Challenges and opportunities in database design and interlaboratory studies on trace evidence fibers

Stephen L. Morgan

Department of Chemistry and Biochemistry,
University of South Carolina, Columbia, SC 29208
morgansl@mailbox.sc.edu

Acknowledgement and disclaimer

The research reported herein was supported by Award No. 2010-DN-BX-K220 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice.

The collaboration and contributions of the following individuals are also recognized:

Nathan C. Fuenffinger, *Department of Chemistry & Biochemistry*
The University of South Carolina, Columbia, SC 29208
John V. Goodpaster, *Forensic and Investigative Sciences Program*
Indiana University Purdue University Indianapolis (IUPUI)
Indianapolis, IN 46202; jvgoodpa.edu@iupui
Edward G. Bartick, *Department of Forensic Sciences, George Washington University, Washington, DC, 20007, ebartick@email.gwu.eduartick@email.gwu.edu*
***Lieutenant Jennifer Nates**, Forensic Services, Trace Evidence, South Carolina Law Enforcement Division, Columbia, SC*

William Edwards Deming: On profound knowledge

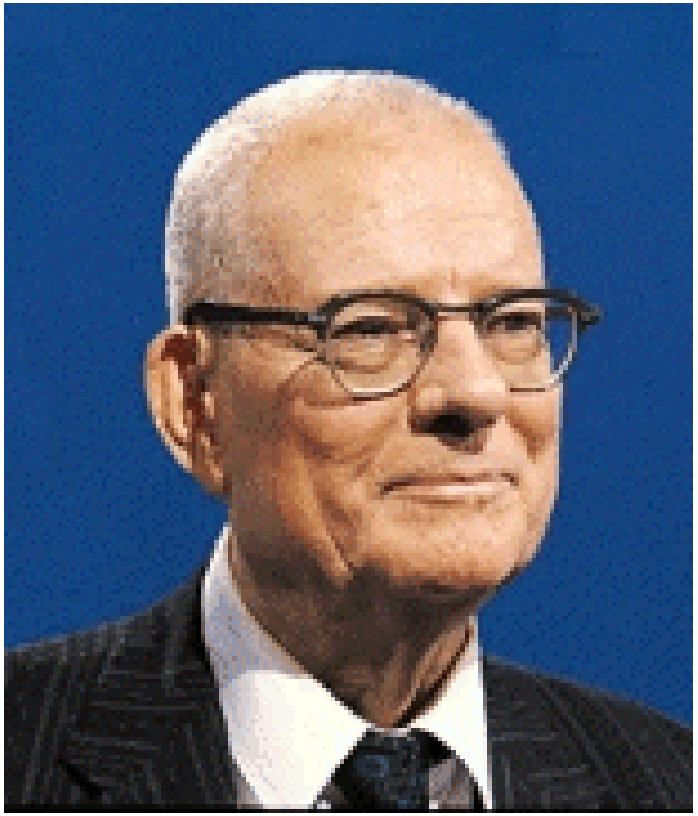
Deming advocated that all managers need to have what he called a “System of Profound Knowledge,” consisting of 4 parts:

Appreciation of a system: understanding the overall processes involving suppliers, producers, and customers (or recipients) of goods and services;

Knowledge of variation: the range and causes of variation in quality, and use of statistical sampling in measurements;

Theory of knowledge: the concepts explaining knowledge and the limits of what can be known.

Knowledge of psychology: concepts of human nature.



Working Hypothesis

If enough is known about the distribution of a population from which questioned and known fibers originate, then knowledge of multiple associated characteristics (physical, optical, or spectroscopic) can be employed to decrease the random probability of a match occurring solely by chance.

Caveat 1: Usable and Realistic

To consider any trace evidence database usable and realistic, it is necessary to have a large number of diverse, and representative samples, that are common in use within the geographic region where the crime occurred.

Caveat 2: Impediment to source

Establishing a collection of fibers that is truly representative is complicated by rapid changes in manufacturing practices and globalization of textile production: the population is a moving target of indeterminate size and evolving diversity.

Caveat 2: Impediment to source matching

"…a 'match' means only that the fibers could have come from the same type of garment, carpet, or furniture; it can provide class evidence…"[34], and that "fiber analyses are reproducible across laboratories because there are standardized procedures for such analyses. [National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*, 2009, National Academy Press: Washington, D.C.]

Understanding the Population

1. How is the item made?
2. Who are the major producers of the item?
3. Who are the major distributors of the item?
4. Where is the item sold?
5. Does the item carry any markings that are traceable to a retailer/distributor/manufacturer?
6. Is the item regulated or approved by a third party?
7. Has the formulation of the item changed over time?
8. How common is the item – how many are distributed/sold?
9. In what regions is the item more common?
10. What is the typical “lifecycle” of an item? How long is it used prior to disposal?

Information sources include: the manufacturing industry, literature, industry representatives, local merchants, and other forensic scientists.

Acquiring a Representative Collection

1. Acquire multiple items from the major manufacturers/distributors;
2. Geographically diverse – cover the area from which your samples originate;
3. Ideally, the make-up of your collection reflects market share and availability;
4. Acquire multiple items of the same type;
5. How many is enough?

Using Multiple Analytical Methods

1. Consult the literature;
2. Assess your available equipment
3. Do not assume that your preferred method is the most discriminating!
4. Use the same sample set and number of replicates for each technique;
5. Ideally, use orthogonal techniques (inorganic/organic, spectroscopy/ chromatography/mass spectrometry, *etc*.).

Reproducibility and differentiability

1. Assess sample size (do smaller samples exhibit more heterogeneity?);
2. Assess instrumental conditions (what conditions give the best precision?);
3. Always acquire replicates (how many is enough?); and,
4. Apply statistical techniques.

Monitor changes in the sample population over time, environmental exposure, or other relevant variables:

1. Changes in manufacturers, distributors and retailers;
2. Changes in formulations; and,
3. Regularly acquire more samples!

Database Design

The crucial issue in designing a database is to start with its purpose and proposed application.

Databases are specific in their content and application. For that content to be applicable, success requires input from all *stakeholders* and potential *users* who will eventually be *owners*.

Target users should be encouraged to discuss *their* needs and asked to provide incremental feedback during development.

Don't forget that users require training to use the database, and the biggest cost may be time – time away from their real jobs.

Developing training materials for a specialized data base takes time and involves costs, and again, the stakeholders should be involved in that process also.

"Quality is *everyone's* responsibility."

W. Edwards Deming

When is the database done?

Software is never done. Code and bits are only provisional and require maintenance.

We've all noticed that software is always being updated these days. It's not any different with databases:

- Perhaps the database structure design, or needs to changes to accommodate new data objects.
- Maybe the software used to create it is outdated, or doesn't talk nicely to newer protocols.

Data bases should be adaptable and capable of change to accommodate new data objects as needed.

- For example, how will missing data be handled in your database?
- Does missing data invalidate a data object or does partial data supply partial information?

Finally, the database must remain relevant to current forensic experience.

It is impossible to make anything foolproof because fools are so ingenious.

