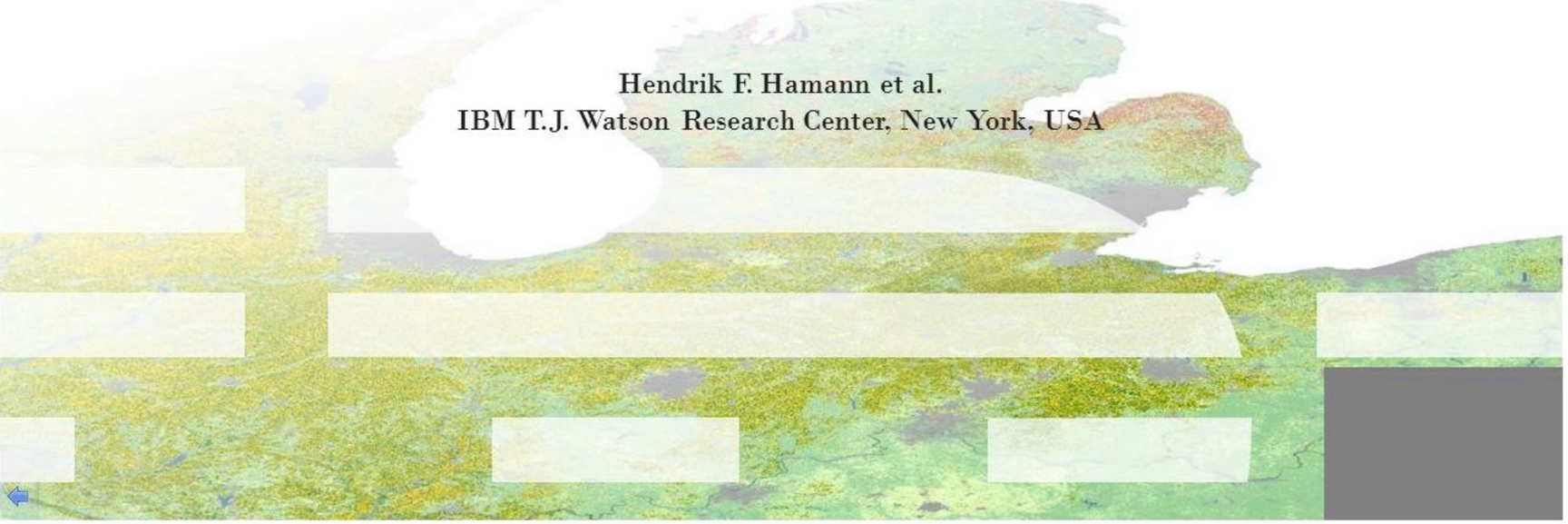
Making the invisible visible:

How to convert data from remote sensing systems and IoT devices to real information

Hendrik F. Hamann et al.
IBM T.J. Watson Research Center, New York, USA



This conference is being recorded

Name	Comment
Amit Srivastava	
Can Wang	
Charles Ernst	
David McKay	
david pinkos	
Floyd Harris	
gerald hanlon	
Grace Harvey	
Guest A10	
Guest A11	
Guest A12	
Guest A15	
Guest A16	
Guest A17	
Guest A18	
Guest A19	

Participants 47 0

Materials

Event Materials

Insert...

AT&T Connect Materials: Ready to Use

Raise Hand | Send Note | Step Out | Whiteboard | Audio | Unmute

Presenter is sharing an application

Meeting in Progress
Rec Connection: Good
No Audio Selected

Making the invisible visible:

How to convert data from remote sensing systems and IoT devices to real information

Hendrik F. Hamann et al.
IBM T.J. Watson Research Center, New York, USA

This conference is being recorded

Name	Comment
Guest A6	
Guest A7	
Hoang Vinh Nguyen	
Iman Morisset	
j f	
jeff carvalho	
Jenny Foley	
Jimmy Hsu	
Jody Schaeffer	
Man Peng	
Maryann Burns	
Neeraj Arora	
nishit pande	
Prashant Katiyar	
Priya Gopinath	
Rajat Kashyap	

Participants 60 0

Materials

Event Materials

Insert...

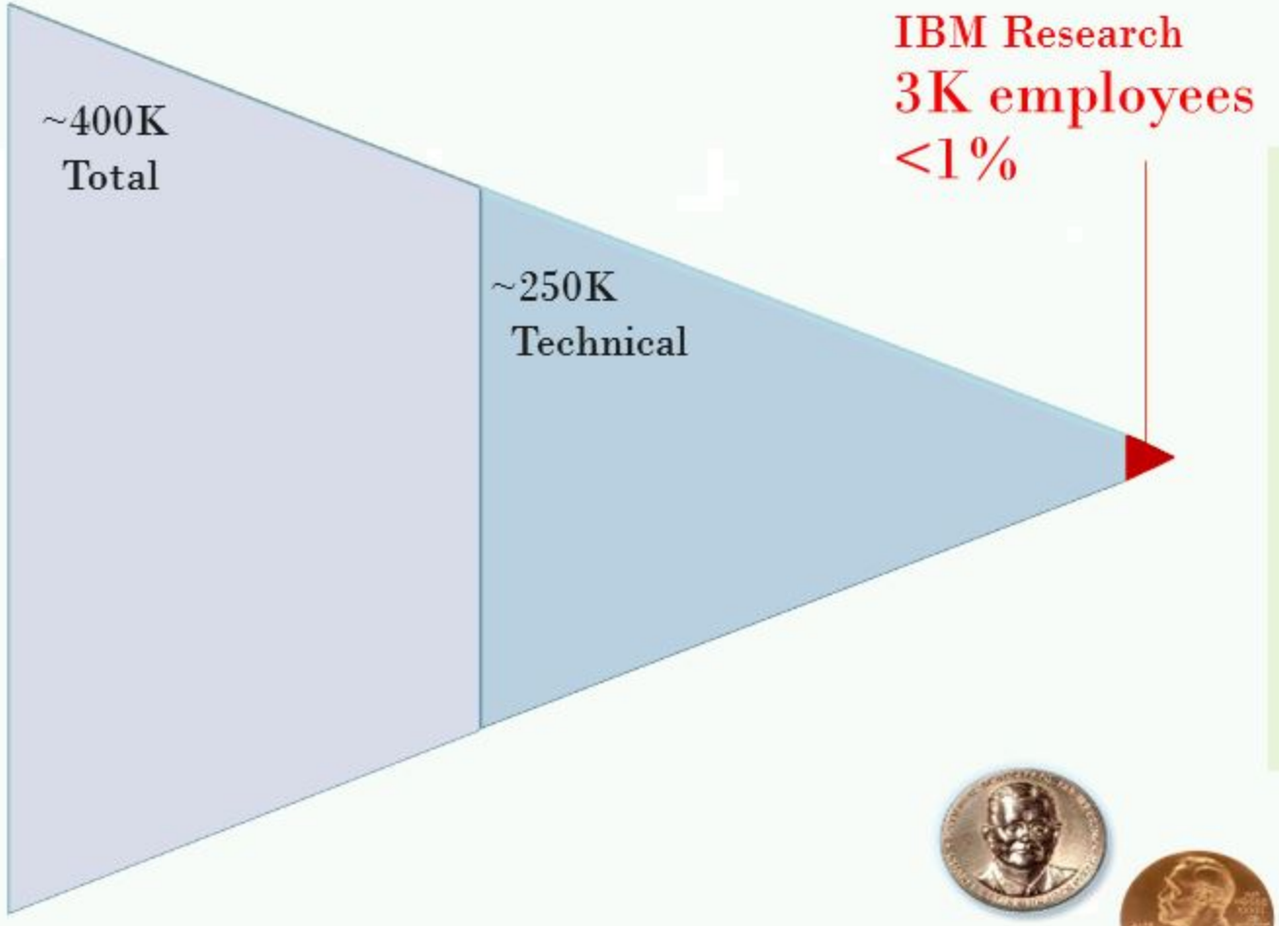
AT&T Connect Materials: Ready to Use

Outline

- IBM Research
- Technology Trends
- Physical Analytics
- Big data platform for spatio-temporal analytics
- Applications

Research is IBM's Innovation Engine

IBM Employee Population



Innovative

- 50% of external honors
- 46% of IBM Fellows
- 25% of all patents filed
- 26% of new academy members

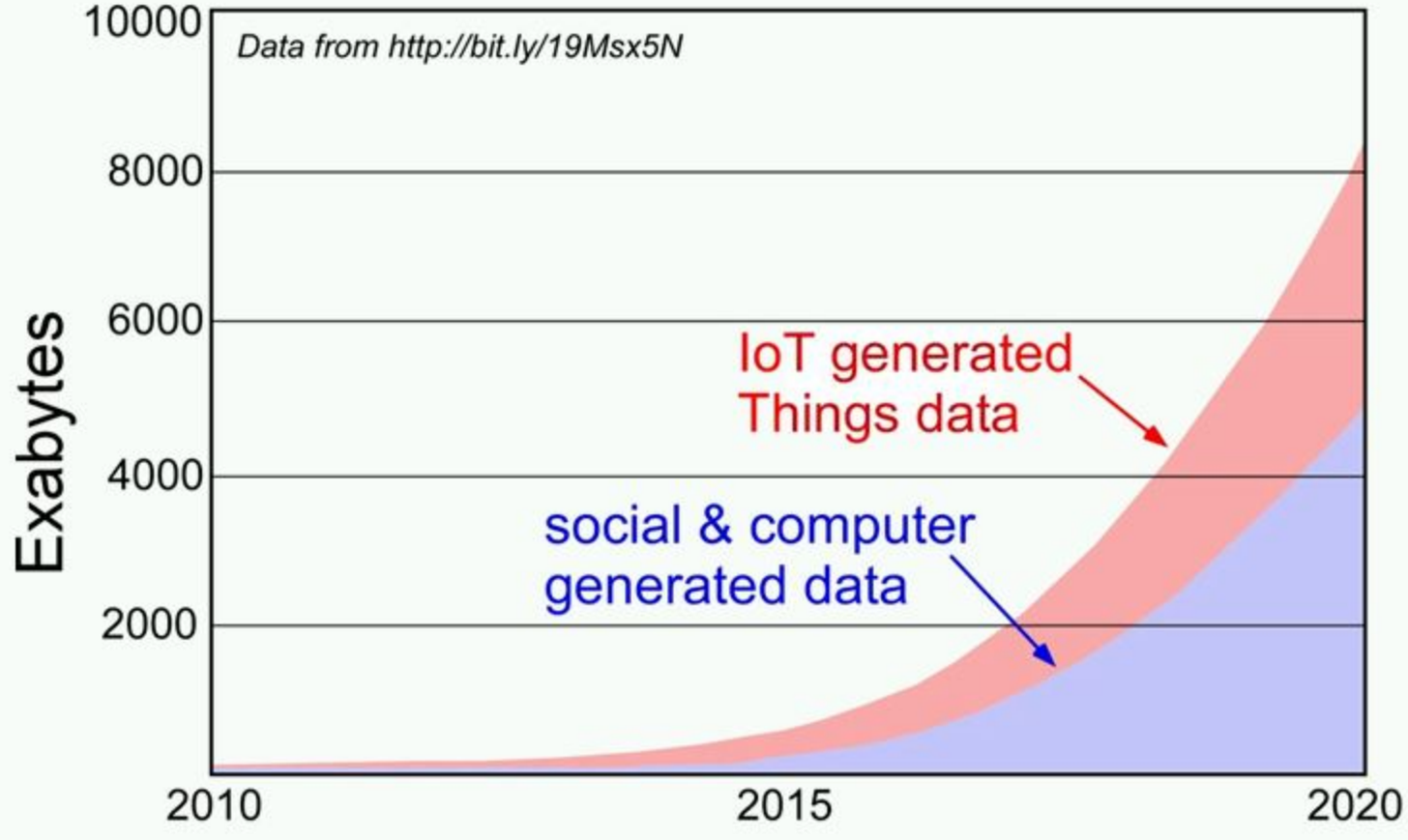
Recognized

- 5 Nobel Laureates
- 13 National Medals
- 6 Turing Awards
- 80 Members, National Academies
- 11 Inductees, National Inventors Hall of Fame

