NOISE

STOP WORDS

	58	100	250	500	1000
Bagging C4.5	96.06	97.58	99.38	99.52	99.52
C4.5	92.99	96.74	99.18	99.48	99.5
Hyperpipes	72.69	84.23	94.91	97.67	97.55
KNN	94.79	97.36	98.12	97.86	97.26
Multilayer P.	96.21	98.06	99.18	99.52	99.62
Naive Bayes	79.31	91.41	96.98	98.61	98.27
NBTree	94.31	96.85	98.8	99.25	99.4
NN	95.2	97.74	98.54	98.27	97.6
SMO-Poly	92.23	97.41	99.09	99.5	99.78
SMO-RBF	77.39	89.87	94.29	97.77	98.97

CURSE OF DIMENSIONALITY

MOST DATA MINING TECHNIQUES NOT EFFECTIVE FOR HIGH-DIMENSION DATA

Google Research Blog

The latest news from Research at Google

All Our N-gram are Belong to You

1,024,908,267,229 running words of text

CURSE OF DIMENSIONALITY

MOST DATA MINING TECHNIQUES NOT EFFECTIVE FOR HIGH-DIMENSION DATA

"The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience."

The Book of Knowledge

"The curse of dimensionality in the k-NN context basically means that Euclidean distance is unhelpful in high dimensions because all vectors are almost equidistant to the search query vector (imagine multiple points lying more or less on a circle with the query point at the center; the distance from the query to all data points in the search space is almost the same)."

