Big Data and Predictive Analytics

New tools for managing delinquency

Jeff Harris, VP Predictive Analytics

Financial Services Group, Xerox Services

January 28, 2013

Agenda

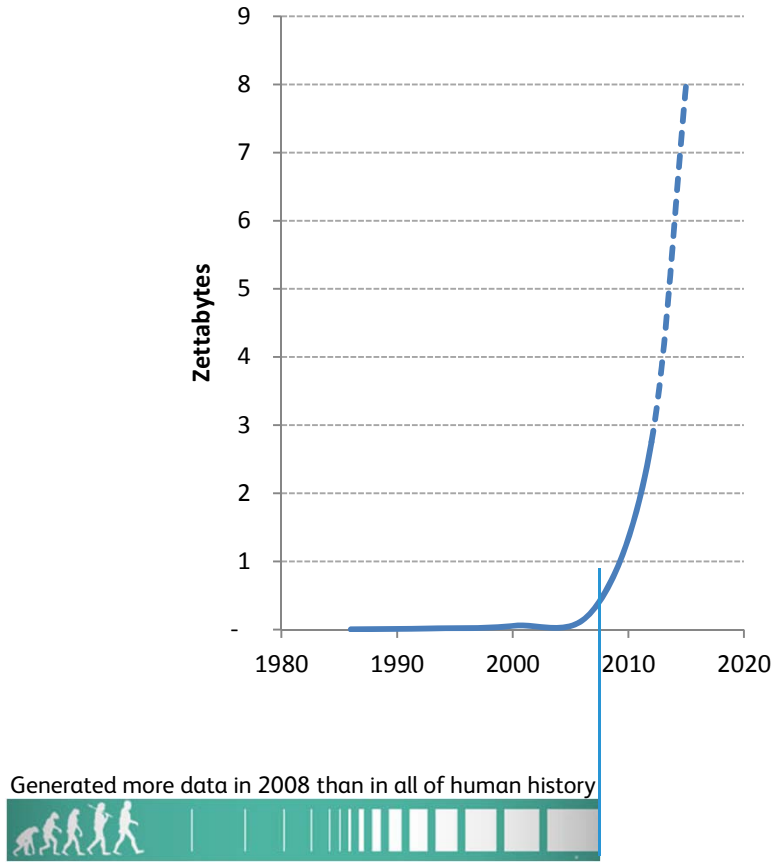
- Big Data
  - What it is
  - Why it matters
  - How to use it
- Data & Privacy
- Walkthrough
- What we do
- Conclusion
- Q&A

WHAT IS BIG DATA

We are in the midst of an information explosion…

- Big Data refers to data sets at scales larger than what traditional databases were designed to handle

Unique data generated annually



A sense of scale

Terabyte	1 terabyte will hold 200,000 mp3 songs or 25 Blu-Ray discs
Petabyte	Will fill 2 datacenter server cabinets
Exabyte	Will fill 2,000 cabinets, in a 4 story data center taking up a city block
Zettabyte	Will fill 1,000 datacenters; about 20 % the size of Manhattan
Yottabyte	Will fill 1,000,000 data centers the size of Rhode Island and Delaware

…driven by the convergence of several trends…

- Progression of Moore's Law
- Proliferation of devices
  - Sensors, smart phones, cameras
- Interconnectedness of things
- Rise of social media
  - Facebook
  - Twitter
  - LinkedIn

… leading to new technologies to handle

- Internet companies (Google, Yahoo!, Facebook) at the forefront
  - Focus on scalability of infrastructure to handle data growth
- Trend toward open source
- Developments in High Performance Computing
  - Encompasses everything from supercomputers (IBM's Blue Gene), to clusters and grid computing
- Virtualization & Cloud Infrastructure

