From ML to ML: Two Disciplines, One Mission
– Making Sense of Imperfect Data

Grace Y. Yi

Canada Research Chair in Data Science (Tier 1)
Department of Statistical and Actuarial Sciences
Department of Computer Science
University of Western Ontario

SSC Gold Medal Address
Statistical Society of Canada Annual Meeting
June 2, 2026, Hamilton, Ontario, Canada

Western

ML = Maximum Likelihood **(Statistics)**

ML = Machine Learning **(AI)**

Western

ML = Maximum Likelihood **(Statistics)**

ML = Machine Learning **(AI)**

Two Disciplines, One Mission → Extracting Knowledge from Data

Western

Part 1: A Brief Comparison of Statistical Science and Machine Learning

What Is Machine Learning?

What Is Machine Learning?

- Wikipedia (accessed on 2026-04-22):
  Machine learning, a term introduced by Arthur Samuel (1959), is a field of study in artificial intelligence concerned with the development and study of
  - statistical algorithms that can learn from data and generalize to unseen data,
  - and thus perform tasks without explicit programming language instructions.

Western

What Is Machine Learning?

- Wikipedia (accessed on 2026-04-22):
  Machine learning, a term introduced by Arthur Samuel (1959), is a field of study in artificial intelligence concerned with the development and study of
  - statistical algorithms that can learn from data and generalize to unseen data,
  - and thus perform tasks without explicit programming language instructions.

- Mohri, Rostamizadeh and Talwalkar (2018):
  Machine learning can be broadly defined as computational methods using experience
  - to improve performance, or
  - to make accurate predictions

What Is Machine Learning?

- Wikipedia (accessed on 2026-04-22):
  Machine learning, a term introduced by Arthur Samuel (1959), is a field of study in artificial intelligence concerned with the development and study of
  - statistical algorithms that can learn from data and generalize to unseen data,
  - and thus perform tasks without explicit programming language instructions.

- Mohri, Rostamizadeh and Talwalkar (2018):
  Machine learning can be broadly defined as computational methods using experience
  - to improve performance, or
  - to make accurate predictions

- Goodfellow, Bengio and Courville (2016):
  Machine learning is essentially a form of applied statistics with
  - increased emphasis on using computers to estimate complicated functions, and
  - decreased emphasis on confidence intervals around these functions.

Western

Over-Simplified Analogy: Learning to Cook

Statistical Science: Cooking with a Recipe

Machine Learning: Cooking without a Recipe

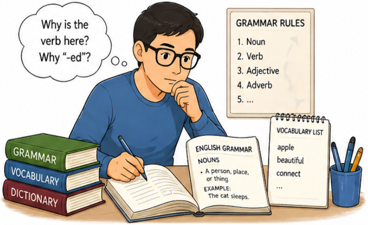
Create A New Dish:

guidance from recipes + experimentation

Western

Over-Simplified Analogy: Acquiring One's Mother Tongue

Statistical Science: Adult Learning



- Structured learning: textbooks, grammar rules, vocabulary lists
- Study rules and grammar
- Focus on understanding "why"

Machine Learning: Infant Learning

- No formal grammar lessons
- Learn from patterns
- Improve with exposure

Patterns + Rules

Learn A New Language:
pattern-based + rule-based learning

How do we learn from data and generalize beyond it?

Statistical Science

modeling + inference

uncertainty + interpretability

model → estimate → quantify uncertainty

Machine Learning

algorithms + optimization

prediction + scalability

represent → optimize → predict

	Statistical Science	Machine Learning

	Statistical Science	Machine Learning
Primary Emphasis	inference and uncertainty quantification	prediction and large–scale learning

Western

	Statistical Science	Machine Learning
Primary Emphasis	inference and uncertainty quantification	prediction and large–scale learning
Data Structure	structured, curated typically numerical	unstructured, digital images, text, speech, graphs

Western

	Statistical Science	Machine Learning
Primary Emphasis	inference and uncertainty quantification	prediction and large–scale learning
Data Structure	structured, curated typically numerical	unstructured, digital images, text, speech, graphs
Parameter Dimension $p = O(10^N)$	small or moderate $N \approx 0$–4 e.g., regression	high or very high $N \approx 5$–12 e.g., neural networks

Western

	Statistical Science	Machine Learning
Primary Emphasis	inference and uncertainty quantification	prediction and large–scale learning
Data Structure	structured, curated typically numerical	unstructured, digital images, text, speech, graphs
Parameter Dimension $p = O(10^N)$	small or moderate $N \approx 0$–4 e.g., regression	high or very high $N \approx 5$–12 e.g., neural networks
Sample Size $n = O(10^N)$	small to moderate $N \approx 1$–5 e.g., clinical trials, surveys	large to massive $N \approx 5$–15 e.g., ImageNet, language

Western

	Statistical Science	Machine Learning

	Statistical Science	Machine Learning
Framework & Scope	frequentist/Bayesian parametric/semiparametric /nonparametric models	supervised/unsupervised /semi-supervised learning reinforcement learning

Western

	Statistical Science	Machine Learning
Framework & Scope	frequentist/Bayesian parametric/semiparametric /nonparametric models	supervised/unsupervised /semi-supervised learning reinforcement learning
Complexity	model/sample	model/sample computation/storage

Western

	Statistical Science	Machine Learning
Framework & Scope	frequentist/Bayesian parametric/semiparametric /nonparametric models	supervised/unsupervised /semi-supervised learning reinforcement learning
Complexity	model/sample	model/sample computation/storage
Identifiability	central requirement	may not be explicitly enforced

Western

	Statistical Science	Machine Learning
Framework & Scope	frequentist/Bayesian parametric/semiparametric /nonparametric models	supervised/unsupervised /semi-supervised learning reinforcement learning
Complexity	model/sample	model/sample computation/storage
Identifiability	central requirement	may not be explicitly enforced
Interpretability	often high	often low

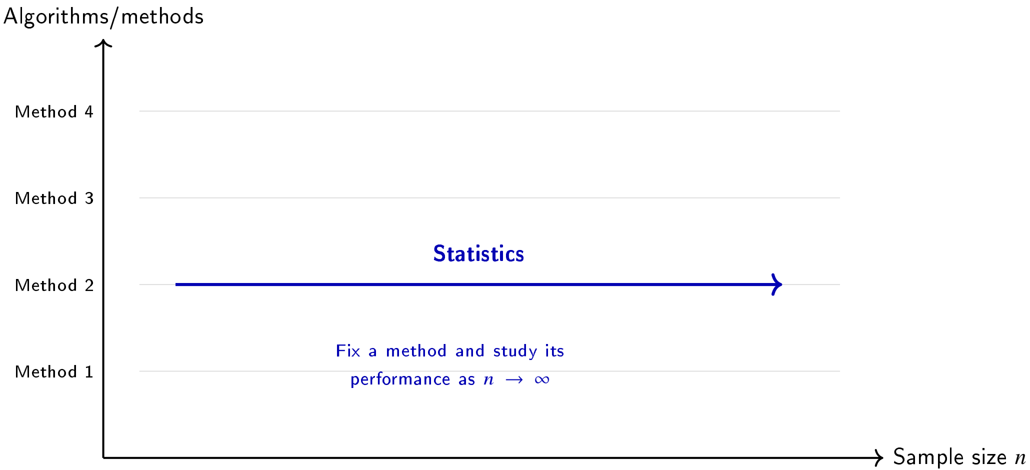
	Statistical Science	Machine Learning
Framework & Scope	frequentist/Bayesian parametric/semiparametric /nonparametric models	supervised/unsupervised /semi-supervised learning reinforcement learning
Complexity	model/sample	model/sample computation/storage
Identifiability	central requirement	may not be explicitly enforced
Interpretability	often high	often low
Terminology	X: covariate/predictor Y: response/outcome	X: input/feature/instance Y: output/label

Western

Validity Justification: Statistics vs Machine Learning

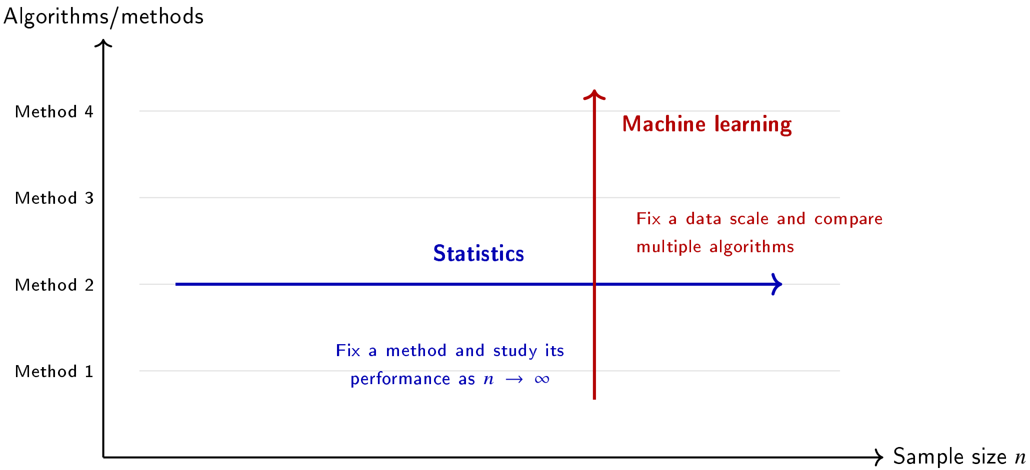
How do we justify that a method works?