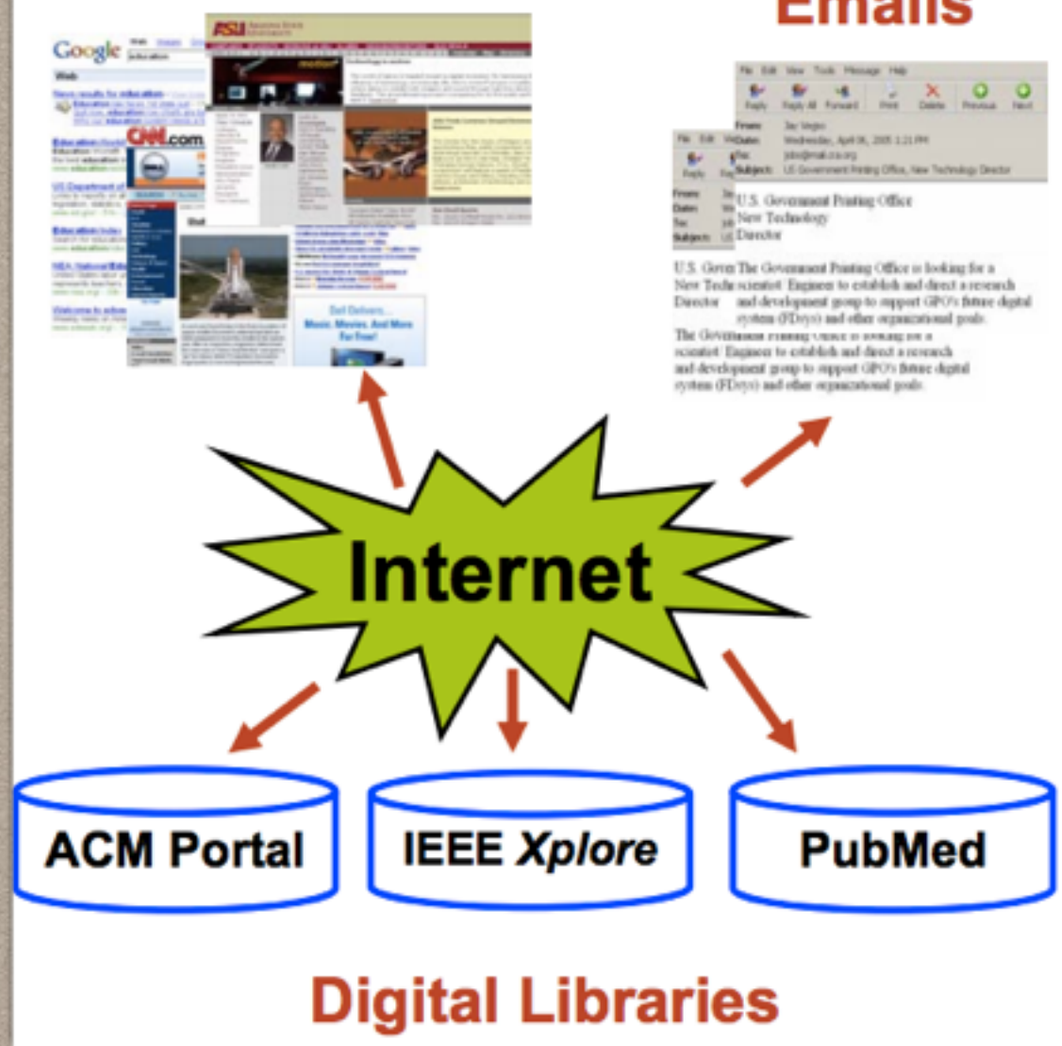
The Book of Knowledge

CURSE OF DIMENSIONALITY

QUERY ACCURACY AND EFFICIENCY DEGRADE RAPIDLY AS THE DIMENSION INCREASES.

APPLICATIONS OF DIMENSIONALITY REDUCTION

- Image retrieval
- text mining
- face recognition
- handwritten digit recognition
- intrusion detection
- microarray data analysis



