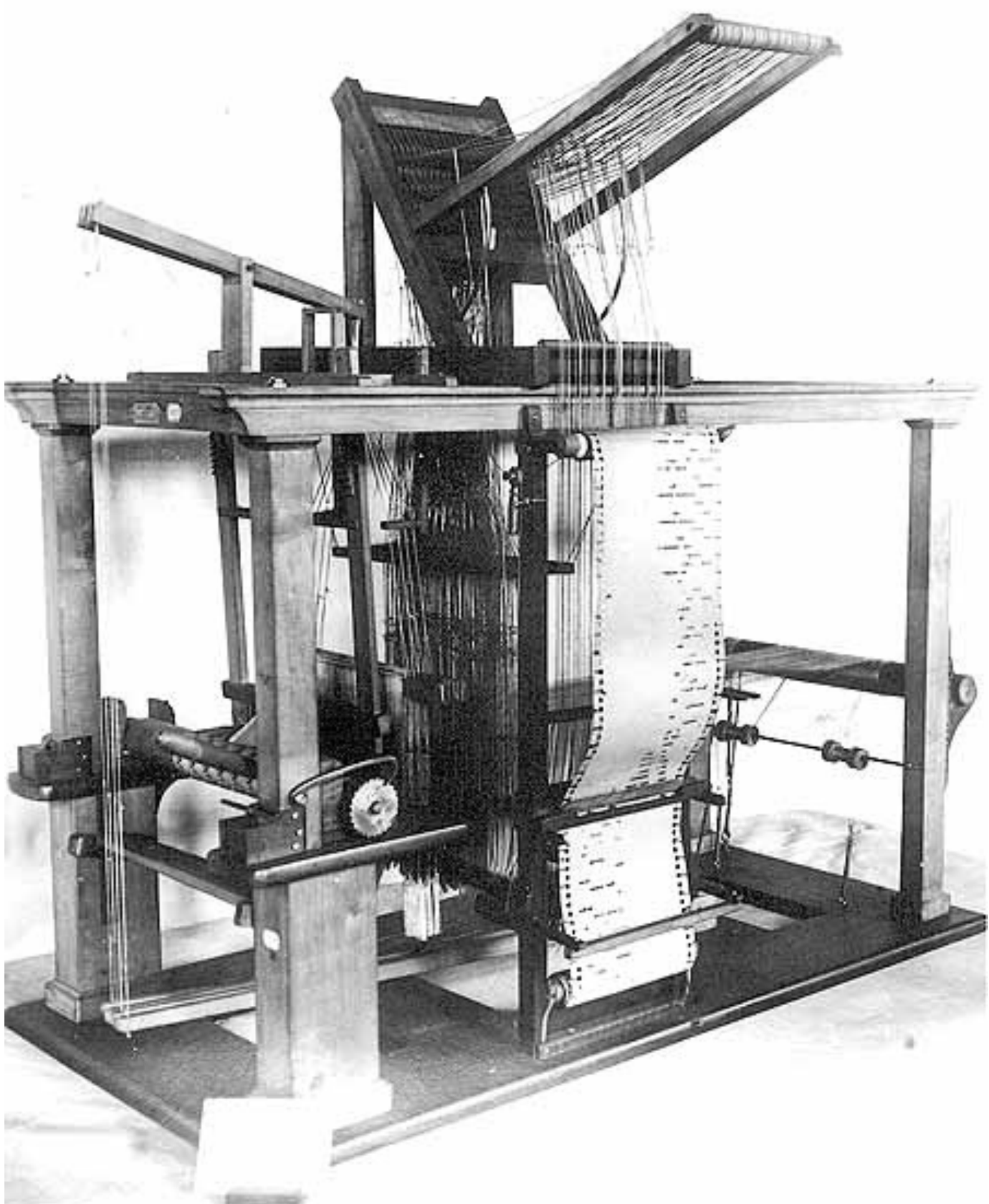
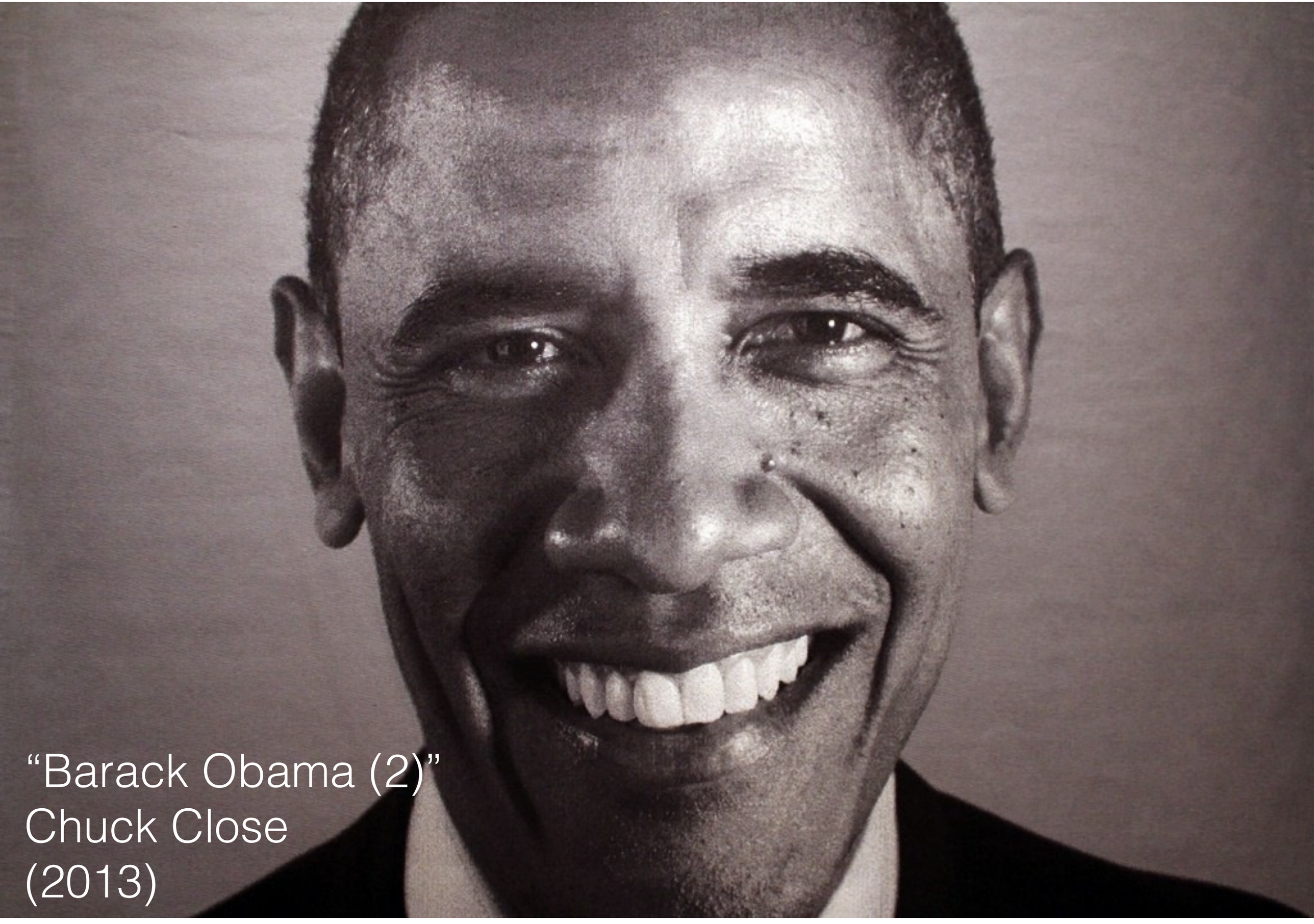
EDCT GE2550: DATA SCIENCE IN EDUCATION