IBM's Blue Gene: 250,000 processors in 72 racks

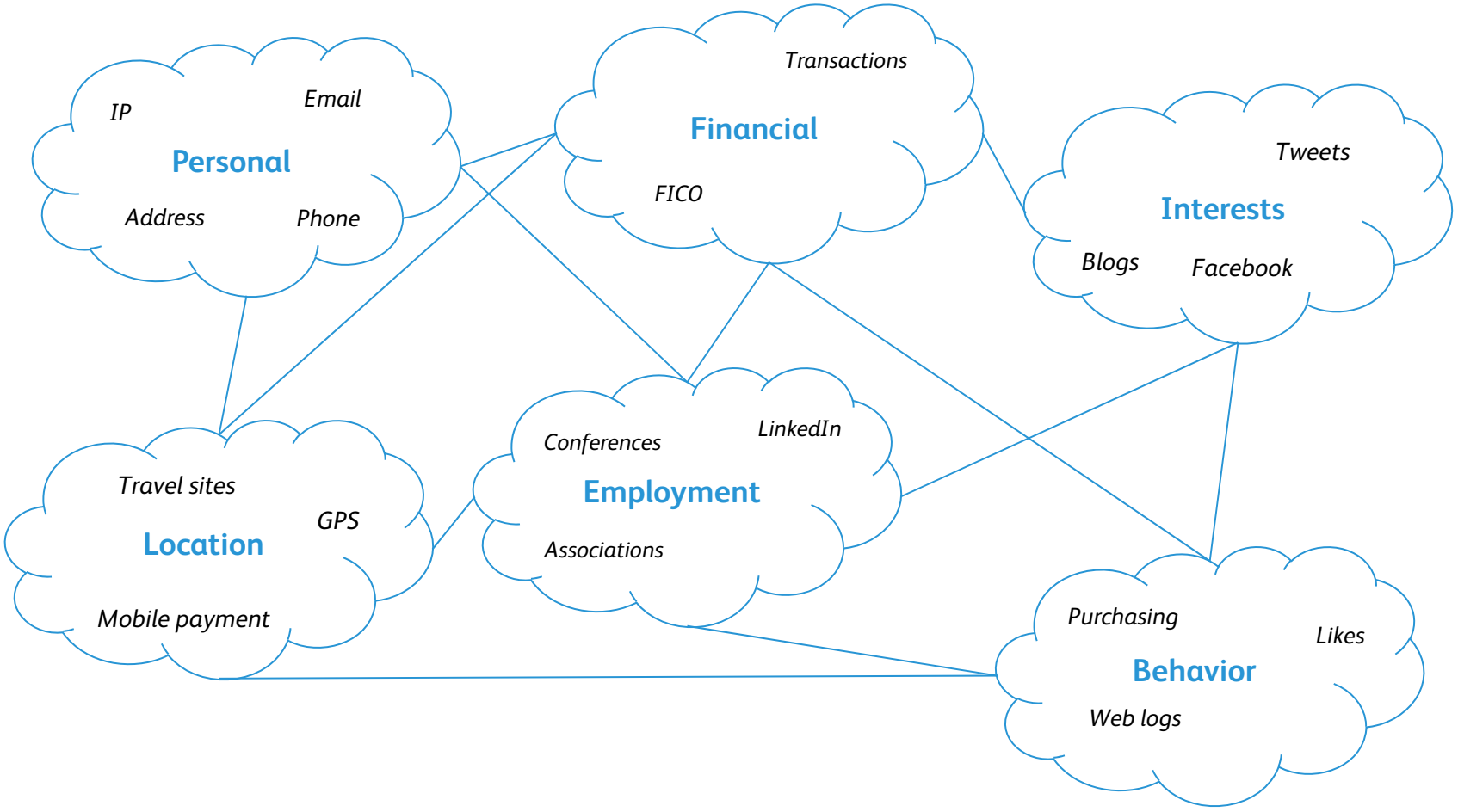
WHY BIG DATA MATTERS

Big Data makes the unit the level of analysis…

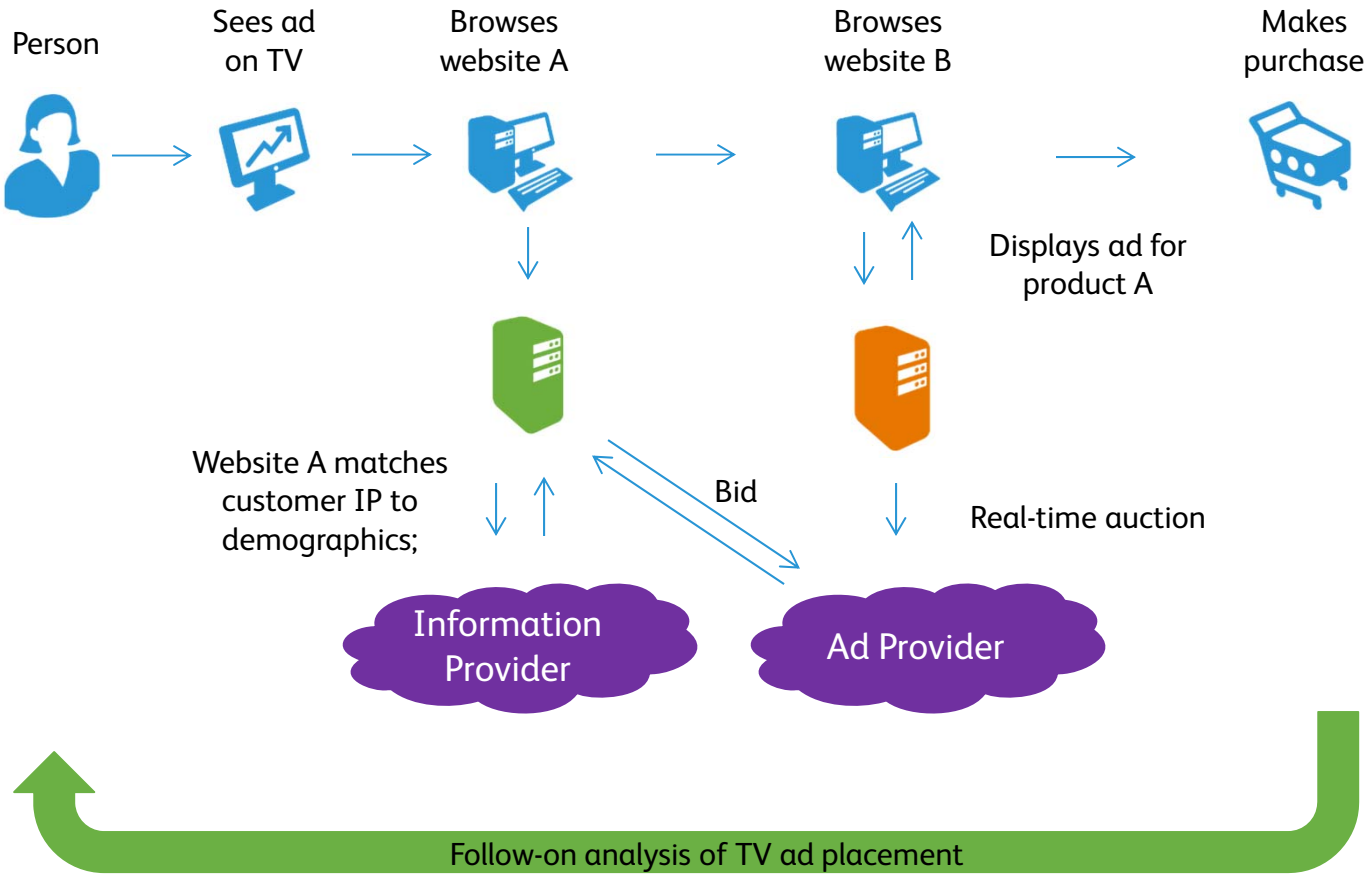
Sector	Old (group; coarse)	New (unit; granular)	Application
Weather	Annual / Monthly	Continuous	Dynamic watering
Agriculture	Farm / acreage	1-foot diameter	Variable rate fertilizer application
Entertainment	Movie critics	Individual viewing (rental) history	Personalized recommendations
Healthcare	Body	DNA	Gene targeted drugs
Finance	Stocks	Individual trades	High frequency trading
Semiconductor	Transistor	Atom	New materials: graphene

