Quantifying Heterogeneous Returns to Genetic Selection:

Evidence from Wisconsin Dairies

Brent Hueth, Jared Hutchins, and Guilherme Rosa

Economics of Research and Innovation in Agriculture - NBER

Genetic Technology in the Dairy Sector

- Biological innovation is a key driver in agricultural productivity (Olmstead and Rhode 2008).
- A largely unstudied sector for biological innovation (by economists) is the dairy sector, which grew because of two innovations:
  - Artificial insemination (AI).
  - Herd testing and genetic evaluations.
- Approach has been less “top down” than other biological innovations, but has delivered large genetic gains in yields.

The Role of Heterogeneous Returns

- Issue: data used for estimating mean returns is **observational** data, so the environment can confound estimates.
  - Animal science: genotypes *interact* with environments (Hayes et. al. 2003, Kearney et. al. 2014).
  - Economics: farmers adopt technology *into* environments based on their expected returns (Grilliches 1957, Suri 2011)
- How might this selection behavior be driving heterogeneity in returns or vice versa?

Research Question

RQ: Are returns to different dairy genetics heterogeneous across producer characteristics?

- Applying a theoretical framework to sire selection.
- Handling herd data to define “choices.”
- Getting accurate counter-factual predictions.

Outline

1. Institutional Framework
2. Theoretical Framework
3. Data Description
4. Empirical Exercises
5. Future Directions

How to Price Genetics for Dairy Bulls

Given one dairy bull:

1. DHIA collects data on all of its offspring.
2. CDCB uses data to evaluate that bull on multiple traits relative to base.
3. Evaluation of bull on multiple traits is published.
4. Use evaluation to price the bull on the market.

An Example Bull

Only God can judge him

TUPAC {4}

$25

1JE00919 JX FARIA BROTHERS TUPAC {4} — 07/04/14 | 840 Reg. 3124526365

- Tank topping production
- Unmatched CFP

CDCB PTA, AJCA PTA, GENEX 8/2017

Net Merit	+$522	69%Rel	**PTAT**	+0.60	74%Rel
Cheese Merit	+$528		**JUI™**	+2.9	
Fluid Merit	+$510		**JPI™**	+152	
Daus. G	**Herds** G		**Fertility (SCR)**	0.0	90%Rel
Milk	+1959	74%Rel	**PregCheck™**	99	89%Rel
Protein	+61	-0.04%	**HCR**	0.5	50%Rel
Fat	+75	-0.08%	**CCR**	-1.5	61%Rel
CFP	+136		**Dtr. Pregnancy Rate**	-3.3	62%Rel
Prod. Life	+3.0		**EFI%**	8.4%	
SCS	+2.93				
LIV	-0.7	41%Rel			
GL	+1.3	57%Rel			

