TRADE-OFFS AMONG AI TECHNIQUES

Christian Kaestner

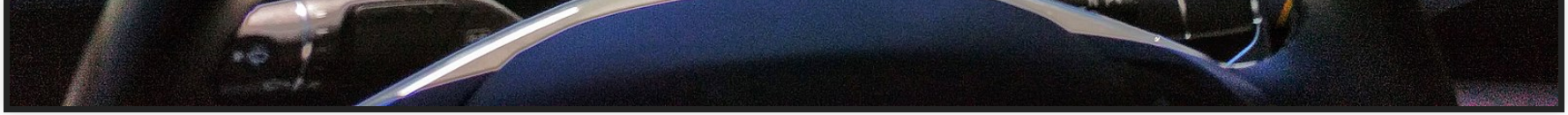
With slides adopted from Eunsuk Kang

Required reading: Vogelsang, Andreas, and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists." In Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2019.

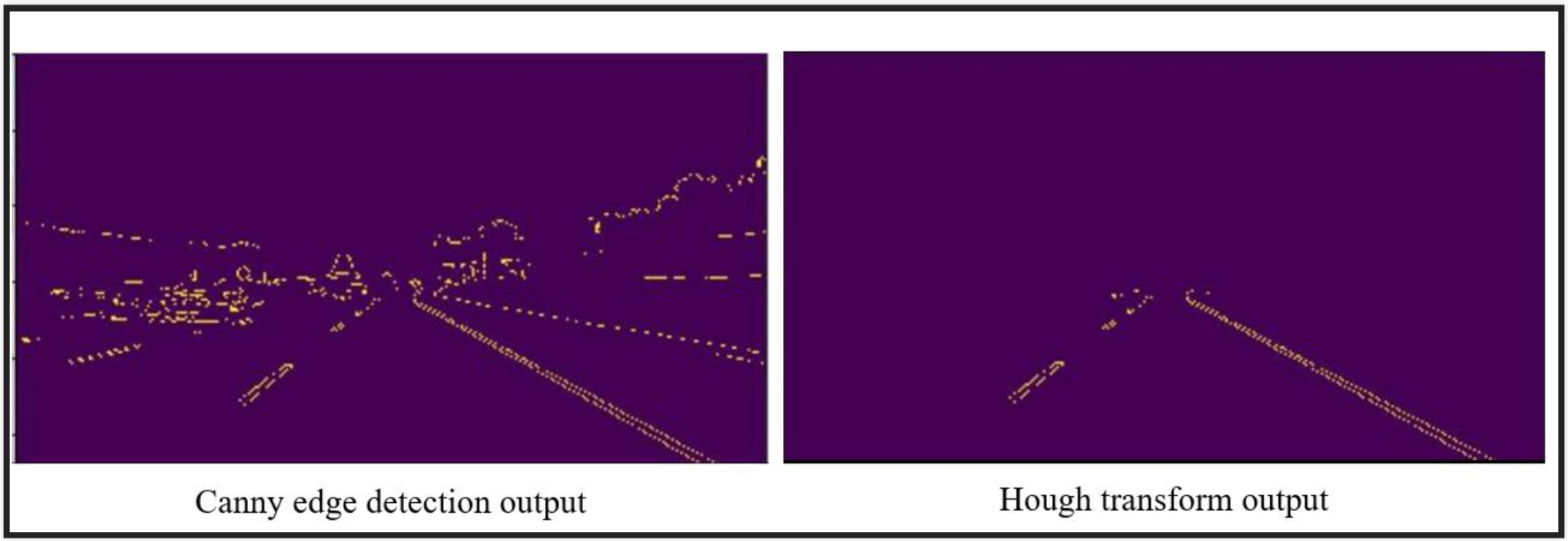
LEARNING GOALS

- Describe the most common models and learning strategies used for AI components and summarize how they work
- Organize and prioritize the relevant qualities of concern for a given project
- Plan and execute an evaluation of the qualities of alternative AI components for a given purpose

TODAY'S CASE STUDY: LANE ASSIST



TODAY'S CASE STUDY: LANE ASSIST



BACKGROUND: LANE ASSIST

From audio, haptic, and visual signal ("lane departure warning") to automated steering ("lane keeping"); often combined with adaptive cruise control

Safety or comfort feature

Multiple inputs: camera, indicators, speed, possibly radar, hands on steering wheel sensor

Multiple AI components: Lane recognition, automated steering, automated breaking

Integrated into larger systems with user interface, sensors, actuators, and other AI and non-AI components, working together with humans

Classic systems based on old line detection techniques in images (no deep learning)

See https://en.wikipedia.org/wiki/Lane_departure_warning_system

QUALITY

VIEWS OF QUALITY

- **Transcendent** – Experiential. Quality can be recognized but not defined or measured
- **Product-based** – Level of attributes (More of this, less of that)
- **User-based** – Fitness for purpose, quality in use
- **Value-based** – Level of attributes/fitness for purpose at given cost
- **Manufacturing** – Conformance to specification, process excellence

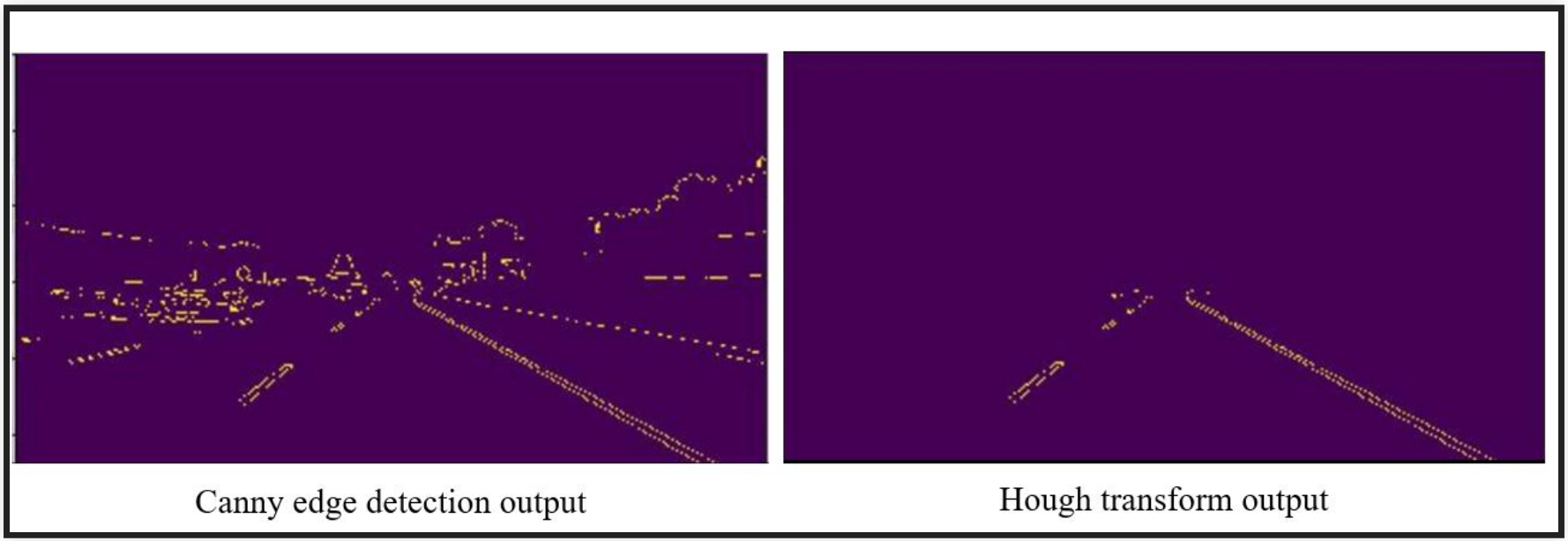
Reference: Garvin, David A., What Does Product Quality Really Mean. Sloan management review 25 (1984).

GARVIN'S EIGHT CATEGORIES OF PRODUCT QUALITY

- Performance
- Features
- Reliability
- Conformance
- Durability
- Serviceability
- Aesthetics
- Perceived Quality

Reference: Garvin, David A., What Does Product Quality Really Mean. Sloan management review 25 (1984).

