Innovation Through Cooperative Data Governance: The Case of the US Dairy Sector

Jared Hutchins*

Preliminary draft, please do not cite or circulate.

Abstract

We discuss a unique approach to innovation and data governance in agriculture using the dairy sector as a case study. One challenge of data governance is making data available to firms for research while still preserving the rights of those who produce the data. Data has become a powerful tool for innovation in agriculture, and most of this data is now produced on the farm. We examine one institutional approach to managing this trade-off used in dairy for nearly a century: cooperative data governance. The National Cooperative Dairy Herd Improvement Program is a decentralized system for collecting dairy farm data to produce valuable information on cattle genetics while keeping ownership of the data in the hands of the farmer. After discussing its history, we discuss how its evolution can inform data governance in agriculture today. The current landscape of digital agriculture could benefit by emphasizing cooperative ownership, setting uniform data standards, and decentralized operation that the NCDHIP has pioneered for the past century.

1 Introduction

An emerging issue in the economy of data is *data governance*, which is the way in which a firm or country manages the use and storage of data. Data is valuable to firms for conducting research and spurring innovation, but the producers of data value privacy and a say in how their data is used (Acquisti et al., 2016; Jones and Tonetti, 2020). Firms and data producers are at odds when firms use data or even sell data in a way that the producers of the data do not approve. How data can be used by firms to spur innovation and economic growth while

*Assistant Professor at the University of Illinois at Urbana-Champaign. Email: jhtchns2@illinois.edu; Corresponding author

retaining the rights of the data producer is a key question for data governance in the digital economy.

This question is especially prescient in the agricultural sector. Historically, Land Grant Universities (LGUs) have played a large role in agricultural innovation by producing data from experiment stations. Thanks to innovations in precision agriculture technology, one farm can potentially produce a quality and variety of data that can rival anything collected from an experiment station (Coble et al., 2018). Data collection is now more decentralized since this wealth of data is being collected by private firms which sell measurement technology. Unfortunately, the rights of the farmer concerning the data they produce with these technologies remain hazy and ill-defined (Carbonell, 2016; Ferris, 2017). Like many parts of the economy, agriculture is need of a data governance model that can help realize the societal benefits of data while also protecting the rights of the farmers that produce the data.

In this paper, we examine one agricultural industry in the United States which has operated with such a data governance model for nearly a century. The National Dairy Herd Improvement Program (NCDHIP) has been the primary data governance structure for data on dairy cattle breeding in the United States since 1925. The NCDHIP is an agreement agreement between dairy farms, genetics companies, and the USDA to collect, manage, and analyze dairy farm data to alleviate information frictions in the market for bull genetics. The system collects on-farm, animal-level production data from US dairy farms and shares the data with USDA scientists who produce performance evaluations of the dairy bulls currently for sale. The NCDHIP is an example of *cooperative data governance* since nearly every stage of the collection, management, and use of the data is controlled by the producers of the data themselves. The data collected by the system made large-scale research on dairy cow breeding possible for nearly a century without sacrificing the rights of dairy farmers to their own data.

The purpose of this article is to examine how the NCDHIP evolved in the dairy sector and how the NCDHIP's experience can inform data governance in agriculture today. The system evolved to address the information asymmetry in the market for dairy bulls. Due to the decentralized nature of animal breeding, there has never been a central authority that can feasibly provide information on all the dairy bulls a farmer might choose from. The NCDHIP grew to replicate this central authority with a decentralized network of mostly farmer-owned institutions with uniform standards for data collection. We outline three important phases of its development: data collection, data standards, and data scaling. Through these phases, the NCDHIP became an efficient mechanism for effectively crowdsourcing research on dairy cow breeding. The system remains an important part of the dairy sector today.

The evolution of the NCDHIP illustrates three important lessons for governance of agricultural data today. First, the NCDHIP is made up of primarily cooperative institutions, a governance structure which alleviates frictions concerning privacy and data ownership. Giving farmers an ownership stake of the institutions collection data is a straightforward way to assign ownership to the data while still providing a means for it to be shared with others. Second, the NCDHIP set standards for how data would be collected, how performance would be measured, and how individual cooperatives in the system were to be organized. This allowed inter-operability of the data, a current issue with current on-farm data collection. Finally, the NCDHIP is a decentralized system which requires minimal government involvement, making it significantly less costly than other government-led efforts to provide information public goods. A decentralized system better fits the needs of agriculture today given how much data collection is now happening on the farm.

Our discussion is at the cross-section of the study of institutions and the study of data in the economy. Data governance is accomplished by institutions, that is sets of constraints to determine property rights, rules of exchange, and ultimately transaction costs (North, 1991). Aspects of data governance such as standards adoption and inter-operability are examples of coordination problems that are often addressed with institutions (Antonelli, 1994; David and Greenstein, 1990). Data governance is increasingly relevant given how important data is in the modern economy. Data is a non-rival good which, like ideas or information has increasing returns when it is shared. (Akcigit et al., 2016; Jones and Tonetti, 2020; Romer, 1990; Stigler, 1961). Jones and Tonetti (2020) highlights how the institutional aspects of data sharing have direct implications for economic growth and welfare. Using a macroeconomic model, they conclude that the highest welfare is achieved when consumers own their data and sell it to firms for research and innovation. Institutional frameworks are a key component of translating the benefits of data collection, namely innovation, to society as a whole.

Our discussion brings this institutional angle to the current conversation concerning data in agriculture. There is tremendous potential for digital agriculture to spur innovation in the sector (Coble et al., 2018). At the same time, this optimism has been tempered by concerns around privacy and ethics (Carbonell, 2016; Ferris, 2017; Kosior, 2020; Sykuta, 2016). This paper adds to the on-going discussion about data institutions in agriculture and what governance models will maximize benefits to the sector (de Beer, 2016). The challenges the NCDHIP faced bear similarities to today's challenges, and it is critical to understand successful models of data governance like the NCDHIP to face these challenges.

Our article proceeds as follows. The first section examines the conceptual framework for understanding the role of the NCDHIP in facilitating learning through data collection. The next section describes how the NCDHIP came to be as a joint effort between dairy farmers and the USDA to further innovation in dairy cattle breeding. The third section describes how the evolution and operation of the NCDHIP can inform current issues in data collection and governance in agriculture today. The last section concludes.

2 Conceptual Framework

The main role of the NCDHIP is addressing information asymmetry in the market for dairy cow genetics. The majority of discussion around adoption of genetic technology in agriculture has focused on adoption of plant varieties (Ciliberto et al., 2019; Foster and Rosenzweig, 1995; Suri, 2011). Adoption and development of new animal "varieties," as it turns out, is

Table 1: Data Collection Time Across Sector

Type	Time from breeding until first production data	
Annual Plants	2 years	1 year to produce seeds + 1 year to harvest.
Dairy Cattle	**5-8 years**	**10 month gestation** **+ 2 years to producing age** **+ average 3-5 years of production**
Beef Cattle	2-3 years	10 month gestation + 1-2 years to slaughter.
Swine	10 months	About 4 months gestation + 6 months to slaughter.
Broilers	3 months	1 month gestation + 2 months to slaughter.
Layers	2-3 years	1 month gestation + 6-8 months maturation + 1-2 years of production.

significantly more complex. In a textbook dating from 1946, Geneticist Arend Hagedoorn said that animal breeding, compared to plant breeding, was "remarkably speculative and economically wasteful" (Hagedoorn, 1946). The speculation and waste are in part because animal breeding has never had the same centralized authority to provide information that plant breeding has historically had.

Animal breeding has not had the same information infrastructure as plant breeding for two reasons. First, acquiring information on new breeds is more time consuming for animals. Many crop varieties can produce offspring in one year, meaning after creating a new variety it will take two years to obtain data on its yield: the first year will produce seeds and the second year will grow the new variety. For dairy farming, producing the same amount of information would take at least five years. Table 1 shows the time it takes to produce new information on breeds in each animal sector. Between hogs, beef cattle, layers, broilers and dairy, dairy is by far the longest. Both LGU research and farmer experimentation is more costly and time consuming for dairy cattle because of these biological constraints.