…and the field of vision panoramic

Big Data provides the information and tools to link data across multiple dimensions…

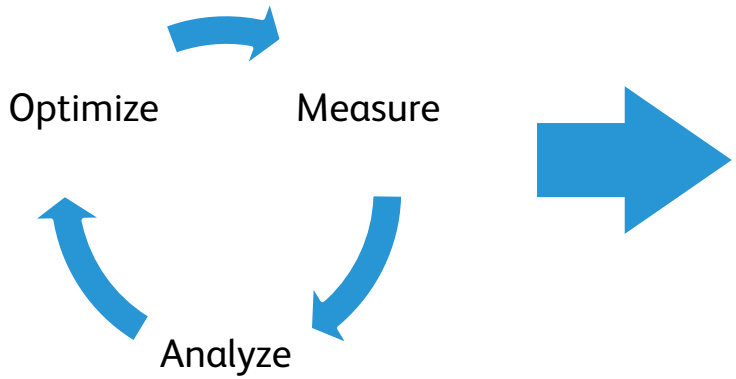


… and to do it virtually instantaneously



Big Data changes the way everything is being done

Continuous Feedback Loops



Leads to

- Economize scarce resources
- Mass personalization
- Rapid experimentation
- Behavioral analysis
- Understanding of what works; what doesn't

HOW TO USE BIG DATA

Big Data in higher education

- Most of us will never encounter Internet scale datasets, but we can use the same techniques to deliver value
- Big Data Concepts:
  - Merging data across information “silos”
  - Predicting individual behavior from group behavior
  - Handling “unstructured” data (e.g., tweets)
- What’s different in education:
  - Longer feedback loops (academic calendar)
  - Regulatory issues
  - More direct measurement and engagement opportunities

Most data analytics projects follow a standard approach

Define goal
- Work backwards
- “green lawn”

Collect data
- Establish system to collect data (75 % - 90 % of effort); collect **everything**
- “Soil ph; moisture; grass variety; rainfall; pests”

Analyze Data
- Iterative process; hypothesis testing

Focus on drivers
- Pareto is your friend
- Optimize the main drivers first (“water before fertilizer”)