Big Data, Learning Analytics & The Information Age

1/30/16 4:39 PM

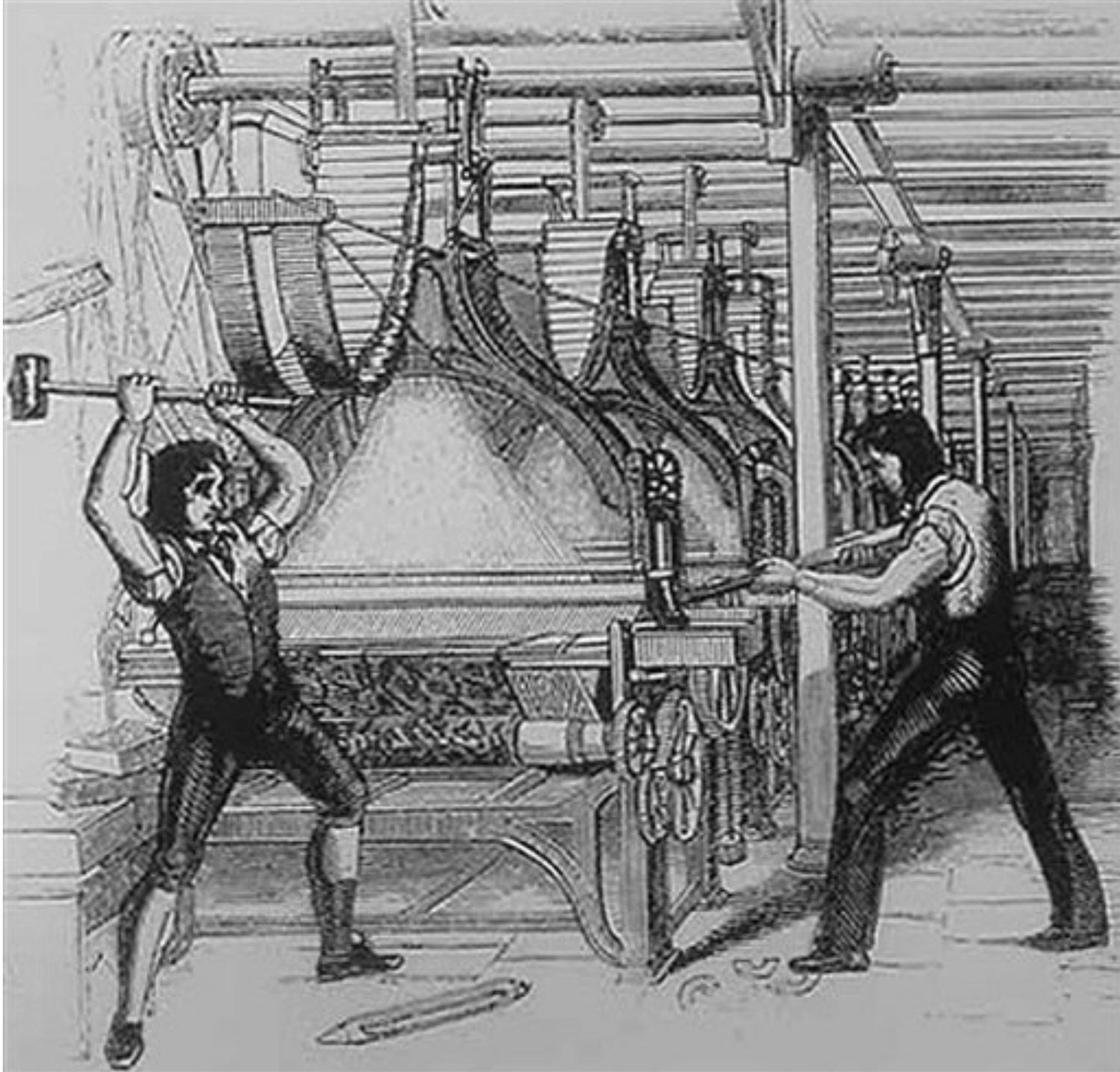
Jacquard Loom
(1801)





“Barack Obama (2)”
Chuck Close
(2013)

Luddites
(1811-17)



Tweets | Media | Likes

leonie haimson @leoniehaimson 2d

Treating kids like cattle @Parents4Privacy

> **Lori Lalama** @TechEducator1
>
> #DataMining #PrivacyAware @leoniehaimson @DianeRavitch wired.com/2012/09/rfid-c...

2

leonie haimson Retweeted

Sheila Resseger @sheilaresseger 2d

@leoniehaimson I had no illusions 2 the contrary! We r up against a relentless & callous assault on human dignity, preventing self-mpowrment

2

Mark Zuckerberg

December 3, 2015 at 5:24pm · Palo Alto, CA ·

I want to thank you all for your heartwarming congratulations on Max's birth and on starting the Chan Zuckerberg Initiative. This whole community has been so loving and supportive.

If you're interested in following the philanthropy work we're doing with the Chan Zuckerberg Initiative, I encourage you to like the page here:

https://www.facebook.com/chanzuckerberginitiative

Since we announced this a couple days ago, many people have asked about what we're planning to focus on and how we're structuring our work.

Our initial focus areas are personalized learning, curing disease, connecting people and building strong communities. We've already made many investments over the past five