©2016 IBM Corporation

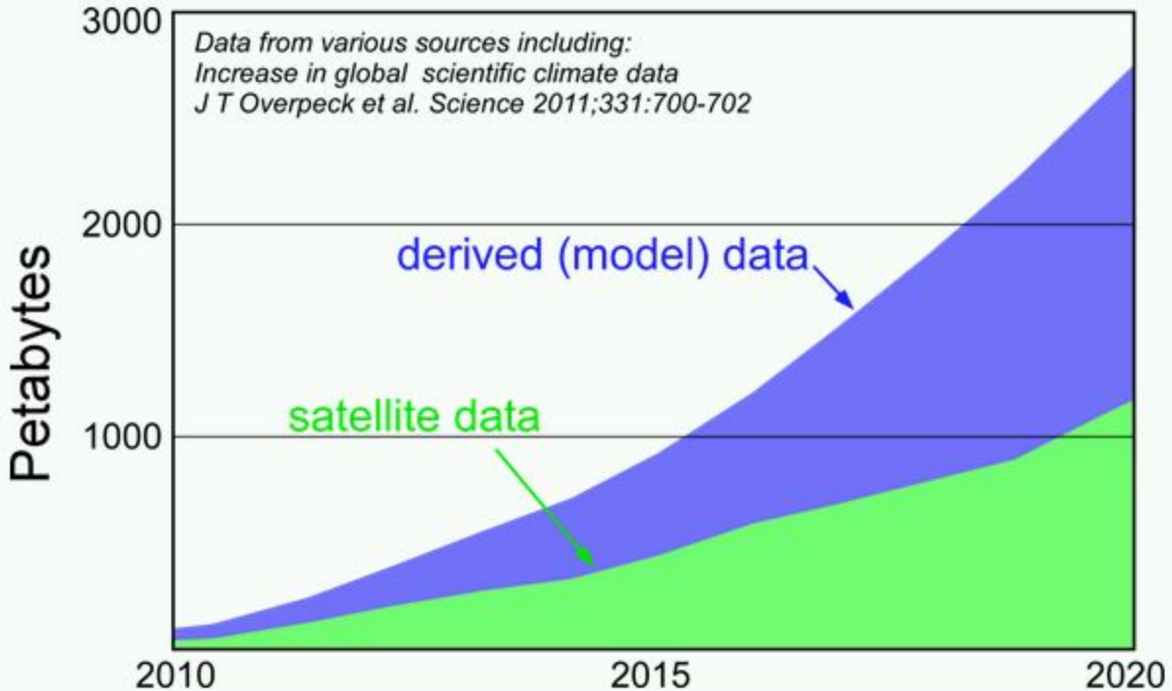
IBM Research is world-wide

"Internet-of-Things" generated data soon bigger than social and transactional data



~ 2x more data growth (currently @ 44EB/month) than social & computer generated da

Data from remote sensors is exploding



- **Number of satellite launches are growing exponentially**
  - Many countries have re-started major satellite programs (China, Europe, India etc.)
  - Cube satellites, Nano-satellites
- **Drones, cell phone with hyper-spectral cameras**

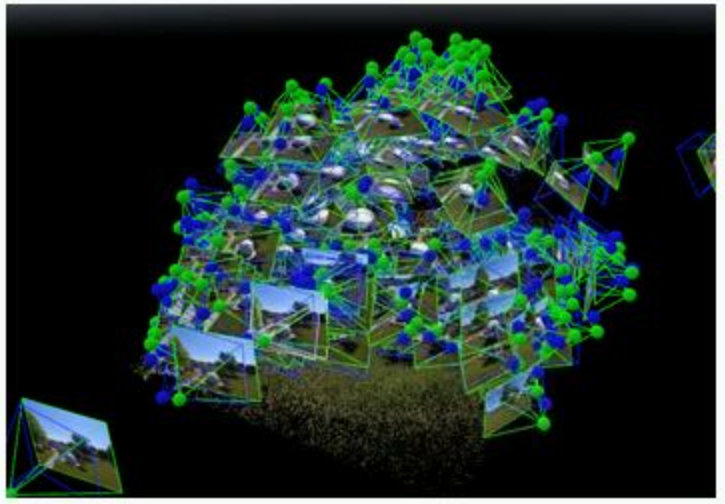
 ©2016 IBM Corporation

The physical world is being digitized….

 ©2016 IBM Corporation

The physical world is being digitized….

Position of the camera and images used for reconstruction



 ©2016 IBM Corporation

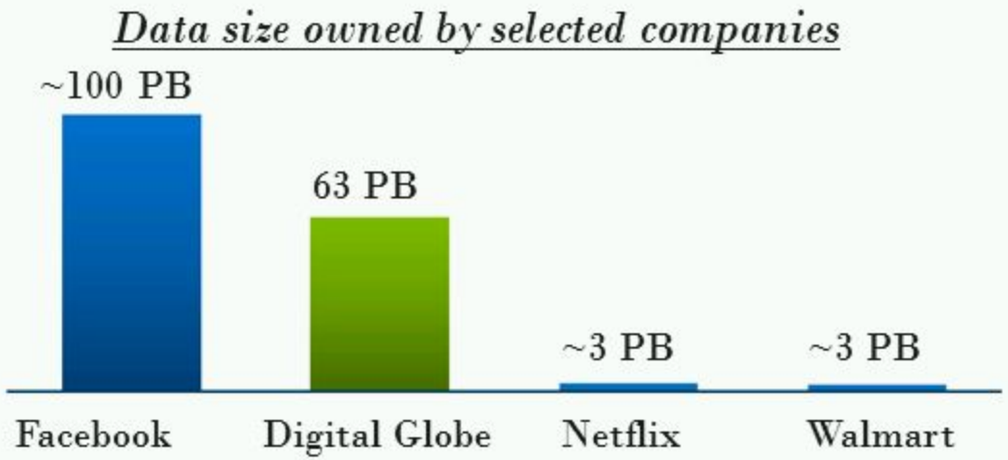
Traditional (physical) industries are being transformed by (IoT) data

Industry	Past – Selling a Product	Future - Service
Healthcare	Diabetes pumps	Diabetes care
Agriculture	Seeds	Crop yields
Consumer	Packaged goods	Nutrition
Automotive	Cars	Transportation
…	…	…
IT Industry	Computers	Computation

McKinsey, GE, IBM, Cisco et al. estimate **hundreds of billions dollar** savings/efficiency improvements in the next 10 years

Big data from the physical world is different than social or transactional data

- Data from the physical world is becoming soon **"mega" big.**
  - 44EByte / months corresponds to **~ 100M hard disks**
  - Processing 44 EByte with 100M servers **would still take more than one hour**
  - **Data throughput** (rather than processing) has become the limiting factor



Sources: Facebook IPO Propectus; Bloomberg May 2013; SAS 2012

- IoT data has more **noise** and is more prone to **error.**
- Data **security**, **verification** etc even more important.
- Data **not always cheap**.

Most analytics on big data is purely data-driven…

	Data-driven Models
Adaptability & deployability	Generic framework, easily adapted to new problems
Domain expertise reliance	Can provide useful results with little domain knowledge
Data availability reliance	Substantial amounts of data to train upon
Fidelity & robustness	Limited to the range of values from the training set and the complexity of the model
Interpretability	Limited
Computation	Big data technologies (high data throughput)

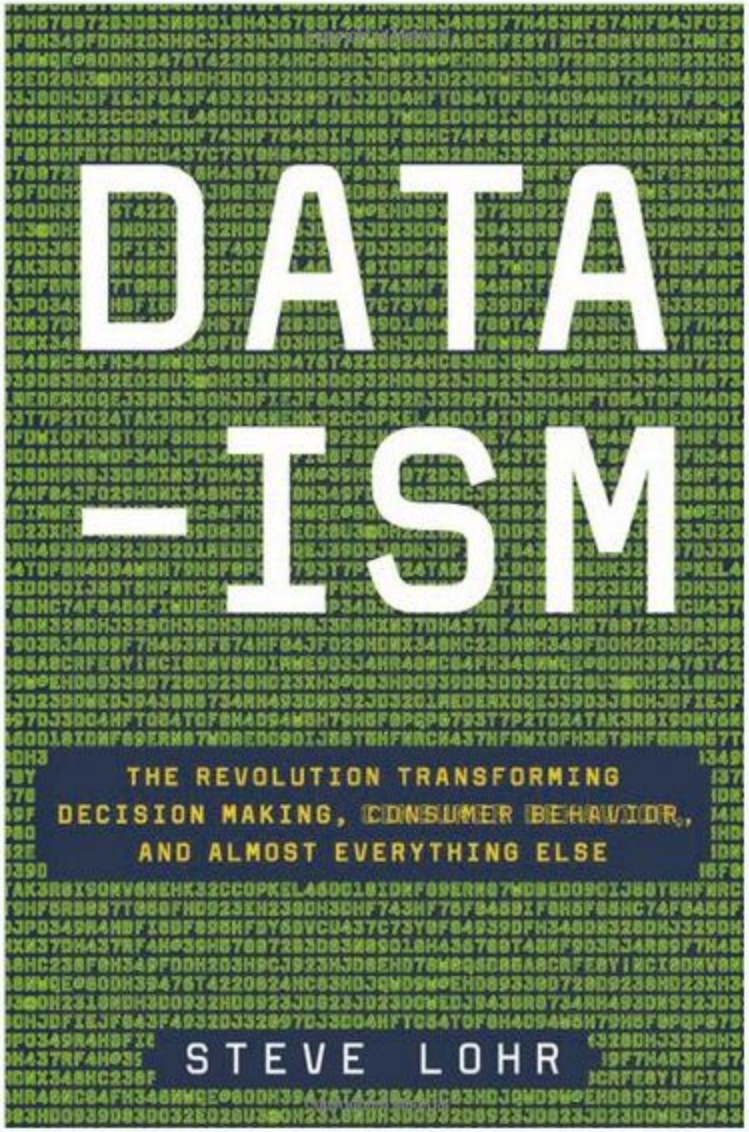
Most analytics on big data is purely data-driven. versus first-principle/domain based

	Data-driven Models	First-principle Models
Adaptability & deployability	Generic framework, easily adapted to new problems	Requires sometimes complex derivations
Domain expertise reliance	Can provide useful results with little domain knowledge	Reliance upon domain expertise
Data availability reliance	Substantial amounts of data to train upon	No / little data
Fidelity & robustness	Limited to the range of values from the training set and the complexity of the model	Universal links can handle highly non-linear and complex relations
Interpretability	Limited	Physically meaningful link between parameters
Computation	Big data technologies (high data throughput)	Traditional HPC (processing throughput)

Physical Analytics Ambition: Combining big da
data-driven and domain models

	Physical Analytics
Adaptability & deployability	Accounts for the relevant physics but still generic enough to easily and rapidly adapt
Domain expertise reliance	Provides useful results with some domain knowledge
Data availability reliance	Right data
Fidelity & robustness	Sufficient robustness and fidelities
Interpretability	Consistent to pinpoint to the relevant physics
Computation	Combination of big data technologies & traditional HPC (high processing & data throughput)

