Human Behavior and Prediction versus Causation with Big Data

Paul D. Mitchell

Professor, Ag and Applied Economics and UW Extension

Director, Renk Agribusiness Institute, UW-CALS

Co-Director, Nutrient & Pest Management Program, UW-Ex

Big Data and Ecoinfomatics in Agricultural Research

University of Wisconsin April 27, 2017

608-265-6514, pdmitchell@wisc.edu, @mitchelluw

NPM

RENK AGRIBUSINESS INSTITUTE
College of Agricultural & Life Sciences

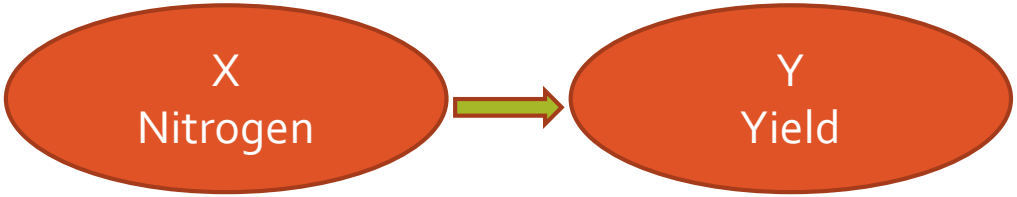
UW Extension
University of Wisconsin-Extension

College of Agricultural & Life Sciences
UNIVERSITY OF WISCONSIN-MADISON

The Problem

- Much of science is testing hypotheses about causation and using the results to make predictions
- Experimental approaches are the standard method
  - Vary a treatment experimentally, replicate and randomize, and then estimate the treatment effect
- The Problem: social and ecological sciences often only have observational data, not experimental data
  - Big data in agriculture, ecology, epidemiology
- Analyzing observational data to make causal inferences for prediction requires appropriate methods, especially if the data are generated by human decisions

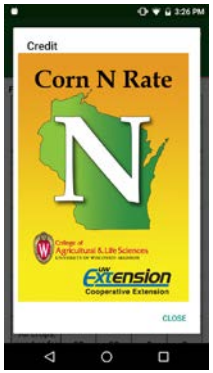
Example: Nitrogen and Corn



- Conduct experiments to trace out the nitrogen response curve for corn yield
- Suppose you wanted to estimate the nitrogen response curve for many farms all over the state
  - To make recommendations, set policies, …
- Survey data on N use and yield for hundreds of farmers to estimate the population's nitrogen response curve
  - Data-driven policy and recommendations
- Problem: These are observational data!

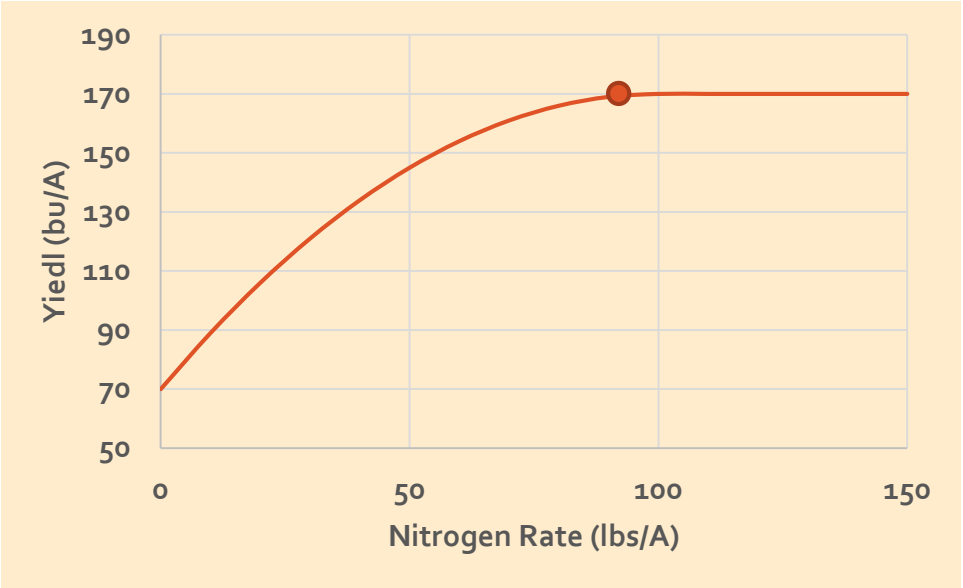
Optimizing the N Rate for Corn

- Suppose a farmer knows his N response curve is a Quadratic Response and Plateau
- Yield = $A + B \times N + C \times N^2$, up to the peak, then flat
- Farmer chooses N rate to maximize net returns
- Max $\pi = p(A + B \times N + C \times N^2) - r \times N$
- $N^* = (r/p - B)/(2C)$ $\quad Y^* = A + B \times N^* + C \times N^{*2}$
- This is the standard MRTN process: use experimental data to estimate coefficients & make recommendations
- http://cnrc.agron.iastate.edu/