1 THE RAPID GROWTH OF GLOBAL DATA

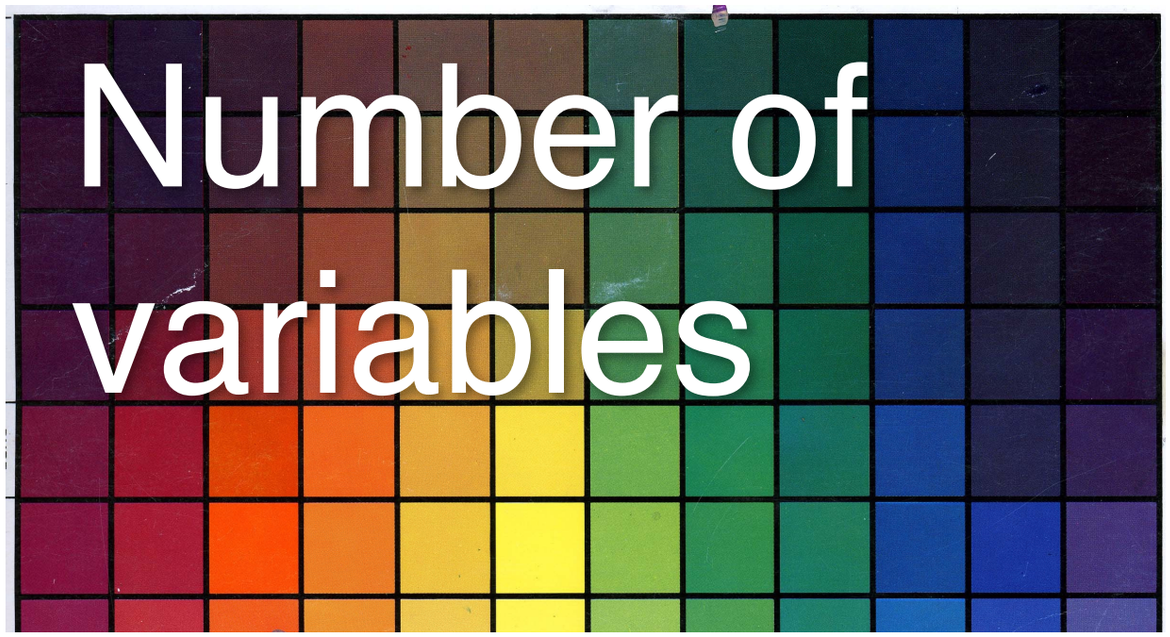
CSC

The production of data is expanding at an astonishing pace. Experts now point to a 4300% increase in annual data generation by 2020. Drivers include the switch from analog to digital technologies and the rapid increase in data generation by individuals and corporations alike.

Computer Sciences Corporation, 2012

What's new?


Number of
data points


Number of
variables


Track
Annotate
News
Zoom
Number of
time points
1999
2000
2001
2002
05/30/03
2004
2005
2006
2007
2008
2009
2010
2011
2012

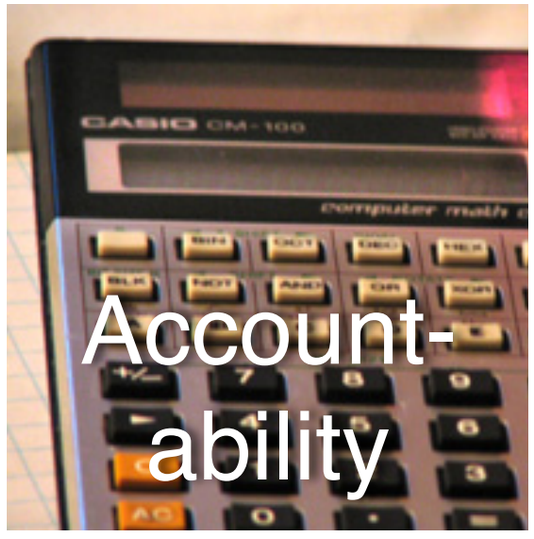

Feed-
back

MR. MESSY

by Roger Hargreaves

- Volume
- Velocity
- Variety
- Veracity

Possibilities



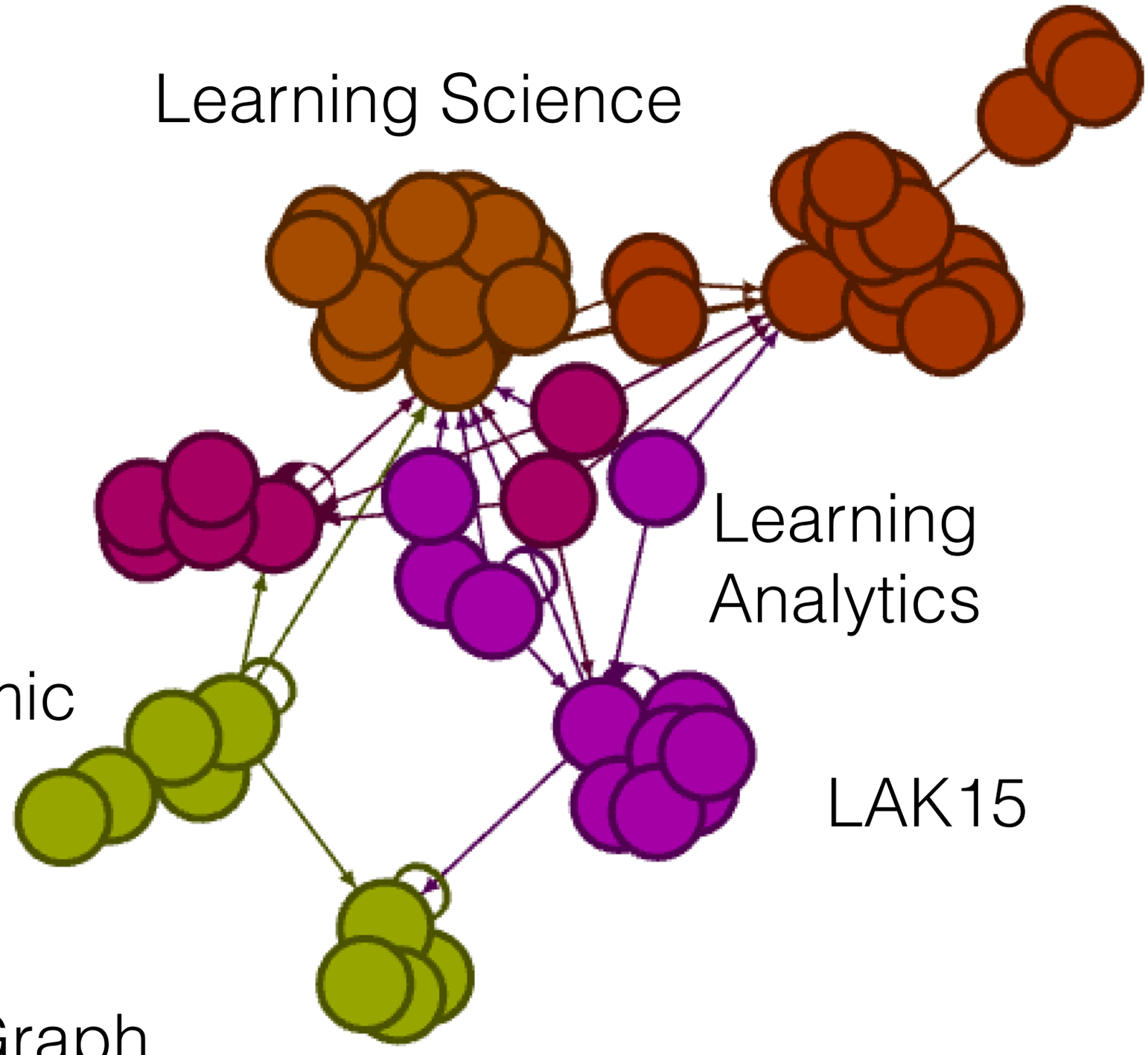
Computer Science

Learning Science

Learning
Analytics

EDM/Academic
Analytics

LAK15



Twitter Social Graph
#learninganalytics
2 weeks worth of tweets

Learning Analytics

- Grew from BI
- Early-mid 2000s in Canada
- Large domain umbrella - EDM, MOOCs, learning sciences, game design
- Systems oriented
- Methodologically broad (qual, quant, CS, stat)
- Journal: JAL
- Society: SOLAR

