Stanford John S. Knight Journalism Fellows May 10, 2017

“Datafication” — the rise of big data and the application of AI to everything

Hugh Dubberly
Dubberly Design Office

presentations.dubberly.com/datafication.pdf

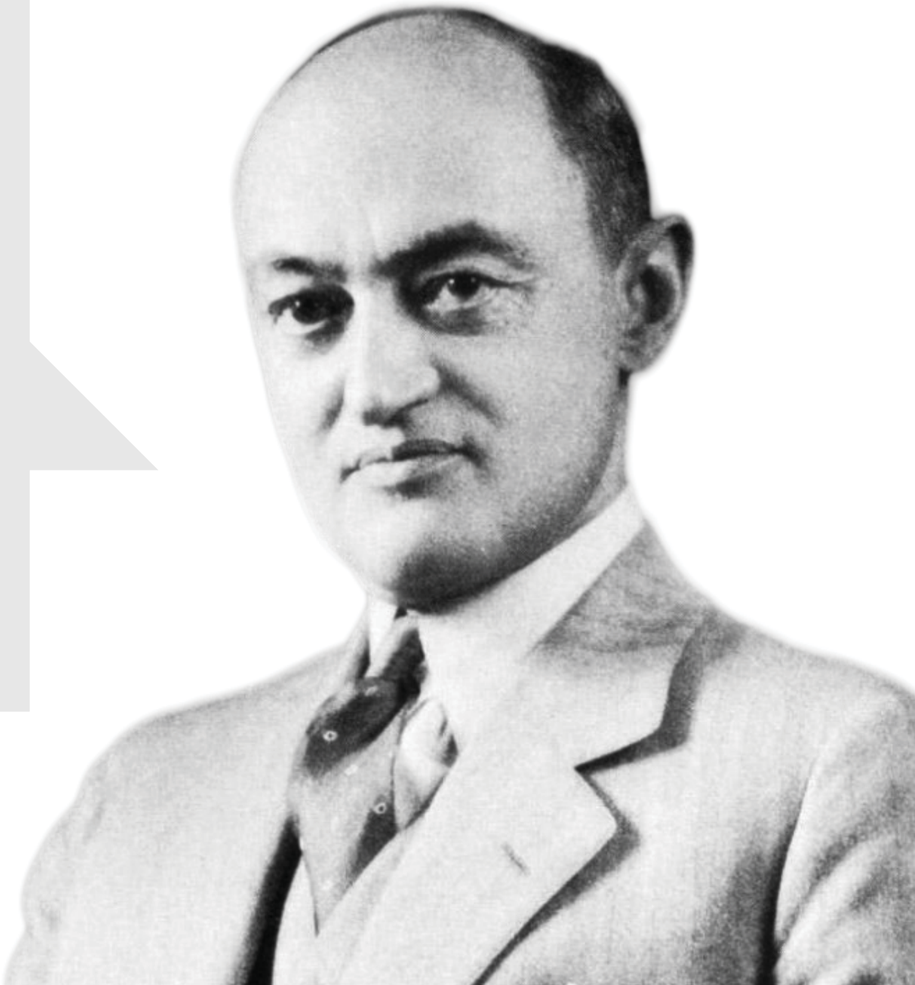
“Creative Destruction is the essential fact about capitalism.”

“Capitalism, then, is by nature a form or method of economic change and not only never is but never can be stationary....

The fundamental impulse that sets and keeps the capitalist engine in motion comes from the new consumers’ goods, the new methods of production or transportation, the new markets, the new forms of industrial organization that capitalist enterprise creates....

The opening up of new markets, foreign or domestic, and the organizational development from the craft shop and factory to such concerns as U. S. Steel illustrate the same process of industrial mutation— if I may use that biological term— that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one.

This process of Creative Destruction is the essential fact about capitalism.”



—Joseph A. Schumpeter, (1942) *Capitalsim, Socialism and Democracy*, pgs 82-83.

We have seen five industrial revolutions; what will be the sixth?

1
1780 - 1840
First industrial revolution
- iron making
- cotton spinning
- water power
- steam power

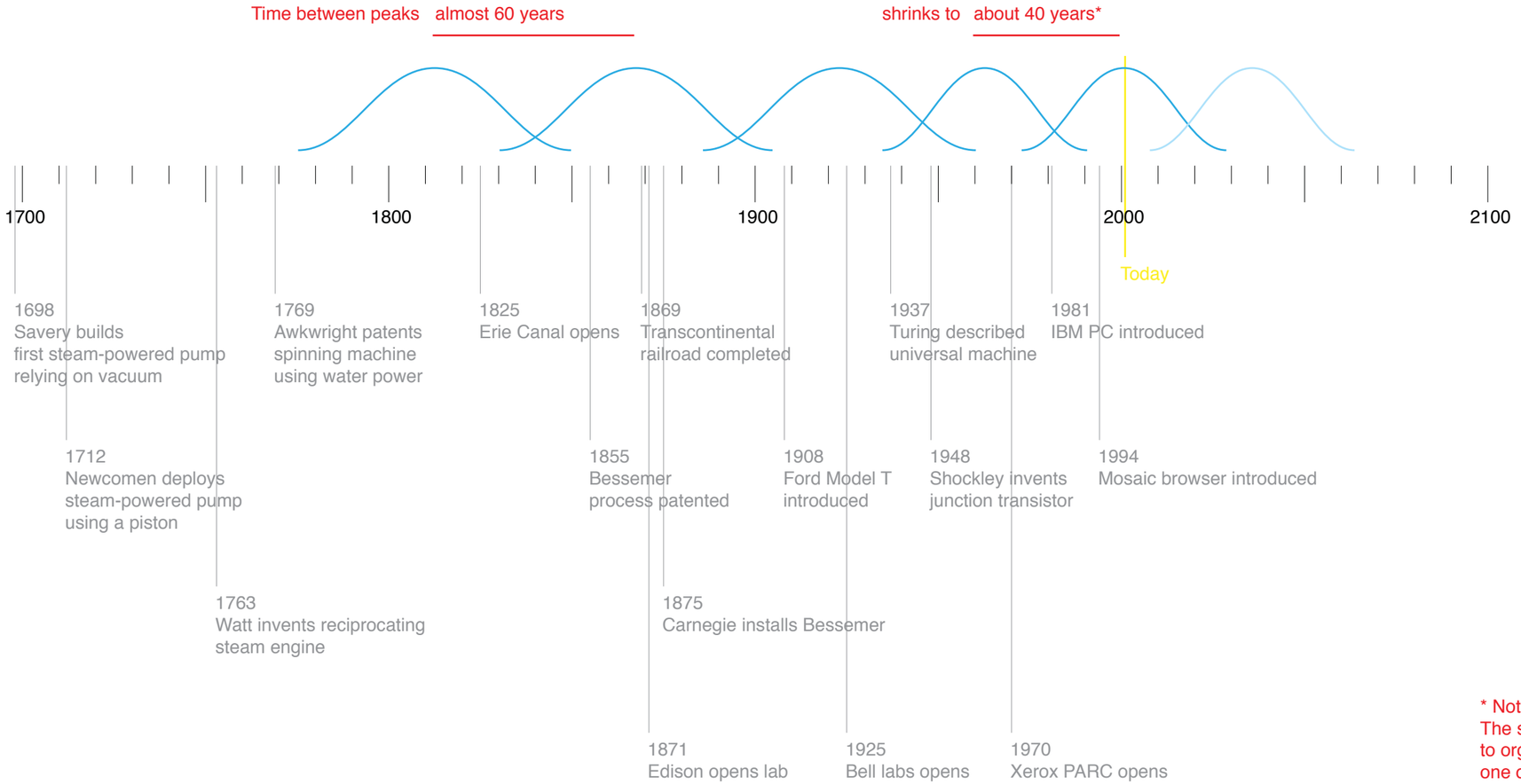
2
1840 - 1900
Second
- steel making
- railroads

3
1900 - 1950
Third
- electrification
- internal combustion

4
1940 - 1980
Fourth
- petrochemicals
- plastics
- electronics
- computing

5
1980 -2020
Fifth
- personal computers
- internet

6
What comes next?



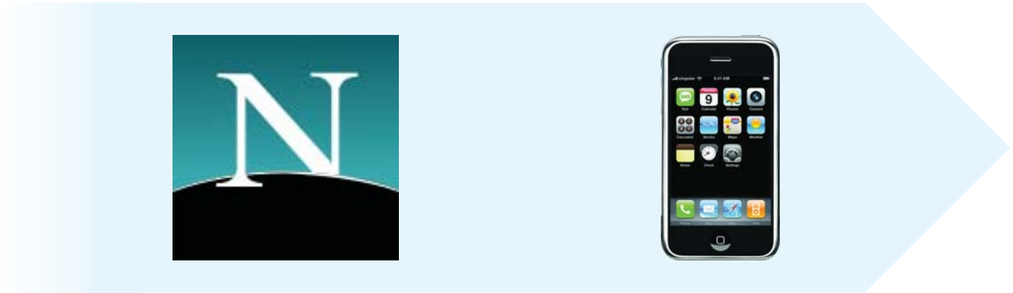
* Note:
The shrinking cycle time may be due to organized research and possibly to one or more “network effects.”

Google CEO Sundar Pichai has predicted “AI First”.

PC Web Mobile “AI First”

Google CEO, Sundar Pichai spoke at the #MadeByGoogle event on October 4, 2016

Siri co-founder Dag Kittlaus is focused on “assistants”.



Web Mobile Assistants

Siri co-founder Dag Kittlaus unveiled Viv
at TechCrunch Distrupt NY 2016

In each era, the dominant technology is a “platform”—a system on which others can build.

Productivity Applications	Web-based Services	Mobile Apps	Monitoring + Prediction Services
			
PC	Web	Mobile	AI

In the early 1980s,
personal computers changed the way business is done.
Think of this as *going digital*; everything is becoming a computer.

> *"As products and the means to create them have become digitized (often referred to as software eating the world), production capability has grown more accessible and portable. And the acceleration of that trend (driven by Moore's Law) means that every single day it gets easier for someone else to compete with your product or service, and to do it better, faster, and cheaper."*

— Aaron Dignan, Undercurrent

In the mid-1990s,
the internet changed the way consumers + business communicate.

Think of this as *getting connected*; everything becomes a web service.

"Millennials don't just want to buy your brand, they want to be part of it. They're looking for ways to participate."

— Jeff Fromm, Barkley

"I envision a 21st century form of business where the everyday consumer is helping shape the social contract … It's a business world that is moving from value-based transactions to values-based partnerships."

— Paul Polman, CEO, Unilver

In 2007,
smartphones made computing ubiquitous–and turned it into communicating.

Think of this as *always connected*; anywhere, anytime.

The iPhone and iPad began to fulfill the vision of the "Dynabook."

"A Personal Computer for Children of All Ages

What we would like to do in this brief note is to discuss some aspects of the learning process which we feel can be augmented through technological media…

We do not feel that technology is a necessary constituent for this process any more than is the book. It may, however, provide us with a better "book", one which is active (like the child) rather than passive. It may be something with the attention grabbing powers of TV, but controllable by the child rather than the networks. It can be like a piano: (a product of technology, yes), but one which can be a tool, a toy, a medium of expression, a source of unending pleasure and delight… and, as with most gadgets in unenlightened hands, a terrible drudge!

This new medium will not 'save the world' from disaster. Just as with the book, it brings a new set of horizons and a new set of problems. The book did, however, allow centuries of human knowledge to be encapsulated and transmitted to everybody; perhaps an active medium can also convey some of the excitement of thought and creation!"

—Alan Kay, 1972

Today is like 1981, 1995 and 2007 all over again.
You can see the next wave coming. It goes by many names:

- Internet of Things (IoT)
- Internet of Everything, Cisco
- Industrial Internet, GE
- Smarter Planet, IBM
- Living Services, Accenture
- Platform World, Sapient.Publicis
- Social CRM or Social Business
- Digital Engagement
- Digital Transformation
- “Datafication”

The eras Pichai + Kittlaus describe can be seen as “waves”, with several trends interacting.

Personal Computers
1975

The Internet
1995

Mobile
2007

“Datafication”
2015

SaaS

Cloud computing
on-demand

Social

Glasses
Goggles
Wearables

CNC
3D printing

Robots
Self-driving cars
Drones

Big data
Linked data
Open data

Virtual reality
Augmented reality

Conversational UI/UX

Chatbots

Predictive analytics
AI 2.0
Machine learning
Deep learning
Computer vision
Natural language processing

Combinatorial innovation explains how trends work together.

"We're in the middle of a period of... 'combinatorial innovation'...
In the 1800's, it was interchangeable parts.
In 1920, it was electronics. In the 1970s, it was integrated circuits.
Now what we see is a period where you have Internet components...
and capabilities to combine these components parts
in ways that create totally new innovations."

—Hal Varian, Google's Chief Economist and UC Berkeley Professor

“Datafication” is a series of trends; none capture the whole.

- **Sensor Revolution** — printing sensors on chips; installing measurement capability all around us.

- **Smart Things** — adding “intelligence” to everything, by building in microprocessors.

- **Internet of Things (IoT)** — connecting sensors and smart things to the cloud.

- **Big Data** — recording everything that happens in the physical world and online.

