Unsupervised learning, Clustering

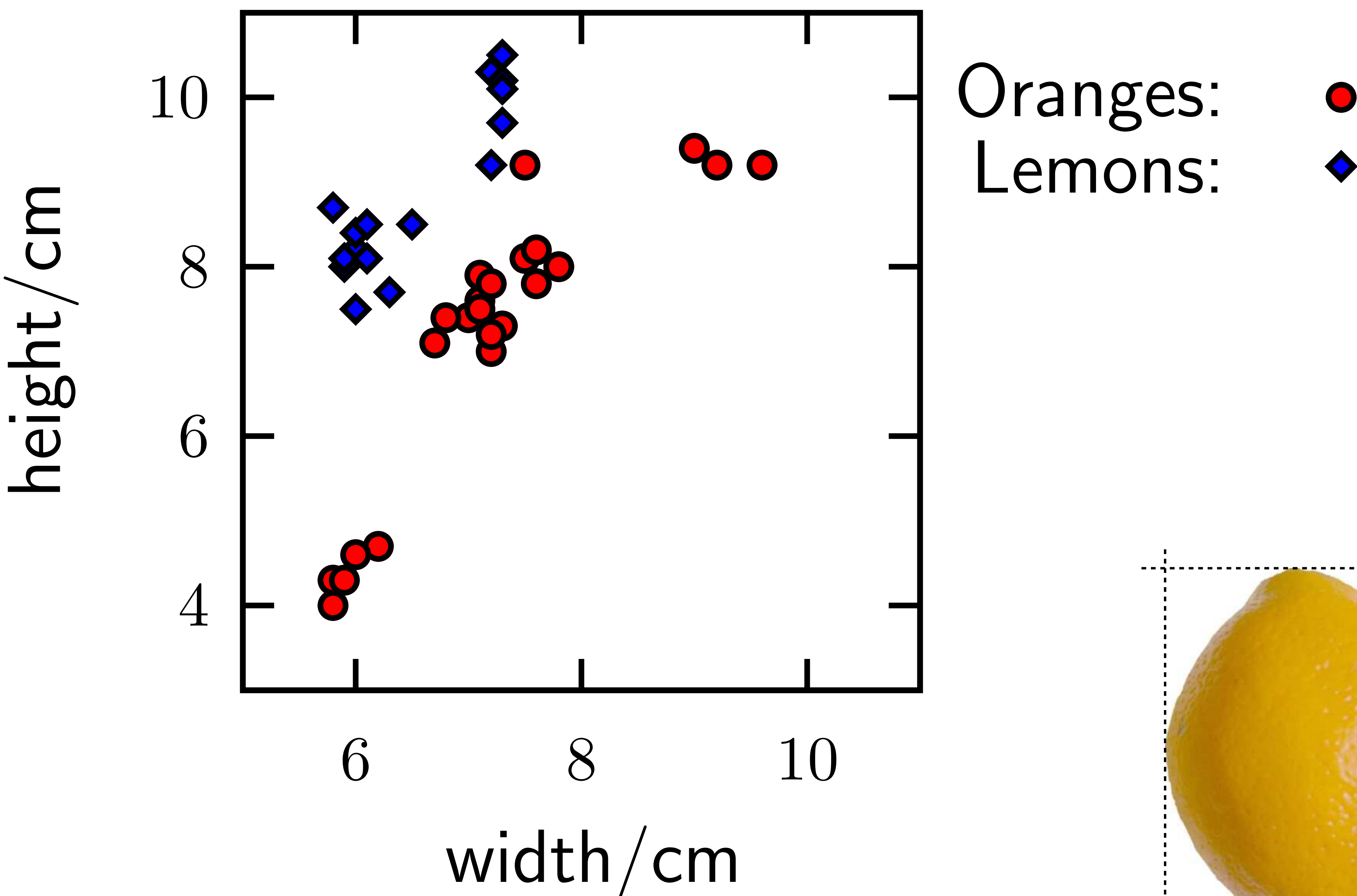
“Human brains are good at finding regularities in data.

