Introduction to Artificial Intelligence
Part II: Methods & Practice

Hsing-Kuo Pao (鮑興國)
National Taiwan University of Science & Technology

Outline

- A brief review of Part I
- Getting started with the data science journey
- Data vs. Models
- Data that Drives
- Welcome to the AI/ML Models
- An Introduction to Deep Learning (Deep Neural Networks)
- Conclusion

A Brief Review of Part I
前情提要

— Introduction to AI —

Confirming the Meaning of AI

"[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning …" (Bellman, 1978) => Thinking humanly	"The study of mental faculties through the use of computational models" (Charniak & McDermott, 1985) => Thinking rationally
"The study of how to make computers do things at which, at the moment, people are better." (Rich & Knight, 1991) => Acting humanly	"The branch of computer science that is concerned with the automation of intelligence behavior" (Luger & Stubblefiel, 1993) => Acting rationally

From Traditional AI to Modern AI

- Out of the long history of AI…
  - Logic & Reasoning
  - Planning
  - **Learning**
  - Acting, communicating, perceiving, etc.
- A Success in Formulating “Induction”
- Machine Learning: using the past experience to improve the performance on certain tasks give a pre-defined measure

LeCun's Comments on Learning Algorithms

- "Pure" reinforcement learning (cherry)
  - The machine predicts a scalar reward given once in a while
  - A few bits for some samples
- Supervised learning (icing)
  - The machine predicts a category or a few numbers for each input
  - 10 to 10,000 bits per sample
- Unsupervised learning (cake)
  - The machine predicts any part of its input for any observed part
  - Millions of bits per sample
- Semi-supervised learning
  - The no. of unlabeled data >> the no. of labeled data!

A General Steps of Data Science

- Collecting data
- Making assumption
- Finding representation
- Selecting model
- Evaluation
- Refinement & Improvement

$\Rightarrow$ Each step is possible to make mistakes and resolving failure

Getting Started with the Data Science Journey
尋找指環的旅程…

— Introduction to AI —

Knowing the Trend
話說當今天下大勢…

What to target for AI (from LeCun):

- Machines need to learn/understand how the world works
  - They need to acquire some level of common sense
- They need to learn a very large amount of background knowledge
  - Through observation and action
- Machines need to perceive the state of the world
  - So as to make accurate predictions and planning
- Machines need to update and remember estimates of the state of the world
  - Paying attention to important events. Remember relevant events
- Machines need to reason and plan
  - Predict which sequence of actions will lead to a desired state of the world

⇒ Most of the knowledge in the world in the future is going to be extracted by machines and will reside in machines.

Knowing the Limits of AI

- The research on the technological singularity
  - "The singularity and the state of the art in AI" by E. Davis, 2014
- Strong AI vs. weak AI
  - Strong AI states that a computer with the right program can in fact be mental (like a human being)
  - Weak AI just aims to solve problems, not necessarily to be mental or model human behavior.

Before Going to Solve Problem via AI/ML Techniques

- Knowing the trend, and knowing the state of the art for different applications
- Knowing how AI can and cannot do (or at least not good at it)
- Knowing the data
  - Respect data by all means
  - Garbage in, garbage out!
- Knowing the models
  - The last part to worry about…

What Can Go Wrong with Data Science

- Mistakes can lead to wrong conclusion
- Mistakes can conclude wrong knowledge
- Mistakes can suggest wrong decision

Ten Most Common Mistakes/Issues by Data Scientists

ASSUMPTION

- Same or different distributions in training and test sets

REPRESENTATION

- Normalization, axis scaling
  - Visualization
  - Methods that need attributes of the same kind: clustering? Unsupervised dimensionality reduction?
- The curse of dimensionality
- Primal space vs. feature space
  - Primal space = attributes, feature space = data

ESTIMATION

- Generative modeling vs. discriminative modeling
  - We want to draw a cat or simply distinguish a cat from a dog?
- Transparent methods vs. black-box methods

Ten Most Common Mistakes/Issues by Data Scientists (cont'd)

EVALUATION

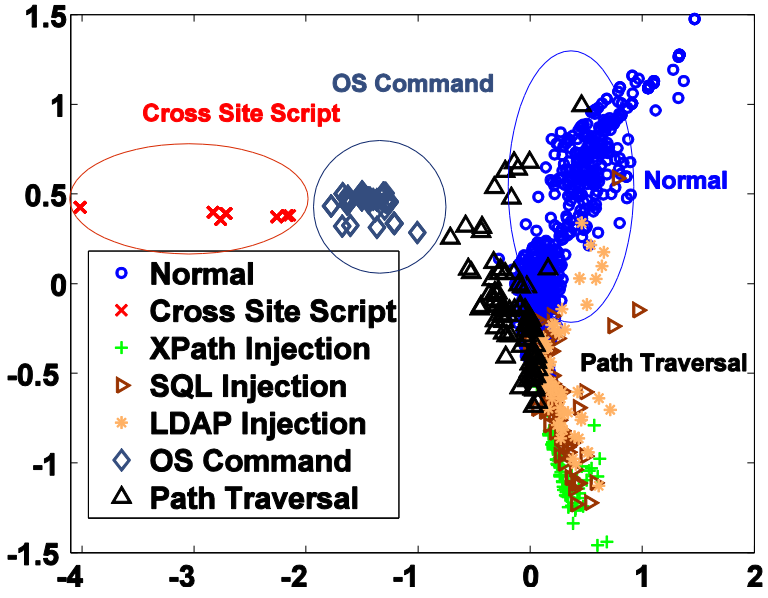
- Effective cross-validation
  - How many folds?
  - stratification
  - Dealing with temporal data
- Training error vs. test error
  - Time series data

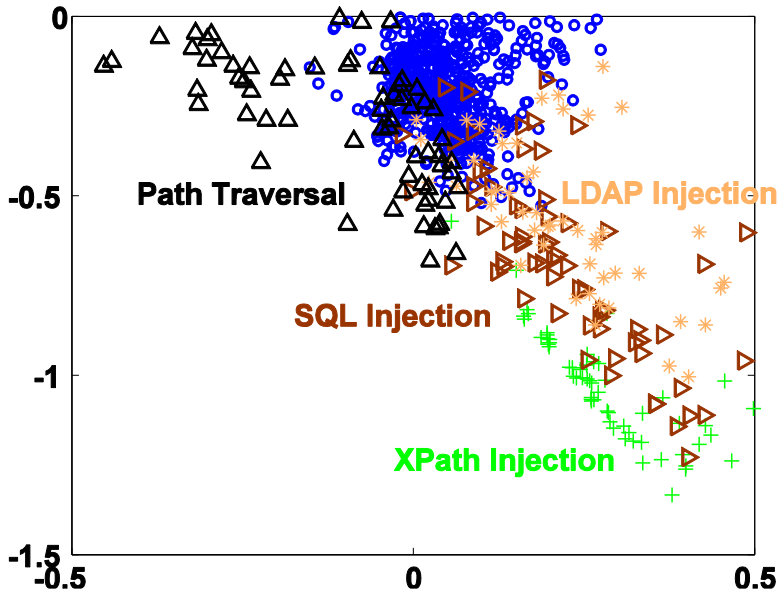
MODEL SELECTION

- Overfitting or underfitting
  - Model complexity vs. data size
  - Model stability
- Correlation ≠ causality

Scales of Graphs

- It is very important to pay attention to the *scale* that you are using when you are plotting
- Fixing the scale in comparison!





Ten Most Common Mistakes/Issues by Data Engineers

DATA

- Acquiring labeled data
  - Crowdsourcing can solve the problems?
  - How to deal with big-data?
  - Anomaly detection
- Too over or too few pre-processing?
- Storing data in a single matrix or database

REPRESENTATION

- Feature engineering: automatic or hand-crafted
  - Deep learning? Dictionary learning? Topic modeling?

