Data driven-computing and graphic display:

Is *Data Science* a science yet?

Hsieh Fushing
University of California at Davis

At Computer Science, NCHU
Taichung, Taiwan
March 9th, 2018

World of Internet & Age of Big Data.

***Data scientist*, co…ol!!!**

- Statistician!
- Data miner!

Data scientists ….two books of interest

- "The ***Internet*** of Us:
- Knowing More and Understanding Less
- at the Age of Big Data"
- By Michael P. Lynch

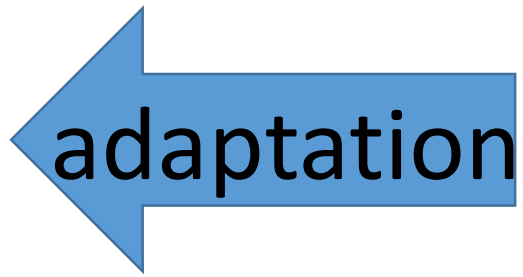
- "Thank You for Being Late:
- An Optimist's Guide to the thriving
- at the **Age of Accelerations**"
- By Thomas L. Friedman

The **age** of accelerations

- The Moore’s Law
- The Market
- The Mother Nature

How to survive under the age of accelerations?

- The Moore's Law
- The Market
- The Mother Nature



AI

Intelligence: the ability to acquire and apply knowledge and skills

- **Artificial “intelligence” (AI)** or **Algorithm-driven “intelligence” (AI)**
- In computer science: study of "intelligent agents“ .
- In general: a machine ***mimics*** "cognitive" functions of human on "learning" and "problem solving“. (***Mimicking ????)***

- **Deep Learning** is a Black Box.

- " A deep-learning system **doesn't** have any **explanatory power**,.. "

- ..... **Geoffrey Hinton** **(U. of Toronto)**

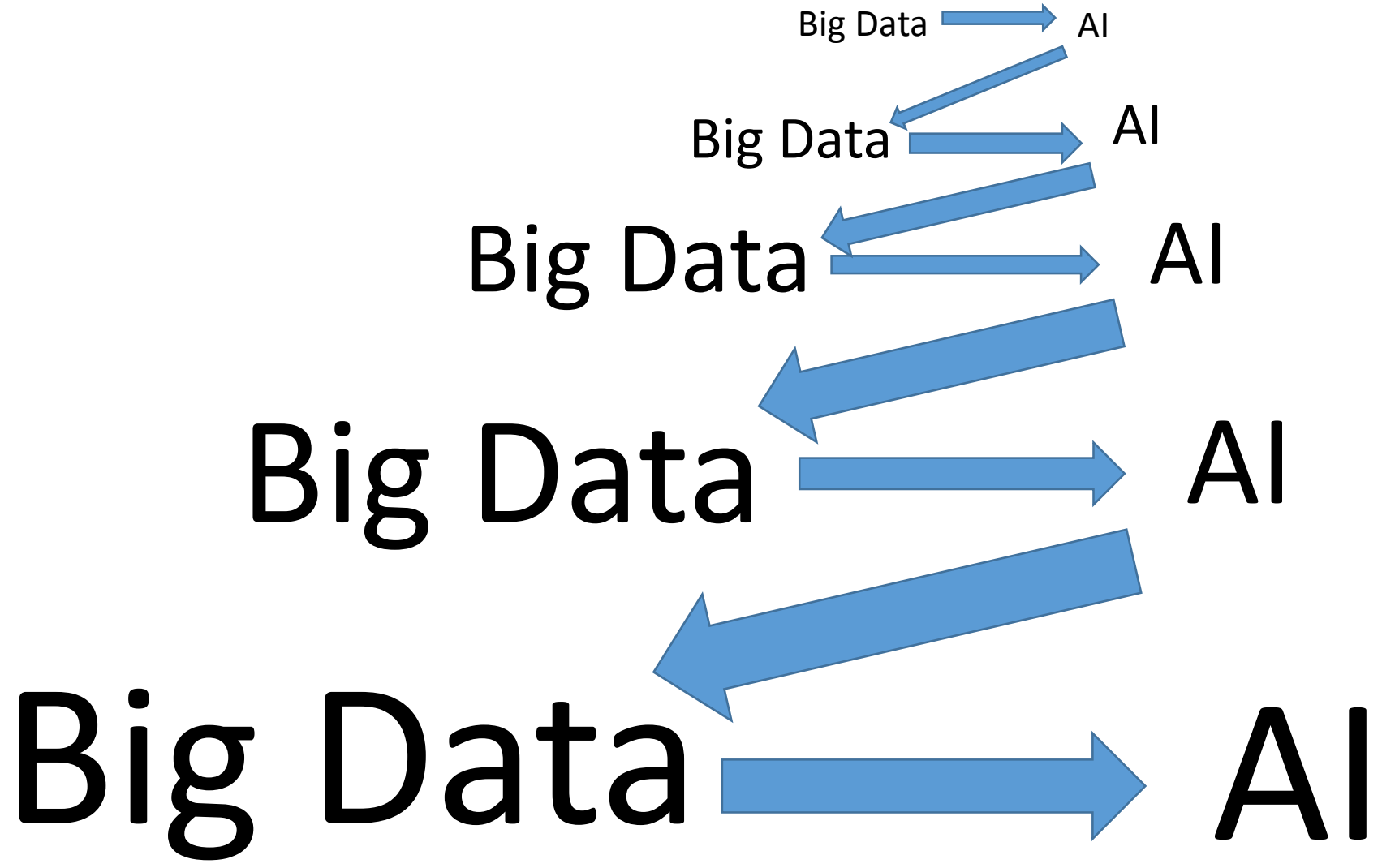
Mimicking…… ?????

- Brain research .. …. still in the age of “ hunting and gathering”

- How about ?

… **Extracted intelligence from**

Big Databases


Big Data
AI
Big Data
AI
Big Data
AI
Big Data
AI
Big Data
AI

From “big data” → *AI*.

This link → is a potential birthplace of “ Data Science”.

What kind of big data? **Database of structural data**

What kind of AI? **Not “Deep Learning”,**

nor “Random forest”.

What is data?

- pattern vs ***code***.
- nature vs **expert**.
- Unstructural vs structural data:
- 1. [high-speed-recoding video] vs [**PITCHf/x data**].
- 2. [**Photo and image**] vs [???]
- 3. [**Genome sequences**] vs [???]

AI and mimicking

- How to **mimic a data big matrix**?
- What if **AI and deep learning** are built on **ensembles of mimicries**?
- An **effective learning** with reinforcement?

Three phases of baseball data analysis on PITCHf/x

- **Pitching mechanics**
- Pitcher-batter interacting patterns
- Pitcher-batter psychological contests

PITCHf/x Database.

- Since **2008** in Major League Baseball (MLB).
- **Sportvision**, a California private company contracted by MLB, has installed cameras in all **30 MLB stadiums**.
- Data of all **pitch-by-pitch** delivered by **all pitchers** in **all MLB games**.
- Available at *http://gd2.mlb.com/components/game/mlb/.*

Sizes of data

- **30** MLB teams;
- Each team has **3-5 (healthy) starting pitchers**;
- Total each team plays **162 games** (81 at home stadium) per season.
- Each starting pitcher pitches around **100 pitches per game**.
- Each pitcher pitches **more than 3000 pitches per season**.

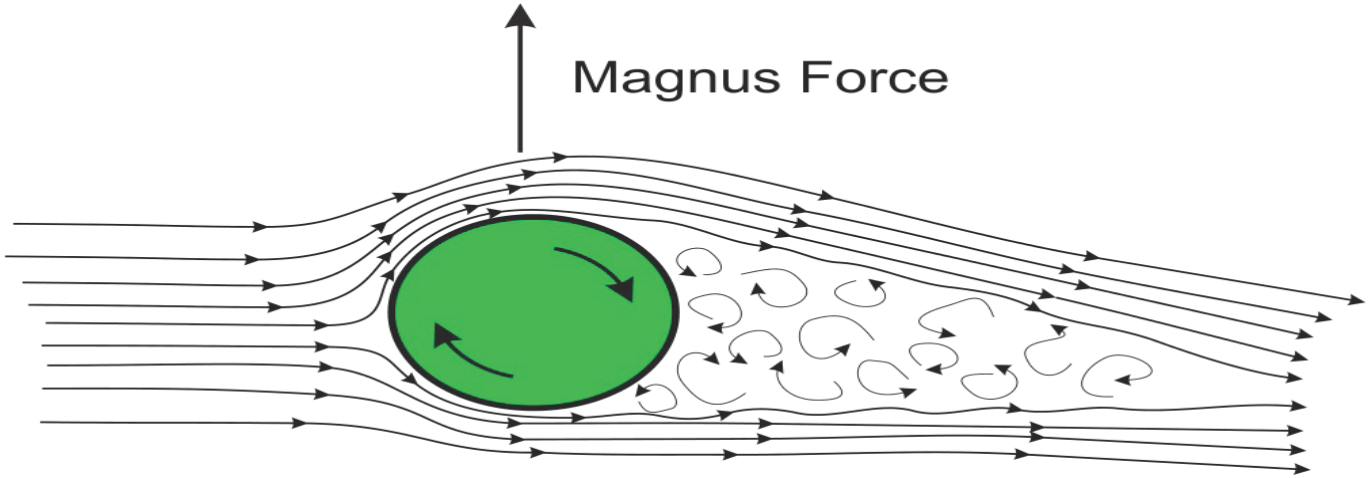
About pitching mechanics……

- **A flight of a baseball** after from pitcher's releasing point to home plate;
- Its **18.35m** trajectory **length** and less than **0.5 sec duration**.
- Its **3D trajectory video** (temporal axis).
- Its focuses on **vertical and horizontal movements**.

Physical Law, aerodynamic *effect* and bio-mechanical parameters.

- **Newton's gravitation**;
- **Magnus effect** caused by speed-spin interactions;
- Bio-mechanistic parameters (of pitcher's physical strengths);
- Weather conditions;
- **baseball seam and surface**.

Magnus effect:



- **Fastball** (FF & FT): back-spin to go **upward**;
- **Curveball**(CU): top-spin to **drop**;
- Cutter (FC), Slider (SL), Sinker(SI), Changeup(CH): side-spins to go **left and right**;

- ***So Magnus effect is shown through:***
- Spin-rate vs. speed vs. vertical movement vs. curvature;
- Spin-direction vs. **horizontal movement** vs. **acceleration**;

Instead of being described by an aerodynamic formula, each pitch is summarized into 21 features.

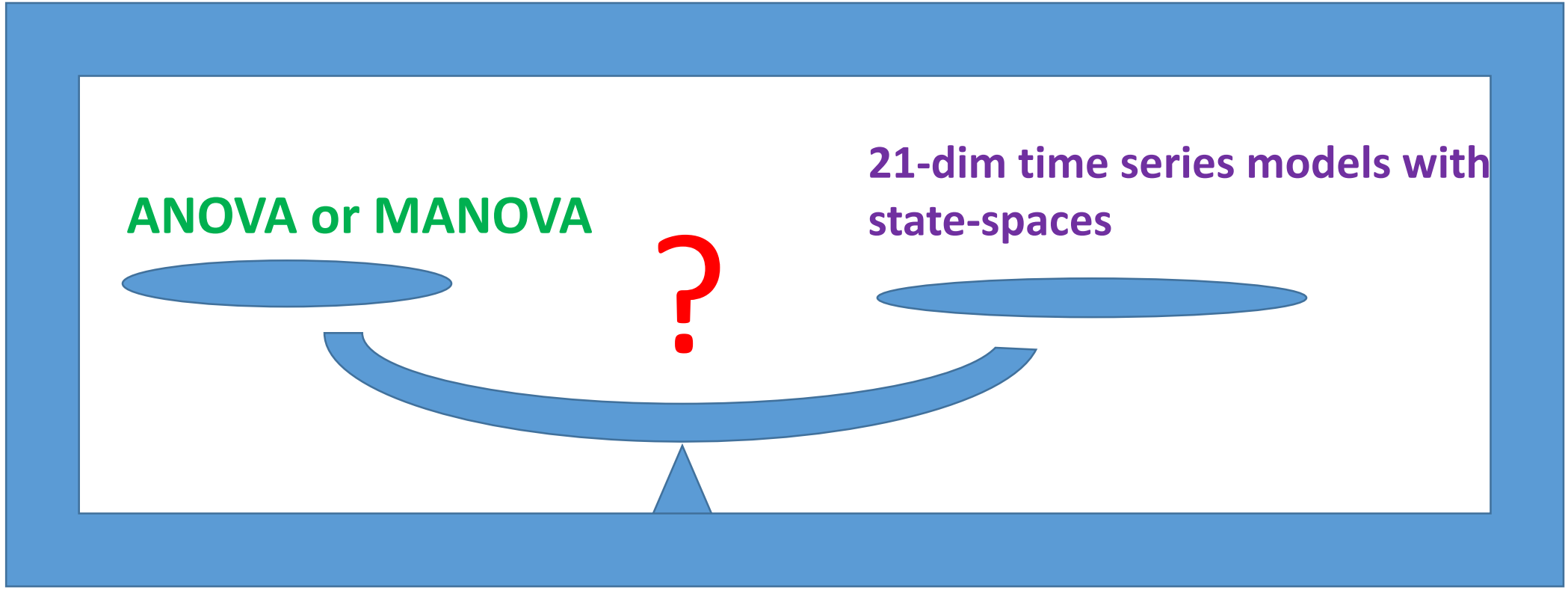
- Releasing point coordinates (x0, z0);
- Speeds at releasing points (vx0, vy0, vz0);
- Accelerations at releasing point (ax, ay, az);
- Start_speed; end_speed;
- Spin_dir; spin_rate;
- Beak_length; Break_angle; beak_y;
- Movements: pfx_x; pfx_z;
- px; pz;
- X, y.

See a pitcher's *pitching mechanics* in PITCHf/x data through .....

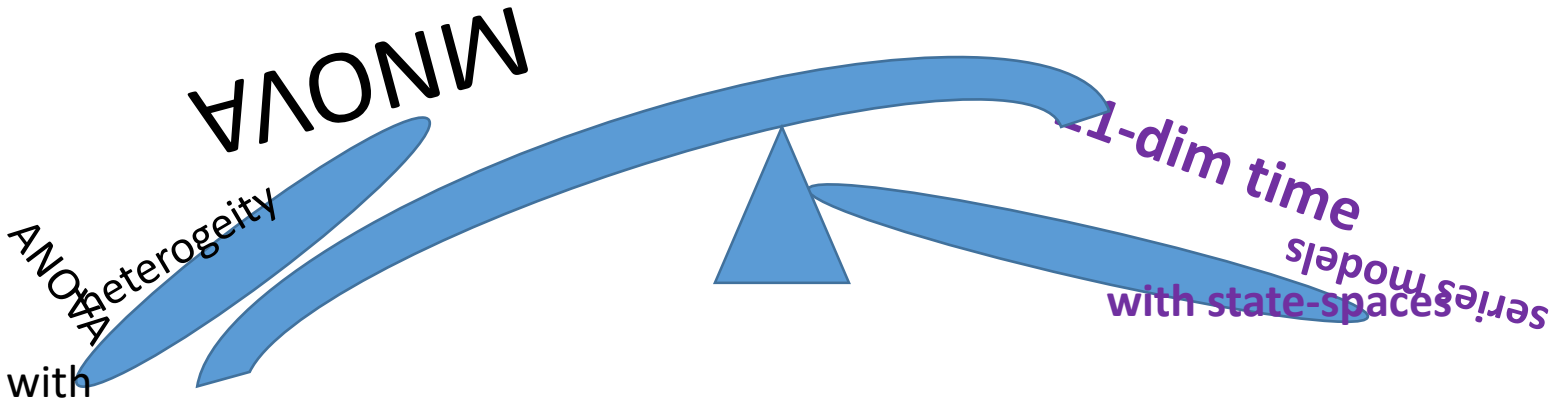
3000^{+} points of **21-dim point-cloud geometry** + ***Time***

- **Scientific problems:** How to compare many **k-dim point-cloud "geometrics"** ***along time axis***?

Thinking within Statistics Box



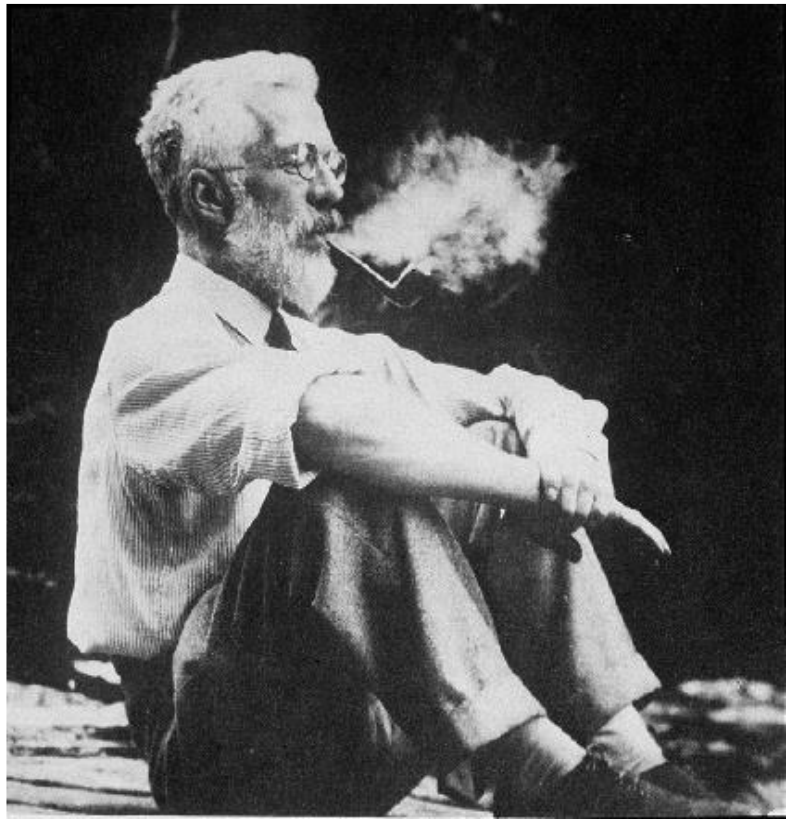
What if without box!



