

Boosting and Additive Models

Boosting

- 1988 Kearns and Valiant: “Can **weak learners** be combined to create a **strong learner**?”

Weak learner definition (informal):

An algorithm \mathcal{A} is a *weak learner* for a hypothesis class \mathcal{H} that maps \mathcal{X} to $\{-1, 1\}$ if for all input distributions over \mathcal{X} and $h \in \mathcal{H}$, we have that \mathcal{A} correctly classifies h with error at most $1/2 - \gamma$

- 1990 Robert Schapire: “Yup!”
- 1995 Schapire and Freund: “Practical for 0/1 loss” AdaBoost
- 2001 Friedman: “Practical for arbitrary losses”
- 2014 Tianqi Chen: “Scale it up!” XGBoost

Additive models

- **Given:** $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$
- **Generate random functions:** $\phi_t : \mathbb{R}^d \rightarrow \mathbb{R}$ $t = 1, \dots, p$
- **Learn some weights:** $\hat{w} = \arg \min_w \sum_{i=1}^n \text{Loss} \left(y_i, \sum_{t=1}^p w_t \phi_t(x_i) \right)$
- **Classify new data:** $f(x) = \text{sign} \left(\sum_{t=1}^p \hat{w}_t \phi_t(x) \right)$

Additive models

- **Given:** $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$
- **Generate random functions:** $\phi_t : \mathbb{R}^d \rightarrow \mathbb{R}$ $t = 1, \dots, p$
- **Learn some weights:** $\hat{w} = \arg \min_w \sum_{i=1}^n \text{Loss} \left(y_i, \sum_{t=1}^p w_t \phi_t(x_i) \right)$
- **Classify new data:** $f(x) = \text{sign} \left(\sum_{t=1}^p \hat{w}_t \phi_t(x) \right)$

An interpretation:

Each $\phi_t(x)$ is a classification rule that we are assigning some weight \hat{w}_t

$$\hat{w}, \hat{\phi}_1, \dots, \hat{\phi}_p = \arg \min_{w, \phi_1, \dots, \phi_p} \sum_{i=1}^n \text{Loss} \left(y_i, \sum_{t=1}^p w_t \phi_t(x_i) \right)$$

is in general computationally hard

Forward Stagewise Additive models

$b(x, \gamma)$ is a function with parameters γ

Examples: $b(x, \gamma) = \frac{1}{1 + e^{-\gamma^T x}}$

Algorithm 10.2 *Forward Stagewise Additive Modeling.*

1. Initialize $f_0(x) = 0$.

2. For $m = 1$ to M :

(a) Compute

$$(\beta_m, \gamma_m) = \arg \min_{\beta, \gamma} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)).$$

(b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.

$$b(x, \gamma) = \gamma_1 \mathbf{1}\{x_3 \leq \gamma_2\}$$

Idea: greedily add one function at a time

Forward Stagewise Additive models

$b(x, \gamma)$ is a function with parameters γ

Examples: $b(x, \gamma) = \frac{1}{1 + e^{-\gamma^T x}}$

Algorithm 10.2 *Forward Stagewise Additive Modeling.*

1. Initialize $f_0(x) = 0$.

2. For $m = 1$ to M :

(a) Compute

$$(\beta_m, \gamma_m) = \arg \min_{\beta, \gamma} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)). = \exp(-y f_{m-1}(x)) \exp(-\beta y b(x_i, \gamma))$$

(b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.

Idea: greedily add one function at a time

AdaBoost: $b(x, \gamma)$: classifiers to $\{-1, 1\}$

$$L(y, f(x)) = \exp(-y f(x))$$

Forward Stagewise Additive models

$b(x, \gamma)$ is a function with parameters γ

Examples: $b(x, \gamma) = \frac{1}{1 + e^{-\gamma^T x}}$

Algorithm 10.2 *Forward Stagewise Additive Modeling.*

1. Initialize $f_0(x) = 0$.

2. For $m = 1$ to M :

(a) Compute

$$(\beta_m, \gamma_m) = \arg \min_{\beta, \gamma} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)).$$

(b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.

$$b(x, \gamma) = \gamma_1 \mathbf{1}\{x_3 \leq \gamma_2\}$$

Idea: greedily add one function at a time

Boosted Regression Trees:

$$L(y, f(x)) = (y - f(x))^2$$

$b(x, \gamma)$: regression trees

Forward Stagewise Additive models

$b(x, \gamma)$ is a function with parameters γ Examples: $b(x, \gamma) = \frac{1}{1 + e^{-\gamma^T x}}$
 $b(x, \gamma) = \gamma_1 \mathbf{1}\{x_3 \leq \gamma_2\}$

Algorithm 10.2 *Forward Stagewise Additive Modeling.*

1. Initialize $f_0(x) = 0$.
2. For $m = 1$ to M :
 - (a) Compute

$$(\beta_m, \gamma_m) = \arg \min_{\beta, \gamma} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)).$$

- (b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.
-

Idea: greedily add one function at a time

Boosted Regression Trees: $L(y, f(x)) = (y - f(x))^2$

$$\begin{aligned} L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)) &= (y_i - f_{m-1}(x_i) - \beta b(x_i; \gamma))^2 \\ &= (r_{im} - \beta b(x_i; \gamma))^2, \quad r_{im} = y_i - f_{m-1}(x_i) \end{aligned}$$

Efficient: No harder than learning regression trees!