HARRIS X RENEGADE X VIBRANT

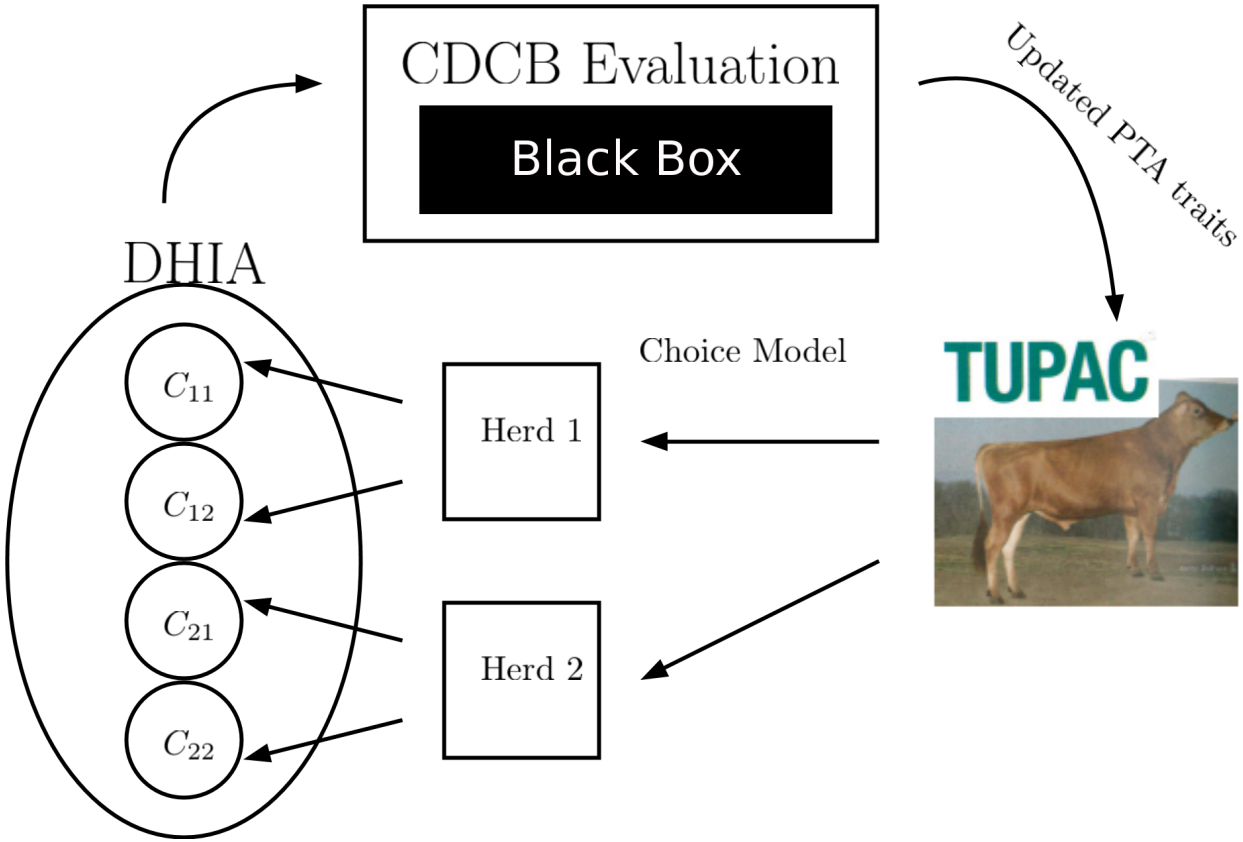
Sire JX SCHULTZ VOLCANO HARRIS {4}
Dam FARIA BROTHERS RENEGADE 215565 {3}, VG-80%
1-08 305d 2x 19,290m 5.1 985f 3.6 702p lbs.
aAa 465 **DMS** 345135
Beta-Casein A1A2 **Kappa-Casein** AB **BBR** 100

AJCA-Trait Profile

Trait		Profile		STA
Stature	Short		Tall	0.2
Strength	Frail		Strong	0.1
Body Depth	Shallow		Deep	N/A
Dairy Form	Tight Rib		Open Rib	0.2
Rump Angle	High Pins		Sloped	0.2
Thurl Width	Narrow		Wide	0.1
Rear Legs-Side Vw.	Posty		Sickle	0.0
Rear Legs-Rear Vw.	Hock-In		Straight	N/A
Foot Angle	Low		Steep	0.1
Feet & Legs Score	Low		High	N/A
Fore Udder Attach.	Loose		Strong	0.2
Rear Udder Height	Low		High	0.4
Rear Udder Width	Narrow		Wide	0.2
Udder Cleft	Weak		Strong	0.9
Udder Depth	Deep		Shallow	0.2
Front Teat Place.	Wide		Close	0.4
Rear Teat Place.	Wide		Close	N/A
Teat Length	Short		Long	0.5
Rear Teat-Rear	Wide		Close	0.0
Rear Teat-Side	Close		Back	0.3

-2 -1 0 1 2

Life and Times of Tupac



- Repeated choices of Tupac will generate a history of evaluations and updated estimates of productivity.

How to Evaluate Tupac

Explain phenotype (output) y as a function of "environment" X and genetics (sire) Z with noise e.

$$y = X\beta + Z\mu + e$$

- For unbiased fixed/random μ, we need:
  - A_{FE}: $E(Ze) = 0$
  - A_{RE}: $Cov(Z, e) = 0$
- The classic case of an "endogenous input" in a production function.

The Henderson Mixed Model (HMM)

Opening the Black Box

$$y = X\beta + Z\mu + e$$

The Henderson Mixed Model (HMM) assumes μ is a random variable and that Tupac's evaluation, PTA $= \hat{\mu}/2$, is a realization from this distribution.

