Spark for Behavior Analysis Research

Ling Jin, Sam Borgeson, Anna Spurlock, Annika Todd
Doris Lee, Alex Sim, John Wu

Lawrence Berkeley National Lab

SPARK SUMMIT EAST 2017

Bringing Science Solutions to the World

In the world of science, Lawrence Berkeley National Laboratory (Berkeley Lab) is synonymous with "excellence." Thirteen Nobel prizes are associated with Berkeley Lab. Seventy Lab scientists are members of the National Academy of Sciences (NAS), one of the highest honors for a scientist in the United States. Thirteen of our scientists have won the National Medal of Science, our nation's highest award for lifetime achievement in fields of scientific research. Eighteen of our engineers have been elected to the National Academy of Engineering, and three of our scientists have been elected into the Institute of Medicine. In addition, Berkeley Lab has trained

BERKELEY LAB VALUES

Overarching commitment to pioneering science
Highest integrity /impeccable ethics
Uncompromising safety
Diversity in people and thought
Sense of urgency

THE LAB AT A GLANCE

- 13 Nobel Prizes
- 15 National Medal of Science recipients
- 1 National Medal of Technology and Innovation recipient
- $700 Million Contributed to the local economy annually
- 3,304 Employees
- 202 Site acreage

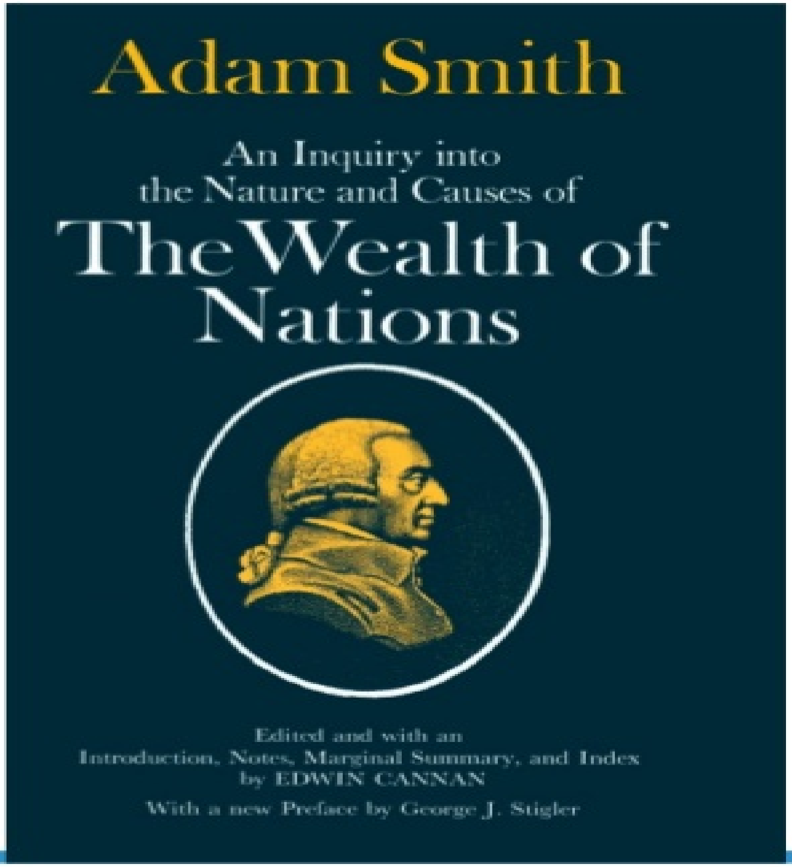
LAB BUDGET

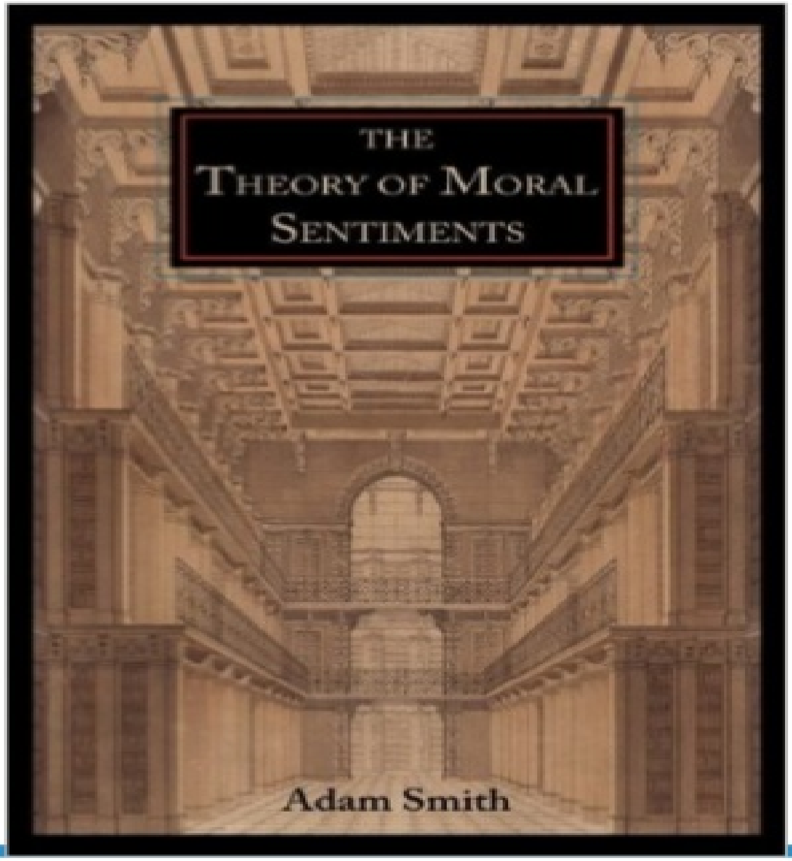
- FY 2015 – $811 million

BERKELEY LAB

www.lbl.gov

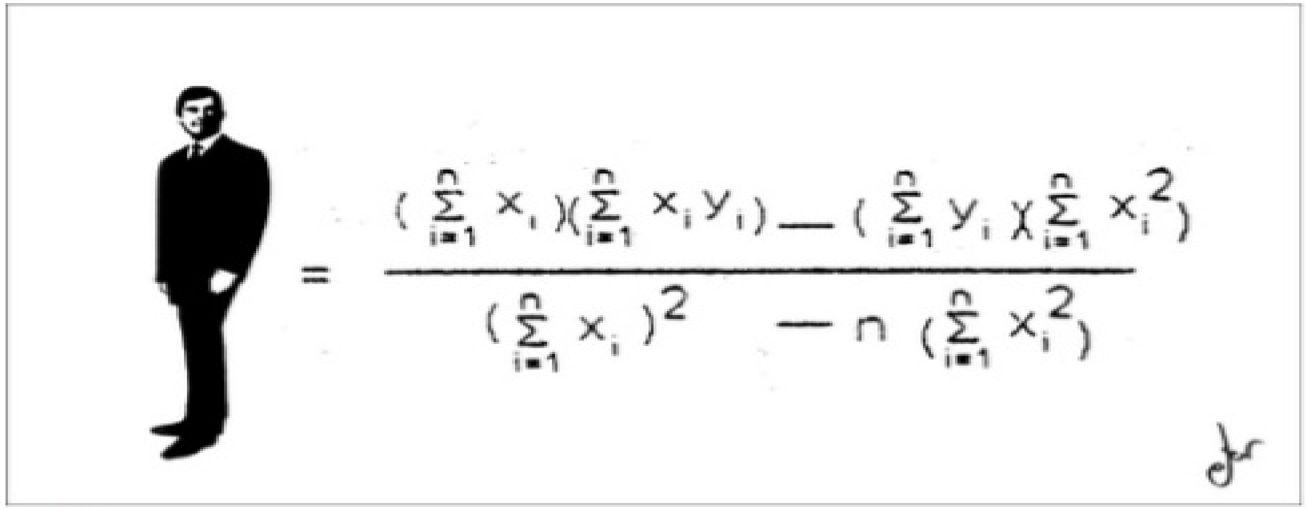
Adam Smith’s Most Famous Books?