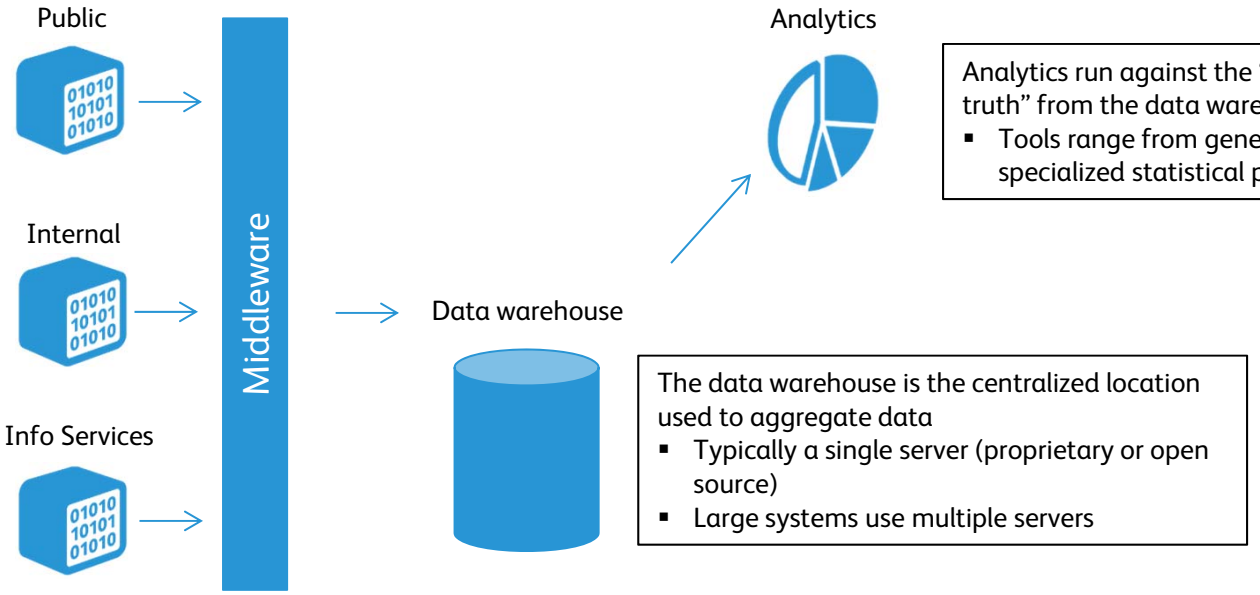
Understand the tradeoffs
- Are we operating under constraints? Is the level of effort to fix worse than the problem?
- “Water more; but I’m on a budget”

Determine actions
- The hardest part of any project is deciding what to do

Validate
- Be willing to be wrong
- “Neighbors dog”

Data Collection is at the heart of Big Data analysis – insights come from merging disparate datasets

Middleware is the interface that moves data between systems
- Includes web/ftp servers; integration servers; etc.

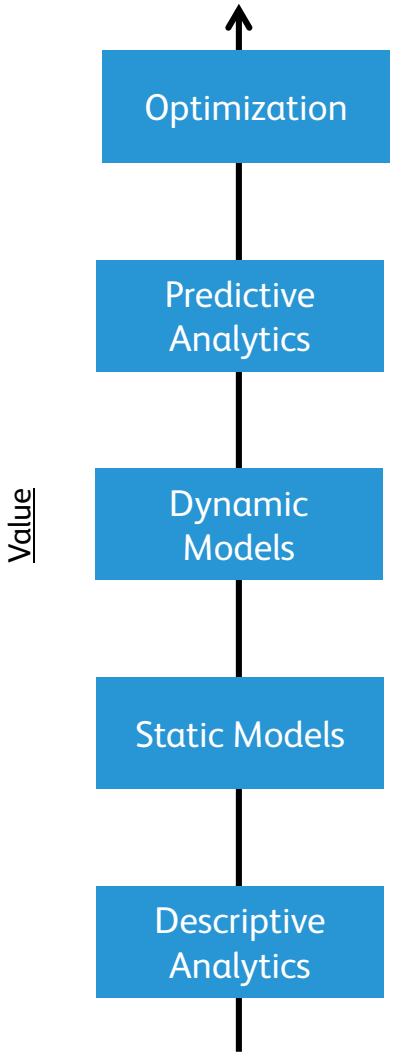


Analytics run against the "single version of the truth" from the data warehouse
- Tools range from general purpose (Excel) to specialized statistical packages (SPSS, R)

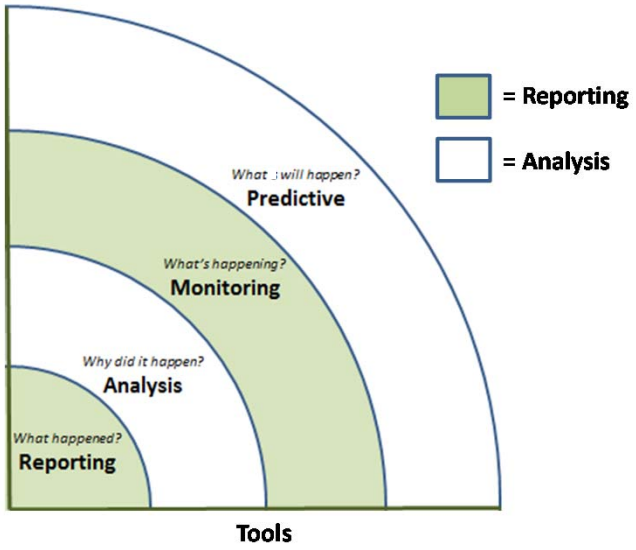
The data warehouse is the centralized location used to aggregate data
- Typically a single server (proprietary or open source)
- Large systems use multiple servers

Most systems have some type of data access
- Direct (ODBC; native driver; Web Service)
- Indirect (CSV, XML, fixed-width)