A second reason is that animal breeding is completely decentralized. Farmers traditionally do not take an active role in plant breeding, instead letting LGUs or private companies find and prove new crop varieties. By buying seeds, crop farmers have perfect control over which variety they use each year and do not need to engage in breeding. In contrast, dairy farmers must breed cows every year to maintain production and produce replacement cows. Since dairy farmers are always producing new, genetically distinct varieties, it would be infeasible for any LGU to prove all the dairy cow genetics that farmers might choose between at an experiment station.

Without a centralized authority, the only option for dairy farmers is to learn from neighbors or from their own experimentation. Both learning-by-doing and learning from social networks are an important mechanism for farmers to learn about technology in the case of crop varieties (BenYishay and Mobarak, 2019; Conley and Christopher, 2001; Foster and Rosenzweig, 1995). However, they are less useful in dairy farming because the returns to using one bull depend so much on the genetics the farm already has. In other words, a neighbor's signals are *much noisier less informative* in dairy farming (not to mention happen over a much longer time frame). To learn the same amount of information as a crop farmer would from their neighbor, a dairy farmer would need *significantly more* neighbors.

The innovation of the NCDHIP is that it provides exactly the kind of information a LGU would provide but does so by aggregating breeding experiments from all over the country. Consider the Foster and Rosenzweig (1995) framework where a farmer is aiming to learn the average, μ, and the standard deviation, σ, of a technology's distribution. In this case, the technology is a dairy bull that the farmer can breed with an existing cow use to obtain a future replacement cow. The farmer can learn from neighbors who have used the same bull, but the signals will take about five years to be realized (see Table 1). Even when the farmer does observe the cow's offspring, the signal is less informative because of all the private information needed to understand it. About 25% of milk yield is determined by genetics, and of this 25% half is determined by the mother of the cow (Herman, 1981, pg. 16). Both management decisions and previous breeding decisions need to be known to disentangle the effect of using that bull from its confounding factors. Since this is private information, it is unlikely that dairy farmers can effectively learn about genetics through their neighbors' experimentation.

One solution to this problem is to somehow increase the number of neighbors that the farmer can learn from. Intuitively, if the farmer can observe several of the bull's offspring in several different environments then they can estimate μ and σ with more precision and less bias. An institution could accomplish this by collecting all the production data of that bull's offspring has and produce and publish empirical estimates $\hat{\mu}$ and $\hat{\sigma}$. As the bull is used in more environments, $\hat{\mu}$ and $\hat{\sigma}$ could be updated to reflect any new information. In this system, every time a farmer uses a bull for breeding every one of their neighbors would be able to learn from their experimentation. This alleviates the information friction in choosing dairy genetics by effectively crowdsourcing the necessary information from every other dairy farm.

While this seems straightforward, such an institution would have to solve some key problems. How would $\hat{\mu}$ and $\hat{\sigma}$ be estimated? How would farmers be incentivized to contribute their data? Who would pay for the costs of data collection? Finally, who would ultimately own and manage the data? In the next section we discuss the history of the NCDHIP and how it solved these problems in pursuit of genetic improvement in dairy cattle.

3 The History of the NCDHIP

The NCDHIP, an institution addressing information asymmetry, was born in a market that is famously fraught with information frictions: milk. In 19th century New England, a major dairy producing region, watering down milk 25-50% before sale was considered an almost universal practice (Olmstead and Rhode, 2008, pg. 344). Before the beginning of the 20th century, milk was easy to adulterate, leading to badly aligned incentives for farmers and processors. In this section we explore how the NCDHIP evolved to provide information public goods in dairy by studying three major phases of its development: data collection,

Table 2: Timeline of the NCDHIP

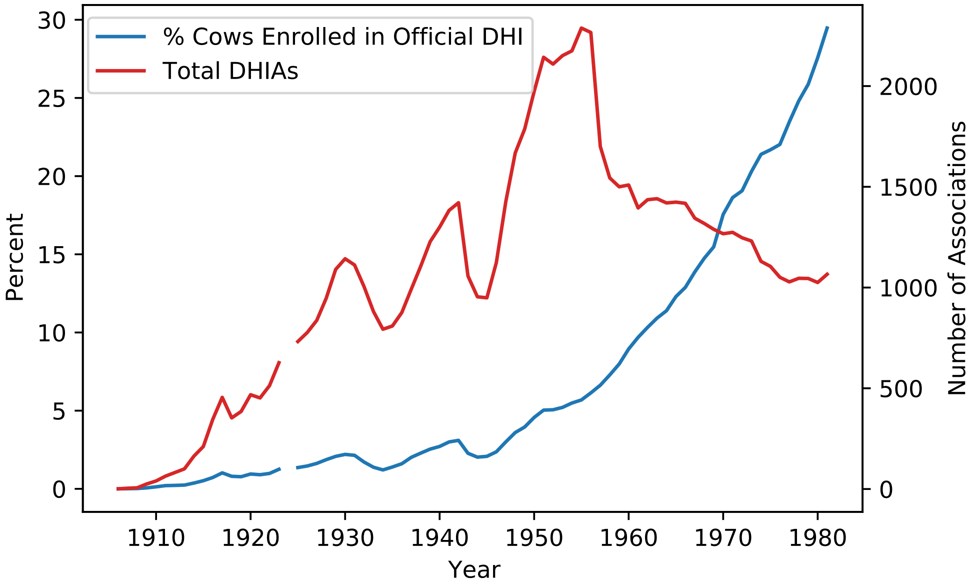
Phases of NCDHIP Evolution	Main Events
Data Collection	• 1890: Charles Babcock invents the Babcock butterfat test. • 1895: Dairy Division in USDA formed. • 1890: Charles Babcock invents the Babcock butterfat test. • 1895: Dairy Division in USDA formed. • 1905: First DHIA formed by Helmer Rabild in Michigan. • 1906: First Bull Cooperative Association formed. • 1908: Helmer Rabild employed by Dairy Division of USDA. • 1914: Smith-Lever Act, Extension aids DHIA formation. • 1917: Breeding research starts at the USDA.
Data Standards	• 1924: ADSA develops testing guidelines for DHIAs. • 1925: DHIA records used for proving sires. • 1935: Nationwide ear-tagging system started by USDA. • 1936: National Sire Proving Program started by USDA. • 1937: USDA publishes its first “sire list.”
Data Scaling	• 1933 (about): AI becomes commercially viable. • 1938: First Cooperative AI Org is formed in New Jersey. • 1946: National Association of Animal Breeders (NAAB) started. • 1952: Official MOU between USDA and DHIA. • 1953: Freezing technology for bull semen viable.

data standards, and data scaling. A timeline with key events can be found in Table 2. We conclude the section by explaining how the NCDHIP operates today.

3.1 Data Collection

The data collection arm of the NCDHIP, called the Dairy Herd Improvement Associations (DHIAs), formed in response to a new incentive created in the dairy market. Charles Babcock, a chemist at the University of Wisconsin, invented a butterfat test for milk in 1890. With the Babcock test, processors could monitor the extent to which dairy farmers adulterated their product by testing the percentage of butterfat in the milk. As one politician put it, the Babcock test “made more dairymen honest than the Bible ever had.” (Olmstead and Rhode, 2008, pg. 344) Since milk buyers now had the ability to measure quality, this created a drastic change in incentives not only for dairy farmers but also for animal breeders. Dairy farmers were incentivized to breed cows that produced the most butterfat rather than the most volume.