ESTIMATION

- Machine learning packages or toolboxes ≠ Software
  - Tuning is highly necessary
  - Shall I use SVM when begins to analyze a dataset?
- C/C++ or Matlab/Python/R?
  - Low-level languages may not guarantee efficiency

Ten Most Common Mistakes/Issues by Data Engineers (cont'd)

ESTIMATION

- How much domain knowledge is enough?
  - How to include domain knowledge in an integrated model?
- Statistics (estimation) vs. Computer science (algorithms)
  - Efficiency vs. effectiveness
- Machine learning (with some possible goal) vs. data mining (no clear goal)
  - How much to know about the data?
- Online methods (real-time modeling) vs. offline methods (non-real-time modeling)

EVALUATION

- Accuracy is not the only thing to care
  - Balanced dataset vs. imbalanced dataset
  - Cost-sensitive methods, downsamping/oversampling methods

Data vs. Models
在對時間遇到對的人

— Introduction to AI —

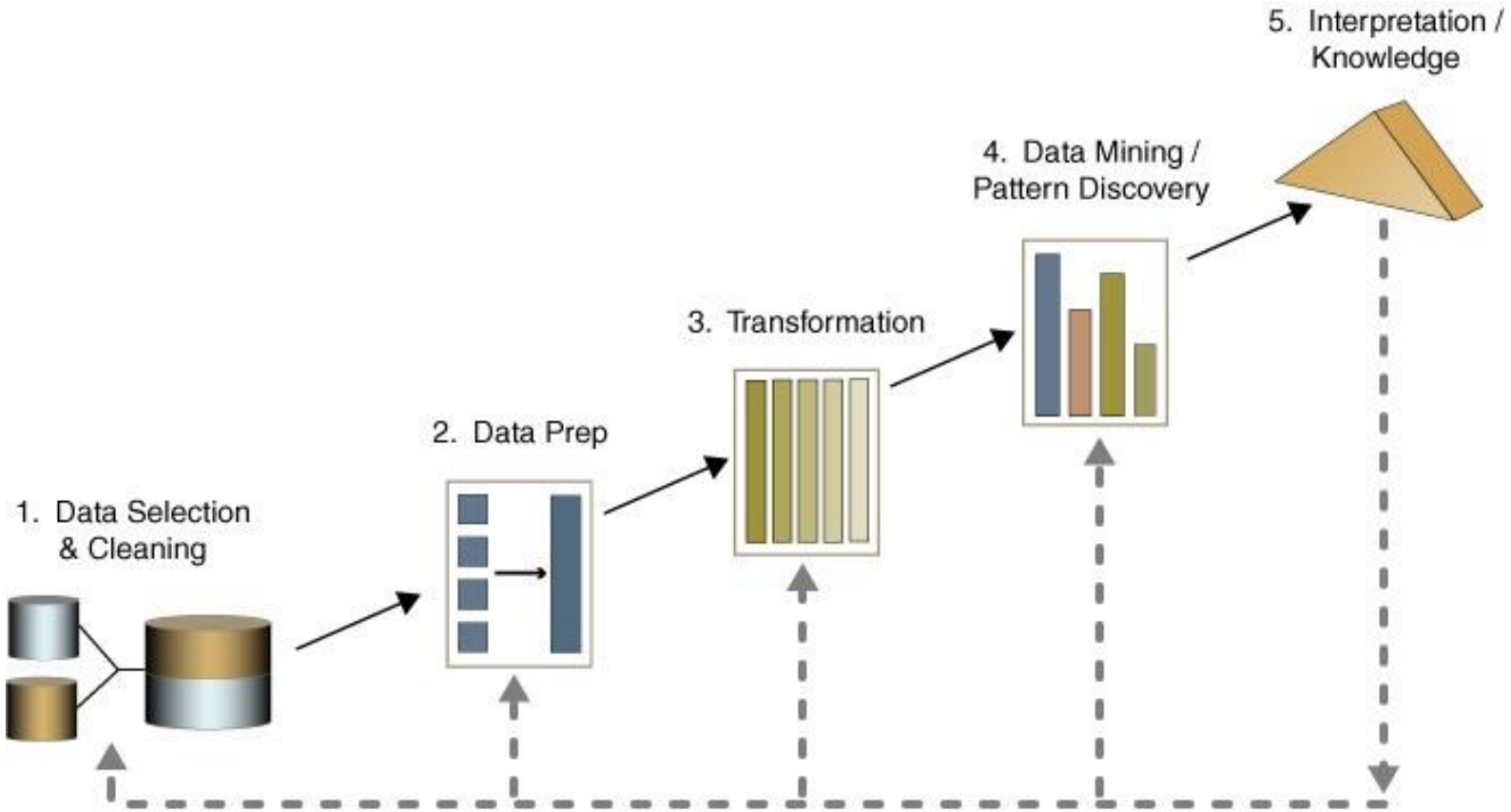
Data vs. Models

- Knowing the data is as important as knowing the models
- There is no best model for everything, but best data could be what we look for!
- The more data, the better model performance that we can expect (in these days)
- The more models that we know, the better model performance that we can expect (due to more capable of finding the right model for the data)

Data that Drives
推動搖籃的手

— Introduction to AI —

Data Drives Knowledge Discovery



- An overview of the steps that compose the knowledge discovery process

Welcome to the Sea of Data

- Values in the data
  - Nominal (categorical), ordinal, counting, numerical
- Types of data
  - A bag of data, sequential data or time series
  - Data w. or w/o dependencies
    - Deciding whether we can't or can use the iid assumption
    - Types of dependencies: Spatial and temporal, tree structured, graph structured, etc.
- Size and dimensionality of data
  - Big data techniques
  - Curse of dimensionality
- Quality of data:
  - How noise? Any outliers (systematic ones or non-systematic ones)?
  - Missing values?
  - Balanced or imbalanced dataset?
- Data statistics

Various Variable/Attribute Types

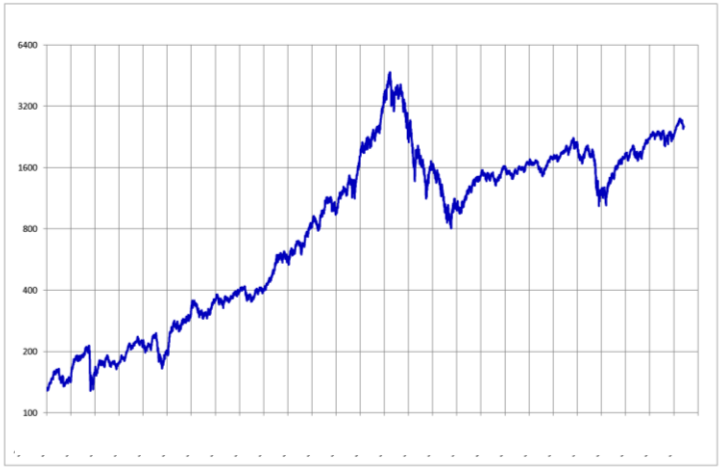
- **Nominal (categorical)**
  - Unordered set (values are names)
  - Operators: =, ≠
  - Example: gender, car origin (Europe, USA, Asia)
- Ordinal
  - Possess a natural order (values are numbers or names)
  - Operators: =, ≠, <, >
  - Example: ratings, school grades
- Counting
  - Just like natural numbers (values are numbers)
  - Operators: =, ≠, <, >, +, –
  - Example: # of childbirths
- **Numerical**
  - Allow for arithmetic operations (values are numbers)
  - Operators: =, ≠, <, >, *, /, +, -
  - Example: weight, acceleration in seconds
- Also subtypes exist: e.g., quantitative geographic (geographic coordinates), quantitative time

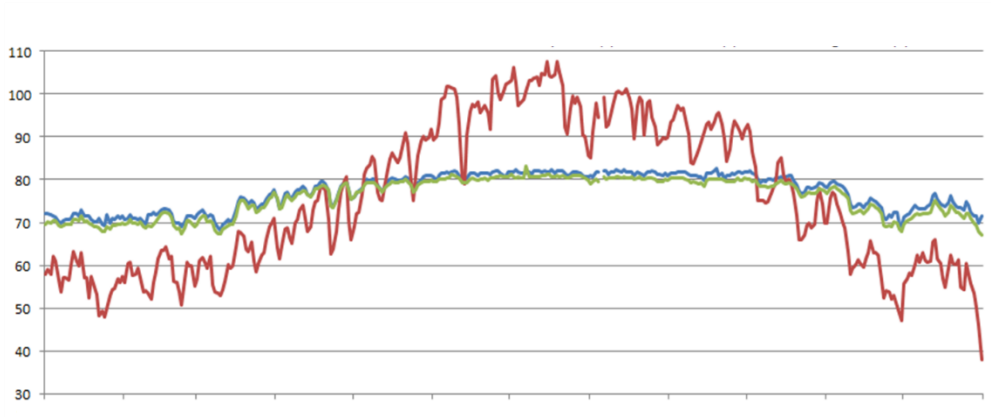
Types of Data

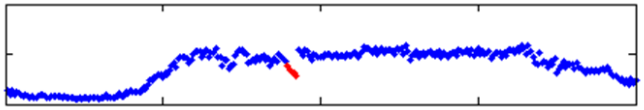
- A bag of data, sequential data or time series
- Data w. or w/o dependencies
  - Deciding whether we can't or can use the iid assumption

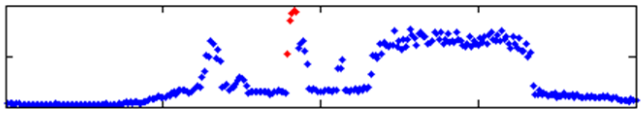
  - Types of dependencies: Spatial and temporal, tree structured, graph structured, etc.

