Data mining

Many definitions

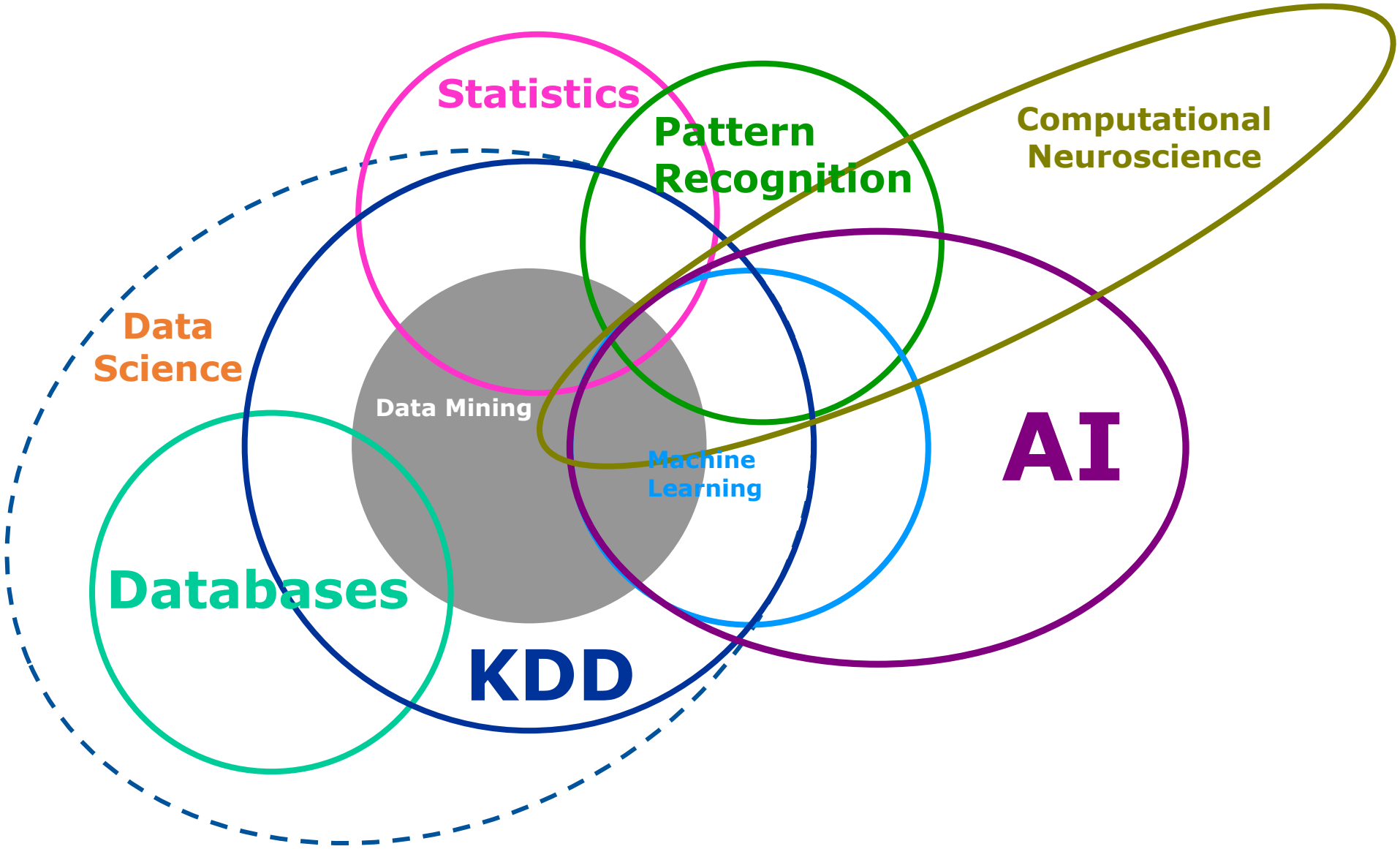
Introduction to Data Mining: “non-trivial extraction of implicit, previously unknown and potentially useful information from data.”

Data mining differs from Statistics due to:

- It’s focus on data storage and data manipulation methodologies
- It’s focus on modeling methods that make few assumptions about the distribution of the training data, but that often have little theoretical support
- It’s focus on commercial applications

In a pop-culture sense, the terms “analytics”, “big data”, “data science”, and “machine learning” are all basically synonyms of data analysis. “Data mining” was perhaps the precursor of these terms.