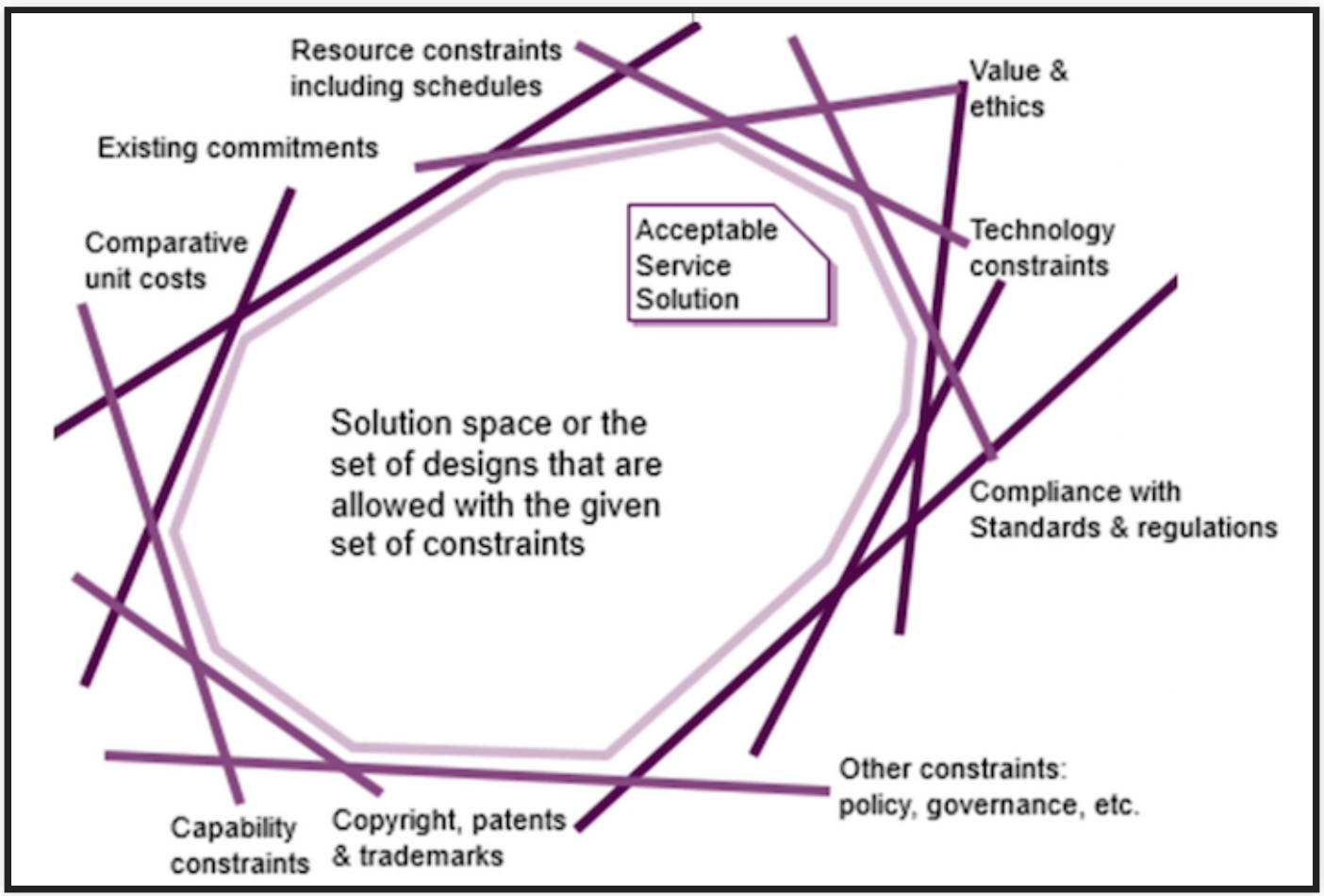
ATTRIBUTES



- **Quality attributes:** How well the product (system) delivers its functionality (usability, reliability, availability, security...)
- **Project attributes:** Time-to-market, development & HR cost...
- **Design attributes:** Type of AI method used, accuracy, training time, inference time, memory usage...

CONSTRAINTS

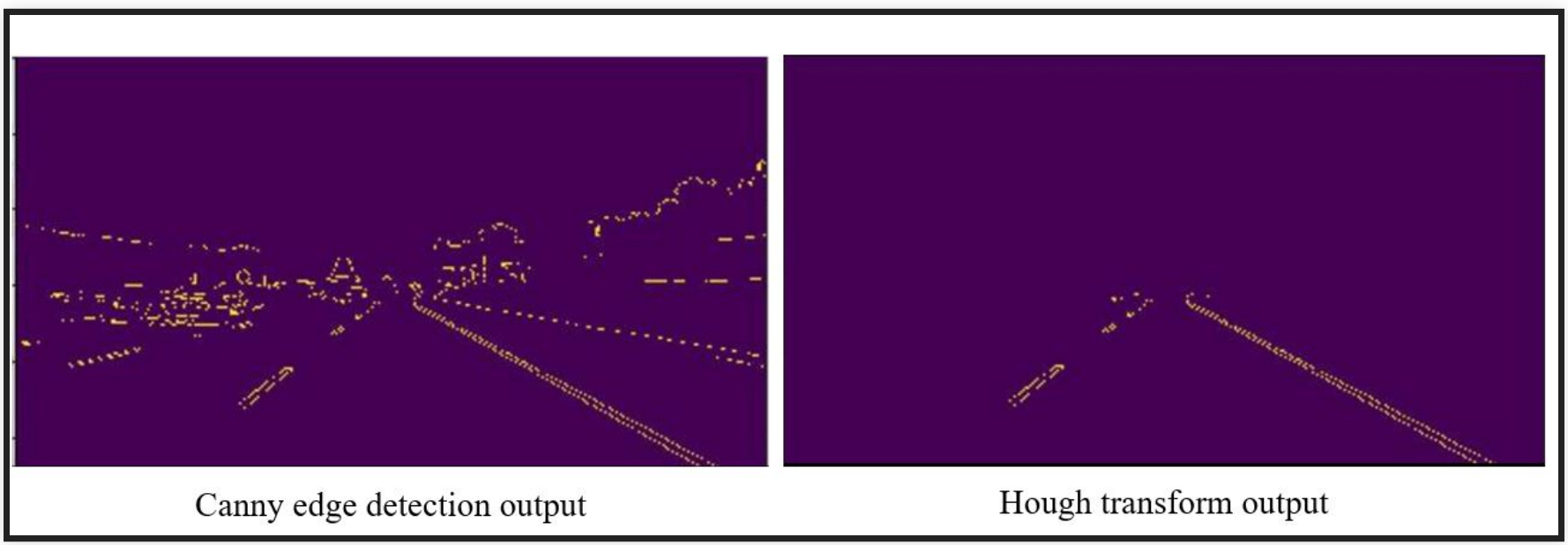
Constraints define the space of attributes for valid design solutions



TYPES OF CONSTRAINTS

- Problem constraints: Minimum required QAs for an acceptable product
- Project constraints: Deadline, project budget, available skills
- Design constraints: Type of ML task required (regression/classification), kind of available data, limits on computing resources, max. inference cost

Plausible constraints for Lane Assist?



AI SELECTION PROBLEM

- How to decide which AI method to use in project?
- Find method that:
    1. satisfies the given constraints and
    2. is optimal with respect to the set of relevant attributes

REQUIREMENTS ENGINEERING:

IDENTIFY RELEVANT QUALITIES OF AI COMPONENTS IN AI-ENABLED SYSTEMS

ACCURACY IS NOT EVERYTHING

Beyond prediction accuracy, what qualities may be relevant for an AI component?

Speaker notes

Collect qualities on whiteboard

QUALITIES OF INTEREST?

Scenario: Component detecting line markings in camera picture

Speaker notes

Which of the previously discussed qualities are relevant? Which additional qualities may be relevant here?

QUALITIES OF INTEREST?

Scenario: Component predicting defaulting on loan (credit rating)

MEASURING QUALITIES

- Define a metric -- define units of interest
    - e.g., requests per second, max memory per inference, average training time in seconds for 1 million datasets
- Operationalize metric -- define measurement protocol
    - e.g., conduct experiment: train model with fixed dataset, report median training time across 5 runs, file size, average accuracy with leave-one-out crossvalidation after hyperparameter tuning
    - e.g., ask 10 humans to independently label evaluation data, report reduction in error from machine-learned model over human predictions
    - describe all relevant factors: inputs/experimental units used, configuration decisions and tuning, hardware used, protocol for manual steps

On terminology: *metric/measure* refer a method or standard format for measuring something; *operationalization* is identifying and implementing a method to measure some factor

EXAMPLES OF QUALITIES TO CONSIDER

- Accuracy
- Correctness guarantees? Probabilistic guarantees (--> symbolic AI)
- How many features? Interactions among features?
- How much data needed? Data quality important?
- Incremental training possible?
- Training time, memory need, model size -- depending on training data volume and feature size
- Inference time, energy efficiency, resources needed, scalability
- Interpretability/explainability
- Robustness, reproducibility, stability
- Security, privacy
- Fairness

ON TERMINOLOGY

- Data scientists seem to speak of *model properties* when referring to accuracy, inference time, fairness, etc
    - ... but they also use this term for whether a *learning technique* can learn non-linear relationships or whether the learning algorithm is monotonic
- Software engineering wording would usually be *quality attributes*, *non-functional requirements*, ...

INTERPRETABILITY/EXPLAINABILITY

"Why did the model predict X?"