Figure 1: DHIA Growth

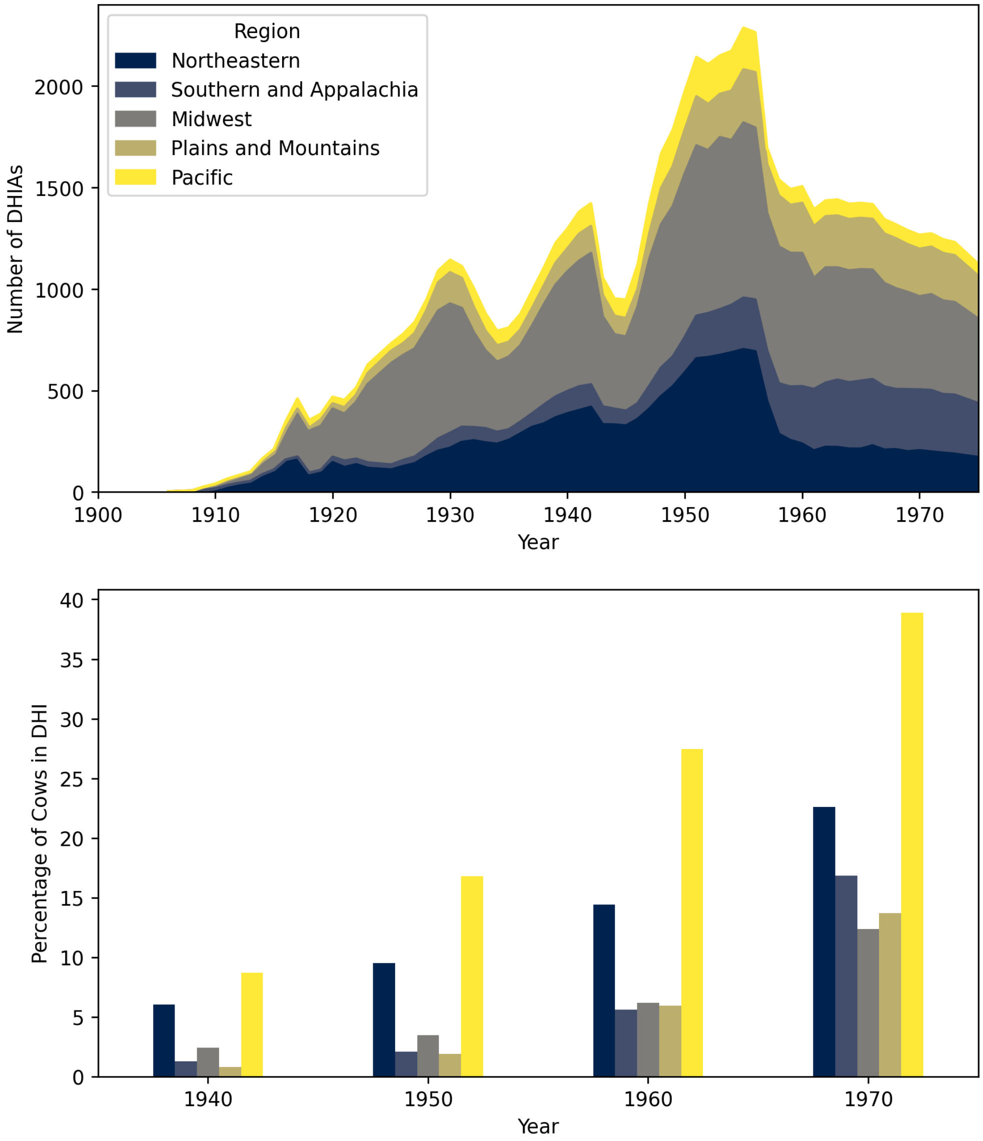


Source: Dairy Herd Improvement Letters, 1925-1980

The DHIAs were cooperatives formed by dairy farmers to test the butterfat producing ability of their cows. Helmer Rabild was a Danish immigrant working for the Michigan Department of Agriculture when he organized the first DHIA, which he called a “cow testing association,” in 1905 (Voelker, 1981). Rabild’s cooperative represented about 14 dairy farms in Newago County, Michigan who together employed one milk tester to take monthly butterfat measurements of their cows using the Babcock test. The cooperative structure was based on the milk testing cooperatives which had been forming for twenty years already in his native Denmark. By the time Rabild formed the first US milk testing cooperative in 1905, there were 400 of these associations in Denmark (Rabild, 1911, pg. 6). The impetus behind forming this cooperative was, at first, only to solve the information friction on the farm. This information allowed farmers to know which cows to remove from the herd. What they still lacked was the ability to tell which bull would produce the offspring they would want as a replacement.

The USDA was quick to catch on to the importance of DHIAs. In 1908, the Dairy Bureau within the USDA employed Rabild to form associations all around the country. With the

Figure 2: Regional Trends in DHIA Growth

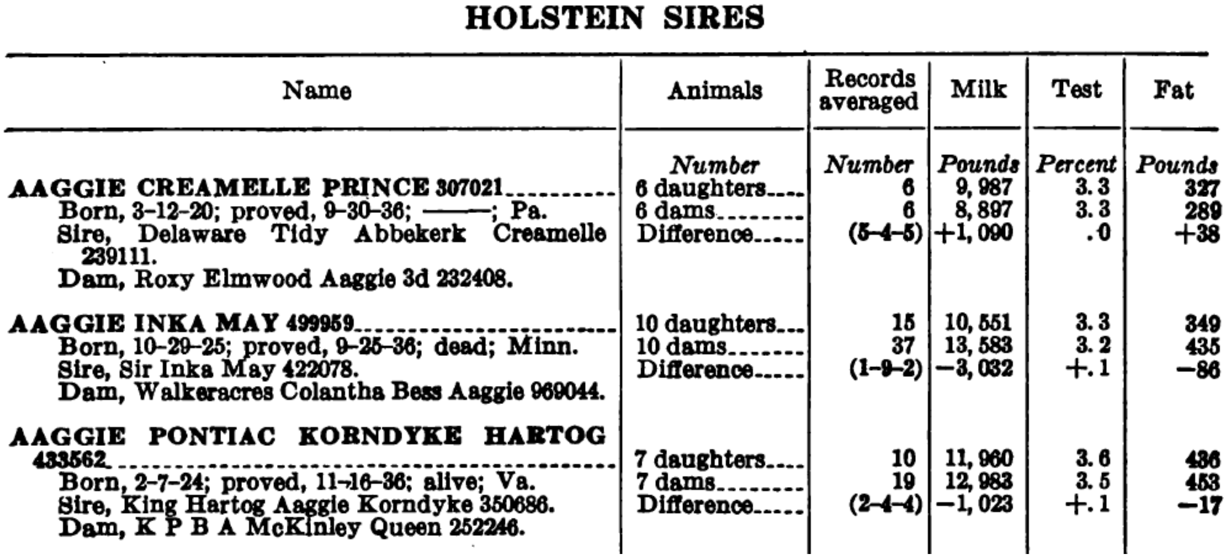


Source: Dairy Herd Improvement Letters, 1925-1980 passage of the Smith-Lever Act in 1914, Rabild obtained the assistance of the Cooperative Extension Service in his mission. Figures 1 and 2 show the rate of growth for DHIAs from 1906 until 1980. Nationally, the number of DHIAs grew until the middle of the 1950s when it began to slowly decline. Conversely, the percentage of US cows enrolled in DHIAs follows a roughly exponential growth pattern: slow growth until about 1945, but accelerating growth onward. Regionally, the number of associations grew the fastest in the Midwest and the Northeast, places where dairy farming already had a long history. The Midwest lagged behind both the Pacific states the Northeast in terms of percentage of cows, however. The Midwest experienced much lower growth in participation than practically any other region. The Pacific region, where the majority of US dairy production would take place in the future, had the highest percentage participation in DHIA even with very few associations.

The DHIAs incentivized participation by providing private benchmarking services which were funded by fees paid by the farmer. Each DHIA was owned by its members, which also elected the board of directors (Bureau of Dairying, 1925). After testing, dairy farmers would be given a monthly report showing the butterfat production of each cow and how their cows compared to other member farms. The testing also revealed how little dairy farmers actually knew about the profitability of different cows. In his report to the USDA, Rabild mentions that the cow farmers thought was the highest producing was often revealed to be one of the poorest producing, highlighting the extent to which information frictions had plagued dairy up to that point (Rabild, 1911, pg. 11).