The data analysis field in general suffers from non-standard vocabulary issues. For instance, see the many different terms used for the rows and columns of a data set.



Machine Learning

Data Mining

SUPERVISED LEARNING	UNSUPERVISED LEARNING	SEMI-SUPERVISED LEARNING

– Regression
  – LASSO regression
  – Logistic regression
  – Ridge regression
– Decision tree
  – Gradient boosting
  – Random forests
– Neural networks
– SVM
– Naïve Bayes
– Neighbors
– Gaussian processes
	– A priori rules
	– Clustering
	  – *k*-means clustering
	  – Mean shift clustering
	  – Spectral clustering
	– Kernel density estimation
	– Nonnegative matrix factorization
	– PCA
	  – Kernel PCA
	  – Sparse PCA
	– Singular value decomposition
	– SOM
		– Prediction and classification*
		– Clustering*
		– EM
		– TSVM
		– Manifold regularization
		– Autoencoders
		  – Multilayer perceptron
		  – Restricted Boltzmann machines

Know y	**Don't know y**	**Sometimes know y**

TRANSFER LEARNING

REINFORCEMENT LEARNING

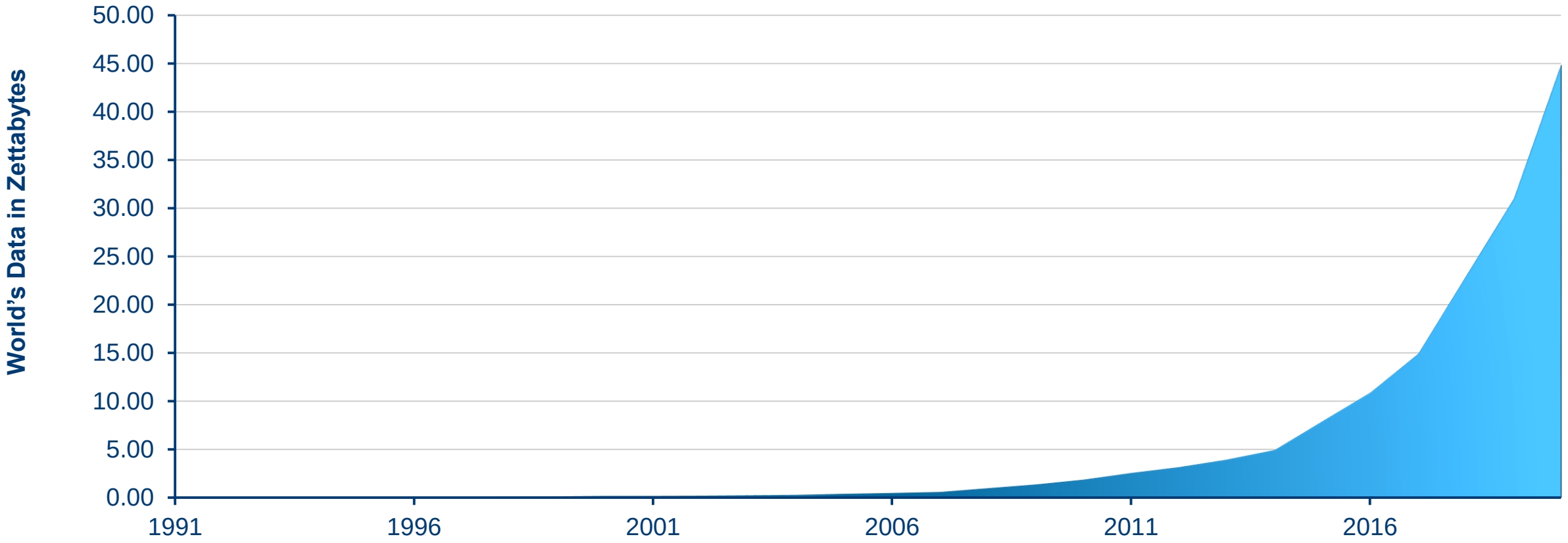
EVOLUTIONARY LEARNING

*In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering.

80/20 rule

Most time is spent cleaning and preprocessing the data!

