STAT 391

5/25/23

Lecture 18

Logistic Regression

Regression examples

Clustering 2 lectures

HW 6 optional

Last HW deadline
Friday 5 pm ← Solutions 5,6

T.B Posted
- notes on linear & logistic regr.
- slides on Clustering

+.b. edited

Logistic Regression

Used for: Classification

$x \in \mathbb{R}^d$

$y \in \{0, 1\}$ "Label"

$\mathcal{D} = \{(x^i, y^i)\ i = 1:n\}$

Model $f(x) = \ln \dfrac{P_{Y|X}(y=1|x)}{P_{Y|X}(y=0|x)} = \beta^T x$ parametrization $\beta \in \mathbb{R}^d$

meaning

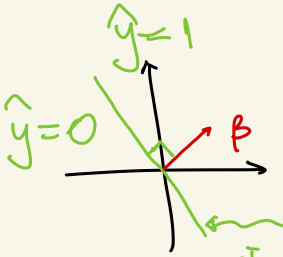
odds $\in (0, \infty)$

log odds $\in (-\infty, \infty)$

Classification with $f = \beta^T x$ (β known)

$\hat{y} = \dfrac{\text{sign } f(x) + 1}{2} \iff f(x) > 0 \iff P_{Y|X}(y=1|x) > P_{Y|X}(y=0|x)$

Probabilistic classifier



- $P_{y|x}[y=1|x] = p$

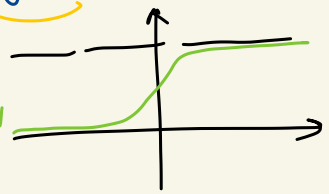
$$f = \ln \frac{p}{1-p} = \ln \frac{P_{y|x}(y=1|x)}{P_{y|x}(y=0|x)} \Rightarrow e^f = \frac{p}{1-p} \Rightarrow p = \frac{e^f}{1+e^f} \in (0,1)$$

$$f = \beta^T x$$

$$1-p = \frac{1}{1+e^f}$$

$$p + (1-p) = 1$$

$\frac{e^f}{1+e^f}$



- Estimating β by Max Likelihood

confidence of classification $P_{y|x}[y|x] = \frac{e^{yf}}{1+e^f}$

Likelihood

$$L(\beta) = Pr[y^{1:n} | x^{1:n}, \beta] = \prod_{i=1}^{n} P_{y|x}[y^i | x^i, \beta] = \prod_{i=1}^{n} \frac{e^{y^i \beta^T x^i}}{1+e^{\beta^T x^i}}$$

p for $y^i=1$, $1-p$ for $y^i=0$

log likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \left[y^i \beta^T x^i - \ln\left(1+e^{\beta^T x^i}\right) \right]$$

maximize over β by Gradient Ascent

$$\ell(\beta) = \sum_{i=1}^{n} \left[y^i \underbrace{\beta^T x^i}_{f} - \ln\left(1 + e^{\beta^T x^i}\right) \right]$$

Gradient $\frac{\partial \ell}{\partial \beta} \in \mathbb{R}^d$

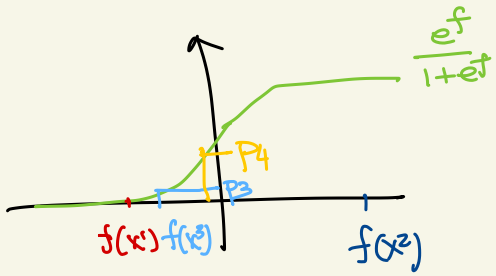
$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial f} \cdot \frac{\partial f}{\partial \beta_j} = \sum_{i=1}^{n} \left[y^i - \frac{e^{f(x^i)}}{1 + e^{f(x^i)}} \right] x^i$$

$$\frac{\partial \ell}{\partial \beta} = \left[\frac{\partial \ell}{\partial \beta_j} \right]_{j=1:d} = \frac{\partial \ell}{\partial f} \left[\frac{\partial f}{\partial \beta_j} \right]_{j=1:p}$$

$$\frac{\partial \ell}{\partial f} = \sum_{i=1}^{n} \left[y^i - \frac{e^{f(x^i)}}{1 + e^{f(x^i)}} \right] \in \mathbb{R}$$