Information = data + meaning + constraints

The wisdom one needs to interpret data correctly does not come directly out of the database. Wisdom requires insight into relationships inherent in the data – how the data elements fit together into a *gestalt*.

The most important maxim for translating information from databases to knowledge is:

context matters.

USC Fiber database

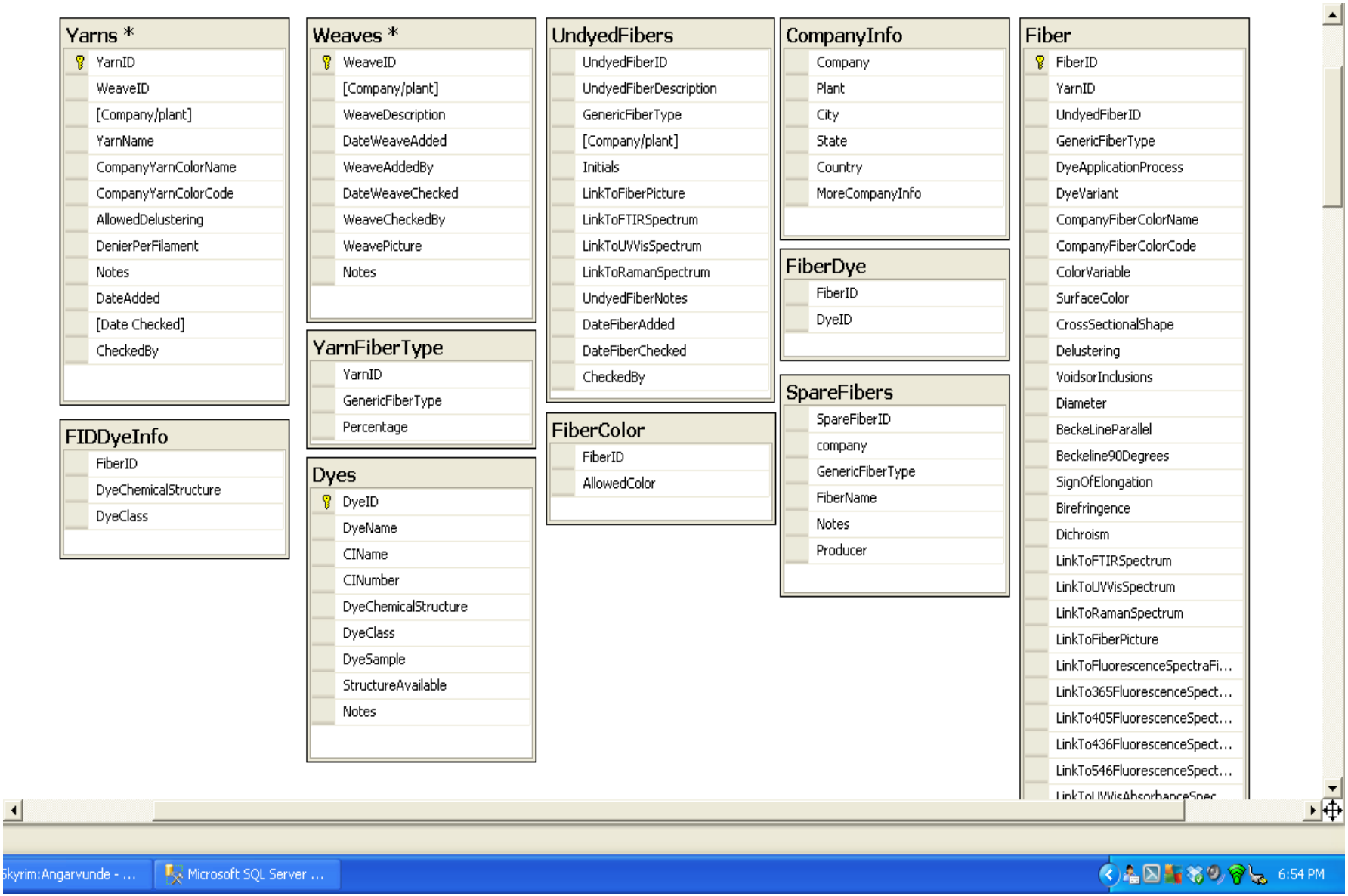
Samples of the four most common fibers encountered in forensic trace evidence: acrylic, cotton, nylon, and polyester; dyes available for most samples. For about 1,000 fibers, optical microscopy data, dye information, and spectra are organized in a web-based database.

More than 500 additional samples including whole swatches, polymer staple materials, and undyed and dyed fibers, but no dye samples.

SLED (Columbia, SC) donated more than 1,400 residential carpet samples obtained from Lowe's consisting of multiple shades of different colored fiber polymers.

Dr. Hal Deadman (GWU) donated samples from a collection of 200 auto carpet fibers collected from junk yards to Northern Virginia. Automobiles models were identified and VIN numbers recorded.

Fiber object database diagram



Fiber Selection Pages

The *Main fiber selection* page displayed upon login is a search page that solicits 1 to 5 search values from the user.

All fibers in the database with the specified characteristics are returned in a *Selected Fiber* window which displays the result set returned from the database—a list of fibers that match the characteristics previously selected. The check list at the bottom left of the screen enables further filtering of the fields that are returned in the results grid.

Search Results for Fiber Diameters

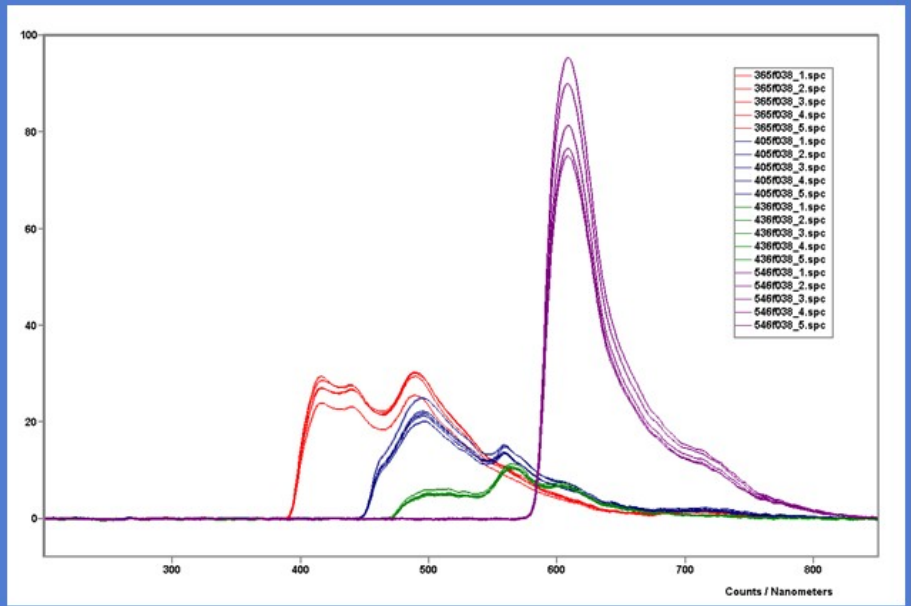
Fiber Details Page for fiber ID 38

Forensic Fiber Database

Current User Profile: david | Logout

Fiber Details:

FiberID	38
YarnID	21
UndyedFiberID	2
GenericFiberType	polyester
DyeApplicationProcess	exhaust
DyeVariant	Regular
CompanyFiberColorName	
CompanyFiberColorCode	
ColorVariable	☐
SurfaceColor	☐
CrossSectionalShape	trilobal
Delustering	many
VoidsorInclusions	no
Diameter	28.75
BeckeLineParallel	G
Beckeline90Degrees	G
SignOfElongation	+
Birefringence	medium
Dichroism	☐
FiberNotes	
DateFiberAdded	1/13/2003 12:00:00 AM



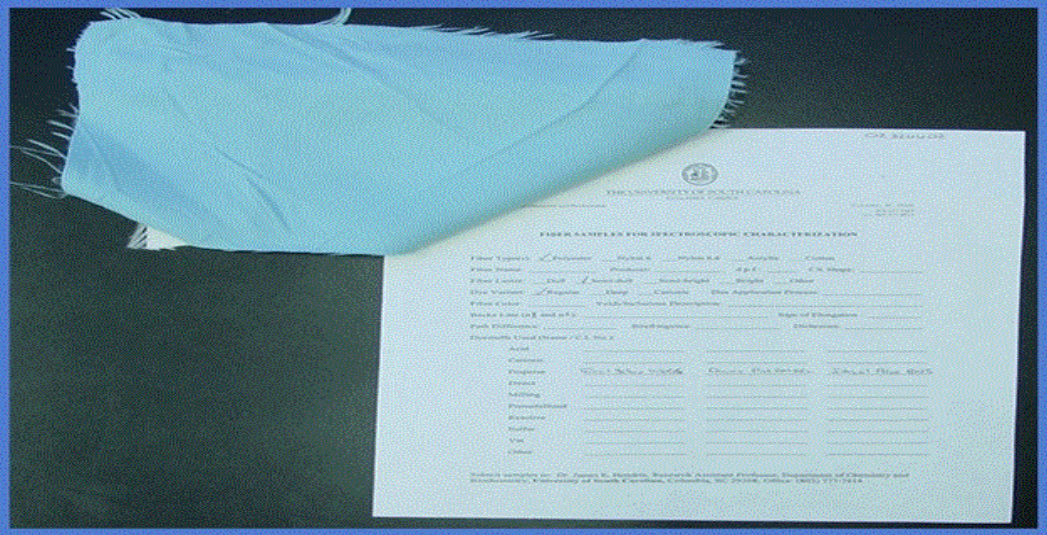
Dye Details Page for Fiber ID 38

Dye Details:

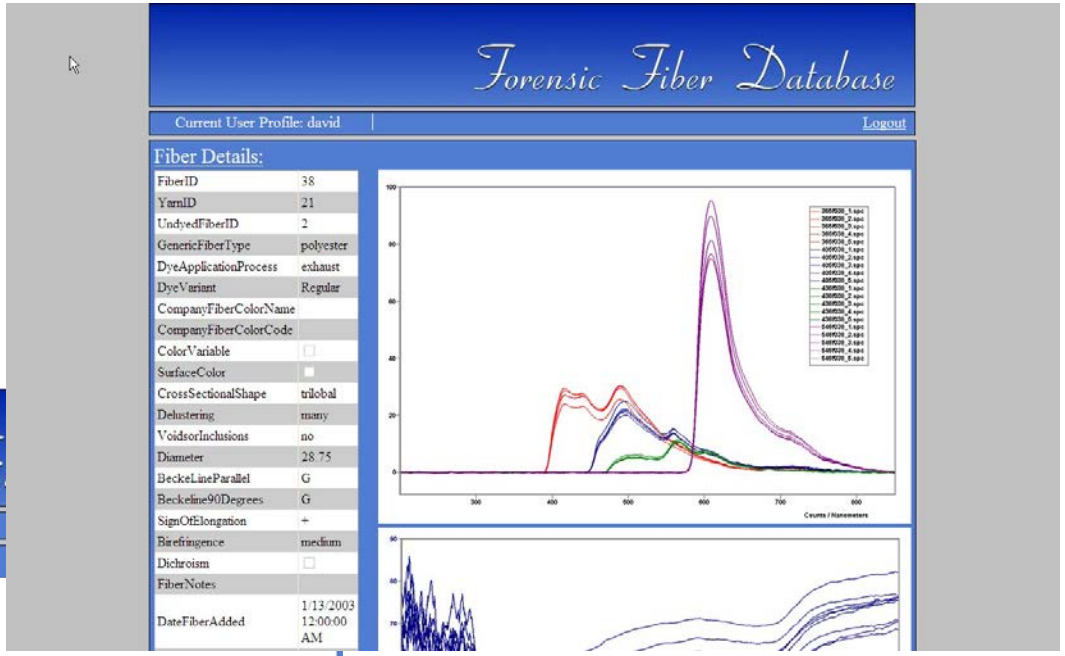
	DyeID	DyeName	CIName	CINumber	DyeChemicalStructure	DyeClass	Notes
Select	20	Dianix Pink AM-REL	not in C.I.			disperse	old name: Palanil Brill Pink E-REL
Select	22	Intrasil (Brill) Blue BNS	not in C.I.			disperse	
Select	109	Terasil Yellow W-6GS	C.I. Disperse Yellow 114			disperse	powder and paste available in our sample library

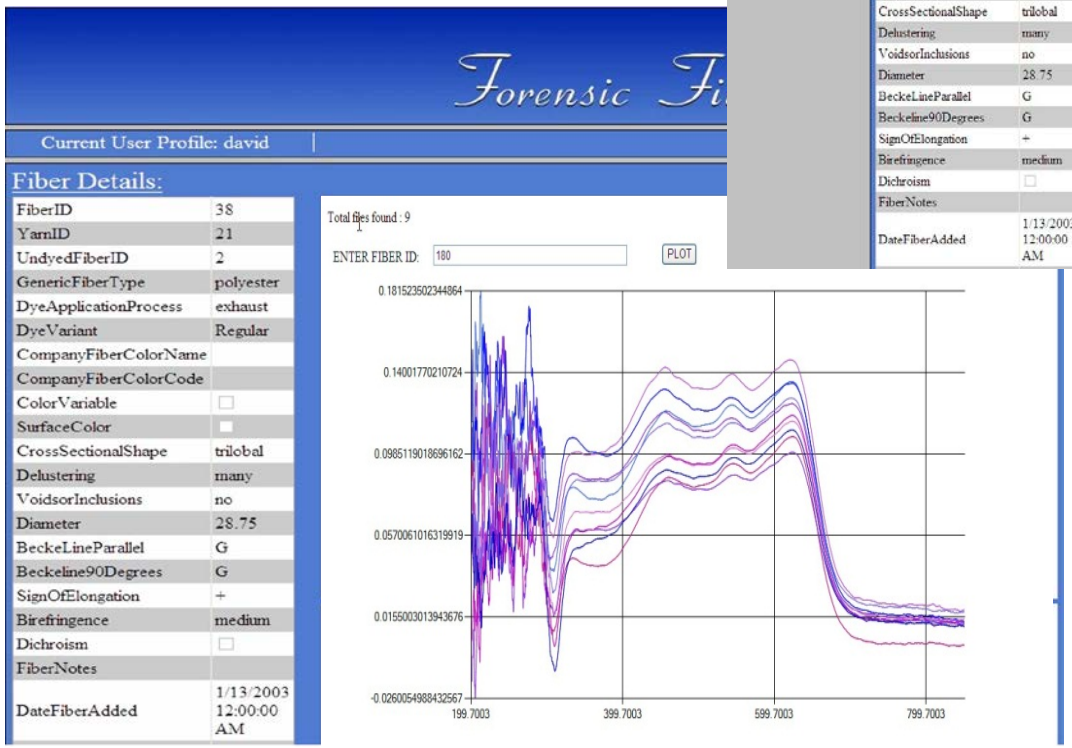
Weave Details:

WeaveID	3
Company/plant	Milliken/CDL
WeaveDescription	
DateWeaveAdded	1/13/2003 12:00:00 AM
WeaveAddedBy	AAN
DateWeaveChecked	6/5/2003 12:00:00 AM
WeaveCheckedBy	JEH
Notes	02326602
YarnID	21
WeaveID1	3
Company/plant1	Unknown
YarnName	
CompanyYarnColorName	
CompanyYarnColorCode	
AllowedDelustering	semi-dull
DenierPerFilament	0



Fiber details, plots, and data export





Information is data distilled

"Having data in a database is not the same thing as knowing what to do with it." [Kay, Roger L. "What is the meaning?" *Computerworld*, 17 October, 1994].

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

– T. S. Eliot, *Choruses from the Rock*

Raw data does not help anyone to make a decision until you can reduce it, using relevance and context, to a higher-level abstraction.

Similar to the number of ways that beer can be brewed, there are a lot of ways one can distill data.

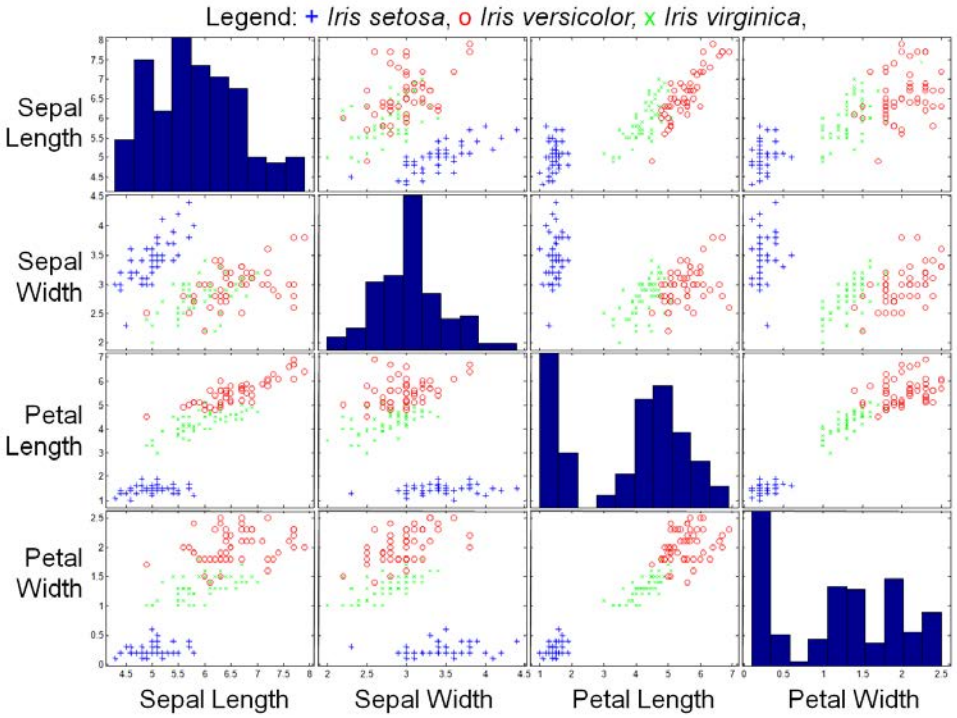
Why not univariate?

http://en.wikipedia.org/wiki/File:Iris_versicolor_2.jpg

QUESTION 1: Instead of dealing with a *multi*-dimensional problem, why not just examine one variable at a time?

ANSWER: Multivariate data can be misleading when examined single variable at a time. With multivariate data, the single variable at a time approach may fail to detect the underlying multivariate structure, whereas a multivariate approach will reveal the truth.

Consider Fisher's *Iris* data set, which has four measurements on each of 50 samples of three types of *Iris*. The individual variable histograms (in blue) may (or may not) show group separation; the two-variable scatter plots hint at the ability to separate groups. Thus, we see trends, or correlations, in plots of Fisher's data.



