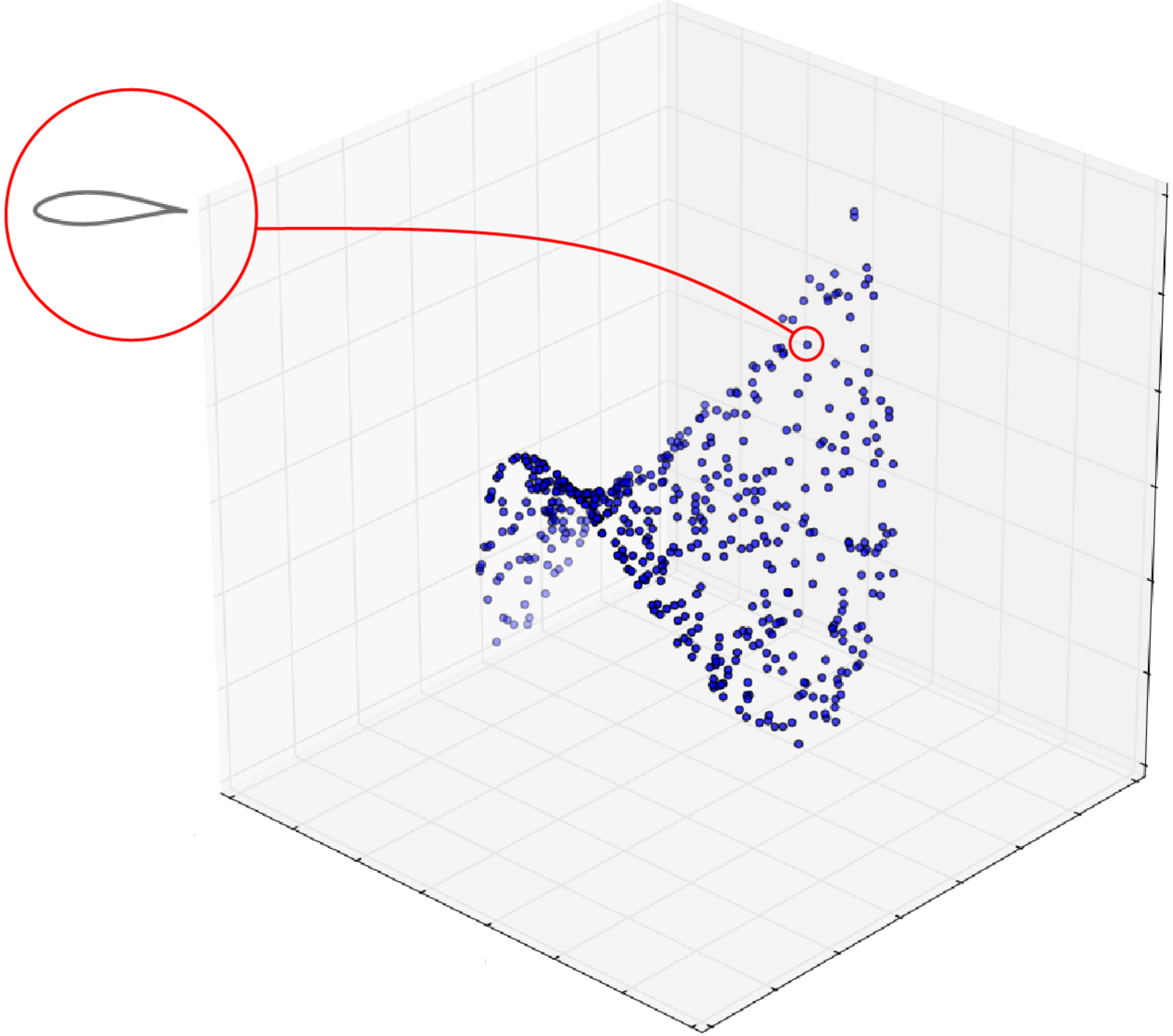
in which

$$\mathbb{P}^\star_{Z|X,Y} = \mathbb{P}_Z(\mathbf{z}) \exp \frac{v^\star([\mathbf{x},\mathbf{y}],[G^\star(\mathbf{z},\mathbf{y}),\mathbf{y}])}{\lambda}.$$

Cost Function:

$$c([\mathbf{x},\mathbf{y},\mathbf{b}],[\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{b}}]) = |\mathbf{x}-\hat{\mathbf{x}}| + |\mathbf{y}-\hat{\mathbf{y}}| + |\mathbf{b}-\hat{\mathbf{b}}|$$

What are Generative Models doing?



$$f: \mathcal{Z} \to \mathcal{X} \qquad \mathbb{P}(\boldsymbol{x}|\boldsymbol{z})$$

$$f^{-1}: \mathcal{X} \to \mathcal{Z} \qquad \mathbb{P}(\boldsymbol{z}|\boldsymbol{x})$$

$$\log \mathbb{P}(\boldsymbol{x}) = \log \mathbb{P}(\boldsymbol{z}) + \log |\det \nabla_{\boldsymbol{x}} f^{-1}(\boldsymbol{x})|$$

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

THEORY AND APPLICATION FOR ENGINEERING

Thurston Sexton

Knowledge Extraction and Application Project

Systems Integration Division

Engineering Laboratory

National Institute of Standards and Technology

U.S. Department of Commerce

DISCLAIMER

The use of any products described in any presentation does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

BACKGROUND: PROJECT/PROGRAM OVERVIEW

Knowledge Extraction** and **Application

- Much of manufacturing know-how is computationally inaccessible, within informally-written documents
- Create human-centric data pipelines to extract value from existing unstructured data at minimal labor cost
- Develop guidelines for using semi-structured data in KPI creation, functional taxonomy prediction, and customized worker training paths

BACKGROUND: MAINTENANCE WORK-ORDER DATA

“HP coolant pressure at 75 psi”

“Hyd leak at saw atachment”

“Major hydraulic leak at Sp#6 horseshoe”

“Clamping spool guard broken”

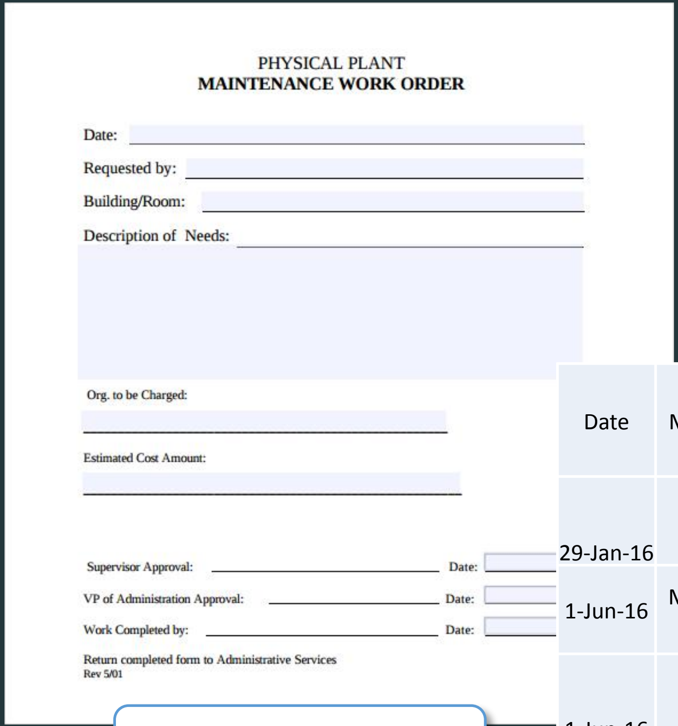
“Replaced seal in saw attachment but still leaking – Reapirs pending with ML”

“Bad Gauge / Low pressure lines cleaned ou”

“Replaced – Operator could have done this!”

“Repaired horseshoe seals”

BACKGROUND: CURRENT MWO DATA ENTRY



PHYSICAL PLANT
MAINTENANCE WORK ORDER

Date:

Requested by:

Building/Room:

Description of Needs:

Org. to be Charged:

Estimated Cost Amount:

Supervisor Approval: Date:

VP of Administration Approval: Date:

Work Completed by: Date:

Return completed form to Administrative Services
Rev 5/01

WORK ORDER FORMS

SPREADSHEETS

Date	Mach	Description	Issued By	Date Up	Maint Tech Assigned	Resolution
29-Jan-16	H15	St#14 tool detect INOP	JS	29-Nov-16	SA	Slug detector at station 14 not working. Would not recognize “Start” signal.
1-Jun-16	Mitsu FT	Brakes worn -Not stopping when in gear	AB	28-Jun-16	Steve A	Repaired
1-Jun-16	H8	St#7 rotator collet broken -wait for Bob B to show him how to remove	JS	8-Jun-16	John Smith	Machine went offline on 6/8 -Mark removed and instructed Bob B on removal/install process

Do “AI” to it! (...?)

Natural Language Processing (et al.) as Engineering Tools

TODAY'S TALK: TAKE-HOME

- NLP “Theory” Basics
    a. Data **models** and engineering **assumptions**
    b. NLP “**Tasks**” and **approaches**
    c. **Metrics** and **Evaluation**

- Contextualize NLP techniques, paradigms
    a. How NLP concepts interface with “Engineering Practice”
    b. Continuous interaction between experts (domain ←→ NLP)

TODAY'S TALK: STRUCTURE

Engineering Practice

- Goal & Approach
- Assumptions
- Measure & Evaluate
- Validate

"State the methods followed and why."

"State your assumptions."

"Apply adequate factors of safety."

"Always get a second opinion."

Hutcheson, M. L. (2003). *Software testing fundamentals: Methods and metrics*. John Wiley & Sons.

TODAY'S TALK: STRUCTURE

Engineering Practice

- Goal & Approach — *"State the methods followed and why."*
- Assumptions — *"State your assumptions."* — Start Here
- Measure & Evaluate — *"Apply adequate factors of safety."*
- Validate — *"Always get a second opinion."*

Hutcheson, M. L. (2003). *Software testing fundamentals: Methods and metrics*. John Wiley & Sons.

ASSUMPTIONS

That turn “Natural Language” into something to “Process”

ASSUMPTIONS: RULE-BASED VS. NUMERICAL

Some very successful ways to “process” natural language involve **rules**.

Assume a language model based on known “logic”:

- Pattern Matching (e.g. regex), “coding”, etc.
- Clear definitions and transparent assumptions (iterate!)
- Can be **powerful** and **efficient**
- Can be **brittle** and **labor**-intensive

Newer techniques assume the text and its **statistical** properties **alone**

ASSUMPTIONS: THE CONTEXT SPECTRUM

- How do we turn text into "numbers"?
- Traditional techniques come in two "flavors"
  a. Bag-of-Words *(Global Frequency and Context)*
  b. Markov Model *(Local Sequence Probability)*
- Opposite answers to the question:

  *"How much does **global** vs. **local** matter to you and/or this text?*

ASSUMPTION: GLOBAL FREQUENCY & CONTEXT

Basic Bag-of-Words

*Words in similar **contexts** <u>are</u> **similar**.*

- *Hydraulic leak at saw attachment*
- *Worn seal caused leak, replaced seal.*
- *Replaced saw, operator could have done this…*

	Hyd.	leak	saw	seal	rep.	…
Doc 1	1	1	1	0	0	…
Doc 2	0	1	0	2	1	…
Doc 3	0	0	1	0	0	…

- Remarkably Powerful
- Similarity is “vector directional”
  - Documents or Terms
  - → Cosine Similarity

ASSUMPTION: GLOBAL FREQUENCY & CONTEXT

Basic Bag-of-Words

*Words in similar **contexts** <u>are</u> **similar**.*

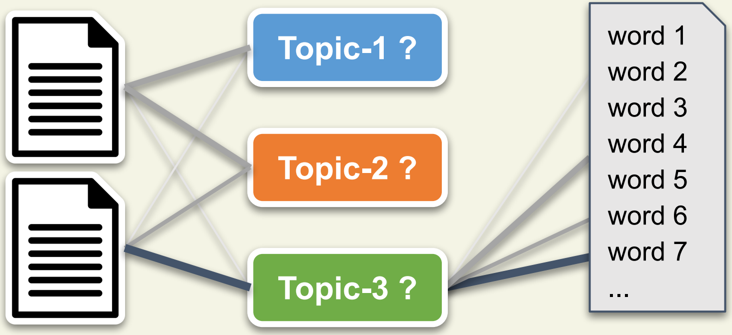
- *Hydraulic leak at saw attachment*
- *Worn seal caused leak, replaced seal.*
- *Replaced saw, operator could have done this...*

	Hyd.	leak	saw	seal	rep.	...
Doc 1	1	1	1	0	0	...
Doc 2	0	1	0	2	1	...
Doc 3	0	0	1	0	0	...

- Remarkably Powerful
- Similarity is "vector directional"
    - Documents or Terms
    - → Cosine Similarity

Modifications

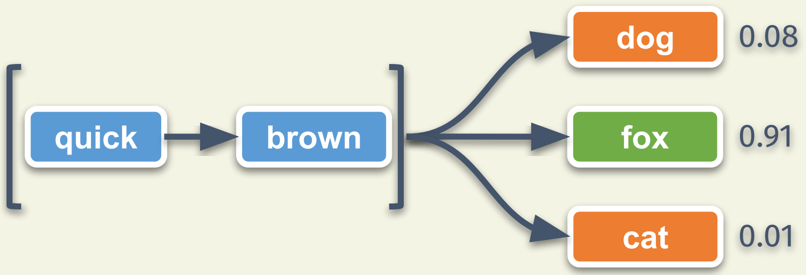
- Re-weighting schemes
    - Normalization, TF-IDF
    - Ties to informational entropy
- Dimension Reduction & **Topics**
    - Some "latent" set of topics:
      "Stuff we talk about" has less variety than "words we have"
    - Acronym soup
      PCA,SVD,LSA,NMF,LDA,TSNE,UMAP



ASSUMPTION: LOCAL SEQUENCE PROBABILITY

Markov Model

Next “states” (read: token/character) is conditionally dependent on the past:

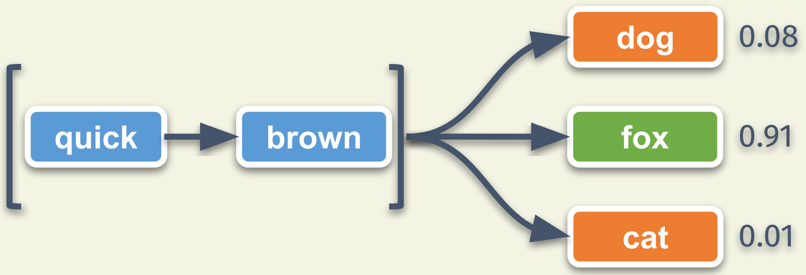


- Useful to generate text and estimate cond. probabilities
- High preference for observed sequences (precision)

ASSUMPTION: LOCAL SEQUENCE PROBABILITY

Markov Model

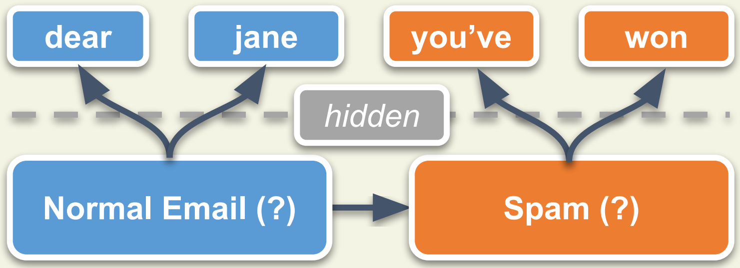
Next "states" (read: token/character) is conditionally dependent on the past:



- Useful to generate text and estimate cond. probabilities
- High preference for observed sequences (precision)

Hidden Markov Model

What we "observe" are emissions from a sequence of states we cannot observe.



- Used for last-gen. language models, bio-informatics, etc.
- Modular! See: GMMs, Bayes-nets...

ASSUMPTIONS: MODERN EMBEDDINGS

But... neural-nets?!

- We like the global context, but also want local sensitivity...
- Neural Nets can be “trained” to find a **vector space** model that **balances** both
    a. **Trained** is the operative term
    b. Packages/tools that let us “embed” text have **already trained** on a textual corpus
- You are assuming your text is “like” *that* text

*Otherwise these are an **approach**—and require proper design!*

ASSUMPTIONS: MORE ON "MODERN EMBEDDINGS"

- *Word2Vec (2013)* trains on a *word*-level
  - Continuous Bag-of-Words (**CBOW**): target word from local context
  - **Skip-Gram**: local context from target word
  - Maintains semantic linearity ("word algebra") — also see GloVe (2014)

lunch + night - day → dinner

better - good + bad → worse

wine + barley - grapes → beer

coffee - drink + snack = pastry

ASSUMPTIONS: MORE ON "MODERN EMBEDDINGS"

- *Word2Vec (2013)* trains on a *word*-level
  - Continuous Bag-of-Words (**CBOW**): target word from local context
  - **Skip-Gram**: local context from target word
  - Maintains semantic linearity ("word algebra") — also see GloVe (2014)

lunch + night - day → dinner

better - good + bad → worse

wine + barley - grapes → beer

coffee - drink + snack → pastry

- *BERT (2018)* is a *sub-word* model...**context** (sentence) dependent!
  - Can capture separate semantic meaning (homophones) and out-of-vocab.
  - State-of-the-art in 2019; used for your Google searches.

GOALS & APPROACH

NLP Tasks and “The Pipeline”

GOALS & APPROACHES: OVERVIEW

- Typical NLP Tasks
  (and their image-processing relatives)
  a. Document Grouping, Classification
  b. Keyword Extraction, Multi-Label Classification
  c. Named Entity Recognition and Parts-of-Speech
- The NLP "Pipeline"
  a. Preprocessing
  b. Analyses

GOAL: DOCUMENT TYPING

- Clustering (Unsupervised)
  - Detect “natural groupings” for analysts to parse
  - Also: interpreting topic models
  - May or may not be relevant, but a useful tool

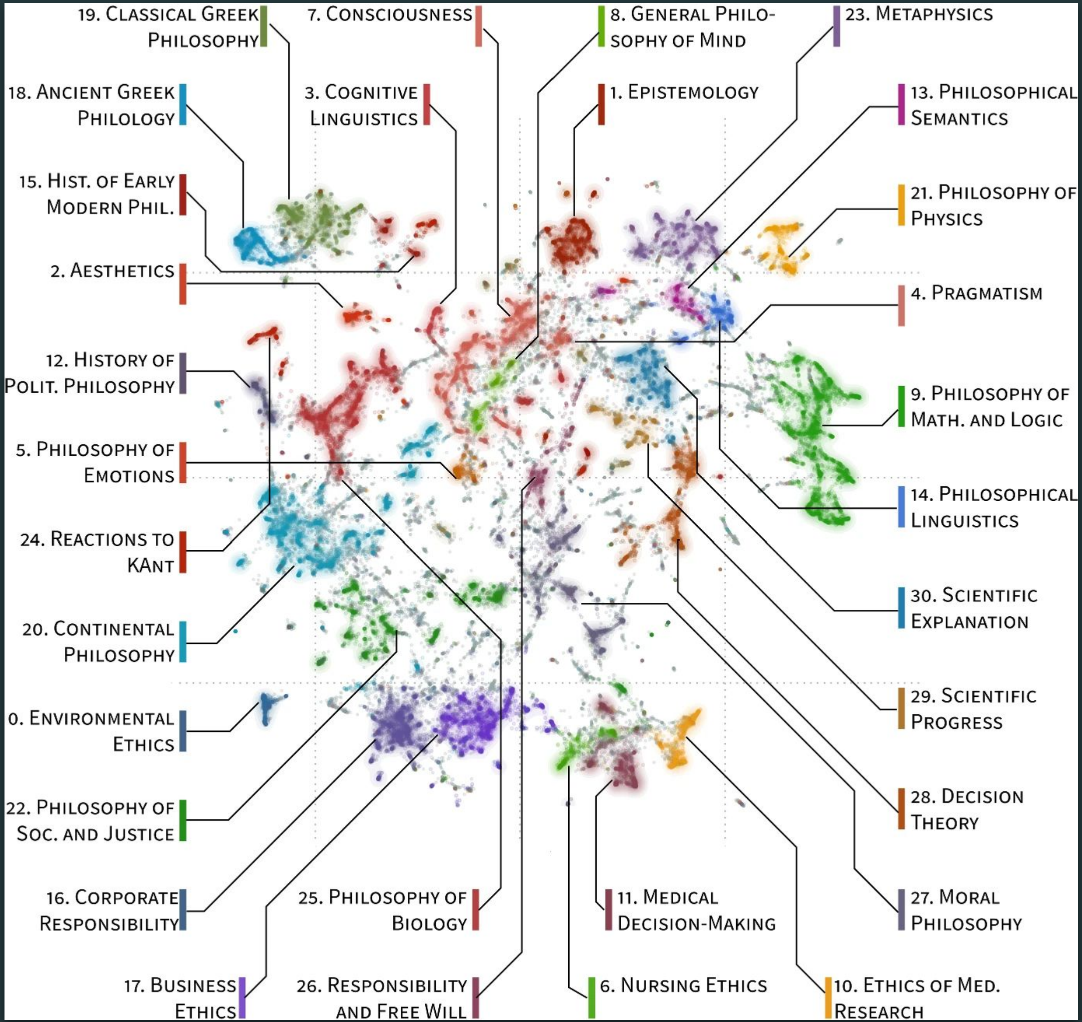
The Structure of Recent Philosophy

Noichl, M. Modeling the structure of recent philosophy. *Synthese* **198**, 5089–5100 (2021).
https://doi.org/10.1007/s11229-019-02390-8

Each “dot” is a paper.