$P_{y^i=1 \mid x^i}$ by the model

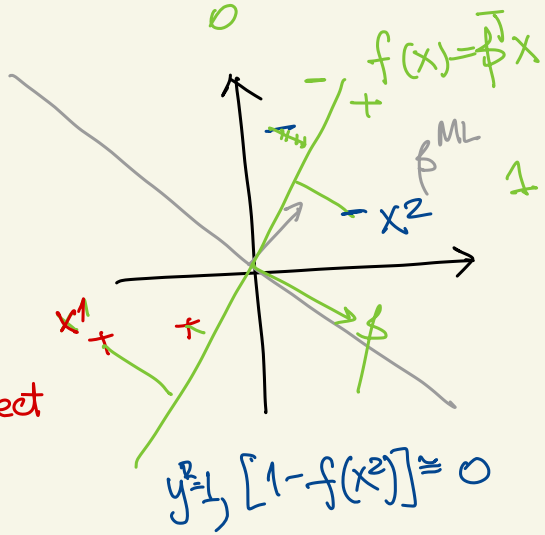
$$\frac{\partial f}{\partial \beta} = \frac{\partial}{\partial \beta}(\beta^T x) = x \quad \leftarrow \text{Exercise prove this} \quad \Leftrightarrow \quad \frac{\partial f}{\partial \beta_j} = x_j$$

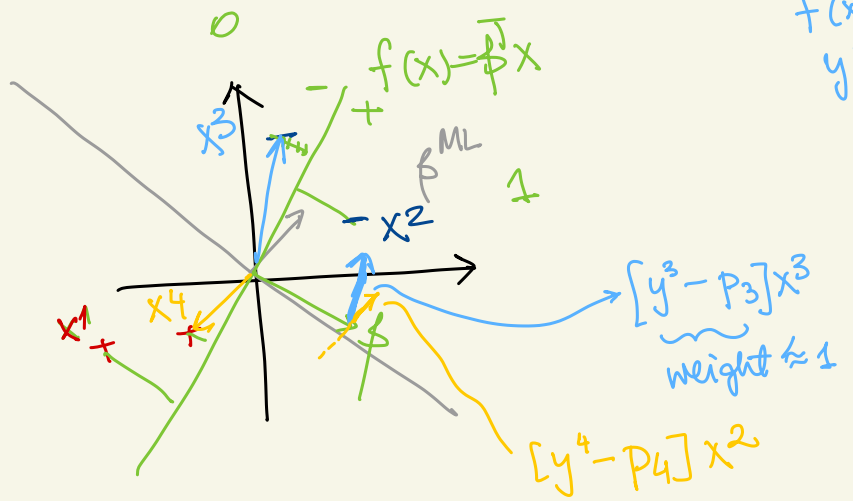


$y^1 = 0$

$$0 \approx \frac{e^{f(x^1)}}{1 + e^{f(x^1)}} = p_1 \begin{cases} |f(x^1)| \text{ large} \\ f(x^1) < 0 \text{ correct} \end{cases}$$

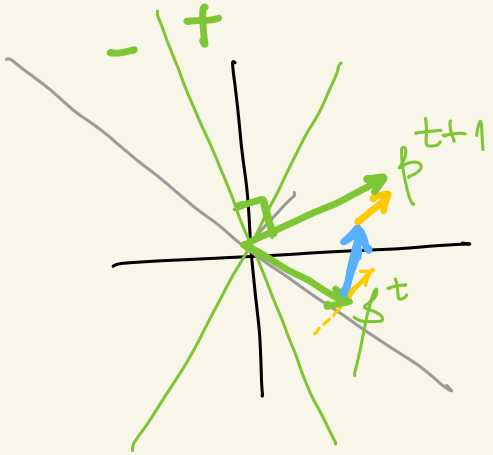
$$0 > [y^1 - p_1] \approx 0$$





$f(x^3) < 0 \Rightarrow y^3 - p_3 > 0$
≈ 1
$y^3 = 1$

$f(x^4) \lesssim 0 \Rightarrow y^4 - p_4 \approx -\frac{1}{2}$
$y^4 = 0$



$$\beta^{t+1} = \beta^t + \eta \frac{\partial \ell}{\partial \beta}$$
$$= \beta^t + \eta \sum_{i=1}^{n} [y^i - p^i] x^i$$

linear combination of x^i

Prediction — regression $y \in \mathbb{R}$ (linear)