NPM's Corn N Rate Calculator App

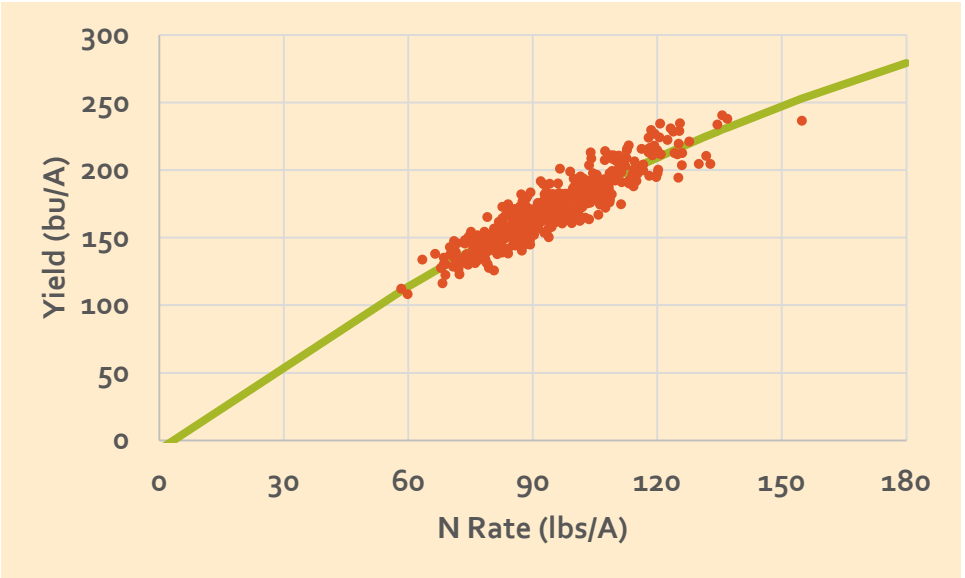
- A = 70, B = 2, C = -0.01
- N price r =$0.40
- Corn price p = $3.50
- Y_{max} = 170 @ N = 100
- N* = 94.3 lbs/A
- Y* = 169.7 bu/A



- Suppose each farmer knows his parameters and optimizes, but A, B and C vary a little bit for each farm
- A~N(70,7), B~N(2,0.2), C ~N(-0.01, 0.001) (CVs = 10%)
- Randomly draw 500 A, B, & C, calculate N* and Y*, then regress $Y^* = A + B \times N^* + C \times N^{*2}$ to recover the average nitrogen response curve for corn for the population

Assumptions

- A~N(70,7), B~N(2,0.2), C~N(-0.01, 0.001)
- r =$0.40, p = $3.50
- Farmers maximize profit
- Survey 500 farmers to collect N* and Y*
- Regression: $Y^* = A + B \times N^* + C \times N^{*2}$
- N* =299.3 , Y* = 343.1



Param	Coeff	St Err	t	p
A	-6.94	13.83	-0.50	0.616
B	2.22	0.28	7.84	0.000
C	-0.0035	0.0014	-2.45	0.015

- Estimated coefficients are biased, do not accurately reveal the underlying causal relation between N and Y because the data are observational: predications will be biased
- **Obseved N and Y are endogenous: co-determined by profit maximizing choices of farmers**