- Embed to 2-dimensions (UMAP)
- Clustering (HDBScan)
- Interpret, synthesize (hard)

Fully interactive online:
https://homepage.univie.ac.at/maximilian.noichl/full/zoom_final/index.html



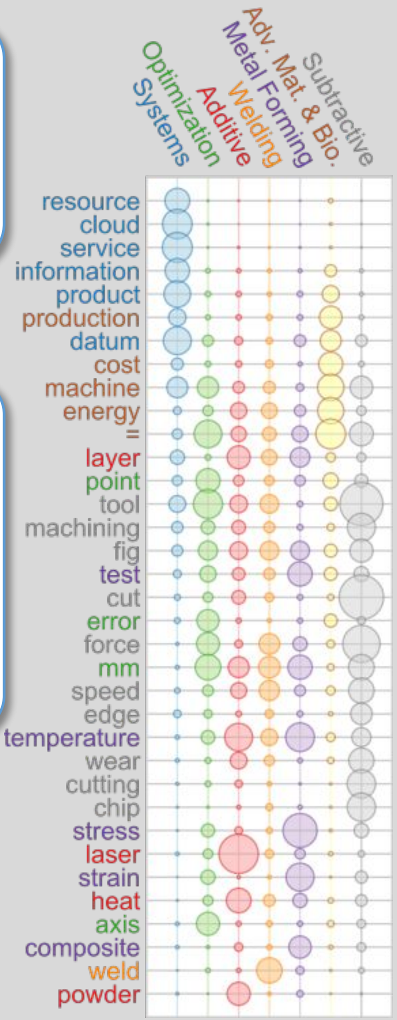
MSEC: A Quantitative Retrospective

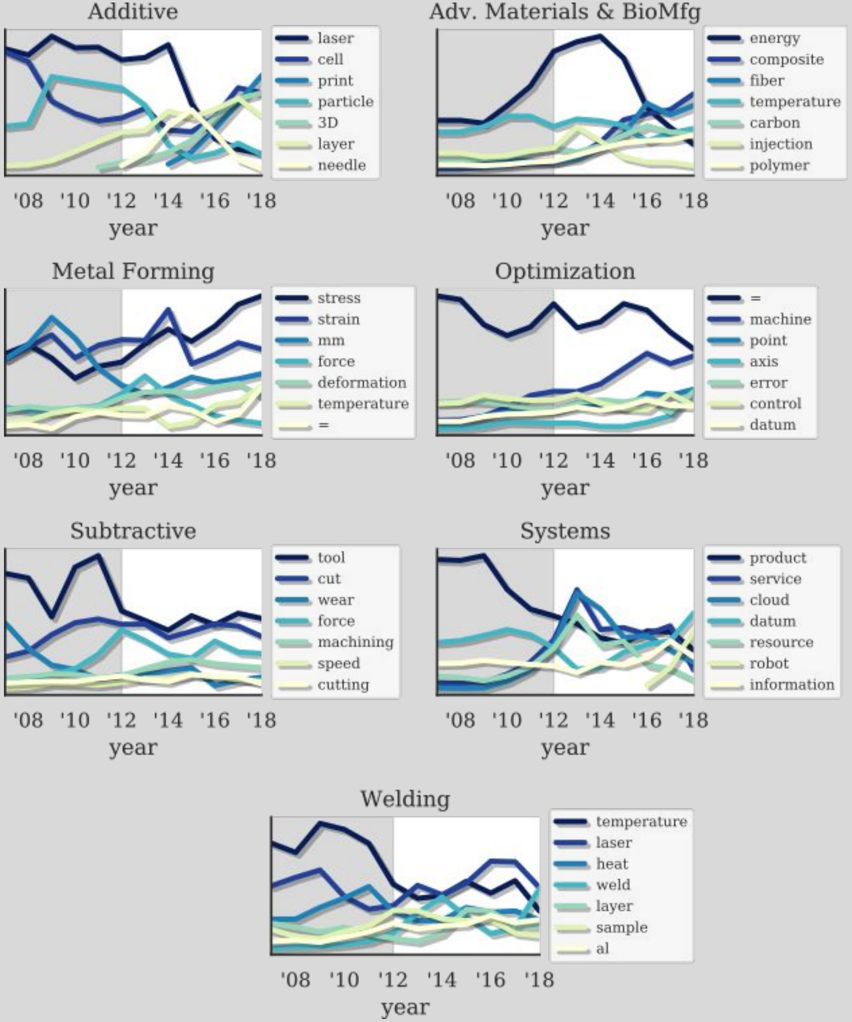
Sexton, T, Brundage, MP, Dima, A, & Sharp, M. "MSEC: A Quantitative Retrospective." September 2020 https://doi.org/10.1115/MSEC2020-8440

Topic Models as an approach to typing:

- Useful understanding
- LDA for static
- Dynamic LDA over time

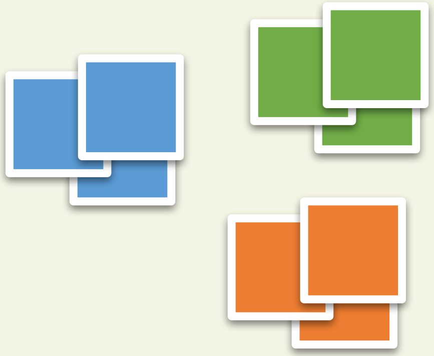
We had to name the topics.



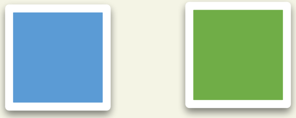


GOAL: DOCUMENT TYPING

- Clustering (Unsupervised)
  - Detect “natural groupings” for analysts to parse
  - Also: interpreting topic models
  - May or may not be relevant, but a useful tool



- Classification (Supervised)
  - Labels required: 1 per category (mutually exclusive)
  - Can be useful for recommendations: “relevant vs. not”
  - Images: “*is this a stoplight?*” or “*which animal?*”, etc.



Cat ? Dog ?

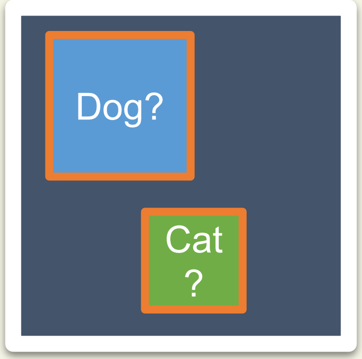
GOAL: DOCUMENT KEYWORDS

- Keyword Extraction (Unsupervised)
  - Use statistical properties to find "important terms"
  - Also see: text summarization
  - TF-IDF (sum), TextRank (graph-based), YAKE, +more

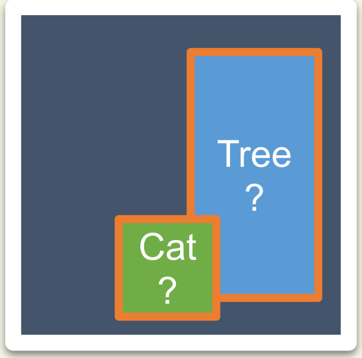
- Multi-Label Classification (Supervised)
  - Labels required: **multiple**-per-document (multiset)
  - Several ways to train, can use domain-knowledge
  - **Harder** problem, but maybe easier to **make** training data...
  - Images: "*What animals are present?*"

GOAL: ENTITY RECOGNITION

- Named Entity Recognition
  - Find **text spans** that contain **keywords**, and **annotate** them
  - Predetermined vocabulary/taxonomy (usually 2-levels)
  - E.g. “I went to New York [LOC]” or “They owe me $25 [CURR]”
  - Images: *“highlight and label the animals...”*



- Parts-of-Speech
  - Automatic determination of **grammar** information
  - SVO triples, dependency parsing, etc.
  - Can be used to “mine” **knowledge graphs**
  - Domain/language-dependent... hard with technical text!



GOALS: OTHERS WORTH MENTIONING

Wide variety of other tasks:

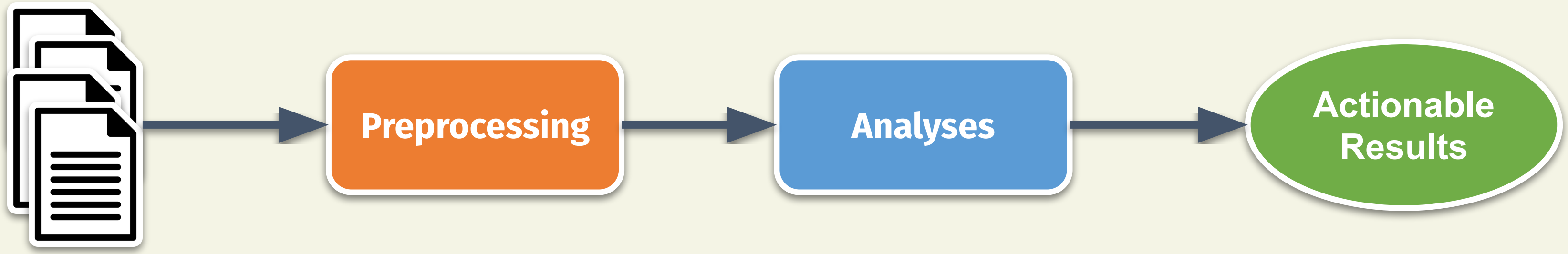
- Sentiment Analysis
- Seq2Seq & Machine Translation
- Reading complexity and writing quality, inclusivity
- Question Answering
- Text Synthesis

What does it take to get to this point?

PROCESS: "THE PIPELINE"

In theory, the NLP Pipeline is a

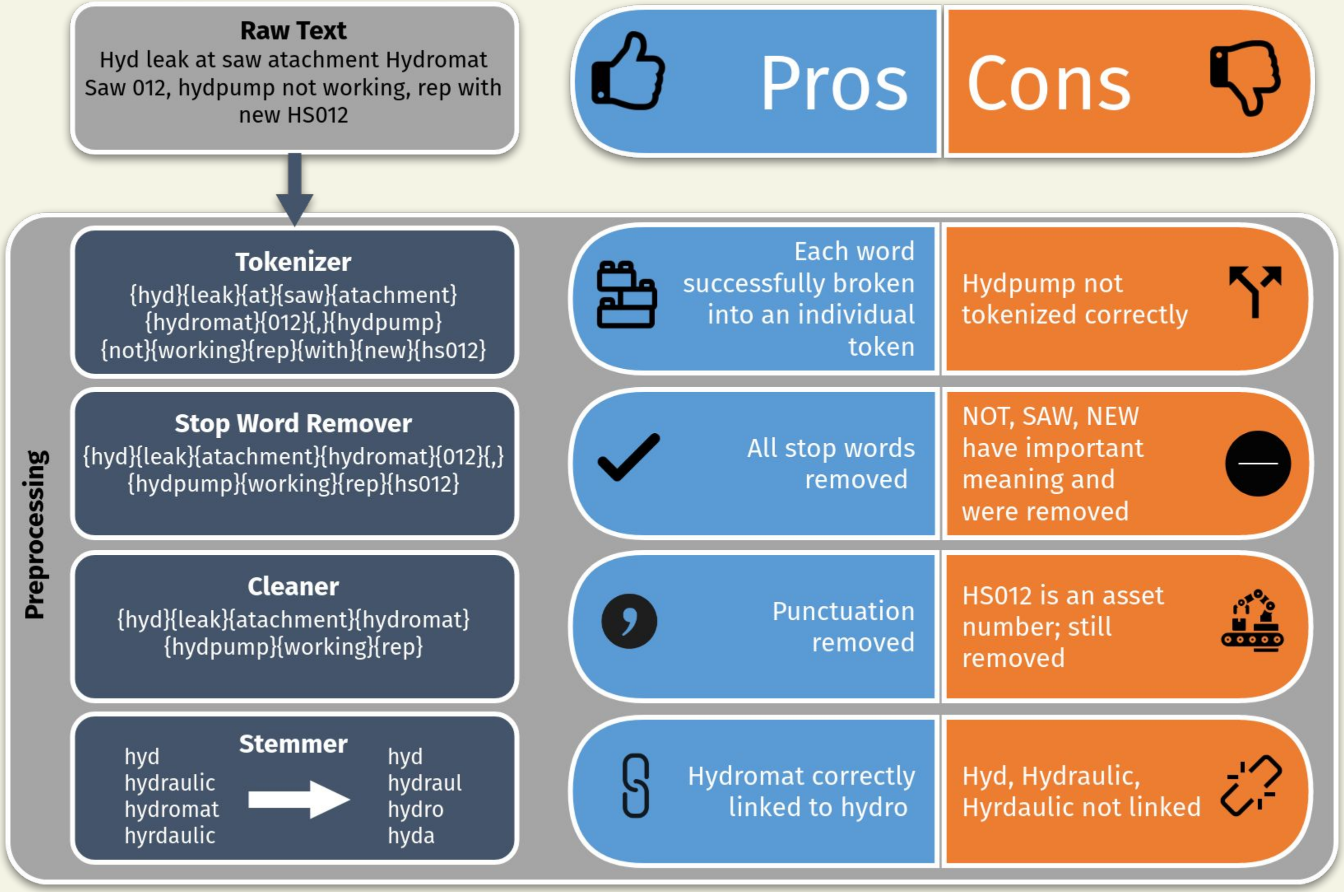
- Sequential progression, that
- Provides usable insight



Impossible to outline the number of variations on this "theme"... Here's:

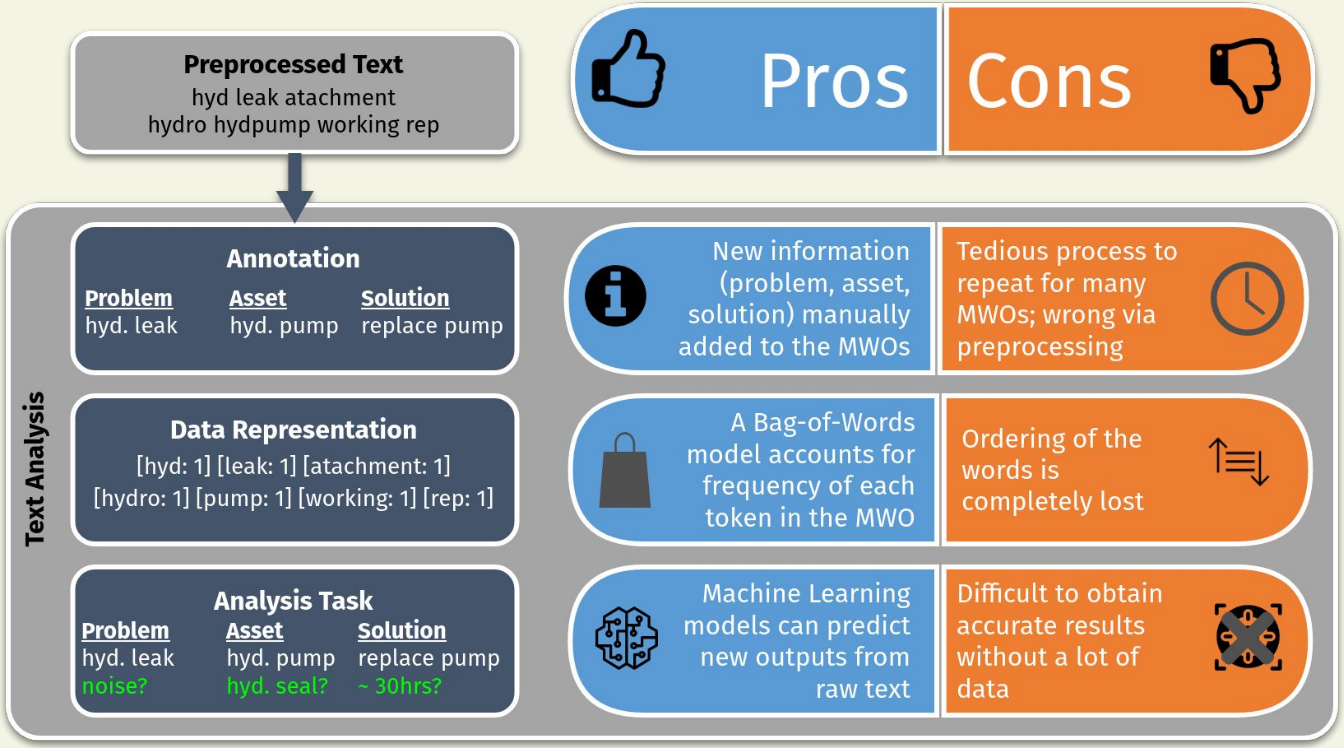
- A common sequence — a *day-in-the-life* of your analyst.
- Benefits and drawbacks of each step

PROCESS: TEXT PREPROCESSING



Technical language processing: Unlocking maintenance knowledge. Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). *Manufacturing Letters*, *27*, 42-46.
Image adapted from original.

PROCESS: TEXT ANALYSES



Technical language processing: Unlocking maintenance knowledge. Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). *Manufacturing Letters*, *27*, 42-46.
Image adapted from original.

MEASURE & EVALUATE

Importance of metrics and knowing what gets evaluated

MEASURE & EVALUATE: OVERVIEW

Key skill of the analyst or engineer is knowing how to **translate:**

Qualitative *needs and constraints →* ***Quantitative*** *metrics and evaluations*

- What do I want to measure?
  - Do **my assumptions** conflict with the measurement?
  - Do the **metric's assumptions** conflict with my goal/process?
  - Will **multiple metrics** provide a broader insight? (yes)
- What constitutes progress toward, or success in, my goal?
  - Have I encoded my (stakeholder) expectations (preferences) sufficiently?
  - Do I have parameters to tune (continuously and/or iteratively)?

Most important: have I **transparently documented** my decisions for **iteration**?

MEASURE

What do I need to measure? Have I "done my homework"?

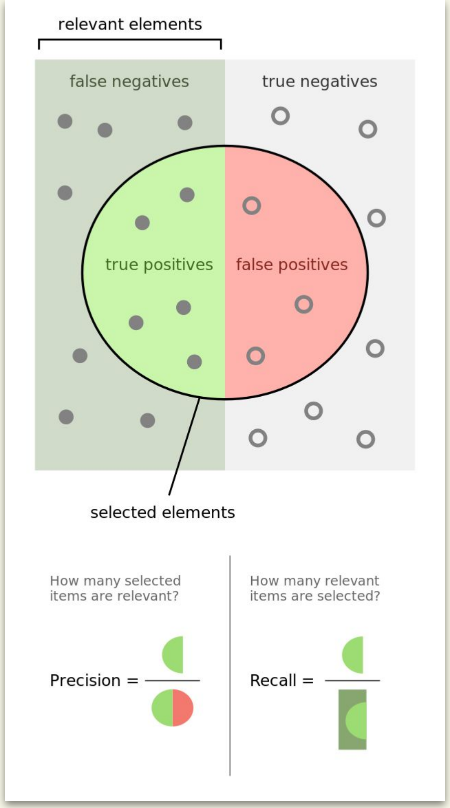
- Similarity or Distance
  - Discrete options, spellings: *Levenstein, Hamming, SymSpell, Jaccard*
  - Vector/Geometry: *Euclidean, Mahalanobis, Minkowski*
  - Distributions: *Kullback-Leibler, Earth-mover/Wasserstein, Cross-Entropy*
- Quality
  - Annotation coverage, label/class imbalance (rare-event?)
  - "Usefulness": *topic perplexity, (B/A) Information Criterion*
  - Inter-rater agreement: *Fleiss' **κ**, Kendall's **τ**, graph-based?*
- Importance
  - Information content: *Shannon Entropy, log-odds, lift, sum-TFIDF*
  - Centrality: *degree, betweenness, spectral (e.g. TextRank),*

EVALUATE: PRECISION & RECALL

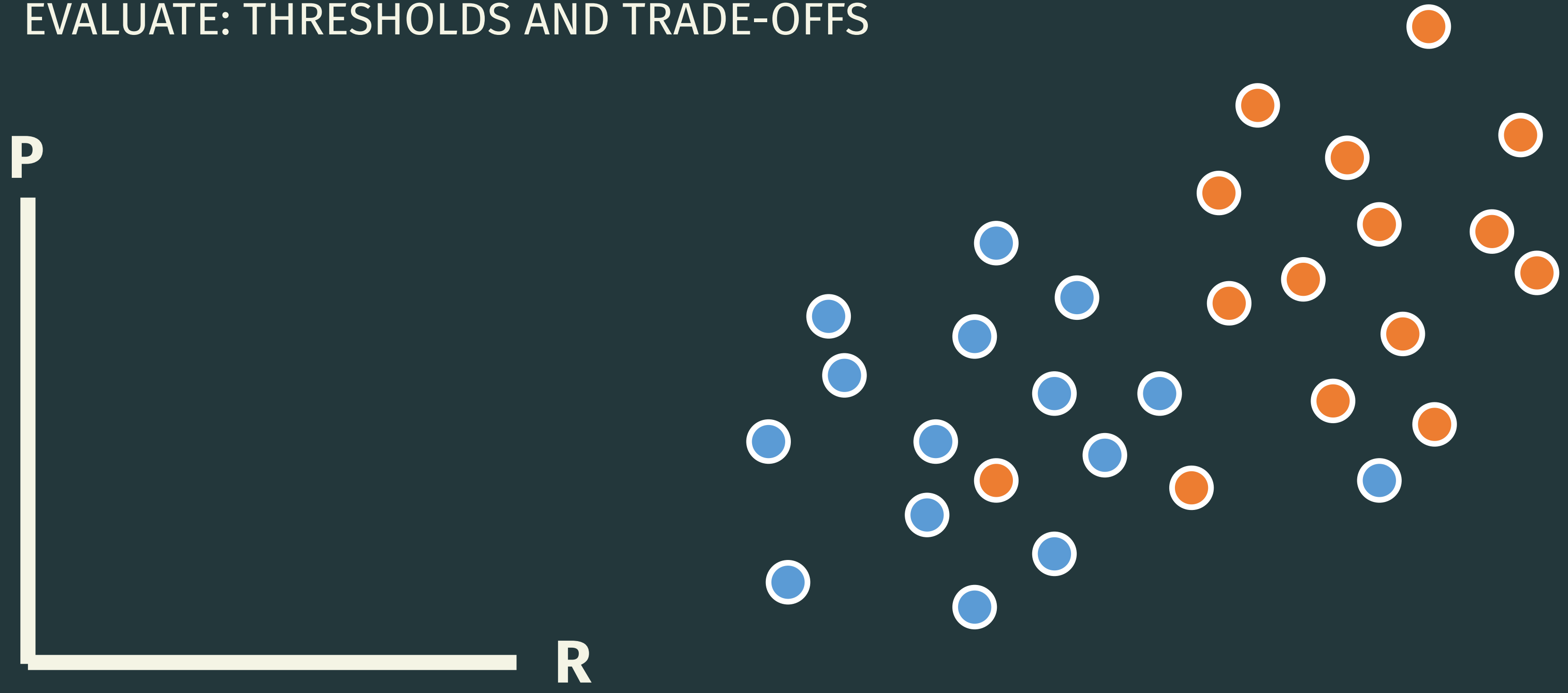
NLP often involves *multilabel* or *imbalanced* classification.