Physical Analytics: (Big) Data gets physical



Copyrighted Material

CONTENTS

1 How Big Is Big Data? 1
2 Potential, Potential, Potential. 13
3 Bet the Company 35
4 Sight and Insight. 61
5 The Rise of the Data Scientist 83
6 Data Storytelling: Correlation and Context . . . 103
7 Data Gets Physical 123
8 The Yin and Yang of Behavior and Data. 143
9 The Long Game 163
10 The Prying Eyes of Big Data. 183
11 The Future: Data Capitalism. 207

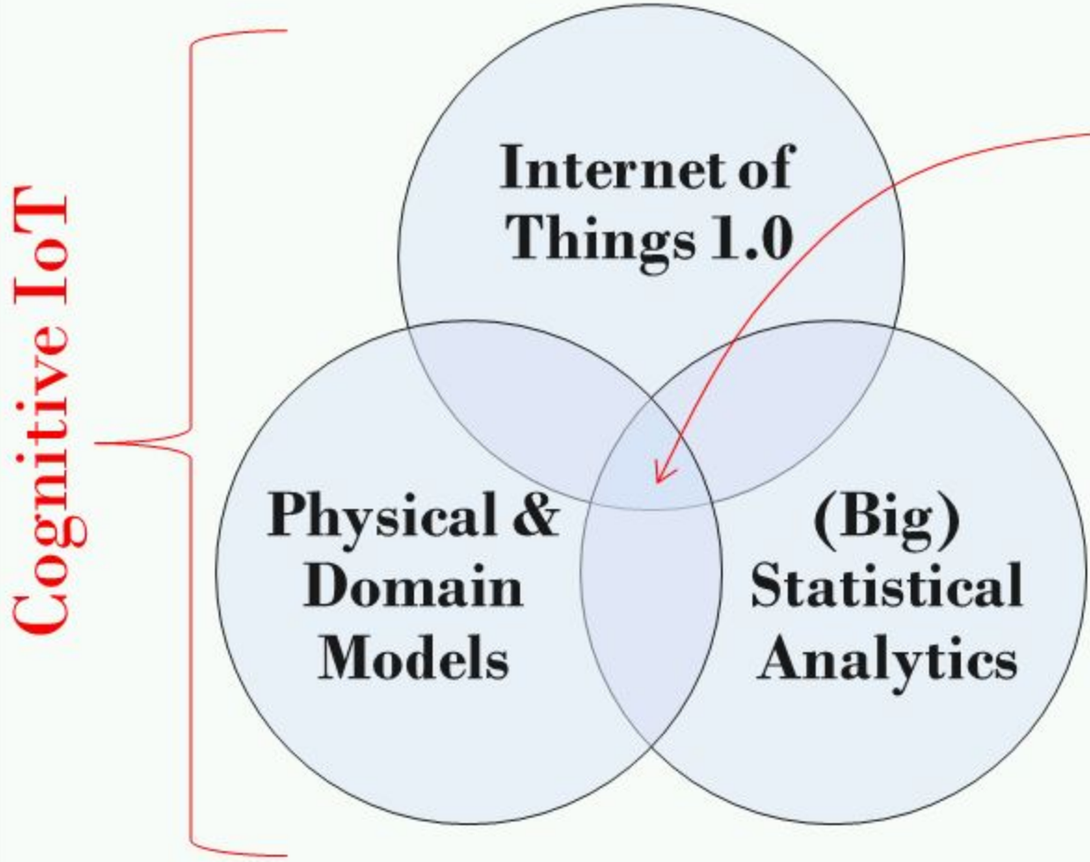
Notes 217
Index 231

x i

Copyrighted Material

 ©2016 IBM Corporation

Physical Analytics lies between IoT 1.0, physics-based modeling and big data statistical analytics



Highly inter disciplinary with many interesting research topics

- Statistical learning under physical constraints
- Model complexity reduction
- Situation-dependent, machine-learnt multi-model blending
- Graph theory and statistical physics
- Parallelizing physical models for data-intensive computation
- …..

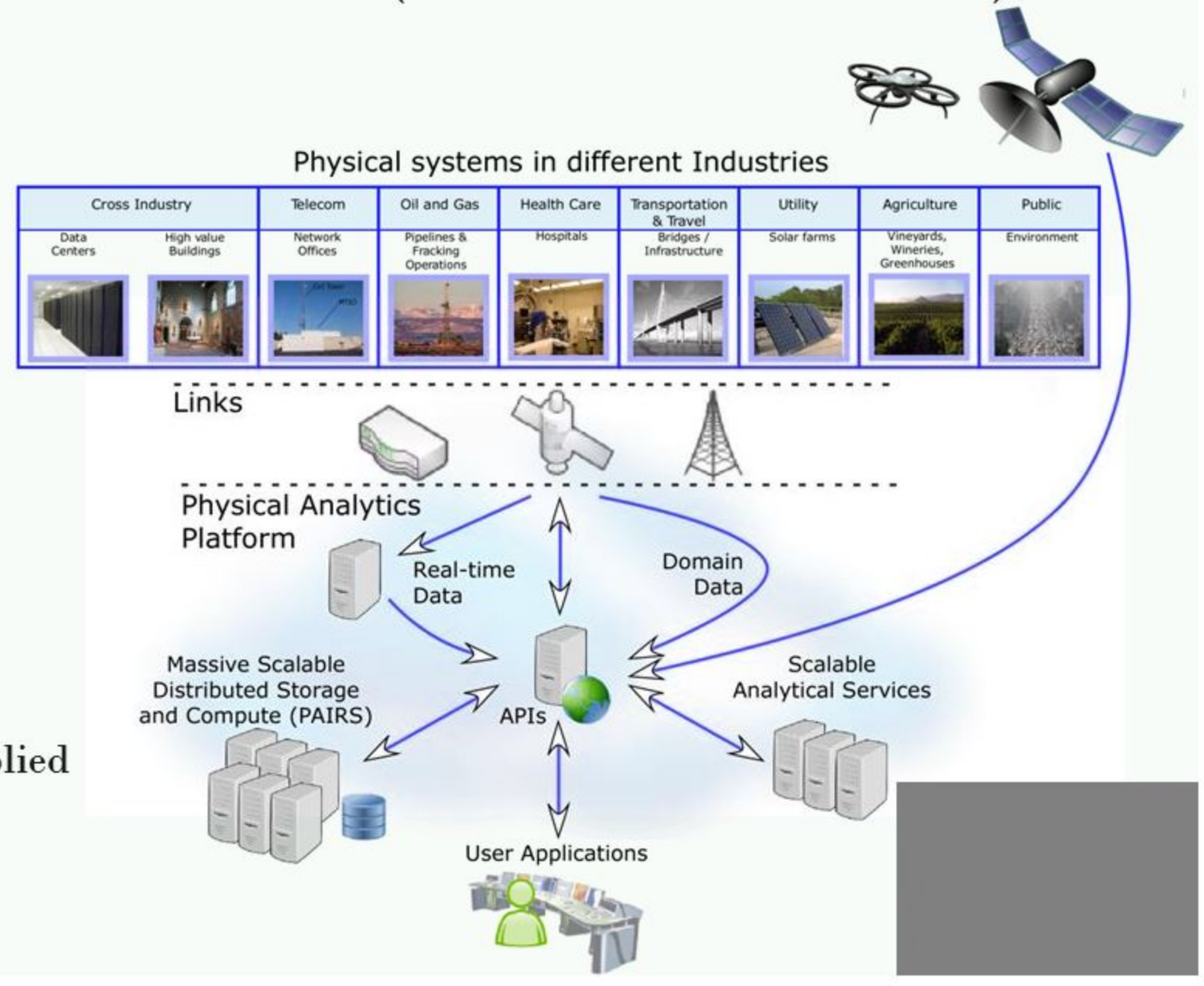
- Feed learning back to improve understanding of the underlying physics

 ©2016 IBM Corporation

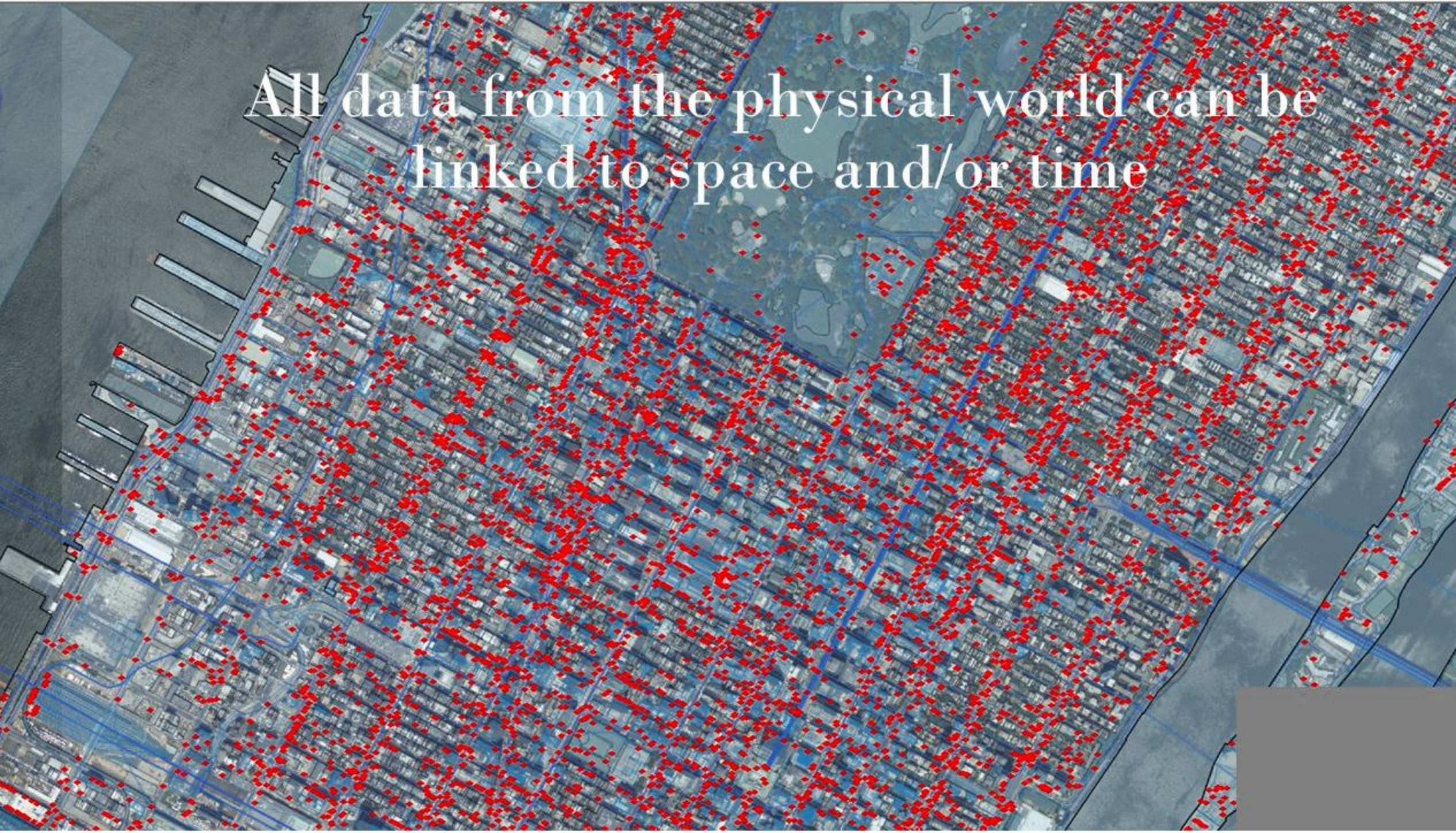
Physical Analytics Platform ("Research IoT")

- Sensor / Actuators
- Communication, Security
- Cloud-based IoT SW Platform
  - Server, Agents, Apps, Clients
- Physical analytics
  - Physical models
  - Reduction of Model Complexity
  - Machine-learnt, multi-model model blending
- (Big) data analytics platform
- Automation and Controls

