Data Analytics: Exploring Unknown Unknowns

Peter Paul, PhD
Engineering Manager
MKS Power Solutions
October 12, 2018

mks

What is Analytics?

- Using data you currently have to make better business decisions

Turning this…

Actionable Information

…into this

Data Analytics Applications are All Around us

- Amazon – Recommendations based on Browse & Purchase History (yours and others)

- Netflix – Recommendations based on Content Already Watched

- Google – Google Flu Trends: Predict Flu Outbreaks based on search queries, among others

- Safeway – Grocery Store Coupons

- Sate Farm – Insurance Rates

Why the Explosion of Data Analytics Applications?

Massive Data Collected & Warehoused
- Web Data
- eCommerce
- Social Media
- Purchase Records
- Financial Transactions
- Cell Phone Data
- IOT

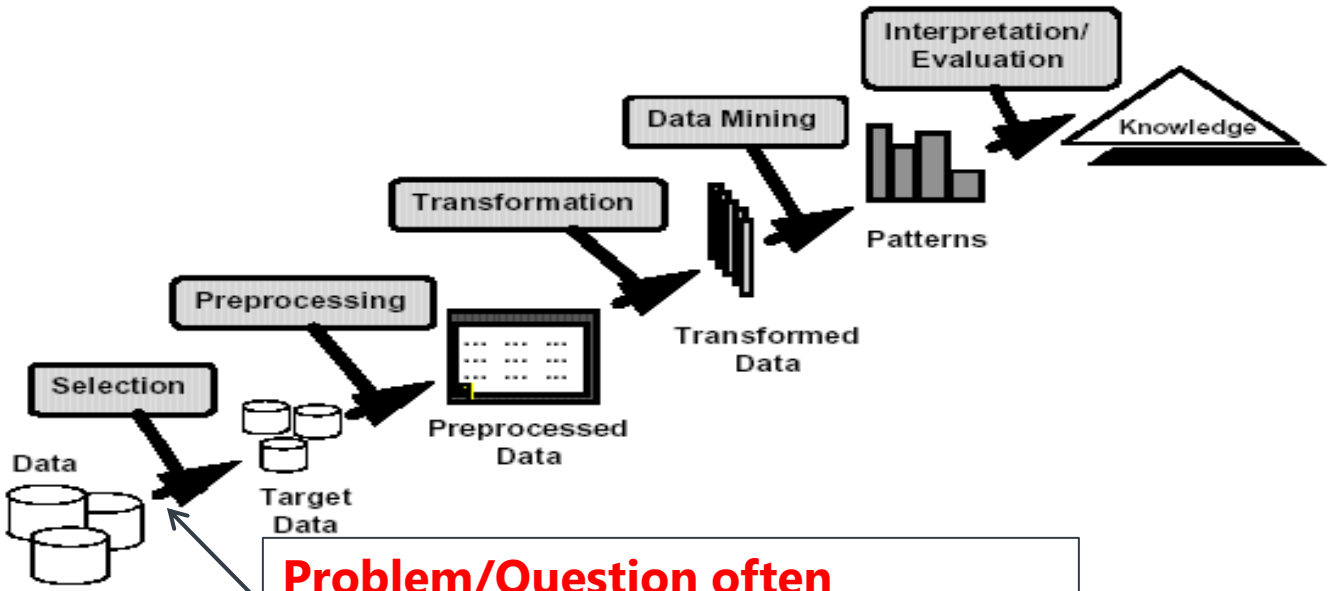
Sophisticated Algorithms
- Machine Learning
- Pattern Recognition
- AI
- Statistics

Extremely Competitive Business Environment
- Mass Customization
- Provide an "Edge"
- Low Cost Business Intelligence

Computer Resources
- Fast & Cheap Processing
- Fast & Cheap Storage
- Fast & Cheap Networking
- On-Demand Computing ("Cloud")

"Textbook" Definition

Automatic exploration & analysis of large quantities of data in order to discover meaningful patterns.



Problem/Question often Formulated *after* data is collected

Data Analytics is Not:
- **Basic Slice & Dice Data Reporting**
- **Only about visualizing the past**
- **Simply another name for statistics**
- **A "Magic Wand"**

Modified From: Tan, Steinbach, Kumar, Introduction to Data Mining.

What kinds of Analytics Operations can be Performed on the Data?

- **Description Methods –** Find human-interpretable patterns that describe the data.
- Clustering – Finding Natural Groups among the Data → Grouping customers into categories.
- Association Rule Discovery – Finding Data Attributes that appear Together → If I buy potato chips, it is likely that I also buy beer?
- Sequential Pattern Discovery – Finding Data Attribute that appear in a sequential pattern → If I buy a book titled: "Golf for Beginners", it is likely that I will buy Golf Lessons sometime in the future?
- **Prediction Methods –** Use some variables to predict unknown or future values of other variables.
- Classification – Predicting which Class an Object Belongs to → What kind of object do I have?
- Regression – Trends, Forecasting → Preventative Maintenance Actions
- Deviation Detection – Anomaly Detection, Finding an Occurrence or an Object that is out of the ordinary. → Fraud Detection