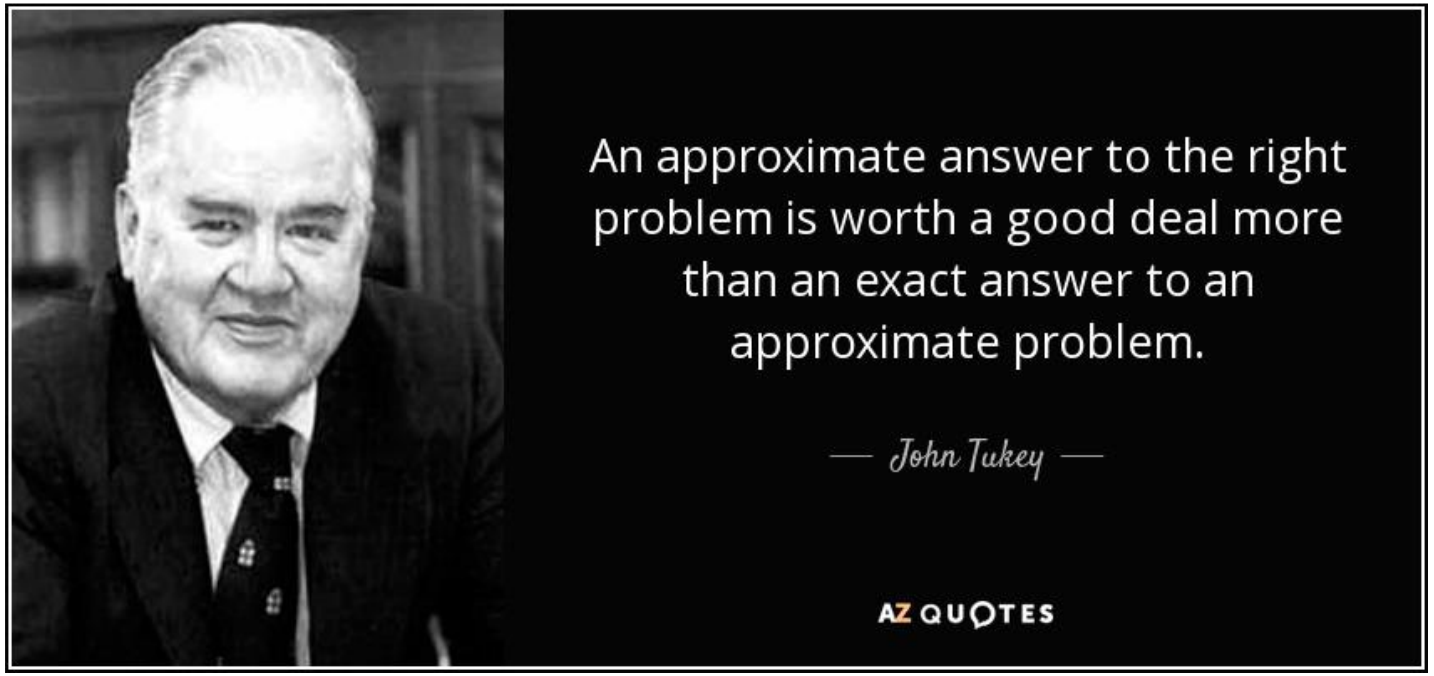
See a pitcher's *pitching mechanics* in PITCHf/x data

via *visible*…….

- **systemic comparison** on **all pitchers** and **idiosyncratic evolution on every single pitcher!!!!!**



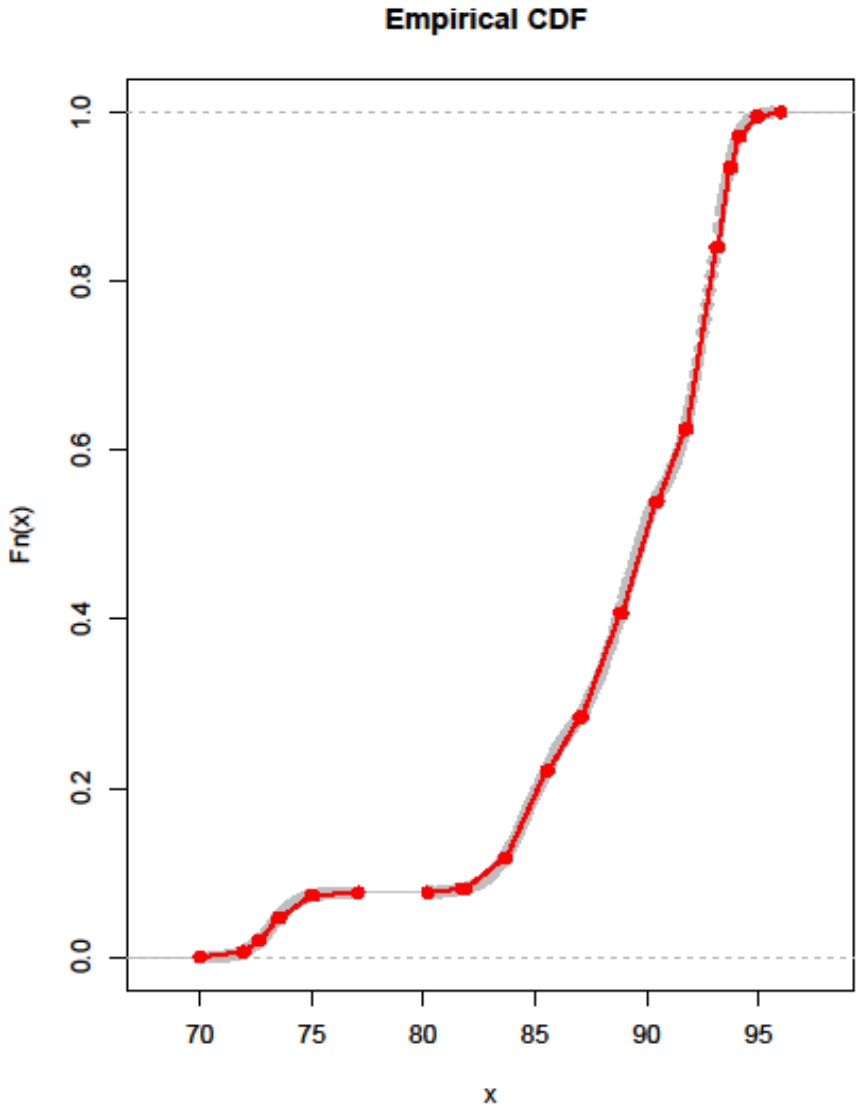
…So ***Data analysis*** ! John! They are your ***problems!!***

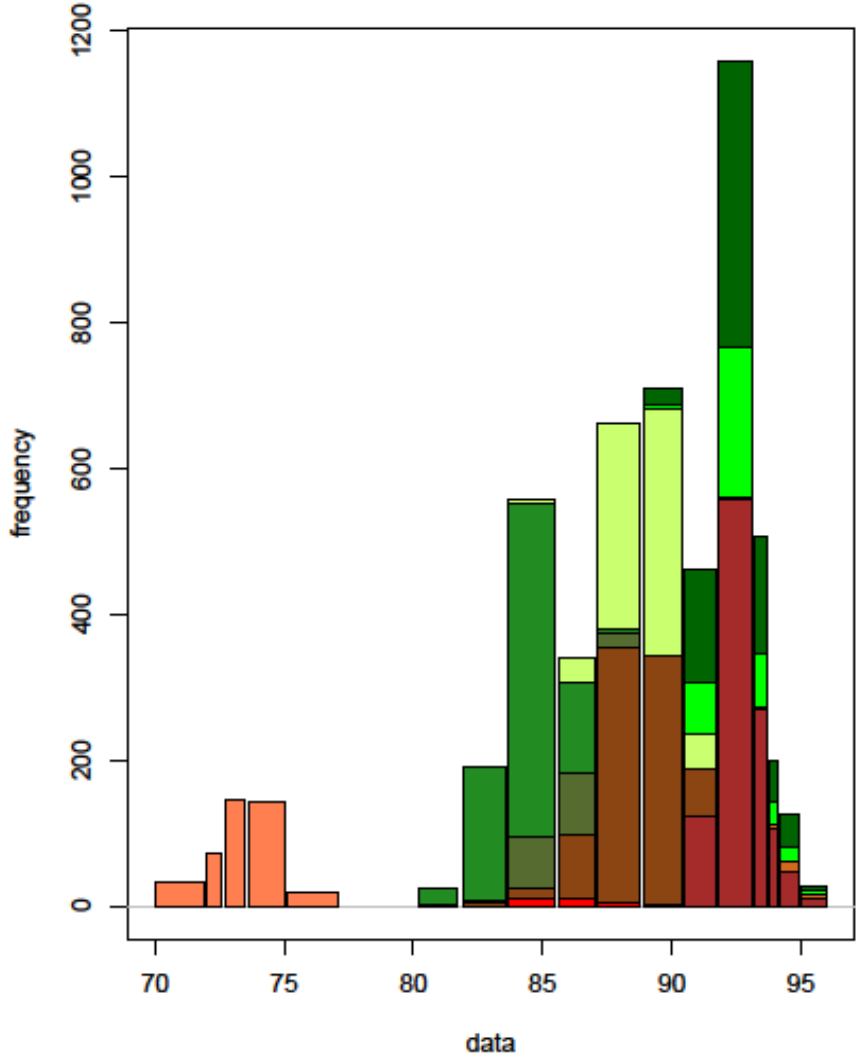


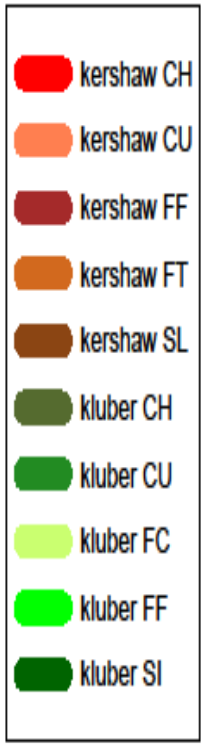

C
ELAND


Dodgers
CLAYTON
ERSHAW 22

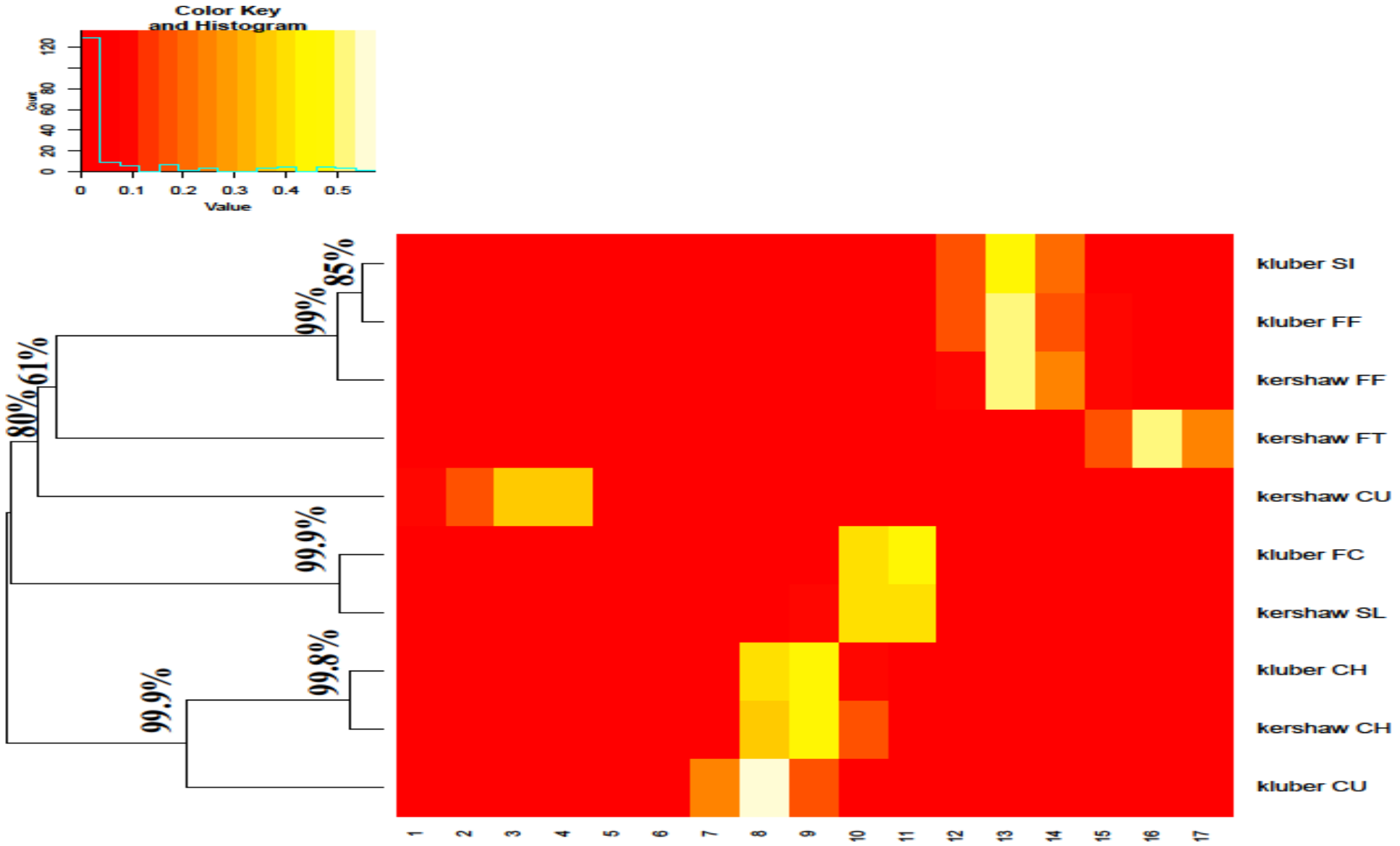
Comparing the two pitchers via Analysis of Histogram (ANOHT) on start-speed



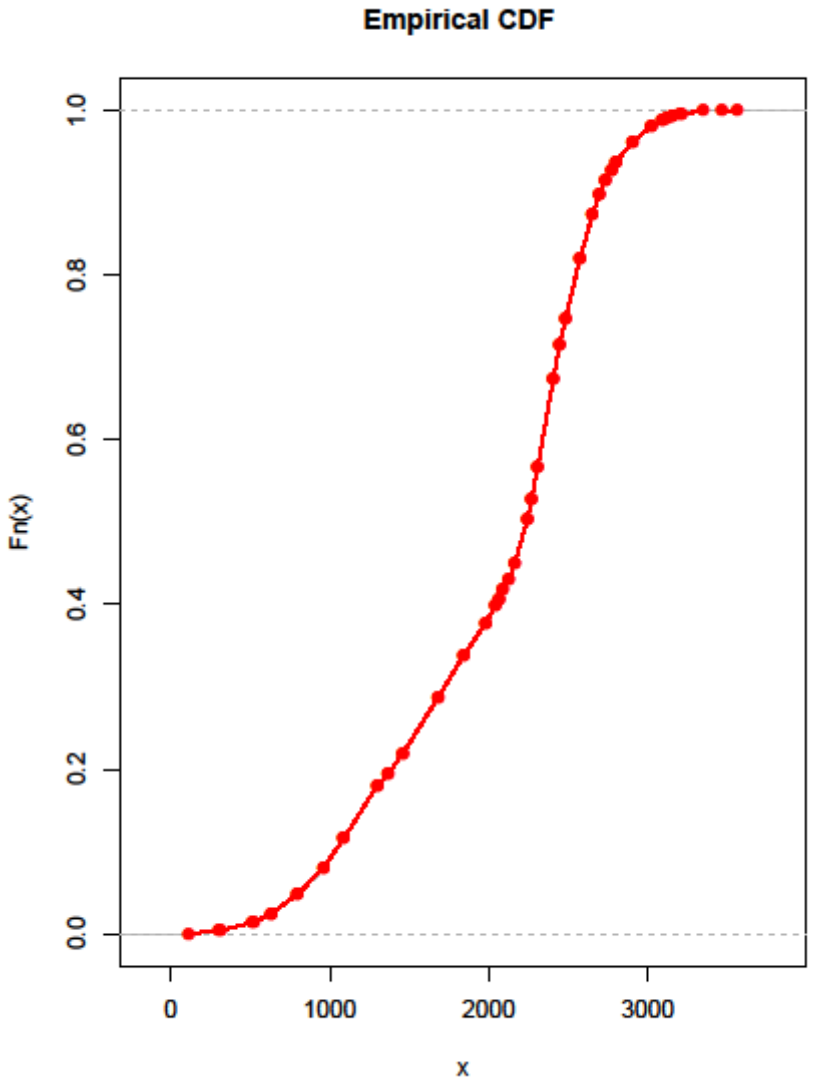


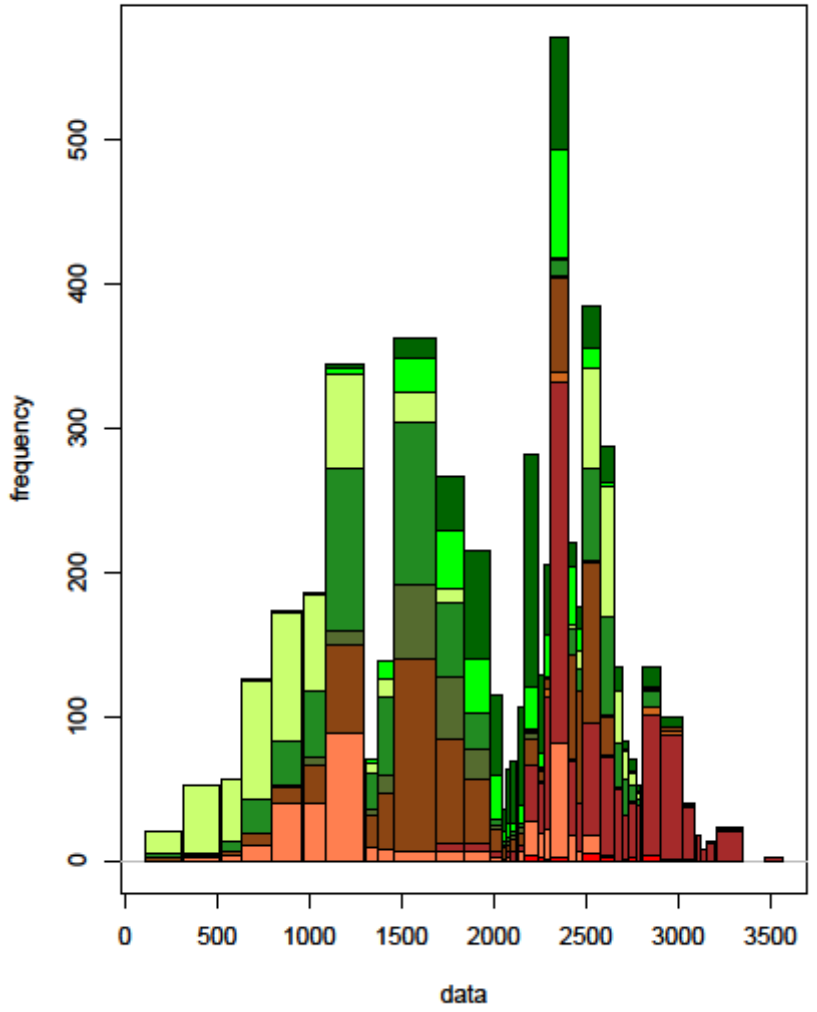


Analysis of Histogram(ANOHT) ….(Fushing and Roy (2018) Royal Society -Open Science)