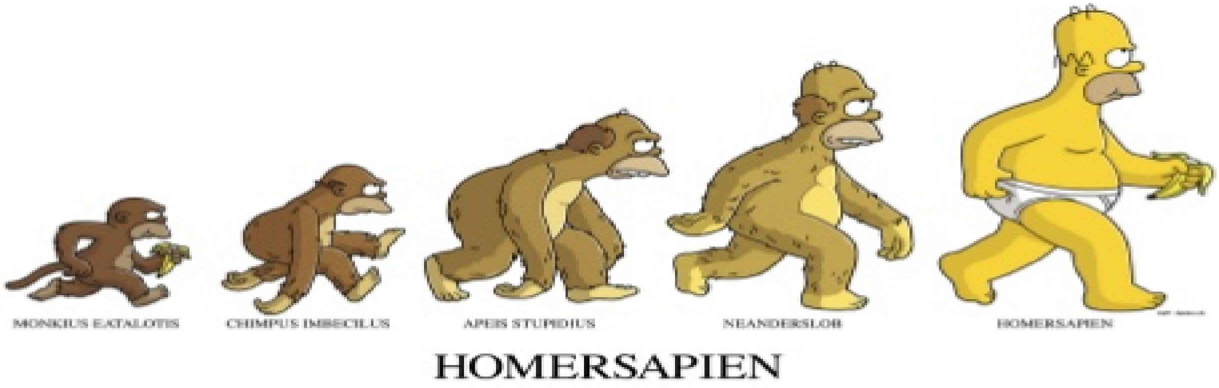


One-minute Behavioral Economics

Simpler to analyze and predict



Harder to predict -- Need massive data and computers



How Are Blackjack and Electric Power Grid Similar?

Bust when demand is larger than supply

2003 North East Blackout: OOPS!!

Bust when over 21

Reducing Peak Demands Through Pricing

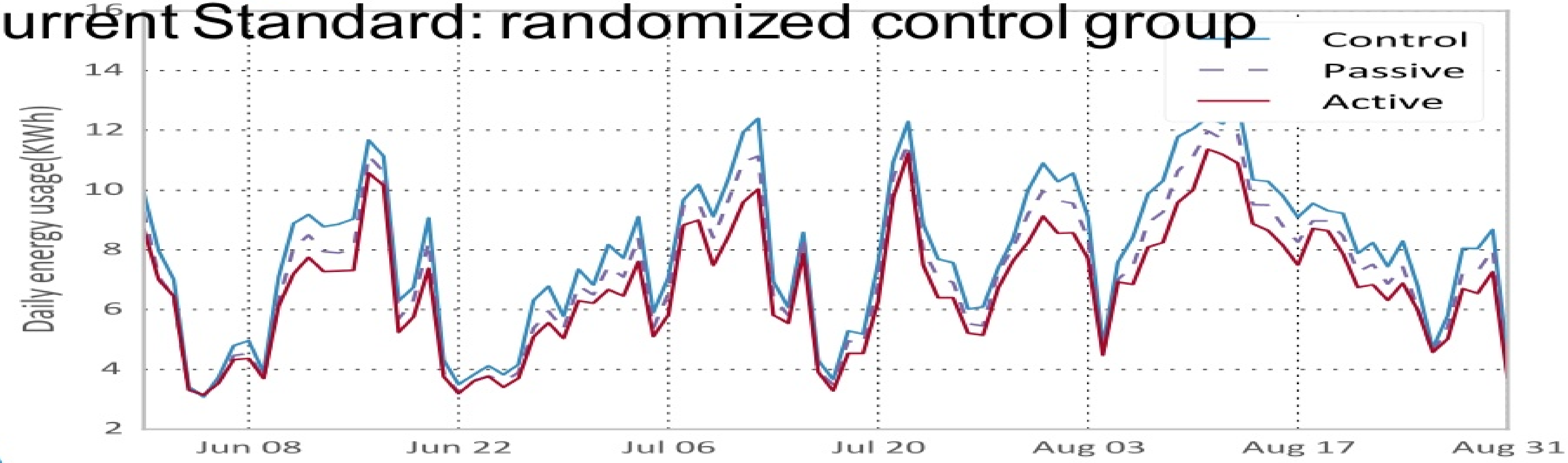
- Residential electricity records: 100,000 households, hourly electricity usage
- A region where electricity usage peaks in the summer time
- Randomized controlled trial of new rate structures - households are randomly placed in different groups
    - Control: using existing fixed rate for electricity
    - Active: households opted in to Time-of-Use Pricing
    - Passive: households defaulted in to Time-of-Use Pricing
- Data collected hourly over three years, one pre-rate, two after (labeled T-1, T, T+1)