While farmers could now find out which cow they might want to replace, how would they know which cow they should replace it with? Before 1926, the DHIAs had only addressed one part of the information friction. The USDA, however, saw potential for these farms to address the information friction in selecting new animals. In 1917, the USDA began dairy cow breeding research at the experiment farm in Beltsville, Maryland, which was in essence the same centralized testing model used for crops (Voelker, 1981). As emphasized in the previous section, animal breeding was not amenable to this model. By the time the USDA

Figure 3: The First "Proved Sires" List



HOLSTEIN SIRES

Name	Animals	Records averaged	Milk	Test	Fat
	Number	*Number*	*Pounds*	*Percent*	*Pounds*
AAGGIE CREAMELLE PRINCE 307021	6 daughters	6	9,987	3.3	327
Born, 3-12-20; proved, 9-30-36; ———; Pa.	6 dams	6	8,897	3.3	289
Sire, Delaware Tidy Abbekerk Creamelle 239111.	Difference	(5-4-5)	+1,090	.0	+38
Dam, Roxy Elmwood Aaggie 3d 232408.					
AAGGIE INKA MAY 499959	10 daughters	15	10,551	3.3	349
Born, 10-29-25; proved, 9-25-36; dead; Minn.	10 dams	37	13,583	3.2	435
Sire, Sir Inka May 422078.	Difference	(1-9-2)	−3,032	+.1	−86
Dam, Walkeracres Colantha Bess Aaggie 969044.					
AAGGIE PONTIAC KORNDYKE HARTOG 433562	7 daughters	10	11,960	3.6	436
Born, 2-7-24; proved, 11-16-36; alive; Va.	7 dams	19	12,983	3.5	453
Sire, King Hartog Aaggie Korndyke 350686.	Difference	(2-4-4)	−1,023	+.1	−17
Dam, K P B A McKinley Queen 252246.					

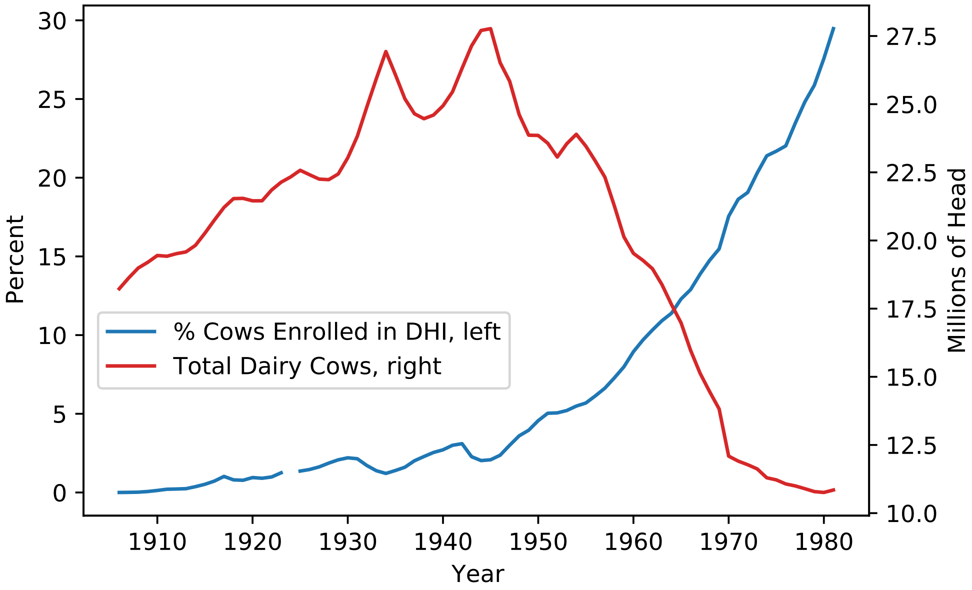
Source: List of Sires Proved in Dairy Herd Improvement Associations (1937), USDA Misc Pub 277

would be able to prove one bull through experimental methods, several more bulls would be adopted and used throughout the dairy sector. The USDA needed a system that could prove genetics at the same pace dairy farmers were breeding.

Using the DHIAs, the USDA essentially crowdsourced animal breeding research by collecting data on the breeding already happening across the country. Using DHIA records, USDA scientists could determine which cows nationwide were producing the most butterfat. Since the breed associations kept detailed records of lineage, the scientists could also determine which bull had sired the best producing offspring. In their first publication in 1937, the USDA listed the average "daughter difference" for each bull, meaning the average difference between the production of the bull's daughters and their mothers (see Figure 3). While these estimates were still only rough estimates of a sire's ability to produce productive offspring, there were a milestone for the US dairy industry. For the first time, dairy farmers had independently verified information on the productivity of different dairy bulls.

Participation in DHIA grew in parallel with the commercialization of the dairy sector. Figure 4 shows that at about the same time that DHIA participation increased the number

Figure 4: Dairy Cows and Participation in DHIA



Source: Dairy Herd Improvement Letters, 1925-1980

of dairy cows began to decline. Cows that used to live in the backyards of farm households began to slowly move onto farms as dairy production became a specialized enterprise. By 1926, two name changes had been made to mark this new era in dairy. First, in 1924 the Dairy Division of the USDA changes its name to the Bureau of Dairy Industry, reflecting the industrialization of the dairy sector. Second, the "cow testing associations" in 1926 became the Dairy Herd Improvement Associations, reflecting a new emphasis on genetic improvement. The next challenge was how to facilitate the growth of this infant data cooperative through coordination in data standards.

3.2 Data Standards

From 1924 to 1937, several efforts were made to coordinate standards in data collection and sire evaluation that helped the NCDHIP grow. Standards are important to economic growth and development in general, and perhaps more so in the case of data (David and Greenstein, 1990; Xu et al., 2012). Standardization is the only way a decentralized system can operate effectively. To do decentralized breeding research with dairy cow records, all dairy farm data had to be collected and treated the same way. While standards often endogenously evolve in economic sectors, the NCDHIP is a particular case where standards were coordinated through both the USDA and the American Dairy Science Association (ADSA), a professional association of dairy scientists. The efforts of the USDA and the ADSA set the foundation for the growth of the NCDHIP by setting standards for data collection, DHIA governance, and the evaluation of bulls.

The first meeting about data collection standards for DHIAs occurred in 1924 at an ADSA annual meeting. The "Dairy Records Committee" was tasked with determining a uniform set of rules across all DHIAs (Voelker, 1981). In their 1925 report, the ADSA specified the equipment that each DHIA had to own, the way the Babcock test was to be administered, the way the data was to be entered, and how the averages had to be calculated. The recommendations of the committee even detail that every dairy farm must "agree to furnish board and lodging for the man employed as tester for at least one day each month." (Bureau of Dairying, 1925, pg. 12)

The committee also provided sample by-laws and contracts that the DHIAs could use for their governance structure. Farmers were to sign contracts with the DHIA to pay the association annual fees in exchange for monthly testing services. The member farmers also elected the board of directors who were usually required to be selected from the existing members (Bureau of Dairying, 1925, pg. 3). The ADSA thought it necessary not only to establish how data collection should be structured but also how each institution should be structured. Standards in data collection were essential to aggregating data across the country, and standards in institutional governance likely helped DHIAs work together more effectively.

While the ADSA set standards for data collection, the USDA set standards for "sire evaluations," or what we referred to as μ and σ in the previous section. The Agricultural Research Service (ARS) became the centralized authority for predicting the performance of dairy bulls. As explained above, estimating μ and σ requires disentangling the effect of genetics from the effects of management from observational data. The goal of their research was to isolate the contribution of the bull to its offspring's productivity from its mother and its environment. To achieve this goal, the USDA enlisted the talent of several influential scientists. Sewall Wright, a scientist who is considered the father of population genetics, worked at the USDA from 1915 until 1925 and helped lay the foundation for the study of animal breeding at the USDA. Another influential scientist was Jay Lush, a geneticist who advocated for a more scientific foundation to animal breeding based on data (Herman, 1981). Lush's work would pave the way for a statistician named Charles R. Henderson who would develop the Henderson Mixed Model, a model which became the standard for data driven animal selection.