Solutions to Address Endogeneity

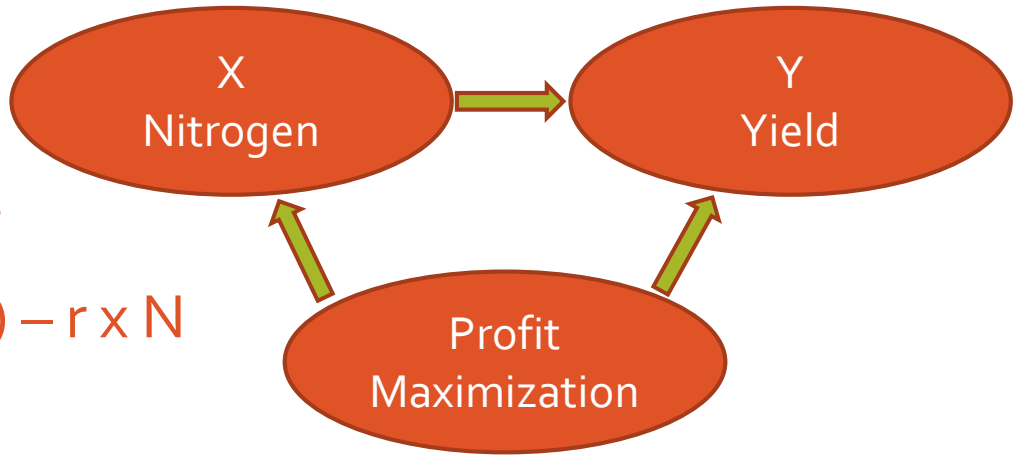
1) Conduct experiments
2) Use "natural experiments"
   - Atrazine prohibition areas: Dong et al. (2016)
3) Model the co-determining process (structural model)

- Panel Data Methods to address endogeneity, omitted variables, and measurement error

4) Instrumental Variables (IV)
   - Used for cross-sectional data
5) Fixed Effects and Control Variables
   - Used for longitudinal data

Model the Co-Determining (Behavioral) Process

Use theory to posit a model of the process that co-determines N* and Y* to derive estimation equations



Max $\pi = p(A + B \times N + C \times N^2) - r \times N$

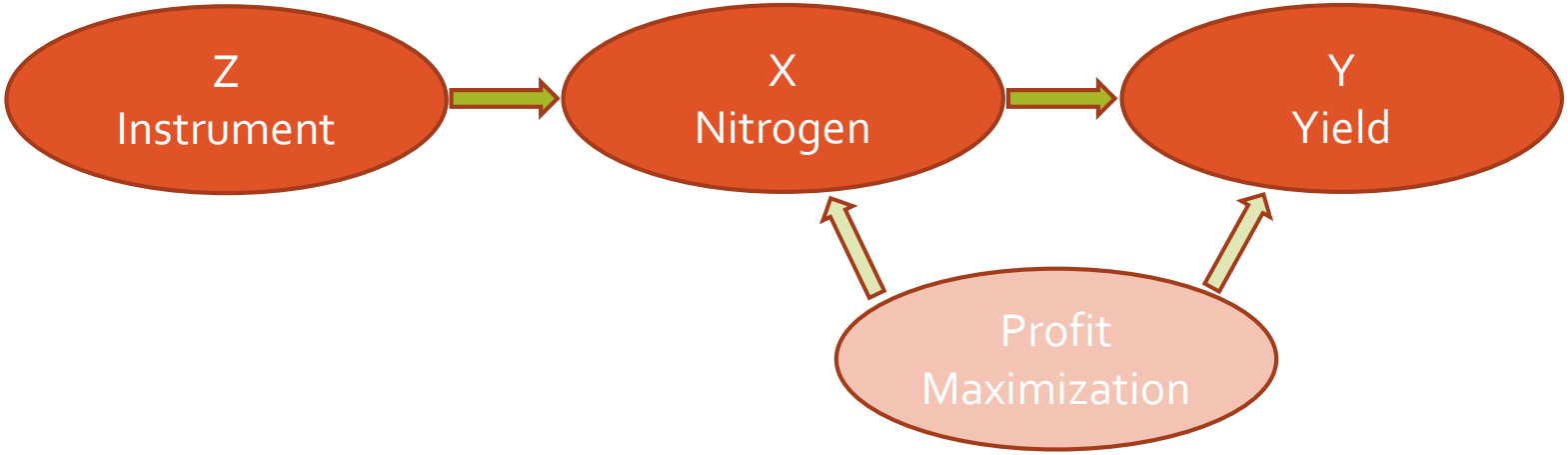
- $N^* = (r/p - B)/(2C) + \varepsilon_1$
- $Y^* = A + B \times N^* + C \times N^{*2} + \varepsilon_2$

- Notice: r/p is the regression variable, yet can recover the coefficients A, B and C of the N response function
- Primal Approaches: profit max, utility max, …
  - Identify production parameters as perceived by the farmers
- Dual Approaches: cost min, revenue max, …
  - Can't recover production function, but it's there