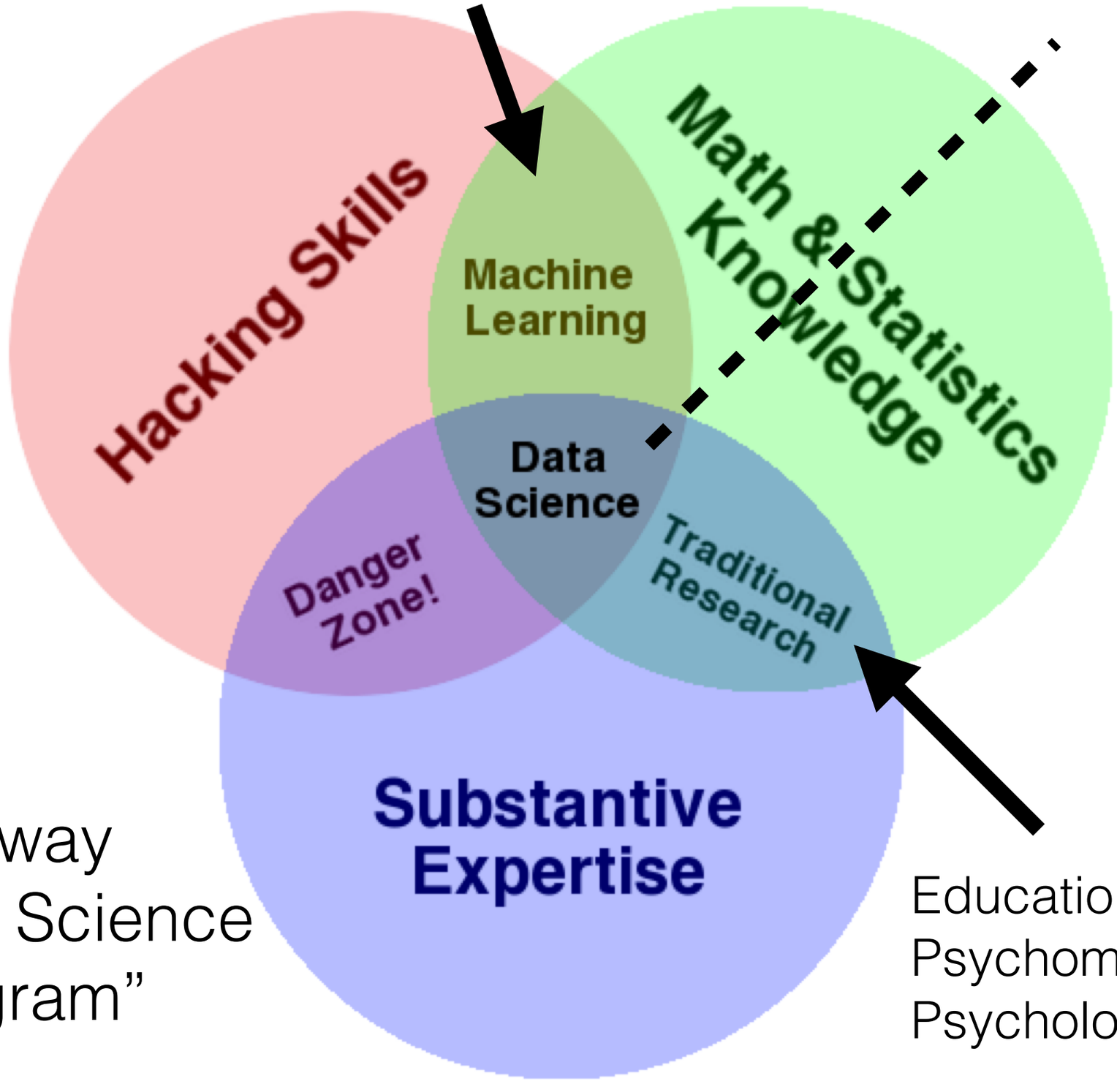
Ed. Data Mining

- Grew from Knowledge Mining or KDD
- Late 90s at CMU
- Intelligent tutoring, log-file
- Software oriented
- CS methods
- Journal: JEDM
- Society: IEDMS

“A Data Scientist is a statistician who lives in San Francisco”

–Sean Owen (Cloudera), 2014

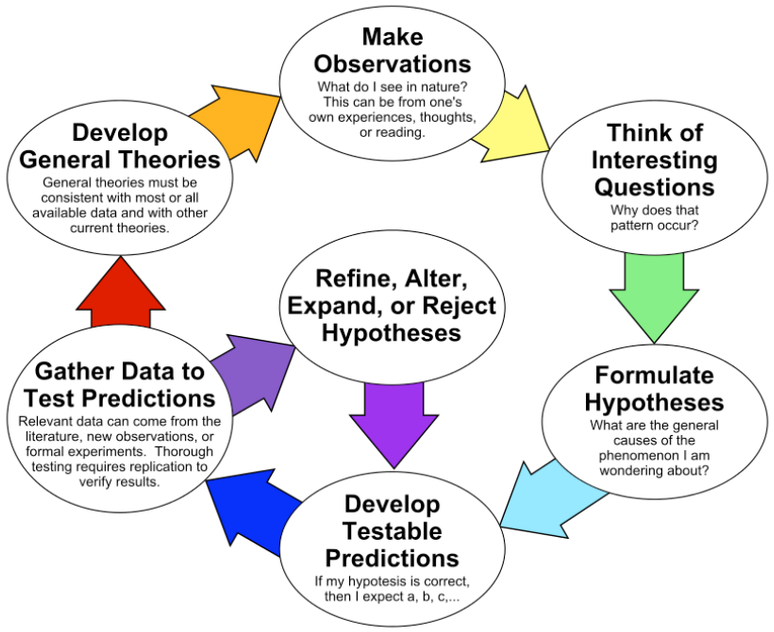
EDM
Academic Analytics



Drew Conway
"The Data Science Venn Diagram"
(2010)

Educational Statistics
Psychometrics
Psychology

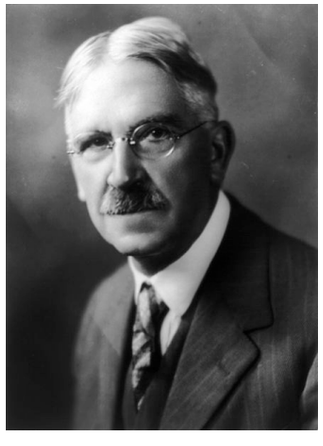
Philosophy of Science Spectrum



EDCT
GE-2550

Mr Vargas

Delores Etter



John Dewey

William James

Paul Feyerabend

Morty

Today

- Introduction to the Field(s)
- Candy
- Course Details
- Q&A
- Quiz
- Defining Educational Goals
- Educational Goals
- Analytic Strategy + Consent
- Pros & Cons of variables

