Statistics and Information Technology

Werner Stuetzle
Professor and Chair, Statistics
Adjunct Professor, Computer Science and Engineering
University of Washington

Prepared for NSF Workshop on *Statistics: Challenges and Opportunities for the 21st Century*.

Introduction

So far, have heard about

- Statistics and the Biological Sciences
- Statistics and the Geophysical and Environmental Sciences
- Statistics and the Social and Economic Sciences
- Engineering and Industrial Statistics

What other --- technology related --- research areas are there, in which statistical ideas are playing an important role?

Note: "Statistical ideas", not "Statisticians"

What other --- technology related --- research areas are there, in which statistical ideas are playing an important role?

(Incomplete and unordered list)

- Computer vision
- Medical imaging
- Speech recognition
- Computer graphics
- Genomics (“Biology is an Information Science”)
- Document organization and retrieval
- Analysis and monitoring of networks
- Customer modeling and transaction analysis
- Finance

(Comment on data mining and machine learning)

Outline of talk

- 3D photography: a case study in the application of ideas from Mathematics and Statistics to problems in computer graphics / computer vision
- Selected examples for use of statistical ideas in other technology related research areas
- Positioning Statistics to take advantage of opportunities

3D Photography: a case study in the application of Mathematics and Statistics to a problem in graphics / vision

3D Photography is an emerging technology aimed at

- capturing
- viewing
- manipulating

digital representations of shape and visual appearance of 3D objects.

3D photograph of fish statuette

3D Photography has potential for large impact because 3D photographs can be

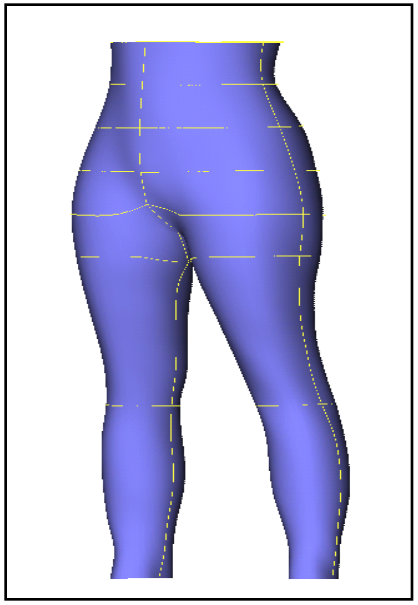
- stored and transmitted digitally
- viewed on CRTs
- used in computer simulations,
- manipulated and edited in software, and
- used as templates for making electronic or physical copies

Will now present some illustrations.

Modeling humans

- Anthropometry
- Ergonomics
- Sizing of garments
- Entertainment (avatars, animation)

Scan of lower body
(Textile and Clothing Technology Corp.)



Fitted template
(Dimension curves drawn in yellow)

Full body scan
(Cyberware)

Modeling artifacts

- Archival
- Quantitative analysis
- Virtual museums

Image courtesy of Marc Levoy and the Digital Michelangelo project

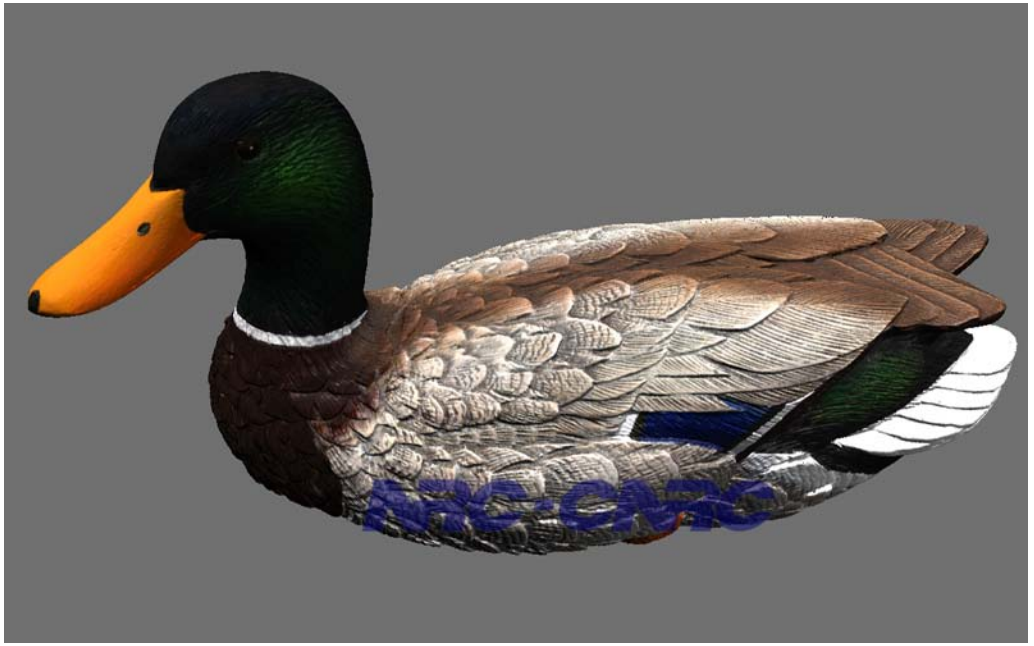
Left: Photo of David's head
Right: Rendition of digital model