→ Accuracy is **unfair** or **overly optimistic**

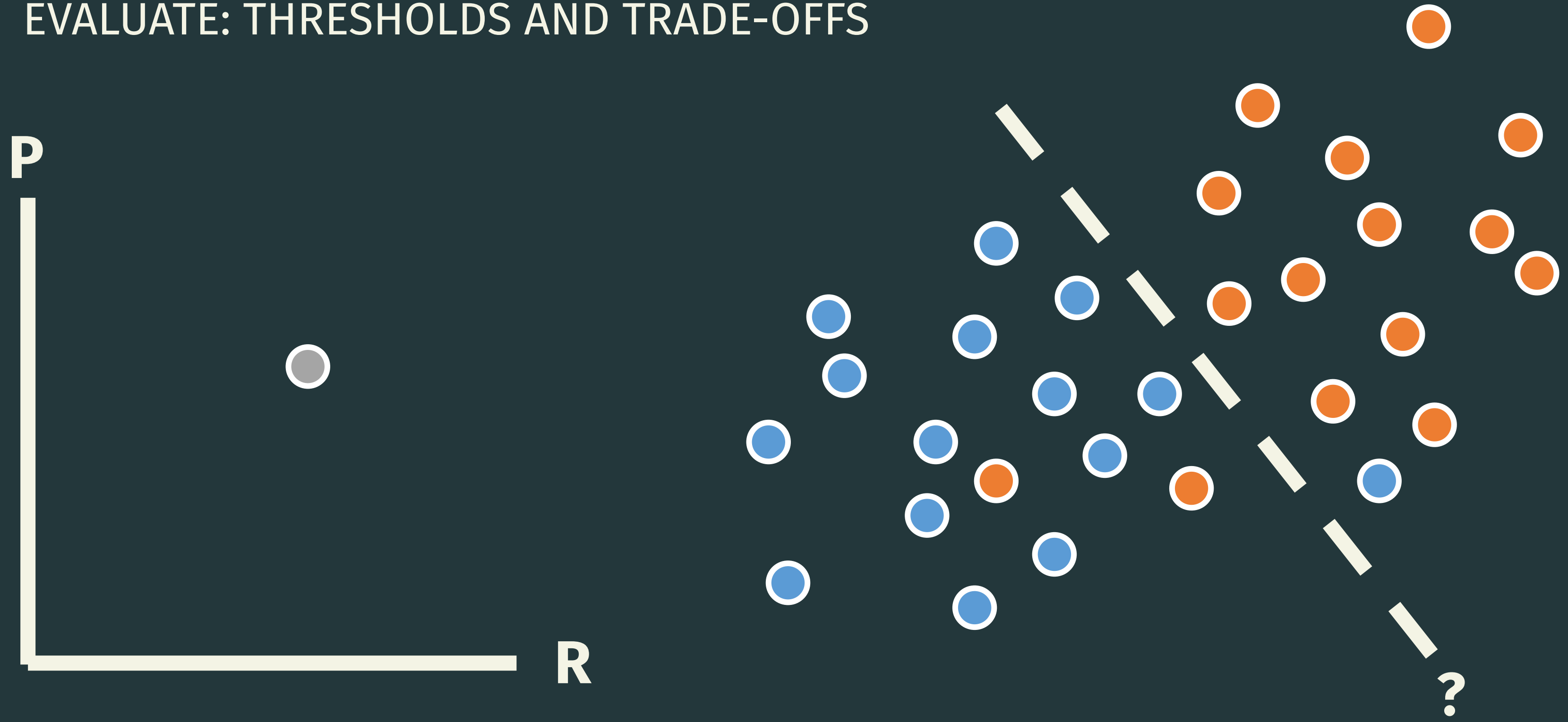
- Precision
  - Also *Positive Predictive Value (PPV):* [TP/(TP+FP)]
  - “Of things **predicted** X, how many **are** X?”
- Recall
  - Also *True Positive Rate* or *Sensitivity:* [TP/(TP+FN)]
  - “Of the things that **are** X, how many were **predicted** X?”
- F-Score
  - Harmonic mean of Precision & Recall:
  - Explicitly combines our preferences for the two
  - Parameter **β** (usually 1) : assign **β**-times more importance to Recall than precision.



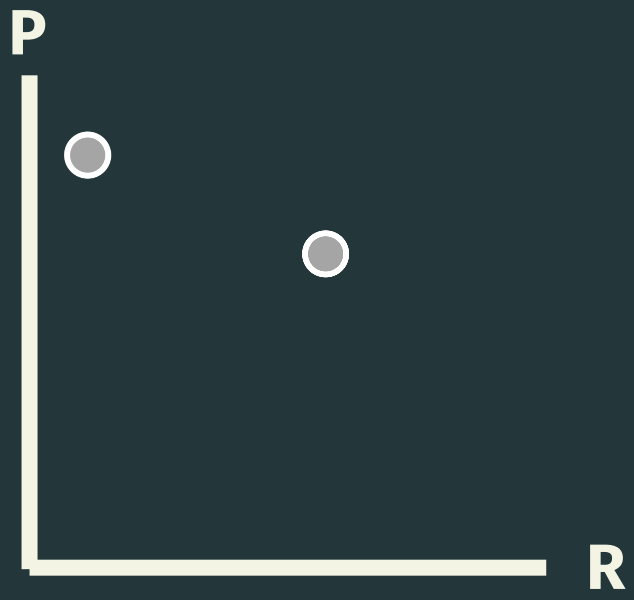
EVALUATE: THRESHOLDS AND TRADE-OFFS

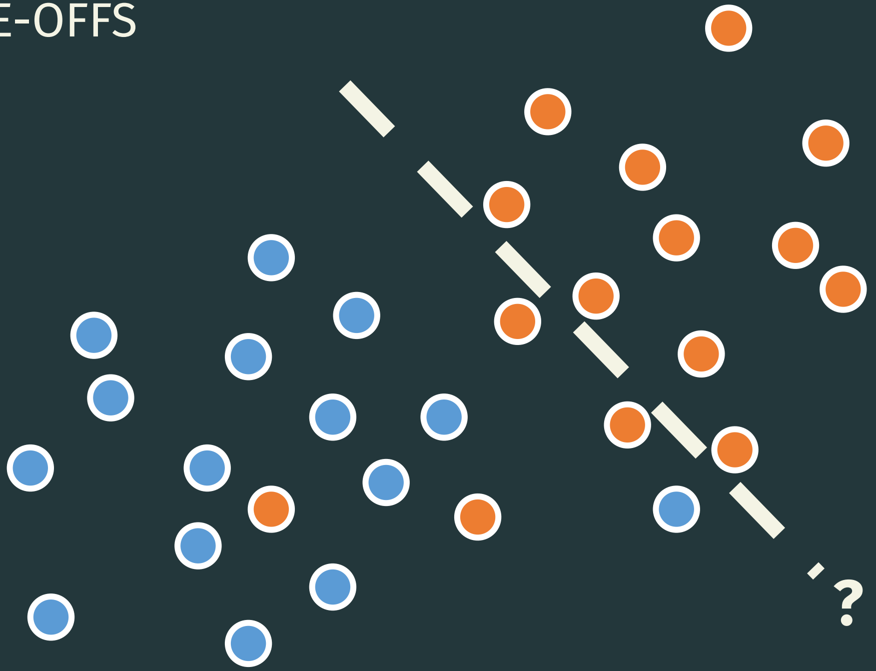


EVALUATE: THRESHOLDS AND TRADE-OFFS

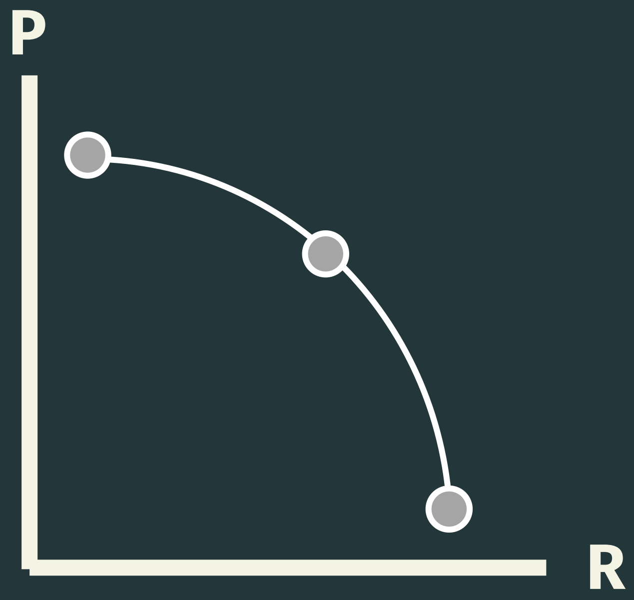


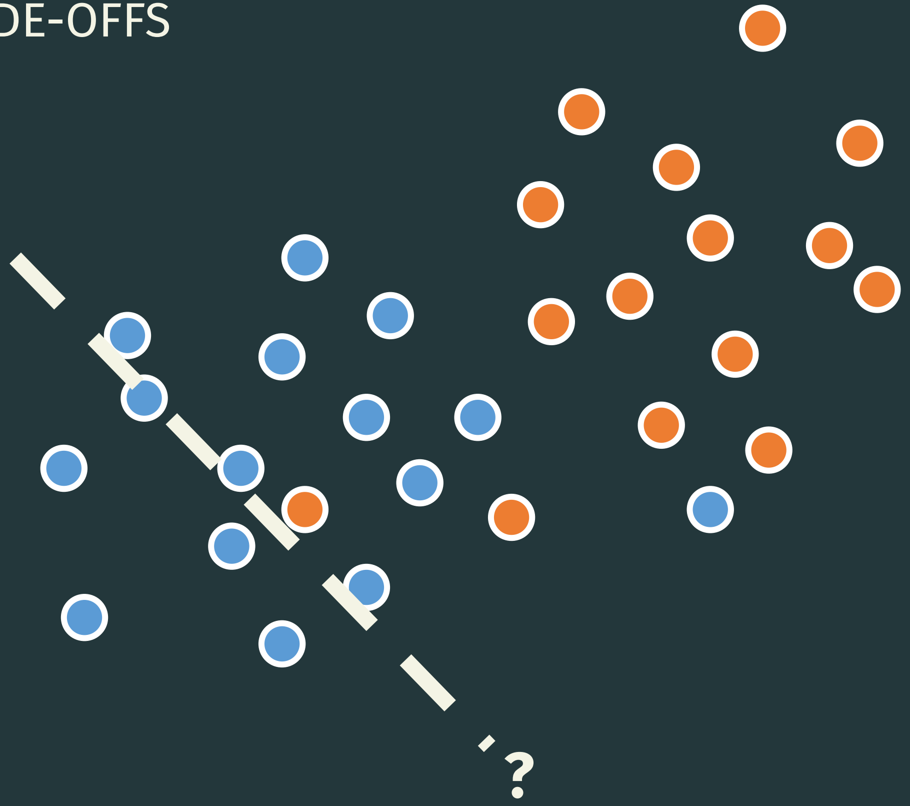
EVALUATE: THRESHOLDS AND TRADE-OFFS



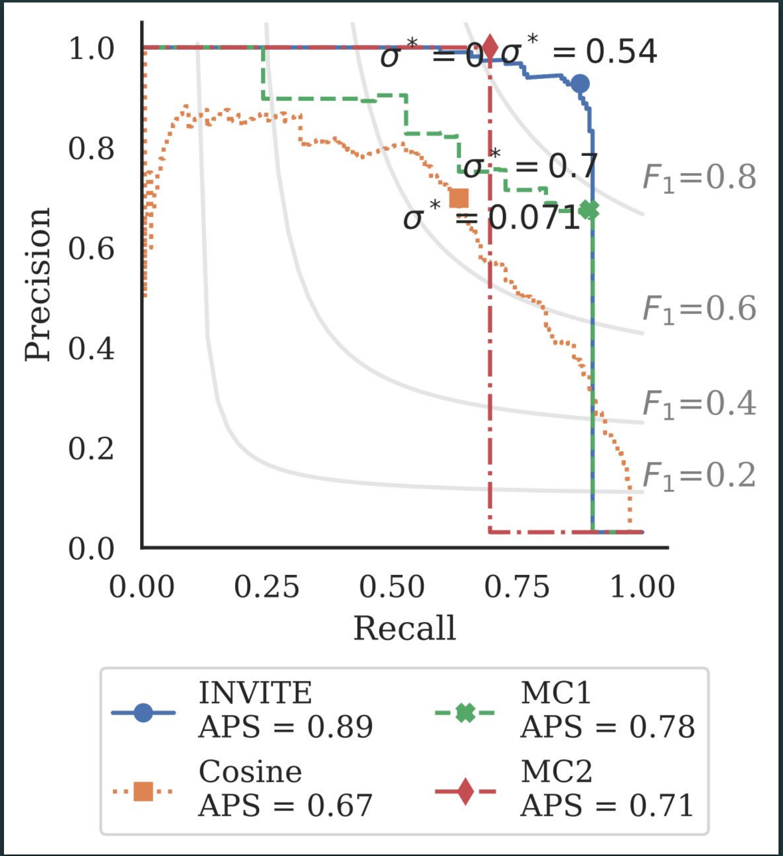


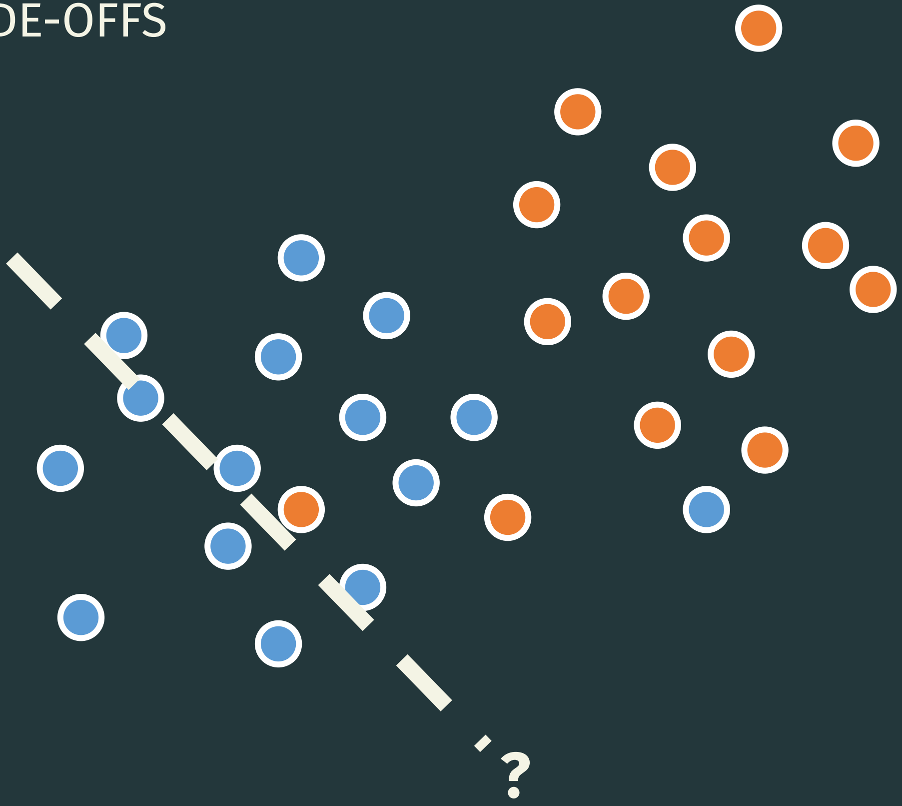
EVALUATE: THRESHOLDS AND TRADE-OFFS





EVALUATE: THRESHOLDS AND TRADE-OFFS





Sexton, T., and Fuge, M. (January 13, 2020). "Organizing Tagged Knowledge: Similarity Measures and Semantic Fluency in Structure Mining." ASME. *J. Mech. Des.* March 2020; 142(3): 031111. https://doi.org/10.1115/1.4045686

EVALUATE: SUMMARY

Do your **homework**

If there's something you want to measure, a metric may exist.

Metrics **evaluate**

Use fundamentals to design metrics that assess what matters.

Metrics **communicate**

Confusion is never the answer; strive for mutual understanding.

Remember that NLP is working on data *for humans, by humans.*

*Be **transparent** and **reproducible**.*

VALIDATION

The “open problem” of human-in-the-loop, domain-specific NLP

VALIDATION: PROBLEMS

So far we have glossed over some very common problems:

- Interpreting topic models can be fraught [1]
- Out-of-the-box tools are pre-trained on *very* different text
- There is not enough data to train custom models
- Too hard to hand-annotate the data we have
- No existing standard annotation to apply, no ontology we agree on
- Events of interest are far too rare (unclear if over-sampling applies)
- ...

In most Engineering Design and Reliability tasks, we *validate:*

> *Sanity checks, second opinions, processes for oversight and collaboration*

[1] *Chang, Jonathan, et al.*
“Reading tea leaves: How humans interpret topic models."
Neural information processing systems. Vol. 22. 2009.

VALIDATION: RE-ASSESSING “THE PIPELINE”

Reality is never as clean as “The Pipeline”.

> *“In practice, the line between input and output are not well defined. An analyst might use intermediary tasks and representations to enrich annotations and cascade into further tasks. A holistic approach to improving one component will inevitably improve the others; a stolid adherence to a given pipeline can prevent progress all-around.*
>
> *[...]*
>
> *By lowering barriers to entry for text analysis through the development of efficiency-boosting tools and a more human-centered annotation approach, engineers have a unique opportunity to simultaneously learn from other domains and improve on their processes. A new approach is needed to adapt NLP methods to industry use cases in a scalable and reproducible way.*[1]

→ View NLP as a socio-technical system rather than as an algorithmic pipeline.

[1]Brundage, Michael P., et al. "Technical language processing: Unlocking maintenance knowledge." *Manufacturing Letters* 27 (2020): 42-46.

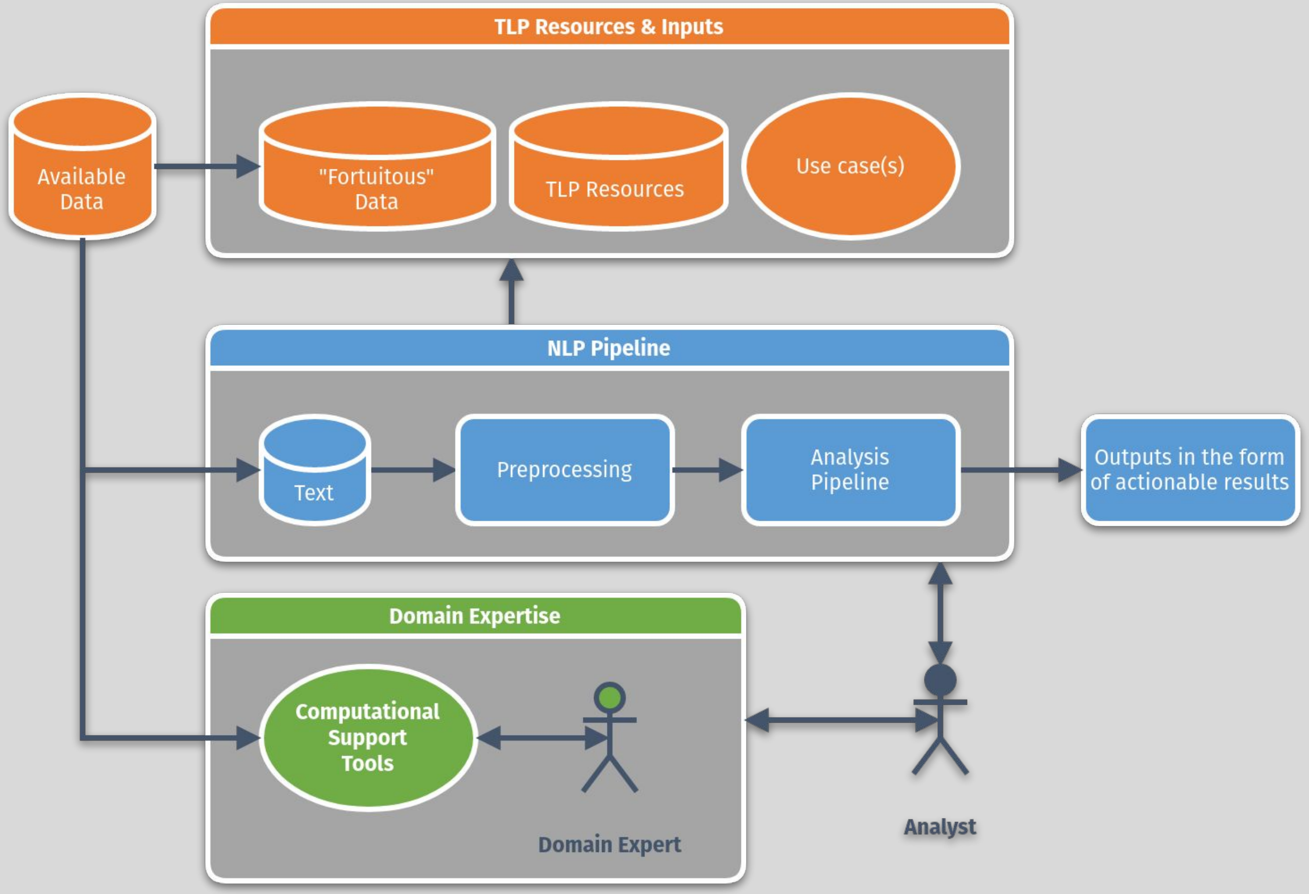
VALIDATION: TECHNICAL LANGUAGE PROCESSING

Enter ***Technical Language Processing***

- NLP Techniques do not *always* adapt well to engineering text
- Current NLP solutions need to be adapted ***correctly*** for use in technical domains
- TLP is a methodology to tailor NLP solutions to engineering text and industry use cases in a scalable and reproducible way

Adapting Natural Language Processing for Technical Text
Dima, Alden, et al.
Applied AI Letters: e33.
Image adapted from original

- How the TLP approach to meaning and generalization differs from NLP
- How data quantity and quality can be addressed
- Potential risks of *not* adapting NLP



VALIDATION: GET INVOLVED

Plan for Distributed Collaboration in the TLP CoI

I. GitHub Organization (just started): **TLP-CoI**
   A. Documentation - best practices for TLP, theory, etc
   B. Networking - curated list for state-of-the-practice: awesome-tlp
   C. Collaboration - base or forks for open tool repositories

II. Events:
   A. Past Workshop (slides):
   B. TLP-COI Slack Workspace - QR code →
   C. Other options? Webinars? Let us know!