Course Details

1/30/16 4:39 PM

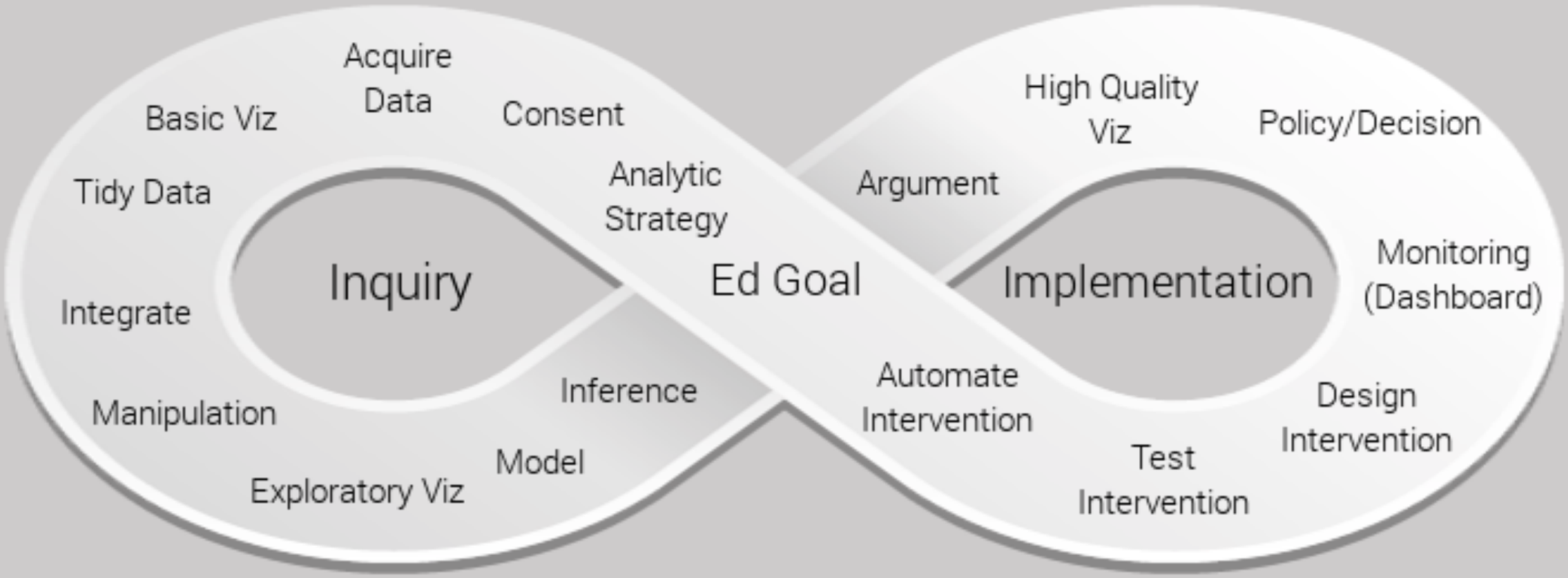
Overall goal

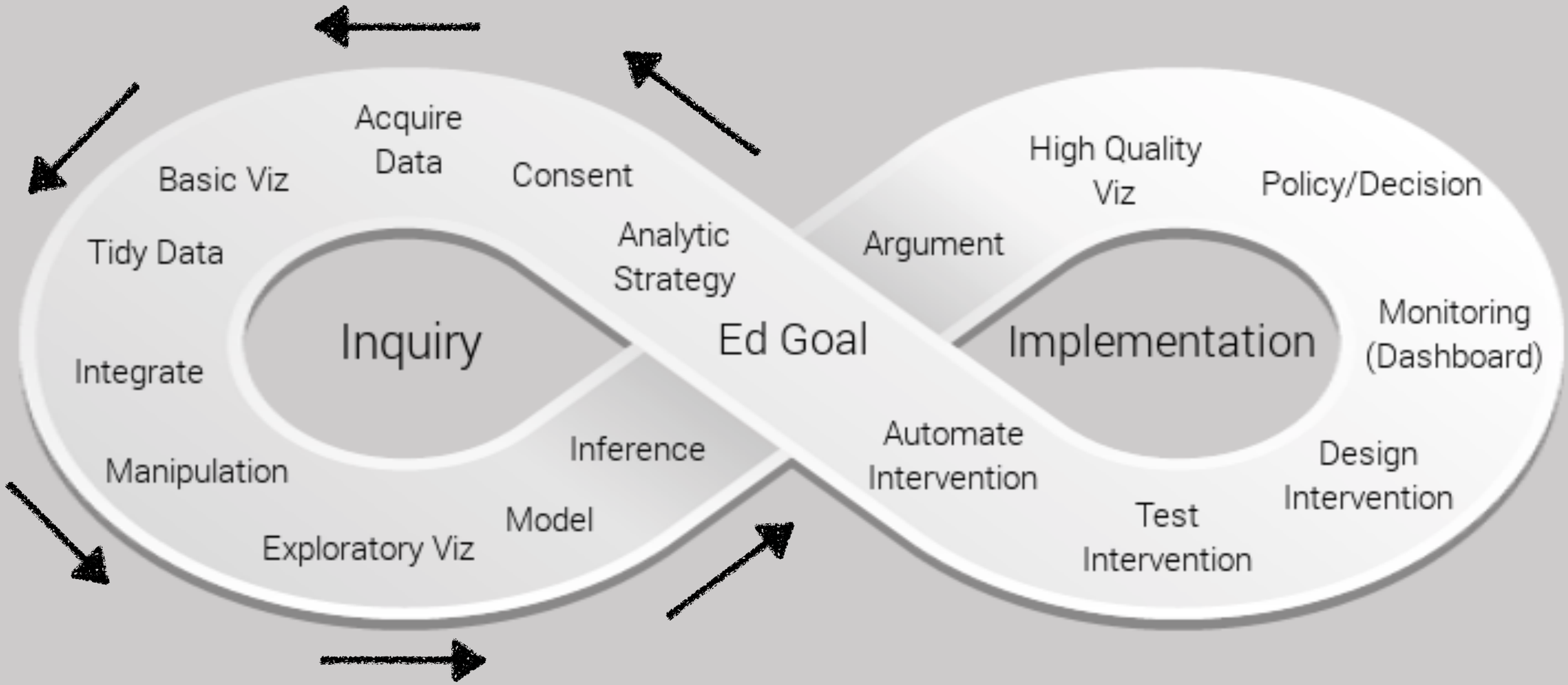
To investigate what the “Data Age” means for learning.