Research Examples

Goal	Method	Policy Implication
Better baseline models of energy use	Gradient tree boosting	**Better program evaluation**
Define relevant household characteristics:		**Classify and segment households using easily accessible data**
• Define representative load shapes	Adaptive K-means clustering	
• Estimate household-specific cooling change points (AC set point)	Piecewise linear regression, bootstrapping	
• Characterize customers into "Lifestyle Groups"	Blend behavioral theory with machine learning techniques	
• Define relevant household energy characteristics	Simple feature algorithms (e.g., mean, min, max, peak usage; variance; entropy, etc.)	

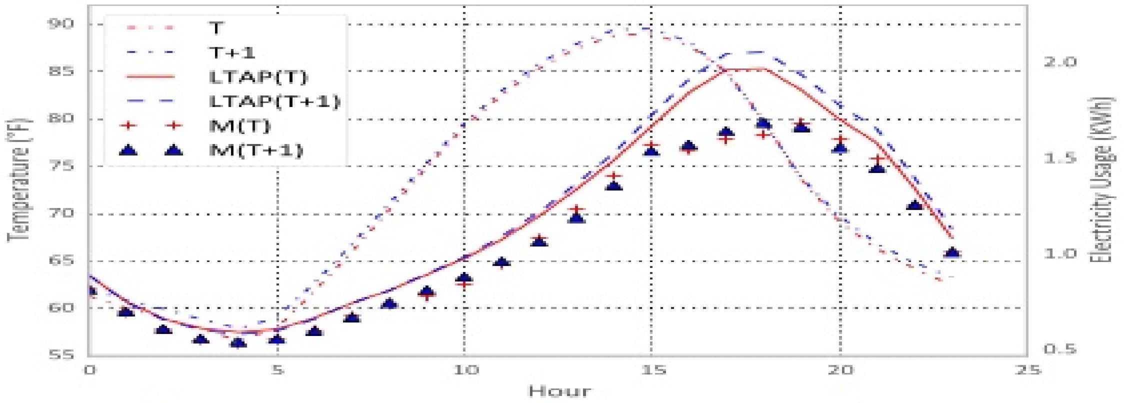
Baselines are Critical for Measuring Changes