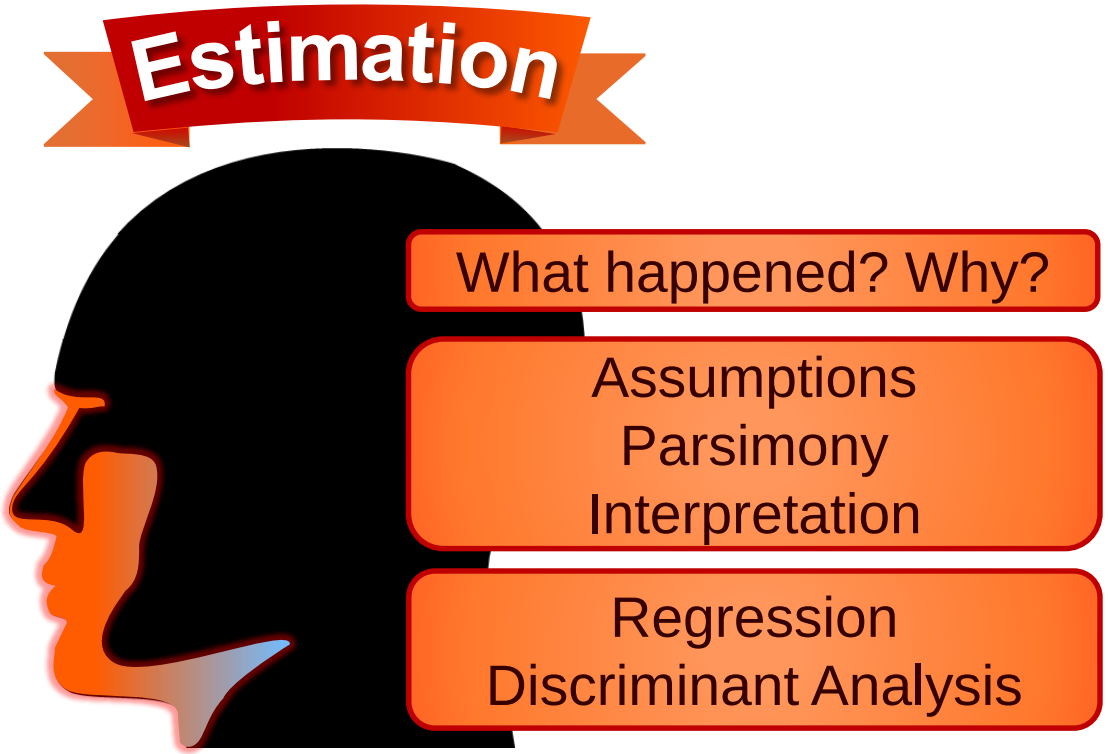
Data growth

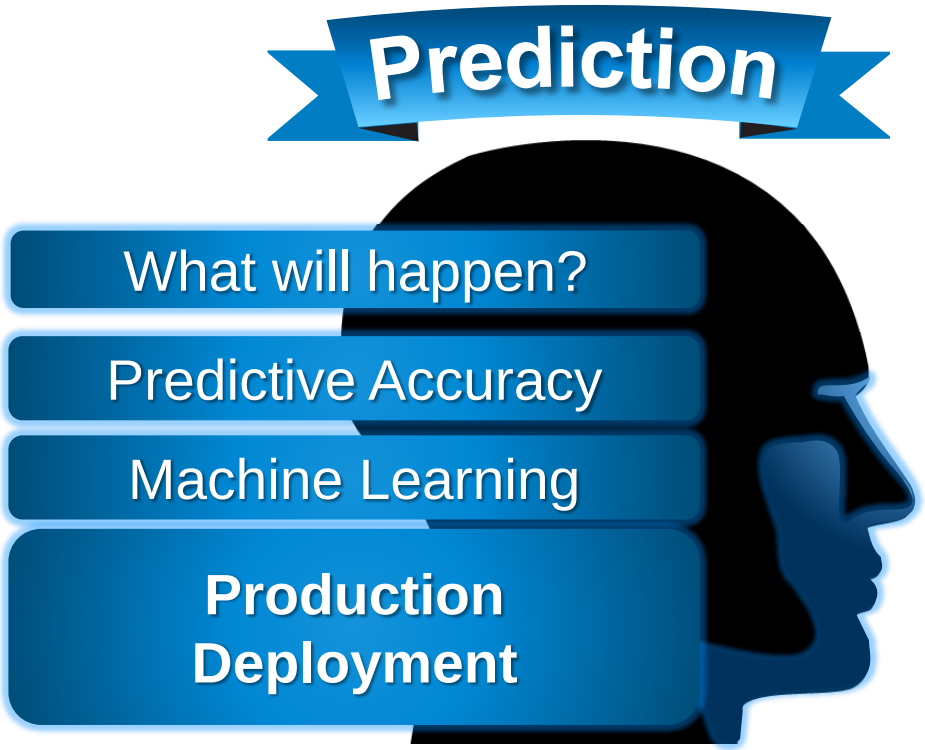


SOURCE: Oracle 2012

ESTIMATION VS. PREDICTION

DIFFERENT GOALS, DIFFERENT MINDSETS