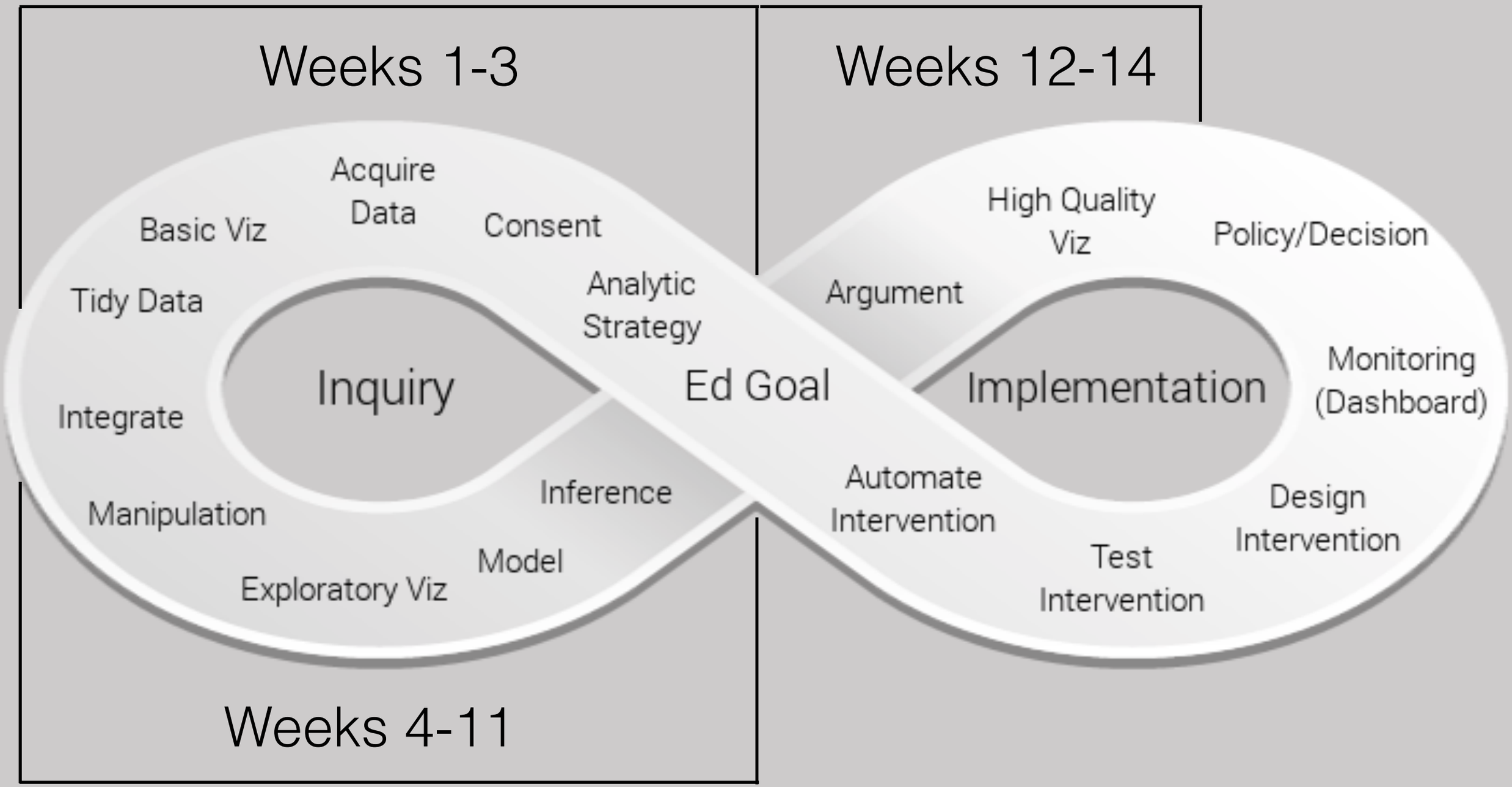
Course Goals

Short Term		Long Term	
Content	***Skills***	***Abstractions***	***Habits***
remember…, understand…	demonstrate…	synthesize…, argue…	organize…, implement…
Conceptual basis of methods, use cases	R, git, application of methods	Evaluate broader implications, have opinions, methods schema	Workflow, documentation

Ed Data Science Cycle




Acquire
Data
Basic Viz
Consent
Tidy Data
Analytic
Strategy
Inquiry
Ed Goal
Implementation
Integrate
Inference
Manipulation
Model
Exploratory Viz
High Quality
Viz
Policy/Decision
Argument
Monitoring
(Dashboard)
Automate
Intervention
Design
Intervention
Test
Intervention


Weeks 1-3
Weeks 12-14
Acquire Data
Basic Viz
Consent
Tidy Data
Analytic Strategy
Inquiry
Ed Goal
Implementation
Integrate
Inference
Manipulation
Model
Exploratory Viz
High Quality Viz
Policy/Decision
Argument
Monitoring (Dashboard)
Automate Intervention
Design Intervention
Test Intervention
Weeks 4-11

Methods

- Data tidying
- Dimensionality Reduction
- Social Network Analysis
- Prediction
- Natural Language Processing

What is the work?

- Attend class
- Weekly readings
- Comment on readings
  - Weekly in class questionnaire
  - Maintain documentation of work (Github, R Markdown, Zotero)
- Ask one question on Stack Overflow
- In person meeting with instructor
- 8 short assignments (including one group assignment)
- Group presentation of Assignment 8
- Produce one argument about learning using data from the class

Assessment

How to assess a course that is all about how difficult it is to measure learning?

- Assess your preparedness to do the work after the course has ended

- Two measures:

    - Contribution: assignments, comments, quizzes

    - Organization: keep a record of what you have done (Zotero, Git, Markdown)

Tools

- Git/Github 
- R/RStudio 
- Twitter 
- Zotero 

- Tools that are worth learning in & of themselves
- Tools that we can easily extract data from
- Tools that you can use unrestricted in the future

Git/Github

- Git is a common version control system
- Github is an online hosting service for Git files

We will be using Git/Github to:

- Keep track of work
- Document problems
- As an LMS (assignment submission, etc)

R/RStudio 

- R is a scientific programming language

- RStudio is an Integrated Development Environment

Why use R?

Pros

- Free
- Platform independent
- Actively developed by a large community of users
- Functionality is VAST
- Help resources are VAST
- Best graphics (At the moment)
- All the cool kids use it

Why use R?

Cons

- Slow (for some tasks)
- Not designed to build tools
- Relies on vectorization
- Non-intuitive structures & syntax with respect to other programming languages

Why use R?

It is designed for people who are *learning* a method *as they apply it*.

Twitter (#DSE16)

Three main uses for Twitter:

- Crowd source answers to your questions
- Comment on readings
- Backchannel in class (#IDKWGO)

Zotero 

- Bibliographic software
- Lives in browser (where you do research)
- Open source

Data

As much as possible we will be generating the data we analyze:

- Gain the experience of being on both sides of the data collection & analysis
- Easier to be invested in the content

But...

- Need to separate what we consider “assessment” and what we consider “activity”
- Need to put limitations on use (consent, time limit)
- Need to develop trust

Last word

If you are having trouble doing the work for whatever reason do not hesitate to reach out (the sooner the better):