Current Standard: randomized control group

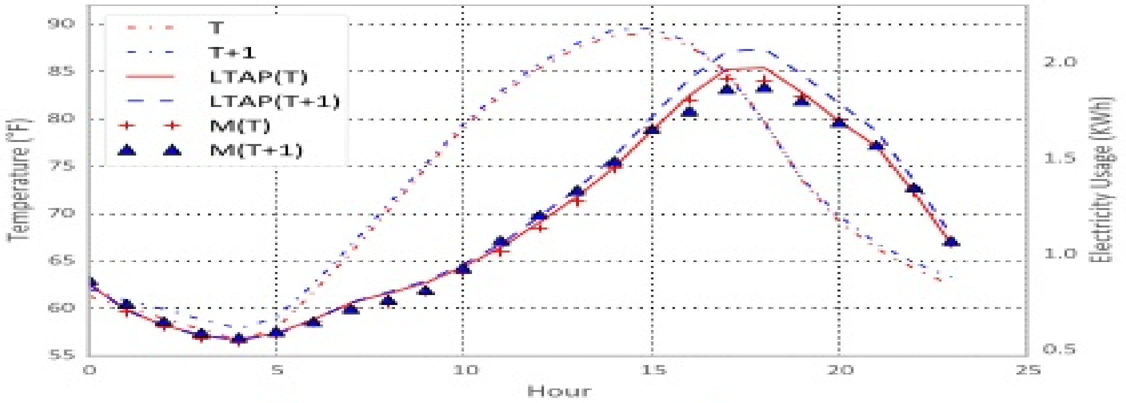


Predictions from new Baseline Technique

Active group



Passive group



Group	P_T	P_{T+1}	$M_T - P_T$	$M_{T+1} - P_{T+1}$
Control	1.960	1.957	0.069	-0.020
Passive	1.849	1.897	-0.027	-0.080
Active	1.860	1.903	-0.164	-0.164

P: predicted
M: measured

Observation: the Active group reduced their uses of electricity consistently over the two yeas of study during the peak-demand hours

Research Examples

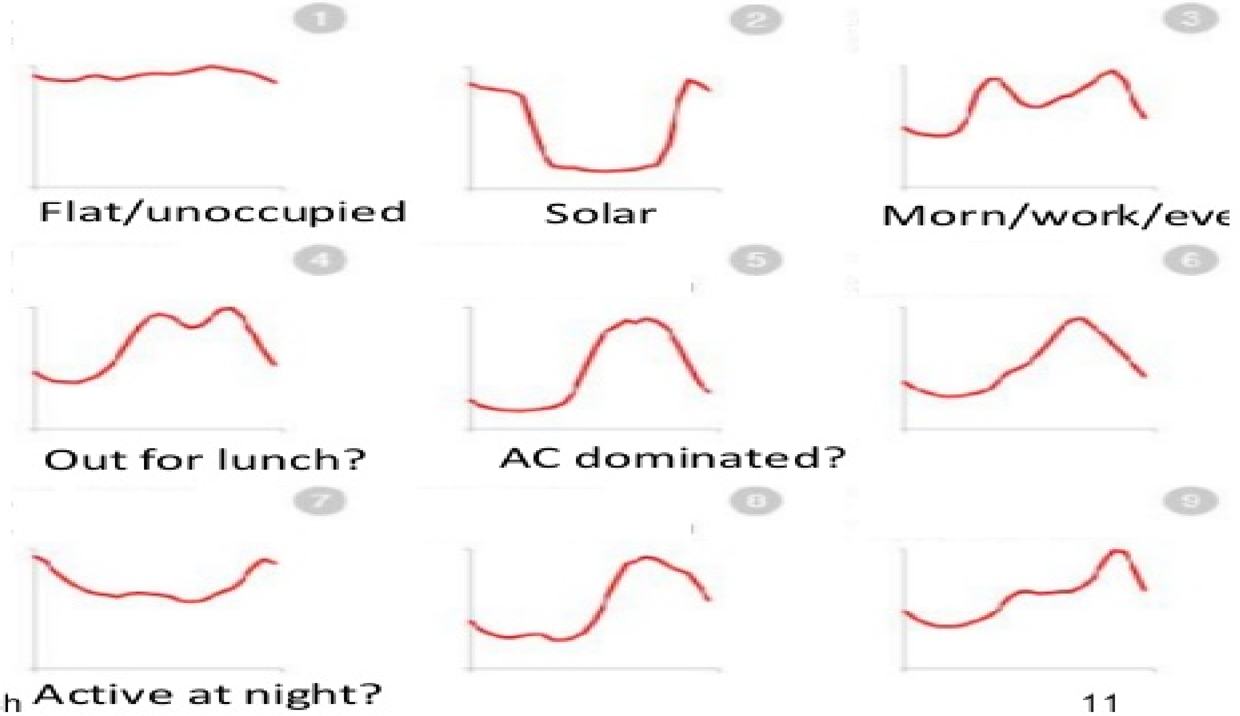
Goal	Method	Policy Implication
Better baseline models of energy use	Gradient tree boosting	**Better program evaluation**
Define relevant household characteristics:		**Classify and segment households using easily accessible data**
• Define representative load shapes	Adaptive K-means clustering	
• Estimate household-specific cooling change points (AC set point)	Piecewise linear regression, bootstrapping	
• Characterize customers into "Lifestyle Groups"	Blend behavioral theory with machine learning techniques	
• Define relevant household energy characteristics	Simple feature algorithms (e.g., mean, min, max, peak usage; variance; entropy, etc.)	