In the first sire list published in 1937, the USDA adopted an intuitive metric for estimating a bull's productivity: the daughter-dam comparison (see Figure 3). The daughter-dam comparison measures the difference in production between the bull's mate (the dam) and the bull's offspring (the daughter). If this difference is positive, then the offspring outperformed the daughter when that bull was used. For example, in Figure 3, the first bull in the list had fat production rating of +38 since on average its daughters produced 38 lbs more fat than their mothers This is a crude measure for netting out variation in production explained by the bull, and was inadequate for indexing bull performance. The method completely ignored the effect of the differing management environments of the daughters and mothers and did not take into account that a mediocre bull could still have a large and positive index if it was mated to an already low-producing cow.

Charles R. Henderson in the 1950s furthered research by comparing cows in the same herd rather than cows with their mothers. Comparing cows in the same herd helped to control for the effect of management and environment when comparing the offspring of different bulls. Henderson eventually published one of the more influential models, the Henderson Mixed Model (Henderson, 1975).[1]. The Henderson Mixed Model outputs a prediction of how well

[1] The Henderson Mixed Model is essentially a random effects model where the effects of bulls are modeled as draws from a normal distribution. The normal distribution has a zero mean and a covariance matrix

the bull will "transmit" its traits to its daughter as well a "reliability" score that indicates the variance of the prediction. The first traits that were studied were milk yield and fat yield, but in the future would be expanded to protein production, health traits, and fertility traits.

Both the USDA and ADSA were instrumental in setting standards for data collection and bull evaluation. The ADSA tackled how to make data flow across state boundaries and through institutions, while the USDA used DHIA data to find new and better ways to publish information on bulls. These efforts were vital to establishing the system because they ensured that data could be collected from a variety of sources and that all bulls would be compared the same way. A unique and critical aspect of the NCDHIP is its connection to scientists through the ADSA and USDA who advised on and implemented these standards to help the system work efficiently. It became a symbiotic relationship whereby scientists obtained data for research through the DHIAs and the farmers received an information good as an output of their research.

While bull proving was underway before 1933, scientists were still limited by the fact that one bull could only produce so much offspring naturally. Without enough offspring per bull, candidate estimates of μ and σ were likely to be very imprecise. The next step for NCDHIP was to scale data collection, which was possible thanks to the commercialization of both artificial insemination (AI) in 1933 and semen freezing in 1952.

3.3 Data Scaling

The introduction of artificial insemination (AI) and the freezing of semen both drastically increased the volume of data collected by the DHIAs and the amount of data for doing breeding research. Before AI, a bull could produce on average 12-13 female calves a year by being physically present on the farm (Olmstead and Rhode, 2008, pg. 346). In AI, the bull

where the genetic relationships determine the pattern of covariance. There is also a "fixed" component to the model which is the contemporary group fixed effects, which makes the regression a "mixed model." For a thorough review of animal evaluation models, see Gianola and Rosa (2015)

Figure 5: Use of AI for Breeding

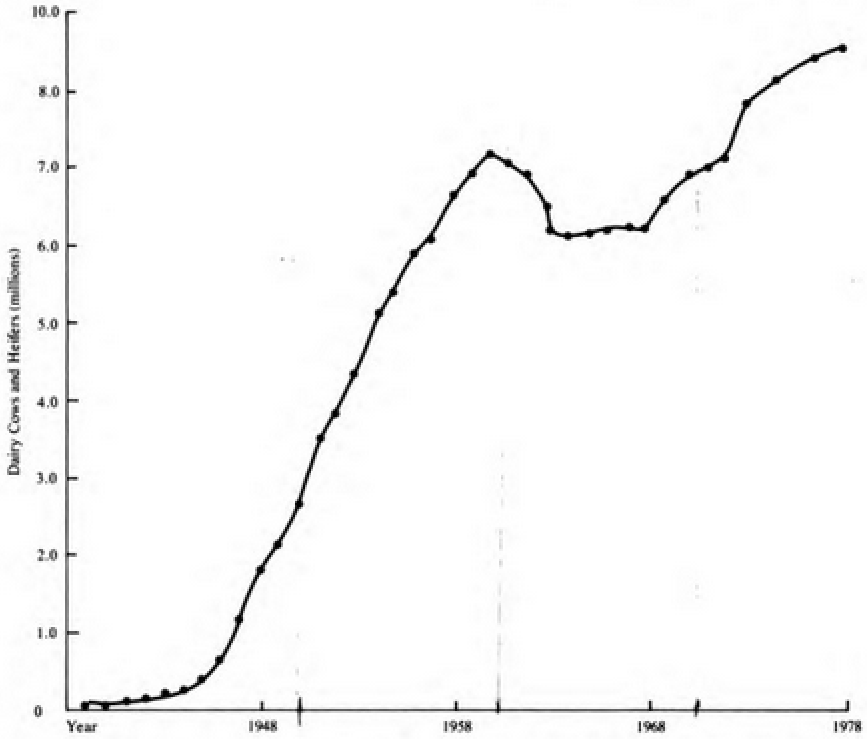


Figure 3.1. **Dairy cows and heifers bred artificially to dairy bulls, 1938–1978. Source: *Dairy Herd Improvement Letter* ARS (1939–1979), USDA; and NAAB reports (1947–1979).**

Source: Herman (1981), pg. 37

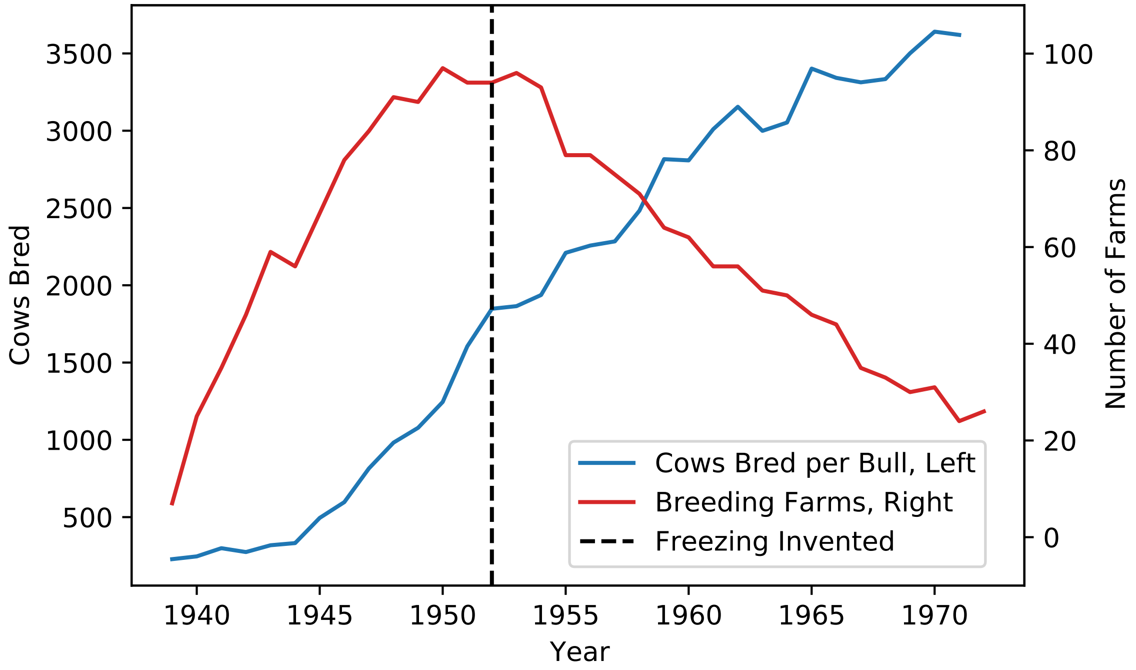
can be bred by multiple farms at the same time. With this method, bulls could produce 5,000 - 20,000 female calves a year, about a thousand-fold increase (Olmstead and Rhode, 2008, pg. 346). After AI, any bull would have about a thousand more data points for estimating evaluations. Figure 5 AI use over time which grew sharply after 1945.