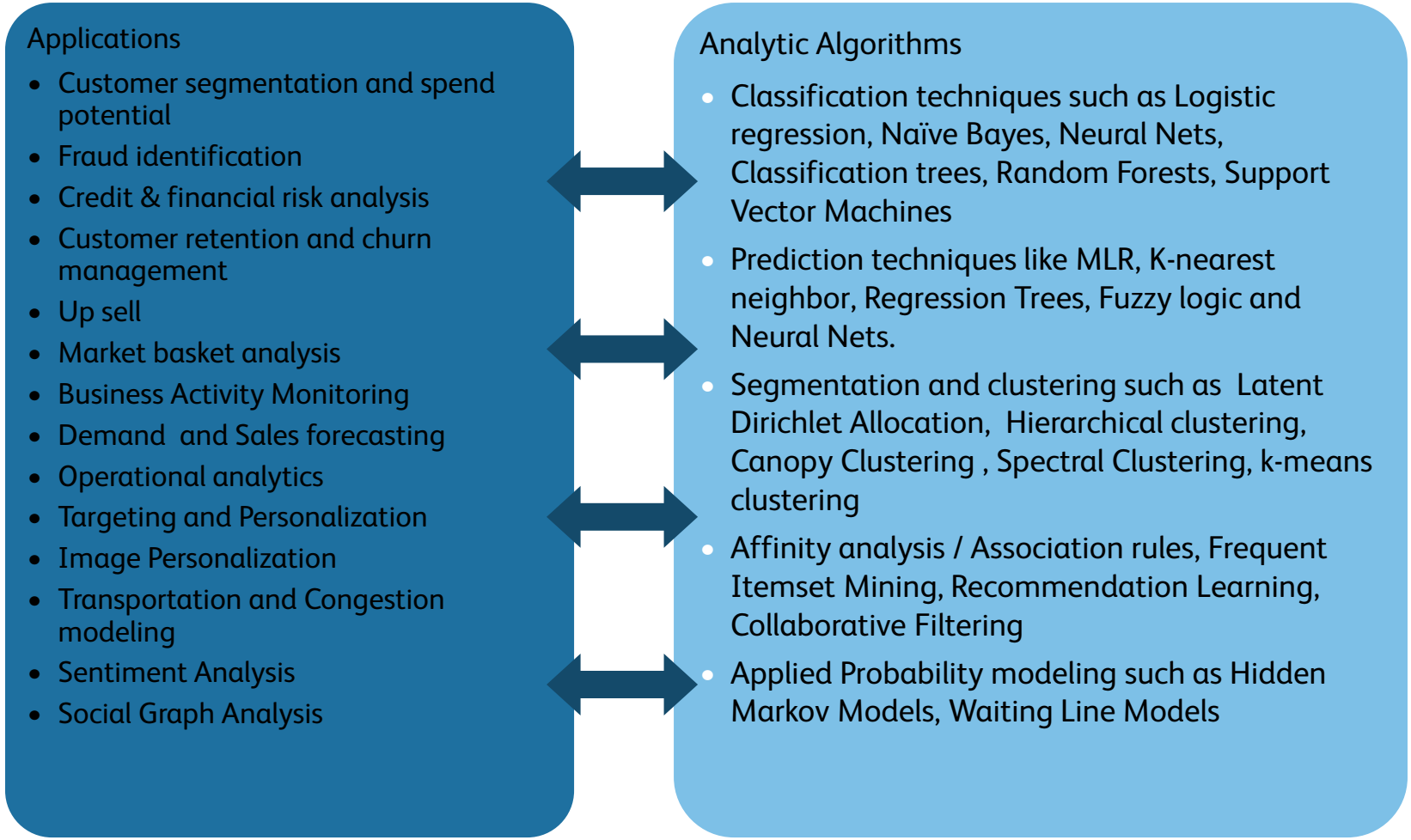
Types of Analytics



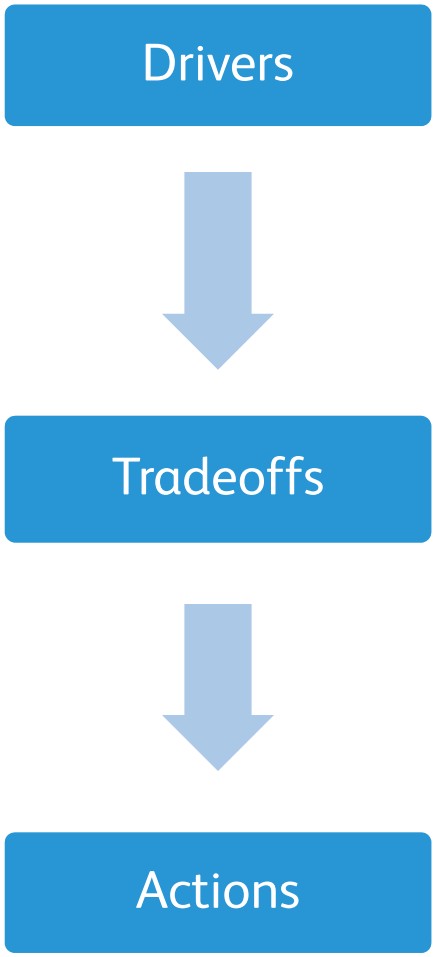
- Real-time data integration
- Optimization and trade-off algorithms
- Tools for decision support and recommendations
- Forecasting based on historic data
- Estimation algorithms
- What-if simulations
- Historical data analysis
- Multi-source data integration
- Reporting and data visualizations



Specialized Predictive Analytics



Drivers / Tradeoffs / Actions



- What are the main drivers? (80/20)
- Are they controllable or influenceable?