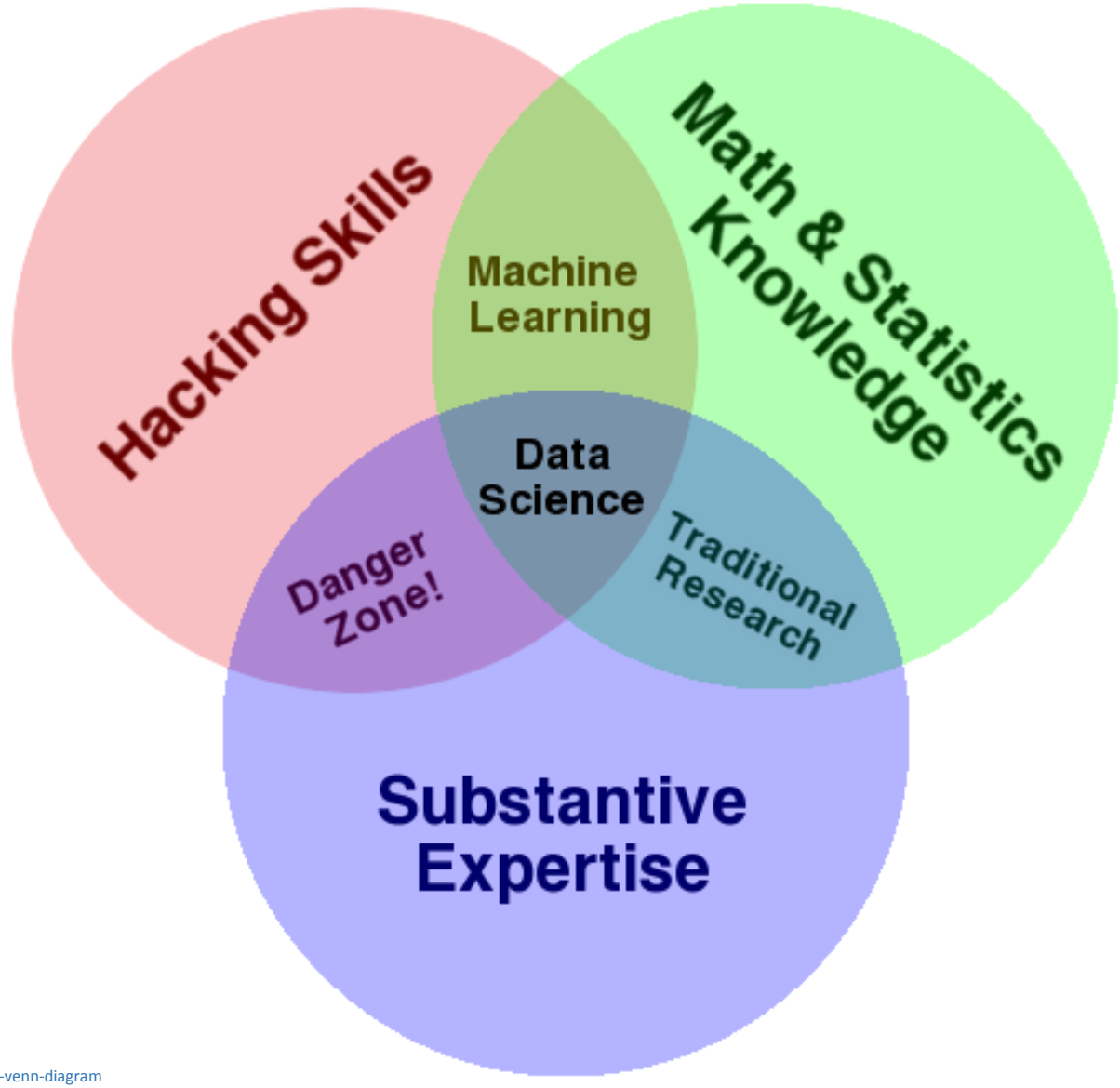


Identify/ Formulate Problem → Data Preparation/ Exploration → Model Building → Deploy Model → Evaluate/ Monitor Model