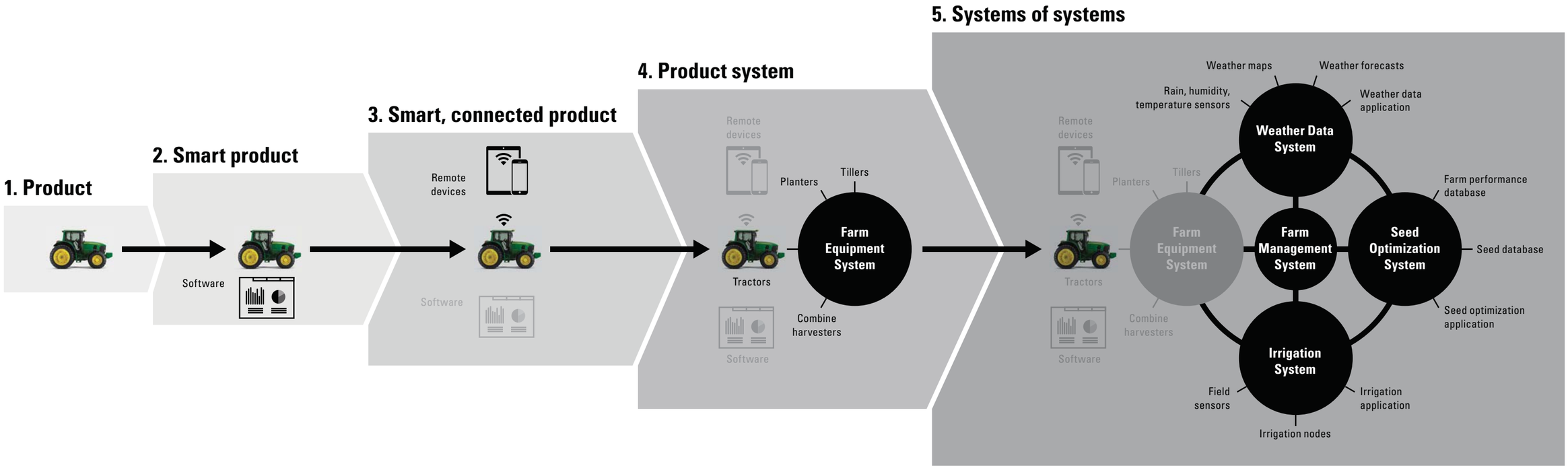
- **Cloud Computing** — putting massive resources online, so that the marginal cost of computation falls to zero.

- **AI, ML, DL, NLP, CV** — algorithms (often run in the cloud), making sense of the measurements we record.

“Datafication”

An example

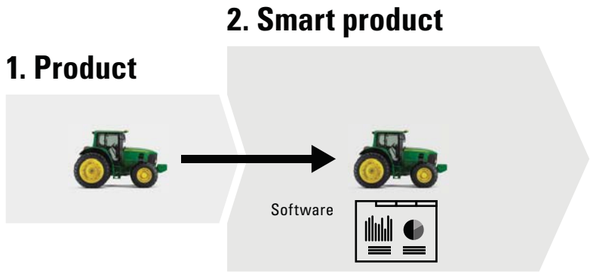
Harvard Business School professor Michael Porter writes about systems of systems.



—Michael Porter and James Heppelmann, How Smart, Connected Products Are Transforming Competition
Harvard Business Review, November 2014
https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

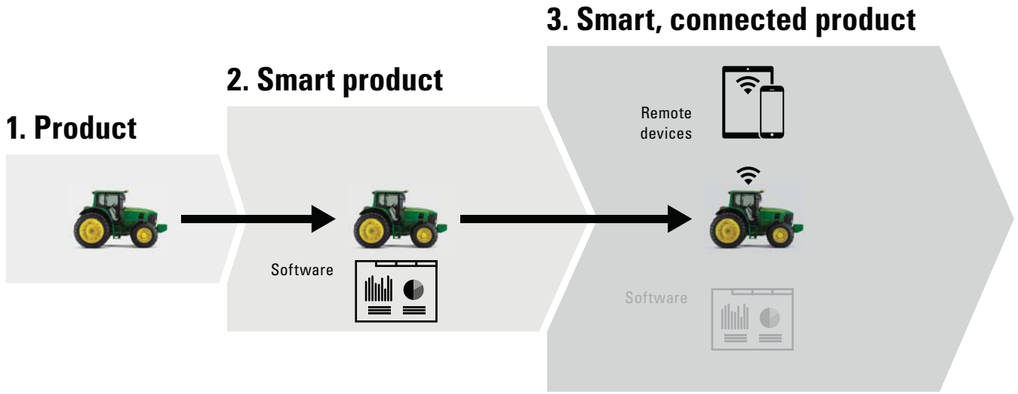
Products are becoming “smart.”

Product
+ Sensor
+ Computer
= Smart Product



Sets of smart products are connecting.

Smart Product
+ Network
+ Cloud Service
= Smart, Connected Product

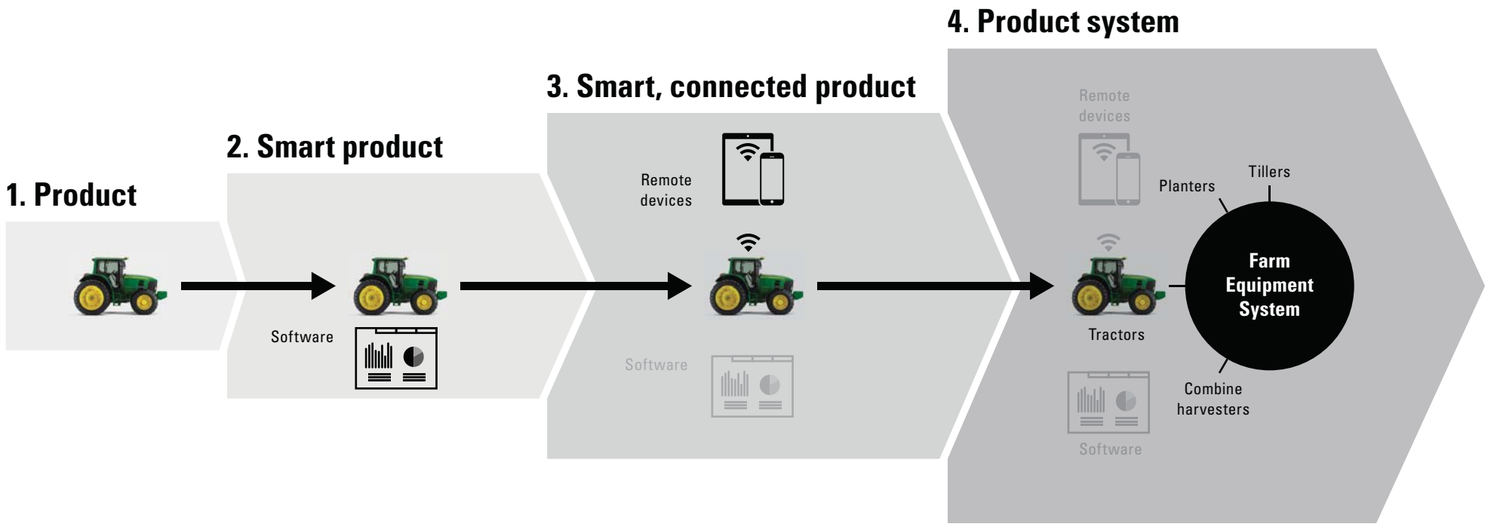


Sets of connected products form product systems.

Smart, Connected Product

\+ other Smart, Connected Products

= Product System

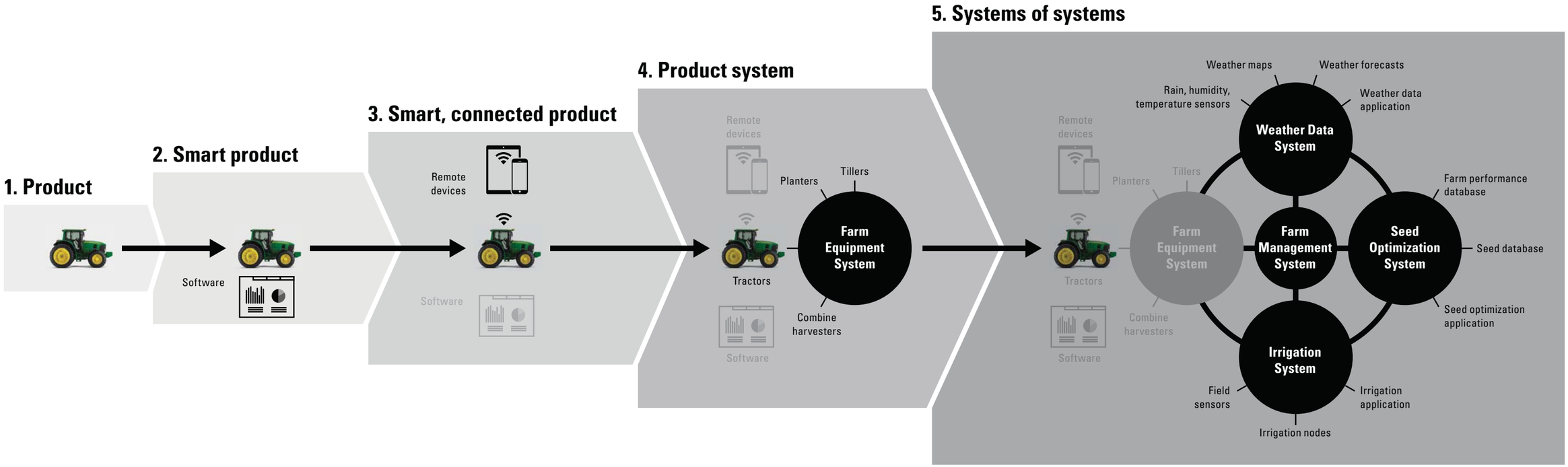


Systems connect to other systems, forming ecologies.

Product Systems

+ other Product Systems

= Product-Services Ecology

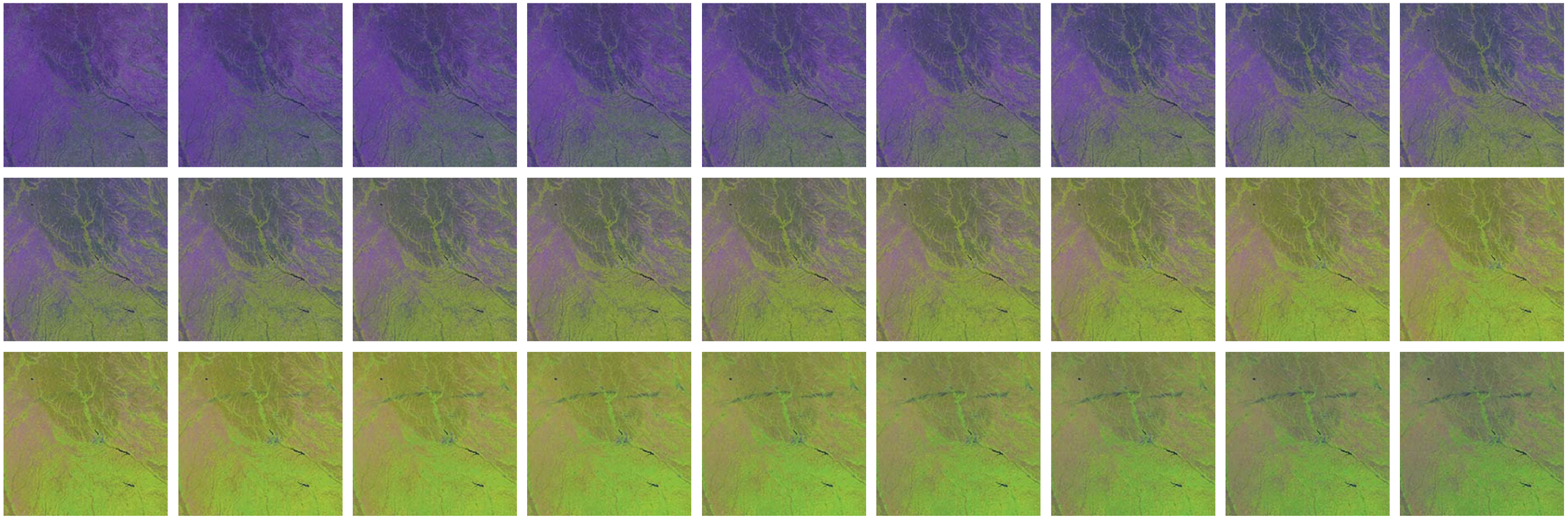


Farms are becoming automated factories.

Plants are attached to sensors, connected to networks, generating data.