Applications of Primal/Dual Approaches

- Chambers and Tzouvelekas (2013) "Estimating population dynamics without population data"
  - Estimated olive fruit fly dynamics using insecticide use data and prices. Results matched observed dynamics.
- Wechsler et al. (2017?) Use survey data and weed management decisions to estimate
  - Average yield loss without weed control: 38-41%
  - Average yield loss with weed control: 4-8%
  - Average glyphosate efficacy: 98%, 95% if resistance
- Someone: Survey data to estimate RW Bt corn efficacy
- Mitchell et al. (201?): GfK Kynetec seed purchase data, prices, and corn seeding rates to estimate yield response curve to seeding rates

Instrumental Variables (IV)



- Problem with endogeneity is correlation between X and error term ε when estimate $Y = \alpha + \beta X + \varepsilon$
  - ε a mix of random error and profit max effects (π)
- Instrumental Variable: Theory suggests a variable that causes X that is not correlated with ε or π
- 1st $X = \gamma + \theta Z + \nu$ 2nd $Y = \alpha + \beta \hat{X} + \varepsilon$ Purge X of problems
- Where do you get instruments?

Instrumental Variable Application

- Hurley and Mitchell (2016) Value of neonicotinoid seed treatments to US soybean farmers *Pest Mgmt Sci*
- Telephone survey of farmers about use and value of neonicotinoid seed treatments in soybeans
- Farmers reported average yield in 2015 and whether or not they used a seed treatment
- Do farmers with higher than average yields buy seed treatments or do seed treatments cause higher than average yields?
- Tried a couple of instruments, but testing showed endogeneity was not a problem, so OLS was fine
- Another application to try: Do farmers with higher than average yields plant cover crops or do cover crops cause higher than average yields?

Fixed Effects and Control Variables

- Perry et al. (2016) Genetically engineered crops and pesticide use in U.S. maize and soybeans *Sci Adv*
- $y_i = \alpha_{t[i]} + \beta_{t[i]} G_i + \gamma_{r[i]} T_{t[i]} + \phi_{f[i]} + e_i$
- y = pesticide use, G = GE use, T = trend, ϕ and e = errors
- Indexes: i = field, t = year, r = region (CRD), f = farmer
- Farmer-specific fixed effects $\phi_{f[i]}$
- Time-specific fixed effects $\alpha_{t[i]}$ and $\beta_{t[i]}$
- Region-specific time trends $\gamma_{r[i]}$

Summary: Econometric Methods and Observational Data

- We can use observational data from surveys to examine some of the same questions as small-plot experiments
- Empirical results are actually reasonable (comparable to small-plot experimental studies) and can achieve wide geographic coverage at lower cost than plot studies
- Need good data and appropriate analytical methods to account for the observational nature of the data
- This is good news for Big Data – we have methods to use observational data for production questions

Prediction versus Causation
Science 3 Feb. 2017

- Machine Learning (ML) algorithms for prediction are 'hot' in Big Data: *Science*: "Prediction and Its Limits"
- Major Point: Using observational data for prediction can lead to major errors, need to focus on causation too and social science has been working on this for some time
- Susan Athey, p.485 "Beyond prediction: Using big data for policy problems"

Overall, for big data to achieve its full potential in business, science, and policy, multidisciplinary approaches are needed that build on new computational algorithms from the SML literature, but also that bring in the methods and practical learning from decades of multidisciplinary research using empirical evidence to inform policy. A nascent but rapidly growing body of research takes this approach: For ex-

Where is Agricultural Economics going in terms of Ag Big Data?

1) Methods development and improvement
   - Need to develop empirical methods with solid theoretical foundations that link machine learning with the econometrics of causal inference
   - We need both prediction and causation
2) Policy questions: What is the effect of …
   - GE crops on pesticide use?
   - Crop rotation on yield?
3) Farm Management recommendations
   - How do we use big data to make management recommendations to improve outcomes on farms?

