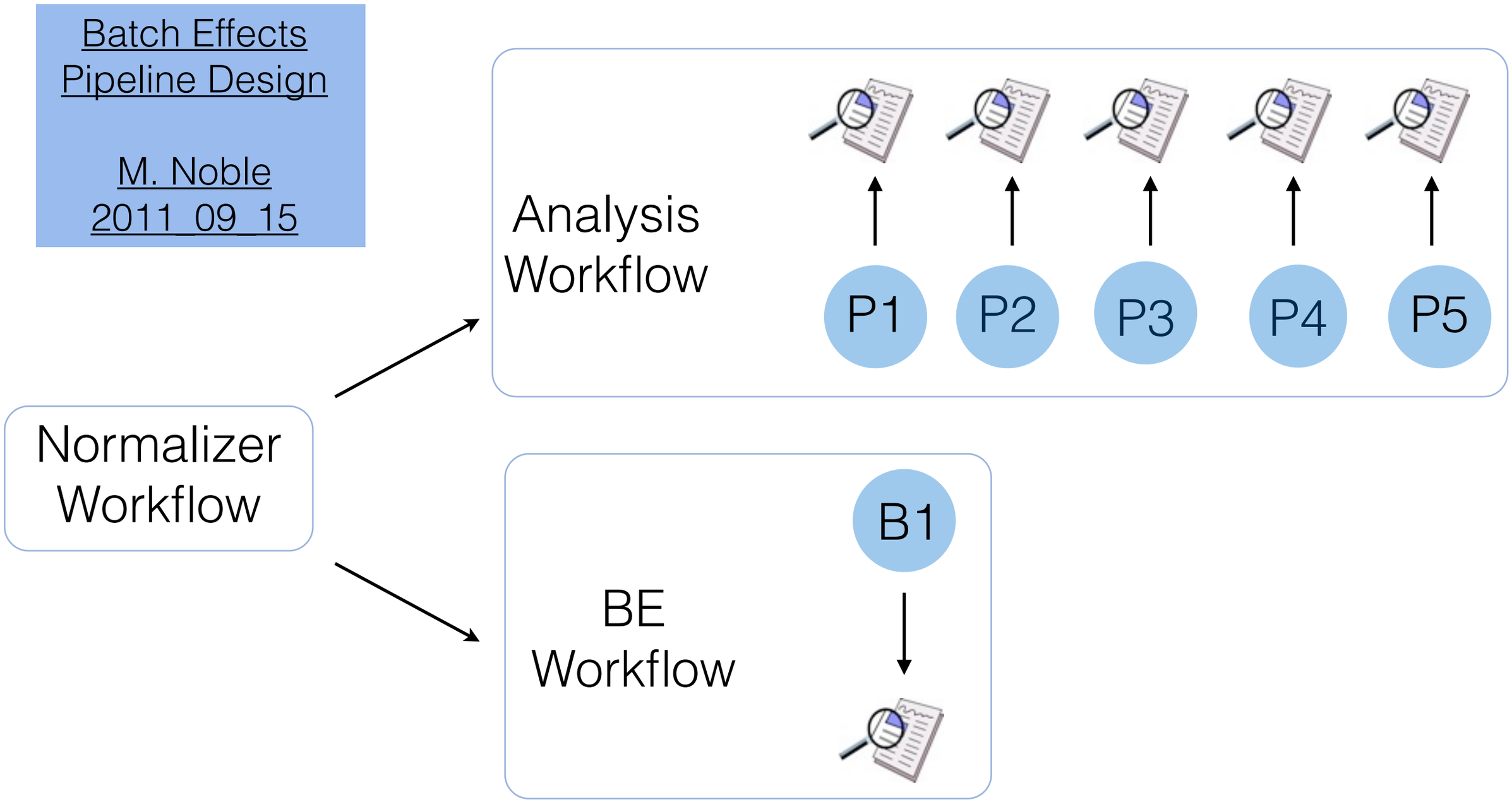
Batch Effects
Pipeline Design

M. Noble
2011_09_15



Scheme 1

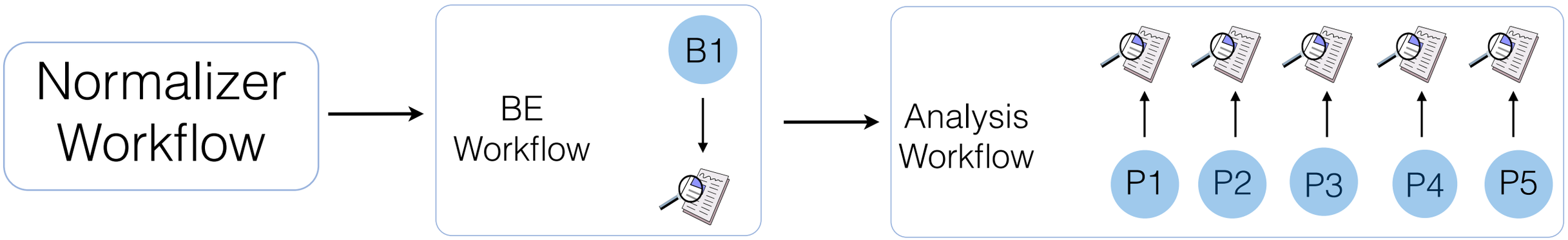
BE run outside of normalizer, in parallel with analyses
BE not flagged as additional column(s) in data
(yields only summary report, “off to the side”)

Cons

analyses do not know data flagged
manual cross-check with BE report needed for each pipe?
consumers of pipe outputs have to replicate cross-check?
scalability issue: 20-ish pipes X 22-ish tumor sets

Scheme 2

BE run prior to analyses
BE flagged as additional column(s) in data
(in addition to summary report, “off to the side”)



Pros

analyses NOW KNOW data has been flagged