THANK YOU

Thurston Sexton

thurston.sexton@nist.gov

U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

AI Enabling Technologies

Grooper and
Watson Content Analytics

June 29, 2021

Overview

What is Grooper?

Software that provides "*Thrilling Automation with Intelligent Document Processing*"*

Use Case: Extraction of data from operator licensing (OL) applications

Forms:
- NRC Form 396 (Certification of Medical Examination by Facility Licensee)
- NRC Form 398 (Personal Qualification Statement – Licensee)

Interfaces:
- Electronic Information Exchange (document ingestion)
- Reactor Program System (authoritative OL data source)

Grooper

NRC Grooper Features

1. Capture Tool

De-skew, Brighten, etc.

2. Image Processing

Optical Character Recognition (OCR)

5. Extraction

Parse and extract data, and write in XML schema

Other features used:

Optical Mark Recognition (OMR) – Recognizes checkmarks

Fuzzy Logic – Dictionary of defined values that can be OCRed or extracted based on a confidence threshold

Grooper

AI Grooper Features

Natural Language Processing and **Machine Learning** finds paragraphs, sentences, or other language elements in documents based on contextual meaning.

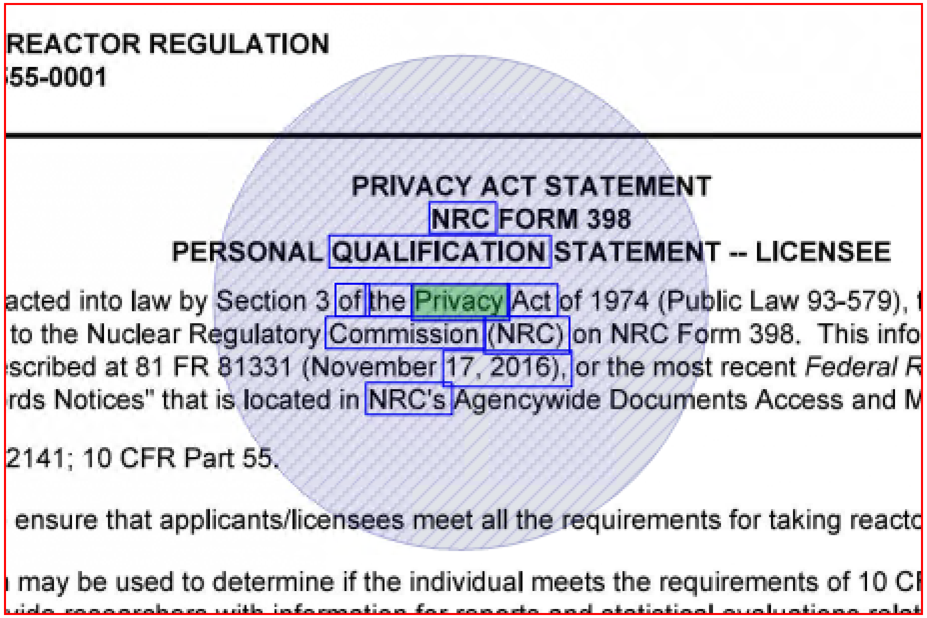
Use Case: Document sensitivity

Method:

- Manually review documents for sensitive keywords and identify true positives (in Grooper client)
- Start to train Grooper to contextually search around area of true positive
- Repeat with several document samples until properly trained

Grooper

AI Grooper Features Cont'd*


REACTOR REGULATION
55-0001

PRIVACY ACT STATEMENT
NRC FORM 398
PERSONAL QUALIFICATION STATEMENT -- LICENSEE

acted into law by Section 3 of the Privacy Act of 1974 (Public Law 93-579), t
to the Nuclear Regulatory Commission (NRC) on NRC Form 398. This info
scribed at 81 FR 81331 (November 17, 2016), or the most recent *Federal R*
rds Notices" that is located in NRC's Agencywide Documents Access and M

2141; 10 CFR Part 55.

ensure that applicants/licensees meet all the requirements for taking reacto

may be used to determine if the individual meets the requirements of 10 CF

Context Scope

Type: *ContextScopeEnum*, ***Default:*** *Zonal*

Determines the scope of context feature extraction. Can be one of the following values:

- **Zonal** - Context features will be extracted from one or more zones, specified relative to the data value.
- **Flow** - Context will include a limited number matching features before and/or after the data value in the text flow.
- **Self** - Context will include all matching features which occur inside of or overlap with the data value.
- **Nearest** - Context will include a limited number of features which are closest to the data value.

Overview

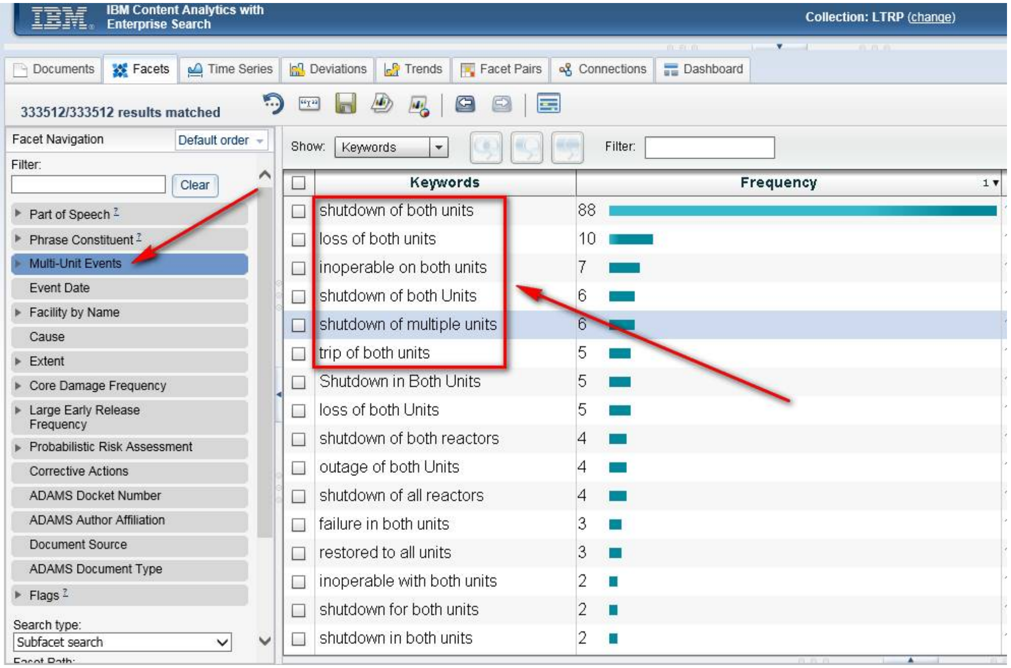
What is Watson Content Analytics?

Software that extrapolates business information from large collections of documents and uses natural language processing to uncover meaningful business insights.

Use Case: RES - Identify Event Reports that included an outage of two or more units

NLP Method:

- Define noun/verb combinations and NLP automatically contrives derivations of those combinations



U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

References

References (Indicated by an *)

BIS, Inc. (2020-2021). AI-Powered Data Integration. Retrieved from https://www.bisok.com/.

AI Enabling Technologies

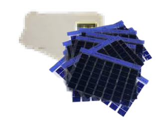
Grooper and
Digitizing Success

June 29, 2021

Digitization Project

- **Digitize Key Docketed Information**
  - Making licensing and design basis information readily available to staff streamline the review process
  - Expand public access to materials
  - Comply with federal records management mandates (e.g., M19-21)
  - Reduce storage cost

- **NUDOCS microform** (1979-1999) – 110K microfiche and 88K aperture cards consist of 2.3M documents or 43M images
  - **109,424** (100%) microfiche, **87,929** aperture cards digitized
  - **43,009,225** (100%) images of 43M fiche/aperture scanned
  - Over **2,355,157** PDFs generated



- **AEC Paper** (pre 1978) – 1,095 boxes of paper records which consist of 205K documents or 3.2M images
  - **191** boxes, **332,879** pages digitized (COVID-19)
  - **13,619** PDFs generated

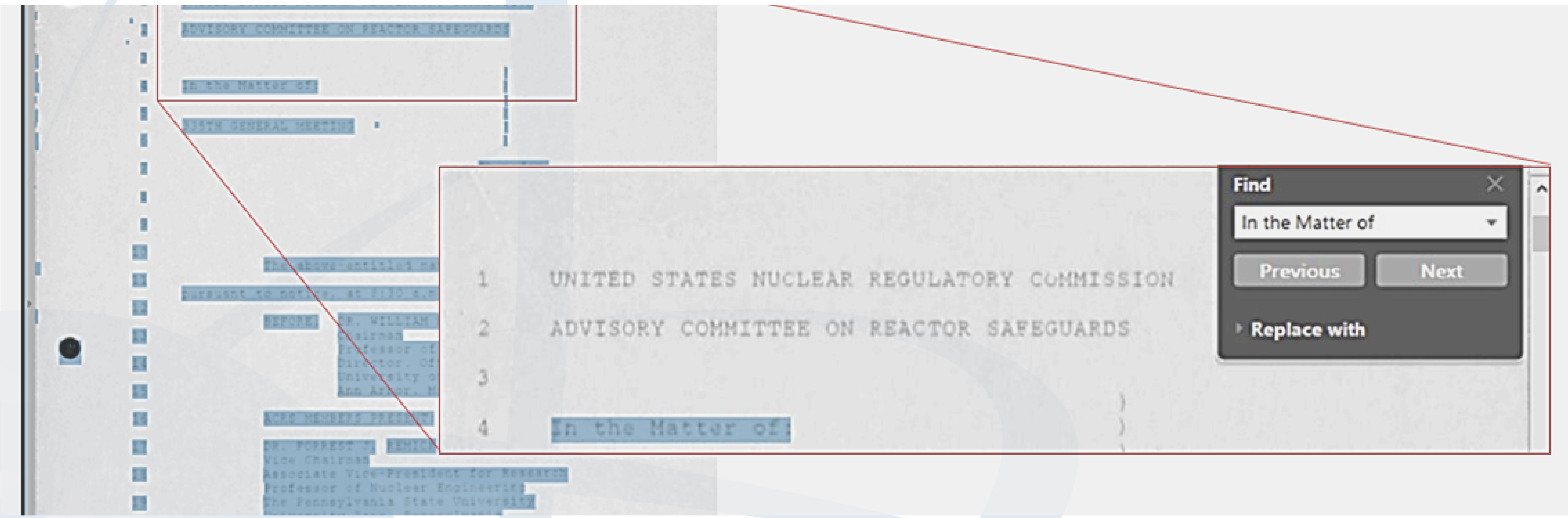
- **Components**
  - Mekel Mach 7
  - Grooper software

U.S.NRC

United States Nuclear Regulatory Commission

Protecting People and the Environment

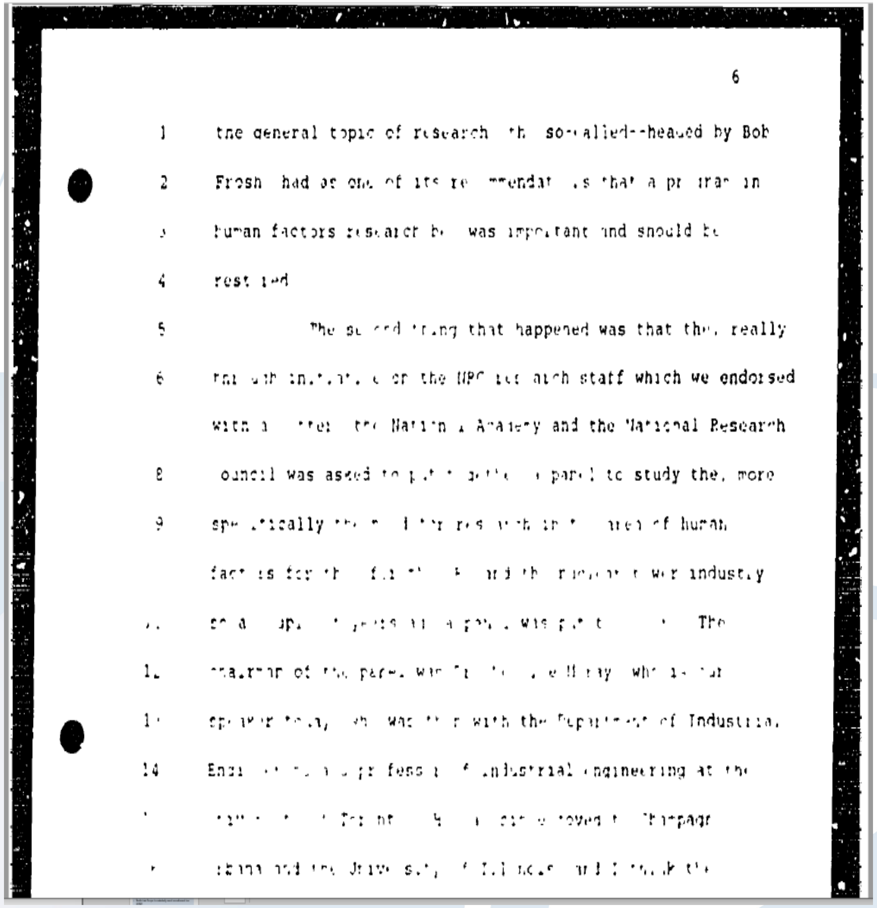
DIGITIZATION PROJECT



New Searchable Image

Processed with Artificial Intelligence

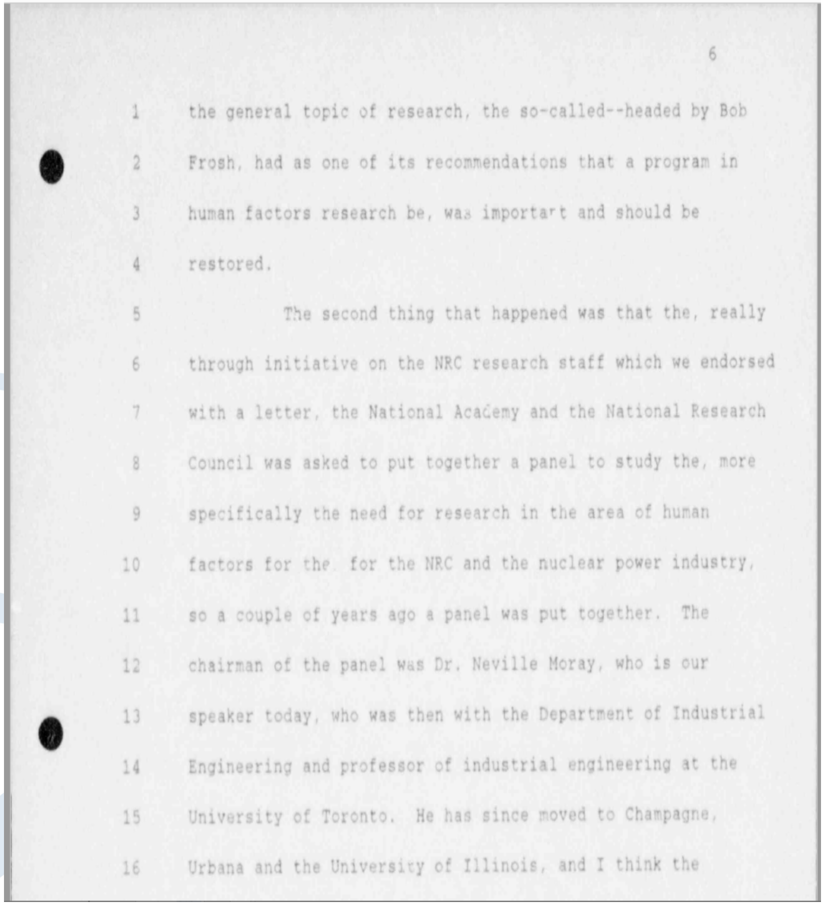
Digitization Project



6

the general topic of research [illegible] so-called--headed by Bob Frosh had as one of its [illegible] that a program in human factors research [illegible] was important and should be [illegible]

The [illegible] thing that happened was that the[illegible] really [illegible] on the [illegible] staff which we endorsed with [illegible] and the National Research [illegible] was asked to [illegible] panel to study the, more [illegible] of human [illegible] industry [illegible] The [illegible] speaker [illegible] with the Department of Industrial [illegible] industrial engineering at the [illegible]

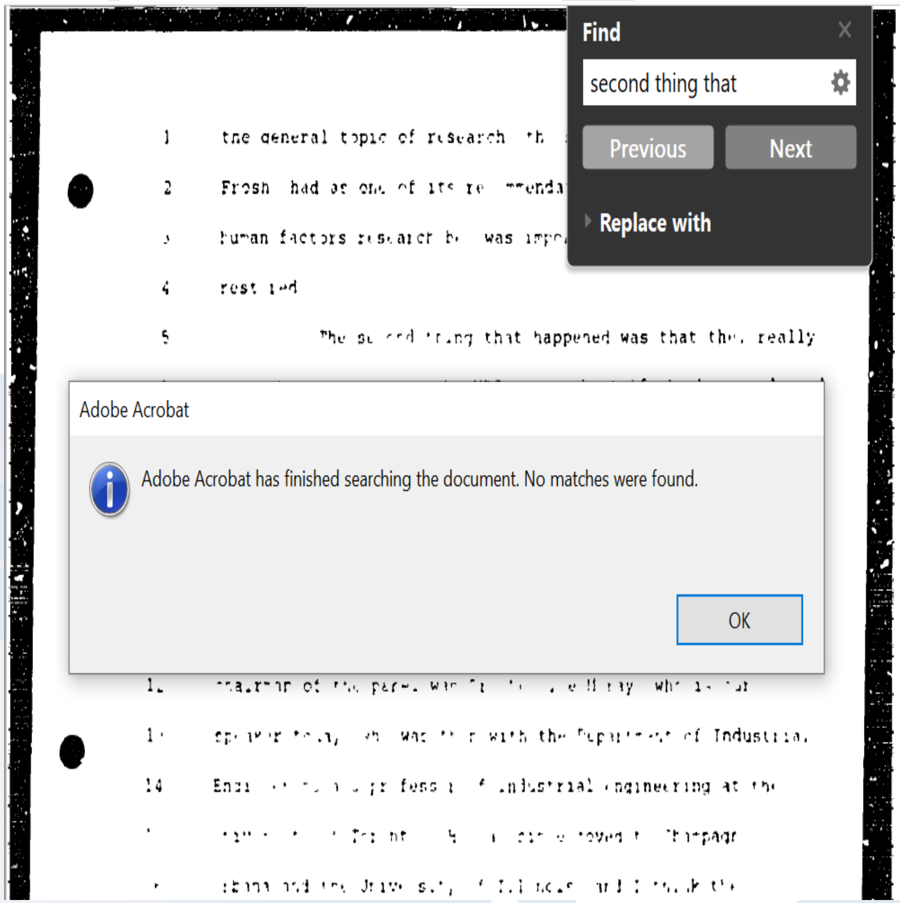


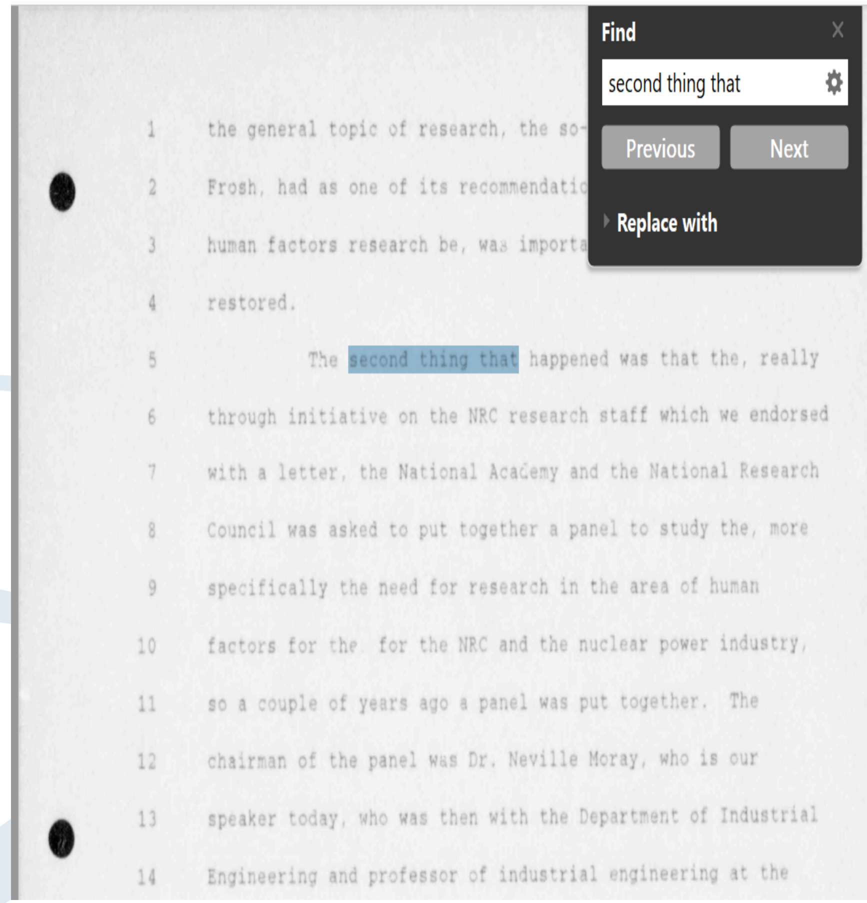
6

the general topic of research, the so-called--headed by Bob Frosh, had as one of its recommendations that a program in human factors research be, was important and should be restored.