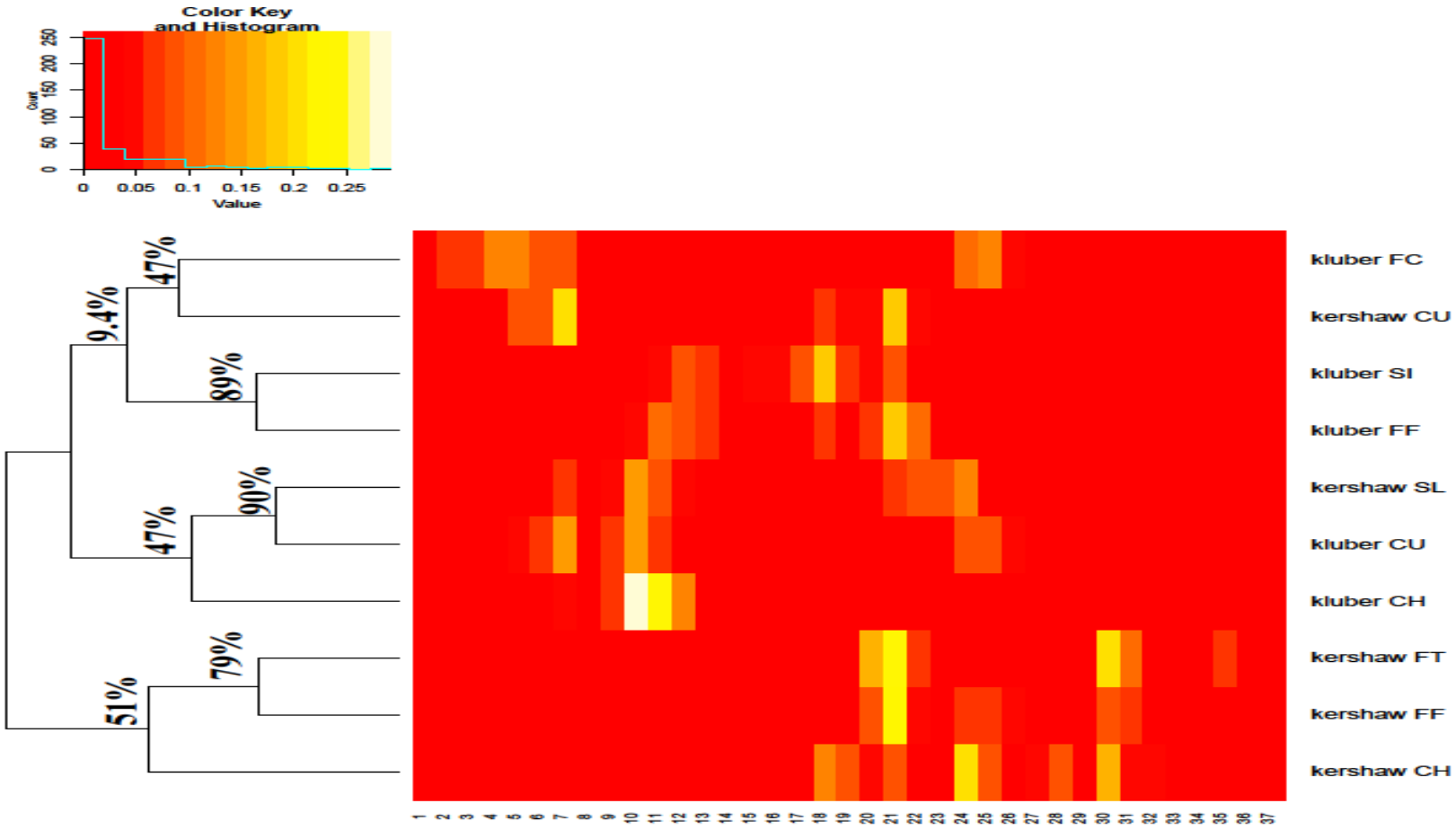


Analysis of Histogram(ANOHT) on spin-rate





Analysis of Histogram(ANOHT) on spin-rate



21 features are created by pitching mechanics, simultaneously.

- **Pitching mechanics ≅ Dependency** among the 21 features.

- ***How to see dependency ?***

- ***Via high-dimensional contingency "hypercube"*** ***upon*** **discrete, categorical or categorized features.**

- ***How to evaluate dependency*** **(asymmetric Association)?**

- via **mutual conditional entropy.[Categorization via possibly-gapped histogram**;

21 features are created by pitching mechanics, simultaneously.

- **Pitching mechanics ≅ Dependency** among the 21 features.

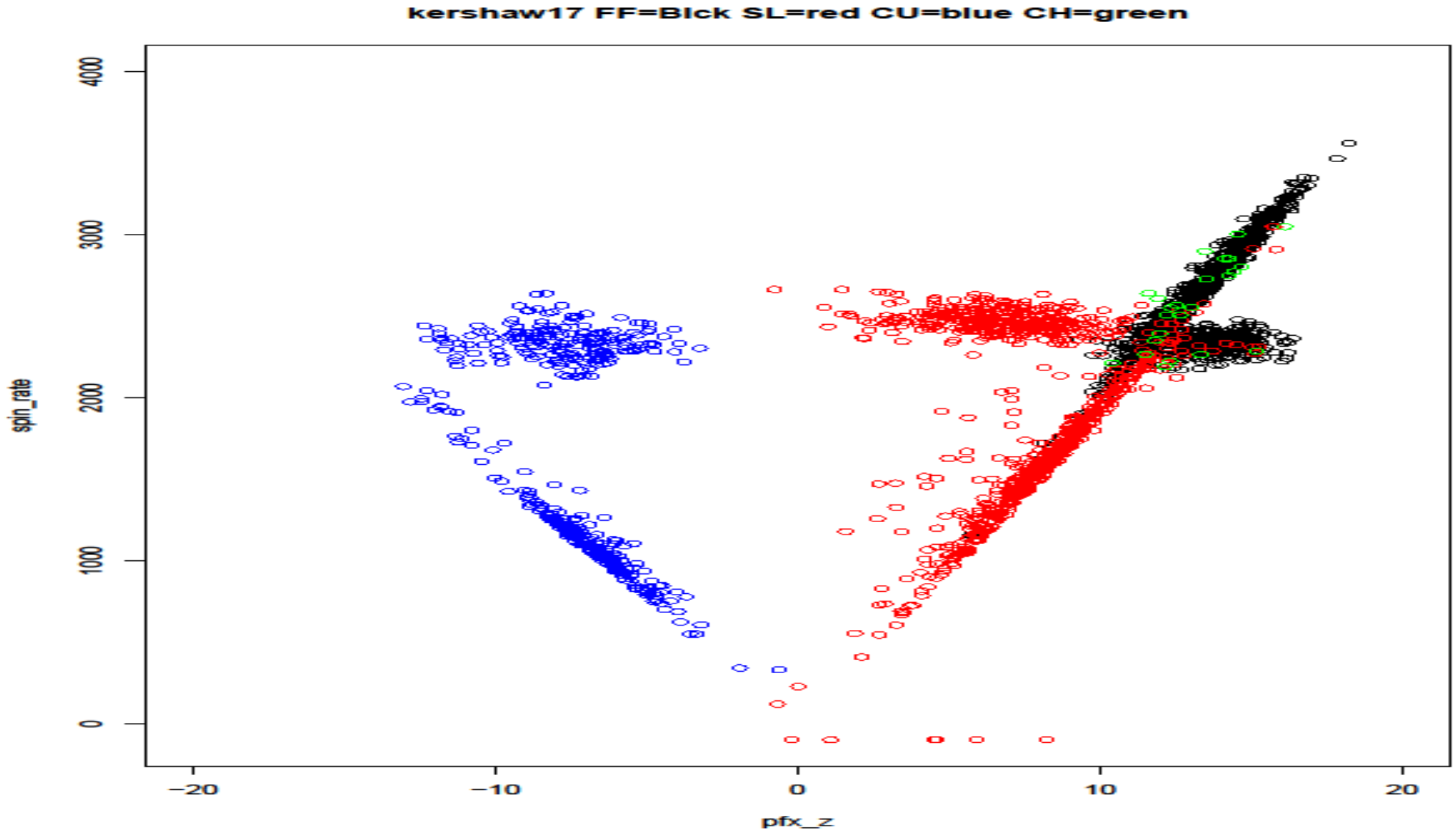
- ***How to see dependency ?***

- ***Via high-dimensional contingency "hypercube"*** ***upon*** **discrete, categorical or categorized features.**

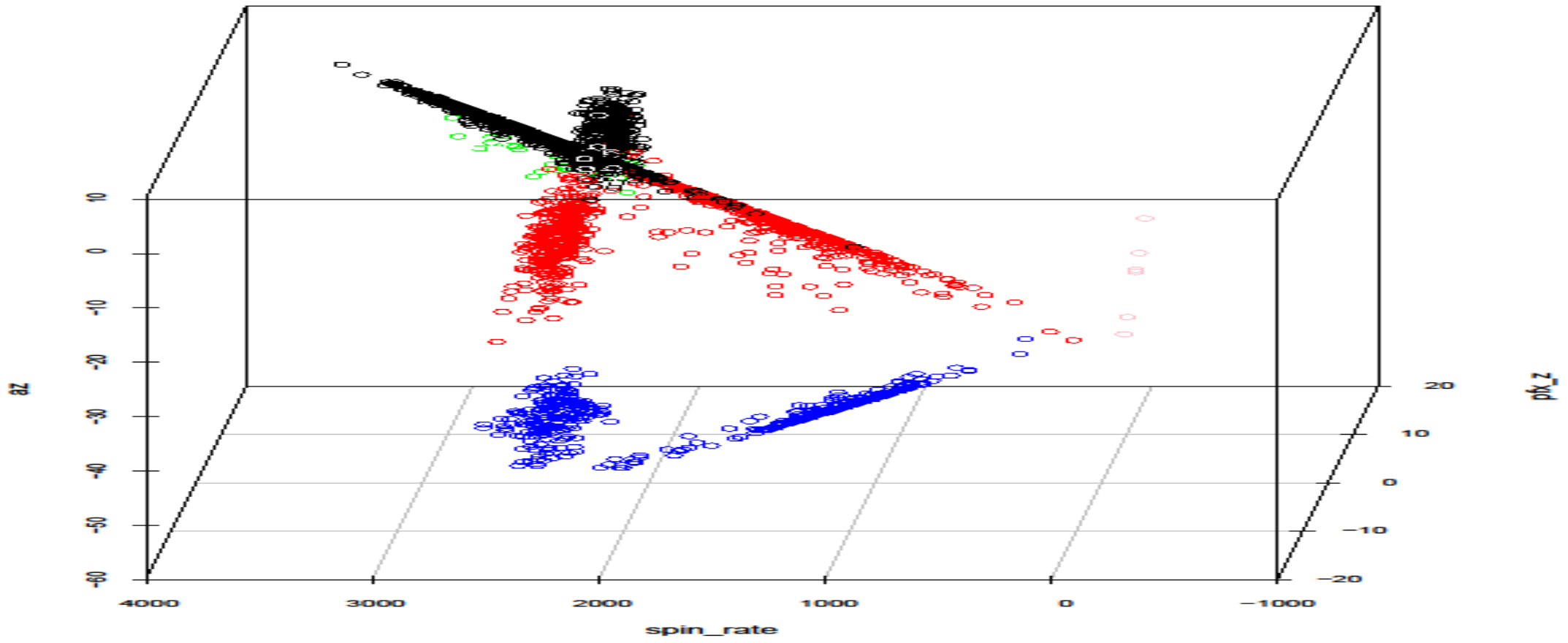
- ***How to evaluate dependency*** **(asymmetric Association)?**

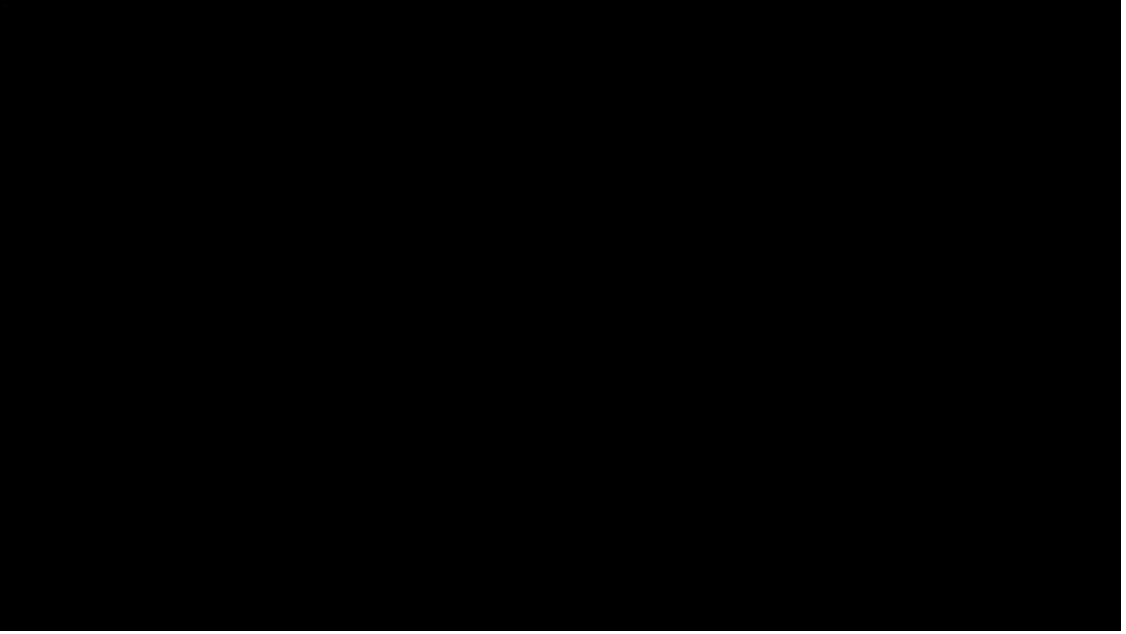
- via **mutual conditional entropy.[Categorization via possibly-gapped histogram**;

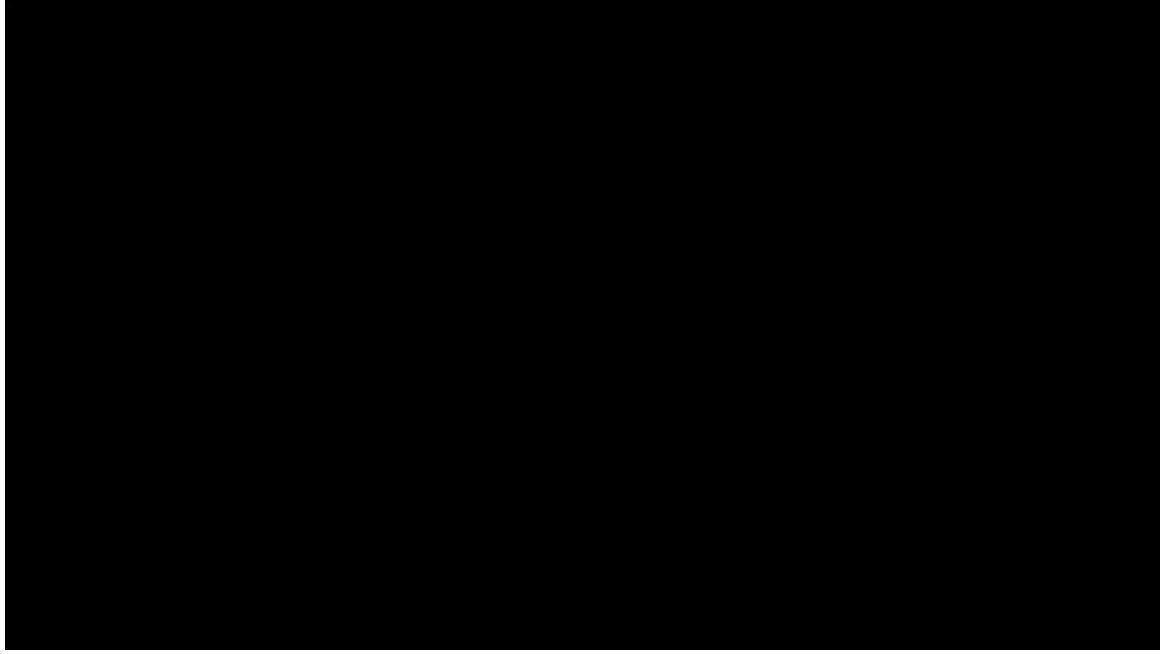
What kind of association measure for dependency?