(1mm spatial resolution, 4 million polygons)

Modeling artifacts

Images courtesy of Marc Rioux and the Canadian National Research Council

Nicaraguan stone figurine



Painted Mallard duck

Modeling architecture

- Virtual walk-throughs and walk-arounds
- Real estate advertising
- Trying virtual furniture

Left image: Paul Debevec, Camillo Taylor, Jitendra Malik (Berkely)

Right image: Chris Haley (Berkeley)

Model of Berkeley Campanile

Model of interior with artificial lighting

Modeling environments

- Virtual walk-throughs and walk arounds
- Urban planning

Two renditions of model of MIT campus
(Seth Teller, MIT)

What does 3D Photography have to do with Statistics?

We have data:

- 3D points acquired by range scanner

  and / or

- Digital images of scene

We want to build a model for scene geometry and “color”.

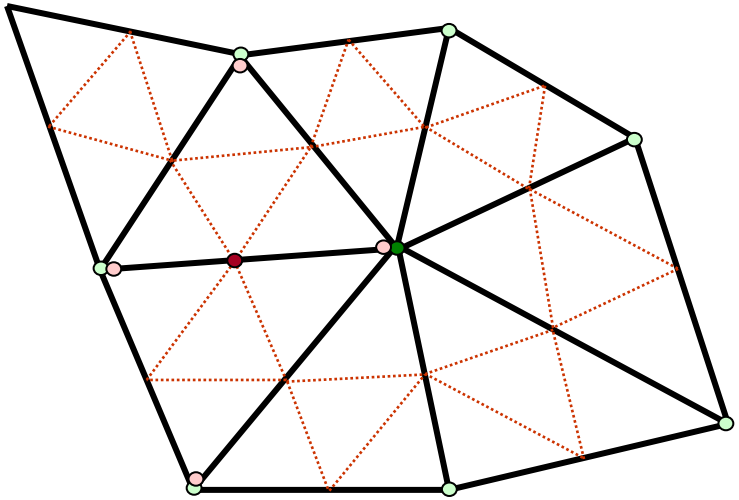
Admittedly:

- Data is not standard “cases by variables” or time series
- Noise is not a dominant aspect of the problem

This does not mean that Statistics has nothing to contribute.

Modeling shape (geometry)

Consider the simplest case where data is a collection of 3D points on object surface.



Model shape by **subdivision surface**

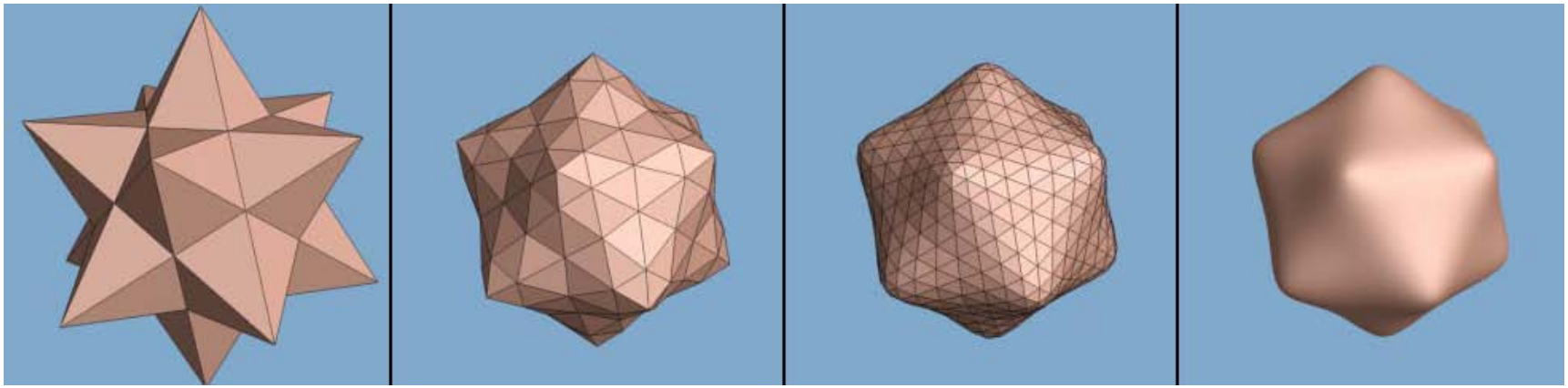
Defined by limiting process, starting with control mesh (bottom left)