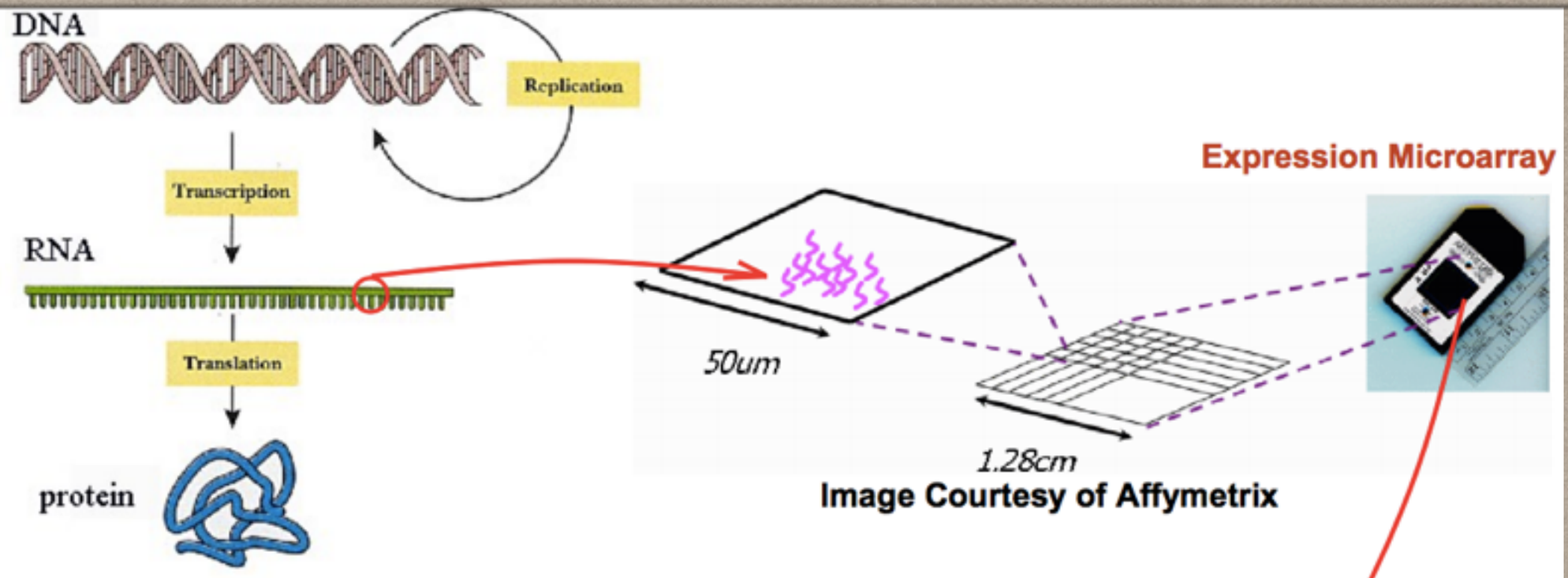
Terms

	T_1	T_2	.....	T_N	C
D_1	12	0	.....	6	Sports
D_2	3	10	.....	28	Travel
⋮	⋮			⋮	⋮
D_M	0	11	.....	16	Jobs

Documents

- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply dimensionality reduction

DOCUMENT CLASSIFICATION



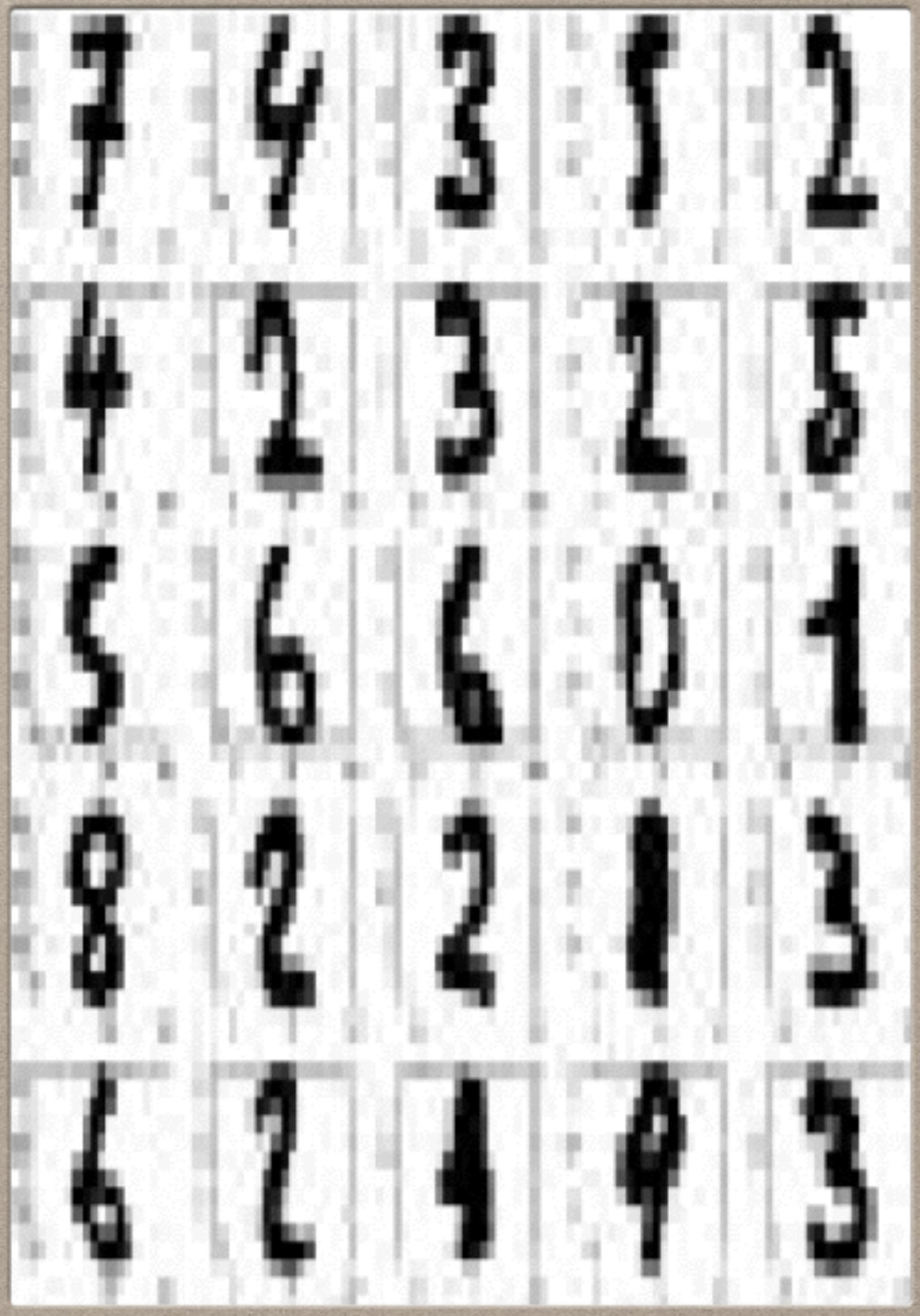
- **Task:** To classify novel samples into known disease types (disease diagnosis)
- **Challenge:** thousands of genes, few samples
- **Solution:** to apply dimensionality reduction