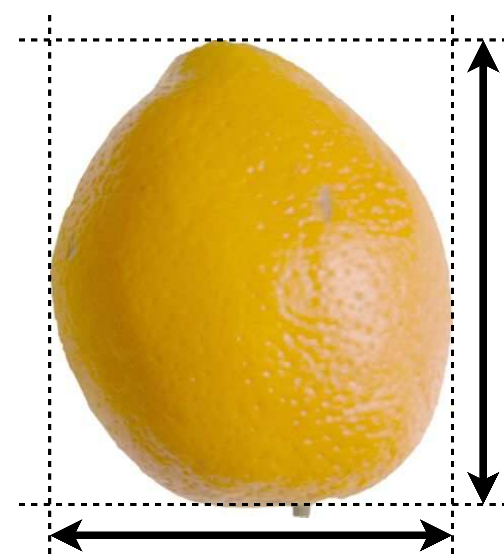
One way of expressing regularity is to put a set of objects into groups that are similar to each other.

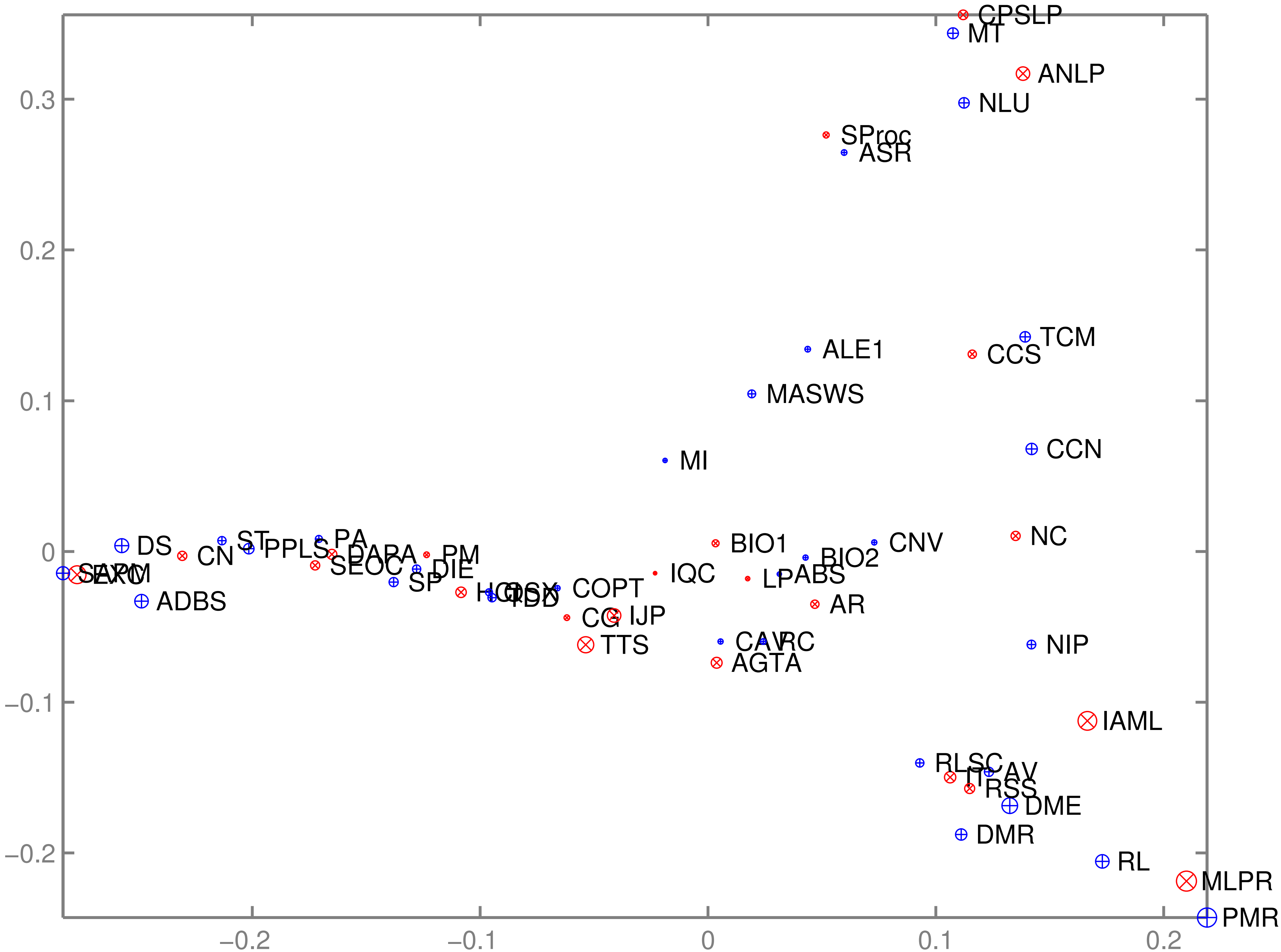
For example, biologists have found that most objects in the natural world fall into one of two categories: things that are brown and run away, and things that are green and don’t run away. The first group they call animals, and the second, plants.”