Daily Load Shape: Definition

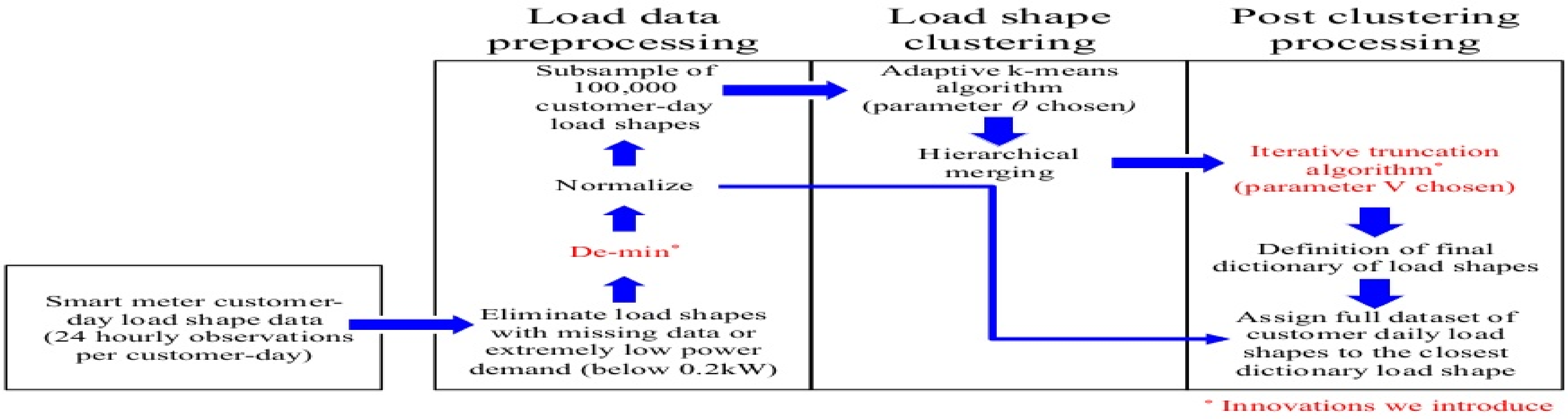
Objective / Definition

- 24-hour electricity usage pattern
- To capture hour variations: don't care about constant usage
- Want to use load shape to study mechanism that might affect electricity usage, therefore concentrate on discretionary demand