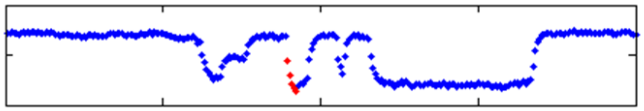
Some Examples of Sequential Data





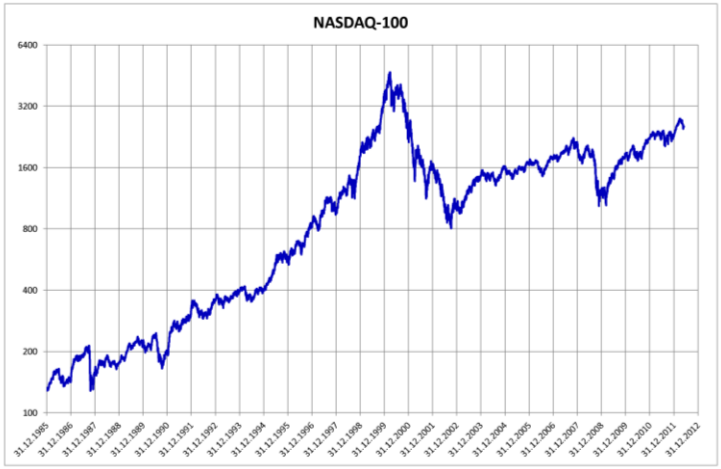




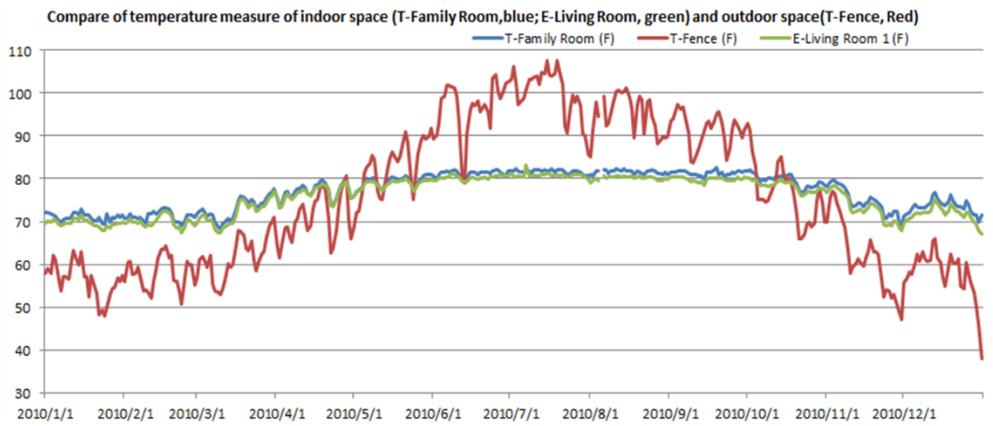


ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGTGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC

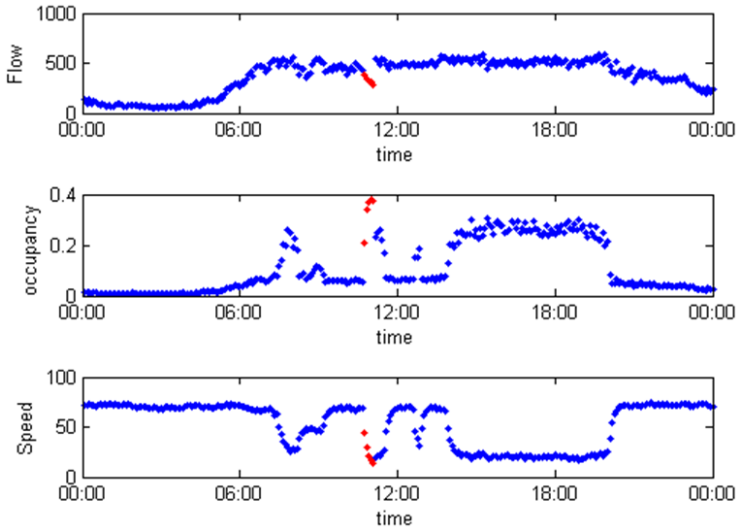
Some Examples of Sequential Data (cont'd)



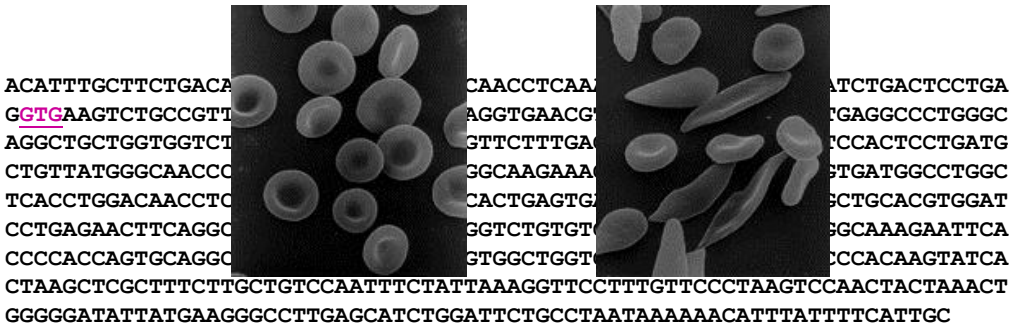
NASDAQ – 100 (1985 – 2012)



Smart Home Data (Mary's home)



Intelligent Transportation System Data



Gene of individual with Sickle Cell Anemia

How to Characterize the Data?

Yes, all are sequential, but…

- Is time involved?
- Is time discrete or continuous?
- What are the data attributes? How many of them?
- What the attributes types are? Categorical or numerical?

- Strong dependency between data in a previous index and later index? How strong it is?
- Is there a beginning and an end in the sequence?
- All attributes are observable? Any hidden information?
- Are the data "generated" by a single source? Any phase transitions? Any anomalies in the data?
- …

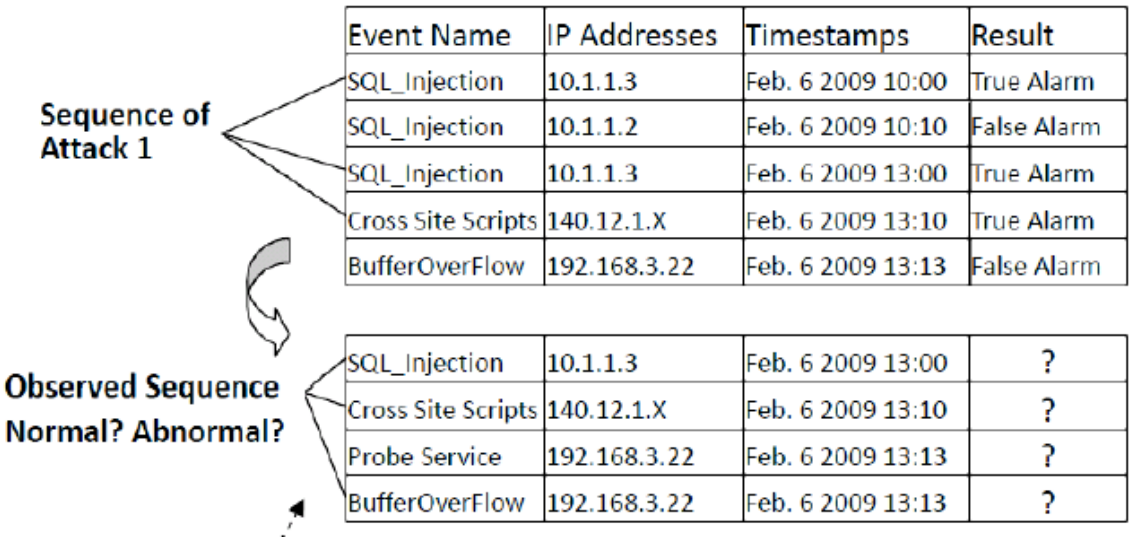
Learning Problems on Various Sequential Data: More Examples

	With Time	Without Time
Discrete Data (with gap)	Intrusion detection based on a series of events like alerts	Part-Of-Speech tagging, prediction function of biological sequence, hyphenation
Continuous Data (without gaps)	Voice recognition, stock market analysis	Analysis on 1-D or 2-D geographical or imaging data

- From now on, we focus mainly on discrete data
- "Discrete" or "continuous" is in the mathematical sense, not referring to "nominal" or "numerical", respectively!

Intrusion Detection based on Alerts

- Given an alert sequence
- Find an attack or anomaly behavior that is associated with a series of alerts.