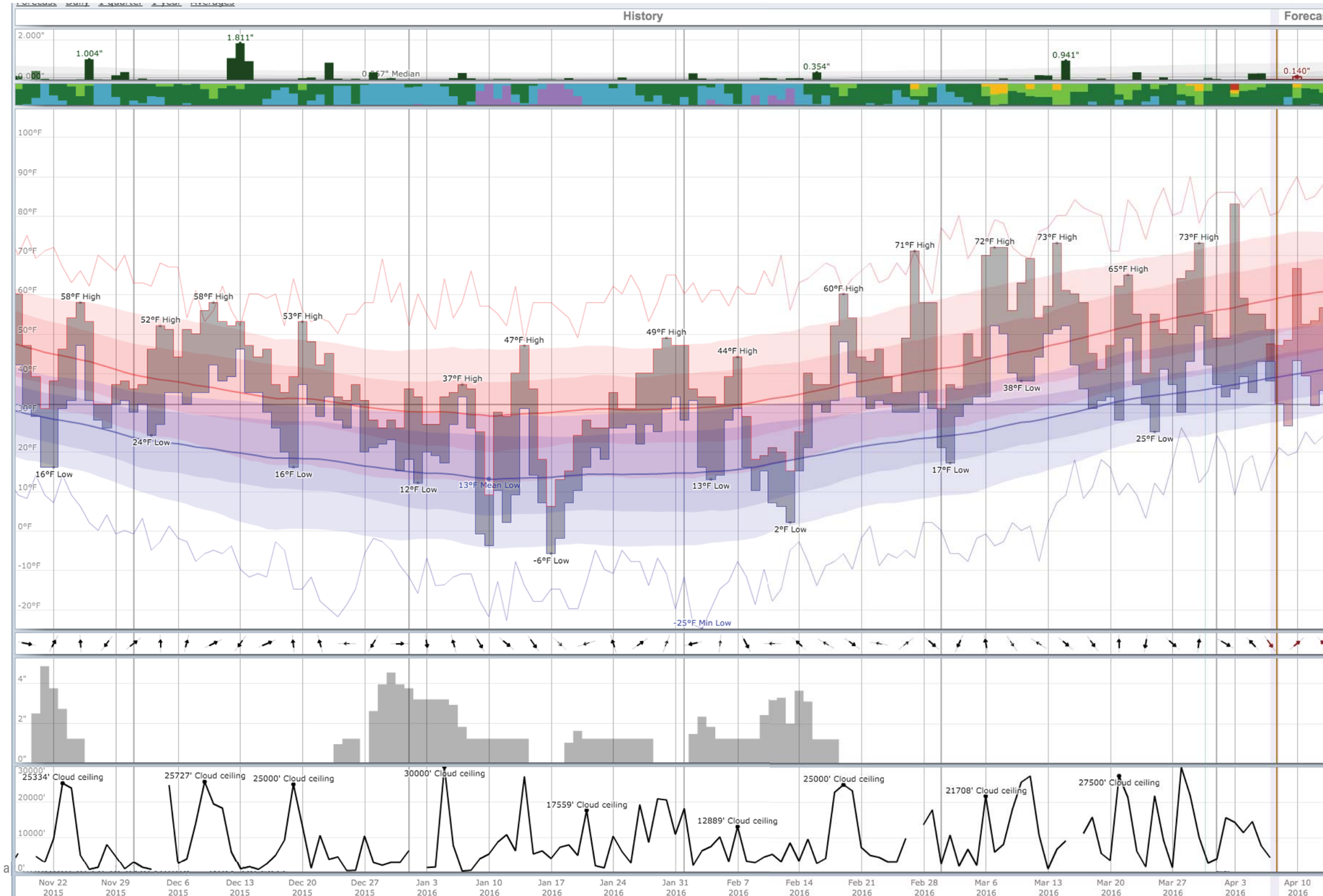
Macro view: processed satellite images of crop growth over time, e.g., central Iowa, March 29 to October 23, in 8 day increments.

Algorithms automatically align images,
remove clouds,
and detect vegetation.

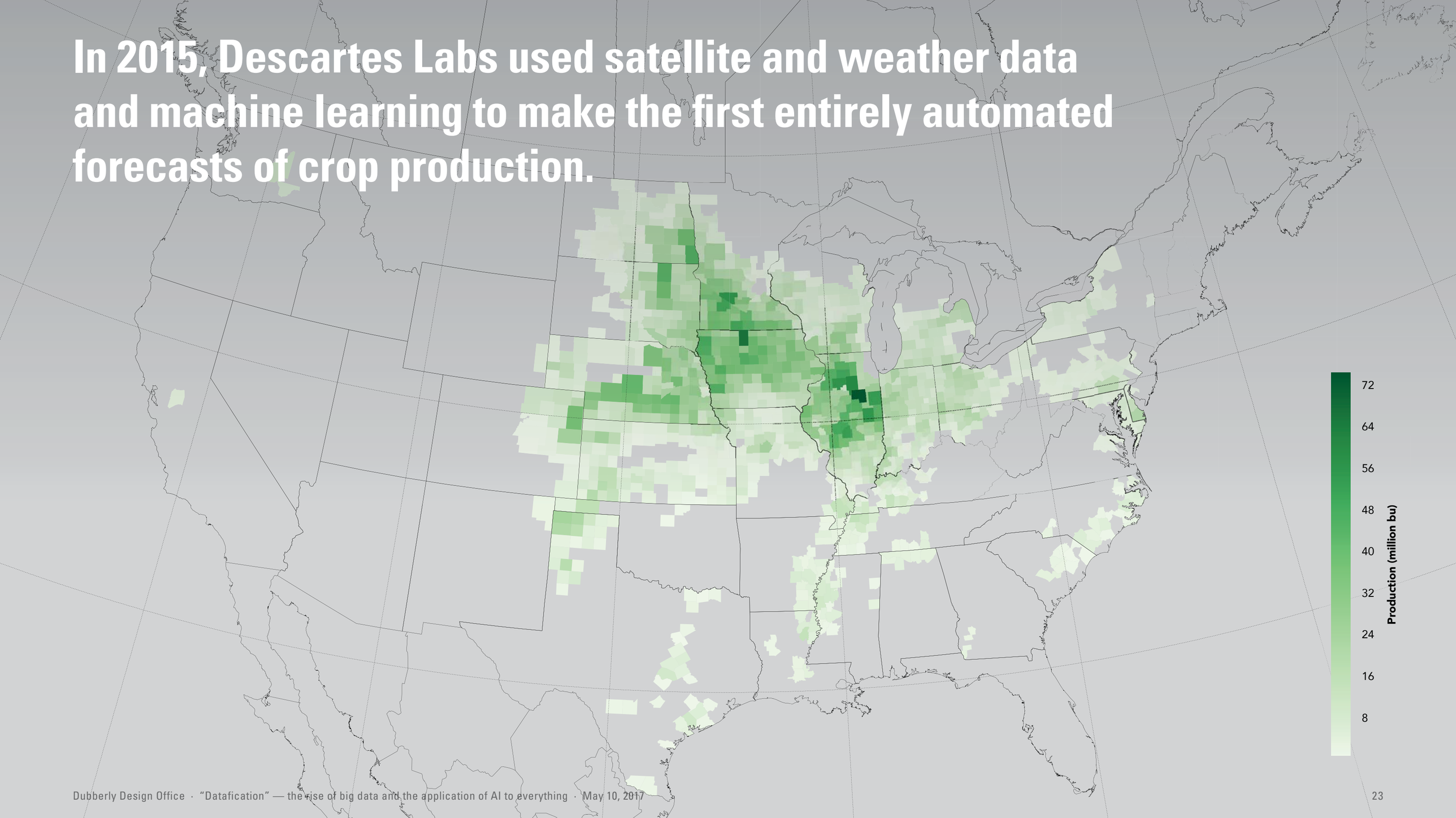


Daily weather data can augment machine learning.

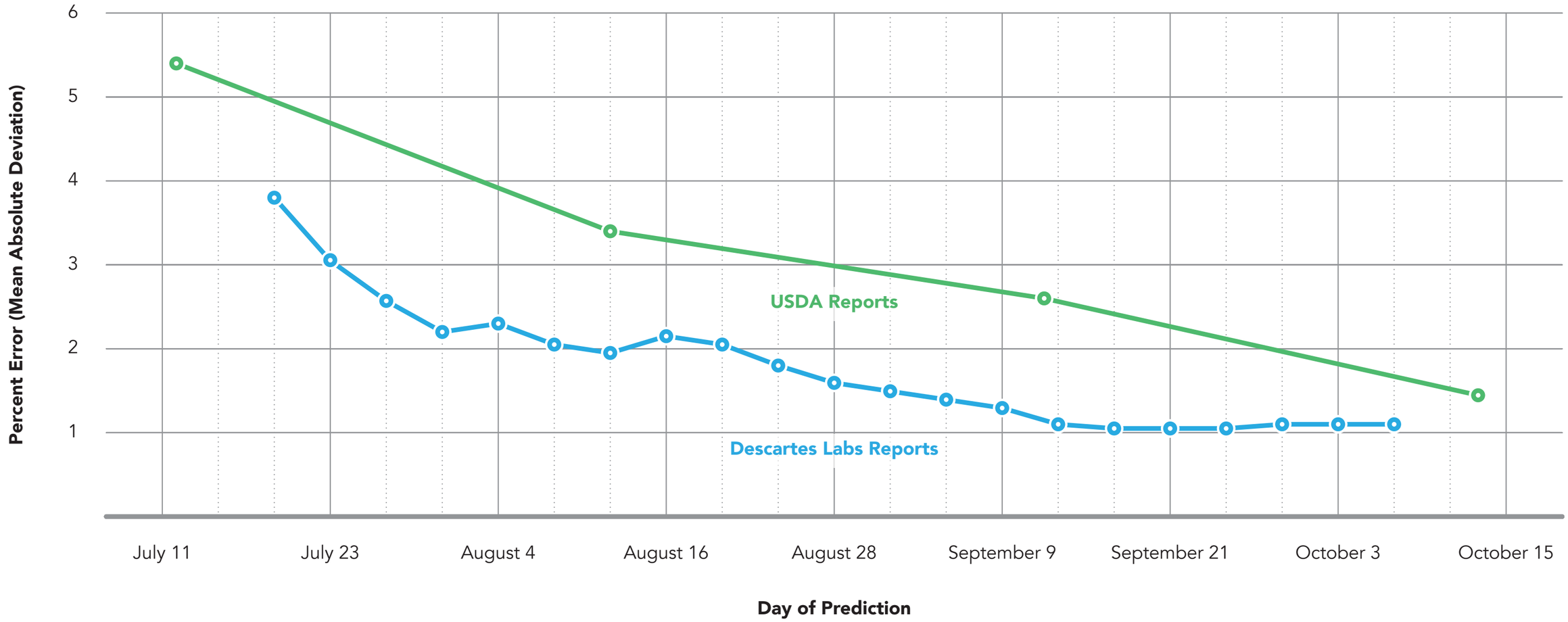
Precipitation,
temperature,
wind direction and speed,
snow cover,
and cloud cover
can aid forecasting.



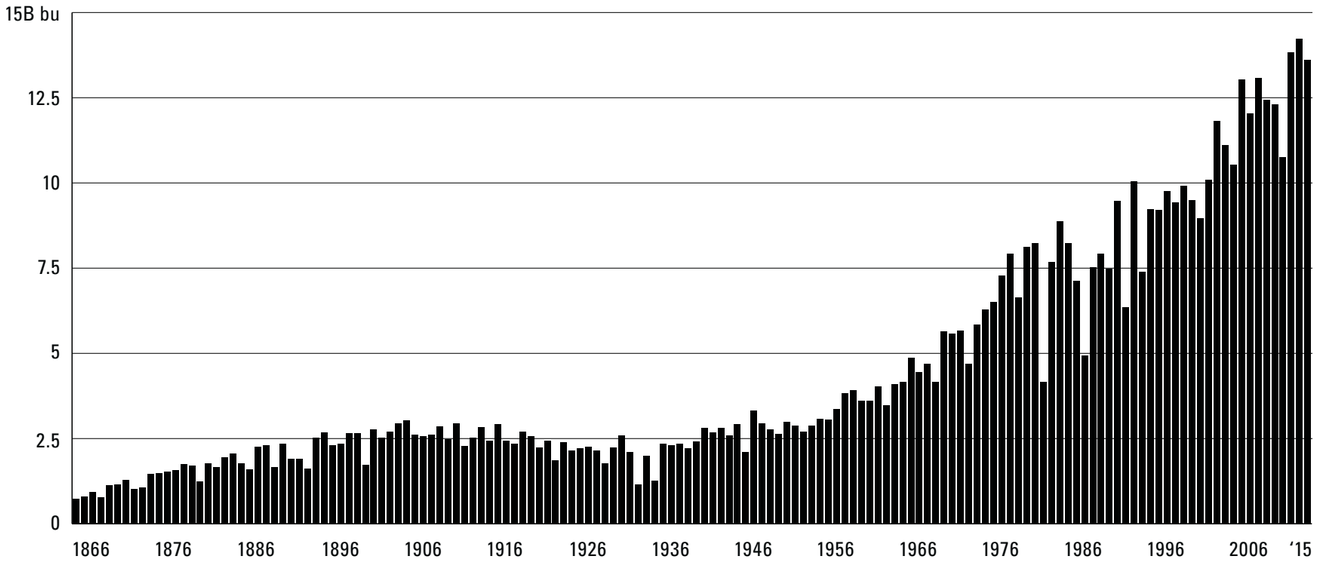
In 2015, Descartes Labs used satellite and weather data and machine learning to make the first entirely automated forecasts of crop production.



Descartes Labs predicted US corn product — within about 1.9% of actual production later reported by USDA.