Much like the DHIAs with the Babcock test, dairy farmers organized cooperatives to take advantage of the new technology. The predecessors to AI cooperatives were "bull associations" which acquired breeding stock and rotated the use of each bull throughout the member farms. The first AI cooperative was formed in 1938 in New Jersey and by 1950 there were 1,500 AI cooperatives in the US (Herman, 1981; Olmstead and Rhode, 2008). Unlike the DHIAs, the USDA was not involved in forming these cooperatives. However, like the DHIAs, the ADSA from 1940 to 1943 laid out recommendations for how to organize an AI cooperative (Herman, 1981, pg. 11). The AI cooperatives eventually formed their own organization, the National Association of Artificial Breeders (NAAB), to collect information on the best practices for good AI. The NAAB still exists today as the main governing body of AI companies, the majority of which remain farmer-owned cooperatives.

AI allowed any farmer to use a bull in their association without excluding other members from using it. However, farmer's still could only choose bulls that were geographically in their area because semen could not be transported long distances. If a bull was particularly productive, the benefits would only be realized at the farms that were geographically close to it. This changed with the commercialization of freezing technology in 1952 (Herman, 1981, pg. 85). Once semen was able to be frozen, AI cooperatives could purchase semen from nearly anywhere in the country. This both expanded the options of dairy farmers and also allowed productivity spillovers across the country.

These two technologies, AI and freezing, led to a growth in bull data and a restructuring of the dairy bull industry. Figure 6 shows the number of cows bred by each bull climbing consistently from about 1944 as AI and freezing allowed one bull to be used anywhere in the country. The dashed line indicates 1952, the year that freezing and transporting semen

Figure 6: Growth in Cows Bred and Breeding Farms, 1938-1972



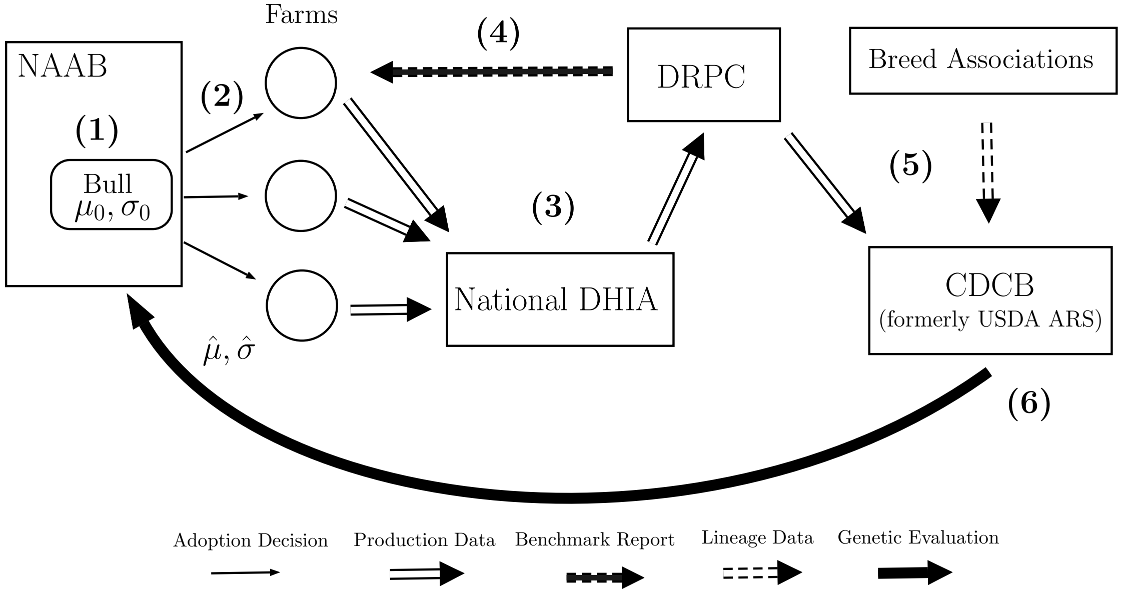
Source: Dairy Herd Improvement Letter 48:4, 1972 USDA

became commercially viable. This parallels the drastic consolidation of AI cooperatives which happened in the 1960s and 1970s. The number of "stud farms," farms specialized in producing dairy bulls for breeding, begins to decline after 1952. AI cooperatives in many areas were formed to give local dairy farmers access to bulls, which was no longer necessary with freezing (Herman, 1981, pg. 182).

3.4 Current Operation

The current NCDHIP system operates via a Memorandum of Understanding (MOU) signed in 1952 between the USDA, the state extension service, and the DHIAs (Service, 1962; Voelker, 1981). This MOU was the first official recognition of the relationship between these institutions and each of their responsibilities in proving sires. The system now includes the National DHIA and the Dairy Records Processing Centers as integral parts of the system. The National DHIA was organized in 1965 to represent all of the DHIAs, which today have been consolidated into just 15 associations covering the country. The data processing is

Figure 7: The NCDHIP System Flow Chart



Steps

(1) Bull has an estimate of its average performance, μ_0, and its standard deviation σ_0.

(2) Farms buy the bull's semen from an NAAB company and use it to produce offspring.

(3) Offspring production data are collected by National DHIA and given to the DRPC.

(4) The DRPC analyzes the data to produce a "benchmark report" for each DHIA member.

(5) The CDCB receives raw production data from the DRPC and lineage data from the Breed Associations.

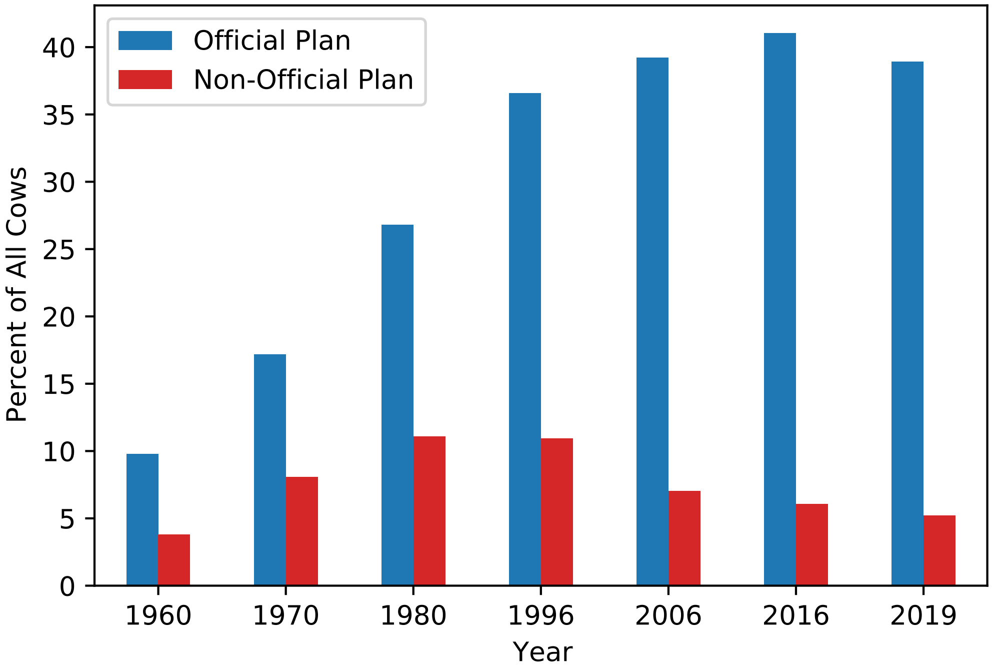
(6) The CDCB produces an updated performance evaluation, $\hat{\mu}$ and $\hat{\sigma}$, for the bull.

handled by Dairy Records Processing Centers (DRPCs) which themselves are farmer owned. Finally, the USDA ARS relinquished its responsibilities estimating sire evaluations in 2013 to the Council on Dairy Cattle Breeding, a member cooperative made up of the NAAB, the DRPCs, the breed associations, and the National DHIA (CDCB, CDCB; Dairyman, 2013).