Event Name	IP Addresses	Timestamps	Result
SQL_Injection	10.1.1.3	Feb. 6 2009 10:00	True Alarm
SQL_Injection	10.1.1.2	Feb. 6 2009 10:10	False Alarm
SQL_Injection	10.1.1.3	Feb. 6 2009 13:00	True Alarm
Cross Site Scripts	140.12.1.X	Feb. 6 2009 13:10	True Alarm
BufferOverFlow	192.168.3.22	Feb. 6 2009 13:13	False Alarm

SQL_Injection	10.1.1.3	Feb. 6 2009 13:00	?
Cross Site Scripts	140.12.1.X	Feb. 6 2009 13:10	?
Probe Service	192.168.3.22	Feb. 6 2009 13:13	?
BufferOverFlow	192.168.3.22	Feb. 6 2009 13:13	?

Part-of-Speech Tagging

- Given an English sentence, can we assign a part of speech to each word?

- “Do you want fries with that?”
- <verb pron verb noun prep pron>

Information Extraction from the Web

- <dl><dt>Srinivasan Seshan (Carnegie Mellon University) <dt>Making Virtual Worlds Real<dt>Tuesday, June 4, 2002<dd>2:00 PM , 322 Sieg<dd>Research Seminar
- * * * name name * * affiliation affiliation affiliation * * * * title title title title * * * date date date date * time time * location location * event-type event-type

Hyphenation

- “Porcupine” ! “001010000”

Welcome to the AI/ML Models

韓信點兵

— Introduction to AI —

We Got Friends

- AI/ML models are not the only tools that we can use for data analysis
- Statistics can also be considered "models", but relatively "static"

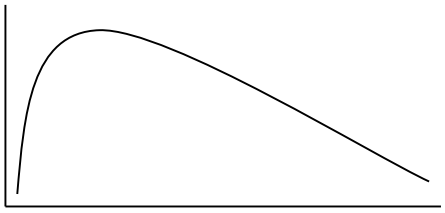
→ AI/ML models are more "dynamic" in the sense of:

  - Using algorithms
  - Finding solution in a converging approach!

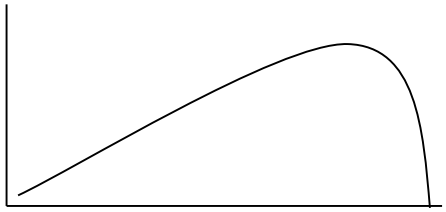
- Other techniques other than AI/ML models may also help:
  - Visualization
  - Human machine interaction
  - Crowdsourcing ("工人智慧")

Statistic to Summarize Data

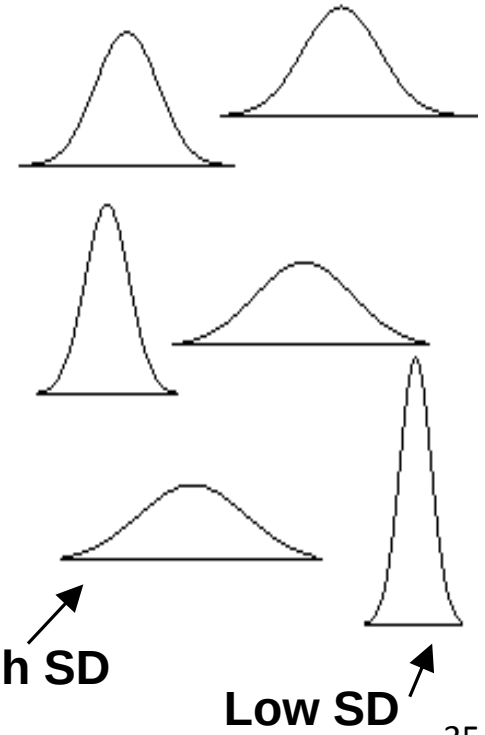
- A statistic is a number summarizing a bunch of values.
  - Simple or univariate statistics summarize values of one variable.
  - Effect statistics summarize the relationship between values of two or more variables.
- Simple statistics for numeric variables…
  - Mean, standard deviation, median, …
- How to deal with multi-dimensional data?



Positive skew

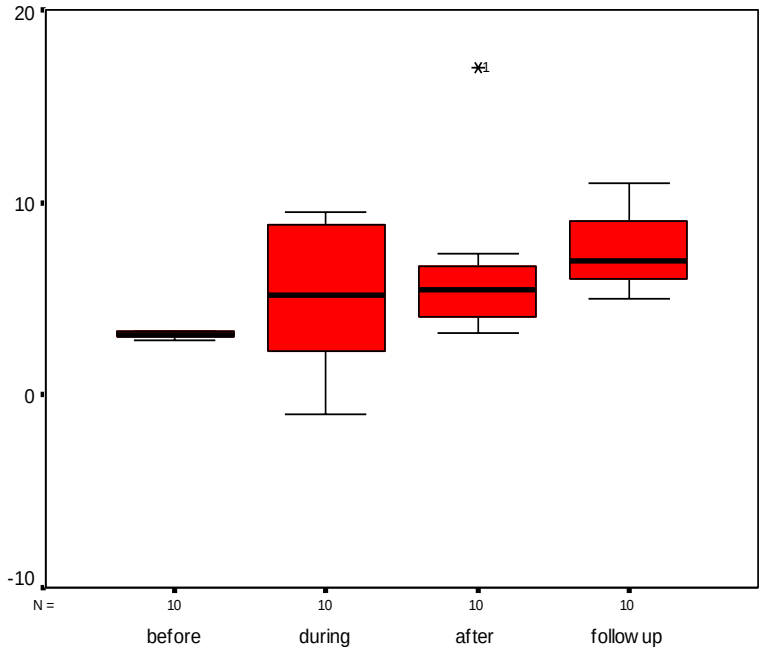


Negative skew



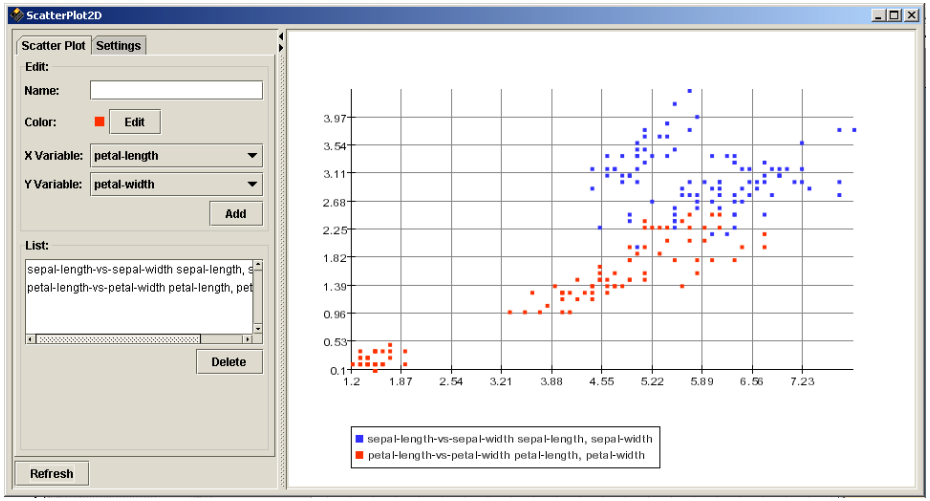
Boxplots

- Upper and lower bounds of boxes are the 25th and 75th percentile (interquartile range)
- Whiskers are *min* and *max* value unless there is an outlier
- An outlier is beyond 1.5 times the interquartile range (box length)