Additive models

- Boosting is popular at parties: Invented by theorists, heavily adopted by practitioners.
- Computationally efficient with “weak” learners. But can also use trees! Boosting can scale.
- Gradient boosting generalization with good software packages (e.g., *XGBoost*). Effective on Kaggle

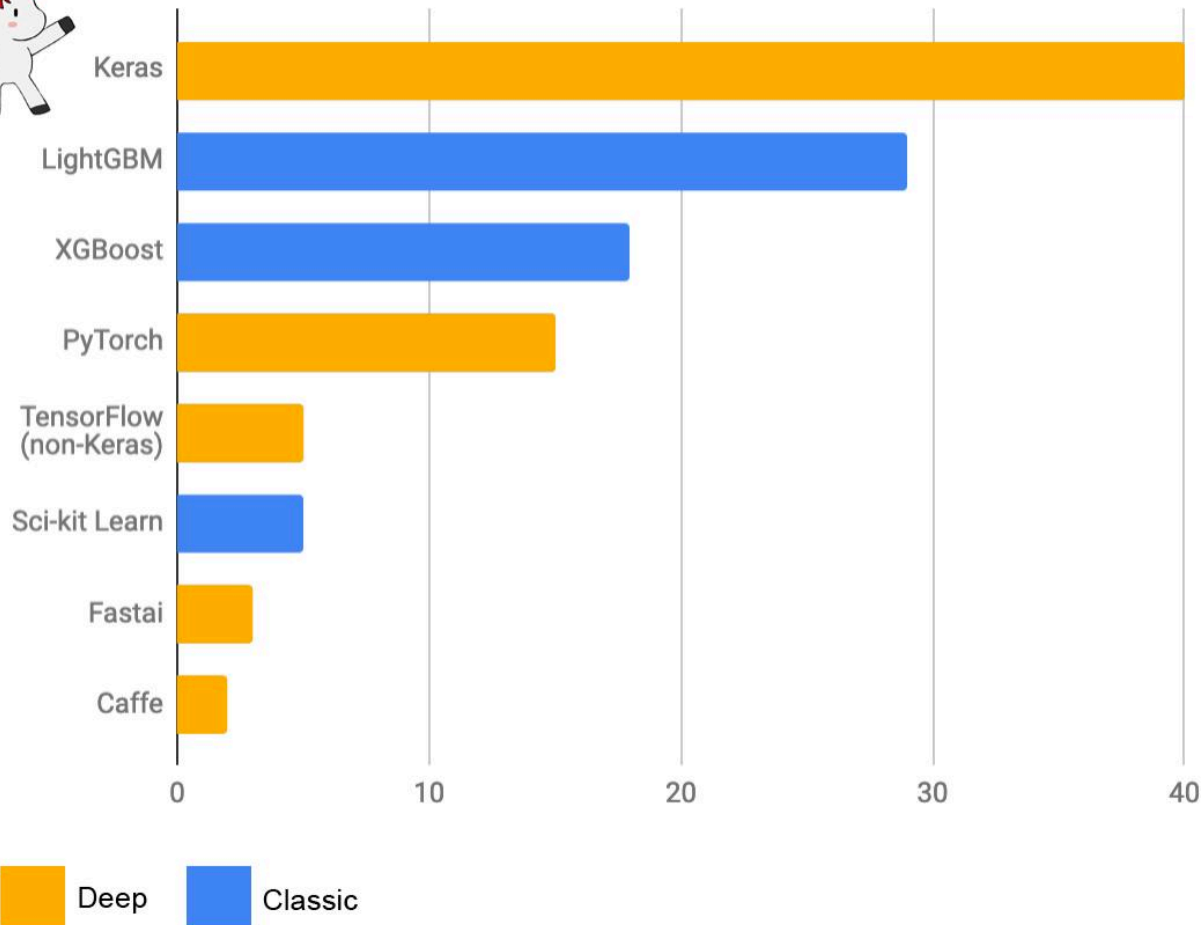
Additive models



François Chollet  @fchollet · Apr 3, 2019

What machine learning tools do Kaggle champions use? We ran a survey among teams that ranked in the *top 5* of a competition since 2016.

Primary ML software tool used by **top-5 teams** on Kaggle in each competition (n=120)