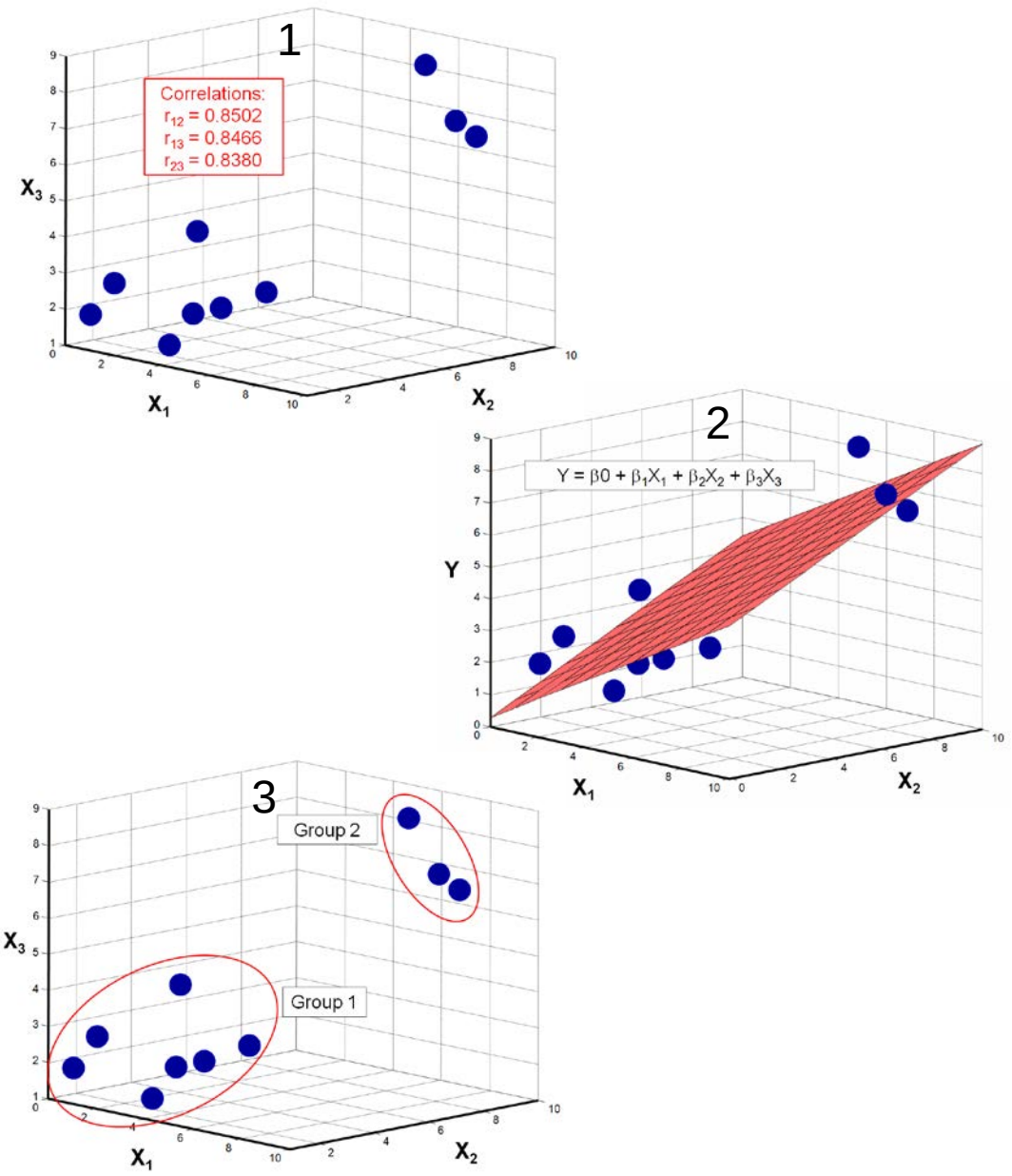
Search for meaningful structure

Typical objectives:

1. Discovering patterns, systematic structure, correlations, trends, or regularities in data involving two or more variables.

2. Testing models that describe relationships among experimental variables and measured responses. Accuracy of prediction.

3. Evaluation of models that describe the relationships among two of more groups (or classes) of objects based on their multivariate patterns.



Research objectives

- **Conduct interlaboratory studies** to evaluate the application of pattern recognition and machine learning tools to forensic fiber examinations based on UV/visible microspectrophotometry

  Provide statistical measures of dissimilarity of fiber spectra along with visualization of comparisons to support decision-making.

- **Determine best performing spectral pre-processing approaches and multivariate methods for fiber discrimination**

  Numerous pre-preprocessing methods have be applied to multivariate data in chemometrics. Which are necessary and what are the effects on discrimination.

Research objectives (continued)

— **Evaluate intra- and interlaboratory variability** associated with microspectrophotometry of textile fibers.

Can consistent conclusions be made from independent analyses? If the same samples are examined in different laboratories, with different instruments, are the results compatible?)

— **Document intra- and inter-laboratory consistency** in UV/visible spectra of fibers with classification error rates.

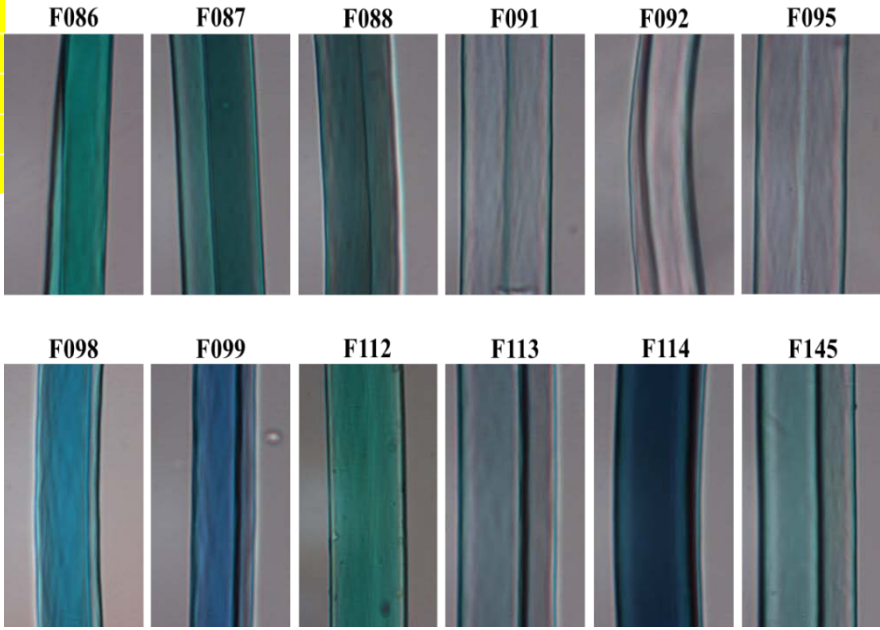
Can classification models be transferred between laboratories, with potential savings in time and resources for forensic analyses? Difficulties in using a model developed in one laboratory to classify data in another laboratory can arise from differences in sample preparation, environmental conditions, and instrumental response.

Comparisons of Interlaboratory Fiber Discrimination

Cationic dye composition for 12 blue acrylic fibers (“Y” indicates dye presence).

	Cationic dye									
Fiber	Blue 3	Blue 41	Blue 60	Blue 147	Red 18	Red 29	Red 46	Yellow 21	Yellow 28	Yellow 29
086	Y				Y				Y	
087		Y					Y		Y	Y
088		Y					Y		Y	
091		Y				Y		Y		
092			Y			Y			Y	
095				Y		Y			Y	
098	Y			Y						
099				Y			Y		Y	
112	Y				Y				Y	
113		Y				Y			Y	
114		Y			Y				Y	
145	Y						Y		Y	

Microscope images of 12 blue acrylic fibers (40×).