Sample \ Gene	M23197_at	U66497_at	M92287_at	...	Class
Sample 1	261	88	4778	...	ALL
Sample 2	101	74	2700	...	ALL
Sample 3	1450	34	498	...	AML
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.

Expression Microarray Data Set

GENE EXPRESSION

MICROARRAY ANALYSIS



MAJOR TECHNIQUES OF DIMENSIONALITY REDUCTION

- feature selection
- feature extraction (aka reduction)

FEATURE SELECTION

- **Definition**: A process that chooses an optimal subset of features according to a objective function
- **Objectives**:
  - To reduce dimensionality and remove noise
  - Improve learning speed
  - Improve accuracy

HOW?

TRYING TO FIND THE OPTIMAL SUBSET

‘OPTIMAL’ FOR ENTIRE WORLD BUT WE ONLY HAVE OUR SMALL DATA SET.

SOLUTION

EXHAUSTIVELY EXAMINE ALL THE SUBSETS AND PICK THE BEST ONE.

PROBLEM

THAT WORD ‘EXHAUSTIVELY’
AND THE COMBINATORIAL NATURE OF
‘EXHAUSTIVELY’

EXAMPLE

- 100 features and I want to select 10

- ??

EXAMPLE

- 100 features and I want to select 10
- 100 x 99 x 98 x 97…
- 62,815,650,955,529,472,000

EXAMPLE

- Let’s say our training set has 10,000 instances
- takes 10 seconds to training and evaluate the classifier.
- take trillions of years to find the optimal subset.

MORE REALISTIC EXAMPLE

- 100,000 features and I want to select 100
- roughly 100k x 100k x 100k x 100k...
- $100{,}000^{100}$

~~OPTIMAL - EXHAUSTIVE~~

PRETTY GOOD - HEURISTIC

FEATURE RANKING

- weighting and ranking individual features
- selecting top-ranked ones for feature selection
- efficiency (not n!) what is it?

FEATURE RANKING

- weighting and ranking individual features
- selecting top-ranked ones for feature selection
- efficiency O(n)
- easy to implement
- Disadvantage: unable to consider correlation between features. (height in feet vs. height in inches). or height and weight work well as a pair.