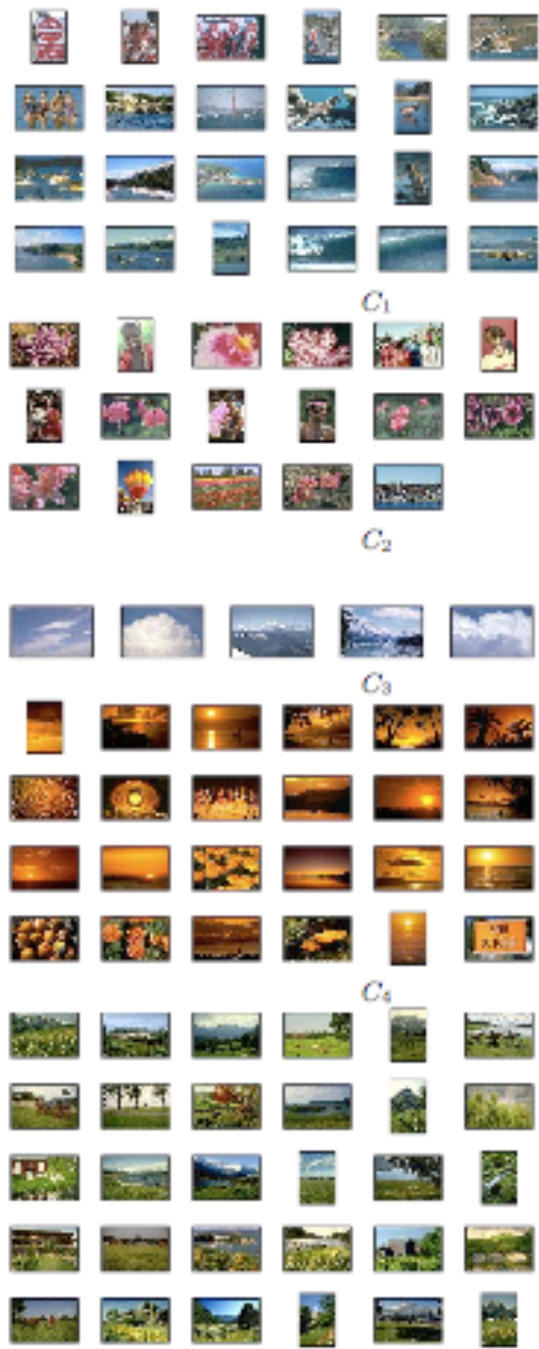
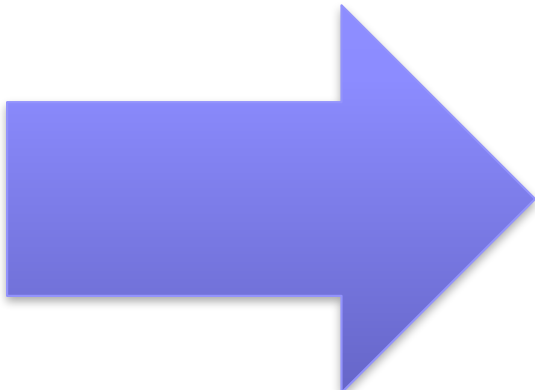
Bagging versus Boosting

- Bagging *averages* many **low-bias, lightly dependent** classifiers to reduce the variance
- Boosting *learns* linear combination of **high-bias, highly dependent** classifiers to reduce error

Clustering K-means



Clustering images



[Goldberger et al.]

Clustering web search results

web news images wikipedia blogs jobs more »

Clusty

race Search advanced preferences

clusters sources sites remix

All Results (238)




- Car (28)
- Race cars (7)
- Photos, Races Scheduled (5)
- Game (4)
- Track (3)
- Nascar (2)
- Equipment And Safety (2)
- Other Topics (7)
- Photos (22)
- Game (14)
- Definition (13)
- Team (18)
- Human (8)
 - Classification Of Human (2)
 - Statement, Evolved (2)
 - Other Topics (4)
- Weekend (8)
- Ethnicity And Race (7)
- Race for the Cure (8)
- Race Information (8)




[more](#) | [all clusters](#)




find in clusters: Find




Cluster **Human** contains **8** documents.




Search Results




- [Race \(classification of human beings\) - Wikipedia, the free ...](#)   




The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **rac**es, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History · Modern debates · Political and ...
[en.wikipedia.org/wiki/Race_\(classification_of_human_beings\)](http://en.wikipedia.org/wiki/Race_(classification_of_human_beings)) - [cache] - Live, Ask
- [Race - Wikipedia, the free encyclopedia](#)   

General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of human beings) **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games
en.wikipedia.org/wiki/Race - [cache] - Live, Ask
- [Publications | Human Rights Watch](#)   

The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
www.hrw.org/backgrounder/usa/race - [cache] - Ask
- [Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...](#)   