— David MacKay, ITILA textbook p284

Oranges and Lemons data







CPSLP
MT
ANLP
NLU
SProc
ASR
TCM
ALE1
CCS
MASWS
CCN
MI
NC
CNV
BIO1
DS
ST
PA
CN
PPLS
DAPA
PM
BIO2
SEOC
DIE
IQC
SP
COPT
ADBS
AR
IJP
CG
TTS
AGTA
NIP
IAML
RLSC
AV
RSS
DME
DMR
RL
MLPR
PMR
0.3
0.2
0.1
0
−0.1
−0.2
−0.2
−0.1
0
0.1
0.2

Stanley

Stanford Racing Team; DARPA 2005 challenge

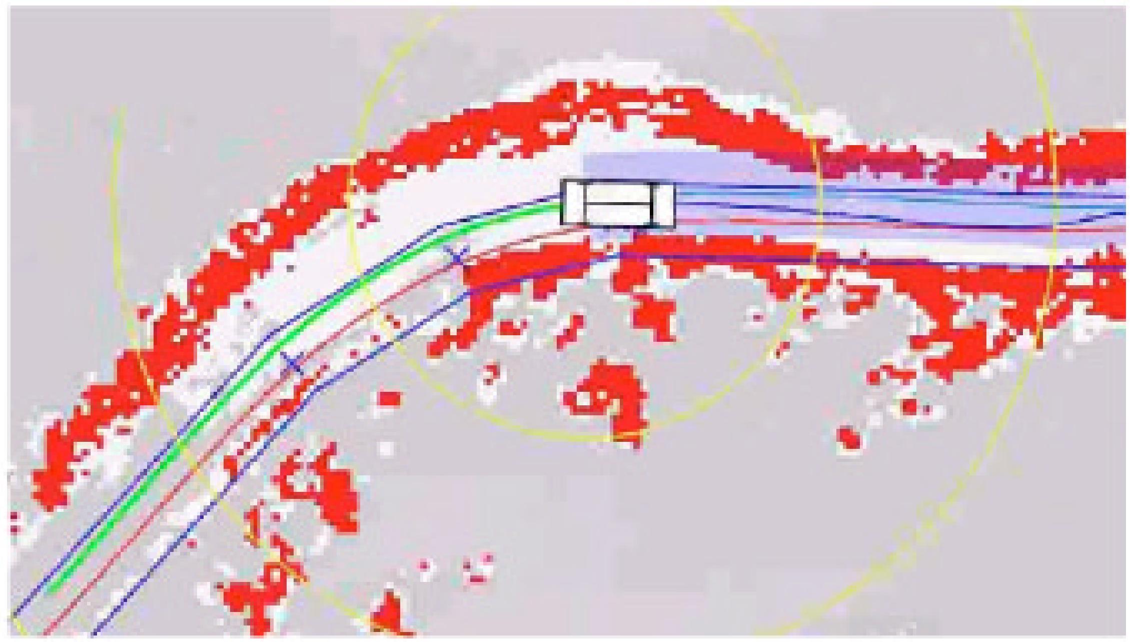
`http://robots.stanford.edu/talks/stanley/`

How to stay on a road?