Validity Justification: Statistics vs Machine Learning



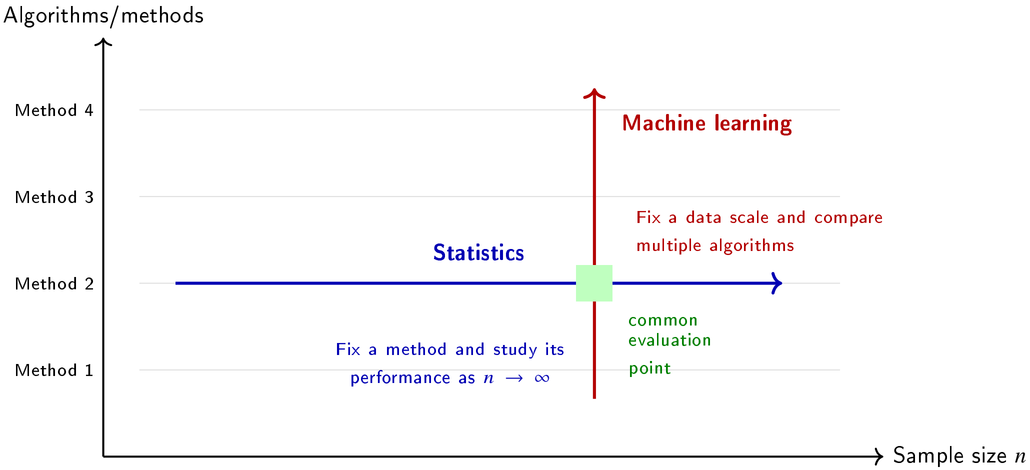
Statistics: often assess a method by considering the hypothetical scenario with $n \to \infty$

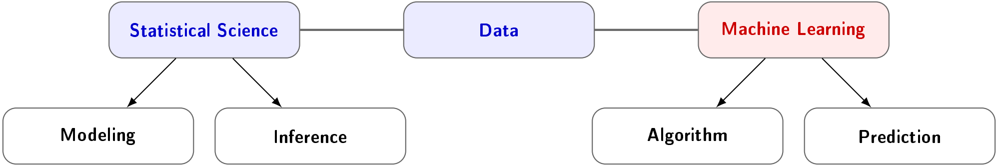
Validity Justification: Statistics vs Machine Learning

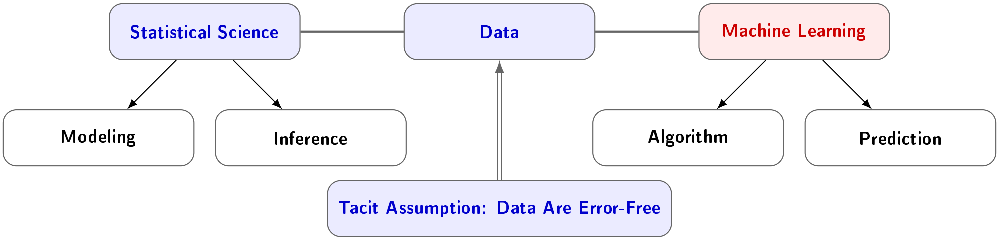


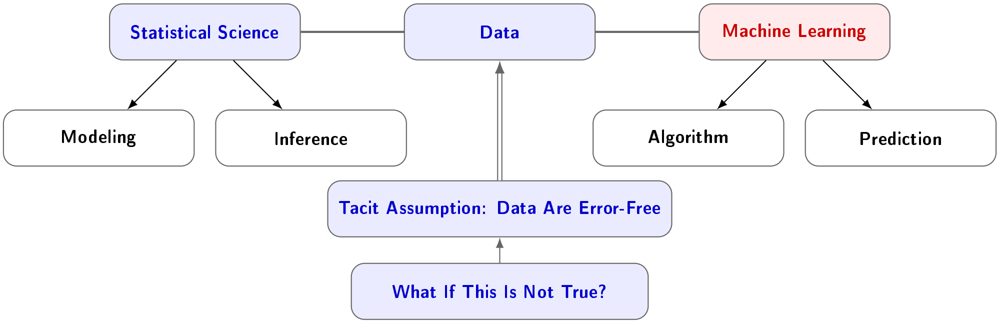
Machine Learning: often compare different methods/algorithms in a hypothetical class by fixing n

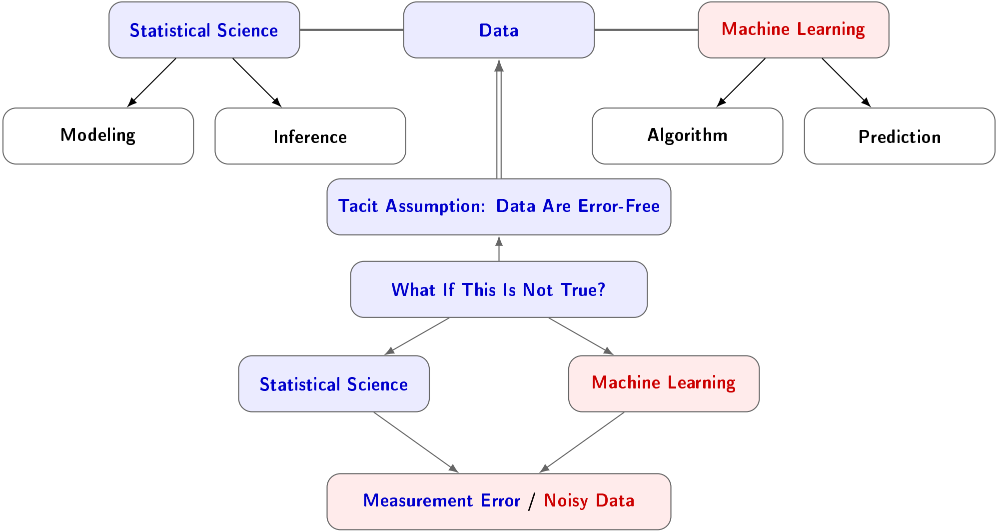
Validity Justification: Statistics vs Machine Learning

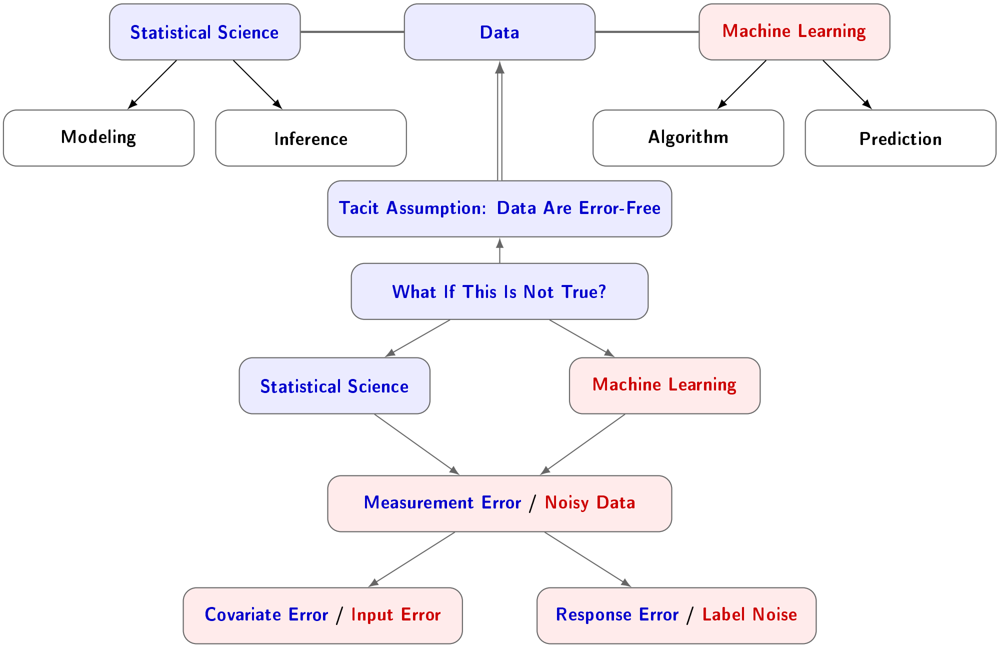



Statistical Science
Data
Machine Learning
Modeling
Inference
Algorithm
Prediction


Statistical Science
Data
Machine Learning
Modeling
Inference
Algorithm
Prediction
Tacit Assumption: Data Are Error-Free


Statistical Science
Data
Machine Learning
Modeling
Inference
Algorithm
Prediction
Tacit Assumption: Data Are Error-Free
What If This Is Not True?


Statistical Science
Data
Machine Learning
Modeling
Inference
Algorithm
Prediction
Tacit Assumption: Data Are Error-Free
What If This Is Not True?
Statistical Science
Machine Learning
Measurement Error / Noisy Data


Statistical Science
Data
Machine Learning
Modeling
Inference
Algorithm
Prediction
Tacit Assumption: Data Are Error-Free
What If This Is Not True?
Statistical Science
Machine Learning
Measurement Error / Noisy Data
Covariate Error / Input Error
Response Error / Label Noise

Part 2: Examples of Measurement Error and Noisy Data

Western

Measurement Error in Statistical Analysis

Western

Example: Reporting Error (Bollinger 1998)

- Reporting errors are typical in survey data. For example, it was found that
    - response error was negatively related to earnings
    - about 10% of the reporters grossly over-reported their income
    - There was greater measurement error among men than among women.
    - for men, a nonlinear relationship between reported earnings and true earnings existed, but for women the relationship was linear

Example: Reporting Error (Bollinger 1998)

- Reporting errors are typical in survey data. For example, it was found that
  - response error was negatively related to earnings
  - about 10% of the reporters grossly over-reported their income
  - There was greater measurement error among men than among women.
  - for men, a nonlinear relationship between reported earnings and true earnings existed, but for women the relationship was linear

- Remarks
  - Response error in income cannot be treated as additive white noise because of its relationship with gender and true earnings.
  - Measurement error is not related to age, education, and weeks worked.