ENTROPY & INFORMATION GAIN

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i))$$

$$IG(X|Y) = H(X) - H(X|Y)$$

METHOD

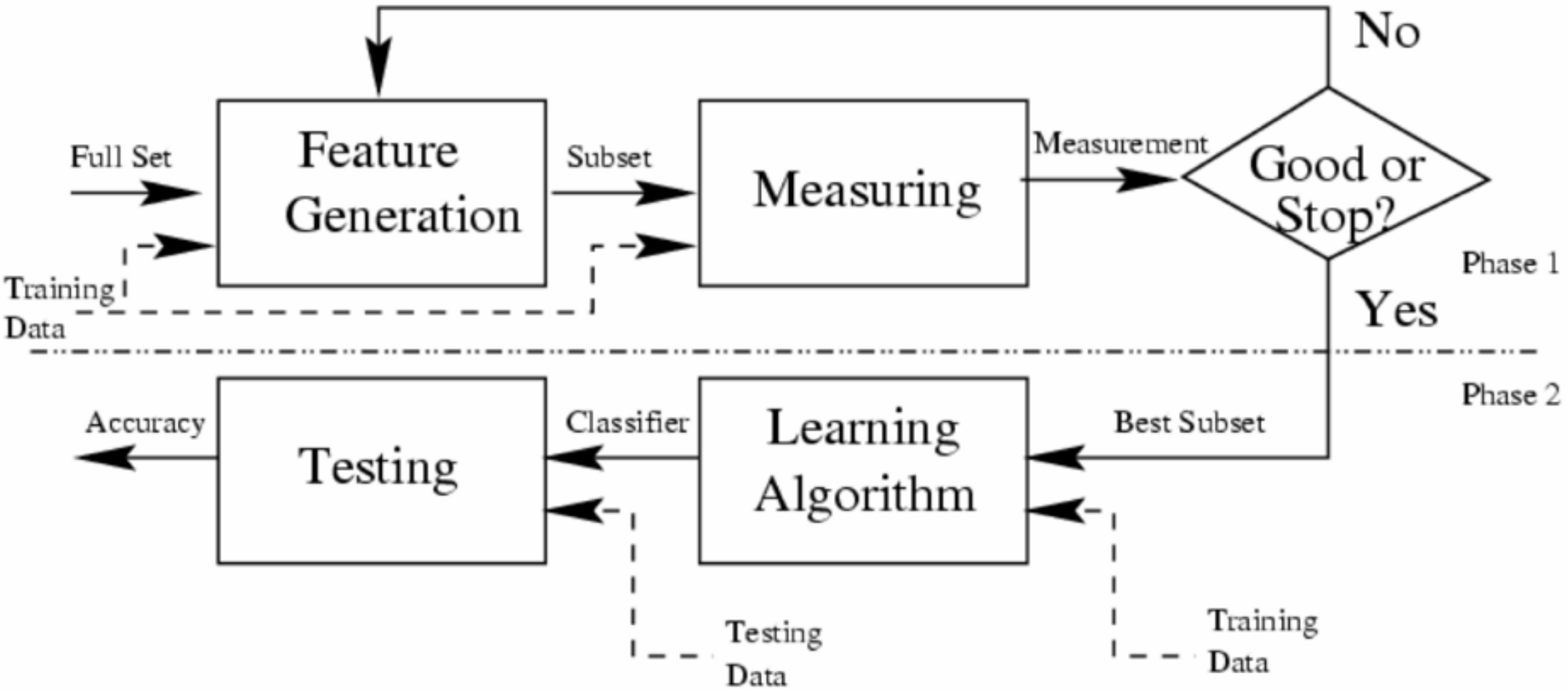
- 100,000 features we want to reduce to 100
- compute information gain of all 100k features
- pick the 100 best

CLASSIFICATION ACCURACY METHOD

- 100,000 features and want to reduce to 100
- run classifier using 1 feature at a time
- pick 100 best performing features (based on accuracy)
- ~1 day of processing
- (improvement over trillions of years)

CLASSIFICATION ACCURACY METHOD

- directly aimed at improving accuracy
- drawback 1: time consuming (maybe)
- drawback 2: dependent on classifier being used (maybe this is a good thing)



CAN AUTOMATE PROCESS

WRAPPER

THAT WAS FEATURE SELECTION

EASY

FEATURE EXTRACTION

FEATURE REDUCTION

FEATURE REDUCTION

- all the features are used
- we map (linear map) a higher dimensional space to a lower dimensional one.
- more on this in a later class.