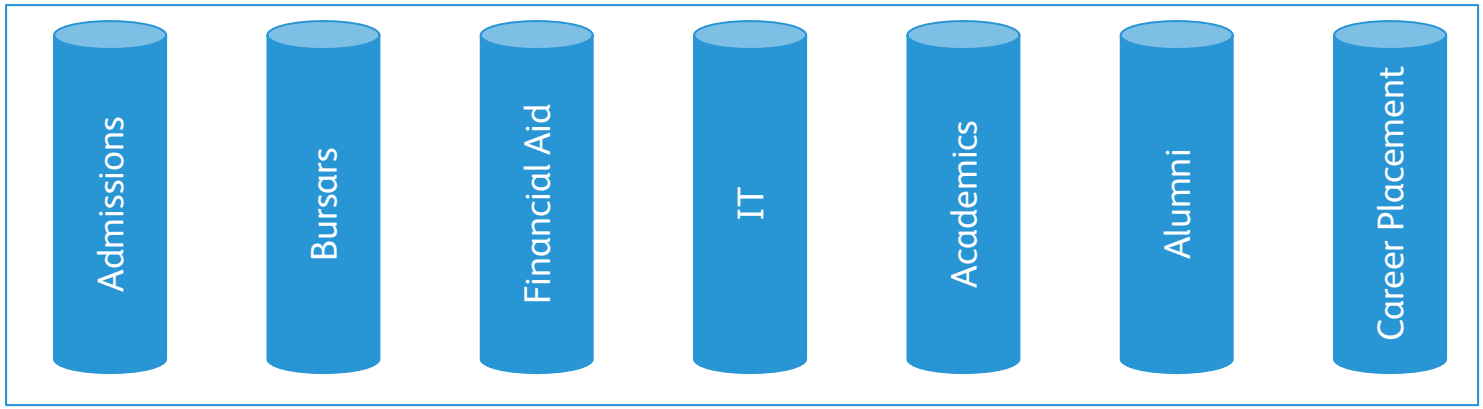
- Complexity/simplicity
- Cost
- Speed
- Internal/external
- Accuracy
- Sustainability

- Incentives matter
- Locally optimal doesn't always equal globally optimal
- Accountability

DATA AND PRIVACY

Available Data

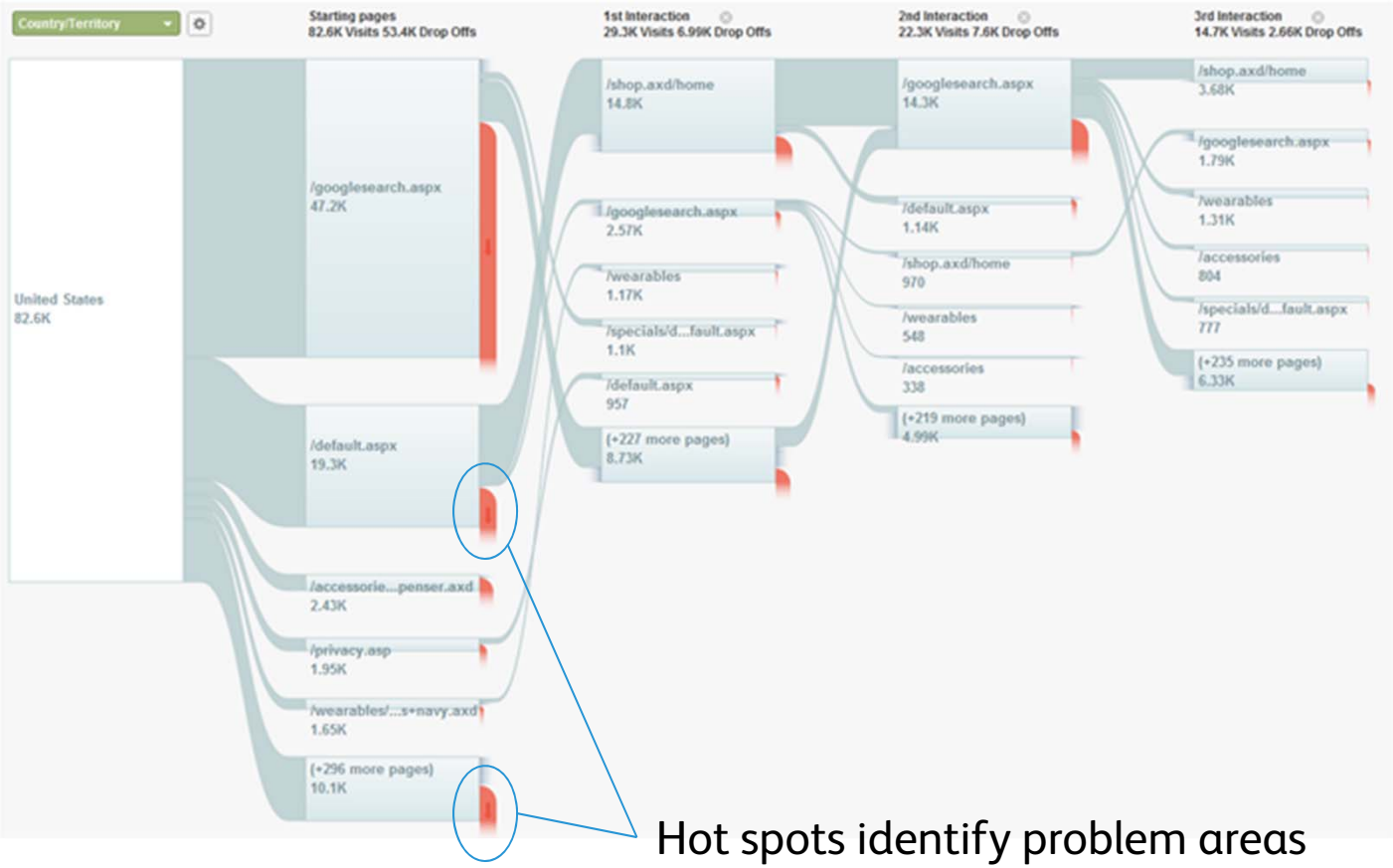
- School's collect a rich dataset on students, but it is typically separated by functional area
- Existing data can be analyzed to provide new insights