Split each face into four (right)

Compute positions of new **edge vertices** as weighted means of **corner vertices**

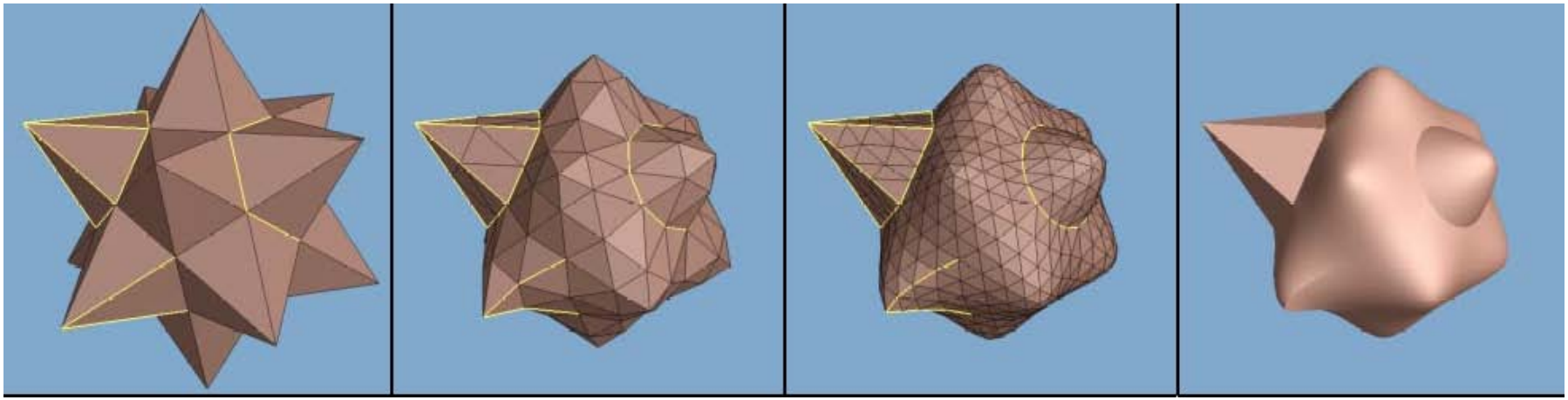
Compute new positions of **corner vertices** as weighted means of their **neighbors**