reducing need for manual cross-check with BE report per pipe?
ditto for consumers of pipe outputs
addresses scalability issue: can be automated

Cons

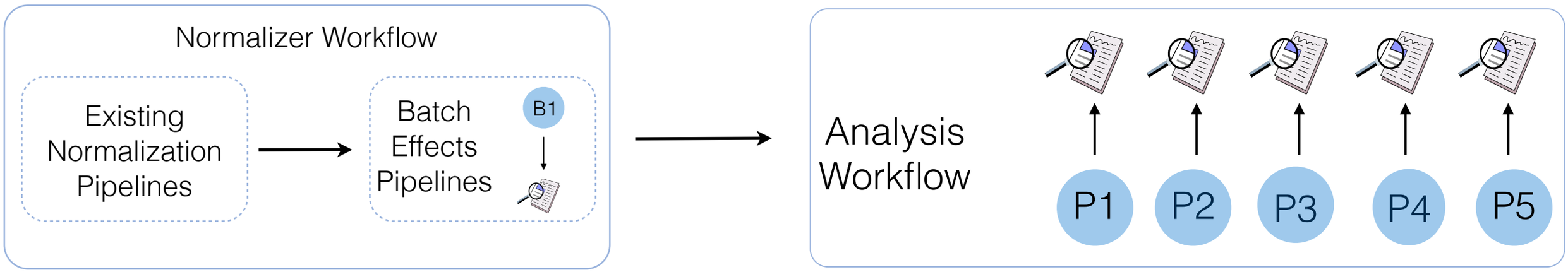
Because outside of normalizer, version-stamped datasets produced by FH NOT BE-flagged

Greatly reducing utility/scope of BE detection.

Scheme 3

Like Scheme 2, BE run prior to analyses
BE flagged as additional column(s) in data
(in addition to summary report, “off to the side”)

But DONE IN NORMALIZER



Pros

All of Scheme 2

But ALSO makes BE flagging more widely accessible

Not just Standard Analyses outputs, but also version-stamped, normed-data

Example: cBIO portal can be “aware” that data are BE-flagged