Explaining predictions + Validating Models + Debugging

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict ar
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Some models inherently simpler to understand

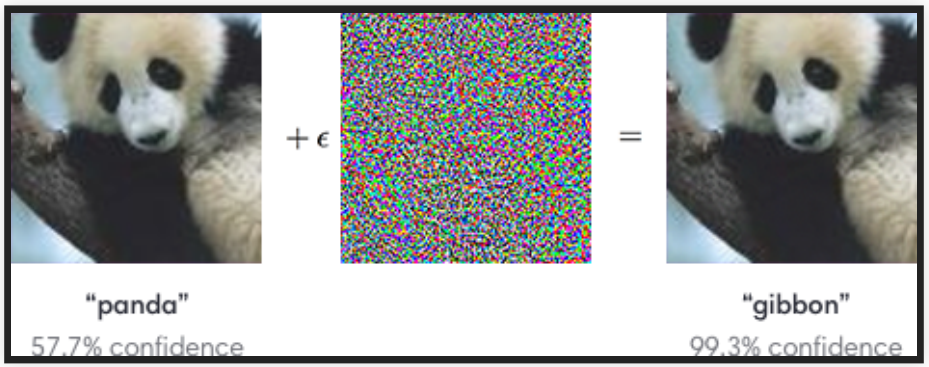
Some tools may provide post-hoc explanations

Explanations may be more or less truthful

How to measure interpretability?

more in a later lecture

ROBUSTNESS



Small input modifications may change output

Small training data modifications may change predictions

How to measure robustness?

more in a later lecture

Image source: OpenAI blog

FAIRNESS

Does the model perform differently for different populations?

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict ar
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Many different notions of fairness

Often caused by bias in training data

Enforce invariants in model or apply corrections outside model

Important consideration during requirements solicitation!

more in a later lecture

REQUIREMENTS ENGINEERING FOR AI-ENABLED SYSTEMS

- Set minimum accuracy expectations ("functional requirement")
- Identify explainability needs
- Identify protected characteristics and possible fairness concerns
- Identify security and privacy requirements (ethical and legal), e.g., possible use of data
- Understand data availability and need (quality, quantity, diversity, formats, provenance)

- Involve data scientists and legal experts
- Map system goals to AI components
- Establish constraints, set goals

Further reading: Vogelsang, Andreas, and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists." In Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2019.

SOME TRADEOFFS OF COMMON ML TECHNIQUES

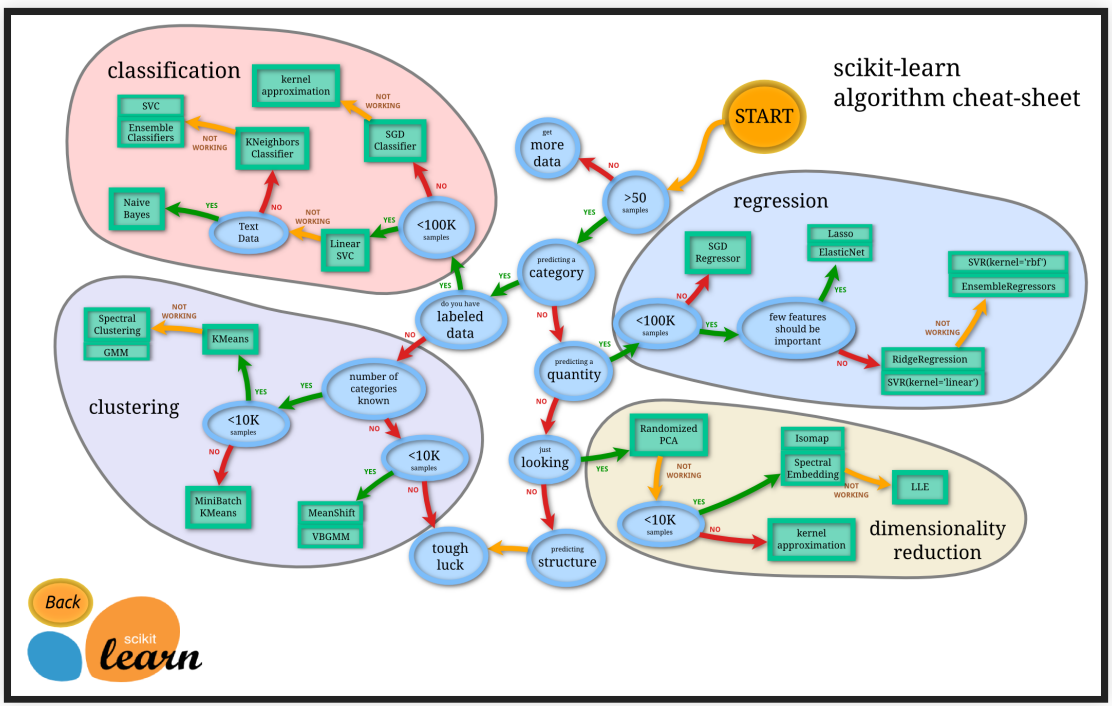
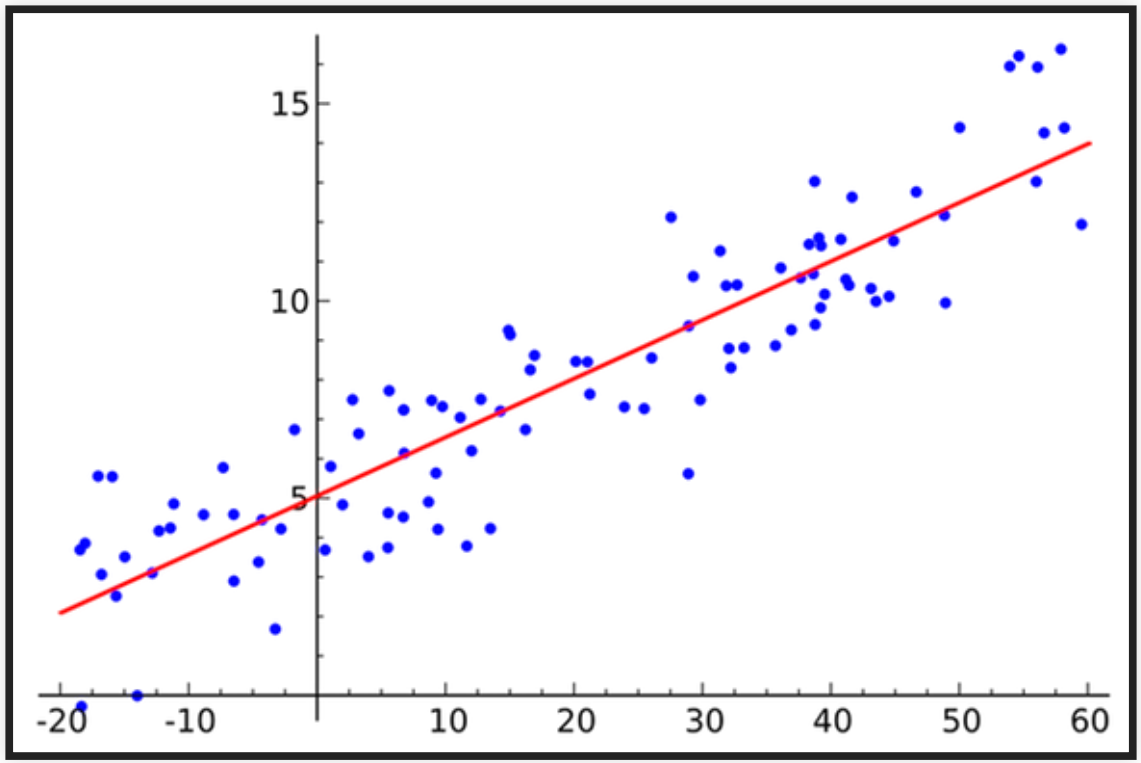


Image: Scikit Learn Tutorial

LINEAR REGRESSION



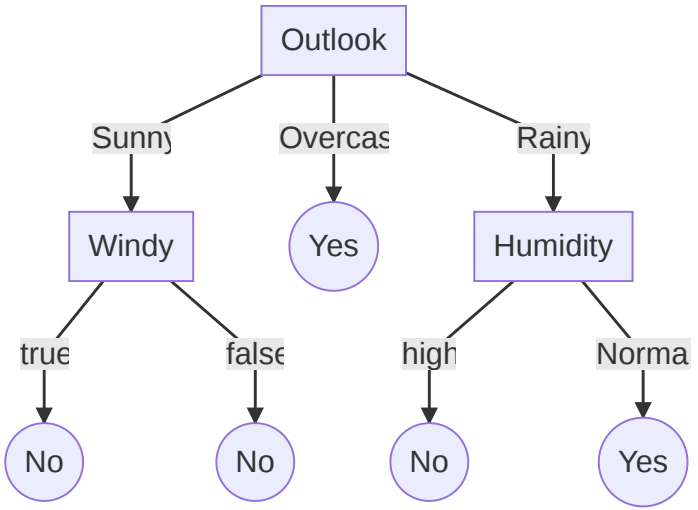
- Tasks: Regression, labeled data
- Linear relationship between input & output variables
- Advantages: ??
- Disadvantages: ??