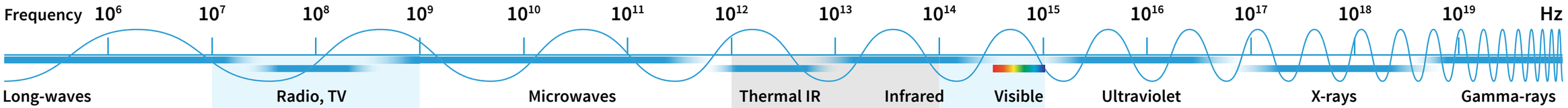
Until now, prediction was based on sampling.
Since 1866, USDA has been measuring corn production—by hand.



Descartes has begun with existing NASA and ESA satellites.

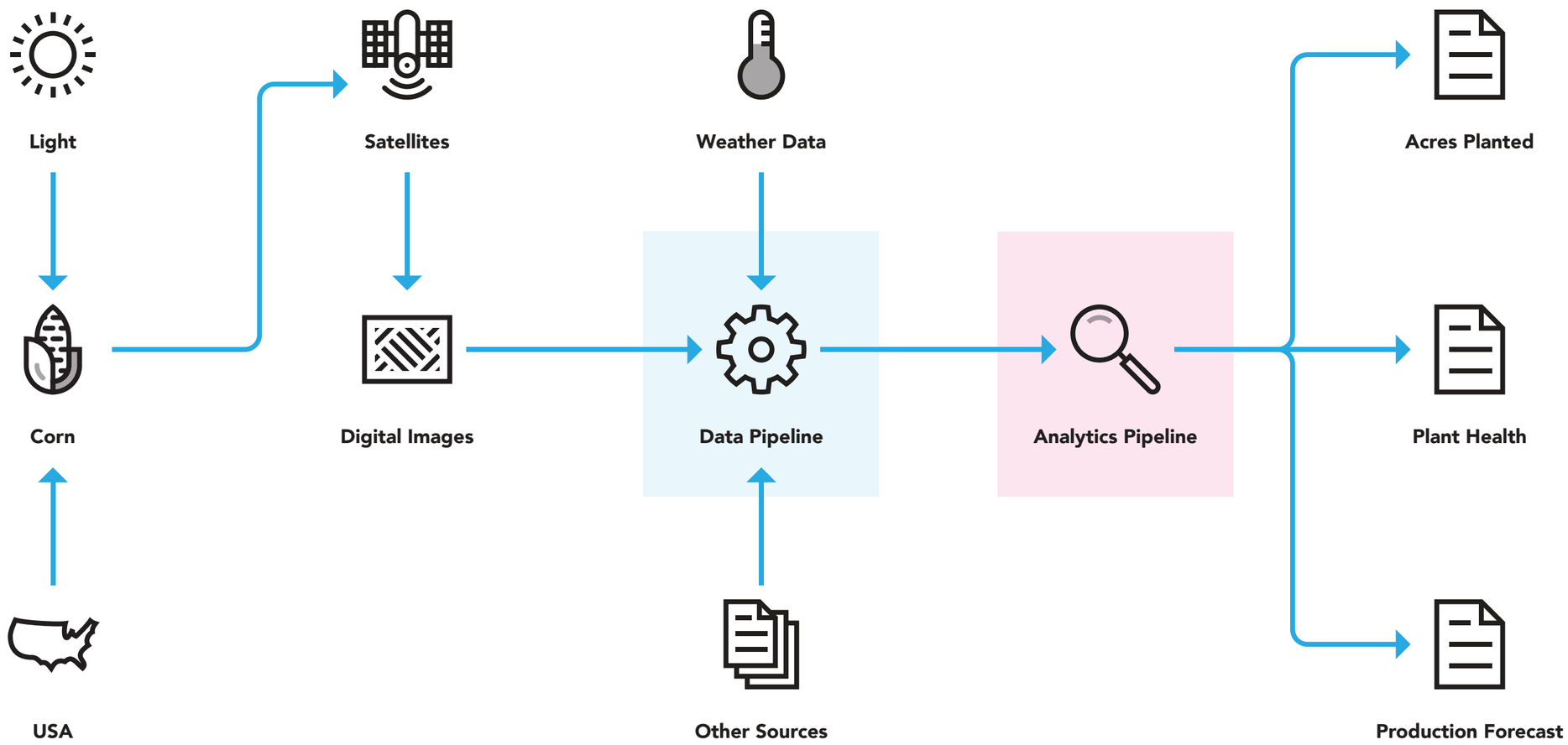
	Spectrum	Resolution	Frequency	History
MODIS	2 bands	250m	Daily	From 1999
Landsat	8 bands	30m	Weekly	From 1972
Sentinel	21 bands	10m	Weekly	From 2015
RapidEye	5 bands	5m	2 Days	From 2009
PlanetScope	5 bands	3-5m	Daily	From 2015
	Broader spectrum enables us to see beyond human vision— Infrared and near infrared indicate plant health.	Resolution detail enables us to see smaller objects—moving towards human scale.	More frequent imaging enables us to track changes/remove clouds close to real time—moving towards seeing weather events.	

The range of aerial sensing platforms is growing.



Traditional satellites	Telesat Kacific		Satellogic Koolock	Satellogic DigitalGlobe Koolock
Micro & nano satellites	Spire GeoOptics (GPS RO) PlanetIq (GPS RO) OneWeb (LTE) Magnitude Space (LPGAN) Axelspace Fleet Kepler (LTE) Astranis Iceye (Radar) Bluefield			Axelspace Planet Hera Systems GHGSat Bluefield Promethean Labs
Balloons	Project Loon (LTE) Facebook High Altitude Drone			
Planes				Promethean Labs TerrAvion
Drones				Promethean Labs GeoVantage

Descartes prediction involved 1 PB of data and roughly 24 hours of computation.



Descartes is a signal of a massive change.

- Self-driving cars, trucks, and drones

- IBM Watson Health

- GE Predix and Siemens MindSphere

- Apple Siri, Viv (now Samsung), Amazon Alexa, Google Assistant, Facebook M, Microsoft Cortana

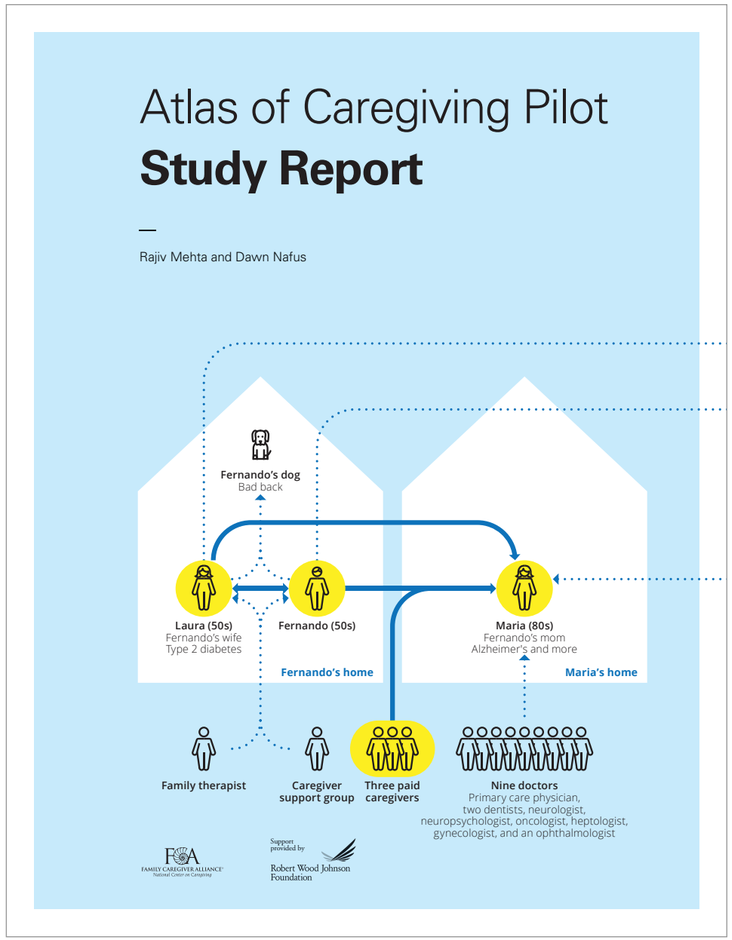
- FBI's Facial Analysis, Comparison, and Evaluation (FACE) Services has access to > 400 million photos.

Large, unique databases are inherently valuable.

data + algorithms = prediction

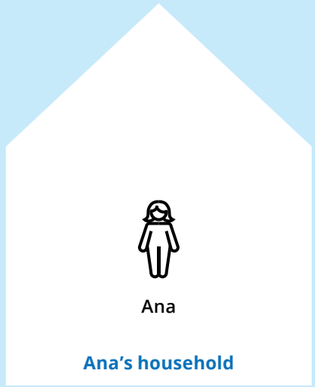
Another example

In 2015, Robert Wood Johnson Foundation funded a pilot study to look at new ways of measuring family caregiving.



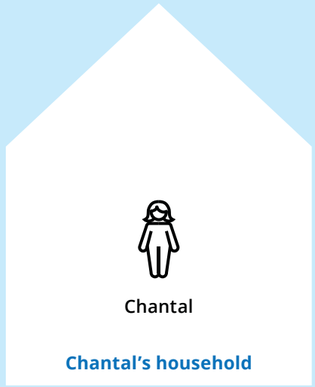
Robert Wood Johnson Foundation

We looked at 14 households, with 20 participants, with 21 chronic conditions.



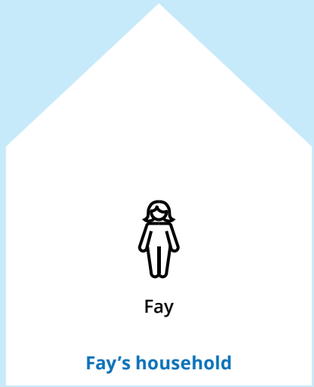
Ana's household

Ana (50s) has had **cystic fibrosis** since birth. She devotes several hours a day to care for her own condition. She also cares for her teenage son Albert, who has **depression**.



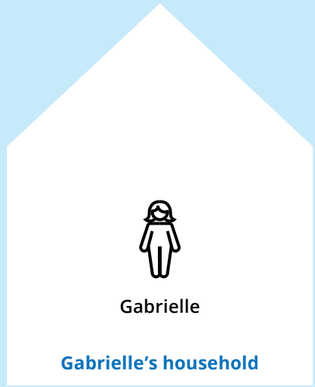
Chantal's household

Chantal (50s) has resigned work to care for her mother Debby (80s) who requires 24x7 care for **dementia**. Additional support comes from a paid home aide and other family members.



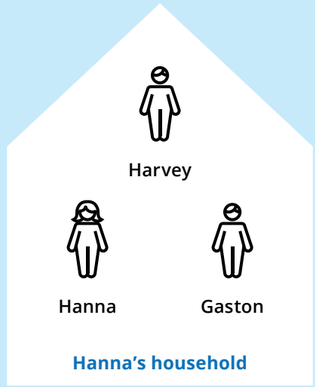
Fay's household

Only-child Fay (30s) cares for her mother Josephine (70s) who has **Alzheimer's**. With no one to help her, she has put PhD studies on hold to provide 24x7 care.



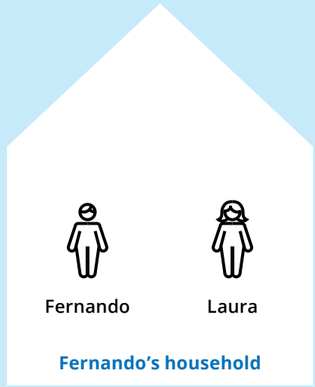
Gabrielle's household

Gabrielle (60s) is the primary caregiver of her mother Karen (101), who has **Alzheimer's**. Gabrielle also has health issues of her own and the sleepless nights and caregiving needs of her mother have taken a toll.



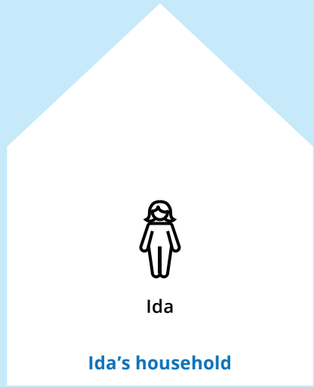
Hanna's household

Hanna and husband Gaston care for her brother Harvey, who has **epilepsy** and **pneumonia/sepsis**. Gaston also cares for his mother, while managing his own **chronic pain** and **edema**. Both Hanna and Gaston also work.



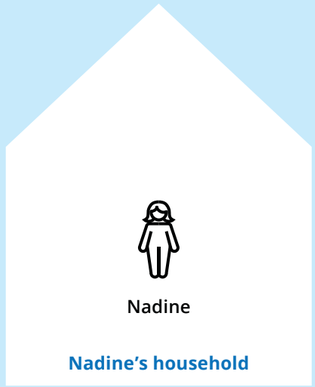
Fernando's household

Fernando and his wife Laura (50s) are the primary caregivers for Fernando's mother Maria (80s) who has **Alzheimer's** disease as well as other health conditions. Together, Fernando and Laura have built a care network to support Maria.