Keep such non-linear relational patterns in mind!!!

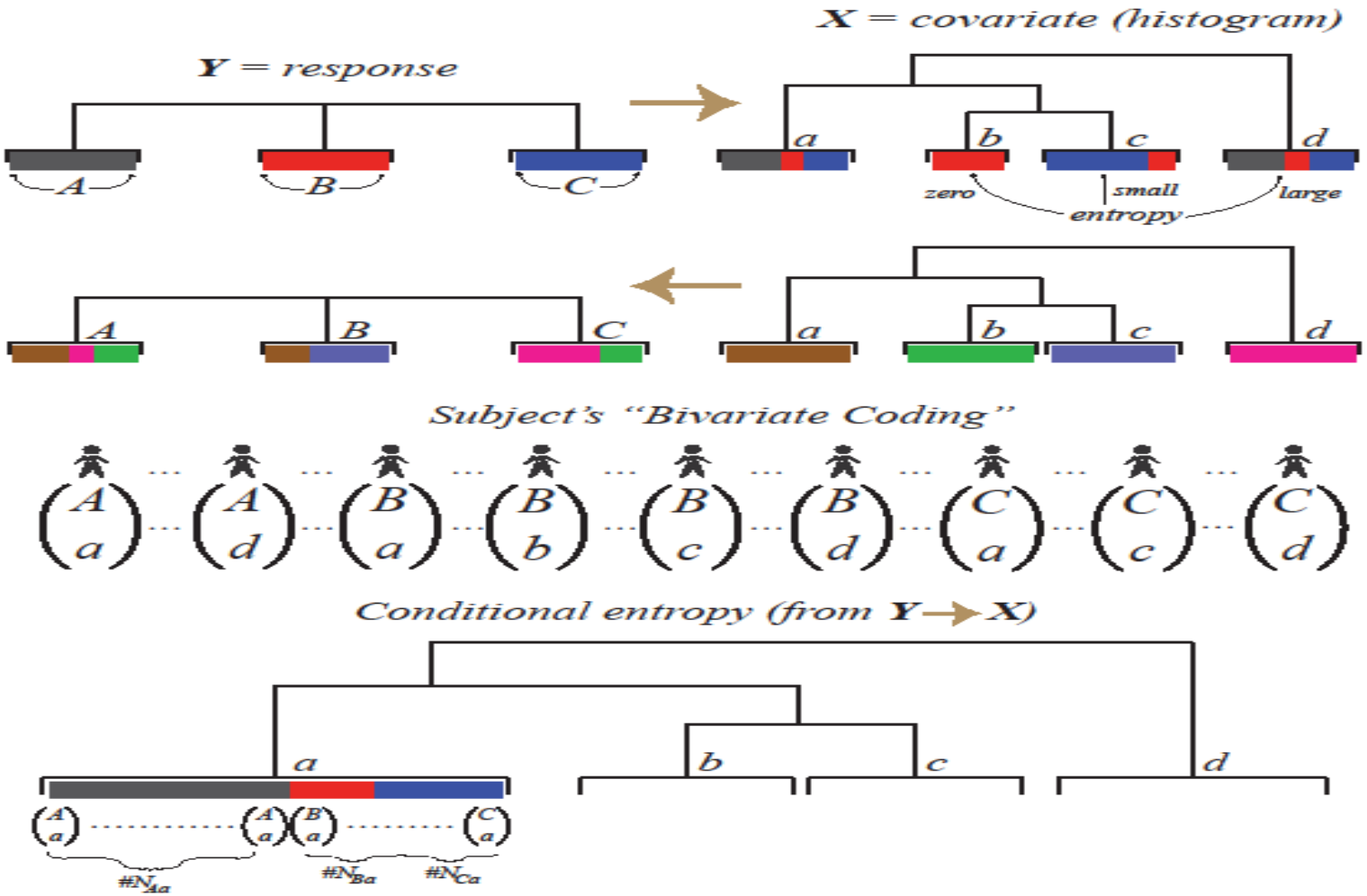






Three videos from the branch A : Break_length vs. Break-angle vs. Start_speed. (Spin-rate as diameter)

Relative mutual conditional entropy (Fushing et al. revised for PLoS One)



Mutual entropy computation

$$\begin{cases} P_{A|a} = \frac{N_{Aa}}{N_{Aa}+N_{Ba}+N_{Ca}} = P_r[A \mid a] = \{Probability\ of\ Y = A\ if\ X = a\} = P_r[A = Y \mid X = a] \\ P_{B|a} = P_r(Y = B \mid X = a) = \frac{N_{Ba}}{N_{.a}} \qquad N_{..} = N_{.a} + N_{.b} + N_{.c} + N_{.d} \\ P_{C|a} = P_r(Y = C \mid X = a) = \frac{N_{Ca}}{N_{.a}} \qquad W_a = \frac{N_{.a}}{N_{..}}, W_b = \frac{N_{.b}}{N_{..}}, \cdots \end{cases}$$

Conditional entropy given $X = a$

$$E_a^{(0)}(Y \to X) = (-1)\{P_{A|a} \cdot \log P_{A|a} + P_{B|a} \cdot \log P_{B|a} + P_{C|a} \cdot \log P_{C|a}\}$$

$$E_a^{(r)}(Y \to X) = \frac{E_a^{(0)}}{E^{(0)}[Y]} \qquad with\ E^{(0)}[Y] = entropy\ of\ Y$$

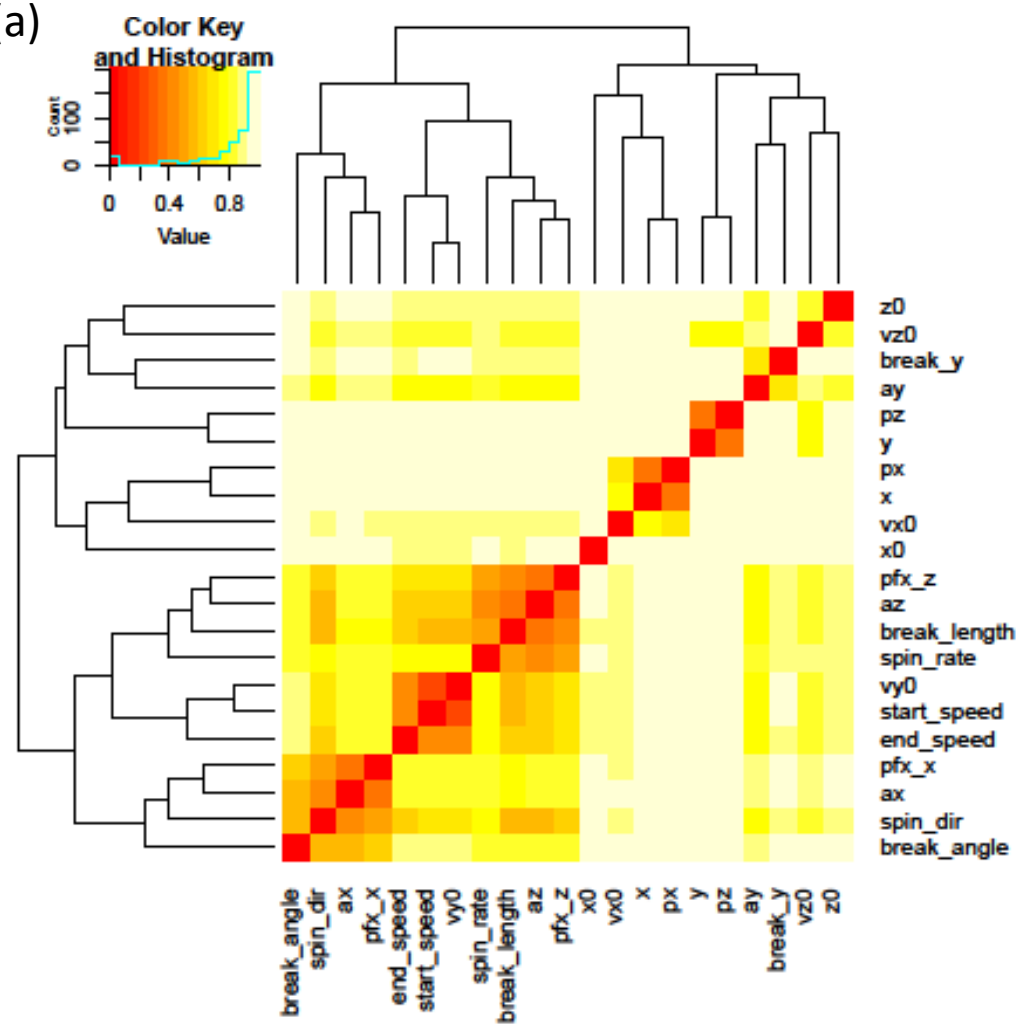
$$E^{(r)}[Y \to X] = W_a \cdot E_a^{(r)}(Y \to X) + W_b \cdot E_b^{(r)}(Y \to X) + W_c \cdot E_c^{(r)}(Y \to X) + W_d \cdot E_d^{(r)}(Y \to X)$$

Mutual (conditional) entropy of Y *and* X

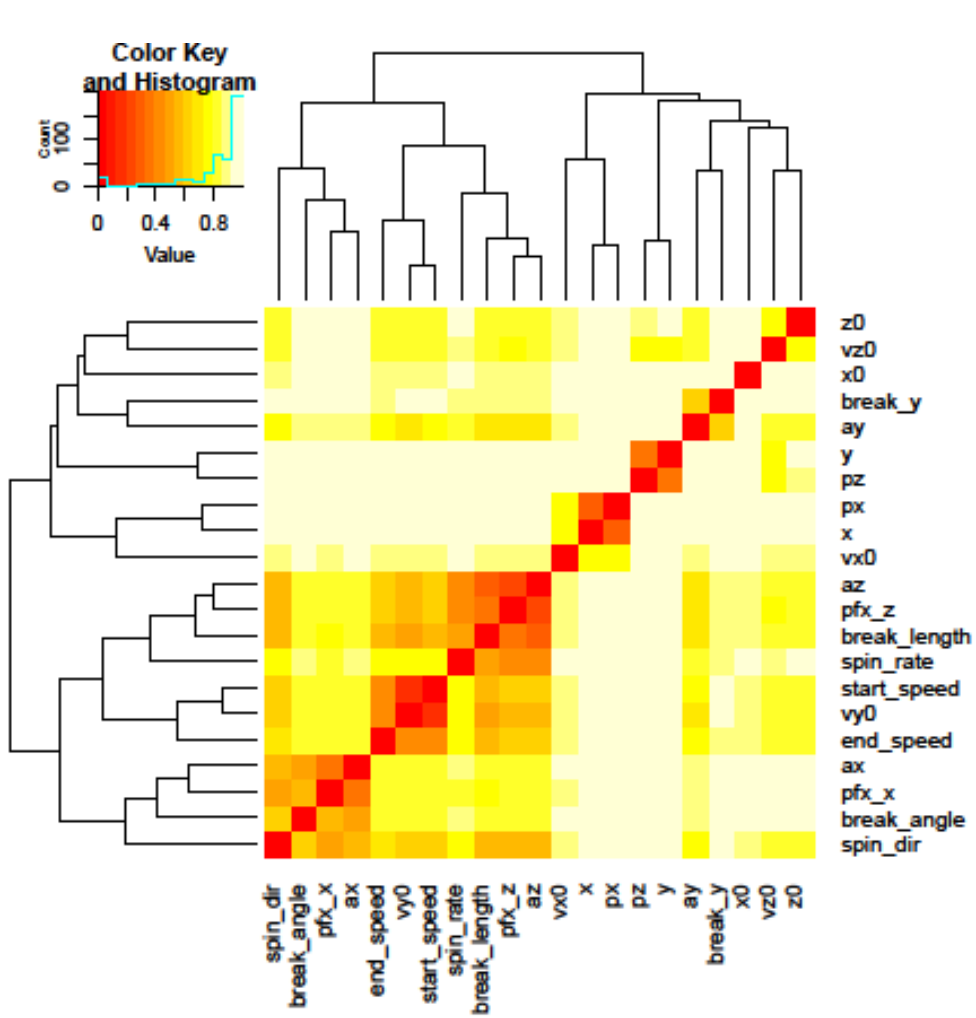
$$E[Y \Longleftrightarrow X] = \frac{(E^{(r)}[Y \to X] + E^{(r)}[X \to Y])}{2}$$

Mutual entropy matrix (Kershaw 13(a) &14(b))

(a)

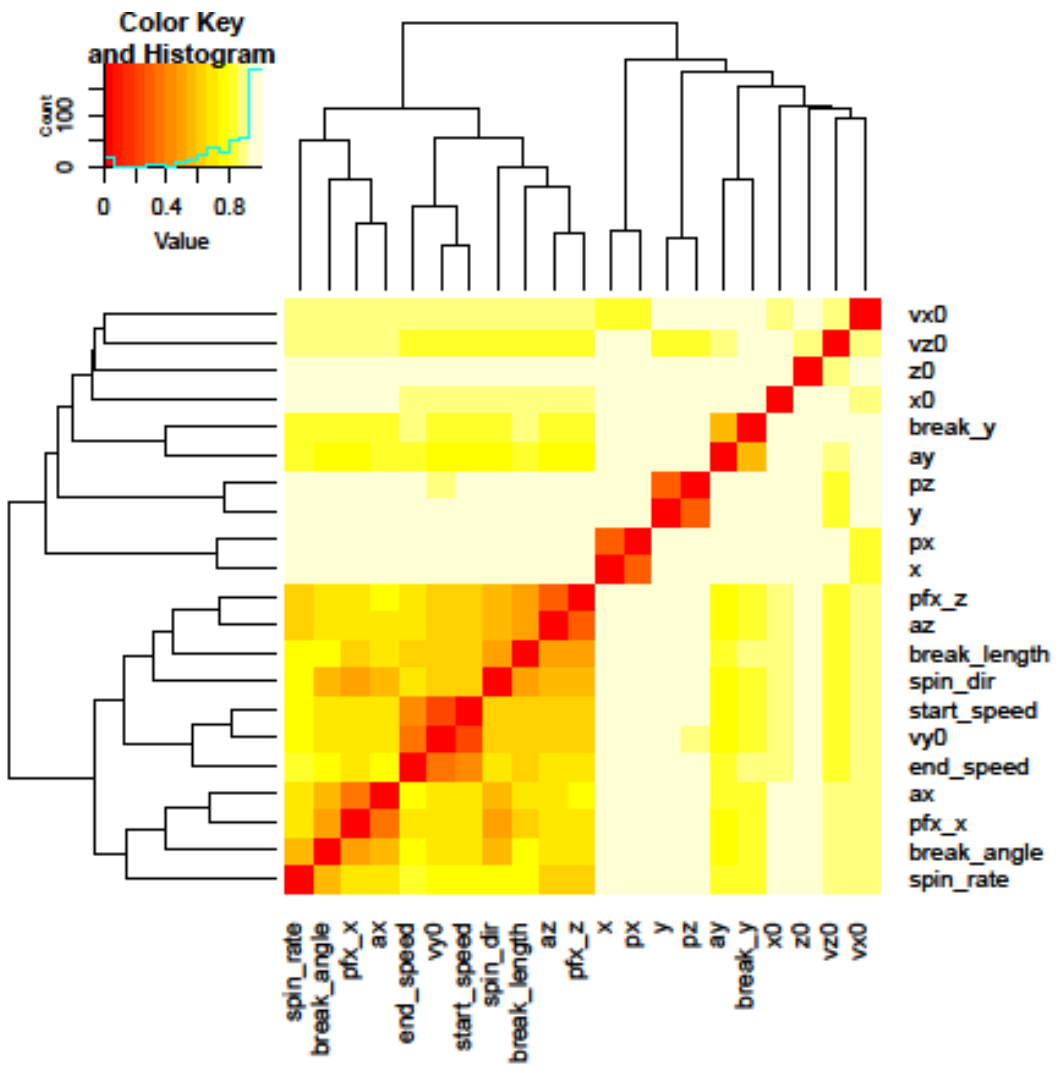


(b)