Speaker notes

- Easy to interpret, low training cost, small model size
- Can't capture non-linear relationships well

DECISION TREE LEARNING

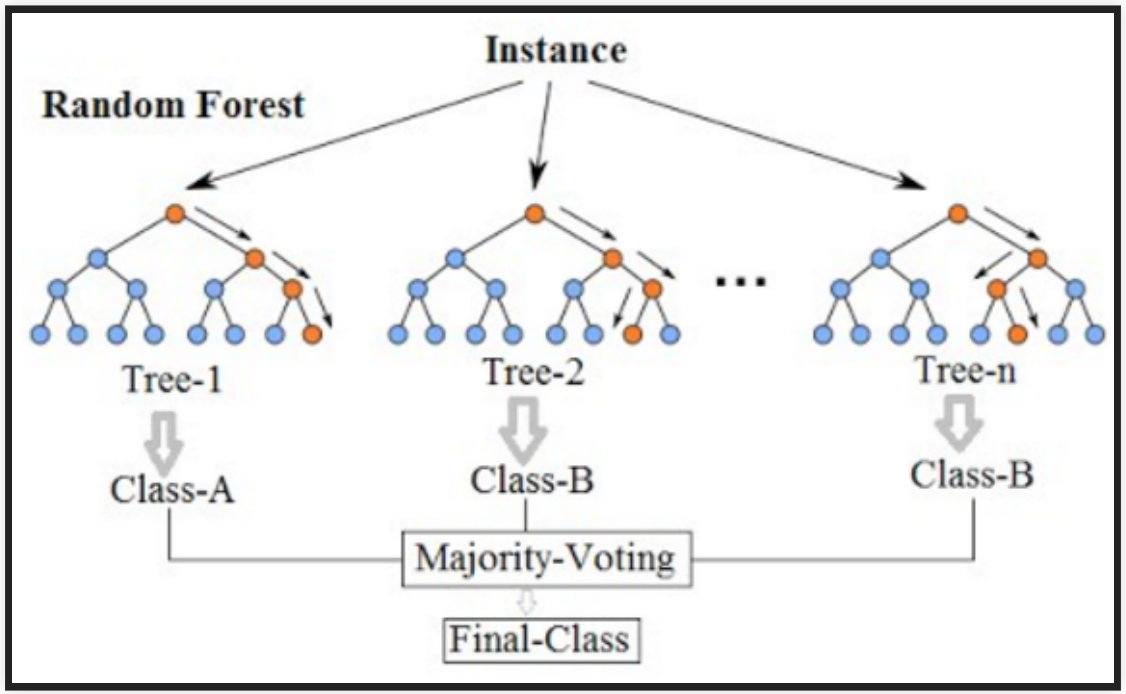
- Tasks: Classification & regression, labeled data
- Advantages: ??
- Disadvantages: ??



Speaker notes

- Easy to interpret (up to a size); can capture non-linearity; can do well with little data
- High risk of overfitting; possibly very large tree size

RANDOM FORESTS



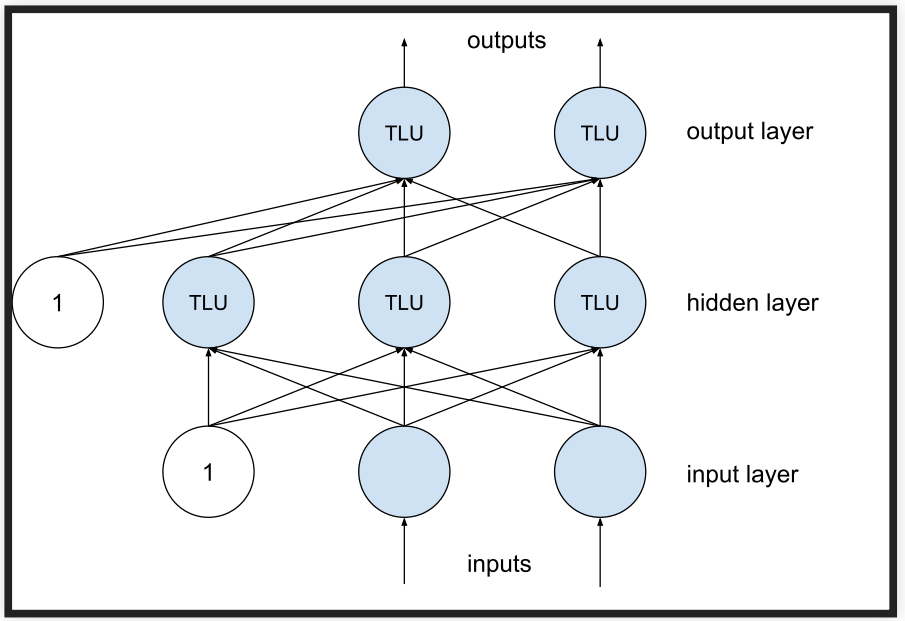
- Construct lots of decision trees with some randomness (e.g., on subsets of data or subsets of features)
- Advantages: ??
- Disadvantages: ??

Speaker notes

- High accuracy & reduced overfitting; incremental (can add new trees)
- Reduced interpretability; large number of trees can take up space

NEURAL NETWORK

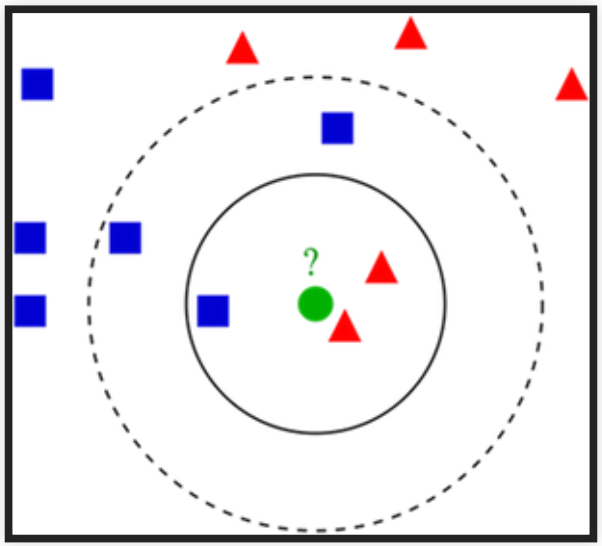
- Tasks: Classification & regression, labeled data
- Advantages: ??
- Disadvantages: ??



Speaker notes

- High accuracy; can capture a wide range of problems (linear & non-linear)
- Difficult to interpret; high training costs (time & amount of data required, hyperparameter tuning)

K-NEAREST NEIGHBORS (K-NN)

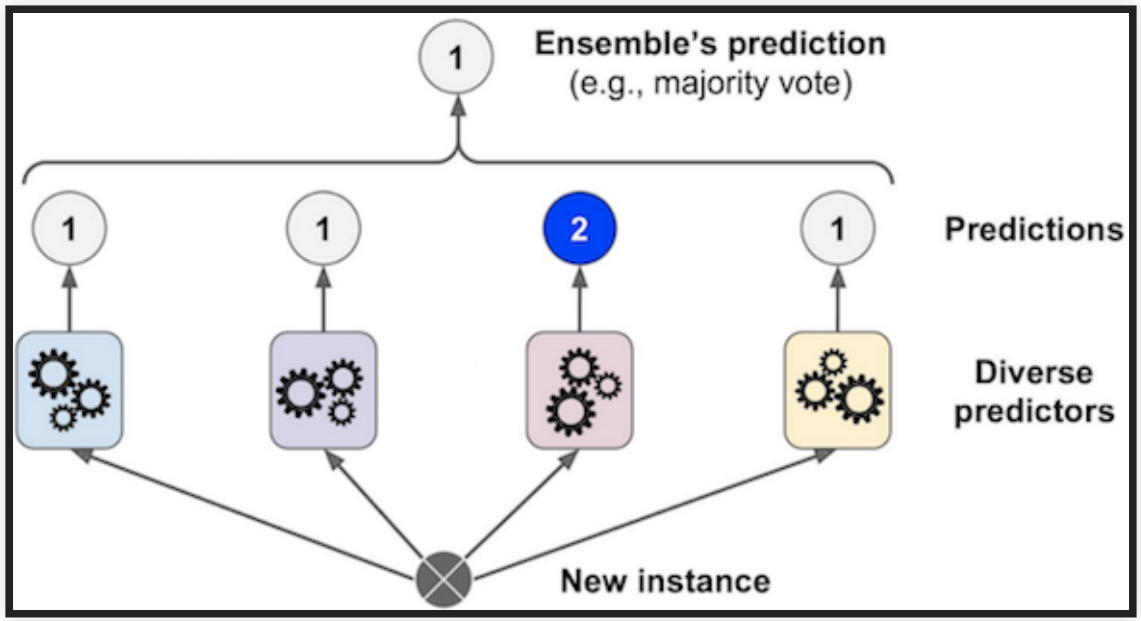


- Tasks: Classification & regression, unsupervised
- Infer the class/property of an object based on that of *k* nearest neighbors
- **Lazy learning**: Generalization is delayed until the inference takes place
- Advantages: ??
- Disadvantages: ??