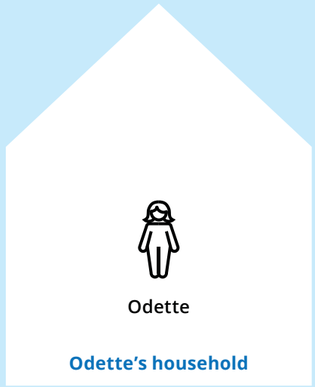
Ida's household

Ida (70s) cares for her husband Ian (70s) who has **Lewy Body Dementia** and **Dysautonomia**. They moved to San Francisco to be nearer to their children two years ago.



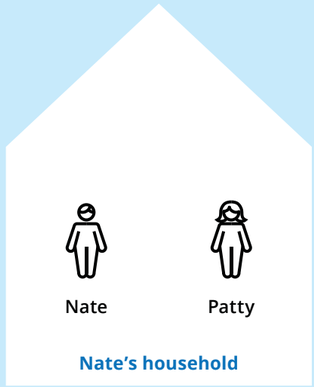
Nadine's household

Nadine (50s) lives with her husband Jerry and two teenage sons, Larry and Karl. Karl has **Type 1 Diabetes**. Nadine is his primary caregiver.



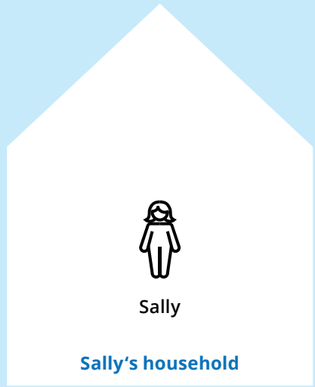
Odette's household

Odette (70s) and her husband Marco (70s) share their home with several other people: their son, son-in-law, and five tenants. Marco has **Parkinson's** disease. Odette is his primary caregiver, but several others are also involved.



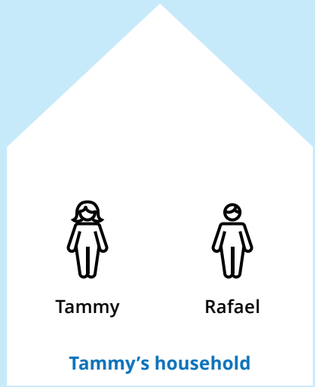
Nate's household

Nate and Patty, both in their 30s, care for each other. Patty has **multiple sclerosis** (MS) and Nate has **glioblastoma**, a terminal condition.



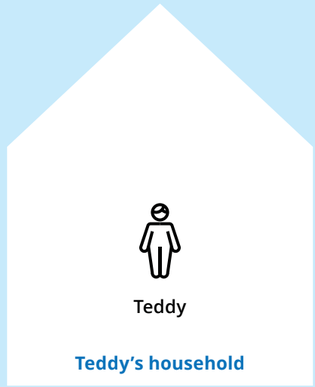
Sally's household

Sally (50s) cares for her son Pablo (20s), who has behavioral and emotional difficulties stemming from **XYY Chromosome Disorder**.



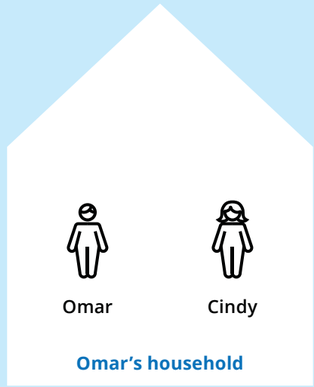
Tammy's household

Tammy (40s) and her husband Rafael (50s) care for their pre-teen children, Wanda and Sam. Wanda has severe **epilepsy** and **cerebral palsy**. She requires 24x7 care. Sam has severe **autism** and also requires a lot of care.



Teddy's household

Teddy (40s) and his wife are the primary caregivers for their two young sons, Van and Walter. Van has **Aspergers (ADHD type)** as well as **encopresis**, and Walter has **cyclical vomiting syndrome**.



Omar's household

Omar (40s) and his separated wife Cindy (40s) share a home with their young son Bob, who has **Aspergers**.

Using 12 sensors

Measuring 16 factors

Photographs
GPS and timestamp

Narrative Clip

Blood Volume Pulse
Calculated to derive heart rate

Acceleration X
Acceleration Y
Acceleration Z
Calculated to derive average motion

Electrodermal Activity
(EDA)

Skin Temperature

Empatica E4

Presence
At home or away

SmartSense Presence Sensor

Motion
In which room and how active

SmartSense Motion Sensors

Humidity
Temperature
Barometric pressure
C02
Noise
Indoor unit

Netatmo Indoor Weather Station

Humidity
Temperature
Outdoor unit

Netatmo Outdoor Weather Station

Over an average of 24 hours

Resulting in 5 GB of data—just from the watch.

The BVP sensor is running at 64 Hz.
That means it makes a reading every
1/64th of a second.
60 seconds comprise a minute;
60 minutes comprise an hour; and
36 hours is the maximum duration of
one of our study sessions.

In other words, one study session
comprises 2,160 minutes,
and just one of the sensors
is collecting 3,840 samples per minute.

That's 8,294,400 samples collected over
the course of one 36-hour session.

8,294,000	samples for BVP (at 64 Hz)
4,147,200	samples for X axis acceleration (at 32 Hz)
4,147,200	samples for Y axis acceleration (at 32 Hz)
4,147,200	samples for Z axis acceleration (at 32 Hz)
518,000	samples for EDA (at 4 Hz)
518,000	samples for skin temperature (at 4 Hz)
21,772,800	samples of raw data for one participant
×19	participants
413,683,200	or nearly half a billion data points

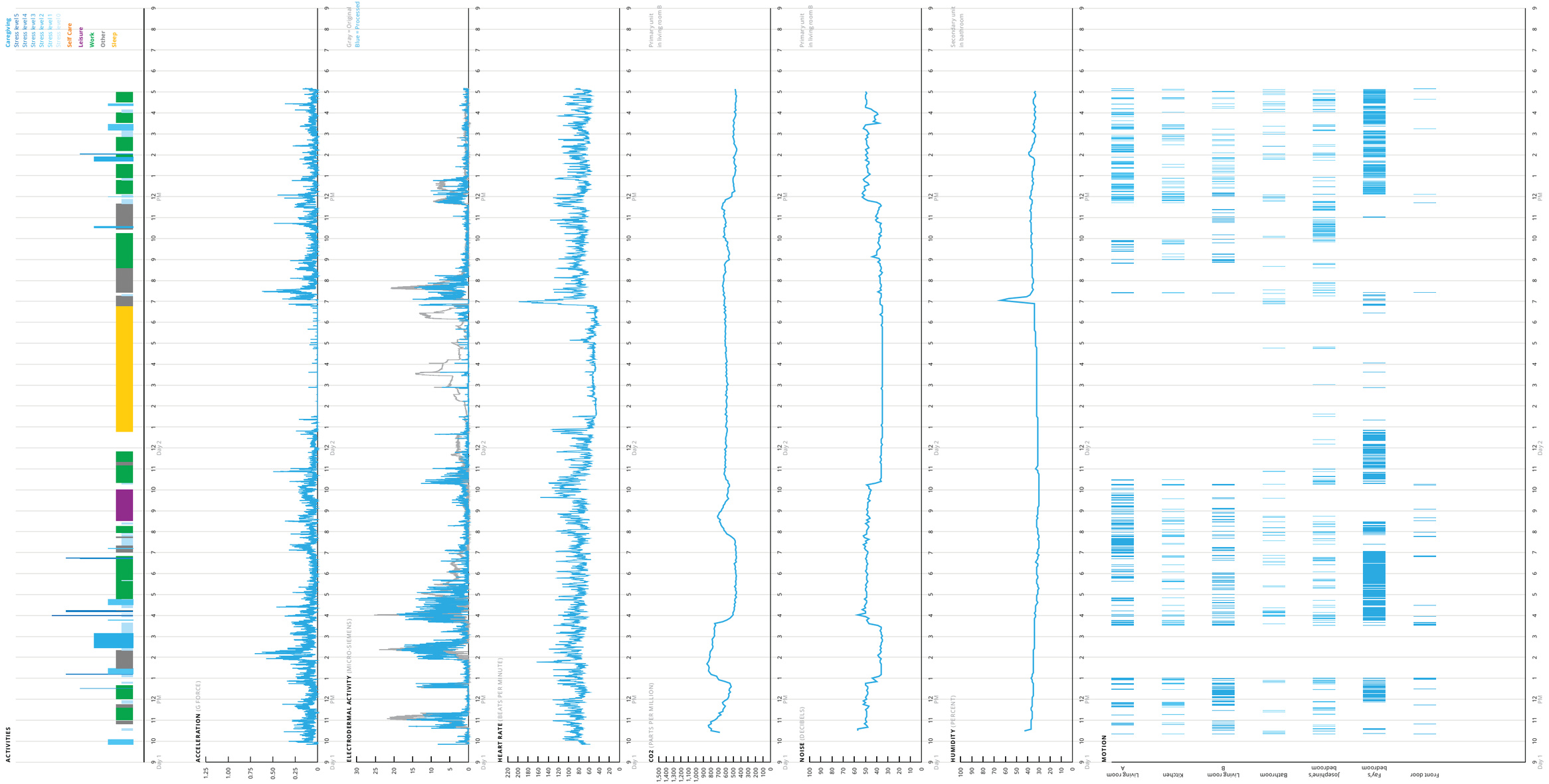
Photo log for Fay ×20 additional participants

Black squares replace recognizable faces to ensure privacy.

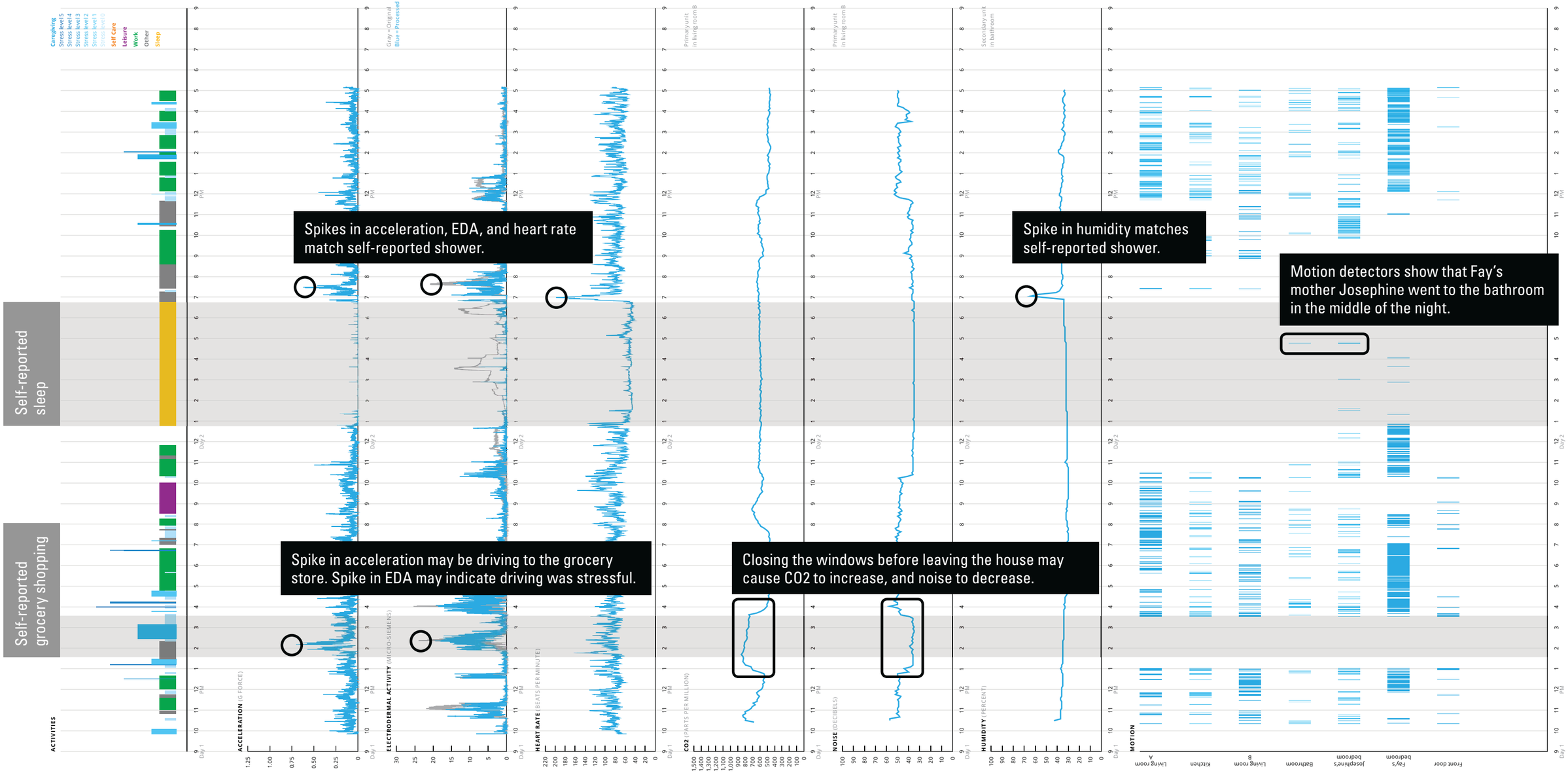
Gray squares indicate when participants turned the camera off.

White squares indicate the start and stop of the study.