- Model specifies a relationship matrix between all sires to attribute the performance of genetic relatives to Tupac.
- If Tupac is used "enough," A_{RE} is satisfied.
- If we think Tupac's value changes in X, make β a random variable ("Reaction Norm" analysis).

An Alternative Approach

- Reasons we might think A_{FE} or A_{RE} is violated:
  - Unobserved "productivity shock" affecting choice and output (production function literature).
  - **Farm specific return to that input affects choice** (technology adoption literature).
- Instead of using genetic relationships, we could use selection behavior to estimate returns to genetics.
- … but this approach requires a model of selection.

A (Rough) Theory of Sire Selection

The return for farmer i choosing sire j is $q_{ij} = \mu_j + u_{ij}$, where μ_j is a random variable and u_{ij} is farm specific heterogeneity. They must pay price p_j and choose j over j' when:

$$
\begin{aligned}
E(q_{ij}|\mathbb{I}) - p_j &> E(q_{ij'}|\mathbb{I}) - p_{j'} \\
E(\mu_{ij}|I) - E(\mu_{ij'}|\mathbb{I}) &> u_{ij'} - u_{ij} + p_j - p_{j'} \\
E(\mu_j - \mu_{j'}|\mathbb{I}) - \Delta p_{jj'} &> \Delta u^i_{jj'}
\end{aligned}
$$

- Expected net return should be larger than the difference in price plus the unobserved "disadvantage": $\Delta u^i_{jj'} = u_{ij'} - u_{ij}$.
- Expected difference conditioned on information $\mathbb{I}$.

A (Rough) Theory of Sire Selection

We claim that ex-ante returns are a function of X and productivity "signals" W, which also influences $\Delta p_{jj'}$.

$$E(\mu_j - \mu_{j'} | X, W) - \Delta p_{jj'}(W) > \Delta u^i_{jj'}$$

$$R(X, W) > U$$

- Some component of match heterogeneity is explained by X in the first term, but the rest is U.
- Here W has two, countervailing effects; improves expected benefit of match but also may increase the price.

Revised Model

Define Z as a binary variable for choosing j over j' which gives return $\mu(X, U) = f(X) + U$, where $f(X)$ is the heterogeneous effect.

$$y = X\beta + Zf(X) + e$$

$$e = ZU + \epsilon$$

$$Z = \begin{cases} 1 & R(X, W) - U \geq 0 \\ 0 & R(X, W) - U < 0 \end{cases}$$

- In general, A_{FE} and A_{RE} are violated: U is unobserved but relates to both Y and Z.
- The function $\mu(X, U)$ is the "marginal treatment effect" identified with instrument W (Heckman et. al. 2006).

Collecting Data

- DHIA data has the lineage of different dairy cows with their outcomes (format 4) as well as what genetics were chosen for breeding (format 5).
- With these two data sources we have a comprehensive picture of past adoption (lineage of current cows) and current adoptions (breeding attempts).
- Using the CDCB website, we collected historical evaluations of dairy sires and matched them to data to know the characteristics at the time they were chosen.

Data Summary

	Format 4	Format 5	Format 4 w/ sire info
Total Obs	15,519,414 (test day)	7,906,774 (total)	
	2,129,699 (lactation)	4,530,960 (breeding)	1,713,102 (lactation)
No. Cows	1,180,447	1,072,263	448,775
No. Sires	26,529	24,555	10,054
No. Herds	4,259	3,161	1,539
State	Wisconsin		
Timeframe	June 2011 - Jan 2015		

Exercise 1:

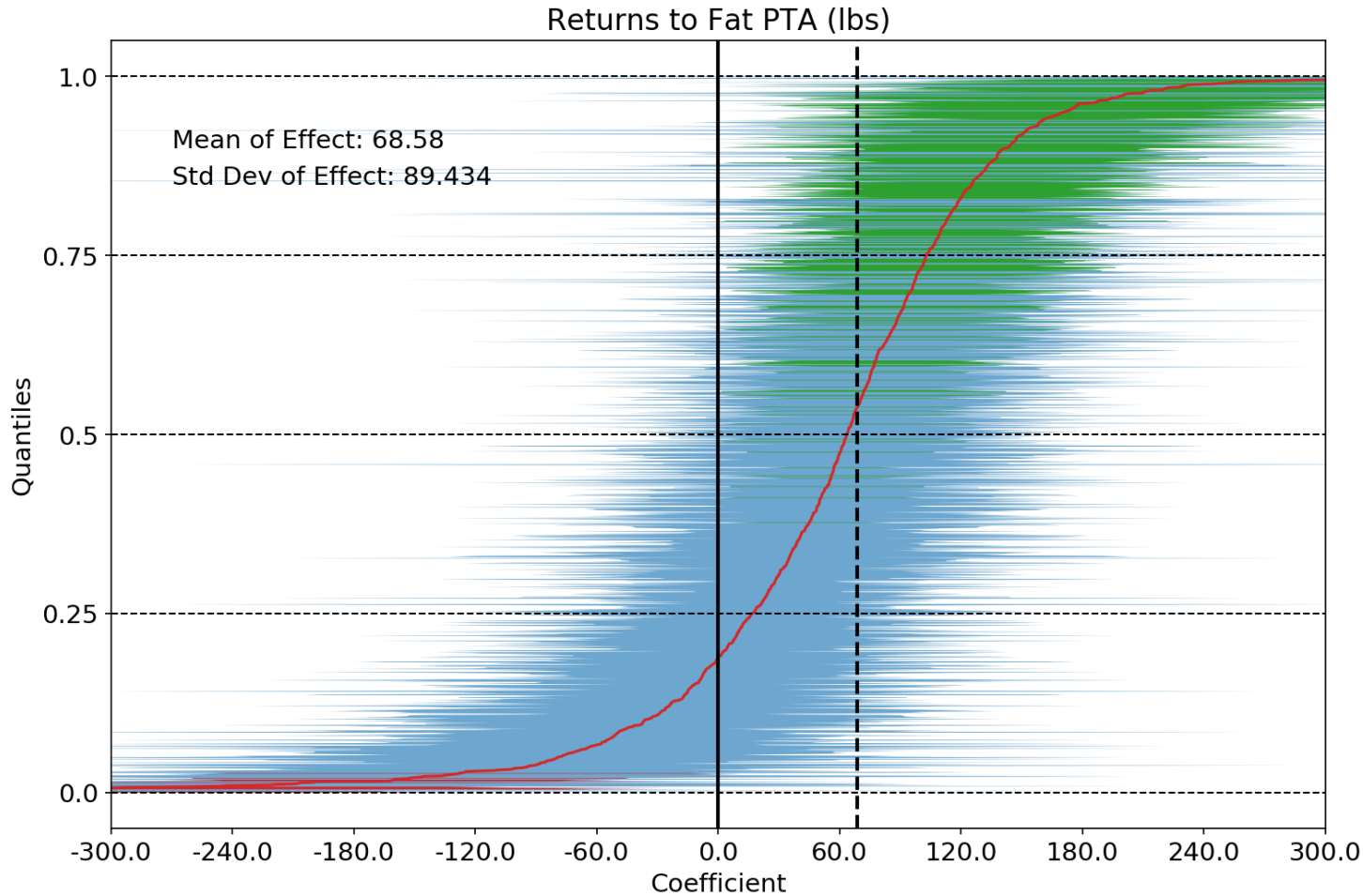
Heterogeneity in Returns to Selecting on Traits

$$y_{iht} = \beta X_{iht} + \tilde{\mu}_h PTA_{iht} + \eta_{iht}$$

- Suppose that every bull is described fully by their PTA number for offspring i on farm h in time t: PTA_{iht}. What are the differences in returns to selecting on just this "input"?
- Controls include lactation number and length, milked 3x, breed, herd fixed effect, ect.
- Next slides are all 1500 $\tilde{\mu}_h$ returns to *PTA* for fat and protein (lactation level).