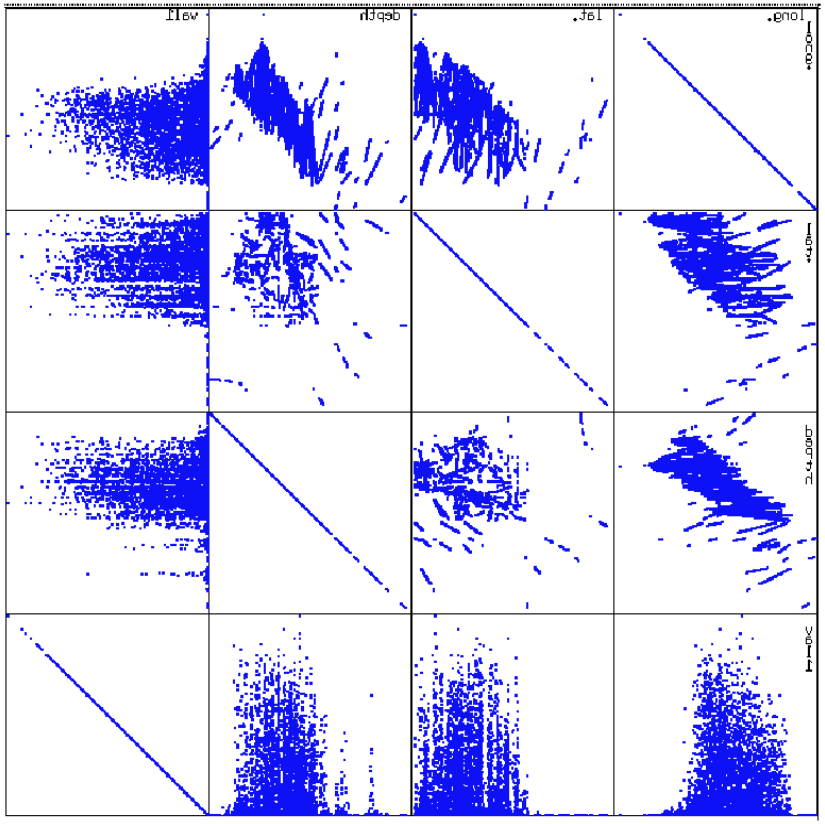


Scatterplots

- A scatterplot of Iris data

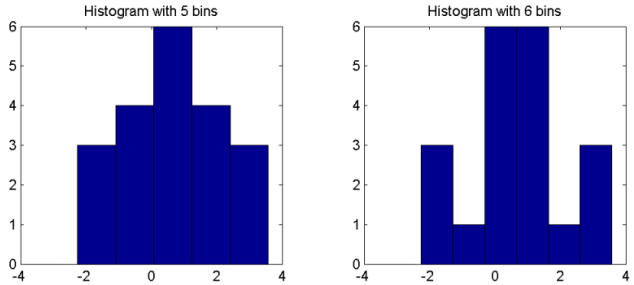


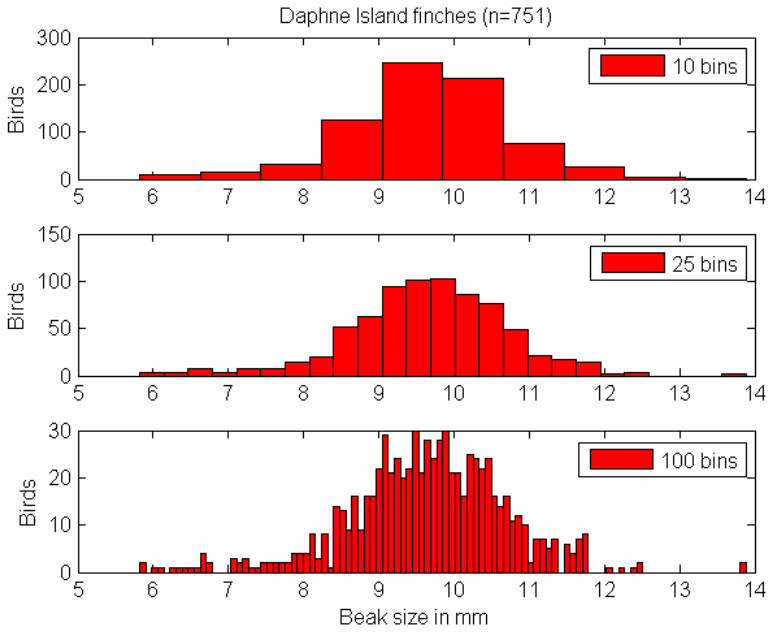
- An example of scatterplot matrix (x-y-diagrams) of a k-dimension data.
- Total of $(k^2 - k)$ or $(k^2/2 - k)$ scatterplots if not including the redundancy

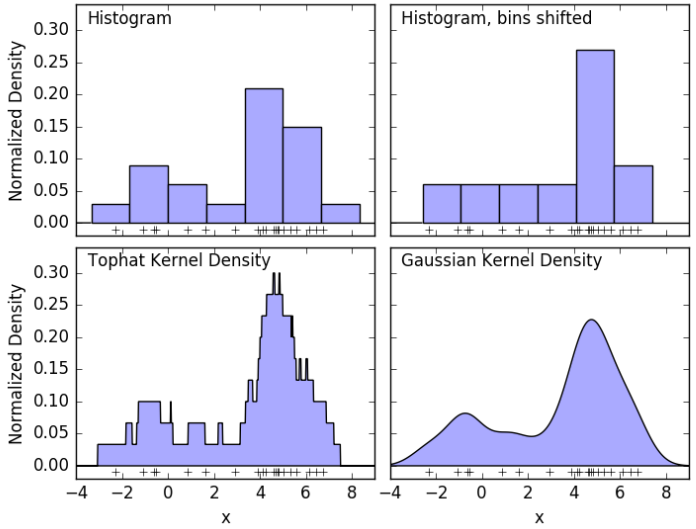


Histograms as Frequency Analysis

- # of bins is an important parameter!

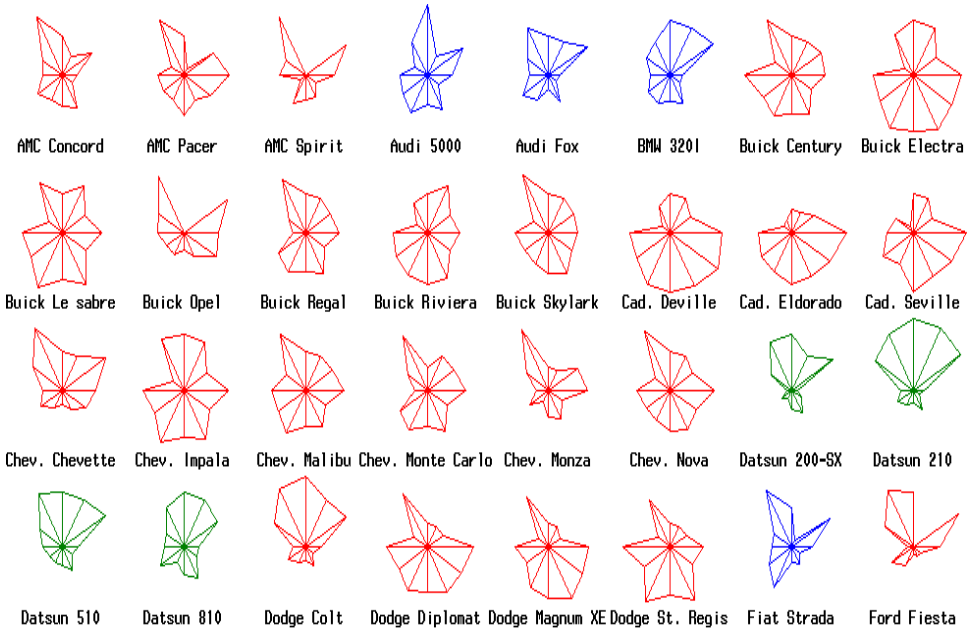


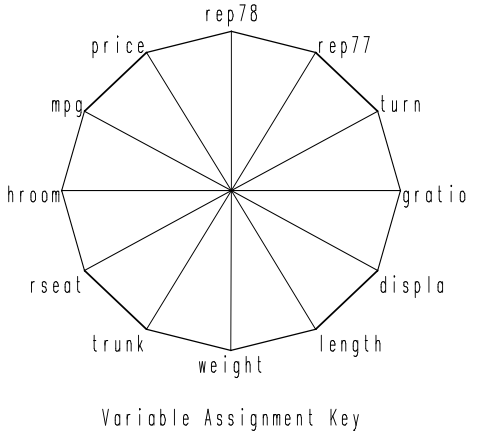




Star Plots (Chambers et al., 1983)

- Each data record is represented as a star-shaped figure with one ray for each variable
- The length of each ray is proportional to the value of its corresponding variable
- Each variable is usually normalized to between a very small number (close to 0) and 1
- The open ends of the rays are usually connected with lines