Example of Blood Pressure (BP) – Long-Term Average Quantity

- The determination of BP is subject to machine and reader variability
  $\Longrightarrow$ reading error

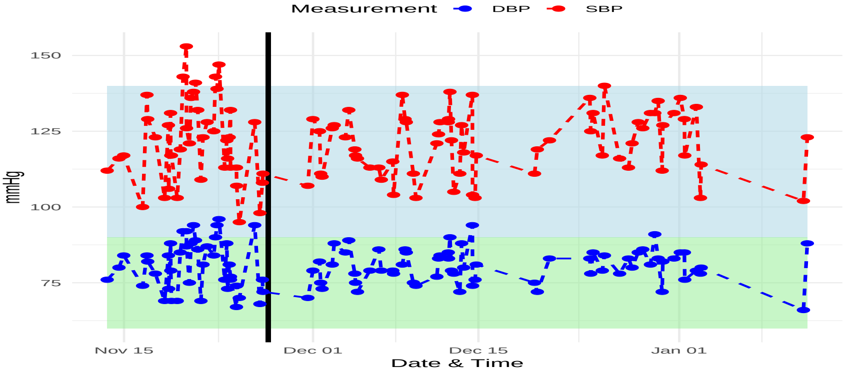
Example of Blood Pressure (BP) – Long-Term Average Quantity

- The determination of BP is subject to machine and reader variability
  $\Longrightarrow$ reading error
- Any specific measurement fluctuates around the true long term average
  $\Longrightarrow$ biological variability

Western

Example of Blood Pressure (BP) – Long-Term Average Quantity

- The determination of BP is subject to machine and reader variability
  $\Longrightarrow$ reading error
- Any specific measurement fluctuates around the true long term average
  $\Longrightarrow$ biological variability



Example: No Direct Measurement in Epidemiological Studies

- Some variables are difficult or even impossible to be measured directly.
- Common Examples:
  - body composition (total body fat or lean mass)
  - dietary intakes
  - physical activity
  - resting energy expenditure

Example: No Direct Measurement in Epidemiological Studies

- Some variables are difficult or even impossible to be measured directly.
- Common Examples:
  - body composition (total body fat or lean mass)
  - dietary intakes
  - physical activity
  - resting energy expenditure

Practical Procedure

- In the absence of direct measurements of some predictors, using surrogate measurements as a replacement seems to be a must.

Warning

- Addressing the induced error is important!

Western

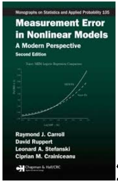
can we?

Data with Measurement Error	$\Longrightarrow$	Useful Conclusions

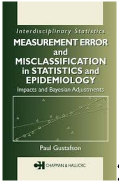
Western

can we?

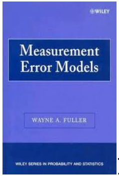
Data with Measurement Error $\Longrightarrow$ Useful Conclusions



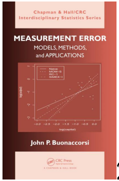
2006



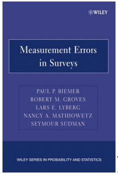
2004



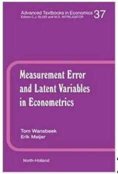
1987



2010



1991

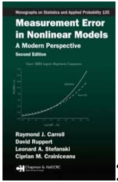


2001

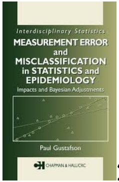
Western

can we?

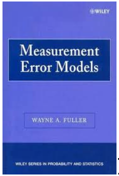
Data with Measurement Error	$\Longrightarrow$	Useful Conclusions



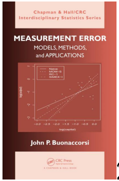
2006



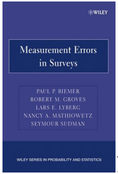
2004



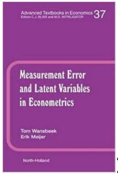
1987



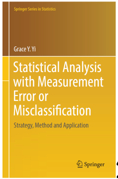
2010



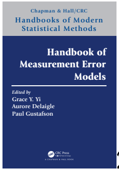
1991



2001



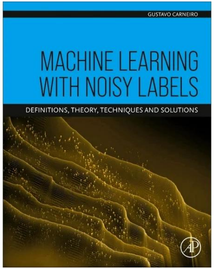
2017



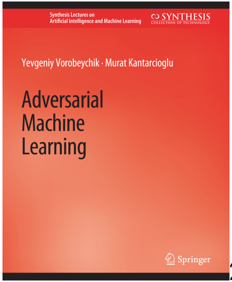
2021

Western

Corrupted Data in Machine Learning



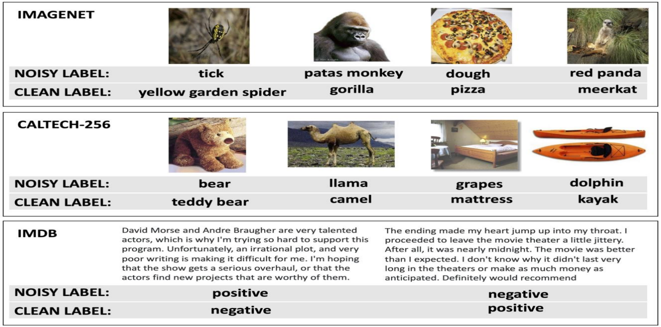
2024



2022

Western

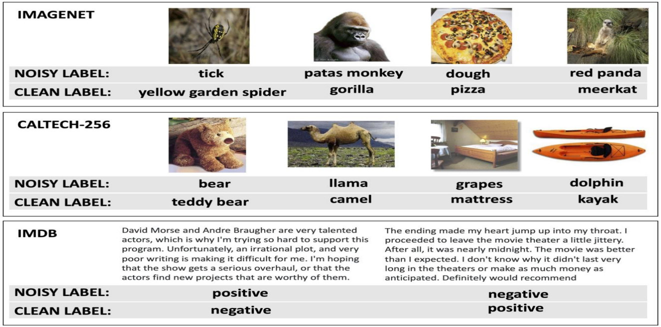
Image Recognition/Text Classification (Source: Carneiro 2024)



Western

Dataset	Label Source	Cleanliness	Noise Source
ImageNet	human annotation via AMT + WordNet	high (in ILSVRC subset)	fine-grained confusion, human error
Caltech-256	manually grouped object categories	relatively clean	background clutter, overlapping categories
IMDB	derived from user rating scores	moderate	heuristic mismatch with sentiment, sarcasm, ambiguity

Image Recognition/Text Classification (Source: Carneiro 2024)



Medical Image

- Medical image labeling needs to be done by radiologists, which is expensive and time-consuming.

Medical Image

- Medical image labeling needs to be done by radiologists, which is expensive and time-consuming.
- A hospital-scale chest X-ray (CXR) dataset (Wang et al., 2017) with over 100,000 images was labeled using natural language processing (NLP) to automatically extract multiple disease labels from the associated radiology reports.

Medical Image

- Medical image labeling needs to be done by radiologists, which is expensive and time-consuming.
- A hospital-scale chest X-ray (CXR) dataset (Wang et al., 2017) with over 100,000 images was labeled using natural language processing (NLP) to automatically extract multiple disease labels from the associated radiology reports.
  - This dataset has a non-negligible proportion of label noise because of
    - the NLP mistakes
    - the difficulty of diagnosing some diseases from CXR images (Oakden-Rayner 2020)

Medical Image

- Medical image labeling needs to be done by radiologists, which is expensive and time-consuming.
- A hospital-scale chest X-ray (CXR) dataset (Wang et al., 2017) with over 100,000 images was labeled using natural language processing (NLP) to automatically extract multiple disease labels from the associated radiology reports.
  - This dataset has a non-negligible proportion of label noise because of
    - the NLP mistakes
    - the difficulty of diagnosing some diseases from CXR images (Oakden-Rayner 2020)

Corrupted Labels

- The ratio of corrupted labels in real-world datasets was reported to range from 8.0% to 38.5% (Song et al. 2023).
- Menze et al. (2014) reported average inter-reader variability of 74%–85% for glioblastoma (a type of brain tumour) segmentation.

Western

Part 3: How Do Statistical Methods and Machine Learning Techniques Interact?

(Some Illustrative Work)

Application 1: Noisy Label in Medical Imaging (Wang, Yi & Cao 2025)

- pneumoconiosis diagnosis from chest X-rays
- high inter-/intra-reader variability

Application 1: Noisy Label in Medical Imaging (Wang, Yi & Cao 2025)

- pneumoconiosis diagnosis from chest X-rays
- high inter-/intra-reader variability

Data Feature

- 1,042 X-rays (NIOSH)
- 1–21 readers; scale 0–3

Label Noise

- Proxy ground truth via consensus
- Mislabeling: 21% (normal), 42% (abnormal)

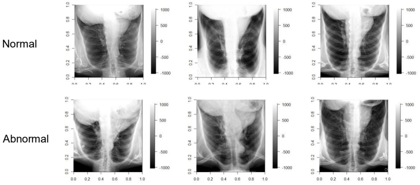


Figure 4: Chest image samples for normal individuals (top row) and abnormal individuals (bottom row).

Application 1: Noisy Label in Medical Imaging (Wang, Yi & Cao 2025)

- pneumoconiosis diagnosis from chest X-rays
- high inter-/intra-reader variability

Data Feature

- $1{,}042$ X-rays (NIOSH)
- 1–21 readers; scale 0–3

Label Noise

- Proxy ground truth via consensus
- Mislabeling: 21% (normal), 42% (abnormal)

Our Method

ML + Statistics

- ML: afDNN learns representations
- Statistics: model annotator noise
- estimate true label distribution

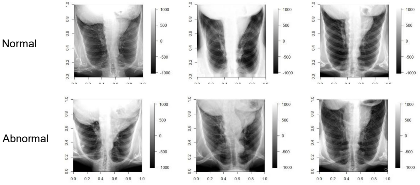


Figure 4: Chest image samples for normal individuals (top row) and abnormal individuals (bottom row).

Western

Application 2: Label Noise in Single-Cell Data (He, Yi & Hu 2026)

- Thousands of genes across hundreds of thousands of cells
- Goal: identify cell types and states

Application 2: Label Noise in Single-Cell Data (He, Yi & Hu 2026)

- Thousands of genes across
  hundreds of thousands of cells
- Goal: identify cell types and states

Sources of Label Noise

- Manual annotation: subjective, expertise-dependent
- Automated methods:
  – different algorithms $\Rightarrow$ different labels
  – inconsistent cluster numbers and assignments

Key Issue

- No reliable ground truth labels

Application 2: Label Noise in Single-Cell Data (He, Yi & Hu 2026)

- Thousands of genes across hundreds of thousands of cells
- Goal: identify cell types and states

Sources of Label Noise

- Manual annotation: subjective, expertise-dependent
- Automated methods:
  – different algorithms ⇒ different labels
  – inconsistent cluster numbers and assignments

Key Issue

- No reliable ground truth labels

Our Method

ML + Statistics

- ML: DNN architecture
- Statistics: treat algorithm–generated labels as crowdsourced annotations
- quantify uncertainty via conformal prediction

Multiple methods → multiple labels
⇓
Crowdsourced noisy labels
⇓
Uncertainty-aware prediction

Supervised Learning/Classification

- In supervised learning, each data point is associated with a label/annotation

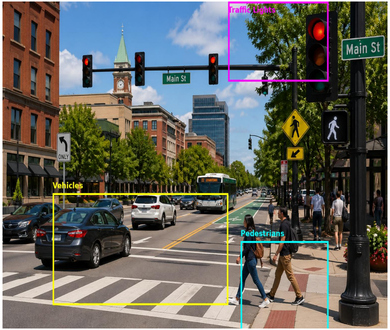
Western

Supervised Learning/Classification

- In supervised learning, each data point is associated with a label/annotation
- Image classification often requires a large amount of labeled data.
  - Acquiring accurately annotated data is expensive, time–consuming, or even impossible.