Samples



Clustering Process



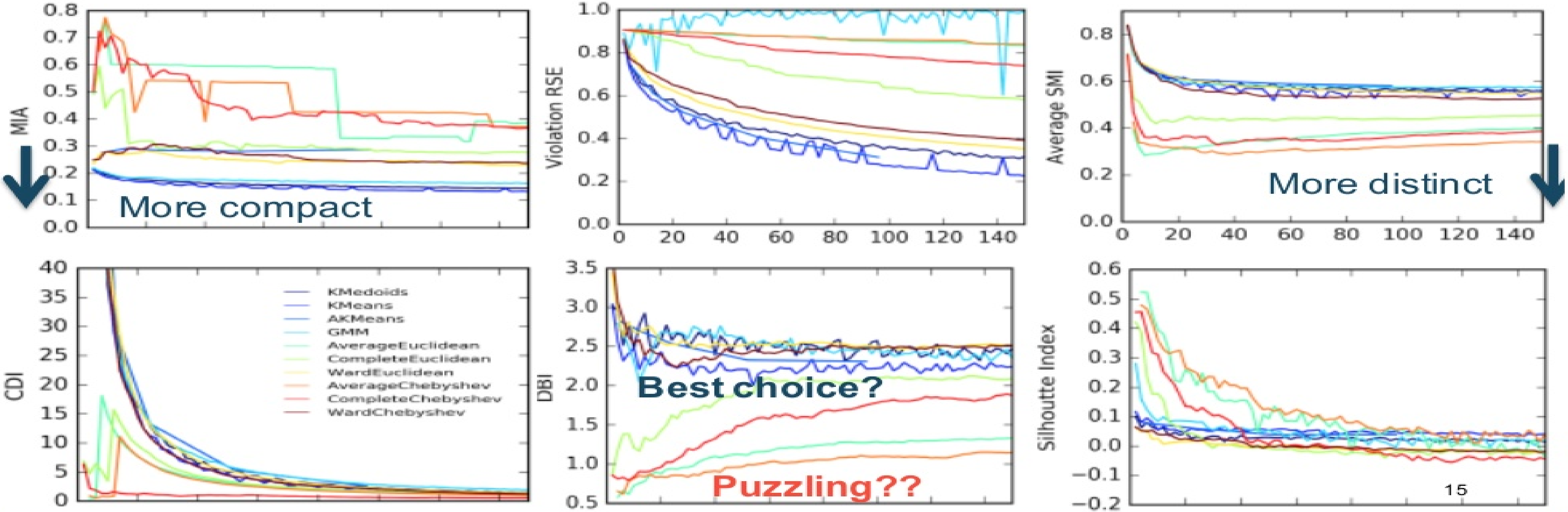
Components of Clustering Process

- **Data cleaning and normalization**
- Remove households with very low demand (<0.2kW)
- Normalization: compute hourly usage to hourly contribution to daily usage
- **Clustering**
- Centroid-based methods: Kmeans, Adaptive Kmeans
- Hierarchical clustering: distance metric, linkage
- Density-based clustering: DBSCAN
- Model-based clustering: GMM
- **Judging cluster quality**
- Compactness: MIA, VRSE
- Distinctness: SMI
- Combined: DBI, Silhouette