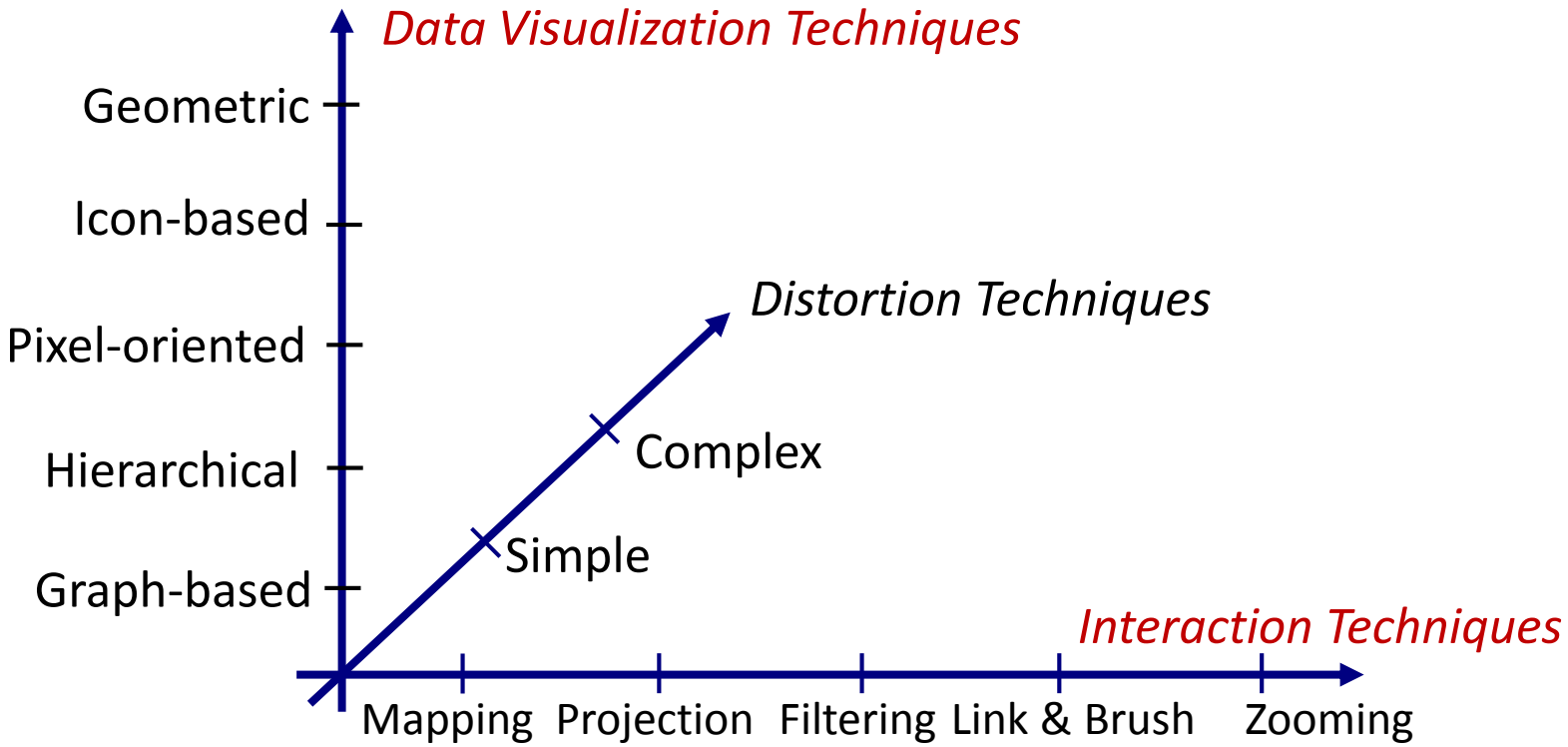
Star plots representation of an auto dataset with 12 variables

Major Categories of Visualization Techniques

- Geometric Techniques: Scatterplots, Landscapes, Projection Pursuit, Prosection Views, Hyperslice, Parallel Coordinates,...
- Icon-based Techniques: Chernoff Faces, Stick Figures, Shape-Coding, Color Icons, TileBars, ...
- Pixel-oriented Techniques: Recursive Pattern Technique, Circle Segments Technique, Spiral- & Axes-Techniques, ...
- Hierarchical Techniques: Dimensional Stacking, Worlds-within-Worlds, Treemap, Cone Trees, InfoCube, ...
- Graph-Based Techniques: Basic Graphs (Straight-Line, Poly-line, Curved-Line), Specific Graphs (e.g., DAG, Symmetric, Cluster), Systems (e.g., Tom Sawyer, Hy+, SeeNet, Narcissus), …
- Hybrid Techniques: arbitrary combinations from above

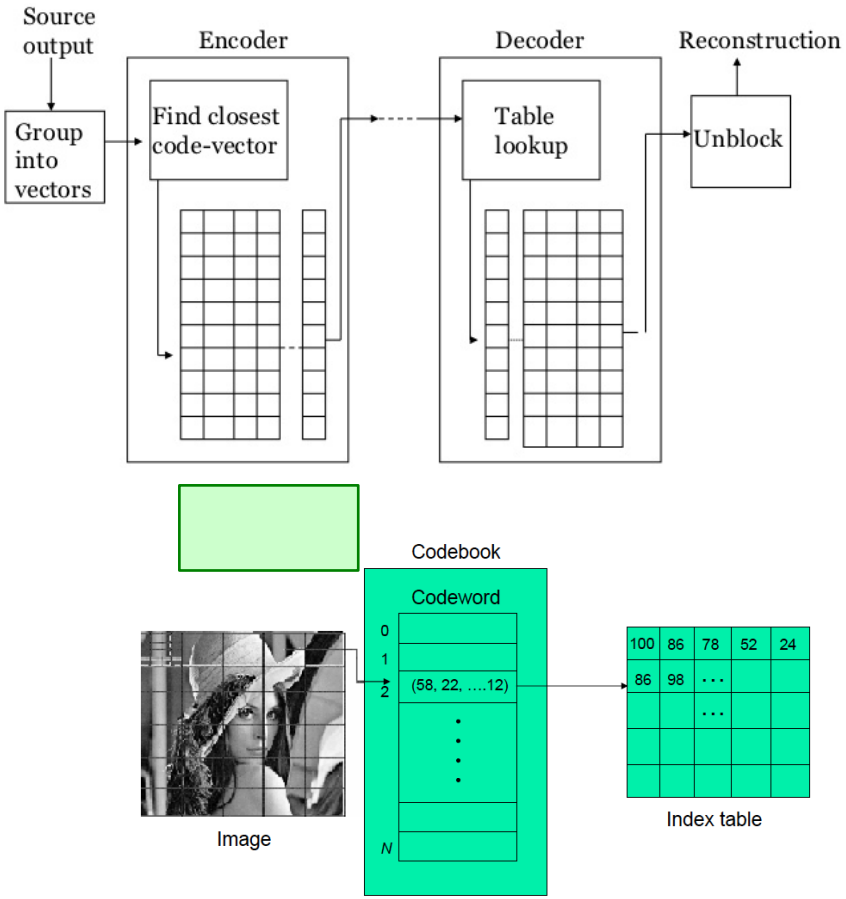
Data Visualization Methods

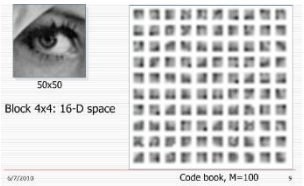
- There's no best visualization methods, but methods that fit to the applications
- Don't be fooled by the visualization result!
- Visualization result may still need objective criteria for evaluation



Vector Quantization

- A lossy compression method
- From analog (numerical) to digital (categorical)
- Some candidate methods
  - Voronoi diagram
  - K-means
  - Gaussian mixtures





Using Models

- Starting with simple models first
  - Some simple statistics may be enough to solve the problem if it is easy!
- Moving from simple models to complex models should be extremely careful!
  - Occam's razor
  - Minimum Description Length
  - You can't let the modeling go forever!

Types of AI/ML Models

- Supervised, unsupervised, semi-supervised, reinforcement learning
- Generative models vs. discriminative models
- Static (offline) models vs. dynamic (online) models
- Transparent models vs. black-box models
  - Models that incorporate with domain knowledge

Generative vs. Discriminative Models

- Algorithms that learn $P(y \mid \mathbf{x})$ directly (aka logistic regression) or algorithms that try to learn mappings directly from the space of inputs $\mathbf{x}$ to the labels $\{0, 1\}$ (such as LTU) are called discriminative models
- Instead, we can try to model $P(\mathbf{x} \mid y)$ and $P(y)$, called the generative model approach, because we can think of $P(\mathbf{x}, y)$ as a model of how the data is generated.
- For example, if y indicates whether an example is a giraffe (0) or an elephant (1), then $P(\mathbf{x} \mid y = 0)$ models the distribution of giraffes' features and $P(\mathbf{x} \mid y = 1)$ models the distribution of elephants' features.