Perception and intelligence

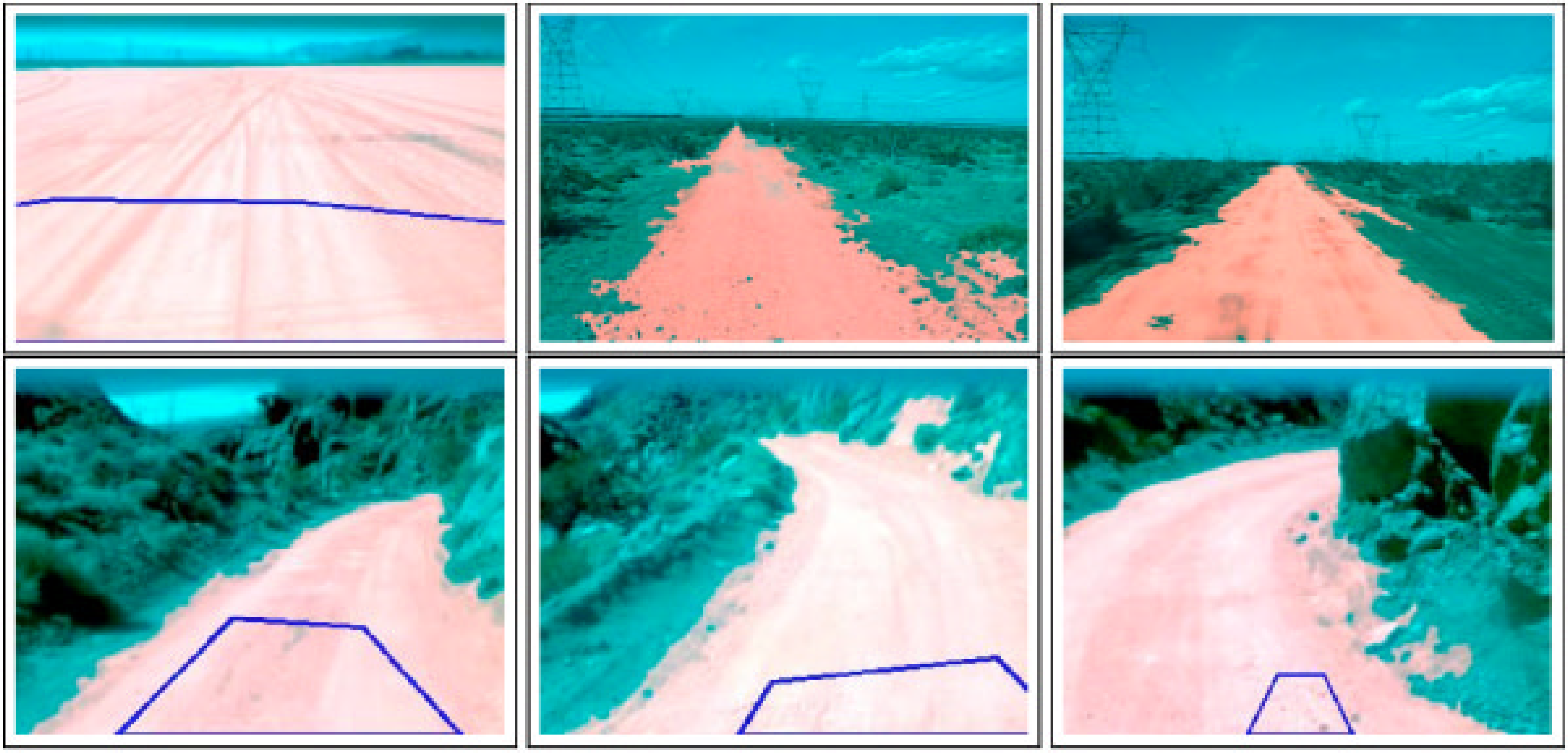
(a) **Beer Bottle Pass**

(b) **Map and GPS corridor**



It would look pretty stupid to run off the road,
just because the trip planner said so.