The second thing that happened was that the, really through initiative on the NRC research staff which we endorsed with a letter, the National Academy and the National Research Council was asked to put together a panel to study the, more specifically the need for research in the area of human factors for the, for the NRC and the nuclear power industry, so a couple of years ago a panel was put together. The chairman of the panel was Dr. Neville Moray, who is our speaker today, who was then with the Department of Industrial Engineering and professor of industrial engineering at the University of Toronto. He has since moved to Champagne, Urbana and the University of Illinois, and I think the

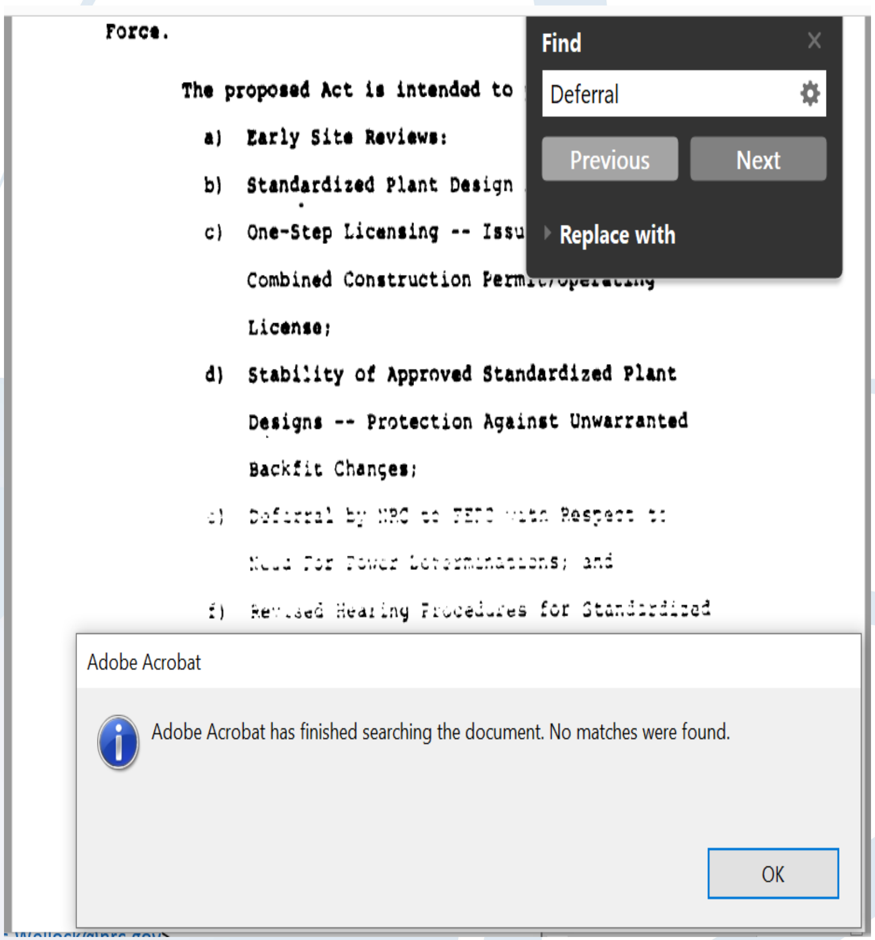
Digitization Project

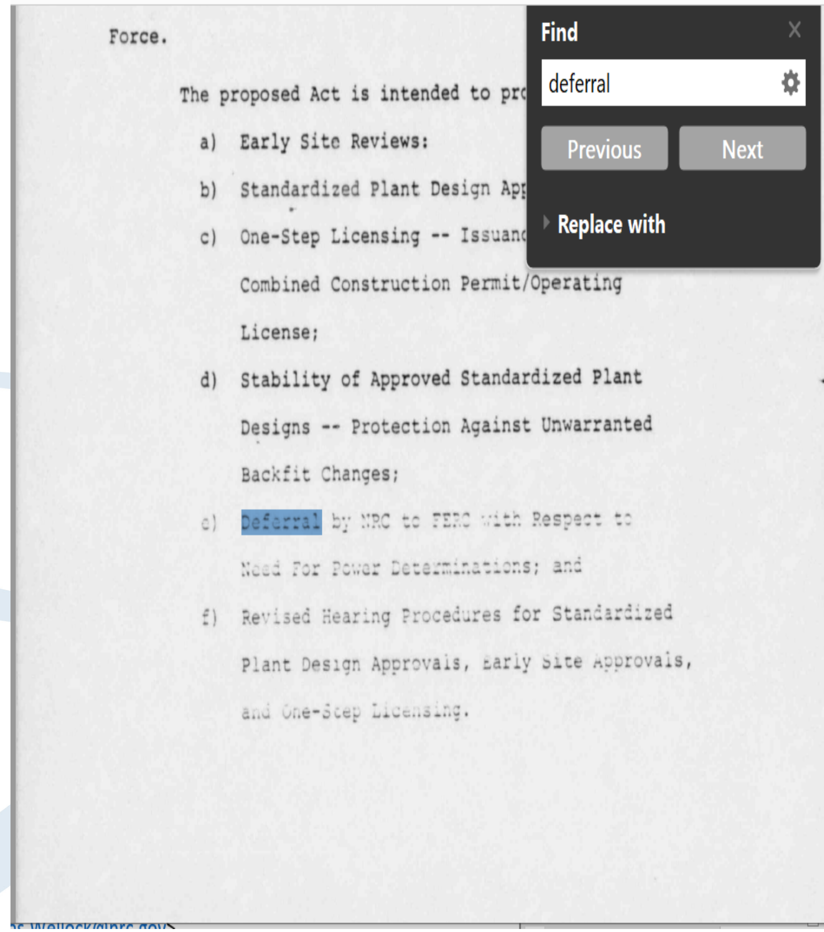




U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

DIGITIZATION PROJECT





AI Enabling Technologies

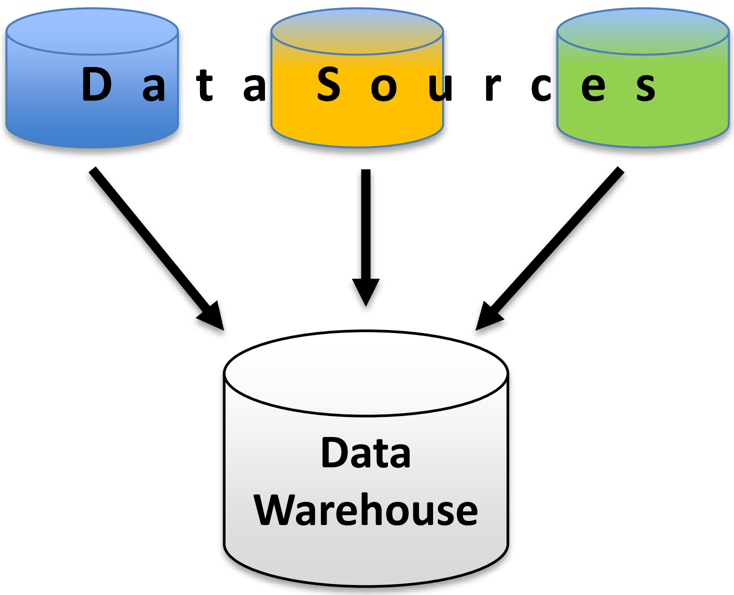
Enterprise Data Warehouse

June 29, 2021

Overview

What is the Enterprise Data Warehouse (EDW)?

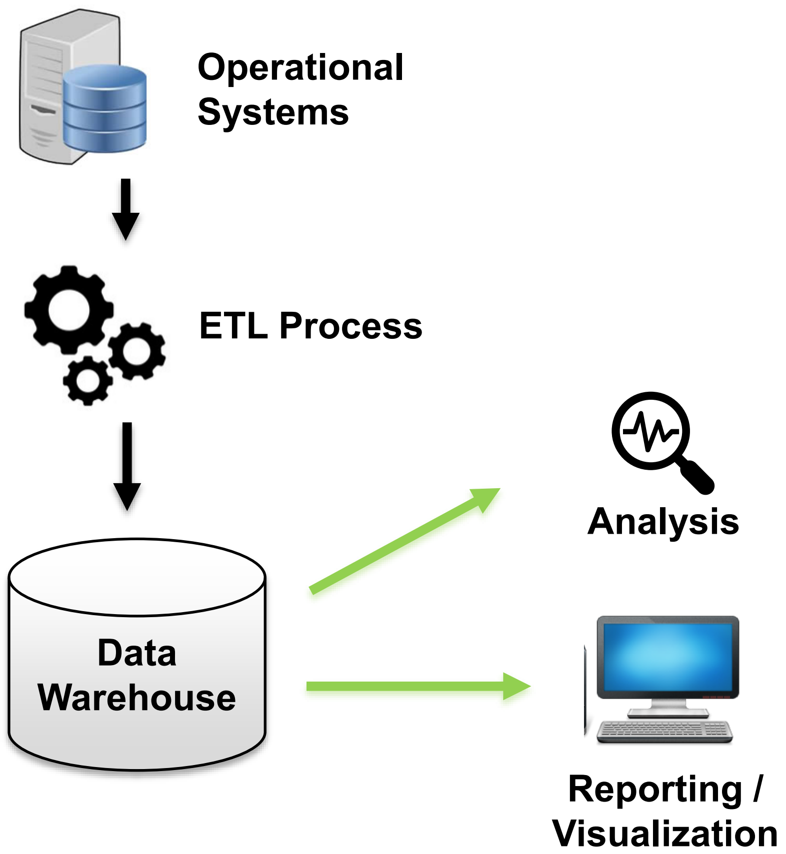
- Central repository of integrated data from NRC's authoritative systems
- Purpose is to provide timely, accurate data from authoritative data sources to be used for reporting and data analytics



Overview

Enterprise Data Warehouse Architecture

- Interfacing systems – NRC's operational systems, authoritative data sources
- ETL Process - The Enterprise Data Warehouse extracts data from authoritative sources, transforms it in a staging area then loads it into the EDW on a scheduled interval
- EDW - Database that stores the data to be used for reporting and data analytics



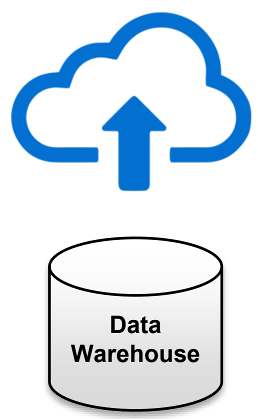
Overview

Benefits of the Data Warehouse

- Improved Reporting Performance and Efficiency
- Improved Data Quality and Consistency
- Empowers Users to Gain Data Insights

Azure Cloud

- Data Warehouse migration to Azure Cloud



- **Azure Analysis Services**
- **Azure Cognitive Services**
- **Azure Machine Learning**

NRC AI Workshop

Event Management Response Tool (EMRT) Project
Relief Request Index Project

Nick Mohr, Senior Technical Leader,
EPRI Welding and Repair Technology Center (WRTC)

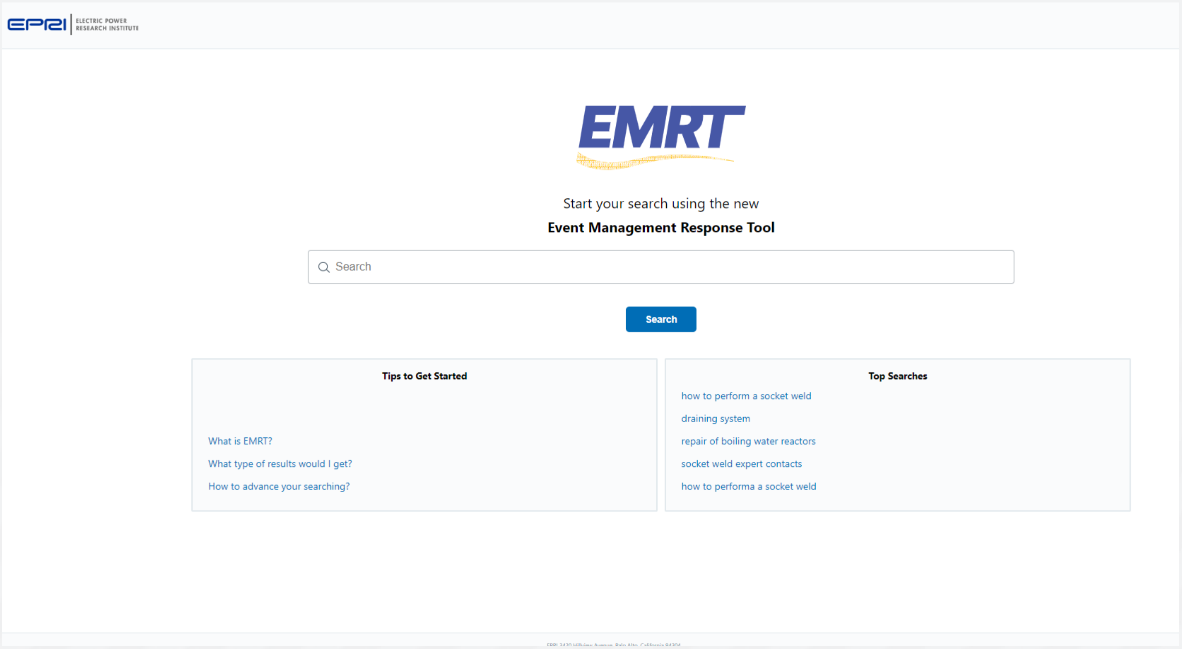
June 29, 2021

Event Management Response Tool (EMRT)

Nick Mohr, Senior Technical Leader, EPRI
Kriti Dhaubhadel, Sparkcognition
Abubaker Sheikh, Sparkcognition
Prateek Jindal, Sparkcognition
Chris Taylor, Sparkcognition
Bryan Corralejo, Sparkcognition
Jaidev Amrite, Sparkcognition

What is the Event Management Response Tool (EMRT)

- Single location to consolidate various data sources for searching and correlation
  - Uses machine learning to refine and make future searches better
- Ingests various file formats (Excel, PDF, PowerPoint, etc.) to make unstructured data structured
- Allows previews of relevant locations within the document to ensure downloading is valuable



Structured Data (e.g. Excel, Access, etc.)
Unstructured Data (Word, PDF, PowerPoint, etc.)

Purpose and Objectives

- Goal to increase productivity by:
  - Reduction of time associated with finding the needed **research products**
    - Display the most relevant information based on a member search within research products
  - Reduction of time associated with finding **Code and Regulatory information** (e.g. regulatory submittals, content within Nuclear Regulatory Research, etc.)
  - Reduction of time associated with find **operating experience and lessons learned** from other EPRI members related to event

EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Value/Objective:
Provide EPRI members the needed information to make informed decisions in one location in reduced time

Event Management Response Tool (EMRT)

EPRI Research

Code and Regulatory

Operating Experience

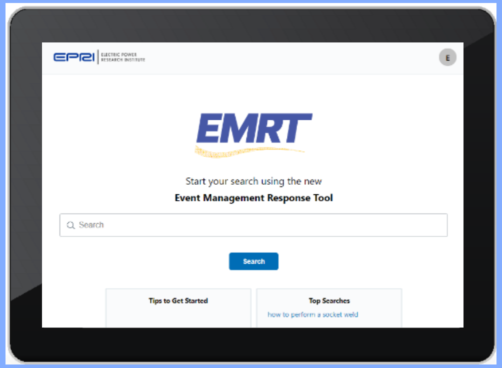
Member Input

Event Management Response Tool

OBJECTIVE: Provide members needed info in one location to make informed decisions in reduced time

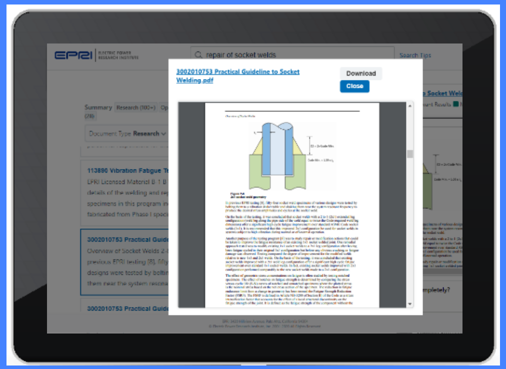
EMRT: Natural Language Processing & Access Full Data Library