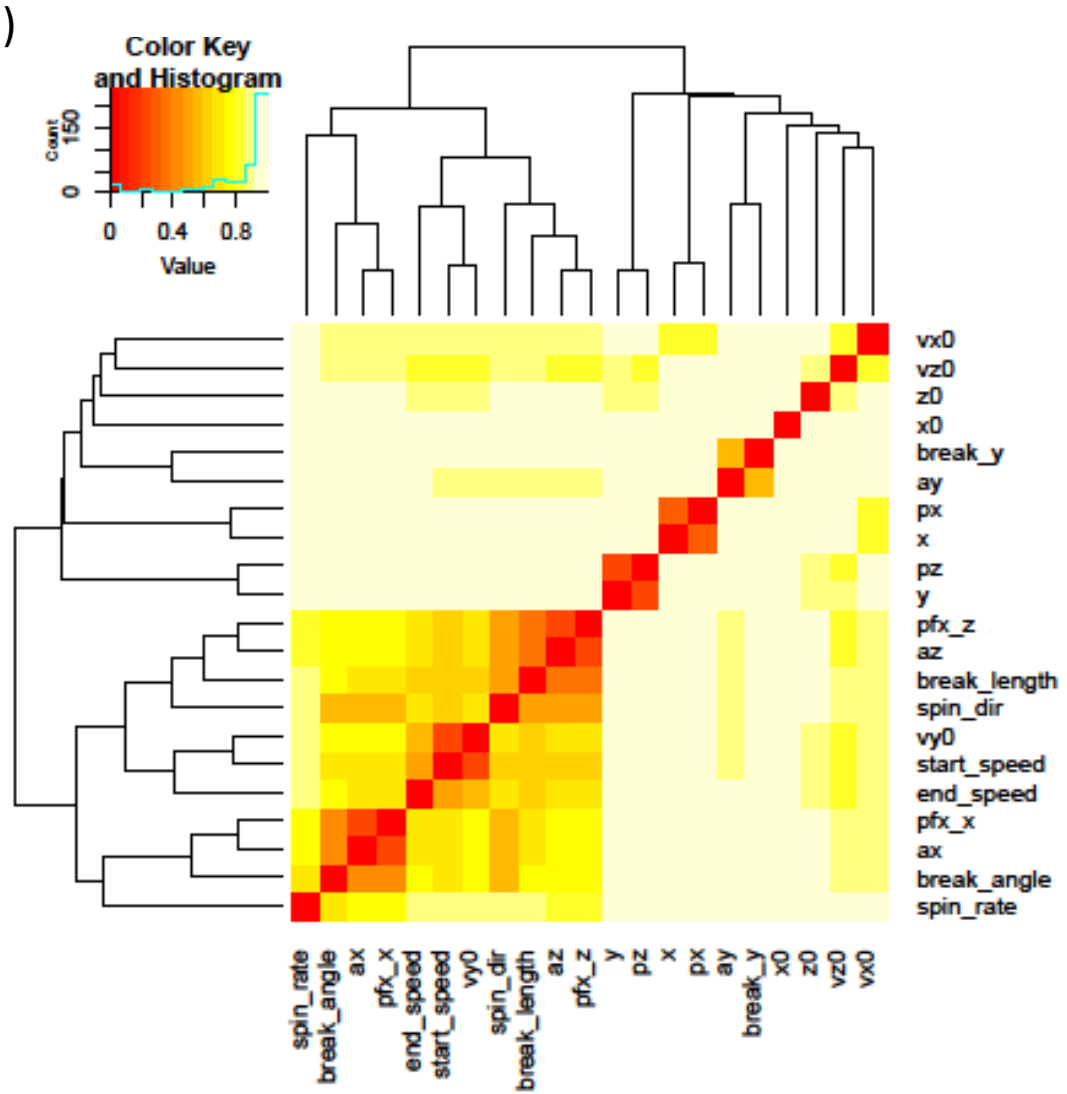


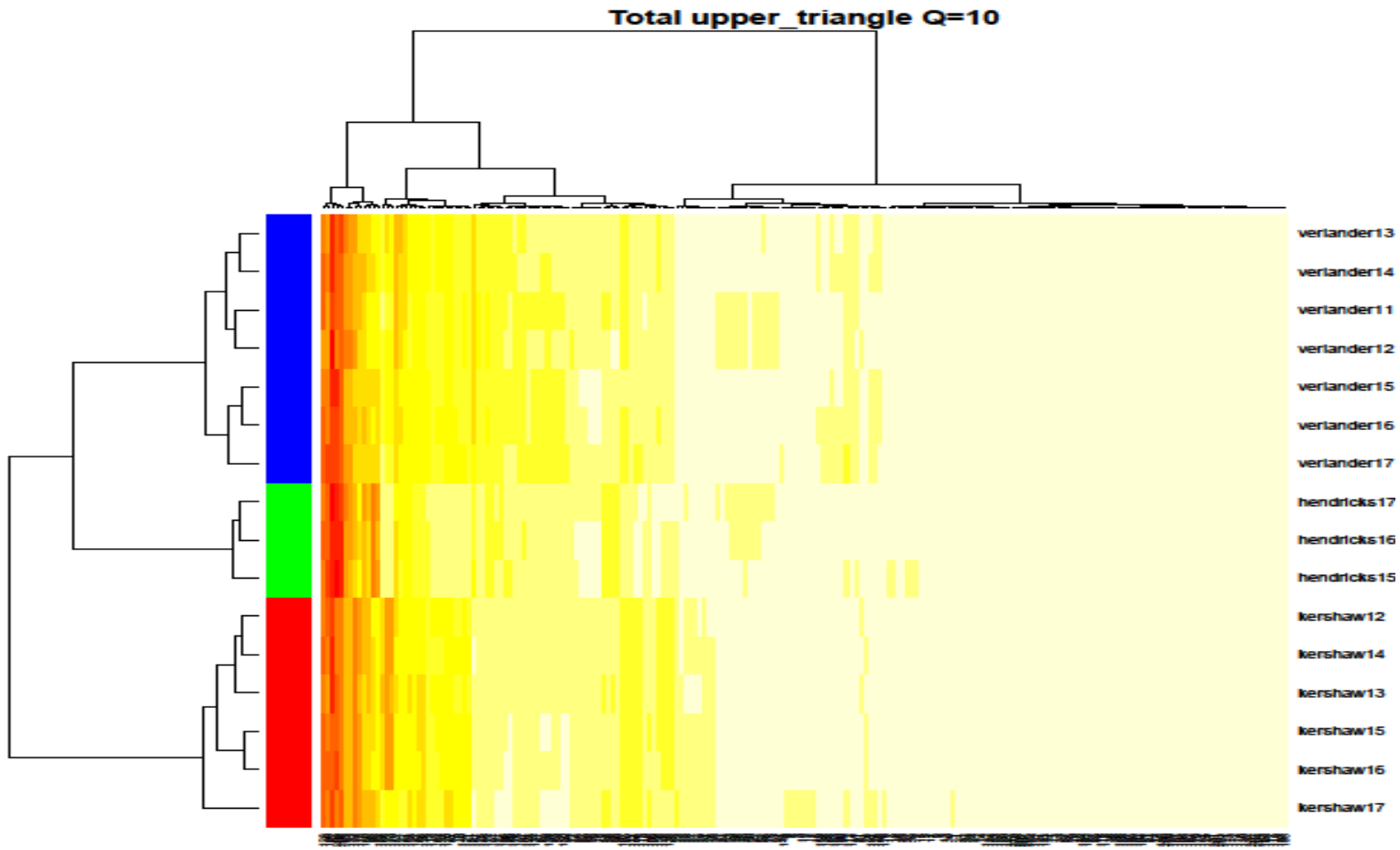
Mutual entropy matrix (Verlander 16(a) and Hendricks16(b)

(a)

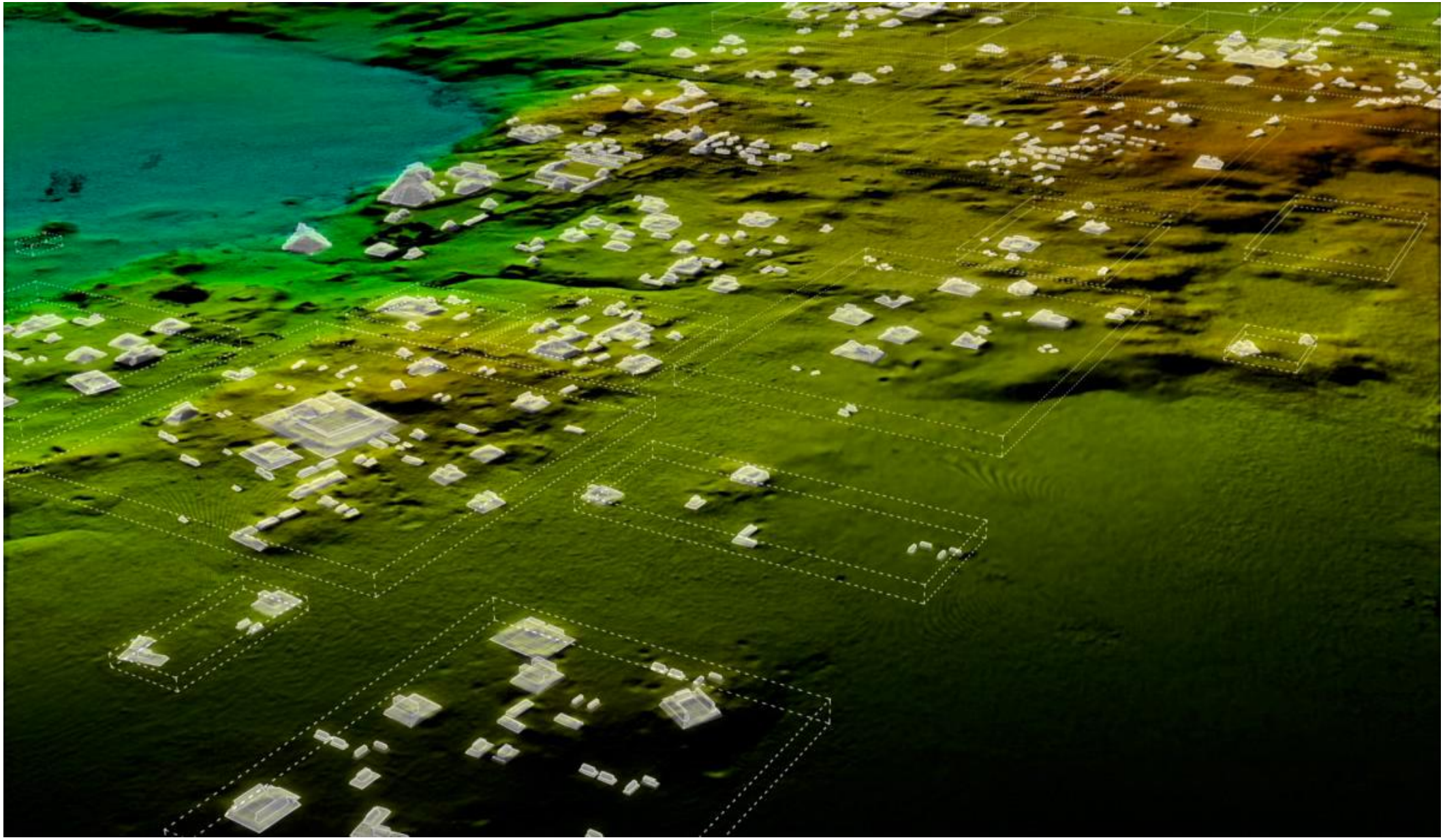


(b)

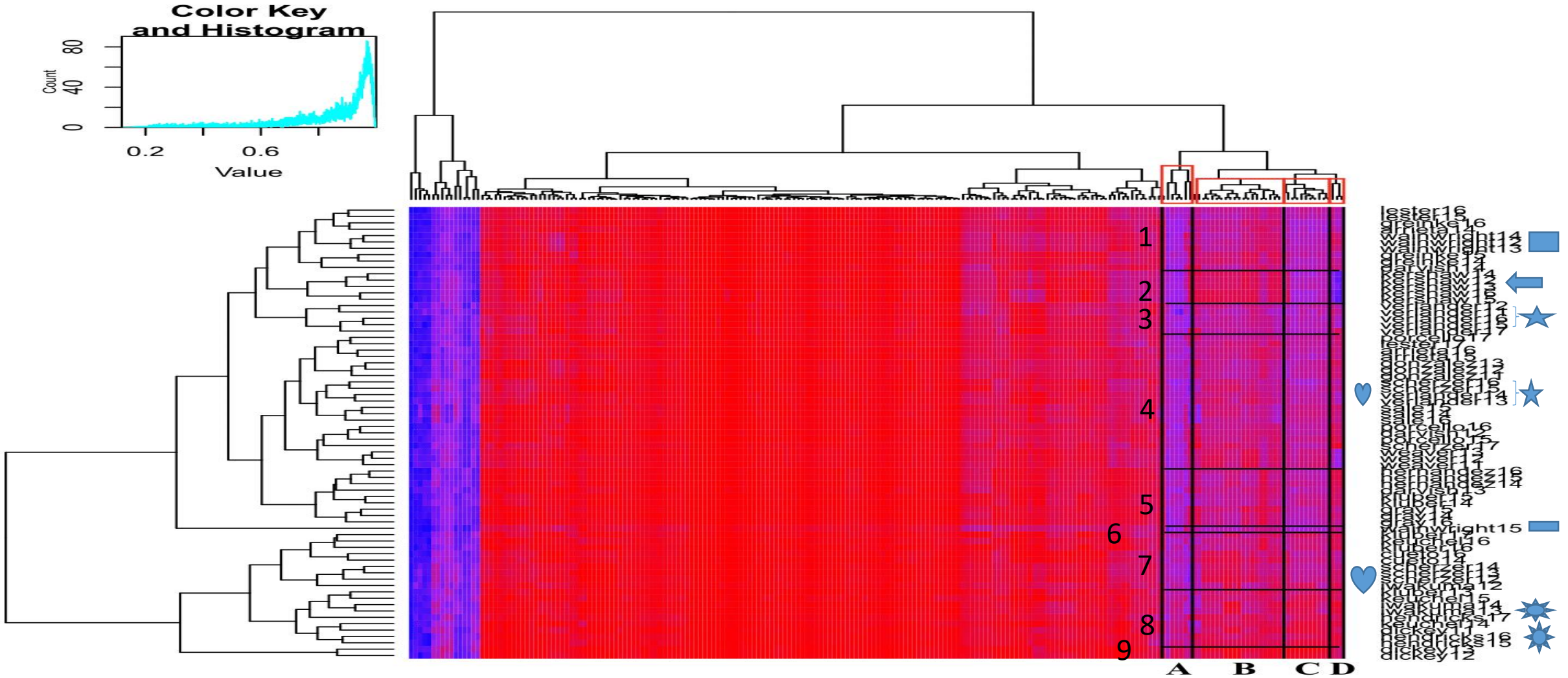



Total upper_triangle Q=10
verlander13
verlander14
verlander11
verlander12
verlander15
verlander16
verlander17
hendricks17
hendricks16
hendricks15
kershaw12
kershaw14
kershaw13
kershaw15
kershaw16
kershaw17

Laser Scans Reveal Maya "Megalopolis" Below Guatemalan Jungle



Universal patterns (embracing law of aerodynamics) in mutual entropy matrices



Systemic characteristics of pitching mechanics

What we see through the heatmap……..

- Universality is confirmed on the very left branch (deep blue color).
- A tree of clusters of pitchers (on row axis).
- Block characteristics (on column axis) of pitching mechanics of each cluster of pitchers.
- What factors (or group of features) to choose in order to effectively compare which groups of pitchers.

Detailed comparison among the three MLB pitchers:

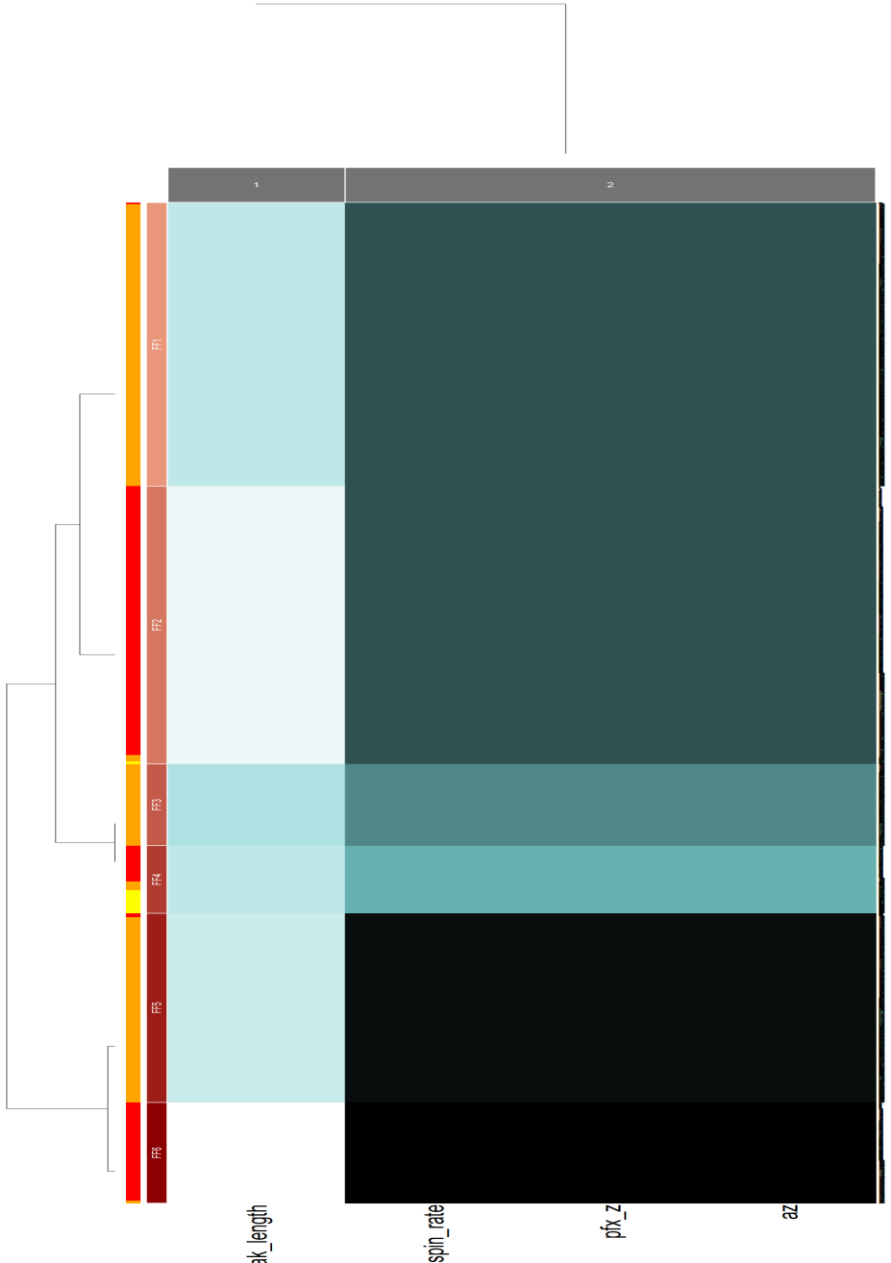
- **Clayton Kershaw** (two time Cy Young award) L.A. Dodgers
- **Justin Verlander** (key pitcher in 2017 World Series) Huston Astros.
- **Kyle Hendricks** (key pitcher in 2016 World Series) Chicago Cubs.