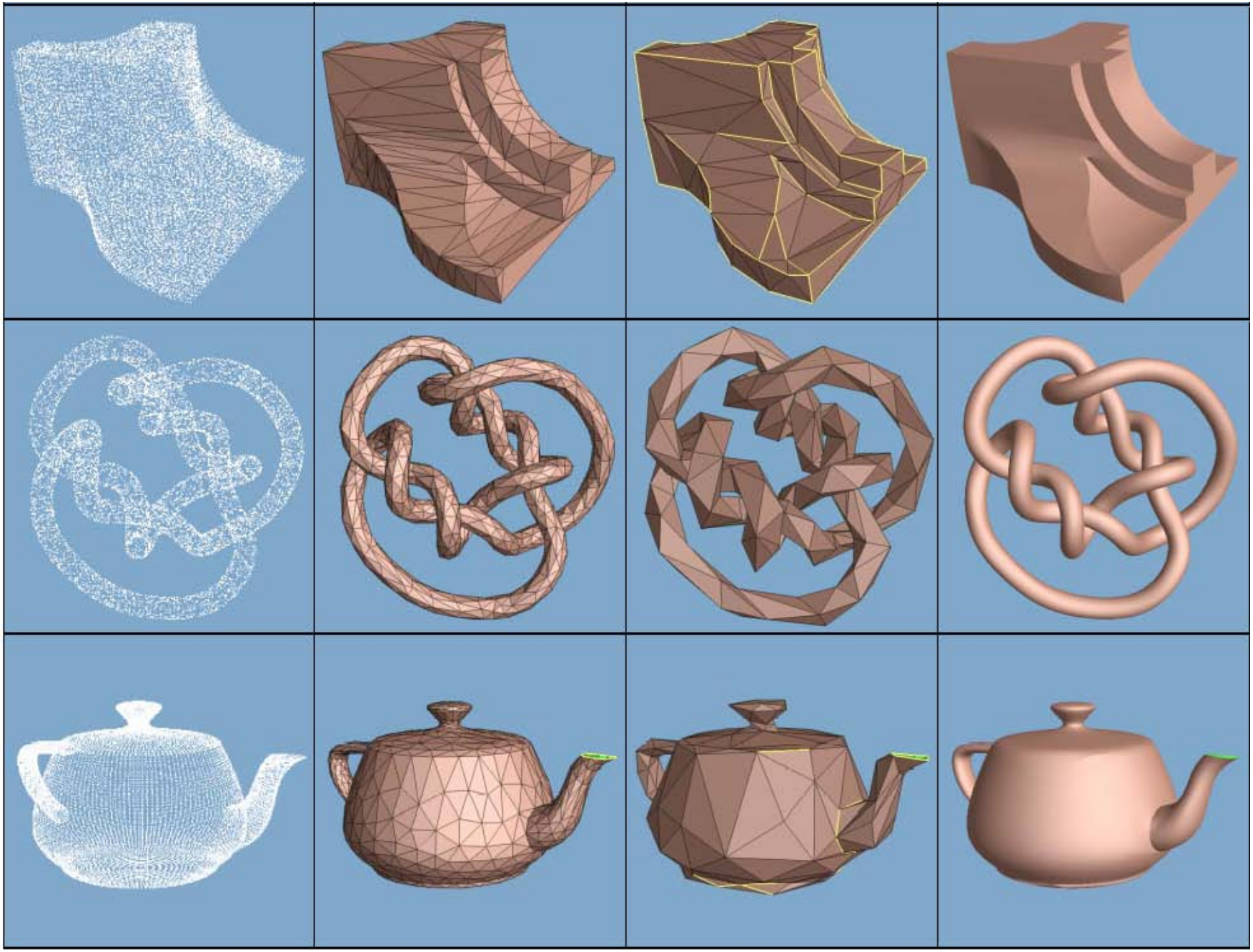
Repeat the process



Remarks

- Subdivision surfaces are a generalization of splines.
- Averaging rules can be modified to allow for sharp edges, creases, and corners (below).
- Shape of limit surface is controlled by positions of control vertices. They are the shape parameters.
- We fit subdivision surfaces to data by solving a penalized nonlinear least squares problem.





Modeling “color”

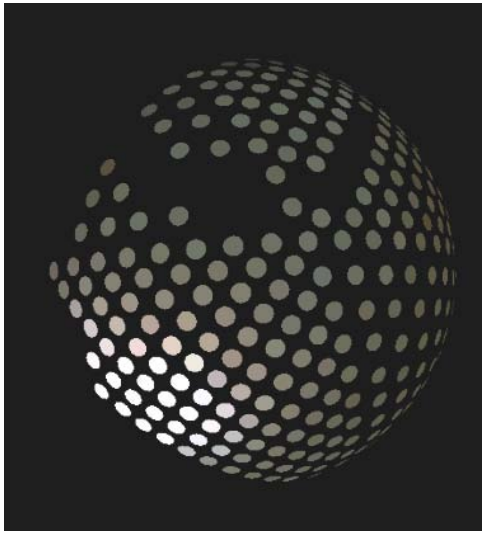
Non-trivial because

- Real objects don’t look the same from all directions (specularity, anisotropy)
- Ignoring these effects makes everything look like plastic

Appearance under fixed lighting is captured by “surface light field” (SLF)

SLF assigns RGB value to each surface point and each viewing direction

SLF is a vector-valued function over (surface x 2-sphere).