≡ Supervised learning — classification $y \in \{0,1\}$ or discrete (logistic)

$P_{y|x}$ wanted

$D = \{(x^i, y^i)\}_{i=1:n}$ labels, output

Clustering : wanted Clusters = groups in data

Unsupervised wanted some feature of P_x distribution of x

$D = \{x^i\}_{i=1:n}$

no output

Ex: estimate - a density
- a discrete distribution

Ex: dimension reduction PCA

Ex: causal inference

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $x_1 \xrightarrow{\text{causes}} x_2$

x_1 = treatment, smoking; x_2 = effect, health cancer

Lecture Notes IX – Clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

May, 2023

Paradigms for clustering ←

+ what is clustering

Parametric clustering algorithms (K given)

- Cost based / hard clustering
- K-means clustering and the quadratic distortion
- Model based / soft clustering

Issues in parametric clustering

- Selecting K

Reading: Ch. 18

What is clustering? Problem and Notation

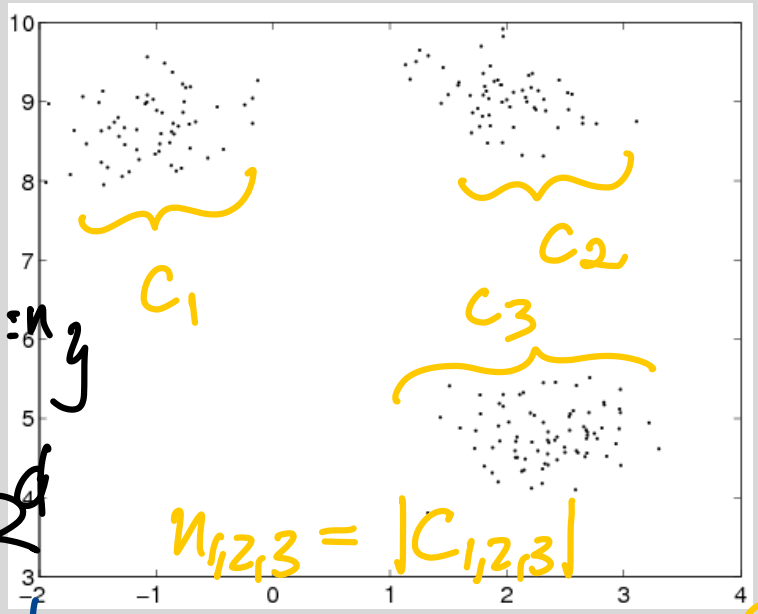
- **Informal definition** **Clustering** = Finding groups in data
- **Notation**

$\mathcal{D}$	=	$\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$ a **data set**
n	=	number of **data points**
K	=	number of **clusters** ($K << n$)
Δ	=	$\{C_1, C_2, \ldots, C_K\}$ a partition of $\mathcal{D}$ into disjoint subsets
$k(i)$	=	the **label** of point i
$\mathcal{L}(\Delta)$	=	cost (loss) of Δ (to be minimized)

- **Second informal definition** **Clustering** = given n **data points**, separate them into K **clusters**
- Hard vs. soft clusterings
  - **Hard** clustering Δ: an item belongs to only 1 cluster
  - **Soft** clustering $\gamma = \{\gamma_{ki}\}_{k=1:K}^{i=1:n}$
    γ_{ki} = the **degree of membership** of point i to cluster k