Summary diagram for Fay ×18 additional participants

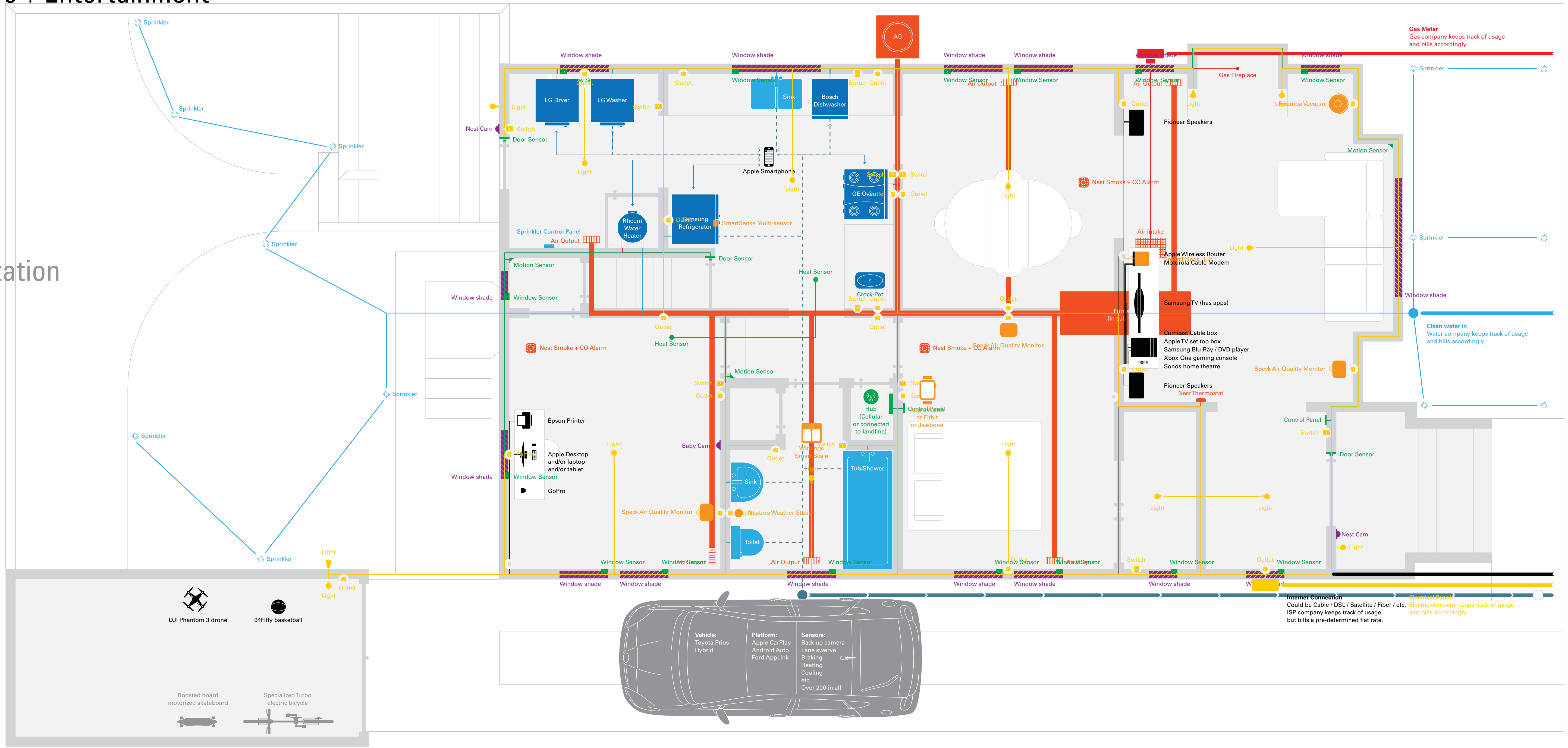


Analyzing Fay's summary diagram for insights—**morning**



IoT devices in homes will produce and collect massive amounts of data:

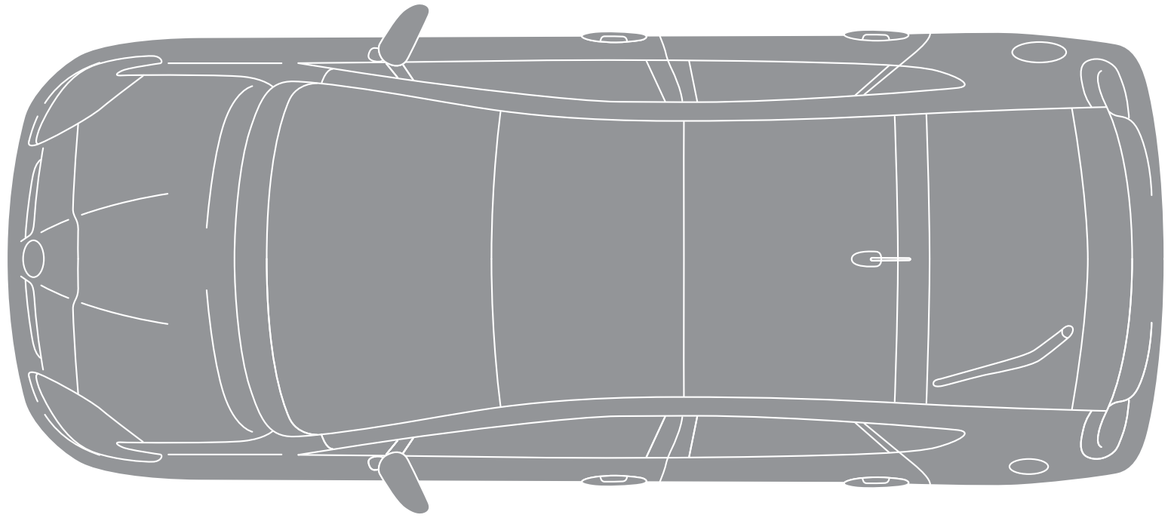
- Appliances
- Computers + Entertainment
- Electrical
- Gas
- Health
- HVAC
- Plumbing
- Privacy
- Security
- Transportation



Today's average car has:

-1 engine
~7 small motors
(windows, wipers, fans)
~30 micro processors
(up to 100 for luxury cars) [1]
~60-100 sensors
(growing to 200 by 2020) [2]
~100 million lines of code
(up from 2 million lines in a generation) [3]

And it produces
"terabytes of data per car per day" [4]



Sources:
[1] http://www.nytimes.com/2010/02/05/technology/05electronics.html
[2] http://www.automotivesensors2015.com/
[3] https://leithporsche.com/news/What+Makes+the+2017+Porsche+Panamera+Different3F+Computer+Code/7659/
[4] Parrish Hanna, Global Director of HMI at Ford (personal communications)

Google + Levi’s connected denim smart jacket

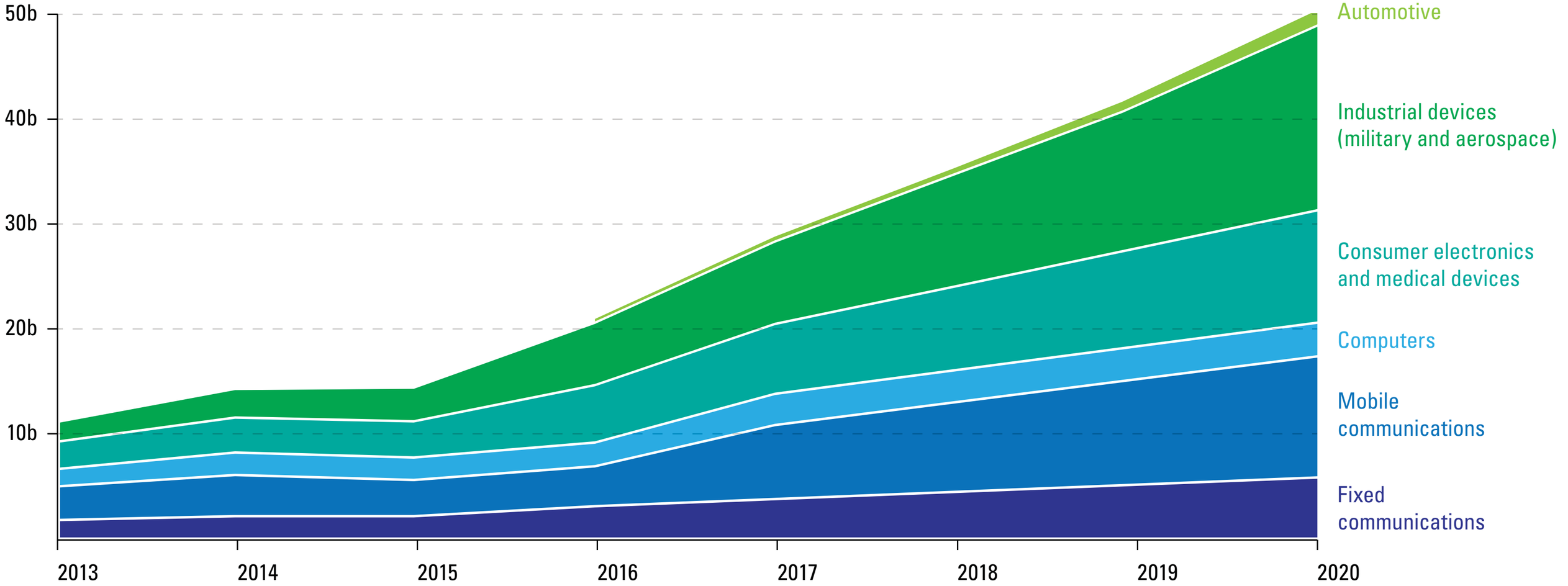
Jacquard Woven Gesture Sensor

Jacquard Tag

Jacquard App

Jacquard Services

By 2020, ~50 billion devices will be connected to the Internet; today, ~7 billion computers and tablets are connected.

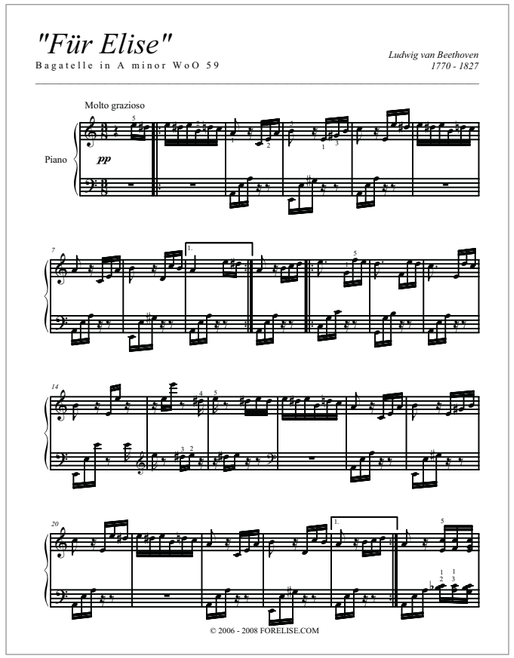


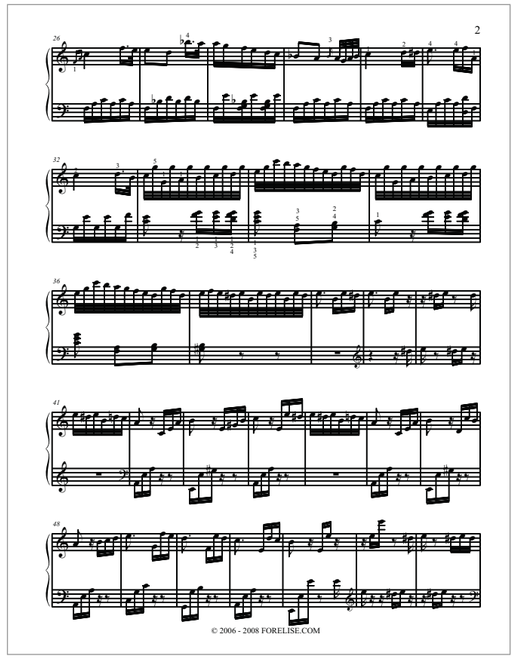
Sources: The Economist and Cisco

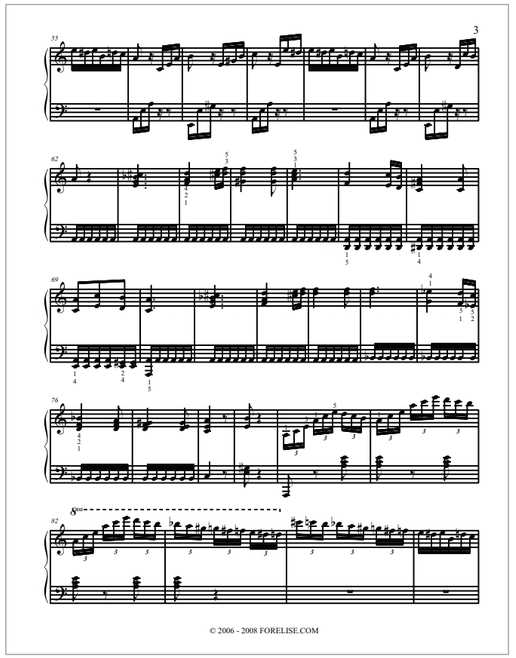
A final example

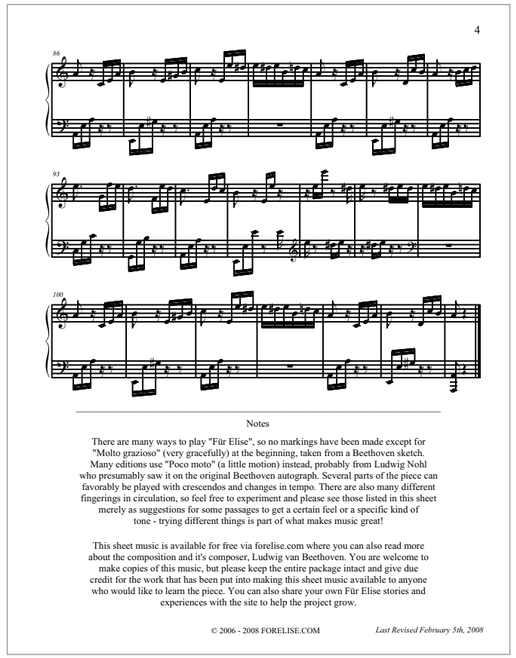
How do you measure “quality” in playing a piano?