Speaker notes

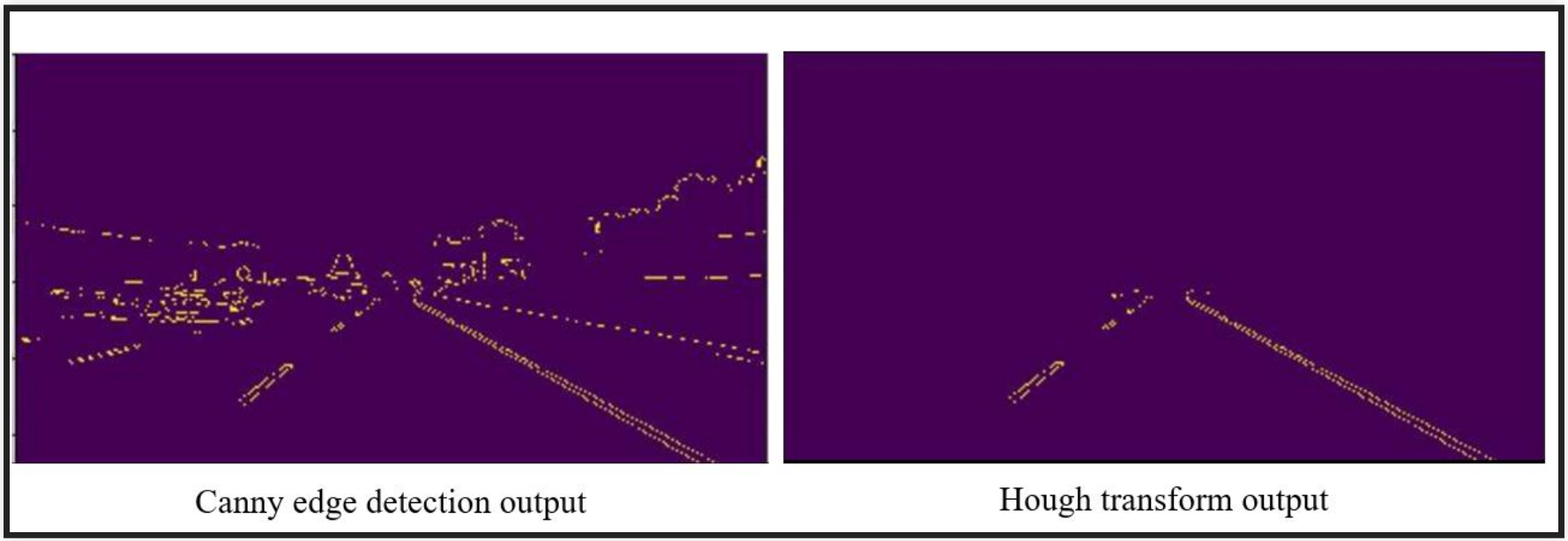
- Easy to interpret; no training required (due to lazy learning); incremental (can continuously add new data)
- Potentially slow inference (again, due to lazy learning); high data storage requirement (must store training instances)

ENSEMBLE LEARNING



- Combine a set of low-accuracy (but cheaper to learn) models to provide high-accuracy predictions

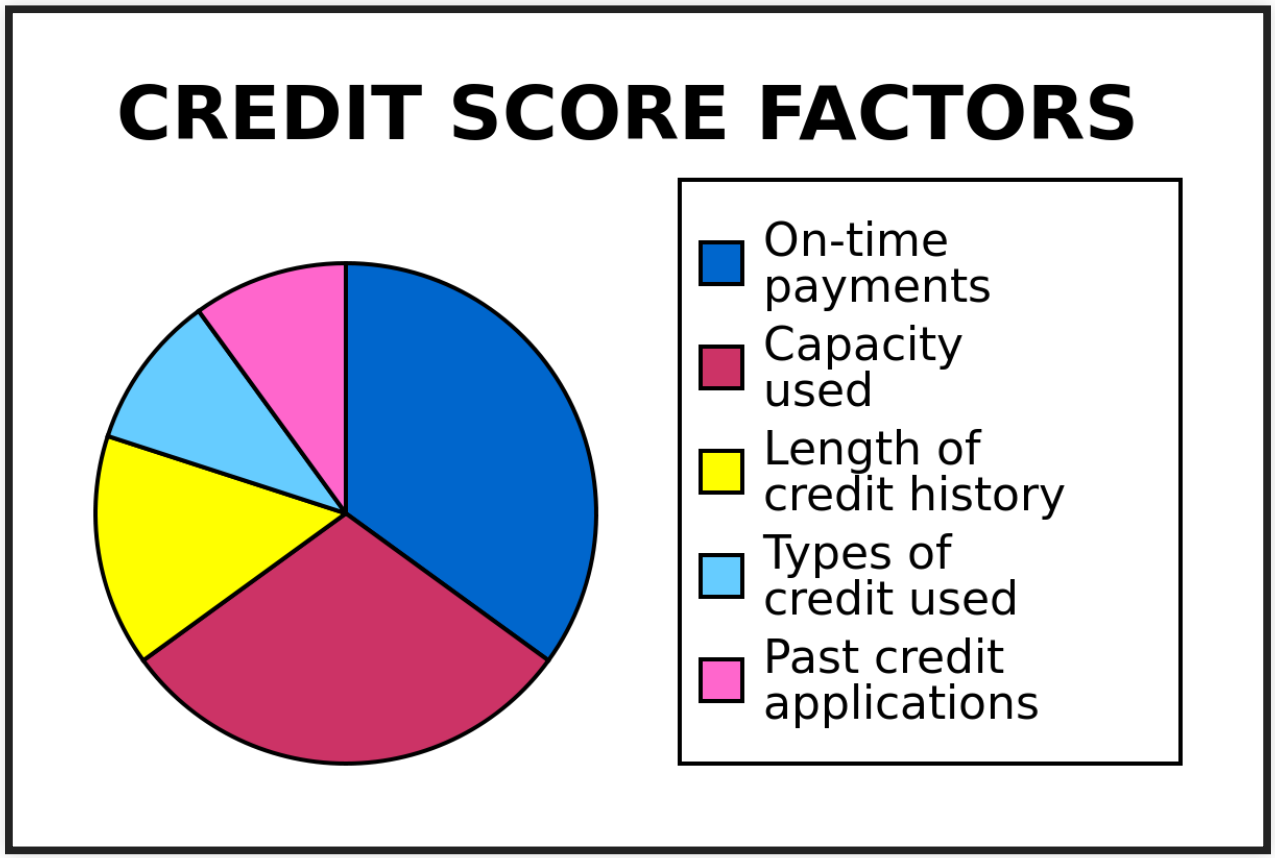
WHICH METHOD FOR LANE DETECTION?



Canny edge detection output

Hough transform output

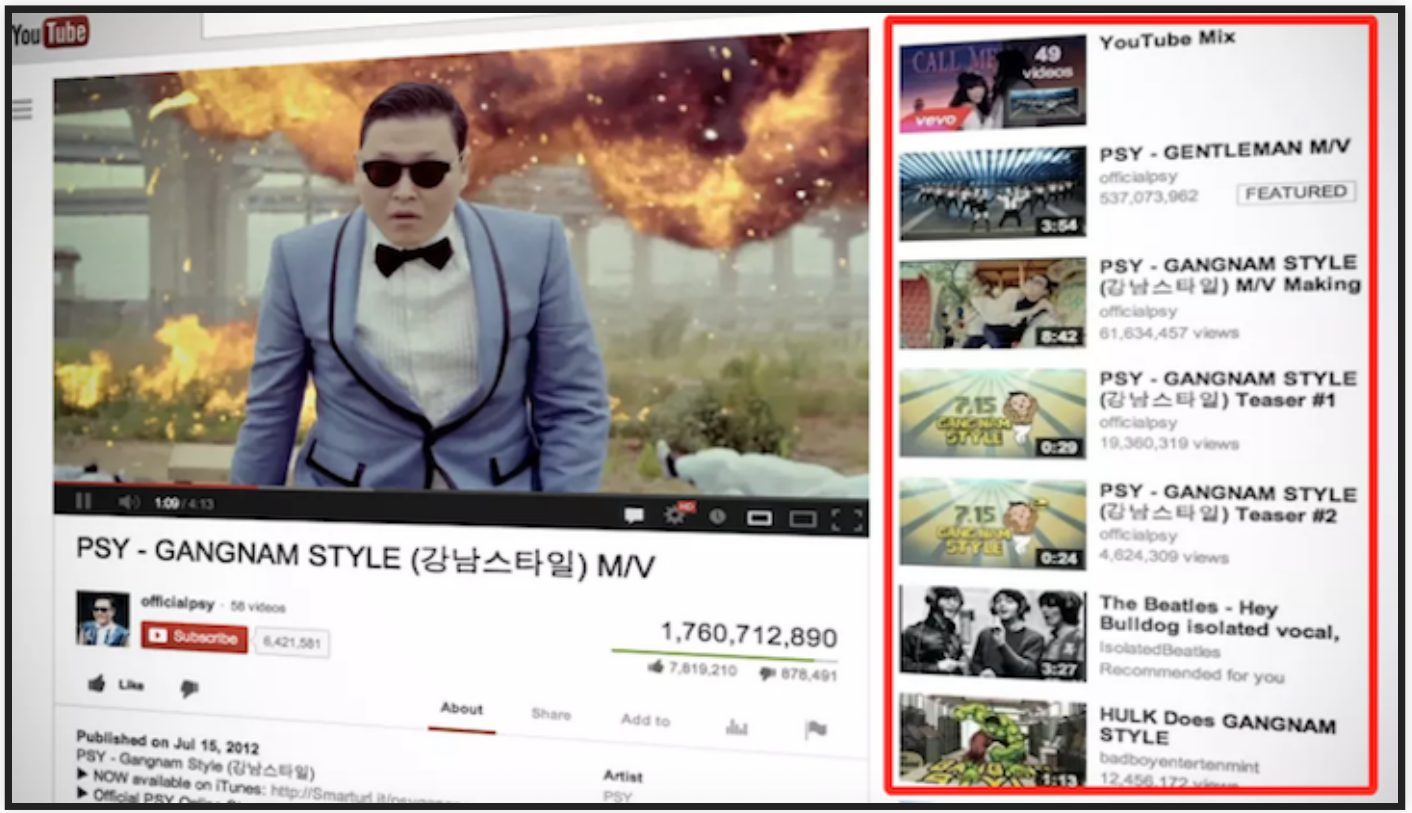
WHICH METHOD FOR CREDIT SCORING?