A Possible Example: Discovery Farms Observational Data

- Yuji Saikai, Matt Ruark, Rebecca Willett
- Edge of field runoff for many fields over years – sediment and nutrient losses (N and P) in water:
  - By day, event, month, season, year
- Farmer management: crop history; method, timing and rate for tillage, fertilizer and manure applications; conservation practices …; plus yield
- Weather data: rainfall, temperature, wind, humidity, …
- Soil parameters: soil series, SOM, slope, …
- **What are the relative contributions of farmer management, weather, and soil factors to soil erosion, nutrient loss and yield?**

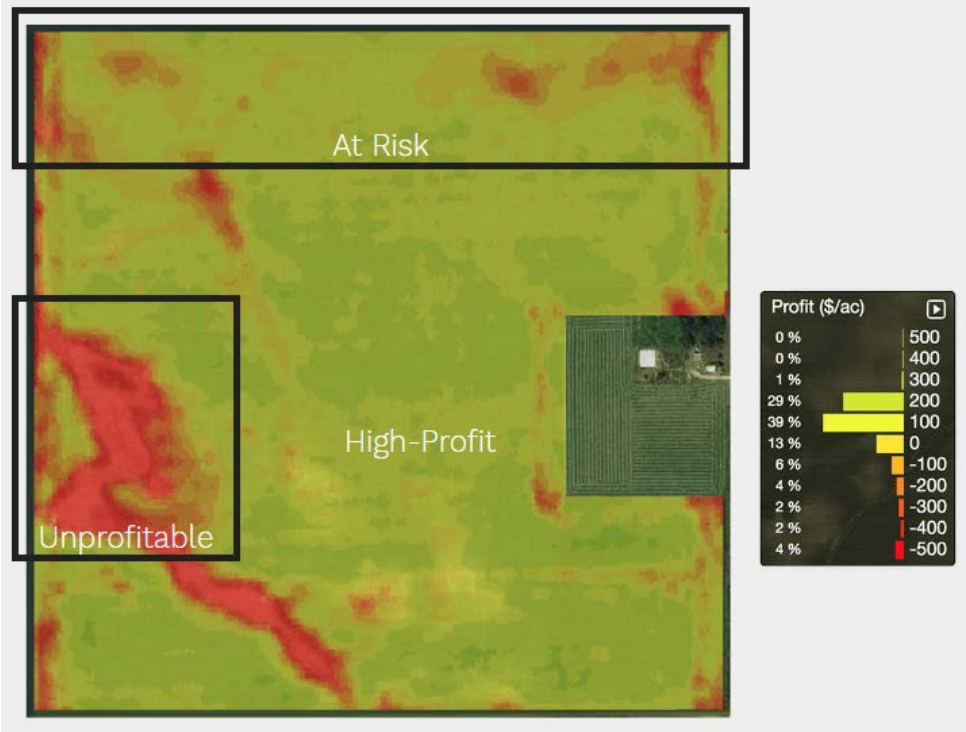
Farm Management Applications

- Ag Funder Ag Tech Investing Report: $1.4 Billion invested in Ag Big Data in 2015
- Lots of private efforts to figure out how to make money applying big data to farm management
- C-FARE Big Ag Data Report: Potential Opportunities
  - Improve farm management, Track food safety, and Enhance sustainability
- Keith Coble: Farm management has become "sexy"
- What will the Big Data Ag Economist look like? (Coble)
  - Work in multidisciplinary teams
  - Can distinguish causation from correlation
  - Trained to use machine learning, to work with less structured and geo-spatial data, and to clean data