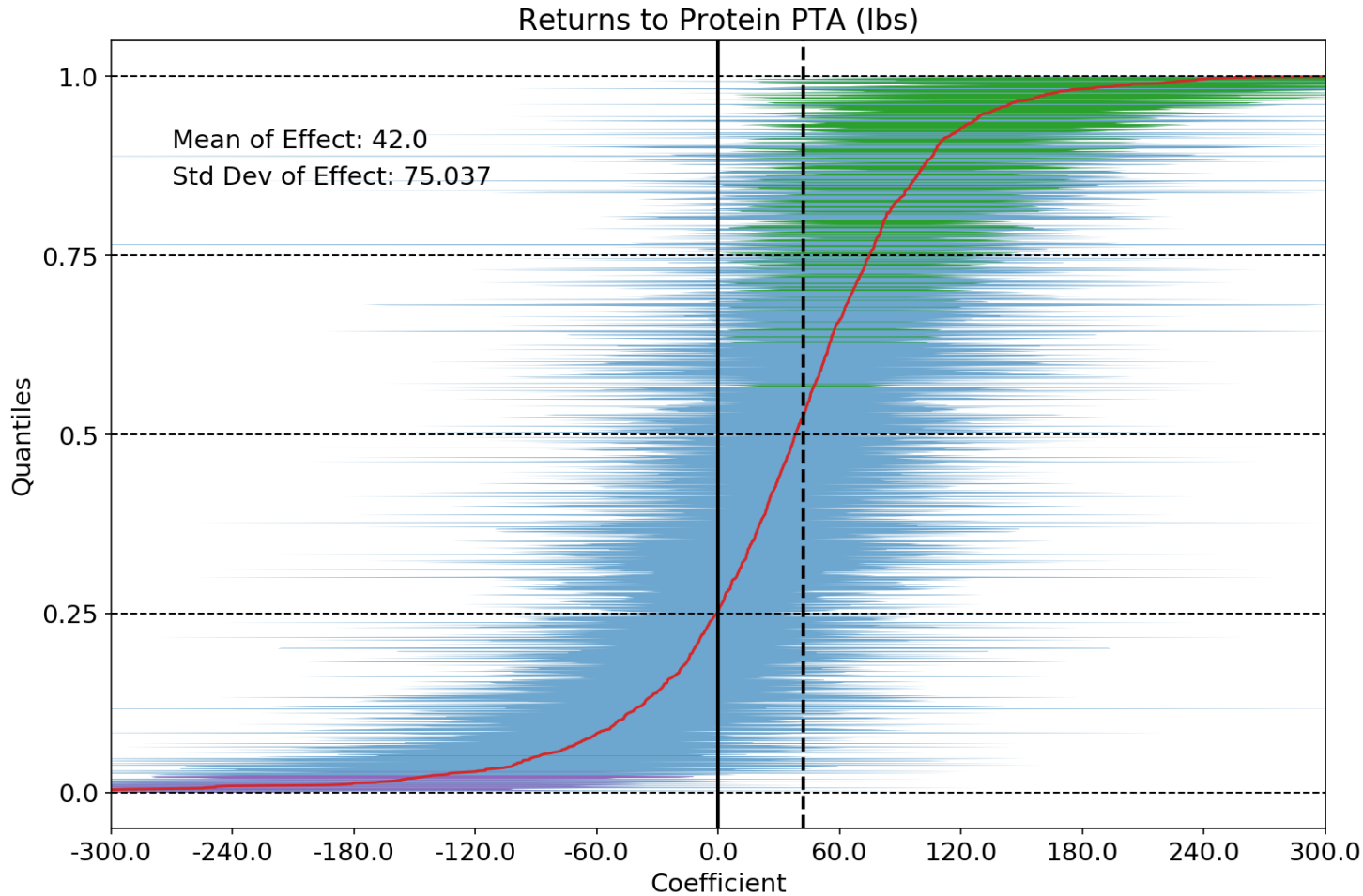
Returns to Fat

Units of 100 PTA



Returns to Protein

Units of 100 PTA



Summary of Exercise 1

- About 1/4 of herds experience significant, positive returns; another 1/4 in point estimate get negative returns.
- The effect of increasing the PTA of a sire, all else equal, is *not homogeneous*.
- A selection model involving several traits could generate negative selection on certain traits depending on objective.