Linear regression, decision tree, neural network, or k-NN?

Image CC-BY-2.0 by Pne

WHICH METHOD FOR VIDEO RECOMMENDATIONS?



Linear regression, decision tree, neural network, or k-NN?

(Youtube: 500 hours of videos uploaded per sec)

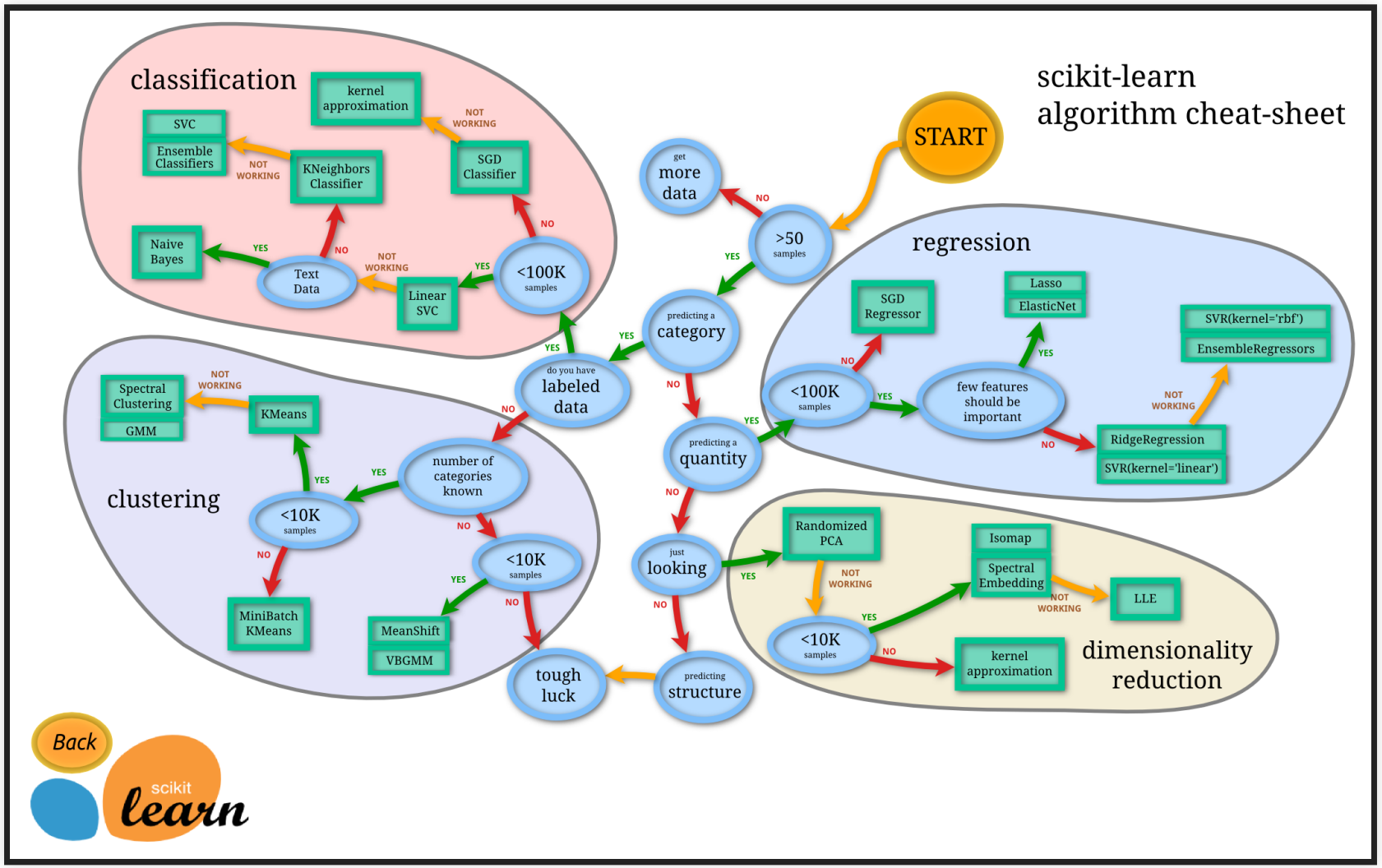
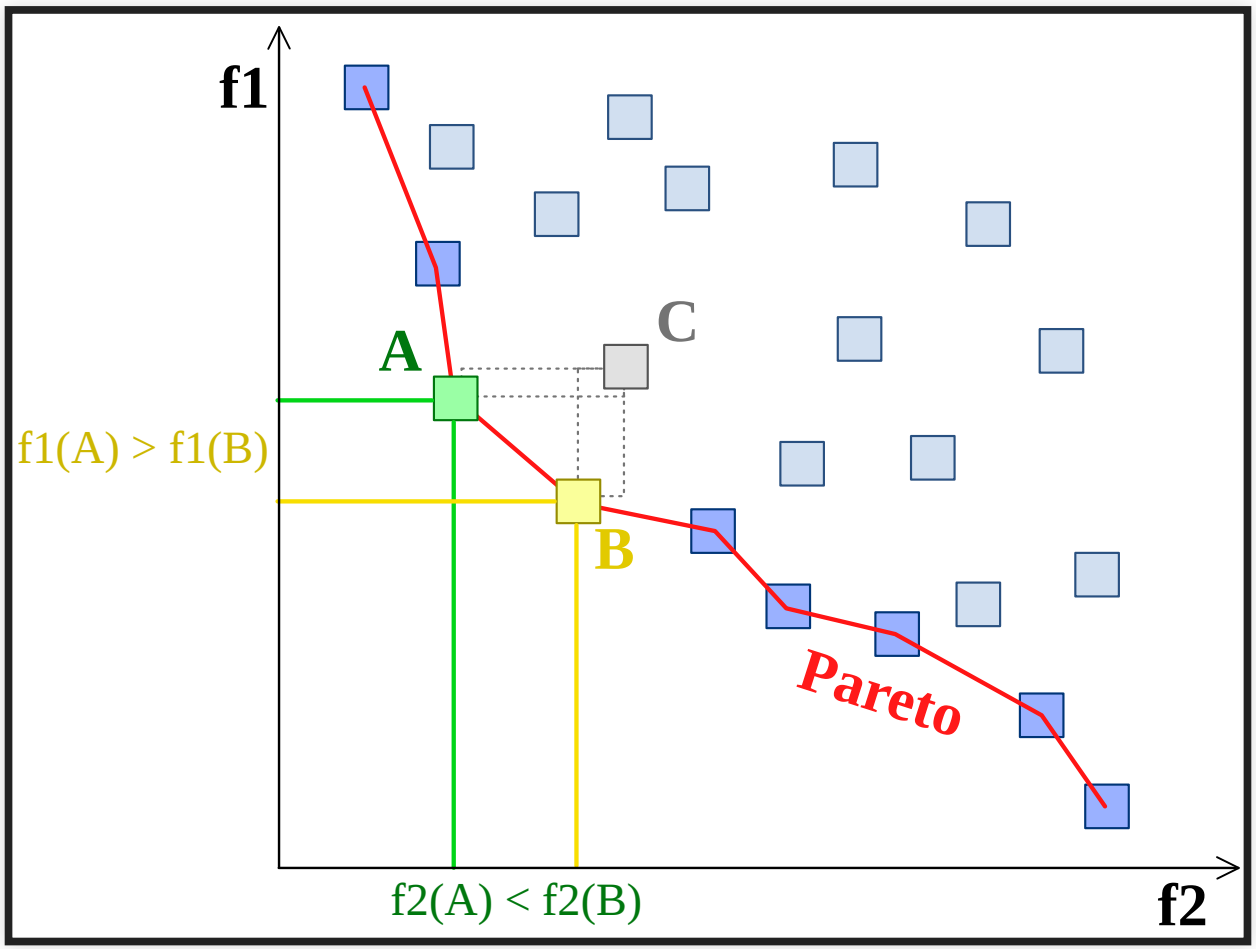
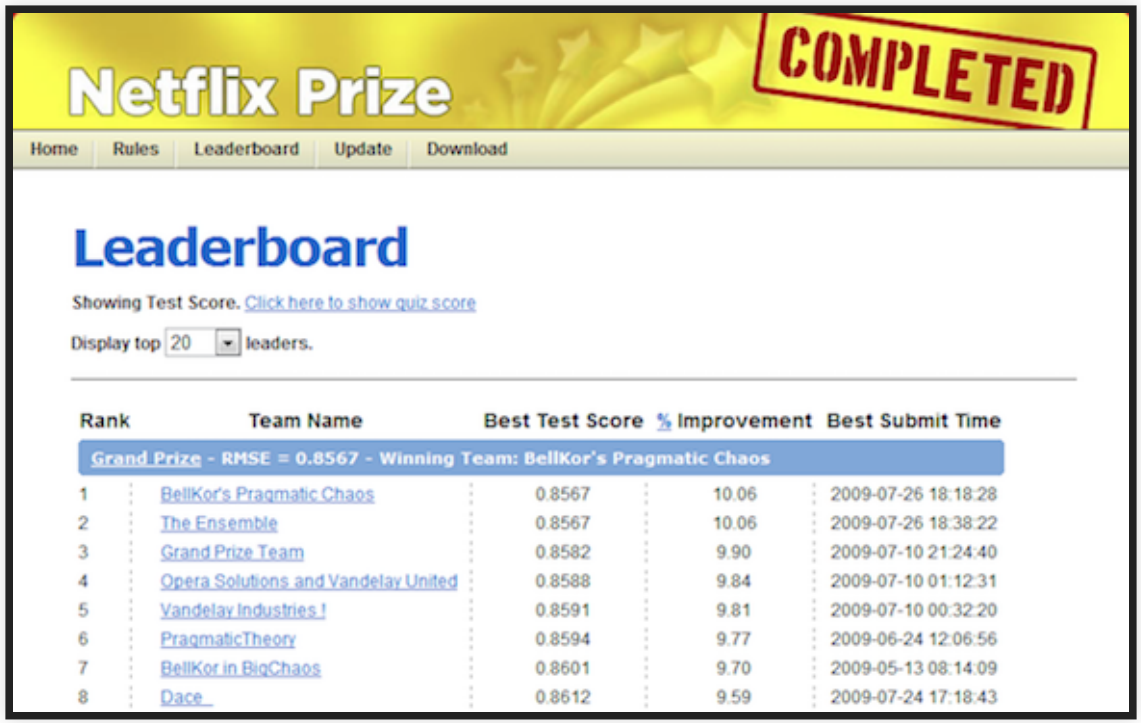