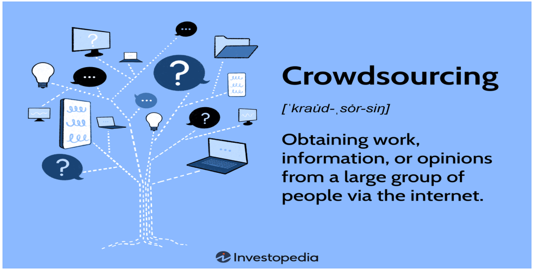
Supervised Learning/Classification

- In supervised learning, each data point is associated with a label/annotation
- Image classification often requires a large amount of labeled data.
  - Acquiring accurately annotated data is expensive, time–consuming, or even impossible.



Crowdsourcing

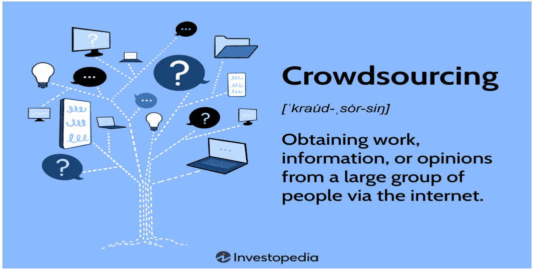
- Crowdsourcing provides an efficient and inexpensive way to collect labels.



Source: https://www.google.com

Crowdsourcing

- Crowdsourcing provides an efficient and inexpensive way to collect labels.

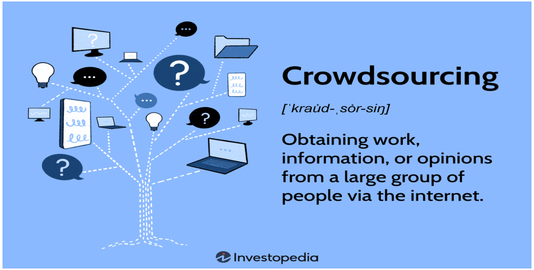


Source: https://www.google.com

- Example: The Amazon Mechanical Turk platform can obtain hundreds of thousands of labels in just a few hours from crowdsourcing workers worldwide, with each label costing only a few cents.

Crowdsourcing

- Crowdsourcing provides an efficient and inexpensive way to collect labels.



Source: https://www.google.com

- Example: The Amazon Mechanical Turk platform can obtain hundreds of thousands of labels in just a few hours from crowdsourcing workers worldwide, with each label costing only a few cents.

Issues and Challenges

- Annotators have varying and imperfect labeling skills, yielding unavoidably noisy labels.

Western

Joint Work with Hui Guo and Boyu Wang (NeurIPS, 2023)

Setup

$$\mathcal{X} \subset \mathbb{R}^p : \quad \text{feature space}$$

$$\mathcal{Y} = \{1, \ldots, K\} \triangleq [K] : \quad \text{label space}$$

$$\mathbf{x}_i \in \mathcal{X} : \quad \text{input data}$$

$$\mathrm{y}_i \in \mathcal{Y} : \quad \text{unobserved true label}$$

Western

Setup

$\mathcal{X} \subset \mathbb{R}^p$: feature space

$\mathcal{Y} = \{1, \ldots, K\} \triangleq [K]$: label space

$\mathbf{x}_i \in \mathcal{X}$: input data

$\mathrm{y}_i \in \mathcal{Y}$: unobserved true label

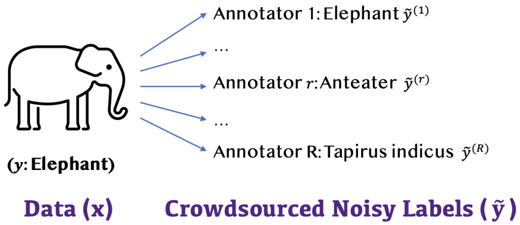
Crowdsourcing

each item is labeled by multiple annotators

$$\mathcal{D} = \{\mathbf{x}_i, \widetilde{\mathrm{y}}_i^{(1)}, \ldots, \widetilde{\mathrm{y}}_i^{(R)}\}_{i=1}^{N}$$

R : #of annotators

$\widetilde{\mathrm{y}}_i^{(r)}$: label given by $r$thannotator



Setup

$$\mathcal{X} \subset \mathbb{R}^p : \quad \text{feature space}$$

$$\mathcal{Y} = \{1, \ldots, K\} \triangleq [K] : \quad \text{label space}$$

$$\mathbf{x}_i \in \mathcal{X} : \quad \text{input data}$$

$$\mathrm{y}_i \in \mathcal{Y} : \quad \text{unobserved true label}$$

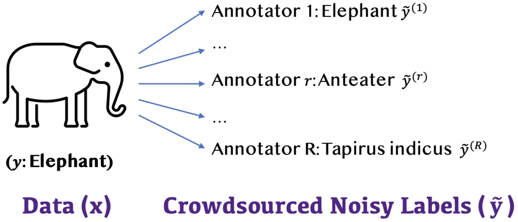
Crowdsourcing

each item is labeled by multiple annotators

$$\mathcal{D} = \{\mathbf{x}_i, \widetilde{\mathrm{y}}_i^{(1)}, \ldots, \widetilde{\mathrm{y}}_i^{(R)}\}_{i=1}^{N}$$

$$R : \quad \#\text{of annotators}$$

$$\widetilde{\mathrm{y}}_i^{(r)} : \quad \text{label given by } r\text{th}\text{annotator}$$



Objective

use data with noisy labels to train a classifier to correctly label new input data

Assumption

R annotators independently label instances

Assumption

R annotators independently label instances

Noisy Label Probability

$$\mathbb{P}(\widetilde{\mathrm{y}}^{(1)}, .., \widetilde{\mathrm{y}}^{(R)} | \mathbf{x}) = \prod_{r=1}^{R} \mathbb{P}(\widetilde{\mathrm{y}}^{(r)} | \mathbf{x})$$

$$= \prod_{r=1}^{R} \sum_{k \in \mathcal{Y}} \left\{ \mathbb{P}(\widetilde{\mathrm{y}}^{(r)} | \mathrm{y} = k, \mathbf{x}) P(\mathrm{y} = k | \mathbf{x}) \right\}$$

Western

Assumption

R annotators independently label instances

Noisy Label Probability

$$\mathbb{P}(\widetilde{\mathrm{y}}^{(1)}, .., \widetilde{\mathrm{y}}^{(R)} | \mathbf{x}) = \prod_{r=1}^{R} \mathbb{P}(\widetilde{\mathrm{y}}^{(r)} | \mathbf{x})$$

$$= \prod_{r=1}^{R} \sum_{k \in \mathcal{Y}} \{\mathbb{P}(\widetilde{\mathrm{y}}^{(r)} | \mathrm{y} = k, \mathbf{x}) P(\mathrm{y} = k | \mathbf{x})\}$$

transition matrix for the rth annotator

base model (true label predictor)

Assumption

R annotators independently label instances

Noisy Label Probability

$$\mathbb{P}(\widetilde{y}^{(1)}, .., \widetilde{y}^{(R)}|\mathbf{x}) = \prod_{r=1}^{R} \mathbb{P}(\widetilde{y}^{(r)}|\mathbf{x})$$

$$= \prod_{r=1}^{R} \sum_{k \in \mathcal{Y}} \left\{ \mathbb{P}(\widetilde{y}^{(r)}|y = k, \mathbf{x}) P(y = k|\mathbf{x}) \right\}$$

transition matrix for the rth annotator

base model (true label predictor)

Remark

- Most available methods require the instance independent assumption:

$$\mathbb{P}(\widetilde{y}^{(r)}|y = k, \mathbf{x}) = \mathbb{P}(\widetilde{y}^{(r)}|y = k) \text{ for each } r$$

Western

Proposed Modeling of Instance-Dependent Transition Matrices

Western

Objective

flexibly reflect two characteristics in the annotation process:

- possibly different effects of the annotator expertise (r)
- influence of the ground truth of class (k)

Western

Objective

flexibly reflect two characteristics in the annotation process:

- possibly different effects of the annotator expertise (r)
- influence of the ground truth of class (k)

Neural Networks

$$\widetilde{\mathrm{y}}^{(r)} \mid \{\mathrm{y} = k, \mathbf{x}\} \sim G\{\mathbf{A}_0^{(r)}\psi_1(\mathbf{x}) + \mathbf{B}_0^{(k)}\psi_2(\mathbf{x})\}$$

Western

Objective

flexibly reflect two characteristics in the annotation process:

- possibly different effects of the annotator expertise (r)
- influence of the ground truth of class (k)

Neural Networks

$$\widetilde{\mathrm{y}}^{(r)} \mid \{\mathrm{y} = k, \mathbf{x}\} \sim G\{\mathbf{A}_0^{(r)} \psi_1(\mathbf{x}) + \mathbf{B}_0^{(k)} \psi_2(\mathbf{x})\}$$

where

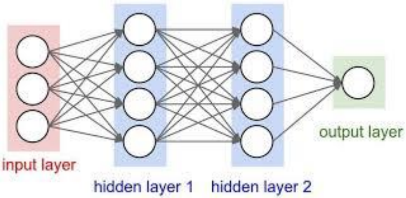
$\mathbf{A}_0^{(r)}$ and $\mathbf{B}_0^{(k)}$: regression weights