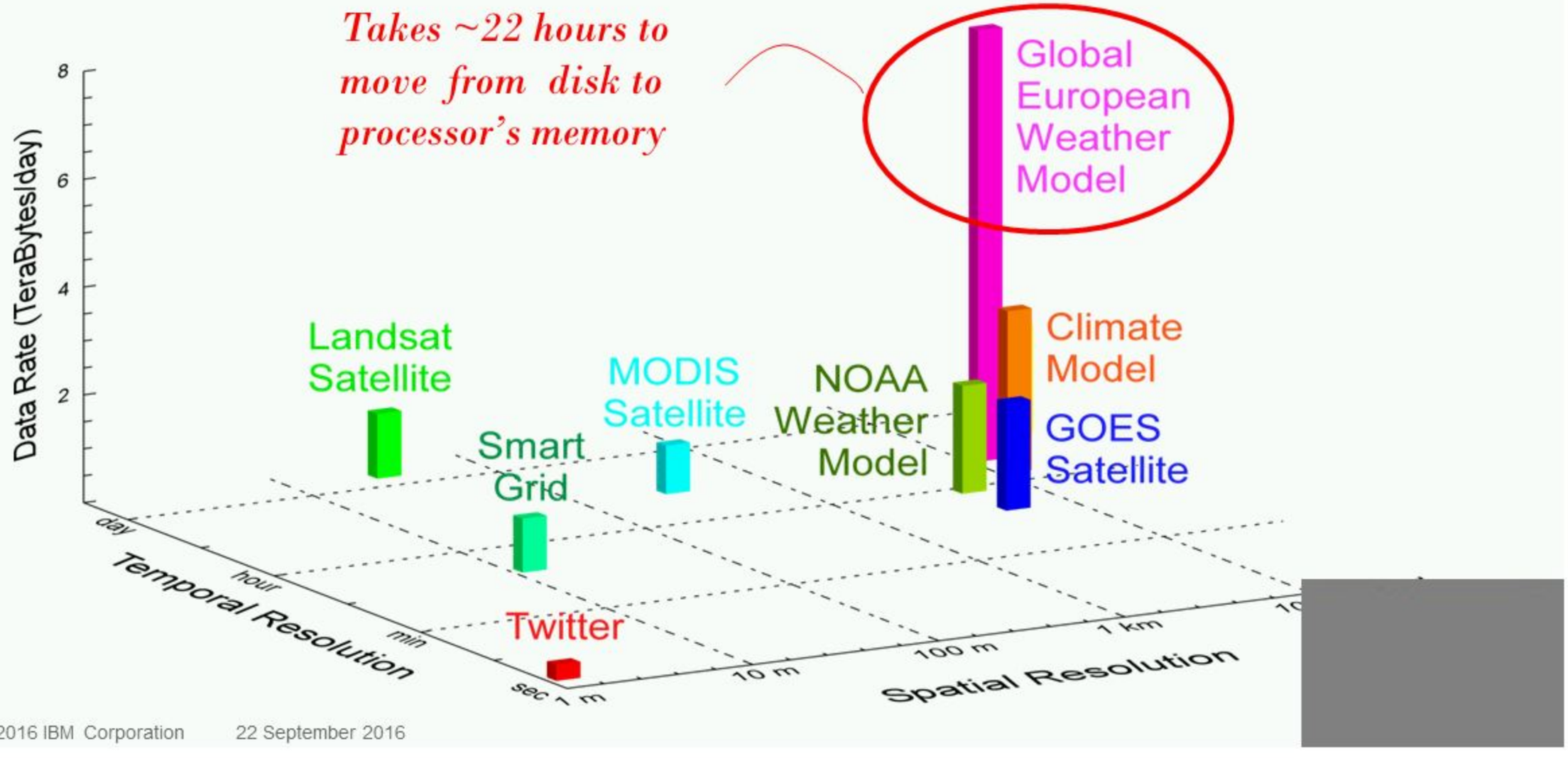
- Platform has been successfully applied in various Industries



 ©2016 IBM Corporation


All data from the physical world can be
linked to space and/or time

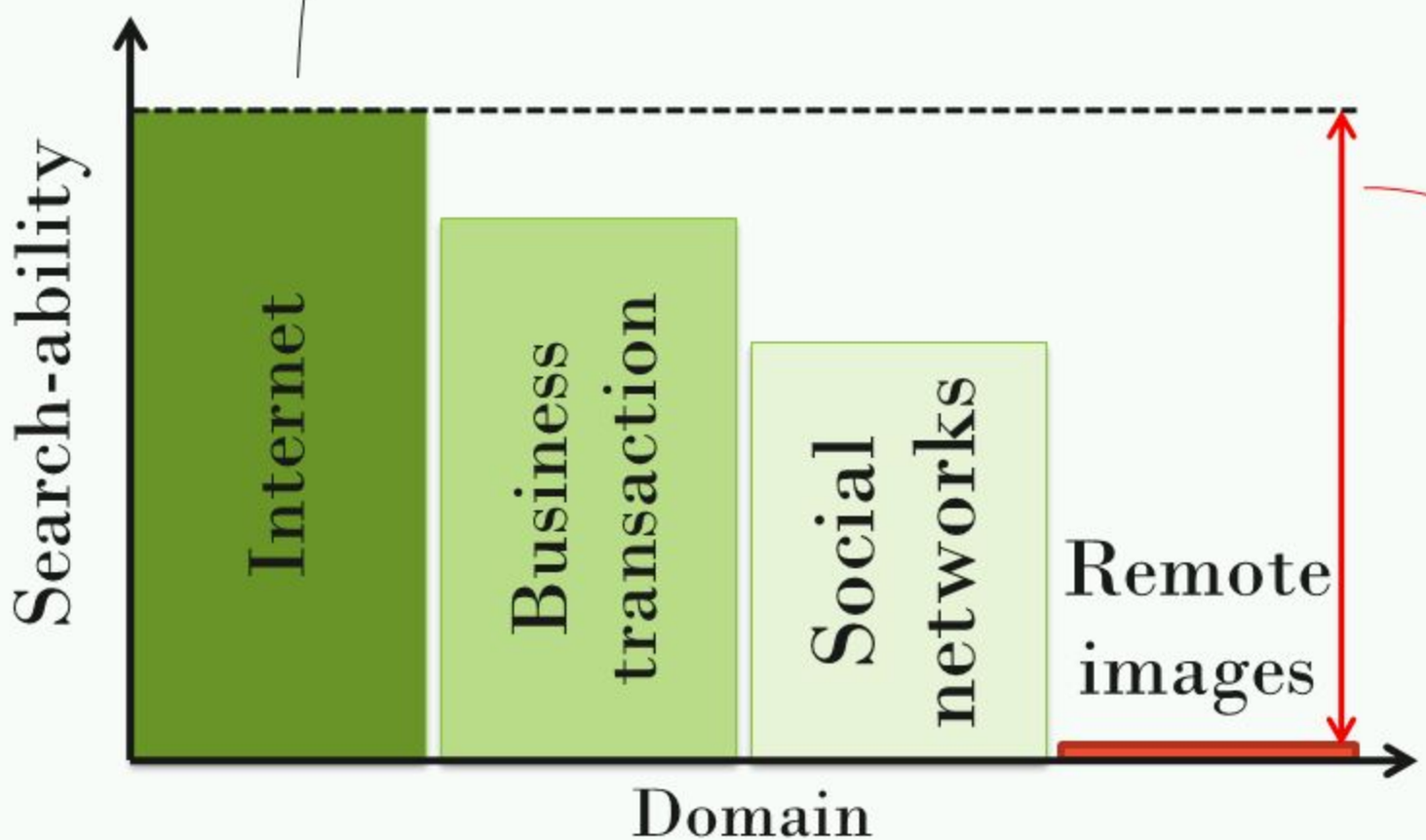
Massive spatio-temporal data sets are generated every day



 ©2016 IBM Corporation

Spatio-temporal data is only coarsely indexed

45 billion web pages have been indexed allowing to search and discover them in ~ 0.5 second



One of the last frontiers for accelerating research and discover lies in the area of remote spatio-temporal images

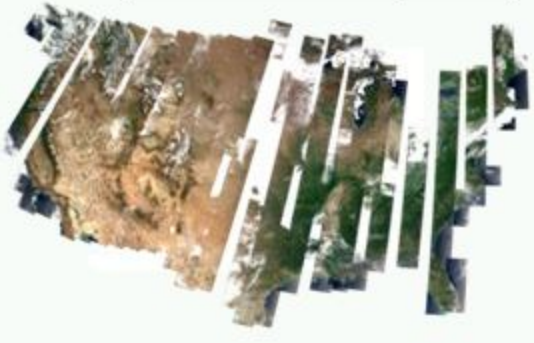
 ©2016 IBM Corporation

Today: Data is moved to the application

TODAY

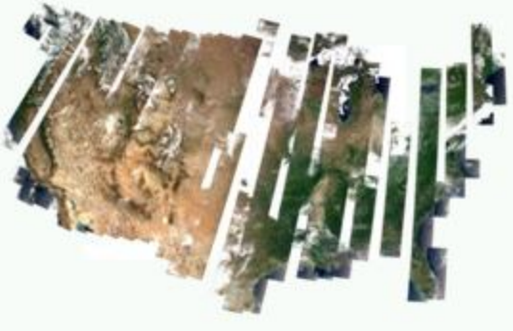
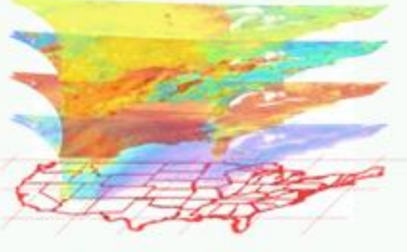
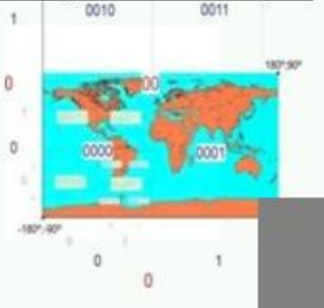
- Each data set is kept at a different place in different formats, projections, units etc.
- Exabytes stored in billions of "individual scenes", or files typically on tape.
- Analysts has to "order" scenes from each source:
  - Download, assemble, re-sample, re-project, align, classify scenes, etc.....
- Data is moved to the application
- Time to value is limited by data curation (90%).