charles.lang@nyu.edu

@learng00d

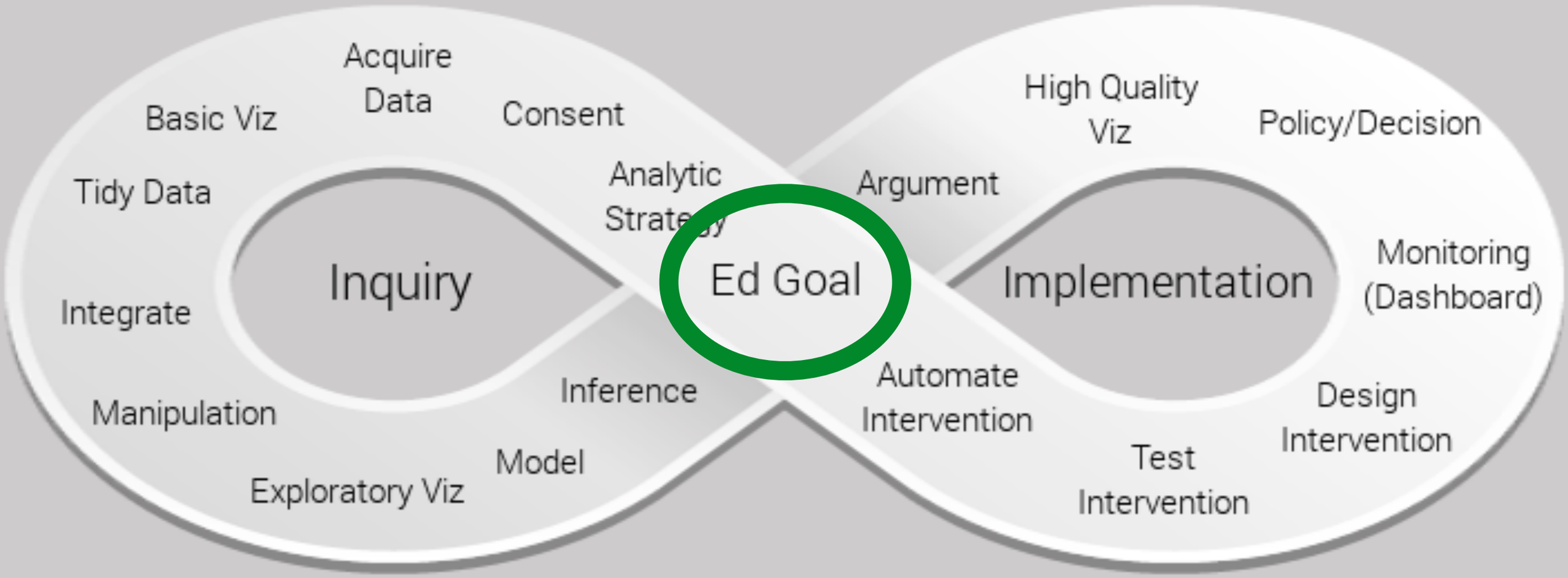
Syllabus

https://github.com/data-science-in-ed/Syllabus

Educational Goals

1/30/16 4:39 PM

Ed Data Science Cycle



“It’s better to be at the bottom of the ladder you want to climb than at the top of the one you don’t.”

–Stephen Kellogg, 2013

Google Books Ngram Viewer

Why?

- Education is controversial
- Concepts cross disciplinary lines
- Vocabulary is slippery
- Systems are complex

Characteristics

- Specific/tightly defined
- No jargon
- Measurable
- Attainable
- Relevant
- Time-limited

- Defined level of inference
- Contextual framework
- They can change (BUT keep a record of the change)
- Can be to answer a question (To determine the time of day students write the most words.)
- Needs to be workshopped with others

Activity 1

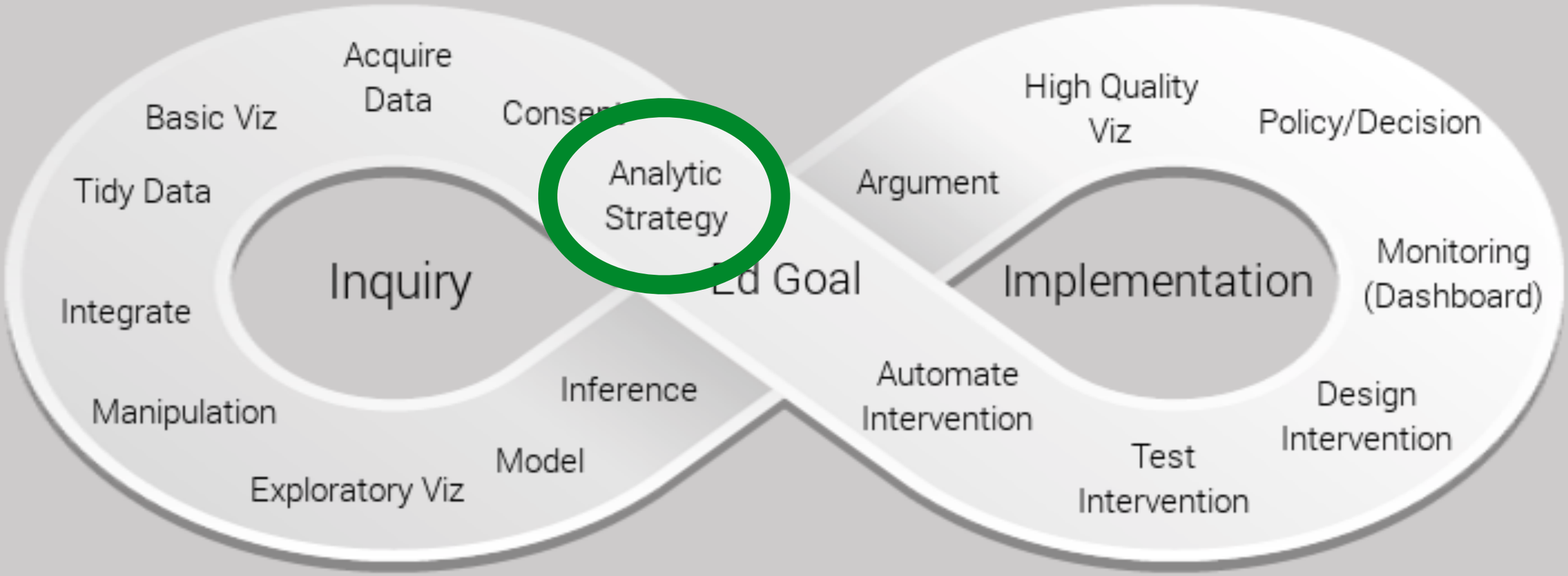
1. Find a partner
2. Determine one educational goal that you would like to define for yourself, for this semester