AgSolver: https://agsolver.com/

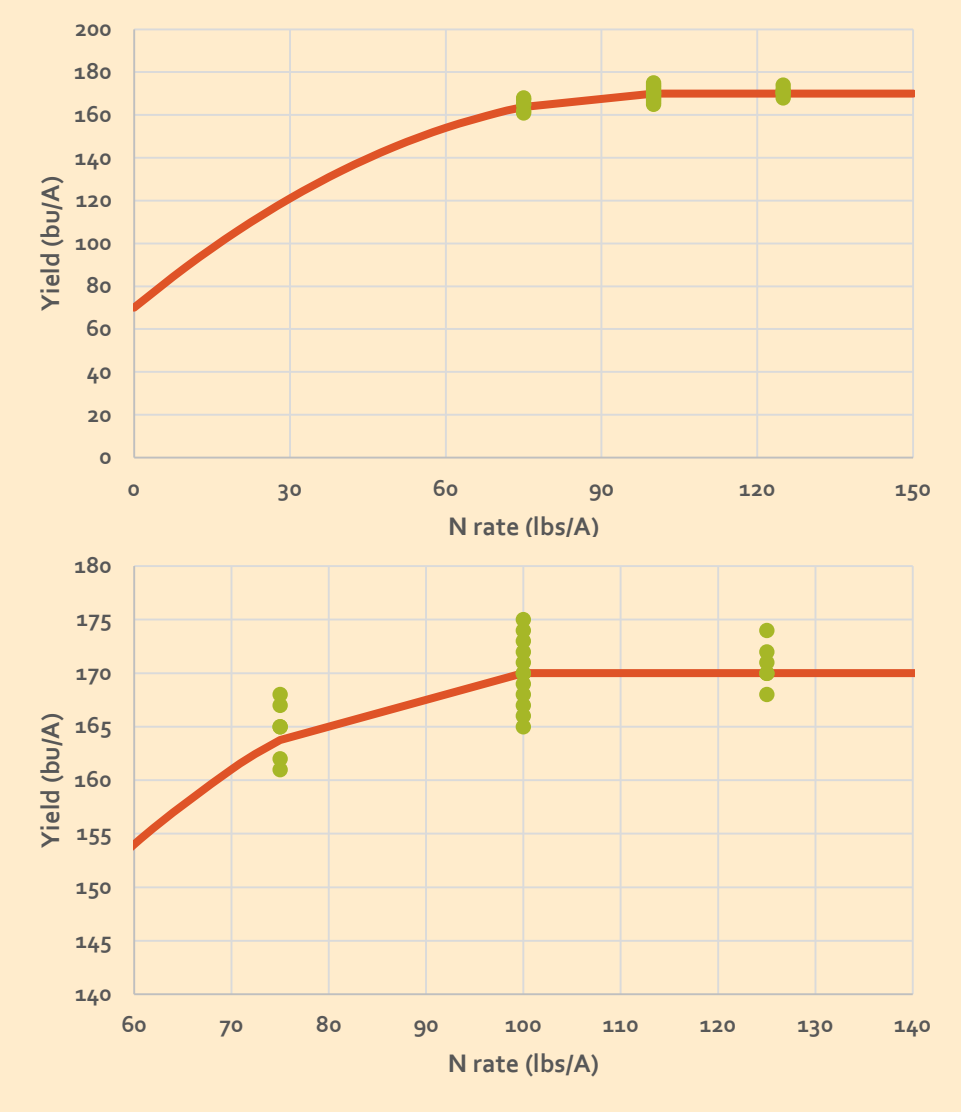
3%-15% of all acres are consistently unprofitable

AgSolver can help you identify these acres and find alternative management techniques to increase your total profit.



Need to Operationalize On-Farm Experimentation

- On-farm experiments needed to trace out the response curve, at least around the optimum
- If N* = 100 lbs/A is your "optimal" rate, then most of the field is at 100 lbs/A, but have six "plots" with ±25 lbs/A to estimate the N response curve
- Link fertilizer program with yield monitor, then automate estimation and updating the fertilizer program



Human Behavior and Prediction versus Causation with Big Data

- Big data has and will make a lot of data available that are observational, not experimental
- Not accounting for the observational nature of the data can lead to serious prediction errors, especially if they are generated by human behavior
- Making more accurate predictions with observational data requires methods that identify causation
- Social scientists have developed several such methods
  - "Natural" experiments
  - Structural models of the decision making process
  - Panel data methods: instrumental variables, fixed effects

Human Behavior and Prediction versus Causation with Big Data

- Application of these methods by ag economists to agricultural production and policy questions remains somewhat limited
- Economists will likely have to learn machine learning and data cleaning algorithms rather than the reverse
- We need multidisciplinary teams to develop methods that merge machine learning algorithms with these social science methods
- Only then can we achieve the potential for Ag Big Data to improve the sustainability of farming

THANKS FOR YOUR ATTENTION

Paul D. Mitchell

Professor, Ag and Applied Economics and UW Extension

Director, Renk Agribusiness Institute, UW-CALS

Co-Director, Nutrient & Pest Management Program, UW-Ex

Big Data and Ecoinfomatics in Agricultural Research

University of Wisconsin April 27, 2017

608-265-6514, pdmitchell@wisc.edu, @mitchelluw

NPM

RENK AGRIBUSINESS INSTITUTE
College of Agricultural & Life Sciences

UW Extension
University of Wisconsin-Extension

College of Agricultural & Life Sciences
UNIVERSITY OF WISCONSIN-MADISON