Exercise 2:

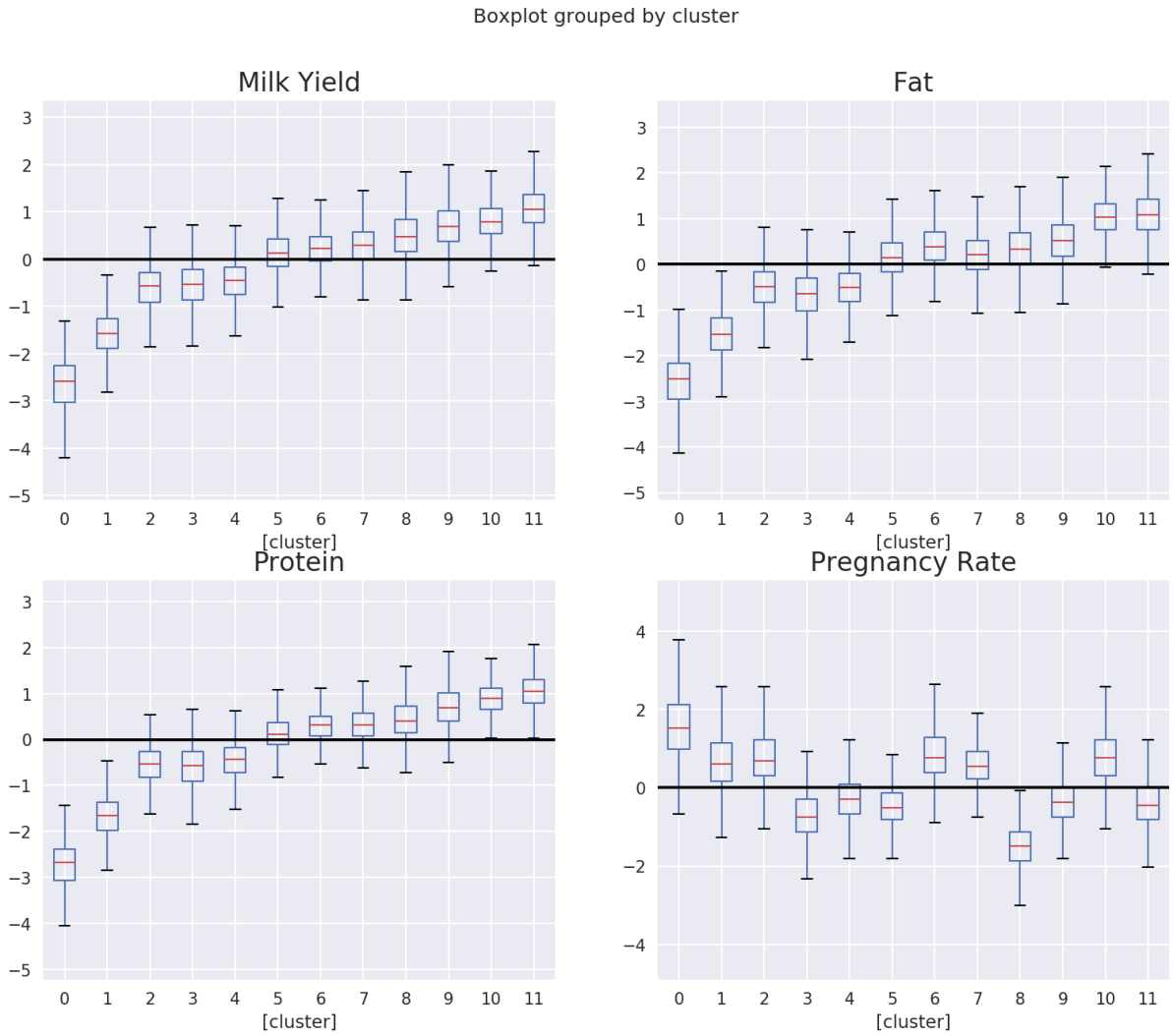
Defining the Choice Set

- Why every sire should not be its own "technology":
  - Anecdotal evidence suggests sire identity is not typically a consideration.
  - Not full support for propensity scores unless every sire is adopted everywhere.
- Alternative: the choice set is described by sire "types."
  - Such a classification can be done with clustering.
  - Instead use $\bar{W}$, the mean NM of each cluster at the time the choice was made.