$\psi_1(\mathbf{x})$ and $\psi_2(\mathbf{x})$: unknown functions

G : softmax function

Modeling $\psi_j(\mathbf{x})$ with Sparse Bayesian DNNs

With $j = 1, 2,$



Modeling $\psi_j(\mathbf{x})$ with Sparse Bayesian DNNs

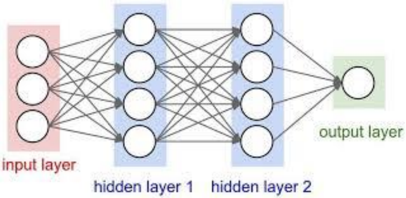
With $j = 1, 2,$

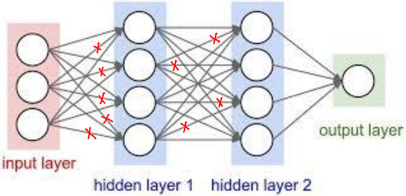
$$\psi_j(\mathbf{x}; \boldsymbol{\theta}^{(j)}) = \mathbf{W}^{(jH_j)}\sigma\left(\ldots\sigma\left[\mathbf{W}^{(jh)}\sigma\left\{\ldots\sigma(\mathbf{W}^{(j1)}\mathbf{x} + \mathbf{b}^{(j1)})\ldots\right\} + \mathbf{b}^{(jh)}\right]\ldots\right) + \mathbf{b}^{(jH_j)}$$

$\mathbf{W}^{(jh)}$: weight matrix for hth layer

$\mathbf{b}^{(jh)}$: bias for hth layer

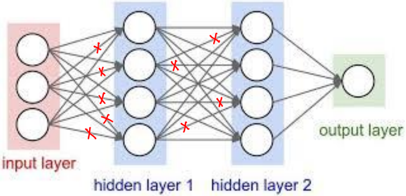
$\sigma(\cdot)$: activation function




input layer
hidden layer 1
hidden layer 2
output layer

Sparsity Indicator

$\gamma_l^{(j)} = I(\theta_l^{(j)} \neq 0)$ with associated parameter vector $\theta_l^{(j)}$

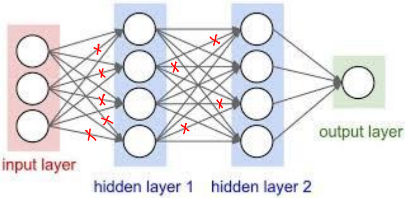


Sparsity Indicator

$\gamma_l^{(j)} = I(\theta_l^{(j)} \neq 0)$ with associated parameter vector $\theta_l^{(j)}$

Prior Distribution

$$\begin{aligned} \theta_l^{(j)} | \gamma_l^{(j)} &\sim \gamma_l^{(j)} \pi_1(\cdot; \sigma_j) + (1 - \gamma_l^{(j)}) \pi_0(\cdot; \alpha_j \sigma_j) \\ \gamma_l^{(j)} &\sim \text{Bernoulli}(\lambda) \end{aligned}$$



Key Idea

- Instance-dependent transition models can be learned by leveraging anchor points.

Detail

- $\mathbf{x}$ is an anchor point of class k if

$$\mathbb{P}(y = k|\mathbf{x}) = 1$$

Western

Key Idea

- Instance-dependent transition models can be learned by leveraging anchor points.

Detail

- $\mathbf{x}$ is an anchor point of class k if

$$\mathbb{P}(\mathrm{y} = k|\mathbf{x}) = 1$$
$$\implies \mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathrm{y} = k, \mathbf{x}) = \mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathbf{x})$$

Western

Key Idea

- Instance-dependent transition models can be learned by leveraging anchor points.

Detail

- $\mathbf{x}$ is an anchor point of class k if

$$\mathbb{P}(\mathrm{y} = k|\mathbf{x}) = 1$$
$$\Longrightarrow \ \mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathrm{y} = k, \mathbf{x}) = \mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathbf{x})$$

- anchor points and their noisy labels:

$$\overline{\mathcal{D}}_0 : \quad \cup_{k,i} \{\{\mathbf{x}_i, \widetilde{\mathbf{y}}_i\} : \mathbb{P}(\mathrm{y}_i = k|\mathbf{x}_i) = 1\}$$
$$n : \quad \text{size of } \overline{\mathcal{D}}_0$$

Western

Learning Transition Model Algorithm

- use majority voting to determine anchor points (Li et al. 2021; Liu & Tao 2016; He, Yi & Hu 2026)
- obtain estimates of the network parameters:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \log p_{\boldsymbol{\theta},i} + \log \pi(\boldsymbol{\theta}) \right\}$$

- determine the sparse network $\Longrightarrow$ resulting estimate for $\mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathrm{y} = k, \mathbf{x})$: $f_{\theta}^{(k,r)}$

Western

Learning Transition Model Algorithm

- use majority voting to determine anchor points (Li et al. 2021; Liu & Tao 2016; He, Yi & Hu 2026)
- obtain estimates of the network parameters:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \log p_{\boldsymbol{\theta},i} + \log \pi(\boldsymbol{\theta}) \right\}$$

- determine the sparse network $\Longrightarrow$ resulting estimate for $\mathbb{P}(\widetilde{\mathrm{y}}^{(r)}|\mathrm{y} = k, \mathbf{x})$: $f_{\theta}^{(k,r)}$

Posterior Probability Measure

$$\Pi(G|\overline{\mathcal{D}}_0) = \frac{\int_G p_{\boldsymbol{\theta}}^n(\overline{\mathcal{D}}_0) d\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^n(\overline{\mathcal{D}}_0) d\Pi(\boldsymbol{\theta})}$$

where

$p_{\boldsymbol{\theta}}^n$: probability density/mass function for $\overline{\mathcal{D}}_0$

$\Pi(\cdot)$: prior for $\boldsymbol{\theta}$

Western

Theorem 1

Let $d(\cdot,\cdot)$ denote the Hellinger distance. Under regularity conditions, there exists a sequence of constants $\{\epsilon_n^2\}_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty} \epsilon_n = 0 \text{ and } \lim_{n\to\infty} n\epsilon_n^2 = \infty$$

such that for $k \in [K]$ and $r \in [R]$, the posterior measure satisfies

$$\Pi\left\{\boldsymbol{\theta} \in \Theta : d(f_{\boldsymbol{\theta}}^{(k,r)}, f_0^{(k,r)}) > M_n\epsilon_n \,\middle|\, \overline{\mathcal{D}}_0\right\} \to 0 \text{ in probability as any } M_n \to \infty$$

Theorem 1

Let $d(\cdot,\cdot)$ denote the Hellinger distance. Under regularity conditions, there exists a sequence of constants $\{\epsilon_n^2\}_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty} \epsilon_n = 0 \text{ and } \lim_{n\to\infty} n\epsilon_n^2 = \infty$$

such that for $k \in [K]$ and $r \in [R]$, the posterior measure satisfies

$$\Pi\left\{\boldsymbol{\theta} \in \Theta : d(f_{\boldsymbol{\theta}}^{(k,r)}, f_0^{(k,r)}) > M_n\epsilon_n \,\middle|\, \overline{\mathcal{D}}_0\right\} \to 0 \text{ in probability as any } M_n \to \infty$$

Implications

- The estimated sparse noise transition model can closely approximate the true transition matrix in Hellinger distance.

Theorem 1

Let $d(\cdot,\cdot)$ denote the Hellinger distance. Under regularity conditions, there exists a sequence of constants $\{\epsilon_n^2\}_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty} \epsilon_n = 0 \text{ and } \lim_{n\to\infty} n\epsilon_n^2 = \infty$$

such that for $k \in [K]$ and $r \in [R]$, the posterior measure satisfies

$$\Pi\left\{\boldsymbol{\theta} \in \Theta : d(f_{\boldsymbol{\theta}}^{(k,r)}, f_0^{(k,r)}) > M_n\epsilon_n \,\middle|\, \overline{\mathcal{D}}_0\right\} \to 0 \text{ in probability as any } M_n \to \infty$$

Implications

- The estimated sparse noise transition model can closely approximate the true transition matrix in Hellinger distance.
- Theorem 1 extends existing results from the i.i.d. setting (e.g., Wang & Rocková 2020; Sun, Song & Liang 2021; Sun, Xiong & Liang 2021) to the i.n.i.d setting.

Western

Proposed Prediction Method
with Label-Noise Effects Accommodated

Prediction with Noisy Annotations

a new input $\mathbf{x}$ receives labels $\widetilde{\mathrm{y}} = (\widetilde{\mathrm{y}}^{(1)}, \ldots, \widetilde{\mathrm{y}}^{(R)})^{\mathrm{T}}$ from R annotators

$$k^* = \arg\max_{k \in [K]} \mathbb{P}(\mathrm{y} = k | \widetilde{\mathrm{y}}^{(r)}, \mathbf{x})$$

Western

Prediction with Noisy Annotations

a new input $\mathbf{x}$ receives labels $\widetilde{\mathrm{y}} = (\widetilde{\mathrm{y}}^{(1)}, \ldots, \widetilde{\mathrm{y}}^{(R)})^{\mathrm{T}}$ from R annotators

$$k^* = \arg\max_{k \in [K]} \mathbb{P}(\mathrm{y} = k | \widetilde{\mathrm{y}}^{(r)}, \mathbf{x})$$

Problem Reformulation

- Step 1:

$$k^* = \arg\max_{k \in [K]} \left\{ \hbar_k \prod_{r=1}^{R} \hat{\tau}^{(r)}_{k, \widetilde{\mathrm{y}}^{(r)}} \right\}$$

estimated probabilities: $\hat{\tau}^{(r)}_{kl} = \mathbb{P}\left(\widetilde{\mathrm{y}}^{(r)} = l \mid \mathrm{y} = k, \mathbf{x}\right)$

class prior: $\hbar_k$ for class k

Western

Prediction with Noisy Annotations

a new input $\mathbf{x}$ receives labels $\widetilde{\mathrm{y}} = (\widetilde{\mathrm{y}}^{(1)}, \ldots, \widetilde{\mathrm{y}}^{(R)})^{\mathrm{T}}$ from R annotators

$$k^* = \arg\max_{k \in [K]} \mathbb{P}(\mathrm{y} = k | \widetilde{\mathrm{y}}^{(r)}, \mathbf{x})$$