LandSat scenes/tiles from different satellite passes



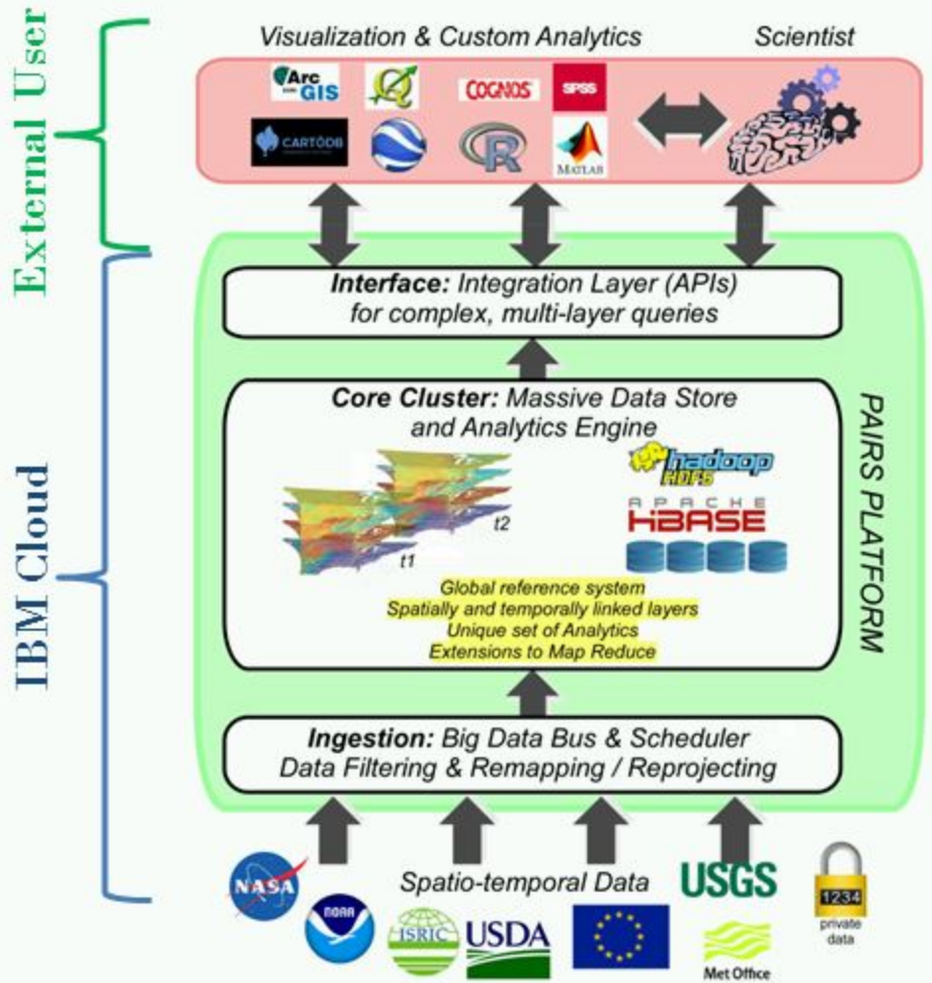
 ©2016 IBM Corporation

PAIRS: Analytics is moved to the data

TODAY	THIS WORK (PAIRS*)
• Each data set is kept at a different place in different formats, projections, units etc. • Exabytes stored in billions of "individual scenes", or files typically on tape. • Analysts has to "order" scenes from each source: – Download, assemble, re-sample, re-project, align, classify scenes, etc..... • Data is moved to the application • **Time to value is limited by data curation (90%).** *LandSat scenes/tiles from different satellite passes* 	• Large scale, pre-processed big data store with pre-aligned layers and data sets – Common formats and projections – Spatial & temporal joins – A global reference system • Accessible as a data service • Analyze without moving data • Access to new layers of analytics. • **Time to value reduced by orders of magnitudes.** *Pre-aligned data sets / layers*  *Global Mastergrid with matching resolution layers*  * Physical Analytics Integrated Repository and Services

 ©2016 IBM Corporation

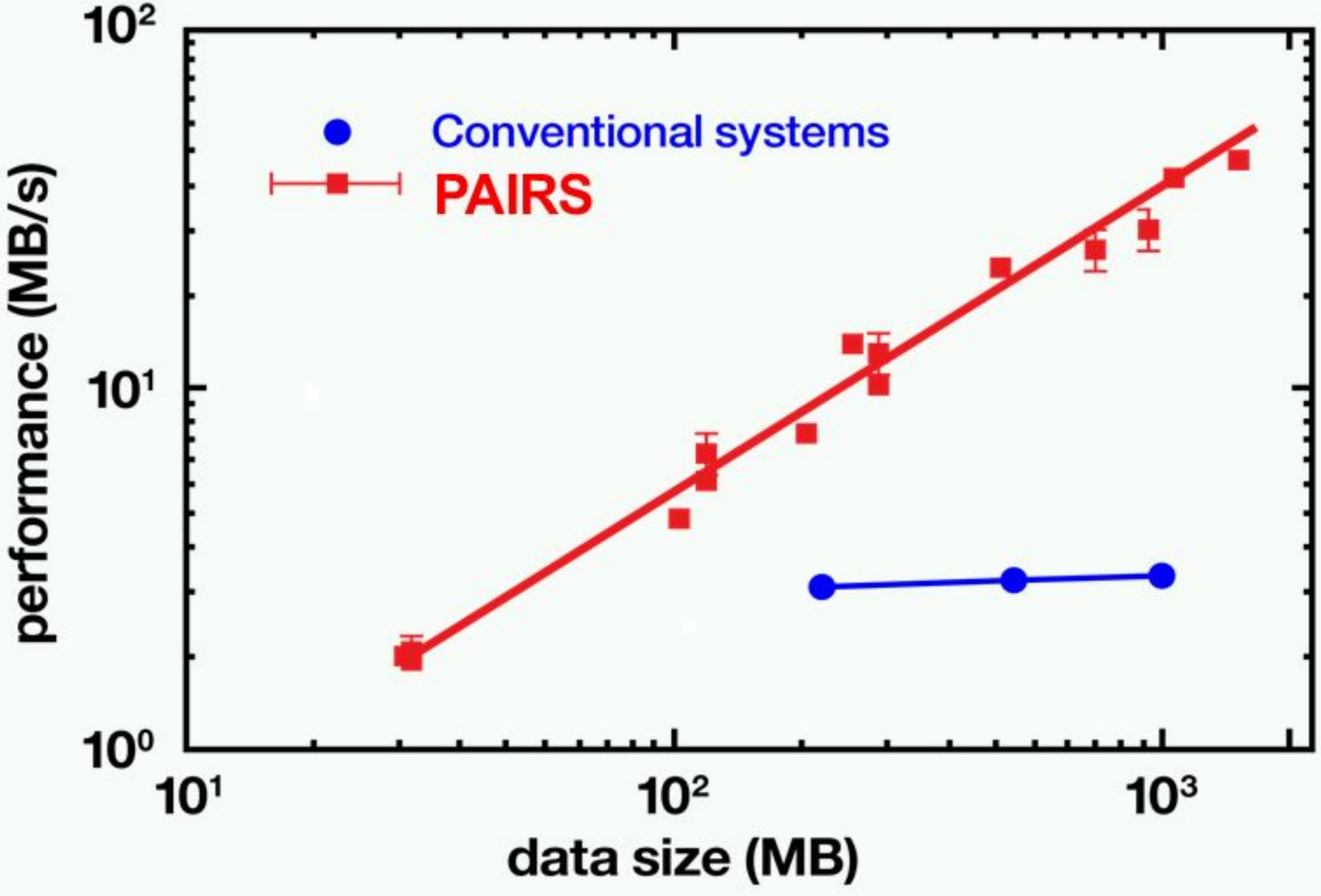
PAIRS: A big data platform for scalable spatio-temporal data and analytics



- Data bus feeds in (near) real-time open spatio-temporal data into PAIRS
- Full data curation process (filtering, classifying, aligning, resampling, reprojecting etc) at ingestion.
- Large-scale Hadoop / Hbase system for efficient distributed data store and processing
- System allows complex queries, e.g.
  - *Find all real estate in California with elevation gradient and high rain fall and certain soil type.*
- **Access to new layers of analytics**
  - Irrigation forecasts
  - Improved weather forecasts
- **Curated data and analytics** accessible **as a Service via** an integration layer
  - REST APIs to run queries
  - Basic web interface to run queries

 ©2016 IBM Corporation

PAIRS "scales" to big data and complex analysis



- PAIRS queries (almost) independent of data size
- Conventional systems require more time of larger data sizes

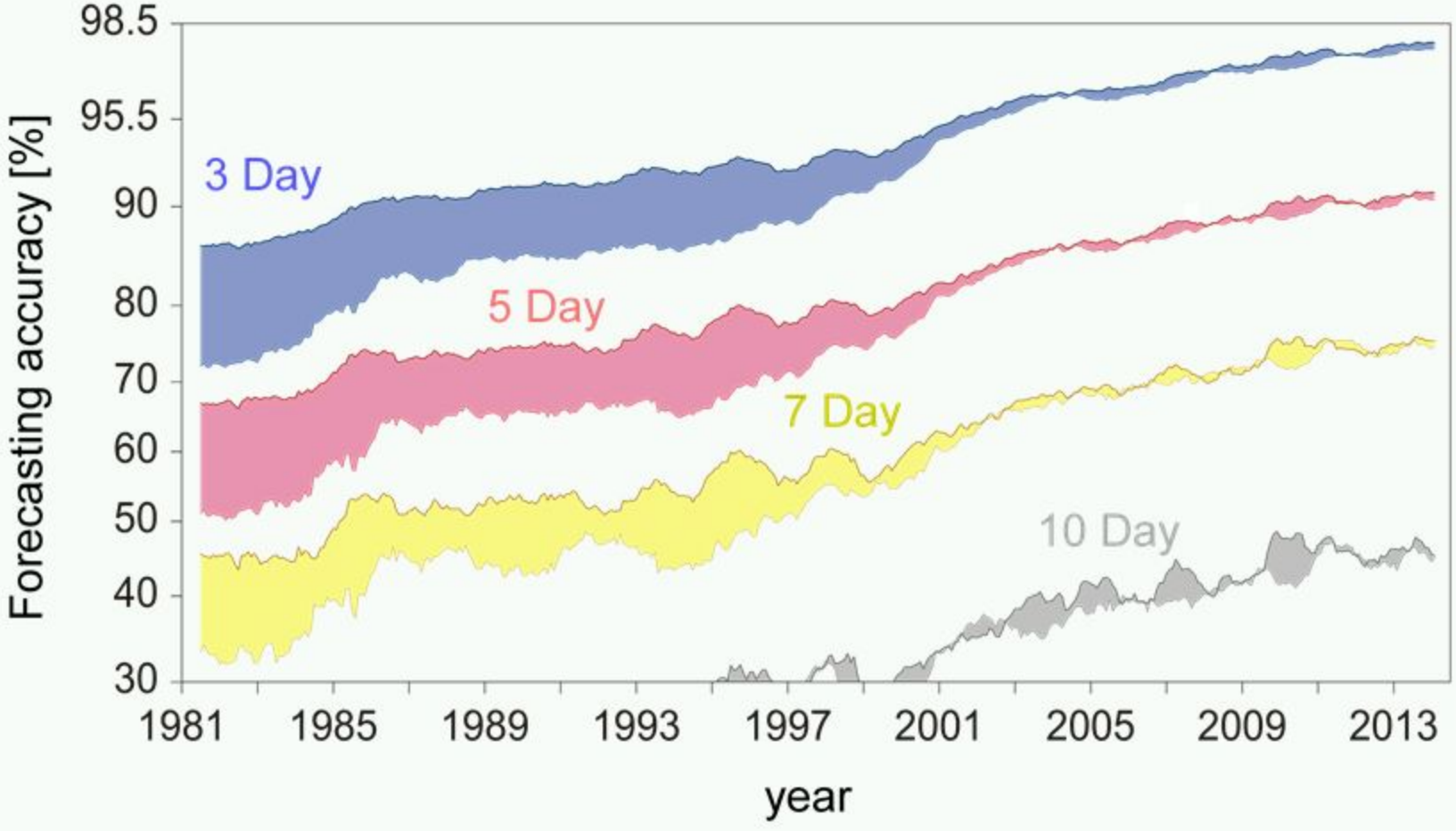
 ©2016 IBM Corporation

To summarize so far….

1. The world is being digitized at an accelerated rate transforming industries
2. A new form of physical analytics - combing statistical and physical models - are required to exploit the opportunity
3. All data from the physical world can be linked in space and time.
4. PAIRS technology allows scalable physical analytics on spatio-temproal data

Applications: Seasonal Weather Forecasting

Historically, NWP* model accuracy improvements have been (only) ~6% per** decade…How can we make a big impact?