$$\sum_k \gamma_{ki} = 1 \quad \text{for all } i$$

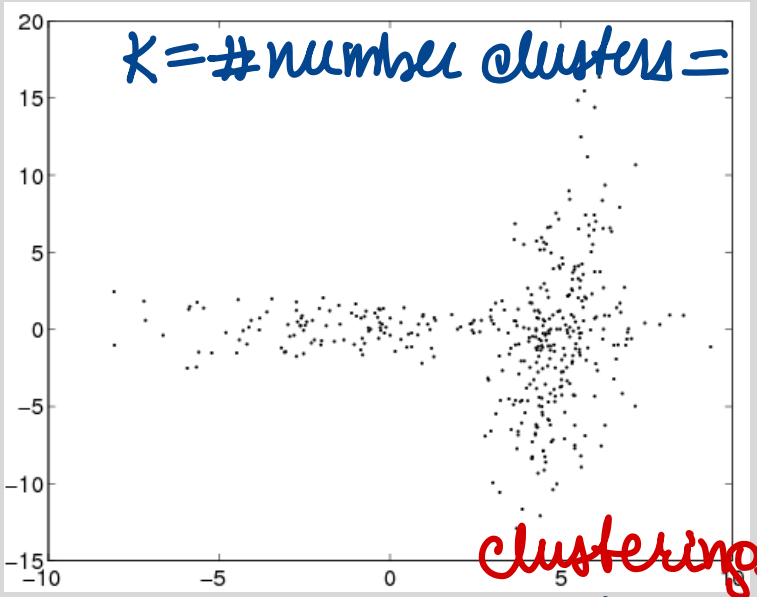
(usually associated with a probabilistic model)



$\mathcal{D} = \{x^{1:n}\}$

$x \in \mathbb{R}^d$

$n = |\mathcal{D}|$

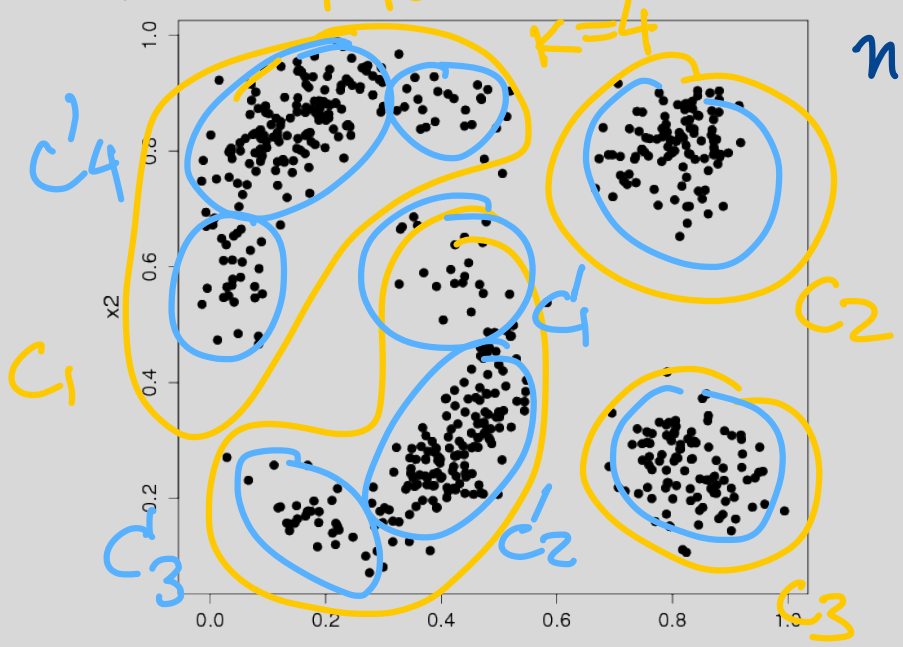


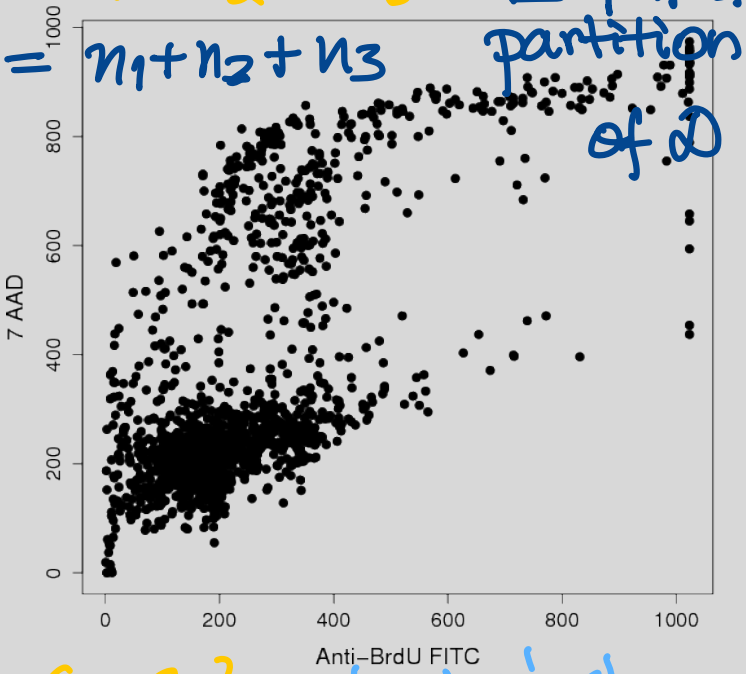
$C_{1,2,3}$ = clusters $\mathcal{D} = C_1 \cup C_2 \cup C_3$