The current process for evaluating a bull is demonstrated in Figure 7. Suppose a bull has an initial evaluation μ_0 and σ_0 and is sold by one of the members of the NAAB. After the bull is adopted on several farms, the goal is to produce an updated evaluation $\hat{\mu}$ and $\hat{\sigma}$ using the data from these farms. Once the offspring begin producing milk, their data is collected by individual DHIAs and sent to the DRPC. The farmer receives their only direct benefit, a benchmarking and analysis report from the DRPC, at this stage. The DRPC then gives the production data to the CDCB for calculation of evaluations. In order to determine which bull sired the cows in DHIA, the CDCB receives lineage data from the breed associations. The CDCB analyzes the production and lineage data together to produce an evaluation for that bull on a number of traits. In the last stage, the NAAB uses the updated evaluation $\hat{mu}$ and $\hat{\sigma}$ to market the bull to farms. The evaluations are publicly available for any bull and are used by NAAB members to market and price different bulls. Through this system, the entire dairy sector benefits from the innovation drive by collection of farm data.

In terms of governance structure, the majority of the members of the NCDHIP are owned by dairy farmers in some capacity. The NAAB is historically made up of AI cooperatives and the majority of NAAB members today are still owned by dairy farmers. All the DHIAs are owned by farmers and so are most of the DRPCs. The breed associations are also owned by farmers and are collectively represented by the Purebred Dairy Cattle Association (PDCA). As stated before, the CDCB is an organization whose members include many of the above. The only parts of the system not directly owned by dairy farmers are the USDA ARS and the extension service. The USDA ARS maintains a relationship with the CDCB via a Non-funded Cooperative Agreement and remains engaged in breeding research with the CDCB.

Figure 8: DHIA Testing Participation, 1960-2019



Source: DHI Letters, 1925-1980, CDCB Reports

In terms of current participation, about 45% of US dairy cows are enrolled in National DHIA. Figure 8 shows the percent of all cows on the "Official Plan" and the "Non-Official Plan" for different years between 1960 and 2019. The "Official Plan" requires a tester to come to the farm to collect the milk samples for analysis while in the "Non-Official Plan" the farmer collects the milk samples themselves and sends it to the DHIA lab for analysis.[2] Since only "Official Plan" measurements are supervised, these are the only records included in the bull's evaluation. Both plans grew until about 1996, after which participation in the Non-Official Plan has dropped from 10% to 5%. Participation in the Official Plan grew to 41% in 2016 before declining to about 38% in 2019.

These trends demonstrate that farmer participation remains somewhat strong in the system. Unlike in 1960, there are a variety of firms today that can provide data analysis services for dairy farms in direct competition with the National DHIA. This may have been one factor

[2] The Non-Official Plan came about as a result of labor shortages in the 1940s for DHIA testers. The Non-Official Plan is also cheaper than the Official Plan since less labor is needed.

in the decline in participation in the Non-Official Plan. Despite growing competition, participation in the Official Plan has not drastically changed since 2006. If the private benefits of participating in National DHIA can easily be found outside the system, it seems likely farmers would not be incentivized to contribute to the public good the NCDHIP provides. Instead, the long history that dairy farmers have with the NCDHIP likely has engendered institutional trust which helps keep participation strong.

4 Lessons Learned from the NCDHIP

The evolution of the NCDHIP was only possible through several research and technological innovations: the Babcock test, innovations in statistical modeling, artificial insemination, and freezing technology. However, technological innovations do not necessarily result in broader innovations benefiting the whole dairy sector. One reason the NCDHIP was able to translate these innovations into benefits for all dairy farmers was the institutional underpinnings of the system. DHIAs and most of the NAAB members are farmer-owned, which allows dairy farmers to have direct control over how their data is used and who it is shared with. In order to maximize the benefits from aggregating data, both government and non-government scientists have helped develop data standards and evaluation methods to streamline research in dairy cow breeding. Finally, the decentralized nature of the system helps it obtain data from a variety of sources with little cost to taxpayers.

In this final section we discuss how these institutional underpinnings can inform data governance in agriculture today. With precision agricultural data becoming easier to collect, there are opportunities to realize benefits from these data never before possible. Institutions determine the extent of the benefits as well as how they are distributed. Data collection in agriculture has become increasingly decentralized, which makes the NCDHIP a relevant case study for seeing how institutions can efficiently realize and distribute the benefits of data aggregation. Three policy relevant attributes of NCDHIP are cooperative ownership, uniform data standards, and decentralization.

4.1 Cooperative Ownership

Who owns and has rights over data is one of the biggest questions concerning data today. High-profile companies engaged in data collection have often engendered bad faith with data producers by selling and using data in a way the producers of the data do not approve. Part of the reason for these problems is that the rights of the data producer are often ill-defined. Since data producers do not always own what they produce, they have little legal recourse when the data collection firm acts in bad faith.

The NCDHIP avoided this problem by having the farmers own the organization that collects and manages the data. When the Babcock test was invented, the collection of milk data was owned and organized by farmers from the very beginning. The DHIAs and the DRPCs have provided a platform for both processing the data and sharing the data with scientists conducting research in animal breeding. The National DHIA represents all of the DHIAs and has the authority to form data sharing agreements with universities and researchers who wish to use dairy farm data collected by the DHIAs. By managing the data through these cooperative institutions, the farmers retain their property rights to their data and still have a mechanism for sharing it with the scientific community.

Data cooperatives can help to manage the trade-off between data use and privacy in other agricultural sectors as well. In a legal landscape where rights over data are not defined, cooperative ownership can be used to better establish these rights (Carbonell, 2016; Ferris, 2017). Organizations like Ag Data Transparent have made great strides in opening up the conversation on data rights, but such voluntary agreements may still not be enough. Until the legal framework around agricultural data changes, cooperative ownership can provide much needed clarity in data ownership that farmers may be needing. Block chain technology makes it even less costly to store data while maintaining privacy and doing so without the need for a centralized intermediary (Davidson et al., 2018; Paik et al., 2019). Using cooperative governance, farmers can both reap the private benefits of their data as well as the indirect benefits from researchers using their data.

4.2 Uniform Data Standards

Adoption of standards is vital for data sharing and realizing the gains of data aggregation. Agriculture faces issues in data standardization because i) agricultural data is complex and multi-facted and ii) precision agriculture data collection currently happens through several, independent private firms. Agricultural data is multi-faceted because it is now acquired through surveys, satellites, written records, and increasingly through a number of different on-farm sensors. Universal, standardized data formats and protocols are needed to be able to aggregate these data sources for analysis, standards agriculture is currently lacking (Anderson et al., 2013; Bahlo et al., 2019). In the absence of a universal data standard, private firms which collect data through precision agriculture technology must invent their own standards. Since the major data collection firms do not manage or collect data in the same way, it is difficult if not impossible to aggregate data from these disparate sources.

The NCDHIP faced a similar issue after DHIAs formed. Data collection happened through thousands of cooperatives across the country who likely all had their separate ways of collecting and storing data. The ADSA played a vital role in making sure both data standards and institutional governance were uniformly defined across the system. Because the ADSA was a professional organization, it could draw on the necessary technical expertise to design standards and change them as needed. The system also received technical expertise from Land Grant scientists, extension agents, and the USDA ARS. The ARS was particularly important since it was responsible for storing DHIA data until the CDCB took over its responsibilities.

Agricultural data could benefit from leadership in data standards to enhance the sharing and use of precision agricultural data. Many efforts are underway to make agricultural data more interoperable, including OpenTEAMS and AgStack. Yet, these efforts may fail to result in coordination if there is not strong leadership in setting these standards. This kind of coordination may not arise endogenously, and the agricultural sector may need something akin to the ADSA to begin the process of setting data standards. Like many standards, there will likely be increasing returns to scale of adoption since it will lower transaction costs substantially. Open standards are a good start given how much the open-source community has contributed to standards and protocols in software development. The trick is initiating the coordination, which may necessitate strong leadership from industry groups, the USDA, or ideally a partnership of industry and government like the NCDHIP had.

4.3 Decentralization

Decentralization can also be an incredibly effective institutional structure with current technology. The modern economy is full of examples of internet platforms being used to crowdsource geospatial data (OpenStreetMap), general knowledge (Wikipedia), and software development (GitHub). Data contributors being geographically disparate no longer matters with current technology the way that it did when the DHIAs were formed. Decentralization can not only be effective for collecting data and conducting research but is even less costly with current technology.