* Numerical weather prediction
** Peter Bauer, Alan Thorpe & Gilbert Brunet doi:10.1038/nature14956

 ©2016 IBM Corporation

Improved global **seasonal** forecasting via machine-learnt, situation-dependent model blending

PAIRS data layers:

- Multiple climate forecast models
- > 1600 weather station data, COSMIC GPS-RO sounding

Physical Analytics:

- Machine-learnt, situation dependent, multi-model blending using historical forecasts and weather data

New PAIRS layer(s):

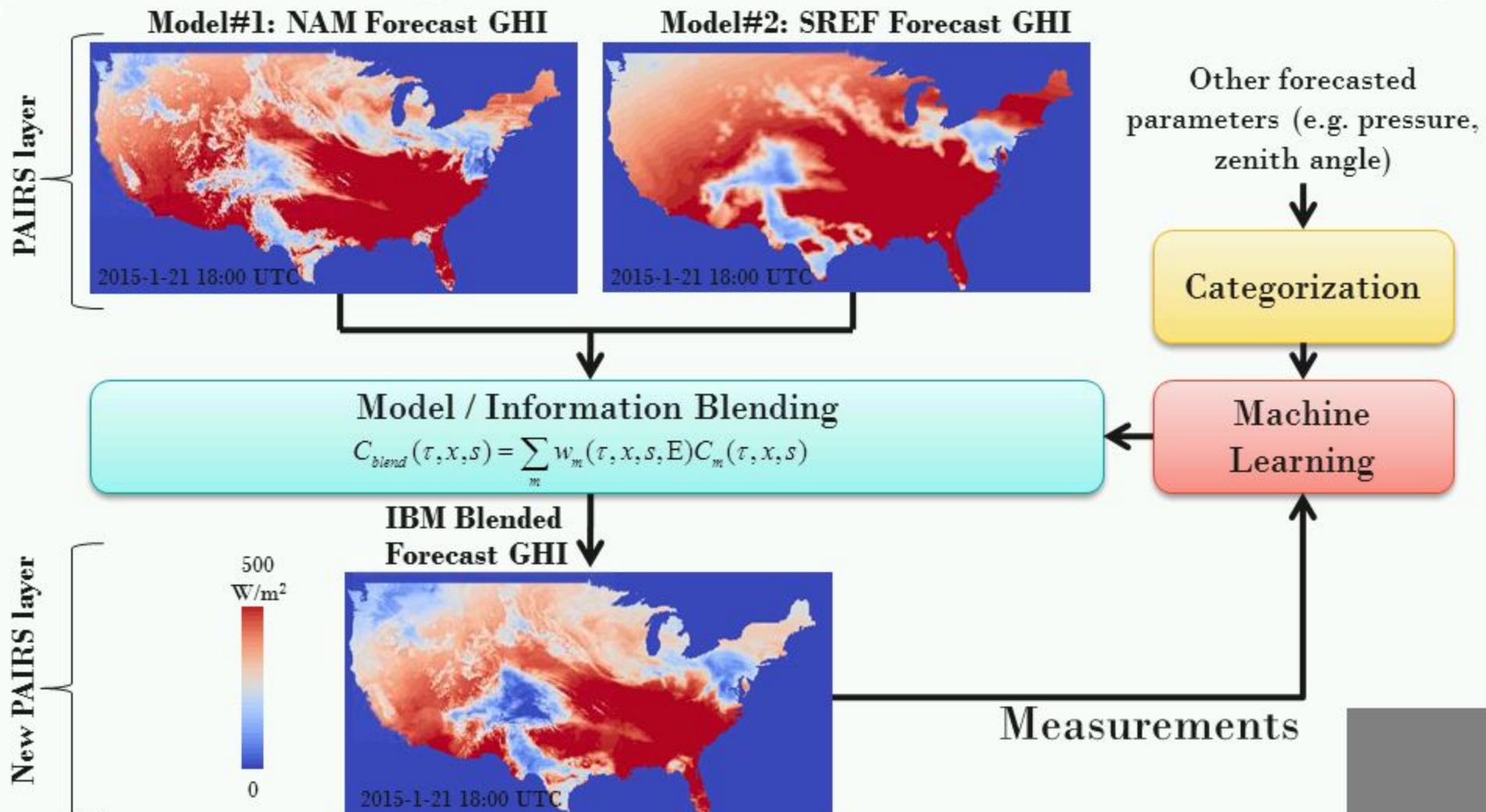
- Improved climate forecasts

Hurricane Ike path forecasts from 8 different weather models*

Climate Models	NOAA CFS v2	ECMWF ENS Extended	ECMWF SEAS	EUROSIP	Beijing CC CGCM	Tokyo CC AGCM	....
Spatial Res & Coverage	0.5 Deg global	0.4 Deg global	0.75 Deg global	2.5 Deg global	2.5 deg global	2.5 deg global	
Temporal Resolution	6 Hourly	6 Hourly	6 Hourly	6 Hourly	Monthly	Weekly	
Forecasting Horizon	0 to 6 months	0 to 45 days	0 to 7 months	0 to 6 month	0 to 11 months	0 to 3 mont	
Ensemble Forecast	4 Members	51 Members	51 Members	41 Members	Ensemble Mean	Ensemble M	

 ©2016 IBM Corporation

Creating a super forecast model via machine-learnt, situation dependent multi-model blending



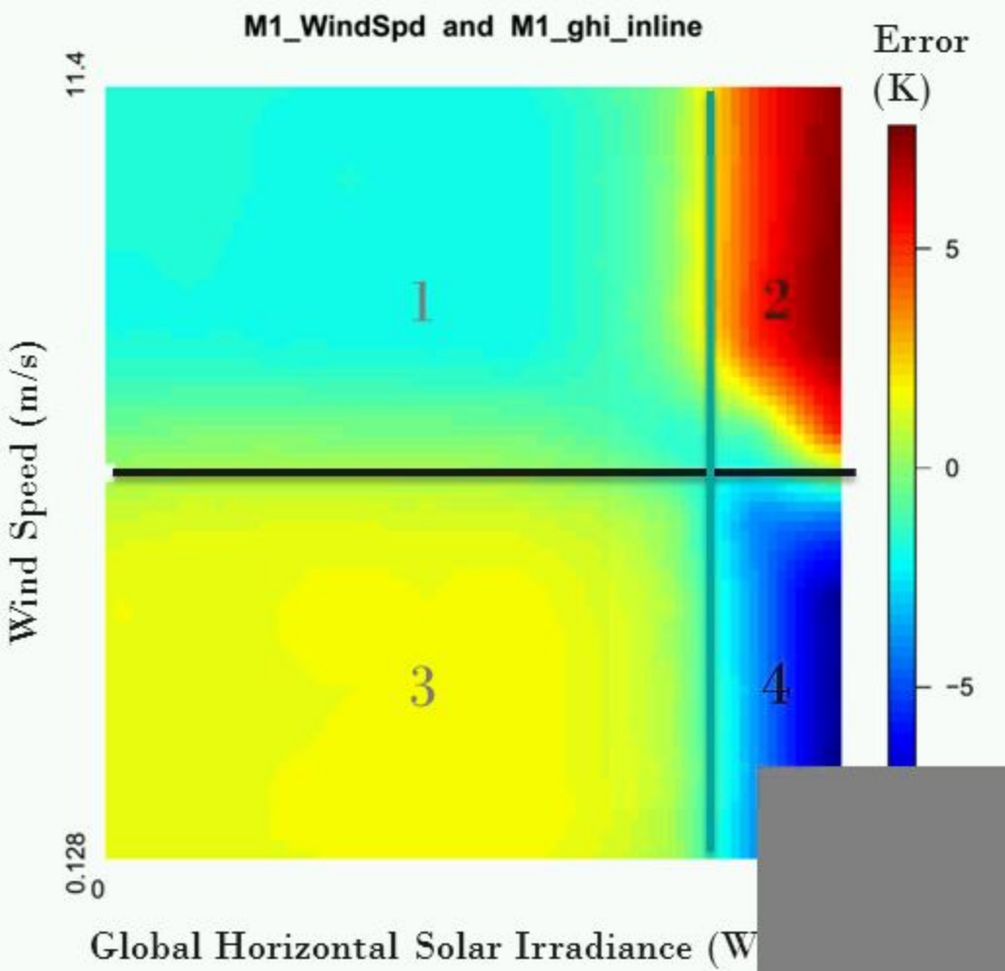
 ©2016 IBM Corporation

Situation-dependent model blending

Which model is more accurate, when, where, under what weather situation?

- Apply functional analysis of variance to understand 1st,2nd,3rd, ….order errors
- Model accuracy can depend strongly on **"weather situation"** category.
- "Weather situation" is determined using a set of parameters including forecasted ones on which model error depends on strongly.

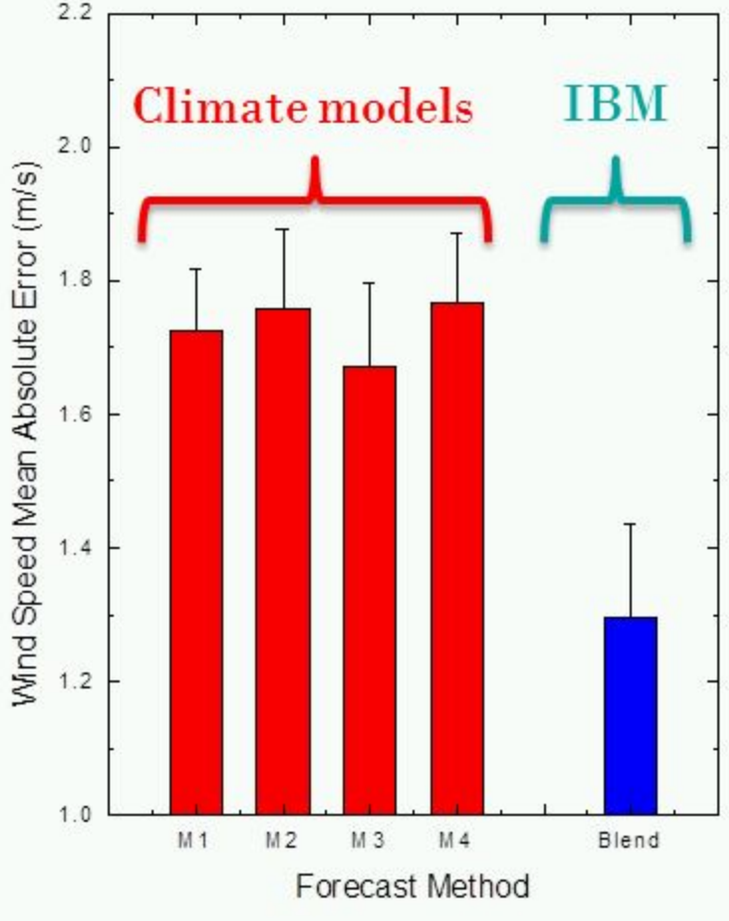
2nd Order Error
NOAA CFSv2 Model 1, Temperature Forecast
30 days ahead, 6 hourly, Surfrad Stations BND