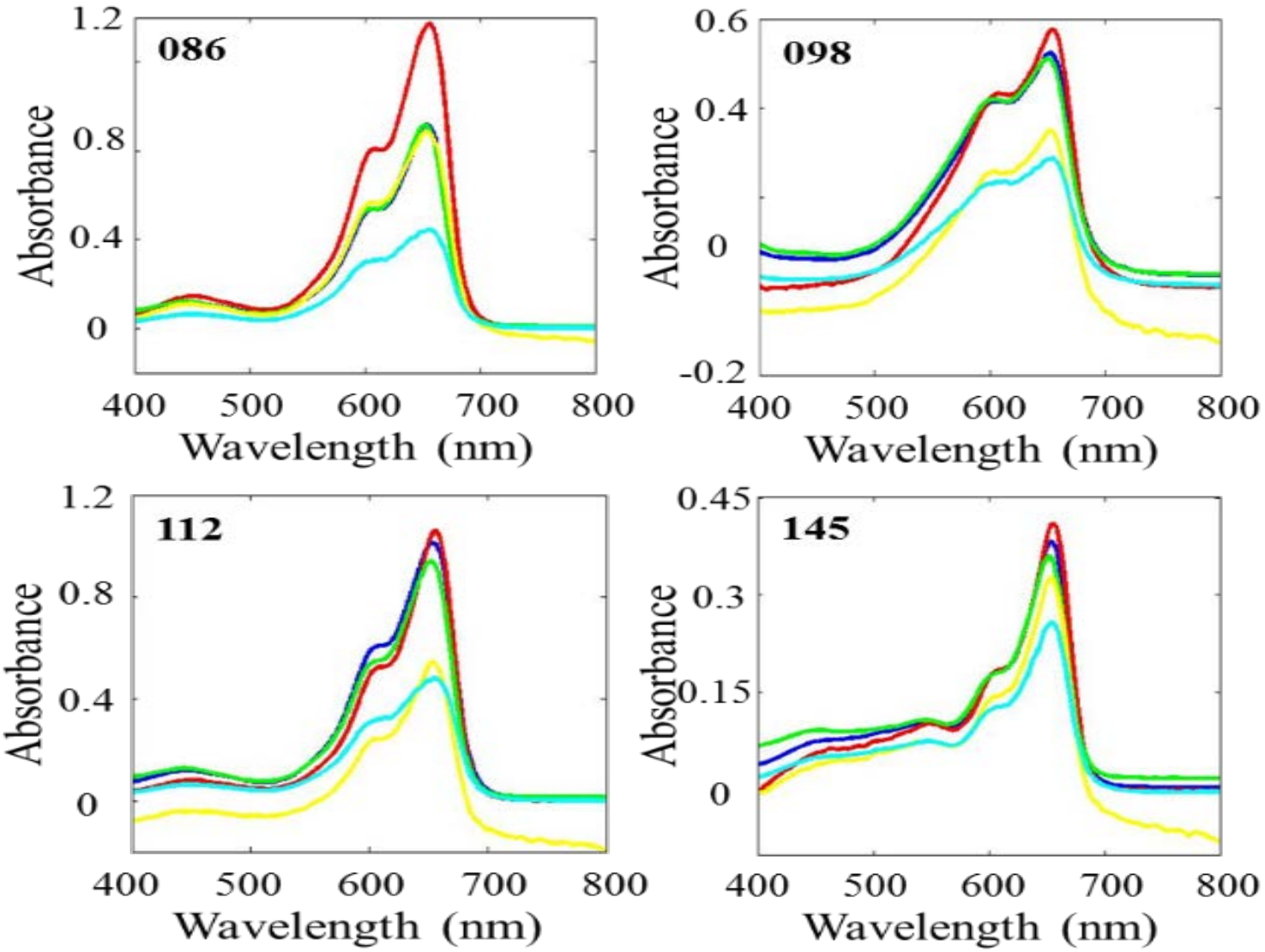
The twelve blue acrylic fibers were characterized by 10 replicate visible spectra taken in at five different laboratories (600 spectra) following the same method protocol using different models of MSP instrumentation (spanning over a decade in age)

Preprocessing is a must for various reasons

Preprocessing	Equation	Purpose
Autoscale	$X_{ij,auto} = \frac{X_{ij} - \overline{X_j}}{s_j}$	Places variables on equal footing to keep scale from dominating analysis
Baseline correction	$X_{ij,base} = X_{ij} - X_{i(min)}$	Corrects baseline offsets
First derivative	$X_{ij,first\ deriv} = \frac{X_{i,j+1} - X_{ij}}{\lambda_{i,j+1} - \lambda_{ij}}$	Corrects baseline effects
Normalization to unit area	$X_{ij,norm} = \frac{X_{ij}}{\sum_{j=1}^{n} \lvert X_{ij} \rvert}$	Removes scaling differences arising from variations in amount of sample as well as instrumental intensity variations caused by changes in fiber thickness
Standard normal variate (SNV)	$X_{ij,SNV} = \frac{X_{ij} - \overline{X_j}}{s_i}$	Removes changes in slope and variability caused by scattering

Definitions: X – Observation; λ - Wavelength; i - Row
s – Standard deviation; n – Number of Variables; j – Column

Blue acrylic fiber spectra from 5 labs

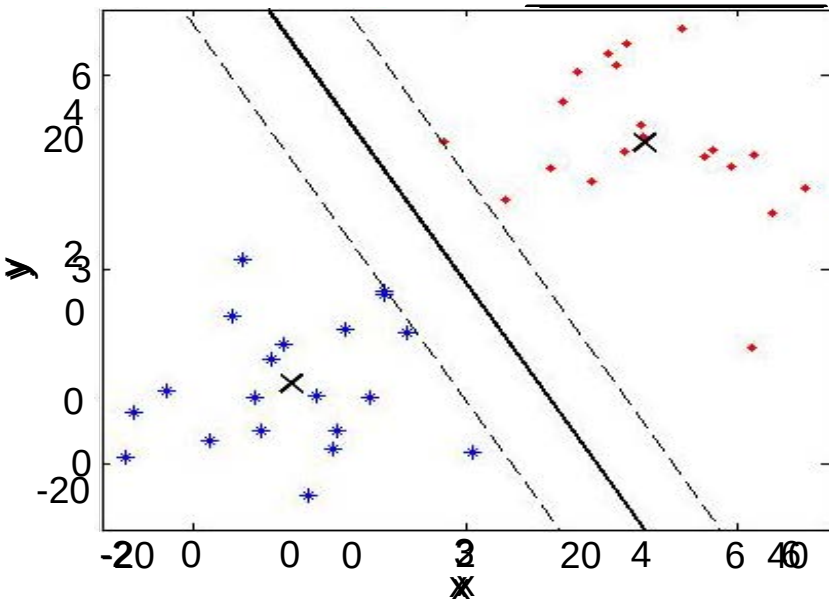


Preprocessing involved: truncation to 400-800 nm; Savitsky-Golay smoothing (21 point, 2nd order polynomial; weighted least squares baseline correction, and mean-centering

Multivariate classification

Linear discriminant analysis (LDA) generates linear decision boundaries which best separate the group means by assuming homogeneity of variances and covariances.

Quadratic discriminant analysis (QDA) is similar to LDA except a separate covariance matrix is estimated for each class. Unequal variance-covariance matrices keep quadratic terms of the multivariate Gaussian function from canceling, as in LDA, resulting in quadratic functions.



Support vector machine discriminant analysis (SVMDA) builds a maximum margin hyperplane in feature space by using kernel functions in a higher dimensional space.

Linear: $K_{(x_i,x_j)} = x_i \times x_j + 1$

Polynomial: $K_{(x_i,x_j)} = (x_i \times x_j + 1)^{d}$

Gaussian: $K_{(x_i,x_j)} = \exp \frac{-\|(x_i - x_j\|^2}{2\sigma^2}$

Between-laboratory comparisons: classification accuracies

METHOD				Percent of correctly classified spectra by sample											
METHOD	Lab.	Acc. (%)	SD	086	087	088	091	092	095	098	099	112	113	114	145
LDA	1	95.9	0.40	89.2	100	100	91.9	100	100	100	100	70	100	100	100
	2	97.3	0.53	80	100	100	100	100	100	100	100	91.2	100	100	100
	3	94.0	0.31	80	100	100	100	100	100	100	90	58.2	100	100	100
	4	92.0	0.83	88.8	100	100	61.6	75.6	91.4	100	100	100	100	100	100
	5	94.9	0.52	70	98.9	100	100	100	98.3	100	100	81.5	90.1	100	100
QDA	1	99.2	0.43	100	100	100	90	100	100	100	100	100	100	100	100
	2	95.3	0.51	87.6	100	100	100	100	100	100	100	55.8	100	100	100
	3	98.2	0.30	100	100	100	100	88.3	100	100	100	100	90	100	100
	4	91.8	0.89	100	95.4	100	72.8	81	79.8	100	100	81.9	90.1	100	100
	5	97.2	0.56	92.5	100	100	100	100	100	100	100	73.8	100	100	100
SVM-DA	1	99.2	0.41	100	100	100	90.9	100	100	100	100	100	100	98.9	100
	2	98.3	0.26	89.9	100	100	100	100	100	100	99.9	90.3	100	100	100
	3	98.3	0.46	99.3	100	100	100	100	100	90.9	90	99.3	100	100	100
	4	88.8	1.02	90.4	99.5	90	33.7	91.3	79.6	100	100	100	100	90.2	98.8
	5	94.9	1.00	70	98.9	100	100	100	98.3	100	100	81.5	90.1	100	100