$\Delta = \{C_1, C_2, C_3\} \Rightarrow K = 3$

$n = n_1 + n_2 + n_3$

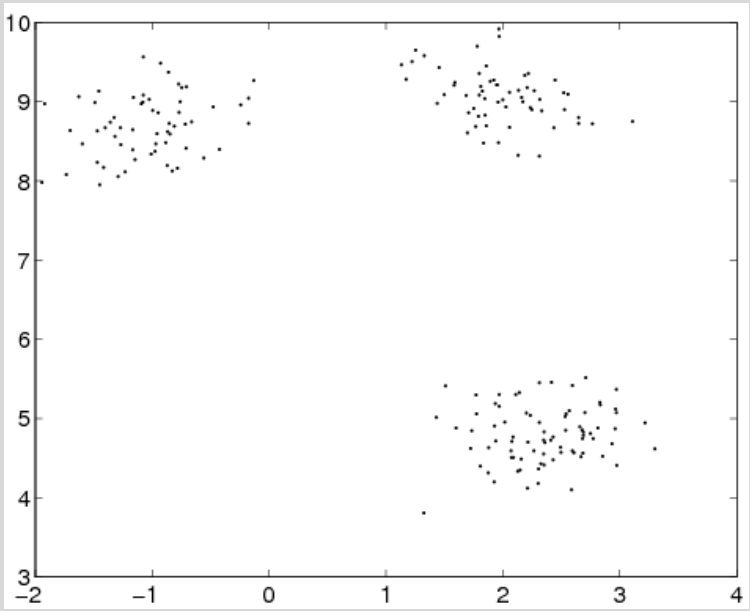
partition of $\mathcal{D}$

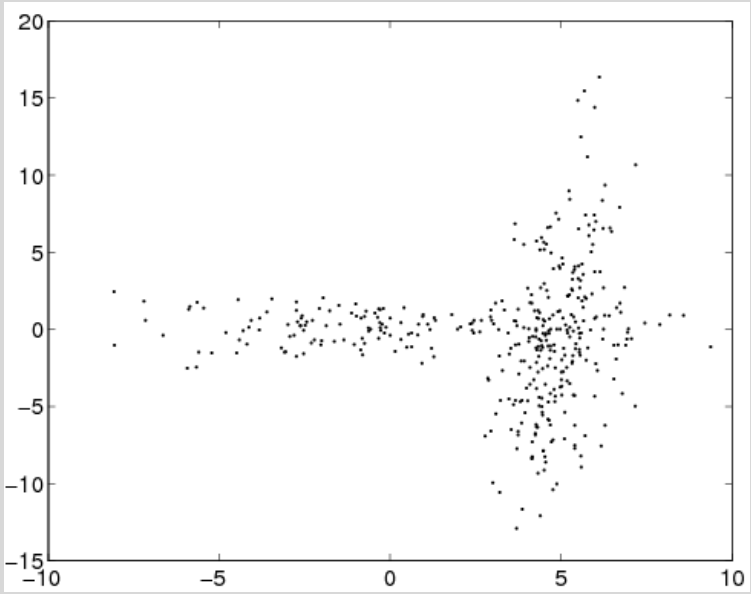


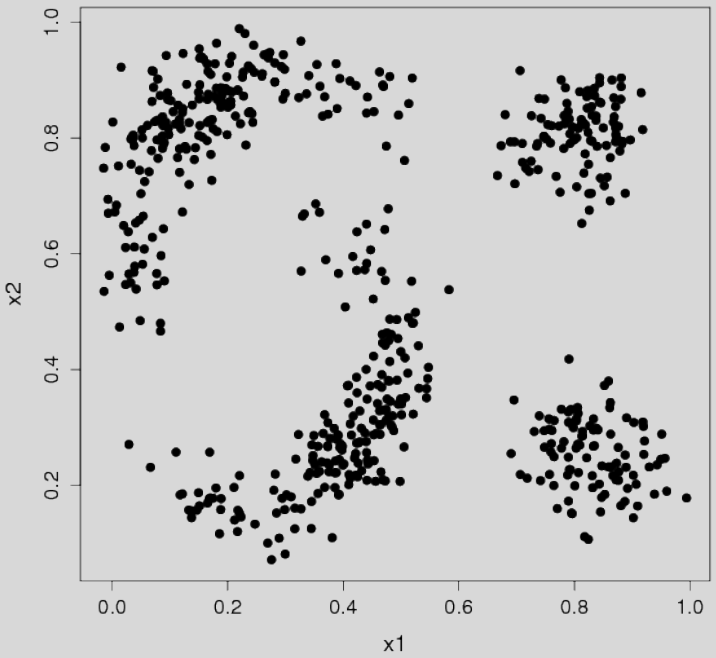


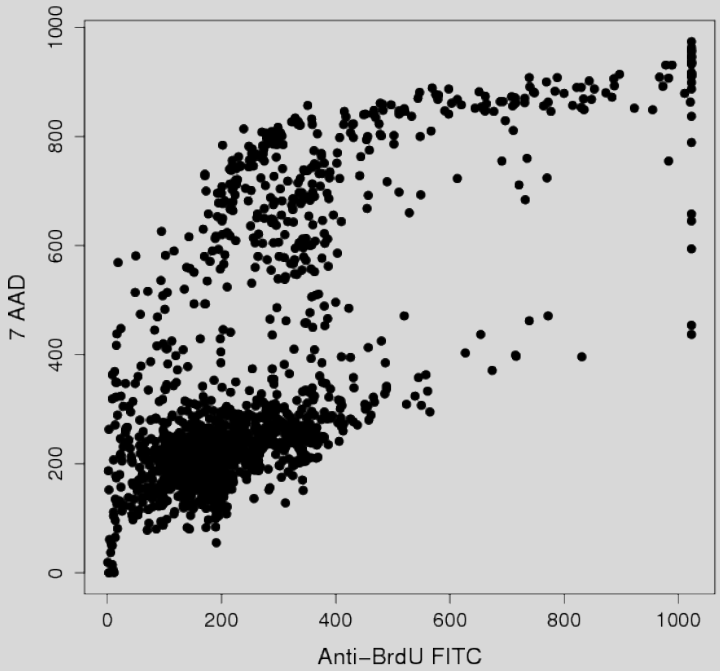
$\Delta = \{C_1, C_2, C_3, C_4\}$ $\Delta' = \{C_1', C_2', \dots\}$

(from [Nugent and Meila, 2010])









(from [Nugent and Meila, 2010])

Paradigms

← each method has own definition of cluster

Depend on type of data, type of clustering, type of cost (probabilistic or not), and constraints (about K, shape of clusters)

- Data = vectors $\{x_i\}$ in $\mathbb{R}^d$

Parametric	Cost based [hard]
(K known)	Model based [soft]
Non-parametric	Dirichlet process mixtures [soft]
(K determined	Information bottleneck [soft]
by algorithm)	Modes of distribution [hard]
	Gaussian blurring mean shift[Carreira-Perpinan, 2007] [hard]

- Data = similarities between pairs of points $[S_{ij}]_{i,j=1:n}$, $S_{ij} = S_{ji} \geq 0$ **Similarity based clustering**

Graph partitioning	spectral clustering [hard, K fixed, cost based]
	typical cuts [hard non-parametric, cost based]
Affinity propagation	[hard/soft non-parametric]

Classification vs Clustering

	Classification	**Clustering**
Cost (or Loss) $\mathcal{L}$	Expectd error	many! (probabilistic or not)
	Supervised	Unsupervised
Generalization	Performance on new data is what matters	Performance on current data is what matters
K	Known	Unknown
"Goal"	Prediction	Exploration Lots of data to explore!
Stage of field	Mature	Still young