3. What information would you need to allow you to determine if you had met that goal

Analytic Strategy

1/30/16 4:39 PM

Ed Data Science Cycle



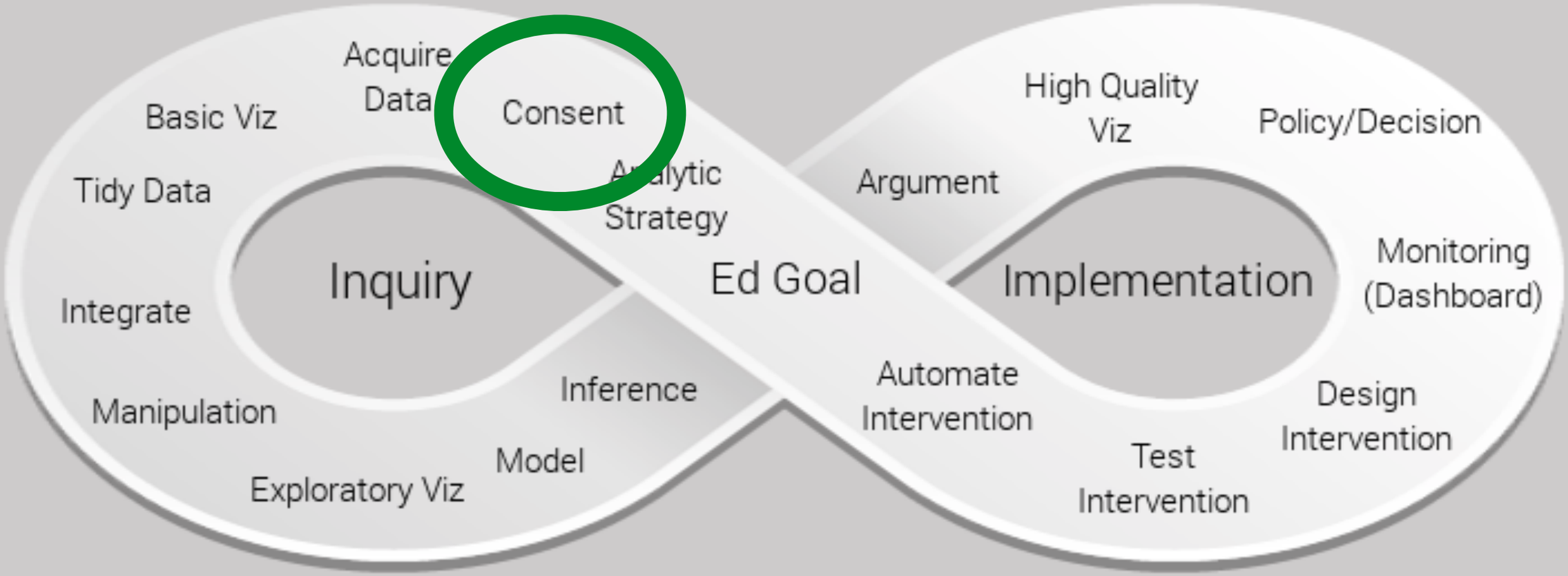
Analytic Strategy

- Going to use a range of methods

- Data will be erased at the end of the course

Consent to Use Data

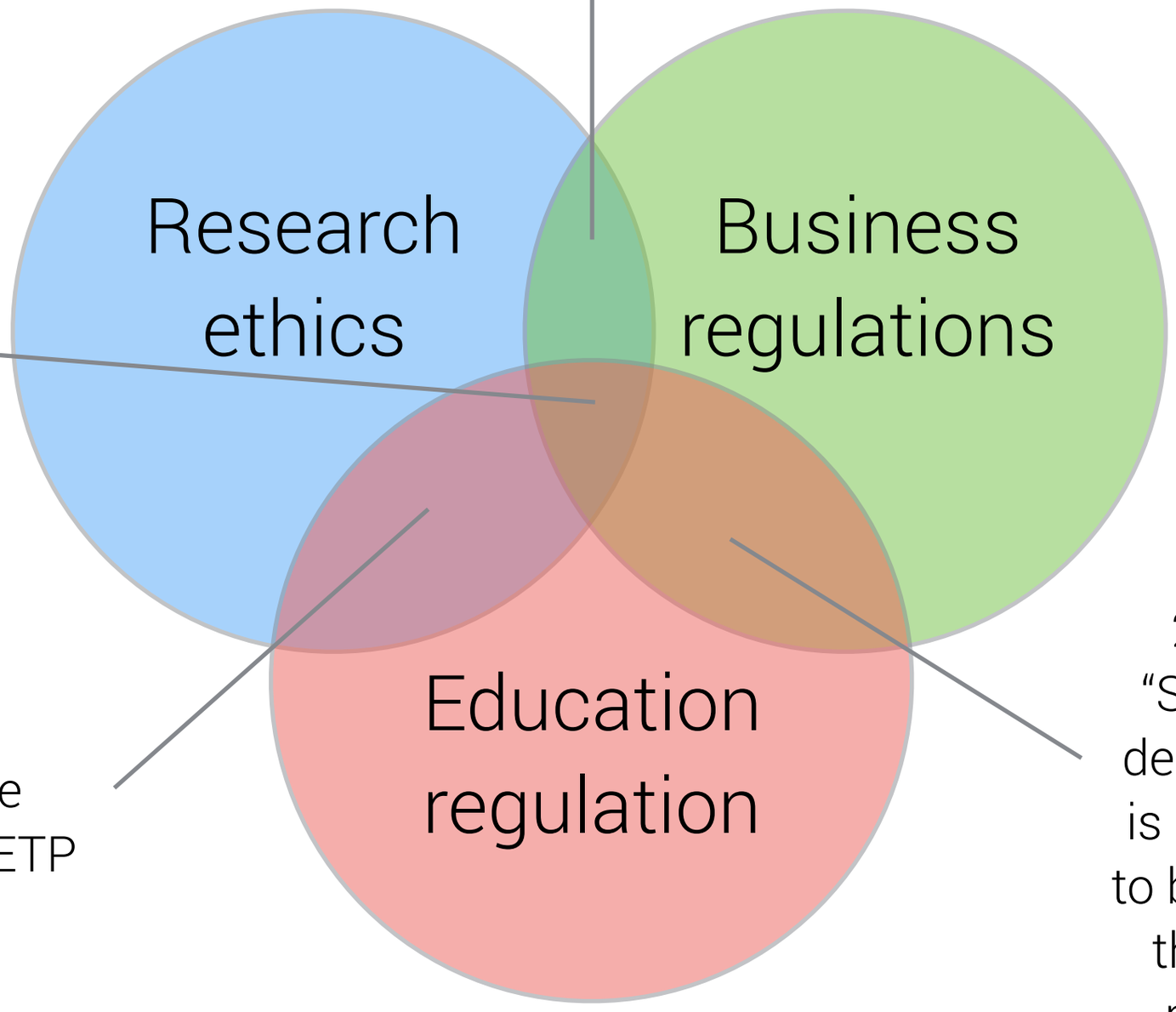
Ed Data Science Cycle



Consent

1. Disclosure
2. Capacity
3. Voluntariness

Class of Civilizations



"We were concerned that exposure to friends' negativity might lead people to avoid visiting Facebook." Adam Kramer (Facebook), 2015

???????

Non-cognitive factors & the NETP (2015)

2008, FERPA "School Official" designation. Data is both supposed to be controlled by the school and proprietary to Google

What are people used to?

Cognitive Dissonance

- Transparency vs. obfuscation
- Siloed information vs. linked functionality
- Defined use vs. mine for insight

Transparency vs. Obfuscation

Terms of Service	Privacy Policy	Community Guidelines

No individual under the age of thirteen (13) may use the Services, provide any personal information to Tumblr, or otherwise submit personal information through the Services (including, for example, a name, address, telephone number, or email address). You may only use the Services if you can form a binding contract with Tumblr and are not legally prohibited from using the Services.

You have to be at least 13 years old to use Tumblr. We're serious: it's a hard rule, based on U.S. federal and state legislation. "But I'm, like, 12.9 years old!" you plead. Nope, sorry. If you're younger than 13, don't use Tumblr. Ask your parents for a Playstation 4, or try books.

Siloed Information vs. Linked Functionality

(2011)

Generation of "Popularity" Indexes

(2016)

Defined Use vs. Mine for Insight

- Restrict use = lose insight
- “Educational use” might be useless
- EG - Time limits on user profiles

Activity 2

1. Get into candy-based groups (Snickers, Milky Ways, etc)
2. Each member of the group is assigned a role:
    - Student = snickers, teacher = MW midnight, parent = MW Regular, ed tech CEO = 3 Musk, researcher = Twix
3. For your given role discuss pros & cons for students allowing the type of data assigned to your group to be collected in an educational setting for the purpose of *improving learning*

Activity 2

1: Environment measures (eg. room temperature)

2: Formative assessments (eg. weekly quizzes)

3: Location data (eg. location in building)

4: Biometric data (eg. Fitbit data)

5: External social website data (eg. Twitter/Facebook)

6: Demographic information (eg. gender, age)