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining

Case Study: Analytics Based Automation of Car Pool Lane Enforcement

Congestion costs US $87 billion/year in wasted fuel and time (2010)

High Occupancy Vehicle lanes (HOV)
High Occupancy Tolling lanes (HOT)

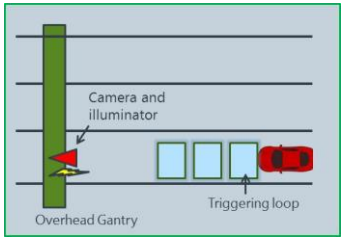
Police Enforcement has Proven Difficult:
- **HOV lane violation rate: up to 65%**
- **Manual HOV enforcement rate: <10%**

HOT Lanes:
HOV Lanes which Single Occupant Vehicles may use if they pay a toll.

Case Study Material from PPaul WACV2014, and PPaul EI2013

→ **Motivates Automated Enforcement**

Computer Vision = Image Processing + Machine Learning



2. OPERATIONAL PHASE: Prediction

Image I've Never Seen Before

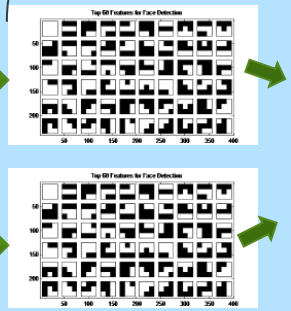
$f(I)$ → Empty

1. TRAINING PHASE: Learn f()

$\hat{L} = f(I)$

Label: Empty

Label: Occupied



Classification Algorithm

Empty, Occupied

Iterate

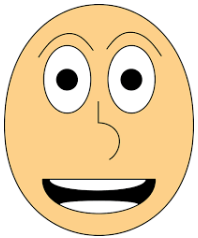
Labeled Training Images

Feature Extraction

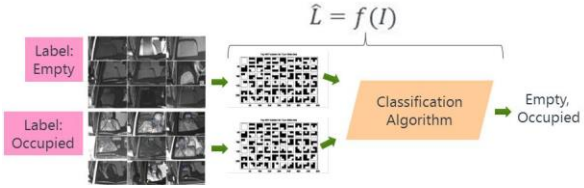
Machine Learning Algorithm

Predicted Label

Evolution of Machine Learning: Heuristics → Engineered Features + Classifiers → Deep Learning

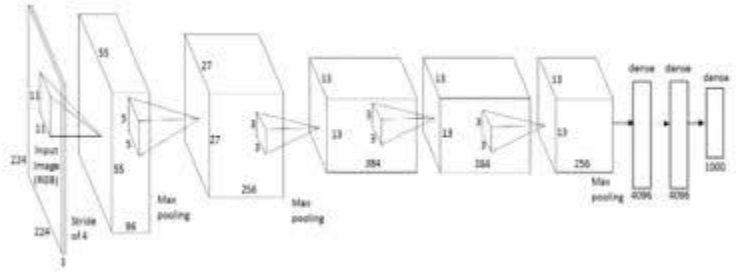


Early AI - Use Human Derived Heuristic: "A Face is Two Eyes a Nose and a Mouth"



More Recent AI:

(1) Define Image Primitives ("engineered features")
(2) Present Labeled Images
(3) Let Computer Determine how it will use primitives to detect a face

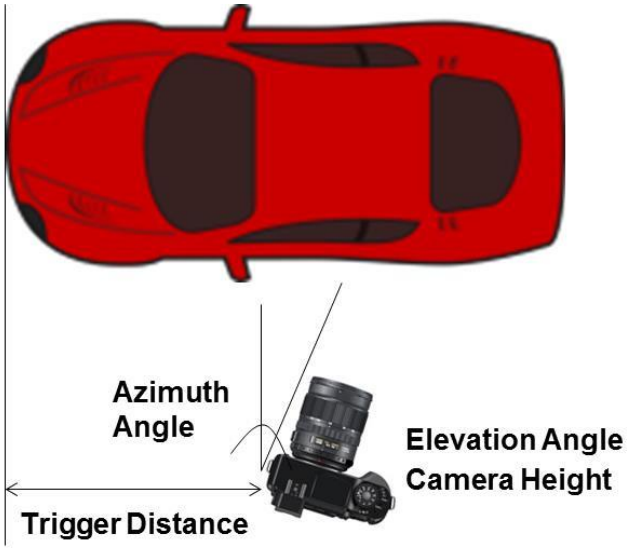


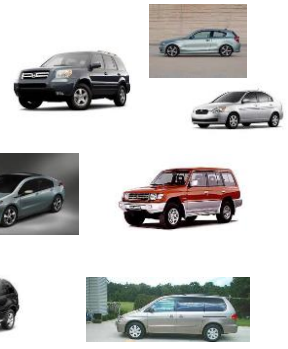
Latest AI – "Deep Learning":

(1) Present Labeled Images
(2) Computer jointly learns best primitives and how to use them

What about different head poses?
Variation among people?
Occlusions from hats & sunglasses?