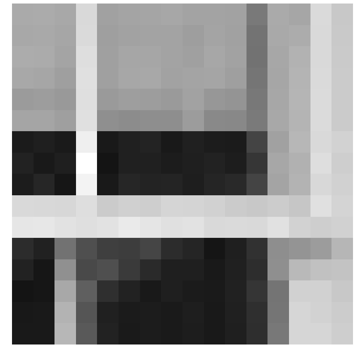
Clustering to stay on the road



Stanley used a Gaussian mixture model.
The cluster just in front is road (unless we already failed).

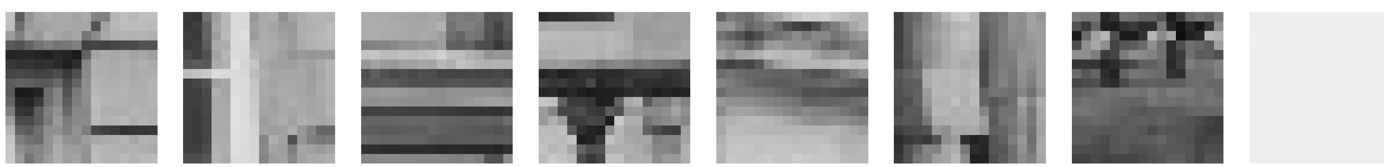
Example: Image denoising

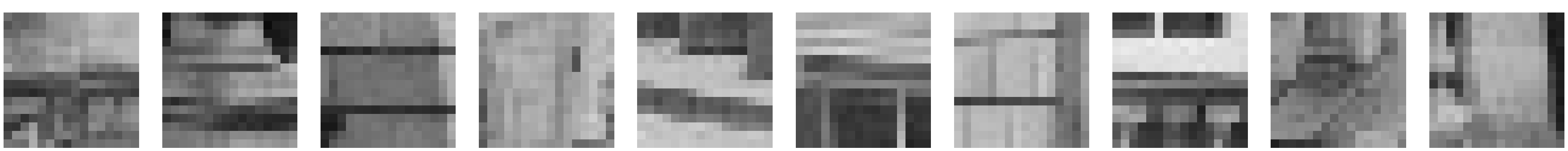
$\mathbf{y}$: 

$\mathbf{x}$: 

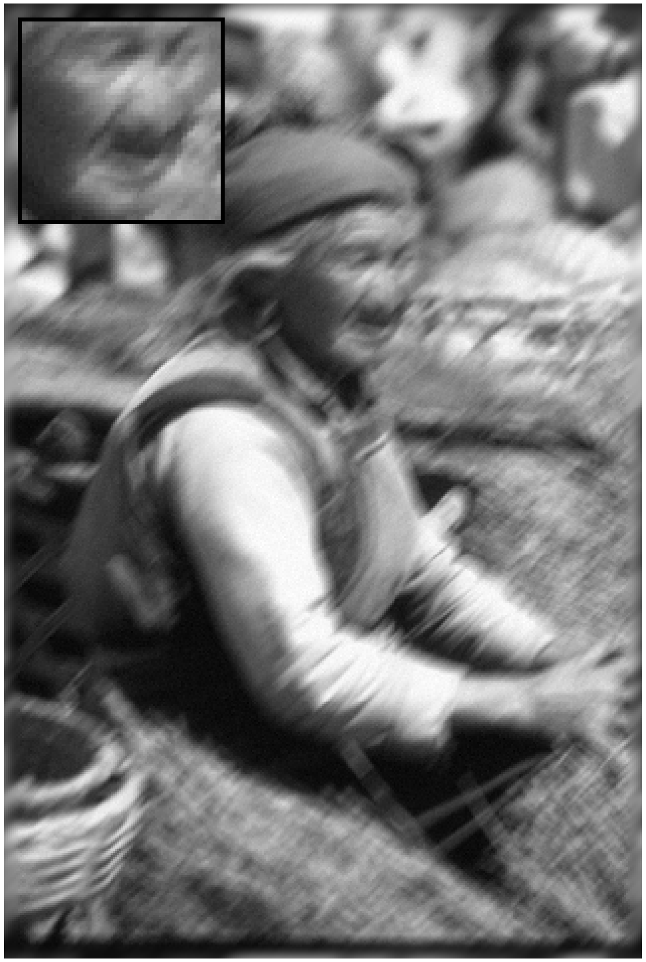
$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$$