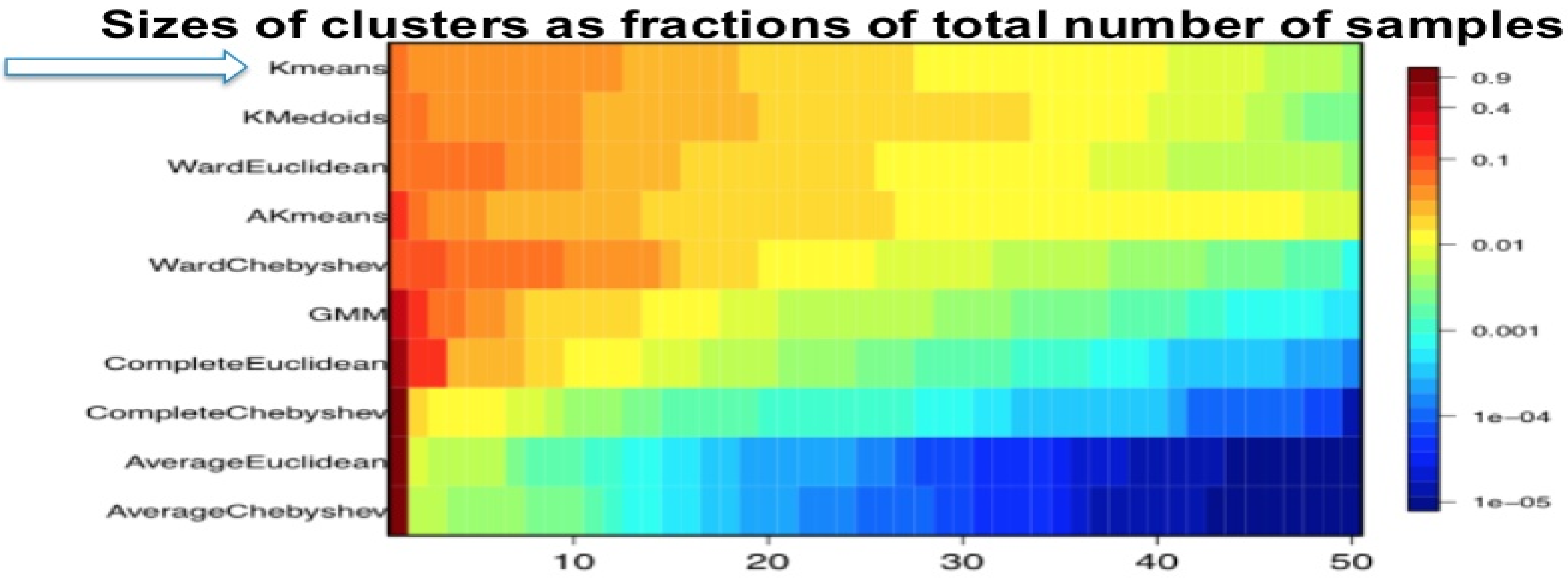
Clustering Quality Index

	Equation	Description	Measure
CDI	$CDI = \frac{1}{\bar{d}(C)} \sqrt{\frac{1}{K} \sum \hat{d}^2(R_k)}$	Cluster Dispersion Indicator	compactness and distinctness
MIA	$MIA = \sqrt{\frac{1}{K} \sum_{K=1}^{K} d^2(r^{(k)}, C^k)}$	Mean Index adequacy	compactness
Silhouette	$SIL = \frac{max(a,b)}{b-a}$ where a = average intra-cluster distance. b = average shortest distance to another cluster	Inverse Silhouette index $\epsilon[-1, 1]$. If SIL¡0, cluster not very compact. Note that this is the inverse of typical definitions of SIL in literature.	compactness and distinctness
Average SMI	$\alpha_{i,j} = \frac{1}{1 \quad \frac{1}{(ln[d(C_i, C_j)]}}$ $< SMI >= \frac{1}{N} \sum_i \sum_j \alpha_{i,j}$	Similarity Matrix Indicator generates a KxK matrix, where K is number of cluster. The farther away the non-diagonal elements are from 1 the better. quantify the measure, averaging over the whole matrix gives us a sense of how non-diagonal elements behave.	distinctness
DBI	$DBI = \frac{1}{K} \sum max \frac{scatter(C_i)+scatter(C_j)}{d(C_i, C_j)}$ where $i \neq j$	Davies-Boulden indicator	compactness and distinctness
VRSE	$E(s, i*(s)) = \sum(s(t) - C_i^*(t))^2 \leq \theta \sum C_{i^*_{(s)}}(t)^2$	Violation rate of RSE threshold. Percentage of data that lie beyond a threshold distance away from the . centroid. With $\theta = 0.3$	compactness

Clustering Quality



Kmeans Produces More Balanced Clusters



Summary

- Examined a good number of clustering methods for identifying daily usage profiles
- Short observation:
  - Centroid-based method (Kmeans) works the best on this set of data
- Longer observations:
  - There are too many choices and no (really) clear winner
  - Centroid-based methods produce more compact clusters
  - Centroid-based methods produce more balanced cluster
  - DBSCAN declare many (up to 90%) data points as background
  - Our observation is different from previous published results (Chicco 2012)
- Future work
  - Maybe a different clustering quality metric would work better?
  - Should examine alternative profile generation methods, other than clustering

Thank You.

John Wu
Email: John.Wu@nersc.gov

SPARK SUMMIT EAST 2017