Problem Reformulation

- Step 1:

$$k^* = \arg\max_{k \in [K]} \left\{ \hbar_k \prod_{r=1}^{R} \hat{\tau}^{(r)}_{k, \widetilde{\mathrm{y}}^{(r)}} \right\}$$

estimated probabilities: $\hat{\tau}^{(r)}_{kl} = \mathbb{P}\big(\widetilde{\mathrm{y}}^{(r)} = l \mid \mathrm{y} = k, \mathbf{x}\big)$

class prior: $\hbar_k$ for class k

- Step 2: Convert to that k^* wins every pairwise comparison for

any $k \neq k^*$

Western

Hypothesis Testing:

$$H_k : \mathbb{P}(\widetilde{\mathrm{y}}|\mathrm{y} = k, \mathbf{x}) \qquad \text{versus} \qquad H_{k'} : \mathbb{P}(\widetilde{\mathrm{y}}|\mathrm{y} = k', \mathbf{x})$$

Western

Hypothesis Testing:

$$H_k : \mathbb{P}(\widetilde{\mathrm{y}}|\mathrm{y} = k, \mathbf{x}) \qquad \text{versus} \qquad H_{k'} : \mathbb{P}(\widetilde{\mathrm{y}}|\mathrm{y} = k', \mathbf{x})$$

Pairwise Test for Label Correction

– using the Neyman-Pearson Lemma

$$\frac{\hbar_{i,k} \prod_{r=1}^{R} \prod_{l=1}^{K} \left\{\tau_{i,kl}^{(r)}\right\}^{\mathbf{1}(\widetilde{y}_i^{(r)}=l)}}{\hbar_{i,k'} \prod_{r=1}^{R} \prod_{l=1}^{K} \left\{\tau_{i,k'l}^{(r)}\right\}^{\mathbf{1}(\widetilde{y}_i^{(r)}=l)}} > \Omega$$

Western

Objective

derive bounds on the Bayes risk

Notation

$$\overline{\mathcal{D}} = \{\mathbf{x}_i, \overline{\mathrm{y}}_i\}_{i=1}^{\overline{n}} : \quad \text{instances with estimated labels } \overline{\mathbf{y}} = \{\overline{\mathrm{y}}_i\}_{i=1}^{\overline{n}}$$

$$\mathbf{y} = \{\mathrm{y}_i\}_{i=1}^{\overline{n}} : \quad \text{true labels}$$

$$\hbar(\mathbf{y}) : \quad \text{prior probability of } \mathbf{y}$$

$$\boldsymbol{\tau} \triangleq \{\boldsymbol{\tau}_i^{(r)} : r \in [R]\}_{i=1}^{\overline{n}} : \quad \text{transition matrices}$$

Western

Objective

derive bounds on the Bayes risk

Notation

$$\overline{\mathcal{D}} = \{\mathbf{x}_i, \overline{\mathrm{y}}_i\}_{i=1}^{\overline{n}} : \quad \text{instances with estimated labels } \overline{\mathbf{y}} = \{\overline{\mathrm{y}}_i\}_{i=1}^{\overline{n}}$$

$$\mathbf{y} = \{\mathrm{y}_i\}_{i=1}^{\overline{n}} : \quad \text{true labels}$$

$$\hbar(\mathbf{y}) : \quad \text{prior probability of } \mathbf{y}$$

$$\boldsymbol{\tau} \triangleq \{\boldsymbol{\tau}_i^{(r)} : r \in [R]\}_{i=1}^{\overline{n}} : \quad \text{transition matrices}$$

Loss Function

$$\mathcal{L}(\overline{\mathbf{y}}, \mathbf{y}) = \frac{1}{\overline{n}} \sum_{i=1}^{\overline{n}} \mathbf{1}(\overline{\mathrm{y}}_i \neq \mathrm{y}_i)$$

Bayes Risk

$$\mathfrak{R}_{\text{Bayes}}(\boldsymbol{\hbar}, \mathcal{L}) = \inf_{\overline{\mathbf{y}}} \left[\sum_{\mathbf{y}\in[K]^{\overline{n}}} \hbar(\mathbf{y}) \mathbb{E}\{\mathcal{L}(\overline{\mathbf{y}}, \mathbf{y}) | \mathbf{y}; \boldsymbol{\tau}\}\right]$$

Western

Theorem 2

Under regularity conditions, we have that

$$
\begin{aligned}
&\frac{1}{\overline{n}}\left[1-\frac{\overline{D}_{KL}(\boldsymbol{\hbar},\boldsymbol{\tau})+\frac{1}{\overline{n}}\log(2-\prod_{i=1}^{\overline{n}}\max_{k\in[K]}\hbar_{i,k})}{\left\{\sum_{i=1}^{\overline{n}}\log(\max_{k\in[K]}\hbar_{i,k})\right\}/\overline{n}}\right] \\
\leq\ & \mathfrak{R}_{\text{Bayes}}(\boldsymbol{\hbar},\mathcal{L}) \\
\leq\ & \frac{K-1}{\overline{n}}\sum_{i=1}^{\overline{n}}\sum_{k=1}^{K}\hbar_{i,k}\exp\left\{-RI_{\Omega}^{(k)}(\boldsymbol{\hbar}_i,\boldsymbol{\tau}_i)\right\}
\end{aligned}
$$

Theorem 2

Under regularity conditions, we have that

$$
\begin{aligned}
&\frac{1}{\bar{n}}\left[1-\frac{\overline{D}_{KL}(\boldsymbol{\hbar},\boldsymbol{\tau})+\frac{1}{\bar{n}}\log(2-\prod_{i=1}^{\bar{n}}\max_{k\in[K]}\hbar_{i,k})}{\left\{\sum_{i=1}^{\bar{n}}\log(\max_{k\in[K]}\hbar_{i,k})\right\}/\bar{n}}\right]\\
\leq\;& \mathfrak{R}_{\text{Bayes}}(\boldsymbol{\hbar},\mathcal{L})\\
\leq\;& \frac{K-1}{\bar{n}}\sum_{i=1}^{\bar{n}}\sum_{k=1}^{K}\hbar_{i,k}\exp\left\{-RI_{\Omega}^{(k)}(\boldsymbol{\hbar}_i,\boldsymbol{\tau}_i)\right\}
\end{aligned}
$$

Remarks

- Theorem 2 establishes bounds on the Bayes risk for arbitrary priors.
- The involvement of $I_{\Omega}^{(k)}(\boldsymbol{\hbar}_i,\boldsymbol{\tau}_i)$ reflects how the identified upper bound of the Bayes risk may be influenced by the prior $\boldsymbol{\hbar}_i$ and the ability of annotators to distinguish between labels.

Western

Numerical Studies

Western

Image Datasets

- two datasets with human annotations:
  - CIFAR-10N: three independent human annotated noisy labels
  - LabelMe: 10,000 training images, 500 validation images, and 1,188 test images

Image Datasets

- two datasets with human annotations:
  - CIFAR-10N: three independent human annotated noisy labels
  - LabelMe: 10,000 training images, 500 validation images, and 1,188 test images

- three datasets with synthetic annotations:
  - MNIST: 10 classes of 28×28 grayscale images
  - CIFAR10: 10 classes of $32 \times 32 \times 3$ color images
  - CIFAR100: 100 fine-grained classes of $32 \times 32 \times 3$ images

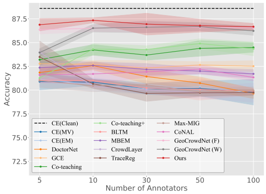
  - consider three groups of annotators with varying expertise
  - an average labeling error rate of about 20%, 35% and 50%

Competing Methods

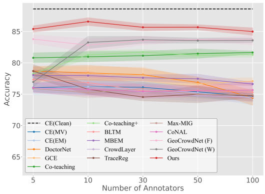
(1). CE (Clean): train the network with the standard cross entopy loss on the clean datasets
(2). CE (MV): train the network using the labels from majority voting
(3). CE (EM): Dawid and Skene (1979)
(4). DoctorNet: Guan, Gulshan, Dai, and (Hinton 2018)
(5). GCE: Zhang and Sabuncu (2018)
(6). Co-teaching: Han, Yao, Yu, Niu, Xu, Hu, Tsang, and Sugiyama (2018)
(7). Co-teaching+: Yu, Han, Yao, Niu, Tsang, and Sugiyama (2019)
(8). BLTM: Yang, Yang, Han, Liu, Xu, Niu, and Liu (2022)
(9). MBEM: Khetan, Lipton, and Anandkumar (2017)
(10). CrowdLayer: Rodrigues and Pereira (2018)
(11). TraceReg: Tanno, Saeedi, Sankaranarayanan, Alexander, and Silberman (2019)
(12). Max-MIG: Cao, Xu, Kong, and Wang (2019)
(13). CoNAL: Chu, Ma, and Wang (2021)
(14). GeoCrowdNet(F): Ibrahim, Nguyen, and Fu (2023)
(15). GeoCrowdNet (W): Ibrahim, Nguyen, and Fu (2023)

Western

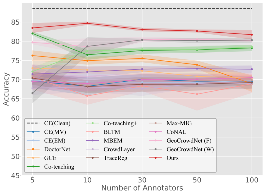
Empirical Results on CIFAR10



(a) IDN-LOW

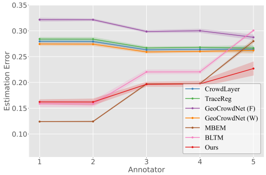


(b) IDN-MID

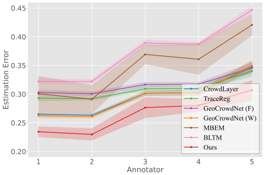


(c) IDN-HIGH

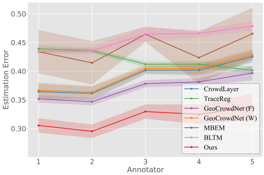
Average Accuracy



(d) IDN-LOW



(e) IDN-MID



(f) IDN-HIGH

Average Estimation Error for Transition Matrices

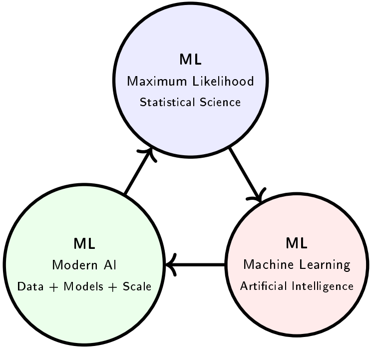
With varying number of annotators, the proposed method

- achieves the highest average test accuracy
- leads to smallest estimation error in most of the cases, especially when the noise rate is high.