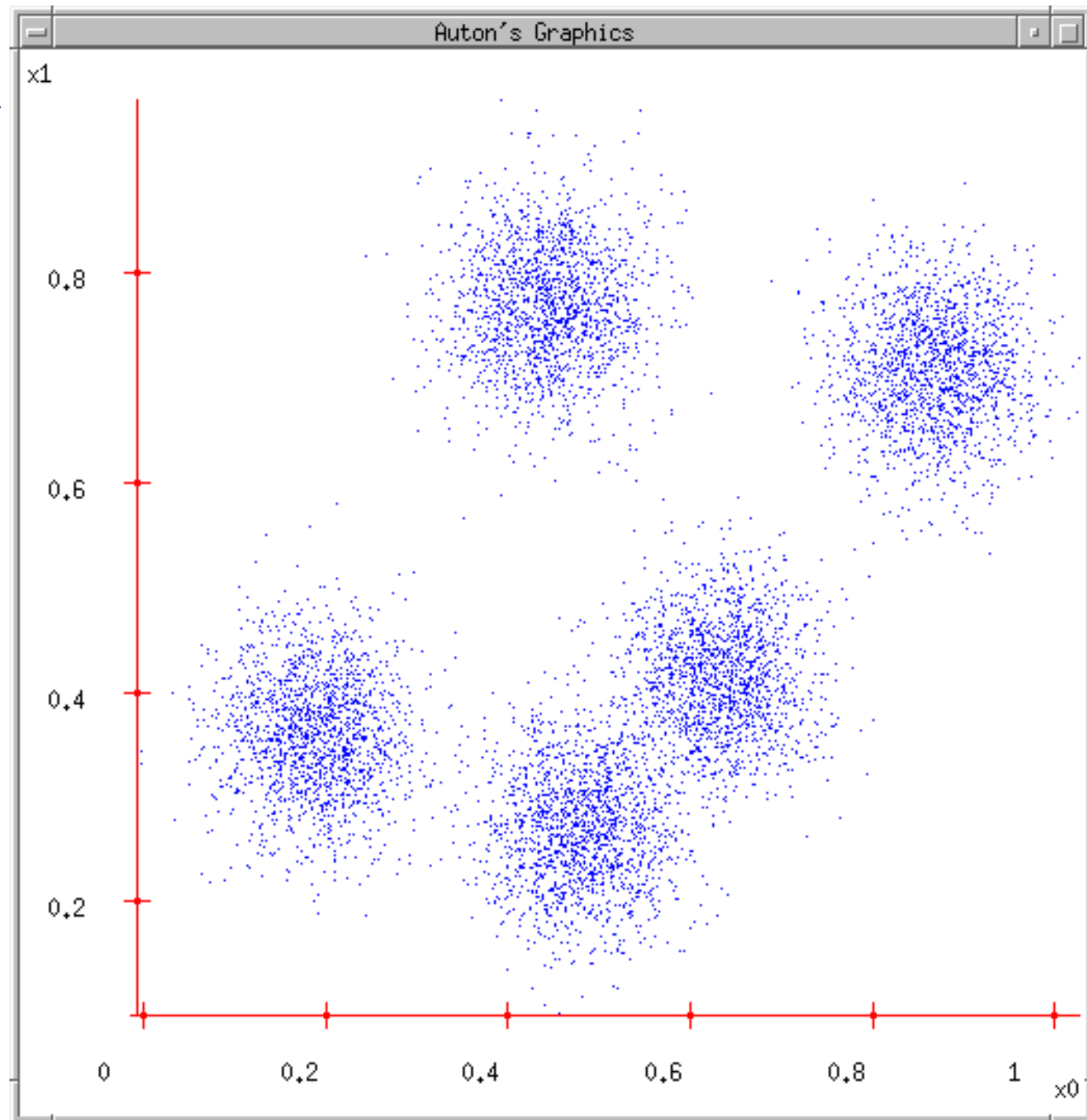
Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...** From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 - [cache] - Live
- [AAPA Statement on Biological Aspects of Race](#)   

AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study **human** evolution and variation, ...
www.physanth.org/positions/race.html - [cache] - Ask
- [race: Definition from Answers.com](#)   

race n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically transmitted physical
www.answers.com/topic/race-1 - [cache] - Live
- [Dopefish.com](#)   

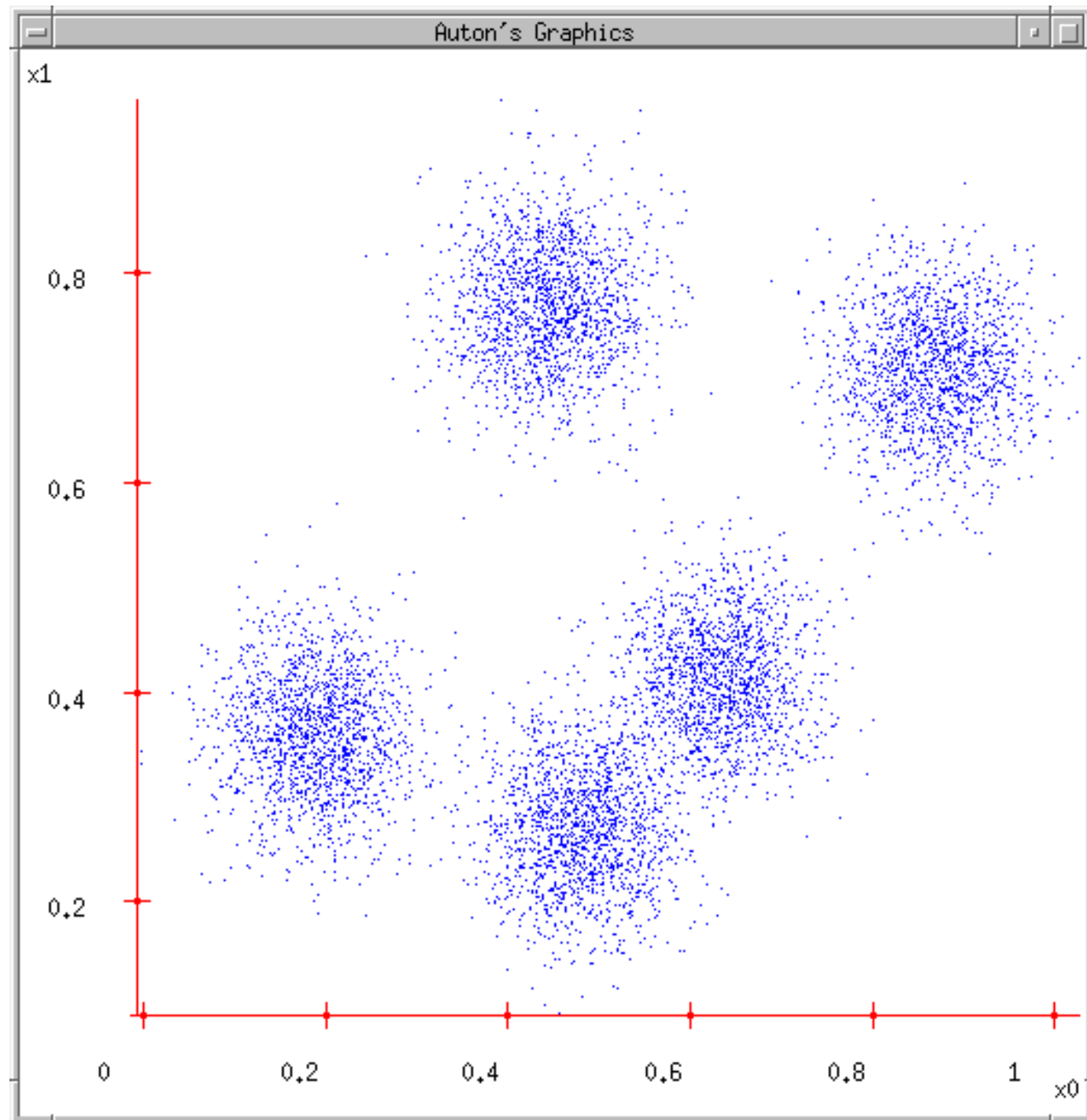
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
www.dopefish.com - [cache] - Open Directory