How to compare the 4D geometries of (Break-length, Spin-rate, Pfx_z, a_z)?

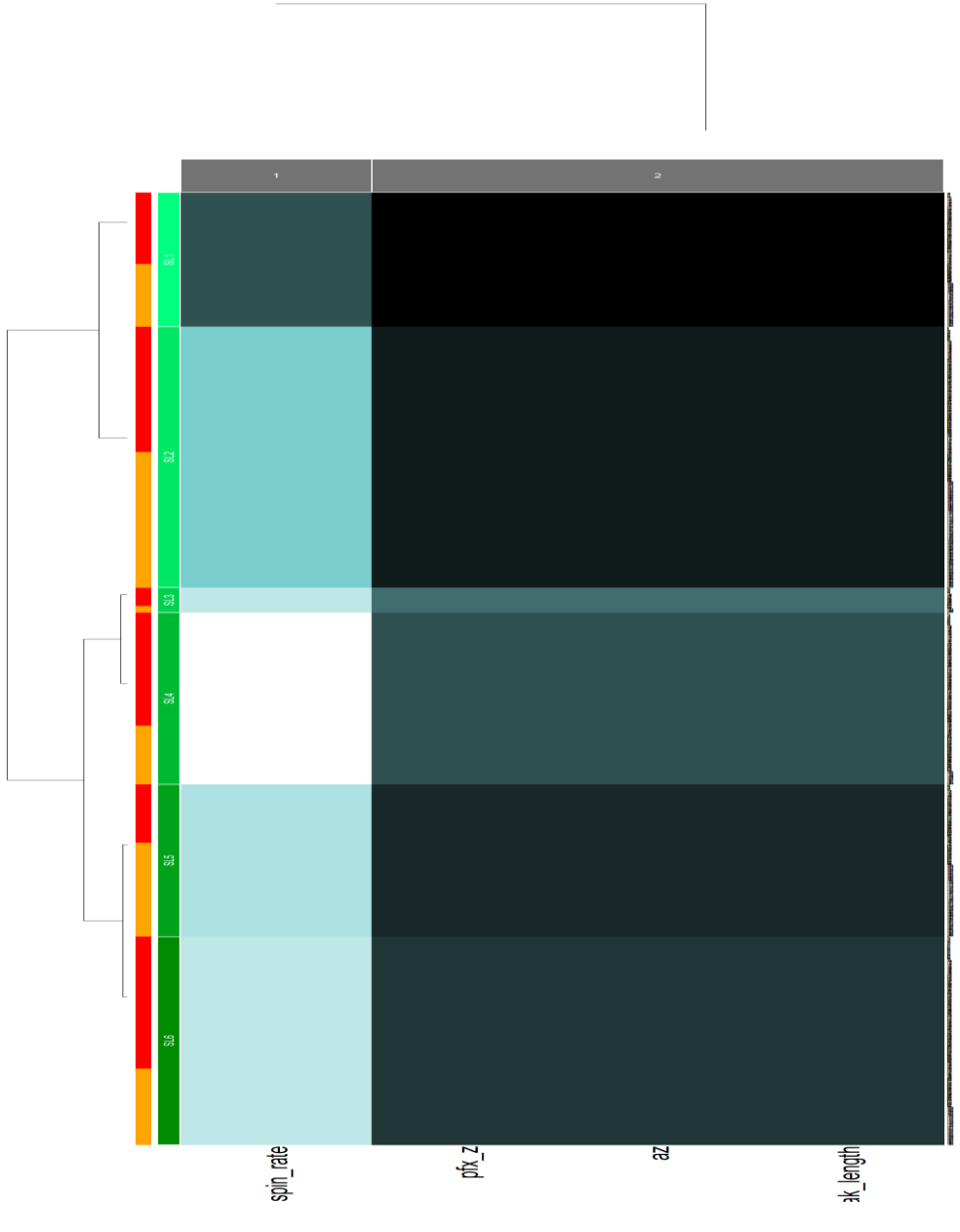
- Pitch-type specific "categorical-pattern-distribution" (with a chosen baseline)
- Step-1: For fastball (FF) and taking Kershaw as the baseline, 6 clusters are computed upon all fastball pitches from the tree pitchers;
- Step-2: Compute the 6 proportions of Kershaw's fastball with respective the 6 clusters.

Kershaw-red, Verlander-orange, Hendricks-Yellow (Break-length, pfx_z, az and spin-rate)

(FF)

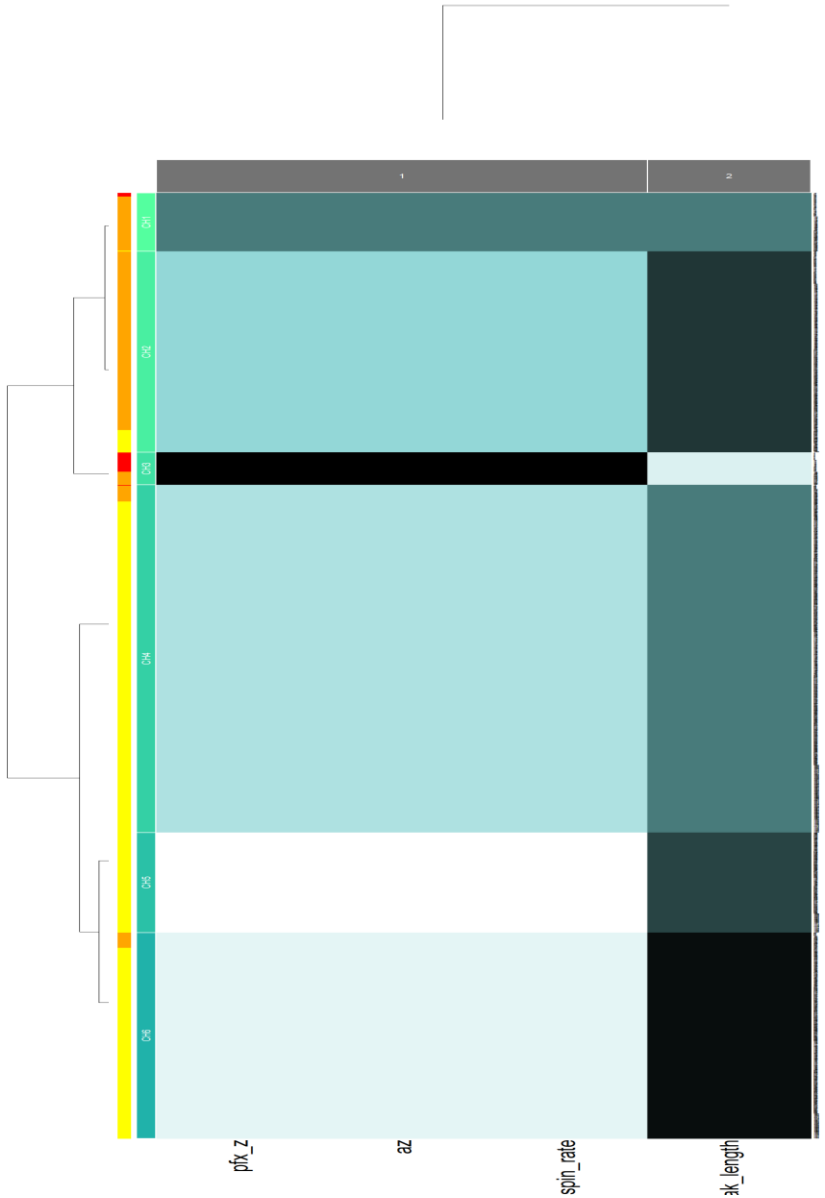


(SL)

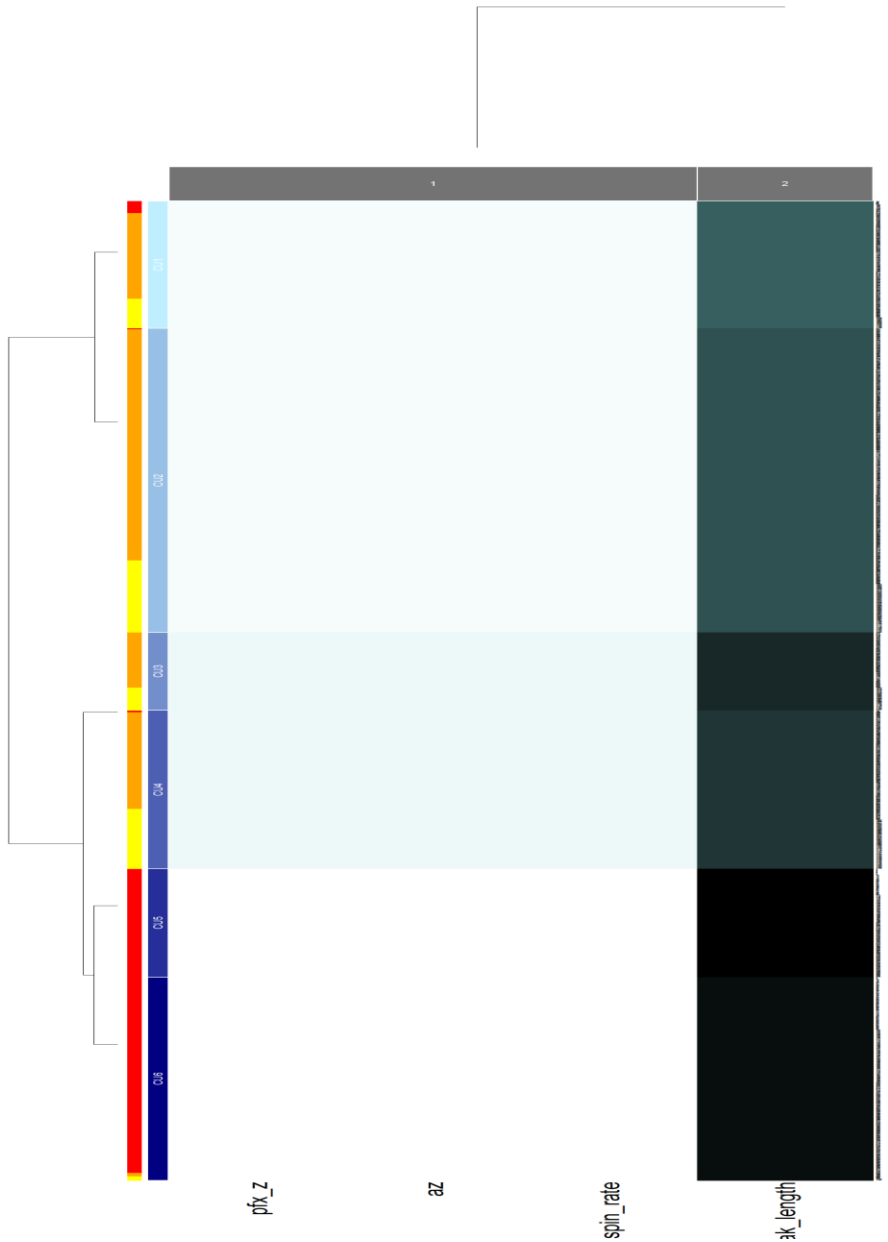


Kershaw-red, Verlander-orange, Hendricks-Yellow (Break-length, pfx_z, az and spin-rate)

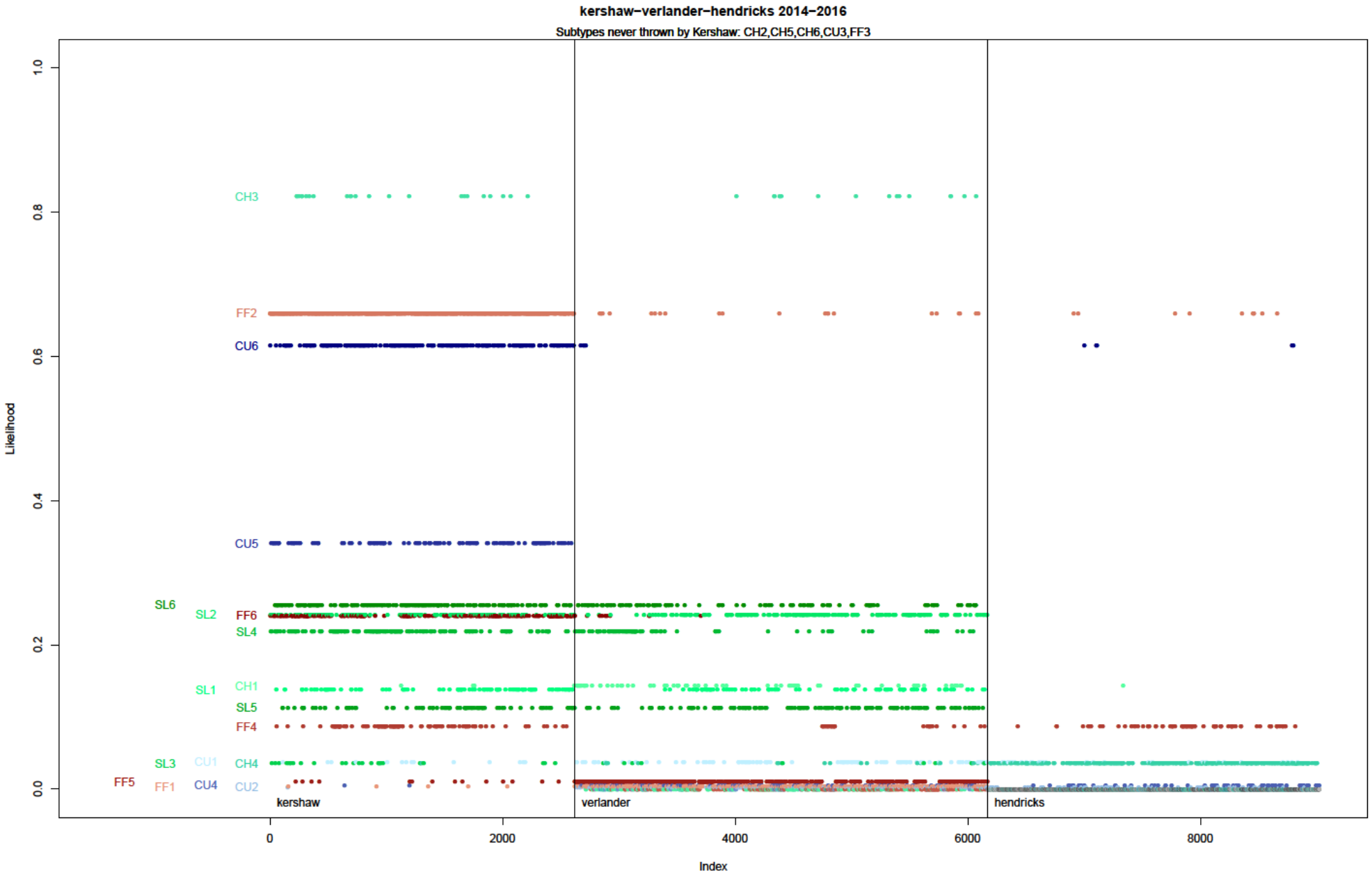
(CH)



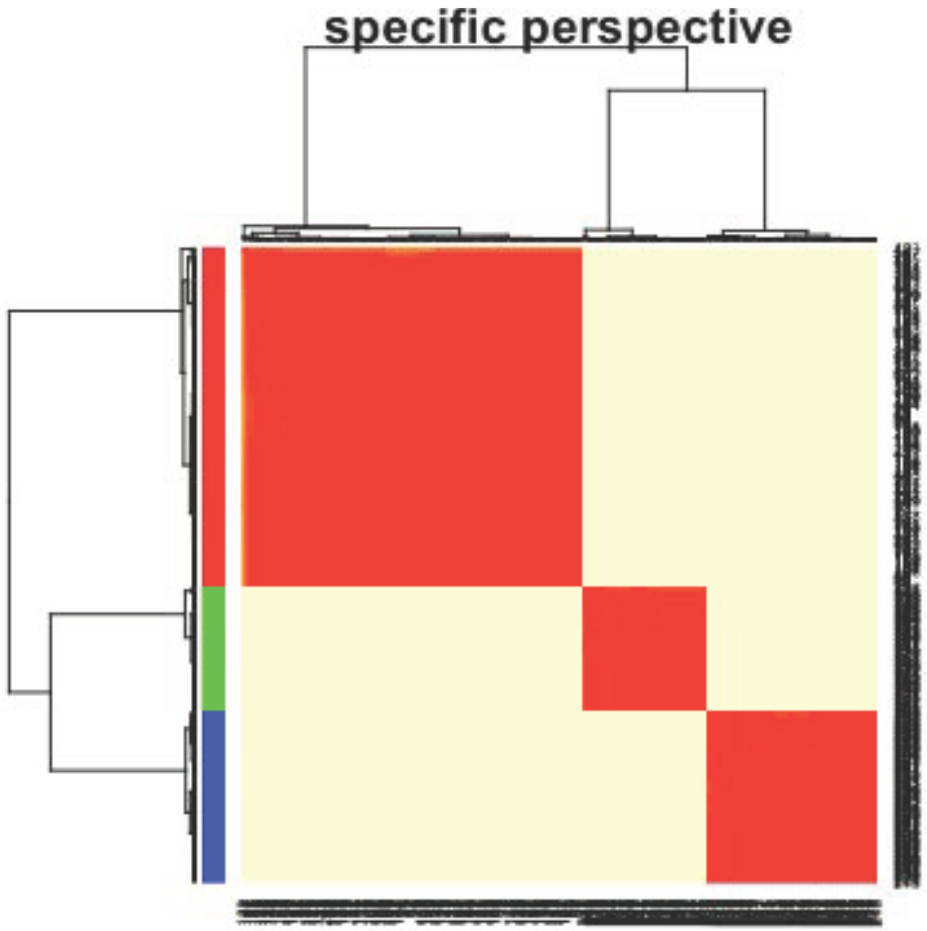
(CU)