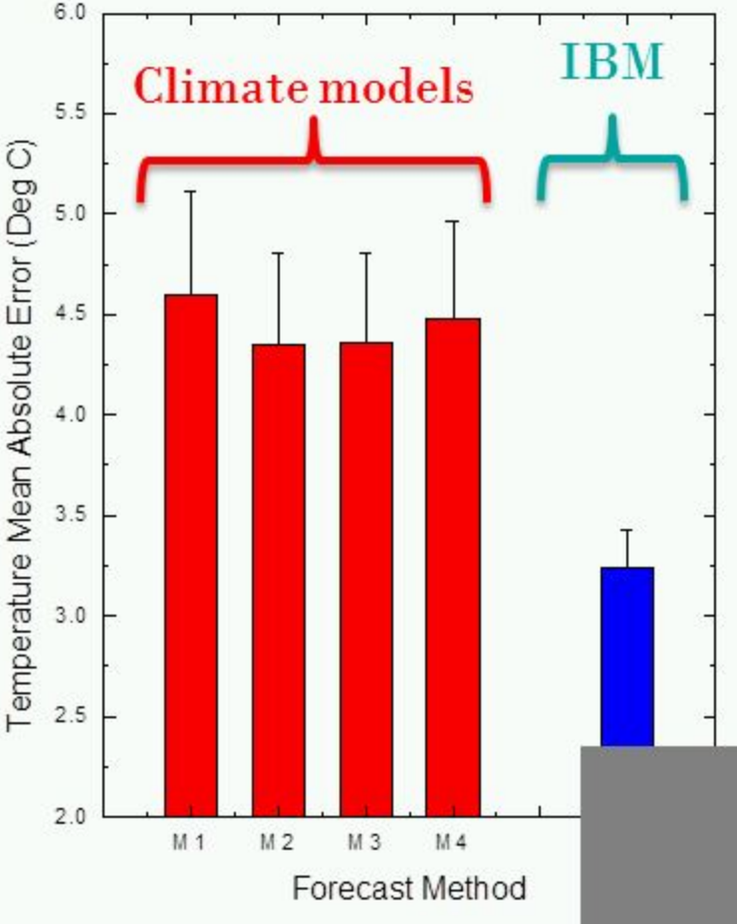
 ©2016 IBM Corporation

More than 30% error reduction for 30 day-ahead forecasting

10m Wind Speed, 30 Day-Ahead



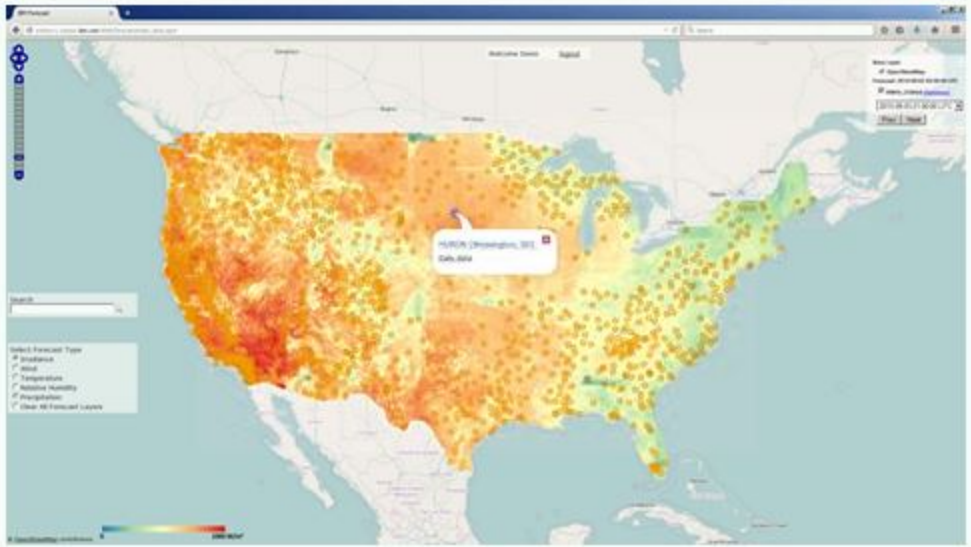
2m Temperature, 30 Day-Ahead



7 Surfrad Stations
Validation
06/20/15 – 09/20/15

©2016 IBM Corporation

We scaled the technology to US continental wide forecasting and beyond



- > 35 % improved accuracy with respect to next best model at 1600 sites across the United States
- PAIRS provides gridded forecasts
- Continuously learns and improves

Performance Metrics

2015-05-10 to 2015-05-24

IBM Blended GHI 0-48hr Ahead

SiteSet	# of Sites	MAE* (W/m^2)	MAPE(%)	RMSE (W/m^2)	NRMSE (%)	MBE (W/m^2)	MaxAE (W/m^2)	KSIPer (%)	pcoe
RAWS_CONUS	1641	135.54	13.55	192.99	19.29	19.84	668.06	16.62	0.78

NOAA NAM GHI 0-48hr Ahead

SiteSet	# of Sites	MAE* (W/m^2)	MAPE(%)	RMSE (W/m^2)	NRMSE (%)	MBE (W/m^2)	MaxAE (W/m^2)	KSIPer (%)	pcoe
RAWS_CONUS	1641	179.06	17.91	253.60	25.37	118.94	814.81	22.27	0.75

NOAA SREF GHI 0-48hr Ahead

SiteSet	# of Sites	MAE* (W/m^2)	MAPE(%)	RMSE (W/m^2)	NRMSE (%)	MBE (W/m^2)	MaxAE (W/m^2)	KSIPer (%)	pcoe
RAWS_CONUS	1641	188.44	18.84	262.97	26.30	129.56	837.83	23.64	0.74

*Metrics Suites: Excluding night time values for solar. Hover mouse over the acronyms to see full definition.

IBM Blended GHI 0-48hr Ahead

 ©2016 IBM Corporation

Applications: Crop Forecasting

Crop recognition

PAIRS data layers:

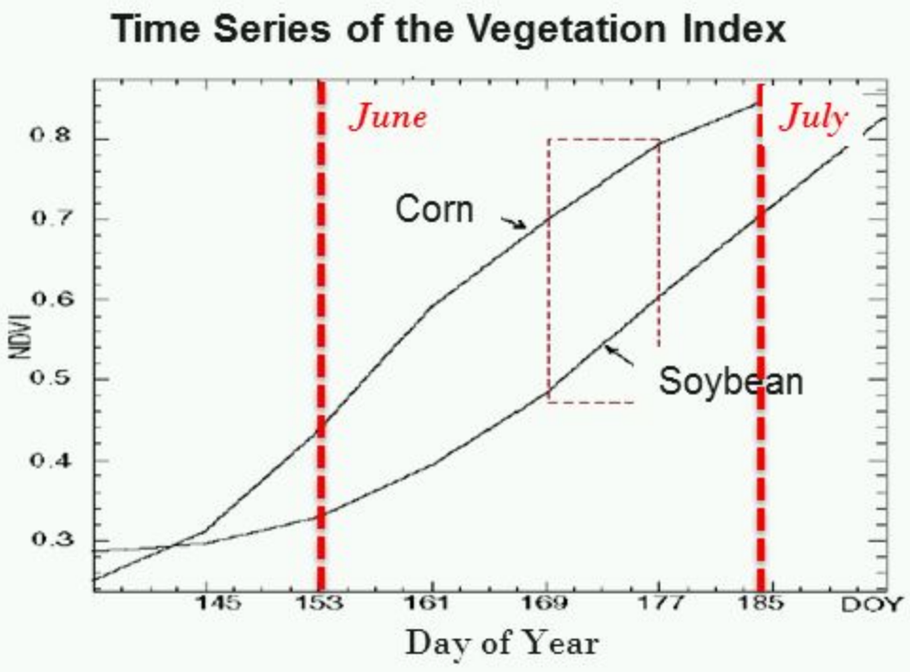
- Satellite data,
- Soil information
- Weather data
- Historical crop maps

Cognitive computing:

- Time-series analysis of NDVI evolution leveraging a 30 year record of satellite and survey data

New PAIRS layer(s):

- Crop maps early in the season

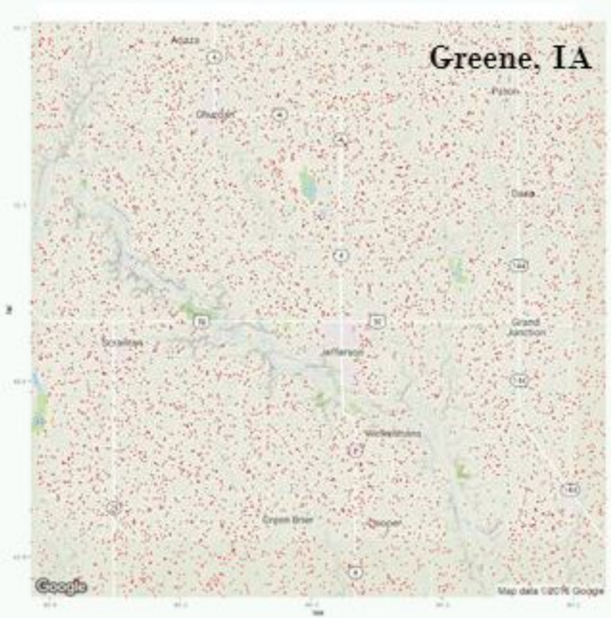


Vegetation Index Map

 ©2016 IBM Corporation

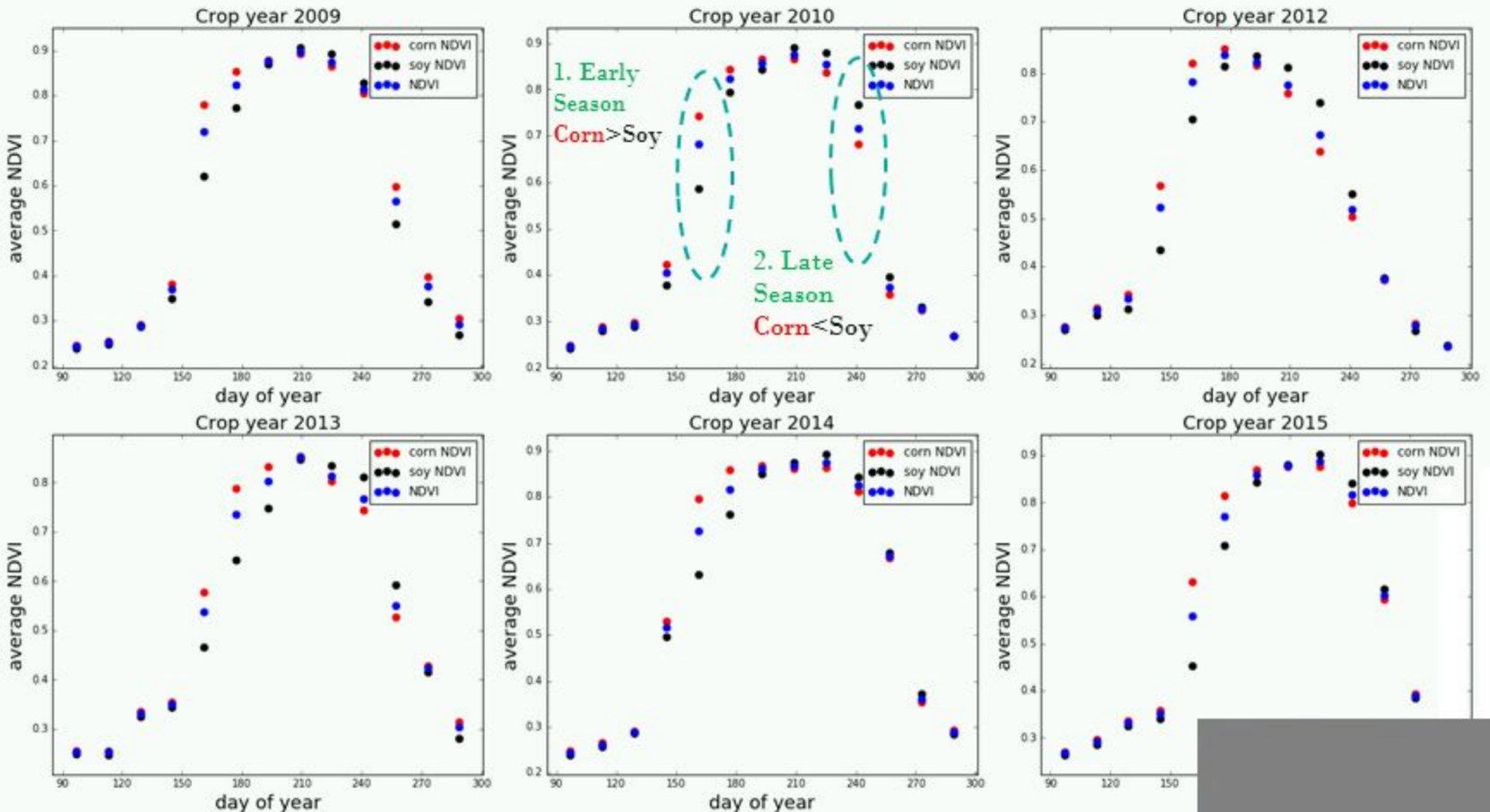
Acreage Prediction using Remote Sensing

Sampling good center-field locations (large enough for MODIS pixels) for Greene and Madison counties.