Likelihood: e.g. $\mathcal{N}(\mathbf{y};\, \mathbf{x},\, \sigma^2 I)$

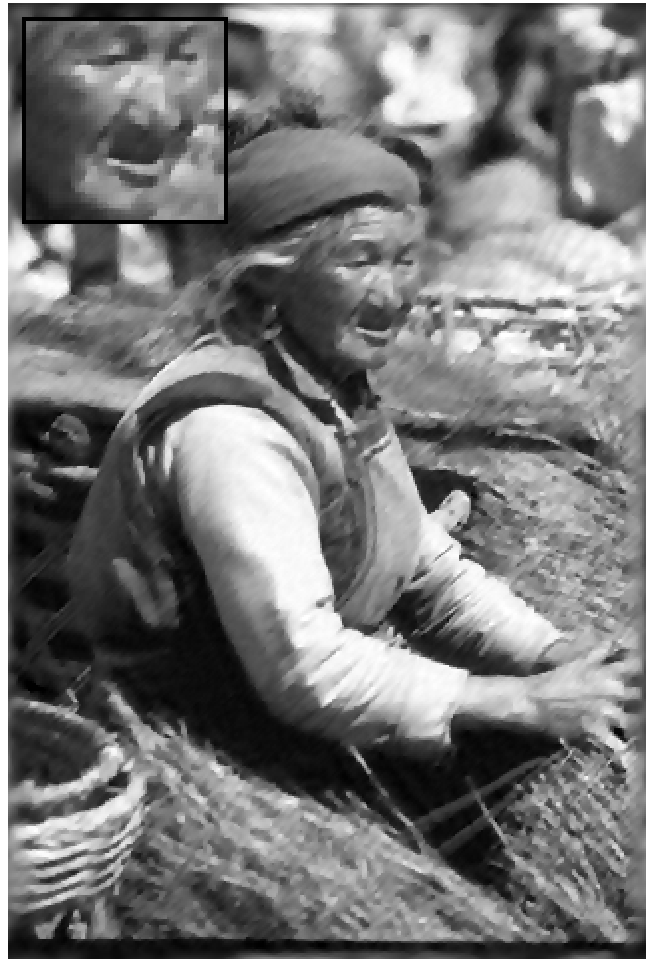
Prior samples: 

 …

Zoran and Weiss, ICCV 2011



(a) Blurred

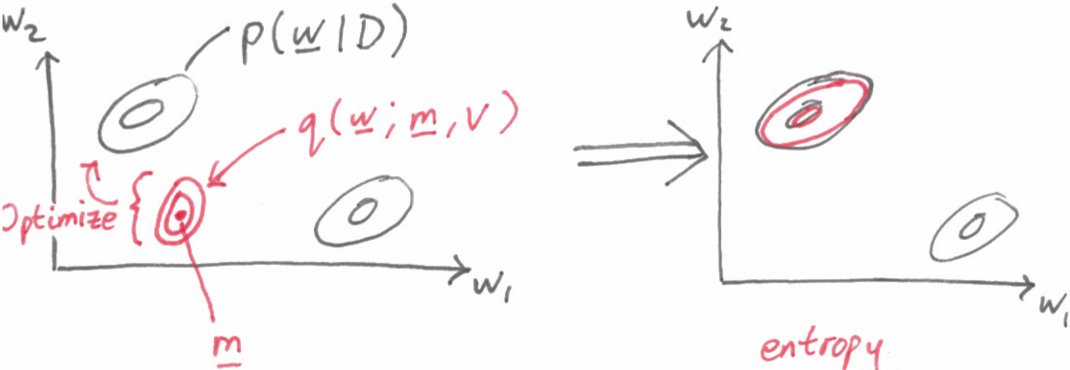


(b) Krishnan et al.