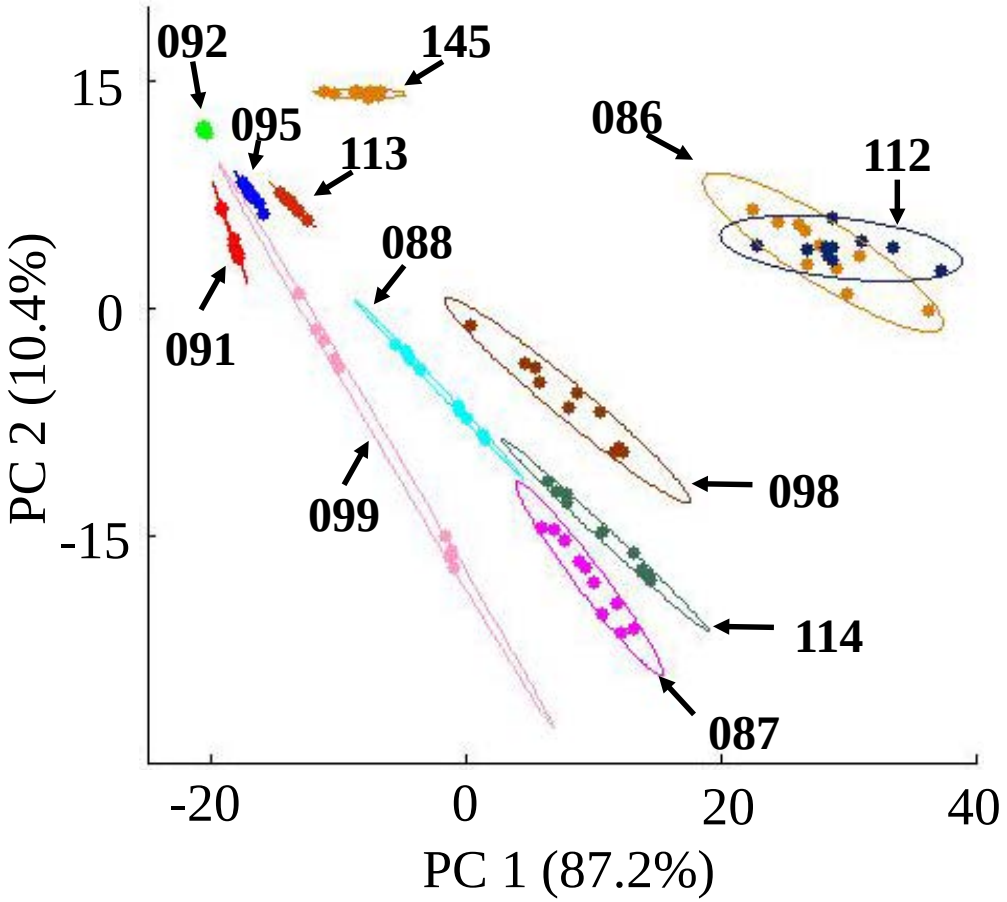
Predictive performances of the discriminant analysis models were determined by internal validation using the stratified 10-fold cross-validation was chosen method, because it is often a good compromise between bias and variance. In stratified 10-fold cross-validation, the data is partitioned into 10 nearly equal sized parts with approximately the same number of samples per class (*i.e.*, per fiber). The discriminant functions are then calculated using the information from all but one of these subsets, and the left-out portion is used to test the classifier. This process is repeated until each subset of samples has been used for testing.

Combined Lab Data Confusion Matrix using QDA

	ACTUAL CLASS											
PREDICTED CLASS	**86**	**87**	**88**	**91**	**92**	**95**	**98**	**99**	**112**	**113**	**114**	**145**
86	85.4								14.6			
87		**100**										
88		0.1	99.9									
91				96	0.3	3.4				0.3		
92				2.3	92.3	5.4						
95				2.2	2.8	95						
98							**100**					
99								**100**				
112	15.8								84.2			
113						0.2				99.8		
114											**100**	
145												**100**

Percentages of correctly classified spectra are in bold and those equaling zero are omitted.

PCA and DA (intra-laboratory)



PCA scores plot of 12 blue acrylic samples (10 replicates each) collected at laboratory *3.* Ellipses around groups of spectra represent distances that are statistically equal from the group mean with 95% confidence.

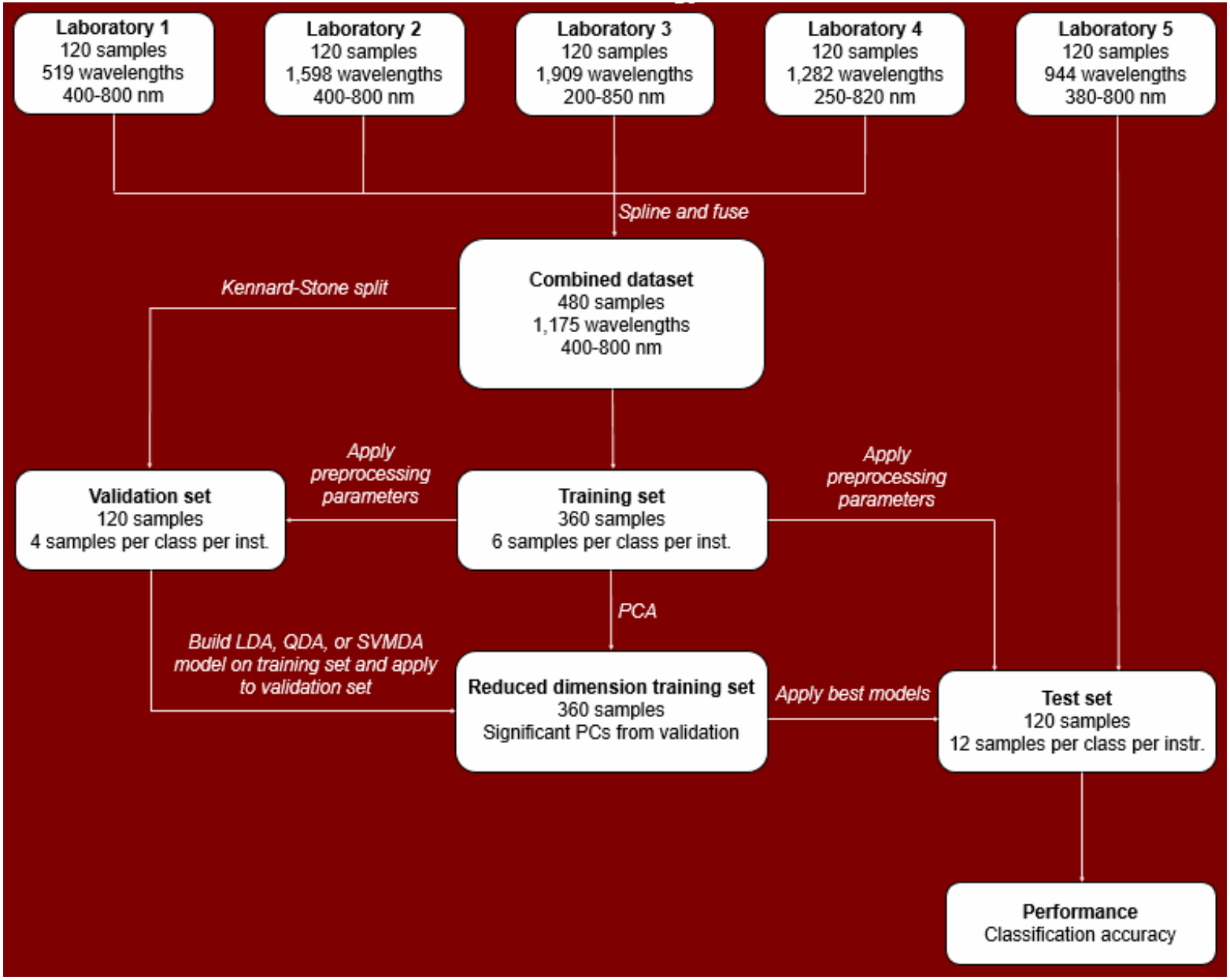
Technique	Accuracy (%)
QDA	96.3 ± 2.9
SVMDA	95.9 ± 4.3
LDA	94.8 ± 2.0