Input

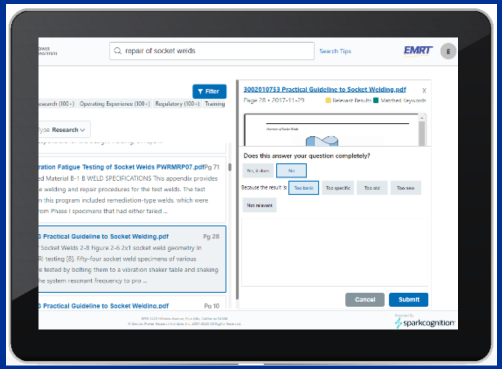


Typeahead search based on prior search terms

Display of Output



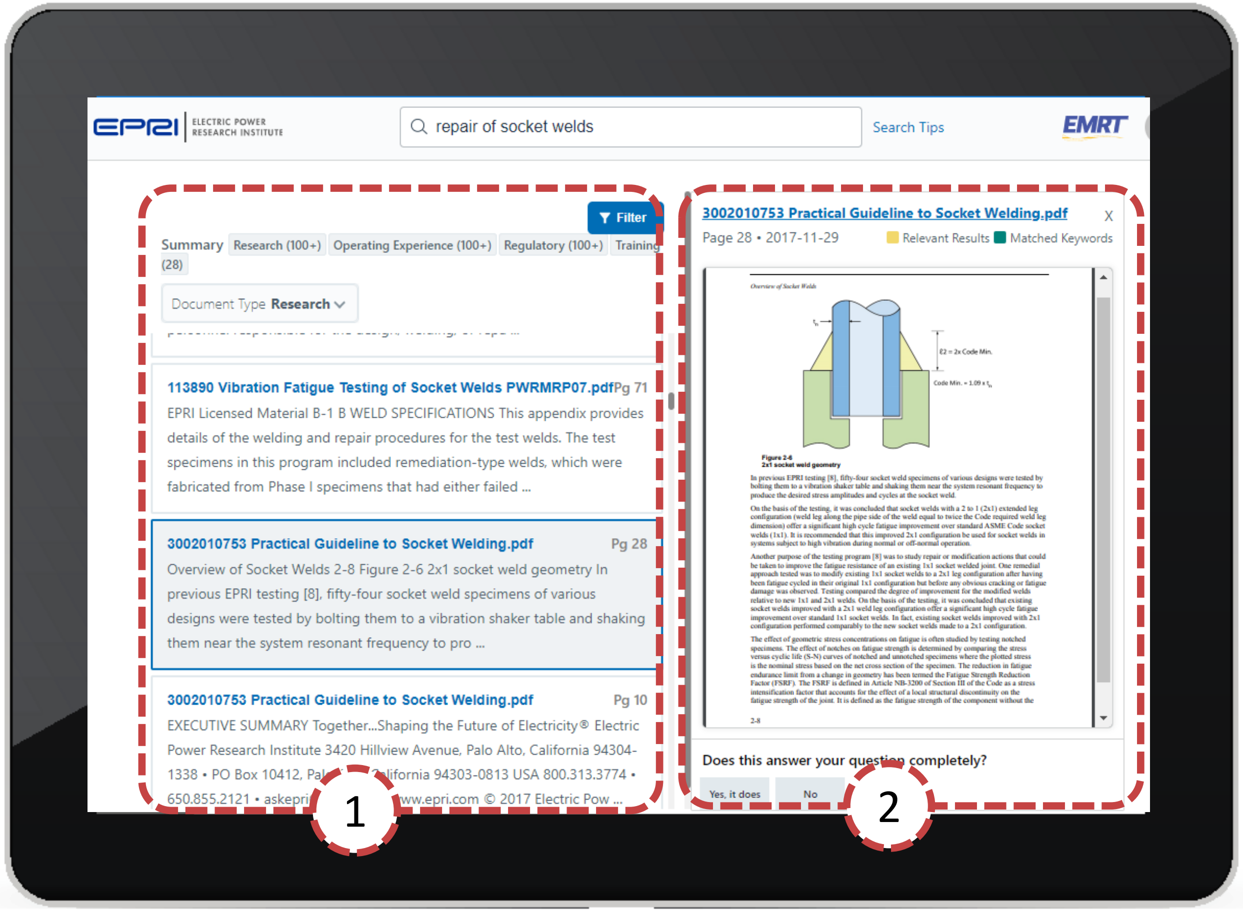
Feedback Loop

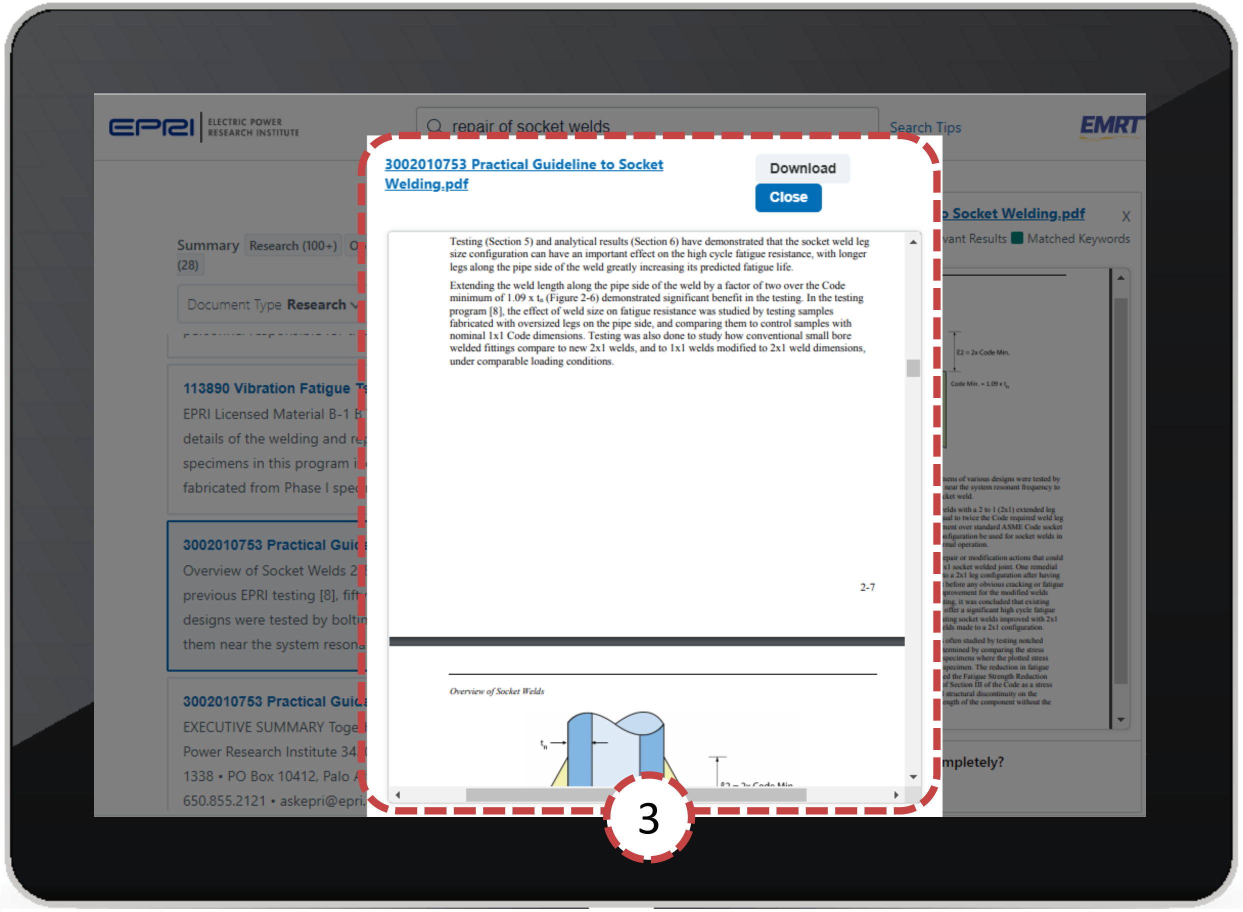


Machine Learning & Refinement of Output

Current input is text string but we would like to also use other input methods in the future

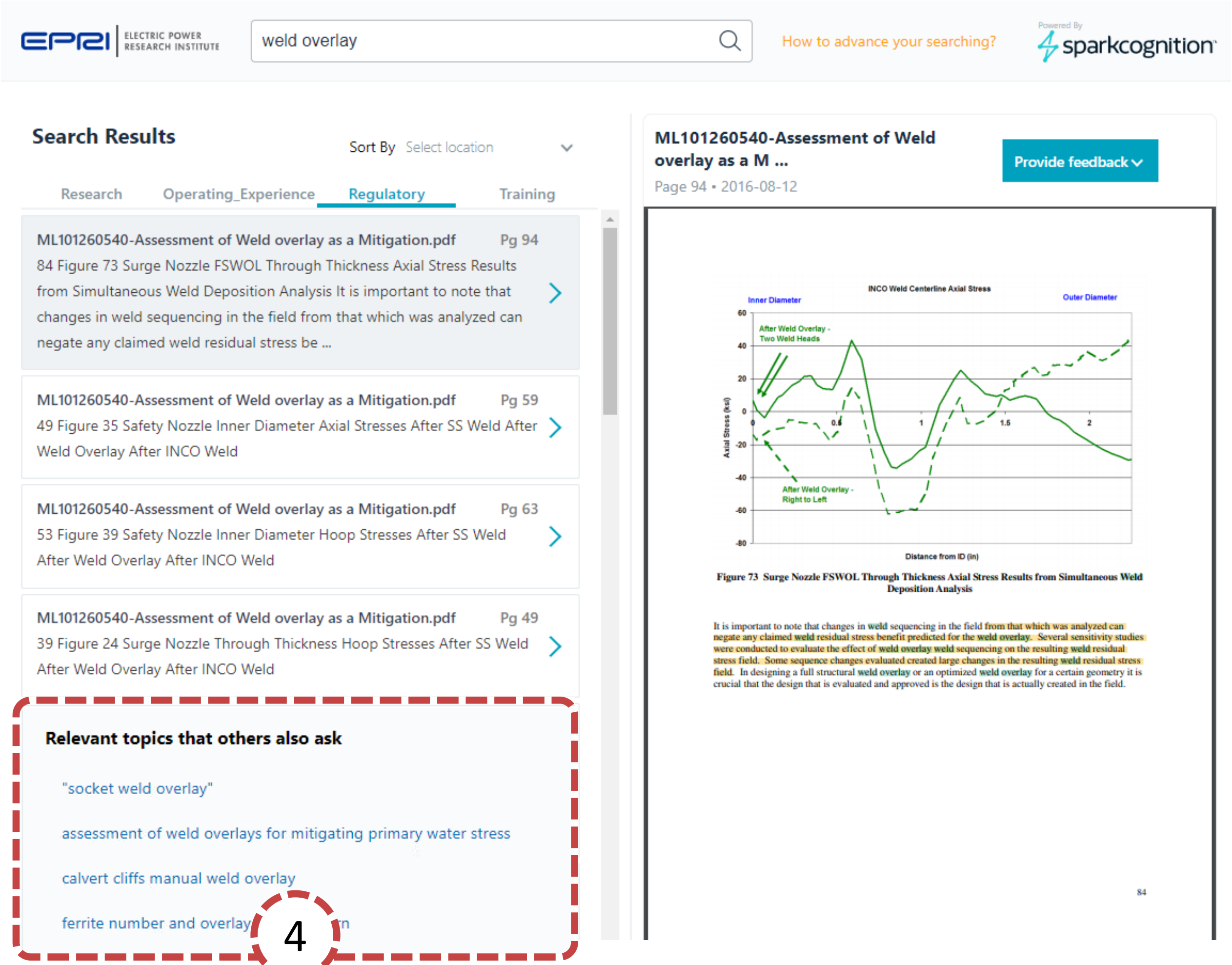
EMRT Search Results – 3 Locations Display Content



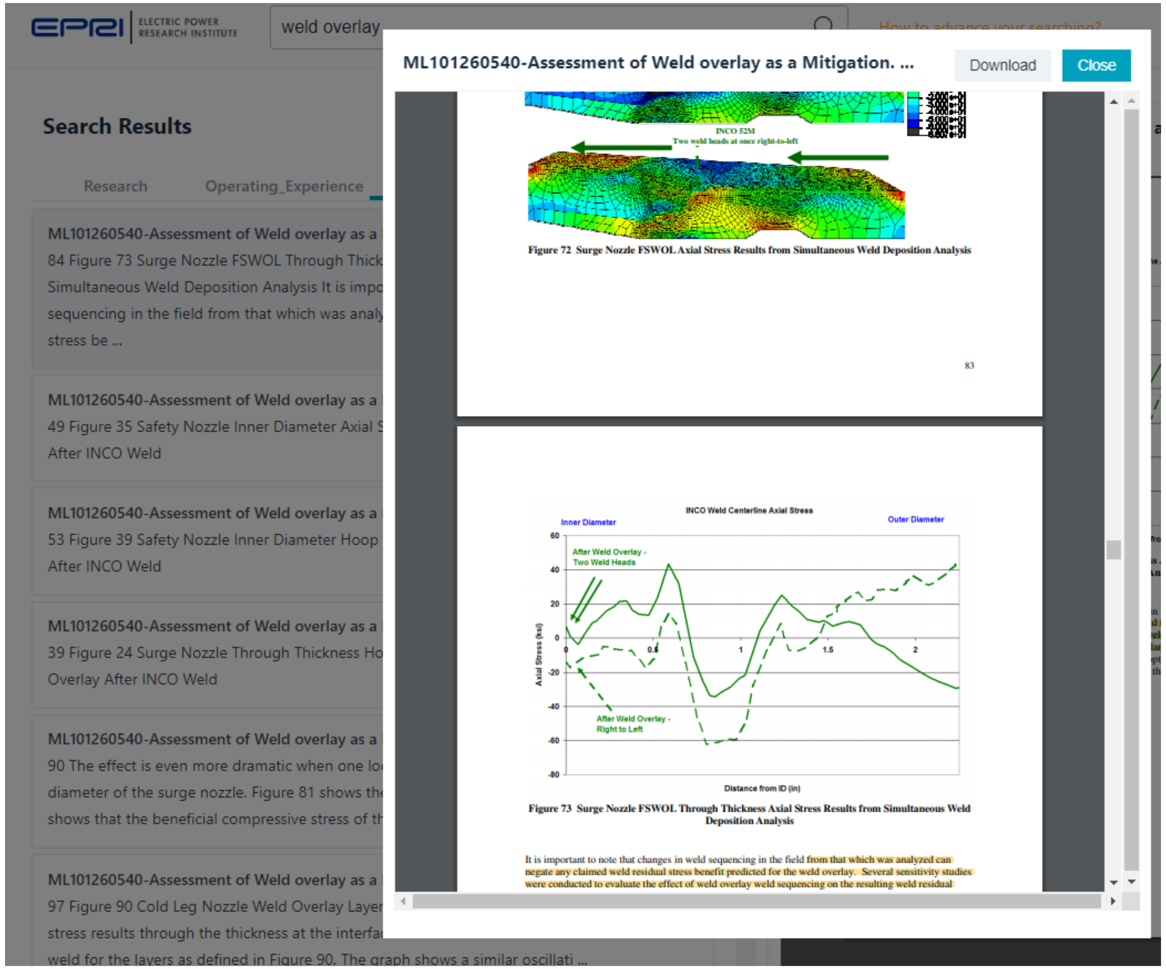


1) Tabulated search results
2) Preview of user-selected search result within the respective document
3) Preview of user-selected search result with ability to scroll to different pages within the respective document

EMRT-Regulatory Information Example



4) Suggested relevant topics on initial search screen and for specific searches



Ability to Scroll within document to determine if it is valuable to download. Download button takes user to NRC site.

EMRT-Regulatory Information-NRC ADAMS Document Library

- Regulatory information is necessary to make decisions
- NRC ADAMS contains a number of **publicly** available documents (subset shown)
- Currently, users search ADAMS but finding data can be difficult
- We can use NLP and machine learning if we ingest and extract the data from these documents
- This would help members search this information more effectively
- Use of existing NRC Application Programming Interface (API) permits filtering by document type

ADAMS Document Types

Order Suspending License
Order, Confirmatory
Organization Chart
Part 21 Correspondence
Performance Indicator
Performance Plan
Performance Planning and Appraisal (SES)
Periodic Monitoring Report (Radiological/Environmental)
Photograph
Planning Call
Plant Issues Matrix
Plant Performance Review
Plant Status Report
Policy and Program Guidance
Policy Statement
Post-Shutdown Decommissioning Activities Report
Pre-decisional Contract Action
Preliminary Safety Analysis Report (PSAR)
Press Release
Privacy Impact Assessment
Privacy Threshold Analysis
Probabilistic Risk Assessment
Program Review
Project Manager (PM) List
Project Plans and Schedules
Project Requirement Document
Proprietary Information Review
Quality Assurance Program
Radiation Overexposure Reports
Records Retention and Disposal Authorization
Records Transmittal and Receipt, SF Form 135
Reference Safety Analysis Report
Reference Safety Analysis Report, Amendment
Regulatory Analysis
Regulatory Guidance
Regulatory Guide
Regulatory Guide, Draft
Report of Proposed Activities in Non-Agreement States, NRC Form 241
Report, Administrative
Report, Miscellaneous
Report, Technical
Request for Access Authorization
Request for Additional Information (RAI)
Request for OMB Review
Request for Procurement Action (RFPA), NRC Form 400
Request for Review of OMB Reporting Requirements
RES Office Letter
Research Information Letter (RIL)
Resume
Reviewer Comments on Conference/Symposium/Workshop Pap
Route Approval Letter to Licensee
Routine Status Report (Recurring Weekly/Monthly)
Rulemaking- Final Rule
Rulemaking- Proposed Rule
Rulemaking-Authority Statement for EDO Signature
Rulemaking-Comment
Rulemaking-Environmental Assessment
Rulemaking-Environmental Impact Statement
Rulemaking-Plan
Rulemaking-Regulatory Analysis
Rulemaking-Regulatory Plan
Safeguard Incident Report
Safeguards Advisory
Safety and Compliance Inspection Record, NRC Form 591
Safety Evaluation
Safety Evaluation Report
Safety Evaluation Report, Draft
Schedule and Calendars
Security Form-Report of Security Infraction, NRC Form 183
Security Form-Security Incident Report, NRC Form 135
Security Frequently Asked Question (SFAQ)
Security Incidence Report
Security Plan
Security Program
Senior Management Meeting (SMM) Results Letter
Significant Event Report
Site Access Letter
Site Characterization Plan
Site Redress Plan
Site Safety Analysis Report (SSAR)
Slides and Viewgraphs
Social Media-Photograph
Social Media-Video Recording
Software Control Documentation
Software Documentation
Space Management
Space Policy
Special Nuclear Material Physical Inventory Summary Report

EMRT-Regulatory Information (focus NRC ADAMS)

NRC API

Use "Document Types" to focus on desired documents

Reference Safety Analysis Report
Reference Safety Analysis Report, Amendment
Regulatory Analysis
Regulatory Guidance
Regulatory Guide
Regulatory Guide, Draft
Report of Proposed Activities in Non-Agreement States, NRC Form 241
Report, Administrative
Report, Miscellaneous
Report, Technical
Request for Access Authorization
Request for Additional Information (RAI)
Request for OMB Review
Request for Procurement Action (RFPA), NRC Form 400
Request for Review of OMB Reporting Requirements
RES Office Letter
Research Information Letter (RIL)
Resume
Reviewer Comments on Conference/Symposium/Workshop Paper
Route Approval Letter to Licensee
Routine Status Report (Recurring Weekly/Monthly)
Rulemaking- Final Rule
Rulemaking- Proposed Rule

example selection (yellow highlighting)

Metadata

erty	XML Property Tag	Type
	MimeType	String
	EstimatedPageCount	Integer
;Number	CaseReferenceNumber	String
	ContentSize	Integer
on	AuthorAffiliation	String
	Keyword	String
e	DocumentDate	Date
r	LicenseNumber	string
r	DocketNumber	string
ber	AccessionNumber	string
er	PackageNumber	String
RS	PublishDatePARS	Date
	DocumentTitle	String
ortNumber	DocumentReportNumber	String
e	DocumentType	String
	AuthorName	String
	CompoundDocumentState	Boolean
iation	AddresseeAffiliation	String
ne	AddresseeName	String
	URI	URI
		String
		String

PDF Documents

Code and Regulatory

Project Overview-High Level

2020

- Prototype was developed with small subset of information
- Alpha Version was completed late 2020 incorporating a large set of data and EPRI member and personnel feedback and suggestions

2021

- Beta Version is currently being developed that will include larger set of information (EPRI Nuclear research, EPRI OE (meeting materials, surveys, etc.), NRC ADAMS data)

2022

- Incorporate feedback from users and consider other sources of Operating Experience, etc.

Relief Request Index Project

Craig Harrington, Technical Executive, EPRI

Nick Mohr, Senior Technical Leader, EPRI

Jacqueline Espinoza, Beyond the Arc

Steven Ramirez, Beyond the Arc

2020-2021: Relief Request Index-Proof Of Concept

Research Question:

Can we apply modern text mining and natural language processing techniques to curate a body of knowledge that would be helpful to plant engineers who are addressing welding repairs and material reliability situations?

NRC ADAMS is a large source of valuable information… but can be difficult to find desired information.

Value:

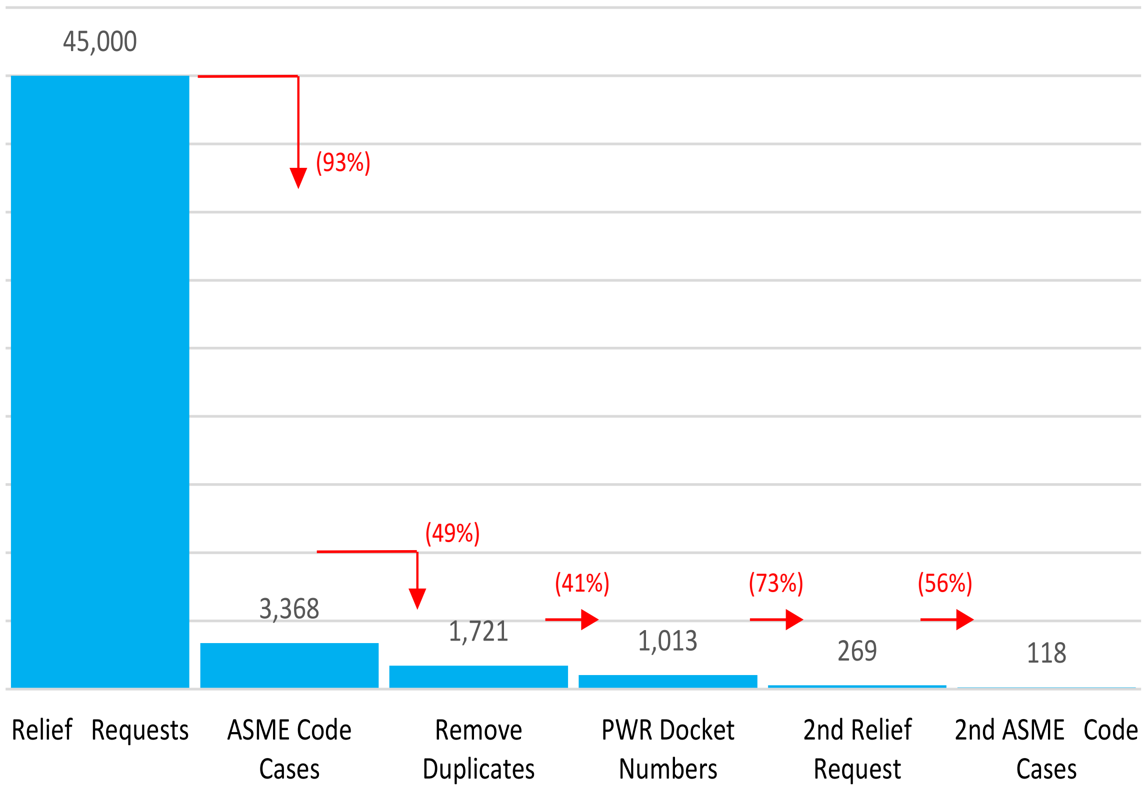
- **Reduce time spent finding complete series of request for alternatives "relief requests"**
- **The curated index assists users in understanding:**
  - **Where code cases have been used**
  - **Any potential conditions that should be addressed when a similar request is being submitted**
  - **Identify new trends**

Background

- EPRI decided to explore a proof of concept in 2021 using subset of desired code cases
- Index to filter by these topics:
  - ASME Code Case Number
  - Systems / Assets
  - Relief Requests for Inspection
  - Relief Requests for Repair
  - Plant Name
  - Operator

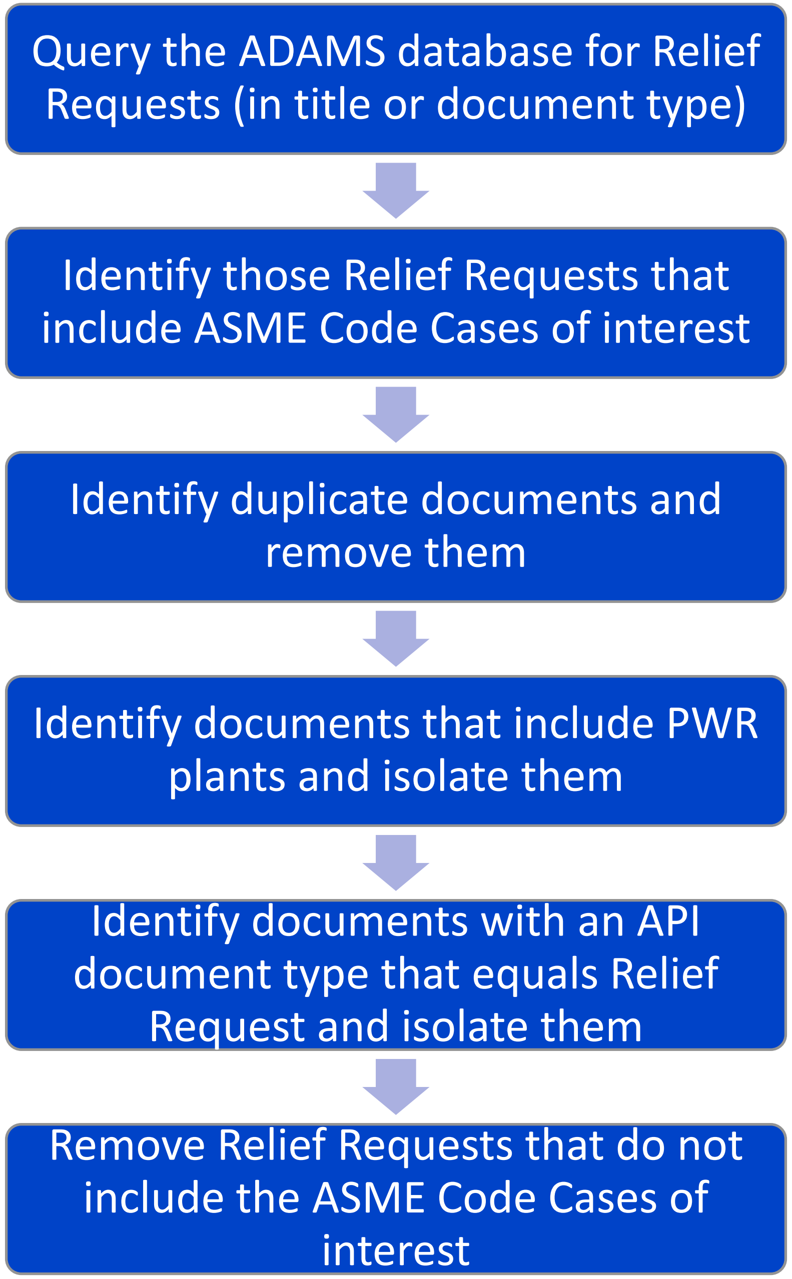
ASME Code Case Number in Series
N-432
N-504
N-562
N-638
N-661
N-666
N-722
N-729
N-740
N-752
N-762
N-766
N-770
N-786
N-789
N-818
N-839
N-853

Process Flow | Creating Code Relief Series



Each bar represents the number of leading documents found after applying the filters described to the right. The objective of the filters is to isolate the most relevant records.