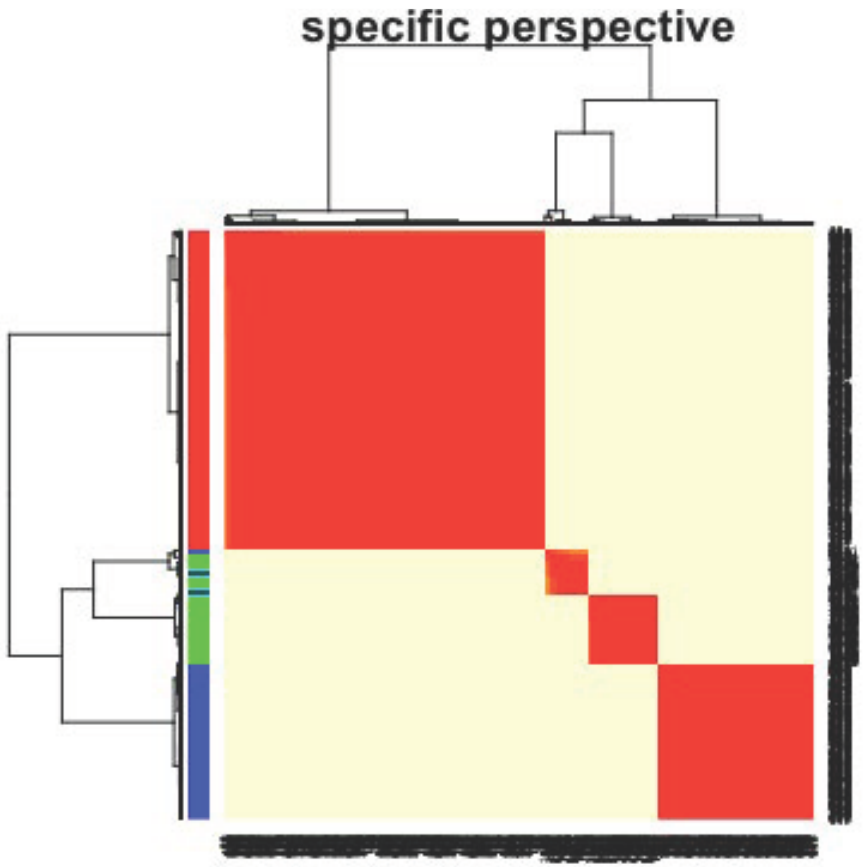
Combined pitch-type specific likelihood time series



CDI on FF & CU (ax(no.6); spin-dir(no.11); break_angle (no.14); pfx_x(no.16))

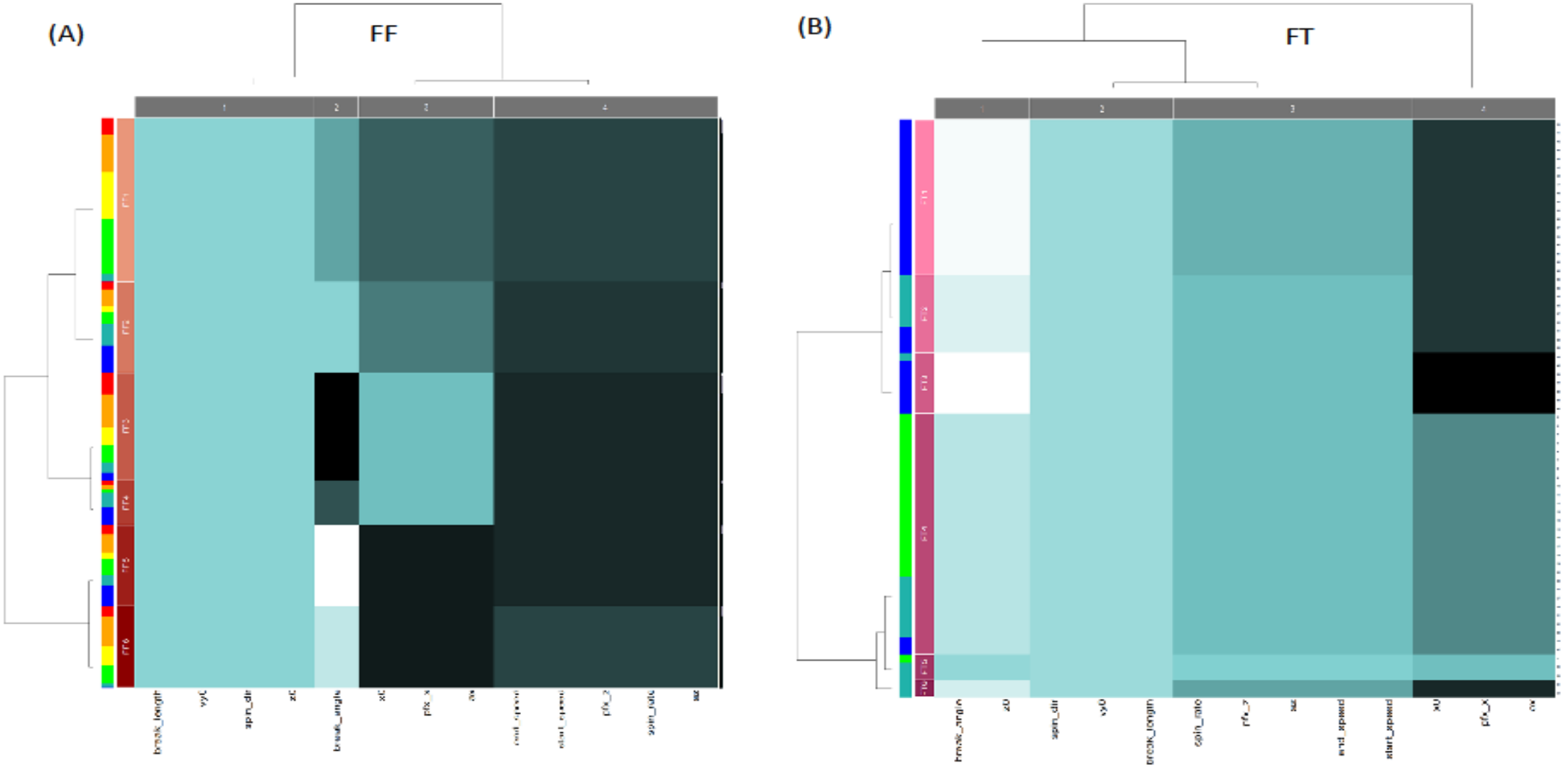


Result: a distance based on single mechanism: {6, 11,16} features. It is perfect on the entire data set as required by CDI.

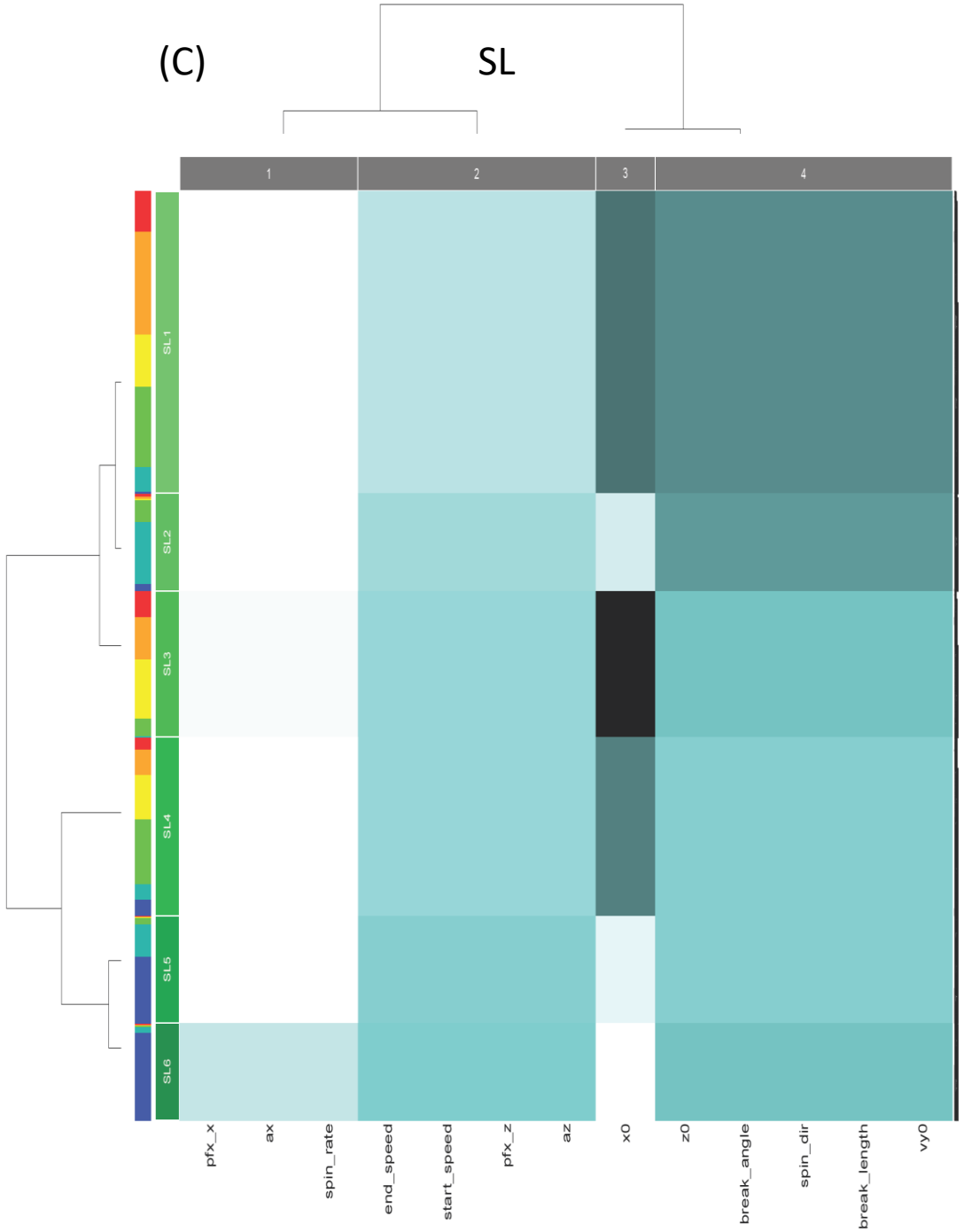


Result: a distance based on single mechanism: {6, 11, 14,16} features. It is not perfect even on the small data set. More explorations are needed for achieving CDI.

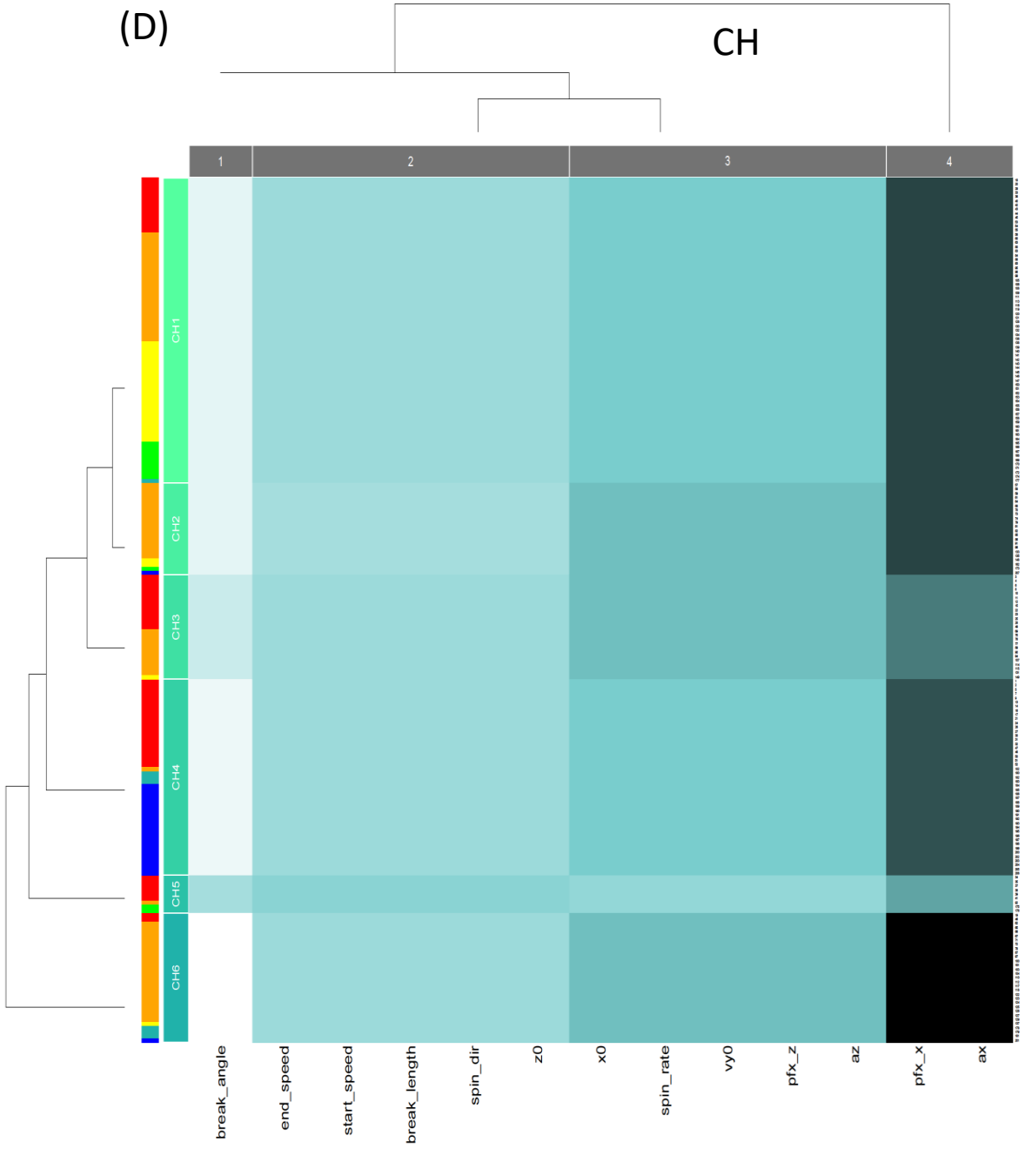
Examples of categorical pattern distributions (Clayton Kershaw: 2012-2017 (categories) and 6 clusters)



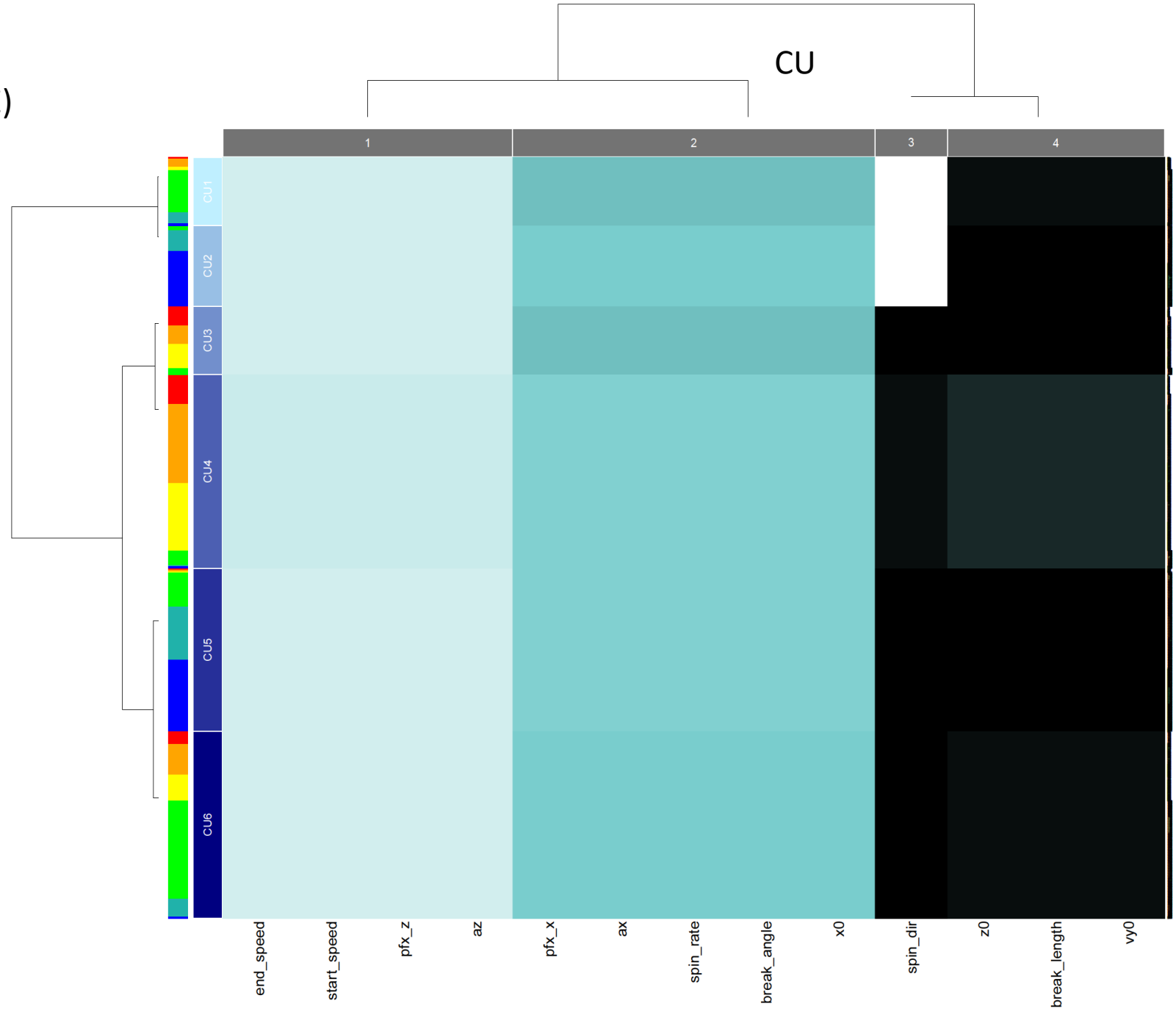
(C)