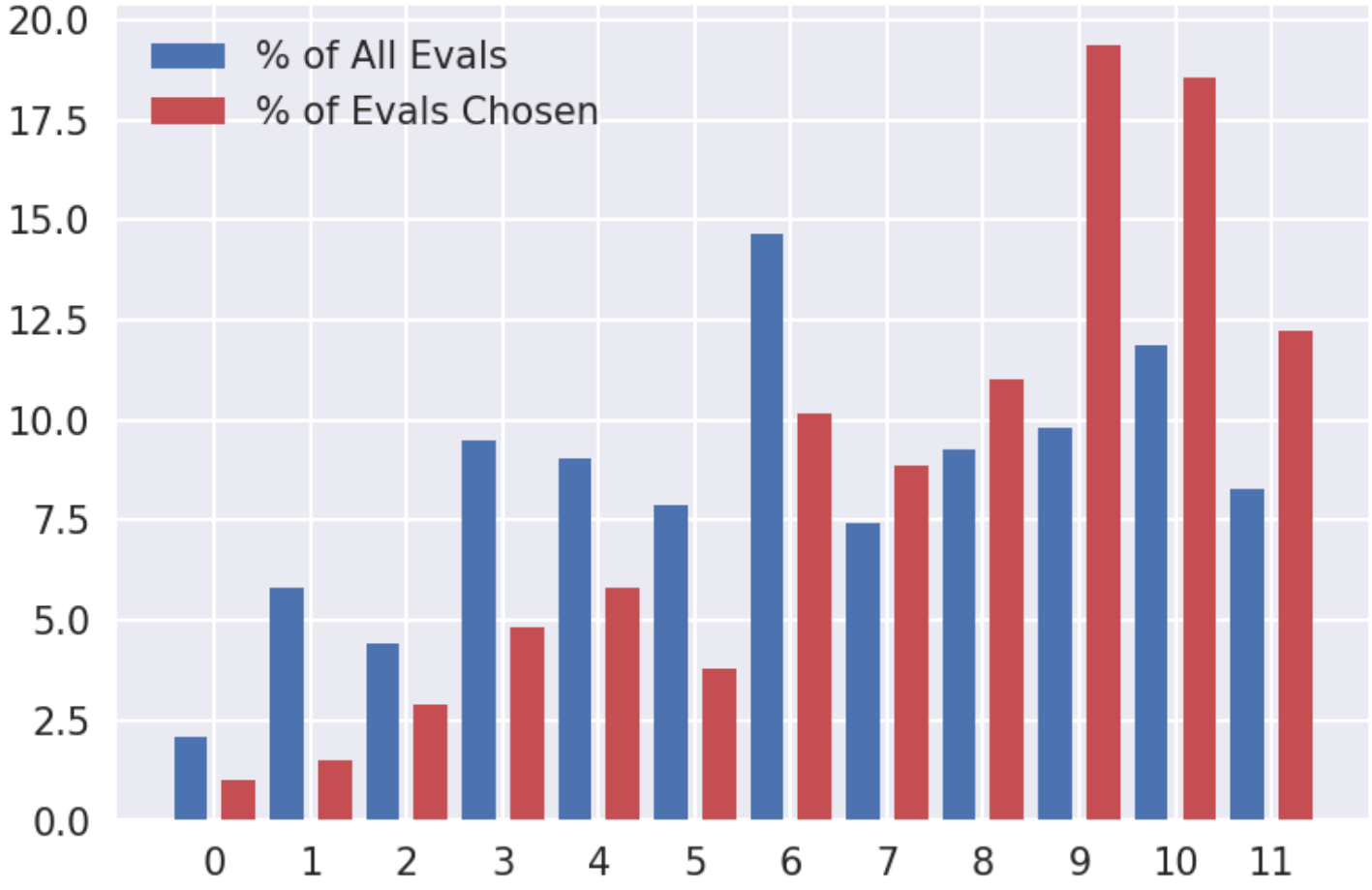
Defining Categories using K-means

- Using K-mean clustering, each bull evaluation was put into one of K categories.
- Choice of *K* must balance:
  - Capturing the diversity of choices (high K).
  - Making sure each category is chosen often enough (low K).
- Silhoutte scores suggested K should be more than between 8 and 12. Here I use 12.
- Features: PTA Milk Yield, PTA Fat, PTA Protein, PTA SCS, PTA DPR.

PTA Characteristics of Categories



Category Frequency



Future Directions

- Estimation of treatment heterogeneity:
  - Semi-Structural approach: “Marginal Treatment Effects” (Heckman et. al. 2006).
  - Nonparametric approach: “Deep IV” (Hartford et. al. 2016)
- More analysis of the first stage to understand how to form groups of sires.
  - Understand stability of preferences.
  - Model the choice more accurately.
- How do we understand “heterogeneity”? What component is other technology choices?

Thank you!