The percentages represent the reduction in records after each filter is deployed.



Process Flow | Creating Code Relief Series

Extract More Records

- Convert PDF files to TXT
- Tag these documents as "Origin"(*) records
- Run NLP algorithm to extract reference numbers, dates, and accession number
- Query ADAMS for additional records based on origin record

Organize the Records

- Group records by the Origin document
- Organize by topical dataset beginning with the oldest date to most recent within dataset
- Assign each dataset a three-digit "Series" number

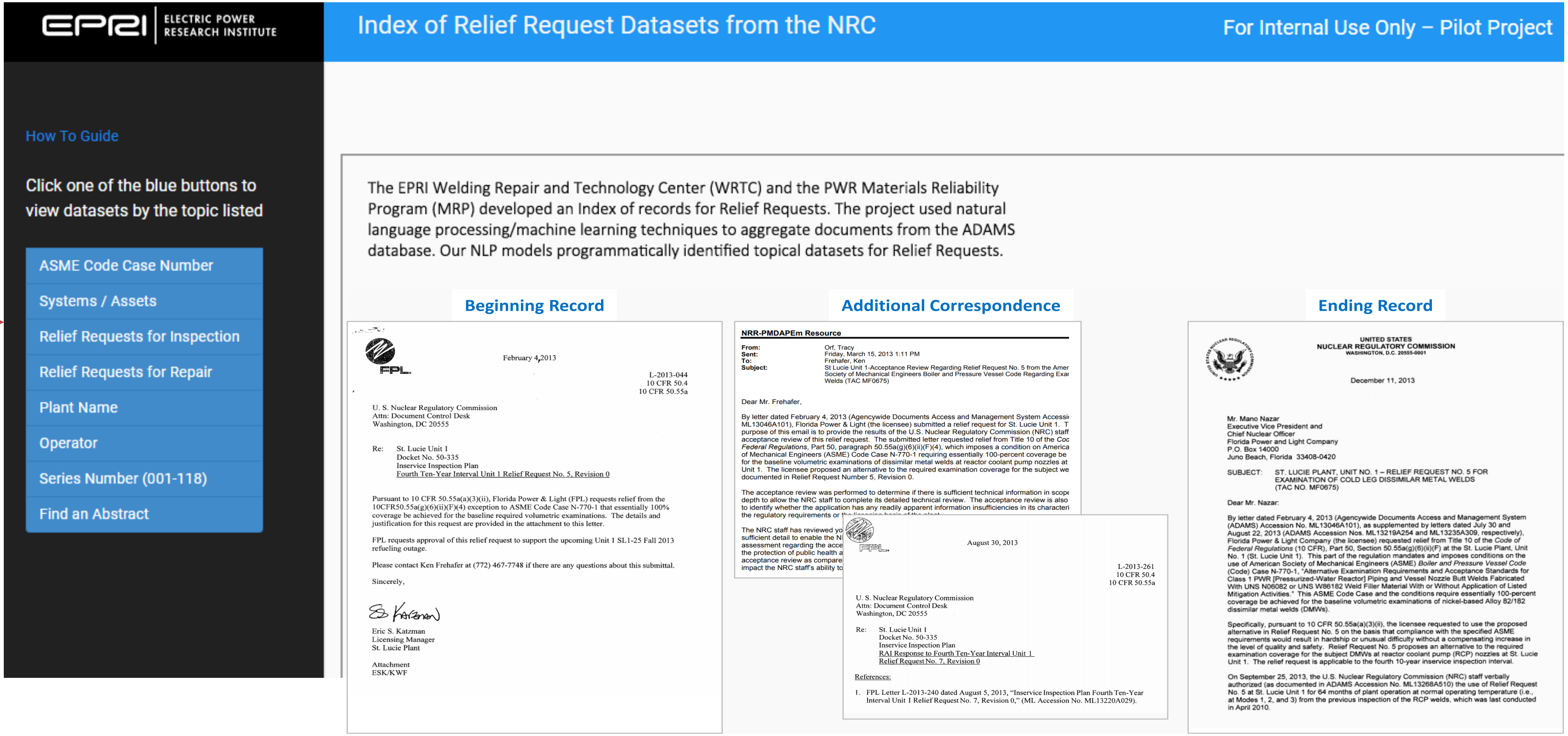
Refine the Topical Datasets

- Remove records within the Series that are not related to the Relief Request for an ASME Code Case
- Remove duplicate series

*** This designation means that these documents are the ones used to expand the search for related records.**

Home Page

Blue buttons lead to different views of the curated data



Datasets organized by the ASME Code Case number that appears in the Relief Request

ELECTRIC POWER RESEARCH INSTITUTE

Index of Relief Request Datasets from the NRC

For Internal Use Only – Pilot Project

How To Guide

Click one of the blue buttons to view datasets by the topic listed

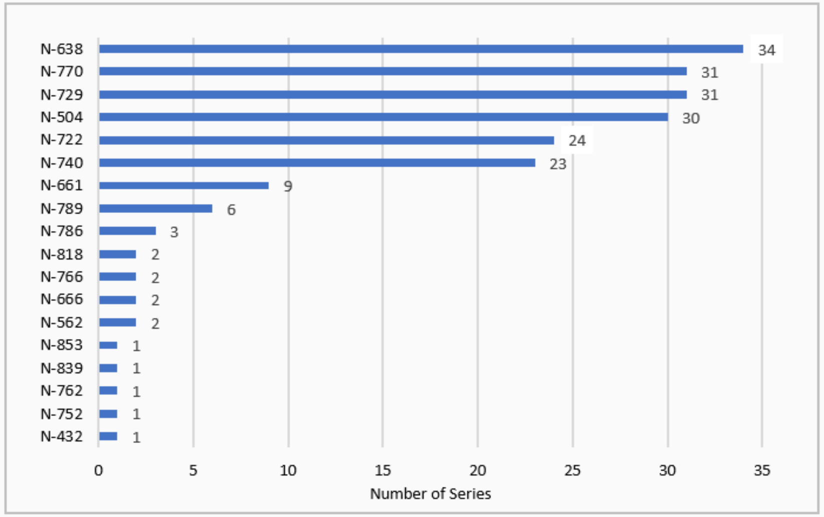
- ASME Code Case Number
- Systems / Assets
- Relief Requests for Inspection
- Relief Requests for Repair
- Plant Name
- Operator
- Series Number (001-118)
- Find an Abstract

In this view, Relief Request datasets are organized by the ASME Code Case number that appears in the beginning or ending record.

Select an ASME Code Case #

Count of Relief Request Series by ASME Code Case

ASME Code Case Number in Series	# of Series w-Code Case
N-432	1
N-504	30
N-562	2
N-638	34
N-661	9
N-666	2
N-722	24
N-729	31
N-740	23
N-752	1
N-762	1
N-766	2
N-770	31
N-786	3
N-789	6
N-818	2
N-839	1



Note: A Relief Request series typically references more than one ASME Code Case number.

A visualization is provided on the home page of each of the topics.

View of datasets for Code Case N-740

In this view, Relief Request datasets are organized by the ASME Code Case number that appears in the beginning or ending record.

N-740

Link to Abstract

Series 105

Series 072

The title headers indicate the start of a unique series.

Plant Name	Operator	Document Title	Date	Author	Code Case #s Appearing in Series	# of Pages	Series #	Link to document
Waterford-3	Entergy Nuclear Operations, Inc.	Waterford Steam Electric Station, Unit 3 - Request for Additional Information Regarding License Amendment Request for Revision of Technical Specification 3/4.7.4,"Ultimate Heat Sink" (EPID L-2018-LLA-0080).	01-28-2019	NRC	['N-504', 'N-638', 'N-740', 'N-770']	6	Series 105	ML19018A010
Waterford-3	Entergy Nuclear Operations, Inc.	Waterford Steam Electric Station, Unit 3 - Proposed Inservice Inspection Program Alternative WF3-RR-19-1 for Application of Dissimilar Metal Weld Full Structural Weld Overlay - Reactor Coolant System Cold Leg Drain Nozzles.	01-28-2019	Entergy Operations, Inc	['N-504', 'N-638', 'N-740', 'N-770']	39	Series 105	ML19028A436
Waterford-3	Entergy Nuclear Operations, Inc.	Waterford, Unit 3, Response to U.S. Nuclear Regulatory Commission Request for Additional Information Regarding Relief Request WF3-RR-19-1 for Application of Dissimilar Metal Weld Full Structural Weld Overlay.	02-04-2019	Entergy Operations, Inc	['N-504', 'N-638', 'N-740', 'N-770']	36	Series 105	ML19035A658
Waterford-3	Entergy Nuclear Operations, Inc.	2019/02/06 NRR E-mail Capture - Verbal Authorization for Relief Request WF3-RR-19-1, Proposed Alternative for ASME Code Section XI, IWA-400 for Waterford Steam Electric Station, Unit 3 (EPID L-219-LLR-0003)	02-06-2019	NRC	['N-504', 'N-638', 'N-740', 'N-770']	4	Series 105	ML19042A298
Waterford-3	Entergy Nuclear Operations, Inc.	Waterford Steam Electric Station, Unit 3 - Authorization of Proposed Alternative to ASME Code Section XI, IWA-4000, "Repair/Replacement Activities" (EPID L-2019-LLR-0003)	08-27-2019	NRC	['N-504', 'N-638', 'N-740', 'N-770']	18	Series 105	ML19232A025
Plant Name	**Operator**	**Document Title**	**Date**	**Author**	**Code Case #s Appearing in Series**	**# of Pages**	**Series #**	**Link to document**
Millstone-2	Dominion Generation	Millstone Power Station, Unit 2 - Alternative Request RR-04-20, Use of Weld Overlays as an Alternative Repair and Mitigation Technique.	04-11-2014	Dominion Nuclear Connecticut, Inc	['N-504', 'N-638', 'N-722', 'N-740', 'N-770']	32	Series 072	ML14112A071
Millstone-2	Dominion Generation	Millstone Power Station Unit 2 - Response To Request For Additional Information Regarding ASME Section XI In-service Inspection Program Alternative Request RR-04-20. Use Of Weld Overlays As An Alternative Repair And Mitigation Technique (TAC No. MF3918)	10-14-2014	Dominion, Dominion Nuclear Connecticut, Inc	['N-504', 'N-638', 'N-722', 'N-740', 'N-770']	7	Series 072	ML14294A453
Millstone-2	Dominion Generation	Millstone Power Station, Unit No. 2 - Alternative Use of Weld Overlay As Repair and Mitigation Technique (TAC No. MF3918).	04-24-2015	NRC	['N-504', 'N-638', 'N-722', 'N-740', 'N-770']	11	Series 072	ML15082A409
Plant Name	**Operator**	**Document Title**	**Date**	**Author**	**Code Case #s Appearing in Series**	**# of Pages**	**Series #**	**Link to document**
Millstone-2	Dominion Generation	Millstone, Unit 2, Fourth 10-Year Interval Inservice Inspection Program and Associated Proposed Alternatives and Relief Request.	07-29-2010	Dominion Nuclear Connecticut, Inc	['N-504', 'N-638', 'N-740', 'N-770']	174	Series 050	ML102580204
Millstone-2	Dominion Generation	Millstone, Unit 2 - Fourth 10-Year Interval Inservice Inspection Program.	08-05-2010	Dominion, Dominion Nuclear Connecticut, Inc, Dominion Resources Services, Inc	['N-504', 'N-638', 'N-740', 'N-770']	124	Series 050	ML102220527
Millstone-2	Dominion Generation	Millstone, Unit 2, Relief Request RR-04-07: Response to Request for Additional Information Regarding Proposed Alternative for Examination Criteria of Weld Overlays.	03-01-2011	Dominion Nuclear Connecticut, Inc	['N-504', 'N-638', 'N-740', 'N-770']	5	Series 050	ML110610207

Each Series has an NLP Developed Abstract

ELECTRIC POWER RESEARCH INSTITUTE

Index of Relief Request Datasets from the NRC

For Internal Use Only – Pilot Project

How To Guide

Click one of the blue buttons to view datasets by the topic listed

- ASME Code Case Number
- Systems / Assets
- Relief Requests for Inspection
- Relief Requests for Repair
- Plant Name
- Operator
- Series Number (001-118)
- Find an Abstract

adams-index.beyondthearc.info/abstracts/Code%20Case%20Series%20105.html

Apps Health & Wellness Inside EPRI - Home ECM Google DeepNLP - EPRI- Pr... DeepNLP-Admin

Plant Name	Abstract - Initial Relief Request	Doc Number
Waterford 3	Note that ASME Code Case N-504-4 has been conditionally approved by the NRC in RG 1.147 with the condition that the provisions of ASME Code, Section XI, Appendix Q be met when using the Code Case. In order to maintain the pressure boundary and structural integrity of the welds, Entergy proposes to perform full structural weld overlays based on ASME Code Case N-740-2. Inc. (Entergy) proposes, as an emergent repair, to mitigate the SCC susceptibility of the Waterford Steam Electric Station, Unit 3 (Waterford 3) reactor coolant system (RCS) cold leg drain nozzle DMWs between the nozzle and safe end by installing a full structural weld overlay (FSWOL) on the DMWs.	ML19028A436

Plant Name	Abstract - Closing NRC Letter	Doc Number
Waterford 3	In lieu of repairing or replacing the subject welds in accordance with the ASME Code, Section XI, the licensee proposed to install a full structural weld overlay (FSWOL) on the affected welds based on the methodology contained in ASME Code Case N-740-2, "Full Structural Dissimilar Metal Weld Overlay for Repair or Mitigation of Class 1, 2, and 3 Items Adherence to Section XI of the ASME Code is mandated by 10 CFR 50.55a(g)(4), "Inservice inspection standards requirement for operating plants," which states, in part, Throughout the service life of a boiling or pressurized water-cooled nuclear power facility, components (including supports) that are classified as ASME Code Class 1, Class 2, and Class 3 must meet the requirements, except design and Enclosure - 2 - access provisions and preservice examination requirements, set forth in Section XI of editions and addenda of the ASME BPV Code.... The licensee compared the proposed alternative to ASME Code Case N-504-4, "Alternative Rules for Repair of Class 1, ,2 and 3 Austenitic Stainless Steel Piping, Section XI, Division 1," and the ASME Code, Section XI, Appendix Q, as shown in Attachment 2 of the relief request dated February 4, 2019.	ML19232A025

Where are we going next

- Easier way to find a complete series of information
- We can now look at data in new ways
  - What does this data mean?
  - Are we seeing initial trends (example: start of degradation in certain components, need for new Code changes, research, etc.)

- Next Steps:
  - Mine a larger NRC ADAMS data set now that the process has been developed and determine if there are any interesting trends
  - Obtain broader member feedback from proof of concept
- Future: Potential to use developed process on structured NRC ADAMS dataset from Event Management Response Tool (EMRT) project and other code cases, requests for alternative

Questions?

Together…Shaping the Future of Energy™

Power Industry Dictionary for Text-Mining and Natural Language Processing Application

Proof of Concept

Karen Kim-Stevens, kkim@epri.com
EPRI Principal Project Manager, Radiation Safety

U.S. NRC Data Science and Artificial Intelligence Regulatory Applications Workshops
Workshop #1 June 29, 2021

www.epri.com

Today, NLP tools will parse words based on their more common usage

NRC.gov text

On April 6, 2006,
a drain noun
cooler noun
relief valve noun
in the
feedwater ? system noun
lifted and
remained open

Word indexing

A Generic NLP dictionary

drain

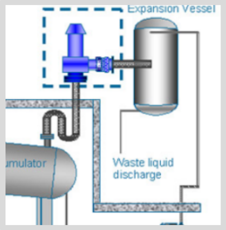
cooler

relief valve

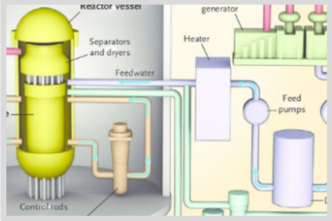
Vs.

B Groundwater Contamination Classifier

drain cooler
relief valve



feedwater system



Objective: Build a Nuclear Industry NLP Dictionary

Our *use case* for this proof of principle

Groundwater Contamination

Use case owner: Karen Kim-Stevens

Goal: Develop a NLP proof of principle that demonstrates the potential benefits of machine learning applied to this domain.

Tasks

- Create a preliminary dictionary to be used for classification.
- Develop a NLP text analytic demo and generate preliminary insights.

Benefits

Natural language text analytics will help the industry enhance preparation and implementation of mitigating actions in the event of inadvertent leaks and spills of radioactive materials.

Scenario

The industry has thousands of text documents from operating experiences, maintenance reports, work orders, regulatory filings, and more that reference Groundwater Contamination. Given the safety significance and the need to find ways to operate more efficiently, all nuclear plants would benefit from extracting and sharing key information from these documents to make quicker, informed decisions, reduce the number of inadvertent spills and leaks, and enhance the safety and response time to a contamination situation.

Potential use cases to develop risk mitigation strategies

Identify Specific SSCs

Identify which SSCs could be associated with a failure and release radioactive liquid into the environment

- How have the sources of SSC leaks and spills changed over time?
- Does the age of the plant impact the components?
- Do certain components leak after a certain amount of time in service?

Work Practices

Identify which work practice tasks could be associated to which jobs or systems that could cause the most release of radioactive liquid into the environment

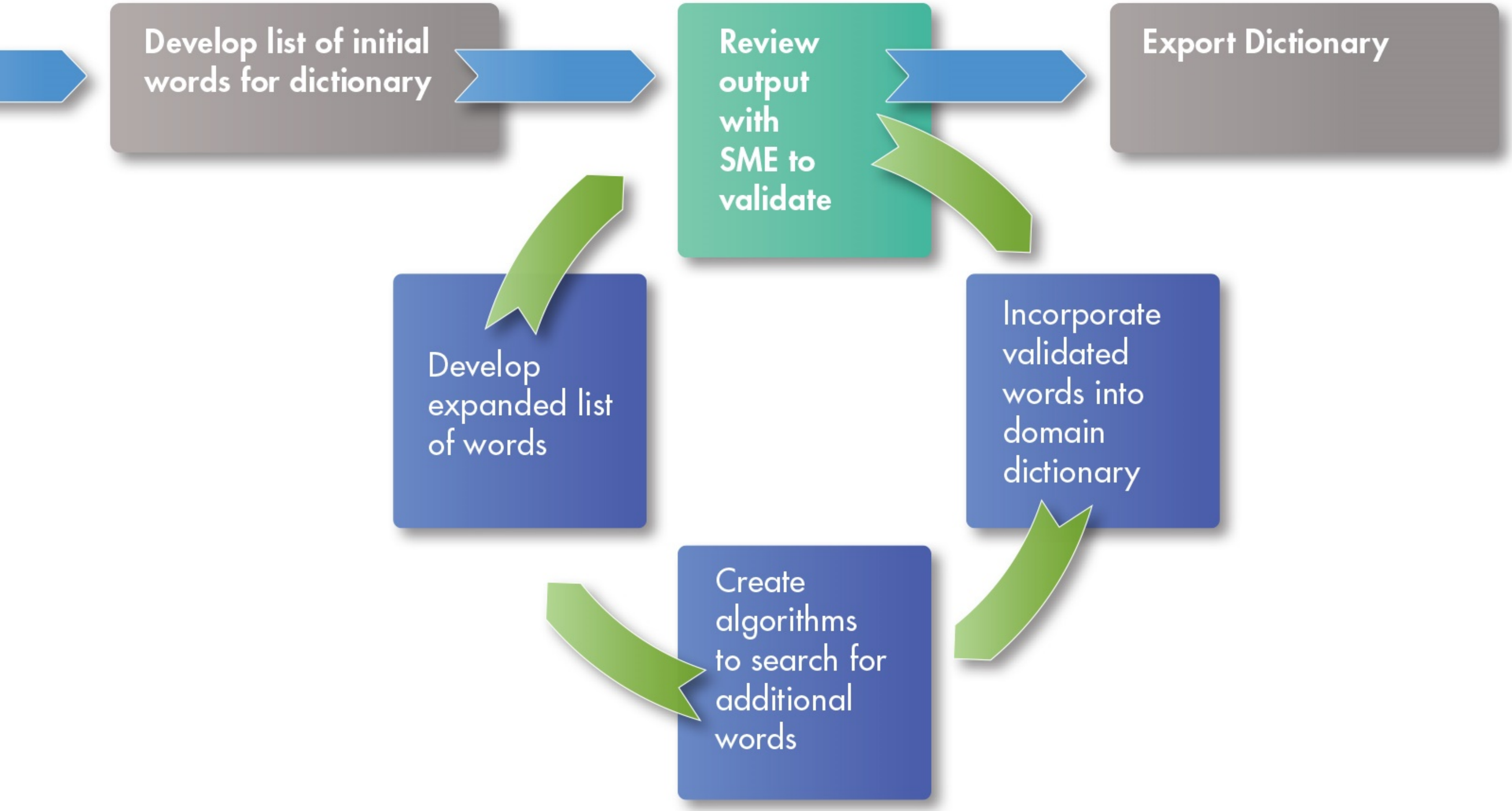
- Do work practices during planned vs. unplanned outages affect the prediction?
- Do routine vs. non-routine affect the prediction?
- How have leaks from work practices changed over time?