Some Data



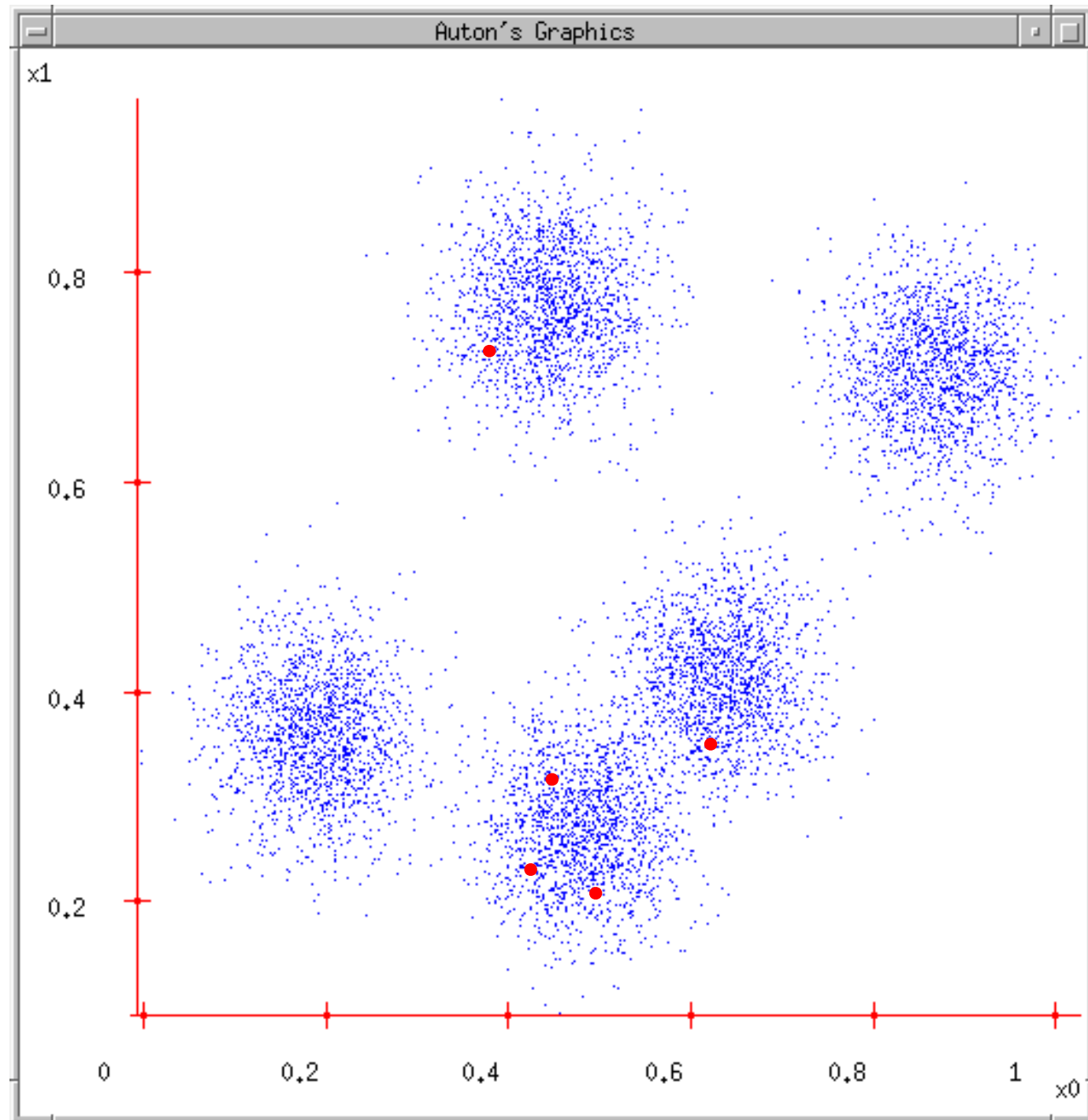
K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)