Data Elements	SAT/ACT H.S. GPA Rank Recommendations	Payment history	Fin Aid Package	Web logs Geolocation	Course roster GPA Major	Giving history	Employment
Meaning	Preparedness	Risk	Risk	Engagement	Effort	Loyalty	Risk

Path Analysis

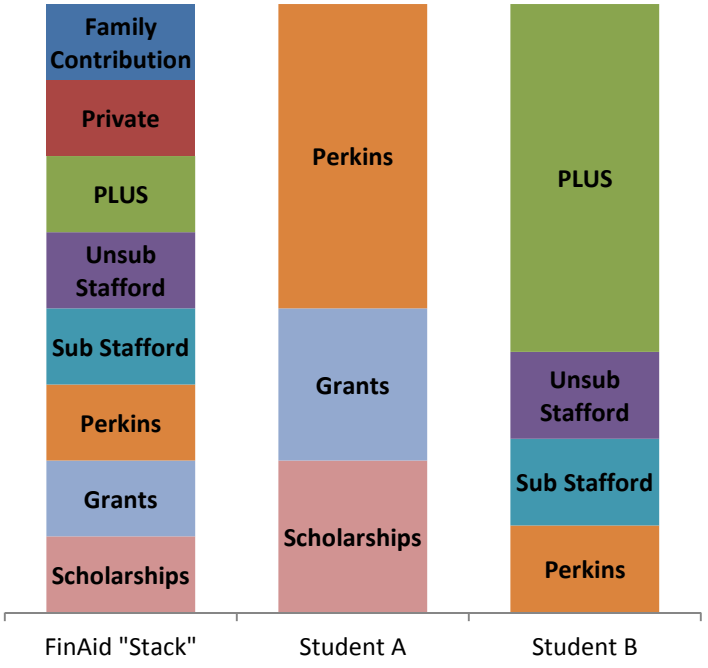
- Used to track how users navigate through a website (pages) to reach a goal (e.g. a purchase)
- Analogous to students taking a courses and reaching graduation



Student Financial Stress

- Financial Aid Offices collect information that can be used to model (predict) student financial stress:
  - Parent financial information
  - Financial Aid Package
- We know from survey data that financial stress is an underlying cause for drop out
- Define a "static model" for stress:

  Stress = $\gamma_s \omega_s + \gamma_g \omega_g + \cdots$
- Factors could be intuitive (stress factor for scholarships=0); defined from descriptive analytics; or dynamically derived from other inputs



- Stress scores could then be shared with faculty advisors; servicers; etc.

We have insights; can we use them?

- Privacy concerns
- FERPA
  - Without consent:
    - Directory information (students can opt-out)
    - Obtained through observation; opinions (risk is an opinion)
  - Undefined terms:
    - School official
    - Legitimate education interest
- Master Promissory Note:
  - DL: "(G) I authorize my schools, ED, and their agents to release information about my loan to each other."
  - Perkins: "The information in your file may be disclosed, on a case by case basis or under a computer matching program, to third parties as authorized under routine uses in the appropriate systems of records notices."
- Every school with have different policies: check with counsel

Details are nice, but some information is better than no information

Privacy concerns lead to “no” and “can’t” but there is usually a useful (and permissible) middle ground