(c) EPLL GMM

$p(\mathbf{x})$ = Mixture of Gaussians fitted to patches

Variational Inference



Cost function:

$$J = \overbrace{\underbrace{-\mathbb{E}_q[\log p(D|\underline{w})]}_{\mathbb{E}[\text{neg. log likelihood}]} - \mathbb{E}_q[\log p(\underline{w})]}^{\mathbb{E}[\text{'Energy'}]} + \overbrace{\mathbb{E}_q[\log q(\underline{w})]}^{-H[q]\ \text{entropy}}$$

$$D_{KL}(q(\underline{w}) \| p(\underline{w}))$$

Marginal Likelihood bound

$$\log p(D) \geqslant -J$$

Minimize J wrt $\underbrace{(\underline{m}, V)}_{\text{Approx. posterior well}}$ and $\underbrace{\sigma_w^2, \sigma_y^2, \ldots}_{\text{Find good model}}$

Minimize J

Like SGD, but need some tricks

Trick #1 Unconstrained Optimization

If we do SGD on σ_y^2 might get -ve values

Optimize $\log \sigma_y^2$ instead

V positive definite, symmetric

$V = LL^T$, L lower triangular
Diagonal is +ve.

We create another matrix

$$\tilde{L}_{ij} = \begin{cases} L_{ij} & i \neq j \\ \log L_{ii} & i = j \end{cases}$$

Optimizer has $\tilde{L} \xrightarrow[\text{exp diag}]{} L \rightarrow V = LL^T \rightarrow \ldots \text{cost}$

SGD on $\tilde{L}$ $\leftarrow$ backprop $\leftarrow$

Evaluating the cost

$D_{KL}(q \| p)$, or "Entropy terms"

We can evaluate these.

Likelihood term:

$$\mathbb{E}_q[\log p(D|\underline{w})]$$

$$= \mathbb{E}_q\left[\sum_{n=1}^{N} \log p(y^{(n)}|\underline{x}^{(n)}, \underline{w})\right]$$

For logistic regression: → 1D integral
do numerically.

Stochastic estimate: Trick #2 Reparameterization trick

$$\mathbb{E}_{N(\underline{w};\,\underline{m},V)}[f(\underline{w})]$$

$$= \mathbb{E}_{N(\underline{\nu};\,0,\mathbb{I})}[f(\underline{m} + L\underline{\nu})]$$

Sample $\underline{w}$, by $\underline{\nu} \sim N(0, \mathbb{I})$

$\underline{w} = \underline{m} + L\underline{\nu}$

Monte Carlo estimate

$$\approx \frac{1}{S} \sum_{s=1}^{S} f(\underline{m} + L\underline{\nu}^{(s)})$$

$$\underline{\nu}^{(s)} \sim \mathcal{N}(0, \mathbb{I})$$

$S = 1$

$$\approx f(\underline{m} + L\underline{\nu}), \quad \underline{\nu} \sim \mathcal{N}(0, \mathbb{I})$$

$$\nabla_{\underline{m}} \; \mathbb{E}_{\mathcal{N}(\underline{w}; \underline{m}, V)}[f(\underline{w})]$$

$$\approx \nabla_{\underline{m}} f(\underline{m} + L\underline{\nu})$$

$$\nabla_L \; \mathbb{E}_{\mathcal{N}(\underline{w}; \underline{m}, V)}[f(\underline{w})]$$

$$\approx \nabla_L f(\underline{m} + L\underline{\nu})$$

$$\nabla_{\underline{w}} f(\underline{w})\Big|_{\underline{w} = \underline{m} + L\underline{\nu}} \; \underline{\nu}^\top$$

Need derivatives of log likelihood as usual.

(Also find $\nabla_{\tilde{L}}$ for SGD from ∇_L)

Mixtures of Gaussians