Types of AI/ML Models

- Fill in gaps in existing knowledge
- Emulate the brain
- Simulate evolution
- Systematically reduce uncertainty
- Notice similarities between old and new

(by Pedro Domingos, U. of Washington)

Types of AI/ML Models

- Fill in gaps in existing knowledge → Symbolists
- Emulate the brain → Connectionists
- Simulate evolution → Evolutionaries
- Systematically reduce uncertainty → Bayesians
- Notice similarities between old and new → Analogizers

The Five Tribes of Machine Learning

Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

The Five Tribes of Machine Learning

Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

- Symbolists: Tom Mitchell, Steve Muggleton, Ross Quinlan
- Connectionists: Yann LeCun, Geoff Hinton, Yoshua Bengio
- Evolutionaries: John Koza, John Holland, Hod Lipson
- Bayesians: David Heckerman, Judea Pearl, Michael Jordan
- Analogizers: Peter Hart, Vladimir Vapnik, Douglas Hofstadter

An Introduction to Deep Learning (Deep Neural Networks)

— Introduction to AI —

Features of DNNs

- Stems from Artificial Neural Networks
- Some old ideas, but realized only recently given the modern computation facilities and breakthrough algorithms
- Deep structures: from 10 to 100 layers in the early stage to > 1000 layers in modern networks
- More than millions to billions of parameters: model complexity is enough to express complex concepts
- Conquered many field including the successful stories on vision, audio and language inputs

What Could be Powerful Machine Learning Algorithms?

- Deep nonlinearity: linear methods could be too simple to be effective for complex concepts
- Ensemble: working with a committee of methods
  - Example of ensemble methods: boosting (Adaboost), random forest, etc.
- Can deal with both labeled and unlabeled data, data with or without dependency
- Actionability: combining domain expert knowledge and model power
- Scalability: can deal with a large scale of data

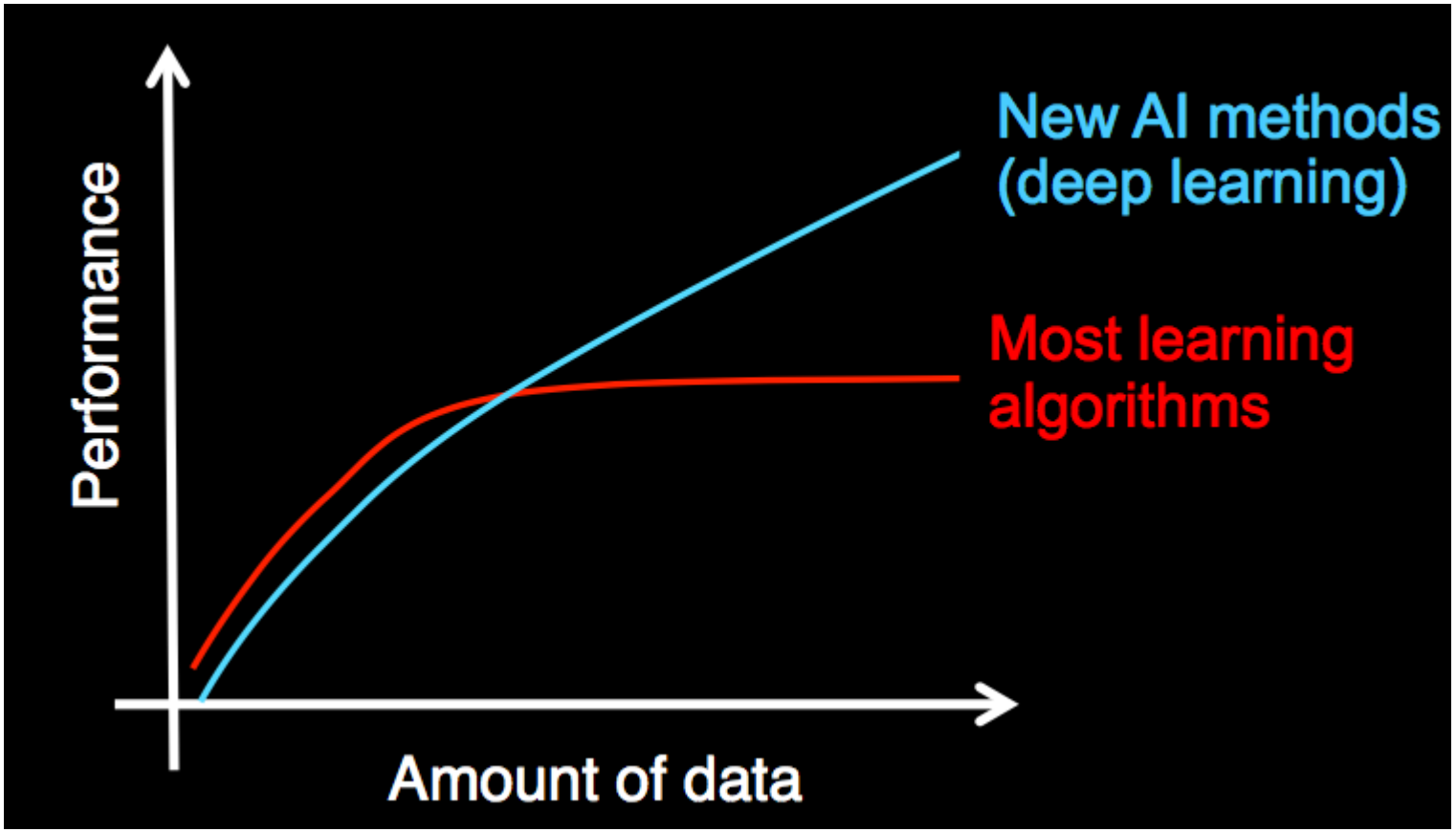
What Makes DNNs Successful?

- Can deal with highly non-linear problems
  - Researchers believe that images, videos, speech content are with highly nonlinear structures
  - Three nonlinear functions to choose from
- Combining unsupervised learning and supervised learning in a single framework
  - Autoencoders for feature learning
- A strategy to be considered as an ensemble approach: dropout
- Providing the possibility to unify generative modeling and discriminative modeling: Deep Bayesian Networks

Deep Learning History

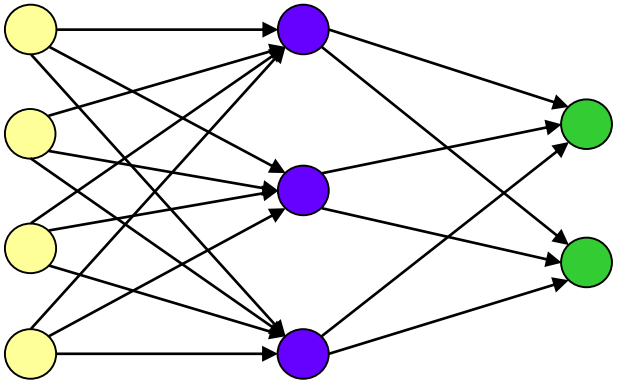
- Inspired by the architectural depth of the brain, researchers wanted for decades to train deep multi-layer neural networks
- No successful attempts were reported before 2006 ...
- Researchers reported positive results with typically one or two hidden layers, but training deeper networks consistently yielded poorer results.
- Exception: convolutional neural networks, LeCun 1998
- SVM: Vapnik and his co-workers developed the SVM in 1993, considered a shallow architecture method
- Digression: In the 1990's, many researchers abandoned neural networks with multiple adaptive hidden layers because SVMs worked better, and there was no successful attempts to train deep networks
- Breakthrough in 2006