An experiment was created to track the hand movement during multiple piano performances of Bethoven’s *Für Elise*.

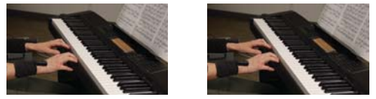
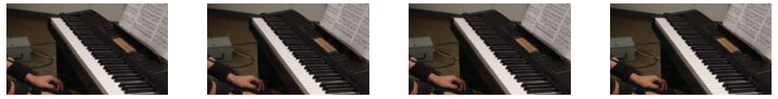
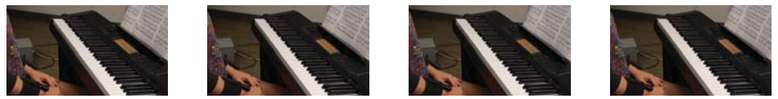
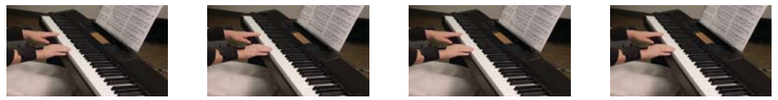






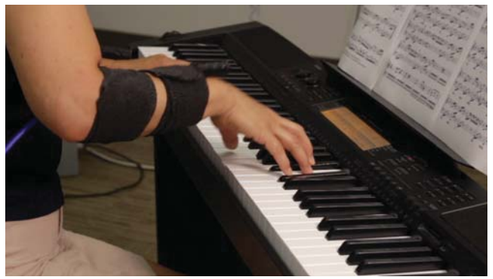


Twenty-five performances by seven performers were tracked and recorded.

- Grace 3 takes 
- Jamie 2 takes 
- Jiarong 4 takes 
- Katie 4 takes 
- Kelsie 4 takes 
- Sachiko 3 takes 
- ShanShan 5 takes 

Hypothesis: advanced players move their wrists to a greater degree.

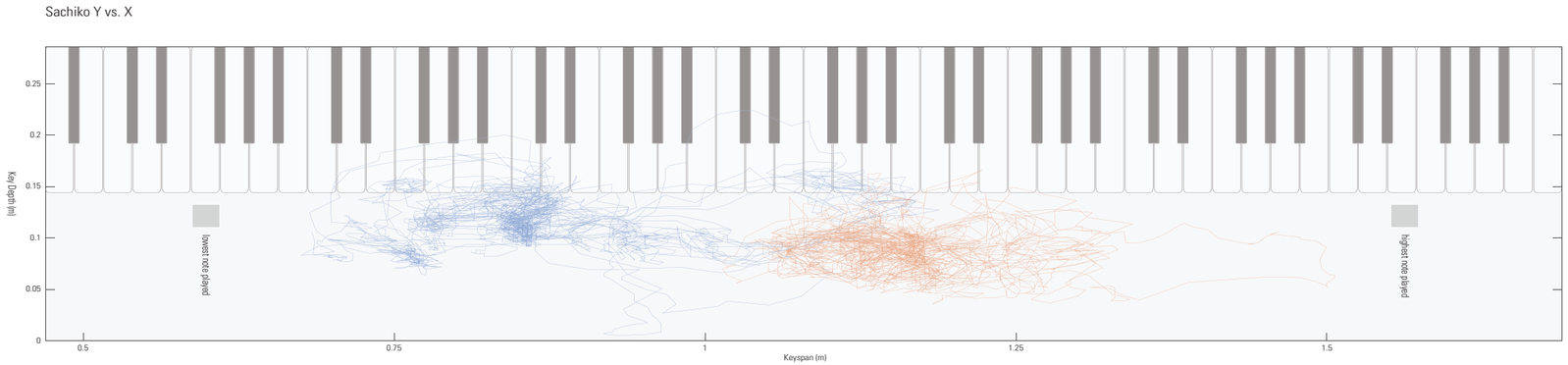
Listening and watching videos of sample performances by Jamie and Sachiko gave a clue.

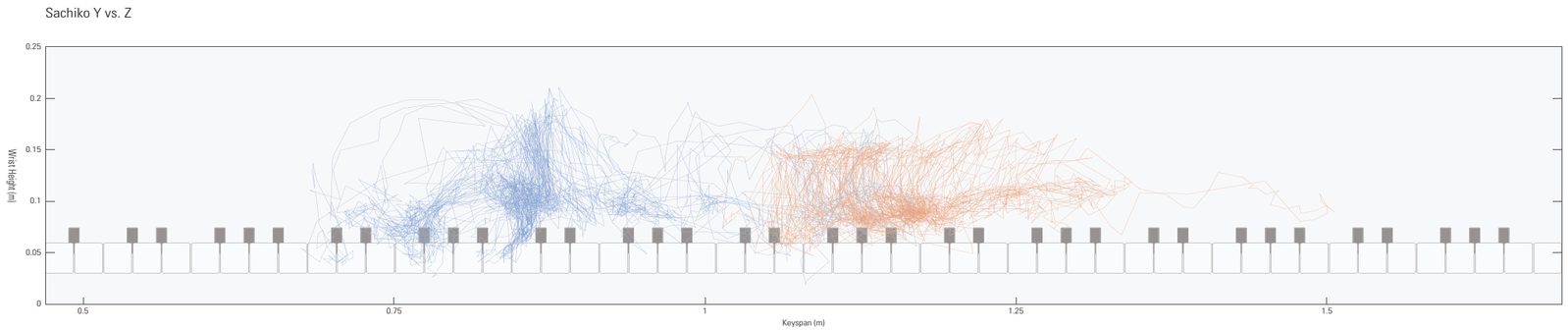


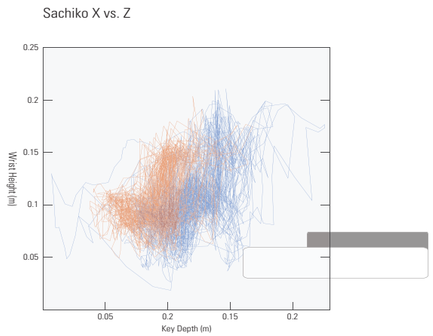
Sachiko’s performance was clearly better. Her wrists moved up and down as she played the piece, while Jamie’s wrists were relatively flat.

The orthographic projections are more revealing.

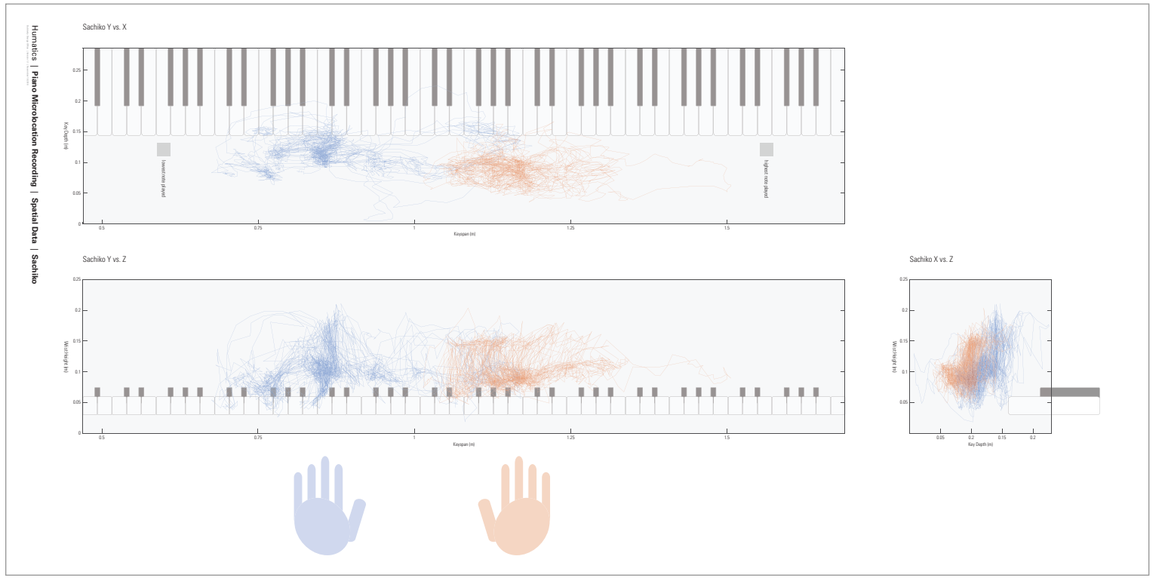
Humatics | Piano Microlocation Recording | Spatial Data | Sachiko



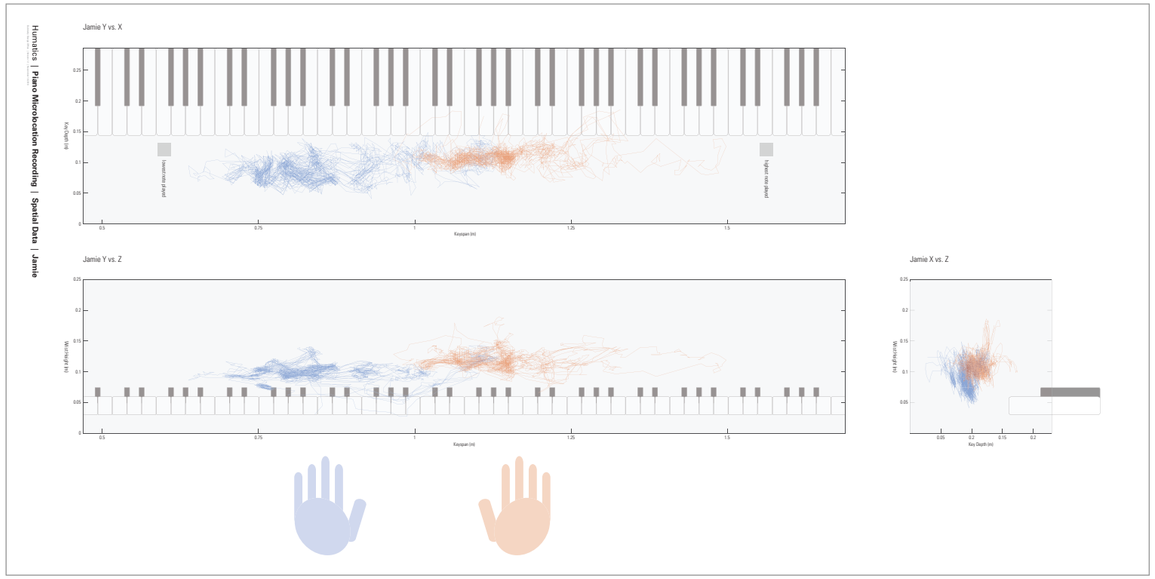




Comparisons show differences in movement in each dimension.