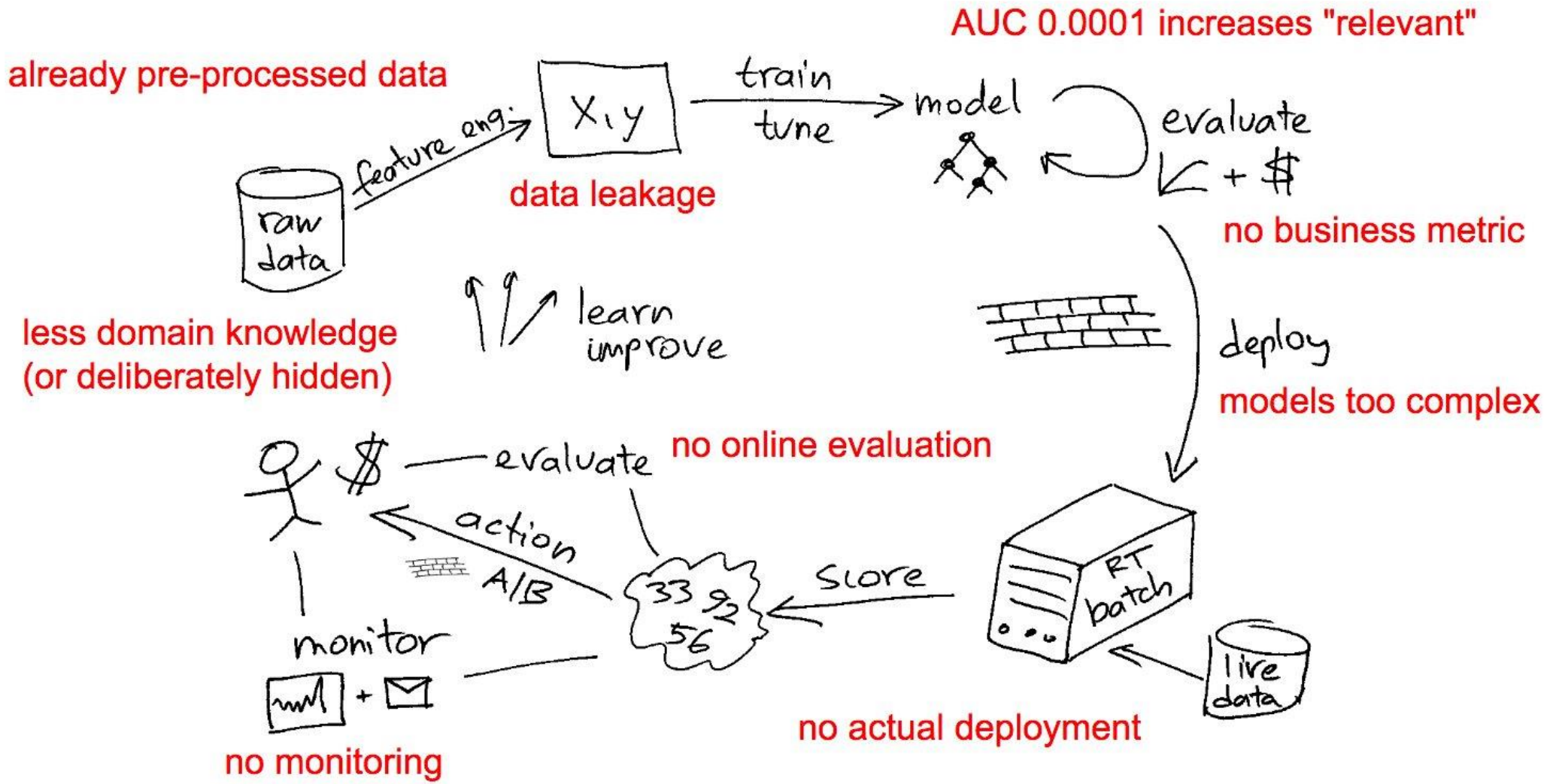
How do we turn our predictions into a production system?

Data science Venn diagram 1.0

Drew Conway, 2010



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



If you do #kaggle to learn #machinelearning, you are missing out on 80% of things you need for ML in real life/production. -- Szilard Pafka

https://twitter.com/DataScienceLA/status/900850613679947777

For lots of business problems GBMs beat deep learning. I was talking about efforts to make GBMs faster (optimized libraries, multicore, GPUs etc.) #machinelearning – Szilard Pafka

https://twitter.com/DataScienceLA/status/936653723568300033

Embrace Automation

- All industries move toward automation
- Algorithms are commodities
- Value-add above and beyond algorithms