Image: Scikit Learn Tutorial

TRADEOFF ANALYSIS



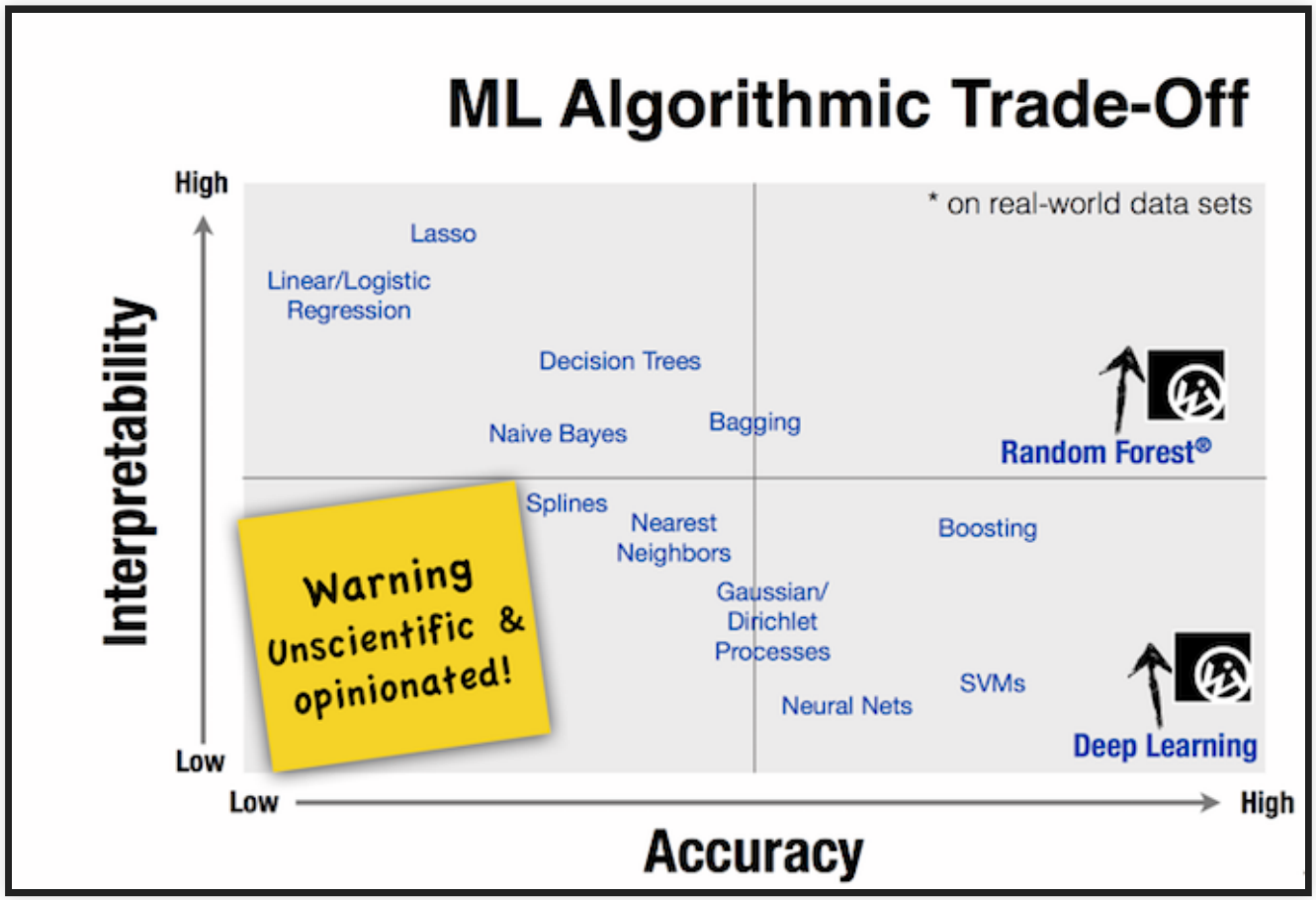
TRADE-OFFS: COST VS ACCURACY



"We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment."

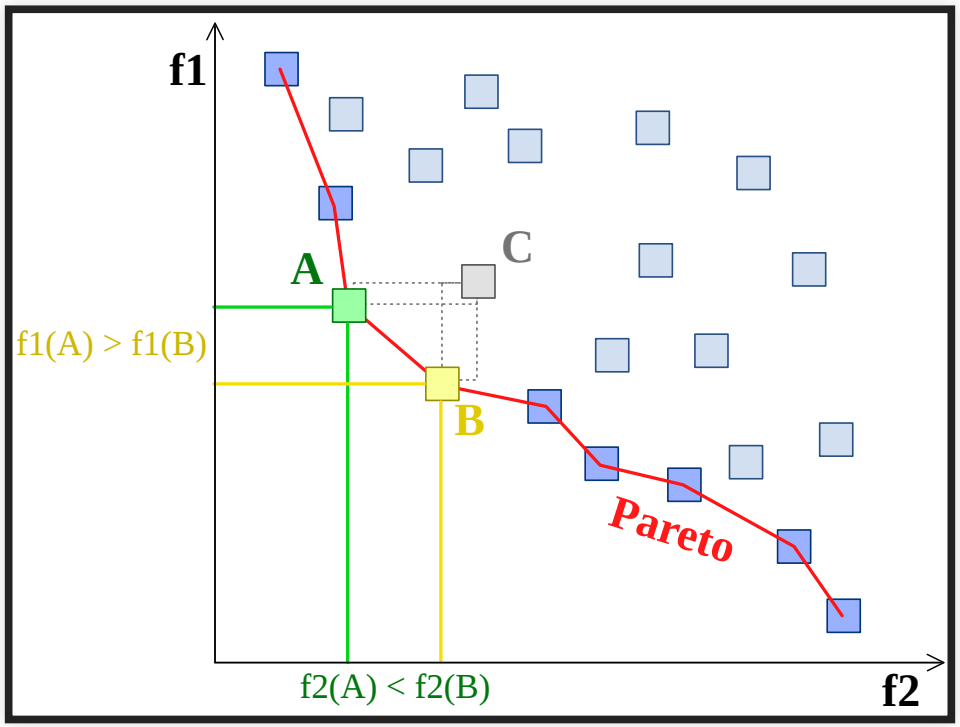
Amatriain & Basilico. Netflix Recommendations: Beyond the 5 stars, Netflix Technology Blog (2012)