Fiber	QDA	SVMDA	LDA
112	82.3	94.1	80.4
086	96.0	89.9	81.6

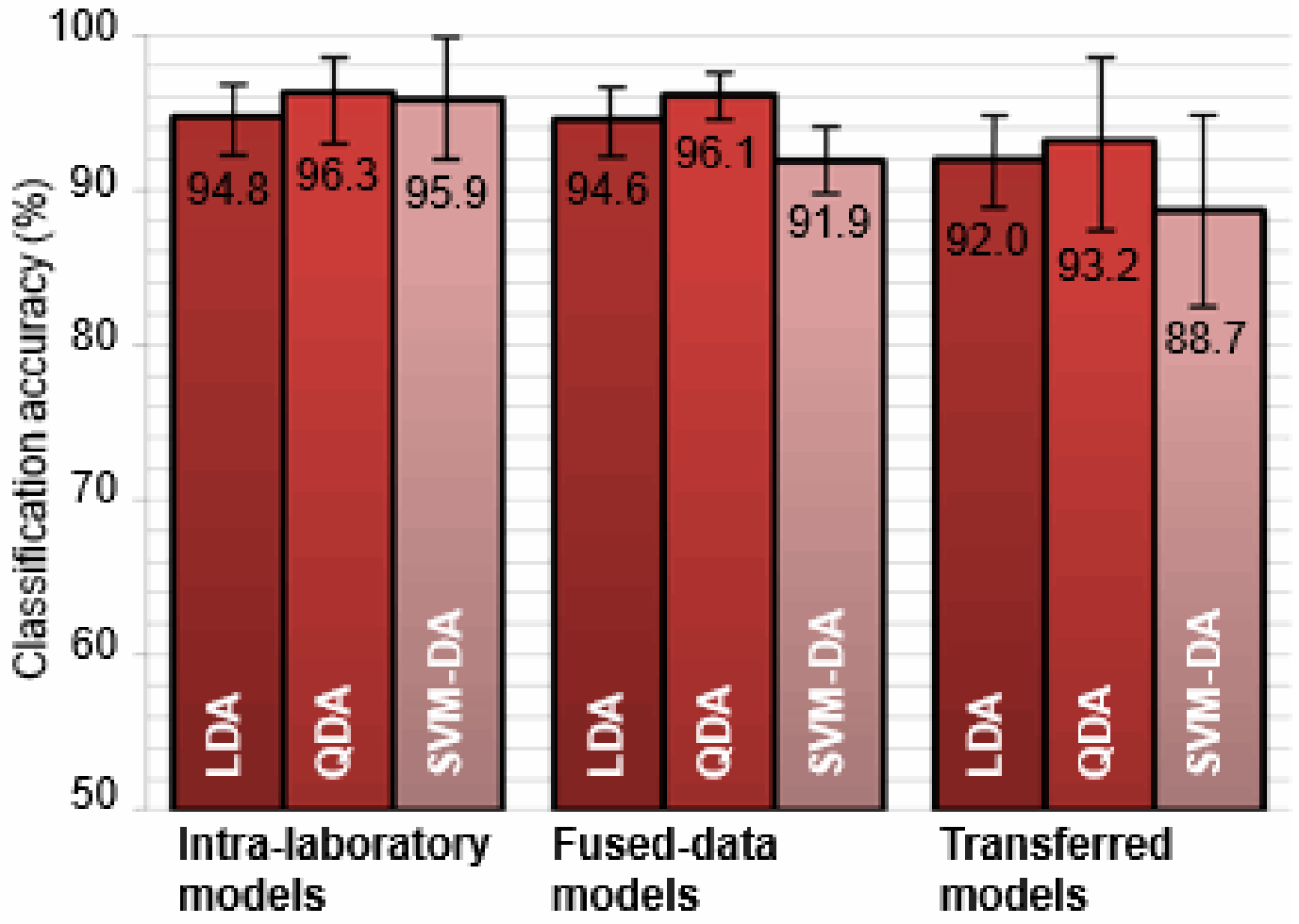
Top: *Average classification accuracy from five laboratories resulting from 100 iterations of stratified 10-fold cross-validation.*

Bottom: *Samples with highest numbers of misclassifications*

Data Fusion Methodology



Accuracy comparisons for multivariate classification



Summary

A prototype fiber database has been developed with the objectives of providing access to fiber characteristics and spectra for statistical comparisons. Understanding the significance of fiber evidence must be based on a thorough background of textile manufacturing practices and of the prevalence of fiber types in various regions of the world.

Mass production has resulted in the presence of textile fibers in numerous different and abundant commercial products. Further, when combinations of polymer types, colors, morphology, *etc*., are all taken into account, enormous numbers of different fibers exist. Establishing a collection of fibers that is representative of all possibilities is complicated by rapid changes in manufacturing practices and globalization of textile production: the population is a moving target of indeterminate size and evolving diversity.

As is often said about the problem of educating scientists to use statistics, the issues most discussed are often about which statistical approaches are 'best'. In fact, the majority of the benefit of statistics, when applied to understanding complex data, arises from the use of simple systematic comparisons with supporting descriptive statistics. It is our belief that if simple graphics do not show discrimination, no amount of statistical machinery will be convincing.