Image Rear Seat Passengers through Side Window



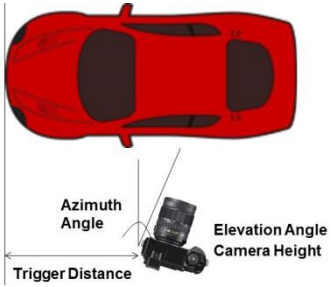


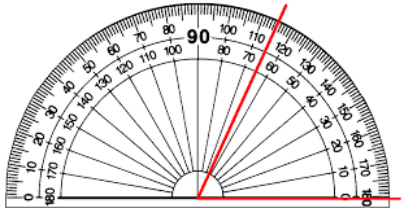
Where do we mount the camera relative to the car?

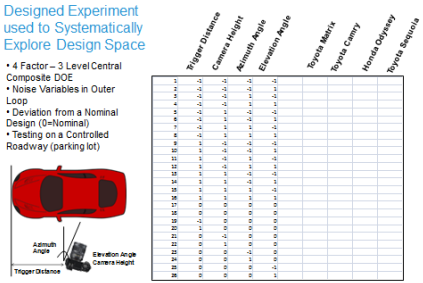
→ Lots of Sizes & Shapes of Cars

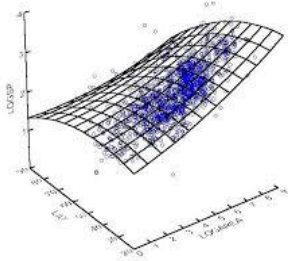
→ Lots of Variety in People

First Principles, Designed Experiments, & Machine Learning: Known Unknowns & Unknown Unknowns









Where do we mount the camera relative to the car?

→ Lots of Sizes & Shapes of Cars

→ Lots of Variety in People

Method 1: Geometry

(1) Define "Average" Car

(2) Define "Average Person"

(3) Perform geometric calculations

Method 2: DOE

(1) Define Mix of Cars

(2) Define Mix of People

(3) Perform Designed Experiment

(4) Determine Experimental Regression Model

(5) Determine Inputs that Maximize Model Output

Method 3: Machine Learning

(1) Mount Camera on Roadway

(2) Collect Data

(3) Assess Occupancy Detection Performance

(4) Adjust Mounting & Repeat

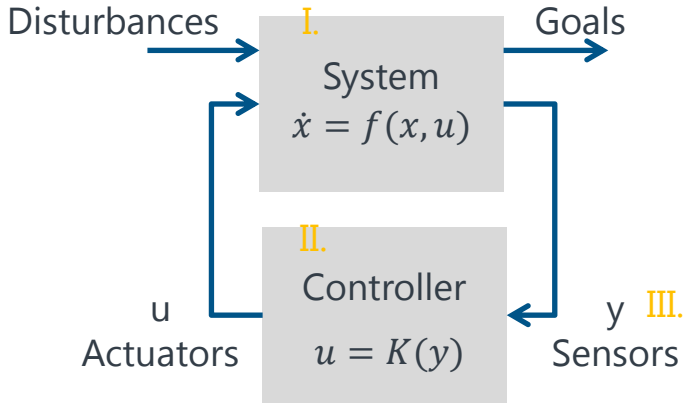
(5) Use Data to built empirical model of Inputs to Outputs

(6) Optimize Model

Data-Driven (Machine Learning) Control

Systems with these Challenges:

- High Dimensional – many distributed actors
- Unknown Dynamics (no First Principles models) – unknown interaction between actors
- Nonlinear – actors perform complex actions
- Typically have Limited Measurements
- Typically have Limited Actuators



I. Data Driven Models
II. Learning Control
III. Sensor & Actuator Placement

Examples:

- Neuroscience: Prevent Seizures
- Turbulent Fluid Systems
- Suppress the spread of Disease
- Regulate Markets
- Transportation Systems
- Power Grid

What is Control?
Optimization Constrained by Dynamics
What Machine Learning Control?
Powerful Nonlinear Optimization based on Data

See: https://www.youtube.com/watch?v=oulLR06lj_E&list=PLMrJAkhIeNNQkv98vuPjO2X2qJO_UPeWR

PROS/CONS of First Principles Models, DOE Models, and Machine Learning Models

- First Principles
  - Analytic Models
  - Finite Element Models
  - Computer Simulation
  - Can Simulate situations that cannot be experimented
  - Insight into parameter trade-offs & Sensitivities
  - Likely Cannot Comprehend all Noises & Variations in Real-World Problem
  - Parameter Optimization before “going live”
- Designed Experiments
  - Can comprehend some noises and variation, but not all
  - “Known Unknowns”
  - Parameter Optimization before “going live”
- Data-Driven (Machine Learning)
  - Big Data needs Big Data
  - Dataset must include all operating conditions, noises, & variations likely to be encountered in operation
  - May need to “go live” to collect data before system parameters are optimized
  - Dataset includes variation distributions and PDFs that are not known a priori
  - Dataset includes interactions not known a priori
  - “Unknown Unknowns”
  - Can be resistant to Human Biases
  - “Let the Data tell the Story”

Thank You!

Contact:
peter_paul@mksinst.com

mks