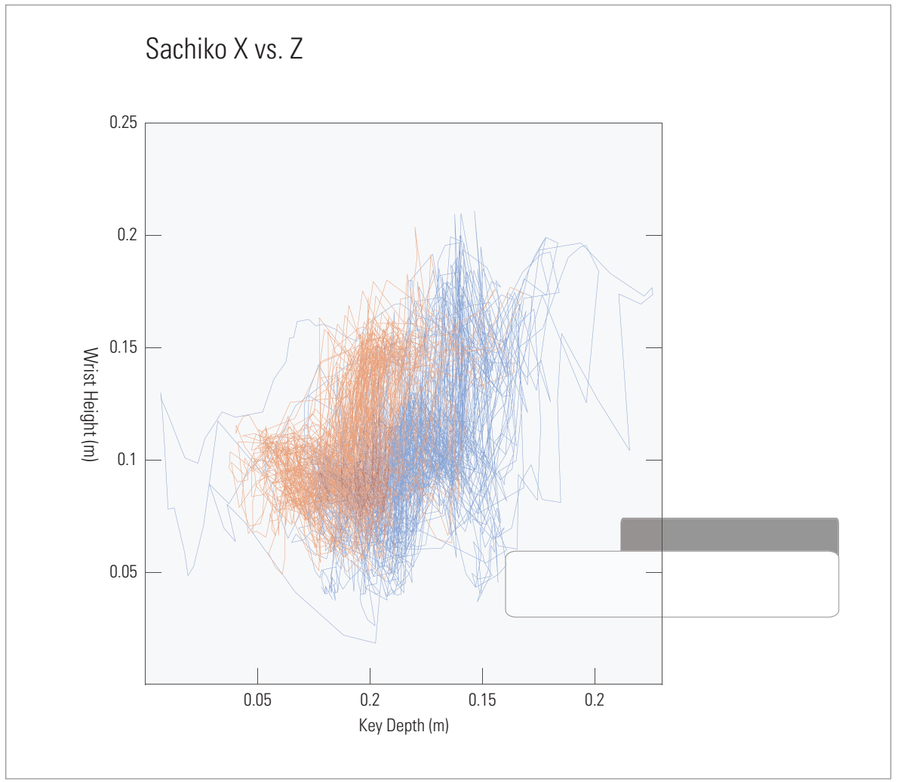


Sachiko

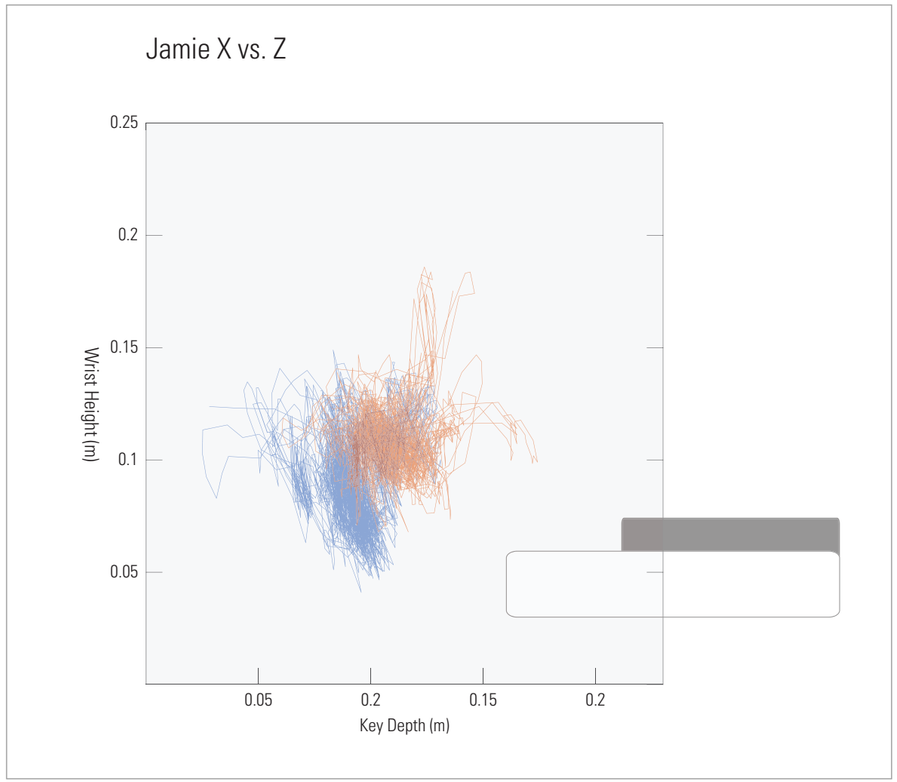


Jamie

The X vs Z plots are the most revealing.



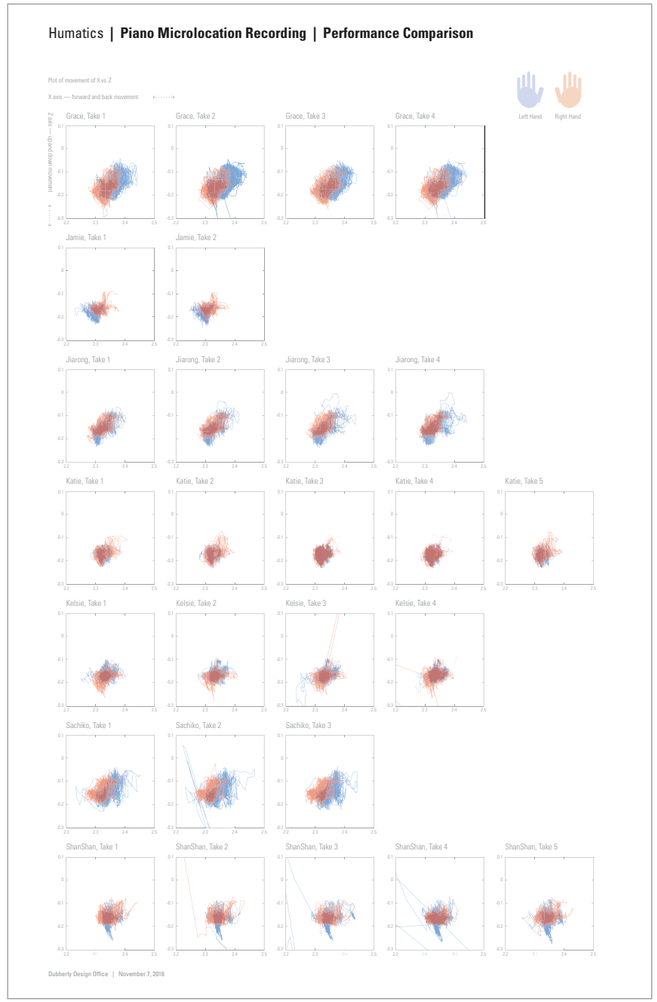
Sachiko



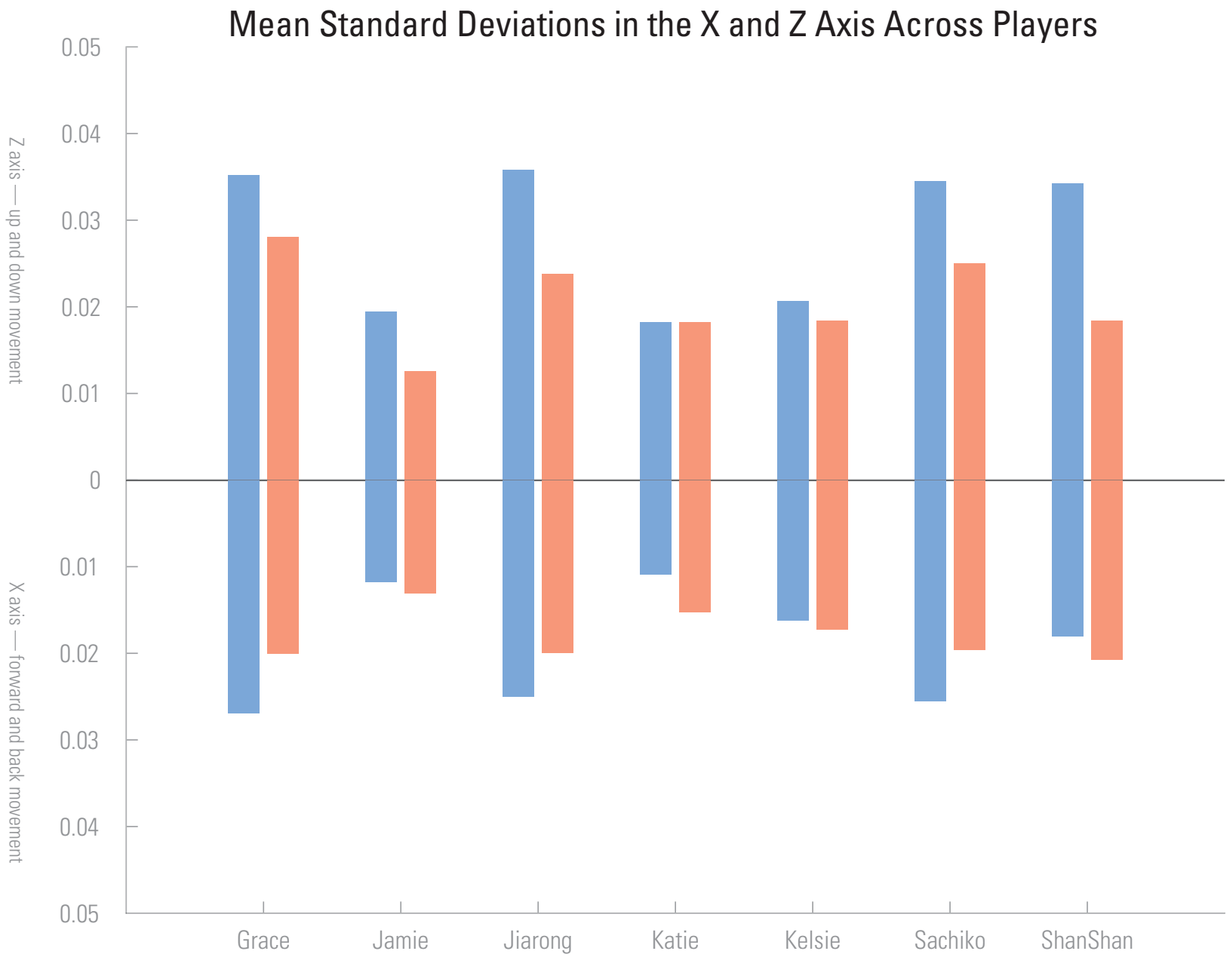
Jamie

X vs Z plots for all performances—
the differences are obvious.

- Grace
- Jamie
- Jiarong
- Katie
- Kelsie
- Sachiko
- ShanShan



Calculating standard deviation shows a clear pattern.



The bar graphs show greater movement (particularly in the left hand) for Grace, Jiarong, Sachiko, and ShanShan—indicating they are advanced performers.

What does this mean?

Each new phase of the information revolution opens a new domain of opportunity — a new seam to mine.

1

Back-office Processing
automating core processes, like finance and HR, with stand-alone, customized applications.

2

Office Automation / PCs
"digitization" of business, applying digital tools to work.

3

Internet
connecting workers, connecting companies, and putting services on-line.

4

Software-as-a-Service
integrating applications, so that data can flow across the enterprise.

5

"Datafication"
making the enterprise aware of itself and its environment, opening the door to reasoning.

“Datafication” offers four successive levels of value,
(based on a model by Michael Porter)

Automation — enabling systems to run autonomously,
(e.g., programmed trading, self-driving cars, etc.).

Optimization — predicting changes (e.g., usage, failure, etc.),
and deploying resources accordingly (i.e., arbitrage).

Control — correcting variables that exceed thresholds,
ensuring that systems operate within bounds.

Monitoring — measuring operations;
sending alerts as variables approach thresholds.

“Datafication” is built on a series of technology layers (a stack); each adding value and creating opportunity

Prediction algorithms — recognizing “patterns of daily living,” reasoning about sequences of events and what is likely to happen.

Change-detection algorithms — recognizing events (changes in objects) and sending alerts when a threshold is reached.

Pattern-recognition algorithms — recognizing objects, teasing “meaning” out of masses of data.

Programmable APIs — making archives accessible for online machine-based queries.

Multi-modal archives — connecting data from multiple sources, so that it can be correlated.

Data pipelines — collecting data in a central repository, cleaning and wrangling it, so that it can be retrieved and used.

Sensor arrays — measuring the environment, by deploying and connecting foundational technology.

“Datafication” also changes the way we interact with computers; each change has increased access and convenience for users

1. Connecting wire patch panels.
2. Submitting batches of punch cards and receiving reams of tele-typed printouts.
3. Remembering and typing commands on keyboards and CRTs.
4. Finding menu items and directly manipulating objects with mice.
5. Touching screens and gesturing.
6. Speaking to virtual assistants and directing autonomous robots.

“Datafication” builds on a long-term trend in science and heralds a new way of doing science

(based on a model by Jeff Leek and Brad Efron)

1. Era of data scarcity— origins of data science, early 19th century	**2. Era of small data sets**— classical statistics developed, late 19th century	**3. Era of mass-produced data**— late 20th century, “macroscopes” emerge	**4. Era of measuring everything**— early 21st century, “macroscopes” become ubiquitous
Data sets were few + infrequent (e.g., census)	Individual scientists working independently	Teams of scientists using computer-controlled instruments	A few large organizations assemble immense data sets (e.g., Google, NSA)
Based on manual sampling	Collect few samples and make many measurements (noise becomes a problem)	Automatic sampling, producing digital data	Millions of measurements of millions of things (much data goes unused)
Producing analog data	Questions remain simple but important (e.g., Which treatment is better?)	Multivariate analysis becomes important	Computing power and band-width gate analysis
Applied to simple but important questions	Correlating an effect with change in a single variable becomes standard of proof	Number and complexity of questions increases	Machine learning comes into its own (overfit becomes a risk)

Until Gutenberg, books had been rare, worth about three years salary, they we're literally chained to shelves in medieval libraries.

In 1455, Gutenberg published his bible, two volumes weighing about 70 pounds. Early tech replicates existing tech, increasing speed + reducing cost.

New tech takes 20 or 30 years to find its own form.
A generation after Gutenberg, Aldus Manutius published a “portable library”.

The result of printing was:

- direct access to “the word of god” — and the Reformation

- nearly universal literacy — and the Enlightenment

- perhaps even the “democratic” nation-state

- and arguably “modernism” itself

We would do well to keep this in mind,
when entrepreneurs promise “disruption”.

Special thanks to
Clara Gonzalez Sueyro
Knut Synstad

hugh@dubberly.com
@DubberlyDesign

Presentation posted at
presentations.dubberly.com/datafication.pdf