Western

Part 4: From ML to ML

Part 4: From ML to ML



Statistics provides principles for ML; ML generates new statistical challenges

From ML to ML:

Examples of “Invisible Victory” of Statistical Thinking

Brief Timeline of Machine Learning

1950s–1960s
Turing
Perceptron

→

1980s
Decision Trees
Backpropagation

→

1990s
CNNs
SVMs

→

2000s
Ensemble Methods
Sparse/Graphical Methods

→

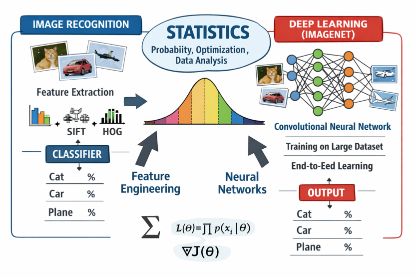
2010s–present
Deep Learning
Transformers
Generative AI

Success Stories: Statistics in Deep Learning

Success Stories: Statistics in Deep Learning

Image Recognition (AlexNet)

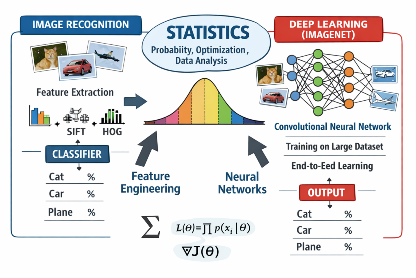
- Large-scale supervised learning (ImageNet) + CNNs



Success Stories: Statistics in Deep Learning

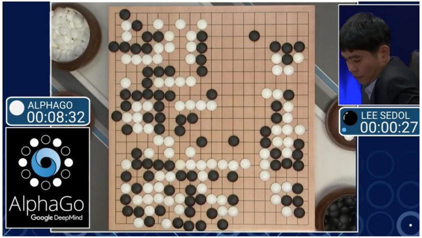
Image Recognition (AlexNet)

- Large-scale supervised learning (ImageNet) + CNNs



Computer Go (AlphaGo)

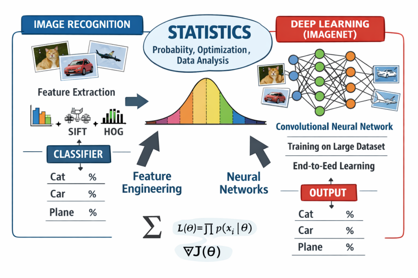
- Deep policy/value networks + Monte Carlo Tree Search



Success Stories: Statistics in Deep Learning

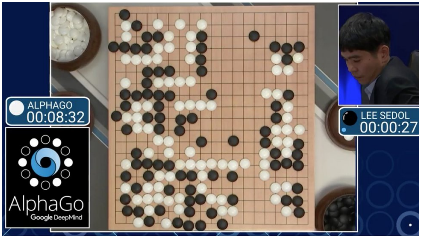
Image Recognition (AlexNet)

- Large-scale supervised learning (ImageNet) + CNNs



Computer Go (AlphaGo)

- Deep policy/value networks + Monte Carlo Tree Search



Where Statistics Is Critical

- Probabilistic modeling
- Loss & likelihood
- Monte Carlo estimation
- Stochastic approximation

Western

Success Story: Statistics in NLP and Language Models

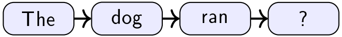
NLP $\longrightarrow$ Statistical Problems:

modeling the joint distribution of word sequences via conditional probabilities

Success Story: Statistics in NLP and Language Models

NLP → Statistical Problems:

modeling the joint distribution of word sequences via conditional probabilities

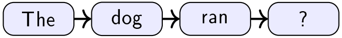


predict the next word

Success Story: Statistics in NLP and Language Models

NLP $\longrightarrow$ Statistical Problems:

modeling the joint distribution of word sequences via conditional probabilities



predict the next word

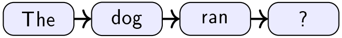
- **Statistical Target:**

$$P(\text{next word} \mid \text{context})$$

Success Story: Statistics in NLP and Language Models

NLP $\longrightarrow$ Statistical Problems:

modeling the joint distribution of word sequences via conditional probabilities



predict the next word

- **Statistical Target:**

$$P(\text{next word} \mid \text{context})$$

- **Estimation:**
  learn this conditional distribution from data

- **Inference:**
  choose likely words, sentences, or translations

Raw Text Is Not Directly Usable

`“The dog ran away”`

Raw Text Is Not Directly Usable

“The dog ran away”

Tokenization

[The, dog, ran, away]

Tokens = words / subwords / characters

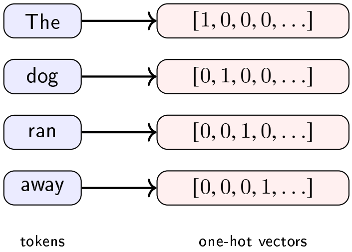
Raw Text Is Not Directly Usable

“The dog ran away”

Tokenization

[The, dog, ran, away]

Tokens = words / subwords / characters



Tokens $\Longrightarrow$ discrete random variables

Raw Text Is Not Directly Usable

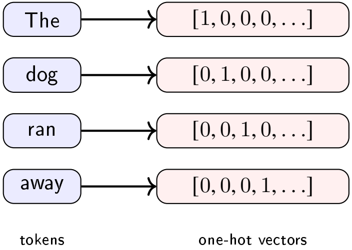
“The dog ran away”

Tokenization

[The, dog, ran, away]

Tokens = words / subwords / characters

- Modern ML/NLP starts with
  text $\longrightarrow$ tokens $\longrightarrow$ vectors $\longrightarrow$ statistics



Tokens $\Longrightarrow$ discrete random variables

- Numerical representation enables statistical modeling.

Western

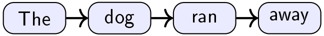
n-gram Models: Model Natural Language as Sequences

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$$

Statistics in NLP: From Words to Probabilities

n-gram Models: Model Natural Language as Sequences

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$$



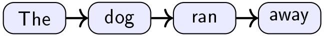
next-word prediction

- MLE: estimate word probabilities
- Smoothing: avoid zero probabilities
- Embeddings: reduce sparsity

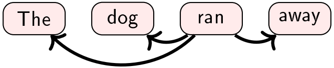
Statistics in NLP: From Words to Probabilities

n-gram Models: Model Natural Language as Sequences

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$$



next-word prediction



transformer attention

- MLE: estimate word probabilities
- Smoothing: avoid zero probabilities
- Embeddings: reduce sparsity
- Attention Weights: data-dependent probabilities
- Softmax: probability normalization
- Training: likelihood/cross-entropy optimization

Transformers improve learning by attending to relevant information across words and contexts.

From ML to ML:

What Machine Learning Brings to Statistics?

Perspective 1: Reliability and Risk – An Analogy with Drug Development and LLMs

Western

Perspective 1: Reliability and Risk – An Analogy with Drug Development and LLMs

Managing Risk and Side Effects in Drug Development

- Drugs undergo rigorous testing before deployment.
- Risks and side effects are explicitly documented.
- Continuously monitor after deployment.
- Severe issues may lead to recall.

Perspective 1: Reliability and Risk – An Analogy with Drug Development and LLMs

Managing Risk and Side Effects in Drug Development

- Drugs undergo rigorous testing before deployment.
- Risks and side effects are explicitly documented.
- Continuously monitor after deployment.
- Severe issues may lead to recall.

LLMs Also Have "Side Effects"

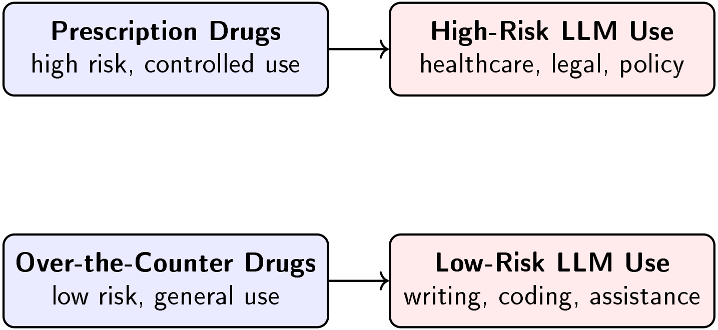
- hallucination/fabricated content: incorrect or nonexistent facts
- data quality dependence: noise/misinformation/outdated knowledge
- bias and fairness issues: social/demographic/cultural bias
- safety and misuse risks: misleading/harmful/manipulative content
- strong at correlation but weak at causal reasoning
- sensitive to prompts
- lack of reliability/calibration: no intrinsic notion of "I don't know"

Not All LLM Usage Should Be Treated Equally

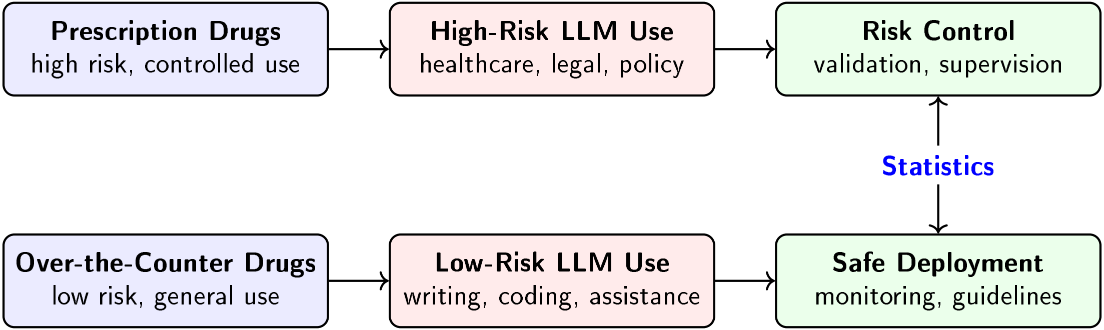
Prescription Drugs
high risk, controlled use