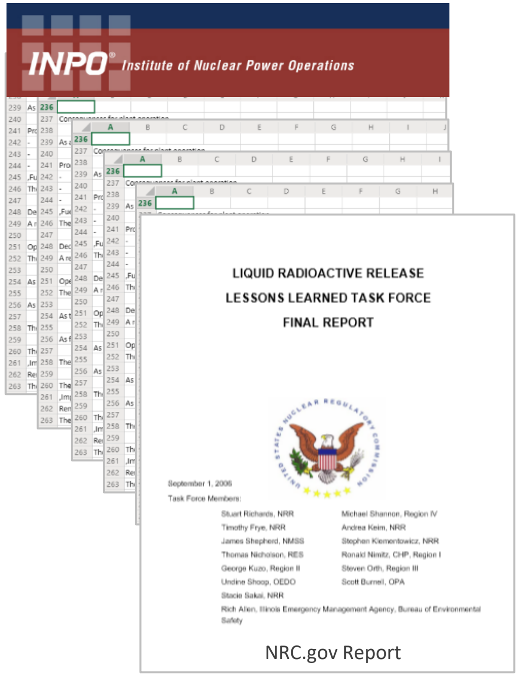
Concentration of radioactive material

Identify how concentration of radioactive material varies by type of leak or spill

- How much does the concentration vary by SSC or WP?
- Does the magnitude vary by for SSCs at BWR vs. PWR plants?
- Can this information be used to help plants identify the source of leaks or spills?

To ensure data integrity and quality results, we followed a structured data science approach

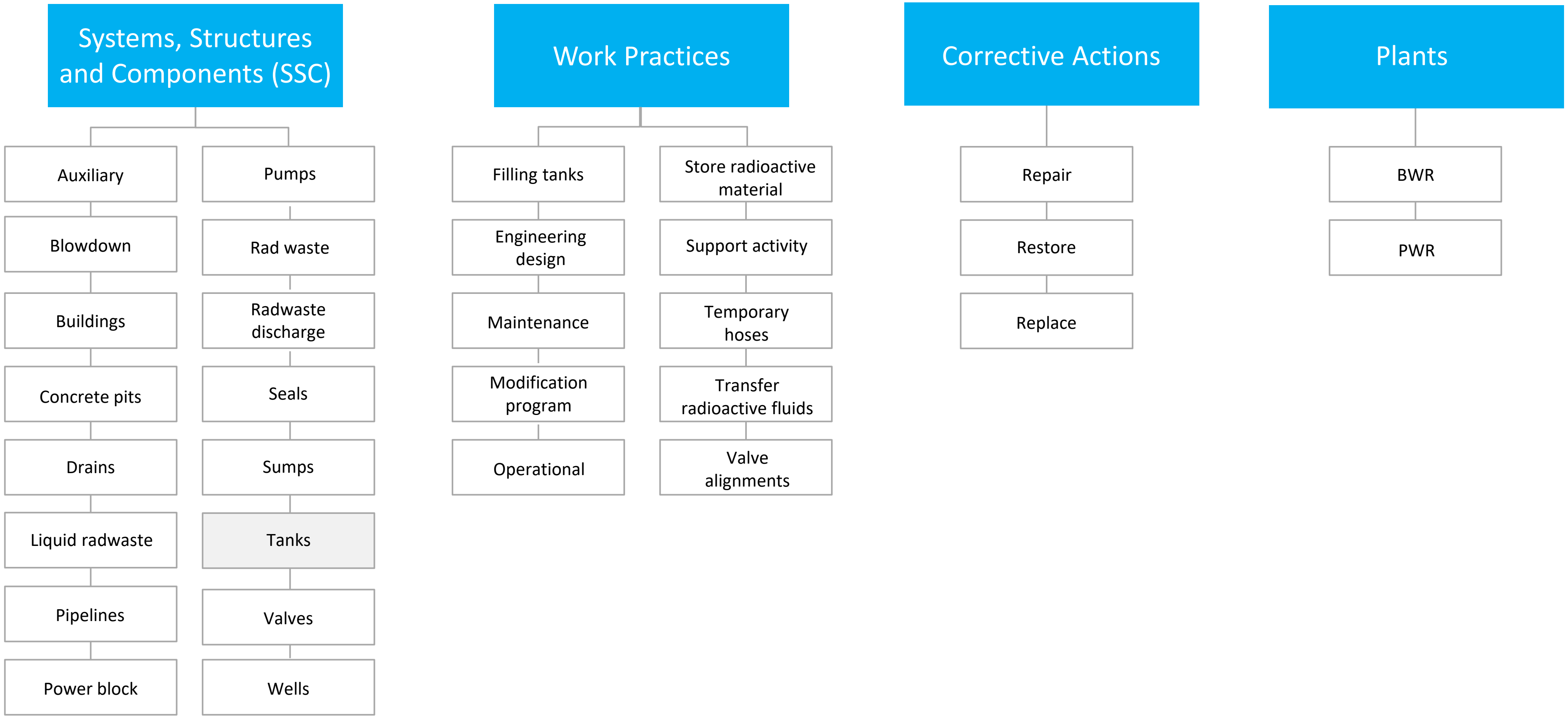




These NLP techniques help to get preliminary results quickly

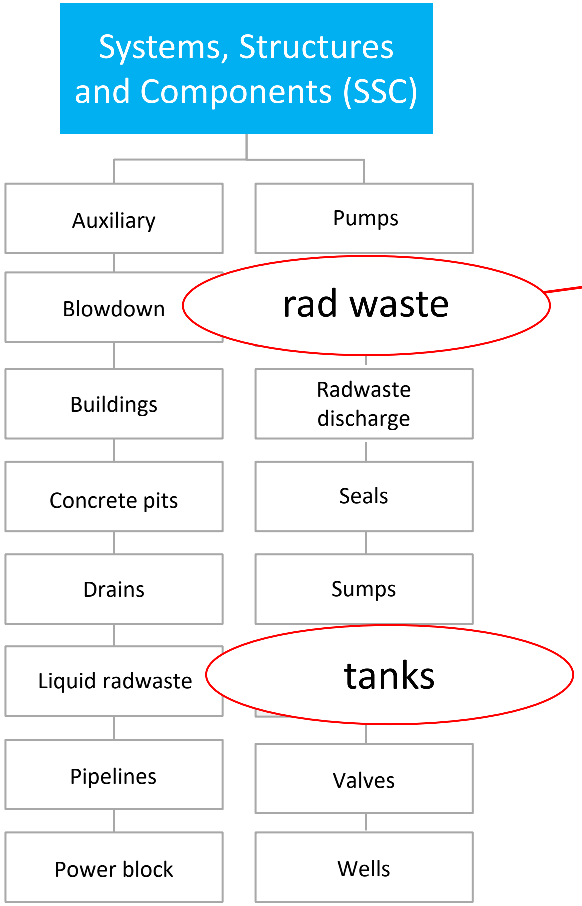
	What is it?	Why?	Example
1 Tokenization	Algorithms to segment text into groupings, phrases, punctuation, called "tokens."	Tokens become the inputs for conducting text mining.	Text: A relief valve in the feedwater system Tokens: "A" "relief valve" "in" "the" "feedwater system"
2 Part of Speech	Statistical models that assign a part of speech to each token - noun, verb, adjective, adverb, etc.	These tags enable modeling to infer the relationships between words in phrases and sentences.	A | relief | valve | in | the | feedwater | system DET ADJ Noun ADP DET Noun Noun
3 Name Entity Recognition	Statistical models that assign labels to tokens such as date, quantity, location and more.	Entity recognition is helpful for information extraction and filtering.	On April 6, 2006 DATE , a drain cooler relief resulting in secondary plant steam being released wall. Approximately 114,000 gallons QUANTITY property LOCATION . The system containing the
4 Rules-based Matching	Algorithms that find phrases, sequences of tokens, and entities.	Improves information extraction and text mining. This approach is one way to directly incorporate SME input.	The cause of the leak was from the liquid radioactive waste processing system. The system containing the feedwater was known to contain tritium. "cause of leak" "liquid radioactive waste processing system" "tritium"

Architecture for the library of dictionaries



The dictionary map provides us with guidance on how topics are organized, overlap and relate to each other. The design evolves based on programmatic exploration and feedback from our SME.

Each word has an additional level of word associations that serve as training topics for machine learning



rad waste
rad waste facility
rad waste floor drain
rad waste peo
rad waste system
rad waste yard
LLRW
low level rad waste
low level rad waste facility
non-rad waste system
NRW

tanks
condensate storage tank
condensate head tank
hydrazine mix tank
mobile tank
sump drain tank
sump pump tank
water storage tank
tank storage area
CST

The lack of a consistent industry nomenclature is a key challenge in building NLP models

For example

groundwater
- gw
- ground water
- ground-water
- gnd water
- g water
- g-water

picocuries
- pCi/L
- pCi
- pCi / L
- picocuries / liter
- picocuries per liter
- pCi/liter
- pCi / liter

pits
- basins
- moats
- motes
- ponds

power block
- auxiliary building
- auxiliary system
- rad waste building
- radwaste building

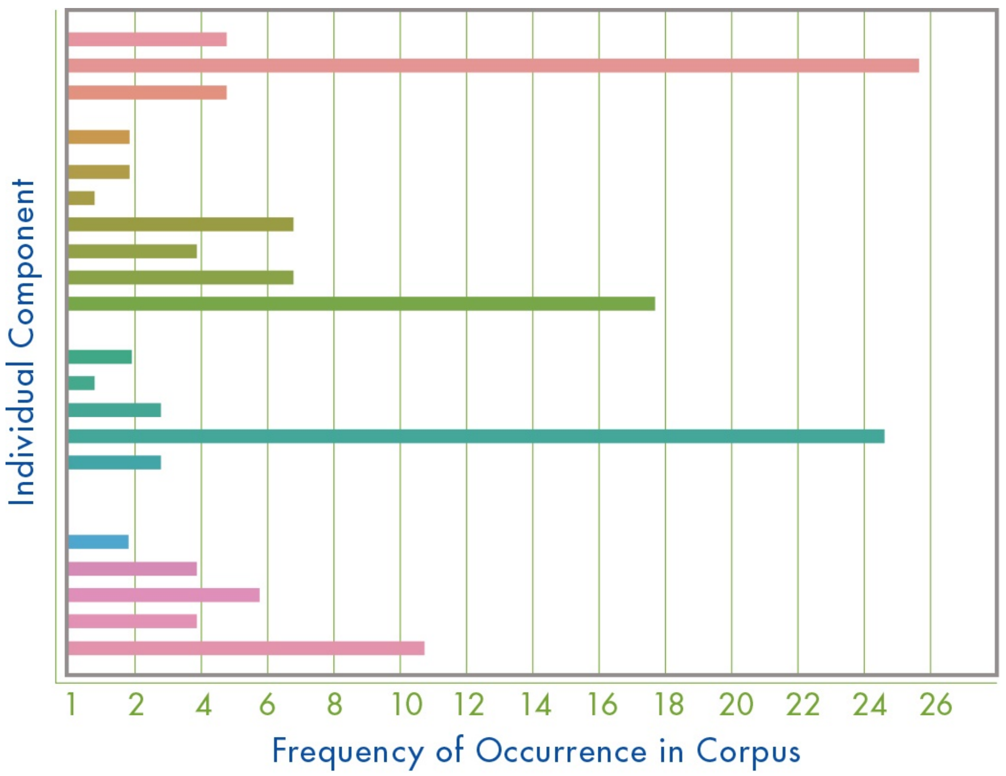
seismic gap
- cracks
- rattle space
- seals
- structural joints

storm drains
- drain systems
- roof drains
- storm systems
- yard drains

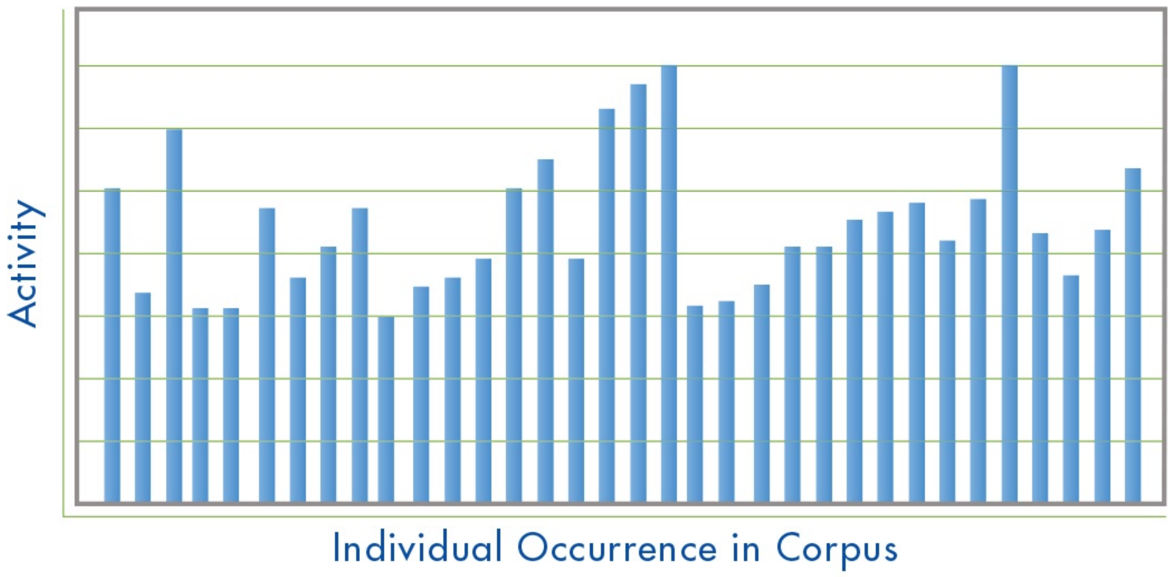
These word variabilities were identified through NLP algorithms and augmented by our subject matter expert (SME). A SME needs to provide guidance as we use the NLP tools.

Preliminary Visualizations

Counts for systems, structures, components associated with groundwater incident reports



Tritium activity (logarithmic scale) of reported events extracted from corpus for one type of reactor design



Elements of our work are transferable
and can help others get started on a similar project

Roles to make this type of project successful

- Project Sponsor
  - The project sponsor is the person or group who owns the project. They hold overall accountability of the project and are responsible for providing resources, support and guidance to enable success. This role ensures that the analysis is aligned with research and business goals.
- Subject Matter Expert (SME)
  - The SME plays a vital role in helping the data scientist understand the data and its nuances. This role will evaluate the text analytic output, ensure it is producing relevant results, and help to describe the specific real world problem that the machine learning project is trying to solve.
- Data Scientist
  - A data scientist collects, analyzes, and interprets large amounts of data. Their skills and expertise in highly advanced analytical tools enables them to understand the data and develop operational models, systems and tools by applying experimental and iterative methods and techniques.
- Data Analyst
  - A data analyst examines the patterns, trends, and other insights extracted from the data. They are responsible for deriving meaningful, actionable insights from the data. They support the project by creating visualizations.

Key Takeaways

- Open-source dictionaries do not understand electric power industry language
- An industry-specific dictionary is needed to conduct text mining and apply NLP-based algorithm
- A workflow template for dictionary construction is repeatable that can be applied to new topics
- The development of an industry specific dictionary will require investment
- However, the nuclear industry will benefit from more efficient ways of digesting and applying industry data and knowledge

For More Information:

Please download “Quick Insight – Power Industry Dictionary for Text-Mining and Natural Language Processing: Proof of Concept.”
https://www.epri.com/research/products/000000003002019609

Together...Shaping the Future of Electricity

High Level Process for NLP Projects

Corpus
- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze
- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Steps Needed to Build a Dictionary (Corpus → Extract)

Process of Project Utilizing Natural Language Processing (Corpus → Analyze)

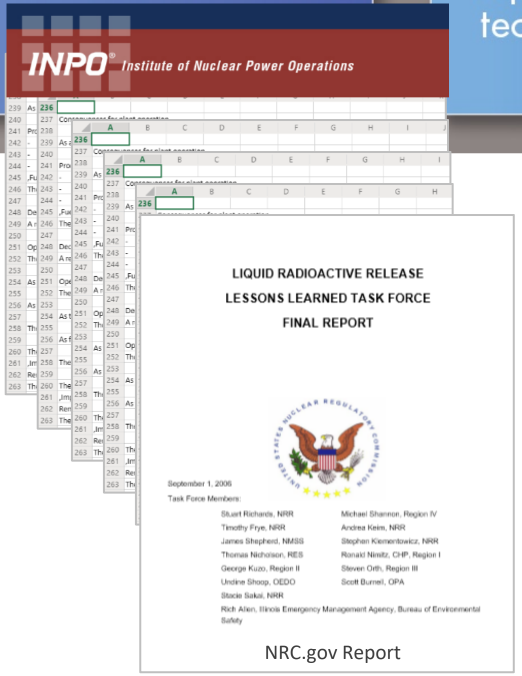
High Level Process for NLP Projects

Corpus
- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze
- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses



NRC.gov Report

High Level Process for NLP Projects

Corpus
- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

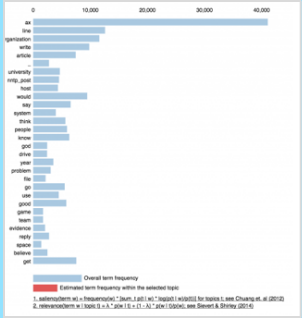
Extract
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze
- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Parts-of-Speech Tagging

N
a |drain cooler relief valve| in the
N V
|feedwater system| lifted and
V Ad
remained open

Word Frequency



Move forward, when you have enough understanding of the text to work with it as data

ELECTRIC POWER RESEARCH INSTITUTE

High Level Process for NLP Projects

Corpus
- Acquire text corpus
- Understand dataset
- Is the corpus adequate

Organize
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

Analyze
- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Pattern Matching Algorithms

determined contamination	due to	Unit 1 Spe
monitoring wells installed	due to	historical
onsite tritium contamination	due to	past oper

Key Terms Algorithms

liquid rad waste tank storage
liquid rad waste processing
temporary systems effluent
aux storm drain systems system
power block rad storage area
cathodic protection system source

Simultaneously search key words and patterns in the corpus

Move forward, when you have enough structure to begin text mining

High Level Process for NLP Projects

Corpus
- Acquire text corpus
- Understand dataset
- Is the corpus adequate

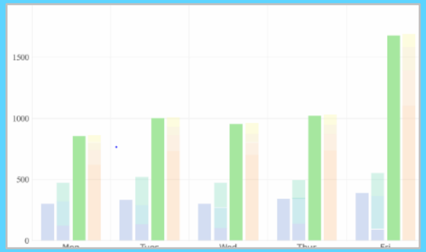
Organize
- Ingest data
- Confirm key attributes such as frequency of terms
- Apply NLP processing techniques to dataset

Extract
- Create structure in unstructured data
- Identify key terms
- Use different NLP methods/algorithms
- Create preliminary visualizations

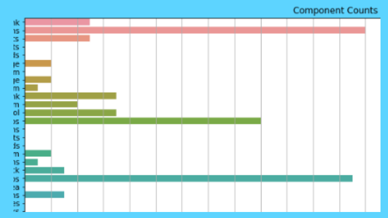
Analyze
- Perform text analytics
- Create visualizations
- Formulate new insights
- Answer hypotheses

Iterate. The analysis will determine if more extractions are needed

Magnitude of Leaks/Spills



High Risk Components PWR plants



INPO and Data Science

Paul Steiner
Manager
Data Management and Industry Trends

Data Science Application

- Supports monitoring station and corporate performance between evaluations/peer reviews
- Informs application of resources

Data Science Tools - Current

- Neural Models Applying Artificial Intelligence

- Models Leveraging Machine Learning

Neural Modeling

- Hundreds of data points are collected monthly
- Experience records are continuously reported
- Thousands of indicators are developed combining the data points and experience records
- Effects based models – limits subjectivity
- Neural modeling identifies patterns within these indicators that correlate to overall and area assessments

Data Science - Future

- Neural Forecasting
- Scram Correlations
- Equipment Failure Correlations
- Predictive, Behavior-Informed Modeling

Questions?