Decentralization is a key aspect of the NCDHIP which is relevant to today. The original reason for decentralization in the NCDHIP was simply that dairy farmers were geographically spread out. The technology did not exist to govern and administer a dairy benchmarking program in a centralized way, and doing so would have been prohibitively costly. Because the DHIAs were farmer owned, the system was always run as a partnership between dairy farmers, scientists, and the government and never as a top-down government program. As a result, the system costs taxpayers less and allows dairy farmers say in how the program operates (through the National DHIA and their membership in the NAAB).

The agricultural sector would benefit from leveraging data from across US agriculture by making use of new technology for coordination used already in crowdsourcing models.

Since much of the data can be used to produce research useful to the sector as a whole, the data would ideally be open access to allow as many researchers as possible to use it (after establishing safeguards for privacy). Some examples of crowdsourcing research via decentralized data collection are MIDATA and SALUS COOP for health research, both of which are themselves cooperatives. The "citizen science movement," a movement which seeks to engage everyday people in collecting data for scholarly research, may also have useful insights for engaging farmers in helping crowdsource agricultural research. Cultivating and curating this kind of data resource would be a tremendous benefit to the sector as a whole.

5 Conclusion

The recent innovations in on-farm measurement technology have revolutionized how farm operations can be run. Information previously unknown to the farm operator can now be integrated into management, which will likely lead to productivity improvements in agriculture. The benefits of digital agriculture can be beyond what happens on the farm, however. Data from farms can be a resource for furthering research and innovation beyond the farm. The past model of collecting data for research at experiment station plots at LGUs is less relevant in a world where the bulk of data collection happens on the farm. Moreover, innovation may be stifled if farmers do not have rights to how their data is used or a say in how the data is governed (Jones and Tonetti, 2020).

The example of the NCDHIP shows how innovation can flourish even when data collection is decentralized. Like many farmers today, dairy farmers were given a measurement technology, the Babcock test, which could provide vital information about their operations. Their data collection proved to be an enormous asset to the dairy sector as a whole thanks to the organization efforts of dairy farmers, the USDA, and other industry actors. The NCDHIP is an example for how on-farm data collection can benefit farmers, encourage innovation, and address privacy and property rights in data. Institutions that practice this kind of cooperative data governance while partnering with the scientific community merit further study in the field of economics. These kinds of institutions demonstrate that innovation and research through data collection and protecting privacy and property rights need not be trade-offs at all.

References

Acquisti, A., C. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 54*(2), 442–92.

Akcigit, U., M. A. Celik, and J. Greenwood (2016). Buy, Keep, or Sell: Economic Growth and the Market for Ideas. *Econometrica 84*(3), 943–984. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12144.

Anderson, D., R. Estell, and A. Cibils (2013). Spatiotemporal cattle data - A plea for protocol standardization. *Positioning 4*(1), 115–136.

Antonelli, C. (1994, December). Localized technological change and the evolution of standards as economic institutions. *Information Economics and Policy 6*(3-4), 195–216. Publisher: North-Holland.

Bahlo, C., P. Dahlhaus, H. Thompson, and M. Trotter (2019, January). The role of interoperable data standards in precision livestock farming in extensive livestock systems: A review. *Computers and Electronics in Agriculture 156*, 459–466.

BenYishay, A. and A. M. Mobarak (2019, May). Social Learning and Incentives for Experimentation and Communication. *The Review of Economic Studies 86*(3), 976–1009.

Bureau of Dairying (1925). Cow Testing Association Letter No. 1.

Carbonell, I. (2016). The ethics of big data in big agriculture. *Internet Policy Review 5*(1), 1–13.

CDCB. Board of Directors.

Ciliberto, F., G. Moschini, and E. D. Perry (2019). Valuing product innovation: genetically engineered varieties in US corn and soybeans. *The RAND Journal of Economics 50*(3), 615–644. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12290.

Coble, K. H., A. K. Mishra, S. Ferrell, and T. Griffin (2018). Big Data in Agriculture: A Challenge for the Future. *Applied Economic Perspectives and Policy 40*(1), 79–96. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1093/aepp/ppx056.

Conley, T. and U. Christopher (2001). Social Learning Through Networks: The Adoption of New Agricultural Technologies in Ghana. *American Journal of Agricultural Economics 83*(3), 668–673. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0002-9092.00188.

Dairyman, H. (2013). USDA-ARS, CDCB sign agreement to transfer genetic evaluations. Section: Page, Industry Buzz.

David, P. A. and S. Greenstein (1990, January). The Economics Of Compatibility Standards: An Introduction To Recent Research. *Economics of Innovation and New Technology 1*(1-2), 3–41. Publisher: Routledge _eprint: https://doi.org/10.1080/10438599000000002.

Davidson, S., P. D. Filippi, and J. Potts (2018, August). Blockchains and the economic institutions of capitalism. *Journal of Institutional Economics 14*(4), 639–658. Publisher: Cambridge University Press.

de Beer, J. (2016). Ownership of Open Data: Governance Options for Agriculture and Nutrition. SSRN Scholarly Paper ID 3015958, Social Science Research Network, Rochester, NY.

Ferris, J. L. (2017). Data privacy and protection in the agriculture industry: Is federal regulation necessary. *Minn. JL Sci. & Tech. 18*, 309. Publisher: HeinOnline.

Foster, A. D. and M. R. Rosenzweig (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of political Economy 103*(6), 1176–1209. Publisher: The University of Chicago Press.

Gianola, D. and G. J. M. Rosa (2015). One Hundred Years of Statistical Developments in Animal Breeding. *Annual Review of Animal Biosciences 3*(1), 19–56.

Hagedoorn, A. L. (1946). *Animal breeding.* Agricultural and horticultural series. London: C. Lockwood and Son Ltd.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics 31*(2), 423–447.

Herman, H. A. (1981). *Improving cattle by the millions: NAAB and the development and worldwide application of artificial insemination.* University of Missouri Press.

Jones, C. I. and C. Tonetti (2020, September). Nonrivalry and the Economics of Data. *American Economic Review 110*(9), 2819–2858.

Kosior, K. (2020). Economic, Ethical and Legal Aspects of Digitalization in the Agri-Food Sector. *Zagadnienia Ekonomiki Rolnej/Problems of Agricultural Economics*.

North, D. C. (1991, March). Institutions. *Journal of Economic Perspectives 5*(1), 97–112.

Olmstead, A. L. and P. W. Rhode (2008). *Creating Abundance.* Cambridge Books.

Paik, H.-Y., X. Xu, H. M. N. D. Bandara, S. U. Lee, and S. K. Lo (2019). Analysis of Data Management in Blockchain-Based Systems: From Architecture to Governance. *IEEE Access 7*, 186091–186107. Conference Name: IEEE Access.

Rabild, H. (1911). *Cow-testing associations.* Washington, D.C. : U.S. Dept. of Agriculture, Bureau of Animal Industry.

Romer, P. M. (1990, October). Endogenous Technological Change. *Journal of Political Economy*. Publisher: The University of Chicago Press.

Service, F. E. (1962). *A Handbook for Extension Workers: National Cooperative Dairy Herd Improvement Program.* Number 248 in Agriculture Handbook. USDA.

Stigler, G. J. (1961). The economics of information. *Journal of political economy 69*(3), 213–225. Publisher: The University of Chicago Press.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica 79*(1), 159–209.

Sykuta, M. E. (2016). Big data in agriculture: property rights, privacy and competition in ag data services. *International Food and Agribusiness Management Review* (Special Issue), 57–73. Publisher: International Food and Agribusiness Management Association.

Voelker, D. E. (1981). Dairy herd improvement associations. *Journal of Dairy Science 64*(6), 1269–1277. Publisher: Elsevier.

Xu, S. X., C. Zhu, and K. X. Zhu (2012, January). Why do firms adopt innovations in bandwagons? Evidence of herd behaviour in open standards adoption. *International Journal of Technology Management 59*(1/2), 63–91. Publisher: Inderscience Publishers.