Over-the-Counter Drugs
low risk, general use

Not All LLM Usage Should Be Treated Equally



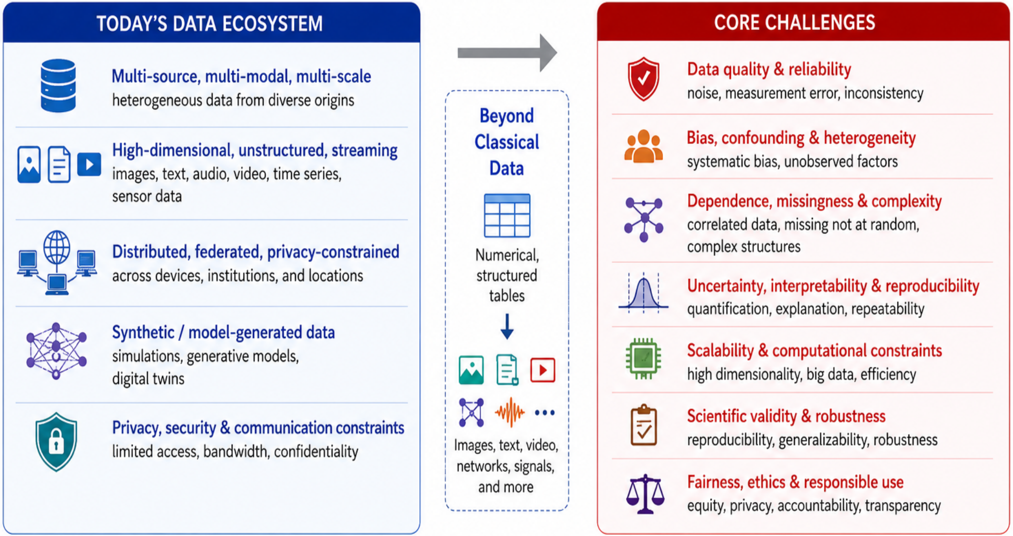
Not All LLM Usage Should Be Treated Equally



Statistical Education and Research Enables Safe and Responsible Use of AI

- quantify uncertainty and risk
- provide transparent "usage warnings"

Perspective 2: The Changing Nature of Data



Western

Beyond Classical Datasets, Data Today Are Richer in Nature

Beyond Classical Datasets, Data Today Are Richer in Nature

Classical Data

- numerical features
- categorical variables
- structured tables
- low/moderate–dimensional
- collected by design/sampling

→

Modern Data

- images, text, video, audio
- networks, signals
- unstructured / multi-modal
- high-dimensional / large–scale
- distributed, federated, & privacy–constrained
- increasingly synthetic & model–generated

Beyond Classical Datasets, Data Today Are Richer in Nature

Classical Data

- numerical features
- categorical variables
- structured tables
- low/moderate–dimensional
- collected by design/sampling

→

Modern Data

- images, text, video, audio
- networks, signals
- unstructured / multi-modal
- high-dimensional / large–scale
- distributed, federated, & privacy–constrained
- increasingly synthetic & model–generated

Remarks

- more data ≠ better inference
- What we can learn depends on how we represent data.
- Data are no longer static objects: – dynamic, streaming, and evolving

Western

Beyond Classical Datasets, Data Today Are Richer in Nature

Classical Data

- numerical features
- categorical variables
- structured tables
- low/moderate–dimensional
- collected by design/sampling

→

Modern Data

- images, text, video, audio
- networks, signals
- unstructured / multi-modal
- high-dimensional / large–scale
- distributed, federated, & privacy–constrained
- increasingly synthetic & model–generated

Remarks

- more data ≠ better inference
- What we can learn depends on how we represent data.
- Data are no longer static objects: – dynamic, streaming, and evolving

Statistical Implications

- quality of data: collection, preprocess, representation
- rethink modeling assumptions
- integrate statistics + ML
- unify theory, methods, scalable computation

Western

Perspective 3: The Changing Nature of Modeling

Western

Perspective 3: The Changing Nature of Modeling

- Classical models give first-order approximations.

Western

Perspective 3: The Changing Nature of Modeling

- Classical models give first-order approximations.
- To perform statistical analysis of modern data, it is necessary to explicitly model **data quality** to provide higher-order approximations:

  second-order, **third-order,** $\cdots$

Western

Perspective 3: The Changing Nature of Modeling

- Classical models give first-order approximations.
- To perform statistical analysis of modern data, it is necessary to explicitly model **data quality** to provide higher-order approximations:

  second-order, **third-order,** $\cdots$

  to accommodate variability/uncertainty arising from

  - data collection process

    e.g., measurement error, missingness, selection bias/fairness, privacy/ethical concerns, etc.

  - data preprocessing and labeling procedures
  - data representation and feature extraction

Western

Perspective 3: The Changing Nature of Modeling

- Classical models give first-order approximations.
- To perform statistical analysis of modern data, it is necessary to explicitly model **data quality** to provide higher-order approximations:

  second-order, **third-order,** $\cdots$

  to accommodate variability/uncertainty arising from

  – data collection process

  e.g., measurement error, missingness, selection bias/fairness, privacy/ethical concerns, etc.

  – data preprocessing and labeling procedures

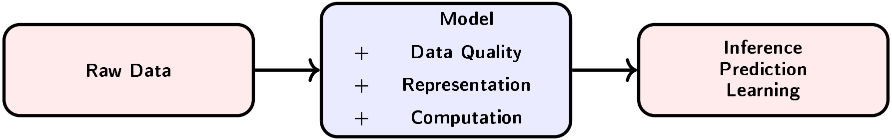
  – data representation and feature extraction

Today, Both Fields Increasingly Share Tools and Goals

- Machine Learning: growing focus on interpretable machine learning.
- Statistical Science: addressing data complexity and imperfections with stronger computational emphasis.

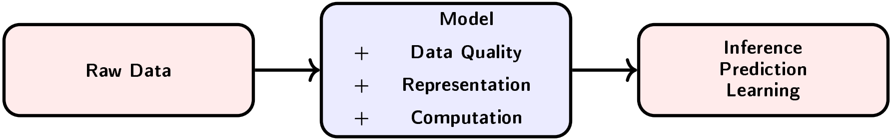
Western

TAKE–HOME MESSAGES



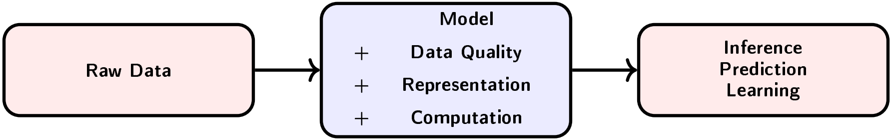
TAKE–HOME MESSAGES

- If the 20th century was about models, the 21st may shift toward – data quality, representation, and modeling, powered by machine learning.



TAKE–HOME MESSAGES

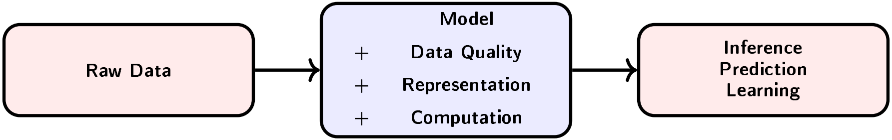
- If the 20th century was about models, the 21st may shift toward
  – data quality, representation, and modeling, powered by machine learning.



- **"All models are wrong, but some are useful." — George Box**

TAKE–HOME MESSAGES

- If the 20th century was about models, the 21st may shift toward
  – data quality, representation, and modeling, powered by machine learning.

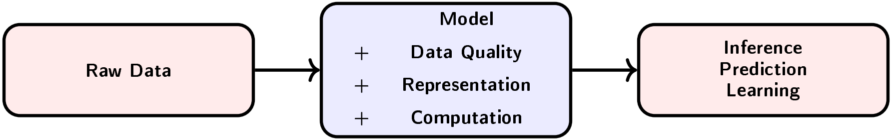


- **"All models are wrong, but some are useful." — George Box**

  "All data are imperfect, but some are useful

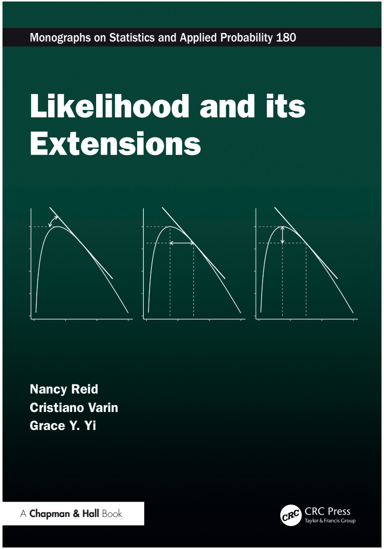
TAKE–HOME MESSAGES

- If the 20th century was about models, the 21st may shift toward
  – data quality, representation, and modeling, powered by machine learning.



- **"All models are wrong, but some are useful." — George Box**

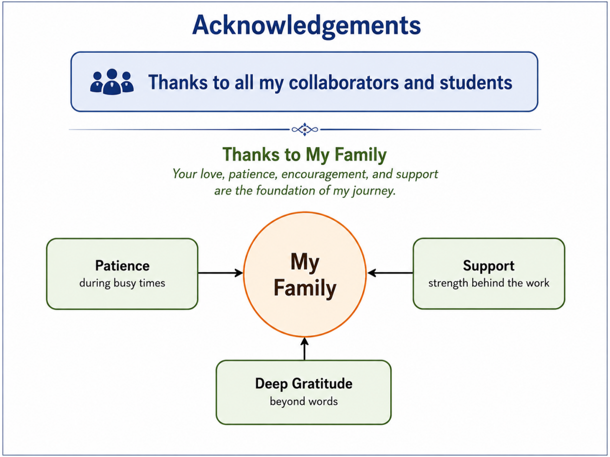
  "All data are imperfect, but some are useful – if properly represented."
  — Grace Y. Yi


Monographs on Statistics and Applied Probability 180
Likelihood and its Extensions
Nancy Reid
Cristiano Varin
Grace Y. Yi
A Chapman & Hall Book
CRC Press
Taylor & Francis Group

Western

Western

Western


Acknowledgements
Thanks to all my collaborators and students
Thanks to My Family
Your love, patience, encouragement, and support
are the foundation of my journey.
Patience
during busy times
My
Family
Support
strength behind the work
Deep Gratitude
beyond words

Acknowledgements

Thanks to all my collaborators and students

Thanks to My Family

Your love, patience, encouragement, and support are the foundation of my journey.

My Family

Patience during busy times

Support strength behind the work

Deep Gratitude beyond words

CANADA RESEARCH CHAIRS
CHAIRES DE RECHERCHE DU CANADA

NSERC
CRSNG

Western