Data and Machine Learning

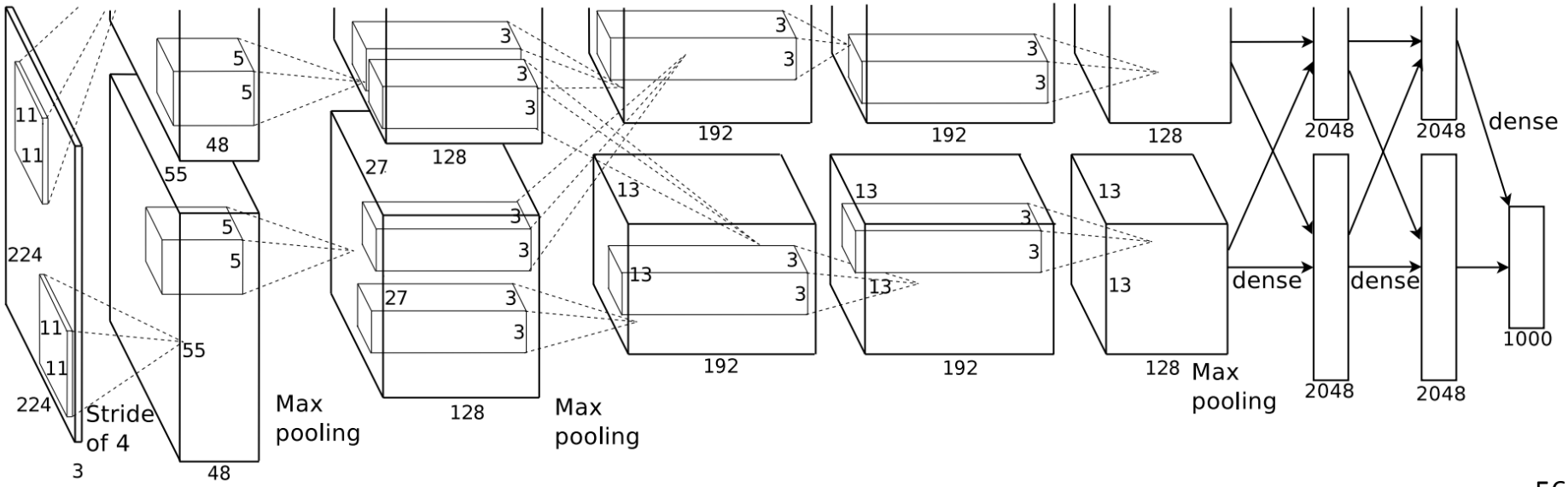


From ANNs to DNNs

- Traditional ANN



- A well-known CNN



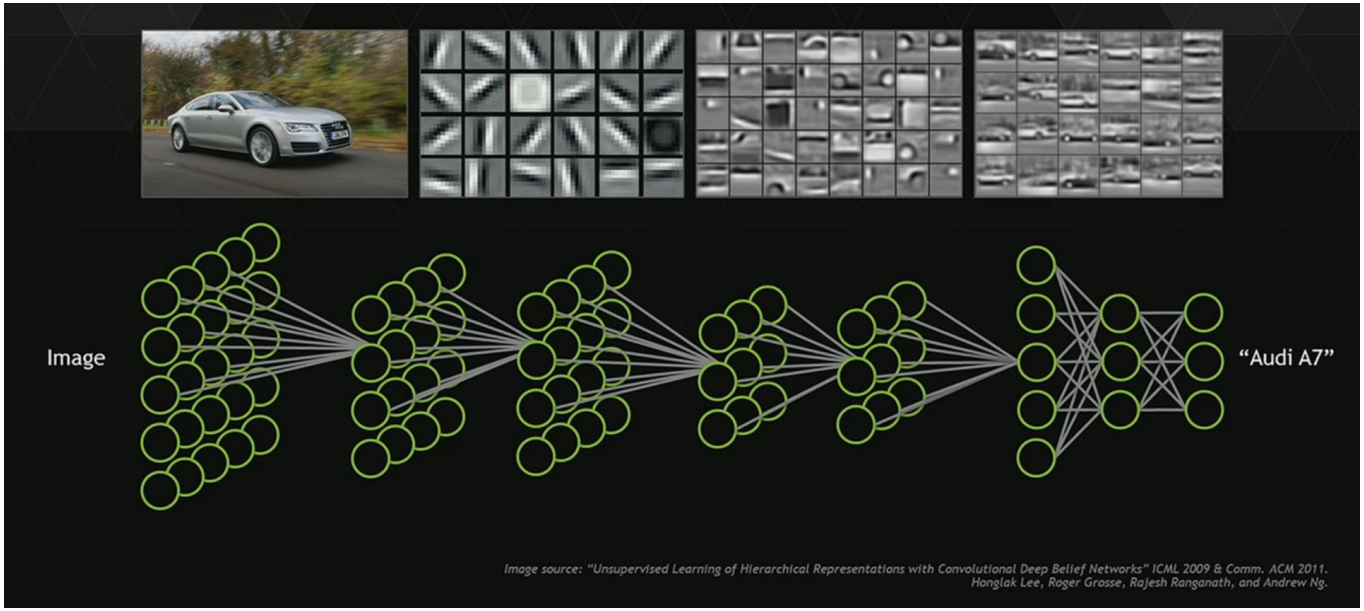
Motivation: Why Go Deep?

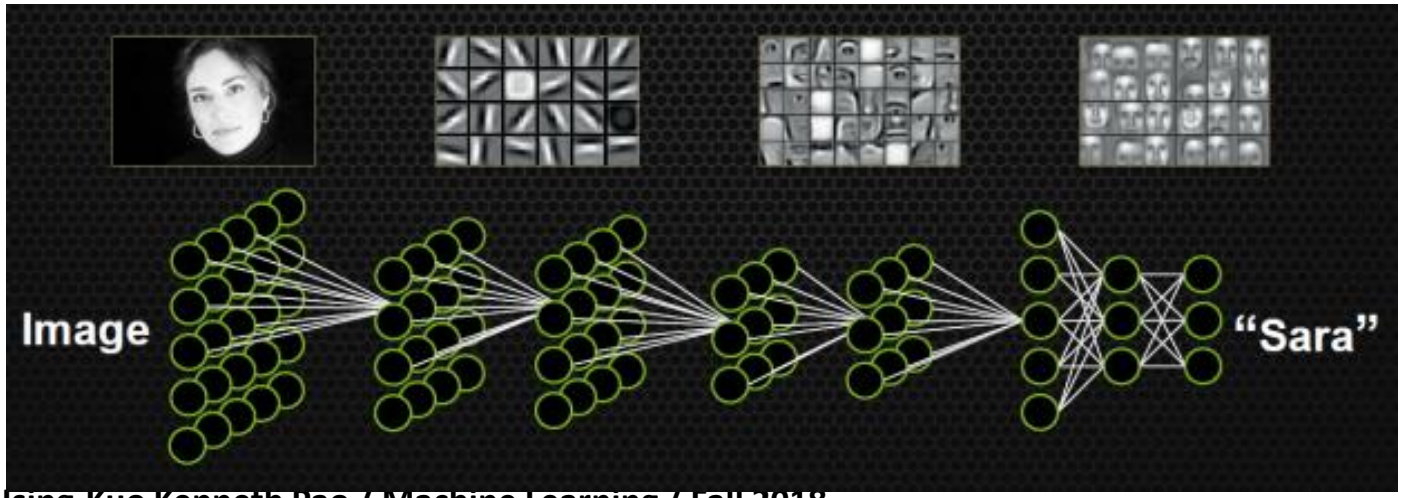
- Deep Architectures can be representationally efficient
  - Fewer computational units for same function
- Deep Representations might allow for a hierarchy or representation
  - Allows non-local generalization
  - Comprehensibility
- Multiple levels of latent variables allow combinatorial sharing of statistical strength
- Deep architectures work well (vision, audio, NLP, etc.)!

Methods other than DNN

- Most machine learning methods are with shallow structures
  - KNN, PCA, logistic regression
  - SVM: $f(\mathbf{x}) = \text{sign}\left(\sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i)\right)$

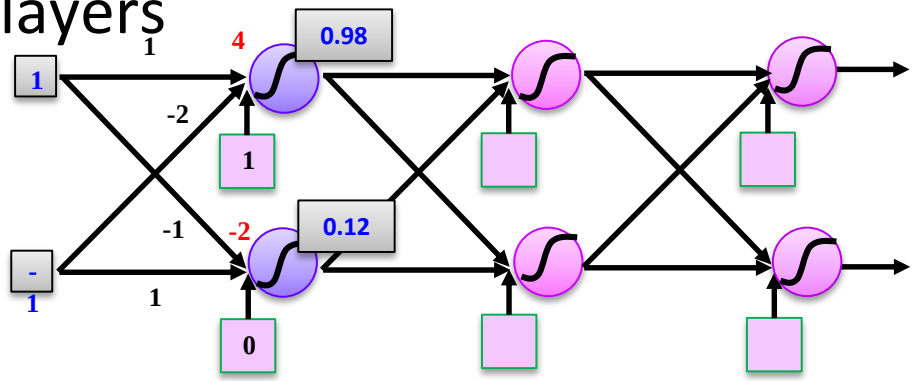
How a Deep Neural Network Sees



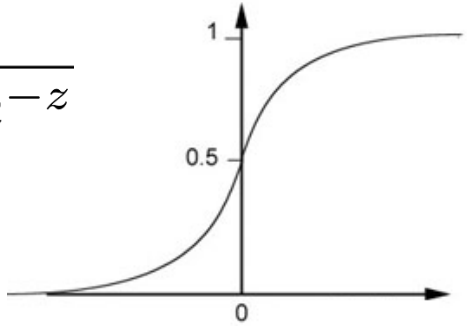


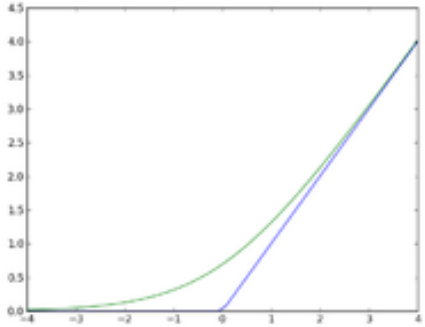
High Nonlinearity

- Networks with deep hidden layers



- Nonlinear functions after inner product computation in
  - ReLU (Rectified Linear Unit) $f(z) = \max(0, z)$
  - sigmoid $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
  - tanh $\tanh(z) = \dfrac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$





- Solving vision, audio needs highly nonlinear functions

q & a