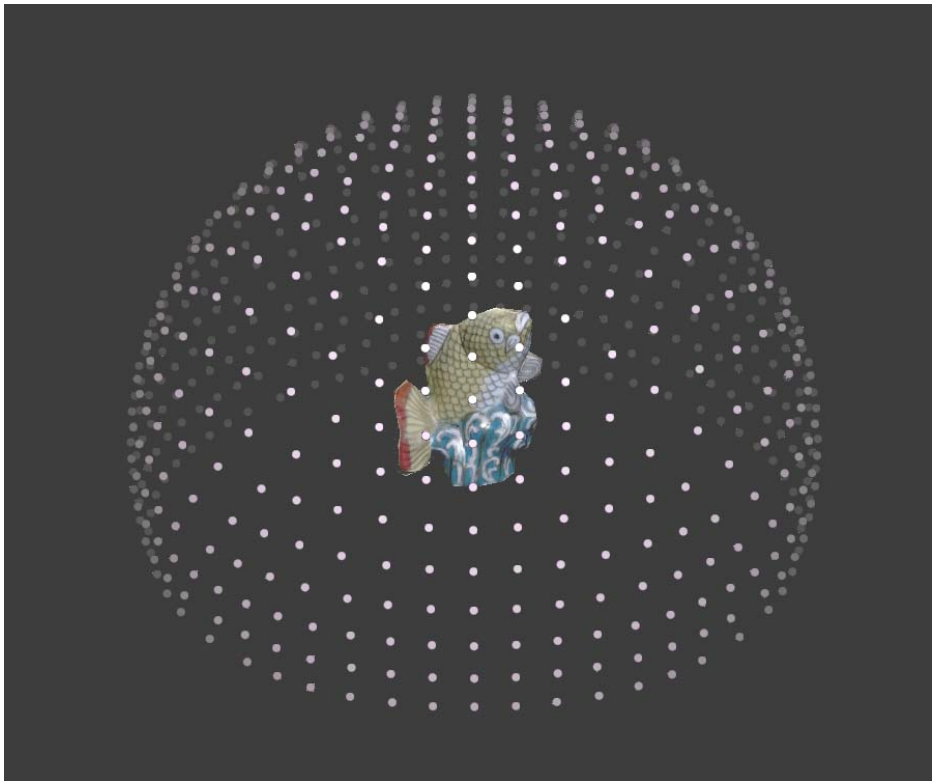


Data lumispere: observed direction - color pairs for single surface point

Raw data:

- 3D points acquired with laser scanner
- Approximately 700 digital images

Camera positions



One of ~ 700 images

After modeling geometry and pre-processing images we have:

- Triangular mesh with 1,000's of vertices
- Collection of (direction, RGB value) pairs for each vertex

Issues

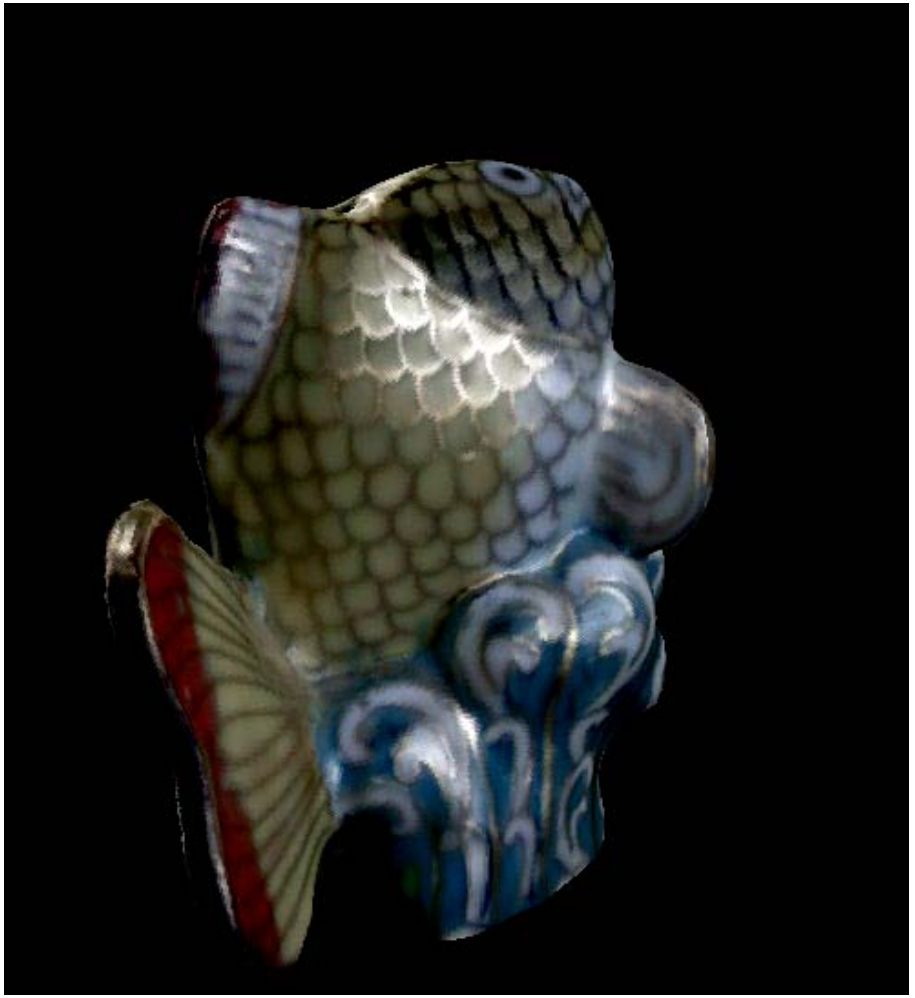
- Interpolation / smoothing on general manifolds
- Compression: uncompressed SLF for fish is about 170 MB
- Real time rendering non-trivial