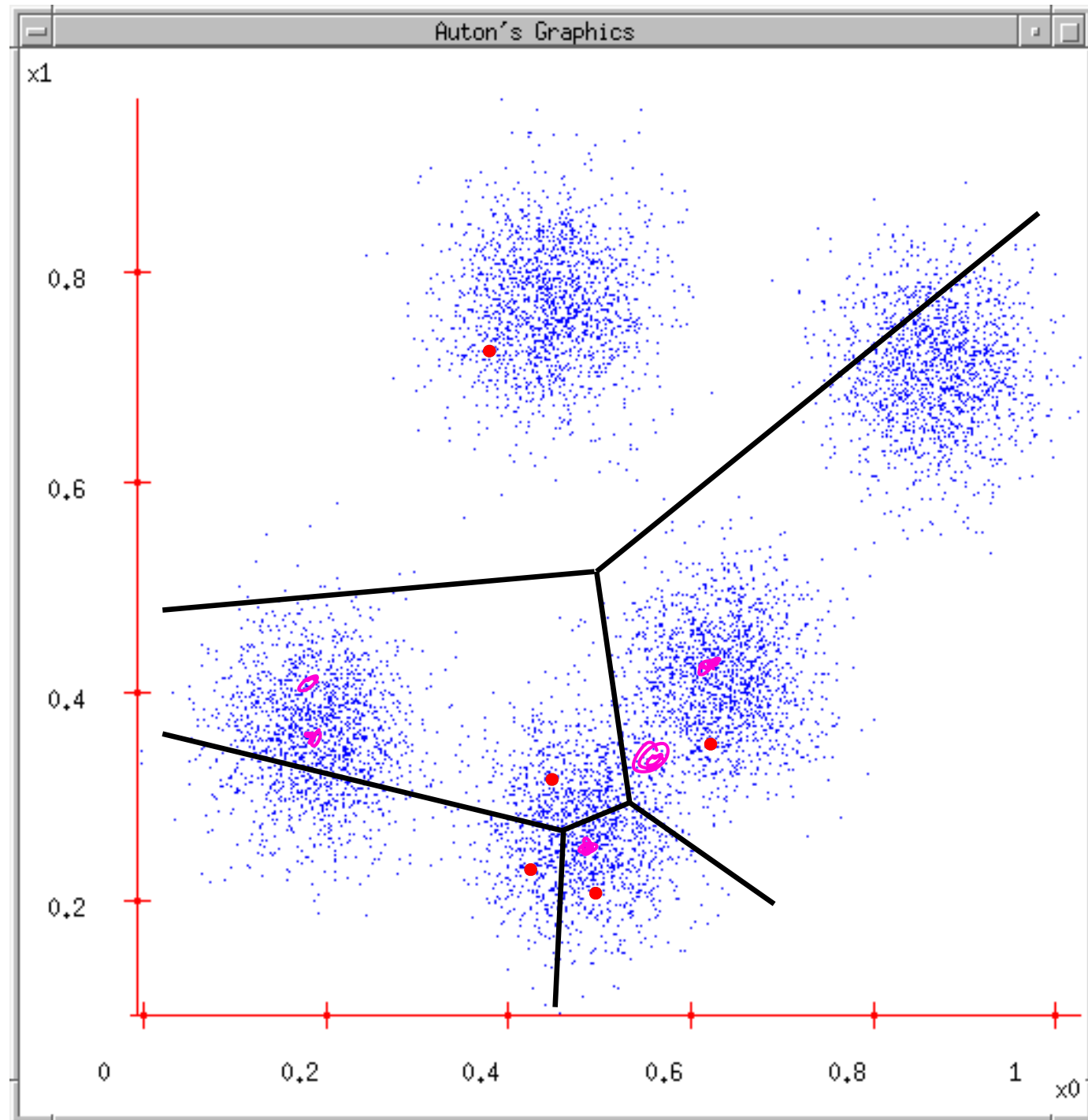
K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations



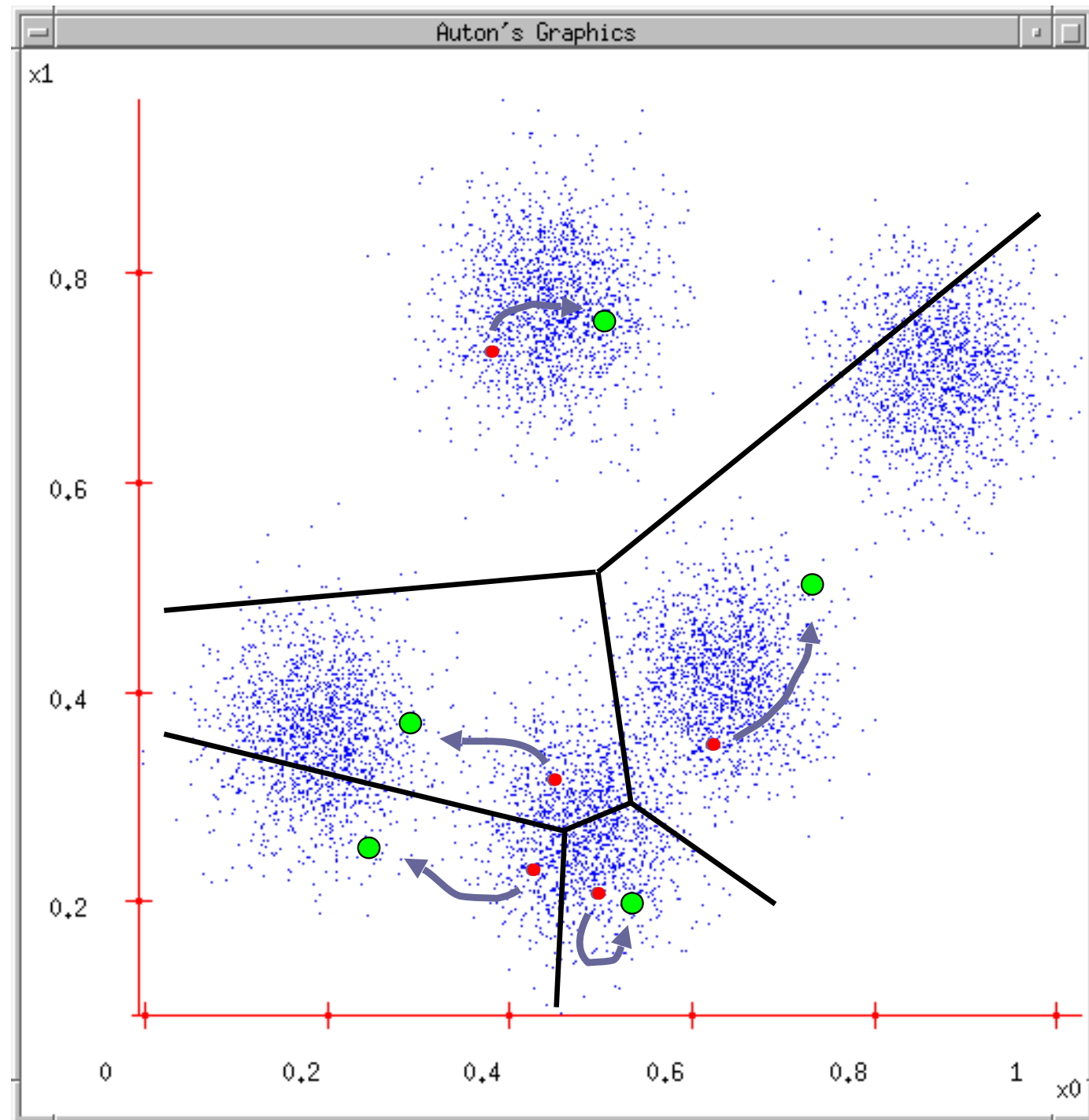
K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