- In California, 5-digit zip codes are considered PII (*Pineda vs. Williams-Sonoma*)
  - 3-digit zip codes can still provide demographic information
- Financial aid package information may not be shareable with servicers
  - Providing serviced loans as a percentage of total financial aid is useful
- Student records may not be shareable between professors
  - Measures of engagement, etc., can be teaching aids

WALKTHROUGH

Reduce Delinquency and Default

Goal: reduce the # of defaults in Perkins program

Collect data:

Admissions	Student Information	Loan History	Academic	Career	Exit Counseling

Analyze data:
We are trying to gain insights about the characteristics of defaulting borrowers, as well as their behaviors and contexts:

- Characteristics (age, starting salary, degree, etc.) already exist in the data
- Behaviors and contexts likely need to be calculated; may be more important in answering "why" or interpreting the results
  - E.g., what was the major vs. how many times did the major change?
  - How many payments did borrowers make before defaulting?
  - What was their starting salary after leaving school?
  - Did their GPA trend up or down during their academic career, or shift dramatically?
- Feature selection algorithms are then used to identify drivers

Reduce Delinquency and Default (cont'd)

Drivers:

- Let's assume that our feature selection algorithm identifies 3 drivers:
  - Math SAT score
  - Degree received
  - Starting salary
- This might give us new insights to explore and to add as data inputs.
- Are any of the drivers controllable? If not, we need to do further analysis.

Tradeoffs:

- We're resource constrained, but we've identified some factors that let us take preemptive action

Actions:

- Provide financial literacy training to low math SAT students receiving Perkins loans (we can't afford to do it for everyone, but we were able to target an at-risk subpopulation)
- Communicate the graduation data point to our Perkins servicer, and begin deeper analysis into our "transition out" points

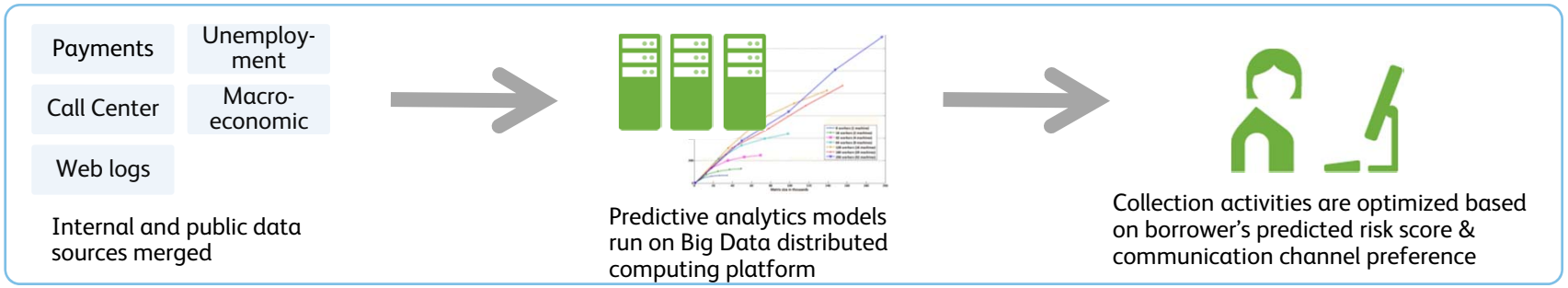
Validation

- Because of the long cohort cycle (6+ years from original disbursement to default), direct validation is difficult

WHAT WE DO

An Early Warning System for Borrower Risk

- **Problem**: The extended economic downturn causes repayment stress for borrowers in Xerox's $60B+ managed student loan portfolio. Scoring the entire portfolio using third-party credit bureau data not economically feasible.
- **Solution**: Xerox data scientists developed an early warning system for borrower risk using Big Data technologies and predictive analytics. The technology identifies future at-risk borrowers so that proactive measures can be taken to prevent delinquency and default.



ADVANTAGES

- Relies on free data sources
- Runs on low-cost computing hardware
- Extensible to new information (social web)

OUTCOMES

- High prediction accuracy
- Significant reduction in delinquency rates
- Lower collection costs

CONCLUSION

Conclusion

- Big Data is changing the analytics landscape
  - New tools and technology for Internet scale datasets
  - Most exist as open source, which democratizes analytics
  - Merging datasets to create a holistic picture is fundamental
- Schools have a rich dataset of student characteristics and behavior that can be mined
  - Many ways to analyze the data for meaning
  - Requires cross-departmental engagement
- Privacy is paramount
  - However, there are ways to share insights, without violating privacy
- Start with the goal in mind
  - How do we help student borrowers?
  - Iterate, iterate, iterate…

Q&A

Contact:
jeff.harris@xerox.com

APPENDIX

Tools for data analysis

	Proprietary	Open Source
Middleware	SAS SQL Server Integration Services	Talend
Data warehousing	Access (small datasets) SQL Oracle	MySQL MongoDB
Visualization	QlikTech Tableau HighCharts (free for schools)	Miso D3
Analysis	Excel SPSS SAS	R, Weka
Machine Learning		Mahout
Web server	IIS WebSphere	Apache Tomcat
Web portals	SharePoint	LifeRay Drupal