TRADE-OFFS: ACCURACY VS INTERPRETABILITY



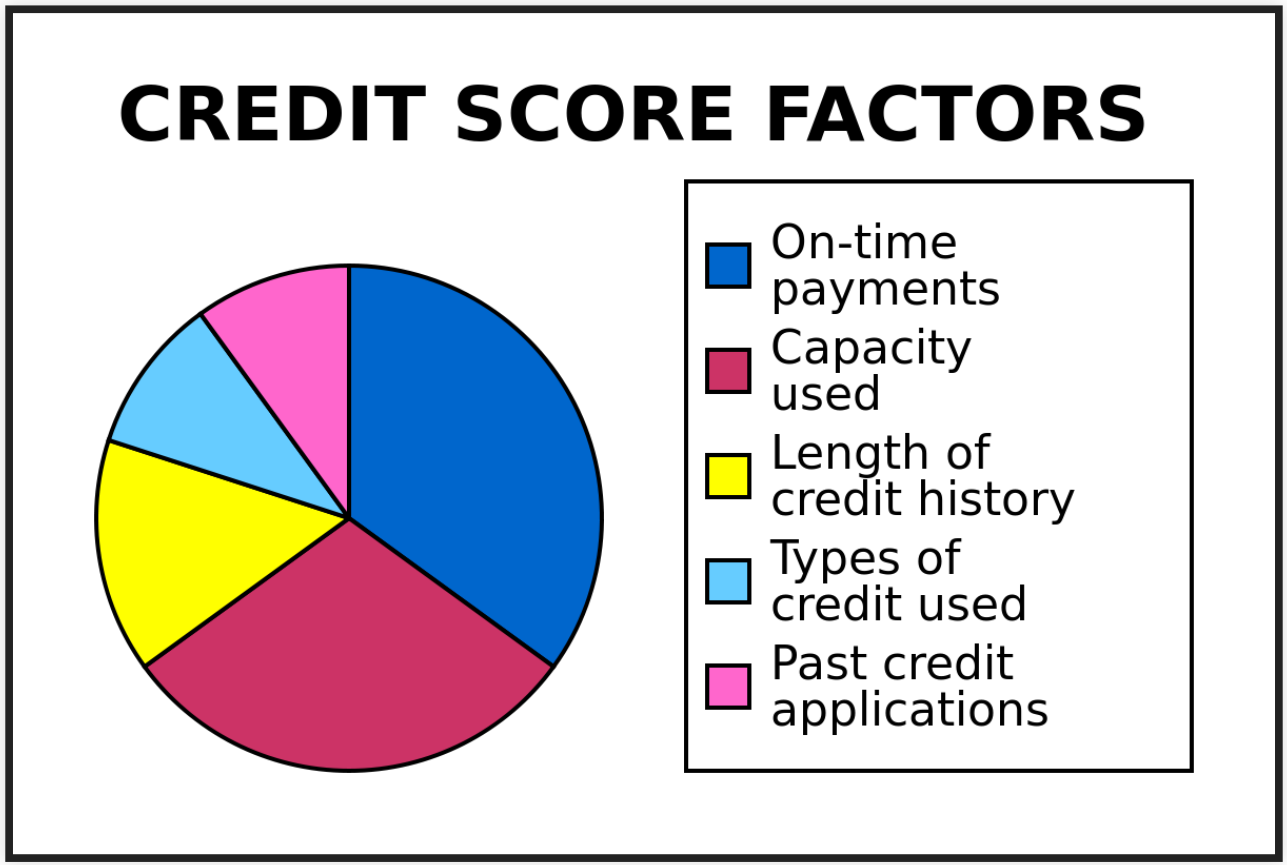
Bloom & Brink. Overcoming the Barriers to Production-Ready Machine Learning Workflows, Presentation at O'Reilly Strata Conference (2014).

MULTI-OBJECTIVE OPTIMIZATION



- Determine optimal solutions given multiple, possibly **conflicting** objectives
- **Dominated** solution: A solution that is inferior to others in every way
- **Pareto frontier**: A set of non-dominated solutions

EXAMPLE: CREDIT SCORING

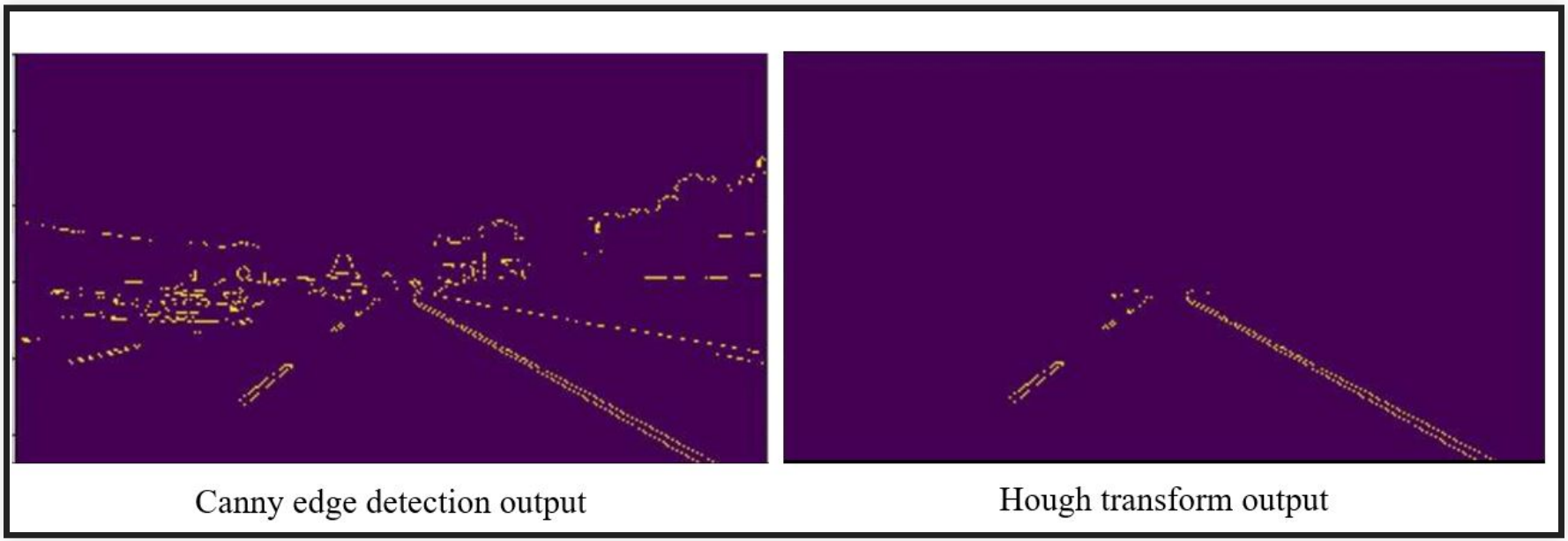


- For problems with a linear relationship between input & output variables:
  - Linear regression: Superior in terms of accuracy, interpretability, cost
  - Other methods are dominated (inferior) solutions

ML METHOD SELECTION AS MULTI-OBJECTIVE OPTIMIZATION

1. Identify a set of constraints
    - Start with problem & project constraints
    - From them, derive design constraints on ML components
2. Eliminate ML methods that do not satisfy the constraints
3. Evaluate remaining methods against each attribute
    - Measure everything that can be measured! (e.g., training cost, accuracy, inference time...)
4. Eliminate dominated methods to find the Pareto frontier
5. Consider priorities among attributes to select an optimal method
    - Which attribute(s) do I care the most about? Utility function? Judgement!

EXAMPLE: LANE DETECTION



Canny edge detection output

Hough transform output

- Constraints: ??
- Invalid solutions: ??
- Priority among attributes: ??

Speaker notes

- Constraints: ML task (classification), inference time (fast, real-time), model size (moderate, for on-vehicle storage)
- Invalid solutions: Linear regression, k-NN
- Priority among attributes: What if accuracy > interpretability = cost?

SUMMARY

- Quality is multifaceted
- Requirements engineering to solicit important qualities and constraints
- Many qualities of interest, define metrics and operationalize
- Survey of ML techniques and some of their tradeoffs
- AI method selection as multi-objective optimization