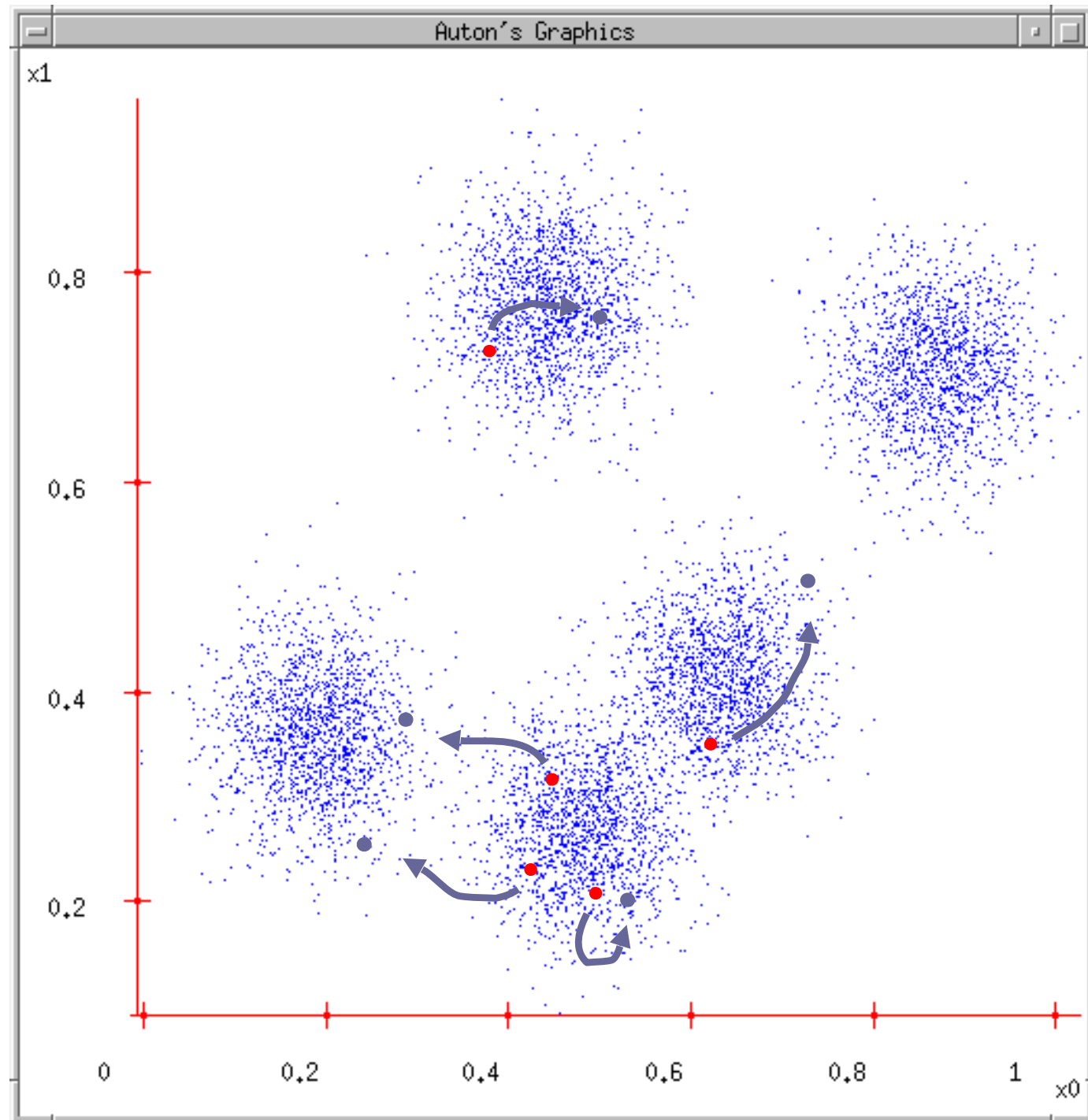
K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



K-means

- Randomly initialize k centers

- $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$

k-means ++

- Classify: Assign each point $j \in \{1, \dots, N\}$ to nearest center:

- $C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$

- Recenter: μ_i becomes centroid of its point:

- $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j:C(j)=i} \|\mu - x_j\|^2$

Does K-means converge???

Part 1

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \| \mu_i - x_j \|^2$$

- Fix μ , optimize C

Does K-means converge???

Part 2

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Fix C, optimize μ

Vector Quantization, Fisher Vectors

Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. *Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel*

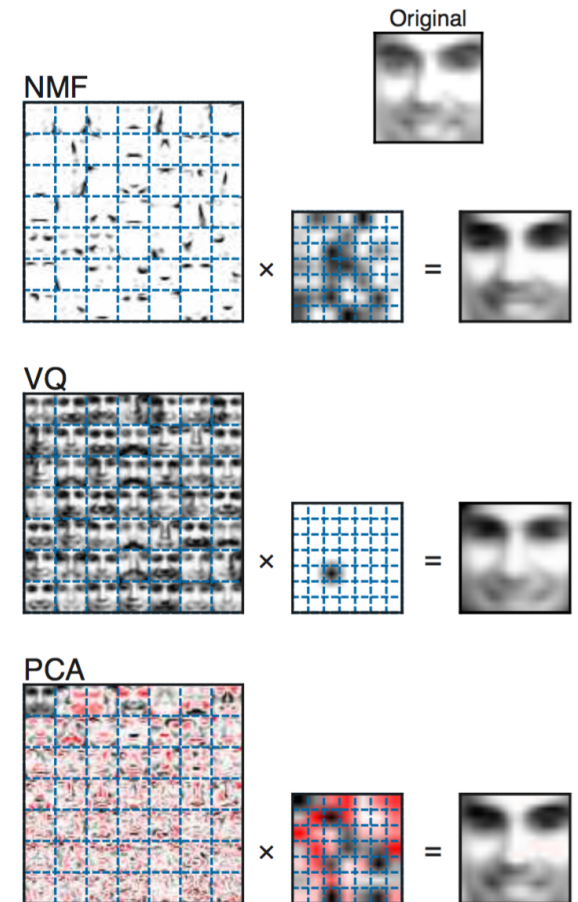
Vector Quantization, Fisher Vectors

Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. *Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel*



Vector Quantization, Fisher Vectors

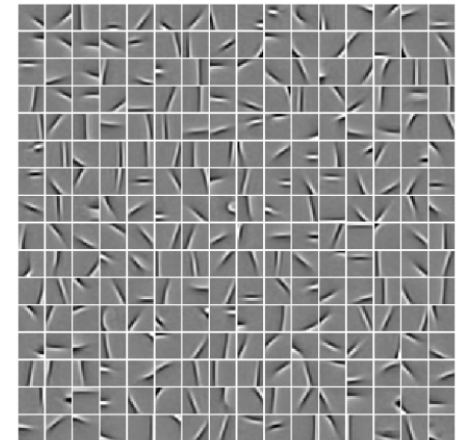
Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



FIGURE 14.9. *Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel*

Typical output of
on patches



Similar reduced representation can be used as a feature vector

Coates, Ng, *Learning Feature Representations with K-means*, 2012