Mesh generated from fish scans

Payoff

Modeling and rendering SLF adds a lot of realism

Incorporating some simple ideas from optics allows for "extrapolation" →



Contributions of Statistics

1. General approach

- We have a set of data - surface points produced by the sensor.
- We want to “fit a parametric model” to these data, in our case a 2D manifold.
- Parameters of model control shape of the manifold.
- We define a goodness-of-fit measure quantifying how well model approximates data.
- We then find the best parameter setting using numerical optimization.

Contributions of Statistics

2. More specific methods and theory

- Linear and nonlinear least squares
- Regularization (ridge regression, penalized PCA, spline smoothing)
- Principal surfaces
- Clustering, vector quantization

Selected examples for use of statistical ideas in other technology related research areas

Medical imaging: Reconstructing the human heart from ultrasound data

- Data that are sparse and noisy
- Build detailed model of imaging process
- Fit surface model instead of working “slice by slice” and then gluing slices together (uses spatial contiguity)
- Incorporate prior information about shape: deformable template + Bayesian analysis using empirical prior distribution for shape parameters

Computer vision: Face recognition in images

- To recognize a face independent of pose probably need 3D model of head
- Want to build model from one or a few photographs
- Prior information on head shape has to be built into the modeling process

Computer graphics: Posing humans

A non-standard prediction problem

Trainig data: High resolution 3D scans of
- few subjects in many poses, and
- many subjects in few poses

Goal: Generate a rule to predict the shape of a new (test) subject in a target pose from the scan in a different pose.

Customer modeling and transaction analysis: Recommender systems

A non-standard prediction problem

Given a sparsely populated matrix

R[i, j] = rating of product j by consumer i,

and attributes for products and consumers, predict the missing elements of R.

Document organization and retrieval: Topic detection and tracking

Given a (dynamic) collection of text documents, partition the collection into topics and track the appearance and disappearance of topics over time.

Positioning Statistics to take advantage of opportunities

Characteristics exhibited by the examples

- Non-standard data --- not cases x variables or time series
  - Images
  - Text
  - Streaming data
- Emphasis on description and prediction, not on inference
- Algorithm development, implementation, and testing is large part of research

To take advantage of opportunities, Statistics has to change its self-image:

- The goal of Statistics is to develop tools for analyzing data.
  Statistics is an Engineering discipline.
  Methodology is its core.

- Development of a tool has to start with a problem that the tool is supposed to address. This indicates the need for collaborative research which can
  - stimulate research in statistical methodology and theory;
  - benefit the application area;
  - create a constituency for Statistics.

- Tools are implemented and assessed on the computer.
  In methodology research Computer Science plays a role that is comparable to the role of Mathematics.

- Assesssment is an integral part of tool development. Tools can be assessed using mathematical analysis, or empirically (Princeton Robustness Study)

- The traditional focus of Statistics on one particular aspect of data analysis, namely dealing with variability, is unnecessarily restrictive as well as unfortunate.

To take advantage of opportunities, Statistics has to change the way in which it recruits and trains students:

- Ph.D. programs currently focus on training and research in statistical theory and applications.
- Therefore, we select for students with background and interest in Mathematics.
- We should also prepare students for methodology research.
- We should select students with background and interest in computing and in collaborative research.

Other disciplines are seizing the opportunity and taking the butter off our bread.

- Either adopt “policy of the small flock” or
- Meet the challenge

Modeling color

- We are given a collection of data lumispheres (direction – color pairs)
- We find a low dimensional subspace of piecewise linear functions on the sphere.
- We aproximaten data lumispheres by their projections on the subspace -> imputation, compression.