- Future work: field boundary detection using LANDSET. Enable sample points to better represent actual farm size.

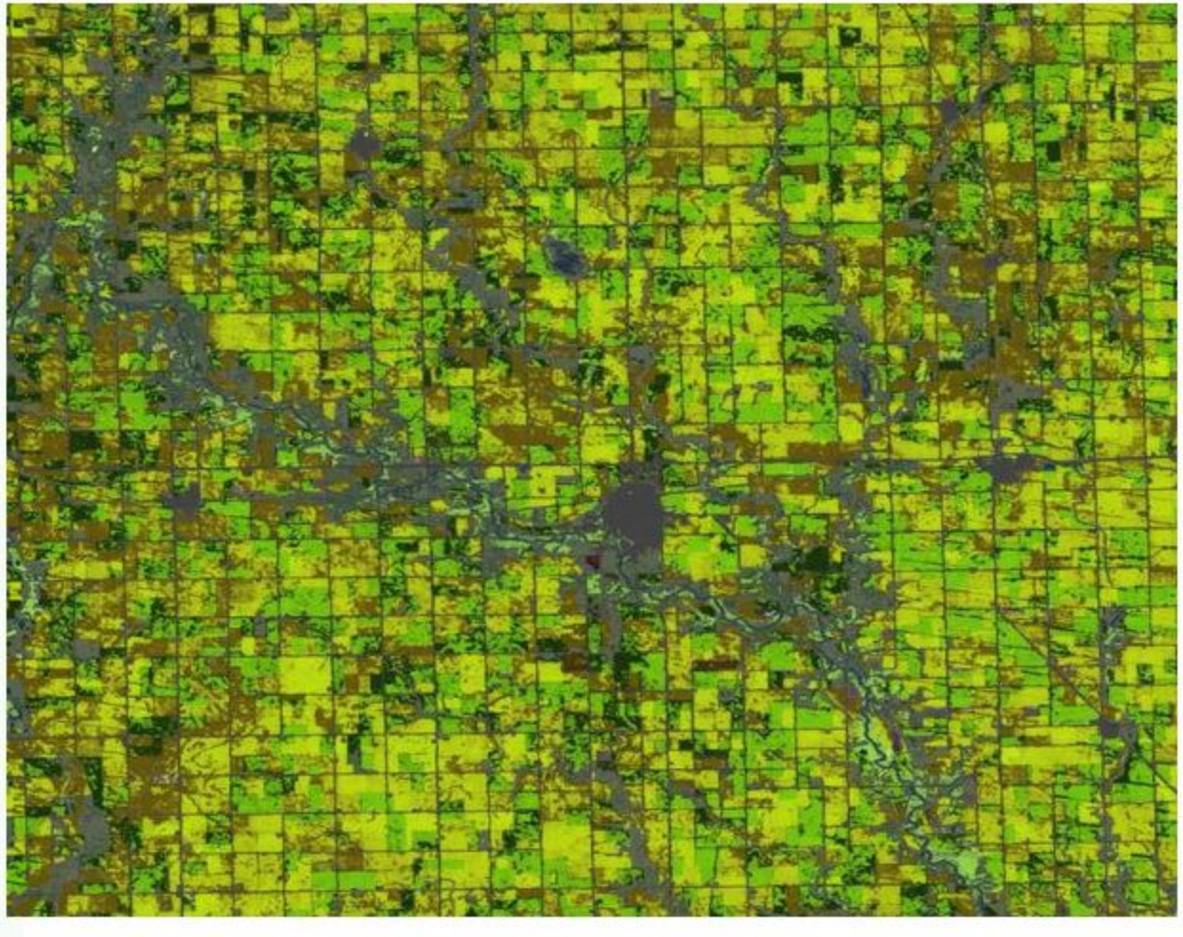
Corn vs. Soy -- NDVI as a function of



- Corn NDVI rises faster than soy in spring. (NDVI 1, early season)
- Corn NDVI falls faster than soy in late summer / autumn. (NDVI 2, later

©2016 IBM Corporation

Early Crop Recognition
Greene County, IA - June 2013 Predictions



NDVI Prediction of early June 2012:

Percentage of Area in Greene County, IA

Corn:	51 %
Soybean:	29 %

 ©2016 IBM Corporation

Early Crop Recognition
Greene County, IA - NASS Results

NDVI Prediction of early June 2012:

Percentage of Area in Greene County, IA

Corn:	51 %
Soybean:	29 %

NASS Results at Jan 2013:

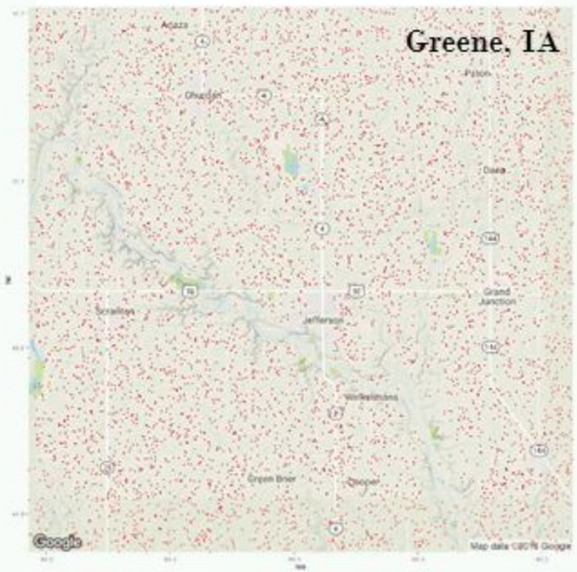
Percentage of Area in Greene County, IA

Corn:	49 %
Soybean:	29 %

 ©2016 IBM Corporation

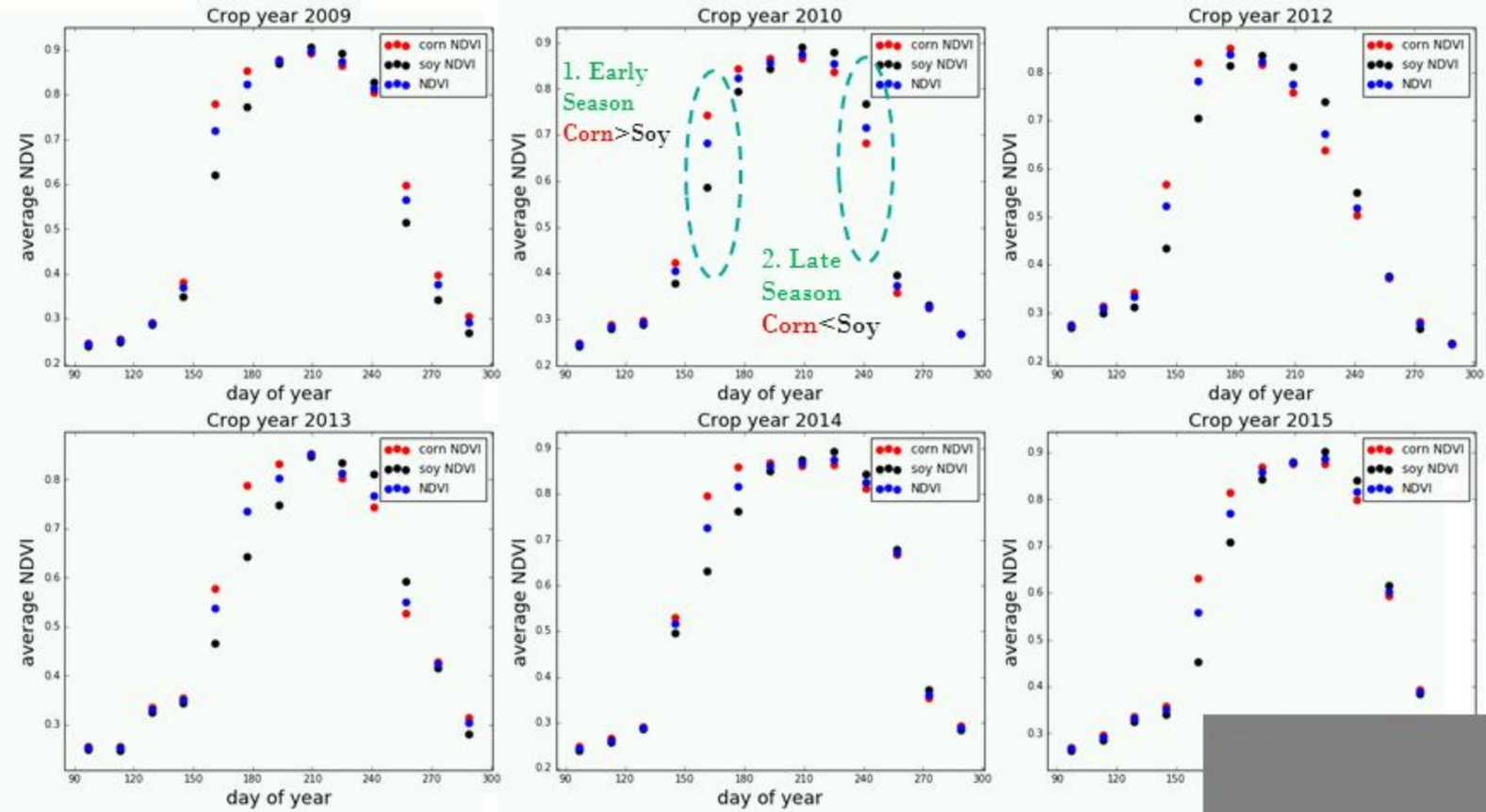
Acreage Prediction using Remote Sensing

Sampling good center-field locations (large enough for MODIS pixels) for Greene and Madison counties.



- Future work: field boundary detection using LANDSET. Enable sample points to better represent actual farm size.

Corn vs. Soy -- NDVI as a function of



- Corn NDVI rises faster than soy in spring. (NDVI 1, early season)
- Corn NDVI falls faster than soy in late summer / autumn. (NDVI 2, later

©2016 IBM Corporation

Acknowledgement

Conrad Albrecht
(Physics and Computation
Heidelberg PhD)

Levente Klein
(Physical Modeling)

Fernando Marianno
(Software Architect)

Michael Schappert
(Embedded System)

Marcus Freitag
(Precision Agriculture)

Hendrik Hamann
(Physical Analytics)

Xiaoyan Shao
(Electrochemistry,
Data scientist)

Vanessa Lopez
(Mathematics,)

Jun Song Wang
(Sensing)

Siyuan Lu
(Machine learning)

Bertrand Marchand
(Solar Forecasting)

Wang Zhou
(Robotics, drones)

Josphine Chang
(Sensor Platform, IoT)

Ramachandran Muralidhar
(Corrosion Science &
Pollution Modeling)

Theodore van Kessel
(Oil and Gas &
Instrumentation)

Bruce Elmegreen
(Astrophysics, Traffic)

Golnaz Badr
(Precision Agriculture)

and many m

©2016 IBM Corporation