SL
1
2
3
4
SL1
SL2
SL3
SL4
SL5
SL6
pfx_x
ax
spin_rate
end_speed
start_speed
pfx_z
az
x0
z0
break_angle
spin_dir
break_length
vy0

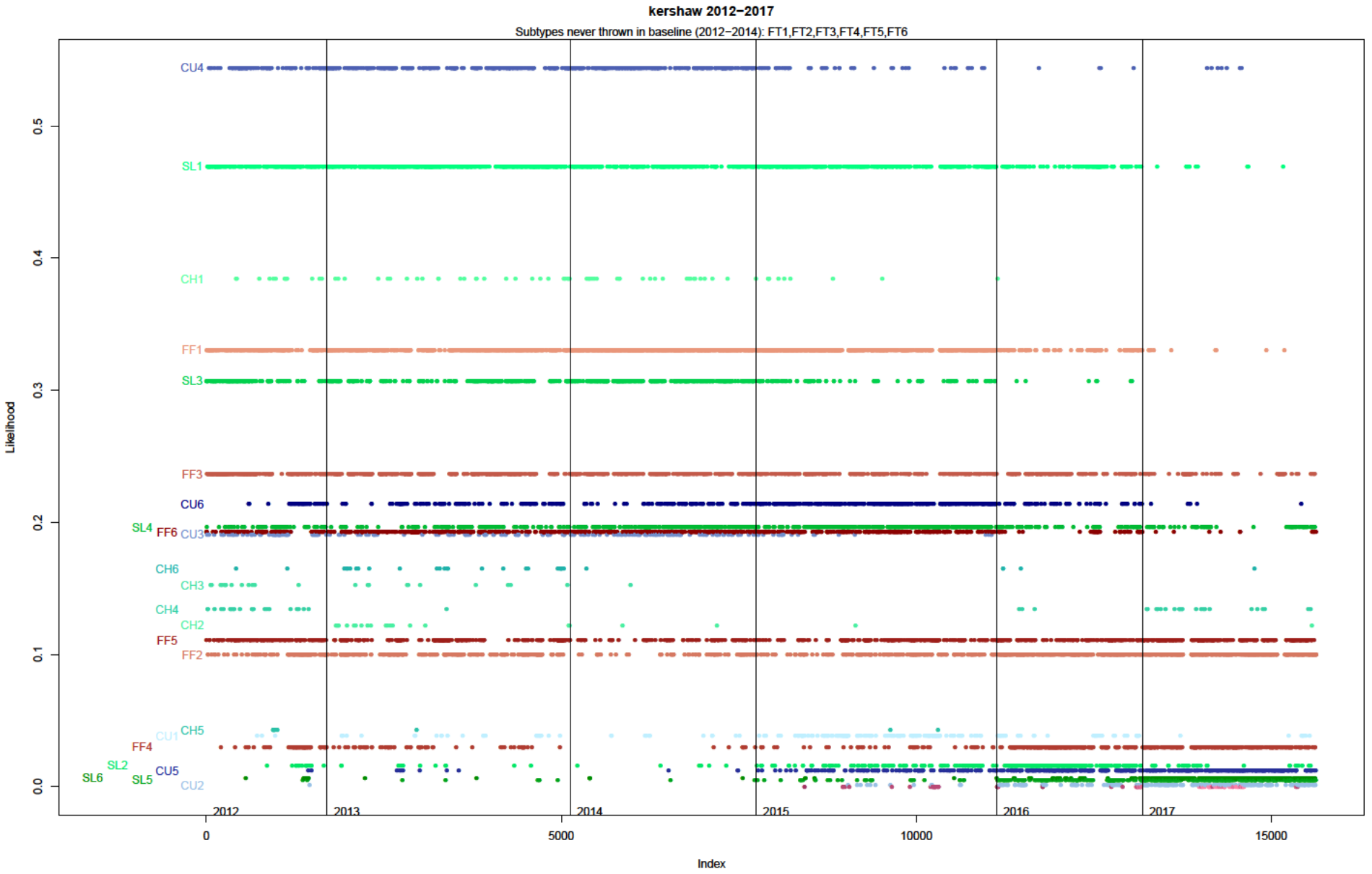
(D)


CH
1
2
3
4
CH1
CH2
CH3
CH4
CH5
CH6
break_angle
end_speed
start_speed
break_length
spin_dir
z0
x0
spin_rate
vy0
pfx_z
az
pfx_x
ax

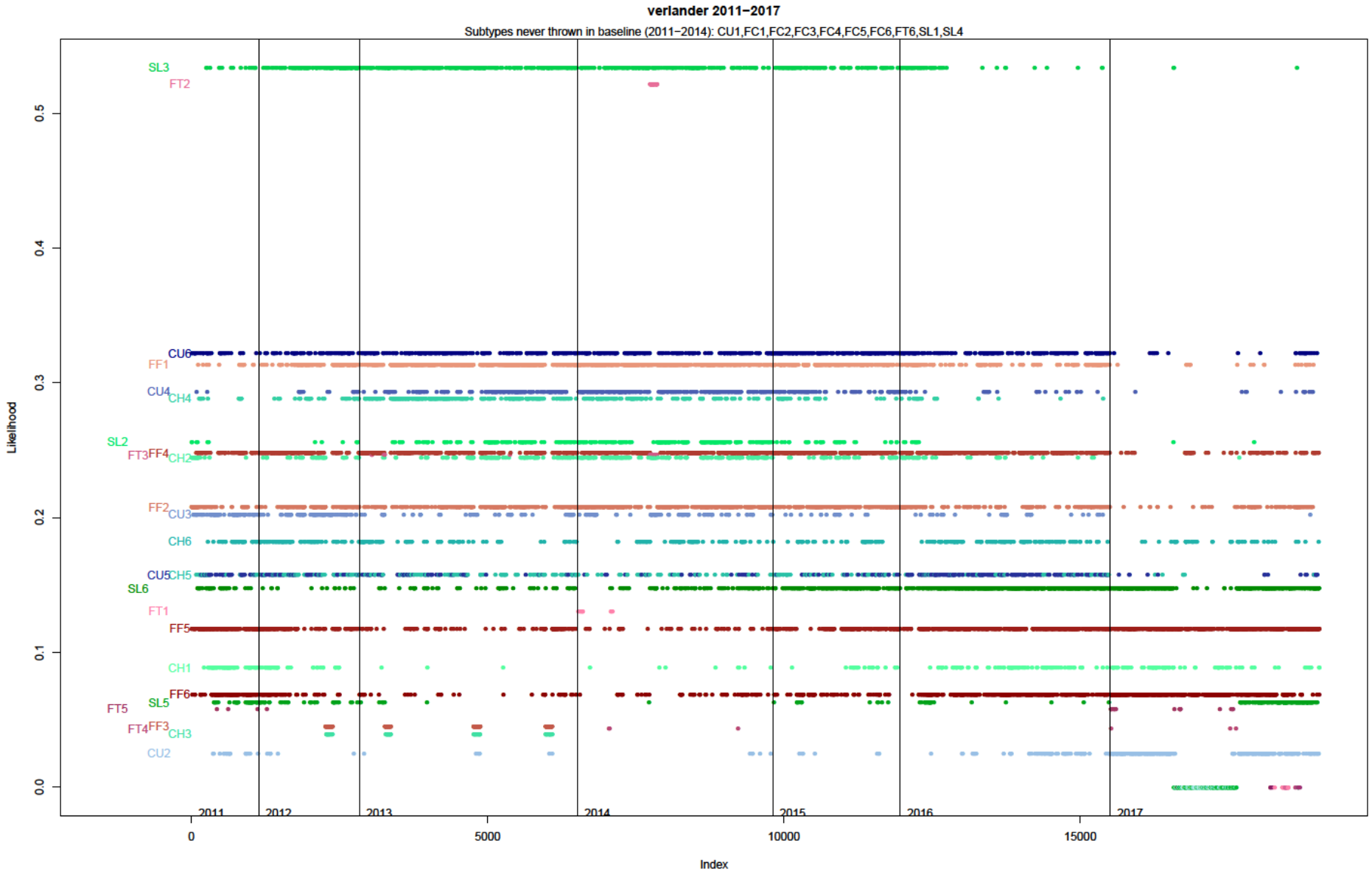
(E)


CU
1
2
3
4
CU1
CU2
CU3
CU4
CU5
CU6
end_speed
start_speed
pfx_z
az
pfx_x
ax
spin_rate
break_angle
x0
spin_dir
z0
break_length
vy0

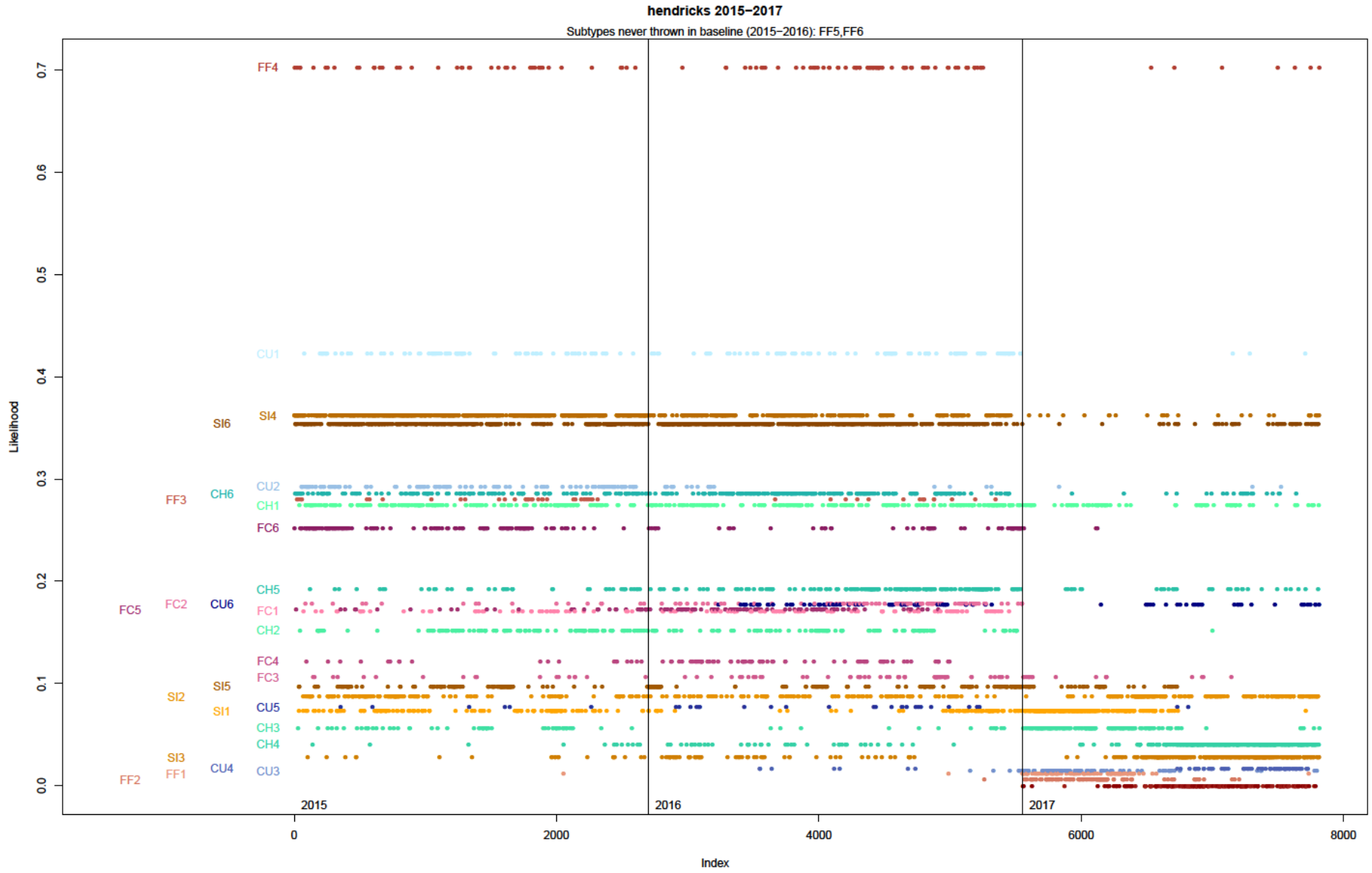
Fine-scale comparisons of Kershaw's pitching mechanics cross 2012-2017 seasons (11D).



Fine-scale comparisons of Verlander's pitching mechanics cross 2011-2017 seasons.



Fine-scale comparisons of Hendricks' pitching mechanics cross 2015-2017 seasons.



Key *merits:*

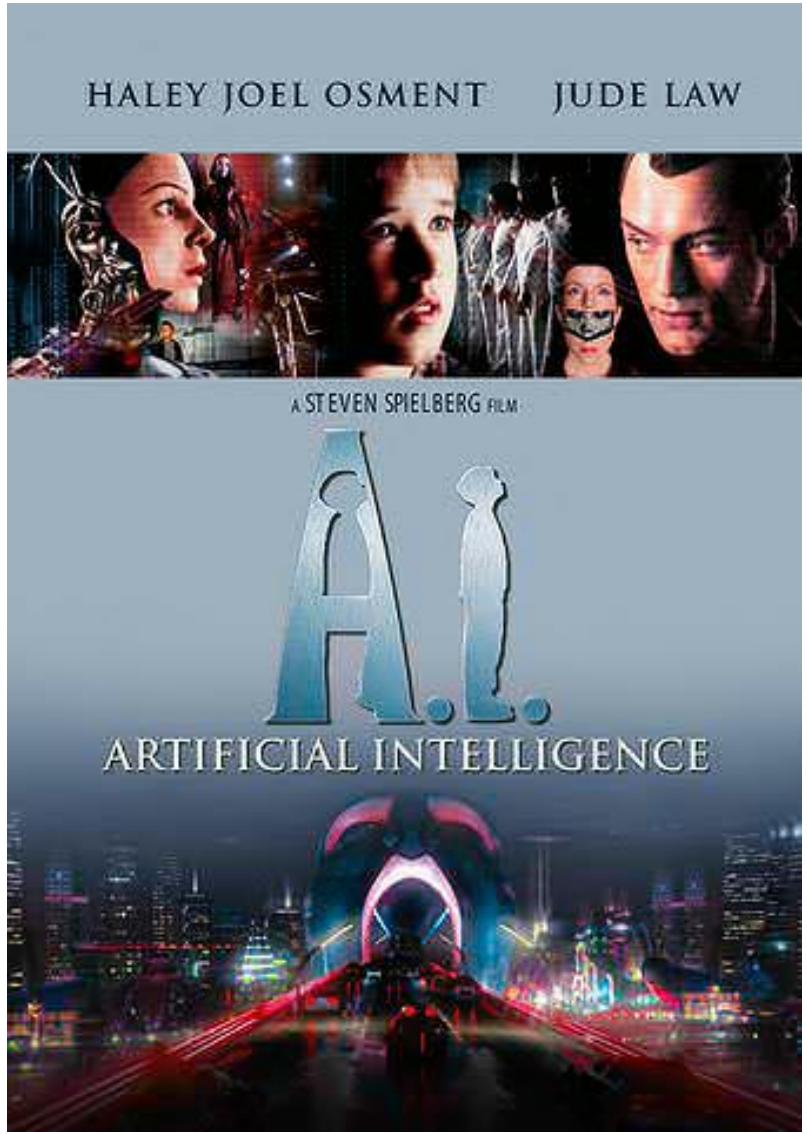
- A **platform** where ***everyone*** can explore information contained in PITCHf/x database;

- For **everyone** to seek real **understanding** and to build true ***knowledge*** on pitching mechanics

- For a pitcher to perform **self-evaluation**;

- For a coach or manager to make **decision**.

What do I really want? <ans> Dr. Know



Youtube

https://www.youtube.com/watch?v=mzsqulKTwO0

What do I really *want*?

- To build an "AI" as an **Open access** platform:
  - **All people can ask MLB pitching related questions;**

and received **informative answers that** embrace **data-driven intelligence**.

- Check out: **dm-mlbpitching.com**